eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Doctoral Thesis

Crop improvement studies based on molecular approaches in interspecific Oil Palm hybrids

Maider Astorkia Amiama

Directed by: Dr. Mónica Hernández Muñoz & Dr. Begoña Marina Jugo Orrantia

Vitoria-Gasteiz, 2020

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Doctoral Thesis

Crop improvement studies based on molecular approaches in interspecific Oil Palm hybrids

Maider Astorkia Amiama

Directed by: Dr. Mónica Hernández Muñoz & Dr. Begoña Marina Jugo Orrantia

Special thanks to: Dr. Enrique Ritter Azpitarte and Dr. Stéphanie Sidibe-Bocs

Vitoria-Gasteiz, 2020

neiker tecnalia

LAFABRIL

Energy & Palma
Energía que evoluciona

Sampoerna Agro

INDEX

Gure hitzak

esan berriz esan

ez daitezela ahaztu

ez daitezela gal,

elur gainean

txori hanka arinek

utzitako arrasto sail

ederra bezalaxe.

Mikel Laboa

AGRADECIMIENTOS

Eusko Jaurlaritzako Ekonomiaren Garapen eta Azpiegitura sailari emandako diru laguntzagatik. Neiker Tecanaliari beraiekin lan egiteko aukera eta beharrezko baliabide guztiak ematearren.

A la Dra. Sonia Castañon por haber visto algo en mí en esa entrevista.

Al Dr. Enrique Ritter por acogerme en su equipo, guiarme, dejarme pensar y permitirme discutir, por enseñarme tanto y por la confianza.

A la Dra. Mónica Hernández por acogerme con los brazos abiertos, enseñarme sin nada a cambio, cuidarme y por el apoyo y el cariño.

À la docteur Stéphanie Sidibe-Bocs pour m'avoir aidé pendant mon séjour, et pas seulement en ce qui concèrne le côté scientifique mais aussi le personnel. Merci d'avoir eu confiance en moi et de m'avoir ouvert les portes de sa maison.

Begoña Jugori bere laguntzagatik tesia zuzentzen eta bideratzen.

To Energy Palma, La Fabril and Sampoerna Agro for letting me be part of the DAMASO and OLIMAS projects. For the trust placed in me.

A Leire, Ana Aragonés, Ana Herrán y Javi por escucharme, responder a todas mis dudas, que han sido muchas, por la paciencia y por las risas. A Maite.

A todos los compañeros de Neiker que están y han pasado. Ánimo a todos, enseguida lo teneís. Os lo mereceis.

A Alba, Itsasne y Sara por que además de compañeras se han convertido en amigas y confidentes.

Betiko laguneri urtez urte nire ondoan egotearren, familia izatearren. Nagoen tokian nagoela beti bisitatzeko edo deitzeko denbora edukitzearren. Zuekin beti izango dut nora bueltatu. Oyi, Nago, Amaia, Lore, Mait, Ira. Ez dadila haria eten.

Jaioneri, nahiz eta kilometrotara egon hurbilen sentitu dudan pertsonetakoa izatearren.

Janire eta Goizargiri. Mila esker zuen bizitzen parte izaten uztearren.

A Rocio, Tere y Aran por acompañarme estos años en Gasteiz.

A mi familia de Montpellier. A Sofia y Sara.

To my Cirad threesome. Fares, Julio and Sara, you made my days shine. Our coffees and the laugths is the best thing I bring from France. Thank you for letting me in.

À Veronique et à mon oncle Victor de m'avoir choyé tant et de m'avoir donné du bon vin.

Izeko eta osabei, bizitza guztian zehar zaintzearren.

Amari, Aitari eta Markeli. Zuek gabe hau ez zen posible izango. Mila esker beti nere erabakiak errespetatzearren, hor egotearren, maitasun guztiagatik. Ezin dezue imaginatu zein zorte oneko sentitzen naizen.

A Xuan. Por todo.

ABSTRACT

Oil Palm (OP) is the crop with the highest oil yield per hectare. The main OP plantations consist on *Elaeis guineensis* (Eg) species, known to produce high amounts of oil. However, in American regions this species is being affected by "Pudrición de Cogollo" disease. Due to this, OP companies started crossing this species with *E. oleifera* (Eo) palms which is resistant to this disease. The obtained interspecific hybrids show interesting characteristic inherited from both parents. However, little work has been done in the improvement of these hybrids. This thesis tries to address this gap applying different molecular approaches. First, an extensive study of a region of the "*Shell-thickness*" (Sh) gene has been conducted on 568 Eg, Eo and hybrid accessions. Then, with the aim to discover promising new Candidate Genes (CG) that could be exploited in further Molecular Assisted Selection Systems (MAS), a large phenotypic study of 25 production and quality traits have been performed within 198 hybrid accesions based on targeted CG and random Restriction site associated RNA sequencing (RARSeq) approaches, followed by Association Mapping (AM) assays.

The large screening performed in the 568 accessions through the Sh gene has enable to detect three new events (OLI1, OLI2, OLI3) and determine to be specific for the studied Eo accessions. Since NK2 SNP was specific for all Eo alleles, two species specific primers (ShG, ShO) were designed, tested and validated in all genotypes. An attempt to relate the detected new Eo events with four phenotypic traits was performed, but since the new variants are in intronic regions, no influence of the Eo Sh alleles in the studied phenotypes was observed.

The phenotypic study consisted of 6 production traits and 19 oil quality traits in which lipids, tocols and carotene content were studied. In the AM study based on CG targets, primer pairs designed in amplicons related to traits of interest were used for library construction and sequencing, while in the RARSeq approach a reduced transcriptome representation was sequenced.

In both assays AM studies were performed using four different models; two generalized linear models and two linear mixed models where in addition a Kinship matrix was added. In order to determine which of the models fitted best for each trait a new equation to calculate the average square distance from the diagonal was developed. In both assays the mixed models showed higher strengths in most of the tested traits. Even though some of the obtained results could be exploited for MAS, the limited portion of the studied genome has limited the results

in both assays and the need of improvement increasing the number of targets has been pointed out.

RESUMEN

La palmera de aceite es el cultivo con mayor rendimiento de aceite por hectárea. El principal cultivo de Palmera de aceite proviene de la especie *Elaeis guineensis* (Eg) conocido por producir grandes cantidades de aceite. Sin embargo, en las regiones Americanas esta especie se ve amenazada por la enfermedad "Pudrición de Cogollo". Para hacer frente a esta situación, las principales compañías de producción de aceite de palma empezaron a cruzar esta especie con palmas *E. oleifera* (Eo) resistentes a la enfermedad. Los híbridos interespecificos obtenidos muestran características interesantes heredadas de ambos progenitores, sin embargo, poco se ha estudiado sobre la mejora de estos híbridos hasta el momento. Esta tesis intenta abordar este espacio mediante diferentes técnicas moleculares. Por una parte se ha realizado un estudio extenso de una parte del gen "*Shell-thickness*" (Sh) en 568 genotipos Eg, Eo e híbridos. También se ha realizado un estudio fenotípico extensivo de 25 caracteres de producción y calidad de aceite en 198 genotipos híbridos seguido por dos estudios de Mapeo por Asociación (MA). Estos últimos se han abordado mediante la búsqueda de Genes Candidato (GC) y Restriction site associated RNA sequencing (RARSeq), con el fin de encontrar nuevos GC que se puedan explotar en sistemas de Mejora Asistida por Marcadores (MAS).

Este trabajo ha permitido determinar tres nuevos alelos (OLI1, OLI2, OLI3) específicos de los genotipos Eo estudiados. El SNP NK2 es único para todos los genotipos Eo estudiados y por lo tanto, se han podido diseñar, probar y validar dos primers específicos de especie (ShG, ShO) en todos los genotipos. A su vez, se ha realizado un estudio con el fin de relacionar estos nuevos alelos Eo a cuatro caracteres de interés. Sin embargo, no se observó relación alguna de los nuevos alelos y los caracteres estudiados.

En el estudio fenotípico se han tratado 6 caracteres de producción y 19 caracteres de calidad de aceite; lípidos, tocoles y contenido de carotenos. El estudio de MA se abordó de dos maneras diferentes. La primera mediante la generación de librerías basadas en GC, donde se amplificaron y secuenciaron regiones del genoma conocidas. La segunda utilizando la tecnología RARSeq donde se construyó una librería del transcriptoma reducido de los genotipos.

Para los estudios de MA se probaron cuatro modelos de asociación; dos modelos lineales generalizados y dos modelos mixtos, en el que además se añadió la matriz de correlación. Para poder determinar cuál de los modelos representaba mejor la asociación entre genotipo-fenotipo se desarrolló una nueva ecuación para calcular el cuadrado promedio de la distancia a

la diagonal. En ambos estudios los modelos mixtos mostraron mejores resultados para la mayoría de los caracteres estudiados. Aunque algunos de los resultados puedan utilizarse potencialmente en MAS, la pequeña proporción del genoma estudiada ha limitado los resultados y apuntan a la necesidad de mejora de los estudios incrementando el número de dianas.

LABURPENA

Olio Palmondoa (OP) hektareako olio-errendimendu handiena duen laborea da. OP-ren labore nagusiena *Elaeis guineensis* (Eg) espeziea da, duen olio produkzio altua dela eta. Amerikako zenbait eskualdetan "Pudrición de Cogollo" izeneko gaixotasunak erasotzen du labore hau. Honi aurre egiteko OP-ko enpresek labore hau *E. oleifera* (Eo) espeziearekin gurutzatu dute gaixotasun horrekiko duen erresistentzia dela eta. Hauen bitartez lortzen diren hibridoek ezaugarri interesgarriak erakusten dituzte bi gurasoengandik oinordetuak, baina hauen hobekuntzari buruz gutxi ikertu da. Hori dela eta, tesi honen bidez dagoen hutsuneari heldu nahi zaio teknika molekular ezberdinak erabiliz. Lehenik, "*Shell-thicknes*" (Sh) izeneko genearen zati bat aztertu da 568 genotipo Eg, Eo eta hibridoetan. Ondoren, Gene Hautagai (CG) berriak ikertzeko eta etorkizunean markagailu bidezko hobekuntza programetan (MAS) erabiltzeko asmoz, bi asoziazio mapa (AM) gauzatu dira CG eta "Restriction site associated RNA sequencing" (RARSeq) bidez 198 genotipo hibridoetan. Bertan, 25 olio ekoizpen- eta kalitate-karaktere ikertu dira.

Ikerketa honek hiru alelo berri (OLI1, OLI2, OLI3) zehaztea ahalbidetu du Eo genotipoetan bakarrik azaltzen direla zehaztuz. NK2 SNP-a ikertutako Eo genotip guztietan ikusi da. Hori dela eta, espezie espezifikoak diren bi primer, ShG eta ShO, diseinatu, probatu eta balidatu dira ikertutako genotipo guztietan. Zehaztutako Eo Sh alelo berriak intereseko karaktereekin erlazionatu nahi izan dira baina emaitzetan ez da inolako erlaziorik detektatu hauen artean, alelo berriak eskualde intronikoan baitaute.

Ikerketa fenotipikoan 6 produkzio-karaktere eta 19 kalitate-karaktere; lipidoak tokolak eta karoteno edukia barne, ikertu dira. AM ikerketa bi eratara bideratu da. Lehenengoa CG bidez, non intereseko genoma zatiak anplifikatu eta sekuentziatu diren. Bigarrenean, RARSeq teknologia erabili da, genotipoen transkriptoma murriztua sekuentziatuz.

Bi ikerketetan lau asoziazio eredu aztertu dira; bi eredu lineal orokor eta bi eredu misto non korrelazio matrizea gehitu den. Aztertutako modeloetatik hoberena aukeratzeko karaktere bakoitzarentzat batez besteko (d^2) distantzia diagonaletik neurtzeko ekuazio berria garatu da. Bi ikerketetan modelo mistoek emaitza hoberenak lortu dituzte aztertutako karaktere gehienetan. Nahiz eta lortutako hainbat emaitza MAS sistemetan aplika daitezkeen, ikertutako genomaren zati txikiak gure ikerketak mugatu ditu eta emaitzak hibetuko lirateke jomuga kopurua handituz.

List of Acronyms and Abbreviations

µL	Micro litre
µM	Micro molar
AFLP	Amplified fragment length polymorphism
Alpha	Alpha compound
Alpha3	Alpha3 compound
AM	Association mapping
ANOVA	Analysis of variance
BN	Bunch number
BW	Bunch weight
BY	Bunch yield
CAPS	Cleaved amplified polymorphic sequences
Car	Carotene content
cDNA	Complementary DNA
CG	Candidate gene
CxL	Coari x La Mé
d^2	Average square distance
ddRAD-Seq	Double digestion restriction site associated DNA
Delta	Delta compound
Delta3	Delta3 compound
DNA	Deoxyribonucleic acid
D	*Dura*
Eg	*Elaeis guineensis*
Eo	*Elaeis oleifera*
FA	Fatty acids
FDR	False discovery rate
Fis	Inbreeding coefficient

Fst	Fixation indices
Gamma	Gamma compound
Gamma3	Gamma3 compound
Gb	Giga base
GLM	Generalized linear model
GWAS	Genome-wide association mapping
ha	Hectares
He	Expected heterozygosity
Ho	Observed heterozygosity
HWE	Hardy–Weinberg equilibrium
IV	Iodine value
K	Kinship matrix
LD	Linkage disequilibrium
MAS	Molecular assisted system
MLM	Linear mixed model
MMT	Million metric tones
Mono-Un	Mono-unsaturated fatty acids %
MPOB	Malaysian palm oil board
mRNA	Messenger RNA
MRRS	Modified reciprocal recurrent selection
MRS	Modified recurrent selection
ng	Nano gram
NGS	Next generation sequencing
OA	Oleic acid %
OilB	Oil percentage in the bunch (%)
OilB	Oil % in bunch
OildM	Oil % in dry mesocarp

OilfM	Oil % in fresh mesocarp
OilM	Oil percentage in the mesocarp (%)
OP	Oil palm
PCA	Principal component analysis
PCR	Polymerase chain reaction
PGM	Personal genome machine
P	*Pisifera*
PisC	Pisifera Congo
PisN	Pisifera Nigeria
PO	Palm oil
Poly-Un	Poly-unsaturated fatty acids %
Q	Structure matrix
QQ	Quantile-quantile
QTL	Quantitative trait loci
RADSeq	Restriction site associated DNA
RAPD	Random amplification of polymorphic DNA
RARSeq	Restriction site associated RNA sequencing
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RNA-Seq	Next generation cDNA sequencing
RSPO	Roundtable on sustainable palm oil
Sat	Saturated fatty acids %
Sh	*Shell* gene allele
SNP	Single nucleotide polymorphism
SSR	Simple sequence repeat
SSS	SSS triglyceride
SUS	SUS triglyceride

SUU	SUU triglyceride
T	*Tenera*
TG	Triglycerides
Toc	Tocols
Toc3	Tocotrienols
Tocph	Tocopherol
TxA(O)	Taisha x Avros (Oleoflores)
TxA(RGS)	Taisha x Avros (RGS)
TxE	Taisha x Ekona
TxY	Taisha x Yangambi
UUU	UUU triglyceride
VCF	Variant calling format

Index of Tables

Index of Tables in Anex

Index of Figures

Index of Figures in Anex

CHAPTER I: INTRODUCTION

CHAPTER I: INTRODUCTION

1. The Oil Palm tree

Oil Palm is a monocotyledon, part of the Arecales order (Cronquist, 1981), Arecaceae family and of *Elaeis* gender. Currently, two *Elaeis* species are accepted; *E. guineensis* (Eg) and *E. oleifera* (Eo).

The main OP plantation comes from Eg also known as African OP and has its origin in Africa, specifically in the Gulf of Guinea, North East Africa (Ergo, 1997). This crop is mainly cultivated in the tropical areas of Africa, South East Asia and Central and South America (Corley, R.H.V. and Tinker et al. 2016) as can be seen in Figure 1.

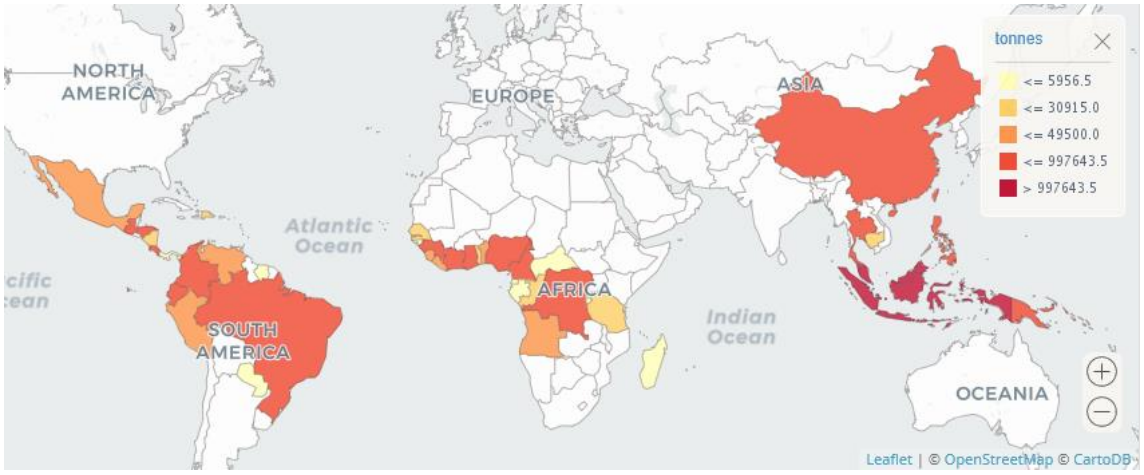


Figure 1: Production average quantities of Oil palm by country from 1994 to 2017 (obtained from FAO (2019)).

Eg can grow to 15-18 meters, even to 30 meters in deep forests. Its leaves can reach lengths of up to 8 meters. Eg can show three different fruit forms determined by the "*Shell-thickness*" (Sh) gene which presents two alleles (Sh+ and sh-):

- *Dura* (D): homozygote (Sh+ Sh+). Shows a thick endocarp (2-8mm) and a small mesocarp (35-65 % mesocarp/fruit) with high oil contents.
- *Pisifera* (P): homozygote (sh- sh-), shows no endocarp and a big mesocarp, but is usually female sterile.
- *Tenera* (T): heterozygote (Sh+ sh-) F1 hybrid from D x P with elevated oil contents. Shows a thin endocarp (0.5-4 mm) and a big mesocarp (55-96 % mesocarp/fruit). The

oil yield is much higher than that of either parent. Commercial varieties are generally *tenera*(Singh et al. 2013a).

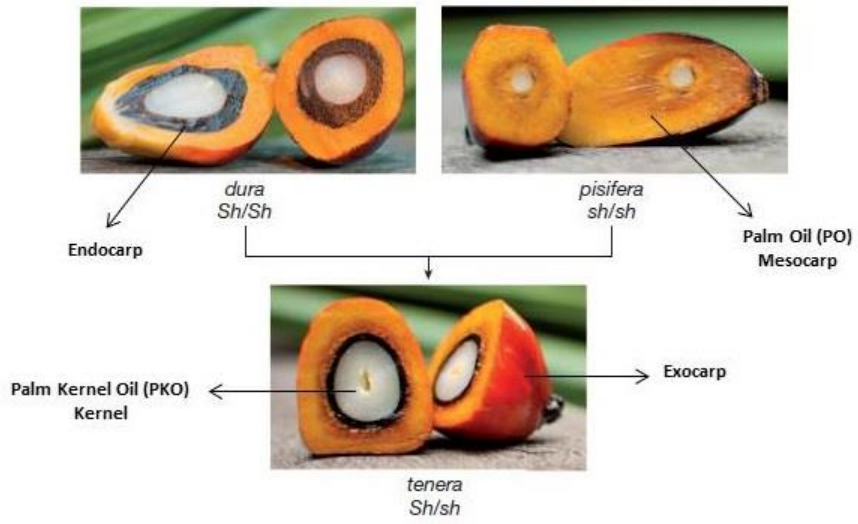


Figure 2: Fruit forms of the African OP (adapted from (Singh et al. 2013a)).

The Sh gene, whose sequence was identified in 2013 by Singh et al. (2013a), represents the most important economic aspect of OP. Two independent mutations in this gene, $Sh^{(AVROS)}$ and $Sh^{(MPOB)}$, codifying *pisifera* haplotypes derived from Congo and Nigeria respectively were found in the DNA-binding domain of the MAD-box-gene. This gene is known to control ovule identity and seed development in *Arabidopsis.* In addition, in 2016 the same group published three novel events, Sh^{MPOB2}; Sh^{MPOB3}; Sh^{MPOB4}, which were found in small frequencies in the population studied (Ooi et al. 2016). Figure 3 summarizes Sh protein sequence together with all those detected variants:

```
wild-type: MGRGKIEIKRIENTTSRQVTFCKRRNGLLKKAYELSVLCDAEVALIVFSSRGRLYEYANN
sh^AVROS: ----------------------------N----------------------------
 sh^MPOB: ----------------------------P----------------------------
sh^MPOB2: ----------------------Q-----------------------------------
sh^MPOB3: ---------------------------------------D------------------
sh^MPOB4: ----------------------N-----------------------------------
```

Figure 3: *Shell* (Sh) MADS box domain mutations associated with the fruit form; wild-type refers to d*ura* and Sh^{AVROS}, Sh^{MPOB}, Sh^{MPOB2}, Sh^{MPOB3} and Sh^{MPOB4} refer to *pisifera* haplotypes (adapted from (Ooi et al. 2016)).

Two fruit types can be distinguished in terms of fruit ripening: (1) the common type *nigrescens* with anthocyanin in exocarp and black or brown fruit, (2) *virences* type with no anthocyanin in exocarp, green color when unripe and orange with greenish tip when ripe (Corley, R.H.V. and Tinker et al. 2016). Since the latter undergoes a greater change of color it is easy to determine the moment when bunches are ripe, thus minimizing yield losses. Singh et al. also identified the sequence of the *virescens* (VIR) gene as the one controlling fruit exocarp color and the indicator of ripeness. VIR is a R2R3-MYB transcription factor with homology to *Lilium* LhMYB12 and similarity to *Arabidopsis* production of anthocyanin pigment1 (PAP1) (Singh et al. 2014).

Figure 4: Fruit types of the African OP. a) Nigrescens fruit types and b) Virescens fruit types (adapted from (Singh et al. 2014)).

Along with Eg, American OP or Eo (Wessels-Boer, 1965) originated in equatorial America, but is of minor commercial interest for PO production. This species is grown in Central and South America and is characterized by lower oil production. However, it shows desirable properties such as better oil quality (Pelaez et al. 2010) (Table 1), resistance to the main diseases affecting palms (Salavarrieta and Jesús 2004) and shorter trunks which prolongs palms productive life (Barba 2019). This species only shows one fruit form with similar morphological characteristics to Eg *dura*. Therefore, Eo is known as *dura* form or Wild form (Montoya et al. 2014; Corley, R.H.V. and Tinker et al. 2016). Regarding the fruit type, 90 % of the palms show similar fruit color to the Eg *virescens* type (Corley, R.H.V. and Tinker et al. 2016).

Since the early 1920s, when the first breakout was reported, Eg has suffered from Bud Rot disease, also known as “Pudrición de Cogollo” in American regions, leading to dead palms

(Sundram and Intan-Nur 2017). The initial symptoms consist in the fall of the young leaves and the loss of the spear leaf. Afterwards, light absorption starts dropping as few or no leaves are present, leading to low biochemical activity and affecting metabolic processes (Moreno-Chacón et al. 2013; Avila-Diazgranados et al. 2016). Even though it is not clear what causes the disease, some experts identify *Phytophthora palmivora* as the causing pathogen (Torres et al. 2010). In order to mitigate economic losses, OP companies started to cultivate hybrids between both species since the Eo parent inherits resistance to the disease (Amblard et al. 2004; Preciado et al. 2011). Breeding programs started to produce Eg x Eo crosses and the first results indicated that by using Eg as the pollen donor (Eo x Eg) oil yields similar to those of Eg *tenera* could be achieved (Corley, R.H.V. and Tinker et al. 2016; Barba 2019). Other desirable characteristics are inherited from Eo. Hybrids show trunk height decrease of around 50 % lower than Eg, thus, prolonging the harvesting period (Torres et al. 2004). Also a better oil quality (Table 1) is obtained compared to Eg due to the increased percentage of oleic acid, lower content of saturated FA and increased iodine values, all interesting qualities from a nutritional point of view (Mozzon et al. 2013; Cadena et al. 2013; Corley, R.H.V. and Tinker et al. 2016). By contrast, oil production is lower than in Eg and assisted pollination is required (Raquel Meléndez and Ponce 2016). The resulting fruit form of the hybrids is determined by the Eg parent, either *dura* or *tenera* (Corley, R.H.V. and Tinker et al. 2016).

Table 1: Oil composition from *E. guineensis* (Eg), hybrids from *E. oleifera* (Eo) x Eg and Eo (adapted from (Corley, R.H.V. and Tinker et al. 2016)).

Fatty acid		*E. guineensis*	*E. oleifera x E. guineensis*	*E. oleifera*
Palmitic	C16:0	27-64	27-41	35-41
Stearic	C18:0	1-13	1-6	1-5
Oleic	C18:1	23-54	43-59	43-48
Linoleic	C18:2	2-18	8-15	9-14
Iodine value		32-65	58-71	58-62

Regarding the genetic composition, both species of the genus *Elaeis*, as well as interspecific hybrids, contain the same number of chromosomes 2n=32. In 2013, Singh et al. (2013b) published the whole-genome sequence of an Eg *pisifera* palm and determined a genome size of 1.8 Giga bases (Gb). 1.535 Gb could be assembled and 34.802 genes were annotated. The guanine-cytosine content (GC %) was settled as 37 % over the whole genome, while in genes it

increased to 50 %. Generally, GC percentage in monocot plant ranges from 33.6 to 48.9 % (Šmarda et al. 2014). The Eo genome was also sequenced with a combination of fragment and linker libraries(Singh et al. 2013b). Furthermore, due to the segmental duplications of chromosome arms they determined the palaeotetraploid origin of the African and American oil palm and positioned them in the evolutionary tree. As can be seen in Figure 5 both *Elaeis* species along with *Phoenix dactylifera* (date) were well separated from other species such as *Musa acuminatea* (banana), *Curcuma longa* (turmeric) or *Zingiber officinale* (ginger). They predicted a divergence time of 65 million years between date and oil palm and 51 million years between Eo and Eg.

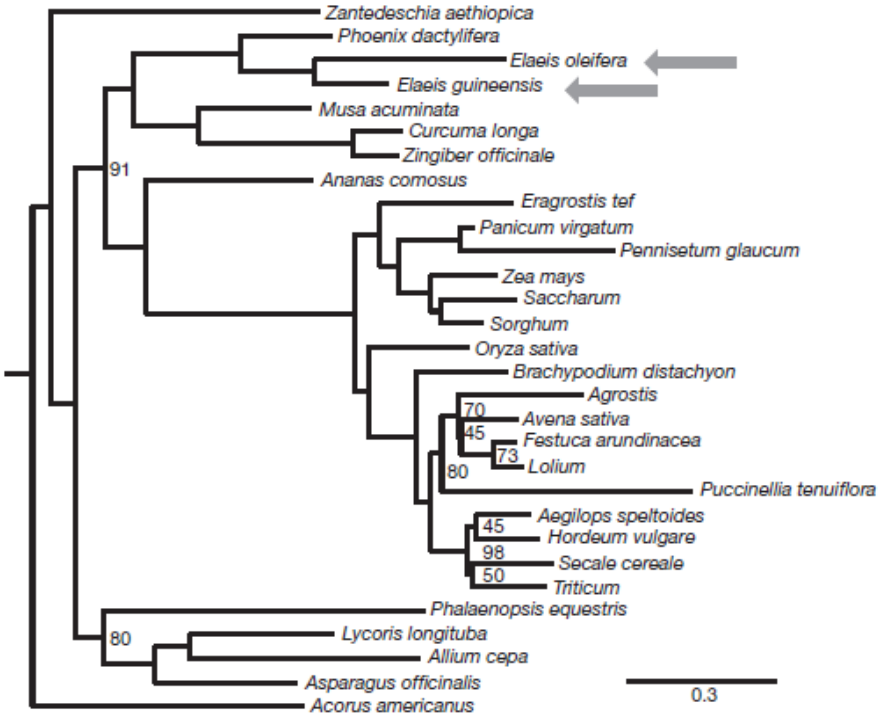


Figure 5: Evolutionary phylogenetic tree of crops (Singh et al. 2013b).

2. General knowledge about Oil Palm

Oil Palm (OP) is the crop with the highest oil yield per hectare as it is able to produce up to ten times more oil than other leading oilseed crops, as can be seen in Figure 6. As a result, its cultive has spread rapidly in tropical regions of Asia, Africa and America with a global production of 84.82 Palm Oil (PO)million metric tons (MMT). 76.01 MMT come from Crude Palm Oil (CPO) which is obtained from the mesocarp of the fruit, while 8.81 MMT are extracted from the kernel of the fruit, Palm Kernel Oil (PKO), according to the United States Department

of Agriculture (USDA) in September 2019/20 (USDA 2019). This makes OP the largest vegetable oil source worldwide.

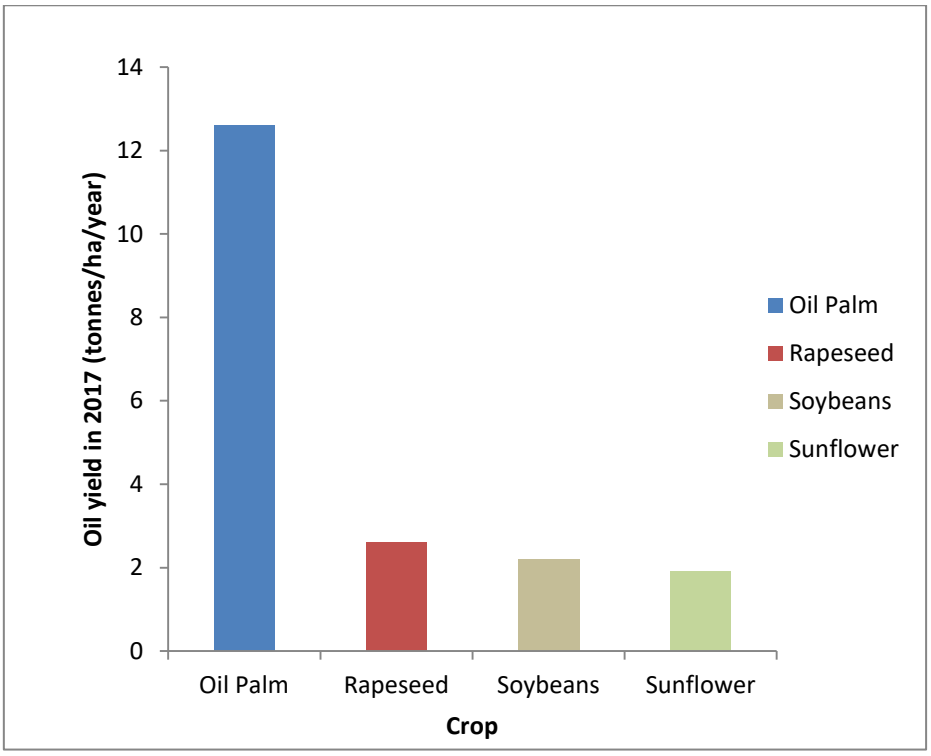


Figure 6: Oil palm (OP) efficiency compared to other vegetable oils (data adapted from FAO(2019)).

Indonesia and Malaysia are the most productive countries producing 71 % of all PO worldwide last year, followed by Thailand, Colombia and Nigeria (USDA 2019). India is, furthermore, the major importer of PO with 10.7 MMT, followed by the European Union with 6.9 MMT and China with 6.7 MMT.

The main components of CPO, at over 95%, are mixtures of triglycerides (TG) made up of different fatty acids (FA) shown in Table 2. The most abundant FAs are myristic (14:0), palmitic (16:0), stearic (18:0), oleic (18:1) and linoleic (18:2) acids. The proportion of saturated and unsaturated acids is approximately equal, with 7 to 10 % saturated TG, and 6 to 12 % fully unsaturated TG(Corley, R.H.V. and Tinker et al. 2016).

Table 2: Typical fatty acid composition (%) of palm oil (PO). (obtained from Sundram et al. (2003)).

Fatty acid chain length	Mean	Range observed	Standard deviation
12:0 lauric	0.3	0-1	0.12
14:0 myristic	1.1	0.9-1.5	0.08
16:0 palmitic	43.5	39.2-45.8	0.95
16:1 palmitoleic	0.2	0-0.4	0.05
18:0 stearic	4.3	3.7-5.1	0.18
18:1 (n-9) oleic	39.8	37.4-44.1	0.94
18:2 (n-6) lionoleic	10.2	8.7-12.5	0.56
18:3 linolenic	0.3	0-0.6	0.07
20:0 arachidic	0.2	0-0.4	0.16

The minor constituents, at less than 1 %, can be divided into FA derivatives such as partial glycerides, sterols and phosphatides and compounds unrelated to FA; free alcohols, pigments (carotenoids and chlorophylls), trace metals or tocopherols. Although these are found in minor proportions they play significant roles and are of nutritional importance. Most of them are summarized in Table 3.

Table 3: Minor PO compounds (adapted from Sundram et al. (2003)).

Carotenoids	%	Vitamin E	%	Sterols	%
α-carotene	36.2	α-tocopherol	28	Cholesterol	4
β-carotene	54.4	β-tocopherol	29	Campesterol	21
γ-carotene	3.3	γ-tocopherol	28	Stigmasterols	21
Lycopene	3.8	δ-tocopherol	14	β-sitosterol	63
Xanthophylls	3.2				
Total (ppm)	500-700		500-800		~300

PKO is similar to coconut oil in that it is rich in saturated fatty acids, mainly lauric acid (about 50 %) and has a smaller proportion of unsaturated fatty acids (Corley, R.H.V. and Tinker et al. 2016).

The use of PO can be divided into two main groups; about 80 % of PO is for edible and 20 % for non-edible products. For food products it can be used as a frying medium due to its resistance to oxidative deterioration, suffering lower polymer formation. It also contains vitamin E and can be mixed with other oils. It can be used for shortening in cooking and baking and for novel food products such as ice cream, whipping cream or cheese. The non-food products have direct uses such as in soaps or printing ink and as oleochemicals such as fatty acids, fatty alcohols or fatty nitrogen. PKO is mainly used for these latter purposes (Basiron and Weng 2004).

Unfortunately, only 10 % of palm production takes the form of oil so replanting also leaves large amounts of trunks and fronds, leading to high levels of biomass waste. To deal with this problem, OP itself, biomass and biogas are used as renewable resources for vehicle propulsion or power generation (Basiron and Weng 2004).

Because of the high demand for PO, its planting has expanded greatly, to the detriment of species diversity and of carbon-rich tropical forests. From year 2000 to 2001, for example, an average of 270.000 ha of forest were converted into OP plantations, leading to deforestation (Vijay et al. 2016). Furthermore, before planting OP plantations need to be cleared either mechanically or by fire, at a high environmental cost (Lal 1996; Schrier-Uijl et al. 2013).

In 2004 the Roundtable on Sustainable Palm Oil (RSPO 2019) was founded with the main objective of protecting environment and society. In 2017, 20 % of global oil production was certified by RSPO (Carlson et al. 2018). Certification of its members is carried out according to RSPO Principles & Criteria (RSPO 2018). These do not require zero deforestation, but their limits enhance high conservation values or high carbon stock forest and set a threshold of 100 ha for peatland areas. They also ban the use of fire for land preparation and for disposal of biomass wastes. Furthermore, certified members have to conserve local community rights by gaining their free, prior and informed consent.

Apart from this controversial matter there is another important ongoing debate, mainly in Europe, about PO´s unhealthy properties. Due to its chemical and physical properties PO has become the ideal component for food industry. However, because PO is naturally bland and light in colour, the food industry makes its refining a mandatory step. This step can be carried out by chemical or physical methods, but in both the remaining oil loses tocols and there is oxidative damage (Gunstone 2011; Dunford 2012). Along with this, the high level of saturated fatty acids, particularly palmitic acid, has underlined the relation between PO intake and different diseases such as obesity, cardiovascular disease, type 2 diabetes mellitus or cancer

(Saadatian-Elahi et al. 2004; Kochikuzhyil et al. 2010; de Wit et al. 2012). However, diverse reviews about this subject disagree. Mancini et al. (2015) published that the different studies relating palmitic acid with the mentioned diseases give controversial results. Also, Marangoni et al. (2017) determined that no data show the relation between palm oil consumption and cancer incidence or mortality. Furthermore, they concluded that the effects should be considered similar to that of other oils or solid fats that are rich in saturated fatty acids and that its replacement by other fats in food production would not necessarily contribute to the overall nutritional profile.

