The genome landscape of *Elaeis guineensis*: development and utility of chromosome-specific cytogenetic markers

A thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

By
Noorhariza Mohd Zaki

April 2019
Department of Genetics and Genome Biology

The history of the earth is recorded in the layers of its crusts

The history of all organisms is inscribed in the chromosomes

-Hitoshi Kihara, 1946-

Declaration

I hereby declare that no part of this thesis has been previously submitted to this or any other university as part of the requirements for a higher degree. The work described here, unless otherwise acknowledged in the text or by reference, was conducted by the undersigned who is fully responsible. The work was conducted at the Department of Genetics and Genome Biology, University of Leicester, during the period from November 2015 to October 2018.

Signed: *Noorhariza*……. Date:**30/04/2019**…….

(Noorhariza Mohd Zaki)

Dedication

This work is dedicated to my family that have been stand by me throughout this PhD invaluable journey;

To my dear husband,
***Norshamsul Md Isa**, thank you for your patience, love, friendship, humor, shoulder to cry on and willingness for takeaway dinner*  . *Sharing our life and love along this journey together is a blessing beyond word.*

To my beautiful children,
***Syahmi Haziq, Safiyyah Hana** and **Syamil Harith**, thank you for being my best cheerleaders that brightened up my day whenever I could not get any results from the lab work.*

-Mama-

26th April 2019

Acknowledgement

First and foremost, I wish to express my sincere gratitude to my supervisors **Prof. J.S (Pat) Heslop-Harrison** from University of Leicester and **Dr. Rajinder Singh** from Malaysian Palm Oil Board for their continuous support, guidance and encouragement over the last three and a half years. It was a great privilege and honour to have guidance from both of experts throughout my PhD years.

I would also wish to thank **Dr. Trude Schwarzacher** and **Dr. John Bailey** for their technical expertise and valuable comments to improve this work.

I would like to thank the **Malaysian Palm Oil Board (MPOB)** for Ph.D. fellowship. My special thanks to MPOB researchers that involved in this work, particularly Dr. Maria Madon for her valuable comments on my work, Nordiana Hanim Mohd Nor, Muhammad Azwan Zulkifli, Zainab Anip and Mohd Razali Mohd Noor who assist me with the collection of plant material from Malaysia.

I am also indebted to Mr. Ramesh Patel, Ms. Heather and my colleagues in Lab 201 and Department of Genetics and Biology; Dr. Osamah Alisawi, Dr. Rubar M. Salih, Dr. Sarbast M. Ihsan, Dr. Stuart Desjardin, Dr. Manosh B. Kumar, and Dr. Paulina Tomaszewska.

Last, and by no means least, I wish to thank my mother, nenek, and parent-in-law that always positively encourage me throughout these years.

The genome landscape of *Elaeis guineensis*: development and utility of chromosome-specific cytogenetic markers

Noorhariza Mohd Zaki

Abstract

Identification of individual *Elaeis guineensis* (2n=32; Arecaceae) chromosomes facilitates further understanding of the genome as well as inheritance of specific agronomic traits in oil palm. While chromosome morphology shows a range of lengths and arm ratios, these do not identify most chromosomes; a tertiary-constriction was noted on the largest chromosome. Here, the first *E. guineensis* reference karyotype with a combination of physical FISH-mapping of repetitive DNA and single copy sequence (EgOligoFISH) was developed. The individual 16 pairs of *E. guineensis* chromosomes could be distinguished using probes from a combination of repetitive DNA (5S rDNA, 18S rDNA) and single copy DNA derived from massive pools of oligonucleotides. Analysis of repetitive DNA from raw Illumina sequence data revealed that, aside from the structural component of repetitive DNA (telomere and rDNAs) and abundance of transposable element superfamily (including *copia*-like Eg9CEN), no newly identified repetitive DNA could distinguish individual *E. guineensis* chromosomes. Exploration of various approaches in developing robust FISH-based chromosome-specific markers from single copy sequence resulted in three sets of massive oligonucleotide (oligo)-based probes (EgOligoFISH; OPAQUE, PPAQUE, and QPAQUE) that are able to identify 16 oil palm chromosomes *via* fluorescent *in situ* hybridisation (FISH). Simultaneous *in situ* hybridization of all three pre-labelled oligo-probes successfully identify about 2/3 of the oil palm chromosomes. The assessment of *E. guineensis* derived massive oligo-probes (EgOligoFISH) on other Arecaceae species was informative. A conserved physical localization of the EgOligoFISH on another *Elaeis* species; *E. oleifera* permits the first proposed *E. oleifera* FISH-karyotype established in this study. The integration of information obtained from *in silico*, as well as physical FISH-mapping of EgOligoFISH on *E. oleifera* mitotic chromosomes, successfully established *E. oleifera* standard karyotype for the first time. As intergeneric markers, the EgOligoFISH probes allowed comparison of chromosome organization in coconut (*Cocos nucifera)* and date palm (*Pheonix dactylifera*) chromosomes.

List of abbreviation

%	Percentage
BAC	Bacterial artificial chromosomes
BCIP	5-bromo-4-chloro-3-indolyl-phosphate
bp	Base pairs
BSA	Bovine Serum Albumin
BLAST	Basic Local Alignment Search Tool
CL	Cluster
cm	Centimetres
DAPI	4',6-diamidino-2-phenylindole
CTAB	Cetyltrimethylammonium bromide
Dig	Digoxigenin
DNA	Deoxyribonucleic acid
dNTPs	Deoxy nucleotide triphosphates
dUTP	2´-Deoxyuridine, 5´-Triphosphate
EDTA	Ethylene diamine tetra-acetic acid
FISH	Fluorescent *in situ* hybridization
FITC	Fluorescein Isothiocyanate
g	Gram
Gb	Giga base pairs
HCl	Hydrochloric acid
hrs	Hour(s)
kb	Kilo base pair
LINEs	Long Interspersed Nuclear Elements
Low Complexity	LC
LTR	Long terminal repeat
M	Molar
mg	Milligram(s)
ml	Millilitre(s)
MPOB	Malaysian Palm Oil Board
MYA	Million years ago
NCBI	National Center for Biotechnology Information
ng/ul	Nanogram/Microlitre
NGS	Next Generation Sequencing
NOR	Nucleolar Organizer Region
°C	Degrees celsius
PBS	Phosphate buffered saline
PCR	Polymerase Chain Reaction
pg	picogram
QTL	Quantitative Trait Loci
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
RNase	Ribonuclease
rpm	Rotations per minute
RT	Room temperature

List of abbreviations continue

SINEs	Short Interspersed Nuclear Elements
SNP	Single Nucleotide Polymorphism
TAREAN	TAndem REpeat Analyzer
TE	Transposable element
t/ha	Metric tonnes per hectare
U	Unit
UV	Ultraviolet
v/v	Volume per volume
w/v	Weight per volume
WGS	Whole Genome Shotgun
µg	Microgram
µg/ml	Microgram per milliliter
µl	Microlitre
µM	Micro molar

Table of content

CHAPTER I

Introduction

1.1 The oil palm

1.1.1 Origin and cultivation

The oil palm is an angiosperm monocotyledon plant belonging to the *Elaeis* genus of the palm family (Arecaceae). The genus *Elaeis* consists of only two species: the African oil palm (*Elaeis guineensis* Jacq.) and the Latin American oil palm (*Elaeis oleifera* H.B.K Cortés) (Corley and Tinker, 2003). The generic name *Elaeis* originates from the ancient Greek word *elaion*, which means 'oil' and the species name '*guineensis*' refers to 'Guinea', the geographic origin of the first discovered oil palm. 'Jacq.' is named after Nicholas Joseph Jacquin, who officially named the oil palm for the first time in 1763.

Previously, the classification of the second species, *E. oleifera*, has been the subject of some argument among taxonomists. In the literature, the *E. oleifera* has been referred as *Elaeis melanococca* and *Corozo oleifera*, when the species first documented in the 1700s (Hardon and Tan, 1969). The latter name was mistakenly used to identify corozo, as was pointed out by Bailey (1933) but has been generally used by non-taxonomists (Blank, 1952). Later, in 1965, Wessel Boer confirmed the classification in the genus *Elaeis* and suggested *E. oleifera* (Kunth) Cortés as a South American species. Both *Elaeis* species can be hybridised, suggesting a close relationship despite their origins in two different continents (Hardon and Tan, 1969).

No historical records formally indicate the geographical origins of the two oil palm species. Archaeological evidence has traced the presence of oil palm as early as 5,000 B.C. in ancient West Africa and Egypt (Hartley, 1967; Corley and Tinker, 2003). The first official documentation showing the existence and trading of palm oil was made by the Portuguese exploration team of Prince Henry the Navigator during their expedition to the Guinea coast

of West Africa in the 14th century (Hartley 1967, Corley and Tinker 2016). In 2013, the first reported divergence prediction of both *E. guineensis* and *E. oleifera* was reported based on the whole genome sequence data (Singh *et al*., 2013). The divergence of both *Elaeis* species 51 million years ago (MYA) was found to coincide with the separation of continents during the Cretaceous period where the formation of deep-water connecting the Central and South Atlantic Ocean separated South America and Africa (Figure 1.1). Consequently, the split of both continents resulted in the separation of the plants of either side of the continents.

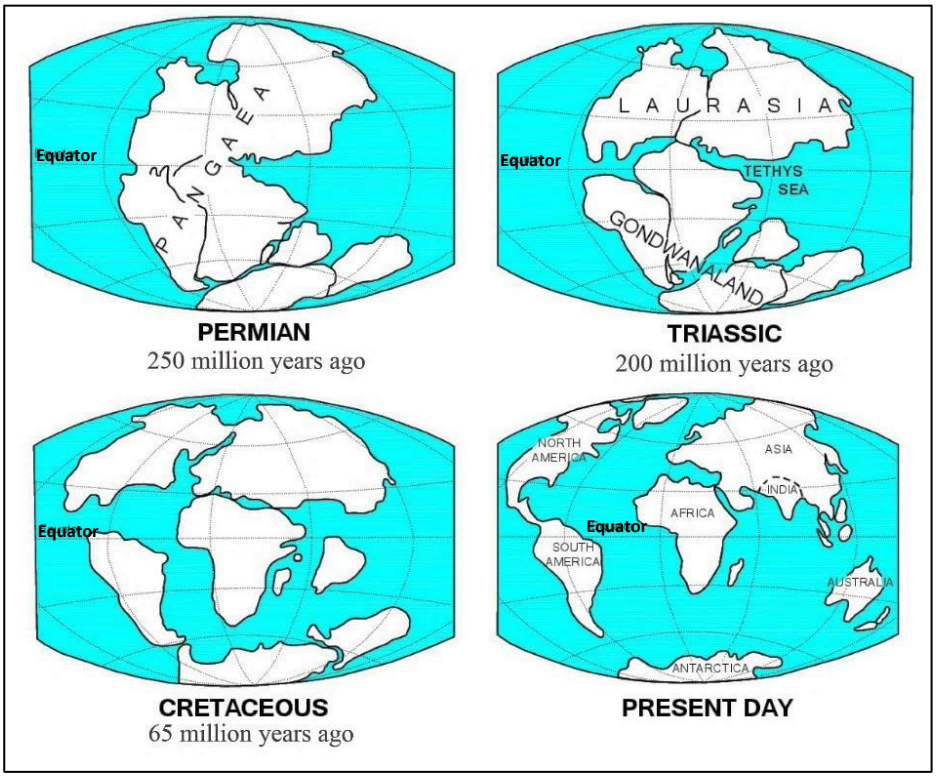


Figure 1.1 Timeline for separation of an ancient supercontinent (Gondwana) into the present world landmasses. Pangaea split apart at the end of the Triassic (200 MYA) into two supercontinents: Laurasia to the north and Gondwana drifting southward. Gondwana itself broke apart about 180 MYA into progressively more separated landmasses that we know today as South America, Africa, Antarctica, Australia, the Arabian Peninsula, and India. (Source: Dittus, 2017)

From Africa, the *E. guineensis* (dura) was exported and grew in The Hortus Garden (Hortus Botanicus) in Amsterdam before 1848. In 1848, two seedlings of the palm were shipped and planted in the Buitenzorg Botanical Gardens (later known as Bogor Botanical Gardens), Indonesia. During the same period, the Buitenzorg Botanical Gardens also received another two oil palm seedlings from Mauritius. These four seedlings were then planted in Deli, Sumatra and named as Deli dura. Later, the Deli dura seedlings were distributed to other parts of Indonesia (Banjar Mas, Java, and Palembang, Sumatra) and Kew Garden in Singapore. Thirty years later (in 1876), the Deli dura was brought from Singapore to Labuan, Malaysia. Subsequently, from 1911 – 1917, commercial plantations of Deli dura were established, both in Indonesia (Sumatra) and Malaysia (Rantau Panjang, Kuala Selangor), and this marks the beginning of the oil palm industry in Southeast Asia (Hartley, 1967; Kushairi and Rajanaidu, 2000). To date, the oil palm breeders and industry members have established that Deli dura, the pioneer planting materials in Southeast Asia, originated from West Africa. In 1917, all the progenies from the four mother palms in Buitenzorg Botanical Gardens were found morphologically similar. Later, molecular genetic diversity studies confirmed that the Deli dura was grouped closely to palms from West Africa and far from Madagascar, which is located next to Mauritius island (Hayati *et al.,* 2004; Ting *et al.,* 2010; Zaki *et al.,* 2012; Bakoumé *et al.,* 2015).

The African oil palm (*E. guineensis*) populations inhabit tropical lowlands with the average annual rainfall of about 1,780–2,280 mm and temperature ranging from 24 °C to 30 °C. The main belt of the palm groves covers regions between 10°N and 10°S from Senegal to Guinea, Sierra Leone, Liberia, Côte d'Ivoire, Ghana, Togo, Benin, Nigeria, Cameroon, Gabon, Congo, Angola and DR Congo (Figure 1.2) (Hartley, 1967; Corley and Tinker, 2015). As for *E. oleifera*, the groves were found along the riverbanks, under shady canopies of tall forest trees and on the areas prone to flooding. The species were distributed widely from Colombia, Suriname, North-West Brazil and the Amazon River basin (Corley and Tinker, 2015; Barcelos *et al*., 2015).

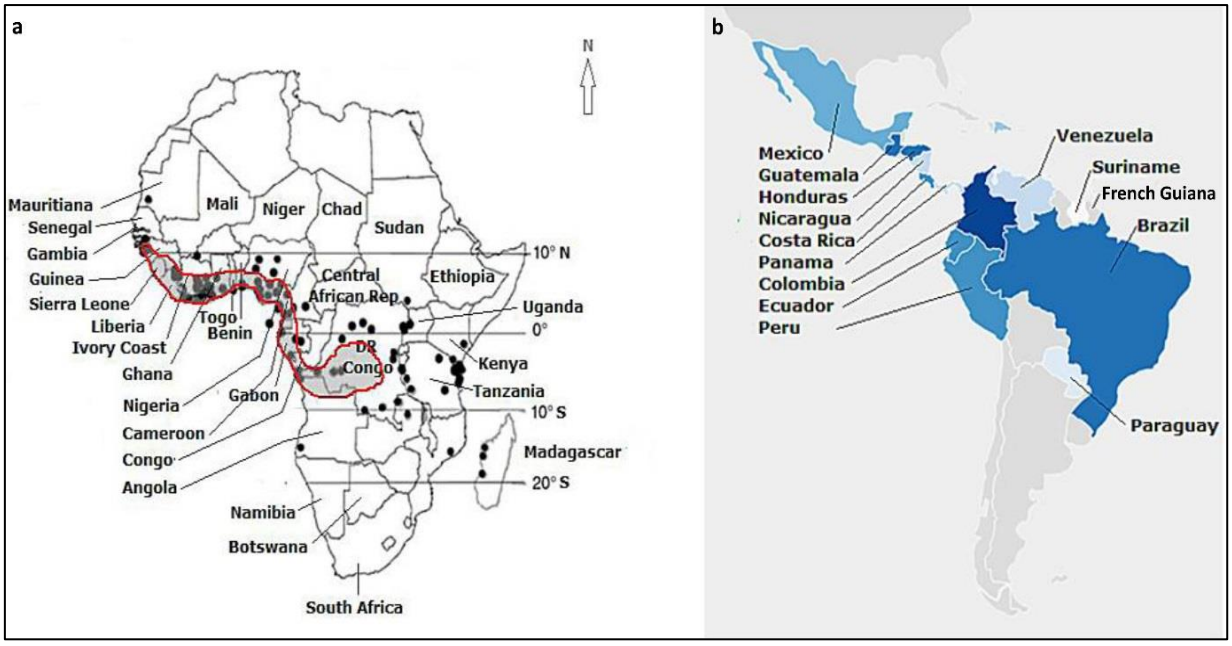


Figure 1.2 Geographical distribution of *E. guineensis* palm groves in Africa (a) and *E. oleifera* in Central and South America (b). The red line area shows the main belt of *E. guineensis* palm groves and blue regions are the current palm oil producing countries in Central and South America. (Source: Corley and Tinker, 2015)

1.1.2 Morphology and growth of oil palm

Both oil palm species are morphologically different; *E. guineensis* can reach 15-18 metres in height, up to 30 metres in a dense forest, with annual height increment of 30-60 cm per year. The *E. oleifera* palm is much shorter with height up to only 8 metres. The annual height increment of the *E. oleifera* is only between 5-10 cm per year, and the trunk tends to lean toward the ground (procumbent) after several years (Figure 1.3a-e).

Oil palm is a temporal dioecious species. It has a single shoot apical meristem with functionally unisexual male and female inflorescences in an alternating cycle on the same plant (Purseglove, 1972; Cruden, 1988). However, occasionally, both the gynoecium (female) and androecium may develop to give a hermaphrodite flower. The male inflorescence has an approximately 40 cm long stalk, with 100-300 finger-like spikelets containing 600-1500 yellow flowers (Figure 1.3f). The basic structure of female inflorescence is similar to the male, but with shorter spikelets, and the mature female

flower produces floral triads (Figure 1.3g-h). The female flowers develop into a bunch of mature fruits after four to six months from pollination.

A matured *E. guineensis* fruit bunch can be classified into two types depending on the colour of the exocarp; *nigrescence* (deep violet to black) and *virescence* (orange) (Figure 1.4), matured fruit bunch of *E. oleifera* having an orange fruit colour resembling the *virescence* type of *E. guineensis*. Oil palm fruits are sessile drupes produced in bunches of up to 3,000 fruits on mature palms, with an average of around 1,500 fruit/bunch. The fruits vary in shape and size and may weigh from 3 g to 30 g (Corley and Tinker, 2003). The pericarp of the oil palm is subdivided into the outer layer exocarp, fleshy mesocarp, and endocarp (shell). Shell coats the seed or kernel (embryo and endosperm) (Figure 1.4b). The crude palm oil and kernel palm oil are extracted from the mesocarp and kernel respectively. The mesocarps yield an edible, orange-red oil commonly known as palm oil and the endosperm or kernel produces a clear yellowish oil that is known as palm kernel oil.

The fruit characteristics of both *E. guineensis* and *E. oleifera* are different (Figure 1.4). *E. guineensis* are classified into three types according to the fruit phenotype; dura, pisifera, and tenera. Dura is characterised by the production of large fruits with thick endocarp (shell) and a small proportion of oil-bearing mesocarp. Pisifera is a shell-less with the oil-bearing mesocarp constituting the entire fruit. Pisifera is a female sterile and used as the male progenitor. The intraspecific hybrids of dura and pisifera, known as tenera palms, have thinner shells surrounded by a distinct fibre ring (Beirnaert and Vanderweyen, 1941). Tenera type, having 30% more mesocarp and, respectively, 30% greater oil content in bunches than dura (Corley and Tinker, 2003), is the commercial material that is widely used for oil extraction. The *E. oleifera* fruits are generally small with thick-shelled fruit. A high proportion of *E. oleifera* parthenocarpic fruits, which may constitute up to 90% of the total, as compared to the African species. The parthenocarpic fruits often abort and contribute to poor palm oil yield (Hartley, 1967; Corley and Tinker, 2003; Barcelos *et al*., 2015) without manual pollination.

The oil palm generation time is lengthy, with seeds taking around 100–120 days to germinate, followed by 10 - 12 months in the nursery before the young seedlings are ready for field planting. The oil palm starts to bear fruit after 2–3 years of field planting and approaches maturity at around ten years. The economic life of plantings varies from 20 - 30 years, depending on local conditions, with excessive palm height being the primary factor for replanting (Corley and Tinker 2003).

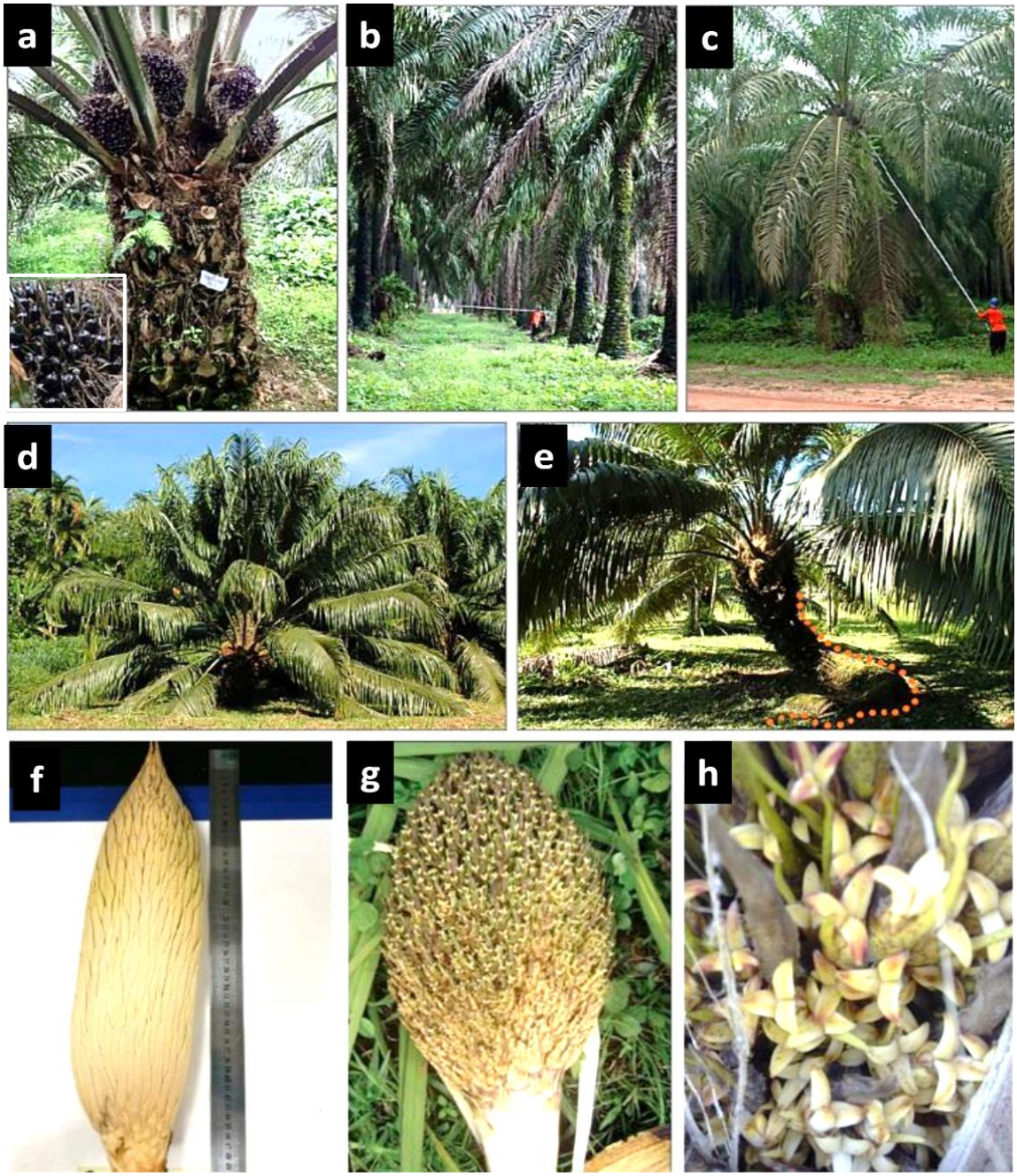


Figure 1.3 Morphology of oil palm tree and flower. (a) Commercial 5-year old tenera oil palm with nigrescens fruit type (inset). (b - c) 26-year old tenera palms with a height of 7-8 metres tall. (d) *E. oleifera* tree (30 years old; Manicore, Amazonas, Brazil). (e) The same tree as in (d) photographed at a different angle to show the procumbent trunk (orange dotted line) (f) immature male inflorescence (g - h) mature female inflorescence bearing floral triad. (Source: Malaysian Palm Oil Board in-house collection; Adam *et al*., 2011; Barcelos *et al*., 2015).

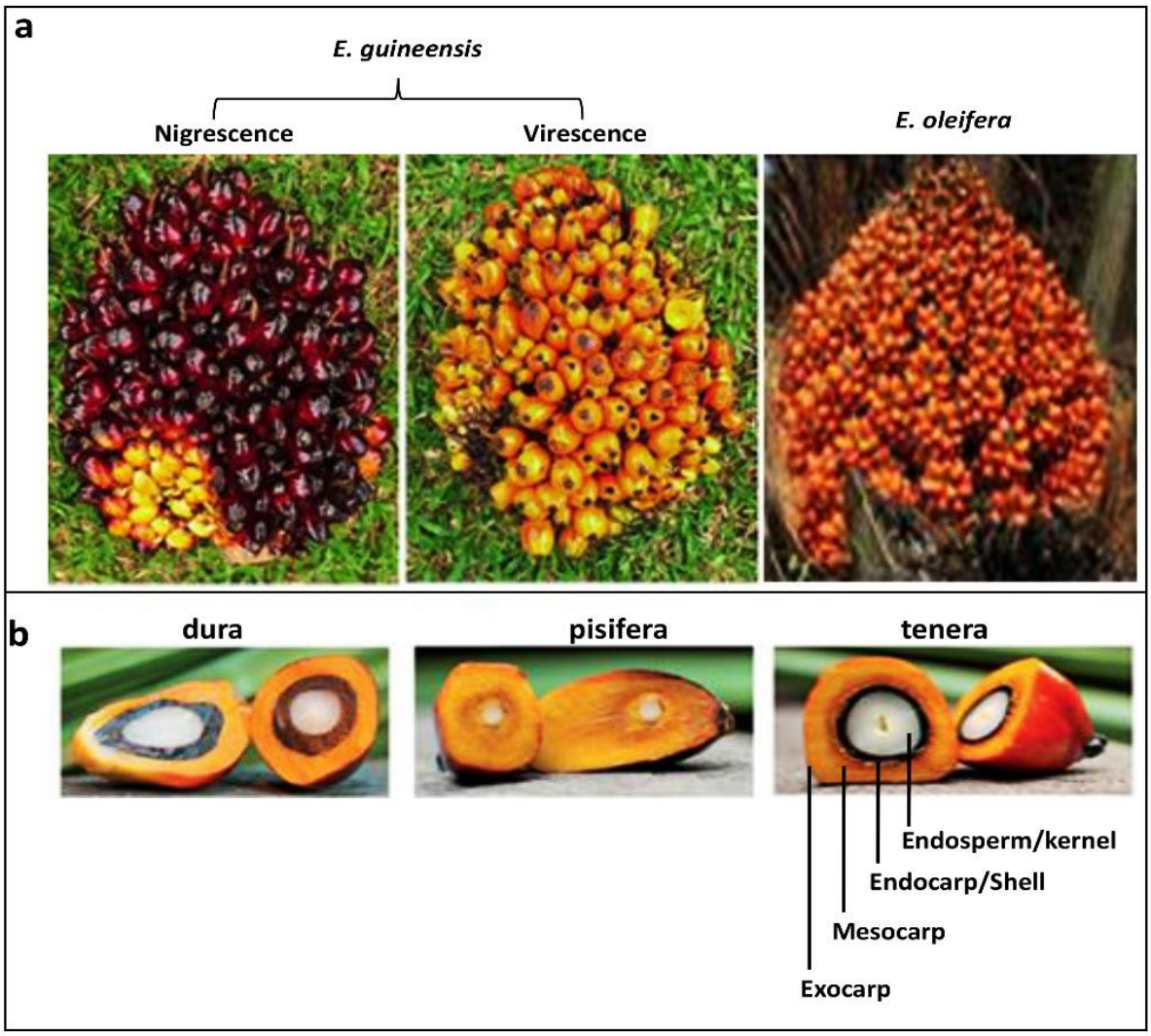


Figure 1.4 Fruit characteristic of oil palm. (a) Two types of matured *E. guineensis* fruit differentiates by the colour of the ripened exocarp (skin). The *nigrescence* phenotype with deep violet to a black colour and the *virescence* orange colour. The colour of matured *E. oleifera* resembles *E. guineensis virescence* type. (b) Three fruit phenotype of *E. guineensis*; dura, pisifera and tenera (commercial intraspecific *E. guineensis*). (Source: Singh *et al*., 2013; 2014; Malaysian Palm Oil Board in-house collection)

1.1.3 Oil palm industry worldwide and its challenges

Oil palm is a unique and one of the most productive oil-bearing tropical crop with a potential palm oil yield capacity over 10 tonnes of oil per hectare (t/ha) (Corley and Tinker, 2015). Among the ten major oilseeds, oil palm accounted for 5.5% of global land use for cultivation and produced 32.0% of global oils and fats output in 2018 (Oil World, 2018). The significantly higher productivity of oil palm has made the crop the most efficient land user and attractive for both domestic and external long-term investments into the countries that grow it.

Since the 1960s, palm oil production has been increasing steadily and has become the largest source of supply to the global vegetable oils and fats market. As of 2018, Indonesia and Malaysia are the major exporters with the production of 26.74 and 16.36 million tonnes, respectively. The Latin America palm oil-producing countries, including Colombia, Guatemala, Honduras, Ecuador, Brazil, Costa Rica, and Peru, altogether supplied approximately 4.6 million tonnes. Production in the African countries contributed about 2.7 million tonnes with Nigeria, Ghana, Côte d'Ivoire, Cameroon and Democratic Republic of the Congo (DR Congo) among the top five producers (Oil World 2018).

Palm oil's unique composition makes it versatile for applications in food manufacturing and the oleochemicals, cosmetic, biodiesel and pharmaceutical industries besides of being used as cooking oil (Choo and Kalanithi, 2014). The oil palm fruit produces mesocarp oil, which is commonly called 'palm oil' (crude palm oil; CPO), and kernel oil (palm kernel oil; PKO) from the kernel. The differences in properties and characteristics in CPO and PKO have made the two oils suitable for broad and discrete applications. The nutritional and oxidative properties, and thermal stability of CPO make it suitable for manufacturing a wide range of products, e.g., cooking oil, snacks, pharmaceuticals, and animal feedstocks. PKO, with a high content of medium-chain saturated fatty acid, is a natural plant-based source for confectionery fats (e.g., butter and margarine). The physical and chemical properties and oxidative stability of PKO also make it a valuable feedstock for the oleochemical industry. CPO and PKO have also been identified as renewable resources for biodiesel (Gunstone and Harwood, 2007).

At present, the world population is approximately 7.7 billion, and the population size is projected to grow to 9.7 billion by 2050 (Worldometer.info; http://www.worldometers.info/world-population/). The fastest-growth rates are expected mainly in India, Africa (e.g., Nigeria, Ethiopia, DR Congo, and the Republic of Tanzania), Pakistan, Indonesia and the United States of America (United Nations, 2017). At present, these are the major countries dominating the consumption of total palm oil produced, hence, suggesting that more palm oil in the form of food and daily consumables will be needed to support the rapid human growth, specifically for these countries.

Estimates from the Food and Agriculture Organisation of the United Nations (FAO) predicted that requirements of palm oil would reach between 93 and 156 million tonnes by 2050 (Corley, 2009). Therefore, it can be assuredly forecast that global demand will remain high and that there will be pressure for yield improvements for decades to come. Hence, there is a constant need to develop new oil palm varieties with enhanced agronomic traits to break the deadlock in palm oil yield, which has been stagnating at 3.5-3.9 t/ha/year for more than 20 years (Murphy, 2014). However, in some breeder's trial, 10-12 tonnes of oil/ha/year are seen for modern planting material of oil palm under best management practice (Soh *et al*., 2017). Corley (1985) predicted a potential oil yield based on physiological modelling of 17 t/ha/year, suggesting that significant breeding progress can still be made to reduce the yield gap without increasing the land use. Hence, production of palms with desired agronomic characters and with high value-added traits *via* molecular breeding is perceived as one of the strategies for sustaining the oil palm industry through marker-assisted breeding.

DNA-based assisted breeding can save time and money in crop breeding programmes. In order to select most characters of interest, it usually is necessary to grow up and analyse each new generation of the crop before it is possible to perform a phenotypic selection of appropriate plants. Using genetic markers, breeders can screen more plants at a very early stage and save several years of laborious work in the development of a new crop variety. This is especially useful for crops like oil palm where it can take three to four years or more for a fruit phenotype to become fully apparent.

The discovery of the genetic marker for *SHELL* gene (Singh *et al*., 2013), which was first identified in 1941 (Beirnaert and Vanderweyen), was one of the success discoveries of the DNA-based MAS application in oil palm breeding which proven to help breeders to increase the palm oil yields. The genetic marker for *SHELL* gene is beneficial for seed producers to reduce or eliminate dura contamination (thick-shelled fruit), and to distinguish the dura, tenera and pisifera plants in the nursery long before they are field planted. This is useful as the pisifera palms have vigorous vegetative growth, and planting them in high density encourages male inflorescence development and pollen production. Accurate genotyping such as demonstrates by *SHELL* gene's marker has a critical implication for a bioeconomy. Enhanced oil yields and other agronomic important traits can optimise and ultimately indicate a clear path towards more intensive use of already planted lands, and, thus, should lessen pressures to expand the land area devoted to oil palm, notably onto endangered rainforest land. Hence, further understanding of the chromosome levels is essential to enhance the efforts in increasing the potential of the oil palm in the future.

1.2 Cytogenetics and chromosome identification in plant

The plant nuclear genome, consisting of the DNA and associated protein, is organised into discrete chromosomes. Each unreplicated chromosome and metaphase chromatid consists of a single DNA molecule that is linear and unbroken from one end to another (Heslop-Harrison and Schwarzacher, 2011; Figure 1.5). The term chromosome means the body (soma) that takes up the colour (chromo) and was introduced by Waldeyer in the late 18th century (Schwarzacher, 2003). Study of the numbers, structure, and organisation of the chromosomes packaging the DNA within the cell nucleus is referred to cytogenetic analysis.

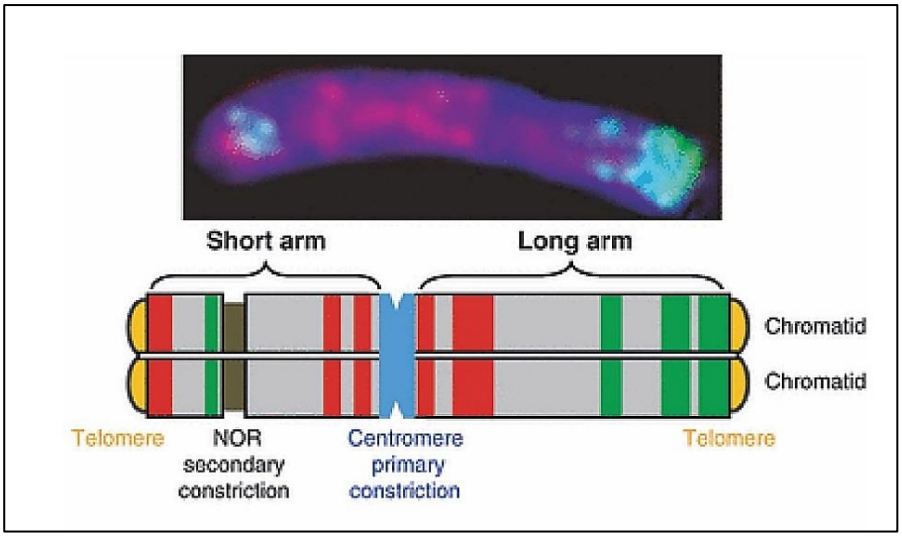


Figure 1.5 Organisation and features of a plant chromosome. Top: A fluorescent light micrograph of a metaphase chromosome stained blue with the DNA-binding fluorochrome 4', 6-diamidino-2-phenylindole (DAPI). *In situ* hybridisation shows the location of two tandemly repeated DNA sequences detected by red and green fluorescence. Bottom: A diagram of the structure of a metaphase chromosome with two chromatids. (Source: Heslop-Harrison and Schwarzacher, 2011).

Chromosomes are most commonly studied by light microscopy using the highest magnification available with oil-immersion optics (objective lens magnification of x64 or x100), in metaphase preparations. Generally, living root tips with many dividing cells from seedlings, plants growing in pots or soil, are pre-treated for up to 24 hours to arrest the cell cycle and accumulate cells at metaphase when the chromosomes are condensed onto the metaphase plate. The roots are then fixed, softened with enzyme, acid or alkali treatments, and squashed to spread metaphase chromosomes on a glass microscope slide before staining and examined under the microscope. Meiotic chromosomes are also studied using preparations made at different stages of meiosis from apical meristems or floral tissue as a source of metaphases in dividing tissues. Nuclei and chromosomes may be stained before spreading of the tissue (typically with Feulgen, which stains DNA bright red), during spread preparation (typically with aceto orcein, also a red stain), or with fluorescent stains for DNA such as DAPI. Chromosomes can also be seen without staining by observing them under phase contrast microscopy (Schwarzacher, 2016).

Morphological study of metaphase chromosomes commonly performed by measuring the absolute or relative sizes (generally taken as the length) of each chromosome. The relative sizes of the two arms divided by the centromere, allows many chromosomes in a species to be identified individually. Relative arm sizes are measured as the "arm ratio" (size of the larger chromosome arm/size of the smaller arm) (Levan *et al*., 1964) or "centromeric index" (size of the shorter arm/size of the whole chromosome), which may be expressed as a percentage (Huziwara *et al*., 1962). Chromosomes may vary in their arm ratio from being telocentric (where the centromere is at the end of the chromosome), through acrocentric and sub-acrocentric, to sub-metacentric and metacentric, where the centromere divides the chromosome into two equal arms. Practically, measurement inaccuracies and unequal condensation of arms during prophase of mitosis can make classification of the long and short arm difficult where the arms are similar sizes. Many species have groups of chromosomes which are too similar in size and arm ratio or show a continuous size distribution from larger to smaller, which makes individual identification of all chromosomes impossible (Braz *et al*., 2018).

The identification of individual chromosomes in a species provides a reference for defining structural differences in both inter- and intra-species and as a platform for developing high-resolution cytogenetic maps. Cytogenetic maps show the physical length of individual chromosomes in micrometres as measured through the microscope and the position of genetically mapped markers relative to cytological landmarks such as centromeres, telomeres, heterochromatin and nucleolar organizer regions (NOR). Genetic linkage maps show the linear order of sequences and markers along the chromosomes, and the amount of recombination between linked markers. With the advancement of DNA sequencing and genomics research, cytogenetic maps are not only valuable for integrating and organising genetic, molecular and cytological information, but also provide a unique insight into genome organisation in the context of chromosomes (Heslop-Harrison and Schwarzacher, 2011). Different kinds of genomic maps differ significantly in the method of production and the ways they are viewed; the integration of the maps is essential to gain a comprehensive view of genome structure and behaviour.

In many cases, the study of the physical chromosomes is the most efficient approach to discover translocations between chromosomes, introgression of chromosomes or chromosome segments from other species, which has been widely discussed in cereals such as wheat, rice and barley (Carvalho *et al*., 2009; Mayer *et al.,* 2011; Kruppa *et al*., 2013; Wang *et al*., 2013; Molnár-Láng *et al.,* 2014; Patokar *et al*., 2016)

One of the potential strategies to catalogue the identity of each chromosome is to directly localise the DNA sequences on physical chromosomes by fluorescence *in situ* hybridisation (FISH). FISH gives reliable and routine results regardless of the chromosome size and the quality of the chromosome preparation (Schwarzacher and Heslop-Harrison, 2000; Jiang and Gill, 2006). This technique, which has been widely used for cytogenetic and genome research, allows us to visualise the physical positions of the associated molecular markers along a given chromosome.

1.3 Fluorescent *in situ* hybridisation (FISH) as a tool for chromosome identification

1.3.1 Principle and application of FISH

FISH is a powerful and unique approach that is able to show the presence and locations of labelled DNA sequences along chromosomes (Schwarzacher, 2003). Before the arrival of fluorescent *in situ* hybridisation in biological fields, the chromosomal distribution of target DNA sequences was one of great challenge in cytogenetic studies. Initially, DNA *in situ* hybridisation with a radioactively labelled probe was first developed to visualise RNA and DNA in mammalian cells (Gall and Pardue, 1969; John *et al.,* 1969). These were relatively expensive and time-consuming and suffered from several drawbacks, including unstable probes; limited resolution; and hazardous materials. The use of radioactive nucleotides for labelling was a cumbersome and slow technique, as slides had to be dipped into a liquid photographic emulsion or covered with a film and the signal was only able to be observed after several days or weeks. Later, with the development of non-radioactive labelling and rapid detection methods, fluorescent *in situ* hybridisation (FISH) was introduced initially

using fluorescein FITC (Langer-Safer *et al*., 1982; Pinkel *et al*., 1986; Schwarzacher *et al*., 1989).

The basic principles of FISH experiments are including chromosome slide preparations, DNA probe preparation (Section 1.3.2 for details), denaturation and hybridisation of the probe and target sequences (chromosomes), washing, detection and microscopy for the signal interpretation. Figure 1.5 gives a diagrammatic overview of the FISH process.

FISH is mainly based upon the same principle as a Southern blot analysis, a cytogenetic equivalent that exploits the ability of single-stranded DNA to anneal to complementary DNA. In the case of FISH, the target is nuclear DNA of either interphase nuclei, metaphase chromosomes or chromatin fibres affixed to a microscope slide (reviewed method in Schwarzacher and Heslop-Harrison, 2000). Once fixed to a microscope slide, the desired nuclear DNA is hybridised to a nucleic acid probe where the DNA probe anneals to its complementary sequence in the specimen DNA. The probe can be labelled with a reporter molecule which is either an attached fluorochrome, enabling direct detection of the probe *via* a coloured signal at the hybridisation site visualised by fluorescence microscopy, or a hapten that can be detected indirectly. This second method relies on immunohistochemistry for probe detection which is based on the binding of antibodies to specific antigens. These molecules are linked to nucleotides and incorporated in the probe by different techniques, including random primer labelling, nick translation, and PCR-based amplification. Once antigen-antibody binding occurs, a coloured histochemical reaction can be observed by fluorescence microscope using a suitable filter with appropriate excitation. For direct detection, FITC, Rhodamine, Texas Red, Cy2, Cy3, Cy5, and ATTO are the most frequently used reporter molecules. Biotin, Digoxigenin, and Dinitrophenol are the reporter molecules typically used for indirect detection methods.

The FISH-based chromosome identification method is more versatile than the traditional chromosome banding techniques. It is widely used for mapping of DNA sequences to their physical location within the genome, for correlating the linkage groups to specific chromosomes, and for understanding the genome organisation of a species. Furthermore,

in situ hybridisation enables identification and characterisation of chromosomes and chromosome segments, providing markers for recent or evolutionary chromosome rearrangements and for changes in sequence abundance during evolution and disease (Sadder *et al*., 2000; Cuadrado *et al*., 2008; Li *et al*., 2015; Alix *et al*., 2017). Moreover, many of the answers obtained with chromosomal *in situ* hybridisation approach are challenging to discover by using any other method. With pure molecular genetics methods, genomic organisation, dynamics and evolution are very hard to interpret even when an abundance of copies of the sequence is present, whereas with *in situ* hybridisation the sequences that represent half a genome and are present in thousands of copies is able to be studied. For example, repeated sequences show multiple bands in gel electrophoresis that are difficult to separate, interpret, and assign to loci. In addition, a large-clone contig and sequencing projects are not able to access long and relatively homogeneous stretches of repetitive sequences, whereas linkage mapping gives limited data about where recombination is occurring in the genome (Jiang and Gill, 2006; Heslop-Harrison and Schwarzacher, 2011; Jiang, 2019).

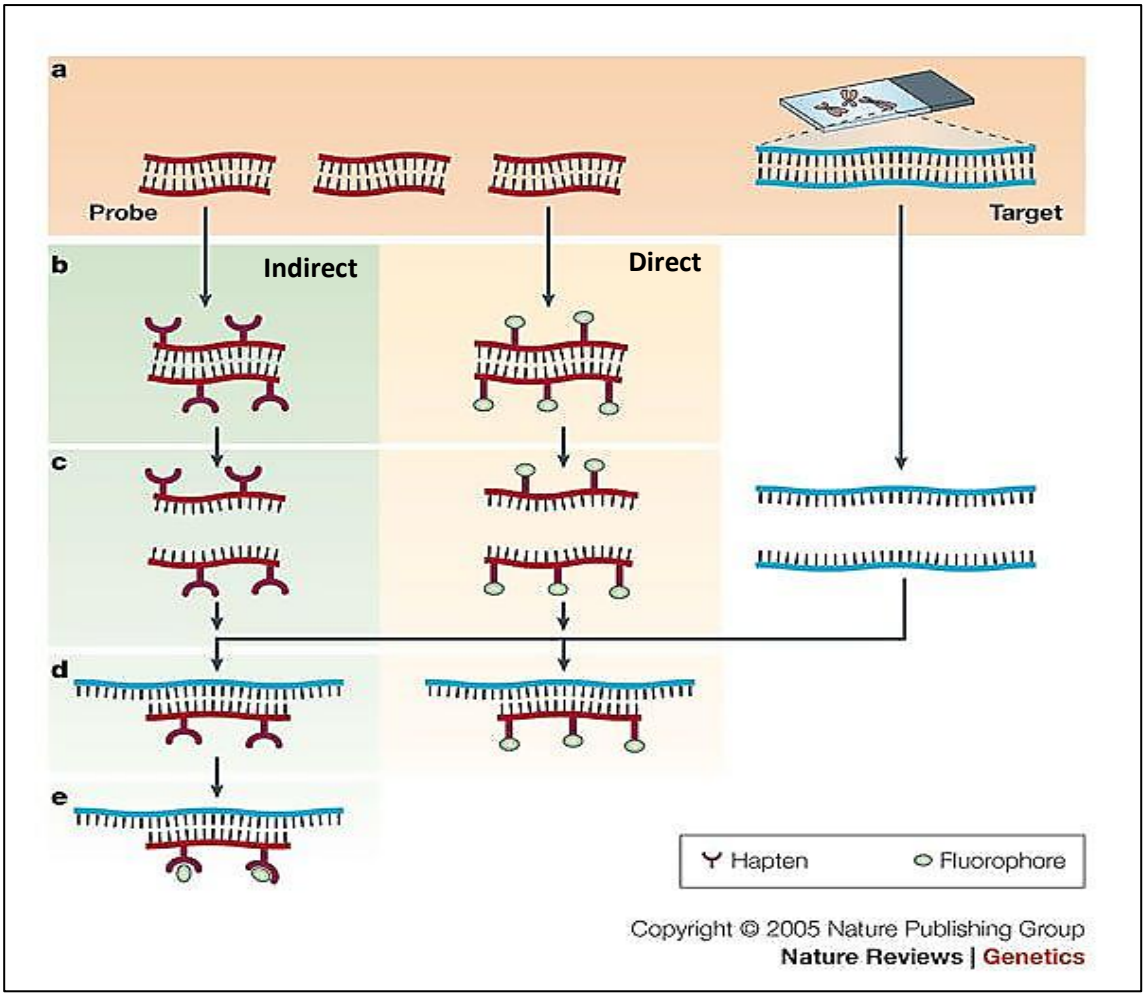


Figure 1.5 Schematic of fluorescent *in situ* hybridisation (FISH). a) The basic elements of FISH are a DNA probe and a target sequence on a chromosome. b) Before hybridisation, the DNA probe is labelled in one of two ways, either indirectly (left column) or directly (right column). In indirect labelling modified nucleotides containing a hapten (e.g., digoxigenin or biotin), while with direct labelling, modified nucleotides containing a fluorophore are used. c) The labelled DNA probe and the target chromosome DNA are denatured to yield single-stranded DNA. d) The probe and the target are mixed in conditions favourable for hybridisation; permitting the re-annealing of complementary DNA sequences. e) Detection of the hybridisation sites. For indirect labelled probes, an antibody or avidin conjugated to a fluorophore is first bound to the labelled probe and then detected with fluorescence (left column). As for direct labelled probe, no immunohistochemistry is needed, and the labelled probe can be visualised directly. (Source: Speicher *et al*., 2005)

1.3.2 Probes for *in situ* hybridisation

FISH is a very straightforward technique that involves hybridisation of DNA molecules (probes) to their complementary sequences on chromosomal preparation. It allows the detection and precise localisation of DNA sequences on interphase nuclei, chromosomes, or chromatin fibres. Apart from a good quality of metaphase preparation that is free from the rigid cell wall, one prerequisite consideration in FISH is the choice of probes. FISH permits rapid cytogenetic characterisation and chromosome identification using a variety of probes. Repetitive DNAs, large-insert genomic DNA clones, such as bacterial artificial chromosomes (BAC) clones and synthetic oligonucleotides, are the most widely used FISH probes in the plant (Jiang and Gill, 2006; Figueroa and Bass, 2010; Heslop-Harrison and Schwarzacher, 2011; Jiang, 2019).

1.3.2.1 Repetitive DNA family

Repeated DNA sequences, composed of units of a few to thousands of base pairs in size, occur in blocks or are dispersed throughout the genome (Schwarzacher, 2003). In the plant genome, the proportions of repetitive DNA could reach up to 90-95% (Heslop-Harrison, 2000). Repetitive elements in eukaryotes can be divided into three classes, in accordance with their organisation, localisation, and functions (Heslop-Harrison and Schmidt, 2012). The first class, which are also the major repeats in plants, is transposable elements (TEs), elements that amplify and reinsert into the nuclear genome (Kumar and Bennetzen, 1999; Wicker *et al.,* 2007) composed of two types that can be distinguished according to their respective mode of transposition. The DNA transposons, or class II TEs move and amplify through DNA, while class I TEs or retrotransposons amplify through an RNA intermediate. LTR-retrotransposon can make up to 50% of plant genomes and can be divided into superfamilies, including *copia* (Pseudovirideae) and *gypsy* (Metavirideae) (Hansen and Heslop-Harrison, 2004) while non-LTR retrotransposons (LINE; Long Interspersed Nuclear Elements and SINE; Short Interspersed Nuclear Elements) were found in lower percentage in plant. The second class of repeats is structural components of chromosomes, including centromeric and telomeric repeats. While telomeric sequences are highly conserved with the repeat motifs of 'TTTAGGG' in most plants, centromeric tandem repeat sequences are not highly conserved between species. The third class of repeats includes other tandem

repeats, such as satellites and microsatellites as well as highly repetitive genes, such as 45SrDNA which is typically 9kb in length comprising of 18S, 5.8S, 26S rRNA genes as well as transcribed and untranscribed spacer regions (Heslop-Harrison and Schmidt, 2012; Biscotti *et al*., 2015) (Figure 1.6)

Repetitive DNA probes can be used singly or in combinations in the FISH experiment for identification of a chromosome of a species. Many repetitive DNA elements generate specific FISH signal patterns on individual chromosomes within a single species (Kato *et al*., 2004; Koo *et al.,* 2005; Paesold *et al.,* 2012; Badaeva *et al*., 2015). FISH signals derived from combinations of repetitive DNA probes (probe cocktails) have been developed in several plant species such as maize, common bean and Asteraceae (Kato *et al.,* 2004; Fonseca *et al*., 2010; Chester *et al.,* 2013; 2015).

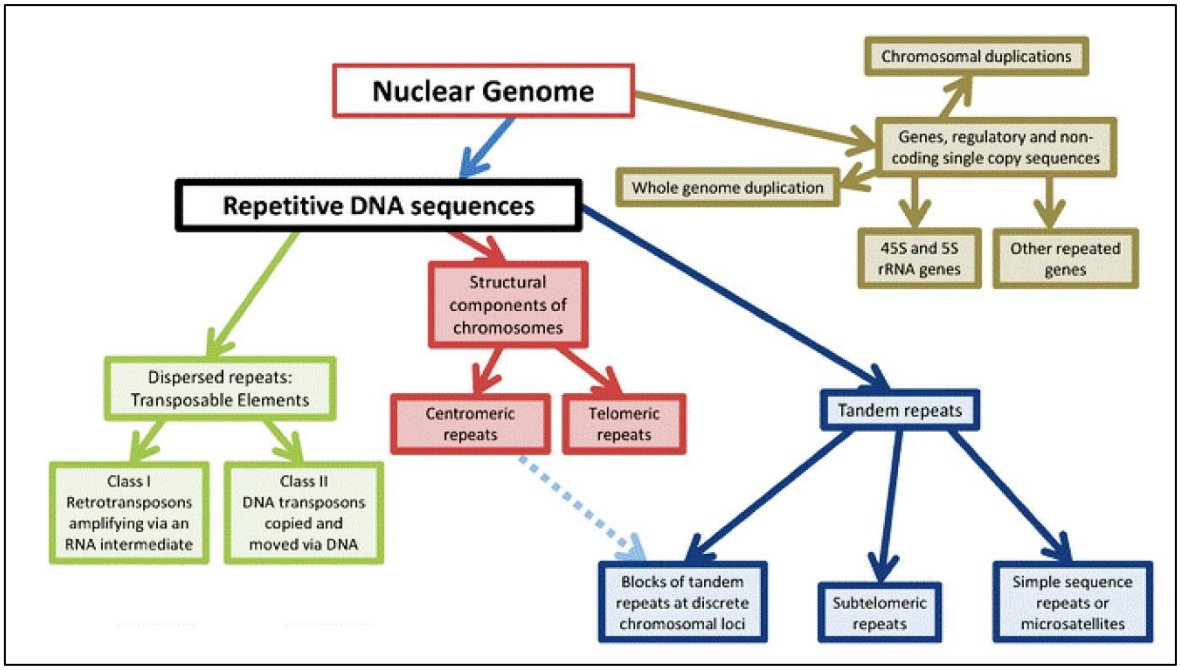


Figure 1.6 Major divisions of repetitive DNA sequences in the plant nuclear genome include dispersed repeat, structural components, tandem repeats, and repeated genes. (Source: Biscotti *et al*., 2015; after Heslop-Harrison and Schmidt, 2012)

1.3.2.2 Large-insert genomic DNA clones (Bacterial Artificial Chromosome; BAC)

As an alternative for repeat-based FISH probes, chromosome-specific cytogenetic DNA markers can be developed for individual chromosomes using large-insert genomic DNA clones, such as BACs. The BAC-FISH based approach has been successfully used to identify individual chromosomes and integrate genetic linkage groups with chromosomes of various numbers of plant species (Lysak *et al*., 2001; Findley *et al*., 2010, 2011; Paesold *et al*., 2012; Mandakova and Lysak, 2016; Zhao *et al.,* 2018). Lysak *et al*. (2001) developed a chromosome-specific painting technique in *Arabidopsis thaliana* by pooling bacterial artificial chromosome (BAC) clones derived from a specific chromosome. The BAC-based probes can be used in the comparative studies, as demonstrated in *Brassicaceae* where the developed BACs from *A. thaliana* was used to study genome duplication further, chromosomal rearrangement and evolution in *Brassicaceae* species (Mandakova and Lysak, 2008; Mandakova *et al*., 2010).

Nevertheless, a few authors have discussed the drawbacks of the BAC-based FISH system in species with a large genome due to the extensive amount of repetitive DNA sequences that could prevent the localisation of the single-copy sequences (Janda *et al*., 2006; Dong *et al.,* 2018). The identification of the chromosome with a BAC clones system is also time-consuming as it requires ordered BAC contigs that cover the entire genome of a plant species. The successful FISH-based approach in *Arabidopsis thaliana* relies on the fact that the *A. thaliana* genome is not only very small (125 Mb) (The Arabidopsis Genome Initiative, 2000), but also composed of a sizeable euchromatic regions (Leutwiler *et al*., 1984, Fransz *et al.,* 2000) reflecting that most of the selected BACs contain almost solely single- or low-copy sequences. This approach was also applied in another model plant, *Brachypodium distachyon*, and its related species (Idziak *et al*., 2011; Betekhtin *et al*., 2014). Similarly, *B. distachyon* has a relatively small genome (c.300 Mb) and ordered BAC contigs covering the entire genome are available.

1.3.2.3 Synthetic oligonucleotide (oligo)

In plant species, synthetic oligonucleotide-based FISH was first introduced by Schwarzacher and Heslop-Harrison (1990) by localising biotinylated $(TTTAGGG)_6$ and $(CCCTAAA)_6$, (*Arabidopsis thaliana t*elomeric repeat) on *Hordeum vulgare* (barley) and *Secale cereale* (rye).

