Gan, Siou Ting (2014) The development and application of molecular markers for linkage mapping and quantitative trait loci analysis of important agronomic traits in oil palm (Elaeis guineensis Jacq.). PhD thesis, University of Nottingham.

THE DEVELOPMENT AND APPLICATION OF MOLECULAR MARKERS FOR LINKAGE MAPPING AND QUANTITATIVE TRAIT LOCI ANALYSIS OF IMPORTANT AGRONOMIC TRAITS IN OIL PALM (*ELAEIS GUINEENSIS* JACQ.)

SIOU TING GAN

B.Sc (Hons) in Life Science (Biomedical Science)

Thesis submitted to the University of Nottingham

for the degree of Doctor of Philosophy

AUGUST 2014

ABSTRACT

Oil palm (*Elaeis guineensis*) produces over five times more oil/year/hectare than oil seed rape and accounted for 33% of world vegetable oil production in 2011. Being a cross-pollinated perennial tree crop with long breeding cycles (typically 12 years) and a large planting area requirement (usually 143 palms/hectare), utilization of molecular technology could greatly improve the efficiency of oil palm breeding. In the present study, various approaches were used to develop molecular markers for genetic linkage mapping and QTL analysis, with the ultimate goal of marker-assisted selection in oil palm.

Firstly, Representational Difference Analysis (RDA) and Amplified Fragment Length Polymorphism (AFLP) were coupled with Bulked Segregant Analysis (BSA) to try to identify marker(s) closely linked to the important shell-thickness gene. A novel combination of RDA with Roche 454 pyrosequencing enabled a more comprehensive study of the enrichment profiles compared to Sanger sequencing. Identification of >35% redundant sequences, repetitive sequences and organelle DNA suggested that subtractive hybridization and target enrichment of RDA were inefficient here, with the lack of elimination of common sequences masking the real difference products. The use of the AFLP method identified 29 primer pairs that yielded 49 putative shell-thickness related-polymorphic bands. A detailed analysis will need to be carried out to fully evaluate and validate these markers.

The use of the relatively new Diversity Array Technology "Genotyping-By-Sequencing" (DArTSeq) platform through genotyping of two closely-related *tenera* self-pollinated F_2 populations, 768 (n=44) and 769 (n=57), generated a total of 11,675

DArTSeq polymorphic markers of good quality. These markers were used in the construction of the first reported DArTSeq based high-density linkage maps for oil palm. Both genetic maps consist of 16 major independent linkage groups (total map length of 1874.8 and 1720.6 cM, with an average marker density of one marker every 1.33 and 1.62 cM, respectively), corresponding well with the 16 homologous chromosome pairs of oil palm ($2n = 2x = 32$; 14/16 chromosomes were confirmed by known location SSR markers). Preliminary quantitative trait loci (QTL) mapping of the yield and vegetative growth traits detected four significant and 34 putative as well as two significant and 30 putative QTLs for these small 768 and 769 populations, respectively. No common significant QTL were detected between the two closely-related controlled crosses which could have allowed combination of QTL across the two populations.

Saturation of the shell-thickness (*Sh*) region with all available DArTSeq markers, as well as map integration around the *Sh* regions for both populations, identified 32 Single Nucleotide Polymorphism (SNP) and DArT markers mapped within a 5 cM flanking region of the *Sh* gene. Homology search of the DArTSeq marker sequence tag (64 bp) against the recently published oil palm genome assembly confirmed that 23 out of the 32 (72%) DArTSeq markers were located on the p5_sc00060 scaffold in which the *SHELL* gene was identified. The identified shell-thickness markers could be useful as molecular screening tools. This study demonstrated the potential and feasibility of using genomic resources available for genetic improvement of oil palm breeding programmes.

ACKNOWLEDGEMENT

TABLE OF CONTENTS

Chapter 3: Approaches to develop Shell-thickness marker(s) using Representational Difference Analysis (RDA) and Next-Generation Sequencing (NGS)

Chapter 4: Approaches to develop Shell-thickness marker(s) using Amplified Fragment Length Polymorphism (AFLP)

Chapter 5: Development and characterisation of DArTSeq and SSR markers for genetic linkage mapping

LIST OF ABBREVIATIONS

AAR	Advanced Agriecological Research Sdn. Bhd.
AFLP	Amplified Fragment Length Polymorphism
AMD	Average marker density
APS	Ammonium persulfate
B	*Bam*HI restriction endonuclease
BC	Backcross
Bno	Bunch number
bp	Base pair
BSA	Bulked Segregant Analysis
Bwt	Bunch weight
CAPS	Cleaved Amplification Polymorphic Sequence
CIM	Composite interval mapping
CIRAD	Centre de Co-operation Internationale en Recherche Agronomique pour le Development, France
cM	CentiMorgan
co	Orphan contig
CP	Cross Pollinator
csRDA	Combined sample representational difference analysis
CTAB	Cetylrimethylammonium Bromide
CV	Coefficient of variation
D	DArT marker

D	*Dura*
DArT	Diversity Array Technology
DArTSeq	Diversity Array Technology "Genotyping-by-Sequencing"
DD	Deli *dura*
DH	Double Haploid
DIECA	Diethyldithiocarbamate sodium
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate (mix of dATP/dTTP/dCTP/dGTP)
DOA	Department of Agriculture
ds	Double-stranded
DWM	Dry to wet mesocarp ratio
E	*Eco*RI restriction endonuclease
EDTA	Ethylenediaminetetraacetic acid
EPPS	2-hydroxyethyl piperizine-N'-3-propene sulfonic acid
EST	Expressed sequence tag
EtOH	Ethanol
FA	Average frond area of frond 17
FB	Fruit to bunch ratio
FDW	Average frond dry weight of frond 17
FFB	Fresh fruit bunch yield
FL	Average frond length of frond 17
Fwt	Average fruit weight
Gb	Gigabase

GbS	Genotyping-by-Sequencing
GCA	General combining ability
GF	Number of green fronds
GW	Genome-wide
GWAS	Genome-wide association studies
H	*Hin*dIII restriction endonuclease
ha	Hectare
Ht	Stem height
IM	Interval Mapping
IPTG	Isopropyl-β-D-thiogalactoside
IRD	Infrared dye
K*	Kruskal-Wallis statistic
KF	Kernel to fruit ratio
K-W	Kruskal-Wallis non-parametric mapping
LAI	Leaf area index
LB	Luria-Bertani
LG	Linkage group
LOD	Logarithm of odds
M	Marker
M	*Mse*I restriction endonuclease
MAS	Marker-assisted selection
MF	Mesocarp to fruit ratio
MgCl$_2$	Magnesium chloride

MIM	Multiple interval mapping
ML	Map length
MPOB	Malaysian Palm Oil Board
MQM	Multiple QTL mapping
MRRS	Modified reciprocal recurrent selection
MRS	Modified recurrent selection
NaCl	Sodium chloride
NaoAc	Sodium Acetate
NCBI	National Centre for Biotechnology Information
NGS	Next-generation sequencing
NIL	Near isogenic line
ns	Not significant
OB	Oil to bunch ratio
OD	Optical density
ODM	Oil to dry mesocarp ratio
OP	Oil palm
OWM	Oil to wet mesocarp ratio
P	*Pisifera*
P	*Pst*I restriction endonuclease
PCR	Polymerase Chain Reaction
PLG	Pseudochromosome linkage group
PT	Permutation test
PTP	Picotiter plate

PVP	Plant variety protection
PVP-40	Polyvinylpyrrolidone-40
Q	Quality
QTL	Quantitative trait loci
RAPD	Random Amplification of Polymorphic DNA
RDA	Representational Difference Analysis
RFLP	Restriction Fragment Length Polymorphism
RIL	Recombinant inbred line
S	SNP marker
SCA	Specific combining ability
SD	Segregation distortion
SF	Shell to fruit ratio
Sh	Shell-thickness gene
SIM	Simple interval mapping
SLS	Sample loading solution
SOC	Super Optimal Broth
SNP	Single Nucleotide Polymorphism
sc	Scaffold
Ss	Single-stranded
SS	Size standard
SSR	Simple Sequence Repeats
T	*Tenera*
TAE	Tris-acetate-EDTA

TEMED	N, N, N', N'-Tetramethylethylenediamine
TM	Total number of markers for linkage group
UV	Ultraviolet
X-Gal	5-Bromo-4-chloro-3-indolyl ß-D-galactopyranoside

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1   The study background

Oil Palm (*Elaeis guineensis* Jacq) is the leading oil crop in the world with production of 45 million tonnes of palm oil in year 2009, constituting 27.5% of world production of oils and fats (MPOB, 2011). Palm oil is the largest internationally traded vegetable oil with India, China, the European Union and Pakistan as the major importers.

There are two species of oil palm, *Elaeis guineensis*, the African oil palm, and *Elaeis oleifera*, the American oil palm. The most cultivated species is *E. guineensis* due to its high oil yield. Even though the plant originated from Africa, it is mostly planted in Southeast Asia, particularly Indonesia and Malaysia, which account for more than 50% of the total oil palm plantation area in the world (MPOB, 2011). Palm oil is mainly used for human consumption (90%) and the remaining 10% is used in the oleochemical industry. With the increasing world population, oil palm with its high oil yield will play a key role in meeting future vegetable oil demands.

Plant breeding generally aims to improve the productivity of domestic crop plants. This process can involve time-consuming and costly processes of repeated backcrossing, self-pollination and progeny evaluation. The development of biotechnology since 1990s has revolutionized plant genetic research, particularly in the area of molecular breeding where conventional breeding is assisted by molecular markers (Rafalski and Tingey, 1993). Advances in molecular marker techniques have facilitated the assessment of genetic diversity, fingerprinting of varieties, linkage map construction, quantitative trait loci (QTL) mapping for desirable traits and marker-assisted selection (Collard *et al.*, 2005). It was shown that two to three cycles of marker-based selection in Maize in only 1

year had led to an average 9% improvement in grain yield (Johnson, 2004). Arguably, marker-assisted selection is of most advantage to perennial tree crops that have long breeding cycles, such as oil palm, than annual crops.

Conventional breeding of oil palm requires 19 years of phenotypic selection, which include 13 years of phenotypic evaluation of testcrosses, 3 years of inter-crossing of best palms to form the next cycle and another 3 years of palms maturity before they can be phenotypic evaluated in the next round (Wong and Bernardo, 2008). Large planting areas are required by oil palm, due to the bulky size of the palms, with standard planting density of 148 palms per hectare (usually hexagons of 9 m between palms) and smaller breeding plot size of 10-20 palms planted in 3-6 replicates are commonly used (Soh *et al.*, 1990). Molecular techniques that can save breeding time, space, cost and effort are very much sought after to improve the efficiency of oil palm breeding. Of these, saving time is the most important.

In view of the great potential of marker-assisted selection towards oil palm breeding, the present project aims to employ approaches to generate molecular markers, to perform genetic linkage mapping and QTL analysis of important agronomic traits Identification of molecular markers for traits of interest would enable early screening of traits using marker-assisted selection before the palms mature or before the trait is expressed, expediting oil palm breeding cycles. With this aim, the present project works on several economically important qualitative and quantitative traits, particularly the monogenic shell-thickness trait, as an example of markers development.

The shell-thickness trait is economically the most important trait in oil palm. Based on this shell-thickness gene, oil palm fruits can be divided into three categories, *dura* (D), *pisifera* (P) and *tenera* (T). Crossing of thick-shelled *dura* with the shell-less *pisifera* results in 100% thin-shelled hybrid *tenera* that contains a higher mesocarp-to-fruit ratio than the *dura* and this translates into higher oil yields. Genetic studies have revealed that the shell gene exhibits co-dominant monogenic inheritance that is exploitable in breeding programmes (Beirnaert and Vanderweyen, 1941; Singh *et al.*, 2013b). Currently, the product of D x P, commonly Deli *dura* x AVROS *pisifera*, is the most common commercial planting material (Soh *et al.*, 2006).

The identity of the fruit form can only be known when the plants start fruiting after 3-4 years of planting. High costs, space, time and effort are incurred during this process before identifying the fruit type within breeding programmes, where crosses can segregate for shell-type. A marker that can identify fruit form at the nursery stage would greatly facilitate and speed up the breeding and planting of wanted material, for example planting of vegetatively vigorous *pisifera* palms could be separated from the slower growing *dura* and *tenera* palms. Additionally, the oil palm industry has been facing an illegitimacy problem in which there is *dura* contamination in commercial D x P seeds due to poor quality control during controlled pollination and also the selling of fake hybrid seedlings by illegal seedling suppliers (Kushairi and Rajanaidu, 2000; Cheyns *et al.*, 2001). Markers can be incorporated into current procedures to identify, refine and even correct legitimacy issues.

In view of the monogenic inheritance of the shell-thickness gene, development of molecular marker(s) linked to this trait would be of great importance. Identified shell-thickness markers could be used to verify the identity and purity of commercial *tenera* hybrid. These markers can also be employed to authenticate the legitimacy of breeding crosses as well as potentially screening out *dura* and *pisifera* progeny from *dura* x *tenera*; *tenera* x *tenera* high value/critical crosses for field testing. The same goes for QTLs of economically important traits which can be utilized for marker-assisted selection in oil palm breeding programmes, saving cost, time and space.

Three different molecular marker approaches were used in the present project, namely Representational Difference Analysis (RDA), Amplified Fragment Length Polymorphism (AFLP) and Diversity Array Technology "Genotyping-by-Sequencing" (DArTSeq), to develop markers closely-linked with the shell-thickness trait. This is the first report on development of shell-thickness markers using RDA, single-enzyme AFLP and DArTSeq techniques.

The present project also aims to employ approaches for the construction of genetic linkage maps and QTL analysis in oil palm using two closely-related F_2 segregating populations, due to the small breeding plot sizes of oil palm breeding. The selection of two populations with full-sibs parents in the present study should allow some map integration for higher accuracy of marker order as well as testing the possibility of combining potential common QTL markers, if any, to increase the power and accuracy of QTL detection. The genetic maps were constructed using microsatellites and markers generated from DArTSeq platform. These genetic maps could in turn facilitate qualitative

and quantitative analysis of economically important traits, such as shell-thickness. The present work reported the first application of DArTSeq method for genetic linkage map construction and QTL analysis in oil palm as well as the construction of oil palm genetic maps using two closely-related populations.

A closely-linked and reliable shell-thickness marker had yet to be developed at the start of this project. Despite that *SHELL* gene was identified recently in July 2013 (Singh *et al.*, 2013b), the present study on shell-thickness trait is important to demonstrate the potential and feasibility of using biotechnology tool and genomic resources available for genetic improvement of oil palm breeding programmes.

1.2   Project overview and thesis structure
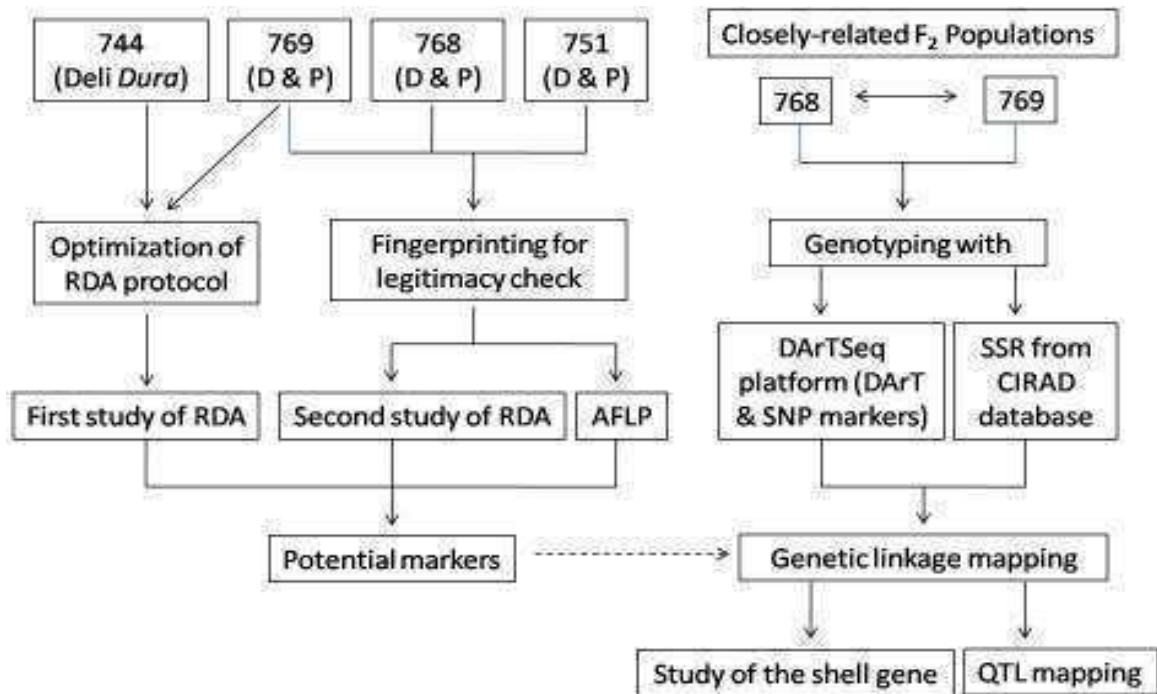
Figure 1.1: Project overview and workflow. D, *dura*; P, *pisifera*; RDA, Representational Difference Analysis; AFLP, Amplified Fragment Length Polymorphism; SSR, Simple Sequence Repeats; DArT, Diversity Arrays Technology; DArTSeq, DArT "Genotyping-by-sequencing"; SNP, Single Nucleotide Polymorphism.

Initially, two different molecular marker approaches were used, Representational Difference Analysis (RDA, chapter 3) and Amplified Fragment Length Polymorphism (AFLP, chapter 4), to identify markers closely linked to the shell-thickness gene. Four different oil palm controlled crosses, 744, 768, 769 and 751, were used. The 744 controlled cross is a self-pollinated Deli *dura* while the 768, 769 and 751 are non-Deli *tenera* selfed-pollinated populations producing all three different fruit types in their progeny, *dura*, *pisifera* and *tenera*. F_2 populations, 768 and 769, are closely-related as their *tenera* parents are from the same *Tenera* x *Pisifera* (T x P) cross of Binga x Yangambi AVROS whereas the *tenera* parent of the 751 was from a T x P cross of

Dumpy AVROS x Yangambi AVROS. The 751 controlled cross is related to the 768 and 769 such that their *pisifera* grandparents of Yangambi AVROS background are siblings of the same cross.

Plant materials used in the present study were made available at different times during the progress of the project. Deli *dura* of the 744 controlled cross as well as *dura* and *pisifera* of the 769 controlled cross were initially used to optimize the RDA protocol and for the first study of RDA (Chapter 3). Upon confirmation of the legitimacy of *dura* and *pisifera* progeny from the 768, 769 and 751 populations through fingerprinting, RDA was repeated in addition to the AFLP study reported in Chapter 4. Bulk Segregant Analysis (BSA) was previously used by Seng *et al.* (2007) to study the monogenic *Virescens* trait (fruit skin colour) in oil palm using the AFLP technique. In the present study, the BSA technique was used to generate distinct *dura* and *pisifera* bulks for study with the RDA and AFLP marker approaches. Only ten *dura* and *pisifera* palms were used for the construction of bulks for each controlled cross. Smaller bulks have higher frequency of false positives (Michelmore *et al.*, 1991) therefore the use of multiple bulks allowed identification of consistent markers to the shell-thickness trait.

Microsatellite markers publicly available from the Centre de Co-operation Internationale en Recherche Agronomique pour le Development (France, CIRAD) database were characterized and the DArT "Genotyping-by-sequencing" (DArTSeq) platform was employed to develop DArT and SNP markers using two closely-related F_2 populations, 768 and 769 (Chapter 5). Coupling of the DArT platform with Illumina short read sequencing in the DArTSeq approach generated both dominant DArT and co-

dominant SNP markers (Chapter 5). Chapter 6 reports the construction of the first high-density DArT- and SNP-based linkage maps of oil palm, one each for the 768 and 769 populations, with SSR markers as anchor markers. These anchor SSR markers were used to uniquely identify the chromosomes of oil palm and allow comparisons with the previous published studies (Chapter 6). QTL analysis of important yield traits, bunch components and vegetative growth traits was then conducted using the framework maps created from the DArT- and SNP-based linkage maps as reported in Chapter 7.

Chapter 8 reports on the identification of closely-linked shell-thickness markers through saturation of the shell-thickness region with DArTSeq markers and an integration of the individual maps. Potential marker(s) generated from the RDA and AFLP analyses could also be used to saturate the genetic linkage maps. However this part of work was not completed due to time constraints.

1.3 Objectives of the study

The objectives of this project are:

(i)   To test the RDA and AFLP marker approaches, targeting the shell thickness gene in the current material

(ii)   To develop and characterise DArTSeq (both DArT and SNP) markers as well as to utilise publicly available SSR markers to provide linkage between current and previously reported results

(iii)   To construct genetic linkage maps using two closely-related F_2 segregating populations using DArTSeq and SSR markers

(iv)   To perform QTL analysis of economically important traits, including fresh fruit bunch yields and its components, along with measures of vegetative growth for the mapped populations.

(v)   To perform qualitative analysis of shell-thickness gene of oil palm in an effort to develop markers useful within the breeding programme.

Chapter 2

Literature Review

2.1   Oil Palm

Oil palm, *Elaeis guineensis* Jacq, is an important tropical perennial oil crop and presently the most productive oil crop in the world per year per hectare. Oil palm belongs to the family *Arecaceae*, tribe *Cocoseae* and subtribe *Elaeidinae*. The genus name, *Elaeis,* is derived from the Greek word *elation*, meaning oil while the species name, *guineensis,* is attributed to the discovery of the tree by Jacquin in the Guinea coast (Jacquemard, 1998).

There are two species of oil palm (Corley and Tinker, 2003). The American oil palm, *Elaeis oleifera,* can be found in tropical countries of Central and South America while the African oil palm, *Elaeis guineensis*, is native to west and central Africa, in a region spanning ±10 ° latitude of the equator. It is postulated that both the African and American oil palm originated from Gondwanaland which disappeared when the American and African continents drifted apart in prehistoric times (Zeven, 1965).

2.1.1   The oil palm plant

Oil palm is a tall plant with an unbranched stem topped by 35-60 pinnate fronds. A mature palm can live up to more than 100 years, but, it is commonly grown for 25-30 years before being replanted as the palm becomes too tall to be harvested economically (Sambanthamurthi *et al.*, 2009).

Oil palm is monoecious with male and female flowers on separate inflorescences on the same plant. The life cycle of male and female flowers fluctuates between four and six months, varying according to genotype and environment (Purseglove, 1972). Every

month a palm can produce one to two fruit bunches (female phase) or one to two male inflorescences. Inflorescences start to appear when the palm reaches maturity at two to three years old (Soh *et al.*, 2003). Detailed investigation has however shown that each flower primordium consists of both male and female organs (Beirnaert, 1935). Sex differentiation occurs at the fourteenth month of florescence initiation in which stigmas are suppressed in the male flowers while stamens are underdeveloped in the female flowers (Hartley, 1988). When both organs develop fully, it gives rise to a hermaphrodite flower. These are common in young plants and during the transitional phase of the floral cycle.

Anthesis and receptivity of male and female inflorescences happens about two month after emergence (Soh *et al.*, 2009). Pollen is usually shed within five days after anthesis and remains viable up to six days. Pollen can be stored as oven-dried pollen in the freezer for about six months or as freeze-dried pollen in vacuum-sealed ampoules in the freezer up to 24 months (Soh *et al.*, 2003). Male inflorescence can produce large amount of pollen, about 30-40 g of pollen for each inflorescence (Rajanaidu *et al.*, 2000). Meanwhile, female flowers are receptive for 36-48 hours after anthesis with stigma exuding moisture to trap pollen grains (Latiff, 2000). The ratio of female to total inflorescences is defined as the sex ratio, an important factor in yield processes. Higher sex ratio indicates higher yield. Sex ratio diminishes with age; young palms can have a sex ratio as high as 98% and decrease to 35% in older palms (Latiff, 2000). Sex ratio is influenced by genetics and environmental factors such as fertiliser application, planting densities and availability of water (Broekmans, 1957; Corley, 1977 and Latiff, 2000).

The oil palm fruit is a sessile drupe varying in shape and length. The weight of each fruit varies from 3 to over 30 g. The fruit changes colour from dark purple or black to reddish-brown when ripe. It consists of three different layers, the soft oily mesocarp or pulp, the shell, and the endocarp or kernel (Figure 2.1). The seed is the remaining part after the mesocarp has been removed from the fruit; it contains only the shell and the kernel.
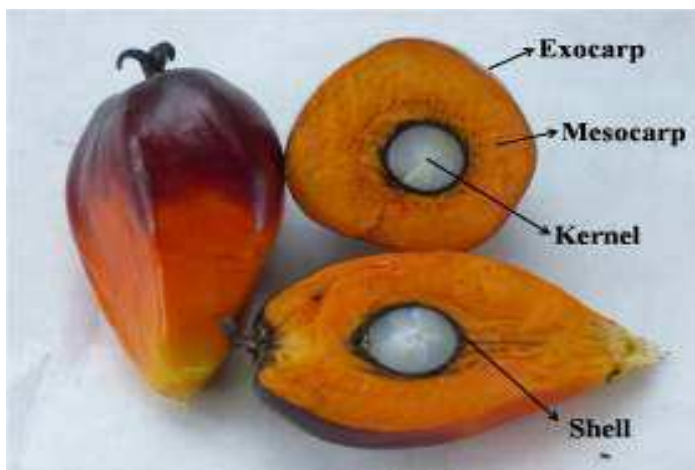
Figure 2.1: Picture of oil palm fruits and its different layers *(Source: Courtesy of Advanced Agriecological Research Sdn Bhd).*

2.1.2   The shell-thickness gene and fruit types

Oil palm is a diploid (2n=32) and has a genome size of around 1.8 billion base pairs (Bennett and Smith, 1991). It is believed that majority of the traits of agronomic importance are polygenic. Of the limited number of monogenic traits that have been identified, shell-thickness is the most important. Shell-thickness is controlled by a single locus, with two alleles, *Sh* and *sh*, showing co-dominant expression (Beirnaert and Vanderweyen, 1941) (Figure 2.2). Oil palms are classified into three fruit types, *dura*, *pisifera* and *tenera*, according to this trait.
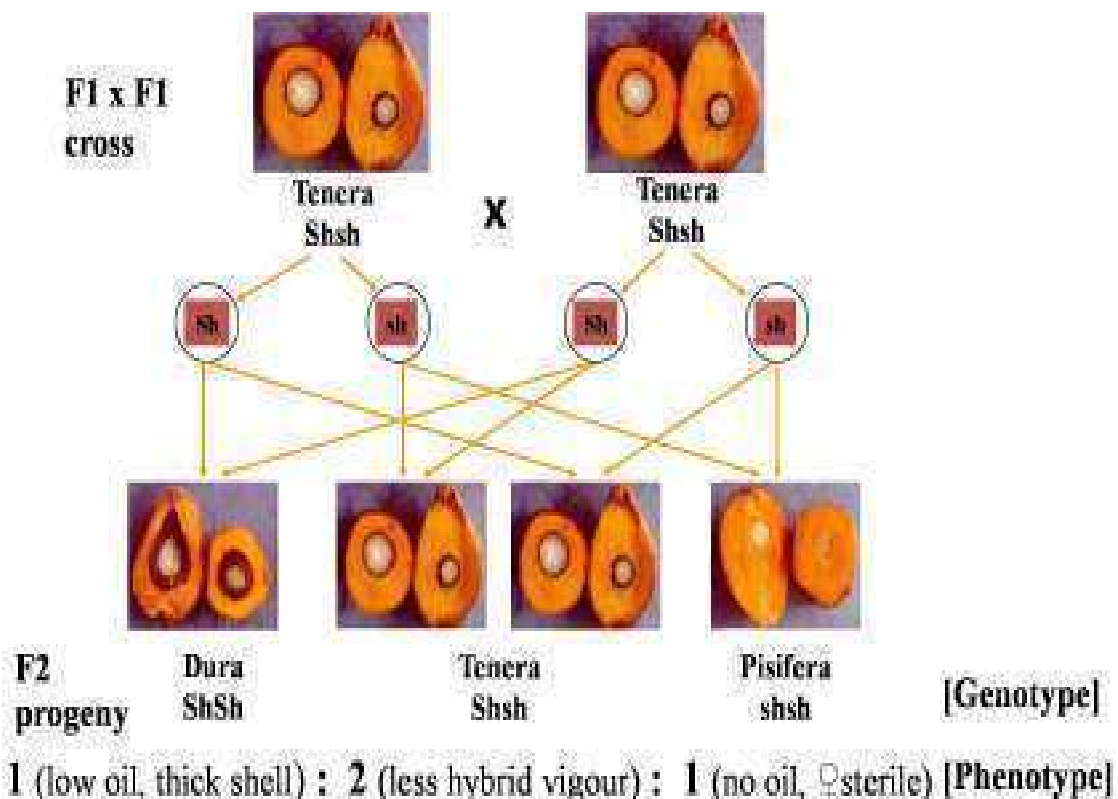
Figure 2.2: Monogenic inheritance of shell-thickness gene. *Tenera* x *Tenera crosses would give rise to segregating progenies in the classical Mendelian ratios, 1 Dura: 2 Tenera: 1 Pisifera (Source: Soh et al., 2009).*

Homozygous *dura* (*ShSh*) has fruits with a shell thickness of 2 to 8 mm and reduced oil-bearing mesocarp. *Pisifera* (*shsh*) with the absence of shell and mesocarp content of 95% might seem to be the ideal planting material. However, it is usually female sterile as its pistillate inflorescence tends to abort during development and hence cannot be used commercially or bred as a female parent. Heterozygous *tenera* (*Shsh*) is produced by the cross of *dura* as mother palm and *pisifera* as pollen donor. It produces fruit with thinner shell, varying from 0.5 to 4 mm, and a greater proportion of mesocarp. In *tenera* fruit, 30% of the shell in a *dura* is replaced by mesocarp which contributes to a 30% increase in oil yield compared to *dura* fruit (Corley and Lee, 1992). Therefore, it is the most commercially cultivated oil palm type. *In situ* hybridization of *sh* gene against

15

fruits between 1 to 5 weeks after anthesis (WAA) detected strong hybridization signal in outer layer of developing kernel of *dura* fruit, as opposed to weak signal in the mesocarp of both *dura* and *pisifera* fruits (Singh *et al.*, 2013b), indicating the earliest stages of shell formation. Characteristics of the three different fruit forms were presented by Sambanthamurthi *et al.* (2009) and are reproduced in Table 2.1.

Table 2.1: Characteristics of *dura*, *tenera* and *pisifera* fruit forms.

Fruit form characteristics	*Dura* (D)	*Tenera* (T)	*Pisifera* (P)
Shell thickness (mm)	2-8	0.5-4	Shell-less
Fiber ring	Absent	Present	Absent
Mesocarp to fruit ratio (%)	35-55	60-96	95
Kernel to fruit ratio (%)	7-20	3-15	
Oil to Bunch (%)	16	26	

(Source: Sambanthamurthi *et al.*, 2009)

Although shell-thickness is under monogenic control, the overlapping of the shell-thickness ranges in *dura* and *tenera* has lead to the postulation that the thickness of the shell is also modified by minor genes (van der Vossen, 1974). Okwuagwu and Okolo (1992, 1994) suggested that there is a kernel-inhibiting factor that is closely linked to the shell-thickness gene and mainly maternally-inherited. With the overlapping shell-thicknesses of *dura* and *tenera* fruits, the ultimate criteria for classifying fruit form is the presence of a fiber ring around the shell in *tenera* fruit, where mesocarp has formed, rather than shell.

2.1.3   Economic importance of oil palm

Among all vegetable oils, palm oil has the greatest versatility in terms of usage (Henderson and Osborne, 2000; Edem, 2002). Oil palm was exploited commercially at the beginning of the twentieth century, as a substitute to animal fat in the production of

candle wax, soap and margarine (Rival, 2007). Oil palm plantations were initiated by European colonists in Southeast Asia and Africa to ensure a steady supply of oil.

A unique property of oil palm lies in the fact that its fruit produces two types of oil, the orange-red palm oil from the mesocarp and the clear yellowish palm kernel oil (Soh *et al.*, 2003). Palm oil and palm kernel oil differ in their fatty acid composition, thus they have different uses. Palm oil and its refined derivates, olein and stearin, are the main commercial products of oil palm. The majority of palm oil is used as food (90%). Refining, fractionation and hydrogenation make a wide range of edible palm oil products available to the market (Yusof, 2007). Palm olein is mostly used in cooking oils, margarines and salad oils due to its low melting point, whereas stearin with its higher melting point is used for shortening, vanaspati and bakery oils (Corley and Tinker, 2003).

Additionally, 10% of palm oil production is used in the oleochemical industry. Carbon chains of palm oil are easily degradable when they enter the natural environment. Palm oil is an environmental friendly oleochemical substitute for mineral oils. Palm oil, as with the other vegetable oils, is also used as a biofuel (Corley and Tinker, 2003). The Malaysian Palm Oil Board (MPOB) has been working on palm oil methyl esters as a diesel fuel (Ong *et al.*, 1990; Choo and Cheah, 2000). The fuel works well with lower carbon emission. But its acceptability is very much dependent on the comparative price of petroleum oil and palm oil. Palm kernel oil is a competitor for coconut oil. Both are the main source of short-chain fatty acids in the world trade. Palm kernel oil is mostly used in the oleochemical industries.

Palm oil contains 50% saturated fatty acids which majority of them are palmitic and stearic that appear to be neutral in their cholestrolemic behaviour (Khosla and Sundram, 1996). Palm oil is being promoted as "balanced" oil (having equal saturated and unsaturated fatty acids) and *trans*-free, with many temperate oils requiring hydrogenation before they can be used in margarine: trans-fats being a side product of catalytic hydrogenation has been shown to be detrimental to health (Wahle and James, 1993; Ascherio, 2002). Palm oil is one of the 17 edible oils which were accepted and certified in meeting the FAO/WHO food standard requirement under the CODEX Alimentarius Commission Programme (CODEX Alimentarius vol. XI). Apart from that, palm oil also contains carotenes (pro-vitamin A), tocopherols and tocotrienols (Pro-vitamin E) that was found to have antioxidant, anti-cancer and cholesterol lowering effect (Nesaretnam *et al.*, 1988; Gutherie *et al.*, 1990; Gutherie *et al.*, 1997; Ong and Goh, 2002). All these beneficial health properties have added value to palm oil as edible oil.

2.1.4   Development of oil palm industry

From its origin in Central and West Africa, cultivation of oil palm has spread throughout Southeast Asia, Gulf of Guinea in Africa and tropical America with the current leading plantation areas in Malaysia and Indonesia. In 2007, the global area planted with oil palm reached 11 million hectare, with around 70% of the plantations belonging to smallholders (Rival, 2007).

In 1848, four African *E. guineensis* seedlings were planted in Bogor Botanic Gardens, Indonesia by the Dutch with two seedlings came from botanic garden in the Netherlands and another two from Mauritius (Hartley, 1988). However, progenies of the

four palms were quite similar, suggesting they may have originated from the same palm or region in West Africa.

Plantations in the East Indies (Indonesia, Malaysia) were started by European colonists using the progenies of the four palms, laying the foundation of the oil palm industry in Southeast Asia. The seedlings, which had thick-shelled fruits or *dura* (D) fruit form, were distributed to the plantations in Deli province in Sumatra and then to Malaysia (Rosenquist, 1986). From these, the uniform, high oil yielding Deli *dura* was developed as the commercial planting material from 1911 until the early 1960s.

Meanwhile in Africa, *dura* fruit were of poor quality and the natural occurrence of thin-shelled *tenera* had led to the early concentration on T x T material for commercial planting in the 1930s. However, by 1938 as much as 25% of sterile palms were found in these *tenera* commercial plantings. An examination of Beirnaert on 29,154 palms in Yangambi has shown that 24.3% were *pisifera*, not significantly different from the expected 25% segregation ratio of a single gene. Examination of D x T crosses showed that there were no *pisifera*, and majority of crosses gave segregation ratio close to 50:50 of *dura*: *tenera* progenies (Corley and Tinker, 2003). In 1956, Pichel reported that several hundred hectares of D x P crosses in Congo were 98% *tenera*, being the first large-scale confirmation of *tenera* production from D x P crosses.

With the understanding of the monogenic inheritance of shell gene by Beirnaert and Vanderweyen (1941), thin-shelled *tenera* with thicker mesocarp were brought into Southeast Asia from West Africa and rapidly became the essential planting material.

However Deli *dura* is still considered the best *dura* and hence widely used for seed production in breeding programmes (Rajanaidu *et al.*, 2000).

With the fast growth of world population and the increase in demand for vegetable oil, the oil palm industry has expanded tremendously over the years. Oil palm is currently the most productive oil crop with an average yield of around 4 tonnes of palm oil/hectare (ha)/year as compared to less than one tonne/ha/year for other oil crops, including soybean (*Glycine max*), sunflower (*Helianthus annuus*) and rapeseed (*Brassica napus*) (Yusof, 2007) (Table 2.2). World production of palm oil surpassed that of soybean oil in 2005 (Figure 2.3) (MPOB, 2011). In 2009, 45 million tonnes of palm oil were produced worldwide, constituting 27.5% of the total of world's oils and fats production (MPOB, 2011).

Table 2.2: Oil productivity of major oil crops in 2007.

Oil Crop	Production (million tonnes)	Average Oil yield (tonnes/ha/year)	Planted Area (million ha)	% of total area
Palm Oil	38.5	3.62	10.55	4.76
Soybean	36.96	0.40	94.15	42.52
Sunflower	10.78	0.46	23.91	10.80
Rapeseed	18.48	0.68	27.22	12.29

(Source: Lam *et al.*, 2009)
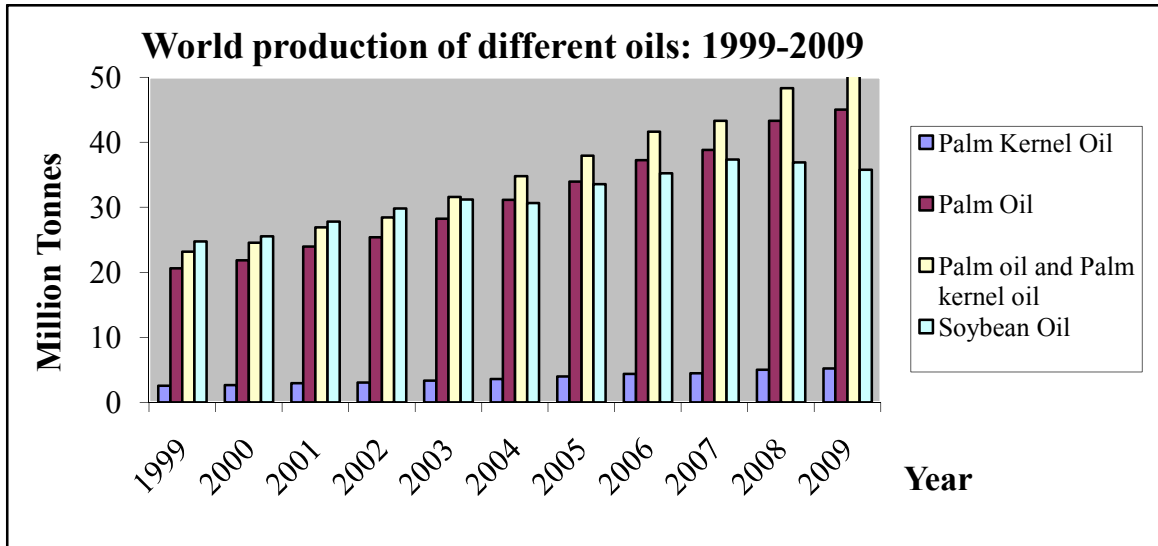
Figure 2.3: World production of palm oil, palm kernel oil and soybean oil from the years 1999 to 2009 [Source: Oil world Annual (1999-2009) and MPOB, 2011].

In terms of production cost, oil palm is the cheapest oil to produce among all vegetable oils, although it is very labour intensive with all harvesting currently carried out manually. For 2007, the production cost of palm oil was USD 228 per tonne in Malaysia compared to USD 400 for soybean oil in the United States and USD 648 of rapeseed oil in Canada, a price that was more than double that of palm oil (Lam *et al.*, 2009). This is further supported by the fact that oil palm has the highest output to input energy ratio, 9.6 to that of 2.5 and 3 for soybean and rapeseed, respectively (Figure 2.4) (Wood and Corley, 1991). The ratio of output to input energy generally gives an indication on how much energy is required (input energy, including fertilizer, milling and others) to produce a certain amount of energy (energy content in the oil). This means that oil palm requires less fertilizers, pesticides and fuel for machinery per unit production of oil when compared to soybean and rapeseed oil.
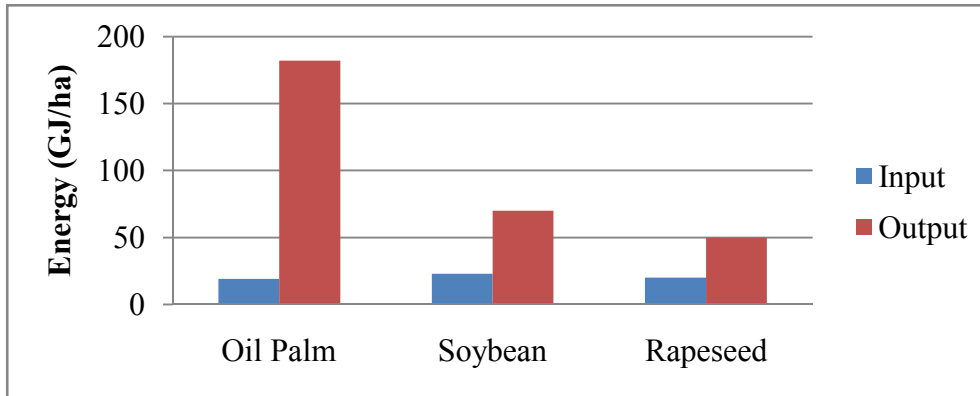
Output/Input	Oil Palm	Rapeseed	Soybean
GJ/ha	9.6	3.0	2.5

Figure 2.4: Comparison of energy output to input ratio of oil palm, soybean and rapeseed oil (Source: Wood and Corley, 1991; Lam *et al*., 2009).

Growing on less than 5% of the world's agricultural land and cheapest in term of production cost, oil palm, however, accounts for almost 28% of the global market share for edible oils production.

2.1.4.1 The oil palm industry in Malaysia

In 1953, Department of Agriculture (DOA) Malaysia started the first D x P plantings in the country by crossing the *pisifera* pollen imported from Nigeria with Deli *dura* to create D x P progenies (Kushairi *et al*., 1999). These progenies were found to perform better than those of Deli *dura* progenies. Planting of D x P materials were quickly adopted by local oil palm industry. Collaboration between DOA and Harrisons *&* Crossfield (now Golden Hope Plantations) found that AVROS *pisifera* from Sumatra had excellent general combining ability (GCA) with the Deli *dura* (Kushairi *et al*., 1999). Since then, the Deli *dura* x AVROS *pisifera* has become the most common planting material in Malaysia and worldwide.

After World Word II, tremendous development of the oil palm industry was achieved in Malaysia (Corley and Tinker, 2003). Today, the oil palm plantation area has increased from 54,000 hectares in 1960 to 4.85 million hectares, an expansion of more than 80-fold with Sabah the largest oil palm planted state, accounting for 29% of the total planted area (Yusof, 2007; MPOB, 2011). In the past few years, the Malaysian industry has started to look for joint ventures in other countries owing to the difficulty in finding suitable sites and the rapidly increasing cost of labour, together with a scarcity of plantation workers in Malaysia. This tremendous expansion of Malaysia's oil palm industry indicates the economic importance of this crop to the country as well as the growing world demand for palm oil.

At present, Malaysia is the world's predominant exporter of palm oil. It exports 45.8% of world needs, while Indonesia, the second largest palm oil exporter, exports 43% (MPOB, 2011). Even though Indonesia has grown to be the biggest world producer of palm oil in recent years, Malaysia would continue to be the leading exporter due to its lower domestic consumption compared to Indonesia.

In Malaysia, palm oil and palm oil-based products are the second largest exports revenue earner after electrical and electronic products; it has a total combined value of RM62.9 billion (approximately £12 billion) and contributed 9.8% to total exports in 2010 (Department of Statistics Malaysia, 2011). It is therefore clear that oil palm industry brings enormous revenue to the country and plays a crucial role in the economics of Malaysia.

2.1.5   Oil palm breeding

Oil palm is a naturally cross-pollinated perennial tree crop. Consequently, the industry has adapted breeding methodologies developed in maize as well as in animal breeding. Major oil palm breeding programmes adopt one of the two basic methods; the modified recurrent selection scheme (MRS) that is commonly practiced by programmes in the Far East influenced by Unilever plantations group and the modified reciprocal recurrent selection scheme (MRRS) that is practiced by programmes mainly in West Africa and Indonesia advised by CIRAD (Centre de Co-operation Internationale en Recherche Agronomique pour le Development, France) (Soh *et al.*, 2003; Soh *et al.*, 2009).

In the MRS, *duras* (D) are selected based on family and individual palm performances, thus the method is also called family and individual selection (FIS) (Rosenquist, 1990). *Pisiferas* (P) being female sterile are selected based on their *tenera* (T) sib performance in the T x T/P family. The selected Ps are then crossed with a number of selected Ds to form a D x P progeny test. If the mean performance of the D x P or T progenies from a P is high, the P is regarded to have a good GCA. This breeding scheme only emphasizes GCA effects, but not specific combining ability (SCA). Both the selected Ds and Ps are then used in commercial D x P seed production. The advantage of this scheme is that more recombinant crosses can be done within a shorter time, saving space and effort needed for extensive progeny-testings. The main disadvantage is that the D parents have not been progeny-tested and that the GCA effects expressed within the

parental D x D and T x T crosses may not be reflected in their D x P hybrid performance (Soh, 1999; Soh and Hor, 2000).

In MRRS, both the D and T parents are identified through the performance of their D x T progeny test. The D and T parents of the best individual crosses in the progeny testing are selfed or sibbed; the resulting D and P palms are used for commercial seed production. To save time, selfs and sibs of the parents are made and planted simultaneously as the progeny test crosses. This scheme exploits both the GCA and SCA effects. The main drawback is the requirement of large experimental areas to test the D x P crosses and selfs/sibs. To produce 3-4 million commercial seeds from the top 15% of crosses, about 500 crosses and 180 selfs have to be planted over 600 haectares and evaluated over 15-25 years (Soh, 1999).

Oil palm has a long generation cycle. The fruit bunch takes about 5 month after controlled pollination to develop and become ripe. Seed germination takes around 100-120 days, including a heat treatment of 40-60 days, followed by 10-12 months in the nursery. The plant will start to bear fruits after 2-3 years of transferring the seedlings to field (Mayes *et al.*, 2008). Only then the fruit type can be identified and recording of bunch yield and bunch analysis be initiated. As a result, from the date controlled pollination is performed, it takes at least 8 to 10 years before recording is complete and palms can be selected for next cycle of breeding.

Rival (2007) has highlighted several problems that oil palm breeding faces. These are (a) the duration of each generation and selection cycle (10-12 years) necessitating vast experimental areas; (b) limited knowledge of the genetic diversity and degree of

heterozygosity of the material tested; (c) the complex phenotypic expression of the desirable quantitative characters; and (d) the impossibility of determining at the nursery stage the variety (*dura/pisifera/tenera*) of individuals to be planted.

Confirmation of fruit form earlier at the nursery stage using molecular marker(s) for shell-thickness gene is important. For example, only 50% of *pisifera* are expected from a T x P crosses. If the identity of P can be confirmed at the nursery stage, only P may be transferred to the field, breeding and testing of P can be performed in a more focused and cost-effective manner.

2.1.6   The illegitimacy problem

In the oil palm industry, the confirmed parentage of the oil palms is of great importance, particularly in breeding programmes and seed production. This can be achieved via controlled pollination with pollen collected from the male inflorescence and crosses made on isolated female inflorescences with tight bagging to ensure no contamination. The identity of the resulting bunch and seeds need to be recorded clearly throughout the whole process of germination, growing in nursery and lastly transfer to the field (Corley and Tinker, 2003).

In the field, *dura* contamination in a commercial D x P cross can be kept below 1% with good quality control during controlled seed pollination. However, with the introduction of pollinating weevil, *Elaidobius kamerunicus*, from its natural home in Africa to the Far East in the early 1980s, illegitimacy became a severe problem with contamination rate as high as 20% (Rao *et al.*, 1994). Stricter controls have been

introduced to circumvent this problem, yet contamination still occurs occasionally (Kushairi and Rajanaidu, 2000).

In addition, there are also illegal seed suppliers in the market who sell fake 'high quality' seeds. Many smallholders still plant palms grown from unselected seed picked up from existing plantings. Report suggests that 40% of smallholders in Ivory Coast had planted unselected T x T seeds (Cheyns *et al.*, 2001). This is expected to give rise to 25% of sterile *pisifera* and 25% of *dura*, hence yield will be at least 30% below that expected from D x P material. The mixture of *dura* and *pisifera* seed also interferes with the efficiency of the extraction mills.

The fact that oil palm is a long duration crop and requires large planting areas of 148 palms per hectare (Soh *et al*. 1990), illegitimate seed could be devastating to the plantation companies. Contamination problems in any crosses cannot be identified until the palms start fruiting and the fruits form can be confirmed. This can lead to the loss of resources and/or reduced yield return, which is highly undesirable.

With the advent of molecular markers, it is possible to detect pollination errors in different crosses. A simple application of marker-assisted selection (MAS) is extremely valuable where individual progeny can be selected using genetic markers related to shell-thickness while they are still in the nursery, allowing only palms with correct fruit forms to be field-planted. Besides, this molecular marker can also be incorporated into any breeding programme where palms with desirable fruit form and other traits of interest can be selected. This can significantly reduce the time and planting cost as well as improve

resource allocation. Therefore identification of markers close to shell-thickness gene and other traits of interest is of important.

2.2   Molecular markers

Development of biotechnology tools has revolutionized plant research, particularly in the area of molecular breeding. Molecular breeding is a concept in which conventional breeding is assisted by molecular markers (Rafalski and Tingey, 1993). In order to achieve marker-assisted selection (MAS), the location or relative distances of a particular marker from the specific trait of interest can be determined by genetic mapping, assuming that the marker inherits different allelic forms that can be distinguished from the parents. The inheritance of these forms can be compared with the inheritance of the trait and strong association may allow the marker to act as a surrogate for the trait in future generations. The main advantage of MAS is that plant can be selected early, even before the trait of interest is expressed and this in turn can greatly reduce the time required to bring new varieties to the market (Mazur and Tingey, 1995).

Genetic markers represent genetic differences between individual organisms or species. Genetic markers can be divided into three major categories: (1) morphological (or phenotypic) markers; (2) biochemical (protein) markers; and (3) DNA-based molecular markers (Winter and Kahl, 1995). Morphological markers are visually detectable plant characteristics such as seed colour, shape and size, flower colour, growth habits or pigmentation. Protein markers are analysed as isozymes, allelic forms of enzymes, which can be separated by molecular weight or isoelectric point on electrophoresis gel. However, both of these markers have several general drawbacks.

These include: limited numbers of markers in most populations and dependence of phenotype or isoenzymes on environmental conditions or the development stage of the plant (Kunert *et al.*, 2001; Collard *et al.*, 2005).

DNA-based molecular markers are the most widely used markers due to their abundance. These markers are practically unlimited in number and are generally independent of environmental conditions, organ specificity and/or the developmental stage of the plant. Molecular markers usually do not have any biological effect and are transmitted by the standard laws of inheritance from one generation to the next. DNA markers are only useful when they can reveal the differences between individuals of the same or different species, termed polymorphism. Polymorphic markers can be further characterised as dominant or co-dominant. Dominant markers are either absent or present while co-dominant markers allow discrimination of homozygotes and heterozygotes (Collard *et al.*, 2005; Mondini *et al.*, 2009). There are currently many types of DNA-based molecular markers systems available for agricultural research, such as Restriction Fragment Length Polymorphism (RFLP; Beckman and Soller, 1986; Tanksley *et al.*, 1989; and Kochert, 1994), Random Amplification of Polymorphic DNA (RAPD; Welsh and McClelland, 1990; Williams *et al.*, 1990; Penner, 1996), Amplified Fragment Length Polymorphism (AFLP; Zabeau and Vos, 1993; Vos *et al.*, 1995); Simple Sequence Repeats (SSR)/microsatellites (Powell *et al.*, 1996; Taramino and Tingey, 1996; McCouch *et al.*, 1997) and Single Nucleotide Polymorphism (SNP; Wang *et al.*, 1998;Beutow *et al.*, 1999; Marth *et al.*, 1999). A comprehensive comparison of the advantages and disadvantages of these marker systems was presented by Mondini *et al.* (2009) and is reproduced in Table 2.3.

Table 2.3: Advantages and disadvantages of commonly-used DNA marker systems.

Molecular Markers	RFLP	RAPD	AFLP	SSR	CAPS	SCAR	IRAP/ REMAP	RAMP	SSCP	SNP	DArT	DArTSeq
Degree of polymorphism	M	M	M	H	L	M	M	M	L	H	M	H
Locus Specificity	Y	N	N	Y	Y	Y	N	N	Y	Y	N	Y
Dominance (D)/Co-dominance (C)	C	D	D	C	C	C	D	D	C	C	D	C
Ease of Replication	H	L	H	H	H	H	H	M	M	H	H	H
Abundance	H	H	H	M	L	L	H	M	L	H	M	H
Sequence information required	N	N	N	Y	Y	Y	Y	N	Y	Y	N	N
Quantity of DNA required	H	L	M	L	L	L	L	L	L	L	L	L
Automation	N	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y
Cost per assay	H	L	M	L/M	M	L	L	M	H	L	L	L
Technical requirement	H	L	M	L/M	H	M	H	H	H	M	M	M

Key: H = High; M = Medium; L = Low; Y = Yes; N = No; RFLP = Restriction Fragment Length Polymorphsim; RAPD = Random Amplification of Polymorphic DNA; AFLP = Amplified Fragment Length Polymorphism; SSR = Simple Sequence Repeats; CAPS = Cleaved Amplification Polymorphic Sequence; SCAR = Sequence Characterised Amplification Regions; IRAP = Inter-Retrotransposon Amplified Polymorphism; REMAP = Retrotransposon-Microsatellite Amplified Polymorphism; RAMP = Randomly Amplified Microsatellite Polymorphism; SSCP = Single Strand Conformation Polymorphism; SNP = Single Nucleotide Polymorphism; DArT = Diversity Array Technology; DArTSeq = DArT "Genotyping-by-Sequencing" (*Source: Mondini et al., 2009).*

2.2.1   Restriction Fragment Length Polymorphism (RFLP)

Restriction fragment length polymorphism (RFLP) was the first molecular marker applied to genome mapping projects (Botstein *et al.*, 1980). The technique relies on the ability of certain bacterial endonucleases to recognize and cleave DNA at specific 4-8 bases palindromic sequences, generating numerous fragments of various lengths, the number of which depends on the number of recognition sequences present in a given genome (Winter and Kahl, 1995). Restriction endonucleases are extremely sensitive to their recognition sequences in which even a single base change will completely abolish the recognition and cleavage of the DNA at that site. Therefore any DNA sequence change such as single nucleotide mutations, small insertions, deletions or DNA rearrangements within the recognition sequence will cause the loss or gain of restriction sites to generate RFLPs (Nguyen and Wu, 2005). In RFLPs, digested DNAs are separated by size using gel electrophoresis and they are then transferred to a membrane and detected by hybridizing the immobilised DNA with labelled and denatured DNA probes. The probes used for hybridization can be genomic DNA, cDNA or expressed sequence tag (ESTs) (Hoeltke *et al.*, 1995; Mansfield *et al.*, 1995).

RFLP markers are co-dominant, relatively high polymorphic and reproducible, thus transferable among laboratories. Despite its usefulness, this technique requires large amount of high quality DNA, depends on the development of probe, and is time consuming, labour-intensive and expensive (Mondini *et al.*, 2009). PCR-based RFLP technique, known as CAPS (cleaved amplification polymorphic sequence), has been introduced to improve the detection of RFLP, avoiding the time-consuming Southern

Blotting step and enabling rapid and high throughput analysis (Lyamichev *et al.*, 1993). However, each CAPS marker must be developed from known sequence.

Being the first generation of DNA-based molecular marker, RFLPs have contributed to the construction of genetic maps in plants (Helentjaris *et al.*, 1986, Chang *et al.*, 1988, McCouch *et al.*, 1988) as well as identification of markers linked to genes of interest (Sarfatti *et al.*, 1989; Barone *et al.*, 1990; Klein-Lankhorst *et al.*, 1991) in the early days. The first genetic map of oil palm was constructed by Mayes *et al.* (1996) using RFLP markers.

2.2.2   **Random Amplification of Polymorphic DNA (RAPD)**

The invention of the Polymerase Chain Reaction (PCR) by Mullis *et al.* (1986) has led to an exponential development of PCR-based molecular markers, the second generation of molecular markers. Random Amplification of Polymorphic DNA (RAPD) was the pioneer of PCR-based markers. The basis of RAPD is the PCR amplification of genomic DNA using short primers, usually 8-10 bp, of arbitrary sequence. RAPD detects DNA polymorphism produced by rearrangement or deletions at or between oligonucleotide primer binding sites in the genome (Williams *et al.*, 1990). The advantages of this technique are it requires no prior knowledge about the genome being analysed, can be employed across species using universal primers, minute amounts of DNA needed, highly polymorphic and can be easily detected on ethidium bromide-stained agarose gels. However the use of short primers has contributed to the major limitation of RAPD as short primers has relatively low annealing temperatures. This reproducibility problem happens not only among laboratories but also can happen within

a laboratory (Jones *et al.*, 1997; Penner *et al.*, 1993). RAPD markers cannot distinguish heterozygous and homozygous individuals, due to its dominant nature (Mondini *et al.*, 2009).

Nevertheless, the quick, simple and efficient nature of RAPD analysis has played a role in high density genetic mapping in many plant species (Kiss *et al.*, 1993; Torres *et al.*, 1993; Hemmat *et al.*, 1994) as well as marker identification for disease resistance genes (Martin *et al.*, 1991; Paran *et al.*, 1991; Adam-Blondon *et al.*, 1994).

2.2.3   **Amplified Fragment Length Polymorphism (AFLP)**

Amplified fragment length polymorphism (AFLP) was initially a DNA fingerprinting method developed by Vos *et al.* (1995). The technique is patented by Keygene NV (Wageningen, The Netherlands) (Zabeau and Vos, 1993). AFLP is a method that employs PCR-based selective amplification of restriction fragments from a digest of total genomic DNA. There are four basic steps in the AFLP technique: (i) digestion of extracted genomic DNA; (ii) ligation of oligonucleotide adaptors; (iii) PCR amplification of adaptors-ligated DNA fragments using primers which sub-sample the product pool of fragments available; and (iv) gel analysis of DNA fragments (Figure 2.5).
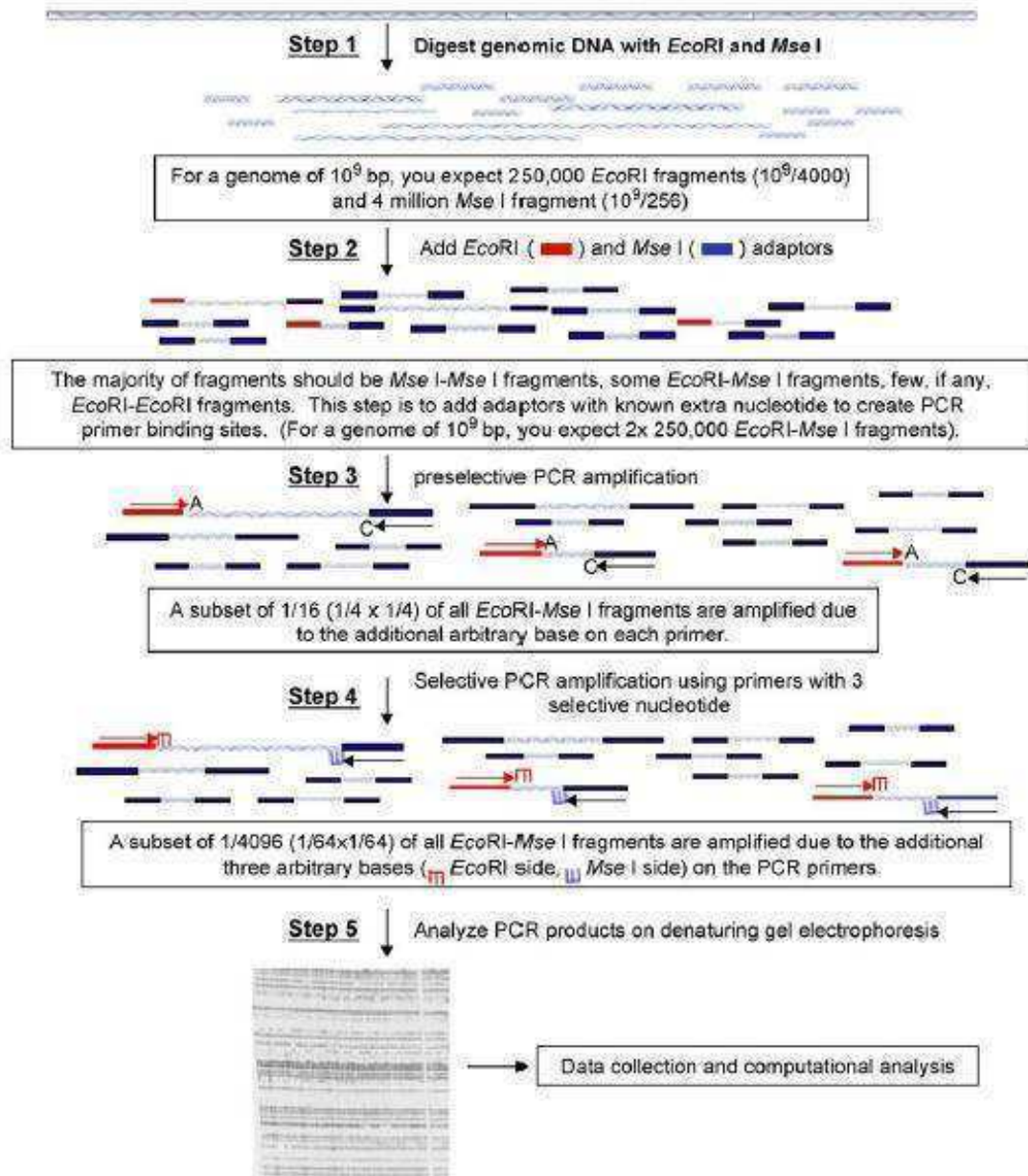
Figure 2.5: Schematic flow chart of the four basic steps of AFLP: digestion, ligation, amplification and gel analysis. *Genomic DNA is digested with two restriction endonucleases, example shown here is frequent cutter Mse*I *enzyme and rare cutter Eco*RI *enzyme. Adaptors are ligated to the restriction fragments with point mutation introduced into the adaptor sequences to prevent repeated digestion after ligation. Preselective amplification is performed using primers with one selective nucleotide at the 3'-end leading to 1/16 of the fragments being amplified, subsequent selective amplification using primers with two additional nucleotide at the 3'-end leads to amplification of 1/4096 of the fragments. PCR products are then resolved on polyacrylamide gel (Source: Liu and Cordes, 2004).*

In the classical AFLP analysis, genomic DNA is digested with two restriction enzymes, a rare-cutting enzyme with 6- to 8- base recognition in combination with a frequent-cutting enzyme of 4-base recognition. The high degree of specificity of restriction enzymes results in the production of a reproducible set of DNA fragments. Double-stranded adaptors (10-30 base pairs long) are ligated to the ends of DNA fragments using T4 DNA ligase. AFLP adaptors consist of a core sequence and an enzyme-specific sequence that is complementary to the sticky ends of the corresponding restriction site. AFLP adaptors are designed in such a way that initial restriction site is not restored after ligation, allowing simultaneous restriction and ligation. With these reactions performed in the same tube, any fragment-to-fragment product is restricted while adaptor-to-adaptor ligation is prevented as the adaptors are not phosphorylated. These two features ensure that adaptors are ligated to virtually all restriction fragments (Blears *et al.*, 1998).

Selective amplification of DNA fragments is achieved using primers complementary to the adaptor and restriction site sequence with additional selective nucleotides at their 3'-end. Only template fragments with complementary nucleotides extending beyond the restriction site will be amplified under stringent annealing conditions. Therefore only a subset of all possible templates is amplified and the number of amplified fragment is reduced approximately four-fold with each additional selective nucleotide, assuming a random base distribution (Vos *et al.*, 1995). The length and nature of the base extension on the 3'-end of the primers can be manipulated to generate fingerprints of the desired complexity. The nucleotide extensions at the 3-end of the primers serve two purposes: (i) a variety of restriction fragment subsets can be amplified;

and (ii) additional possibilities of polymorphism can be detected beyond the restriction itself (Blears *et al.,* 1998).

Two-step amplification is generally involved in AFLP fingerprinting analysis of complex genomes (10^8-10^9 bp). The first PCR amplification, named pre-amplification, is performed with primers having a single or no selective nucleotide and these primers are not radioactively or fluorescently-labelled. PCR products from pre-amplification are diluted and used as templates for the second amplification reaction using primers having more selective extensions. Pre-amplification reduces the overall complexity of mixture up to 16-fold if a single selective base is used on each primer, reducing the background noise. Selective amplification with three selective bases in each primer leads to final amplification of only 1/4096 of fragments in the mixure (Blears *et al.*, 1998; Liu and Cordes, 2004).

Polymorphism can be detected by electrophoretic separation of the amplified fragment on a denaturing polyacrylamide gel. Typically, the primer corresponding to the rare-cutter will be labelled radioactively or with a fluorescent dye to allow detection of the amplified fragments (Vos *et al.*, 1995). Polymorphism happens when there are (i) mutations in the restriction sites, (ii) mutations in the sequences adjacent to the restriction sites and complementary to the selective primer extensions, and (iii) insertions or deletions within the amplified fragments (Savelkoul *et al.*, 1999).

The AFLP technique offers several advantages. This technique can be applied to any DNA samples regardless of the origin and complexity with no prior sequence information needed. Only a small quantity of genomic DNA is needed (1-2 μg) and it is

found to be insensitive to template concentration (Vos *et al.*, 1995). The AFLP technique is reliable and robust as stringent conditions are used for primer annealing; this technique combines the reliability of RFLP with the power of the PCR technique. AFLP is easier to perform than RFLP as tedious manipulations of southern blotting hybridization used for RFLP studies is not needed for AFLP (Valsangiacomo *et al.*, 1995). The markers produced by AFLP technique are also reliable and reproducible within and between laboratories (Blears *et al.*, 1998). In plants, AFLP analysis has been found to be more informative than RAPD and RFLP analysis (Powell *et al.*, 1996; Russell *et al.*, 1997).

Despite its high reproducibility and reliability, the AFLP technique has its own disadvantages. In general, the polymorphism level of AFLP is lower than that of other molecular techniques, such as RFLP and SSR. However, AFLP markers show the highest marker effective multiplex ratio; the ability to analyse a large number of polymorphic loci simultaneously (Ridout and Donini, 1999; Varshney *et al.*, 2007). AFLP can produce up to four times more polymorphic loci per primer combination than RAPD, RFLP and SSR system (Mba and Tohme, 2005). This characteristic feature of AFLP marker system confirms the highly informative value of the technique. Another major disadvantage of AFLP is scoring of presence and absence of AFLP bands yields dominant markers (Mba and Tohme, 2005). Nevertheless, AFLP fragments of the same size from different individuals that show obvious differences in intensity can be quantified to be co-dominant markers (Meudt and Clarke, 2007). Intensity differences are predicted to be positively correlated with allelic copy number (van Eck *et al.*, 1995; Piepho and Koch, 2000). Special software is required for accurate quantitation of band intensity to differentiate homozygotic and heterozygotic for co-dominat scoring (Savelkoul *et al.*,

1999; Meudt and Clarke, 2007) and poorly amplified samples could give incorrect calls between one or two copies.

Since its publication in 1995, AFLP has been used extensively in many studies in plants and, more recently, for animals, fungi and bacteria, spanning numerous disciplines in genetics, evolution and ecology (Meudt and Clarke, 2007). AFLP has been used for DNA fingerprinting of prokaryotes and eukaryotes, construction of high-density molecular linkage maps as well as genetic diversity studies in animals and plants, and positional cloning of genes of interest. AFLP is also useful for characterisation and strain identification of bacteria and fungi, as well as eukaryotic pathogens of plants and animals (Blears *et al.*, 1998; Savelkoul *et al.*, 1999; Meudt and Clarke, 2007). In plants, the AFLP technique has four major applications, (i) genetic analysis and variety identification; (ii) germplasm management; (iii) indirect selection of agronomically important traits; and (iv) marker-assisted backcross breeding.

2.2.4   Simple Sequence Repeat (SSR)

Simple Sequence Repeat (SSR) (Tautz *et al.*, 1986), also known as microsatellites (Litt and Luty, 1989), are tandem repeats of short DNA sequence, typically 1-6 bases in length, that are widespread in all prokaryotic and eukaryotic genomes analysed to date (Zane *et al*, 2002). Plant genomes are rich in AT repeats whereas the AC repeat is more common in animal genomes (Lagercrantz *et al.*, 1993). Microsatellites are present in both coding and noncoding regions and distributed throughout the nuclear genome. Besides nuclear SSR, there are also mitochondrial (mtSSR) (Soranzo *et al.*, 2001; Rahendrakumar *et al.*, 2007) and chloroplastic SSRs (cpSSR) (Provan *et al.*, 2001; Chung

et al., 2006). In plants, microsatellites frequency is inversely correlated with the genome size, but the percentage of repetitive DNA was reported to be the same in coding regions (Morgante *et al.*, 2002)

SSR polymorphism mainly derives from variability in amplified fragment length, which in turn depends on the number of repeat units contained by alleles at a given locus. SSR are assayed by PCR amplification using the unique sequences of flanking region as primers. The PCR protocol of SSR can employ either unlabelled primer pairs or primer pairs with one labelled primer. Analysis of unlabelled PCR products can be carried out using 3% agarose gel stained with ethidium bromide or on a 6% polyacrylamide gel. Concentrated agarose gel is only suitable for SSR PCR products that differ in size by at least 10 bp. Polyacrylamide gel electrophoresis using labelled primer or silver staining of unlabelled products is more suitable for detection of polymorphism less than 10 bp (Nguyen and Wu, 2005), with a thin (0.4 mm) 6% denaturing polyacrylamide sequencing gel having the greatest resolution (1-2 bp). Fluorescence dyes offer several advantages over other labelling methods which include longer shelf life of fluorescent compound than radioisotopes, safer and easier to handle as well as faster detection and higher sensitivity (Nguyen and Wu, 2005). Multiplexing can be achieved using different markers with different dyes analysed in a single gel lane and/or PCR products of different size with the same dye analyzed in the same lane. Nevertheless, it can be very costly to fluorescently-label one of the primers in all primer pairs and a 'poor-man' approach has been introduced (Schuelke, 2000). In this approach, the forward primer is designed such that the M13 sequence is added at the 5'-end and the fluorescent dye is incorporated separately into a 'M13 primer'. The fluorescently-labelled M13 primer, the forward

primer with the M13 tail and reverse primer are all added into a single PCR reaction. The tagged forward and normal reverse primers are incorporated to give the specific product, before the limited forward-tag primer runs out. The PCR reaction is continued by the M13 primer which now primes the reaction from the 5'-M13 tag already incorporated, incorporating fluorescent dye at the same time into the product. With this technique, one single fluorescent dye-labelled M13 primer can be used for all primer pairs, an inexpensive alternative especially beneficial to small research groups. Therefore, fluorescently-based genotyping system using a sequence analyser has been widely adopted in recent years.

SSRs are a highly popular molecular marker in most areas of molecular genetics due to their co-dominant inheritance, ubiquitous occurrence, multi-allelic nature, high reproducibility, small locus size, ease of accessing size variation through PCR with flanking primers and the requirement of low amounts of DNA. SSRs are also excellent markers for fluorescent techniques, multiplexing and easily automated for high throughput screening (Agarwal *et al.*, 2008). However the development of microsatellites is tedious, costly and requires extensive knowledge of DNA sequence information which can be of major obstacle for the majority of species, particularly minor, underutilized crop species. Large scale isolation of microsatellites in plant is also cumbersome due to relatively low frequency of microsatellites in plant genomes compared to animal genomes (Powell *et al.*, 1996). EST projects of several plant species for gene discovery have generated a wealth of publicly available sequence data; these data can be utilized for identification of SSRs, referred to as EST-SSRs. Generation of EST-SSRs is relatively easy and cheap, although limited to those species or close relatives for which there is

sufficient number of ESTs available (Varshney *et al.*, 2005a). This is particularly true with the development of Next Generation Sequencing (NGS) to generate transcriptome data (Zalapa *et al.*, 2012).

Among all the different molecular techniques, SSR has been the most extensively exploited class of markers. The advent of new technologies has not affected the use of SSR. Application of microsatellite in plants can be categorized into 4 groups, (i) genome mapping; (ii) cultivar identification and marker-assisted selection; (iii) genetic diversity and phylogenetic studies; and (iv) population and evolutionary studies (Wang *et al.*, 2009; Kalia *et al.*, 2011).

Genome mapping is an area where SSR are heavily exploited. Together with other marker systems, SSRs have been applied for genetic mapping of many different plant species, including trees, major and minor crops, fruits and vegetables, ornamentals and turf grass (Wang *et al.*, 2009). Comparative mapping using SSR has been successfully performed in many plant species and this facilitates our understanding of evolutionary processes as well as identification of "linkage block" and gene syntenies which will in turn lead to DNA markers development for marker-assisted selection and/or cross species homologous cloning (Wang *et al.*, 2009; Kalia *et al.*, 2011). SSR markers have also been used to anchor and construct physical maps of soybean (Shultz *et al.*, 2007; Shoemaker *et al.*, 2008) and Arabidopsis (Wang *et al.*, 1997) which is important for assembly of genome DNA sequences and positional cloning.

Transferability of SSR markers to related species, particularly from major species to minor species, has enabled construction of genetic maps in minor crops and

improvement of chromosomal regions as well as study of genetic diversity and phylogenetic relationship of many minor species (Wang *et al.*, 2009). SSR markers from barley have been employed for generation of genetic linkage map for rye and wheat (Varshney *et al.*, 2005b) whereas the genetic diversity of USDA *Lespedeza* germplasm and its phylogenetic relationship with the genus *Kummerowia* were assessed using SSR markers derived from *Medicago*, cowpea (*Vigna unguiculata*) and soybean (Wang *et al.*, 2009).

Microsatellites markers have been used in many areas of oil palm study. Billotte *et al.* (2001) first reported the development and characterisation of microsatellite markers from oil palm and the use of these SSR for genetic diversity of the genus *Elaeis* as well as phylogenetic studies across palm taxa. From there, the same research group published the first high density map of oil palm using SSR markers from oil palm and coconut (Billotte *et al.*, 2005) and performed QTL analysis on SSR-based multi-parent linkage mapping in oil palm (Billotte *et al.*, 2010).

SSR markers have also been employed for genome analysis and DNA fingerprinting of oil palm tissue culture clones as a means of quality control (Singh *et al.*, 2007). Development of EST-databases has further enabled the mining of SSRs. SSR markers derived from a small collection of ESTs have been shown to be useful for genetic analysis of *E. guineensis* germplasm (Singh *et al.*, 2008a). Further work on a larger collection of ESTs has also developed more informative SSR for genetic diversity studies between *E. guineensis* and *E. oleifera* germplasm as well as transferability across palm taxa (Ting *et al.*, 2010). Recent publications also included SSR derived from tissue

culture ESTs (Low *et al.*, 2008) and ESTs from cDNA libraries of developing vegetative and reproductive tissues (Tranbarger *et al.*, 2012).

2.2.5   **Single Nucleotide Polymorphism (SNP)**

Single nucleotide variation that occurs in the genome sequence of individuals of a population is known as SNP. SNPs are the new generation of markers. They are the most abundant molecular markers in any organism; they can reveal hidden polymorphism that cannot be detected by other markers and methods. SNPs are widely distributed throughout genomes although they are more widespread in the non-coding regions of the genome (Mondini *et al.*, 2009). Theoretically, SNP can produce up to four alleles, containing either one of the four bases, A, T, C, and G. Practically, bi-allelic SNPs are more prevalent and most often it is either the two pyrimidines C/T or the two purines A/G (Kahl *et al.*, 2005). Although the Polymorphic Information Content (PIC) of SNP is not as high as multi-allelic microsatellites, this limitation is balanced by their sheer number; therefore SNPs will be marker of choice in future.

Various methodologies have been applied for discovery of SNPs in the plant genome. These include discovery of SNPs from EST libraries, array analysis, re-sequencing of PCR amplicons and also using Next-Generation Sequencing (NGS) approach (Ganal *et al.*, 2009). Numerous EST databases have been generated for many plant species. SNPs are a free by-product from these expanding databases with SNP being screened using bioinformatics tools, although they also need to be validated in the lab. The low quality of EST sequences has impeded the identification of true SNPs with validation rate published so far of only 50-85% (Batley *et al.*, 2003; Yamamoto *et al.*,

2005). The basis of SNP identification using arrays with short oligonucleotides is that they very sensitive to sequence variation, especially when the variations are located in the middle of the oligonucleotide. This approach can cover many genes (1,000-20,000) in one go without expression level bias but the false discovery rate is very high, 25-50% (Ganal *et al.*, 2009).

Amplicon resequencing is the most direct way of SNP discovery. It involves PCR amplication of DNA fragments from several lines. The PCR products are then fully sequenced and the resulting sequences are aligned and compared for SNP identification (Rafalski, 2002). SNPs identified using this approach is highly reliable with a false rate of less than 5% (Ganal *et al.*, 2009). The limitations of this technique are mainly the cost and it is a tedious technique. Advancement in sequencing technology has accelerated the discovery of SNPs. SNPs can be directly mined from sequenced genome but only applied to major crops that have been fully sequenced, such as Arabidopsis (The Arabidopsis Genome Initiative, 2000), rice (*Oryza sativa*; International Rice Genome Sequencing Project, 2005), maize (*Zea mays*; Schnable *et al.,* 2009), soybean (Schmutz *et al.*, 2010), to name a few, with the latest addition of tomato (*Solanum lycopersicum*; The Tomato Genome Consortium, 2012).

Over the years, SNPs technology has advanced and been employed extensively in the field of human and animal genetics study. However, the research into SNPs in plant genomes has been slower and mostly focused on major crops that are economically more important. SNPs are useful for high density genetic mapping, QTL analysis, association studies, germplasm characterisation, molecular breeding and population studies in plants

(Rafalski, 2002). With the falling costs and increased accessibility of genotyping technologies, SNPs markers have also been utilized to expand the resolution and throughput of genetic analysis in less-domesticated plant species such as cowpea (Muchero *et al.*, 2009), grapevine (*Vitis*; Myles *et al.*, 2010) and cottonwood (*Populus trichocarpa*; Wegrzyn *et al.*, 2010).

2.2.6   Representational Difference Analysis (RDA)

RDA was first published by Lisitsyn *et al.* (1993) to study the differences between two complex genomes. It is a technique in which subtractive hybridization and selective amplification are used to isolate the unique DNA fragments present in one DNA sample but absent from another (Figure 2.6).
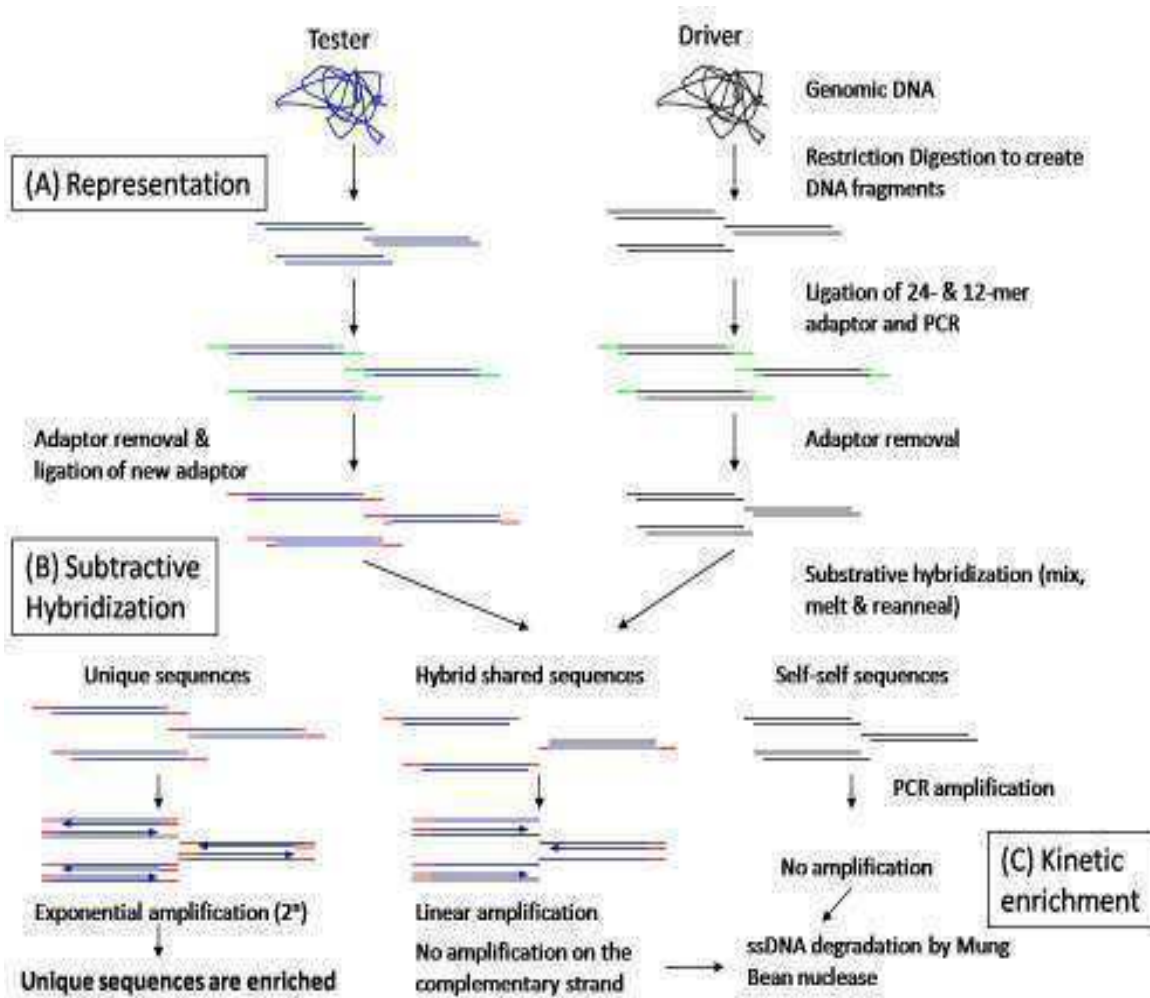
Figure 2.6: Schematic flow of Representational Difference Analysis (RDA). *RDA can be divided into three phases, (A) representation, (B) subtractive hybridization, and (C) kinetic enrichment. (A) DNA of both tester and driver are digested with restriction endonuclease, ligated to an adaptor pair and amplified with the long adaptor (24-mer) to produce a subpopulation of the original genomic DNA. This representational phase is important to reduce the complexity of the starting material. (B) After removal of the original adaptor pair, tester alone is ligated and amplified with another adaptor pair. This newly ligated-tester is hybridized with an excess amount of driver. During the subtractive hybridization, common sequences between tester and driver form hybrids due to the high abundance of driver and only the unique sequences (target sequences) in the tester re-anneal to themselves. (C) During the kinetic enrichment step, only re-annealed tester sequences have adaptors on both strands and hence will be amplified exponentially. Hybrid sequences with only one adaptor and one tester strand undergo only linear enrichment. These single-stranded DNAs are degraded by mung bean nuclease. Eventually, the unique sequences in the tester (difference products) are produced (Source: Lisitsyn, 1995).*

RDA belongs to a common class of DNA subtractive hybridization techniques in which one DNA sample (tester) is mixed with an excess of another DNA sample (driver), so that the common sequences in tester form hybrids predominantly with the driver, thereby enriching "target" sequences that are unique to the tester (Lisitsyn *et al.*, 1993; Lisitsyn and Wigler, 1995). Traditional subtractive hybridization was found to have two major problems when applied to complex genomes. The first problem was incomplete reassociation due to genome complexity. Target sequences usually are unique sequences with their reassociation is normally slow and not going to completion (Lewin, 1994). This impedes isolation of the unique sequences. The second problem was insufficient enrichment of target sequences. Although subtractive hybridization can be repeated for a few rounds, the total enrichment of target sequences is usually only about 100-fold (Wieland *et al.*, 1990) and only relatively long or abundant sequences (representing 0.1-1% of the genome) can be purified using this method. Purification of smaller target sequences is not favoured, particularly from complex genomes.

The above limitations can be circumvented by an RDA approach that has two additional components: representation and kinetic enrichment. Representation refers to any means of reproducibly generating a subset of DNA fragments, reducing the sequence complexity (Lisitsyn, 1995). According to Lisitsyn and Wigler (1995), at least a 10-fold reduction in mammalian genome complexity is required for success in subtractive hybridization. This can be achieved by digestion of the genomic DNA with a restriction endonuclease, ligation to a defined set of oligonucleotide adaptors and subsequently amplification of the DNA fragments with PCR. Smaller fragments, below 1 kb, would be more effectively amplified than large fragments, generating a subpopulation of small

restriction fragments called amplicons. Several different restriction endonucleases can be used to sample the whole genome.

After the construction of tester and driver amplicons, adaptors are removed and a new set of adaptors is linked to the tester amplicons only. PCR amplification is applied for selective enrichment of the double-stranded tester. Excess driver DNA acts as a competitive inhibitor for self-reannealing of those common sequences found in both tester and driver. Therefore only target sequences that are present in tester alone can self-reanneal and subsequently be enriched selectively at an exponential rate in the PCR reaction. These hybridization and kinetic enrichment steps can be repeated to achieve sufficient target enrichment. Lisitsyn (1995) revealed that a combination of subtraction and kinetic enrichment leads to the high degree of target-sequence purification, more than 10^7-fold after three rounds.

RDA does not require any prior knowledge of the location of the gene of interest nor the availability of a pre-existing genetic map. It also offers the advantage of providing exact sequence information about the final differences product. Additionally, Oh *et al.* (2007) suggested that a complexity of about 5 x 10^8 base pairs of DNA can be screened in each subtraction of RDA, which is greater than can be accomplished by other techniques.

RDA was first applied to detect genetic lesions in tumor (Lisitsyn *et al.*, 1993; Lisitsyn *et al.*, 1995) and since then has been widely exploited for oncology (Kaneda *et al.*, 2003; Hollestelle and Schutte, 2005; Chung *et al.*, 2008) as well as other medical studies (Kornblum and Geschwind, 2001; Shiao *et al.*, 2005; Kisielow and Cebrat, 2007;

Chang *et al.*, 2008; Molenaar *et al.*, 2009.) Given its robustness, the RDA technique has been successfully applied to a variety of organisms. These include viruses (Lammens *et al.*, 2009), bacteria [*Pseudomonas aeruginosa* (Choi *et al.*, 2002); *Actinobacillus pleuropneumoniae* (Xie *et al.*, 2009)], plants [soybean (Ling *et al.*, 2003); rice (Park *et al.*, 2007; Sperotto *et al.*, 2008); tomato (Kok *et al.*, 2007); Pea (*Pisum sativum*; Li *et al.*, 1998); liver wort (*Marchantia polymorpha*; Fujisama *et al.*, 2001)] and animal studies [marsupials (Brown *et al.*, 2008); dog (*Canis lupus*; Everts *et al.*, 2000); rat (*Rattus*; Toyota *et al.*, 1996)]. It can be used to identify genomic deletions, rearrangements, insertion, amplification or point mutation between any two complex genomes.

In plants, the Cullis group from Case Western Reserve University, Cleveland, United States has employed the RDA technique extensively to study somaclonal variation in tissue culture (Cullis and Kunert, 2000) as well as diversity study of date palm (*Phoenix dactylifera*; Voster *et al.*, 2002) and flax (*Linum usitatissimum*; Oh and Cullis, 2003). A DNA microchip technology that was developed from RDA and useful in plant tissue culture industry was introduced by Kunert *et al.* (2002). In 2007, Oh *et al.* successfully identified a DNA fragment located in the labile region of banana genome that is highly susceptible to stress imposed during tissue culture and associated with higher rearrangement and mutation rates. The authors suggested that this DNA fragment has the potential to be developed into a detection kit for somaclonal variation. The examples mentioned above explain the robustness and usefulness of this technique in the study of differences between two samples.

2.2.7 **Diversity Array Technology (DArT)**

Diversity Array Technology (DArT) is a microarray hybridization-based marker system that allows simultaneous typing of several hundreds to thousands loci in a single assay without relying on the sequence information. It generates whole genome profiling by scoring presence versus absence of DNA fragments in representations of genomic DNA samples (Jaccoud *et al.*, 2001; Wenzl *et al.*, 2004). This technique can overcome some of the limitations of other molecular marker techniques such as capacity, speed and cost (Akbari *et al.*, 2006).

DArT assays DNA polymorphism through generation of genomic representations. Genomic representations can be produced by restriction digestion of genomic DNA using any combinations of restriction endonuclease and hence reproducibly reducing the complexity of genomic DNA of samples. Microarray is built once for each species and contains representation fragments produced from a set of genotypes that cover the gene pool of the species (Hutter *et al.*, 2007). The number of markers available for a particular species is therefore governed by the level of genetic variation within the species (or gene pool) and the number of complexity reduction methods screened (Mondini *et al.*, 2009). This approach was described in detail by Jaccoud *et al.* (2001) and is reproduced here (Figure 2.7).
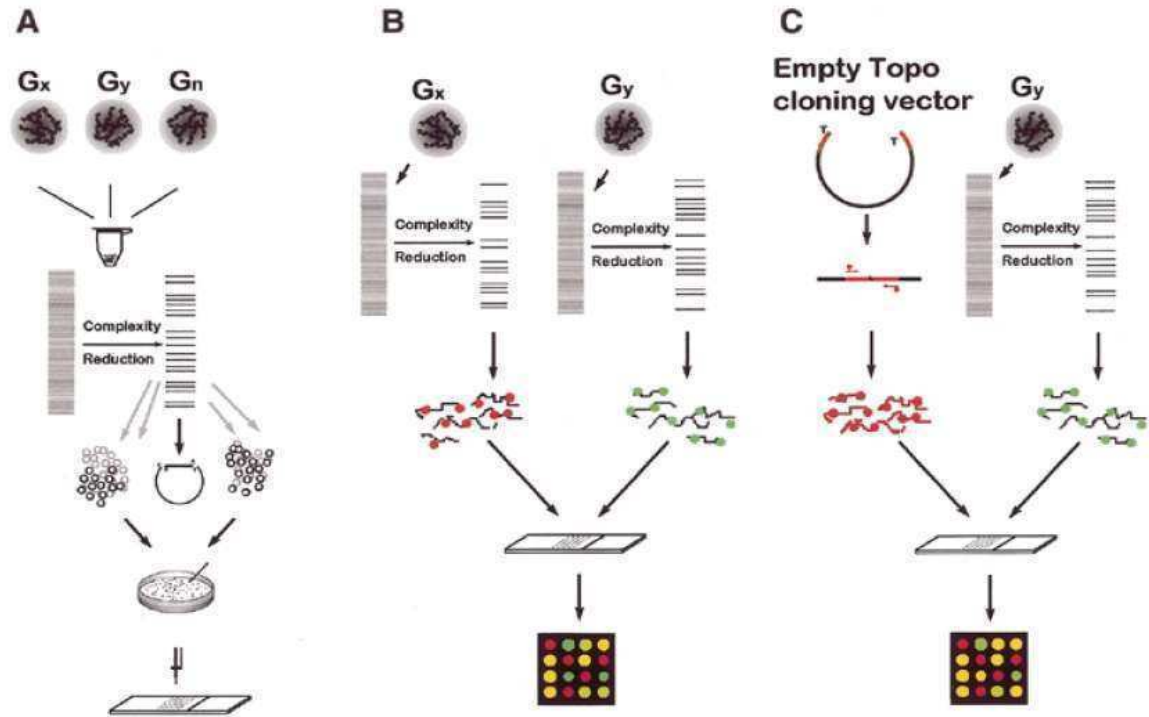
Figure 2.7: Schematic representation of DArT. *(A) Generation of diversity panels. Genomics DNAs of specimens to be studied are pooled together. The DNA is cut with chosen restriction enzyme(s) and ligated to adaptors. The genome complexity is reduced in this case by PCR using primers with selective overhangs. The fragments from representations are cloned. Cloned inserts are amplified using vector-specific primers, purified and arrayed onto a solid support. (B) Contrasting two samples using DArT. Two genomic samples are converted to representations using the same method as in (A). Each representation is labelled with a green or red fluorescent dye, mixed and hybridized to the diversity panel. The ratio of green:red intensity is measured at each array feature. Significant differences in the signal ratio indicate array elements (and the relevant fragment of the genome) for which the two samples differ. (C) Genetic fingerprinting using DArT. The DNA samples for analysis is converted to a representation using the methods as in (A) and labelled with green fluorescent dye. Fragments of the cloning vector, which are common to all elements of the array (polylinker of PCR2.1-TOPO vector, marked red), are labelled with red fluorescent dye and hybridized to a Diversity Panel together with green fluorescent-labelled representation. First the ratio of signal intensity is measured at each array feature for each input genotype used to generate Diversity Panels. Polymorphic spots are identified by binary distribution of signal ratios among input samples. Any new specimen can be assayed on arrays of polymorphic features to generate a genetic fingerprint (Source: Jaccoud et al., 2001).*

The major advantages of the DArT technique is that it does not require prior sequence information for the species to be studied; this enables study of minor or 'orphan' crops that have limited DNA sequence information. This technique also offers highly parallel, fast, reproducible and comprehensive genome coverage analysis which is cost effective (Semagn *et al.*, 2006a). It is estimated that the cost of DArT markers are tenfold lower than SSR markers per data point (Xia *et al.*, 2005). Nonetheless, this technique has its own limitation. It is a dominant marker system which might limit its application (Semagn *et al.*, 2006a).

The DArT approach has been performed in plant species of virtually any ploidy level. It was first developed for diploid rice with small genome size of 430 Mbp (Jaccoud *et al.*, 2001). The technology was then expanded to barley with 5000 Mbp genome (Wenzl *et al.*, 2004), hexaploid wheat (Akbari *et al.*, 2006) and sugarcane, one of the most genetically complex plant genome (x=5-14) (Heller-Uszynska *et al.*, 2006), to name a few.

To date, DArT arrays have been established for more than 120 plants species (www.diversityarrays.com) with more than 3000 and 7000 DArT markers were developed for barley and wheat, respectively (Varshney *et al.*, 2010). DArT has been employed extensively for genetic mapping studies, including Barley (Wenzl *et al.*, 2004), Arabidopsis (Wittenberg *et al.*, 2005), Wheat (Akbari *et al.*, 2006), Sorghum (Mace *et al.*, 2008). Integration of DArT markers with other marker systems has improved genetic mapping of certain species, allowing construction of high density genetic maps. Wenzl *et al.* (2006) reported a high density consensus map of barley comprising 2085 DArTs and

850s other markers (SSRs, RFLPs and STSs) with an average inter-bin distance of 0.71 ± 1.01 cM when co-segregating loci were grouped into bins. The first dense genetic map of banana (*Musa acuminata*) was published in 2010 (Hippolyte *et al.*, 2010). This reference map has the expected eleven linkage groups containing 167 SSR and 322 DArT loci with an average density of 2.8 cM per marker.

In parallel with genetic mapping, the DArT technique has been utilized for identifying trait-marker associations and QTL analysis, for example water logging tolerance, net blotch resistance and drought tolerance in Barley (Li *et al.*, 2008; Grewal *et al.*, 2008; Varshney *et al.*, 2012) and Ergot resistance in sorghum (Parh *et al.*, 2008). The large number of markers being assayed concurrently by the DArT technique has also contributed to high resolution assessment in genetic diversity studies in cassava (Xia *et al.*, 2005), rice (Xie *et al.*, 2006), pigeonpea (Yang *et al.*, 2006), oat (Nicholas *et al.*, 2009), banana (Risterucci *et al.*, 2009), rye (Bolibok-Bragoszewska *et al.*, 2009), *Eucalyptus* (Steane *et al.*, 2011) and rapeseed (Raman *et al.*, 2012) as well as association mapping (Crossa *et al.,* 2007; Neumann *et al.*, 2010).

2.2.7.1   DArT "Genotyping-by-sequencing" (DArTSeq)

Genome complexity reduction for genotyping, a crucial step in DArT technology, has now been taken to another level when combined with next-generation sequencing (NGS) technologies, a method generally termed as Genotyping-by-Sequencing (GbS) (Sansaloni *et al.*, 2011). The use of genome complexity reduction combined with multiplex sequencing was first demonstrated through restriction-site associated DNA (RAD) tagging (Baird *et al.*, 2008; Miller *et al.*, 2007). GbS was developed as a simple

but robust approach for complexity reduction in large complex genomes for high density SNP discovery and genotyping (Elshire *et al.*, 2011; Poland *et al.*, 2012). The GbS approach is suitable for population studies, germplasm characterisation, breeding and trait mapping in diverse organism.

DArT "Genotyping-by-sequencing" (DArTSeq) is a new marker platform developed by DArT Ptd Ltd, in which DArT platform is coupled with Illumina short read sequencing to generate DArT (presence/absence) and SNP markers. This technology has been successfully applied for genetic mapping of *Eucalyptus* (Sansaloni *et al.*, 2011) and genetic diversity assessment study of *Lesquerella* and related species (Cruz *et al.*, 2013).

2.2.8   Next-Generation Sequencing (NGS)

Since its first publication in the late 1970s by Nobel laureates Frederick Sanger and Walter Gilbert (Sanger and Coulson, 1975; Maxam and Gilbert, 1977) and subsequent development of chain termination method by Sanger and colleagues (Sanger *et al.*, 1977), Sanger or dideoxy sequencing, has been the most commonly used DNA sequencing technique to date and was used to complete human genome sequencing project.

Despite its wide range of application, Sanger sequencing method has several limitations such as (1) the need for gels or polymers to separate the fluorescently-labelled DNA fragments by size, (2) the relatively low number of samples that could be analysed in parallel and (3) the difficulty of total automation of the sample preparation methods in which clonal populations of DNA are currently produced using *Escherichia coli*, which is

labour-, robotics- and space-intensive for large-scale operations (Ansorge, 2009; Varshney *et al*., 2009). Advancement in sequencing technologies have delivered the next-generation sequencing (NGS) approaches which are capable of processing millions of sequence reads in parallel in a single run. Availability of NGS techniques has rapidly changed the landscape of life sciences. Currently, three main systems are available in the market: Roche/ 454 FLX (www.454.com), Illumina/ Solexa Genome Analyzer (www.illumina.com), and the Applied Biosystems SOLiD^{TM} System (www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html). Recently, another two new systems have been launched in the market: the Helicos Heliscope^{TM} (www.helicosbio.com) and Pacific Biosciences SMRT (www.pacificbiosciences.com).

Although these NGS platforms are quite diverse in their configurations and sequencing biochemistry (Table 2.4), they share many common features. The sequencing reaction is performed on *in vitro* clonally amplified single strands of a fragment library, avoiding the need for the bacterial cloning step as well as the associated cloning bias issues. Helicos and Pacific Biosystems mentioned above are "single molecule" sequencers that do not require any amplification of DNA fragments prior to sequencing, so should avoid the inherent bias in PCR amplification. Relatively little input DNA (a few micrograms at most) is needed to produce a library. Most importantly, array-based NGS enables a much higher degree of parallelism than conventional capillary-based sequencing; hundreds of millions of reads can be processed in parallel rather than 96 at a time and they require only one or two instrument runs to complete an experiment. Collectively, these differences dramatically reduce the sequencing cost per base by

several orders of magnitude. The main limiting factor of the new technology is shorter

read lengths (35-500 bp, depending on the platform) compared to Sanger sequencer (650-

1000 bp), raw accuracy (base-calls generated by the new platforms are at least ten-fold

less accurate than those generated by Sanger sequencers); and also lower reading

accuracy in homopolar stretches of identical bases. The huge amount of data generated by

these systems (over a gigabase per run) in the form of short reads presents another

challenge for developments of software and more efficient computer algorithms. These

technologies will continue to improve to overcome such limitations (Mardis, 2008a;

Mardis, 2008b; Shendure and Ji, 2008; Ansorge, 2009).

Table 2.4: Comparison of the Next-Generation DNA Sequencing Platforms.

