A DISSERTATION FOR THE DEGREE OF DOCTOR OF PHYLOSOPHY

Genomic and phenotypic characterization of sesame germplasm for phylogenomics and breeding

By

YEDOMON ANGE BOVYS ZOCLANCLOUNON

MAJOR IN CROP SCIENCES

DEPARTMENT OF CROP SCIENCES

COLLEGE OF AGRICULTURAL AND LIFE SCIENCES

GRADUATE SCHOOL

JEONBUK NATIONAL UNIVERSITY

2023. 02. 22.

Under the Guidance of Dr. Jeong-Gu, Kim & Professor Dr. Youngjun, Mo

Genomic and phenotypic characterization of sesame germplasm for phylogenomics and breeding

By

YEDOMON ANGE BOVYS ZOCLANCLOUNON

MAJOR IN CROP SCIENCES

DEPARTMENT OF CROP SCIENCES

COLLEGE OF AGRICULTURAL AND LIFE SCIENCES

GRADUATE SCHOOL

JEONBUK NATIONAL UNIVERSITY

2022.09.28.

A Thesis Submitted to Jeonbuk National University

In Partial Fulfillment of the Requirements

For the Degree of DOCTOR OF PHYLOSOPHY

Genomic and phenotypic characterization of sesame germplasm for phylogenomics and breeding

Committee Chair Dr. Nam-Jin, Chung

Committee Vice-Chair Dr. Byoung Ohg, Ahn

Committee Member Dr. Yeisoo, Yu

Committee Member Dr. Jeong-Gu, Kim

Committee Member Dr. Youngjun, Mo

APPROVED AS QUALIFIED DISSERTATION OF YEDOMON ANGE BOVYS ZOCLANCLOUNON FOR THE DEGREE OF DOCTOR OF PHYLOSOPHY BY COMMITTEE MEMBERS

2023.1.3.

Genomic and phenotypic characterization of sesame germplasm for phylogenomics and breeding

Name: Yedomon Ange Bovys Zoclanclounon

Department: Crop Sciences

Advisors: Dr. Jeong-Gu, Kim & Prof. Dr. Youngjun, Mo

Abstract

Accelerating the Sesame (*Sesamum indicum* L.) breeding in Korea requires the development of genomic resources as well as a set of pre-breeding core accessions. While Korea is the third country in the world, after India and China, with a high number of sesame genetic resources, active genome-assisted breeding is still slowing down due to the lack of high-quality genomic resources. In the last decade, sesame enters in genomics era with the sequencing of Chinese cultivars (Mishuozhima, Baizhima) and varieties (Yuzhi11, Zhongzhi13). This enables the discovery of major genes involved in sesame oil biosynthesis, drought stress-responsive genes, and metabolic pathways of specialized metabolites including sesamin and sesamolin. Despite this tremendous achievement, the application of these genomic results is hampered by the absence of a native Korean cultivar's genome project. Therefore, under the initiative of the Korea Genome Project, a high-quality genome project of the Korean cultivar *S. indicum* var. Goenbaek, an oil and lignan rich sesame, has been launched. Meanwhile, the origin of the diversity and the ancestor of the cultivated sesame is still a mystery since sesame is known as the oldest oil crop in the world. Thus, wild relatives constitute important resources since they harbor genes of interest that can help to prevent biotic and abiotic stresses, boost oil, and nutraceutical components including sesamin, sesamol, and sesamolin. With the increasing number of patents regarding specialized metabolites from sesame with health-promoting effects, it is urgent to produce sufficient data to understand the evolutionary history behind sesamum speciation and develop genetic resources for sesame breeding in the Republic of Korea. Our thesis covers a broad spectrum by aiming at (i) the construction of a core collection sesame accessions that can be useful for sesame breeding; (ii) generating the first chromosomal-scale genome assembly and annotation of the oil and lignan-enriched Korean cultivar Goenbaek; (iii)

identify potential enzymes involved in upstream step of lignans biosynthesis pathway; (iv) investigating the fundamental evolutionary basis of the speciation in the Sesamum genus. Firstly, a large panel of 506 sesame accessions has been screened for agronomic, oil, protein, and lignans traits. A core collection of 102 accessions covering 35 countries has been proposed. Ultimately this core collection could serve as material for genome-wide association analysis for the discovery of genes of interest regarding yield, oil, protein, sesamin, and sesamolin parameters. Knowing that a reference genome is required for active genome-assisted breeding, we, secondly took advantage of long reads sequencing and chromosome conformation capture technologies to construct a chromosome-level genome assembly of the Korean cultivar Goenbaek. By a comparative genomic and transcriptomic approaches, species-specific genes cluster coding for oil and specialized metabolites have been pulled out. Therefore, in-depth investigation of this gene pool could allow the design of oil boosted or lignans-boosted genotypes via gene editing strategy. Meanwhile, with the intention to contribute to the basic background behind sesamum speciation, comparative plastomics study has been carried out on a set of six African native sesamum species. The results showed that the speciation in sesame is likely linked to two major events including ascending disploidy and recent hybridization. The first event (ascending disploidy) leads to the creation of new species (*Sesamum angolense*, *Sesamum pedaloides*, *Ceratotheca sesamoides*, and *Ceratotheca triloba*) with 2n = 2x = 32 from the common ancestor (2n =2x = 26). The second event (hybridization) occurred relatively recently with the formation of species with 2n = 2n = 64 (*Sesamum radiatum*). Moreover, we clearly identified the genome A and B of the hybrid *S. radiatum* as *C. sesamoides* and *Sesamum angolense* respectively. Altogether, we provided in the current work, sufficient foundational data to dive into genome-assisted breeding in sesame for yield, oil, protein, sesamin, and sesamolin breeding.

On the fundamental research side, we provided for the first time, the evidence of key master events that drive the speciation in the Sesamum genus. The impact of these resources will dramatically impact sesame breeding in Korea by boosting the design of new varieties that combine both agronomic and health benefits.

Contents

List of Tables

List of Figures

Introduction

Sesame (*Sesamum indicum* L.) is widely consumed in the Republic of Korea. It is used as recipe ingredients and its oil served for food preparation. Moreover, due to its richness in antioxidants content, sesame is considered as a nutraceutical crop and anti-aging dietary food (Rodríguez-García et al., 2019). In fact, sesame seed is known to be rich in oil (44~58%) (Alyemeni et al., 2011), lignans including sesamin, sesamolin, and sesamol; all as health benefits metabolites (Andargie et al., 2021). In fact, lignans from sesame have exhibited promising properties for coping with cardio vascular disease (Dalibalta et al., 2020), prostate and breast cancer (Liu et al., 2006). Moreover, sesame seed oil is rich in fatty acids (96%) including palmitic (11%), linoleic (35%), oleic acid (43%), and stearic acids (Elleuch et al., 2007).

Owing the importance of this crop in Korea, tremendous efforts have been done to create a genetic resource that can be used for sesame breeding. At early stage, Kang et al., (2006) also employed 2,246 accessions to extract a core collection (475 accessions) based on 12 agronomic traits from a collection originated from ten Korean agro-climatic zones. Asekova et al., (2018) worked on 129 Korean landraces and cultivars to investigate their genetic variability through simple sequence repeats (SSR) markers. Besides, Park et al., (2015) assessed a set of 2,751 world-wide sesame accessions collected from 15 countries using 15 phenotyping traits. Thus, a core collection of 278 accessions have been created and maintained in Agrobiodiversity Center of National Institute of Agricultural Sciences. However, this previous study only took into account one African country ie Egypt. Meanwhile, several wild species including *S. radiatum*, *S. pedaloides*, *S. alatum*, *S. sesamoides syn. Ceratotheca sesamoides*, and *S. trlobum syn. Ceratotheca triloba*, *S. schinzianum*, are known to be African native (POWO, 2022). Therefore, it is worth

including African accessions to represent appropriately a realistic genetic pool that cover the African continent. Although disputes between scientists confronted Africa and Indian subcontinent as origin of the cultivated sesame *S. indicum*, the evidence from taxonomic expeditions revealed that 98% of accepted nomenclature in sesame genus are originated from Africa (Dossa et al., 2017b). In the current era of genomics, phylogenomics study offers a path to get insights into the phylogenetic relationships between wild relatives and the cultivated species.

Despite the availability of five genomes from others countries (China and India) (Yu et al., 2019b), there is a lack of sufficient genomic resources that can be used for assisting the Korean sesame breeding program. Meanwhile, Kim et al., (2018) released an elite Korean variety containing high oil, fatty acid, and lignans constituents. *S. indicum* var. *Goenbaek* exhibited disease tolerance against phytophtora disease and have been recently used to map a gene of interest regarding its resistance (Asekova et al., 2021). Nevertheless, the lack of Goenbaek genome oriented the scientists to rely on the Chinese cultivar reference genome for quantitative trait loci (QTL) and genome-wide association analysis (GWAS) investigation. Thus, the necessity to create a high-quality chromosome level genome of a representative and agronomically important Korean genotype is highly important for accurate and fast molecular breeding. Therefore, it urged to construct a genomic resource database for accelerating the sesame breeding for yield and yield related traits as well as health promoting metabolites.

Taking into account the previous observations, several objectives were proposed: (i) develop a representative genetic pool of sesame accessions integrating African countries that can serve for breeding purpose; (ii) generate a high-quality genome assembly of an oil- and lignan-rich variety named Goenbaek; (iii) find out candidate genes potentially involved in

lignans pathway; (iv) investigate the evolutionary background underlying the speciation within sesamum species.

The following sections are divided according to the three global objectives here above mentioned. We presented: (i) a review relative to the sesame order (Lamiales) genome projects that have been conducted so far; (ii) an assessment of agronomic traits and construction of core collection from a worldwide panel (Chapter 1); (iii) a high-quality chromosome-grade genome assembly of *Sesamum indicum* var. Goenbaek (Chapter 2); (iv) characterization of peroxidase and laccase gene families and *in silico* identification of potential genes involved in upstream steps of lignan formation in sesame (Chapter 3) (v) an insight into the phylogenetics in sesamum species by taking into account wild relatives (Chapter 4).

Literature Review

Introduction

Lamiales is one of the largest and most diversified orders of angiosperm (flowering) plants and has more than 23,755 species belonging to 24~25 families in the asterids clade with representatives found all over the world (Stevens, 2012). Like most of the angiosperm members, Lamiales species play a crucial role for human being through providing oxygen, food and others resources. They are widely cultivated in tropical and sub-tropical areas of Asia owing to high demand for its essential oil, medicinal and woody species. Some well-known species in Lamiales include medicinally and economically important plants such as pot-herbs like mint, sage, oregano (basil), patchouli (*Lamiaceae*) (He et al., 2016a; Vining et al., 2017), olives, ash tree (*Oleaceae*) (Cruz et al., 2016; Sollars et al., 2017), sesame (*Pedaliaceae*) (Wang et al., 2014) and bladderwort (*Lentibulariaceae*) (Ibarra-Laclette et al., 2013). Lamiales presents variable genome size and specialized life strategies including carnivory, parasitism, epiphytism and desiccation tolerance (Schäferhoff et al., 2010).

Among lamiales clade, 62 plant genomes have been sequenced using Illumina, PacBio, Nanopore and Roche technologies. Precisely, eight species including medicinal, cave (adaptive), oil and woody plants have assembled at chromosomal level (Figure 1).

In recent years, there are huge research efforts focusing on genome assembly for Lamiales order. For example, the reference genomes of *Scutellaria baicalensis* and *Andrographis paniculata* have provided evolutionary insights of specific flavone biosynthetic pathways. The genomic data paved the way for enhancing bioactive productivity, antiviral properties, bioactive diterpenoid lactone for *S. baicalensis* and *A. paniculata*, respectively (Sun et al.,

2019; Zhao et al., 2019).

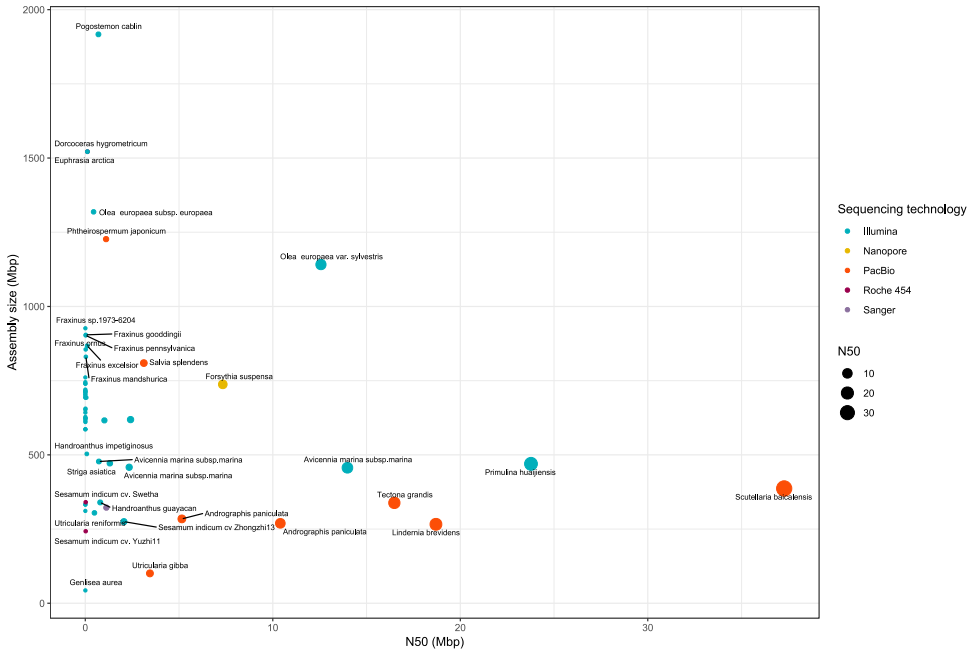

Figure 1. Overview of genome contiguity of Lamiales species assemblies following sequencing technologies. Data were collected from NCBI

Similarly, the reference genome of a cave plant (adaptive) *Primulina huaijiensis* is a good starting resource for the 180 members of this genus with limestone karst habitat. This genome explored gene family proliferation of WRKYs, bZIPs, adaptation to the high salinity and drought stress and gene retention following whole genome duplications (WGDs) events (Li et al., 2020b). The olive plant (*Olea europaea var. sylvestris*) genome investigation revealed that genes went through expansion and neo-functionalization mechanisms. It also provides a relevant overview of gene families involved in oil biosynthesis. For *Sesamum indicum* genome, researchers took advantage of a high density genetic map for unraveling seat coat color, growth habit, plant height and other valuable

candidate genes involved in biotic and abiotic stresses.

The most widely distributed gray mangrove species *Avivennia marina* genome has provided functional candidate genes underlying phenotypic and environmental divergence among the mangrove species. Also, salinity tolerance and adaptive mechanism at the transcriptional and metabolomics levels were covered (Friis et al., 2020). Similarly, the genome of *Tectona grandis* (teak), known for its strong wood, enables identifying phenylpropanoid pathway genes that lead to lignin formation and main sources for natural durability. The genome availability will surely facilitate traits specific candidate genes, anti-insecticidal natural production for sustainable production of teak (Zhao et al., 2019).

The genome of the carnivorous plant *Utricularia gibba* served as the first plant model for genome miniaturization phenomenon and to understand how plants evolve to acquire an adaptive genome architecture (Lan et al., 2017).

This review mainly focuses on the tremendous achievements of genome sequencing projects in Lamiales, their contributions for crop improvement and the challenges that need to be resolved in complex genomes.

Overview of whole genome sequencing projects for Lamiales species

The sequencing platform matters

The last decade, high-throughput genome sequencing advent enabled the investigation of Lamiales representative genomes through both short and long-reads sequencing technologies. At early stage of Lamiales genomics, the Roche 454 sequencing and Sanger platforms (in 2014) were employed to perform a genome skimming of some species mainly represented by the Penstemon members (Blischak et al., 2014). Despite the low coverage of the genomic data (~0.005x), relevant markers were generated for accurate identification of Penstemon and relative's species (Blischak et al., 2014). Most of the conducted sequencing projects in Lamiales order rely on Illumina short-reads sequencers including Illumina HiSeq and Illumina MiSeq (Figure 1). Meanwhile, third generation long-reads technologies is getting more interest from the genome project managers. The usage of long-reads alone or coupled with recent scaffolding methods resulted in more contiguous genome for some Lamiales species including *S. baicalensis*, *Lindernia brevidens*, *T. grandis*, *Forsythia suspensa*, and *A. paniculata* (Figure 1). Interestingly, scientific community seems to be more aware about Lamiales member values since the sequencing data undoubtedly increased during the last decade with a peak 73 Gbp of raw data produced in 2020 (Figure 2).

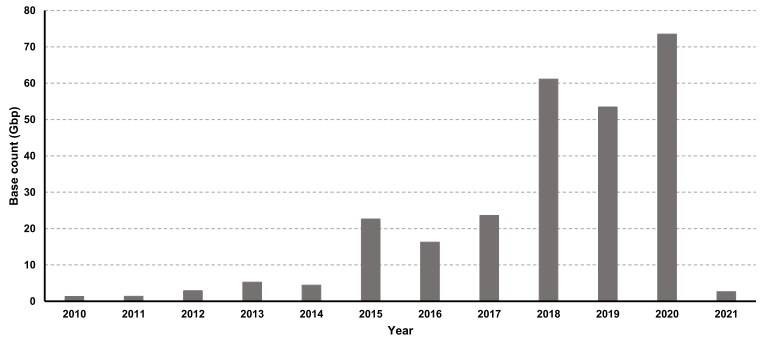

Figure 2. Sequence Read Archive (SRA) base counts of Lamiales species. Data were collected from NCBI.

Why Lamiales representatives are getting more attention?

Whole genome projects for Lamiales order included up to now, a total of 25 genus covering 11 families (Figure 3). The most representative families were *Oleaceae* and *Lamiaceae*, and *Acanthaceae* accounting for 52%, 11% and 10% respectively. Interestingly, the members of these families possess some valuable characteristics such as oil production and pharmacological properties. For example, Fraxinus species (Ash tree) exhibited antioxidant, antimicrobial, anti-inflammatory properties (Sarfraz et al., 2017; Attanzio et al., 2019; Bouguellid et al., 2020) where Olive tree is very famous for its highly valuable essential oil use in cosmetics industries. The olive also demonstrated antioxidant, cytoprotective, and cholinesterase's inhibition capabilities (Bahadori et al., 2018). Besides health benefits, Lamiales also provide energy fossil for human as well as raw materials for construction. Timber plant such as *T. grandis* and Ash tree provide wood for diverse applications in the wood industries. This highlighted the economic values of Lamiales members for human activities.

Genome characteristics complexity in Lamiales

Among the existing genomes assemblies, only 8% were assembled at chromosomal scale level for 14% at scaffold and 40% at contigs level. Reach a chromosomal level for some plants species is sometimes challenging due to multiple factors including the genome size, ploidy, repeat content and heterozygosity (Claros et al., 2012). Among the sequenced Lamiales members, the ploidy level varied from 2n = 2x = 16 to 2n = 4x = 128. Our prospective studies revealed the lack of information on ploidy's level regarding polyploid candidates.

Therefore, karyotyping appears as a compulsory step to figure out chromosome arrangement in those species. Besides, most of polyploid candidates were not yet assembled at chromosomal scale, confirming the assembly challenge especially for polyploidy organisms. For example, the chromosomal scale of the octaploid *Pogostemon cablin* is not yet available. However, its transcriptome landscape was provided by He et al. (2018). A key point among Lamiales members is the significant variability of the genome size. In fact, size fluctuation is responsible of genetic diversity and adaptation patterns (Kraaijeveld, 2010). Therefore, some plant models were discovered in Lamiales order *i.e.*, the plant miniature *Genlisea aurea* and the biggest Lamiales genome sequenced until this review *P. cablin*. Interestingly, some adaptation plant model relative to ecological environment, living style was also highlighted. For example, *U. gibba* and *U. reniformis*, two carnivorous plant genomes were scrutinized. Similarly, two major plants (*L. brevidens* and *Dorcoceras hygrometricum*) that exhibit desiccation tolerance were also sequenced and constitute a valuable resource for drought tolerance mitigation.

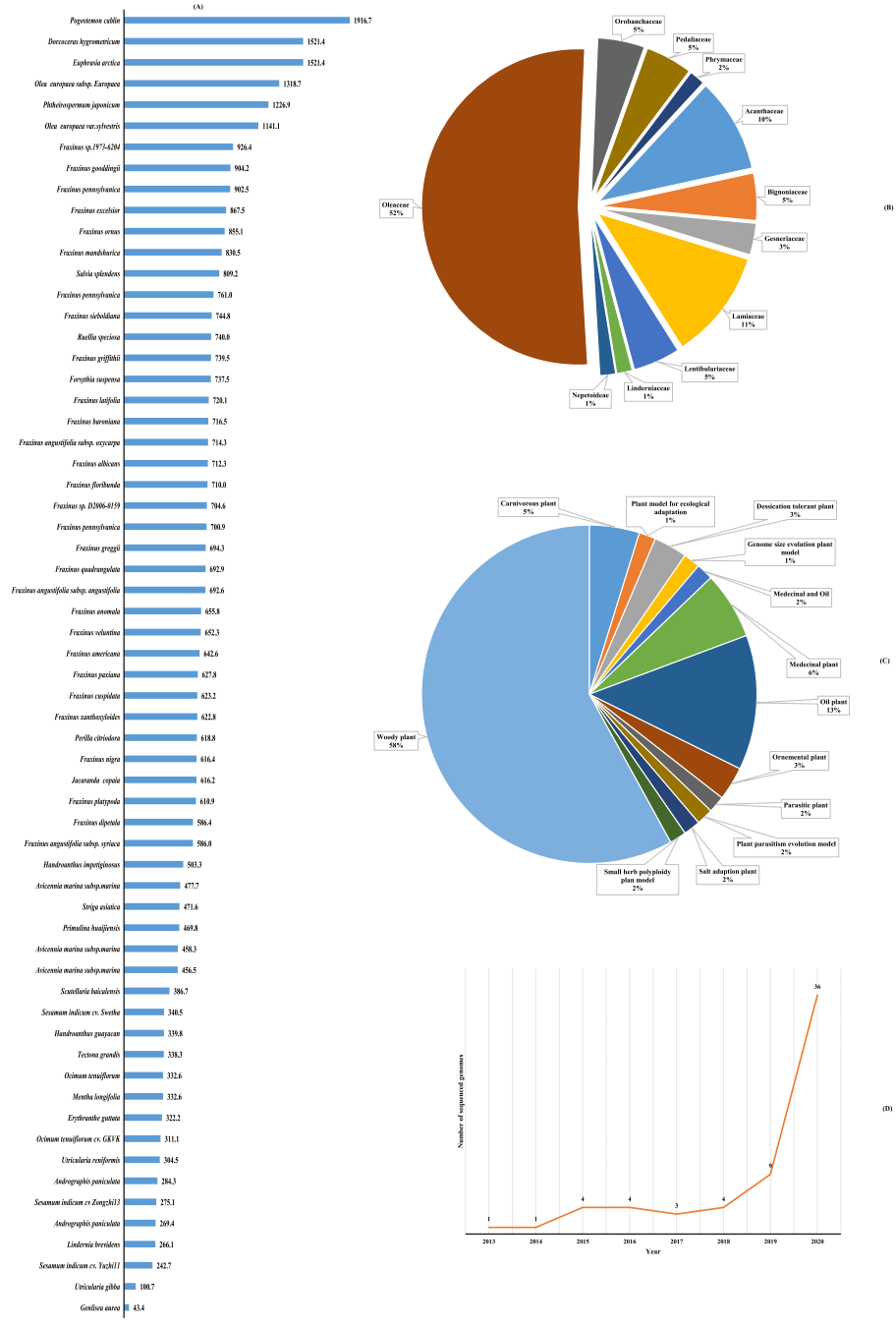

Figure 3. Overview of Lamiales species whole genome sequence. A) Size distribution of Lamiales species assemblies, B) Proportion of sequenced plants families in Lamiales order, C) Representatives

of plant type used in Lamiales genome sequence projects, D) Evolution of whole genome sequence projects from 2013 to 2020.

Contributions at fundamental research and plant breeding levels

The advent of genome sequencing technologies provides considerable advances in Lamiales. The contribution of genome sequencing in Lamiales (Table 1) order is classified in two groups: fundamental research and crop improvement aspects.

Contribution to fundamental research

Mechanism of genome size variation in plants

Genome sequencing of the carnivorous plants *G. aurea* and *U. gibba* help to understand the underlying mechanism of genome size variation in plants. Despite the intricacy of genome contraction phenomenon, the reduction of genome size derived from gene loss, introns and intergenic regions reduction (Ibarra-Laclette et al., 2013; Leushkin et al., 2013). It was proposed that genome size reduction occurred after multiple WGDs events marked by a purging mechanism of the excess DNA through genome fractionalization (Michael, 2014). In contrast, the bloating mechanism is proposed for big genome species (Baidouri and Panaud, 2013). The bloating mechanism involved the proliferation of long terminal retrotransposons (LTRs) with no purging event over time (Tenaillon et al., 2010). Interestingly, no whole genome duplication occurred for one of the biggest plants (~19Gbp), the Norway spruce (Nystedt et al., 2013). This suggests that the extension of the epigenome with no purging mechanism might be responsible of the genome size increase.

Table 1. An overview of the contribution of plant genomics in Lamiales order

Scientific Name	Contribution to the empiric knowledge	Major achievements	Reference
Genlisea aurea	- Contribution to understanding of genome contraction phenomenon during evolutionary process of genome size - Contribution to phylogenetic dating of whole genome duplication events understanding	Release of the first genome sequence for fundamental research	
Utricularia gibba	- Contribution to understanding of genome contraction phenomenon during evolutionary process of genome size - Contribution to phylogenetic dating of whole genome duplication events understanding	Release of the first genome sequence for fundamental research	
Mimulus guttatus	- Comprehension of basic properties subtended the meiotic recombination rate patterns in the genome at nucleotide-level resolution	Release of the first genome sequence for fundamental research	

Table 1. *Continued*

	Contribution to crop improvement	Major achievements
Fraxinus excelsior	- Discovery of genes responsible of resistance to the epidemic ash dieback disease through associative transcriptomics - Discovery of two groups of putative orthologues of 5 GEMs loci involve in flower development and stress response, bud set, growth cessation and dormancy	- Twenty gene expression markers (GEMs) associated the ash dieback disease composed of eight MADS-box proteins and two cinnamoyl-CoA reductase 2 (CCR2) genes involved in the hypersensitive response of ash tree - SVP/StMADS11 group of type II MADS-box genes
Sesamum indicum	- Deep understanding of oil biosynthesis pathway - Identification of QTL involved in desirable traits (color, height, yield related traits) - Contribution to genetic association studies for interested traits in sesame (growth, drought and waterlogging tolerance)	- Unraveling molecular foundation of high oil content in sesame through discovery of lipid genes families especially the type 1 lipid transport genes involved in high oil content - Discovery of gene involve in sesamin biosynthesis (SIN_1015471) and root biomass encoding gene called BIG ROOT BIOMASS gene

Table 1. *Continued*

Boea hygrometrica	- Deep understanding of drought tolerance patterns in desiccation tolerant plant model, *B. hygrometrica* in particular the protection of photosystem II during desiccation tolerance mechanism	- 9888 differentially expressed genes (DEGs) genes involved in responses of the plant to desiccation at the transcript level - Eight positive DEGs encode enzymes directly involved in ABA bio-synthesis and catabolism, indicating tight control of ABA levels during dehydration - Presence of PLD-1α, PLD-γs, PLD-β, and PLD-P1/Z1 (single phospholipase D genes), involved in the control of dehydration response - Two proteins derived from LEA1s (late embryogenesis abundant proteins), Bhs4_093 and Bhs4_094, play crucial role in the stabilization of the photosynthetic proteins during dehydration and rehydration process

Table 1. *Continued*

Mentha longifolia	- Contribution for *M. longifolia* crop improvement through molecular approach - Discovery of disease resistance gene	- Two genes involved in biosynthesis pathway were discovered: (4S)-limonene synthase and (-) – limonene 3-hydroxylase - Development of more than 2 million SSR markers for molecular breeding, metabolic engineering for essential oil production - Discovery of a glandular trichome-specific promoter involved in oil biosynthesis which can enable to focus on accumulation of monoterpenoids during oil biosynthesis - 17 putative disease resistance gene homologs were found
Ocimum tenuiflorum	- Contribution to the deep comprehension of pathway of major medicinal metabolites production and genes involved in their biosynthesis - Contribution to the leaves coloration pigmentation in Ocimum species.	- Discovery of ursolic acid, a sesquiterpenes member known to have anti-inflammatory, anti-microbial, anti-tumour and anti-cancer properties - Elucidation of leave pigmentation due to Chalcone synthase (CHS) - Anthocianin related genes were also found

Table 1. *Continued*

Pogostemon cablin	- Provide insight of molecular mechanism of the sesquiterpenoid biosynthesis pathway - Provide molecular tools for selection of *P. cablin* with improved medicinal and pharmaceutical traits	- Gamma-cadeninene synthase gene responsible of aroma was found - A total of 458 genes involved in the synthesis of metabolites were found - Discovery of 43 genes involve involved in sesquiterpenoid biosynthesis
Salvia miltiorrhiza	- Identification of genes involved in medicinal metabolite biosynthesis	- Discovery of 29 genes associated to the biosynthesis of transhinone and phenolic acid - Discovery of 82 terpene synthase genes involved in production of hemi-, mono- sesqui- or di-terpenes - Discovery of 5 genes involved in salvianolic acid biosynthesis and its expression in the root phloem and xylem tissues

Table 1. *Continued*

Tectona grandis	Contribution to the knowledge of genetic diversity and conservation of the woody tree *T. grandis*	- Generation of 182712 SSRs useful for population genetics studies - Generation of basic genomic data for comprehension of genetic basis of wood properties, pest tolerance, adaptive traits, germplasm movement and genetic resource conservation
Olea europaea	- Deep understanding high level accumulation of acid oleic in olive	- Increased expression of SACPD (stearoyl-ACP Desaturase) genes and decreased expression of FAD2 (fatty acid desaturase) explains the high accumulation of oleic acid in olive
	- Deep clarification of oil biosynthesis pathway in olive	- Discovery of 94 unique genes specific to oleaster involved in oil biosynthesis metabolic pathway
Handroanthus impetiginosus	- Generation of preliminary genomic resources for breeding, photochemistry, molecular studies and conservation of the neglected timber forest tree	- Discovery of 7 enzymes involved in lapachol synthesis (a bioactive component engaged in resistance of the tree against rotting fungi and insects)

Table 1. *Continued*

Ruellia speciosa	- Generation of SSR markers for population genetics studies	- Two putative copies of the enzyme F3H, eight of F3”H, one of F35H and three DFR were found to be involved in flower coloration
	- Deep comprehension of genes involved in anthocyanin biosynthesis pathway	- Identification of 96 homologues MYB transcription factors involved in flavonoid biosynthesis, a large part of anthocyanin biosynthesis pathways

In Lamiales, the genetic background of the biggest assembled genome *P. cablin* (He et al., 2016b, 2018) is not yet elucidated. As one the key factors making plant genome more complex, the assessment of the epigenome dynamics might definitively path a new avenue for unravelling genome variation mechanism.

Genetic background of desiccation tolerance in plants

The Lamiales member *Boea hygrometrica* genome revealed two WGD events and the presence of more than 9,000 unique genes which confer the capacity to the *B. hygrometrica* to dry without dying (Xiao et al., 2015). Early light-inducible proteins (ELIP) and 5S ribosomal RNA have been found to be involved in the protection of photosynthetic apparatus and the rapid activation of the protein synthesis. This attribute gives to the *B. hygrometrica* its resurrection function. The comparative genomics among a sensitive and desiccation tolerant Lindernia showed evidence of WGD as well as network rewiring (Van Buren et al., 2018). Besides, the comparative transcriptome of both Lindernia species exhibited the abundance of the Late Embryogenesis-Abundant (LEA) Protein Family genes. The LEA genes are known to play an important role for the mitigation of drought tolerance in some crops such as potato (Chen et al., 2019), wheat (Liu et al., 2019) and pepper (Lim et al., 2018).

Genetic background of carnivorous plants adaptation

Among Lamiales species, three carnivorous plants were sequenced: the miniature plant *G. aurea* (Leushkin et al., 2013), *U. gibba* (Ibarra-Laclette et al., 2013; Lan et al., 2017) and *U. reniformis* (Silva et al., 2020). Long-reads sequencing of the humped bladderwort enables

the detection of genome duplation marked coupled with the existence of trapping and digestive genes involved in peptide transporter, ATPase, hydrolase and chitinase activities. Tandem duplication events were therefore suggested as main drivers of the carnivorous speciation (Lan et al., 2017).

Contribution for crop improvement

The contribution of the genomics for plant breeding with a focus on four plants (*Sesamum indicum*, *Olea europaea*, *Mentha longiflora, and Andrographis paniculata*) representing the top four most sequenced family's members (Figure 3-B) are discussed in this section.

Genomics contribution to the molecular breeding in S. indicum

Sesame seeds contained high amount of oil, natural antioxidant including sesamol, sesamin and sesamolin (Bedigian et al., 1985a). Sesame is traditionally used in Africa and Asia as food and for medicinal purposes (Bedigian, 2018). *S. indicum* genome project has been initiated by the Chinese scientists from Henan Sesame Research Center (HSRC) of the Henan Academy of Agricultural Sciences and Oil Crops Research Institute (OCRI) of Chinese Academic of Agricultural Sciences. The two research teams generated draft genome of two Chinese cultivars Yuzhi 11 (Zhang et al., 2013b) for Henan Sesame Research Center, and Zhongzhi13 (Wang et al., 2014) for Oil Crops Research Institute. Later on, the OCRI team released an updated version with 13 pseudomolecules based on linkage mapping strategy (Wang et al., 2016). Besides Yuzhi11 and Zhongzhi13, two Chinese landraces Baizhima and Mishuozhima and one elite Indian modern cultivar Swetha were re-assembled based on the Zhongzhi13 reference genome. The Table 1 shows the main characteristics of the five sesame genome assemblies available up to date. A pan-genome was then constructed

(Yu et al., 2019a) leading to the detection of 22343, 22146, 27557, 20876 and 23372 core genes for Baizhima, Mishuozhima, Swetha, Yuzhi11 and Zhongzhi13 respectively (Table 1). Seed yield and quality traits were more dominant in the modern varieties whereas environmental adaptation related genes were abundant in landraces. The availability of the sesame genome laid the foundation for genomic-assisted sesame breeding.

The table 1 presented the main achievements of sesame genome sequencing project in advanced sesame breeding. At early stage, an important attention was paid to the germplasm genetic diversity studies. For plant breeding, knowing the genetic diversity of the plant material is a pre-requisite step. Therefore, several types of markers were designed and applied.

Random amplified polymorphic DNA (RAPD) (Abdellatef et al., 2008), amplified fragment length polymorphism (AFLP) (Laurentin and Karlovsky, 2007) were proposed and tested. Genome simple sequence repeats (g-SSR) (Dixit et al., 2005; Wei et al., 2014; Dossa et al., 2017c; Yu et al., 2017; Teklu et al., 2021), transcriptome-derived simple sequence repeats (Wei et al., 2014), expressed sequence tags simple sequence repeats (EST-SRR) (Wei et al., 2011; Badri et al., 2014; Sehr et al., 2016), cDNA simple sequence repeats (cDNA-SRR) (Spandana et al., 2012; Wang et al., 2012a), and inter-simple sequence repeats (ISSR) (Kim et al., 2002a) were also tested for genetic diversity studies as well as agronomic-oriented markers development. Besides, high-throughput markers using genotyping by sequencing (Uncu et al., 2016), specific length amplified fragment sequencing (SLAF) (Zhang et al., 2013e; Cui et al., 2017; Mei et al., 2017), restriction site-associated DNA sequencing (RAD) (Wu et al., 2014; Wang et al., 2016) and whole genome sequencing (Wang et al., 2015a; Wei et al., 2015; Zhang et al., 2016) approaches enabled genetic mapping of wide range of traits such as: oil content, sesamin production, capsule length, growth habit, number of capsules,

seed color, plant height, root biomass, semi-dwarf and yield related traits (Table 1).

The sesame genome sequencing project has the merit to provide useful databases that can be used for sesame breeding. One of the latest important achievement in sesame breeding was the discovery of an orphan gene 'Big Root Biomass' that governs root biomass in sesame (Dossa et al., 2020). Knowing that root plays a crucial function for the management of plant stress factors (drought stress, salinity), this recent discovery path is the way for sesame breeding for abiotic stress tolerance. In the same vein, a total of 47 core genes regulating waterlogging tolerance in sesame were recently identified. Taking advantage of a time course dense transcriptome data, WRKYs and ERF transcription factors were found to be the key master players in sesame response to waterlogging stress (Wang et al., 2021).

Meanwhile, combining QTL and GWAS approaches, Sovetgul *et al*. (2020) were able to highlight the gene SIN_1019016 as a potential candidate gene involved in *Phytophtora* blight disease resistance in sesame.

Genetic improvement for the olive tree O. europae

The olive tree is one of the oldest cultivated species mainly for its oil. Before the olive whole genome project initiative, multiple types molecular marker types were developed for the assessment of olive tree diversity studies and genetic mapping including RAPD(Fabbri et al., 1995; Mekuria et al., 1999; Wu et al., 2004) AFLP (Angiolillo et al., 1999), RFLP (Besnard and Bervillé, 2002; Bronzini de Caraffa et al., 2002), SSR (Rallo et al., 2000; Sefc et al., 2000; Cipriani et al., 2002), and sequence characterized amplified regions (SCAR) (Hernández et al., 2001a, 2001b; Mekuria et al., 2002; Wu et al., 2004). The whole genome sequence of the cultivated olive tree (*O. europae subsp. europaea*) was released in 2016, spanning 1.31 Gb with an N50 of 443 Kb (Cruz et al., 2016). Both fosmid library and

Illumina paired-end sequencing were employed. It is valuable to mention that with the third generation sequencing technologies; a better assembly quality can be achieved. Interestingly, wild olive genome was also generated (Unver et al., 2017); providing a useful resource to infer the evolutionary background of oil biosynthesis. With more than 50,000 genes, the wild olive genome investigation resulted in the divergence of some oil biosynthesis genes (fatty-acid desaturase (FAD2), stearoyl-ACP desaturase (SACPD), enoyl-ACP reductase (EAR), and ACP-hydrolase/thioesterase (ACPTE)) via duplication event. The accumulation of SACPD gene was suggested to be responsible of high content in oleic acid. The availability of both genomes enabled the identification of 14 leaf and fruit-related candidate genes (S1_1842014, S1_7858740, S1_403246, S1_12085523, S1_13002224, S1_7336035, S1_10019163, S1_11074838, S1_13767032, S1_984251, S1_13164923, S1_6292562, S1_138350, S1_904125) based on genotyping by sequencing approach (Kaya et al., 2019). Taking advantage of high-throughput sequencing of diverse panels, the domestication process of the cultivated olive tree (Julca et al., 2020), the genetic diversity (Taranto et al., 2018) as well as the functional variation polymorphism (Cultrera et al., 2019; Mariotti et al., 2020) were evaluated at SNP level in order to speed up olive tree breeding for high quality oil content.

Genomics contribution to M. longifolia breeding

The mint genome (*M. longifolia* CMEN 585) was released in 2017 (Vining et al., 2017) based on a combination of Illumina PacBio technologies. The contigs were anchored into 12 pseudo-chromosomes through linkage mapping approach. The assembly genome reached 353 Mb representing 92% based on the flow cytometry size method estimation (Bennett and Leitch, 2005). Among the 35,597 predicted genes, eight candidate monoterpene

biosynthesis-related genes were identified. Besides, 292 resistance genes were inferred, providing putative genes for mint Verticillum disease resistance. The release of the wild mint *M. longiflora*, one of the ancestral species of the cultivated mint might help to understand the rise of the hexaploid hybrid *M.* × *piperita*. Besides, the genomic data is a valuable resource to enhance mint breeding for oil and disease resistance.

Genomics contribution to A. paniculata breeding

A. paniculata is medicinal plant widely distributed and used in tropical and subtropical regions of China, India and Southeast Asian countries. The plant has highly acclaimed crucial therapeutic and pharmacological properties due to a high content in andrographolide and neoandrographolide.

At early stage of the genome project, molecular markers were employed for diversity studies including RAPD (Sharma et al., 2009), ISSR, SCoT, CBDP, SSR (Tiwari et al., 2016), and AFLP (Wijarat P, 2012). Later on, transcriptome profiling of diterpenes genes in leaf and root was reported (Garg et al., 2015). A total of 48 diterpenes related genes were highly expressed, providing a key resource for the characterization of diterpenes biosynthesis. However, the biosynthesis pathway was still unclear. The Chloroplast genome was released a year after with 114 unique genes (Ding et al., 2016). The Whole genome sequencing project of this medicinal plant was conducted by two researcher's groups in China. The first released genome was generated with Illumina, PacBio and Hi-C technologies. It results in a 269 Mb genome with an N50 of 388 Kb. The anchoring produced 23 pseudo-chromosomes and 25,428 genes were predicted (Sun et al., 2019). The second released genome (Liang et al., 2020) exhibited more contiguity with a N50 size of 5.14 Mb spanning 284 Mb. Both genomic data shed light on the terpenoid biosynthetic pathway in *A. paniculata*. As a result,

UDP-dependent glycosyltransferases cytochrome P450 monooxygenases, and 2-oxoglutarate-dependent dioxygenases were predicted to play an important role in diterpenoid lactone synthesis.

Besides, using contrasting materials with regards to andrographolide content, Patel *et al*. (2020) pointed out the Isopentenyl-diphosphate delta-isomerase, 13-hydroxy-3-methylglutaryl-coenzyme reductase (HMGR), geranyl diphosphate synthase small subunit, genes as responsible of the andrographolide accumulation of in leaf.

Conclusion and future perspectives

In the present review, we reported an overview of the completed genome projects conducted in Lamiales order, by highlighting their contribution for fundamental and applied crop sciences. Generate more accurate genome at chromosomal scale is pre-requisite for plant breeding or detection of valuable biological metabolism pathway. Therefore, a particular accent should be put on the usage of third generation technologies for precise assembly. Recent long-reads technologies such as Pac-Bio Hifi, Nanopore ProMethion, and Hi-C scaffolding technologies have been used (from Dovetails and Phase Genomics companies) (Hon et al., 2020; Kadota et al., 2020; Lang et al., 2020). Bionano technology constitutes a good alternative to cope with polyploidy and genome size issues as well as structural variant detection while 10x Genomics is preferred for linking reads from short-reads sequencing platform in order to improve the contiguity (Kyriakidou et al., 2018). Among the available assembly, numerous gaps were present. The usage of long-reads sequencing technologies might help to provide near complete gap-filled assembly. In subsequent, more valuable genes can be recovered for application in plant breeding, human health or commercial

purpose. As polyploidy, heterogeneous or big plant genome is challenging to assemble, a systematic evaluation of different assembly algorithms should be performed. Hence, assembly benchmarking datasets represent a valuable resource to deal with further assembly curation.

Alongside the Vertebrate Genomes Project established by the Genome 10K consortium, a typical workflow to reach a golden standard diploid genome assembly was successfully tested (Rhie et al., 2021). The workflow starts by the generation of haplotype-resolved contigs assembly using PacBio Continuous Long-Reads (PacBio CLR) followed by three scaffolding strategies including 10× Genomics linked reads, Bionano optical maps and Phase Genomics Hi-C consecutively. Hence, gap filling, polishing and manual refinement closed this approach.

In addition, full-length transcripts are now possible to be sequenced in one shot with Isoseq technology. This will be beneficial for precise detection of molecular markers of candidate gene for genomic-assisted breeding.

Altogether, improving genome of Lamiales members will dramatically help for a better comprehension of metabolites synthesis, functional attributes of important agronomic traits and enable an efficient genomic-assisted plant breeding.

CHAPTER 1: Agronomic Traits Diversity Assessment from a Worldwide Sesame Accessions and Extraction of a Core Collection

Summary

Accurate knowledge of phenotypic diversity in a germplasm is a paramount for effective genomic-assisted breeding. A worldwide sesame panel was screened for both agronomic and nutritional characters as part of the South Korean sesame genomic-assisted program. The aim is to identify valuable materials to boot sesame breeding program. Taking advantage of a 506 worldwide accessions, a phenotyping was carried out at Jeonju and Miryang research stations during the cropping season 2018. A significant natural variation of agronomic traits was highlighted for some major traits including plant height, productive axis, length, seed color, number of locules per capsule, and dried seed weight. Following a multivariate analysis, a set of 32 accessions were identified as candidates for yield improvement. Moreover, a core collection composed of 102 accessions was developed, offering a starting point material for genome-wide association studies. From the core collection, protein, oil, sesamolin, sesamin, and fatty acid content were determined for 72 accessions due to the seeds quantity limitation. The accessions T218, T077, T419, T148, TN03, and T414 were pinpointed for lignans, oil content, and fatty acid-oriented sesame breeding. Interestingly, highly leafy-type accessions were also identified, representing a valuable resource for nutraceutical values of sesame leaf investigation.

Introduction

Over the last few years, global hunger has begun a rising challenge in the world due to the increasing population growth. By 2050, the global population is expected to exceed 9 billion, increasing the food demand about 70% (Tripathi et al., 2018). When coupled with the current adverse effects of climate change, the Zero Hunger objective of the Food and Agricultural Organization (FAO) is seriously jeopardized with the addition of more than 80 million of undernourished people in the recent COVID-19 pandemic context (FAO, IFAD, UNICEF, WFP, 2020). The projections of the climate change variability on agricultural sector threaten the attainment of the food security and poverty reduction ambitions in developing countries (Nhemachena et al., 2020; Molotoks et al., 2021). Therefore, there is an urgent need to develop and deploy crops that combine high yield, nutritional values and strong ability to grow in harsh environments. Besides, crops that can improve the human health through the mitigation of diseases are getting more importance (Yu and Li, 2021). The example of the barley cultivar BARLEY^{TM} revealed the presence of resistant starch metabolite that has positive effects on the alleviation of the type-2 diabetes and coronary heart diseases (Morell et al., 2003).

Considered as an orphan crop (Dossa et al., 2017a), sesame (*Sesamum indicum*) is a nutritional food (Bedigian, 2018) and an excellent source of lignans (Moazzami et al., 2007) that showed a wide range of benefits for human health including lowering blood cholesterol and glucose, cardiovascular disease prevention (Huang et al., 2021), tumor growth suppression (Harikumar et al., 2010), anti-carcinogenic properties (Yokota et al., 2007) and metabolic syndrome alleviation (Farajbakhsh et al., 2019). Lignans become marketable compounds with a high economical value estimated to $351.6 million dollars in 2019. The

lignans market might approximate more than $500 million dollars[1] by 2027. This interest for lignans is noticeable with a growing number of patents in healthy food additives as well as skin care sectors (Forse and Chavali, 2001b; You et al., 2011a; Kojima et al., 2020; Yamada et al., 2020).

Genebank is a reservoir of genetic diversity allowing the identification of promising sources for crop improvement (Smale and Jamora, 2020). South Korea has the second largest sesame genebank in the world with 7,853 accessions[2]. At the early stage of the Korean genomic-assisted sesame breeding program, the selection of valuable high-quality nutritional and health-beneficial materials is a crucial prerequisite.

From the Korean genbank, sesame core collection development using 2,751 worldwide accessions, was previously performed by including only one African country (Park et al., 2015). As broad geographic coverage is, genetic resources origin can provide a heuristic view for detection of novel desirable traits. Therefore, we enlarge geographic basis of genotypes covering 35 countries throughout the world by reaching 22 African countries. Despite the characterization of some African accessions from the Chinese genebank (Dossa et al., 2018) for lipid content, the two major lignans *viz* sesamin and sesamolin assessment have been neglected.

The present study aimed to i) assess the variability of agronomic traits in a worldwide sesame panel, ii) infer a core collection, iii) screen the core collection for sesamolin, sesamin, oil, and protein contents. Ultimately, this research global goal is to provide relevant resources that can serve as starting materials for the Korean genomic-assisted selection

[1] https://www.grandviewresearch.com/industry-analysis/lignans-market

[2] http://genebank.rda.go.kr/plantMain.do

initiative for boosting sesame yield, nutritional quality and health benefits.

Materials and Methods

Plant material and field experiment

A total number of 506 accessions (Table 2) provided by the Korean genebank were tested during summer season 2018 (May-September). The experiment was laid out following Federer's augmented design (Federer and Raghavarao, 1975) with 8 blocks at Jeonju (35° 49' 50.37"N latitude, 127° 3' 52.79"E longitude) and Miryang (35° 29' 29.70"N latitude, 128° 44' 31.98"E longitude). The inter-row and inter-plant distances were 0.7m and 0.2m respectively.