3. Classical crop improvement

Breeding programs are implemented in most crops in order to obtain improved characteristics for commercial purposes. For this purpose breeders start with a population where a genetic variation exists and then select which desirable characters to work on. In the case of OP big differences exist between the two species.

3.1. *Elaeis guineensis* (Eg)

In the case of Eg, since the early 1920s crop improvement programs have focused on maximizing oil and oil kernel yield to deal with increasing demand for PO. Early improvement programs took place in the Congo and in the Far East where large scale plantations were cultivated and selection and breeding programs carried out. The approach was different for each of the locations: in Africa they focused on improvement in high-quality T material and in Asia on Deli D populations.

In Africa the first plantations established in Gabon were unsuccessful. However, subsequent plantations in Belgian Congo (nowadays Dmocratic Republic of Congo), French territories (Ivory Coast) and Nigeria led to increased exports of PO and kernels. In 1940 Beirnaert determined the inheritance of the Sh gene when he examined a D x T population in Yangambi (Africa) and saw no P genotypes. Furthermore, he detected that the majority of those crosses segregated close to 50:50 D:T (Beirnaert 1941). This discovery changed the methods implemented beforehand and breeding programs focused on *dura* x *pisifera* crosses with *dura* as female palm and *pisifera* as pollen donor. *Tenera* palms obtained from these crosses showed higher oil content with 30 % more oil extraction compared to *dura,* but also higher bunch yield as well as higher growth vigor. The differences between the three fruit forms are explained in a previous section (Section 1.2). At the end of the 1920s the most productive plantations were located in the Ivory Coast, with the help of the French government (Corley, R.H.V. and Tinker et al. 2016).

In Asia, by contrast, the first seedlings introduced were planted in the Buitenzorg (now Bogor) Botanic Gardens in 1848 in Java. Four different origins were planted, two from a botanic garden in Amsterdam and two from 'Bourbon or Mauritius' in the Indian Ocean. Since the palms from those seeds were quite similar it is thought that they all had their origin in Africa, possibly from a single parent palm. The progeny of these palms were transferred to Sumatra in 1875 and became the foundation stock for South East Asia, while some were planted near Delhi as ornamental palms. The latter were found to be productive and eventually they became the first productive materials in Sumatra and Malaya (Henson 2012).

Once the Sh gene was discovered, most of the breeding programs all over the world were based on two basic population improving schemes: the modified recurrent selection (MRS) explained in Figure 7 and the modified reciprocal recurrent selection (MRRS) shown in Figure 8 (Soh et al. 2017).

In MRS scheme, the parents, *dura* (D) (usually Deli D) and *tenera* (T) or *tenera*/*pisifera* (P), are selected for recurrent cycle trials on D x D and T x T crosses. Depending on their performance, D and P top genotypes are then crossed in a D x P top-cross progeny-test. The remaining selected D and P genotypes are afterwards used for commercial D x P hybrid production.

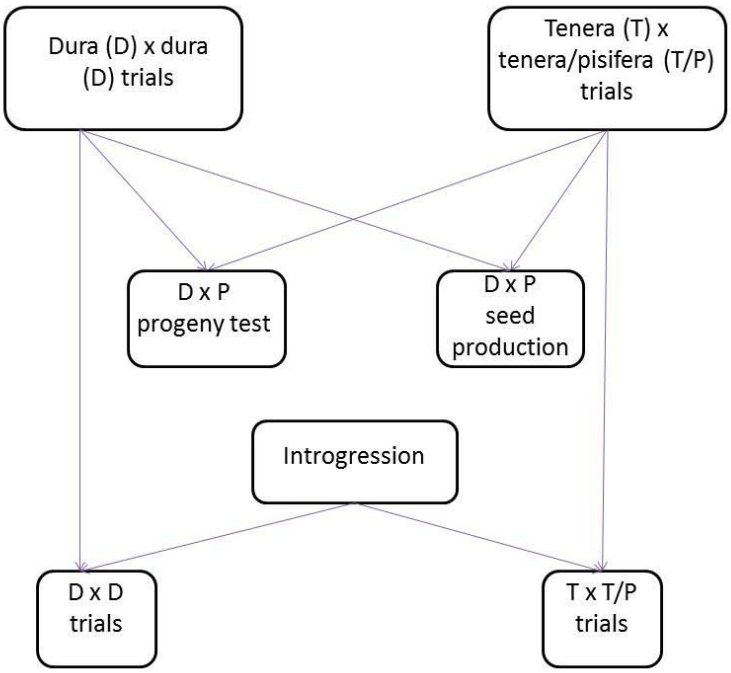


Figure 7: Generic modified recurrent selection (MRS) scheme in OP (obtained from (Soh et al. 2017)).

Moreover, in MRRS selected top D and P/T parents from selected D x T progeny test are self-pollinated (self) and germplasm is used for breeding and hybrid seed production. Usually this self-pollination is limited to two cycles due to severe inbreeding depression. Also a recurrent recombinant cycle is developed within parental populations for improvement and longer term populations. MRRS takes more planting space since both progeny and self-pollination crosses have to be tested.

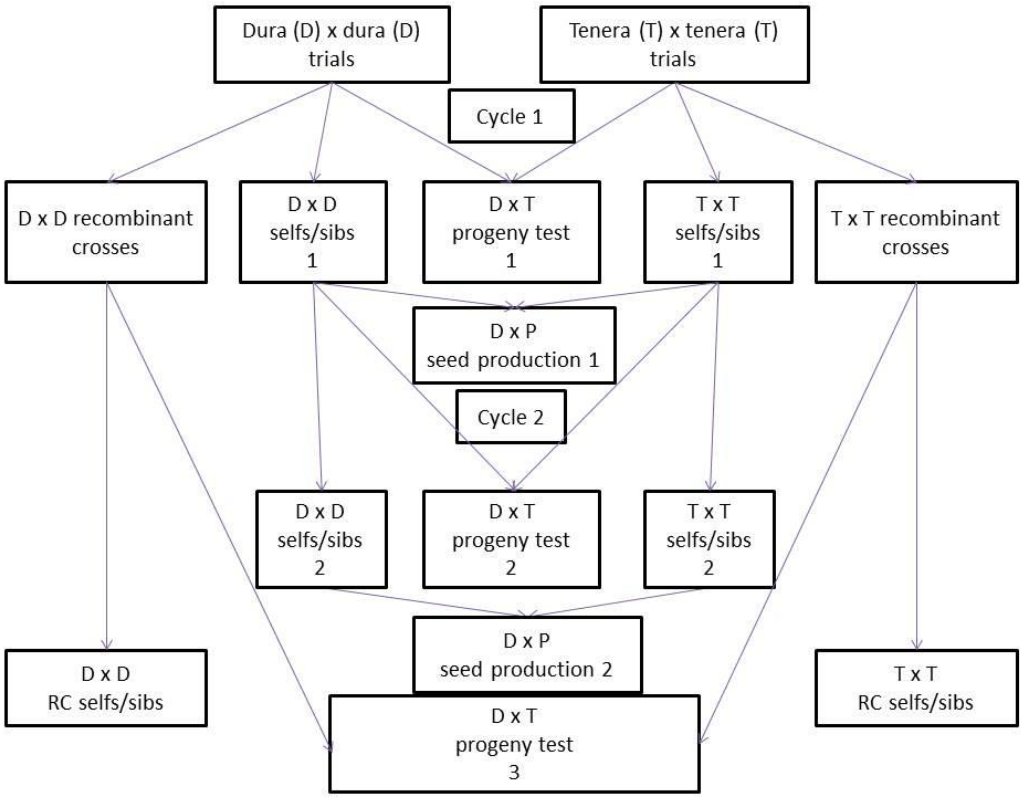


Figure 8: Generic modified reciprocal recurrent selection (MRRS) scheme in OP (obtained from (Soh et al. 2017)).

3.2. *Elaeis oleifera* (Eo)

Regarding the second species, less information exists about the breeding programs applied to Eo.

Native populations of Eo have been found in Brazil, Colombia, Costa Rica, Ecuador, French Guinea, Honduras, Nicaragua, Panama, Peru, Suriname and Venezuela (Corley, R.H.V. and Tinker et al. 2016).

In Ecuador, 39 ha plantation of Eo had been established by INIAP (Instituto Nacional de Investigaciones Agropecuarias) by 1965. This institute maintains a seed bank but there is very little public information about it. First botanical collections in Ecuador date from 1986 when native Ecuadorian palms were first found in the Taisha locality in the province of Morona Santiago (Ecuadorian Amazon). In 1990, wild Eo palms were found in Pastaza and Orellana provinces (Montúfar et al. 2018).

In Costa Rica a breeding program in Coto OP Plantation has existed since 1967, with Eo genotypes comprising 43 regions and covering seven countries. This breeding program has achieved successful results commercializing Eo x Eg F1 hybrids and backcross hybrids named "Compactas".

IRHO (now Cirad, France) started systematic studies and collections of Eo from Central America (Mexico, Nicaragua, Guatemala, Costa Rica, Panama) and South America (Colombia) in 1968 and an important collection process has taken place in Africa since 1975.

Also, the Malaysian Palm Oil Board (MPOB) extended its prospection program to Eo, collecting seeds in Colombia, Panama, Costa Rica, Honduras, Nicaragua, Brazil and Surinam in 1981-1982. From this collection, material from Brazil and Surinam showed much lower fresh fruit bunches, oil yields and iodine values. However, Brazilian material showed high oil extraction ratios (Montúfar et al. 2018).

3.3. Interspecific hybrids

In 1981 the first commercial plantation of interspecific Eg x Eo hybrids was implemented in Colombia due to the resistance of hybrids to "lethal yellowing" (Turner and Incorporated Society of Planters. 1981). Furthermore, in 1999 Amblard et al. (Seminar on Worldwide performance of DXP oil palm planting materials et al. 1999) determined Eo x Eg hybrids as resistant to Bud Rot disease. These properties as well as others explained in Section 1.2 converted these hybrids into promising source of new genetic variation.

Even though the first plantation trials were Eg x Eo crosses most of the present-day trials are Eo x Eg, as yields can be compared with Eg *tenera* palms. At first, Hardon (1969) and Meunier and Boutin (1975) quoted poor oil extraction rates for Eo x Eg leading to low oil yields in the bunch (OilB) of only 17-18 %, when in good commercial *tenera* 22-23 % were collected. However, recent results show improved results of OilB with 24 % and oil percentage in mesocarp (OilM) of 47 % (Corley, R.H.V. and Tinker et al. 2016).

4. Molecular crop improvement

Even though conventional methods have been widely and effectively used over the years, some difficulties still exist; 1) the inheritability of undesirable traits, 2) polygenic behaviour of most of the desired traits; 3) interaction with the environment, 4) long breeding cycles additionally prolonged due to the need of using manual pollination and 5) the large amount of land needed to run the trials.

Since 1980s, molecular marker technology has been used to avoid these problems and to study large amounts of diversity through genes or proteins. The different markers have been used as confirmation of pedigree or legitimacy, to assess genetic diversity or in "marker-assisted" selection systems in order to select desired genotypes. This technology has evolved greatly and nowadays new, fast and cost effective sequencing technologies enable the sequencing of whole genomes and transcriptomes. The large amount of data obtained from these assays has highlighted bioinformatics as a large new field to work in now that tools able to manage all this information have become of high importance.

4.1. Molecular markers

Due to their stability and their capability to distinguish between genotypes, molecular markers are highly valued in genetics based research. The first molecules to be used as markers were secondary metabolites but their instability restricted their use (Grover and Sharma 2016). After these, enzyme markers (isoenzymes) were used (Tanksley et al. 1981). Genotypes are distinguished for having allelic variations of enzymes due to genetic or epigenetic differences. Isoenzymes were successfully used for a short time for the detection of genetic diversity or population structure. However, their use was limited due to the small number of enzymatic dyes and number of markers to cover a whole genome (Tanksley 1993). DNA markers rely on the polymorphism present between different individuals. The bases of these polymorphisms are insertion, deletion, point mutations, duplication and translocation found across the genome. Even though different types of markers have emerged over the years, only some have received global acceptance. These markers are summarized and characterized in Table 4.

Table 4: Characteristics of the most popular molecular markers used in plants. (adapted from (Nadeem et al. 2018)).

Characteristics	Restriction Fragment Length Polymorphism (RFLP)	Random Amplification of Polymorphic DNA (RAPD)	Amplified Fragment Length Polymorphism (AFLP)	Simple Sequence Repeat (SSR)	Single Nucleotide Polymorphism (SNP)
Co-dominant /Dominant	Co-dominant	Dominant	Dominant	Co-dominant	Co-dominant
Reproducibility	High	High	High	High	High
Polymorphism level	Medium	Very High	High	High	High
Required DNA quality	High	High	High	Low	High
Required DNA quantity (ng)	10000	20	500-1000	50	50
Cost	High	Low	High	High	Variable
Sequencing	Yes	No	No	Yes	Yes
Status	Past	Past	Past	Present	Present
PCR requirement	No	Yes	Yes	Yes	Yes
Visualization	Radioactive	Agarose gel	Agarose gel	Agarose gel	Bioinformatic tools

These types of DNA markers have been widely used in OP. Billote et al. (2005) developed a high-density linkage map based on 255 SSR and 688 AFLP markers of 16 linkage groups corresponding to the 16 Eg chromosomes. In order to study their genetic diversity, Arias et al. (2014) used 20 SSR within different populations of Eg and Taeprayoon et al. (2015) used 96 SSR in 121 Eg genotypes from three breeding populations. Ritter et al. (2016) developed a molecular system to distinguish between the *dura*, *pisifera* Congo (PisC) and *pisifera* Nigeria (PisN) origin oil palm Sh forms. An external primer pair called ShEx was designed in order to amplify the whole Sh region. Then, an allele specific primer ShDC was designed to amplify only *dura* or PisC genotypes, but not PisN genotypes. Finally, with the amplification product of the previous PCR a digestion was settled with *Hind III* enzyme which only digest *dura* genotypes. Following with the Sh gene, Babu et al. (2017) designed three Cleaved Amplified Polymorphic

Sequences (CAPS) markers to discriminate between the D, P and T forms. However in this study just PisC origin was taken into account. In Eo populations, Barcelos et al. (2002) used 37 complementary DNA (cDNA) probes that produced 248 RFLP polymorphic fragments for genetic diversity study of 241 Eo accessions in which four groups were identified. Zaki et al. (2010) developed 10 SSR markers from 1500 sequences and assessed the genetic diversity of germplasm collections from four South American countries. Some studies have also been carried out with Eo x Eg hybrids; for their genetic and phenotypic diversity study Arias et al. (2015) used 29 Eo x Eg genotypes based on 13 SSR. Montoya et al. (2013) and Singh et al. (2009) both constructed dense linkage maps in order to study quantitative trait loci (QTL) related to fatty acids in hybrids. In the first study a first generation (Eo x Eg) x Eg population was used and 364 SSR were amplified to construct the map. In the second study, AFLP, RFLP and SSR markers were used in an interspecific cross involving a Colombian Eo and a Nigerian Eg parent.

4.2. Next Generation Sequencing (NGS)

The development of Next Generation Sequencing (NGS) platforms has revolutionized genomic and transcriptomic methods since enormous amounts of data are created with cheap, fast and cost-effective practices so nowadays most approaches focus mainly on SNP genotyping. New genotyping methods, chemistries and platforms are constantly being developed to answer with great accuracy challenging questions such as recombination breakpoints for linkage mapping, Association Mapping assays (AM) or Genome-Wide Association Studies (GWAS) for complex traits, QTL mapping or quantifying rare transcripts without prior knowledge (Davey et al. 2011).

There are many assays for SNP typing and the decision as to which is most appropriate is a challenge, but the final choice mainly depends on the user´s capability and criteria. Here follows an introduction to the two different approaches used in this thesis: i) Sequence analysis of Candidate Gene (CG) amplicons explained with more detail in Chapter 3 and ii) Restriction site associated RNA Sequencing (RARSeq) presented in Chapter 4.

The CG approach relies on the study of already known regions of interest. These candidate genes can be identified by several sources; (i) literature searches related to known genes with proven influence on the characteristic of interest, (ii) exploration of relevant metabolic pathways, and (iii) analyses of published QTL and co-located transcripts with a relevant biological meaning. Oligonucleotides are designed to anneal to the regions of interest which are known to be associated with particular traits and then these regions are selected or

enriched prior to sequencing. Mosquera et al. (2016), for example, developed a targeted approach for the discovery of diagnostic SNP in potato for resistance to the late blight disease. In this study they designed primer pairs for nine candidate gene of the jasmonate biosynthesis pathway which is already known to be involved in the resistance of this disease. Li et al. (2016) also performed a CG driven approach in *Pinusradiate,* where a total of 209 CG for wood density and growth were selected from diverse sources.

In 2015 Alabady et al. (2015) for the first time described a RARSeq approach for population genomics and mapping analyses based on the combination of two well-known approaches; ddRADSeq or RADSeq and RNA-Seq. In RADSeq assays one or two enzymes are used for digestion coupled with posterior size selection of genomic regions adjacent to the restriction sites. Meanwhile, RNA-Seq allows sequencing whole transcriptomes in most of the populations and tissues and it is mainly used to measure gene expression. The novel RARSeq assay bypasses the main gaps which ddRADSeq and RNA-Seq face: in the first case the massive amount of data consisting largely of nongenic sequences and in the second, the presence of alternative transcripts which hinder inferring the genotype.

For the construction of RARSeq libraries total RNA of each genotype is extracted and messenger RNA (mRNA) is obtained using poly-A tail. Following this, single strand complementary DNA (cDNA) and afterwards double strand cDNA are synthesized. cDNA of each genotypes is then digested by two enzimes, one rare enzyme and one frequent, cutting restriction enzyme. This point can be critical and previous *in silico* assays are necessary to select the best enzymes that will best represent the transcriptome of interest. Once digestion is over, specific adapters are ligated to the restriction fragments, library is size selected, and gel slices are cleaned. Only targeted sequences will be sent for sequencing. In Figure 9 a scheme for RARSeq approach is displayed.

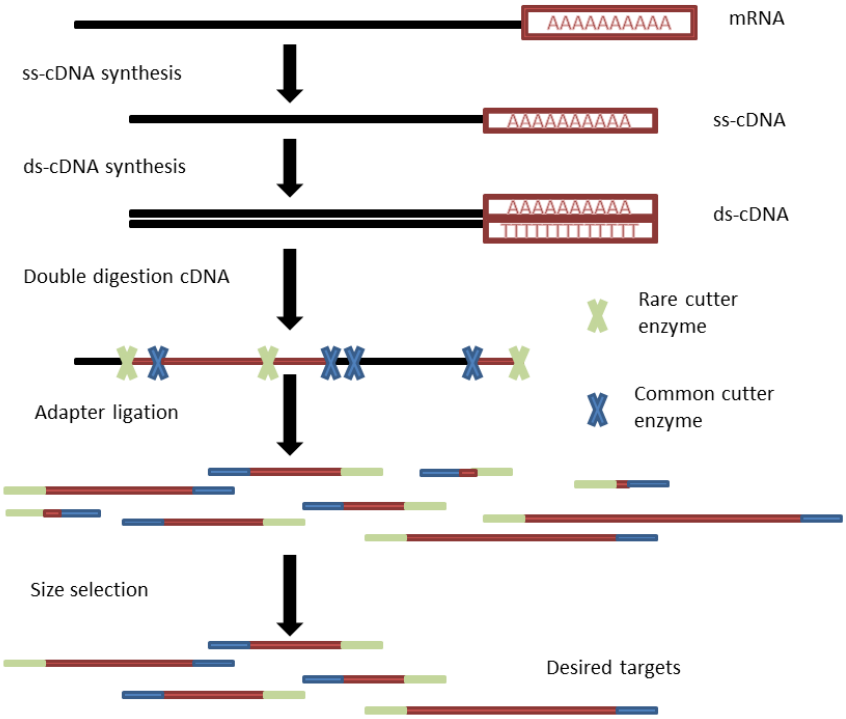


Figure 9: Schematic visualization of RARSeq steps.

While no publications are available for RARSeq assay in OP, a great many studies of RADSeq or RNA-Seq assays can be found for Eg. However no studies are available for Eo or Eo x Eg hybrids. Lei et al. (2014) and Somyong et al. (2018), for example, performed a RNA-Seq assay in order to study the different C-Repeat Binding Factor (CBF) under cold stress and to look for candidate genes involved in parthenocarpy, respectively. Using RADSeq Bai et al. (2018) constructed an ultrahigh-density linkage map for mapping quantitative trait loci (QTL) of important traits. Along the same lines, these authors (Bai et al. 2017) produced 16 linkage groups for selecting QTL for higher oil content in oil palms.

With respect to the chosen sequencing platforms, Thermo Fisher´s Ion Torrent™ Personal Genome Machine (PGM) was used for the CG approach and the Illumina MiSeq sequencer for the RARSeq approach.

PGM Ion Torrent™ chemistry relies on emulsion PCR. Protons are released when a dNTP is incorporated and the resulting change in the pH is detected by a complementary metal-oxide-semiconductor and an ion–sensitive field-effect-transistor. In each cycle only one type of dNTP is detected, however, more than one identical dNTP can be incorporated into the same cycle (Figure 10).

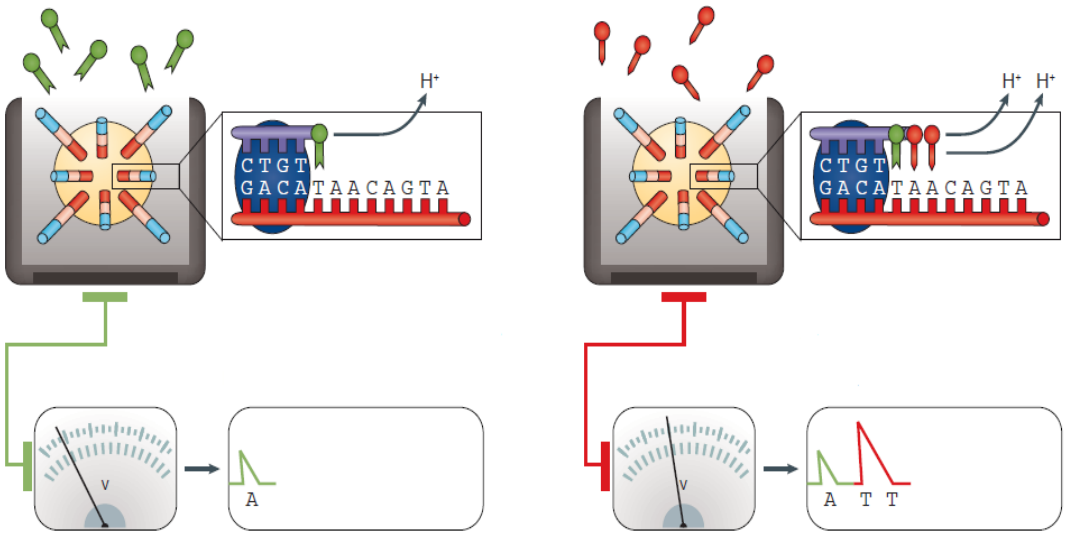


Figure 10: Ion Torrent sequencer chemistry (adapted from (Goodwin et al. 2016).

Illumina, by contrast, works with solid-phase amplification and identifies fluorophored dNTP based on total internal reflection fluorescence microscopy using either two or four laser channels (Goodwin et al. 2016). Once the image is detected, fluorophores are washed away and the 3´-OH group is regenerated allowing the start of a new cycle.

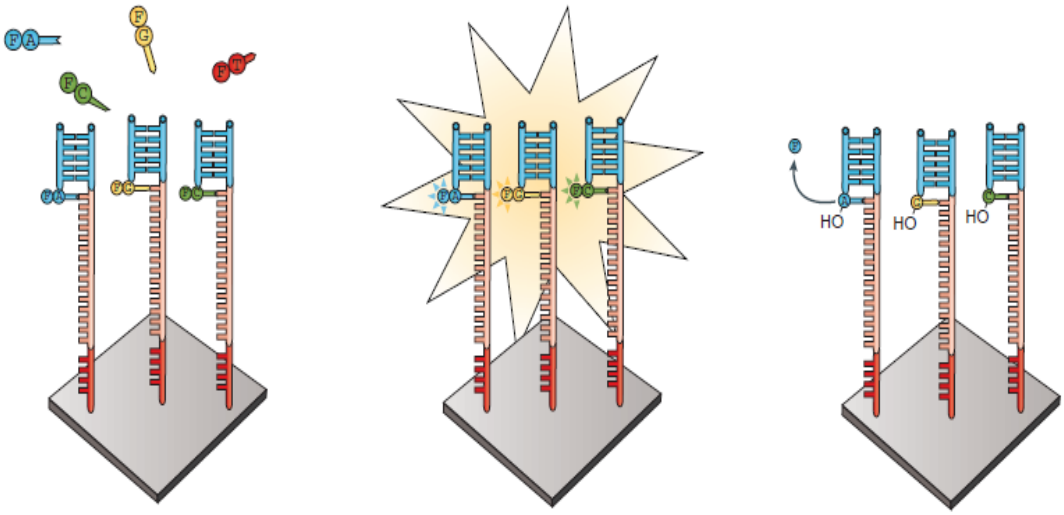


Figure 11: Illumina sequencer chemistry (adapted from (Goodwin et al. 2016).

4.3. Bioinformatics

In parallel with the advances in sequencing technologies, data analysis tools, methods and algorithms have been growing and becoming more and more important in enabling the interpretation of the large amount of information. Programming environments such as R (The R Development Core Team 2008) and Python (Python Software Foundation 2001) have become popular choices and software packages for these languages keep being released under open source licenses. Depending on the approach, many pipelines and tools can be found to suit each researcher's needs and web repositories such as GitHub (2019) compile a wide amount of pipelines and scripts for many genomic purposes.

Figure 12 shows a typical workflow for data analysis. Workflows with similar structures have been used to analyze our data in this thesis and the scheme presented could be extrapolated for similar approaches such as RADSeq or GBS. Each step of the workflow is explained and the different software modules are indicated (de Carvalho et al. 2019).

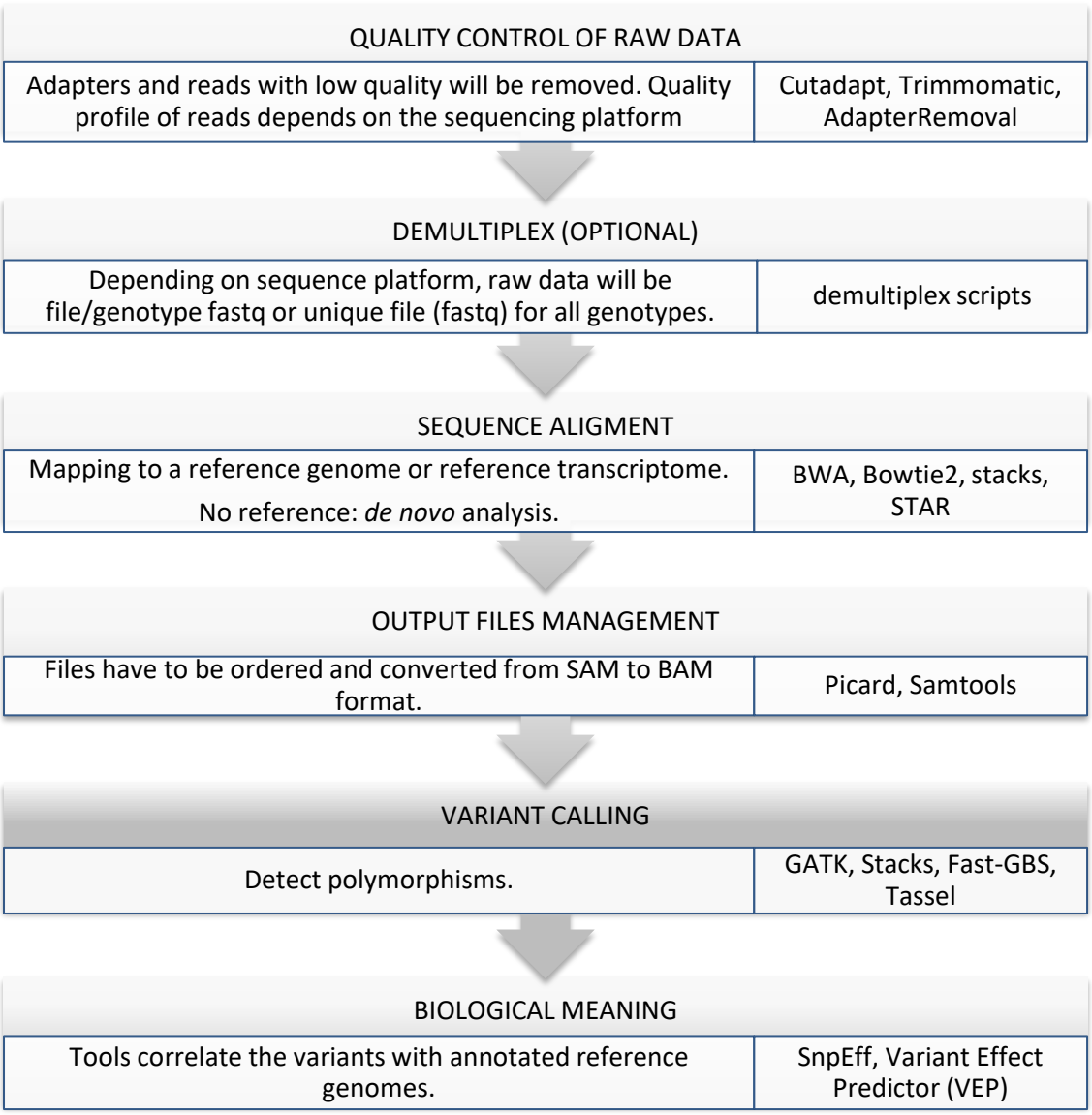


Figure 12: Data analysis workflow used to treat the reads from NGS libraries.

5. Association Mapping

Even though technology has evolved substantially, a good approach is still necessary in order to obtain reliable results. Most of the traits with environmental or agricultural value are quantitative in nature and influenced by more than one gene, by the environment, and by interactions between genes and environments. Characterizing and understanding these functional loci can help to improve breeding programs. The two main assays used for this approach are Linkage Mapping (LM) and Association Mapping (AM) studies. Linkage analysis is typically conducted in structured populations derived from a bi-parental cross in which the shared inheritance of functional polymorphisms markers within families is exploited (Yu and Buckler 2006). However, the mapping resolution of this assay can be limited by the low

number of recombinations for identifying the causative QTL, by the low number of polymorphisms or by the small population size (J. and Cloutier 2012). Furthermore, since these studies are performed in specific mapping populations, the results may not be valid for other genetic backgrounds. In order to deal with these gaps, AM studies have gained the attention of the scientific community. AM relies on the shared historical recombination accumulated in individuals of unobserved ancestry. The main advantages of this approach are the higher mapping resolution, higher allele number and time saving (Buckler and Thornsberry 2002). In the following Figure 13 a typical methodology for an AM study is shown.

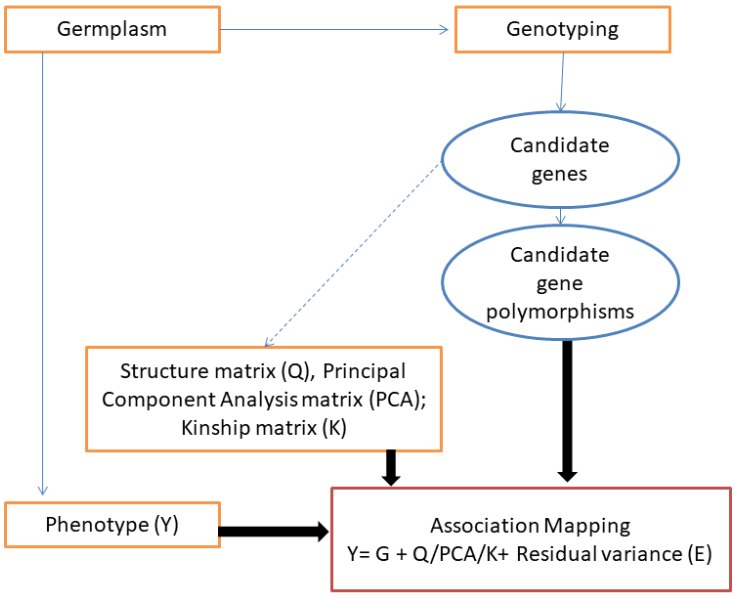


Figure 13: Schematic diagram of a typical methodology for an Association Mapping approach (adapted from (Zhu et al. 2008).

The bases of AM studies rely on Linkage Disequilibrium (LD) which represents a non-random association between two markers or two genes in a population. Therefore, the markers tested for association must be the causal variant for that trait or be in highly correlated LD to the causal marker (Hirschhorn and Daly 2005). While alleles in linkage equilibrium would segregate following Mendel's law (Castle 1903), alleles in LD will segregate as haplotypes within different probabilities. A representation of this can be seen in Figure 14.

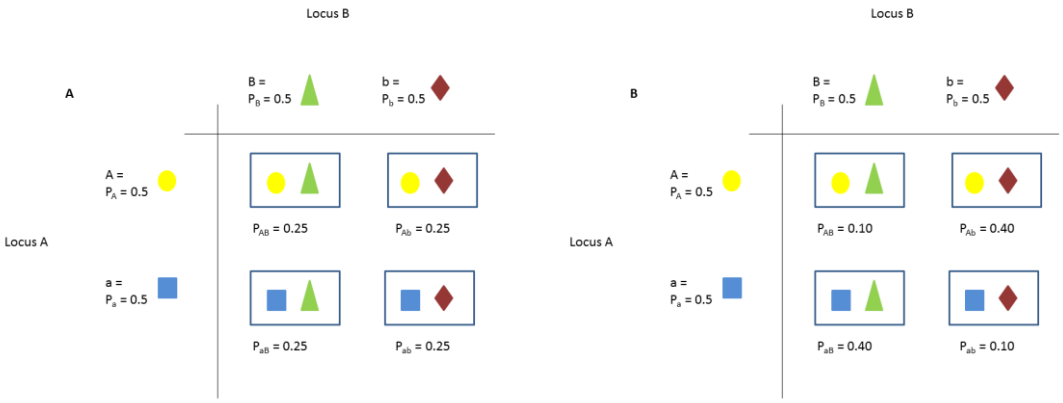


Figure 14: Linkage equilibrium (A) and Linkage disequilibrium (B).

The quantification of this value is commonly given in two statistics; *D* and r^2. The first explains the difference between the gametic frequencies of haplotypes observed and the expected gametic frequencies of haplotype under linkage equilibrium (J. and Cloutier 2012). It is given in the following way:

$$D = P_{AB} - P_A P_B \quad (1)$$

Equation 1: Quantification of Linkage Disequilibrium (LD) where P_{AB} is the frequency of haplotype AB and P_A and P_B are the frequencies of the A and B alleles, respectively.

Moreover, r^2 reflects the square correlation of gene frequencies in the sample summarizing both recombinational and mutational history (Sved and Hill 2018). Nowadays, r^2 is the most relevant LD measurement and is calculated as follows:

$$\boldsymbol{r^2 = \frac{D^2}{P_A P_a P_B P_b}} \quad (2)$$

Equation 2: Quantification of the correlation of gene frequencies, where D is the quantified LD value, and P_A, P_a, P_B and P_b are the frequencies of alleles A, a, B and b respectively.

Typically, LD is estimated plotting r^2 values from a data set against distance. 0.1 and 0.2 values of LD decay are usually used to describe LD decay (Zhu et al. 2008) and determine the coverage marker needed for an efficient mapping. For example, if LD decays rapidly a greater amount of markers will be needed to capture the causal marker or a marker in high LD to the causal marker related to the trait of interest. However, if LD extends to a great distance, fewer variants will be needed.

Together with LD, the number of genotypes, the genomic diversity, the relationship within families and an efficient and well collected phenotypic data will be key points for reliable and strong mapping.