Platforms	Roche (454) GS-FLX	Illumina Genome Analyzer	ABI SOLiD
Starting DNA (μg)	3-5	0.1-1	0.1-20
Amplification	Emulsion PCR	Bridge PCR	Emulsion PCR
Sequencing method	Pyrosequencing	Sequencing by synthesis	Sequencing by ligation
Read length (bases)	500	32-40	35
Throughput capability (Gb per run)	0.1	1.3	4
Reagent cost per run (list prices)	$8,500	$3,000	$3, 400
Run time	7.5 h	3 d	7 d
Paired reads/Span	Yes/3 kb	Yes/200-400 bp	Yes/3-20 kb

(Source: Liu, 2009)

2.2.8.1 454 (Roche) Pyrosequencing

Roche 454 sequencing was the first NGS system introduced into the market

(Margulies *et al.*, 2005) and it works on the principle of "pyrosequencing" in which

incorporation of a nucleotide by DNA polymerase results in the release of pyrophosphate

to fuel a series of downstream reactions that ultimately produces light from the cleavage of luciferin by firefly enzyme luciferase (Figure 2.8).
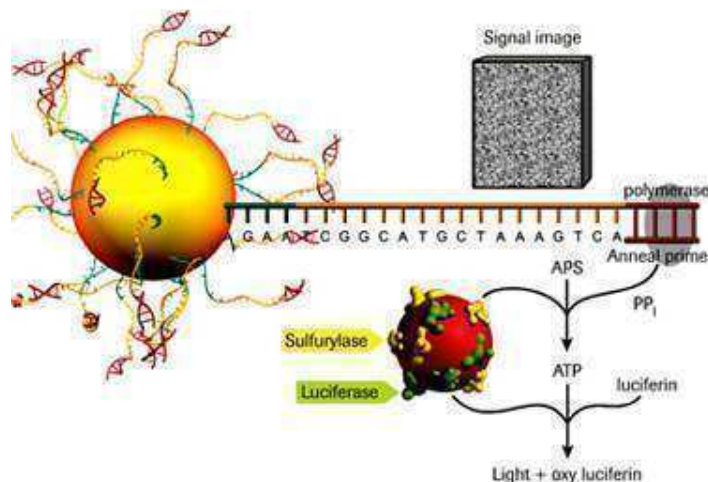
Figure 2.8: Schematic illustration of the pyrosequencing reaction which occurs on nucleotide incorporation to report sequencing-by-synthesis in 454 sequencing technology. *Incorporation of the complementary base generates inorganic pyrophosphate (PPi), which is converted to ATP by sulfurylase. Luciferase uses the ATP to convert luciferin to oxyluciferin, producing light (Source: Rothberg and Leamon, 2008).*

In 454 sequencing system (Figure 2.9), DNA fragments are ligated with specific adaptors and then mixed with a population of 28 μm beads carrying complementary oligonucleotides, resulting in binding of one fragment to each bead. Emulsion PCR is carried out for fragment amplification, with each bead isolated into individual oil-to-water micelles that also contain PCR reagents, producing around one million copies of each fragment on the surface of each bead. Amplification is necessary to obtain sufficient light signal intensity for reliable detection in next sequencing-by-synthesis reaction steps (Ansorge, 2009).
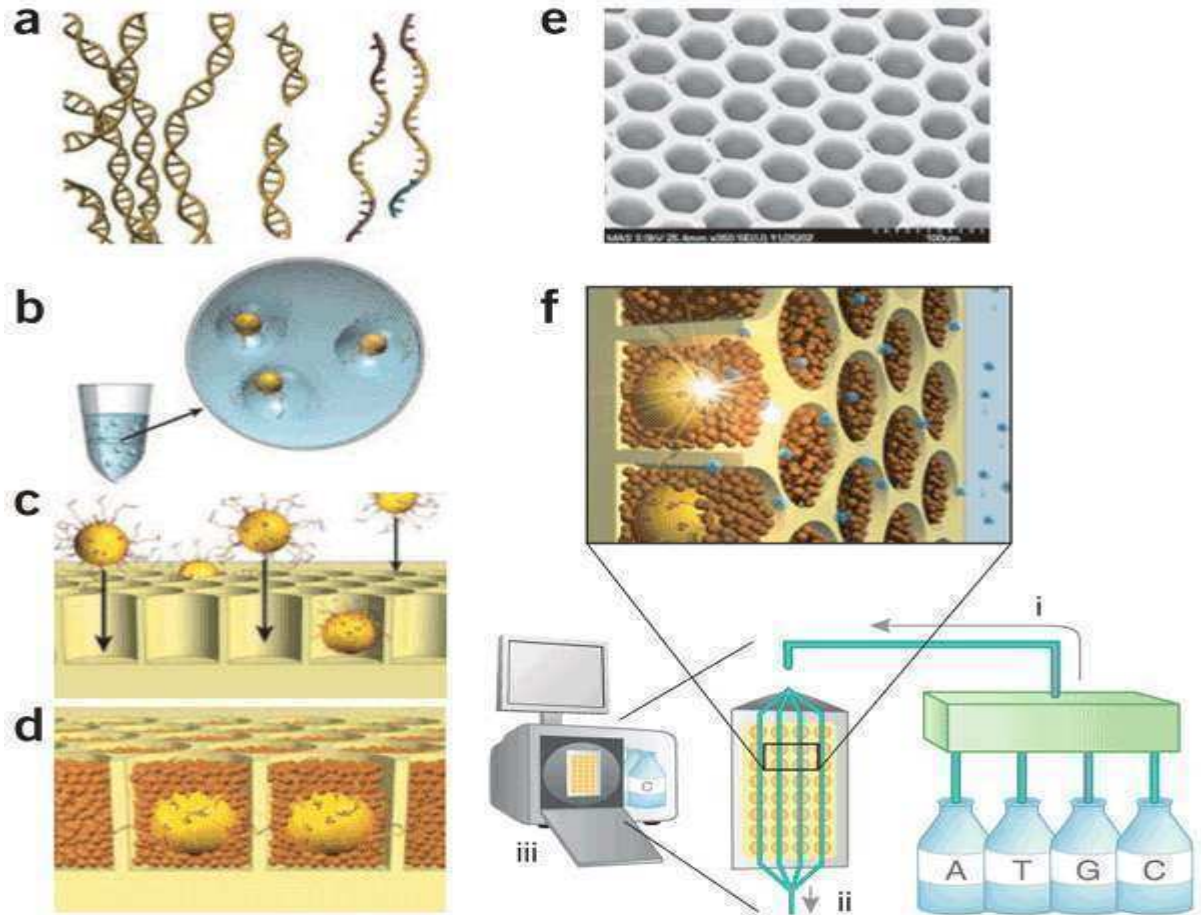
Figure 2.9: Overview of the 454 sequencing technology. *(a) Genomic DNA is isolated, fragmented, ligated to adaptors and separated into single strands. (b) Fragments are bound to beads under conditions that favour one fragment per bead; the beads are isolated and compartmentalized in the droplets of a PCR-reaction-mixture-in-oil emulsion and PCR amplification occurs within each droplet, resulting in beads each carrying ten million copies of a unique DNA template. (c) The emulsion is broken, the DNA strands are denatured, and beads carrying the single-stranded DNA templates are enriched (not shown) and deposited into wells of a fiber-optic slide. (d) Smaller beads carrying immobilized enzymes required for a pyrophosphate sequencing reaction are deposited into each well. (e) Scanning electron micrograph of a portion of a fiber-optic slide, showing fiber-optic cladding and wells before bead deposition. (f) The 454 sequencing instrument consists of the following major subsystems; a fluidic assembly (object i), a flow cell that includes the well-containing fiber-optic slide (object ii), a CCD camera-based imaging assembly with its own fiber-optic bundle used to image the fiber-optic slide (part of object iii), and a computer that provides the necessary user interface and instrument control (part of object iii) (Source: Margulies et al., 2005; Rothberg and Leamon, 2008).*

Beads with million copy of fragments attached are then deposited on a microfabricated array of picotiter plate (PTP) that hold a single bead in each of several hundred thousand single wells, providing a fixed location at which each sequencing reaction can be monitored. Smaller beads containing immobilized enzymes for downstream pyrosequencing are also added surrounding the fragments:bead. During sequencing, one side of the PTP acts as a flow cell which nucleotides and reagent solution are delivered in a sequential fashion, whereas the other side of PTP is bounded to a fiber-optic bundle for CCD (charge-coupled device) –based signal detection (Shendure and Ji, 2008). Knowing the identity of the nucleotide supplied in each step, the presence of a light signal indicates the base incorporated into the sequence of the growing DNA strand. This sequencing is "asynchronous" in that some features may get ahead or behind other features depending on their sequence relative to the order of base addition.

The major limitation of 454 technology is that it cannot accurately interpret long stretches of the same nucleotide (a homopolymer run). Therefore the dominant error type for 454 is insertion-deletion rather than substitution (Shendure and Ji, 2008). The key advantage of 454 is it can achieve reading lengths of 400-500 base range, paired-end reads, and hence it is suitable for *de novo* assembly and metagenomics.

The GS-FLX Titanium instrument produces an average read length of 450 bp per bead, with a throughput of ~450 Mb of sequence data during a 10-h run. By contrast, a single ABI 3730 can sequence 24 x 96-well plates per day producing ~ 440 kb of sequence data in 7 h, with an average read length of 650 bp per sample (Mardis, 2008). The newly upgraded 454 FLX Titanium XL+ increase data output from 450 Mb to about

700 Mb with read length up to 1 kb. The 454 technology has been the most widely published next-generation technology, having almost 3,000 research publications (www.454.com).

2.3 Genetic linkage mapping and QTL analysis

2.3.1 Genetic linkage mapping

Genetic mapping, also known as linkage mapping, is a process of determining the relative position and distances between markers along chromosomes. Genetic linkage was first discovered in 1905 in the sweet pea (*Lathyrus odoratus*) by Bateson and colleagues, although that time linkage between loci was referred as 'coupling'. Following Morgan's observation that the amount of crossing over between genes might indicate the distance between them on a chromosome, Morgan's student, Sturtevant used these ideas to develop the first genetic map of chromosome X of *Drosophila melanogaster* in 1913. At that time, genetic maps were generated by just a few to several tens of phenotypic markers obtained one by one by observing morphological and biochemical variations of an organism, mainly following mutation (Wu *et al.*, 2008). Development of wide range of molecular markers that reveals differences at DNA level over the past two decades has resulted in extensive genetic mapping in many species and also generation of much more densely populated genetic maps, generally into the range of several hundreds to more than a thousand markers per genome (Semagn *et al.*, 2006b; Wu *et al.*, 2008).

Genetic maps are vital for identification of chromosomal locations containing genes and QTLs associated with traits of interests which in turn facilitates marker-

assisted selection. Genetic mapping enables comparative mapping between related species, provides a framework for anchoring physical maps and facilitates positional or map-based cloning of gene of interest (Semagn *et al.*, 2006b).

During meiosis, recombinations occur when homologous chromosome pairs form chiasma and exchange sections of chromosome which leads to production of recombinant gametes. In a segregating population, there is a mixture of parental and recombinant genotypes. The frequency of recombinant genotypes gives an estimate of the distance between two markers on a chromosome; on the assumption probability of crossing over is proportional to the distance between two markers. The closer the two markers are located on a chromosome, the lower the frequency of recombination between the markers while markers situated far apart on a chromosome or on different chromosomes assort independently. Linked markers have a recombination frequency that is less than 50% while unlinked markers have a recombination frequency of 50% (Collard *et al.*, 2005; Semagn *et al.*, 2006b).

There are three important steps in constructing genetic linkage maps: (1) production of the mapping population; (2) identification of marker polymorphism; and (3) linkage analysis of markers using computer software. Generation of a genetic map is a conceptually simple yet computationally complex process (Collard *et al.*, 2005).

(i)   Production of the mapping population

One of the most critical steps in constructing a genetic map is to develop an appropriate mapping population. Mapping populations should be segregating populations

derived from two genetically divergent parents, differing for one or more traits of interest. The parent lines should be genetically divergent enough to enable identification of large number of polymorphic markers that are well-distributed across the genome but at the same time they should not be genetically too distant to avoid sterility and/or high levels of segregation distortion during linkage analysis (Semagn *et al.*, 2006).

The choice of a mapping population could vary based upon the objectives of the experiment, the time frame and resources available. Various types of mapping population can be produced from the heterozygous F_1 hybrids and each of these mapping populations has its own advantages and disadvantages.

(i) F_2 population: Self-pollination of F_1 hybrids

(ii)   Backcross (BC) population: Crossing of F_1 plants back into one of the parents

(iii)   Recombinant Inbred lines (RIL): Single-seed selection from individual plants of an F_2 population continue for 6-8 generations

(iv)   Near isogenic lines (NIL): Backcrossing for at least six to seven generations followed by self-pollination of selected individuals produce lines that are homogenous for the target gene and nearly isogenic with the recipient parents

(v)   Double haploids (DH): Doubling of gametes from F_1 or F_2 plants

Both F_2 and BC population are the simplest form of a mapping population as they are easy to construct and require only short time. However, F_2 and BC population are highly heterozygous and cannot be easily preserved; they are considered to be temporary populations. On the other hand, RIL, NIL and DH populations constitute a permanent source that can be replicated indefinitely without genetic-change occurring and shared by

many groups in the research community (Schneider, 2005; Semagn *et al.*, 2006b). The length of time required to produce RIL and NIL is the major constraint in mapping studies. They usually take six to eight generations to achieve homozygosity, a very time-consuming process. DH populations are faster to generate than RIL and NIL but their production is only possible in species that are amenable to tissue culture (Collard *et al.*, 2005).

The type of populations to be used in mapping studies also depends on the reproductive mode of the plant to be analysed. Self-pollinating species allow the generation of lines displaying a maximum degree of homozygosity, hence all population types can be used as mapping populations. However, it is difficult to produce pure lines for self-incompatible plants due to inbreeding depression. Mapping populations such as F_1 and BC are more suitable for map construction (Schneider, 2005).

Simulation studies performed by Ferreira *et al.* (2006) using a sample size of 50-1,000 individuals of F_2, BC, RIL and DH populations have shown that a total of 200 individuals were required to construct reasonably accurate linkage maps for all population types. In practice, population size ranging from 50-250 individuals is generally used in preliminary genetic mapping studies (Mohan *et al.*, 1997). Larger population size will be useful for high resolution mapping (Collard *et al.*, 2005).

(ii)   Identification of polymorphism

The second step in the construction of a linkage map is identification of polymorphic markers, markers that can reveal differences between parents. Construction

of a genetic map requires sufficient polymorphism between parents of a cross (Young, 1994). Overall, cross-pollinating species show higher genetic heterozygosity as compared to inbreeding species, therefore distantly related parents should be selected for mapping of inbreeding species. The entire mapping population including the parents must then be genotyped with the selected polymorphic markers (Collard *et al.*, 2005; Semagn *et al.*, 2006b).

The choice of DNA markers used for mapping depends on the availability of characterised markers, resources available as well as the type of mapping population. Dominant marker systems, such as AFLP and RAPD, are unable to show differences between homozygous and heterozygous individuals, hence these markers are not ideal for mapping F_2 or BC populations. On the other hand, RIL and DH populations can maximize the information obtained from dominant markers. F_2 populations are best to be exploited using co-dominant marker systems, such as SSRs and SNPs (Ferreira *et al.*, 2006; Semagn *et al.*, 2006b).

(iii)   Linkage analysis of markers

Linkage analysis of markers is the final step of the construction of a genetic map in which each DNA marker of each individual of a population is converted to coding data and linkage of the markers is analysed using computer programmes (Collard *et al.*, 2005). Commonly used software programmes that are freely available from the internet include Mapmaker/EXP (Lander *et al.*, 1987; Lincoln *et al.*, 1993a), MapManager QTX (Manly *et al.*, 2001) and CarthaGene (Schiex and Gaspin, 1997; de Givry *et al.*, 2005) while

JoinMap (Stam, 1993; van Ooijen, 2006) is a commercial programme that is also widely used.

Linkage analysis of markers can essentially be split into 3 parts: locus grouping, locus ordering and distance estimation. The first part, locus grouping, divides DNA marker into candidate linkage groups using the odds ratios, which refers to the ratio of the probability that two loci are linked with a given recombination value over a probability that two are not linked. This ratio is called a logarithm of odds (LOD) value or LOD score (Risch, 1992). LOD values of >3 are typically used to construct linkage groups (Collard *et al.*, 2005). A LOD value of 3 between two markers indicates that linkage is 1000 times more likely than no linkage (1000:1). Higher LOD value will result in fragmented linkage groups while lower LOD value tend to create few linkage groups with large number of markers per group which might lead to unstable locus orders and fusion between different linkage groups. Ideally, linkage groups obtained should be the same as the haploid chromosome numbers of the species under study (Nelson, 2005; Semaign *et al.*, 2006b).

The second part, ordering, takes each of the linkage groups in turn and aims to find the relative orders of the markers within the group. For a linkage group of *m* markers, there are *m!/2* possible orders and there is no sure way to find the best possible order even for groups with modest size. Therefore ordering is the central problem in linkage mapping and most effort in genetic mapping algorithm development has been spent researching the marker ordering problem (Semagn *et al.*, 2006b; Cheema and Dicks, 2009). Various locus-ordering criteria have been adopted by different statistical

programmes including minimum sum of adjacent recombination fractions (Falk, 1989), minimum weighted least squares marker order (Stam, 1993) and maximum likelihood (Lander *et al.*, 1987; Jansen *et al.*, 2001).

Once the order of the markers has been obtained, the final step is to find the length of the linkage group which is the sum of all inter-marker distances. Map distance is measured in terms of the frequency of recombination between marker loci (Paterson, 1996). Recombination fractions are converted to map units, centiMorgans (cM), by mapping functions. One cM is equal to one percent recombination, but for longer distances, recombination fraction (Rf) is not linearly related to centiMorgan distances. Two commonly used mapping functions are Haldane and Kosambi mapping function. The Kosambi mapping function assumes that recombination events influence the occurrence of adjacent recombination events (Kosambi, 1944) while Haldane mapping function assumes absence of interference between crossover events (Haldane, 1919).

2.3.2   Quantitative trait loci (QTL) analysis

Quantitative characters have been a major area of genetic study for more than a century because they are a common feature of natural variation in populations, particularly for commercially important traits in plants (Kearsay and Farquhar, 1998). Quantitative trait shows a continuous range of variation in a population, which is more or less normally distributed. There is no obvious discontinuities in the distribution as might be expected from a single gene trait (Kearsay 1998). The genetic variation underlying quantitative characters results from the segregation of numerous quantitative trait loci (QTL), each explaining a portion of the total variation, and whose expression is modified

by interactions with other genes and also the environment (Mackay, 2001). Therefore mapping quantitative trait is difficult since the genotype cannot be unequivocally determined from phenotype.

Quantitative trait loci (QTL), first termed by Gelderman (1975), is a region of the genome that is associated with an effect on a quantitative trait. QTL analysis is looking for associations between the quantitative trait and the marker alleles segregating in the population (Kearsey and Farquhar, 1998). The main objectives of QTL analysis are to identify the regions of genome that affect the trait of interest and to explore the effects and interactions of these regions (Kearsey, 1998). QTL analysis involves two essential steps, mapping of the markers and the association of the trait with the markers. Usefulness of genetic linkage map for localization of QTL for a quantitative trait was first demonstrated by Paterson *et al.* (1988). Establishment of large collections of molecular markers has enabled construction of detailed genetic maps which laid the foundation for QTL analysis (Doerge, 2002). Various statistical techniques have been employed to analyse the association between the markers and quantitative trait, ranging from simple single-marker analysis to models that include multiple markers and interactions.

Single-marker analysis is a simple approach in which *t*-test, analysis of variance (ANOVA) or linear regression is used to test if the differences between the marker means are significant for the trait and hence point to the existence of potential QTL (Kearsey, 1998). Linear regression is most commonly used because the coefficient of determination (R^2) from the marker explains the phenotypic variations arising from the QTL linked to the marker (Collard *et al.*, 2005). Single-marker analysis investigates individual markers

independently without reference to their position and order, therefore this method does not require linkage map and can be performed using basic statistical software programme. Although computationally simple, this approach suffers several major limitations: (i) difficult to conduct separate estimates of QTL location and effect; (ii) the likelihood of QTL detection decreases significantly when the distance between the marker and QTL increases; (iii) the effect of QTL are likely to be underestimated as they are confounded with recombination frequency (Tanksley, 1993; Doerge, 2002; Collard *et al.*, 2005). Commonly, single-marker analysis is performed using computer programme QGene (Nelson, 1997) and MapManagerQTX (Manly *et al.*, 2001).

Simple interval mapping (SIM), was first proposed by Lander and Botstein (1989), makes use of linkage map and explores the interval between pairs of markers for the presence of QTL (Kearsey, 1998). Intervals between adjacent pairs of markers along a chromosome are scanned in a systemic, linear (also referred to as one-dimensional) fashion and the likelihood profile of a QTL being at any particular point in each interval is determined (Kearsey and Farquhar, 1998). The resulting LOD scores are plotted along a chromosome map and the peak of LOD exceeds some significance threshold indicates the likely location of the QTL and provides information on its confidence interval (Churchill and Doerge, 1994; Mangin *et al.*, 1994). Interval mapping is statistically more powerful than single-marker approach to detect QTL, but it is still a single-QTL model and the one-dimensional search of QTL does not consider the interactions between multiple QTLs (Doerge, 2002). SIM is commonly conducted using software MapMaker/QTL (Lincoln *et al.*, 1993b), Windows QTL Cartographer (Wang *et al.*, 2012), MapQTL (van Ooijen, 2009) and QGene.

Composite Interval Mapping (CIM) introduced by Zeng (Zeng, 1993) and Multiple QTL mapping (MQM) introduced by Jansen (Jansen, 1993) in the same year, were developed to overcome some of the shortcomings of SIM. Both methods extend the ideas of interval mapping to include additional markers as cofactors outside a defined window of analysis. The inclusion of co-factors is used to eliminate the background genetic noise (QTLs elsewhere on the genome) and neutralize the effects of linked QTLs (from outside the window of analysis) resulting in an increase in the power and reduction of interference due to linked QTLs (Zeng, 1993; Zeng, 1994; Jansen 1993; Jansen and Stam, 1993). However, these two approaches are still a one-dimensional search and hence are unable to accommodate a multiplicity of potential epistatic QTL effects (Doerge, 2002). The CIM and MQM method have been implemented in the QTL Cartographer and MapQTL, respectively.

Multiple intervals mapping (MIM), as the names implies, uses multiple interval simultaneously to fit multiple QTLs into the model (Kao et al., 1999). MIM implemented in QTL Cartographer, uses a stepwise selection method to add and remove QTLs from a model first arrived at by CIM, then estimate simultaneously the QTL genotypes and their likelihoods, and finally searches for epistatic effects between modelled QTLs and each other or the unoccupied QTL positions on the map. MIM is well situated to the identification and estimation of genetic architecture parameters, including the number, genomic positions, effects and interactions of significant QTL and their contribution to the genetic variance (Nelson, 2005).

Locating multiple interacting QTL that are associated with multiple traits is the goal of many current scientific investigations. Continuous advancement in molecular marker technology coupled with evolving sophisticated statistical analyses and modelling, are expected to enable greater power and precision in the detection of QTL and contributes to the application of QTL in crop improvement, such as marker-assisted selection (MAS).

2.4   Biotechnology and molecular research in oil palm

Oil palm is a perennial tree crop with long breeding and selection cycles, so molecular breeding is of great interest as this could save time, cost and effort as well as utilise the limited resources (land and labour) more effectively. Molecular research activities on oil palm started in the early 1990s. Since then, different molecular techniques have been used to determine and isolate markers in oil palm, namely isozyme (Ghesquiére, 1984; Ghesquiére, 1985; Baudouin, 1992; Rajanaidu *et al.*, 1993; Choong *et al.*, 1996), RFLP (Cheah, 1990; Jack *et al.*, 1995; Mayes *et al.*, 1996), RAPD (Shah *et al.*, 1994), AFLP (Cheah, 2000; Kulratne *et al.*, 2000) and SSR (Billotte *et al.*, 1999; Billotte *et al.*, 2001; Billotte *et al.*, 2005).

In oil palm, two monogenic inherited traits of importance are, fruit colour gene (*Vir*) and shell-thickness gene (*Sh*). Two RFLP markers linked to the fruit colour gene were identified, MET16 (3 cM) and KT3 (4 cM), in the linkage map constructed by MPOB (Sambanthamurthi *et al.*, 2009). The markers were found to not only be able to distinguish the *nigrescens* and *virescens* fruit but also able to distinguish the homozygous and heterozygous forms of *Vir* fruit. The *Vir* fruits are green in colour when unripe and

change to bright orange when ripe due to absence of carotenoids in the exocarp. This profound change in colour allows the easy identification of ripened bunches and hence reduces crop loss through fallen fruits (Jack *et al.*, 1998).

Many molecular approaches have been employed to study shell-thickness gene (*Sh*) and/or marker(s). Mayes *et al.* (1997) developed a RFLP genetic linkage map for oil palm using a population derived from a self-pollinated *tenera* palm that segregated for the shell-thickness character, enabling the discovery and mapping of a RFLP marker (pOPgSP1282) linked to *Sh* at a distance of 9.8 cM. This marker is rather far away from *Sh* to be used for identification of fruit types as the chances of recombination between marker allele and gene are still high.

RAPD work by Moretzsohn *et al.* (2000) on linkage mapping of the shell-thickness locus also revealed two different markers, R11-1282 and T19-1046, that were 17.5 cM and 23.9 cM, respectively on either side of the sh^+ locus. Although the two markers were even further away from *Sh* gene, the authors claimed that the use of flanking marker-based assay would allow *tenera* and *pisifera* palms resulting from D x P cross to be identified correctly with an error rate of only 4% (0.175 x 0.239 = 0.042). Hence, more precise and fast identification of fruit form is possible with these two markers. However, these markers have not been verified, validated and would need to be converted to a different format for use in selection programmes.

A high density microsatellite-based linkage map of oil palm was published by Billotte *et al.* in 2005. It is the first linkage map that has 16 independent linkage groups corresponding to the 16 homologous chromosome pairs of oil palm. This integrated map

covered 1,743 cM of the plant's genome by using 944 SSR and AFLP markers. An AFLP marker, E-AGG/M-CAA132 was discovered to map at 4.7 cM from the *Sh* locus and to be located at terminal region of LG4. This AFLP marker is the closest marker to the shell-thickness gene that had been published. However, there are no reports of its validation in commercial material or its use in selection programmes. Sambanthamurthi *et al.* (2009) commented that the marker is still too far to allow for an error free selection of the trait in the nursery. Ideally, a marker with a distance of 1 cM or two close flanking markers are preferred.

On the other hand, most agronomically important traits, such as oil yield and quality, are controlled by many genes. Quantitative trait locus (QTL) analysis is used to study polygenic traits (Collard *et al.*, 2005). Using the same population as Mayes *et al.* (1994)*,* Rance *et al.* (2001) reported the first quantitative trait loci mapping for yield components in oil palm. This study identified several putative markers associated with fruit weight, petiole cross section, rachis length, and ratios of shell:fruit, mesocarp:fruit and kernel:fruit. MPOB also conducted QTL studies associated with oil quality in oil palm in which 11 QTLs were detected for Iodine Value, C14:0, C16:0, C16:1, C18:0, C18:1 and C18:2 in four different linkage groups using a framework map consisting of AFLP, RFLP and SSR markers (Singh *et al.*, 2009). A recent report by Montoya *et al.* (2013) revealed the detection of 19 QTLs associated with palm oil fatty acid composition using an interspecific pseudo-backcross of *E. guineensis* and *E. oleifera.*

Billotte *et al.* (2010) published the first QTL analysis on multi-parent population in oil palm. In this study, within-family and across-family analysis were performed for

QTL searches and a total of 76 QTLs were identified from 24 quantitative traits which proves that across-family analysis is efficient with interconnected families and can partially solve the small family size issue of classical genetic trials of oil palm.

Due to the narrow genetic basis of oil palm planting material, extensive collection of wild and semi-wild material has been made by MPOB to improve current commercial germplasm (Lawrence *et al.*, 1995; Rajanaidu *et al.*, 2000; Mohd Din *et al.*, 2005) and germplasm diversity has been assessed by different molecular markers (Hayati *et al.*, 2004; Maizura *et al.*, 2006; Singh *et al.*, 2008a; Ting *et al.*, 2010). Genetic diversity studies can estimate the genetic distance of different breeding materials and in turn help to identify new elite material suitable for interogression into breeding programmes.

In recent years, several Malaysian companies have embarked on oil palm genome sequencing projects to enhance the understanding of the crop with the aim of developing high-yielding and more disease-resistance oil palm. The Asiatic Centre for Genome Technology Sdn. Bhd. (ACGT) and its partner Synthetic Genomics Inc. (SGI) announced the completion of the first draft of the assembly and annotation of the oil palm genome in May, 2008 (Lee and Cheah, 2009). A year later (May 2009) another private company, Sime Darby Berhad announced that they have successfully sequenced, assembled and annotated the oil palm genome with 93.8% completeness through collaboration with Synamatix Sdn. Bhd. (The Star Online, 2009; Sime Darby, 2009).

Being the leading oil palm research and development centre of Malaysia, MPOB has carried out extensive research work on oil palm. They constructed a linkage map and discovered an RFLP marker for *Sh* gene by using progenies derived from a self-

pollinated *tenera E. guineensis* palm (Palm T128) (Singh and Cheah, 2004). Through its collaboration with Orion Genomics for sequence analysis, a subset of 30,000 non-repetitive high quality SNPs were identified and selected by MPOB. Based on progenies from self-pollination of the same *tenera* (Palm T128), a linkage map with 16 linkage groups was constructed using the selected SNPs, RFLP, AFLP and SSR markers. Four different SNP markers were found to map on either side of the *Sh* gene with the closest marker (SNPM00310) at a distance of 2.2 cM (Singh, 2010).

Recently, MPOB published the 1.8 gigabase (Gb) genome sequence of the African oil palm *E. guineensis* and the draft sequence of South American oil palm *E. oleifera* (Singh *et al.*, 2013a) as well as the identification of *SHELL* gene (Singh *et al.*, 2013b). The combined total length of the assembly is 1.535 Gb which comprises nearly 35,000 genes, including the oil biosynthesis gene and other transcriptional regulators highly expressed in the kernel. The authors commented that the genome sequence will facilitate identification of genes responsible for important yield and quality traits as well as somaclonal epigenetic alterations.

Chapter 3

Approaches to develop

Shell-thickness marker(s) using

Representational Difference

Analysis (RDA) and Next-

Generation Sequencing (NGS)

3.1 Introduction and objective

Subtractive hybridization has been used to find the difference between two samples or genomes of interest in which DNA from sample A was hybridized against an excess DNA from sample B to remove common sequences between the two samples, thereby enabling enrichment of "target" sequences unique to sample A. Sample A is termed the tester while sample B is the driver. However, subtractive hybridization was found to be inefficient for comparison of high complexity genomic DNAs (Lisitsyn, 1995; Lisitsyn and Wigler, 1995). Representational Difference Analysis (RDA) was devised in 1993 to overcome this limitation (Lisitsyn *et al.*, 1993).

RDA consists of three important steps, which are production of the genomic representations, subtractive hybridization and kinetic enrichment. Representation is a process of generating a subpopulation of the genomes of interest with reduced complexity through restriction digestion of genomic DNA, adaptor ligation followed by 'whole-genome' amplification of the representation. Subsequent subtractive hybridization will eliminate common fragments present in both tester and driver populations, leaving only the differences present for further kinetic enrichment. Successive iterations of the subtraction and PCR amplification allow enrichment of the target sequence of interest (Lisitsyn *et al.*, 1993; Hubank and Schatz, 1994; Lisitsyn, 1995).

In order to locate polymorphism related to a gene of interest using the RDA technique, it is crucial to have tester and driver samples that differ primarily in the region of the target gene (Lisitsyn *et al.*, 1995). For self-compatible plant species, this can be achieved by production of near-isogenic lines. However, it is not simple to

produce pure lines for out-crossing plants due to inbreeding depression (Schneider, 2005). Bulked Segregant Analysis (Michelmore *et al.*, 1991) was developed to overcome the problem of lack of availability of near-isogenic lines for both inbreeding and outbreeding species.

The Bulked Segregant Analysis (BSA) approach was developed by Michelmore *et al.* in 1991 in which DNA samples of individuals derived from a segregating population of a single cross are pooled so that within each pool, the individuals have the same trait or gene of interest but are arbitrary for all other genes. Therefore, the two bulked samples differ only for the region of interest and surrounding DNA which has not undergone genetic recombination during meiosis, but are heterozygous for all other regions and the contrasting bulks can be analysed by comparison to identify markers for that particular region. This approach has been shown to work well for genes with major effects in which markers tightly-linked to the gene of interest will show significant differences in allele frequency between the two DNA bulks (Quarrie *et al.*, 1999).

The BSA method has previously been used successfully to identify RAPD (Moretzsohn *et al.*, 2000) and AFLP markers (Billotte *et al.*, 2001a, b) linked to the *Sh* gene. This indicates that the method is suitable to use in combination with any marker system for the study of shell-thickness trait. BSA approach was also used in combination with AFLP for the study of *Virescens* trait (fruit skin color) in oil palm by creating two different DNA bulks of ten palms each (Seng *et al.*, 2007).

Introduction of massively parallel DNA sequencing platforms, termed Next-Generation Sequencing (NGS), has striking impact on recent scientific discoveries. NGS approaches reduce the cost and speed of DNA sequencing by several orders of

magnitude allowing sequencing of genome-wide scale and ultra-resolution of single base precision (Shendure and Ji, 2008). Illumina SOLEXA and Applied Biosystem SOLiD sequencing platform produce short-read sequences (35-100 bases) that are frequently used for resequencing in which reads can be aligned against a reference genome or transcriptome. Meanwhile, Roche 454 pyrosequencing is more commonly applied for non-model organism sequencing projects as the longer reads generated (1 kb) are more amenable for *de novo* assembly (Kumar and Blaxter, 2010).

The objective of this study is to exploit RDA together with BSA to develop markers for the shell-thickness gene that determines the segregation pattern of *dura*, *pisifera* and *tenera* fruit forms. *Dura* and *pisifera* samples can be bulked together according to their fruit form. In the present study, four different controlled crosses were exploited and each *dura* and *pisifera* bulk was consisted of ten palms from the same controlled cross. The use of multiple bulks allows the identification of consistent markers to the shell-thickness gene.

This project also aims to identify RDA difference products using NGS approach, 454 pyrosequencing. Conventionally, RDA difference products are cloned into plasmid, transformed into bacteria and sequenced by Sanger sequencing. Combination of RDA approach with high sensitivity 454 pyrosequencing would allows more comprehensive understanding of the enrichment profile generated besides eliminating the laborious transformation procedure.

3.2 Materials and Methods

3.2.1 Plant materials

Four different oil palm controlled crosses were exploited in this study, 744, 768, 769 and 751. Oil palm fruit were characterised phenotypically and frond one leaf was sampled from oil palm crosses obtained from the Paloh Estate of Advanced Agriecological Research Sdn. Bhd. (AAR) in Johore, Malaysia. The Deli *dura* from the 744 and *dura* and *pisifera* from the 769 controlled crosses were collected in 2008 while the *dura* and *pisifera* from the 768 and 751 controlled crosses were collected in November 2009. Ten samples were collected for each fruit category, Deli *dura*, *dura* and *pisifera*. The list of samples is shown in Table 3.1.

After cutting the leaves from individual palms, the leaves were cleaned with 70% ethanol (EtOH), cut into small pieces and packed into plastic bags. All plastic bags were clearly labelled and stored at -80 ºC.

Table 3.1: List of samples collected from the Paloh Estate of AAR in Johore, Malaysia.

	744	769		768		751	
No.	Deli *Dura*	*Dura*	*Pisifera*	*Dura*	*Pisifera*	*Dura*	*Pisifera*
1.	744/131	769/8	769/1	768/28	768/32	751/7	751/26
2.	744/132	769/12	769/19	768/31	768/34	751/8	751/27
3.	744/133	769/23	769/21	768/35	768/43	751/22	751/29
4.	744/134	769/24	769/27	768/41	768/45	751/25	751/30
5.	744/135	769/35	769/40	768/42	768/46	751/28	751/31
6.	744/150	769/36	769/44	768/44	768/50	751/39	751/34
7.	744/152	769/39	769/52	768/49	768/51	751/40	751/43
8.	744/153	769/43	769/53	768/56	768/52	751/42	751/44
9.	744/154	769/49	769/54	768/57	768/58	751/45	751/48
10.	744/162	769/55	769/57	768/60	768/59	751/46	751/49

3.2.2 **Extraction of genomic DNA**

Genomic DNA was extracted from leaves using the cetyltrimethylammonium bromide (CTAB) method described by Doyle and Doyle (1987) with minor modifications. Four grams of leaf sample were ground into fine powder in liquid nitrogen in a pre-chilled mortar and pestle. The samples were then transferred into a 50 ml falcon tube containing 10 ml of modified CTAB lysis buffer [2% (w/v) CTAB, 100 mM Tris-HCl pH 8, 20 mM ethylenediaminetetraacetic acid (EDTA) pH 8, 140 mM sodium chloride (NaCl), 2% (w/v) polyvinylpyrrolidone-40 (PVP-40), 5 mM ascorbic acid, 4 mM diethyldithiocarbamate sodium (DIECA) and 0.4% (v/v) 2-β mercaptoethanol]. The tubes were incubated at 60 ℃ in a water bath for an hour with tubes being inverted and mixed 5-6 times during this interval to ensure complete lysis. After that, the tubes were left to cool down to room temperature for about 15 min before an equal volume of chloroform: iso-amylalcohol (24:1) was added. The tubes were gently inverted by putting on a shaker (N-Biotek, Inc), at 60 rpm, for 30 min and followed by centrifugation (Sigma 3-18K) at 4700 rpm, 25 ℃ for 15 min. The upper aqueous phase was transferred to a new tube and chloroform: iso-amylalcohol (24:1) extraction was repeated until a clear interface was obtained. The aqueous phase was transferred to a new tube and 0.6 volumes of ice-cold isopropanol was added and mixed well. The mixture was incubated at -80 ℃ for at least half an hour for DNA precipitation. After that, centrifugation was performed at 4700 rpm, 4 ℃ for 15 min. The pellet was washed with 10 ml wash buffer [76% (v/v) absolute ethanol (EtOH) and 10 mM ammonium acetate] followed by shaking at room temperature for 30 min and centrifugation at 4700 rpm, 4 ℃, for 15 min. The washing step was repeated once before the pellet was vacuum dried (Concentrator plus, Eppendorf). The pellet was

resuspended in 2 ml of TE buffer (10 mM Tris-HCl, pH 8.0 and 1 mM EDTA) at 4 ºC, overnight.

The following day, extractions were checked to ensure that the pellet had resuspended before addition of 1.25 µl of RNase (10 µg/ml) into each tube. The tubes were incubated at 37 ºC for at least 30 min. A 0.5 volumes of 7.5 M filtered-sterile ammonium acetate (pH 7.7) was added to the solution, mixed well and the tubes were left on ice for 30 min. Then, centrifugation was carried out at 4600 rpm, 4 ºC, for 15 min. The RNase-treated samples were transferred to a new tube and 2.5 volumes of cold absolute EtOH were added. The mixture was mixed gently by inverting the tube and incubated at -80 ºC for an hour. DNA was precipitated by centrifugation at 4700 rpm, 4 ºC, for 15 min. The pellet was washed twice with 70% (v/v) EtOH and vacuum-dried. The pellet was then re-dissolved in 500 µl of TE buffer. The quality and quantity of the extracted genomic DNA was checked by electrophoresis on a 1% agarose gel (Vivantis) containing SYBR® Safe DNA gel stain (Invitrogen) in 1x TAE buffer (40 mM Tris, 20 mM acetic acid and 1 mM EDTA, pH 8.0) followed by visualization on a FluorChem HD2 Multi Image II (Alpha Innotech) and an OD measurement at 260 nm wavelength (Nanodrop 1000 Spectrophotometer, Thermo Scientific). DNA was then stored at -20 ºC.

3.2.3 Fingerprinting analysis of samples and pooling of samples

Samples of the same controlled cross were pooled according to their shell-thickness trait for further analysis. It is important to ensure all palms come from the same cross and no outcross is present before sample bulking. For this reason, all genotypes were fingerprinted before bulking.

Thirteen oil palm SSR primer sets were selected from the LINK2PALM (L2P) EU FP5 ICO-DEV (http://www.neiker.net/link2palm/) (Appendix A) to fingerprint all the *dura* and *pisifera* samples from the 768, 769 and 751 controlled crosses. Parents of the three controlled crosses, 228/05, 228/06 and 138/04, were included to serve as controls for identification of outcrosses. Samples with bands that were different from their parent were regarded as outcrosses. All samples were sent to the University of Nottingham Sutton Bonington UK campus for genotyping.

Samples of each controlled cross with proven identity were bulked by mixing the same total DNA amount of each individual to form the respective *dura* and *pisifera* bulk of the 769, 768 and 751 controlled crosses. The Deli *dura* parent of the 744 controlled cross was no longer available; hence progeny were pooled to form Deli *dura* bulk without a legitimacy test. Bulked samples were checked on 1% agarose gel electrophoresis and OD measurement at 260 nm wavelength.

3.2.4 Representational Difference Analysis (RDA)

Whilst DNA samples were fingerprinted in Nottingham UK, optimization of RDA protocol and the first RDA analysis were performed using unproven Deli *dura* bulk from the 744 and unproven *dura* and *pisifera* bulk from the 769 controlled crosses. At this stage, the samples were bulked without identity verification. RDA analysis was repeated afterward using bulked samples with confirmed identity.

3.2.4.1 Optimization of the RDA protocol

a) <u>Optimization of restriction digestion</u>

Six different restriction endonucleases, *Bam*HI, *Eco*RI, *Hin*dIII, *Hpa*II, *Mse*I and *Pst*I (New England Biolabs), were tested for their suitability to use for the RDA analysis by digesting 400 ng of the pooled 744 Deli *dura* sample with 10 U/µg of enzyme in a total volume of 10 µl at 37 °C under various incubation times of 1, 3, 6 and 16 (overnight) h. Digestion profiles were analyzed on a 1% agarose gel with same amount of the undigested 744 Deli *dura* bulk loaded as negative control.

b)   Optimization of primer concentration in Polymerase Chain Reaction (PCR)

After genomic digestion of the DNA samples, adaptor primers were ligated to DNA fragments and the same adaptor primers were used to amplify the DNA ligation products.

Each set of adaptor primers (Bioneer) contains one long 24-mer oligonucleotide and one short 12-mer oligonucleotide (Table 3.2). The 24-mer adaptors are ligated to the 5'-ends of DNA fragments while the 12-mer adaptors are used to generate a double-stranded ligation template with the 24-mer adaptor. These 12-mer adaptors are not ligated to the DNA fragments due to lack of 5' phosphate group on the oligonucleotide. Therefore, during incubation at 72 °C, the 12-mer oligonucleotide within the adaptors are dissociated from DNA fragments and the 3'-ends of DNA are filled up subsequently by *Taq* DNA polymerase, generating the priming sites for the 24-mer adaptors. These 24-mer adaptors are used as the forward and reverse primer in each PCR condition. All adaptors are designed in such a way that they can be removed by the restriction endonuclease after amplification of DNA fragments.

Table 3.2: Sequences of oligonucleotides (adaptors) used in RDA.

Adaptor Pair Set	Name	Sequence
1	R Hind 24	5'- AGC ACT CTC CAG CCT CTC ACC GCA -3'
	R Hind 12	5'- AGC TTG CGG TGA -3'
2	J Hind 24	5'- ACC GAC GTC GAC TAT CCA TGA ACA -3'
	J Hind 12	5'- AGC TTG TTC ATG -3'
3	N Hind 24	5'- AGG CAG CTG TGG TAT CGA GGG AGA -3'
	N Hind 12	5'- AGC TTC TCC CTC -3'
1	R Bam 24	5'- AGC ACT CTC CAG CCT CTC ACC GAG -3'
	R Bam 12	5'- GAT CCT CGG TGA -3'
2	J Bam 24	5'- ACC GAC GTC GAC TAT CCA TGA ACG -3'
	J Bam 12	5'- GAT CCG TTC ATG -3'
3	N Bam 24	5'- AGG CAA CTG TGC TAT CCG AGG GAG -3'
	N Bam 12	5'- GAT CCT CCC TCG -3'

In order to amplify ligated-DNA fragments, optimization of primer concentration in the polymerase chain reaction (PCR) was performed using 0.8, 1, 2, 3 and 4 µM of the 24-mer adaptor as the forward and reverse primer in the reaction. Previously optimized *Hin*dIII-digested and R Hind adaptor-ligated pooled Deli *dura* from the 744 controlled cross was used as template and the PCR reaction was prepared using 8 ng of ligated DNA, 320 µM each of dNTPs mix (dATP, dGTP, dCTP and dTTP) (Promega), 0.1 mg/ml of bovine serum albumin (BSA) (New England Biolabs), 2 µl of 10x PCR buffer [100 mM potassium chloride, 100 mM ammonium sulphate, 20 mM Tris-HCl, 2 mM magnesium sulphate and 1% Triton X-100, pH 8.8] (New England Biolabs) and respective amount of R Hind 24 primer in a total volume of 20 µl. The tubes were incubated for 3 min at 72 ºC in a preheated thermal cycler (G-storm Thermal Cycler, Gene Technologies) for dissociation of the 12-mer oligonucleotide. To fill in the 3'-recessed ends of the ligated fragments, 3 U of *Taq* DNA polymerase (New England Biolabs) was added to each tube, mixed by

pipetting and incubated at 72 ºC for another 5 min. The mixture was immediately amplified by PCR for 20 cycles of denaturation at 95 ºC for 1 min and annealing/extension at 72 ºC for 3 min, followed by a final extension of 10 min at 72 ºC and an indefinite soak at 4 ºC. PCR profiles were separated and analyzed on a 1.5 % agarose gel electrophoresis.

3.2.4.2 Generation of first amplicons (representation)

a) Digestion of genomic DNA

To produce representations (amplicons) for RDA, 2 µg of both tester and driver DNA were digested with 10 U/µg of the restriction endonuclease, *Bam*HI or *Hin*dIII, in a 40 µl mixture and incubated overnight at 37 ºC. Digested DNA was then purified using a GeneAll Expin™ Clean Up SV Mini kit (GeneAll Biotechnology) and eluted using 50 µl of pre-warmed EB buffer according to manufacturer's instructions. DNA was analyzed on a 1% agarose gel and the DNA concentration was quantified by measuring OD at 260 nm wavelength.

b) Ligation of the R 24- and 12-mer adaptor set

For the adaptor ligation reaction, 1 µg of *Bam*HI- or *Hin*dIII-digested DNA was mixed with 0.5 nmol of adaptor pair R Bam 12 and R Bam 24 or R Hind 12 and R Hind 24, respectively (Table 3.2, primer set 1) in a total volume of 30 µl T4 DNA ligase buffer (New England Biolabs). The ligation mixture was incubated at 55 ºC for 5 min in a heating block followed by gradual cooling of the mixture to 10 ºC for approximately 1 h to allow annealing of the oligonucleotides. Condensation was collected by a brief spin. Four hundred units of T4 DNA ligase were then added to the mixture and incubated overnight at 16 ºC for ligation of adaptors to DNA fragments.

c) <u>**Amplification of the tester and driver DNA**</u>

After overnight ligation, DNA was diluted with 970 µl of TE buffer. To generate the first-round amplicons, tester and driver ligation products were amplified by PCR with the R Bam 24 or R Hind 24 as primer. PCR reactions were set up with each PCR tube containing 40 ng of ligated DNA, 1 µM of R 24-mer adaptor, 320 µM each of dNTPs mix, 0.1 mg/ml of BSA in 1x PCR buffer of 100 µl total volume. As previously described in section 3.2.4.1(b), the tubes were incubated for 3 min at 72 ºC in a preheated thermal cycler before adding 15 U of *Taq* polymerase into each tube to fill in the 3'-recessed ends of the ligated fragments. After a further 5 min incubation at 72 ºC, the mixture was immediately amplified by 20 cycles of PCR.

Tester and driver amplicons were purified separately using GeneAll® Expin™ PCR SV Mini kit (GeneAll Biotechnology). The quality of PCR products were analyzed on 1% agarose gel and the DNA concentration was again measured by OD measurement at 260 nm wavelength.

Prior to the subtractive hybridization step, adaptors were removed from both driver and tester amplicons to prevent driver amplicons from forming end-annealed complexes during hybridization. To remove R adaptors, all first round tester and driver amplicons were subjected to overnight restriction digestion at 37 ºC using 10 U of *Bam*HI or *Hin*dIII enzyme for each µg of amplicon. Digested DNA fragments were then purified by a GeneAll® Expin™ Cleanup SV Mini kit and quantified by OD measurement at 260 nm wavelength.

3.2.4.3 Subtractive hybridization

a) <u>**Change of adaptors for tester amplicons**</u>

Another two different sets of adaptors (J and N) (Table 3.2, primer sets 2 and 3) were designed for every restriction endonuclease and these two sets of adaptors were used alternatively in the hybridization steps, meaning the J adaptor set (Table 3.2, primer set 2) was used for round 1 and round 3 of enrichment while the N adaptor set (Table 3.2, primer set 3) was used for round 2 instead, preventing any potential carry over before subsequent rounds of RDA.

Tester amplicons were ligated to J Hind or J Bam adaptors and amplified using the same protocol mentioned in sections 3.2.4.2 (b) and 3.2.4.2 (c). Amplified tester amplicons were purified by GeneAll® Expin™ PCR SV Mini kit and no removal of adaptors was required. The annealing/extension temperature for the J Hind 24 primer was set at 70 ºC while 72 ºC was used for J Bam 24 primer. It should be noted that only the tester amplicons are ligated to defined oligonucleotides prior to the hybridization step, but not the driver amplicons.

b) **First round of subtractive hybridization**

Five hundred ng of the J adaptor ligated-tester amplicons were mixed with 40 µg of driver amplicons (tester: driver ratio of 1:80) for the first round of subtractive hybridization. A one-tenth volume of 3M sodium acetate (NaOAC), pH 5.2 and 3 volume of absolute EtOH were added to the mixture followed by incubation at -80 ºC for 60 min. DNA precipitation was carried out by centrifugation at 13,000 rpm, 4 ºC, for 30 min. The pellet was washed twice with 70% (v/v) EtOH and centrifuged at 13,000 rpm, 4 ºC, for 15 min. The DNA pellet was vacuum dried before resuspension with 4 µl of 3x EE buffer [30 mM (2-hydroxyethyl piperizine)-N'-(3-propene sulfonic acid) (EPPS), pH 8 and 3 mM EDTA]. DNA solutions were denatured at 98 ºC for 5

min in a thermal cycler and 1 µl of 5 M NaCl was added. The mixture was incubated at 67 ºC for 20 h to allow hybridization process to occur.

c) **Selective amplification**

At the end of the hybridization, DNA was diluted to a 0.1 µg/µl concentration by adding 395 µl of TE buffer. Four tubes of PCR reaction were set up for each subtractive hybridization reaction, containing 40 µl of diluted hybridized DNA (4 µg), 0.32 mM dNTPs mix, 0.1 mg/ml BSA and 10 µl of 10x PCR buffer in a total volume of 100 µl. The reaction mixture was incubated without primer at 72 ºC for 3 min in a preheated thermal cycler before addition of 15 U of *Taq* DNA polymerase and incubation at 72 ºC for another 5 min. This step is necessary to fill in and reform the adaptor ends of re-annealed tester that is the priming site for exponential enrichment of difference products. Hybridized DNA was amplified for 10 cycles (1 min at 95 ºC, 3 min at 72 ºC, and held for 10 min more for the last cycle) after addition of 1.55 µM of J Hind 24 or J Bam 24 primer, according to the reaction. For the J Hind 24 primer, annealing/polymerization temperature of 70 ºC was used instead.

PCR products were purified with the kit and DNA was eluted with 50 µl of elution buffer. Twenty microliters of amplified product were incubated with 20 U of mung bean nuclease (New England Biolabs) at 30 ºC for 30 min in a total volume of 40 µl to degrade single-stranded DNA molecules present after amplification. The reaction was stopped by adding 160 µl of 50 mM Tris-HCl (pH 8.0) and the nuclease was heat inactivated for 5 min incubation at 98 ºC. Forty microliters of the nuclease-treated products were amplified for another 20 cycles under the same conditions as before the mung bean treatment. The PCR products were purified by kit and the resulting amplicons were termed the First Difference Product.

d)   Subsequent subtractive hybridization

For the second round of subtractive hybridization, difference products from the first round were digested with the original restriction endonuclease (*Bam*HI or *Hin*dIII), ligated and amplified with the N adaptor pair (Table 3.2, primer set 3). One hundred nanograms of the N adaptor-ligated difference products were mixed with 40 µg driver and the hybridization and kinetic enrichment process were repeated as in the first cycle. This second hybridization was done at a tester: driver ratio of 1: 400. For the third round, J adaptor set (Table 3.2, primer set 2) was ligated to restriction enzyme digested-products from round two. Two hundred pictograms of these J adaptor-ligated difference products were then mixed with 40 µg of driver, a tester: driver ratio of 1: 200,000. Subtraction hybridization and amplification were repeated again with the first kinetic amplification done at 15 cycles instead of 10; and after mung bean degradation, 30 cycles of final amplification was carried out compared to the previous 20 cycles due to the low amount of tester present.

Difference products from all three rounds of subtractive hybridization were electrophoresed on a 3% agarose gel to analyse the enrichment profile.

3.2.4.4   Cloning and sequencing of difference products

Difference products from the round 3 subtractive hybridization were cloned and sequenced to examine their nucleotide composition and identify any potential sequence which could be used as possible markers close to the shell-thickness gene.

a)   Gel extraction

Difference products from round 3 were electrophoresed on a 2% agarose gel in 1x TAE. Difference products within the size range of 200 to 450 bp were excised using a clean razor blade under longwave UV-light. Exposure to shortwave UV-light was minimized to prevent formation of pyrimidine dimers. The excised agarose gel was purified using GeneAll® Combo Gel and PCR purification (GeneAll Biotechnology) according to manufacturer's instruction.

b)   TA cloning

Based on a 3:1 insert:vector molar ratio, the purified difference products were ligated into pGEM-T easy vector (Promega) using T4 DNA ligase. This ligation mixture was incubated overnight at 4 °C. The ligated products were transformed into JM109 competent cells. In brief, frozen competent cells were placed in an ice bath for about 5 min until just thawed and the cells were mixed by gentle flicking. Fifty microliters of competent cells were transferred into a sterile 1.5 ml microcentrifuge tube on ice follow by addition of 2 μl ligation reaction. After mixing them by flicking the tubes, the mixture was placed on ice for 20 min. The cells were then heat-shocked in a 42 °C water bath for 45 s and immediately returned to ice for 2 min. Room-temperature Super Optimal Broth (SOC medium) containing 2% (w/v) bactotryptone, 0.5% (w/v) bacto-yeast extract, 10 mM NaCl, 2.5 mM KCl, 20 mM Mg^{2+}stock and 20 mM glucose was added to the tubes to a total volume of 1 ml. The tubes were incubated at 37 °C for 1.5 h with 150 rpm shaking. One hundred microliters (10%) of each transformation culture was plated on LB plates with ampicillin/IPTG/X-Gal [15 g agar in 1 L of Luria-Bertani (LB) containing 1% (w/v) bacto-tryptone, 0.5% (w/v) bacto-yeast extract and 0.5% (w/v) NaCl, pH 7.0 with 100 μg/ml ampicillin, 0.5 mM isopropyl-β-D-thiogalactoside (IPTG) and 80 μg/ml 5-bromo-4-chloro-3-indolyl ß-D-

galactopyranoside (X-Gal)]. The plates were inverted and incubated overnight at 37 °C.

c)   <u>**Screening of positive transformants**</u>

Based on blue-white colony screening, white colony transformants were picked using a sterile toothpick and mixed into a 20 µl PCR reaction. The same toothpick was used to steak on LB plates with 100 µg/ml ampicillin for purification of colonies. The presence of inserts in the vector was confirmed by colony-PCR screening using the J Bam 24 or J Hind 24 primer. PCR was performed with an initial denaturation at 95 °C for 5 min, 30 cycles of denaturation at 95 °C for 1 min, annealing/extension at 72 °C (for *Bam*HI amplicons) or 70 °C (for *Hin*dIII amplicons) for 2 min and a final extension of 72 °C for additional 10 min. PCR products were analyzed on a 2% agarose gel to check the existence of insert. The LB plates were incubated overnight at 4 °C. Clones with inserts were inoculated in 3 ml of LB medium with 100 µg/ml ampicillin the day after and incubated overnight at 37 °C, with shaking at 200 rpm.

Plasmids were extracted from the overnight bacterial culture using the GeneAll® Exprep™ Plasmid Quick Kit (GeneAll Biotechnology, Korea). The presence of an insert and the insert size was again confirmed by PCR with the respective J Bam 24 or J Hind 24 primer. Sixty nanograms of plasmid in 50 µl of total reaction mixture were amplified with PCR conditions of initial denaturation at 95 °C for 2 min, 25 cycles of denaturation at 95 °C for 1 min, annealing/extension at 72 °C (for *Bam*HI amplicons) or 70 °C (for *Hin*dIII amplicons) for 2 min and final extension of 72 °C for an additional 10 min. PCR products were analyzed on a 3% agarose gel together with the corresponding round three difference products.

Plasmids with an insert size in the expected size range were then sent to Macrogen Inc. (South Korea) for sequencing using the T7 primer located on the pGEMeasy vector.

d) <u>**Analysis of sequences obtained**</u>

The resulting sequences of round 3 reciprocal subtractive hybridization of both *Bam*HI and *Hin*dIII amplicons were analyzed using a multiple sequence alignment program, ClustalW2, available on the European Bioinformatics Institute website of European Molecular Biology Laboratory (EMBL-EBI) (http://www.ebi.ac.uk/) to study the identities, similarities and differences between sequences.

Homology searches against sequences available in the GenBank database were also performed for each sequence using the BlastN procedure (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (Altschul *et al.*, 1990). Adaptor sequences (24-mer) were removed from the sequences before the homology search except for the recognition site of the *Bam*HI or *Hin*dIII restriction enzymes.

3.2.4.5 **Assessment of the RDA technique with positive control**

The effectiveness of the RDA technique was tested using a positive control. *Hin*dIII digestion of *Lambda* DNA gives rise to nine distinct bands, which are 125, 564, 2027, 2322, 4361, 6557, 9416 and 23130 bp. The 125 bp fragment was selected as the positive control for RDA and this fragment was used to spike the control sample containing the tester DNA.

*Hin*dIII-digested *Lambda* DNA ladder (Fermentas) was electrophoresed on a 1% agarose gel containing 1x SYBR Safe DNA stain in 1x TAE buffer. Band 125 bp was excised using a clean razor blade under longwave UV-light. The excised agarose gel was purified using the GeneAll® Combo Gel and PCR purification kit according to manufacturer's instruction. The 125 bp fragment was added into control samples of the *Hin*dIII-digested genomic bulks of the legitimate *dura* and *pisifera* tester of the 769 controlled cross at a molecular level of 1, 10, 100 and 1000 copies. Reciprocal RDA analysis was performed as mentioned in sections 3.2.4.2 and 3.2.4.3 for three rounds of subtractive hybridization. Difference products were analyzed through electrophoresis on 3% agarose gel and visualized using UV.

A specific primer pair, Lambda125F (5'-AAG CTT GGC TTG GAG CCT G-3') and Lambda125R (5'-GAG CTT AGA ACC TTT ACC AAA GG-3'), were designed for the 125 bp fragment from *Hin*dIII-digested *Lambda* DNA. This primer pair was used to detect the presence of the positive control, 125 bp fragment, in the tester as well as the difference products after subtractive hybridization. PCR was carried out with each tube containing 0.5 µl of tester amplicons or round 3 difference products, 0.5 µM each of forward and reverse primer, 0.32 mM dNTP, 1.5 mM MgCl$_2$, 1 U of *Taq* polymerase and 1x PCR buffer in total volume of 25 µl. The excised 125 bp fragment and round 3 difference products of *dura* tester against *pisifera* driver of the 768 controlled cross (without *Lambda* DNA added) were used as positive and negative control, respectively. PCR amplification was performed with initial denaturation at 95 ºC for 3 min, 25 cycles of denaturation at 95 for 30 s, annealing at 70.5 ºC for 1 min and extension at 72 ºC for 30 s, followed by a final extension at 72 ºC for 10 min. PCR products were checked by electrophoresis on a 3% agarose gel.

3.2.5 **454 pyrosequencing of round 2 and 3 difference products**

454 pyrosequencing was employed to sequence the round 2 and 3 difference products alongside conventional bacterial cloning followed by Sanger sequencing. Three different reciprocal RDA analyses were studied; these include Deli *dura* of the 744 against *pisifera* of the 769 controlled cross (first RDA analysis); reciprocal analysis of the 769 controlled cross, with and without outcrosses (first and second RDA analysis).

In order to combine all the difference products of RDA into a single sample for 454 pyrosequencing, the J and N adaptor sequences were modified such that by having a single base change, a series of adaptor sequences were generated (Table 3.3, Bioneer) and could be used to amplify the final products for sequencing. This allows individual reactions to be identified after sequencing (Table 3.4). This principle has been proven (Mayes *et al.*, unpublished data; Ho *et al.*, 2013).

Phusion polymerase was used to create blunt-end products for 454 sequencing. Final products were generated with each PCR tube containing 40 µl of nuclease-treated products, 1.25 µM primer, 0.32 mM dNTP, 1 U Phusion High- Fidelity polymerase (Finnzymes) and 20 µl of 5x Phusion HF buffer in total volume of 100 µl. PCR amplification was carried out as follows, initial denaturation at 98 ºC for 30 s, 35 cycles of denaturation at 98 ºC for 20 s and annealing/extenstion at 72 ºC for 90 s, with final extension at 72 ºC for 10 min. PCR products were purified with the kit. Quality and quantity of purified products were checked by electrophoresis on a 1.5% agarose gel and OD measurement at 260 nm wavelength. Equal amount of purified PCR products were pooled with each contributing 2 µg DNA.

Table 3.3: Series of oligonucleotide primers for 454 sequencing based on single base modification of N and J 24 primers.

Primer Name	Primer sequences
NBam24A	5'-AGG CAA CTG TGC TAT CCG AGG GAG-3'
NBam24B	5'-AGG CAA CTG TGC TA<u>A</u> CCG AGG GAG-3'
NBam24C	5'-AGG CAA CTG TGC <u>A</u>AT CCG AGG GAG-3'
NBam24D	5'-AGG CAA CTG TG<u>G</u> TAT CCG AGG GAG-3'
NBam24E	5'-AGG CAA CT<u>C</u> TGC TAT CCG AGG GAG-3'
NBam24F	5'-AGG CAA CTG TGC TAT C<u>G</u>G AGG GAG-3'
NHind24A	5'-AGG CAG CTG TGG TAT CGA GGG AGA-3'
NHind24B	5'-AGG CAG CTG TGG TA<u>A</u> CGA GGG AGA-3'
NHind24C	5'-AGG CAG CTG TGG <u>A</u>AT CGA GGG AGA-3'
NHind24D	5'-AGG CAG CTG TG<u>C</u> TAT CGA GGG AGA-3'
NHind24E	5'-AGG CAG CT<u>C</u> TGG TAT CGA GGG AGA-3'
NHind24F	5'-AGG CAG CTG T<u>C</u>G TAT CGA GGG AGA-3'
JBam24A	5'-ACC GAC GTC GAC TAT CCA TGA ACG-3'
JBam24B	5'-ACC GAC GTC GAC <u>A</u>AT CCA TGA ACG-3'
JBam24C	5'-ACC GAC GTC GAC TA<u>A</u> CCA TGA ACG-3'
JBam24D	5'-ACC GAC GTC GA<u>G</u> TAT CCA TGA ACG-3'
JBam24E	5'-ACC GAC GT<u>G</u> GAC TAT CCA TGA ACG-3'
JBam24F	5'-ACC GA<u>G</u> GTC GAC TAT CCA TGA ACG-3'
JHind24A	5'-ACC GAC GTC GAC TAT CCA TGA ACA-3'
JHind24B	5'-ACC GAC GTC GAC <u>A</u>AT CCA TGA ACA-3'
JHind24C	5'-ACC GAC GTC GAC TA<u>A</u> CCA TGA ACA-3'
JHind24D	5'-ACC GAC GTC GA<u>G</u> TAT CCA TGA ACA-3'
JHind24E	5'-ACC GAC GT<u>G</u> GAC TAT CCA TGA ACA-3'
JHind24F	5'-ACC GA<u>G</u> GTC GAC TAT CCA TGA ACA-3'

(Underling indicates the base change)

Table 3.4: Corresponding template DNA for each modified 454 primers.

Primer	Template DNA (Mung bean treated-products)
A	744 Deli *dura* (tester) against 769 *pisifera* (driver)
B	769 *dura* (tester) against 769 *pisifera* (driver), with outcross D36
C	769 *dura* (tester) against 769 *pisifera* (driver), without outcross D36
D	769 *pisifera* (tester) against 744 Deli *dura* (driver)
E	769 *pisifera* (tester) against 769 *dura* (driver), with outcross D36
F	769 *pisifera* (tester) against 769 *dura* (driver), without outcross D36

Pooled samples were analyzed on a 2% agarose gel in 1x TAE buffer and bands within the size range of 200 to 700 bp were excised and purified using the kit. This gel-purified sample was sent for a 1/16th run 454 sequencing using Roche Genome Sequencer (Centre for Genetics and Genomics, University of Nottingham).

3.2.5.1   454 pyrosequencing data analysis

Raw sequencing data was divided into clusters based on the modified J and N adaptor sequences by the service provider before the result was received. *De novo* assembly was performed using CLC Genomics Workbench v5.0 (CLC Bio). Adaptor sequences were removed and contigs from different pools of RDA analysis were compared with the use of BioEdit Sequence Alignment Editor Version 7.1.11 (Hall, 1999).

By using CLC Genomics Workbench, homology searches for each R3 contig were performed against an in-house database of oil palm mesocarp transcriptome (Mayes *et al.*, unpublished data), *Phoenix dactylifera* (date palm) genome (Al-Dous *et al.*, 2011) as well as *Oryza sativa* (rice) and *Arabidopsis thaliana* database available in NCBI website using BlastN, with probability scores below 1e-10 considered to be potentially significant. In addition, all the R3 contig sequences were also sent to MPOB for homology analysis against MPOB *Pisifera* Assembly V5 and the repetitive DNA elements database TIGR and RepBase. MPOB has previously anchored the *pisifera* genome assembly to their genetic linkage map T128. For any particular scaffold that a contig has significant hits to, the linkage group that the scaffold belongs to was also disclosed.

3.3   Results

3.3.1   Fingerprinting analysis and generation of DNA bulks

Fingerprinting analysis using 13 polymorphic CIRAD SSR markers identified one illegitimate sample from each controlled cross (Tables 3.5-3.7). The outcrosses were *dura* 36 from the 769, *dura* 28 from the 768 and *pisifera* 48 from the 751 controlled cross. They were considered as outcrosses because at least one of the primers screened was found to have bands that were not present in their respective parents. In fact, D36 from the 769 controlled cross had 10 out of 13 primer pairs showing incorrect banding (Table 3.5) while D28 from the 768 (Table 3.6) and P48 from the 751 controlled cross (Table 3.7) had seven and six primer pairs showing incorrect bands, respectively. The rest of the samples were consistent with being derived from the appropriate *tenera* self-pollinated. The three outcrosses were excluded from the DNA bulks for further study.

Table 3.5: Fingerprinting analysis of *dura* and *pisifera* samples from the 769 controlled cross using 13 CIRAD SSR primers.

	OP1	OP5	OP13	OP12	OP11	OP24/6	OP2	OP7	OP20	OP18	OP21	OP29	Overall
F1 228/06	205/219	236/(257?)	314/320	184/192/206	253	177/181/190	152/154	242/246	NA	292/294	225/240	122/134	
769/8 (D)	205/219 ✓	236/257 ✓	314/320 ✓	184/192/206 ✓	253 ✓	177/181/190 ✓	154 ✓	242/246 ✓	225 ✓	292/294 ✓	225/240 ✓	134 ✓	✓
769/12 (D)	205/219 ✓	236 ✓	314/320 ✓	184/192/206 ✓	253 ✓	181/190 ✓	154 ✓	246 ✓		292 ✓	225 ✓	122/134 ✓	✓
769/23 (D)	205 ✓	236 ✓	314/320 ✓	184/206 ✓	253 ✓	190 ✓	152/154 ✓	242 ✓	225 ✓	292 ✓	225/240 ✓	122/134 ✓	✓
769/24 (D)	205/219 ✓	236 ✓	314 ✓	184/206 ✓	253 ✓	177/181/190 ✓	152/154 ✓	242/246 ✓	225 ✓	292 ✓	225/240 ✓	134 ✓	✓
769/35 (D)	205	236 ✓	320 ✓	184/206 ✓	253	177/181 ✓	152 ✓	242/246 ✓	NA	294	225/240 ✓	122/134 ✓	✓
769/36 (D)	205/(217) ✗	236/257 ✓	314 ✓	192/(200)/206 ✗	253 ✓	177/181 ✓	160/164 ✗	242/246 ✓	NA	292/294 ✓	240 ✓	122/134 ✓	✗
769/39 (D)	205 ✓	236 ✓	314 ✓	184/192/206 ✓	253 ✓	177/181/190 ✓	152 ✓	242 ✓	NA	292/294 ✓	225/240 ✓	134 ✓	✓
769/43 (D)	205/219 ✓	236 ✓	314 ✓	192/206 ✓	253 ✓	177/181/190 ✓	154 ✓	242 ✓	NA	294	225/240 ✓	122/134 ✓	✓
769/49 (D)	205 ✓	236 ✓	320 ✓	192/206 ✓	253 ✓	177/181 ✓	154 ✓	242 ✓	NA	292 ✓	225 ✓	122 ✓	✓
769/55 (D)	205/219 ✓	236/257 ✓	320	184/192/206 ✓	253 ✓	177/181/190 ✓	154 ✓	242/246 ✓	225 ✓	292/294 ✓	225 ✓	134 ✓	✓
769/1 (P)	205/219 ✓	236 ✓	NA	184/192/206 ✓	253 ✓	181/190 ✓	152 ✓	246 ✓	225 ✓	292/294 ✓	NA	122/134 ✓	✓
769/19 (P)	219 ✓	236/257 ✓	314 ✓	184/206 ✓	253 ✓	181/190 ✓	152 ✓	246 ✓	225 ✓	292/294 ✓	NA	122/134 ✓	✓
769/21 (P)	NA	236/257 ✓	314 ✓	184/192/206 ✓	253 ✓	177/190 ✓	152 ✓	246 ✓	225 ✓	294	225/240 ✓	134 ✓	✓
769/27 (P)	205/219 ✓	236 ✓	320 ✓	192/206 ✓	253 ✓	177/181/190 ✓	?	246 ✓	225 ✓	292 ✓	225/240 ✓	122 ✓	✓
769/40 (P)	219 ✓	236 ✓	314/320 ✓	192/206 ✓	253 ✓	181/190 ✓	152/154 ✓	246 ✓	NA	292/294 ✓	240 ✓	134 ✓	✓
769/44 (P)	205/219 ✓	236 ✓	314 ✓	192/206 ✓	253 ✓	177/181/190 ✓	152/154 ✓	242/246 ✓	NA	292 ✓	225/240 ✓	122/134 ✓	✓
769/52 (P)	205 ✓	236 ✓	NA	184/192/206 ✓	253 ✓	177/181/190 ✓	152 ✓	242/246 ✓	NA	294	225/240 ✓	122/134 ✓	✓
769/53 (P)	205/219 ✓	236/257 ✓	314/320 ✓	184/192/206 ✓	253 ✓	177/181 ✓	152 ✓	246 ✓	NA	294	225/240 ✓	134 ✓	✓
769/54 (P)	205/219 ✓	236 ✓	314/320 ✓	184/192/206 ✓	253 ✓	177/181/190 ✓	154 ✓	242/246 ✓	225 ✓	292/294 ✓	240	122/134 ✓	✓
769/57 (P)	205/219 ✓	236 ✓	320 ✓	184/206 ✓	253 ✓	177/181/190 ✓	152 ✓	246 ✓	NA	294		134 ✓	✓

√= consistent with the parent; x = not consistent with the parent, outcross

Table 3.6: Fingerprinting analysis of *dura* and *pisifera* samples from the 768 controlled cross using 13 CIRAD SSR primers.

	OP1		OP5		OP13		OP12		OP11		OP24/6		OP2		OP7		OP20		OP18		OP21		OP29		Overall
F1 228/05	213/219		236		314		192/200/206		240/253		177/181/190		164		240		225/246		292/303		225		116/122		
768/28 (D)	203	x	236/257	x	322	x	192	✓	253/258	x	181/190	✓	158/166	x	NA		222	x	294	x	225	✓	133	x	x
768/31 (D)	219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181/190	✓	164	✓	240	✓			292	✓	225	✓	116/122	✓	✓
768/35 (D)	213/219	✓	236	✓	314	✓	192/(200)/206	✓	(240)/253	✓	177/181/190	✓	164	✓	240	✓	246	✓	303	✓	225	✓	122	✓	✓
768/41 (D)	213/219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓	164	✓	240	✓	NA		292	✓	225	✓	116/122	✓	✓
768/42 (D)	213/219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181	✓	164	✓	240	✓	NA		292	✓	225	✓	116/122	✓	✓
768/44 (D)	219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓	164	✓	240	✓	246	✓	292	✓	225	✓	116/122	✓	✓
768/49 (D)	219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181/190	✓	164	✓	240	✓	225/246	✓	303	✓	225	✓	116	✓	✓
768/56 (D)	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	181/190	✓	164	✓	NA		NA		292	✓	225	✓	122	✓	✓
768/57 (D)	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	177/181	✓	164	✓	240	✓	NA		292	✓	225	✓	122	✓	✓
768/60 (D)	213	✓	236	✓	314	✓	192/(200)/206	✓	(240)/253	✓	177/181	✓	164	✓	240	✓	225	✓	292/303	✓	225	✓	122	✓	✓
768/32 (P)	213/219	✓	236	✓	314	✓	192/206	✓	253	✓	181/190	✓	164	✓	240	✓	NA		292/303	✓	225	✓	122	✓	✓
768/34 (P)	213/219	✓	236	✓	314	✓	192/200	✓	253	✓	181/190	✓	164	✓	240	✓	225/246	✓	292/303	✓	225	✓	116/122	✓	✓
768/43 (P)	213	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/(181)	✓	164	✓	240	✓	225	✓	292/303	✓	225	✓	116	✓	✓
768/45 (P)	219	✓	236	✓	314	✓	192/200	✓	253	✓	177/181	✓	164	✓	240	✓	NA		303	✓	225	✓	122	✓	✓
768/46 (P)	219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	181/190	✓	164	✓	240	✓	NA		292/303	✓	225	✓	122	✓	✓
768/50 (P)	219	✓	236	✓	314	✓	192/(200)/206	✓	240/253	✓	177/181/190	✓	164	✓	240	✓	NA		NA		303	✓	225	✓	116	✓	✓
768/51 (P)	213/219	✓	236	✓	314	✓	192/200	✓	253	✓	177/181	✓	164	✓	240	✓	NA		292	✓	225	✓	122	✓	✓
768/52 (P)	219	✓	236	✓	314	✓	192/200/206	✓	240/253	✓	177/181	✓	164	✓	240	✓	NA		303	✓	225	✓	116	✓	✓
768/58 (P)	213/219	✓	236	✓	314	✓	192/206	✓	240/253	✓	177/181	✓	164	✓	240	✓	225/246	✓	292/303	✓	225	✓	116	✓	✓
768/59 (P)	213/219	✓	236	✓	314	✓	192/(200)/206	✓	253	✓	177/181/190	✓	164	✓	240	✓	246	✓	292/303	✓	225	✓	122	✓	✓

√= consistent with the parent; x = not consistent with the parent, outcross

Table 3.7: Fingerprinting analysis of *dura* and *pisifera* samples from the 751 controlled cross using 13 CIRAD SSR primers.

	OP1		OP5		OP13		OP12		OP11		OP24/6		OP2		OP7		OP20		OP18		OP21		OP29		Overall
F1 138/04	217		257		314		206		253/258		183		154/160		240		NA		294		225/240		122		
751/7 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	160	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/8 (D)	217	✓	257	✓	314		206	✓	253/258	✓	183	✓	154/160	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/22 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	160	✓	240	✓	NA		294	✓	240	✓	122	✓	✓
751/25 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154	✓	240	✓	240		294	✓	225/240	✓	122	✓	✓
751/28 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154	✓	240	✓	240		294	✓	240	✓	122	✓	✓
751/39 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154/160	✓	240	✓	240		294	✓	240	✓	122	✓	✓
751/40 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	160	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/42 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/45 (D)	217	✓	257	✓	314		206	✓	253/258	✓	183	✓	154/160	✓	240	✓	NA		294	✓	240	✓	122	✓	✓
751/46 (D)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/26 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154/160	✓	240	✓	240		294	✓	225	✓	122	✓	✓
751/27 (P)	217	✓	257	✓	314	✓	206		253/258	✓	183	✓	154	✓	240	✓	240		294	✓	225/240	✓	122	✓	✓
751/29 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154/160	✓	240	✓	240		294	✓	240	✓	122	✓	✓
751/30 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154	✓	240	✓	240		294	✓	225	✓	122	✓	✓
751/31 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	160	✓	240	✓	NA		294	✓	240	✓	122	✓	✓
751/34 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154/160	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/43 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154	✓	240	✓	NA		294	✓	225/240	✓	122	✓	✓
751/44 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	154/160	✓	240	✓	NA		294	✓	240	✓	122	✓	✓
751/48 (P)	217	✓	236/257	x	314	✓	192/206	x	253		183/190	x	152/165	x	240/242	x	225		292/294	x	225	✓	122	✓	x
751/49 (P)	217	✓	257	✓	314	✓	206	✓	253/258	✓	183	✓	160	✓	240	✓	240		294	✓	225/240	✓	122	✓	✓

√= consistent with the parent; x = not consistent with the parent, outcross

After confirming the legitimacy of all samples, *dura* and *pisifera* of the same controlled cross, meaning that they are derived from a single segregating population, were pooled together according to their shell-thickness. On the other hand, the identity of *Deli* dura progenies of the 744 controlled cross was not fingerprinted due to the lack of an available parent. Figure 3.1 shows that all DNA bulks were of good quality and could be used for both RDA and AFLP analysis.
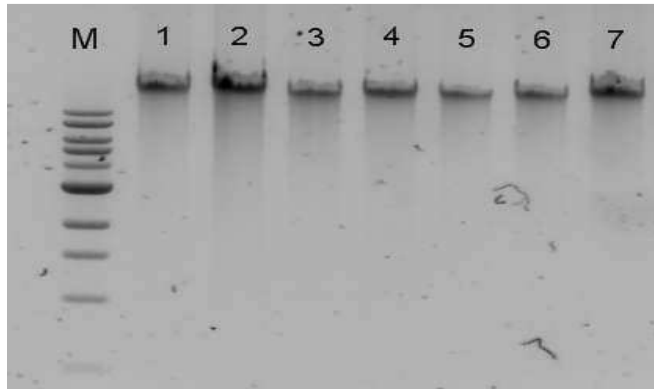
Figure 3.1: Electrophoresis profiles of DNA bulks generated. (1) Deli *dura* from the 744 controlled cross; (2) *dura* and (3) *pisifera* from the 769 controlled cross; (4) *dura* and (5) *pisifera* from the 768 controlled cross; (6) *dura* and (7) *pisifera* from the 751 controlled cross. **M**, 1 kb ladder (New England Biolabs).

3.3.2 Optimization of the RDA protocol

a) Optimization of restriction digestion

*Hin*dIII, *Bam*HI, *Eco*RI and *Mse*I seemed to digest oil palm DNA well (Figure 3.2). *Mse*I having a 4 bp recognition site cut DNA into smaller average fragments of less than 3 kb in size [Figure 3.2 (E)]. In contrast, DNA was not digested or minimally digested by the *Hpa*II and *Pst*I enzyme [Figure 3.2 (D) and (F)]; the digested products had a similar gel profile to the undigested negative control, indicating that the majority of the DNA was still in its intact form.
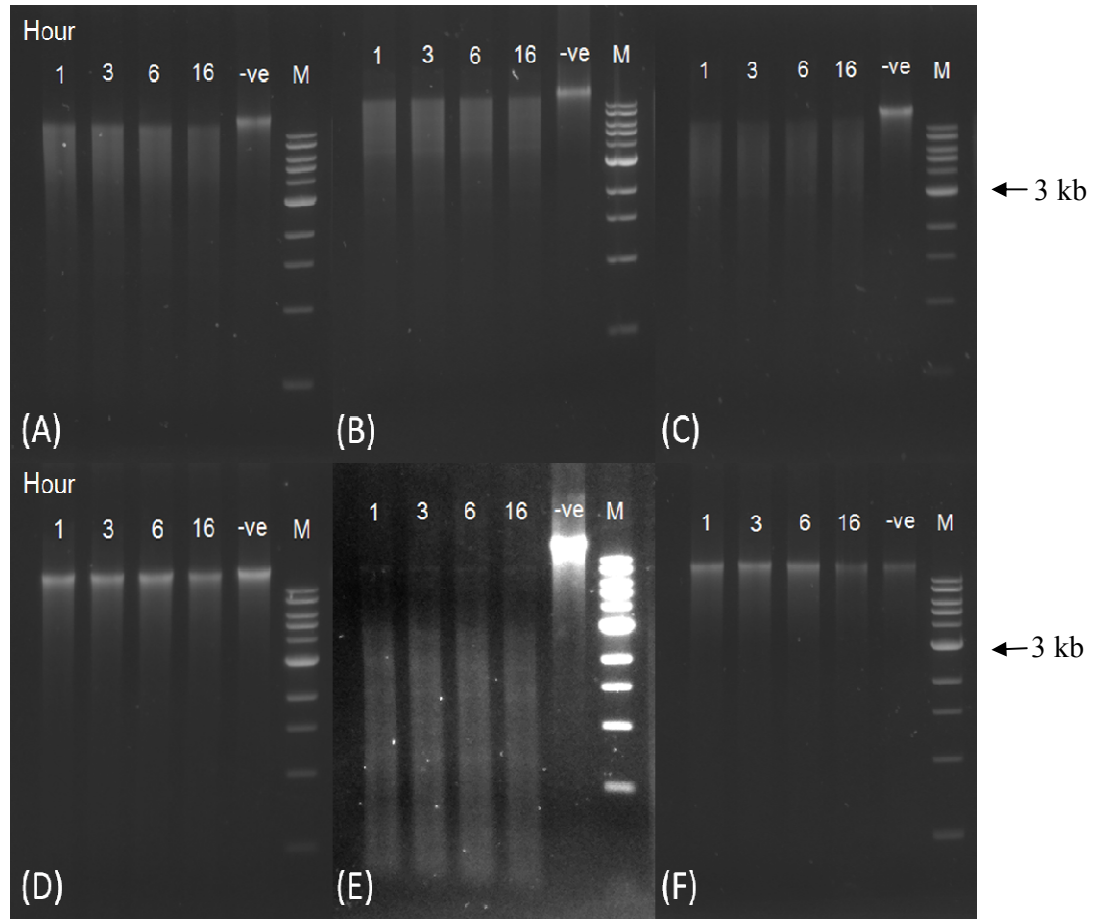
Figure 3.2: Restriction digestion profiles of six different restriction endonucleases. Pooled 744 Deli *dura* sample was digested with (A) *Bam*HI, (B) *Eco*RI, (C) *Hin*dIII, (D) *Hpa*II, (E) *Mse*I, and (F) *Pst*I for 1, 3, 6, and 16 hours, respectively. −ve, undigested negative control; **M**, 100 bp ladder (New England Biolabs).

Regardless of the incubation time, the restriction enzymes produced highly similar cleavage profiles. The present work focused on two different enzymes, *Bam*HI and *Hin*dIII, with overnight incubation of 16 h at 37 °C to ensure complete digestion.

b)   **Optimization of primer concentration in PCR**
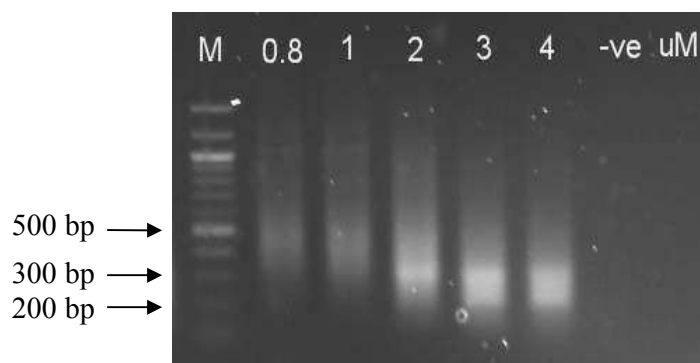
Figure 3.3: Amplification profiles of PCR using five different primer concentrations (0.8, 1, 2, 3 and 4 µM). –ve, negative control of PCR reaction; **M**, 100 bp marker (New England Biolabs).

Figure 3.3 demonstrates that DNA fragments smaller than 500 bp were preferentially enriched with increasing amounts of primer in the PCR despite the background whole genome amplification of fragments less than 1 kb. Dense DNA bands of about 300 to 200 bp started to appear when 2 µM of primer was used. This preferential enrichment of DNA fragments less than 500 bp is not desirable as all DNA fragments smaller than 1,500 bp should receive same degree of enrichment before going into subtractive hybridization in which the unique target fragments in the tester will be preferentially amplified. Therefore, a primer concentration of 1 µM was used in subsequent PCR reactions of RDA analysis.

3.3.3 First RDA analysis

a) Generation of first round *Bam*HI and *Hin*dIII amplicons (representations)

After digestion of genomic DNA, ligation of adaptor primers to DNA fragments and amplification of ligated DNA fragments, first round *Bam*HI and *Hin*dIII amplicons were generated. Gel-electrophoresis of all amplicons shows patterns of multiple bands (Figure 3.4). All six amplicons contained predominantly small restriction DNA fragments between the size of 200 to 1500 bp, representing a

subpopulation of the initial tester and driver sequences, hence the complexity of the genomic DNAs was successfully reduced.
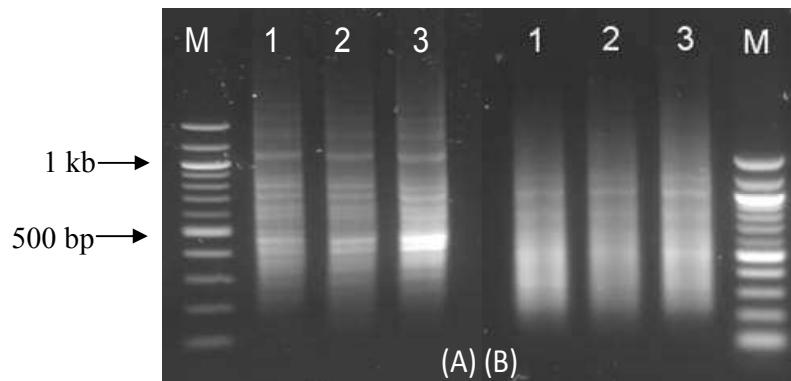
Figure 3.4: Electrophoresis profiles of the first round RDA amplicons. (A) *Bam*HI and (B) *Hin*dIII amplicons. 1, 744 Deli *dura*; 2, 769 *dura*; 3, 769 *pisifera*. **M**, 100 bp marker (New England Biolabs).

b)   Reciprocal subtractive hybridization of *Bam*HI and *Hin*dIII amplicons

Three rounds of subtractive hybridization were performed for both *Bam*HI and *Hin*dIII amplicons with increased stringency; tester: driver ratio of 1:80, 1:400 and 1:200,000 for round 1, 2 and 3, respectively. Reciprocal analysis was performed in which the 744 Deli *dura* or the 769 *dura* was used as tester and the 769 *pisifera* as driver in one experiment while in another experiment, the 769 *pisifera* was the tester with the 744 Deli *dura* or the 769 *dura* as the driver.

Figures 3.5 and 3.6 illustrate that a smear of DNA fragments from the initial representations was gradually replaced by distinct DNA bands of discrete length after three rounds of subtractive hybridization; indicating that significant enrichment of target sequences had been achieved for both *Bam*HI and *Hin*dIII amplicons. Different enrichment profiles were observed between reciprocal subtractive hybridization of the same amplicons (red arrow, Figures 3.5 and 3.6). There were no obvious differences in enrichment profiles between the 744 Deli *dura* and the 769 *dura* samples,

regardless of whether they were used as tester or driver (blue arrow, Figures 3.5 and

3.6).
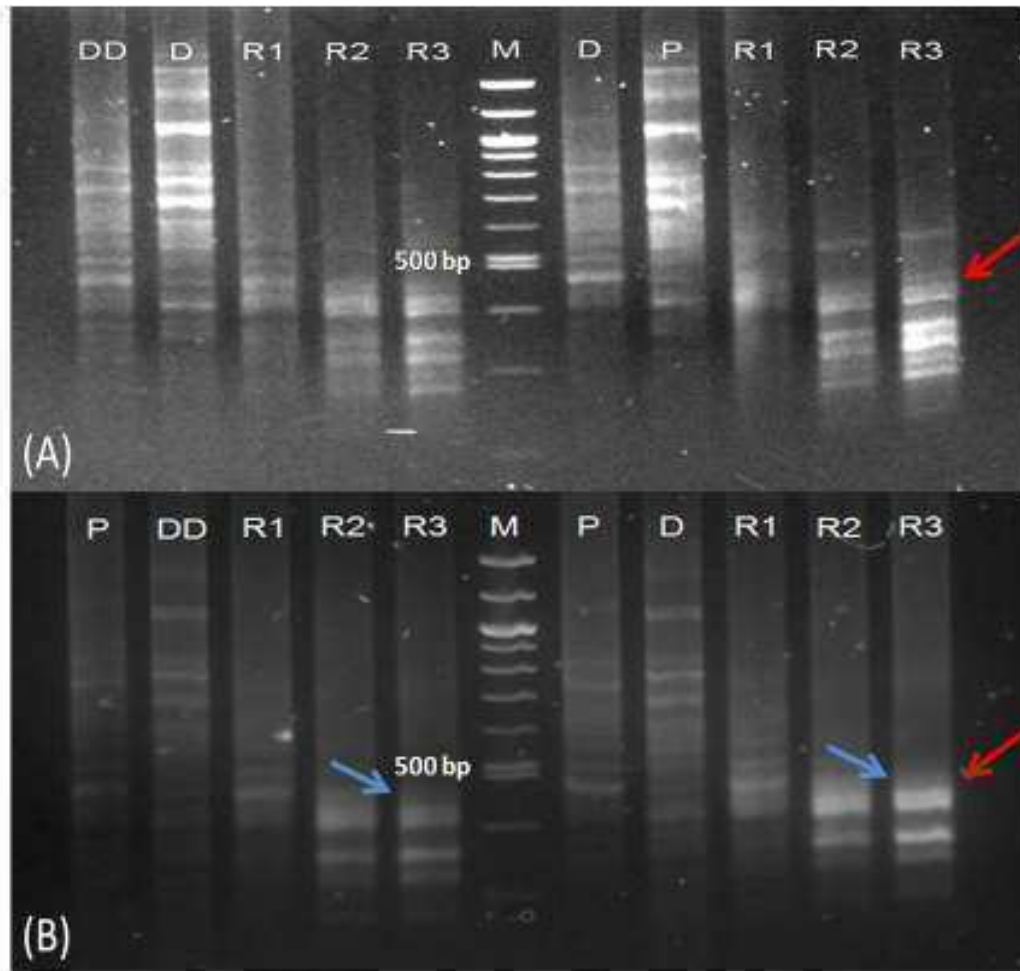
Figure 3.5: *Bam*HI enrichment profiles of three rounds of reciprocal subtractive hybridization. (A) The 744 Deli *dura* and 769 *dura* were used as tester to hybridize against driver 769 *pisifera* and (B) *vice versa* with three rounds of subtractive hybridization at tester: driver ratio of 1:80 for R1, 1:400 for R2 and 1:200,000 for R3, respectively. DD, 744 Deli *dura*; D, 769 dura; P, 769 *pisifera*. **M**, 100 bp ladder (New England Biolabs). Red arrows indicate different enrichment profile between reciprocal analyses while blue arrows indicate no difference in enrichment profile between 744 Deli *dura* and 769 *dura*.
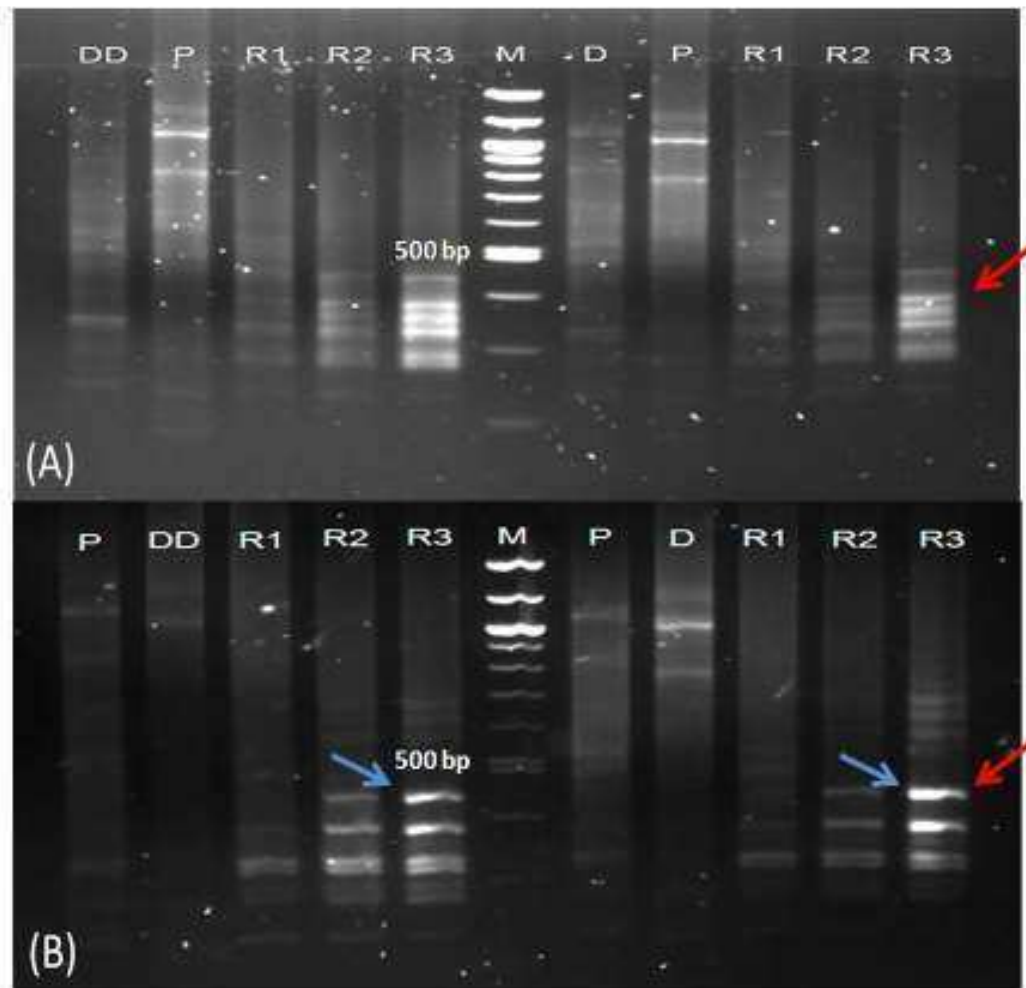
Figure 3.6: *Hin*dIII enrichment profiles of three rounds of reciprocal subtractive hybridization. (A) The 744 Deli *dura* and 769 *dura* were used as tester to hybridize against driver 769 *pisifera* and (B) *vice versa* with three rounds of subtractive hybridization at tester: driver ratio of 1:80 for R1, 1:400 for R2 and 1:200,000 for R3, respectively. DD, 744 Deli *dura*; D, 769 dura; P, 769 *pisifera*. **M**, 100 bp ladder (New England Biolabs). Red arrows indicate different enrichment profile between reciprocal analyses while blue arrows indicate no difference in enrichment profile between 744 Deli *dura* and 769 *dura*.

c)   **Sanger sequencing and characterization of difference products**

        Positive clones from round 3 difference products were obtained through colony screening of bacterial transformants. All clones were denoted as XYZ with X for tester, Y for driver and Z for restriction endonuclease used followed by a number. The 744 Deli *dura* bulk was represented by DD, 769 *dura*

as D, 769 *pisifera* as P while *Bam*HI and *Hin*dIII enzymes were simplified as B

and H, respectively. Thus, when the 744 Deli *dura* was used as tester to hybridize

against the driver 769 *pisifera* of *Bam*HI amplicons, the 15 clones being screened

were named as DDPB-1 to DDPB-15, respectively.

Screening of the plasmids by PCR amplification showed that the majority of

plasmids contained inserts of different sizes ranging from 200 to 450 bp,

corresponding well with the respective difference products sizes targeted. Only

inserts of similar size to the difference products were sent for sequence analysis.

Multiple sequence alignment using ClustalW revealed that each category of

difference products had at least two clones that were identical (data not shown). In

addition, many clones within the 744 Deli *dura* and the 769 *dura* analysis had

identical sequences, regardless of whether the 744 Deli *dura* or 769 *dura* were

used as tester or driver. Such identical clones were noticed in both *Bam*HI and

*Hin*dIII representations (Table 3.8). It was noticed that 5 identical clones of PDH

(PDH-2, -4, -6, -13 and -14) were actually the same clone as the 4 identical clones

of PDDH (PDDH-3, -4, -8 and -11) with an alignment score of at least 96%. Besides

the PDDH-3 clone family, clone PDDH-5, -6 and -9 were also found to have an

alignment score of more than 96% and PDH-15 also fell into this family. It was

not expected that identical clones would be located within reciprocal analysis of

the same representations (Table 3.9), suggesting that the subtraction of common

sequences did not go to completion.

Table 3.8: Identical clones within the 744 Deli *dura* and the 769 *dura* analysis.

Restriction enzymes	Tester	Driver	Clones ID	Length (bp)	Alignment Score	Sequence alignment
***Bam*HI**	Deli *dura*	*Pisifera*	DDPB-3	283	≥88	Appendix B1
	Deli *dura*	*Pisifera*	DDPB-6	283		
	Dura	*Pisifera*	DPB-15	281		
***Bam*HI**	Deli *dura*	*Pisifera*	DDPB-9	356	87	Appendix B2
	Dura	*Pisifera*	DPB-5	356		
***Bam*HI**	Deli *dura*	*Pisifera*	DDPB-11	289	90	Appendix B3
	Dura	*Pisifera*	DPB-11	289		
***Bam*HI**	*Pisifera*	Deli *dura*	PDDB-1	359	96	Appendix B4
	Pisifera	*Dura*	PDB-5	359		
***Bam*HI**	*Pisifera*	Deli *dura*	PDDB-7	420	97	Appendix B5
	Pisifera	*Dura*	PDB-10	420		
***Bam*HI**	*Pisifera*	Deli *dura*	PDDB-13	331	≥95	Appendix B6
	Pisifera	*Dura*	PDB-4	330		
	Pisifera	*Dura*	PDB-12	330		
***Hind*III**	Deli *dura*	*Pisifera*	DDPH-8	337	89	Appendix B7
	Dura	*Pisifera*	DPH-8	337		
***Hind*III**	Deli *dura*	*Pisifera*	DDPH-10	395	94	Appendix B8
	Dura	*Pisifera*	DPH-15	394		
***Hind*III**	*Pisifera*	Deli *dura*	PDDH-1	385	96	Appendix B9
	Pisifera	*Dura*	PDH-8	385		
***Hind*III**	*Pisifera*	Deli *dura*	PDDH-3	448	≥97	Appendix B10
	Pisifera	Deli *dura*	PDDH-4	448		
	Pisifera	Deli *dura*	PDDH-8	448		
	Pisifera	Deli *dura*	PDDH-11	448		
	Pisifera	*Dura*	PDH-2	448		
	Pisifera	*Dura*	PDH-4	448		
	Pisifera	*Dura*	PDH-6	448		
	Pisifera	*Dura*	PDH-13	448		
	Pisifera	*Dura*	PDH-14	448		
***Hind*III**	*Pisifera*	Deli *dura*	PDDH-5	323	≥97	Appendix B11
	Pisifera	Deli *dura*	PDDH-6	323		
	Pisifera	Deli *dura*	PDDH-9	323		
	Pisifera	*Dura*	PDH-15	323		

Table 3.9: Identical clones in reciprocal analysis of *Bam*HI and *Hind*III amplicons.

Restriction enzymes	Tester	Driver	Clones ID	Length (bp)	Alignment Score	Sequence alignment
***Bam*HI**	Deli *dura*	*Pisifera*	DDPB-1	330	95	Appendix C1
	Pisifera	Deli *dura*	PDDB-13	331		
	Pisifera	*Dura*	PDB-4	330		
	Pisifera	*Dura*	PDB-12	330		
***Bam*HI**	*Dura*	*Pisifera*	DPB-6	456	94	Appendix C2
	Pisifera	*Dura*	PDB-6	456		
***Bam*HI**	Deli *dura*	*Pisifera*	DDPB-9	355	≥87	Appendix C3
	Dura	*Pisifera*	DPB-5	356		
	Pisifera	Deli *dura*	PDDB-6	356		
	Pisifera	Deli *dura*	PDDB-9	356		
***Hind*III**	Deli *dura*	*Pisifera*	DDPH-2	280	≥84	Appendix C4
	Deli *dura*	*Pisifera*	DDPH-13	281		
	Pisifera	Deli *dura*	PDDH-14	282		
	Pisifera	Deli *dura*	PDDH-15	281		
***Hind*III**	*Dura*	*Pisifera*	DPH-7	323	≥97	Appendix C5
	Pisifera	Deli *dura*	PDDH-5	323		
	Pisifera	Deli *dura*	PDDH-6	323		
	Pisifera	Deli *dura*	PDDH-9	323		
	Pisifera	*Dura*	PDH-15	323		
***Hind*III**	*Dura*	*Pisifera*	DPH-11	385	≥94	Appendix C6
	Dura	*Pisifera*	DPH-12	385		
	Pisifera	Deli *dura*	PDDH-1	385		
	Pisifera	*Dura*	PDH-8	385		

Homology searches against the GenBank database revealed that PDDH-3 clone family with length of 448 bp had significant homology with a mitochondria DNA sequence from plant species (Table 3.10). The E-value of the BlastN search was as high as 0.0 followed by 5e-177. Meanwhile, PDDH-5 family (323 bp) was also found to be homologous to the chloroplast gene of *Elaeis oleifera*, coding for 23S ribosomal RNA (rrn23), with an E- value of 4e-125 (Figure 3.7). The discovery of mitochondria and chloroplast DNA rather than nuclear DNA for both clone families was rather unexpected.

Table 3.10: Homology search of PDDH-3 family using GenBank database.