Table 2. List of sesame accessions used in the present study

Field Code	IT Code/Name	Biological status	Country of origin
T001	IT 165009	Wild	Ghana
T002	IT 165010	Wild	Ghana
T003	IT 165633	Unknown	Egypt
T004	IT 165634	Unknown	Egypt
T005	IT 165635	Unknown	Egypt
T006	IT 165636	Unknown	Egypt
T007	IT 165793	Unknown	Egypt
T008	IT 167146	Unknown	Kenya
T009	IT 167147	Unknown	Kenya
T010	IT 167148	Wild	Kenya
T011	IT 169148	Unknown	Egypt
T012	IT 169149	Unknown	Egypt
T013	IT 169150	Unknown	Egypt
T014	IT 169151	Unknown	Egypt
T015	IT 169152	Unknown	Egypt
T016	IT 169153	Unknown	Egypt
T017	IT 169154	Unknown	Egypt
T018	IT 169155	Unknown	Egypt

Table 2. *Continued*

T019	IT 169156	Unknown	Egypt
T020	IT 169157	Unknown	Egypt
T021	IT 169158	Unknown	Egypt
T022	IT 169159	Unknown	Egypt
T023	IT 169160	Unknown	Egypt
T024	IT 169161	Unknown	Egypt
T025	IT 169162	Unknown	Egypt
T026	IT 169163	Unknown	Egypt
T027	IT 169164	Unknown	Egypt
T028	IT 169165	Unknown	Egypt
T029	IT 169166	Unknown	Egypt
T030	IT 169167	Unknown	Egypt
T031	IT 169168	Unknown	Egypt
T032	IT 169169	Unknown	Egypt
T033	IT 169170	Unknown	Egypt
T034	IT 169171	Unknown	Egypt
T035	IT 169172	Unknown	Egypt
T036	IT 169173	Unknown	Egypt
T037	IT 169174	Unknown	Egypt
T038	IT 169175	Unknown	Egypt
T039	IT 169176	Unknown	Egypt
T040	IT 169177	Unknown	Egypt
T041	IT 169178	Unknown	Egypt
T042	IT 169179	Unknown	Egypt
T043	IT 169180	Unknown	Egypt
T044	IT 169211	Unknown	Egypt
T045	IT 169212	Unknown	Egypt
T046	IT 169213	Unknown	Egypt
T047	IT 169214	Unknown	Egypt
T048	IT 169215	Unknown	Egypt
T049	IT 169216	Unknown	Egypt

Table 2. *Continued*

T050	IT 169217	Unknown	Egypt
T051	IT 169218	Unknown	Egypt
T052	IT 169219	Unknown	Egypt
T053	IT 169220	Unknown	Egypt
T054	IT 169221	Unknown	Egypt
T055	IT 169222	Unknown	Egypt
T056	IT 169223	Unknown	Egypt
T057	IT 169224	Unknown	Egypt
T058	IT 169225	Unknown	Egypt
T059	IT 169226	Unknown	Egypt
T060	IT 169227	Unknown	Egypt
T061	IT 169228	Unknown	Egypt
T062	IT 169229	Unknown	Egypt
T063	IT 169230	Unknown	Egypt
T064	IT 169231	Unknown	Egypt
T065	IT 169232	Unknown	Egypt
T066	IT 169233	Unknown	Egypt
T067	IT 169532	Unknown	Nigeria
T068	IT 169533	Unknown	Nigeria
T069	IT 169534	Unknown	Nigeria
T070	IT 169536	Unknown	Cameroon
T071	IT 169537	Unknown	Cameroon
T072	IT 169538	Unknown	Zaire (Democratic Repuplic of the Congo)
T073	IT 169579	Unknown	Ethiopia
T074	IT 169609	Unknown	Lybia
T075	IT 169610	Unknown	Lybia
T076	IT 169611	Unknown	Lybia
T077	IT 169615	Unknown	Ethiopia
T078	IT 169622	Unknown	Mozambique
T079	IT 169623	Unknown	Mozambique
T080	IT 169624	Unknown	Mozambique

Table 2. *Continued*

T081	IT 169625	Unknown	Mozambique
T082	IT 169708	Unknown	Zaire (Democratic Repuplic of the Congo)
T083	IT 169725	Unknown	Mozambique
T084	IT 169726	Unknown	Mozambique
T085	IT 169727	Unknown	Mozambique
T086	IT 169731	Unknown	Egypt
T087	IT 169732	Unknown	Egypt
T088	IT 169733	Unknown	Egypt
T089	IT 169734	Unknown	Egypt
T090	IT 169735	Unknown	Egypt
T091	IT 169736	Unknown	Egypt
T092	IT 169740	Unknown	Egypt
T093	IT 169741	Unknown	Egypt
T094	IT 169742	Unknown	Egypt
T095	IT 169743	Unknown	Egypt
T096	IT 169744	Unknown	Egypt
T097	IT 169840	Unknown	Ethiopia
T098	IT 169995	Unknown	Nigeria
T099	IT 169999	Unknown	Nigeria
T100	IT 170000	Unknown	Nigeria
T101	IT 170001	Unknown	Nigeria
T102	IT 170002	Unknown	Nigeria
T103	IT 170003	Unknown	Nigeria
T104	IT 170006	Unknown	Nigeria
T105	IT 170008	Unknown	Nigeria
T106	IT 170009	Unknown	Nigeria
T107	IT 170010	Unknown	Nigeria
T108	IT 170011	Unknown	Nigeria
T109	IT 170012	Unknown	Nigeria
T110	IT 170013	Unknown	Nigeria
T111	IT 170014	Unknown	Nigeria

Table 2. *Continued*

T112	IT 170015	Unknown	Nigeria
T113	IT 170016	Unknown	Nigeria
T114	IT 170017	Unknown	Nigeria
T115	IT 170018	Unknown	Nigeria
T116	IT 170019	Unknown	Nigeria
T117	IT 170028	Unknown	Sudan
T118	IT 170029	Unknown	Sudan
T119	IT 170030	Unknown	Sudan
T120	IT 170088	Unknown	Sudan
T121	IT 170089	Unknown	Sudan
T122	IT 170091	Unknown	Sudan
T123	IT 170094	Unknown	Sudan
T124	IT 170095	Unknown	Sudan
T125	IT 170097	Unknown	Sudan
T126	IT 170099	Unknown	Sudan
T127	IT 170100	Unknown	Sudan
T128	IT 170101	Unknown	Sudan
T129	IT 170112	Unknown	Sudan
T130	IT 170113	Unknown	Sudan
T131	IT 170114	Unknown	Sudan
T132	IT 170115	Unknown	Sudan
T133	IT 170116	Unknown	Sudan
T134	IT 170117	Unknown	Sudan
T135	IT 170118	Unknown	Sudan
T136	IT 170119	Unknown	Sudan
T137	IT 170120	Unknown	Sudan
T138	IT 170121	Unknown	Sudan
T139	IT 170122	Unknown	Sudan
T140	IT 170123	Unknown	Sudan
T141	IT 170126	Unknown	Sudan
T142	IT 170128	Unknown	Sudan

Table 2. *Continued*

T143	IT 170129	Unknown	Sudan
T144	IT 170131	Unknown	Sudan
T145	IT 170132	Unknown	Sudan
T146	IT 170133	Unknown	Sudan
T147	IT 170134	Unknown	Sudan
T148	IT 170135	Unknown	Sudan
T149	IT 170136	Unknown	Sudan
T150	IT 170137	Unknown	Sudan
T151	IT 170138	Unknown	Sudan
T152	IT 170139	Unknown	Sudan
T153	IT 170140	Unknown	Sudan
T154	IT 170142	Unknown	Sudan
T155	IT 170143	Unknown	Sudan
T156	IT 170144	Unknown	Sudan
T157	IT 170145	Unknown	Sudan
T158	IT 170146	Unknown	Sudan
T159	IT 170148	Unknown	Sudan
T160	IT 170152	Unknown	Sudan
T161	IT 170153	Unknown	Sudan
T162	IT 170155	Unknown	Sudan
T163	IT 170156	Unknown	Sudan
T164	IT 170157	Unknown	Sudan
T165	IT 170158	Unknown	Sudan
T166	IT 170159	Unknown	Sudan
T167	IT 170161	Unknown	Angola
T168	IT 184270	Unknown	Egypt
T169	IT 184294	Unknown	Tanzania
T170	IT 184295	Unknown	Tanzania
T171	IT 184296	Unknown	Tanzania
T172	IT 184297	Unknown	Tanzania
T173	IT 184298	Unknown	Tanzania

Table 2. *Continued*

T174	IT 184299	Unknown	Tanzania
T175	IT 184300	Unknown	Tanzania
T176	IT 184318	Unknown	Egypt
T177	IT 184319	Unknown	Egypt
T178	IT 184320	Unknown	Egypt
T179	IT 184321	Unknown	Egypt
T180	IT 184322	Unknown	Egypt
T181	IT 184324	Unknown	Egypt
T182	IT 184325	Unknown	Egypt
T183	IT 184326	Unknown	Egypt
T184	IT 184328	Unknown	Somalia
T185	IT 184330	Unknown	Somalia
T186	IT 184332	Unknown	Somalia
T187	IT 184333	Unknown	Somalia
T188	IT 184334	Unknown	Somalia
T189	IT 184335	Unknown	Somalia
T190	IT 184336	Unknown	Somalia
T191	IT 184337	Unknown	Somalia
T192	IT 184339	Unknown	Ethiopia
T193	IT 184340	Unknown	Ethiopia
T194	IT 184341	Unknown	Ethiopia
T195	IT 184342	Unknown	Ethiopia
T196	IT 184343	Unknown	Ethiopia
T197	IT 184344	Unknown	Ethiopia
T198	IT 184345	Unknown	Ethiopia
T199	IT 184346	Unknown	Ethiopia
T200	IT 184379	Unknown	Zaire (Democratic Republic of the Congo)
T201	IT 184380	Unknown	Zaire (Democratic Republic of the Congo)
T202	IT 184381	Unknown	Zaire (Democratic Republic of the Congo)

Table 2. *Continued*

T203	IT 184382	Unknown	Zaire (Democratic Republic of the Congo)
T204	IT 184383	Unknown	Zaire (Democratic Republic of the Congo)
T205	IT 184437	Unknown	Somalia
T206	IT 184438	Unknown	Angola
T207	IT 184439	Unknown	Angola
T208	IT 184518	Unknown	Egypt
T209	IT 184519	Unknown	Egypt
T210	IT 184520	Unknown	Egypt
T211	IT 184553	Unknown	Mozambique
T212	IT 184554	Unknown	Mozambique
T213	IT 184556	Unknown	Mozambique
T214	IT 184557	Unknown	Nigeria
T215	IT 184589	Unknown	Tanzania
T216	IT 184649	Unknown	Egypt
T445	IT 184704	Unknown	Sudan
T217	IT 184733	Unknown	Liberia
T218	IT 184738	Unknown	Mozambique
T219	IT 184745	Unknown	Egypt
T220	IT 184764	Unknown	Sudan
T221	IT 193999	Unknown	Senegal
T222	IT 194000	Unknown	Senegal
T223	IT 194001	Unknown	Senegal
T224	IT 194002	Unknown	Senegal
T225	IT 194003	Unknown	Senegal
T226	IT 194005	Unknown	Senegal
T227	IT 194008	Unknown	Sudan
T228	IT 194010	Unknown	Sudan
T229	IT 194011	Unknown	Sudan
T230	IT 194012	Unknown	Sudan
T231	IT 194013	Unknown	Sudan

Table 2. *Continued*

T232	IT 194015	Unknown	Sudan
T233	IT 194016	Unknown	Sudan
T234	IT 194017	Unknown	Sudan
T235	IT 194356	Unknown	Ethiopia
T236	IT 194357	Unknown	Ethiopia
T237	IT 216913	Unknown	Zaire (Democratic Republic of the Congo)
T238	IT 218014	Unknown	Sudan
T239	IT 218263	Unknown	Kenya
T240	IT 218264	Unknown	Kenya
T241	IT 219069	Unknown	Kenya
T242	IT 219070	Unknown	Kenya
T243	IT 219071	Unknown	Kenya
T244	IT 238548	Landrace	Kenya
T245	IT 238549	Landrace	Kenya
T246	IT 238557	Landrace	Sudan
T247	IT 238558	Landrace	Sudan
T248	IT 242948	Landrace	Kenya
T249	IT 242949	Landrace	Kenya
T250	IT 242950	Landrace	Kenya
T251	IT 242951	Landrace	Kenya
T252	IT 271179	Unknown	Kenya
T253	IT 271180	Unknown	Kenya
T254	IT 271181	Unknown	Kenya
T255	IT 271182	Unknown	Kenya
T256	IT 271192	Cultivar	Sudan
T257	IT 271209	Landrace	Kenya
T258	IT 271210	Landrace	Kenya
T259	IT 271211	Landrace	Kenya
T260	IT 275753	Landrace	Tanzania
T261	IT 286849	Unknown	Sudan

Table 2. *Continued*

T262	IT 286850	Unknown	Sudan
T263	IT 286851	Unknown	Sudan
T264	IT 286852	Unknown	Sudan
T265	IT 286853	Unknown	Sudan
T266	IT 286854	Unknown	Sudan
T267	IT 286855	Unknown	Sudan
T268	IT 286856	Landrace	Eritrea
T269	IT 286876	Unknown	Morocco
T270	IT 286877	Unknown	Morocco
T271	IT 286878	Unknown	Morocco
T272	IT 286884	Unknown	Morocco
T273	IT 286885	Unknown	South Africa
T274	IT 286889	Unknown	Uganda
T275	IT 286890	Unknown	Sudan
T276	IT 286891	Unknown	Sudan
T277	IT 286892	Unknown	Sudan
T278	IT 286893	Unknown	Sudan
T279	IT 286894	Unknown	Sudan
T280	IT 286896	Unknown	Nigeria
T281	IT 286899	Unknown	Senegal
T282	IT 286906	Unknown	Libya
T283	IT 286909	Unknown	Egypt
T284	IT 290165	Unknown	Nigeria
T285	IT 299405	Unknown	Egypt
T286	IT 299433	Unknown	Egypt
T287	IT 299435	Unknown	Morocco
T288	IT 299436	Unknown	Morocco
T289	IT 299440	Unknown	Morocco
T290	IT 299444	Unknown	Uganda
T291	IT 299445	Unknown	Tunisia
T292	IT 311597	Unknown	Kenya

Table 2. *Continued*

T293	IT 311676	Landrace	Morocco
T294	IT 169244	Unknown	Turkey
T295	IT 169248	Unknown	Turkey
T296	IT 169397	Unknown	Turkey
T297	IT 169404	Unknown	Turkey
T298	IT 169409	Unknown	Turkey
T299	IT 184593	Unknown	Turkey
T300	IT 184602	Unknown	Turkey
T301	IT 184615	Unknown	Turkey
T302	IT 184623	Unknown	Turkey
T303	IT 184636	Unknown	Turkey
T304	IT 184519	Unknown	Egypt
T305	IT 184520	Unknown	Egypt
T306	IT 184292	Unknown	India
T307	IT 184425	Unknown	India
T308	IT 184466	Unknown	India
T309	IT 184472	Unknown	India
T310	IT 184473	Unknown	India
T311	IT 184475	Unknown	India
T312	IT 184476	Unknown	India
T313	IT 184492	Unknown	India
T314	IT 156334	Unknown	Korea
T315	IT 156362	Unknown	Korea
T316	IT 193010	Unknown	Korea
T317	IT 193692	Unknown	Korea
T318	IT 195373	Unknown	Korea
T319	IT 184354	Unknown	Mexico
T320	IT 184356	Unknown	Mexico
T321	IT 184366	Unknown	Mexico
T322	IT 184546	Unknown	Mexico
T323	IT 196091	Unknown	Mexico

Table 2. *Continued*

T324	IT 169338	Unknown	China
T325	IT 169359	Unknown	China
T326	IT 196047	Unknown	China
T327	IT 196049	Unknown	China
T328	IT 196068	Unknown	China
T329	IT 184318	Unknown	Egypt
T330	IT 184321	Unknown	Egypt
T331	IT 184326	Unknown	Egypt
T332	IT 184311	Unknown	Philippines
T333	IT 184313	Unknown	Philippines
T334	IT 184314	Unknown	Philippines
T335	IT 184317	Unknown	Philippines
T336	IT 169949	Unknown	Russia
T337	IT 184651	Unknown	Russia
T338	IT 184653	Unknown	Russia
T339	IT 184654	Unknown	Russia
T340	IT 184747	Unknown	Russia
T341	IT 184753	Unknown	Russia
T342	IT 102678	Unknown	Korea
T343	IT 102975	Unknown	Korea
T344	IT 103159	Unknown	Korea
T345	IT 113273	Unknown	Korea
T346	IT 113593	Unknown	Korea
T347	IT 184600	Unknown	India
T348	IT 184718	Unknown	India
T349	IT 184726	Unknown	India
T350	IT 196116	Unknown	India
T351	IT 184510	Unknown	Iran
T352	IT 184511	Unknown	Iran
T353	IT 184514	Unknown	Iran
T354	IT 184515	Unknown	Iran

Table 2. *Continued*

T355	IT 184516	Unknown	Iran
T356	IT 184517	Unknown	Iran
T357	IT 169955	Unknown	Afghanistan
T358	IT 184426	Unknown	Afghanistan
T359	IT 184427	Unknown	Afghanistan
T360	IT 184428	Unknown	Afghanistan
T361	IT 184429	Unknown	Afghanistan
T362	IT 184430	Unknown	Afghanistan
T363	IT 184431	Unknown	Afghanistan
T364	IT 184432	Unknown	Afghanistan
T365	IT 184434	Unknown	Afghanistan
T366	IT 184436	Unknown	Afghanistan
T367	IT 184736	Unknown	Iran
T368	IT 184527	Unknown	Japan
T369	IT 184529	Unknown	Japan
T370	IT 184532	Unknown	Japan
T371	IT 192443	Unknown	Japan
T372	IT 209652	Unknown	Japan
T373	IT 184757	Unknown	Nepal
T374	IT 200589	Unknown	Nepal
T375	IT 169941	Unknown	Pakistan
T376	IT 184569	Unknown	Pakistan
T377	IT 184570	Unknown	Pakistan
T378	IT 184579	Unknown	Pakistan
T379	IT 184586	Unknown	Pakistan
T380	IT 184305	Unknown	Philippines
T381	IT 184308	Unknown	Philippines
T382	IT 184642	Unknown	Turkey
T383	IT 184643	Unknown	Turkey
T384	IT 184644	Unknown	Turkey
T385	IT 184727	Unknown	Turkey

Table 2. *Continued*

T386	IT 184269	Unknown	USA
T387	IT 184272	Unknown	USA
T388	IT 184399	Unknown	USA
T389	IT 184686	Unknown	USA
T390	IT 192412	Unknown	USA
T391	IT 184409	Unknown	Venezuela
T392	IT 184715	Unknown	Venezuela
T393	IT 184734	Unknown	Venezuela
T394	Daheuk	Cultivar	Korea
T395	Yupung	Cultivar	Korea
T396	Aleum	Cultivar	Korea
T397	Suji	Cultivar	Korea
T398	Gangheug	Cultivar	Korea
T399	Geumok	Cultivar	Korea
T400	Gangan	Cultivar	Korea
T401	Goenbaek	Cultivar	Korea
T402	Sangbaek	Cultivar	Korea
T403	Ohsan	Cultivar	Korea
T404	Milseong	Cultivar	Korea
T405	CC229	Unknown	Korea
T406	CC218	Unknown	Korea
T407	Baekseol	Cultivar	Korea
T408	Miho	Cultivar	Korea
T409	Maluhime	Cultivar	Korea
T410	Jobaek	Cultivar	Korea
T411	Gomaoh	Cultivar	Korea
T412	Maluaeong	Cultivar	Korea
T413	Dodam	Cultivar	Korea
T414	Nuri	Cultivar	Korea
T415	Chamhwang	Cultivar	Korea
T416	Dt-sel	Mutant	Turkey

Table 2. *Continued*

T417	Dt-45	Mutant	Turkey
T418	CC215	Unknown	Korea
T419	655	Unknown	Korea
T420	227	Unknown	Korea
T421	UD335	Unknown	Korea
T422	UD347	Unknown	Korea
T423	Yonginlocal	Cultivar	Korea
T424	Suwon195	Cultivar	Korea
T425	E.R.	Unknown	Korea
T426	HS445-1-1-2-2-4	Unknown	Korea
T427	UD106	Unknown	Korea
T428	Gyeongbuk 22	Cultivar	Korea
T429	*Sesamum calysimum subsp. baumii*	Wild	Zimbabwe
T430	*Sesamum alatum*	Wild	Zimbabwe
T431	*Sesamum schinzianum*	Wild	Namibia
T432	*Sesamum radiatum*	Wild	Angola
T433(C1)	Goenbaek	Cultivar	Korea
T434(C2)	Kopoom	Cultivar	Korea
T435(C3)	Daheuk	Cultivar	Korea
T436(C4)	Pyungan	Cultivar	Korea
T437(C5)	Yoomi	Cultivar	Korea
T438	GBK040641	Unknown	Kenya
T439	GBK040736	Unknown	Kenya
T440	HB168	Mutant lines	Senegal
T441	LC162	Mutant lines	Senegal
T442	LC164	Mutant lines	Senegal
T443	EF147	Mutant lines	Senegal
T444	EF153	Mutant lines	Senegal
T445	IT 184704	Unknown	Sudan
TN01	HUNo_1991_2478	Landrace	Korea
TN02	Kochangsuchib	Landrace	Korea

Table 2. *Continued*

TN03	Kyonbuk33	Landrace	Korea
TN04	Hukchangkun	Landrace	Korea
TN05	Bekmi	Landrace	Korea
TN06	Pongnek	Landrace	Korea
TN07	Kuansan	Landrace	Korea
TN08	Danbek	Landrace	Korea
TN09	Ansan	Landrace	Korea
TN10	Yusong	Landrace	Korea
TN11	Hansom	Landrace	Korea
TN12	Jinju	Landrace	Korea
TN13	Jinbeak	Landrace	Korea
TN14	Suwon	Landrace	Korea
TN15	Annam	Landrace	Korea
TN16	Yanbeak	Landrace	Korea
TN17	Pungsan	Landrace	Korea
TN18	Yanghok	Landrace	Korea
TN19	Keonhok	Landrace	Korea
TN20	Huangbeak	Landrace	Korea
TN21	Sodong	Landrace	Korea
TN22	Hoahuk	Landrace	Korea
TN23	Namsak	Landrace	Korea
TN24	Seongbuk	Landrace	Korea
TN25	Hansan	Landrace	Korea
TN26	Dasak	Landrace	Korea
TN27	Sunhok	Landrace	Korea
TN28	Namda	Landrace	Korea
TN29	Hukseok	Landrace	Korea
TN30	Pungnam	Landrace	Korea
TN31	Pungan	Landrace	Korea
TN32	Poonan	Landrace	Korea
TN33	Nambeak	Landrace	Korea

Table 2. *Continued*

TN34	Dubeol	Landrace	Korea
TN35	Seonbeak	Landrace	Korea
TN36	Manhuk	Landrace	Korea
TN37	Kangbeak	Landrace	Korea
TN38	Manhuk	Landrace	Korea
TN39	Jinki	Landrace	Korea
TN40	Kopum	Landrace	Korea
TN41	Mihuk	Landrace	Korea
TN42	Yunbuk	Landrace	Korea
TN43	Yubeak	Landrace	Korea
TN44	Yoonhuk	Landrace	Korea
TN45	Yangan	Landrace	Korea
TN46	Seonhuk	Landrace	Korea
TN47	Chungmo5002	Landrace	Korea
TN48	Chongmo5003	Landrace	Korea
TN49	Yumi	Landrace	Korea
TN50	Kalmi	Landrace	Korea
TN51	Chongmo5005	Landrace	Korea
TN52	Chongmo5006	Landrace	Korea
TN53	Chongmo5007	Landrace	Korea
TN54	Jinyul	Landrace	Korea
TN55	Pungseong	Landrace	Korea
TN56	Anbeak	Landrace	Korea
TN57	Yean	Landrace	Korea
TN58	Manri	Landrace	Korea
TN59	Ouarung	Landrace	Korea
TN60	Hokeon	Landrace	Korea
TN61	Beakson	Landrace	Korea

The length of a row was 1.4 m with eight plants. During the experiment, 16 quantitative and five qualitative traits (Table 3 See Results section) were measured on 5 randomly-selected

healthy plants. However, for flowering, maturity, biomass and yield data were recorded on the unit plot basis. NPK fertilization was provided by the ratio 2.9:3.1:3.2 kg per acre. The recommended cultivation practices were followed during the experiment.

It is worth mentioning that the implementation and collection of the experimental data at Miryang was carried out by the research scientist Kim Sung-Up and colleagues from Department of Southern Area Crop, National Institute of Crop Sciences, Rural Development Administration.

Preparation of samples for content analysis

The harvested seeds were immediately washed with sterile water and air-dried for 3 days at room temperature, and then stored at 4℃ prior to analysis. Due to the limitation of seed quantity, a total of 72 accessions were processed from the developed core collection (102 accessions, See Results section).

Oil quantification

Oil content was determined by the Soxhlet method using a Büchi B-811 extraction system (Büchi Labortechnik AG, Flawil, Switzerland). Briefly, a 2 g pulverized seeds (60 mesh) was weighed into an extraction thimble (25 × 100 mm) covered by glass wool. The loaded thimble was then inserted into the Büchi B-811 extraction system with an addition of *n*-hexane (150 mL). The mixture was boiled at 105℃ for 180 minutes followed by 30 minutes cooling step in a dessicator. The total oil content was calculated on the sesame seeds dry weight basis.

Fatty acid, sesamin and sesamolin quantification

Chemicals. Acetic acid, palmitic acid methyl ester, stearic acid methyl ester, oleic acid methyl ester, linoleic acid methyl ester, linolenic acid methyl ester, sesamin, sesamolin, and trifluoroacetic acid were purchased from Sigma Chemical Company (St. Louis, MO, USA). Analytical grade n-hexane, methanol, toluene and water were purchased from J.T. Baker (Phillipsburg, NJ, USA).

Instruments. High Performance Liquid Chromatography (HPLC) was performed using a Dionex Ultimate 3000 RSLC system equipped with degasser, binary pump, diode-array detector, auto-sampler (Thermo Scientific, Germering, Germany) for sesamin and sesamolin analysis. Lipid was determined using a Buchi B-811 (Büchi, Switzerland) soxhlet extraction system. Nitrogen content was determined by a rapid N exceed® (Elementar Analysensysteme GmbH, Germany). Gas Chromatography (GC) was performed using an Agilent 7890A series (Santa Clara, USA) equipped with flame ionization detector (FID) for fatty acid analysis.

Preparation of sample and sesamin and sesamolin calibration curve. The dried seeds of

sesame were pulverised (60 mesh) for 3 min using a HR 2860 coffee grinder (Philips, Drachten, Netherlands), and each sample (1.0 g) extracted in 20 ml of 80% methanol for 24 h at room temperature in a shaking incubator. The supernatant was centrifuged at 3,000g for 3 min, and then filtered through a 0.2 μm syringe filter (Whatman Inc., Maidstone, UK) prior to HPLC analysis. For quantification, the peak areas of the isolated compounds were integrated from the HPLC chromatogram at 330 nm using Dionex software. The standard stock solutions were prepared by dissolving in methanol to obtain a 1 mg/ml concentration. Calibration curves were obtained with methanol at eight different concentrations (0, 5, 10, 20, 40, 60, 80, and 100 μg/ml). All calibration curves had coefficients of linear correlation r^2 > 0.999.

HPLC determination of sesamin and sesamolin contents. The quantification of sesamin and sesamolin contents in the seeds of sesame accessions was carried out using Ultimate 3000 HPLC analysis. A 10μl sample of the 80% methanol extract was injected into an analytical YMC-Triart C18 column (50 mm × 2 mm, 2 μm, YMC Co., LTD, Kyoto, Japan). The mobile phase was composed of 0.1% TFA in 60% methanol. The column temperature was maintained at 25 ˚C and the flow rate was 0.3 mL/min. The detector was held at a fixed wavelength of 290 nm.

Gas chromatography determination of fatty acid contents. Fatty acid components were elucidated using the gas chromatography Agilent 7890A (Santa Clara, USA) machine. Before analysis, fatty acid methyl esters (FAMEs) were prepared for gas chromatographic analysis by methylation of the extracted fat using water: methanol: toluene (1:20:10, v/v). The FAMEs were extracted with 2 mL hexane and 1 ℓL was injected into the gas-chromatograph, in split mode (split ratio 1:50). Fatty acid analysis was carried out on an Agilent gas chromatograph (Model 7890A GC) fitted with an automatic sampler (Model

7683B Injector) and FID detector. The conditions used were the following: HPFFAP capillary column (30 m × 0.318 mm I.D., 0.25 µm film thickness; Agilent Technologies), temperature programmed from 150 ℃ for 1 min, then 150 to 230 ℃ at 2.5 ℃/min, then held for 2 min. Carrier gas was nitrogen, column flow 1.0 mL/min, inlet and detector were set at 250 and 260 ℃, respectively.

Protein content quantification

The protein quantification was performed following Dumas combustion method in a rapid N exceed® (Elementar Analysensysteme GmbH, Germany) analyzer with 1 g sample weight. The crude protein was determined by multiplying the total nitrogen content by a factor of conversion 6.25 as described by Biancarosa et al., (2017)

Data analysis

The collected data were fully analyzed with the open-source statistical software R v.4.0.2(R Core Team, 2020). For reproducibility purpose, the data as well as the accompanying R code are made accessible (https://github.com/Yedomon/field_data_analysis/tree/main/Field_Paper_Final/Data_and_code_availability).

Data diagnosis and descriptive statistics. Using the dlookr v.0.4.2 (Ryu, 2021) and pastecs(Grosjean et al., 2018) packages, the outliers identification, data frequency distribution, Shapiro-wilk normality test and descriptive statistics were performed using *find_outliers()*, *plot_normality()*, *normality()*, and *stats.desc()* functions, respectively.

Diversity index analysis. For the qualitative traits, (branching type, capsule hairiness, flower color, inflorescence type and seed color) Shannon-Weaver (Shannon and Weaver, 1949) (Eq.

(1) and Simpson(Simpson, 1949) (Eq. (2) diversity indexes were calculated with the function *diversity()* of the package vegan (Oksanen et al., 2020). The equations of the two indexes are:

$$\text{Shannon index (H)} = -\sum_{i=1}^{p} \left(\frac{n_i}{N}\right) \log \left(\frac{n_i}{N}\right) \quad (1)$$

and

$$\text{Simpson index (D)} = \frac{1}{\sum_{i=1}^{n} \left(\frac{n_i}{N}\right)^2} \quad (2)$$

where n is the number (n) of observations regarding one particular qualitative traits modality i divided by the total number of observations (N)

Variability of agronomic traits: Considering that Africa or Asia continents are predicted to be the center of diversity candidates, we performed a one-way analysis of variance with the factor continent of origin. A generalization of Welch's method using trimmed means was employed since homoscedasticity assumption of our dataset was not satisfied. The one-way analysis of variance was run using the function *ggbetweenstats()* of the ggstatsplot v.0.7.0 package (Patil, 2018) with the option type = "robust". For categorical variables, a Pearson Chi-square test was carried out with the function *ggpiestats(),* in order to depict the variation between Africa and Asian continents.

Correlation among traits. In order to assess correlation between agronomic parameters, Spearman correlation test were performed with the function *ggcorrmat()* of the package ggstatsplot v.0.70 (Patil, 2018).

Path coefficient analysis for yield and yield relative components. As correlation alone doesn't automatically mean causative effect, we executed the path coefficient analysis following Dewey and Lu (1959), to unravel direct or indirect effect between dried seed weight and relative yield components with the function *path.analysis ()* of the package agricolae v.1.3-3 (de Mendiburu, 2020).

Classification of the accessions based on the agronomic traits. In order to group accessions based on agronomic traits, a principal component analysis followed by hierarchical agglomerative clustering were done using the function *PCA()* and *HCPC()* of the packages FactoMineR v.2.4 (Lê et al., 2008) and factoextra v.1.0.7(Kassambara and Mundt, 2019) respectively. The Euclidean distance-based similarity followed by Ward classification method was employed for the clustering stage. To delineate the traits that characterize each cluster, a v test was carried out as described by Lê et al., (2008). Prior the principal component analysis, data were standardized using the *scale()* function of the rstats (R Core Team, 2020) package. The visualization of the multivariate analysis was rendered using the function *fviz_pca_biplot()* of the factoextra package v.1.0.7 (Kassambara and Mundt, 2019).

Core collection inference and quality evaluation. The R version of Core Hunter 3 (De Beukelaer et al., 2018) *viz* corehunter v3.2.1 was employed to determine a core collection by applying the average-entry-to-nearest-entry distance scheme based on Gower's distance metric (Gower, 1971). The Core Hunter phenotypic data was generated from the comma-separated values excel file format of the data using the *phenotypes()* function. Then, the core collection was inferred with the function *samplecore()*. The quality of the inferred core collection with regard to the worldwide panel was assessed with the following metrics suggested by Hu et al. (2000) and Kim et al. (2007):

- the coincidence rate of range CRR (%) **(Eq.**(3**),**

$$CRR\ (\%) = \frac{1}{n} \sum_{i=1}^{n} \frac{R_{Ci}}{R_{Wi}} \times 100 \qquad (3)$$

where R_{Ci} is the range of the core collection for the agronomic trait i, and R_{Wi} is the range of the worldwide panel for the trait i;

- the variable rate VR (%) **(Eq.** (4**)**

$$VR\ (\%) = \frac{1}{n} \sum_{i=1}^{n} \frac{CV_{Ci}}{CV_{Wi}} \times 100 \qquad (4)$$

where CV_{Ci} is the coefficient of variation of the core collection for the agronomic trait i, and CV_{Wi} is the coefficient of variation of the worldwide panel for the trait i;

- the variance difference percentage VDP (%) **(Eq.** (5**)**

$$VDP\ (\%) = \frac{1}{n} \sum_{i=1}^{n} \frac{|\sigma_{Wi} - \sigma_{Ci}|}{\sigma_{Ci}} \times 100 \qquad (5)$$

where σ_{Ci} is the variance of the core collection for the agronomic trait i, and σ_{Wi} is the variance of the worldwide panel for the trait i;

- the mean difference percentage MDP (%) **(Eq.** (6 **)**

$$\text{MDP (\%)} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\mu_{Wi} - \mu_{Ci}|}{\mu_{Ci}} \times 100 \quad (6)$$

where μ_{Ci} is the variance of the core collection for the agronomic trait i, and μ_{Wi} is the variance of the worldwide panel for the trait i.

Besides, the means difference significance between the core and the whole accessions sets were computed following a Student t-test (for productive axis length), Wilcoxon test (for capsule number, plant height, branch number, stem diameter, dried biomass, dried seed weight, thousand seed weight, number of days to 50% flowering, number of days to maturity, number of days between flowering and maturity, capsule length, capsule width) or generalized linear model with a poisson error distribution (for harvest index, number of capsule per leaf axil, number of locules per capsule).

Geographical map: The map was rendered with sf v.0.9-8 (Pebesma et al., 2021), ggspatial v.1.1.5 (Dunnington and Thorne, 2021), and ggplot2 v.3.3.3(Wickham, 2016) packages. The world shape file was retrieved from the University of California UC DAVIS geographic map data web repository[3].

[3] https://biogeo.ucdavis.edu/data/gadm3.6/gadm36_shp.zip

Results

Natural traits variation in the worldwide sesame panel

The range, mean, standard deviation, and coefficient of variation of measured traits are presented in the Table 3. The highest coefficient of variation was observed for dried seed weight (74.09%) followed by harvest index (72.39%), number of branches per plant (71.53%), number of capsules per plant (64.49%), and number of capsule per leaf axil (53.87%). Most of yield-related traits exhibited a wide range of variation indicating that phenotypic-based selection is appropriate for those traits.

Plant architecture including branching type is economically important trait that can affect crop productivity (Teichmann and Muhr, 2015). In the present panel, 61.73% highly branched (n > 10 branches), 5.12% bi-branched and only 0.2 % (the accession TN42) unbranched accessions were recorded (Table 3).

Table 3. Descriptive statistics of agronomic traits

	Traits	Trait code	Reference	Unit	Range	Mean	Standard deviation	CV(%)
Quantitative traits	Dried seed weight	DSW	IPGRI and NBPGR 2004	g	153.58	32.82	24.32	74.09
	Harvest index	HI	IPGRI and NBPGR 2004	-	0.48	0.12	0.09	72.39
	Number of branches per plant	BNU	UPOV 2014	count	42	5.47	3.91	71.53
	Number of capsules per plant	CNU	IPGRI and NBPGR 2004	count	1195	113.21	73.01	64.49
	Number of capsule per leaf axil	CNLA	IPGRI and NBPGR 2004	count	2	1.41	0.76	53.87
	Dried Biomass weight	DBI	Defined in this study	g	1141	302.63	138.80	45.86
	Productive axis length	PAL	UPOV 2014	cm	205	77.23	29.51	38.20
	Stem diameter	DIA	Defined in this study	mm	53.40	15.98	5.58	34.95
	Days to 50% flowering	FLO	IPGRI and NBPGR 2004	day	83	51.50	12.86	24.97
	Plant height	PHE	IPGRI and NBPGR 2004	cm	224	155.36	35.66	22.95

Table 3. *Continued*

Quantitative traits	Days from 50% flowering to 50% physiological maturity	FTM	Defined in this study	day	50	48.95	8.73	17.85
	Capsule length	CLE	IPGRI and NBPGR 2004	mm	37	25.08	4.47	17.83
	1000-seeds weight	TSW	IPGRI and NBPGR 2004	g	2.86	2.76	0.48	17.62
	Capsule width	CWI	IPGRI and NBPGR 2004	mm	15.2	7.75	1.18	15.28
	Days to 50% maturity	MAT	Defined in this study	day	99	100.4679443	13.15873489	13.09
	Number of locules per capsule	NLC	IPGRI and NBPGR 2004	1-3 scale	Four locules (99.21%), 6 locules (0.39%), 8 locules (0.39%)			

Table 3. *Continued*

Qualitative traits	Branching type	Branching_type	Defined in this study	1-6 scale	Highly branched [number of branches > 10] (61.73%), Moderately branched [3 < number of branches < 9] (18.93%),Extremely branched [Bushy-type] (10.84%), Bi-branched [two branches] (5.12%), Mono-branched [1 branch] (3.15%), unbranched (0.2%)
	Capsule hairiness	Cap_Hairiness	IPGRI and NBPGR 2004	1-4 scale	Sparse (81.06%), Medium (16.96%), Profuse (1.18%), Glaborous (0.78%)
	Flower colour	Flower_Colour	Defined in this study	1-3 scale	White (97.04%), Pink (2.76%), Purple (0.2%)
	Inflorescence type	Inflorescence_type	IPGRI and NBPGR 2004	1-2 scale	Indeterminate (99.60%), determinate (0.4%)
	Seed colour	Seed_Colour	Defined in this study	1-10 scale	White (33.72%), Brown (21.49%), Composite (20.31%), Black (9.66%), Light Brown (7.88%), Light gray (2.76%), Light yellow (1.57%), Dark brown (1.57%), Red brown (0.39%), Yellow (0.39%)

Most of the tested accessions exhibited white flower (97.4%), followed by pink (2.76%) and purple (0.4%) flowers (Figure 4A-C). The purple color was observed for the wild relative *Sesamum radiatum* (Figure 4A) whereas a typical pronounced pink color was showed by the wild *Sesamum alatum* (Figure 4B).

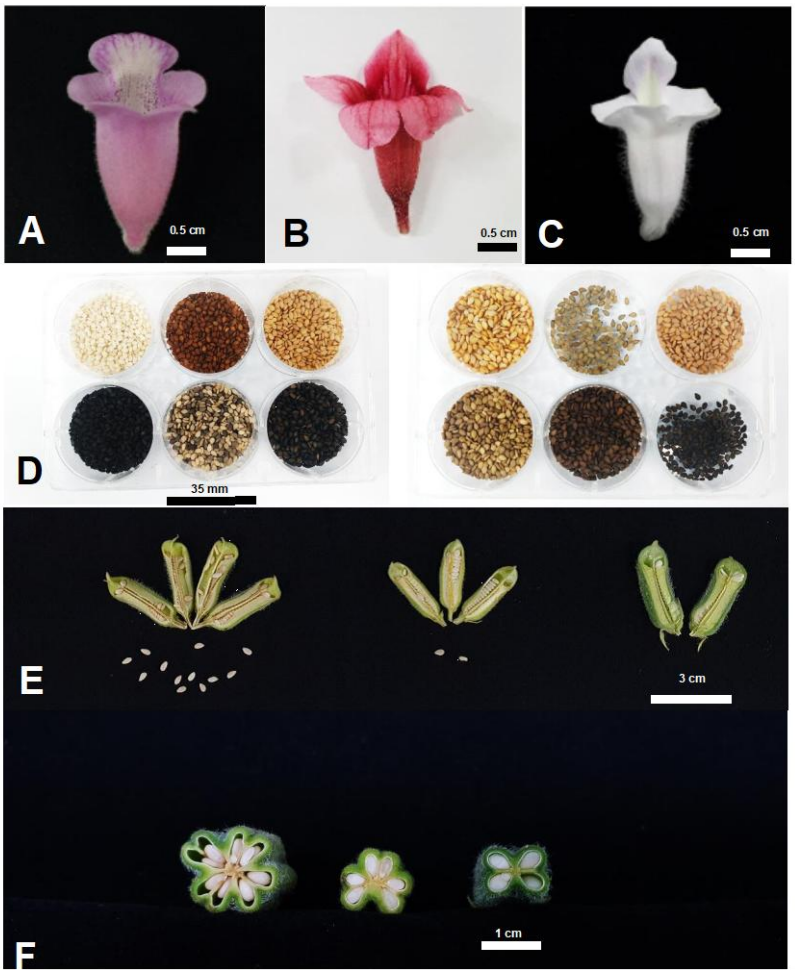

Figure 4. Photographs showing some sesame morphological characteristics.; Flower color of (A) *Sesamum radiatum*, (B) *Sesamum alatum*, (C) *Sesamum indicum* cv. Goenbaek; an overview of the seed color diversity (D); Vertical (E) and horizontal section view of capsule harboring eight, six four locules (from left to right) respectively.

A remarkable diverse (H = 1.72, D = 0.78) seed color was also noted (Table 4, Figure 4D), with 33.72% of white seed followed by 21.49% and 9.66 % of black seeds (Table 3). Interestingly, we identified some accessions that present six (T418) and eight (T109 and T324) locules per capsule (Figure 4E-F).

Table 4. Shanon-Weiner and Simpson diversity index for five qualitative traits

Traits	Shanon-Weiner Index (H)	Simpson's Index (D)
Inflorescence type	0.026	0.008
Flower color	0.141	0.057
Seed color	1.722	0.781
Branching type	1.127	0.568
Capsule hairiness	0.562	0.314

None of the accessions showed determinate growth habit except the induced determinate mutants dt-sel and dt-45 originated from Turkish sesame breeding program (Uzun and Çağırgan, 2006).

Relationship among traits

The coefficients of correlation matrix among agronomic traits were summarized in the Figure 5. A total of 41 positive ($p < 0.05$) and 45 negative ($p < 0.05$) coefficients of correlation were highlighted. For yield aspect, the strongest positive correlation was found between dried seed weight and harvest index ($r = 0.85$, $p < 0.001$). Similarly, number of days to flowering and number of days to maturation also exhibited a high relationship ($r = 0.85$, $p < 0.001$). In term of biomass, a positive relationship was revealed between stem diameter and dried biomass ($r = 0.69$, $p < 0.001$). The similar tendency was confirmed between plant

height and stem diameter (r = 0.73, p < 0.001) and plant height and dried biomass (r = 0.61, p < 0.001).

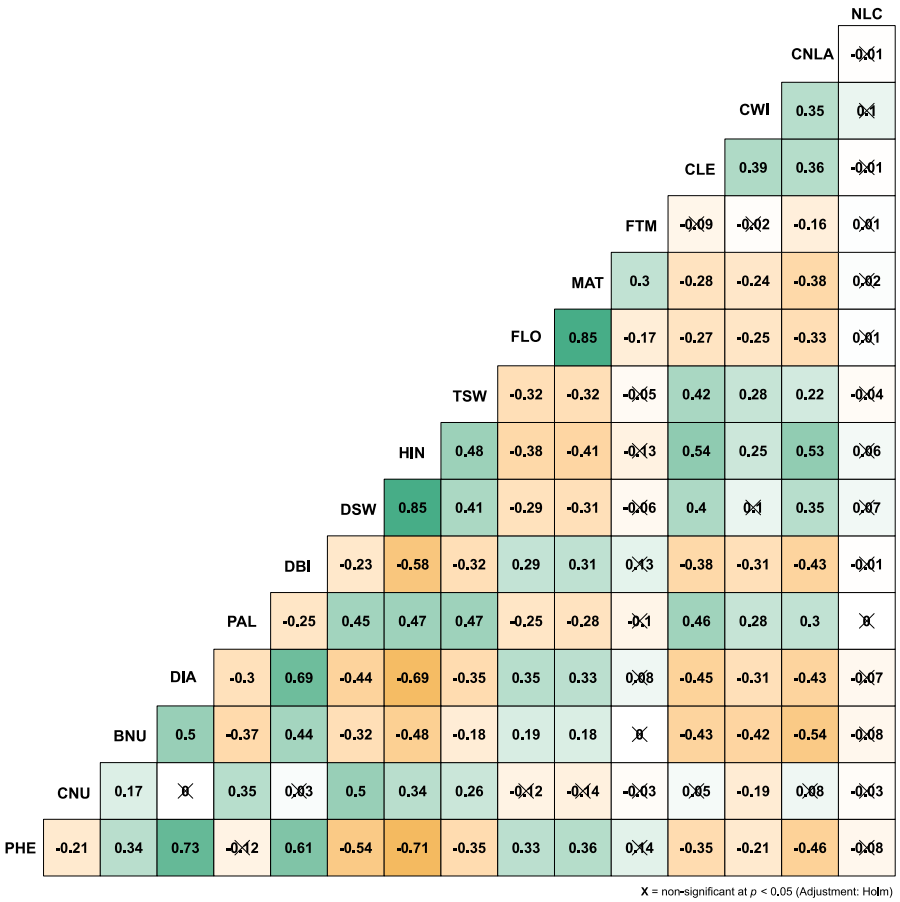

Figure 5. Correlation among quantitative traits.Significant correlations have not cross symbol on the values. Green and yellow square represents positive and negative correlation respectively.

Meanwhile, the highest negative relationship was detected between plant height and harvest index (r = -0.71, p < 0.001), followed by stem diameter and harvest index (r = -0.69, p < 0.001), and plant height and dried seed weight (r = -0.54, p < 0.001). Overall, biomass traits

augmentation seems to have a reduction effect on yield and yield components traits.

In order to clarify the effect of the studied traits with a focus on dried seed weight, a path coefficient analysis was carried out. Details results were presented in the Table 5. The traits, number of capsules (0.18), productive axis length (0.14), thousand seed weight (0.02), and number of locules per capsule (0.02) exerted a positive direct effect on dried seed weight.

Table 5. Direct and indirect effect of various traits on dried seed weight showed by path coefficient analysis

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
PHE	**-0.091**	-0.030	-0.021	-0.012	-0.007	0.199	-0.578	-0.007	-0.314	0.392	-0.075	0.001	0.014	0.050	-0.001
CNU	0.016	**0.177**	-0.007	-0.001	0.044	0.007	0.215	0.006	0.157	-0.162	0.006	0.000	0.009	-0.008	0.000
BNU	-0.033	0.020	**-0.060**	-0.009	-0.051	0.148	-0.396	-0.005	-0.191	0.196	0.006	0.001	0.029	0.056	-0.001
DIA	-0.062	0.009	-0.030	**-0.017**	-0.040	0.228	-0.537	-0.007	-0.314	0.346	-0.017	0.001	0.019	0.043	-0.001
PAL	0.005	0.057	0.022	0.005	**0.139**	-0.090	0.330	0.011	0.269	-0.312	0.034	-0.001	-0.018	-0.030	0.000
DBI	-0.050	0.004	-0.024	-0.011	-0.035	**0.361**	-0.438	-0.006	-0.280	0.335	-0.046	0.001	0.019	0.041	0.000
HIN	0.064	0.046	0.029	0.011	0.055	-0.192	**0.826**	0.010	0.359	-0.450	0.080	-0.001	-0.017	-0.060	0.001
TSW	0.029	0.050	0.013	0.006	0.067	-0.101	0.355	**0.022**	0.336	-0.358	0.017	-0.001	-0.019	-0.025	-0.001
FLO	-0.026	-0.025	-0.010	-0.005	-0.033	0.090	-0.264	-0.007	**-1.121**	0.993	0.121	0.001	0.014	0.032	0.000
MAT	-0.031	-0.025	-0.010	-0.005	-0.037	0.105	-0.322	-0.007	-0.964	**1.154**	-0.184	0.001	0.015	0.041	0.000
FTM	-0.012	-0.002	0.001	-0.001	-0.008	0.029	-0.116	-0.001	0.235	0.369	**-0.574**	0.000	0.002	0.016	0.000
CLE	0.026	0.005	0.024	0.007	0.061	-0.112	0.396	0.009	0.258	-0.289	0.029	**-0.003**	-0.023	-0.039	-0.001
CWI	0.021	-0.025	0.027	0.005	0.040	-0.112	0.231	0.007	0.258	-0.277	0.023	-0.001	**-0.062**	-0.047	0.002
CNLA	0.043	0.012	0.031	0.007	0.039	-0.137	0.462	0.005	0.336	-0.439	0.086	-0.001	-0.027	**-0.107**	-0.001
NLC	0.006	-0.002	0.003	0.001	-0.003	-0.004	0.025	-0.001	-0.011	0.023	-0.006	0.000	-0.005	0.003	**0.020**

In bold and positive: direct effect; in bold and negative: indirect effect; See Table 3 for trait details.

However, number of days to flowering (-1.12), plant height (-0.09), flowering to maturity days (-0.57), branch number (-0.06), and stem diameter (-0.01) exhibited indirect effect on dried seed weight.

Altogether, both correlation and path coefficient analysis pinpointed the positive contribution of the number of capsules, number of locules per capsule, and the productive axis length for dried seed weight.

Phenotypic-based clustering

The hierarchical classification of the accessions resulted in three clusters (Figure 6). The cluster 1, 2, and 3 grouped 35.37%, 41.50%, and 23.12% of the total number of accessions, respectively.

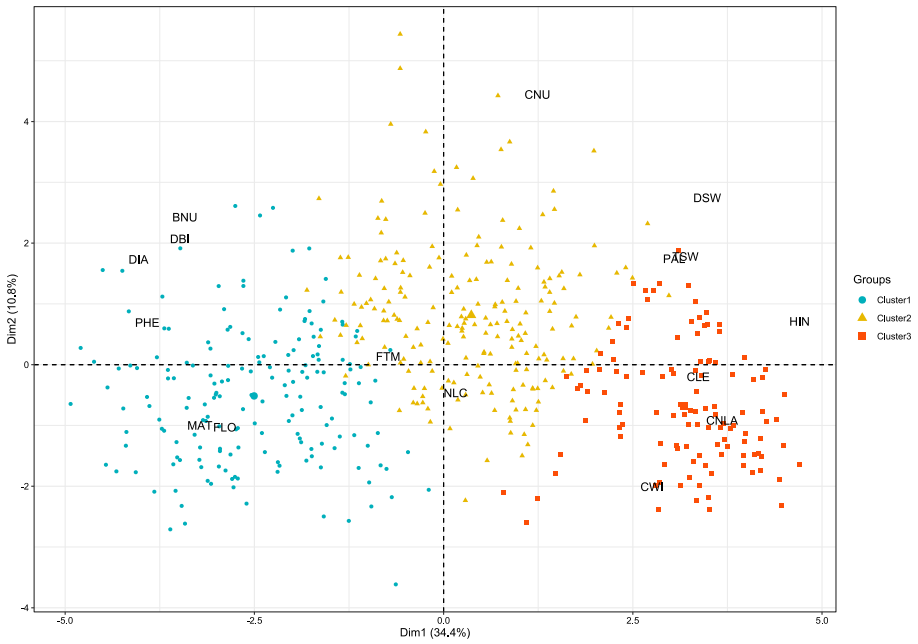

Figure 6. Clusters representation of the accessions following quantitative traits.

Traits details are provided in the Table 3.

The quantitative traits that described each cluster are summarized in the Table 6.

Table 6. Quantitative traits associated to each cluster from the worldwide panel following the v test

Clusters	Traits	Mean in the cluster	sd in the cluster	Overall mean	Overall sd	v test	p-value
Cluster 1	DIA	18.91	3.25	16.08	3.66	12.86	p < 0.01
	PHE	178.45	19.32	156.09	29.86	12.46	p < 0.01
	DBI	393.75	124.73	316.25	127.62	10.10	p < 0.01
	FLO	60.82	10.69	54.16	11.20	9.88	p < 0.01
	MAT	107.43	10.56	100.88	11.46	9.49	p < 0.01
	BNU	6.86	2.32	5.50	2.48	9.04	p < 0.01
	CNU	137.95	45.94	112.68	45.63	10.49	p < 0.01
Cluster 2	DSW	38.07	18.81	28.93	20.22	8.56	p < 0.01
	TSW	2.81	0.32	2.64	0.43	7.45	p < 0.01
	PAL	81.62	17.12	73.38	21.53	7.24	p < 0.01
	FTM	48.03	6.66	46.71	6.04	4.13	p < 0.01
	HIN	0.12	0.05	0.11	0.08	3.08	p < 0.01
	CNLA	2.44	0.82	1.42	0.76	16.41	p < 0.01
Cluster 3	HIN	0.21	0.07	0.11	0.08	14.46	p < 0.01
	CWI	8.63	1.05	7.72	0.93	12.12	p < 0.01
	CLE	28.16	3.37	24.95	3.73	10.61	p < 0.01
	DSW	40.30	14.57	28.93	20.22	6.93	p < 0.01
	TSW	2.87	0.28	2.64	0.43	6.77	p < 0.01
	PAL	84.06	18.33	73.38	21.53	6.11	p < 0.01

sd: standard deviation, See Table 3 for trait codes details.

The cluster 1 encompassed the accessions that exhibited high biomass and low yield. The cluster 2 is characterized by late maturing and moderately yield-performing accessions while the cluster 3 representing the elite accessions harboring high-yield attributes (Table 6, Figure 6). Most accessions of the cluster 3 (69.23%) originated from eastern Asia (Figure 7).

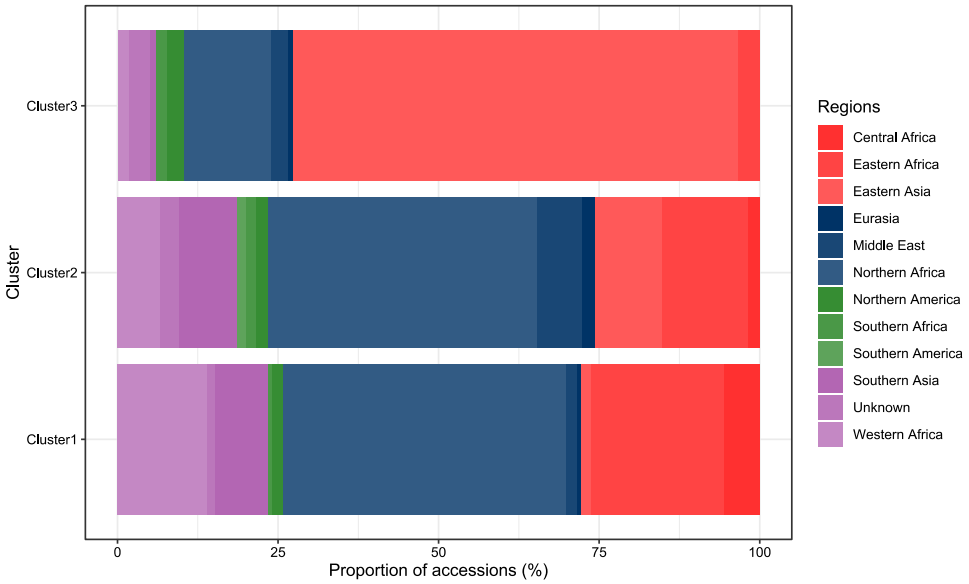

Figure 7. Relative contribution of region of origin per cluster

Interestingly, African representatives (13.67% for northern Africa, 3.42% for eastern Africa, 1.71% for western Africa, 1.71 % for southern Africa) are the second largest group that exhibited high-yield performance (Figure 7).

Core collection inference

From 506 accessions, the Core Hunter 3 program generated a core collection encompassing 102 accessions. The number of retained accessions following geographical position is presented in the Figure 8. The top 3 most contributing regions were eastern Asia (n = 34), followed by the northern Africa (n = 30) and the eastern Africa (n = 10) (Figure 8).

