Compendium of Plant Genomes
Series Editor: Chittaranjan Kole

Maizura Ithnin
Ahmad Kushairi *Editors*

The Oil Palm Genome

Springer

Compendium of Plant Genomes

Series Editor

Chittaranjan Kole, Raja Ramanna Fellow, Government of India,
ICAR-National Research Center on Plant Biotechnology, Pusa,
New Delhi, India

Whole-genome sequencing is at the cutting edge of life sciences in the new millennium. Since the first genome sequencing of the model plant Arabidopsis thaliana in 2000, whole genomes of about 100 plant species have been sequenced and genome sequences of several other plants are in the pipeline. Research publications on these genome initiatives are scattered on dedicated web sites and in journals with all too brief descriptions. The individual volumes elucidate the background history of the national and international genome initiatives; public and private partners involved; strategies and genomic resources and tools utilized; enumeration on the sequences and their assembly; repetitive sequences; gene annotation and genome duplication. In addition, synteny with other sequences, comparison of gene families and most importantly potential of the genome sequence information for gene pool characterization and genetic improvement of crop plants are described.

Interested in editing a volume on a crop or model plant? Please contact Prof. C. Kole, Series Editor, at ckoleorg@gmail.com

More information about this series at http://www.springer.com/series/11805

Maizura Ithnin • Ahmad Kushairi
Editors

The Oil Palm Genome

Springer

Editors
Maizura Ithnin
Advanced Biotechnology
and Breeding Centre
Malaysian Palm Oil Board (MPOB)
Kajang, Malaysia

Ahmad Kushairi
Malaysian Palm Oil Board (MPOB)
Kajang, Malaysia

ISSN 2199-4781 ISSN 2199-479X (electronic)
Compendium of Plant Genomes
ISBN 978-3-030-22548-3 ISBN 978-3-030-22549-0 (eBook)
https://doi.org/10.1007/978-3-030-22549-0

© Springer Nature Switzerland AG 2020
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

This book series is dedicated to my wife Phullara and our children Sourav and Devleena

Chittaranjan Kole

Preface to the Series

Genome sequencing has emerged as the leading discipline in the plant sciences coinciding with the start of the new century. For much of the twentieth century, plant geneticists were only successful in delineating putative chromosomal location, function, and changes in genes indirectly through the use of a number of "markers" physically linked to them. These included visible or morphological, cytological, protein, and molecular or DNA markers. Among them, the first DNA marker, the RFLPs, introduced a revolutionary change in plant genetics and breeding in the mid-1980s, mainly because of their infinite number and thus potential to cover maximum chromosomal regions, phenotypic neutrality, absence of epistasis, and codominant nature. An array of other hybridization-based markers, PCR-based markers, and markers based on both, facilitated construction of genetic linkage maps, mapping of genes controlling simply inherited traits, and even gene clusters (QTLs) controlling polygenic traits in a large number of model and crop plants. During this period, a number of new mapping populations beyond F2 were utilized and a number of computer programmes were developed for map construction, mapping of genes, and mapping of polygenic clusters or QTLs. Molecular markers were also used in the studies of evolution and phylogenetic relationship, genetic diversity, DNA fingerprinting, and map-based cloning. Markers tightly linked to the genes were used in crop improvement employing the so-called marker-assisted selection. These strategies of molecular genetic mapping and molecular breeding made a spectacular impact during the last one and a half decades of the twentieth century. But still they remained "indirect" approaches for elucidation and utilization of plant genomes since much of the chromosomes remained unknown and the complete chemical depiction of them was yet to be unravelled.

Physical mapping of genomes was the obvious consequence that facilitated the development of the "genomic resources" including BAC and YAC libraries to develop physical maps in some plant genomes. Subsequently, integrated genetic–physical maps were also developed in many plants. This led to the concept of structural genomics. Later on, emphasis was laid on EST and transcriptome analysis to decipher the function of the active gene sequences leading to another concept defined as functional genomics. The advent of techniques of bacteriophage gene and DNA sequencing in the 1970s was extended to facilitate sequencing of these genomic resources in the last decade of the twentieth century.