Accession number	Description	Query coverage	E-value	Maximum identities
EU365401.1	*Bambusa oldhamii* mitochondrion,complete genome	100 %	0.0	95 %
FM179380.1	*Vitis vinifera* complete mitochondrial genome, cultivar Pinot noir clone ENTAV115	100 %	5e-177	96 %
AP011077.1	*Oryza sativa* Indica Group mitochondrial DNA, complete genome, cultivar: Lead rice	100 %	7e-176	94 %
AP011076.1	*Oryza rufipogon* mitochondrial DNA, complete genome	100 %	7e-176	94 %
BA000029.3	*Oryza sativa* Japonica Group mitochondrial DNA, complete genome	100 %	7e-176	94 %
EU431224.1	*Carica papaya* mitochondrion, complete genome	100 %	3e-173	94 %
BA000042.1	*Nicotiana tabacum* mitochondrial DNA, complete genome	99 %	4e-166	92 %
Y08501.2	*Arabidopsis thaliana* mitochondrial genome	100 %	6e-145	88 %
AP006444.1	*Brassica napus* mitochondrial DNA, complete genome	100 %	6e-145	88 %
AP009381.1	*Cycas taitungensis* mitochondrial DNA, complete genome	82 %	1e-122	89

>gb|EU016908.1| Elaeis oleifera 23S ribosomal RNA (rrn23) gene, partial sequence; chloroplast
Length=2810

Score = 455 bits (504), Expect = 4e-125
Identities = 260/265 (98%), Gaps = 0/265 (0%)
Strand=Plus/Plus

```
Query   1    GGGTGAAACTAAGTGGAGGTCCGAATCGACCGATGTTGAAGAATCAGCGGATGAGTTGTG   60
             |||||||||||||||||||||||| ||||  ||||||||||||||||||||||||||||||
Sbjct   722  GGGTGAAACTAAGTGGAGGTCCGAACCGACTGATGTTGAAGAATCAGCGGATGAGTTGTG   731

Query   61   GTTAGGGGTGAAATGCCACTCGAACCCAGAGCTAGCTGGTTCTCCCCGAAATGCGTTGAG   120
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct   782  GTTAGGGGTGAAATGCCACTCGAACCCAGAGCTAGCTGGTTCTCCCCGAAATGCGTTGAG   841

Query   121  GTGCAGCAGTTGACTGGACATCTAGGGGTAAAGCACTGTTTCGGTGCGGGCTGCGAGAAC   180
             | |||||||||||||||||||||||||||||||||||||||||||||||| ||||||| |
Sbjct   842  GCGCAGCAGTTGACTGGACATCTAGGGGTAAAGCACTGTTTCGGTGCGGGCCGCGAGAGC   931

Query   181  GGTACCAAATCGAGGCAAACTCTGAATACTAGATATGACCCCAAAATAACAGGGGTCAAG   240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct   902  GGTACCAAATCGAGGCAAACTCTGAATACTAGATATGACCCCAAAATAACAGGGGTCAAG   961

Query   241  GTCGGCCAGTGAGACGATGGGGGAT   265
             |||||||||||||||||||||||||
Sbjct   962  GTCGGCCAGTGAGACGATGGGGGAT   936
```

Figure 3.7: BlastN search of PDDH-5 family using GenBank database. E- value and identities of the search are shown.

3.3.4   Second RDA analysis

Fingerprinting results indicated that each controlled cross had one outcross sample that should be excluded from bulk construction. Consequently, RDA analysis was repeated using pooled legitimate *dura* and *pisifera* samples of the 769, 768 and 751 controlled crosses.

a)   <u>Generation of *Bam*HI and *Hin*dIII representations</u>

Genomic digestion of pooled DNA using *Bam*HI or *Hin*dIII restriction endonucleases displayed smearing of DNA as compared to the single bands of intact

genomic DNA (Figure 3.8). This indicates that both *Bam*HI and *Hin*dIII enzymes cut

the pooled samples efficiently.
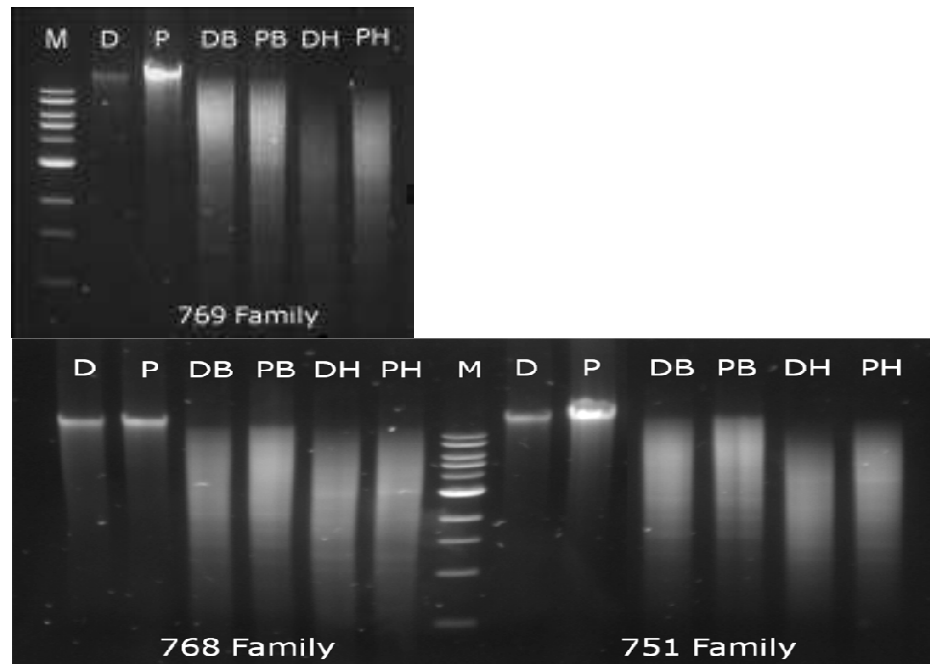
Figure 3.8: Restriction digestion profiles of DNA bulks from the 769, 768 and 751

controlled crosses using *Bam*HI and *Hin*dIII restriction endonuclease. D, *dura*; P,

pisifera; DB, *Bam*HI-digested *dura*; PB, *Bam*HI-digested *pisifera*; DH, *Hin*dIII-digested *dura*;

PH, *Hin*dIII-digested *pisifera*. **M**, 1 kb ladder (New England Biolabs).

Figure 3.9 displays the generation of representations through amplification of

R primers, removal of R adaptors and amplification of J primers for tester,

sequentially. The amplification and cleavage pattern were the same for both *dura* and

pisifera pools of the same controlled cross in which the same repeat bands were

consistently shown. This is expected as there had been no subtraction performed and

the visible repeat bands were likely to be those repetitive DNA that exist in both the

dura and *pisifera* pools. There was also a reduction of 48 bp in size for digested PCR

products (Figure 3.9, lanes 3 and 4), corresponding well with the removal of 24-mer R

adaptors from both ends of the DNA fragments. Ligation and amplification of tester

using J primers (Figure 3.9, lanes 5 and 6) showed an increase in size of 48 bp, giving

a similar pattern as the earlier R primer amplification. Overall, the pre-enrichment steps worked well to proceed for subtractive hybridization.
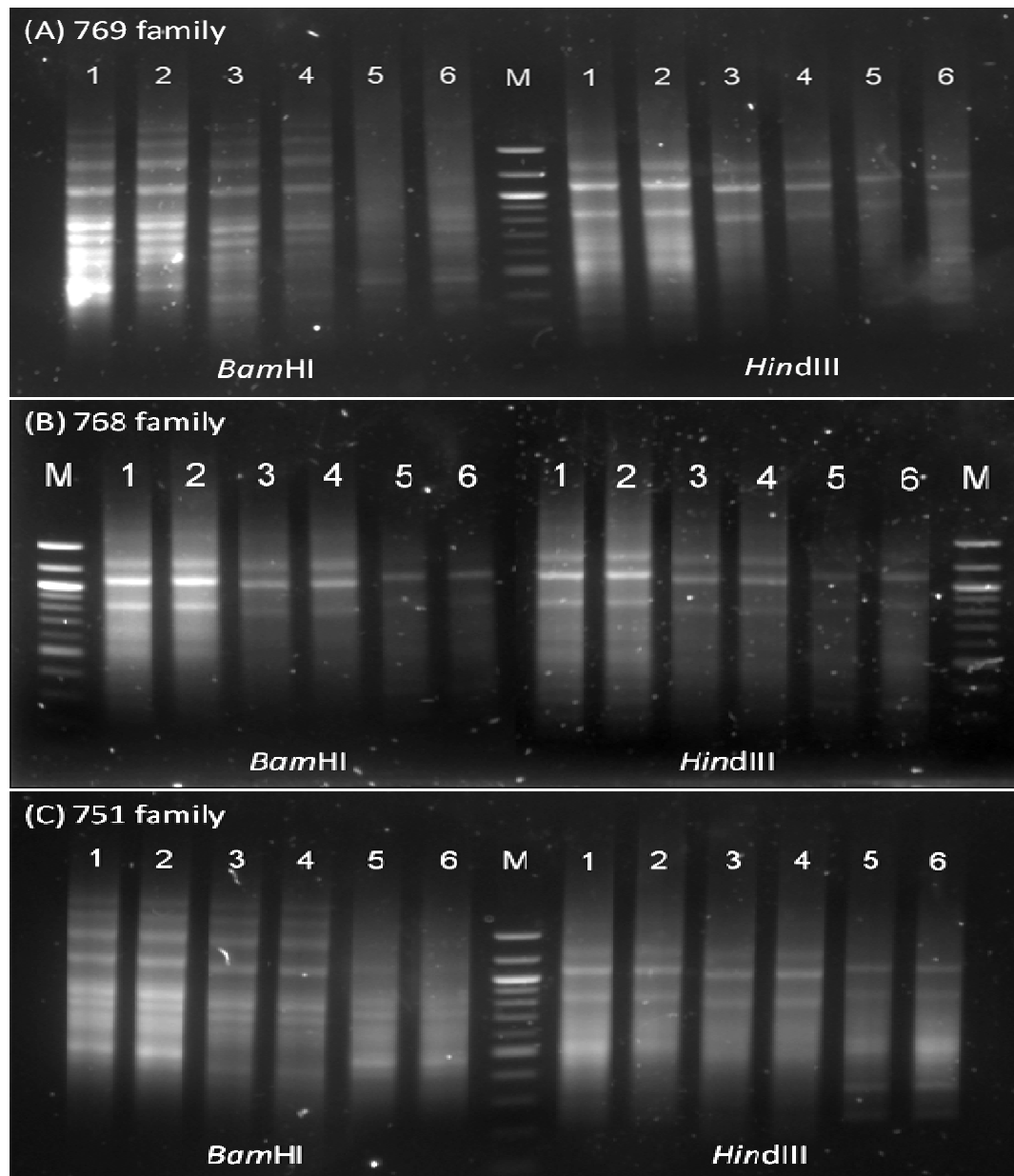
Figure 3.9: Amplification and digestion profiles of *Bam*HI and *Hin*dIII amplicons for pooled *dura* and *pisifera* of the 769, 768 and 751 controlled crosses. R24-primer amplification of (1) *dura* and (2) *pisifera*; removal of R adaptors from (3) *dura* and (4) pisifera; J24-primer amplification of (5) *dura* and (6) *pisifera*. **M**, 100 bp ladder (New England Biolabs).

b)   <u>**Reciprocal subtractive hybridization of *Bam*HI and *Hin*dIII amplicons**</u>

After three rounds of subtractive hybridization with increased stringency, progressive enrichment of tester was observed in Figure 3.10; the smearing of DNA fragments from round 1 difference products was slowly replaced by discrete DNA bands of round 3 difference products. However, unexpected highly similar enrichment profiles were obtained for reciprocal analyses. The banding pattern of RDA analysis of *Bam*HI representations was the same for all the three controlled crosses, but different from those of RDA analysis of *Hin*dIII amplicons. In the previous RDA analysis using the 769 controlled cross with D36 outcross included, different enrichment profiles between reciprocal analyses were observed (Figures 3.5 and 3.6; Figure 3.11, lanes 9 and 10). RDA study on the 744 Deli *dura* against the 769 *pisifera* was also found to be enriched differently between reciprocal (Figures 3.5 and 3.6; Figure 3.11, lanes 7 and 8).
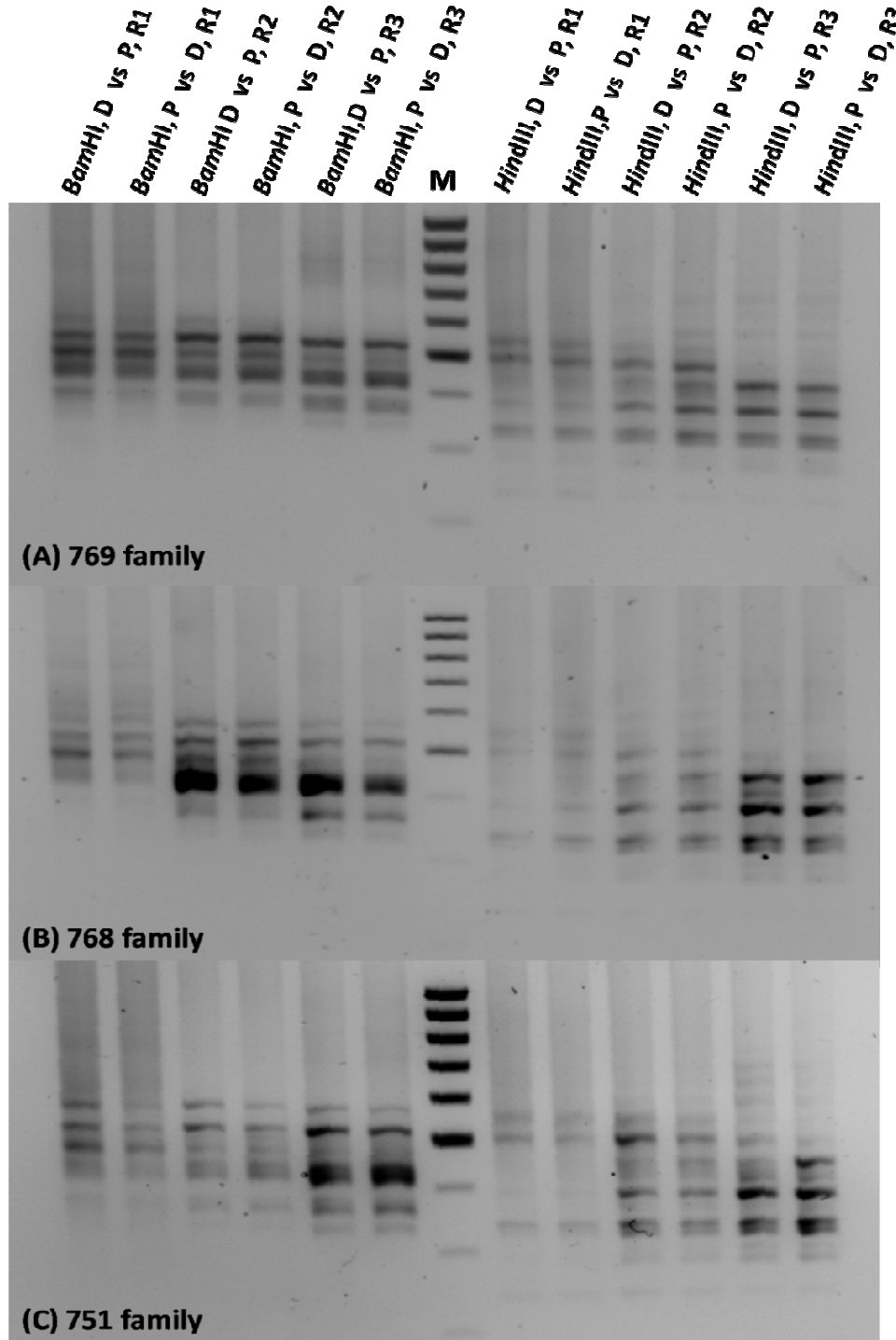
Figure 3.10: Reciprocal subtractive hybridization profiles of *Bam*HI and *Hin*dIII representations of the 769, 768 and 751 controlled crosses. Reciprocal analysis was compared side by side. D, *dura*; P, *pisifera*; R1, round 1 subtractive hybridization at tester:driver ratio of 1:80; R2, round 2 at tester:driver ratio of 1:400; R3, round 3 at tester:driver ratio of 1:200,000. **M**, 100 bp ladder (New England Biolabs).
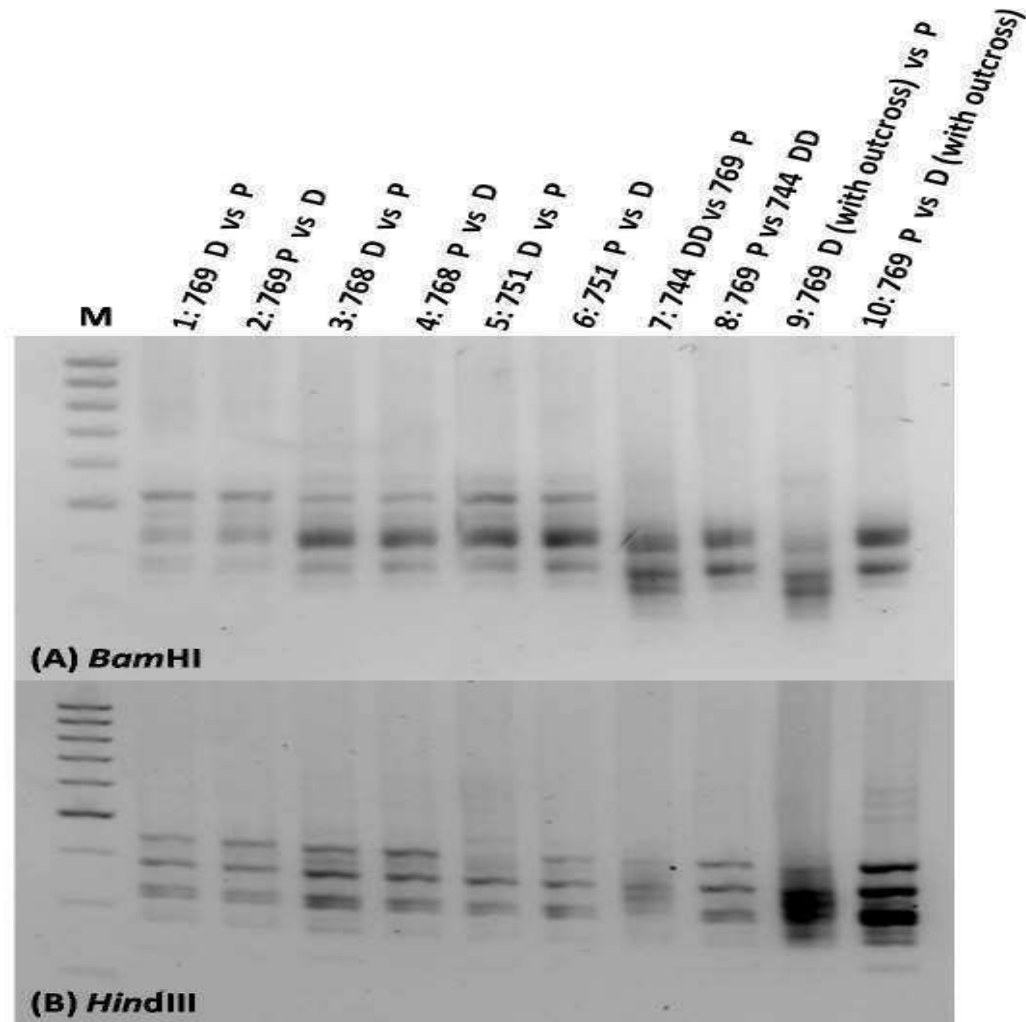
Figure 3.11: Enrichment profiles of round 3 difference products from both first and second reciprocal RDA analyses. Reciprocal analysis was compared side by side. D, *dura*; P, *pisifera*; DD, Deli *dura*. **M,** 100 bp ladder (New England Biolabs).

3.3.5 Assessment of the RDA technique with positive control

In view of the highly similar enrichment profiles in reciprocal analyses, the effectiveness of the RDA technique was tested with the 125 bp fragment from *Hin*dIII-digested *Lambda* DNA added into the tester as positive control spike (red arrow, Figure 3.13). After three rounds of reciprocal subtractive hybridization analysis of the legitimate 769 pooled samples, it was demonstrated that highly similar

enrichment profiles were once again attained and 125 bp band was not seen in the difference products even at a starting level of 1,000 copies (Figure 3.12).
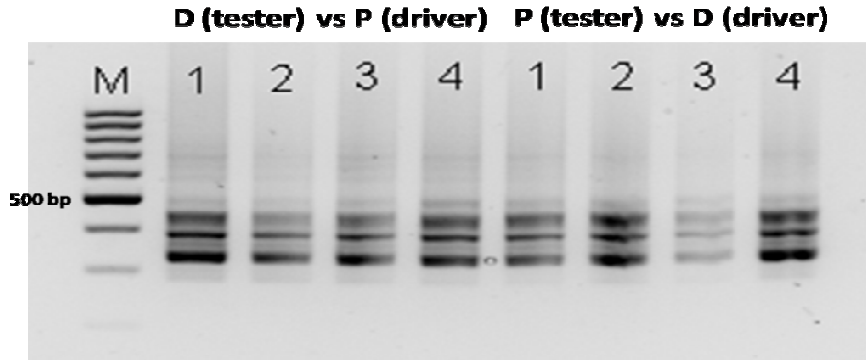
Figure 3.12: Round 3 enrichment profiles of the legitimate 769 pooled samples with *Lambda* DNA added as positive control. A 125 bp fragment excised from *Hin*dIII-digestion of *Lambda* DNA was added at molecular level of (1) 1, (2) 10, (3) 100, and (4) 1000 copies, respectively, into the tester. D, *dura*; P, *pisifera*. **M**, 100 bp ladder (New England Biolabs).

However, when round 3 difference products were amplified with the specific Lambda125 primer pair, faint bands at size of 125 bp were observed [Figure 3.13 (A)] suggesting that enrichment was occurring, but inefficiently. Amplification of a tester that contains one copy number of 125 bp fragment also showed a very faint band and the band intensity increased with increasing copy number of positive control [Red arrow, Figure 3.13 (B)]. This implies that positive control was successfully being picked up during subtractive hybridization but existed in relatively low quantities.
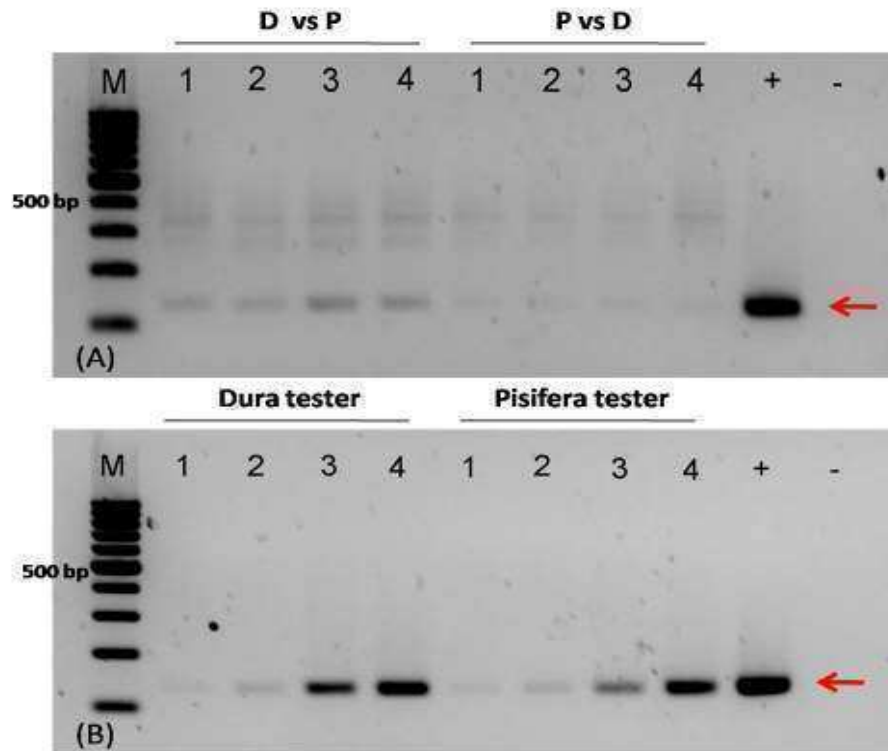
Figure 3.13: Amplification profiles of the positive control 125 bp fragment using Lambda125 primer pair. (A) Round 3 difference products of reciprocal analysis of the legitimate 769 pooled samples and (B) tester DNA were subjected to Lambda125 primer pair amplification. Positive control spike was added into the tester in molecular level of (1) 1, (2) 10, (3) 100, and (4) 1000 copies, respectively. The excised 125 bp fragment was the positive control while the previous round 3 difference products of the 768 controlled cross was used as negative control. Arrow indicates the 125 bp fragment. D, *dura*; P, *pisifera*. **M**, 100 bp ladder (New England Biolabs). Arrow indicates the amplification of positive control.

3.3.6   454 pyrosequencing of round 2 and 3 difference products

a)   <u>Tagging of RDA pools for 454 pyrosequencing</u>

Round 2 and 3 difference products of the reciprocal 744 Deli *dura* against 769 *pisifera* and reciprocal analysis of *dura* against *pisifera* of the 769 controlled cross with and without the D36 outcross were successfully amplified by N and J 24 primer with single nucleotide modifications to allow labelling of products within the 454 sequencing pools (Figure 3.14). This allows each difference product to be identified

after obtaining the combined sequencing data. Figure 3.15 shows the pooled difference products with a majority of bands less than 500 bp, corresponded well with the read length of 454 sequencing.
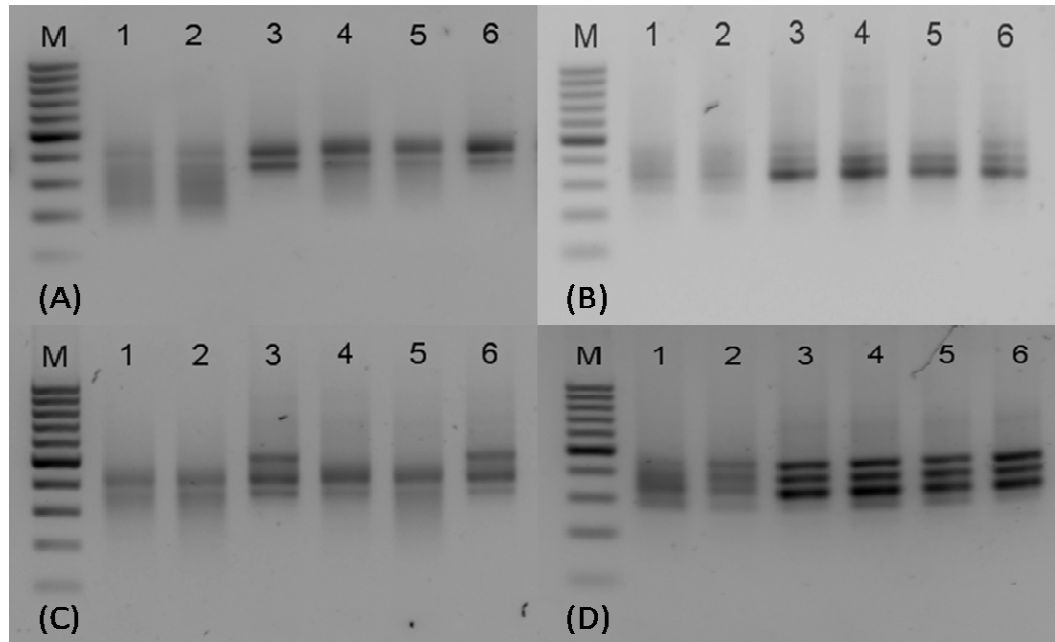
Figure 3.14: Amplification profiles of round 2 and 3 difference products using N or J primers with single nucleotide modifications. (A) NBam24 A-F primer; (B) NHind24 A-F primer; (C) JBam24 A-F primer; and (D) JHind24 A-F primer. 1, 744 DD vs 769 P; 2, 769 D (with outcross) vs P; 3, legitimate 769 D vs P; 4, 769 P vs 744 DD; 5, 769 P vs D (with outcross); 6, legitimate 769 P vs D; tester vs driver. **M**, 100 bp ladder (New England Biolabs).
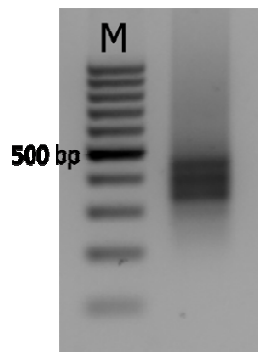
Figure 3.15: Electrophoresis profile of pooled round 2 and 3 difference products ready for 454 pyrosequencing. M, 100 bp ladder (New England Biolabs).

b) ***De novo* assembly of 454 pyrosequencing data of RDA pools**

Sequences generated from 454 pyrosequencing were clustered into their original RDA pools according to the respective adaptor sequences. A total of 37,239 sequences were generated for all the 24 RDA pools and from there, 1,103 contigs were assembled, as summarised in Table 3.11. In general, RDA analysis using the *Hin*dIII restriction enzyme generated fewer contigs compared to RDA analysis with the *Bam*HI enzyme. Significant reduction of contigs were also observed in the RDA analysis using fully legitimate samples (pools C and F) compared to samples pools containing an outlier (pools B and D), particularly those of round 3 analysis.

Table 3.11: Number of sequences and assembled contigs for each RDA pool.

Restriction Enzyme	RDA round	Tester	Driver	Name of RDA pool	No. of reads	No. of contigs assembled
*Bam*HI	R2	744 Deli *Dura*	769 *Pisifera*	NBamA	675	58
		769 *Dura* (outcross D36)	769 P*isifera*	NBamB	714	49
		769 *Dura* (x outcross)	769 *Pisifera*	NBamC	1148	29
		769 *Pisifera*	744 Deli *Dura*	NBamD	1103	60
		769 P*isifera*	769 *Dura* (outcross D36)	NBamE	739	62
		769 *Pisifera*	769 *Dura* (x outcross)	NBamF	1303	54
	R3	744 Deli *Dura*	769 *Pisifera*	JBamA	990	79
		769 *Dura* (outcross D36)	769 P*isifera*	JBamB	1700	90
		769 *Dura* (x outcross)	769 *Pisifera*	JBamC	1384	23
		769 *Pisifera*	744 Deli *Dura*	JBamD	2007	81
		769 P*isifera*	769 *Dura* (outcross D36)	JBamE	3233	102
		769 P*isifera*	769 *Dura* (x outcross)	JBamF	1871	27
*Hin*dIII	R2	744 Deli *Dura*	769 *Pisifera*	NHindA	801	37
		769 *Dura* (outcross D36)	769 P*isifera*	NHindB	736	35
		769 *Dura* (x outcross)	769 *Pisifera*	NHindC	1432	16
		769 *Pisifera*	744 Deli *Dura*	NHindD	1711	16
		769 *Pisifera*	769 *Dura* (outcross D36)	NHindE	1497	31
		769 *Pisifera*	769 *Dura* (x outcross)	NHindF	1792	30
	R3	744 Deli *Dura*	769 *Pisifera*	JHindA	1124	52
		769 *Dura* (outcross D36)	769 *Pisifera*	JHindB	979	47
		769 *Dura* (x outcross)	769 *Pisifera*	JHindC	1843	26
		769 *Pisifera*	744 Deli *Dura*	JHindD	2258	22
		769 P*isifera*	769 *Dura* (outcross D36)	JHindE	3071	52
		769 *Pisifera*	769 *Dura* (x outcross)	JHindF	3128	23

c)   Homology search within the RDA pools

Due to time constraints, further homology search analysis was focused on contigs obtained from round 3 RDA but not round 2. Contigs were categorized into three different classes according to their homology search result (Table 3.12). Contigs that were present only in their own pool were unique contigs. Contigs that can be found in at least two of the *dura* tester pool but not *pisifera* pool, or *vice versa*, were *dura*/*pisifera*-specific contigs whereas universal contigs were contigs that were present in both *dura* and *pisifera* RDA pools.

A majority of contigs, about 70%, were universal contigs that could be identified in both direction of reciprocal RDA analysis, followed by 22% of unique contigs.   Only a small portion of contigs was regarded as *dura*/*pisifera*-specific contigs that could be of particular interest. These 17 *dura*/*pisifera*-specific contigs could be potentially shell-thickness related-markers that are worth further investigation. A total of 15 *dura* and *pisifera*-specific contigs were generated from the *Bam*HI analysis while the *Hin*dIII analysis had only produced two *dura*-related contigs but not *pisifera*-related contigs.

Table 3.12: Classification of round 3 RDA contigs.

JBam	A	B	C	D	E	F
Unique	25	29	0	7	22	1
universal	48	54	22	56	65	21
Dura/*Pisifera*-specific	6	7	1	18	15	5
Total	79	90	23	81	102	27

JHind	A	B	C	D	E	F
Unique	14	19	3	0	14	2
Universal	36	27	22	22	38	21
Dura/*Pisifera*-specific	2	1	1	0	0	0
Total	52	47	26	22	52	23

The list of putative contigs is presented in Table 3.13. Each group of contigs contains different numbers of transcripts with variable lengths; hence the longest contigs representative of each group were used for further analysis. The restriction recognition sites [*Bam*HI (GGATCC) and *Hin*dIII (AAGCTT)] could be recovered and highlighted in some of the contigs.

Table 3.13: Putative shell-thickness related-contigs and their sequences.

Category	Restriction Enzyme	Representative	Members of contigs	Sequence
Dura-specific	*Bam*HI	JBam_B17	JBam_A15	GGTGGTGGTGGAGGTGGAGGTGGAG GGGGTTCCGGATATGGGAGTGGTGG CGGGAGTGGCTCTGGTTATGGATCGG GATATGGTGATGGTTCAGGCTACGGC AGCGGTACGGGTGGAGGACATGGTG AAGGAGGAGGAGGAGGTGGCGGCGG CGGTGGCGGTGGGAGTTACGGCGGT GGGGGGAGTTACGGCGGAGGTGGCG GCGGCGGCGG
		JBam_A19	JBam_B57, JBam_B77	**GGATCC**CACCATCATCATCCCCGCAT TAAAGAAGCCCTGATGCCCTTCTCCT CCCACCATCCCCCCCTGAGTAAAACG ACCCCGACTACCACCCCCTTCCAGAA ACGCGCACGACGGCGACTGCACCGC CGCCTCCAACCCCACGAGCCCGCCCG GCATCGCCGCCAGGCACTCCTCCTCG TACGCCGGCGCCTGCTGCTGCGCCGA CGGGCCCCCTATCTGGATCGGCACCA TGGACTCGGGCGAGTAGGAGCCGTA CCCGCCAATTGGGATC**GGATCC**
		JBam_B90	JBam_A26	TCCTACGAGGCCGTGCTCGATGACCC GGCCGTCGACGCCGTCTACGTGCCGC TTCCCACGGGCCTACACGTCCGCTGG GCGGTGGCCGCGGCGGAGCACGGGA AGCACGTGCTCCTCGAGAAGCCCAC GGCGCTGTGCGCCGC**GGATCC**
		JBam_B67	JBam_A32, JBam_A31	**GGATCC**AACGGTTGATCAACACGAA CTCGAGATTCTCCACTCCATTGCTCC TCGTGCAACTATAAATAACCACCTCC CTCCTAAGTCCTAGGGCGCCACAAAT TCCTCCCAAATCCGGGGGAGAGAGA TTAAAAAAACGGCGAAGAAGAACGA AGGCGGACGTCGATGCAGTCGATGG ACGTGGAGAAGATCCCCGCCGGCGG CGTGGAGAAGATCCCCGGCCGCGGC GTGGAGGATGAGGAGGACTCGCCGA TCGAGCAGGTGCGGCTGACGGCGTT GACGACGGACGACCCGACGCTCCCG GTGTGGACGTTCCGGATGTGGTTCAT CGGGGTCCGTTCATGGATTGTCGACG TCGGCTCCCTCGGATAGCAGAGTTGC CT
		JBam_B23	JBam_A38, JBam_C21	**GGATCC**TATTGGTGATCCGGGGTAG CAATTCTGCTCCTGGATGACGAGCTT TAGCCCCGCGAGAACGGCGGCGAAG CACACCGGGGCGGAGACGTAGGCAC CGGTTCCGACGACGAGGTGGGGCGG AGGGGCGGAGGAGATTCCAGCTGGC AAAGATGGAGCGGAGGAGGGCGAGG GGTAGGAGGAGGTTCTGTGGGGAGA GGAAGGGACGGACCAGGCGGGACTT

				GGGGACCGGGAGGAACTCGTAGCCG GCGGCGGGGACGAGGTCGTGCTCCA TTCCAGTGCCGGTACCGAGGAAGAC AATGCGGGTTCCCGGGCAGGCGGCT C**GGATCC**
	*Hin*dIII	JHind_A23	JHind_C22	AGCTTTGGAAGTAGGGATTTAAGCG AAGAATAGCCATAGTCTGATAGCGG AGTTAAGTGGAGAAGTAACTAGGAG AATCAAGACAAGCAATACAGAAAGG AAGAGATTAGGGGTTTATTTAGCTAA GCACTGGATGCTCCATTTCTTTGGCG TTCATGAGACCAACAAGGGGAAAAG GGTGAAGTTAAAGTCCAGTTCAGGTA TCCTTCCACCGAAGGAACATTCCTCT GGCTACCAAGTACAAGCTATCACTTG AGCCCTCTCCTTGCG**AAGCTT**
		JHind_A50	JHind_B26	AGCTTGGCTATGTGGCTAGACTGCTG CGTGCTTATTGGTATGTCCGAGGAAA GGACTCGAGAGGGCTGGCCTTGGTGT GTGGCATGCTGGGGCGCCGCACGGG TGTTTGGGTAAGAAAACAATCATCCC GTGACAATATGTCGCCGCCGCCGCCA AAGTGGAGGGGCCGTTGTTACTGGG GCCCGCCCATTTGTAACAACAAGCTC ACCGGAGCCATGGCTTTGAGGCAGC GTCGGTACCCTAGCCCAAAAGACAC AGATGGCCATGGGACTCACACGGCT AGCACTGCCGTGGG**AAGCTT**
Pisifera-specific	*Bam*HI	JBam_D29	JBam_F2	AGGTAGCGGCTCTGGCTCTGGGTCAG GAAGCGGGTACGGGTCTGGCTCTGG ATATGGCACCGGTGGAGCGCATGCT GGAGGCTATGGAAGCGGTGGAGGTG GCGG
		JBam_D24	JBam_E36	TCCAACGAATAGGAGGCTCGAGGAG GCGGCCGAGGGGAGGGCAGACCGGG GCGTCGAAGAGACAGCCGGAGACTC AGGGAAACCCAATCTAGCGACGGGG AAGAGGGGACCGGTGGCAGCTGGGG AGAGATGGGCCGAGAGCTCAGGGAG GTGGTCAGGGCTCGGGGAGGTAGGA GGTGGCCAAGGGGCTCAGTGTCCGA TGGAATGGGCTCGGGAAGCGAGATC CGATGGTCAGTGGCTCAGGGAGGCG GCTGGTGGCTCGAGGCAGTG**GGATC C**
		JBam_D68	JBam_E62, JBam_E61	TCCGGTCATAGCCCTCGCGGCTCCGA CGGTCCAAGTGGAGGAGAGGCCGAC AGAGGAGGTGGCCGAAGGAATATCG GCGGCTTCGTCGGTGCGAGTGGAGCC GGATGACGTTCGGGAAACCGAACAT CATCCGGCGGCGTCCATTGGCGCAAA GGGGGGCGTCGGGTCGAACTCCAGC GTGCCGTCGCTGCCAGTCCCGTCGGT CGGGGCAGCCGGTCGGGGGAAAGCC

		CCTATGGAGCCCGCGGAGGAAGACA GGTCGGGGAGCCGCTCAACGTCTCCC AGCGCATACTACCCCGAGGGCGCGT CGGCGTTGGCCAAGCACAACCTTGCG AGGAGGTTGTGCCAAG**GGATCC**
JBam_E57	JBam_F18	**GGATCC**ATCGGTTGAAGTCCGACTG GGATGCCTGATTGAGCAGAATGGCTC TTCACATCATCGTTCTGGGAGAGGCT CACATGTTCCAAGATCGCTGACGTTC TAGACGGGAGTCACCTACCGCAGGA GCTCGGATGGAGATTTTCTGTTGGCG GAGTCCGGGGAAGTCCGATCGCTAG GGAAGTCCGTCTGGGATTTATCTGAT GTGGGGGCTCATCCGAAGATCGTCCG CTGGGGAAGCCCGATTTTACGGGGA GCCTGATAAGAGGTCGGTCGGCGAT GGACTTCAGATAATGCTGGGGAGGC TCGGCAGACATCGGGGGCGATCGGC CATCGCAGGAACCCAGCTGCTGGAG CTCATCCAT
JBam_E43	JBam_D35	TCCCGAGGGGGAGGACAGGGAGTGG GTGCCGGAGGAGACGATCGGAGGCA GCTCCGTTGAGCGCTCCTTTAAGAAC AAGGGCGCTGCGAAGATTCGCCCCTT CCCTCTAGCGCCAATCCTGTTGGTGC AAAAATCCGCCTGCACCGGAGAAGC TGGAGTCGGGGAAGCCGCGGTCGCC GCCGGGACCTGCAAGGGAAGTCTAA ACCGGAGGTGGGGTTGCTCCGGCAA GACCCTCCGACGCTCAAGTCAGTTCT CTGCCTCAACAAGAATGGAGTGCTCG AACGGAGAATTTAGCAGAGTTTTTGA GATAAGAAATGAGCTTAGAGAATAA CGTATCT**GGATCC**
JBam_E77	JBam_D60	ACCGACGTGGACTATCCATGAACGG ACCCTTCCGACCCCTCCTCGGCACCG TCCGCCTCCTCGTCGGCAACCTCCGC CTCCTCTCCCGCCGCCGCGATTGCCC CGCCTTCGACGCCGTCGGCATTCGCC GCCCTCCGGCCGCTTCCACGGCGTCC TTAACGTCGGCGCCACGATTCTCCGC TGCGTCTCTCTTGTGGCCGCCAAGGT CCTCGCCACCTGCCCCGCCGTCAGCT ACCGCGACCTCATGGGGAAGGAGGC CTCCAAGATCCGGCGCTTCCAGCGAC TCCGGCGGCGGCGGCCAGGCGTGTG GATTCGTTCAT
JBam_E82	JBam_D25	ATCCGCCATCAGTTCCAGCAGCGGCC GCACGATCCCGTCCCTTGACGGCTAG GATCTTGTTCTCCCGAGTGGAGCACA GCAAGAAGAGCGCCGTGGAGGCGTC CTTCTTCCCCCTCTGGCCGCCGGTCTC GAGGAGGTTGACAAGGTGGGGGATC GCGCCGGAGCGGCCGATGGCGATCT

			TGTGCTCCTCGATCTTGGAGAGGCGG AGGAGGGCGCAGGCGGCGTTCTCGC GGGCGGCGGGGGTACCGGTCTTGAG GACAAGGATGAGGGGCCGGATGGCG CCAGCGGCGGCGATG**GGATCC**
	JBam_E88	JBam_D32, JBam_D44, JBam_F10	**GGATCC**CCGGAGGGTCACTCCTGTC ATGAACTTCGGCTGGGGGGGTATTTTA TACCCAACACCAGTCCCCCTACTTTC GAGTTCGAATTTCGAATGAAGGAAG TACAGAAAAATTTTACTACTGCCGAA GTTGTCCCCTTGAACCCTGTGCATGA TCGCCCCCCAGATATTTTTGGCAATT AAATGCGCGCGTGCTGGAGTCTTTTC GAATCGGGGCGATTCGAAGAGGGAC CCTTCGAAATTCTCGATGGTACGCTG GCCCAAGTACGGTGCAGTAATGGCTC CGTCAGCTGTCAGCCGCTTTTAGCCG CCTGCCGTGGCGAGTGGGATACGCAT CGAGCGCAGGTCGACCTGGGGGAGA TTCACGATCATTATGGCGCCGGAT
	JBam_F13	JBam_D27	AATCCAGCGTCCAGGGCAGCAGCGG TGGCCGGAGGGAGTGGTTATGGTGG TGGAGGCAGCCGCAGCGGCGCTACC CGTGGAGGGTATGGGTTCAACGGTG GGGGCAGCGGCGGTGGTGGGAGAGG AGGGGGTGCGAGAGGCGGTGGCTGG GGTTACGGAGGCGGTGGTGCCTGCTA TAACTGTGGTGAGACTGGTCACATCG CTAGGAATTGCTACCAAGGAGGCGG AGGCGGTGGGAGGTACGGCGGCGGT GGCG
	JBam_F19	JBam_D78, JBam_D79, JBam_E9	TCCGGTGCATTAGTGCTGGTGTGATC GCACCCACAATGATTTGTTCGAGATT CGTCGATATAACGTCGCGGTCGTCGC ACGCCATCTGTAACCCACCCACAGTC CTGGCTGGTCGGGTACCGGACCCATC AAGTGGGTCCCGCGACCTCGCACGG CACTGTCGGGCTCCAGACTCAGTTTT TTCTGAGAAAAACGTTACCCGCGGCA GAAGAAAGAGATCTCCATAAAATTA ATGAAAAAAGTAACTTGAATAAAGT AAAAGGGAACGAAGATTAAAAGGGT AGGCAACACGAGGACTTCCGACGGG TGGTCACCCACTCCCACGTACGACTC GTGCCCGACGCACGCTCGACTGCGG AGT

d) **Homology search against four different genome and transcriptome databases**

Homology analysis of all R3 contigs revealed that an average of 40% of contigs showed significant homology ($E<10^{-10}$) to the in-house oil palm mesocarp transcriptome and date palm genome while only 15% and 6% of contigs had significant homology to rice and Arabidopsis genomes that are freely accessible through GenBank, respectively (Figure 3.16). This result correlates well with the genetic distance of oil palm against date palm, rice and Arabidopsis.
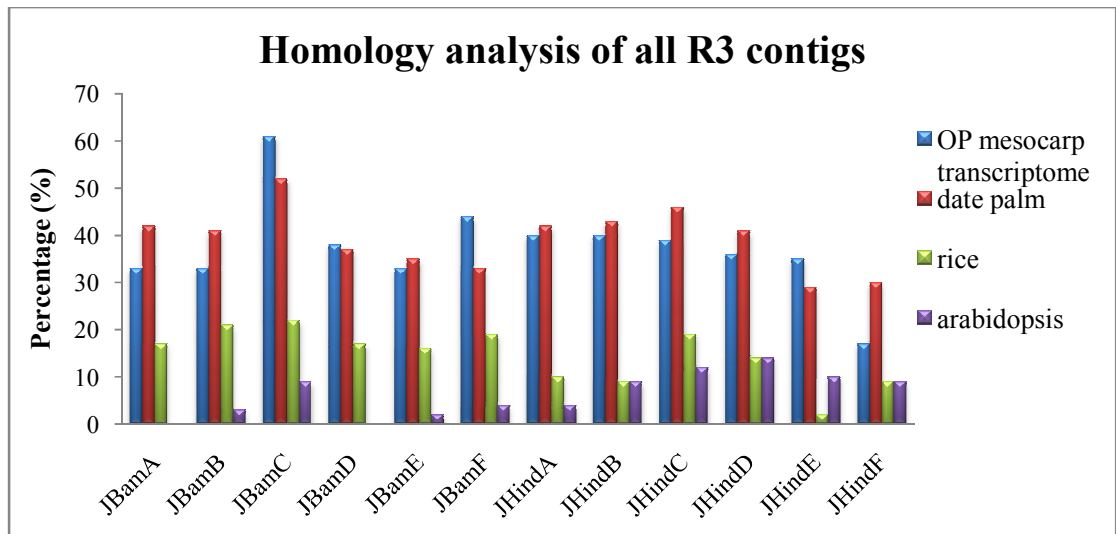
Figure 3.16: Percentage of contigs within each RDA pool with significant homology search (E-value $<10^{-10}$). Homology analysis was conducted against oil palm mesocarp transcriptome (Mayes *et al*., unpublished data), date palm genome (Al-Dous *et al*., 2011) as well as rice and Arabidopsis genome available on GenBank.

e) **Homology search against MPOB oil palm *pisifera* genome assembly V5**

Better results were obtained from homology analysis against the oil palm *pisifera* genome assembled by MPOB (Table 3.14). More than 60% of round 3 contigs had significant homology ($E<10^{-10}$) to the *pisifera* genome with 41.3% and

51.4% of contigs from the *Bam*HI and *Hin*dIII RDA analysis, respectively, having E-values less than 10^{-100}. Redundant sequences were contigs that hit multiple locations in the genome with similar E–value which constituted 27.1% and 34.2% of the *Bam*HI and *Hin*dIII contigs, respectively. A contig was categorised as no-hit when the hit region was <50% that of the contig's length. The *Bam*HI analysis returned a much higher portion (11.7%) of "no-hit" contigs than the *Hin*dIII analysis (2.4%).

Table 3.14: Homology search of round 3 contigs against oil palm *pisifera* genome assembled by MPOB. Contigs were categorised according to their E-value.

JBam contigs (%)	A	B	C	D	E	F
E-value < 10^{-100}	48.1	35.6	34.8	40.7	41.2	48.1
$10^{-100} \leq$ E-value $\leq 10^{-50}$	20.3	17.8	4.3	11.1	12.7	14.8
E-value > 10^{-50}	5.1	7.8	4.3	2.5	5.9	3.7
Redundant sequence	12.7	28.9	43.5	30.9	28.4	33.3
No-hit	13.9	10	13	14.8	11.8	0
JHind contigs (%)	A	B	C	D	E	F
E-value < 10^{-100}	53.8	55.3	34.6	40.9	57.7	52.2
$10^{-100} \leq$ e-value $\leq 10^{-50}$	15.4	8.5	11.5	9.1	11.5	0
E-value > 10^{-50}	1.9	2.2	0	0	3.8	0
Redundant sequence	26.9	29.8	53.8	45.5	25	47.8
No-hit	1.9	4.3	0	4.5	1.9	0

f)   Homology search against retroelements databases and GenBank

Homology search against the retroelements database TIGR and Repbase as well as GenBank successfully identified a proportion of contigs with significant homology to repetitive DNA. A total of 11 contigs from both *dura* and *pisifera* pools of the *Bam*HI analysis were homologous to oil palm repetitive DNA deposited in GenBank while another six and five contigs from the *Bam*HI and *Hin*dIII analysis, respectively, showed significant homology against the TIGR database at E-value <10[-10]. Meanwhile, an additional seven and nine contigs from the *Bam*HI and *Hin*dIII analysis were also found to be significant hits against the retroelements Repbase

database at a lower E-value of $<10^{-5}$. In short, a portion of contigs generated from the current RDA analysis were concluded to be repetitive DNA.

Interestingly, BlastN search of the sequences using GenBank revealed that a small portion of contigs showed significant homology to organelle DNA (chloroplast and mitochondria). There were seven and six contigs in the round 3 *Bam*HI and *Hin*dIII RDA analysis, respectively, that were homologous to chloroplast DNA. Surprisingly, a total of 33 contigs from the *Hin*dIII RDA analysis were found to be homologous to mitochondrial DNA of date palm while another six contigs from the *Hin*dIII analysis were homologous to mitochondria of genus Sorghum. The majority of these contigs were disclosed to be located in all the *dura* and *pisifera* pools. This identification of organelle DNA is consistent with previous finding of Sanger sequencing of the first study of RDA analysis (RDA pools A, B, D and E) (Table 3.9 and Figure 3.7). The isolation of common repetitive and organelle DNA from the difference products of RDA confirmed the postulation that common sequences were somehow being selected and enriched after subtractive hybridization and possibly mask the presence of real difference products.

g)   <u>Homology analysis of putative shell-thickness related-contigs</u>

Followed up from the identification of shell-thickness related RDA contigs (Table 3.13), homology result of these putative contigs against MPOB *pisifera* genome assembly as well as GenBank is summarised in Table 3.15. It was discovered that four of the *pisifera*-specific contigs were redundant sequences when compared to the *pisifera* genome assembly. Homology analysis against GenBank further confirmed that two of them were repetitive and ribosomal DNA.

On the other hand, B57 from the *Bam*HI and A23 from the *Hin*dIII analysis were found to be located on orphan contigs ("co"), contigs that cannot be ordered into a scaffold ("sc") during the genome assembly process. Contig JHind_A23 was significantly homologous to the mitochondrial DNA of date palm. The six putative contigs that were redundant sequences or located on orphan contigs were not suitable for identification of the shell-thickness gene.

Meanwhile, the *pisifera* genome assembly had been anchored to MPOB T128 genetic linkage map. Nine putative contigs were found located on scaffold that could be anchored to a particular linkage group and these contigs were located on different pseudochromosomes. Three of the contigs, JBam_B90, JBam_B23 and JBam_D24, were located on the same PLG01 while contigs JHind_A50 and JBam_D29 were located at PLG11.

According to recent publication of MPOB, *SHELL* gene was located on tpseudochromosome PLG04 and mapped by sequence similarity to assembly scaffold p5-sc00060 (Singh *et al.*, 2013b). None of the putative or common RDA contigs isolated from the present project were mapped by sequence similarity to assembly scaffold p5_sc00060 and none of the putative RDA contigs were mapped to scaffold located on PLG04. Details examination of all RDA contigs that mapped to scaffold located on PLG04 revealed those contigs were either present in reciprocal analysis or only in one out of three of the *dura/pisifera* pools which cannot be declared as putative RDA markers.

Table 3.15: Homology analysis of putative shell-thickness-related RDA markers against MPOB *pisifera* genome assembly as well as GenBank. Sc, scaffold; co, orphan contig; PLG, pseudochromosomes.

Category	Representative	*Pisifera* genome assembly	Linkage group[a]	Putative function revealed by GenBank Search (Organism, E-value, Identity)
Dura-specific	JBam_B17	p5_sc00441	-	-
	JBam_A19	p5_sc00071	PLG14	-
	JBam_B90	p5_sc00054	PLG01	Oxidoreductase (*Zea mays*,1e-29, 81%)
	JBam_B67	p5_co387164	-	-
	JBam_B23	p5_sc00001	PLG01	-
	JHind_A23	p5_co794290	-	Mitochondrion (*Phoenix dactylifera*, 1E-71, 55%)
	JHind_A50	p5_sc00332	PLG11	-
Pisifera-specific	JBam_D24	p5_sc00041	PLG01	-
	JBam_D29	p5_sc00085	PLG11	-
	JBam_D68	Redundant	-	-
	JBam_E57	p5_sc04008	-	-
	JBam_E43	Redundant	-	Repetitive DNA (*Elaeis guineensis*, 5E-103, 70%)
	JBam_E77	p5_sc00126	PLG08	-
	JBam_E82	p5_sc00090	PLG09	-
	JBam_E88	Redundant	-	-
	JBam_F13	p5_sc00014	PLG13	-
	JBam_F19	Redundant	-	5S ribosomal RNA gene (*Arabidopsis Thaliana*, 3e-11, 91%)

[a]**Linkage group**= The *pisifera* genome assembly was anchored to the T128 genetic linkage map by MPOB

In short, reciprocal RDA analysis using different *dura* and *pisifera* pools had not been successful in identifying markers closely linked with the shell-thickness region. Instead, repetitive and organelle DNA sequences were isolated from the difference products of RDA which suggested that during subtractive hybridization, common sequences were being selected and enriched for unknown reasons, most likely to be technical, and possibly mask the presence of real difference products.

3.4 Discussion

3.4.1 Fingerprinting and generation of DNA bulks

Three different types of samples were used in Representational Difference Analysis (RDA), Deli *dura*, *dura* and *pisifera*. Deli *duras* are normally used as female parents in almost all major oil palm commercial hybrid seed production programmes in Malaysia and Indonesia (Soh *et al.*, 2006). Deli *dura* palms were originated from the progenies of the four palm seedlings planted in Bogor Botanic Gardens in Indonesia in 1848 and later distributed to the plantations in Deli province in Sumatra and thence to Malaysia (Soh *et al.*, 2009). Meanwhile, self-pollination of non-Deli *tenera* parents of the 768, 769 and 751 crosses are expected to produce 25% segregants of *dura* and *pisifera* fruit types, respectively, in each cross, according to Mendelian inheritance. Therefore, *dura* and *pisifera* from the same controlled cross are siblings with the same genetic base, enabling different fruit types to be pooled for Bulked Segregant Analysis (BSA) coupling with RDA or AFLP technique.

Besides the study of *sh* gene, BSA approach had also previously been used in combination with AFLP markers for the monogenic *Virescens* trait (fruit skin color) in oil palm by creating two different DNA bulks, "*Nigrescens* bulk" and "*Virescens* bulk" (Seng *et al.*, 2007). Although no publication can be found for RDA coupled with the BSA method, Oh and Cullis (2003) have previously performed a combined sample representational difference analysis (csRDA) with DNA from four different genotypes of flax. Using this approach they successfully isolated DNA sequences that have undergone physical rearrangements in the flax genome. Therefore, BSA should work well for the study of the shell-thickness gene in combination with RDA.

For the present study, BSA approach was undertaken by bulking the respective *dura* and *pisifera* samples of the same segregating population followed by reciprocal RDA analysis with the aim to identify marker(s) closely-linked to the shell-thickness gene that controls fruit type. A non-classical BSA approach was also exploited concurrently by bulking the 744 *Deli* dura samples and analyzed against *pisifera* samples from the 769 controlled cross. In view of the different genetic base of both the 744 and 769 controlled crosses, it is expected that non-identical genomic regions will be enriched alongside with the target shell-thickness region. For the present study, only 10 palms were used to create the Deli *dura*, *dura* and *pisifera* bulks of different controlled crosses. It should be noted that smaller bulks have higher frequency of false positives (Michelmore *et al.*, 1991) and hence the use of multiple bulks were to identify consistent markers to the shell-thickness gene.

Before constructing the bulks, it is critical to ensure the true-identity of each and every sample for the BSA approach to work properly. All samples within the same bulk should truly come from the same controlled cross. Existence of any out-crosses in either or both bulks will lead to identification of false positive differences/polymorphism between the *dura* and *pisifera* bulks. This will be particularly a problem with small bulk sizes. Therefore, all the samples need to be fingerprinted before construction of the bulks to remove any out-cross/mis-sampled palms from the bulks.

In this project, legitimacy checking of samples was performed by fingerprinting all the samples together with their respective self-pollinated parents. Legitimacy of samples can be determined by comparing their genetic profile with those of their respective parents. In this study, there is only one parent for each

controlled cross as the samples were derived from self-pollination. Thirteen oil palm SSR markers were used in this study to determine the inheritance pattern of the samples and it was found that each controlled cross contained one outcross and these outcrosses were discarded from bulk generation (Tables 3.5-3.7).

The appearance of outcrosses in a controlled cross is rather common. Previous cDNA-RFLP molecular mapping in oil palm had shown similar results in which some of the palms were found to have bands not found in the parental palms (Singh *et al*., 2008b). The authors had attributed the appearance of extra bands to "illegitimate" palms caused by pollen contamination. Similar observation was made during RFLP genetic mapping of an oil palm controlled cross and was also attributed to pollen contamination (Mayes *et al*., 1997). Occurrence of pollen contamination has previously been reported for controlled crosses of oil palm (Chin, 1995). Therefore, it is sensible for samples that contain fingerprint bands not found in their respective parents be regarded as outcross due to contamination and to be discarded from construction of the bulks.

Complete sets of samples (*dura* and *pisifera* from the 768, 769 and 751 controlled crosses) were sent for genotyping at the end of the first year of the current study after completion of the harvesting of leaves samples for the 768 and 751 controlled crosses. Therefore optimization of the RDA protocol and first RDA analysis were performed using Deli *dura* from the 744 controlled cross as well as *dura* and *pisifera* samples from the 769 controlled cross without proven legitimacy. Later on, RDA analysis was repeated again using the legitimate *dura* and *pisifera* pools of the 768, 769 and 751 controlled crosses. This was termed as the second RDA analysis.

3.4.2   Optimization of the RDA protocol

a)   <u>Optimization of restriction digestion</u>

The success of the RDA protocol is highly dependent on the reduced complexity of the genome sequence through creation of representations or subgroups of DNA populations to ensure complete re-association of rare target sequence during subtractive hybridization. Choice of the restriction endonuclease used in RDA greatly determines the complexity of the generated subpopulation or representation. Bishop *et al.* (1983) working on a model for restriction fragment length distributions showed that *Bam*HI generated a mean fragment length of 5,534 bp in the human genome with 16.5% of the fragments less than 1 kb, while *Eco*RI and *Hin*dIII has a mean fragment length of 3,013 and 1,873 bp, respectively, and 28.2% and 41.3% of fragments smaller than 1 kb. Lisitsyn and Wigler (1995) suggested that the complexity of human *Hin*dIII amplicons is close to the limit of RDA and on the other hand, any restriction endonuclease producing mean fragment lengths larger than *Bam*HI, the preparation of amplicons becomes irreproducible. Therefore, all restriction endonucleases with mean fragment length between these two extremities could be used for difference analysis of human DNAs. Even though the above research dealt with the human genome and the fact that oil palm genome is distinctly different from the human genome, a basic rule can be deduced about the desirable fragment size distribution. RDA might not work well with restriction enzymes that cut any particular genome too frequently or, on the other extreme, rarely.

Six different restriction endonucleases were tested in the present study to examine their ability to cut the oil palm genome and the fragment distribution generated, *Bam*HI, *Eco*RI, *Hin*dIII, *Hpa*II, *Mse*I and *Pst*I. The tested *Bam*HI, *Eco*RI,

*Hin*dIII restriction enzymes were previously shown to cut the oil palm DNA well (Cheah *et al.*, 1993) and *Pst*I was used to construct the first RFLP oil palm genetic linkage map by Mayes *et al.* (1997).

Figure 3.2 noticeably illustrates that complexity of the representations generated by frequent cutter *Mse*I is not low enough to allow complete re-association of rare sequences during hybridization while *Hpa*II enzyme which minimally cut the 744 Deli *dura* DNA might produce amplicons that contain too small a fraction of the original DNA population to allow isolation of sufficient target sequences. This suggests that *Mse*I and *Hpa*II might not be ideal restriction endonuclease for RDA of oil palm genome. On the contrary, *Bam*HI, *Eco*RI and *Hin*dIII seem to produce representations with reasonable genomic complexity. Despite that *Pst*I enzyme seems to cut oil genome minimally, *Pst*I enzyme with its methylation sensitivity targets hypomethylated gene-rich regions of the genome (Schouten *et al.*, 2012); it could be a potential candidate for RDA analysis.

One concern about RDA is that the use of representations implies that not all of the potential differences between two genomes can be isolated. This has led to the suggestion from Hollestelle and Schutte (2005) that several representations generated using different restriction enzymes may be analyzed to isolate more of the target sequences present in the original tester population when necessary. Therefore two different enzymes, *Bam*HI and *Hin*dIII, were selected for the present study.

Partial digestion is detrimental as it introduces artefactual difference products due to different cleavage pattern among tester and driver populations generated by incomplete digestion of the bulk DNA (Hollestelle and Schutte, 2005). Therefore

samples were incubated with *Bam*HI or *Hin*dIII enzymes for overnight 16 h in the present work to ensure complete digestion.

b) **Optimization of the PCR**

Amplicons of RDA are generated through a "whole genome" PCR that amplifies an entire population of DNA sequences instead of a single sequence. However with the inherent limitation of PCR, small DNA fragments are more efficiently amplified than large fragments. Therefore DNA fragments smaller than 1,500 bp will be preferentially amplified (Kinzler and Vogelstein, 1989), generating amplicons that can represent a subpopulation of the original tester and driver DNA populations. Figures 3.3, 3.4 and 3.9 correlate well with this statement in which amplification of average DNA fragment sizes of less than 1 kb was observed.

A whole genome PCR generates more PCR products and thus requires more PCR reagents and enzyme. Exhaustion of reagents affects mainly the extension of long DNA fragments and this can be visualized on agarose gel as smearing of PCR products toward the well indicating the existence of ssDNA products (Hollestelle and Schutte, 2005). Baldocchi and Flaherty (1997) demonstrated that optimal number of cycles can vary as a function of polymerase activity, choice of restriction enzyme, template concentration, digestion and ligation efficiency, freshness of PCR buffer and annealing temperature. Consequently they suggested that the number of PCR cycles should be optimized by performing a pilot PCR with various numbers of cycles and selection of cycle number that gives optimal yield of 0.05-0.10 μg/2 μl PCR product.

There was no sign of reagent exhaustion during generation of the *Bam*HI and *Hin*dIII amplicons in the present study (Figures 3.4 and 3.9). This may suggest that

the 20 cycles of PCR is suitable, if not optimal, for amplicons production. It is, however, advisable to incorporate a pilot PCR before generation of tester and driver amplicons to ensure maximum recovery of PCR products without exhausting the reagents.

3.4.3   Reciprocal subtractive hybridization of amplicons

Representational difference analysis was first introduced as a tool for finding the difference between two samples (Lisitsyn *et al.*, 1993). Since then this technique has undergone several technical improvements for different applications, such as study of differential gene expression using cDNA-RDA (Hubank and Schatz, 1994); shortened-protocol by eliminating the representation steps for small and less complex genome (Strathdee and Johnson, 1995); methylation-sensitive RDA (Ushijima *et al.*, 1997); generation of representation using arbitrarily primed-PCR (Yoshida *et al.*, 1999) and a simplified protocol introduced by Felske (2002).

No matter how RDA has evolved for different purposes, the subtractive hybridizations step is core to this technique in which one DNA population (tester) is hybridized against an excess amount of another DNA population (driver). The target sequences are the differences between these two DNA populations present in tester but not in driver. During the hybridization, the DNA mixture is denatured and then allowed to randomly re-associate. Three types of hybrids are formed during this process, tester-tester homoduplexes, tester-driver heteroduplexes and driver-driver homoduplexes. Excess driver acts as a competitive inhibitor for the re-annealing of tester DNA which is also common to driver DNA, hence common sequences between these two DNA populations are subtracted out.

In order to ensure only self-reannealed unique tester amplicon is amplified after hybridization, it is necessary to change the adaptors on tester so that amplifiable primer binding only occurs in the tester-tester hybrid. Adaptors from the representation step should never be used in the subtractive hybridization step to prevent driver amplification due to uncleaved primers. Furthermore, the same adaptors should never be used in two consecutive rounds of RDA in order to achieve highest possible enrichment. Therefore, three different sets of adaptor pairs were employed in the present study; R adaptor pair (Table 3.2, primer set 1) was solely for amplicons generation while J and N adaptor pairs (Table 3.2, primers set 2 and 3) were used alternatively for iterative rounds of subtractive hybridization (Lisitsyn *et al.*, 1993). Felske (2002) introduced another set of RDA adaptors, the well known T7/SP6 primers, for studies of microdiversity. This primer sets are known not to cross-react with the bacterial genomic DNA (Kimmerly *et al.*, 1994). Nonetheless, the main point is different adaptors should be used for tester and driver regardless of the types of adaptor used.

During subtractive hybridization, tester-driver heteroduplexes contain only one adaptor on the tester strand but not the complementary driver strand. DNA polymerization creates the primer binding site on the complementary driver strand and hence leads to linear amplification of the driver strand during kinetic enrichment. This background level of linear amplification is reduced by degrading the single-stranded driver DNA with mung bean nuclease. Hollestelle and Schuttle (2005) commented that mung bean nuclease should be used but not S1 nuclease even though both nucleases are active against ssDNA. This is because S1 nuclease is sensitive to nicks and nucleotide mismatches. RDA is heavily dependent on PCR amplification and various rounds of PCR might introduce mutation in DNA sequences which would

leads to formation of hybrids of not exactly complementary that would be recognized and degraded by S1 nuclease. Therefore, S1 nuclease is not recommended.

For the present study, three rounds of reciprocal subtractive hybridization were carried out for *Bam*HI and *Hin*dIII amplicons of *dura* and *pisifera* DNA bulks of first and second RDA analysis. A stepwise reduction of the complexity of the products in each subtractive hybridization was observed with clear bands of difference products were visible in the third round of reciprocal analyses (Figures 3.5, 3.6 and 3.10). This confirms the core principle of RDA of successive enrichment of potential target sequences through rounds of subtractive hybridization with increased stringency (Lisitsyn *et al.*, 1993). It was also noted here and elsewhere (Bowler *et al.*, 1999; Allen *et al.*, 2003) that RDA preferentially enriches for sequences between 200 to 450 bp in length.

However, it was unexpected to observe that the second reciprocal RDA analysis of both *Bam*HI and *Hin*dIII representations gave rise to highly similar enrichment profiles (Figures 3.10 and 3.11). Different enrichment profiles were observed between RDA analysis of *Bam*HI and *Hin*dIII amplicons.

a)   **Troubleshooting of highly enrichment profiles between reciprocal RDA analyses**

In view of the highly similar profiles between reciprocal analyses, examination of the effectiveness of RDA using a positive control (Figures 3.12 and 3.13) indicates that positive control was successfully selected during subtractive hybridization step but the amount was too low to be detected, suggesting inefficient

enrichment. Repetitive sequences that exist abundantly in the oil palm genome are likely to be transmitted if subtractive hybridization is incomplete.

Different enrichment profiles for RDA using different restriction endonucleases as well as for reciprocal analysis have been published by other researchers. For example, reciprocal RDA analysis on two date palm varieties using two different restriction enzymes (Voster *et al.*, 2002) resulted in only one of the four subtractions produced a DNA difference product after one round of hybridization, which was with the Barhee *Bam*HI-digested DNA as tester and Medjool as driver, not *vice versa* nor in *Hin*dIII-digested DNA. Similarly, RDA libraries generated from reciprocal analysis of honey bee worker and queen larvae differed considerably in term of percentage of sequences with similarity to predicted genes or unpredicted genes as well as the functions of the predicted genes (Humann and Hartfelder, 2011). Differential expression of selected gene sets was confirmed by quantitative RT-PCR, representing candidates of modulators of caste-specfic development of honey bee ovary. Therefore it was considered unusual in the present study for obtaining highly similar enrichment pattern between reciprocal analyses.

3.4.4 Sequencing of RDA difference products

Characterization of difference products obtained from the first RDA analysis was initially performed using conventional Sanger sequencing. It is unexpected to find out that the two families of clone, PDDH-3 and PDDH-5, were homologous to mitochondrial and chloroplast DNA, respectively. Characterization using Sanger sequencing in the present study is deemed time consuming as it involves laborious cloning, transformation followed by propagation and plasmid extraction of individual clones prior to sending for sequencing. This

141

approach clearly has a limited ability to sample all of the potential difference products present.

In order to understand the composition of RDA difference products that were obtained, especially those that exist in small quantities, the 454 pyrosequencing technique was introduced into the present study. Round 2 and 3 difference products from the reciprocal analysis of the 744 Deli *dura* against the 769 *pisifera* as well as reciprocal analysis of the 769 controlled cross, with and without the outcross included, were sent for 454 deep-sequencing. The sensitivity of the GS FLX Titanium 454 sequencing was proved to be able to detect rare genetic variants constituting as little as 1% of the molecular population (Simen *et al.*, 2009; Le *et al.*, 2009). A comparative study was performed by researchers in Yale University School of Medicine who found that conventional Sanger sequencing failed to detect 95% of the low-abundance HIV drug-resistant variants, whereas 454 sequencing detected all mutations found by Sanger sequencing as well as additional low-abundance variants, with 62% detected at levels 1 to 5% and 38% were detected at levels 5 to 20% (Le *et al.*, 2009). With its proven high sensitivity, 454 sequencing can help to detect potential difference products that were suspected to be present in low amounts after 2 and 3 rounds of subtractive hybridization. The average read length of 454 sequencing is 400 to 500 base pair read which matches well with the observed RDA preferential products of between 200 to 450 bp in length.

Combining RDA approach with next-generation sequencing (NGS) is novel. Ho *et al.* (2013) successfully employed methylation-sensitive RDA coupled with NGS approach to identify candidate biomarker associated with embryogenic competency in oil palm. The authors commented that replacement of Sanger

sequencing with pyrosequencing not only abolishes the need of lengthy transformation and potentially isolating sequencing which are not clonable, but also enables the generation of large numbers of sequences including those present at low abundance, allowing examination of more comprehensive representation of difference between samples. Therefore 454 pyrosequencing was utilised to study the representation profile of round 2 and 3 difference products in the present study.

As expected, contigs generated from round 3 RDA analyses shared greater similarity with the database for the oil palm mesocarp transcriptome (Mayes *et al.*, unpublished data) and date palm genome (Al-Dous *et al.*, 2011) followed by rice and lastly Arabidopsis genome (Figure 3.16). Date palm and oil palm are members of the palm family hence they share greater similarity. Meanwhile, rice, like all grasses, is a monocotyledon, hence the rice genome has higher similarity with the monocotyledon oil palm than the dicotyledon Arabidopsis. This result corresponds well with the X-species analysis of the same set of oil palm DNA in which better signal was obtained from hybridization of oil palm DNA to the rice Affymetrix chip than to Arabidopsis affymetrix chips (ATH1) (Chai *et al.*, unpublished data). Availability of an annotated date palm genome would definitely be helpful in annotating the current set of RDA contigs.

Homology search of contigs against the MPOB *pisifera* genome assembly identified around 40% of the contigs, regardless of *Bam*HI and *Hin*dIII, as redundant sequences (Table 3.14). Redundant sequences are contigs that have multiple hits in the genome assembly, implying that they are common repetitive sequences in the genome. Repetitive sequences were estimated to make up 57% of the recently released 1.8 Gb of oil palm *E. guineensis pisifera* genome (Singh *et al.*, 2013a). The

identity of redundant sequences found in the present study were further confirmed to be repetitive DNA when BlastN search was conducted against retroelements database TIGR and Repbase as well as GenBank. The existence of repetitive DNA as well as organelle DNA, particularly in the *Hin*dIII analysis, constituted a substantial percentage of the contigs generated. This provides evidence for the hypothesis that common sequences, such as repetitive and organelle DNA, were being either enriched or insufficiently reduced during selective hybridisation, masking the presence of real difference products.

Nevertheless, pyrosequencing of RDA difference products has made the handling of large number of sequences and their identification feasible. The RDA technique should detect polymorphisms related to presence/absence of restriction recognition site as well as indel mutations and/or translocation. Contigs were classified as *dura*/*pisifer*a-specific with the criteria that the contigs had to be found in at least two of the *dura* pools, but not any of the *pisifera* pools and *vice versa*. A total of seven and ten contigs were selected as *dura* and *pisifera*-specific, respectively (Table 3.13). Out of these 17 contigs, only three of the *dura*-specific contigs appeared in all three of the *dura* pools.

Surprisingly, four of the *pisifera*-specific contigs were detected to be redundant sequences while one *dura*-specific contig was a homologue of mitochondria DNA (Table 3.15). Contig JBam_B67 was also found to be located on orphan contigs that cannot be assembled into any scaffolds. All these contigs were deemed not suitable for further characterization. The remaining putative contigs were homologous to scaffolds that anchored to different pseudochromosomes, mainly PLG01 and PLG11. However, the *SHELL* gene was recently reported to be located at

PLG04 and mapped to scaffold p5-sc00060 (Singh *et al.*, 2013b). None of the putative contigs were mapped to this particular scaffold nor located at pseudochromosome PLG04.

The successful application of RDA technique in addressing different embryogenic potential of oil palm explants using methylation-sensitive restriction endonuclease, *Hpa*II (Ho *et al.*, 2013), suggested that methylaiton-sensitive enzyme, such as *Pst*I, could be a potential enzyme for RDA analysis of the shell-thickness gene. *Pst*I enzyme with its preferential targeting of hypomethylated gene-rich regions of chromosomes (Schouten *et al.*, 2012) might potentially be useful in eliminating common repetitive sequences in the oil palm genome, allowing generation of RDA representation profile from gene-rich regions. Meanwhile, it is advisable to include a positive control in future RDA study to examine the effectiveness of the enrichment analysis.

In conclusion, this chapter reported the first attempt to isolate oil palm shell-thickness marker(s) using the Representational Difference Approach (RDA). An unexpected highly similar enrichment profile was obtained between reciprocal analyses in the second RDA analysis. Assessment of RDA technique with positive control indicated the amount of positive control in the enrichment profile was too low to be detected, suggesting enrichment was occurring, but inefficiently. Characterization of contigs assembled from pyrosequencing revealed substantial portions of mitochondria, chloroplast and repetitive DNA existed in both directions of subtraction analyses. This leads to the speculation that common sequences were being either enriched or insufficiently reduced for unknown reasons, most likely to be technical, masking the presence of real difference products. The reciprocal RDA

approach had failed to identity marker(s) closely-linked with the shell-thickness region in the present study.

Nevertheless, this chapter reported the significance of coupling of RDA with the NGS technique in generating large numbers of sequences covering those present in low abundance, allowing more comprehensive understanding of the representational profile(s) generated. Therefore this novel combinational method is comparatively more useful than conventional transformation and the Sanger sequencing-based method.

Chapter 4

Approaches to develop

Shell-thickness marker(s) using

Amplified Fragment Length

Polymorphism (AFLP)

4.1 Introduction and Objective

Amplified Fragment Length Polymorphism (AFLP) is a well established molecular marker technique which was first published in 1995. AFLP is the selective PCR amplification of restriction fragments from a digest of total genomic DNA. It combines the ubiquity of endonuclease restriction sites throughout the genome (as for RFLP) with the flexibility and ease of PCR-based technology (such as RAPD) (Mba and Tohme, 2005). Molecular genetic polymorphisms are identified by the presence or absence of DNA fragments. With its high reproducibility, robustness and informativeness, AFLP has a wide range application in genetic diversity, population genetics, linkage mapping, parentage analyses and single-locus PCR marker development (Meudt and Clarke, 2007).

AFLP is a popular DNA fingerprinting technique for plant study. The fingerprints are produced without prior sequence knowledge which is very useful for analysis of orphan crops or plants with no available genome sequences. The AFLP technique has demonstrated its usefulness in the study of the oil palm genome. It has been previously used in somaclonal variation studies of oil palm tissue culture (Matthes *et al.*, 2001; Cheong *et al.*, 2006) as well as in the genetic linkage analysis of oil palm (Billotte *et al*., 2005). In the same paper, Billotte *et al.* (2005) also reported an AFLP marker of the shell-thickness gene, E-Agg/M-CAA132, with a distance of 4.7 cM from the *Sh* gene.

Therefore, in addition to testing the RDA method, this study aims to employ AFLP as a second molecular marker method to explore its effectiveness in identifying potential marker(s) close to the shell-thickness gene. In this study, a modified AFLP method based on the use of single enzyme and a single adaptor was applied. Five

different restriction enzymes, *Bam*HI, *Eco*RI, *Hin*dIII, *Mse*I and *Pst*I, were tested. At the same time, conventional AFLP using a combination of the frequent cutting *Mse*I restriction enzyme and the rarer cutting *Eco*RI enzyme was also performed. Although this combination of *Eco*RI/*Mse*I enzyme has been utilized by Billotte *et al.* (2005) to develop the shell-thickness AFLP marker, the oil palm population that they were working on was the CIRAD golden cross of *tenera* from the La Mé population (LM2T) and *dura* from the Deli population (DA10D), a totally different genetic background from current set of samples. Meanwhile, there are no reports of validation of this AFLP marker, E-Agg/M-CAA132, across different breeding programmes. Therefore, it is interesting and worthwhile to repeat their approach using current set of samples. This particular selective primer pair of *Eco*RI/*Mse*I was exploited in the present study.

4.2   Materials and Methods

The same legitimate *dura* and *pisifera* bulks from the 768, 769 and 751 controlled crosses were used in both the RDA and AFLP study. Five different single-enzyme AFLP using *Bam*HI, *Eco*RI, *Hin*dIII, *Mse*I and *Pst*I restriction enzymes and conventional *Eco*RI/*Mse*I AFLP were performed.

4.2.1   Restriction digestion-ligation

Restriction endonuclease digestion and ligation of adaptors were carried out using the modified method described by Gibson *et al.* (1988). Each bulked sample was digested overnight (16 h) at 37 °C in an aliquot containing 1 μg DNA and 20 U restriction enzyme with the buffer provided in a final volume of 20 μl.

An aliquot of 10 µl of digested DNA was added to 5 pmol of *Bam*HI, *Eco*RI, *Hin*dIII, *Pst*I or 50 pmol of *Mse*I adaptor pair (Table 4.1, Bioneer, South Korea), 1 U of T4 DNA ligase and ligase buffer in total volume of 50 µl. For conventional AFLP, 5 pmol *Eco*RI and 50 pmol *Mse*I as the adaptor pair were used. The mixture was incubated at 37 °C for 3 h. After ligation, the samples were diluted to 100 µl with TE buffer and were incubated at 65 °C for 15 min to inactive the T4 ligase.

To prepare the adaptors, oligonucleotides F and R (Table 4.1) were mixed in equal molar amounts in distilled water and were allowed to anneal at room temperature for 10 min.

4.2.2   Pre-amplification PCR

Pre-amplification PCR for single-enzyme AFLP was performed in a final volume of 50 µl containing 5 µl of ligated DNA fragments, 0.25 mM dNTPs, 1.5 mM MgCl$_2$, 150 ng of single primer with selective nucleotide A and 1.25 U of *Taq* polymerase in 1x PCR buffer provided by the manufacturer. For conventional *Eco*RI/*Mse*I AFLP, 100 ng each of *Eco*RI and *Mse*I selective primers were used instead. The amplification was performed with an initial denaturation at 94 ºC for 5 min, followed by 25 cycles of denaturation at 94 ºC for 30 s, annealing at 56 ºC for 1 min and extension at 72 ºC for 1 min, with a final extension at 72 ºC for 10 min. The PCR primers used in the single-enzyme AFLP had the same sequence as the adaptor F with one additional selective nucleotide A at the 3'-end while the *Eco*RI and *Mse*I primers for conventional AFLP had one additional selective nucleotide A and C at their 3'-end, respectively (Table 4.1). In the adaptor pair, adaptor R was not ligated to the DNA fragments as the adaptor was not phosphorylated and hence the adaptor R

dissociated from DNA fragments during the initial denaturation step in the PCR. *Taq*

polymerase filled in the overhang region in the first stage of the reaction.

Table 4.1: Sequences (5'-3') of adaptors and primers used for pre-amplification.

Name	Type	Sequences (5'-3')
BamHI-F	Adaptor F	GACGATGAGTCCTGAA
BamHI-R	Adaptor R	GATCTTCAGGACTCAT
EcoRI-F	Adaptor F	GACGATGAGTCCTGAT
EcoRI-R	Adaptor R	AATTATCAGGACTCAT
HindIII-F	Adaptor F	GACGATGAGTCCTGAC
HindIII-R	Adaptor R	AGCTGTCAGGACTCAT
MseI-F	Adaptor F	GACGATGAGTCCTGAG
MseI-R	Adaptor R	TACTCAGGACTCAT
PstI-F	Adaptor F	CTCGTAGACTGCGTACATGCA
PstI-R	Adaptor R	TGTACGCAGTCTAC
*EcoRI(C)-F	Adaptor F	CTCGTAGACTGCGTACC
*EcoRI(C)-R	Adaptor R	AAT TGGTACGCAGTCTAC
*MseI(C)-F	Adaptor F	GACGATGAGTCCTGAG
*MseI(C)-R	Adaptor R	TACTCAGGACTCAT
BamHI+A	Primer+1	GACGATGAGTCCTGAAGATCC<u>A</u>
EcoRI+A	Primer+1	GACGATGAGTCCTGATAATTC<u>A</u>
HindIII+A	Primer+1	GACGATGAGTCCTGACAGCTT<u>A</u>
MseI+A	Primer+1	GACGATGAGTCCTGAGTAA<u>A</u>
PstI+A	Primer+1	GACTGCGTACATGCAG<u>A</u>
*EcoRI(C)+A	Primer+1	GACTGCGTACCAATTC<u>A</u>
*MseI(C)+C	Primer+1	GATGAGTCCTGAGTAA<u>C</u>

(* indicates adaptors and primers used for conventional *Eco*RI/*Mse*I AFLP, underline indicates the selective nucleotide)

The amplification products were separated by electrophoresis on a 1% agarose

gel containing 1x SYBR Safe DNA stain in 1x TAE buffer. The gels were visualized

under UV illumination. The pre-amplification reaction products were diluted 500-fold

with TE buffer. These diluted products served as the templates for the final selective

amplification reactions.

4.2.3 Selective amplification PCR

In order to achieve higher resolution, selective amplification PCR products of

AFLP analysis were separated on LICOR 4300 DNA analyzer (LI-COR Biosciences).

The LICOR system uses highly sensitive infrared fluorescence detection technology

and thus a fluorescent primer has to be incorporated into final products. Generally,

forward primers were labelled with the fluorescent dye.

In this project, instead of labelling all forward primers in the reaction which is

very costly, a fluorescent labelled-universal primer (Schuelke, 2000) was used. This

approach works in such a way that there are three primers in the PCR system, a

forward primer with addition of a universal primer sequence at the 5'-end, a reverse

primer and a fluorescent dye-labelled universal primer. For the single-enzyme AFLP

approach, single primers A with two additional selective nucleotides work as both

forward and reverse primer, hence only one primer is used in a single PCR reaction.

These primers were modified with addition of 24 bp of universal M13 (-41) sequence

at the 5-end (Table 4.2).

Selective amplification PCR was set up in which each tube contains 2.5 µl of

diluted pre-amplification products, 0.2 mM each of dNTPs, 1.5 mM of $MgCl_2$, 0.4

µM of single primer A with two additional selective nucleotides (Table 4.2), 0.05 µM

of IRD700-M13 (-41) primer, 0.2 U of *Taq* polymerase in 1x PCR buffer in a final

volume of 10 µl. For conventional AFLP, 0.2 µM each of *Eco*RI and *Mse*I selective

primers (Table 4.3) were added to the PCR mix and a total of sixteen primer

combinations were tested (Table 4.4). The PCR reaction was performed with an initial denaturation of 94 ºC for 5 min, followed by 36 cycles with the following cycle conditions: denaturation at 94 ºC for 30 s, annealing for 30 s follow by extension at 72 ºC for 1 min. The annealing temperature in the first cycle was 65 ºC, subsequently reduced each cycle by 0.7 ºC for the next 12 cycles, and was continued at 56 ºC for the remaining 23 cycles. With two selective nucleotides at the 3'-end of primer A, a total of 16 primers for each restriction endonuclease were tested in the selective amplification step.

The fluorescent dye on the M13 (-41) IRDye-700 primer is light sensitive and has to be handled carefully to minimise light exposure. Therefore, all primer tubes as well as PCR tubes were wrapped with aluminium foil for storage until further analysis.

4.2.4 Electrophoresis using LICOR 4300 DNA Analyzer

Gel apparatus of LICOR 4300 DNA Analyzer was assembled according to manufacturer's application manual (LI-COR Biosciences). A 6% (v/v) Long Ranger gel [72 ml of 50% Long Ranger gel solution (Lonza Rockland, Inc.), 7 M urea (Hamburg, Germany), in 1x TBE buffer (Fermentas)] was cast by adding 150 µl of 10% ammonium persulfate (APS) and 15 µl N,N,N',N'-Tetramethylethylenediamine (TEMED) (Merck, USA) last to the solution. The gel was allowed to polymerize at room temperature for around 1 h.

Table 4.2: Sequences (5'-3') of primers used for selective amplification of single-enzyme AFLP with addition of M13 (-41) at the 5'-end and two additional selective nucleotides at the 3'-end.

Name	Type	Primer Sequence (5'-3')
BamHIAXX	Primer+3	<u>CGCCAGGGTTTTCCCAGTCACGAC</u>GTCCTGAAGATCC*AXX*
EcoRIAXX	Primer+3	<u>CGCCAGGGTTTTCCCAGTCACGAC</u>GTCCTGATCCTTC*AXX*
HindIIIAXX	Primer+3	<u>CGCCAGGGTTTTCCCAGTCACGAC</u>GTCCTGACAGCTT*AXX*
MseIAXX	Primer+3	<u>CGCCAGGGTTTTCCCAGTCACGAC</u>GAGTCCTGAGTAA*AXX*
PstIAXX	Primer+3	<u>CGCCAGGGTTTTCCCAGTCACGAC</u>TGCGTACATGCAG*AXX*

[Underlining represents universal M13 (-41) sequence; italics indicate the selective nucleotide; X can be A, C, G or T, a total of 16 combinations]

Table 4.3: Sequences (5'-3') of primers used for selective amplification of conventional AFLP with addition of M13 (-21) at the 5'-end and two additional selective nucleotides at the 3'-end.

Name	Type	Sequence (5'-3')
EcoAAC21	Primer+3	<u>TGTAAAACGACGGCCAGTA</u>GACTGCGTACCAATTC*AAC*
EcoACG21	Primer+3	<u>TGTAAAACGACGGCCAGTA</u>GACTGCGTACCAATTC*ACG*
EcoACT21	Primer+3	<u>TGTAAAACGACGGCCAGTA</u>GACTGCGTACCAATTC*ACT*
EcoAGC21	Primer+3	<u>TGTAAAACGACGGCCAGTA</u>GACTGCGTACCAATTC*AGC*
EcoAGG21	Primer+3	<u>TGTAAAACGACGGCCAGTA</u>GACTGCGTACCAATTC*AGG*
MseCAA	Primer+3	GATGAGTCCTGAGTAACAA
MseCAT	Primer+3	GATGAGTCCTGAGTAACAT
MseCTA	Primer+3	GATGAGTCCTGAGTAACTA
MseCTT	Primer+3	GATGAGTCCTGAGTAACTT

[Underlining represents universal M13 (-21) sequence; italics indicate the selective nucleotide]

Table 4.4: Primer combinations used for selective amplification of conventional *Eco*RI/*Mse*I AFLP.

*Eco*RI primer	*Mse*I primer	*Eco*RI primer	*Mse*I primer
EcoAAC21	MseCAA	EcoAGC21	MseCAA
	MseCAT		MseCAT
	MseCTA		MseCTT
	MseCTT		
		EcoAGG21	MseCAA
EcoACG21	MseCAT		MseCAT
	MseCTA		MseCTT
	MseCTT		
EcoACT21	MseCAA		
	MseCAT		
	MseCTA		

Gel apparatus was mounted onto the instrument and buffer tanks were filled with 1x TBE running buffer. The gel was pre-run for 25 min at 45 W, 1500 V, 40 mA and 45 °C to warm up and stabilize the gel. Meanwhile, 10 μl of formamide loading buffer [98% formamide (Fisher Scientific, USA), 10 mM EDTA (pH 8.0) and 0.1% bromophenol blue (Fisher Scientific, USA)] was added to the samples. Samples were then denatured at 95 °C for 5 min and immediately placed on ice. Half microliters of samples were loaded into the well. Electrophoresis was performed at 45 W, 1500 V, 40 mA and 45 °C for 210 min. Digital images were produced in real time by the sequencer.

4.3 Results

4.3.1 Pre-amplification PCR

Digestion of pooled genomic DNA with different restriction endonucleases followed by pre-amplification PCR with one additional selective nucleotide at the 3'-end of primer gave rise to different amplification profiles within the size range of 100 bp to 3 kb (Figure 4.1). The majority of bands from the *Bam*HI representation had band

size within 500 bp to 2 kb as compared to *Hin*dIII representations which had bands in the size range of 200 bp to 1.5 kb. It was observed that *Eco*RI, *Mse*I, *Pst*I subpopulations had bands ranging from 200 bp to 2 kb, 500 bp to 1.5 kb and 900bp to 3kb, respectively. Combinational digestion of *Eco*RI/*Mse*I enzymes produced a subpopulation of smaller size fragments, ranging from 100 bp to 1kb. All of these substantial smearing profiles indicate a successful pre-amplification PCR.
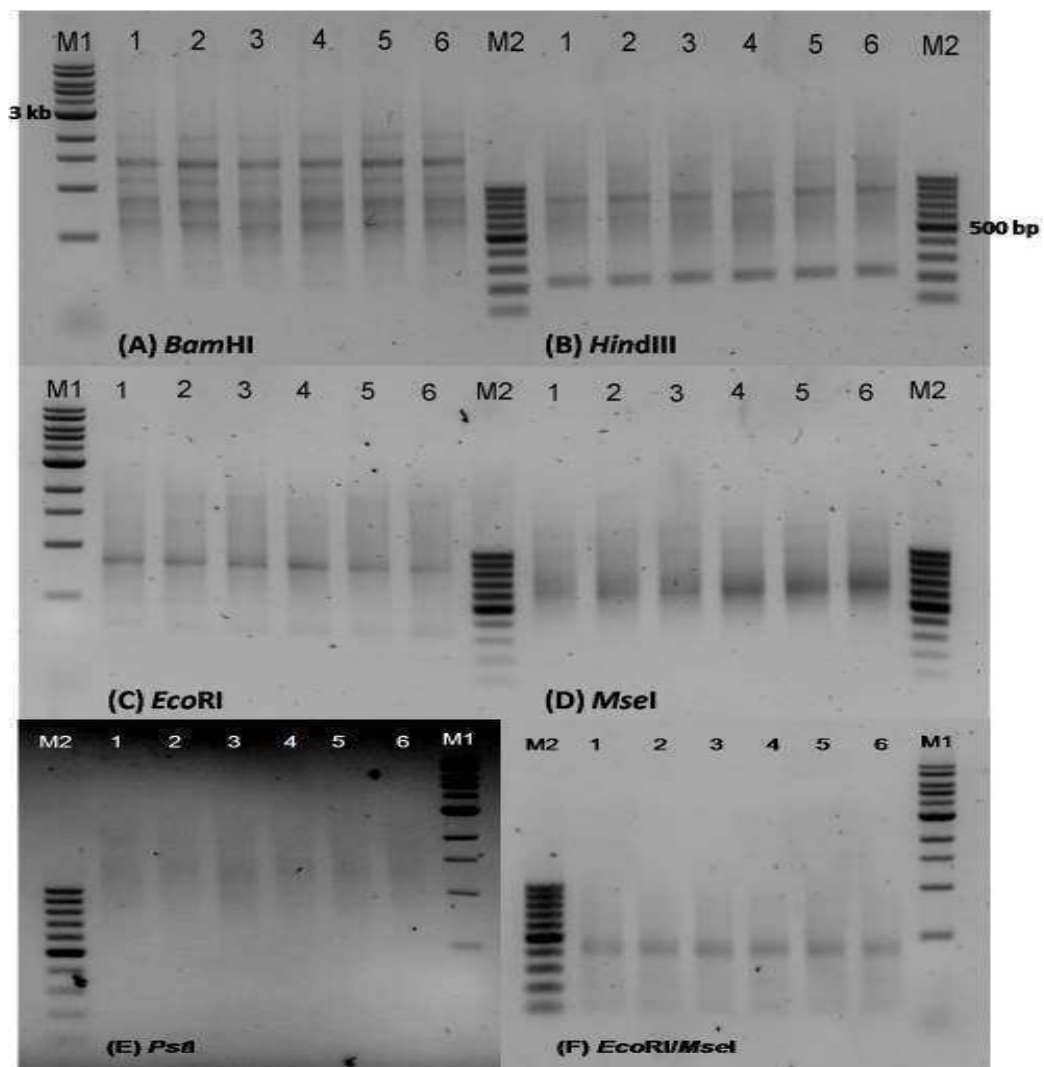
Figure 4.1: Pre-amplication profiles of pooled samples from the 769, 768 and 751 controlled crosses using primers with one additional selective nucleotide. (1) *dura* and (2) *pisifera* of 769; (3) *dura* and (4) *pisifera* of 768; (5) *dura* and (6) *pisifera* of 751 controlled cross. **M1**, 1 kb ladder; **M2**, 100 bp ladder (New England Biolabs).

4.3.2   **Selective amplification**

Electrophoresis for selective amplification with 2 additional selective nucleotides on a LICOR DNA analyzer revealed high levels of polymorphism for AFLP analyses (Figure 4.2; Appendices D1-D10). Polymorphism was observed within controlled cross (between *dura* and *pisifera* of the same controlled cross, not necessarily due to the shell-thickness gene, rather due to small bulk size), between controlled crosses as well as between populations (self-pollinated parents of the 769 and 768 cross are from the same cross while parent of the 751 cross is from another related cross). It was also noticed that the size range of selective amplified products for all single-enzyme AFLPs and conventional *Eco*RI/*Mse*I AFLP matched well with the size range of their pre-amplification profiles, suggesting that selective amplification had generated a subset of the original fingerprints.

Figure 4.2 shows that the combination *Eco*RI/*Mse*I had generated AFLP profile with a majority of bands smaller than 350 bp, whereas the selective profile of the single-enzyme *Eco*RI AFLP shows bands with size spanning the range 300 bp to 1 kb. The dense bands of the standard *Eco*RI/*Mse*I AFLP were difficult to score for polymorphism on the LICOR system.
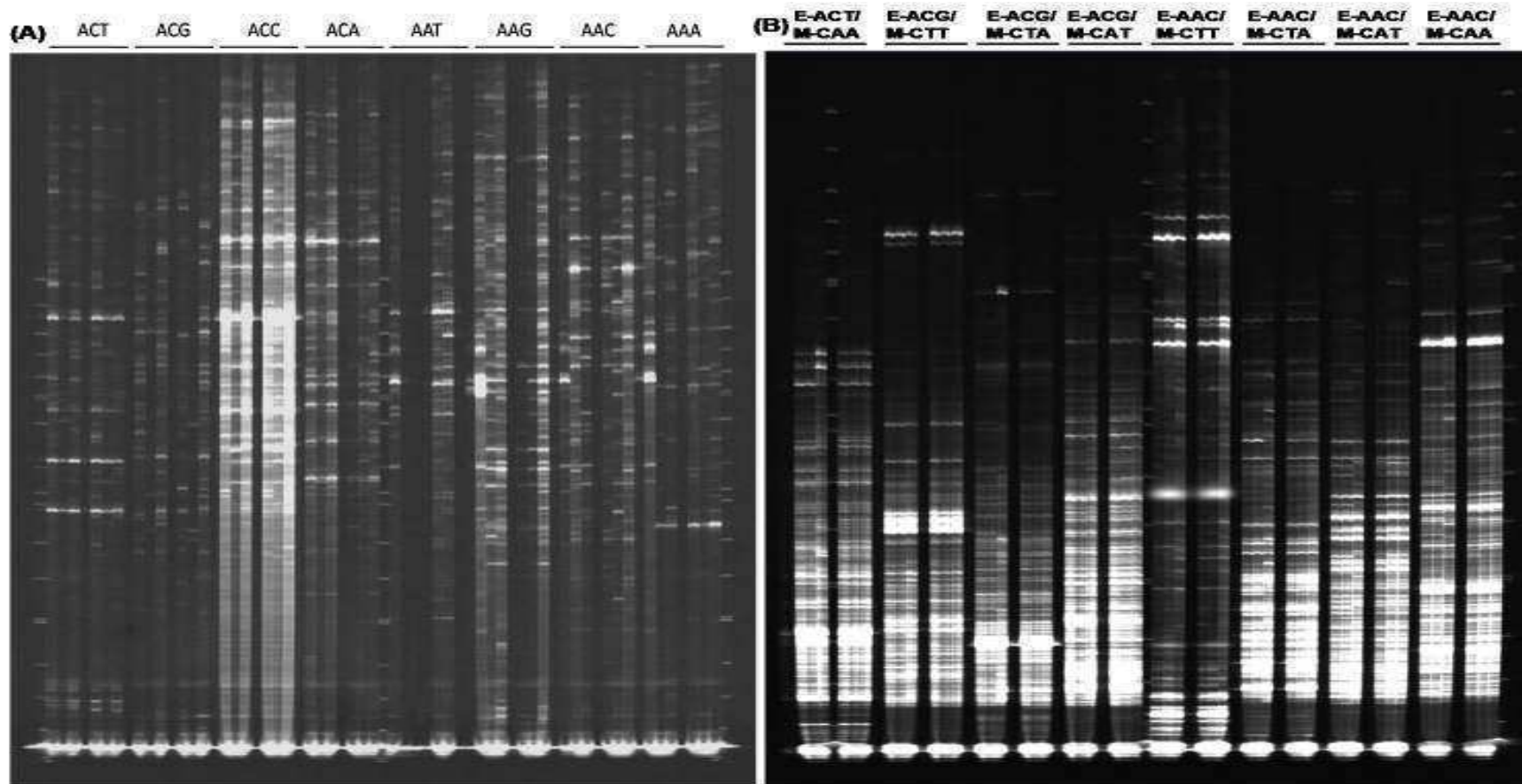
Figure 4.2: Example of electrophoresis of selective amplification profiles for AFLP. (A) Single-enzyme AFLP using *Eco*RI restriction enzyme with 3'-end of the primers having selective nucleotides from AAA to ACT and; (B) Conventional *Eco*RI/*Mse*I AFLP with their respective primer combinations.

4.3.3 **Identification of shell-thickness related-polymorphic bands**

Out of all the polymorphisms, 87 different polymorphic bands were found to be potentially related to the shell-thickness gene. A band of estimated size was only taken into consideration when it appears in at least two out of three of the pooled *dura* samples, but not in *pisifera* samples, or *vice versa*. It was found that AFLP analysis using the *Hin*dIII enzyme had the highest number of shell-thickness related-polymorphism (24), followed by *Eco*RI (21), *Eco*RI/*Mse*I (18), *Pst*I (9), *Bam*HI (8) and *Mse*I enzyme (7). These 87 polymorphic bands were further categorized according to their band intensity into three different groups (Table 4.5). *Bam*HI, *Eco*RI, *Hin*dIII, *Mse*I and *Pst*I, enzymes were simplified as B, E, H, M and P, respectively. The majority of polymorphic bands from the *Eco*RI analysis fell into either the strong or moderate signal groups while the weak signal group contains an abundance of bands from *Bam*HI and *Hin*dIII analysis. Figure 4.3 illustrates some bands with strong intensity.

Table 4.5: Grouping of shell-thickness related-polymorphic bands according to the signal intensity of bands.