The synthetic oligos can be end-labelled with biotin-dUTP or digoxigenin-dUTP (Schwarzacher and Heslop-Harrison, 1990; Cuadrado and Schwarzacher, 1998), or conjugated with a fluorochrome during the synthesis (Danilova *et al*., 2012; Waminal *et al*., 2018). FISH mapping with synthetic oligos was widely used to investigate the chromosomal organisation of repetitive DNA in the genome as well as their roles in the identification of individual chromosomes in *Triticaceae*; wheat, rye and barley (Cuadrado and Schwarzacher, 1998; Cuadrarado and Jouve, 2002; Danilova *et al*., 2012; Tang *et al*., 2014; Danilova *et al*., 2017; Ruban and Badaeva, 2018). The use of oligo-SSR as a probe in FISH also has been demonstrated in *Panax ginseng* (Waminal *et al*., 2018), *Vicia faba* (Fuchs *et al*., 1998) and oil palm (Castilho *et al*., 2000). Various advantages of synthetic oligo probes compared to traditionally prepared probes from cloned satellite repeats have been discussed by several authors. These include a consistent probe quality and the reduction of time and cost for the whole *in situ* hybridisation process. Moreover, a fully sequenced reference genome is not required to develop such probes, as the synthetic oligo probes can be designed directly from computationally identified satellite repeats from genomic sequence data (Lang *et al*., 2018; Waminal *et al*., 2018).

The oligo synthesises probes can also be custom-designed from single-copy DNA sequences. Although a large number of single copy oligos may be required to visualise a specific chromosomal region (Boyle *et al*., 2011; Yamada *et al*., 2011; Beliveau *et al*., 2012), oligos specific to a chromosomal region or to an entire chromosome can be computationally identified and synthesised in parallel as a pool (Beliveau *et al*., 2012; Han *et al*., 2015). Each oligo in the pool can be added with sequence tags at both ends during synthesis, which allows PCR-based amplification of the entire pool (Beliveau *et al*., 2012; Han *et al*., 2015). Subsequently, FISH probes can be generated from the pool *via* amplification of oligos labelled directly with a fluorochrome or indirectly with biotin-dUTP

or digoxigenin-dUTP (Beliveau *et al*., 2012; Han *et al*., 2015; Albert *et al*., 2019). Thus, each synthesised oligo pool can be used as an infinite probe resource since the synthesised DNA (< 500ng) can be used for up to a million FISH applications (Han *et al*., 2015).

1.4 Bioinformatics techniques for repetitive DNA identification

To date, remarkable progress has been made in understanding repetitive DNA in genomes due to the introduction of next-generation sequencing (NGS) technologies and corresponding bioinformatic approaches. Lerat (2010) and Girgis (2015) summarised the bioinformatics approaches and databases for the identification and classification of repetitive DNA sequences and classified them into four general groups based on the usage and approaches.

The first group referred to as '*homology-based methods*'. This method compares input read sequences with databases of known repetitive sequences such as Repbase, RepeatMasker and PLOTREP (Bao *et al.,* 2015; Smit *et al.,* 2015).

The second group is '*signature-based methods'* which exploit the common structural landscapes of repetitive elements to identify DNA repeats. Each class of repetitive DNA has a set of unique features such as a target site duplication, a poly-A tail, terminal inverted repeats, long terminal repeats, and/or a hairpin loop. The signature of a class of repetitive DNA consists of a subset of these features. A signature-based tool searches a sequence for features comprising the signature of the class of interest. Good examples of signature-based approaches are LTR_STRUC, LTRharvest and RetroTector (McCarthy and McDonald, 2003; Ellinghaus *et al*., 2008; Sperber *et al.* 2009).

The third group is referred as '*de novo* methods'. This method is mainly based on the repetitive nature of transposable elements and other repeats in order to identify new families of repeats (Janicki *et al.,* 2011). All repeat families are assembled by *de novo* methods those meeting thresholds of their copy numbers. These methods either built on *k*-mer frequency (the occurrence of small strings) or self-alignment (self-comparison) based approaches. Examples for self-alignment approaches are ReCon (Bao and Eddy,

2002) and PILER (Edgar and Myers 2005). Repetitive DNA identification using *k*-mer frequencies are based on counting the occurrence of short identical motifs that are present in genome sequences in multiple copies. Examples of tools that are based on *k*-mer analysis are RepeatScout (Price *et al.,* 2005) and Tallymer (Kurtz *et al.,* 2008).

The last group are called '*consensus methods*'. These methods combine repetitive DNA identified by a group of different tools. For example, the REPET program (Flutre *et al*., 2011) utilizes both *de novo* and signature-based methods in its pipeline. RepeatModeler (http://www.repeatmasker.org) is based on ReCon and RepeatScout for the identification of repetitive DNA. Moreover, RepeatExplorer and TAREAN (Novak *et al*., 2010, 2017) also based on combination of signature-based and *de novo* methods in their pipeline

Section 1.4.1 and 1.4.2 reviews the three informatics approaches implemented in this study in order to identify repetitive DNA in the oil palm genome that has potential as a chromosome-specific marker.

1.4.1 *k*-mer analysis

Repetitive sequence content in a genome can be analysed by using the *k*-mer frequency (Bergman and Quesneville, 2007; Marçais and Kingsford, 2011). *k*-mer means length *k* sequence included in the analysed dataset, for example, the sequence CCGTGAAT is an 8-mer and it is only one of the 8-mers positioned in the sequence chain of TTGCTCCGTGAATTGAT. We can explore these sequences in the genome by counting all *k*-mers. Interestingly, *k*-mer counting can be a useful approach for estimation of a repeat libraries completeness and further explore sequences that could not be found in the libraries (Krassovsky and Henikoff, 2014). *k*-mer analysis has been used to count the frequency of DNA sequences of length *k* from raw reads data. It is a suitable tool for measuring genome sizes and correcting sequence errors by using available informatics tools such as Jellyfish (Marçais and Kingsford, 2011) and Tallymer (Kurtz *et al*., 2008) and *findGSE* (Sun *et al.,* 2017). This method is considered an unbiased tool for counting repetitive sequences due to its independence of the genome assembly process (Marçais and Kingsford, 2011).

k-mer has been applied for identifying highly repeated structures from unassembled genome sequences and the correlation between these sequences and the centromeric regions of several mammalian genomes (Alkan *et al*., 2011). Williams *et al*. (2013) counted the repeated DNA sequences in bacteria using this tool. In *Drosophila melanogaster*, *k*-mer frequencies were used for counting repetitive sequences, identifying known transposons and short repeats (Krassovsky and Henikoff, 2014). Recently, various lengths of *Taraxacum* microspecies motifs have been analysed using frequency analysis of all possible sequences, evaluating different lengths and complementing the graph-based outcomes (Salih, 2017). Using NGS data from the sheep genome, major classes of dispersed, tandemly repeated elements and endogenous retroviruses-related repetitive sequences were identified by frequency analysis of short motifs (Mustafa 2018).

1.4.2 Graph-based clustering of the raw read sequence

1.4.2.1 RepeatExplorer

Sequences represented in multiple reads can be clustered using graph-based approaches. RepeatExplorer (Novak *et al.,* 2010; 2013) is a collection of software tools for the characterisation of repetitive elements and is accessible *via* a web interface (www.repeatexplorer.org). Using the algorithm of graph-based clustering developed in the RepeatExplorer, the characterisation of the repetitive DNA can be performed in the computational pipeline without any demand for known reference genome databases. A schematic illustration of the RepeatExplorer components and workflow is shown in Figure 1.7.

The input of the RepeatExplorer pipeline is millions of short reads from next-generation sequencing (NGS) and which are random and non-selective. It conducts an all-to-all pairwise comparison of the reads and groups those reads which share significant sequence similarities into 'clusters.' These clusters mostly represent repeats, because only the reads derived from sequences present in the genome multiple times can produce sufficient number of similarity hits in the low-pass sequencing data (0.01–0.50 x genome coverage is typically used). In principle, the number of reads in each cluster is proportional to the genomic abundance of the corresponding repeat, thus enabling its quantification.

RepeatExplorer has proved to be particularly efficient for repeat identification and characterisation in many eukaryote species; for instance, in *Musaceae family, Fabae* tribe, olive, radish, *Taraxacum* as well as in mammals (Barghini *et al*., 2014; Novak *et al.*, 2013; He *et al*., 2015; Macas *et al*., 2015; Salih 2017;Mustafa, 2018).

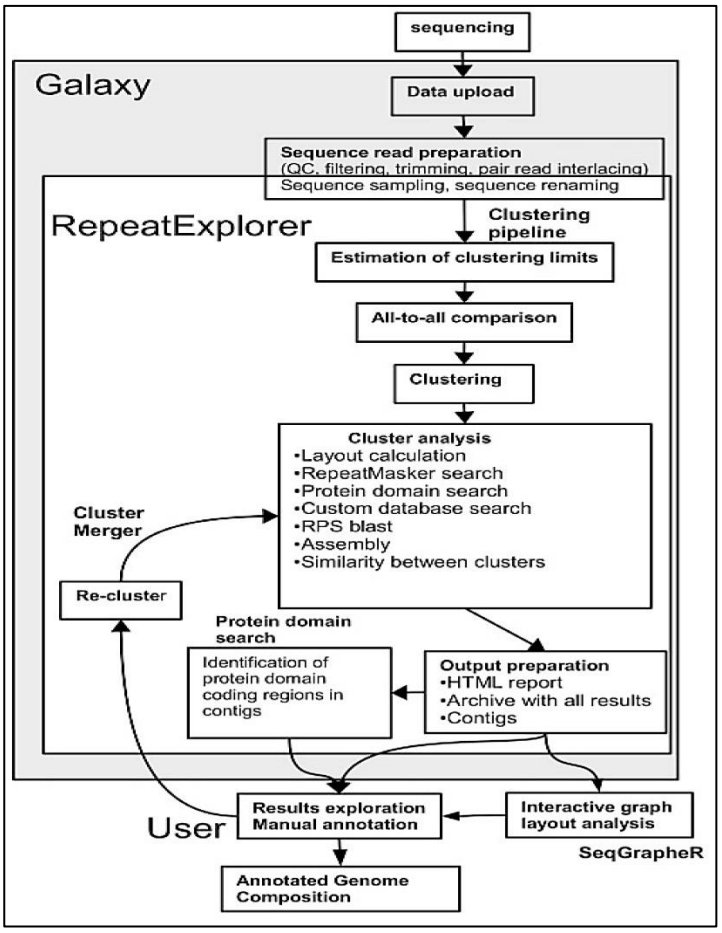


Figure 1.7 Graph-based clustering of repetitive raw reads using the RepeatExplorer pipeline. The pipeline runs on Galaxy, an open source, web-based platform. (Reproduced from Novak *et al.,* 2013)

1.4.2.2 Tandem Repeat Analyser (TAREAN)

Tandem repeat analyser (TAREAN) is a novel pipeline running under the Galaxy environment available *via* the RepeatExplorer server (http://www.repeatexplorer.org/) that effectively detects satellite repeats in the unassembled short reads. The TAREAN builds on the principle of repeat identification by graph-based clustering of NGS reads (RepeatExplorer; Novak *et al*., 2010) (Figure 1.8). The recognition of repetitive DNA clusters is based on circular structures in the graph-based clusters. Repeat monomers from the most frequent *k*-mers are reconstructed through destructing read sequences from their clusters. TAREAN has been efficiently examined through low-pass genome reads of various plant species.

The results from graph-based or *k*-mer identification of satellite repeats reveal the presence and exact genomic abundance of repetitive motifs but give no information about their chromosomal distribution. In some cases, the repeats can be identified in whole genome sequence assemblies (despite the collapse in the number of copies during assembly), but *in situ* hybridisation (FISH) is probed essentially to give detailed information about locations, number of sites and relative abundance between sites. An example of such characterisation was given in *Vicia faba* where three repeats were detected and their loci identified on chromosomes (Novák *et al.,* 2017).

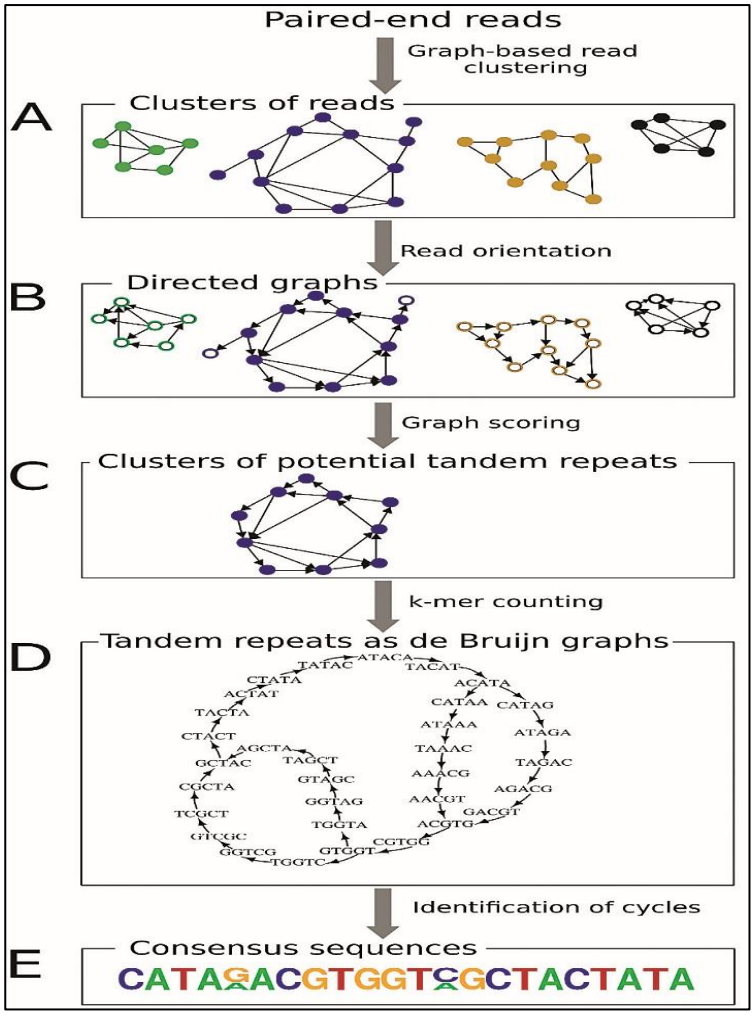


Figure 1.8 Graphic illustration of the identification of candidate tandem repeat sequences using TAREAN analysis workflow. (Source: Novák *et al*., 2017)

1.5 Oil palm genome and chromosomes

Both *Elaeis* species are diploid with n= 16 (Sato *et al*., 1949; Sharma and Sarkar, 1956; Madon *et al.,* 1995; 1998). The 2C nuclear DNA content of *E. guineensis* was within the range of 3.76 – 4.32 pg (Rival *et al*., 1997; Srisawat *et al*., 2005; Madon *et al*., 2008; Camillo *et al*., 2014) slightly lower compared to the *E. oleifera* with 4.43 pg (Camillo *et al*., 2014).

The individual chromosomes of oil palm are challenging to distinguish cytogenetically due to their morphological similarity. There is a limited history of oil palm chromosome identification available. Most of the published works on the identification of *E. guineensis* chromosomes were based on the chromosome length (Sato *et al*., 1949; Sharma and Sarkar, 1956; Madon *et al*., 1995).

Sato *et al*. (1949) defined the species as having 32 somatic chromosomes of which there are four pairs of long chromosomes having sub-median constrictions and twelve pairs of short chromosomes with sub-median or sub-terminal constrictions. However, the same authors also mentioned that the analysed species was only found in tropical America and the chromosome complement is comparatively small. Sharma and Sarkar (1956) graded the 32 *E. guineensis* chromosome with a size range from 1.15 µm to 2.97 µm into three classes. Class 1: three pairs of comparatively long chromosomes, Class 2: four pairs of comparatively medium-sized chromosomes and Class 3: nine pairs of comparatively short chromosomes. Furthermore, the author also described the presence of the secondary constriction on two pairs of the *E. guineensis* longest chromosomes aside of 14 pairs having a median to sub-median primary constriction. Subsequently, 40 years later, Madon *et al*. (1998) assigned both *E. guineensis* and *E. oleifera* chromosomes to Group 1: one pair of the longest chromosomes, Group 2: eight pairs of medium length chromosomes and Group 3: seven pairs of short chromosomes. Madon *et al.* (1998) also found that there was no significant difference in chromosome length of both *Elaeis* species, supporting the ability of the two species to form hybrids (Hardon and Tan, 1969).

Using *in situ* hybridisation of repetitive sequences and the 45S rDNA sequence, Castilho *et al*. (2000) assigned the chromosomes into four groups, comprising one largest

chromosome (which hybridises to 5S rDNA genes), a group of eight medium chromosomes, a group of six smaller chromosomes and the smallest chromosome carrying the 18S-25S rDNA genes and a secondary constriction at the nucleolar organizer region (NOR).

The whole genome sequence of *E. guineensis* has assigned the individual *E. guineensis* chromosomes according to the size of sequence scaffolds which correspond to the linkage group in the selected oil palm mapping population (Singh *et al.,* 2013). The reported length of the assembled oil palm genome was 1.535 Gb, representing some 85% of total genome size although only 43% of this was assembled and assigned to chromosomes. As for *E. oleifera*, the published draft genome was for comparison purposes (Singh *et al*., 2013). Compared to other oil crops, the *E. guineensis* genome is larger than soybean (*Glycine max*; 1.12 Gb) and rapeseed (*Brassica napus*; 0.63 –1.13 Gb), but smaller than corn (*Zea mays*; 2.3Gb), peanut (*Arachis hypogaea*; 2.7 Gb) and coconut (*Cocos nucifera*; 2.42 Gb) (Schmutz *et al*., 2010; Chalhoub *et al*., 2014; Schnable *et al*., 2009; Bertioli *et al*., 2016; Bayer *et al*., 2017; Xiao *et al.,* 2017).

1.6 Challenges and problem statement

To date, the oil palm (*Elaeis guineensis*) is one of the most productive oil-bearing crops worldwide, accounting for 5.5% of global land use for cultivation and producing 32.0% of global oils and fats. The demand for palm oil will continue to rise to meet the per capita intake of the increasing world population. Therefore, it can be assuredly forecast that global demand will remain high and that there will be pressure for yield improvements for decades to come. Hence, there is a constant need to develop new oil palm varieties with enhanced agronomic traits to increase the palm oil yield.

Recent developments, particularly associated with sequence analysis and oil palm genomics, have the potential to continue the highly successful breeding history of oil palm, helping to confirm it as a mainstay of food security as the world population grows. There are also significant challenges ahead to ensure that oil palm plays its part in the future of agriculture while recognising that sustainability and protection of the environment must be balanced with the potential of the oil palm to produce more oil per unit area than any

other crop. Hence, through genetic studies, the continuous cytogenetic manipulation of the chromosome complement in crop plants is one of the valuable approaches available to plant breeders for introducing entirely new variation into crop varieties.

Knowledge of the structures and organisation of the oil palm chromosomes (*E. guineensis* and *E. oleifera)*, as well as in many other economically important crop species, is important for the development of new lines. Such information helps the production of elite hybrids in the way of understanding the causes of some abnormalities or infertility and the characterisation of differences between related species or even breeding lines. The accomplishment of these goals closely relies on having an in-depth understanding of the individual chromosome where the consistent numbering of individual chromosomes within and between the oil palm (*E. guineensis and E. oleifera*) genomes will allow characterisation of any translocations or recombinants in hybrids and integrating the knowledge of the physical chromosomes with the sequence data.

1.7 Aims and objectives

Aim: The overall aim of this thesis is to investigate the structures and organization of the chromosomes of *E. guineensis* by developing robust chromosome markers that able to distinguish 16 pair of *E. guineensis* chromosomes.

Objectives:

1) To characterize large scale architecture of the chromosomes for repetitive DNA to answer the following question: Are there any novel repeats identified from the unassembled (raw reads) sequence compared to the previous study that has not been mapped on the chromosome? Are there any unique chromosome markers that can distinguish the *E. guineensis* chromosome individually? (Discussed in Chapter III)

2) To establish *E. guineensis* FISH-based physical map that serves as a reference karyotype for the species by determining the best approach to develop a robust chromosome-specific cytogenetic marker from a single and low copy sequence in *E. guineensis* genome. (Discussed in Chapter IV)

3) To determine the utility of the developed *E. guineensis* chromosome-specific cytogenetic markers in identifying chromosomes and further establish the karyotype of *E. oleifera.* (Discussed in Chapter V)

4) To determine the utility of the developed chromosome-specific cytogenetic markers across Arecaceae; *Cocos nucifera* and *Phoenix dactylifera*. (Discussed in Chapter V)

CHAPTER II

Materials and Method

2.1 Materials

2.1.1 Plant Materials

The young leaves for total genomic DNA isolation were collected from a field experiment established at Kluang, Johor, Malaysia. The leaves were frozen with liquid nitrogen and kept in -80 °C until needed. The roots were collected either directly from the green house at the University of Leicester (United Kingdom), from a field experiment and oil palm nursery established at Kluang, Johor, Malaysia, and Malaysian Palm Oil Board, Bangi, Malaysia respectively. The plant materials relevant to each chapter are listed therein. This can be found in section 3.3.1, 4.2.1 and 5.2.1

2.1.2 Standard Solutions

Standard solutions used in the study as summarized in Table 2.1. The company names of some chemicals were indicated to show the specificity of some chemicals used in specific experiments.

Table 2.1 List of solutions used in the study. Unless indicated, the solutions were autoclaved and stored at room temperature. The reagents used in this study should be 'molecular biology grade' (have been tested in molecular experiments) or an equivalent high-quality grade for buffer salts and acids.

Experiment: DNA Extraction	
Solutions	**Component and preparation**
CTAB buffer (pH 7.5 - 8.0)	2 % (w/v) cetyltrimethylammonium bromide 100 mM Tris–HCL (pH 8.0) 1.4 M NaCL 20 mM EDTA (pH 8.0) 0.5 M ascorbic acid 0.4 M DIECA (diethyldithiocarbamic acid) No autoclaving, preferably use the freshly prepared
DNA wash buffer	76 % (v/v) ethanol 10mM ammonium acetate No autoclaving
10x TE buffer (pH 8.0)	10 mM Tris (tris-hydroxymethylamino-methane)-HCl 0.1 mM EDTA (ethylene-diamine-tetra-acetic acid)
0.5 M ascorbic acid	88.05 g/L. Filter sterilise and store at 4 °C
0.4M DIECA	68.52 g/L. Filter sterilise and store at 4 °C
Experiment: Gel electrophoresis	
Solutions	**Component and preparation**
6x gel loading buffer	60 % (v/v) glycerol 0.25 % (w/v) bromophenol blue 0.25 % (w/v) xylene cyanol FF No autoclaving and stored at 4 °C Diluted to 1x in deionised H_2O
50x TAE (pH 8.0)	2 M Tris-HCl 50 mM EDTA (ethylenediaminetetraacetic acid; pH 8.0) 5.71 % (v/v) glacial acetic acid Diluted to 1X in deionised H_2O
Ethidium bromide (10 mg/ml)	1 g Ethidium bromide dissolved in 100 ml of sterile distilled water. No autoclaving and stored at 4°C
Experiment: Chromosome preparation	
Solutions	**Component and preparation**
10x enzyme buffer (pH 4.6)	40 mM citric acid 60 mM tri-sodium-citrate No autoclaving and stored at 4°C Diluted to 1X in deionised H2O
1x enzyme solution	1.8 % (w/v) cellulase (72 U/ml; 21947; Calbiochem) 0.2 % (w/v) cellulase (10 U/ml; Onzuka RS) 3 % (v/v) pectinase (13.5 U/ml; P4716; Sigma-Aldrich) Prepared in 1X enzyme buffer No autoclaving and stored at -20 °C

Table 2.1 continue

Experiment: Chromosome preparation	
Solutions	**Component and preparation**
2 mM 8-hydroxyquinoline	0.29 g of 8-hydroxyquinoline in 1 L of distilled water. No autoclaving and stored in the dark at 4°C for up to 1 year Note: the powder might take several hours to dissolve.
Experiment: Colourimetric dot-blot	
Solutions	**Component and preparation**
Buffer 1	100 mM Tris-HCl (tris(hydroxymethyl)aminomethane)(pH 7.5) 15 mM NaCl
Buffer 2	0.5 % (w/v) blocking reagent (Roche Diagnostic) Prepared in buffer 1
Buffer 3	100 mM Tris-HCl (pH 9.5) 100 mM NaCl 50 mM MgCl2
Experiment: Fluorescent *in situ* hybridisation	
Solutions	**Component and preparation**
20 % SDS (Sodium dodecyl sulphate)	2 g (SDS) in 10 ml distilled water
20x SSC (pH 7.0) (Saline Sodium Citrate)	0.3 M NaCl, 0.03 M Sodium citrate.
50 % Dextran sulphate	50 g dextran sulphate in 100 ml distilled water Filter sterilized and stored at -20°C.
100 µg/ml DAPI (4', 6-diamidino-2-phenylindole	5g DAPI dissolved in 50 ml molecular grade water (stock; 100 µg/ml) For final working concentration (4µg/ml) stock was diluted with McIlvaine's buffer No autoclaving and stored at -20°C
McIlvaine's buffer (pH 7.0)	0.1 M citric acid 0.2 M di-sodium hydrogen phosphate
Detection buffer	4X SSC, 0.2 % (v/v) Tween 20 No autoclaving and preferably use the freshly prepared
Blocking solution	5 % BSA prepared in detection buffer No autoclaving and preferably use the freshly prepared

2.2 Methods

2.2.1 Isolation of genomic DNA

Total genomic DNA was isolated from plant spear leaves using a modified Cetyltrimethylammonium bromide (CTAB) method (Doyle and Doyle, 1990). Approximately 5 g leaves were ground to a fine powder in liquid nitrogen using mortar and pestle and placed in 50 ml tubes containing 25 ml of 2x pre-heated modified CTAB lysis buffer. The mixture was then incubated at 60 °C for 30 minutes in a shaking water bath and left to cool at room temperature for 15 minutes. 10 ml of chloroform: isoamyl alcohol (24:1) was added, and the mixture was mixed gently by inverting the tube. This is followed by centrifugation at 10 000 rpm for 15 minutes at 25 °C. The upper aqueous phase was carefully transferred into a new 50 ml tube, and subsequently, 0.6 volume of cold isopropanol was added and mixed by inverting the tube a few times and kept at -20 °C for at least 1 hour.

The following day, the extract was centrifuged at 12 000 rpm at 4 °C for 15 minutes. The supernatant was discarded, and the drained pellet was re-suspended in 5 ml of wash buffer and left at room temperature for 1 hour. The wash buffer was carefully poured off, and the pellet was dried in a speed vacuum for 20-30 minutes. The dried DNA pellet was then dissolved in 2-4 ml TE buffer, depending on the size of the pellet, followed by incubation at 50 °C in a shaking water bath until it dissolves (about 2-3 hours). 2.5 µl of RNase (10 mg/ml) was added to 2 ml of DNA sample and further incubated at 37 °C for 30 minutes followed by addition of 2 ml ammonium acetate (7.5 M, pH 7.7). The tube content was then mixed and left on ice for 20 minutes followed by centrifugation at 4 °C, 12 000 rpm for 15 minutes. The upper aqueous was transferred to a new tube. Two volumes of cold ethanol (99.8%) were added and mixed by inverting the tube. The sample was stored at -80 °C for 45 minutes and centrifuged at 10 000 rpm for 15 minutes at 4°C. The supernatant was discarded, and the obtained pellet was washed with 10 ml of 70% ethanol and further centrifuged at 10 000 rpm for 15 minutes at 4 °C. The drained pellet was dried in a speed vacuum for 30 minutes and re-suspended in 1-2 ml of TE buffer pH8.0. The DNA was kept at 4 °C for further use.

2.2.2 Gel electrophoresis

Multipurpose agarose gels (Bioline) were used for routine gel electrophoresis. The agarose, 0.8 % or 1 % (w/v), was dissolved in 1X TAE with a microwave oven. The dissolved agarose was cooled down before adding the ethidium bromide (10 mg/ml) inside the fume hood to a final concentration of 0.5 µg/ ml. The gel was then cast and immersed in a gel electrophoresis tank with 1 x TAE as the running buffer. Samples were then mixed 2.5:1 (v/v) with 1 x loading dye, pipetted into the wells and the gel was run at 80-100 V for 1-2 hours. The gel was visualized with a gel documentation system (Gene Flash; Syngene Bio Imaging).

2.2.3 Quantitation of DNA

2.2.3.1 Gel electrophoresis

The concentration and integrity of DNA extracts was determined by running 5 µL of each sample on 0.8 % (w/v) standard agarose (Bioline), alongside HyperLadder^{TM} 1kb (Bioline)

2.2.3.2 Spectrophotometry

The concentration (ng/µl) and purity ratios (A260/A280; A260/A230) of DNA extracts were measured using a NanoDrop® ND-2000 spectrophotometer (Thermo Scientific). The spectrophotometer was first blanked using 1 µL TE buffer (pH 8.0) and readings were taken using 1 µL of each sample. The genomic DNA regarded as good quality if the OD 260/280nm ratio between the range of 1.8-2.0.

2.2.3.3 Dilution and storage

Isolated genomic DNA was diluted with TE buffer to a working concentration of 50 ng/uL and stored at -20 °C until use.

2.2.4 Primer design and polymerase chain reaction (PCR)

2.2.4.1 PCR amplification

The PCR primer pairs were designed using Primer 3 (Rozen and Skaletsky, 1999) within Geneious program. The designed primers relevant to each chapter are listed therein. These can be found in section 3.3.4 (Table 3.1) and section 4.2 (Table 4.1 and Table 4.2)

The final concentration of the PCR reaction components are described explicitly in Chapter III (Section 3.3.4) and Chapter IV (section 4.2.2). Unless mentioned, generally the amplification of genomic DNA by PCR was performed using the following programme: Initial denaturation for 3 minutes at 95 °C; followed by 30 cycles of denaturation at 95 °C for 30 seconds (s), annealing for 30 s at optimized annealing temperature (depending on the primers) and extension at 72 °C for 1 minute 30 s; finally followed by final extension for 1 minute at 72 °C. PCR products were then separated on a 1 % agarose gel electrophoresis.

2.2.4.2 PCR amplification of telomeres

For amplification of telomeres, the reaction was set up in 50 µl total volume containing 38.7 µl ddH2O, 5 µl of 10 x Buffer A (Kapa Biosystems), 2 µl 10 mM dNTP Mix, 2 µl of each 10 µM telomere forward and reverse primers and 0.3 µl of 5 U/µL KAPA Taq DNA Polymerase (Kapa Biosystems). The PCR cycling conditions used a 'touch down' programme and consisted of 3 minutes initial denaturation at 95 °C, followed by 7 cycles of denaturation (94 °C, 2 minutes), annealing (66 °C, 30 s decreased 1°C during each of the six consecutive cycles until it reaches 60 °C) and primer extension (72 °C, 45 s). The final step was a 2 minutes extension at 72 °C and held at 16 °C.

2.2.4.2 Purification of PCR products

PCR products were visualised on a 1 % (v/w) standard agarose gel. Amplicons were purified and other PCR components removed from the reaction mixture using the NucleoSpin® Gel and PCR Clean-up kit (Machery-Nagel), following the manufacturer's instructions. Purified amplicons were then assayed using a NanoDrop® ND-2000 spectrophotometer and stored at -20 °C until use, up to a maximum of 1 month.

2.2.5 Sequencing of PCR amplicons

Selected clean PCR products were sent along with custom primers to GATC Biotech (Germany). 5 µl (final concentration within 15-30 ng/µl) of purified PCR products were pre-mixed with 5 µl (5 µM/µl) of either forward or reverse primers added together into an Eppendorf tube. The sequencing results were further analysed using Geneious (Kearse *et al.* 2012; http://www.geneious.com).

2.2.6 DNA probe labelling

2.2.6.1 Random priming

Total genomic DNA and probes larger than 500 bp in size were labelled with digoxigenin-11-dUTP (Roche) or biotin-16-dUTP (Roche) using the Bioprime® Array CGH random priming kit (Invitrogen), following the manufacturer's instructions.

2.2.6.2 Labelling by PCR

Occasionally, clones smaller than 500 bp in size were labelled by PCR. Amplifications were conducted in a reaction mixture containing: 1x Buffer A (KAPA Biosystem), 1.5 mM $MgCl_2$ (Bioline), 0.4 µM of each M13 forward and reverse primers, 0.4 mM dNTP mix (Bioline), 20 µM digoxigenin-11-dUTP or biotin-16-dUTP (Roche), 0.5 U Taq DNA polymerase (KAPA Biosystem), and 100 ng probe, made up to 50 µL with ddH_2O.

2.2.6.3 Purification

Labelled probes were purified using the BioPrime® Purification Module (Invitrogen), following the manufacturer's instructions. They were then stored at -20 °C and were stable for at least a year.

2.2.6.4 Testing the incorporation of labelled nucleotides (dot-blot)

The incorporation of labelled nucleotides (digoxigenin-11-dUTP or biotin-16-dUTP) within probes was tested using a colourimetric dot-blot, according to Schwarzacher & Heslop-Harrison (2000). The nylon Hybond-N^+ membrane (Amersham) was soaked in buffer 1 for 5 minutes and then dried between the filter paper. The samples were bound to the pre-washed charged nylon membrane by applying 1 µl of the probe onto it. In the petri dish, the membrane was then washed twice with gentle shaking, first in buffer 1 for 1 minute and then in buffer 2 for 30 minutes. The buffer was poured off and the probes were then exposed to alkaline phosphatase conjugates by applying 0.5 ml antibody solution (1.5 U/ml anti-digoxigenin-AP (Roche) and 2 U/ml streptavidin-AP-conjugate (Life Technologies) in buffer 1 to the membrane, covered with the petri dish lid, and incubated for 30 minutes at 37 °C in the dark, gently shaken. The membrane was then washed twice, first in buffer 1 for 15 minutes and then in buffer 3 for 2 minutes. The conjugated alkaline phosphatase was then provided with a substrate by applying 1.5 ml detection solution (0.33 mg INT/BCIP

in buffer 3; Roche) to the membrane and left for 10 minutes at room temperature in total darkness. The incorporation of labelled nucleotides was then detected by the degree of coloured product released by the enzymatic reaction.

2.2.7 Chromosome preparation

Chromosome spreads of oil palm were prepared from adult palm planted in the field and seedlings from the oil palm nursery. For the adult palm, the palm tree needs to be mulched at least one month before collecting the roots to allow newly generated roots to grow. As for oil palm seedlings, daily watering for at least five days is needed to allow good quality of new roots to regenerate. Mulching process and root collection from adult palm trees as well the seedlings as shown in Figure 2.1. For coconut and date palm, re-potting and daily watering were performed for at least two months before the roots were suitable to harvest.

2.2.7.1 Accumulation and fixation of metaphase chromosomes

Chromosome preparation of the plant materials was prepared using a modified technique adapted from Schwarzacher & Heslop-Harrison (2000) and Madon *et al.* (1995). Actively growing, white, root tips were collected from the potted plant (seedlings) or the field (adult palm) between 8.30 am- 10.30 am. The roots were then pre-treated in 2 mM 8-hydroxyquinoline (BDH Chemicals) for 5-6 hours at 18 °C to arrest the metaphase. The roots were then fixed in 3:1 (v/v) ethanol: glacial for 24 hours at 4 °C. After the fixation, the roots were then transferred to 70 % ethanol and kept at -20 °C for at least three months.

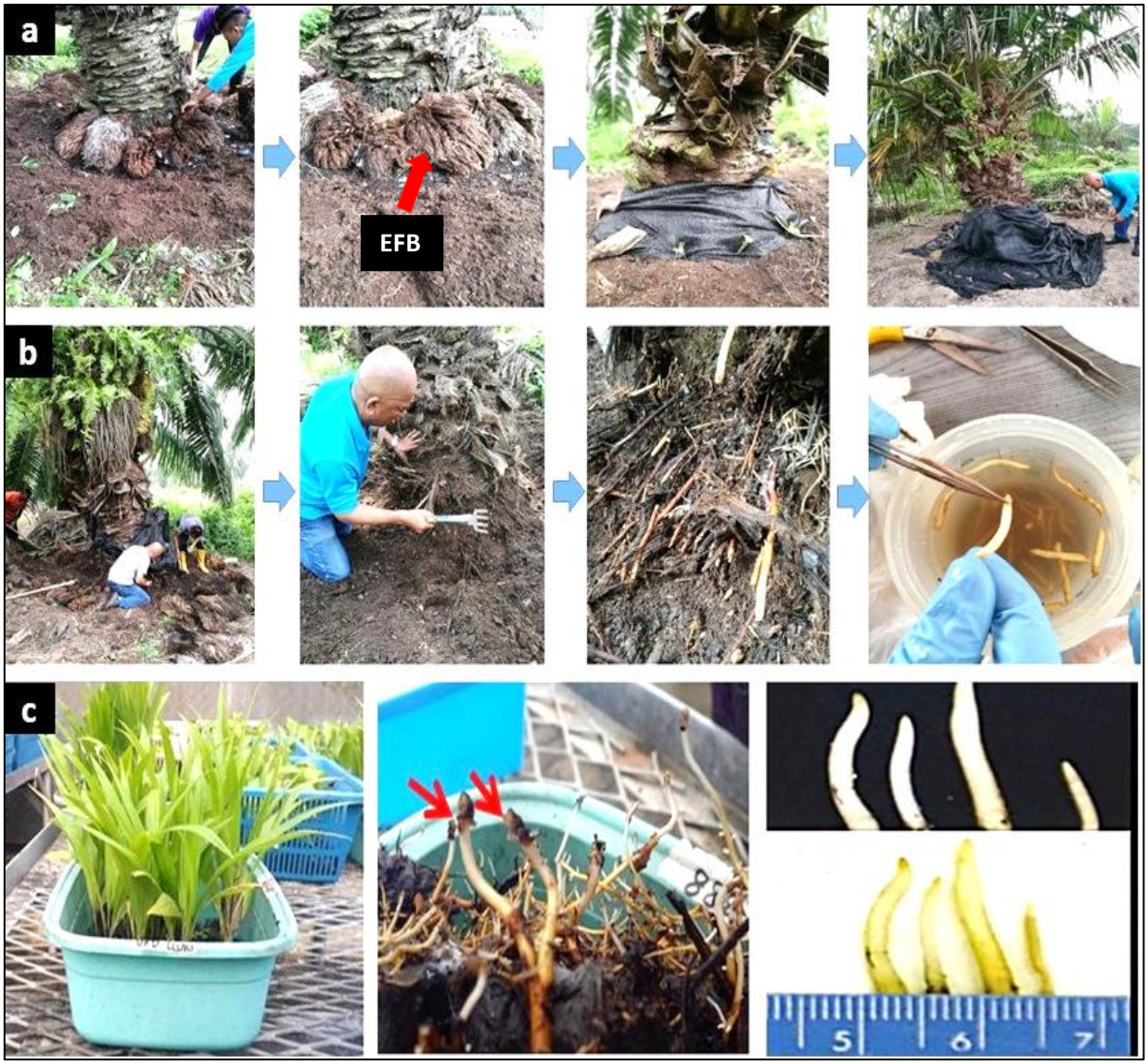


Figure 2.1 Collection of root tips from oil palm. a) Mulching process of adult oil palm. Oil palm basal area was dug (c. 1 ft.) and placed with empty fruit bunch (EFB). The EFB enclosed area was covered with plastic mesh to avoid any disruption from animal or environment. b) Collection of the roots from the mulched oil palm roots after one month. New roots were regenerated from the secondary oil palm roots. c) Example of oil palm seedling used for the root collection. Red arrow showing the meristematic region of oil palm roots covered with the root cap. The last image showed the good quality of the root tips (in scale) with a whitish/opaque colour.

2.2.7.2 Squash preparation of plant chromosomes

i) Aceto-orcein staining

Aceto-orcein squashes of root apical meristems were conducted according to Bailey & Stace (1992). Root tips were hydrolysed in 5 N HCl at room temperature for 10 minutes and then transferred to 70 % ethanol until use. Next, the root tips were dissected, stained, and squashed in aqueous 2 % (w/v) aceto-orcein (Sigma-Aldrich). Cells were observed under bright field on a Zeiss Universal microscope and somatic chromosome number recorded in at least five well-spread metaphases from different root tips.

ii) Air-dried preparations

Root tips were washed twice in a citric acid-citrate buffer until the root sink (5-10 minutes each) to remove all the fixative. The root was then digested in enzyme solution for 3-4 hours at 37 °C. Root tips were then placed on glass slides in a drop of acetic acid (60 %) and left for 1 minute to increase the dispersion of the cytoplasm. About 1-2 mm of the root tips was dissected and transferred to new Polysine^{TM} glass slide in a drop of 60 % acetic acid. The cells were teased out from the remaining meristematic terminal and spread using the fine needle. The coverslip was applied to the material carefully, and the cells were dispersed by tapping the coverslip gently with a flat back of pencil (or any stick that has a flat end). The cells were briefly heated under the flame and squashed using thumb pressure. Slides were scanned for quality under a phase contrast microscope (Zeiss Universal) and preserved by freezing on dry ice, removing the coverslips and allowing them to air-dry. Chromosome spread preparation can be used after overnight or can be stored desiccated in -20 °C until use.

2.2.8 Fluorescent *in situ* hybridisation

Fluorescent *in situ* hybridisation (FISH) was carried out according to Schwarzacher and Heslop-Harrison (2000) with modifications that are suitable for the species used in this study.

2.2.8.1 Pre-treatment of chromosome preparation

i) Re-fixation

Chromosome preparations were re-fixed in fresh 3:1 (v/v) ethanol to glacial acetic acid for 30 minutes at room temperature, washed twice in 100% ethanol for 5 minutes each and left to air-dry.

ii) RNase treatment

Extraneous RNA was removed by applying 200 µl RNase solution (100 µg/ml in 2x SSC; Sigma-Aldrich) to the chromosome preparations, placing a large plastic coverslip (25 x 30 mm) on top, and incubated for 1 hour at 37 °C in a humid chamber. The chromosome preparations were then washed twice in 2X SSC. (Note: coverslips were always removed in the first washing step following treatments)

iii) Pepsin treatment

The slides were pre-treated with 10 mM HCl for 2 minutes and excess cytoplasm was removed by applying 200 µl pepsin solution (5-10 µg/ml in 10 mM HCl; 3200-4500 U/mg; Sigma-Aldrich) to the chromosome preparations, covered with plastic coverslip, and incubated for 20-30 minutes at 37 °C in a humid chamber. The chromosome preparations were then washed in distilled water for 1 minute followed by washing with 2X SSC twice for 5 minutes.

iv) Formaldehyde fixation

The chromosome preparations were fixed in 4 % formaldehyde (Fisher Scientific) for 10 minutes at room temperature. The chromosome preparations were then washed twice in 2X SSC for 5 minutes.

v) Dehydration

The chromosome preparations were dehydrated through an ethanol series (70 %, 85 % and 100 %, 2 minutes each). The slides were then left to air-dry and checked under a phase contrast microscope. The suitable slides (contains 15 and more metaphases) were further used for *in situ* hybridisation.

2.2.8.2 *In situ* hybridisation

A probe mixture was prepared, containing: 50 % (v/v) formamide (Sigma-Aldrich), 10 % (w/v) dextran sulphate (Sigma-Aldrich), 0.125 % (w/v) sodium dodecyl sulphate (SDS; Sigma-Aldrich), 0.025 µg/mL salmon sperm DNA (Sigma-Aldrich), 2X saline sodium citrate (SSC) and 200 ng of each probe, made up to 40 µl with ddH_2O.

The hybridisation mixture was then denatured for 10 minutes at 80 °C in a water bath or in a PCR machine followed by snap cooling on ice for 10-15 minutes to prevent re-annealing of the single-stranded DNA. The probe mixture was then applied to the chromosome preparations and a small plastic coverslip (22 x 22 mm) placed on top. The chromosome preparations were then denatured for 5-7 minutes at 73 °C on a heated flatbed thermal cycler (Thermo Scientific), allowed to cool to 37 °C, and held there for at least 20 hours to enable the hybridisation of labelled probes to complementary targets on the chromosome preparations.

The temperature of denaturation, the formamide concentration and Na^+ ion amount in SSC limits the hybridisation stringency. The salmon sperm DNA and blocking DNA reducing or removing the non-specific hybridisation. The dextran sulphate used to increase the volume of the mixture without decreasing the concentration of the probe. SDS improves the penetration of probe while the EDTA stops the nucleases. The concentrations of salt and formamide permitted the sequences with homology 75-80 % to form duplexes (Schwarzacher and Heslop-Harrison, 2000).

2.2.8.3 Post-hybridisation washes

The chromosome preparations were given a series of post-hybridisation washes to remove the hybridisation mixture and weakly bounded/unbound probe that will contribute to the background signal on the slide. Prior to washing, the post-hybridisation washing solutions (2X SSC and 0.1X SSC) were prepared and pre-heated to 42-45 °C in a water bath. The slides were briefly washed in 2X SSC at 35-40 °C to remove the coverslip from the slide. This followed by washing the slides once in 2X SSC for 2 minutes at 42 °C, once in a stringent wash solution (in the current study just low stringency was used: 0.1X SSC) for 10 minutes at 42 °C, then in 2X SSC for 5 minutes at room temperature. The slides were then washed once in a detection buffer for 5 minutes at room temperature.

2.2.8.4 Detection of hybridisation sites

Detection of hybridisation sites allows the visualization of the indirectly labelled probes (see section 1.3 for details). Slides were incubated in detection buffer (4X SSC, 0.2 % (v/v) Tween-20) for 5 minutes. This was followed by applying 200 µl of blocking solution (5 % BSA in detection buffer) onto the chromosome preparation to block any non-specific sites that could potentially bind detection reagents. A large plastic coverslip was applied onto the slides and followed by incubation at 37 °C for 30 minutes in a humid chamber. Hybridisation sites for biotin and digoxigenin-labeled probes were detected with 2 µg/ml streptavidin conjugated to Alexa Fluor® 594 (Molecular Probes, Thermo Fisher Scientific) and 4 µg/ml anti-digoxigenin conjugated to fluorescein isothiocyanate (FITC, Roche Diagnostic) respectively. A 50 µl of the antibody in blocking solution was then applied to the chromosome preparations, a small coverslip applied, and the slides incubated for 1 hour at 37 °C. The chromosome preparations were then washed twice in detection buffer for 10 minutes each at 40 °C.

2.2.8.5 Nuclear counterstaining and mounting of slides

The cells were simultaneously counterstained and mounted with mixture of DAPI (4, 6-diamidino-2-phenylindole) and CITIFLUOR^{TM} AF1 antifade (Chem Lab) [DAPI; 100 µg/ml (6µl) + antifade (97 µl) + ddH2O (97 µl)]. A large glass coverslip (No. 0, 24 x 40mm) was

then placed on top and gently pressed with filter paper to remove the extra antifade. Slides were then stored at 4 °C at least for overnight before imaging.

2.2.8.7 Image acquisition, processing, and analysis

Slides were examined using a Nikon Eclipse 80i fluorescent microscope. Three Nikon filters were used for the observation, UV-2E/C (emission at 435-485) for DAPI, B-2E/C (emission at 515-555) for fluorescein and G-2E/C (emission at 590-650) for Alexa 594. Images were acquired with Nikon DS-Qi1 Digital camera and NIS elements AR, version 3.2 Software. The individual channels were pseudo-coloured to visualise the sites of probe hybridisation. The images were processed using Adobe Photoshop CS5 software (Adobe System Inc., http://www.adobe.com) using cropping and functions that affect the whole image equally.

2.2.8.8 Re-probing of chromosome slides

Occasionally, the chromosome preparations could be used for the second or even third time by re-probing the slides. For re-probing the used chromosome preparations, the slides were incubated at 37 °C for at least 10 minutes to reduce the viscosity of the antifade, and this follows by removal of the coverslip. The preparations were washed twice in detection buffer for 1 hour at room temperature and then incubated twice in 2 x SSC for 5 min at room temperature. The chromosome preparations were then dehydrated using an ethanol series (70 %, 85 %, and 100 %), and left to air-dry. The FISH procedure then continued as above (see section 2.2.8.2 onwards).

CHAPTER III

Analysis of *Elaeis guineensis* repetitive DNA from chromosome cytogenetics through sequence assemblies to raw reads

---

3.1 Introduction

Eukaryotic genomes, including most angiosperm plants, are known to include a high proportion of repetitive elements. An established classification has grouped this repetitive DNA component of the genome as class I or II transposable elements (TEs), tandemly repeated DNA, endogenous retroviruses and simple repetitive DNA (Heslop-Harrison and Schmidt, 1998; Wicker *et al*., 2007; Biscotti *et al.,* 2015). The specificity of repetitive DNA is widely discussed in various species, being a chromosome-specific, species-specific or genus-specific and also presence on a centromeric or sub-telomeric in nature (Cermak *et al*., 2008; Klemme *et al*., 2013; Mehrotra *et al.,* 2014; Santos *et al.*, 2015; Garrido-Ramos 2015; Said *et al*., 2018)

Over the last few decades, molecular and cytological analysis approaches have been the key methods for characterization of various repetitive DNA sequence elements in plants. This includes identification of abundant restriction satellite DNA fragments and characterization of abundant clones from a clone library sequences and using degenerated primers for amplifying the reverse transcriptase (RT) genes (Schmidt, 1999; Kurtz *et al.,* 2008; Biscotti *et al.,* 2015). Nevertheless, continuous development of next-generation sequencing (NGS) technologies which can generate up to gigabases of sequence data has presented new opportunities for the investigation of repetitive elements in plant genomes (reviewed in Weiss-Schneeweiss *et al*., 2015). *In silico* analysis of repetitive DNA with various bioinformatics tools is becoming increasingly important with non-selective, high throughput shotgun sequencing (NGS) technology. Analysis of repetitive elements from

genome assemblies often collapsed tandem repeats (Kurtz *et al.,* 2008; Moreira-Filho *et al*., 2011; Trangen and Salzberg, 2011). Identification of repetitive DNA from unassembled (raw) NGS data is cloning-free, thus avoiding the potential bias caused by difficulties in propagating some repeat types in bacteria (Song *et al*., 2001). Theoretically, repeat detection from the unassembled genome is based on evaluation of the identical or similar sequence reads frequencies, which increase as the genomic copy numbers of corresponding repetitive elements increase. This approach proved to be efficient for identification of repeats in 23 species of the monophyletic legume tribe *Fabeae* (Macas *et al*., 2015), olive (Barghini *et al*., 2014), radish (He *et al*., 2015), *Solanaceae* (Bombarley, 2017) and apomictic *Taraxacum* (Salih 2017).

Elaeis guineensis (oil palm) is a major cultivated crop and the world's largest source of edible vegetable oil. Very little had been known about the repetitive DNA present in oil palm. Castilho *et al.* (2000) pioneered the first investigation of repetitive DNA component in oil palm. A number of major repetitive sequences families were isolated from the oil palm genome and characterized by sequencing, Southern and *in situ* hybridisation. The obtained results were referred for the whole genome sequencing of the oil palm (Singh *et al*., 2013). Thirteen years apart from the first documented repetitive DNA to whole genome sequence data, only two research reported on the significance of the retrotransposon in oil palm genome (a very positive reference to the work of Price *et al*., 2003 and Kubis *et al*., 2003). Remarkably, the significance of the retrotransposon in the oil palm breeding was proven when DNA hypomethylation of a LINE retrotransposon showed relatedness to rice *Karma* gene in all oil palm mantled clones (Ong-Abdullah *et al.,* 2015)

Since the release of the oil palm genome, the large-scale characterization of the oil palm repetitive DNA was based on the *in silico* characterization of the assembled genome (Beule *et al*., 2015; Filho *et al*., 2017). Here, the first phase of this study aimed to build a picture of large scale architecture of the chromosomes from repetitive DNA to answer the following question: Are there any novel repeats identified from the unassembled (raw reads) sequence compared to the previous study that has not been mapped on the chromosome? Are there any unique chromosome markers that can distinguish the *E. guineensis* chromosome individually?

3.2 Materials and methods

3.2.1 Plant materials

The oil palm (*E. guineensis*) materials used for developing the DNA probes were published by Singh *et al.* (2013) and are currently maintained at the MPOB Research Station, Kluang, Johor, Malaysia. Meristematic root tips were collected from three Pisifera palms (0.182/77, 0.182/30 and 0.182/7). Genomic DNA (0.182/77) was extracted and purified from a spear leaf using the modified CTAB method as described by Doyle and Doyle (1990).

3.2.2 DNA sequence data

The *E. guineensis* raw DNA sequence paired-end reads used in this study were from shotgun genomic fragment libraries with a total size of 38.3 gigabytes (GB) (127,310,950 reads) generated with an Illumina HiSeq 2500 Rapid Mode Sequencer.

3.2.3 Identification and characterization of repetitive elements

A combination of manual approaches and automated programs were used to identify and classify repeated sequences from *E. guineensis* genome data. Basic analyses of the assemblies (contig assembly and nucleotide sequence alignment) were performed on Ubuntu Linux 13.10 with Geneious (Kearse *et al.*, 2012; http://www.geneious.com/). The programmes RepeatExplorer (Novák *et al.,* 2013), and Tandem Repeat ANalyser (TAREAN) (Novák *et al.,* 2017) were used for graph-based clustering of repeated sequences in the raw reads. Additionally, (BLAST); Basic Local Alignment Search Tool (Altschul *et al.,* 1990), and the database of conserved protein motifs of retroelements (Hansen and Heslop-Harrison, 2004) were used for characterising repeat sequences.

3.2.3.1 RepeatExplorer

The whole genome raw data (38.3 GB) was split to several sub-sets containing approximately 2 GB data due to the size limitation of the program. A sub-set of data containing 1,717,952 reads was randomly selected and uploaded on to the Galaxy/RepeatExplorer server for graph-based sequence clustering and repeat identification.

3.2.3.2 *k*-mer analysis

This analysis performed *de-novo* estimation of the highly frequent repeats from the whole 38.3 GB sequence reads. The *k*-mer analysis was performed on Ubuntu Linux 13.10. The Jellyfish *k*-mer counting program Version 2.1.3 (Marcais and Kingford 2011), was used to count canonical *k*-mers where *k* was equal to 16, 32, or 64. *k*-mer analysis output which consists of the most abundant motifs for each *k*-value (at the threshold of 100, 1000 and 10 000 or more) were further assembled (*de-novo*) using Geneious (Kearse *et al.*, 2012) to generate longer contigs. Generally, the longer contigs represent overlapping *k*-mers from the short, abundant repeat motifs.

Homology search for the contigs was further performed with BLASTn against the NCBI database and a customized retroelement motif database (Hansen and Heslop Harrison, 2004) for conserved retroelement domains and internal protein motifs

Three basic commands to numerate *k*-mers and their occurrence counts (e.g., 16 mer):

1) **Counting all *k*-mers of the whole paired-end raw reads**

   jellyfish count -m 16 -s 100M -t 10 -C reads.fasta

   This will count canonical (-C) 16mers (-m 16), using a hash with 100 million elements (-s 100M) and 10 threads (-t 10) in the sequences in the file reads.fasta. The output is written in the file 'mer_counts.jf' by default

2) **Converting the *k*-mer count output from binary format to readable file**

   jellyfish dump mer_counts.jf > mer_counts_dumps.fasta

3) **Extracting the most abundant motifs for each value of *k*-mer repeated at selected thresholds (e.g. 1,000 or more) times in the raw reads**

   grep -A 1 --no-group-separator '[1-9][0-9][0-9][0-9]' mer_counts_dumps.fasta > output.fasta

3.2.4 Conversion of repetitive sequences into FISH probes

The PCR primer pairs derived designed using Primer 3 (Rozen and Skaletsky, 1999) within Geneious program. Details of the nuclear gene, transposable element, RepeatExplorer repetitive clusters, and *k*-mer derived repeats as listed in Table 3.1. All the DNA probes were amplified from *E. guineensis* genomic DNA by PCR using the following programme: 3 minutes at 95 °C followed by 30 cycles of (30 seconds at 95 °C; 30 seconds of optimized annealing temperature; 1 minute 30 seconds at 72 °C) and final extension for 1 minute at 72 °C. PCR products were separated on a 1 % agarose gel electrophoresis and isolated with the E.Z.N.A Gel Extraction Kit (Omega) as described by the manufacturer. The purified amplicon was stored at -20 °C until use. PCR products were labelled with digoxigenin-11-dUTP (Roche Diagnostic, Basel, Switzerland) or biotin-16-dUTP (Roche Diagnostic) using the BioPrime® Array CGH Labelling System (Invitrogen, California, USA) according to the manufacturer's instructions.

Table 3.1 Nuclear gene, transposable element, RepeatExplorer repetitive clusters, and *k*-mer derived repeats use in the study. Primer/oligo name, sequence, expected band sizes, annealing temperature and source of the primers/oligos are shown.