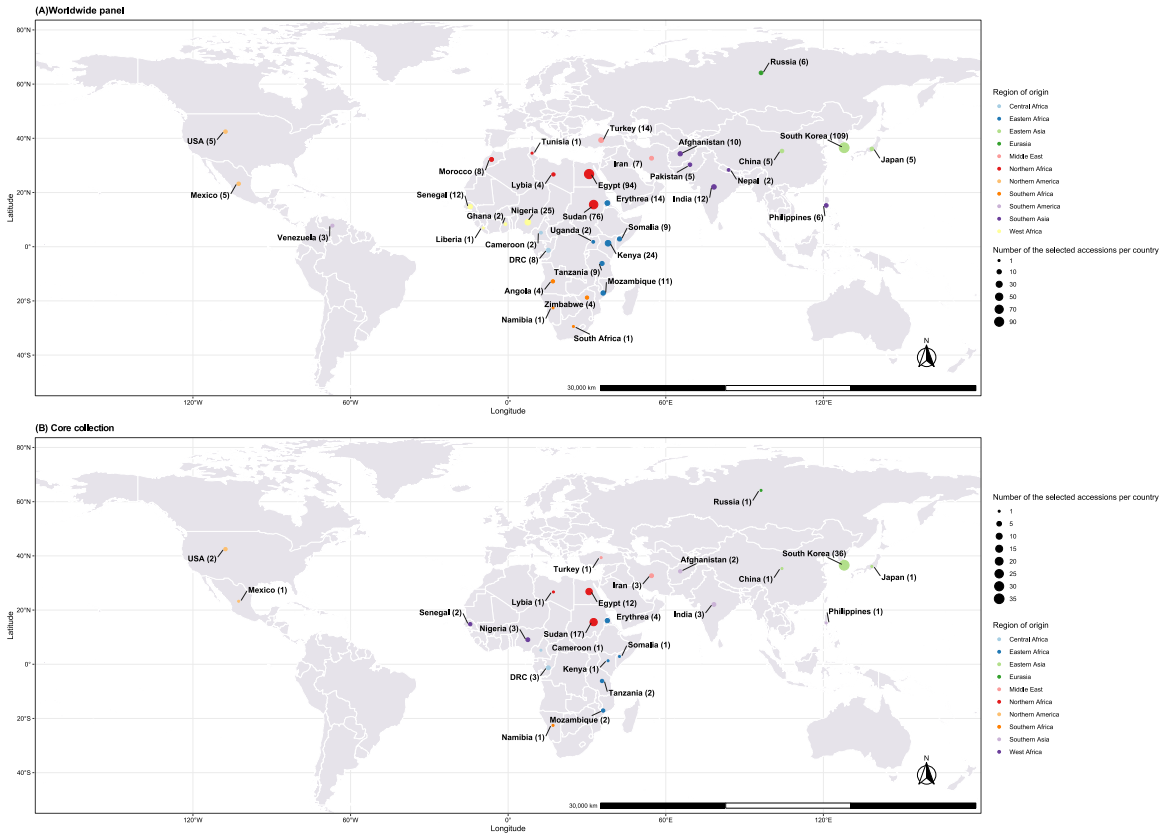

Figure 8. Map showing the quantitative distribution of accessions for the worldwide panel (A) and the inferred core collection (B).

The evaluation of the core collection quality revealed a variation of the coincidence rate of range (CRR) per trait from 50% to 100% with an overall value of 78.04% (Table 7).

The variable rate per (VR) trait was ranging from 89.65% to 116.87% with an overall VR of 100.49%. More interestingly, there was no significant difference ($p > 0.05$) between the core and the worldwide collection for all traits. This result was supported by the low overall mean difference (3.61%) and variance difference (14.39%) percentages (Table 7).

Table 7. Metrics showing the quality of the inferred core collection

	Range			CV			Mean			Variance			
Traits	Whole	Core	CRR (%)	Whole	Core	VR(%)	Whole	Core	MDP(%)	Whole	Core	VDP(%)	p-value*
PHE	171.30	168.5	98.37	0.19	0.21	108.38	156.12	151.00	3.39	891.91	980.00	8.99	p = 0.18
CNU	394.70	231.7	58.70	0.41	0.36	89.65	112.80	111.17	1.47	2089.04	1630.70	28.11	p = 0.98
BNU	15.50	11.36	73.28	0.45	0.52	114.42	5.51	5.05	8.99	6.18	6.81	9.27	p = 0.10
DIA	21.20	19.49	91.94	0.23	0.25	107.83	16.08	15.59	3.19	13.40	14.63	8.42	p = 0.18
PAL	123.22	102	82.78	0.29	0.34	114.42	73.41	71.02	3.36	464.06	568.71	18.40	p = 0.35
DBI	757.50	549.50	72.54	0.40	0.37	90.74	316.27	303.22	4.30	16287.15	12328.18	32.11	p = 0.52
DSW	114.36	102.81	89.90	0.70	0.64	91.13	28.90	30.62	5.65	409.34	381.83	7.20	p = 0.38
HIN	0.39	0.31	77.94	0.75	0.70	92.80	0.11	0.12	10.91	0.01	0.01	7.83	p = 0.71
TSW	2.42	2.10	86.59	0.16	0.15	93.02	2.64	2.63	0.53	0.18	0.16	16.81	p = 0.51
FLO	53	36	67.92	0.21	0.19	90.75	54.15	54.97	1.49	125.63	106.62	17.82	p = 0.31
MAT	59.50	42	70.59	0.11	0.11	95.92	100.88	101.47	0.58	131.36	122.27	7.43	p = 0.62
FTM	50	40.50	81	0.13	0.15	116.87	46.72	46.49	0.49	36.56	49.45	26.06	p = 0.34
CLE	27	19.40	71.85	0.15	0.15	97.10	24.94	24.96	0.06	13.91	13.13	5.94	p = 0.71

Table 7. *Continued*

CWI	7.38	5.55	75.25	0.12	0.13	105.22	7.72	7.89	2.21	0.87	1.00	13.62	p = 0.07
CNLA	2	2	100	0.54	0.54	100.51	1.42	1.59	10.71	0.58	0.74	21.09	p = 0.20
NLC	4	2	50	0.07	0.07	99.08	4.02	4.04	0.38	0.08	0.08	1.08	p = 0.94
			78.04			100.49			3.61			14.39	-

*Whole versus core means comparison p-value for each trait, CV: Coefficient of variation, CRR: Coincidence rate of range, VR: Variable rate, VDP: Variance difference percentage

Identification of candidate genotypes for oil, protein, fatty acid, sesamin and sesamolin contents

Out of 102 accessions of the core collection, we were able to extracted and quantify oil, fatty acid and lignans contents for 72 accessions due to the seed quantity limitation. The hierarchical classification based on oil, lignans, fatty acid and the agronomic traits resulted in three clusters (Figure 9A). The cluster 1 grouped the accessions that show not only high-yield attributes but also are rich in proteins and alpha linoleic acid content (Table 8). The cluster 2 highlighted the accessions that exhibit higher yield while the cluster 3 is characterized by oil rich accessions with higher sesamin, sesamolin, and linoleic acid contents (Table 8).

Based on their relative contributions to the construction of the first factorial plan, the accessions T218, T077, T419, and T148 (Figure 9B) appeared as valuable candidates for lignan-oriented breeding. The accession TN03 and T415 were highlighted for oleic acid and protein contents respectively (Figure 9B).

Table 8. Quantitative traits associated with each cluster from the core collection

Clusters	Traits	Mean in the cluster	sd in the cluster	Overall mean	Overall sd	v test	p-value
Cluster1	CNLA	2.59	0.67	1.74	0.90	6.55	$p < 0.01$
	HIN	0.22	0.05	0.15	0.07	6.05	$p < 0.01$
	CWI	8.83	0.79	8.01	1.06	5.31	$p < 0.01$
	Protein	28.90	1.23	26.97	2.57	5.19	$p < 0.01$
	Oleic acid	44.78	2.52	42.08	3.73	5.01	$p < 0.01$
	CLE	28.03	2.23	25.80	3.23	4.78	$p < 0.01$
	TSW	2.86	0.22	2.73	0.34	2.72	$p < 0.01$
	Alpha linoleic acid	0.41	0.19	0.35	0.16	2.50	$p < 0.01$
	PAL	83.59	20.12	77.71	19.88	2.05	$p < 0.01$
Cluster 2	CNU	150.52	30.91	122.07	33.67	4.42	$p < 0.01$
	BNU	5.76	2.00	4.39	2.37	3.01	$p < 0.01$
	DSW	45.82	18.51	36.85	16.85	2.78	$p < 0.01$
	DBI	325.84	83.47	280.53	101.93	2.32	$p < 0.01$

Table 8. *Continued*

Clusters	Traits	Mean in the cluster	sd in the cluster	Overall mean	Overall sd	v test	p-value
	PHE	173.35	19.11	146.12	28.25	5.56	p < 0.01
	Linoleic acid	46.37	2.54	43.01	3.59	5.39	p < 0.01
	MAT	107.03	9.68	99.75	10.48	4.01	p < 0.01
	Oil	52.85	3.29	50.15	3.88	4.01	p < 0.01
Cluster3	DIA	17.11	3.75	14.75	3.46	3.94	p < 0.01
	Sesamin	4.39	1.67	3.08	1.99	3.83	p < 0.01
	FTM	50.72	7.88	46.29	6.75	3.79	p < 0.01
	DBI	340.18	96.74	280.53	101.93	3.38	p < 0.01
	BNU	5.66	1.79	4.39	2.37	3.10	p < 0.01

sd: standard deviation, See Table 3 for trait codes details

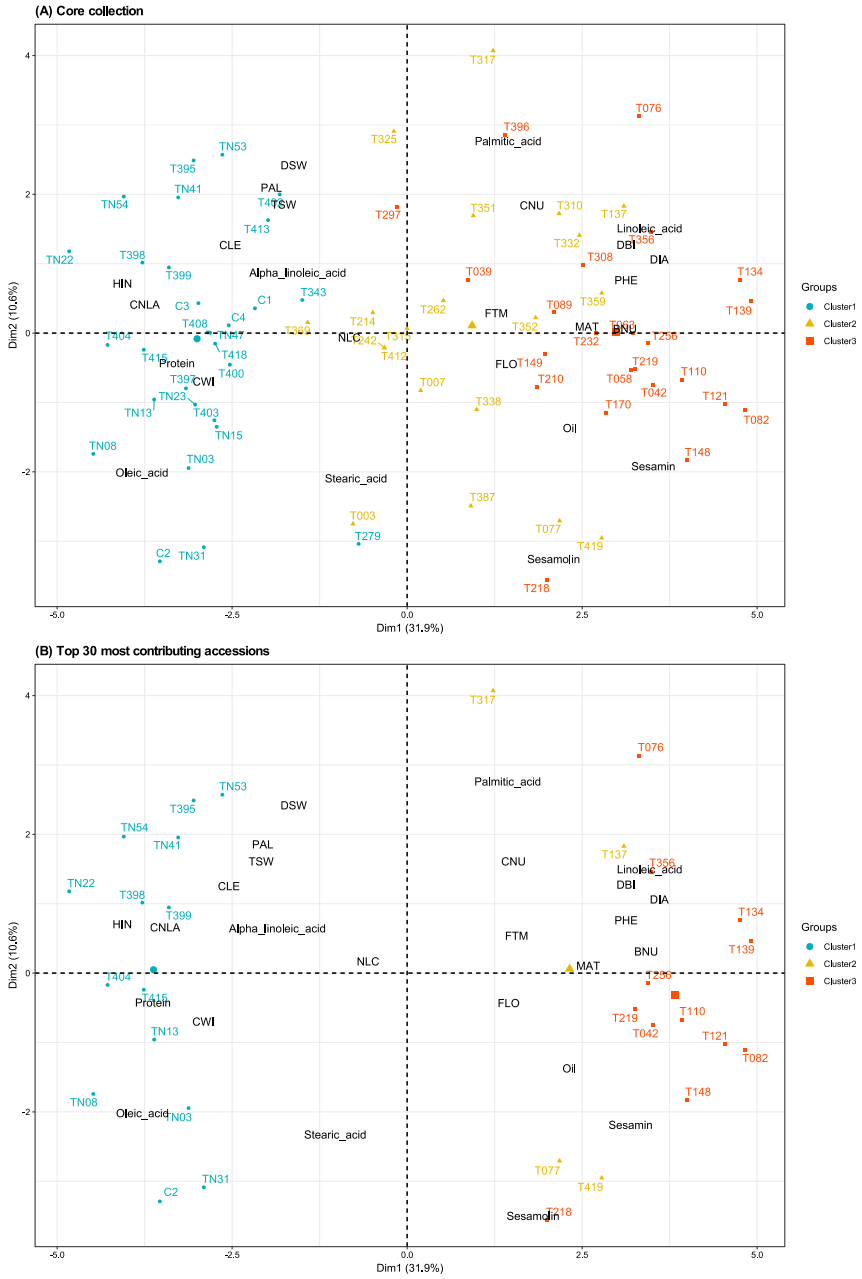

Figure 9. Projection map showing accessions, traits, and clusters for the core collection (A) and the top 30 most contributive accessions (B). Traits details are provided in the Table 3

Discussion

The present study reports a comprehensive view of the phenotypic variability of a worldwide sesame panel from the Korean genebank and the development of a multi-purpose core collection regarding agronomic and nutritional traits.

The wide range variability observed among the accessions for the studied traits provide a scope for selection and set a path for the identification of novel genotypes with desirable traits.

The study showed that some accessions (most from western and central African continent) exhibited important leafy biomass, are taller and less productive compared to the Asian representatives. Despite sesame leaves nutritional values have been neglected (mainly due to the oilseed trait) by the scientific community (Bedigian, 2018), it is widely consumed in some African countries including Benin, Togo, Niger, Burkina-Faso, Nigeria, Sudan as leafy-vegetables and employed as remedy (Bedigian, 2018).

The delay of the flowering and maturation days for these African accessions was also observed. More interestingly, certain did not enter in the maturity stage or even not flower. These observations indicate the photoperiodism sensitivity of some African accessions in our experimental environment (35° N latitude). Therefore, the photosensitivity appears as an adaptative trait for discriminating some African genotypes. Similarly, Yingzhong and Yishou (2002) reported some tropical accessions that did not flower at the latitude of 40°N. Despite this phenomenon, valuable African representative performed well at 35° N Latitude with early flowering and high yield and nutritional characteristics. Most of those accessions are from eastern and northern Africa, representing acceptable resources as parental genotype for population development in the tested environment.

Wide-range seed color was also observed in the worldwide panel. As suggested by Andargie et al. (2021), sesame seed color may undergo intensive selection by human so far. As a result, extensive seed color variability occurred. This observation was in line with others agronomic traits including number of locules per capsule, branching type and number of capsule per leaf axil.

For yield-oriented breeding, the knowledge about the yield components traits that has a direct or indirect impact of yield is paramount for efficient yield-related breeding. Herein, we investigated the cause and effect relationship among yield and yield-component traits. The results highlighted the number of capsules, number of locules per capsule, and the productive axis length as key traits that has a direct effect on dried seed weight. Similar studies carried out in India (Subramanian and Subramanian, 1994), Turkey (Uzun and Çağırgan, 2006), and China (Yingzhong and Yishou, 2002) support the present findings. Therefore, these traits may be considered as index for parental material selection for yield improvement. Specifically, the accessions T109 and T324 that harbor eight locules per capsule constitute valuable candidates as parental genotype.

Meanwhile, we were able to define a core collection that preserve the phenotypic variability from the whole set. The core collection size is about 20% from the initial worldwide set, suggesting that the inferred core set contains the minimum of repetitiveness. Comparable core set size was previously reported for sorghum (Girma et al., 2020) and safflower (Kumar et al., 2016) with 24% and 31% respectively. Moreover, a non-significant difference between whole and core set for all traits was noted, supporting the fact that the core set maintained the genetic diversity. It is also valuable to mention a geographical broad representativeness of the inferred core collection in contrast with the previous core collection developed by Park et al. (2015). These observations support the good quality of

the inferred core collection for effective usage in sesame breeding through genome-wide association studies for the dissection of the genetic basis of the desirable traits.

The study presents the first investigation of lignans content from a worldwide set of sesame accessions. The lignans is known to have multiple health benefits for human (Andargie et al., 2021). The candidate genotypes provided by this study constitute a valuable resource regarding lignan-oriented breeding. Besides, for nutritional purposes, we also identified candidate rich-protein and lipid content accessions that can serve as parental material for population development. We also suggest as further investigation, to screen the high leafy biomass accessions for leaf nutritional content. It may be a valuable fiber dietary alternative that can help to feed people in the current context of the increasing population.

As part of the Korean genomic-assisted sesame breeding, these initial results pave the way for the identification of genomic regions responsible of the expression of the desirable agronomic traits. Ultimately, the proposed core collection would lay a foundation for association mapping studies for effective sesame breeding regarding oil, protein and lignans contents.

CHAPTER 2: Construction of a High Quality Chromosome-Scale Genome of the Korean Variety *Sesamum indicum* var. Goenbaek

Summary

Sesamum indicum seed is omnipresent in Korean dietary habit as food decoration and the oil is used in traditional medicine due to the natural antioxidant content. Although genomic resource is available mainly from Chinese and India genotypes, the lack of Korean genomic resource hampered the design of novel genotypes with valuable agronomic importance as well as food and health enhancing properties. The present study aimed to generate a high-quality chromosome-level genome assembly and annotation of the Korean elite cultivar Goenbaek that is richer in oil, sesamin and sesamolin contents, and demonstrated resistance to Phytophtora blight disease. The combination of short reads, long reads, and chromosome conformation capture helped to assemble a highly contiguous genome of 13 chromosomes spanning 262.5 Mbp with an N50 value of 19.9 Mbp. From the assembled genome and intrinsic RNA sequencing data, a total of 23,539 protein-coding genes were predicted. Phylogenetic analysis placed Goenbaek in a distinct branch suggesting geographical or adaptive pattern leading to genetic diversity. The orthology analysis revealed that Goenbaek exhibited a set of 1,171 unique gene families enriched in lipid metabolism and biosynthesis, fatty acid metabolism and phenylpropanoid biosynthesis; which confirmed the rich oil content of Goenbaek. Besides, health beneficial gene repertoire including sphingolipid and selenocompounds metabolism also contributed to the Goenbaek dispensable genes. Furthermore, genome-wide investigation of late abundant embryogenesis (LEA) protein-coding genes revealed the presence of 71 LEA proteins including 56, 5, 4, 4, 1, and 1 member(s) for LEA2, SMP, LEA1, LEA4, LEA3 and LEA5 respectively. The organ-wide expression of LEA2 genes suggests their potential key role in plant growth promotion and abiotic stress mitigation. We constructed in the present study, a new reference-grade sesame

genome that will serve for the Korean genomics-assisted sesame breeding program for agronomic traits improvement, oil and specialized health promoting metabolites investigation.

Introduction

Sesamum indicum L. is one of the major oil crop traditionally used in Korean food culture as cooking oil, seasoning or decoration food (Kim et al., 2016). This member of mint family harbors valuable nutritional and health benefits to human mainly due to its lignans (sesamin, sesamolin, sesaminol and sesamol). Based on the recent Korean nutritional and health survey, the daily lignan intake from sesame oil is up to 77% with 18.39 mg and 13.26 mg per person for males and females respectively (Kim et al., 2020). From last decade, tremendous effort brought this orphan crop into the genomics era (Dossa et al., 2017b). Initial genome was rendered using BAC clone and short-read approaches with Yuzhi11 cultivar (Zhang et al., 2013c). Meanwhile, the Oil Crop Research Institute (OCRI) generated a contig-level assembly (Wang et al., 2014) that have been updated in linkage groups later on (Wang et al., 2016). The availability of the reference Chinese cultivar Zhongzhi13 facilitated the identification and validation of molecular marker relative to seed coat color, plant height (Wang et al., 2016), seed yield components (Wei et al., 2015; Zhou et al., 2018), drought and salinity stress (Li et al., 2018; Dossa et al., 2019), root biomass development (Dossa et al., 2020), and water logging stress (Wang et al., 2021).

Moreover, a pangenome was constructed based on one Indian cultivar (*S. indicum* cv. Swetha) and four Chinese genotypes including two cultivars (*S. indicum* var. Zhongzhi13 and *S. indicum* var. Yuzhi11) and two landraces (*S. indicum* cv. Baizhima and *S. indicum* cv. Mishuozhima) (Yu et al., 2019a). The strategy employed for the pangenome construction relied on reference-based assembly with Zhongzhi13 as a backbone representative. Therefore, the need of reference-free high-quality genome resources may help to get insights into the variability among genotypes relative to the traits of interest.

The Korean modern cultivar Goenbaek exhibited three capsules per node, unbranched pattern, high seed yield characteristics (1.6 ton per hectare), oil (50.2% of crude fat) and lignans-rich contents (sesamin: 3.96 mg/g, sesamolin: 2.57 mg/g), and resistance to the sesame Phytophtora blight disease (Kim et al., 2018; Asekova et al., 2021). Having a good quality genome will drastically improve the ongoing Korean sesame breeding activities for oil, lignans (Ha et al., 2017), and disease resistance (Asekova et al., 2021). Thus, a chromosome-grade genome assembly of this cultivar was constructed to accelerate the genomics-assisted sesame breeding program.

Materials and Methods

Tissue sampling, DNA and RNA extractions

Young leaves of *S. indicum* var. Goenbaek were collected, flash-frozen in liquid nitrogen and stored at -80°C prior to extraction. The DNA was extracted based on a modified CTAB protocol (Allen et al., 2006). Regarding the RNA extraction, samples from roots, stem, leaves and apical meristem were extracted using RNeasy® Plant Mini Kit (Qiagen, Germany). All samples were checked for quality using Nanodrop ND 1000 (Thermo Fisher, Waltham, MA, USA) and stored at -80°C prior sequencing.

Short-Reads Sequencing

After a second quality check with Qubit Fluorometer (Thermo Fisher, Waltham, MA, USA), 1µg of genomic DNA was sheared to ~550 bp fragments. Then, the library preparation was performed with a TruSeq Nano DNA Prep Kit (Illumina, San Diego, CA, USA) following manufacturer's protocol. Taking advantage of the NovaSeq 6000 System (Illumina, San Diego, CA, USA), a 100 bp paired-ends sequencing data was generated.

RNA sequencing library preparation was performed using the Illumina TruSeq RNA sample preparation kit following the manufacturer's instructions. Afterwards, 151 bp paired-ends sequencing using Novaseq6000 platform has been executed for seed, root, apical meristem, stem and leaf organs.

Genome Size Estimation

Using the short-reads data, a k-mer based genome size estimation was performed with GenomeScope v.2.0 (Ranallo-Benavidez et al., 2020) tools with k-mer depth setting k = 19. The estimated genome size was 321 Mbp (Figure 10). Based on this indicative genome size information, we conducted a consequent PacBio long-reads sequencing.

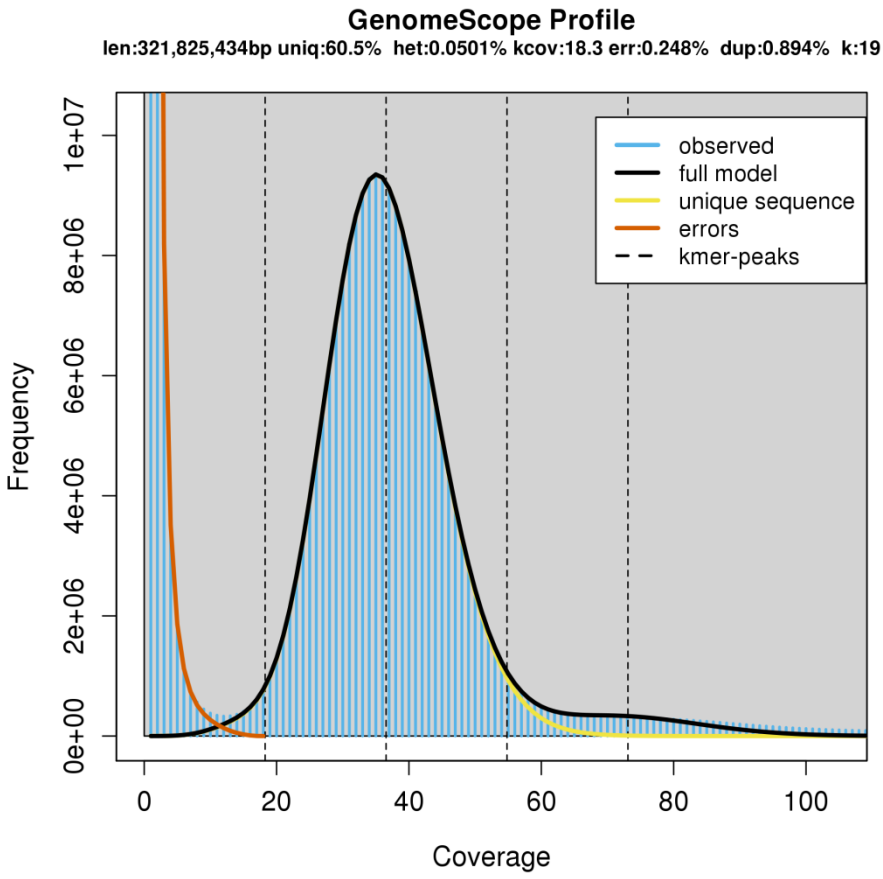

Figure 10. GenomeScope profile of *Sesamum indicum* var. Goenbaek using Illumina short reads via k-mer prediction at k = 19

Based on this indicative genome size information, we conducted a consequent PacBio long-reads sequencing.

PacBio Long Reads Sequencing and Initial Assembly

Using a high molecular weight of genomic DNA, a ~20 kb SMRT bell library was constructed following SMRTbell™ Libraries' protocol (Pacific Biosciences, Melon Parl, CA, USA). The sequencing was performed on PacBio® Sequel I system (PacBio CLR) with five Single Molecule Real Time (SMRT®) cells using P6-C4 chemistry.

Initial assembly was performed using FALCON package (Chin et al., 2016). The resulting primary assembly was error-corrected with short reads paired-end data following Kang et al. (2020) approach. Briefly, after mapping the short reads data onto the assembly using the BWA-MEM algorithm of Burrows-Wheeler Aligner (BWA) ver. 0.7.12 (Li and Durbin, 2010), the HaplotypeCaller and FastaAlternateReferenceMaker modules from the Genome Analysis Toolkit (GATK) ver. 3.54 (McKenna et al., 2010) were employed to refine the base accuracy with default parameters.

Hi-C Sequencing and Scaffolding

In order to scaffold the long reads-based assembly into pseudomolecules, one Hi-C proximity ligation library using the Dovetail™'s Hi-C Kit v.1.0 has been prepared. The restriction enzyme DpnII has been employed. The prepared library was shipped to Dovetails Genomics (Scotts Valley, California, USA) for sequencing and scaffolding via Dovetail's HiRise™ scaffolding pipeline (Putnam et al., 2016).

Genome Assembly Quality Assessment

The reference-free assembly quality metric LTR Assembly Index (LAI) have been calculated using LTR Retriever v.2.9.0 (Ou et al., 2018). Besides, genome assembly completeness was evaluated with BUSCO v.5.2.2-0 (Simão et al., 2015) with Embryophyta odb10 data set (accessed 15th September, 2021). Assembly contiguity was checked with QUAST-LG v.5.2.2-0 (Mikheenko et al., 2018). In addition, MUMmer v4.0.0 package (Marçais et al., 2018) has been employed to check the genome structure concordance of the assembly compared with the reference Zhongzhi13.

Chromosome Assignment, Structural and Functionnal Genome Annotation

We assigned the chromosome to the generated chromosomes by aligning the Hi-C assembly onto Zhongzhi13 with nucmer package of MUMmer v4.0.0 (Marçais et al., 2018) The repeats analysis was performed using RepeatModeler v. 1.0.8[4] and RepeatMasker v. 4.0.5[5] tools.

Non-coding RNA including tRNA, rRNA, miRNA, and snRNA were identified using Infernal v1.1.4 (Nawrocki et al., 2009) with the Rfam database (Kalvari et al., 2021) (accessed the 24th, December 2021). A cross-validation of tRNA and rRNA was conducted with tRNAscan-SE v2.0.8 (Lowe and Eddy, 1997) and Barrnap v0.9 (Seeman and Booth, 2013) respectively.

[4] http://www.repeatmasker.org/RepeatModeler/

[5] http://www.repeatmasker.org

Ab initio, evidence and homology based annotation pipeline is illustrated in the Figure 11. Briefly, a genome-guided transcriptome assembly was performed by mapping trimmed RNA-seq data onto the softmasked genome assembly using TopHat2 (Kim et al., 2013). The mapped reads were employed to construct Goenbaek transcripts with StringTie (Pertea et al., 2015). The resulting transcripts assembly was used to train BRAKER v1.11 (Hoff et al., 2016) annotation pipeline which makes use of AUGUSTUS (Stanke et al., 2008) and GeneMark-ET (Lomsadze et al., 2014) for gene prediction.

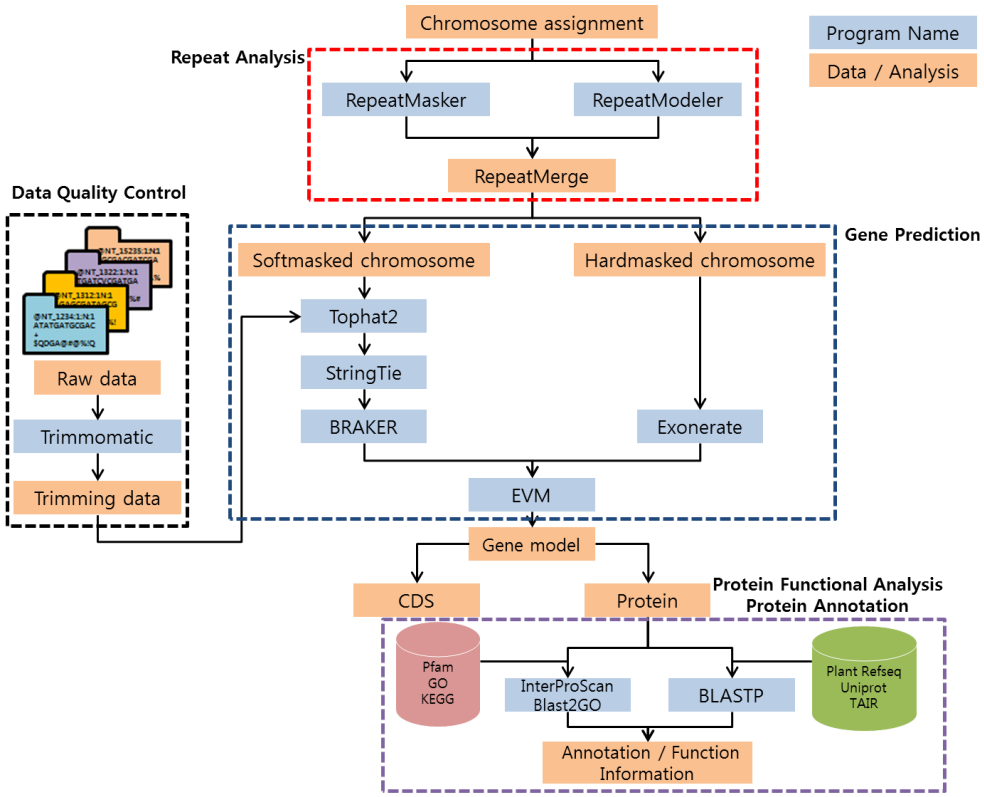

Figure 11. Diagrammatic of the genome annotation of *Sesamum indicum* var. Goenbaek.

In parallel, Exonerate v2.2.0 (Slater and Birney, 2005) served to refine coding gene structure and predict near exact intron/exon boundaries using the hard-masked genome assembly. Both evidence-and *ab initio* annotations were integrated into EVidenceModeler (EVM) (Haas et al., 2008) to generate the gene models.

Functional annotation was executed by doing BLASTp (E-value: 1e-5) to Plant Refseq, Uniprot, TAIR and the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein databases. Proteins domain were inferred by searching against Pfam, GO and KEGG databases using InterProScan ver. 5.34-73.0 (Jones et al., 2014) and Blast2GO Command Line ver. 1.4.1 (Götz et al., 2008). The resulting annotation was manually curated based on the evidence data sets and by comparing with the reference genome Zhonghi13.

Comparative Genomics and Phylogeny Construction

In order to find out unique genes specific to Goenbaek, we retrieved from Oil Crops Research Institute (OCRI) pan-genome repository[6] (accessed the 15th, November, 2018) the protein sequences of *S. indicum* var. Zhongzhi13, S. indicum var. Yuzhi11, *S. indicum* cv. Baizhima, *S. indicum* cv. Mishuozhima, and *S. indicum* var. Swetha. By combining with the newly generated Goenbaek proteins file to the above-mentioned data set, we inferred the genes clustering using OrthoFinder v.2.3.12 (Emms and Kelly, 2019). Afterwards, we took advantage of STAG (Emms and Kelly, 2018) algorithm and STRIDE (Emms and Kelly,

[6] http://www.sesame-bioinfo.org/pan-genome

2017) implemented in OrthoFinder v.2.3.12 (Emms and Kelly, 2019) to infer and rooted the phylogenetic tree respectively.

Two datasets have been used for orthology and phylogenetic analyses. The first data set encompassed only Sesamum species while the second one included close relative species *i.e.* *Perilla frutescens* var. frutescens (PF40) (GenBank: QFCC00000000.2), *Perilla citriodora* PC002 (GenBank: QIYW00000000.2) and PC099 (GenBank: SDAM00000000.2) genotypes (Zhang et al., 2021), *Mimulus guttatus* (GenBank: APLE00000000.1) (Hellsten et al., 2013), *Solanum lycopersicum* (GenBank: AEKE00000000.3) (Sato et al., 2012) and *Arabidopsis thaliana* (GenBank: GCA_000001735.2) (Lamesch et al., 2012) as outgroup. Using Blast2Go v.5.2.5 GUI version (Götz et al., 2008), we predicted functional attribute of Goenbaek-specific genes in order to figure out the key variety-specific biological functions.

Assembly-Based Structural Variants Calling

To find out structural variants, Goenbaek were aligned onto the reference genome Zhongzhi13 using nucmer module from MUMmer package v4.0.0 (Marçais et al., 2018) with the following setting: -maxmatch -l 100 -c 500. Afterwards, the structural variants including insertion, deletion, repeat expansion, repeat contraction, tandem expansion, and tandem contraction were detected using Assemblytics web-server (Nattestad and Schatz, 2016).

Identification of Late Embryogenesis Abundant (LEA) Genes

Candidate LEA genes representing one of the protective gene repertoire against desiccation or osmotic stresses (Hundertmark and Hincha, 2008) were investigated in Goenbaek genome using hmmsearch with the following setting -E 1e-5 --domE 1e-5. The PFAM domain of the

different classes of LEA genes including Dehydrin (PF00257.21), LEA1 (PF03760.17), LEA2 (PF03168.15), LEA3 (PF03242.15), LEA4 (PF02987.18), LEA5 (PF00477.19), LEA6 (PF10714.11), and SMP (PF04927.14) were retrieved from Pfam website (http://pfam.xfam.org/). Candidate LEA genes were screened for LEA domain search using PfamScan command line tool v1.6 (Madeira et al., 2019). The conserved domain of each gene was double-checked with the help of Interproscan web-server[7].

The physico-chemical properties and sub-cellular localization were assessed with ProtParam[8] (Gasteiger et al., 2005) and WoLF PSORT[9] web-servers (Horton et al., 2007). To assess the phylogenetic relationships of sesame LEA with well-known LEA genes from *Oryza sativa* (Wu et al., 2019b) and *A. thaliana* (Hundertmark and Hincha, 2008), a maximum likelihood phylogenetic tree inference was performed using IQ-TREE v1.6.12 (Nguyen et al., 2015a) with 1,000 bootstrap iterations after multiple sequence alignment and poorly aligned regions trimming steps handled by MAFFT v.7.464-0 (Katoh and Standley, 2013) and trimAl v.1.4.1 (Capella-Gutiérrez et al., 2009) respectively. Besides, the expression count of LEA genes in Goenbaek organs was assessed using HISAT2 v2.2.1 (Kim et al., 2015a) for mapping the RNA-seq data, and FeatureCounts module from Subread v. 2.0.1 package (Liao et al., 2013) for gene abundance count. The expression pattern within the organs was rendered using pheatmap v.1.0.12 (Kolde, 2019) R package following Ward's clustering method (clustering_method = "ward.D") and genes-centered scaling (scale = "row").

[7] https://www.ebi.ac.uk/interpro/search/sequence/

[8] http://web.expasy.org/protparam/

[9] https://wolfpsort.hgc.jp/

Results and Discussion

High-Quality Chromosomal-Level Genome Assembly

The result of the genome assembly presented in the Figure 12A showed a set of 13 chromosome sets; which is consistent with the previous karyotyping of the Chinese cultivar Yuzhi11 (Zhang et al., 2013c) and the produced Hi-C contact map (Figure 12B).

From the long-reads sequencing, a total of 29.11 Gbp (90.4-fold) was generated. Initial assembly resulted in 433 contigs spanning 282 Mbp with an N50 of 2.5 Mbp. The contigs anchoring with chromosome conformation capture Hi-C data drastically improved the contiguity of the draft assembly by 7.6 fold. A total of 249 sequences have been reached with high signal contact map. The identified 13 chromosomes set size was 282 Mbp with an N50 value of 19.8 Mbp (Table 9).

Table 9. Comparative statistics of 13 chromosomes assemblies of varieties and landraces

	Goenbaek	**Zhongzhi13**	**Yuzhi11**	**Swetha**	**Mishuozhima**	**Baizhima**
Data type	PacBio+Hi-C	I+HDLP*	R-B**	R-B	R-B	R-B
Sequences number	13	13	13	13	13	13
Largest contig (bp)	26,666,084	26,180,356	18,405,807	31,370,383	22,313,205	22,640,701
Total length (bp)	262,484,163	259,726,592	171,572,600	309,289,683	217,618,843	217,824,408
N50 (bp)	19,877,096	20,257,639	12,915,212	24,078,102	17,134,316	16,884,392
L50	6	6	6	6	6	6
N70 (bp)	17,217,227	16,756,707	11,343,251	23,333,770	14,404,523	13,939,448
L70	9	9	9	8	9	9
Gap	183	8,906	16,372	88,929	26,813	24,869
Coding-genes count	23,539	34,451	21,360	37,650	26,737	27,015

*High density linkage map, **Reference-based

By comparing the chromosomal assembly set of Goenbaek and the reference zhongzhi13 (Table 9), a substantial improvement has been completed at gap level. In fact, we counted 183 gaps for Goenbaek for 8,906 gaps for Zhonghi13.

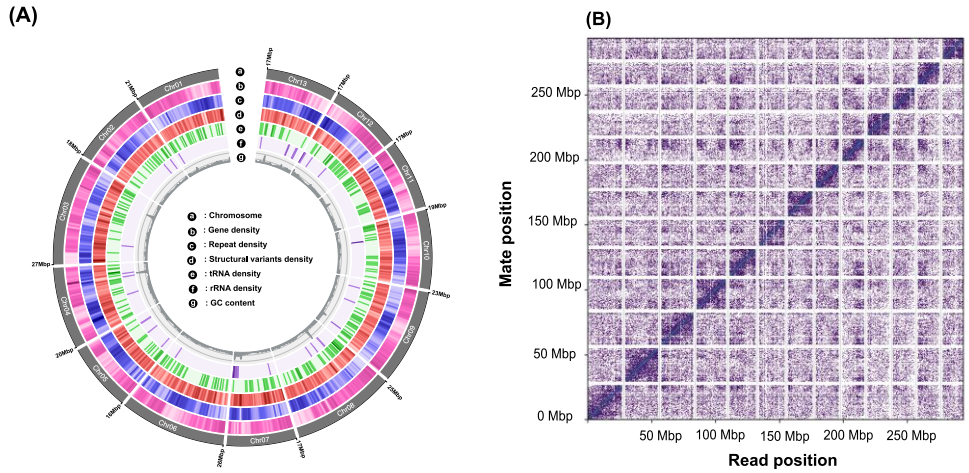

Figure 12. Genome landscape of the Korean *Sesamum indicum* var. cultivar Goenbaek.(A)-Circos plot showing the genome features including from the outermost to innermost tracks, a) Chromosome, b) Genes density, c) Repeats density, d) Structural variants density, e) tRNAs density, f) rRNAs density. Tracks were summed in 500 Kbp windows with 25 Kbp step size. (B)- Dovetail Genomic's HiC linkage density histogram showing the mapping positions of the first (x-axis) and second (y-axis) read in the read pair grouped into bins. The color of each square gives the number of read pairs within that bin. White vertical and black horizontal lines have been added to show the borders between scaffolds. Scaffolds less than 1 Mb are excluded.

Several metrics have been used to check the quality of our assembly including assembly contiguity (N50), assembly completeness (BUSCO), assembly continuity (LTR Assembly Index) and colinearity (whole genome alignment with MUMmer package onto Zhonghzhi13). Although the contiguity is a similar range (Table 9), the BUSCO analysis using embryophyta odb10 dataset revealed a better coverage of single copy orthologues in Goenbaek (92.8%) compared to Zhongzhi13 (24.5%). When considering LAI index (Figure 13), the overall value (17.47) in Goenbaek was higher than Zonghzhi13 one. Similar

observation was noted by comparing each paired-chromosomal region. The whole-genome alignment of Goenbaek onto Zhongzhi13 revealed high similar genome structure (Figure 14).

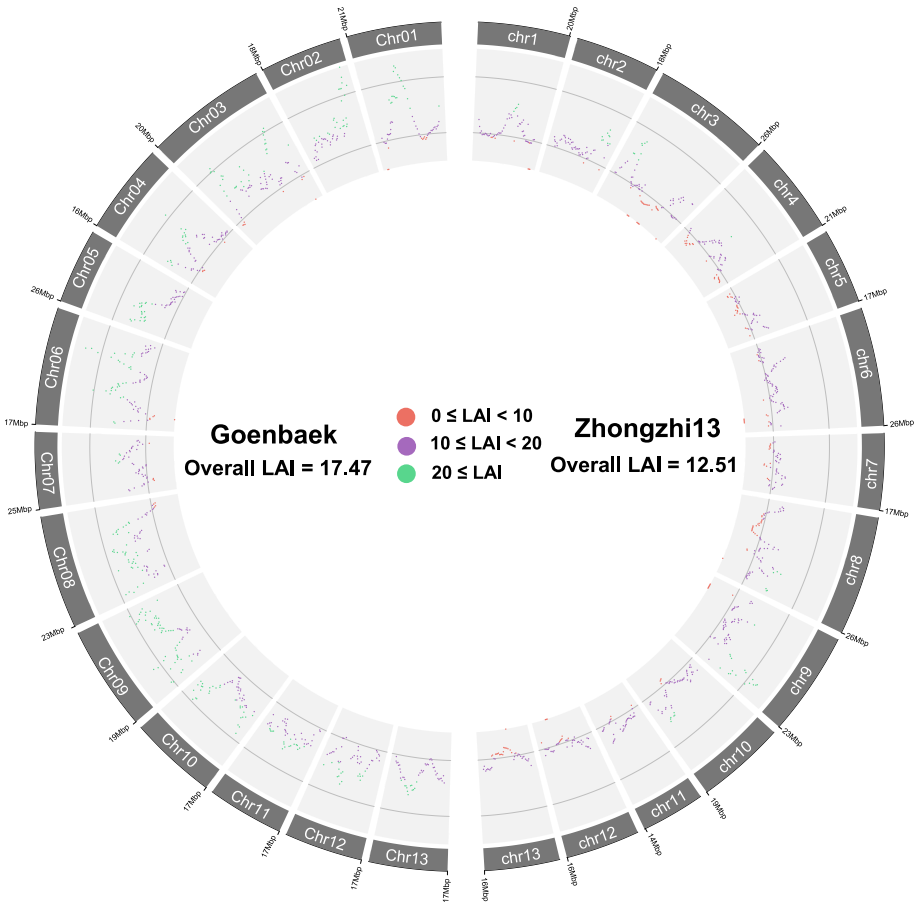

Figure 13. Dot plot showing the LTR Assembly Index (LAI) distribution within each chromosome of Goenbaek and Zhongzhi13. The LAI values are colored as follows: red for LAI values between 0 and 10, purple for LAI values between 10 and 20, green for LAI values superior to 20.

However, the inversions (colored in red in the Figure 14) of some chromosome fragments

have been noted.

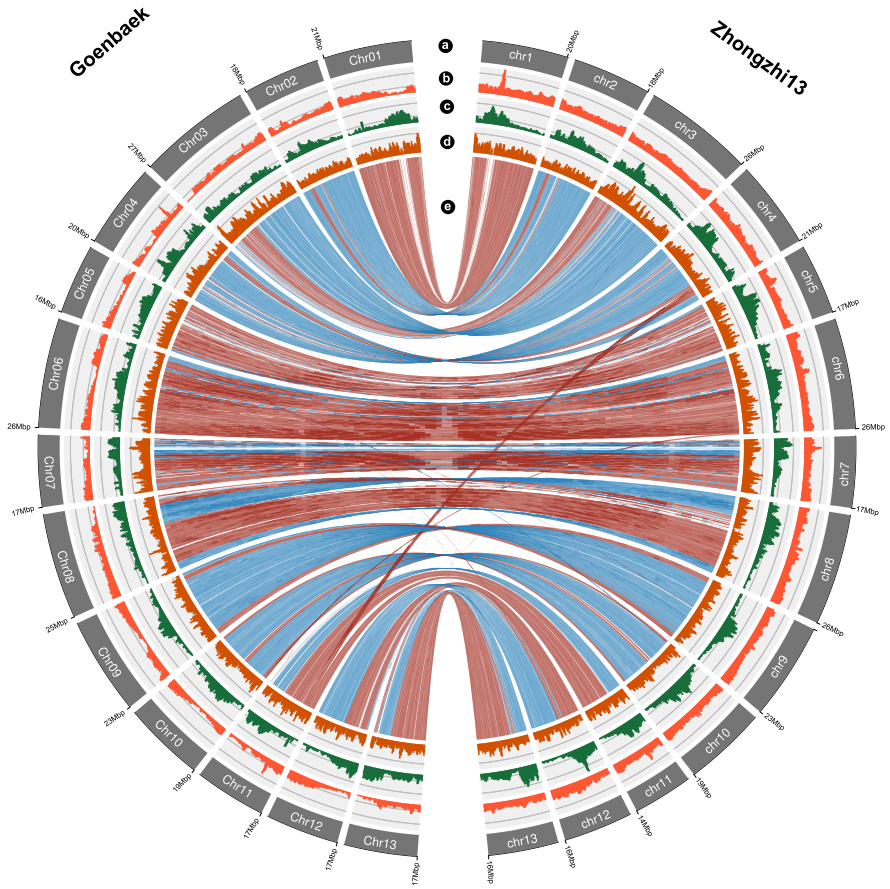

Figure 14. Comparative genome map of Goenbaek and Zhongzhi13 showing chromosome size (a), gene density (b), repeats density (c), (structural variants density (c), whole-genome alignments of Goenbaek onto Zhongzi13 chromosomes with blue ribbons indicating forward-strand alignments and red ribbons indicating reverse-strand alignments (inversions). Tracks were summed in 500 Kbp windows with 25 Kbp step size.

Overall, the quality assessment results imply that Goenbaek exhibits reliable contiguity and completeness; making it useful resource for Korean sesame breeding program.

Genome annotation

Repeat sequences in Goenbaek genome spanned approximately 128.1 Mbp, which represents 48.83% of the assembled genome (Table 10). Long terminal repeat (LTR) retrotransposons are prominent, occupying 20.2% of the genome. Among LTR, Copia and Gypsy types are the most abundant, accounting for 9.1%, and 7.3% respectively. The repeat elements in *S. indicum* var. Goenbaek are higher (48.83%) than *S. indicum* var. Zhongzhi13 (28.5%) (Wang et al., 2014), *S. indicum* var Yuzhi11 (31.65%) (Zhang et al., 2013c) but lower than the close relatives *Perilla frutescens* (64.1%), and *Perilla citriodora* (56.7%) (Zhang et al., 2021).

Table 10. Repeat content information in the sesame variety Goenbaek

Class	Elements	Count	Length (bp)	Percent
DNA	Uncharacterized	813	258,832	0.10%
	CMC-EnSpm	6,716	3,696,805	1.41%
	CMC-Transib	1,060	52,414	0.02%
	Crypton	586	304,925	0.12%
	Dada	316	68,115	0.03%
	En-Spm	2,809	235,867	0.09%
	Harbinger	61	9,347	0.00%
	MULE-MuDR	13,205	3,915,495	1.49%
	Mite	89	4,562	0.00%
	MuDR	377	40,420	0.02%

Table 10. *Continued*

	PIF-Harbinger	1,320	871,778	0.33%
	Sola	1,000	487,100	0.19%
	TcMar	1,544	313,493	0.12%
	TcMar-Pogo	859	291,316	0.11%
	TcMar-Stowaway	275	35,287	0.01%
	hAT	360	89,733	0.03%
	hAT-Ac	4,696	1,346,792	0.51%
	hAT-Charlie	1,984	385,492	0.15%
	hAT-Tag1	2,366	576,646	0.22%
	hAT-Tip100	54	14,449	0.01%
Total DNA		**40,490**	**12,998,868**	**4.95%**
LINE	Uncharacterized	722	116,263	0.04%
	Jockey	42	17,361	0.01%
	L1	6,135	6,155,849	2.35%
	R1	214	53,257	0.02%
	R2	420	233,710	0.09%
Total LINE		**7,533**	**6,576,440**	**2.51%**
LTR	Uncharacterized	26,562	6,731,082	2.56%
	Caulimovirus	2,245	2,895,561	1.10%
	Copia	30,561	24,013,267	9.15%
	Gypsy	29,582	19,223,979	7.32%
	Ngaro	1,083	118,297	0.05%
Total LTR		**90,033**	**52,982,186**	**20.18%**
SINE	Uncharacterized	9	657	0.00%
	Alu	61	20,325	0.01%

Table 10. *Continued*

Total SINE	**70**	**20,982**	**0.01%**
MobileElement	535	101,176	0.04%
Other	364	31,414	0.01%
Helitron	353	67,806	0.03%
Retroelement	140	28,681	0.01%
Unknown	194,737	55,237,760	21.04%
nonLTR	112	14,570	0.01%
Satellite	4	483	0.00%
Simple_repeat	479	115,181	0.04%
Total	**334,850**	**128,175,547**	**48.83%**
Total non-redundant		**118,036,267**	**44.97%**

The inference of non-coding RNA returned 841 tRNA, 416 rRNA, 123 miRNA, and 134 snRNA, spanning 62.55 Kbp, 311.85 Kbp, 16.79 Kbp, and 13.65 Kbp respectively (Table 11).

Table 11. Predicted non-coding RNA statistics

Type	Number	Total length (bp)	Percentage in the genome (%)
tRNA	841	62,549	0.024%
rRNA	416	311,851	0.119%
miRNA	123	16,785	0.006%
snRNA	134	13,645	0.005%

For evidence based annotation, a total of 37.3 Gb transcriptome data from leaf, root, stem, apical meristem, and seed were generated. The gene prediction resulted in 23,539 protein-

coding genes with an average gene, exon, and intron length of 3,155 bp, 245 bp, and 472 bp respectively. Compared to the Zhongzhi13 (34,451 coding-genes), the genes count was lower. However, it is comparable with Yuzhi11 and Baizhima gene sets which exhibited 21,360 and 27,015 genes respectively (Table 9).

From functional annotation analysis, 91.1% of Goenbaek genes shared abundant sequence similarity with Plant RefSeq sequences; while 73.9% and 84.2% of sequence similarity was found for UniProt and *A. thaliana* databases respectively (Table 12).

Table 12. Functional annotation information from the predicted protein coding-genes of Goenbaek

Annotation category		**Hit count**	**Percent(%)**
BLASTP	Plant RefSeq	21,438	91.1%
	UniProt	17,389	73.9%
	Arabidopsis thaliana	19,823	84.2%
InterProScan	GO	10,968	46.6%
	KEGG	1,061	4.5%
	Pfam	17,119	72.7%
Transcriptome	mRNA-Seq read count	16,368	69.5%
Total hit		21,875	92.9%
Known		19,708	83.7%
Uncharacterized		2,027	8.6%
Hypothetical		1,804	7.7%
Total		23,539	100.0%

Domain search revealed that 72.7%, 46.6%, and 4.5% of the predicted genes have known domains when using Pfam, GO and KEGG databases respectively. Altogether, 83.7% of the predicted Goenbaek genes were functionally annotated based on the known protein sets while uncharacterized and hypothetical genes accounted for 8.6% and 7.7% respectively.

Species-Specific Genes Repertoire and Phylogenetic Placement of Goenbaek

Predicted proteins from Goenbaek and available cultivars and landraces protein sets were compared to find out Goenbaek-specific gene clusters (Figure 15). Additional protein dataset from Lamiales species including Perilla, Mimulus and Solanum genera were added with *A. thaliana* as outgroup in order to infer the phylogenetic position of Goenbaek (Figure 16).

The gene clustering analysis revealed a total of 9,919 gene clusters shared by all *S. indicum* species while 1,171 genes clusters were unique to Goenbaek (Figure 15). The KEGG analysis revealed that Goenbaek unique gene clusters were enriched some biological functions including alpha-linoleic acid metabolism (ko00592), fatty acid biosynthesis (ko0061), penose and glucuronate interconversions (ko00040), glycerolipid metabolism (ko00561), galactose metabolism (ko00052), and phenylpropanoid biosynthesis (ko00940).