Group	Polymorphic bands
Strong	E-AAA390, E-AAC400, E-AGA650, E-AGC900, E-AGC700, E-ATA660, E-ATA800, E-ATT345, E-ATT640, H-ACA700, H-ACA900, H-ACG625, H-AGA610, M-ACG410, P-ATT580, P-AGA780, P-AGG530, E-AGG/M-CAA250, E-AGG/M-CAA270, E-AAC/M-CAT325, E-AAC/M-CTT225, E-ACG/M-CAT325
Moderate	B-AGC565, E-AAA750, E-AAA600, E-AAC450, E-ACA555, E-ACG670, E-AGC620, E-AGC800, H-AAA850, H-ACA510, H-ACA700, H-ACA900, H-ACC515, H-AGG550, H-AGG600, H-ATT680, H-ATT850, H-ATT900, M-ACC580, P-AAA330, P-ATT675, P-ATT515, E-AAC/M-CAA525, E-AAC/M-CAT300, E-AAC/M-CTT380, E-AGC/M-CAA250, E-AGC/M-CAT300
Weak	B-AAA730, B-AAA620, B-AAC730, B-AAC740, B-AAT730, B-ACA600, B-ATT364, E-AAA395, E-AAA800, E-AAC450, E-AAC740, E-AAG580, H-AAA545, H-ACA560, H-ACC700, H-ACC770, H-ACC950, H-ATA580, H-ATA660, H-ATT675, H-ATT380, H-ATT440, M-AAA780, M-AAA355, M-AAG950, M-ACA880, M-ACT485, P-AGA520, P-ATG430, P-ATT480, E-AAC/M-CAA380, E-AAC/M-CAA370, E-AAC/M-CTA380, E-AAC/M-CTT335, E-AGC/M-CAA315, E-AGC/M-CAT230, E-AGC/M-CTT245, E-AGC/M-CTT410

(B = *Bam*HI; E = *Eco*RI; H = *Hin*dIII; M = *Mse*I; P = *Pst*I)
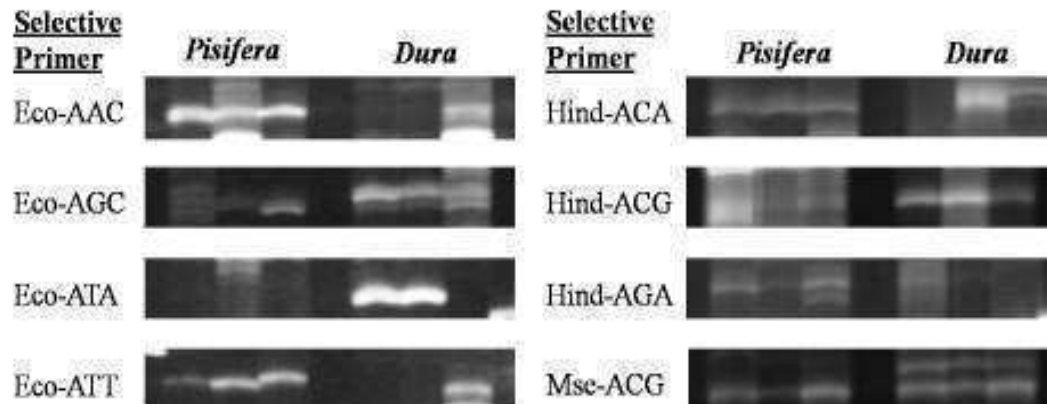
Figure 4.3: Examples of potential shell-thickness related-polymorphic bands that have strong intensity.

4.4 Discussions

Two different molecular marker techniques have been employed to identify marker(s) closely-linked with the shell-thickness trait. Representation Difference Analysis (RDA) allows enrichment and identification of target sequences through DNA hybridization and PCR enrichment (Lisitsyn *et al.*, 1993) whereas Amplified Fragment Length Polymorphsim (AFLP) is a whole genome profiling molecular technique allowing rapid evaluation of many thousands of loci for polymorphism.

Both AFLP and RDA are different in principal. Despite the differences, both techniques share similar working steps involving genomic digestion, adaptor ligation and sub-sample amplification. As for RDA, the initial step in AFLP is the digestion of total genomic DNA with incomplete digestion being detrimental; it results in detection of false positive differences in banding pattern that do not reflect the true DNA polymorphisms (Blears *et al.*, 1998). As previously discussed, RDA relies heavily on the choice of restriction nucleases as this determines the size of sub-population generated and hence the effectiveness of rare target sequence to completely re-associate

during subtractive hybridization. In AFLP, two restriction enzymes, a 6-bases rarer cutter and a 4-bases frequent cutter, are commonly employed for genomic digestion. The majority of researchers use the combination of *Eco*RI (as the rarer cutting) and *Mse*I (as the frequent cutting) enzymes for AFLP analysis. These are also the enzymes included in the commercial AFLP kits from Invitrogen, LICOR and Applied Biosystems companies. The complexity of analysed fragments is further reduced by a two-step amplification strategy that uses primers with selective nucleotides at the 3'-end. It was found that there is an inverse relationship between the number of fragments that are amplified and the number of nucleotides that are added to the primers, due to increased sub-sampling of the molecular population when additional selective bases are added (Vos *et al.*, 1995). Therefore, it is obvious that both restriction endonuclease and primer selection contributes to the effectiveness of AFLP analysis. Indeed, Robinson and Harris (1990) inferred that the choice of restriction enzymes and primers greatly affect both the quality and quantity of data generated.

Meanwhile, it was also noted that both RDA and AFLP utilize adaptor ligation to restricted fragments for fragment amplification. However, adaptors in AFLP are designed in such a way that a base change is introduced into the restriction recognition site, so that the original restriction sites are not restored during ligation, enabling restriction and ligation to be completed in the same tube (Vos *et al.*, 1995). Whereas, restoration of original restriction sites is important for the RDA protocol to enable the adaptors on tester to be changed during each successive rounds of subtractive hybridization, ensuring that primer binding sites are only formed in the tester-tester hybrids (Lisitsyn, 1995).

PCR amplification using *Taq* polymerase is another common key feature in both RDA and AFLP methods. For RDA, the PCR mixture is pre-incubated at 72 ºC for dissociation of unphosphorylated 12-mer oligonucleotide before addition of *Taq* polymerase to fill in the 3'-recessed ends of the ligated fragments. On the other hand, Vos *et al.* (1995) commented that the filling in of the 3'-recessed ends by *Taq* polymerase during the heating step is a matter of only seconds or less or *Taq* polymerase may have even immediately displaced the non-ligated strands at low temperatures during assembly of reaction mixture. Therefore, *Taq* polymerase is directly added to the pre-amplification and selective amplification mixture in AFLP.

In the present study, the use of *dura* and *pisifera* bulks together with an AFLP-based method allowed the detection of polymorphisms which could be linked to the shell-thickness gene. This approach has been previously applied in oil palm research for identification of markers to the *Virescens* (Seng *et al.*, 2007) and the shell-thickness traits (Billotte *et al.*, 2001a, b). Furthermore, AFLP in combination with the BSA method has also been successfully employed in numerous plant species for marker discovery. To name a few, maize (Cai *et al.*, 2003); barley (Altinkut *et al.*, 2003); rice (Liu *et al.*, 2010); wheat (Zhang *et al.*, 2011) and sorghum (Chang *et al.*, 2012). This proves that AFLP coupled with BSA is a very useful and powerful technique for identifying markers that are tightly linked, or co-segregate with, genes underlying monogenic and quantitatively inherited traits.

The single-enzyme AFLP technique was first introduced and commonly used for microorganism studies (Gaafar *et al.*, 2003; Giammanco *et al.*, 2007; Gibson *et al.*, 1998; Valsangiacomo *et al.*, 1995). It is believe that this study was the first to exploit

the effectiveness of single-enzyme AFLP for isolation of genes of interest for plant genome, namely oil palm genome.

The use of single enzyme instead of combination of two restriction endonucleases had raised concern about an insufficient reduction of genome complexity and high background level of amplified fragments. Figure 4.1 shows that combinational *Eco*RI/*Mse*I digestion yielded a profile with fragments of smaller size (100-1000 bp) compared to genomic digestion with a single enzyme *Eco*RI and *Mse*I alone, which produced larger bands with size ranging from 200 bp to 2 kb and 500 bp to 1.5 kb, respectively. To ensure further reduction of genomic complexity, a total of six selective nucleotides were used in the present study, three selective bases for primers of both ends of fragments. This is consistent with the experiment set up of Vos *et al.* (1995) in which six selective nucleotides were also used for the larger genomes of maize (2500 Mbases) and human (3400 Mbases) as compared to oil palm genome size of 1800 Mbases. Blears *et al.* (1998) suggested that 1-2 selective nucleotides on 3'-end of each primer may be sufficient for small genomes of 10^6-10^7 base pairs (bp) while more complex genome of 10^8-10^9 bp will require additional selective nucleotides to reveal polymorphism.

In this study, a single-base extension at the 3'-end of primers was used during pre-selective amplification followed by two additional selective nucleotides for selective PCR amplification. All 16 possible combinations derivable from four nucleotides were exploited. Mba and Tohme (2005) estimated that a 256-fold complexity reduction can be achieved through selective annealing of the PCR primer to only the subset of restricted fragments that carry the specific three selective nucleotides. Four-base extensions were not recommended as Vos *et al.* (1995) had demonstrated that

164

the tolerance of mismatches can occur during amplification of primers with 4-base extensions, indicating a loss of selectivity. Therefore, only primers with at most three selective nucleotides were used in this study.

Single-enzyme AFLP analysis on typing and epidemiological studies of *Legionella pneumphila* (Valsangiacomo *et al.*, 1995) found that the choice of suitable restriction enzyme is crucial to generate informative profiles with a reasonable number of polymorphic bands. *Pst*I enzyme (GC-rich) was found to be appropriate for typing of *L. pneumophila* but not *Borrelia burgdorferi*; less informative patterns were obtained with *Borrelia* strains due to its low GC content. With the large genome size of oil palm, restriction digestion of genomic DNA with a single enzyme would definitely generate large numbers of potentially polymorphic bands. Therefore, five different enzymes were exploited in the present study. *Bam*HI and *Hin*dIII enzymes were employed in previous RDA analysis (chapter 3) while *Pst*I enzyme was used for genome complexity reduction in development of oil palm DArTSeq markers as reported in Chapter 5. These DArTSeq markers were then employed to construct the high density genetic linkage maps of the 768 and 769 populations (Chapter 6). Among the five enzymes, it was demonstrated that the majority of the shell-thickness related-polymorphic bands with strong intensity were generated from the analyses using *Eco*RI enzyme (9), follow by combinational *Eco*RI/*Mse*I enzyme (5) while AFLP analysis using single-enzyme *Hin*dIII and *Bam*HI gave rise to considerable numbers of weak signal potential polymorphic bands.

It was found that around 100 polymorphic bands per primers pair per sample were observed in the selective amplification profile of AFLP analysis in the present study (Appendices D1-D8). Typically, AFLPs produce 50-100 fragments (Vos *et al.*,

1995) and around 100-150 bands can be separated on a standard length sequencing gel of 40-50 cm (Ridout and Donini, 1999). This reveals that AFLP analyses in the present study had generated reasonable numbers of polymorphic bands. Nevertheless, research had shown that organisms with large amounts of repetitive DNA and retrotransposons frequently give rise to profiles with many low-intensity peaks that are difficult to score (Kardolus *et al.*, 1998; Fay *et al.*, 2005). This could explain the appearance of quite a number of less intense bands in the current set of gel profiles as oil palm genome, like most plant genome, consists of large numbers of repetitive DNA sequences (Castilho *et al.*, 2000).

One of the major concerns for single-enzyme AFLP analysis is the formation of an inverted repeat at the ends with base-pairing of the ends of the fragments forming a stem-loop structure that competes with primer annealing (Vos *et al.*, 1995). The use of a single primer could also cause the occurrence of "doublets" on the gels due to unequal mobility of the two strands of the amplified fragments which can be observed in high resolution analysis systems (Vos *et al.*, 1995). For the present study, these two problems should not be neglected although reasonable numbers of polymorphic bands had been generated and differences in banding pattern between *dura* and *pisifera* bulks were successfully identified.

In the present study, AFLP method was deem promising in identifying shell-thickness related polymorphic bands, further recovery and molecular cloning of the candidate bands are necessary to identify the genetic markers as the sequence content of the AFLP fragments were unknown throughout the process. Various studies have shown that polymorphic AFLP detected were mostly non-coding fragments closely

linked with the gene rather than inside the gene sequence itself (Butlin, 2010; Minder and Widmer A, 2008; Paris and Despres, 2012).

On the other hand, genetic linkage mapping of the candidate AFLP specific primer pairs would identify a map location for the AFLP markers; allow better understanding of the relative distance of the markers to the trait of interest, the shell-thickness gene. AFLP is commonly used in the construction of high density genetic linkage maps and in position cloning of gene of interest because of its versatility (Blears *et al.*, 1998). AFLP markers are often complementary to other molecular marker techniques and in some cases, the resulting linkage maps were better resolved (Meudt and Clarke, 2007). AFLP has been generally employed for genetic linkage mapping in oil palm. Besides the genetic mapping of CIRAD golden cross by Billotte *et al.* (2005), AFLP, together with RFLP and microsatellites marker systems, was used for the construction of genetic map for FELDA high yielding DA41 cross (Seng *et al.*, 2011) and interspecific *E. guineensis* and *E. oleifera* cross (Singh *et al.*, 2009). The latter enabled identification of 11 QTLs for iodine value and six different components of fatty acid composition that control oil quality.

Meanwhile, together with DArT and SSR markers, AFLP markers have also been employed to construct the genetic linkage mapping of other plant species, for such as wheat (Semagn *et al.*, 2006), triticale (Tyrka *et al.*, 2011) and ryegrass (Julie et *al.*, 2013). In the present study, selective AFLP primer pairs that were identified would be useful for genotyping of closely-related populations of the 768 and 769 followed by saturation of the genetic linkage maps constructed using DArTSeq and SSR markers, as reported in hapter 6, to facilitate the identification of the shell-thickness marker(s) and

possibly characterization of the shell gene as well as QTL study of important agronomic traits. However time was lacking for this part of work.

Although AFLP is more robust than RFLP and RAPD, the technique requires technical skills and DNA of high quality. The quality of the extracted DNA and the method of extraction could affect the profiles obtained (Jones *et al.*, 1997, Benjack *et al.*, 2006). Benjack *et al.* (2006) and Mikulášková *et al.* (2012) reported that commercial DNA extraction kits can give better quality DNA than some other methods. DNAs used in the present study were extracted conventionally using modified CTAB method. It is believed that contaminants during purification of DNA such as chloroform, ethanol and EDTA will interfere with the performance of restriction endonucleases (Fuchs and Blakesley, 1983) and this is detrimental as fragments generated from incomplete restriction of genomic DNA may be misinterpreted as false polymorphism (Blears *et al.*, 1998). Therefore it is vital to use DNA of high purity with OD measurement of 260/280 ratio of 1.8-2, regardless of the extraction methods.

Genotyping errors are another issue that is often neglected. In comparing four different case studies of population genetic study, Bonin *et al.* (2004) estimated a 2.6% of genotyping error for AFLP loci from a study of *Betula nana* and it was also found that human factors were non-negligible error generators. Therefore the author suggested that systematic pilot study should be performed before any extensive investigation, to provide opportunity to acquire experience with the technique and to achieve reproducibility. Pompanon *et al.* (2005) classified the main cause of genotyping errors into four groups, which are variation in DNA sequence, sample quality, biochemical artefacts and equipment, and lastly human factor. The use of appropriate number of positive and negative controls, 5 to 10% replication of samples as well as experience

and rigor in the laboratory work are necessary to maintain the consistency of profiling and reduce the genotyping error (Pompanon *et al.*, 2005). Therefore for the current AFLP analysis, it will be beneficial to have a replicate study together with negative controls for the assessment of the reproducibility of the technique and error estimation.

In conclusion, the use of single-enzyme and conventional *Eco*RI/*Mse*I AFLP analyses identified 29 primer pairs that yielded 49 polymorphic bands with good intensity between *dura* and *pisifera* bulks. AFLP technique was deemed promising for the identification of marker(s) closely linked with the shell-thickness gene through saturation of the constructed high density DArT- and SNP-based genetic linkage maps. However, due to the time constraints, this part of work was not pursued. Future studies should involve an extensive AFLP analysis with appropriate controls to reduce genotyping errors and acquire experience with the experimental techniques.

Chapter 5

Development and characterisation

of DArTSeq and SSR markers for

genetic linkage mapping

5.1   Introduction and objective

All organisms are subjected to mutation as a result of normal cellular operations or interactions with the environment, leading to genetic variations (polymorphism) between and within species. Molecular marker technology can be utilised to reveal these naturally occurring polymorphisms (Nguyen and Wu, 2005). A molecular marker is defined as a particular region of DNA that reveals differences at the genome level (Agarwal *et al.*, 2008).

The development of molecular markers has revolutionized plant genetic research. Molecular markers are commonly used in plant genetic analyses, such as assessment of genetic diversity, fingerprinting of varieties, linkage map construction, QTL mapping for desirable traits and marker-assisted selection (Collard *et al.*, 2005; Semagn *et al.*, 2005).

Microsatellites or SSRs are one of the commonly used molecular markers in plant studies. Microsatellites are regions of DNA that consist of short tandem repeating nucleotide units that can be found throughout the genomes of eukaryotic species (Powell *et al.*, 1996). Because of their high reproducibility, multi-allelism and co-dominant inheritance, SSRs are the marker of choice for plant genetics and breeding applications (Gupta and Varshney, 2000). Billotte *et al.* (2001) reported the development of the first set of oil palm microsatellite markers. The authors used these markers to construct a high density linkage map and QTL analysis in oil palm (Billotte *et al.*, 2005; Billotte *et al.*, 2010).

A single nucleotide polymorphism (SNP) is a single-point mutation in the DNA in which one nucleotide at a particular locus is substituted with another one. In recent years,

development and use of SNPs in plant genetics and breeding has gained popularity compared to SSRs. SNPs are highly abundant in genomes, amenable to high-throughput screening, co-dominant and usually bi-allelic (Kahl *et al.*, 2005). Development of Next Generation Sequencing (NGS) technologies has also catalysed the development of SNP even in crops with little or no sequence information (Varshney *et al.*, 2009).

Classical Diversity Array Technology (DArT) is a microarray-based marker system utilising genome complexity reduction to simultaneous type several hundreds to thousands of loci in a single assay. DArT markers enable sequence-independent and cost-effective whole-genome profiling (Jaccoud *et al.*, 2001). This technique has been successfully applied for various studies and diversity arrays are currently available for over 120 different plant species, including oil palm (www.diversityarray.com). DArT "Genotyping-by-sequencing" (DArTSeq) is a new marker platform in which the DArT complexity reduction approach is coupled with Illumina short read sequencing to generate dominant DArT markers and co-dominant SNP markers (Sansaloni *et al.*, 2011).

The aim of the research reported in this chapter is to develop and characterise DArTSeq (both DArT and SNP) markers as well as characterise publicly available SSR markers using two closely related *tenera* self-pollinated oil palm crosses, namely 768 and 769. These markers were used in the construction of genetic linkage maps (See chapter 6) to identify markers closely linked to economically important traits, particularly the shell-thickness trait (See chapter 7 and 8). This chapter reports the first attempt to employ the DArTSeq platform to genotype oil palm populations. Publicly available SSR markers

reported in Billotte *et al.* (2005, 2010) were also screened for their polymorphism in the present study.

5.2   Materials and Methods

5.2.1   Plant Materials

Two populations, 768 and 769, from the AAR oil palm breeding programme were selected to generate DArTSeq markers and characterise SSR markers in the present study. The progenies of the 768 and 769 populations are derived from self-pollination of *tenera* palm 228/05 and 228/06, respectively. Both 228/05 and 228/06 are full-sibs from the same *tenera* x *pisifera* cross of Binga x Yangambi AVROS origin. A total of 48 and 58 offspring from the 768 and 769 controlled crosses, respectively, together with their *tenera* parents were available and used for screening of SSR markers and development of DArTSeq markers reported in this chapter and subsequent genetic mapping and QTL analysis reported in chapters 6 and 7. The previously identified outliers of the 768 and 769 controlled crosses (Chapter 3), 768/28 and 769/36, were excluded from markers development and characterisation.

The mapping populations are planted at AAR breeding research station in Paloh Estate, Johore, Malaysia. Sampling of frond one leaves from the progeny palms was carried out in October 2011. The leaves were cleaned with 70% EtOH, cut into small pieces, packed and stored at -80 °C.

5.2.2 **Extraction of genomic DNA and quality check**

Genomic DNA was extracted from freeze-dried leaves using NucleoSpin® Plant II kit according to manufacturer's instruction (Macherey-Nagel, Germany) which could be divided to four basic steps, lysis, binding of DNA samples to column, washing and elution of DNA. The quality and quantity of DNA was determined by agarose gel visualisation under UV light. Five microlitres of eluted DNA was mixed with 5 µL 6x loading dye and loaded onto a 1% agarose gel in 0.5x TBE buffer. At the same time, a series of known concentrations of uncut Lambda bacteriophage (50-500 ng) and 2-log DNA ladder (New England Biolabs) were also loaded into the same gel. The gel was run at 100 V for 60 min and visualised under UV light. Quantification of DNA was achieved by comparing the band intensity of eluted DNAs with those of the lambda DNA standards.

The integrity and purity of DNA samples were checked by incubating 1 µL of genomic DNA in a total of 5 µL of *Eco*RI restriction buffer with and without *Eco*RI restriction enzyme at 37 °C for 2 hours. The DNA was then visualised by 1% agarose gel electrophoresis. DNA was stored at -20°C.

5.2.3 **Development and characterisation of DArT and SNP markers from the DArTSeq platform**

A total of 106 progenies from the 768 and 769 controlled crosses together with their *tenera* parents were sent to Diversity Arrays Technology Pty Ltd in Yarralumla, Australia, for the genotyping service with DArTSeq platform. Twenty microlitres of

DNA with concentration of about 50 to 100 ng/µL were pipetted into a fully skirted 96-well plates. The plates were capped by strips of eight caps and sealed with parafilm prior to shipment for DArT services.

A detailed account of DNA genotyping using DArTSeq platform has been reported earlier by Sansaloni *et al.* (2011) and Cruz *et al.* (2013). In brief, the procedure involves generation of genomic representations of individual samples using restriction enzymes combinations that involve *Pst*I. A *Pst*I-RE site specific adaptor is tagged with 96 different barcodes enabling a plate of encoded DNA samples to run within a single lane on an Illumina Genome Analyzer IIx. A sequencing primer is included in the *Pst*I adaptor so that the tags generated are always reading into the genomic fragments from the *Pst*I sites. After the sequencing run, the FASTQ files are quality filtered using a threshold of 90% confidence for at least half of the bases and with more stringent filtering for the barcode sequences. The filtered data are then split into their respective targets (genotypes) using a barcode splitting script. After producing various QC statistics and trimming of the barcode, the sequences are aligned against the reference created from the tags identified in the sequence reads generated from all the samples. The output files from the alignment generated using the Bowtie software are processed using an in-house analytical pipeline to produce a "DArT score" (presence/absence) and "SNP" tables. Additionally, several parameters are computed by DArTsoft for evaluating the quality of markers, for example parameter call rates (percentage of genotypes able to be called) and Q (a quality score that measures signal to noise ratio).

Upon receiving the DArT score and SNP tables from DArT Pty Ltd, the percentage of missing data and allele ratio of DArTSeq markers (DArT and SNP) in both the 768 and 769 populations were calculated. The rate of missing data is the ratio of individuals with missing data to the total number of individuals in the population, while the allele ratio was calculated as the segregation ratio of individual alleles in the population. Subsequently, DArT and SNP markers were selected for mapping work with the following criteria: firstly, markers with less than or equal to 5% missing data were selected; secondly, selection of DArT and SNP markers with allele ratio of 0.15-0.85. Lastly, genotyping data of *tenera* parents was used as a quality control in which inconsistent results between expected segregation patterns based on the parental scores and the observed population scores were eliminated from the dataset.

5.2.4 Characterisation of SSR markers

The present study aimed to identify at least two polymorphic SSR markers, available at the CIRAD public database, from both ends of the oil palm chromosomes, using the 768 and 769 controlled crosses. These SSR markers were used as anchor loci in the construction of genetic maps reported in chapter 6. The latest genetic map published by Billotte *et al.* (2010) served as a reference for the location of SSR markers while the primer sequences of the markers were retrieved from http://www.neiker.net/link2palm/OilP/for1-6a.htm.

A three primer labelling system was adopted for SSR genotyping of the current mapping populations (Schuelke, 2000). All forward SSR primers were designed with a M13 sequence (5'-CACGACGTTGTAAAACGAC-3') added to the 5'-end giving rise to

the 'Tagged-Forward" primer. In each PCR reaction, the amount of M13-tagged forward primer would be roughly 1/10th of the reverse primer with the remaining 9/10th of the forward primer being a fluorescently-labelled M13 sequence primer. This allows incorporation of fluorescent dye into final PCR products when the locus-specific M13-tagged forward primer is exhausted and the dye-labelled M13 primer takes over during PCR reaction. All the M13-tagged forward primers and reverse primers were synthesized by MWG Eurofins while the blue, green or black dye-labelled M13 primers were from WellRED primers, Sigma. Extracted genomic DNAs were diluted to 10 ng/μL for PCR reactions.

5.2.4.1   Optimization of primer annealing temperature by gradient PCR

The three primer amplification system necessitates finding the optimal temperature which favours amplification with all three primers. A range of annealing temperatures was tested using gradient PCR. DNA of all individual palms was mixed in equal amounts to form the template DNA for the gradient PCR.

The PCR reaction was set up in 96-well plates (Thermo Scientific) by mixing 20 ng of temperature DNA, 4 mM dNTPs mix, 0.4 μM M13-tagged forward primer, 4 μM reverse primer, 1x dye-labelled M13 primer, 0.1 μl of *taq* polymerase and 2 μl of 10x PCR buffer in a total volume of 20 μl. The plate was sealed with Thermowell® sealing mat (Fisher Scientific) and briefly centrifuged to bring down the contents. PCR reaction was performed with an initial denaturation of 94 °C for 3 mins, followed by 35 cycles of denaturation at 94 °C for 1 min, 6 different annealing temperatures of 50, 53, 56, 59, 62

and 65 °C for 1 min, extension at 72 °C for 2 min, and final extension of 72 °C for 10 min.

Five microlitres of 6x loading buffer were added into each PCR reaction and 10 μl of the sample was then analysed on a 2% agarose gel alongside a 2-log ladder. The optimal annealing temperature would be expected to show a single specific band of expected size with the strongest band intensity and little background.

5.2.4.2   Screening of polymorphic SSR markers

After optimizing the annealing temperature of each primer set, SSR markers were screened for their polymorphism using the *tenera* parents of both the 768 and 769 controlled crosses, 228/05 and 228/06. PCR reactions were prepared with each tube containing 20 ng of template DNA, 0.4 μM M13-tagged forward primer, 4 μM reverse primer, 1x dye-labelled M13 primer, *taq* polymerase in 1x PCR buffer with total volume of 20 μl. For the same primer pair, the reactions involving the DNA of 228/05 and 228/06 used different coloured dyes, so to allow both products to be run in the same capillary size evaluation, and also to allow any coincidence of allele sizes between the genotypes to be resolved. Amplifications were carried out with the following programme: 94 °C for 3 mins, 35 cycles of 94 °C for 1 min, selected annealing temperatures for 1 min and 72 °C for 2 min, and final extension of 72 °C for 10 min. The PCR products were then checked on a 2% agarose gel before running on the capillary sequencer.

For the same primer pair, both PCR reactions of 228/05 and 228/06 were pooled together for fragment size analysis using a Beckman CEQ 8000 Genetic Analyzer

(Beckman coulter inc, USA). The blue dye gives a stronger signal than the green dye on the CEQ machine. Therefore, a larger volume of green dye-labelled PCR products were added to the pool, normally 5 μl green-labelled products were mixed with 2 μl blue-labelled PCR products.

Prior to fragment size analysis, sample loading solution (SLS) and size standard (SS) (Beckman Coulter Inc, Fullerton, USA) were mixed in the ratio of 1:100 (v/v). Four microlitres of pooled PCR products were loaded into a new PCR plates, mixed with 25 μl of SLS:SS mix and covered with a drop of mineral oil. The mixture was then loaded into a Beckman CEQ 8000 Genetic Analyzer, voltage was applied for gel electrophoresis and the samples were analysed. The fragments sizes of genotypes can be manually scored using the CEQ$_{TM}$ 8000 Fragments Analysis Software Version 8 to calibrate for size.

Figure 5.1 illustrates an example of an amplification profile produced using the Fragment Analysis Software. The software converts the banding pattern into a plot with the height of peaks corresponding to the intensity of each band. The position of the peak along the x-axis corresponds to the size of the band. One colour is used for the size standard to calibrate the band positions of the microsatellite amplification product. Here red is used for the size standard. SSR primer pairs are considered to be polymorphic when both parents were found to have at least two alleles.
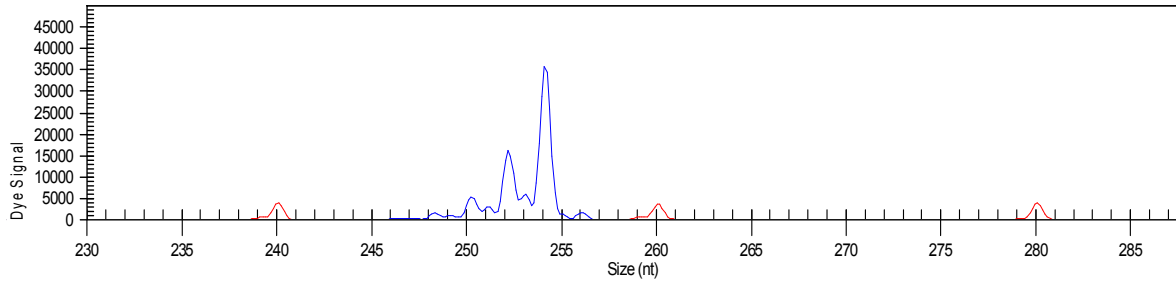
Figure 5.1: Examples of a fragment amplification profile analysed using the CEQ$_{TM}$ 8000 Fragments Analysis Software Version 8.

5.3.4.3 Genotyping of the mapping populations

All individuals in the mapping population were genotyped with the selected polymorphic SSR markers with the parents serving as positive control. The procedures of PCR and fragment size analysis were performed in the same way as mentioned in section 5.3.4.2.

To save cost, a pool of four different SSR marker PCR products were mixed to send for multiplexed fragment size analysis using the CEQ fragment analyzer. When the size differences between different marker-alleles of the same pool were sufficient to be distinguished (at least 30 bp apart, based on parental alleles), all PCR reactions were labelled with the strong blue dye. Visualised band intensities during agarose gel electrophoresis were used to determine the amount of products to be added to the pool before capillary analysis. Products with stronger amplification would contribute less to the pool while weak amplification products would contribute more to the pool to have an overall balance of intensity of signal during fragment analysis. For primer pairs in the same pool with products of similar size (<30 bp), one of the PCR products would be

labelled with the green dye instead and more green dye-labelled products would be loaded for fragment analysis due to the weak signal of green dye.

5.3   Results

5.3.1   Quality checking of the extracted genomic DNA

Prior to the shipment of DNA samples for DArT genotyping service, quality and integrity check of the kit-extracted genomic DNA samples was performed using restriction endonuclease digestion reaction, to ensure that inhibitors were not present, as complete digestion is essential to provide robust differences between genotypes. Figure 5.2 illustrates the digestion profile of genomic DNA without (a, control) and with (b) restriction endonuclease *Eco*RI. Intact high molecular weight genomic DNA were observed for all DNA samples that were not subjected to restriction enzyme digestion while smearing of DNAs were observed after digestion with the restriction enzyme, suggesting complete digestion of genomic DNA. This demonstrates that DNA samples were intact and were likely to be of good enough quality for subsequent DArT analysis.
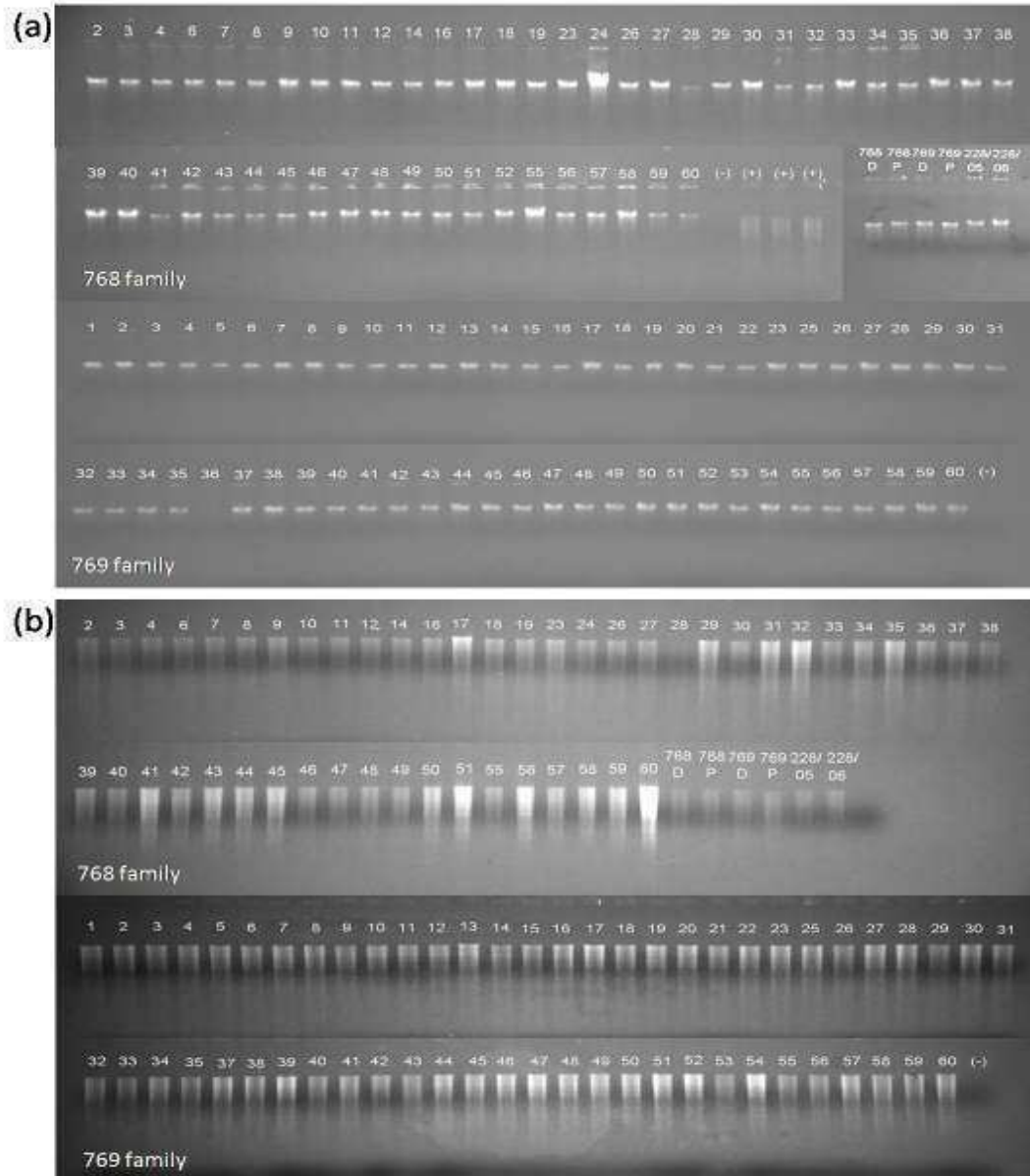
Figure 5.2: Digestion profiles of genomic DNA samples from the 768 and 769 controlled crosses (a) without and (b) with the addition of restriction endonuclease *Eco*RI. (+), positive control with addition of the *Eco*RI enzyme; **(-)**, negative control without *Eco*RI enzyme.

5.3.2   **Characterisation of DArT and SNP markers from the DArTSeq**
               platform

A total of 11,675 DArTSeq markers, constituting of 6,764 DArT and 4,911 SNP, were generated from genotyping of the 768 and 769 mapping populations. DArT markers are dominant and were scored as either 1 (present) or 0 (absent). SNP markers are biallelic with scoring for both alleles generated, thus homozygosity and heterozygosity of individual samples could be distinguished using SNP markers.

The quality score (Q) for DArT markers ranged from 1.50-38.65 with an average of 5.35 whereas call rate was in the range of 0.73-1 with a mean of 0.90. Better call rate were attained by SNP markers, ranging from 0.75-1 with a mean of 0.95. The call rate of the *tenera* parents was 82.4% and 98.4% for DArT and SNP markers, respectively. The Q score is a direct measure of the quality of genotyping while the call rate essentially reflects the percentage of missing data tolerated. The DArT and SNP markers generated were generally of good quality and high polymorphism.

Initial analysis discovered that four and one samples from the 768 and 769 controlled crosses, respectively, were distinct from their cross and parents. These samples, namely 768/26, 768/27, 768/29, 768/30 and 769/18, were discarded as illegitimate samples. These atypical genotyping results were consistent between both DArT and SNP fingerprinting. Therefore these five samples were removed from the subsequent analysis, marker selections as well as genetic mapping and QTL analysis.

5.3.2.1   Genotyping using DArT markers

Figure 5.3 illustrates the percentage of missing data and allele ratio of DArT markers genotyped. It was noticed that 5,907 DArT markers from the 769 controlled cross (87%) had no missing data compared to only 1,537 in the 768 controlled cross (22.7%). The majority of DArT markers obtained from the 768 controlled cross had a missing data rate less than 0.2. Closer inspection has revealed that 769/7, 768/40 and 769/13 are the samples with the highest number of missing data, more than 15% of the total DArT markers genotyped. These three samples had 15.1%, 15.7%, and 17.7% of missing data, respectively.

DArT markers were regarded as monomorphic when all the individuals in the population were scored either 1 or 0. Out of the 6,764 DArT markers, a total of 2,423 (35.80%) and 2,314 (34%) monomorphic DArT markers were present in the 768 and 769 controlled crosses, respectively, but polymorphic between crosses or the five outcrosses identified. This is due to the software pipeline looking for polymorphism across the entire set of samples, which includes two (related) crosses and outcrosses.

Due to the diversity of DArT markers in terms of percentage of missing data and allele ratio, a subset of relatively good DArT markers were selected with the stringent criteria of less than 5% missing data and allele ratio of 0.15-0.85. In total, 948 and 958 DArT markers were chosen for the 768 and 769 controlled crosses, respectively, for subsequent linkage mapping analysis.
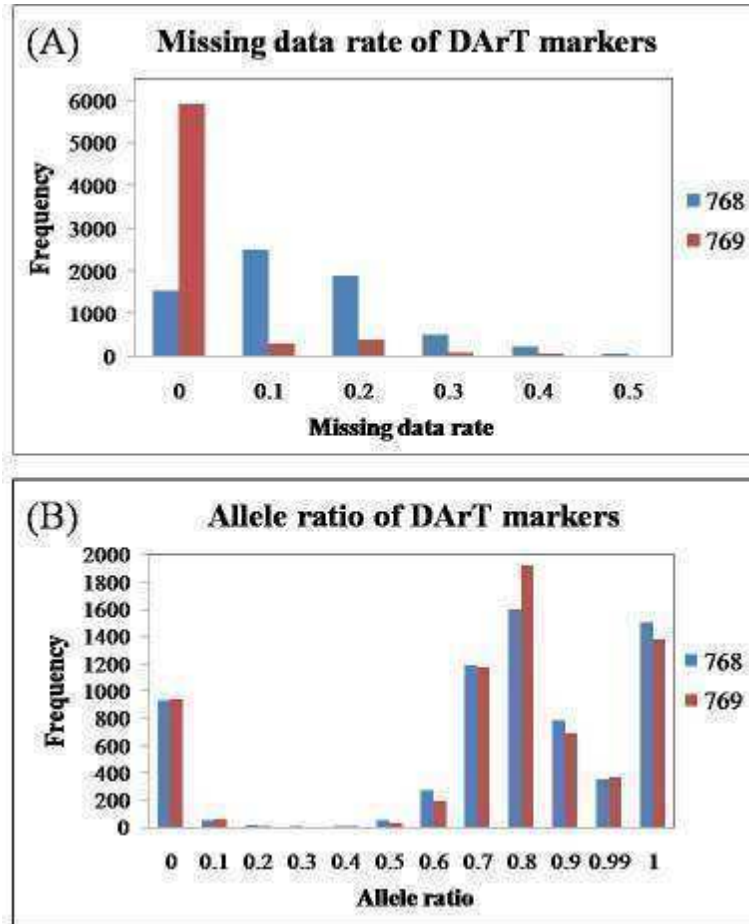
Figure 5.3: Missing data rate and allele ratio of DArT markers for the 768 and 769 populations. Missing data rates were calculated as the number of individuals with missing data over total number of individuals in the population. Allele ratio were calculated as the average ratio of presence:absence of bands in the population.

5.3.2.2   Genotyping using SNP markers

Both populations showed similar trends in the rate of missing data and allele ratio of the SNP alleles genotyped (Figure 5.4). A large proportion of the SNP alleles from both the 768 and 769 controlled crosses were found to be free of missing data (68.6% and 64.2%, respectively). None of the individual genotypes had more than 15% missing data. Samples with the greatest numbers of missing data were 769/29 and 769/7 with percentages of 12.8 (626/4911) and 12.7 (621/4911), respectively.

SNP markers are bi-allelic where scoring of both alleles is possible for all SNP markers. Studies on the segregation patterns of SNP alleles revealed that around 40% of SNP alleles have the same segregation patterns across all the individuals of the same population, being either present or absent. The other alleles of the same SNP marker could have various segregation patterns; however these highly distorted SNP markers are not suitable for mapping. It was found that the 768 and 769 controlled crosses have as high as 1,538 and 1,537 monomorphic SNP markers, respectively, which is slightly greater than 30% of the total SNP markers genotyped.

As for the DArT markers, only a proportion of relatively good SNP markers were selected for further linkage mapping and QTL analysis. The selection criteria were set as follow: less than or equal to 5% missing data and allele ratio of 0.15-0.85 for both alleles of the SNP markers. The parental genotyping data also served as a quality control with the parental palm expected to be heterozygous for any segregating SNP marker. If this was not the case, then the marker was eliminated. Eventually, 719 and 729 SNP markers were selected for the 768 and 769 populations, respectively, for the construction of genetic linkage maps.

As a result, a total of 1,667 and 1,687 markers were selected from DArTSeq platform for map construction of the 768 and 769 populations, respectively.
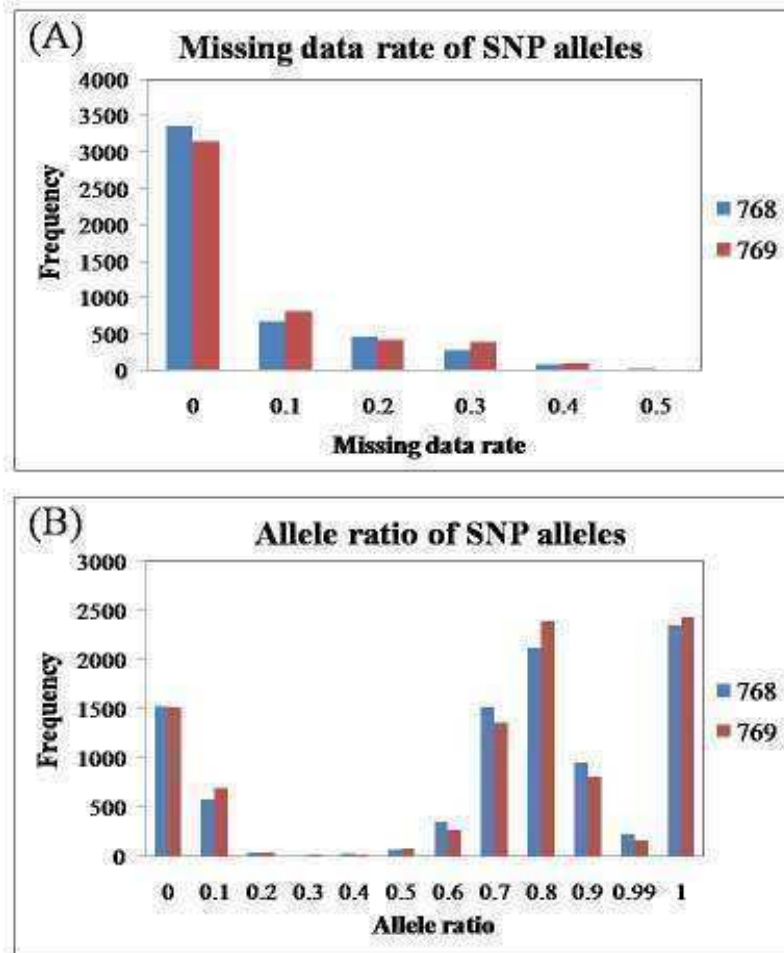
Figure 5.4: Missing data rate and allele ratio of SNP alleles for the 768 and 769 populations. Missing data was calculated as the number of individuals with missing data over total number of individuals in the population. Allele ratio was calculated as the average ratio of presence:absence of an allele in the population.

5.3.3   **Characterisation of SSR markers**

5.3.3.1   **Determination of optimal annealing temperature using gradient PCR**

Oil palm SSR markers developed by CIRAD (Billotte *et al.*, 2005, 2010) were selected to screen for polymorphism in the 768 and 769 mapping populations. In view of amplification using three primers, gradient PCR was performed to determine the optimal annealing temperature for each SSR primer pair. Figure 5.5 illustrates examples of

gradient PCR products of six different primer pairs. Vague or no amplification was observed for mEgCIR3747 and mEgCIR2029 SSR primer pairs while the mEgCIR3383 primer pair displayed multiple bands. These three primers were deemed not suitable for progeny genotyping. As for the mEgCIR3358, mEgCIR2600 and mEgCIR0555 primers, clear single bands were observed, although the amplification of mEgCIR3358 was relatively weaker than the other two. The highest temperature of good band amplification was chosen as the optimal annealing temperature. Therefore the following temperatures, 53, 59 and 56 °C, were chosen as optimal annealing temperature for mEgCIR3358, mEgCIR2600 and mEgCIR0555 primer pairs, respectively.
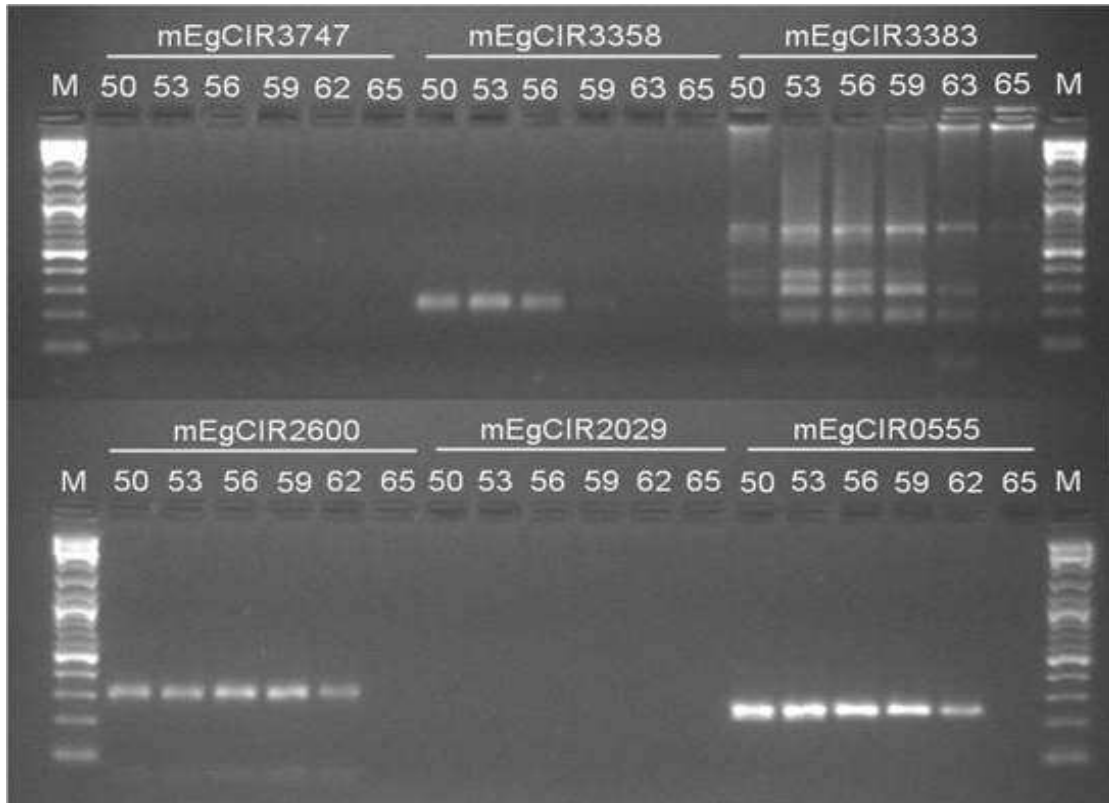
Figure 5.5: Example of the gel electrophoresis profiles of gradient PCR for six SSR markers, using the three primer labelling approach. Gradient PCR was performed using six different annealing temperatures, 50, 53, 56, 59, 62 and 65 °C. **M**, 2-log DNA ladder (New England Biolabs).

5.3.3.2   Determination of polymorphism using parental genotypes

Following the optimization of annealing temperatures, the selected primer pairs were screened against *tenera* self-pollinated parents of both controlled crosses, 228/05 and 228/06. An example of an agarose gel electrophoresis of the amplification of parental material using SSR markers is shown in Figure 5.6. Clear single bands with good product yield for both parents indicate that amplification was successful and the PCR product could be further analysed using the CEQ 8000 Genetic Analyzer. For the primer mEgCIR3213, faint bands were observed suggesting weak PCR amplification. For those primers producing faint bands, the PCR reaction was repeated with increased DNA and agarose gel electrophoresis was used to confirm amplification before proceeding for fragment analysis.
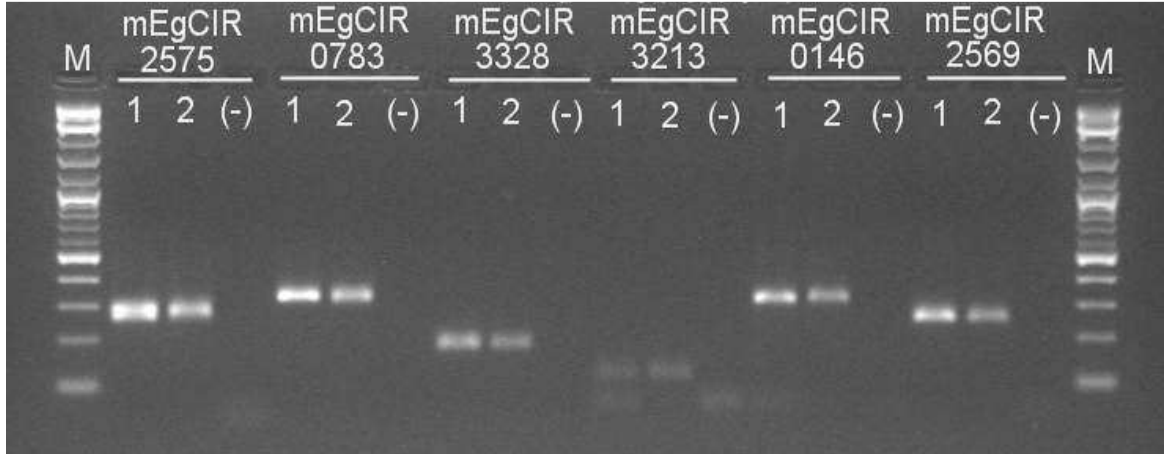
Figure 5.6: Electrophoresis profiles of SSR amplification of the 228/05 and 228/06 parents. 1, 228/05; 2, 228/06; (-), negative control. **M**, 2-log DNA ladder (New England Biolabs).

Figure 5.7 illustrates five examples of fragment profiles amplified from 228/05 and 228/06 using CEQ$_{TM}$ 8000 Fragments Analysis Software Version 8. A single peak was observed for both parents for mEgCIR0555 primer pairs [Figure 5.7 (a)], this primer

product is concluded to be monomorphic in each parents (although polymorphic between

parents) while screening of primer mEgCIR3590 showed that 228/06 is homozygous but

228/05 is heterozygous [Figure 5.7 (b)]. Meanwhile, the mEgCIR3477 primer was shown

to be polymorphic for both *tenera* parents with clear display of two peaks, 2 alleles, with

size of 244 and 261 bp, respectively. Both samples share the same alleles for this SSR

primer pairs. The peaks with its shutter bands were shown to be evenly spaced with

decreasing height to the left of the peak and larger alleles show slightly shorter peak than

smaller alleles.

Both mEgCIR3809 and mEgCIR2215 primers were also polymorphic in both

parents but with different fragment profiles [Figure 5.7 (c) and (d)]. For the mEgCIR3809

primer, parent 228/05 had two alleles of 115 and 122 bp in size while the two alleles of

228/06 were 121 and 133 bp. Primer mEgCIR2215 exhibits a complex profile [Figure 5.7

(e)]. Both samples were heterozygous, but the two alleles were only one repeat unit

different in size. Overlapping of the second stutter peak of the larger allele with the

smaller allele increased the intensity of smaller allele.
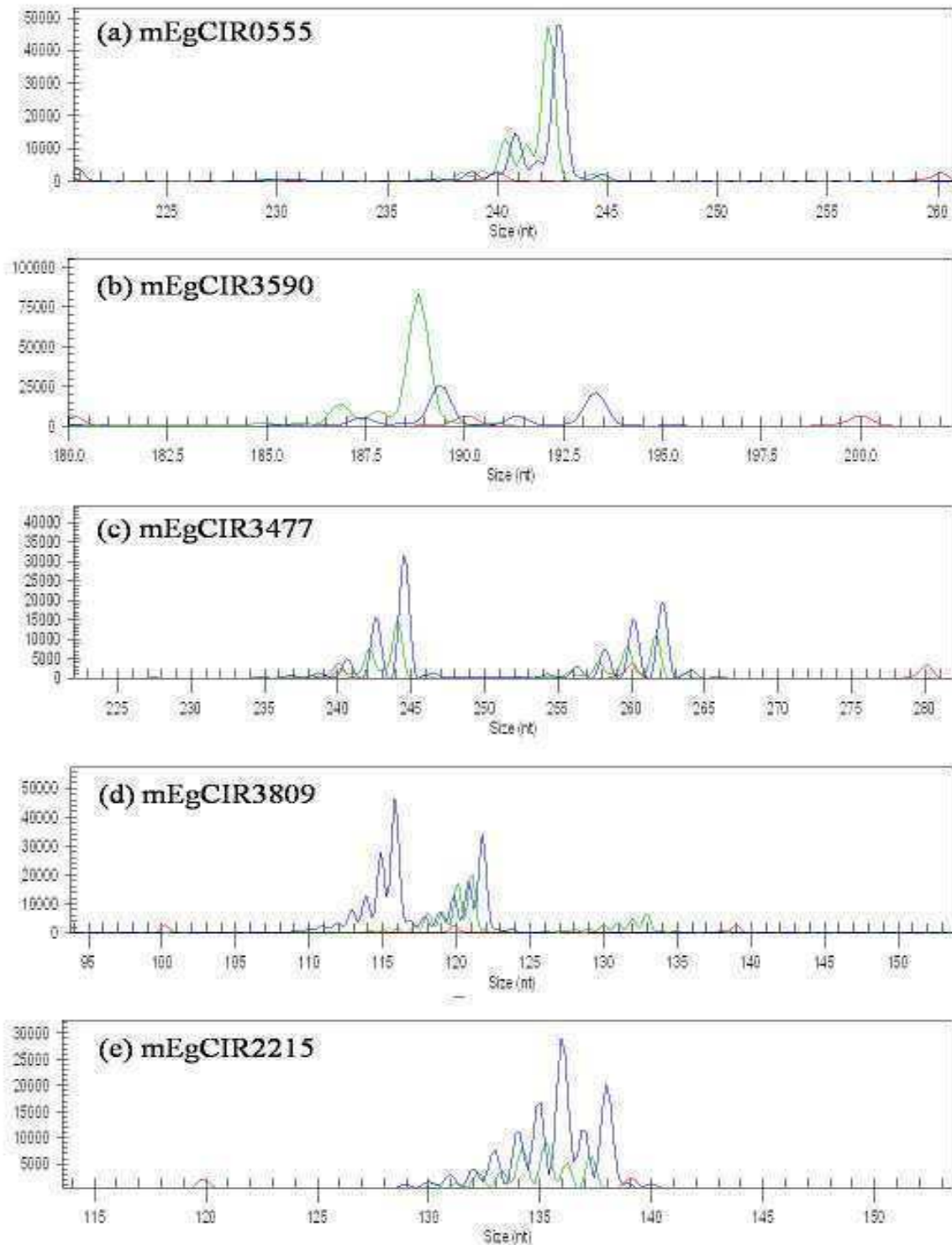
Figure 5.7: Examples of the fragment analysis profiles of polymorphism screening of SSR markers using two oil palm parental materials. Blue colour peaks indicate 228/05 and green colour peaks indicate 228/06. Red colour is the size standard.

Initially, 64 oil palm SSR markers were screened for their amplification and polymorphism. More markers were screened subsequently to ensure at least one polymorphic SSR marker was identified for each end of the linkage groups for both populations. A total of 102 markers were eventually screened and 36 polymorphic markers were identified. This suggests a 35% polymorphism level for CIRAD SSR markers on current set of mapping populations from AAR breeding programme. Note that no polymorphic markers were identified for LG 5, 13 and one terminal end of LG 4 where the *Sh* gene is located despite screening of all the available SSR markers on those particular chromosomes (Billotte *et al.*, 2010). Table 5.1 presents the selected polymorphic markers with their forward and reserve primer sequences, linkage group according to previously published genetic maps by Billotte *et al.* (2010), optimal annealing temperature and allele size in both the 768 and 769 populations.

Table 5.1: Selected polymorphic SSR markers for genotyping of the 768 and 769 populations.

No.	Primer Name	Primer Sequence (5'-3')	LG	Tm (°C)	Allele size 228/05 (768)	Allele size 228/06 (769)
1	mEgCIR3788	F: M13- TTGTATGACCAAAGACAGC R: AGCGCAACATCAGACTA	1	56	182/184	182/184
2	mEgCIR3809	F: M13-CCTTGCATTCCACTATT R: AGTTCTCAAGCCTCACA	1	53	115/122	121/133
3	mEgCIR3392	F: M13-AGCAAGGGAGAAAGATG R: CGAGCAATCAACCTGACTA	1	56	251/280	251/281
4	mEgCIR2215	F: M13-GAACTTGGCGTGTAACT R: TGGTAGGTCTATTTGAGAGT	2	56	136/138	136/138*
5	mEgCIR0793	F: M13-GTACTTCGCAACTATTCCTTTTCTT R: AGTTGATCGTGGTGCCTGAC	2	56	170/176	172/176
6	mEgCIR2575	F: M13-GGGACTTCGCAAACTGTAGCA R: CGGTGGCGTATGGTGGATT	2	62	259/274*	271/274
7	mEgCIR3649	F:M13-TTTAGAGGACAAGGAGATAAG R: CGACCGTGTCAAGAGTG	2	62	306/311	307/311
8	mEgCIR3683	F: M13-GTAGCTTGAACCTGAAA R:AGAACCACCGGAGTTAC	2	56	157/161	157/161
9	mEgCIR2518	F: M13-GATCCCAATGGTAAAGACT R: AAGCCTCAAAAGAAGACC	3	53	291/303	291/293
10	mEgCIR3301	F: M13-GCACTTGGTGGTTATGA R: AGCTGCTGATGGATATC	3	50	148/161/ 230/244	148/156/ 230/238
11	mEgCIR3477	F: M13-CCTTCAAGCAAAGATACC R: GGCACCAAACACAGTAA	4	56	244/262	244/262
12	mEgCIR3526	F: M13-GGGAGAGGAAAAAATAGAG R: CCTCCCTGAGACTGAGAAG	4	56	227/240*	225/227/ 240/246
13	mEgCIR0783	F: M13-GAATGTGGCTGTAAATGCTGAGTG	6	62	322/324	314/324

		R: AAGCCGCATGGACAACTCTAGTAA				
14	mEgCIR3358	F: M13-CCAAGGAACAACATAGA	6	53	235/245	214/235
		R: GTTCCCATCCTATTAGAC				
15	mEgCIR2600	F: M13-GGGGATGAGTTTGTTTGTTC	7	59	292/297	285/297
		R: CCTGCTTGGCGAGATGA				
16	mEgCIR0894	F: M13-TGCTTCTTGTCCTTGATACA	7	56	211/217	204/217
		R: CCACGTCTACGAAATGATAA				
17	mEgCIR2887	F: M13-CTACGGACTCACACCTATAT	8	50	107/109	107/109
		R: ATGGTTCATCAATGAGATC				
18	mEgCIR3622	F: M13-GCCAGTTAGGAATACAA	8	50	170/174	154/174
		R: GTCACGCATTTTTCTTG				
19	mEgCIR3592	F: M13-GAGCCAAAACAGACTTCAA	9	56	200/206	200/202
		R: ACCGTATATGACCCCTCTC				
20	mEgCIR3663	F: M13-AGCAAAATGGCAAAGGAGAG	9	56	235/247	235/247
		R: GGTGTGTGCTATGGAAGATCATAGT				
21	mEgCIR0446	F: M13-CCCCTTCGAATCCACTAT	10	53	225/229	225/229
		R: CAAATCCGACAAATCAAC				
22	mEgCIR3826	F: M13-AAACCAAGTCAAGTTCAGTT	10	50	263/267	267/269
		R: TTTTTTTAATTGATGGATAG				
23	mEgCIR3362	F: M13-CCCATCATCTGCTCAGGATAGAC	11	59	165/195	165/195
		R: ACCCTCTCCTCTTGGGAAGA				
24	mEgCIR3653	F: M13-CATGAGATGGTATATAATCTATAC	11	56	149/165	139/165
		R: ACGAGATCTGCTTCATTGT				
25	mEgCIR1730	F: M13-AATTTCAAATACAGCATAGC	12	56	263/273	263/273
		R: CATAGTAAGTTTTGGATGATTATTA				
26	mEgCIR0906	F: M13-TTTTATTTTCCCTCTCTTTTGA	12	56	155*	155/177
		R: ATTGCGTCTCTTTCCATTGA				
27	mEgCIR0465	F: M13-TCCCCCACGACCCATTC	12	56	142/151	151*
		R: GGCAGGAGAGGCAGCATTC				
28	mEgCIR0772	F: M13-TATAATCCACCCAGCACAAC	14	53	166/182	166/184

			LG	Tm		
		R: CCAATTATACAATCCCACAAAG				
29	mEgCIR3350	F: M13-GGAATAAAGCTTCCAACAAC	14	62	309/311	309/311
		R: CCTGGTCGTTTGGTAGAGA				
30	mEgCIR2409	F: M13-TAATTCATGAGTGCCCAACA	15	59	176/186	184*
		R: TATGGTCCCACAAACTTCTC				
31	mEgCIR0230	F: M13-CCCTGGCCCCGTTTTTC	15	59	341/346	345/359*
		R: AGCGCTATATGTGATTCTAA				
32	mEgCIR1729	F: M13-TACGTGAAAGGCTTGCTTAT	15	56	132/134	121/134
		R: ATGGATTCATTTCGTTCACA				
33	mEgCIR0773	F: M13-GCAAAATTCAAAGAAAACTTA	15	62	288/290	263/312
		R: CTGACAGTGCAGAAAATGTTATAGT				
34	mEgCIR3346	F: M13-CTTCAAGGATTATGAAGTTA	15	56	190/196	186/198
		R: ATTGTGTCGAGAGCTATGA				
35	mEgCIR0353	F: M13-AGAGAGAGAGAGTGCGTATG	16	59	106/110	110/114
		R: GTCCCTGTGGCTGCTGTTTC				
36	mEgCIR0782	F: M13-CGTTCATCCCACCACCTTTC	16	62	176/189	176/189
		R: GCTGCGAGGCCACTGATAC				

F: Forward primer; R: Reverse primer; M13: 5'-CACGACGTTGTAAAACGA C-3';

LG: Linkage group; Tm: Optimal annealing temperature

* Monomorphic primer or highly distorted segregation for that population

5.3.3.3 **Genotyping of the 768 and 769 populations**

Large scale genotyping of the 768 and 769 populations were performed using the selected 36 polymorphic SSR markers. Figure 5.8 shows the gel electrophoresis profiles obtained using the mEgCIR2518 and mEgCIR0772 primers in which amplification is consistent for most genotypes, although the polymorphism is unresolved on the agarose system. Figure 5.9 shows examples of fragment size analysis of the mEgCIR2518 primer. The size of PCR products of SSR primers was noted through gel electrophoresis. This allowed four different primer pairs to be multiplexed for CEQ fragment analysis [Figure 5.9 (a)]. Successful fingerprinting allows the recognition of heterozygous [Figure 5.9 (b)], homozygous for allele A [Figure 5.9 (c)], and homozygous for allele B [Figure 5.9 (d)] genotype of each sample.

It is important to note that samples 768/26, 768/27 and 768/29 were not amplified by the mEgCIR0772 primer pairs [Red arrow, Figure 5.8 (b)]. Primer mEgCIR2518 amplified two alleles of 291 and 303 bp in all the samples of the 768 population, except for sample 768/29 which produced two alleles of 294 and 301 bp [Figure 5.9 (e)]. This difference was not detectable with a low resolution agarose gel electrophoresis [Red arrow, Figure 5.8 (a)].
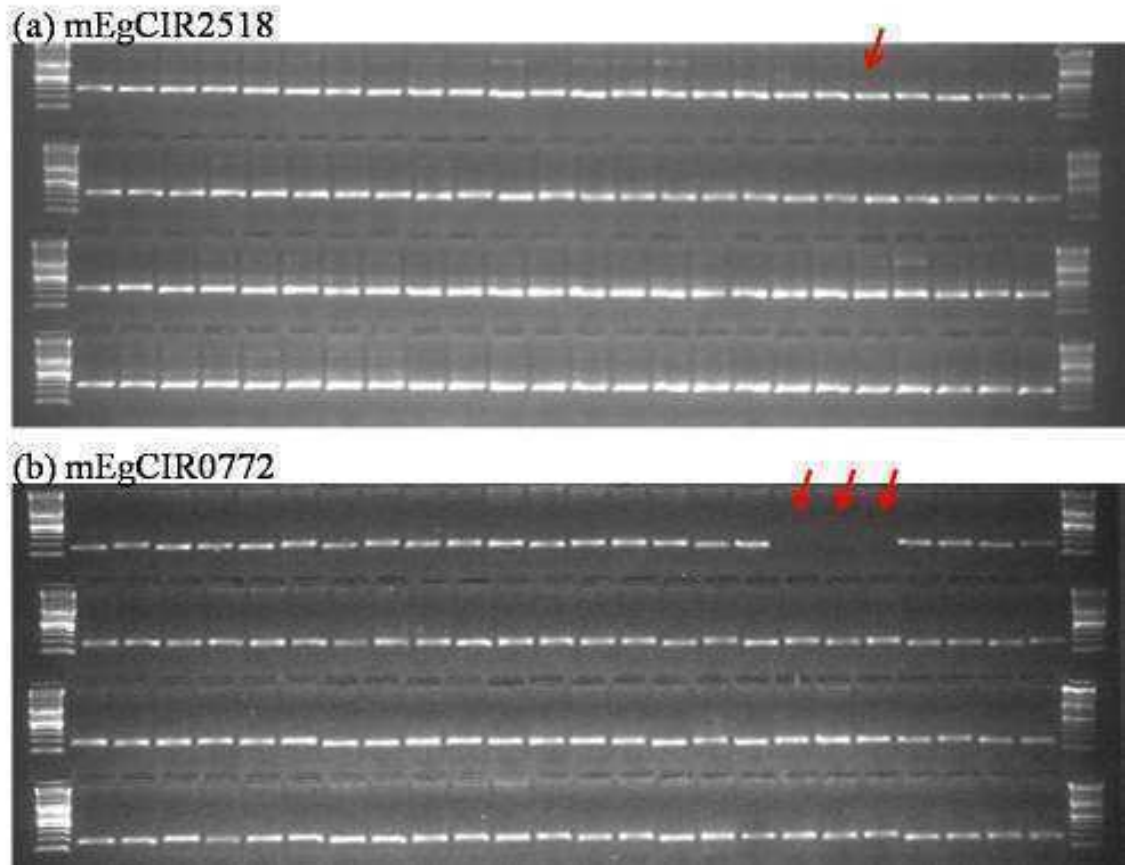
Figure 5.8: Gel electrophoresis profiles of the 768 and 769 populations amplified by (a) mEgCIR2518 and (b) mEgCIR0772 primers. Arrows indicate the outliers identified from DArTSeq genotyping.
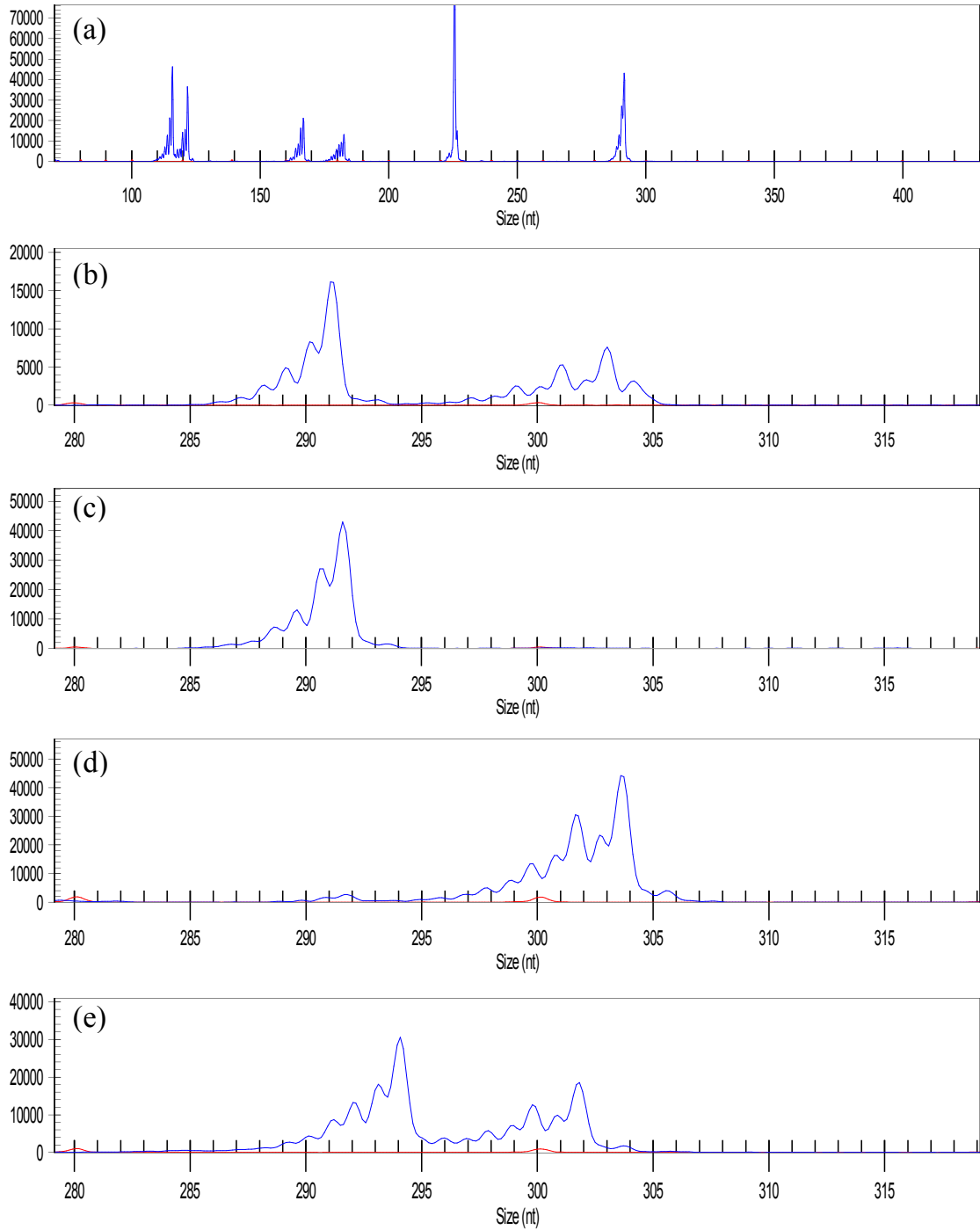
Figure 5.9: Fragment analysis profiles of four samples from the 768 controlled cross.
(a) Multiplexed analysis of mEgCIR3809, mEgCIR2518, mEgCIR0446 and mEgCIR0772 primers for sample 768/3. Upon amplification by mEgCIR2518 primer, (b) 768/2 was found to be heterozygous; (c) 768/3 was homozygous for allele 291 bp; (d) 768/45 was homozygous for allele 303 bp; (e) 768/29, as an outlier, had two totally different alleles.

Genotyping using all available markers revealed that samples 26, 27, 29 and 30 from the 768 and sample 18 from the 769 controlled cross are outcrosses, corresponding with those identified by genotyping using DArT and SNP markers. They were found to have different alleles than the populations for the majority of the SSR primers tested. Outliers from the 768 controlled cross were successfully identified by 30 out of the 36 SSR primers. Further inspection of SSR amplification profiles suggested that outlier samples 26, 27 and 29 from 768 cross were from the same (incorrect) controlled cross, while 768/30 was from another cross.

SSR fingerprinting of the 768 and 769 populations also revealed that mEgCIR2575, mEgCIR3526 and mEgCIR0906 primers are monomorphic in the 768 population while mEgCIR0465 and mEgCIR2409 are monomorphic in the 769 population. Highly distorted segregation patterns were also observed when the 769 controlled cross was genotyped using the mEgCIR2215 and mEgCIR0230 primers. Only 769/50 was found to be homozygous for allele 136 bp of the mEgCIR2215 primer while 769/22 was the only heterozygous individual for primer EgCIR0230, suggesting that these primers are not truly polymorphic in the 769 controlled cross. Despite some of the SSR being monomorphic or having unusual segregation pattern, all 36 SSR markers were included for linkage analysis of markers and determination of segregation distortion using mapping software.

5.4   Discussion

The introduction of molecular markers in the early 1980s enabled unlimited detection and exploitation of DNA polymorphism at any chromosomal location. In the

field of plant studies, hybridization-based markers (such as RFLP) were the first developed and employed followed by amplification-based markers, for example RAPD, AFLP and SSR markers. Recent development of next-generation-sequencing (NGS) and microarray platforms has accelerated the generation of markers, such as SNP and DArT markers (Henry, 2013). Development of molecular markers allows construction of plant genetic maps that are fundamental for understanding the organization of plant genomes and for genetics study (Collard *et al.*, 2005).

The objective of the present study is to develop and characterise both dominant DArT and co-dominant SNP markers from DArTSeq platform by genotyping two closely-related *tenera* self-pollinated 768 and 769 populations. This is the first report of genotyping oil palm crosses with the new DArTSeq platform. Meanwhile, the present study also reports the screening and characterisation of publicly available CIRAD SSR markers through genotyping of the two 768 and 769 populations. The polymorphic markers developed and characterised in this study would be used in the construction of dense genetic linkage maps of oil palm as reported in chapter 6.

5.4.1 Development and characterisation of DArT and SNP markers from the DArTSeq platform

Diversity Array Technology (DArT) is a relatively new marker system that was first reported in early 2000 (Jaccoud *et al.*, 2001). This technique is based on genome complexity reduction using restriction endonucleases which are highly specific to their recognition sequence. Classical DArT provides a microarray hybridization-based high-throughput whole-genome profiling platform for genotyping of hundreds to thousands of

polymorphic loci without any need for prior sequence information (Jaccoud *et al.*, 2001; Wenzl *et al.*, 2004). The development of NGS techniques has also increased the discovery of markers, particularly SNP markers, in many plant species (Davey *et al.*, 2012).

The DArTSeq platform is the newest genotyping technology offered by DArT Pty Ltd in which the platform was developed based on the genome complexity reduction of the DArT array coupling with NGS enabling rapid SNP discovery alongside the generation of 'classical' DArT markers. The use of DArTSeq technology was first published by Sansaloni *et al.* in 2011 for a genetic mapping study of *Eucalyptus*. This technology has also been applied to study the genetic diversity of oilseed crop *Lesquerella* and related species by Cruz *et al.* (2013).

A crucial step in the Diversity Arrays Technology is the complexity reduction of genomic representations. Complexity reduction is a process which generates a defined fraction of genomic fragments reproducibly (Schouten *et al.*, 2012). Numerous efforts have been put in to optimize genomic complexity reduction methods through combinations of restriction enzymes to maximize the number of polymorphic clones/markers (Jaccoud *et al.*, 2001; Wenzl *et al.*, 2004; Schouten *et al.*, 2012; Cruz *et al.*, 2013). The complexity reduction method used most often in DArT involves digestion with the rare methylation-sensitive 6 bp cutter, *Pst*I, together with a frequent cutter, such as *Alu*I, *BstN*I, *Taq*I or *Mse*I (Wenzl *et al.*, 2004). The *Pst*I/*Taq*I combination is one of the routinely used enzyme combinations. As for sequencing-based DArT genotyping, a third restriction enzyme was introduced to eliminate a subset of the fragments (Sansaloni

et al., 2011). In view of the importance of complexity reduction to the success of both DArT microarray and DArTSeq platform genotyping, it is therefore essential to check the quality of samples by prior restriction digestion to ensure that the samples were completely digested by restriction enzymes (Figure 5.2).

Genotyping of 106 progenies from the 768 and 769 populations generated 11,675 DArTSeq markers, including 6,764 and 4,911 of DArT and SNP markers, respectively. The number of markers obtained in the present study is higher than that generated for *Eucalyptus* (Sansaloni *et al.*, 2011). In *Eucalyptus*, 2,835 polymorphic DArT and over 1,500 SNP markers were obtained from the screening of a segregating population of 89 individuals derived from an intra-specific cross. Diversity analysis of *Lesquerella* germplasm generated 27,748 markers from the DArTSeq platform (Cruz *et al.*, 2013), far greater than the number of markers generated in the present study, probably due to the far broader survey of germplasm carried out in *Lesquerella* germplasm compared with the two controlled crosses of oil palm analysed here.

Several parameters were evaluated to determine the quality of the markers generated by DArT technology. One of them is call rate which is the percentage of definite "0" or "1" alleles, compared to 'missing' data. In the present study, an average call rate of 90% and 95% were attained for the DArT and SNP markers with 49% and 78.9% of the markers having call rates of more than 90%, respectively. This call rate is better than the one published by Petroli *et al.* (2012) for his work on *Eucalyptus*, in which they showed that only 36% of DArT markers had a call rate $\geq$ 90%. However, higher call rates have been observed in DArT study of various plant species such as apple (96.7%;

Schouten *et al.*, 2012), olive (96.9%; Domínguez-García *et al.*, 2012) and einkorn wheat (99.2%; Jing *et al.*, 2009) using the classical DArT microarray methodology.

A call rate of $\geq$ 95%, equal to missing data $\leq$ 5%, was set to select both SNP and DArT markers for mapping of the 768 and 769 populations. Due to the small sample size of both populations, a more stringent selection criterion was applied to ensure mapping results were good with high quality data. Most mapping studies using the DArT platform did not set a high threshold of call rate for selecting markers, although this was based upon the original array method where far fewer markers were generated. Map construction of 91 olive seedlings derived from the cross of "Picual" x "Arbequina" was performed using DArT markers selected with quality parameter Q greater than 77 and call rate greater than 80 (Domínguez-García *et al.*, 2012). Meanwhile, in two different genetic mapping exercises using an F1 hybrid from inter-specific cross of *Eucalyptus grandis* and *Eucalyptus urophylla* have chosen DArT markers based on call rates $\geq$ 75% together with others parameters on reproducibility and Q score (Kumar Kullan *et al.*, 2012; Petroli *et al.*, 2012). Petroli *et al.* (2012) commented that marker call rates of $\geq$ 75% would still yield good quality data for map construction hence this less stringent threshold was adopted to maximize the number of markers positioned on the linkage map.

In conclusion, it is believed that stringent marker selection criteria in the present study for genetic maps construction using the DArTSeq platform would produce high quality data. The high number of markers selected (more than 1,600 markers for both

populations) would facilitate satisfactory marker linkage and ordering analysis during the genetic linkage map construction.

5.4.2 Characterisation of SSR markers

Microsatellites or SSR markers are the most widely applied class of molecular markers used in genetic studies. SSRs are tandem repeats of the DNA in the form of iterations of repeat units of 1-6 nucleotides (Ellegren, 2004).

SSR markers were first reported in oil palm by Billotte and his group in 2001. Billotte *et al.* (2001b) reported the development and characterisation of the first set of 21 CIRAD SSR markers as well as their utility across oil palm species. The first microsatellite-based high-density oil palm linkage map was published in 2005 (Billotte *et al.*, 2005) and followed by QTL analysis of important yield traits on a multi-parent linkage maps constructed solely based on SSR markers (Billotte *et al.*, 2010). The Malaysian Palm Oil Board (MPOB), the biggest oil palm research institute in South East Asia, has also developed SSR markers, particularly through mining of EST databases (Singh *et al.*, 2008, Ting *et al.*, 2010).

Microsatellite markers have been extensively used for map construction due to their ubiquitous occurrence, multi-allelic nature, high levels of polymorphism, transferability between populations, easily automation and exchanged between laboratories. However, development of SSR markers requires high levels of expertise and the availability of sequence information; it is a tedious and cost intensive project (Kalia *et al.*, 2011). Therefore in the present study, no development of SSR markers was involved

but rather SSR markers available in public database were screened. SSR markers, particularly those developed by CIRAD, were exploited as anchor loci in this first reported DArTSeq marker-based genetic linkage study of oil palm. The same set of CIRAD SSR markers, together with those isolated by FELDA and MPOB, were also utilized to construct the first genetic maps of FELDA oil palms (Seng *et al.*, 2011). There is a widespread use of SSR markers as anchor loci and/or assigning linkage groups in the genetic mapping of other plant species using DArT markers, for example rapeseed (*Brassica napus* L.; Raman *et al.*, 2012), banana (*Musa acuminata.*; Hippolyte *et al.*, 2010) and einkorn wheat (*Triticum monococcum*; Jing *et al.*, 2009).

In the present study, the polymorphism of the SSR primer pairs was determined using CEQ^{TM} 8000 Genetic Analysis System (Beckman Coulter, USA), instead of agarose gel electrophoresis. A 3% agarose gel has been used by several researchers to screen for polymorphism of SSR markers with at least 20 bp difference in allele size (Beyene *et al.*, 2005; Legesse *et al.*, 2007; Ashkani *et al.*, 2012). However, this resolution is too low for detection of allele differences in the present study given that the majority of SSR alleles had less than 20 bp difference between alleles, with the smallest only two bp difference. CEQ^{TM} 8000 Genetic Analysis System generates electropherograms peak profiles, allowing manual scoring of individual SSR product sizes. Although CEQ^{TM} 800 contains an automated allele binning wizard, Nariman (2013) commented that visual inspection of fragment size is recommended to avoid mis-reporting of automated sizing caused by scoring stutters as consistent major peaks in automated scoring. Heterozygosity of individual progeny was also confirmed by direct comparison of their entire microsatellite profile with those of the parental alleles, with which the automated calling

system is unable to cope with changes in the relative peak height between alleles. Manual scoring allows an inspection of the shift in overall microsatellites "shape", which is generally more informative than scoring individual peaks.

Manual scoring of SSR alleles was performed according to advice highlighted by Selkoe and Toonen (2006). The major potential scoring error that might be encountered was predominantly due to stutter peaks. During PCR amplification, some products are one, two, or three repeats short in sequence due to errors in PCR amplification and these show up in the electropherograms as evenly spaced peaks with decreasing height to the left of the true peak and are called stutter peaks [Figure 5.7 (d)]. Stutter peaks are of use to differentiate SSR alleles from non-specific products. However, scoring can be difficult or confusing when two alleles of the same primer with stutter peaks are only one repeat unit different in size. In this case, the second peak is higher than the first peak as overlapping of the first stutter peak of the larger allele with the true peak of smaller allele increases the signal intensity of the true peak of smaller allele [Figure 5.7 (e)]. Meanwhile, larger alleles usually have slightly shorter peaks than smaller alleles due to less efficient PCR amplification of longer repeat units. This was discerned in Figure 5.7 (c) and (d) of the present study.

In order to reduce the cost of using fluorescently-labelled dyes, a three primers system (Schuelke, 2005) was utilized for the current SSR polymorphism screening, the same approach was adopted for the AFLP analysis using the LICOR electrophoresis system that were previously discussed in chapter 4. The cost of screening a large number of microsatellites and progeny in the present study was further reduced by multiplexing

PCR products of different SSR markers that showed size differences of at least 30 bp. Four SSR markers were pooled in the current study with all the primers being labelled with only the fluorescent blue dye D4. The green dye D3 was used when the allele size of different microsatellites was similar (< 30 bp). The comparatively weak D2 black dye was avoided. Previously, Molosiwa (2012) reported the generation of false peaks due to colour bleed through when multiplexing different PCR products of similar size with D3 green and D4 blue dyes. This was not observed in the present study. However to counteract the relatively weaker signal of D3 green dye, compared to D4 blue dye, the amount of PCR products labelled with D3 green dye was doubled or tripled when added to the mixture.

A total of 102 CIRAD SSR markers were screened in the present study and 36 markers were found to be polymorphic in current mapping populations. This polymorphism level (35%) is much lower than the one identified by Seng *et al.* (2010) for a high yielding cross in FELDA oil palm breeding programme. In this exercise, Seng *et al.* screened 255 CIRAD SSR markers and 144 (56.5%) markers were mapped. This is most likely due to the *tenera* self-pollinated controlled crosses used in the present study having narrower origins, compared to *dura* x *pisifera* cross of FELDA (Seng *et al.*, 2010). Different polymorphism levels were also detected when SSR markers were used together with DArT markers for genetic mapping of other plant species, namely 45.8% for banana (Hippolyte *et al.*, 2010), 32.8% for einkorn wheat (Jing *et al.*, 2009) and 34.3% for hexaploid wheat (Semagn *et al.*, 2006).

Nevertheless, the 36 polymorphic CIRAD SSR markers were used in the genetic linkage map construction of the 768 and 769 populations together with the SNP and DArT markers generated from DArTSeq platform. These anchor SSR loci were used to assign and orientate linkage groups with reference to oil palm genetic maps published by Billotte *et al.* (2010). The construction of the genetic linkage maps is reported in chapter 6.

In conclusion, this chapter reported the development of first set of DArTSeq markers in oil palm and characterization of a subset of more than 1600 high quality polymorphic DArTSeq marker. These SNP and DArT markers together with the 36 polymorphic SSR markers were used in the construction of high density genetic linkage maps (chapter 6) as well as to identify markers closely linked with important agronomic traits, such as Shell-thickness gene (*Sh*) (Chapter 8) and yield traits (Chapter 7).

Chapter 6

Construction of genetic linkage maps using DArTSeq and SSR markers

6.1   Introduction and objective

The development of a wide range of molecular markers that reveal differences at the DNA level has contributed to extensive genetic mapping in many species. Genetic mapping, also known as linkage mapping, is a process of assigning the available markers to different groups based on two-point analysis, ordering the markers along the linkage groups and determining the relative genetic distances between them on the basis of their recombination frequency. Genetic maps are the graphic representations of the arrangement of genes or markers on the chromosomes. Recombination events, the naturally-occurring "breaking and rejoining" of segments of chromosomes during meiosis, is the fundamental basis for construction of genetic maps (Grant and Shoemaker, 2001).

For orphan plants whose genomes are yet to be sequenced, genetic maps are vital for understanding the order and spacing of markers as well as the relative order to those of other plants through comparative mapping. A genetic map can also provide a scaffold for genome sequence assembly and validation. Most importantly, genetic maps underpin the study of key plants genes and quantitative trait loci which in turn facilitates marker-assisted selection in plant breeding programmes (Cheema and Dicks, 2009). Ultimately, breeding programmes depend upon the patterns of genetic recombination to produce the new combinations of trait genes for selection.

Oil palm genome mapping based on DNA markers began in late 1990s, and since then several genetic maps of oil palm have been constructed. RFLP was the first molecular marker developed for mapping in the human genome (Botstein *et al.*, 1980)

and subsequently plant genomes (Bernatzky and Tanksley 1986; Weber and Helentjaris, 1989), including oil palm (Mayes *et al.*, 1997). This first linkage map was constructed from 97 co-dominant RFLP loci which gave 24 linkage groups. Despite the reliability of RFLP markers, it is very tedious and costly to develop maps based on them; hence later genetic maps have utilised PCR-based molecular markers. The second genetic map of oil palm was constructed using RAPD markers (Moretzsohn *et al.*, 2000). The first high density linkage map of oil palm was created by Billotte *et al.* (2005) using SSR and AFLP markers. Since then SSR and AFLP markers have become the marker of choice for the construction of oil palm genetic linkage maps, with a contribution from other marker types (Singh *et al.*, 2009; Billotte *et al.*, 2010; Singh *et al.*, 2010; Seng *et al.*, 2011). The majority of the above mentioned genetic studies worked with controlled cross populations that segregated for the shell-thickness gene (*Sh*), allowing screening for markers closely linked to this economically important trait.

The objective of this chapter is to construct the first genetic linkage map of oil palm using DArTSeq (DArT 'Genotyping-by-sequencing' generating DArT and SNP) and SSR markers. In chapter 5, DArT and SNP markers generated from the DArTSeq platform as well as SSR markers from the CIRAD public database were characterized, selected and discussed. These subsets of markers were subsequently used to generate genetic linkage maps for the 768 and 769 controlled cross populations, with SSR markers being the anchor markers for assigning linkage group identities to their putative linkage group, through reference to previously reported studies (Billotte *et al.*, 2005, 2010). This is the first study generating genetic linkage maps for Advanced Agriecological Research Sdn. Bhd. (AAR) breeding materials. The populations used in

this study are segregating for the fruit shell-thickness trait and thus the constructed maps could be of value to search for markers linked to the shell-thickness gene (*Sh*) as well as others economically important quantitative traits.

6.2   Materials and methods

Two *tenera* self-pollinated F_2 populations, 768 and 769, were used for the construction of two genetic linkage maps. Both the *tenera* parents, 228/05 and 228/06, were siblings from the same *tenera* x *pisifera* cross. The 768 and 769 populations, consisting of 44 and 57 legitimate progenies, respectively, were genotyped with selected polymorphic SSR and DArTSeq markers as reported in chapter 5, and the marker scores were used for the construction of the genetic maps in the current study. The fruit variety of each progeny was determined phenotypically and scored as a morphological marker to allow mapping of the shell-thickness gene (*Sh*). The JoinMap 4.1 Software (Van Ooijen, 2006) was used to construct the genetic maps for the two F_2 segregating populations.

6.2.1   Coding of genotype data and preparation of data files

It is important to firstly determine the linkage phases of the markers, either in coupling or repulsion. However, the *tenera* and *pisifera* grandparents of both the 768 and 769 controlled cross populations were no longer available in the field and hence there was no parental data available for phase determination. Because of this, both self-pollinated populations were first analysed using a Cross Pollinator (CP) designation within JoinMap 4.1 and a genotype coding of <hkxhk>. Genotyping data of all markers

were converted as follows; for dominant DArT markers, presence of a defined allele was scored as "k-" with "hh" indicating the absence of an allele – the fully informative state. Both SSR and SNP are co-dominant markers in which genotyping data for two alleles (A and B) of each marker are available. The data were scored as "hh" for presence of allele A only, "hk" for presence of both allele A and B; and "kk" for presence of allele B only. The locus designations used by Diversity Array Technology Ltd for DArT and SNPs generated were adopted with modification in this study. The first 4 universal digits ("1000") were removed and a prefix "D." or "S." was added to indicate DArT or SNP, respectively.

Data for analysis were prepared according to the format given in the manual of JoinMap 4.1. The main file is a plain text *locus genotype file*, also known as a *loc-file*, can be prepared using a text editor program, such as *Windows WordPad*. This *loc-file* contains the genotype codes for all the loci of a segregation population and has a sequential structure. The file contains four instructions as the header defines the name of the population, the type of the population, the number of loci and the number of individuals, followed by a data body that contains the genotype information of all loci for all individuals in the population. For this study, the population type was initially set to CP.

6.2.2   Linkage analysis and phase determination of markers

The CP *loc-files* were loaded into the JoinMap 4.1 software for analysis. Before beginning mapping, coding data was checked for errors and highlighted errors were corrected where possible or data marked as missing. Segregation patterns and the

presence of any segregation distortion were calculated by the software. For each segregating marker, a chi-square goodness-of-fit analysis was performed to test for deviation from the expected segregation ratio, 1:3 for dominant DArT and 1:2:1 for co-dominant SSR and SNP markers for significance p-values of 0.05, 0.01, 0.005, 0.001, 0.0005 and 0.0001. Markers with segregation distortion at $p<0.0005$ significance level were excluded from further analysis.

Markers in both populations were grouped according to "independence LOD". The groups were manually selected using thresholds from LOD 4-5 so as to ensure that the SSR loci that have been assigned to a particular chromosome in previously published genetic maps (Billotte *et al.*, 2005, 2010) were in the same group, where possible. The phase of markers was then determined by the software within each linkage group, inferring phase for dominant markers from surrounding co-dominant markers.

6.2.3   Phase conversion of markers and preparation of data files for conventional F_2 mapping

Once the phase was determined by the software, all markers in the linkage groups were collected into a new file and the genotyping data converted back for analysis as a "F_2" population type. According to the JoinMap 4.1 manual, the genotype codes for F_2 population type are as follows:

Code	Description
a	Homozygote as the first parent
b	Homozygote as the second parent
h	Heterozygote (as the F_1)
c	Not genotype a
d	Not genotype b
-	Genotype unknown

The coding data was then converted as follows:

For dominant DArT markers with phase (0,0), the conversion was as below:

'k-' converted to 'c'

'hh' converted to 'a'

For dominant DArT marker with phase (1,1), the conversion was as below:

'k-' converted to 'd'

'hh' converted to 'b'

For co-dominant SSR and SNP markers showing phase (0,0), they were converted as below:

'hh' converted to 'a'

'kk' converted to 'b'

'hk' converted to 'h'

For co-dominant markers showing phase (1,1), they were converted as follows:

'hh' converted to 'b'

'kk' converted to 'a'

'hk' converted to 'h'

The plain text *loc-files* were prepared in the same way as described in section 6.2.1, except that the population type was changed to F_2.

6.2.4   Linkage analysis of makers and map construction

Linkage groups of F_2 populations were established using LOD scores from 4-10. Linkage groups were assigned to chromosomes according to the known location of SSR markers in Billotte *et al.* (2005, 2010). For linkage groups that belong to the same chromosome but did not group, an attempt was made to combine groups using a lower LOD by preparing a new *loc-file* only including the suspected fragments of the linkage groups marked by microsatellites from the same linkage group.

Mapping of markers was performed using the regression mapping algorithm (Stam, 1993) at the default value of recombination frequency ≤ 0.4; LOD score ≥ 1; goodness-of-fit jump threshold = 5, ripple value = 1. Regression mapping builds the map by adding loci one by one, starting from the pair of loci with the most evidence for linkage (highest LOD). The best position of each subsequent locus is determined by comparing the goodness-of-fit of the calculated map for each tested position. The locus is removed when the goodness-of-fit decreased too sharply (the Chi-square score for the overall map increases by a jump of more than the set threshold of 5) or when the

locus give rise to negative distance estimates in the map. This process is continued until all loci are handled once and this becomes map 1. The order of the markers is fixed. In the second round, a new attempt was made to map all removed loci into the fixed order from map 1. Loci are removed when the jump was too large or negative distances are encountered. At the end of the second round, the mapping order is fixed and this becomes map 2. In the third round, all previously rejected markers are added to the map without constraints, giving rise to map 3. Therefore, map 3 often includes most markers but can be undesirable as it can contain potentially poor markers or those leading to potential conflicts with other markers. Haldane's mapping function was used to convert recombination frequencies into map distances in units of centiMorgan (cM).

6.3 Results

Inheritance and segregation analysis of markers as well as the construction of two genetic maps for the two small closely-related controlled cross populations, 768 and 769, are reported in this chapter. A total of 1,704 markers, consisting of 36 SSR, 948 DArT, 719 SNP markers and one morphological marker, the *Sh* gene, were selected for linkage mapping of the 768 controlled cross while 1,724 markers were selected for the 769 controlled cross population containing 36 SSR, 958 DArT, 729 SNP markers and the *Sh* gene.

6.3.1 Inheritance and segregation analysis of markers

Segregation distortion was calculated using JoinMap 4.1 and several markers were found to be distorted at a significance level *p*<0.05 (Table 6.1). Three and four

SSR markers were distorted at $p<0.005$, respectively, for the 768 and 769 controlled cross populations and removed.

Table 6.1: Segregation distortion of markers for both the 768 and 769 controlled cross populations at different significance levels.

Type of marker	Mendelian Segregation ratio	768 population			769 population		
		Total no. of markers	Distortion at level		Total no. of markers	Distortion at level	
			p<0.05	*p*<0.0005		*p*<0.05	*p*<0.0005
SSR	1:2:1	36	6 (16.7%)	3	36	10 (27.8%)	4
DArT	1:3	948	104 (11%)	2	958	99 (10.3%)	5
SNP	1:2:1	719	53 (7.4%)	6	729	81 (11.1%)	11
Sh	1:2:1	1	-	-	1	-	-
Total	-	1704	163 (9.6%)	11	1724	190 (11%)	20

The majority of the loci in both the 768 and 769 controlled crosses segregated in the expected Mendelian ratio of 1:2:1 for SSR and SNP markers or 1:3 for DArT markers. The 769 controlled cross was found to have more distorted SSR markers, ten as compared to six SSR for the 768 controlled cross at the 5% significance level. The percentage of segregation-deviated DArT markers was similar for both populations, 10.3% and 11% of the total DArT markers. Both the 768 and 769 controlled crosses have 53 and 81 SNP markers, respectively, that showed segregation distortion at $p<0.05$; a notably higher percentage for the 769 controlled cross. No significant deviation was found from the 1:2:1 segregation ratio expected for *dura*: *tenera*: *pisifera* within each cross for the shell-thickness (*Sh*) major Mendelian gene. Overall, the 769 controlled cross had more distorted markers, 190 compared to 163 markers for the 768 controlled cross.

Only markers showing very significant distortion ($p<0.005$) were excluded from further mapping analysis. This included three SSR, two DArT and six SNP markers for the 768 controlled cross and four SSR, five DArT and eleven SNP markers for the 769 controlled cross (Table 6.2). Therefore a total of 1,693 and 1,704 markers from the 768 and 769 populations, respectively, were used for the further phase determination and linkage analysis of markers.

Meanwhile, analysis of individual progenies revealed that samples 768/34 and 769/53 had the highest number of missing data for each family. Both 768/34 and 769/53 had 118 and 153 missing data, respectively, which constituted 6.9% and 8.9% of the total polymorphic markers selected for each population.

6.3.2   Phase determination and map construction

Due to the lack of parental data, the phase of segregating markers was determined by the JoinMap 4.1 software through an initial analysis of the population as a Cross Pollinator (CP), in which groupings of markers using LOD 4-5 produced 21 linkage groups with 1,645 markers and 17 linkage groups with 1,690 markers for the 768 and 769 controlled crosses, respectively. Those groups with only 1-2 markers were discarded and formed a total of 48 markers ungrouped in the 768 controlled cross and 14 markers in the 769 controlled cross (Table 6.2). The majority of these ungrouped markers were dominant DArT markers. Linkage phase was then determined by the software for all markers within each and every linkage group.

Table 6.2: Summary of marker elimination during the construction of genetic maps.

Marker Type	768 population				
	Total	Excluded	Ungrouped	Unmapped	Total mapped
SSR	36	3	-	-	33
DArT	948	2	47	58	839
SNP	719	6	1	32	682
Morphological	1	-	-	-	1
Total Markers	1704	11	48	90	1555

Marker Type	769 population				
	Total	Excluded	Ungrouped	Unmapped	Total mapped
SSR	36	4	-	-	32
DArT	958	5	14	103	836
SNP	729	11	-	52	666
Morphological	1	-	-	-	1
Total Markers	1724	20	14	155	1535

On the basis of the determined linkage phases, the two controlled crosses were reanalysed as F_2 populations with the coding data of markers converted to the genotype codes for F_2 populations. Grouping of markers using LOD 4-10 again produced 21 and 17 linkage groups for the 768 and 769 mapping populations, respectively. Detailed inspection of the linkage groups generated by CP and F_2 analyses revealed that the grouping of markers into linkage groups was the same for both analyses, indicating that marker conversion was successful. Linkage groups were then assigned to chromosomes using the known location of anchoring SSR markers in the reference genetic map published by Billotte *et al.* (2010). The results showed that a number of linkage groups were assigned to the common chromosomes. Separate analysis on these linkage groups showed that the combining of the groups could be achieved using lower LOD of 2.5 – 3.9, as highlighted in Table 6.3. Linkage group 8 of the 768 controlled cross was

observed to be fragmented into three different linkage groups and they were regrouped using LOD score of 2.8. Through this approach an initial set of 16 linkage groups were obtained for both the 768 and 769 populations.

Table 6.3: LOD score, number of map rounds to include all markers and the final adopted map for each linkage group.

	768 population			769 population		
Linkage groups	LOD score	No. of round for regression mapping	Map used	LOD score	No. of round for regression mapping	Map used
1	5.4	3	2	5.1	3	2
2	3.5*	3	2	5.4	3	2
3	4.6	3	2	6.9	3	2
4	3.0*	A-3; B-3	A-2; B-2	6.9	3	2
5/13	4.5	1	1	4.9	1	1
6	4.9	2	2	5.3	3	2
7	4.3	3	2	8.3	3	2
8	2.8*	3	2	4.3	3	2
9	5.7	3	2	5.7	3	2
10	2.5*	A-3; B-3	A-2; B-2	4.5	A-3; B-3	A-2; B-2
11	4.9	3	2	5.4	3	2
12	4.5	3	2	3.7*	3	2
13/5	4.8	3	2	10	3	2
14	5.7	3	3	9.4	3	2
15	4.5	3	2	6.3	3	2
16	6.6	2	2	8.8	3	2

* Linkage groups that were combined separately using lower LOD scores.