Primer ID	Primer Sequence (5'-3')/ Oligo Sequence (F: Forward; R: Reverse)	Annealing Ta (°C)	Product size (bp)	Source
Ty1	F: ACNGCNTTYYTNCAYGG R: ARCATRTCRTCNACRTA	42	270	Flavell *et al*. 1992
GyRT-1 GyRT-4	F: MRNATGTGYGTNGAYTAYMG R: RCAYTTNSWNARYTTNGCR	39,44,46	420	Kubis *et al.* 1998
BEL-1MF	F: RVNRRANTTYCGNCCNNATHAG R: GACARRGGRTCCCCCTGNCK	48,49	500-600	Kubis *et al.* 1998
LC_Cl12	F: CCTATGATCATTTTGGTCAAGGGG R: TGTCCGTTACACTCGTTTGC	56	1400	RepeatExplorer
LC_Cl42	F: GAAGAGAAAGTGGAGCGTGC R: GTGACTGGTATGCGACTTCACG	58	1200	RepeatExplorer
LC_Cl61	F: CCTAATGTTCTGAGACACGTTCG R: ATGGCTTCGATGCGATGGG	56	1350	RepeatExplorer
LC_Cl83	F: TACAACCCCAAGGCTTGCC R: CCAAGGAGCGGCATCACC	60	1153	RepeatExplorer
pEgKB9_319bp	F: AGGACCTATGTGAACGAGGC R: GGGCGTATCAGCTATCTCACC	58	293	*k*-mer analysis
pEgKB17	F: TGATCGGATCACTTGACTATCGAGC R: CCCACTACCAGACGGTTACAGG	58	598	*k*-mer analysis
pEgKB23	F: CCATCTGATGGATTACCCTGGC R: TTCTATCGAGCATAGGTCACTGTAGG	58	400	*k*-mer analysis
Eg9CEN	F: CCATATGGGTTGGTTGTCC R: ACAGCGACTCATTCTTCTCC	58	350	Malaysian Palm Oil Board (MPOB)
18 rRNA	F: CGAACTGTGAAACTGCGAATTG R: TAGGAGCGACGGGCGGTGTGAA	66	1500	*k*-mer analysis and RepeatExplorer
M13 (5S rDNA from clone pTA794)	F:GTAAAACGACGGCCAGT R: GGAAACSGCTATGACCATG	55	410	Gerlach and Dyer 1980
5S rDNA	[Cyanine3]GTTAAGCGTGCTTGGGTGAGAGTAG TACTACGATGGGTGACC	not applicable	not applicable	*k*-mer analysis
Telomere	[6-Fam] CCCTAAACCCTAAACCCTAAACCCTAAACCCTAA ACCCTAAA	not applicable	not applicable	*k*-mer analysis

Y=C+T; R=A+G; M=A+C; K=G+T; S=G+C; W=A+T; H=A+T+C; D=G+A+T; B=G+T+C; V=G+A+C; N=A+G+C+T

3.2.5 Chromosome preparations

Metaphase spreads were prepared from root tips of *E. guineensis* according to Madon *et al*. (1995) and Schwarzacher and Heslop-Harrison (2000) with modifications. Root tips were pre-treated with 2 mM 8-hydroxyquinoline for 5-6 hr at 18 °C and fixed in 3:1 ethanol: glacial acetic acid (v/v). The fixed roots were then stored in 70 % ethanol at 4 °C until further use. The root tips were washed several times with citric acid-citrate buffer (pH 4.6) and digested at 37 °C for up to 4 hr in enzyme solutions containing 2 % - 4 % (w/v) cellulase (Sigma C1184; final concentration 10-20 U/ml), 0.2 % (w/v) 'Onozuka' RS cellulase (final concentration of 10 U/ml), 3 % (v/v) pectinase (Sigma P4716 from *Aspergillus niger*; solution in 40 % glycerol, final concentration 15-20 U/ml) in citric acid-citrate buffer. Mitotic chromosomes were spread by squashing and heating onto a pre-cleaned glass slide in a drop of 60 % acetic acid under a coverslip, frozen before flicking off the coverslip, and left to air-dry before using for FISH.

3.2.6 Fluorescent *in situ* hybridisation (FISH)

In situ hybridisation was performed according to Schwarzacher and Heslop-Harrison (2000) with minor modifications. A total of 40 µl probe was applied per slide, containing 50 % (v/v) formamide, 20 % (w/v) dextran sulphate, 2X saline sodium citrate (SSC, 0.3 M NaCl, 0.03 M sodium citrate), 0.05 µg of salmon sperm DNA, 0.25 % (w/v) sodium dodecyl sulphate, 0.25 mM ethylenediamine-tetraacetic acid and 25-100 ng probe. The probe mixture was denatured for 10 minutes at 80 °C and immediately transferred to ice. Probe and chromosomal DNA were denatured together on a heated block (Thermo Fisher Scientific) at 72 °C for 5 minutes under plastic coverslips, allowed to cool to hybridisation temperature of 37 °C overnight (minimum 16 hours). A series of post hybridisation washes were carried out with 2X SSC and 0.1X SSC at 42 °C. Hybridisation sites for biotin and digoxigenin-labeled probes were detected with 2 µg/ml streptavidin conjugated to Alexa 594 (Molecular Probes, Thermo Fisher Scientific) and 4 µg/ml anti-digoxigenin conjugated to fluorescein isothiocyanate (FITC, Roche Diagnostic) respectively. DAPI (4, 6-diamidino-2-phenylindole) in CITIFLUOR AF1 (Chem Lab,) antifade solution was used to counterstain the chromosomes. At least two slides with 15 high-quality metaphases were hybridised and analysed for each probe and species combination.

3.2.7 Image acquisition, processing, and analysis

Slides were examined using a Nikon Eclipse 80i fluorescent microscope. Three Nikon filters were used for the observation, UV-2E/C (emission at 435-485) for DAPI, B-2E/C (emission at 515-555) for fluorescein and G-2E/C (emission at 590-650) for Alexa 594. Images were acquired with a Nikon DS-Qi1 Digital camera and NIS elements AR, version 3.2 Software. The individual channels were pseudo-coloured to visualise the sites of probe hybridisation. The images were processed using Adobe Photoshop CS5 software (Adobe System Inc., http://www.adobe.com) using cropping and functions that affect the whole image equally.

3.3 Results

3.3.1 Identification of repetitive DNA in the unassembled *E. guineensis* genome with independent analysis of RepeatExplorer, *k*-mer analysis and TAndem REpeat ANalyzer (TAREAN)

A total of 1.7 million sequence reads were randomly selected from the unassembled sequence of *E. guineensis* to generate repeat clusters using RepeatExplorer software (Novak et *al*., 2013). This analysis resulted in a total of 68,063 repeat clusters and 559,049 single/non-clustered reads. The 68,063 clusters represented different repeat families in *E. guineensis* genome that accounted for 67.5 % of the analysed 1.7 million reads. Among these, 106 clusters that accounted for 57.1 % of the genomic reads were relatively enriched in the *E. guineensis* genome (genome proportion > 0.01 %) (Figure 3.1a).

The LTR retrotransposons were the most abundant repeat families, accounting for 40.67 % of *E. guineensis* genome (Figure 3.1b). Among them, the LTR/*copia* retrotransposon were the most abundant representing 30.31 % of the genome, followed by LTR/*gypsy*, with 10.34 % and long interspersed nuclear elements (LINEs) with the lowest genome proportion (0.02 %). LTR/*copia* retrotransposon are mainly represented by *Angela* lineage (20 %). The rest of the lineages (*Maximus, Tork, Ivana/Oryco, AleI/Retrofit,* and *AleII)* occupies approximately 7.57 % in *E. guineensis* genome. LTR/*gypsy* retrotransposon was represented with three lineages namely *Ogre/Tat* (5.43 %), *Athila* (1.91 %) and C*hromovirus* (1.85 %). Five DNA transposons (DNA/hAT-Tag1, DNA/hAT-Ac, DNA/hAT-

Tip100, DNA/CMC-EnSpm, and DNA/PIF-Harbinger) were identified representing 1.25 % of the genome in which all the individual transposons showed less than 1% of genome proportions. The rest of the repeat families, including simple repeat, ribosome DNA showed a relatively low genome proportion of <1 %. In addition, "unclassified repeats" were also categorised *via* RepeatExplorer with a significantly high percentage (13.75 %).

With *k*-mer analysis, the repetitive DNA sequences were identified from a total of 38.3 GB unassembled sequence of *E. guineensis*. Identification of DNA sequence substrings with different values of motif (*k*) of 16, 32 and 64 mer that is repeated for 100, 1,000 and 10,000 times was performed with Jellyfish (Marcais and Kingford, 2011). All identified short motifs were assembled into longer contigs. The number of generated contigs were 1713, 1800 and 165 for 16-mer, 32-mer and 64-mer respectively. The first top 100 contigs with a length of more than 40 mer were further compared against the NCBI. Homology search against both databases characterises the abundant repeated motif sequence falls into four distinct categories of repetitive DNA namely rDNA, *copia*-like, microsatellite and pEgKB family (Figure 3.2). Telomere repeat, a structural component of the chromosome which consist of the 7 bp of telomeric sequences CCCTAAA/TTTAGGG was only identified from 64-mer.

With the default parameter setting, TAREAN programme was unable to identify any putative satellite and putative LTR from the whole 38.3 GB unassembled sequence of *E. guineensis* genome (Figure 3.3).

Taken together results obtained from RepeatExplorer and *k*-mer derived contigs analysis, the 38.3 GB of unassembled *E. guineensis* genome sequence can be categorized into seven major groups of repetitive DNA namely; LTR/*copia,* LTR/*gypsy,* LINE, DNA transposon, simple repeat/microsatellite, rDNA, pEgKB family, and telomere.

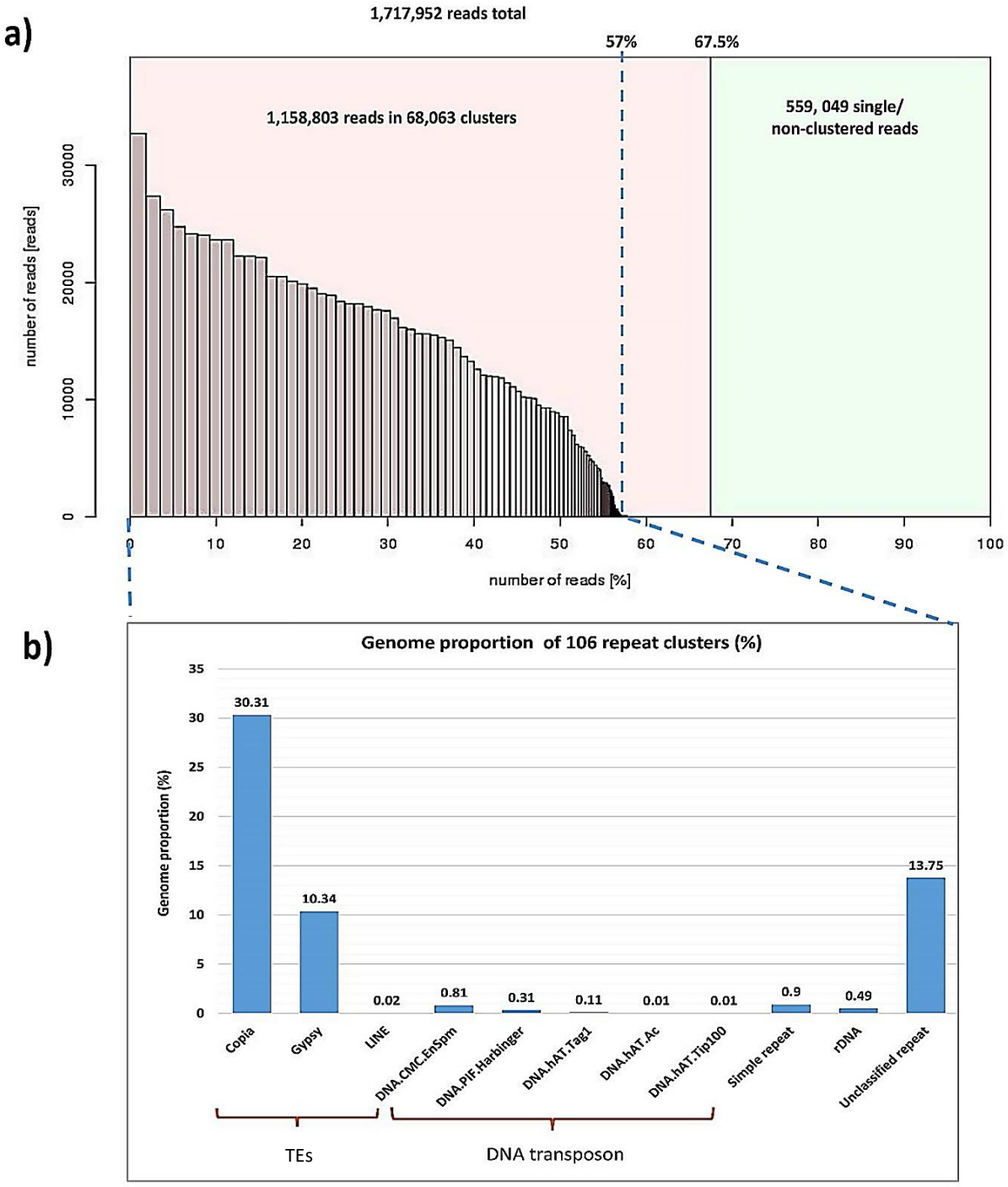


Figure 3.1 Composition and annotation of repetitive DNA in the unassembled genome sequence of *E. guineensis* as identified by RepeatExplorer programme. a) Summary of the contents of repetitive DNA and the single copy reads among which 106 repeat clusters (57%) were analysed further. Bars on the histogram represent individual clusters; bar sizes correspond to the number of reads in the clusters. b) Annotation and the genome proportion of the 106 repeat clusters (TEs; Transposable Elements)

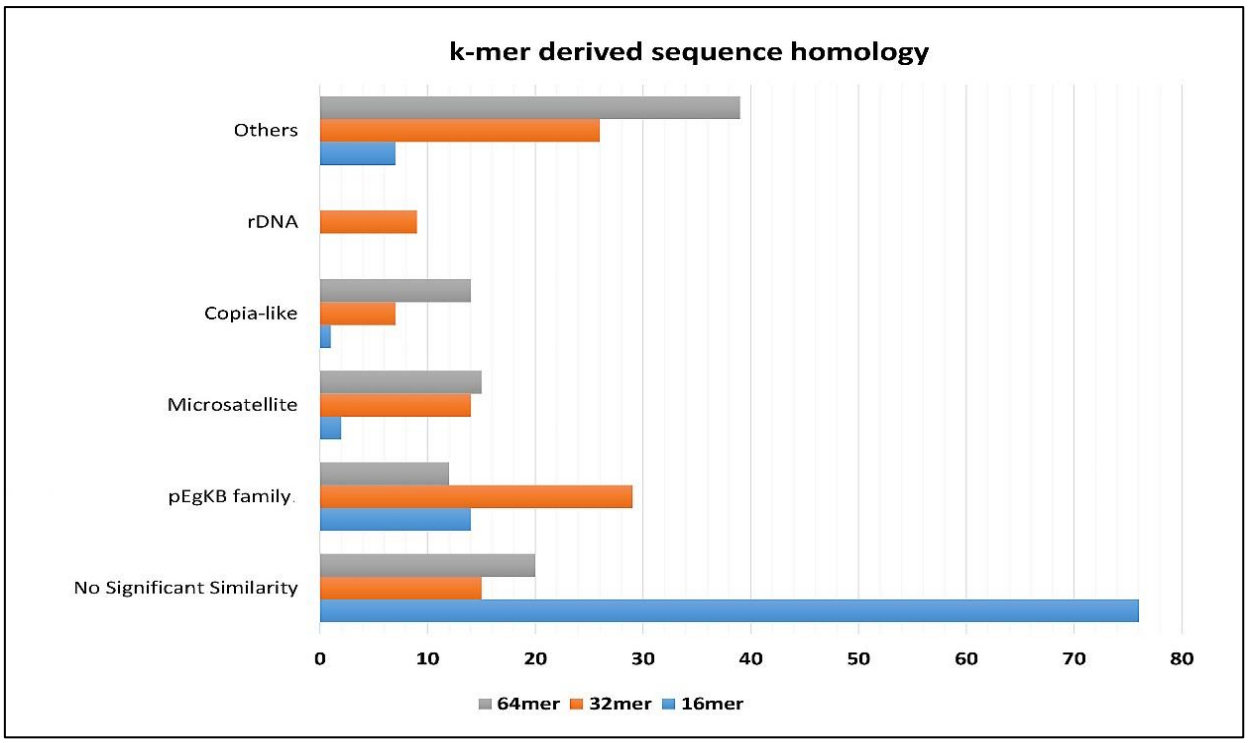
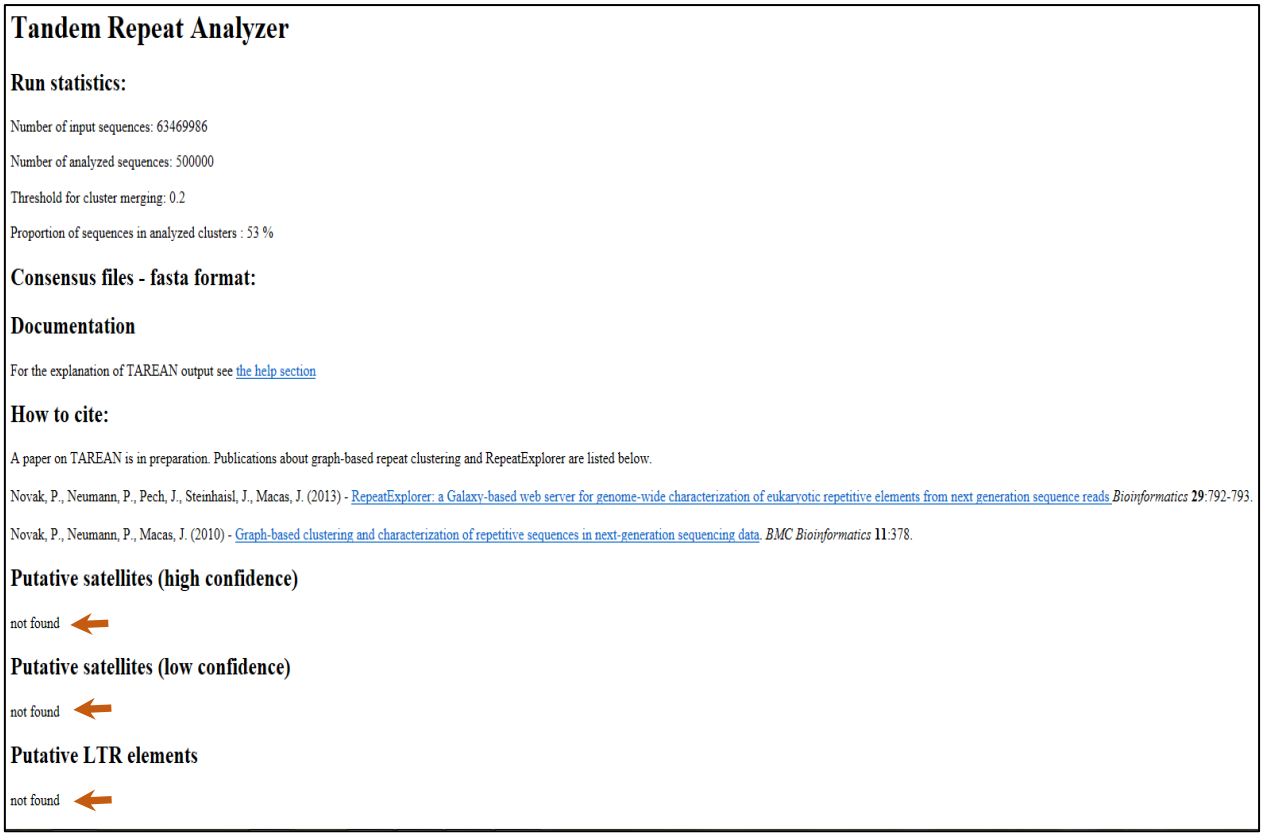


Figure 3.2 Characterization of top 100 of the most abundant nucleotide motif generated by 16, 32 and 64 *k*-mers analysed against NCBI and customized retroelement motif database (Hansen and Heslop Harrison, 2004).



Tandem Repeat Analyzer

Run statistics:

Number of input sequences: 63469986

Number of analyzed sequences: 500000

Threshold for cluster merging: 0.2

Proportion of sequences in analyzed clusters : 53 %

Consensus files - fasta format:

Documentation

For the explanation of TAREAN output see the help section

How to cite:

A paper on TAREAN is in preparation. Publications about graph-based repeat clustering and RepeatExplorer are listed below.

Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. (2013) - RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads *Bioinformatics* **29**:792-793.

Novak, P., Neumann, P., Macas, J. (2010) - Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* **11**:378.

Putative satellites (high confidence)

not found

Putative satellites (low confidence)

not found

Putative LTR elements

not found

Figure 3.3 First page of RepeatExplorer showing the absence of tandem repeat DNA (orange arrow) from unassembled *E. guineensis* genome identified by TAREAN analysis.

3.3.2 Chromosomal localization of retrotransposons in *E. guineensis* chromosome with universal repetitive primer

The retrotransposons of *E. guineensis* were characterised by analysing the Ty3-*gypsy*, Ty1-*copia*, and LINE retrotransposon using PCR primers specific for conserved domains of Reverse Transcriptase (RT) genes of *copia*-like, *gypsy*-like and LINE retroelements.

3.3.2.1 *Copia*

A fragment of the reverse transcriptase gene of the *copia* retrotransposon group was amplified from genomic DNA by PCR using synthetic degenerate oligonucleotide primers corresponding to the peptide sequences TAFLHG and YVDDML (Flavell *et al*., 1992). *In situ* hybridisation of the *copia* PCR product showed signal on all chromosomes at broad centromeric region (Figure 3.4).

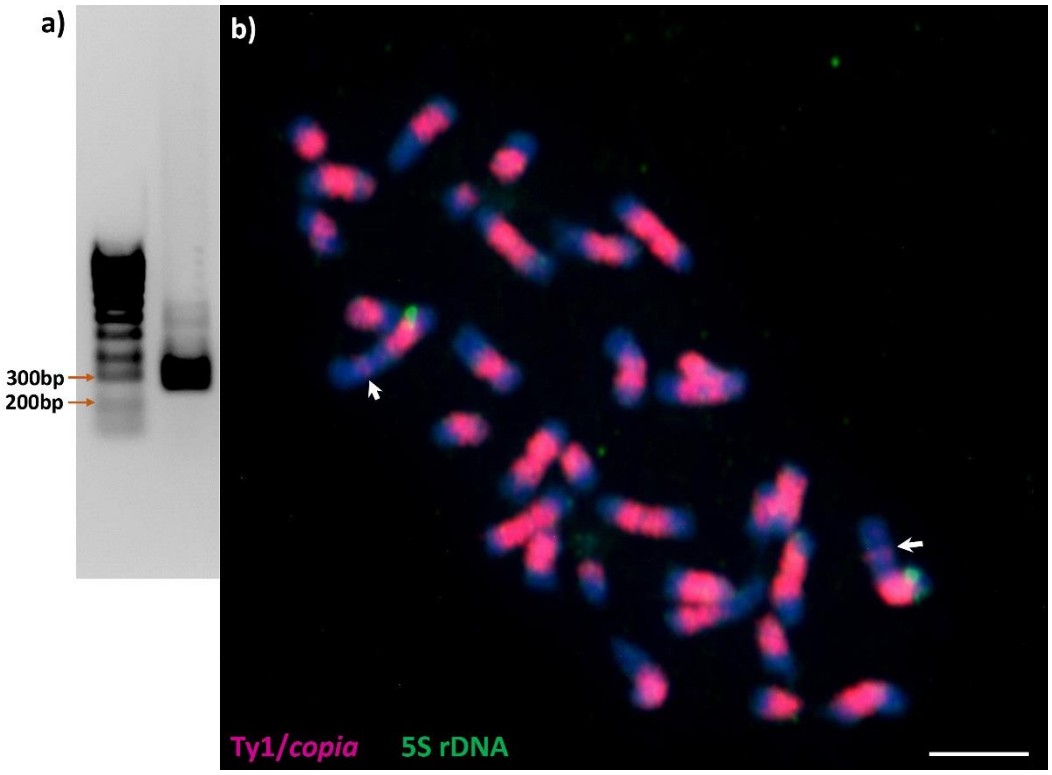


Figure 3.4 Chromosomal localization of *copia* retrotransposon a) abundant PCR fragment (c. 300 bp; ladder left) used as a probe. b) Hybridisation of *copia* retrotransposon PCR product (pink) showing broad centromeric region an additional intercalary signal (white arrow) on the opposite arm of the chromosome with 5S rDNA (green). Bar: 5µm

3.3.2.2 *gypsy*

PCR with primers GyRT1 and GyRT4 as described by Kubis *et al.* (1998) generated the expected fragment of about 420 bp. *In situ* hybridisation of *gypsy*-like probes to metaphase chromosomes showed a dispersed localization on all 32 *E. guineensis* chromosomes (Figure 3.5).

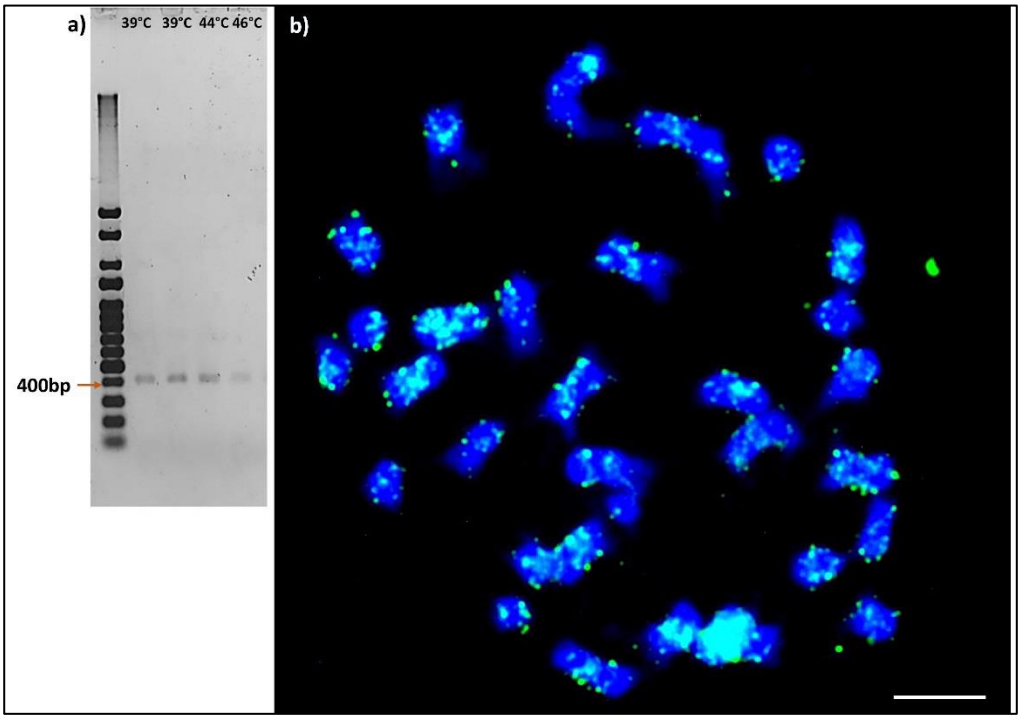


Figure 3.5 Chromosomal localization of *gypsy* retrotransposon a) PCR fragment (c. 400 bp at three annealing temperatures; ladder left) used as a probe. b) *Gypsy* retrotransposon (green) showed a dispersed hybridisation pattern on all *E. guineensis* chromosome. Bar: 5µm

3.3.2.3 LINE

PCR with the primers BEL-1MF and BEL-2MR (Kubis *et al*., 1998) generated two fragments (c. 500 bp and c. 600 bp) larger from the expected size (410 bp) (Figure 3.6a). PCR fragment close to 500 bp was purified, labelled and further used as a probe in the *in situ* hybridisation. Chromosomal localization of the with LINE probes showed a similar dispersed hybridisation signal on all 16 chromosome pairs as observed with *gypsy*-like probe (Figure 3.6b).

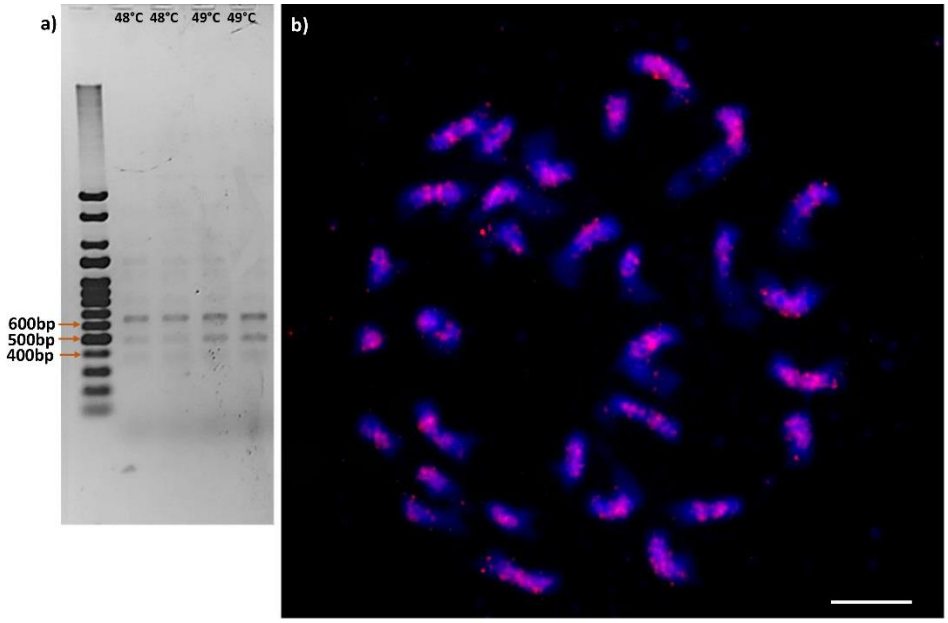


Figure 3.6 Chromosomal localization of *LINE*-like retrotransposon a) PCR fragment (c. 500 bp and variant c.650 bp, at two annealing temperatures; ladder left) used as a probe. b) *LINE*-like retrotransposon (pink) showed a dispersed hybridisation pattern on all *E. guineensis* chromosome with some clustering on the proximal regions of the chromosomes. Bar: 5µm

3.3.3 Chromosomal localization of 'unclassified' repetitive DNA on *E. guineensis* chromosome

The automated classification reported "Unclassified repeat" (or "low complexity") class with 13.75 % from the whole set data analysed with RepeatExplorer analysis (Figure 3.1b). Probes were designed from four randomly selected clusters and hybridized on the *E. guineensis* chromosomes. *In situ* hybridisation result showed the unspecific hybridisation pattern of the putative chromosome-specific probes on the *E. guineensis* chromosome (Figure 3.7). LC_Cl12 (b) showed broad hybridisation pattern on the proximal or painted whole arm of the 32 chromosomes. LC_Cl83 (a) and LC_Cl42 (d) with scarce but localized hybridisation pattern and LC_Cl61 (c) showed hybridisation signals on only a few chromosomes without a specific pattern that could distinguish the chromosome.

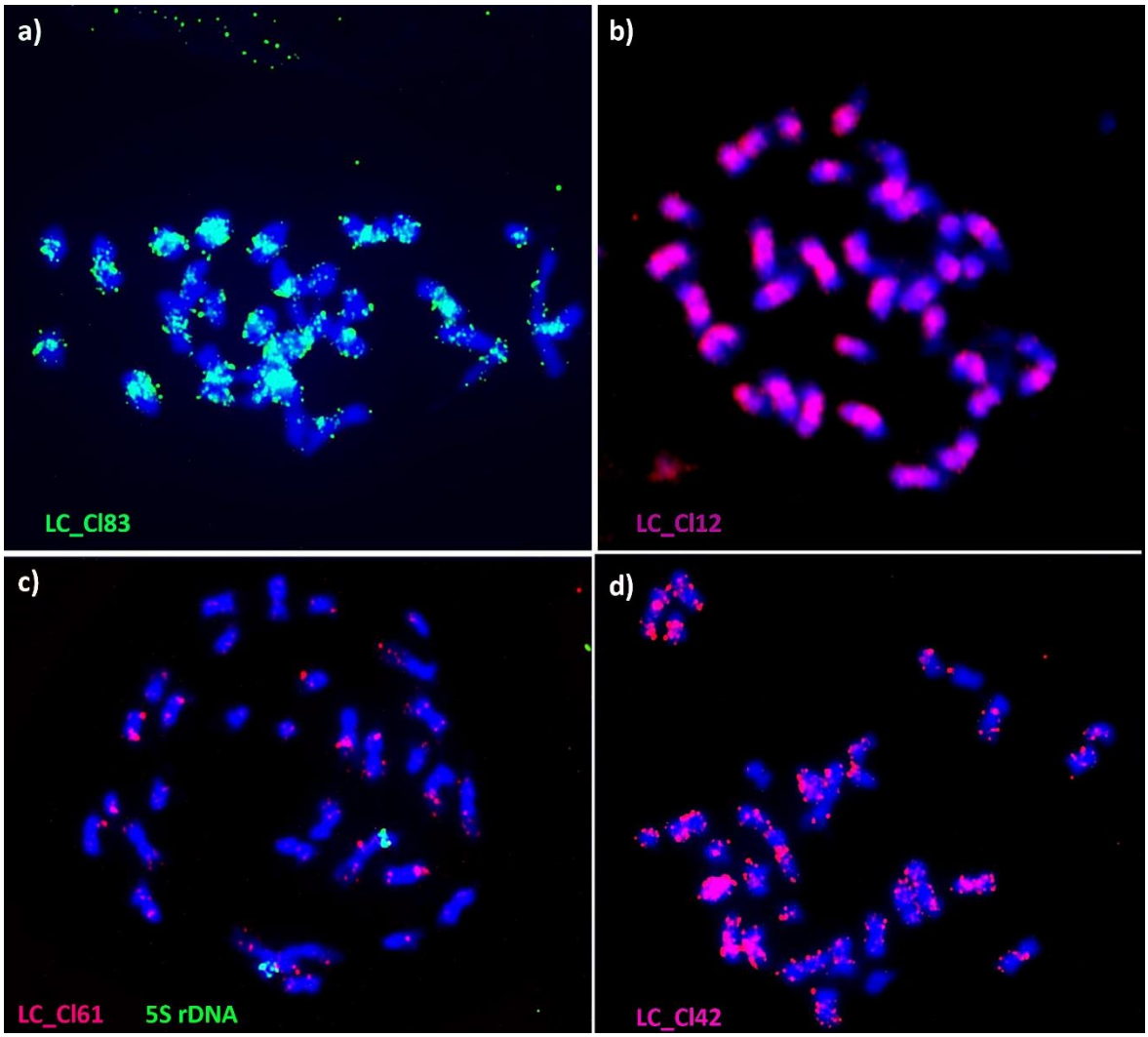


Figure 3.7 Chromosomal localization of probes derived from "unclassified/Low complexity" clusters generated from RepeatExplorer analysis. The selected, designed probes showed dispersed hybridisation signal across 16 pairs of oil palm chromosomes with a certain degree of hybridisation signals strength. The probes used in the *in situ* are indicated on individual FISH images respectively.

3.3.4 Chromosomal localization of unique repetitive DNA in oil palm genome

3.3.4.1 pEgKB family

Homology searches of the first top 100 contigs derived from *k*-mer analysis resulted in abundant hits to the repetitive DNA belonging to the pEgKB family. The pEgKB family is comprised of 14 clones (AJ271996-AJ272001 and AJ271979-AJ271981) with fragments of the reverse transcriptase domain of *copia*-like retrotransposon published by Castilho *et al.* (2000). Further sequence alignment of the pEgKB family resulted in three distinct clusters consisting of the pEgKB23, pEgKB17 and pEgKB19 families as depicted in Figure 3.8. Assembly of the three clusters to the 38.3 GB *E. guineensis* unassembled genome sequence revealed a genome proportion of 0.27 %, 0.37 % and 0.89 % for pEgKB9, pEgKB17, and pEgKB23 respectively.

Further similarity searches of the three sequences against *E. guineensis* assembled genome sequence (Eg5) were performed. A unique 307 bp fragment from the pEgKB9 sequence was found specific to chromosome 3 and chromosome 10 of *E. guineensis*. A designed primer pair of pEgKB9_319bp (section 3.2.4; Table 3.1) showed a single and clear amplified PCR fragment with the size of c. 300 bp (Figure 3.9a). However, chromosomal localization of the pEgKB9_319bp showed dispersed hybridisation signals on all the 32 *E. guineensis* chromosomes (Figure 3.9b).

No unique region was identified from both pEgKB17 and pEgKB23 sequence. Nevertheless, primers were designed from both sequences (Section 3.2.3; Table 3.1). The amplified PCR product was further hybridized to the oil palm chromosome to observe the chromosomal localisation of the probes as it has not been documented in Castilho *et al.* (2000). PCR of pEgKB17 derived primer pairs resulted in the amplification of two clear bands with the size of 400 bp and 600 bp and multiple faint bands with larger size (Figure 3.10). Hence no probes were used from pEgKB17. As for pEgKB23, a clear PCR amplified fragments with a size of c. 400 bp was further extracted, purified, labelled and hybridized on the *E. guineensis* chromosomes (Figure 3.11). Interestingly, the 400bp fragment of pEgKB23 painted all the 32 *E. guineensis* chromosome in the pericentromeric region with one of the largest chromosomes marked with a single intercalary hybridisation band.

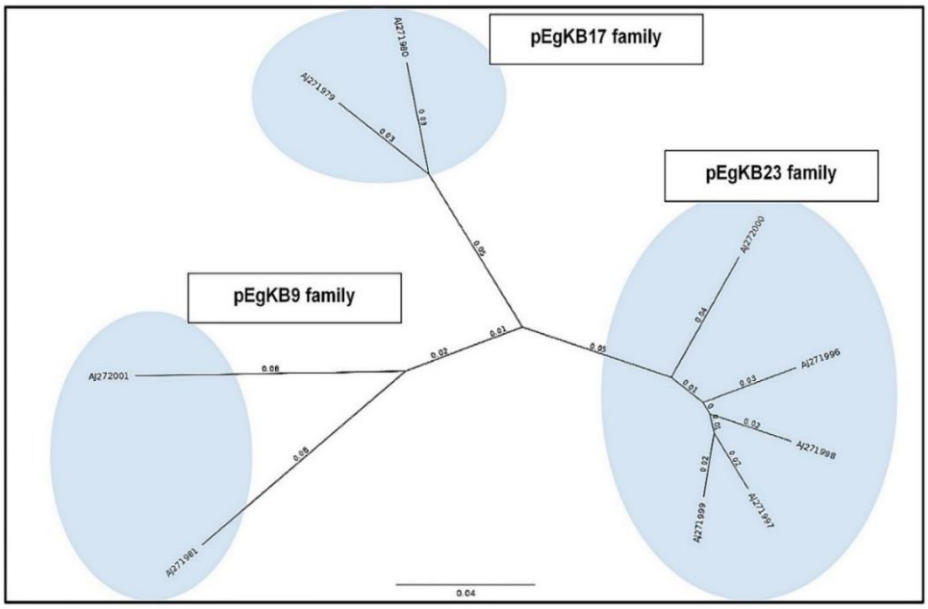


Figure 3.8 Unrooted UPGMA tree of pairwise relatedness of the pEgKB family. Nine pEgKB sequences clustered to three distinct groups; pEgKB9, pEgKB17, and pEgKB23. The unrooted UPGMA tree was generated with Tamura-Nei genetic distances within Geneious programme (Geneious Tree Builder). UPGMA, Unweighted Pair Group Method with Arithmetic Mean.

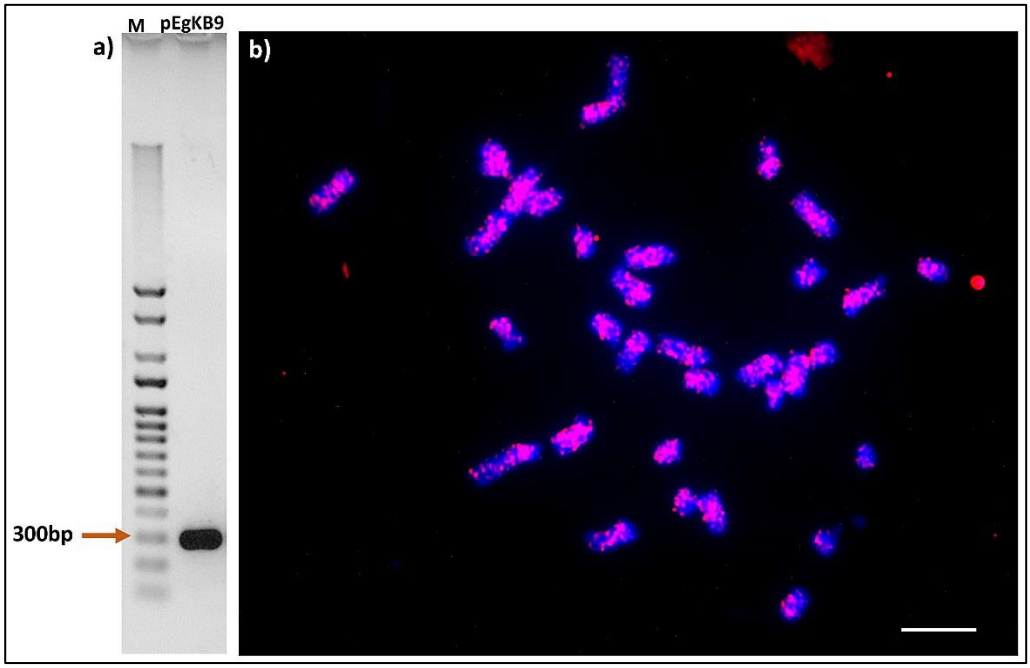


Figure 3.9 Characterization of the 309 bp specific-pEgKB9 region. a) PCR amplified of 309 bp length region of pEgKB9 specific probes. b) *In situ* hybridisation of the 309 bp fragment showed scattered localization across the 32 *E. guineensis* chromosomes. Bar: 5µm

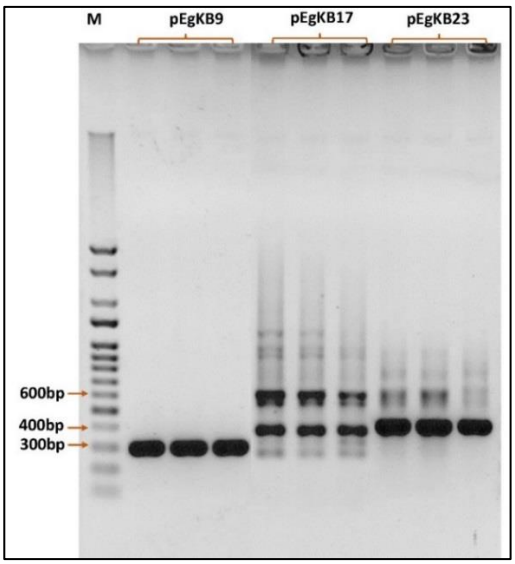


Figure 3.10 PCR amplification of pEgKB family. The single fragment was amplified for pEgKB9 and pEgKB23. PCR of pEgKB17 derived primer pairs resulted with the amplification of two bright bands with the size of 400bp and 600bp and multiple faint bands with the larger size.

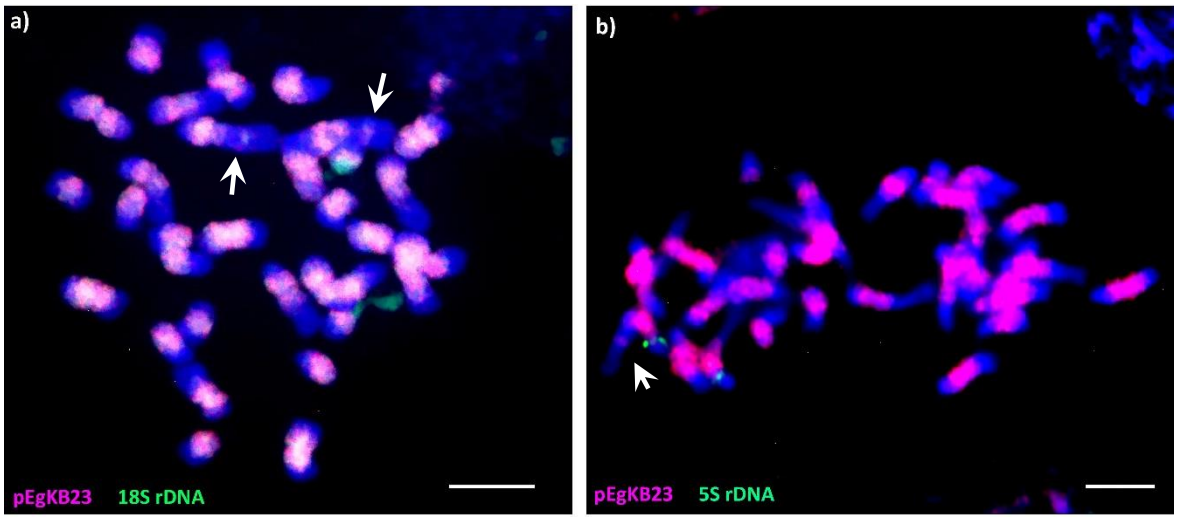


Figure 3.11 *In situ* hybridisation of pEgKB23 on the pericentromeric region of 32 *E. guineensis* chromosomes with one intercalary band on the long arm of the largest chromosome (white arrow). a) pEgKB23 displayed in pink and 18S rDNA displayed in green, arrow showed intercalary sites hybridized by pEgKB23 on one pair of the largest chromosome. b) pEgKB23 displayed in pink and 5S rDNA displayed in green, arrow showed intercalary site hybridized by pEgKB23 on opposite arm carrying 5S rDNA. Only one intercalary hybridisation sites visually observed Bar: 5µm

3.3.4.2 rDNA

Analysis of unassembled genome sequence by RepeatExplorer and *k*-mer analysis enabled the discovery of gene sequences related to ribosomal DNA repeat unit. From RepeatExplorer output, rDNA genes composed of 0.49 % (0.43 % of 45S rDNA and 0.06 % of 5S rDNA) of *E. guineensis* genome. Homology searches of the *k*-mer derived contigs were only able to identify 18 rDNA (a subunit of 45S rDNA).

Alignment of *E. guineensis* putative 45S rDNA sequences derived from RepeatExplorer (Cluster 63; Cl63 and Cluster 68; Cl68) and *k*-mer derived putative rDNA (*k*-mer32, Contig3 and Contig 5) with complete 45S rDNA from rice (KM036282) and partial 18S rDNA of *Elaeis oleifera* (AY012395) showed a conserved region of the 18S rDNA (Figure 3.12a). The designed primer pair flanking the conserved region amplified a 1.7 kb fragment that further use for *in situ* hybridisation (section 3.2.3; Table 3.1). Chromosomal localization of the PCR-amplified 18S rDNA region showed a similar pattern and strength as seen with wheat 45S rDNA (pTa71 clone) on *E. guineensis* chromosome (Castilho *et al*., 2000; Madon *et al.,* 2005) (Figure 3.12b). Remarkably, the similar hybridisation pattern and strength was also observed on the *Elaeis oleifera* (*E. oleifera*), another species of oil palm (Figure 13.2c). Moreover, chromosomal hybridisation of the oligo designed from the conserved region of 18S rDNA also showed consistent hybridisation region on one pair of small acrocentric *E. guineensis* chromosome (section 4.3.1; Figure 4.6)

A synthetic oligo (42mer) was designed from the conserved region of the aligned sequence of 5S rDNA derived from RepeatExplorer (Cluster 77) with 5S rDNA from other 30 species. However, in addition to the expected 5S rDNA hybridization signal on one arm of the largest chromosome, the designed oligo of the 5S rDNA showed dispersed signals on the telomeric region across the 32 chromosomes (Figure 3.13).

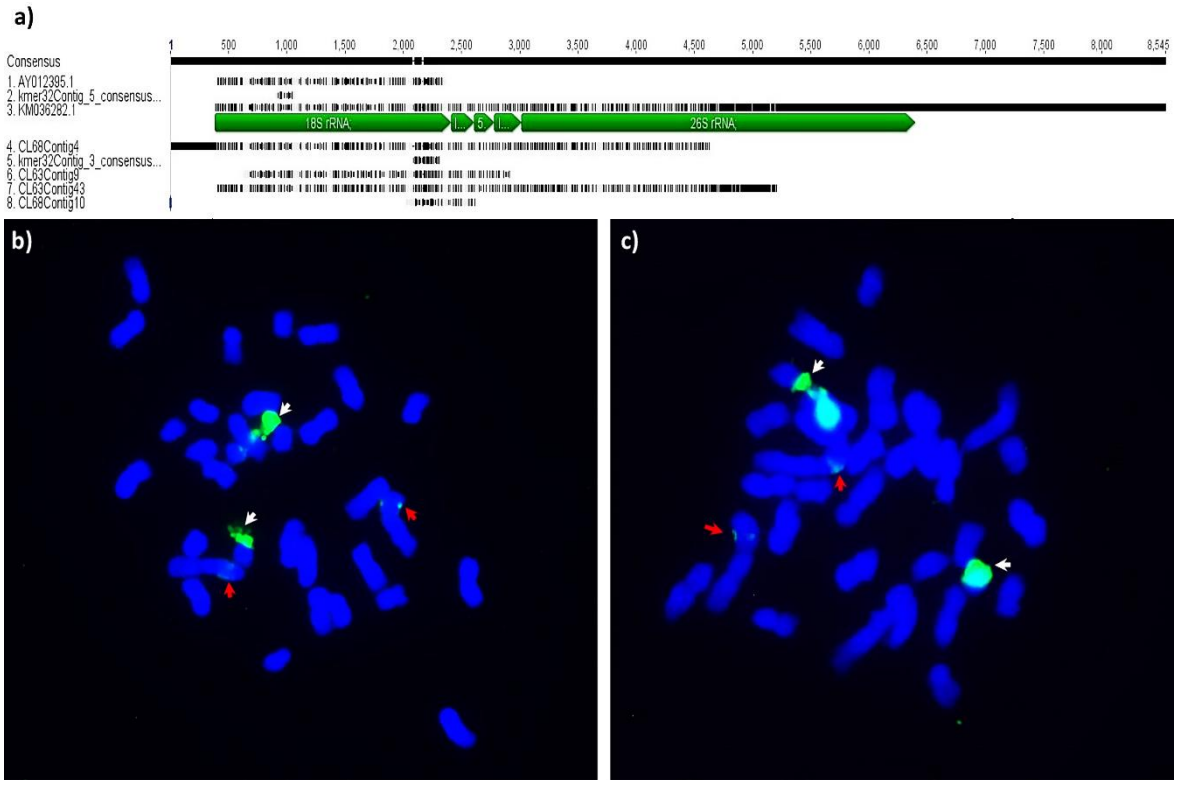


Figure 3.12 Sequence alignments and chromosomal localization of 18S rDNA derived from RepeatExplorer and *k*-mer analysis. a) Alignment of *E. guineensis* putative 45S rDNA sequences derived from RepeatExplorer (Cl63 and Cl68) and *k*-mer derived putative rDNA (*k*-mer32, Contig 3 and Contig 5) with complete 45S rDNA from rice (KM036282) and partial 18S rDNA of *Elaeis oleifera* (AY012395) showed a conserved region of the 18S rDNA. Figure (b) and (c) showed *in situ* hybridisation of the PCR amplified 18s rDNA on *E. guineensis* (b) and *E. oleifera* (c).Two broad sites of strong hybridisation signal (green signal; white arrow) on one pair of the small chromosome of *E. guineensis* and *E. oleifera* respectively. Red arrow showed 5S rDNA localization on one pair of largest chromosomes for both *Elaeis* species with a probe derived from clone pTa71 (Gerlach and Dyer 1980).

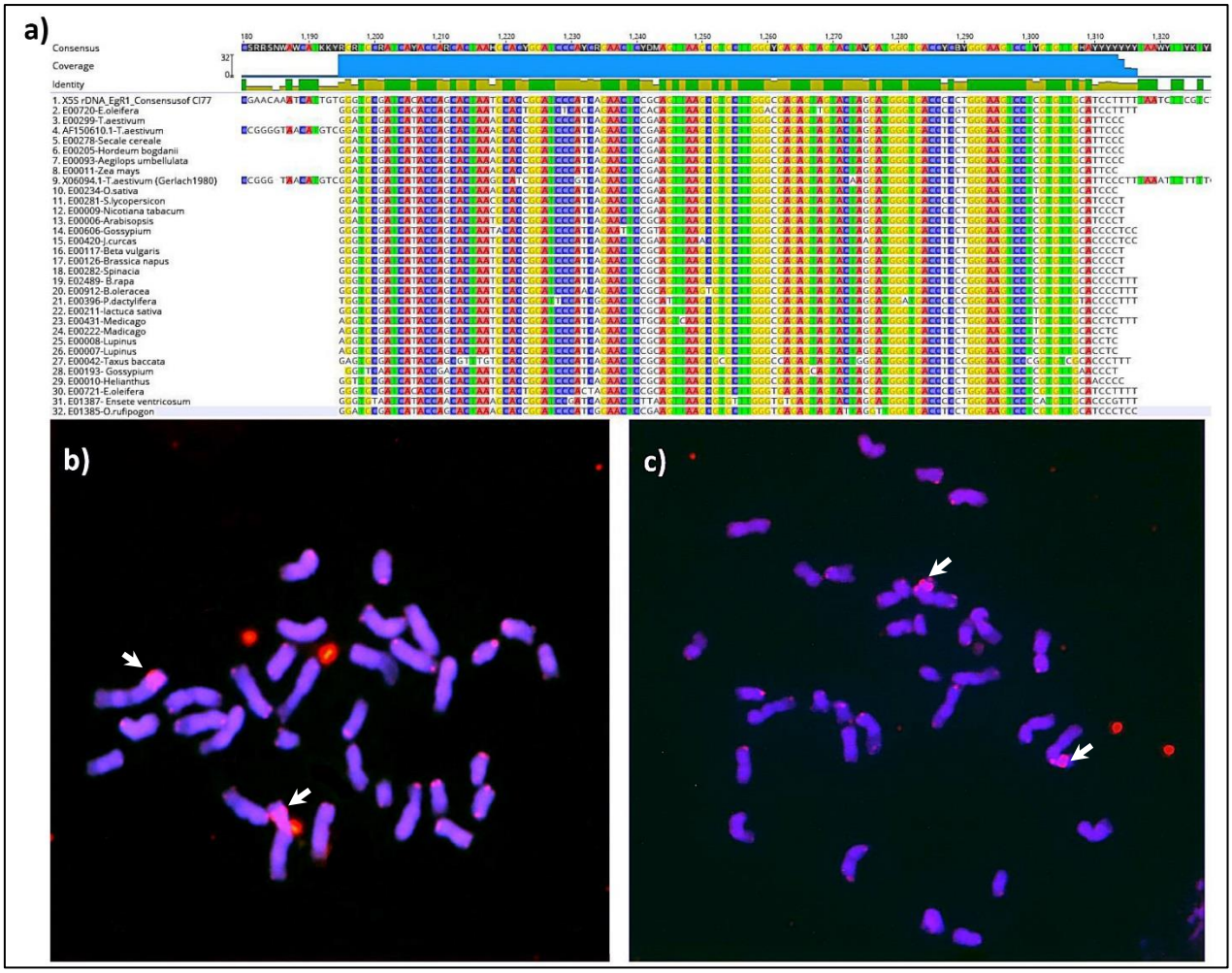


Figure 3.13 Localization of the 5S rDNA synthetic on oil palm chromosome. a) A synthetic oligo (42mer) was designed from the conserved region of the aligned sequence of 5S rDNA derived from RepeatExplorer (Cl77) with 5S rDNA from other 30 species. b and c) *In situ* hybridisation of the 5S oligo (red) showed dispersed signals on the telomeric region across the 32 chromosomes in addition to a clear signal of 5S rDNA on one arm of the largest chromosome (white arrow).

3.3.4.3 Telomere

Seven canonical sequence of telomere (TTTAGGG) were obtained from 64-mer contig (k-mer analysis) with genome proportion of 0.02 % in the *E. guineensis* genome (Figure 3.14a). *In situ* hybridisation of the synthetic oligonucleotide telomere $(CCCTAAA)_6$ showed a strong paired signal on the telomeric sites at both end of all 32 chromosomes with no intercalary sites (Figure 3.14b).