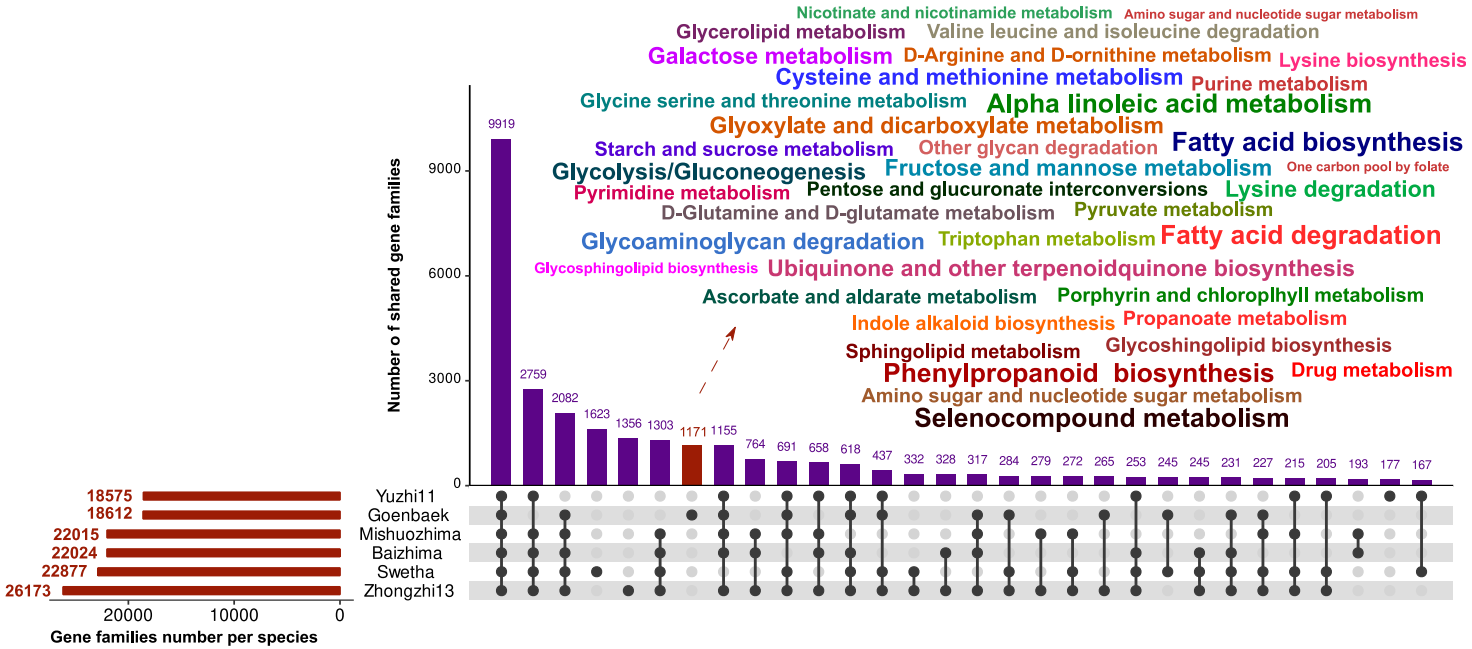

Figure 15. Gene conservation status in sesame pangenome showed by an UpSet plot depicting the number of shared gene families. The red vertical bar highlights Goenbaek-specific gene families with the enriched functional attributes.

The abundant presence of lipid metabolism related functions suggests that Goenbaek have undergone important seed quality-oriented selection. In fact, Goenbaek exhibited high oil and lignan contents of Goenbaek in Korea (Kim et al., 2018). Similar biological oil-related functions were also predicted to be unique to the other cultivars Zhongzhi13, Yuzhi11 and Swetha while landraces showed environmental adaptive gene repertoire including plant-pathogen interaction, protein folding, signal transduction, protein processing in endoplasmic reticulum, and PI3K-Akt signaling (Yu et al., 2019a).

Furthermore, Goenbaek unique genes also encompassed health promoting related proteins active in Drug metabolism (ko00983), selenocompound metabolism (ko00450), sphingolipid metabolism (ko00600), glycosphingolipid biosynthesis–ganglio series (ko00604), indicating the presence of natural health-promoting genes. It is worth mentioning that the variety Goenbaek exhibited high content in the specialized metabolites sesamin and sesamolin (Kim et al., 2018). Sesame lignans (sesamin, sesamol, sesamolin, and sesaminol) are well known as having anti-inflammatory anti-microbial, antioxidant, and anti-cancer properties (Mahendra Kumar and Singh, 2015; Majdalawieh and Mansour, 2019; Wu et al., 2019b). Moreover, sphingolipid metabolism and selenocompounds biosynthesis have been also identified as anti-cancer (Ogretmen, 2018) and antioxidant (Tapiero et al., 2003) agents respectively.

The inferred phylogenetic tree placed Goenbaek in Sesamum clade but genetically distant from the other *S. indicum* species (Figure 16). The tree topology was consistent with the results from Yu et al., (2019a) that showed that Zhongzhi13 and Yuzhi11 formed a sub-clade (Chinese cultivars) while the Chinese landraces (Baizhima and Mishuozima) also clustered together. The Indian and Korean cultivars are distinct from its congeners, implying the geographical and/or environmental effect on genetic diversity.

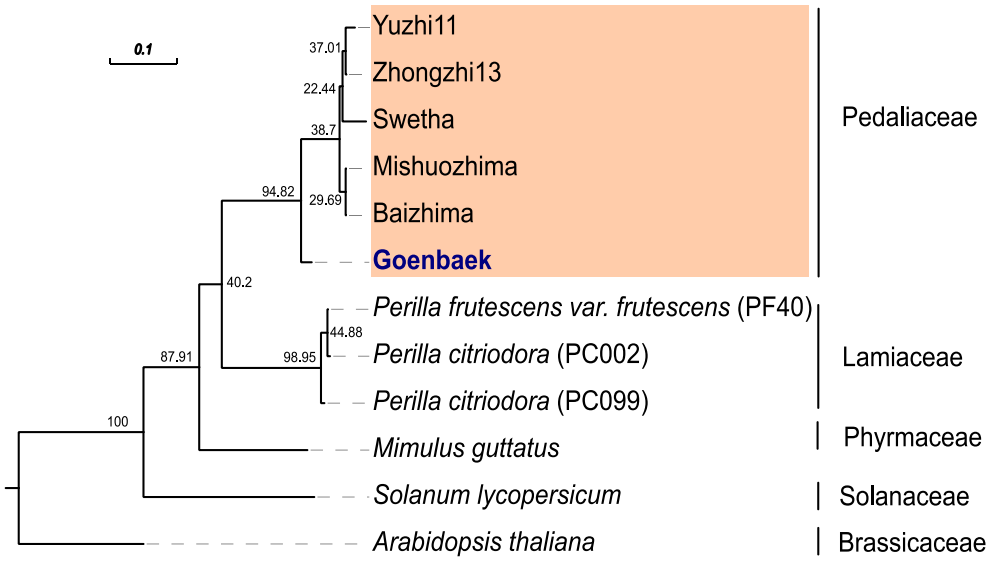

Figure 16. Phylogenetic tree inferred from *S. indicum* species and Lamiales close relative species including *P. frutescens* var. *frutescens*, *P. citriodora*, *M. guttatus* and *S. lycopersicum*. *A. thaliana* was set as outgroup.

The observed placement of Goenbaek might be linked to its species-specific genes since the number of genes in species-specific orthogroups is 3.5 times higher when compared to Zhongzhi13. This rate is drastically superior when compared to Yuzhi11 (47.5 times), Baizhima (21.75 times), Mishuozhima (37.3 times), Swetha (1.23 time); confirming the phylogenetic topology.

Genome-wide Identification and Phylogenetic Analysis of SiLEA Genes

LEA genes have been functionally correlated to response in abiotic stress by acting as molecular shield (Furuki et al., 2011), hydradation buffer (Hundertmark et al., 2012), interacting with cell membranes for signaling (Olvera-Carrillo et al., 2011), as enzyme protection (Hand et al., 2011), or regulating phytohormones (Lim et al., 2018).

A total of 71 LEA genes have been identified and classified into six classes including LEA1, LEA2, LEA3, LEA4, LEA5, SMP, and Dehydrin (Figure 17A). The most abundant class was LEA2, which encompassed 56 proteins, while the other classes, SMP, LEA1, LEA4, LEA3 and LEA5 contained 5, 4, 4, 1, and 1 member(s) respectively. The different classes were consistently confirmed via phylogenetic inference using not only sesame LEA genes (Figure 17A) but also, well-characterized LEA genes from *Oryza sativa* and *A. thaliana* (Figure 17B).

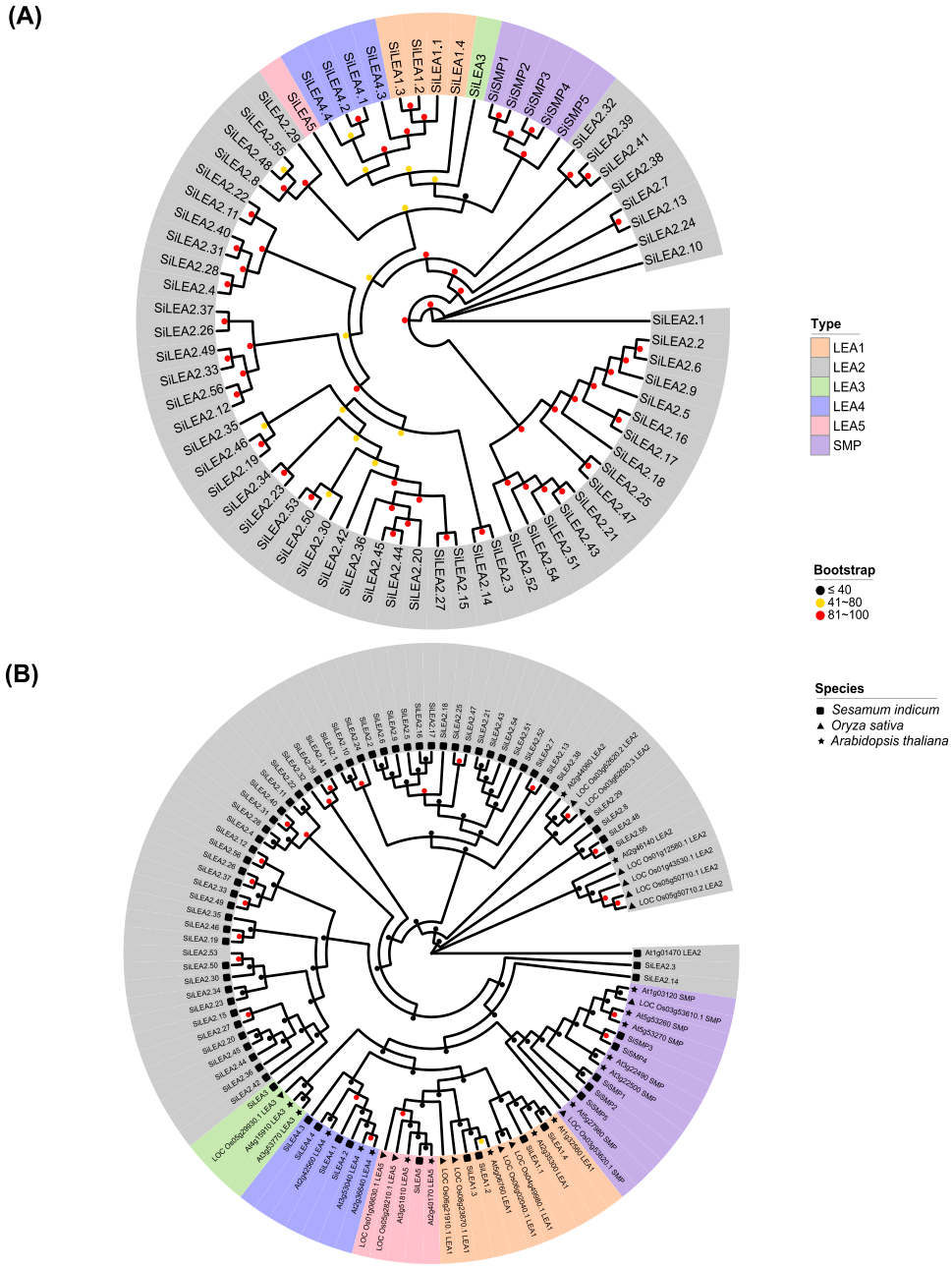

Figure 17. Phylogenetic trees of Late embryogenesis abundant (LEA) genes. (A)-Unrooted maximum likelihood phylogenetic tree of *Sesamum indicum* var. Goenbaek. (B)-Unrooted maximum likelihood

of LEA genes from *S. indicum* var. Goenbaek, *Oryza sativa* and *Arabidopsis thaliana*. Black (<40), yellow (41-80), and red (81-100) dot represents the clades support values in the phylogenetic trees.

Expression profile of SiLEA genes in Goenbaek

The expression profiles of different tissues (leaf, stem, seed, apical meristem, and root) (Figure 18) revealed that SiLEA2 genes class is present within all organs while SiLEA1 genes are exclusively highly expressed in seed. Most of SiLEA4 genes were also expressed in seed only except SiLEA4 which is also expressed in the apical meristem. The unique SiLEA5 was also preferentially expressed in seed while the SiSMP genes were present not only in seed but also in apical meristem, leaf and stem. The relative presence of LEA within all organs suggests their contribution to the plant growth and development. The specific prominence of SiLEA2 genes in all organs implies that Goenbaek have a rich LEA2 repertoire that can play active role in abiotic stress context including drought stress. Specifically, the relative high LEA2 genes expression in root suggests its functional role in root development (Magwanga et al., 2018).

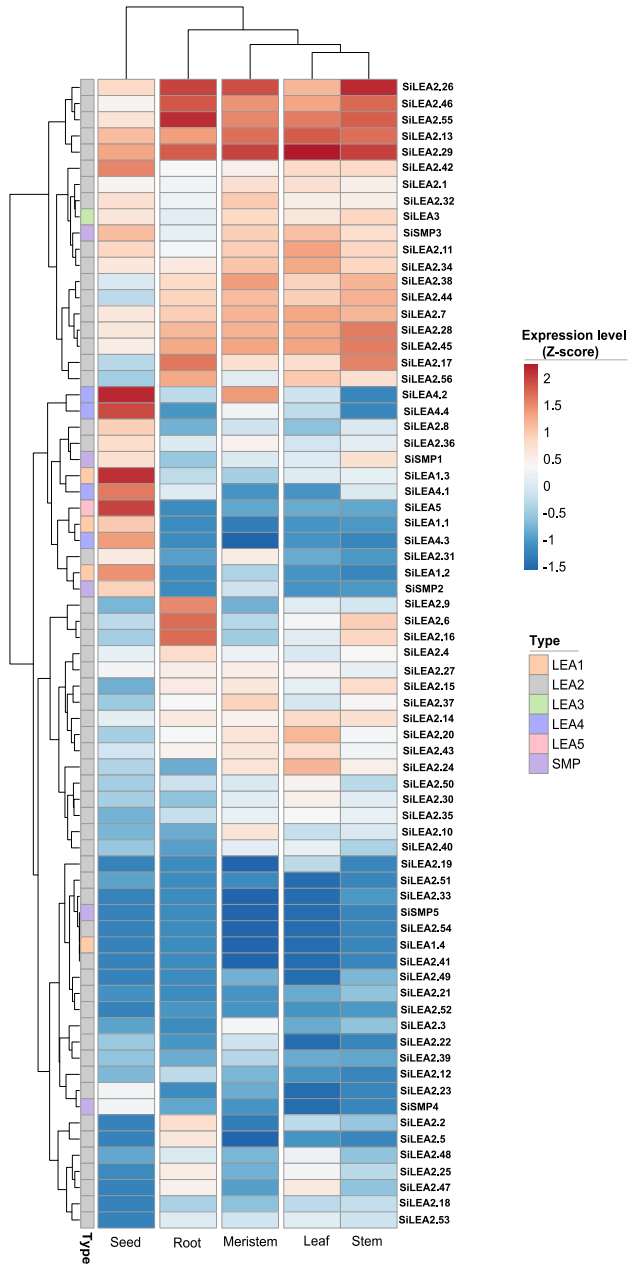

Figure 18. Heatmap depicting the expression of LEA genes in *S. indicum* var. Goenbaek leaf, root, stem, apical meristem, and root. Different color of arcs represents different types of LEA genes.

As initial part of the Korean sesame breeding program, we generated a high-quality chromosomal grade genome of the elite cultivar Goenbaek with a total of 23,539 protein-coding genes. Comparative genomics highlighted a strong selection pressure from human regarding seed quality as source for oil and lignan production. The phylogenetic inference exhibited that Goenbaek is genetically distinct from its congeners, implying geographical or adaptive-driven genetic diversity. Moreover, we investigated LEA gene repertoire of Goenbaek, providing initial molecular resources for abiotic stress management. Given the economic importance of sesame oil as well as its health promoting properties, the present genomic resource offers a path for sesame breeding regarding not only routinely targeted agronomic traits, but also food and health promoting attributes.

CHAPTER 3: Characterization of Peroxidase and Laccase Gene Families and *in silico* Identification of Potential Genes involved in Upstream Steps of Lignan Formation in Sesame

Summary

Peroxidases and laccases are oxidative enzymes involved in physiological processes in plants, covering responses to biotic and abiotic stress as well as biosynthesis of health-promoting specialized metabolites. Although they are thought to be involved in the biosynthesis of (+)-pinoresinol, a comprehensive investigation of this class of enzymes has not yet been conducted in the emerging oil crop sesame and no information is available regarding the potential (+)-pinoresinol synthase genes in this crop. In the present study, we conducted a pan-genome-wide identification of peroxidase and laccase genes coupled with transcriptome profiling of diverse sesame varieties. A total of 83 and 48 genes have been identified as coding for sesame peroxidase and laccase enzymes, respectively. Based on their protein domain and *Arabidopsis thaliana* genes used as bait, the genes were classified into nine and seven groups of peroxidase and laccase genes, respectively. The expression of the identified genes was evaluated using dynamic transcriptome sequencing data from six sesame varieties, including the elite cultivars vs landraces, white vs black seed varieties, high vs low oil content varieties. Two peroxidases (*SiPOD52* and *SiPOD63*) and two laccases (*SiLAC1* and *SiLAC39*), well conserved within the sesame pan-genome, exhibiting a consistent expression patterns within the sesame varieties and matching the biosynthesis kinetic of (+)-pinoresinol in the seeds, have been identified as potential (+)-pinoresinol synthase genes. The findings from this study pave the way for lignans-oriented sesame bio-engineering with wide applications in food, health and medicinal domains.

Introduction

Sesame (*Sesamum indicum* L.), a member of Pedaliaceae family, is an oil crop whose seed contain lignans including (+)- sesamin, (+)- sesamolin, and (+)- sesaminol (Murata et al., 2017). Therapeutical properties of sesame lignans against neurodegenerative diseases (Katayama et al., 2016; Shimoyoshi et al., 2019), prostate, and breast cancers (Liu et al., 2006) have been reported. Besides, lignans represent an emerging perspective for health care and disease prevention as functional foods and nutraceuticals (Peterson et al., 2010; Peñalvo and López-Romero, 2012; Zamora-Ros et al., 2012; Durazzo et al., 2014; Sun et al., 2014; Rodríguez-García et al., 2019). The lignan market is exploding and may reach over USD 610 Million by 2028 (Anonymous, 2021) with multiple applications covering food, pharmaceutics, and cosmetics industries. Meanwhile, several patents have been deposited dealing with the extraction, purification, and transformation of lignans from sesame (Hardwicke et al., 1959; Markus, 1962; Forse and Chavali, 2001a; Namiki et al., 2001; You et al., 2011b; Sok et al., 2012; Chami et al., 2013; Yamada et al., 2014; Kojima et al., 2017); showing the growing interest for this class of plant specialized metabolites.

So far, tremendous works have been done to elucidate the lignans biosynthesis pathway in sesame by using wild (*S. alatum* and *S. radiatum*) and cultivated (*S. indicum*) materials (Ono et al., 2006, 2019; Noguchi et al., 2008; Murata et al., 2017; Harada et al., 2020). A total of six enzymes involved in lignan biosynthesis in sesame have been characterized, including two cytochrome P450 coding genes (*CYP81Q1* and *CYP92B14*), four glycosyltransferases (*UGT71A9*, *UGT94D1*, *UGT94AG1*, and *UGT94AA2*) (Murata et al., 2017; Ono et al., 2019) (Figure 19). The connections between the identified enzymes and their respective targets have been depicted in the Figure 19. At the initial step of the lignan biosynthesis pathway, an

oxidative coupling reaction involving two molecules of conyferol alcohol takes place. With the help of a dirigent protein (*DIR*), the primary precursor of ligan, (+)-pinoresinol, is generated. The latter is then sequentially catalyzed by *CYP81Q1* to produce (+)-piperitol and (+)-sesamin, respectively. Further, (+)-sesamolin and (+)-sesaminol synthesis is guided by *CYP92B14*.

It is worth mentioning that the mechanism for the synthesis of the central precursor *ie* (+)-pinoresinol, possibly involves a combinatory action of oxidase enzymes (laccases and/or peroxidases) with dirigent protein (Kim et al., 2015b). Oxidation by peroxidases and/or laccases followed by a stereo-selective radical coupling guided by a dirigent protein is assumed to be determinant at early stage of the lignan biosynthesis (Davin et al., 1997; Kim et al., 2002b, 2015b; Pickel et al., 2010). In brief, one electron oxidation on conyferyl alcohol is catalyzed by peroxides and/or laccasses. As a result, an intermediate molecule, a conyferyl alcohol (1)-derived free radical, is formed. Then an intramolecular cyclization guided by a dirigent protein induces the (+)-pinoresinol molecule (Kim et al., 2015b).

To the best of our knowledge, evidence of oxidases activity involved at the critical precursor step of sesame lignan biosynthesis has not been yet established. While a dirigent protein (*XP_011080883*) has been detected (Andargie et al., 2021) in sesame via by sequence homology approach, the fundamental information relative to peroxidases and laccases remain unexplored.

Peroxidases and laccases are multifunctional enzymes playing a wide panel of roles in plant covering biotic (Bindschedler et al., 2006; Daudi et al., 2012; Hu et al., 2018) and abiotic responses (Cai et al., 2006; Roy et al., 2017; Kidwai et al., 2020; Li et al., 2021; Aleem et al., 2022), and other biological processes such as fiber initiation (Hu et al., 2020); cell elongation (Jackson and Ricardo, 1998), lignification (Gabaldón et al., 2005; Liang et al.,

2006), seed setting and panicle branching (Zhang et al., 2013d), pigmentation (Ring et al., 2013; Fang et al., 2015), and flavonoid oxidation (Pourcel et al., 2007). They are also useful for therapeutic and industrial applications. For instance, peroxidases are used in industrial enzymatic reactions, diagnostic tests, and enzyme immunoassays (Yoshida et al., 2003). Meanwhile, laccase has a great importance in paper industry due to its capacity of delignification (Poppius-Levlin et al., 2001; Hussain et al., 2022). Besides, laccases are also useful in ethanol production, wine clarification, industrial effluents treatment, herbicide degradation, dyes decoloration, and drug analysis (Mayer and Staples, 2002).

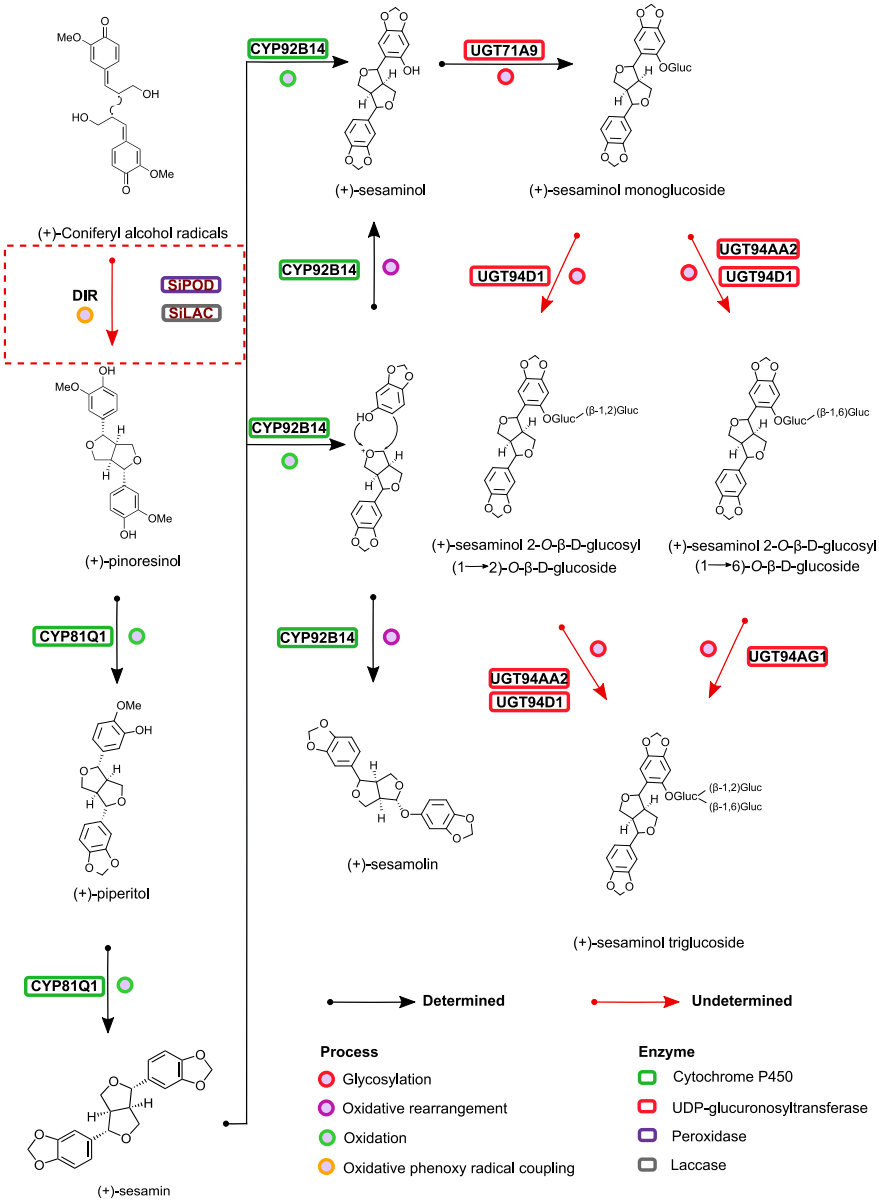

Figure 19. A simplified lignans' biosynthesis pathway in sesame showing the route of synthesis of (+)- sesamin, (+)- sesamolin, and (+)- sesaminol. *CYP81Q1* and *CYP92B14* triggered the biosynthesis of (+)- sesamin, (+)- sesamolin and (+)- sesaminol, while *UTG71A9*, *UGT94D1*, *UGT94AA2*, and

UGT94AG1 is suggested to catalyze the synthesis of (+)- sesaminol momoglucoside, (+)-sesaminol 2-O-β-D-glucosyl (1 2)-O-β-D-glucoside, (+)-sesaminol 2-O-β-D-glucosyl (1 6)-*O*-β-D-glucoside, and (+)-sesaminol triglucoside. The target step in the present study is marked by dashed red rectangular box. The pathway is adapted from Ono et al. (2006).

Owing to the importance of the peroxidases and laccases in plants, the present study was carried out to comprehensively characterize them in sesame, and outline candidate genes likely involved in at the early steps of sesame lignans biosynthesis.

Materials and Methods

Genome-wide identification of Peroxidase and Laccase genes and core genes inference

Peroxidase and laccase genes were screened out using the genome data from *S. indicum* var. Zhongzhi13 (NCBI RefSeq accession: GCF_000512975.1), *S. indicum* var. Goenbaek (https://zenodo.org/record/6350881, accessed on 15 March 2022), *S. indicum* cv Mishouzhima, *S. indicum* cv. Baizhima, *S. indicum* var. Yuzhi11, and *S. indicum* var. Swetha (http://www.sesame-bioinfo.org/pan-genome, accessed on 15 November 2018) respectively. NCBI HMM accession TIGR03390.1 (EC 1.10.3.2) served for the identification of candidate laccase genes, whereas PFAM HMM accession PF00141.26 was employed to detect peroxidase genes. After a hit search using hmmsearch (-E 1e-5 --domE 1e-5), a domain verification was executed with PfamScan v1.6 (Mistry et al., 2007) to check the presence of the POD and LAC respective domains. Spurious genes were filtered out. An additional check of the presence of the domains was performed using the InterProScan v5 (Jones et al., 2014).

In order to infer the core conserved POD and LAC genes within the sesame pangenome dataset, Orthofinder v2.3.12 (Emms and Kelly, 2019) was run with the default settings.
The identified POD and LAC genes from the reference genome of *S. indicum* var Zhongzhi13 were retained for downstream analysis.

Chromosome location and synteny analyses

Genome mapping of SiPOD and SiLAC was rendered using MG2C V2.1 (Chao et al., 2021) based on the annotation information. Sesame-to-sesame and sesame-to-Arabidopsis synteny blocks evaluation were investigated with the MCScanX toolkit (Wang et al., 2012b). Regarding the evolutionary origin of the duplicated genes, the duplicate genes classifier perl script from MCScanX, helped to distinguish between genes evolved by tandem or segmental duplication.

Phylogenetic tree construction

Peroxidase and laccase genes from *A. thaliana* were added to SiPOD and SiLAC for the construction of the phylogenetic tree. Prior to the tree inference, the genes were aligned using MAFFT v7.464-0 (Katoh and Standley, 2013). The resulting alignment was trimmed with trimAl v1.4.1 (Capella-Gutiérrez et al., 2009). Subsequently, the trees were constructed with IQ-TREE v1.6.12 (Nguyen et al., 2015b). Peroxidase and laccase maximum likelihood trees were inferred following models LG+R5 and LG+I+G4 with 1,000 iterations, respectively. The tree model detection was estimated using the ModelFinder package (Kalyaanamoorthy et al., 2017).

RNA-Seq Data Retrieval

A set of six sesame varieties (Zhongzhi13, Zhongfengzhi No.1, Zhongzhi No.33, ZZM4728, ZZM2161, and ZZM3495) were selected for the peroxidase and laccase expression investigation using RNA sequencing data (Table 13).

Table 13. Information relative to the SRA accessions of RNA sequencing data used in the present study

Variety name	Key characteristic	NCBI Project	NCBI SRA accession number	Reference
Zhongzhi13	High oil content (59g/100g seed)-Reference genome	PRJNA186669	SRR1055252,SRR1055255,SRR1055253,SRR1055260,SRR1055262,SRR1055261,SRR1055254,SRR1055257,SRR1055259,SRR1055256,SRR1055258,SRR1055263	Wang et al. (2014)
Zhongfengzhi No1	White seed	PRJNA679682	SRR13089686,SRR13089687,SRR13089688,SRR13089689,SRR13089690,SRR13089691,SRR13089692,SRR13089693,SRR13089694,SRR13089695,SRR13089696,SRR13089697,SRR13089698,SRR13089699,SRR13089700,SRR13089701,SRR13089702,SRR13089703,SRR13089704,SRR13089705,SRR13089706,SRR13089707,SRR13089708,SRR13089709,SRR13089710,SRR13089711,SRR13089712,SRR13089713,SRR13089714,SRR13089715,SRR13089716,SRR13089717,SRR13089718,SRR13089719,SRR13089720,SRR13089721	Wang et al. (2020)
Zhongzhi No33	Black seed			
ZZM4728	High oil content (59g/100g seed)	PRJNA400575	SRR6010084,SRR6010085,SRR6010086,SRR6010087,SRR6010088,SRR6010089,SRR6010090,SRR6010091,SRR6010092,SRR6010093,SRX396185,SRX396186,SRX396187,SRX396188,SRX396189,SRX396190,SRX396191,SRX396192,SRX396193,SRX396194,SRX396195,SRX396196	Wang et al. (2019)
ZZM2161	Low oil content (48g/100g seed)			
ZZM3495	Low oil content (51g/100g seed)			

The criteria for selection were (a) oil content production (high oil-producing variety versus low oil-producing variety) and (b) seed color (white versus black). Thus, in addition to the multi-organs transcriptome data of the reference genome Zhongzhi13 (Wang et al., 2014), the seed RNA-Seq of two sesame pure lines Zhongfengzhi No.1 and Zhongzhi No.33 (Wang et al., 2020), exhibiting white and black seed color respectively, were used. The seed transcriptome sequences of one high (59.1%) oil content (ZZM4728) and two low oil content ZZM2161 (48.4%) and ZZM3495 (50.95%) varieties (Wang et al., 2019), were also downloaded from NCBI. The RNA samplings were performed at 10, 20, 25, and 30 days after anthesis for ZZM2161, ZZM3495, and ZZM4728, while Zhongfengzhi No.1 and Zhongzhi No.33 samplings were performed at 5, 8, 11, 14, 17, 20, 23, 26, and 30 days after anthesis. Detailed information regarding the materials and RNA-Seq raw data SRAs were provided in Table 13.

Expression Profile Quantification and Candidate Genes Selection

The RNA-Seq raw data were quality-checked using FastQC v0.11.2 (Andrews, 2010). Sequencing adapters and low-quality (Q < 30) reads were filtered out with Trimmomatic v0.36 (Bolger et al., 2014). The clean data were then mapped to the reference genome using HISAT v2.2.1 (Kim et al., 2015a). The gene expression profile following each tissue was assessed with the RSEM package (Li and Dewey, 2011) as fragments per kilobase of transcript per million fragments mapped (FPKM). The resulting heatmaps showing the differential expression within diverse tissues were plotted with TBTools v1.098746 (Chen et al., 2020) with $\log_2$(FPKM) values.

To select candidate genes, we applied three major filtering criteria. Firstly, from the reference genome dataset, the gene of interest should be preferentially expressed in the seed

tissue compared to root, leaf, stem, and capsule. Secondly, the candidate genes from the first filtering step have been checked for their expression in high versus low oil content varieties with a particular emphasis on those which are early expressed during the seed development stage. The early-stage criterion was included since pinoresinol is one of the precursors of the lignans pathway (Figure 19). Besides, the kinetic of the biosynthesis of the pinoresinol previously described by Ono et al. (2006) implied an increasing expression activity at early stage of the seed development that will enable later, the accumulation of downstream lignans in the seed at the maturity stage. Thirdly, from the second filtering step, the same approach was applied to the white versus black seed dataset.

Overall, the retained candidate genes belonged to the core gene repertoire from the sesame pan-genome, were preferentially expressed in seed tissues, and expressed higher at the early stage of the seed development.

Conserved Motifs, Gene structure, GO annotation, and Orthologs Detection of Candidate SiPOD and SiLAC genes

The candidate genes from the three steps filtering were screened for conserved motifs using MEME Suite v5.0.4 (Bailey et al., 2015) with a maximum number of motifs set to 20. A 2Kbp promoter sequence of candidate genes was submitted to the PlantCARE database (http://bioinformatics.psb.ugent.be/webtools/plantcare/html/) (Lescot et al., 2002) to find out cis-acting regulatory elements. Besides, the InterProScan v5 (Jones et al., 2014)) was employed to find out the molecular functions associated with the candidate genes. Meanwhile, their orthologs search was executed with SHOOT (https://www.shoot.bio/) (Emms and Kelly, 2022) using the plant database option.

Transcription factor enrichment analysis

Taking advantage of the Plant Transcriptional Regulatory Map (PlantRegMap) platform (Tian et al., 2020), we performed a transcription factor (TF) enrichment analysis (http://plantregmap.gao-lab.org/tf_enrichment.php) in order to estimate the most contributive TF families potentially involved in the regulation of peroxidase and laccase genes.

Results and Discussion

Peroxidase and Laccase Genes Variability in Sesame Pangenome and Phylogenetic Analyses

From the sesame pan-genome gene sets, there were 83, 82, 60, 59, 54, and 54 peroxidase genes counted in Zongzhi13, Swetha, Mishuozima, Baizhima, Goenbaek, and Yuzhi11 genomes, respectively (Figure 20).

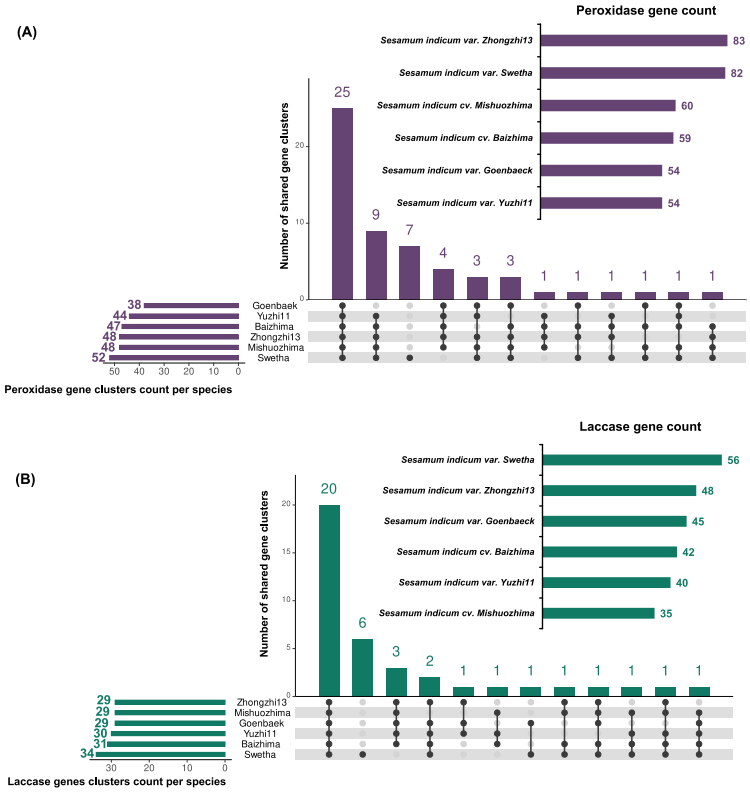

Figure 20. Gene count and conservation analysis of peroxidase (A) and laccase (B) genes in sesame pange-nome. Horizontal bar charts summarize the gene peroxidases/laccases count. Upset plots show the core conserved count of peroxidases/laccases within sesame pan-genome.

The peroxidase count variability observed at intra-species level within the sesame pangenome, is also noted at inter-species scale with 138, 119, 102, 90, 73, and 47 peroxidase genes counted in *Oryza sativa* (Passardi et al., 2004), *Zea mays* (Wang et al., 2015b), *Solanum tuberosum* (Yang et al., 2020), *Bentula pendula* (Cai et al., 2021), *A. thalina* (Tognolli et al., 2002), and *Vitis vinifera* (Xiao et al., 2020), respectively. Most of the gene clusters (25) are shared by all varieties while only Swetha exhibited species specific gene clusters (7).

An average of 44 ± 7 laccase genes were identified in the sesame pangenome. The most abundant number of laccases were observed for Swetha (56), followed by Zhongzhi13 (48), Goenbaek (n = 45), Baizhima (n = 42), Yuzhi11 (n = 40), and Mishuozhima (n = 35). Similar laccases count was observed in land plants including *Prunus persica* (n = 48) (Qui et al., 2022), *Panicum virgatum* (n = 49) (Li et al., 2022), *Solanum melongena* (n = 42) (Wan et al., 2022) but lower in *A. thaliana* (n = 17) (Turlapati et al., 2011). The orthology analysis revealed that the core laccase genes were grouped into 20 clusters while Swetha showed 6 specific gene clusters. With some exception to Yuzhi11 and Goenbaek, the peroxidase and laccase genes were globally abundant in modern varieties (Zhongzhi13, Yuzhi11, Swetha, and Goenbaek) compared to landraces (Mishuozima and Baizhima); suggesting the influence of the human oil-oriented selection. In fact, at whole-genome scale, landraces (Mishuozima and Baizhima) exhibited specific genes coding for environmental adaption while modern varieties showed preferential genes with oil-related functional attributes (Yu et al., 2019a).

Interestingly, only Swetha showed a specific gene cluster suggesting a singular peroxidase and laccase gene repertoire in Swetha. However, this should be analyzed with caution since

the Swetha chromosome-scale genome was constructed based on the reference genome of Zhongzhi13 (Yu et al., 2019a) with short-reads assembly as initial contigs-level assembly. For downstream analyses peroxidases and laccases from the reference genome Zhongzhi13 have been used (Table 14).

Table 14. List of identified peroxidase (*SiPOD*) and laccase (*SiLAC*) genes in *Sesamum indicum* var. Zhongzhi13

GeneID Code	NCBI gene locus tag	NCBI Protein ID	Linkage group	Start	End	Strand	Protein length (aa)
SiPOD1	LOC105177799	NP_001306621.1	LG15	4027009	4029056	+	330
SiPOD2	LOC105155321	XP_011069499.1	LG2	5075827	5080756	+	412
SiPOD3	LOC105155321	XP_011069505.1	LG2	5075827	5080756	+	363
SiPOD4	LOC105155173	XP_011069538.1	LG1	219480	222922	+	319
SiPOD5	LOC105155705	XP_011069927.1	LG2	10275638	10277587	-	336
SiPOD6	LOC105155907	XP_011070192.1	LG2	10724463	10726439	+	331
SiPOD7	LOC105156219	XP_011070595.1	LG2	14999942	15001479	+	322
SiPOD8	LOC105156220	XP_011070596.1	LG2	15004642	15006195	+	322
SiPOD9	LOC105156221	XP_011070597.1	LG2	15007577	15009014	-	322
SiPOD10	LOC105156222	XP_011070598.1	LG2	15015312	15016789	+	322
SiPOD11	LOC105156223	XP_011070599.1	LG2	15019876	15021509	+	321
SiPOD12	LOC105156224	XP_011070600.2	LG2	15027834	15029531	+	324
SiPOD13	LOC105156635	XP_011071133.1	LG2	17910231	17911979	+	433
SiPOD14	LOC105156835	XP_011071372.1	LG3	309851	311429	-	358
SiPOD15	LOC105159001	XP_011074235.1	LG3	18672137	18674658	+	346
SiPOD16	LOC105159307	XP_011074620.1	LG3	23364676	23366477	-	340

Table 14. *Continued*

SiPOD17	LOC105159364	XP_011074707.1	LG3	23754091	23755576	-	319
SiPOD18	LOC105159461	XP_011074839.1	LG3	24327365	24331541	+	355
SiPOD19	LOC105159461	XP_011074841.1	LG3	24327365	24331541	+	354
SiPOD20	LOC105160267	XP_011075883.1	LG4	8577270	8578961	+	325
SiPOD21	LOC105160817	XP_011076613.1	LG4	12163218	12165342	-	321
SiPOD22	LOC105161067	XP_011076942.1	LG4	14810782	14814226	+	377
SiPOD23	LOC105161067	XP_011076943.1	LG4	14810782	14814226	+	358
SiPOD24	LOC105161067	XP_011076944.1	LG4	14810782	14814226	+	356
SiPOD25	LOC105161240	XP_011077167.1	LG4	16328654	16330182	+	327
SiPOD26	LOC105161241	XP_011077168.1	LG4	16344174	16345709	+	327
SiPOD27	LOC105161265	XP_011077204.1	LG4	16345855	16348226	+	223
SiPOD28	LOC105162056	XP_011078263.1	LG5	4994731	4997357	-	318
SiPOD29	LOC105162057	XP_011078264.1	LG5	5006658	5010427	-	317
SiPOD30	LOC105162255	XP_011078554.1	LG1	10578194	10579889	-	339
SiPOD31	LOC105162609	XP_011078985.1	LG5	12713099	12714915	-	332
SiPOD32	LOC105162642	XP_011079026.1	LG5	14490145	14493286	-	332
SiPOD33	LOC105163909	XP_011080738.1	LG6	6655746	6663339	-	329
SiPOD34	LOC105163910	XP_011080739.1	LG6	6646504	6649894	-	328
SiPOD35	LOC105164889	XP_011082007.1	LG6	18174254	18176194	+	340
SiPOD36	LOC105165620	XP_011082989.2	LG6	23430520	23433568	-	320
SiPOD37	LOC105165732	XP_011083137.1	LG6	24275260	24277252	-	325
SiPOD38	LOC105165752	XP_011083175.1	LG6	24496720	24498335	-	331
SiPOD39	LOC105165788	XP_011083218.1	LG6	24995591	24997111	+	321
SiPOD40	LOC105166002	XP_011083477.1	LG7	2511417	2513602	+	318
SiPOD41	LOC105167693	XP_011085801.1	LG8	3834989	3836347	-	328
SiPOD42	LOC105167987	XP_011086187.1	LG8	8728567	8730311	-	330
SiPOD43	LOC105168207	XP_011086488.1	LG8	10610639	10612217	-	332

Table 14. *Continued*

SiPOD44	LOC105169185	XP_011087829.1	LG8	17891314	17893156	+	331
SiPOD45	LOC105169211	XP_011087857.1	LG8	18032762	18034594	-	352
SiPOD46	LOC105169211	XP_011087858.1	LG8	18032762	18034594	-	329
SiPOD47	LOC105169784	XP_011088597.1	LG8	20970684	20974843	-	288
SiPOD48	LOC105169899	XP_011088753.1	LG8	19827804	19829758	-	337
SiPOD49	LOC105170229	XP_011089203.2	LG9	743154	744955	+	335
SiPOD50	LOC105170682	XP_011089855.1	LG9	5144699	5147939	+	250
SiPOD51	LOC105172108	XP_011091748.1	LG10	5742934	5746837	-	351
SiPOD52	LOC105172589	XP_011092399.1	LG10	14707344	14710521	+	320
SiPOD53	LOC105172805	XP_011092685.1	LG10	13442527	13444592	+	335
SiPOD54	LOC105174327	XP_011094690.1	LG11	13459068	13461004	+	323
SiPOD55	LOC105174346	XP_011094725.1	LG11	13552835	13556074	-	250
SiPOD56	LOC105174640	XP_011095107.1	LG11	15140686	15142588	+	321
SiPOD57	LOC105174641	XP_011095108.1	LG11	15144165	15145538	+	324
SiPOD58	LOC105175345	XP_011096066.1	LG12	3305255	3307261	+	330
SiPOD59	LOC105175398	XP_011096136.1	LG12	2432345	2434061	-	343
SiPOD60	LOC105175601	XP_011096394.1	LG12	4620863	4623538	-	342
SiPOD61	LOC105175960	XP_011096905.1	LG12	6286486	6287926	+	326
SiPOD62	LOC105176241	XP_011097275.1	LG13	3162487	3164343	-	326
SiPOD63	LOC105176439	XP_011097539.1	LG13	4388157	4390528	-	332
SiPOD64	LOC105177169	XP_011098522.1	LG15	530850	532188	+	335
SiPOD65	LOC105177177	XP_011098533.1	LG15	563733	565085	+	327
SiPOD66	LOC105177323	XP_011098747.1	LG2	2443657	2449297	-	346
SiPOD67	LOC105177673	XP_011099203.1	LG15	3188852	3191073	-	327
SiPOD68	LOC105178505	XP_011100296.1	LG16	957325	959584	-	327
SiPOD69	LOC105178606	XP_011100419.1	LG16	981195	991536	-	324
SiPOD70	LOC105179284	XP_011101199.1	LG2	4067865	4069707	-	353

Table 14. *Continued*

SiPOD71	LOC105164521	XP_020547868.1	LG1	10951538	10958337	-	283
SiPOD72	LOC105161263	XP_020548845.1	LG4	16340170	16341560	+	339
SiPOD73	LOC105161262	XP_020548851.1	LG4	16325516	16326870	+	339
SiPOD74	LOC105161067	XP_020548915.1	LG4	14810782	14814226	+	382
SiPOD75	LOC105162172	XP_020549401.1	LG5	5459904	5462322	+	324
SiPOD76	LOC105163979	XP_020549974.1	LG6	6011516	6014681	-	320
SiPOD77	LOC110012114	XP_020550076.1	LG6	5999248	6002541	-	323
SiPOD78	LOC105165620	XP_020550649.1	LG6	23430520	23433568	-	326
SiPOD79	LOC105170828	XP_020552323.1	LG9	2603778	2607483	-	141
SiPOD80	LOC105171593	XP_020553136.1	LG10	1047079	1048584	+	329
SiPOD81	LOC105175601	XP_020553919.1	LG12	4620863	4623538	-	264
SiPOD82	LOC105177323	XP_020554732.1	LG2	2443657	2449297	-	346
SiPOD83	LOC105178605	XP_020555028.1	LG16	967250	970663	-	324
SiLAC1	LOC105158065	XP_011072992.1	LG3	11683186	11688476	-	583
SiLAC2	LOC105158919	XP_011074134.2	LG3	19998216	20008384	-	1094
SiLAC3	LOC105159584	XP_011074997.2	LG3	23660270	23663303	-	540
SiLAC4	LOC105160591	XP_011076336.1	LG4	10837166	10841534	-	589
SiLAC5	LOC105160846	XP_011076646.1	LG4	12343777	12346486	+	578
SiLAC6	LOC105161083	XP_011076960.1	LG4	14345676	14349168	+	571
SiLAC7	LOC105161178	XP_011077090.1	LG4	15942526	15944824	-	558
SiLAC8	LOC105161557	XP_011077581.1	LG5	1326619	1329463	-	557
SiLAC9	LOC105161599	XP_011077639.1	LG5	1555983	1558799	+	578
SiLAC10	LOC105163722	XP_011080467.1	LG6	4594330	4602428	+	583
SiLAC11	LOC105163777	XP_011080540.1	LG6	5030461	5038524	+	595
SiLAC12	LOC105163916	XP_011080750.1	LG6	6772908	6775663	-	568
SiLAC13	LOC105163993	XP_011080832.2	LG6	6739865	6742257	-	481
SiLAC14	LOC105164010	XP_011080855.1	LG6	7495715	7512811	-	554

Table 14. *Continued*

SiLAC15	LOC105164284	XP_011081204.1	LG6	10032137	10036492	-	592
SiLAC16	LOC105166647	XP_011084388.2	LG7	7883756	7893878	+	618
SiLAC17	LOC105167012	XP_011084865.1	LG7	10564762	10567822	+	580
SiLAC18	LOC105167055	XP_011084917.2	LG1	14072620	14077387	+	580
SiLAC19	LOC105167080	XP_011084946.1	LG7	11040221	11042719	-	555
SiLAC20	LOC105167465	XP_011085509.1	LG8	2609921	2613220	+	538
SiLAC21	LOC105169197	XP_011087840.1	LG8	17967232	17969842	-	570
SiLAC22	LOC105169650	XP_011088411.1	LG8	20310353	20312751	+	559
SiLAC23	LOC105169830	XP_011088669.1	LG8	21246987	21250543	-	538
SiLAC24	LOC105170430	XP_011089492.1	LG9	3282633	3285226	+	578
SiLAC25	LOC105170577	XP_011089702.1	LG9	4296118	4298003	+	555
SiLAC26	LOC105170621	XP_011089763.1	LG9	4572419	4574294	+	555
SiLAC27	LOC105170622	XP_011089764.1	LG9	4576660	4578543	+	554
SiLAC28	LOC105170623	XP_011089765.1	LG9	4581842	4583628	+	554
SiLAC29	LOC105170624	XP_011089766.1	LG9	4587067	4588974	-	555
SiLAC30	LOC105170625	XP_011089767.1	LG9	4589956	4591783	+	555
SiLAC31	LOC105173860	XP_011094064.2	LG1	18419180	18422936	+	562
SiLAC32	LOC105174465	XP_011094885.2	LG11	14135034	14137439	-	575
SiLAC33	LOC105174466	XP_011094887.1	LG11	14139187	14141881	-	583
SiLAC34	LOC105174467	XP_011094888.1	LG11	14145378	14147817	-	578
SiLAC35	LOC105174540	XP_011094977.1	LG11	14549546	14552701	+	558
SiLAC36	LOC105174721	XP_011095211.1	LG11	14420150	14422933	+	569
SiLAC37	LOC105175333	XP_011096053.1	LG12	3235193	3238630	-	541
SiLAC38	LOC105176483	XP_011097600.1	LG13	4906457	4908971	+	570
SiLAC39	LOC105176975	XP_011098284.1	LG2	1800238	1806833	+	572
SiLAC40	LOC105177684	XP_011099220.1	LG15	3257180	3260595	+	539
SiLAC41	LOC105177685	XP_011099223.1	LG15	3264150	3266987	+	535

Table 14. *Continued*

SiLAC42	LOC105178005	XP_011099634.1	LG15	5802610	5805209	+	565
SiLAC43	LOC105178053	XP_011099705.1	LG2	2699177	2704090	+	538
SiLAC44	LOC105179101	XP_011100999.1	LG2	3887053	3893030	-	563
SiLAC45	LOC105179108	XP_020547488.1	LG2	3911499	3918040	-	570
SiLAC46	LOC105161083	XP_020549108.1	LG4	14345676	14349168	+	466
SiLAC47	LOC105167465	XP_020551736.1	LG8	2609921	2613220	+	538
SiLAC48	LOC105174713	XP_020553486.1	LG11	14122512	14124992	+	576

In addition to the chromosomal location of the identified genes (Figure 21), we dived into the assessment of the evolutionary determinants of the *SiPOD* and *SiLAC* genes distribution within the sesame genome.

Thus, two types of genes have been found (Figure 22, Figure 23): the paralogous genes that are adjacent on the same chromosome (tandem duplicated genes) and those which are far away from each other on different chromosomes (segmental duplicated genes). For *SiPOD*, the results highlighted 18 pairs paralogous genes that underwent segmental duplication while one pair derived from tandem duplication (Figure 22).