Even though AM studies rely on the study of polymorphisms of unobserved ancestry some authors have pointed out the difficulty of replication of significant results in independent studies. Population stratification has emerged as the most serious systematic bias producing type I error (Pritchard and Rosenberg 1999; Marchini et al. 2004; Hirschhorn and Daly 2005). To confront this, statistical methods based on correction of population structure and familial relatedness have been proposed. In 2016 Teh et al. (2016a) performed a GWAS approach to identify key loci for high mesocarp oil content in OP. For this purpose, they adopted a compressed mixed linear model (MLM) approach in which they took into account population parameters based on a principal component analysis (PCA) and a group Kinship matrix (K). Also in 2016 and with OP, Kwong et al. (2016) performed a GWAS approach in which they studied the shell percentage in fruit (S/F %) in T genotypes. Since all genotypes were T form the Sh effect was removed from the analysis and another CG influencing that trait was targeted. In order to correct for the population structure, as in the previous study, MLM method was applied. Finally, recently, Babu et al. (2019) also performed a GWAS approach based on GBS libraries to find markers related to stem height increment in OP. In this case, two different models were tested; 1) general linear model (GLM) based method where structure matrix (Q) was used as covariate and 2) MLM method where K matrix was used in addition. Following these studies, GLM and MLM methods will be tested in this thesis.

6. General Objectives

This thesis was realized within the frame of a collaborative project between Neiker Tecnalia (Spain), La Fabril (Ecuador), Energy Palma (Ecuador) and Sampoerna Agro (Indonesia) with the main aim of finding markers related to oil production and quality in collection of *Elaeis oleifera*(Eo) accessions and hybrids of Eo and *Elaeis guineensis* (Eo x Eg) species. The discovery of such loci will help to produce elite seeds supported by marker assisted selection systems (MAS) for decreasing production costs as well as improving crop sustainability.

Therefore, the objective of this thesis is the discovery of candidate genes related to the principal production and quality traits that will be exploited in the near future in MAS programs. For that aim the specific objectives are the following:

1. Given the importance of the *Shell* gene (Sh), analysis of the Sh SNP in *E. oleifera* and interspecific hybrids since no information is available for Eo SNP.
2. Discovery of molecular markers related to oil production and oil quality traits based on the target analysis of Candidate Genes (CG) known to be involved in important metabolic pathways related to these traits.
3. Discovery of molecular markers related to oil production and oil quality traits based on the transcriptome analysis in a population of unobserved ancestry.

The hypothesis of this thesis is:

“Association mapping based on targeted and random candidate gene analysis allows the detection and identification of molecular markers related to traits of interest in interspecific *Elaeis oleifera* x *Elaeis guineensis* oil palm hybrids”

CHAPTER II: SHELL (Sh) GENE SCREENING AND ALLELE SPECIFIC PRIMER DESIGN

This chapter has been published:

Astorkia M, Hernandez M, Bocs S, Ponce K, León O, Morales S, Quezada N, Orellana F, Wendra F, Sembiring Z, Asmono D, Ritter E (2020). Analysis Of The Allelic Variation In The Shell Gene Homolog Of E. oleifera And Design Of Species Specific Shell Primers. Euphytica (DOI: 10.1007/s10681-019-2538-7).

CHAPTER II: SHELL (Sh) GENE SCREENING AND ALLELE SPECIFIC PRIMER (ASP) DESIGN

1. Introduction

Oil Palm is the crop with the highest oil yield per hectare. In 2018 the production of palm oil was 72.76 million tones according to USDA. Oil palm is the mayor source of vegetable oil with 35.6 % of all vegetable oil produced in the world. More than half of the oil production 49.82 million metric tons were used for food uses, while 18.05 million tones went to the industry (USDA 2018).

The main species for oil production is the African oil palm (*E. guineensis* Jacq. (Eg)), which is cultivated in three main areas of the tropics: Africa, South Asia and Central and South America (Corley, R.H.V. and Tinker et al. 2016). The African oil palm can grow to 15-18 m height, but almost to 30 m in deep forests. This species shows three different fruit types that are determined by one single gene known as "Shell thickness (Sh gene)" (Singh et al. 2013a). The *dura* fruit type is homozygous for the Sh gene allele Sh+, has a thick endocarp of about 2-8 mm and a thin mesocarp with high oil content. The *pisifera* fruit type is homozygous for the Sh gene allele Sh- with almost no endocarp and a thick mesocarp. However, this contains only a small quantity of oil. The third fruit type is *tenera* which is heterozygous for the Sh gene (Sh+ sh-), has a thin endocarp of about 0.5-4 mm and a thick mesocarp of 60-96% with high oil contents. Current oil palm varieties are all *tenera* since they have the highest oil production. They are obtained by crossing female *dura* parents with male *pisifera* parents, since *pisifera* are generally female sterile (Corley, R.H.V. and Tinker et al. 2016).

Another oil producing species of minor commercial interest is the American oil palm (*Elaeis oleifera* Kunth (Eo)*)* found in the tropics of Central and South America. This species has a low oil production, but has some favourable properties, such as a much better oil quality due to high unsaturated fatty acids contents (Pelaez et al. 2010), resistance to the main diseases affecting palms (Salavarrieta and Jesús 2004) and a shorter trunk. The morphological fruit characteristics of Eo resemble a *dura* fruit type. Therefore, the American oil palm is considered as an Eo *dura* type or Wild type (Montoya et al. 2014; Corley, R.H.V. and Tinker et al. 2016).

Central and South American Eg plantations are suffering from "Pudrición del Cogollo" disease, confronting important economic losses, since the affected palms will generally die (Sundram and Intan-Nur 2017). In highly infested areas Eg cultivation is almost impossible. This disease was first reported in the early twenties, but yet it is not clear what causes the disease. Some

experts point to *Phytophthora palmivora* as the causing agent (Torres et al. 2010), but the answer is still challenging.

In order to face this new scenario, different breeding companies have started to work with hybrids between the two *Elaeis* species using a *pisifera* genotype as pollen donor. These hybrids represent "*tenera*" genotypes and most important, they show resistance to the "Pudrición de Cogollo" disease which is inherited from the Eo parent (Amblard et al. 2004; Preciado et al. 2011). They have also other interesting characteristics, such as a decrease in height and improved oil quality. They show a higher percentage of oleic acid, as well as a lower content of saturated fatty acids, both interesting qualities from a nutritional point of view (Mozzon et al. 2013; Cadena et al. 2013; Corley, R.H.V. and Tinker et al. 2016). On the other hand, oil production is lower than in Eg, but still considerable. Twenty tons of fresh fruit bunches per hectare and year can be collected and the oil extraction rate is around 21% in ripe bunches (Torres et al. 2004). In compensation, since the height of these hybrids is around 50% lower than that of Eg, their productive life is longer and this crop can be harvested over a longer period in the plantation (Torres et al. 2004).

The *Sh*-thickness gene was identified by the Malaysian Palm Oil Board (MPOB) (Singh et al. 2013a) as a homologue of the MADS-box SEEDSTICK gene from *Arabidopsis*. The paper describes also the allelic variations between *dura* and two *pisifera* origins (Nigeria and Congo) in the first exon. Recently, a second paper published by MPOB determined three new *pisifera* events for Sh also in the first exon, showing different frequencies (Ooi et al. 2016). Two of them within a six amino acid stretch of the *pisifera* Nigeria event generating a lysine to glutamine and lysine to asparagine substitution, respectively. The third event is found ten amino acid substitutions away from the *pisifera* Congo event, resulting in an alanine to aspartate substitution. Details about the characteristics of the published allelic variation of the partial Sh gene in Eg are included in Table 7.

To facilitate the selection process of the desired plant material, fruit type specific primers have been developed. Ritter et al. (2016) presented primer pairs that distinguish between *dura*, *pisifera* Congo and *pisifera* Nigeria genotypes. Babu et al. (2017) presented CAPS marker (Cleaved amplified Polymorphic Sequences) that allow the differentiation between *dura*, *tenera* and *pisifera* Congo genotypes. However, these CAPS marker cannot distinguish between the different *pisifera* origins. Reyes et al. (2015) also developed allele specific primers in order to differentiate between the three fruit forms. However, no differentiation between *pisifera* origins is possible with these primers.

While several Sh studies are available for Eg, actually no publications on this topic exist for Eo. Therefore, the aim of the presented work was to study the allelic variation of the Sh gene in broader Eo germplasm and available interspecific hybrids that could be exploited for potential downstream applications.

2. Material and Methods

2.1. Plant material

Table 5: Plant material screened for allelic variability within a partial amplicon of the *Shell* gene (Sh).

E. oleifera

No.	Accession	No. Ac	Origin
1.1	O-PASTAZA	79	Ecuador
1.2	O-ERENE	17	Peru
1.3	O-MORONA	58	Ecuador
1.4	O-SERRA	14	Peru
1.5	O-CUCHILLO COCHA	24	Peru
1.6	O-SINU	13	Colombia
1.7	Taisha x Sinu (Eo x Eo)	4	Oleoflores, Colombia
	Subtotal:	**209**	

Interspecific hybrids (*Eo* x *Eg*)

No.	Accession	No. Ac	Origin
2.1	Coari x La Mé (Cabaña)	40	Hacienda La Cabaña, Colombia
2.2	Taisha x Avros (Oleoflores)	74	Oleoflores, Colombia
2.3	Taisha x Avros (RGS)	37	RGS, Ecuador
2.4	Taisha x Yangambi (RGS)	19	RGS, Ecuador
2.5	Taisha x Ekona (RGS)	25	RGS, Ecuador
2.6a	OxT_OL-Dura (from 7 families)	123	Oleoflores, Colombia
2.6b	OxT_ OL-Tenera (from 6 families)	9	Oleoflores, Colombia
	Subtotal:	**327**	

E. guineensis pisifera

No.	Accession	No. Ac	Origin
3.1	Ghana (4), Avros (3), Nigeria (3), Ekona (4)	14	BSM, Indonesia
3.2	TxT (Avros x DAMY-Las Flores); 3 families	18	Oleoflores, Colombia
	Subtotal:	**32**	
	TOTAL:	**568**	**accessions**

Legend:

No. Ac = Number of accessions, Eo = *E. oleifera*, Eg = *E. guineensis, T =tenera*

OxT_OL-Dura = Hybrids from Eo x Eg *tenera* crosses with one *E. oleifera* "*dura*" and one *E. guineensis dura* allele

OxT_OL-Tenera = Hybrids from *Eo* x Eg *tenera* crosses with one *E. oleifera* "*dura*" and one *E. guineensis pisifera* allele

Three different types of palms representing breeding material from Energy & Palma were screened for allelic variation of the Sh gene. In total 209 Eo genotypes, 327 interspecific hybrids and 32 Eg *pisifera* genotypes were analysed in this study. Characteristics and origins of the different accessions are shown in Table 5.

2.2. DNA extraction and library construction

DNA extractions were performed from young leaflet tissue samples using the Analytik JenaLife extraction kit (Science Products, Germany) according to the manufacturer instructions.

All PCR primers were designed using Primer3 software (Untergasser et al. 2012). All amplification products were visualized via gel-electrophoresis on a 1.5 % TAE agarose gel stained with GelRed® (Biotium).

An amplicon library of a partial Sh gene was constructed in the mentioned plant materials. Amplicons were generated in a 2-step PCR reaction as shown in the scheme in Figure 15, separately for each genotype.

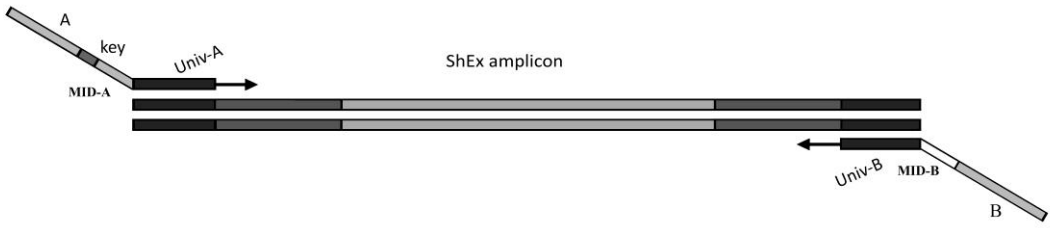


Figure 15: Scheme used for generating barcoded amplicons within the 1st Sh exon. See text and Table 6 for details.

For the first PCR reaction fusion primers were used which were composed of a universal part (UniA, UniB) and a part common to all Sh gene alleles published by Ritter et al. (2016) as ShEx primers (UniA_ShEx, UniB_ShEx, Table 6, No. 1a,b). These primers produced a 237 bp amplicon within the Sh gene.

20 ng of each genomic DNA, Invitrogen™ Platinum™ SuperFi™ PCR Master Mix (Life Technologies), and 0.16 µM primer-mix were used per 25 µl amplification reaction. The PCR conditions were as follows: 98 °C denaturation for 30s, followed by 30 cycles of 98 °C for 10 s, 58 °C for 30 s, 72 °C for 60 s, and a final elongation step of 72 °C for 5 minutes. PCR reactions

were performed in a Thermal Cycler ABI 2720 (Applied Biosystems, Foster, USA). Amplification products were visualized as described and the PCR products were purified using Agencourt AMPure XP(Beckman Coulter).

The purified PCR products were then used in a second PCR reaction to barcode each genotype. For this purpose fusion primers were designed which were composed of one part complementary to the universal part of UniA and UniB, a genotype specific MID part, the key part [ACGT] to calibrate the sequencing machine and the specific key sequences A and B used by the sequencing platform. These primers are shown in Table 6, No. 2a, b as well as the forward and reverse MID sequences in Table 6, No. 3A-P.

The combination of the MID sequences with UniA and UniB sequences, respectively, allow to identify unambiguously each genotype. By using a combination of forward and reverse MID a large number of genotypes can be barcoded with a relative small number of primers. With 2n MID primers n^2 genotypes can be discriminated.

Table 6: Primers and MID sequences used for generating barcoded amplicons within amplicon of the Sh gene and species specific primers (SSP). The Sh specific parts of the fusion primers are marked in bold and the universal UniA and UniB parts in *italics*.

No	Name	Primer Sequence	Ta(°C)	bp
1a	**UniA_ShEx**	Fw: *GCAAGACTCGAGCATCTCCA***GGATCGAGAACACCACAAGC**	58	277
1b	**UniB_ShEx**	Rv: *GCGATCGTCACTGTTCTCCA***AATTTGGCTTGGCCATAGAA**	58	
2a	**Barcode_ShA**	Fw: CATCTCATCCCTGCGTGTCTCCGA*CTCAG***[MID1]***GCAAGACTCGAGCA TCTCCA*	58	345
2b	**Barcode_ShB**	Rv: CCTCTCTATGGGCAGTCGGTGAT**[MID2]***GCGATCGTCACTGTTCTCCA*	58	

3 MID Primers and their reverse complements (rev com)

	MID (5'->3')	rev com		MID (5'->3')	rev com		MID (5'->3')	rev com
A	ACGCTCAG	CTGAGCGT	G	TATGCTAGA	TCTAGCATA	L	GCTATGACAG	CTGTCATAGC
B	TACATCAT	ATGATGTA	H	CGCACTGAG	CTCAGTGCG	M	ATACATAGCT	AGCTATGTAT
C	CGCGACTA	CTGAGCGT	I	CAGACTCTA	TAGAGTCTG	N	GACAGCGCGT	ACGCGCTGTC
D	AGCTAGTC	GACTAGCT	J	TGTGAGCAC	GTGCTCACA	O	CATGTCAGTA	TACTGACATG
E	CGAGATCA	TGATCTCG	K	GCGATAGTAC	GTACTATCGC	P	ACGCACTCGC	GCGAGTGCGT
F	TCAGTGCTG	CAGCACTGA						

4	**SSP Primers** *(1)*			
4a	**ShG**	Fw: TGACGCCTTCTCTTCCTT**C** Rv: TGTGATAATTTGAAAGGGTAATTTT	56	195
4b	**ShO**	Fw: TGACGCCTTCTCTCCCTT**T** Rv: GATCACTTGATCATTCCTTCCT	56	281

Legend: Ta (°C) = annealing temperature; bp = base pair of the amplification. Species specific primers ShG amplify only in *E. guineensis* genotypes, ShO only in *E. oleifera* genotypes. Both primer pairs amplify in interspecific hybrids.

For each barcoding reaction, a 25 µL reaction volume was prepared containing 1 µL of the purified PCR product, 0.2 µM forward and reverse barcoding primer and Invitrogen™ Platinum™ SuperFi™ PCR Master Mix. PCR reactions and visualization were performed as described before in the first step.

PCR products of each barcoded genotype were individually quantified with a Qubit 2.0 device, using the Qubit dsDNA HS assay (Life Technologies). Equal concentrations of genotype specific PCR products were mixed in one tube.

This pool was purified with columns using the GeneRead Size Selection Kit (Qiagen). The quality of the library was verified on an Agilent 2100 Bioanalyzer using DNA Chips with HS DNA Kit reagents according to the manufacturer's protocol (Agilent Technologies). The library was sent for sequencing to the Center for Applied Medical Research (CIMA, Spain), using the Ion Torrent Personal Genome Machine (PGM) with the 318 Chip. Sequencing was performed unidirectional.

2.3. Species specific PCR primers

Based on the obtained results, two primer pairs were designed which amplify either in only Eo and interspecific hybrid genotypes (ShO; Table 6, No. 4a) or in Eg and hybrid genotypes (ShG; Table 6, No. 4b). These two primer pairs were tested and validated in all genotypes.

The PCR reactions were performed in a final volume of 25 µl. Reactions contained 10X PCR Reaction buffer composed of 160 mM $(NH_4)_2SO_4$, 670 mM Tris–HCl pH 8.8, 0.1 % Tween-20, 25 mM $MgCl_2$, 200 µM of each dNTP, 0.08 µM of each forward and reverse primer, 0.5 U of Taq polymerase from Bioron (DNA Free Sensitive Taq DNA polymerase, BIORON GmbH, Germany), and 20 ng of genomic DNA. The PCR program consisted of an initial denaturation step at 94 °C for 5 min, followed by 33 cycles of denaturation at 94 °C for 30 s, primer annealing at 56 °C for 30 s and primer extension at 72 °C for 45 s, followed by a final elongation at 72 °C for 10 min. Amplification products were visualized as described.

2.4. Bioinformatics analyses

Analyses of the obtained sequences were performed using the South Green Bioinformatics Platform http://southgreen.cirad.fr/, (South Green Collaborators 2016), which provides different bioinfomatic tools and methods for sequence analysis.

In order to obtain clean Sh candidate gene amplicon sequences, trimming and demutliplexing steps were done. First, each genotype was identified by the combination of MIDs in each read. Sequences were separated in genotype specific files. For this purpose the public

"demultiplex.py" script (Flutre et al. 2017) was used. Then, the "Cutadapt trimming tool" (Martin 2011) was applied to remove universal primer parts (UniA, UniB) and the MID. Once all sequences were clean and assigned to the corresponding genotypes, double demultiplexing was performed searching for the first 10 nucleotides of the ShEX forward and reverse primers with the same "demultiplex.py" script.

Finally, the "Snakemake" script (Soriano et al. 2018) of the South Green bioinformatics platform (South Green Collaborators 2016) was used to map the reads using BWA (Li and Durbin 2010), clean the alignments with Samtools (Li et al. 2009), sort the reads with Picard-tools (Broad Institue 2015) and to call the SNP using GATK haplotype caller (McKenna et al. 2010). The MPOB Eg *pisifera* genome sequence (Singh et al. 2013b), was used as reference. Reads with a quality score below 10 were discarded using VCFTools (Danecek et al. 2011) in order to avoid bad quality reads derived from sequencing errors. The allelic variants were visualized with IGV software (Robinson et al. 2011).

2.5. Trait recording and Phenotypic Data analyses

For the Eo palms and the interspecific hybrids fruit weight (FW) and fruit component data were available. This latter considered kernel to fruit (KF), shell to fruit (SF) and mesocarp to fruit (MF) ratios as percentages. Representative portions of fruits from each palm were collected, following standard bunch analysis procedures in oil palm (Babu 2008). Each fruit was weighted, mesocarp, shell and kernel were separated manually, the components were weighted and the corresponding ratios were calculated.

Saphiro Wilk tests were applied in order to check for non-normal distributed data. The traits that showed a significant deviation were normalized by z-score correction and the normalized data were used for further processing.

Analyses of variances of the traits were performed using SAS procedure Proc GLM (SAS Institute Inc.). For the Eo accessions the model considered origin (Ori) and allele combinations (AC) as main effects, as well as their interactions (Ori *AC). For the hybrids two different models were applied. Initially, the effects of the Eo Sh allele and the Eg Sh allele were separated and used as main effects together with Ori. Also the interactions between Ori and Sh alleles were included in the model (Ori*Eo, Ori*Eg).

In addition, the same model was applied as for the Eo accessions, considering Ori and AC as main effects and their interaction.

Separation of means for traits with significant effects was performed always using Duncan tests.

3. Results

3.1. Sequence analysis for SNP detection and defining events

Three types of plant material have been screened for analysing the allelic variation within a partial Exon 1 and adjacent intron amplicon of the Sh gene: (i) a total of 209 Eo accessions from seven origins of Peru, Ecuador and Colombia, including two intraspecific hybrids (Taisha x Sinu) and (ii) interspecific hybrids from 6 different origins (327 accessions in total). Five of them were obtained by crossing Eo accessions with different *pisifera* pollen donors and one origin from crosses between Eo (Pastaza) and *tenera* (Avros), separated in the resulting *dura* and *tenera* genotypes (2.6 a,b in Table 5). In addition, (iii) a total of 32 *pisifera* genotypes from 2 sources were screened; *pisifera* derived from four origins (3.1) and *pisifera* derived from three families of TxT crosses (3.2). In total 568 accessions were evaluated.

It is worth to notice that other Eo origins are targeted indirectly in the hybrids, such as Coari from Brazil and Taisha from Peru, as well as Eg *dura* from the OxT (*tenera*) crosses.

The sequence of the 237 bp amplicon within the Sh gene is shown in Table 7, separated in several parts. This amplicon implies the well-known Exon 1 of the Sh gene and part of the adjacent intron sequence. They correspond to the sequence stretch from 3.077.982 to 3.078.218 of the MPOB reference genome, chromosome 2. Even though mapping of the sequences was done in EG5 version after running blast into the last EG5.1 version coordinates for the Sh amplicon were the same.

After cleaning and filtering the sequences as described above, around 23000 reads for the partial Sh gene were available in the mentioned plant materials. A total of 7 SNP were detected by the "Snakemake" workflow. Table 7 shows the detected allelic variation of the partial Sh amplicon in the screened plant materials. With respect to Eg events, we observed in our plant material the already known *dura*, *pisifera* Congo (PisC), *pisifera* Nigeria (PisN) and MPOB3 events. They are defined by three SNP at relative bp positions 58, 65 and 95 in the amplicon, respectively, and imply the nucleotide changes [T/C], [A/T] and [C/A] with respect to the *dura* reference sequence. The resulting amino acid (AA) changes are leucine (L) to proline (P) in *pisifera* Nigeria, lysine (K) to asparagine (N) in *pisifera* Congo and alanine (A) to aspartic acid (D) for MPOB3. No MPOB2 or MPOB4 events were found in our samples. All Eo sequences were identical to the *dura* sequences in this region.

With respect to the Eo sequences, we found in the screened material four SNP further downstream of the Eg SNP in the intronic part. They are located at relative bp positions 165

(NK1), 184 (NK2), 188 (NK3a) and 192 (NK3b), respectively. The SNP NK1 was defined by a nucleotide change [A/G], NK2 by [C/T], NK3a by [C/G] and NK3b by [T/C] changes. The SNP screening of the analysed plant material revealed that NK2 appeared in all Eo accessions and in the interspecific hybrids, but was absent in all Eg accessions, indicating that this SNP is characteristic for Eo, at least in the analysed materials. NK3a and NK3b co-segregated in all cases and can be considered as a "double" SNP (=NK3). The SNP NK1 and NK3 segregated independently, could be both absent, but appeared never together.

Based on these findings three specific Eo events can be defined: (i) OLI1, where only the common Eo SNP NK2 occurs, (ii) OLI2 where in addition the NK1 SNP is present and (iii) OLI3 where beside NK2 also NK3a,b were present. All Eg sequences had the same sequence as the *dura* reference sequence in this region.

Table 7: Detected allelic variation in the amplicon of the Sh gene in the screened plant materials and resulting events.

		***E. guineensis* SNP**				***E. oleifera* SNP**		
	SNP	**PisN(*58*) PisC(*65*)**			**MPOB3(*95*)**	**NK1(*165*)**	**NK2(*184*)**	**NK3a(188) NK3b(*192*)**
	Pos	57	67	77	87	165	175	185
		|	|	|	|	|	|	|
	Dura	**CTGAAGAAAG**	**CTTATGAGTT**	**GTCTGTCCTT**	**TGTGATGCTG**	**ATGACGCCTT**	**CTCTTCCTTC**	**GCTCATATCA**
Eg events	**PisN**	C**C**GAAGAAAG	CTTATGAGTT	GTCTGTCCTT	TGTGATGCTG	ATGACGCCTT	CTCTTCCTTC	GCTCATATCA
	PisC	CTGAAGAA**T**G	CTTATGAGTT	GTCTGTCCTT	TGTGATGCTG	ATGACGCCTT	CTCTTCCTTC	GCTCATATCA
	MPOB3	CTGAAGAAAG	CTTATGAGTT	GTCTGTCCTT	TGTGATG**A**TG	ATGACGCCTT	CTCTTCCTTC	GCTCATATCA
Eo events	**Oli1**	CTGAAGAAAG	CTTATGAGTT	GTCTGTCCTT	TGTGATGCTG	ATGACGCCTT	CTCTTCCTT**T**	GCTCATATCA
	Oli2	CTGAAGAAAG	CTTATGAGTT	GTCTGTCCTT	TGTGATGCTG	**G**TGACGCCTT	CTCTTCCTT**T**	GCTCATATCA
	Oli3	CTGAAGAAAG	CTTATGAGTT	GTCTGTCCTT	TGTGATGCTG	ATGACGCCTT	CTCTTCCTT**T**	GCT**G**ATA**C**CA
Amino Acids	**Ref AA**	**-L-—K-—K—-**	**A—-Y—-E—-L**	**—-S—-V-—L—**	**-C—-D-—A—-**			
	Alt AA	-**P**-—K-—**N**—-	A**—-**Y-—E—-L	**—-**S—-V-—L—	-C—-D-—**D**—-			

Common amplicon sequences:

C1 <1-56 bp> GGATCGAGAACACCACAAGCCGGCAGGTCACTTTCTGCAAACGCCGAAATGGACTG

C2 <97-164 bp> AGGTTGCCCTTATTGTCTTCTCCAGCCGGGGCCGCCTCTAAATAACAGGTATGCTTTG

C3 <195-237 bp AAGTTAATTTTATGGCTTCATTTGTTCTATGGCCAAGCCAAATT

3.2. Distribution of *Shell* events in the analysed plant materials

Table 8 shows the frequencies of distribution of the different Eg and Eo events in the analysed *Elaeis* germplasm. The OLI2 event was absent in all Peruvian Eo accessions (Erene, Serra, Cuchillo Cocha). In these genotypes always OLI1 and OLI3 events were detected. In Pastaza, Morona and Sinu origins only OLI2 events were found. In Taisha x Sinu accessions OLI1 and OLI2 events were detected, suggesting that the Taisha origin contains at least OLI1 and OLI2 events.

Table 8: Frequencies and distribution of the different *E. guineensis* and *E. oleifera* events in the analysed *Elaeis* germplasm.

A)	***E. oleifera***			**Allele composition of the Sh locus**				
No.	**Origin and observed events**	**No. Ac**	**MV**	**OLI1 OLI1**	**OLI2 OLI2**	**OLI3 OLI3**	**OLI1 OLI2**	**OLI1 OLI3**
1.1	O-PASTAZA - (OLI2)	79	3		76			
1.2	O-ERENE - (OLI1,3)	17	1			10		6
1.3	O-MORONA - (OLI2)	58	7		51			
1.4	O-SERRA - (OLI1,3)	14	1	4		4		5
1.5	O-CUCHILLO COCHA – (OLI1,3)	24		5		7		12
1.6	O-SINU - (OLI2)	13			13			
1.7	Taisha x Sinu (Eo x Eo) - (OLI1,2)	4			2		2	
	Subtotal:	**209**	**9**	**9**	**142**	**24**	**2**	**23**

B)	**Interspecific hybrids**									
No.	**Accessions and observed *E. oleifera* events**	**No. Ac**	**MV**	**OLI1 MPOB 3**	**OLI2 PisC**	**OLI2 PisN**	**OLI2 MPOB3**	**OLI3 PisC**	**OLI3 MPOB 3**	**OLI2 DURA**
2.1	Coari x La Mé (Cabaña) – OLI1,2,3	40	5	3	13				17	2
2.2	Taisha x Avros (Oleoflores) – OLI2	74	4		59	7				4
2.3	Taisha x Avros (RGS) – OLI1,2,3	37	2	5	4	10		1	10	5
2.4	Taisha x Yangambi (RGS) – OLI2	19	1		11					7
2.5	Taisha x Ekona (RGS) – OLI2	25	2		12	11				
2.6a	OxT_OL-Dura	123	6		3	1	6	1		106
2.6b	OxT_ OL-Tenera	9	1		5	1	2			
	Subtotal:	**327**	**21**	**8**	**107**	**30**	**8**	**2**	**27**	**124**

C)	***E. guineensis pisifera***					
No.	**Accessions**	**No. Ac**	**MV**	**PisC PisC**	**PisN PisN**	**PisC PisN**
3.1	Ghana (4), Avros (3), Nigeria (3), Ekona (4)	14	2	3	6	3
3.2	TxT (Avros x DAMY-Las Flores); 3 families	18	1	16	1	
	Subtotal:	**32**	**3**	**19**	**7**	**3**

Legend: No. Ac = number of accessions; MV = accessions with missing values

With respect to the interspecific hybrids we found besides the expected "*tenera*" genotypes with one Eo "*dura*" allele, in 4 of 5 origins also true *dura* genotypes with one Eg *dura* allele reaching in the Taisha x Yangambi genotypes over 35 %. This is apparently due to pollination errors with *tenera* or contamination of the *pisifera* pollen with *tenera* pollen.

Independent of this, in the Coarí x La Mé accessions we found OLI1, 2 and 3 events as well as the PisC and MPOB3 events. In the two Taisha x Avros origin all Eo and *pisifera* events were detected. In the Taisha x Yangambi and Taisha x Ekona hybrids only OLI2 events occurred. In the first case only PisC events were present, while the Ekona parents in the second case contributed PisC and PisN alleles.

In the screened Eg *pisifera* accessions we detected in the *pisifera* from BSM three genotypes which were homozygous for PisN and PisC, respectively, and three which were heterozygous for these events. In Ekona and Nigeria origins from BSM we found the homozygous genotypes for PisN, while Avros genotypes showed homozygous PisC events. All Ghana genotypes were heterozygous for PisN and PisC (results not shown).

In the *pisifera* derived from the TxT crosses all were homozygous, 16 for the PisC event and one for PisN event indicating that the parents of each progeny genotypes had the same *pisifera* allele. No MPOB3 event was detected in all analysed *pisifera*.

3.3. SSP primer validation

The particular SNP NK2 which appears in all Eo events and consequently in all genotypes which have Eo Sh alleles (Eo and interspecific hybrids) and are absent in all pure Eg events was used to develop species specific primers (SSP) which discriminate between Eo and Eg alleles. Their sequences are also integrated in Table 6.

The SSP designed for each species were validated in all accessions of Table 5. All genotypes showed the expected results. ShG primers only amplified in Eg and hybrids, while ShO primers amplified in Eo and in hybrids. Figure 16 shows one example of the visualized amplification products for each type of plant material of Table 5.

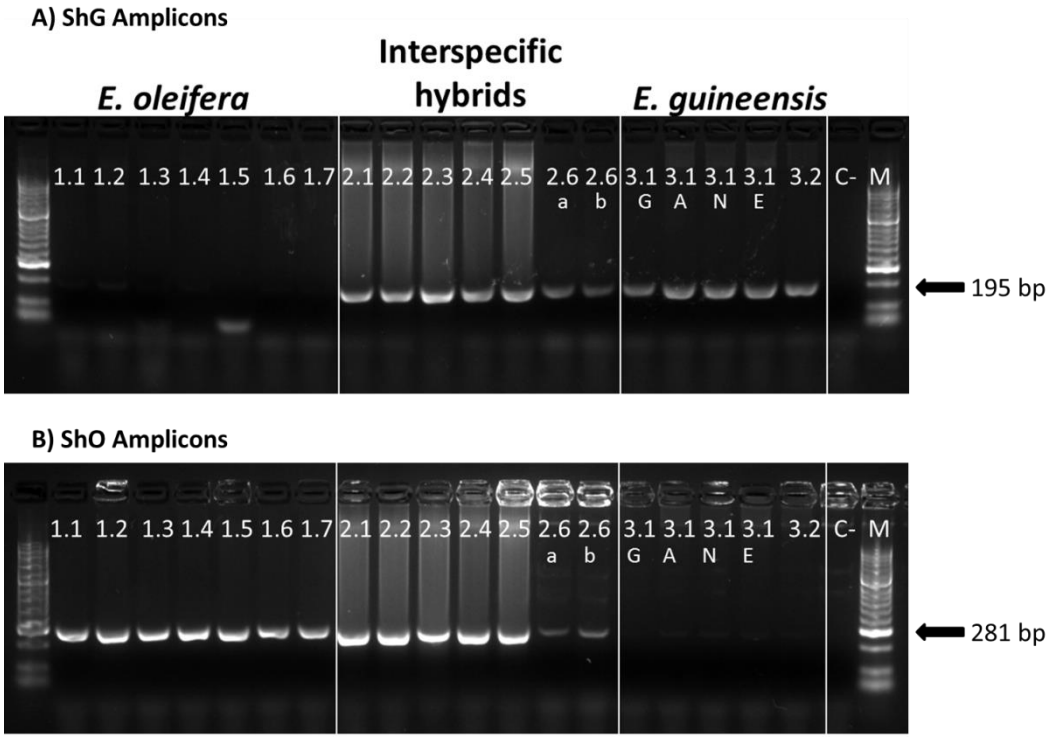


Figure 16: Examples for amplification products of ShG and ShO primers in one sample of each type of screened plant material (see Table 5). G=Ghana, A=Avros, N=Nigeria, E=Ekona accession; M=ladder.

3.4. Phenotypic Data Analyses

All fruit component data were not normally distributed and had to be transformed for further analyses.

The results of the analyses of variances (ANOVA) in Eo accessions are summarized in Table 9A. Identical results would have been obtained for models assuming a nested design (allele combination (AC) within origins (Ori)) instead of the interaction (Ori*AC).

For fruit weight (FW) only an effect of the Ori was detected and for kernel to fruit ratio (KF) only a significant Ori*AC interaction. Shell to fruit ratio (SF) and mesocarp to fruit ratio (MF) showed both a significant effect of the Ori and a significant Ori*AC interaction, but no AC effects.

Table 10 presents the mean values and coefficients of variations for FW and fruit components in the Eo accessions and integrates the separation of means results for significant effects. Considering FW, Sinu and Taisha x Sinu (TxS) accessions revealed significant lower FW than all other accessions. Erene, Serra and Cuchillo Cocha (CC) accessions showed the significant highest FW.

With respect to KF, the significant highest ratio occurs in Serra OLI1/OLI1 genotypes and the lowest in Erene homozygous OLI3 and Pastaza homozygous OLI2 accessions.

Table 9: Summary of the analysis of variance results for A) E. oleifera accessions and B) interspecific hybrid accessions.

A) Eo accessions		**Fruit Weight (FW)**		**Kernel to fruit (KF)**		**Shell to fruit (SF)**		**Mesocarp to fruit (MF)**	
Source	DF	SS III	Pr > F	SS III	Pr > F	SS III	Pr > F	SS III	Pr > F
Ori	4	202.653	**<.0001**	5.707	0.1601	6.925	**0.0184**	8.138	**0.0068**
AC	2	1.564	0.7691	2.129	0.2910	2.513	0.1120	0.842	0.4696
Ori*AC	3	3.907	0.7261	11.863	**0.0040**	14.619	**<.0001**	5.834	**0.0167**
Error	165	654.134		154.300		149.943		159.186	
Total	174	862.258		174.000		174.000		174.000	
B1) Hybrid accessions		**Fruit Weight (FW)**		**Kernel to fruit (KF)**		**Shell to fruit (SF)**		**Mesocarp to fruit (MF)**	
Source	DF	SS III	Pr > F	SS III	Pr > F	SS III	Pr > F	SS III	Pr > F
Ori	4	80.405	**0.0030**	0.001	0.7947	0.431	**<.0001**	0.449	**<.0001**
Eo	1	13.953	0.0900	<.0001	0.9848	0.007	0.2749	0.006	0.3557
Eg	2	47.173	**0.0085**	0.003	0.1346	0.204	**<.0001**	0.161	**<.0001**
Ori*Eo	1	9.146	0.1691	<.0001	0.7240	0.008	0.2280	0.010	0.2569
Ori*Eg	4	35.163	0.1251	0.007	**0.0468**	0.273	**<.0001**	0.263	**<.0001**
Error	146	765.877		0.102		0.751		1.070	
Total	158	951.717		0.113		1.674		1.959	
B2) Hybrid accessions		**Fruit Weight (FW)**		**Kernel to fruit (KF)**		**Shell to fruit (SF)**		**Mesocarp to fruit (MF)**	
Source	DF	SS III	Pr > F	SS III	Pr > F	SS III	Pr > F	SS III	Pr > F
Ori	4	80.302	**0.0031**	0.001	0.842	0.426	**<.0001**	0.443	**<.0001**
AC	4	58.380	**0.0190**	0.003	0.3434	0.213	**<.0001**	0.171	**0.0003**
Ori*AC	6	44.310	0.1684	0.007	0.1303	0.281	**<.0001**	0.274	**<.0001**
Error	144	768.725		0.102		0.750		1.071	
Total	158	951.717		0.113		1.674		1.959	

Legend: Source: source of variation; Ori: origin; AC: allele combinations; DF: degrees of freedom; SS III: Type III sum of squares; Pr > F: effect of the classification variable on the response.