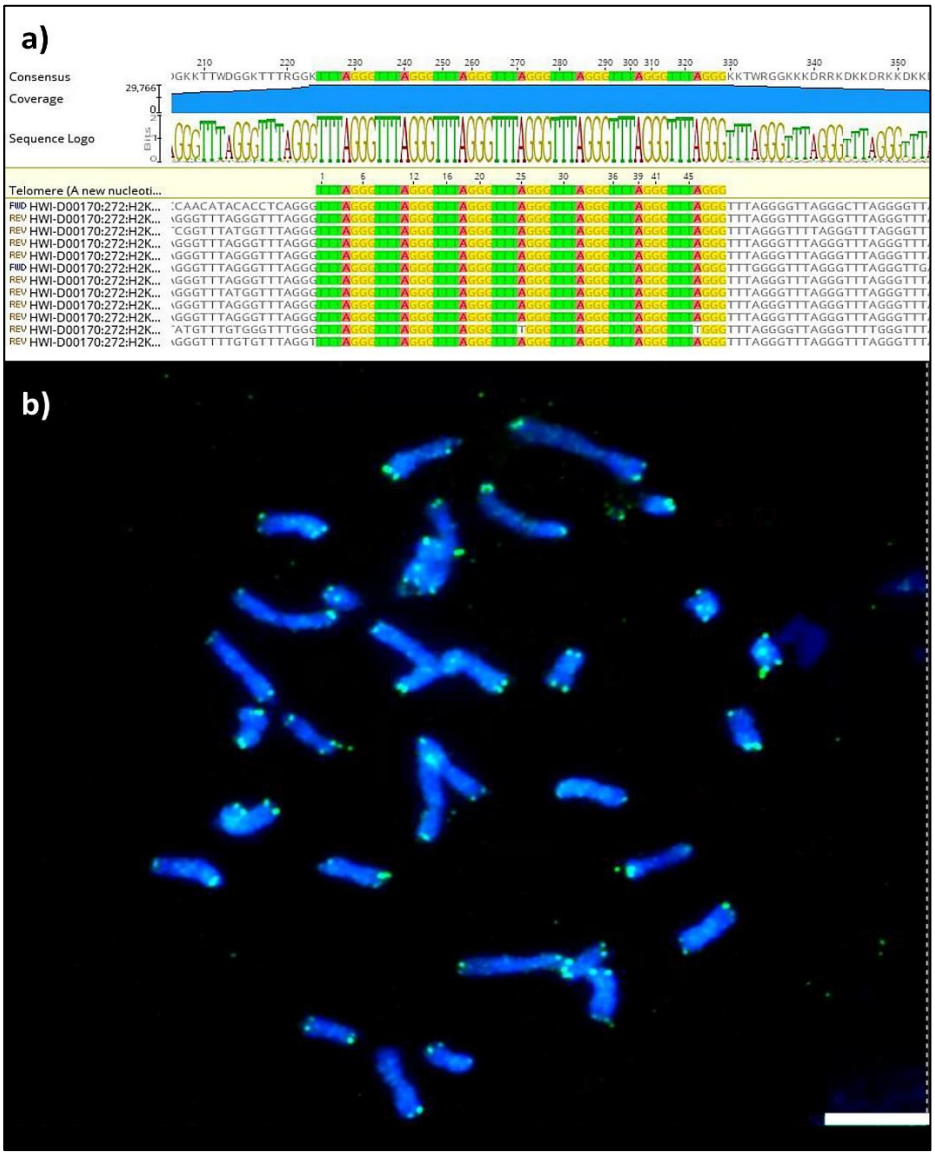


Figure 3.14 *E. guineensis* telomere. a) Characterization of telomere sequence (TTTAGGG) in *E. guineensis* unassembled sequence and b) Characterization of telomere sequence on *E. guineensis* physical chromosome (green paired dots on the terminal region of the 32 chromosomes) Bar: 5µm

3.3.4.4 Putative centromeric sequence

A putative centromeric region (named Eg9CEN) of 354 bp identified from *E. guineensis* Eg9 sequence assembly (unpublished; kindly provided by MPOB and collaborators based on an updated version of the assembly in Singh *et al*., 2013) with 1% genome proportion in the *E. guineensis* genome. The amplification within the fragment by PCR showed a single band amplicon with the expected size of c.330 bp (Figure 3.15b). *In situ* hybridisation with Eg9CEN showed a broad centromeric hybridisation signal observed on all 32 chromosomes, notably with a conserved and distinct intercalary site of hybridisation detected on the opposite arm to the 5S rDNA site (Figure 3.15b; 3.15c).

DAPI staining of *E. guineensis* chromosomes showed the localization of the intercalary hybridized site of Eg9CEN next to a tertiary constriction site (Figure 3.15b; 3.15c). The localization of the pericentromeric hybridisation site of the retrotransposon on the physical chromosome is in agreement with the *E. guineensis* genome assembly (Figure 3.15d-circos plot; Singh *et al*. 2013). Interestingly, for chromosome 2, apart from a similar dense region of retroelement density on both pseudo- and physical-chromosome (track II; image III), no intercalary site of retrotransposons was observed on the pseudo-chromosome (track II). Intercalary telomeric positioning on the assembly (track I) showed different orientation compared to the tertiary constriction identified on the *E. guineensis* physical chromosome.

Remarkably, the hybridisation pattern generated by Eg9CEN showed a resemblance to the hybridisation pattern of pEgKB23, a *copia*-like retrotransposon identified from the *k-mer* derived contigs (Figure 3.11). Sequence homology searches against NCBI database showed 95% to 96% similarity to five of the pEgKB family; pEgKB1, pEgKB23, pEgKB19, pEgKB14 and pEgKB14 (Figure 3.16). Interestingly, the stretch of the c. 300 bp of Eg9CEN showed 81 % similarity with *Phoenix dactylifera* clone dpB3Y sex-determination region sequence, one of the repeat-rich male-specific BAC clones published recently by Torres *et al*. (2018).

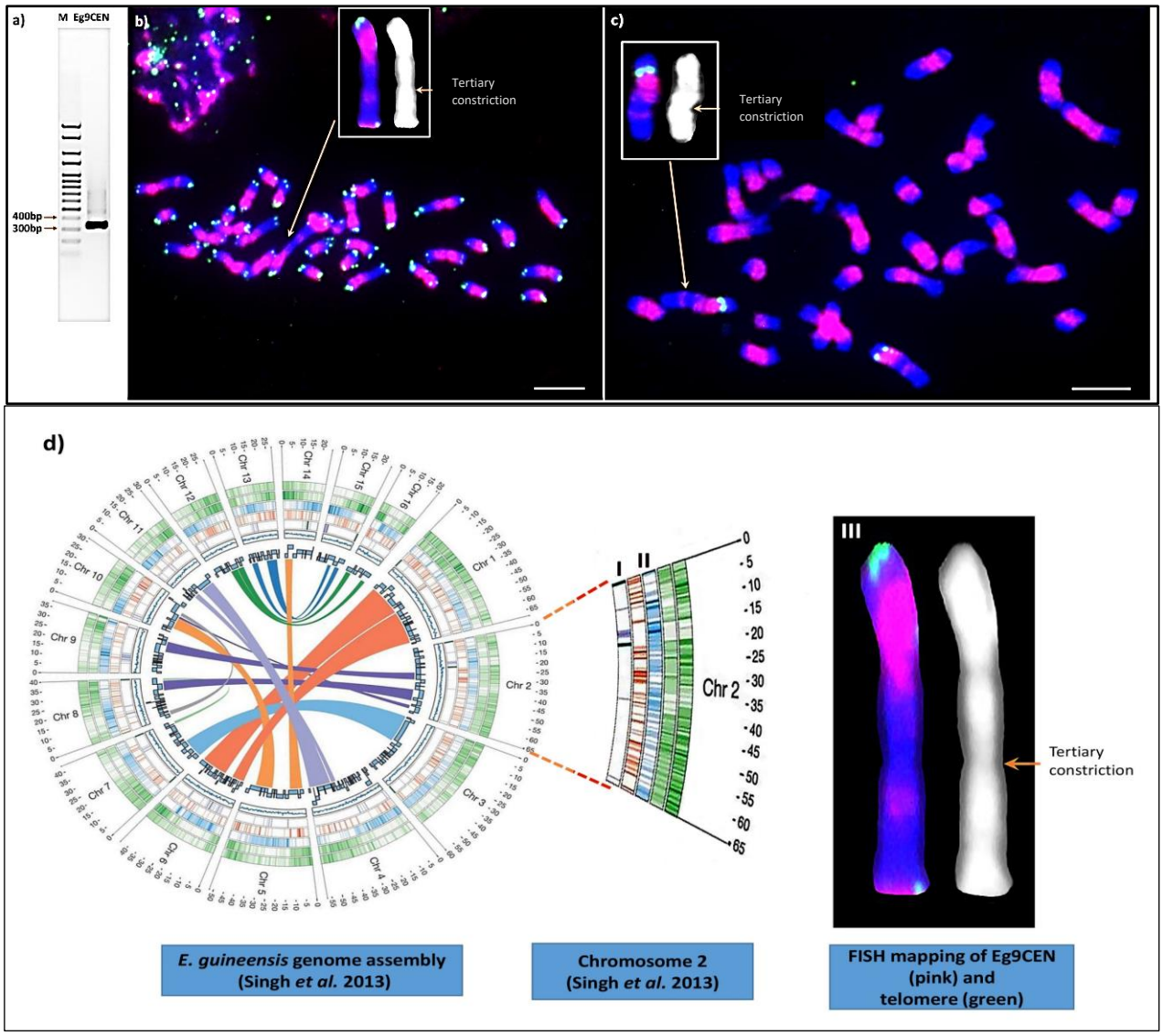


Figure 3.15 Characterization of putative centromeric sequence-*copia* like (Eg9CEN) identified from updated unpublished *E. guineensis* assembly (Eg9). a) PCR amplification of Eg9CEN fragment (c.300bp) from genomic DNA of *E. guineensis*. b) *In situ* hybridisation of Eg9CEN (displayed in pink) and telomere (displayed in green). Boxed chromosome showed the intercalary hybridisation site near to newly identified tertiary constriction site of the largest chromosome. c) *In situ* hybridisation of Eg9CEN (displayed in pink) and 5S rDNA (displayed in green). Boxed chromosome showed the intercalary hybridisation site near to newly identified tertiary restriction site of the largest chromosome opposite to the arm carrying 5S rDNA.

d) Localization of the retroelement density and telomere (Singh *et al*. 2013; track II and track I respectively) compared to physical location of the tertiary constriction and intercalary hybridized region (image III) of the Eg9CEN *copia*-like probes on the Chromosome 2. (Pink mark on the terminal region due to the minor overlapped with the other chromosome). Bar: 5µm

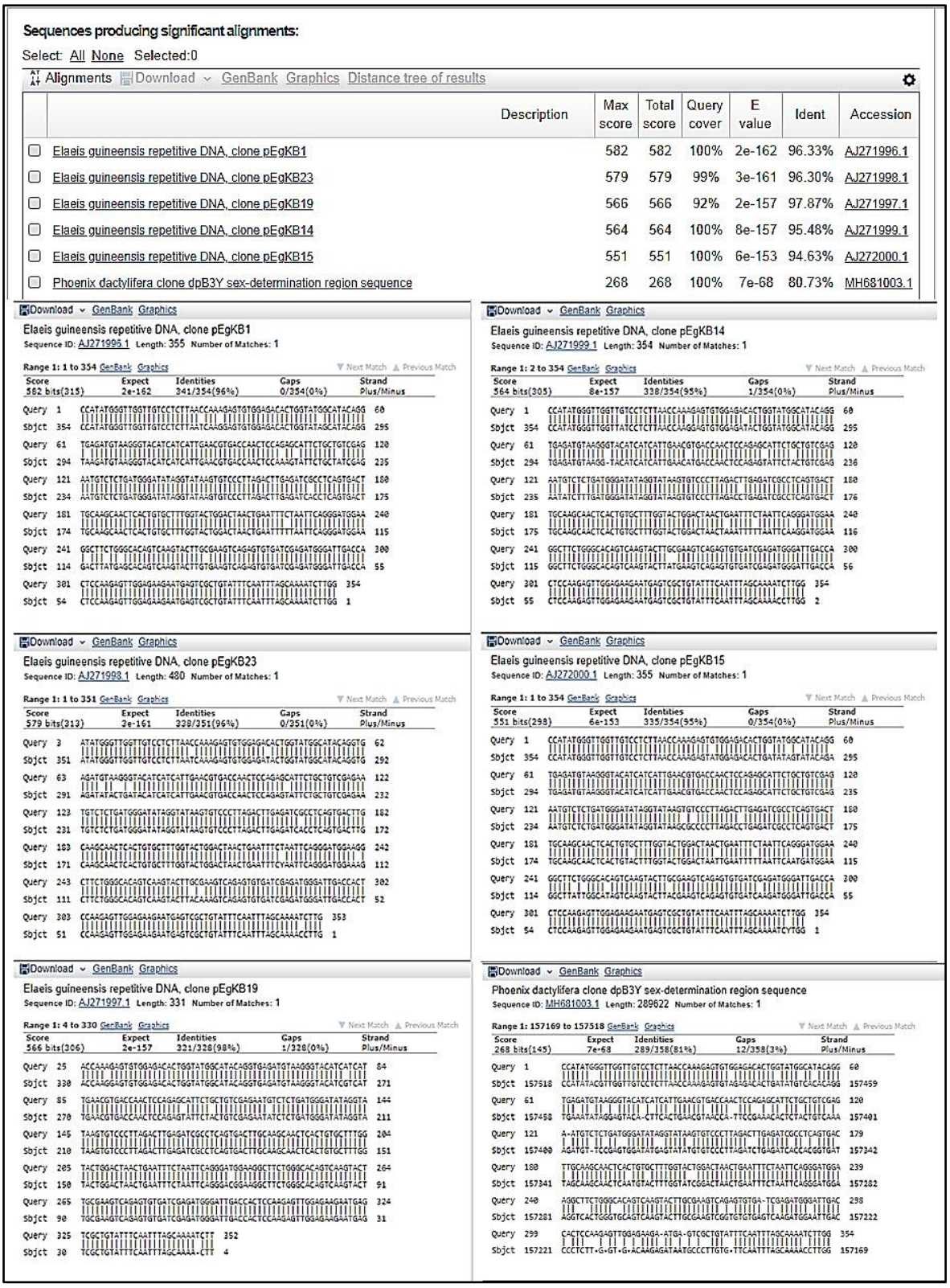


Figure 3.16 Similarity search of the Eg9CEN sequence against the NCBI database. BLAST result of the 354bp putative centromere sequence from *E. guineensis* Eg9 assembly showed a high similarity (95-96%) to pEgKB family and clone dpB3Y sex determination regions sequence from *Phoenix dactylifera* (81 %).

3.4 Discussion

3.4.1 Repetitive DNA landscape of the *E. guineensis* genome from unassembled genome sequence and chromosomal *in situ* hybridisation

In the original paper describing oil palm genome, the repeat content of *E. guineensis* had been estimated to be approximately 57 % of its 1.8 Gb genome with a substantial occurrence of LTR retrotransposon (Singh *et al*., 2013). Here, the major repeat families identified by both *de novo* approaches (RepeatExplorer and *k*-mer analysis) showed no newly identified major repetitive DNA class compared to known repeats in *E. guineensis* from random cloning, PCR, restriction digestions, and cytogenetic analysis (Castilho *et al.,* 2000). Thus the high-volume sequence analysis (Singh *et al*., 2013; Beule *et al*., 2015; Filho *et al*., 2017) and specific analysis for repetitive DNA carried out here with unassembled raw sequence reads (Figure 3.1 and 3.2) showed that the older studies (Castilho *et al.,* 2000; Kubis *et al.,* 2003; Price *et al*., 2003) had identified all major repeats. Furthermore, there were no major repetitive DNA families that were abundant at a few chromosomal sites or with chromosome-specific distributions (Figure 3.4-Figure 3.7; Figure 3.9; Figure 3.11). Domination of transposable elements in the unassembled *E. guineensis* genome (40.66 %) identified by RepeatExplorer was slightly higher compared to the identified from assembled genome which ranges from 39.41 % (Filho *et al*., 2017) to 39.5 % (Beule *et al*., 2015) presumably because some reads were collapsed during assembly. The finding showed that the underestimation of repeat analysis using available genome contig sequences compared to using unassembled sequence reads is not an issue in determining the repetitive genome proportion from the whole genome.

Chromosomal localization of the most abundant transposable element (LTR-*copia* and LTR-*gypsy*) with universal primers showed a differential localization of both LTR super-families (Figure 3.4 and Figure 3.5). *Copia* hybridisation was more concentrated on the broad region of proximal chromosome while *gypsy* exhibited a random and dispersed hybridisation pattern all over the 32 *E. guineensis* chromosomes. Contrary to the present finding, in some cereal, such as barley, wheat, and rice, the substantial occurrence of LTR-g*ypsy* elements was commonly observed on centromeric and pericentromeric locations (Cheng and Murata, 2003; Nagaki *et al.,* 2005; Divashuk *et al*., 2016). Moreover, the two well

investigated centromere-specific retrotransposons in *Poaceae*, the CRR from rice (Cheng *et al.,* 2002) and CRM from maize (Ananiev *et al*., 1998) were Ty3-*gypsy* type transposable elements. The physical distribution of the transposable element supported by the identification of a preferential insertion of full length *copia* elements in relatively gene-poor regions compared to randomly distributed *gypsy* element of the *E. guineensis* assembled pseudo-chromosomes (Beule *et al*., 2015), agreeing with the *in situ* hybridisation results here (Figure 3.4 - Figure 3.6) and shown by Castilho *et al*. (2000). Similar differential distribution of both super-families was also reported in gymnosperm (Friesen *et al*., 2001) and in a number of the higher plant (Heslop-Harrison *et al*., 1997; Brandes *et al*., 2001; Karlov *et al*., 2010). Comparisons performed between partially or entirely sequenced plant genomes have shown that LTR elements are mostly concentrated in gene-poor regions, with variation according to superfamily or lineage (Brandes *et al*., 1997; Vitte *et al*., 2005; Bennetzen and Wang, 2014).

The raw-read analyses revealed some sequences which could not be easily classified by comparison with repetitive DNA databases. Four of these from RepeatExplorer clustering were selected for *in situ* hybridisation (Figure 3.7) to see whether any hybridisation patterns were different from known repeat families. The results showed that the range of patterns was similar to those of other repeats, with dispersed location or broad centromeric hybridisation.

Based on the *in situ* hybridisation results (Figure 3.4, 3.11 and 3.15b-c), it can be concluded that putative centromere Eg9CEN is a *copia*-like sequence having similarity with pEgKB23. Notably, the Eg9CEN, a consensus extracted from the *in silico* analysis rather than being a single cloned sequences, showed a more uniform hybridisation pattern than the 95 % similar pEgKB23 clone published by Castilho *et al*. (2000). Previously, the intercalary site on the largest chromosomes carrying 5S rDNA was not distinguished when the clones were used in the *in situ* hybridisation, despite the similarity. Given the similarity (95 %) between the sequences, even with highly stringent hybridisation and wash conditions, *in situ* (or

indeed Southern) hybridisation is unlikely to distinguish robustly between the probes, so it is interesting that the cloned probes gave a less specific signal.

Using the strategy implemented here, aside from the structural component of repetitive DNA (telomere and rDNA) and abundance of transposable element superfamily (including copia-like Eg9CEN), no newly identified repetitive DNA could distinguish individual *E. guineensis* chromosome. After compilation of the results obtained in this chapter, a new karyotypic data comprising morphological markers as well as repetitive DNA is proposed (Figure 3.16). Furthermore, the obtained repetitive DNA library of *E. guineensis* derived from the raw read will be useful in screening the whole genome for obtaining chromosome specific marker in Chapter IV. By merging reads from clusters containing the same repeat types/families, resulting in repeat-specific library that can be used as reference in various types of sequence similarity searches. The advantage of using such library instead of a few selected consensus sequences is that they capture the full range of the repeat sequence variation and thus provide better sensitivity in the detection of less conserved repeat variants.

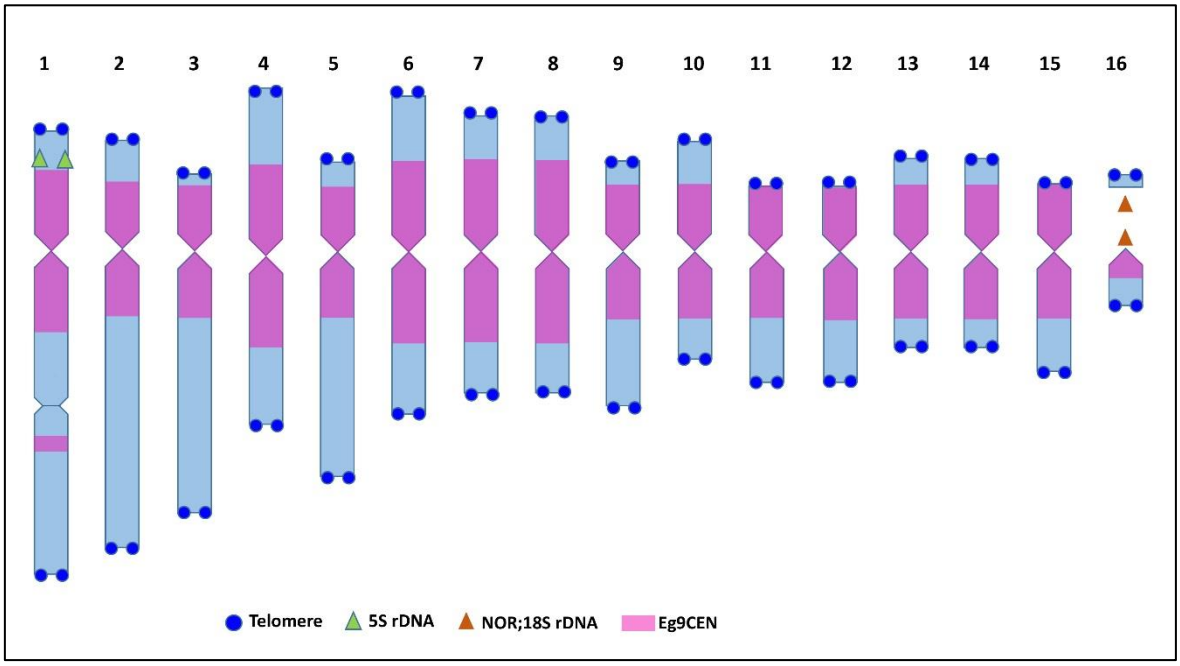


Figure 3.16 Proposed oil palm karyotype with physical localization and relative distribution of repetitive DNA. Telomere (blue circle), 5S rDNA (green triangle), 18S rDNA (brown triangle) and Eg9CEN-*copia* like (pink). Tertiary constriction of the largest chromosome also shown. The chromosomes in the karyotype were arranged by order of decreasing size. Centromeric constrictions are drawn as a cross; secondary constriction (Chromosome 16) at the NOR as a gap; tertiary constriction (Chromosome 1) as a constriction.

3.4.2 Re-visiting the tertiary constriction structure of *E. guineensis* chromosomes proof of Robertsonian translocation?

Chromosomes show a primary constriction morphologically at the centromere, and a secondary constriction is recognized in many species at sides of 45S rDNA (Nucleolar Organizer Region, NOR). Here the identification of physical tertiary constriction on the q-arm of one pair of *E. guineensis*'s largest chromosomes supported the Robertsonian fusion that has been speculated for *E. guineensis* chromosome 2 through genome sequencing (Singh *et al*., 2013). Interestingly, a minor hybridisation signal observed near to the tertiary constriction site showed by a 354bp putative centromeric (*copia*-like retrotransposon) marker. Searches of the NCBI database indicated 81 % similarity of the sequence with *Phoenix dactylifera* clone dpB3Y sex determination region sequence containing a cytidine deaminase like gene (Torres *et al*., 2018). Cytidine deaminase was the only genus-wide male-specific gene in *Phoenix* in which both X and Y chromosome copies were identified. Phylogenetic analysis of this gene indicates that the male and female alleles form separate clusters, suggesting that their sequences diverged before the speciation events (Torres *et al*., 2018). The fact that the cytidine deaminase contains sequences found in all of investigated *Phoenix* males suggests that the rearrangement (fusion) of the region concerning the ancestral oil palm may have played a role in the formation of the Y chromosome though this remains to be investigated.

Since the first identification of the 32 *E. guineensis* chromosomes with two dicentric chromosomes (Sharma and Sarkar, 1956) (Figure 3.17) there was no further documentation or investigation of the *E. guineensis* tertiary constriction reported. Dicentric chromosomes are products of genomic rearrangements that place two centromeres on the same chromosome. Due to the presence of two primary constrictions, they are inherently unstable and overcome their instability by epigenetically inactivating and deleting one of the two centromeres, thus resulting in functionally monocentric chromosomes that segregate normally during cell division (Chiantante *et al*., 2017). In a mammal, a Robertsonian evolutionary fusion led to the formation of human chromosome 2, explaining the only chromosome number difference between humans (46 chromosomes) and great apes (48 chromosomes) Lejeune *et al*., 1973; Yunis and Prakash, 1982; Ijdo *et al*., 1991. The Robertsonian event is also critical in the bovids where all with

60 autosomal chromosome arms but fusing to include metacentric autosomes (Escudeiro *et al*., 2019; Chaves *et al*., 2003). Robertsonian fusion has been demonstrated as one of the important mechanisms of karyotype evolution in several genera of the monocotyledonous flowering plant belongs to *Tradescantieae* (*Commelinaceae* family) where it can be regarded as a type of chromosome ortho-selection (Jones *et al*., 1998). Other species groups where chromosomes are rearranging include the *Brassicaceae* (with n=9, n=10, and n=8) and the methods using oligonucleotide probes (Chapter IV) are being used to examine the nature of fusions and translocations in the *Brassicaceae* (unpublished from MolCyt lab).

As for *E. guineensis* (oil palm), the synteny of the 18 date palm linkage group with one of the 16 oil palm chromosomes was reported by Mathew *et al*. (2014). How the two genomes diverged from 16 to 18 or vice versa is of interest as synteny determined between the date palm genetic map and the oil palm chromosomes suggests that oil palm chromosome 2 constitutes a fusion of date palm chromosome 1 and 10. Re-visiting the existing of tertiary constriction of the *E. guineensis* largest chromosome (identified as Chromosome 2 in Eg9) was interesting as it is open new question marks, whether there is a plausible explanation for the karyotype evolution concerning the sex-determining region in both species of monocot that have different natures of sex determination. Repetitive DNA sequences are frequently found to be different between autosomes and sex chromosomes both in animals (Mustafa, 2018; Chaves *et al*., 2005) and plants (Navajaz-Perez *et al*., 2006). Furthermore, it is interesting to investigate the connection of the identified tertiary constriction with retrotransposon in oil palm speciation event with regards to the evolution of monoecious and dioecious in Arecaceae.

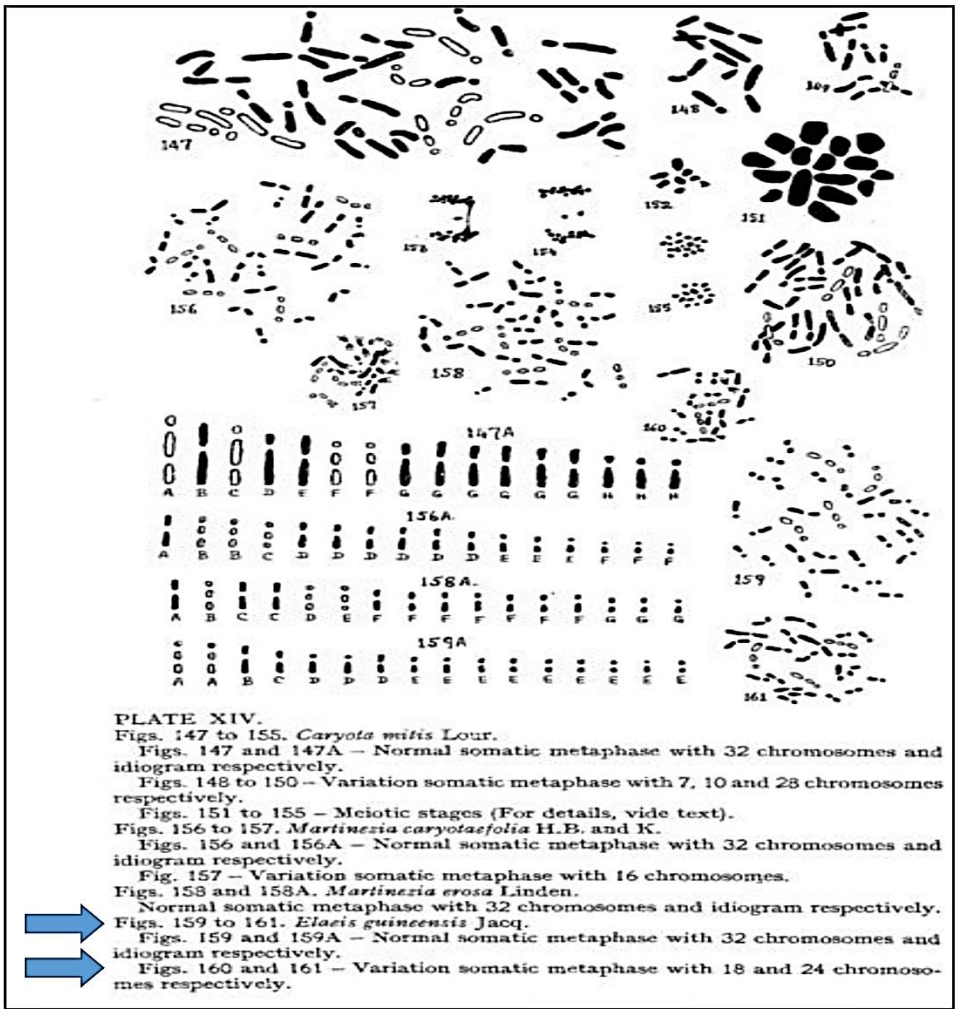


Figure 3.17 Identification of two dicentric *E. guineensis* chromosomes. Arrow showing the legend of the 16 haploid *E. guineensis* chromosomes karyotype with two dicentric chromosomes in displayed figure 159-161 (Sharma and Sarkar 1956)

CHAPTER IV

An *Elaeis guineensis* reference karyotype using unique-single copy massive oligonucleotide pools as chromosome-specific markers

---

4.1 Introduction

The 1.8 gigabase (Gb) oil palm (*E. guineensis* Jacq.) genome sequence was published in 2013 (Singh *et al.*) with approximately 43 % of genome successfully assembled. Comparison of the genome build to genetic linkage maps resulted in 16 genetic scaffolds representing 16 pseudo-chromosomes of oil palm. The unraveling of the world's most important oil yielding crop genome serves as the foundation of numerous studies. However, even when the sequences of the entire genome are available, this massive amount of genetic and physical sequence data cannot be directly associated with chromosome structure. Furthermore, assembly algorithms may not join contigs into scaffolds correctly, giving inversions or translocations for example, in regions where genetic marker density is low and repetitive DNA is abundant (Findley *et al*., 2010; Burton *et al*., 2013; Shearer *et al*., 2014).

A genetic linkage map is a good indicator of the linear marker order and the amount of recombination between linked markers. Nevertheless, it is essential to link the exact physical position of DNA sequences for whole genomes to individual chromosomes. Crossover giving rise to the meiotic recombination is not uniformly distributed over chromosome arms. As a result, loci that are physically far apart on the chromosomes can be tightly linked on linkage maps and vice versa (Heslop-Harrison, 1991; Wang *et al*., 2006; Sun *et al*., 2013; Shearer *et al*. 2014). Cytogenetic maps or physical maps show the positions of genetically mapped markers on individual chromosomes, relative to cytological

landmarks such as centromeres, telomeres, heterochromatin, and nucleolar organizer regions (NOR). With the progression of genome research, cytogenetic maps will not only be valuable for integrating and organizing genetic, molecular, and cytological information, but it will also provide a unique insight into genome organization in the context of the chromosomes (Heslop-Harrison and Schwarzacher, 2011). Furthermore, cytogenetic maps are essential to define translocations (and less frequently, inversions), and also to study introgression by exploiting genetic diversity from related species.

As reviewed in Chapter I, direct localization of DNA sequences on physical chromosomes by fluorescent *in situ* hybridisation (FISH) is commonly used to construct a cytogenetic map in the plant. Various approaches or combination of strategy have been developed to assign linkage groups to physical chromosomes based on the FISH technique. Nevertheless, a successful and efficient FISH experiment primarily relies on two major parameters; the high quality of chromosome preparation and robust DNA probes (reviewed in Jiang and Gill, 2006; Figueroa and Bass, 2010; Jiang, 2019). Development of chromosome markers to distinguish individual chromosomes in plant species is often done with repetitive DNA probes. Because of their high repetition in the genome, repetitive DNA provides useful markers to identify chromosomes by *in situ* hybridisation either being used singly or a cocktail probe (Leitch and Heslop-Harrison, 1992; 1993, Lengerova *et al*., 2005; Jiang and Gill, 2006; reviewed in Biscotti *et al*., 2015).

In *E. guineensis*, the high-volume sequence analysis (Singh *et al*., 2013; Beule *et al*., 2015; Filho *et al*., 2017) and specific analysis for repetitive DNA with unassembled raw sequence reads (discussed in Chapter III) showed that the older studies (Castilho *et al.,* 2000; Kubis *et al.,* 2003; Price *et al.,* 2003) had identified all major repeats. Furthermore, there were no major repetitive DNA families that were abundant at a few chromosomal sites or with chromosome-specific distributions except for Eg9CEN that giving additional intercalary hybridisation marker that is close to identified tertiary constriction (Chapter III). Hence, another effective strategy to identify *E. guineensis* individual chromosome is by anchoring the physical chromosome with a low or single copy sequence directly localized on chromosomes by FISH.

Physical mapping of low/single-copy genes remains a problem in lots of plant species and genera. The detection of small unique sequences on plant chromosomes has been challenging because the debris of cell wall and cytoplasm reduces the accessibility of target DNA and increases background and consequently results in a relatively low signal-to-noise ratio (Lehfer *et al*., 1993; Jiang and Gill, 2006). Yet, successful detection of single-copy genes with a size range of 2-10 kb has been reported on mitotic chromosomes of *Petunia hybrida* (Fransz *et al*., 1996), rice (Ohmido *et al*., 1998), maize (Lamb *et al*., 2007; Danilova *et al.,* 2008), wheat (Danilova *et al*., 2012; Perez *et al*., 2009) as well as *Rosa wichurana* (Kirov *et al*., 2014). However, the use of short single-copy sequences as chromosomal probes has not become a routine method and has not been successful in many species.

With the plummeting sequencing cost, the gene assignment on the physical chromosomes is often done with the assembly of whole genome sequences. By exploiting the genome assembly and advanced technique in DNA synthesis, a more efficient approach for direct visualization of genetically mapped markers on chromosomes using fluorescent *in situ* hybridisation (FISH) is possible. Most recently, a massively synthesized oligonucleotide demonstrates the ability marking the portions or entire chromosome *via* FISH approach. The work has been documented in mammals (Beliveau *et al*., 2012), *Drosophila (*Yamada *et al*., 2011*)*, Cucurbitaceae (Han *et al*., 2015, Li *et al*., 2016), Solanaceae (Braz *et al.*, 2018) and maize (Albert *et al*. 2019). As the unique sequence can be customized chosen as a target, this approach provides a straightforward means for integrating the genetic maps and physical maps on the plant.

The current study aimed to develop chromosome-specific cytogenetic markers from single and low copy sequence in the *E. guineensis* genome. This chapter discusses the approaches that have been deployed towards achieving the dedicated aim, as well as describes the development of massively synthesized pre-labeled oligo pools toward establishing *E. guineensis* FISH-based physical map that serves as a reference karyotype for the species.

4.2 Materials and methods

4.2.1 Plant materials

The oil palm (*E. guineensis*) materials used for developing the DNA probes were published by Singh *et al.* (2013) and are currently maintained at the MPOB Research Station, Kluang, Johor, Malaysia. Meristematic root tips were collected from three Pisifera palms (0.182/77, 0.182/30 and 0.182/7). Genomic DNA (0.182/77) was extracted and purified from a spear leaf using the modified CTAB method (Doyle and Doyle, 1990; section 2.2.1) and was used to amplify the 18S rDNA regions.

4.2.2 Development of chromosome-specific cytogenetic markers

The genome assemblies or builds of oil palm are designated by Eg followed by a serial number. Build Eg5 was published (Singh *et al*., 2013) and used at the start of work here. Later, Eg9 was available for the analysis. Genome assemblies with lower numbers are generally robust in low-copy-rich regions but may include areas with low read coverage. Assembly errors, involving both missed-joins (translocations) between single copy regions, and the collapse of sequences, may occur, particularly at the ends of repetitive DNA regions, but are expected to be better resolved in later builds.

4.2.2.1 Optimization towards developing chromosome-specific markers from single and low copy regions of *E. guineensis*

All the optimization steps were performed using sequences of pseudo-chromosome 1 derived from Eg5 assembly (Singh *et al*., 2013).

i) Development of chromosome-specific markers from unique, long and low copy region DNA sequences

The unique and low copy regions DNA sequences with length 10 kb and 5 kb were obtained by Orion Biosciences (collaborator) using designated proprietary bioinformatics protocols. The primers were designed using Primer3 (Table 4.1). The 25 µl PCR reaction comprised of 200 ng genomic DNA template, 0.2 µM of forward and reverse primer, 1 X Hi Fidelity buffer, 0.2 mM dNTP mixture, 2.0 mM $MgSO_4$ and 1U of Platinum Taq DNA Polymerase. PCR was performed as follows: denaturation at 94 °C for 2 minutes, 30 cycles of 94 °C for 30 seconds,

54–57 °C for 30 seconds (annealing temperature depending on the primers requirement) and 68 °C for 30 seconds followed by a final extension at 68 °C for 30 minutes. PCR products were checked with agarose gels, purified, labelled with biotin or digoxigenin-dUTP and were used as a probe in the FISH experiments (Chapter II; Section 2.2).

Table 4.1 Details of primers designed from 10 kb and 5 kb low copy unique sequence

Primer name	Primer Sequence (Forward)	Primer sequence (Reverse)	Annealing (°c)	Expected size (bp)
Eg5k_1p1	TATCTGTGGCCCGAGTTCTT	AAAAGACTAAATTCTTTGCCCAAC	55	5002
Eg5k_1p2	AACTAGGGCACAACCCCTTT	GTATGGGCAATCCCTCCTTT	55	5010
Eg5k_1p3	CCTTCAAAAGAATGAGTCCTTCAA	CCACCCACTTCCCCAATG	55	4998
Eg5k_1p4	ACCAAGAAATTGCACTGAGAA	GCAAGGCGTAGATAAGGGAAA	55	4998
Eg10k_1p1	TCGGTTCTGAAATTTATTGGCAG	TGCAATCGTCAATAATCGCAAG	57	10,001
Eg10k_1p2	CCATCCATCAGTTCCGTCCT	GGTAGCTTGTTTCTCCTTTCCA	57	10,004

ii) Development of short and low copy probes

Low copy region of Eg5k_1p3 was analysed by aligning the sequence against *E. guineensis* raw reads using Geneious (Kearse *et al*., 2012; http://www.geneious.com). Three primer pairs flanking the two identified low copy regions (Ex750 and Ex1332) were designed (Table 4.2). The 25 µl PCR reaction comprised of 100 ng genomic DNA template, 0.2 µM of forward and reverse primer, 1 mM KAPA buffer (Mg^{+}), 0.2 mM dNTP mixture, and 0.5U of KAPA Taq DNA Polymerase. PCR products were checked with agarose gels, purified, labelled with biotin or digoxigenin and were used as a probe in the FISH experiments (Chapter II; Section 2.2).

Table 4.2 Details of primers designed for Ex750 and Ex1322

Primer name	Primer Sequence (Forward)	Primer sequence (Reverse)	Annealing temperature (°c)	Expected size (bp)
Ex750_1	GGTCCCATCTCTTTATCGAGG	CCACCAGAATTACGAGGC	56	700
Ex1322_1	TCGGACAATAGCTACTGTACCG	GATTGTATGTGGATGGCTCCG	58	1250
Ex1322_2	CTCGGACAATAGCTACTGTACC	ATTGTATGTGGATGGCTCCG	57	1250

iii) Development of synthetic oligonucleotides (oligo) and low copy probes

Both of Ex750 and Ex1322 PCR amplified fragments were sent for sequencing. The sequencing data confirmed the Ex750 is a part of Ex1322 with a minimal single nucleotide polymorphism (SNP) region. *In situ* hybridisation of Ex750 showed a hybridisation on some regions on the chromosome, hence refinement of the unique oligo sequence was performed with Ex750 (discussed in section 4.3.1).

The schematic of the synthetic oligo development as shown in Figure 4.1. Three regions with the length of 61, 72 and 80 bp were selected from Ex750 sequence and synthesize as pre-labelled oligo. The pre-labelled oligos were hybridized on the *E. guineensis* metaphase chromosome singly and in a combination of the all three in a pool.

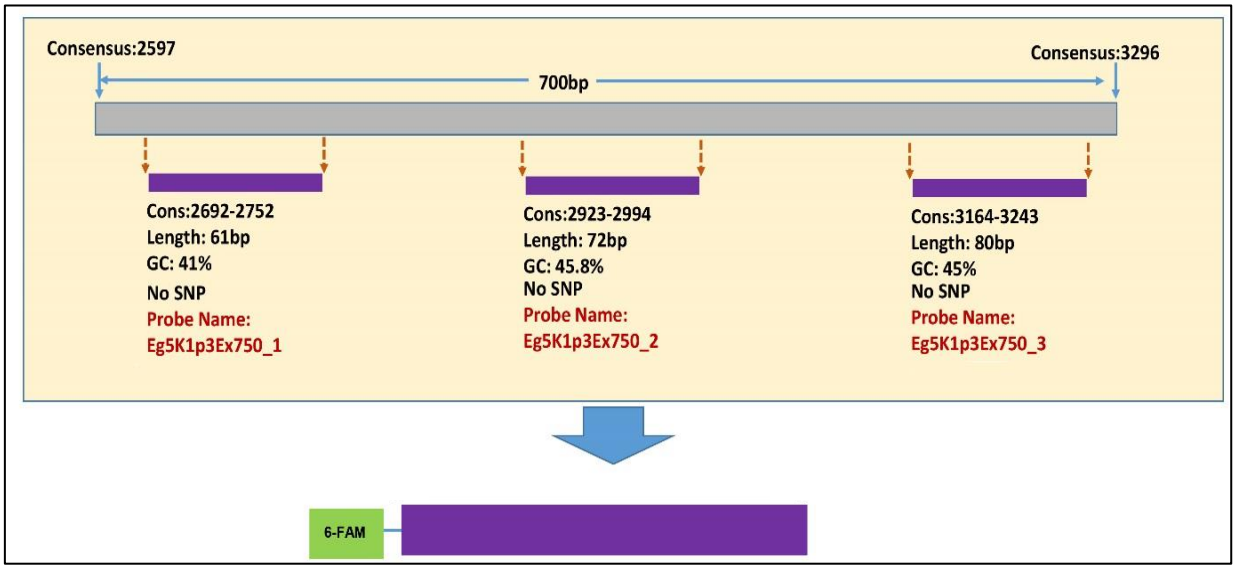


Figure 4.1 Schematic of the development of shorter synthetic oligonucleotide probes from Ex750. The selection criteria for individual oligo was indicated. 'Cons' representing the coordinate of Eg5k_1p3 where the designed oligo sequence was extracted. The selected region (Eg5K1p3Ex750_1, 2 and 3) were attached with a fluorescent dye (6-carboxyfluorescein; 6-FAM).

iv) Development of pooled, short, single copy synthetic oligonucleotide probes

A total of 60 and 35 oligos were designed within the single copy regions of Eg5k_1p3 sequence (Figure 4.2a, displayed as a blue font) for two sets of oligo libraries respectively; Set 1: 25 mer length and Set 2: 50 mer length. All the designed oligos are listed in APPENDIX 1.

Generally, the designed scheme for both sets as follows:

i) Oligo was designed from both strands (forward and reverse),

ii) The oligo must contain all ending bases (A, G, C or T) and

iii) The average GC content c.40 %.

The synthetic oligos were end-labelled with biotin using BioArray Terminal Labeling Kit (Enzo Life Sciences) following the manufacturer's instructions. The end-labelled oligos were then hybridized on oil palm chromosome with six combination schemes of the probes (Figure 4.2b).

Preliminary assessment of the approach was performed using ten oligos (45 mer length; APPENDIX 1 for the sequence) derived from 18S rDNA sequence (highly repeated genes) using the outlined strategy. The *in situ* hybridisation of the oligo (Figure 4.6) showed a similar hybridisation pattern and strength as showed by the PCR amplified 18S rDNA (Figure 3.12b), and pTA71 clone (Castilho *et al*., 2000, Madon *et al.,* 2005) indicating the efficiency of the approach at least with a repetitive DNA.

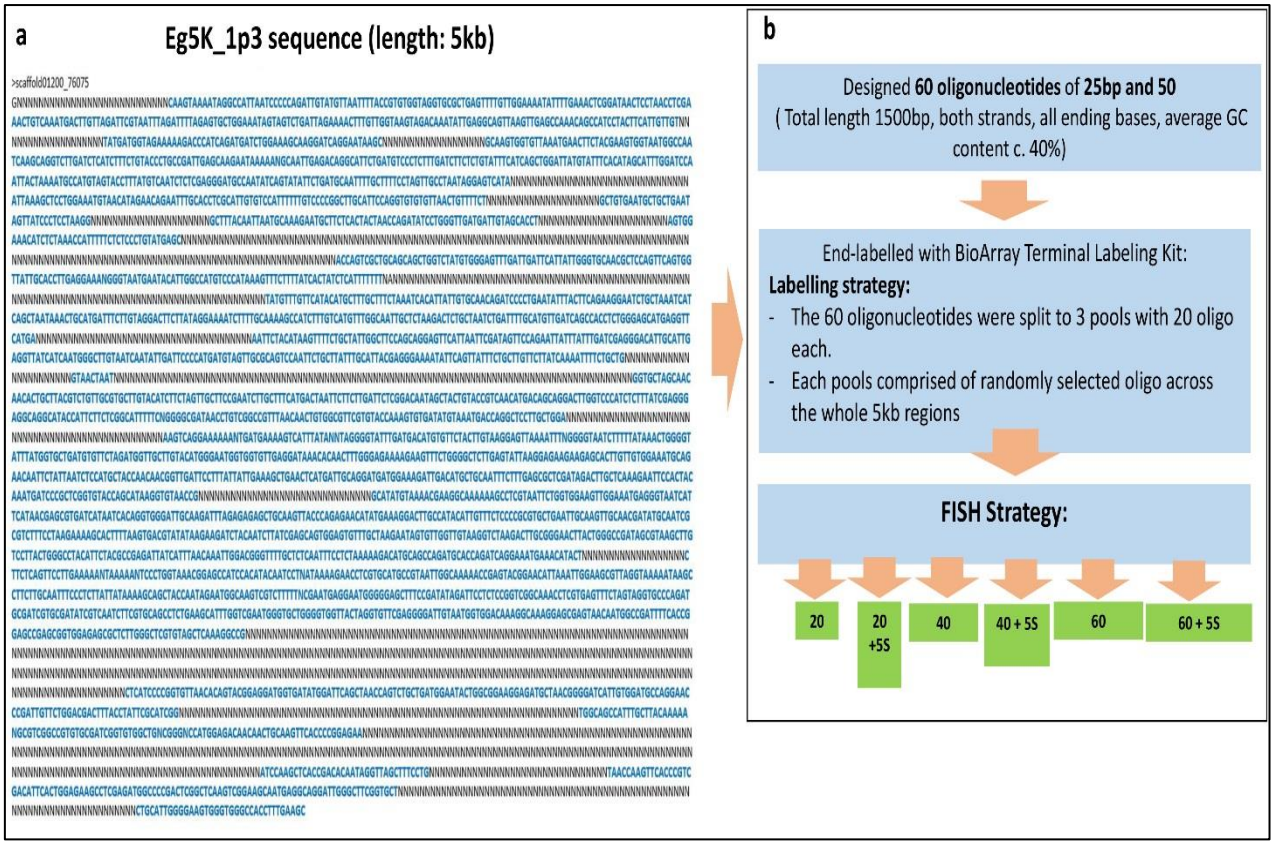


Figure 4.2 Development of pooled, short, single copy synthetic oligo probes for chromosomes identification with FISH. a) Unique regions of Eg5k_1p3 displayed as a blue font. b) An example of a schematic workflow showing the development of a larger pool single-copy DNA sequence as synthetic oligo DNA probes. The end-labelled oligo pools were used as probes for the FISH experiment as indicated in the green box. '5S' represent 5S rDNA that were used in dual-probe *in situ* hybridisation.

4.2.2.2 Development of chromosome-specific cytogenetics markers from massively synthesized, single copy oligonucleotide (oligo) pools

Development of the chromosome-specific oligonucleotide (oligo) pool followed the flowchart in Figure 4.3. Target chromosome regions (200 to 500 kb long) for probe design were chosen based on: presence of genomic regions with candidate loci of interest; and/or predicted chromosomal positions giving a unique combination of location and large differences in chromosome arm morphology, along with some redundancy. Preliminary selection of single-copy, short length oligonucleotides (45-60 mers) from the selected target regions of oil palm genome assembly (unpublished updated builds from Singh *et al*., 2013). For identification of repetitive sequences, we used a sample of 2 GB of 250bp (1.7 million reads) Illumina HiSeq reads (approximately 1x genome coverage). The oligonucleotides related to repetitive sequences were then eliminated (Chapter III) by screening against repeats identified by graph-based read clustering (RepeatExplorer; Novak *et al.,* 2010, 2013) and high-frequency *k*-mer analysis. Sequences having extreme AT/GC ratios were discarded. The repeat-filtered oligos were aligned to the oil palm reference genome to filter out those with duplicated locations in the genome (>80% similarity over the oligo sequence). Target regions with, typically, a probe density averaging ≥3 oligo/kb was selected as the final oligo set, giving between 1375 and 5598 oligonucleotides per locus, spanning 200 to 500kb of the genome assembly (Table 4.3). Three independent oligo libraries were synthesized (probe synthesis scale 700pM; total number of oligos was 52,508) with different fluorescent probes: OPAQUE (ATTO550), PPAQUE (ATTO488), QPAQUE (ATTO647) (Arbor Biosciences, Michigan, USA).

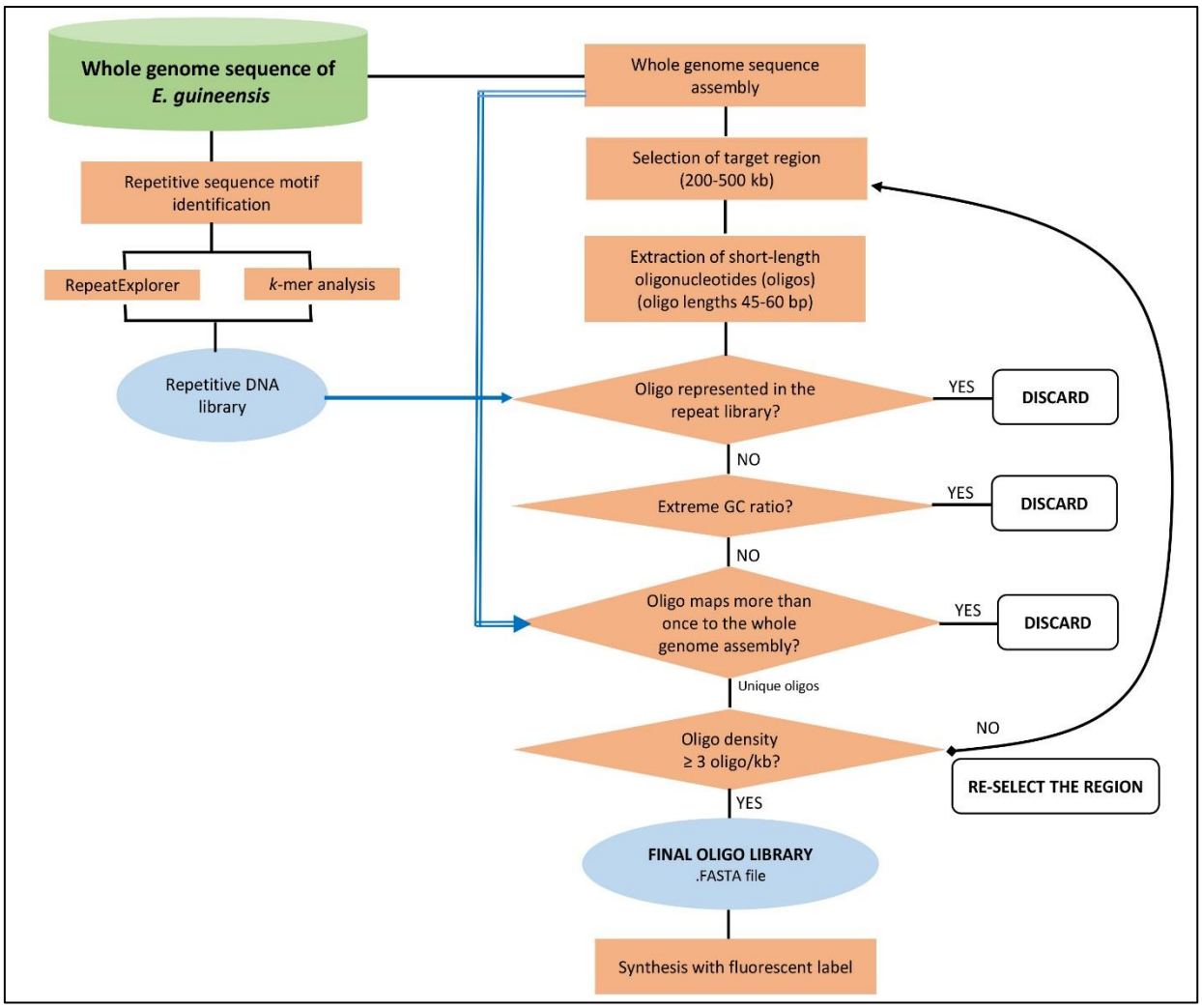


Figure 4.3 Workflow showing the development of *E. guineensis* chromosome-specific probes from the massive single copy oligo probes.

4.2.3 Preparation of chromosome spreads.

Chromosome spreads were prepared from plant root tips by using a technique adapted from Madon *et al*. (1995) and Schwarzacher and Heslop-Harrison (2000). In brief, the harvested root tips were pre-treated with 2 mM 8-hydroxyquinoline for 5-6 hours at 18 °C and fixed in 3:1 ethanol: glacial acetic acid (v/v) and stored in 70 % ethanol at 4 °C. The root tips were washed several times with citric acid-citrate buffer and digested at 37 °C for up to 4 hours in enzyme solutions containing 2-4% (w/v) cellulase (Sigma C1184; final concentration 10-20 U/ml), 0.2 % (w/v) 'Onozuka' RS cellulase (final concentration 10 U/ml), 3 % (v/v) pectinase (Sigma P4716 from *Aspergillus niger*; solution in 40% glycerol, final concentration 15-20 U/ml) in citric acid-citrate buffer. Mitotic chromosomes were

spread by squashing and heating onto a pre-cleaned glass slide in a drop of 60 % acetic acid under a coverslip, frozen before flicking off the coverslip, and left to air-dry before using for FISH.

4.2.4 Fluorescent *in situ* hybridisation.

Standard *in situ* hybridisation was performed according to Schwarzacher and Heslop-Harrison (2000) with slight modifications (section 2.2.8 for details).

For *in situ* hybridisation of massive oligo probes, the optimization was carried out based on the standard *in situ* hybridisation (Schwarzacher and Heslop-Harrison, 2000), Han *et al*. (2015) and Braz *et al*. (2018). A total of 40 µl probe was applied per slide, containing 50 % (v/v) formamide, 20 % (w/v) dextran sulphate, 2X SSC, 0.25 % (w/v) sodium dodecyl sulphate, 0.25 mM ethylenediamine-tetraacetic acid and 20 pmol probe. Probe and chromosomal DNA were denatured together on a heated block (Thermo Fisher Scientific) at 73 °C for 5 min under plastic coverslips and incubated in a moisture chamber at 37 °C for two days. A series of post-hybridisation washes were carried out with 2x SSC and 0.1x SSC at 42 °C. DAPI (4,6-diamidino-2-phenylindole) in CITIFLUOR AF1 (Chem Lab,) antifade solution was used to counterstain the chromosomes. At least two slides with 15 high-quality metaphases were hybridized and analyzed for each probe and species combination.

4.2.5 Fluorescence microscopy and imaging

Photographs were captured with a Nikon Eclipse N80*i* fluorescence microscope equipped with a DS-QiMc monochromatic camera (Nikon, Tokyo, Japan). Each metaphase was captured using four different filter sets; 1) UV-2E/C (excitation at 240-380; emission at 435-485 nm) for DAPI, 2) B-2E/C (excitation at 465-495; emission at 515-555 nm) for fluorescein and ATTO 488, 3) G-2E/C (excitation at 528-553; emission at 590-650 nm) for Alexa 594, ATTO 550 and ATTO 594 and 4) 31023 (excitation at 630-650 nm; emission at 665-695 nm) for ATTO 647N. The individual channels were pseudo-coloured to visualise the sites of probe hybridisation. The images were overlaid and further analyzed with Adobe Photoshop CS5 (Adobe Systems, San Jose, CA, USA) or NIS-Elements BR3.1 software (Nikon) using only cropping, and functions affecting the whole image equally.

4.3 Results

4.3.1 Optimizations towards developing *E. guineensis* chromosome-specific markers from single and low copy DNA

By using informatics approaches, four different strategies were adopted to develop chromosome-specific DNA probes from single and low copy sequences of oil palm. Validation with fluorescent *in situ* hybridisation showed that not any of the developed DNA probes were able to distinguish individual chromosomes. Below is the summary of the four strategies (*i-iv*):

i. Unique, long (10 kb and 5kb) and low copy DNA sequence

The unique and low copy regions DNA sequences with length 10 kb and 5 kb kindly provided by Malaysian Palm Oil Board (MPOB) and collaborators. Sequences that have been identified as a unique, low copy DNA from informatics approaches hybridized all over the oil palm chromosomes on broad centromeric region and painted the whole arm of the chromosome regions. Example of the *in situ* hybridisation with the probes as shown in Figure 4.4 (a) and (b).