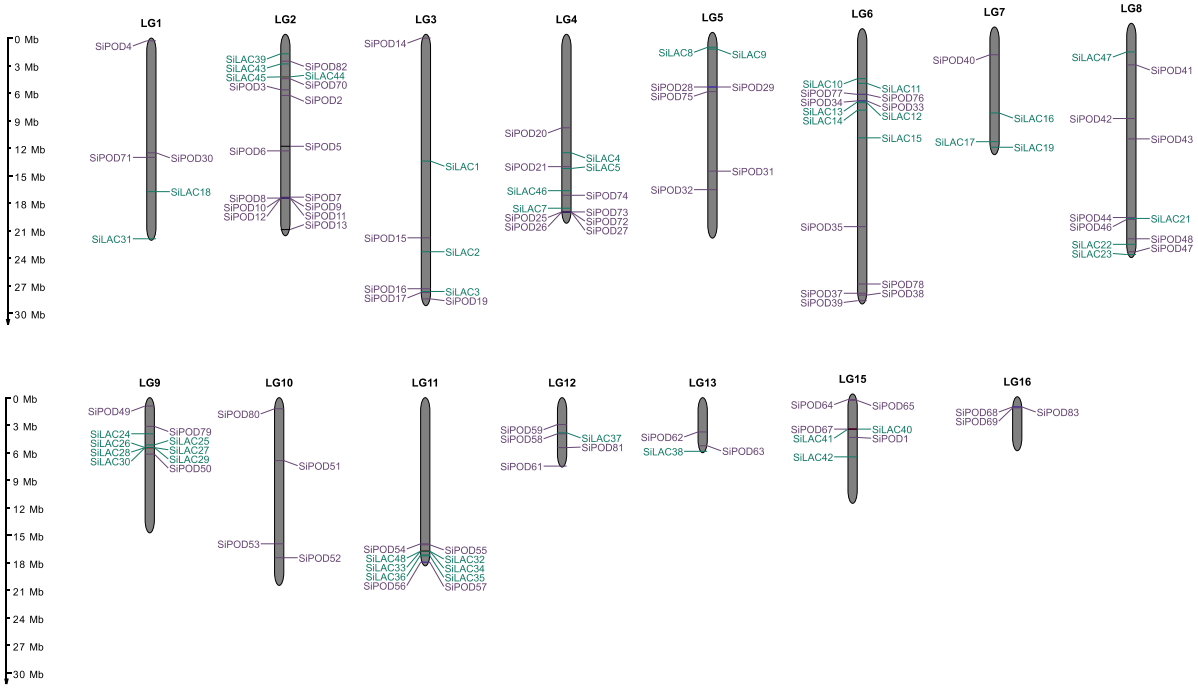

Figure 21. Chromosome location of peroxidase and laccase genes in sesame

Tandem and segmental duplications are considered as an evolutionary driving force resulting in gene families' expansion (Moore and Purugganan, 2003; Cannon et al., 2004). Peroxidase genes duplication via tandem and segmental duplications have been extensively reported in plants including soybean (Aleem et al., 2022), cassava (Wu et al., 2019a), carrot (Meng et al., 2021), potato (Yang et al., 2020), cotton (Duan et al., 2019), watermelon (Yang et al., 2022), maize (Wang et al., 2015b), tomato (Huang et al., 2022), Chinese pear (Cao et al., 2016) and others. In our study, peroxidase gene expansion is mainly affected by segmental duplication; which is consistent with the findings of Cao et al (Cao et al., 2016) in pear. However, in trees such as *Betula pendula* (Cai et al., 2021) and *Populus trichocarpa* (Ren et al., 2014), tandem duplication is the main drivers of peroxidase genes expansion.

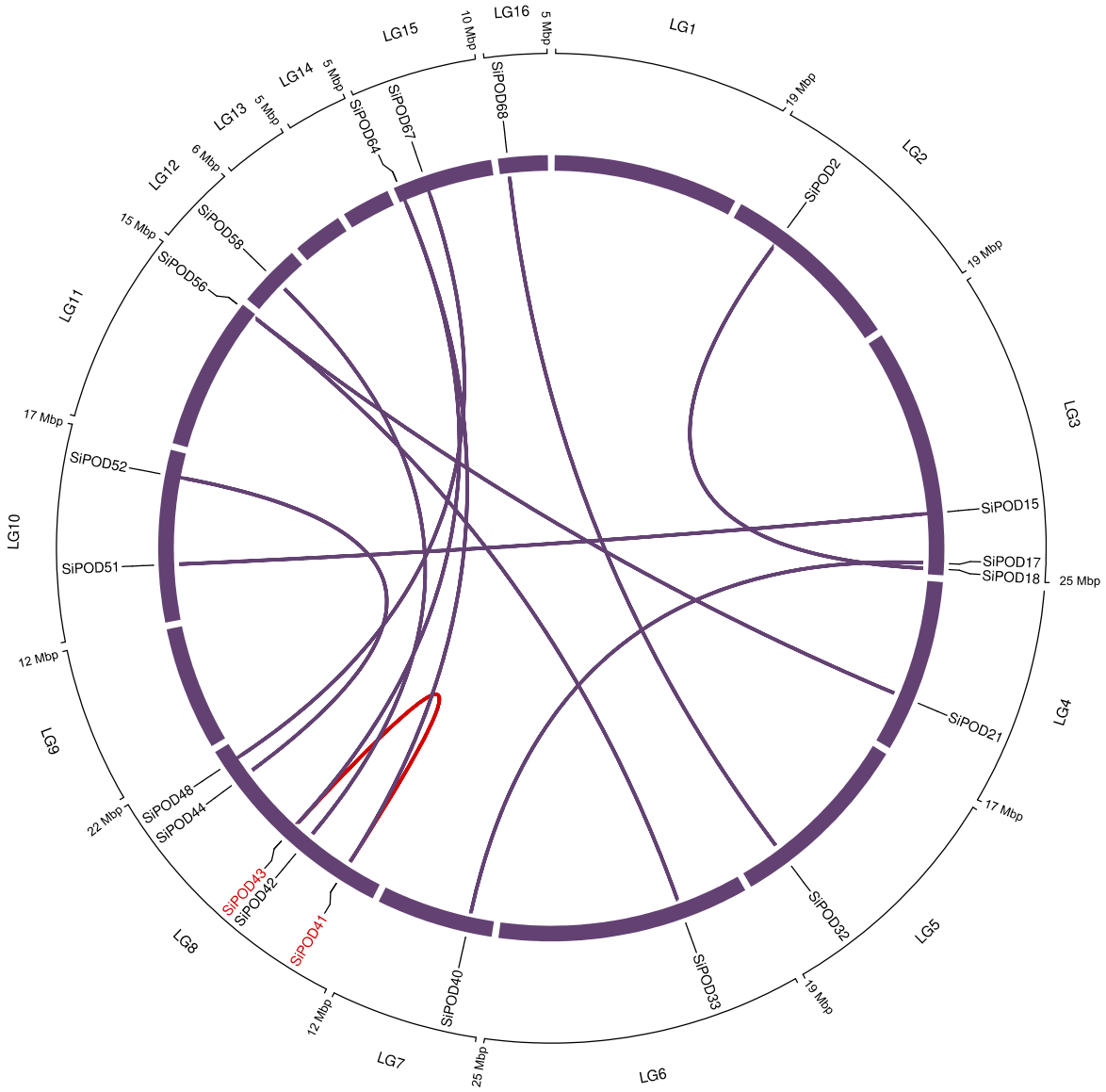

Figure 22. Circos plot showing paralogous peroxidase genes exhibiting segmental duplication in sesame. Genes colored in red are tandem duplicated.

Similarly, a total of 20 *SiLAC* genes belonged exclusively to segmental duplication in sesame (Figure 23), while only tandem duplication pattern was detected in *S. melongena* with 16 laccase genes (Wan et al., 2022). For *P. virgatum*, both tandem and segmental duplications were highlighted (Li et al., 2022).

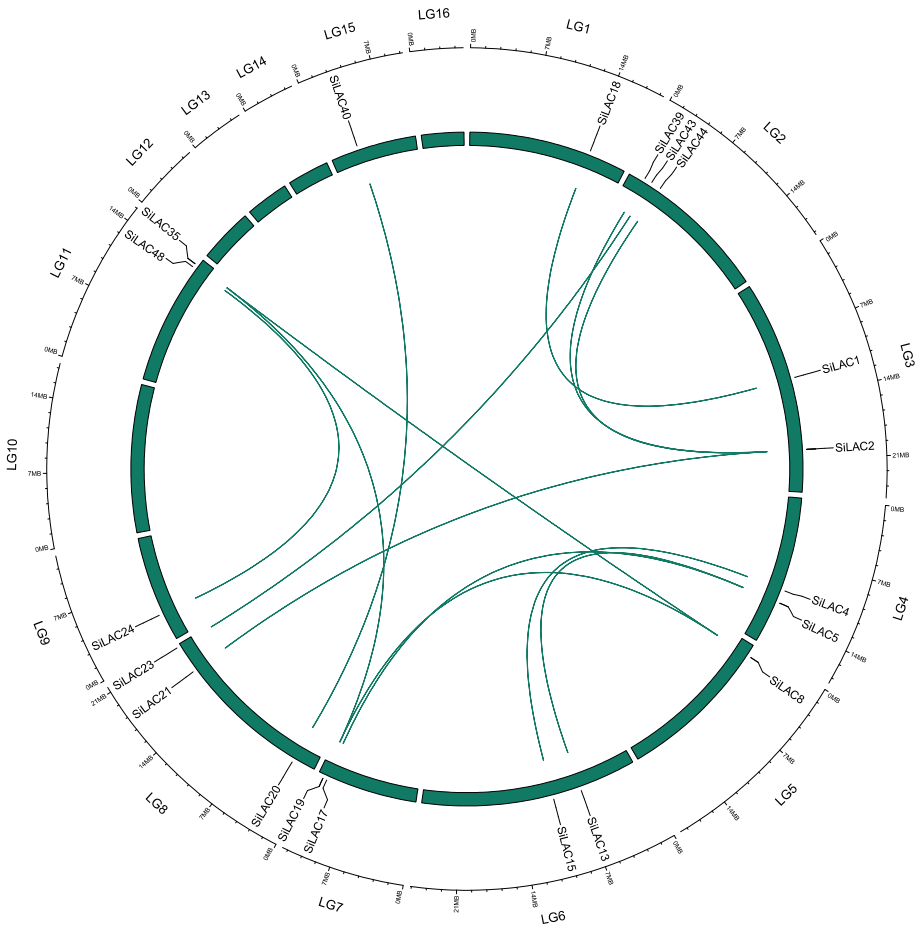

Figure 23. Circos plot showing paralogous laccase genes belonging to segmental duplication pattern.

The diversity of gene duplication pattern for peroxidases and laccases indicates different

routes for the oxidases expansion following the species.

The synteny analysis revealed that 31 (37%) *SiPOD* and 26 (54%) *SiLAC* genes showed synteny with *A. thalina* respective gene sets (Figure 24 and Figure 25), suggesting that they are conserved within both species.

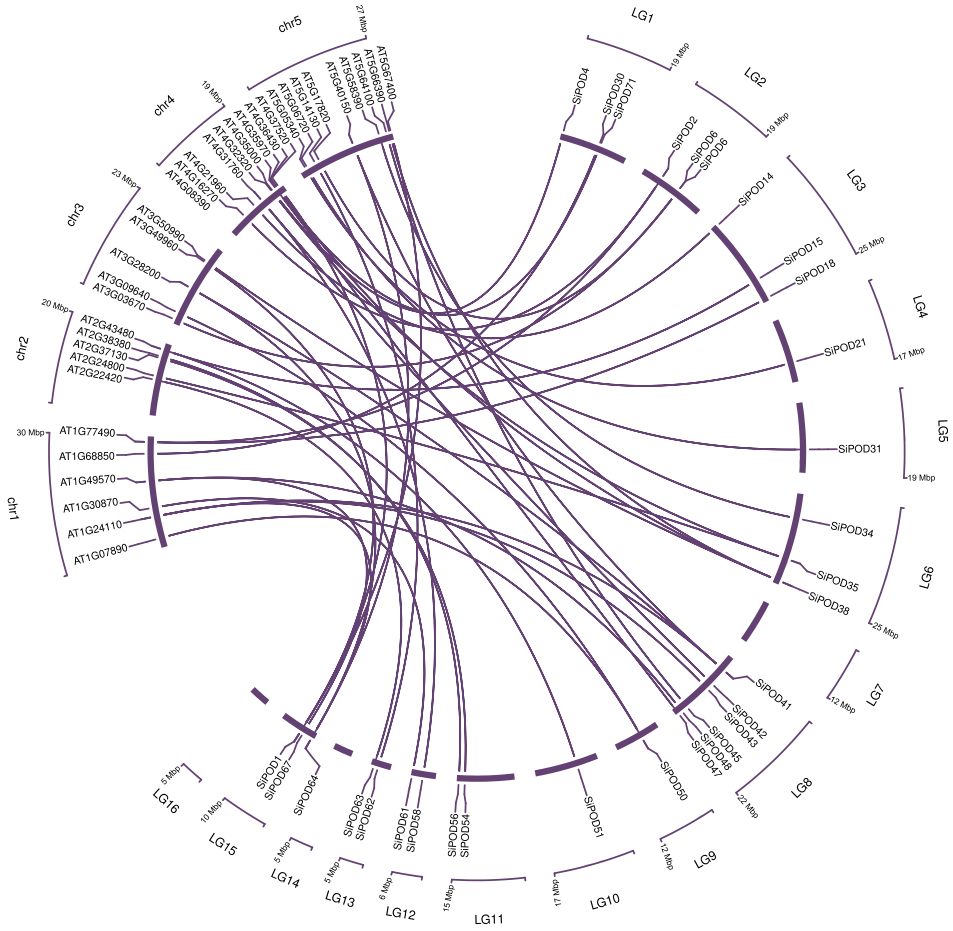

Figure 24. Circos plot showing synthenic peroxidase genes between *Sesamum indicum* and *Arabidopsis thaliana*.

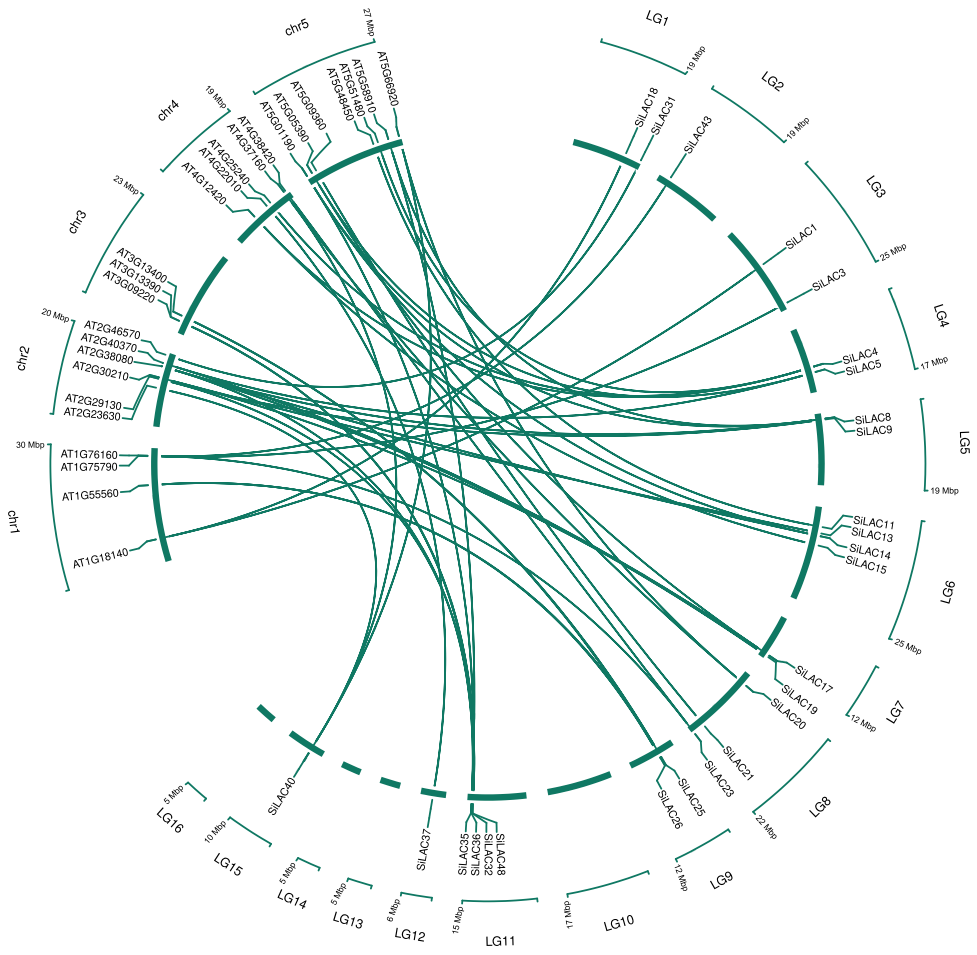

Figure 25. Circos plot showing synthenic laccase genes between *Sesamum indicum* and *Arabidopsis thaliana*.

To classify the identified peroxidase and laccase genes, phylogenetic trees were constructed using *A. thaliana* genes as baits. The results revealed nine and seven groups for peroxidase and laccase gene families, respectively (Figure 26 and Figure 27).

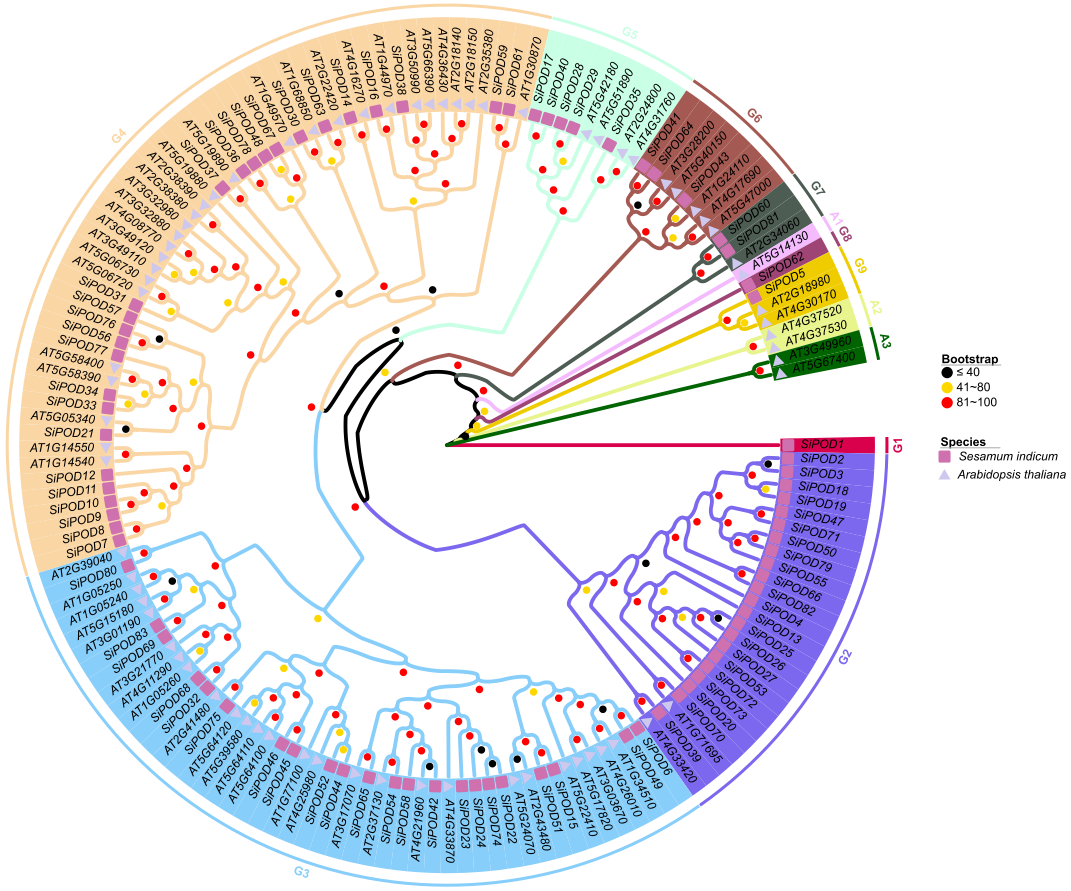

Figure 26. Unrooted maximum likelihood phylogenetic tree of peroxidases. The tree was constructed following the LG+R5 model with a total of 1,000 iterations. Square shape represents the genes from *S. indicum* while triangle shape stands for *A. thaliana* genes. Black (<40), yellow (41-80), and red (81-100) dot represents the clades support values. Arabidopsis-specific phylogenetic groups were noted with the letter A (A1-A3). Groups containing *SiPOD* start with the letter G (G1-G9).

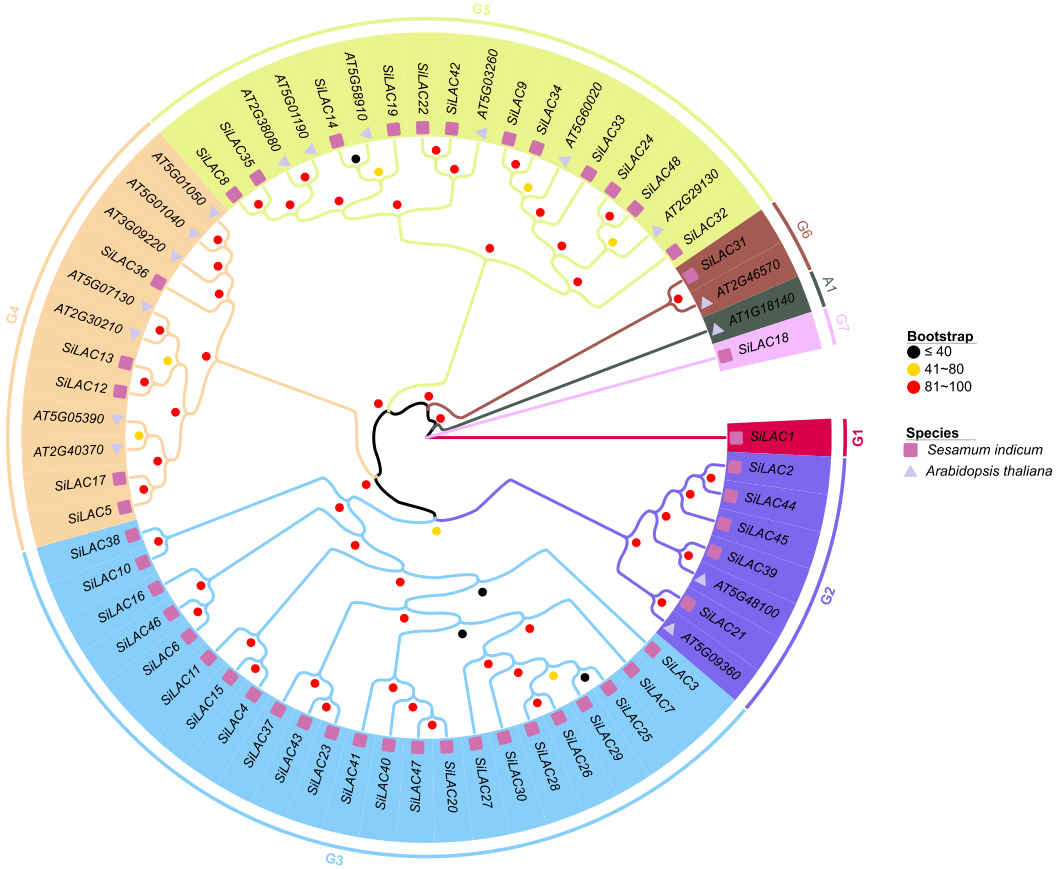

Figure 27. Unrooted maximum likelihood phylogenetic tree of laccases. The tree was constructed following the LG+I+G4 model with a total of 1,000 iterations. Square shape represents the genes from *S. indicum* while triangle shape stands for *A. thaliana* genes. Black (<40), yellow (41-80), and red (81-100) dot represents the clades support values. Arabidopsis-specific phylogenetic groups were noted with the letter A (A1). Groups containing *SiLAC* start with the letter G (G1-G7)

Expression Profiles of SiPOD and SiLAC genes within diverse tissues and Candidate Genes Selection

To find out robust potential (+)-pinoresinol synthase genes, a multi-varieties comparative transcriptome approach has been utilized. RNA-Seq data from six sesame varieties (Table 13) have been inspected following the three filtering steps including (i) preferential expression in seed, (ii) expression at least at early stage of the seed development, and (iii) expression of the gene across all tested varieties. High oil content and white seed varieties were employed as positive control for the gene selection step.

For peroxidase genes (Figure 28), a set of eight *SiPOD* (*SiPOD41*, *SiPOD42*, *SiPOD47*, *SiPOD50*, *SiPOD52*, *SiPOD8*, *SiPOD63*, and *SiPOD65*) genes were preferentially expressed in the seeds of Zhongzhi 13 (Figure 28A).

The selected genes were mined for their expression in the seeds of low (LOS) versus high (HOS) oil content varieties (Figure 28B). The results pinpointed *SiPOD42*, *SiPOD52*, *SiPOD58*, *SiPOD63*, *SiPOD65*, and *SiPOD50*. *SiPOD42* was most expressed in HOS at 10 days after compared to the two LOS varieties. Its expression was maintained within all the seed development stages preferentially in the HOS. Similar observations were noted for *SiPOD52*, *SiPOD58*, and *SiPOD50* with an ascending expression from 10 to 20 days after anthesis followed by a decline from 25 days after anthesis. Interestingly, *SiPOD63* and *SiPOD65* were expressed at early stage (10 days after anthesis) in all varieties before the expression fell in the following development stages.

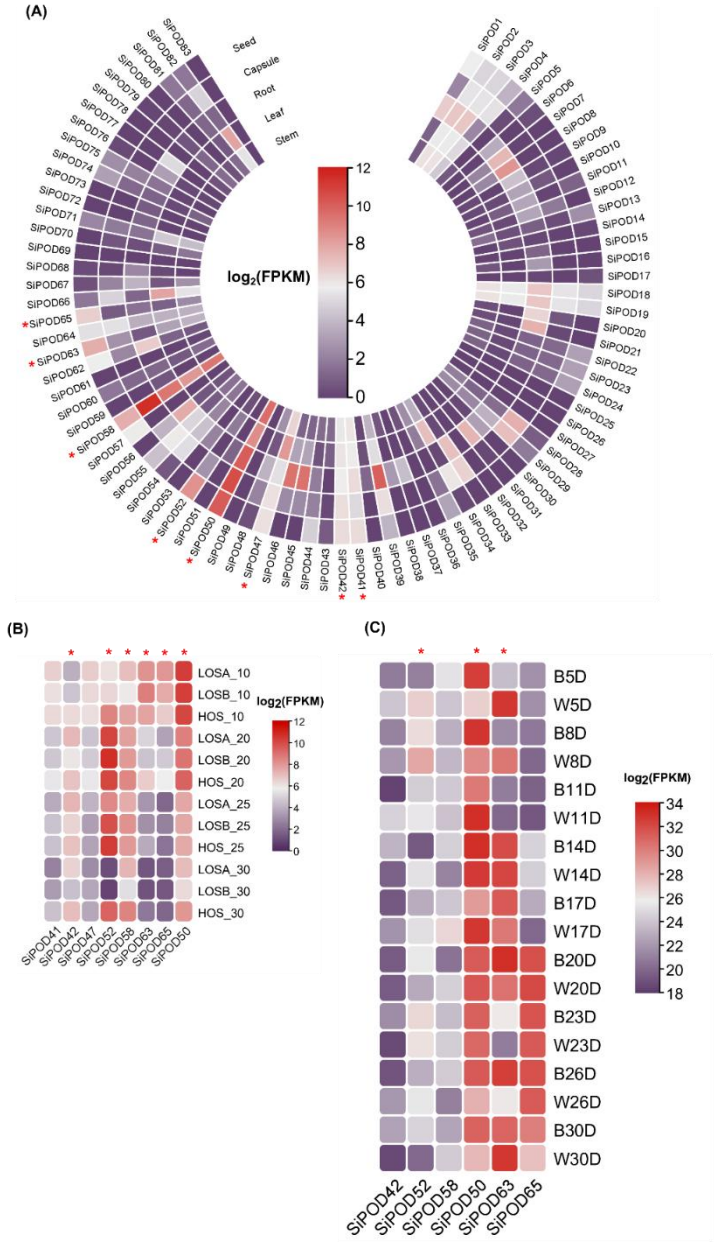

Figure 28. Expression profile of sesame peroxidase genes within diverse tissues: (A) seed, root, leaf, stem and capsule from Zhongzhi13; (B) Seed from ZZM4728 (HOS), ZZM3495 (LOSA), and ZZM2161 (LOSB); (C) Seed from black seed Zhongzhi No. 33 (B) and white seed Zhongzhi No.1 (W) varieties. Candidate genes for each set of tissues are highlighted with a red star symbol.

The candidate genes from the second filtering round were screened for their preferential high expression in white seed or at least their expression at earlier stage of the seed development regardless the seed coat color (Figure 28C). Thus, the genes *SiPOD52*, *SiPOD63*, *SiPOD50*, and *SiPOD65* came up to be the potential candidates. *SiPOD52* and *SiPOD63* showed a higher expression in white seed at early stage (five and eight days after anthesis). Intriguingly higher expression of *SiPOD50* was noted in black seed at five and eight days after anthesis. However, the expression of the gene was quite stable within all development stages regardless the seed color.

As for the laccase genes (Figure 29), *SiLAC1*, *SiLAC12*, and *SiLAC39* came out on the top in the first filtering round (Figure 29A). From this set, *SiLAC39* showed the peak of expression in both low and high oil content varieties at 20 days after anthesis. Similarly, *SiLAC1* expression was higher at 25 days after anthesis regardless of the type of variety (Figure 29B). Furthermore, the expression of the two later genes was checked in the black versus white varieties (Figure 29C). Interestingly, *SiLAC1* exhibited a higher expression in white seed variety (Zhongzhi No.1) at early stage of the seed development (8 days after anthesis) compared to the black one (Zhongzhi No.33). Similarly, *SiLAC39* expression was differentially higher in white seed variety (Zhongzhi No.1) at both 11 and 17 days after anthesis.

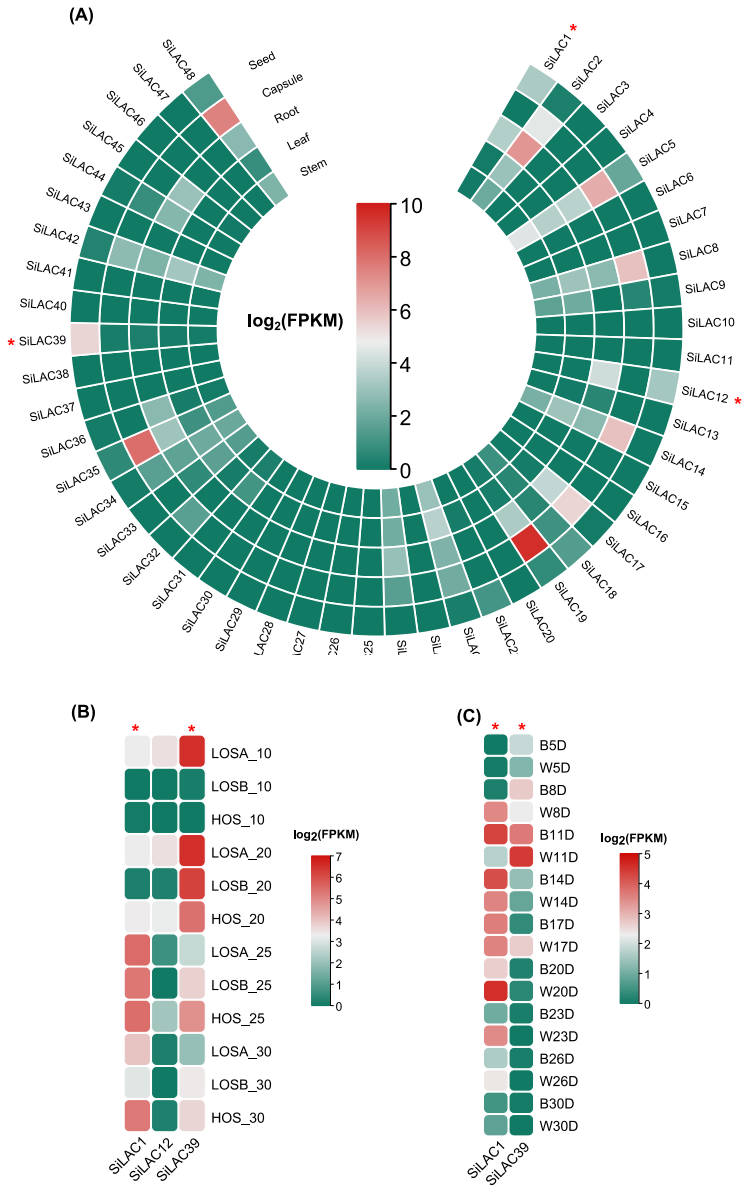

Figure 29. Expression profile of sesame laccase genes within diverse tissues: (A) seed, root, leaf, stem and capsule from Zhongzhi13; (B) Seed from ZZM4728 (HOS), ZZM3495 (LOSA), and ZZM2161 (LOSB); (C) Seed from black seed Zhongzhi No. 33 (B) and white seed Zhongzhi No.1 (W) varieties. Candidate genes for each set of tissues are highlighted with a red star symbol.

It is worth noting that all candidate genes fall into the principle according to which they should belonged to the core gene repertoire of the sesame pan-genome and also, followed the piperosinol synthesis kinetic as described by Ono et al., (2006) and depicted in the Figure 30 .

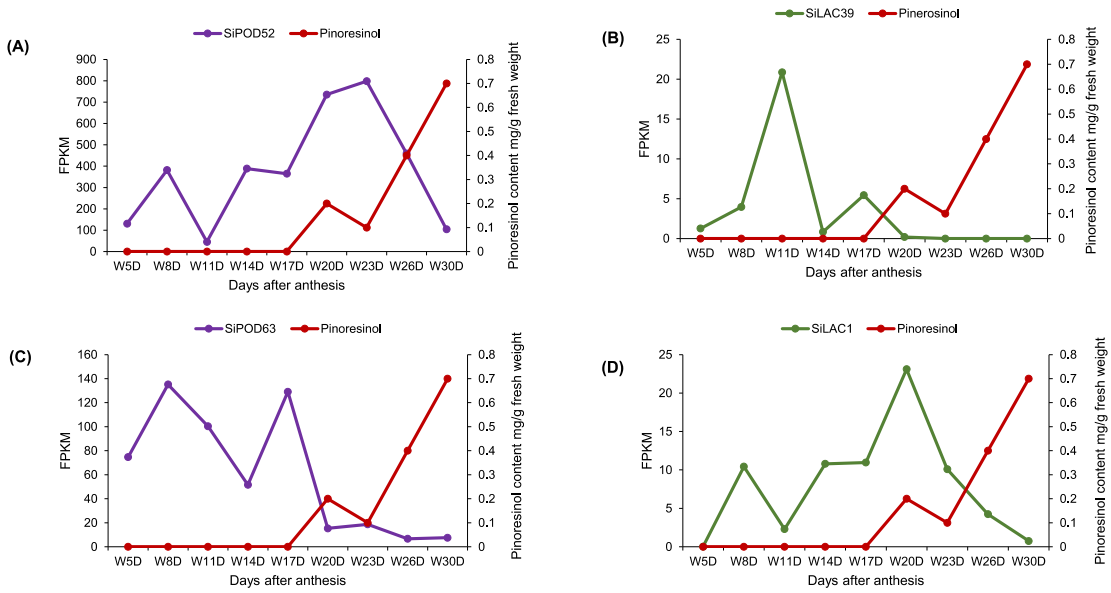

Figure 30. FKPM variation of candidate peroxidase and laccase genes following sesame seed development and pinoresinol content in the white seed pure line genotype Zhongfengzhi No1. Pinoresinol content values following seed development stages are collected from Ono et al. (2006) study.

Briefly, the kinetic of the lignans biosynthesis suggests the expression of oxidative enzymes (potentially here peroxidase and/or laccase) at upstream stage (from +(-)coniferyl alcohol to +(-)pinoresinol synthesis) of the lignans biosynthesis. Therefore, the pinoresinol synthase is supposed to be expressed at early stage of the seed development to enable the downstream biosynthesis resulting into the +(-) sesamin, +(-) sesamolin, +(-) sesaminol at maturity stage

(~30 days after anthesis). Thus, by combining the pinoresinol content extracted from sesame seed by Ono et al. (2006) and the transcript data sets, we picked up two peroxidase (*SiPOD52*, *SiPOD63*) and two laccase (*SiLAC1*, *SiLAC39*) genes. Globally, the FPKM values of the four candidate genes were declining at 23 days after anthesis (approximatively one week before the maturation stage) Figure 30. The *SiPOD50* gene was filtered out since it was constantly expressed at all development stages (Figure 28C) of the seed, which does not match the lignan biosynthesis kinetic. Therefore, for downstream bioinformatic analyses (cis-acting elements, orthologs identification and gene ontology analyses), only *SiPOD52*, *SiPOD63*, *SiLAC1*, and *SiLAC39* genes were processed.

Cis-Acting Elements, Related Transcription Factors, and Functionnal Attributes

Although the candidate genes shared similar gene structure (Figure 31A-C), a wide diversity of regulation and functional characteristics were highlighted through the cis-element analysis (Figure 31D), including hormonal response, light response, abiotic stress response, and physiological development. This suggests that *SiPOD* and *SiLAC* genes might involve in a broad spectrum of biological processes in sesame. Knowing the important role of transcription factors in gene regulation, we performed a transcription factor oriented enrichment analysis by using all *SiPOD* and *SiLAC* genes, to find out the most contributive TF that potentially are involved in peroxidase and laccase gene regulation. A panel of TF families have been predicted among which v-myb avian myeloblastosis viral oncogene homolog (*MYB*), *NAM* (no apical meristem), *ATAF1-2* (Arabidopsis thaliana activating factor), CUC2 (cup-shaped cotyledon) (*NAC*), Basic leucine zipper (*bZIP*), Heat shock factors (*HSF*), Homeodomain-leucine zipper (*HD-ZIP*), MIKC-type MADS-box (*MIKC_MADS*) were the most abundant. As depicted in the Figure 31 E and F, MYB are the

most predominant TFs indicating their putative regulatory role on both peroxidases and laccases. Using transgenic *A. thaliana* lines, Shen et al. (2017) showed that the sweet cherry (*Prunus avium* cv. Hong Deng) *R2R3 MYB* was able to alleviate salt stress and bacterial resistance with an the accumulation of peroxidase and anthocyanin. Furthermore, co-expression of *A. thaliana* laccases (*lac4* or *lac17*) with *MYB63* genes is known to rescue dwarfism in *A. thaliana* mutant lines (Perkins et al., 2020).

Figure 31. Candidate peroxidase and laccase genes from transcriptome profiling. Protein motifs (A), domain (B) and gene structure (C), cis-acting elements functional attributes (D), transcription factor enrichment word cloud of laccase (E), and peroxidase (F) candidate genes.

Looking at the orthologs of the candidate genes in other taxa by using a phylogenetic-based approach with SHOOT (Emms and Kelly, 2022), we found homologous sequences in oil- and non-oil crops including *Solanum lycopersicum*, *Arabidopsis thaliana*, *Brassica oleracea*, *Gossypium raimondii*, *Glycine max*, *Triticum aestivum*, *Oryza sativa*, and *Zea mays* (Table 15).

Table 15. List of candidate genes and their orthologous sequences

	SiPOD52	***SiPOD63***	***SiLAC1***	***SiLAC39***
NCBI gene locus tag	LOC105172589	LOC105176439	LOC105158065	LOC105176975
NCBI protein ID	XP_011092399.1	XP_011097539.1	XP_011098284.1	XP_011098284.1
Arabidopsis thaliana	AT4G25980.1*	AT2G22420.1	AT5G48100.1	AT5G48100.1
Solanum lycopersicum	Solyc10g047110.2.	Solyc11g010120.2.	Solyc04g058040.2.	Solyc04g058000.2.
Brassica oleracea	Bol039581	Bol044748	Bol033117	Bol033117
Gossypium raimondii	Gorai.002G208800	Gorai.008G090200	Gorai.009G260400	Gorai.009G260400
Glycine max	Glyma.15G050800	Glyma.09G002500	Glyma.10G219200	Glyma.20G189800
Triticum aestivum	Traes_2AL_3FB8B316E.1	Traes_5BL_881FB6FDE.1	Traes_5DL_CB39AD4BA.1	Traes_5DL_CB39AD4BA.1
Oryza sativa	LOC_Os12g08920	LOC_Os09g29490	LOC_Os10g30140	LOC_Os10g30140
Zea mays	GRMZM2G081928_P01	GRMZM5G843748_P02	GRMZM2G320786_P01	GRMZM2G320786_P01

*Orthologs available in Phytozome database https://phytozome-next.jgi.doe.gov/

The GO annotation confirmed that peroxidase genes are involved in hydrogen peroxidase catabolic process (GO:0042744) with a heme binding (GO:0020037) and peroxidase activity (GO:0004601) as main molecular functions. Regarding laccase, the GO annotation revealed that *SiLAC1* and *SiLAC39* are related to lignin catabolic biological process (GO:0046274) with copper ion binding (GO:0005507), oxidase activity (GO:0016491), and hydroquinone: oxygen oxidase activity (GO:0052716) as major molecular functions. From the GO annotation results, both *SiPOD* and *SiLAC* genes were predicted to have an oxidative role; which is a key requirement that may contribute to the transformation of the (+)- coniferyl alcohol into the (+)- pinoresinol. As matter of fact, peroxidase and laccase genes were reported to be able to oxidize coniferyl and p-coumaryl alcohols; acting as catalysts during cell wall lignification in *Zinnia elegans* (Gavnholt and Larsen, 2002; Ros Barceló et al., 2004; Barceló et al., 2007; Novo-Uzal et al., 2013; Tugbaeva et al., 2021). Therefore, the suggested genes are valuable candidate for functional validation and ultimately, for usage in pharmaceutical and food industries through bioengineering. Since sesame is a recalcitrant material for genetic transformation via classical methods (Andargie et al., 2021), hairy roots method might be a valuable alternative path for in-vitro production of sesame lignans as demonstrated by Ogasawara et al. (1998).

The present study highlighted the peroxidase and laccase genes in the sesame reference genome Zhongzhi13. A set of 83 peroxidase and 48 laccase genes have been identified. A variability of gene count was noted within the sesame pan-genome. Taking advantage of a large panel of transcriptome data, and a stringent filtering approach, four candidate genes (*SiPOD52*, *SiPOD63*, *SiLAC1*, and *SiLAC39*) have been proposed to play the function of (+)-pinoresinol synthase. The peroxidase and laccase genes were predicted to interact with a wide panel of transcription factors and involved in diverse molecular processes. Taking

together, their presumed oxidative function may be crucial for the pinoresinol biosynthesis, a precursor of the sesame lignans. Thus, the findings from this study provide valuable information for a functional investigation of the candidate genes and the genetic improvement of specialized metabolites biosynthesis in sesame.

CHAPTER 4: Insights into the Speciation in Sesamum Genus via a Phylogenomics Approach

Summary

In Sesamum species complex, the lack of wild species genomic resources for the phylogenetic relationship elucidation hinders the evolutionary background comprehension of speciation. In the present study, we generated complete chloroplast genomes of six wild relatives (*Sesamum alatum*, *Sesamum angolense*, *Sesame pedaloides*, *Ceratotheca sesamoides*, *Ceratotheca triloba*, *Sesamum radiatum*) and one Korean cultivar *Sesamum indicum cv*. Goenbaek. A typical quadripartite chloroplast structure including two inverted repeats (IR), a large single copy and a small single copy was observed. A total of 114 unique genes encompassing 80 coding-genes, four ribosomal RNAs, and 30 transfer RNAs were counted. The chloroplast genomes (152,863-153,338 bp) exhibited IR contraction/expansion phenomenon and were quite conserved in both coding and non-coding regions. However, the highest nucleotide diversity index was found in ycf1 gene, followed by ndhA, ndhE, psaC-ndh-D and ndhF, providing useful markers for taxon discrimination. The phylogenetic inference, the time divergence dating, and nuclear whole-genome alignment suggests that *S. radiatum* (2n = 64) have occurred by the hybridization of *C. sesamoides* (2n = 32) and *S. angolense* (2n = 32) about 0.05 million years ago (Mya). Besides, *S. alatum* was clearly discriminated by forming a single clade, showing its long genetic distance and potential early speciation event in regards to the others. Comparative chloroplast genomes assessment unveiled that potential ascending dysploidy followed by hybridization contributed to the speciation in Sesamum genus. Altogether, we propose to rename *C. sesamoides* and *C. triloba* as *S. sesamoides* and *S. trilobum* respectively, as suggested earlier based on morphological description. This study provides the first insight into the phylogenetic relationships among the cultivated and wild African native relatives. The chloroplast

genomes data lay a foundation for speciation genomics in Sesamum species complex.

Introduction

Sesamum L. genus belongs to the Pedaliaceae family with around 80 species grouped in 17 genera (Cronquist, 1981). The leaves are alternate or opposite and the inflorescence appears generally as a solitary type in leaf axils with the presence of extra-floral nectaries (Bedigian, 2015). The number of species in this genus is under constant revision since the classification criteria was quite diverse according to the authors. The index Kewensis listed 34 species (Nayar and Mehra, 1970). Later on, Kobayashi (1991) reported 38 species. Through the construction of the *Sesamum* spp. descriptor, the number was revised to 20 species (IPGRI and NBPGR, 2004; Bedigian, 2015). Based on the Plants of World database, a total of 31 species have been accepted including 22 wild species native from Africa (POWO, 2022).

The wild relatives are mainly distributed across the tropical Africa (From Senegal to Somalia), central and southern Africa, and in dry-prone Indian subcontinent areas (Bedigian, 2015). Both dietary habits and traditional medicine practices are marked by the usage of cultivated and wild relatives (Ntwenya et al., 2017; Aworh, 2018; Bedigian, 2018). Among the therapeutic virtues of sesame, the cholesterol rate lowering is one of the important that was reported to prevent high blood pressure disease (Hsu and Parthasarathy, 2017). This function was imputed to the presence of the singular lignans known as sesamolin and sesamin (Visavadiya and Narasimhacharya, 2008).

The progenitor and the domestication history underpinnings the cultivated sesame has been a subject of debate. Despite the high number of wild relatives in Africa, the investigations based on the interspecific hybridization ability (Bedigian, 2014), the presence or absence of sesamolin (Bedigian et al., 1985b; Bedigian, 2003), external transcribed spacer-based phylogeny reconstruction (Gormley et al., 2015), suggested the Indian native wild *Sesamum*

malabaricum as the probable progenitor. Besides, the scientific controversy relative to the center of origin of the cultivated sesame opposed the Africa and the Indian subcontinent (Bedigian, 2003).

Moreover, several species belonging to Josephinia, Dicerocaryum, and Ceratotheca genus were reported to be closely related to Sesamum species based on phenotypic data and limited number of plastid markers (trnL-trnF and ndhF) and external transcribed spacer sequences (Gormley et al., 2015). More specifically, these genus form a species complex with Sesamum, making it difficult to clearly delineate species boundaries (Manning and Magee, 2018).

The chloroplast organelle is referred to be a chemical factory of the plant cells involved in the crucial metabolism in green plants known as photosynthesis (Kirchhoff, 2019). Due to its uniparental inheritance and non-recombination intrinsic characteristics, the chloroplast is widely used to infer the phylogenetic relationships at inter- and intra-taxon level (Biju et al., 2019; Köhler et al., 2020; Zhou et al., 2020). To the best of our knowledge, only the *S. indicum* taxon chloroplast genome have been assembled (Yi and Kim, 2012; Zhang et al., 2013a). Due to the lack of wild relative's chloroplast genome, a comprehensive study of the phylogenetic relationships between *Sesamum* and *Ceratotheca* has not yet been elucidated. Interestingly, a total of three types of chromosome number (2n = 26, 2n = 32 and 2n = 64) has been reported in the *Sesamum* genus, and only 2n = 32 for *Ceratotheca;* suggesting potential polyploidy event occurrence (Raghavan and Krishnamurthy, 1947; Kobayashi, 1991; Patil and Hiremath, 2002, 2004). Besides, the chromosome number of some Sesamum species including *S. pedaloides* is still undetermined. Therefore, complete chloroplast genome and karyotyping offer a relevant opportunity to investigate the evolutionary background leading to the diverse chromosome number set and, in the same line, the

speciation among the two sister species.

This study was designed to address two main questions: (1) what is the chloroplast genome organization variation between *Sesamum* and *Ceratotheca* representatives? (2) How phylogenetically related are the sister species regarding the diverse chromosome number sets? We took advantage of whole genome sequencing data for (i) assembling and annotating the first complete chloroplast genome of *S. alatum*, *S. pedaloides*, *S. angolense*, *S. radiatum*, *C. sesamoides*, and *C. triloba*; (ii) investigating their sequence polymorphism and divergence, and (iii) inferring the phylogenetic relationships and time divergence among the species.

Materials and Methods

Taxon Sampling and DNA Extraction

A total of six wild sesame species *v.i.z S. alatum* (2n = 26), *S. angolense* (2n = 32), *S. radiatum* (2n = 64), *S. pedaloides* (2n = indeterminate), *C. sesamoides*, and *C. triloba* were provided by the Korean genebank. The natural distribution of wild relatives is summarized in the Figure 32. Besides, the Korean elite cultivar Goenbaek (2n = 26) was also employed for *de novo* chloroplast genome assembly.

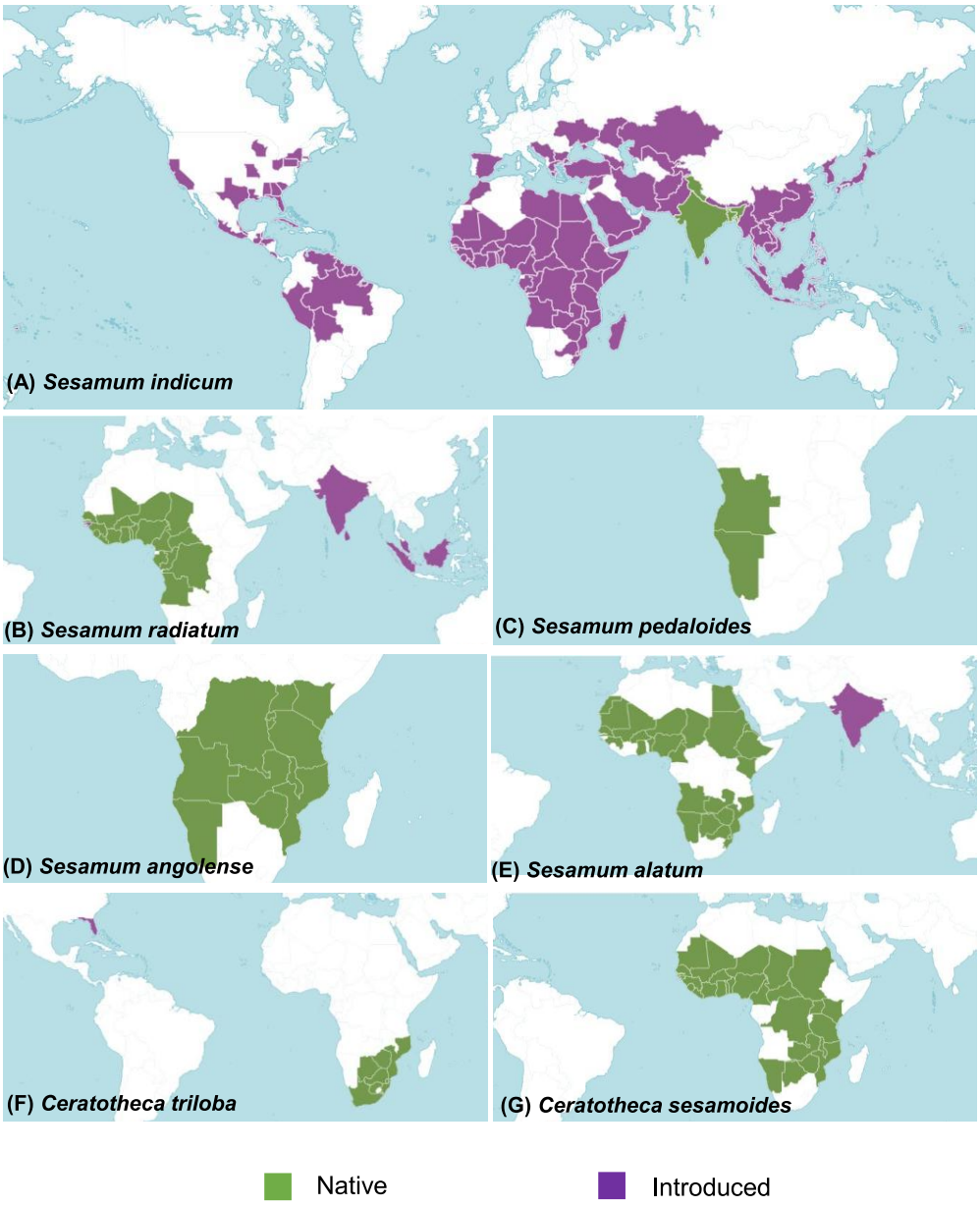

Figure 32. Distribution map showing native and introduced areas of sesame species:*Sesamum indicum* (A), *Sesamum radiatum* (B), *Sesamum pedaloides* (C), *Sesamum angolense* (D), *Sesamum alatum* (E), *Ceratotheca triloba* (F), and *Ceratotheca sesamoides* (G)

Young leaves for each species were sampled and the DNA was extracted following a modified CTAB protocol (Allen et al., 2006). Afterwards, the DNA purity was checked in 1 % agarose gel (1X TAE) and NanoDrop® ND-1000 UV-Vis spectrophotometer (Thermo Fisher Scientific, USA). The extracted DNA samples were stored at -20℃ prior to further usage.

Library preparation and Sequencing

The TruSeq DNA Nano library preparation kit (Illumina, San Diego, USA) was used to construct the library by fragmenting 1 µg high-quality genomic DNA of each samples followed by 5' and 3' adapter ligation. The NovaSeq 6000 machine (Illumina, San Diego, USA) served as platform for the short-reads sequencing.

Assembly and Annotation

For the *de novo* chloroplast assemblies, we used GetOrganelle with default parameters (Jin et al., 2020). The Rubisco subunit gene (Genbank accession: HQ384882.1) of the reference chloroplast genome (Genbank accession: NC_016433.2) from *S. indicum* cv Ansanggae was provided as seed. Besides, primers flanking the junction sites were designed using primer3 v2.3.6 (Untergasser et al., 2012) and a PCR-based validation was carried out in the conditions described as follows: The total volume of 20 µL encompassed 15 ng of DNA, 10 pmol of each primer and dried SafeDry Taq LTP-480 premix (CellSafe Co., Ltd., Gyeonggi-do, Korea). The PCRs experiments were conducted in 8 strip tubes in a C-1000 Thermal Cycler (Bio-Rad, Hercules, CA, USA). The PCR cycles were 95℃ (3 min), 35 cycles of 95℃ (30 s), 55℃ (30 s), 72℃ (30 s), followed by the extension step for 5 min at 72 ℃. The amplified products were separated in 1 % agarose gel (1X TAE) and visualized in a UVP

GelSolo M-26XV imager (Analytik Jena, CA, USA).