Considering SF the significant highest average values can be found in CC accessions and the lowest in Morona, Pastaza and TxS genotypes. Accordingly, the significant lowest and highest values can be found in the corresponding combined Ori*AC data. There is only one AC for Pastaza, Morona and TxS, but for CC the highest value occurs for CC OLI1/OLI1 genotypes. In addition, a significant low value can be found in Serra genotypes with the same allele combination.

For MF significant lowest and highest values are exactly inversed compared to SF; the highest values becoming the lowest and vice versa. One exception is the homozygous OLI1 genotypes in the Serra accessions.

Table 10: Mean values and coefficients of variation for fruit weight and fruit components in the E. oleifera accessions.

AC/Origin	PASTAZA	ERENE	MORONA	SERRA	C. COCHA	SINU	TXS	Mean	CV [%]
A) Fruit Weight (FW)									
OLI1/OLI1				12.51	11.37			**11.80**	14.8
OLI2/OLI2	10.97		9.29			7.03		**10.01**	20.2
OLI3/OLI3		12.91		12.37	11.71			**12.28**	17.3
OLI1/OLI2							6.02	**6.02**	0.7
OLI1/OLI3		12.03		11.60	11.96			**11.92**	16.8
Mean	**10.97 AB**	**12.39 A**	**9.29 B**	**12.16 A**	**11.74 A**	**7.03 C**	**6.02 C**	**10.45**	21.3
CV [%]	15.4	20.4	17.7	17.1	13.5	11.5	0.7	21.3	
B) Kernel to fruit (KF)									
OLI1/OLI1				11.45 A	7.38 BC			**8.91**	66.0
OLI2/OLI2	6.59 C		7.01 BC			8.54 ABC		**6.92**	31.2
OLI3/OLI3		6.52 C		7.27 BC	10.69 AB			**8.49**	44.4
OLI1/OLI2							8.65 ABC	**8.65**	19.1
OLI1/OLI3		10.27 ABC		10.56 AB	8.38 ABC			**9.37**	41.9
Mean	**6.59**	**8.51**	**7.01**	**9.76**	**8.91**	**8.54**	**8.65**	**7.42**	39.1
CV [%]	30.4	53.5	33.8	54.8	40.3	16.6	19.1	39.1	
C) Shell to fruit (SF)									
OLI1/OLI1				32.52 E	50.15 A			**43.54**	26.9
OLI2/OLI2	31.67 E		31.25 E			35.09 CDE		**31.83**	14.4
OLI3/OLI3		44.98 AB		45.96 AB	40.46 BCD			**43.19**	19.6
OLI1/OLI2							32.81 E	**32.81**	25.5
OLI1/OLI3		34.48 DE		41.91 BCD	42.50 BC			**39.73**	20.7
Mean	**31.67 C**	**39.14 AB**	**31.25 C**	**40.13 AB**	**43.64 A**	**35.09 BC**	**32.81 C**	**34.23**	21.3
CV [%]	12.1	26.8	17.7	22.7	18.4	9.7	25.5	21.3	
D) Mesocarp to fruit (MF)									
OLI1/OLI1				56.02 ABC	42.47 E			**47.55**	27.3
OLI2/OLI2	62.26 A		61.74 A			56.37 AB		**61.25**	8.2
OLI3/OLI3		48.50 BCDE		46.77 DE	48.84 BCDE			**48.32**	16.2
OLI1/OLI2							58.55 A	**58.55**	11.5
OLI1/OLI3		55.25 ABCD		47.53 CDE	49.13 BCDE			**50.90**	17.2
Mean	**62.26 A**	**52.35 BC**	**61.74 A**	**50.11 C**	**47.45 C**	**56.37 AB**	**58.55 A**	**58.35**	13.9
CV [%]	9.7	16.6	9.6	21.9	18.4	7.7	11.5	13.9	
Overall CV [%]	**23.9**								

*Means with the same letter are not statistically different ($\alpha > 0.05$). Legend: CV: relative coefficients of variation [%]; AC: allele combination; C. COCHA: Cuchillo Cocha; TxS: Taisha x Sinu.

Coefficient of variation (CV) values showed large variations depending on trait, Ori and AC. The average CV value for traits ranged from 14 % in MF to 39 % in KF. With respect to the AC across traits the minimum value was 0.7 % for FW in OLI1/OLI2 genotypes and the maximum value of 66% was observed for KF in homozygous OLI1 genotypes. With respect to the origins across traits the minimum value was also 0.7 % for FW in TxS accessions, since only one AC occurs in this origin. The maximum value of 55 % was observed for KF in Serra accessions.

The results of the ANOVA test in interspecific hybrids considering the model which separates Sh gene allele effects, Eo and Eg, are summarized in Table 9-B1. Since in several cases specific allele combinations are linked to specific origins, the design is quite unbalanced and it was not possible to compute Eo*Eg interactions, neither the triple interaction Ori*Eo*Eg. The corresponding sum of squares (SS) was zero.

For all four traits no effect of the Eo Sh allele was detected, neither significant interactions Ori*Eo. For FW significant effects of hybrid Ori and Eg were observed without interaction. For KF only a slightly significant interaction Ori*Eg was detected. Both, SF and MF revealed highly significant Ori and Eg effects, as well as significant Ori*Eg interactions.

This model was chosen particularly for analysing potential effects of the Eo Sh alleles on the different traits in interspecific hybrids. However, in order to reduce complexity, to improve the balance of the design and for comparison with the Eo accessions, ANOVA were performed also with models considering besides Ori only the AC and Ori*AC interactions. The results are summarized in Table 9-B2.

For FW significant effects of hybrid Ori and AC were detected, but no Ori*AC interaction effect, and for KF no significant effects were observed. SF and MF showed both significant effects of the Ori and AC, as well as, significant Ori*AC interactions.

Table 11 presents the mean values and coefficients of variations for FW and fruit components in interspecific hybrids, with integrated separation of means results.

Table 11: Mean values and coefficients of variation for fruit weight and fruit components in the interspecific hybrid accessions.

AC/Origin	CXL	TXA(O)	TXA(RGS)	TXY	TXE	Mean	CV [%]
A) Fruit Weight (FW)							
OLI1/MPOB3	10.12		11.86			**11.21 A**	20.4
OLI2/ PisC	9.77	11.68	8.83	10.81	11.67	**11.21 A**	19.7
OLI2/ PisN		11.07	10.09		14.33	**12.05 A**	26.0
OLI2/ DURA		12.34	12.49	12.77		**12.59 A**	20.0
OLI3/MPOB3	9.82		9.02			**9.54 B**	18.6
Mean	**9.83 C**	**11.66 B**	**10.19 C**	**11.61 B**	**13.16 A**	**11.61**	22.2
CV [%]	23.0	16.9	24.6	21.2	20.6	21.2	
B) Kernel to fruit (KF)							
OLI1/MPOB3	7.25		7.32			**7.30**	34.3
OLI2/ PisC	8.27	7.02	8.56	7.07	8.61	**7.38**	38.7
OLI2/ PisN		7.96	7.81		5.94	**7.13**	34.7
OLI2/ DURA		5.62	4.30	8.17		**6.62**	41.4
OLI3/MPOB3	7.67		6.95			**7.41**	31.4
Mean	**7.87**	**7.03**	**7.13**	**7.52**	**7.11**	**7.28**	36.8
CV [%]	43.7	35.8	36.2	26.3	35.2	36.8	
C) Shell to fruit (SF)							
OLI1/MPOB3	19.50 CD		26.99 BC			**24.18 B**	37.2
OLI2/ PisC	15.14 DEF	13.41 DEF	11.14 DEF	15.54 DEF	17.00 DEF	**14.11 C**	47.1
OLI2/ PisN		9.31 EF	17.82 CDE		31.47 AB	**20.57 B**	55.4
OLI2/ DURA		7.96 F	34.65 AB	40.06 A		**29.73 A**	58.5
OLI3/MPOB3	19.86 CD		19.78 CD			**19.83 B**	42.1
Mean	**17.91 B**	**12.66 C**	**21.30 AB**	**25.64 A**	**25.14 A**	**17.83**	57.7
CV [%]	45.9	51.6	51.0	53.7	42.5	57.7	
D) Mesocarp to fruit (MF)							
OLI1/MPOB3	73.25 BCD		65.69 CDE			**68.52 BC**	12.9
OLI2/ PisC	76.59 ABC	79.57 AB	80.30 AB	77.39 AB	74.39 BC	**78.51 A**	10.6
OLI2/ PisN		82.73 AB	74.36 BC		62.59 DE	**72.30 B**	15.1
OLI2/ DURA		86.42 A	61.04 EF	51.76 F		**63.65 C**	29.4
OLI3/MPOB3	72.48 BCD		73.27 BCD			**72.76 AB**	13.3
Mean	**74.22 B**	**80.31 A**	**71.57 BC**	**66.84 C**	**67.75 C**	**74.90**	14.9
CV [%]	13.1	10.1	15.4	21.7	15.1	14.9	
Overall CV [%]	**32.8**						

*Means with the same letter are not statistically different ($\alpha > 0.05$). Legend: CV: relative coefficients of variation [%]; AC: allele combination; CxL: Coari x La Mé (Cabaña); TxA(O): Taisha x Avros (Oleoflores); TxA(RGS): Taisha x Avros (RGS); TxY: Taisha x Yangambi (RGS); TxE: Taisha x Ekona (RGS).

The significantly highest average FW were detected in Taisha x Ekona accessions and the lowest in Coari x La Mé and Taisha x Avros (Oleoflores) accessions. Considering AC across origins, the significant lowest FW were found for OLI3/MPOB3 combinations present in two hybrid origins, while all other AC revealed significant higher values without differences among them. With respect to SF the significantly highest values were detected for Taisha x Yangambi and Taisha x Ekona accessions and the lowest for Taisha x Avros (Oleoflores). For AC across origins the highest value occurred as expected for OLI2/DURA and the lowest for OLI2/PisC. Considering combinations of Ori*AC the significantly lowest and highest values were observed in Taisha x Avros (Oleoflores) and Taisha x Yangambi accessions, respectively, but surprisingly both for accessions with the same AC (OLI2/DURA). As for Eo accessions the significant highest and lowest MF of hybrids were also exactly inverted compared to the SF values.

On average the overall CV values of hybrids (32.8 %) were somewhat higher than for the Eo accessions (23.9 %). They range for traits from 15 % in MF to 58 % in SF. With respect to the AC across traits the minimum value was 10.6 for MF in OLI2/PisC genotypes and the maximum value of 59 % was observed for SF in OLI2/DURA genotypes. With respect to the Ori across traits the minimum value was 10 % for MF in Taisha x Avros (Oleoflores) accessions and the maximum value of 56 % was observed for SF in Taisha x Yangambi accessions.

4. Discussion

We present the first extensive study of the allelic variation of the Sh gene homolog in a broader germplasm collection of Eo and interspecific hybrids, which can be exploited for detecting the fruit type and the origin of oil palm accessions.

In this content it is important to realize, that the appearance of the fruits in a bunch is the same as the fruit type of the mother palm. The planted seeds, however, can develop depending on the pollination into *dura*, *tenera* or even *pisifera* genotypes (Corley, R.H.V. and Tinker et al. 2016).

The identification of the fruit type is important for seed certification purposes of commercial *tenera* seeds by identifying the degree of contamination with *dura* seeds (Donini et al. 2000). Other applications arise from prospections of new or introductions of unknown plant material.

In classical oil palm breeding usually TxT or TxP crosses are performed. In both cases either for recovering *pisifera* genotypes or in the first case for evaluating *dura* genotypes. The fruit type of the planted seeds can be already determined in seedling stage and only the desired genotypes can be planted (Ritter et al. 2016), saving space and money in this way.

In this paper, we present for the first time specific primers which are able to distinguish the Sh gene homologs of Eo and Eg, irrespective of the particular species specific events.

Current breeding tendencies with interspecific hybrids go to backcross programs using repeatedly *pisifera* genotypes as pollen donors in order to maintain favourable characteristics such as shortness, oil quality and resistances to different diseases from Eo and at the same time improve the oil yield through Eg properties. While the interspecific hybrid is uniform, the BC1 would segregate with respect to the Sh gene in pure *pisifera* and "*tenera*" genotypes containing one *pisifera* allele and one Eo *dura* allele. Only the availability of specific primers for the Eo *dura* allele allows to distinguish between these two genotypes, since the *dura* primers from Eg will not amplify. The "*tenera*" genotypes can be used for BC2 crosses, while the pure *pisifera* can be used to build up an own *pisifera* collection.

The specific Eo primers (ShO) were evaluated in an extended germplasm collection from Ecuador, Peru and Colombia. Their suitability was also confirmed for material from Brazil indirectly in the Coari x La Mé hybrids. However, some other important Eo origins have not been tested. These are for example other important origins from Brazil, such as Manaus or Manicoré and also Eo germplasm from Surinam which has not been targeted. It would be convenient to validate in further studies the proposed species specific primers also in these materials.

The analyses of variances revealed no effects of Eo Sh allele combinations in Eo accessions on FW and fruit morphology traits. Also no effect of Eo Sh alleles for the same traits was detected in interspecific hybrids. These findings are not surprising, considering that the three Eo alleles are located in an intronic region, although Jo and Choi (2015) have pointed out potential influences of intron sequences such as alternative splicing or gene regulation.

According to our data in the analysed plant materials particularly the origin influences FW and morphology in Eo accessions and hybrids, except for kernel to fruit ratio (KF). In Eo accessions significant effects for combinations of Origins (Ori) and allele combinations (AC) can be found for all fruit component traits, as reflected in the significant Ori*AC interactions.

In interspecific hybrids the significant AC effects observed for FW, SF and mesocarp to fruit (MF) is mainly due to the effects of the Eg Sh allele. In addition, significant effects for combinations of Ori*AC can be found for SF and MF. For these two traits the significant highest and lowest mean values were inverted, which is reflected by the large negative correlation of -

97.2 % between the individual values of these traits. The same occurs also in Eo accessions and the correlation is 93.4 %

Unfortunately, Eo accessions and interspecific hybrids do not share any common Eo origin and cannot be compared directly. Nevertheless, if we compare the total mean values between Eo accessions and hybrids, we can see that the average FW are somewhat higher in hybrids (11.19 vs 10.45), KF are similar (7.28 vs 7.42), SF are reduced in the hybrids (17.83 vs 34.23) and the MF values are higher here (74.90 vs 58.35), as expected.

Perhaps larger differences would have been expected, but one has to consider that the hybrids descend from random samples of available materials which could be acquired from different companies, while the Eo palms descended from intensive prospections in specific areas, where fruits from palms with favourable characteristics, such as big fruits, were chosen. On the other hand, also considerable variation exist within both, Eo and hybrid origins.

In conclusion, although the specific oleifera alleles cannot be used for selecting favourable fruit morphology characteristics, they still represent markers for the applications mentioned above.

CHAPTER III: TARGETED CANDIDATE GENE APPROACH

This chapter has been published:

Astorkia M, Hernandez M, Bocs S, Lopez de Armentia E, Herran A, Ponce K, León O, Morales S, Quezada N, Orellana F, Wendra F, Sembiring Z, Asmono D, Ritter E (2019). Association Mapping between Candidate Gene SNP and Production and Oil Quality Traits in Interspecific Oil Palm Hybrids. Plants (DOI: 10.3390/plants8100377).

CHAPTER III: TARGETED CANDIDATE GENE APPROACH

1. Introduction

East-Asian countries address most of the oil palm production. Actually, Indonesia, Malaysia, and Thailand together produce almost 90% of the palm oil worldwide. Latin-American countries have started climbing positions in production few years ago, since Asian countries suffer lack of space due to increased oil demand and restricted cultivation areas (Seto and Reenberg 2014). Colombia, for example, has produced 1.68 million metric tons in 2019 (USDA 2019) and ranks now fourth in the list of most productive countries. Moreover, two other Latin-American countries can be found among the top 10 palm oil producing countries in 2017; Ecuador and Honduras which produced 273,364 and 201,665 tons of oil, respectively (Food and Agriculture Organization of the United Nations 2019).

However, the main oil palm species *Elaeis guineensis* (Eg) is suffering from bud rot disease "Pudrición de Cogollo" in these countries (Pelaez et al. 2010; Food and Agriculture Organization of the United Nations 2019) leading to important economic losses, since most of the infected palms die. In order to face this situation, seed companies work now with interspecific hybrids between *Elaeis oleifera* and Eg (Eo × Eg) (Barba 2016). These hybrids combine desirable characteristics of both species; high oil production inherited from Eg and higher amounts of oleic and linoleic acids, vitamins, sterols, and iodine values, as well as resistance to different diseases descending from Eo (Din 2000; Torres et al. 2004). Cadena et al. (2013) reported an average of 71.5 % oil in dry mesocarp of Eo × Eg interspecific hybrids, for commercial varieties of Eg var. *tenera* an average of 78 % oil content and an average of 26.3 % oil for Eo palms. They also reported the measured iodine values for these materials. Eo × Eg hybrid palms revealed an average iodine value of 66.3 g I_2 100 g^{-1}, Eg palms showed 52 g I_2 100 g^{-1} and Eo palms an average of 77.4 g I_2 100 g^{-1}.

Many breeding and seed companies have started breeding programs to get elite hybrid palms. Marker-assisted selection has emerged as a useful technology for this purpose, particularly for traits controlled by multiple genes, such as those related to oil quality and oil quantity. However, until now only a few studies have been published on this topic. Montoya et al. (2013) identified 19 quantitative trait loci (QTL) associated with fatty acid composition in an interspecific pseudo-backcross (Eo × Eg) × Eg. Singh et al. (2009) constructed a linkage map using AFLP, RFLP, and SSR markers in an interspecific cross of a Colombian Eo and a Nigerian Eg accession and detected 11 QTL for iodine value and for six components of the fatty acid composition. Since these two studies were performed in specific mapping populations, the

results may not be valid for other genetic backgrounds. Association Mapping (AM) based on linkage disequilibrium (LD) represents a way to avoid this problem, since a random population with unobserved ancestry can be studied (Risch and Merikangas 1996; Augusto and Garcia 2001). While this technique is widely used in other crops, only a few articles have been published in Eg (The et al. (2016a), Kwong et al. (2016), or Xia et al. (2018)) and none in interspecific crosses of *Elaeis* species. Therefore, in the current study a broader collection of Eo × Eg hybrids was analyzed for different traits, divided in two big groups; production and quality traits. Production traits cover agronomic performance in terms of bunch number, bunch weight, and bunch yield and the oil contents in mesocarp and bunch. The analyzed oil quality traits considered different components of lipids and tocols, as well as carotenoids. Even though these last two represent only minor components, they are of nutritional importance (Corley, R.H.V. and Tinker et al. 2016). The quality traits are described in detail under Material and Methods. The aim of this study was to determine via amplicon sequencing the allelic variation of potential candidate genes (CG) influencing these traits and to determine the effects of their particular single nucleotide polymorphisms (SNP) on trait expression, in order to exploit promising CG SNP for downstream applications in molecular breeding.

2. Material and Methods

2.1. Plant Material

A broader collection of 198 Eo × Eg F1 genotypes from five different origins were evaluated in the Energy and Palma plantation in San Lorenzo (Ecuador; 1.122980, −78.763190 GPS coordinates). These consisted of 40 hybrid genotypes from Coari × La Mé origin (Hacienda La Cabaña, Bogotá, Colombia), 75 accessions from Taisha × Avros (Oleoflores, Barranquilla, Colombia), 37 genotypes from Taisha × Avros (RGS, Quito, Ecuador), 21 genotypes from Taisha × Yangambi (RGS, Ecuador), and 25 genotypes from Taisha × Ekona (RGS, Ecuador).

2.2. Candidate Gene (CG) Selection

Partial amplicons from 167 CG related to oil production and oil quality were used for the analysis. These CG were identified randomly by *in silico* mining using different sources: (i) literature searches related to known genes from oil palm or other species with proven influence on the trait of interest, (ii) relevant patent sequences in oil palm and other species, (iii) exploration of relevant metabolic pathways such as palm oil biosynthesis for potentially useful enzymes, and (iv) analyses of published QTL and co-located transcripts with a relevant biological meaning. Amplicon primers for these CG were designed only in exons, but not in adjacent regulatory regions (López de Armentia 2017). The CG name, the Gene ID from NCBI,

the CG position according to the MPOB reference genome obtained by BLAST searches, the putative function of the CG and the forward and reverse primers used to obtain the partial amplicons can be found in Anex 1 Table A 1.

2.3. Trait Recording

Eo × Eg genotypes were planted in 2010 and phenotypic data recording started in 2014. In total, six production traits and 19 quality traits were studied.

The evaluated production traits were bunch number (BN; (nº)), bunch weight (BW; (kg)), bunch yield (BY = BN*BW; (kg)), oil percentage in fresh mesocarp (OilfM; (%)), oil percentage in dry mesocarp (OildM; (%)), and oil percentage in the bunch (OilB; (%)). BN and BW data were collected over four years and cumulative data were used for the analysis. OildM data was determined by Soxhlet extractions. OilfM and OilB were calculated according to García and Yañez (2000) as modified by Arias et al. (2015).

The analyzed oil quality traits considered different components of lipids and tocols, as well as carotenoids. Lipid components included percentages of oleic acid (OA), of saturated acids (Sat), mono-unsaturated acids (Mono-Un), and poly-unsaturated acids (Poly-Un) and were measured using the AOCS Official Ce-1h-05 (2017a) method. The iodine value (IV) in cg_{iodine}/g was measured using the AOCS Official Da 15-48 method (2017b) and the percentages of the different types of triglycerides (SSS, SUS, SUU, UUU) were analyzed using the AOCS Official Ce-5C-93 method (2017c). The nomenclature of the triglycerides indicate the saturation level of fatty acids at each of the three positions (S = saturated, U = unsaturated). Tocols (Toc) considered the sum of individual alpha, beta, gamma tocopherol´s (Tocph, Alpha, Beta, Gamma), and the sum of alpha3, beta3, gamma3 tocotrienols (Toc3, Alpha3, Beta3, Gamma3). All compounds were determined using the AOCS Official Ce 8-89 method (2017d) and are expressed in ppm. The carotene contents (Car; (ppm)) were measured using the PORIM p2.6 method (Siew and Tang 1995).

Saphiro–Wilk tests were applied in order to check for non-normal distributed data. The traits that showed a significant deviation were normalized by z-score correction and the normalized data were further used for analysis of variance (ANOVA) analyses. ANOVA analyses of the different traits and origins were performed in order to see how the origin of the different accessions affects oil production and quality. Separation of means for traits with significant differences was performed using a Tukey post hoc test. All analyses were performed using R language.

2.4. DNA Extraction and Library Construction

DNA extractions were performed from young leaflet tissue samples using the Analytik JenaLife extraction kit (Science Products, Jena, Germany) according to the manufacturer instructions.

All PCR primers were designed in exons of the CG by blasting the CG against the oil palm genome sequence from MPOB (Singh et al. 2013b) and using Primer3 software (Untergasser et al. 2012). All amplification products were visualized via gel-electrophoresis in 1.5% TAE agarose gel stained with GelRed® (Biotium, Fremont, CA, USA).

Three amplicon libraries were constructed with a total of 167 CG in the mentioned plant materials. First and second libraries were constructed with 55 CG each, while the third had 57 CG. The CG for each library were chosen randomly. The library number in which a particular CG was included is indicated in Anex 1 Table A 1. Amplicons for each CG were generated in a two-step PCR reaction as shown schematically in Figure 17, separately for each genotype.

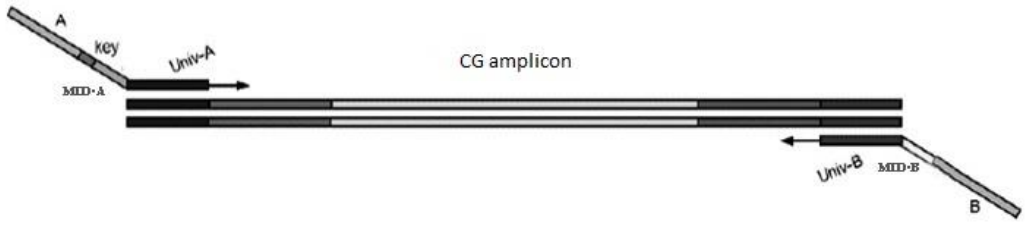


Figure 17: Scheme of the procedure for generating barcoded CG amplicons in oil palm hybrids.

For the first multiplex PCR reaction fusion primers were used which were composed of a universal part (UniA, UniB) and a part common to the CG of interest. These primers produced 120–300 bp amplicons. For each library several multiplex reactions were performed. For selecting appropriate primers for these multiplex reactions, each primer pair was tested with all others for Self-Dimers and Cross Primer Dimers formation using Thermo Fisher Multiple Primer Analyzer (2016). Sets of primers without dimer formation were used for each multiplex reaction.

A total of 20 ng of each genomic DNA, Invitrogen™ Platinum™ SuperFi™ PCR Master Mix (Life Technologies, Carlsbad, CA, USA), and 0.16 µM primer-mix were used per 25 µL amplification reaction. The PCR conditions were as follows: 98 °C denaturation for 30 s, followed by 30 cycles of 98 °C for 10 s, 58 °C for 30 s, 72 °C for 60 s, and a final elongation step of 72 °C for 5 min. PCR reactions were performed in a Thermal Cycler ABI 2720 (Applied Biosystems, Foster,

USA). Amplification products were visualized as described and the PCR products were purified using Agencourt AMPure XP (Beckman Coulter, Indianapolis, IN, USA).

All purified multiplex PCR products of a specific genotype were combined in one pool and used in a second PCR reaction to barcode each genotype. For this purpose, fusion primers were designed which were composed of one part complementary to the universal part of UniA and UniB, a genotype specific MID part, the key part (ACGT) to calibrate the sequencing machine and the specific key sequences A and B used by the sequencing platform. All primers as well as the forward and reverse MID sequences are shown in Table 12.

Table 12: Universal adapters and MID sequences used for generating barcoded amplicons of the different Candidate Genes (CG). The CG specific parts of the fusion primers are specified and replace by "X" in 1a and 1b primers below. Universal UniA and UniB parts are in italics.

No	Name	Primer Sequence
1a	**UniA_X(CG)**	Fw: *GCAAGACTCGAGCATCTCCA***X**
1b	**UniB_X(CG)**	Rv: *GCGATCGTCACTGTTCTCCA***X**
2a	**Barcode_UniA**	Fw: CATCTCATCCCTGCGTGTCTCCGA*CTCAG***[MID1]***GCAAGACTCGAGCATCTCCA*
2b	**Barcode_UniB**	Rv: CCTCTCTATGGGCAGTCGGTGAT**[MID2]***GCGATCGTCACTGTTCTCCA*

3 MID Primers and their reverse complements (rev com)

	MID (5'->3')	rev com		MID (5'->3')	rev com		MID (5'->3')	rev com
A	ACGCTCAG	CTGAGCGT	**G**	TATGCTAGA	TCTAGCATA	**L**	GCTATGACAG	CTGTCATAGC
B	TACATCAT	ATGATGTA	**H**	CGCACTGAG	CTCAGTGCG	**M**	ATACATAGCT	AGCTATGTAT
C	CGCGACTA	CTGAGCGT	**I**	CAGACTCTA	TAGAGTCTG	**N**	GACAGCGCGT	ACGCGCTGTC
D	AGCTAGTC	GACTAGCT	**J**	TGTGAGCAC	GTGCTCACA	**O**	CATGTCAGTA	TACTGACATG
E	CGAGATCA	TGATCTCG	**K**	GCGATAGTAC	GTACTATCGC	**P**	ACGCACTCGC	GCGAGTGCGT
F	TCAGTGCTG	CAGCACTGA						

The genotype specific combinations of the MID sequences with UniA and UniB sequences, respectively, allow to identify unambiguously each genotype. By using a combination of forward and reverse MID a large number of genotypes can be barcoded with a relatively small number of primers. With 2n MID primers n^2 genotypes can be discriminated.

For each barcoding reaction, a 25 µL reaction volume was prepared containing 1 µL of the purified PCR product, 0.2 µM forward and reverse barcoding primer, and Invitrogen™ Platinum™ SuperFi™ PCR Master Mix. PCR reactions and visualization were performed as described before in the first step.

PCR products of each barcoded genotype were individually quantified with a Qubit 2.0 device, using the Qubit dsDNA HS assay (Life Technologies, Carlsbad, CA, USA). Equal concentrations of genotype specific PCR products were mixed in one tube.

Each pool was purified with columns using the GeneRead Size Selection Kit (Qiagen, Hilden, Germany). The quality of the libraries was verified on an Agilent 2100 Bioanalyzer using DNA Chips with HS DNA Kit reagents according to the manufacturer's protocol (Agilent

Technologies). The libraries were sent for sequencing to the Center for Applied Medical Research (CIMA, Pamplona, Spain), using the Ion Torrent PGM. Emulsion PCR was performed with Ion PGM™ Template OT2 400 Kit according to the manufacturer's protocol. All libraries were sequenced using the 318 Chip v2 with the Ion PGM™ Sequencing 400 Kit. Sequencing was performed unidirectionally.

2.5. Sequence Processing and Association Analysis

Analyses of the obtained sequences were performed using the South Green Bioinformatics Platform http://southgreen.cirad.fr/ (South Green Collaborators 2016), which provides different bioinfomatic tools and methods for sequence analysis.

The Fastq files of the three libraries were combined and processed together, since all genotypes had the same MID combination in the three libraries. In order to obtain clean amplicon sequences, trimming and demutliplexing steps were performed. First, each genotype was identified by the combination of MIDs in each read. Sequences were separated in genotype specific files. For this purpose the public "demultiplex.py" Python script (Flutre et al. 2017) was used. Then, the "Cutadapt trimming tool" v1.8.1(Martin 2011) was applied to remove universal primer parts (UniA, UniB) and the MIDs. The cleaned, genotype specific sequences were processed using the "Snakemake-capture" script (Soriano et al. 2018) of the South Green bioinformatics platform to map the reads using BWA v0.7.15 (Li and Durbin 2010), to clean the alignments with Samtools v1.3 (Li et al. 2009), to sort the reads with Picard-tools v2.7.0 (Broad Institue 2015) and to call the SNP using GATK haplotype caller v3.7-0 (McKenna et al. 2010). The MPOB *E. guineensis pisifera* genome sequence (Singh et al. 2013b) was used as reference.

The SNP of the obtained Variant Calling Format (VCF) file were filtered using VCFtools software v4.2 (Danecek et al. 2011). Markers were filtered for only biallelic SNP with a minimum allele frequency of 0.05 and a maximum of 0.95, markers below q < 30 were eliminated as well as indels. Additionally, variants with more than 20% of missing data were eliminated for the following analyses. Genetic diversity was studied in terms He and Ho of the markers using the adegenet (Jombart 2008) and hierfstat (Goudet and Jombart 2015) packages in R. Monomorfic markers were eliminated for the following analyses. For studying genetic variances between and within origins, Fst obtained from VCFtools and Fis obtained from the hierfstat package were used. We tested also for HWE using the pegas package(Paradis et al. 2018). The null hypothesis (Ho = 0; *p* value < 0.05) was that the population is in equilibrium and pairing occurs randomly. fastStructure software v1.0 (Raj et al. 2014) was applied to analyze the population structure. Allele frequencies of each cluster from 1 to 9 were estimated with a 10-fold cross-

validation (CV). In order to choose the appropriate number of model components explaining the structure in our dataset, thechooseK.py script of the fastStructure software was run. The distruct.py script from the fastStructure was used for drawing the distruct plot.

Association studies were performed on a single marker basis using GAPIT v 3.0 (Wang and Zhang 2018) in R environment. Initially, fixed effects GLM were applied to test associations between segregating markers and phenotype for each trait. For this purpose, either Q matrix obtained from fastStructure (K = 6) was used as covariate, or PCA matrix with three components derived from GAPIT was used as covariate (GLM_Q, GLM_PCA). In addition, MLM analyses were performed in order to include both fixed and random effects. In this case, the IBS K matrix obtained from Tassel (v5.2.44) was incorporated into the previous models (MLM_Q+K, MLM_PCA+K) in order to reflect relationships among individuals with either the Q matrix or the PCA matrix. Multiple testing was also considered, since GAPIT provides beside unadjusted *p* values also FDR using the method of Benjamini and Hochberg (1995) adjusted *p* values.

The resulting observed and expected *p* values of each model were visualized separately for each trait in a QQ plot, in order to get a first impression on the fitting of different alternative models. In addition, an equation was developed to measure the average square distance (d^2) of the CG data points from the diagonal of the QQ plot for each model:

$$d^2 = \left(\sum_{i=1}^{n} {P_o}^2 + {P_e}^2 - \left(\frac{P_o+P_e}{2}\right)^2\right)/n \quad (3)$$

Equation 3: P_o and P_e are the expected and observed –log(p) values, respectively and n the number of CG data points. The model with the smallest d^2 value was considered as the best fitting model for our data.

3. Results

3.1. Phenotype Analysis

Saphiro–Wilk tests revealed 16 traits which were not normally distributed. They are marked with "*" in Table 13. The ANOVA results for testing the influence of origins on the traits are presented in Anex 1 Table A 2. Transformed data were used for non-normal distributed traits. Observed mean values, standard deviations (SD), minimum and maximum values, and the significance levels of the F tests are shown for each analyzed trait in Table A 2. All production traits showed significant differences at significance level $p < 0.001$ as well as 16 quality traits.

The SSS triglyceride (SSS), Delta compound (Delta), and Gamma compound (Gamma) traits did not reveal significant differences between origins.

The results of the Tukey post hoc tests are presented in Table 13. Production traits oil % in fresh mesocarp (OilfM), oil % in dry mesocarp (OildM), and oil % in bunch (OilB) revealed large values for the Coari × La Mé origin, while the Taisha × Ekona genotypes showed the lowest values for all production traits. On the other hand, Taisha × Avros (Oleoflores) revealed the highest values for bunch number (BN), bunch yield (BY), and bunch weight (BW) traits. For quality traits also a large difference was detected between Coari × La Mé and the other four origins. The Coari × La Mé origin showed statistically significant higher values for mono-unsaturated fatty acids % (Mono-Un), oleic acid % (OA), iodine value (IV), SUU triglyceride (SUU), or UUU triglyceride (UUU), but significant lower values than the other origins for saturated fatty acids % (Sat), poly-unsaturated fatty acids % (Poly-Un), SUS triglycerides (SUS), and tocopherol (Tocph) and tocotrienols (Toc3) compounds.

Table 13: Mean values of the studied traits for each origin and significant levels obtained by Tukey post hoc tests.