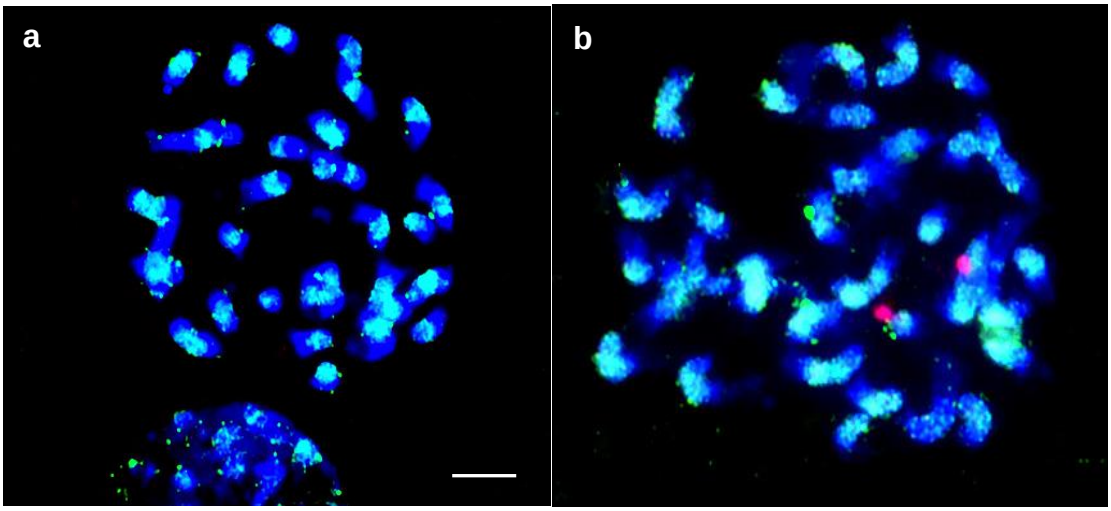


Figure 4.4 Development of chromosome-specific markers from low copy and unique sequence region. (a) and (b) are examples of hybridisation of the unique and low copy sequence of 5 kb DNA on *E. guineensis* chromosome (green: 5 kb probes; red: 18S rDNA). Hybridisation is specific to broad regions of chromosomes and is little assistance to chromosome arm identification.

ii. Short, low copy and unique sequence regions

Development of DNA probes in this strategy involved a re-examination of the low copy regions within the 5 kb and 10 kb sequences provided by the collaborator (section 4.3.1-i). The sequences were aligned against the 38.3 GB *E. guineensis* unassembled whole genome sequence (Singh *et al.,* 2013) data. Among the six investigated sequences, two low copy regions from Eg5K_1p3 were found to be unique based on the informatics pre-selection criteria (Ex750 and Ex1323). Both region Ex750 with the length of 700bp (Figure 4.5a) and Ex1323 (1300 bp) (Figure 4.5b) was not showed to have any homology to the repetitive DNA library (Chapter III and a library of retrotransposon amino acid motif from MolCyt Laboratory), as well as NCBI.

PCR amplification of both fragments with the designed primers resulted with the single band at the expected size (Table 4.2; Figure 4.5c). *In situ* hybridisation of both fragments (Ex750 and Ex1323) showed a different distribution hybridisation pattern across 32 *E. guineensis* chromosomes. Notably, the absence of Ex750 probes on some of the *E. guineensis* chromosome (5 pairs) able to distinguish oil palm chromosomes for the first time (4.5d). In contrast, the Ex1323 probes (4.5e) painted the whole chromosome as a similar pattern shown in Figure 4.4a and 4.4b. The obtained result showed the potential of the approach as well as the low copy regions probe Ex750 to be further developed as a chromosome-specific marker. Hence, the next strategy is discussed in (4.3.1-iii).

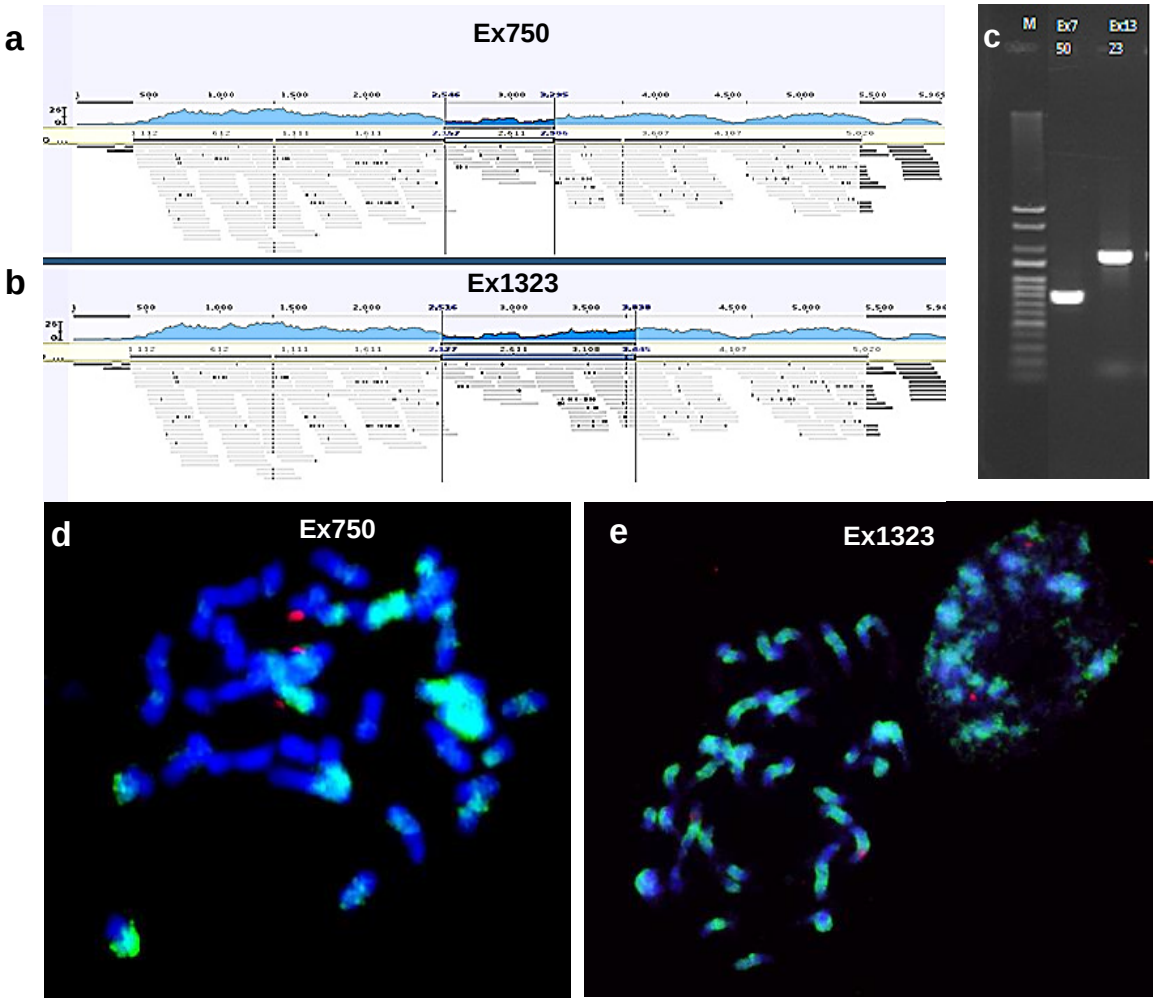


Figure 4.5 Development of chromosome-specific markers from short, low copy and unique sequence region. Two low copy regions (a: Ex750 and b: Ex1323) extracted from 5kb sequence probes and the amplified product showed a single band for both regions (c). *In situ* hybridisation of the probes showed the ability of Ex750 (d; green signal) to distinguish some of the chromosomes by being absent on the chromosomes compared to Ex1323 (e; green signal). The red signal on both *in situ* images (d and e) is 5S rDNA.

iii. Development of synthetic low-copy probes from Ex750

Based on *in-situ* hybridisation evidence showed from figure 4.5d, the low copy probe namely Ex750 that is unique to *E. guineensis* was able to distinguish about 2/3 of oil palm chromosomes, although the degree of certainty in pairing homologues was considered rather low. It was hypothesized whether a shorter piece of the low copy region Ex750 would be able to distinguish at least fewer number of *E. guineensis* individual chromosomes. To test the hypothesis, three short synthetic probes with length in between of 61-80 bp, GC content between 41-46 % and free from any SNPs were designed from the Ex750 sequence (Figure 4.1). The oligo was individually synthesized and labeled with 6-Carboxyfluorescein (6-FAM) fluorescent.

A few combinations of the three pre-labelled oligos were hybridized to oil palm mitotic chromosomes using the standard FISH method developed for *E. guineensis* as well as with various optimization on the critical parameter in FISH (*e.*g., Concentration of the hybridisation mixture components, length of hybridisation in 37 °C and stringency washes). However, no trace of the hybridisation signal was observed on the chromosomes. It was postulated, short length (c.200mer if combining all oligo) and low copy nature of the probes limits the probe ability to be observed as a clear signal.

iv. Development of pool, short, low copy synthetic oligonucleotide probes

The strategy developed in 4.3.1-iii was extended to develop a bigger pool of single copy probes covers a more extended stretch of DNA region. Workflow summarizes the probe development are as shown in Figure 4.2. A total of 60 single-copy DNA sequences with a length of 50 bp and 25 bp from 5 kb regions of Eg5K_1p3 sequence were synthesized. The oligo was end-labeled with terminal deoxynucleotidyl transferase (TdT) and labeled with biotin and further used as a probe for FISH.

Preliminary assessment of the approach with a pool of ten oligos with 45 mer length derived from 18S rDNA sequence (Figure 4. 6) showed a similar hybridisation pattern and strength as obtained by the PCR amplified 18S rDNA (Figure 3.12b) and pTA71 clone (Castilho *et al*., 2000, Madon *et al.,* 2005) indicates the efficiency of the approach at least with a highly repetitive DNA.

The *in-situ* hybridisation with pools of 60-short oligomer (1500mer in total) was able to show traces of dot signals on some chromosomes with high exposure time; with the current microscope facility, the average exposure needed is about 60s-80s (no figures attached as the traces of the signal is not visible in the printed image). However, the observed hybridisation signal was not conclusive as the signals were dispersed on several chromosomes, and not only on the expected chromosome from where the oligos were designed (chromosome 1; carrying 5S rDNA). Nevertheless, the visibility of the hybridisation signals supports the previous hypothesis that using a pool of short, and low copy/single copy sequence would be able to give specific hybridisation signal on the chromosomes.

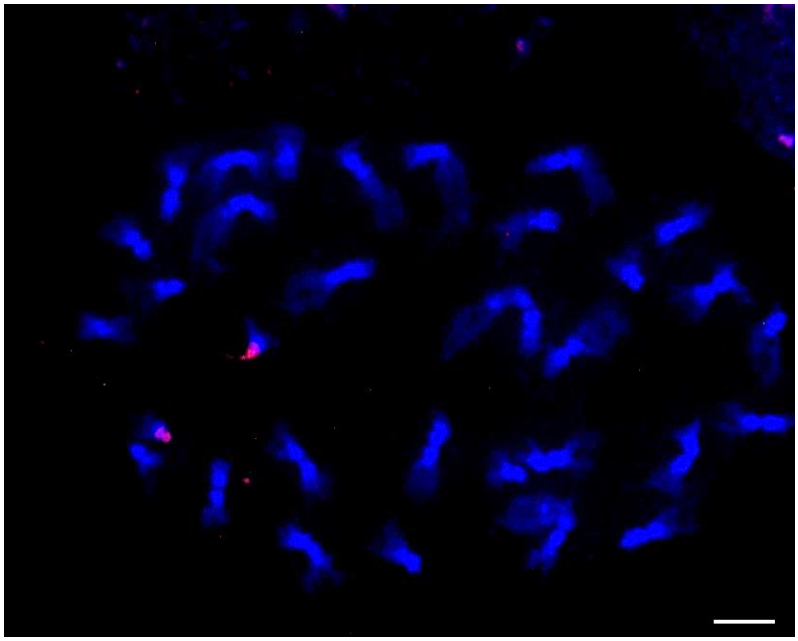


Figure 4.6 *In situ* hybridisation of biotin end-labelled oligo probe designed from 18S rDNA sequence. Oligo pools of 18S rDNA displayed in red on blue DAPI stained chromosome. Bar 5µm

4.3.2 Development of single copy short oligonucleotide (oligo) massive pools for *E. guineensis* chromosome identification

Taking advantage of the knowledge gathered from results of Chapter III and the strategies deployed in the optimization section (Section 4.3.1 (i-iv)), there was no repetitive DNA that was able to distinguish *E. guineensis* chromosome individually except for 18S rDNA, 5S rDNA, and Eg9CEN. Moreover, the optimization outcome has lead to a premise that postulating the potential of a pool of short length and single copy sequence able to show specific hybridisation signal on the *E. guineensis* chromosomes. Hence further enhancement of the strategy was performed by developing a massive pool of single copy short oligo pools for *E. guineensis* chromosome identification.

The oligo probes were selected from single and low copy sequences in the oil palm genome (Eg9, unpublished data; after Singh *et al.,* 2013). The selection of Eg9 was due to the denser assembled regions compared to the published assembly (Eg5), hence giving better density coverage within 1 kb region to design the oligo probes.

Through the workflow illustrated in Figure 4.3, three chromosome-specific oligo-FISH probes; OPAQUE, PPAQUE, and QPAQUE were developed (henceforth the whole libraries occasionally will be referred as EgOligoFISH). Each probe library contains 16,123 (OPAQUE), 18,685 (PPAQUE) and 17,700 (QPAQUE) different short-oligos (43-48 base sequence) derived from 19 different regions for 13 of the oil palm chromosomes (Table 4.3; Figure 4.7). There was no oligo designed from Chromosome 2 (another large chromosome aside from Chromosome 1), Chromosome 14 (small chromosome) and Chromosome 16 (the only acrocentric chromosome with specific chromosome marker; 18S rDNA). The selected regions spanned 200 kb-500 kb, and each chromosomal region is covered by 1,375 to 5,598 oligos. The final selection of the oligo was based on three parameters, i) probe density/kb window of the selected genome region, ii) a percentage of GC content, iii) homology to repetitive DNA (repetitive library from Chapter III). Oligos were discarded from the designed pool if the probe density was less than 3/kb, GC content exceeded from the range of 30-40% and showed similarity with the repetitive library of the oil palm genome.

The oligo libraries were designed by batches. OPAQUE (Oil Palm Analysis Queue) was the first designed library and hybridized on the *E. guineensis* mitotic chromosomes. All the optimizations with the *in-situ* hybridisation protocols were established using the OPAQUE library. Subsequently, PPAQUE and QPAQUE were designed to complete the identification of all 16 oil palm chromosomes. The stringency in eliminating the repeats content in both PPAQUE and QPAQUE was increased by screening the candidates' oligo to larger pools of repeats library. Both libraries (PPAQUE and QPAQUE) were explicitly selected from coordinates of the QTL linked traits published by Maizura *et al.*, 2017; Ting *et al*., 2016; Ting *et al.,* 2018) and detected *Ganoderma* resistance QTL co-localized with oil palm predicted R-genes (Tisne *et al.*, 2017) (Table 4.4).

Table 4.3 Design of oligo library from 19 regions for oligo-FISH probe development

Chromosome (Eg9 assembly)	Start position (kb)	End position (kb)	Region length (kb)	Number of oligos	Probe density/kb	GC content (%)	Oligo library
2a	60,000	60,500	500	2370	4.74	31.7	OPAQUE
3i	7,500	7,800	300	1690	5.63	34.3	QPAQUE
3ii	53,000	53,300	300	2666	8.89	34.0	QPAQUE
4i	32,300	32,600	300	2038	6.79	34.1	PPAQUE
4ii	174,000	174,500	500	3604	7.21	33.7	PPAQUE
5	80	300	220	2876	13.07	36.6	QPAQUE
5	16,700	17,000	300	3114	10.38	33.9	PPAQUE
6a	19,000	19,300	300	2904	9.68	34.1	QPAQUE
7a	3,100	3,300	200	2080	10.4	33.8	PPAQUE
7b	1,350	1,550	200	2365	11.83	34.8	QPAQUE
8	40,000	40,500	500	1617	3.23	31.2	OPAQUE
9	98,000	98,300	300	2943	9.81	35.3	PPAQUE
10a	1,750	2,250	500	1983	3.97	32.2	PPAQUE
10b	11,400	11,800	400	1375	3.44	35.4	QPAQUE
11	86,000	86,300	300	3824	12.75	34.6	QPAQUE
12	1,000	1,500	500	2737	5.47	31.7	OPAQUE
12	9,000	9,500	500	3801	7.6	31.7	OPAQUE
13	30,000	30,500	500	5598	11.2	33.4	OPAQUE
15	10,500	10,700	200	2923	14.62	33.4	PPAQUE

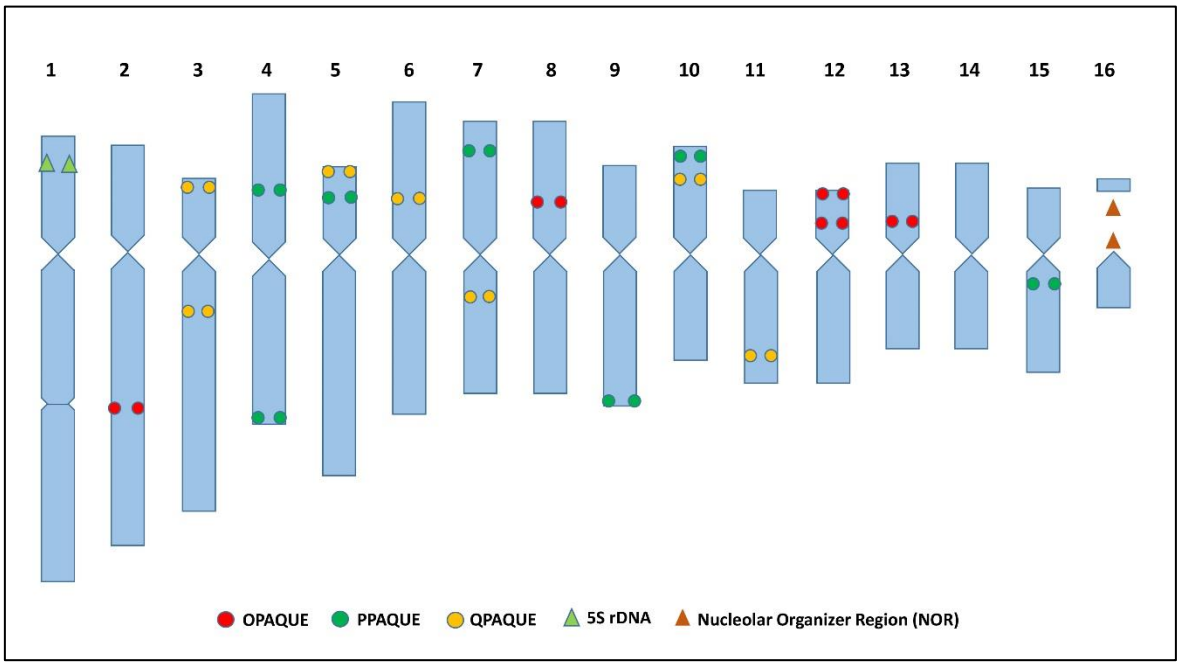


Figure 4.7 Expected chromosomal position of the designed oligo-FISH probe from *E. guineensis* reference genome (Eg9; unpublished; after Singh *et al.,* 2013). OPAQUE; red circle, PPAQUE; green circle and QPAQUE; yellow circle), rDNA (5S; displayed as a green triangle).

Table 4.4 Details of PPAQUE and QPAQUE library linked with *E. guineensis* QTL. The QTL for the respective Eg5 coordinates were translated to Eg9 coordinate. The selected coordinate for PPAQUE and QPAQUE was based on the Eg9 assembly.

Chr	PPAQUE coordinate	QPAQUE coordinate	Eg5 coordinate	QTL
3		7,500,000 – 7,800,000	1,528,660-11,115,373	IV; C16:0, C18:1, C14:0 and C18:0 (Ithnin *et al*., 2017)
3		71,100,000- 71,300,000	1,528,660-11,115,373	IV; C16:0, C18:1, C14:0 and C18:0 (Ithnin *et al*., 2017)
4	32,500,000 - 32,800,000		38,753,958-38,754,198	MTF, MFW, OTB and RL (Ithnin *et al.,* 2017)
4	174,000,000 -174,500,000			Nil
5	16,500,000 – 17,000,000		34,828,628-40,396,733	OTDP (Ting *et al.,* 2016)
5		80,000 – 400,000	34,828,628-40,396,733	OTDP (Ting *et al*., 2018)
6a		18,700,000 - 19,900,000	18,700,000 - 19,900,000	disease resistance protein rga4-like (Tisne *et al*., 2017)
7a	3,100,000 - 3,300,000		34,504,016-34,504,256	HT, KTB, MTF and OY (Ithnin *et al.,* 2017)
7b		1,350,000 – 1,550,000	1,590,892-1,591,132	HT, KTB, MTF and OY (Ithnin *et al.,* 2017)
9	36,600,000-36,700,00		36,600,000-36,700,00	disease resistance protein rpm1-like (Tisne *et al*., 2017)
10a	500,000 - 800,000	11,500,000 - 11,800,000	233,687-233,927	RL (Ithnin *et al.,* 2017)
11		86,000,000 - 86, 300,000	25,189,079-25,189,319	MTF (Ithnin *et al*., 2017)
15	15,000,000 - 15,250,000		15,000,000 - 15,250,000	disease resistance rpp13-like protein 1-like (*Tisne et al*., 2017)

Legend:

Chr (Chromosome), Iodine value (IV), mesocarp-to-fruit (MTF), mean fruit weight (MFW), oil-to-bunch (OTB), rachis length (RL), oil-to-dry mesocarp (OTDP), height (HT), kernel-to-bunch (KTB) and oil yield (OY)

4.3.3 *E. guineensis* FISH-based reference karyotype with EgOligoFISH

4.3.3.1 Optimization of the *in-situ* hybridisation with massive single copy oligo probes

A series of optimization were performed using one of the libraries (OPAQUE) to ensure the ideal *in-situ* hybridisation environment for the developed massive oligo probes (Table 4.5). From the observation, 20 pmol probes with two days incubation in 37 °C is an optimum condition to obtain a bright signal on the chromosomes for all three oligo probes set. Denaturation of the chromosome before applying the hybridisation mixture as suggested by Han *et al.,* 2015 and Braz *et al*., 2018 did not give any significant difference compared to the established method published by Schwarzacher and Heslop-Harrison (2000) (and in this study with Arecaceae species) which applying hybridisation mixture contained formamide on the non-denatured chromosomes.

Table 4.5 *In situ* hybridisation optimization of the massive single copy oligo on *E. guineensis* mitotic chromosomes.

Concentration of single probe (pmol)	**Days of incubation in a humid environment at 37°C**	**Application of formamide for chromosome denaturation**	***In situ* hybridisation signal observation**
10	1	Mixed with the hybridisation mixture	The signal from regions with lower probe density could not be observed clearly for certain probe
10	2	Mixed with the hybridisation mixture	The signal from regions with lower probe density could not be observed clearly for certain probe
10	2	Chromosome denatured with formamide before applying the hybridisation mixture (Han *et al.*, 2015; Braz *et al*., 2018)	The signal from regions with lower probe density could not be observed clearly for certain probe
20	1	Mixed with the hybridisation mixture	The signal from regions with lower probe density could not be observed clearly for certain probe

Table 4.5 continue

Concentration of single probe (pmol)	Days of incubation in a humid environment at 37°C	Application of formamide for chromosome denaturation	*In situ* hybridisation signal observation
20	2	Mixed with the hybridisation mixture	Bright signals observed for all probes
20	2	Chromosome denatured with formamide before applying the hybridisation mixture (Han *et al.*, 2015; Braz *et al.,* 2018)	Bright signals observed for all probes

4.3.3.2 Identification of individual *E. guineensis* mitotic chromosomes with EgOligoFISH

EgOligoFISH probes were hybridized on the *E. guineensis* mitotic chromosomes separately (Figure 4.8 a-c) and in a combinations of three libraries (Figure 4.9) to confirm the assignment of the individual *E. guineensis* chromosomes identity with the designated oligo probes. In general, nearly all the three designed probes showed FISH signals as expected from the probe design with some discrepancies from the expected designed location.

The next few paragraphs describes the ability of the designed oligonucleotide probes to identify *E. guineensis* chromosomes individually. Figure 4.8a to Figure 4.8c showed the *in situ* hybridization of the OPAQUE, PPAQUE and QPAQUE on the *E. guineensis* chromosomes and Figure 4.8d summarised the identification of the individual chromosomes based on the expected designed oligo and the observed *in situ* signal on the physical chromosomes. Wherever possible, the description includes the combination of both morphological features of *E. guineensis* chromosome (large-, medium- and small-size; acrocentric; tertiary constriction) and localization of repetitive DNA (5S rDNA) with the oligo to support the description. (Note: *E. guineensis* comprised of two pairs of comparatively long chromosomes; six pairs of comparatively medium-sized chromosomes; seven pairs of comparatively short-sized chromosomes).

The OPAQUE library was expected to distinguish four chromosomes; Chromosome 2a, 8, 12 and 13 with five different hybridisation regions (Table 4.4). The *in situ* hybridisation of the OPAQUE probe differentiated five individual chromosomes (Figure 4.8a) with eight pair clear signals on five chromosomes. An additional terminal hybridisation signal was observed on one arm of the small chromosome (orange arrow). Remarkably, the expected single intercalary sites of Chromosome 2 were observed as three pairs of hybridisation signals on one of the largest chromosomes; two signals closely localized on the p-arm (one pair intercalary; 1 pair sub-terminal) and one signal on the terminals of the q-arm. Dual hybridisation with 5S rDNA showed the localization of the rDNA probes in between of both two OPAQUE signals on the p-arm chromosome (Figure 4.8a; boxed chromosome). Hybridisation signals of Chromosome 8, 12 and 13 were as expected (Figure 4.8d).

In figure 4.8b, seven hybridisation sites were able to distinguish six individual chromosomes with PPAQUE probe as expected with noticeable difference on the observed location of the hybridisation sites compared to a designated location. Chromosome 15 (small-sized chromosome) showed a shift from the designed proximal region to terminal hybridisation sites (Figure 4.8d).

The QPAQUE oligo probe was designed to distinguish six chromosomes with seven different regions. *In-situ* hybridisation of the oligo probe showed the expected hybridisation sites with an additional faint signal on the terminal region of one large sub-metacentric chromosome (Figure 4.8c; orange arrow). The chromosome further identified as Chromosome 2 based on the chromosome morphological feature. Two hybridisation sites on one of the medium sizes chromosomes confirm Chromosome 3. Nevertheless, it was interesting to note that the region that was supposed to be separated by 45.5 Mb showed a clear paired signal close to each other. Two designed proximal regions from Chromosome 6 and Chromosome 7 observed as distinct sub-terminal signals. Dual hybridisation of the QPAQUE and 5S rDNA does not show any co-hybridisation of both probes on the same chromosome.

Triple hybridisation of OPAQUE, PPAQUE, and QPAQUE was performed (Figure 4.9) to assess the simultaneous use of the three libraries with different fluorochrome labels.

Furthermore, the usage of the three probes in one *in situ* hybridisation experiment further validate the uniqueness of hybridisation signals generated by the three oligo probes in distinguishing individual *E. guineensis* chromosome. Only one site of co-hybridisation was observed on the terminal region of one small chromosome (Figure 4.9a, orange arrow) where both signals of OPAQUE (displayed in red) and PPAQUE (displayed in green) were observed. The co-hybridisation of the OPAQUE and PPAQUE confirmed the identity of the additional small chromosome with the terminal hybridisation site identified by the OPAQUE probe as Chromosome 15. Moreover, additional faint hybridisation site of QPAQUE probe were observed on one of the large sub-metacentric chromosomes (identified as Chromosome 2) was not visible with simultaneous hybridisation.

Remarkably, simultaneous *in situ* hybridisation of three oligo libraries successfully distinguish 13 oil palm chromosomes including three chromosomes without any hybridisation sites but able to be differentiated with additional morphological features; one large chromosome (Chromosome 2), ii) one small chromosome (Chromosome 14) and iii) Chromosome 16, small acrocentric chromosomes. However, one of the chromosomes (Chromosome 10; Figure 4.8a; white star labelled) that have been identified with individual hybridisation could not be observed with this simultaneous hybridisation. This indicates that the low density (c. 3probe within 1 kb region) is not ideal to be used for simultaneous *in situ* hybridisation at least in *E. guineensis*. Furthermore, the ability of three simultaneous short, pre-labeled oligo probes in identifying the chromosome in this study is the first reported in the plant.

The built ideogram is based on consistent oligo hybridisation on at least three metaphase chromosome spreads from two mitotic slides of the individual and simultaneous *in situ* hybridisation (Figure 4.10). The unique FISH signals derived from the three oligo probes uniquely distinguished the 16 *E. guineensis* individual chromosomes along with the rDNA (5S) and other morphological features.

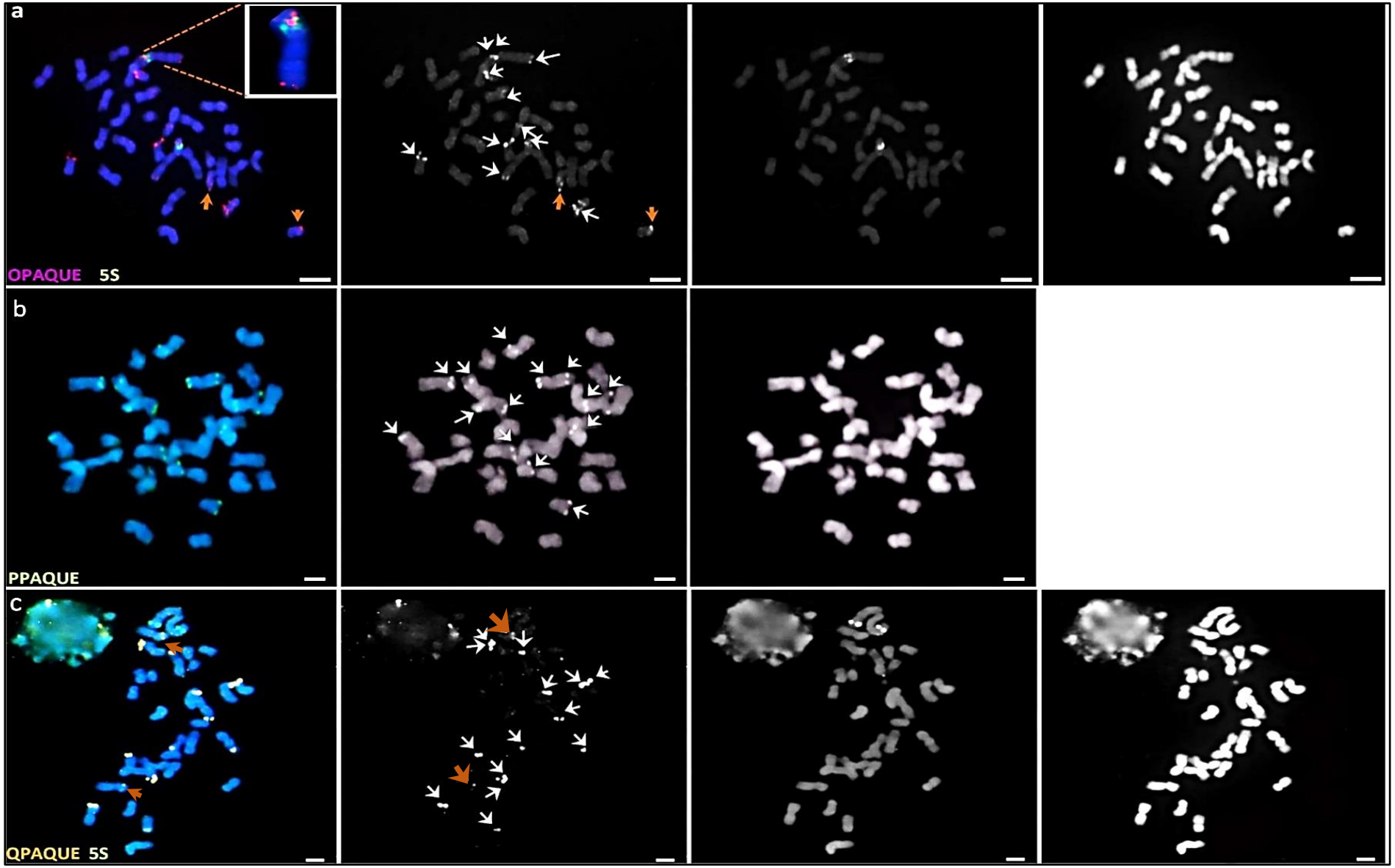

a
OPAQUE 5S
b
PPAQUE
c
QPAQUE 5S

Figure 4.8 continue

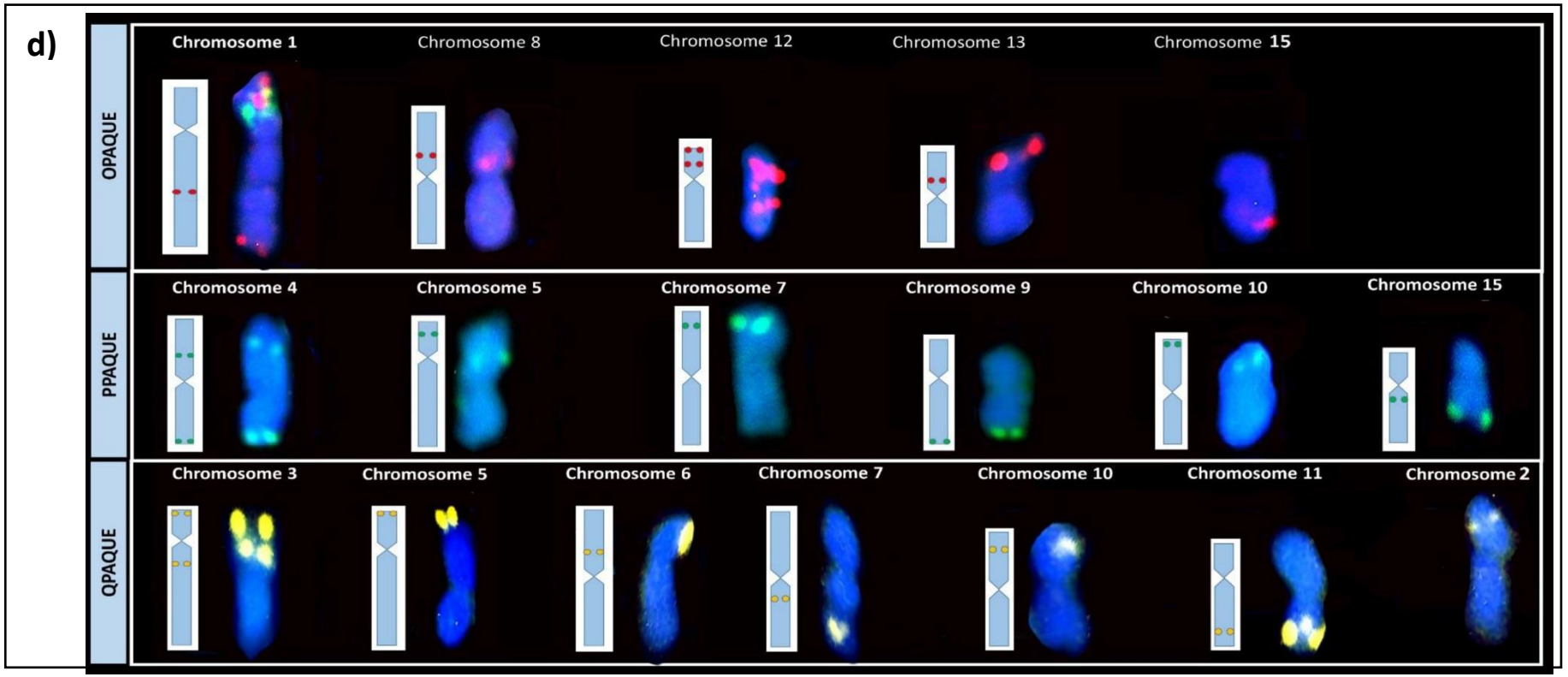


Figure 4.8 Individual FISH of the developed pre-labelled massive oligo probe on *E. guineensis* mitotic chromosome. White arrow showing the hybridized region of oligo probes (a-OPAQUE; b- PPAQUE and c- QPAQUE) on the individual *E. guineensis* chromosomes. Orange arrow showing additional hybridisation sites detected by the respective probes. Boxed chromosome shows the enlarged Chromosome 2. d) Comparison of expected location of the designed oligo (ideogram) and the physical localization of the respective oligo probes on *E. guineensis* chromosome. Scale bar: 5µm

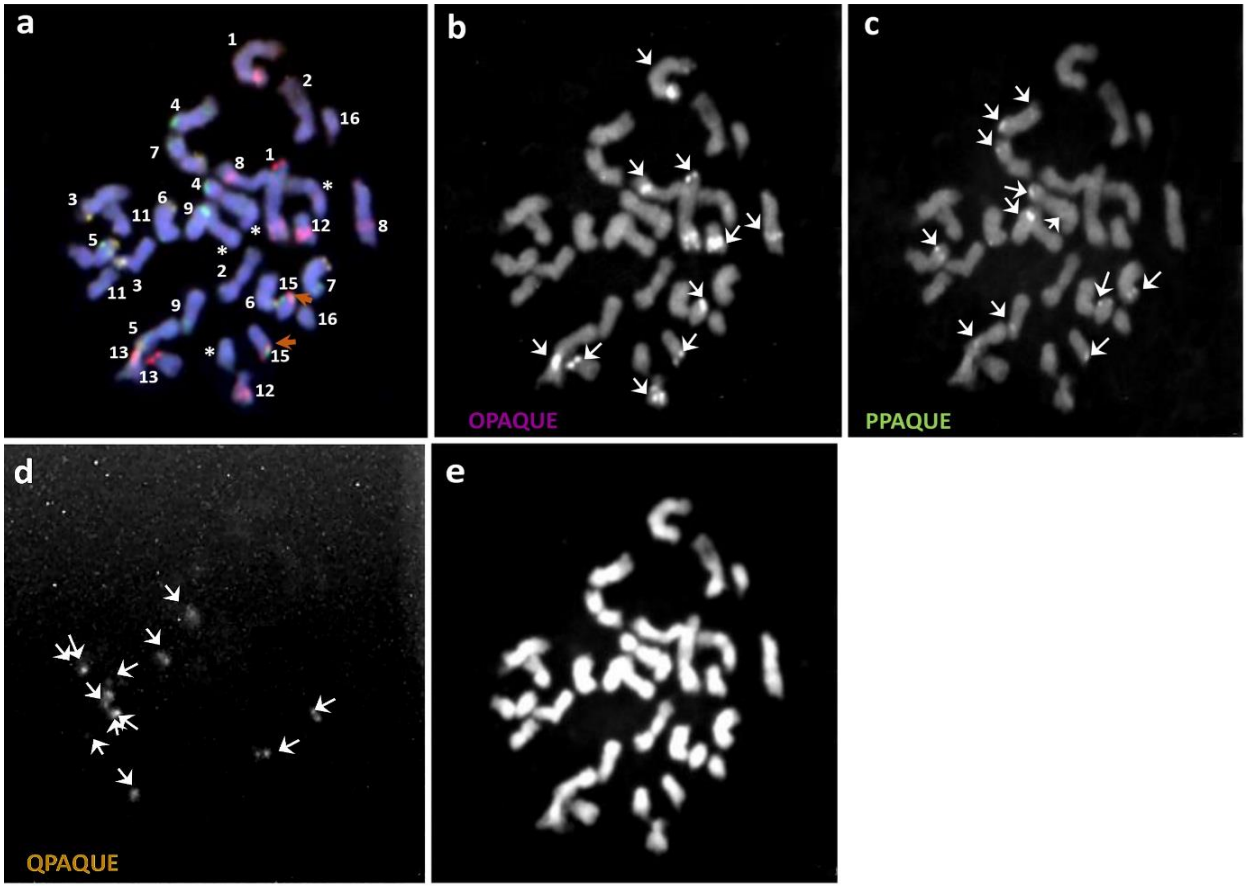


Figure 4.9 Simultaneous FISH with three pre-labeled oligo libraries (OPAQUE, PPAQUE, and QPAQUE) identified 16 pairs of *Elaeis guineensis* chromosomes as shown in (a, b, c and d). Figure 4.7e showing all 32 *E. guineensis* chromosomes (the quality of the chromosome was not the highest as it has been used for several times for re-probing). White arrow point to the regions hybridized by individual oligo probe (b: OPAQUE; c: PPAQUE; d: QPAQUE). Cross hybridisation of OPAQUE (displayed in red) and PPAQUE (displayed in green) confirm the identity of Chromosome 15 (a, orange arrow). Star labelled chromosomes are chromosomes that could not be identified with triple-FISH but identified with single-FISH (Chromosome 10) and Chromosome 14 that does not have any designated hybridisation region. Scale bar: 5µm.

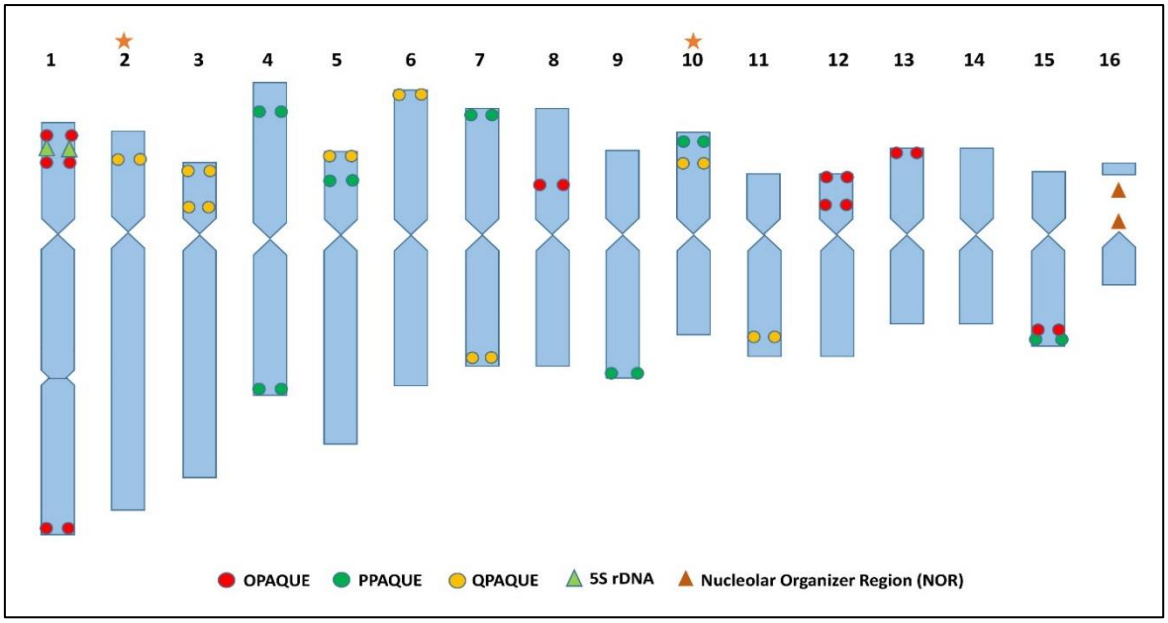


Figure 4.10 Propose *E. guineensis* karyotype with the physical location and relative distribution of the massive single copy oligo probes (OPAQUE; displayed in red, PPAQUE; displayed in green and QPAQUE; displayed in yellow) rDNA (5S; displayed as a green triangle) and NOR (brown triangle). The chromosomes in the karyotype were arranged by order of decreasing size. Assignment of Chromosome 2 and Chromosome 10 (orange star labelled) was based on the individual (single-probe) *in situ* hybridisation. Centromeric constrictions are drawn as a cross; secondary constriction (Chromosome 16) at the NOR as a gap; tertiary constriction (Chromosome 1) as a constriction.

4.4 Discussion

The massive oligonucleotide pools designed to single copy regions of the oil palm genome were able to identify each chromosome or chromosome arm (Table 4.3; Figure 4.8; Figure 4.9).

Several approaches have been experimented with to achieve the main goal of the study for identifying individual chromosome arms. Informatics analysis of the repetitive DNA which was presumed to be more than 50 % in the genome (Singh *et al*., 2013) did not reveal additional abundant repeat families except for what has been discovered by Castilho *et al.,* 2000 and Kubis *et al.,* 2003 (See Chapter III). The effort was further extended in analysing the low copy region of the assembled genome (Eg5). A specific hybridisation pattern of unique 5 kb and 10 kb probes to broad regions of chromosomes resemble of hybridisation pattern of some of the repetitive DNA used in Chapter III. Generally, the broad hybridisation pattern can be used to characterize the chromosome. Such findings were reported for repetitive DNA in oil palm (Castilho *et al*., 2000), carrot (Nowicka *et al*., 2015), meadow fescue (Krivankova *et al*., 2017) and black mustard (Wang *et al*., 2017), but it was little assistance in identifying the individual chromosomes of the species.

Three speculated factors that could explain the specific broad hybridisation of the unique low copy region probes are; 1) Possibility of the occurrence of repetitive regions within the probes; Even though PCR successfully amplified a single fragment from the whole genome, there are also chances of short regions within the probes that were not effectively masked. 2) Variation of stringency along the probe length and 3) Technical issues; Usage of high concentration of probes allows more hybridisation or less stringent washes which allow unspecific bound probes to remain on the chromosome.

As shown in the results, considerable effort was made to design probes spanning regions of the genome with 5 kb to 10 kb target regions. While low copy *in situ* hybridisation has been regularly reported in plants, results have in general not proved reproducible or scalable to make them widely applicable. Here, some of the low copy synthesized or PCR amplified probes showed inconsistency in reproducing the signals (4.3.1; ii-iv), and the

observed signals were not able to identify chromosomes robustly. Nevertheless, increasing the pool size of the synthetic oligo does give significant improvement to the observed signal. From a pool of three synthetic pre-labelled oligo (61 bp- 80bp) to a pool of 60 oligos with the length of 25 bp or 50 bp, it was noticeable that the observed hybridisation signals were better with a very minimal background. This is proven the strategy of using the bigger oligo-pools significantly working. However, the main issue was to obtain unique FISH probes that could generate a reproducible and robust signal that able to identify the individual 16 *E. guineensis* chromosomes.

Probe amplification approaches have been considered to enhance the faint signals as it has been successfully mapped single copy gene *via* FISH in onion (Romanov *et al*., 2015), *Rosa* (Kirov *et al*., 2014), oat (Sanz *et al*., 2012) and wheat (Perez *et al*., 2009). However, the parallel work carried out in our laboratory (Mr. Rafiq; MolCyt Lab) using the Tyramide Signal Amplification (TSA) system with low-copy probes on onion metaphase chromosome had proven unsuccessful. The optimizations resulted either gaining in the background or not showing any signal improvement.

Therefore, as the new technology of massive oligonucleotide pool synthesis became available, this was applied using about 1,000,000 bases of fluorochrome-labelled probes as 20,000 individual 43-48 mers, designed to about six different single copy regions of the genome. Based on the knowledge that have been gathered from the optimization stage and available synthetic massive oligos technology applied on various species (Yamada *et al.*, 2011, Beliveau *et al*., 2012, Han *et al*., 2015, Li *et al*., 2016, Braz *et al*., 2018), the pre-labeled, short, single copy oligo-pools probes were developed massively from the refined unpublished Eg9 assembly (after Singh *et al.*, 2013). The developed FISH-oligo probes (EgOligoFISH) were proven successfully identified 16 individual *E. guineensis* chromosome along with rDNA and further established *E. guineensis* karyotype for the very first time. Probe densities of more than 3 oligo/kb are sufficient to generate an observable FISH signal intensity on *E. guineensis* somatic metaphase chromosomes. It was noticeable brighter signals were observed when higher probes density used. Therefore, probes with high oligo density are suitable for future work in painting the pachytene chromosomes as has been

suggested by Han *et al*. (2015) in cucumber and Albert *et al*. (2019) in maize. Care was taken to remove all repetitive sequences from the designed probe pool, and the minimal background signal obtained confirmed that there was no significant dispersed repeat contamination in the synthesized pools. Nevertheless, it was perhaps surprising that the dispersed signal was so low, given that some of the PCR-amplified low-copy regions, designed on a similar strategy but much lower scale, did show signal across whole chromosomes.

The unpredicted hybridisation that was detected on three of the chromosomes (Chromosome 1, 2 and 15) may reflect assembly errors, or potentially translocations between chromosomes. In addition, collapsing of the sequences particularly at the ends of repetitive DNA regions was speculated to be the contribution of the hybridisation seen, and it is possible that genome duplications may be collapsed and are responsible for the multiple signals seen on chromosome 15. Missed-orientation of some of the probes is also likely to arise during assembly. Nevertheless, the main objective to distinguish individual *E. guineensis* chromosomes was successfully achieved using the designated massive oligosynthetic probes. Furthermore, this is also the first report in plants showing the ability of three simultaneous short, pre-labeled oligoprobes to identify 2/3 of the chromosome.

The collection of FISH probes developed in this study able to narrow down the specific genes on the physical chromosomes as some of it developed specifically from QTL linked regions (See 4.3.2; Table 4.4). Development of chromosome-specific markers enriched with genes is beneficial in facilitating the introgression of beneficial species for crop improvement. In the case of oil palm, *E. oleifera* (2n=2x=32) is an *E. guineensis* sister species which has an important pool of genes for oil palm improvement, including various agronomic traits and resistance to biotic and abiotic stresses (Murphy, 2014; Barcelos *et al*., 2015). In rye, chromosome-specific markers are proven effective in identifying alien chromosome aberrations which were presumed to facilitate the utilization of disease resistance genes from rye in wheat improvement (Wu *et al*., 2017). By using chromosome, specific markers developed from transcriptome sequences, the lines that are carrying chromosome aberrations can be identified.

The system/technique developed here is beneficial to any lab that have limited microscopy resources. Beliveau *et al.* (2012, 2015 and 2017) have reported that the development of the same type of oligo, however using the STORM system to observe the signal will be a limitation for certain laboratory. Here, a standard fluorescence microscope is sufficient enough to visualize the developed oligo-FISH probes hybridized genomic regions ranging in size from tens of kilobases to megabases in Arecaceae family. The method also will become a powerful tool in constructing a cytogenetic map of any species as it gives researchers precise control over the location and patterning of each probe set by manipulating information from the sequence data. The direct-labelled oligonucleotide probe pools currently cost approximately $USD 1,750 for 700 pmole, only suitable for about 70 slides. It is possible that the synthesis price will be lower in the future, or alternatively, that routine amplification and labelling methods (including end-labelling with TdT as used here) may be developed. As discussed, the reasons that the PCR low copy probes failed to work as robust labels for single chromosomal loci or regions (at a probe cost of c. $400, but usable for more assays), were unclear.

As a conclusion, this finding demonstrates the very first established *E. guineensis* reference karyotype using a robustly developed single copy oligo probes (EgOligoFISH). This is also the first report in plant showing the ability of three simultaneous short, pre-labeled oligoprobes to identify 2/3 of the chromosome. The ability of developed EgOligoFISH as a basis for further comparative cytogenetics research through cross-species chromosome painting will be discussed in Chapter V.

CHAPTER V

Utility of the developed massive single-copy oligo probes across Arecaceae

---

5.1 Introduction

The family Arecaceae (Palmae) is one of the oldest flowering plants families and consists of approximately 181 genera (Govaerts *et al*., 2015) with fossils dating from the Cretaceous period (Purseglove, 1972). The palm family is the third most economically important family after the grasses (Poaceae) and legumes (Leguminosae) (http://www.fao.org/faostat/en/#data/QC). Many Arecaceae species are exploited in some way for human purposes such as oil palm (*Elaeis guineensis*), date palm (*Phoenix dactylifera*) and coconut (*Cocos nucifera*) which occupy a particularly high profile, due to their economic importance (Balick and Beck, 1990).

As introduced in Chapter I, the genus *Elaeis* (tribe Cocoseae) consists of two species, *E. guineensis* from West Africa and *E. oleifera* from Central and South America. The commercial *E. guineensis* has a higher yield compared to *E. oleifera*. Nevertheless, *E. oleifera* has remarkable breeding traits of interest, e.g., higher unsaturated fatty acid content, lower height increment, and resistance to diseases (Cochard *et al.,* 2005). Published phylogenies all suggest *Phoenix* is a sister to the branch with *Cocos* and *Elaeis*. However, most published dates of separation have very wide confidence intervals and differ between publications, sometimes being estimated based on very few gene sequences, despite the availability of whole genome sequences. Singh *et al*. (2013) predicted a divergence of 51 million years ago (MYA) between *E. oleifera* and *E. guineensis* and 65 MYA between oil palm (*Elaeis*) and dates (*Phoenix*) (Figure 5.1). However, remarkably, *E. guineensis* and *E. oleifera* give rise to a fertile hybrid (Hardon and Tan, 1969). Xiao *et al.* (2017) reported the divergence time of *Cocos nucifera* and *Elaeis guineensis* as about 46.0 MYA (25.4–83.3) is more recent than *Cocos nucifera* and *Phoenix dactylifera* at

about 71 MYA (46.8-107.5), suggesting a closer relationship between *C. nucifera* and *E. guineensis* (Figure 5.1).

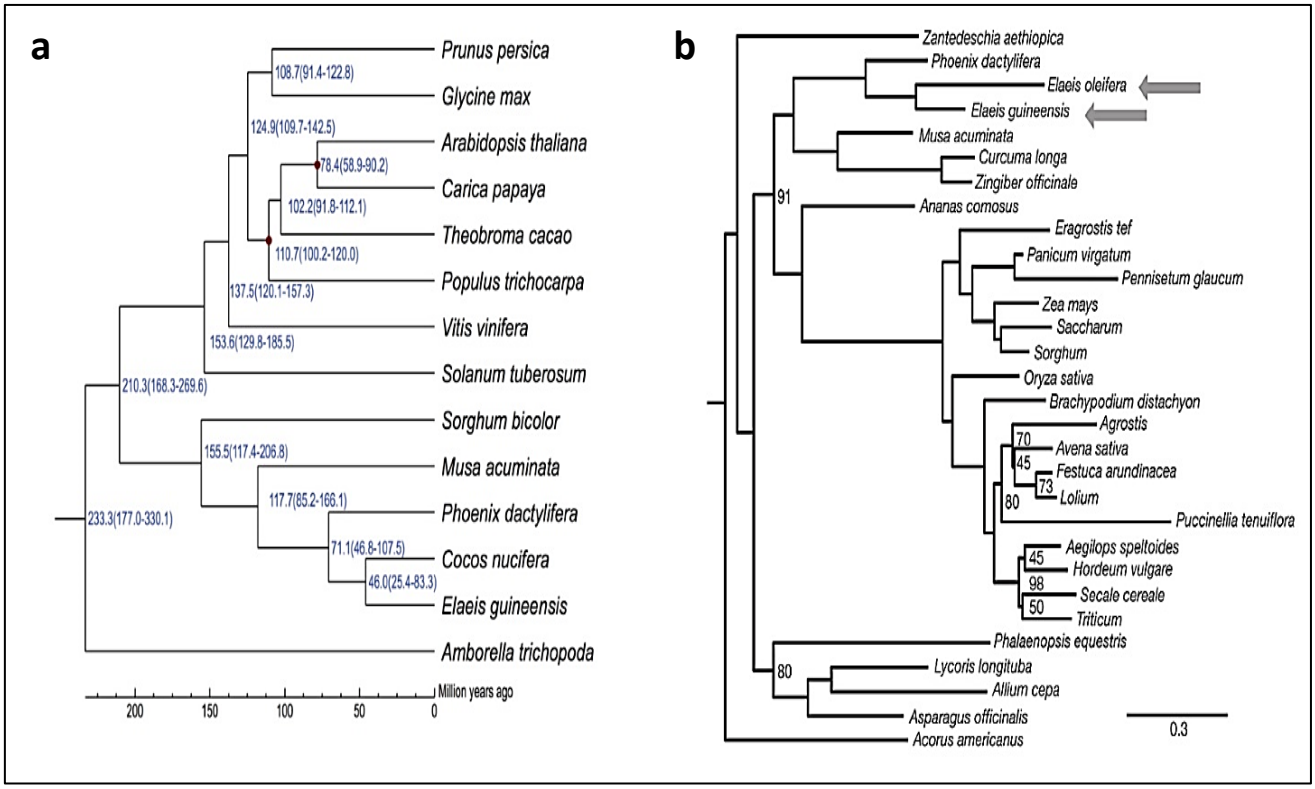


Figure 5.1 Estimation of divergence time of *E. guineensis*, *E. oleifera*, *C. nucifera,* and *P. dactylifera*. a) Reproduced from Xiao *et al*. (2017). The blue numbers on the nodes are the divergence time from present (million years ago with confidence bounds); the red nodes indicate the previously published calibration times. The Bayesian relaxed molecular clock approach was used to estimate species divergence based on the four degenerate sites. b) Reproduced from Singh *et al*. (2013), a maximum likelihood tree of monocotyledonous taxa is shown along with bootstrap values. Scale bar indicates the mean number of substitutions per site. Phylogenetic dating using conservative constraints predicted a divergence 65 million years ago (MYA) between date and oil palm and 51 MYA between *E. oleifera* and *E. guineensis*.