As expected, sequencing of chromosomal regions would have led to too much data to store, characterize, and utilize with the-then available computer software could handle. But the development of information technology made the life of biologists easier by leading to a swift and sweet marriage of biology and informatics, and a new subject was born—bioinformatics.

Thus, the evolution of the concepts, strategies, and tools of sequencing and bioinformatics reinforced the subject of genomics—structural and functional. Today, genome sequencing has travelled much beyond biology and involves biophysics, biochemistry, and bioinformatics!

Thanks to the efforts of both public and private agencies, genome sequencing strategies are evolving very fast, leading to cheaper, quicker, and automated techniques right from clone-by-clone and whole-genome shotgun approaches to a succession of second-generation sequencing methods. The development of software of different generations facilitated this genome sequencing. At the same time, newer concepts and strategies were emerging to handle sequencing of the complex genomes, particularly the polyploids.

It became a reality to chemically—and so directly—define plant genomes, popularly called whole-genome sequencing or simply genome sequencing.

The history of plant genome sequencing will always cite the sequencing of the genome of the model plant Arabidopsis thaliana in 2000 that was followed by sequencing the genome of the crop and model plant rice in 2002. Since then, the number of sequenced genomes of higher plants has been increasing exponentially, mainly due to the development of cheaper and quicker genomic techniques and, most importantly, the development of collaborative platforms such as national and international consortia involving partners from public and/or private agencies.

As I write this Preface for the first volume of the new series *Compendium of Plant Genomes*, a net search tells me that complete or nearly complete whole-genome sequencing of 45 crop plants, eight crop and model plants, eight model plants, 15 crop progenitors and relatives, and three basal plants is accomplished, the majority of which are in the public domain. This means that we nowadays know many of our model and crop plants chemically, i.e. directly, and we may depict them and utilize them precisely better than ever. Genome sequencing has covered all groups of crop plants. Hence, information on the precise depiction of plant genomes and the scope of their utilization are growing rapidly every day. However, the information is scattered in research articles and review papers in journals and dedicated Web pages of the consortia and databases. There is no compilation of plant genomes and the opportunity of using the information in sequence-assisted breeding or further genomic studies. This is the underlying rationale for starting this book series, with each volume dedicated to a particular plant.

Plant genome science has emerged as an important subject in academia, and the present compendium of plant genomes will be highly useful to both students and teaching faculties. Most importantly, research scientists involved in genomics research will have access to systematic deliberations on the plant genomes of their interest. Elucidation of plant genomes is of interest not only for the geneticists and breeders, but also for practitioners of an array of plant science disciplines, such as taxonomy, evolution, cytology, physiology, pathology, entomology, nematology, crop production, biochemistry, and obviously bioinformatics. It must be mentioned that information regarding each plant genome is ever-growing. The contents of the volumes of this compendium are, therefore, focusing on the basic aspects of the genomes and their utility. They include information on the academic and/or economic importance of the plants, description of their genomes from a molecular genetic and cytogenetic point of view, and the genomic resources developed. Detailed deliberations focus on the background history of the national and international genome initiatives, public and private partners involved, strategies and genomic resources and tools utilized, enumeration on the sequences and their assembly, repetitive sequences, gene annotation, and genome duplication. In addition, synteny with other sequences, comparison of gene families, and, most importantly, the potential of the genome sequence information for gene pool characterization through genotyping by sequencing (GBS) and genetic improvement of crop plants have been described. As expected, there is a lot of variation of these topics in the volumes based on the information available on the crop, model, or reference plants.

I must confess that as the series editor, it has been a daunting task for me to work on such a huge and broad knowledge base that spans so many diverse plant species. However, pioneering scientists with lifetime experience and expertise on the particular crops did excellent jobs editing the respective volumes. I myself have been a small science worker on plant genomes since the mid-1980s, and that provided me the opportunity to personally know several stalwarts of plant genomics from all over the globe. Most, if not all, of the volume editors are my long-time friends and colleagues. It has been highly comfortable and enriching for me to work with them on this book series. To be honest, while working on this series I have been and will remain a student first, a science worker second, and a series editor last. And I must express my gratitude to the volume editors and the chapter authors for providing me the opportunity to work with them on this compendium.