Origin	**Coari × La Mé**		**Taisha × Avros (RGS)**		**Taisha × Avros (Oleoflores)**		**Taisha × Ekona**		**Taisha × Yangambi**	
Production Traits	Mean Value	Level	Mean Value	Level	Mean Value	Level	Mean Value	Level	Mean Value	Level
BN (nº)*	52.49	B	39.81	C	63.00	A	32.75	C	40.10	BC
BW (kg)	9.44	B	11.04	B	13.22	A	9.81	B	9.91	B
BY (kg)*	501.67	B	469.32	B	845.66	A	334.30	B	444.75	B
OilfM (%)	34.69	A	28.99	B	29.31	B	24.74	C	28.71	BC
OildM (%)*	65.23	A	51.20	B	53.70	B	45.69	C	51.14	BC
OilB (%)	22.66	A	17.38	BC	19.67	B	14.09	C	17.04	BC
Oil Quality Traits										
Sat (%)*	32.07	B	37.91	A	38.64	A	39.39	A	40.00	A
Mono-Un (%)*	56.06	A	48.74	B	46.54	B	46.66	B	46.05	B
Poly-Un (%)	12.35	C	13.01	BC	14.31	A	13.68	AB	13.62	AB
OA (%)*	54.84	A	47.21	B	44.84	B	44.98	B	44.16	B
IV (cg/g)*	68.87	A	63.25	B	63.56	B	61.97	B	61.62	B

SSS (%)*	1.08	-	1.48	-	1.12	-	1.18	-	1.64	-
SUS (%)*	17.76	B	24.38	A	25.41	A	25.80	A	25.99	A
SUU (%)	35.82	A	31.95	B	31.40	B	31.77	B	29.44	B
UUU (%)*	21.06	A	12.01	B	10.28	B	10.23	B	9.90	B
Tocph (ppm)*	164.37	C	247.15	AB	198.63	BC	290.47	A	255.70	AB
Alpha (ppm)*	115.22	B	178.43	A	130.78	B	211.37	A	203.34	A
Delta (ppm)*	40.28	-	44.17	-	40.93	-	54.31	-	43.10	-
Gamma (ppm)*	39.49	-	46.64	-	47.29	-	47.35	-	42.15	-
Toc3 (ppm)	874.15	C	1087.74	B	1338.07	A	1159.80	AB	1065.70	BC
Alpha3 (ppm)	203.71	C	313.70	B	396.75	A	320.28	AB	314.24	AB
Delta3 (ppm)*	66.96	B	98.57	B	143.11	A	95.68	B	80.82	B
Gamma3 (ppm)	605.39	B	675.47	B	806.15	A	743.84	AB	670.64	B
Toc (ppm)	1038.52	B	1334.90	A	1543.12	A	1450.27	A	1321.40	AB
Car (ppm)*	785.89	BC	832.09	B	671.91	C	900.20	AB	1068.65	A

* Means with the same letter are not statistically different ($\alpha > 0.05$). Traits marked with "*" did not follow a normal distribution according to Saphiro–Wilk tests. Production traits: bunch number (BN), bunch weight (BW), bunch yield (BY), oil % in fresh mesocarp (OilfM), oil % in dry mesocarp (OildM) and oil % in bunch (OilB). Quality traits: oleic acid % (OA), saturated fatty acids % (Sat), mono-unsaturated fatty acids % (Mono-Un), poly-unsaturated fatty acids % (Poly-Un), iodine value (IV), carotene contents (Car), different types of triglycerides in % (SSS, SUS, SUU, UUU), tocopherol (Tocph) compounds; Alpha, Delta, Gamma, tocotrienol (Toc3) compounds; Alpha3, Delta3, Gamma3, tocols (Toc).

3.2. Genotype Analysis

Three separate amplicon libraries were constructed with a total of 167 candidate genes. The first library was constructed from 56 candidate genes and yielded over 13.9 million raw reads. The second library from 55 CG produced around 9.2 million raw reads and the third library from 56 CG generated around 9.6 million raw reads. This total number of 32.7 million reads was reduced to 9.8 million clean reads after the filtering steps. Approximately 83% of the reads mapped to the Eg reference genome. The Snakemake-capture workflow identified initially 12,200 potential SNP. However, after the mentioned filtering steps, only 115 potential SNP

remained for the following analyses. The average observed (Ho) and expected heterozygosity (He) were 0.61 and 0.37, respectively. Bartlett's test revealed a significant difference between expected and observed heterozygosity. The fixation indices (Fst) values revealed no discriminant differentiation between populations as can be seen in Table 14, since all values were close to zero. With respect to the Fst values, the largest distances between origins were observed between Coari × La Mé and Taisha × Avros (Oleoflores) or Taisha × Yangambi, while the closest distances were observed between Coari × La Mé and Taisha × Avros (RGS) and between Taisha × Ekona and Taisha × Yangambi. The inbreeding coefficients (Fis) values revealed no relatedness between individuals of the same origin since all obtained values were negative, suggesting a high diversity within origins. The Chi square tests indicated that only 38 of the markers were in Hardy–Weinberg equilibrium (HWE), while the other 77 showed significant deviations.

Table 14: Genetic diversity studies in terms of inter cross Fixation indices (Fst) and intra cross Inbreeding coefficients (Fis).

Inter-Cross Fst Value	**Taisha × Yangambi**	**Taisha × Ekona**	**Taisha × Avros (Oleoflores)**	**Taisha × Avros (RGS)**	**Coari × La Mé**
Taisha × Yangambi	-	0.028876	0.055139	0.068303	0.10416
Taisha × Ekona	-	-	0.051121	0.064635	0.083617
Taisha × Avros (Oleoflores)	-	-	-	0.10259	0.10992
Taisha × Avros (RGS)	-	-	-	-	0.012305
Intra-Cross Fis Values	−0.7447191	−0.69170213	−0.72402062	−0.46477064	−0.46522124

Cluster analysis of the 115 markers by fastStructure for determining ancestry indicated that six sub-populations (K = 6) exists in our germplasm. These six cluster are represented in Figure 18 as distruct plot. This parameter was also used for association mapping analyses.

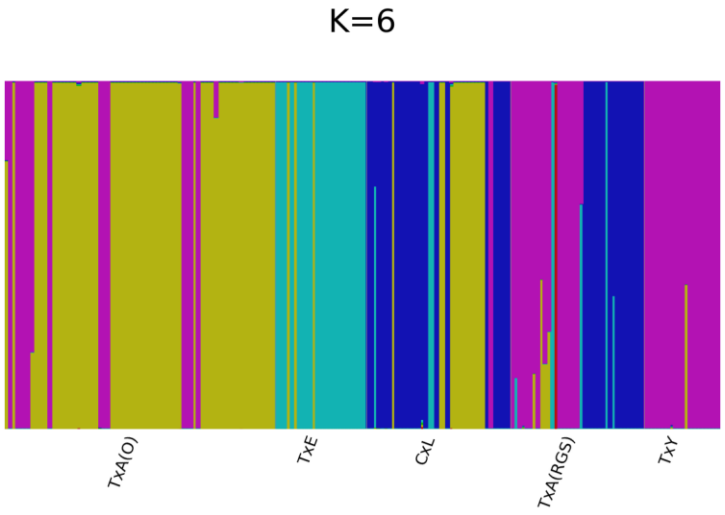


Figure 18: distruct plot of the 6 clusters used to explain our population structure. Each genotype is represented by one line and the colors indicate the estimated fraction of each individual to each sub-population.

3.3. Association Analysis

The remaining 115 SNP belong to 62 of the 167 initial CG used in the study and four of them showed multi locus mapping at two loci. SNP numbers for each candidate gene varied between one and four. The remaining CG are shown in Anex 1 Table A 3. Internal names for these 62 CG, the NCBI Gene ID, the CG position on the Malaysian Palm Oil Board (MPOB)'s reference genome, as well as the putative function of the CG are indicated in that Table.

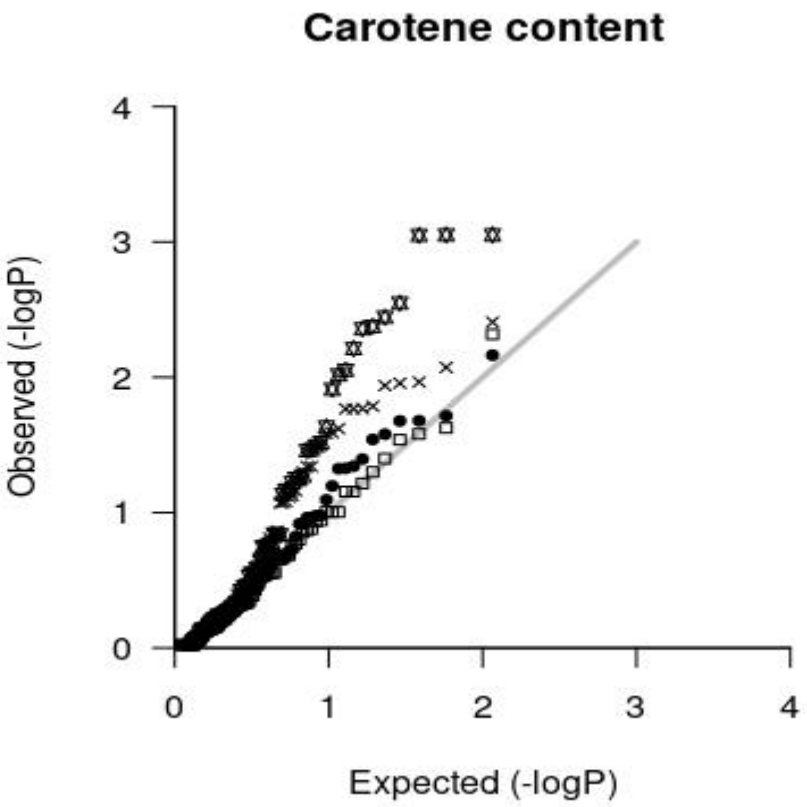


Figure 19: Example for a Quantile-Quantile (QQ) plot for Carotene contents (Car). Candidate gene (CG) data points of alternative generalized linear model (GLM) with structure matrix (Q) or principle component analysis matrix (PCA) as covariates: GLM_Q, GLM_PCA, respectively, and linear mixed models (MLM) incorporating in addition the IBS Kinship matrix (K) into the models: MLM_Q+K, MLM_PCA+K. They are represented by different symbols. (black circles: MLM_PCA+K; white squares: MLM_Q+K; stars: GLM_Q; crosses: GLM_PCA).

After running Association Mapping using GAPIT, expected and observed p values of each model were drawn as a Quantile-Quantile (QQ) plot for each trait. Figure 19 shows an example of a QQ plot for carotene contents (Car), reflecting the fitting of different alternative fixed generalized linear model (GLM) and fixed and random linear mixed models (MLM). The QQ plots for each trait are shown in Anex 1 Figure A1 1. The below described formula for calculating the average square distance (d^2) of the CG data points from the diagonal of the QQ plot was applied for determining the best fitting model for each trait, even though in several cases the differences in the values for alternative models are very small. The results are shown in Table 15. For all production traits except OilfM MLM gave the best results. The OilfM trait fitted best with the GLM taking into account the structure matrix (Q) model; GLM_Q. OildM and OilB traits fitted best with the MLM using principle component analysis matrix (PCA) and

IBS Kinship matrix (K); MLM_PCA+K and the three bunch related traits with MLM_Q+K models. Additionally, for most quality traits, mixed models were found to be the best fitting models, but some traits such as Alpha3 compound (Alpha3), Gamma, tocols (Toc), and Toc3 revealed better results with fixed effect models. Eight of the quality traits fitted better with MLM_PCA+K models and the other seven with MLM_Q+K models.

Table 15: Average square distance (d^2) values of the CG data points from the diagonal of the QQ plot for determining the best fitting model for each trait.

Production Traits	GLM_PCA	GLM_Q	MLM_PCA+K	MLM_Q+K
BN	0.4349	0.335	0.350	**0.286**
BY	0.369	0.335	0.298	**0.289**
BW	0.377	0.383	0.357	**0.332**
OilfM	0.294	**0.293**	0.294	0.294
OildM	0.281	0.285	**0.281**	0.327
OilB	0.331	0.337	**0.303**	0.458
Oil Quality Traits				
Sat	0.301	0.332	**0.270**	0.442
Mono-Un	0.305	0.352	**0.298**	0.317
Poly-Un	0.348	0.385	**0.347**	0.381
OA	0.333	0.365	**0.323**	0.426
IV	0.434	0.376	**0.327**	0.753
SSS	0.310	0.312	0.295	**0.292**
SUS	0.286	0.319	**0.285**	0.314
SUU	0.272	0.279	**0.271**	0.282
UUU	0.313	0.348	**0.306**	0.355
Tocph	0.333	0.355	0.323	**0.322**
Alpha	0.359	0.394	0.330	**0.327**
Delta	0.341	0.319	0.341	**0.317**
Gamma	0.265	**0.260**	0.265	0.266
Toc3	0.315	**0.306**	0.315	0.311
Alpha3	0.284	**0.264**	0.284	0.270

Delta3	0.329	0.382	0.315	**0.295**
Gamma3	0.342	0.339	0.337	**0.333**
Toc	0.325	**0.309**	0.325	0.316
Car	0.486	0.645	0.359	**0.334**

The best fitting model with smallest d^2 value is indicated in bold and underlined for each CG.

Table 16 presents the results of association mapping. The detected associations based on observed unadjusted *p* values < 0.05 between CG SNP and traits are displayed, as well as the genome location of the significant SNP, the applied model, the significance level of the association, the explained variance, and the effect of the marker. The significant SNP which belong to a particular CG were grouped.

SNP belonging to a total of seven CG influenced significantly six production traits. Three CG revealed significant effects on two different production traits, while the other four CG influenced only one trait each, leading to a total of 10 significant associations for production traits. The BW trait was influenced by three different CG, OildM and OilB by two CG and BN, BY, and OilfM by only one CG. The explained variances by the model ranged from 8.9% to over 26% for the different CG.

Table 16: Results of association mapping between CG Single nucleotide polymorphisms (SNP) and production and oil quality traits in oil palm hybrids.

CG	**SNP Position**	**Production Traits**	**AM Model**	***p* Value**	**%VA**	**Effect**
BKACPII_1	C10: 22949607	BW	MLM_Q	0.013	13.9	6.812
		BY	MLM_Q	0.037	26.2	538.811
EgNAC	C05: 40852639	OildM	MLM_PCA	0.044	18.3	-5.524
		OilB	MLM_PCA	0.046	8.9	-3.256
LIPOIC	C07: 18432097	OilfM	GLM_Q	0.042	10.9	-2.387
M2200	C13: 12503450	OildM	MLM_PCA	0.009	19.9	13.384
PKP-ALPHA	C01: 40816686	OilB	MLM_PCA	0.007	10.8	-9.339
SEQUI	U02: 19591286	BW	MLM_Q	0.015	14.1	2.319
TO1	U02: 79752170	BN	MLM_Q	0.020	24.4	-45.134
		BW	MLM_Q	0.033	14.8	-6.218

CG Name	SNP Position	Quality Traits	AM Model	*p* Value	%VA	Effect
ATAGB1_ML*	C13: 103569	SSS	MLM_Q	0.022	7.2	-0.614
		Mono-Un	MLM_PCA	0.008	20.4	-5.291
		Poly-Un	MLM_PCA	0.047	10.7	1.136
ATP3	U05: 50035832	Mono-Un	MLM_PCA	0.050	18.7	-5.726
		Poly-Un	MLM_PCA	0.003	13.7	2.549
atpB	CT: 54552	Delta	MLM_Q	0.046	11.6	-6.913
BnC8_761	C08: 4351912	Delta3	MLM_Q	0.008	17.3	33.287
		OA	MLM_PCA	0.025	20.1	-2.488
		UUU	MLM_PCA	0.048	21.8	-2.250
CA3	C02: 35978226	Delta	MLM_Q	0.045	11.6	15.740
EgNAC	C05: 40852136	OA	MLM_PCA	0.015	20.6	2.890
		Sat	MLM_PCA	0.042	17.2	-1.990
		SUS	MLM_PCA	0.014	23.2	-2.189
		SUU	MLM_PCA	0.019	14.3	2.125
		UUU	MLM_PCA	0.007	23.5	3.246
	C05: 40852594	Mono-Un	MLM_PCA	0.044	18.8	3.568
		OA	MLM_PCA	0.005	21.7	4.826
		Poly-Un	MLM_PCA	0.010	12.3	-1.310
		SUS	MLM_PCA	0.009	23.6	-3.376
		UUU	MLM_PCA	0.003	24.3	5.081
	C05: 40852639	Car	MLM_Q	0.026	26.9	-173.576
EOCHYB	C04: 37534489	Alpha	MLM_Q	0.027	14.6	-50.440
GLUT1	C12: 28135330	OA	MLM_PCA	0.040	19.7	-2.823
	C12: 28135361	OA	MLM_PCA	0.040	19.7	-2.823
	C12: 28135379	OA	MLM_PCA	0.040	19.7	-2.823
HtC2_11412	C08: 25294023	Delta3	MLM_Q	0.036	15.7	27.356
		SUU	MLM_PCA	0.047	13.4	-1.552
	C08: 25294107	Delta3	MLM_Q	0.015	16.7	29.133

		SSS	MLM_Q	0.049	6.1	0.290
HtC2_1255C2-411	C02: 43975856	SSS	MLM_Q	0.046	6.1	0.529
	C02: 43975982	SSS	MLM_Q	0.046	6.1	0.529
HtC7_9200	C06: 41269483	Toc	GLM_Q	0.042	13.6	157.269
		Tocph	MLM_Q	0.046	13.2	35.249
	C06: 41269559	Car	MLM_Q	0.005	28.3	-158.848
JC35	C13: 22806955	Car	MLM_Q	0.024	27.0	-109.838
JC55	C05: 14759308	IV	MLM_PCA	0.007	15.1	7.826
LIPOIC	C07: 18431998	Gamma	GLM_Q	0.041	7.0	-7.841
		Mono-Un	MLM_PCA	0.039	18.9	-2.700
		OA	MLM_PCA	0.024	20.2	-2.842
		Poly-Un	MLM_PCA	0.037	11.0	0.782
	C07: 18432097	Gamma	GLM_Q	0.003	11.6	-11.135
		Toc	GLM_Q	0.043	13.5	-184.481
		Toc3	GLM_Q	0.027	16.3	-173.161
		Delta3	MLM_Q	0.033	16.0	-27.292
PAT_2	C09: 34725045	Alpha	MLM_Q	0.014	15.5	49.496
		Delta	MLM_Q	0.027	12.1	9.758
		Tocph	MLM_Q	0.005	15.4	70.245
PAT_2_ML	C02: 23775894	Poly-Un	MLM_PCA	0.035	11.0	-0.877
PAT_6	C08: 27075521	Car	MLM_Q	0.040	26.6	-163.806
PDHB	C01: 51857834	IV	MLM_PCA	0.027	13.8	3.407
PKP-ALPHA	C01: 40816686	UUU	MLM_PCA	0.034	22.1	-7.962
SEQUI	U02: 19591232	Toc	GLM_Q	0.031	13.9	260.781
		Toc3	GLM_Q	0.040	15.9	212.755
		Gamma3	MLM_Q	0.015	14.1	142.540

		SSS	MLM_Q	0.028	6.6	0.504
		IV	MLM_PCA	0.037	13.5	-2.828
		Poly-Un	MLM_PCA	0.020	11.6	-1.139
	U02: 19591286	IV	MLM_PCA	0.020	14.1	−3.630
SHELL	C02: 3078054	Alpha	MLM_Q	0.029	14.6	66.367
		Delta	MLM_Q	0.028	12.1	17.751
		Tocph	MLM_Q	0.019	14.1	90.283
		Toc	GLM_Q	0.031	13.9	213.715
		Toc3	GLM_Q	0.046	15.8	169.457
	C02: 3078154	Alpha	MLM_Q	0.048	14.2	37.502
		Delta3	MLM_Q	0.026	16.1	32.474
		Gamma3	MLM_Q	0.023	13.7	109.420
		Tocph	MLM_Q	0.029	13.7	51.563
TO1	U02: 79752182	Gamma3	MLM_Q	0.030	13.6	136.526
	U02: 79752184	Gamma3	MLM_Q	0.030	13.6	136.526
TO3	C03: 13885419	Car	MLM_Q	0.029	27.0	-157.239

Legend: CG Name: internal name of the CG; SNP position: genome location of the SNP; Trait: associated trait; Association Mapping (AM) Model: best fitting model for AM; *p* value: observed error probability value for the model; %VA: percentage of the total variance explained by the model; Effect: effect of the marker.

For quality traits SNP belonging to a total of 23 CG showed potential significant associations with 18 out of the 19 quality traits using unadjusted *p* values. Alph3 did not show any association with any of the studied CG SNP. The explained variances by the models ranged from 6.1% to over 28% of the total variance. For nine CG more than one SNP showed associations with different traits. Poly-Un showed associations with six of the studied CG and the Car trait revealed associations with five CG. Four potential associations were observed for the Delta, Delta3 compound (Delta3), Mono-Un, OA, SSS, and Toc traits and three associations for Alpha compound (Alpha), Gamma3 compound (Gamma3), IV, Toc3, Tocph, and UUU. SUU revealed two potential associations and Gamma, Sat and SUS showed only one potential association. It is also worth to notice, that four of the CG—EgNAC, SEQUI, LIPOIC, and TO1—

showed also potential effects on different production traits. However, considering FDR adjusted *p* values, all detected associations are not significant anymore.

4. Discussion

4.1. Phenotypic Data Analysis

The analyses of production traits revealed larger differences between Coari × La Mé genotypes and the other four origins where Taisha was involved. The Coari × La Mé origin presented on average a higher oil to bunch (OilB) percentage and higher oil percentages in fresh and dry mesocarp (OilfM, OildM). Peláez et al. (2010) observed that Coari palms as well as their hybrids with *Eg* had higher CO_2 fixation capacities, which are positively correlated with an increase in oil contents (Corley, R.H.V. and Tinker et al. 2016). On the other hand, Taisha palms have been described by Barba (2019) as "Oleifera Guineensis palms", since they have similar morphological characteristics as *guineensis* palms. In our study we also found higher bunch weights (BW) in all origins involving Taisha and a higher bunch yield (BY) in the Taisha × Avros (Oleoflores) accessions. However, Arias et al. (2015) studied different Eo origins and detected the highest total oil-per-bunch ratios [%] for Taisha accessions followed by Coari accessions, indicating that there may be considerable variation between the particular accessions of the origins. From a commercial point of view (CPO, crude palm oil yield), also the industrial extraction rates have to be considered, which according to Soh et al. (2017) are lower for hybrids involving Taisha.

Considering analyses of quality traits, some studies are available from Montoya et al. (2013), Singh et al. (2009), and Cadena et al. (2013). These authors analyzed beside iodine value particularly the fatty acid composition in interspecific hybrids from controlled crosses and established linkage maps with integrated QTL for these traits. Cadena et al. studied the lipase activity, oil contents in fresh mesocarp, and iodine values in a collection of Eg, Eo, and Eo × Eg genotypes. However, we present here the first detailed oil quality analyses for oil palm involving 19 different quality traits. These include traits related to lipids where the saturation level of fatty acids was measured, considering the percentages of saturated (Sat), mono-unsaturated (Mono-un), and poly-unsaturated (poly-un) fatty acids. The mono-unsaturated fatty acids are considered as the healthiest (Tierney and Roche 2007; Qian et al. 2016). We also analyzed the percentage of oleic acid in the oil (OA) which was classified as mono-unsaturated omega-9 fatty acid, the iodine value (IV) indicating the global degree of unsaturated fatty acids, and particularly the different types of triglycerides which can be formed from three fatty acids (SSS > SUS > SUU > UUU). We found large differences between Coari × La Mé and the

other origins. Coari × La Mé accessions showed desirable characteristics such as high contents of mono-unsaturated acids, oleic acid, high iodine values, and UUU and SUU triglycerides, while the saturated acid levels were significantly below those of the other origins. Pelaez et al. (2010) also determined higher oleic acid contents and iodine values in Coari palms.

We performed also a detailed study for tocols contents which are composed of tocotrienols and tocopherols. These components represent different forms of vitamin E and can be found in oil palm as beneficial phytonutrients (Nesaretnam et al. 1995). Both, tocotrienols and tocopherols have four isomers each (α-, β-, γ-, δ-) and have unique benefits (Yuen May and Nesaretnam 2014). Here we studied three of them (α-, β-, γ-). In contrary to what has been observed above, Coari × La Mé accessions showed significantly less contents of tocols. The α isomers from tocopherols and tocotrienols revealed lower quantities compared to the other four origins. Finally, carotenoids contents were measured in the five origins. These pigments are responsible for the orange-red brilliant color of the oil and are precursors of vitamin A(Yuen May and Nesaretnam 2014). For this trait the Taisha × Yangambi origin revealed the highest content.

4.2. SNP Detection and Genetic Diversity Analysis

We used the Ion Torrent Personal Genome Machine (PGM) sequencing platform for convenience, based on previous experiences in other studies and ease of access. Similar studies using the PGM platform were also performed by other authors (Guo et al. 2015; Singh et al. 2019).

Mapping of the sequenced reads were performed using the published *Eg* var. *pisifera* genome sequence as reference. The decision to use this genome relied on the fact that actually no reference genome exists for Eo even though Singh et al. (2013b) published a draft. Nevertheless, Camillo et al. (2014) analyzed genome sizes of Eg, Eo, and interspecific hybrids with the intention to reveal in the near future the genome sequence of Eo. When available, the genome sequences of both *Elaeis* species could be used as reference for mapping the sequence reads.

In our analysis 83% of the reads could be mapped onto the reference genome and 12,200 SNP were identified initially. According to Singh et al. (2013b) 73 % of the transposable element contents differ between Eg and Eo and could decrease the SNP numbers, since the reads in the hybrids descend from both *Elaeis* species. The high number of SNP was reduced drastically after the filtering steps and only 115 potential markers remained. The 62 targeted CG included two CG with multi-locus characteristics (PAT_2, ATAGB1), since they mapped to different

chromosomes on the genome. These results suggest that the corresponding CG primers were specific for gene families rather than for individual CG.

Random seed samples were received descending from multiple crosses made by Oleflores and RGS. However, nothing was known about the population structure a priori. Therefore, we performed some global genetic analyses. The Ho = 0.61 was significantly higher than the He = 0.37 in the accessions of all five origins. This high Ho value is in accordance with Arias et al. (2015) who evaluated phenotypic and genetic diversity in two assays using of 13 and 19 SSR markers to characterize different Eo origins, including two Eo × Eg accessions and calculated Ho values of even 0.70 and 0.77 in the two assays, respectively. They also observed that 27% and 32% of the detected alleles in the study represented specific alleles of the different Eo origins and that one of the Eo × Eg accessions had the largest number of specific alleles. Arias et al. (2014) found also for Eg accessions higher observed heterozygosity levels than the expected ones in most of the 23 analyzed origins. This can explain also the findings in our study since Eo origins from Brazil (Coari) and from Ecuador (Taisha), as well as Eg origins from La Mé, Ekona, Yangambi, and Avros are incorporated into our hybrids. Furthermore, due to the nature of our F1 hybrids, it is expected to observe a higher Ho value.

According to Johnson and Shaw (2015) the high Ho value is also coherent with the observed negative values of the computed Fis values in each of the five origins indicating high levels of genetic variability(Johnson and Shaw 2015). The observed high Ho value leads consequently also to high deviations from HWE (77 markers out of 115).

4.3. Association Mapping Results

Many studies have been published for the important oil palm crop Eg with the objective of crop improvement. However, the hybrids between the *Elaeis* species, which are so important in Latin-American regions, have been studied far less so far. Actually, only some QTL studies have been performed in order to improve the crop (Singh et al. 2009; Montoya et al. 2013; Ting et al. 2014, 2016). However, these studies consider structured (mapping) populations Here we performed a genotype-phenotype association study where the germplasm represents a random population with unobserved ancestry.

In total four different models were used for association mapping. Two GLM models with population structure (GLM_Q) and principal component analysis (GLM_PCA) as covariates and two MLM models where in addition a K matrix between individuals was included (MLM_Q+K, MLM_PCA+K). After the analysis, the coincidence of observed and expected *p* values was visualized in a QQ plot for each trait. Several authors have used these QQ plots to determine the best fitting models visually (Gamazon et al. 2015; Álvarez et al. 2017; Lin et al. 2017).

When looking to the example of a QQ plot for carotenes contents in Figure 19, it can be seen clearly that the GLM_Q model represented by “stars” is the worst for fitting our data, while deciding visually between the other three models is impossible. Therefore, we developed an equation to calculate the average square distance (d^2) of the CG data points from the diagonal of the QQ plot which represents an objective method for determining the best fitting model for each trait.

In our study the mixed effects models fitted better for most of our traits, while only a few traits were found to have better associations with GLM models where the K matrix was not taken into account. These findings are in accordance with those of Wang et al. (2012), Nigro et al. (2019), or Lin et al. (2017), who reported that MLM models were more appropriate for association studies in maize and wheat.

As pointed out by Gao et al. (2016) the output of FDR adjusted *p* values from GAPIT is highly stringent, leading to the loss of the detected significant associations using unadjusted *p* values. *A p* value of 0.05 was set as threshold for identifying potential CG with potential significant influence on a trait as also in other studies with similar approaches (Pasam et al. 2012; Zegeye et al. 2014; Gao et al. 2016; Li et al. 2016). In total, seven CG were found to be related to six production traits and 23 CG to 18 quality traits (Table 16). With respect to the CG with significant effects, special attention has to be paid to four of them (LIPOIC, SEQUI, TO1, EgNAC) with a potential relevant biological meaning.

If not considering FDR adjusted *p* values, LIPOIC revealed potential associations with one production traits (OilfM,) and six quality traits (Gamma, OA, Mono-Un, Poly-Un, Toc, Toc3) It represents a lipoyl synthase gene, responsible for the synthesis of lipoic acid a universal antioxidant under oxidative stress conditions. This gene is required for cell growth, mitochondrial activity, and coordination of fuel metabolism and uses multiple mitochondrial 2-ketoacid dehydrogenase complexes (Solmonson and Deberardinis 2017) for the catalysis. Together with LIP2 it is essential for mitochondrial protein lipoylation during seed development (Ewald et al. 2014). It is known to be of high importance for obtaining high yielding plants (Schoen et al. 2010).

Using unadjusted *p* values, also TO1 may influence the production traits BN and BW and one quality trait Gamma3. This CG represents a gamma-tocopherol methyltransferase which catalyzes the conversion of gamma-tocopherol into alpha-tocopherol. In *Arabidopsis* the overexpression of this enzyme resulted in more than 80-fold increase of α-tocopherol at the expense of γ-tocopherol without changing the total tocopherol contents (Shintani and DellaPenna 1998).

The candidate gene SEQUI showed potential influence with one production trait, BW, and six quality traits; Toc, Toc3, Gamma3, SSS, IV, and Poly-Un. It is an alpha-humulene synthase transcript related to zerumbone biosynthesis. This compound is known as an essential oil of *C. verbenacea* and *Cannabis sativa* L. (Fernandes et al. 2007; Benelli et al. 2018) and has healing effects as a multi-anticancer agent(Yu et al. 2008) and anti-inflamatory effects (Fernandes et al. 2007). This compound also mediates the formation of beta-caryophyllene, another oil compound related to reduce systemic inflammation and oxidative stress(Ames-Sibin et al. 2018).

Finally, EgNAC showed that it could be associated with seven quality traits and one production trait. NAC transcription factors have been studied widely in different crops. They are known to regulate different plant functions in plants, such as fruit ripening in tomato (Kou et al. 2016), variations in the protein content of wheat (Hu et al. 2013), increase in seed yield (Liang et al. 2014), and regulative functions for biotic and abiotic stress responses (Nuruzzaman et al. 2013).

These findings indicate that many significant candidate genes could be involved in complex biological pathways, but there is still a lot of information missing. Fully understanding these metabolic pathways can help to discover the precise role of these genes influencing particular characters and can be a good starting point to obtain higher yielding oil palm varieties with increased oil contents. Association mapping results could be exploited in potential downstream applications by selecting genotypes with superior alleles of different significant candidate genes in Marker Assisted Selection systems.

Production traits are the most interesting characters from a commercial point of view. However, quality traits are becoming more and more important in recent years. Breeding Companies look for high quality oil properties in order to satisfy customer's preferences. Components such as high levels of unsaturated acids, high carotene contents, or high amount of tocols are becoming more and more important traits for taking into account. Our association mapping approach and whole understanding of the function of these detected candidate genes could help to obtain improved palms with these desired qualities.

In our study we only considered partial amplicons from a reduced number of candidate genes, limiting the scope of our approach. Further studies should be conducted in the future to improve the results, considering other molecular tools such as whole genome resequencing, transcriptome sequencing, or bait sequencing in order to increase the number of targets.

CHAPTER IV: RESTRICTION SITE ASSOCIATED RNA SEQUENCING (RARSeq) APPROACH

This chapter has been published:

Astorkia M, Hernandez M, Bocs S, Ponce K, León O, Morales S, Quezada N, Orellana F, Wendra F, Sembiring Z, Asmono D, Ritter E (2020). Detection of significant SNP associated with production and oil quality traits in interspecific oil palm hybrids using RARSeq. Plant Science (DOI: 10.1016/j.plantsci.2019.110366)

CHAPTER IV: RANDOM RESTRICTION SITE ASSOCIATED RNA SEQUENCING (RARSEQ) APPROACH

1. Introduction

The constant improvement in Next Generation Sequencing (NGS) platforms is leading to lower sequencing prizes as well as higher sequencing efficiency in terms of read numbers and shorter processing times. Molecular approaches such as Restriction Associated DNA sequencing (RADSeq) (Davey et al. 2010) and Genotyping by sequencing (GBS) (He et al. 2014) have also taken advantage of the NGS improvements for obtaining a massive amount of variants in populations of hundreds or even thousands of genotypes. These two techniques rely on the digestion of several genotypes with one or two enzymes and the posterior selection of part of the genome of interest. Moreover these techniques do not require previous information about the subject of interest, and the library preparation is quick, easy and cheap. Association mapping (AM) studies are being constantly performed with data obtained from RADSeq or GBS laboratory assays (Xu et al. 2014; Verma et al. 2015; Alipour et al. 2017; Yu et al. 2017; Bai et al. 2017; Sant'Ana et al. 2018). AM strategies have been used in a wide range of plants for identifying genes responsible for quantitative variation of desired traits (Li et al. 2016; Zhao et al. 2017; Swamy et al. 2017). In contrast to quantitative trait loci (QTL) (Paterson et al. 1988) analyses in linkage mapping populations, AM is defined by a marker-trait association based on linkage disequilibrium (LD) in unstructured germplasm, avoiding associations based on stratification. Furthermore, it allows to analyse random germplasm and helps to obtain higher amounts of alleles for each gene (Buckler and Thornsberry 2002; Yu and Buckler 2006).

Nevertheless, RADSeq and GBS assays work with DNA leading mostly to intragenic variants (Su et al. 2017; Carrasco et al. 2018). A solution to this problem was proposed by Alabady et al. (2015) who performed a population genomics and mapping analysis based on a Restriction Site Associated RNA Sequencing (RARSeq) assay. They concluded that this novel approach holds several benefits, such as enrichment for functional markers, a balanced depth of coverage of the samples, obtainment of transcriptome-wide unbiased markers, robustness and sufficient overlap across individuals. The study published by Alabady et al. opens up a new door to explore AM strategies in a more efficient way, since most of the markers are functional markers.

Crops like Oil Palm reach stable performance six to 12 years after planting (Ismail and Mamat 2002), extending considerably classical breeding programs and forcing breeders to work with

molecular markers assisted selection strategies. For the main Oil palm species *Elaeis guineensis* (Eg) several AM studies have been published (Teh et al. 2016b; Xia et al. 2018).

In Central America Eg palms are dying because of the devastating "Pudrición de Cogollo disease (PC)" (Sundram and Intan-Nur 2017), but cultivation of hybrids between *Elaeis oleifera* (Eo) and Eg (Barba 2016) represent an alternative in recent years. These hybrids show tolerance to the disease and bring with them other desirable characteristics such as better oil quality and longer productive life (Din 2000; Torres et al. 2004).

Palm oil is the most used vegetable oil worldwide and the oil palm market is highly competitive and shows growing tendencies (USDA 2019). Therefore, Central American companies have also started to implement molecular breeding programs with hybrids, in order to improve the oil production along with oil quality. Until now only a few studies have been published about the improvement of the desired characteristics of these hybrids. Montoya et al. (2013) identified 19 QTL related to fatty acid composition in an interspecific pseudo-backcross (Eo x Eg) x Eg. Singh et al. (2009) constructed a linkage map using AFLP, RFLP and SSR markers in an interspecific cross of a Colombian Eo and a Nigerian Eg accession and detected 11 QTL for iodine value and for six components of the fatty acid composition. However these two studies were performed in specific mapping populations and the results may not be valid for other genetic backgrounds.

We have performed a RARSeq based Association Mapping study using different models in a broader population of Eo x Eg hybrid genotypes from different origins. The aim was to identify functional markers associated to different production and oil quality traits which could be exploited for downstream applications in molecular breeding programs.

2. Material and Methods

2.1. Plant material

We have evaluated a population of 104 *Elaeis oleifera* (Eo) x *Elaeis guineensis* (Eg) genotypes of five different origins in the Energy & Palma plantation (San Lorenzo, Ecuador, GPS coordinates: 1.122980, -78.763190). These palms consisted of 17 Taisha x Ekona (RGS, Quito, Ecuador) accessions, 21 Coari x La Mé (Hacienda La Cabaña, Bogotá, Colombia) accessions, 23 Taisha x Avros (RGS, Ecuador), 31 Taisha x Avros (Oleoflores, Barranquilla, Colombia) accessions and 12 Taisha x Yangambi (RGS, Ecuador) accessions.