Comparative studies are beneficial in explaining the function of biological structures and in providing markers for evolutionary investigation, whether in the context of plant breeding, ecology, or biodiversity (Heslop-Harrison, 2000). Moreover, comparative genomics is an important and expanding field of research, and the genome-wide comparison of the chromosome constitution of different species makes a significant contribution to this field. The increasing amount of plant genome sequence data enables robust comparative analyses in a plant kingdom in answering biological questions by transferring knowledge from a model plant to another genome of interest. Combination of sequence data that become handy now as well as comparative fluorescent *in situ* hybridisation (FISH) mapping lead to a powerful approach for establishing chromosome homology maps, defining the sites of chromosome fusions and fissions, investigating chromosome rearrangements during evolution and constructing ancestral karyotypes (Young *et al.,* 2011; Betekhtin et al., 2014; Lou *et al*., 2014; Braz *et al.,* 2018, Albert *et al.,* 2019).

The long growth period before reproductive years hampers conventional breeding progress of the three palm species that account for the vast majority of the Arecaceae family's economic importance; relatively limited genomics study has been carried out. Most of the developed molecular studies focus on characterizing the germplasm, and, to a lesser extent, developing quantitative trait loci (QTL) (Meerow *et al*., 2012). Previous reported comparative genome studies within Arecaceae species have been based on genetic analysis of polymorphic DNA markers (Billotte *et al.,* 2001; 2004, Akkak *et al.,* 2009; Ting *et al.,* 2010; Zaki *et al*., 2012; Ting *et al.,* 2014; Filho *et al*., 2017), as well as genome sequence data (Dous *et al*., 2011; Singh *et al*., 2013; Matthew *et al*., 2014; Xiao *et al*., 2019). With the availability of genome sequence of oil palm (*Elaeis guineensis* and *Elaeis oleifera*), date palm (*Phoenix dactylifera*) and coconut palm (*Cocos nucifera*), integration of *in silico* and physical comparative mapping *via* FISH may further elucidate the synteny and co-linearity of the genes within this economically important crops.

Comparative genetic mapping by cross-hybridizing the massive-oligo pools synthetic probes in different species is promising in examining genome architecture and genome relationships within the genus (Meng *et al*., 2018; Hou *et al*., 2018; Braz *et al*., 2018; Qu *et*

al., 2017; Han *et al*., 2015). This chapter discusses the utility of the *E. guineensis* chromosome-specific cytogenetic markers developed from massive-oligo pools of single-copy DNA (see Chapter IV) in identifying chromosomes and further establish karyotype of *E. oleifera.* Moreover*,* the utility of the developed massive oligo pools probes was extended in identifying physical chromosomes of *Cocos nucifera*, another species in the tribe of Cocoseae as well as *Phoenix dactylifera*.

5.1.1 *Elaeis oleifera*

Elaeis oleifera (2n=2x=32) is a species in the oil palm genus along with the commercial *Elaeis guineensis*. The *E. oleifera* was referred to as *Elaeis melanococca* and *Corozo oleifera* when the species first documented in 1700 century (Hardon and Tan, 1969). Later, Wessel Boer (1965) supported the classification in the genus *Elaeis* and suggested *E. oleifera* (Kunth) Cortés as a South American species. *E. oleifera* populations occur naturally in South-Central America, from Honduras to Colombia and in the Amazon region, growing in both shaded and flooding conditions, suggesting broader environmental adaptability compared to the *E. guineensis* (Corley and Tinker, 2015). There are no documented historical indications of artificial selection for improved yield in *E. oleifera* which remains significantly lower compared to *E. guineensis* (Barcelos *et al*., 2015)

E. oleifera agronomic potential gained oil palm breeders' interests at the beginning of the 1900 century. In the 1920s, *E. oleifera* was introduced to Africa and 1950s to Asia. However, it is only in 1975 that *E. oleifera* natural populations have been thoroughly sampled to establish *ex situ* germplasm collections in Malaysia, from Ivory Coast, Costa Rica and Brazil (Escobar, 1981; Rajanaidu, 1986; Barcelos *et al*., 2002)

This American species is seen as a promising genetic resource for oil palm improvement and is currently used in oil palm hybrid (*E. guineensis* × *E. oleifera*) breeding programs (Barcelos *et al*., 2015). Despite its lower yield, it has attracted oil palm breeder attention due to several interesting traits which can have significant economic implications if introgressed into commercial *E. guineensis.*

The most important traits of *E. oleifera* are:

a) Shorter height, due to a slow trunk annual growth height (5-10 cm), which facilitates harvesting and ensures a longer lifespan of plantations (Corley and Tinker, 2003).

b) A higher proportion of unsaturated fatty acids and vitamins A and E content, improving the oil nutritional value (Nagendran *et al.,* 2000; Montoya *et al*., 2014)

c) Lower lipase activity in the mature fruit mesocarp, beneficial in extending the time between harvest and fruit processing (Sambanthamurthi *et al*., 1995; Cadena *et al*., 2013)

d) High level of pest and disease tolerance; resistance to *Fusarium* wilt and bud-rot caused by *Phytophthora palmivora* (Barcelos, 1986; Corley and Tinker, 2003; Torres *et al*. 2016)

There is limited genomic-related research carried out for *E. oleifera* despite the species' economically important traits. The F_1 of *E. oleifera* X *E. guineensis* or *OG* hybrids are planted in a large area of Latin America due to the disease resistance, and several *E. oleifera* palm with traits of interest were introgressed into high yielding *E. guineensis* variety. However, due to poorly understood cytogenetics problem, the F_1 of *OG* hybrids still faces reproductive limitations, notably lower natural fertility that results in reduced pollen production with lower viability and poorer dispersion (Corley and Tinker, 2003). Consequently, hybrids exhibit fruit abortion and lower oil production. Assisted pollination is used to overcome this limitation in the plantation with consequent of high costs and labor inconvenience (Barcelos *et al*., 2015).

In 1969, Hardon and Tan investigated the *E. guineensis* X *E. oleifera* F_1 hybrid vigor, cytology, and fertility which resulted in the evidence of the crossability of the two species. Nevertheless, the same study observed the reduced viability and fertility on the obtained F_1 plantlet. Later, Madon *et al*. (1998) cytologically characterized 16 *E. oleifera* diploid chromosomes to three major groups based on the physical length which contains two subterminal (acrocentric), thirteen submedian (submetacentric) and one telocentric. As *E. guineensis* and *E. oleifera* can be crossed to produce interspecific hybrids, the compatibility

of the chromosome length similarity, as well as genome compositions potentially, allows the gene exchange between these *Elaeis* species as suggested by Hardon and Tan (1969) and Madon *et al*. (1998). Nevertheless, works that involved *E. oleifera* genomics only focusing on the analysis of various polymorphic marker (Zaki *et al*., 2012; Beule *et al*., 2015; Filho *et al*., 2017) and QTL mapping of several interesting traits of interspecific hybrids (Montoya *et al*., 2014; Ting *et al*., 2014; Lee *et al*., 2015). The released *E. oleifera* sequence genome (Singh *et al*., 2013) was only for comparative purposes. Apart from presented evidence of retained *E. guineensis* segmental duplication in *E. oleifera* which support the pre-dated divergence of the African and American oil palm, minimal information was revealed.

5.1.2 *Cocos nucifera*

Coconut palm (*Cocos nucifera* L., 2n=32), is one of the monocotyledon oil crop species in *Cocoseae* along with *E. guineensis* and *E. oleifera*. It is widely cultivated in 93 tropical countries with more than 12 million hectares of growth area due to its wide application in agriculture and industry (www.fao.org/faostat/en/). Numerous home gardens also grow a few coconut palms each, to provide food; 'water' and sap to drink; oil for cooking and non-edible uses; coir for insulation, matting, manufacturing and as planting medium; leaves for fencing; sugar, vinegar and alcoholic beverages from sap; timber and wood for construction; fuel from the husk, leaves and shells; materials for artefacts, traditional medicine as well as ritual purposes (Johnson *et al*., 2018).

C. nucifera is generally categorized into 'Tall' which flowers after 8-10 years of planting and 'Dwarf,' that flowers earlier (4-6 years after planting) (Xiao *et al.,* 2017). These two groups of coconut varieties have very distinct characteristics; the Tall coconut with height reach to 30 metres, allogamous (allogamy: cross-fertilisation) with medium to large sized fruits. Dwarf coconut is shorter (12 m), autogamous (autogamy; self-fertilisation) and generally classified into three groups according to its fruit color: yellow, red, and green (Aragão *et al*., 2010) that is controlled by two loci, R and G (Bourdeix, 1988). Several theories are explaining the origin of Dwarf coconut; one of them establishes that it is a variant of the Tall coconut that arose by mutation or inbreeding (Swaminathan and Nambiar, 1961). South-east Asia (Cambodia, Hainan Island, Indonesia, Malaysia, Philippines, Thailand, and

Vietnam) is one of the most diverse regions for coconut cultivation with highly domesticated Dwarf coconut and in the Pacific coast of the Americas, coconuts were native to Panama. Recent studies have shown that the local Panama coconut is closely related to those from the Philippines (Baudouin *et al.,* 2013). The literature reports that both types are diploids with 2n=2x=32 chromosomes (Sisunandar *et al*., 2007) with minor differences in the karyotype. The karyotype was considered asymmetric, with 11 metacentrics and five sub-metacentric chromosomes pair chromosome length ranged from 5.57 µm to 2.13 µm (Pereira *et al*., 2017). Recently, the coconut genome has been published with an estimated genome size of 2.42 Gb (Xiao *et al*., 2017).

5.1.3 *Phoenix dactylifera*

The date palm (*Phoenix dactylifera* L.) is a perennial monocot, belonging to the Arecaceae family and is widely cultivated in arid and semi-arid countries (Alami-Saeid *et al.,* 2014). The genus comprises of 14 recognized species (2n=36) in which some of the species extensively used for ornamental purposes (*Phoenix roebelenii* and *Phoenix canariensis*), food (sap from *Phoenix sylvestris*), clothing, construction, fiber, feed for livestock, as well as having cultural importance (Barrow *et al*., 1998).

Together with the olive, grape, and fig, date palms were amongst the first fruit crops domesticated in the Old World (Zohary and Speigel-Roy, 1975). The earliest cultivation of *P. dactylifera* was recorded in 3700 BC in the area between the Euphrates and the Nile River (Munier *et al*., 1973). This oldest domesticated tree is capable of living over 100 productive years (Al-Mssalem *et al.,* 2013). The fruit of the date palm can be eaten fresh or dried or transformed into a large variety of products such as syrup or paste. Each year, more than eight million tons are produced worldwide, and this number is continuously growing owing to substantial scale efforts to increase the numbers of fruit-producing trees (http://faostat.fao.org/default.aspx). The sap, with high sugar content, may also be harvested and used as a sweetener or for fermentation.

The date palm is dioecious with separate male and female trees with the female bears the fruit. The late initial reproductive age (5-10 years) is the major constraint for genetic improvement, and for centuries the production of dates relies on the clonal propagated

female palms (Adawy *et al*., 2015). Despite the apparent importance of the date palm, few genetic resources exist due to its long generation time. Previous research on the date palm chromosome number suggests it has 18 chromosome pairs (2n=2x=36) (Beal, 1937) even though some evidence for other numbers have been presented (Salih *et al*., 1987, http://www.actahort.org/books/882/882_28.htm).

A large amount of genomic data has been generated for the date palm in the past ten years. The first draft sequence of a female commercial cultivar of date palm published in 2011 (Dous *et al*., 2011), and further improved in 2013 (Al-Mssalem *et al*., 2013) with reported genome size approximately 670 Mb distributed on 18 chromosomes. Nevertheless, due to the highly heterozygous nature of the date palm, the improved version of the assembled genome is still highly fragmented (>80 000 scaffolds) (Gros-Balthazard *et al*., 2018).

5.2 Materials and methods

5.2.1 Plant materials

The oil palm (*E. guineensis* and *E. oleifera*) materials used for chromosome preparation were published by Singh *et al.* (2013) and are currently maintained at the MPOB Research Station, Kluang, Johor, Malaysia. Meristematic root tips of *E. guineensis* were collected from three Pisifera palms (0.182/77, 0.182/30 and 0.182/7) and *E. oleifera* were collected from Costa Rica germplasm. *Cocos nucifera* (garden center source) and *Phoenix dactylifera* (*cv* Medjool from seed from Nouf Alsayeid) were maintained in University of Leicester greenhouse. Genomic DNA (0.182/77) was extracted and purified from a spear leaf using the modified CTAB method as described by Doyle and Doyle (1990).

5.2.2 *In silico* comparative analysis across Arecaceae

The 52,508 oligonucleotide sequences (henceforth referred to as EgOligoFISH; Chapter IV) were aligned to the reference genome of *E. oleifera* (Singh *et al*. 2013; 26,769 scaffolds), *Phoenix dactylifera* (Dous *et al.,* 2011.; 57,277 scaffolds, Al-Mssalem *et al*., 2013; 80,315 scaffolds) and *Cocos nucifera* (Xiao *et al*., 2017; 11, 694 scaffolds) by Geneious program (Kearse *et al*. 2012) to investigate the homology of the oligo derived *E. guineensis* on other Arecaceae family member.

5.2.3 Preparation of chromosome

Chromosome spreads were prepared from *E. guineensis*, *E. oleifera*, *Phoenix dactylifera* and *Cocos nucifera* root tips as described in Section 4.2.3 with minor modifications for *P. dactylifera* and *C. nucifera*.

5.2.4 Fluorescent *in situ* hybridisation, microscopy, and imaging.

In situ hybridisation was performed according to Schwarzacher and Heslop-Harrison (2000) and established approach developed for massive oligo probes (section 2.2.8 and 4.2.4). Briefly, a total of 40 µl probe was applied per slide, containing 50 % (v/v) formamide, 20 % (w/v) dextran sulphate, 2x SSC, 0.25 % (w/v) sodium dodecyl sulphate, 0.25 mM ethylenediamine-tetraacetic acid and 20 pmol of oligo probe. Probe and chromosomal DNA were denatured together on a heated block (Thermo Fisher Scientific) at 73 °C for 5 min under plastic coverslips and incubated in a moisture chamber at 37 °C for two days. A series of post-hybridisation washes were carried out with 2x SSC and 0.1x SSC at 42 °C. DAPI (4,6-diamidino-2-phenylindole) in CITIFLUOR AF1 (Chem Lab,) antifade solution was used to counterstain the chromosomes.

Photographs were taken on a Nikon Eclipse N80*i* fluorescence microscope equipped with a DS-QiMc monochromatic camera (Nikon, Tokyo, Japan). Each metaphase was captured using four different filter sets; 1) UV-2E/C (excitation at 240-380; emission at 435-485 nm) for DAPI, 2) B-2E/C (excitation at 465-495; emission at 515-555 nm) for fluorescein and ATTO 488, 3) G-2E/C (excitation at 528-553; emission at 590-650 nm) for Alexa 594, ATTO 550 and ATTO 594 and 4) 31023 (excitation at 630-650 nm; emission at 665-695 nm) for ATTO 647N. The individual channels were pseudo-coloured to visualise the sites of probe hybridisation. The images were overlaid and further analyzed with Adobe Photoshop CS5 (Adobe Systems, San Jose, CA, USA) or NIS-Elements BR3.1 software (Nikon) using only cropping, and functions affecting the whole image equally.

5.3 Results

5.3.1 *In silico* comparative analysis of *E. guineensis* derived synthetic oligonucleotide probe pools (EgOligoFISH) in Arecaceae

In silico analysis was performed to assess the synteny of the EgOligoFISH sequences derived from *E. guineensis* genome with other three Arecaceae species; *E. oleifera* and *C. nucifera* from tribe *Cocoseae* as well as *P. dactylifera* from tribe *Phoenix*. Alignment of the 52,508 EgOligoFISH against the three species revealed the highest sequence similarity to *E. oleifera* with 64.4 % similarity followed by *C. nucifera* (15.5 %), *P. dactylifera-a* (5.2 %) and *P. dactylifera*-b (5.5 %) (Figure 5.2). Between the individual oligo library, QPAQUE library consistently showing the highest sequence similarity on all analyzed species followed by PPAQUE and OPAQUE. Notably, alignment of the 52,508 *E. guineensis* derived oligo sequences against *E. oleifera, C. nucifera,* and *P. dactylifera* genome sequence resulted with about 14 % of *E. guneensis* oligo-specific sequence.

Details of the *in silico* comparative mapping of EgOligoFISH across the analysed species as shown in Table 5.1. For *P. dactylifera*, only one reference genome was used (Dous *et al*., 2011) for comparison purposes. A total of 34, 35 and 32 scaffolds (with at least of 25 aligned oligos) from *E. oleifera*, *C. nucifera,* and *P. dactylifera* respectively showed a synteny block to a specific chromosome region of *E. guineensis*. For example, the 2,370 oligo sequence from Chromosome 2a (OPAQUE library) showed a block of synteny with scaffold o8_sc00769 (*E. oleifera* genome). A total of 1,247 of OPAQUE (derived from Chromosome 2a) were conserved on one continuous region of the scaffold o8_sc00769. Interestingly, the same oligos were also found conserved on a block region of specific scaffolds of *C. nucifera* genome (Scaffold 4237) and *P. dactylifera* genome (PDK30s793851). Example of the synteny as illustrated in Figure 5.3 shows the oligonucleotide sequences shared between the homoeologous scaffolds in the four Arecaceae species.

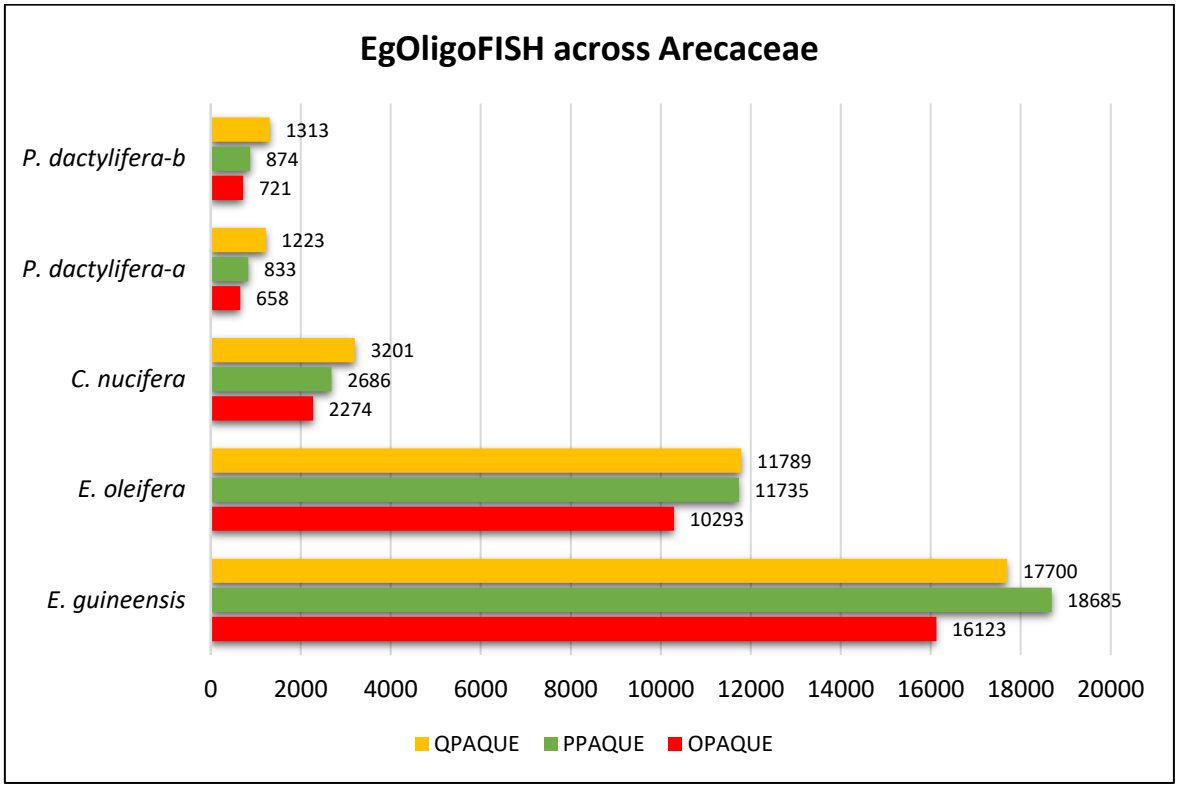


Figure 5.2 Overview of the EgOligoFISH contents in the *E. oleifera*, *C. nucifera* genome and *P. dactylifera* (a; Dous *et al*., 2011 and b; Al-Mssalem *et al*., 2013). Bars represent the number of EgOligoFISH sequences that is conserved in the respective species (all are present in *E. guineensis*). OPAQUE, red bar; PPAQUE, green bar; QPAQUE, yellow bar.

Table 5.1 Details of *in silico* analysis of EgOligoFISH probe sequence across *E. oleifera*, *C. nucifera* and *P. dactylifera*

EgOligoFISH library	Chromosome number	Total number of oligos	*Cocoseae*				*Phoenix*	
			Elaeis oleifera		*Cocos nucifera*		*Phoenix dactylifera (Dous et al. 2011)*	
			Total oligo mapped to *E. oleifera* assembly	Scaffolds with high-density oligos (≥ 25)	Total oligo mapped to *C. nucifera* assembly	Scaffolds with high-density oligos (≥ 25)	Total oligo mapped to *P. dactylifera* assembly	Scaffolds with high-density oligos (≥ 25)
OPAQUE	2a	2370	1518	o8_sc00769 o8_sc00319	418	Scaffold4237	82	PDK_30s793851
	8	1617	951	o8_sc00923 o8_sc04429 o8_sc04463	106	Scaffold1471	14	*nil*
	12i	2737	1499	o8_sc01293 o8_sc00339 o8_sc00744 o8_sc06300	282	Scaffold2102 Scaffold645	64	PDK_30s814661
	12ii	3801	2365	o8_sc00987 o8_sc00073 o8_sc01208	520	Scaffold2614 Scaffold3237 Scaffold177275	196	PDK_30s942751 PDK_30s762671 PDK_30s922111
	13	5598	3960	o8_sc00037 o8_sc00226 o8_sc21541	948	Scaffold2986 Scaffold8389 Scaffold2324 Scaffold912	302	PDK_30s767921 PDK_30s1116681 PDK_30s1045061 PDK_30s755911
PPAQUE	4i	2038	1449	o8_sc00021	299	Scaffold13836 Scaffold9737 Scaffold7518	69	*nil*
	4ii	3604	2037	o8_sc01117 o8_sc01875 o8_scoo372	544	Scaffold46	191	PDK_30s675901 PDK_30s696631 PDK_30s971241
	5	3114	1866	o8_sc00176	441	Scaffold736	89	PDK_30s65509269
	7a	2080	1326	o8_sc00064	353	Scaffold2219	118	PDK_30s1060371 PDK_30s65509167

Table 5.1 continue

OligoFISH library	Chromosome number	Total number of oligos	*Cocoseae*				*Phoenix*	
			Elaeis oleifera		*Cocos nucifera*		*Phoenix dactylifera (Dous et al. 2011)*	
			Total oligo mapped to *E. oleifera* assembly	Scaffolds with high-density oligos (≥ 25)	Total oligo mapped to *C. nucifera* assembly	Scaffolds with high-density oligos (≥ 25)	Total oligo mapped to *P. dactylifera* assembly	Scaffolds with high-density oligos (≥ 25)
PPAQUE	9	2943	1942	o8_sc00221	429	Scaffold4653 Scaffold10918 Scaffold7279 Scaffold3661	145	PDK_30s662861 PDK_30s918151
	10a	1983	1165	o8_scoo549	110	Scaffold1340	11	*nil*
	15	2923	1950	o8_sc00003	510	Scaffold12165	210	PDK_30s1031881 PDK_30s1063071 PDK_30s6550926 PDK_30s940621
QPAQUE	3i	1690	712	o8_sc00193 o8_sc00206	151	Scaffold2957	28	*nil*
	3ii	2666	1662	o8_sc00334 o8_sc01671	581	Scaffold3028 Scaffold2055	221	PDK_30s828471 PDK_30s809731
	5	2876	2157	o8_sc00016	601	Scaffold2026 Scaffold12388 Scaffold17361	210	PDK_30s763731 PDK_30s1033231
	6a	2904	1950	o8_sc00388	461	Scaffold10017	202	PDK_30s929081 PDK_30s745091
	7b	2365	1670	o8_sc00080	354	Scaffold18158 Scaffold103281	210	PDK_30s848901 PDK_30s841481
	10a	1375	694	o8_sc00824	105	Scaffold10987	29	*nil*
	11	3824	2944	o8_sc00031 o8_sc00219	948	Scaffold2239 Scaffold4085	323	PDK_30s928781 PDK_30s679881 PDK_30s758411

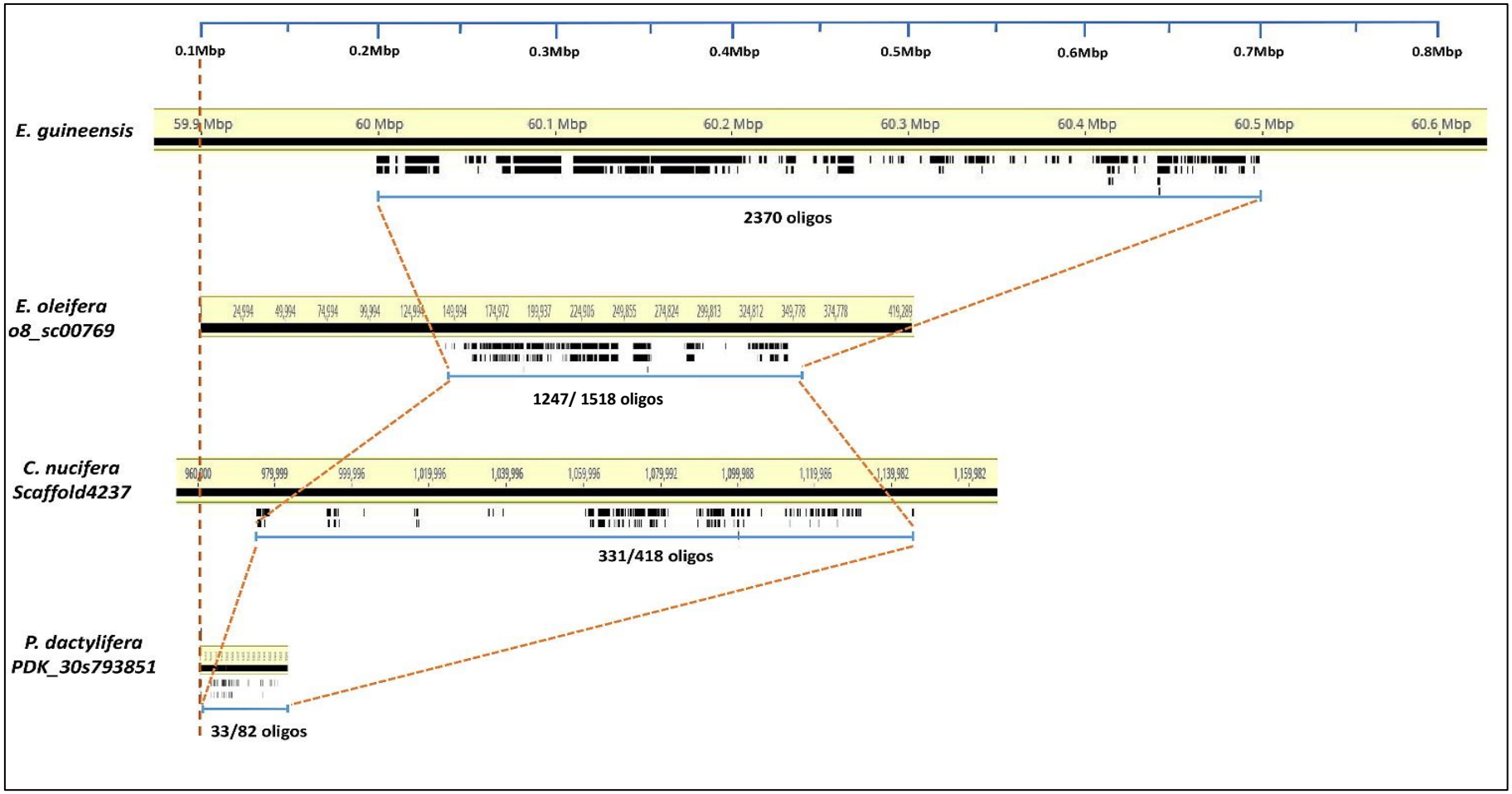


Figure 5.3 EgOligoFISH synteny in Arecaceae. Illustration showing an example of the oligonucleotide sequences shared between the homoeologous scaffolds in the four Arecaceae species (*E. guineensis*, *E. oleifera*, *C. nucifera,* and *P. dactylifera*)

5.3.3 Comparative *in situ* hybridisation of individual EgOligoFISH in Arecaceae

In parallel to the *in silico* analysis, comparative *in situ* hybridisation of the EgOligoFISH probe was performed on *E. oleifera*, *C. nucifera,* and *P. dactylifera* mitotic chromosome to assess the ability of the developed probe in distinguishing the physical chromosome of the analysed species respectively.

5.3.3.1 Karyotyping of *E. oleifera* chromosomes with EgOligoFISH

The significantly high similarity of EgOligoFISH sequence on the *E. oleifera* assembled genome (64.3%) with *in silico* analysis suggests the ability of EgOligoFISH probes in identifying *E. oleifera* chromosome. Therefore, physical validation of EgOligoFISH in distinguishing *E. oleifera* individual chromosome was assessed by hybridising individual OPAQUE, PPAQUE and QPAQUE probe on *E. oleifera* mitotic chromosomes (Figure 5.4).

Remarkably, all the *E. oleifera* chromosomes were successfully identified with clear and unique signals of EgOligoFISH probes including one large-size, one medium-size and one acrocentric chromosome without any hybridisation signal. The *in situ* hybridisation of the OPAQUE library on *E. guineensis* mitotic chromosomes distinguished five individual chromosomes with eight pair signals (Figure 5.4a; section 4.3.3.2 for details). In *E. oleifera*, nearly all of the chromosome's hybridisation pattern showed in *E. guineensis* was observed, except for the *E. oleifera* largest chromosome, which, interestingly showed only one pair signals on each of the chromosome arms (Figure 54a-i; boxed chromosome). In comparison to the *E. guineensis,* the largest chromosome of *the* species was identified with three pairs of OPAQUE hybridisation sites; two on the p-arm and one signal on the sub-terminal of the q-arm (Figure 5.4a; boxed chromosome). The signals observed on intercalary of p-arm and terminal of q-arm of *E. oleifera* is identical with two signals observed in *E. guineensis*. Dual hybridisation of the *E. oleifera* chromosome with 5S rDNA and OPAQUE probes showed the location of the 5s rDNA with a broad region signal on the short arm of the longest chromosome that has one pair of the OPAQUE signal. Alignment of oligo probe sequence of Chromosome 2 to *E. oleifera* genome assembly indicates a total of 852 oligo sequence was unique to *E. guineensis* (Chromosome 2), hence, speculated to explain the missing hybridisation region on *E. oleifera* largest chromosome.

In situ hybridisation of PPAQUE probe on *E. oleifera* chromosome showed an identical hybridisation pattern as observed in *E. guineensis*. Seven hybridisation sites were able to distinguish six *E. oleifera* individual chromosomes with some differences on the chromosomal location of the hybridised regions. For instance, a distinct hybridisation difference was observed for Chromosome 4 of *E. guineensis* in which a pair of hybridisation signal was observed from each chromosome arm (Figure 5.4b; boxed chromosome), however, in *E. oleifera*, both of the signals were observed on the same arm of *E. oleifera* medium-sized chromosome (Figure 5.4b-i; boxed chromosome).

The QPAQUE library (Figure 5.4c-i) distinguished six *E. oleifera* chromosomes with six pairs of hybridisation signals. Interestingly, unlike *E. guineensis*, no additional large-size chromosome was hybridized in *E. oleifera* (Figure 5.4c). Other than that, all of the hybridisation signals of QPAQUE library on *E. oleifera* chromosome resembling the *E. guineensis* hybridisation pattern.

A standard karyotype was developed for *E. oleifera* based on the FISH physical mapping of the individual EgOligoFISH probes on the *E. oleifera* mitotic chromosomes (Figure 5.5). The built ideograms are based on consistent oligo hybridisation on at least three metaphase chromosome spreads from two mitotic slides of the individual *in situ* hybridisation. The signals formed a bar code that uniquely labels the 16 *E. oleifera* chromosomes. This includes Chromosome 16 as the only nucleolar organizer region (NOR) chromosome in the *E. oleifera* genome which also carries an 18S rDNA (see Section 3.1) and one of the longest sub-metacentric and one small-size chromosome that was not hybridized with any of the probes. The co-hybridized signals of OPAQUE and QPAQUE on *E. guineensis* Chromosome 15 could not be observed in *E. oleifera* as there was no simultaneous *in situ* hybridisation of the three oligo probes was performed for the species. Nevertheless, since the additional similar hybridisation sites of the OPAQUE probe was observed in *E. oleifera*, the assignment chromosome 15 of *E. oleifera* was as same as *E. guineensis*. Moreover, based on the *in silico* analysis result (section 5.3.1), the *E. oleifera* scaffolds were assigned on the *E. oleifera* physical chromosome correspond to the *in situ* hybridisation signal.

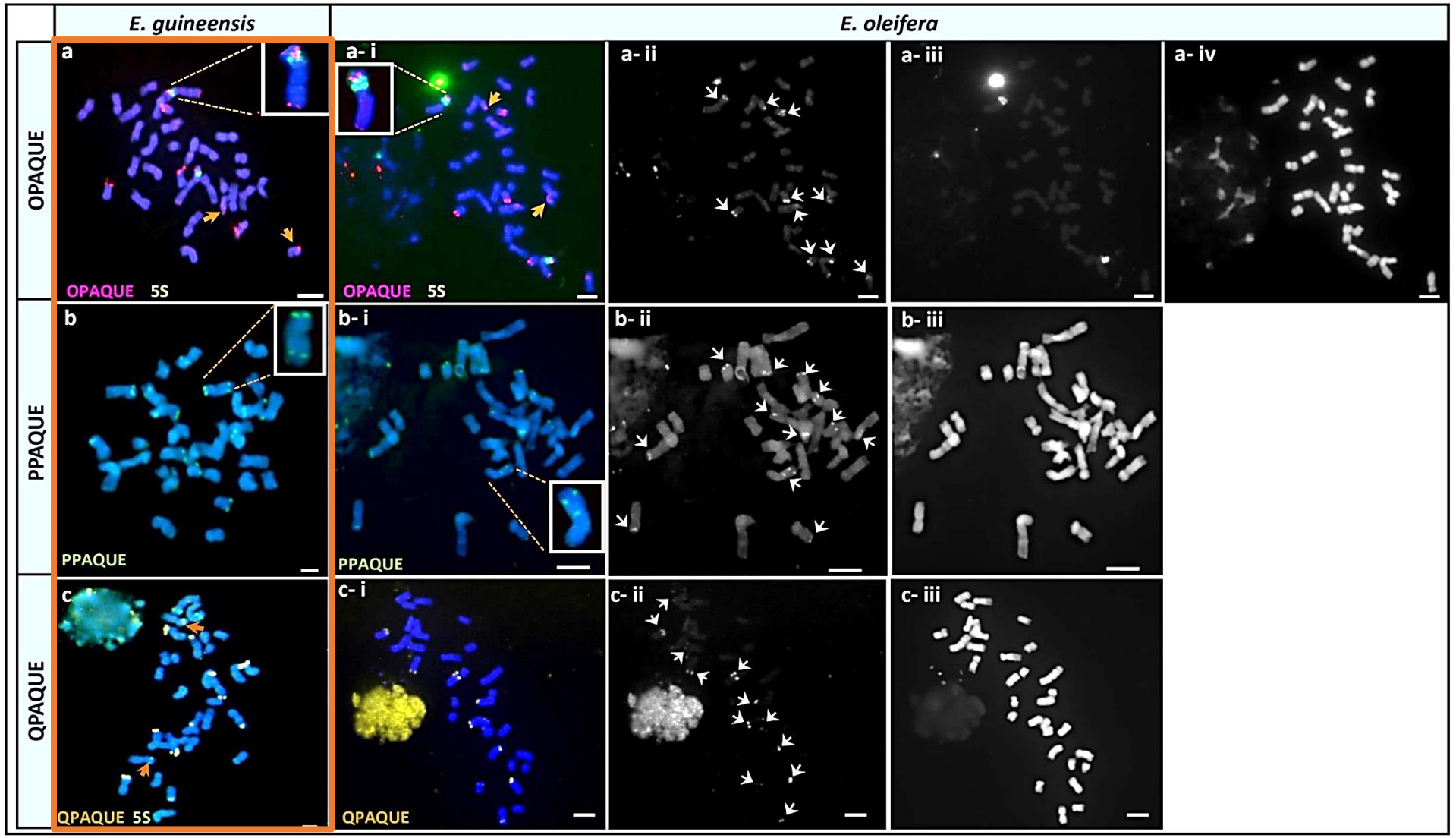


Figure 5.4 continue

Figure 5.4 continue

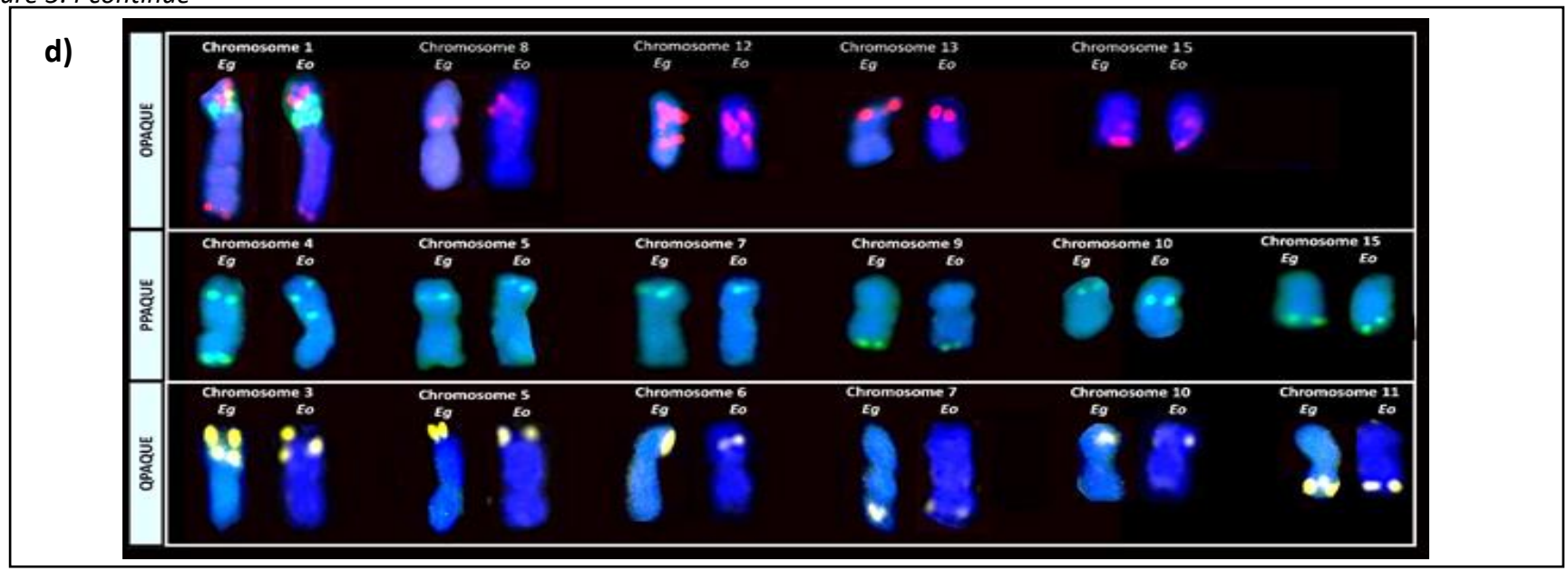


Figure 5.4 Comparative FISH mapping of *E. guineensis* (orange box) and *E. oleifera* with the developed pre-labelled massive oligo probe (EgOligoFISH). White arrow showing the individual *E. oleifera* chromosomes hybridised with EgOligoFISH (a-ii OPAQUE; b-ii PPAQUE and c-ii QPAQUE). a) *In situ* hybridisation of OPAQUE probe on *E. guineensis* chromosome and *E. oleifera* chromosome (a-i to a-iv). The orange arrow shows additional chromosome identified by OPAQUE on both *E. guineensis* (a) and *E. oleifera* (a-i). Boxed chromosome showed the large-size chromosome of both *E. guineensis* (a) and *E. oleifera* (b) with different hybridisation signal of OPAQUE (details in the text). b) *In situ* hybridisation of PPAQUE probe on *E. guineensis* chromosome and *E. oleifera* chromosome (b-i to b-iii). Boxed chromosome showed the enlarged chromosome of *E. guineensis* (a) and *E. oleifera* (b) with a distinct difference on the hybridisation signal. c) *In situ* hybridisation of QPAQUE probe on *E. guineensis* chromosome and *E. oleifera* chromosome (c-i to c-iii). Orange arrow showing additional hybridisation sites identified by QPAQUE probe on *E. guineensis* but absent in *E. oleifera* (5.4c and 5.4c-i). d) Nearly identical *in situ* hybridization signals observed between *E. guineensis* and *E. oleifera* chromosomes. *Eg; E. guineensis*, *Eo: E. oleifera*. Scale bar: 5µm

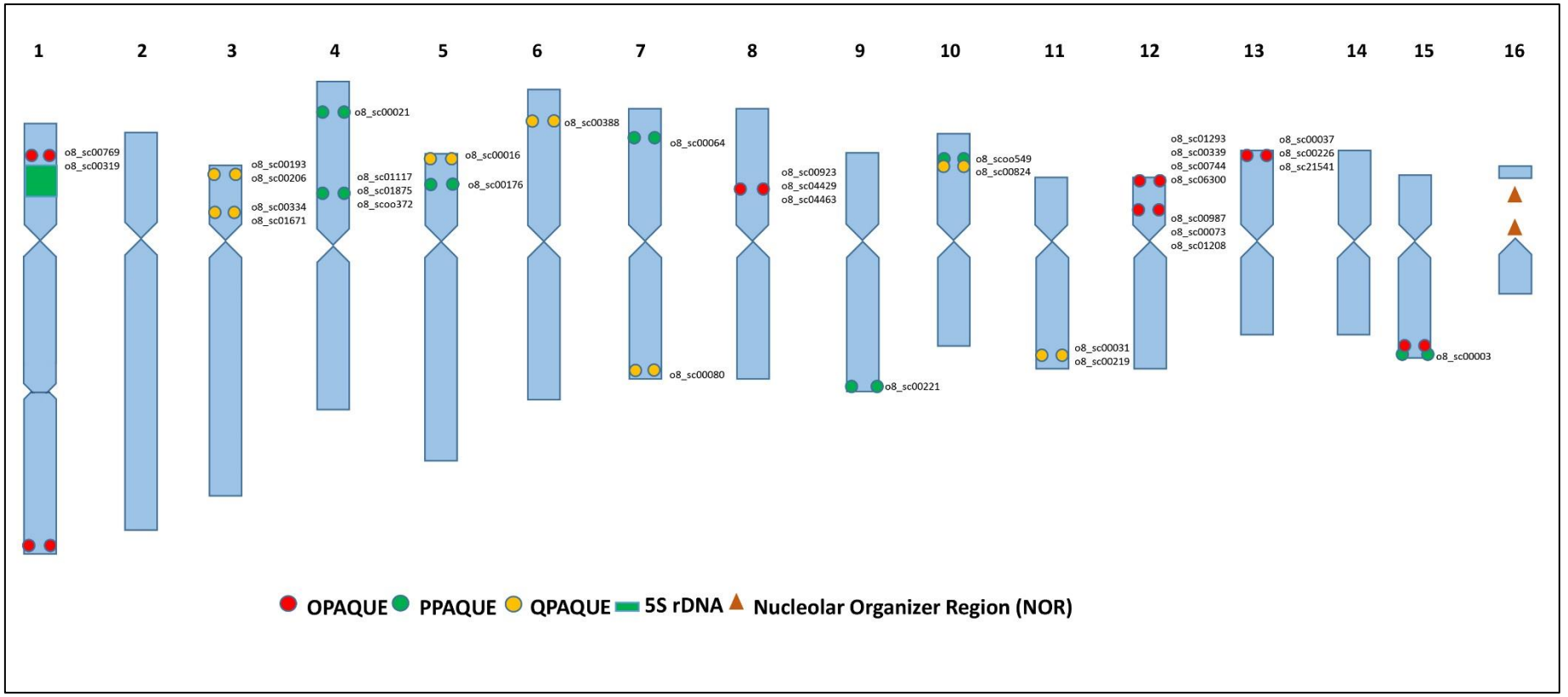


Figure 5.5 Identification of *E. oleifera* mitotic chromosomes with EgOligoFISH probes. Proposed standard karyotype of *E. oleifera* based on FISH mapping of EgOligoFISH. The assignment of the scaffolds on the karyotype is based on the *in silico* alignment of EgOligoFISH sequence against *E. oleifera* whole genome sequence (Table 5.1). Centromeric constrictions are drawn as a cross; secondary constriction (Chromosome 16) at the NOR as a gap; tertiary constriction (Chromosome 1) as a constriction.

5.3.2.2 Utility of EgOligoFISH probes in *Cocos nucifera* and *Phoenix dactylifera*

The EgOligoFISH probes were hybridized on *C. nucifera* and *P. dactylifera* mitotic chromosomes to assess the ability of the developed massive oligo probe in identifying chromosomes across the genus. The reliability of the observed signal was based on the observation of at least three metaphase cells on one FISH experiment. All the FISH experiment environment was based on the established FISH protocol for massive oligo probes that have been described in Chapter II (section 2.2.8) and Chapter IV (section 4.2.4). FISH was performed with the individual as well as simultaneous hybridisation (for *P. dactylifera*) with OPAQUE, PPAQUE and QPAQUE probes.

In *C. nucifera*, only *in situ* hybridisation with the QPAQUE probe consistently distinguished five pairs of the coconut palm chromosome (Figure 5.6). Compared to *E.guineensis*, two hybridisation sites of QPAQUE were absent in *C. nucifera* genome. All the hybridisation signals were observed on the terminal/sub-terminal domains of the chromosomes. The individual chromosomes distinguished by QPAQUE were assigned as Cn1-Cn5 based on the combination of the chromosome morphology and the pattern as well as the intensity of the signals displayed on the chromosome. The observed physical hybridisation was in agreement with low similarity (17.6%) of EgOligoFISH probes with *C. nucifera* genome from *in silico* analysis with QPAQUE probes giving the highest percentage of similarity among all three oligo probes (Figure 5.2). Furthermore, the differences in chromosomes and cytoplasm composition could also be considered to the lack of hybridisation other oligo probes on *C. nucifera* chromosome (see discussion).

The individual *in situ* hybridisation of *P. dactylifera* with OPAQUE probes was able to distinguish *P. dactylifera* individual chromosome but with inconsistent and indistinguishable signal distribution (Figure 5.7a). Twelve clear paired dot signals were observed on six pairs of the chromosome (5.7a-i), and also, a smudge of the OPAQUE signal also labelled some of the chromosomes. Similarly, the dispersed pattern of hybridisation signals of OPAQUE was observed in the interphase of *P. dactylifera* (5.6c-ii). Contrary to OPAQUE, FISH mapping of *P. dactylifera* with QPAQUE showed eight clear hybridisation

signals distinguished four pairs of *P. dactylifera* chromosomes (Figure 5.6b). With a combination of morphological features of the *P*. *dactylifera* chromosome and the intensity of the signals, the four paired chromosomes were identified as Pd1-Pd4. Hybridisation with PPAQUE probe was not able to observe on *P. dactylifera* metaphase chromosomes neither with individual nor simultaneous *in situ* hybridisation of EgOligoFISH probes (Figure 5.6c-i). Nevertheless, PPAQUE signals were observed on the interphase with simultaneous hybridisation of the probes (Figure 5.6c-ii). The ability of *in situ* hybridisation of the *E. guineensis* derived oligo to physically identified at least some of the *P. dactylifera* given the lower percentage of the similarity of EgOligoFISH (3.6-5.2%) when aligned to *P. dactylifera* genome sequence was interesting (Figure 5.1). Simultaneous FISH mapping of the three probes on the *P. dactylifera* interphase chromosome indicated the potential of EgOligoFISH derived from *E. guineensis* in further analysing genome constitution of date palm *via* fluorescent *in situ* hybridisation approach that could not be analysed only by in *silico* analysis.

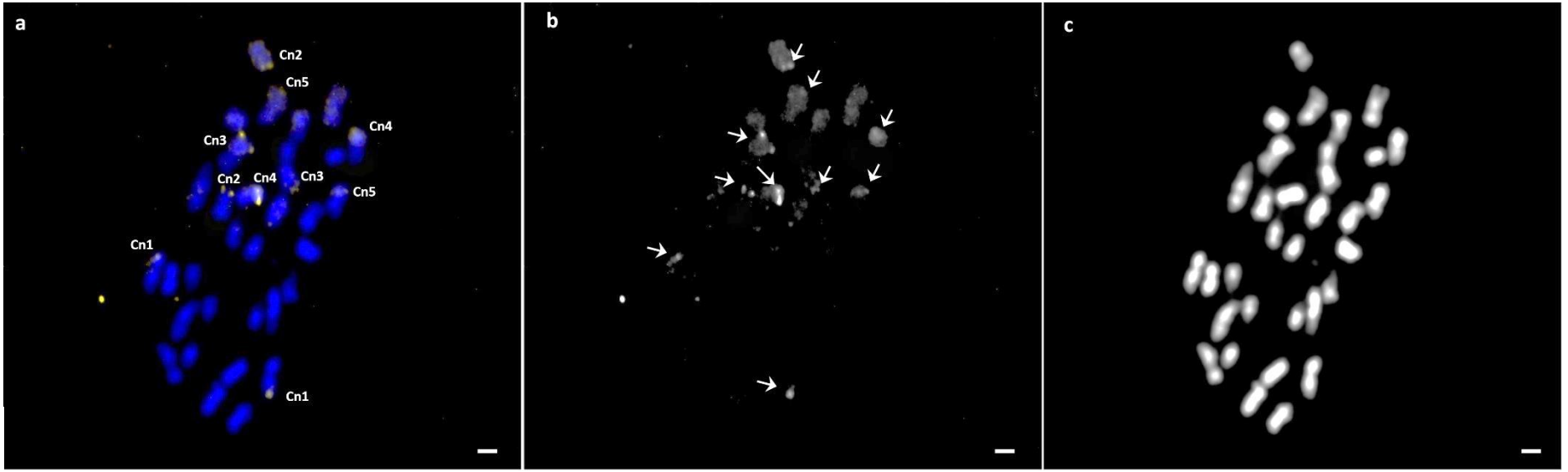


Figure 5.6 FISH mapping of EgOligoFISH on *C. nucifera* mitotic chromosome. a) *In situ* hybridisation of QPAQUE probe (yellow) on *P. dactylifera* chromosome (blue). Five pair chromosomes hybridized by QPAQUE probe were labelled as Cn1 –Cn5. b) White arrows are pointing on the OPAQUE probe signals hybridized on the *C. nucifera* chromosomes. c) 32 *C. nucifera* chromosomes. Bar 5µm.

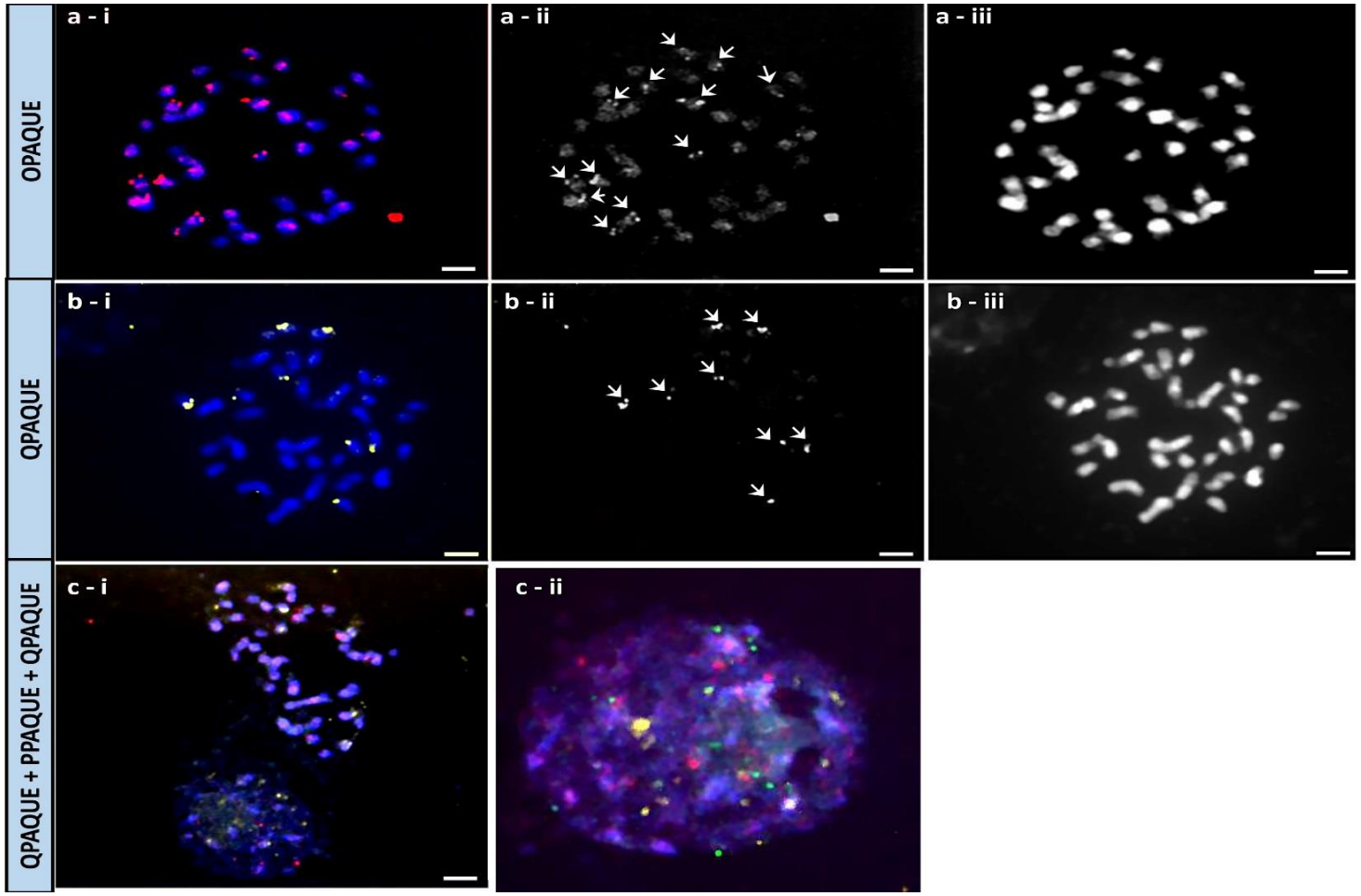


Figure 5.7 FISH mapping of EgOligoFISH on *P. dactylifera*. a) *In situ* hybridisation of OPAQUE (red) on *P. dactylifera* chromosome (blue). b) FISH of QPAQUE (yellow) on *P. dactylifera* chromosome (blue). a-iii and b-iii) 36 *P. dactylifera* chromosomes, c) Simultaneous *in situ* hybridisation of EgOligoFISH probe on *P. dactylifera*, metaphase (c-i) and interphase (c-ii). The OPAQUE signal displayed in red, PPAQUE signal displayed in green and QPAQUE displayed in yellow. White arrows are pointing on the signals hybridized on the metaphase chromosomes (a-ii and b-ii). Bar 5µm.

5.4 Discussion

5.4.1 Oligonucleotide pools establish *E. oleifera* standard karyotype

In silico analysis of massive oligo probes (EgOligoFISH) sequence performed in this study demonstrate a high level of sequence homology with *E. oleifera* genome (Figure 5.2, Table 5.1) supported by the ability of EgOligoFISH probes derived from *E. guineensis* in labelling *E. oleifera* chromosome (Figure 5.4). This suggests the potential of probes developed from the same genus in identifying individual chromosome of species that has been diverged more than 51 million years ago (MYA; Figure 5.1). Previous studies showed the ability of the oligo derived probes in distinguishing related species of *Cucumis* that have diverged 12 MYA (Han *et al*., 2015), potato and tomato from *Solanum* genus that have been diverged 7 MYA (Braz *et al*., 2018) and *Saccharum* which the divergence period among different species has been proposed to be less than 2 MYA (Meng *et al*., 2018).