I also wish to mention here my thanks and gratitude to the Springer staff, particularly Dr. Christina Eckey and Dr. Jutta Lindenborn for the earlier set of volumes and presently Ing. Zuzana Bernhart for all their timely help and support.

I always had to set aside additional hours to edit books beside my professional and personal commitments—hours I could and should have given to my wife, Phullara, and kids, Sourav and Devleena. I must mention that they not only allowed me the freedom to take away those hours from them but also offered their support in the editing job itself. I am really not sure whether my dedication of this compendium to them will suffice to do justice to their sacrifices for the interest of science and the science community.

New Delhi, India Chittaranjan Kole

Preface

Genomics is a combined discipline of classical genetics and computational science. Genomics aims at understanding the structure, function, evolution, genetic mapping, and genome editing of an organism using DNA sequence data. This book showcases significant breakthroughs and updates in oil palm genomic research and related fields. Meant for a wide spectrum of readers, a chapter on the history and economic importance of the crop is included. With oil palm genome sequence, genes or markers associated with agronomic traits of interest were identified while new tools were developed. Certain tools were applied to estimate genetic diversity of the oil palm. Such information is crucial to breeders in designing appropriate breeding schemes to enrich the narrow genetic base of current planting materials for sustainable development of the industry. The genetic control of economically important phenotypes, such as fruit forms, fruit colours, and tissue culture-related abnormalities, is among the principle outcomes from genomic research. Subsequent chapters highlight the introgression of wild germplasm with current genetic materials and applying modified reciprocal recurrent selection scheme in breeding programmes and production of improved planting materials. Genetic improvements were further enhanced by means of molecular cytogenetics tools and marker-assisted selection, developed from genome-wide association studies (GWAS) and genomic selection (GS). This method is to select palms carrying specific chromosomes and favourable QTLs more precisely for breeding programmes, leading to the development of elite planting materials. Such high-yielding materials with niche characteristics are tissue cultured to obtain a large number of uniformed planting materials. An array of innovations were developed including fine-tuning the propagation protocols to more user-friendly and efficient with fast result. Genomic-derived quality control tool was developed and applied across tissue culture materials to minimize yield-affecting somaclonal variations. Complementing the breeding techniques, genetic engineering is used to diversify palm oil applications vis-a-vis higher value-added products. Researchers continuously working towards optimizing genetic transformation systems of the oil palm and challenges faced during the process deliberated. The final chapter presents the state-of-the-art post-genomics tools such as transcriptomics, proteomics, and metabolomics which are embraced as phenotyping tools to elucidate the mechanisms in fruit ripening and fatty acid synthesis, among others. On the account of the indispensable need to unravel diseases in oil palm, post-genomic tools are exploited to advance knowledge in plant–pathogen interaction for novel biomarker discoveries. The chapters in this book were contributed by experts in their respective fields of research. This book provides a comprehensive reference material in genomics research for both oil palm and non-oil palm scientists.

Kajang, Malaysia

Maizura Ithnin
Ahmad Kushairi

Contents

The History and Economic Importance of the Oil Palm

1

Sean Mayes

Abstract

Oil palm is a remarkable crop which in 2017 produced 34% of the world's vegetable oil from 0.36% of the world's agricultural land. The fact that it is a perennial 'tree' crop growing in the humid tropics under high light intensity accounts for the tenfold advantage in oil yield it has over many annual temperate oil crops, per hectare, per year. However, slow breeding cycles in a manually intensive industry are beginning to erode this advantage, and the challenges of climate change have yet to be addressed. The history, economic importance and how research has been applied to improve the productivity and value of the oil palm are reviewed, along with some of the future challenges which research needs to address.