All chloroplast assemblies were annotated with GeSeq (Tillich et al., 2017). The setting parameters were defined as follows: HMMER profile search for coding genes and ribosomal RNA annotation, ARAGORN v1.2.38 (Laslett and Canback, 2004) and tRNAscan-SE v2.0.5 (Chan and Lowe, 2019) for transfer RNA genes detection and the *S. indicum* L. cv Ansanggae chloroplast as a reference for homology-based annotation purpose. Chloë v1.1, a stand-alone chloroplast annotator[10] served as additional third party annotator for comparison. Using the reference chloroplast annotation, the pseudo-genes as well as trans-spliced genes were manually inspected. The chloroplast genome map was rendered with OrganellarGenomeDRAW (OGDRAW) version 1.3.1(Greiner et al., 2019).

Comparative Chloroplast Genome Analysis

The annotated chloroplast genomes were compared with mVISTA webserver[11] (Frazer et al., 2004) with the cultivar Ansanggae chloroplast genome as reference. The Shuffle-LAGAN was selected as the alignment mode. In order to identify putative gene rearrangements or synteny patterns within the chloroplast genomes a whole genome alignment was executed in AliTV (Ankenbrand et al., 2017) and Mauve v.2.4.0 with the progressiveMauve algorithm option (Darling et al., 2004) respectively.

The IR/LSC and IR/SSC boundaries of the chloroplast genomes were visualized with IRscope RShiny web app[12] (Amiryousefi et al., 2018). *Arabidopsis thaliana* chloroplast

[10] https://chloe.plantenergy.edu.au/annotate.html

[11] http://genome.lbl.gov/vista/mvista/submit.shtml

[12] https://irscope.shinyapps.io/irapp/

genome (Genebank accession: NC_000932.1) was included as outgroup. The nucleotide diversity (π) among the assembled chloroplast genomes was calculated using DnaSP v.6.12.3 (Rozas et al., 2017).

Repeat Analysis

The palindrome, complement, forward and reverse sequences identification was carried out with the REPuter webserver[13] (Kurtz et al., 2001). The minimal repeat size and hamming distance were set to 30 bp, and 3 respectively. Simple Sequence Repeats (SSRs) were found out using MISA-web program (https://webblast.ipk-gatersleben.de/misa/) with default parameters (Beier et al., 2017).

Phylogenetic Inference

To infer the phylogeny between Sesamum, Ceratotheca and close members from Lamiales order, chloroplast genome datasets (*S. indicum* cv. Ansangae: NC_016433.2; *S. indicum* cv Yuzhi11: KC569603.1; *Echinacanthus attenuates*: NC_039762.1; *Echinacanthus lofouensis*: NC_035876., *Echinacanthus longipes*: NC_039761.1, *Torenia concolor*: NC_045272.1, *Torenia fournieri*: NC_056129.1, *Torenia benthamiana*: NC_045273.1) were retrieved from NCBI. *Vitis vinifera* (NC_007957.1) was used as outgroup. A total of 75 common coding protein genes (atpH , psbK , ndhI , rpl20 , matK , ndhF , atpF , psbF , petN , rps7 , psbE , petB , rpl16 , petG , petA , rps12 , rpl2 , ndhH , rps3 , rps2 , atpB , ndhE , psaA , psbT , psbM , infA , rpoB , rpl36 , psbC , ndhD , rpoC1 , atpE , accD , rps11 , rps4 , psaI , rps18 ,

[13] https://bibiserv.cebitec. uni-bielefeld.de/reputer

ndhC , psaJ , psbL , rpl33 , ndhA , petL , rpl23 , rps19 , ndhG , ndhK , rpl32 , rpl14 , psaC , psbJ , rpl22 , rps8 , ndhB , psbH , rpoC2 , ndhJ , clpP1 , rbcL , atpI , psbB , rpoA , psbD , rps16 , cemA , ccsA , pbf1 , psbZ , and rps14) served for the phylogeny inference. A multiple sequence alignment was performed with MAFFT v. 7.471-0 (Katoh and Standley, 2013). The multiple sequences alignment files were trimmed with trimAl (Capella-Gutiérrez et al., 2009) to remove poorly aligned regions. Afterwards, the maximum-likelihood tree was constructed with IQ-TREE v2.0.3 (Nguyen et al., 2015a) following the automatically selected best-fit model. The ultrafast bootstrap method (Minh et al., 2013) with 1000 iterations was applied. Bayesian inference was also employed to infer the phylogenetic relationship with MrBayes v3.2.6 (Huelsenbeck and Ronquist, 2001). Additionally, we followed the same approach for the phylogenetic tree construction based on all coding sequence genes using only the newly generated chloroplast genomes.

Divergence Time Estimation

To estimate the divergence time among the studied species, the RelTime method and the general time reversible model were performed in MEGA X following the procedure defined by Mello (2018). Based on the TimeTree database (Kumar et al., 2017), two calibration constraints were set as follows: a) Pedaliaceae versus Acanthacae: 34-70 Mya, and b) Linderniaceae versus Pedaliaceae: 41-66 Mya.

Selection Pressure Analysis

The nonsynonymous-to-synonymous substitution ratio (Ka/Ks) for each orthologous pair (with *S. indicum* as reference), was computed using the codeml package from PAML tool (Yang, 1997). Prior to the calculation, a codon-based nucleic acid alignment was obtained from the initial coding proteins multiple sequence alignment file using PAL2NAL (Suyama et al., 2006). For a reliable interpretation, the ratios with $Ks < 0.01$ or $Ks > 2$ were filtered out.

Results

General Features of the Assembled Chloroplast Genomes

The assembled chloroplast genomes resulted in a single circular quadripartite genome with typical two IRs separeted by LSC and SSC (Figure 34).

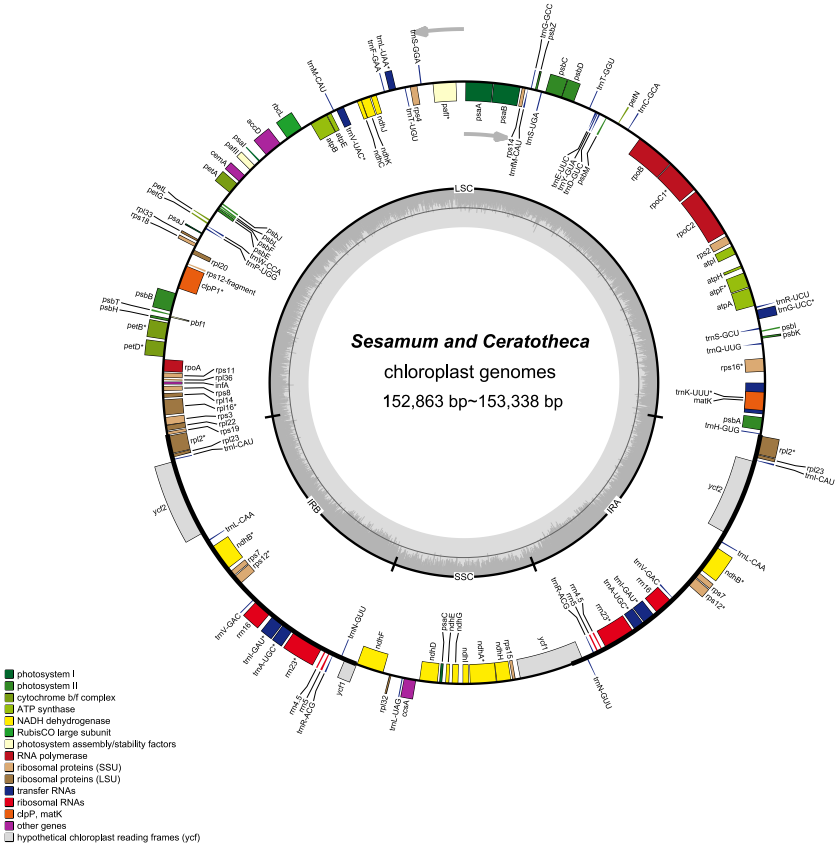

Figure 33. The chloroplast genome map of Sesamum and Ceratotheca species

The plastome sizes were 153,089 bp, 153,096 bp, 153,217 bp, 153,285 bp, 153,338 bp, 153,287 bp, and 152,863 bp for *S. angolense*, *S. alatum*, *S. pedaloides*, *S. radiatum*, *S. indicum cv* Goenbaek, *C. sesamoides*, and *C. triloba* respectively (Table 16).

Table 16. Complete chloroplast genome statistics of wild and cultivated sesame species.

	Sesamum indicum cv Goenbaek	***Sesamum alatum***	***Sesamum angolense***	***Sesamum pedaloides***	***Sesamum radiatum***	***Ceratotheca sesamoides***	***Ceratotheca triloba***
Length (bp)	153,338	153,096	153,089	153,217	153,285	153,287	152,863
GC content (%)	38.2	38.27	38.25	38.26	38.27	38.27	38.29
IR length (bp)	25,142	25,150	25,157	25,190	25,131	25,131	25,096
LSC length (bp)	85,180	85,004	85,106	85,118	85,183	85,185	84,872
SSC length (bp)	17,874	17,792	17,669	17,719	17,840	17,840	17,799
CDS count	80	80	80	80	80	80	80
tRNA count	30	30	30	30	30	30	30
rRNA count	4	4	4	4	4	4	4
Genes count	114	114	114	114	114	114	114

The seven genomes contained 114 unique genes including 80 coding-proteins, 30 transfer RNA and four ribosomal RNA genes as previously found by Yi and Kim (2012) and Zhang et al. (2013) using *S. indicum* cv Ansanggae and *S. indicum* Yuzhi 11 as plant model, respectively. Among the assembled chloroplast genomes, ten each coding sequence genes (atpF, rpoC1, rps12, petB, petD, rps16, rpl2, rpl16, ndhA, and ndhB) harbor a single intron while two genes (clpP and ycf3) contain two introns. The GC contents were similar in IR (43.39 ± 0.01 %), LSC (36.40 ± 0.03 %), SSC (32.62 ± 0.08 %), and whole chloroplast genome scale (38.26 ± 0.03 %) (Table 16).

Phylogenetic relationships between *Sesamum* and *Ceratotheca* species

To figure out the evolutionary relationship between Sesamum, Ceratotheca, and the closest related species belonging to Acanthacae and Linderniaceae, a tree (Figure 35) was constructed using 75 shared protein sequences. *Vitis vinifera* was employed as an outgroup. As expected, Sesamum and Ceratotheca representatives clustered in a single clade, resolving the sister species relationship.

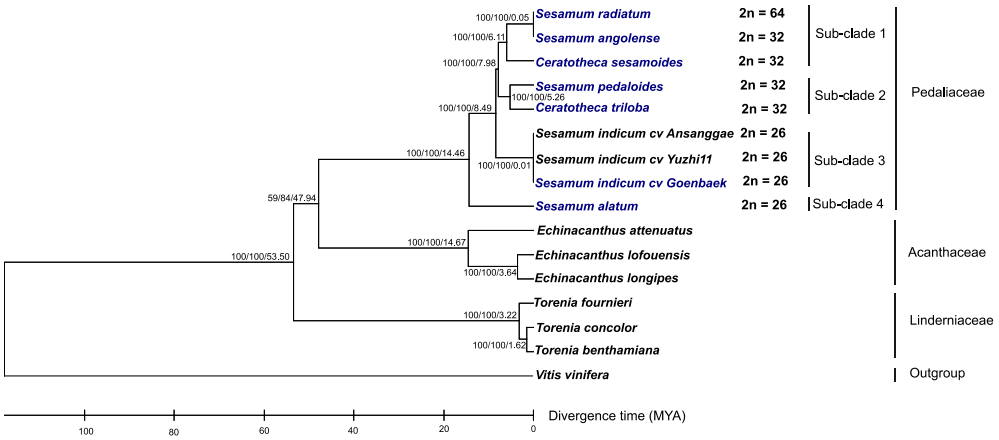

Figure 34. Phylogenetic tree depicting the evolutionary relationship between Sesamum, Ceratotheca and others Lamiales species

A close view of the Pedaliaceae clade highlighted four major sub-clades. The sub-clade 1 represents a mixture of two ploidy levels with *S. radiatum* (2n = 64), *S. angolense* (2n = 32) and *C. sesamoides* (2n = 32). Interestingly, the sub-clade 2 encompassed *S. pedaloides* (2n = 32) and *C. triloba* (2n = 32), confirming the ploidy of *S. pedaloides*. The sub-clade 3 grouped the cultivated species *S. indicum* (2n = 26) while the sub-clade 4 clearly spotted the wild relative *S. alatum* (2n = 26) as genetically distant from the cultivars and other wild relatives. For additional investigation, we constructed a tree based on 80 coding sequence genes shared by our taxa of interest (Figure 36).

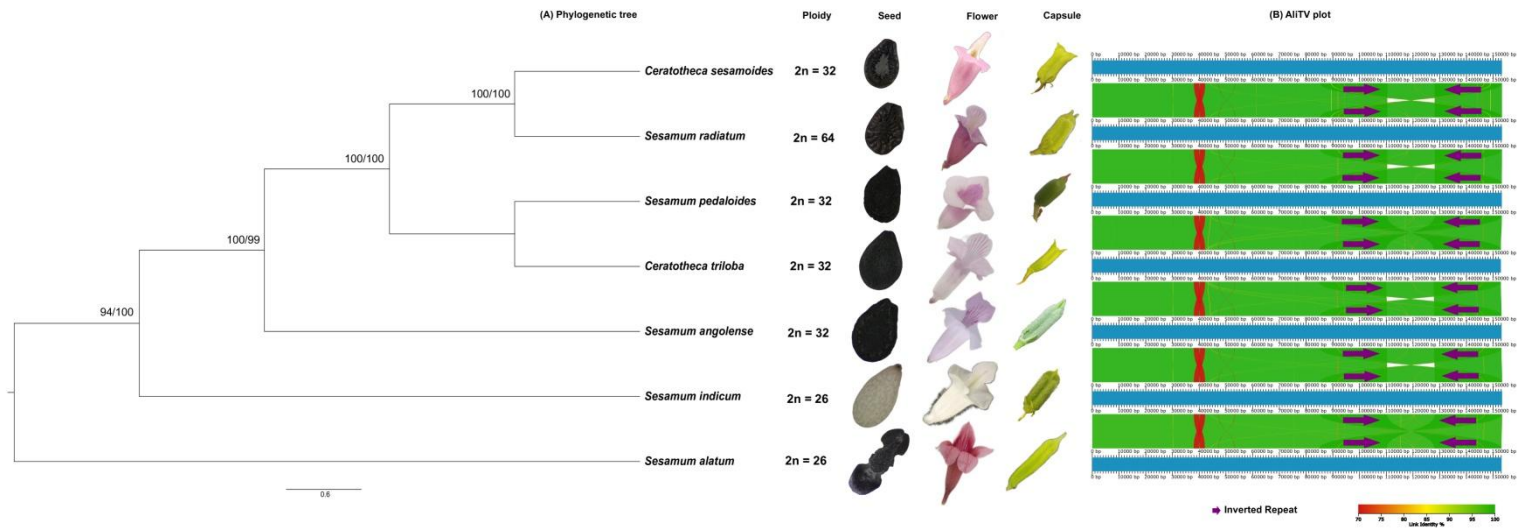

Figure 35. Phylogenetic tree, morphological variations and synthenic view between Sesamum and Ceratotheca species.

As a result, the tree topology was consistent with the previous one confirming the intricate relationship among both genera.

Divergence Time Estimation

In order to understand the speciation occurrence time among Sesamum and Ceratotheca species, a time divergence analysis was performed (Figure 35). Firstly, *S. alatum* (Sub-clade4 in the tree) exhibiting unique seed morphology with wings in our plant material (Figure 36 A), is estimated to have occurred 14.46 Million years ago (Mya). Secondly, the sub-clade 2 members (*S. pedaloides* and *C. triloba*) split concomitantly from the Sub-clade 1 and 2 at about 5.26 Mya. Thirdly, *C. sesamoides* (member of the sub-clade1) was inferred to have occurred 6.11 Mya, a little bit earlier than *S. pedaloides* and *C. triloba*.

As expected, the cultivated type (*S. indicum*) have been occurred lastly at 0.01 Mya. Interestingly, the time divergence inference revealed that *S. radiatum* (2n = 64) and *C. sesamoides* (2n = 32) have concomitantly occurred recently at about 0.05 Mya. Altogether, the time divergence dating indicates recent hybridization between *C. sesamoides* (2n = 32) and *S. angolense* (2n = 32) leading to the new species *S. radiatum* (2n = 64).

Comparative plastome analysis

Although the morphological features of the species are distinct, the chloroplast genome structure is highly conserved (Figure 36B). However, IR contraction and expansion has been revealed with length ranging from 25,096 bp to 25,190 bp while LSC varied from 84,872 bp to 85,185 bp. SSC sized from 17,719 bp to 17,874 bp.

By comparing the chloroplast genomes boundaries of Sesamum and Ceratotheca species, we noted that the IRb/LSC junction is sited between rpl2 and rsp19 genes (Figure 37). The pseudogene ycf1 is located exclusively in IRb for *S. indicum cv* Ansanggae, *S. indicum cv*

Yuzhii11 and at the border IRb/SSC the other species. The ndhF gene is mainly in the SSC region for all species excepted for *S. angolense*. The ycf1 gene of all species was located at the SSC/IRa junctions with a gene size ranging from 5312 to 5369 bp. The trnH genes were localized in the LSC region, 1-16 bp away from the the IRa-LSC border. Overall, the chloroplast genome structure at different junctions was highly conserved among *Sesamum* and *Ceratotheca* species.

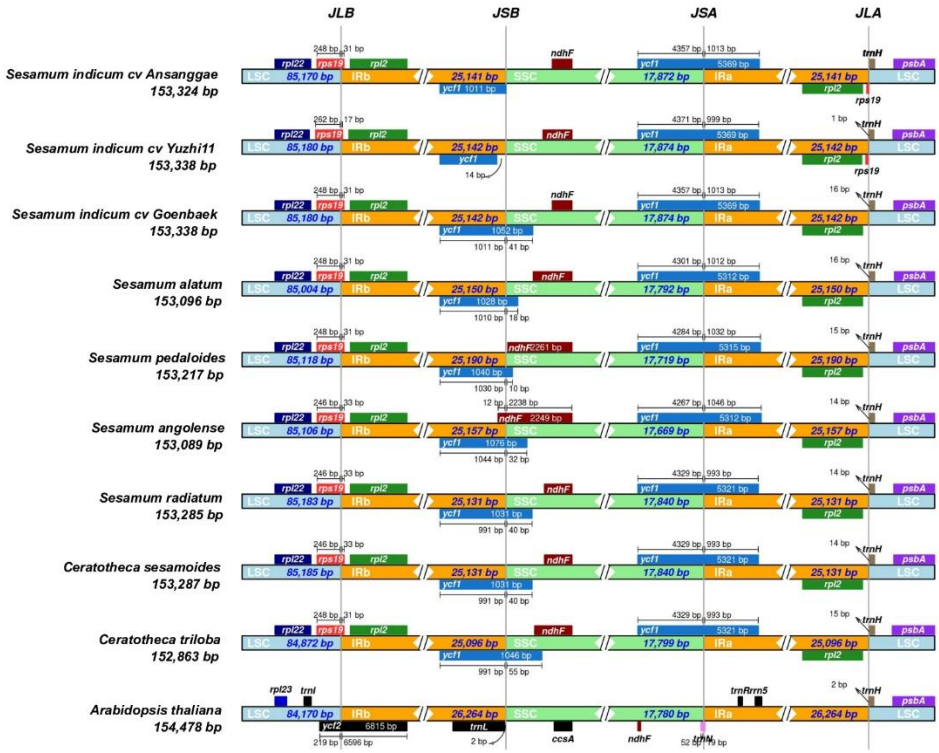

Figure 36. Chloroplast junction's sites view showing genes distribution alongside the boundaries and IR contraction and expansion within the Sesamum and Ceratotheca species. *Arabidopsis thaliana* was set as outgroup.

Variation Hotspots within Chloroplast Genomes of *Sesamum* and *Ceratotheca* species

Despite the high-colinearity of the chloroplast genomes within *Sesamum* and *Ceratotheca* species, substantial variations were noted mainly in SSC regions (Figure 38).

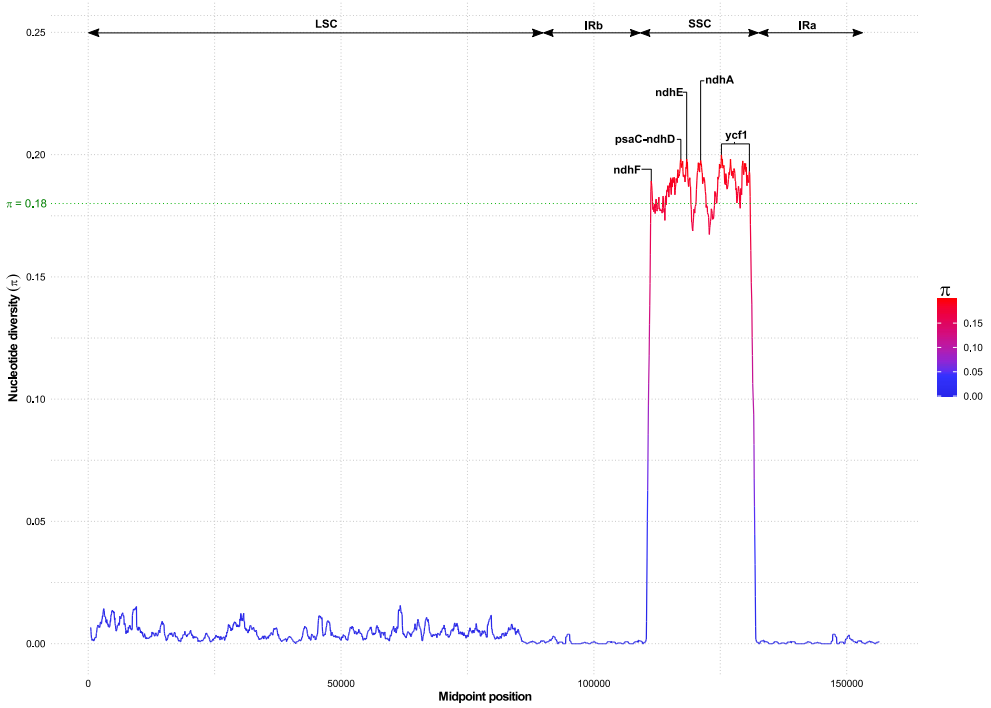

Figure 37. Nucleotide diversity variation with chloroplast genomes of Sesamum and Ceratotheca species. Highest values above 0.18 (in green) indicate candidate genes for population genetics purpose.

The nucleotide diversity calculation revealed a peak value located in the ycf1 followed by ndhA, ndhE, psaC-ndh-D and ndhF regions. To estimate their discriminatory power, we inferred the phylogenetic tree using each gene. As a result, ycf1, psaC-ndhD, and ndhF sequences clearly distinguished the taxa (Figure 39) as depicted earlier with a set of 75 (Figure 35) and 80 (Figure 36A) coding genes, respectively.

Therefore, they could be used as a marker to delineate *Sesamum* and *Ceratotheca* species since several species are not yet well characterized at both morphologic and cytogenetic levels. Besides, *C. sesamoides* and *S. radiatum* exhibited a consistent monophyletic pattern for all divergent genes except *ndhE*, confirming the close evolutionary relationship between these two specific species.

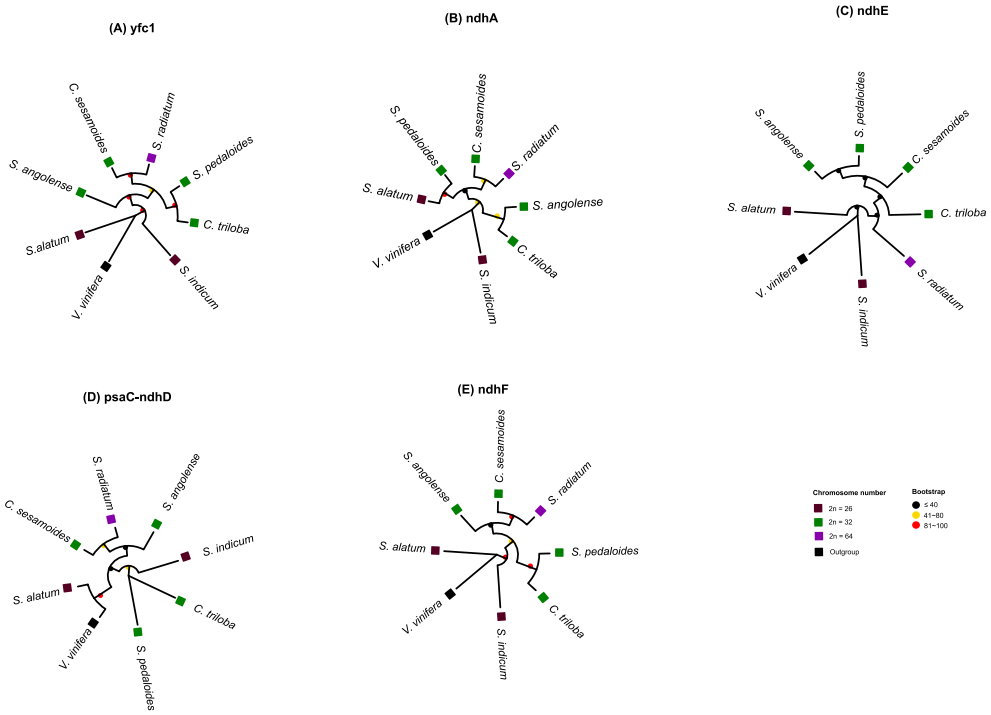

Figure 38. Evaluation of the discriminatory power of the candidate regions with the high nucleotide diversity index. The tree topology of ycf1 (A), psaC-ndhD (D), and ndhF (E) was consistent with the maximum likehood phylogenetic tree based on the 75 coding-genes. Discrepancies were noted for ndhA- and ndhE-based tree inference. *Vitis vinifera* was set as an outgroup. Squares and circles were colored following the ploidy and the bootstrap values respectively.

Codon Usage Analysis

Codon usage bias was examined by computing the Relative Synonymous Codon Usage (RSCU) (Sharp and Cowe, 1991). RSCU represents the observed frequency of a codon divided by the expected frequency. A lack of bias referred to the codons with RSCU values close to 1. Globally, slight variation of RSCU was found within Sesamum and Ceratotheca species (Figure 40) values among Sesamum and Ceratotheca species.). A total of 27 core codons exhibited RSCU > 1, of which 24 were adenine/thymine-ending codons, one guanine-ending codon and two cytosine-ending codons. In contrast, guanine- or cytosine-ending codons mostly exhibited RSCU < 1. The most biased codon was found for the stop codon TAA (RSCU = 1.55 +/- 0.01) while the less biased was detected for the stop codon TAG (RSCU = 0.69 +/- 0.01). Similar trend of A-T biased codon usage was observed for others plant species (Wang et al., 2017; Biju et al., 2019).

Interestingly, by examining the species cluster from the heat map, the phylogenetic tree topology is concordant with the coding sequence based tree inference; indicating a robust estimation of phylogenetic relationship based on codon usage as observed in a wide range of families (Gao et al., 2019; Chi et al., 2020; Wu et al., 2021).

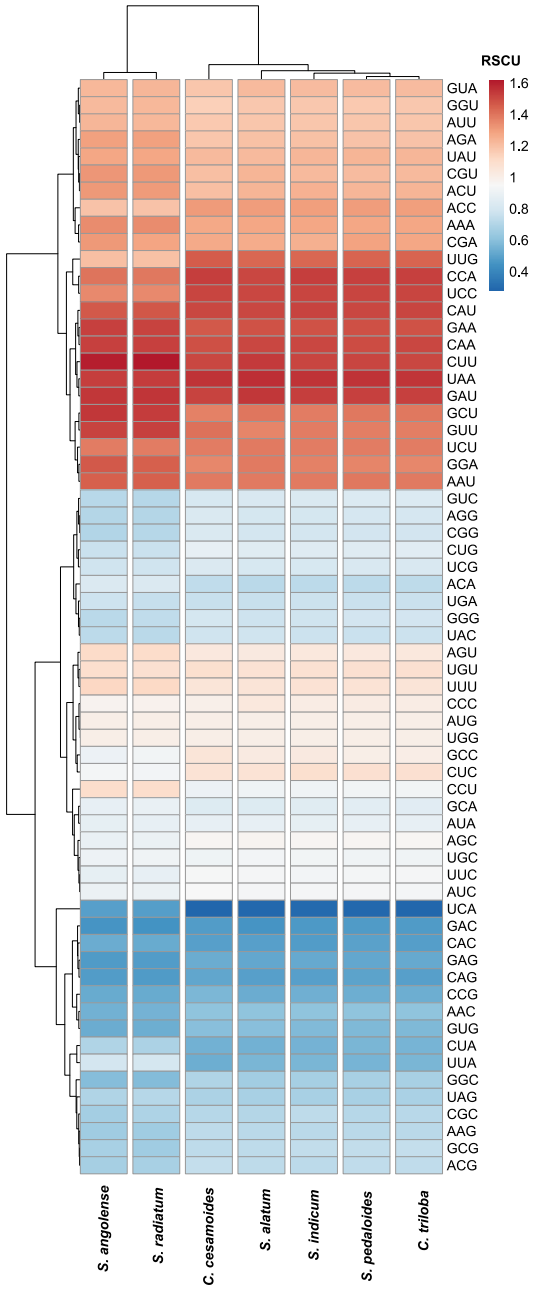

Figure 39. Heat map of relative synonymous codon usage (RSCU) values among Sesamum and Ceratotheca species.

Repeats analysis

Long repeats constitute a driving force for chloroplast genome rearrangement and has been used for phylogenetic inference between species (Park et al., 2017; Luo et al., 2021). It induces genetic diversity by promoting intermolecular recombination in the chloroplast genome (Park et al., 2018). Long repeats encompass forward, reverse, palindrome, and complement types. In the present study, the mean count of long repeats was 25.71 +/- 3.24 bp. The number of long repeats ranged from 21 to 31, among which palindromic (12-18) and forward (9-13) types are the most abundant. Besides, the size of the repeats was mainly within the range of 30–39 bp. Only *S. indicum* exhibited repeats in the range 60-69 bp (Table 17).

Table 17. Microsatellites identification report in the assembled chloroplast genomes

Species	ID	Region	Type	Motif	Repeat	Start	End	Feature	Gene name
Sesamum indicum	SIC1	LSC	Mono	A	10	239	248	Intergenic	-
Sesamum indicum	SIC2	LSC	Mono	A	10	4381	4390	Intergenic	-
Sesamum indicum	SIC3	LSC	Mono	A	11	8578	8588	Intergenic	-
Sesamum indicum	SIC4	LSC	Mono	T	11	13441	13451	Intergenic	-
Sesamum indicum	SIC5	LSC	Mono	T	10	16684	16693	Intergenic	-
Sesamum indicum	SIC6	LSC	Mono	T	11	18898	18908	CDS	rpoC2
Sesamum indicum	SIC7	LSC	Mono	T	10	29739	29748	Intergenic	-
Sesamum indicum	SIC8	LSC	Mono	C	11	36703	36713	Intergenic	-
Sesamum indicum	SIC9	LSC	Di	AT	6	42988	42999	Intergenic	-
Sesamum indicum	SIC10	LSC	Mono	T	11	43919	43929	Intergenic	-
Sesamum indicum	SIC11	LSC	Mono	T	10	44803	44812	Intergenic	-
Sesamum indicum	SIC12	LSC	Mono	T	10	46896	46905	Intergenic	-
Sesamum indicum	SIC13	LSC	Di	TA	6	47045	47056	Intergenic	-
Sesamum indicum	SIC14	LSC	Mono	A	10	47691	47700	Intergenic	-
Sesamum indicum	SIC15	LSC	Mono	T	10	49162	49171	Intergenic	-
Sesamum indicum	SIC16	LSC	Mono	C	10	51909	51918	Intergenic	-
Sesamum indicum	SIC17	LSC	Mono	T	11	55422	55432	CDS	atpB
Sesamum indicum	SIC18	LSC	Mono	T	10	59980	59989	Intergenic	-

Table 17. *Continued*

Sesamum indicum	SIC19	LSC	Mono	A	10	61792	61801	Intergenic	-
Sesamum indicum	SIC20	LSC	Mono	C	10	66533	66542	Intergenic	-
Sesamum indicum	SIC21	LSC	Mono	T	10	71254	71263	Intergenic	-
Sesamum indicum	SIC22	LSC	Mono	A	10	72472	72481	Intergenic	-
Sesamum indicum	SIC23	LSC	Mono	T	11	81169	81179	Intergenic	-
Sesamum indicum	SIC24	LSC	Mono	T	10	82872	82881	Intergenic	-
Sesamum indicum	SIC25	LSC	Mono	A	10	82938	82947	Intergenic	-
Sesamum indicum	SIC26	IR	Mono	T	10	102985	102994	Intron	trnA-UGC
Sesamum indicum	SIC27	SSC	Mono	T	10	112964	112973	Intergenic	-
Sesamum indicum	SIC28	SSC	Mono	T	11	113885	113895	Intergenic	-
Sesamum indicum	SIC29	SSC	Mono	A	10	117212	117221	Intergenic	-
Sesamum indicum	SIC30	SSC	Mono	A	10	121280	121289	Intergenic	-
Sesamum indicum	SIC31	SSC	Mono	T	10	126406	126415	CDS	ycf1
Sesamum indicum	SIC32	IR	Mono	A	10	135525	135534	Intron	trnA-UGC
Sesamum alatum	SA1	LSC	Mono	T	10	240	249	Intergenic	-
Sesamum alatum	SA2	LSC	Mono	A	10	16670	16679	Intergenic	-
Sesamum alatum	SA3	LSC	Mono	T	11	18878	18888	CDS	rpoC2
Sesamum alatum	SA4	LSC	Mono	T	11	35800	35810	Intergenic	-
Sesamum alatum	SA5	LSC	Mono	A	11	36714	36724	Intergenic	-
Sesamum alatum	SA6	LSC	Di	AT	6	42901	42912	Intergenic	-
Sesamum alatum	SA7	LSC	Mono	A	10	45524	45533	Intergenic	-
Sesamum alatum	SA8	LSC	Mono	A	14	47567	47580	Intergenic	-
Sesamum alatum	SA9	LSC	Mono	T	11	55273	55283	CDS	atpB
Sesamum alatum	SA10	LSC	Mono	T	10	64780	64789	Intergenic	-
Sesamum alatum	SA11	LSC	Tri	ATA	5	67857	67871	Intergenic	-
Sesamum alatum	SA12	LSC	Mono	T	10	71647	71656	Intergenic	-
Sesamum alatum	SA13	LSC	Mono	A	15	75104	75118	Intergenic	-
Sesamum alatum	SA14	LSC	Mono	A	11	76127	76137	Intergenic	-
Sesamum alatum	SA15	LSC	Mono	T	11	80996	81006	Intergenic	-
Sesamum alatum	SA16	LSC	Mono	A	10	82762	82771	Intergenic	-
Sesamum alatum	SA17	IR	Mono	T	10	99060	99069	Intergenic	-
Sesamum alatum	SA18	IR	Mono	G	11	104034	104044	intron	rrn23
Sesamum alatum	SA19	SSC	Mono	A	12	113307	113318	Intergenic	-
Sesamum alatum	SA20	SSC	Mono	T	11	113685	113695	Intergenic	-

Table 17. *Continued*

Sesamum alatum	SA21	SSC	Tri	TAA	5	123313	123327	Intergenic	-
Sesamum alatum	SA22	SSC	Mono	T	10	126194	126203	CDS	ycf1
Sesamum alatum	SA23	IR	Mono	C	11	134057	134067	intron	rrn23
Sesamum alatum	SA24	IR	Mono	A	10	139032	139041	Intergenic	-
Sesamum angolense	SAG1	LSC	Mono	A	13	237	249	Intergenic	-
Sesamum angolense	SAG2	LSC	Mono	A	10	4838	4847	Intergenic	-
Sesamum angolense	SAG3	LSC	Mono	T	10	8357	8366	Intergenic	-
Sesamum angolense	SAG4	LSC	Mono	A	12	8593	8604	Intergenic	-
Sesamum angolense	SAG5	LSC	Mono	T	10	13439	13448	Intergenic	-
Sesamum angolense	SAG6	LSC	Mono	A	11	16657	16667	Intergenic	-
Sesamum angolense	SAG7	LSC	Mono	T	11	18862	18872	CDS	rpoC2
Sesamum angolense	SAG8	LSC	Mono	T	10	29707	29716	Intergenic	-
Sesamum angolense	SAG9	LSC	Mono	T	16	46856	46871	Intergenic	-
Sesamum angolense	SAG10	LSC	Mono	A	10	47640	47649	Intergenic	-
Sesamum angolense	SAG11	LSC	Mono	T	10	55359	55368	CDS	atpB
Sesamum angolense	SAG12	LSC	Mono	T	10	59916	59925	Intergenic	-
Sesamum angolense	SAG13	LSC	Mono	T	10	61854	61863	Intergenic	-
Sesamum angolense	SAG14	LSC	Mono	T	11	81098	81108	Intergenic	-
Sesamum angolense	SAG15	LSC	Mono	A	12	82866	82877	Intergenic	-
Sesamum angolense	SAG16	IR	Mono	T	10	85161	85170	Intergenic	-
Sesamum angolense	SAG17	IR	Mono	T	10	102893	102902	intron	trnA-UGC
Sesamum angolense	SAG18	SSC	Mono	A	10	111964	111973	CDS	ycf1
Sesamum angolense	SAG19	SSC	Mono	T	10	124828	124837	Intergenic	-
Sesamum angolense	SAG20	IR	Mono	A	10	135294	135303	intron	trnA-UGC
Sesamum angolense	SAG21	IR	Mono	A	10	153026	153035	Intergenic	-
Sesamum pedaloides	SP1	LSC	Mono	A	10	8567	8576	Intergenic	-
Sesamum pedaloides	SP2	LSC	Mono	A	10	16701	16710	Intergenic	-
Sesamum pedaloides	SP3	LSC	Mono	T	11	18905	18915	CDS	rpoC2
Sesamum pedaloides	SP4	LSC	Mono	T	12	29746	29757	Intergenic	-
Sesamum pedaloides	SP5	LSC	Mono	T	10	43934	43943	Intergenic	-
Sesamum pedaloides	SP6	LSC	Mono	T	10	46917	46926	Intergenic	-
Sesamum pedaloides	SP7	LSC	Mono	A	11	47709	47719	Intergenic	-
Sesamum pedaloides	SP8	LSC	Mono	A	10	51691	51700	Intergenic	-
Sesamum pedaloides	SP9	LSC	Mono	T	10	55429	55438	CDS	atpB

Table 17. *Continued*

Sesamum pedaloides	SP10	LSC	Mono	T	10	59988	59997	Intergenic	-
Sesamum pedaloides	SP11	LSC	Mono	T	10	64548	64557	Intergenic	-
Sesamum pedaloides	SP12	LSC	Mono	T	10	70679	70688	Intergenic	-
Sesamum pedaloides	SP13	LSC	Mono	T	11	81117	81127	Intergenic	-
Sesamum pedaloides	SP14	IR	Mono	T	11	85171	85181	Intergenic	-
Sesamum pedaloides	SP15	IR	Mono	T	10	102952	102961	intron	trnA-UGC
Sesamum pedaloides	SP16	SSC	Mono	A	10	112868	112877	Intergenic	-
Sesamum pedaloides	SP17	SSC	Mono	A	11	113433	113443	Intergenic	-
Sesamum pedaloides	SP18	SSC	Mono	T	11	113810	113820	Intergenic	-
Sesamum pedaloides	SP19	SSC	Mono	T	10	126310	126319	CDS	ycf1
Sesamum pedaloides	SP20	IR	Mono	A	10	135375	135384	intron	trnA-UGC
Sesamum pedaloides	SP21	IR	Mono	A	11	153155	153165	Intergenic	-
Sesamum radiatum	SR1	LSC	Mono	A	10	237	246	Intergenic	-
Sesamum radiatum	SR2	LSC	Mono	G	11	5038	5048	Intergenic	-
Sesamum radiatum	SR3	LSC	Mono	A	10	8577	8586	Intergenic	-
Sesamum radiatum	SR4	LSC	Mono	T	13	13444	13456	Intergenic	-
Sesamum radiatum	SR5	LSC	Mono	A	11	16698	16708	Intergenic	-
Sesamum radiatum	SR6	LSC	Mono	T	11	18903	18913	CDS	rpoC2
Sesamum radiatum	SR7	LSC	Mono	T	11	29735	29745	Intergenic	-
Sesamum radiatum	SR8	LSC	Mono	A	11	36724	36734	Intergenic	-
Sesamum radiatum	SR9	LSC	Mono	A	11	36829	36839	Intergenic	-
Sesamum radiatum	SR10	LSC	Mono	A	10	45677	45686	Intergenic	-
Sesamum radiatum	SR11	LSC	Mono	T	11	46929	46939	Intergenic	-
Sesamum radiatum	SR12	LSC	Di	TA	6	47079	47090	Intergenic	-
Sesamum radiatum	SR13	LSC	Mono	A	10	47724	47733	Intergenic	-
Sesamum radiatum	SR14	LSC	Mono	C	10	51908	51917	Intergenic	-
Sesamum radiatum	SR15	LSC	Mono	T	10	55422	55431	CDS	atpB
Sesamum radiatum	SR16	LSC	Mono	T	10	61923	61932	Intergenic	-
Sesamum radiatum	SR17	LSC	Mono	T	10	64604	64613	Intergenic	-
Sesamum radiatum	SR18	LSC	Mono	T	10	66690	66699	Intergenic	-
Sesamum radiatum	SR19	LSC	Mono	A	10	68074	68083	Intergenic	-
Sesamum radiatum	SR20	LSC	Mono	T	10	71244	71253	Intergenic	-
Sesamum radiatum	SR21	LSC	Mono	T	10	71827	71836	Intergenic	-
Sesamum radiatum	SR22	LSC	Mono	A	10	72469	72478	Intergenic	-
Sesamum radiatum	SR23	LSC	Mono	T	11	81178	81188	Intergenic	-

Table 17. *Continued*

Sesamum radiatum	SR24	IR	Mono	T	10	102997	103006	intron	trnA-UGC
Sesamum radiatum	SR25	SSC	Mono	A	10	112056	112065	CDS	ycf1
Sesamum radiatum	SR26	SSC	Mono	T	10	117192	117201	Intergenic	-
Sesamum radiatum	SR27	SSC	Mono	T	10	121259	121268	Intergenic	-
Sesamum radiatum	SR28	IR	Mono	A	10	135463	135472	intron	trnA-UGC
Ceratotheca sesamoides	CS1	LSC	Mono	A	10	237	246	Intergenic	-
Ceratotheca sesamoides	CS2	LSC	Mono	G	12	5038	5049	Intergenic	-
Ceratotheca sesamoides	CS3	LSC	Mono	A	10	8578	8587	Intergenic	-
Ceratotheca sesamoides	CS4	LSC	Mono	T	13	13445	13457	Intergenic	-
Ceratotheca sesamoides	CS5	LSC	Mono	A	11	16699	16709	Intergenic	-
Ceratotheca sesamoides	CS6	LSC	Mono	T	11	18904	18914	CDS	rpoC2
Ceratotheca sesamoides	CS7	LSC	Mono	T	12	29736	29747	Intergenic	-
Ceratotheca sesamoides	CS8	LSC	Mono	C	10	36717	36726	Intergenic	-
Ceratotheca sesamoides	CS9	LSC	Mono	A	10	36727	36736	Intergenic	-
Ceratotheca sesamoides	CS10	LSC	Mono	A	11	36831	36841	Intergenic	-
Ceratotheca sesamoides	CS11	LSC	Mono	A	10	45679	45688	Intergenic	-
Ceratotheca sesamoides	CS12	LSC	Mono	T	10	46931	46940	Intergenic	-
Ceratotheca sesamoides	CS13	LSC	Di	TA	6	47080	47091	Intergenic	-
Ceratotheca sesamoides	CS14	LSC	Mono	A	10	47725	47734	Intergenic	-
Ceratotheca sesamoides	CS15	LSC	Mono	T	10	51749	51758	Intergenic	-
Ceratotheca sesamoides	CS16	LSC	Mono	C	10	51910	51919	Intergenic	-
Ceratotheca sesamoides	CS17	LSC	Mono	T	10	55424	55433	CDS	atpB
Ceratotheca sesamoides	CS18	LSC	Mono	T	10	61925	61934	Intergenic	-
Ceratotheca sesamoides	CS19	LSC	Mono	T	10	64606	64615	Intergenic	-
Ceratotheca sesamoides	CS20	LSC	Mono	T	10	66692	66701	Intergenic	-
Ceratotheca sesamoides	CS21	LSC	Mono	A	10	68076	68085	Intergenic	-
Ceratotheca sesamoides	CS22	LSC	Mono	T	10	71246	71255	Intergenic	-
Ceratotheca sesamoides	CS23	LSC	Mono	T	10	71829	71838	Intergenic	-
Ceratotheca sesamoides	CS24	LSC	Mono	A	10	72471	72480	Intergenic	-
Ceratotheca sesamoides	CS25	LSC	Mono	T	11	81180	81190	Intergenic	-
Ceratotheca sesamoides	CS26	IR	Mono	T	10	102999	103008	intron	trnA-UGC
Ceratotheca sesamoides	CS27	SSC	Mono	A	10	117203	117212	Intergenic	-
Ceratotheca sesamoides	CS28	SSC	Mono	A	10	121270	121279	Intergenic	-
Ceratotheca sesamoides	CS29	SSC	Mono	T	10	126406	126415	CDS	ycf1

Table 17. *Continued*

Ceratotheca sesamoides	CS30	IR	Mono	A	10	135465	135474	intron	trnA-UGC
Ceratotheca triloba	CT1	LSC	Mono	A	10	219	228	Intergenic	-
Ceratotheca triloba	CT2	LSC	Mono	T	10	16658	16667	Intergenic	-
Ceratotheca triloba	CT3	LSC	Mono	A	11	16669	16679	Intergenic	-
Ceratotheca triloba	CT4	LSC	Mono	T	11	18874	18884	CDS	rpoC2
Ceratotheca triloba	CT5	LSC	Mono	A	10	30889	30898	Intergenic	-
Ceratotheca triloba	CT6	LSC	Mono	T	10	46874	46883	Intergenic	-

Microsatellites are referred to short tandem repeat sequences of one to six nucleotide repeats (Fan and Chu, 2007). SSRs are widely present in chloroplast genome and have been extensively used as molecular markers for population genetics, phylogenetic relationships inference, and species identification (Powell et al., 1995; Huang et al., 2018; Lee et al., 2019; Li et al., 2020a). We counted 21 to 32 chloroplast SSR within the assembled chloroplast genomes (Table 17), of which most are monomeric (> 87%). The majority of SSRs are located in LSC sequence and mainly in intergenic regions (>76%) for all species. The most dominant motif (A) count ranged from 19 to 27 and spanning 208-277 bp. Trinucleotide repeats (AAT) were only detected for *S. alatum* occupying 30 bp of the chloroplast genome length (Table 17).

Selection pressure analysis

The pairwise ratio of non-synonymous substitutions (Ka) to the rate of synonymous substitutions (Ks) analysis is presented in the Figure 41. Ka/Ks ratios with Ks < 0.1 or K > 0.2 were changed into zero for a reliable estimation of the selection pressure. The results revealed that the NAD(P)H-quinone oxidoreductase subunit I (ndhI) has undergone strong positive selection in *S. angolense* and *S. radiatum*. Similar trends were observed for the photosystem I gene ycf4 mainly in *S. radiatum*, *S. alatum* and *C. sesamoides*. The rpl20 gene also exhibited positive selection only in *S. alatum*. However, the *matK*, *ndhF*, and *ycf1* Ka/ks values were around 1, implying neutral selection pressure.

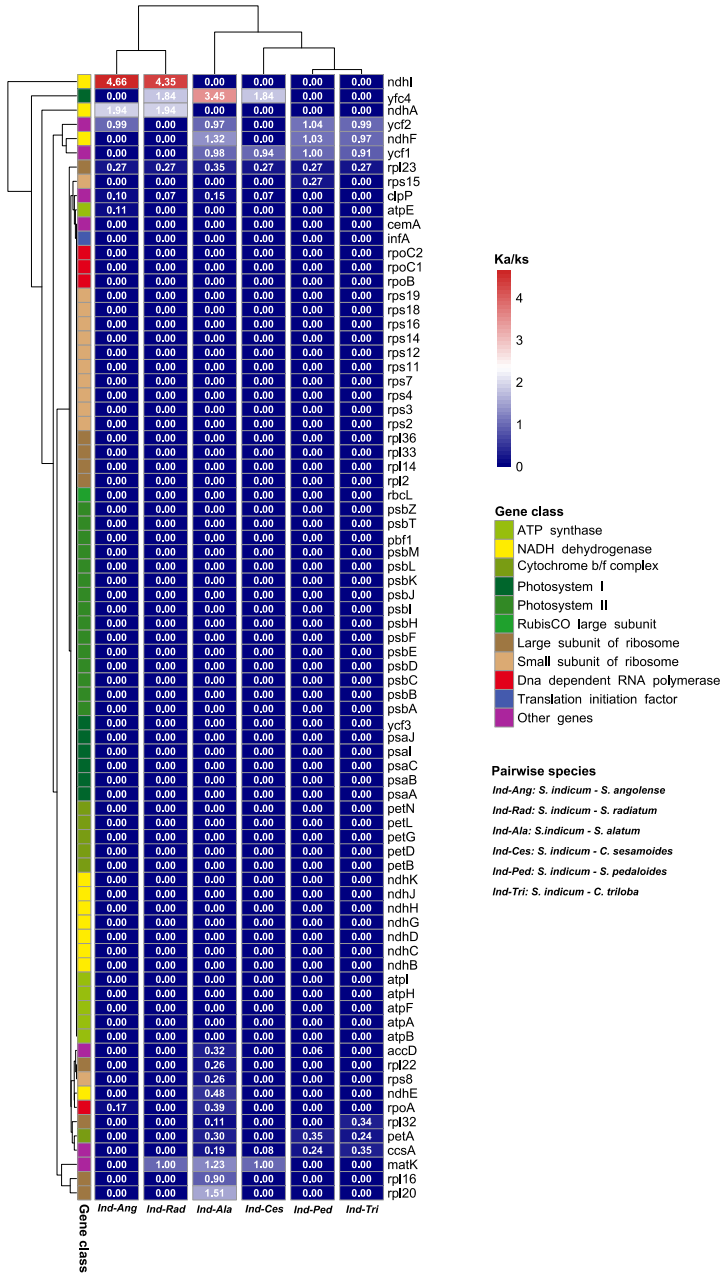

Figure 40. Heat map of pairwise Ka/ks values among wild relatives and the cultivated type *Sesamum indicum*. Non-applicable (with Ks = 0) Ka/Ks and ratio with Ks < 0.1 and Ks> 2 were changed into zero.

Discussion

Plastome evolution between Sesamum and Ceratotheca genus

We reported for the first time, whole chloroplast genomes of six African native wild sesame species including *S. alatum*, *S. angolense*, *S. pedaloides*, *S. radiatum*, *C. sesamoides*, and *C. triloba*. The information from the generated plastome sequences served for comparative analysis. A typical quadripartite chloroplast structure including LSC, SSC and two IR was observed. In-depth comparative analysis revealed contraction and expansion events within all species. IR expansion and contraction is a common phenomenon observed in land plants resulting in the variation of chloroplast length at both intra- and inter- species level (Asaf et al., 2020; Guo et al., 2021). As expected, the chloroplast genome was highly conserved among Sesamum and Ceratotheca despite the morphological differences. The conserved structure is consistent with the two previously published *S. indicum* chloroplast genomes (Yi and Kim, 2012; Zhang et al., 2013a).

The variation of microsatellites copy numbers in chloroplast genome is helpful for population genetics, and polymorphism assessment. Single-nucleotide repeats SSR datasets provided in the current study constitute a useful resource for further population polymorphism study within Sesamum, Ceratotheca and potentially close relative species in Pedaliaceae. Palindromic repeats are known to constitute mutational hotspots contributing to plastome expansion (Smith, 2020). By mining the chloroplast genomes, we detected that palindromic repeats are prominent in all chloroplast genomes. Therefore, they represent a suitable resource for markers development in regard to genetic diversity investigation in Sesamum species continuum.

Chloroplast genes generally evolved under purifying selection mainly to maintain functional continuity of the genes over a long period of time (Matsuoka et al., 2002; Jiang et al., 2018). However, previous comparative plastome studies identified some genes that underwent positive selection including the photosynthetic genes *rbcL* (Ivanova et al., 2017) and *ycf2* (Jiang et al., 2018) among others.