2.2. Trait recording

The hybrids were planted in 2010 and phenotypic data recording started in 2014. In total 6 production traits and 19 quality traits were studied (Table 18).

The Production traits were bunch number (BN; [no]), bunch weight (BW; [Kg]), bunch yield (BY=BN*BW; [kg]), oil percentage in fresh mesocarp (OilfM; [%]),oil percentage in dry mesocarp (OildM; [%]) and oil percentage in bunch (OilB; [%]). BN and BW data were collected during four years and cumulative data were used for the analysis. OildM data was determined by Soxhlet extractions. OilfM and OilB were calculated according to García and Yañez (2000) as modified by Arias et al. (2015).

Oil quality traits can be divided in three groups; lipids, tocols and carotenoids. Lipids were characterized as percentages of oleic acid (OA), of saturated acids (Sat), mono-unsaturated acids (Mono-Un) and poly-unsaturated acids (Poly-Un) and measured using the AOCS Official Ce-1h-05 (2017a) method. The iodine value (IV) in cg iodine g^{-1} was measured using the AOCS Official Da 15-48 method (2017b) and the percentages of the different types of triglycerides (SSS, SUS, SUU, UUU) using the AOCS Official Ce-5C-93 method (2017c). The nomenclature of the triglycerides indicate the saturation level of fatty acids at each of the three positions (S=saturated, U=unsaturated). Tocols (Toc) were calculated as the sum of tocopherols (Tocph) and tocotrienols (Toc3), which in turn were computed as the sum of individual alpha, beta and gamma tocopherols (Alpha, Beta, Gamma) and the sum of alpha3, beta3, gamma3 tocotrienols (Alpha3, Beta3, Gamma3), respectively. The contents of all these compounds were determined using the AOCS Official Ce 8-89 method (2017d) and are expressed in ppm. The carotene contents (Car; [ppm]) were determined according to the PORIM p2.6 method (Siew and Tang 1995).

Saphiro Wilk tests were applied in order to check for non-normally distributed trait data. The traits that showed a significant deviation were normalized by z-score correction and the normalized data were used for further analyses. For each trait an ANOVA was performed for evaluating the effect of the origin of the different accessions on the analysed trait. Separation of means was performed using Tukey post hoc tests. All analyses were computed using R language.

2.3. RNA extraction and library construction

Library construction, including mRNA isolation and cDNA synthesis, was based on the cDNA-AFLP methodology of Bachem et al. (1998), using double digested restriction fragments of ds-cDNA. Barcoded amplicons were generated from each genotype and size selection of the final amplification products was performed. All adapters and primers used in this study are shown in Table 17.

Samples of young leaflet tissue were used for total RNA extractions with the Plant/Fungi RNA Purification Kit (Norgen, Thorold, ON, Canada), according to the manufacturer instructions. Each sample was then purified by DNAsa RNA clean & Concentrator™ columns (Zymo Research, USA) and resuspended in 25 µL DNase/RNase-free H_2O for mRNA isolation, following the Bachem et al. (1998) protocol.

For single strand cDNA synthesis 21 µL of mRNA, 6 µL of reaction mix and 3 µL of enzyme from the SuperScript™ VILO™ cDNA Synthesis Kit (Invitrogen™) were mixed. Samples were incubated at 25 °C for 10 minutes, followed by 42 °C for 1 hour and 85 °C for 5 minutes. Double stranded cDNA was obtained with a mix of 15 µL 10 x cDNA Buffer from the Bachem et al. protocol, 3.5 µL of DNA Polymerase I (Invitrogen™), 1.5 µL of RNASEe H (Invitrogen™), 1 µL of dNTP (25mM) and 98 µL of DNase/RNase-Free H_2O for each 119 µL of cDNA. This mix was incubated for 2 hours at 16 °C.

Purification of the obtained ds-cDNA was performed by adding one volume of Phenol:Chloroform:Isoamylalcohol (25:24:1). The mix was gently shaken and centrifuged for 10 minutes at 12000 g. The upper phase was recovered and two volumes of cold absolute etOH was added and incubated overnight at -20 °C. Samples were centrifuged for 30 minutes at 4 °C and 14000 g. The obtained pellets were dried and resuspended in 23 µL of DNase/RNase-free H_2O. One µL of this product was used to quantify cDNA with a Qubit® 2.0 Fluorometer (Thermo Fisher, Carlsbad, CA, USA) using the Qubit™ RNA HS Assay Kit.

Digestion of the cDNA was performed by double-digestion using AseI (FastDigest VspI, Thermo Scientific™) and TaqI (FastDigest TaqI, Thermo Scientific™). The digestion conditions were according to the manufacturer instructions. Afterwards, top and bottom adapters (TaqI Adapter; AseI Adapter, Table 17; No. 1a, 1b) with specific overhangs for the restriction fragments and a complementary part of the Illumina sequencing primers were ligated to the restriction fragments. The top adapter of AseI was biotinylated at the 5´ end. The reaction mix was prepared as follows: 0.5 µL of 10 x Fast Digest Buffer Thermo Scientific™, 1 µL of a mix of top and bottom TaqI adapters (50 µM), 1 µL of a mix of top and bottom AseI adapters (5 µM),

1 µL of ATP (10 mM), 1 µL of T4 DNA Ligase (1U/µL) (Invitrogen) and 0.5 µL nuclease free water. The products of this ligation reaction were purified with Ampure beads (Agencourt AMPure XP Kit, Beckman Coulter, Nyon, Switzerland) following manufacturer instructions.

AseI-AseI and AseI-TaqI biotinylated ligation products were then collected using streptavidin coated magnetic beads (Dynabeads™ M-280 Streptavidin), removing in this way the TaqI-TaqI fragments.

A first PCR was performed for amplifying the captured ligation products. The PCR mix consisted of 4 µL of the captured ligation product, 2.5 µL of 10 x PCR Buffer (Bioron, Römerberg, Germany), 2 µL dNTP (2.5 mM), 0.2 µL of AseI-Ilu (100 ng/µL) primer (Table 17, No 2a), 2 µL dNTP (2.5 mM), 0.2 µL of TaqI-Ilu (100 ng/µL) primer (Table 17, No 2b), 0.1 µL of Taq Polymerase (Bioron) and 16 µL of nuclease free water. The thermocycler conditions were denaturation at 94 °C for 5 minutes, followed by 30 cycles of 94 °C for 10 s, 56 °C for 30 s, 72 °C for 60 s, and a final elongation step of 72 °C for 10 minutes. All PCR reactions were performed in a Thermal Cycler ABI 2720 (Applied Biosystems, Foster, USA).

A second PCR was performed for indexing the genotypes with unique barcodes. Illumina indexes (Table 17, No. 4) were used in order to unambiguously identify each genotype. The PCR mix for each genotype was composed of 2.5 µL 10 x PCR Buffer (Bioron), 2 µL of dNTP (2.5 mM), 0.2 µL of P5 (10 µM) primer (Table 17, No 3a), 0.2 µL of P7 (10 µM) primer (Table 17, No 3b), 0.1 µL of Taq Polymerase (Bioron), 18 µL of nuclease free water and 2 µL of the previous enrichment reaction. PCR conditions were the same as in the first PCR reaction.

The reactions of all indexed genotypes were combined in one pool. This pool was migrated in a 2% agarose gel (1 x TAE) in order to choose the desired sequence lengths. Agarose was cut between 225-380 bp and the gel slice was cleaned with NucleoSpin® Gel and PCR Clean Kit (MACHEREY-NAGEL GmbH & Co., Germany).

The quality of the library was verified on an Agilent 2100 Bioanalyzer using DNA Chips with HS DNA Kit reagents, according to the manufacturer's protocol (Agilent Technologies). The library was sent for sequencing to StarSEQ GmbH (Germany), using the Illumina MiSeq sequencer (250 bp, paired-end reads). Three runs were performed with aliquots of the same library.

Table 17: Ligation adapters, amplification primers, Illumina primers and index sequences used for generating barcoded amplicons of restriction fragments.

No	Name	Primer Sequence
1a	TaqI Adapter	Top: 5' CGGATCGGAAGAGCA 3' Bottom: 5' GTGTGCTCTTCCGATC 3'
1b	AseI Adapter	Top: 5' [BIO] CGACGCTCTTCCGATCTTAAT 3' Bottom: 5´GAGAAGGCTAGAATTA 3´
2a	AseI-Ilu	5' ACACGACGCTCTTCCGATCTTA 3`
2b	TaqI-Ilu	5' AGACGTGTGCTCTTCCGATC 3`
3a	Illumina P5	5'AATGATACGGCGACCACCGAGATCTACAC[SA5ii]ACACTCTTTCCCTAC ACGACGCTCTTCCGATCT 3'
3b	Illumina P7	5'CAAGCAGAAGACGGCATACGAGAT[SA7ii]GTGACTGGAGTTCAGACGT GTGCTCTTCCGATC 3`

4 Illumina barcode indexes [5' > 3']

	SA5ii	**SA7ii**		**SB5ii**	**SB7ii**
01	ATCGTACG	AACTCTCG	01	CTACTATA	AAGTCGAG
02	ACTATCTG	ACTATGTC	02		ATACTTCG
03	TAGCGAGT	AGTAGCGT	03		AGCTGCTA
04	CTGCGTGT	CAGTGAGT	04		CATAGAGA
05	TCATCGAG	CGTACTCA	05		CGTAGATC
06	CGTGAGTG	CTACGCAG	06		CTCGTTAC
07	GGATATCT	GGAGACTA	07		GCGCACGT
08	GACACCGT	GTCGCTCG	08		GGTACTAT
09		GTCGTAGT			
10		TAGCAGAC			
11		TCATAGAC			
12		TCGCTATA			

2.4. Sequence processing and Association analysis

In order to get an impression of the expected distribution of restriction fragments using AseI and TaqI for digestion, an *in silico* study was conducted using the "fragmatic" script (Chafin 2016) with published oil palm cDNA from Malaysian Oil Palm Board (MPOB) (2010). The sequences of the expected restriction fragments were blasted against the annotated oil palm gene database from MPOB using BLAST software (Altschul et al. 1990) in order to count the number of targeted genes.

Analyses of the sequences obtained from the sequencing platform were performed using the South Green Bioinformatics Platform (South Green Collaborators 2016), which provides different bioinfomatic tools and methods for sequence analysis.

Each of the three runs of the library yielded forward and reverse fastq files for each genotype on the Illumina MiSeq platform. Files of the same genotype of each run were concatenated in order to obtain two final forward and reverse fastq files per genotype. These files were processed by the "Snakemake-capture" script (Soriano et al. 2018) of the South Green bioinformatics platform. "Cutadapt" (Martin 2011) was used to clean the reads and for removing the adapters and common sequence parts of the reads. Sequences were filtered for quality below q < 10 and sequences shorter than 35 pair bases were discarded. "BWA" (Li and Durbin 2010) was used to map the reads, "Samtools" (Li et al. 2009) to clean the alignments, "Picard-tools" (Broad Institue 2015) to sort the reads and "GATK haplotype caller" (McKenna et al. 2010) to call the variants using the Variant Calling Format (VCF). The MPOB *E. guineensis pisifera* genome sequence (Singh et al. 2013b) was used as reference.

In order to count the number of genes which actually were targeted by the obtained sequences, "snpEff" software (Cingolani et al. 2012) was used with the raw vcf files from the "Snakemake-capture" workflow and the annotated gene database from MPOB as reference.

The obtained variants from the VCF file were filtered again using "VCFtools" software (Danecek et al. 2011). Markers were filtered for only biallelic SNP with a minimum allele frequency of 0.05 and a maximum of 0.95. Markers below q < 25 and depth < 6 were eliminated as well as indels. Variants with more than 30% of missing data were also discarded to obtain the first VCF file (VCF_1). Missing values of remaining markers were imputed using "LinkImputeR" software (Money et al. 2017). This final VCF file (VCF_2) was used for performing the genotype-phenotype association studies.

Genetic diversity study in our germplasm was performed with the genotypic data of the VCF_1 file. Diversity was studied in terms of expected (He) and observed heterozygosis (Ho) of the markers using the "adegenet" (Jombart 2008) and "hierfstat" (Goudet and Jombart 2015) R packages following Nei´s statistics (Nei 1987). The "Ho" value was calculated as follows:

$$\mathrm{H_o} = 1 - \sum_{\mathrm{k}} \sum_{\mathrm{i}} \mathrm{Pkii/np}$$

where Pkii is the homozygote proportion i in sample k and np the number of samples. The "He" value was calculated using the following equation:

$$H_e = \tilde{n}/(\tilde{n}-1)\left[1 - \sum_i \bar{p_i^2} - H_o/2\tilde{n}\right]$$

where $\tilde{n} = np/\sum_k 1/n_k$ and $\bar{p_i^2} = \sum_k p_{ki}^2/np$.

For studying genetic variances between and within origins, Fixation Indices (Fst) obtained from "VCFtools" and Inbreeding coefficients (Fis) obtained from the "hierfstat" package were computed.

The VCF_1 file was also used to obtain the structure (Q), Principal Component Analysis (PCA) and IBS Kinship matrixes (K) for the association studies. "fastStructure" software (Raj et al. 2014) was used to analyse the population structure and the "chooseK.py" script to determine the population structure matrix (Q). Allele frequencies of each cluster from 1 to 9 were estimated with a 10 fold cross-validation (CV). The PCA matrix was obtained from TASSEL (version 5.2.44) (Bradbury et al. 2007) and the K matrix from GAPIT software (Wang and Zhang 2018) (version 3.0). Darwin Software (Perrier, X., Jacquemoud-Collet 2006) was used to draw a dendogram reflecting the population structure which was derived from the "identity by state" (IBS) distance matrix from TASSEL and using the nearest neighbour clustering method.

Association studies were performed using GAPIT software. Fixed effects linear models (GLM) and mixed linear models (MLM) were applied to test associations between segregating markers and phenotype. For GLM, either the population structure matrix Q or the PCA matrix was used as covariate (GLM_Q; GLM_PCA). For MLM the Kinship matrix (K) was incorporated into the models with either Q matrix or PCA matrix, in order to reflect the relationships among individuals (MLM_Q+K; MLM_PCA+K). The percentage of variation (%VA) explained by each SNP for each trait was calculated as the difference between the %VA of the model with and without the SNP.

The resulting observed and expected *p* values of each model were visualized separately for each trait in a Quantile-Quantile (QQ) plot, in order to get a first impression on the fitting of different alternative models. In addition, the average squared distances of the data points from the diagonal (d^2) of the QQ plot was computed for each model and trait, based on the Equation 3.

The observed *p* values were sorted by chromosome and location for each trait. For all significant SNP within a 25 kb distance, only one single SNP with smallest *p* value was chosen

as representative. Associations with unadjusted false discovery rate (FDR) *p* values below 0.01 and FDR adjusted *p* values below 0.05, respectively, were considered as true. Sequences with significant SNP were blasted against the nucleotide database at NCBI (Geer et al. 2009) using BLASTN optimized for highly similar sequences (megablast), in order to reveal the biological meaning of the underlying genes.

In order to visualize in a comprehensive way the different steps described above under the Materials and Methods section, we present in Figure 20 a scheme which summarizes the overall experimental approach.

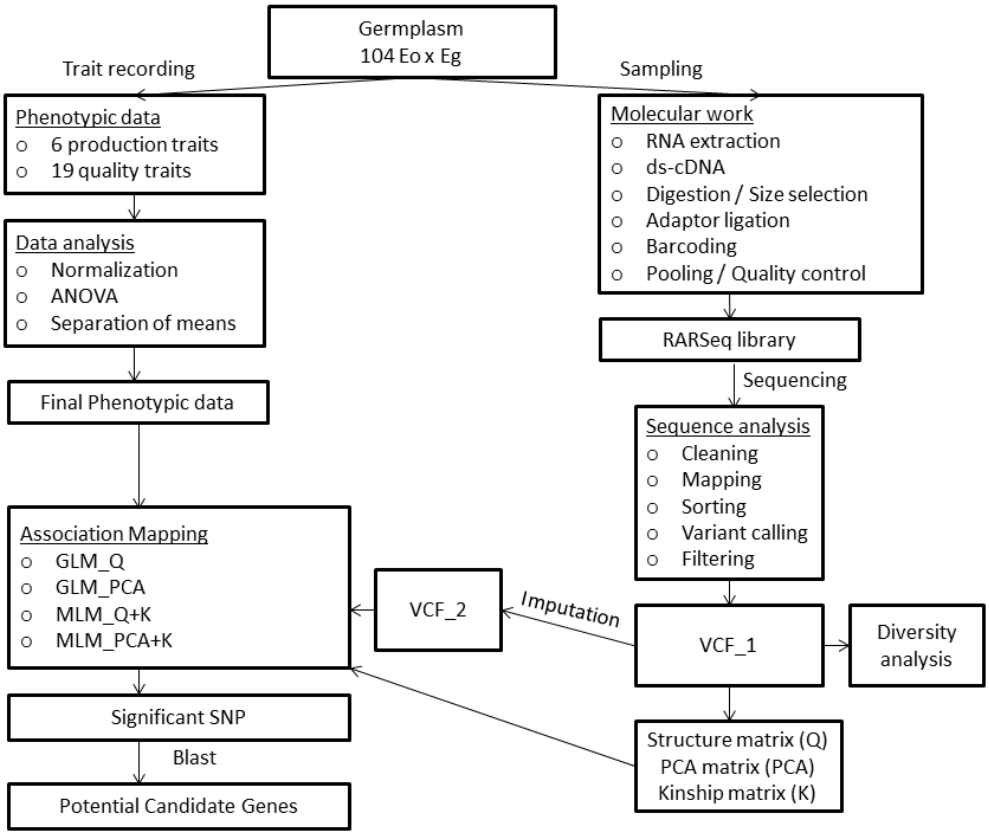


Figure 20: Overall scheme of the procedure.

3. Results

3.1. Phenotypic data analysis

Initial Saphiro-Wilk tests revealed that 15 of the 25 traits were not normally distributed. They are marked with "*" in Table 18. Transformed data were used for processing non-normal distributed traits.

Table 18: Mean values, standard deviations (SD), minimum and maximum values of each analysed trait, and ANOVA significance levels between the different origins of oil palm hybrids.

Production traits	**Mean**	**SD**	**Max**	**Min**	**ANOVA**
BN [nº]*	49.087	20.437	82	1	***
BW [Kg]	11.070	3.219	17.408	3.428	***
BY [Kg]	569.767	313.573	1280.4	10.5	***
OilfM [%]	29.045	5.769	45.177	15.3	***
OildM [%]*	53.168	8.697	77.68	35.67	***
OilB [%]	18.475	4.748	29.181	5.71	***
Oil quality traits	**Mean**	**SD**	**Max**	**Min**	**ANOVA**
Sat [%]*	37.476	4.445	45.26	20.07	***
Mono-Un [%]*	48.512	5.295	65.15	37.46	***
Poly-Un [%]	13.588	1.683	17.33	9.93	**
OA [%]*	47.198	5.872	65.2	35	***
IV [cg/g]*	64.305	4.534	81.39	54.62	***
SSS [%]*	1.229	0.707	3.98	0.098	*
SUS [%]*	23.659	4.486	33.2	8.925	***
SUU [%]	32.195	4.177	45.224	21.4	*
UUU [%]*	12.813	5.888	31.95	3.34	***
Toc [ppm]	1355.682	436.397	2157.5	392.9	**
Tocph [ppm]*	213.579	93.472	475.3	18.3	*
Alpha [ppm]*	152.905	71.827	387.7	18.3	*
Delta [ppm]*	42.078	15.346	89.2	10.5	*
Gamma [ppm]*	45.182	10.519	73	30.8	
Toc3 [ppm]	1142.103	374.765	1882	306.8	***
Alpha3 [ppm]	325.742	136.269	651	55.3	***
Delta3 [ppm]*	105.516	55.897	262.9	18.6	***
Gamma3 [ppm]	711.945	210.910	1172.1	211.7	*
Car [ppm]*	800.023	245.045	1469	353	

Significance levels: $p < 0.001$***; $p < 0.01$** and $p < 0.05$*. Traits marked with "*" revealed non-normally distributed data in Saphiro-Wilk test.

The ANOVA results for testing the influence of origins on the traits are presented in Table 18. Observed mean values, standard deviations (SD), minimum and maximum values and the significance levels of the F tests are shown for each analysed trait. All production traits showed significant differences at significance level $p < 0.001$, as well as nine quality traits. Two quality traits, Poly-Un and Toc, showed significant differences < 0.01 and six quality traits < 0.05. Only the Car and Gamma traits did not reveal significant differences between origins.

The results of the Tukey post hoc tests for separation of means are presented in Table 19. Production traits OilfM, OildM and OilB revealed large values for the Coari x La Mé origin,

while the Taisha x Ekona genotypes showed the lowest values for all production traits. On the other hand, Taisha x Avros (Oleoflores) revealed the highest values for BN, BY and BW traits. For quality traits also a large difference was detected between Coari x La Mé and the other four origins. The Coari x La Mé origin showed significant higher values for Mono-Un, OA, IV, SUU or UUU, but significant lower values than the other origins for the traits Sat, Poly-Un, SUS and different tocol compounds.

Table 19: Mean values of the studied traits for each of the accessions and significance levels obtained by Tukey post hoc tests.

Origin	**Coari x La Mé**		**Taisha x Avros (RGS)**		**Taisha x Avros (Oleoflores)**		**Taisha x Ekona**		**Taisha x Yangambi**	
Production traits	Mean value	Level	Mean value	Level	Mean value	Level	Mean value	Level	Mean value	Level
BN [nº]	54.62	AB	45.22	B	61.40	A	29.13	C	41.08	BC
BY [Kg]	497.76	BC	538.26	B	822.16	A	292.28	C	464.02	BC
BW [Kg]	9.14	C	11.72	AB	13.08	A	9.25	BC	10.36	BC
OilfM [%]	33.56	A	28.63	B	28.49	B	22.66	C	29.02	AB
OildM [%]	62.98	A	50.80	B	52.43	B	42.96	C	51.71	B
OilB [%]	22.49	A	16.93	B	18.91	B	12.15	C	18.19	B
Oil quality traits	Mean value	Level	Mean value	Level	Mean value	Level	Mean value	Level	Mean value	Level
Sat [%]	33.54	B	38.12	A	38.51	A	39.00	A	38.78	A
Mono-Un [%]	53.36	A	48.59	B	46.30	B	47.02	B	47.05	B
Poly-Un [%]	12.85	B	13.01	B	14.37	A	13.73	AB	13.92	AB
OA [%]	53.62	A	47.02	B	44.65	B	44.97	B	45.45	B
IV [cg/g]	68.35	A	63.55	B	63.42	B	62.69	B	62.75	B
SSS [%]	1.27	AB	1.58	A	0.97	B	1.03	AB	1.51	AB
SUS [%]	18.80	B	24.12	A	25.04	A	26.18	A	24.38	A
SUU [%]	34.12	A	31.82	AB	31.64	AB	32.56	AB	30.00	B
UUU [%]	19.58	A	12.18	B	10.38	B	10.18	B	11.82	B
Tocph [ppm]	154.68	B	227.67	AB	213.78	AB	264.37	A	215.03	AB
Alpha [ppm]	109.62	B	168.26	AB	144.06	AB	183.34	A	172.09	AB
Delta [ppm]	40.91	AB	41.19	AB	44.24	AB	50.80	A	30.25	B
Gamma [ppm]	49.40	-	43.19	-	45.54	-	47.72	-	42.54	-
Toc3 [ppm]	914.08	A	1074.54	AB	1310.15	A	1292.49	A	1081.15	AB
Alpha3 [ppm]	223.37	B	313.40	AB	400.33	A	346.09	A	318.05	AB
Delta3 [ppm]	68.38	B	98.75	AB	134.27	A	122.35	A	87.33	AB
Gamma3 [ppm]	626.13	B	662.39	AB	775.54	AB	824.06	A	675.78	AB
Toc [ppm]	1068.77	B	1302.21	AB	1523.92	A	1556.86	A	1296.18	AB
Car [ppm]	845.05	-	804.56	-	691.26	-	836.27	-	904.08	-

* Means with the same letter are not statistically different ($\alpha > 0.05$).

3.2. *In silico* digestion assay

An *in silico* assay of the expected AseI-TaqI restriction fragments was performed using the "fragmatic" script with the oil palm cDNA sequences from MPOB (2010). The expected restriction fragment frequencies are displayed in Figure 21. Our library was cut in an agarose gel between 225-380 bp with an average value of 310 bp, leading to an average insert size of 169 bp after discarding primer and adapter sequences. For this length a total of 38,282 restriction fragments are expected. TaqI-TaqI fragments predominate with 34,098 fragments (89 %), while 3,616 AseI-TaqI tags (9.5 %) and only 568 AseI-AseI fragments (1.5 %) are expected. In total, 2.11% of the translated sequences were supposed to be targeted in this assay which corresponds to 2,190 annotated genes (8.32%) from the MPOB database.

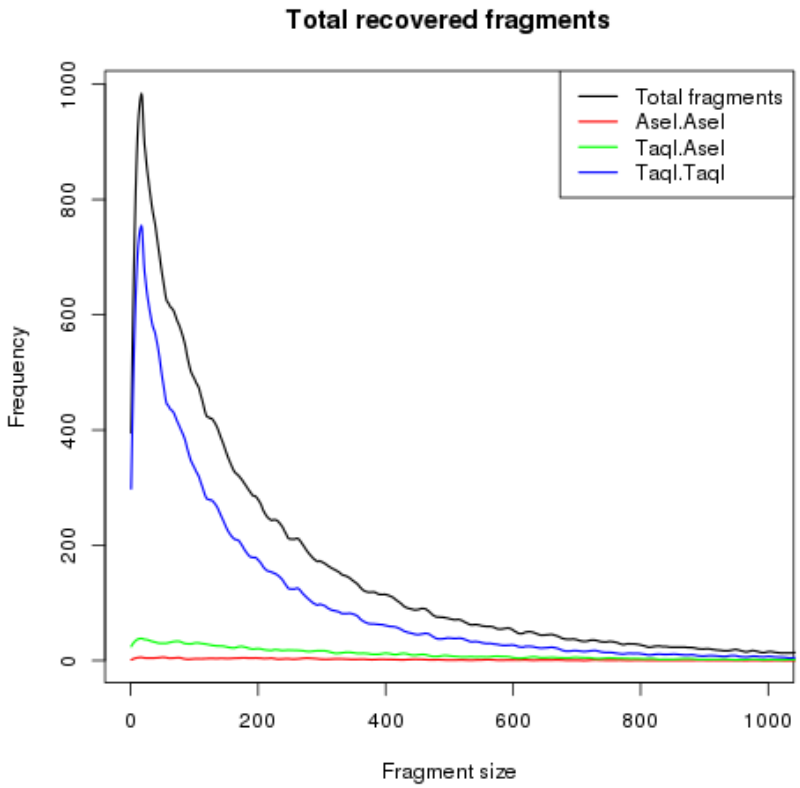


Figure 21: Frequency of fragment sizes derived from the in silico assay of double enzyme digestion (AseI, TaqI) using the Oil Palm cDNA from MPOB.

3.3. Sequencing and filtering results

Three sequencing runs were performed with the purified library on an Illumina MiSeq platform. The first and second runs yielded 18.5 million reads each, and the third run obtained 20.5 million raw reads. After the cleaning steps by Cutadapt software, 31.6 million clean reads remained for the mapping steps. A total of 97.8 % of the reads mapped to the *Elaeis guineensis*

genome (Singh et al. 2013b). Initially the “Snakemake-capture” workflow identified 2,992,736 million variants. The SnpEff software determined that these SNP targeted a total of 575 genes of the MPOB annotated gene database (2.18 % of all genes. After the first filtering step 599,754 SNP were left. However, after filtering for 30 % missing values only 310 SNP remained.

The “fastStructure” software determined that seven sub-populations (K=7) exist in our germplasm using the VCF file without imputing missing values (VCF_1). This parameter was used in further genotype-phenotype analysis.

The VCF_1 file was also used to perform the genetic diversity study. The results are shown in Table 20. The average observed (Ho) and expected heterozygosity (He) values were 0.346 and 0.288, respectively. Bartlett´s test detected a significant difference between expected and observed heterozygosity. Fst values revealed no discriminant differentiation between populations as shown in Table 20, since all values were close to zero or negative.

Table 20: Genetic diversity studies in terms of inter cross Fixation indices (Fst) and intra cross Inbreeding coefficients (Fis).

Inter-cross Fst value	**Taisha x Yangambi**	**Taisha x Ekona**	**Taisha x Avros (Oleoflores)**	**Taisha x Avros (RGS)**	**Coari x La Mé**
Taisha x Yangambi	-	-0.00266	0.00137	-0.00159	0.06903
Taisha x Ekona	-	-	0.01290	0.00920	0.06343
Taisha x Avros (Oleoflores)	-	-	-	0.00913	0.06557
Taisha x Avros (RGS)	-	-	-	-	0.07383
Intra-cross Fis value	-0.15329	-0.12511	-0.17101	-0.22990	-0.02630

The largest distances between origins were always observed between Coari x La Mé and the other origins involving Taisha. These origins showed always much smaller distances between them. The Fis values reported in Table 20 revealed no relatedness between individuals of the same origin, since all obtained values were negative, suggesting a high diversity within origins. The dendogram in Figure 22 reflects these relationships between the different origins. Eighteen of the 21 Coari x La Mé origins cluster apart from the other accessions involving Taisha, which are more intermixed.

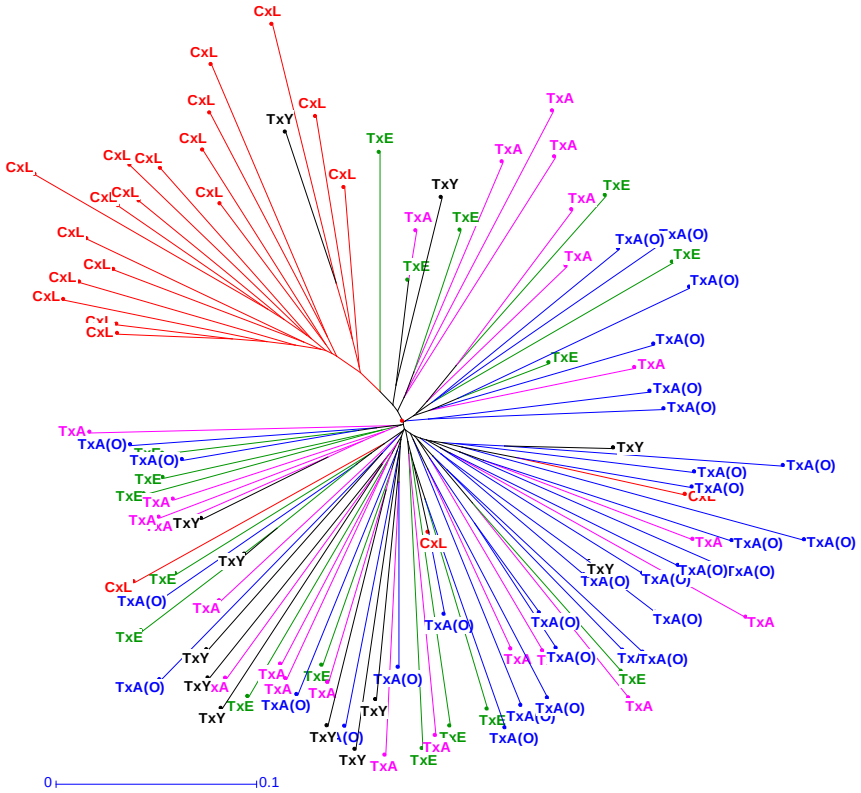


Figure 22: Dendogram derived from IBS (identity by state) distance matrix in Tassel and using the nearest neighbour clustering method where CxL: Coari x La Mé is in red, TxA: Taisha x Avros (RGS) is in pink, TxE: Taisha x Ekona is in green, TxY: Taisha x Yangambi is in black and TxA(O): Taisha x Avros (Oleoflores) is in blue.

3.4. Association mapping results

The VCF_1 file was filtered for a call rate of 30% and missing values were imputed using LinkImputeR software (VCF_2). After performing Association Mapping using GAPIT, expected and observed *p* values of each model and trait were drawn in QQ plots to get a first impression of the fitting of alternative models. The average distance of the data points from the diagonal (d^2) explained in Materials and Methods was used to determine for each trait the particular model which fitted best based on the smallest d^2 value. The results are shown in Table 21. For production traits only OilB fitted best with a GLM_Q model, while the other four traits fitted best with mixed models; OilfM with MLM_Q+K and BN, BY and BW with MLM_PCA+K. For quality traits 12 out of 19 traits, revealed better fitting with mixed models, but seven traits (Mono-Un, Poly-Un, OA, SUU, UUU, Gamma, Toc) showed better results with generalized linear models.

Table 21: Average square distance (d^2) values of the CG data points from the diagonal of the QQ plot for determining the best fitting model for each trait.

Production traits	GLM_Q	GLM_PCA	MLM_Q+K	MLM_PCA+K
BN [no]	0.382	0.368	0.376	**0.355**
BY [Kg]	0.375	0.352	0.363	**0.339**
BW [Kg]	0.338	0.315	0.350	**0.314**
OilfM [%]	0.327	0.330	**0.326**	0.333
OildM [%]	0.353	0.337	0.343	**0.336**
OilB [%]	**0.332**	0.335	0.333	0.335
Oil quality traits				
Sat [%]	0.550	0.689	**0.515**	0.620
Mono-Un [%]	**0.499**	0.688	0.646	0.566
Poly-Un [%]	**0.348**	0.381	0.354	0.376
OA [%]	**0.589**	0.847	0.641	0.768
IV [cg/g]	0.584	0.485	0.477	**0.465**
SSS [%]	0.385	0.371	0.371	**0.369**
SUS [%]	0.572	0.725	**0.566**	0.608
SUU [%]	**0.367**	0.410	0.382	0.410
UUU [%]	**0.590**	0.786	0.599	0.691
Tocph [ppm]	0.355	0.337	0.357	**0.329**
Alpha [ppm]	0.342	0.325	0.344	**0.321**
Delta [ppm]	0.357	0.346	0.357	**0.337**
Gamma [ppm]	**0.326**	0.350	0.328	0.350
Toc3 [ppm]	0.389	0.389	**0.388**	0.389
Alpha3 [ppm]	0.370	0.365	**0.364**	0.365
Delta3 [ppm]	0.389	0.389	**0.383**	0.389
Gamma3 [ppm]	0.390	0.387	0.390	**0.386**
Toc [ppm]	**0.383**	0.384	0.384	0.384
Car [ppm]	0.400	0.391	0.395	**0.390**

The best fitting model with smallest d^2 value is indicated in bold and underlined for each CG.

Table 22 presents a summary of the detected significant associations between SNP and production and oil quality traits in oil palm hybrids. The particular model, FDR adjusted and unadjusted *p* values, the explained variance, the effect of allele substitution and the reference number and biological meaning of the SNP as retrieved from NCBI are indicated for each SNP. Under unadjusted conditions ($p < 0.01$) five production traits and 15 quality traits showed potential associations with 24 SNP representing 23 loci. The traits IV, SUS, UUU, OA, Sat and Mono-Un revealed the highest number of associations with different SNP, ranging between 6 and 12, while the traits Tocph, Alpha, Delta, Gamma, Delta3, Toc3, Toc, SUU, Poly-Un, BY, BN, BW, OilfM and OildM showed only one or two significant associations. In total 78 potential associations were detected. The explained variances by the model ranged from 5% to 16.6%.

Blast searches against the nucleotide database from NCBI revealed significant homologies for 11 out of the 23 loci with *Elaeis* mRNA sequences having a biological meaning, while five SNP showed significant homologies with mRNA sequences of the closely related *Phoenix dactylifera* species. For seven SNP no significant homologies were detected (Table 22).

When applying FDR multiple testing ($p < 0.05$), only six quality traits showed significant associations with eight SNP. The trait UUU trait revealed seven associations, OA six associations, SUS five associations, Sat three associations and the traits IV and Mono-Un two significant associations. In total 25 promising associations were detected. Six of the SNP revealed annotated functions in *Elaeis,* while two SNP revealed no annotated functions (Table 22).

Table 22: Significant associations between SNP and production and oil quality traits in Oil palm hybrids.