1.1 Introduction

The oil palm is a recent crop success story, particularly in Southeast Asia. While it was widely known and traded during the European industrial revolution, it has only really become a major crop in the last century. This is largely due to the establishment of a wide range of food and nonfood uses and of organised plantations which now produce the majority of the world's traded palm oil and vegetable oil as well. The fundamental basis for this success has been productivity, with oil yields now more than fourfold higher than at plantation establishment in the 1920s (Corley and Lee 1992; Henson 2012—Table 1c). With the current average yields in Malaysia, for example, running at between 3.5 and 4 tonnes oil per hectare per year, this is also around tenfold the yield obtained by some annual oil crops. Moreover, global production of vegetable oil has increased from 90.5 million metric tonnes in 2000/01 to 195.1 million metric tonnes in 2017/18, an increase which has been driven by rising demand and led to palm oil being the most traded global oil (Statistica; 27-08-2018).

While the recent past of oil palm cultivation has been a success story, the future is more uncertain. The yield gap between average yield actually obtained and potential yield is stubbornly static in many countries with the best breeding trials achieving 10–12 tonnes of oil per hectare per year (Henson 2012), and with plantations facing significant palm disease and labour shortage (and cost) threats, oil palm as an industry will need to change significantly in the future if it is to remain competitive and relevant. This is further exacerbated by the poor press particularly relating to deforestation (Vijay et al. 2016) that the oil palm industry receives in some

S. Mayes (✉)
University of Nottingham, Sutton Bonington
Campus, Loughborough, Leicestershire LE12 5RD,
UK
e-mail: sean.mayes@nottingham.ac.uk

M. Ithnin and A. Kushairi (eds.), *The Oil Palm Genome*,
Compendium of Plant Genomes, https://doi.org/10.1007/978-3-030-22549-0_1

of its biggest markets, such as Europe, and the current predictions for climate change and their potential impacts (Paterson et al. 2015).

Corley and Tinker (2015) cover just about everything to do with oil palm and the industry, while Henson (2012) also provides a detailed review of the history of oil palm. Henderson and Obsorne (2000) provide an interesting take on some of the drivers for the industry developing.

The current chapter provides a brief overview of the history and importance of this crop and looks at some of the research history, while also identifying some of the broad challenges for the future.

1.2 Taxonomy, Biology and Distribution

Botanically, 'African oil palm' is classified in order, Arecales; family, Arecaceae; subfamily, Arecoideae; tribe, Cocoseae and in the genus 'Elaeis' with the specific name reflecting the presumed origin in African Guinea, hence *Elaeis guineensis*. It is a monoecious palm with a single growing point, which produces pinnate leaves and bunches of fruit composed of drupes with an endocarp, shell and exocarp. Palm oil comes from the exocarp and kernel oil from the endocarp. For mature palms, bunches can be composed of thousands of fruits and weigh around 50 kg. Individual palms cycle between the production of male and female inflorescences, naturally promoting outcrossing, although some bunches may have small numbers of hermaphroditic flowers (Corley and Tinker 2015).

The wild/semi-wild palm belt in Africa stretches from roughly +10° N to 10° S (although populations are also found in Madagascar, some 20° S, although appear to be relatively genetically distinct (Corley and Tinker 2015—Fig. 1; Bakoume et al. 2015). While *E. guineensis* has been found growing 'wild' in South America, pollen and fossil evidence suggest an African origin (Zeven 1964), with potential transfer of African oil palm to South America in historical time, where it was only picked up industrially in recent years.

A related palm, *E. oleifera,* is also within the genus of Elaeis and is the 'South American oil palm'. The consensus is that the two species split with the breakup of an ancient landmass around 60 million years ago, leading to a speciation by geographical separation without genetic barriers. While hybrids are possible, they can have limited fertility, requiring backcrossing. Production based on F_1 hybrids requires hand-pollination to ensure bunch set. Both species have 16 chromosomes ($2n = 32$), and cytogenetic studies have shown three classes of chromosomes according to length, with similarity between the two species (Maria et al. 1995, 1998; Castilho et al. 2000) although Genomic In Situ Hybridisation (GISH) can be used to distinguish the species (Cheah et al. 2000; Zaki et al. 2017). The recent publication of the genome sequence for *E. guineensis* and comparisons to *E. oleifera* show extensive genomic differences, with substantial variation in the types and quantities of repeat elements between species (Price et al. 2002, 2003; Singh et al. 2013a). Divergence tests suggested a separation time consistent with the geographical separation theory (51 MYR) (Singh et al. 2013a) arguing for a very ancient divergence point, with subsequent genomic evolution.