In our study, mainly photosynthesis related genes including *ndhA*, *ndhI*, and *ycf4* exhibited strong positive selection. Subsequently, owing the specific distribution patterns of the studies samples in tropical Africa (See Figure 32) and the drought-prone habitat preference in nature, we postulate that the selection of this category of genes might related to adaptation to environment changes including photosynthetic rate, drought, temperature, carbon dioxide level, or ecological niche (Piot et al., 2018). Moreover, the capability of photosynthetic-oriented genes selection may contribute the drought tolerance strength of the cultivated sesame *S. indicum* as revealed by previous extensive functional genomics studies (Dossa et al., 2019; Yu et al., 2019b).

Our data provided a high-resolution view for delineation of Pedaliaceae species in the sesame speciation continuum

Delineating species is a challenging aspect in taxonomy since the methodology is quite heterogeneous depending on not only of the taxa but also the scientists. Pedaliaceae s.l. family encompasses several tribes including Sesamothamneae Ihlenf., Sesamaea (Endl.) Meisn., and Pedaliae Dumort. Using plastid and nuclear markers, Gormley et al. (2015) provided the evidence of the monophyletic pattern of these tribes. However, the authors highlighted that Sesamum is paraphyletic in regards with Ceratotheca, Josephinia, and

Dycerocaryum genus. This latter observation is in contrast with our results that showed that Sesamum and Ceratotheca formed a complex. The low number of markers and the used taxa in the previous study might explain the divergence of the tree topology. In fact, there were the absence of 2n = 64 chromosome set representatives. Therefore, our study provided the first insight regarding the chromosome number variation criterion.

Chloroplast genome is generally well conserved in land plants (Trösch et al., 2018). Despite the highly conserved chloroplast genome arrangement within both genus, remarkable sequence divergence was noted specifically in *ycf1* gene. This gene previously demonstrated its power to discriminate species at both low and large scale (Handy et al., 2011; Dong et al., 2015). The *ycf1*-based tree topology was consistent with the protein coding genes tree, confirming that *ycf1* is a powerful gene with promising potential for DNA barcoding purpose.

RNA editing sites is referred to be a driving force leading to phenotypic plasticity, genetic diversity in plants as well as in mammals (Gommans et al., 2009). A comprehensive classification based on the occurrence of RNA editing sites in chloroplast genomes was consistent with the inferred tree, validating our tree topology.

Interestingly, a relatively long branch of *S. alatum* was observed implying its long evolutionary occurrence compared to the other ones; which is consistent with tree topology from Gormley et al. (2015), biogeographic and morphological data (Bedigian, 2018). It is noteworthy that *S. alatum* seed exhibits a singular characteristic with winged-seeds; which is absent for the other ones. The presence of wings is one of the key ecological adaptive trait for ancient wild crop that ensure seed dispersal by wind (Willson and Traveset, 2000). To the best of our knowledge, this trait seems to be lost during the evolution in Sesamum genus since no other member (described so far) of the genus harbors it. Moreover, the time

divergence estimation revealed that *S. altatum* diverged earlier (14.46 MYA) from others species, supporting the previous findings.

Ascending Disploidy and Interspecific hybridization drive speciation in Sesamum complex

In Sesamum genus, the chromosome number variation from the diploid 2n = 26 to 2n = 32 is still a mystery. While the ancient species in the present study (*S. alatum*) and the cultivated type (*S. indicum*) are also diploid with 2 = 26 chromosomes, the group of 2n = 32 might have undergone chromosomal rearrangements, altering their karyotype. We speculate here that the 2n = 32 members including *S. pedaloides*, *S. angolense*, *C. sesamoides*, and *C. triloba* are the result of an ascending dysploidy event leading to their chromosome number change (Figure 42). Chromosome fission, fusion, or acquisition can leads to the increase of ploidy as termed as ascending dysploidy or, in contrast, descending dysploidy (Udall et al., 2019; Mayrose and Lysak, 2021). As a consequence, this evolutionary event might be one of the source of speciation leading to the sister species *C. sesamoides* and *C. triloba* also known as *S. sesamoides* and *S. triolobum* (POWO, 2022). Similar ascending disploidy events have been also observed in Marantaceae (Winterfeld et al., 2020) and Fabaceae (Ta et al., 2021; Waminal et al., 2021) species complex.

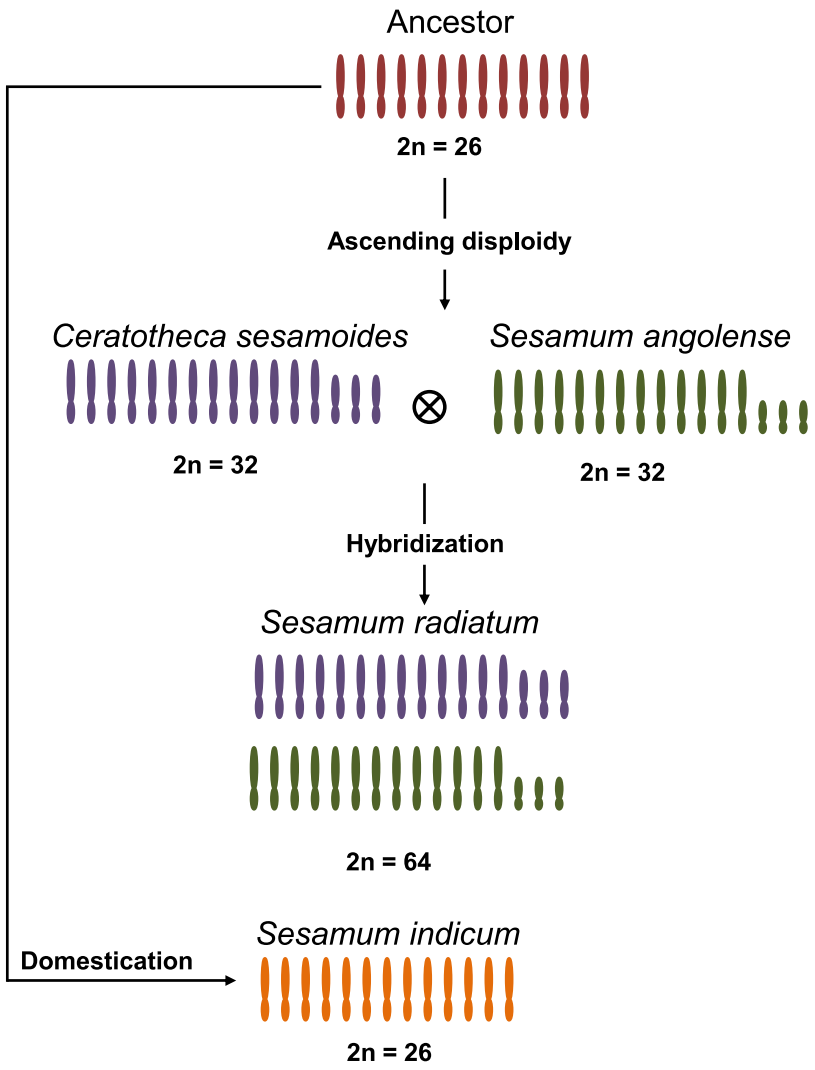

Figure 41. Proposed sketch depicting the speciation in Sesamum complex with a focus on *Sesamum radiatum* originated from the hybridization of *Sesamum angolense* and *Ceratotheca sesamoides*. The diagrammatic is based on the cytogenetic, plastome phylogenomics and comparative nuclear genome (unpublished) data.

Hybridization event is frequent in plants resulting in the advent of new species (Alix et al., 2017). In regards with the number of chromosome set, a clear classification was spotted with grouping diploid 2n = 26 as well as 2n = 32 and 2n = 64 in distinct sub-clades. In this study, Ceratotheca is robustly resolved as sister species of Sesamum genus. Based on the time divergence inference, we observed that *S. radiatum* and *S. angolense* appeared almost concomitantly alongside 2 = 32 representatives (*C. sesamoides* more specifically); suggesting a hybridization event leading to 2n = 64 might involve both *Sesamum angolense* and *Ceratotheca sesamoides*. To test this hypothesis, we aligned long-reads high-contiguous nuclear genomes of the all wild relatives onto a chromosomal-scale assembly of *S. radiatum* (2n = 64) (unpublished data). As a result, we detected the presence of two sub-genomes corresponding to *S. angolense* and *Ceratotheca sesamoides* as sub-genome A and B respectively. This genomic evidence supports karyotyping and successful field hybridization test previously performed between 2n = 32 and 2n = 64 representatives (Subramanian, 2003). Besides, the overlapping distribution areas of *C. sesamoides* and *S. angolense* in Southern Africa, covering Namibia, Zimbabwe, Mozambique, Democratic Republic of Congo, Tanzania, and Kenya (Figure 32), might also facilitated the natural hybridization.

Implications for systematic placement of Ceratotheca in Sesamum tribe (Sesamaea)

In Sesamum tribe, the classification of species has evolved following the descriptor. From whole genome (unpublished), cytogenetic and plastome analyses, we postulate that the nomenclature *C. sesamoides* and *C. triloba* might change into *S. sesamoides*, and *S. trilobum* as suggested by Bruce (1953) and (POWO, 2022). Consequently, the section Ceratotheca (Endl.) J.C. Manning & Magee, might be merged into the section Sesamum.

Limits and prospects

The present study has the merit to provide new datasets for elucidating the delineation between Sesamum and Ceratotheca. However, inaccurate species boundaries can negatively impact fundamental and applied research. Our study was circumscribed to only African native Sesamum species. The native Indian subcontinent wild species (*S. mulayanum*, *S. malabaricum* for example) are missing, narrowing a full comprehension of the phylogeny dynamics in Sesamum genus and its congeners. Although *S. malabaricum*, an Indian native wild is suggested as progenitor (Pathak et al., 2015), genomics-based evidence is required since many others Sesamum species are still not well characterized. It is worth to mention that Josephinia and Dicerocaryum sections form a species complex with Sesamum section with some representatives such as *S. africanum* Tod. and *S. zanguebarium* (Lour.) J.C. Manning & Magee renamed as *Josephinia africana* Vatke and *Dicerocaryum sinuatum* Bojer respectively (Manning and Magee, 2018). Therefore, additional characterization at organelle and whole genome levels of these close relatives should be explored to fully resolve the taxonomic puzzle in Sesamaea. Sesamum pan-plastome and pan-genome assessment will provide a path to resolve the origin of *Sesamum indicum* as well as a clear delineation among Sesamaea members.

The chloroplast genomes were highly conserved in respect with gene orientation, GC content, and gene content. However, the most divergent sequence was detected in the coding gene *ycf1*. The results from this study provide the first insight into evolutionary events underpinning the chromosome changes and speciation in Sesamum species complex. Ascending disploidy and hybridization were suggested as key speciation factors leading to the sister species *C. sesamoides* and *C. triloba*. Ultimately, the chloroplast sequence from

this study lay a foundation for DNA barcoding markers development and genomics species-centered research.

Conclusion

The present work took advantage of new sequencing technologies coupled with phenotyping strategies to successfully provide: (i) an African inclusive and worldwide core collection database resource for sesame breeders regarding not only agronomic traits, but also oil, and lignans oriented sesame breeding in Republic of Korea; (ii) the first chromosome-grade genome assembly of one of the well-established elite Korean variety *S. indicum* var. Goenbaek to accelerate genome-assisted sesame breeding; (iii) the first comprehensive characterization and identification of peroxidases and laccases potentially involved in upstream steps of lignan production in sesame; (iv) the first evidence of phylogenetics relationships among sesamum species complex, shedding light into genome ploidy evolution, nomenclature update as well as species time occurrence.

The produced resources offer an avenue for parental lines selection for population development, marker-assisted breeding, population genomics and speciation genomics. From this work, it is worth noting that we could not find evidence regarding the parental ancestor of the cultivated sesame due to the lack of native Indian wild relative's materials. However, we provide strong evidence for speciation events (ascending dysploidy and hybridization) leading to a rich African sesame species. The evolutionary analysis pointed out at least that the African continent is a source of genetic diversity when it comes to sesame genus. Knowing that wild relative are precious sources of key genes that can be used for climate change resilience, disease resistance, drought, and salinity resistance, the attention should now be focused on the numerous wild genes potentials. We suggest bringing sesame into super-pangenome era by integrating both African and Indian wild species. Altogether, the importance of solving the ancestor of sesame is capital for understanding the evolution of sesame and effective sesame breeding for trait of interest

beneficial for human food consumption, economy and health.

Abstract in Korean

국내에서 참깨(*Sesamum indicum* L.) 육종을 가속화하기 위해서는 유전자원 개발뿐만 아니라 유전체 정보 개발도 필요하다. 한국은 인도, 중국에 이어 세계에서 참깨 유전자원이 3 번째로 많은 국가지만, 여전히 양질의 유전체 정보가 부족해 활발한 유전체 기반 육종이 이루어지고 있지 않다. 최근 10 년간 참깨는 중국 재배품종(Mishuozhima, Baizhima)과 Yuzhi11, Zhongzhi13 품종의 염기서열 분석으로 유전체 정보가 구축되었다. 이를 통해 참기름 생합성과 관련된 주요 유전자, 가뭄 스트레스에 반응하는 유전자, 세사민(sesamin), 세사몰린(sesamolin) 등 특정 대사산물의 대사경로에 관여하는 유전자 등의 발견이 가능했다. 이와 같은 여러 성과에도 불구하고, 한국 토종 참깨 품종에 대한 유전체 정보 기반 육종은 한국 참깨 유전체 프로젝트가 없기 때문에 적용하는데 어려움이 있다. 따라서 기름과 리그난 함유량이 풍부한 한국 참깨 품종 건백을 대상으로 하여고품질 유전체 프로젝트가 시작되었다. 한편, 참깨는 세계에서 가장 오래된 유지작물로 알려져 있으나 그 다양성의 기원과 재배종 참깨의 조상은 여전히 수수께끼이다. 야생종 참깨 유전자원들은 생물학적, 비생물학적 스트레스를 예방하고 기름 함량을 증가시키며 세사민, 세사몰린과

같은 유용 성분 증대를 위한 유전자를 가지고 있기 때문에 중요한 자원이 된다. 건강증진 효과가 있는 참깨의 특화대사물에 대한 특허가 늘고 있는 가운데 참깨의 진화 및 분화 역사를 파악할 수 있는 충분한 자료 생산과 대한민국 참깨 육종을 위한 유전자원 개발이 시급하다. 본 논문은 (i) 참깨 육종에 활용할 수 있는 핵심수집단 구축을 목표로 하여 광범위한 유전자원을 평가하였다. (ii) 기름과 리그난 함량이 풍부한 한국 품종 건백의 염색체 수준 유전체 정보를 구축하였다. (iii) 유전체 정보를 활용하여 참깨속 종분화의 근본적인 진화 기반을 조사하였다. 우선, 506 개의 참깨 유전자원으로 이루어진 다양성 집단을 농업형질, 기름함량, 단백질 및 리그난 특성에 대해 선별하여 35 개국을 아우르는 102 개 유전자원으로 구성된 핵심수집단을 구축하였다. 이 핵심수집단은 기름수율, 단백질, 세사몰린 등 유용 형질에 관련된 유전자의 탐색을 위한 전유전체연관분석을 위한 재료로 사용될 수 있다. 활발한 유전체 기반 육종을 위해서는 표준 유전체가 필요하므로, long-read 기반 염기서열분석기술과 염색체구조포착(chromosome conformation capture) 기술을 활용해 한국 품종 건백의 염색체 수준의 유전체를 구축했다. 또한 비교 유전학적 접근법에 의해 기름함량과 특수 대사산물을 암호화하는 종별 유전자 클러스터가 추출되었다.

따라서 이 유전자 풀에 대한 심층 연구를 기반으로 유전자 편집 전략을 통해 기름 및 리그난 함량을 증대시키기 위한 유전자형의 설계가 가능할 것으로 기대된다. 한편, 참깨 종분화의 기초 연구에 기여하기 위하여, 아프리카 토종 참깨 6 종에 대한 비교 색소체학 및 세포유전학 연구가 수행되었다. 이를 통하여 참깨의 종 분화가 ascending disploidy 및 최근의 배수체화를 포함한 두 가지 주요 사건과 연관되어 있음을 밝혔다. 첫 번째 사건은 공통 조상으로부터 2n = 2x = 32(2n = 2x = 26)을 가진 새로운 종(*Sesamum angolense*, *Sesamum pedaloides*, *Ceratotheca sesamoides*, *Ceratotheca triloba*)의 탄생이다. 2n = 2x = 64 (*Sesamum radiatum*)인 종의 형성과 함께 두 번째 사건(유사화)이 비교적 최근에 발생했다. 또한, 우리는 *S. radiatum* 의 게놈 A 와 B 를 각각 *C. sesamoides* 와 *Sesamum angolense* 로 명확히 확인했다. 본 연구는 수확량, 기름함량, 단백질, 세사몰린 등 주요 형질 개선을 목표로 하는 육종사업에 활용하기 위한 참깨의 고품질 유전체 정보를 제공하며, 기초연구 측면에서는 참깨속의 종 분화를 주도하는 주요 진화적 이벤트의 증거를 제공했다. 이러한 정보는 농업형질과 기능성을 개선한 새로운 참깨 품종 개발에 유용하게 활용됨으로써 한국의 참깨 육종 발전에 긍정적 기여를 할 것으로 기대된다.

Literature cited

Abdellatef, E., Sirelkhatem, R., Mohamed Ahmed, M. M., Radwan, K. H., and Khalafalla, M. M. (2008). Study of genetic diversity in Sudanese sesame (*Sesamum indicum* L.) germplasm using random amplified polymorphic DNA (RAPD) markers. *African J. Biotechnol.* doi:10.4314/ajb.v7i24.59603.

Aleem, M., Riaz, A., Raza, Q., Aleem, M., Aslam, M., Kong, K., et al. (2022). Genome-wide characterization and functional analysis of class III peroxidase gene family in soybean reveal regulatory roles of GsPOD40 in drought tolerance. *Genomics* 114, 45–60. doi:10.1016/j.ygeno.2021.11.016.

Alix, K., Gérard, P. R., Schwarzacher, T., and Heslop-Harrison, J. S. (Pat) (2017). Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Ann. Bot.* 120, 183–194. doi:10.1093/aob/mcx079.

Allen, G. C., Flores-Vergara, M. A., Krasynanski, S., Kumar, S., and Thompson, W. F. (2006). A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. *Nat. Protoc.* 1, 2320–2325. doi:10.1038/nprot.2006.384.

Alyemeni, M. N., Basahy, A. Y., and Sher, H. (2011). Journal of medicinal plant research. *J. Med. Plants Res.* 5, 270–274. Available at: http://www.academicjournals.org/journal/JMPR/article-abstract/BDC1CCC18667.

Amiryousefi, A., Hyvönen, J., and Poczai, P. (2018). IRscope: an online program to visualize the junction sites of chloroplast genomes. *Bioinformatics* 34, 3030–3031. doi:10.1093/bioinformatics/bty220.

Andargie, M., Vinas, M., Rathgeb, A., Möller, E., and Karlovsky, P. (2021). Lignans of Sesame (*Sesamum indicum* L.): A Comprehensive Review. *Molecules* 26, 883.

doi:10.3390/molecules26040883.

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Angiolillo, A., Mencuccini, M., and Baldoni, L. (1999). Olive genetic diversity assessed using amplified fragment length polymorphisms. *Theor. Appl. Genet.* doi:10.1007/s001220051087.

Ankenbrand, M. J., Hohlfeld, S., Hackl, T., and Förster, F. (2017). AliTV—interactive visualization of whole genome comparisons. *PeerJ Comput. Sci.* 3, e116. doi:10.7717/peerj-cs.116.

Anonymous (2021). Lignans Market Share, Size, Trends, Industry Analysis Report. Polaris Market Research: San Franscisco, CA, USA; pp. 1–145.

Asaf, S., Khan, A. L., Lubna, Khan, A., Khan, A., Khan, G., et al. (2020). Expanded inverted repeat region with large scale inversion in the first complete plastid genome sequence of Plantago ovata. *Sci. Rep.* 10, 3881. doi:10.1038/s41598-020-60803-y.

Asekova, S., Kulkarni, K. P., Oh, K. W., Lee, M.-H., Oh, E., Kim, J.-I., et al. (2018). Analysis of Molecular Variance and Population Structure of Sesame (*Sesamum indicum* L.) Genotypes Using Simple Sequence Repeat Markers. *Plant Breed. Biotechnol.* 6, 321–336. doi:10.9787/PBB.2018.6.4.321.

Asekova, S., Oh, E., Kulkarni, K. P., Siddique, M. I., Lee, M. H., Kim, J. I., et al. (2021). An Integrated Approach of QTL Mapping and Genome-Wide Association Analysis Identifies Candidate Genes for Phytophthora Blight Resistance in Sesame (Sesamum indicum L.). *Front. Plant Sci.* 12, 1–15. doi:10.3389/fpls.2021.604709.

Attanzio, A., D'Anneo, A., Pappalardo, F., Bonina, F. P., Livrea, M. A., Allegra, M., et al.

(2019). Phenolic composition of hydrophilic extract of manna from sicilian Fraxinus angustifolia vahl and its reducing, antioxidant and anti-inflammatory activity in vitro. *Antioxidants* 8, 1–13. doi:10.3390/antiox8100494.

Aworh, O. C. (2018). From lesser-known to super vegetables: the growing profile of African traditional leafy vegetables in promoting food security and wellness. *J. Sci. Food Agric.* 98, 3609–3613. doi:10.1002/jsfa.8902.

Badri, J., Yepuri, V., Ghanta, A., Siva, S., and Siddiq, E. A. (2014). Development of microsatellite markers in sesame (*Sesamum indicum* L.). *Turkish J. Agric. For.* doi:10.3906/tar-1312-104.

Bahadori, M. B., Zengin, G., Bahadori, S., Dinparast, L., and Movahhedin, N. (2018). Phenolic composition and functional properties of wild mint (*Mentha longifolia var. calliantha* (Stapf) Briq.). *Int. J. Food Prop.* 21, 198–208. doi:10.1080/10942912.2018.1440238.

Baidouri, M. El, and Panaud, O. (2013). Comparative genomic paleontology across plant kingdom reveals the dynamics of TE-driven genome evolution. *Genome Biol. Evol.* 5, 954–965. doi:10.1093/gbe/evt025.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Res.* 43, W39–W49. doi:10.1093/nar/gkv416.

Barceló, A. R., Ros, L. V. G., and Carrasco, A. E. (2007). Looking for syringyl peroxidases. *Trends Plant Sci.* 12, 486–491. doi:10.1016/j.tplants.2007.09.002.

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

Bedigian, D. (2014). A new combination for the Indian progenitor of sesame, *sesamum indicum* (Pedaliaceae). *Novon* 23, 5–13. doi:10.3417/2012062.

Bedigian, D. (2015). Systematics and evolution in Sesamum L. (Pedaliaceae), part 1: Evidence regarding the origin of sesame and its closest relatives. *Webbia J. Plant Taxon. Geogr.* 70, 1–42. doi:10.1080/00837792.2014.968457.

Bedigian, D. (2018). Feeding the Forgotten: Wild and Cultivated Ceratotheca and Sesamum (Pedaliaceae) That Nourish and Provide Remedies in Africa. *Econ. Bot.* 72, 496–542. doi:10.1007/s12231-018-9437-z.

Bedigian, D., Seigler, D. S., and Harlan, J. R. (1985a). Sesamin, sesamolin and the origin of sesame. *Biochem. Syst. Ecol.* 13, 133–139. doi:10.1016/0305-1978(85)90071-7.

Bedigian, D., Seigler, D. S., and Harlan, J. R. (1985b). Sesamin, sesamolin and the origin of sesame. *Biochem. Syst. Ecol.* 13, 133–139. doi:10.1016/0305-1978(85)90071-7.

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. *Bioinformatics* 33, 2583–2585. doi:10.1093/bioinformatics/btx198.

Bennett, M. D., and Leitch, I. J. (2005). Nuclear DNA amounts in angiosperms: Progress, problems and prospects. in *Annals of Botany* doi:10.1093/aob/mci003.

Besnard, G., and Bervillé, A. (2002). On chloroplast DNA variations in the olive (Olea europaea L.) complex: Comparison of RFLP and PCR polymorphisms. *Theor. Appl. Genet.* doi:10.1007/s00122-001-0834-8.

Biancarosa, I., Espe, M., Bruckner, C. G., Heesch, S., Liland, N., Waagbø, R., et al. (2017). Amino acid composition, protein content, and nitrogen-to-protein conversion factors of 21 seaweed species from Norwegian waters. *J. Appl. Phycol.* 29, 1001–1009. doi:10.1007/s10811-016-0984-3.

Biju, V. C., P.R., S., Vijayan, S., Rajan, V. S., Sasi, A., Janardhanan, A., et al. (2019). The Complete Chloroplast Genome of *Trichopus zeylanicus* , And Phylogenetic Analysis

with Dioscoreales . *Plant Genome* 12, 190032. doi:10.3835/plantgenome2019.04.0032.

Bindschedler, L. V., Dewdney, J., Blee, K. A., Stone, J. M., Asai, T., Plotnikov, J., et al. (2006). Peroxidase-dependent apoplastic oxidative burst in Arabidopsis required for pathogen resistance. *Plant J.* 47, 851–863. doi:10.1111/j.1365-313X.2006.02837.x.

Blischak, P. D., Wenzel, A. J., and Wolfe, A. D. (2014). Gene Prediction and Annotation in Penstemon (Plantaginaceae): A Workflow for Marker Development from Extremely Low-Coverage Genome Sequencing . *Appl. Plant Sci.* 2, 1400044. doi:10.3732/apps.1400044.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bouguellid, G., Russo, C., Lavorgna, M., Piscitelli, C., Ayouni, K., Wilson, E., et al. (2020). Antimutagenic, antigenotoxic and antiproliferative activities of *Fraxinus angustifolia* Vahl. leaves and stem bark extracts and their phytochemical composition. *PLoS One* 15, 1–21. doi:10.1371/journal.pone.0230690.

Bronzini de Caraffa, V., Maury, J., Gambotti, C., Breton, C., Bervillé, A., and Giannettini, J. (2002). Mitochondrial DNA variation and RAPD mark oleasters, olive and feral olive from Western and Eastern Mediterranean. *Theor. Appl. Genet.* doi:10.1007/s00122-002-0883-7.

Bruce, A. (1953). *Flora of Tropical East Africa:Pedaliaceae*.

Cai, K., Liu, H., Chen, S., Liu, Y., Zhao, X., and Chen, S. (2021). Genome-wide identification and analysis of class III peroxidases in *Betula pendula*. *BMC Genomics* 22, 1–19. doi:10.1186/s12864-021-07622-1.

Cai, X., Davis, E. J., Ballif, J., Liang, M., Bushman, E., Haroldsen, V., et al. (2006). Mutant

identification and characterization of the laccase gene family in Arabidopsis. *J. Exp. Bot.* 57, 2563–2569. doi:10.1093/jxb/erl022.

Cannon, S. B., Mitra, A., Baumgarten, A., Young, N. D., and May, G. (2004). The roles of segmental and tandem gene duplication in the evolution of large gene families in *Arabidopsis thaliana*. *BMC Plant Biol.* 4, 1–21. doi:10.1186/1471-2229-4-10.

Cao, Y., Han, Y., Meng, D., Li, D., Jin, Q., Lin, Y., et al. (2016). Structural, evolutionary, and functional analysis of the class III peroxidase gene family in Chinese pear (*Pyrus bretschneideri*). *Front. Plant Sci.* 7, 1–12. doi:10.3389/fpls.2016.01874.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972–1973. doi:10.1093/bioinformatics/btp348.

Chami, A., Pillai, C., Balachandran, M., Rhesma, V., Sundaresan, A., Thomas, S., et al. (2013). Process for the extraction of bioactive lignans with high yield and purity from sesame oil. U.S. Patent No. 8,350,066 B2, 8 January.

Chan, P. P., and Lowe, T. M. (2019). "tRNAscan-SE: Searching for tRNA Genes in Genomic Sequences," in *Gene Prediction: Methods and Protocols, Methods in Molecular Biology*, ed. K. Martin, 1–14. doi:10.1007/978-1-4939-9173-0_1.

Chao, J., Li, Z., Sun, Y., Aluko, O. O., Wu, X., Wang, Q., et al. (2021). MG2C: a user-friendly online tool for drawing genetic maps. *Mol. Hortic.* 1, 1–4. doi:10.1186/s43897-021-00020-x.

Chen, C., Chen, H., Zhang, Y., Thomas, H. R., Frank, M. H., He, Y., et al. (2020). TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* 13, 1194–1202. doi:10.1016/j.molp.2020.06.009.

Chen, Y., Li, C., Zhang, B., Yi, J., Yang, Y., Kong, C., et al. (2019). The Role of the Late

Embryogenesis-Abundant (LEA) Protein Family in Development and the Abiotic Stress Response: A Comprehensive Expression Analysis of Potato (*Solanum tuberosum*). *Genes (Basel).* 10, 148. doi:10.3390/genes10020148.

Chi, X., Zhang, F., Dong, Q., and Chen, S. (2020). Insights into Comparative Genomics, Codon Usage Bias, and Phylogenetic Relationship of Species from Biebersteiniaceae and Nitrariaceae Based on Complete Chloroplast Genomes. *Plants* 9, 1605. doi:10.3390/plants9111605.

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods* 13, 1050–1054. doi:10.1038/nmeth.4035.

Cipriani, G., Marrazzo, M. T., Marconi, R., Cimato, A., and Testolin, R. (2002). Microsatellite markers isolated in olive (*Olea europaea* L.) are suitable for individual fingerprinting and reveal polymorphism within ancient cultivars. *Theor. Appl. Genet.* doi:10.1007/s001220100685.

Claros, M. G., Bautista, R., Guerrero-Fernández, D., Benzerki, H., Seoane, P., and Fernández-Pozo, N. (2012). Why assembling plant genome sequences is so challenging. *Biology (Basel).* 1, 439–459. doi:10.3390/biology1020439.

Cronquist, A. (1981). *An Integrated System of Classification of Flowering Plants*. New York: Colombia University Press.

Cruz, F., Julca, I., Gómez-Garrido, J., Loska, D., Marcet-Houben, M., Cano, E., et al. (2016). Genome sequence of the olive tree, *Olea europaea*. *Gigascience*. doi:10.1186/s13742-016-0134-5.

Cui, C., Mei, H., Liu, Y., Zhang, H., and Zheng, Y. (2017). Genetic diversity, population structure, and linkage disequilibrium of an association-mapping panel revealed by

genome-wide SNP markers in sesame. *Front. Plant Sci.* doi:10.3389/fpls.2017.01189.

Cultrera, N. G. M., Sarri, V., Lucentini, L., Ceccarelli, M., Alagna, F., Mariotti, R., et al. (2019). High levels of variation within gene sequences of Olea europaea L. *Front. Plant Sci.* doi:10.3389/fpls.2018.01932.

Dalibalta, S., Majdalawieh, A. F., and Manjikian, H. (2020). Health benefits of sesamin on cardiovascular disease and its associated risk factors. *Saudi Pharm. J.* 28, 1276–1289. doi:10.1016/j.jsps.2020.08.018.

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Implicitfunction.Pdf. *Genome Res.* 14, 1394–1403. doi:10.1101/gr.2289704.tion.

Daudi, A., Cheng, Z., O'Brien, J. A., Mammarella, N., Khan, S., Ausubel, F. M., et al. (2012). The apoplastic oxidative burst peroxidase in Arabidopsis is a major component of pattern-triggered immunity. *Plant Cell* 24, 275–287. doi:10.1105/tpc.111.093039.

Davin, L. B., Wang, H. Bin, Crowell, A. L., Bedgar, D. L., Martin, D. M., Sarkanen, S., et al. (1997). Stereoselective bimolecular phenoxy radical coupling by an auxiliary (dirigent) protein without an active center. *Science (80-.).* 275, 362–366. doi:10.1126/science.275.5298.362.

De Beukelaer, H., Davenport, G. F., and Fack, V. (2018). Core Hunter 3: flexible core subset selection. *BMC Bioinformatics* 19, 203. doi:10.1186/s12859-018-2209-z.

de Mendiburu, F. (2020). agricolae: Statistical procedures for agricultural research. R package version 1.3-3. http://CRAN.R-project.org/package=agricolae.

Dewey, D. R., and Lu, K. H. (1959). A Correlation and Path-Coefficient Analysis of Components of Crested Wheatgrass Seed Production 1. *Agron. J.* 51, 515–518. doi:10.2134/agronj1959.00021962005100090002x.

Ding, P., Shao, Y., Li, Q., Gao, J., Zhang, R., Lai, X., et al. (2016). The complete chloroplast

genome sequence of the medicinal plant *Andrographis paniculata*. *Mitochondrial DNA*. doi:10.3109/19401736.2015.1025258.

Dixit, A., Jin, M. H., Chung, J. W., Yu, J. W., Chung, H. K., Ma, K. H., et al. (2005). Development of polymorphic microsatellite markers in sesame (*Sesamum indicum* L.). *Mol. Ecol. Notes*. doi:10.1111/j.1471-8286.2005.01048.x.

Dong, W., Xu, C., Li, C., Sun, J., Zuo, Y., Shi, S., et al. (2015). ycf1, the most promising plastid DNA barcode of land plants. *Sci. Rep.* 5, 8348. doi:10.1038/srep08348.

Dossa, K., Diouf, D., Wang, L., Wei, X., Zhang, Y., Niang, M., et al. (2017a). The emerging oilseed crop *Sesamum indicum* enters the "Omics" era. *Front. Plant Sci.* 8, 1–16. doi:10.3389/fpls.2017.01154.

Dossa, K., Diouf, D., Wang, L., Wei, X., Zhang, Y., Niang, M., et al. (2017b). The Emerging Oilseed Crop *Sesamum indicum* Enters the "Omics" Era. *Front. Plant Sci*. 8, 1–16. doi:10.3389/fpls.2017.01154.

Dossa, K., Li, D., Zhou, R., Yu, J., Wang, L., Zhang, Y., et al. (2019). The genetic basis of drought tolerance in the high oil crop *Sesamum indicum*. *Plant Biotechnol. J.* 17, 1788–1803. doi:10.1111/pbi.13100.

Dossa, K., Wei, X., Niang, M., Liu, P., Zhang, Y., Wang, L., et al. (2018). Near-infrared reflectance spectroscopy reveals wide variation in major components of sesame seeds from Africa and Asia. *Crop J.* 6, 202–206. doi:10.1016/j.cj.2017.10.003.

Dossa, K., Yu, J., Liao, B., Cisse, N., and Zhang, X. (2017c). Development of Highly Informative Genome-Wide Single Sequence Repeat Markers for Breeding Applications in Sesame and Construction of a Web Resource: SisatBase. *Front. Plant Sci.* 8, 1–10. doi:10.3389/fpls.2017.01470.

Dossa, K., Zhou, R., Li, D., Liu, A., Qin, L., Mmadi, M. A., et al. (2020). A novel motif in

the 5’-UTR of an orphan gene ‘ Big Root Biomass’ modulates root biomass in sesame . *Plant Biotechnol. J.* doi:10.1111/pbi.13531.

Duan, P., Wang, G., Chao, M., Zhang, Z., and Zhang, B. (2019). Genome-wide identification and analysis of class III peroxidases in allotetraploid cotton (*Gossypium hirsutum* L.) and their responses to PK deficiency. *Genes (Basel).* 10. doi:10.3390/genes10060473.

Dunnington, D., and Thorne, B. (2021). ggspatial: Spatial data framework for ggplot2. version 1.1.5 http://CRAN.R-project.org/package=ggspatial.

Durazzo, A., Carcea, M., Adlercreutz, H., Azzini, E., Polito, A., Olivieri, L., et al. (2014). Effects of consumption of whole grain foods rich in lignans in healthy postmenopausal women with moderate serum cholesterol: A pilot study. *Int. J. Food Sci. Nutr.* 65, 637–645. doi:10.3109/09637486.2014.893283.

Elleuch, M., Besbes, S., Roiseux, O., Blecker, C., and Attia, H. (2007). Quality characteristics of sesame seeds and by-products. *Food Chem.* 103, 641–650. doi:10.1016/j.foodchem.2006.09.008.

Emms, D. M., and Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol. Biol. Evol.* 34, 3267–3278. doi:10.1093/molbev/msx259.

Emms, D. M., and Kelly, S. (2018). STAG: Species Tree Inference from All Genes. *bioRxiv 267914*, 1–29. doi:https://doi.org/10.1101/267914.

Emms, D. M., and Kelly, S. (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 20, 238. doi:10.1186/s13059-019-1832-y.

Emms, D. M., and Kelly, S. (2022). SHOOT: phylogenetic gene search and ortholog inference. *Genome Biol.* 23, 1–13. doi:10.1186/s13059-022-02652-8.

Fabbri, A., Hormaza, J. I., and Polito, V. S. (1995). Random amplified polymorphic DNA analysis of olive (*Olea europaea* L.) cultivars. *J. Am. Soc. Hortic. Sci.*

doi:10.21273/jashs.120.3.538.

Fan, H., and Chu, J. Y. (2007). A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics Bioinforma.* 5, 7–14. doi:10.1016/S1672-0229(07)60009-6.

Fang, F., Zhang, X. L., Luo, H. H., Zhou, J. J., Gong, Y. H., Li, W. J., et al. (2015). An intracellular Laccase is responsible for epicatechin-mediated Anthocyanin degradation in litchi fruit Pericarp. *Plant Physiol.* 169, 2391–2408. doi:10.1104/pp.15.00359.

FAO, IFAD, UNICEF, WFP, W. (2020). *The State of Food Security and Nutrition in the World 2020*. Rome: FAO, IFAD, UNICEF, WFP and WHO doi:10.4060/ca9692en.

Farajbakhsh, A., Mazloomi, S. M., Mazidi, M., Rezaie, P., Akbarzadeh, M., Ahmad, S. P., et al. (2019). Sesame oil and vitamin E co-administration may improve cardiometabolic risk factors in patients with metabolic syndrome: a randomized clinical trial. *Eur. J. Clin. Nutr.* 73, 1403–1411. doi:10.1038/s41430-019-0438-5.

Federer, W. T., and Raghavarao, D. (1975). On Augmented Designs. *Biometrics* 31, 29. doi:10.2307/2529707.

Forse, A. R., and Chavali, S. R. (2001a). Sesamol inhibition of Δ-5-desaturase activity and uses therefor. U.S. Patent No. 6,172,106 B1, 9 January.

Forse, A. R., and Chavali, S. R. (2001b). Sesamol Inhibitor of Delta-5-Desaturase Activity and Uses Therefor. U.S. Patent No. 2001/0031275.

Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* 32, 273–279. doi:10.1093/nar/gkh458.

Friis, G., Vizueta, J., Smith, E. G., Nelson, D. R., Khraiwesh, B., Qudeimat, E., et al. (2020). A high-quality genome assembly and annotation of the gray mangrove, *Avicennia marina*. *G3 Genes Genomes Genetics*. doi:10.1093/g3journal/jkaa025.

Furuki, T., Shimizu, T., Kikawada, T., Okuda, T., and Sakurai, M. (2011). Salt Effects on the Structural and Thermodynamic Properties of a Group 3 LEA Protein Model Peptide. *Biochemistry* 50, 7093–7103. doi:10.1021/bi200719s.

Gabaldón, C., López-Serrano, M., Pedreño, M. A., and Barceló, A. R. (2005). Cloning and molecular characterization of the basic peroxidase isoenzyme from *Zinnia elegans*, an enzyme involved in lignin biosynthesis. *Plant Physiol.* 139, 1138–1154. doi:10.1104/pp.105.069674.

Gao, B., Yuan, L., Tang, T., Hou, J., Pan, K., and Wei, N. (2019). The complete chloroplast genome sequence of *Alpinia oxyphylla* Miq. and comparison analysis within the Zingiberaceae family. *PLoS One* 14, e0218817. doi:10.1371/journal.pone.0218817.

Garg, A., Agrawal, L., Misra, R. C., Sharma, S., and Ghosh, S. (2015). *Andrographis paniculata* transcriptome provides molecular insights into tissue-specific accumulation of medicinal diterpenes. *BMC Genomics*. doi:10.1186/s12864-015-1864-y.

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., et al. (2005). "Protein Identification and Analysis Tools on the ExPASy Server," in *The Proteomics Protocols Handbook* (Totowa, NJ: Humana Press), 571–607. doi:10.1385/1-59259-890-0:571.

Gavnholt, B., and Larsen, K. (2002). Molecular biology of plant laccases in relation to lignin formation. *Physiol. Plant.* 116, 273–280. doi:10.1034/j.1399-3054.2002.1160301.x.

Girma, G., Nida, H., Tirfessa, A., Lule, D., Bejiga, T., Seyoum, A., et al. (2020). A comprehensive phenotypic and genomic characterization of Ethiopian sorghum germplasm defines core collection and reveals rich genetic potential in adaptive traits. *Plant Genome* 13, 1–17. doi:10.1002/tpg2.20055.

Gommans, W. M., Mullen, S. P., and Maas, S. (2009). RNA editing: A driving force for

adaptive evolution? *BioEssays* 31, 1137–1145. doi:10.1002/bies.200900045.

Gormley, I. C., Bedigian, D., and Olmstead, R. G. (2015). Phylogeny of Pedaliaceae and Martyniaceae and the Placement of Trapella in Plantaginaceae s. l. *Syst. Bot.* 40, 259–268. doi:10.1600/036364415x686558.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., et al. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* 36, 3420–3435. doi:10.1093/nar/gkn176.

Gower, J. C. (1971). A General Coefficient of Similarity and Some of Its Properties. *Biometrics* 27, 857. doi:10.2307/2528823.

Greiner, S., Lehwark, P., and Bock, R. (2019). OrganellarGenomeDRAW (OGDRAW) version 1.3.1: expanded toolkit for the graphical visualization of organellar genomes. *Nucleic Acids Res.* 47, W59–W64. doi:10.1093/nar/gkz238.

Grosjean, P., Ibanez, F., and Etienne, M. (2018). pastecs: Package for Analysis of Space-Time Ecological Series. R package version 1.3.21. https://CRAN.R-project.org/package=pastecs.

Guo, Y.-Y., Yang, J.-X., Bai, M.-Z., Zhang, G.-Q., and Liu, Z.-J. (2021). The chloroplast genome evolution of Venus slipper (Paphiopedilum): IR expansion, SSC contraction, and highly rearranged SSC regions. *BMC Plant Biol.* 21, 248. doi:10.1186/s12870-021-03053-y.

Ha, T. J., Lee, M. H., Seo, W. D., Baek, I. Y., Kang, J. E., and Lee, J. H. (2017). Changes occurring in nutritional components (phytochemicals and free amino acid) of raw and sprouted seeds of white and black sesame (*Sesamum indicum* L.) and screening of their antioxidant activities. *Food Sci. Biotechnol.* 26, 71–78. doi:10.1007/s10068-017-0010-9.

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* 9, 1–22. doi:10.1186/gb-2008-9-1-r7.

Hand, S. C., Menze, M. A., Toner, M., Boswell, L., and Moore, D. (2011). LEA proteins during water stress: Not just for plants anymore. *Annu. Rev. Physiol.* 73, 115–134. doi:10.1146/annurev-physiol-012110-142203.

Handy, S. M., Parks, M. B., Deeds, J. R., Liston, A., De Jager, L. S., Luccioli, S., et al. (2011). Use of the chloroplast gene ycf1 for the genetic differentiation of pine nuts obtained from consumers experiencing dysgeusia. *J. Agric. Food Chem.* 59, 10995–11002. doi:10.1021/jf203215v.

Harada, E., Murata, J., Ono, E., Toyonaga, H., Shiraishi, A., Hideshima, K., et al. (2020). (+)-Sesamin-oxidising CYP92B14 shapes specialised lignan metabolism in sesame. *Plant J.* 104, 1117–1128. doi:10.1111/tpj.14989.

Hardwicke, J. E., King, J., and Terrell, R. C. (1959). Synthesis of sesamol acetate and sesamol. U.S. Patent No. 2,885,407, 5 May. doi:10.1145/178951.178972.

Harikumar, K. B., Sung, B., Tharakan, S. T., Pandey, M. K., Joy, B., Guha, S., et al. (2010). Sesamin Manifests Chemopreventive Effects through the Suppression of NF-κB–Regulated Cell Survival, Proliferation, Invasion, and Angiogenic Gene Products. *Mol. Cancer Res.* 8, 751–761. doi:10.1158/1541-7786.MCR-09-0565.

He, Y., Peng, F., Deng, C., Xiong, L., Huang, Z. Y., Zhang, R. Q., et al. (2018). Data descriptor: Building an octaploid genome and transcriptome of the medicinal plant pogostemon cablin from lamiales. *Sci. Data* 5, 1–11. doi:10.1038/sdata.2018.274.

He, Y., Xiao, H., Deng, C., Xiong, L., Nie, H., and Peng, C. (2016a). Survey of the genome

of Pogostemon cablin provides insights into its evolutionary history and sesquiterpenoid biosynthesis. *Sci. Rep.* doi:10.1038/srep26405.

He, Y., Xiao, H., Deng, C., Xiong, L., Nie, H., and Peng, C. (2016b). Survey of the genome of *Pogostemon cablin* provides insights into its evolutionary history and sesquiterpenoid biosynthesis. *Sci. Rep.* 6, 1–10. doi:10.1038/srep26405.

Hellsten, U., Wright, K. M., Jenkins, J., Shu, S., Yuan, Y., Wessler, S. R., et al. (2013). Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci.* 110, 19478–19482. doi:10.1073/pnas.1319032110.

Hernández, P., De La Rosa, R., Rallo, L., Dorado, G., and Martín, A. (2001a). Development of SCAR markers in olive (*Olea europaea*) by direct sequencing of RAPD products: Applications in olive germplasm evaluation and mapping. *Theor. Appl. Genet.* doi:10.1007/s001220100603.

Hernández, P., De la Rosa, R., Rallo, L., Martín, A., and Dorado, G. (2001b). First evidence of a retrotransposon-like element in olive (*Olea europaea*): Implications in plant variety identification by SCAR-marker development. *Theor. Appl. Genet.* doi:10.1007/s001220000515.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32, 767–769. doi:10.1093/bioinformatics/btv661.

Hon, T., Mars, K., Young, G., Tsai, Y. C., Karalius, J. W., Landolin, J. M., et al. (2020). Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci. Data* 7, 1–11. doi:10.1038/s41597-020-00743-4.

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* 35, 585–587.

doi:10.1093/nar/gkm259.

Hsu, E., and Parthasarathy, S. (2017). Anti-inflammatory and Antioxidant Effects of Sesame Oil on Atherosclerosis: A Descriptive Literature Review. *Cureus* 9. doi:10.7759/cureus.1438.

Hu, J., Zhu, J., and Xu, H. M. (2000). Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor. Appl. Genet.* 101, 264–268. doi:10.1007/s001220051478.

Hu, Q., Min, L., Yang, X., Jin, S., Zhang, L., Li, Y., et al. (2018). Laccase GhLac1 modulates broad-spectrum biotic stress tolerance via manipulating phenylpropanoid pathway and jasmonic acid synthesis. *Plant Physiol.* 176, 1808–1823. doi:10.1104/pp.17.01628.

Hu, Q., Xiao, S., Guan, Q., Tu, L., Sheng, F., Du, X., et al. (2020). The laccase gene GhLac1 modulates fiber initiation and elongation by coordinating jasmonic acid and flavonoid metabolism. *Crop J.* 8, 522–533. doi:10.1016/j.cj.2019.11.006.

Huang, H., Zhou, G., Pu, R., Cui, Y., and Liao, D. (2021). Clinical evidence of dietary supplementation with sesame on cardiovascular risk factors: An updated meta-analysis of randomized controlled trials. *Crit. Rev. Food Sci. Nutr.* 0, 1–11. doi:10.1080/10408398.2021.1888689.

Huang, L. S., Sun, Y. Q., Jin, Y., Gao, Q., Hu, X. G., Gao, F. L., et al. (2018). Development of high transferability cpSSR markers for individual identification and genetic investigation in Cupressaceae species. *Ecol. Evol.* 8, 4967–4977. doi:10.1002/ece3.4053.

Huang, Y. le, Zhang, L. kui, Zhang, K., Chen, S. min, Hu, J. bin, and Cheng, F. (2022). The impact of tandem duplication on gene evolution in Solanaceae species. *J. Integr. Agric.*

21, 1004–1014. doi:10.1016/S2095-3119(21)63698-5.

Huelsenbeck, J. P., and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17, 754–755. doi:10.1093/bioinformatics/17.8.754.

Hundertmark, M., and Hincha, D. K. (2008). LEA (Late Embryogenesis Abundant) proteins and their encoding genes in *Arabidopsis thaliana*. *BMC Genomics* 9, 1–22. doi:10.1186/1471-2164-9-118.

Hundertmark, M., Popova, A. V., Rausch, S., Seckler, R., and Hincha, D. K. (2012). Influence of drying on the secondary structure of intrinsically disordered and globular proteins. *Biochem. Biophys. Res. Commun.* 417, 122–128. doi:10.1016/j.bbrc.2011.11.067.

Hussain, A., Bilal, M., Rafeeq, H., Jabeen, Z., Afsheen, N., Sher, F., et al. (2022). "Role of laccase in the pulp and paper industry," in *Nanotechnology in Paper and Wood Engineering* (Elsevier), 35–60. doi:10.1016/B978-0-323-85835-9.00006-4.

Ibarra-Laclette, E., Lyons, E., Hernández-Guzmán, G., Pérez-Torres, C. A., Carretero-Paulet, L., Chang, T. H., et al. (2013). Architecture and evolution of a minute plant genome. *Nature* 498, 94–98. doi:10.1038/nature12132.

IPGRI, and NBPGR (2004). *Descriptors of Sesame (Sesamum spp.)*. International Plant Genetic Resources Institute, Rome, Italy; and National Bureau of Plant Genetic Resources, New Delhi, India.

Ivanova, Z., Sablok, G., Daskalova, E., Zahmanova, G., Apostolova, E., Yahubyan, G., et al. (2017). Chloroplast genome analysis of resurrection tertiary relict Haberlea rhodopensis highlights genes important for desiccation stress response. *Front. Plant Sci.* 8, 1–15. doi:10.3389/fpls.2017.00204.

Jackson, P., and Ricardo, C. P. P. (1998). The changing peroxidase polymorphism in Lupinus

albus during vegetative development. *Funct. Plant Biol.* 25, 261. doi:10.1071/PP97083.

Jiang, P., Shi, F.-X., Li, M.-R., Liu, B., Wen, J., Xiao, H.-X., et al. (2018). Positive Selection Driving Cytoplasmic Genome Evolution of the Medicinally Important Ginseng Plant Genus Panax. *Front. Plant Sci.* 9, 1–11. doi:10.3389/fpls.2018.00359.

Jin, J.-J., Yu, W.-B., Yang, J.-B., Song, Y., DePamphilis, C. W., Yi, T.-S., et al. (2020). GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol.* 21, 241. doi:10.1186/s13059-020-02154-5.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics* 30, 1236–1240. doi:10.1093/bioinformatics/btu031.

Julca, I., Marcet-Houben, M., Cruz, F., Gómez-Garrido, J., Gaut, B. S., Díez, C. M., et al. (2020). Genomic evidence for recurrent genetic admixture during the domestication of Mediterranean olive trees (*Olea europaea* L.). *BMC Biol.* doi:10.1186/s12915-020-00881-6.

Kadota, M., Nishimura, O., Miura, H., Tanaka, K., Hiratani, I., and Kuraku, S. (2020). Multifaceted Hi-C benchmarking: What makes a difference in chromosome-scale genome scaffolding? *Gigascience* 9, 1–15. doi:10.1093/gigascience/giz158.

Kalvari, I., Nawrocki, E. P., Ontiveros-Palacios, N., Argasinska, J., Lamkiewicz, K., Marz, M., et al. (2021). Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* 49, D192–D200. doi:10.1093/nar/gkaa1047.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* 14, 587–589. doi:10.1038/nmeth.4285.

Kang, C. W., Kim, S. Y., Lee, S. W., Mathur, P. N., Hodgkin, T., Zhou, M. De, et al. (2006). Selection of a Core Collection of Korean Sesame Germplasm by a Stepwise Clustering Method. *Breed. Sci.* 56, 85–91. doi:10.1270/jsbbs.56.85.

Kang, S. H., Kim, B., Choi, B. S., Lee, H. O., Kim, N. H., Lee, S. J., et al. (2020). Genome Assembly and Annotation of Soft-Shelled Adlay (*Coix lacryma-jobi* Variety *ma-yuen*), a Cereal and Medicinal Crop in the Poaceae Family. *Front. Plant Sci.* 11, 1–14. doi:10.3389/fpls.2020.00630.

Kassambara, A., and Mundt, F. (2019). factoextra: Extract and visualize the results of multivariate data analyses. R package version 1.0.7. http://CRAN.R-project.org/package=factoextra. Available at: https://rdrr.io/github/kassambara/factoextra/.

Katayama, S., Sugiyama, H., Kushimoto, S., Uchiyama, Y., Hirano, M., and Nakamura, S. (2016). Effects of Sesaminol Feeding on Brain Aβ Accumulation in a Senescence-Accelerated Mouse-Prone 8. *J. Agric. Food Chem.* 64, 4908–4913. doi:10.1021/acs.jafc.6b01237.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010.

Kaya, H. B., Akdemir, D., Lozano, R., Cetin, O., Sozer Kaya, H., Sahin, M., et al. (2019). Genome wide association study of 5 agronomic traits in olive (*Olea europaea* L.). *Sci. Rep.* doi:10.1038/s41598-019-55338-w.

Kidwai, M., Ahmad, I. Z., and Chakrabarty, D. (2020). Class III peroxidase: an indispensable enzyme for biotic/abiotic stress tolerance and a potent candidate for crop improvement. *Plant Cell Rep.* 39, 1381–1393. doi:10.1007/s00299-020-02588-y.