The regression algorithm available in the JoinMap 4.1 software was employed for map generation. This generated three rounds of mapping for the majority of the linkage groups in both populations. As map 3 involved removing the constraints which were in place for round 1 and 2, it was not accepted as the final version of the linkage group and round two maps (map 2) were selected instead (Table 6.3). All the markers in linkage group (LG) 5, the smallest linkage group, were ordered into the map during the first round of mapping (Map 1). For LG 6 and 16 of the 768 controlled cross, all the

"jumped markers" from Map 1 could be added into Map 2 without a need to relax the criteria and hence Map 2 were selected to assemble the linkage map of the 768 population. At the end of linkage mapping, a total of 103 DArT and 52 SNP markers were unmapped from the linkage groups of the 769 controlled cross, greater than the number of markers that were removed from the 768 controlled cross, 58 DArT and 32 SNP markers only (Table 6.2).

LG 10 was divided into A and B, although these two parts can be grouped at LOD of 2.5 and 4.5, for the 768 and 769 controlled cross, respectively. Interestingly, part B of the 768 controlled cross achieved a high LOD of 3.7 when linked to LG 11. Therefore, this portion of markers was regarded as part B of LG 10 tentatively in this project. Map generation for LG 10 using the regression algorithm failed while mapping using the maximum likelihood algorithm resulted in undesirably large gaps between the groups in the map (data not shown). In view of the problems encountered, LG 10 was analysed as 2 separate groups of LG 10A and 10B for both populations and maps were generated separately with unknown orientation. The same combining problem using the regression algorithm was also observed on LG 4 of the 768 controlled cross. Orientation of the maps of LG 4A and 4B in the 768 controlled cross was determined using the map of LG 4 of the 769 controlled cross as a reference and the framework microsatellites.

In summary, this study generated 18 and 17 linkage groups for mapping populations 768 and 769, respectively. These linkage groups were combined together

into 16 independent linkage groups, which corresponded well to the 16 homologous chromosome pairs of oil palm.

6.3.3   Evaluation of markers on the maps

Figures 6.1 and 6.2 present the genetic maps generated for the 768 and 769 controlled-cross populations, respectively, and Table 6.4 and 6.5 summarize the characteristics of each linkage group. No polymorphic SSR markers were successfully identified for linkage groups 5 and 13. Therefore the identity of linkage groups without SSR marker cannot be confirmed and the 2 linkage groups were named as 5/13 and 13/5.

6.3.3.1   Map length and genome coverage

The genetic linkage map of the 768 controlled cross contains 1,555 polymorphic marker loci (33 SSR, 839 DArT, 682 SNP and the *Sh* locus) assigned into 18 linkage groups with 15-158 markers per group, with an average of 86.4 markers per group. The map spanned 1,874.81 cM with an average length of 104.16 cM per group and an average marker density of one marker every 1.33 cM. The linkage map of the 769 controlled cross was produced using 1,535 markers (32 SSR, 836 DArT, 666 SNP and allele *Sh*) distributed on 17 linkage groups with 6-226 markers per group, giving an average of 90.3 markers per group. This genetic map has a total map length of 1,720.61 cM giving an average length of 101.21 cM per group and an average of one marker for every 1.62 cM. The genetic maps constructed for both the 768 and 769 controlled

crosses have comparable genome coverage, average map length and average marker density.

LG 2

LG 3

LG 5/13

LG 4A

LG 4B(1)

LG 4B (2)

LG 6

Position	Marker
0	D.16711
	D.14296
10.35	mEgCIP3358
14.87	S.18686|22:T>A
20.51	D.04783
21.23	
24.70	S.12350|16:T>C
25.67	S.19016|68:T>G
25.91	D.09194
26.46	S.23021|18:G>C**
27.09	D.02317
27.20	S.21835|58:A>G
27.97	D.00855, D.14536
28.00	D.09069
28.17	D.17620
28.22	D.17720**
28.66	S.16244|45:C>A
29.23	S.09746|51:T>C
29.83	D.14501**, D.26023**
30.04	B.07328
30.23	D.15812
30.79	D.19527
32.48	S.10894|29:T>A
54.59	D.14522**
34.61	D.02777**, D.00668**
46.57	D.07581
46.62	D.14239**
48.70	D.25049
59.12	D.11390
63.95	D.21290**
64.36	D.22370**
68.66	D.04504
68.76	D.17439
77.62	D.19914, D.11582
79.62	D.17338
80.33	D.12067**
80.71	D.11465
84.47	D.04375
84.61	D.18458
84.73	D.00064
85.27	D.06752
87.93	S.22962|65:A>G
89.48	S.21620|26:G>C
90.55	D.08636
92.09	mEgCIP0783
92.85	D.23837
93.19	S.08476|55:G>C
95.53	D.22248
96.50	D.16644
96.76	S.02582|16:A>T
98.84	S.18455|52:T>G
103.50	S.15339|33:T>G

LG 7

Position	Marker
0	S.24228|57:C>T
1.59	S.14086|29:C>T
2.54	D.05259
2.56	S.34329|37:T>C
2.99	S.24934
3.02	S.17984|17:C>T
6.06	mEgCIP0894
7.46	S.17259|12:G>A
12.81	S.13578|20:A>G
13.72	D.01313
16.47	D.07798
16.78	S.15814|18:T>C
17.04	D.10933
17.34	D.26889
17.60	D.22530
17.61	D.03765
20.25	S.25028|56:C>A
20.27	D.20845
20.28	D.10696
20.33	D.10275
20.83	S.01098|27:T>G
21.85	D.24722
22.46	D.07575, D.02163
23.73	D.08830
27.97	D.18769
23.10	S.19643|45:T>A
23.31	S.10238|37:C>G
24.21	S.13252|47:G>T
24.63	D.02823
25.10	S.17598|56:C>T
27.93	S.12108|7:G>A
28.09	D.01679
29.32	S.16626|32:G>A
30.29	D.02580
31.55	D.02087
32.67	D.19302
36.72	S.25915|61:T>C**
37.51	D.16682
38.40	D.14832
46.61	S.08137|64:T>C
49.31	S.19253|6:T>C
49.90	D.09407
51.07	S.05791|42:G>T
52.42	D.03711
52.97	D.20639
55.79	S.10224|12:A>T
56.22	S.24593|32:A>T
56.97	D.17583
56.99	D.19967
57.00	D.22519, D.00889
57.80	D.12670
57.83	D.13360
60.34	D.19389
60.47	D.24375
62.50	D.15734
64.70	D.15269
64.93	D.14850
66.54	S.13901|31:C>A
71.49	S.08651|53:C>T
75.25	D.03395
77.43	S.04110|39:T>C
83.15	D.42551**
84.12	S.12908|20:A>G
85.49	D.17245
85.65	mEgDR2600

LG 9

LG 11

LG 12

LG 14

Figure 6.1: Genetic linkage map of the 768 population. The map consists of 1,555 markers loci (33 SSR, 839 DArT, 682 SNP and *Sh* locus). Marker names are shown to the right of each LG, with map distances in centiMorgans (Haldane units) to the left. D: DArT marker, S: SNP marker, mEgCIR: *E. guineensis* SSR marker. *: Skewed markers at *p*= 0.1, **: Skewed markers at *p*= 0.05, ***: Skewed markers at *p*= 0.01, ****: Skewed markers at *p*= 0.005, *****: Skewed markers at *p*= 0.001.

LG 1 (1)   LG 1 (2)

LG 2

LG 3

LG 5/13

LG 4 (1)

Position	Marker
0	S.25001|53:T>C
2.88	S.08677|64:A>G
3.25	S.06301|43:C>T
4.35	D.16025
4.39	D.02369
5.26	D.17430
6.47	D.03252
6.74	S.20683|53:C>T
7.85	D.01380
8.36	D.07554
8.43	D.06444
8.87	S.06735|57:G>A
9.25	S.22560|37:G>C
9.61	SH
10.49	S.18695|31:T>A
11.26	S.04059|60:C>T
12.31	D.03941
14.99	D.07313
18.19	S.18672|16:T>C
20.00	S.07830|65:G>A
21.12	D.02902
21.57	S.23082|46:T>C
22.87	S.05801|46:T>C
24.32	S.23961|37:T>C
25.36	S.04102|60:T>C
28.69	D.04891
29.02	D.06039
29.07	D.03616
29.11	D.01759
29.73	S.37363|46:A>G
32.19	S.09220|53:T>A
32.54	D.12453
33.58	D.12070
33.60	D.18343
33.73	D.13429
33.78	D.04100
34.20	S.05222|15:A>G
38.57	D.03234, D.15866
38.87	S.15448|30:T>G
38.96	D.16796
40.48	D.00958
40.54	D.21912
40.62	D.07171
40.77	D.19563
41.34	S.15976|19:A>T
41.85	S.15357|20:C>T
42.20	S.13743|26:C>T
43.48	mEgCIR3526
45.10	D.06865
45.12	D.03299
45.26	D.15497
45.54	D.17816
45.65	D.03339
45.69	D.20085
48.58	D.02443
50.65	S.27294|42:G>T
53.04	D.35465
53.44	D.09504
54.02	D.08043
55.59	D.07214
59.95	D.01225
60.20	D.16873**
61.26	D.26924
61.31	D.36686
61.37	D.00854
61.63	D.14223
61.66	D.10057
63.43	D.01192
63.48	D.11584
65.72	D.15694
66.82	D.10431
67.15	D.01596
67.43	D.09240
68.14	D.12317
68.65	D.09866
70.00	D.18294
71.47	D.03497
71.68	D.14271
73.28	D.13505
74.84	S.24917|32:G>A
76.06	D.06478
77.67	D.17682
79.08	D.08767
79.13	D.01729
79.48	S.21204|23:G>A
80.60	D.02942
81.14	D.01713
81.20	D.11497
81.54	S.07674|50:T>C
82.10	S.18252|54:C>G
83.19	S.09116|27:C>A
86.56	D.02930
86.94	S.24598|62:G>A
87.98	S.09456|22:G>T
89.32	S.19691|40:A>T
89.38	S.16973|58:A>G, D.15459
89.43	D.03568
89.53	D.12726
89.54	D.18936
89.87	S.17703|29:C>A, S.10290|7:T>A
90.27	S.10488|20:T>C
90.55	D.05823
90.56	D.04921
92.30	S.23040|13:T>A
92.66	D.14396
92.87	D.13422
93.45	S.20396|19:T>C
94.13	S.21007|28:T>A

LG 4 (2)

Position	Marker
94.13	S.21007|28:T>A
94.81	S.21681|20:A>G
95.19	S.24530|51:T>C
95.23	D.08974
96.16	S.20442|14:G>T
96.76	S.21852|19:C>T
97.67	S.19466|24:C>T
98.83	D.10469
99.75	S.28492|51:C>T
100.43	D.13853
100.82	D.03588
101.91	S.07919|65:G>T
102.69	S.16554|25:C>T
103.35	S.12273|40:G>A
103.78	D.07788
104.63	S.17817|22:T>A
105.54	S.19191|67:T>C
106.93	D.16391
107.36	S.23845|55:T>G
110.59	D.18530
111.08	D.00689
112.02	S.20800|6:T>G
113.37	D.11044**
113.62	D.03555
115.65	S.15083|57:C>T
116.31	S.27060|49:C>G
117.12	S.19925|58:T>C
118.39	S.05650|59:A>G
119.40	S.22836|44:A>T
121.37	S.25512|48:G>A
122.12	D.01170
122.54	S.09120|66:A>G
123.39	D.04081
124.25	S.25616|11:T>C
126.01	D.31338
126.87	S.13484|18:A>G
127.19	D.09238
127.21	D.07099, D.12154
129.20	S.21321|53:C>G
129.23	D.12251
130.08	S.19878|59:A>T
130.73	S.01465|36:C>T
132.55	S.14471|61:A>C
132.56	D.17892
132.98	S.15683|37:T>G
133.41	S.07945|G:A>G
134.17	D.12691
134.53	D.17057
134.67	D.16399
135.23	S.19396|24:C>T, S.16868|17:A>T
135.63	S.16132|25:T>C
135.89	S.04979|35:T>C
136.65	S.08758|29:A>C
136.76	D.11298
137.24	D.08552
137.86	S.29094|30:C>T
138.40	D.01267
139.63	D.17410
140.24	D.17590
140.31	D.02196
141.50	S.09170|35:C>T
142.55	S.01908|67:A>C
143.21	S.15878|14:T>G
143.64	D.03547
143.85	D.17322
143.96	D.22192
144.18	D.06954
144.21	D.16038
144.47	S.12531|51:T>C
144.72	D.22493
144.95	D.16585
145.73	S.09697|54:A>G
146.19	S.13147|47:G>A
147.04	D.01245
147.60	S.20108|60:C>A
148.64	S.29311|32:A>G
150.25	S.15964|58:A>G
151.03	D.04652
151.45	D.03535
152.15	D.11232
152.88	D.30808
153.62	S.28364|42:G>A
153.82	S.07720|37:T>G
154.12	mEgCIR3477, S.17498|24:G>A
154.44	D.09878
155.19	S.24627|20:T>C
155.80	S.26844|12:T>A
156.60	S.08069|44:A>T
157.18	S.20651|47:T>C
158.69	D.19819
160.83	S.24639|51:G>C
161.60	S.16089|61:A>G
162.18	S.04762|43:C>T
162.87	D.01963
164.13	D.03544
166.00	D.17896
166.16	S.06832|43:T>A
166.32	D.24271
167.67	D.07692
168.51	S.21065|66:G>A**
171.58	S.14494|12:T>G
172.48	D.11472
173.45	D.05858
173.55	D.07413
174.06	S.14856|17:A>G
174.84	D.22746
176.10	S.13175|38:T>G
176.39	D.08229
177.74	S.19892|5:T>C
178.83	S.20353|9:A>G
180.53	S.23579|39:G>A
184.11	S.24287|21:A>C
187.11	S.26956|36:A>T****

LG 8 (1)

Position	Marker
0	0.05705
2.75	S.06021|30:A>G
8.33	0.19648
9.78	0.19247
9.88	0.23543, 0.09750
10.11	0.06227
11.91	0.04686
13.50	S.12812|56:A>G
15.56	S.26622|19:A>G
17.86	0.10308
18.28	0.05524
18.62	S.20016|64:A>G
19.76	0.03476
20.01	0.04099
21.78	0.16530
22.39	S.17960|8:T>C, S.15686|49:A>C
22.71	0.03592
23.34	0.04786
25.41	S.05438|15:G>A
27.34	S.22028|13:A>T
28.35	0.18604
29.37	S.22780|32:T>A
33.16	0.06567**, 2.05136***
33.24	0.14679***
35.58	mLgCIR3622***
40.35	0.17635****
41.29	0.02218
41.96	0.23868**
42.06	0.10937**
42.67	0.17038**
44.54	0.24476
47.41	S.12859|42:T>C
47.80	0.10687
50.75	0.06129, 0.05778
50.76	0.06704, 0.11401
50.82	0.07755
50.83	0.10418
50.85	0.16285
51.77	0.03851
51.86	0.04205**
57.65	S.78630|13:C>A****
58.43	0.01748**
58.77	0.04761
61.25	0.09470
61.97	0.07708
62.04	0.10041
65.13	0.18915
65.68	0.06541**
66.48	0.07796
66.70	0.06153
67.59	0.08689**
67.81	0.02191
67.85	0.11860**
68.71	0.10481
69.27	0.17699***
70.89	0.00972
72.65	S.27202|64:G>A**
73.86	S.51724|38:T>C**
75.12	0.03901
76.04	0.03634
76.10	0.22125
76.13	0.16986
76.19	0.24130
76.22	0.18105
76.23	0.12139
76.25	0.12508
76.39	0.09302
76.45	0.36706
77.23	0.13041
78.51	S.04255|6:A>T
78.31	0.05105, 0.05287
78.39	0.04630
79.79	0.16338

LG 8 (2)

Position	Marker
79.79	0.16339
80.29	S.10848|32:A>G, S.06268|30:A>G
81.71	0.10989
87.24	0.12720
83.47	0.12354
83.07	0.03070
83.14	S.01727|6:A>G
84.74	0.01529
84.76	0.14112
85.56	0.07789
88.70	S.19013|29:T>C
89.00	0.16883
92.00	S.15325|31:T>C
92.84	S.20889|8:T>A
93.26	S.19|34|37:T>C
93.53	S.04089|6:A>G
94.83	S.17814|39:G>A
96.70	S.22855|24:G>A
98.88	S.35119|12:C>T
101.17	0.13964
102.78	0.01226
102.99	0.14171
104.24	0.01930
104.49	S.04256|43:T>A
104.75	0.10864
104.74	S.14058|32:T>A, S.37758|54:G>A
105.89	S.13230|43:G>A
112.33	S.02341|60:A>G
115.80	S.19046|51:G>A
116.38	0.16519
116.68	S.20867|55:T>C
117.94	0.14196
118.13	0.05550
118.15	0.13675
118.58	0.05065
118.92	0.21726
119.41	0.06154
119.72	S.20167|28:T>G
120.40	S.21914|51:G>A
121.52	S.06057|63:A>G
121.78	0.08058
122.61	S.17694|23:T>C
124.20	S.29502|45:T>C
125.34	S.19836|17:T>G
127.64	S.27512|19:T>G
129.00	0.17635
129.46	S.11245|68:G>A
130.44	S.06557|15:G>A
131.46	S.71093|47:C>T
132.41	S.14874|50:T>C
134.71	S.24516|46:C>A
138.63	0.16099
140.54	0.12567
142.77	mLgCIR2887
141.07	0.14284
141.20	0.00902
141.53	S.53517|12:C>A
42.04	S.05126|9:A>G
44.11	S.04994|25:A>C
143.75	S.13984|64:T>G
144.06	0.04984
144.33	0.02658
144.31	S.14594|22:T>C
144.58	0.13911
145.30	S.34716|8:T>C
145.74	0.22054
145.87	0.06623
146.72	S.38462|49:A>C
147.93	0.16305
148.86	S.11845|20:T>C**
153.18	S.33855|66:A>T

LG 9

Position	Marker
0	S.20806|8:C>A
4.31	0.10844
4.50	mEgCIR3592**
9.50	S.21224|15:G>A
11.23	0.05151**
14.58	0.15418
15.50	S.19515|16:A>G
19.62	0.02715
19.85	0.10529
19.66	0.15144**
19.76	S.21386|48:G>C
21.04	0.02420, 0.06529
21.52	0.04941
21.53	0.13571
22.81	0.05289
23.90	S.13828|83:T>C
26.80	0.13182**
28.12	S.24687|37:G>C
37.23	S.09422|62:C>T
39.55	S.11022|21:A>T
40.77	S.07170|54:G>C**
41.77	S.01623|48:T>C
42.77	S.07786|47:C>C
43.08	0.07232
43.91	0.04698
44.67	0.07655
45.81	S.20397|23:G>A
46.62	S.23767|49:T>G
48.86	S.27284|56:C>A
50.79	S.22141|61:G>A
51.37	S.22624|41:T>C
52.63	S.12541|57:G>A
53.01	S.18008|5:C>T
53.48	S.17588|39:A>T
53.79	S.17806|63:C>T
54.13	S.12024|49:C>T
54.33	0.13056
54.35	0.18839
54.58	S.10473|20:T>C
54.80	S.11987|13:T>C
55.17	0.07089
55.44	S.25612|43:T>C
56.36	S.24413|26:G>T
56.60	mEgCIR5663, S.14007|15:T>G
56.86	S.186146:A>C
57.22	S.14228|32:G>A
57.48	0.05294
57.66	0.06648
57.83	0.12846
58.44	0.16963
58.52	S.18120|13:G>A
58.55	0.19339
58.56	0.05896
58.90	S.22855|54:G>C
59.15	0.08124
59.38	0.04541
60.44	0.13713
60.87	S.25652|35:C>G
61.78	S.19724|68:C>G
61.98	0.13114
62.72	0.06260|37:T>C
63.07	S.06929|16:T>C
63.53	S.01144|33:A>T
64.37	S.01704|17:C>A
65.52	S.25010|27:G>A, S.22655|32:C>G, S.06280|22:T>A
65.63	S.13209|50:C>T, S.06282|8:T>G, S.05953|17:A>C
67.81	S.18088|63:T>C
67.96	S.10666|51:C>C
69.04	S.01246|56:T>G
73.19	S.21117|23:G>C
74.61	S.22057|48:T>A
76.07	S.19392|44:C>G
82.48	0.03220
86.81	0.19782
88.34	0.01134
90.17	S.18953|57:T>C
90.28	0.14973
93.14	S.07948|9:C>T
93.90	S.26398|51:C>A
94.87	0.05090
96.29	S.02166|48:G>A
96.38	S.21700|53:C>G
103.26	S.21514|27:G>A**

LG 11

Position	Marker
0	S.23536|39:A>T
4.43	0.01231
6.79	S.26465|47:C>A
10.03	S.08989|43:G>A
10.75	0.05236
12.67	mEgCIR3362
13.85	0.06542
13.96	0.09555
13.97	0.13267
14.28	S.13935|46:G>C
14.84	S.06465|27:G>A
16.32	0.05595
16.49	0.10321
16.59	0.22830
16.80	0.14996
17.02	S.09107|50:C>A
17.37	S.07642|62:G>A
17.75	S.05984|50:G>A
18.39	0.09039
19.31	0.11199
23.19	S.10240|24:G>A
26.30	0.38047
27.14	0.05787
28.34	S.06200|39:A>C**
30.47	S.08354|40:C>T
31.27	S.23048|24:G>T
32.89	0.08075
33.97	0.17149
34.01	0.02996
35.38	S.29914|43:T>C
36.34	S.18707|57:T>C
37.14	0.12709
40.60	0.05125
41.55	mEgCIR3653
43.95	0.21910
47.64	0.19832
48.99	S.11187|50:G>C
49.36	0.14687
49.40	0.05676
50.51	0.07470, 0.14602
51.02	S.08915|39:A>G
52.61	S.14129|38:A>C
72.66	0.01026
73.41	0.02991
75.77	S.06552|67:C>A
76.79	0.04864
77.25	0.02912
77.29	0.06303
78.39	0.10903
79.97	S.21535|12:T>G

LG 14

Figure 6.2: Genetic linkage map of the 769 population. The map consists of 1,535 markers loci (32 SSR, 836 DArT, 666 SNP and *Sh* locus). Marker names are shown to the right of each LG, with map distances in centiMorgans (Haldane units) to the left. D: DArT marker, S: SNP marker, mEgCIR: *E. guineensis* SSR marker. *: Skewed markers at *p*= 0.1, **: Skewed markers at *p*= 0.05, ***: Skewed markers at *p*= 0.01, ****: Skewed markers at *p*= 0.005, *****: Skewed markers at *p*= 0.001.

Table 6.4: Characteristics of the genetic linkage groups of the mapping population 768.

Linkage group	TM	SSR	DArT	SNP	SD (%)	ML	AMD
1	158	3	87	68	11 (7%)	161.38	1.03
2	89	4	44	41	12 (13.5%)	114.93	1.31
3	45	2	30	13	7 (15.6%)	72.73	1.65
4A	59	0	31	27	1 (1.7%)	81.91	1.41
4B	158	1	77	80	3 (1.8%)	171.20	1.09
5/13	15	0	11	4	0	32.23	2.30
6	55	2	39	14	11 (20%)	103.50	1.92
7	67	2	40	25	2 (3%)	86.65	1.31
8	145	2	86	57	28 (19.3%)	176.66	1.23
9	96	2	33	61	0	105.46	1.11
10A	122	2	67	53	3 (2.5%)	126.58	1.05
10B	48	0	29	19	0	62.53	1.33
11	76	2	55	19	4 (5.3%)	62.09	0.83
12	95	2	43	50	4 (4.2%)	142.13	1.51
13/5	53	0	21	32	0	75.66	1.45
14	122	2	63	57	10 (8.2%)	129.49	1.07
15	70	5	47	18	3 (4.3%)	103.99	1.51
16	82	2	36	44	0	65.69	0.81
Total	**1555**	**33**	**839**	**682**	**99 (6.4%)**	**1874.81**	**-**
Mean	86.39	1.83	46.61	37.89	5.5 (6.4%)	104.16	1.33
Min	15	0	11	4	0	32.23	0.81
Max	158	5	87	80	28 (19.3%)	176.66	2.30

TM = Total number of markers for each linkage group
SD (%) = Number of markers that have significant segregation distortion at $p<0.05$ level and percentage as compared to the total number of markers in the group
ML = Map length in centiMorgans (cM)
AMD = Average marker density in cM

Table 6.5: Characteristics of the genetic linkage groups of the mapping population 769.

Linkage group	TM	SSR	DArT	SNP	SD (%)	ML	AMD
1	157	3	88	66	28 (17.8%)	143.12	0.92
2	86	4	41	41	10 (11.6%)	100.12	1.18
3	55	2	40	13	4 (7.3%)	62.82	1.16
4	226	2	123	100	4 (1.8%)	187.11	0.83
5/13	6	0	3	3	0	42.23	8.45
6	77	2	54	21	6 (7.8%)	111.04	1.46
7	66	2	37	27	11 (16.7%)	85.58	1.32
8	151	2	96	53	19 (12.6%)	153.18	1.02
9	89	2	33	54	6 (6.7%)	103.26	1.17
10A	123	2	65	56	5 (4.1%)	129.62	1.06
10B	44	0	29	15	14 (31.8%)	56.40	1.31
11	51	2	30	19	1 (2%)	79.97	1.6
12	100	2	49	49	1 (1%)	144.96	1.46
13/5	65	0	30	35	9 (13.8%)	49.75	0.78
14	113	2	53	58	17 (15%)	85.74	0.77
15	56	3	38	15	1 (1.8%)	91.91	1.67
16	70	2	27	41	1 (1.4%)	93.80	1.36
Total	**1535**	**32**	**836**	**666**	**137 (8.9%)**	**1720.61**	**-**
Mean	90.29	1.88	49.18	39.18	8.1 (9%)	101.21	1.62
Min	6	0	3	3	0	42.23	0.77
Max	226	4	123	100	28 (17.8%)	187.11	8.45

TM = Total number of markers for each linkage group
SD (%) = Number of markers that have significant segregation distortion at $p<0.05$ level and percentage as compared to total number of markers in the group
ML = Map length in centiMorgans (cM)
AMD = Average marker density in cM

6.3.3.2 Marker distribution among chromosomes

Markers were well distributed over all the linkage groups with various numbers of DArT and SNP markers mapped into individual chromosome linkage groups. The linkage groups formed were found to have similar number of markers in both the 768 and 769 controlled crosses, except for linkage groups 3, 5/13, 6, 11, 13/5 and 15 (Table 6.4). The largest linkage group was LG 4 with a map length of around 200 cM and the

highest numbers of DArT and SNP markers; whereas LG 5/13 had the lowest number of molecular markers, shortest map length and lowest marker density for both mapping populations.

Map distance between two consecutive markers of the 768 population varied from 0 to 15.4 cM with only 57 out of the 1,536 intervals (3.7%) more than 5 cM apart. The largest gaps were observed on LG 13 between DArT markers D.22679 and D.08221, a gap of 15.4 cM in distance. LG 6 has the highest number of intervals with gaps larger than 10 cM, three gaps in total, with distances of 12.9, 12 and 10.4 cM, respectively. Further analysis of the 768 maps showed that several intervals with gaps more than 5 cM were located at the terminal end of linkage groups, for example LG 5 (10.5 cM), LG 10B (6.5 cM), LG 15 (5.5 cM) and both terminals of LG 4B (12.2 and 11.5 cM). Interestingly, the first seven markers at the top terminal end of LG 4B have large intervals gaps of between 7.4-12.2 cM.

Compared to the 768 population, 769 has fewer intervals with gaps greater than 5 cM, only 36 out of the 1,518 intervals in the map (2.4%). Despite that, the two largest intervals between markers were 20 and 16.7 cM in distance and located on LGs 11 and 12, respectively. Again, several intervals with greater map distances were detected at the terminal ends of linkage groups, including LG 1 (12.6 cM), LG 5 (11.7 and 6.8 cM for both ends, respectively), LG 16 (9.2 cM), LG 10A (7.2 cM) and LG 15 (5.8 cM). The smallest linkage group, LG 5, constituting only six markers, had the lowest average marker density of 8.45 cM with map distances ranging from 3.4 to13.6 cM.

During the development of the linkage groups, the same molecular markers were found to be grouped into the same linkage groups between the 768 and 769 controlled crosses. There was some variation in the final map order of these markers within each genetic linkage group, but further analysis is needed to confirm whether these are a true reflection of underlying differences in genetic order, or an effect of limited population size and noise within the dataset.

6.3.3.3   Segregation distortion of markers

A total of 99 and 137 skewed markers were mapped into the final genetic maps of the 768 and 769 controlled crosses, respectively, representing 6.4% and 8.9% of total mapped loci. The 99 distorted markers in the 768 controlled cross were 64 DArT (64.6%), 32 SNP (32.3%) and 3 SSR (3%) whereas the 769 controlled cross had 80 DArT (58.4%), 51 SNP (37.2%) and 6 SSR (4.4%) displaying significant segregation distortion ($p<0.01$).

Distorted markers were not evenly distributed across linkage groups. LG 6 and 8 of the 768 controlled cross and LG 10B and 1 of the 769 controlled cross were the two groups with the highest numbers of markers displaying strong allelic frequency distortion within each controlled cross. They contained 20% (11/55), 19% (28/145), 32% (14/44) and 18% (28/157) of skewed markers, respectively. In contrast, five linkage groups of the 768 controlled cross do not contain any segregation distorted markers; these are LG 5/13, 9, 10B, 13/5 and 16. The 769 map had only one linkage group, LG 5/13, with no marker alleles exhibiting frequencies skewed from their Mendelian expectations.

Distorted markers were quite diffuse within several linkage group while some of the linkage groups displayed clustering of segregation-distorted markers, as might be expected from genuine genetic effects (being aware of that markers closely mapping to a distorted marker are also likely to be distorted). This arrangement is more apparent for linkage groups that have the highest number of markers with significant segregation distortion. Four linkage groups of the 769 controlled cross showed regions of strong allelic frequency distortion, including LG 1, 7, 8 and 10B. Distorted markers of LG 1 of the 769 controlled cross were observed to be distributed into 2 clusters, 12 out of 16 markers from map distances 97.51 to 118.6 cM and 9 out of 13 markers from distance 123.29 to 143.12 cM were significantly distorted. It was also found that 8 out of 9 markers within a short map distance of 8.91 cM (33.16 – 42.07 cM) in LG 8 displayed strong allelic frequency distortion. For LG 7 and 10B, 10 and 12 distorted markers were located within a map distance of 12.67 cM (47.66 to 60.33 cM) and 16.17 cM (40.23 to 56.40 cM), respectively.

The 768 controlled cross had three linkage groups with minor regions of segregation distortion, LG 2, 6 and 8. Skewed markers were located at a different region of LG 8 of the 768 controlled cross, compared to the 769 controlled cross. These 13 distorted markers were located at map distances of 69.33 to 86.53 cM. Meanwhile, clustering of markers with significant segregation distortion was located at map distances 48.49 - 70.97 cM and 34.59 - 64.36 cM of LG 2 and 6, respectively.

6.4   Discussion

Genetic linkage maps are fundamental tools in plant genetic research for understanding the biological basis of complex traits and dissecting genetic determinants underlying the expression of agronomically important traits. This could facilitate marker-assisted selection in breeding programmes in the short term and gene isolation through map-based positional cloning in the long term (Semagn *et al.*, 2006b; Wu *et al.*, 2008). Therefore this chapter reports the construction of the first genetic linkage maps using AAR breeding materials and the DArTSeq method in oil palm. This is useful for the isolation of markers closely linked with the shell-thickness gene as well as other economically important quantitative traits.

6.4.1   Mapping population and population size

Choice of mapping population is one of the most critical criteria in the construction of genetic linkage maps (Collard *et al.*, 2005). Oil palm being an out-breeding species with a long generation cycle and large planting area requirements is particularly difficult to attain suitable mapping population sizes. In addition, most oil palm breeding programmes focus on the assessment of small cross numbers to identify the best families in the trial, rather than individual crosses. In this project, populations derived from the self-pollination of *tenera* palms were employed as mapping populations. The use of self-pollinated *tenera* palms was also reported by Mayes *et al.* (1997) and Singh *et al.* (2010). These controlled self-pollinated populations are segregating for the shell-thickness trait, the trait of interest in the present study. Other researchers have used different mapping populations, for example *tenera* x *pisifera*

248

(Moretzsohn *et al.*, 2000), *dura* x *tenera* (Billotte *et al.*, 2005) and *dura* x *pisifera* (Seng *et al.*, 2011). An interspecific cross of *Elaeis oleifera* and *Elaeis guineensis tenera* palm was also used as a mapping population (Singh *et al.*, 2009).

The large size of oil palm and family-focused breeding approach has rendered limited numbers of crosses available for mapping exercises with sufficient offspring per cross. Most breeding trials have plot sizes of 10-20 palms planted in 3-6 replicates at most (Soh *et al.*, 1990), thus only 40-120 palms are available per cross. For this project, 49 and 59 palms were available from the self-pollinations of the 228/05 and 228/06 *tenera* palms, respectively, from the populations 768 and 769. The population size of these crosses is considered small as compared to the 98 palms of self-pollination of *tenera* A137/30 (Mayes *et al.*, 1997) or 192 palms of T128 *tenera* self-pollinated (Singh *et al.*, 2010). Population sizes of 50-250 individuals are sufficient for genetic mapping studies; however, larger populations are essential for high-resolution mapping and in turn the low numbers affect the power of QTL detection (Mohan *et al.*, 1997). Liu (1998) reported that confidence levels of a detected QTL in a genetic map declined from 90% to 60% when the population size decreased from 100 to 50 individuals, indicating that power for detecting QTL as well as the accuracy of the mapping can be effected.

In order to increase the power of QTL analysis, two genetically-related crosses, 768 and 769, were selected as the mapping populations in this study. Their *tenera* self-pollinated parents, 228/05 and 228/06, are full-sibs derived from the same cross of *tenera* and *pisifera*. Therefore these two parents together with their progenies share

similar genetic backgrounds which might enable combination of traits of interest to increase the significance of detected QTL.

6.4.2   Phase determination

Phase determination of markers, either in coupling or repulsion, is essential prior to map construction. According to Singh *et al.* (2010), a high degree of heterozygosity can be expected in the oil palm genome due to its out-breeding nature. Progenies derived from self-pollinated palms are expected to behave like an F_2 population. The *tenera* and *pisifera* grandparents are no longer available in the field, so it was not possible to determine the phases of markers for the current mapping populations. Similar problems were encountered by Singh *et al.* (2010). To overcome this limitation, they performed their mapping by first analysing the self-pollinated population as a Cross Pollinator (CP) in which linkage phase of markers was determined by the JoinMap software automatically after loci were grouped. The phases of markers were then converted to the F_2 coding and map construction was performed with the self-pollinated populations as F_2 populations. The linkage mapping reported in this chapter was performed following protocol of Singh *et al.* (2010). The same numbers of linkage groups were obtained and markers were grouped similarly for both CP and F_2 phases, indicating that phase conversion of markers was performed successfully in the present study.

During the preliminary stage of mapping, some groups had insufficient linkage when analysed with regression mapping, while mapping using maximum likelihood algorithm gave rise to a map with a gap of 10,000 cM between two distant groups of

markers, indicating repulsion between these two groups of markers on the same chromosome, due to incorrect marker phase. Where this occurred, the problem was resolved by reversing the genotypes codes of one of the linkage group from (a, c) to (b, d) or *vice versa* as has been reported by Nariman (2013) for a segregating F_2 population of Bambara groundnut (*Vigna subterranea*).

6.4.3   Map construction

Preliminary linkage mapping in the present study identified 21 and 17 linkage groups for the 768 and 769 controlled crosses, respectively, a greater number than the 16 haploid chromosomes of oil palm (Maria *et al.*, 1995). This could be due to the relatively small sample size of the F_2 progenies used in this study. Another possible reason could be several areas of the genome remained undetected with the current set of selected molecular markers, due to lack of markers located in those regions or due to lack of polymorphism. The latter might occur if there were regions of the genome which were identical by descent in the parent. Singh *et al.* (2009) also constructed 21 linkage groups from an interspecific cross for QTL analysis of fatty acid compositions in oil palm and a number of published maps for oil palm have indicated that there may be genetic effects leading to breaks in the expected linkage groups. Many mapping studies have emphasized the need for a large number of markers and/or mapping populations to reduce the linkage group numbers to haploid chromosome numbers and increase map accuracy (Sharma *et al.*, 2002; Crane and Crane, 2005; Semagn *et al.*, 2006). In fact, preliminary mapping of the 768 and 769 controlled crosses using a smaller subset of DArT and SNP markers selected under very stringent criteria

produced 24 and 20 linkage groups, respectively (data not shown). The inclusion of more markers to present study successfully reduced the number of linkage groups for both mapping populations, although still more than the haploid chromosomes number. The lack of polymorphic markers in particular chromosomal regions could be due to the marker systems employed in the present study, as stringent criteria were used to screen and select DArT and SNP markers for map construction. For examples, pre-screening and exclusion of highly skewed markers could have potentially eliminated regions of high segregation distortion, leading to no markers bridging region of lower segregation distortion. Selection of more markers using relaxed criteria could be useful to reduce the linkage group numbers to haploid chromosomes number. Increasing the number of SSR markers could also be of useful and the use of multiple marker systems can complement each other (Adawy *et al.*, 2005). Besides, the larger population size of the 769 population (57 palms) as compared to the 768 population (44 palms) might contribute to the lower number of linkage groups obtained, 17 linkage group for the 769 as opposed to 21 linkage groups in the 768 population.

Mayes *et al.* (1997) also commented that using a LOD score of 4 is likely to fragment potentially genuine associations. This is also illustrated by the fact that different linkage groups from the same chromosomes can be linked when lower LOD scores were used (Table 6.3), for example LG 8 of the 768 population were fragmented into three groups under LOD 4 and were regrouped at LOD 2.8 while the two groups of markers from LG 12 of the 769 population were regrouped at LOD 3.7. To some extent this is a consequence of the difficulty of separating large numbers of markers on the basis of a single LOD score, with interaction across groups or noise within datasets

leading to high LODs fragmenting groups. Nevertheless, the possibility that more markers were needed for these particular chromosomal regions cannot be ruled out.

In this project, SSR markers from Billotte *et al.* (2010) were used as anchoring loci to allow two or more linkage groups that belong to the same chromosome to be combined and remapped, bringing the linkage groups nearer to the expected 16 independent groups in both the 768 and 769 mapping populations; the same number as basic chromosome set of oil palm. This approach is not uncommon for mapping studies reported in other plant species (Chetelat *et al.*, 2000; Song *et al.*, 2005; Loridon *et al.*, 2005; Semagn *et al.*, 2006). In rye, a total of 43 linkage groups were initially generated from 1,965 DArT markers, with 367 DArT and 20 SSR markers as anchor loci, these linkage groups were reassembled into 7 larger linkage groups representing rye chromosomes (Bolibok-Bragoszewska *et al.*, 2009). In the present study, the use of SSR markers with known map locations facilitates the assignment of linkage groups to previous published groups and the determination of the orientation of linkage group relative to the reference map. A similar approach was used in rapeseed by Raman *et al.* (2012) and in Brassica by Wang *et al.* (2011).

LG 4A and 4B in the 768 population were not mapped into the same linkage group as in the 769 population, most likely due to the relatively smaller population size of the 768 controlled cross or potentially due to a genetic effect, possibly a translocation, rather than a problem in the methodology. The same applies to LG 10A and 10B in which both parts were linked when a LOD score of 2.5 was used for the 768 population, while a higher LOD score of 3.7 was achieved when part B was linked with

LG 11. During the preliminary genetic linkage mapping of both populations using stringently selected sets of markers, LG 10B was observed to be linked with LG 15 at a much lower LOD score of 1.6 and 2.7 for the 768 and 769 populations, respectively. It is anticipated that increasing population size and/or number of markers will improve the mapping of both regions of chromosomes.

In the present study, no polymorphic CIRAD SSR markers were identified for LG 5, 13 and one terminal end of LG 4, where shell-thickness gene is located. This is not surprising as CIRAD SSR markers were only mapped into 14 linkage groups of the genetic maps of FELDA's breeding material (Seng *et al.*, 2011). The author commented that the lack of complete congruence is due to the very different genetic backgrounds of the planting materials used.

6.4.4   Marker evaluation and distribution

Three different mapping studies of oil palm have successfully generated 16 linkage groups. Among them, the first was published by CIRAD group (Billote *et al.*, 2005) using a *tenera* x *pisifera* cross. This map contains 944 markers, including 255 SSR, 688 AFLP and the *sh* allele, with a total map length of 1,743 cM, average chromosome length of 109 cM and the average marker density of one marker every 1.8 cM. The *tenera* T128 self-pollinated mapping population from MPOB was used to constructed another linkage map consisting of 351 AFLP, 124 RFLP, 17 SSR, 23 SNP and the *sh* trait, a total of 516 markers (Singh *et al.*, 2010). The total length of this map is 1,599.5 cM with an average chromosome length of 100 cM and marker density of one every 3 cM. More recently, FELDA, the biggest oil palm plantation operator in

Malaysia published a genetic linkage map for its high yielding Deli x Yangambi cross (Seng *et al.*, 2011). This integrated map was generated from 479 loci (331 SSR, 142 AFLP and 6 PCR-RFLP). The total map length was 2,247.5 cM with an average length of 137 cM and a marker density of 4.7 cM.

In comparison, the genetic map constructed in this study for 768, an F_2 population, integrated 33 SSR, 839 DArT, 683 SNP and the *sh* allele. This map spanned over 1,874.81 cM with an average linkage group length of 104.16 cM and the average marker density was one marker per 1.33 cM. On the other hand, the genetic map of the 769 population was constructed from 32 SSR, 836 DArT, 666 SNP and the *sh* allele with a total map length of 1,720.61 cM, an average group length of 101.21 cM and an average marker density of one marker every 1.61 cM. In general, genetic maps generated in the present project are larger than the map of MPOB, comparable to CIRAD but shorter than FELDA's. The present study achieved a greater overall marker density compared to the other three previously published mapping projects. The observed differences in length and marker densities for different oil palm genetic linkage maps are expected to be due to differences in recombination frequencies owing to mapping population type and size, marker systems employed, and/or the algorithms and mapping functions used to compute genetic distances (Jing *et al.*, 2009). For instance, Billotte *et al.* (2005) used Kosambi's mapping function, whereas Haldane's mapping function was employed in the present study. The Kosambi mapping function assumes recombination events influence the occurrence of adjacent recombination events (interference) while the Haldane mapping function assumes no interference

between crossover events, thus maps generated using the Kosambi mapping function are shorter than those from Haldane (Collard *et al.*, 2005).

The genetic maps constructed using the 768 and 769 populations shared a number of similarities in terms of grouping of markers in the same linkage groups, location of markers in terms of chromosomes position (telomere vs centromere), total length of map, average map length and average marker density. This is likely to be due to the full-sibs background of their respective 228/05 and 228/06 parents. However, differences in marker order among these two genetic maps were observed. These differences are expected to be due to genetic differences between the two full-sibs parents, although perturbation of the mapping algorithms by missing data or differences in population size/data quality could also have an influence on the final map orders. Each of the siblings will receive different sets of genetic materials from the maternal *tenera* and paternal *pisifera* grandparent palms and there will be regions between the two populations where the alleles inherited are completely different. Emma *et al.* (2009) reported that differences in local recombination frequencies between populations can affect marker ordering between maps. Genetic mapping only gives an indication of the relative positions of the markers to each other which is influenced by the recombination frequencies and hence population size, with accurate local order of markers only achieved with very large populations (Sourdille *et al.*, 2004). Despite relatively small sample sizes of progenies used, this study managed to produce 16 independent linkage groups with high genome coverage and marker density with the large number of markers mapped.

Comparison of SSR markers mapped in FELDA's linkage map (Seng *et al.*, 2011) and the present study revealed that markers grouping were inconsistent in places. Both FELDA's map and maps generated from the present study shared 22 CIRAD SSR markers. Linkage group assignment of half of these shared SSR markers agreed between the maps. For example mEgCIR2215 and mEgCIR3683 in LG 2 and mEgCIR3592 and mEgCIR3592 in LG 9 of present study were found to be located at LG VIII and VII of FELDA's map, respectively. However, discrepancy was observed for the other half of the shared SSR markers. Marker mEgCIR3788 in LG 1 and mEgCIR3477 in LG 4 were located at LG V of FELDA's map. Marker mEgCIR3826 and mEgCIR0446 in LG 10 of the present study were separated into two linkage groups in FELDA's map, LG XII and XVI, respectively. Assignment of SSR markers in the present study was in accordance to CIRAD map but not with the FELDA map.

Despite the high marker density attained in the present study, large gaps were observed between adjacent markers in both the 768 and 769 genetic linkage maps. Large gaps were also reported in other oil palm mapping studies (Billotte *et al.*, 2005; Billotte *et al.*, 2010; Seng *et al.*, 2011). The two largest intervals of FELDA's genetic map were reported to be 26.9 cM in group III and 25.6 cM in group IX (Seng *et al.*, 2011) whereas the two largest gaps of the first microsatellite-based high density oil palm map published by CIRAD were 18 and 14 cM on LG 9 and 12, respectively (Billotte *et al.*, 2005). Regions of low marker density have previously been reported, even on the ultra-dense genetic linkage map with >10,000 loci constructed from a heterozygous diploid potato population (van Os *et al.*, 2006) in which a gap spanning 14 and 20 cM was found on linkage group VIII of the maternal and paternal parental

maps. The authors postulated that this could be either due to recombination hot spots or could also indicate fixation (homozygosity) of the potato genome in that particular region. Castiglioni *et al.* (1999) also commented that large gaps observed between loci could be due to homozygosity of the genome studied or the non-uniform distribution of recombination events as reflected by the mapping algorithm.

In the present study, several of the large intervals were located at the terminal regions of the linkage groups. Similar observations were found within both oil palm genetic maps published by the CIRAD research group (Billotte *et al.*, 2005; Billotte *et al.*, 2010). Studies on maize have revealed that the most severe recombination suppression occurred in centromeric regions with the recombination frequencies of telomeric region up to 100 times higher than centromeric regions (Farkhari *et al.*, 2011). Occurrence of large marker intervals at the terminal region of the linkage groups could be due to non-uniform recombination frequencies or common descent of the regions.

There were no intervals greater than 25 cM in any of the linkage groups between the two maps in the present study, indicating that the maps are relatively homogeneous with regards to marker distribution and is likely to have good coverage of the genome (Singh *et al.*, 2009; Seng *et al.*, 2011). These two maps provide a useful resource for the search and tagging of traits of economic importance.

6.4.5   Segregation distortion

Segregation distortion is defined as the deviation of the observed genotype frequencies from their expected Mendelian segregation ratios. It is reported that the occurrence of segregation distortion is very likely in a population of an out-crossing crop suffering from inbreeding depression due to several cycles of self-pollination (Bolibok-Bragoszewska *et al.*, 2009).

Segregation distortion has been previously reported in others oil palm studies. The level of markers distortion (9.6% and 11% for the 768 and 769 populations, respectively) observed in the present study is slightly higher than the one published by Billotte *et al.* (2005) but lower than the one reported by Singh *et al.* (2009) at 21%. This segregation distortion was also very much lower than that which has been observed in other crops, for example 43.8% and 20.4% for wheat (Jing *et al.*, 2009; Semagn *et al.*, 2006), 42% for tomato (Truong *et al.*, 2010), and 36.7% for rye (Bolibok-Bragoszewska *et al.*, 2009).

Regions of LG 8 of the first microsatellite-based CIRAD genetic maps (Billotte *et al.*, 2005) were noticed to contain clusters of skewed markers. Similar observations were obtained in the present study although distorted markers were found on different regions of LG 8 of the 768 and 769 genetic maps (Figure 6.2 and 6.3). Markers deviating from the expected segregation ratio may be attributed to linkage with closely positioned genes subject to direct selection or displaying lethal alleles, particularly when they were located in a common region of the genome, in this case LG 8 (Billotte *et al.*, 2005; Truong *et al.*, 2010). Examples are segregation distortion in *Populus* spp.

caused by a lethal allele affecting embryo development (Bradshaw and Stettler, 1994) and segregation of markers co-segregating with the Melampsora resistance gene showed significant deviation (Cervera *et al.*, 2001). Truong *et al.* (2010) suggested including distorted markers in the mapping process to avoid missing parts of the linkage groups. Most importantly, it has been confirmed that segregation distortion does not affect the quality of mapping results, both with simulated (Hackett and Broadfoot, 2003) and experimental data (Sharopova *et al.*, 2002; Bolibok-Bragoszewska *et al.*, 2009). In the present study, distorted markers were mapped, except those markers with very significant segregation distortion (at $p<0.0005$) which were excluded from the mapping process. Therefore, 6.4% and 8.9% of distorted markers were mapped in the 768 and 769 populations, respectively.

In conclusion, the present study reported the first high density DArT- and SNP-based genetic maps for both the 768 and 769 populations using the new DArTSeq platform with SSR markers from public database (Billotte *et al.*, 2010) as anchor loci. The genetic maps generated contain 16 independent linkage groups, corresponding well to the 16 homologous chromosome pairs of oil palm. These maps will be useful for the analysis of qualitative and quantitative traits of interest in oil palm.

Chapter 7

Quantitative trait loci (QTL) mapping of economically important traits

7.1   Introduction and objective

Quantitative characters are a common feature of genetic variation in nature, in which traits do not fall into discrete classes but show a continuous range of variation in a population, often with a more or less normal distribution. Many of the commercially important traits in crop plants, such as plant yield and height, exhibit quantitative inheritance. Genetic variation underlying quantitative traits results from segregation of numerous quantitative trait loci (QTL), each explaining a portion of the total variation, and whose expression is modified by interactions with other genes and by the environment (Paran and Zamir, 2003).

The term "Quantitative Trait Loci" (QTL) was first coined by Gelderman (1975) as a region of the genome that is associated with an effect on a quantitative trait. Using molecular markers, QTLs can be described by their chromosomal location, dosage effect, phenotypic effect(s) and sensitivity to the environment (Paterson *et al.*, 1991). One can employ powerful statistical methods to determine likelihood intervals for the locations of QTLs by comparing the alleles inherited at a locus with the average trait value of individuals clustered by the allele version that they carry (Semagn *et al.*, 2010).

QTL mapping of oil palm was first reported by Rance *et al.* (2001) in which QTLs associated with vegetative and yield components were detected and mapped. Billotte *et al.* (2010) performed and tested a QTL analysis designed for multi-parent linkage mapping for traits including fruit yield and its components and measures of vegetative growth. By mapping and analysis in an interspecies cross of *E. oleifera* x *E.*

guineensis, Singh *et al.* (2009) and Montoya *et al.* (2013) published QTL mappings for fatty acid composition. The latest QTL mapping publication in oil palm by Ting *et al.* (2013) identified QTLs associated with callogenesis and embryogenesis of oil palm tissue culture process. All the work reported above represents important developments towards the application of marker-assisted selection (MAS) in oil palm breeding programmes.

This chapter reports an initial analysis of QTLs associated with fruit yield and its components as well as measures of vegetative growth in the 768 and 769 populations by using the high density DArT- and SNP- based genetic linkage maps reported in chapter 6. In view of the small population sizes available in the present study, QTL analysis on two closely-related F_2 segregating populations would allow us to make a comparison and possibly combine any potential QTLs identified in both populations. This is the first QTL mapping study reported on AAR breeding materials. The ultimate objective of mapping QTLs in commercial populations is to utilize molecular breeding strategies such as marker-assisted selection (MAS).

7.2   **Materials and Methods**

As for chapter 5 and 6, closely-related *tenera* self-pollinated 768 and 769 populations were used in the present study for QTL mapping analysis.

7.2.1   **Phenotypic data**

Phenotypic data for 21 yield and vegetative traits were available for both the 768 and 769 segregating F_2 populations. Fruit yield and its components, bunch number and

bunch weight, were individually recorded over two periods: an immature period from 3-5 years after planting and a mature period from 6-10 years after planting. The physical bunch components were recorded at random intervals over 5-13 years after planting for the *tenera* and *dura* palms (as *pisifera* palms are female infertile). Vegetative growth measurements were made for the surviving palms at 10 years old.

7.2.2   Statistical analysis of phenotypic traits

All statistical analyses were performed using Genstat 15th Software (VSN International). The range and distribution of the quantitative data was tested by the normality test of Shapiro-Wilk with an α threshold of 5% (Shapiro and Wilk, 1965). Traits showing non-normal distribution were test-transformed using various transformation approaches (power, log or square root) followed by retesting to determine whether the transformed trait was normally distributed.

All trait data were also explored to determine the significance of the shell-thickness genotype on the measures of quantitative traits using non parametric Mann-Whitney *U* (Mann and Whitney, 1947) or Kruskal-Wallis (Kruskal and Wallis, 1952) test for the two or three fruit types, respectively. If significant differences were detected for a given trait, the individual phenotypic data for the *dura* fruits were corrected based on those of *tenera* fruits by a mean correction as follows:

$$\text{Dura}_{\text{standardized data}} = \text{Dura}_{\text{raw}} + (\text{Tenera}_{\text{mean}} - \text{Dura}_{\text{mean}})$$

The relationships between phenotypic traits, at individual palm level, were estimated by calculating the Spearman rank-order correlation coefficients (Spearman, 1904).
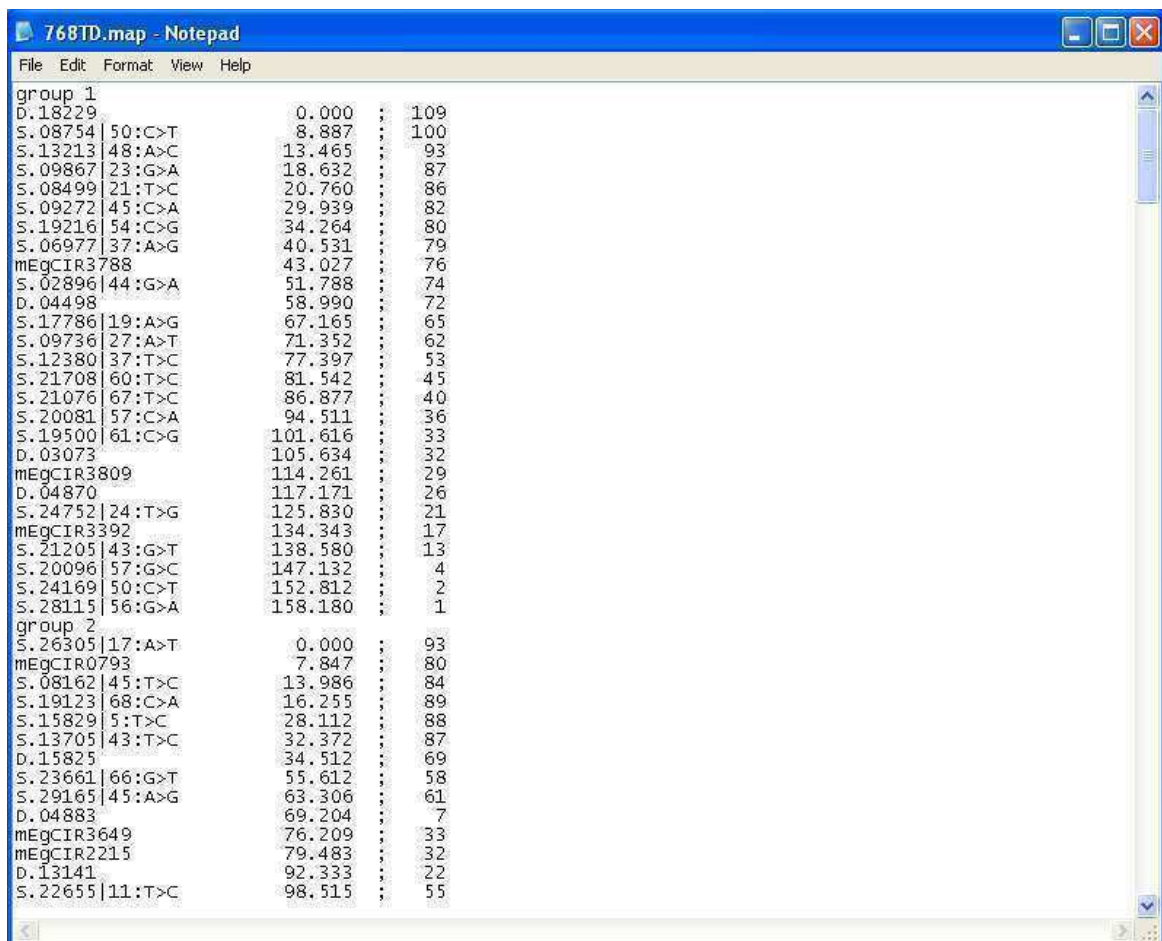
7.2.3   Preparation of data files for QTL mapping

QTL analysis using the constructed genetic linkage maps and all available genotypic and phenotypic data was performed using MapQTL®6 software (Van Ooijen, 2009) for both the 768 and 769 populations. Prior to QTL analysis, a framework genetic map was constructed from the high density DArT and SNP genetic linkage maps generated in Chapter 6. Markers with missing data and/or double recombination events were removed one by one and construction of the genetic map was repeated until a framework map of one marker every 5-10 cM for each linkage group was obtained. By generating a framework map with highest quality spaced markers, any conflicts within the dataset can be resolved and more confidence in the genetic order of markers used to evaluate the quantitative traits gained.

Just like JoinMap, MapQTL uses plain text file to load data to be analysed. Three types of data files are required for QTL analysis:

1) *Locus genotype file* (also called *loc-file*): contains the genotype codes of all the loci of a segregating population. This file has the same format as the one used in JoinMap for the construction of genetic linkage map.

2) *Map-file*: contains the map positions of all loci after generation of the genetic linkage map. Map positions are important in interval mapping as they are used to calculate

265

recombination frequencies which are necessary for calculation of likelihood. Whereas

the non-parametric Kruskal-Wallis quantitative trait analysis method in MapQTL

analyses the loci one by one, with map positions used to sort the loci, but not

imparting any additional information. The *map-file* does not contain any header and

starts with the group number. Loci and their map position must be given in ascending

order on the subsequent lines. Figure 7.1 shows an example of *map-file*.

Figure 7.1: The *map-file* of the 768 controlled cross used for QTL mapping.

3) *Quantitative data file* (also called *qua-file*): contains the data of quantitative traits of

all individuals and has a sequential structure. The file contains three instructions at

the header; follow by names of the traits and data body that contains the information

of each trait for all the individuals (Figure 7.2). The syntax of the three header

instructions are as follow:

ntrt = NTRT

nind = NIND

miss = MISS

where NTRT and NIND are the number of traits and individuals, respectively, and

MISS is the missing value indicator, in this case " * ". NIND must be equal to the

value of NIND in the corresponding *loc-file.*

Figure 7.2: Part of the *qua*-file for the 768 controlled cross for QTL mapping.

7.2.4 **QTL analysis**

7.2.4.1 **Non-parametric Kruskal-Wallis Mapping (K-W)**

As a first step, the non-parametric Kruskal-Wallis (K-W) test was performed for all normally and non-normally distributed traits to identify significant marker-trait associations at $p<0.005$.

Non-parametric mapping makes no assumption of the probability distribution(s) of quantitative traits; hence it is suitable for analysis of both normally and non-normally distributed quantitative traits. The K-W test is regarded as the non-parametric equivalent of one-way analysis of variance. The test ranks all individuals according to the quantitative trait, while it classifies them according to their marker genotype. A segregating QTL (with big effect) linked closely to the tested marker will result in large differences in average rank of the marker genotype classes. A test statistic based on the ranks of the genotype classes is calculated. For individuals in ties, i.e. several individuals have equal values of the quantitative trait, the average rank (mid-rank) is used, while for the test the statistic adjusted for ties is used (indicated by K*) (Lehmann, 1975).

The K-W statistic (K*) is distributed as a chi-square distribution in which the degree of freedom is the number of genotype classes minus one and the significance level (p-value) is indicated in asterisks (* = 0.1, ** = 0.05, *** = 0.01, **** = 0.005, ***** = 0.001, ****** = 0.0005, ******* = 0.0001).

7.2.4.2   Interval Mapping (IM)

A so-called QTL likelihood map or profile is calculated in interval mapping (IM), meaning that for each position on the genome (say every centiMorgan) the likelihood for the presence of a segregating QTL is determined (the likelihood under the alternative hypothesis, H1). This likelihood (H1) is compared to the likelihood for the situation when a locus with zero genetic effects would segregate, i.e. there is no QTL (the null hypothesis, H0). This comparison is done with a likelihood ratio statistic called the LOD (or LOD score).

For all normally distributed trait data in the present study, IM was performed using a LOD statistic test with a mapping step size of 1 cM and a maximum number of five neighboring markers being considered. Framework maps with one marker every 5-10 cM were used as the map file for the interval mapping. Estimates of QTL position were obtained at the point where the LOD score assumes its maximum. For each trait, the genome wide empirical LOD thresholds for QTL detection ($p<0.05$) was estimated using a permutation test (PT) of 10,000 iterations of the trait data. In the present study, LOD score ≥ 3 are presented as a potential/indicative QTL whereas a LOD score $\geq$ the significant threshold value was used to declare a QTL as significant. The confidence interval of each QTL was determined by a one LOD decrease on each side of the LOD peak, representing around a 90% confidence interval.

Luo *et al.* (2003) and Xu (2003) proposed a correction method to correct the bias in overestimation of phenotypic variances associated with identified QTL using limited

269

population sizes and this formula was adopted by Montoya *et al.* (2013) and in the present study.

The variance explained by an identified QTL is, as estimated by MapQTL,

% variance explained = 100 (σ_a^2/σ_p^2)

where σ_a^2 is the genetic variance due to additive effect and σ_p^2 is the phenotypic variance.

The corrected variance explained by this identified QTL was re-estimated as:

% Corrected variance explained = 100 (σ_a^2/σ_p^2) [1- 1/(2Ln(10) x LOD)]

$$= 100 \ (\sigma_a^2/\sigma_p^2) \ [1- 1/(4.605 \ x \ LOD)]$$

where LOD is the LOD value of the identified QTL.

7.3   Results

7.3.1   Statistical analysis of phenotypic data

7.3.1.1   Descriptive statistics of quantitative traits

Quantitative traits were divided into three categories, namely production, bunch components and vegetative growth traits. Table 7.1, 7.2 and 7.3 illustrate the mean, variances, percentage coefficient of variation (CV), maximum value, minimum value and results of the test for normality for all fruit types, *tenera*, *dura* and *pisifera*, respectively, for the 21 phenotypic traits collected from both the 768 and 769 controlled crosses. In

general, quantitative traits were observed to have higher CVs in the 768 controlled cross as opposed to the 769 controlled cross. Production traits of both controlled crosses were associated with CVs between 19.94% to 67.98% while CVs of vegetative growth traits were in the range of 7.93% to 23.57%, indicating a higher degree of variation in both populations for production traits. Indeed, the majority of *pisifera* palms in both controlled crosses are affected by female infertility with no or mininal fruit being borne, contributing to the substantial lower mean value in all production traits, as compared to *tenera* and *dura* palms. As expected, extremely high variation was observed for the *shell to fruit* trait (*SF*) accounted for CVs of 88.71% and 110.30% for the 768 and 769 populations, respectively, due to the distinctive shell-thickness for all three fruit types.

Table 7.1: Descriptive statistics for the production quantitative traits measured in the 768 and 769 F$_2$ populations.

No	Trait	Acronym	Population	Fruit Type	Mean	Variance	CV (%)[a]	Minimum	Maximum	Normality[b]
1	Average bunch number/palm/year at 3-5 years	Bno3_5	768	Total	10.27	42.42	63.45	0.00	21.00	0.9225**
				Tenera	13.87	21.39	33.35	5.33	21.00	
				Dura	12.21	15.99	32.76	7.00	18.33	
				Pisifera	1.42	5.74	168.30	0.00	6.67	
			769	Total	13.08	20.96	35.01	1.00	21.00	0.9705[ns]
				Tenera	14.65	9.31	20.83	7.00	19.67	
				Dura	14.19	35.47	41.96	1.00	21.00	
				Pisifera	8.18	7.46	33.39	2.33	12.00	
2	Fresh fruit bunch yield/palm/year at 3-5 years (kg/palm/year)	FFB3_5	768	Total	29.05	390.10	67.98	0.00	60.07	0.9275**
				Tenera	39.21	229.47	38.63	11.10	59.43	
				Dura	35.15	189.30	39.14	14.50	60.07	
				Pisifera	3.37	34.03	173.00	0.00	17.68	
			769	Total	52.20	579.11	46.10	5.07	105.78	0.9824[ns]
				Tenera	62.21	344.08	29.82	27.63	105.78	
				Dura	57.09	524.17	40.10	5.07	82.62	
				Pisifera	23.05	121.29	47.79	5.48	43.68	
3	Average bunch weight at 3-5 years (kg)	Bwt3_5	768	Total	2.52	1.16	42.71	0.00	4.47	0.9260**
				Tenera	2.85	0.66	28.43	1.45	4.47	
				Dura	2.87	0.45	23.26	1.99	4.25	
				Pisifera	1.51	1.70	86.40	0.00	3.60	
			769	Total	3.87	0.83	23.53	1.60	6.22	0.9850[ns]
				Tenera	4.21	0.48	16.53	3.13	6.22	
				Dura	4.16	0.40	15.23	3.33	5.50	
				Pisifera	2.75	0.50	25.69	1.60	3.78	
4	Average bunch number/palm/year at 6-10 years	Bno6_10	768	Total	13.78	59.51	55.99	0.00	28.80	0.8897**
				Tenera	17.34	26.02	29.42	7.00	28.80	

No.	Trait	Code	N	Category						
				Dura	17.51	12.21	19.96	11.40	25.20	
				Pisifera	2.89	17.97	146.60	0.00	14.60	
			769	Total	13.92	35.15	42.60	0.40	23.80	0.9141**
				Tenera	16.58	6.61	15.51	11.60	21.80	
				Dura	16.90	15.37	23.20	11.60	23.80	
				Pisifera	4.63	11.74	73.99	0.40	10.60	
5	Fresh fruit bunch yield/palm/year at 6-10 years (kg/palm/year)	FFB6_10	768	Total	86.57	2712.00	60.15	0.00	171.48	0.8749**
				Tenera	106.81	1190.00	32.30	32.86	153.78	
				Dura	116.76	879.95	25.41	51.78	171.48	
				Pisifera	14.10	562.90	562.90	0.00	82.78	
			769	Total	127.04	3331.91	45.44	2.44	213.72	0.8740**
				Tenera	157.85	638.25	16.00	115.30	213.72	
				Dura	147.61	728.03	18.28	83.50	177.82	
				Pisifera	32.21	536.86	71.93	2.44	59.92	
6	Average bunch weight at 6-10 years (kg)	Bwt6_10	768	Total	5.80	3.74	33.32	0.00	9.53	0.9304*
				Tenera	6.15	1.85	22.11	4.23	8.86	
				Dura	6.69	1.82	20.14	4.52	9.53	
				Pisifera	4.13	6.04	59.53	0.00	7.33	
			769	Total	8.87	3.13	19.94	4.07	12.41	0.9816^{ns}
				Tenera	9.60	1.50	12.78	6.41	12.41	
				Dura	8.93	3.00	19.40	6.58	12.34	
				Pisifera	7.03	2.87	24.11	4.07	9.92	

[a] Percentage of coefficients of variation

[b] Level of significance corresponding to * $p<0.05$ and ** $p<0.01$, *ns* not significant

Table 7.2: Descriptive statistics for the bunch component quantitative traits measured in the 768 and 769 F$_2$ populations.

No	Trait	Acronym	Population	Fruit Type	Mean	Variance	CV (%)[a]	Minimum	Maximum	Normality[b]
1	Average fruit weight (g)	Fwt	768	Total	11.27	5.88	21.52	6.51	16.99	0.9414[ns]
				Tenera	10.04	2.27	15.02	6.51	13.70	
				Dura	13.61	4.45	15.50	11.14	16.99	
			769	Total	11.06	3.39	16.66	7.77	15.11	0.9671[ns]
				Tenera	10.79	2.86	15.69	7.77	14.79	
				Dura	12.73	4.16	16.02	9.67	15.11	
2	Fruit to bunch ratio (%)	FB	768	Total	64.81	37.53	9.45	53.41	73.52	0.9465[ns]
				Tenera	63.13	36.39	9.56	53.41	72.98	
				Dura	68.00	26.73	7.60	58.44	73.52	
			769	Total	61.14	30.14	8.98	47.01	72.46	0.9729[ns]
				Tenera	60.48	29.45	8.97	47.01	72.46	
				Dura	65.22	18.71	6.63	60.10	70.02	
3	Kernel to fruit ratio (%)	KF	768	Total	5.13	3.07	34.16	2.39	9.68	0.9408[ns]
				Tenera	5.03	3.43	36.85	2.78	9.68	
				Dura	5.33	2.63	30.43	2.39	7.40	
			769	Total	5.13	1.31	22.28	2.79	9.21	0.9142**
				Tenera	5.22	1.42	22.81	2.79	9.21	
				Dura	4.63	0.46	14.69	3.50	5.34	
4	Shell to fruit ratio (%)	SF	768	Total	15.37	185.97	88.71	3.59	37.82	0.7077**
				Tenera	5.76	1.47	21.01	3.59	8.22	
				Dura	33.63	10.42	9.60	29.63	37.82	
			769	Total	9.64	113.15	110.30	2.26	41.31	0.5670**
				Tenera	5.67	4.06	35.56	2.26	12.50	
				Dura	34.26	79.70	26.06	18.75	41.31	
5	Mesocarp to fruit ratio (%)	MF	768	Total	79.50	195.66	17.60	55.98	93.63	0.7729**
				Tenera	89.21	8.66	3.30	82.10	93.63	
				Dura	61.04	13.88	6.10	55.98	67.13	

#	Trait	Abbr	N	Group						
			769	Total	85.23	112.34	12.44	53.99	94.95	0.6336**
				Tenera	89.11	7.92	3.16	80.70	94.95	
				Dura	61.11	79.48	14.59	53.99	76.50	
6	Oil to dry mesocarp ratio (%)	ODM	768	Total	79.20	4.95	2.81	72.60	82.00	0.8591**
				Tenera	79.08	4.67	2.73	72.60	82.00	
				Dura	79.44	5.94	3.07	73.27	81.55	
			769	Total	77.27	3.27	2.34	73.97	81.55	0.9753^{ns}
				Tenera	77.11	2.96	2.23	73.97	81.25	
				Dura	78.25	5.03	2.87	76.00	81.55	
7	Dry to wet mesocarp ratio (%)	DWM	768	Total	69.69	17.30	5.97	63.10	78.30	0.9576^{ns}
				Tenera	67.66	7.76	4.12	63.10	73.20	
				Dura	73.55	13.04	4.91	67.30	78.30	
			769	Total	64.36	26.58	8.01	50.80	75.90	0.9890^{ns}
				Tenera	63.62	23.88	7.68	50.80	71.33	
				Dura	68.90	23.50	7.04	63.70	75.90	
8	Oil to wet mesocarp ratio (%)	OWM	768	Total	55.23	14.37	6.86	49.83	63.57	0.9520^{ns}
				Tenera	53.54	6.61	4.80	49.83	58.56	
				Dura	58.45	14.00	6.40	51.21	63.57	
			769	Total	49.80	23.53	9.74	37.57	61.90	0.9847^{ns}
				Tenera	49.13	20.89	20.89	37.57	57.59	
				Dura	53.96	24.08	9.09	49.69	61.90	
9	Oil to bunch ratio (%)	OB	768	Total	28.12	15.87	14.17	18.85	34.52	0.9694^{ns}
				Tenera	30.12	7.13	8.87	25.01	34.52	
				Dura	24.31	10.53	13.35	18.85	29.38	
			769	Total	25.81	17.02	15.99	17.31	32.54	0.9670^{ns}
				Tenera	26.49	13.75	14.00	18.86	32.54	
				Dura	21.58	19.87	20.65	17.31	28.57	

[a] Percentage of coefficients of variation

[b] Level of significance corresponding to * $p<0.05$ and ** $p<0.01$, *ns* not significant

Table 7.3: Descriptive statistics for the vegetative growth quantitative traits measured in the 768 and 769 F_2 populations.

No	Trait	Acronym	Population	Fruit Type	Mean	Variance	CV (%)[a]	Minimum	Maximum	Normality[b]
1	Average frond length of frond 17 (cm)	FL	768	Total	506.80	1900.00	8.60	430.00	580.00	0.9612^{ns}
				Tenera	498.60	1595.00	8.01	430.00	570.00	
				Dura	512.50	2606.00	9.96	430.00	580.00	
				Pisifera	514.91	1762.89	8.15	458.00	580.00	
			769	Total	481.85	1472.00	7.93	365.00	570.00	0.9762^{ns}
				Tenera	488.90	1143.00	6.91	432.00	570.00	
				Dura	474.20	348.00	3.94	450.00	515.00	
				Pisifera	480.90	3897.00	12.98	365.00	542.00	
2	Average frond dry weight of frond 17 (kg)	FDW	768	Total	2.83	0.34	20.52	1.86	4.23	0.9661^{ns}
				Tenera	2.56	0.20	17.56	1.86	3.54	
				Dura	2.64	0.12	12.90	2.10	3.27	
				Pisifera	3.55	0.17	11.69	2.88	4.23	
			769	Total	2.97	0.49	23.57	1.96	5.53	0.9264^{**}
				Tenera	2.87	0.27	18.13	1.99	3.94	
				Dura	2.67	0.29	20.19	1.96	3.55	
				Pisifera	3.59	0.95	27.10	2.40	5.53	
3	Average frond area of frond 17 (m^2)	FA	768	Total	13.23	3.92	14.96	8.53	16.39	0.9592^{ns}
				Tenera	12.98	3.05	13.45	8.53	15.44	
				Dura	13.44	4.08	15.04	9.33	16.39	
				Pisifera	13.44	5.91	18.08	9.59	15.96	
			769	Total	15.94	6.60	16.11	10.68	20.99	0.9772^{ns}
				Tenera	16.01	6.40	15.81	10.68	20.90	
				Dura	15.27	7.96	18.48	11.63	20.99	
				Pisifera	16.48	6.30	15.23	12.45	19.99	
4	Number of green fronds	GF	768	Total	38.98	22.35	12.13	30.00	49.00	0.9807^{ns}
				Tenera	38.00	13.05	9.51	31.00	45.00	

				Dura	37.08	26.74	13.95	30.00	45.00	
				Pisifera	43.00	14.80	8.95	37.00	49.00	
			769	Total	41.27	20.80	11.05	34.00	53.00	0.9536*
				Tenera	40.23	14.12	9.34	34.00	51.00	
				Dura	41.55	29.67	13.11	34.00	51.00	
				Pisifera	44.10	24.54	11.23	38.00	53.00	
5	Leaf area index	LAI	768	Total	7.11	1.76	18.69	4.61	9.50	0.9642^{ns}
				Tenera	6.78	0.94	14.32	4.71	8.31	
				Dura	6.92	2.87	24.46	4.61	9.50	
				Pisifera	7.91	1.39	14.91	5.95	9.25	
			769	Total	9.03	2.14	16.20	6.19	13.81	0.9712^{ns}
				Tenera	8.83	1.61	14.37	6.19	10.99	
				Dura	8.69	2.39	17.78	6.35	11.88	
				Pisifera	9.98	2.77	16.67	8.02	13.81	
6	Stem height (cm)	Ht	768	Total	229.20	2516.00	21.88	159.00	355.00	0.9386*
				Tenera	207.20	844.90	14.03	159.00	275.00	
				Dura	222.40	2876.00	24.11	160.00	350.00	
				Pisifera	277.27	2193.00	16.89	210.00	355.00	
			769	Total	265.00	2489.00	18.82	187.00	415.00	0.9290**
				Tenera	247.70	1137.00	13.62	187.00	345.00	
				Dura	258.20	2653.00	19.95	195.00	396.00	
				Pisifera	314.10	2789.00	16.82	210.00	415.00	

[a] Percentage of coefficients of variation

[b] Level of significance corresponding to * *p*<0.05 and ** *p*<0.01, *ns* not significant

Table 7.4: Descriptive statistics for the production traits of the 768 and 769 populations after removal of *pisifera* palms.

No	Trait	Acronym	Population	Mean	Variance	CV (%)[a]	Minimum	Maximum	Normality[b]
1	Average bunch number/palm/year at 3-5 years	Bno3_5	768	13.21	19.38	33.32	5.33	21.00	0.9423^{ns}
			769	14.52	15.82	27.39	1.00	21.00	0.9421^{*}
2	Fresh fruit bunch yield/palm/year at 3-5 years (kg/palm/year)	FFB3_5	768	37.61	211.30	38.65	11.10	60.07	0.9570^{ns}
			769	60.81	387.50	32.37	5.07	105.80	0.9905^{ns}
3	Average bunch weight at 3-5 years (kg)	Bwt3_5	768	2.86	0.56	26.12	1.45	4.47	0.9781^{ns}
			769	4.20	0.45	16.02	3.13	6.22	0.9688^{ns}
4	Average bunch number/palm/year at 6-10 years	Bno6_10	768	17.41	20.03	25.71	7.00	28.80	0.9273^{*}
			769	16.66	8.72	17.72	11.60	23.80	0.9791^{ns}
5	Fresh fruit bunch yield/palm/year at 6-10 years (kg/palm/year)	FFB6_10	768	110.70	1061.00	29.42	32.86	171.50	0.8946^{**}
			769	155.10	667.70	16.66	83.50	213.70	0.9868^{ns}
6	Average bunch weight at 6-10 years (kg)	Bwt6_10	768	6.36	1.85	21.38	4.23	9.53	0.9617^{ns}
			769	9.42	1.94	14.80	6.41	12.41	0.9844^{ns}

[a] Percentage of coefficients of variation

[b] Level of significance corresponding to * $p<0.05$ and ** $p<0.01$, *ns* not significant

Table 7.5: Descriptive statistics for several vegetative growth traits of the 768 and 769 populations after removal of *pisifera* palms.

No	Trait	Acronym	Population	Mean	Variance	CV (%)[a]	Minimum	Maximum	Normality[b]
1	Average frond dry weight of frond 17 (kg)	FDW	768	2.59	0.17	15.68	1.86	3.54	0.9815^{ns}
			769	2.81	0.28	18.67	1.96	3.94	0.9714^{ns}
2	Number of green fronds	GF	768	37.64	17.99	11.27	30.00	45.00	0.9655^{ns}
			769	40.59	18.00	10.45	34.00	51.00	0.9556^{ns}
3	Stem height (cm)	Ht	768	213.20	1637.00	18.98	159.00	350.00	0.9145^{*}
			769	250.50	1521.00	15.57	187.00	396.00	0.9135^{**}

[a] Percentage of coefficients of variation

[b] Level of significance corresponding to * *p*<0.05 and ** *p*<0.01, *ns* not significant

Table 7.6: Descriptive statistics of traits showing normal distribution after transformation and the type of transformation applied.

Population	Trait	Transformation	Mean	Variance	CV (%)[a]	Minimum	Maximum	Normality[b]
768	Bno6_10	Not successful	-	-	-	-	-	-
	FFB6_10	Power of 2	13290	41989106	48.76	1080	29405	0.9434^{ns}
	SF	Not successful	-	-	-	-	-	-
	MF	Not successful	-	-	-	-	-	-
	ODM	Not successful	-	-	-	-	-	-
	Ht	Square root	14.54	1.79	9.21	12.61	18.71	0.9414^{ns}
769	Bno3_5	Power of 2	226.4	10570	45.42	1	441	0.9892^{ns}
	KF	Square root	5.13	1.31	22.28	2.79	9.21	0.9433^{ns}
	SF	Not successful	-	-	-	-	-	-
	MF	Not successful	-	-	-	-	-	-
	Ht	Log (base 10)	2.39	0.004	2.663	2.272	2.598	0.9652^{ns}

[a] Percentage of coefficients of variation

[b] *ns* not significant

The results also indicated that all the production traits of the 768 population as well as the *average bunch weight* and *fresh fruit bunch yield* at 6-10 years after planting (*Bno6_10* and *FFB6_10*) of the 769 population deviated significantly from a normal distribution at $p<0.05$ (Table 7.1). As for individual bunch components, *shell to fruit* and *mesocarp to fruit* ratios (*SF* and *MF*) of both populations were found to be none normally distributed, as were *oil to dry mesocarp ratio* (*ODM*) of the 768 and *kernel to fruit ratio* (*KF*) of the 769 population (Table 7.2). *Stem height* at the 9th year after planting (*Ht*) of both populations is also deviated from a normal distribution (Table 7.3). Non-normal distribution data were also observed for two additional vegetative traits of the 769 population, *frond dry weight* (*FDW*) and *number of green fronds* (*GF*).

In view of the female sterility characteristic of *pisifera* palms which is believed to have affected the distribution of the production traits, all *pisifera* individuals were excluded from further analysis and descriptive trait analyses were repeated (Table 7.4). *Pisifera* palms were eliminated from both populations for all production traits so that each trait was consistently described between the two populations. Upon removal of *pisifera* palms, repetition of the Shapiro-Wilk normality test showed that the production traits were normally distributed at an α-threshold of 5%, except for trait *Bno6_10* and *FFB6_10* of the 768 and *Bno3_5* of the 769 population. Using the same criteria, individual *pisifera* palms were also removed from vegetative traits *FDW*, *GF* and *Ht* (Table 7.5). Both *FDW* and *GF* traits of the 768 and 769 populations were normally distributed, but *Ht* was not.

Various transformation methods were tested on traits showing non-normal distribution to transform data into a normal distribution. *Stem height* (*Ht*) of the 768 and *KF* of the 769 populations were successfully transformed into normal distributions using square roots while the production traits *FFB6_10* of the 768 and *Bno3_5* of the 769 populations were transformed using power of two (Table 7.6). Figure 7.3 illustrates an example of the residual plot before and after transformation. As seen in Figure 7.3, the histogram before transformation was bell-shaped with a long left tail that was changed into a better fit after transformation with the normal plot of actual residual against expected value was in an approximate straight line after transformation, indicating the transformed *Bno3_5* trait of the 769 population was normally distributed and had an equal distribution of residuals. Nevertheless, several traits could not be transformed to normal distribution despite numerous attempts. These traits include bunch components *SF* and *MF* of both populations and *ODM* and *Bno6_10* of the 768.
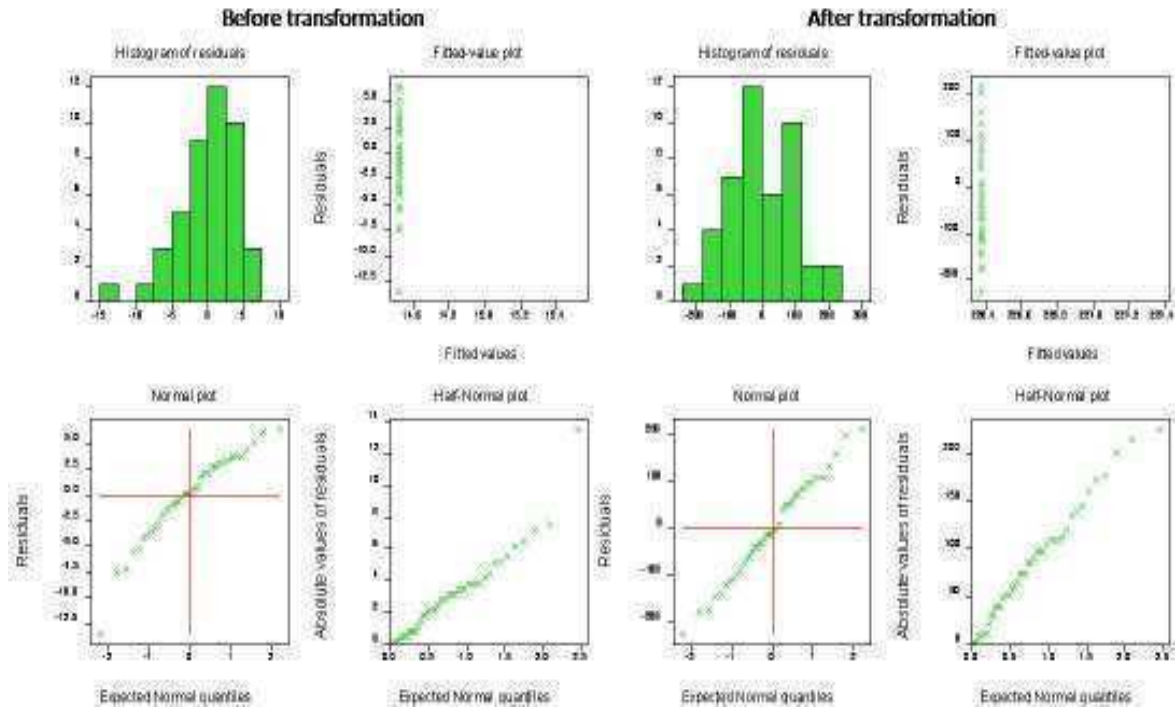
Figure 7.3: Residual plots of trait *Bno3_5* of the 769 population before and after transformation.

7.3.1.2   Effect of the *Sh* gene on quantitative phenotypic traits

Given the distinctive characteristic of oil palm fruit types, it is essential to examine the effect of the *Sh* gene on phenotypic traits and to eliminate this major gene effect that could cause bias in the QTL search results. Indeed, a non-parametric equivalent of the *t*-test, *Mann-Whitney* U test, showed significant mean differences between *tenera* and *dura* fruits for most bunch component traits, except *KF* and *ODM*, at the 5% limit (Table 7.7). The remaining production and vegetative growth traits were not dependent on the fruit type. The result obtained in the two F_2 populations was largely consistent, except for *fruit to bunch ratio* (*FB*), *dry to wet mesocarp ratio* (*DWM*) and *oil to wet mesocarp ratio* (*OWM*), in which significant differences ($p<0.05$) were detected in

the 768 but not the 769 controlled cross. The individual phenotypic data of traits with significant differences between fruit types were corrected as mentioned in section 7.2.2. To maintain the consistency within each trait, *FB*, *DWM* and *OWM* of the 769 controlled cross were also corrected.

Table 7.7: Determination of the *Sh* gene effect on quantitative phenotypic traits measurement.

Traits	Fruit type	768	769
Production traits			
Bno3_5	*Tenera, Dura*	0.311^{ns}	0.682^{ns}
FFB3_5	*Tenera, Dura*	0.439^{ns}	0.825^{ns}
Bwt3_5	*Tenera, Dura*	0.906^{ns}	0.781^{ns}
Bno6_10	*Tenera, Dura*	0.657^{ns}	0.984^{ns}
FFB6_10	*Tenera, Dura*	0.598^{ns}	0.458^{ns}
Bwt6_10	*Tenera, Dura*	0.221^{ns}	0.204^{ns}
Bunch components			
Fwt	*Tenera, Dura*	<0.001**	0.041*
FB	*Tenera, Dura*	0.040*	0.066^{ns}
KF	*Tenera, Dura*	0.286^{ns}	0.282^{ns}
SF	*Tenera, Dura*	<0.001**	<0.001**
MF	*Tenera, Dura*	<0.001**	<0.001**
ODM	*Tenera, Dura*	0.383^{ns}	0.307^{ns}
DWM	*Tenera, Dura*	<0.001**	0.053^{ns}
OWM	*Tenera, Dura*	0.001**	0.082^{ns}
OB	*Tenera, Dura*	<0.001**	0.016*
Vegetative growth traits			
FL	Tenera, *Dura, Pisifera*	0.478^{ns}	0.280^{ns}
FDW	*Tenera, Dura*	0.573^{ns}	0.355^{ns}
FA	Tenera, *Dura, Pisifera*	0.641^{ns}	0.540^{ns}
GF	*Tenera, Dura*	0.616^{ns}	0.480^{ns}
LAI	Tenera, *Dura, Pisifera*	0.084^{ns}	0.128^{ns}
Ht	*Tenera, Dura*	0.630^{ns}	0.626^{ns}

Level of significance corresponding to * *p*<0.05 and ** *p*<0.01, *ns* not significant

The Gaussian distribution of traits was checked with a normality test, Shapiro-Wilk, after correction (Table 7.8). The distribution of all normally distributed traits was maintained after means correction. Two of the non-normal traits of the 768 controlled cross, *SF* and *MF*, turned into normal distributions after correction while the same traits in the 769 and *ODM* of the 768 were still significantly non-normally distributed at *p<*0.05.

Table 7.8: Normality test on bunch component traits after means correction.

Traits	768	769
Fwt	0.9814ns	0.9784ns
FB	0.9630ns	0.9800ns
SF	0.9774ns	0.7809**
MF	0.9577ns	0.8712**
DWM	0.9687ns	0.9728ns
OWM	0.9659ns	0.9824ns
OB	0.9719ns	0.9576ns

Level of significance corresponding to * *p<*0.05 and ** *p<*0.01, *ns* not significant

7.3.1.3   Correlation of quantitative traits

Phenotypic correlations between quantitative production, bunch components and vegetative growth traits were computed for both the 768 and 769 populations at the individual palm level and significant Spearman rank-order correlations between individual traits (*p<*0.05) are presented in tables 7.9 and 7.10, respectively.

Production traits: For both the 768 and 769 populations, a significant positive correlation(*p<*0.05) was found for the same production traits, *Bno*, *FFB* and *Bwt*, between immature (3-5 years) and mature (6-10 years) phases as well as a positive and significant correlation (*p<*0.01) between *FFB* with its respective *Bno* and *Bwt* components. *Bwt3_5*

and *Bwt6_10* of the 768 population showed significant positive correlations ($p<0.05$) with bunch components *KF*, *SF* and negative correlations with *MF*, *ODM*, *OWM*. In contrast, *Bwt3_5* and *Bwt6_10* of the 769 population were positively correlated with vegetative growth traits *FL* and *FDW* instead.

Bunch components: Both populations shared similar a trend of correlations between bunch components traits. The highest positive correlation was observed between *DWM* and *OWM* while the highest negative correlation was between *SF* and *MF* traits. *KF* was positively correlated with *SF* while negatively correlated with *MF*. *OB* exhibited significant positive correlations ($p<0.01$) with *OWM*, *DWM* and *FB*.

Vegetative growth traits: Significant positive correlations ($p<0.01$) were exhibited between *FL* and *FDW* as well as *LAI* and *FA* of both the 768 and 769 populations. Additionally, *LAI* of the 769 population was positively correlated with *FL* and *FDW*.

Table 7.9: Spearman rank-order correlation coefficients between individual phenotypic traits of the 768 population.

Trait	Bno3_5	FFB3_5	Bwt3_5	Bno6_10	FFB6_10	Bwt6_10	Fwt	FB	KF	SF	MF	ODM	DWM	OWM	OB	FL	FDW	FA	GF	LAI
Bno3_5																				
FFB3_5	**0.823**																			
Bwt3_5	0.052	**0.535**																		
Bno6_10	0.404*	0.393*	0.108																	
FFB6_10	0.315	**0.530**	**0.580**	**0.564**																
Bwt6_10	-0.075	0.320	**0.774**	-0.026	**0.664**															
Fwt	0.088	-0.048	-0.252	0.130	-0.004	-0.174														
FB	0.093	0.165	0.212	0.019	0.047	0.247	-0.098													
KF	-0.006	0.182	0.468*	-0.405*	0.036	**0.541**	**-0.550**	0.466*												
SF	0.183	0.445*	**0.660**	0.031	0.398*	0.403*	-0.218	0.365	0.3922*											
MF	-0.135	-0.445*	**-0.732**	0.196	-0.257	**-0.522**	0.394*	-0.455*	**-0.721**	**-0.886**										
ODM	-0.007	-0.345	**-0.646**	0.159	-0.366	**-0.584**	0.295	-0.092	-0.344	-0.302	0.440*									
DWM	-0.090	-0.137	-0.281	0.221	0.039	-0.166	0.272	0.126	-0.306	-0.122	0.303	0.096								
OWM	-0.065	-0.326	**-0.572**	0.201	-0.192	-0.441*	0.401*	0.012	-0.454*	-0.303	**0.521**	**0.550**	**0.839**							
OB	-0.001	-0.150	-0.309	0.223	-0.118	-0.149	0.250	**0.739**	-0.049	-0.155	0.188	0.315	0.491**	**0.550**						
FL	-0.225	-0.204	0.206	-0.148	0.180	0.193	0.054	-0.076	0.000	0.319	-0.164	-0.080	-0.027	-0.088	-0.184					
FDW	-0.381*	-0.234	0.198	-0.273	0.175	0.296	0.107	-0.083	0.016	0.356	-0.212	-0.191	0.214	0.070	-0.171	**0.661**				
FA	0.111	0.251	0.308	0.022	0.251	0.215	-0.404*	-0.035	0.248	0.321	-0.363	-0.246	0.042	-0.110	-0.272	0.221	0.153			
GF	0.253	0.001	-0.432*	-0.085	-0.162	-0.382*	0.137	-0.047	-0.192	-0.086	0.120	0.059	-0.052	0.075	0.059	-0.359*	-0.379*	0.044		
LAI	0.263	0.180	-0.022	-0.065	0.100	-0.047	-0.183	-0.032	0.092	0.173	-0.203	-0.157	0.050	0.014	-0.129	0.000	-0.036	**0.751**	**0.607**	
Ht	-0.024	-0.012	0.122	0.077	0.290	0.200	0.051	0.115	0.104	0.335	-0.250	-0.382*	0.155	-0.061	0.040	0.142	0.020	-0.073	0.102	0.016

Asterisks indicate significant values at $p<0.05$ (*) and $p<0.01$ (**); correlations < -0.5 and >0.5 are highlighted in bold.

Table 7.10: Spearman rank-order correlation coefficients between individual phenotypic traits of the 769 population.

Trait	Bno3_5	FFB3_5	Bwt3_5	Bno6_10	FFB6_10	Bwt6_10	Fwt	FB	KF	SF	MF	ODM	DWM	OWM	OB	FL	FDW	FA	GF	LAI
Bno3_5																				
FFB3_5	0.881**																			
Bwt3_5	0.056	0.458**																		
Bno6_10	0.316*	0.111	-0.410**																	
FFB6_10	0.332*	0.334*	0.070	0.681**																
Bwt6_10	0.042	0.297	0.508**	-0.400**	0.329*															
Fwt	0.027	-0.158	-0.367*	0.019	-0.064	-0.013														
FB	0.093	-0.080	-0.296	0.074	0.072	-0.163	0.111													
KF	-0.136	-0.114	-0.155	-0.009	0.059	-0.001	-0.055	0.131												
SF	0.026	-0.050	-0.187	0.048	0.167	0.128	0.125	-0.067	0.348*											
MF	0.064	0.129	0.240	-0.034	-0.109	-0.064	-0.065	0.085	-0.610**	-0.925**										
ODM	0.010	0.071	0.270	-0.157	-0.107	0.059	0.060	0.014	-0.449**	-0.220	0.324									
DWM	0.054	0.187	0.264	-0.057	0.153	0.171	0.033	-0.005	-0.199	-0.268	0.309	0.596**								
OWM	0.021	0.159	0.292	-0.090	0.095	0.151	0.055	0.017	-0.250	-0.294	0.345*	0.710**	0.979**							
OB	0.089	0.093	0.102	-0.107	-0.028	-0.020	0.040	0.593**	-0.295	-0.507**	0.577**	0.569**	0.650**	0.687**						
FL	0.281	0.391*	0.311*	0.006	0.367*	0.390*	-0.089	-0.064	-0.100	-0.293	0.286	0.218	0.566**	0.539**	0.283					
FDW	0.342*	0.529**	0.451**	-0.077	0.296	0.534**	-0.157	-0.245	-0.177	-0.146	0.203	0.162	0.376*	0.357*	0.145	0.679**				
FA	0.204	0.264	0.128	-0.058	0.288	0.417**	0.114	-0.175	-0.109	-0.056	0.107	-0.058	0.321	0.258	0.008	0.684**	0.628**			
GF	-0.082	-0.194	-0.212	0.164	0.030	-0.240	0.073	0.087	-0.016	0.151	-0.170	0.017	-0.208	-0.202	-0.085	-0.344*	-0.471**	-0.383*		
LAI	0.177	0.167	-0.063	0.066	0.271	0.237	0.168	-0.187	-0.145	-0.016	0.045	-0.029	0.251	0.189	-0.046	0.541**	0.433**	0.837**	0.144	
Ht	0.334*	0.326*	0.184	0.156	0.183	-0.041	-0.321	-0.065	-0.298	-0.186	0.241	0.110	-0.151	-0.124	0.009	-0.029	0.032	-0.079	0.077	-0.078

Asterisks indicate significant values at *p*<0.05 (*) and *p*<0.01 (**); correlations < -0.5 and >0.5 are highlighted in bold.

7.3.2   The construction of the framework maps for QTL analysis

Framework linkage maps with markers spaced every 5-10 cM were generated separately for the 768 and 769 populations to allow analysis of quantitative phenotypic traits. Figure 7.4 illustrates the comparison between the framework maps of the 768 and 769 populations.

Framework maps of the 768 and 769 populations consisted of 340 and 295 markers including the morphological marker for the shell-thickness gene (*Sh*), with total map lengths of 1,843 and 1,753 cM, respectively (Tables 7.11 and 7.12). Due to the co-dominant nature of SNP and SSR markers, both of these markers were preferably selected during the process of constructing the framework maps, hence as high as 73% of total markers in the framework maps were these two marker types.

The average marker density was one marker in every 5.81 and 6.42 cM for the 768 and 769 populations, respectively. Both the framework maps of the 768 and 769 populations have four (1.2%) and seven (2.5%) intervals greater than 15 cM. The greatest intervals were observed on LG 2 of the 768 and LG 6 of the 769 populations, with distances of 21.1 and 27.1 cM, respectively. A total of 145 common markers were obtained between the two framework maps. The locus order of common markers was in general concordant between the two maps except for minor local inversions observed in LG 4, 5/13 and 10.

(768) LG 1

(769) LG 1

(768) LG 4

Position	Marker
0	S.20522|68:A>G
3.77	S.25001|55:T>C
6.15	S.07685|56:C>G
7.84	S.06301|43:C>T
9.04	S.04069|60:C>T
10.04	S.18695|31:T>A
12.37	SH
13.38	S.06735|57:G>A
14.35	S.22560|37:G>C
18.07	0.15745
26.64	0.11065
31.81	S.18672|16:T>C
36.90	S.07830|65:C>A
37.91	0.02902
40.48	S.04102|60:T>C
48.87	S.09220|53:T>A
51.30	S.05222|15:A>G
57.47	S.15448|30:T>G
64.97	S.15876|19:A>T
73.61	S.07674|52:T>C
76.19	S.20125|52:G>A
76.70	S.21204|23:G>A
77.08	0.04921
77.10	0.05823
77.56	S.09456|22:G>T
78.85	S.10468|20:T>C
79.97	S.17703|29:C>A, S.10290|7:T>A
81.29	S.21007|28:T>A
82.73	0.00702
88.92	0.05568
92.85	S.07919|65:G>T
96.90	S.12273|40:G>A
101.64	0.00689
110.57	0.05555
119.11	S.19925|58:T>C
122.89	S.16868|17:A>T
126.69	S.04979|35:T>C
132.95	S.15878|14:T>G
135.85	S.20108|60:G>A
140.72	S.15964|38:A>G
146.43	S.07720|37:T>G
150.30	mEgCIR3477
155.37	S.04762|43:C>T
164.20	S.06852|43:T>A
172.02	S.20355|9:A>B
189.02	S.28068|53:T>C

(769) LG 4

Position	Marker
0	S.25001|55:T>C
2.66	S.06301|43:C>T
6.29	S.20685|53:C>T
7.59	S.18695|31:T>A
9.12	S.04059|60:C>T
10.35	S.06735|57:G>A
11.75	SH
12.43	S.22560|37:G>C
21.05	S.18672|16:T>C
27.50	S.04102|60:T>C
36.51	S.09220|53:T>A
39.30	S.05222|15:A>G
46.17	S.15448|30:T>G
53.00	mEgCIR3526
56.59	0.02443
67.69	S.07214
74.25	0.14223
84.40	0.09240
88.54	0.03497
95.03	S.24917|32:G>A
101.95	S.18252|54:C>G
107.93	S.09456|22:G>T
109.63	S.17703|29:C>A
115.68	S.19468|24:C>T
121.06	S.12273|40:G>A
129.48	0.00689
158.04	S.27060|49:C>G
142.27	0.04081
149.10	S.21321|53:C>G
153.22	S.07945|6:A>G
156.80	S.08758|29:A>C
164.46	S.12531|51:T>C
169.82	S.15964|38:A>G
173.84	mEgCIR3477
180.70	S.16089|61:A>G
190.48	S.14494|12:T>G
196.26	S.13175|38:T>G
202.84	S.24287|21:A>C

(768) LG 5/13

(769) LG 5/13

(768) LG 6

(769) LG 6

(768) LG 11

```
0          D.13027
3.12       D.08335
9.06       D.01231
12.37      D.15846
15.96      S.08989|43;G>A
17.54      mEgDR3362
20.71      S.05984|50;G>A
25.90      D.05787
28.67      S.06354|40;C>T
32.74      D.02996
34.81      mEgDR3653
41.61      S.25562|15;T>C
49.85      D.04884
55.27      S.06552|67;C>A
61.47      S.21535|12;T>G
```

(769) LG 11

```
0          S.23556|39;A>T
5.88       D.01231
12.36      mEgDR3362
17.11      S.09107|50;C>A
22.56      S.10240|24;G>A
25.52      D.05787
29.94      S.23048|24;G>T
32.44      D.02996
40.05      mEgDR3653
48.15      S.11187|50;G>C
50.51      S.08915|39;A>G
75.68      S.06552|67;C>A
83.10      S.21535|12;T>G
```

(768) LG 13/5

```
0          D.17319
15.26      D.06221
23.95      D.11426
28.67      S.02534|24;C>A
32.71      S.14633|10;C>T
38.72      D.05622
43.15      S.17272|47;T>G
47.46      D.07848
50.54      D.03179
55.90      S.13745|36;A>G
59.71      S.08963|17;C>A
63.65      S.14069|28;C>T
70.80      S.02680|67;A>G
75.11      S.23847|24;C>T
```

(769) LG 13/5

```
0          S.22128|29;a>g
6.11       S.02534|24;C>A
10.24      S.14633|10;C>T
13.48      D.05622
18.06      D.04730
23.55      S.17272|47;T>G
28.81      D.07848
36.57      S.13745|36;A>G
41.80      S.16375|53;G>A
45.77      S.07455|42;G>C
47.99      S.02680|67;A>G
```

297

(768) LG 12

(769) LG 12

(768) LG 14

(769) LG 14

Figure 7.4: Comparison of the framework maps constructed for the 768 and 769 populations with the Haldane mapping function. Marker names are shown to the right of each LG, with map distances (in cM) to the left. Common markers between the two maps are linked with a line. D: DArT marker, S: SNP marker, mEgCIR: *E. guineensis* SSR marker.