Interestingly, the nearly identical strong hybridisation signals in *E. oleifera* chromosome compared to *E. guineensis* (Figure 5.4d) reflects the conserved sequence within these two *Elaeis* species on specific chromosome regions (Table 1, Figure 5.3). The integration of information obtained from *in silico*, as well as physical FISH-mapping of EgOligoFISH on *E. oleifera* mitotic chromosome, allows the establishment of *E. oleifera* standard karyotype for the first time. The proposed arrangement of the scaffolds on *E. oleifera* physical chromosome will benefit further refine assembly of *E. oleifera* in arranging the orientation of the assembled genome. Nevertheless, further investigation of the major variations of hybridisation sites observed on three of the *E. oleifera* chromosomes (putative Chromosome 1, 4 and 10) is essential to narrow down the potential explanation of such differences.

In the future, the extended design of massive oligo pools has potential in assisting in defining recombination landscape in meiotic OxG hybrid to further investigate lower natural fertility issues in the interspecific hybrid. Han *et al.* (2015) reported the potential of the massive single-copy oligo to be used in tracking chromosomes during early meiosis of cucumber. The same author suggested a combinational approach using FISH with bulked

oligo and immunolocalization of meiotic protein to further elucidate the relationship between recombination and homeologous pairing in species.

5.4.2 Massive oligo pools probe (EgOligoFISH) in distinguishing chromosome across-genus of Arecaceae

Use of *in situ* hybridisation probes across-genus is a powerful approach for chromosome and genome-wide comparison of the chromosome structures and constitution of different species. It identifies ancient syntenies shared by widely divergent species and clearly defines the chromosomal rearrangements among species. The syntenic regions indicate the likely structure of ancestral chromosomes, whereas the chromosome rearrangements indicate the dynamics and mechanisms of chromosome change during evolution (Rens *et al.,* 2006).

In this study, *in silico* analysis indicates a low homology with the conserved sequence present in regions spanning less than 500 kb regions within a limited number of scaffolds of both *C. nucifera* and *P. dactylifera* genomes (Figure 5.2, Table 5.1). Al-Mssalem *et al.,* (2013) and Matthew *et al.,* (2014) have described a significant macro-synteny/ long-range synteny between *E. guineensis* and *P. dactylifera*. Matthew *et al*., (2014) reported that most of the 18 date palm linkage group were syntenic with one of the 16 oil palm chromosomes. In the case of *C. nucifera*, most markers from each *C. nucifera* linkage group were found aligned to the same oil palm genome (Xiao *et al*., 2014). The shared sequence conserved in a block region of the analysed Arecaceae species (Figure 5.3, Table 5.1) that have been diverged more than 51 MYA suggest the potential of oligo pools-based probes derived *E. guineensis* as a source in preliminary genomics study of other Arecaceae under-studied species.

In plants, cross-species/genus chromosome identification using BAC (Bacterial Artificial Chromosome) based probes has been demonstrated for several taxa such as *Brassicaceae* (Xiong and Pires, 2011), Solanaceae (Szinay *et al.,* 2012) and *Brachypodium* (Hasterok *et al*., 2006, Lusinska *et al.,* 2018). Nevertheless, BAC-FISH based approach is challenging or partially successful in providing reproducible landmarks for the individual chromosome of a plant with large genome size. This mainly due to the abundance of the repeats in the

genome (Ma *et al*., 2010; Majka *et al*., 2017). Given our knowledge of the repeat distribution across *Elaeis* genomes (Castilho *et al*., 2000*,* Kubis *et* al., 2003; Singh *et al.,* 2013, Chapter III), and the similarity between the two species *E. guineensis* and *E.oleifera* (Table 5.1), neither genomic DNA nor BAC probes would be suitable to discriminate the genomes, chromosomes or chromosome segments using *in situ* painting in *Elaeis*.

The physical FISH mapping of EgOligoFISH in both *C. nucifera* and *P. dactylifera* compared to the *in silico* data, showed a smaller number of oligo sequences mapping in both genomes (Figure 5.2). The numbers of oligos from *E. guineensis* mapping to the three species was consistent with their evolutionary divergence. The oligo pools-based probe showed its ability in identifying four pairs of chromosomes in *P. dactylifera* (Pd1-Pd4) and five chromosomes in *C. nucifera* (Cn1-Cn5a). Moreover, the localized hybridized region on the interphases of *P. dactylifera* regardless of the inability of some probes to hybridize the mitotic chromosomes potentially due to technical reasons (see Section 5.4.3). The finding suggests the ability of *in situ* hybridisation of the *E. guineensis* derived oligo to physically identify at least some of the *P. dactylifera* and *C. nucifera* chromosome given the lower percentage of the similarity of EgOligoFISH in both genome sequence (Table 5.1). Simultaneous FISH mapping of the three probes on the interphase chromosome indicated FISH is the powerful approach in analysing, complementing and validating the genome structure and constitution of a species, that built based on the *in silico* analysis. As pointed out by Murphy *et al*. (2005), the resolution of synteny declines with increasing phylogenetic distance. Only 1–2% of the genome is actually transcribed, conservation of the sequences stretches to chromosome-specific noncoding sequences, indicating their functional significance and explains why chromosome probes of one species hybridize to DNA of a distantly diverged species (Murphy *et al*., 2005). The ability of EgOligoFISH in labelling chromosomes of other genera indicates the utility of the probe set to be used in other species of the same genus in *Cocoseae* as well as *Phoenix*.

The utility of the EgOligoFISH described in this chapter, suggesting its importance as interspecific or intergeneric markers to provide some useful genetic information for poorly studied related species, contributing to conservation, genetic assessment, and construction of linkage maps. The value has also been established in Brassicaceae species

(Lysak *et al*., 2005, 2006; Mandakova and Lysak 2008; Mandakova *et al*., 2010) where BAC-derived probes are appropriate; as well as in the first tests of oligonucleotide pool probes in the Solanaceae (Braz *et al*., 2018). Furthermore, this finding strongly indicates the utility of massive oligo probes developed from an established assembly to anchor physical chromosomes across the taxa showing a minimal similarity *via* informatics analysis (Figure 5.2, Table 5.1). Also, the combination of the *in silico* data and physical FISH mapping can assist in upgrading these scaffold-based genome assemblies (*E. oleifera, C. nucifera,* and *P. dactylifera*) to the chromosomally assembled genomes.

5.4.3 Robustness of oligonucleotide pools probes across taxa: from the technical perspective

The pools of synthetic oligonucleotides with c. 200,000 bp of homologous sequence to the reference sequence proved generally robust in the detection of homologous hybridisation sites on chromosomes of the various Arecaceae species (Figure 5.4 – Figure 5.7). Fluorescent *in situ* hybridisation (FISH) for visualization of nucleic acids was developed as an alternative to older methods with radiolabeled probes (Gall and Pardue, 1969). Over three decades, extensive work in *in situ* hybridisation has been optimized for various type of probes, e.g., labelled clones, repetitive sequences and total genomic DNA probes (reviewed in Schwarzacher and Heslop-Harrison 2000, Schwarzacher 2003, Jiang and Gill 2006, Figueroa and Bass 2010, Huber *et al*., 2018). Recently, the use of synthesized oligonucleotide probes built on a defined set of unique sequences chosen from available genome sequence have been became interests of a plant researchers (Meng *et al*., 2018; Hou *et al*., 2018; Braz *et al*., 2018; Qu *et al*., 2017; Han *et al*., 2015). With contrasting copy numbers (each probe only once in the target, always to one DNA strand), probe lengths (always 47 to 50bp) and other properties (GC content, single-copy chromosome regions rather than repetitive), chromosomal target denaturation, hybridisation, and washing conditions may benefit from optimization.

The *in situ* hybridisation has been established for oil palm (Madon *et al*., 1996, 1998) adapted from Heslop-Harrison (1991). In previous experiments, the ideal *in situ* hybridisation was established with repetitive DNA probes (Chapter III) and optimization with single/low copy probes (Chapter IV). The differences in the *in situ* hybridisation

patterns have been found between batches of probes, fixations, and chromosome preparations, relating to the cytoplasm, denaturation temperatures, and probe concentrations. Given that the synthetic oligo probes are defined in length, base composition labelling, and homology to the target, they might be expected to be much more robust and reproducible in their hybridisation compared to enzymatically labelled probes.

Preparation of a clean chromosome spreads is crucial in the *in situ* hybridisation as the non-specific signal can be deposited on cytoplasm and cell debris. The cytoplasm will further mask the chromosomes and restrict the access of probe and detection reagents (Schwarzacher 1989, Kato *et al*. 2011, Kirov 2014). In this study, two major parameters that give positive effect in getting the ideal free-cytoplasm condition of *Elaeis* metaphase chromosome are; 1) Enzymatic treatment for cell wall digestion and 2) Pepsin treatment conditions. Based on the *in situ* hybridisation consistency observed in both *C. nucifera* and *P. dactylifera,* even though some probes successfully hybridized on the chromosome, further optimization of the chromosome preparations is still required. The inconsistent quality of the metaphase chromosome preparations of both *C. nucifera* and *P. dactylifera* was speculated to be due to the thicker cytoplasm compared to *Elaeis* species. Also, the established pepsin treatment for *E. guineensis* was probably not ideal for both *C. nucifera* and *P. dactylifera*.

Re-probing of slides is valuable (Heslop- Harrison 1992, Schwarzacher 2000). As is well known, the extra washing and denaturation steps lead to deterioration of chromosome morphology and sometimes to the loss of whole chromosomes. It may be possible to improve techniques, perhaps by alteration of formamide concentrations in the hybridisation solution or reduction in wash temperatures, specifically for oligonucleotide hybridisation where it may be acceptable to retain previous probe signal.

The *in situ* hybridisation results, though, show that the labelled synthetic oligonucleotide pools are robust probes for *in situ* hybridisation. They give cross-species signal even in chromosome preparations that have not been fully optimized. As predicted from the *in silico* analysis, no adjustment of probe concentration nor hybridisation and washing

stringency was required to obtain the signal. Where comparisons are required across genera, in the future, it is likely that the probe design pipeline can be further optimized.

5.5 Conclusion

Combination of EgOligoFISH and other chromosome-labelling probes lays the foundation for future studies of the structure, organization, and evolution of genome in Arecaceae. The integration of information obtained from *in silico*, as well as physical FISH-mapping of EgOligoFISH on *E. oleifera* mitotic chromosomes, successfully establish *E. oleifera* standard karyotype for the first time. Furthermore, the utility of the EgOligoFISH described in this chapter suggests its importance as interspecific or intergeneric markers in providing some useful genetic information for poorly studied related species in Arecaceae. Nevertheless, there is more to gain from the comparative *in situ* study if optimization of the coconut (*C. nucifera)* and date palm (*P. dactylifera*) chromosome preparations are considered in future work.

CHAPTER VI

Summary and prospects of the study

Knowledge of the structures and organization of the chromosomes is valuable for the development of new lines and creating hybrids. Moreover, the information also beneficial in understanding the grounds of some abnormalities or infertility and characterizing differences between related species or even breeding lines. Hence, the development of a robust, reliable and easy to use markers and techniques for chromosome identification are crucial for such studies.

As for *E. guineensis,* little is known on the physical structure of species apart from morphologically identified 16 pairs chromosome with broad size-clustering characterization (Sharma and Sarkar, 1956; Madon *et al*., 1998) and localization of repetitive DNA on a specific part of a chromosome from prior studies (Castilho *et al*., 2000; Kubis *et al*., 2003). However, neither analysing high volumes of DNA sequence (a technology unavailable at the time of the cited publications) nor improvements to *in situ* hybridization protocols and microscopy, did not identify specific repetitive or low-copy DNA sequences that could identify all chromosome arms (Aim 1, Chapter I).

This study has met the designated challenge/Aim 2 (Chapter I) by establishing the first *E. guineensis* reference karyotype with a combination of physical FISH-mapping of repetitive DNA and massive pools of synthetic oligonucleotides from single copy sequence (EgOligoFISH). Here, a reference karyotype for 16 pair of *E. guineensis* with a combination of repetitive DNA (*copia*-like Eg9CEN, telomere, 5S rDNA, 18S rRNA) and massive oligonucleotide pools is proposed (Aim 2, Chapter I; Chapter III; Chapter IV; Figure 6.1). Remarkably, the conserved physical localization of the oligo derived *E. guineensis* on another *Elaeis* species, *E. oleifera* metaphase chromosomes is informative. The *E. oleifera* proposed FISH-karyotype (Figure 5.5; Chapter V) with scaffolds assignment will benefit in positioning and refining the assembly of the *E. oleifera* draft genome (Singh *et al*., 2013) at the chromosome level.

Information gathered from repetitive analyses of unassembled raw sequence data in this study showed no newly identified major repetitive DNA class compared to known repetitive in *E. guineensis* from random cloning, PCR, restriction digestions, and cytogenetic analysis. Specific analyses carried out here revealed that *E. guineensis* lack of tandem repeat compared to other monocots supported the findings reported by Castilho *et al*., 2000. (Chapter III). Nevertheless, the superior quality of *E. guineensis* chromosome preparation obtained in this study allows the physical identification of a tertiary constriction on one pair of secondary constriction chromosomes that has a site for 18S rRNA and NOR region. The newly identified tertiary constriction on the long arm of the largest chromosome showed that having a superior quality of the chromosome preparation is significant as it confirms the assembled genome. Re-visiting the existence of tertiary constriction of the *E. guineensis* largest chromosome with minor hybridisation site *copia*-like putative centromeric sequence that having similarity with the date palm sex-determination region is interesting. The work can be extended to investigate the proposed Robertsonian fusion by designing oligo flanking the sex determination region sequence from date palm and further hybridized on the oil palm chromosomes.

The combination of EgOligoFISH and other chromosome-labeling probes lays the foundation for future studies of the structure, organization, and evolution of genome in Arecaceae (Aim3 and Aim 4, Chapter I). The collection of massive oligo FISH probes developed in this study able to narrow down the specific genes on the physical chromosomes as some of it specifically developed from QTL linked regions. The utility of the EgOligoFISH in *Cocos nucifera* and *Phoenix dactylifera*, suggests its importance as interspecific or intergeneric markers in providing some useful genetic information for poorly studied related species in Arecaceae. Nevertheless, more to gain from the comparative *in situ* study if further minor optimization of the coconut (*C. nucifera)* and date palm (*P. dactylifera*) chromosome preparations considered in future work.

Development of chromosome-specific markers enriched with genes is beneficial in facilitating the introgression of beneficial species for crop improvement. Wild relatives of crops known to have valuable traits for crop improvement that can be introgressed into

crops by crossing the crops with the wild species and breeding the backcross progenies. In Malaysia for example, an extensive oil palm germplasm collections of *E. guineensis* (Nigeria, Cameroon, Zaire, Tanzania, Madagascar, Angola, Senegal, Gambia, Sierra Leone, Guinea and Ghana) and *E. oleifera* (Honduras, Panama, Costa Rica, Suriname, Ecuador, Peru and Columbia) has been carried out with intention to broaden the genetic base of current breeding materials. By using the *E. guineensis* and *E. oleifera* germplasms, the OxG hybrids and backcrosses can be created to obtain a good combination of attractive traits including various agronomic traits and resistance to biotic and abiotic stresses (Chapter V). However, these efforts can take a long time and involve multiple resources due to long breeding cycle, large planting areas, intensive labour, and high maintenance cost. Hence, the transferability of the oligo sequence information between both *Elaeis* species as well as physical FISH-mapping chromosome information will further assist in a hybrid of *E. oleifera* and *E. guineensis* (OxG). Knowledge of the structures and organization of the *Elaeis guineensis* chromosomes, as in any economically important crop, is valuable for development of new lines, allowing identification of translocations, duplications and inversions. With knowledge of causes of some abnormalities or infertility, characterization of differences between related species or breeding lines, and improvement of genome assemblies at the chromosome level, elite lines including hybrids with another species, *E. oleifera* (American oil palm), can be selected more efficiently.

In conclusion, with the robust oligo-FISH probes and established molecular cytogenetic techniques of developed here combined with the advent of molecular markers will be beneficial in facilitating oil palm traditional breeding.

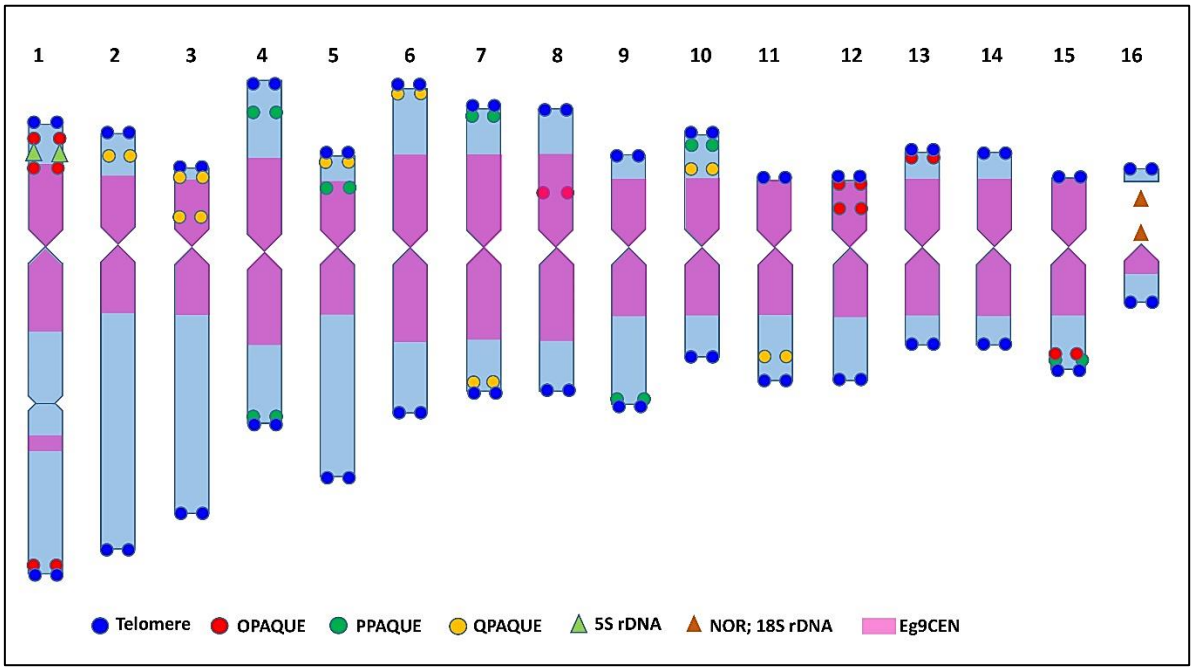


Figure 6.1 The genome landscape of oil palm (*Elaeis guineensis* Jacq). Propose standard FISH-karyotype of *E. guineensis* with repetitive DNA and single-copy DNA sequence. Centromeric constrictions are drawn as a cross; secondary constriction (Chromosome 16) at the NOR as a gap; tertiary constriction (Chromosome 1) as a constriction.

References

ADAM, H., COLLIN, M., TREGEAR, J. W., RICHAUD, F., BEULÉ, T., CROS, D., OMORÉ, A., NODICHAO, L. & NOUY, B. 2011. Environmental regulation of sex determination in oil palm: current knowledge and insights from other species. *Annals of Botany,* 108**,** 1529-1537.

ADAWY, S., ATIA, M. & EL-ITRIBY, H. 2015. Sex-differentiation based on fluorescence *in situ* hybridization (FISH) with 5s and 45s rDNA of Egyptian date palm trees. *Int J Adv Biotechnol Res* 6.

AKKAK, A., SCARIOT, V., TORELLO MARINONI, D., BOCCACCI, P., BELTRAMO, C. & BOTTA, R. 2009. Development and evaluation of microsatellite markers in *Phoenix dactylifera* L. and their transferability to other *Phoenix* species. *Biologia Plantarum,* 53**,** 164-166.

ALBERT, P. S., ZHANG, T., SEMRAU, K., ROUILLARD, J.-M., KAO, Y.-H., WANG, C.-J. R., DANILOVA, T. V., JIANG, J. & BIRCHLER, J. A. 2019. Whole-chromosome paints in maize reveal rearrangements, nuclear domains, and chromosomal relationships. *Proceedings of the National Academy of Sciences,* 116**,** 1679-1685.

ALISAWI, O. N. K. 2019. Virus integration and tandem repeats in the genomes of *Petunia.* PhD Thesis, University of Leicester.

ALIX, K., GÉRARD, P. R., HESLOP-HARRISON, J. S. & SCHWARZACHER, T. 2017. Polyploidy and interspecific hybridization: partners for adaptation, speciation and evolution in plants. *Annals of Botany,* 120**,** 183-194.

AL-DOUS, E. K., GEORGE, B., AL-MAHMOUD, M. E., AL-JABER, M. Y., WANG, H., SALAMEH, Y. M., AL-AZWANI, E. K., CHALUVADI, S., PONTAROLI, A. C., DEBARRY, J., ARONDEL, V., OHLROGGE, J., SAIE, I. J., SULIMAN-ELMEER, K. M., BENNETZEN, J. L., KRUEGGER, R. R. & MALEK, J. A. 2011. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology,* 29**,** 521.

AL-MSSALLEM, I. S., HU, S., ZHANG, X., LIN, Q., LIU, W., TAN, J., YU, X., LIU, J., PAN, L., ZHANG, T., YIN, Y., XIN, C., WU, H., ZHANG, G., BA ABDULLAH, M. M., HUANG, D., FANG, Y., ALNAKHLI, Y. O., JIA, S., YIN, A., ALHUZIMI, E. M., ALSAIHATI, B. A., AL-OWAYYED, S. A., ZHAO, D., ZHANG, S., AL-OTAIBI, N. A., SUN, G., MAJRASHI, M. A., LI, F., TALA, WANG, J., YUN, Q., ALNASSAR, N. A., WANG, L., YANG, M., AL-JELAIFY, R. F., LIU, K., GAO, S., CHEN, K., ALKHALDI, S. R., LIU, G., ZHANG, M., GUO, H. & YU, J. 2013. Genome sequence of the date palm *Phoenix dactylifera* L. *Nature Communications,* 4**,** 2274.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *J Mol Biol,* 215**,** 403-10.

ANANIEV, E. V., PHILLIPS, R. L. & RINES, H. W. 1998. Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. *Proceedings of the National Academy of Sciences,* 95**,** 13073-13078.

BAILEY, J. P. & STACE, C. A. 1992. Chromosome banding and pairing behaviour in *Festuca* and *Vulpia* (Poaceae, Pooideae). *Plant Systematics and Evolution,* 182**,** 21-28.

BAČOVSKÝ, V., HOBZA, R. & VYSKOT, B. 2018. Technical Review: Cytogenetic Tools for Studying Mitotic Chromosomes. *In:* BEMER, M. & BAROUX, C. (eds.) *Plant Chromatin Dynamics: Methods and Protocols.* New York, NY: Springer New York.

BAKER, W. J., CLARKSON, J. J., CHASE, M. W., WILMOT, T., NORUP, M. V., SAVOLAINEN, V., COUVREUR, T. L. P., DOWE, J. L., LEWIS, C. E. & PINTAUD, J.-C. 2011. Phylogenetic relationships among arecoid palms (Arecaceae: Arecoideae). *Annals of Botany,* 108**,** 1417-1432.

BAKOUMÉ, C., WICKNESWARI, R., SIJU, S., RAJANAIDU, N., KUSHAIRI, A. & BILLOTTE, N. 2015. Genetic diversity of the world's largest oil palm (*Elaeis guineensis* Jacq.) field genebank accessions using microsatellite markers. *Genetic Resources and Crop Evolution,* 62**,** 349-360.

BALICK, M. & BECK, H. 1990. Useful palms of the world – a synoptic bibliography. New York: Columbia University Press.

BAO, W., KOJIMA, K. K. & KOHANY, O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA,* 6**,** 11.

BAO, Z. & EDDY, S. R. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res,* 12**,** 1269-76.

BARCELOS, E., AMBLARD, P., BERTHAUD, J. & SEGUIN, M. 2002. Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. *Pesquisa Agropecuária Brasileira,* 37**,** 1105-1114.

BARCELOS, E., RIOS, S. D. A., CUNHA, R. N. V., LOPES, R., MOTOIKE, S. Y., BABIYCHUK, E., SKIRYCZ, A. & KUSHNIR, S. 2015. Oil palm natural diversity and the potential for yield improvement. *Frontiers in Plant Science,* 6.

BARGHINI, E., NATALI, L., GIORDANI, T., CAVALLINI, A., COSSU, R. M., PINDO, M., VELASCO, R., CATTONARO, F., SCALABRIN, S. & MORGANTE, M. 2014. The Peculiar Landscape of Repetitive Sequences in the Olive (*Olea europaea* L.) Genome. *Genome Biology and Evolution,* 6**,** 776-791.

BARROW, S. C. 1998. A monograph of *Phoenix* L. (Palmae: Coryphoideae). *Kew Bulletin,* 53**,** 513-575.

BAYER, P. E., HURGOBIN, B., GOLICZ, A. A., CHAN, C.-K. K., YUAN, Y., LEE, H., RENTON, M., MENG, J., LI, R., LONG, Y., ZOU, J., BANCROFT, I., CHALHOUB, B., KING, G. J., BATLEY, J. & EDWARDS, D. 2017. Assembly and comparison of two closely related *Brassica napus* genomes. *Plant Biotechnology Journal,* 15**,** 1602-1610.

BEIRNAERT, A. D. S. F. O. & VANDERWEYEN, R. 1941. *Contribution à l'étude genetique et biometrique des variétés d'Elaeis Guineensis Jacquin,* [Nairobi], [East African Standard].

BELIVEAU, B. J., BOETTIGER, A. N., NIR, G., BINTU, B., YIN, P., ZHUANG, X. & WU, C. T. 2017. *In Situ* Super-Resolution Imaging of Genomic DNA with OligoSTORM and OligoDNA-PAINT. *Methods in molecular biology (Clifton, N.J.),* 1663**,** 231-252.

BELIVEAU, B. J., BOETTIGER, A. N., AVENDAÑO, M. S., JUNGMANN, R., MCCOLE, R. B., JOYCE, E. F., KIM-KISELAK, C., BANTIGNIES, F., FONSEKA, C. Y., ERCEG, J., HANNAN, M. A., HOANG, H. G., COLOGNORI, D., LEE, J. T., SHIH, W. M., YIN, P., ZHUANG, X. & WU, C.-T. 2015. Single-molecule super-resolution imaging of chromosomes and *in situ* haplotype visualization using Oligopaint FISH probes. *Nature Communications,* 6**,** 7147.

BELIVEAU, B. J., JOYCE, E. F., APOSTOLOPOULOS, N., YILMAZ, F., FONSEKA, C. Y., MCCOLE, R. B., CHANG, Y., LI, J. B., SENARATNE, T. N., WILLIAMS, B. R., ROUILLARD, J.-M. &

WU, C.-T. 2012. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proceedings of the National Academy of Sciences,* 109**,** 21301-21306.

BENNETZEN, J. L. & PARK, M. 2018. Distinguishing friends, foes, and freeloaders in giant genomes. *Current Opinion in Genetics & Development,* 49**,** 49-55.

BENNETZEN, J. L. & WANG, H. 2014. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology,* 65**,** 505-530.

BERTIOLI, D. J., CANNON, S. B., FROENICKE, L., HUANG, G., FARMER, A. D., CANNON, E. K., LIU, X., GAO, D., CLEVENGER, J., DASH, S., REN, L., MORETZSOHN, M. C., SHIRASAWA, K., HUANG, W., VIDIGAL, B., ABERNATHY, B., CHU, Y., NIEDERHUTH, C. E., UMALE, P., ARAUJO, A. C., KOZIK, A., KIM, K. D., BUROW, M. D., VARSHNEY, R. K., WANG, X., ZHANG, X., BARKLEY, N., GUIMARAES, P. M., ISOBE, S., GUO, B., LIAO, B., STALKER, H. T., SCHMITZ, R. J., SCHEFFLER, B. E., LEAL-BERTIOLI, S. C., XUN, X., JACKSON, S. A., MICHELMORE, R. & OZIAS-AKINS, P. 2016. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat Genet,* 48**,** 438-46.

BETEKHTIN, A., JENKINS, G. & HASTEROK, R. 2014. Reconstructing the Evolution of Brachypodium Genomes Using Comparative Chromosome Painting. *PLoS One,* 9**,** e115108.

BEULÉ, T., AGBESSI, M. D. T., DUSSERT, S., JALIGOT, E. & GUYOT, R. 2015. Genome-wide analysis of LTR-retrotransposons in oil palm. *BMC Genomics,* 16**,** 795.

BILLOTTE, N., MARSEILLAC, N., BROTTIER, P., NOYER, J.-L., JACQUEMOUD-COLLET, J.-P., MOREAU, C., COUVREUR, T., CHEVALLIER, M.-H., PINTAUD, J.-C. & RISTERUCCI, A.-M. 2004. Nuclear microsatellite markers for the date palm (*Phoenix dactylifera* L.): characterization and utility across the genus *Phoenix* and in other palm genera. *Molecular Ecology Notes,* 4**,** 256-258.

BISCOTTI, M. A., OLMO, E. & HESLOP-HARRISON, J. S. 2015. Repetitive DNA in eukaryotic genomes. *Chromosome Res,* 23**,** 415-20.

BLANK, S. D. 1952. A reconnaissance of the American oil palm. *Trop Agric [Trinidad],* 29**,** 90-101.

BOMBARELY, A., MOSER, M., AMRAD, A., BAO, M., BAPAUME, L., BARRY, C. S., BLIEK, M., BOERSMA, M. R., BORGHI, L., BRUGGMANN, R., BUCHER, M., D'AGOSTINO, N., DAVIES, K., DRUEGE, U., DUDAREVA, N., EGEA-CORTINES, M., DELLEDONNE, M., FERNANDEZ-POZO, N., FRANKEN, P., GRANDONT, L., HESLOP-HARRISON, J. S., HINTZSCHE, J., JOHNS, M., KOES, R., LV, X., LYONS, E., MALLA, D., MARTINOIA, E., MATTSON, N. S., MOREL, P., MUELLER, L. A., MUHLEMANN, J., NOURI, E., PASSERI, V., PEZZOTTI, M., QI, Q., REINHARDT, D., RICH, M., RICHERT-PÖGGELER, K. R., ROBBINS, T. P., SCHATZ, M. C., SCHRANZ, M. E., SCHUURINK, R. C., SCHWARZACHER, T., SPELT, K., TANG, H., URBANUS, S. L., VANDENBUSSCHE, M., VIJVERBERG, K., VILLARINO, G. H., WARNER, R. M., WEISS, J., YUE, Z., ZETHOF, J., QUATTROCCHIO, F., SIMS, T. L. & KUHLEMEIER, C. 2016. Insight into the evolution of the Solanaceae from the parental genomes of Petunia hybrida. *Nature Plants,* 2**,** 16074.

BOYLE, S., RODESCH, M. J., HALVENSLEBEN, H. A., JEDDELOH, J. A. & BICKMORE, W. A. 2011. Fluorescence *in situ* hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis. *Chromosome Research,* 19**,** 901-909.

BRANDES, A., HESLOP-HARRISON, J. S. & FRIESEN, N. 2001. Diversity, Origin, and Distribution of Retrotransposons (gypsy and copia) in Conifers. *Molecular Biology and Evolution,* 18**,** 1176-1188.

BRANDES, A., THOMPSON, H., DEAN, C. & HESLOP-HARRISON, J. S. 1997. Multiple repetitive DNA sequences in the paracentromeric regions of *Arabidopsis thaliana* L. *Chromosome Research,* 5**,** 238-246.

BRAZ, G. T., HE, L., ZHAO, H., ZHANG, T., SEMRAU, K., ROUILLARD, J. M., TORRES, G. A. & JIANG, J. 2018. Comparative Oligo-FISH Mapping: An Efficient and Powerful Methodology To Reveal Karyotypic and Chromosomal Evolution. *Genetics,* 208**,** 513-523.

BURTON, J. N., ADEY, A., PATWARDHAN, R. P., QIU, R., KITZMAN, J. O. & SHENDURE, J. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology,* 31**,** 1119.

CAFASSO, D. & CHINALI, G. 2014. An ancient satellite DNA has maintained repetitive units of the original structure in most species of the living fossil plant genus Zamia. *Genome,* 57**,** 125-35.

CAMILLO, J., LEÃO, A. P., ALVES, A. A., FORMIGHIERI, E. F., AZEVEDO, A. L., NUNES, J. D., DE CAPDEVILLE, G., DE A MATTOS, J. K. & SOUZA, M. T., JR. 2014. Reassessment of the Genome Size in *Elaeis guineensis* and *Elaeis oleifera*, and Its Interspecific Hybrid. *Genomics insights,* 7**,** 13-22.

CARVALHO, A., MARTÍN, A., HESLOP-HARRISON, J. S., GUEDES-PINTO, H. & LIMA-BRITO, J. 2009. Identification of the spontaneous 7BS/7RL intergenomic translocation in one F1 multigeneric hybrid from the *Triticeae* tribe. *Plant Breeding,* 128**,** 105-108.

CASTILHO, A., VERSHININ, A. & HESLOP-HARRISON, J. S. 2000. Repetitive DNA and the Chromosomes in the Genome of Oil Palm (*Elaeis guineensis*). *Annals of Botany,* 85**,** 837-844.

CERMAK, T., KUBAT, Z., HOBZA, R., KOBLIZKOVA, A., WIDMER, A., MACAS, J., VYSKOT, B. & KEJNOVSKY, E. 2008. Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. *Chromosome Res,* 16**,** 961-76.

CHALHOUB, B., DENOEUD, F., LIU, S., PARKIN, I. A. P., TANG, H., WANG, X., CHIQUET, J., BELCRAM, H., TONG, C., SAMANS, B., CORRÉA, M., DA SILVA, C., JUST, J., FALENTIN, C., KOH, C. S., LE CLAINCHE, I., BERNARD, M., BENTO, P., NOEL, B., LABADIE, K., ALBERTI, A., CHARLES, M., ARNAUD, D., GUO, H., DAVIAUD, C., ALAMERY, S., JABBARI, K., ZHAO, M., EDGER, P. P., CHELAIFA, H., TACK, D., LASSALLE, G., MESTIRI, I., SCHNEL, N., LE PASLIER, M.-C., FAN, G., RENAULT, V., BAYER, P. E., GOLICZ, A. A., MANOLI, S., LEE, T.-H., THI, V. H. D., CHALABI, S., HU, Q., FAN, C., TOLLENAERE, R., LU, Y., BATTAIL, C., SHEN, J., SIDEBOTTOM, C. H. D., WANG, X., CANAGUIER, A., CHAUVEAU, A., BÉRARD, A., DENIOT, G., GUAN, M., LIU, Z., SUN, F., LIM, Y. P., LYONS, E., TOWN, C. D., BANCROFT, I., WANG, X., MENG, J., MA, J., PIRES, J. C., KING, G. J., BRUNEL, D., DELOURME, R., RENARD, M., AURY, J.-M., ADAMS, K. L., BATLEY, J., SNOWDON, R. J., TOST, J., EDWARDS, D., ZHOU, Y., HUA, W., SHARPE, A. G., PATERSON, A. H., GUAN, C. & WINCKER, P. 2014. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. *Science,* 345**,** 950-953.

CHAVES, R., ADEGA, F., HESLOP-HARRISON, J. S., GUEDES-PINTO, H. & WIENBERG, J. 2003. Complex satellite DNA reshuffling in the polymorphic t(1;29) Robertsonian translocation and evolutionarily derived chromosomes in cattle. *Chromosome Research,* 11**,** 641-648.

CHAVES, R., GUEDES-PINTO, H. & HESLOP-HARRISON, J. S. 2005. Phylogenetic relationships and the primitive X chromosome inferred from chromosomal and satellite DNA analysis in Bovidae. *Proceedings. Biological sciences,* 272**,** 2009-2016.

CHENG, Z., DONG, F., LANGDON, T., OUYANG, S., BUELL, C. R., GU, M., BLATTNER, F. R. & JIANG, J. 2002. Functional Rice Centromeres Are Marked by a Satellite Repeat and a Centromere-Specific Retrotransposon. *The Plant Cell,* 14**,** 1691-1704.

CHENG, Z.-J. & MURATA, M. 2003. A centromeric tandem repeat family originating from a part of Ty3/gypsy-retroelement in wheat and its relatives. *Genetics,* 164**,** 665-672.

CHESTER, M., RILEY, R. K., SOLTIS, P. S. & SOLTIS, D. E. 2015. Patterns of chromosomal variation in natural populations of the neoallotetraploid T*ragopogon mirus* (Asteraceae). *Heredity,* 114**,** 309-317.

CHIATANTE, G., GIANNUZZI, G., CALABRESE, F. M., EICHLER, E. E. & VENTURA, M. 2017. Centromere Destiny in Dicentric Chromosomes: New Insights from the Evolution of Human Chromosome 2 Ancestral Centromeric Region. *Molecular biology and evolution,* 34**,** 1669-1681.

COMAI, L., MAHESHWARI, S. & MARIMUTHU, M. P. A. 2017. Plant centromeres. *Current Opinion in Plant Biology,* 36**,** 158-167.

CONTENTO, A., HESLOP-HARRISON, J. S. & SCHWARZACHER, T. 2005. Diversity of a major repetitive DNA sequence in diploid and polyploid Triticeae. *Cytogenet Genome Res,* 109**,** 34-42.

COCHARD, B., AMBLARD, P. & DURAND-GASSELIN, T. 2005. Oil palm genetic improvement and sustainable development. *OCL,* 12**,** 141-147.

CORLEY, R. H. V. & TINKER, P. B. H. 2015. *The Oil Palm,* New York, John Wiley & Sons.

CORLEY, R. H. V. 2009. How much palm oil do we need? *Environmental Science & Policy,* 12**,** 134-139.

CORLEY, R. H. V. & TINKER, P. B. 2003. *The Oil Palm,* Oxford, Blackwell Science.

CRUDEN, R. W. 1988. Temporal Dioecism: Systematic Breadth, Associated Traits, and Temporal Patterns. *Botanical Gazette,* 149**,** 1-15.

CUADRADO, A., CARDOSO, M. & JOUVE, N. 2008. Increasing the physical markers of wheat chromosomes using SSRs as FISH probes. *Genome,* 51**,** 809-815.

CUADRADO, A., SCHWARZACHER, T. & JOUVE, N. 2000. Identification of different chromatin classes in wheat using *in situ* hybridization with simple sequence repeat oligonucleotides. *Theoretical and Applied Genetics,* 101**,** 711-717.

CUADRADO, A. & SCHWARZACHER, T. 1998. The chromosomal organization of simple sequence repeats in wheat and rye genomes. *Chromosoma,* 107**,** 587-594.

D'HONT, A., DENOEUD, F., AURY, J.-M., BAURENS, F.-C., CARREEL, F., GARSMEUR, O., NOEL, B., BOCS, S., DROC, G., ROUARD, M., DA SILVA, C., JABBARI, K., CARDI, C., POULAIN, J., SOUQUET, M., LABADIE, K., JOURDA, C., LENGELLÉ, J., RODIER-GOUD, M., ALBERTI, A., BERNARD, M., CORREA, M., AYYAMPALAYAM, S., MCKAIN, M. R., LEEBENS-MACK, J., BURGESS, D., FREELING, M., MBÉGUIÉ-A-MBÉGUIÉ, D., CHABANNES, M., WICKER, T., PANAUD, O., BARBOSA, J., HRIBOVA, E., HESLOP-HARRISON, P., HABAS, R., RIVALLAN, R., FRANCOIS, P., POIRON, C., KILIAN, A., BURTHIA, D., JENNY, C., BAKRY, F., BROWN, S., GUIGNON, V., KEMA, G., DITA, M., WAALWIJK, C., JOSEPH, S., DIEVART, A., JAILLON, O., LECLERCQ, J., ARGOUT, X., LYONS, E., ALMEIDA, A., JERIDI, M., DOLEZEL, J., ROUX, N., RISTERUCCI, A.-M., WEISSENBACH, J., RUIZ, M., GLASZMANN, J.-C., QUÉTIER, F., YAHIAOUI, N. &

WINCKER, P. 2012. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature,* 488**,** 213.

DANILOVA, T. V., FRIEBE, B. & GILL, B. S. 2012. Single-copy gene fluorescence *in situ* hybridization and genome analysis: Acc-2 loci mark evolutionary chromosomal rearrangements in wheat. *Chromosoma,* 121**,** 597-611.

DANILOVA, T. V. & BIRCHLER, J. A. 2008. Integrated cytogenetic map of mitotic metaphase chromosome 9 of maize: resolution, sensitivity, and banding paint development. *Chromosoma,* 117.

DITTUS, W. 2017. The biogeography and ecology of Sri Lankan mammals points to conservation priorities. *Ceylon Journal of Science,* 46**,** 33-64.

DIVASHUK, M. G., KHUAT, T. M., KROUPIN, P. Y., KIROV, I. V., ROMANOV, D. V., KISELEVA, A. V., KHRUSTALEVA, L. I., ALEXEEV, D. G., ZELENIN, A. S., KLIMUSHINA, M. V., RAZUMOVA, O. V. & KARLOV, G. I. 2016. Variation in Copy Number of Ty3/Gypsy Centromeric Retrotransposons in the Genomes of Thinopyrum intermedium and Its Diploid Progenitors. *PLoS One,* 11**,** e0154241.

DOYLE, J. J. & DOYLE, J. L. 1990. Isolation of plant DNA from fresh tissue. *Focus,* 12**,** 39-40.

DONG, G., SHEN, J., ZHANG, Q., WANG, J., YU, Q., MING, R., WANG, K. & ZHANG, J. 2018. Development and Applications of Chromosome-Specific Cytogenetic BAC-FISH Probes in *S. spontaneum*. *Frontiers in Plant Science,* 9.

DU, P., LI, L., LIU, H., FU, L., QIN, L., ZHANG, Z., CUI, C., SUN, Z., HAN, S., XU, J., DAI, X., HUANG, B., DONG, W., TANG, F., ZHUANG, L., HAN, Y., QI, Z. & ZHANG, X. 2018. High-resolution chromosome painting with repetitive and single-copy oligonucleotides in Arachis species identifies structural rearrangements and genome differentiation. *BMC Plant Biology,* 18**,** 240.

EDGAR, R. C. & MYERS, E. W. 2005. PILER: identification and classification of genomic repeats. *Bioinformatics,* 21 Suppl 1**,** i152-8.

ELLINGHAUS, D., KURTZ, S. & WILLHOEFT, U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics,* 9**,** 18.

ESCOBAR, R. 1981. Preliminary results of the collection and evaluation of the American oil palm *Elaeis oleifera* (H.B.K.) Cortes in Costa Rica, Panama and Colombia*,* Kuala Lumpur, Incorporated Society of Planters.

ESCUDEIRO, A., FERREIRA, D., MENDES-DA-SILVA, A., HESLOP-HARRISON, J. S., ADEGA, F. & CHAVES, R. 2019. Bovine satellite DNAs – a history of the evolution of complexity and its impact in the Bovidae family. *The European Zoological Journal,* 86**,** 20-37.

FLAVELL, A. J., SMITH, D. B. & KUMAR, A. 1992. Extreme heterogeneity of Ty1-copia group retrotransposons in plants. *Molecular and General Genetics MGG,* 231**,** 233-242.

FIGUEROA, D. M. & BASS, H. W. 2010. A historical and modern perspective on plant cytogenetics. *Briefings in Functional Genomics,* 9**,** 95-102.

FILHO, J. A. F., DE BRITO, L. S., LEÃO, A. P., ALVES, A. A., FORMIGHIERI, E. F. & JÚNIOR, M. T. S. 2017. In Silico Approach for Characterization and Comparison of Repeats in the Genomes of Oil and Date Palms. *Bioinformatics and biology insights,* 11**,** 1177932217702388-1177932217702388.

FINDLEY, S. D., PAPPAS, A. L., CUI, Y., BIRCHLER, J. A., PALMER, R. G. & STACEY, G. 2011. Fluorescence *in situ* hybridization-based karyotyping of soybean translocation lines. *G3 (Bethesda, Md.),* 1**,** 117-129.

FINDLEY, S. D., CANNON, S., VARALA, K., DU, J., MA, J., HUDSON, M. E., BIRCHLER, J. A. & STACEY, G. 2010. A fluorescence *in situ* hybridization system for karyotyping soybean. *Genetics,* 185.

FLUTRE, T., DUPRAT, E., FEUILLET, C. & QUESNEVILLE, H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLoS One,* 6**,** e16526.

FONSECA, A., FERREIRA, J., DOS SANTOS, T. R., MOSIOLEK, M., BELLUCCI, E., KAMI, J., GEPTS, P., GEFFROY, V., SCHWEIZER, D., DOS SANTOS, K. G. & PEDROSA-HARAND, A. 2010. Cytogenetic map of common bean (Phaseolus vulgaris L.). *Chromosome Res,* 18**,** 487-502.

FRANSZ, P., SOPPE, W. & SCHUBERT, I. 2003. Heterochromatin in interphase nuclei of Arabidopsis thaliana. *Chromosome Research,* 11**,** 227-240.

FRANSZ, P. F., STAM, M., MONTIJN, B., HOOPEN, R. T., WIEGANT, J., KOOTER, J. M., OUD, O. & NANNINGA, N. 1996. Detection of single-copy genes and chromosome rearrangements in *Petunia hybrida* by fluorescence in situ hybridization. *The Plant Journal,* 9**,** 767-774.

FRIEBE, B., JIANG, J., RAUPP, W., MCINTOSH, R. & GILL, B. 1996. Characterization of wheat-alien translocations conferring resistance to diseases and pests: current status. *Euphytica,* 91**,** 59-87.

FU, S., CHEN, L., WANG, Y., LI, M., YANG, Z., QIU, L., YAN, B., REN, Z. & TANG, Z. 2015. Oligonucleotide Probes for ND-FISH Analysis to Identify Rye and Wheat Chromosomes. *Scientific Reports,* 5**,** 10552.

GALL, J. G. & PARDUE, M. L. 1969. Formation and detection of RNA-DNA hybrid molecules in cytological preparations. *Proc Natl Acad Sci U S A,* 63**,** 378-83.

GARCIA-PEREZ, J. L., WIDMANN, T. J. & ADAMS, I. R. 2016. The impact of transposable elements on mammalian development. *Development (Cambridge, England),* 143**,** 4101-4114.

GARRIDO-RAMOS, M. A. 2015. Satellite DNA in Plants: More than Just Rubbish. *Cytogenetic and Genome Research,* 146**,** 153-170.

GARRIDO-RAMOS, M. A. 2017. Satellite DNA: An Evolving Topic. *Genes,* 8**,** 230.

GERLACH, W. L. & DYER, T. A. 1980. Sequence organization of the repeating units in the nucleus of wheat which contain 5S rRNA genes. *Nucleic acids research,* 8**,** 4851-4865.

GILL, N., FINDLEY, S., WALLING, J. G., HANS, C., MA, J., DOYLE, J., STACEY, G. & JACKSON, S. A. 2009. Molecular and chromosomal evidence for allopolyploidy in soybean. *Plant Physiol,* 151.

GILL, N., HANS, C. S. & JACKSON, S. 2008. An overview of plant chromosome structure. *Cytogenetic and Genome Research,* 120**,** 194-201.

GIRGIS, H. Z. 2015. Red: an intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale. *BMC bioinformatics,* 16**,** 227-227.

GOVAERTS R, DRANSFIELD J, ZONA S, HODEL DR & HENDERSON, A. 2015. World checklist of Arecaceae.: Facilitated by the Royal Botanic Gardens, Kew. Retrieved 22 February 2019.

GREILHUBER, J. 1977. Why plant chromosomes do not show G-bands? *Theor Appl Genet,* 50.

GROS-BALTHAZARD, M., HAZZOURI, K. M. & FLOWERS, J. M. 2018. Genomic Insights into Date Palm Origins. *Genes,* 9**,** 502.

GUNSTONE, F. D. & HARWOOD, J. L. 2007. *Occurrence and Characterization of Oils and Fat,* Boca Raton, CRC.

HAN, Y., ZHANG, T., THAMMAPICHAI, P., WENG, Y. & JIANG, J. 2015. Chromosome-Specific Painting in *Cucumis* Species Using Bulked Oligonucleotides. *Genetics,* 200**,** 771-9.

HANSEN, C. & HESLOP-HARRISON, J. S. 2004. Sequences and Phylogenies of Plant Pararetroviruses, Viruses, and Transposable Elements. *Advances in Botanical Research.* Academic Press.

HARDON, J. J. 1969. Interspecific hybrids in the genus Elaeis II. vegetative growth and yield of F1 hybrids *E. guineensis* x *E. oleifera*. *Euphytica,* 18**,** 380-388.

HARDON, J. J. & TAN, G. Y. 1969. Interspecific hybrids in the genus *Elaeis* I. crossability, cytogenetics and fertility of F1 hybrids of *E. guineensis* x *E. oleifera*. *Euphytica,* 18**,** 372-379.

HARTLEY, C. W. S. 1967. *The Oil Palm (Elaeis guineensis Jacq.),* London, Longmans Green & Co. Ltd.

HAYATI, A., WICKNESWARI, R., MAIZURA, I. & RAJANAIDU, N. 2004. Genetic diversity of oil palm (*Elaeis guineensis* Jacq.) germplasm collections from Africa: implications for improvement and conservation of genetic resources. *Theoretical and Applied Genetics,* 108**,** 1274-1284.

HE, Q., CAI, Z., HU, T., LIU, H., BAO, C., MAO, W. & JIN, W. 2015. Repetitive sequence analysis and karyotyping reveals centromere-associated DNA sequences in radish (*Raphanus sativus* L.). *BMC Plant Biology,* 15**,** 105.

HE, Y., WANG, C., HIGGINS, J. D., YU, J., ZONG, J., LU, P., ZHANG, D. & LIANG, W. 2016. MEIOTIC F-BOX Is Essential for Male Meiotic DNA Double-Strand Break Repair in Rice. *The Plant Cell,* 28**,** 1879-1893.

HESLOP-HARRISON, J. S. 2000. Comparative Genome Organization in Plants: From Sequence and Markers to Chromatin and Chromosomes. *The Plant Cell,* 12**,** 617-635.

HESLOP-HARRISON, J. S., BRANDES, A., TAKETA, S., SCHMIDT, T., VERSHININ, A. V., ALKHIMOVA, E. G., KAMM, A., DOUDRICK, R. L., SCHWARZACHER, T., KATSIOTIS, A., KUBIS, S., KUMAR, A., PEARCE, S. R., FLAVELL, A. J. & HARRISON, G. E. 1997. The chromosomal distributions of Ty1-copia group retrotransposable elements in higher plants and their implications for genome evolution. *Genetica,* 100**,** 197-204.

HESLOP-HARRISON, J. S., HARRISON, G. E. & LEITCH, I. J. 1992. Reprobing of DNA: DNA in situ hybridization preparations. *Trends in Genetics,* 8**,** 372-373.

HESLOP-HARRISON, J. S. & SCHMIDT, R. 2012. Plant Nuclear Genome Composition. *eLS.*

HESLOP-HARRISON, J. S. & SCHWARZACHER, T. 2011. Organisation of the plant genome in chromosomes. *Plant J,* 66**,** 18-33.

HESLOP-HARRISON, J. S., SCHWARZACHER, T., ANAMTHAWAT-JDNSSON, K., LEITCH, A. R., SHI, M. & LEITCH, I. J. 1991. In-situ hybridization with automated chromosome denaturation. *Technique,* 3.

HOFFMANN, M. P., DONOUGH, C. R., COOK, S. E., FISHER, M. J., LIM, C. H., LIM, Y. L., COCK, J., KAM, S. P., MOHANARAJ, S. N., INDRASUARA, K., TITTINUTCHANON, P. & OBERTHÜR, T. 2017. Yield gap analysis in oil palm: Framework development and application in commercial operations in Southeast Asia. *Agricultural Systems,* 151**,** 12-19.

HOU, L., XU, M., ZHANG, T., XU, Z., WANG, W., ZHANG, J., YU, M., JI, W., ZHU, C., GONG, Z., GU, M., JIANG, J. & YU, H. 2018. Chromosome painting and its applications in cultivated and wild rice. *BMC plant biology,* 18**,** 110-110.

HUANG, Y., LUO, L., HU, X., YU, F., YANG, Y., DENG, Z., WU, J., CHEN, R. & ZHANG, M. 2017. Characterization, Genomic Organization, Abundance, and Chromosomal Distribution of Ty1-copia Retrotransposons in *Erianthus arundinaceus*. *Frontiers in Plant Science,* 8.

HUBER, D., VOITH VON VOITHENBERG, L. & KAIGALA, G. V. 2018. Fluorescence in situ hybridization (FISH): History, limitations and what to expect from micro-scale FISH? *Micro and Nano Engineering,* 1**,** 15-24.

HUZIWARA, Y. 1962. Karyotype Analysis in Some Genera of Compositae. VIII. Further Studies on the Chromosomes of Aster. *American Journal of Botany,* 49**,** 116-119.

IDZIAK, D., BETEKHTIN, A., WOLNY, E., LESNIEWSKA, K., WRIGHT, J., FEBRER, M., BEVAN, M. W., JENKINS, G. & HASTEROK, R. 2011. Painting the chromosomes of Brachypodium—current status and future prospects. *Chromosoma,* 120**,** 469-479.

IJDO, J., BALDINI, A., WARD, D., REEDERS, S. & WELLS, R. 1991. Origin of human chromosome 2: an ancestral telomere-telomere fusion. *Proceedings of the National Academy of Sciences,* 88**,** 9051-9055.

ITHNIN, M., XU, Y., MARJUNI, M., SERDARI, N. M., AMIRUDDIN, M. D., LOW, E.-T. L., TAN, Y.-C., YAP, S.-J., OOI, L. C. L., NOOKIAH, R., SINGH, R. & XU, S. 2017. Multiple locus genome-wide association studies for important economic traits of oil palm. *Tree Genetics & Genomes,* 13**,** 103.

JALIGOT, E., HOOI, W. Y., DEBLADIS, E., RICHAUD, F., BEULÉ, T., COLLIN, M., AGBESSI, M. D. T., SABOT, F., GARSMEUR, O., D'HONT, A., ALWEE, S. S. R. S. & RIVAL, A. 2014. DNA Methylation and Expression of the EgDEF1 Gene and Neighboring Retrotransposons in mantled Somaclonal Variants of Oil Palm. *PLOS ONE,* 9**,** e91896.

JANICKI, M., ROOKE, R. & YANG, G. 2011. Bioinformatics and genomic analysis of transposable elements in eukaryotic genomes. *Chromosome Res,* 19**,** 787-808.

JANDA, J., ŠAFÁŘ, J., KUBALÁKOVÁ, M., BARTOŠ, J., KOVÁŘOVÁ, P., SUCHÁNKOVÁ, P., PATEYRON, S., ČÍHALÍKOVÁ, J., SOURDILLE, P., ŠIMKOVÁ, H., FAIVRE-RAMPANT, P., HŘIBOVÁ, E., BERNARD, M., LUKASZEWSKI, A., DOLEŽEL, J. & CHALHOUB, B. 2006. Advanced resources for plant genomics: a BAC library specific for the short arm of wheat chromosome 1B. *The Plant Journal,* 47**,** 977-986.

JIANG, J. 2019. Fluorescence in situ hybridization in plants: recent developments and future applications. *Chromosome Research*.

JIANG, J., BIRCHLER, J. A., PARROTT, W. A. & KELLY DAWE, R. 2003. A molecular view of plant centromeres. *Trends in Plant Science,* 8**,** 570-575.

JIANG, J. & GILL, B. S. 2006. Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. *Genome,* 49**,** 1057-68.

JIANG, J., NASUDA, S., DONG, F., SCHERRER, C. W., WOO, S.-S., WING, R. A., GILL, B. S. & WARD, D. C. 1996. A conserved repetitive DNA element located in the centromeres of cereal chromosomes. *Proceedings of the National Academy of Sciences,* 93**,** 14210-14213.

JOHN, H. A., BIRNSTIEL, M. L. & JONES, K. W. 1969. RNA-DNA hybrids at the cytological level. *Nature,* 223**,** 582-7.

JOHNSON, V., BOURDEIX, R., PRADES, A. & PICQ, C. 2018. Global strategy for the conservation and use of coconut genetic resources 2018 -2028: summary brochure. Bioversity International; Cogent.

JOHNSTON, J. S., PEPPER, A. E., HALL, A. E., HODNETT, G., PRICE, H. J., DRABEK, J., LOPEZ, R. & CHEN, Z. J. 2005. Evolution of Genome Size in Brassicaceae. *Annals of Botany,* 95**,** 229-235.

JONES, K. 1998. Robertsonian fusion and centric fission in karyotype evolution of higher plants. *The Botanical Review,* 64**,** 273-289.

KARLOV, G. I., FESENKO, I. A., ANDREEVA, G. N. & KHRUSTALEVA, L. I. 2010. Chromosome organization of Ty1-copia-like retrotransposons in the tomato genome. *Russian Journal of Genetics,* 46**,** 677-681.

KATO, A., ALBERT, P. S., VEGA, J. M. & BIRCHLER, J. A. 2006. Sensitive fluorescence in situ hybridization signal detection in maize using directly labeled probes produced by high concentration DNA polymerase nick translation. *Biotech Histochem,* 81.

KATO, A., LAMB, J. C., ALBERT, P. S., DANILOVA, T., HAN, F., GAO, Z., FINDLEY, S. & BIRCHLER, J. A. 2011. Chromosome Painting for Plant Biotechnology. *In:* BIRCHLER, J. A. (ed.) *Plant Chromosome Engineering: Methods and Protocols.* Totowa, NJ: Humana Press.