Genetic diversity analysis from the collections made by Palm Oil Research Institute of Malaysia (PORIM; now Malaysian Palm Oil Board—MPOB) shows the highest level of variation for *E. guineensis* within a Nigerian collection, when compared with 11 other African countries in the palm belt (Maizura et al. 2006) suggesting this region as a possible centre of origin or diversity.

1.3 History and Economic Importance

The first recorded sample of what has been tentatively identified as palm oil was rin from funeral goods interned with a Pharaoh of the Early Dynastic period, around 5000 BCE (Raymond 1961). This suggests that palm oil was recognised to be of value and that human exploitation of palm oil was already underway at that point. Excavations in Ghana have identified preserved samples of palm oil from around 4500 BCE indicating that local exploitation in West Africa was underway at a similar period to the Abydos sample in Egypt. The Abydos sample also argues for the 'trade'or at least movement of oil palm, given that oil palm is unlikely to have ever grown in Egypt. It has been suggested that Arab traders were responsible for taking it to Egypt (Obahiagbon 2012).

The first botanical description was produced by Jacquin (1763) after whom the species is named, with more detailed description by Gaertner (1788) with a recognition of the ability of the palm to produce inflorescences of different sexes. Likely written descriptions by Portuguese explorers date from 1435 to 1460 (Crone 1937). Palm oil was used as part of the food for slaves during the slave trade on the long journey to the New World, but it was really developments in Europe which secured the future of the African oil palm trade. The Industrial Revolution led to the development of saponification and created a major demand for oil palm for lubricants, detergents (e.g. Palmolive, being a well-known soap brand)(Henderson and Obsorne 2000) and other industrial products, and this demand saw trade (still as a cottage industry) increase significantly. Eventually, the trade in oil palm was promoted as a way to suppress the continuing trade in slaves, which was banned in Britain and the British Empire in 1807 through the efforts of William Wilberforce and the abolitionists, but continued clandestinely for many years (Stilliard 1938), with the trade in palm oil seen as a legitimate replacement for shipping slaves from Africa.

The quality of the initial palm oil was poor—it was still essentially a cottage industry, and oil and kernels would be prepared by the villages along the coast or imported from further inland. It was only really with the establishments of the first forts that a more coherent approach to oil palm trade begins, and eventually, this led to the establishment of the first plantations in colonial Africa. In the 1880s, around 75% of Ghana's international trade was in palm oil (http://mofa.gov.gh/site/?page_id=8819). One of the most notable early plantations was set up on a Belgian concession in the Congo (Democratic Republic of Congo at Binga) in 1935 by Sir William Lever, who founded the British part of Unilever (Henderson and Obsorne 2000).

Since these initial stages of industrial development, considerable progress has been made, although the focus of the industry shifted after World War II towards Southeast Asia, with significant expansion in the 1960s and 1970s, where plantations established in Malaysia and Indonesia were highly productive and now produce 85% of palm oil in the world.

A concern of the current oil palm industry is a relatively restriction origin for some of the most important origins in the current commercial use. The final commercial palm is a hybrid between thick-shelled *dura* palms and shell-less *pisifera* pollen parents. This single gene trait has had more influence on palm breeding than any other (Beirnaert and Vanderweyen 1941; Singh et al. 2013b). According to pedigree, the main Deli *dura* population (seed parent) is derived from four palms introduced into the Bogor Botanic Garden in Indonesia in 1848. Offspring of these were grown in the Economic Garden and were used by many plantations for selection, although selection criteria adopted were often different, leading to a number of Breeding Populations of Restricted Origin (BPRO) (Rosenquist 1986). On the pollen side, the offspring of a palm called 'Djongo'—the 'best'—was imported from Congo and eventually came to represent the AVROS *pisifera* origin with 75% of the genetic material derived from a single palm. This is now the most important *pisifera* source (Corley and Tinker 2015). However, as oil palm is naturally outcrossing, going through cycles of male and female inflorescence production, it seems very likely that early attempts at controlled pollination involved relatively high levels of pollen contamination. The introduction of the pollinating weevil, *Elaeidobius kamerunicus* into Southeast Asia in 1981 (Syed 1982) led to high levels of *dura* contamination in controlled commercial seed production, so the original African sources of germplasm, where the pollinating weevils were endemic, are likely to represent a range of material from different pollen sources, even if the maternal palm was correctly selected.