Table 7.11: Characteristics of the framework genetic map of the 768 population.

Linkage group	TM	SSR	DArT	SNP	ML	AMD	CM (%)
1	27	3	4	20	158.18	6.08	12 (44.4%)
2	22	4	4	14	138.65	6.60	8 (36.4%)
3	13	2	5	6	77.12	6.43	6 (46.2%)
4	47	1	9	36	189.02	4.11	18 (38.3%)
5/13	6	0	3	3	28.97	5.79	4 (66.7%)
6	17	2	10	5	108.15	6.76	7 (41.2%)
7	15	2	3	10	83.33	5.95	9 (60%)
8	27	2	8	17	182.87	7.03	10 (37%)
9	20	2	3	15	95.89	5.05	9 (45%)
10A	22	2	8	12	123.39	5.88	11 (50%)
10B	11	0	5	6	61.01	6.10	6 (54.5%)
11	15	2	7	6	61.47	4.39	7 (46.7%)
12	28	2	5	21	158.79	5.88	10 (35.7%)
13/5	14	0	6	8	75.11	5.78	7 (50%)
14	23	2	4	17	127.70	5.80	7 (30.4%)
15	20	5	5	10	108.24	5.70	8 (40%)
16	13	2	2	9	65.02	5.42	7 (53.8%)
Total	**340**	**33**	**91**	**215**	**1843**	**-**	**145 (42.6%)**
Mean	20.00	1.94	5.35	12.65	108.41	5.81	9.06 (48.08%)
Min	6	0	2	3	28.97	4.11	4 (30.4%)
Max	47	5	10	36	189.02	7.03	18 (66.7%)

TM = Total number of markers for each linkage group

ML = Map length in centiMorgan (cM)

AMD = Average marker density in cM

CM = Number and percentage of common markers between the 768 and 769 populations

Table 7.12: Characteristics of the framework genetic map of the 769 population.

Linkage group	TM	SSR	DArT	SNP	ML	AMD	CM (%)
1	23	3	6	14	139.81	6.35	12 (52.2%)
2	18	4	6	8	97.07	5.71	8 (44.4%)
3	12	2	2	8	74.23	6.75	6 (50%)
4	38	2	7	29	202.84	5.48	18 (47.4%)
5/13	6	0	3	3	42.23	8.45	4 (66.7%)
6	17	2	7	8	124.01	7.75	7 (41.2%)
7	15	2	2	11	90.28	6.45	9 (60%)
8	21	2	9	10	142.18	7.11	10 (47.6%)
9	20	2	3	15	106.96	5.63	9 (45%)
10A	21	2	7	12	136.58	6.83	11 (52.4%)
10B	14	0	8	6	60.81	4.68	6 (42.9%)
11	13	2	3	8	83.10	6.92	7 (53.8%)
12	22	2	6	14	134.94	6.43	10 (45.5%)
13/5	11	0	3	8	47.99	4.8	7 (63.5%)
14	16	2	1	13	89.83	5.99	7 (43.8%)
15	14	3	3	8	104.61	8.05	8 (57.1%)
16	14	2	3	9	74.93	5.76	7 (50%)
Total	**295**	**32**	**79**	**184**	**1752**	-	145 (49.2%)
Mean	17.35	1.88	4.65	10.82	103.08	6.42	9.06 (53.31%)
Min	6	0	1	3	42	5	4 (41.2%)
Max	38	4	9	29	203	8	18 (66.7%)

TM = Total number of markers for each linkage group
ML = Map length in centiMorgan (cM)
AMD = Average marker density in cM
CM = Number and percentage of common markers between the 768 and 769 populations

7.3.3   Quantitative trait loci (QTL) analysis

QTL analysis of important quantitative yield and vegetative traits was performed for both the 768 and 769 populations using Kruskal-Wallis (K-W) and the Interval Mapping (IM) method implemented in the MapQTL6 software. Several QTLs were detected as presented in Tables 7.13 and 7.14. Significant LOD thresholds for QTL determination were estimated at the genome-wide (GW) global risk α of 5% for each trait using a 10,000 permutation test and were found to range from LOD 4.0 to 4.4 and 3.8 to

4.2 for the different traits studied in the 768 and 769 populations, respectively. The LOD thresholds of the 769 population were generally lower than those of the 768 population for the same traits. A QTL was regarded as putative/potential QTL when its LOD score ≥ 3 while those with LOD score ≥ the significant threshold value were declared as significant QTL.

A total of 4 significant and 19 putative QTLs were detected by IM for the 768 population with an average 17.2 cM confidence intervals for the positions of the QTLs (minimum 3 cM and maximum 45.1 cM). An additional 15 markers were also identified from K-W analysis at p=0.005 with significant marker-trait associations (Table 7.13). As for the 769 population, two significant and 13 putative QTLs were detected by IM and 17 extra markers were associated significantly with a particular trait at p=0.005 (Table 7.14). The minimum and maximum confidence intervals for the positions of QTLs identified in the 769 population were 4 cM and 99.2 cM, respectively, with a mean of 22.4 cM. The QTL results of the 21 phenotypic traits collected from both the 768 and 769 populations are presented in the following section individually.

***Average bunch number/palm/year at 3-5 years (Bno3_5)*:** A single putative QTL was detected for the 768 population at position 155.4 cM of LG 4 by IM at a LOD score of 3.9. SNP marker S.04762|43:C>T was located at this position and this significant marker-trait association was confirmed by K-W analysis (K*=11.3, p=0.005) (Table 7.13, Figure 7.5). This QTL explains 39.9% of the total phenotypic variance. No QTLs were detected for the 769 population by either K-W or IM analysis for trait *Bno3_5*.

Table 7.13: QTLs identified by Kruskal-Wallis method (at $p<0.005$) and interval mapping method for the 768 population.

Trait	Acronym	LG[a]	Position (cM)[b]	Marker[c]	Interval Mapping Analysis Maximum LOD[d]	GW (5%)	Additive	Dominance	% Corrected Var[e]	Confidence Interval (cM)	Kruskal-Wallis Analysis K*	Df	Significance level[f]
Average bunch number/palm/year at 3-5 years	Bno3_5	4	155.4	S.04762|43:C>T	3.9	4.2	-3.19	3.84	39.9	153.3-160.4	11.4	2	****
Fresh fruit bunch yield/palm/year at 3-5 years (kg/palm/year)	FFB3_5	4	157.4	S.04762|43:C>T[δ]	3.5	4.3	-5.06	17.21	35.8	154.3-162.4	10.7	2	****
		8	100.5	S.05265|46:A>T[δ]	4.4*		-14.31	5.45	43.3	93.5-110.9	13.0	2	****
Average bunch weight at 3-5 years (kg)	Bwt3_5	3	51.7	S.23581|9:C>T[δ]	3.7	4.2	-0.67	0.71	37.9	35.8-54.7	13.6	2	****
		14	81.7	S.15886|6:T>C[δ]	4.2*		-0.67	0.36	41.9	74.4-91.5	15.6	2	******
		8	64.4	D.24476							8.0	1	****
Fresh fruit bunch yield/palm/year at 6-10 years (kg/palm/year)	FFB6_10	8	99.5	S.05265|46:A>T[δ]	3.0	4.0	-5614.24	2252.40	31.6	61.9-107.0	9.3	2	***
		14	67.4	S.09875|5:G>C							11.7	2	****
Average bunch weight at 6-10 years (kg)	Bwt6_10	10B	50.9	D.02392[δ]	3.1	4.2	0.07	1.77	32.8	43.6-57.4			
		13/5	0.0	D.17319	3.8		1.17	-0.62	38.6	0.0-10.0	10.9	1	*****
		14	79.7	S.15886|6:T>C	4.5*		-1.23	0.57	44.6	75.4-82.7	16.2	2	******
		1	101.6	S.19500|61:C>G							11.5	2	****
			105.6	D.03073							7.9	1	****
Average fruit weight (g)	Fwt	1	7.0	S.08754|50:C>T[δ]	3.3	4.1	1.49	-0.96	38.2	0.0-23.8	10.9	2	****
Kernel to fruit ratio (%)	KF	8	147.6	S.06557|15:G>A[δ]	3.5	4.4	-1.33	-0.89	40.0	123.7-167.7	12.6	2	****
		16	55.0	S.20722|60:G>A[δ]	3.0		-0.17	2.25	35.2	50.1-65.0	10.3	2	***
Shell to fruit ratio (%)	SF	13/5	50.5	D.03179							8.8	1	****
			75.1	S.23947|24:C>T							10.7	2	****

Trait	Abbr	LG	Dist	Marker						Interval			Sig
Mesocarp to fruit ratio (%)	MF	13/5	75.1	S.23947|24:C>T							11.1	2	****
Oil to dry mesocarp ratio (%)	ODM	3	29.1	D.03237							8.4	1	****
		14	87.4	mEgCIR0772							11.9	2	****
			90.5	S.19424|28:C>T							13.8	2	****
			93.5	D.05369							9.5	1	****
Dry to wet mesocarp ratio (%)	DWM	4	56.3	S.15448|30:T>G$^\delta$	3.0	4.3	0.12	3.81	34.9	48.5-60.5	10.2	2	***
		8	121.6	S.19046|51:G>A$^\delta$	3.2		1.01	3.38	36.7	111.9-126.7	11.0	2	****
		12	141.3	S.09183|33:A>G	3.4		-2.98	1.51	39.2	136.9-145.3	10.9	2	****
Oil to wet mesocarp ratio (%)	OWM	3	52.7	S.23581|9:C>T$^\delta$	3.2	4.2	2.45	-4.11	36.9	38.8-55.7	10.0	2	***
		4	57.3	S.15448|30:T>G$^\delta$	3.2		0.65	3.55	37.3	48.9-60.5	11.5	2	****
		8	123.7	S.19046|51:G>A$^\delta$	4.2*		1.51	3.63	45.7	113.9-127.7	14.2	2	*****
Average frond length of frond 17 (cm)	FL	6	32.1	S.09746|51:T>C							10.9	2	****
			37.0	D.14522							8.2	1	****
		7	73.0	S.04110|39:T>C							10.8	2	****
Number of green fronds	GF	1	143.6	S.20096|57:G>C$^\delta$	3.1	4.2	-1.93	-4.23	32.6	130.8-157.8	10.4	2	***
		4	91.9	S.07919|65:G>T$^\delta$	3.3		-3.69	1.41	34.2	83.7-106.6	11.6	2	****
		10A	6.4	S.13650|39:C>A	3.6		-3.25	-2.46	37.3	2.0-24.3	12.4	2	****
Leaf area index	LAI	16	65.0	S.01165|10:T>C	3.0	4.1	-0.33	1.31	25.0	62.0-65.0	11.6	2	****
		14	127.7	S.24206|47:T>C							11.2	2	****
Stem height (cm)	Ht	3	16.3	mEgCIR2518$^\delta$	3.0	4.1	1.09	-1.51	31.8	10.0-25.8	7.6	2	**

[a] LG = Linkage group

[b] Cumulative distance from the top marker of the linkage group

[c] δ = Neighbour locus if not at the QTL position

[d] * = α significance threshold at 5%

[e] Percentage of the phenotypic variance explained at the QTL corrected according to Luo *et al.* (2003)

[f] Significance level of K* values: * = 0.1, ** = 0.05, *** = 0.01, **** = 0.005, ***** = 0.001, ****** = 0.0005, ******* = 0.0001

Table 7.14: QTLs identified by Kruskal-Wallis method (at *p*<0.005) and interval mapping method for the 769 population.

Trait	Acronym	LG	Position (cM)	Marker	Interval Mapping Analysis						Kruskal-Wallis Analysis		
					Maximum LOD	GW (5%)	Additive	Dominance	% Corrected Var.	Confidence Interval (cM)	K*	Df	Significance level
Fresh fruit bunch yield/palm/year at 3-5 years (kg/palm/year)	FFB3_5	15	44.1	mEgCIR3346							10.9	2	****
Average bunch weight at 3-5 years (kg)	Bwt3_5	15	62.0	S.04962|59:A>G							10.8	2	****
Average bunch number/palm/year at 6-10 years	Bno6_10	9	40.7	S.09422|62:C>T[δ]	3.4	4.0	-2.32	-2.16	28.2	23.9-51.3	10.8	2	****
		11	12.4	mEgCIR3362							11.4	2	****
			17.1	S.09107|50:C>A							11.6	2	****
Fresh fruit bunch yield/palm/year at 6-10 years (kg/palm/year)	FFB6_10	14	66.6	S.34182|5:G>A							11.6	2	****
Fruit to bunch ratio (%)	FB	7	23.8	D.02806[δ]	3.1	3.9	3.41	4.17	30.6	14.7-29.9			
		13/5	0.0	S.22128|29:A>G	3.4		3.20	5.06	33.0	0.0-4.0	11.7	2	****
		14	8.7	mEgCIR3350							12.4	2	****
Kernel to fruit ratio (%)	KF	4	149.1	S.21321|53:C>G							11.5	2	****
Shell to fruit ratio (%)	SF	4	10.4	S.06735|57:G>A							8.2	1	****
Mesocarp to fruit ratio (%)	MF	4	153.2	S.07945|6:A>G							10.8	2	****
Oil to dry mesocarp ratio (%)	ODM	1	89.8	mEgCIR3809[δ]	3.4	4.0	2.42	-1.87	33.1	70.6-97.6	6.9	2	**
		9	85.4	D.03220	3.2		-1.33	-0.66	31.1	78.9-90.7	10.1	1	****
Dry to wet mesocarp ratio (%)	DWM	6	0.0	S.28644|63:C>T	3.0	4.0	-0.89	6.35	29.3	0.0-8.0	12.0	2	****
		9	85.7	D.03220							9.9	1	****
Oil to wet mesocarpratio (%)	OWM	6	1.0	S.28644|63:C>T[δ]	3.2	3.8	-0.96	6.45	31.2	0.0-9.0	11.6	2	****
		9	85.7	D.03220							10.1	1	****

Trait	Abbr	LG	Dist	Marker	K*	Thr	Add	Dom		Range	PVE		Sig
Oil to bunch ratio (%)	OB	6	81.2	D.04504							8.1	1	****
Average frond length of frond 17 (cm)	FL	10	91.3	D.30978							7.9	1	****
		A	108.6	D.01967							8.1	1	****
Average frond dry weight of frond 17 (kg)	FDW	6	41.7	D.00853	5.3*	4.0	0.31	-0.46	43.1	38.5-45.7	16.8	1	*******
		10 A	99.9	S.17427|35:C>T$^{\delta}$	3.0		0.27	0.38	26.2	84.2-118.4	10.0	2	***
Average frond area of frond 17 (m^2)	FA	3	36.9	D.15904$^{\delta}$	3.1	4.0	-1.73	-1.11	22.9	24.9-45.1			
		6	46.7	S.29059|24:A>C$^{\delta}$	4.8*		2.03	-0.003	33.7	34.6-60.8	15.8	2	******
		12	20.5	S.16761|49:A>G							11.3	2	****
Number of green fronds	GF	12	11.0	mEgCIR1730$^{\delta}$	3.6	4.2	-3.16	-2.40	30.9	6.0-35.1	10.8	2	****$^{\delta}$
Leaf area index	LAI	3	34.9	D.15904$^{\delta}$	3.4	3.8	-0.94	-0.94	24.8	22.9-41.1	8.9	1	****
		6	50.5	S.29059|24:A>C	3.3		0.92	0.20	23.8	39.5-62.8	12.6	2	****
		10 A	91.3	D.30978							9.0	1	****
Stem height (cm)	Ht	1	108.1	S.24752|24:T>G$^{\delta}$	3.2	4.0	0.05	-0.04	26.4	101.6-112.1	9.9	2	***
		16	43.3	D.04855							8.6	1	****

[a] LG = Linkage group

[b] Cumulative distance from the top marker of the linkage group

[c] δ = Neighbour locus if not at the QTL position

[d] * = α significance threshold at 5%

[e] Percentage of the phenotypic variance explained at the QTL corrected according to Luo *et al.* (2003)

[f] Significance level of K* values: * = 0.1, ** = 0.05, *** = 0.01, **** = 0.005, ***** = 0.001, ****** = 0.0005, ******* = 0.0001

Figure 7.5: Thirty-eight significant and putative QTLs for the yield and vegetative traits identified in the 768 population. Only linkage groups for which "putative" or "significant" QTL were found are shown. Marker names are shown to the right of each LG, with map distances (in cM) to the left. QTL acronym is described in Tables 7.1 to 7.3.

LG 11

0	S.23536|39:A>T
5.88	0.01231
12.36	mEgCIR5362
17.11	S.09107|50:C>A
22.56	S.10240|24:G>A
25.52	0.05787
29.94	S.23048|24:G>T
32.44	0.02996
40.06	mEgCIR3653
48.15	S.11187|50:G>C
50.51	S.08915|39:A>G
75.68	S.06552|67:C>A
83.10	S.21535|12:T>G

Bno 6_10
Bno 6_10

LG 12

0	S.23327|65:A>G
10.03	mEgCIR1730
14.73	S.22184|36:G>T
20.48	S.16761|49:A>G
31.06	S.19847|67:A>G
47.10	0.03964
52.05	0.09501
57.39	S.15938|49:C>G
65.88	0.22534
68.38	S.14142|18:T>A
72.94	S.16259|21:A>G
77.83	S.11057|23:G>C
80.63	mEgCIR0906
83.70	S.18531|54:G>C
96.91	S.10001|49:A>G
99.26	0.05128
107.63	0.13193
111.72	S.18210|30:C>T
114.63	S.15824|41:A>G
118.26	0.15334
124.88	S.09183|33:A>G
134.94	S.21015|40:T>A

FA
GF

LG 13/5

0	S.22128|29:a>g
6.11	S.02534|24:C>A
10.24	S.14633|10:C>T
13.48	0.05622
18.06	0.04730
23.55	S.17272|47:T>G
28.81	0.07848
36.57	S.13745|38:A>G
41.80	S.18375|53:G>A
45.77	S.07455|42:G>C
47.99	S.02680|67:A>G

FB

Figure 7.6: Thirty-two significant and putative QTLs for the yield and vegetative traits identified in the 769 population. Only linkage groups for which "putative" or "significant" QTL were found are shown. Marker names are shown to the right of each LG, with map distances (in cM) to the left. QTL acronym is described in Tables 7.1 to 7.3.

316

Fresh fruit bunch yield/palm/year at 3-5 years (kg/palm/year) (FFB3_5): IM analysis revealed one significant and one putative QTL each on LG 8 and 4, respectively, for trait *FFB3_5* collected from the 768 population (Table 7.13, Figure 7.5). Both QTLs accounted for 43.3% and 35.8% of the total phenotypic variation, respectively. The significant QTL peaked at position 100.5 cM of LG 8 with LOD score of 4.4, slightly higher than the genome-wide permutation test threshold (4.3; 10,000 permutation tests) whereas the putative QTL (LOD score = 3.5) was detected at 157.4 cM of LG 4. A strong marker-trait association was identified between the SSR locus mEgCIR3346 at 44.1 cM of LG 15 and trait *FFB3_5* for the 769 population by K-W analysis (K*=10.9, *p*=0.005) (Table 7.14, Figure 7.6). A maximum LOD value (2.44) was also observed at the nearby location (45.1 cM) with IM analysis, despite not reaching the significance threshold.

Average bunch number/palm/year at 6-10 years (Bwt3_5): Two QTLs were detected by IM analysis for trait *Bwt3_5* on LG 3 and 14 of the 768 population (Table 7.13 and Figure 7.5). Together the QTLs explained about 80% of total phenotypic variation and the estimates of additive effects and dominance variation of both QTLs were in the same direction. These two QTLs were linked with nearby SNP marker S.23581|9:C>T and S.15886|6:T>C, respectively, as was evidenced by significant marker-trait association analysis using K-W mapping (K*=15.6, *p*=0.0005 and K*=13.6, *p*=0.005, respectively). A significant marker-trait association was also identified between D.24476 marker on LG 8 and trait *Bwt3_5* of the 769 population by K-W analysis (K*=8, *p*= 0.005). For the 769 population, S.04962|59:A>G marker on LG 15 was found to be significantly associated

with trait *Bwt3_5* through K-W analysis (K*= 10.8, *p*=0.005) (Table 7.14 and Figure 7.6).

Average bunch number/palm/year at 6-10 years (Bno6_10): Trait data collected from the 768 population was not normally distributed despite transformations attempted, hence only non-parametric K-W analysis was performed and no significant marker-trait association was detected. A putative QTL at position 40.7 cM of LG 9 was identified for the 769 population and this accounted for 28.2% of the total phenotypic variation (Table 7.14 and Figure 7.6). SNP marker S.09422|62:C>T was located nearby to this QTL peak. Two additional markers mEgCIR3362 and S.09107|50:C>A of LG 11 were revealed to be significantly associated with *Bno6_10* trait of the 769 population (K*=11.4 and 11.6, respectively, *p*=0.005) (Table 7.14 and Figure 7.6).

Fresh fruit bunch yield/palm/year at 6-10 years (kg/palm/year) (FFB6_10): A putative QTL for the *FFB6_10* trait of the 768 population was detected on LG 8 with LOD peak at position 99.5 cM and S.05265|46:A>T was the closest marker to the QTL peak (Table 7.13 and Figure 7.5). This SNP marker was also linked to a QTL for *FFB3_5* with both the additive QTL effects in the same direction, corresponding well to the significant positive correlation between *FFB3_5* and *FFB6_10* traits (Table 7.9). Trait *FFB6_10* was also found to be significantly associated with marker S.09875|5:G>C and S.34182|5:G>A on LG 14 of the 768 and 769 populations, respectively, by K-W analysis (K*=11.7 and 11.6, respectively, *p* = 0.005) (Tables 7.13 and 7.14; Figures 7.5 and 7.6).

*Average bunch weight at 6-10 years (kg) (Bwt6_10)***:** A total of three QTLs were revealed by IM analysis of *Bwt6_10* trait data collected from the 768 population. These included one significant QTL on LG 14 and two putative QTLs on LG 10B and 13/5 (Table 7.13 and Figure 7.5). The LOD score of the significant QTL peaked at the location of the S.15886|6:T>C marker and this marker was also associated with the significant QTL of *Bwt3_5* trait. Both QTL results were supported by highly significant marker-trait association detected in K-W analysis (*p*=0.0005). Meanwhile, K-W analysis revealed that two closely-located markers on LG 1, S.19500|61:C>G and D.03073, were also associated with *Bwt6_10* trait of the 768 population (K*=11.5 and 7.9, respectively, *p*=0.005) (Table 7.13, Figure 7.5). There was no QTL or significant marker-trait association detected for the *Bwt6_10* trait of the 769 population.

Average fruit weight (g) (Fwt): Trait data analysis of the 768 population identified a single putative QTL located at 7 cM of LG 1 with LOD score of 3.3 (Table 7.13 and Figure 7.5). SNP marker S.08754|50:C>T was nearest to this locus and the QTL explained 38.2% of total phenotypic variation. Again, no QTLs were detected for the *Fwt* trait analysed in the 769 population.

Fruit to bunch ratio (%) (FB): Both K-W and IM analysis revealed no significant marker-trait association or QTLs for trait *FB* in the 768 population. In contrast, two putative QTLs were identified on LG 7 and 13/5 of the 769 population with LOD scores of 3.1 and 3.4, respectively (Table 7.14 and Figure 7.6). DArT marker D.02806 was closely-linked with the putative QTL on LG 7 while the QTL on LG 13/5 peaked on SNP marker S.22128|29:A>G. Both QTLs accounted for about 64% of the total phenotypic

variation. K-W analysis detected an additional SSR marker on LG 14, mEgCIR3350, to be significantly associated with the trait *FB* (K*=12.4, *p*=0.005) (Table 7.14 and Figure 7.6).

Kernel to fruit ratio (%) (KF): IM analysis of the 768 population detected two putative QTLs controlling the trait *KF*. These two QTLs were located on LG 8 and 16 at 147.6 and 55 cM with LOD scores of 3.5 and 3, respectively (Table 7.13 and Figure 7.5). The closest markers to these putative QTLs were SNP markers S.06557|15:G>A and S.20722|60:G>A for LG 8 and 16, respectively. Meanwhile, K-W analysis of the 769 population estimated a strong marker-trait association between S.21321|53:C>G marker on LG 4 with trait *KF* (K*=11.5, *p*=0.005) (Table 7.14 and Figure 7.6).

Shell to fruit ratio (%) (SF): *SF* trait of both the 768 and 769 populations were analysed by non-parametric K-W analysis due to the non-normal distribution of the trait. It was found that marker D.03179 and S.23947|24:C>T located at 50.5 and 75.1 cM on LG 13/5 of the 768 (Table 7.13 and Figure 7.5) and marker S.06735|57:G>A on LG 4 of the 769 population (Table 7.14 and Figure 7.6) were significantly associated with the *SF* trait (*p*=0.005). It has to be noted that the data for *SF* trait for both populations were corrected for the shell-thickness gene before QTL analysis.

Mesocarp to fruit ratio (%) (MF): MF trait is another non-normally distributed bunch component trait that could only be analysed by non-parametric K-W analysis. Strong marker-trait association was detected for marker S.23947|24:C>T on LG 13/5 of the 768 population (*p*=0.005), which was also strongly associated with *SF* trait (Table 7.13 and

Figure 7.5). This correlates well with the significant negative correlation between *SF* and *MF* traits of the 768 population (Table 7.9). Interestingly, the *MF* trait of the 769 population was found to be strongly associated with S.07945|6:A>G marker on LG 4 (K^*=10.8, *p*=0.005) (Table 7.14 and Figure 7.6). This marker was closely linked with the S.21321|53:C>G marker that was found to be associated with the *KF* trait, corresponding to the significant negative correlation between *MF* and *KF* traits of the 769 population (Table 7.10).

Oil to dry mesocarp ratio (%) (ODM): K-W analysis identified significant association between this trait of the 768 population with D.03237 marker on LG 3 and mEgCIR0772, S.19424|28:C>T and D.05369 markers on LG 14 (*p*=0.005) (Table 7.13 and Figure 7.5). IM analysis of trait data for the 769 population revealed two putative QTLs located at positions 89.8 and 85.4 cM of LG 1 and LG 9 with LOD scores of 3.4 and 3.2, respectively (Table 7.14 and Figure 7.6). The estimated additive effects of both QTLs were in opposite direction, accounting for about 33% and 31% of the total phenotypic variation.

Dry to wet mesocarp ratio (%) (DWM): Three putative QTLs were detected for the trait *DWM* in the 768 population. These QTLs mapped on 56.3, 121.6 and 141.3 cM of LG 4, 8 and 12, respectively (Table 7.13 and Figure 7.5). QTLs on LG 4 and 8 had additive QTL effects in the same direction but opposite to the effect of QTL on LG 12. Only one putative QTL was identified by IM for the 769 population for trait *DWM*. It was located on LG 6 at position 0 cM with S.28644|63:C>T (Table 7.14 and Figure 7.6). Additional DArT marker, D.03220 on LG 9, was found to be significantly associated with the trait

DWM of the 769 population through K-W analysis (*p*=0.005) (Table 7.14 and Figure 7.6).

Oil to wet mesocarp ratio (%) (OWM): IM analysis revealed one significant QTL on LG 8 and two putative QTLs on LG 3 and 4 of the 768 population for trait *OWM*. A SNP marker, S.19046|51:G>A, was the nearest marker to the significant QTL locus on LG 8 and this was supported by a strong association result obtained from K-W analysis (K*=14.2, *p*=0.001) (Table 7.13 and Figure 7.5). This marker together with S.15448|30:T>G on LG 4 were also identified as QTLs linked with the *DWM* trait of the 768 population. Similarly, markers S.28644|63:C>T and D.03220 on LG 6 and 9, respectively, of the 769 population were identified as putative QTLs for both the *OWM* and *DWM* traits and their associated additive QTL effects were in the same negative direction (Table 7.14 and 7.6). This supported the strong positive correlation relationship between the *DWM* and *OWM* trait of both the 768 and 769 populations (Tables 7.9 and 7.10).

Oil to bunch ratio (%) (OB): Non-parametric K-W analysis indicated that significant marker-trait association was found between this trait and the DArT marker D.04504 on LG 6 of the 769 population at 81.2 cM (K*=8.1, *p* = 0.005) (Table 7.14 and Figure 7.6).

Average frond length of frond 17 (cm) (FL): Significant marker-trait association was revealed by K-W analysis between the trait *FL* with markers S.09746|51:T>C and D.14522 on LG 6 and S.04110|39:T>C on LG 7 of the 768 population (Table 7.13 and

Figure 7.5) as well as DArT markers D.30978 and D.01967 on LG 10A of the 769 population ($p = 0.005$) (Table 7.14 and Figure 7.6).

Average frond dry weight of frond 17 (kg) (FDW): Trait data analysis of the 769 population identified two QTLs located on LG 6 and 10A associated with this trait. The QTL on LG 6 was declared as a significant QTL and was evidenced by the strongest marker-trait association detected by K-W analysis (K*=16.8, *p*=0.0001) (Table 7.14 and Figure 7.6). This significant QTL explained about 43% of the total phenotypic variation.

Average frond area of frond 17 (m^2) (FA): One significant and one putative QTL were identified for the trait *FA* of the 769 population on LG 6 and 3 with LOD scores of 4.8 and 3.1, respectively, by IM analysis (Table 7.14 and Figure 7.6). The SNP marker S.29059|24:A>C was the closest marker to the QTL locus on LG 6 and K-W analysis supported a strong marker-trait association (K*=15.8, *p*=0.0005). An additional marker S.16761|49:A>G on LG 12 was also identified by K-W analysis to be significantly associated with this trait at a lower *p*-value (K* = 11.3, *p* = 0.005).

Number of green fronds (GF): IM analysis of trait data for the 768 population revealed three putative QTLs on LG 1, 4, 10A associated with this trait. The LOD score achieved by the three QTLs were 3.1, 3.3 and 3.6 with SNP markers S.20096|57:G>C, S.07919|65:G>T and S.13650|39:C>A closely-linked with the QTL locus, respectively (Table 7.13 and Figure 7.5). A different QTL of the trait *GF* was detected for the 769 population. This putative QTL was located at 11 cM of LG 12 and accounted for approximately 31% of the total phenotypic variation (Table 7.14 and Figure 7.6).

Leaf area index (LAI): A single putative QTL was identified by IM analysis at 65 cM of LG 16 of the 768 population which was linked with SNP marker S.01165|10:T>C (Table 7.13 and Figure 7.5). Marker-trait association by K-W analysis also indicated an association between the trait *LAI* with the S.24206|47:T>C marker on LG 14. Trait data analysis from the 769 population revealed that position 34.9 cM on LG 3 and 50.5 cM on LG 6 were putative QTLs of the *LAI* trait (Table 7.14 and Figure 7.6). Both QTLs explained 24.8% and 23.8% of the total phenotypic variation, respectively. Overlapping of the QTL confidence intervals was observed for traits *FDW*, *FA* and *LAI* on LG 6 with positive additive QTL effects as well as the trait *FA* and *LAI* on LG 3 with negative effects. DArT marker D.30978 on LG 10A was also found to be significantly associated with both *FL* and *LAI* traits of the 769 population by K-W analysis (*p*=0.005) (Table 7.14 and Figure 7.6). This corresponds well with the significant positive correlation between *FL*, *FDW* and *FA* with *LAI* trait of the 769 population (*p*<0.01) (Table 7.10).

Stem height (cm) (Ht): IM analysis of the 768 population detected a putative QTL on LG 3 at 16.3 cM with LOD score of 3 (Table 7.13 and Figure 7.5). SSR marker mEgCIR2518 was linked to this locus which explained 31.8% of the total phenotypic variation. A different putative QTL for the trait *Ht* was found on LG 1 of the 769 population with marker S.24752|24:T>G linked closely to this QTL peak (Table 7.14 and Figure 7.6). Marker-trait association by K-W analysis also revealed significant association between DArT marker D.04855 on LG 16 with the trait *Ht* (K* = 8.6, *p*=0.005).

Despite the initial objective of this chapter being to combine any common QTLs between the two closely related *tenera* self-pollinated populations, none of the QTLs

identified in both populations were common for the same quantitative traits in the initial

QTL analysis reported here.

7.4   Discussion

The majority of important traits in plant breeding, for example yield, height, drought and disease resistance, are quantitative traits that exhibit continuous distribution of phenotypes that do not follow clear patterns of Mendelian inheritance. The study of quantitative traits is complex as these traits are the cumulative effect of many genes and their interaction with the environment or they may be under simpler genetic control, but are significantly affected by environment; hence one cannot infer the genotype from the phenotype (Collard *et al.*, 2005). Construction of high density genetic maps of any species of interest using molecular marker technologies has enabled plant geneticists to detect and estimate the effects of quantitative trait loci through QTL mapping. QTL mapping is a powerful tool for studying the inheritance and genetic architecture of quantitative traits and provides information on number and chromosomal location of QTLs affecting a trait, magnitude and direction of effect of each QTL, dominant and/or additive gene action of each QTL as well as interaction between different QTLs (epistasis) and between genotypes and environments (Semagn *et al.*, 2010).

After construction of two genetic maps using the DArTSeq platform for the 768 and 769 F_2 populations (chapter 6), this chapter reports the QTL study of 21 quantitative yield traits, bunch components and also vegetative growth traits. Due to the small size of both the 768 and 769 populations, this chapter presents a preliminary QTL study. The

size of the controlled crosses is an issue throughout oil palm breeding, due to the schema adopted for classical breeding.

7.4.1 Statistical analysis of quantitative phenotypic traits

Quantitative phenotypic data were available for the yield of fruit and its components, bunch number and bunch weight, fruit bunch components and measures of vegetative growth. A series of statistical investigations have been conducted to study the nature of the trait variation collected from the field. The Gaussian distribution of traits was assessed using a Shapiro-Wilk test and it was found that majority of production and vegetative traits were non-normally distributed when *pisifera* palms were included in the dataset (Tables 7.1 and 7.3). In general, the distribution of traits was normal after the exclusion of *pisifera* individuals (Tables 7.4 and 7.5). *Pisifera* palms are usually female-sterile (Singh *et al.*, 2013). The majority of *pisifera* palms in the present study did not produce fruit and thus contributed significantly to the non-normal distribution of yield traits as well as the lack of available data for various fruit bunch components. Similar observations were reported by Rance *et al.* (2001) in their QTL analysis of yield components of oil palm using a population segregating for the shell-thickness gene. The authors commented that all *pisifera* palms should be discarded from further QTL analysis due to their female-infertile character as well as unequal variances among shell-thickness genotypes, with *pisifera* individuals associated with lower levels of phenotypic variance due to sterility.

In the present study, shell-thickness genotypes (*dura* or *tenera*) were also found to have a significant effect on phenotypic traits *Fwt*, *SF*, *MF* and *OB* of both self-pollinated populations as well as *FB*, *DWM* and *OWM* of the 768 population (Table 7.7). The study by Rance *et al.* (2001) detected the same genotype effect on the phenotypic traits *FB*, *SF*, *MF* and *OB*, but not *Fwt*, and these trait values were corrected prior to interval mapping analysis. QTL analysis on the initial phenotypic data by Billotte *et al.* (2010) identified the strong influence of the *Sh* locus region on traits *Fwt*, *FB*, *MF*, *KF* and palm oil industrial extraction rate (*OER*) with *Sh* effect accounted for as high as 90% of the phenotypic variation in *MF*. No vegetative trait was found dependent on the *dura* and *tenera* variety of the palms, in a similar way to the results reported in the present study. Identification and removal of the *Sh* locus effect from quantitative traits is important so that variation due to the shell-thickness genotypes were accounted for prior QTL analysis (Table 7.8). Nevertheless, it should be noted that corrections of trait values for the shell-thickness genotypes prior to QTL analysis might result in the erroneous allocation of phenotype variance (Rance *et al.*, 2001).

The phenotypic traits studied in the present study were found to have complex relationships with each other. Strong correlations were observed between yield, bunch components and vegetative traits. Fresh fruit bunch yield (*FFB*) is a product of total bunch number (*Bno*) and bunch weight (*Bwt*). Therefore, strong correlations of these yield traits were reported previously by Billotte *et al.* (2010) for both immature and mature phase with a classic negative correlation between *Bno* and *Bwt*, presumably due to source limitation. These strong positive correlations were also detected in the present

327

study but negative correlations were only found between *Bno6_10* and *Bwt6_10*, but not *Bno3_5* and *Bwt3_5*, of both the 768 and 769 population (Tables 7.9 and 7.10). Correlation of *Bno6_10* and *Bwt6_10* for the 768 population was not significant, despite being negatively correlated (Table 7.9).

Yield traits of the 769 population were significantly correlated with various vegetative growth traits (*FDW*, *FL* and *Ht*; Table 7.10). This result was inconsistent with the QTL mapping in multi-parent population by Billotte *et al.* (2010) as well as the correlation result of the 768 population in the present study. No significant correlation between individual vegetative traits and yield traits were reported by Billotte *et al.* (2010) while the yield traits of the 768 population were significantly correlated with fruit bunch components, but not vegetative growth traits (Table 7.9).

The purpose of bunch analysis is to estimate the contents of oil and kernel in the bunch which were then used to calculate oil and kernel yield. The NIFOR method as described by Blaak *et al.* (1963) is the standard bunch analysis protocol that is generally used by most research institutes. Using this method, oil to bunch ratio (OB), the oil extraction ratio at laboratory scale, is calculated as *OB = ODM* x *DWM* x *WM/F* x *FB* (*ODM* = oil to dry mesocarp ratio; *DWM* = dry to wet mesocarp ratio; *WM/F* = wet mesocarp to fruit ratio; *FB* = Fruit to bunch ratio) and *OWM = ODM* x *DWM* (*OWM* = oil to wet mesocarp ratio) (Corley and Tinker, 2003). In the present study, significant positive correlations of *OB* with *OWM*, *ODM*, *DWM*, *FB* were found in both populations, except that the *OB* and *ODM* traits of the 768 was positively correlated, but below the level of significance (Tables 7.9 and 7.10). On the other hand, the trait *MF* was found to

have significant negative correlation with both *KF* and *SF* which correlates well with the results of Billotte *et al.* (2010). This correlation is apparent; within a fruit, mesocarp size decreases when shell and kernel size increases.

Various vegetative growth traits were measured from oil palm leaves at the 10th year after planting, including *frond dry weight* (*FDW*), *frond length* (*FL*), *frond area* (*FA*), *number of green fronds* (*GF*) and *leaf area index* (*LAI*). Significant correlations between traits were found to be inconsistent between the 768 and 769 populations. *LAI* is the product of area per leaf (*FA*), number of palms per hectare and number of leaves per palm (= *GF*) (Corley and Tinker, 2003). Positive correlations between *LAI* and *FA* were observed for both the 768 and 769 populations. Positive significant correlations were also detected for *LAI* and *GF* in the 768 population, but this correlation was not significant for the 769 population. On the other hand, the significant negative correlation of *GF* with *FL*, *FDW* and *FA* suggested that with a limited amount of energy and space available for frond development, the increase in frond number could reduce the length and width growth of individual fronds.

7.4.2 Quantitative Traits Mapping

7.4.2.1 Population size

Population size plays a major role in the power of QTL detection as well as the accuracy and precision of QTL analysis. It is well known that small sample size in quantitative trait loci (QTL) mapping can lead to an underestimation of the QTL number, overestimation of QTL effects and a failure to quantify any QTL interactions (Melchinger

329

et al., 1997; Vales *et al.*, 2005). Beavis (1998) observed that even 200 offspring may be too few for reliable QTL detection. Most published QTL experiments have employed between 100-200 offspring (Semagn *et al.*, 2010).

Choice of population size and marker method employed is often dependent on cost of marker genotyping and/or cost of trait phenotyping. The standard planting density of oil palm in the fields is 143 palms per hectare. Classical breeding trials also involve relatively small number of palms, usually between 60 and 75 per family (Billotte *et al.*, 2010). It is believed that this trial size is sufficient to estimate the characteristics of planting population, allowing selection of the best families, from which the best individuals can be identified for further breeding. However, such Family and Individual Selection (FIS) breeding systems are less appropriate for QTL detection.

In the present study, a total of 44 and 57 legitimate palms were available for the 768 and 769 populations, respectively, and these were further reduced to 33 and 44 palms after removal of the *pisifera* palms. Indeed, Rance *et al.* (2001) remarked that population size would have to be large enough to allow exclusion of *pisifera* individuals if QTL mapping projects were to be performed on a *tenera* self-pollinated population. Alternatively, *dura* x *tenera* or *dura* x *dura* crosses that do not give rise to *pisifera* could be employed. QTL analysis of most traits was performed by exclusion of *pisifera* palms in the present study, except for vegetative growth traits *FL*, *FA* and *LAI*. In view of the limited population size available for the 768 and 769 populations, preliminary QTL analyses were performed independently on both closely-related crosses to test the

possibility of combining potential common QTL markers, if any, to increase the power and accuracy of QTL detection.

7.4.2.2   **Framework maps and QTL detection**

Framework maps with markers spaced every 5-10 cM for the 768 and 769 populations were established in the present study for QTL analysis, instead of using the high density linkage maps reported in chapter 6. A relative sparse framework map with evenly spaced markers is adequate for QTL detection and previous reports have proven that the power of QTL detection was the same for maps with marker spacing of 10 cM compared to highly saturated maps and this detection power was only slightly decreased with marker spacing of 20 cM or even 50 cM (Darvasi *et al.*, 1993). The framework maps constructed for the 768 and 769 populations have an average marker density of one marker every 5.8 cM and 6.4 cM with a range of 4.1 to 7.0 cM and 5 to 8 cM, respectively (Tables 7.11 and 7.12). The total map length of the framework linkage maps is comparable to those of the high density full genetic maps. The marker orders of framework maps and the full maps were generally consistent, except that some local inversion were observed on LG 2, 4, 5, 8, 9, 12 and 16 of the 768 and LG 4, 10B, 12 and 14 of the 769 populations.

With the reasonable spacing of markers and comparable map length, the framework maps constructed in the present study are deemed suitable, if not optimal, for further QTL analysis. The use of spaced, highest quality markers, also avoids potential inflation of map distances because of poor data or conflict between markers. Thus it is

expected to be a truer representation of the underlying 'ideal' map, compared with the high density marker map. A total of 145 markers were common between the two framework maps, allows potential combination of common QTL found between the two closely-related populations.

Both interval mapping (IM) and Kruskal-Wallis (K-W) tests were performed for the QTL analysis. The non-parametric K-W test is (strictly speaking) more appropriate for non-normally distributed traits as the IM test can be biased by deviations from normality and uneven residuals (Montoya *et al.*, 2013). The single-QTL model of interval mapping performs a likelihood ratio test at even positions in the genome, say every centiMorgan, to determine the presence of segregating QTL. The result are plotted as a likelihood-ratio test statistics (LOD scores) against the chromosomal map distance (Van Ooijen, 2009). Collard *et al.* (2005) reported dense markers may pose problems for linkage analysis software to correctly order the marker and can lead to erroneous QTL mapping results. Therefore, a framework map was established and used in the present preliminary QTL mapping.

MapQTL offers a permutation test (PT) for interval mapping to determine the significance threshold of the LOD score based on actual data rather than on assumed normality distributed data (Van Ooijen, 2009). Before permutation tests were widely accepted as an appropriate method to determine significance thresholds, LOD score between 2.0 to 3.0 (most commonly 3.0) were considered as the significance threshold. The significance thresholds in the present study were determined using 10,000 iteration permutation tests and were found to range between LOD 3.8 to 4.4 for different traits of

different mapping populations. QTLs with LOD scores greater than the 5% genome-wide (GW) significant threshold were considered as "significant QTL" for traits studied while QTLs with LOD score ≥ 3 but lower than the calculated threshold value were regarded as "putative QTL".

Kruskal-Wallis (K-W) tests using the framework maps might not identify the most significant marker-trait association compared to high density maps, due to the limited number of markers in the framework maps.

7.4.2.3   QTLs detected

There were no common significant QTL identified between the 768 and 769 populations for any trait. However, common markers could be found at a lower LOD threshold, for example both populations have a peak at the same region of LG 8 for trait *FFB6_10* although the maximum LOD score for the 769 population was only 2.2; The same peaks at LG 6 were also observed for the trait *FDW* with a maximum LOD score of 5.3 for the 769 but a LOD score of only 2.4 for the 768 population (data not shown).

As previously mentioned, limited sample size will cause downward bias of the number of QTL detected but lowering of LOD threshold to allow identification of more "putative" QTL in the current analysis is unfavorable as it will inevitably contribute to substantial Type I error; identification of false positive QTLs.

Comparison of the present study with those previous publications is not straightforward, particularly with the report of Rance *et al.* (2001). The linkage groups

produced by Rance *et al.* for QTL analysis of yield components consisted of 22 linkage groups using solely RFLP markers. Due to the lack of common markers, no direct comparison is feasible between the study of Rance *et al.* (2001) and the present study. The same QTLs on linkage group 11 were identified for both *MF* and *KF* traits in the study of Rance *et al.* (2001) and their additive QTL effects were in different directions, as predicted by the negative phenotypic correlation between these traits. Trait *MF* in the present study was found to have a significant negative correlation with *KF* and *SF* traits for both mapping populations. However, the traits *MF* and *SF* of the 768 population shared the same QTLS.23947|24:C>T on LG 13/5 while the traits *MF* and *KF* of the 769 population were correlated with QTL around the same regions of LG 4.

The framework maps of the 768 and 769 mapping populations consisting of 33 and 32 SSR markers, respectively, with the original aim to have each linkage group with one SSR markers at each end. This was to make comparison with the multi-parent QTL mapping study conducted by Billotte *et al.* (2010) possible. Several traits were found to have QTL on the same linkage groups, either around the same region or different regions of the group. Trait *Bno3_5* of the 768 population was mapped to S.04762|43:C>T of LG 4 with a confidence interval of 153.3 to 160.4 cM, which is 'downstream' of SSR marker mEgCIR3477, whereas the QTL for trait *Bno3_5* across the families identified by Billotte *et al.* (2010) was mapped to a region of LG 4 'upstream' of marker mEgCIR3477. Meanwhile, a common QTL between trait *Bwt6_9* and *Bwt3_5* across the families was revealed to be located in a region in between SSR marker mEgCIR3788 and

mEgCIR3809 of LG 1 by Billotte *et al.* (2010) while QTL of *Bwt6_10* of the 768 population was mapped to a region just 'upstream' of mEgCIR3809 SSR marker.

Both the present study and that by Billotte *et al.* (2010) reported a QTL for trait *Fwt* on LG 1 but the QTL in the 768 population was located at the top of LG 7 while the QTL identified by Billotte *et al.* (2010) was mapped to the other end of LG 7. Similar observations were also obtained for the QTL of the trait *GF* on LG 4. This QTL was mapped directly 'downstream' of SSR marker mEgCIR3477 by Billotte *et al.* (2010) but the QTL in the 768 population was located at least 40 cM distance away from marker mEgCIR3477. Comparison of the QTL mapping of the 769 population with the study by Billotte *et al.* (2010) did not identify common QTL marker and/or QTL marker on the same LG.

The distinctive QTL mapping result obtained from the 768 and 769 populations was not surprising. Although the self-pollinated parents of both mapping populations were derived from the same T x P cross, both parents are expected to have a certain degree of similarity and difference in their genetic makeup. Earlier work by Melchinger *et al.* (1998) using two independent samples of F_2 populations with different sample sizes of 344 and 107 detected 107 and 39 QTLs, respectively, of which only 20 were common. In the multi-parental QTL mapping of oil palm by Billotte *et al.* (2010), the authors reported that only one QTL was significantly present in three out of the four crosses and all the other significant QTLs were specific to one or another of the crosses while another 16 out of 44 QTLs detected by the across-family model were not identified by the within-family analyses. The authors commented that small numbers of individuals per cross

335

contributed to the fact that QTLs could only be identified by one model but not the other. Both publications imply that sample size indeed played a significant role in the power of QTL detection and even with large numbers of individuals, the statistical power of QTL detection remains modest for QTLs with limited effects.

Most of the QTLs detected for the 768 population were for yield traits while most QTLs for the 769 population were for vegetative growth traits. The 769 population has a relatively larger sample size, particularly with vegetative growth traits *FL, FA* and *LAI* with *pisifera* palms included. There is a big difference in terms of sample size between the yield traits of the 768 and the vegetative growth traits of the 769 populations, 33 against 44 or 57 palms. The power and accuracy of QTL detection would be slightly higher for vegetative growth traits of the 769 population.

The small sample size can also lead to the overestimation of the additive variance associated with correctly detected QTL. The bias can be due to sampling error and Beavis effect (Beavis 1994, 1998). The bias due to sampling error which is a contribution of the environmental variance to the estimate of the additive variance of QTL could be corrected as suggested by Luo *et al.* (2003) and Xu (2003) and adopted by Montoya *et al.*(2013) and the present study. However, the major part of the overestimation is due to the Beavis effect itself and cannot be corrected. Indeed, using the approach suggested by Luo *et al.* (2003), the correction of bias is minimal, in the range of only 2% to 2.7% and 1.6% to 2.3% for QTLs identified in the 768 and 769 populations, respectively. The Beavis effect refers to the overestimation of the effect size of the QTL as a result of small sample sizes in QTL studies. In a simulation study, Beavis (1998) showed that average

336

estimates of the phenotypic variances associated with correctly identified QTL were greatly overestimated if only 100 offspring were evaluated, slightly overestimated if 500 offspring were evaluated and fairly close to the actual magnitude when 1000 offspring were evaluated. When the sample size was small, say 100, the statistical power of detecting a small QTL was as low as 3% and the estimated effects were typically inflated 10-fold. This phenomenon has since been called the Beavis effect.

Lande and Thompson (1990) discussed the bias of QTL effects estimated and suggested performing QTL mapping with one data set and based on the information obtained estimate QTL effects in an independent data set so as to obtain unbiased estimates of QTL effects. QTL mapping using different population sizes conducted by Melchinger *et al.* (1998) revealed the large upward bias of estimates of the QTL effects and the authors agreed that inflated QTL effects could result in an overly optimistic assessment of the efficiency of marker-assisted selection.

In conclusion, this chapter has reported preliminary QTL mapping of yield traits, bunch components and vegetative growth traits of two small populations. No common QTL were identified in these closely-related F_2 populations. Due to the small sample size available, interpretation of the results obtained should be done with caution and further validation/analysis is needed to confirm the accuracy of QTL detected in larger populations as well as the estimated phenotypic variance explained by the QTLs identified. Upon validation, the identified QTLs would be useful for marker-assisted recurrent selection (MARS) of oil palm breeding in which phenotypic evaluations of

crosses can be eliminated, accelerating the time per selection cycle to only 6 years compared to 19 years per cycle of conventional phenotypic selection (Wong *et al.*, 2008).

Chapter 8

Study of the Shell-thickness region

in oil palm

8.1   Introduction and objective

Oil palm fruits can be divided into three different types, which are thick-shell *dura* fruits, shell-less, but female-sterile, *pisifera* fruits and the hybrid *tenera*. *Tenera* fruits have an intermediate fruit form, thinner shell with a greater proportion of mesocarp and a distinct fibre ring around the shell. The later allows unambiguous identification of *tenera* fruits in the field, as there is substantial variation for shell-thickness and the ranges of *dura* and *tenera* shell-thicknesses from different germplasm sources overlap.

In oil palm, the majority of agronomically important traits are controlled by multiple genes (quantitative), except for the shell-thickness trait. The shell gene which controls fruit type shows monogenic co-dominant inheritance (Beirnaert and Vanderweyen, 1941). In *tenera* fruit, 30% of the shell in a *dura* fruit is replaced by mesocarp which contributes to a 30% increase in oil yield, as compared to the *dura* fruit (Corley and Lee, 1992). Therefore *tenera* palms are the most commercially cultivated oil palm genotype and the basis for modern oil palm breeding through recurrent selection of maternal *dura* pools and paternal *pisifera* pools (Soh and Hor, 2000).

Various genetic mapping exercises on oil palm using different molecular markers such as RFLP (Mayes *et al.*, 1997), RAPD (Moretzsohn *et al.*, 2000) and AFLP together with SSR (Billotte *et al.*, 2005) have identified molecular markers linked to the *Sh* gene at genetic distances ranging from 4.7 to 23.9 cM.

This chapter reports the study of the shell-thickness region through saturation of the genetic linkage maps with the aim of identifying closely-linked shell-thickness marker(s) which could be used in marker-assisted selection for early selection of fruit type.

8.2 Materials and methods

Genetic grouping of the 768 and 769 populations was repeated with all available DArT and SNP markers generated from the DArTSeq platform and the linkage groups containing the *Sh* locus were identified for further mapping analysis, to try to saturate this shell-thickness region, using the JoinMap 4.1 Software. After obtaining saturated linkage maps around the *Sh* region separately for both mapping populations, map integration was performed by selecting both group nodes that contain the *Sh* region. Map calculation of integrated map was based on mean recombination frequencies and combined LOD scores. Recombination frequencies and LOD scores were estimated for each pair of markers in individual maps, which in turn were used to calculate the virtual number of recombinant and non-recombinant gametes in each population. Mean recombination frequency and combined LOD scores were obtained by totalling the numbers of recombinant and non-recombinant gametes in both the 768 and 769 populations.

DArT and SNP markers mapped within a 5 cM flanking region of the *Sh* gene were identified from the saturated individual maps of both the 768 and 769 populations, as well as the integrated map. Homology search using the 64 bp sequence tag associated with each DArTSeq marker was performed against MPOB *pisifera* genome

assembly V5, as well as with the retroelement databases Repbase and TIGR Plant Repeats.

8.3   Results

In order to gain a better understanding of the shell-thickness region, genetic mapping was repeated with all available DArTSeq markers and higher density maps around this region were obtained for both the 768 and 769 populations (Figure 8.1). Overall, the majority of markers flanking the *Sh* gene were common to the two populations, although different local marker order and map distances were observed. By using mean recombination frequencies and combined LOD scores, an integrated map was generated for the *Sh* region (Figure 8.1). Marker order in the integrated map was different from that of the individual maps of the 768 and 769 populations and markers were more densely arranged on one side of the *Sh* region than the other.

A total of 32 DArT and SNP markers were identified as flanking the *Sh* gene within 5 cM. Homology search of these markers against the MPOB *pisifera* genome assembly revealed that despite the short sequence of the marker tags (64 bp), significant homology (E-value $\leq 10^{-25}$) were obtained for all markers, except for one DArT marker D.08807 with no hit. All the markers with significant homology has only a single hit with no sub-alignment score produced, indicating the markers were aligned to a single locus in the available genome sequence. Furthermore, 23 out of the 32 markers (72%) were found to be located on the same scaffold p5_sc00060, another three markers were on scaffold p5_sc00263 and four others markers were aligned to orphan contigs (Table 8.1). Orphan contigs are contigs that cannot be assembled to any scaffold. The

homology search of the 32 markers against Repbase and TIGR plant repeats databases returned no significant hits even at E-value of 10^{-5}. This indicates that DArT system generated markers from gene-rich region of the genome, as would be expected for a marker system based on the methylation-sensitive restriction endonuclease, *Pst*I.

Closer inspection of the hit region of the DArTSeq markers against p5_sc00060 and p5_sc00263 scaffolds enabled identification of a sequential arrangement of markers according to the scaffold sequence order (Figure 8.2). It was observed that the overall arrangement of markers on the saturated linkage maps was broadly consistent with the scaffold sequence order, but with considerable local inconsistency. In the present study, the *Sh* gene was found to be located within scaffold p5_sc00060, nearer to one end of this scaffold and next to the second scaffold p5_sc00263. As MPOB had anchored the *pisifera* genome assembly to their T128 genetic linkage map, scaffolds p5_sc00060, p5_sc00263 and p5_sc00051 were revealed to be located at pseudochromosome PLG04, sequentially ordered as the first, second and third scaffold. The arrangement of DArTSeq markers on the present saturated genetic maps corresponds well with the scaffold arrangement on PLG04.

Meanwhile, detailed examination of locus arrangement within linkage group 4 around shell-thickness region of both the 768 and 769 populations revealed potential mis-phenotyping of fruit forms. Figure 8.3 shows that sample no. 55, 769/58, was phenotyped as *pisifera* in the field (genotype "a"). However according to molecular marker mapping, the correct genotype should be "h", a *tenera* fruit form instead.
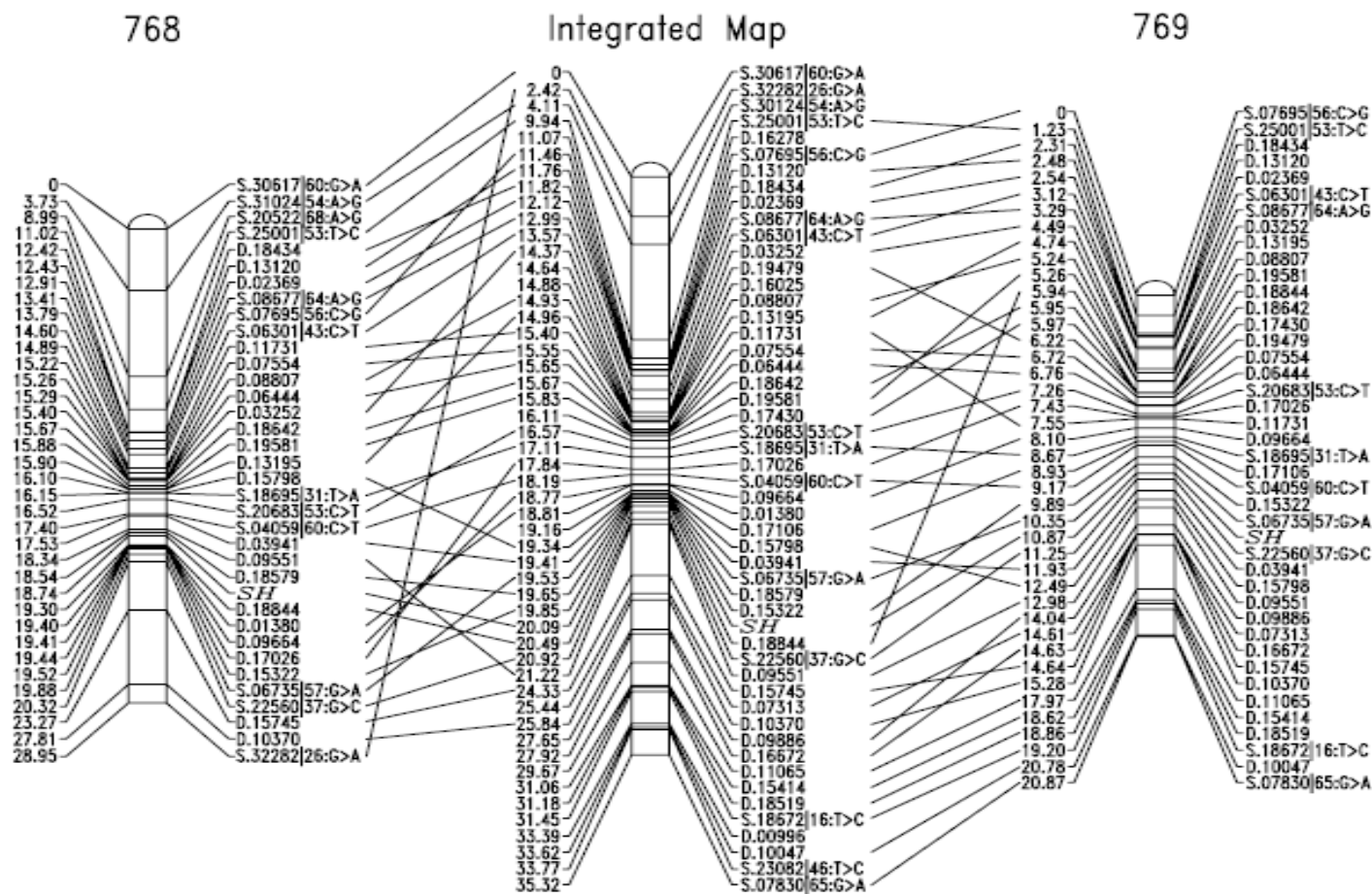
Figure 8.1: Saturation of the shell-thickness region for the 768 and 769 populations and integration of map. Marker names are shown to the right of each LG, with map distances (in cM) to the left. Common markers between the two maps were linked. D: DArT marker, S: SNP marker, mEgCIR: *E. guineensis* SSR marker.

Table 8.1: Homology search of the DArTSeq markers close to the *Sh* gene against the MPOB *pisifera* genome assembly.

No.	Query Markers	Query Length	Subject	Subject Start	Subject End	Subject Length	% Identity	E-Value
	Hit on scaffolds							
1	S.08677|64:A>G	69	p5_sc00060	1883454	1883386	69	98	9.00E-29
2	S.07695|56:C>G	69	p5_sc00060	2341873	2341941	69	100	4.00E-31
3	S.06301|43:C>T	69	p5_sc00060	2020936	2021004	69	100	4.00E-31
4	S.18695|31:T>A	69	p5_sc00060	964795	964727	69	100	4.00E-31
5	S.20683|53:C>T	69	p5_sc00060	1162278	1162210	69	100	4.00E-31
6	S.04059|60:C>T	69	p5_sc00060	931022	930954	69	100	4.00E-31
7	S.06735|57:G>A	69	p5_sc00060	590712	590780	69	98	9.00E-29
8	D.11731	69	p5_sc00060	1247526	1247458	69	100	4.00E-31
9	D.07554	69	p5_sc00060	1391164	1391230	67	98	1.00E-27
10	D.03252	69	p5_sc00060	1660903	1660958	56	100	2.00E-33
11	D.18642	69	p5_sc00060	2622690	2622758	69	100	4.00E-31
12	D.19581	69	p5_sc00060	1637564	1637632	69	98	9.00E-29
13	D.15798	69	p5_sc00060	609934	609999	66	98	6.00E-27
14	D.03941	69	p5_sc00060	573996	574049	54	96	2.00E-17
15	D.09551	69	p5_sc00060	419720	419660	69	100	4.00E-31
16	D.18579	69	p5_sc00060	414468	414524	57	100	5.00E-24
17	D.18844	69	p5_sc00060	224296	224232	69	100	9.00E-29
18	D.01380	69	p5_sc00060	647655	647715	69	98	9.00E-29
19	D.09664	69	p5_sc00060	582349	582281	69	100	4.00E-31
20	D.17026	69	p5_sc00060	689507	689575	69	100	4.00E-31
21	D.15322	69	p5_sc00060	419717	419785	69	100	4.00E-31
22	D.17430	69	p5_sc00060	3182584	3182516	69	100	4.00E-31
23	D.17106	69	p5_sc00060	324907	324975	69	100	4.00E-31
24	D.07313	69	p5_sc00263	175755	175823	69	100	4.00E-31
25	S.22560|37:G>C	69	p5_sc00263	623298	623365	68	98	4.00E-28
26	D.15745	69	p5_sc00263	472927	472862	66	96	1.00E-24
27	D.16672	69	p5_sc00051	3533842	3533774	69	100	4.00E-31
	Hit on orphan contigs							
28	D.06444	69	p5_co354336	1481	1548	68	97	3.00E-26
29	D.13195	69	p5_co354336	1481	1549	69	100	1.00E-31
30	D.09886	69	p5_co661878	381	449	69	95	2.00E-24
31	D.10370	69	p5_co859148	172	240	69	100	1.00E-31
	No hits							
32	D.08807	69	Not hits	-	-	-	-	-
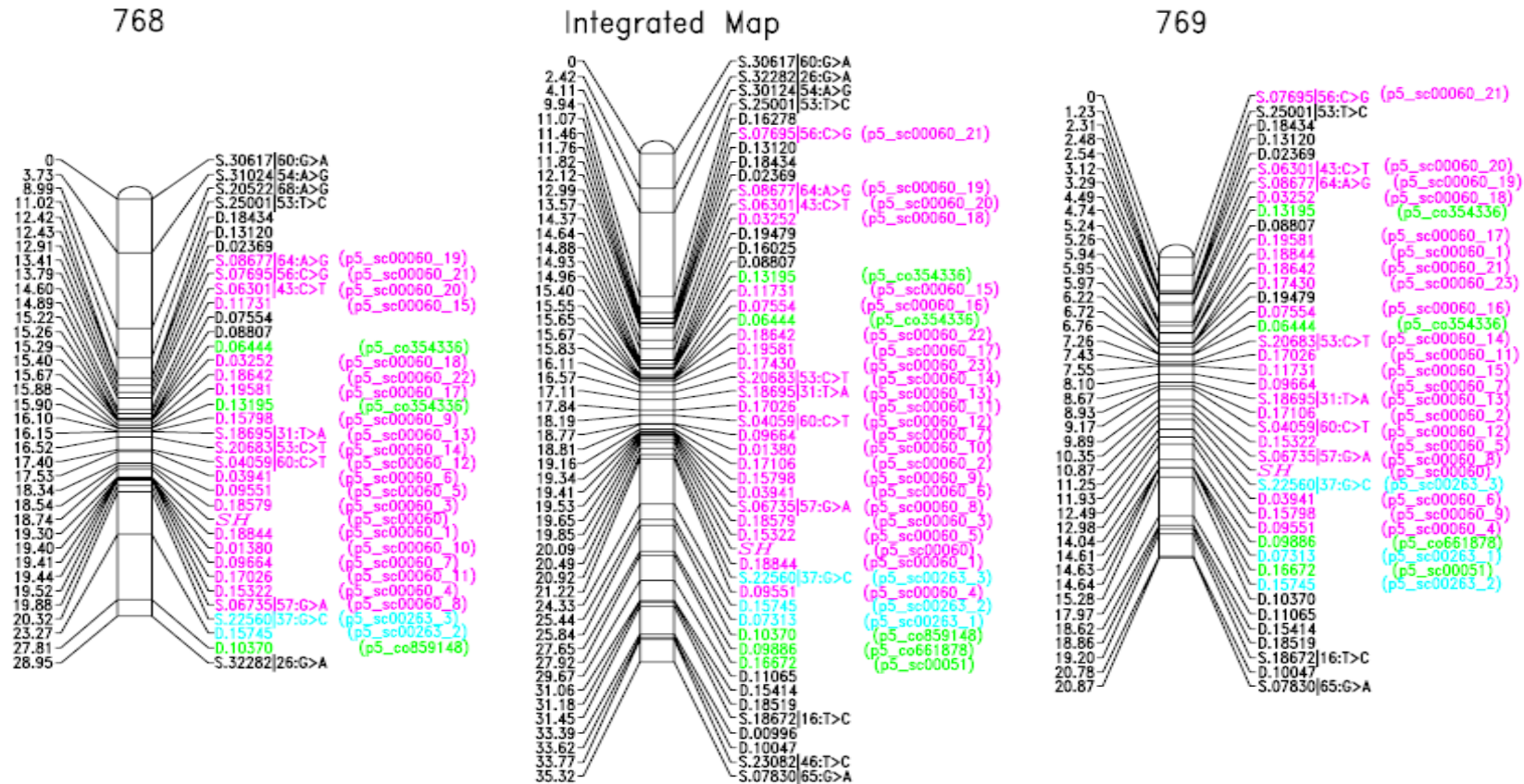
Figure 8.2: The sequential arrangement of DArTSeq markers flanking the *Sh* gene within 5 cM against MPOB *pisifera* genome assembly. Marker names and their corresponding scaffold name are shown to the right of each LG, with map distances (in cM) to the left. Scaffold p5_sc00060 with the shell gene was identified and scaffold p5_sc00263 were highlighted in margenta and blue, respectively, while other scaffold and orphan contigs were in green. Numbers were given according to the sequence location of DArTseq markers against genome assembly.

Figure 8.3: Identification of potential mis-phenotyping of fruit form through examination of locus arrangement of linkage group 4 with saturation of markers around the *Sh* region. The blue arrow highlights the genotype "a" of the 769/58 sample which should be a genotype "h" according to molecular marker genotyping.

8.4 Discussion

Saturation of the shell-thickness region was performed by repeating the genetic mapping process with all available DArTSeq markers, regardless of missing data and quality scores. Denser maps around the *Sh* gene were produced for both F_2 mapping populations separately as well as an integrated map (Figure 8.1). Apparent inversion and/or non-colinearities between these saturated maps were detected and the alignment of markers around the *Sh* gene region as compared to the *pisifera* genome assembly also indicated locally inconsistent marker order with the genome sequence. This is likely to be an effect of variable missing data, lower quality markers and the difficult of mapping high density markers in small genetic intervals with any accuracy, particularly in small mapping populations. The potential mistyping of one of the populations may have further complicated the mapping, by making the *Sh* data essentially incompatible with the marker data.

This locally inconsistent marker order was observed in other mapping studies of sorghum (Mace *et al.*, 2009) as well as *Eucalyptus* (Petroli *et al.*, 2012). Mace *et al.* (2009) commented that inversion is a common feature of closely spaced markers and this arrangement could be real, error in one of the small populations or statistical uncertainty of orders at the cM-scale that is inherent in datasets with small number of individual genotypes. Sample size is indeed vital for resolution of marker order during genetic map construction as it affects the power of linkage detection as well as the accuracy of recombination frequency estimation, particularly when a large number of markers are mapped with a limited progeny size (Alheit *et al.*, 2011), as in the case of

the present study. Map integration of the closely-related 768 and 769 populations has partially resolved the ordering inconsistencies with better alignment of DArTSeq markers according to the genome sequence compared to individual maps.

Meanwhile, Petroli *et al.* (2012) reported that several scattered DArT markers showing discrepancy between genetic maps and genome sequence were borderline in terms of marker quality and call rate parameters which could possibly contribute to the inconsistent marker order. It is believed that the inclusion of poor quality DArTSeq markers in the current high density map around the shell-thickness region could have played a significant role in the observed ordering inconsistencies.

Petroli *et al.* (2012) revealed that 96% of *Eucalyptus* DArT markers from conventional DArT microarray could be successfully aligned to the assembly of genome sequence with 97.1% of the probes confidently aligned to a single locus in the genome. Meanwhile, for the genetic mapping study of an F_2 pesudo-backcross of *Eucalyptus grandis* x *E. urophylla*, mapping of classical DArT markers to genome sequence assembly successfully covered 87% of the sequenced genome with highly conserved marker order between the genetic map and genome scaffold (Kullan *et al.*, 2012). Petroli *et al.* (2012) commented that the genome-coverage attributes of markers derived from the DArTSeq platform should remain essentially the same as those from the classical DArT array because the same genome complexity reduction method is applied in both platforms. Indeed, preliminary analysis of the 32 markers around the *Sh* gene in the present study has shown that all, except one, markers unambiguously hit a unique position in the genome despite the short 64 bp tag sequence of the markers. It is

believed that this is the first study that reported alignment of DArTSeq markers to genome sequence, albeit at a small scale.

Recently MPOB published the identification of the *SHELL* gene, reported to be a homologue of the MADS-box gene *SEEDSTICK* (*STK*, also known as *AGAMOUS-LIKE* 11) (Singh *et al.*, 2013). It was reported that the *SHELL* gene was positioned in T128 linkage group 7 and was mapped by sequence similarity to the assembly scaffold p3-sc00043. Assembly p5-sc00060 was an improved version of scaffold 43 and p5-sc00263 was reassembled from scaffolds p3-sc00191, p3-sc00203 and p3-sc02216 which were also associated with linkage group 7. This publication has confirmed the presence of the *Sh* gene in scaffold p5-sc00060 reported in the present study (Figure 8.2) and thus the identification of 24 closely-linked DArTSeq markers (<5 cM) to the *Sh* gene.

The identification of a palm with discordant phenotype and genotype in the present study (Figure 8.3) has also been reported by Singh *et al.* (2013). Singh *et al.* (2013) genotyped a total of 336 individual palms through sequencing of exon 1 of *SHELL* gene in which 96.7% of palms showed to have concordant genotypes and phenotype but the remaining 11 palms (3.3%) had discordant phenotypes. The authors commented that phenotyping error in the field is believed to be in excess of 5%, emphasizing the importance for a molecular assay to predict fruit form more accurately. Due to variation in fertility, accurate phenotyping of fruit forms for oil palm trees grown in plantation could be difficult for some samples. Indeed, the fruit form of both mapping populations used in the present study was determined by the breeder in the

field and the fruit form was confirmed again during sampling. However, not all the palms were bearing fruits during sampling and hence not all palms identity were confirmed, including the 769/59 palm. The ability of the current set of DArTSeq markers around the *Sh* locus to pinpoint potential mis-phenotyping of fruit forms indicates the usefulness of these markers as a molecular screening tool for fruit form.

In conclusion, a set of closely-linked shell-thickness markers was successfully identified in the present study through saturation of the *Sh* region with all available DArTSeq markers. The homology search of DArTSeq markers flanking the *Sh* gene against the MPOB *pisifera* genome assembly confirmed their close relationships with *SHELL* gene. The identified shell-thickness markers could be valuable as molecular screening tool for fruit form determination, and possible reveal the patterns of recombination in the region surrounding the *Sh* locus in different sources of germplasm.

Additionally, preliminary analysis of the DArTSeq markers against the *pisifera* genome assembly produced a high level of unique hits, despite the short 64 bp tag sequence of the markers. This suggests that DArTSeq markers have a great potential in assisting the anchoring of genetic maps to genomic sequence.

Chapter 9

General Discussion and Future

Directions

9.1 General Discussion

Oil palm (*Elaeis guineensis*) is the most productive oil crop in the world. Palm oil production accounted for 33% of total world vegetable oil production in 2011 with an average yield of 4 tonnes of oil per hectare per year (Montoya *et al.*, 2013). Oil palm fruit can be divided into three different fruit types based on their shell-thickness trait, namely thick-shelled *dura*, shell-less, but often female-sterile, *pisifera* and thin-shelled *tenera*. *Tenera*, a hybrid from a cross between *dura* and *pisifera*, is more productive for palm oil than *dura* due to its thinner shell and increased mesocarp. Therefore the *tenera* is almost exclusively the commercially planted oil palm genotype. The shell-thickness gene is the single most important gene of economic importance in oil palm breeding.

Oil palm is an out-crossing perennial tree crop with long breeding cycles and requires large planting areas for breeding trials. The identity of the fruit form can only be determined when the palms start fruiting 3-5 years after planting and each breeding selection cycle requires at least 12 years of phenotypic evaluation of testcrosses followed by inter-crossing of the best palms to form the basis of the next breeding cycle (Wong and Bernardo, 2008). Therefore the study of any economically and/or agronomically important trait can be very costly, time-consuming and labour-intensive.

The employment of molecular biotechnology tools such as DNA markers and genetic mapping would greatly facilitate and expedite the identification and isolation of important genes/quantitative trait loci (QTL) for desirable traits which could in turn improve the efficiency of palm selection through marker-assisted selection (MAS).

Wong and Bernardo (2008) commented that molecular marker technology has the greatest potential for increasing gain per unit time and reducing the cost of oil palm breeding than in annual crops, such as maize.

The objectives of the present project are to develop approaches for generating molecular markers, linkage mapping and QTL analyses in the oil palm, working on the important shell-thickness trait as well as other yield and vegetative growth traits, with the ultimate goal of developing tools to improve oil palm breeding efficiency.

9.1.1   Approaches to develop molecular markers

Three different molecular marker approaches were explored in the present project, namely Representational Difference Analysis (RDA) (Chapter 3), Amplified Fragment Length Polymorphism (AFLP) (Chapter 4) and DArT "Genotyping-by-sequencing" (DArTSeq) (Chapter 5).

Chapter 3 reported not only the first RDA study coupled with Bulked Segregant Analysis (BSA) for identification of markers closely-linked with the shell-thickness gene but also the novel approach of combining RDA with the NGS technique that was reported by Ho *et al.* (2013). Compared to characterization of the RDA difference products through Sanger sequencing [Section 3.3.3 (c)], it was proven that this new approach generated large numbers of sequences that allow identification of those present at relatively low abundance as well as avoiding the need to go through laborious cloning and transformation process [section 3.3.6 (b)]. This is in accordance

with the comparative study of Sanger and 454 sequencing reported by Lee *et al.* (2009) as discussed in chapter 3.

The RDA study using the *Bam*HI and *Hin*dIII restriction endonucleases generated highly similar enrichment profiles between reciprocal analyses after three rounds of selective hybridization with increased stringency, although different enzymes gave different profiles (Figure 3.10 and 3.11). Assessment of the RDA technique with *Hin*dIII-digested *Lambda* DNA as a positive control spike suggested that enrichment of the target was happening, but not efficiently enough (Figures 3.12 and 3.13). Homology search of the assembled contigs generated from 454 deep-sequencing against the MPOB *pisifera* genome assembly (Singh *et al.*, 2013a; Singh *et al.*, 2013b) revealed that none of the putative contigs were close to the shell-thickness region whereas a search of the public database GenBank identified a significant number of repetitive sequences and organelle DNA such as mitochrondria and chloroplast DNA [section 3.3.6 (d)]. Characterization of common sequences from the RDA difference products indicates that the selective hybridization of the current study was ineffective in which common sequences were not excluded and potentially masked the presence of any real difference products, although the use of NGS did allow a deeper reading of the sequences present.

Chapter 4 reported approaches to develop shell-thickness marker(s) using single-enzyme and conventional AFLP on legitimate *dura* and *pisifera* pools of the 768, 769 and 751 controlled crosses. This is the first attempt to exploit single-enzyme AFLP to study traits of interest in oil palm genome. Around 100 polymorphic bands

355

per primer pair per sample were identified in the selective amplification profile in the present study, a typical observation for AFLP analysis (Vos *et al.*, 1995). Previous studies reported that choice of suitable restriction enzyme is vital for generating informative profiles with reasonable numbers of polymorphic bands (Valsangiacomo *et al.*, 1995). Among the five restriction enzymes tested for the present single-enzyme and conventional AFLP study, single-enzyme *Eco*RI, *Hin*dIII and conventional *Eco*RI/*Mse*I AFLP analyses generated the majority of the potential shell-thickness related-polymorphic bands (Table 4.5). Using the same combination of *Eco*RI/*Mse*I enzyme, a shell-thickness AFLP marker was generated from a La Mé (LM2T) x Deli *dura* (DA10D) genetic background as reported by Billotte *et al.* (2005). The informativeness and usefulness of AFLP as an approach was proven in the present study, as the materials used were from different genetic backgrounds of Binga x Yangambi-AVROS (768 and 769) and Dumpy AVROS x Yangambi AVROS (751).

Both single-enzyme and conventional AFLP analyses generated promising polymorphic bands/profiles. The putative polymorphic bands could be further characterized and/or progenies from the 768 and 769 populations could be genotyped using the selected primer pairs to saturate the genetic linkage maps constructed in the present study (Chapter 6). This would allow the position of the putative AFLP markers to be determined. However due to time limitation, this part of work was not accomplished. It is, however, advisable to include appropriate positive and negative controls to reduce genotyping errors in further AFLP analysis (Pompanon *et al.*, 2005).

The third molecular marker approach developed was the DArT "Genotyping-by-sequencing" (DArTSeq) platform, a relatively new high-throughput genome profiling method. This technique combines the use of the classical DArT genome complexity reduction method with next generation sequencing (NGS) to generate both dominant DArT markers and co-dominant SNP markers (Sansaloni *et al.*, 2011). Chapter 5 presented the first report on genotyping of oil palm populations using the new DArTSeq platform.

In the present study, the markers generated using DArTSeq platform were chosen for progeny genotyping and the construction of genetic linkage maps (Chapter 5 and 6). Classical DArT microarray has clear advantages in terms of cost and genotyping time as shown in various crops (Kilian *et al.*, 2005; Wittenberg *et al.*, 2005; Xia *et al.*, 2005) whereas the DArTSeq platform was able to deliver more dominant DArT markers than the conventional microarray DArT method with an additional set of co-dominant SNP markers located outside the recognition site of the RE (Sansaloni *et al.*, 2011; Cruz *et al.*, 2013). Since both microarray DArT and DArTSeq platforms have the same development costs, a significant decrease in cost per data point for the DArTSeq platform and increase in the speed of analysis was reported (Cruz *et al.*, 2013). Indeed, the Illumina sequencing platform along with sample barcoding has allowed multiplexing of experiments such that many individual mapping projects can be performed in parallel, reducing the cost and effort needed to complete a mapping project (Blair *et al.*, 2008), such is the case in the present study.

Development of SNP markers from other technologies requires DNA sequence and hence most SNP assays were developed for model organisms or major crops where large amounts of DNA sequence information are available (Rafalski, 2002). However, the genotyping-by-sequencing (GbS) approach such as DArTSeq platform requires no reference genome. The consensus of the read clusters across the sequences tagged sites becomes the reference for scoring of SNP markers (Elshire *et al.*, 2013). As the SNPs are scored in a segregating population, they are partially validated, particularly if the segregation patterns permits mapping of the marker associated with the sequence tag. Generation of co-dominant SNP markers from DArTSeq platform in the present study was accomplished prior to the publication of the oil palm genome sequence by Singh *et al.* (2013a). Therefore the current study has proven that DArTSeq platform is suitable for SNP markers development followed by genetic mapping without relying on pre-existing sequence information of the species of interest.

As presented in chapter 5, genotyping of the 768 and 769 F_2 populations with the DArTSeq platform generated a total of 11,675 DArTSeq markers, 6,764 DArT and 4,911 SNP, of good quality which were polymorphic. The markers generated not only allow the construction of high density genetic linkage map for analysis of qualitative/quantitative traits of interest, as in the case of the present study, but the high marker density and observed good genome coverage could also be employed for Genome-Wide Association Studies (GWAS) in oil palm. GWAS is a form of linkage disequilibrium (LD) mapping that investigates the genetic variation in the whole genome to detect signals of association for complex traits (Zhu *et al.*, 2008).

9.1.2   **Genetic linkage mapping**

Genetic linkage mapping is an important tool for the analysis of qualitative and quantitative traits. The type and size of the mapping population can greatly affect the accuracy of genetic map constructed. Based on a simulation study, Ferreira *et al.* (2006) revealed that the higher the number of individuals the better the estimate of genome size. Two hundred individuals were required to construct reasonably accurate linkage maps for all population while F_2 populations with co-dominant markers and RIL populations were more efficient populations for estimating recombination frequency. However populations with high number of individuals might not be feasible due to increased costs and labour required or may simply not exist for perennial and tree species, such as oil palm. Due to space limitations, a typical oil palm breeding programme has plot sizes of 10-20 palms planted in 3-6 replicates (Soh *et al.*, 1990).

In the present study, two closely-related *tenera* self-pollinated populations, 768 and 769, with population sizes of 44 and 57 legitimate progeny, respectively, were chosen for the construction of the genetic maps. In order to increase the accuracy of genetic mapping, genotyping was performed with the high throughput DArTSeq platform to generate large numbers of dominant DArT and co-dominant SNP markers. In fact, Ferreira *et al.* (2006) commented that employment of more informative populations and markers would allow the use of lower number of individuals and maintain the efficiency of genetic mapping. The selection of two populations with full-sibs parents in the present study should also allow some map integration for higher accuracy of marker order. This is the first study that has employed the DArTSeq

platform for genotyping and genetic mapping in oil palm species. Chapter 6 reported the generation of the first high density DArT- and SNP-based genetic maps in oil palm for both the 768 and 769 populations.

The genetic linkage maps of the 768 and 769 populations spanned 1,874.81 cM and 1,720.61 cM, respectively (Table 6.4 and 6.5), comparable with the oil palm high density microsatellite-based genetic linkage map reported by Billotte *et al.* (2005; 1,743 cM). The genome size of *E. guineensis* is estimated to be 2C = 3.86 ± 0.26 pg which is equivalent to 1,887.54 ± 127 Mbp (Madon *et al.*, 2008). It is estimated that the present genetic maps have high genome coverage although genetic distance is not linearly related to physical distance. The high genome coverage reported here correlates well with previous observations and the assumption that DArT markers from both the microarray and GbS platform display a reasonably uniform distribution throughout the genome with preferential targeting of gene-rich region (Kilian *et al.*, 2012; Petroli *et al.*, 2012).

Nevertheless, comparison of the genetic maps in the present study with the recently published genome sequence assembly of oil palm (Singh *et al.*, 2013a) revealed discrepancies in linkage group size and possibly inconsistency in marker arrangement. Ranking of linkage groups generated in the present study according to their map lengths showed that LG 10 (cumulative of both A and B groups) was the second largest LG in both maps whereas LG 10 is the sixth largest chromosome according to published genome assembly. This might indicate a possible chromosomal-assignment disagreement of markers from group B of LG 10 since mapping of LG 10

360

was not successful with the regression algorithm and the maximum likelihood mapping generated undesirably large gap in between part A and B of LG 10 (Chapter 6). Apart from the discrepancies found, LG 4 and 8 were the two largest and LG 5/13 was the smallest LGs generated from both genetic maps, concordant with their chromosomes size in the genome assembly.

The inconsistencies between the genetic maps and sequence assembly were not unexpected (DeWan *et al.*, 2002a). Genetic mapping determines the relative position and distance of markers based on recombination frequencies. Recombination frequency between two markers depends on the informativeness of the markers and the number of individuals typed with limited numbers of meiotic events causing poor estimates of recombination frequencies and incorrect ordering of markers in small genetic regions (DeWan *et al.*, 2002b). Indeed, mapping of multiple markers to the same location was more apparent in the map of the 768 population than the 769 population due to the lower number of individuals genotyped, hence the lower number of meiotic events analysed.

The accuracy of genetic maps is vital for fine mapping and for the isolation of genes for traits of interest. Given that the majority of the DArT markers unambiguously aligned to a unique position in the genome (Table 8.1; Kullan *et al.*, 2012; Petroli *et al.*, 2012) and the availability of a genome assembly of oil palm (Singh *et al.*, 2013a), the reported high density DArT- and SNP-based genetic maps in the present study can be improved. The information from both genetic maps and physical maps can be combined to correctly order the markers, particularly those closely-linked markers for

fine mapping. To achieve this, DeWan *et al.* (2002b) suggested that markers can be selected based upon (i) if its sequence-based physical map agrees with the genetic map; (ii) heterozygosity of the marker; and (iii) its map position was supported by a likelihood ratio ≥ 3. Based on these suggestions, a better framework map of the 768 and 769 populations can be established by selecting highly informative markers, particularly co-dominant SNP and SSR markers, which are common to both populations and the map position tested for concordance to the published *pisifera* genome assembly. A full genetic map can then be constructed by fixing the marker order of the framework map to allow addition of more DArTSeq markers, without disturbing the best framework order of markers. This genetic map would offer the possibility of assigning unanchored scaffolds to assembled pseudochromosomes of the published genome assembly (Petroli *et al.*, 2012), although for genomes with substantial scaffold fragmentation, it would require large numbers of progeny to allow fine order mapping to be accurate.

The high density DArT- and SNP-based genetic maps reported in the present study have the highest marker density compared to all the previously reported genetic maps reported in oil palm. By using the classical DArT microarray, a study of an F_2 pseudo-backcross of *Eucalyptus grandis* x *E. urophylla* also reported the densest genetic map in which a consensus linkage map was constructed using 2,229 DArT markers and 61 SSR loci resulting in an average marker density of 0.48 cM (Kullan *et al.*, 2012). In another study of F_1 cross of the same *E. grandis* x *E. urophylla* species, a

map was constructed from 2,484 markers (2,274 DArT markers and 210 microsatellite) with an average inter-marker distance of 0.5 cM (Petroli *et al.*, 2012).

Clustering of DArT markers as well as the co-segregation of a number of DArT markers were observed in the above studies as well as other mapping studies using DArT markers (Akbari *et al.*, 2006; Semagn *et al.*, 2006d; Wenzl *et al.*, 2006; Mace *et al.*, 2009). This was also observed in the present study. Clustering of markers could be caused by an unbalanced distribution of recombination events along chromosomes (Lou *et al.*, 2013; Mace *et al.*, 2009) or may also be indicative of gene-rich regions in the genome (Alheit *et al.*, 2011; Semagn *et al.*, 2006d). *Pst*I, the most commonly used restriction enzyme in the DArT assay for genome complexity reduction, is a CXG methylation-sensitive enzyme that cuts hypomethylated sequences which are often low-copy and occur primarily in gene-rich regions of the genome (Schouten *et al.*, 2012). Nevertheless, some local clustering of the markers in the present map could be due to the limited number of individual genotyped and hence limited recombination events for distinguishing between individual markers during map construction.

Although both genetic maps of the 768 and 769 populations shared high similarity in terms of grouping of markers into the same linkage groups and the location of such markers in terms of chromosomes (telomere vs centromere), the linear marker order between the maps was not completely congruent (Figures 6.1 and 6.2). The majority of observed marker order inconsistencies involved closely-spaced markers, covering about 1-5 cM, but in a few cases different marker orders also occurred over longer distances. Inconsistencies of marker order were commonly observed in plant

species especially when individual maps are integrated into the consensus map and it is believed that this phenomenon is mainly due to differences in recombination frequencies of marker pairs in populations of different sizes and type, probably due to the stochastic nature of recombination (Khan *et al.*, 20112; Mace *et al.*, 2009; Studer *et al.*, 2010). These marker orders could be explained by statistical uncertainty of orders at the cM-scale due to the limited number of progenies in the datasets (Mace *et al.*, 2009) or they could be caused by local rearrangements or segmental duplications of the genome. Khan *et al.* (2012) in their study of a multi-population consensus genetic map of apple observed this. All of the above explanations could possibly play a role in inconsistencies of marker order between the 768 and 769 maps. The construction of framework maps using highly informative common markers or anchored loci would have great potential for studying the presence of any real inversion or rearrangement of marker order between the 768 and 769 populations.

In the mean time, it is anticipated that whole genome re-sequencing of the parents of 768 and 769 populations, 228/05 and 228/06 respectively, would allow better understanding of the genome arrangement between the two closely-linked populations. With the recent publication of oil palm genome sequence by MPOB, whole genome re-sequencing can be performed using economical massively-parallel next-generations sequencing technologies that generate short sequence read of 35-100 bases, such as Illumina SOLEXA and ABI SOLiD (Yann and Juan, 2010). The short reads can be mapped against the reference genome, allowing discovery of any genetic variations between the closely-related *tenera* self-pollinated populations and the reference

genome. Comparison of DArTSeq markers to the whole-genome sequencing would be beneficial to address the discrepancies in linkage group as well as inconsistency in marker arrangement.

9.1.3   Quantitative trait loci (QTL) Mapping

Chapter 7 reported a preliminary QTL study of important quantitative yield traits, bunch components as well as vegetative growth traits of the 768 and 769 populations. Different sets of significant and putative QTLs were identified for each cross with no common significant QTL for any particular trait (Tables 7.13 and 7.14). Direct comparison with the study of Rance *et al.* (2001) working on similar phenotypic traits was difficult because of different marker types, mapping population structure and density of genetic maps used in the studies. Nevertheless, comparison with the study of Billotte *et al.* (2010) is possible with the presence of common anchored SSR loci deliberately included in the genetic maps of the 768 and 769 populations. No congruence of QTL locations was found between the two studies.

The size of the mapping population is the most important factor influencing QTL detection. The present QTL study has the smallest population size among all previously reported QTL studies on yield components as well as fatty acid components of oil palm (Billotte *et al.*, 2010; Montoya *et al.*, 2013; Rance *et al.*, 2005; Singh *et al.*, 2009). The initial population size of the present study is limited due to the lack of larger available populations in the current breeding programme, while removal of the *pisifera* palms from QTL analysis due to their female-infertility character further reduced the

365

population size to 33 and 44 progenies for the 768 and 769 controlled crosses, respectively.

Limited population size not only led to low QTL detection power, particularly those QTLs with small or medium effects, but also over-estimation of the effects of detected QTLs (Beavis, 1998; Melchinger *et al.*, 1997; Vales *et al.*, 2005). This well-known fact is further supported by a recent QTL mapping study by Pelgas *et al.* (2011) using two populations of 260 and 500 progenies. This study revealed that only 29% of QTL detected from 500 progenies were also identified using 260 progenies and the percentage of phenotypic variance explained for these 29% QTL in the 260 progenies were approximately twice as large as those obtained from 500 progenies. Although correction of bias due to sampling error was performed in the present study as suggested by Luo *et al.* (2003), the correction is minimal and hence it is believed that the corrected phenotypic variance explained by QTL identified in the present study could be overestimated.

In view of the small family size of classical breeding trials of oil palm, multi-parent QTL mapping like that tested in oil palm by Billotte *et al.* (2010) might be more effective in the detection and evaluation of the effects of QTLs. In fact, 16 out of the 44 QTLs detected by the across-family model were not identified by the within- family analyses in the study by Billotte *et al.* (2010). The authors commented that the larger population size of multi-parent mapping design allows greater detection power for QTL of a given parent shared by several crosses but does not alleviate the effect of small number of individuals per cross, hence QTLs could be detected by one model but not

the other. This is obviously compounded by the fact that different effects will be segregating in the different populations within the cross design. It is anticipated that larger mapping population size, around 100 palms per family, coupling with multi-parent QTL mapping approach would be beneficial for QTL analysis in oil palm.

Despite the numerous QTL mapping studies on various agronomically important traits in diverse crop species, QTL studies have shown limited application as markers used have not been reliable in predicting the desired phenotype due to low accuracy of QTL mapping studies and/or inadequate validation (Young, 1999; Semagn *et al.*, 2006c). Therefore, identification of reliable QTL is a preliminary step in developing a successful marker-assisted selection (MAS). Marker should be validated in independent populations of different genetic backgrounds so that they can reveal polymorphism in a wide range of parental genotypes to be useful in breeding programmes (Collard *et al.*, 2005). Also, QTL should be detected in populations which are intimately part of the current breeding programme.

The limitations of pedigree-based QTL mapping can be solved by Genome-wide Association Studies (GWAS). By using large number of lines and varieties or the entire germplasm, GWAS detects common genetic variation segregating in the populations with higher mapping resolution, often down to single genes or individual nucleotides level (Korte and Farlow, 2013; Zhu *et al.*, 2008). GWAS requires the availability of large numbers of polymorphic markers with density higher than the extent of linkage disequilibrium of the species of interest to detect significant associations between genotype and phenotype (Brachi *et al.*, 2011; Semagn *et al.*,

2010). High density parallel genotyping technologies such as DArTSeq platform reported in the present study could be a potential approach for GWAS. So far, GWAS have primarily focused on plants with short lifecycles, such as *Arabidopsis*, maize and rice, and there are no reported applications of this approach in oil palm, GWAS is believed to be particularly suited for perennial tree crops with long generation time (Iwata *et al.*, 2013; Khan and Korban, 2012).

9.1.4 Study on the shell-thickness region

The construction of genetic maps using two closely-related *tenera* self-pollinated crosses, 768 and 769, allowed a study of the shell-thickness region as well as other agronomically important yield traits. The high density genetic maps of the 768 and 769 crosses were further saturated around the shell-thickness region with all DArTSeq markers regardless of their quality followed by integration of the two population maps (Figure 8.1). Homology search of markers flanking the *sh* locus within 5 cM against the MPOB *pisifera* genome assembly revealed that 72% of tested DArTSeq markers were located to scaffoldp5_sc00060 in which the *SHELL* gene is reported to be located (Singh *et al.*, 2013b). This confirmed the identification of markers closely-linked with the shell-thickness region in the present study. The unambiguous alignment of DArTSeq markers to unique positions in the genome despite being short in sequence (64 bp) has proven that DArTSeq markers generated in the present study using the *Pst*I enzyme are also likely to be targeting the gene-rich or unique sequence regions, similar to microarray DArT-based studies (Kullan *et al.*, 2012; Petroli *et al.*, 2012).

368

Based on the reference *pisifera* genome sequence, the *SHELL* gene has recently been mapped and sequenced in which allelic differences were found between *dura* and *pisifera* as reported by Singh *et al.* (2013b). *Pisifera* fruit forms were found to be caused by two independent disruptive SNPs that affect the highly conserved DNA binding and dimerization domain of a MADS-box gene. These mutations are accounted for the single-gene heterosis of *tenera* palms via heterodimerization.

The oil palm industry has been facing a *dura* contamination problem in commercial D x P seed due to poor crossing quality control (Kushairi and Rajanaidu, 2000; Cheyns *et al.*, 2001). It is of utmost importance to have a marker for the shell-thickness gene to allow the industry to confirm the purity of commercial planting materials. Markers for the shell-thickness gene would allow fruit forms to be distinguished at the nursery stage before they are field planted and hence facilitate planting of the wanted fruit type and improve resource allocation. Accurate genotyping by the shell-thickness marker would also prevent mis-phenotyping of palm type as reported by Singh *et al.* (2013b). The authors commented that the fruit form phenotyping error is believed to be in excess of 5%, highlighting need for a diagnostic kit that can accurately identify fruit form. Genetic mapping analysis in the present study has also identified a potential mis-phenotyped palm, 769/59. The identity of fruits in the present study was previously determined by the breeder and confirmed again during leaf sampling for the project. However the fruit form of this particular palm together with some other palms were not confirmed as the palms were not bearing fruits during sampling. Therefore further examination and confirmation is required. As the incorrect

phenotyping is for a *pisifera* palm, whereas the genotype of the surrounding markers is for a *tenera* fruit form, it is also possible that the mis-phenoyping has occurred because of the presence of an infertile *tenera*, which are believed to occur at a low frequency.

The present study has successfully identified a set of 32 DArTSeq markers closely mapped to the flanking region of the shell-thickness gene. The identified SNP and DArT markers would be useful as in-house molecular screening tool of fruit form prior availability of commercialized kit develop from the *SHELL* gene. The published *SHELL* marker could be used as positive control for validation of the identified DArTSeq markers.

9.2   Conclusions

A number of approaches were used in this project for the development of molecular markers, linkage mapping and QTL analyses in oil palm, working on the economically important shell-thickness trait and other quantitative yield and vegetative traits. Major findings from the present study are listed below:

1. The Representational Difference Analysis (RDA) approach was employed to develop markers linked to the shell-thickness gene using two different restriction enzymes, *Bam*HI and *Hin*dIII, which generated highly similar enrichment profiles in reciprocal analyses. Novel coupling of next-generation sequencing to RDA enabled a more comprehensive study of the enrichment profiles compared to Sanger sequencing. Identification of repetitive elements and organelle DNAs from the difference products indicates that common

sequences were not efficiently excluded during selective hybridization, masking the presence of real difference products (Chapter 3).

2. Both single-enzyme and conventional Amplified Fragment Length Polymorphism (AFLP) were employed to develop markers linked to shell-thickness gene and both generated putative shell-thickness related-polymorphic bands that require further validation (Chapter 4).

3. Development of the first set of DArTSeq markers in oil palm through genotyping of two closely-related F_2 populations, the 768 and 769, generated a total of 6,764 dominant DArT and 4,911 co-dominant SNP markers of good quality and were polymorphic (Chapter 5).

4. Characterization and identification of a subset of 948 and 958 high quality polymorphic DArT markers as well as 719 and 729 SNP markers from the DArTSeq platform for the 768 and 769 populations, respectively (Chapter 5).

5. Characterization of 36 polymorphic SSR markers derived from the public CIRAD database as anchored loci for genetic mapping (Chapter 5).

6. Construction of the first high density DArTSeq genetic linkage maps of oil palm for the 768 and 769 populations (Chapter 6).

7. Preliminary QTL mapping of 21 quantitative production, bunch components and vegetative growth traits identified four and two significant QTLs as well as 34 and 30 putative QTLs for the 768 and 769 populations, respectively (Chapter 7).

8. Preliminary alignment of DArTSeq markers to the genome sequence indicated that DArTSeq markers are highly enriched for genic regions and can be used to

identify corresponding scaffolds to develop single locus markers for MAS (Chapter 8).

9. Identification of markers closely-linked with the shell-thickness gene through saturation of maps around this region and alignment to oil palm *pisifera* genome assembly (Chapter 8).

9.3   Limitations of the study

Nevertheless, the present study has several noteworthy limitations that will be discussed with recommendations made for future research.

Development of shell-thickness marker(s) using Representation Difference Analysis (RDA) approach has gave rise to highly similar enrichment profile between reciprocal analyses as well as identification of >35% of repetitive sequences and organelle DNA. Successfully generation of high quality DArTSeq markers in the present study using *Pst*I enzyme as genome complexity reduction approach suggesting that RDA analysis using *Pst*I enzyme might potentially be useful in eliminating common repetitive sequences in the genome through its preferential targeting of hypomethylated gene-rich regions of chromosomes. It is also sensible to include a positive control in RDA study for enrichment assessment.

The major limitation in the present study is the small population size of oil palm breeding available for genetic linkage mapping and QTL mapping. Despite the use of two closely-related *tenera* self-pollinated populations to allow combination of any

potential QTLs identified, different QTLs were identified in both populations with variable amount of variation are accounted for by the QTLs. It is anticipated that larger mapping populations, at least 100 palms, should be prioritized together with QTL mapping using multi-parent approach in order for QTL mapping to be beneficial for breeding of perennial tree crops such as oil palm. Meanwhile, establishment of framework maps using highly informative markers from both the 768 and 769 genetic linkage maps as well as whole genome re-sequencing of both parents would enable study of local inconsistency due to translocations and/or inversions.

9.4 Future directions

In view of the recent publication of *SHELL* gene by MPOB (Singh *et al.*, 2013b), the RDA approach to identify markers linked to the shell-thickness gene should not be pursued any further. However methylation-sensitive RDA analysis using the *Pst*I enzyme together with a positive control for enrichment assessment could be of potential for marker development and coupling of RDA and NGS would allow more comprehensive study of the enrichment profiles.