SNP	**Trait**	***p* value**	**FDR Adjusted *p* value**	**%VA**	**Effect**	**Model**	**Annotation / Biological Function**
C01: 2788341	IV	0.000112	**0.017**	12.60	-4.413	MLM_PCA	NM_001304427.1: holocarboxylase synthetase [HCS], mRNA
	SUS	0.000144	**0.010**	12.00	4.750	MLM_Q	
	UUU	0.000315	**0.016**	10.50	-5.811	GLM_Q	
	OA	0.000479	**0.030**	9.40	-5.345	GLM_Q	
	Sat	0.001100	0.075	8.60	3.852	MLM_Q	
C01: 34080563	OA	0.003380	0.116	6.50	-3.404	GLM_Q	XM_010934437.2: thioredoxin H1 [LOC105053337], mRNA
	Mono-Un	0.007040	0.273	5.80	-2.890	GLM_Q	
	UUU	0.007230	0.198	5.60	-3.112	GLM_Q	
	Sat	0.009990	0.258	5.60	2.314	MLM_Q	
C01: 49567798	SUU	0.000564	0.175	13.10	-8.732	GLM_Q	XM_010942156.3: fructose-bisphosphate aldolase 1, cytoplasmic [LOC105059022], mRNA
	SUS	0.002000	0.078	8.10	7.211	MLM_Q	
	UUU	0.000055	**0.004**	13.50	-4.570	GLM_Q	
	OA	0.000223	**0.017**	10.60	-4.103	GLM_Q	
	Sat	0.000389	**0.040**	10.30	3.058	MLM_Q	
	IV	0.001770	0.110	8.10	-2.694	MLM_PCA	
	Delta 3	0.005360	0.676	7.80	33.339	MLM_Q	
	Toc	0.005860	0.685	7.60	253.760	GLM_Q	
	Toc3	0.007770	0.684	7.10	211.116	MLM_Q	
C02: 17522364	OA	0.000053	**0.005**	13.00	-8.316	GLM_Q	XM_010913829.3: PTI1-like tyrosine-protein kinase 1 [LOC105038129], mRNA
	IV	0.000059	**0.017**	13.80	-6.331	MLM_PCA	
	UUU	0.000067	**0.004**	13.20	-8.562	GLM_Q	
	SUS	0.000324	**0.017**	10.90	5.947	MLM_Q	
	Mono-Un	0.006370	0.273	5.90	-5.073	GLM_Q	

	Sat	0.008170	0.258	5.40	4.111	MLM_Q	
C03: 58738016	OA	0.000021	**0.003**	14.50	-4.911	GLM_Q	No annotation
	UUU	0.000068	**0.004**	13.10	-4.760	GLM_Q	
	SUS	0.000106	**0.010**	12.60	3.566	MLM_Q	
	Sat	0.006750	0.258	5.80	2.340	MLM_Q	
	Mono-Un	0.009960	0.343	5.30	-2.668	GLM_Q	
C05: 38969899	Sat	0.008470	0.258	5.00	-1.696	MLM_Q	XM_010923961.2: chaperone protein DnaJ (LOC105045611), mRNA
	BY	0.008190	0.917	7.30	-181.230	MLM_PCA	
	BN	0.008630	0.779	7.20	-11.671	MLM_PCA	
C07: 21833956	Oild M	0.004820	0.960	7.20	-3.646	MLM_PCA	XM_010928354.3 : 40S ribosomal protein S5 [LOC105048878], mRNA
C10: 26013376	UUU	0.005130	0.159	6.20	2.357	GLM_Q	XM_010934283.2: ADP-ribosylation factor-like protein 8c (LOC105053201), mRNA
	SUS	0.007390	0.232	6.00	-1.743	MLM_Q	
C11: 22661050	Mono-Un	0.002490	0.154	7.30	6.080	GLM_Q	XM_010935353.2 : uncharacterized LOC105053988, transcript variant X4, mRNA
	Sat	0.003080	0.159	6.40	-4.959	MLM_Q	
C15: 20356039	OA	0.000006	**0.002**	16.60	-10.324	GLM_Q	XM_010941582.3 : oligouridylate-binding protein 1B [LOC105058606], transcript variant X4, mRNA
	UUU	0.000028	**0.004**	14.70	-9.911	GLM_Q	
	SUS	0.000063	**0.010**	13.90	7.357	MLM_Q	
	Mono-Un	0.000146	**0.023**	11.90	-7.904	GLM_Q	
	Sat	0.000287	**0.040**	12.40	6.479	MLM_Q	
	IV	0.000917	0.071	9.10	-5.843	MLM_PCA	
	OilfM	0.008920	0.997	7.10	-6.478	MLM_Q	
C16: 2055062	UUU	0.008310	0.198	5.40	2.212	GLM_Q	XM_010942259.3 : protein SRC2 [LOC105059087], mRNA
CU05: 76475493	UUU	0.000022	**0.004**	15.10	-9.951	GLM_Q	No annotation
	SUS	0.000167	**0.010**	12.60	6.731	MLM_Q	
	OA	0.000644	**0.033**	9.00	-7.195	GLM_Q	
	SUU	0.005000	0.640	8.50	-5.268	GLM_Q	
CU06: 31292942	SUS	0.008220	0.232	5.50	2.483	MLM_Q	No annotation
	UUU	0.008300	0.198	5.40	-3.154	GLM_Q	
CU06: 44960284	Mono-Un	0.000047	**0.014**	13.90	-6.695	GLM_Q	XM_010910976.3 : 4-hydroxy-3-methylbut-2-enyl diphosphate
	Sat	0.000052	**0.016**	15.10	5.641	MLM_Q	
	IV	0.000775	0.071	9.40	-4.505	MLM_PCA	

	OA	0.002010	0.078	7.30	-5.372	GLM_Q	reductase, chloroplastic [LOC105035427], transcript variant X2, mRNA
	UUU	0.003460	0.134	6.80	-5.044	GLM_Q	
	SUS	0.008080	0.232	5.10	3.543	MLM_Q	
CU07: 25160367	Tocph	0.003180	0.492	9.30	-101.879	MLM_PCA	No annotation
	Gamma	0.004460	0.988	14.80	-12.008	GLM_Q	
	Alpha	0.008190	0.981	7.40	-70.616	MLM_PCA	
CU07: 9224455	Delta	0.009410	0.946	8.40	-14.282	MLM_PCA	MH681000.1: *Phoenix dactylifera* clone dpBGPATlike sex-determination region sequence
CU08: 33353724	UUU	0.000463	**0.021**	9.90	-7.619	GLM_Q	MH680999.1: *Phoenix dactylifera* clone dpBCYPlike sex-determination region sequence
	SUS	0.001300	0.057	7.90	5.328	MLM_Q	
	Toc3	0.003310	0.684	8.70	451.481	MLM_Q	
	Mono-Un	0.003800	0.197	6.70	-5.790	GLM_Q	
	OA	0.005460	0.154	5.80	-5.988	GLM_Q	
	Toc	0.006830	0.685	7.30	483.411	GLM_Q	
CU09: 20495654	Alpha	0.000523	0.162	13.10	-78.420	MLM_PCA	MH680999.1: *Phoenix dactylifera* clone dpBCYPlike sex-determination region sequence
	Tocph	0.000947	0.294	11.90	-95.528	MLM_PCA	
	BN	0.009870	0.779	6.90	13.927	MLM_PCA	
CU09: 20495700	Poly-Un	0.004500	0.885	8.00	0.812	GLM_Q	
CU10: 38760019	OA	0.006670	0.172	5.50	-1.750	GLM_Q	MH681002.1: *Phoenix dactylifera* clone dpB2Y sex-determination region sequence
CU10: 44270240	BW	0.006530	0.974	7.20	1.370	MLM_PCA	XM_008781472.1: *Phoenix dactylifera* uncharacterized LOC103699453, partial mRNA
CU10: 7083762	Sat	0.006470	0.258	5.80	1.284	MLM_Q	No annotation
CU10: 79084791	Sat	0.001220	0.075	8.90	4.562	MLM_Q	No annotation
	Mono-Un	0.002400	0.154	7.40	-5.126	GLM_Q	
	OA	0.009180	0.219	5.10	-4.715	GLM_Q	
CU10: 83647765	IV	0.008580	0.354	5.60	-4.260	MLM_PCA	No annotation

Legend: SNP: name of the SNP based on the MPOB genome location of the SNP; Trait: associated trait; *p* values: observed unadjusted (< 0.01) and FDR adjusted error probability values (< 0.05) for the model; %VA: explained variance; Effect: effect of the marker; AM Model: best fitting model for AM; Reference

number and biological function: reference number at NCBI and biological function based on blast searches at NCBI. SNP with significant FDR adjusted p values are indicated in bold.

4. Discussion

4.1. Phenotypic data analysis

The analysis of production traits revealed high values for Taisha x Avros (Oleoflores) with respect to BW, BY and BN, while the highest oil contents ratios were found for Coari x La Mé accessions. These results are in accordance with those of Pelaez et al. (2010) who observed that Coari palms as well as their hybrids with Eg had higher CO_2 fixation capacities, which are positively correlated with an increase in oil contents (Corley, R.H.V. and Tinker et al. 2016). On the other hand, Taisha palms have been described by Barba (2019) as "Oleifera Guineensis palms", since they have similar morphological characteristics as the Eg palms.

Considering the analyses of quality traits, studies are available from Montoya et al. (2013) and Singh et al. (2009). These authors analysed beside iodine value, particularly the fatty acid composition in interspecific hybrids from controlled crosses and established linkage maps with integrated QTL for these traits. Also Cadena et al .(2013) studied the lipase activity, oil contents in fresh mesocarp and iodine values in collections of Eg, Eo and Eo x Eg genotypes. However, we present here a detailed oil quality analyses for oil palm involving 19 different quality traits. Here we include traits related to lipids, where the saturation level of fatty acids were measured, considering the percentages of saturated (Sat), mono-unsaturated (Mono-un) and poly-unsaturated (Poly-un) fatty acids. The mono-unsaturated fatty acids are considered as the healthiest (Tierney and Roche 2007; Qian et al. 2016). We have also analysed the percentage of oleic acid in the oil (OA) which was classified as monounsaturated omega-9 fatty acid, the iodine value (IV) indicating the global degree of unsaturated fatty acids, and particularly the different types of triglycerides which can be formed from three fatty acids (SSS > SUS > SUU > UUU). We found large differences between Coari x La Mé and the other origins. Coari x La Mé accessions showed desirable characteristics such as high contents of mono-unsaturated acids, oleic acid, high iodine values and UUU triglycerides, while the saturated acid levels were significantly below those of the other origins. Pelaez et al. (2010) also determined higher oleic acid contents and iodine values in Coari palms.

We performed a detailed study for tocol contents which are composed of tocotrienols and tocopherols. These components represent different forms of vitamin E and can be found in oil palm as beneficial phytonutrients (Nesaretnam et al. 1995). Both, tocotrienols and tocopherols have four isomers each (α-, β-, γ-, δ-) and have unique benefits(Yuen May and Nesaretnam

2014). Here we studied three of them (α-, β-, γ-). In this case the Coari x La Mé accessions showed lower tocols contents. Finally, carotenoids contents were measured in the five origins. These pigments are responsible for the orange-red brilliant colour of the oil and are precursors of vitamin A (Yuen May and Nesaretnam 2014). However, for this trait no significant difference was observed between origins.

4.2. *In silico* digestion assay

AseI (AT↓TAAT) and TaqI (T↓CGA)) restriction enzymes have been widely used in different crops (LaO et al. 2008; Zhang et al. 2008; Xiao et al. 2017) for digestion in cDNA-AFLP assays. Therefore, we also selected these two enzymes to construct our library, and an *in silico* assay was performed to determine the expected restriction fragments using oil palm cDNA from MPOB as reference. Due to the nature of the enzymes, one being a common cutter (TaqI) and the other a rare cutter (AseI), the desired AseI-TaqI tags were in a much lower proportion compared to the TaqI-TaqI tags. Therefore, a biotinylated AseI adapter was used for the adapter ligation reaction of the restriction fragments. Afterwards, all AseI-TaqI and AseI-AseI tags were captured with streptavidin coated beads to get rid of the TaqI-TaqI tags. Biotinylated adapters are commonly used in single and dual enzyme restriction RADSeq libraries in order to capture the desired fragments (Krück et al. 2013; Matsumura et al. 2014; Marrano et al. 2017).

4.3. SNP detection and genetic diversity analysis

A total of 57.5 million reads were obtained from the three sequencing runs, and after applying the cleaning and quality steps, 55 % of the raw reads were left for the following analysis.

Almost 98 % of the reads could be mapped to the *Elaeis guineensis* (Eg) *pisifera* genome from MPOB, indicating that BWA represents efficient mapping software and our reference genome is suitable. In this context it is worth to mention, that currently no genome for *Elaeis oleifera* (Eo) x Eg hybrids is available.

Even though the “Snakemake-capture” workflow detected initially 2,992,736 million SNP, after applying the first filtering steps the number of variants were reduced to 599,754. Moreover, when filtering by a call rate of 30 % this number decreased drastically to 310 SNP.

Libraries based on enzyme digestion and following selection of a reduced representation of the genome are known to have large amounts of missing data as well as heterozygote undercalling, caused by individual genetic divergence or technical issues when preparing the libraries (Swarts et al. 2014; Brouard et al. 2017). Since we work with expression tags, it is expected to find differential gene expression between accessions, incrementing in this way the

missing call rates. With respect to the integrity of the RNA, Reiman et al. (2019) concluded, that even though RNA integrity is critical for estimating the expression level due to the degradation of samples, biologically meaningful analyses can be still performed on degraded RNA samples. Furthermore, Romero et al. (2014) concluded, that working with samples with RNA integrity number of four (RIN=4) is still possible, and in addition they added that approaches are more effective, when including all samples, regardless of quality.

Considering the global genetic analysis, the observed heterozygosity (Ho) = 0.346 was significantly higher than the expected heterozygosity (He) = 0.288 in the accessions of all five origins. This finding is expected, since the study was conducted on hybrids of different species and parents from different geographical regions. Moreover, the Fst and the negative Fis values also indicated an excess of heterozygosity.

4.4. Association Mapping Results

Eo x Eg hybrids are used all over Central America due to their tolerance to the "Pudrición de Cogollo" disease. However, only a few studies to improve desirable traits of these hybrids have been published so far and were performed mostly in structured germplasm (Singh et al. 2009; Montoya et al. 2013; Ting et al. 2014, 2016).

For performing the association mapping study with this high rate of missing values an imputation step was included in the analysis. Missing variants in the genotypes were estimated by LinkImputeR. This software performs missing data imputation based on the k-nearest neighbour algorithm. There is no need for physical or genetic maps, and the software is designed to work also on non-model and heterozygous species. The linkage disequilibrium (LD) between SNP is taken into account when choosing the nearest neighbours and samples are chosen which share an evolutionary history at the SNP to be imputed (Money et al. 2015). Nazzicari et al. (2016) observed in alfalfa (autotetraploid with high heterozygosity) that the K-nearest neighbour imputation showed the highest imputation accuracy compared to other studied imputation software. They found an average accuracy of 82.20 %. The same authors noted also a decrease in imputation accuracy when extreme missing value ratios of 0.4-0.7 were present. Money et al. (2017) determined an accuracy of around 94.6 % with LinkImputeR software when threshold values were depth below 8 and call ratios below 0.5.

Teh et al. (2016b) determined linkage disequilibrium decay rates for Deli x AVROS and Nigeria x AVROS accessions on 25 Kb and 20 Kb segments as average pairwise correlation coefficients (r^2) of 0.12 and 0.15, respectively, indicating that all SNP within this region would provide the same information. Therefore, also in our study consecutive SNP markers within a distance

below 25 Kb were represented by only one SNP with the lowest *p* value. The SNP with the smallest *p* value in the segments were the same for all traits, except in one case where two adjacent SNP (CU09: 20495654; CU09: 20495700) revealed the lowest *p* values for different traits. Therefore, we determined 24 significant SNP at 23 loci.

In total four different models were used for association mapping. Two generalized linear models (GLM) with population structure (GLM_Q) and principal component analysis (GLM_PCA) matrixes as covariates, and two mixed linear models (MLM) where in addition a Kinship relatedness matrix (K) between individuals was included in the model (MLM_Q+K, MLM_PCA+K). After the analysis, the best fitting model was determined from QQ plots using the average squared distance from the diagonal as described in Materials and Methods.

Unadjusted $p < 0.01$ values were used as threshold to determine potential significant associations. A total of 78 potential associations were observed under these conditions between 25 SNP and 20 traits. However, when applying FDR multiple testing ($p < 0.05$) this value decreased to 25 associations and only six traits were related to eight SNP. This FDR ($p < 0.05$) threshold has been used by several authors(Cheng et al. 2015; Branham et al. 2016). However, some studies have pointed out that these values might be too conservative and potential useful associations may be lost (Pasam et al. 2012; Gao et al. 2016). Zegeye et al. (2014), for example, used in a GWAS study unadjusted *p* values of ≤ 0.005 or ≤ 0.05 *p* as thresholds to determine significant QTL for stripe rust in synthetic wheat lines.

Adjusted *p* values revealed associations with the traits IV, Mono-Un, OA, UUU, Sat and SUS traits. Positive effects were observed for IV, Mono-Un, OA and UUU, while negative effect values were detected for Sat and SUS and vice versa. These findings are not surprising since the first group is related to unsaturated fatty acids, while Sat and SUS consider the complementary, saturated fatty acids composition.

We have detected for five of the eight FDR adjusted SNP associated with the six quality traits (C01: 2788341, C01: 49567798, C02: 17522364, C15: 20356039, CU06: 44960284) a potential relevant biological meaning related to oil biosynthesis.

SNP C01: 2788341 was detected within a holocarboxylase synthetase [HCS]. In humans, as well as in plants, this ligase attaches biotin to apocarboxylases. Biotin is essential for metabolism and survival, since biotin-dependant carboxylases catalyse key reactions such as fatty acids synthesis, gluconeogenesis and amino acid catabolism (Puyaubert et al. 2008; León-Del-Río et al. 2017).

SNP C01: 49567798 is encoded within a fructose-bisphosphate aldolase 1. Fructose 1,6-biphosphate aldolase (FBA) is a key enzyme in plants and it is involved in diverse reactions; glycolysis, gluconeogenesis and Calvin cycle (Lu et al. 2012). In Tea oil tree (Camellia oleifera) it is known to hydrolyze fructose-1,6-bisphosphate into dihydroxyacetone phosphate and glyceraldehyde 3-phosphate which are two critical metabolites for oil biosynthesis (Zeng et al. 2014).

SNP C02: 17522364 is located within a PTI1-like tyrosine-protein kinase 1. In Arabidopsis, Ghelis et al. (2008) determined that tyrosine phosphorylation may regulate oil biogenesis or oleosin targeting and therefore modulate the hydrolysis of lipid reserves. Moreover, Ramachandiran et al. (2018) stated that the oil content in Arabidopsis seeds is regulated by serine/threonine/tyrosine protein kinase.

SNP C15: 20356039 encodes an oligouridylate-binding protein 1B (UBP1b). It is a stress granule protein known to be involved in stress tolerance in plants (Nguyen et al. 2016). Nguyen et al. (2017) showed also the sensitivity of ABA signaling-pathway genes when overexpressing UBP1b. Especially, mitogen-activated protein kinase cascade genes showed overexpression profiles. These genes are known to regulate pollen lipid body biogenesis in Arabidopsis (Zheng et al. 2018), as well as fatty acid, triglyceride, phospholipid and cholesterol synthesis (Jeong et al. 2014).

SNP CU06: 44960284 4-hydroxy-3-methylbut-2-enyl diphosphate reductase is part of the DOXP/MEP pathway for isoprenoid biosynthesis. Kizer et al. determined that the accumulation of 3-hydroxy-3-methylglutaryl-coenzymeA, another gene of, the DOXP/MEP pathway, influences fatty acid biosynthesis, increasing palmitic acid and to less extent the oleic acid levels (Kizer et al. 2008).

Considering the results for unadjusted p values, potential SNP with a relevant biological meaning related to quality and production traits were detected.

SNP C01: 34080563, encoding thioredoxin proteins, showed a potential association with several quality traits. Kozaki et al. (2000) determined that thioredoxin-f proteins play key roles in de novo fatty acid biosynthesis since they regulate Acetyl-CoA carboxylase known to catalyze the formation of malonyl-CoA from acetyl-CoA. Also, Lemaire et al. (2004) arrived to the same conclusion.

SNP C05: 38969899 revealed a potential association with two production traits (BN, BY) and with the quality trait Sat. This SNP is located in a chaperone DnaJ protein. These proteins are

known as essential to abiotic and biotic stress responses, but Salas-Muñoz et al. detected that the loss of AtDjA3 gene function was associated with reduced seed production in Arabidopsis (Salas-Muñoz et al. 2016).

SNP C07: 21833956 is encoded through a ribosomal protein (RP). This SNP influences the OildM trait. Tatematsu et al. (2008) indicated that a large number of RP are involved in the regulation of protein synthesis-related genes during seed germination, while Li et al. and Liu et al. determined the relation of these RP in the regulation of lipid metabolism (Liu et al. 2014; Li et al. 2015).

SNP C10: 26013376, potentially influencing the UUU and SUS traits, represents an ADP-ribosylation factor-like protein, associated with several processes such as DNA repair, protein turnover, inflammatory regulation, aging or metabolic regulation (Vida et al. 2017). Furthermore, Gariani et al (2017) stated that by inhibiting poly ADP-ribosylation, the fatty acid oxidation increases in mice. Also the activity of Phospholipase D (PLD) is under the control of ADP-ribosylation (Basiouni et al. 2013) and PLD is a key in the production of free choline and phosphatidic acid, this latter known to be the intracellular lipid mediator of many biological functions (Jenkins and Frohman 2005).

The suppression of protein SRC2 (SNP C16: 2055062), potentially influencing UUU, inhibits hepatocellular glucose and lipid and cholesterol biosynthetic pathways (Madsen et al. 2015).

Often the identification of genes with a relevant biological meaning influencing a trait is more difficult and perhaps speculative, since in part complex biological pathways influenced by many genes are involved, and still a lot of information is missing.

Our association mapping approach was able to detect significant associations between SNP and up to five production traits and 15 quality traits, depending on the stringency of the analyses. These findings could help breeding companies to detect or to develop oil palm hybrids with improved productivity or higher oil quality by applying marker assisted selection in their breeding programs. Increased productivity can make oil palm cultivation more sustainable, while increased oil quality would satisfy the increasing demand of customers for quality oil with higher contents of unsaturated fatty acids (Aprile et al. 2012).

In our RARSeq assay we could target only a reduced number of SNP associated with some candidate genes, limiting the scope of our approach. This was mainly due to the high ratio of missing values. For future applications more sequencing runs should be performed to obtain a reasonable number of variants. In addition, further studies should be conducted to improve

the results, considering also other molecular tools such as whole genome resequencing, transcriptome sequencing, or bait sequencing in order to increase the number of targets.

CHAPTER V: GENERAL DISCUSSION AND CONCLUSIONS

CHAPTER V: GENERAL DISCUSSION AND CONCLUSIONS

5.1. General discussion

Palm Oil (PO) is the most used vegetable oil worldwide with a production of 84.82 million metric tons according to the last report of the United States Department of Agriculture (USDA 2019). Moreover, the Oil Palm (OP) tree produces up to 10 times more oil than any other oil crop since it is the highest yielding oil crop (Sumathi et al. 2008). Due to this, since early 1900 different breeding programs were focused on the improvement of oil production. An inflexion point occurred in 1940, when Beirnaert identified the *Shell* thickness (Sh) gene and saw that crosses between *dura* (D) fruit form (female) and *pisifera* (P) fruit form (male) generated *tenera* (T) fruit form palms with 30 % more oil production. Since then, all breeding programs have been focused on the improvement of T palms.

At the end of the twentieth century molecular breeding emerged as the perfect tool to face problems of classical breeding such as the long breeding cycles, the use of large amount of land or the polygenic behaviour of some traits. Furthermore, the advancements in technology enable nowadays the sequencing of whole genomes or transcriptomes in cost effective ways. Malaysian Palm Oil Board (MPOB) emerged as pioneer in OP related genomics. In 2013 Singh et al. (2013b) published the assembled genome of OP and determined genome size of 1.8 Gb. In that same year, this group published also the protein encoding the Sh gene and showed two independent mutations in exon 1 in which P haplotypes from Congo (PisC) and Nigeria (PisN) origin were determined. Following this publication, Ritter et al. (2016) and Babu et al. (2017) designed different molecular marker systems to distinguish between D and P forms in seedling stage. In 2014, also Singh et al. (2014), determined the *virescens* (VIR) gene controlling the fruit exocarp color which is an indicator of ripeness.

Even though these two characters are encoded by only one gene most of the traits are quantitative in nature. The study of these traits has been based on two main assays; Linkage Mapping (LM) and Association Mapping (AM). While the first one is based on structured population, the second relies on the linkage disequilibrium (LD) of genotypes of unobserved ancestry (J. and Cloutier 2012). First OP studies were based on LM. Seng et al. (2011), for example, constructed a linkage map based on single sequence repeat (SSR) and amplified fragment length polymorphism (AFLP) markers for oil yield components in Deli origin D and Yangambi origin P. However, these results may not be valid in other background due to the

stratification of the populations and these last years AM approaches have become the most popular (Teh et al. 2016b; Xia et al. 2018).

All these studies were mainly focused on the *Elaeis guineensis* (Eg) species which addresses most of the oil production. However, Eg palms are suffering from Bud rot disease, also known as "Pudrición de Cogollo" in American regions in which devastating scenarios with many dead palms are found. Due to this, *Elaeis oleifera* (Eo), less productive, have gained the attention of American Palm companies since the interspecific cross between the two species Eo x Eg has shown desirable characteristics such as better oil quality, resistance to diseases and competitive oil production (Preciado et al. 2011; Mozzon et al. 2013; Barba 2016; Corley, R.H.V. and Tinker et al. 2016). Specific studies to improve these species are few: Montoya et al. (2013) established a linkage map with 364 SSR and developed QTL for fatty acid composition, while Singh et al. (2009) also worked on fatty acid composition and developed a linkage map using AFLP, restriction fragment length polymorphism (RFLP) and SSR markers.

With the main aim to enrich the huge gap that exists in this field, a screening of 209 Eo genotypes, 32 Eg genotypes and 327 Eo x Eg genotypes from different origins have been performed across a 237 pb amplicon of Sh gene. Furthermore, an exhaustive analysis of 6 oil production traits and 19 oil quality traits have been done across 198 Eo x Eg genotypes from 5 origins. Moreover, related to these traits two AM studies based on 1) targeted candidate gene (CG) approach and 2) random Restriction site Associated RNA Sequencing (RARSeq) have been performed.

The first objective of the thesis focused on the exploitation of the well-known *Shell* (Sh) gene. Sh gene is known to encode the endocarp development, thus, the fruit forms, allowing the selection of desired varieties in nursery stage (Singh et al. 2013a). Even though extensive analysis have been performed in Eg (Singh et al. 2013a; Ooi et al. 2016), no studies are available for Eo Sh alleles. A broader analysis of a partial Sh amplicon addressing part of exon and intron sequences was performed using 209 Eo accessions, 327 interspecific hybrid accessions and 32 Eg *pisifera* genotypes. This large screening determined apart from the already known D, PisC, PisN and MPOB3 alleles, three new events (OLI1, OLI2, OLI3) derived from four new SNP (NK1, NK2, NK3a, NK3b), all located in the adjacent intron of exon 1.

While the different Eg genotypes showed the already known Sh Eg alleles, all Eo genotypes had the same SNP as the *dura* type. In contrary, the newly found four SNP were absent in all Eg accessions in this region, suggesting to be specific for Eo accessions.

NK2 variant appeared in all screened Eo and interspecific cross accessions. NK3a and NK3b cosegregated together (=NK3) and NK1 and NK3 never appeared together. The 3 Eo specific events are defined by the presence of only NK2 (OLI1), the occurrence of NK2 together with NK1 (OLI2) and the presence of NK2 together with NK3 (OLI3), respectively.

Since the NK2 SNP was specific to all Eo accessions, two species specific primers (SSP), ShG-ShO, were designed, tested and validated. The results showed the discriminant application between Eg and Eo genotypes of these SSP; ShG only amplify in Eg and hybrid genotypes and ShO in Eo and hybrid genotypes. Moreover, breeding OP companies use *pisifera* genotypes as pollen donors in backcross programs with interspecific hybrids in order to improve oil production and inherit Eg favourable characteristics. These SSP will allow the detection of desired alleles in BC1 crosses where50 % pure Eg *pisifera* and 50 % "*tenera*" (Eo *dura*, Eg *pisifera*) are expected. The "*tenera"* genotypes could be used for further back crosses and the pure *pisifera* to build up an own *pisifera* collections. These primers can be of high importance in regions highly infected with "Pudrición de Cogollo" where most Eg palms die and have to be replaced by selected hybrids. Moreover, the pollen market of *pisifera* varieties can reach up to 800-1000 USD/gr of pollen for selected materials, becoming essential having own *pisifera* populations.

With the purpose of analysis if these new Eo events affect also the fruit weight (FW), kernel percentage in the fruit (KF), shell percentage in the fruit (SF) and mesocarp percentage in the fruit (MF) different ANOVA tests were applied in Eo and hybrid populations. However, since the new variants are in intronic regions the expected results showed null influence of the Eo Sh alleles in the studied phenotypes.

Second and third objectives relied on the discovery of new molecular markers related to traits of interest. When conducting an AM approach correct phenotyping is a key point since exhaustive and rigorous data are needed to perform reliable and strong mapping(Rafalski 2010). Furthermore, the size of the population of study is also of high importance; larger germplasm will provide more power. Most of the studies in OP have been focused on the improvement of oil yield (Kwong et al. 2016; Bai et al. 2017) and few have been done for oil quality (Mozzon et al. 2013; Morcillo et al. 2013). Here we have studied 6 production traits (bunch number [BN], bunch weight [BW], bunch yield [BY], oil percentage in fresh mesocarp [OilfM], oil percentage in dry mesocarp [OildM], and oil percentage in the bunch [OilB]), and for the first time an extensive study of 19 quality traits has been performed in which lipids (oleic acid [OA], saturated acids [Sat], mono-unsaturated acids [Mono-Un], poly-unsaturated

acids [Poly-Un], iodine value [IV], percentages of different types of triglycerides [SSS, SUS, SUU, UUU]), tocols ([Toc], sum of individual alpha, beta, gamma tocopherol´s [Alpha, Beta, Gamma, Tocph], sum of alpha3, beta3, gamma3 tocotrienols [Alpha3, Beta3, Gamma3, Toc3]) and carotene content [Car] have been studied in a broader collection of Eo x Eg genotypes. These results can contribute to deliver high-quality oil palm varieties according to the market demands.

Palms were planted in 2010 and data recording started in 2014. BN and BW data were collected over four years and cumulative data were used. The quantification of the rest of the traits was performed once. In total 198 Eo x Eg genotypes from five different origins were evaluated; 40 Coari x La Mé (CxL) accessions, 37 Taisha x Avros (RGS) (TxA(RGS)) accessions, 75 Taisha x Avros (Oleoflores) (TxA(O)) accessions, 25 Taisha x Ekona (TxE) accessions and 21 Taisha x Yangambi (TxY) accessions. ANOVA tests were applied to check the influence of the origin on the phenotype. CxL accessions showed overall larger oil extraction ratios, while Taisha accessions showed higher BN and BW. When looking to quality traits, also large differences were observed between CxL and Taisha accessions. CxL accessions showed significant higher unsaturated fatty acid levels, while Taisha accessions showed higher saturated values. These results are in accordance with Pélaez et al. (2010) and Barba (2019); the first study showed high OA, IV and oil extraction for Coari palms and the second revealed Taisha palms similar to “Guineensis” palms which are known to have higher saturated fatty acids levels. Nevertheless, Arias et al. (2015) detected the highest total oil-per-bunch ratios [%] in Taisha accessions. When looking to tocols CxL showed the lowest values. TxY accessions showed the highest Carotene contents.

AM in this study was based on two different approaches; targeted CG approach and random RARSeq approach in the mentioned 198 hybrids. In the first approach CG primer pairs related to traits of interest from a previous study were used (López de Armentia 2017). In total 167 CG addressing 1.86 % of the OP genome were tested. 32.7 million reads were obtained by a Personal Genome Ion Torrent Machine (PGM) using the 318 chip, but after the first filtering step 9.8 million clean reads were left. 83 % of them mapped to the reference genome, and GATK haplotype-caller detected 12,200 single nucleotide polymorphisms (SNP). However after applying the filtering steps and using a missing value threshold of 20 % only 115 SNP remained comprising 62 CG. In a similar approach with *Epinephelus coioides* (Guo et al. 2015) four genes of interest addressing 19.438 pb were studied by CG approach and PGM sequencing. After applying similar filtering steps 338 SNP were called and tested for association. Also Singh et al. (2019) performed in rice a CG based AM approach using a PGM platform and detected 155

SNP within 13 genes. As can be seen, these results show that the obtained results vary significantly in each study.

In the second approach for the first time a RARSeq approach has been applied in an AM study. A reduced transcriptome representation of a subset of the same genotypes (104 Eo x Eg) was sequenced three times by Illumina MiSeq platform and 31.6 million clean reads were obtained. 97.8 % of the reads mapped to the reference genome and GATK identified initially 2,992,736 variants, addressing 2.18 % of all genes. However, a great decrease in the number of SNP (310) was observed after cleaning filtering and using call ratios of 30 %. Double digestion Restriction site Associated DNA (ddRADSeq) libraries are known to have large amounts of missing data, mainly due to mutations at restriction sites (Arnold et al. 2013), due to individual genetic divergence or technical biases when preparing the library (Swarts et al. 2014; Brouard et al. 2017). Pyne et al. (2017) developed the first linkage map of *Ocimum basilicum* using a ddRADSeq approach. The library was sequenced twice on an Illumina HiSeq 2000 platform. From the initial 25,363 polymorphic loci, 3,492 polymorphic loci with less than 20 missing individuals remained. Friedline et al. (2015) observed an average of 79.4 % of missing haploid genotypes in four sequenced libraries using an Illumina HiSeq platform. These studies confirm, that the reduced representation of the genome-transcriptome assays leads to high missing value ratios. Moreover, since our study is based on in expression tags, it is expected to have an increment of missing values.

These two sequencing platforms have been used widely in diverse assays. Quail et al. (2012) studied their performance in a set of microbial genomes. *Staphylococcus aureus* was the closest to OP since it contains 33 % guanine-cytosine content. After sequencing and mapping to reference genomes; using “tmap” software for PGM and BWA for Illumina, they concluded that both sequencers were suitable for genomes with these characteristics. Furthermore, our results suggest that the chosen reference genome was suitable for our genotype data since most of the reads could be mapped. The same publication stated also that the number of SNP was higher in PGM, but contained also a higher number of false positives.

Concerning genetic diversity analysis, the two assays showed slightly different results. Significant excess of heterozygosity was detected in the population, since an observed heterozygosity (Ho) of 0.61 was determined in the CG approach and Ho = 0.35 in the RARSeq approach. Since we studied hybrid accessions from two different species, it is expected to have this excess of heterozygosity. However, the low Ho in RARSeq could be surprising. This phenomena may be explained by; 1) working with expression tags in which the most

heterozygous markers won´t be expressed in all genotypes and will be removed when applying missing values filtering, 2) less genotypes were tested, 3) a higher missing value threshold was applied and 4) in the CG approach primers are designed in gene amplicons expected to affect the phenotype. Looking to other published studies, Babu et al. (2019) determined a Ho value of 0.38 in Eg palms and claimed that this value represents a high heterozygosity level. Montoya et al. (2013) tested 700 SSR and determined the number of heterozygous loci in different *Elaeis* palms; 161 (28 %) in Eo palms, 147 (25 %) in Eg palm and 457 (78 %) and in interspecific Eo x Eg hybrid palms, suggesting a really high heterozygosity in hybrids. Arias et al. (2015) showed Ho values of up to 0.77 in Eo genotypes and up to 0.60 in hybrid Careté x La Mé genotypes.

Genetic diversity between accessions and within accessions were determined by Fixation indices (Fst) and Inbreeding coefficient (Fis) values, respectively. Here, both assays revealed similar results. Fst values showed high diversity in the whole population, since all values were close to zero. In both assays, the hybrids involving Brazilian accessions (Coari x La Mé) showed to be the most different from the rest of the Ecuadorian accessions. This finding agrees with the ones found by Arias et al. (2015) who studied genetic and phenotypic differences between Brazilian, Peruvian, Ecuadorian and Colombian Eo. They evaluated 13 microsatellites (SSR) and detected 112 alleles of which 27 % were specific alleles for each country. However, in the CG approach the closest distances were observed between Taisha x Avros (RGS) and Coari x La Mé and a large distance was detected between Taisha x Avros (RGS) and Taisha x Avros (Oleoflores) accessions. Fis values, on the other hand, were close to zero or even negative indicating highly diverse populations in which no inbreeding mating system was detected (French et al. 2005).