KATO, A., LAMB, J. C. & BIRCHLER, J. A. 2004. Chromosome painting using repetitive DNA sequences as probes for somatic chromosome identification in maize. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 13554-13559.

KHRUSTALEVA, L. & KIK, C. 2001. Localization of single copy T-DNA insertion in transgenic shallots (Allium cepa L.) by using ultra-sensitive FISH with tyramide signal amplification. *Plant J,* 25.

KHRUSTALEVA, L., KIROV, I., ROMANOV, D., BUDYLIN, M., LAPITSKAYA, I., KISELEVA, A., FESENKO, I. & KARLOV, G. 2012. The chromosome organization of genes and some types of extragenic DNA in Allium. *Acta Hort,* 969.

KIROV, I., DIVASHUK, M., VAN LAERE, K., SOLOVIEV, A. & KHRUSTALEVA, L. 2014. An easy "SteamDrop" method for high quality plant chromosome preparation. *Molecular Cytogenetics,* 7**,** 21.

KLEMME, S., BANAEI-MOGHADDAM, A. M., MACAS, J., WICKER, T., NOVAK, P. & HOUBEN, A. 2013. High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytol,* 199**,** 550-8.

KOO, D. H., CHOI, H. W., CHO, J., HUR, Y. & BANG, J. W. 2005. A high-resolution karyotype of cucumber (*Cucumis sativus* L. 'Winter Long') revealed by C-banding, pachytene analysis, and RAPD-aided fluorescence in situ hybridization. *Genome,* 48**,** 534-40.

KRASSOVSKY, K. & HENIKOFF, S. 2014. Distinct chromatin features characterize different classes of repeat sequences in Drosophila melanogaster. *BMC Genomics,* 15**,** 105.

KRIVANKOVA, A., KOPECKY, D., STOCES, S., DOLEZEL, J. & HRIBOVA, E. 2017. Repetitive DNA: A Versatile Tool for Karyotyping in Festuca pratensis Huds. *Cytogenet Genome Res,* 151**,** 96-105.

KRUPPA, K., SEPSI, A., SZAKÁCS, É., RÖDER, M. S. & MOLNÁR-LÁNG, M. 2013. Characterization of a 5HS-7DS.7DL wheat-barley translocation line and physical mapping of the 7D chromosome using SSR markers. *Journal of Applied Genetics,* 54**,** 251-258.

KUBIS, S. E., CASTILHO, A. M. M. F., VERSHININ, A. V. & HESLOP-HARRISON, J. S. 2003. Retroelements, transposons and methylation status in the genome of oil palm (Elaeis guineensis) and the relationship to somaclonal variation. *Plant Molecular Biology,* 52**,** 69-79.

KUBIS, S. E., HESLOP-HARRISON, J. S., DESEL, C. & SCHMIDT, T. 1998. The genomic organization of non-LTR retrotransposons (LINEs) from three Beta species and five other angiosperms. *Plant molecular biology,* 36**,** 821-831.

KUMAR, A. & BENNETZEN, J. L. 1999. Plant retrotransposons. *Annual Rev Genet,* 33**,** 479-532.

KURTZ, S., NARECHANIA, A., STEIN, J. C. & WARE, D. 2008. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. *BMC Genomics,* 9**,** 517.

KUSHAIRI, A. & RAJANAIDU, N. 2000. Breeding Population, Seed Production and Nursery Management. Kuala Lumpur: Malaysian Palm Oil Board.

LANGER-SAFER, P. R., LEVINE, M. & WARD, D. C. 1982. Immunological method for mapping genes on Drosophila polytene chromosomes. *Proc Natl Acad Sci U S A,* 79**,** 4381-5.

LANGRIDGE, P. & WAUGH, R. 2019. Harnessing the potential of germplasm collections. *Nature Genetics,* 51**,** 200-201.

LAWRENCE, G. J. & APPELS, R. 1986. Mapping the nucleolus organizer region, seed protein loci and isozyme loci on chromosome 1R in rye. *Theor Appl Genet,* 71.

LEITCH, I. J., LEITCH, A. R. & HESLOP-HARRISON, J. S. 1991. Physical mapping of plant DNA sequences by simultaneous in situ hybridization of two differently labelled fluorescent probes. *Genome,* 34**,** 329-333.

LEJEUNE, J., DUTRILLAUX, B., RETHORÉ, M. O. & PRIEUR, M. 1973. Comparaison de la structure fine des chromatides d'Homo sapiens et de Pan troglodytes. *Chromosoma,* 43**,** 423-444.

LENGEROVA, M., KEJNOVSKY, E., HOBZA, R., MACAS, J., GRANT, S. R. & VYSKOT, B. 2004. Multicolor FISH mapping of the dioecious model plant, *Silene latifolia*. *Theor Appl Genet,* 108**,** 1193-9.

LERAT, E. 2010. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity (Edinb),* 104**,** 520-33.

LEUTWILER, L. S., HOUGH-EVANS, B. R. & MEYEROWITZ, E. M. 1984. The DNA of *Arabidopsis thaliana*. *Molecular and General Genetics MGG,* 194**,** 15-23.

LEVAN, A., FREDGA, K. & SANDBERG, A. A. 1964. Nomenclature for centromeric position on chromosome.

LEVSKY, J. M. & SINGER, R. H. 2003. Fluorescence in situ hybridization: past, present and future. *Journal of Cell Science,* 116**,** 2833-2838.

LI, J., WEBSTER, M. A., WRIGHT, J., COCKER, J. M., SMITH, M. C., BADAKSHI, F., HESLOP-HARRISON, P. & GILMARTIN, P. M. 2015. Integration of genetic and physical maps of the *Primula vulgaris* S locus and localization by chromosome in situ hybridization. *New Phytol,* 208**,** 137-48.

LI, K., WANG, H., WANG, J., SUN, J., LI, Z. & HAN, Y. 2016. Divergence between C. melo and African Cucumis Species Identified by Chromosome Painting and rDNA Distribution Pattern. *Cytogenet Genome Res,* 150**,** 150-155.

LI, S.-F., SU, T., CHENG, G.-Q., WANG, B.-X., LI, X., DENG, C.-L. & GAO, W.-J. 2017. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes,* 8**,** 290.

LIU, S., ZHENG, J., MIGEON, P., REN, J., HU, Y., HE, C., LIU, H., FU, J., WHITE, F. F., TOOMAJIAN, C. & WANG, G. 2017. Unbiased K-mer Analysis Reveals Changes in Copy Number of Highly Repetitive Sequences During Maize Domestication and Improvement. *Scientific Reports,* 7**,** 42444.

LOH, S. K. & CHOO, Y. M. 2013. Prospect, Challenges and Opportunities on Biofuels in Malaysia. *In:* POGAKU, R. & SARBATLY, R. H. (eds.) *Advances in Biofuels.* Boston, MA: Springer US.

LOW, E., NAGAPPAN, J., KUANG-LIM, C., NIK SHAZANA, N. M. S., MOHD AMIN, A. H., ROZANA, R., NORAZAH, A., NADZIRAH, A., LEE, P. L. A., ONG-ABDULLAH, M., SINGH, R., MOHAMAD ARIF, A. M., RAVIGADEVI, S., PARVEEZ, G. K. A. & KUSHAIRI, A. 2017. The oil palm genome revolution. *Journal of Oil Palm Research,* 29**,** 456-468.

LUO, S., MACH, J., ABRAMSON, B., RAMIREZ, R., SCHURR, R., BARONE, P., COPENHAVER, G. & FOLKERTS, O. 2012. The Cotton Centromere Contains a Ty3-gypsy-like LTR Retroelement. *PLOS ONE,* 7**,** e35261.

LUSINSKA, J., MAJKA, J., BETEKHTIN, A., SUSEK, K., WOLNY, E. & HASTEROK, R. 2018. Chromosome identification and reconstruction of evolutionary rearrangements in *Brachypodium distachyon*, *B. stacei* and *B. hybridum*. *Annals of botany,* 122**,** 445-459.

LYSAK, M. A., BERR, A., PECINKA, A., SCHMIDT, R., MCBREEN, K. & SCHUBERT, I. 2006. Mechanisms of chromosome number reduction in *Arabidopsis thaliana* and related Brassicaceae species. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 5224-5229.

LYSAK, M. A., KOCH, M. A., PECINKA, A. & SCHUBERT, I. 2005. Chromosome triplication found across the tribe Brassicaceae. *Genome Research,* 15**,** 516-525.

MA, Y., ISLAM-FARIDL, M. N., CRANE, C. F., STELLY, D. M., PRICE, H. J. & BYRNE, D. H. 1996. A new procedure to prepare slides of metaphase chromosomes of roses. *HortScience,* 31.

MACAS, J., KEJNOVSKÝ, E., NEUMANN, P., NOVÁK, P., KOBLÍŽKOVÁ, A. & VYSKOT, B. 2011. Next Generation Sequencing-Based Analysis of Repetitive DNA in the Model Dioceous Plant *Silene latifolia*. *PLOS ONE,* 6**,** e27335.

MACAS, J., NOVÁK, P., PELLICER, J., ČÍŽKOVÁ, J., KOBLÍŽKOVÁ, A., NEUMANN, P., FUKOVÁ, I., DOLEŽEL, J., KELLY, L. J. & LEITCH, I. J. 2015. In Depth Characterization of Repetitive DNA in 23 Plant Genomes Reveals Sources of Genome Size Variation in the Legume Tribe *Fabeae*. *PLOS ONE,* 10**,** e0143424.

MADON, M., ARULANDOO, X. & ZAKI 2018. Short communication: Genomic constitution of oil palm interspecific hybrid crosses monitored by Genomic in situ Hybridization (GISH). *Journal of Oil Palm Research*.

MADON, M., CLYDE, M. M. & CHEAH, S. C. 1998. Cytological analysis of *Elaeis guineensis* and *Elaeis oleifera* chromosomes. *Journal of Oil Palm Research,* 10**,** 68-91.

MADON, M., CLYDE, M. M. & CHEAH, S. C. 1999. Application of genomic in situ hybridization (GISH) on *Elaeis* hybrids. *Journal of Oil Palm Research***,** 74-80.

MADON, M. & HESLOP-HARRISON, J. S. 2001. Physical mapping of rRNA genes on *Elaeis* chromosomes. *Journal of Oil Palm Research*.

MADON, M., PHOON, L., CLYDE, M. & MOHD DIN, A. 2008. Application of flow cytometry for estimation of nuclear DNA content in *Elaeis*. *Journal of Oil Palm Research,* 20**,** 447-452.

MALUSZYNSKA, J. & HESLOP-HARRISON, J. S. 1993. Molecular Cytogenetics of the Genus Arabidopsis: In situ Localization of rDNA Sites, Chromosome Numbers and Diversity in Centromeric Heterochromatin. *Annals of Botany,* 71**,** 479-484.

MANDÁKOVÁ, T., JOLY, S., KRZYWINSKI, M., MUMMENHOFF, K. & LYSAK, M. A. 2010. Fast Diploidization in Close Mesopolyploid Relatives of Arabidopsis. *The Plant Cell,* 22**,** 2277.

MANDÁKOVÁ, T. & LYSAK, M. A. 2016. Painting of Arabidopsis Chromosomes with Chromosome-Specific BAC Clones. *Current Protocols in Plant Biology,* 1**,** 359-371.

MARÇAIS, G. & KINGSFORD, C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics,* 27**,** 764-770.

MATHEW, L. S., SPANNAGL, M., AL-MALKI, A., GEORGE, B., TORRES, M. F., AL-DOUS, E. K., AL-AZWANI, E. K., HUSSEIN, E., MATHEW, S., MAYER, K. F., MOHAMOUD, Y. A., SUHRE, K. & MALEK, J. A. 2014. A first genetic map of date palm (*Phoenix dactylifera*) reveals long-range genome structure conservation in the palms. *BMC Genomics,* 15**,** 285.

MAY, C. Y. & NESARETNAM, K. 2014. Research advancements in palm oil nutrition. *European Journal of Lipid Science and Technology,* 116**,** 1301-1315.

MAYER, K. F. X., MARTIS, M., HEDLEY, P. E., ŠIMKOVÁ, H., LIU, H., MORRIS, J. A., STEUERNAGEL, B., TAUDIEN, S., ROESSNER, S., GUNDLACH, H., KUBALÁKOVÁ, M., SUCHÁNKOVÁ, P., MURAT, F., FELDER, M., NUSSBAUMER, T., GRANER, A., SALSE, J., ENDO, T., SAKAI, H., TANAKA, T., ITOH, T., SATO, K., PLATZER, M., MATSUMOTO, T., SCHOLZ, U., DOLEŽEL, J., WAUGH, R. & STEIN, N. 2011. Unlocking the Barley Genome by Chromosomal and Comparative Genomics. *The Plant Cell,* 23**,** 1249-1263.

MAYES, S., HAFEEZ, F., PRICE, Z., MACDONALD, D., BILLOTTE, N. & ROBERTS, J. 2008. Molecular Research in Oil Palm, the Key Oil Crop for the Future. *In:* MOORE, P. H. & MING, R. (eds.) *Genomics of Tropical Crop Plants.* New York, NY: Springer New York.

MAYES, S., JACK, P. L., CORLEY, R. H. V. & MARSHALL, D. F. 1997. Construction of a RFLP genetic linkage map for oil palm (*Elaeis guineensis* Jacq.). *Genome,* 40**,** 116-122.

MCCARTHY, E. M. & MCDONALD, J. F. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics,* 19**,** 362-7.

MEHROTRA, S. & GOYAL, V. 2014. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. *Genomics Proteomics Bioinformatics,* 12**,** 164-71.

MENG, Z., ZHANG, Z., YAN, T., LIN, Q., WANG, Y., HUANG, W., HUANG, Y., LI, Z., YU, Q., WANG, J. & WANG, K. 2018. Comprehensively Characterizing the Cytological Features of *Saccharum spontaneum* by the Development of a Complete Set of Chromosome-Specific Oligo Probes. *Frontiers in Plant Science,* 9.

MOLNÁR-LÁNG, M., LINC, G. & SZAKÁCS, É. 2014. Wheat–barley hybridization: the last 40 years. *Euphytica,* 195**,** 315-329.

MONTOYA, C., COCHARD, B., FLORI, A., CROS, D., LOPES, R., CUELLAR, T., ESPEOUT, S., SYAPUTRA, I., VILLENEUVE, P., PINA, M., RITTER, E., LEROY, T. & BILLOTTE, N. 2014. Genetic Architecture of Palm Oil Fatty Acid Composition in Cultivated Oil Palm

(*Elaeis guineensis* Jacq.) Compared to Its Wild Relative E. oleifera (H.B.K) Cortés. *PLOS ONE,* 9**,** e95412.

MORALLI, D., YUSUF, M., MANDEGAR, M., KHOJA, S., MONACO, Z. L. & VOLPI, E. V. 2011. An improved technique for chromosomal analysis of human ES and iPS cells. *Stem Cell Rev,* 7.

MOREIRA-FILHO, O., KUHN, G. C. S., KÜTTLER, H. & HESLOP-HARRISON, J. S. 2011. The 1.688 Repetitive DNA of Drosophila: Concerted Evolution at Different Genomic Scales and Association with Genes. *Molecular Biology and Evolution,* 29**,** 7-11.

MORENO, R., CASTRO, P., VRÁNA, J., KUBALÁKOVÁ, M., CÁPAL, P., GARCÍA, V., GIL, J., MILLÁN, T. & DOLEŽEL, J. 2018. Integration of Genetic and Cytogenetic Maps and Identification of Sex Chromosome in Garden Asparagus (*Asparagus officinalis* L.). *Frontiers in Plant Science,* 9.

MOSIER, N. S., SARIKAYA, A., LADISCH, C. M. & LADISCH, M. R. 2001. Characterization of dicarboxylic acids for cellulose hydrolysis. *Biotechnol Prog,* 17.

MUNIER, P. 1973. *The date palm,* Paris, G.-P. Maisonneuve et Larose.

MURATA, M. 1983. Staining air-dried protoplasts for study of plant chromosomes. *Stain Technol,* 58.

MURPHY, W. J., LARKIN, D. M., DER WIND, A. E.-V., BOURQUE, G., TESLER, G., AUVIL, L., BEEVER, J. E., CHOWDHARY, B. P., GALIBERT, F., GATZKE, L., HITTE, C., MEYERS, S. N., MILAN, D., OSTRANDER, E. A., PAPE, G., PARKER, H. G., RAUDSEPP, T., ROGATCHEVA, M. B., SCHOOK, L. B., SKOW, L. C., WELGE, M., WOMACK, J. E., O'BRIEN, S. J., PEVZNER, P. A. & LEWIN, H. A. 2005. Dynamics of Mammalian Chromosome Evolution Inferred from Multispecies Comparative Maps. *Science,* 309**,** 613-617.

MURPHY, D. J. 2007. Future prospects for oil palm in the 21st century: Biological and related challenges. *European Journal of Lipid Science and Technology,* 109**,** 296-306.

MURPHY, D. J. 2014. The future of oil palm as a major global crop: Opportunities and challenges. *Journal of Oil Palm Research,* 26.

MUSTAFA, S. I. 2018. *Mitochondrial and repetitive DNA defining the sheep genome landscape.* PhD Thesis, University of Leicester.

NAGAKI, K., NEUMANN, P., ZHANG, D., OUYANG, S., BUELL, C. R., CHENG, Z. & JIANG, J. 2004. Structure, divergence, and distribution of the CRR centromeric retrotransposon family in rice. *Molecular biology and evolution,* 22**,** 845-855.

NAGARAJAN, N. & POP, M. 2013. Sequence assembly demystified. *Nature Reviews Genetics,* 14**,** 157.

NAGENDRAN, B., UNNITHAN, U. R., CHOO, Y. M. & SUNDRAM, K. 2000. Characteristics of Red Palm Oil, a Carotene- and Vitamin E–rich Refined Oil for Food Uses. *Food and Nutrition Bulletin,* 21**,** 189-194.

NANI, T. F., SCHNABLE, J. C., WASHBURN, J. D., ALBERT, P., PEREIRA, W. A., SOBRINHO, F. S., BIRCHLER, J. A. & TECHIO, V. H. 2018. Location of low copy genes in chromosomes of Brachiaria spp. *Molecular Biology Reports,* 45**,** 109-118.

NAVAJAS-PEREZ, R., SCHWARZACHER, T., RUIZ REJON, M. & GARRIDO-RAMOS, M. A. 2009. Characterization of RUSI, a telomere-associated satellite DNA, in the genus Rumex (Polygonaceae). *Cytogenet Genome Res,* 124**,** 81-9.

NEUMANN, P., NOVÁK, P., HOŠTÁKOVÁ, N. & MACAS, J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. *Mobile DNA,* 10**,** 1.

NICK, C., LANES, É. C. M., KUKI, K. N., FREITAS, R. D. & MOTOIKE, S. Y. 2014. Molecular Characterization and Population Structure of the Macaw Palm, *Acrocomia aculeata* (Arecaceae), Ex Situ Germplasm Collection Using Microsatellites Markers. *Journal of Heredity,* 106**,** 102-112.

NOVAK, P., AVILA ROBLEDILLO, L., KOBLIZKOVA, A., VRBOVA, I., NEUMANN, P. & MACAS, J. 2017. TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Res,* 45**,** e111.

NOVÁK, P., NEUMANN, P., PECH, J., STEINHAISL, J. & MACAS, J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics,* 29**,** 792-793.

NOVÁK, P., NEUMANN, P. & MACAS, J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics,* 11**,** 378.

NOWICKA, A., GRZEBELUS, E. & GRZEBELUS, D. 2016. Precise karyotyping of carrot mitotic chromosomes using multicolour-FISH with repetitive DNA. *Biologia Plantarum,* 60**,** 25-36.

NUSSBAUMER, T., MARTIS, M. M., ROESSNER, S. K., PFEIFER, M., BADER, K. C., SHARMA, S., GUNDLACH, H. & SPANNAGL, M. 2012. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Research,* 41**,** D1144-D1151.

OBAHIAGBON, F. I. 2012. A review: Aspects of the African oil palm (*Elaeis guineensis* Jacq.) and the implication of its bioactives in human health. *American Journal of Biochemistry and Molecular Biology* 2**,** 1-14

OHMIDO, N., IWATA, A., KATO, S., WAKO, T. & FUKUI, K. 2018. Development of a quantitative pachytene chromosome map and its unification with somatic chromosome and linkage maps of rice (*Oryza sativa* L.). *PLOS ONE,* 13**,** e0195710.

OIL WORLD. 2018. *Oil World Annual 2018* [Online]. Hamburg: ISTA Mielke GmbH. Available: http://www.oilworld.de [Accessed 22 March 2019].

OLIVEIRA, L. C., DE OLIVEIRA, M. D. S. P., DAVIDE, L. C. & TORRES, G. A. 2016. Karyotype and genome size in *Euterpe* Mart. (Arecaceae) species. *Comparative cytogenetics,* 10**,** 17-25.

ONG-ABDULLAH, M., ORDWAY, J. M., JIANG, N., OOI, S.-E., KOK, S.-Y., SARPAN, N., AZIMI, N., HASHIM, A. T., ISHAK, Z., ROSLI, S. K., MALIKE, F. A., BAKAR, N. A. A., MARJUNI, M., ABDULLAH, N., YAAKUB, Z., AMIRUDDIN, M. D., NOOKIAH, R., SINGH, R., LOW, E.-T. L., CHAN, K.-L., AZIZI, N., SMITH, S. W., BACHER, B., BUDIMAN, M. A., VAN BRUNT, A., WISCHMEYER, C., BEIL, M., HOGAN, M., LAKEY, N., LIM, C.-C., ARULANDOO, X., WONG, C.-K., CHOO, C.-N., WONG, W.-C., KWAN, Y.-Y., ALWEE, S. S. R. S., SAMBANTHAMURTHI, R. & MARTIENSSEN, R. A. 2015. Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature,* 525**,** 533.

PAESOLD, S., BORCHARDT, D., SCHMIDT, T. & DECHYEVA, D. 2012. A sugar beet (*Beta vulgaris* L.) reference FISH karyotype for chromosome and chromosome-arm identification, integration of genetic linkage groups and analysis of major repeat family distribution. *The Plant Journal,* 72**,** 600-611.

PATOKAR, C., SEPSI, A., SCHWARZACHER, T., KISHII, M. & HESLOP-HARRISON, J. S. 2016. Molecular cytogenetic characterization of novel wheat-*Thinopyrum bessarabicum* recombinant lines carrying intercalary translocations. *Chromosoma,* 125**,** 163-172.

PEREIRA, T. N. S., NETO, M. F., DE SOUZA, M. M., GERONIMO, I. G. D. C., MELO, C. A. F. D. & PEREIRA, M. G. 2017. Cytological Characterization of Brazilian Green Dwarf Coconut (*Cocos nucifera* L.) via Meiosis and Conventional and Differential Karyotyping. *Cytologia,* 82**,** 167-174.

PEŠKA, V., MANDÁKOVÁ, T., IHRADSKÁ, V. & FAJKUS, J. 2019. Comparative Dissection of Three Giant Genomes: *Allium cepa, Allium sativum, and Allium ursinum*. *International Journal of Molecular Sciences,* 20**,** 733.

PETERSON, D. G. 2014. Chapter Two - Evolution of Plant Genome Analysis. *In:* PATERSON, A. H. (ed.) *Advances in Botanical Research.* Academic Press.

PIJNACKER, L. P. & FERWERDA, M. A. 1984. Giemsa C-banding of potato chromosomes. *Canadian Journal of Genetics and Cytology,* 26**,** 415-419.

PINKEL, D., STRAUME, T. & GRAY, J. W. 1986. Cytogenetic analysis using quantitative, high-sensitivity, fluorescence hybridization. *Proceedings of the National Academy of Sciences,* 83**,** 2934-2938.

PISUPATI, R., VERGARA, D. & KANE, N. C. 2018. Diversity and evolution of the repetitive genomic content in Cannabis sativa. *BMC genomics,* 19**,** 156-156.

PRESTING, G. G. 2018. Centromeric retrotransposons and centromere function. *Current Opinion in Genetics & Development,* 49**,** 79-84.

PRESTING, G. G., MALYSHEVA, L., FUCHS, J. & SCHUBERT, I. 1998. A TY3/GYPSY retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *The Plant Journal,* 16**,** 721-728.

PRICE, Z., SCHULMAN, A. H. & MAYES, S. 2003. Development of new marker-example from oil palm. *Plant Genetics Res,* 1**,** 103-113.

PURSEGLOVE, J. W. 1972. *Tropical Crops: Monocotyledons Longman (1972),* London, Longman.

QU, M., LI, K., HAN, Y., CHEN, L., LI, Z. & HAN, Y. 2017. Integrated Karyotyping of Woodland Strawberry (*Fragaria vesca*) with Oligopaint FISH Probes. *Cytogenetic and Genome Research,* 153**,** 158-164.

RAJANAIDU, N. *Elaeis oleifera* collection in Central and South America. Proceedings of the International Workshop: Oil palm germplasm and utilisation, 1986 Kuala Lumpur. Palm Oil Research Institute of Malaysia.

RENS, W., FU, B., O'BRIEN, P. C. M. & FERGUSON-SMITH, M. 2006. Cross-species chromosome painting. *Nature Protocols,* 1**,** 783.

RIVAL, A., BEULE, T., BARRE, P., HAMON, S., DUVAL, Y. & NOIROT, M. 1997. Comparative flow cytometric estimation of nuclear DNA content in oil palm (*Elaeis guineensis* Jacq) tissue cultures and seed-derived plants. *Plant Cell Reports,* 16**,** 884-887.

ROGERS, S. O. & BENDICH, A. J. 1985. Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol Biol,* 5.

ROMANOV, D., DIVASHUK, M., HAVEY, M. J. & KHRUSTALEVA, L. 2015. Tyramide-FISH mapping of single genes for development of an integrated recombination and cytogenetic map of chromosome 5 of *Allium cepa*. *Genome,* 58**,** 111-119.

RONNE, M. 1990. Chromosome preparation and high resolution banding (review). *In Vivo,* 4**,** 337-65.

ROTHFELS, K. H. & SIMINOVITCH, L. 1958. An Air-Drying Technique for Flattening Chromosomes in Mammalian Cells Grown In Vitro. *Stain Technology,* 33**,** 73-77.

SADDER, M. T., PONELIES, N., BORN, U. & WEBER, G. 2000. Physical localization of single-copy sequences on pachytene chromosomes in maize (*Zea mays* L.) by chromosome in situ suppression hybridization. *Genome,* 43**,** 1081-1083.

SAID, M., HŘIBOVÁ, E., DANILOVA, T. V., KARAFIÁTOVÁ, M., ČÍŽKOVÁ, J., FRIEBE, B., DOLEŽEL, J., GILL, B. S. & VRÁNA, J. 2018. The *Agropyron cristatum* karyotype, chromosome structure and cross-genome homoeology as revealed by fluorescence in situ hybridization with tandem repeats and wheat single-gene probes. *Theoretical and Applied Genetics,* 131**,** 2213-2227.

SALIH, R. H. M. 2017. *Nuclear and chloroplast genome diversity in apomictic microspecies of Taraxavum.* PhD Thesis, University of Leicester.

SANTOS, F. C., GUYOT, R., DO VALLE, C. B., CHIARI, L., TECHIO, V. H., HESLOP-HARRISON, P. & VANZELA, A. L. 2015. Chromosomal distribution and evolution of abundant retrotransposons in plants: gypsy elements in diploid and polyploid *Brachiaria* forage grasses. *Chromosome Res,* 23**,** 571-82.

SATHE, S. R. & MANEKAR, S. C. 2018. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience,* 7.

SATO 1949. Karyotype Alteration and Phylogeny, VI Karyotype analysis in Palmae. *CYTOLOGIA,* 14**,** 174-186.

SCHMIDT, T. 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant Molecular Biology,* 40**,** 903-910.

SCHMUTZ, J., CANNON, S. B., SCHLUETER, J., MA, J., MITROS, T., NELSON, W., HYTEN, D. L., SONG, Q., THELEN, J. J., CHENG, J., XU, D., HELLSTEN, U., MAY, G. D., YU, Y., SAKURAI, T., UMEZAWA, T., BHATTACHARYYA, M. K., SANDHU, D., VALLIYODAN, B., LINDQUIST, E., PETO, M., GRANT, D., SHU, S., GOODSTEIN, D., BARRY, K., FUTRELL-GRIGGS, M., ABERNATHY, B., DU, J., TIAN, Z., ZHU, L., GILL, N., JOSHI, T., LIBAULT, M., SETHURAMAN, A., ZHANG, X.-C., SHINOZAKI, K., NGUYEN, H. T., WING, R. A., CREGAN, P., SPECHT, J., GRIMWOOD, J., ROKHSAR, D., STACEY, G., SHOEMAKER, R. C. & JACKSON, S. A. 2010. Genome sequence of the palaeopolyploid soybean. *Nature,* 463**,** 178.

SCHNABLE, P. S., WARE, D., FULTON, R. S., STEIN, J. C., WEI, F., PASTERNAK, S., LIANG, C., ZHANG, J., FULTON, L., GRAVES, T. A., MINX, P., REILY, A. D., COURTNEY, L., KRUCHOWSKI, S. S., TOMLINSON, C., STRONG, C., DELEHAUNTY, K., FRONICK, C., COURTNEY, B., ROCK, S. M., BELTER, E., DU, F., KIM, K., ABBOTT, R. M., COTTON, M., LEVY, A., MARCHETTO, P., OCHOA, K., JACKSON, S. M., GILLAM, B., CHEN, W., YAN, L., HIGGINBOTHAM, J., CARDENAS, M., WALIGORSKI, J., APPLEBAUM, E., PHELPS, L., FALCONE, J., KANCHI, K., THANE, T., SCIMONE, A., THANE, N., HENKE, J., WANG, T., RUPPERT, J., SHAH, N., ROTTER, K., HODGES, J., INGENTHRON, E., CORDES, M., KOHLBERG, S., SGRO, J., DELGADO, B., MEAD, K., CHINWALLA, A., LEONARD, S., CROUSE, K., COLLURA, K., KUDRNA, D., CURRIE, J., HE, R., ANGELOVA, A., RAJASEKAR, S., MUELLER, T., LOMELI, R., SCARA, G., KO, A., DELANEY, K., WISSOTSKI, M., LOPEZ, G., CAMPOS, D., BRAIDOTTI, M., ASHLEY, E., GOLSER, W., KIM, H., LEE, S., LIN, J., DUJMIC, Z., KIM, W., TALAG, J., ZUCCOLO, A., FAN, C., SEBASTIAN, A., KRAMER, M., SPIEGEL, L., NASCIMENTO, L., ZUTAVERN, T., MILLER, B., AMBROISE, C., MULLER, S., SPOONER, W., NARECHANIA, A., REN, L., WEI, S., KUMARI, S., FAGA, B., LEVY, M. J., MCMAHAN, L., VAN BUREN, P., VAUGHN, M. W., et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science,* 326**,** 1112-5.

SCHWARZACHER, T. 2003. DNA, chromosomes, and in situ hybridization. *Genome,* 46**,** 953-962.

SCHWARZACHER, T., AMBROS, P. & SCHWEIZER, D. 1980. Application of Giemsa Banding to Orchid Karyotype Analysis. *Plant Systematics and Evolution,* Volume 134, 293.

SCHWARZACHER, T. & HESLOP-HARRISON, J. S. 1991. In situ hybridization to plant telomeres using synthetic oligomers. *Genome,* 34**,** 317-323.

SCHWARZACHER, T. & LEITCH, A. R. 1994. Enzymatic Treatment of Plant Material to Spread Chromosomes for In Situ Hybridization. *In:* ISAAC, P. G. (ed.) *Protocols for Nucleic Acid Analysis by Nonradioactive Probes.* Totowa, NJ: Humana Press.

SCHWARZACHER, T., LEITCH, A. R., BENNETT, M. D. & HESLOP-HARRISON, J. S. 1989. In Situ Localization of Parental Genomes in a Wide Hybrid. *Annals of Botany,* 64**,** 315-324.

SHARMA, A. K. & SHARKAR, S. K. 1956. CYTOLOGY OF DIFFERENT SPECIES OF PALMS AND ITS BEARING ON THE SOLUTION OF THE PROBLEMS OF PHYLOGENY AND SPECIATION *Genetica,* 28**,** 361-488.

SHEARER, L. A., ANDERSON, L. K., DE JONG, H., SMIT, S., GOICOECHEA, J. L., ROE, B. A., HUA, A., GIOVANNONI, J. J. & STACK, S. M. 2014. Fluorescence *In Situ* Hybridization and Optical Mapping to Correct Scaffold Arrangement in the Tomato Genome. *G3: Genes|Genomes|Genetics,* 4**,** 1395-1405.

SILJAK-YAKOVLEV, S., CERBAH, M., SARR, A., BENMALEK, S., BOUNAGA, N., COBA DE LA PENA, T. & BROWN, S. C. 1996. Chromosomal sex determination and heterochromatin structure in date palm. *Sexual Plant Reproduction,* 9**,** 127-132.

SINGH, R., LOW, E.-T. L., OOI, L. C.-L., ONG-ABDULLAH, M., NOOKIAH, R., TING, N.-C., MARJUNI, M., CHAN, P.-L., ITHNIN, M., MANAF, M. A. A., NAGAPPAN, J., CHAN, K.-L., ROSLI, R., HALIM, M. A., AZIZI, N., BUDIMAN, M. A., LAKEY, N., BACHER, B., VAN BRUNT, A., WANG, C., HOGAN, M., HE, D., MACDONALD, J. D., SMITH, S. W., ORDWAY, J. M., MARTIENSSEN, R. A. & SAMBANTHAMURTHI, R. 2014. The oil palm VIRESCENS gene controls fruit colour and encodes a R2R3-MYB. *Nature Communications,* 5**,** 4106.

SINGH, R., LOW, E.-T. L., OOI, L. C.-L., ONG-ABDULLAH, M., TING, N.-C., NAGAPPAN, J., NOOKIAH, R., AMIRUDDIN, M. D., ROSLI, R., MANAF, M. A. A., CHAN, K.-L., HALIM, M. A., AZIZI, N., LAKEY, N., SMITH, S. W., BUDIMAN, M. A., HOGAN, M., BACHER, B., VAN BRUNT, A., WANG, C., ORDWAY, J. M., SAMBANTHAMURTHI, R. & MARTIENSSEN, R. A. 2013. The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature,* 500**,** 340.

SINGH, R., ONG-ABDULLAH, M., LOW, E.-T. L., MANAF, M. A. A., ROSLI, R., NOOKIAH, R., OOI, L. C.-L., OOI, S. E., CHAN, K.-L., HALIM, M. A., AZIZI, N., NAGAPPAN, J., BACHER, B., LAKEY, N., SMITH, S. W., HE, D., HOGAN, M., BUDIMAN, M. A., LEE, E. K., DESALLE, R., KUDRNA, D., GOICOECHEA, J. L., WING, R. A., WILSON, R. K., FULTON, R. S., ORDWAY, J. M., MARTIENSSEN, R. A. & SAMBANTHAMURTHI, R. 2013. Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature,* 500**,** 335.

SMIT, A., HUBLEY, R., GREEN, P. & 2013-2015 RepeatMasker Open-4.0. http://www.repeatmasker.org

SOH, A. E., MAYES, S. E. & ROBERTS, J. E. (2017). *Oil Palm Breeding,* Boca Raton, CRC Press.

SONG, J., DONG, F., LILLY, J. W., STUPAR, R. M. & JIANG, J. 2001. Instability of bacterial artificial chromosome (BAC) clones containing tandemly repeated DNA sequences. *Genome,* 44**,** 463-9.

SPEICHER, M. R. & CARTER, N. P. 2005. The new cytogenetics: blurring the boundaries with molecular biology. *Nature Reviews Genetics,* 6**,** 782-792.

SPERBER, G., LÖVGREN, A., ERIKSSON, N.-E., BENACHENHOU, F. & BLOMBERG, J. 2009. RetroTector online, a rational tool for analysis of retroviral elements in small and medium size vertebrate genomic sequences. *BMC Bioinformatics,* 10**,** S4.

SPURBECK, J. L., ZINSMEISTER, A. R., MEYER, K. J. & JALAL, S. M. 1996. Dynamics of chromosome spreading. *Am J Med Genet,* 61**,** 387-93.

SRISAWAT, T., PATTANAPANYASAT, K. & DOLEZEL, J. 2012. Flow cytometric classification of oil palm cultivars. *African Journal of Biotechnology*.

SUN, H., DING, J., PIEDNOËL, M. & SCHNEEBERGER, K. 2017. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics,* 34**,** 550-557.

SUN, J., ZHANG, Z., ZONG, X., HUANG, S., LI, Z. & HAN, Y. 2013. A high-resolution cucumber cytogenetic map integrated with the genome assembly. *BMC Genomics,* 14**,** 461.

TANG, Z., YANG, Z. & FU, S. 2014. Oligonucleotides replacing the roles of repetitive sequences pAs1, pSc119.2, pTa-535, pTa71, CCS1, and pAWRC.1 for FISH analysis. *Journal of Applied Genetics,* 55**,** 313-318.

THE ARABIDOPSIS GENOME, I. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature,* 408**,** 796.

TING, N.-C., JANSEN, J., MAYES, S., MASSAWE, F., SAMBANTHAMURTHI, R., OOI, L. C.-L., CHIN, C. W., ARULANDOO, X., SENG, T.-Y., ALWEE, S. S. R. S., ITHNIN, M. & SINGH, R. 2014. High density SNP and SSR-based genetic maps of two independent oil palm hybrids. *BMC Genomics,* 15**,** 309.

TING, N.-C., MAYES, S., MASSAWE, F., SAMBANTHAMURTHI, R., JANSEN, J., SYED ALWEE, S. S. R., SENG, T.-Y., ITHNIN, M. & SINGH, R. 2018. Putative regulatory candidate genes for QTL linked to fruit traits in oil palm (*Elaeis guineensis* Jacq.). *Euphytica,* 214**,** 214.

TING, N.-C., YAAKUB, Z., KAMARUDDIN, K., MAYES, S., MASSAWE, F., SAMBANTHAMURTHI, R., JANSEN, J., LOW, L. E. T., ITHNIN, M., KUSHAIRI, A., ARULANDOO, X., ROSLI, R., CHAN, K.-L., AMIRUDDIN, N., SRITHARAN, K., LIM, C. C., NOOKIAH, R., AMIRUDDIN, M. D. & SINGH, R. 2016. Fine-mapping and cross-validation of QTLs linked to fatty acid composition in multiple independent interspecific crosses of oil palm. *BMC Genomics,* 17**,** 289.

TING, N.-C., ZAKI, N. M., ROSLI, R., LOW, E.-T. L., ITHNIN, M., CHEAH, S.-C., TAN, S.-G. & SINGH, R. 2010. SSR mining in oil palm EST database: application in oil palm germplasm diversity studies. *Journal of Genetics,* 89**,** 135-145.

TING, N.-C., ZAKI, N. M., ROSLI, R., LOW, E.-T. L., ITHNIN, M., CHEAH, S.-C., TAN, S.-G. & SINGH, R. 2010. SSR mining in oil palm EST database: application in oil palm germplasm diversity studies. *Journal of Genetics,* 89**,** 135-145.

TISNE, S., POMIES, V., RIOU, V., SYAHPUTRA, I., COCHARD, B. & DENIS, M. 2017. Identification of Ganoderma Disease Resistance Loci Using Natural Field Infection of an Oil Palm Multiparental Population. *G3 (Bethesda),* 7**,** 1683-1692.

TORRES, M. F., MATHEW, L. S., AHMED, I., AL-AZWANI, I. K., KRUEGER, R., RIVERA-NUÑEZ, D., MOHAMOUD, Y. A., CLARK, A. G., SUHRE, K. & MALEK, J. A. 2018. Genus-wide sequencing supports a two-locus model for sex-determination in *Phoenix*. *Nature Communications,* 9**,** 3969.

TÓTH, G., DEÁK, G., BARTA, E. & KISS, G. B. 2006. PLOTREP: a web tool for defragmentation and visual analysis of dispersed genomic repeats. *Nucleic acids research,* 34**,** W708-W713.

TREANGEN, T. J. & SALZBERG, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

UNITED NATIONS, D. O. E. A. S. A., POPULATION DIVISION 2017. World Population Prospects: The 2017 Revision, Key Findings and Advance Tables. Working Paper No. ESA/P/WP/248. New York: United Nation.

VITTE, C. & PANAUD, O. 2005. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenetic and Genome Research,* 110**,** 91-107.

WAMINAL, N. E., PELLERIN, R. J., KIM, N.-S., JAYAKODI, M., PARK, J. Y., YANG, T.-J. & KIM, H. H. 2018. Rapid and Efficient FISH using Pre-Labeled Oligomer Probes. *Scientific Reports,* 8**,** 8224.

WANG, G. X., HE, Q. Y., MACAS, J., NOVAK, P., NEUMANN, P., MENG, D. X., ZHAO, H., GUO, N., HAN, S., ZONG, M., JIN, W. W. & LIU, F. 2017. Karyotypes and Distribution of Tandem Repeat Sequences in *Brassica nigra* Determined by Fluorescence in situ Hybridization. *Cytogenet Genome Res,* 152**,** 158-165.

WANG, C.-J. R., HARPER, L. & CANDE, W. Z. 2006. High-Resolution Single-Copy Gene Fluorescence in Situ Hybridization and Its Use in the Construction of a Cytogenetic Map of Maize Chromosome 9. *The Plant Cell,* 18**,** 529-544.

WANG, G., LI, H., CHENG, Z. & JIN, W. 2013. A novel translocation event leads to a recombinant stable chromosome with interrupted centromeric domains in rice. *Chromosoma,* 122**,** 295-303.

WANG, N. & DAWE, R. K. 2018. Centromere Size and Its Relationship to Haploid Formation in Plants. *Molecular Plant,* 11**,** 398-406.

WEISS-SCHNEEWEIS, H., LEITCH, A. R., MCCANN, J., JANG, T. & MACAS, J. 2015. Employing next generation sequencing to explore the repeat landscape of the plant genome. *In:* HÖRANDL E, A. M. (ed.) *Next generation sequencing in plant systematics.* Königstein, Germany: Koeltz Scientific Books.

WESSELS BOER, J. G. 1965. *The indigenous palms of Surinam.* Thesis. Univ. of Utrecht.

WICKER, T., SABOT, F., HUA-VAN, A., BENNETZEN, J. L., CAPY, P., CHALHOUB, B., FLAVELL, A., LEROY, P., MORGANTE, M., PANAUD, O., PAUX, E., SANMIGUEL, P. & SCHULMAN, A. H. 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet,* 8**,** 973-82.

WILLIAMS, D., TRIMBLE, W. L., SHILTS, M., MEYER, F. & OCHMAN, H. 2013. Rapid quantification of sequence repeats to resolve the size, structure and contents of bacterial genomes. *BMC Genomics,* 14**,** 537.

WORLDOMETERS.INFO. 2019. Dover, Delaware, U.S.A. Available: http://www.worldometers.info/world-population/ [Accessed 01 April 2019].

WU, N., LI, M., SUN, H., CAO, Z., LIU, P., DING, T., XU, H., CHU, C., ZHUANG, L. & QI, Z. 2017. RNA-seq facilitates development of chromosome-specific markers and transfer of rye chromatin to wheat. *Molecular Breeding,* 38**,** 6.

XIAO, Y., XIA, W., MASON, A. S., CAO, Z., FAN, H., ZHANG, B., ZHANG, J., MA, Z., PENG, M. & HUANG, D. 2019. Genetic control of fatty acid composition in coconut (Cocos nucifera), African oil palm (*Elaeis guineensis*), and date palm (*Phoenix dactylifera*). *Planta,* 249**,** 333-350.

XIAO, Y., XU, P., FAN, H., BAUDOUIN, L., XIA, W., BOCS, S., XU, J., LI, Q., GUO, A., ZHOU, L., LI, J., WU, Y., MA, Z., ARMERO, A., ISSALI, A. E., LIU, N., PENG, M. & YANG, Y. 2017. The genome draft of coconut (*Cocos nucifera*). *GigaScience,* 6.

YAMADA, N. A., RECTOR, L. S., TSANG, P., CARR, E., SCHEFFER, A., SEDERBERG, M. C., ASTON, M. E., ACH, R. A., TSALENKO, A., SAMPAS, N., PETER, B., BRUHN, L. & BROTHMAN, A. R. 2011. Visualization of Fine-Scale Genomic Structure by Oligonucleotide-Based High-Resolution FISH. *Cytogenetic and Genome Research,* 132**,** 248-254.

YUNIS, J. & PRAKASH, O. 1982. The origin of man: a chromosomal pictorial legacy. *Science,* 215**,** 1525-1530.

ZAKI, N. M., SINGH, R., ROSLI, R. & ISMAIL, I. 2012. *Elaeis oleifera* genomic-SSR markers: exploitation in oil palm germplasm diversity and cross-amplification in arecaceae. *International journal of molecular sciences,* 13**,** 4069-4088.

ZHANG, L., YANG, X., TIAN, L., CHEN, L. & YU, W. 2016. Identification of peanut (Arachis hypogaea) chromosomes using a fluorescence in situ hybridization system reveals multiple hybridization events during tetraploid peanut formation. *New Phytologist,* 211**,** 1424-1439.

ZHAO, R., MIAO, H., SONG, W., CHEN, C. & ZHANG, H. 2018. Identification of sesame (*Sesamum indicum* L.) chromosomes using the BAC-FISH system. *Plant Biology,* 20**,** 85-92.

ZHOU, H. C., PELLERIN, R. J., WAMINAL, N. E., YANG, T.-J. & KIM, H. H. 2019. Pre-labelled oligo probe-FISH karyotype analyses of four Araliaceae species using rDNA and telomeric repeat. *Genes & Genomics*.

ZOHARY, D. & SPIEGEL-ROY, P. 1975. Beginnings of fruit growing in the old world. *Science,* 187**,** 319-27.

APPENDICES

APPENDIX 1 Single copy oligonucleotide (25 mer) from Eg5K_1p3

Oligo name	Sequence
Eg5K1p3a(F1)	CAAGTAAAATAGGCCATTAATCCCC
Eg5K1p3a(F2)	TCGGATAACTCCTAACCTCGAAACT
Eg5K1p3a(R1)	ACAACAATGAAGTAGGATGGCTGTT
Eg5K1p3a(R2)	AATCAGACTACTATTTCCAGCACTC
Eg5K1p3b(F1)	GACCCATCAGATGATCTGGAAAGCA
Eg5K1p3c(F1)	GCAAGTGGTGTTAAATGAACTTCTA
Eg5K1p3c(F2)	CAATTGAGACAGGCATTCTGATGTC
Eg5K1p3c(F3)	TGGTAATGGCCAATCAAGCAGGTCT
Eg5K1p3c(R1)	GTAATTGGATCCAAATGCTATGTGA
Eg5K1p3c(R2)	TATTGGCATCCCTCGAGAGATTGAC
Eg5K1p3c(R3)	TATGACTCCTATTAGGCAACTAGGA
Eg5K1p3d(F1)	GCTCCTGGAAATGTAACATAGAACA
Eg5K1p3d(R1)	AACAGTTAACACACACCTGGAATGC
Eg5K1p3e(R1)	ATAACTATTCAGCAGCATTCACAGC
Eg5K1p3f(F1)	CTCACTACTAACCAGATATCCTGGG
Eg5K1p3g(F1)	ACCATTTTTCTCTCCCTGTATGAGC
Eg5K1p3h(F1)	ATACATTGGCCATGTCCCATAAAGT
Eg5K1p3h(R1)	TAGACCAGCTGCTGCAGCGACTGGT
Eg5K1p3i(F1)	ATCACATTATTGTGCAACAGATCCC
Eg5K1p3i(F2)	GCATGATTTCTTGTAGGACTTCTTA
Eg5K1p3i(R1)	CAGATTAGCAGAGTCTTAGAGCAAT
Eg5K1p3i(R2)	TCATGAACCTCATGCTCCCAGAGGT
Eg5K1p3j(F1)	TTCTGCTATTGGCTTCCAGCAGGAG
Eg5K1p3j(F2)	GTTATCATCAATGGGCTTGTAATCA
Eg5K1p3j(R1)	TGATAAGAACAAGCAGAAATAACTG
Eg5K1p3k(F1)	GGTGCTAGCAACAACACTGCTTACG
Eg5K1p3k(F2)	GCTTCCGAATCTTGCTTTCATGACT
Eg5K1p3k(F3)	ATGACAGCAGGACTTGGTCCCATCT
Eg5K1p3k(R1)	GTTAAACGGCCGACAGGTTATCGCC
Eg5K1p3k(R2)	TCCTAGTTGCCTAATAGGAGTCATA
Eg5K1p3m(F1)	TAGGGGTATTTGATGACATGTGTTC
Eg5K1p3m(F2)	GTGTTGAGGATAAACACAACTTTGG
Eg5K1p3m(F3)	AGAAGAGCACTTGTTGTGGAAATGC
Eg5K1p3m(F4)	CTTTGAGCGCTCGATAGACTTGCTC
Eg5K1p3m(R1)	AGAACACATCAGCACCATAAATACC
Eg5K1p3m(R2)	TCCTGCAATCATGAGTTCAGCTTTC

APPENDIX 1 Single copy oligonucleotide (25 mer) from Eg5K 1p3

Oligo name	Sequence
Eg5K1p3m(R3)	CGGTTACACCTTATGCTGGTACACC
Eg5K1p3m(F1)	GCCTCGTAATTCTGGTGGAAGTTGG
Eg5K1p3m(F2)	TGATCATAATCACAGGTGGGATTGC
Eg5K1p3m(F3)	CAGAGAACATATGAAAGGACTTGCC
Eg5K1p3m(F4)	AACGATATGCAATCGCGTCTTTCCT
Eg5K1p3m(R1)	GTCTTAGACCTTACAACCAACACTA
Eg5K1p3m(R2)	ATGATAATCTCGGCGTAGAATGTAG
Eg5K1p3m(R3)	CATTTCCTGATCTGGTGCATCTGGC
Eg5K1p3n(F1)	TCCCTGGTAAACGGAGCCATCCACA
Eg5K1p3n(F2)	CCGAGTACGGAACATTAAATTGGAA
Eg5K1p3n(F3)	AGCAGCTACCAATAGAATGGCAAGT
Eg5K1p3n(F4)	GATATAGATTCCTCTCCGGTCGGCA
Eg5K1p3n(R1)	ATGCTTCAGAGGCTGCACGAAGATT
Eg5K1p3n(R2)	TGCCTTTGTCCACCATTACAATCCC
Eg5K1p3n(R3)	CGGCCTTTGAGCTACACGAGCCCAA
Eg5K1p3o(F1)	CCGGTGTTAACACAGTACGGAGGAT
Eg5K1p3o(F2)	TACTGGCGGAAGGAGATGCTAACGG
Eg5K1p3o(R1)	CCGATGCGAATAGGTAAAGTCGTCC
Eg5K1p3p(F1)	GCGTCGGCCGTGTGCGATCGGTGTG
Eg5K1p3p(R1)	TTCTCCGGGGTGAACTTGCAGTTGT
Eg5K1p3q(R1)	ACCTATTGTGTCGGTGAGCTTGGAT
Eg5K1p3r(F1)	TAACCAAGTTCACCCGTCGACATTC
Eg5K1p3r(R1)	AGCACCGAAGCCCAATCCTGCCTCA
Eg5K1p3s(R1)	GCTTCAAAGGTGGCCCACCCACTTC

APPENDIX 1 Single copy oligonucleotide (50 mer) from Eg5K_1p3

Oligo name	Sequence
Eg5K1p3a(F1)-50nt	GGCCATTAATCCCCCAGATTGTATGTTAATTTTACCGTGTGGTAGGTGCG
Eg5K1p3a(R1)-50nt	ACAACAATGAAGTAGGATGGCTGTTTGGCTCAACTTAACTGCCTCAATAT
Eg5K1p3b(F1)-50nt	TATGATGGTAGAAAAAGACCCATCAGATGATCTGGAAAGCAAGGATCAGG
Eg5K1p3c(F1)-50nt	GCAAGTGGTGTTAAATGAACTTCTACGAAGTGGTAATGGCCAATCAAGCA
Eg5K1p3c(F2)-50nt	CAGGCATTCTGATGTCCCTCTTTGATCTTCTCTGTATTTCATCAGCTGGA
Eg5K1p3c(R1)-50nt	TGCTTGATTGGCCATTACCACTTCGTAGAAGTTCATTTAACACCACTTGC
Eg5K1p3d(F1)-50nt	GCTCCTGGAAATGTAACATAGAACAGAATTTGCACCTCGCATTGTGTCCA
Eg5K1p3f(F1)-50nt	GAATGCTTCTCACTACTAACCAGATATCCTGGGTTGATGATTGTAGCACC
Eg5K1p3h(F1)-50nt	GGGTAATGAATACATTGGCCATGTCCCATAAAGTTTCTTTTATCACTATC
Eg5K1p3h(R1)-50nt	ATAATGAATCAATCAAACTCCCACATAGACCAGCTGCTGCAGCGACTGGT
Eg5K1p3i(F1)-50nt	ACATTATTGTGCAACAGATCCCCTGAATATTTACTTCAGAAGGAATCTGC
Eg5K1p3i(R1)-50nt	CATGAACCTCATGCTCCCAGAGGTGGCTGATCAACATGCAAAATCAGATT
Eg5K1p3j(F1)-50nt	GCAGGAGTTCATTAATTCGATAGTTCCAGAATTATTTATTTGATCGAGGG
Eg5K1p3j(R1)-50nt	TTCCCTCGTAATGCAAATAAGCAGAATTGGACTGCGCAACTACATCATGG
Eg5K1p3k(F1)-50nt	GGTGCTAGCAACAACACTGCTTACGTCTGTTGCGTGCTTGTACATCTTCT
Eg5K1p3k(F2)-50nt	CTTGATTCTCGGACAATAGCTACTGTACCGTCAACATGACAGCAGGACTT
Eg5K1p3k(R1)-50nt	TCCAGCAAGGAGCCTGGTCATTTACATATCACACTTTGGTACACGAACGCC
Eg5K1p3l(F1)-50nt	CTGGGGTATTTATGGTGCTGATGTGTTCTAGATGGTTGCTTGTACATGGG
Eg5K1p3l(F2)-50nt	GAAAGCTGAACTCATGATTGCAGGATGATGGAAAGATTGACATGCTGCAA
Eg5K1p3l(R1)-50nt	CGGTTACACCTTATGCTGGTACACCGAGCGGGATCATTTGTAGTGGAATT
Eg5K1p3m(F1)-50nt	TCACAGGTGGGATTGCAAGATTTAGAGAGAGCTGCAAGTTACCCAGAGAA
Eg5K1p3m(F2)-50nt	GCAAGTTGCAACGATATGCAATCGCGTCTTTCCTAAGAAAAGCACTTTTA
Eg5K1p3m(F3)-50nt	CGCCGAGATTATCATTTAACAAATTGGACGGGTTTTGCTCTCAATTTCCT
Eg5K1p3m(R1)-50nt	TCGTTATGAATGATTACCCTCATTTCCAACTTCCACCAGAATTACGAGGC
Eg5K1p3n(F1)-50nt	CCTCGTGCATGCCGTAATTGGCAAAAACCGAGTACGGAACATTAAATTGG
Eg5K1p3n(F2)-50nt	CCGATATAGATTCCTCTCCGGTCGGCAAACCTCGTGAGTTTCTAGTAGGT
Eg5K1p3n(F3)-50nt	TCGTGCAGCCTCTGAAGCATTTGGTCGAATGGGTGCTGGGGTGGTTACTA
Eg5K1p3n(R1)-50nt	TTGAGCTACACGAGCCCAAGAGCGCTCTCCACCGCTCGGCTCCGGTGAAA
Eg5K1p3o(F1)-50nt	CTCATCCCCGGTGTTAACACAGTACGGAGGATGGTGATATGGATTCAGCT
Eg5K1p3o(R1)-50nt	CCGATGCGAATAGGTAAAGTCGTCCAGAACAATCGGGTTCCTGGCATCCA
Eg5K1p3r(F1)-50nt	TAACCAAGTTCACCCGTCGACATTCACTGGAGAAGCCTCGAGATGGCCCC

APPENDIX 1 Oligonucleotide from 18S rDNA

Oligo name	Sequence
18s-1	ACCTGGTTGATCCTGCCAGTAGTCATATGCTTGTCTCAAAGATTA
18s-2(R)	CGGAAGTCGGGGTTTGTTGCACGTATTAGCTCTAGAATTACTACG
18s-3	CATGGTGGTGACGGGTGACGGAGAATTAGGGTTCGATTCCGGAGA
18s-4	TGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTT
18s-5(R)	TTGCTTTGAGCACTCTAATTTCTTCAAAGTAACGGCGCCGGAGGC
18s-6	AGCATTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTGGG
18s-7	CTGAGAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTT
18s-8(R)	GCGGCCCAGAACATCTAAGGGCATCACAGACCTGTTATTGCCTCA
18s-9	TCAGCTCGCGTTGACTACGTCCCTGCCCTTTGTACACACCGCCCG
18s-10	GGAAGGAGAAGTCGTAACAAGGTTTCCGTAGGTGAACCTGCGGAA