As for the AFLP approach, the single-enzyme *Eco*RI, *Hin*dIII and conventional *Eco*RI/*Mse*I approach with specific primer pairs that generated the putative shell-thickness polymorphic profile could be used to genotype both the 768 and 769 F_2 populations, to complement genetic maps and identify a map location for the AFLP markers identified by BSA. It is important, however, to include both positive and negative controls to reduce/evaluate genotyping errors in further AFLP analysis.

373

Establishment of the new framework maps utilizing highly informative markers, such as co-dominant SSR and SNP markers, which are common between the 768 and 769 populations would allow a better comparison between the two populations for study of potential inversions and genome arrangements.

Lastly, it is of great interest to validate and convert the closely-linked DArTSeq markers identified in the present study to PCR format that can be utilized as a molecular tool for verification and/or determination of fruit forms early in the nursery stage. The identity of potentially mis-phenotyped 769/59 palm should also be confirmed once the proposed PCR-based screening tool and/or the published *SHELL* marker is available commercially.

References

ADAM-BLONDON, A.F., SEVIGNAC, M., BANNEROT, H. and DRON, M. (1994) SCAR, RAPD and RFLP markers linked to a dominant gene (Are) conferring resistance to anthracnose in common bean. *Theoretical and Applied Genetics* 88, 865-870.

ADAWY, S.S., HUSSEIN, E.H.A., ISMAIL, S.E.M.E., EL-ITRIBY, H.A. (2005) Genomic diversity in date palm (*Pheonix dactylifera* L.) as revealed by AFLP's in comparison to RAPD's and ISSR's. *Arab Journal of Biotechnology* 8, 99-114.

AKBARI, M., WENZL, P., CAIG, V., CARLING, J., XIA, L., YANG, S., USZINSKI, G., MOHLER, V., LEHMENSIEK, A., KUCHEL, H., HAYDEN, M.J., HOWES, N., SHARP, P., VAUGHAN, P., RATHMELL, B., HUTTNER, E. and KILIAN, A. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theoretical and Applied Genetics* 113, 1409-1420.

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. and MALEK, J.A. (2011) De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nature Biotechnology* 29, 521-527.

ALHEIT, K.V., REIF, J.C., MAURER, H.P., HAHN, V., WEISSMANN, E.A., MIEDANER, T. and WÜRSCHUM, T. (2011) Detection of segregation distortion loci in triticale (x *Triticosecale* Wittmack) based on high-density DArT marker consensus genetic linkage map. *BMC Genomics* 12, 380.

ALLEN, N.L., PENN, C.W. and HILTON, A.C. (2003) Representational difference analysis: critical appraisal and method development for the identification of unique DNA sequences from prokaryotes. *Journal of Microbiological Methods* 55, 73-81.

ALTINKUT, A., KAZAN, Z. and GOZUKIRMIZI, N. (2003) AFLP marker linked to water-stress-tolerant bulk in barley (*Hordeum vulgare* L.). *Genetics and Molecular Biology* 26, 77-82.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. and LIPMAN, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215, 403-410.

ANSORGE, W.J. (2009) Next-generation DNA sequencing techniques. *New Biotechnology* 25, 195-203.

ARGAWAL, M., SHRIVASTAVA, N. and PADH, H. (2008) Advances in molecular marker techniques and their applications in plant sicences. *Plant cell Report* 27, 617-631.

ASCHERIO, A. (2002) Diet fat and risk of CHD: New developments. In: *Enhancing Oil Palm Industry Development through environmentally technology*, pp. 60-65 (Eds. Z. Poeloengan *et al.*). Indonesian Oil Palm Research Institute.

ASHKANI, S., RAFII, M.Y., RUSLI, I., SARIAH, M., ABDULLAH, S.N.A., RAHIM, H.A. and LATIF, M.A. (2012) SSRs for marker-assisted selection for blast resistance in rice (*Oryza sativa* L.). *Plant Molecular Biology Reporter* 30, 79-86.

BACHEM, C.W., VAN DER HOEVEN, R.S., DE BRUIJIN, S.M., VREUGDENHIL, D., ZABEAU, M. and VISSER, R.G. (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant Journal* 9, 745-753.

BALDOCCHI, R.A. and FLAHERTY, L. (1997) Isolation of genomic fragments from polymorphic regions by representational difference analysis. *Methods: A companion to Methods in Enzymology* 13, 337-346.

BARONE, A. RITTER, E., SCHACHTSCHNABEL, U., DEBENER, T., SALAMINI, F. and GEBHARDT, C. (1990) Localization of restriction fragment length polymorphism mapping in potato of a major dominant gene conferring resistance to the potato cyst nematode *Globodra schachtii*. *Molecular and General Genetics* 224, 177-182

BASIRON, Y. and SALMIAH, A. (1994) Potential new value-added products from palm oil and palm kernel oil. In: *Management for enhanced profitability in plantations,* pp. 409-418 (Eds. K.H. Chee). The Incorporated Society of Planters, Kuala Lumpur.

BATLEY, J., BARKER, G., O'SULLIVAN, H., EDWARDS, K. and EDWARDS, D. (2003) Mining for single nucleotide polymorphisms in insertions/deletions in maize expressed sequence tag data. *Plant Physiology* 132, 84-91.

BAUDOUIN, L. (1992) Use of molecular markers for oil palm breeding. I. Protein Markers. *Oleagineux* 47, 681-691.

BEAVIS. W. D. (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies. In: *Proceedings of the Forty-Ninth Annual Corn & Sorghum Industry Research Conference*, pp. 250-266. American Seed Trade Association, Washington, D. C.

BEAVIS, W. D. (1998) QTL analysis: power, precision, and accuracy. In; *Molecular Dissection of Complex Traits*, pp. 145-162 (Ed. A.H. Paterson) CRC Press, New York.

BECKMANN, J. and SOLLER, M. (1986) Restriction fragment length polymorphisms in plant genetic improvement. *Oxford Surveys of Plant Molecular and Cell Biology* 3, 197-250.

BEIRNAERT, A. (1935) Intoduction à la biologie florale du palmier à huile (*Elaeis guineensis* Jacquin). *Publ. Inst. Nat. Etude agron, Congo Belge, Ser. Sci.* 5, 3-42.

BEIRNAERT, A. and VANDERWEYEN, R. (1941) Contribution à l'étude génétique et biométrique des variétés d'*Elaeis guineensis* Jacq. Publications de l'institut national pour l'étude agronomique du Congo Belge, *série scientifique* 27.

BENJAK, A., KONRADI, J., BLAICH, R. and FORNECK, A. (2006) Different DNA extraction methods can caused different AFLP profiles in grapevine *(Vitis vinifera* L.). *Vitis* 45, 15-21.

BENNETT, M.D. and SMITH, J.B. (1991) Nuclear DNA amounts in angiosperm. *Philosophical Transactions of the Royal Society London B* 334, 309-345.

BERNATZKY, R., and TANKSLEY, S.D. (1986) Toward a saturated linkage map in tomato based on isozyme and random cDNA sequences. *Genetics* 112, 887–898.

BEUTOW, K.H., EDMONSON, M.N. and CASSIDY, A.B. (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genetics* 21, 323-325.

BEYENE, Y., BOTHA, A.M., and MYBURG, A.A. (2005) A comparative study of molecular morphological methods of describing genetic relationships in traditional Ethiopian highland maize. *African Journal of Biotechnology* 4, 586-595.

BILLOTTE, N., FRANCES, L., AMBLARD, P., DURAND-GASSELIN, T., NOYER, J.L. and COURTOIS, B. (2001a) Search for AFLP and microsatellite molecular markers of the *Sh* gene in oil palm (*Elaeis guineensis* Jacq.) by bulk segregant analysis (BSA) and genetic mapping. In: *Proceeding of PIPOC International Palm Oil Congress 2001,* pp. 442-445. Malaysian Palm Oil Board (MPOB), Kuala Lumpur, Malaysia.

BILLOTTE, N., JOURJON, M.F., MARSEILLAC, N., BERGER, A., FLORI, A., ASMADY, H., ADON, B., SINGH, R., NOUY, B., POTIER, F., CHEAH, S.C., RODHE, W., RITTER, E., COURTOIS, B., CHARRIER, A. and MANGIN, B. (2010) QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics* 120, 1673-1687.

BILLOTTE, N., LAGODA, P.J.L., RISTERUCCI, A-M. and BAURENS F-C. (1999) Microsatellite-enriched libraries: applied methodology for the development of SSR markers in tropical crops. *Fruits* 54, 277-288.

BILLOTTE, N., MARSEILLAC, N., RISTERUCCI, A–M., ADON, B., BROTTIER, P., BAURENS, F –C., SINGH, R., HERRÁN, A., ASMADY, H., BILLOT, C., AMBLARD, P., DURAND-GASSELIN, T., COURTOIS, B., ASMONO, D., CHEAH, S.C., ROHDE, W., RITTER, E., and CHARRIER, A. (2005) Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis*) Jacq. *Theoretical and Applied Genetics* 110, 754-765.

BILLOTTE, N., RISTERUCCI, A.M., BARCELOS, E., NOYER, J.L., AMBLARD, P. and BAURENS, F.C. (2001b) Development, characterization, and across-taxa utility of oil palm (*Elaeis guineensis* Jacq.) microsatellite markers. *Genome* 44, 413-425.

BISHOP, D.T., WILLIAMSON, J.A. and SKOLNICK, M.H. (1983) A model for restriction fragment length distributions. *The American Journal of Human Genetics* 35, 795-815.

BLAAK, G., SPARNAAIJ, L.D. and MENENDEZ, T. (1963) Breeding and inheritance in the oil palm (*Elaeis guineensis* Jacq.) Part II. Methods of bunch quality analysis. *Journal of West African Institute for Oil Palm Research* 4, 146-155.

BLEARS, M.J., DE GRANDIS, S.A., LEE, H. and TREVORS, J.T. (1998) Amplified fragment length polymorphism (AFLP): a review of the procedure and its applications. *Journal of Industrial Microbiology and Biotechnology* 21, 99-114.

BOLIBOK-BRAGOSZEWSKA, H., HELLER-USZYNSKA, K., WENZL, P., USZYNSKI, G., KILIAN, A. and RAKOCZY-TROJANOWSKA, M. (2009) DArT markers for the rye genome – genetic diversity and mapping. *BMC Genomics* 10, 578.

BONIN, A., BELLEMAIN, E., BRONKEN EIDESEN, P., POMPANON, F., BROCHMANN, C. and TABERLET, P. (2004) How to track and assess genotyping errors in population genetic studies. *Molecular Ecology* 13, 3261-3273.

BOTSTEIN, D., WHITE, R.L., SKOLNICK, M. and DAVIS, R.W. (1980) Construction of a genetic linkahe map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* 32, 314-331.

BOWER, R. and BIRCH, R.G. (1992) Transgenic sugarcane plants via microprojectile bombardment. *Plant Journal* 2, 409-416.

BOWLER, L.D., HUBANK, M. and SPRATT, B.G. (1999) Representational difference analysis of cDNA for the detection of differential gene expression in bacteria: development using a model of iron regulated gene expression in *Neisseria meningitidis*. *Microbiology* 145, 3529-3537.

BRACHI, B., MORRIS, G.P. and BOREVITZ, J.O. (2011) Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology* 12, 232.

BRADSHAW, H.D. and STETTLER, R.F. (1994). Molecular genetics of growth and development in populous. II. Segregation distortion due to genetic load. *Theoretical and Applied Genetics* 89, 551-558.

BROEKMANS, A.F.M. (1957) Growth flowering and yield of the oil palm in Nigeria. *Journal of West African Institute of Oil Palm Research* 2, 187-220.

BROWN, J.D., GOLDEN, D. and O'NEILL, R.J. (2008) Methylation perturbations in retroelements within the genome of a Mus interspecific hybrid correlate with double minute chromosome formation. *Genomics* 91, 267-273.

BUTLIN, R.K. (2010) Population genomics and speciation. *Genetica* 138, 409–418.

CAI, H.-W., GAO, Z.-S., YUYAMA, N., and OGAWA, N. (2003) Identification of AFLP markers closely linked to the *rhm* gene for resistance to Southern Corn Leaf Blight in maize by using bulked segregant analysis. *Molecular Genetics and Genomics* 269, 299-303.

CASTIGLIONI, P., AJMONE-MARSAN, P., VAN WIJK, R., and MOTTO, M. (1999). AFLP markers in a molecular linkage map of maize: co-dominant scoring and linkage group distribution. *Theoretical and Applied Genetic* 99, 425-431.

CASTILHO, A., VERSHININ, A. and HESLOP-HARRISONa, J.S. (2000) Repetitive DNA and the chromosomes in the genome of oil palm (*Elaeis guineensis*). *Annals of Botany* 85, 837-844.

CERVERA, M-T., STORME, V., IVENS, B., GUSMAO, J., LIU, B.H., HOSTYN, V., VAN SLYCKEN, J., VAN MONTAGU, M. and BOERJAN, W. (2001). Dense genetic linkage maps of three populus species (*Populus deltoids, P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. *Genetics* 158, 787-809.

CHANG, C., BOWMAN, A.W., LANDER, E.S. and MEYEROWITZ, E.W. (1988) Restriction fragment length polymorphism linkage map of *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences the United States of America* 85, 6856-6860.

CHANG, J.–H., CUI, J.–H., XUE, W., and ZHANG, Q.–W. (2012). Identification of molecular markers for a Aphid resistance gene in Sorghum and selective efficiency using these markers. *Journal of Integrative Agriculture* 11, 1086-1092.

CHANG, Y., STOCKINGER, M.P., TASHIRO, H. and LIN, C.L. (2008). A novel noncoding RNA rescues mutant SOD1-mediated cell death. *FASEB Journal* 22, 691-702.

CHEAH, S.C. (1990) Restriction fragment length polymorphism (AFLP) in oil palm. *Paper presented at Third National Molecular Biology Seminar*. 25-29 September 1990, Serdang, Malaysia.

CHEAH, S.C. (2000) Biotechnological strategies for improving plantation tree crops: the oil palm – a case study. In: *Proceedings of the International Planters Conference*, pp. 59-76, Kuala Lumpur, Malaysia.

CHEAH, S.C., NOR AKMAR, S.A., OOI, L.C.L., RAHIMAH, A.R. and MARIA, M. (1993) Detection of DNA variability in the oil palm using RFLP probes. In: *Proceedings of the PORIM International Palm Oil Conference-Agriculture*, pp. 140-150 (Eds. Y. Basiron and B.S. Jalani). Palm Oil Research Institute of Malaysia (PORIM), Bangi, Malaysia.

CHEEMA, J. and DICKS, J. (2009) Computational approaches and software tools for genetic linkage map estimation in plants. *Briefing in Bioinformatics* 10, 595-608.

CHEONG, P.K., KOH, S.J., CHEAH, S.C. and SINGH, R. (2006) Analysis of tissue culture-derived regenerants using methylation sensitive AFLP. *Asia Pacific Journal of Molecular Biology and Biotechnology* 14, 47-55.

CHETELAT, R.T., MEGLIC, V. and CISNEROS, P. (2000). A genetic map of tomato based on BC_1 *Lycopersicon esculentum X Solanum lycopersicoides* reveals overall synteny but suppressed recombination between these homeologoues genomes. *Genetics* 154, 857-867.

CHEYNS, E., KOUAME, Y.S. and NAI, S. (2001) Itinéraires techniques et nature du materiel vegetal: diversité des forms socials et techniques de production en Côte d'Ivoire. *Ol Corps Gras Lipides* 8, 524-528.

CHIN, C.W. (1995) Oil Palm planting materials and quality control. In: *Technologies in the Plantation-"The Way Forward", Proceedings PORIM National Oil Palm Conference*, pp. 38-47 (Eds. N. Rajanaidu and B.S. Jalani). Palm Oil Research Insititute of Malaysia (PORIM), Bangi, Malaysia.

CHOI, J.Y., SIFRI, C.D., GOUMNEROV, B.C., RAHME, L.G., AUSUBELI, F.M. and CALDERWOOD, S.B. (2002) Identification of virulence genes in a pathogenic strain of Pseudomonas aeruginosa by representational difference analysis. *The Journal of Bacteriology* 184, 952-961.

CHOO, Y.M. and CHEAH, K.Y. (2000) Biofuel. In: *Advances in oil palm research*, Vol 2, pp. 1293-1345 (Eds. Y. Basiron, B.S. Jalani and K.W. Chan). Malaysian Palm Oil Board, Kuala Lumpur.

CHOONG, C.Y., SHAH, F.H., RAJANAIDU, R. and ZAKRI, A.H. (1996) Isozyme variation of Zairean oil palm (*Elaeis guineensis* Jacq.) germplasm collection. *Elaeis* 8, 45-53.

CHUNG, A.M., STAUB, J.E. and CHEN, J.F. (2006) Molecular phylogeny of *Cucumis* species are revealed by consensus chloroplast SSR marker length and sequence variation. *Genome* 49, 219-229.

CHUNG, W., KWABI-ADDO, B., ITTMANN, M., JELINEK, J., SHEN, L., YU, Y. and ISSA, J.P. (2008) Identification of novel tumor markers in prostate, colon and breast cancer by unbiased methylation profiling. *PLoS One* 3, e2079.

CHURCHILL, G.A. and DOERGE, R.W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* 138, 963-971.

COLLARD, B.C.Y., JAHUFER, M.Z.Z., BROUWER, J.B. and PANG, E.C.K. (2005) An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica* 142, 169-196.

CORLEY, R.H.V. (1977) Oil palm yield components and yield cycles. In: *International Development in Oil Palm*, ISP, Kuala Lumpur.

CORLEY, R.H.V. and GRAY, B.S. (1976) Growth and morphology. In: *Oil Palm Research*, pp. 7-21. Elsevier, Amsterdam.

CORLEY, R.H.V. and LEE, C.H. (1992) The physiological basis for genetic improvement of oil palm in Malaysia. *Euphytica* 60, 179-184.

CORLEY, R.H.V. and TINKER, P.B. (2003) *The Oil Palm* (4th ed.). Blackwell Science.

CRANE, C.F. and CRANE, Y.M. (2005). A nearest-neighboring-end algorithm for genetic mapping. *Bioinformatics* 21, 1579-1591.

CROSSA, J., BURGUENO, J., DREISIGACKER, S., VARGAS, M., HERRERA-FOESSEL, S.A., LILLEMO, M., SINGH, R.P., TRETHOWAN, R., WARBURTON, M., FRANCO, J., REYNOLDS, M., CROUCH, J.H. and ORTIZ, R. (2007) Association analysis of historical bread wheat germplasm using additive genetic covariance of relatives and population structure. *Genetics* 177, 1889-1913.

CRUZ, V.M.V., KILIAN, A. and DIERIG, D.A. (2013) Development of DArT marker platforms and genetic diversity assessment of the U. S. Collection of the new oilseed crop Lesquerella and related species. *PLoS One* 8, e64062.

CULLIS, C.A. and KUNERT, K.J. (2000) Isolation of tissue culture-induced polymorphisms in babanas by representational difference analysis. *Acta Horticulturae* 530, 421-428.

DARVASI, A., WEINREB, A., MINKE, V., WELLER, J.I. and SOLLER, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* 134, 943-951.

DE GIVRY, S., BOUCHEZ, M., CHABRIER, P., MILAN, T. and SCHIEX, T. (2005) CarthaGene: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics* 21, 1703-1704.

DEWAN, A.T., PARRADO, A.R., MATISE, T.C. and LEAL, S.M. (2002a) The map problem: a comparison of genetic and sequence-based physical maps. *The American Journal of Human Genetics* 70, 101-107.

DEWAN, A.T., PARRADO, A.R., MATISE, T.C. and LEAL, S.M. (2002b) Map error reduction: using genetic and sequence-based physical maps to order closely linked markers. *Human Heridity* 54, 34-44.

DOERGE, R.W. (2002) Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3, 43-52.

DOMÍNGUEZ-GARCÍA, M.C., BELAJ, A., DE LA ROSA, R., SATOVIC, Z., HELLER-USZYNSKA, K., KILIAN, A., MARTÍN, A. and ATIENZA, S.G. (2012) Development of DArT markers in olive (*Olea europaea* L.) and usefulness in variability studies and genome mapping. *Scientia Horticulturae* 136, 50-60.

DOYLE, J.J. and DOYLE, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemistry* 19, 11-15.

EDEM, D.O. (2002) Palm oil: Biochemical, physiological, nutritional haematological and toxicological aspects: a review. *Plant Foods for Human Nutrition* 57, 319-341.

EDWARDS, A., CIVITELLO, A., HAMMOND, H.A., CASKEY, C.T. (1991) DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *American Journal of Human Genetics* 49, 746–756.

ELLEGREN, H. (2004) Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics* 5, 435-445.

ELSHIRE, R.J., GLAUBITZ, J.C., SUN, Q., POLAND, J.A., KAWAMOTO, K., BUCKLER, E.S. and MITCHELL, S.E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379.

EVERTS, R.E., VERSTEEG, S.A., RENIER, C., VIGNAUX, F., GROOT, P.C., ROTHUIZEN, J. and VAN OOST, B.A. (2000) Isolation of DNA markers informative in purebred dog families by genomic representational difference analysis (gRDA). *Mammalian Genome* 11, 741-747.

FALK, C.T. (1989) A simple scheme for preliminary ordering of multiple loci: application to 45 CF families. *Progress in Clinical Biology Research* 329, 17-22.

FARKHARI, M., LU, Y., SHAH, T., ZHANG, S., NAGHAVI, M.R., RONG, T. and XU, Y. (2011). Recombination frequency variation in maize as revealed by genomewide single-nucleotide polymorphisms. *Plant Breeding* 130, 533-539.

FAY, M.F., COWAN, R.S. and LEITCH, I.J. (2005) The effects of nuclear DNA content (C-value) on the quality and utility of AFLP fingerprints. *Annals of Botany* 95, 237-246.

FELSKE, A. (2002) Streamlined representational difference analysis for comprehensive studies of numerous genomes. *Journal of Microbiological Methods* 50, 305-311.

FERREIRA, A., DA SILVA, M.F., SILVA, L.D.C.E. and CRUZ, C.D. (2006) Estimating the effects of population size and type on the accuracy of genetic maps. *Genetics and Molecular Biology* 29, 187-192.

FUCHS, R. and BLAKESLEY, R. (1983) Guide to the use of type II restriction endonucleases. *Methods in Enzymology* 101, 3.

FUJISAMA, M., HAYASHI, K., NISHIO, T., BANSO, T., OKADA, S., YAMATO, K.T., FUKUZAWA, H. and OHYAMA, K. (2001) Isolation of X and Y chromosome-specific DNA markers from a liverwort, Marchantia polymorpha, by representational difference analysis. *Genetics* 159, 981-985.

GAAFAR, A., UNZAGA, M.J., CISTERNA, R., CLAYO, F.E., URRA, E., AYARZA, R. and MARTIN, G. (2003) Evaluation of a modified single-enzyme amplified fragment length polymorphism technique for fingerprinting and differentiating of *Mycobacterium kansasii* type I isolates. *Journal of Clinical Microbiology* 41, 3846-3850.

GANAL, M.W., ALTMANN, T. and RÖDER, M.S. (2009) SNP identification in crop plants. *Current Opinions in Plant Biology* 12, 211-217.

GANCHEVA, A., POT, B., VAN HONACKER, K., HOSTE, B. and KERSTERS, K. (1999) A polyphasic approach towards the identification of strains belonging to *Lactobacillus acidophilus* and related species. *Systemic and Applied Microbiology* 22, 573-585.

GELDERMAN, H. (1975) Investigation on inheritance of quantitative characters in animals by gene markers. I. Methods. *Theoretical and Applied Genetics* 46, 300-319.

GHESQUIÉRE, M. (1984) Enzyme polymorphism in oil palm (*Elaeis guineensis* Jacq) I. Genetic control of 9 enzyme-systems. *Oleagineaux* 39, 561-574.

GHESQUIÉRE, M. (1985) Enzyme polymorphism in oil palm (*Elaeis guineensis* Jacq) II Variability and genetic structure of seven origins of oil palm. *Oleagineaux* 40, 529-540.

GIAMMANCO, G.M., MAMMINA, C., ROMANI, C., LUZZI, I., DIONISI, A.M. and NASTASI, A. (2007) Evaluation of a modified single-enzyme amplified fragment length polymorphism (SE-AFLP) technique for subtyping *Salmonella enteric* serotype Enteritidis. *Research in Microbiology* 158, 10-17.

GIBSON, J.R., SLATER, E., XERRY, J., TOMPKINS, D.S. and OWEN, R.J. (1998) Use of an amplified-fragment length polymorphism technique to fingerprint and differentiate isolates of *Helicobacter pylori. Journal of Clinical Microbiology* 36, 2580-2585.

GRANT, D. and SHOEMAKER, R.C. (2001) Plant gene mapping techniques. *Encyclopedia of Life Sciences.* John Willey and Son, Ltd.

GREWAL, T.S., ROSSNAGEL, B.G., POZNIAK, C.J. and SCOLES, G.J. (2008) Mapping quantitative trait loci associated with barley net blotch resistance. *Theoretical and Applied Genetics* 116, 529-539.

GUPTA, P.K, and VARSHNEY, R.K. (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica* 113,163–185.

GUTHERIE, N., GAPOR, A., CHAMBERS, A.F. and CARROLL, K.K. (1990) Inhibition of proliferation of Ostrogen receptor-negative MDA-MB-435 and positive MCF-7 human breast cancer cells by palm oil tocotrienols and tomoxifen, alone and combination. *Journal of Nutrition* 127, 544-548.

GUTHERIE, N., GAPOR, A., CHAMBERS, A.F. and CARROLL, K.K. (1997) Palm oil tocotrienols and plant flaranoids act synergistically with each other and with tomoxifen in inhibiting proliferation and growth of oestrogen receptor-negative MDA-MB-435 and positive MCF-7 human breast cancer cells in culture. *Asia Pacific Journal of Clinical Nutrition* 6, 41-45.

HACKETT, C.A. and BROADFOOT, L.B. (2003). Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. *Heredity* 90, 33-38.

HALDANE, J.B.S. (1919) The combination of linkage values, and the calculation of distances between the loci of linked factors. *Journal of Genetics* 8, 299-309.

HALL, T.A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41, 95-98.

HARDON, J.J., CORLEY, R.H.V. and LEE, C.H. (1987) Breeding and selecting the oil palm. In: *Improving vegetatively propagated crops*, pp. 63-81 (Eds. A.J. Abbot and R.K. Atkin). Academic Press, London.

HARTLEY, C.W.S. (1988) *The oil palm* (*Elaeis guineensis* Jacq.). Longman Scientific and Technology Publication. Wiley, New York.

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

HELENTJARIS, T., SLOCUM, M., WRIGHT, S., SCHAEFER, A. and NIENHUIS, J. (1986) Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphism. *Theoretical and Applied Genetics* 72, 761-769.

HELLER-USZYNSKA, K., CAIG, V., CARLING, J., EVERS, M., USZYNSKI, G., PIPERIDIS, G., GILMOUR, R., AITKEN, K., JACKSON, P., HUTTNER, E. and KILIAN, A. (2006) Diversity Arrays Technology (DArT) for high throughput, whole-genome molecular analysis in sugarcane. Tropical Crops Biotechnology Conference. Cairns, Australia, 16-19 August

HEMMAT, M., WEEDEN, N.F., MANGANARIS, A.G. and LAWSON, D.M. (1994) Molecular marker linkage map for apple. *Journal of Heredity* 84, 4-11.

HENDERSON, J. and OSBORNE, D.J. (2000) The oil palm in our lives: how this came about. *Endeavour* 24, 63-68.

HENRY, R. J. (2013) Evolution of DNA marker technology in plants. In: *Moleculer Markers in Plants*, pp. 1-19 (Ed. R. J. Henry). John Wiley & Sons, Inc.

HIPPOLYTE, I., BAKRY, F., SEQUIN, M., GARDES, L., RIVALLAN, R., RISTERUCCI, A.-M., JENNY, C., PERRIER, X., CARREEL, F., ARGOUT, X., PIFFANELLI, P., KHAN, I., MILLER, R.N.G., PAPPAS, G.J., MBEGUIE-A-MBEGUIE, D., MATSUMOTO, T., DE BERNARDINIS, V., HUTTNER, E., KILIAN, A., BAURENS, F.-C., D'HONT, A., COTE, F., COURTOIS, B. and GLAZMANN, J.-C. (2010) A saturated SSR/DArT linkage map of Musa

acuminate addressing genome rearrangements among bananas. *BMC Plant Biology* 10, 65.

HO, W.K. (2012) Oil palm tissue culture and DNA biomarkers for embryogenic potential. PhD Thesis. University of Nottingham.

HO, W.K., OOI, S.E., MAYES, S., NAMSIVAYAM, P. ONG-ABDULLAH, M. and CHIN, C.F. (2013) Methylation levels of a novel genetic element, EgNB3 as a candidate biomarker associated with the embryogenic competency of oil palm. *Tree Genetics and Genomes* 9, 1099-1107.

HO, Y.W. (2000) 2000 Annual Report United Plantations Berhad Research Department.

HOELTKE, H.J., ANKENBAUER, W., MUHLEGGER, K., REIN, R., SAGNER, G., SEIBEL, R., WALTER, T. (1995) The Digoxigenin (DIG) system for non-radioactive labeling and detection of nucleic acids – an overview. *Cell Molecular Biology* 41, 883-905.

HOLLESTELLE, A. and SCHUTTE, M. (2005) Representational Difference Analysis as a tool in the search for new tumor suppressor genes. In: *Methods in Molecular Medicine: Pancreatic Cancer: Methods and Protocols*, Vol 103, pp. 143-159 (Eds. G. Su). Humana Press Inc., Totowa, NJ.

HUBANK, M. and SCHATZ, D.G. (1994) Identifying differences in mRNA expression by representational difference analysis of cDNA. *Nucleic Acids Research* 22, 5640-5648.

HUMANN, F.C. and HARTFELDER, K. (2011) Representational Difference Analysis (RDA) reveals differential expression of conserved as well as novel genes during caste-specific development of the honey bee (*Apis mellifera* L.) ovary. *Insert Biochemistry and Molecular Biology* 41, 602-612.

IWATA, H., HAYASHI, T., TERAKAMI, S., TAKADA, N., SAWAMURA, Y. AND YAMAMOTO, T. (2013) Potential assessment of genome-wide association study and genomic selection in Japanese pear *Pyrus pyrifolia*. *Breeding Science* 63, 125-140.

JACCOUD, D., PENG, K., FEINSTEIN, D. and KILIAN, A. (2001) Diversity Arrays: a solid state technology for sequence information independent genotyping. *Nucleic Acids Research* 29, e25

JACK, P.L., DIMITRIJEVIC, T.A.F. And MAYES, S. (1995) Assessment of nuclear, mitochondrial and chloroplast RFLP markers in oil palm (*Elaeis guineensis* Jacq). *Theoretical and Applied Genetics* 90, 643-649.

JACK, P.L., JAMES, C., PRICE, Z., RANCE, K., GROVES, L., CORLEY, R.H.V., NELSON, S. and RAO, V. (1998) Application of DNA markers in oil palm breeding. In: *Proceedings of the 1998 International Oil Palm Congress-Commodity of the Past, Today and Future*, pp. 315-324 (Eds. A. Jatmika). Indonesian Oil Palm Research Institute, Medan, Indonesia.

JACQUEMARD, J.C. (1998) *The tropical agriculturist; oil palm*. MacMillan Education Ltd, London and Basingstoke, UK.

JANSEN, J., DE JONG, A.G., and VAN OOIJEN, J.W. (2001) Constructing dense genetic linkage maps. *Theoretical and Applied Genetics* 102, 1113-1122.

JANSEN, R.C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics* 135, 1457-1468.

JANSEN, R.C. and STAM, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* 136, 1447-1455.

JING, H.-C., BAYON, C., KANYUKA, K., BERRY, S., WENZL, P., HUTTNER, E., KILIAN, A. and HAMMOND-KOSACK, K.E. (2009). DArT markers: diversity analyses, genome comparison, mapping and integration with SSR markers in *Triticum monococcum*. *BMC Genomics* 10**,** 458.

JOHNSON, R. (2004) Marker-assisted selection. *Plant Breeding Review* 24, 293-309.

JONES, C.J., EDWARDS, K.J., CASTAGLIONE, S., WINFIELD, M.O., SALA, F., VAN DE WIEL, C., BREDEMEIJER, G., VOSMAN, B., MATTHES, M., DALY, A., BRETTSCHNEIDER, R., BETTINI, P., BUIATTI, M., MAESTRI, E., MALCESCHII, A., MARMIROLI, N., AERT, R., VOLCKAERT, G., RUEDA, T., LINACERO, R., VAZQUES, A. and KARP. A. (1997) Reproducibility testing of RAPD, AFLP, SSR markers in plants by a network of European Laboratories. *Molecular Breeding* 3, 381-390.

JULIE, K., ANN T., CARON, J., IAN, K. and IAN, A. (2013) A DArT marker genetic map of perennial ryegrass (*Lolium perenne* L.) integrated with detailed comparative mapping information; comparison with existing DArT marker genetic maps of *Lolium perenne*, *L. multiflorum* and *Festuca pratensis*. *BMC Genomics* 14, 437.

KAHL, G., MAST, A., TOOKE, N., SHEN, R., and VAN DEN BOOM, D. (2005) Single Nucleotide Polymporhism: Detection techniques and their potential for genotyping and genome mapping. In: *The Handbook of Plant Genome Mapping. Genetic and Physical Mapping,* pp.75-107 (Eds. K. Meksem and G. Kahl). WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

KALIA, R.K., RAI, M.K., KALIA, S., SINGH, R. and DHAWAN, A.K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* 177, 309-334.

KANAGARAJ, P., PRINCE, K.S.J., SHEEBA, J.A., BIJI, K.R., PAUL, S.B., SENTHIL, A. and BABU, R.C. (2010) Microsatellite markers linked to drought resistance in rice (*Oryza sativa* L.) *Current Science* 98, 836-839.

KANEDA, A., TAKAI, D., KAMINISHI, M., OKOCHI, E. and USHIJIMA, T. (2003) Methylation-sensitive representational difference analysis and its application to cancer research. *Annals of the New York Academy of Sciences* 983, 131-141.

KARDOLOUS, J.P., ECK, H.J. and BERG, R.G. (1998) The potential of AFLPs in biosystematics: a first application in *Solanum* taxonomy (Salanaceae). *Plant Systematics and Evolution* 210, 87-103.

KAU, C.–H., ZENG, Z.–B. and TEASDALE, R.D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics* 152, 1203-1216.

KEARSEY, M.J. (1998) The principles of QTL analysis (a minimal mathematics approach). *Joournal of Experimental Botany* 49, 1619-1623.

KEARSEY, M.J. and FARQUHAR, A.G.L. (1998) QTL analysis in plants; where are we now? *Heridity* 80, 137-142.

KHAN, M.A., HAN, Y., ZHAO, Y.F., TROGGIO, M. and KORBAN, S.S. (2012) A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. *PLoS One*, 7, e47864.

KHAN, M.A. and KORBAN, S.S. (2012) Association mapping in forest trees and fruit crops. *Journal of Experimental Botany* 63, 4045-4060.

KHOSLA, P. and SUNDRAM, K. (1996) Effects of dietary fatty acid composition on plasma cholesterol. *Progress in Lipid Research* 35, 93-132.

KILIAN, A., HUTTNER, E., WENZL, P., JACCOUD, D., CARLING, J. CAIG, V., EVERS, M., HELLER-USZYNSKA, K., USZYNSKI, G., CAYLA, C, PATARAPUWADOL, S., XIA, L., YANG, S. and THOMSON, B. (2005) The fast and the cheap: SNP and DArT-based whole genome profiling for crop improvement. In: *Proceedings of the international congress "In the wake of the double helix: from the green evolution to the gene evolution"*, pp. 443-661 (Eds. R. Tuberosa, R.L. Phillips, M. Gale) Avenue Media, Bologna, Italy.

KILIAN, A., WENZL, P., HUTTNER, E., CARLING, J., XIA, L., BLOIS, H., CAIG, V., HELLER-USZYNSKA, K., JACCOUD, D., HOPPER, C., ASCHENBRENNER-

KILIAN, M., EVERS, M., PENG, K., CAYLA, C., HOK, P. and USZYNSKI, G. (2012) Diversity Arrays Technology: A generic genome profiling technology on open platforms. *Methods in Molecular Biology* 888, 67-88.

KIMMERLY, W.J., KYLE, A.L., LUSTRE, V.M., MARTIN, C.H., PALAZZOLO, M.J. (1994) Direct sequencing of terminal regions of genomic PI clones: a general strategy for the design of sequence-tagged site markers. *GATA* 11, 117-128.

KINZLER, K.W. and VOGELSTEIN, B. (1989) Whole genome PCR: Application to the identification of sequences bound by gene regulatory proteins. *Nucleic Acids Research* 17, 2645-3653.

KISIELOW, P. and CEBRAT, M. (2007) Identification of genes involved in positive selection of CD4+8+ thymocytes: expanding the inventory. *Immunological Investigations: A Journal of Molecular and Cellular Immunology* 36, 353-369.

KISS, G.B., CSANADI, G., KALMAN, K., KALO, P. and OKRESZ, L. (1993) Construction of a basic linkage map for alfalfa using RFLP, RAPD, isozyme and morphological markers. *Molecular and General Genetics* 238, 129-137.

KLEIN-LANKHORST, R., RIETVELD, P., MACHIELS, B., VERKERK, R., WEIDE, R., GEBHARDT, C. and KORNEEF, M. (1991) RFLP markers linked to the root knot nematode gene Mi in tomato. *Theoretical and Applied Genetics* 81, 661-667

KOCHERT, G. (1994) RFLP technology. In: *DNA-based Markers in Plants*, pp. 8-38 (Eds. R.L. Phillips and I.K. Vasil). Kluwer Academic Publishers, Dordrecht.

KOELEMAN, J.G.M., STOOF, J., BIESMANS, D.J., SAVELKOUL, P.H.M., and VAN DER BROUCKE-GRAULS, C.M.J.E. (1998) Comparisons of ARDRA, RAPD and AFLP fingerprinting for identication of *Acinetobacter* genomic species and typing of *Acinetobacter baumannii. Journal of Clinical Microbiology* 36, 2522-2529.

KOK, E.J., FRANSSEN-VAN HAL, N.L., WINNUBST, L.N., KRAMER, E.H., DIJKSMA, W.T., KUIPER, H.A. and KEIJER, J. (2007) Assessment of representational difference analysis (RDA) to construct informative cDNA microarrays for gene expression analysis of species with limited transcriptome information, using red and green tomatoes as a model. *Journal of Plant Physiology* 164, 337-349.

KORNBLUM, H. and GESCHWIND, D. (2001) The use of representational difference analysis and cDNA microarrays in neural repair research. *Restorative Neurology and Neuroscience* 18, 89-94.

KORTE, A. and FARLOW, A. (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* 9, 29.

KOSAMBI, D.D. (1944) The estimation of the map distance from recombination values. *Annals of Eugenics* 12, 172-175.

KRIKORIAN, A.D. (1989) The context and strategies for tissue culture of date, African oil and coconut palms. In: *Applications of biotechnology in forestry and horticulture*, pp. 119-144 (Eds. V. Dhawan). Plenum, New York.

KRUSKAL, W.H. and WALLIS, W.A. (1952) Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47, 583-621.

KULLAN, A.R.K., VAN DYK, M.M., JONES, N., KANZLER, A., BAYLEY, A. and MYBURG, A.A. (2012) High-density genetic linkage amps with over 2,400 sequence-anchored DArT markers for genetic dissection in an F2 pseudo-backcross of *Eucalyptus grandis* x *E. urophylla*. *Tree Genetics and Genomes* 8, 163-175.

KULRATNE, R.S., SHAH, F.H. and RAJANAIDU, N. (2000) Investigation on genetic diversity in African natural oil palm populations and Deli dura using AFLP primers. In: *Proceedings of the International Symposium on Oil Palm Genetic Resources and Utilization*, pp. X1-X39 (Eds. N. Rajanaidu and D. Ariffin). Malaysian Palm Oil Board, Kuala Lumpur.

KUMAR, S. and BLAXTER, M.L. (2010) Comparing *de novo* assemblers for 454 transcriptome data. *BMC Genomics* 11, 571.

KUNERT, K.J., BAAZIZ, M. and CULLIS, C.A. (2001) Techniques for determination of True-to-type Date Palm (*Phoenix dactylifera* L.) plants: A literature review. *Emirates Journal of Agricultural Science* 15, 1-16.

KUNERT, K.J., VORSTER, J., BESTER, C. and CULLIS, C.A. (2002) DNA microchip technology in the plant tissue culture industry. In: *Crop Biotechnology,* pp. 86-96 (Eds. K. Rajasekaran, T.J. Jacks, J.W. Finley) American Chemical Society Symposium Series 829.

KUSHAIRI, A. and RAJANAIDU, N. (2000) Breeding populations, seed production and nursery management. In: *Advances in oil palm research,* Vol. I, pp. 39-98 (Eds. Y. Basiron, B.S. Jalani and K.W. Chan) Malaysian Palm Oil Board, Kuala Lumpur.

KUSHAIRI, A., RAJANAIDUI, N., JALANI, B.S., RAFIL, M.Y. and MOHD, D.A. (1999). PORIM Oil Palm Planting Materials. *PORIM Bulletin*, Number 38, pp. 1-13.

LAGERCRANTZ, U., ELLEGREN, H. and ANDERSSON, L. (1993) The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. *Nature Genetics* 30, 194-200.

LAM, M.K., TAN, K.T., LEE, K.T. and MOHAMED, A.R. (2009) Malaysian palm oil: Surviving the food versus fuel dispute for a sustainable future. *Renewable and Sustainable Energy Reviews* 13, 1456-1464.

LAMMENS, E., CEYSSENS, P.J., VOET, M., HERTVELDT, K., LAVIGNE, R. and VOLCKAERT, G. (2009) Representational Difference Analysis (RDA) of bacteriophage genomes. *Journal of Microbiological Methods* 77, 207-213.

LANDE, R. and THOMPSON, R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* 124, 743-756.

LANDER, E.S., GREEN, P., ABRAHAMSON, J., BARLOW, A., DALY, M.J., LINCOLN, S.E. and NEWBURG, L. (1987) Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1, 174-181.

LATIFF, A. (2000) The Biology of the Genus *Elaeis*. In: *Advances in Oil Palm Research*, Volume 1, pp. 19-38 (Eds. Y. Basiron, B.S. Jalani and K.W. Chan) Malaysian Palm Oil Board.

LAWRENCE, M.R., MARSHALL, D.F. and DAVIES, P. (1995) Genetics of genetic conservation. II. Samples size when collecting seed of cross-pollinating species and the information that can be obtained from the evaluation of material held in gene banks. *Euphytica* 84, 101-107.

LAZZI, C., BOVE, C.G., SGARBI, E., GATTI, M., LA GIOIA, F., TORRIANI, S. and NEVIANI, E. (2009) Application of AFLP fingerprint analysis for studying the biodiversity of *Streptococcus thermophilus*. *Journal of Microbiological Methods* 79, 48-54.

LE, T., CHIARELLA, J., SIMEN, B.B., HANCZARUK, B., EGHOLN, M., LANDRY, M.L., DIECKHAUS, K., RESEN, M.I. and KOZAL, M.J. (2009) Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS One* 4, e6079.

LEBRUN, P., BAUDOUIN, L., BOURDEIX, R., KONAN, J.L., BARKER, J.H., ALDAM, C., HERRÁN, A. and RITTER, E. (2001) Construction of a linkage map of the Rennell Island Tall coconut type (Cocos nucifera L.) and QTL analysis for yield characters. *Genome* 44, 962-970.

LEE, W.W. and CHEAH, S.C. (2009) Decoding the oil palm genome. In: *International seminar on oil palm genomics and its application to oil palm breeding*, (Eds. C. Bakoume, N. Rajanaidu and A. Mohd. Din) International Society for Oil palm breeders (ISOPB).

LEGESSE, B.W., MYBURG, A.A., PIXLEY, K.V. and BOTHA, A.M. (2007) Genetic diversity of African maize inbred lines revealed by SSR markers. *Hereditas* 144, 10-17.

LEHMANN, E.L. (1975) Nonparametrics. McGraw-Hill, New York.

LEWIN, B. (1994) Genome size and genetic content. In: *Genes V.*, pp. 657-676. Oxford University Press, Oxford.

LI, H., VAILLANCOURT, R., MENDHAM, N. and ZHOU, M. (2008) Comparative mapping of quantitative trait loci associated with waterlogging tolerance in barley (*Hordeum vulgare L.*). *BMC Genomics* 9, 401.

LI, H.Y., GUO, Z.F. and ZHU, Y.X. (1998) Molecular cloning and analysis of a pea cDNA that is expressed in darkness and very rapidly induced by gibberellic acid. *Molecular and General Genetics* 259, 393-397.

LINCOLN, S., DALY, M. and LANDER, E. (1993a) Constructing genetic linkage maps with MAPMAKER/EXP. Version 3.0. Whitehead Institute for Biomedical Research Technical Report, 3rd Edn.

LINCOLN, S., DALY, M. and LANDER, E. (1993b) Mapping genes controlling quantitative traits using MAPMAKER/QTL. Version 1.1. Whitehead Institute for Biomedical Research Technical Report, 2nd Edn.

LING, J.Q., KOJIMA, T., SHIRAIWA, M. and TAKAHARA, H. (2003) Cloning of two cysteine proteinase genes, CysP1 and CysP2, from soybean cotyledons by cDNA representational difference analysis. *Biochimica et Biophysica Acta* 1627, 129-139.

LISITSYN, N.A. (1995) Representational difference analysis: finding the difference between genomes. *Trends in Genetics* 11, 303-307.

LISITSYN, N.A., LISITSINA, N.M., DALBAGNI, G., BARKER, P., SANCHEZ, C.A., GNARRA, J., LINEHAN, W.M., REID, B.J. and WIGLER, M.H. (1995) Comparative genomic analysis of tumors: Detection of DNA losses and amplification. *The Proceedings of the National Academy of Sciences of the United States of America* 92, 151-155.

LISITSYN, N.A., LISITSYN, N., and WIGLER, M.H. (1993) Cloning the differences between two complex genomes. *Science* 259, 946-951.

LISITSYN, N. and WIGLER, M.H. (1995) Representational difference analysis in the detection of genetic lesions in cancer. *Methods in Enzymology* 254, 291-304.

LITT, M. and LUTY, J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *American Journal of Human Genetics* 44, 397-401.

LIU, G.E. (2009) Applications and case studies of the next-generation sequencing technologies in food, nutrition and agriculture. *Recent Patents on Food, Nutrition and Agriculture* 1, 75-79.

LIU, H.B. (1998). Statistical Genomics, Linkage, Mapping and QTL analysis. pp. 611, CRC, Boca Raton, Florida.

LIU, X., HUA, Z., and WANG, Y. (2010) Quantitative trait locus (QTL) analysis of percentage grains chalkiness using AFLP in rice (*Oryza sativa* L.). *African Journal of Biotechnology* 10, 2399-2405.

LIU, Z.J. and CORDES, J.F. (2004) DNA marker technologies and their applications in aquaculture genetics. *Aquaculture* 238, 1-37.

LORIDON, K., MCPHEE, K., MORIN, J., DUBREUIL, P., PILET-NAYEL, M.L., AUBERT, G., RAMEAU, C., BARANGER, A., COYNE, C., LEJEUNE-HÈNAUT, I. and BURSTIN, J. (2005). Microsatellite marker polymorphism and mapping in pea (*Pisum sativum* L.). *Theoretical and Applied Genetics* 111, 1022-1031.

LOW, E.T.L., ALIAS, H., BOON, S.B., SHARIFF, E.M., TAN, C.Y.A., OOI, L.C.L., CHEAH, S.C., RAHA, A.R., WAN, K.L. and SINGH, R. (2008) Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs:Identifying genes associated with callogenesis and embryogenesis. *BMC Plant Biology* 8, 62.

LUO, L., MAO, Y. and XU, S. (2003) Correcting the bias in estimation of genetic variances contributed by individual QTL. *Genetica* 119, 107-113.

LUO, Q., HE, Y., CHENG, C., ZHANG, Z., LI, J. HUANG, S. and CHEN, J. (2013) Integration of high-resolution physical and genetic map reveals differential recombination frequency between chromosomes and the genome assembling quality in cucumber. *PLoS One* 8, e62676.

LYAMICHEV, V., BROW, M.A.D. and DAHLBERG, J.E. (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* 260, 778-783.

MACE, E.S., RAMI, J-F., BOUCHET, S., KLEIN, P.E., KLEIN, R.R., KILIAN, A., WENZL, P., XIA, L., HALLORAN, K. and JORDAN, D.R. (2009). A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array technology (DArT) markers. *BMC Plant Biology* 9, 13.

MACE, E.S., XIA, L., JORDAN, D.R., HALLORAN, K., PARH, D.K., HUTTNER, E., WENZL, P. and KILIAN, A. (2008) DArT markers: diversity analyses and mapping in *Sorghum bicolor*. *BMC Genomics* 9, 26.

MACKAY, T.F.C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* 35, 303-339.

MAHERAN, A.B., AW, K.T., ABU ZARIN, O. and CHIN, C.W. (1995) Vegetative propagation of oil palm (*Elaeis guineensis*) from laboratory to field-FELDA's experience. In: *Proceedings of the 1993 PORIM International Palm Oil Congress on Update and Vision*, pp. 99-113 (Eds. S., Jalani, D. Ariffin, N. Rajanaidu, M.D. Tayeb, K. Paranjothy, M.W. Basri, I.E. Hanson and K.C. Chang) Palm Oil Research Institute of Malaysia, Kuala Lumpur.

MAIZURA, I., RAJANAIDU, N., ZAKRI, A.H. and CHEAH, S.C. (2006) Assessment of genetic diversity in oil palm (*Elaeis guineensis* Jacq.) using restriction fragment length polymorphism (RFLP). *Genetics Research Crop Evolution* 53, 187-195.

MANGIN, B., GOFFINET, B. and REBAI, A. (1994) Constructing confidence intervals for QTL location. *Genetics* 138, 1301-1308.

MANLY, K.F., CUDMORE, R.H.Jr., and MEER, J.M. (2001) Map Manager QTX, cross-platform software for genetic mapping. *Mammalian Genome* 12, 930-932.

MANN, H.B. and WHITNEY, D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18, 50-60.

MANSFIELD, E.S., WORLEY, J.M., MCKENZIE, S.E., SURREY, S., RAPPAPORT, E. and FORTINA, P. (1995) Nucleic acid detection using non-radioactive labeling methods. *Molecular and Cellular Probes* 9, 145-156.

MARDIS, E.R. (2008a) Next-Generation DNA Sequencing Methods. *Annual Review Genomics Human Genetics* 9, 387-402.

MARDIS, E.R. (2008b) The impact of next-generation sequencing technology on genetics. *Trends in Genetics* 24, 133-141.

MARGULIES, M., EGHOLM, M., ALTMAN, W.E., ATTIYA, S., BADER, J.S., BEMBEN, L.A., BERKA, J., BRAVEMAN, M.S., CHEN, Y.-J., CHEN, Z., DEWELL, S.B., DU, L., FIERRO, J.M., GOMES, X.V., GODWID, B.C., HE, W., HELGESEN, S., HO, C.H., IRZYK, G.P., JANDO, S.C., ALENQUER, M.L.I., JARVIE, T.P., JIRAGE, K.B., KIM, J.-B., KNIGHT, J.R., LANZA, J.R., LEAMON, J.H., LEFKOWITZ, S.M., LEI, M., LI, J., LOHMAN, K.L., LU, H., MAKHIJANI, V.B., MCDADE, K.E., MCKENNA, M.P., MYERS, E.W.,

NICKERSON, E., NOBILE, J.R., PLANT, R., PUC, B.P., RONAN, M.T., ROTH, G.T., SARKIS, G.J., SIMONS, J.F., SIMPSON, J.W., SRINIVASAN, M., TARTARO, K.R., TOMASZ, A., VOGT, K.A., VOLKMER, G.A., WANG, S.H., WANG, Y., WEINER, M.P., YU, P., BEGLEY, R.F. and ROTHBERG, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376-380.

MARIA, M., CLYDE, M.M. and CHEAH, S.C. (1995). Cytological analysis of *Elaeis guineensis* (*tenera*) chromosomes. *Elaeis* 7, 122-134.

MARTH, G.T., KORF, I., YANDELL, M.D., YEH, R.T., GU, Z., ZAKERI, H., STITZIEL, N.O., HILLIER, L., KWOK, P.Y. and GISH, W.R. (1999) A general approach to single nucleotide polymorphism discovery. *Nature Genetics* 23, 452-456.

MARTIN, G.B., WILLIAMS, J.G.K. and TANKSLEY, S.D. (1991) Rapid identification of markers linked to a *Pseudomonas* resistance gene in tomato by using random primers and near-isogenic lines. *Proceedings of the National Academy of Sciences the United States of America* 88, 2336-2340.

MATTHES, M., SINGH, R., CHEAH, S.C. and KARP, A. (2001) Variation in oil palm (*Elaeis guineensis* Jacq.) tissue culture-derived regenerants revealed by AFLP with methylation-sensitive enzymes. *Theoretical Applied Genetics* 102, 971-979.

MAXAM, A.M. and GILBERT, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* 74, 560-564.

MAYES, S., HAFEEZ, F., PRICE, Z., MACDONALD, D., BILLOTTE, N. and ROBERTS, J. (2008) Molecular research in oil palm, the key oil crop for the future. In: *Genomics of Traopical Crop Plants*, pp. 371-404 (Eds. P. H. Moore and R. Ming) Springer.

MAYES, S., JACK, P.L., MARSHALL, D.F. and CORLEY, R.H.V. (1997) The construction of an RFLP linkage map for oil palm. *Genome* 40, 116-122.

MAYES, S., JAMES, C.M., HORNER, S.F., JACK, P.L. and CORLEY, R.H.V. (1996) The application of restriction fragment length polymorphism for the genetic fingerprinting of oil palm (*Elaeis guineensis* Jacq.). *Molecular Breeding* 2, 175-180.

MAZUR, B.J. and TINGEY, S.V. (1995) Genetic mapping and introgression of genes of agronomic importance. *Current Opinion in Biotechnology* 6, 175-182.

MBA, C. and TOHME, J. (2005) Use of AFLP markers in surveys of plant diversity. *Methods in Enzymology* 395, 177-201.

MCCOUCH, S.R., CHEN, X., PANAUD, O., TEMNYKH, S., XU, Y., CHO, Y., HUANG, N., ISHII, T., and BLAIR, M. (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. *Plant Molecular Biology* 35, 89-99.

MCCOUCH, S.R., KOCHERT, G., YU, Z.H., WANG, Z.Y., KHUSH, G.S., COFFMAN, W.R. and TANKSLEY, S.D. (1988) Molecular mapping of rice chromosomes. *Theoretical and Applied Genetics* 76, 815-829.

MELCHINGER, A.E., UTZ, H.F., SCHÖN, C.C. (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics* 149, 383-403.

MEUDT, H.M. and CLARKE, A.C. (2007) Almost forgotten or latest practice? AFLP applications, analyses and advances. *Trends in Plant Science* 12, 106-117.

MEUNIER, J. and GASCON, J.P. (1972) Le schéma général d'amélioration du palmier à huile à l'I.R.H.O. *Oleagineux* 27, 1-12.

MICHELMORE, R.W., PARAN, I. and KESSELI, R.V. (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detech markers in specific genomic regions by using segregating populations. *The Proceedings of the National Academy of Science of the United States of America* 88, 9828-9832.

MIKULÁŠKOVÁ, E., FÉR, T and KUČABOVÁ, V. (2012) The effect of different DNA isolation protocol and AFLP fingerprinting optimizations on error rate estimates in the bryophyte *Campylopus introflexus. Lindbergia* 35, 7-17.

MILLER, M.R., DUNHAN, J.P., AMORES, A., CRESKO, W.A. and JOHNSON, E.A. (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research* 17, 240-248.

MILLER, N.J., KIFT, N.B. and TATCHELL, G.M. (2005) Host-associated populations in the lettuce root aphid, *Pemphigus bursarius* (L.). *Heredity* 94, 556–564.

MINDER, A.M. and Widmer, A. (2008) A population genomic analysis of species boundaries: neutral processes, adaptive divergence and introgression between two hybridizing plant species. *Molecular Ecology* 17, 1552–1563.

MOHAN, M., NAIR, S., BHAGWAT, A., KRISHNA, T.G., YANO, M., BHATIA, C.R. and SASAKI, T. (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding* 3, 87-103.

MOHD DIN, A., RAJANAIDU, N. and KUSHAIRI, A. (2005) Exploitation of genetic variability in oil palm. In: *Proceedings of MOSTA best practices workshops: agronomy and crop* management, pp.19-42. Malaysia Oil Science Technology Association.

MOLENAAR, A.J., HARRIS, D.P., RAJAN, G.H., PEARSON, M.L., CALLAGHAN, M.R., SOMMER, L., FARR, V.C., ODEN, K.E., MILES, M.C., PETROVA, R.S., GOOD, L.L., SINGH, K., MCLAREN, R.D., PROSSER, C.G., KIM, K.S., WIELICZKO, R.J., DINES, M.H., JOHANNESSEN, K.M., GRIGOR, M.R., DAVIS, S.R. and STELWAGEN, K. (2009) The acute-phase protein serum amyloid A3 is expressed in the bovine mammary gland and plays a role in host defence. *Biomarkers* 14, 26-37.

MOLOSIWA, O.O. (2012) Genetic diversity and population structure analysis of bambara groundnut [*Vigna subterranean* (L.) Verdc.] landraces using morpho-agronomic characters and SSR markers. University of Nottingham, PhD Thesis.

MONDINI, L., NOORANI, A. and PAGNOTTA, M.A. (2009) Assessing plant genetic diversity by molecular tools. *Diversity* 1, 19-35.

MONEY, T., READER, S., QU, L.J., DUNFORD, R.P. and MOORE, G. (1996) AFLP-based mRNA fingerprinting. *Nucleic Acids Research* 24, 2616-2617.

MONTOYA, C., LOPES, R., FLORI, A., CROS, D., CUELLAR, T., SUMMO, M., ESPEOUT, S., RIVALLAN, R., RISTERUCCI, A.-M., BITTENCOURT, D., ZAMBRANO, J.R., ALARCÓN, G.W.H., VILLENEUVE, P., PINA, M., NOUY, B., AMBLARD, P., RITTER, E., LEROY, T. and BILLOTTE, N. (2013) Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross from *Elaeis oleifere* (H. B. K.) Cortés and oil palm (*Elaeis guineensis* Jacq.) *Tree Genetics and Genomes* 9, 1207-1225.

MORETZSOHN, M.C., NUNES, C.D., FEREIRRA, M.E. and GRATTAPAGLIA, D. (2000) RAPD linkage mapping of the shell thickness locus in oil palm (*Elaeis guineensis* Jacq). *Theoretical and Applied Genetics* 100, 63-70.

MORGANTE, M., HANAFEY, M. and POWELL, W. (2002) Microsatellites are preferentially associate with nonrepetitive DNA in plant genomes. *Nature Genetics* 30, 194-200.

MUCHERO, W., DIOP, N., BHAT, P., FENTON, R., WANAMAKER, S., POTTORFF, M., HEARNE, S., CISSE, N., FATOKUN, C., EHLERS, J., ROBERTS, P.A. and CLOSE, T.J. (2009) A consensus genetic map of cowpea [Vigna unguiculata (L) Walp.] and synteny based on EST-derived SNPs. *Proceedings of the National Academy of Sciences of the United States of America* 106, 18159-18164.

MULLIS, K.B., FALOONA, F.A., SCHARF, S.J., SAIKI, S.K., HORN, G.T. and ERLICH, H.A. (1986) Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harbour Symposia on Quantitative Biology* 51, 263-273.

MYLES, S., CHIA, J.M., HURWITZ, B., SIMON, C., ZHONG, G.Y., BICKLER, E. and WARE, D. (2010) Rapid Genomic Characterization of the Genus Vitis**.** *PLoS One* 5**,** e8219.

NARIMAN, S.A. (2013) Genetic Analysis of Plant Morphology in Bambara Groundnut [*Vigna subterranean* (L.) Verdc.]. University of Nottingham, PhD Thesis.

NEERAJA, C., MAGHIRANG-RODRIGUEZ, R., PAMPLONA, A., HEUER, S., COLLAR, B.C., SEPTININGSIH, E.M., VERGARA, G., SANCHEZ, D., XU, K., ISMAIL, A.M. and MACKILL, D.J. (2007) A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. *Theoretical and Applied Genetics* 115, 767-776.

NELSON, J.C. (1997). QGENE: Software for marker-based genomic analysis and breeding. *Molecular Breeding* 3, 239-245.

NELSON, J.C. (2005). Methods and Software for Genetic Mapping. In: *The Handbook of Plant Genome Mapping. Genetic and Physical Mapping*, pp.53-74 (Eds. K. Meksem & G. Kahl) WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

NESARETNAM, K., STEPHEN, R., DILS, R. and DARBREY, P. (1988) Tocotrienols inhibit the growth of human breast cancer cells irrespective of oestrogen receptor status. *Lipids* 33, 461-469.

NEUMANN, K., KOBILISKI, B., DENČIĆ, S., VARSHNEY, R.K. and BORNER, A. (2010) Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum* L.) *Molecular Breeding* 27, 37-58.

NGUYEN, H.T. and WU, X. (2005) Molecular marker Systems for Genetic Mapping. In: *The Handbook of Plant Genome Mapping. Genetic and Physical Mapping*, pp. 25-51 (Eds. K. Meksem & G. Kahl) WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

NICHOLAS, A.T., KILIAN, A., WIGHT, C.P., HELLER-USZYNSKA, K., WENZL, P., RINES, H.W., BJORNSTAD, A., HOWARTH, C.J., JANNINK, J-L., ANDERSON, J.M., ROSSNAGEL, B.G., STUTHMAN, D.D., SORRELLS, M.E., JACKSON, E.W., TUVESSON, S., KOLB, F.L., OLSSON, O., FEDERIZZI, L.C., CARSON, M.L., OHM, H.H., MOLNAR, S.J., SCOLES, G.J., ECKSTEIN, P.E., BONMAN, J.M., CEPLITIS, A. and LANGDON, T.

(2009) New DArT markers for oat provide enhanced map coverage and global germplasm characterization. *BMC Genomics* 10, 39.

OH, T.J. and CULLIS, C.A. (2003) Labile DNA sequences in flax identified by combined sample representational difference analysis (csRDA). *Plant Molecular Biology* 52, 527-536.

OH, T.J., CULLIS, M.A., KUNERT, K., ENGELBORGHS, I., SWENNEN, R. and CULLIS, C. A. (2007) Genomic changes associated with somaclonal variation in banana (*Musa* spp.). *Physiologia Plantarum* 129, 766-774.

OKWUAGWU, C.O. and OKOLO, E.C. (1992) Maternal inheritance of kernel size in the oil palm (*Elaeis guineensis* Jacq). *Elaeis* 4, 72-78.

OKWUAGWU, C.O. and OKOLO, E.C. (1994) Genetic control of polymorphism for kernel to fruit ratio in oil palm (*Elaeis guineensis* Jacq). *Elaeis* 6, 75-81.

ONG. A.M. and OOI, S.E. (2006) Biomarkers: Finding a Niche in Oil Palm Tissue Culture. Submitted to Oil Palm Bulletin.

ONG, A.S.H. and GOH, S.H. (2002) Palm oil: a healthful and cost effective dietary omponent. In: *Food and Nutrition Bulletin*, Volume 23, no. 1, pp. 11-22. United Nations University Press.

ONG, A.S.H., HASSAN, A.H., BASIRON, Y., CHOO, Y.M. and MOHD TOP, A.G. (1990) Palm vitamin E and palm diesel pilot plants. In: *Proc. Symp. New developments in palm oil,* pp. 45-55 (Eds. K. G. Berger) Palm Oil Research Institute Malaysia, Kuala Lumpur.

PARAN, I., KESSELI, R. and MICHELMORE, R. (1991) Identification of restriction-fragment-length-polymorphism and random amplified polymorphic DNA markers linked to downy mildew resistance genes in lettuce, using near isogenic lines. *Genome* 34, 1021-1027.

PARAN, I. and ZAMIR, D. (2003) Quantitative traits in plants: beyond the QTL. *Trends in Genetics* 19, 303-306.

PARH, D.K., JORDAN, D.R., AITKEN, E.A.B., MACE, E.S., JUN-AI, P., MCINTYRE, C.L. and GODWIN, I.D. (2008) QTL analysis of ergot resistance in sorghum. *Theoretical and Applied Genetics* 117, 369-382.

PARIS, M. and DESPRES, L. (2012) Identifying insecticide resistance genes in mosquito by combining AFLP genome scans and 454 pyrosequencing. *Molecular Ecology* 21, 1672–1686.

PARK, D.S., LEE, S.K., LEE, J.H., SONG, M.Y., SONG, S.Y., KWAK, D.Y., YEO, U.S., JEON, N.S., PARK, S.K., YI, G., SONG, Y.C., NAM, M.H., KU, Y.C. and JEON, J.S. (2007) The identification of candidate rice genes that confer resistance to the brown planthopper (Nilaparvata lugens) through representational difference analysis. *Theoretical and Applied Genetics* 115, 537-547.

PATERSON, A.H. (1996) Making genetic maps. In: *Genome mapping in plants*, pp. 23-39 (Ed. A.H. Paterson) Academic Press, Austin, Texas.

PATERSON, A.H., TANKSLEY, S.D. and SORRELLS, M.E. (1991) DNA markers in plant improvement. *Advances in Agronomy* 46, 39-90.

PELGAS, B., BOUSQUET, J., MEIRMANS, P.G., RITLAND, K. and ISABEL, N. (2011) QTL mapping in white spruce: gene maps and genomic regions underlying adaptive traits across pedigrees, years and environments. *BMC Genomics* 12, 145.

PENNER, G. (1996) RAPD analysis of plant genomes, In: *Methods of Genome Analysis in Plants*, pp. 251-268 (Ed. P.P. Jauhar). CRC Press, Boca Raton.

PENNER, G.A., BUSH, A., WISE, R., KIM, W., DOMIER, L., KASHA, K., LAROCHE, A., SCOLES, G., MOLNAR, S.J. and FEDAK, G. (1993) Reproducibility of random amplified polymorphic DNA (RAPD) analysis among laboratories. *Genome Research* 2, 341-345.

PETROLI, C.D., SANSALONI, C.P., CARLING, J., STEANE, D.A., VAILLANCOURT, R.E., MYBURG, A.A., DA SILVA, O.B.J., PAPPAS, G.J.J., KILIAN, A. and GRATTAPAGLIA, D. (2012) Genomic characterization of DArT markers based on high-density linkage analysis and physical mapping to the *Eucalyptus* genome. *PLoS One* 7, e44684.

PIEPHO, H.-P. and KOCH, G. (2000) Codominant analysis of banding data from a dominant marker system by normal mixtures. *Genetics* 155, 1459-1468.

POLAND, J.A., BROWN, P.J., SORRELLS, M.E. and JANNINK, J –L. (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. *PLoS One* 7, e32253.

POMPANON, F., BONIN, A., BELLEMAIN, E. and TABERLET, P. (2005) Genotyping errors: causes, consequences and solutions. *Nature Review Genetics* 6, 847-846.

POWELL, W., MACHRAY, G., and PROVAN, J. (1996) Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* 1, 215-222.

PROVAN, J., POWELL, W. and HOLLINGSWORTH, P.M. (2001) Chloroplast microsatellites: new tools for studies in plant ecology and evolution. *Trends in Ecology Evolution* 16, 142-147.

PURSEGLOVE, J.W. (1972) Tropical crops. In: *Monocotyledons*, pp. 607. Longman, London.

QUARRIE, S.A., LAZIĆ-JANČIĆ, V., KOVAČEVIĆ, D., STEED, A. and PEKIC, S. (1999) Bulk segregant analysis with molecular markers and its use for improving drought resistance in maize. *Journal of Experimental Botany* 50, 1299-1306.

RAFALSKI, A. (2002) Application of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* 5, 94-100.

RAFALSKI, J.A. and TINGEY, S.V. (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. *Trends in Genetics* 9, 275-280.

RAJANAIDU, N., KUSHAIRI, A., RAFIL, M., MOHD DIN, A., MAIZURA, I. and JALANI, B.S. (2000) Oil palm breeding and genetic resources. In: *Advances in Oil Palm Research*, Volume 1, pp. 171-237 (Eds. Y. Basiron, B.S. Jalani and K.W. Chan) Malaysian Palm Oil Board.

RAJANAIDU, N., SAPURAH, R., RAO, V., ZAKRI, A.H. and EMBI, M.N. (1993) Isoenzyme variation in Nigerian oil palm (*Elaeis guineensis*) germplasm. In: *Proceedings on the ISOPB International Symposium on Recent Development in Oil Palm Tissue Culture and Biotechnology*, pp. 209-216. Kuala Lumpur, Malaysia.

RAJENDRAKUMAR, P., BISWAL, A.K., BALACHANDRAN, S.M., SRINIVASARAO, K. and SUNDARAM, R.M. (2007) Simple sequence repeats in organellar genomes of rice: frequency and distribution in genic and intergenic regions. *Bioinformatics* 23, 1-4.

RAMAN, H., RAMAN, R., NELSON, M.N., ASLAM, M.N., RAJASEKARAN, R., WRATTEN, N., COWLING, W.A., KILIAN, A., SHARPE, A.G. and SCHONDELMAIER, J. (2012) Diversity Array Technology Markers: Genetic diversity analyses and linkage map construction in rapeseed (*Brassica napus* L.). *DNA Research* 19, 51-65.

RANCE, K.A., MAYES, S., PRICE, Z., JACK, P.L. and CORLEY, R.H.V. (2001) Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). *Theoretical and Applied Genetics* 103, 1302-1310.

RAO, V., JALANI, B.S. and RAJANAIDU, N. (1994) Effect of dura contamination on oil extraction rate (OER). In: *Proc. Natn. Seminar "Palm oil extraction rate: problems and issues"*, pp. 58-60 (Eds. D. Arrifin and B.S. Jalani) Palm Oil Research Institute Malaysia, Kuala Lumpur.

REYNA-LOPEZ, G.E., SIMPSON, J. and RUIZ-HERRERA, J. (1997) Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi

by amplification of restriction polymorphism. *Molecular and General Genetics* 253, 703-710.

RIDOUT, C.J. and DONINI, P. (1999) Use of AFLP in cereal research. *Trends in Plant Science* 4, 76-79.

RISCH, N. (1992) Genetic linkage: Interpreting LOD scores. *Science* 255, 803-804.

RISTERUCCI, A-M., HIPPOLYTE, I., PERRIER, X., XIA, L., CAIG, V., EVERS, M., HUTTNER, E., KILIAN, A. and GLASZMANN, J-C. (2009) Development and assessment of Diversity Arrays Technology for high-throughput DNA analyses in Musa. *Theoretical and Applied Genetics* 119, 1093-1103.

RIVAL, A. (2007) Oil Palm. In: *Transgenic Crops VI*, *Biotechnology in Agriculture and Forestry,* Volume 61, pp. 59-80 (Eds. E.C. Pua and M.R. Davey) Springer-Verlag Berlin Heidelberg, New York.

ROBINSON, J.P. and HARRIS, S.A. (1999) Amplified fragment length polymorphisms and microsatellites: A phylogenetic perspective. In: *Which DNA Marker for Which Purpose? Final Compendium of the Research Project Developmnent, Optimisation and Validation of Molecular Tools for Assessment of Biodiveristy in Forest Trees* (Eds. E.M. Gillet) European Union DGXII Biotechnology FW IV Research Programme Molecular Tools for Biodiversity.

ROSENQUIST, E.A. (1986) The genetic base of oil palm breeding populations. In: *Proceedings of the International Workshop on Oil Palm Germplasm and Utilization*, pp. 16-27 (Eds. A.C. Soh, N. Rajanaidu and M. Nasir) Palm Oil Research Institute Malaysia, Kuala Lumpur.

ROSENQUIST, E.A. (1990) An overview of breeding technology and selection in *Elaeis guineensis*. In: *Proceedings of the 1989 International Palm Oil Development Conference – Agriculture*, pp. 5-26. Palm Oil Research Institute Malaysia, Kuala Lumpur,

ROTHBERG, J.M. and LEAMON, J.H. (2008) The development and impact of 454 sequencing. *Nature* 26, 1117-1124.

RUSSELL, J.R., FULLER, J.D., MACAULAY, M., HATZ, B.G., JAHOOR, A., POWELL, W. and WAUGH, R. (1997) Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *Theoretical and Applied Genetics* 95, 714-722.

SAMBANTHAMURTHI, R., RAJINDER, S., PARVEEZ, G.K.A., ONG, M. and KUSHAIRI, A. (2009) Opportunities for the oil palm via breeding and biotechnology. In: *Breeding Plantations Tree Crops: Tropical Species*, pp. 377-421 (Eds. S.M. Jain and P.M. Priyadarshan) Springer Science.

SANGER, F. and COULSON, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology* 94, 441-448.

SANGER, F., NICKLENS, S. and COULSON, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* 74, 5463-5467.

SANSALONI, C., PETROLI, C., JACCOUD, D., CARLING, J., DETERING, F., GRATTAPAGLIA, D and KILIAN, A. (2011) Diversity Arrays Technology (DArT) and next generation sequencing combined: genome-wide high throughput, highly informative genotyping for molecular breeding of *Eucalyptus*. *BMC Proceedings* 5, P54.

SARFATTI, M., KATAN, J., FLUHR, R. and ZAMIR, D. (1989) An RFLP marker in tomato linked to the *Fusarium oxysporum* resistance gene 12. *Theoretical and Applied Genetics* 78, 755-759.

SAVELKOUL, P.H.M., AARTS, H.J.M., DE HAAS, J., DIJKSHOORN, L., DUIM, B., OTSEN, M., RADEMAKER, J.L.W., SCHOULS, L. and LENSTRA, J.A. (1999) Amplified-fragment length polymorphism analysis: the state of art. *Journal of Clinical Microbiology* 37, 3083-3091.

SCHIEX, T. and GASPIN, C. (1997) CarthaGene: constructing and joining maximum likehood genetic maps. In: P*roceedings of the International Conference on Intelligent Systems for Molecular Biology*, Abstract 5, AAAI Press, pp. 258-267.

SCHMUTZ, J., CANNON, S., SCHLUETER, J., MA, J., MITROS, T., NELSON, W., HYTEN, D., SONG, Q., THELEN, J., CHENG, J., XU, D., HELLSTEN, U., MAY, G., YU, Y., SAKURAI, T., UMEZAWA, T., BHATTACHARYYA, M., 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., SHINOZAKI, K., NGUYEN, H., WING, R., CREGAN, P., SPECHT, J., GRIMWOOD, J., ROKHSAR, D., STACEY, G., SHOEMAKER, R., and JACKSON, S. (2010). Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178-183.

SCHNABLE, P., WARE, D., FULTON, R.S., STEIN, J.C., WEI, F., PASTERNAK, S., LIANG, C., ZHANG, J., FULTON, L., GRAVES, T.A., MINX, P., REILEY, 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., WARE, D., 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., LOPOZ, G., CAMPOS, D., BRAIDOTTI, M., ASHLEY, E., GELSER, 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., YING, K., YEH, C-T., EMRICH, S.J., JIA, Y., KALYANARAMAN, A., HSIA, A-P., BARBAZUK, W.B., BAUCOM, R.S., BRUTNELL, T.P., CARNITA, N.C., CHAPARRO, C., CHIA, J-M., DERAGON, J-M., ESTILL, J.C., FU, Y., JEDDELOH, J.A., HAN, Y., LEE, H., LI, P., LISCH, D.R., LIU, S., NAGEL, D.H., MCCANN, M.C., MIGUEL, P.S., MYERS, A.M., NETTLETON, D., NGUYEN, J., PENNING, B.W., PONNALA, L., SCHNEIDER, K.L., SCHWARTZ, D.C., SHARMA, A., SODERLUND, C., SPRINGER, N.M., SUN, Q., WANG, H., WATERMAN, M., WESTERMAN, R., WOLFGRUBER, T.K., YANG, L., YU, Y., ZHANG, L., ZHOU, S., ZHU, Q., BENNETZEN, J.L., DAWE, R.K., JIANG, J., JIANG, N., PRESTING, G.G., WESSLER, S.R., ALURU, S., MARTIENSSEN, R.A., CLIFTON, S.W., MCCOMBIE, W.R., WING, R.A., WILSON, R.K. (2009) The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326, 1112-1115.

SCHNEIDER, K. (2005) Mapping populations and principles of genetic mapping. In: *The Handbook of Plant Genome Mapping. Genetic and Physical Mapping,* pp. 3-21 (Eds. K. Meksem, G. Kahl) WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany.

SCHOUTEN, H.J., VAN DE WEG, W.E., CARLING, J., KHAN, S.A., MCKAY, S.J., VAN KAAUWEN, M.P.W., WITTENBERG, A.H.J., VAN PUTTEN, H.J.J.K., NOORDIJK, Y., GAO, Z., REES, D.J.G., VAN DYK, M.M., JACCOUD, D., CONSIDINE, M.J. and KILIAN, A. (2012) Diversity arrays technology (DArT) markers in apple for genetic linkage maps. *Molecular Breeding* 29, 645-660.

SCHUELKE, M. (2000) An economic method for the fluorescent labelling of PCR fragments. *Nature Biotechnology* 18, 233-234.

SELKOE, K.A. and TOONEN, R.J. (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecology Letters* 9, 615-629.

SEMAGN, K., BJØRNSTAD, Å. and NDJIONDJOP, M.N. (2006a) An overview of molecular marker methods for plants. *African Journal of Biotechnology* 5, 2540-2568.

SEMAGN, K., BJØRNSTAD, Å. and NDJIONDJOP, M.N. (2006b) Principles, requirements and prospects of genetic mapping in plants. *African Journal of Biotechnology* 5, 2569-2587.

SEMAGN, K., BJØRNSTAD, Å. and NDJIONDJOP, M.N. (2006c) Progress and prospects of marker assisted backcrossing as a tool in crop breeding programs. *African Journal of Biotechnology* 5, 2588-2603.

SEMAGN, K., BJØRNSTAD, Å., SKINNES, H., MARØY, A.G., TARKEGNE, Y. and WILLIAM, M. (2006d). Distribution of DArT, AFLP, and SSR markers in a genetic linkage map of a double-haploid hexaploid wheat population. *Genome* 49, 545-555.

SEMAGN, K., BJØRNSTAD, A. and XU, Y. (2010) The genetic dissection of quantitative traits in crops. *Electronic Journal of Biotechnology* 13 (5).

SENG, T.-Y. MOHAMED SAAD, S.H., CHIN, C.-W., TING, N.-C., SINGH, R.S.H., ZAMAN, F.Q., TAN, S.-G. and SYED ALWEE, S.S.R. (2011) Genetic linkage map of a high yielding FELDA Deli x Yangambi oil palm cross. *PLos One* 6, e265293.

SENG, T.Y., ZAMAN, F.Q., HO, C.L., ITHNIN, M. and RAO, V. (2007) Flanking AFLP markers for the *Virescens* trait in oil palm. *Journal of Oil Palm Research* 19, 381-392.

SHAH, F.H., RASHID, O., SIMONS, A.J. and DUNSDON, A. (1994) The utility of RAPD markers for the determination of genetic variation in oil palm (*Elaeis guineensis*). *Theoretical and Applied Genetics* 89, 713-718.

SHAPIRO, S.S. and WILK, M.B. (1965) An analysis of variance test for normality (complete samples). *Biometrika* 52, 591-611.

SHARMA, R., AGGARWAL, R.A.K., KUMAR, R., MOHAPATRA, T., and SHARMA, R.P. (2002). Construction of an RAPD linage map and localization of QTLs for oleic acid level using recombinant inbreds in mustard (*Brassica juncea)*. *Genome* 45, 467-472.

SHAROPOVA, N., MCMULLEN, M.D., SCHULTZ, L., SCHROEDER, S., SANCHEZ-VILLEDA, H., GARDINER, J., BERGSTROM, D., HOUCHINS, K., MELIA-HANCOCK, S., MUSKET, T., DURU, N., POLACCO, M., EDWARDS, K., RUFF, T., REGISTER, J.C., BROUWER, C., THOMPSON, R., VELASCO, R., CHIN, E., LEE, M., WOODMAN-CLIKEMAN, W., LONG, M.J., LISCUM, E., CONE, K., DAVIS, G. and COE, E.H.Jr (2002). Development and mapping of SSR markers for maize. *Plant Molecular Biology* 48, 463-481.

SHEN, X., ZHOU, M., LU, W. and OHM, H. (2003) Detection of fusarium head blight resistance QTL in a wheat population using bulked segregant analysis. *Theoretical and Applied Genetics* 106, 1041-1047.

SHENDURE, J. and JI, H. (2008) Next-generation DNA sequencing. *Nature Biotechnology* 26, 1135-1145.

SHIAO, Y.H., CRAWFORD, E.B., ANDERSON, L.M., PATEL, P. and KO, K. (2005) Allele-specific germ cell epimutation in the spacer promoter of the 45S ribosomal RNA gene after Cr(III) exposure. *Toxicology and Applied Pharmacology* 205, 290-296.

SHOBA, D., MANIVANNAN, N., VINDHIVAVARMAN, P. and NIGAM, S.N. (2012) SSR markers associated for late leaf spot disease resistance by bulked segregant analysis in groundnut (*Arachis hypogaea* L.). *Euphytica* 188, 265-272.

SHOEMAKER, R.C., GRANT, D., OLSON, T., WARREN, W.C., WING, R., YU, Y., KIM, H., CREGAN, P., JOSEPH, B., FUTRELL-GRIGGS, M., NELSON, W., DAVITO, J., WALKER, J., WALLIS, J., KREMITSKI, C., SCHEER, D., CLIFTON, S.W., GRAVES, T., NGUYEN, H., WU, X., LUO, M., DVORAK, J., NELSON, R., CANNON, S., TOMKINS, J., SCHMUTZ, J., STACEY, G. and JACKSON, S. (2008) Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. *Genome* 51, 294-302.

SHULTZ, J.L., KAZI, S., BASHIR, R., AFZAL, J.A. and LIGHTFOOT, D.A. (2007) The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. *Theoretical and Applied Genetics* 114, 1081-1090.

SIMEN, B.B., SIMONS, J.F., HULLSIEK, K.H., NOVAK, R.M., MACARTHUR, R.D., BAXTER, J.D., HUANG, C., LUBESKI, C., TURENCHALK, G.S., BRAVERMAN, M.S., DESANY, B., ROTHBERG, J.M., EGHOLM, M. and KOZAL, M.J. (2009) Low-abundance drug-resistant viral variants in chronically HIV-infected, antiretroviral treatment-naïve patients significantly impact treatment outcomes. *The Journal of Infectious Diseases* 199, 693-701.

SINGH, R.S.H., ABDUL RAHMAN, R., OOI, C.L., and LOW, E.T. (2010) Method for identification of a molecular marker linked to the shell gene of oil palm. International application published under The Patent Cooperation Treaty (PCT). World Intellectual Property Organization. International Publication Number: WO 2010/056107 A2

SINGH, R. and CHEAH, S.C. (2004) Genetic linkage mapping in oil palm. In: *Plant and Animal Genomes XII Conference,* poster 745.

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., ABDUL MANAF, M.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. and MARTIENSSEN, R.A. (2013b) The oil palm *SHELL* gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature* 500, 340-344.

SINGH, R., NAGAPPAN, J., TAN, S.G., PANANDAM, J.M. and CHEAH, S.C. (2007) Development of simple sequence repeat (SSR) markers for oil palm and their application in genetic mapping and fingerprinting of tissue culture clones. *Asian Pacific Journal of Molecular Biology and Biotechnology* 15, 121-131.

SINGH, R., NOORHARIZA, M.Z., TING, N.-C., ROZANNA, R., TAN, S.–G., LOW, E.-T., ITHNIN, M. and CHEAH, S.–G. (2008a) Exploiting an oil palm EST database for the development of gene-derived SSR markers and their exploitation for assessment of genetic diversity. *Biologia* 63, 227-235.

SINGH, R., ONG-ABDULLAH, M., LOW, E-T. L., ABDUL MANAF, M.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. and SAMBANTHAMURTHI, R. (2013a) Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. *Nature* 500, 335-339.

SINGH, R., TAN, S.G., PANANDAM, J., RAHMAN, R.A. and CHEAH, S.C. (2008b) Identification of cDNA-RFLP markers and their use for molecular mapping in oil palm (*Elaeis guineensis). Asia Pacific Journal of Molecular Biology and Biotechnology* 16, 53-63.

SINGH, R., TAN, S.G., PANANDAM, J.M., RAHMAN, R.A., OOI, L.C.L., LOW, E-T.L., SHARMA, M., JANSEN, J. and CHEAH, S-C. (2009) Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. *BMC Plant Biology* 9, 114.

SOH, A.C. (1999) Breeding plants and selection methods in oil palm. In: *Proceedings of the Symposium on the Science of Oil Palm Breeding*, pp.65-95 (Eds. N. Rajanaidu and B.S. Jalani). Palm Oil Research Institute, Kuala Lumpur, Malaysia.

SOH, A.C. and HOR, T.Y. (2000) Combining ability correlations for bunch yield and its component components in outcrossed populations in oil palm. In: *Proceedings of the International Symposium on Oil Palm Genetic Resources and Utilization,* pp. M1-M14 (Eds. N. Rajanaidu and D. Ariffin) Malaysian Palm Oil Board, Kuala Lumpur.

SOH, A.C., KEE, K.K. and GOH, K.J. (2006) Research and innovation towards sustainable palm oil production. *Journal of Science and Technology in the Tropics* 2, 77-95.

SOH, A.C., LEE, C.H., YONG, Y.Y., CHIN, C.W., TAN, Y.P., RAJANAIDU, N. and PHUAH, P.K. (1990) The precision of oil palm breeding trials in Malaysia. In: *International Symposium Application of statistics to perennial tree crops research*, pp. 41-50 (Eds. A. C. Soh, N. Rajanaidu and M.N.H. Basri) Palm Oil Research Institute Malaysia, Kuala Lumpur.

SOH, A.C., WONG, C.K., HO, Y.W. and CHOONG, C.W. (2009) Oil Palm. In: *Oil Crops*, *Handbook of Plant Breeding 4,* Chapter 11, pp. 333-368 (Eds. J. Vollmann and I. Rajcan) Springer Science.

SOH, A.C., WONG, G., HOR, T.Y., TAN, C.C. and CHEW, P.S. (2003) Oil palm genetic improvement. In: *Plant Breeding Reviews*, pp. 165-219 (Eds. J. Janick) John Wiley & Sons, Inc.

SOH, A.C., WONG, G., TAN, C.C., CHEW, P.S., HOR, T.Y., CHONG, S.P. and GOPAL, K. (2001) Recent advances towards commercial production of elite oil palm clones. In: *Proceedings 2001 PIPOC International Palm Oil Congress*, pp. 33-44. Malaysian Palm Oil Board, Kuala Lumpur.

SONG, Q.J., SHI, J.R., SINGH, S., FICKUS, E.W., COSTA, J.M., LEWIS, J. (2005). Development and mapping of microsatellite (SSR) markers in wheat. *Theoretical and Applied Genetics* 110, 550-560.

SORANZO, N., PROVAN, J. and POWELL, W. (1999) An example of microsatellite length variation in the mitochondrial genome of conifers. *Genome* 42, 158-161.

SOURDILLE, P., SINGH, S., CADALEN, T., BROWN-GUERDIRA, G.L., GAY, G., QI, L., GILL, B.S., DUFOUR, P., MURIQNEUX, A and BERNARD, M. (2004). Microsatellite-based deletion bin system for the establishment of genetic-physical map relationships in wheat (*Triticum aestivum* L.). *Functional and Integrative Genomics* 4, 12-25.

SPEARMAN, C. (1904) The proof and measurement of association between two things. *American Journal of Psychology* 15, 72-101.

SPEROTTO, R.A., BOFF, T., DUARTE, G.L. and FETT, J.P. (2008) Increased senescence-associated gene expression and lipid peroxidation induced by iron deficiency in rice roots. *Plant Cell Reports* 27, 183-195.

STAM, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. *Plant Journal* 3, 739-744.

STEANE, D.A., NICOLLE, D., SANSALONI, C.P., PETROLI, C.D., CARLING, J., KILIAN, A., MYBURG, A.A., GRATTAPAGLIA, D. and VAILLANCOURT, R.E. (2011) Population genetic analysis and phylogeny reconstruction in *Eucalyptus* (Myrtaceae) using high-throughput, genome wide genotyping. *Molecular Phylogenetics and Evolution* 59, 206-224.

STRATHDEE, C.A. and JOHNSON, W.M. (1995) Identification of epidemiology markers for *Neisseria meningitides* using difference analysis. *Gene* 166, 105-110.

STUDER, B., KÖLLIKER, R., MUYLLE, H., ASP, T., FREI, U., ROLDÁN-RUIZ, I., BARRE, P., TOMASZEWSKI, C., MEALLY, H., BARTH, S., SKØT, L., ARMSTEAD, I.P., DOLSTRA, O. and LÜBBERSTEDT, T. (2010) EST-derived SSR markers used as anchor loci for the construction of a consensus linkage map in ryegrass (*Lolium* spp.). *BMC Plant Biology* 10, 177.

STURTEVANT, A.H. (1913) The linear arrangement of six sex-linked factors in *Drosophila*, as shown by their mode of association. *Journal of Experimental Zoology* 14, 43-59.

TANKSLEY, S.D. (1993) Mapping polygenes. *Annual Reviews of Genetics* 27, 205-233.

TANKSLEY, S.D., YOUNG, N.D., PATERSON, A.H. and BONIERBALE, M. (1989) RFLP mapping in plant breeding: New tools for an old science. *Biotechnology* 7, 257-264.

TARAMINO, G., and TINGEY, S. (1996) Simple sequence repeats for germplasm analysis and mapping in maize. *Genome* 39, 277-287.

TAUTZ, D., TRICK, M. and DOVER, G. (1986) Cryptic simplicity in DNA is a major source of genetic variation. *Nature* 322, 652-656.

TING, N-C., JANSEN, J., NAGAPPAN, J., ISHAK, Z., CHIN, C-W., TAN, S-G., CHEAH, S-C. and SINGH, R. (2013) Identification of QTLs associated with callogenesis and embryogenesis in oil palm using genetic linkage maps improved with SSR markers. *PLoS ONE* 8, e53076.

TING, N.-C., NOORHARIZA, M.Z., ROZANA, R. LOW, E.–T.L., MAIZURA, I. CHEAH, S.–C., TAN, S.–G. and SINGH, R. (2010) SSR mining in oil palm EST database: application in oila palm germplasm diversity studies. *Journal of Genetics* 89, 125-145.

TORRES, A.M., WEEDEN, N.F. and MARTIN, A. (1993) Linkage among isozyme, RFLP, and RAPD markers. *Plant Physiology* 101, 394-452.

TOYOTA, M., CANZIAN, F., USHIJIMA, T., HOSOYA, Y., KURAMOTO, T., SERIKAWA, T., IMAI, K., SUGIMURA, T. and NAGAO, M. (1996) A rat

genetic map constructed by representational difference analysis markers with suitability for large-scale typing. *The Proceedings of the National Academy of Sciences of the United States of America* 93, 3914-3919.

TRANBARGER, T.J., KLUABMONGKOL, W., SANGSRAKRU, D., MORCILLO, F., TREGEAR, J.W., TRAGOONRUNG, S. and BILLOTTE N. (2012) SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*. *BMC Plant Biology* 12, 1.

TRUONG, H.T.H., GRAHAM, E., ESCH, E., WANG, J-F. and HANSON, P. (2010) Distribution of DArT markers in a genetic linkage map of Tomato. *Korean Journal of Horticultural Science and Technology* 28, 664-671.

TYRKA, M., BEDNAREK, P.T., KILIAN, A., WEDZONY, M., HURA, T. and BAUER, E. (2011) Genetic map of triticale compiling DArT, SSR and AFLP markers. *Genome* 54, 391-401.

TYRKA, M., PEROVIC, D., WARDYŃSKA, A. and ORDON, F. (2008) A new diagnostic SSR marker for selection of the *Rym4/Rym5* locus in barley breeding. *Journal of Applied Genetics* 49, 127-134.

USHIJIMA, T., MORIMURA, K., HOSOYA, Y., OKONOGI, H., TATEMATSU, M., SUGIMURA, T. and NAGAO, M. (1997) Establishment of methylation-sensitive–representational difference analysis and isolation of hypo- and hypermethylated genomic fragments in mouse liver tumors. *The Proceedings of the National Academy of Sciences of the United States of America* 94, 2284-2289.

VALES, M.I., SCHÖN, C.C., CAPETTINI, F., CHEN, X.M., COREY, A.E., MATHER, D.E., MUNDT, C.C., RICHARDSON, K.L., SANDOVAL-ISLAS, J.S., UTZ, H.F. and HAYES, P.M. (2005) Effect of population size on the estimation of QTL: a test using resistance to barley stripe rust. *Theoretical and Applied Genetics* 111, 1260-1270.

VALSANGIACOMO, C., BAGGI, F., GAIA, V., BALMELLI, T., PEDUZZI, R. and PIFFARETTI, J.-C. (1995) Use of amplified fragment length polymorphism in molecular typing of *Legionella pneumophila* and application to epidemiological studies. *Journal of Clinical Microbiology* 33, 1716-1719.

VAN ECK, H.J., VAN DER VOORT, J.R., DRAAISTRA, J., VAN ZANDVOORT, P., VAN ENCKEVORT, E., SEGERS, B., PELEMAN, J., JACOBSEN, E., HELDER, J. and BAKKER, J. (1995) The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. *Molecular Breeding* 1, 397-410.

VAN DER VOSSEN, H.A.M. (1974) *Towards more efficient selection for oil yield in the oil palm (Elaeis guineensis* Jacq.*).* Ph. D. Thesis. University Wageningen, The Netherlands.

VAN OOIJEN, J.W. (2006) Joinmap ® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma, B. V., Wageningen, Netherlands.

VAN OOIJEN, J.W. (2009) MapQTL ® 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma, B. V., Wageningen, Netherlands.

VAN OS, H., ANDRZEJEWSKI, S., BAKKER, E., BARRENA, I., BRYAN, G.J., CAROMEL, B., GHAREEB, B., ISIDORE, E., DE JONG, W., VAN KOERT, P., LEFEBVRE, V., MILLNOURNE, D., RITTER, E., ROUPPE VAN DER VOORT, J.N.A.M., ROUSELLE-BOUGEOIS, F., VAN VLIET, J., WAUGHT, R., VISSER, R.G.F., BAKKER, J., VAN ECK, H.J. (2006). Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* 173, 1075-1087.

VARSHNEY, R.K., CHABANE, K., HENDRE, P.S., AGGARWAL, R.K. and GRANER, A. (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. *Plant Science* 173, 638-649.

VARSHNEY, R.K., GLASZMANN, J.-C., LEUNG, H. and RIBAUT, J.-M. (2010) More genomic resources for less-studied crops. *Trends in Biotechnology* 28, 452-260.

VARSHNEY, R.K., GRANAR, A. and SORRELLS, M.E. (2005) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* 23, 48-55.

VARSHNEY, R.K., NAYAK, S.N., MAY, G.D. and JACKSON, S.A. (2009) Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* 27, 522-530.

VARSHNEY, R.K., PAULO, M.J., GRANDO, S., VAN EEUWIJK, F.A., KEIZER, L.C.P., GUO, P., CECCARELLI, S., KILIAN, A., BAUM, M. and GRANER, A. (2012) Genome wide association analyses for drought tolerance related traits in Barley (*Hordeum vulgare* L.) *Field Crops Research* 126, 171-180.

VARSHNEY, R.K., SIGMUND, R., BORNER, A., KORZUN, V., STEIN, N., SORRELLS, M.E., LANGRIDGE, P. and GRANER, A. (2005b) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Science* 168, 195-202.

VIKRAM, P., SWAMY, B.P.M., DIXIT, S., STACRUZ, M.T., AHMED, H.U., SINGH, A.K. and KUMAR, A. (2011) *qDTY₁.₁*, a major QTL for rice grain yield under reproductive-stage drought stress with a consistent effect in multiple elite genetic backgrounds. *BMC Genetics* 12, 89.

VORSTER, B.J., KUNERT, K.J. and CULLIS, C.A. (2002) Use of representational difference analysis for the characterization of sequence differences between date palm varieties. *Plant Cell Reports* 21, 271-275.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., VAN DE LEE, T., HOERNES, M., FRIJTERS, A., POT, J., PELEMAN, J., KUIPER, M. and ZABEAU, M. (1995) AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Research* 23, 4407-4414.

WAHID, M.B. (2009) Sequencing the oil palm genome: The beginning. In: *Palm Oil-Balancing Ecologics with Economics, Proceedings of Agriculture, Biotechnology and Sustainability Conference* Vol. I, pp. 3. International Palm Oil Congress 2009, Malaysian Palm Oil Board, Kuala Lumpur.

WAHLE, K.W.J. and JAMES, W.P.T. (1993) Isomeric fatty acids and human health. *European Journal of Clinical Nutrition* 47, 828-839.

WANG, D.G., FAN, J.B., SIAO, C.-J., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E., SPENCER, J., KRUGLYAK, L., STEIN, L., HSIE, L. TOPALOGLO, T., HUBBELL, E., ROBINSON, E., MITTMANN, M., MORRIS, M.S., SHEN, N., KILBURN, D., RIOUX, J., NUSBAUM, C., ROZEN, S., HUDSON, T.J., LIPSHUTZ, R., CHEE, M. and LANDER, E.S. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077-1082.

WANG, M.L., BARKLEY, N.A. and JENKINS, T.M. (2009) Microsatellite markers in plants and insects. Part I: Application of Biotechnology. *Genes Genomes Genomics* 3, 54-67.

WANG, M.L., HUANG, L., BONGARD-PIERCE, D.K., BELMONTE, S., ZACHGO, E.A., MORRIS, J.W., DOLAN, M. and GOODMAN, H.M. (1997) Construction of an approximately 2 Mb contig in the region around 80 cM of *Arabidopsis thaliana* chromosome 2. *Plant Journal* 12, 711-730.

WANG, S., BASTEN, C.J. and ZENG, Z-B. (2012) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC. (http://statgen.ncsu.edu/qtlcart/WQTLCart.htm)

WEBER, D., and HELENTJARIS, T. (1989) Mapping RFLP loci in maize using B – A translocations. *Genetics* 121, 583–590.

WEGRZYN, J.L., ECKERT, A.J., CHOI, M., LEE, J.M., STANTON, B.J., SYKES, R., DAVIS, M.F., TSAI, C.–J. and NEALE, D.B. (2010) Associations genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa,* Salicaceae) secondary xylem. *New Phytologist* 188, 515-532.

WELSH, J. and MCCLELLAND, M. (1990) Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Research* 18, 7213-7218.

WENZ, H.M., ROBERTSON, J.M., MENCHEN, S., OAKS, F., DEMOREST, D.M., SCHEIBLER, D., ROSENBLUM, B.B., WIKE, C., GILBERT, D.A. and EFCAVITCH, J.W. (1998) High-precision genotyping by denaturing capillary electrophoresis. *Genome Research* 3, 69-80.

WENZL, P., CARLING, J., KUDRNA, D., JACCOUD, D., HUTTNER, E., KLEINHAFS, A. and KILIAN, A. (2004) Diversity Arrays Technology (DArT) for whole-genome profiling of barley. *Proceedings of the National Academy of Sciences of the United States of America* 101, 9915-9920.

WENZL, P., LI, H., CARLING, J., ZHOU, M., RAMAN, H., PAUL, E., HEARNDEN, P., MAIER, C., XIA, L., CAIG, V., OVESNA, J., CAKIR, M., POULSEN, D., WANG, J., RAMAN, R., SMITH, K.P., MUEHLBAUER, G.J., CHALMERS, K.J., KLEINHOFS, A., HUTTNER, E. KILIAN, A. (2006) A high-density consenses map of barley linking DArT markers to SSR, RFLP, STS loci and phenotypic traits. *BMC Genomics* 7, 206.

WIELAND, I., BOLGER, G., ASOULINE, G. and WIGLER, M. (1990) A method for difference cloning: Gene amplification following subtractive hybridization. *The Proceedings of the National Academy of Sciences of the United States of America* 87, 2720-2724.

WILLCOX, M.C., KHAIRALLAH, M.M., BERGVINSON, D., CROSSA, J., DEUTSCH, J.A., EDMEADES, G., GONALEZ-DE-LOEN, D., JIANG, C., JEWELL, D.C., MIHM, J.A., WILLIAMS, W.P. and HOISINGTON, D. (2002) Selection for resistance to southwestern corn borer using marker-assisted and conventional backcrossing. *Crop Science* 42, 1516-1528.

WILLIAMS, J., KUBELIK, A., LIVAK, K., RAFALSKI, J. and TINGEY, S. (1990) DNA polymorphism amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* 18, 6531-6535.

WINTER, P. and KAHL, G. (1995) Molecular marker technologies for plant improvement. *World journal of Microbiology and Biotechnology* 11, 438-448.

WITTENBERG, A.H.J., VAN DER LEE, T., CAYLA, C., KILIAN, A., VISSER, R.G.F. and SCHOUTEN, H.J. (2005) Validation of the high-throughput marker

technology DArT using the model plant Arabidopsis Thaliana. *Molecular Genetics and Genomics* 274, 30-39.

WONG, C.K. and BERNARDO, R. (2008) Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. *Theoretical and Applied Genetics* 116, 815-824.

WOOD, B.J. and CORLEY, R.H.V. (1991) The energy balance of oil palm cultivation. In: *Proceeding of 1991 PORIM International Palm Oil Conference*, pp.130-143. Malaysia.

WU, Y., BHAT, P.R., CLOSE, T.J. and LONARDI, S. (2008) Efficient and Accurate Construction of Genetic Linkage Maps from the Minimum Spanning Tree of a Graph. *PLoS Genetics* 4, e1000212.