Despite a technically very narrow pedigree, the performance of the introduced thick-shelled *dura* fruit type was sufficiently good (or the tested African thin-shelled *tenera* sufficiently bad) that Malaysia and Indonesia only moved reluctantly from planting thick-shelled *dura* as commercial material to thin-shelled *tenera* in the 1960s. This shift increases oil yields through reducing shell thickness and increasing mesocarp oil by 30%. It certainly is also the case that the environment in Southeast Asia is more favourable than much of Africa, being humid with high rainfall, stable temperatures, good light and, perhaps most importantly, little in the way of seasonal variation, so that palms can produce all year round. Today, Southeast Asia accounts for around 90 per cent of world production and 61% of world trade. In 2017, 67.92 million tonnes of palm oil were produced, together with 7.25 million tonnes of palm kernel oil, compared with 53.94 million tonnes of soybean oil, the second-largest oil producer. Oil palm is planted on 19.04 million hectares of world agricultural land (0.36%). Oil palm alone accounts for 34% of world production (Kushairi et al. 2018).

Oil palm is also unusual in that it produces two oils: mesocarp oil from the fruit pericarp and kernel oil from the kernels. These are extracted separately and have different compositions, with 'crude palm oil': CPO—mesocarp oil—being roughly 50% palmitic, 40% oleic and 10% other unsaturated oils. This makes it more solid at (nontropical) room temperatures, and it is often fractionated into oleic and stearic fractions. The kernel oil is formed of saturated short-chain fatty acids, which provide a composition quite similar to coconut oil, and it is often used in ice creams and coffee whiteners. Because of the desire to separately extract the two oils, the processing is slightly more complicated than for temperate oilseed crops. In addition to the oil itself, CPO is high in carotenoids and tocopherols and tocotrienols. The first can be used for Vitamin A production and the latter ones for Vitamin E production. The crushed seed cake can be fed to animals (but is a relatively poor feed), and palm oil can be converted to biodiesel, while there is also some use of the palm trunk for construction materials (Soh et al. 2017a).

1.4 The Challenges

Breeding in any tree crop is often a long-term process, and oil palm is no exception. The understanding that the control of the fruit shell thickness was under control of a single gene means that almost all commercially planted material today is of *tenera* hybrids. In practice, this has led to variations on Recurrent Reciprocal Selection (RRS) and Family and Individual Selection (FIS) schemes, or their combinations (Soh et al. 2017b). Because of the requirement for hybrids and the difficulty predicting breeding values in the *pisifera* line (many origins are female sterile due to bunch abortion), the actual breeding cycle for oil palm is long (between 12 and 19 years). Despite the long cycles, an experiment evaluating different rounds of selection of material suggests that in four generations, the yield had quadrupled, compared to unimproved material. Roughly half of this was attributed to genetic improvement and half to improvements in management and agronomy (Corley and Lee 1992).

The main factors influencing response to breeding selection are the heritability of the trait and the selection intensity imposed. 'Oil yield per palm' has a lower heritability than 'oil-to-mesocarp ratio', and greater response to selection would be expected by focusing on more heritable components of the desired oil yield trait. This led to the idea of bunch analysis (Blaak 1965) which breaks down oil yield into the component traits which underlie it. This approach to assessing palm value in a breeding programme has been largely unchanged since it was originally adopted. In a breeding trial, it is possible to measure fresh fruit production from individual palms, but it is not possible to bunch analyse all of the bunches from a palm. This leads to a sampling approach, where the value of a palm is only judged when a minimum of three bunches have been analysed (and preferably over five) and often only as a part of assessing the quality of the parental palms of a family. While this appears potentially limiting in terms of selection accuracy, it clearly works in practice, with yield