Kim, A. Y., Yun, C. I., Lee, J. G., and Kim, Y. J. (2020). Determination and daily intake estimation of lignans in sesame seeds and sesame oil products in Korea. *Foods* 9, 1–11. doi:10.3390/foods9040394.

Kim, D. H., Zur, G., Danin-Poleg, Y., Lee, S. W., Shim, K. B., Kang, C. W., et al. (2002a). Genetic relationships of sesame germplasm collection as revealed by inter-simple sequence repeats. *Plant Breed.* doi:10.1046/j.1439-0523.2002.00700.x.

Kim, D., Langmead, B., and Salzberg, S. L. (2015a). HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. doi:10.1038/nmeth.3317.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *ACS, Div. Environ. Chem. - Prepr. Ext. Abstr.* 14. doi:https://doi.org/10.1186/gb-2013-14-4-r36.

Kim, K.-W., Chung, H.-K., Cho, G.-T., Ma, K.-H., Chandrabalan, D., Gwag, J.-G., et al. (2007). PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23, 2155–2162. doi:10.1093/bioinformatics/btm313.

Kim, K. W., Smith, C. A., Daily, M. D., Cort, J. R., Davin, L. B., and Lewis, N. G. (2015b). Trimeric structure of (+)-pinoresinol-forming dirigent protein at 1.95 Å resolution with three isolated active sites. *J. Biol. Chem.* 290, 1308–1318. doi:10.1074/jbc.M114.611780.

Kim, M. K., Jeon, J. H., Fujita, M., Davin, L. B., and Lewis, N. G. (2002b). The western red cedar (Thuja plicata) 8-8′ DIRIGENT family displays diverse expression patterns and conserved monolignol coupling specificity. *Plant Mol. Biol.* 49, 199–214. doi:10.1023/A:1014940930703.

Kim, S.-U., Lee, M.-H., Pae, S.-B., Oh, E.-Y., Kim, J.-I., and Ha, T.-J. (2018). A Sesame Variety “Goenbaek” with Phytophthora Blight Disease Resistance and High Yield. *Korean J. Breed. Sci.* 50, 256–260. doi:10.9787/KJBS.2018.50.3.256.

Kim, S. H., Kim, M. S., Lee, M. S., Park, Y. S., Lee, H. J., Kang, S. ah, et al. (2016). Korean diet: Characteristics and historical background. *J. Ethn. Foods* 3, 26–31. doi:10.1016/j.jef.2016.03.002.

Kirchhoff, H. (2019). Chloroplast ultrastructure in plants. *New Phytol.* 223, 565–574. doi:10.1111/nph.15730.

Kobayashi, T. (1991). “Cytogenetics of Sesame (*Sesamum indicum*),” in *Chromosome Engineering in Plants Genetics, Breeding, Evolution, Part B*, eds. T. Tsuchiya and P. K. Gupta (Elsevier B.V.), 581–592. doi:https://doi.org/10.1016/B978-0-444-88260-8.50036-7.

Köhler, M., Reginato, M., Souza-Chies, T. T., and Majure, L. C. (2020). Insights into chloroplast genome variation across Opuntioideae (Cactaceae). *Front. Plant Sci.* 11. doi:10.1101/2020.03.06.981183.

Kojima, A., Yuasa, I., Kiyomoto, K., and Omura, A. (2017). Composition for promoting collagen production, for promoting elastin production and/or for promoting keratinocyte migration and usage therefor. US Patent No. 9,629,823 B2, 25 April.

Kojima, A., Yuasa, I., Kiyomoto, K., and Omura, A. (2020). Composition for promoting collagen production, for promoting keratinocyte migration and usage therefor. U.S. Patent No. 2016/0175280.

Kolde, R. (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap.

Kraaijeveld, K. (2010). Genome Size and Species Diversification. *Evol. Biol.* 37, 227–233.

doi:10.1007/s11692-010-9093-4.

Kumar, S., Ambreen, H., Variath, M. T., Rao, A. R., Agarwal, M., Kumar, A., et al. (2016). Utilization of Molecular, Phenotypic, and Geographical Diversity to Develop Compact Composite Core Collection in the Oilseed Crop, Safflower (*Carthamus tinctorius* L.) through Maximization Strategy. *Front. Plant Sci.* 7, 1–14. doi:10.3389/fpls.2016.01554.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* 34, 1812–1819. doi:10.1093/molbev/msx116.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001). REPuter: The manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.* 29, 4633–4642. doi:10.1093/nar/29.22.4633.

Kyriakidou, M., Tai, H. H., Anglin, N. L., Ellis, D., and Strömvik, M. V. (2018). Current strategies of polyploid plant genome sequence assembly. *Front. Plant Sci.* 871, 1–15. doi:10.3389/fpls.2018.01660.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res.* 40, 1202–1210. doi:10.1093/nar/gkr1090.

Lan, T., Renner, T., Ibarra-Laclette, E., Farr, K. M., Chang, T. H., Cervantes-Pérez, S. A., et al. (2017). Long-read sequencing uncovers the adaptive topography of a carnivorous plant genome. *Proc. Natl. Acad. Sci. U. S. A.* doi:10.1073/pnas.1702072114.

Lang, D., Zhang, S., Ren, P., Liang, F., Sun, Z., Meng, G., et al. (2020). Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *Gigascience* 9,

1–7. doi:10.1093/gigascience/giaa123.

Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res.* 32, 11–16. doi:10.1093/nar/gkh152.

Laurentin, H., and Karlovsky, P. (2007). AFLP fingerprinting of sesame (*Sesamum indicum* L.) cultivars: Identification, genetic relationship and comparison of AFLP informativeness parameters. *Genet. Resour. Crop Evol.* doi:10.1007/s10722-006-9128-y.

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. *J. Stat. Softw.* 25, 1–18. doi:10.18637/jss.v025.i01.

Lee, K. J., Raveendar, S., Choi, J. S., Gil, J., Lee, J. H., So, Y. S., et al. (2019). Development of chloroplast microsatellite markers for identification of Glycyrrhiza species. *Plant Genet. Resour. Characterisation Util.* 17, 95–98. doi:10.1017/S1479262118000308.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van De Peer, Y., et al. (2002). PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res.* 30, 325–327. doi:10.1093/nar/30.1.325.

Leushkin, E. V., Sutormin, R. A., Nabieva, E. R., Penin, A. A., Kondrashov, A. S., and Logacheva, M. D. (2013). The miniature genome of a carnivorous plant *Genlisea aurea* contains a low number of genes and short non-coding sequences. *BMC Genomics* 14, 11. doi:10.1186/1471-2164-14-476.

Li, B., and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323. doi:10.1186/1471-2105-12-323.

Li, C., Zheng, Y., and Huang, P. (2020a). Molecular markers from the chloroplast genome of rose provide a complementary tool for variety discrimination and profiling. *Sci. Rep.* 10, 1–15. doi:10.1038/s41598-020-68092-1.

Li, D., Dossa, K., Zhang, Y., Wei, X., Wang, L., Zhang, Y., et al. (2018). GWAS uncovers differential genetic bases for drought and salt tolerances in sesame at the germination stage. *Genes (Basel).* 9. doi:10.3390/genes9020087.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698.

Li, L. F., Cushman, S. A., He, Y. X., and Li, Y. (2020b). Genome sequencing and population genomics modeling provide insights into the local adaptation of weeping forsythia. *Hortic. Res.* doi:10.1038/s41438-020-00352-7.

Li, R., Zhao, Y., Sun, Z., Wu, Z., Wang, H., and Fu, C. (2022). Genome-Wide Identification of Switchgrass Laccases Involved in Lignin Biosynthesis and Heavy-Metal Responses.

Li, W., Huai, X., Li, P., Raza, A., Mubarik, M. S., Habib, M., et al. (2021). Genome-wide characterization of glutathione peroxidase (GPX) gene family in rapeseed (*Brassica napus* L.) revealed their role in multiple abiotic stress response and hormone signaling. *Antioxidants* 10. doi:10.3390/antiox10091481.

Liang, M., Davis, E., Gardner, D., Cai, X., and Wu, Y. (2006). Involvement of AtLAC15 in lignin synthesis in seeds and in root elongation of Arabidopsis. *Planta* 224, 1185–1196. doi:10.1007/s00425-006-0300-6.

Liang, Y., Chen, S., Wei, K., Yang, Z., Duan, S., Du, Y., et al. (2020). Chromosome Level Genome Assembly of *Andrographis paniculata*. *Front. Genet.* doi:10.3389/fgene.2020.00701.

Liao, Y., Smyth, G. K., and Shi, W. (2013). The Subread aligner: Fast, accurate and scalable

read mapping by seed-and-vote. *Nucleic Acids Res.* 41. doi:10.1093/nar/gkt214.

Lim, J., Lim, C. W., and Lee, S. C. (2018). The Pepper Late Embryogenesis Abundant Protein, CaDIL1, Positively Regulates Drought Tolerance and ABA Signaling. *Front. Plant Sci.* 9, 1–12. doi:10.3389/fpls.2018.01301.

Liu, H., Xing, M., Yang, W., Mu, X., Wang, X., Lu, F., et al. (2019). Genome-wide identification of and functional insights into the late embryogenesis abundant (LEA) gene family in bread wheat (Triticum aestivum). *Sci. Rep.* 9, 13375. doi:10.1038/s41598-019-49759-w.

Liu, Z., Saarinen, N. M., and Thompson, L. U. (2006). Sesamin Is One of the Major Precursors of Mammalian Lignans in Sesame Seed (*Sesamum indicum*) as Observed In Vitro and in Rats. *J. Nutr.* 136, 906–912. doi:10.1093/jn/136.4.906.

Lomsadze, A., Burns, P. D., and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* 42, e119–e119. doi:10.1093/nar/gku557.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* 25, 0955–0964. doi:10.1093/nar/25.5.0955.

Luo, C., Huang, W., Sun, H., Yer, H., Li, X., Li, Y., et al. (2021). Comparative chloroplast genome analysis of Impatiens species (Balsaminaceae) in the karst area of China: insights into genome evolution and phylogenomic implications. *BMC Genomics* 22, 1–18. doi:10.1186/s12864-021-07807-8.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi:10.1093/nar/gkz268.

Magwanga, R. O., Lu, P., Kirungu, J. N., Dong, Q., Hu, Y., Zhou, Z., et al. (2018). Cotton late embryogenesis abundant (LEA2) genes promote root growth and confer drought stress tolerance in transgenic Arabidopsis thaliana. *G3 Genes, Genomes, Genet.* 8, 2781–2803. doi:10.1534/g3.118.200423.

Mahendra Kumar, C., and Singh, S. A. (2015). Bioactive lignans from sesame (Sesamum indicum L.): evaluation of their antioxidant and antibacterial effects for food applications. *J. Food Sci. Technol.* 52, 2934–2941. doi:10.1007/s13197-014-1334-6.

Majdalawieh, A. F., and Mansour, Z. R. (2019). Sesamol, a major lignan in sesame seeds (Sesamum indicum): Anti-cancer properties and mechanisms of action. *Eur. J. Pharmacol.* 855, 75–89. doi:10.1016/j.ejphar.2019.05.008.

Manning, J. C., and Magee, A. R. (2018). Additional new combinations in Sesamum L. (Pedaliaceae: Sesameae). *Bothalia* 48, 1–2. doi:10.4102/abc.v48i1.2363.

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Comput. Biol.* 14, e1005944. doi:10.1371/journal.pcbi.1005944.

Mariotti, R., Belaj, A., De La Rosa, R., Leòn, L., Brizioli, F., Baldoni, L., et al. (2020). EST–SNP study of olea Europaea L. Uncovers functional polymorphisms between cultivated and wild olives. *Genes (Basel).* doi:10.3390/genes11080916.

Markus, R. L. (1962). Process for Producing Sesamol. U.S. Patent No. 3,058,995, 16 October. Available at: https://patentimages.storage.googleapis.com/4f/0d/c2/eddb447fcc5cf9/US3058995.pdf.

Matsuoka, Y., Yamazaki, Y., Ogihara, Y., and Tsunewaki, K. (2002). Whole chloroplast genome comparison of rice, maize, and wheat: Implications for chloroplast gene diversification and phylogeny of cereals. *Mol. Biol. Evol.* 19, 2084–2091.

doi:10.1093/oxfordjournals.molbev.a004033.

Mayer, A. M., and Staples, R. C. (2002). Laccase: New functions for an old enzyme. *Phytochemistry* 60, 551–565. doi:10.1016/S0031-9422(02)00171-1.

Mayrose, I., and Lysak, M. A. (2021). The Evolution of Chromosome Numbers: Mechanistic Models and Experimental Approaches. *Genome Biol. Evol.* 13, 1–15. doi:10.1093/gbe/evaa220.

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

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

Mekuria, G. T., Collins, G. G., and Sedgley, M. (1999). Genetic variability between different accessions of some common commercial olive cultivars. *J. Hortic. Sci. Biotechnol.* doi:10.1080/14620316.1999.11511114.

Mekuria, G. T., Sedgley, M., Collins, G., and Lavee, S. (2002). Development of a sequence-tagged site for the RAPD marker linked to leaf spot resistance in olive. *J. Am. Soc. Hortic. Sci.* doi:10.21273/jashs.127.4.673.

Mello, B. (2018). Estimating TimeTrees with MEGA and the TimeTree Resource. *Mol. Biol. Evol.* 35, 2334–2342. doi:10.1093/molbev/msy133.

Meng, G., Fan, W., and Rasmussen, S. K. (2021). Characterisation of the class III peroxidase gene family in carrot taproots and its role in anthocyanin and lignin accumulation. *Plant Physiol. Biochem.* 167, 245–256. doi:10.1016/j.plaphy.2021.08.004.

Michael, T. P. (2014). Plant genome size variation: Bloating and purging DNA. *Briefings Funct. Genomics Proteomics* 13, 308–317. doi:10.1093/bfgp/elu005.

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* 34, i142–i150. doi:10.1093/bioinformatics/bty266.

Minh, B. Q., Nguyen, M. A. T., and Von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* 30, 1188–1195. doi:10.1093/molbev/mst024.

Mistry, J., Bateman, A., and Finn, R. D. (2007). Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* 8, 1–14. doi:10.1186/1471-2105-8-298.

Moazzami, A. A., Haese, S. L., and Kamal-Eldin, A. (2007). Lignan contents in sesame seeds and products. *Eur. J. Lipid Sci. Technol.* 109, 1022–1027. doi:10.1002/ejlt.200700057.

Molotoks, A., Smith, P., and Dawson, T. P. (2021). Impacts of land use, population, and climate change on global food security. *Food Energy Secur.* 10, 1–20. doi:10.1002/fes3.261.

Moore, R. C., and Purugganan, M. D. (2003). The early stages of duplicate gene evolution. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15682–15687. doi:10.1073/pnas.2535513100.

Morell, M. K., Kosar-Hashemi, B., Cmiel, M., Samuel, M. S., Chandler, P., Rahman, S., et al. (2003). Barley sex6 mutants lack starch synthase IIa activity and contain a starch with novel properties. *Plant J.* 34, 173–185. doi:10.1046/j.1365-313X.2003.01712.x.

Murata, J., Ono, E., Yoroizuka, S., Toyonaga, H., Shiraishi, A., Mori, S., et al. (2017). Oxidative rearrangement of (+)-sesamin by CYP92B14 co-generates twin dietary lignans in sesame. *Nat. Commun.* 8, 1–10. doi:10.1038/s41467-017-02053-7.

Namiki, M., Kobayashi, T., and Hara, H. (2001). Process of producing sesame lignans an/or

sasame flavors. U.S. Patent No. 6,278,005 B1, 21 August. doi:Aug. 21, 2001.

Nattestad, M., and Schatz, M. C. (2016). Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* 32, 3021–3023. doi:10.1093/bioinformatics/btw369.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics* 25, 1335–1337. doi:10.1093/bioinformatics/btp157.

Nayar, N. M., and Mehra, K. L. (1970). Sesame: Its uses, botany, cytogenetics, and origin. *Econ. Bot.* 24, 20–31. doi:10.1007/BF02860629.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015a). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. (2015b). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* 32, 268–274. doi:10.1093/molbev/msu300.

Nhemachena, C., Nhamo, L., Matchaya, G., Nhemachena, C. R., Muchara, B., Karuaihe, S. T., et al. (2020). Climate Change Impacts on Water and Agriculture Sectors in Southern Africa: Threats and Opportunities for Sustainable Development. *Water* 12, 2673. doi:10.3390/w12102673.

Noguchi, A., Fukui, Y., Iuchi-Okada, A., Kakutani, S., Satake, H., Iwashita, T., et al. (2008). Sequential glucosylation of a furofuran lignan, (+)-sesaminol, by *Sesamum indicum* UGT71A9 and UGT94D1 glucosyltransferases. *Plant J.* 54, 415–427. doi:10.1111/j.1365-313X.2008.03428.x.

Novo-Uzal, E., Fernández-Pérez, F., Herrero, J., Gutiérrez, J., Gómez-Ros, L. V., Bernal, M. Á., et al. (2013). From Zinnia to Arabidopsis: Approaching the involvement of

peroxidases in lignification. *J. Exp. Bot.* 64, 3499–3518. doi:10.1093/jxb/ert221.

Ntwenya, J. E., Kinabo, J., Msuya, J., Mamiro, P., Mamiro, D., Njoghomi, E., et al. (2017). Rich Food Biodiversity Amid Low Consumption of Food Items in Kilosa District, Tanzania. *Food Nutr. Bull.* 38, 501–511. doi:10.1177/0379572117708647.

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., et al. (2013). The Norway spruce genome sequence and conifer genome evolution. *Nature* 497, 579–584. doi:10.1038/nature12211.

Ogasawara, T., Chiba, K., and Tada, M. (1998). "*Sesamum indicum* L. (Sesame): In Vitro Culture, and the Production of Naphthoquinone and Other Secondary Metabolites," in, 366–393. doi:10.1007/978-3-642-58833-4_19.

Ogretmen, B. (2018). Sphingolipid metabolism in cancer signalling and therapy. *Nat. Rev. Cancer* 18, 33–50. doi:10.1038/nrc.2017.96.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., et al. (2020). vegan: Community Ecology Package.

Olvera-Carrillo, Y., Reyes, J. L., and Covarrubias, A. A. (2011). Late embryogenesis abundant proteins: Versatile players in the plant adaptation to water limiting environments. *Plant Signal. Behav.* 6, 586–589. doi:10.4161/psb.6.4.15042.

Ono, E., Nakai, M., Fukui, Y., Tomimori, N., Fukuchi-Mizutani, M., Saito, M., et al. (2006). Formation of two methylenedioxy bridges by a Sesamum CYP81Q protein yielding a furofuran lignan, (+)-sesamin. *Proc. Natl. Acad. Sci. U. S. A.* 103, 10116–10121. doi:10.1073/pnas.0603865103.

Ono, E., Waki, T., Oikawa, D., Murata, J., Shiraishi, A., Toyonaga, H., et al. (2019). Glycoside-specific glycosyltransferases catalyze regio-selective sequential glucosylations for a sesame lignan, sesaminol triglucoside. *Plant J.* 101, 1221–1233.

doi:10.1111/tpj.14586.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46, e126. doi:10.1093/nar/gky730.

Park, I., Yang, S., Choi, G., Jin Kim, W., and Moon, B. C. (2017). The complete chloroplast genome sequences of *Aconitum pseudolaeve* and *Aconitum longecassidatum*, and development of molecular markers for distinguishing species in the *Aconitum subgenus lycoctonum*. *Molecules* 22. doi:10.3390/molecules22112012.

Park, I., Yang, S., Kim, W. J., Noh, P., Lee, H. O., and Moon, B. C. (2018). The complete chloroplast genomes of six ipomoea species and indel mar-ker development for the discrimination of authentic pharbitidis semen (Seeds of I. nil or I. Purpurea). *Front. Plant Sci.* 9, 1–14. doi:10.3389/fpls.2018.00965.

Park, J.-H., Suresh, S., Raveendar, S., Baek, H.-J., Kim, C.-K., Lee, S., et al. (2015). Development and Evaluation of Core Collection Using Qualitative and Quantitative Trait Descriptor in Sesame (*Sesamum indicum* L.) Germplasm. *Korean J. Crop Sci.* 60, 75–84. doi:10.7740/kjcs.2014.60.1.075.

Passardi, F., Longet, D., Penel, C., and Dunand, C. (2004). The class III peroxidase multigenic family in rice and its evolution in land plants. *Phytochemistry* 65, 1879–1893. doi:10.1016/j.phytochem.2004.06.023.

Patel, A. A., Shukla, Y. M., Kumar, S., Sakure, A. A., Parekh, M. J., and Zala, H. N. (2020). Transcriptome analysis for molecular landscaping of genes controlling diterpene andrographolide biosynthesis in *Andrographis paniculata* (Burm . f.) Nees. *3 Biotech*. doi:10.1007/s13205-020-02511-y.

Pathak, N., Bhaduri, A., Bhat, K. V., and Rai, A. K. (2015). Tracking sesamin synthase gene expression through seed maturity in wild and cultivated sesame species - a

domestication footprint. *Plant Biol.* 17, 1039–1046. doi:10.1111/plb.12327.

Patil, C. G., and Hiremath, S. C. (2002). Genome Relations among Octaploid Species of Sesamum L. (Pedaliaceae). *Cytologia (Tokyo).* 67, 403–409. doi:https://doi.org/10.1508/cytologia.67.403.

Patil, C. G., and Hiremath, S. C. (2004). Karyotypic studies in octaploid species of Sesamum L. *J. Cytol. Genet.* 5, 73–76.

Patil, I. (2018). ggstatsplot: "ggplot2" based plots with statistical details. R package version 0.7.0. https://CRAN.R-project.org/package=ggstatsplot.

Pebesma, E., Bivand, R., Racine, E., Sumner, M., Cook, I., Keitt, T., et al. (2021). sf: Simple Features for R. R package version 0.9-7. http://CRAN.R-project.org/package=sf.

Peñalvo, J. L., and López-Romero, P. (2012). Urinary enterolignan concentrations are positively associated with serum HDL cholesterol and negatively associated with serum triglycerides in U.S. adults. *J. Nutr.* 142, 751–756. doi:10.3945/jn.111.150516.

Perkins, M. L., Schuetz, M., Unda, F., Smith, R. A., Sibout, R., Hoffmann, N. J., et al. (2020). Dwarfism of high-monolignol Arabidopsis plants is rescued by ectopic LACCASE overexpression. *Plant Direct* 4, 1–16. doi:10.1002/pld3.265.

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122.

Peterson, J., Dwyer, J., Adlercreutz, H., Scalbert, A., Jacques, P., and McCullough, M. L. (2010). Dietary lignans: Physiology and potential for cardiovascular disease risk reduction. *Nutr. Rev.* 68, 571–603. doi:10.1111/j.1753-4887.2010.00319.x.

Pickel, B., Constantin, M.-A., Pfannstiel, J., Conrad, J., Beifuss, U., and Schaller, A. (2010). An Enantiocomplementary Dirigent Protein for the Enantioselective Laccase-

Catalyzed Oxidative Coupling of Phenols. *Angew. Chemie Int. Ed.* 49, 202–204. doi:10.1002/anie.200904622.

Piot, A., Hackel, J., Christin, P.-A., and Besnard, G. (2018). One-third of the plastid genes evolved under positive selection in PACMAD grasses. *Planta* 247, 255–266. doi:10.1007/s00425-017-2781-x.

Poppius-Levlin, K., Tamminen, T., Kalliola, A., and Ohra-aho, T. (2001). "Characterization of Residual Lignins in Pulps Delignified by Laccase/ N -Hydroxyacetanilide," in, 358–372. doi:10.1021/bk-2001-0785.ch022.

Pourcel, L., Routaboul, J. M., Cheynier, V., Lepiniec, L., and Debeaujon, I. (2007). Flavonoid oxidation in plants: from biochemical properties to physiological functions. *Trends Plant Sci.* 12, 29–36. doi:10.1016/j.tplants.2006.11.006.

Powell, W., Morgante, M., McDevitt, R., Vendramin, G. G., and Rafalski, J. A. (1995). Polymorphic simple sequence repeat regions in chloroplast genomes: Applications to the population genetics of pines. *Proc. Natl. Acad. Sci. U. S. A.* 92, 7759–7763. doi:10.1073/pnas.92.17.7759.

POWO (2022). Plants of the World Online. Facilitated by the Royal Botanic Gardens, Kew. Published on the Internet; http://www.plantsoftheworldonline.org/ Retrieved 01 January 2022.

Putnam, N. H., O'Connell, B. L., Stites, J. C., Rice, B. J., Blanchette, M., Calef, R., et al. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* 26, 342–350. doi:10.1101/gr.193474.115.

Qui, K., Zhou, H., Pan, H., Sheng, Y., Yu, H., Xie, Q., et al. (2022). Genome-wide identification and functional analysis of the peach (*P. persica*) laccase gene family reveal members potentially involved in endocarp lignification. *Trees - Struct. Funct.*

doi:10.1007/s00468-022-02296-y.

R Core Team (2020). R: A Language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Raghavan, T. S., and Krishnamurthy, K. V. (1947). Cytogenetical studies in sesamum. Part I. Cytology of the parents, *Sesamum orientale* Linn, and *Sesamum prostratum* Retz. and the cytology of the sterile hybrid between them and of the fertile amphidiploid. in *Proceedings of the Indian Academy of Sciences - Section B*, 236–275. doi:https://doi.org/10.1007/BF03051810.

Rallo, P., Dorado, G., and Martín, A. (2000). Development of simple sequence repeats (SSRs) in olive tree (*Olea europaea* L.). *Theor. Appl. Genet.* doi:10.1007/s001220051571.

Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* 11, 1432. doi:10.1038/s41467-020-14998-3.

Ren, L. L., Liu, Y. J., Liu, H. J., Qian, T. T., Qi, L. W., Wang, X. R., et al. (2014). Subcellular relocalization and positive selection play key roles in the retention of duplicate genes of Populus class III peroxidase family. *Plant Cell* 26, 2404–2419. doi:10.1105/tpc.114.124750.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592, 737–746. doi:10.1038/s41586-021-03451-0.

Ring, L., Yeh, S. Y., Hücherig, S., Hoffmann, T., Blanco-Portales, R., Fouche, M., et al. (2013). Metabolic interaction between anthocyanin and lignin biosynthesis is associated with peroxidase FaPRX27 in strawberry fruit. *Plant Physiol.* 163, 43–60.

doi:10.1104/pp.113.222778.

Rodríguez-García, C., Sánchez-Quesada, C., Toledo, E., Delgado-Rodríguez, M., and Gaforio, J. (2019). Naturally Lignan-Rich Foods: A Dietary Tool for Health Promotion? *Molecules* 24, 917. doi:10.3390/molecules24050917.

Ros Barceló, A., Gómez Ros, L. V., Gabaldón, C., López-Serrano, M., Pomar, F., Carrión, J. S., et al. (2004). Basic peroxidases: The gateway for lignin evolution? *Phytochem. Rev.* 3, 61–78. doi:10.1023/B:PHYT.0000047803.49815.1a.

Roy, J., Blervacq, A. S., Créach, A., Huss, B., Hawkins, S., and Neutelings, G. (2017). Spatial regulation of monolignol biosynthesis and laccase genes control developmental and stress-related lignin in flax. *BMC Plant Biol.* 17, 1–20. doi:10.1186/s12870-017-1072-9.

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J. C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S. E., et al. (2017). DnaSP 6: DNA sequence polymorphism analysis of large data sets. *Mol. Biol. Evol.* 34, 3299–3302. doi:10.1093/molbev/msx248.

Ryu, C. (2021). dlookr: Tools for data diagnosis, exploration, transformation. R package version 0.4.2. https://CRAN.R-project.org/package=dlookr. Available at: https://cran.r-project.org/package=dlookr.

Sarfraz, I., Rasul, A., Jabeen, F., Younis, T., Zahoor, M. K., Arshad, M., et al. (2017). Fraxinus: A Plant with Versatile Pharmacological and Biological Activities. *Evidence-based Complement. Altern. Med.* 2017. doi:10.1155/2017/4269868.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485, 635–641. doi:10.1038/nature11119.

Schäferhoff, B., Fleischmann, A., Fischer, E., Albach, D. C., Borsch, T., Heubl, G., et al.

(2010). Towards resolving lamiales relationships: Insights from rapidly evolving chloroplast sequences. *BMC Evol. Biol.* 10. doi:10.1186/1471-2148-10-352.

Seeman, T., and Booth, T. (2013). Barrnap: Basic Rapid Ribosomal RNA Predictor [Internet]. http://github.com/tseemann/barrnap. Accessed April 15, 2020.

Sefc, K. M., Lopes, M. S., Mendonça, D., Rodrigues Dos Santos, M., Laimer Da Câmara Machado, M., and Da Câmara Machado, A. (2000). Identification of microsatellite loci in olive (Olea europaea) and their characterization in Italian and Iberian olive trees. *Mol. Ecol.* doi:10.1046/j.1365-294X.2000.00954.x.

Sehr, E. M., Okello-Anyanga, W., Hasel-Hohl, K., Burg, A., Gaubitzer, S., Rubaihayo, P. R., et al. (2016). Assessment of genetic diversity amongst Ugandan sesame (*Sesamum indicum* L.) landraces based on agromorphological traits and genetic markers. *J. Crop Sci. Biotechnol.* doi:10.1007/s12892-015-0105-x.

Shannon, C. E., and Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.

Sharma, S. N., Sinha, R. K., Sharma, D. K., and Jha, Z. (2009). Assessment of intra-specific variability at morphological, molecular and biochemical level of Andrographis paniculata (Kalmegh). *Curr. Sci.*

Sharp, P. M., and Cowe, E. (1991). Synonymous codon usage in *Saccharomyces cerevisiae*. *Yeast* 7, 657–678. doi:10.1002/yea.320070702.

Shen, X., Guo, X., Guo, X., Zhao, D., Zhao, W., Chen, J., et al. (2017). PacMYBA, a sweet cherry R2R3-MYB transcription factor, is a positive regulator of salt stress tolerance and pathogen resistance. *Plant Physiol. Biochem.* 112, 302–311. doi:10.1016/j.plaphy.2017.01.015.

Shimoyoshi, S., Takemoto, D., Ono, Y., Kitagawa, Y., Shibata, H., Tomono, S., et al. (2019).

Sesame lignans suppress age-related cognitive decline in senescence-accelerated mice. *Nutrients* 11. doi:10.3390/nu11071582.

Silva, S. R., Moraes, A. P., Penha, H. A., Julião, M. H. M., Domingues, D. S., Michael, T. P., et al. (2020). The terrestrial carnivorous plant utricularia reniformis sheds light on environmental and life-form genome plasticity. *Int. J. Mol. Sci.* 21, 1–24. doi:10.3390/ijms21010003.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi:10.1093/bioinformatics/btv351.

Simpson, E. H. (1949). Measurement of Diversity. *Nature* 163, 688–688. doi:10.1038/163688a0.

Slater, G. S. C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 1–11. doi:10.1186/1471-2105-6-31.

Smale, M., and Jamora, N. (2020). Valuing genebanks. *Food Secur.* 12, 905–918. doi:10.1007/s12571-020-01034-x.

Smith, D. R. (2020). Can Green Algal Plastid Genome Size Be Explained by DNA Repair Mechanisms? *Genome Biol. Evol.* 12, 3797–3802. doi:10.1093/gbe/evaa012.

Sok, D.-E., Cui, H., and Kim, M. (2012). Isolation and Boactivities of Furfuran Type Lignan Compounds from Edible Plants. *Recent Patents Food, Nutr. Agric.* 1, 87–95. doi:10.2174/2212798410901010087.

Sollars, E. S. A., Harper, A. L., Kelly, L. J., Sambles, C. M., Ramirez-Gonzalez, R. H., Swarbreck, D., et al. (2017). Genome sequence and genetic diversity of European ash trees. *Nature* 541, 212–216. doi:10.1038/nature20786.

Sovetgul, A., Oh, E., Kulkarni, K. P., Lee, M. H., and In, J. (2020). A Combinatorial Approach of Biparental QTL Mapping and Genome-Wide Association Analysis Identifies Candidate Ge. *bioRxiv*. doi:10.1101/2020.03.18.996637.

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

Stanke, M., Diekhans, M., Baertsch, R., and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24, 637–644. doi:10.1093/bioinformatics/btn013.

Stevens, P. F. (2012). Angiosperm Phylogeny Website. *Version 12, July 2012*, Page last updated: 05/12/2015. Available at: http://www.mobot.org/MOBOT/research/APweb/.

Subramanian, M. (2003). Wide crosses and chromosome behaviour in Sesamum. *Madras Agric. J.* 90, 1–15.

Subramanian, S., and Subramanian, M. (1994). Correlation Studies and Path Coefficient Analysis in Sesame (*Sesamum indicum* L.). *J. Agron. Crop Sci.* 173, 241–248. doi:10.1111/j.1439-037X.1994.tb00560.x.

Sun, Q., Wedick, N. M., Pan, A., Townsend, M. K., Cassidy, A., Franke, A. A., et al. (2014). Gut microbiota metabolites of dietary lignans and risk of type 2 diabetes: A prospective investigation in two cohorts of U.S. women. *Diabetes Care* 37, 1287–1295. doi:10.2337/dc13-2513.

Sun, W., Leng, L., Yin, Q., Xu, M. M., Huang, M., Xu, Z., et al. (2019). The genome of the medicinal plant *Andrographis paniculata* provides insight into the biosynthesis of the bioactive diterpenoid neoandrographolide. *Plant J.* doi:10.1111/tpj.14162.

Suyama, M., Torrents, D., and Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34, W609–W612. doi:10.1093/nar/gkl315.

Ta, T. D., Waminal, N. E., Nguyen, T. H., Pellerin, R. J., and Kim, H. H. (2021). Comparative FISH analysis of Senna tora tandem repeats revealed insights into the chromosome dynamics in Senna. *Genes Genomics* 43, 237–249. doi:10.1007/s13258-021-01051-w.

Tapiero, H., Townsend, D. ., and Tew, K. . (2003). The antioxidant role of selenium and seleno-compounds. *Biomed. Pharmacother.* 57, 134–144. doi:10.1016/S0753-3322(03)00035-0.

Taranto, F., D'Agostino, N., Pavan, S., Fanelli, V., di Rienzo, V., Sabetta, W., et al. (2018). Single nucleotide polymorphism (SNP) diversity in an olive germplasm collection. in *Acta Horticulturae* (International Society for Horticultural Science (ISHS), Leuven, Belgium), 27–32. doi:10.17660/ActaHortic.2018.1199.5.

Teichmann, T., and Muhr, M. (2015). Shaping plant architecture. *Front. Plant Sci.* 6, 1–18. doi:10.3389/fpls.2015.00233.

Teklu, D. H., Shimelis, H., Tesfaye, A., Mashilo, J., Zhang, X., Zhang, Y., et al. (2021). Genetic Variability and Population Structure of Ethiopian Sesame (*Sesamum indicum* L.) Germplasm Assessed through Phenotypic Traits and Simple Sequence Repeats Markers. *Plants* 10, 1129. doi:10.3390/plants10061129.

Tenaillon, M. I., Hollister, J. D., and Gaut, B. S. (2010). A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* 15, 471–478. doi:10.1016/j.tplants.2010.05.003.

Tian, F., Yang, D. C., Meng, Y. Q., Jin, J., and Gao, G. (2020). PlantRegMap: Charting

functional regulatory maps in plants. *Nucleic Acids Res.* 48, D1104–D1113. doi:10.1093/nar/gkz1020.

Tillich, M., Lehwark, P., Pellizzer, T., Ulbricht-Jones, E. S., Fischer, A., Bock, R., et al. (2017). GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* 45, W6–W11. doi:10.1093/nar/gkx391.

Tiwari, G., Singh, R., Singh, N., Choudhury, D. R., Paliwal, R., Kumar, A., et al. (2016). Study of arbitrarily amplified (RAPD and ISSR) and gene targeted (SCoT and CBDP) markers for genetic diversity and population structure in Kalmegh [*Andrographis paniculata* (Burm. f.) Nees]. *Ind. Crops Prod.* doi:10.1016/j.indcrop.2016.03.031.

Tognolli, M., Penel, C., Greppin, H., and Simon, P. (2002). Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*. *Gene* 288, 129–138. doi:10.1016/S0378-1119(02)00465-1.

Tripathi, A. D., Mishra, R., Maurya, K. K., Singh, R. B., and Wilson, D. W. (2018). "Estimates for world population and global food availability for global health," in *The role of functional food security in global health*, eds. R. Watson, R. Singh, and T. Takahashi (Academic Press, Cambridge), 3–24.

Trösch, R., Barahimipour, R., Gao, Y., Badillo-Corona, J. A., Gotsmann, V. L., Zimmer, D., et al. (2018). Commonalities and differences of chloroplast translation in a green alga and land plants. *Nat. Plants* 4, 564–575. doi:10.1038/s41477-018-0211-0.

Tugbaeva, A., Ermoshin, A., Plotnikov, D., Wuriyanghan, H., and Kiseleva, I. (2021). Role of Class III Peroxidases in Stem Lignification of *Zinnia elegans* Jacq. 22. doi:10.3390/iecps2020-08847.

Turlapati, P. V., Kim, K. W., Davin, L. B., and Lewis, N. G. (2011). The laccase multigene family in *Arabidopsis thaliana*: Towards addressing the mystery of their gene

function(s). *Planta* 233, 439–470. doi:10.1007/s00425-010-1298-3.

Udall, J. A., Long, E., Ramaraj, T., Conover, J. L., Yuan, D., Grover, C. E., et al. (2019). The Genome Sequence of *Gossypioides kirkii* Illustrates a Descending Dysploidy in Plants. *Front. Plant Sci.* 10, 1–10. doi:10.3389/fpls.2019.01541.

Uncu, A. O., Frary, A., Karlovsky, P., and Doganlar, S. (2016). High-throughput single nucleotide polymorphism (SNP) identification and mapping in the sesame (*Sesamum indicum* L.) genome with genotyping by sequencing (GBS) analysis. *Mol. Breed.* doi:10.1007/s11032-016-0604-6.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., et al. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res.* 40, e115–e115. doi:10.1093/nar/gks596.

Unver, T., Wu, Z., Sterck, L., Turktas, M., Lohaus, R., Li, Z., et al. (2017). Genome of wild olive and the evolution of oil biosynthesis. *Proc. Natl. Acad. Sci.*, 201708621. doi:10.1073/pnas.1708621114.

Uzun, B., and Çağırgan, M. İ. (2006). Comparison of determinate and indeterminate lines of sesame for agronomic traits. *F. Crop. Res.* 96, 13–18. doi:10.1016/j.fcr.2005.04.017.

Van Buren, R., Wai, C. M., Pardo, J., Giarola, V., Ambrosini, S., Song, X., et al. (2018). Desiccation tolerance evolved through gene duplication and network rewiring in Lindernia. *Plant Cell* 30, 2943–2958. doi:10.1105/tpc.18.00517.

Vining, K. J., Johnson, S. R., Ahkami, A., Lange, I., Parrish, A. N., Trapp, S. C., et al. (2017). Draft Genome Sequence of *Mentha longifolia* and Development of Resources for Mint Cultivar Improvement. *Mol. Plant* 10, 323–339. doi:10.1016/j.molp.2016.10.018.

Visavadiya, N. P., and Narasimhacharya, A. V. R. L. (2008). Sesame as a hypocholesteraemic and antioxidant dietary component. *Food Chem. Toxicol.* 46,

1889–1895. doi:10.1016/j.fct.2008.01.012.

Waminal, N. E., Pellerin, R. J., Kang, S.-H., and Kim, H. H. (2021). Chromosomal Mapping of Tandem Repeats Revealed Massive Chromosomal Rearrangements and Insights Into *Senna tora* Dysploidy. *Front. Plant Sci.* 12, 1–13. doi:10.3389/fpls.2021.629898.

Wan, F., Zhang, L., Tan, M., Wang, X., Wang, G. L., Qi, M., et al. (2022). Genome-wide identification and characterization of laccase family members in eggplant (*Solanum melongena* L.). *PeerJ* 10, 1–24. doi:10.7717/peerj.12922.

Wang, L., Dossa, K., You, J., Zhang, Y., Li, D., Zhou, R., et al. (2021). High-resolution temporal transcriptome sequencing unravels ERF and WRKY as the master players in the regulatory networks underlying sesame responses to waterlogging and recovery. *Genomics* 113, 276–290. doi:10.1016/j.ygeno.2020.11.022.

Wang, L., Dossou, S. S. K., Wei, X., Zhang, Y., Li, D., Yu, J., et al. (2020). Transcriptome dynamics during black and white sesame (*Sesamum indicum* L.) seed development and identification of candidate genes associated with black pigmentation. *Genes (Basel).* 11, 1–14. doi:10.3390/genes11121399.

Wang, L., Xia, Q., Zhang, Y., Zhu, X., Zhu, X., Li, D., et al. (2016). Updated sesame genome assembly and fine mapping of plant height and seed coat color QTLs using a new high-density genetic map. *BMC Genomics* 17, 31. doi:10.1186/s12864-015-2316-4.

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

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

doi:10.1186/gb-2014-15-2-r39.

Wang, L., Zhang, Y., Li, D., Dossa, K., Wang, M. L., Zhou, R., et al. (2019). Gene expression profiles that shape high and low oil content sesames. *BMC Genet.* 20, 1–11. doi:10.1186/s12863-019-0747-7.

Wang, L., Zhang, Y., Qi, X., Gao, Y., and Zhang, X. (2012a). Development and characterization of 59 polymorphic cDNA-SSR markers for the edible oil crop *Sesamum indicum* (pedaliaceae). *Am. J. Bot.* doi:10.3732/ajb.1200081.

Wang, W., Yu, H., Wang, J., Lei, W., Gao, J., Qiu, X., et al. (2017). The Complete Chloroplast Genome Sequences of the Medicinal Plant *Forsythia suspensa* (Oleaceae). *Int. J. Mol. Sci.* 18, 2288. doi:10.3390/ijms18112288.

Wang, Y., Tang, H., Debarry, J. D., Tan, X., Li, J., Wang, X., et al. (2012b). MCScanX: A toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* 40, 1–14. doi:10.1093/nar/gkr1293.

Wang, Y., Wang, Q., Zhao, Y., Han, G., and Zhu, S. (2015b). Systematic analysis of maize class III peroxidase gene family reveals a conserved subfamily involved in abiotic stress response. *Gene* 566, 95–108. doi:10.1016/j.gene.2015.04.041.

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

Wei, X., Liu, K., Zhang, Y., Feng, Q., Wang, L., Zhao, Y., et al. (2015). Genetic discovery for oil production and quality in sesame. *Nat. Commun.* 6, 1–10. doi:10.1038/ncomms9609.

Wei, X., Wang, L., Zhang, Y., Qi, X., Wang, X., Ding, X., et al. (2014). Development of

simple sequence repeat (SSR) markers of sesame (*Sesamum indicum*) from a genome survey. *Molecules* 19, 5150–5162. doi:10.3390/molecules19045150.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Second edi. Springer International Publishing.

Wijarat P (2012). Genetic evaluation of *Andrographis paniculata* (Burm. f.) Nees revealed by SSR, AFLP and RAPD markers. *J. Med. Plants Res.* doi:10.5897/jmpr11.1025.

Willson, M. F., and Traveset, A. (2000). "The ecology of seed dispersal," in *Seeds. The Ecology of Regeneration in Plant Communities*, ed. M. Fenner (CABI Publishing, Wallingford, UK.), 410. doi:10.1079/SSR2003142.

Winterfeld, G., Ley, A., Hoffmann, M. H., Paule, J., and Röser, M. (2020). Dysploidy and polyploidy trigger strong variation of chromosome numbers in the prayer-plant family (Marantaceae). *Plant Syst. Evol.* 306, 36. doi:10.1007/s00606-020-01663-x.

Wu, C., Ding, X., Ding, Z., Tie, W., Yan, Y., Wang, Y., et al. (2019a). The class III peroxidase (POD) gene family in cassava: Identification, phylogeny, duplication, and expression. *Int. J. Mol. Sci.* 20, 1–17. doi:10.3390/ijms20112730.

Wu, K., Liu, H., Yang, M., Tao, Y., Ma, H., Wu, W., et al. (2014). High-density genetic map construction and QTLs analysis of grain yield-related traits in Sesame (*Sesamum indicum* L.) based on RAD-Seq techonology. *BMC Plant Biol.* 14, 1–14. doi:10.1186/s12870-014-0274-7.

Wu, L., Nie, L., Wang, Q., Xu, Z., Wang, Y., He, C., et al. (2021). Comparative and phylogenetic analyses of the chloroplast genomes of species of Paeoniaceae. *Sci. Rep.* 11, 1–16. doi:10.1038/s41598-021-94137-0.

Wu, M.-S., Aquino, L. B. B., Barbaza, M. Y. U., Hsieh, C.-L., De Castro-Cruz, K. A., Yang, L.-L., et al. (2019b). Anti-Inflammatory and Anticancer Properties of Bioactive

Compounds from *Sesamum indicum* L.—A Review. *Molecules* 24, 4426. doi:10.3390/molecules24244426.

Wu, S. B., Collins, G., and Sedgley, M. (2004). A molecular linkage map of olive (*Olea europaea* L.) based on RAPD, microsatellite, and SCAR markers. *Genome*. doi:10.1139/g03-091.

Xiao, H., Wang, C., Khan, N., Chen, M., Fu, W., Guan, L., et al. (2020). Genome-wide identification of the class III POD gene family and their expression profiling in grapevine (*Vitis vinifera* L). *BMC Genomics* 21, 1–13. doi:10.1186/s12864-020-06828-z.

Xiao, L., Yang, G., Zhang, L., Yang, X., Zhao, S., Ji, Z., et al. (2015). The resurrection genome of *Boea hygrometrica* : A blueprint for survival of dehydration. *Proc. Natl. Acad. Sci.* 112, 5833–5837. doi:10.1073/pnas.1505811112.

Yamada, D., Kato, M., Ono, Y., Tomimori, N., Nishiumi, T., and Nakahara, K. (2014). Oil-in-water emulsions containing lignan-class compounds and compositions containing the same. U.S. Patent No. 8,685,455 B2, 1 April.

Yamada, D., Kato, M., Ono, Y., Tomimori, N., Nishiumi, T., and Nakahara, K. (2020). Oil-in-Water Emulsions Containing Lignan-Class Compounds and Compositions Containing the Same. U.S. Patent No. 8,685,455.

Yang, T., Zhang, P., Pan, J., Amanullah, S., Luan, F., Han, W., et al. (2022). Genome-Wide Analysis of the Peroxidase Gene Family and Verification of Lignin Synthesis-Related Genes in Watermelon. *Int. J. Mol. Sci.* 23. doi:10.3390/ijms23020642.

Yang, X., Yuan, J., Luo, W., Qin, M., Yang, J., Wu, W., et al. (2020). Genome-Wide Identification and Expression Analysis of the Class III Peroxidase Gene Family in Potato (*Solanum tuberosum* L.). *Front. Genet.* 11, 1–15.

doi:10.3389/fgene.2020.593577.

Yang, Z. (1997). Paml: A program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* 13, 555–556. doi:10.1093/bioinformatics/13.5.555.

Yi, D. K., and Kim, K. J. (2012). Complete chloroplast genome sequences of important oilseed crop *Sesamum indicum* L. *PLoS One* 7. doi:10.1371/journal.pone.0035872.

Yingzhong, Z., and Yishou, W. (2002). Genotypic correlations and path coefficient analysis in sesame. *Sesame safflower Newsl.* 17, 10–12.

Yokota, T., Matsuzaki, Y., Koyama, M., Hitomi, T., Kawanaka, M., Enoki-Konishi, M., et al. (2007). Sesamin, a lignan of sesame, down-regulates cyclin D1 protein expression in human tumor cells. *Cancer Sci.* 98, 1447–1453. doi:10.1111/j.1349-7006.2007.00560.x.

Yoshida, K., Kaothien, P., Matsui, T., Kawaoka, A., and Shinmyo, A. (2003). Molecular biology and application of plant peroxidase genes. *Appl. Microbiol. Biotechnol.* 60, 665–670. doi:10.1007/s00253-002-1157-7.

You, J. W., Rho, H. S., Kim, D. H., Chang, I. S., and Lee, O. S. (2011a). Sesamol Derivatives and Their Salts, the Process for Preparing the Same, and the Skin External Composition Containing the Same. U.S. Patent No. 7,943,599.

You, J. W., Rho, H. S., Kim, D. H., Chang, I. S., and Lee, O. S. (2011b). Sesamol derivatives or their salts, the process for preparing the same, and the skin external composition containing the same. U.S. Patent No. 7,943,599 B2, 17 May.

Yu, H., and Li, J. (2021). Short- and long-term challenges in crop breeding. *Natl. Sci. Rev.* 8, 2021. doi:10.1093/nsr/nwab002.

Yu, J., Dossa, K., Wang, L., Zhang, Y., Wei, X., Liao, B., et al. (2017). PMDBase: A database for studying microsatellite DNA and marker development in plants. *Nucleic*

Acids Res. doi:10.1093/nar/gkw906.

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., et al. (2019a). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol. J.* 17, 881–892. doi:10.1111/pbi.13022.

Yu, J., Golicz, A. A., Lu, K., Dossa, K., Zhang, Y., Chen, J., et al. (2019b). Insight into the evolution and functional characteristics of the pan-genome assembly from sesame landraces and modern cultivars. *Plant Biotechnol. J.* 17, 881–892. doi:10.1111/pbi.13022.

Zamora-Ros, R., Agudo, A., Luján-Barroso, L., Romieu, I., Ferrari, P., Knaze, V., et al. (2012). Dietary flavonoid and lignan intake and gastric adenocarcinoma risk in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *Am. J. Clin. Nutr.* 96, 1398–1408. doi:10.3945/ajcn.112.037358.

Zhang, H., Li, C., Miao, H., and Xiong, S. (2013a). Insights from the complete chloroplast genome into the evolution of *Sesamum indicum* L. *PLoS One* 8. doi:10.1371/journal.pone.0080508.

Zhang, H., Miao, H., Li, C., Wei, L., Duan, Y., Ma, Q., et al. (2016). Ultra-dense SNP genetic map construction and identification of SiDt gene controlling the determinate growth habit in *Sesamum indicum* L. *Sci. Rep.* 6. doi:10.1038/srep31556.

Zhang, H., Miao, H., Wang, L., Qu, L., Liu, H., Wang, Q., et al. (2013b). Genome sequencing of the important oilseed crop *Sesamum indicum* L. *Genome Biol.* 14, 1–9. doi:10.1186/gb-2013-14-1-401.

Zhang, H., Miao, H., Wang, L., Qu, L., Liu, H., Wang, Q., et al. (2013c). Genome sequencing of the important oilseed crop *Sesamum indicum* L. *Genome Biol.* 14, 401.

doi:10.1186/gb-2013-14-1-401.

Zhang, Y. C., Yu, Y., Wang, C. Y., Li, Z. Y., Liu, Q., Xu, J., et al. (2013d). Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nat. Biotechnol.* 31, 848–852. doi:10.1038/nbt.2646.

Zhang, Y., Shen, Q., Leng, L., Zhang, D., Chen, S., Shi, Y., et al. (2021). Incipient diploidization of the medicinal plant Perilla within 10,000 years. *Nat. Commun.* 12, 5508. doi:10.1038/s41467-021-25681-6.

Zhang, Y., Wang, L., Xin, H., Li, D., Ma, C., Ding, X., et al. (2013e). 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.* doi:10.1186/1471-2229-13-141.

Zhao, Q., Yang, J., Cui, M.-Y., Liu, J., Fang, Y., Yan, M., et al. (2019). The Reference Genome Sequence of Scutellaria baicalensis Provides Insights into the Evolution of Wogonin Biosynthesis. *Mol. Plant* 12, 935–950. doi:10.1016/j.molp.2019.04.002.

Zhou, R., Dossa, K., Li, D., Yu, J., You, J., Wei, X., et al. (2018). Genome-Wide Association Studies of 39 Seed Yield-Related Traits in Sesame (*Sesamum indicum* L.). *Int. J. Mol. Sci.* 19, 2794. doi:10.3390/ijms19092794.

Zhou, T., Zhu, H., Wang, J., Xu, Y., Xu, F., and Wang, X. (2020). Complete chloroplast genome sequence determination of Rheum species and comparative chloroplast genomics for the members of Rumiceae. *Plant Cell Rep.* 39, 811–824. doi:10.1007/s00299-020-02532-0.

감사의 글

이 논문을 쓰는 동안 많은 지원과 도움을 받았습니다. 박사학위를 받을 수 있게 해준 국립농업과학원에 감사드리고, 안병옥 과장님과 강상호 연구관님의 많은 도움과 리더십 노력에 감사드립니다. 또한, 연구에 대한 질문과 연구의 전문성을 배우는데 도움을 주신 이근표 연구관님과 김정구 연구사님에게 감사의 말씀드립니다. 이 논문을 완성하는데 기술지원을 해주신 김성업 연구사님과 (주)DNA케어 유의수 박사님께 감사드리고, 박사학위 과정을 성공적으로 마칠 수 있도록 소중한 지도를 해주신 전북대학교의 정남진 교수님, 모영준 교수님께 감사의 말씀드립니다. 여러분들의 통찰력 있는 피드백은 연구에 대한 생각을 더 날카롭게 하고 연구를 더 높은 수준으로 끌어올릴 수 있도록 해주셨습니다. 마지막으로, 우리 연구실 선생님들과 농업생명자원부 유전체과 선생님들 모두에게 감사인사 드립니다.

2023년 1월

예도몬 올림