XIA, L. PENG, K., YANG, S., WENZL, P., DE VICENTE, C., FREGENE, M. and KILIAN, A. (2005) DArT for high-throughput genotyping of cassava (*Manihot esculenta*) and its wild relatives. *Theoretical and Applied Genetics* 110, 1092-1098.

XIE, F., LEI, L., DU, C., LI, S., HAN, W. and REN, Z. (2009) Genomic differences between *Actinobacillus pleuropneumoniae* serotypes 1 and 3 and the diversity distribution among 15 serotypes. *FEMS Microbioogy Letters* 333, 147-155.

XIE, Y., MCNALLY, K., LI, C.Y., LEUNG, H. and ZHU, Y.Y. (2006) A high-throughput genomic tool: diversity array technology complementary for rice genotyping. *Journal of Integrative Plant Biology* 48, 1069-1076.

XIONG, L.Z., XU, C.G., MAROOF, M.A.S. and ZHANG, Q. (1999) Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive amplification polymorphism technique. *Molecular and General Genetics* 261, 439-446.

XU, S. (2003) Theoretical basis of Beavis effect. *Genetics* 165, 2259-2268.

YAMAMOTO, N., TSUGANE, T., WATANABE, M., YANO, K., MAEDA, F., KUWATA, C., TORKI, M., BAN, Y., NISHIMURA, S. and SHIBATA, D. (2005) Expressed sequence tags from the laboratory-grown miniature tomato (*Lycopersicon esculentum*) cultivar Micro-Tom and mining for single nucleotide polymorphism and insertions/deletions in tomato cultivars. *Gene* 356, 127-134.

YANG, S., PANG, W., ASH, G., HARPER, J., CARLING, J., WENZL, P., HUTTNER, E. and KILIAN, A. (2006) Low level of genetic diversity in cultivated pigeonpea compared to its wild relatives is revealed by Diversity Arrays Technology (DArT). *Theoretical and Applied Genetics* 113, 585-595.

YANN, S-G. and JUAN, I.M-B. (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research* 20, 1432-1440.

YOSHIDA, Y., USHIJIMA, T., YAMASHITA, S., IMAI, K., SUGIMURA, T. and NAGAO, M. (1999) Development of the arbitratily primed-representational difference analysis method and chromosomal mapping of isolated high throughput rat genetic markers. *The Proceedings of the National Academy of Sciences of the United States of America* 96, 610-615.

YOUNG, N.D. (1999) A cautiously optimistic vision for marker-assisted breeding. *Molecular Breeding* 6, 505-510.

YOUNG, N.D. (2000) Constructing a plant genetic linkage map with DNA markers. In: *DNA-based Markers in Plants*, pp. 31-47 (Eds. R.L. Philips and J.K. Vasil) Kluwer Academic Publishers, Dordrecht, The Netherlands.

YUSOF, B. (2007) Palm oil production through sustainable plantations. *European Journal of Lipid Science and Technology* 109, 289-295.

ZABEAU, M. and VOS, P. (1993) Selective restriction fragment amplification: a general method for DNA fingerprinting. Publication 0 534 858 A1, buletin 93/13. European Patent Office, Munich, Germany.

ZALAPA, J.E., CUEVAS, H., ZHU, H., STEFFAN, S., SENALIK, D., ZELDIN, E., MCCOWN, B., HARBUT, R. and SIMON, P. (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. *American Journal of Botany* 99, 193-208.

ZAMZURI, I., ISA, Z.A. and ROHANI, O. (2005) Field evaluation of MPOB clones. In: *Proceedings of the 1005 National Seminar on Advances in Breeding and Clonal Technologies for Super Yielding Planting Materials*, pp. 237-250. Malaysian Palm Oil Board.

ZANE, L., BARGELLONI, L. and PATARNELLO, T. (2002) Strategies for microsatellite isolation: a review. *Molecular Ecology* 11, 1-16.

ZENG, Z-B. (1993) Theoretical basis of precision mapping of quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* 90, 10972-10976.

ZENG, Z-B. (1994) Precision mapping of quantitative trait loci. *Genetics* 136, 1457-1468.

ZEVEN, A.C. (1965) The origin of the oil palm. *Journal of Nigeria Institute for Oil Palm Research* 4, 218-225.

ZHANG, N., YANG, W.-X., LIU, D.–Q. (2011). Identification and molecular tagging of Leaf Rust Resistance Gene (*Lr24*) in wheat. *Agriculture Sciences in China* 10, 1898-1905.

ZHU, C., GORE, M., BUCKLER, E.S. and YU, J. (2008) Status and prospects of association mapping in plants. *The Plant Genome* 1, 5-20.

Appendices

Appendix A: Oil palm microsatellite primer sets selected from the LINK2PALM (L2P) EU FP5 ICO-DEV for fingerprinting of *dura* and *pisifera* samples from the 768, 769 and 751 controlled crosses.

No.	Primer Name	CIRAD Name	EMBL acc.	Library	Repeat motif	Primer Sequence (5'-3')	Tm (°C)
1	OP 1	MEgCIR0874	AJ578558	GA PstI	(CA)11(GA)18	F: TCCAGTTGTCGAGTTGTAGT R: ATTATGGGGTTATGCTTTCA	52
2	OP 2	mEgCIR3809	AJ578733	GA RsaI	(GA)22	F: CCTTGCATTCCACTATT R: AGTTCTCAAGCCTCACA	52
3	OP 5	mEgCIR0369	AJ578516	GT Son	(GT)26	F: GGGTAGCAAACCTTGTATTA R: ACTTCCATTGTCTCATTATTCT	52
4	OP 6	mEgCIR2518	AJ578605	GA PstI	(GT)6(GA)32	F: GATCCCAATGGTAAAGACT R: AAGCCTCAAAAGAAGACC	52
5	OP 7	mEgCIR1753	AJ578573	GT RsaI	(GT)21	F: GCAGGGATTAAGTTTGATAT R: TTTGATGTTGCTTCTTTGAT	52
6	OP 11	mEgCIR0195	AJ578506	GA Son	(GA)21	F: CCCACCACCCCTAGCTTCTC R: ACCCCGGTCCAAATAAAATC	58
7	OP 12	mEgCIR3869	AJ578741	GA RsaI	(GA)20	F: CCAATGCAGGGGACATT R: CCACGTCTACGAAATGATAA	52
8	OP 13	mEgCIR0894	AJ578562	GA PstI	(GA)18	F: TGCTTCTTGTCCTTGATACA R: AGAACCACCGGAGTTAC	52
9	OP 18	mEgCIR3886	AJ578743	GA RsaI	(GA)5gt(GA)20	F: TTCTAGGGTCTATCAAAGTCATAAG R: AGCCACCACCACCATCTACT	52
10	OP 20	mEgCIR3519	AJ578672	GA RsaI	(GA)15(GT)8	F: CCACTGCTTCAAATTTACTAG R: GCGTCCAAAACATAAATCAC	52
11	OP 21	MEgCIR0878	AJ578559	GA PstI	(GA)22	F: CAAAGCAACAAAGCTAGTTAGTA R: CAAGCAACCTCCATTTAGAT	52
12	OP 24	mEgCIR1730	AJ578571	GT RsaI	(CT)17(GT)5	F: AATTTCAAATACAGCATAGC R: CATAGTAAGTTTTGGATGATTATTA	52

| 13 | OP 29 | mEgCIR2670 | AJ578616 | GA PstI | (GA)20 | F: AGCTCTCATGCAAGTAAC | 52 |
| | | | | | | R: TTCAACATACCGTCTGTA | |

Appendix B: Multiple sequence alignment for identical clones between Deli *dura* and *dura* analysis of the first RDA analysis.

Figure B1: Sequence alignment of clone DDPB-3, DDPB-6 and DPB-15 using ClustalW.

```
SeqA Name       Len(nt)  SeqB Name       Len(nt)  Score
=======================================================
1    DDPB-3   283      2     DDPB-6   283      91
1    DDPB-3   283      3     DPB-15   281      91
2    DDPB-6   283      3     DPB-15   281      88
=======================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPB-3     ACCGACGTCGACTATCCATGAACGGATCCATCCAGCTTGGCTCGATCGTCAGTTGTAGCA 60
DPB-15     ACCGACGTCGACTATCCATGAACGGATCCATCCAGCTTGGCTCGGCCGTCAGCTGTAGCA 60
DDPB-6     ACCGATGTCGACTATCCATGAACGGATCCATCCAGCTTGGCTCGGTCGTTAATTGTAGCA 60
           ***** *********************************** *** *   *******


DDPB-3     CCTCCTCAGCCTTGTCGATGCTCGATTGCTCGAGACTCTCTACGAATGTCCGACCCAAAG 120
DPB-15     CCTCCTCAGCCTTGTCGACGCTCAGTTGCTCGAGACTCTTTACGAGTGTCCGACCCAAAG 120
DDPB-6     CCTCCTCAGCCTTGTCGACGCTCGATTGCTCGAGACTCTCTATGAGTGTCTGGTCCAAAG 120
           *****************  ****  *************  ** **  **** *   ******


DDPB-3     CACCGTAGGTGGATGTCGCAAGCCTAGAGAGTGCATCGGCCCGAGCATTCTCCATCCTGA 180
DPB-15     CGCCATAGGCGGATGTCGCAAGCCTGGAG--TGCGTCGGCCCGAGCGTTCTCCATCCTGG 178
DDPB-6     CTCCGTAGGCAGATGTCGCAAGCCTGGAGAGTGCATCGGCCCGAGCATTCTCTATCCTGA 180
           *  ** ****   **************  ***   *** **********  *****  ******


DDPB-3     GAATGTGGGAGATCTCAAAATACTTGAGGTGCGCCACGAGACCTCTCACTTTCTGAAAAT 240
DPB-15     GAATGTGAGAGATCTCGAAATACTTGAGGTGCGCCACGAGATCCCTCACCTTCTGGAAAT 238
DDPB-6     GAATATGGGAGATCTCAAAATACTTGAGGTGTGTTACAAGGTCTCTCACTTTCTGAAAAT 240
           **** ** ******** ************** *   ** **  * ***** ***** ****


DDPB-3     ATTTCGCCATGGTCGGATCCGTTCATGGATAGTCGACATCGGT 283
DPB-15     ATTTCGCCATAGTCGGATCCGTTCATGGATAGTCGACGTCGGT 281
DDPB-6     ATTTCACCATGGTCGGATCCGTTCATGGATAGTCGACGTCGGT 283
           ***** **** ************************** *****
```

Figure B2: Sequence alignment of clone DDPB-9 and DPB-5 using ClustalW.

```
SeqA Name      Len(nt)  SeqB Name      Len(nt)  Score
=====================================================
1    DDPB-9    355      2    DPB-5     356      87
=====================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPB-9     ACCGACGTC-ACTATCCATGAACGGATCCCAGACGAGCTCCTACAACGGGACAGGCTCCA 59
DPB-5      ACCGATGTCGACTATCCATGAACGGATCCCAGACGAGCTTCTACAACGGGACAGGCTCCG 60
           ***** *** ****************************** ******************

DDPB-9     TCGGCCAGGTCACTACTCCGAGCTCCCACGACAACCGATCTTCGATCGAGCTTCTACAAT 119
DPB-5      TTAGCCAGATCATTACTCCGAGCTTCCATAACAACTGGTCTTCGATCGAGCTTGTACAAC 120
           *   ***** *** ********** ***   ***** * ************** *****

DDPB-9     AAGCGGCCTCCATTCAGCCTTCTACAAGAATCGGACTCCATCTGAACTTCCATAGTGGAT 179
DPB-5      AAGCGGCCTCCTTTCAGTCTTCTGCAAGAACCGGATTCCGTCCGAACTTCCATAGTGGAT 180
           *********** ***** ***** ****** **** *** ** ****************

DDPB-9     GGATTCTGGACAAACTTCTACAACAGACGGGCTTCAGCAGCTGAATTTCTACAACGATCG 239
DPB-5      AGATTCCGGACAAGCTTCTACAACAGATGGACTCCAGCAGCTGGATTTCTACAATGGCCG 240
            ***** ****** ************* ** ** ********* ********** **  **

DDPB-9     ACCACCTCCGATGTCTGCAGTGCCTCCCCAGCCATCAACCTTCAATAGTGTCAGTCGGGC 299
DPB-5      ATCGCCTCCGGTGTCTGCCGAACCTTCCCAGCCATCAACCTTCAATAGCATCAATCAGAC 300
           * * ****** ******* *   *** ********************  *** ** * *

DDPB-9     TTCAACAACGCCAACCAAACTTCCACAGGATCCGTTCATGGATAGTCGACGTCGGT 355
DPB-5      TTCAACAACGCCATCCAGACTTCCACAGGATCCGTTCATGGATAGTCGACGTCGGT 356
           ************* *** **************************************
```

Figure B3: Sequence alignment of clone DDPB-11 and DPB-11 using ClustalW.

```
SeqA Name        Len(nt)   SeqB Name        Len(nt)   Score
=====================================================
1    DDPB-11  289      2    DPB-11   289       90
=====================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPB-11     ACCGACGTCGACTATCCATGAACGGATCCCAGATGAATTTCGACAATGGAAGGGCTCCAC 60
DPB-11      ACCGACGTCGACTATCCATGAACGGATCCCAGACAAATTTCGACAATGAAAGAGCTCCAC 60
            ********************************  ************* *** *******

DDPB-11     CAGCTGGATCTCTACTCCGAACTTCTACCACAAGCAGTCTACTCTGAACTTCTAATGCAG 120
DPB-11      CAGTCAGATCTCTACTCCAAACTTCTACTGCAAGCAGTCTCCTCCGAACTTCTACTGCAG 120
            ***    ************* ********  ********** *** ********* *****

DDPB-11     GCGGTCTACTCCGGATCTCTACTGTAAGCGAATTCCATCCGAGCCTCTACTGTAAACAAA 180
DPB-11      GCGGTCTACTCTGGATCTCTACTGTAAGTGAATTCCATCTGAGCCTCTACTGTAAATGAA 180
            *********** **************** ********* ****************   **

DDPB-11     TTCCTTCCGAACTTCTATTACAGGTAGACTTCAGCCGAGCTTCTTCATAGTCGAATCCCA 240
DPB-11      TTCCTTTTGAACTTCTATTACAGACAAACTTCAGCTGAGCTTCTTCATAGCCAGATCCCA 240
            ******  *************** * ******** *************  *  ******

DDPB-11     ATCGAGCTCCTACAGTGGATGGATCCGTTCATGGATAGTCGACGTCGGT 289
DPB-11      GTCGAGCTCCTACAGTGGATGGATCCGTTCATGGATAGTCGACGTCGGT 289
             ***********************************************
```

Figure B4: Sequence alignment of clone PDDB-1 and PDB-5 using ClustalW.

```
SeqA Name     Len(nt)  SeqB Name     Len(nt)  Score
===================================================
1   PDDB-1   359     2    PDB-5     359      96
===================================================


CLUSTAL 2.0.12 multiple sequence alignment


PDDB-1     ACCGACGTCGACTATCCATGAACGGATCCCTGTGCCCCTGATCGGCTTTGCCGGAGACAC 60
PDB-5      ACCGACGTCGACTATCCATGAACGGATCCCTGTGCCCCTGATCGGCTTTGCCGGAGACAC 60
           ************************************************************


PDDB-1     CGTCACGATAGAGGGAGAAATTACCCTGCCCATGACGGTCGGCACCGAACCACGGCAAAG 120
PDB-5      CGTCACGACAGAGGGAGAAATTACCCTGCCCGTGACGGTCGGCATCGAACCACGGCAAAG 120
           ******** ********************** *********** ***************


PDDB-1     CACGGTCTCCCTCACTTTCGCGGTCGCCCAAGTTCCTTCGGCCTACAACGCCATACTCGG 180
PDB-5      TACGGTCTCCCTCACTTTCGCGGTCGCCCAAGTTCCTTCGGCCTACAACGCCATACTCGG 180
           ***********************************************************


PDDB-1     ACGACCCGGATTGAACGCCCTCAAGGCGATCGTCTCGACGTACCATCTCCTTGTTCGATT 240
PDB-5      GCGACCCGGATTGAACGCCCTCAAGGCGATCGTCTCGACGCACCATCTCCTTGTTCGATT 240
           **************************************** *******************


PDDB-1     CCCGACCAAAAATGGAGTCGGGGAGATGCGCGGAGATCAACAGCTCGCCCGACGATGCTT 300
PDB-5      CCCGACCAACAACGGAGTCGGGGAGATGCGCGGAGATCAACAGCTCGCTCGCCGATGCTT 300
           ********* ** ***************************************** ** ********


PDDB-1     CCAAATCTCCGCTCAAAACGACGAGACAAAGGATCCGTTCATGGATAGTCGACGTCGGT 359
PDB-5      CCAAATCTCCGCTCAAAGCGACGGGACGAAGGATCCGTTCATGGATAGTCGACGTCGGT 359
           ***************** ***** *** *******************************
```

Figure B5: Sequence alignment of clone PDDB-7 and PDB-10 using ClustalW.

```
SeqA Name      Len(nt)  SeqB Name      Len(nt)  Score
=====================================================
1    PDDB-7    420       2    PDB-10    420       97
=====================================================


CLUSTAL 2.0.12 multiple sequence alignment


PDDB-7     ACCGACGTCGACTATCCATGAACGGATCCGGTGCATTAGCGCTGGTGTGATCGCACCCAC 60
PDB-10     ACCGACGTCGACTATCCATGAACGGATCCGGTGCATTAGTGCTGGTGTGATCGCGCCCTC 60
           **************************************** ************** *** *


PDDB-7     AATGATTTGTTCAAAATTCGTCGATATAACGTCGCGGCCGTCGCACGCCATCTGTAACCC 120
PDB-10     AACGATTTGTTCGAAATTCGTCGATATAACGTCGCGGCCATCGCACGCCATCTGTAACCC 120
           ** ********* ************************** *******************


PDDB-7     ACCCACAGTCCTGGCTGGTCGGGTACCGGGCCCATCAAGTGGGTCCCGCGAGCTCGCACG 180
PDB-10     ACCCACAGTCCTGGCTGGTCGGGTACCGGACCCATCAAGTGGGTCCCGCGAGCTCGCACG 180
           ***************************** ******************************


PDDB-7     GCACTGTCGGACTCCAGACTCACTTTTTTCTAGAGAAAAAACGTTACCTACGGCAGAAGA 240
PDB-10     GCACTGTCGGACTCCAGACTCAGTTTTTTCTGGAGAAAAAACGTTACCTACGGCAGAAGA 240
           ********************** ******* ***** ***********************


PDDB-7     AAGAGATCTCCATAAAAATTATGAAAAAAAAGTCTTGAAATAAAAATTAAAGGGACGAAG 300
PDB-10     AAGAGATCTCCATAAAAATTATGAAAAAAAAGTCTTGAAATAAAAATTAAAGGGACGAAG 300
           ************************************************************


PDDB-7     ATTAAAAGGGGTGCAACACGAGGACTTCCAAGGGTGTCACCCATCCCAGTACGACTCGCA 360
PDB-10     ATTAAAAGGGGTGCAACACGAGGACTTCCAAGGGTGTCACCCATCCCAGTACGACTCGCG 360
           ***********************************************************


PDDB-7     CCCAAGCACGCTCGACCGCGGAGTTCTGATGGGATCCGTTCATGGATAGTCGACGTCGGT 420
PDB-10     CCCGAGCACGCTCGACTGCGGAGTTCTGATGGGATCCGTTCATGGATAGTCGACGTCGGT 420
           *** *********** *******************************************
```

Figure B6: Sequence alignment of clone PDDB-13, PDB-4 and PDB-12 using ClustalW.

```
SeqA Name        Len(nt)   SeqB Name        Len(nt)   Score
============================================================
1    PDDB-13   331       2    PDB-4      330       95
1    PDDB-13   331       3    PDB-12     330       96
2    PDB-4     330       3    PDB-12     330       95
============================================================


CLUSTAL 2.0.12 multiple sequence alignment


PDDB-13     ACCGACGTCGACTATCCATGAACGGATCCATCGGCGCACCATAAACCAGTCAATCAGCGA 60
PDB-12      ACCGACGTCGACTATCCATGAACGGATCCATCGGCGCACTATAAACCAGCCAATCAGCGA 60
PDB-4       ACCGACGTCGACTATCCATGAACGGATCCATCGGAGCACCATAAACCAACCAATCAGCGA 60
            *********************************** **** ********  *********


PDDB-13     GTCTCGGCTCACCGCAATGCATCGCGAGCCTGATCCACGCGCGCGGTCCAGGGCTCATGA 120
PDB-12      GTCTCGGCTCACCGCAACGCATCGCGAGCCCGATCCACGCGCGCGGTCCAGGGCTCATGA 120
PDB-4       GTCTCGGCTCACCGCAACGCATCGCGAGCCCGATCCACGCGCGCGGTCCAGGGCTCATGA 120
            ***************** *********** **************************


PDDB-13     GGAGAGAGAAGAAGGTCCCAAGGGCAACCTGGCAACTTCACATCACTCGATGGTCCGATC 180
PDB-12      GGAGAGAGAAGAAGGTCCGAAGGGCAACCTGGTAACTTCACATCACTCGATGGTCCGATC 180
PDB-4       GGAGAGAGAAGAAGGTTCGAAGGGCAACCTGGTAACTTCACATCACTCGATGGTCCGATC 180
            *************** * * ************* **************************


PDDB-13     TTCTCTGATCAAGATCCAACGCTCCATATTTTTTGTGCCATCGGACGGCCACCATCGCAG 240
PDB-12      TTCTTTGATCAAGATCCAACGCTCCATATTTTT-GCACCATCGGACGGCCACCATTGCAG 239
PDB-4       CTCTCTGACCAAGATCCAATGCTCCACATTTTT-GCGCCATCGGACGGCCACCATCGTAG 239
             *** *** ********** ****** ****** *   ***************** * **


PDDB-13     CCGGTCAAGATCCAACCGTAATCAAGGCAATTGCAAGAATCAATAAACACTAGGACAACA 300
PDB-12      CCGGTCAAGATCCAACCGTAATCAAGGCCATTGCAAGAATCAATAAACACTAGGACAACA 299
PDB-4       CCGGTCAAGATCCAACCATAATCAAGGCAATTGCAAGAATCAATAAACACTAAGACAACA 299
            ***************** ********* ************************ *******


PDDB-13     CGGGATCCGTTCATGGATAGTCGACGTCGGT 331
PDB-12      CGGGATCCGTTCATGGATAGTCGACGTCGGT 330
PDB-4       CGGGATCCGTTCATGGATAGTCGACGTCGGT 330
            ******************************
```

Figure B7: Sequence alignment of clone DDPH-8 and DPH-8 using ClustalW.

```
SeqA Name     Len(nt)  SeqB Name     Len(nt)  Score
====================================================
1    DDPH-8   337      2    DPH-8    337      89
====================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPH-8      ACCGACGTCGACTATCCATGAACAAGCTCCGTCCGGACTCCTATGGGAGCCGGATTTCAC 60
DPH-8       ACCGACGTCGACTATCCATGAACAAGCTTCGTTCGGACTCCTACGGAAGCCAGACTTCAC 60
            **************************** *** ********** ** **** ** *****


DDPH-8      CCCTGACTTTCATTGCAGGAAGACTCCACCCGGACCCCTATGGGAGCCGGGCCTCGTCTC 120
DPH-8       CCCTGACTTTAATTGCAGGAAAACTCCGCCCGGACCCCTACGGGAGCCGGGCCTCGTCCC 120
            ********** ********** ***** ************ **************** *


DDPH-8      CGACTCCGACTGCAGACGGACCTTGTCCGGACTCCTACGGGAGCCGGACTCCACCCACAA 180
DPH-8       CGACCCCGGCTGCAGACAGACCTCGTCTGGACTCCTACGCGAGCCGGACTCCGCCCACAA 180
            **** *** ******** ***** *** *********** ************ *******


DDPH-8      CTCTGACCGCAAGCAGACTCCGCCTGGACTCCTACGGGAGCTGGGCTCCGTCCCCAACTC 240
DPH-8       CTCCAACTGCAAGTAGACTCCACTTGGACTCCTACGGGAGCTGGGCTCCGCCCACAACTC 240
            ***   ** ***** ******* * ************************ ** ******


DDPH-8      CAACTGCCAGTAGACTCCGTCTGGACTCCTACAGGAGCCGGGCTCCATCACCAACTTCAA 300
DPH-8       CAACTGCAAGTAGACTCCGCCCAGACTCCTACGGGAGCCAGGCTCCGTCCCCAGCATCAA 300
            ******* ********** *   ********* ****** ****** ** *** * ****


DDPH-8      CTGCTGGTAAGCTTGTTCATGGATAGTCGACGTCGGT 337
DPH-8       CTGCTGGTAAGCTTGTTCATGGATAGTCGACGTCGGT 337
            *************************************
```

Figure B8: Sequence alignment of clone DDPH-10 and DPH-15 using ClustalW.

```
SeqA Name       Len(nt)   SeqB Name       Len(nt)   Score
============================================================
1    DDPH-10   395       2    DPH-15    394       94
============================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPH-10     ACCGACGTCGACTATCCATGAACAAGCTTCTAGGCTGACCAAGATCCTGAGGCTGAGCTT 60
DPH-15      ACCGACGTCGACTATCCATGAACAAGCTTCTAGGCTGACCAAGATCCCGAGGCTGAGCTT 60
            ************************************************ ************

DDPH-10     CTAGGTCTCAGGTTTGAAGCATCCAGGCTGACCAAGATTTTGAGGCCAAGCTCCAAGGTC 120
DPH-15      CCAGGTCTCAGGTTCGAAGCATCCGGGCTGACCAAGATTTCAAGGCCAAGCTCCAAGGTC 120
            * ************ ********* ***************  ******************

DDPH-10     TCACATCAAGGCCGACCAAGACCATGAGACCAAGCTCCTAGGTCTTGCATCCGAAGCATC 180
DPH-15      TCGCATCAAGGCCGACCAAGACCACGAGACCAAGCTCCTAGGTCTTGCATCCGAAGCATC 180
            ** ********************** ***********************************

DDPH-10     TAGGTAGACTAGGCCAAAGTCCTGAGACCGAGCTCCGATGTCTCAAATCCAAAGCATCCA 240
DPH-15      TAGGTAGACTAGGCCGAAGTCCTGAGACCGAGCTCCGATGTCTCAAATCCGAAGCATCCA 240
            *************** ********************************** *********

DDPH-10     GGCAAGATTGAGCTCCTAGGTCTTGGATTCAAAGCATCCAGGCCAAGGAAAATCTCGAGA 300
DPH-15      GGCAAGATTGAGCTCTTAGGTCTTGGATTCGAAGCATCCAGGTCAAGGAAAATCTCAAGA 300
            *************** ************** *********** ************* ***

DDPH-10     TTGAGCCAGAGGTGAGTCCTAAGATCGAGCTCTGAGGTCTTGAATCTGAAGCATCCAGGC 360
DPH-15      TTGAGC-AGAGGTGAGTCCTAAGACCGAGCTCTGAGATATCGAATCTGAAGCATCCAGGC 359
            ****** **************** *********** * *  ******************

DDPH-10     GAGACCAAGCTTGTTCATGGATAGTCGACGTCGGT 395
DPH-15      GAGACCAAGCTTGTTCATGGATACTCGACGTCTGT 394
            ********************** ******** **
```

Figure B9: Sequence alignment of clone PDDH-1 and PDH-8 using ClustalW.

```
SeqA Name     Len(nt)  SeqB Name     Len(nt)  Score
===================================================
1    PDDH-1   385      2    PDH-8    385      96
===================================================


CLUSTAL 2.0.12 multiple sequence alignment


PDDH-1      ACCGATGTCGACTATCCATGAACAAGCTTCCTCGAGGTCGGCAACATGATTTCCGGAAAA 60
PDH-8       ACCGACGTCGACTATCCATGAACAAGCTTCCTCGAGGTCGGCAACATGATTTCCGGAAGA 60
            ***** ********************************************* *


PDDH-1      TCTACTTTTCACGGGCATATCGTCCACGTACACCTCCATGTTTCGGCCGATTTGTTCTTT 120
PDH-8       TCTACTTTTCACGAGCATATCGTCCACATACACCTCCATGTTTCGGCCGATTTATTCTTT 120
            ************* ************* ************************** ******


PDDH-1      GAAGATTTTATTTACCAACCGCTGATGGGTTGCGCCTGCGTTCTTGAGGCCGAACGGCAT 180
PDH-8       GAAGATTTTATTTACCAACCGCTGATAGGTTGCGCCTGCATTCTAGAGGCCGAACGGCAT 180
            ************************** *********** **** **************


PDDH-1      AACCCTGTAGCAGTAAAGGCCGCGACTAGTGGTAAAAGCAGTCTTTTCTTCATCTTCCGA 240
PDH-8       AACCCTGTAGCAGTAAAGGCCGCGACTAGTGATAAAAGCAGTCTTTTCTTCATCTTCCGA 240
            ******************************* ****************************


PDDH-1      TGCCATCCTAATCTGGTTATAGCCGGAGAATGCGTCCATAAAAGTGAGAAGCTCGTGGCC 300
PDH-8       TGCCATCCTAATCTGGTTATAGCCGGAGAATGCATCCATAAAAGTGAGAAGCTCGTGGCC 300
            ********************************* **************************


PDDH-1      CGAGGTGGCGTCCACCAGTTGGTCGATCCTGGGTAGGGGGTAGCTGTCCTTTGGGCAAGC 360
PDH-8       CGAGGTAGCGTCCACCAGTTGGTCGATCCTGGGCAGGGGGTAGCTGTCCTTTGGGCAAGC 360
            ****** ************************** **************************


PDDH-1      TTGTTCATGGATAGTCGACGTCGGT 385
PDH-8       TTGTTCATGGATAGTCGACGTCGGT 385
            *************************
```

Figure B10: Sequence alignment of clone PDDH-3, PDDH-4, PDDH-8, PDDH-11, PDH-2, PDH-4, PDH-6, PDH-13 and PDH-14 using ClustalW.

```
SeqA Name       Len(nt)   SeqB Name       Len(nt)   Score
==========================================================
1    PDDH-3     448       2    PDDH-4     448       98
1    PDDH-3     448       3    PDDH-8     448       98
1    PDDH-3     448       4    PDDH-11    448       97
1    PDDH-3     448       5    PDH-2      448       97
1    PDDH-3     448       6    PDH-4      448       97
1    PDDH-3     448       7    PDH-6      448       98
1    PDDH-3     448       8    PDH-13     448       97
1    PDDH-3     448       9    PDH-14     448       97
2    PDDH-4     448       3    PDDH-8     448       99
2    PDDH-4     448       4    PDDH-11    448       98
2    PDDH-4     448       5    PDH-2      448       98
2    PDDH-4     448       6    PDH-4      448       99
2    PDDH-4     448       7    PDH-6      448       99
2    PDDH-4     448       8    PDH-13     448       99
2    PDDH-4     448       9    PDH-14     448       99
3    PDDH-8     448       4    PDDH-11    448       98
3    PDDH-8     448       5    PDH-2      448       98
3    PDDH-8     448       6    PDH-4      448       99
3    PDDH-8     448       7    PDH-6      448       99
3    PDDH-8     448       8    PDH-13     448       99
3    PDDH-8     448       9    PDH-14     448       99
4    PDDH-11    448       5    PDH-2      448       97
4    PDDH-11    448       6    PDH-4      448       98
4    PDDH-11    448       7    PDH-6      448       98
4    PDDH-11    448       8    PDH-13     448       97
4    PDDH-11    448       9    PDH-14     448       97
5    PDH-2      448       6    PDH-4      448       98
5    PDH-2      448       7    PDH-6      448       98
5    PDH-2      448       8    PDH-13     448       97
5    PDH-2      448       9    PDH-14     448       97
6    PDH-4      448       7    PDH-6      448       99
6    PDH-4      448       8    PDH-13     448       98
6    PDH-4      448       9    PDH-14     448       98
7    PDH-6      448       8    PDH-13     448       99
```

```
7    PDH-6    448     9    PDH-14    448     99
8    PDH-13   448     9    PDH-14    448     98
=====================================================
```

CLUSTAL 2.0.12 multiple sequence alignment


```
PDH-2      ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDH-14     ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDH-4      ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDH-6      ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDDH-8     ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDDH-4     ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDDH-11    ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACAATTGGTGCCGACCACA 60
PDH-13     ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGATTGGTGCTGACCACA 60
PDDH-3     ACCGACGTCGACTATCCATGAACAAGCTTCGTTGCCCGAACACGGTTGGTGCTGACCACA 60
           *********************************************  ******* *******


PDH-2      CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDH-14     CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDH-4      CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDH-6      CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDDH-8     CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDDH-4     CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDDH-11    CTAGGTGCTACCGTGGTAGCAAGAGAGGCCAGGCAATGACAATTGAGAGGTTGTCACTGA 120
PDH-13     CTAGGTGCTACCGTGGTAACAAGAGAGGCCAGGCAGTGACAATTGAGAGGTTGTCACTGA 120
PDDH-3     CTAGGTGCTACCGTGGTAGCAAGGGAGGCCAGGCAGTGACAATTGACAGGTTGCCACTGA 120
           ***************** **** ********** ********** ****** ******


PDH-2      GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDH-14     GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDH-4      GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDH-6      GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDDH-8     GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDDH-4     GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDDH-11    GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDH-13     GCATTCCGTCTCACACGGGAAGAGAGGTCAAATGGCAAGGCAAAAGGCCATACGCCCGTG 180
PDDH-3     GCATTCCGTCTCACACGGGAAGAGAGGTCAAGTGGCAAGGCAAAAGGCCATACGCCCGTG 180
           ******************************* **************************
```

```
PDH-2      TGGCTCCTCGCGGAGTATAGCTCACACCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
PDH-14     TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
PDH-4      TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
PDH-6      TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
PDDH-8     TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGAGGCAACGCCC 240
PDDH-4     TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
PDDH-11    TGGCTCCTCGCGGAGTATAGCTCACATCCAAACGTCTGATCGGGGAACGGGGCAACGCCC 240
PDH-13     TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
PDDH-3     TGGCTCCTCGCGGAGTATAGCTCACATCCAAACATCTGATTGGGGAACGGGGCAACGCCC 240
           ************************** ****** ****** ******** **********


PDH-2      ATGAAGTTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDH-14     ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDH-4      ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDH-6      ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDDH-8     ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDDH-4     ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDDH-11    ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDH-13     ATGAAGCTCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
PDDH-3     ATGAAGCCCCGGCGGAAAGGGAAGGCCTGCCAGGCCGTATGCCCATGGGTGCAGGATTCT 300
           ******  ***************************************************


PDH-2      TCAAAAAAGCGCGGGCTGACTCGGAGACCTGGGACCTTGGCTTAGCAACGAATGTAGGGA 360
PDH-14     TCGAAAAAGCGCGGGCTGACTCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDH-4      TCGAAAAAGCGCGGGCTGACCCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDH-6      TCGAAAAAGCGCGGGCTGACCCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDDH-8     TCGAAAAAGCGCGGGCTGACTCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDDH-4     TCGAAAAAGCGCGGGCTGACTCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDDH-11    TCGAAAAAGCGCGGGCTGACTCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDH-13     TCGAAAAAGCGCGGGCTGATTCGGAGACCTGAGACCTTGGCTTAGCAACGAATGAAGGGA 360
PDDH-3     TCGAAAAAGCGCGGGCTGACTCGGAGACCTGGGACCTTGGCTTAGCAACGAATGAAGGGA 360
           ** ***************  ********** ********************* *****
```

```
PDH-2      AGCTCGAAGAGCTTTCTCTGCCACCGGCTTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDH-14     AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTACTGGTCGGCCAACTAAAGCTCGCTA 420
PDH-4      AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDH-6      AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDDH-8     AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDDH-4     AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDDH-11    AGCTCGAAGAGCTTTCTCCGCCAGCGGCCTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDH-13     AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTAGTGGTCGGCCAACTAAAGCTCGCTA 420
PDDH-3     AGCTCGAAGAGCTTTCTCCGCCAGCGGCTTATGTAGTGGTTGGCCAACTAAAGCTCGCTA 420
           *****************  ****  ****  ******  ****  ******************


PDH-2      AGCTTGTTCATGGATAGTCGACGTCGGT 448
PDH-14     AGTTTGTTCATGGATAGTCTACGTCGGT 448
PDH-4      AGCTTGTTCATGGATAGTCGACATCGGT 448
PDH-6      AGCTTGTTCATGGATAGTCGACGTCGGT 448
PDDH-8     AGCTTGTTCATGGATAGTCGACGTCGGT 448
PDDH-4     AGCTTGTTCATGGATAGTCGACGTCGGT 448
PDDH-11    AGCTTGTTCATGGATAGTCGACGTCGGT 448
PDH-13     AGCTTGTTCATGGATAGTCGACGTCGGT 448
PDDH-3     AGCTTGTTCATGGATAGTCGACGTCGGT 448
           **  ***************  **  *****
```

Figure B11: Sequence alignment of clone PDDH-5, PDDH-6, PDDH-9 and PDH-15 using ClustalW.

```
SeqA Name       Len(nt)  SeqB Name       Len(nt)  Score
=======================================================
1    PDDH-5     323       2    PDDH-6     323       97
1    PDDH-5     323       3    PDDH-9     323       97
1    PDDH-5     323       4    PDH-15     323       97
2    PDDH-6     323       3    PDDH-9     323       98
2    PDDH-6     323       4    PDH-15     323       98
3    PDDH-9     323       4    PDH-15     323       98
=======================================================


CLUSTAL 2.0.12 multiple sequence alignment



PDDH-9      ACCGACGTCGACTATCCATGAACAAGCTTATCCCCGATCGTCTCACTGGCCGACCTTGAC 60
PDH-15      ACCGACGTCGACTATCCATGAACAAGCTTATCCCCCATCGCCTCACTGGCCGACCTTGAC 60
PDDH-6      ACCGACGTCGACTATCCATGAACAAGCTTACCCCCCATCGTCTCACTGGCCGACCTTGAC 60
PDDH-5      ACCGACGTCGACTATCCATGAACAAGCTTATCCCCCATCGTCTCACTGGCCGACCTTGAC 60
            ****************************** **** **** ******************

PDDH-9      CCCTGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGCTCTC 120
PDH-15      CCCTGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGCTCTC 120
PDDH-6      CCCTGTTATTTTGGGGTCGTATCTAGTATTCAGAGCTTGCCTCGATTTGGTACCGCTCTC 120
PDDH-5      CCCTGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGTTCTC 120
            ****************** **************** ****************** ****

PDDH-9      GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
PDH-15      GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
PDDH-6      GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
PDDH-5      GCAGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCACCTCAAC 180
            ** ************************************************ *******

PDDH-9      GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
PDH-15      GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
PDDH-6      GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
PDDH-5      GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
            ************************************************************
```

434

```
PDDH-9     TCATCCGCTGATTCTTCAACATCAGTCGGTTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
PDH-15     TCATCCGCTGATTCTTCAACATCAGTCGGTTCGGGCCTCCACTTAGTTTCACCCAAGCTT 300
PDDH-6     TCATCCGCCGATTCTTCAACATCAGTCGGTTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
PDDH-5     TCATCCGCTGATTCTTCAACATCGGTCGATTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
           ******** ************* **** ***** ***********************


PDDH-9     GTTCATGGATAGTCGACATCGGT 323
PDH-15     GTTCATGGATAGTCGACGTCGGT 323
PDDH-6     GTTCATGGATAGTCGACGTCGGT 323
PDDH-5     GTTCATGGATAGTCGACGTCGGT 323
           ***************** *****
```

Appendix C: Multiple sequence alignment for identical clones within reciprocal analysis for the first RDA analysis.

Figure C1: Sequence alignment of clone DDPB-1, PDDB-13, PDB-4 and PDB-12 using ClustalW.

```
SeqA Name       Len(nt)  SeqB Name       Len(nt)  Score
========================================================
1    DDPB-1     330      2    PDDB-13    331      97
1    DDPB-1     330      3    PDB-4      330      96
1    DDPB-1     330      4    PDB-12     330      97
2    PDDB-13    331      3    PDB-4      330      95
2    PDDB-13    331      4    PDB-12     330      96
3    PDB-4      330      4    PDB-12     330      95
========================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPB-1      ACCGACGTCGACTATCCATGAACGGATCCATCGGCGCACCATAAACCAGCCAATCAGCGA 60
PDDB-13     ACCGACGTCGACTATCCATGAACGGATCCATCGGCGCACCATAAACCAGTCAATCAGCGA 60
PDB-12      ACCGACGTCGACTATCCATGAACGGATCCATCGGCGCACTATAAACCAGCCAATCAGCGA 60
PDB-4       ACCGACGTCGACTATCCATGAACGGATCCATCGGAGCACCATAAACCAACCAATCAGCGA 60
            ********************************** **** ********  **********


DDPB-1      GCCTCGGCTCACCGCAACGCATCGCGAGCCCGATCCACGCGCGCGGTCCAGGGCTCGTGA 120
PDDB-13     GTCTCGGCTCACCGCAATGCATCGCGAGCCTGATCCACGCGCGCGGTCCAGGGCTCATGA 120
PDB-12      GTCTCGGCTCACCGCAACGCATCGCGAGCCCGATCCACGCGCGCGGTCCAGGGCTCATGA 120
PDB-4       GTCTCGGCTCACCGCAACGCATCGCGAGCCCGATCCACGCGCGCGGTCCAGGGCTCATGA 120
            * ************** *********** ************************** ***


DDPB-1      GGAGAGAGAAGAAGGTCCGAAGGGCAACCTGGTAACTTCACATCACTCGATGGTCCGATC 180
PDDB-13     GGAGAGAGAAGAAGGTCCCAAGGGCAACCTGGCAACTTCACATCACTCGATGGTCCGATC 180
PDB-12      GGAGAGAGAAGAAGGTCCGAAGGGCAACCTGGTAACTTCACATCACTCGATGGTCCGATC 180
PDB-4       GGAGAGAGAAGAAGGTTCGAAGGGCAACCTGGTAACTTCACATCACTCGATGGTCCGATC 180
            *************** * ************ **************************
```

```
DDPB-1      TTCTCTGATTAAGATCCAACGCTCCATATTTTT-GCGCCATCGGACGGCCACCATCGCAG 239
PDDB-13     TTCTCTGATCAAGATCCAACGCTCCATATTTTTTGTGCCATCGGACGGCCACCATCGCAG 240
PDB-12      TTCTTTGATCAAGATCCAACGCTCCATATTTTT-GCACCATCGGACGGCCACCATTGCAG 239
PDB-4       CTCTCTGACCAAGATCCAATGCTCCACATTTTT-GCGCCATCGGACGGCCACCATCGTAG 239
             *** ***  ********* ****** ****** *  ***************** * **


DDPB-1      CCGGTCAAGATCCAACCGTAATCAAGGCAATTGCAAGAATCAATAAACACTAGGACAACA 299
PDDB-13     CCGGTCAAGATCCAACCGTAATCAAGGCAATTGCAAGAATCAATAAACACTAGGACAACA 300
PDB-12      CCGGTCAAGATCCAACCGTAATCAAGGCCATTGCAAGAATCAATAAACACTAGGACAACA 299
PDB-4       CCGGTCAAGATCCAACCATAATCAAGGCAATTGCAAGAATCAATAAACACTAAGACAACA 299
            **************** ********** *********************** *******


DDPB-1      CGGGATCCGTTCATGGATAGTCGACGTCGGT 330
PDDB-13     CGGGATCCGTTCATGGATAGTCGACGTCGGT 331
PDB-12      CGGGATCCGTTCATGGATAGTCGACGTCGGT 330
PDB-4       CGGGATCCGTTCATGGATAGTCGACGTCGGT 330
            ******************************
```

Figure C2: Sequence alignment of clone DPB-6 and PDB-6 using ClustalW.

```
SeqA Name    Len(nt)  SeqB Name    Len(nt)  Score
==================================================
1   DPB-6   456     2    PDB-6    456      94
==================================================


CLUSTAL 2.0.12 multiple sequence alignment


DPB-6       ACCGACGTCGACTATCCATGAACGGATCCGTCGGCTATTTAAACCCCCTAATCCCCCTTC 60
PDB-6       ACCGACGTCGACTATCCATGAACGGATCCGTCGGCTACTTAAACCCCCTAATCTCCCTTC 60
            ************************************* *************** ******


DPB-6       GCGGCTTCATCCTGTTGATAGGATGGTACCCTTCTTTCGCCTGAGAGCCTAGCTACTCAG 120
PDB-6       GTGGCTTCATCCTGTCGATAAGACGGTACCTTTCTTTCGCCTGAAAGCCTAGCTACTCAG 120
            * ************* **** ** ****** ************* **************


DPB-6       CTACTTCCCTCGTACCAGTCCTTGCCTTCTTCAGCTCCCCAATTCTTCTTCGAGGGTTCC 180
PDB-6       CTACTTCCCTCGCACCAGTCCTTGCCTTCTTCAGCCCCCCAATTCTTCTCTAAGGGTTCC 180
            ************ ********************** *************   ********


DPB-6       CACACCATGTCCTCTACCCCGGAGGCCATTCTTGAAGGGATGTCTAGGGCTTGGAGGAAG 240
PDB-6       CACACCATGTCCTCTACCCCGGAGGCCATTCTTGAAGGGATGTCTAGGGCTTGGAGGGAG 240
            ********************************************************* **


DPB-6       GTGAGGAGGAGAACGGCGGCCGGTGAGGATCTCCGTCGTTTTAAAGGGGGCTCAAGGAAG 300
PDB-6       GCGAGGAGGAGAACGGCGGTCGGTGAGGATCTCCGTCGTCTTAAAAGGGGCTCAAGGAAG 300
            * ***************** *************** ***** ***** **************


DPB-6       AATTCCTTCAACAACAGGGGGTTTAATAACGGGGGCCATCGATAAAGTTATTCTGCCCGAG 360
PDB-6       AATTCCTTCAACAACAGGGGGTTTAATAACGGGGGCCACCGGTAAAGTTATTCTGCCCGAG 360
            ************************************* ** *****************


DPB-6       AATTTTGGAGTAGCAAGGCGAGGTGATACCGGTCAAGAGGGCAGAGCTGTCGTCACAAGC 420
PDB-6       AATTTCGGAGTAGCAAGGCAGGGTGATACCGGTCAAGAGGGCAGAGCTGTCGTCACAAGC 420
            ***** *************   **************************************


DPB-6       TGTGGCGGGATCCGTTCATGGATAGTCGACGTCGGT 456
PDB-6       TATGGCGGGATCCGTTCATGGATAGTCGACGTCGGT 456
            * **********************************
```

Figure C3: Sequence alignment of clone DDPB-9, DPB-5, PDDB-6 and PDDB-9 using ClustalW.

```
SeqA Name      Len(nt)  SeqB Name      Len(nt)  Score
=====================================================
1    DDPB-9    355      2    DPB-5     356      87
1    DDPB-9    355      3    PDDB-6    356      88
1    DDPB-9    355      4    PDDB-9    356      88
2    DPB-5     356      3    PDDB-6    356      88
2    DPB-5     356      4    PDDB-9    356      87
3    PDDB-6    356      4    PDDB-9    356      88
=====================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPB-9      ACCGACGTC-ACTATCCATGAACGGATCCCAGACGAGCTCCTACAACGGGACAGGCTCCA 59
PDDB-9      ACCGACGTCGACTATCCATGAACGGATCCCAAACGAGCTTCTACAACAGGACAGGCTCCA 60
PDDB-6      ACCGACGTCGACTATCCATGAACGGATCCCAGACGAGCTTCTACAACGGGATAGGCTCCA 60
DPB-5       ACCGATGTCGACTATCCATGAACGGATCCCAGACGAGCTTCTACAACGGGACAGGCTCCG 60
            ***** *** ******************** ******* ******* *** *******


DDPB-9      TCGGCCAGGTCACTACTCCGAGCTCCCACGACAACCGATCTTCGATCGAGCTTCTACAAT 119
PDDB-9      CCGGTCAGGTCACTGCTCCGAGCCTCCACAACAATCGATCTTCGATCGAGCTTCTACAAT 120
PDDB-6      TCGGCCAGGTCACTACTCCGAGCTTCCACGACAATCGATCTTCGACCGAGCTTCTACAAT 120
DPB-5       TTAGCCAGATCATTACTCCGAGCTTCCATAACAACTGGTCTTCGATCGAGCTTGTACAAC 120
             * *** *** * ********   ***   ****   * ******* ******* *****


DDPB-9      AAGCGGCCTCCATTCAGCCTTCTACAAGAATCGGACTCCATCTGAACTTCCATAGTGGAT 179
PDDB-9      AAGCGGCCTCCTTTCAGCCTTCTGTAAGAATTGGGCTCCGTTCGAACTTCCATAGCGAAT 180
PDDB-6      TAGTGGCCTCCTTTCTGTTTTTTTGCAAGAATCGAACTCCGTCCGAACTTCCACAGCGGAT 180
DPB-5       AAGCGGCCTCCTTTCAGTCTTCTGCAAGAACCGGATTCCGTCCGAACTTCCATAGTGGAT 180
            ** ******* *** *   ** *   *****  *    *** *   ********* ** * **


DDPB-9      GGATTCTGGACAAACTTCTACAACAGACGGGCTTCAGCAGCTGAATTTCTACAACGATCG 239
PDDB-9      GGATTCCGGACGAGCTTCTACAACAGACGGGCCCCAACAGTTGGATCTCTACAACGGTCG 240
PDDB-6      GGATTCCGGACGAGCTTCTACAACAGACGGGCTCCAGCAGCTGGATTTCTACAACGGCCG 240
DPB-5       AGATTCCGGACAAGCTTCTACAACAGATGGACTCCAGCAGCTGGATTTCTACAATGGCCG 240
             ***** **** * ************* ** *  ** *** ** ** ******* *  **
```

```
DDPB-9      ACCACCTCCGATGTCTGCAGTGCCTCCCCAGCCATCAACCTTCAATAGTGTCAGTCGGGC 299
PDDB-9      ATCACCTCCGATGTCTGCCGAACCTCCCCAGCCATCAACCTTCAATAGCACCAGTCGGGC 300
PDDB-6      GTCGCCTCCGGTGTCTGCCGAACCTCCCCAGCCATCAGCCTTCAATAGCGTCAGCCGTAC 300
DPB-5       ATCGCCTCCGGTGTCTGCCGAACCTTCCCAGCCATCAACCTTCAATAGCATCAATCAGAC 300
              * ****** ******* *   *** ********** **********   **  *   *


DDPB-9      TTCAACAACGCCAACCAAACTTCCACAGGATCCGTTCATGGATAGTCGACGTCGGT 355
PDDB-9      TTCAACAATGCCATCCAGACTTCCACAGGATCCGTTCATGGATAGTCGACATCGGT 356
PDDB-6      TTCCACAACACCATCCAGACTTCCACAGGATCCGTTCATGGATAGTCGACGTCGGT 356
DPB-5       TTCAACAACGCCATCCAGACTTCCACAGGATCCGTTCATGGATAGTCGACGTCGGT 356
            *** ****   *** *** ***************************** *****
```

Figure C4: Sequence alignment of clone DDPH-2, DDPH-13, PDDH-14 and PDDH-15 using ClustalW.

```
SeqA Name        Len(nt)  SeqB Name        Len(nt)  Score
==========================================================
1    DDPH-2      280       2    DDPH-13     281       84
1    DDPH-2      280       3    PDDH-14     282       92
1    DDPH-2      280       4    PDDH-15     281       91
2    DDPH-13     281       3    PDDH-14     282       86
2    DDPH-13     281       4    PDDH-15     281       86
3    PDDH-14     282       4    PDDH-15     281       92
==========================================================


CLUSTAL 2.0.12 multiple sequence alignment


DDPH-2      ACCGACGTCGACTATCCATGAACAAGCTCCGTCCGGACTCTTACGGGAGCCGGACTTCGC 60
PDDH-14     ACCGACGTCGACTATCCATGAACAAGCTTCGTCCGGACTCCTACGGGAGCCGGACTTCGC 60
PDDH-15     ACCGACGTCGACTATCCATGAACAAGCTTCGTCCGGACTCCTATGGGAGTCGGATTTCGC 60
DDPH-13     ACCGACGTCGACTATCCATGAACAAGCTTCGTCCGGACTCCTACGGGGGCCGGACTTCAC 60
            ***************************** *********** ** *** * **** *** *


DDPH-2      CCCTGACTTTAATTGCAGGAAGACTCCACCCGGACCCCTACGGGAGCTGGGCCTCATCCC 120
PDDH-14     CCCTGACTTTAATTGCAGGAAGGCTCCGCCTGGACCCCGACGGGAGCCGGGCCTCGTCCC 120
PDDH-15     CCCTGACTTTAATTGCAGGAAGACTCTGCCCCGGACCCCTACGGGAGCCGGACCTCGTCCC 120
DDPH-13     CCCTGACTGTGATTACAGGTAGGCTTCGCCCGAGCTCCTACGGGAGTCGGATCTCGTCCC 120
            ******** * *** **** ** **    ** *   * ** *******   **  *** ****
DDPH-2      C-GACTCTGGCTGCAGACAGACCTCGTCCGGACTCCTACGGGAGCCAGACTCCGCCCGTA 179
PDDH-14     CCGACTCCGGCCACAGACAGACCTCGTCCGGACTCCTACGGGAGTCGGACTCCGCCCACA 180
PDDH-15     C-GACTCCAACTGCAGACAGACCTCGTCCGGACTCCTACGGGAGCCAGACTCCACCCACA 179
DDPH-13     C-AACTCCGAATGCAGAAAGGCTTCACCCGGACTTCTACGGGAGCAGGACTCCGCCCATA 179
            *   ****        **** ** * **   ******* *********    ****** *** *


DDPH-2      ACTCCGACTGC-AAGCAGACTCCGCCCGGACTCCTACGGGAGCCAGGCTCTGTCACCAAC 238
PDDH-14     ACTCCAACTACCAAGCAGACTCCGCCCGGACTCCTACGGGAGCCGGGCTCCGTCACCAAC 240
PDDH-15     ACTCCGACTGCCAAGTAGACTCTGCCCGGACTTCTACGGGAGCCAGACTCCGTCACCAAC 239
DDPH-13     ACTCCAACTGCCAAGTAGACTCTGCCCGGACTCCTAGGAAAGCCGGGGTCCGTCACCAAC 239
            ***** *** * *** ****** ********* *** *   **** *   ** *********
```

```
DDPH-2     TTCAATCGCTGGTAAGCTTGTTCATGGATAGTCGACGTCGGT 280
PDDH-14    TTCAACTGCTGGTAAGCTTGTTCATGGATAGTCGACGTCGGT 282
PDDH-15    TTCAACTGCTGGTAAGCTTGTTCATGGATAGTCGACGTCGGT 281
DDPH-1     TTCAACTGCTGGTAAGCTTGGTCATGGATAGTCGACGTCGGT 281
           *****  ************* *******************
```

Figure C5: Sequence alignment of clone DPH-7, PDDH-5, PDDH-6, PDDH-9 and PDH-15 using ClustalW.

```
SeqA Name      Len(nt)  SeqB Name      Len(nt)  Score
======================================================
1    DPH-7     323      2    PDDH-5    323      98
1    DPH-7     323      3    PDDH-6    323      98
1    DPH-7     323      4    PDDH-9    323      99
1    DPH-7     323      5    PDH-15    323      99
2    PDDH-5    323      3    PDDH-6    323      97
2    PDDH-5    323      4    PDDH-9    323      97
2    PDDH-5    323      5    PDH-15    323      97
3    PDDH-6    323      4    PDDH-9    323      98
3    PDDH-6    323      5    PDH-15    323      98
4    PDDH-9    323      5    PDH-15    323      98
======================================================


CLUSTAL 2.0.12 multiple sequence alignment

PDDH-9     ACCGACGTCGACTATCCATGAACAAGCTTATCCCCGATCGTCTCACTGGCCGACCTTGAC 60
PDH-15     ACCGACGTCGACTATCCATGAACAAGCTTATCCCCCATCGCCTCACTGGCCGACCTTGAC 60
PDDH-6     ACCGACGTCGACTATCCATGAACAAGCTTACCCCCCATCGTCTCACTGGCCGACCTTGAC 60
DPH-7      ACCGACGTCGACTATCCATGAACAAGCTTATCCCCCATCGTCTCACTGGCCGACCTTGAC 60
PDDH-5     ACCGACGTCGACTATCCATGAACAAGCTTATCCCCCATCGTCTCACTGGCCGACCTTGAC 60
           ****************************** **** **** ******************


PDDH-9     CCCTGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGCTCTC 120
PDH-15     CCCTGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGCTCTC 120
PDDH-6     CCCTGTTATTTTGGGGTCGTATCTAGTATTCAGAGCTTGCCTCGATTTGGTACCGCTCTC 120
DPH-7      CCCCGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGCTCTC 120
PDDH-5     CCCTGTTATTTTGGGGTCATATCTAGTATTCAGAGTTTGCCTCGATTTGGTACCGTTCTC 120
           *** ************* ************** ****************** ****


PDDH-9     GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
PDH-15     GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
PDDH-6     GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
DPH-7      GCGGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCGCCTCAAC 180
PDDH-5     GCAGCCCGCACCGAAACAGTGCTTTACCCCTAGATGTCCAGTCAACTGCTGCACCTCAAC 180
           ** ************************************************* *******
```

```
PDDH-9    GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
PDH-15    GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
PDDH-6    GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
DPH-7     GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
PDDH-5    GCATTTCGGGGAGAACCAGCTAGCTCTGGGTTCGAGTGGCATTTCACCCCTAACCACAAC 240
          ************************************************************


PDDH-9    TCATCCGCTGATTCTTCAACATCAGTCGGTTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
PDH-15    TCATCCGCTGATTCTTCAACATCAGTCGGTTCGGGCCTCCACTTAGTTTCACCCAAGCTT 300
PDDH-6    TCATCCGCCGATTCTTCAACATCAGTCGGTTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
DPH-7     TCATCCGCTGATTCTTCAACATCAGTCGGTTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
PDDH-5    TCATCCGCTGATTCTTCAACATCGGTCGATTCGGACCTCCACTTAGTTTCACCCAAGCTT 300
          ******** ************** **** ***** ***********************


PDDH-9    GTTCATGGATAGTCGACATCGGT 323
PDH-15    GTTCATGGATAGTCGACGTCGGT 323
PDDH-6    GTTCATGGATAGTCGACGTCGGT 323
DPH-7     GTTCATGGATAGTCGACGTCGGT 323
PDDH-5    GTTCATGGATAGTCGACGTCGGT 323
          ***************** *****
```

444

Figure C6: Sequence alignment of clone DPH-11, DPH-12, PDDH-1 and PDDH-8 using ClustalW.

```
SeqA Name      Len(nt)  SeqB Name      Len(nt)   Score
======================================================
1    DPH-11    385       2    DPH-12    385       94
1    DPH-11    385       3    PDDH-1    385       94
1    DPH-11    385       4    PDH-8     385       95
2    DPH-12    385       3    PDDH-1    385       96
2    DPH-12    385       4    PDH-8     385       96
3    PDDH-1    385       4    PDH-8     385       96
======================================================


CLUSTAL 2.0.12 multiple sequence alignment


DPH-12    ACCGACGTCGACTATCCATGAACAAGCTTCCTCGAGGTCGGCAACATGATTTCCGAAAGA 60
PDDH-1    ACCGATGTCGACTATCCATGAACAAGCTTCCTCGAGGTCGGCAACATGATTTCCGGAAAA 60
PDH-8     ACCGACGTCGACTATCCATGAACAAGCTTCCTCGAGGTCGGCAACATGATTTCCGGAAGA 60
DPH-11    ACCGACGTCGACTATCCATGAACAAGCTTCTTCGAGGTCGGCAACATGATTTTCGGAAGA 60
          *****  ********************** ****************** ** ** *


DPH-12    TCTACTTTTCACGAGCATATCGTCCACGTACACCTCCATGTTTCGGCCGATTTGTTCTTT 120
PDDH-1    TCTACTTTTCACGGGCATATCGTCCACGTACACCTCCATGTTTCGGCCGATTTGTTCTTT 120
PDH-8     TCTACTTTTCACGAGCATATCGTCCACATACACCTCCATGTTTCGGCCGATTTATTCTTT 120
DPH-11    TCTACTTTTCATGAGCATATCGTCCACGTACACCTCCATGTTTCGGCCGATTTGTTCTTT 120
          *********** *  ************* ************************* ******


DPH-12    GAAGATTTTGTTTACCAACCGCTGATAGGTTGCGCCTGCGTCCTTGAGGCCGGACGGCAT 180
PDDH-1    GAAGATTTTATTTACCAACCGCTGATGGGTTGCGCCTGCGTTCTTGAGGCCGAACGGCAT 180
PDH-8     GAAGATTTTATTTACCAACCGCTGATAGGTTGCGCCTGCATTCTAGAGGCCGAACGGCAT 180
DPH-11    GAAGATTTTATTTACCAACCGCTAGTAGGTTGCGCCTGCGTTCTTGAGGCCAAACGGCAT 180
          ********* *************   * *********** * ** ******  *******


DPH-12    AACCCTGTAGCGGTAAAGGCCGCGACTAGTGATAAAAGCAGTCTTTTCTTCATCTTCCGA 240
PDDH-1    AACCCTGTAGCAGTAAAGGCCGCGACTAGTGGTAAAAGCAGTCTTTTCTTCATCTTCCGA 240
PDH-8     AACCCTGTAGCAGTAAAGGCCGCGACTAGTGATAAAAGCAGTCTTTTCTTCATCTTCCGA 240
DPH-11    AACCCTGTGGCAGTAGAGGCTGCGACTAGTGATAAAAGCAGTCTTTTCTTCATCTTCCGA 240
          ******** ** *** **** *********  ****************************
```

445

```
DPH-12      TGCCATCCTAACCTGGTTATAGCCGGAGAATGTGTCCATAAAAGTGAGAAGCTCGTGGCC 300
PDDH-1      TGCCATCCTAATCTGGTTATAGCCGGAGAATGCGTCCATAAAAGTGAGAAGCTCGTGGCC 300
PDH-8       TGCCATCCTAATCTGGTTATAGCCGGAGAATGCATCCATAAAAGTGAGAAGCTCGTGGCC 300
DPH-11      TGCCATCCTAATCTGGTTATAGCCGGAGAATGCATCCATGAAAGTGAGAGGCTCGTGGCC 300
            *********** ******************* ***** ********* **********


DPH-12      CGAGGTAGCGTCCACCAGTTCGTCGATCCTGGGCAGGGGGTAGCTGTCCTTTGGGCAAGC 360
PDDH-1      CGAGGTGGCGTCCACCAGTTGGTCGATCCTGGGTAGGGGGTAGCTGTCCTTTGGGCAAGC 360
PDH-8       CGAGGTAGCGTCCACCAGTTGGTCGATCCTGGGCAGGGGGTAGCTGTCCTTTGGGCAAGC 360
DPH-11      CGAAGTAGAGTCCACCAGTTGGTCGATCCTGGGCAGGGGGTAGCTGTCCTTTGGGCAAGC 360
            *** ** * ********** ************ ************************


DPH-12      TTGTTCATGGATAGTCGACGTCGGT 385
PDDH-1      TTGTTCATGGATAGTCGACGTCGGT 385
PDH-8       TTGTTCATGGATAGTCGACGTCGGT 385
DPH-11      TTGTTCATGGATAGTCGACGTCGGT 385
            *************************
```

Appendix D: Electrophoresis profiles of AFLP analysis.
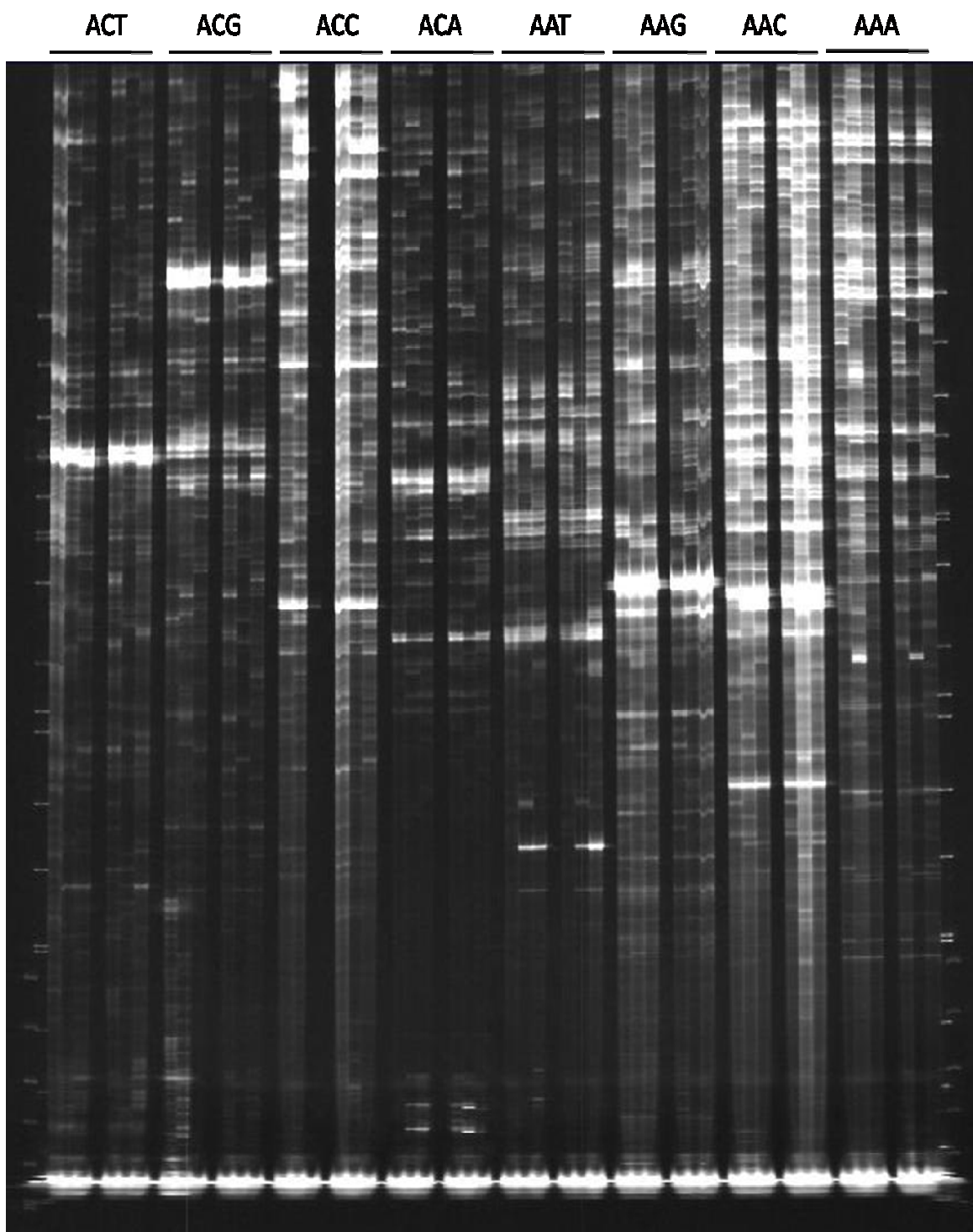
Figure D1: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Bam*HI restriction enzyme with 3'-end of the primers having selective nucleotides from AAA to ACT.
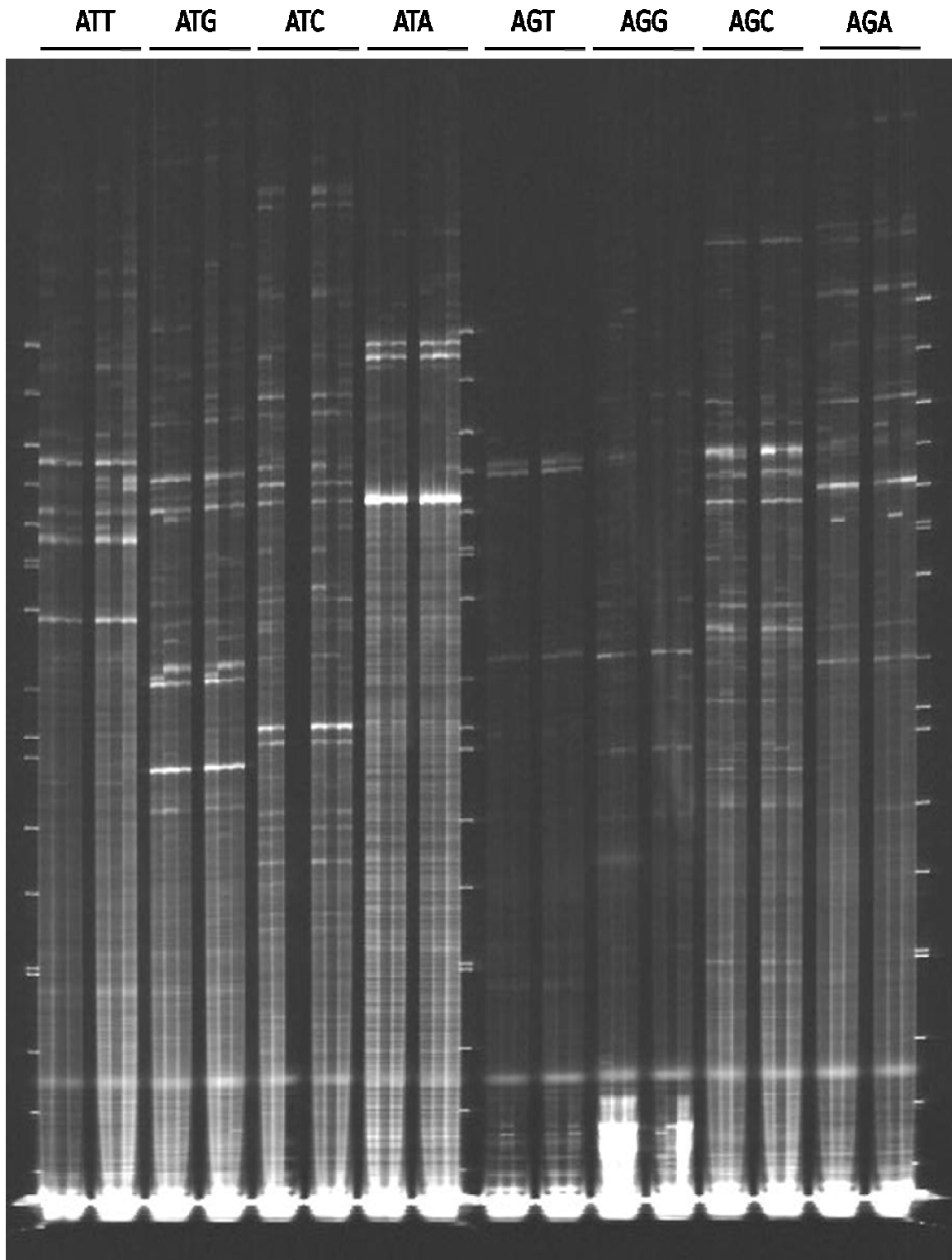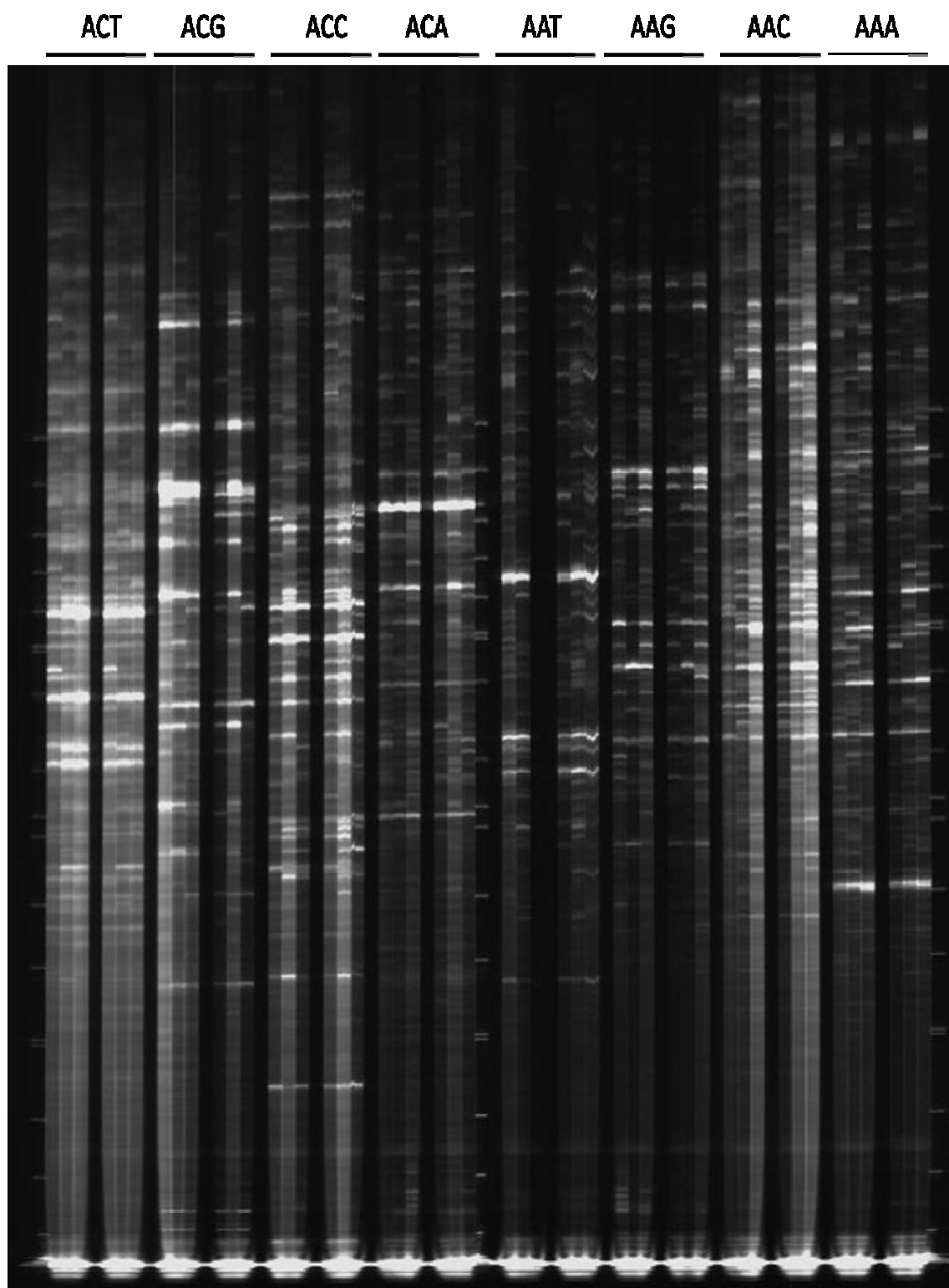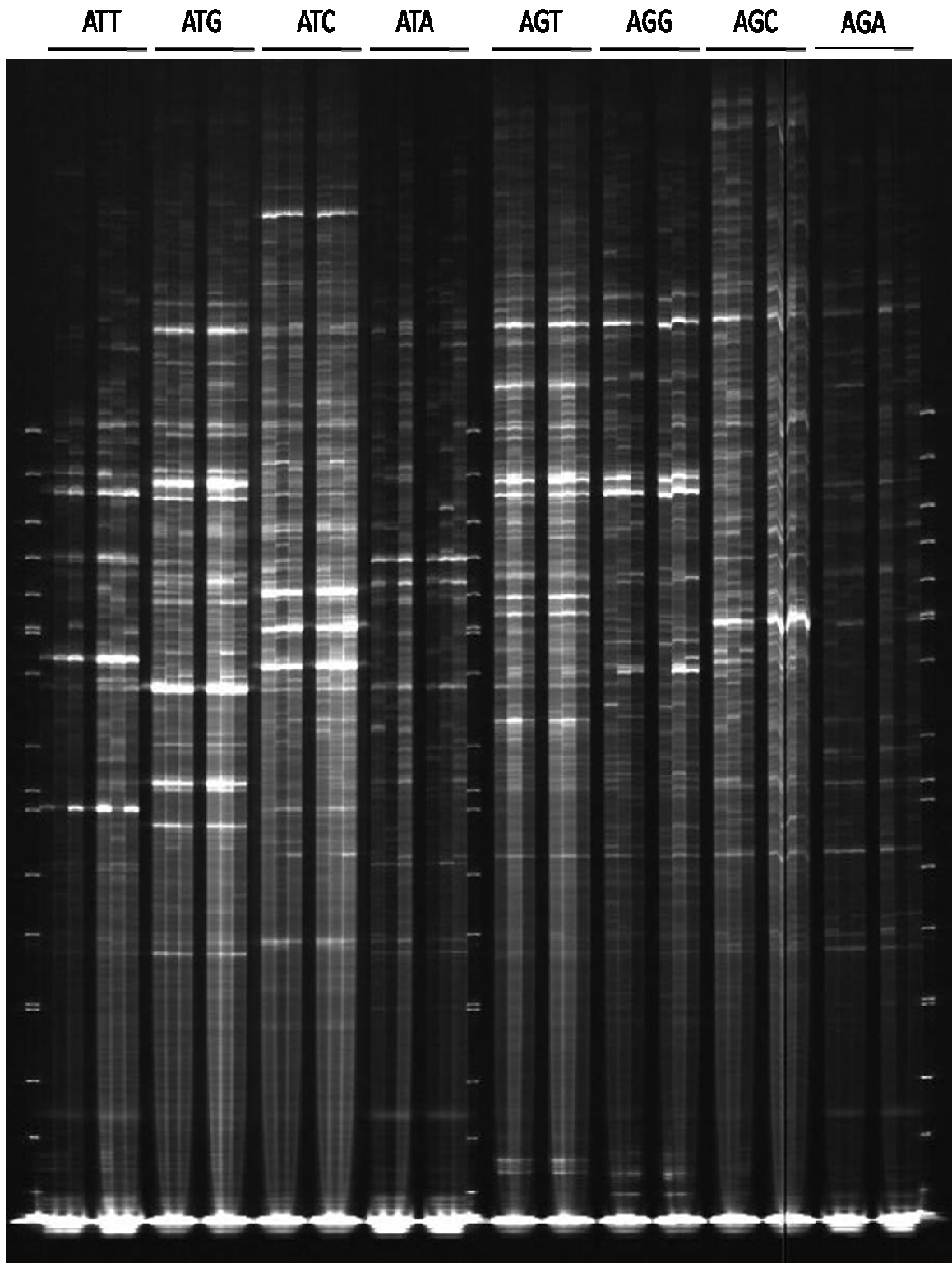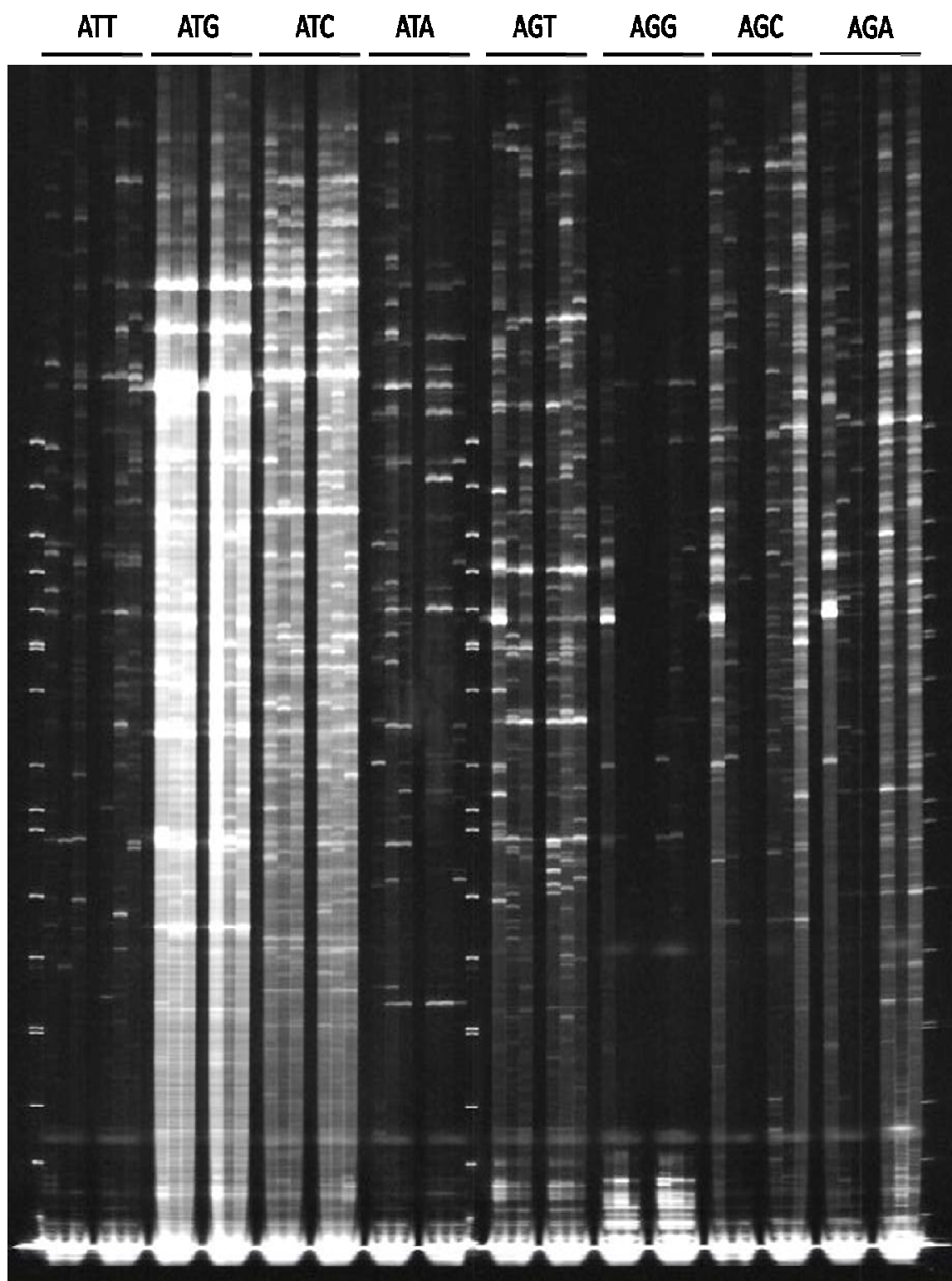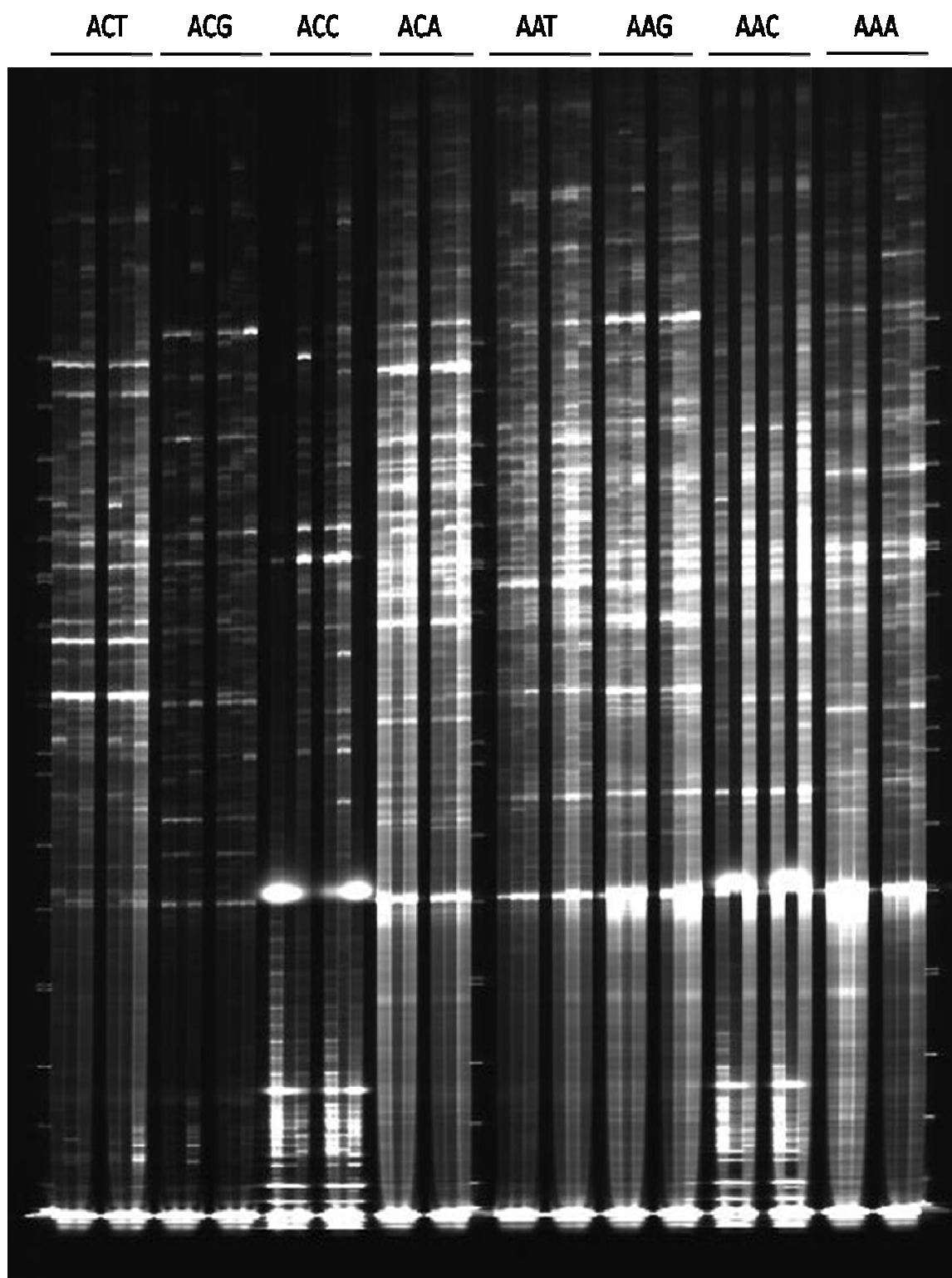
Figure D2: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Bam*HI restriction enzyme with 3'-end of the primers having selective nucleotides from AGA to ATT.

Figure D3: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Hin*dIII restriction enzyme with 3'-end of the primers having selective nucleotides from AAA to ACT.
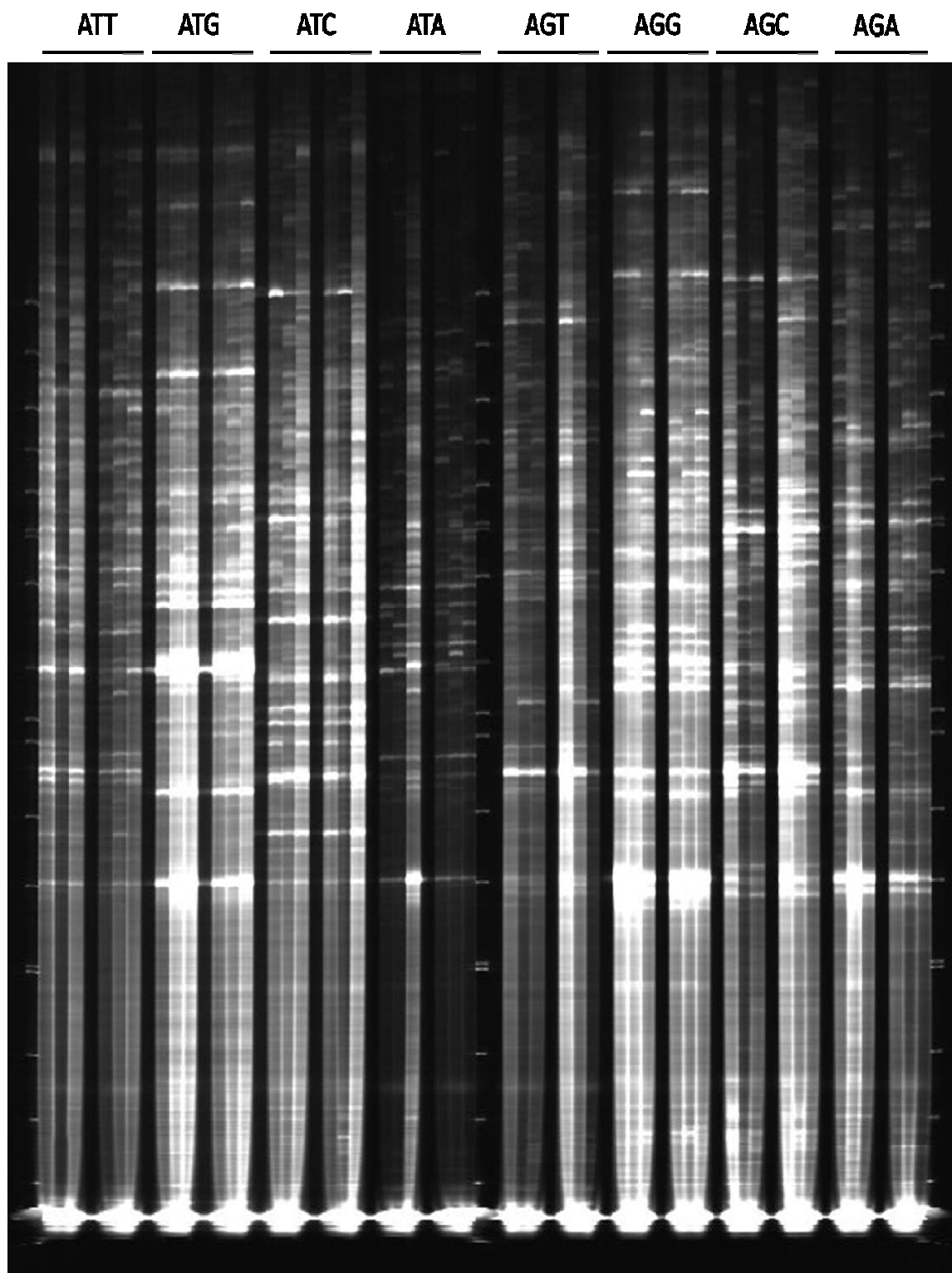
Figure D4: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Hin*dIII restriction enzyme with 3'-end of the primers having selective nucleotides from AGA to ATT.

Figure D5: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Eco*RI restriction enzyme with 3'-end of the primers having selective nucleotides from AGA to ATT.
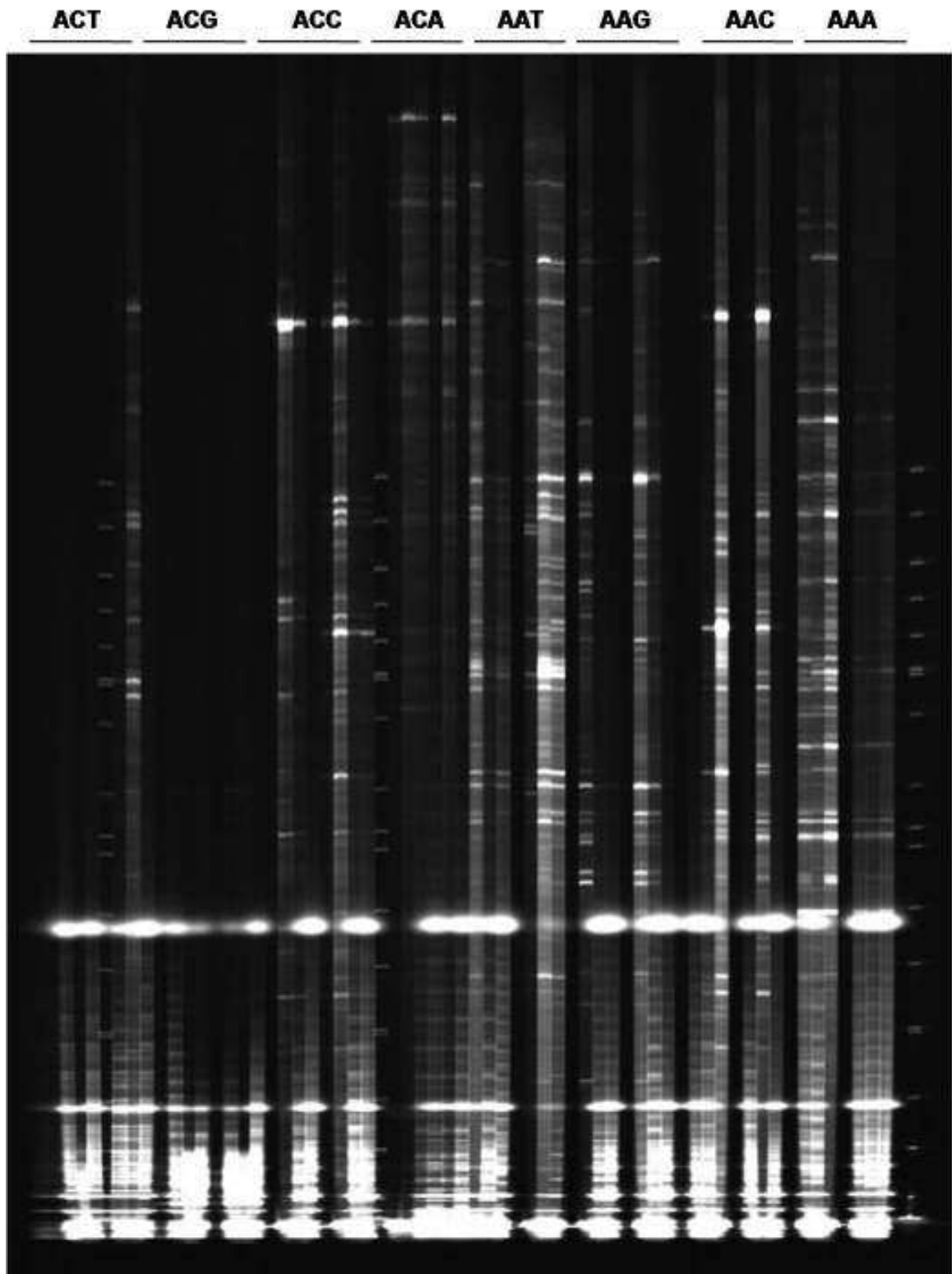
Figure D6: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Mse*I restriction enzyme with 3'-end of the primers having selective nucleotides from AAA to ACT.

Figure D7: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Mse*I restriction enzyme with 3'-end of the primers having selective nucleotides from AGA to ATT.

Figure D8: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Pst*I restriction enzyme with 3'-end of the primers having selective nucleotides from AAA to ACT.
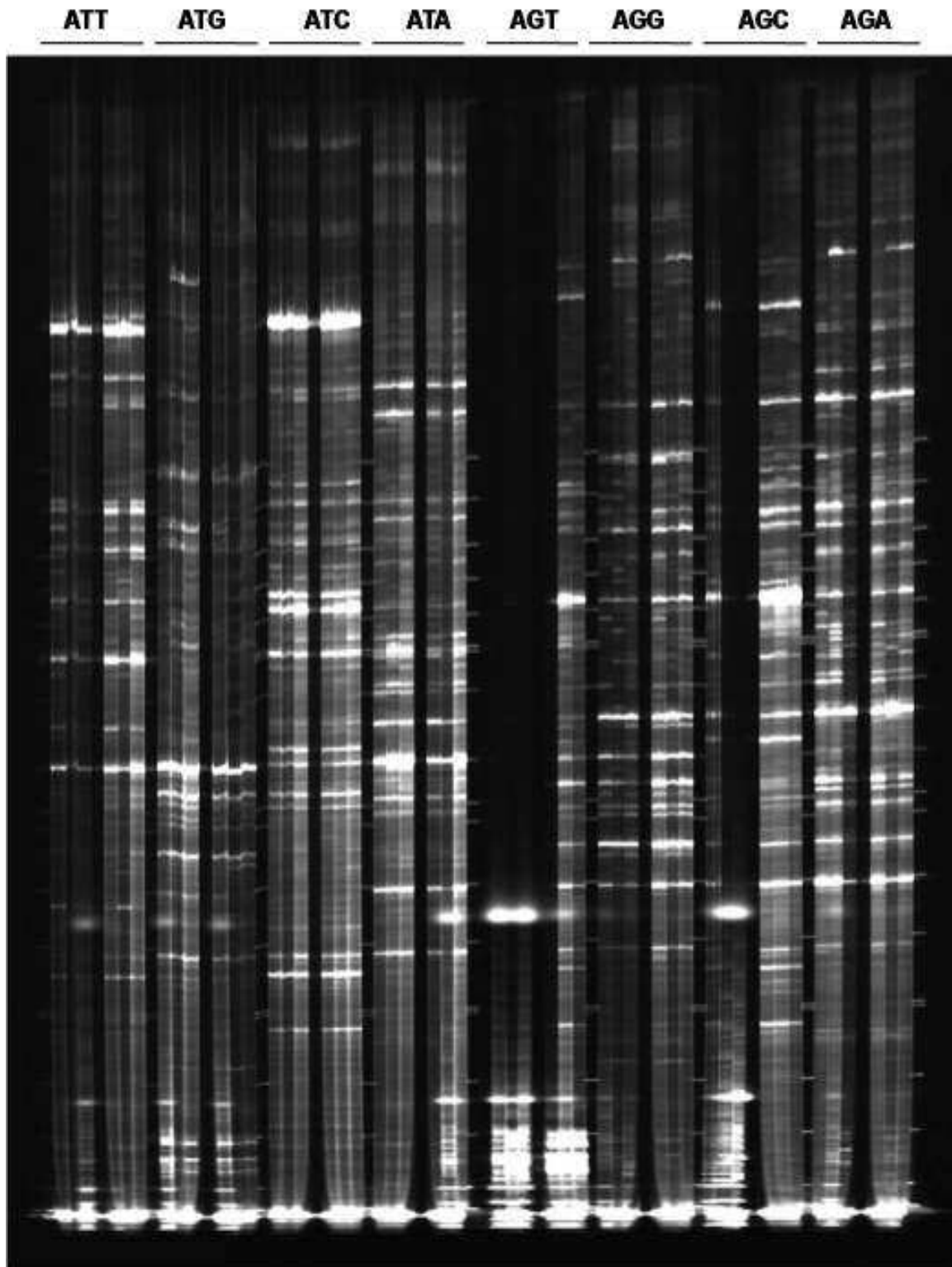
Figure D9: Electrophoresis profile of single-enzyme amplified fragment length polymorphism (AFLP) using the *Pst*I restriction enzyme with 3'-end of the primers having selective nucleotides from AGA to ATT.
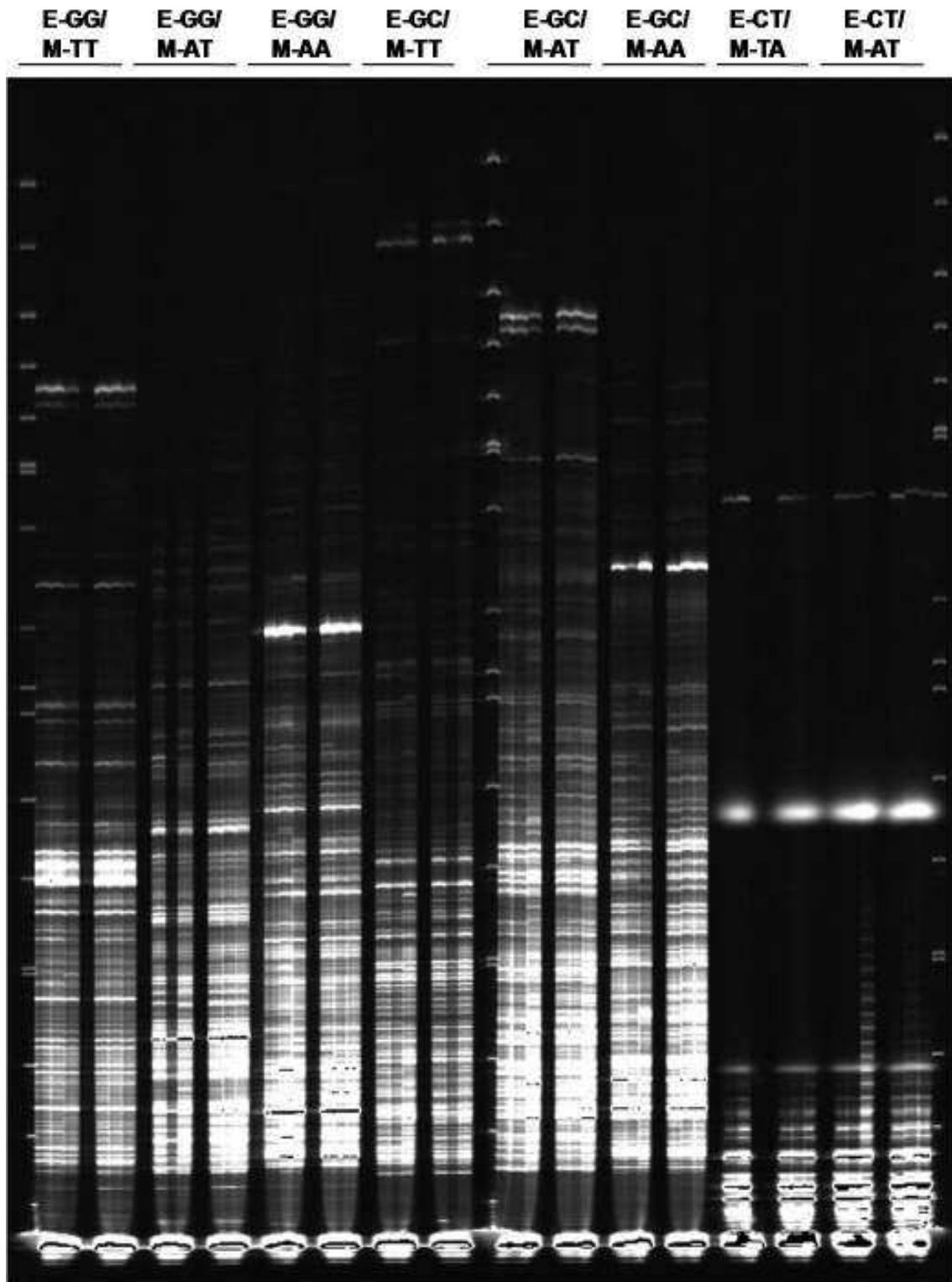
Figure D10: Electrophoresis profile of conventional amplified fragment length polymorphism (AFLP) using the combination *Eco*RI/*Mse*I restriction enzyme with 3'-end of the primers having three selective nucleotides each.