Four models were tested for phenotype-genotype association in the two assays. Two generalized linear models (GLM) where structure matrix (Q) and principal component analysis matrix (PCA) were taken into account (GLM_Q; GLM_PCA)and linear mixed models (MLM) where in addition an identity by descent Kinship matrix (K) explaining relatedness between individuals was added (MLM_Q+K; MLM_PCA+K). These four models have been widely used in diverse fields such as humans, forensics, plant or animal breeding. Liu et al. performed a genome-wide association study (GWAS) for colon carcinogenesis in mice and used a mixed model in which population structure and genetic relatedness were applied (Liu et al. 2012). In maize, Wang et al. (2012) tried 6 different models in their GWAS assay; three GLM models with Q, PCA and no covariate, and three MLM models comprising K matrix, Q + K and PCA + K. These models have also been applied in OP as described in the Introduction.

The observed *p* values of the four models were drawn in a Quantile-Quantile plot (QQ) based on the expected *p* values for each trait. A QQ plot is a graphical representation of the deviation of the observed *p* values from the null hypothesis in which SNP are sorted and plotted against the expected *p* values from a theorical χ^2-distribution (Ehret 2010). However, from the plots of our models, it was hard to choose visually the model that fitted best our data, since in many cases the models showed similar graphs. Therefore, an equation to calculate the average square distance (d^2) from the diagonal was developed to choose objectively the best model. In this equation, the smallest d^2 shows the model that deviates less from the null hypothesis, and in our assays the models with smallest d^2 for each trait were used for further analysis. Since the markers in each assay were different, QQ plots for each trait were drawn in both CG and RARSeq approaches, even though the traits were the same. Jung et al. stated that different results are achieved when using different markers, since the power to detect allelic association depends on the specific properties of the markers (Jung et al. 2005). In general and in accordance with the literature (Wen et al. 2014, 2015), mixed models showed higher strength, since smaller d^2 values were obtained for these models for most of the studied traits. In the CG approach 20 of the studied trait showed to fit better with mixed models, while in RARSeq this was the case for 17.

To call for true associations *p* value thresholds were used for each assay. Multiple testing uncorrected ($p < 0.05$ in CG assay, $p < 0.01$ in RARSeq assay) values were used, as well as, False Discovery Rate (FDR) (Benjamini and Hochberg 1995) adjusted $p < 0.05$ values. Most of the genomic and proteomic studies deal with false positive results, Type I errors, when many candidates are statistically tested (Noble 2009). Due to this, multiple correction tests are applied to test statistical confidence measures based on the number of test performed. The Bonferroni adjustment (Holm 1979) is the most applied test (Yu et al. 2017; Zhao et al. 2017) in which the family-wise error rate is corrected. However, this test might be too strict (Riedelsheimer et al. 2012; Nigro et al. 2019) and also, false negative (Type II errors) might be neglected. Thus, the tendency has shifted to the less restrictive FDR adjustment (Romero et al. 2014; Sant'Ana et al. 2018) in which the expected proportion of falsely rejected hypothesis is controlled (Narum 2006). But, here too, the stringency of FDR correction has been pointed out by some studies and accepted *p* values in the bottom of 0.1 percentile of the distribution have been used (Chan et al. 2010). As mentioned before, AM studies rely on LD. Due to the small amount of variants observed in both assays no LD study was conducted. Babu et al. (2019) determined a LD decay of 25 Kb at 0.45 of average pair wise coefficient of correlation (r^2) and Teh et al. (2016a) an LD decay of 25 Kb and 20 Kb at 0.12 and 0.15 r^2, respectively, suggesting

high decay ratios for OP. In consequence, CG assay results were given as CG-haplotype results and in RARSeq assay only one SNP within a 25Kb distance was used. With these cut offs 7 potential CG related to 6 production traits and 23 potential CG associated to 18 quality traits were detected in the CG approach. Four of them showed a relevant biological meaning for further analysis. However, when FDR multiple test was applied no significant results were achieved. In the RARSeq assay, using unadjusted *p* values, 23 loci seemed to be related to 5 production traits and 15 quality traits, but when applying, FDR multiple correction with $p < 0.05$, only 8 loci were left related to 6 quality traits. Six of the detected loci presented potential relevant functions for further studies.

These results have emphasized the quantitative nature of these traits since several CG involving complex metabolic routes could be associated with single traits. Even though the preferences of OP breeding companies focus on production traits, the results obtained in these two assays have shown only few useful results for these traits. This might be due to the short age of the palms. In contrary, the assays have detected some interesting SNP related to quality traits with low *p* values that could be exploited.

In both assays, a limited portion of the genome has been studied limiting the scope of our approaches. In the CG assay no significant results were obtained using corrected tests indicating the need improvement of our studies using alternative techniques such as transcriptome sequencing or bait sequencing. In RARSeq better results were obtained, but yet the number of SNP left after applying filtering steps suggest a need of amelioration. In conclusion, the number of targets has to be improved in order to make meaningful analyses.

5.2. Conclusions

In the following section the main findings of this thesis are summarized under the assumption: "Association mapping based on targeted and random candidate gene analysis allows the detection and identification of molecular markers related to traits of interest in interspecific *Elaeis oleifera* x *Elaeis guineensis* oil palm hybrids".

1. A broader analysis of an amplicon of the "*Shell thickness*" gene have been performed using 209 Eo accessions, 327 interspecific hybrid accessions and 32 Eg P genotypes. This analysis has been able to detect the already known Eg *dura*, PisC, PisN and MPOB3 events in Exon 1 and 4 new additional SNP (NK1, NK2, NK3a, NK3b) in the adjacent intronic region resulting in three new Eo events (OLI1, OLI2, OLI3). While all Eo accessions behave as *dura* type with respect to the Eg events, the Eg genotypes behave as *dura* in all new Eo variants, suggesting that this new variants are specific for Eo.
2. Two species specific primers (ShG, ShO) have been designed, tested and validated. These primers allow the identification of pure *pisifera* genotypes in BC1 crosses between interspecific Eo x Eg hybrids and Eg *pisifera* palms. These primers can become of high importance in regions highly infected with "Pudrición de Cogollo" where most Eg palms die and owing *pisifera* populations is essential.
3. Fruit weight (FW), kernel percentage in the fruit (KF), shell percentage in the fruit (SF) and mesocarp percentage in the fruit (MF) data have not shown significant association with the Eo alleles. However, the origin of the genotypes has shown significant influence on all the traits but KF.
4. An extensive study of 25 oil palm traits related to production and quality has been developed in a broader germplasm of 198 interspecific Eo x Eg hybrids from five origins (Coari x La Mé [CxL]; Taisha x Avros (RGS) [TxA]; Taisha x Avros (Oleoflores) [TxA(O)]; Taisha x Ekona [TxE]; Taisha x Yangambi [TxY]). CxL accessions have shown higher oil % in fresh mesocarp, oil % in dry mesocarp, and oil % in bunch levels, while TxA(O) has revealed the highest values for bunch number, bunch yield, and bunch weight traits. For quality traits, CxL accessions have shown significally higher unsaturated fatty acid levels in terms of unsaturated fatty acids %, oleic acid %, iodine value, SUU triglyceride and UUU triglyceride. These Interesting associations between phenotype and origin could be used for the selection of desired phenotypes.
5. One targeted candidate gene (CG) approach and one random Restriction site Associated RNA Sequencing (RARSeq) analysis have been performed for the discovery

of new SNP to be used in association mapping (AM) studies. CG approach have been able to detect 115 SNP, while RARSeq approach 310 SNP. Markers from both, targeted and random assays have shown high heterozygosity (CG Ho= 0.61; RARSeq Ho= 035) of the studied materials. Furthermore, Fixation index and Inbreeding coefficient have confirmed the high diversity that exists in the studied germplasm.

6. In both AM studies four statistical models have been applied successfully. Two generalized linear models where structure matrix or principal component analysis matrix (GLM_Q, GLM_PCA) have been evaluated as covariates, as well as, two mixed linear models where in addition a Kinship matrix (MLM_Q+K; MLM_PCA+K) has taken into account.
7. A new equation has been developed in order to measure the average square distance from the diagonal in AM assays, which allow to determine in an objective way the best model that fits our traits of interest. This equation has determined the preference of mixed models for 20 traits in CG approach and for 17 traits in RARSeq assay.
8. CG AM approach has not been able to detect any significant association using false discovery rate (FDR) adjusted *p* values. However, with unadjusted *p* values, 7 CG have been detected with significant influence on six production traits and 23 CG influenced 18 quality traits. Four of the detected CG, LIPOIC-SEQUI-TO1-EgNAC, have revealed interesting relevant biological meanings that could be exploited within markers assisted selection (MAS) programs.
9. The RARSeq approach has been able to detect significant FDR adjusted associations between 8 SNP and 6 quality traits and 25 SNP influence significantly 20 traits using unadjusted *p* values. 10 SNP have shown a potential relevant biological meaning related to oil biosynthesis and seed production.
10. Even though these two approaches have been able to detect CG that could be exploited in the near future, the general coverage of both assays and the low number of SNP left after filtering steps have pointed out the need of increasing the percentage of the studied genome.

ANEX

ANEX

Table A 1: Characteristics of all 171 Candidate genes analysed initially by Amplicon sequencing in oil palm hybrids.

No	CG Name	**GeneID_NCBI**	**CGpos MPOB (Amplicon_length)**	**CG function**	**Amplicon Pimers**	**Library No**
1	ACAD	XM_010942754.2	C16: 15433126-15432973 (154 bp)	peroxisomal acyl-coenzyme A oxidase 1	Fw:TCCAGCTATAAAAGGACAAGGAA Rv: TTTGGGATCAAATGTTGCAG	2
2	ACYL-ACPF	XM_010926998.2	C07: 1586207-1586053 (155 bp)	palmitoyl-acyl carrier protein thioesterase	Fw: AAGCAGTGGACCCTTCTTGA Rv:TCTATAGAAGCCGTCCGATCA	3
3	AG1	XM_019849509.1	C03: 8101407-8101187 (221 bp)	MADS BOX trancription factor	Fw: AGGAGGAGCCAAGAGGTAGC Rv: TTGTTGGCGTATTCGTAGAGG	3
4	AIL5	XM_010930367.2	C02: 47130315-47130524 (210 bp)	AP2-like ethylene-responsive transcription factor	Fw: TCATGAACAGTGACCTCCCC Rv: CTTCAAAACCCAACGCCAGA	3
5	ANT	XM_019849201.1	U10: 19445997-19446205 (209 bp)	AP2-like ethylene-responsive transcription factor	Fw: TTGGCTCTTGGTTCATTGGC Rv: AGGATTCAAGTCACGGCTGA	3
6	ASP2	No annotation	C16: 3316490-3316701 (212 bp)	asparagine synthetase-related protein	Fw:GGAAAATAATATCTTGAGTTCCACA Rv: TGGATCATAGATCGGA	3
7	ATAGB1	XR_002165879	C12: 17556353-17556542 (190 bp)	GTP binding protein beta 1	Fw: ATTCACTGGATTGGGCTCCT Rv: CATGCACTATCAAGACCACCAC	1
8	ATP1	EU016946.1	CT: 15520-15741 (222 bp)	ATP synthase CF0 subunit IV	Fw: TTCGGAATCCACAAACCATT Rv: TCGTAGGCGCAGCTAACTCT	2
9	ATP2	EU016918.1	CT:	ATP synthase CF1 epsilon subunit	Fw: CAATGGTTAACGGTGGCTCT	2

			53908-54078 (171 bp)		Rv: TGCATGTCTCTTACCCTCAGC	
10	ATP3	EU016883.1	U05: 50035784-50035933 (150 bp)	ATP synthase CF0 subunit III	Fw: TTGCTGTAGGCCAAGCTGTA Rv: AAAAGGATTCGCCAACAAAA	2
11	atpB	EU016907.1	CT: 54576-54425 (152 bp)	ATP synthase beta subunit (atpB)	Fw: TGCGCAAAGAGTTAAGCAAA Rv: CCACGAAGAAGGGTTGTGAT	2
12	AUX2	XM_010920561.2	C04: 33587161-33587383 (223 bp)	probable indole-3-pyruvate monooxygenase	Fw: CAGGGTGTTCCTTTCGTGAT Rv: ACTGGTTGAATCTCGGGTTG	2
13	BAK1	XM_010938735.2	C13: 25407350-25407528 (179 bp)	somatic embryogenesis receptor kinase 2-like	Fw: ACAGCAGTCCGTGGAACAAT Rv: ACCCAGTCAAGCAACATCAC	3
14	BKACPII_1	FJ940767.1	C10: 22949664-22949486 (179 bp)	beta-ketoacyl-ACP synthase II	Fw: TGACCTTTGCATCATTCAGC Rv: ACGCAGCTTCTTTGTTGGT	1
15	BKACPIII	AF169015.1	C15: 19066681-19066878 (198 bp)	beta ketoacyl synthase III	Fw: TGCAACAGTCAAGGATGAGG Rv: AGCCAGTCAATGCTGGAACT	1
16	BnC10_7131	XM_010934039.2	C10: 22721817-22722099 (283 bp)	cell division control protein 48 homolog B isoform X2	Fw: GGCAGGCATGGTAGCTCTTA Rv: TGGCTGAAGGTGTTGTCAAG	1
17	BnC12_2975	XM_019854065.1	C12: 13275500-13275773 (274 bp)	nucleolar GTP-binding protein 2-like	Fw: ATCATCCCAGCAGCTAATGC Rv: GCAAAACATTTCCCACCTT	1
18	BnC13_gi191204957	XM_010938111.2	C13: 5330168-5330456 (289 bp)	TPD1 protein homolog 1-like	Fw: AGGACGACGTGGTGGTGTA Rv: CAGGAGACGGAGGAGACGTA	1
19	BnC2_10C3-629	XM_019847956.1	C02: 28517586-28517885	MADS-box transcription factor 21 isoform X2	Fw: TGAAGCAGGACATGAGTTTGA Rv: TGCAACAAAAGTCATGCAAT	1

			(300 bp)			
20	BnC2_1289	XM_010914447.1	C02: 29220936-29220683 (254 bp)	FRIGIDA-like protein 3	Fw: TAAGGTCACAGAGGCCCAAG Rv: CATTTTCAGCAGGCTTCACA	3
21	BnC3_792	XM_010918923.1	C03: 32128327-32128119 (209 bp)	oryzain gamma chain	Fw: GACTGGGAAAGGCATCTCTCT Rv: CGACAACCTTAACACCCACA	3
22	BnC7_3962	XM_010927796.2	C07: 12214023-12213768 (256 bp)	alpha,alpha-trehalose-phosphate synthase	Fw: TGTCTGTGATGCAGGAGAGG Rv: TGCGAGCTTCTTACTGCTTG	1
23	BnC8_761	XM_010929225.2	C08: 4351998-4351827 (172 bp)	60S ribosomal protein L23	Fw: CAACTTGTTAGTTTGTGTTTTGGAA Rv: ACAGGTTGTTCCCGAATCAG	1
24	BRI1	XM_010929461.2	C01: 2799339-2799525 (187 bp)	systemin receptor SR160-like	Fw: CTGCAAGGTTGGAGAAGAGC Rv: GTGAATTATGTGGGGAATGC	2
25	CA3	XM_010914762.2	C02: 35978110-35978321 (212 bp)	15-cis-zeta-carotene isomerase	Fw: CAGATGGTTGGGCAGGTAAT Rv: TGACGTCCATCAAGAATTGC	2
26	CA4	XR_002166117.1	C01: 37606005-37605886 (120 bp)	zeta-carotene desaturase	Fw: GATGTGGGATCCTGTTGCTT Rv: TCAGGAGAACCCTTCAGCAT	2
27	DDB1_CUL4	XM_010913578.2	C02: 9118490-9118731 (242 bp)	DDB1-CUL4 associated factor 1	Fw: ATTCTCAGGCGAAATCCAGA Rv: TTCCTCCCAACCTCAAACAG	3
28	DEF1	NM_001303583.1	U03:955283-955535 (253 bp)	MADS box transcription factor	Fw: CGAGCACCCAGTTTATGGTT Rv: GCGGATAGAGAGGCTTACCA	1
29	DWARF7	XM_010918132.2	C03: 21298109-21298302 (194 bp)	delta(7)-sterol-C5(6)-desaturase	Fw: ACCTTCAAACAAAGCCATGC Rv: TCCAAACTCCACAAAGACCA	1
30	DXS2	NM_001303573.	C14:	1-deoxy-D-xylulose-5-phosphate	Fw: ATGTTGGTGGGAATGGAGGT	3

		1	21869601-21869411 (191 bp)	synthas	Rv: TCCCTTGTTCTTGCGTCTCT	
31	EgAcp	XM_010926732.1	C06: 42406831-42406636 (196 bp)	1-aminocyclopropane-1-carboxylate oxidase 1	Fw: ACAAGGATGGAAGCCGTCTA Rv: CGTCGTCGTCACTTGAACAT	3
32	EgARF1	JN003543.1	C09: 32491569-32491287 (283 bp)	Auxin response factor as transcription factor	Fw: AGCCCTTGATTTGTCGATGC Rv: ACTGGACTTATCTTGGGGTTGT	3
33	EgBRX	XM_019853109.1	C10: 23849149-23848990 (160 bp)	Brevis radix for regulating brassinosteroid-biosynthesis	Fw: ATGTGTAACCCATCGCTCCA Rv: ACTCTTTCGGGCCTTGGTTA	3
34	EgDSI	AY182168.1	C04: 56149385-56149241 (145 bp)	opsc112 protein disulphide isomerase	Fw: GTGCGGATAGTCAGGAGCTT Rv: GCCTCCTTCTTCAGCAACAC	3
35	EgEBF	XM_010908977.2	U03: 51006442-51006615 (174 bp)	EIN3-binding F-box protein 1-like	Fw: TGCTGCCCAAAGTTGAAGTC Rv: AGCAACAGCCTTCCCATAGT	3
36	EgFATB1.2	XM_010926998.2	C07: 1581809-1581616 (194 bp)	palmitoyl-acyl carrier protein thioesterase, chloroplastic	Fw: CTAATGACTGCACTGGTGGC Rv: CCTGCCCCAATTGAGAATGG	3
37	EgFATB2.2	XM_010916714.2	C03: 1846524-1846322 (203 bp)	palmitoyl-acyl carrier protein thioesterase	Fw: GGAATGTGGGACCGAACTTG Rv: CTCCATCCATCCTAGCACGT	1
38	EgMBAGL2-3	XM_010914716.2	C02: 33901209-33901053 (157 bp)	agamous-like MADS-box protein AGL9 homolog (AGL2-3)	Fw: CTGGGCCTAGCGTGAGTAAT Rv: AACAGTTGCCAATTACAGACCA	1
39	EgNAC	DQ267443.1	C05: 40852033-40852228 (196 bp)	1 NAC protein	Fw: CGTTGTTCGGGCTAAAAGAG Rv: ACTTGGTGCCTTCCCAGAGTA	1
40	EgPINF3-6_PIN1	XR_002165602.1	C10: 11001094-11000825	probable auxin efflux carrier component 1c	Fw: CTACCGAGCTCCTCCCAAAA Rv: GCAGGCATTTCAACATCCCA	3

			(270 bp)			
41	EgPINF3-9_PIN4	XM_010910943.2	U06: 37424604-37424811 (208 bp)	probable auxin efflux carrier component 1c	Fw: ATGGAGCGAGAGGACTTCAG Rv: TGAGGCTGGAGTAGGTGTTG	3
42	EgPPGL	XR_833625.1	C12: 8353843-8353638 (206 bp)	opsc155 putative 6-phosphogluconolactonase	Fw: GTGCTGTGCACAAGGCTCTA Rv: GCTAGTCCCGCAGAAAGTTG	1
43	EgTPase	AJ507416.1	C03: 36154746-36154995 (250 bp)	partial TPase pseudogene	Fw: TAGGCCTCATTTGCACTCGA Rv: GGCACCATACCTCGAAAAGC	3
44	EgWRI1-2.2	XM_010924633.2	C05: 39784587-39784821 (235 bp)	ethylene-responsive transcription factor WRI1	Fw: ACCAACTTTGCTAGCAGTGC Rv: GGCCATCATCAATTGGGACA	3
45	EIN4	XM_019850079.1	C04: 31379185-31379020 (166 bp)	ethylene receptor 2-like	Fw: AGGTGGGACACCAGAGATTG Rv: CTGCGATTCCTCCAAAACT	2
46	ELO2	XM_010939372.2	C14: 1363060-1363273 (214 bp)	elongator complex protein 1	Fw: ATCCAGCTGAGGCTGCTAAA Rv: TTCGCCACTTTCTCTGTTCC	3
47	EPS134312	XM_010934072.2	C10: 22943253-22943135 (119 bp)	3-oxoacyl-[acyl-carrier-protein] synthase II	Fw: AACGAATGGACAAATTCATGC Rv: TATGAGCACTCCGCATTTTG	2
48	EPS168	XM_019849021.1	C01: 43301684-43301536 (149 bp)	MADS box transcription factor	Fw: GGCGGATCGAGAACAAGATA Rv: GGCGTACTCGTAGAGCTTGC	2
49	EPS3	XM_010923043.2	C05: 10242828-10242630 (199 bp)	Peroxirredoxin 1-Cys	Fw: AACATGGAGGAGGTGGTCAG Rv: TTGTGAAGCGGAGGTAATCC	2
50	EPS50987A	U68756.1	C07: 23767648-23767541 (108 bp)	Esteroil-ACP- desaturase	Fw: CCAGGATCAGGAGGATTGAA Rv: TTGGTCATGCTCAGAGTTGC	2

51	ETR1	XM_010923296.2	C05: 14758531-14758659 (129 bp)	probable ethylene response sensor 1	Fw: ACAATGGCCTGCTGAAGAAC Rv: AGATGGGTTGCTCCACAAAG	2
52	ETR2	XM_010937466.2	C12: 25045240-25045404 (165 bp)	ethylene receptor 3	Fw: GATCCCGCAACTTCTGAGAG Rv: AAGATGGTATGCCGGTCAAG	2
53	FA1	XM_010923348.2	C05: 16828712-16828892 (181 bp)	glyoxysomal fatty acid beta-oxidation multifunctional protein	Fw: TTGTCCCTCCAACTGATGTG Rv: ATCCGATTCACAGCAAAACC	2
54	FA2	XM_010906903.1	C09: 6276571-6276327 (245 bp)	stearoyl-[acyl-carrier-protein] 9-desaturase 5	Fw: ATGAGAAGCGCCATGAGACT Rv: GTTCCACCTTCGGACAAGAA	2
55	FA4	XM_010928432.1	C07: 23768995-23768846 (150 bp)	stearoyl-[acyl-carrier-protein] 9-desaturase 5	Fw: GTGCATCTGAAGCCTGTTGA Rv: CATATCTCCAACCAAGCAAACA	2
56	FA6	XM_010928432.1	C07: 23768994-23768762 (233 bp)	stearoyl-[acyl-carrier-protein] 9-desaturase 5	Fw: TGCATTTGAAGCCTGTTGAG Rv: AGTAAGGCTTGCCCCTGTCT	2
57	FA8	XM_010906903.1	U02: 21412859-21413024 (166 bp)	stearoyl-[acyl-carrier-protein] 9-desaturase 5	Fw: GGGATGAAACAGGTGCAAAC Rv: CCATTCCCGAACCAATTAGA	2
58	FFB1_CL1016_S1.2	XM_010925967.2	C01: 21685911-21685668 (244 bp)	2-hydroxyacyl-CoA lyase	Fw: CCCCAACTTCTTTTGTTCCA Rv: TGGAATGGAGAAACTGCAAA	3
59	FFB11_C1_S1	XM_010935502.2	C11: 20181476-20181728 (253 bp)	protein maternally expressed gene 5 isoform X2	Fw: GAGCATGACCGAGATTCAGC Rv: CCTCCTGATCTGACCGTGTT	1
60	FFB13_C2168_S1	XM_010938902.2	C13: 22806564-22806378 (187 bp)	peptidyl-prolyl cis-trans isomerase NIMA-interacting 4	Fw: GCGAATGGAAAGCATGTGTT Rv: TGGTGTCTCTTCATTACCCACA	3
61	FFB2_C2_S1	XM_010914588.1	C02:	cytochrome P450 71A9-like	Fw: CTGCTTCCTCGAGAGTCCAT	3

			31455509-31455742 (234 bp)		Rv: CCCCAAATTCATTCCAGGGC	
62	FFB2_C3566_S9	XM_019847934.1	C02: 31308535-31308364 (172 bp)	transport protein Sec61 subunit alpha-like	Fw: GGTGCACAAAAGTTACTTGGC Rv: TCTGCAGAAGTTCATCCAAACA	3
63	FFB2_C4663_S1.2	XM_010913692.2	C02: 13183406-13183579 (174 bp)	tubby-like F-box protein 8 isoform X2	Fw: GCACTGCATCATGTACTCCA Rv: TCAGAGAAGCGAGAACTGCT	1
64	FFB2_C4741_S3	XM_010914549.2	C02: 31047775-31048026 (252 bp)	3-oxoacyl-[acyl-carrier-protein] reductase 4-like	Fw: TAATGTAAGCTCGGGTGGCT Rv: TCCCGCGTAACATAGATCGG	1
65	FFB2_C8_S1.2	XM_010914588.1	C02: 31455556-31455849 (294 bp)	cytochrome P450 71A1-like	Fw: TATCCTGCCAAACACGAGAG Rv: ATTCCTGGTGCCTCGTTCAT	1
66	FFB6_C2082_S1	XM_010925974.2	C06: 33052533-33052692 (160 bp)	3-ketoacyl-CoA synthase 4	Fw: TCCTTCTTCACAGCTGCAGA Rv: ATGAAGCACACATGAGGGCA	3
67	FFB6_C3684_S1	XM_019851300.1	C06: 33630243-33630033 (211 bp)	aspartic proteinase oryzasin-1-like	Fw: ACCTGAGAGTTGGATTTGCAG Rv: AACGGGCAGAACAACATACA	3
68	FFB8_C1455_S3.4.5.6	XM_010929243.2	C08: 4631353-4631609 (257 bp)	E3 ubiquitin-protein ligase RGLG2-like isoform X2	Fw: AGGGCCTTGAGTTATGTCCC Rv: TCCAGATTACCGCAACACCA	3
69	FFB8_C545_S1	XM_010929225.2	C08: 4354803-4354604 (200 bp)	60S ribosomal protein L23	Fw: GGGGCGAAGAACCTCTACAT Rv: TCAAAGTACATATAGACGCCGTC	1
70	GID1	XM_010940559.2	C14: 22469857-22470046 (190 bp)	gibberellin receptor	Fw: CGAGTCGGAGAAGAGATTGG Rv: AAGATCCAGTCCTGCCACTG	1
71	GLO2	XM_019846334.1	U02: 30414183-30414396	MADS box transcription factor 2	Fw: TTGCATGCCAGATTCCAATA Rv: CAGCCATCTATTAGCCCATCA	3

			(214 bp)			
72	GLUT1	AF261691	C12: 28135291-28135449 (159 bp)	glutelin	Fw: TTCCAATCCTTCCAGCAATC Rv: AGTATCGGCATCGGTGTAGC	1
73	HDAC3	XM_010916856.2	C03: 4083972-4084211 (240 bp)	histone deacetylase 19-like	Fw: TCCTGGAACGGGAGACATAC Rv: ACACACTCTGCATGGCCTTT	3
74	HOLOS	NM_001304427.1	C01: 2788014-2788219 (206 bp)	holocarboxylase synthetase	Fw: GCATCCACATGTCCAAACTG Rv: CATAATATCCGCTCCCCTGA	2
75	HtC10_11102	XM_010934308.2	C10: 26439083-26439382 (300 bp)	uncharacterized LOC105053217	Fw: CCATTGATACCTAAAGCTGGAGA Rv: GGACATGACCAAGCTTGAAA	3
76	HtC2_11412	XM_010929998	C08: 25294193-25293999 (195 bp)	pyrophosphate-energized vacuolar membrane proton pump	Fw: AGGCAGTTCAACACCATTCC Rv: AACAAGAGAGCCTGCAAGGA	1
77	HtC2_1255C2-411	XR_830848.2	C02: 43975808-43976030 (223 bp)	ferredoxin-thioredoxin reductase catalytic chain	Fw: AGGCAAACTTGGAGCACAGA Rv: TCAGAAATGTTTCCCGCATC	1
78	HtC2_7081	XM_010915183.2	C02: 44068212-44067983 (229 bp)	peroxidase 63	Fw: GGAGGTGCGACAACTTCAAT Rv: TGTGGTGACAAACTGCAGAA	3
79	HtC4_2106	XM_010919763.2	C04: 4078977-4079257 (281 bp)	3-isopropylmalate dehydratase large subunit	Fw: CCATATACAGCTGCGGCTTC Rv: ACCCTACAGTCCCCACCTCT	1
80	HtC4_240	XM_010920216.1	C04: 24573578-24573296 (283 bp)	membrane steroid-binding protein 2	Fw: AGACCATGGCGCTAAATCAG Rv: CCTCCCCTCATCTGAAAAGA	3
81	HtC4_4489	XM_010919778.2	C04: 4462811-4462530 (282 bp)	probable methyltransferase PMT21	Fw: ATCCTGCGACCAAATGGATA Rv: TGCACTGAATTGCATATTTCC	1

82	HtC7_9200	XM_010913888.1	C06: 41269444-41269579 (136 bp)	PP2A regulatory subunit TAP46	Fw: TGCAACTTTTGCTCAGGATG Rv: TGCCAATCTTTCCCTCTCAC	1
83	HtC8_1026C1-144	XM_010929898.1	C08: 23228408-23228684 (277 bp)	GPI-anchored protein LLG1	Fw: CTCTGCTGTGCTGCTCTCAC Rv: ATGCCAGAACCATTCCAGAT	1
84	HtC8_11217	XM_010929897.2	C08: 23297262-23297520 (259 bp)	uncharacterized WD repeat-containing protein C2A9.03-like isoform X2	Fw: CCTCTCCAAATCCATTGCTG Rv: GCGGCTGTATTGAAGGAGAC	1
85	IA2	XM_010925757.2	C06: 27105094-27104931 (164 bp)	jasmonic acid-amido synthetase JAR1	Fw: GCAGGCGGAAGCTAATACTG Rv: GATCACGTAGTGGCCTGGAT	2
86	JC19	XM_010922207.2	C05: 513397-513194 (204 bp)	Tropinone reductase	Fw:TGCTGGGACAAATATAAGGAAAA Rv: GTCGCTGCATAAATGGTTCC	1
87	JC35	XM_010938902.2	C13: 22806861-22807087 (227 bp)	peptidyl-prolyl cis-trans isomerase nima-interacting 4-like	Fw: CCATGTGGGGTAAATCCTTG Rv: GAATGCCCATCAGGAAAGAA	2
88	JC41	XM_010933280.1	C10: 11707966-11707723 (244 bp)	zerumbone synthase-like	Fw: AGAAAGCACGGTGCAAAGAT Rv: ATTCATTGAAGTCGGCATCC	1
89	JC47	XM_010912944.2	C01: 5753178-5753378 (201 bp)	Vacuolar Processing Enzyme	Fw: CATCAGGGCCACTATCAACC Rv: TTGTGCCCCCTTTTTGTTAG	2
90	JC55	XM_010923296.2	C05: 14759213-14759438 (226 bp)	probable ethylene response sensor 1	Fw: CTTTGTGGAGCAACCCATCT Rv: GTGTCCGTATCAAGCCCATT	1
91	JC59	XM_010920313.2	C04: 27102747-27102946 (200 bp)	serine carboxypeptidase-like	Fw: GACATTAGGAAGCAGTGCGAAG Rv: CCATCCTCAAGAAGAGCAGGA	2
92	JC8	XM_010926273.2	C06:	malonate--CoA ligase-like	Fw: AGGGCACTGATGGAATGAGA	2

			36747648-36747487 (162 bp)		Rv: TCATCACATGTAGAAGCTCTGC	
93	LCY-B	XM_010944289.2	U01: 25121988-25122248 (261 bp)	lycopene-b-cyclase	Fw: TGGAAGAATCTGTGGCCCAT Rv: AACCAAAGGAAGCGTACCCT	3
94	LF_a	XM_010910529.2	C09: 30683275-30683428 (154 bp)	lipase-like PAD4	Fw: GCAGTCCGAATTTGAATGGT Rv: TTTTGGAGTCCTTCCTCGAA	2
95	LIPOIC	XM_010927965.2	C07: 18431940-18432130 (191 bp)	lipoyl synthase	Fw: AATATGCTCCTCCGGGAACT Rv: GTCGAATGCTTCTGGGGTAA	1
96	M14540	XM_010934428.1	C01: 34062235-34062416 (182 bp)	ubiquitin-conjugating enzyme 15-like	Fw: TGAGCTTCACAAGCTTTTCGT Rv: TGTGACACAGGAGTATTAGGTGA	3
97	M2200	XM_010938349.2	C13: 12503327-12503494 (168 bp)	uncharacterized protein	Fw: AGGATCACCAATGCCTACAA Rv: AGCAAGTATGTTGGCACTTCA	2
98	M2252	XR_002165175.1	C07: 12193406-12193211 (196 bp)	two-component response regulator ORR21	Fw: CTTGCATTGAGGCTTGTTGA Rv: AATTCTTTGTCTGGCGTTGG	3
99	M23551	XM_010937402.2	C12: 25645701-25645899 (199 bp)	mannan synthase 1-like isoform X1	Fw: TACGGCATCGTTCTCATCAA Rv: TGGTGGAGTTGGAGTGTCAG	3
100	M3117	XM_010925967.2	C01: 21688440-21688636 (197 bp)	2-hydroxyacyl-CoA lyase-like	Fw: GCCTCGGATAAAACCCTAGC Rv: GATCAGATGGCTGCTGTGAA	3
101	M3256	XR_002165148.1	C07: 12405669-12405851 (183 bp)	T-complex protein 1 subunit delta	Fw: ATCCCAGCACTGATCTCACC Rv: TGTCAAGGCAACTAGGAGCA	1
102	M3609	XM_010938346.2	C13: 12827505-12827646	glycosyltransferase, putative	Fw: GCAGAAATCAGGGTGACCTC Rv: CACGACGGAGATTGTTGTGA	2

			(142 bp)			
103	M43144	XM_010917491.2	C03: 12945319-12945139 (181 bp)	GTPase Der	Fw: GGGAAGGCACAACTTTCAGA Rv: TTTTGTCTGATCGGGCTCTC	3
104	M43696	XM_010928111.2	C07: 16118939-16118778 (162 bp)	ethylene-responsive transcription factor ERF014-like	Fw: TATGATCGGGAGGGTGGTAG Rv: GGAGTGAGGATGAGGAGCTG	3
105	M43898	XM_010926282.2	C06: 36855343-36855184 (160 bp)	protein tesmin/TSO1-like CXC 5-like	Fw: TACAGGGCATTTCCATGGTC Rv: GATGCCATTGCAAGAAGTCA	2
106	M4585	XR_831108.2	C03: 5078622-5078468 (155 bp)	DEAD-box ATP-dependent RNA helicase 20	Fw: GGAGCACCTCGACCCATT Rv: CGCATTGGTCGAACAGGAAG	2
107	M4883	XM_019850224.1	C04: 4188649-4188886 (238 bp)	protein enchanced disease resistance 2	Fw: TCCTGTTCAGGCTCTGGAAT Rv: CAGGGTTGCGTGAATGGTT	3
108	M6256	XM_019846597.1	U02: 80801398-80801584 (187 bp)	DUO pollen 3 (DUO3) gene	Fw: TCCTCCTGACCTGTCTGTCC Rv: GTATCTGCAAACGCACGAGA	2
109	M6ASA	XM_010930131.2	C08: 27391101-27391282 (182 bp)	Microsome localized omega-6-desaturase	Fw: AAGACGCCCTTCTTCTCTCC Rv: GCGCACCACCTCTTCTCTAC	1
110	M7467	XR_003388267.1 (Date palm)	U07: 44621406-44621548 (143 bp)	pentatricopeptide repeat-containing protein	Fw: TACTTGACATGGTGCGTGGTT Rv: CCTCCCTTGATTGCTTCATC	1
111	M8373	XM_010940580.2	C14: 23021215-23021047 (169 bp)	polyadenylate-binding protein RBP47B	Fw: CGCCCAGTATTTGGATCAGT Rv: ATTGGGCCTCTTTTGGTCTT	2
112	M847	XM_010927799.2	C07: 12154005-12153789 (217 bp)	microtubule-associated protein 70-1-like	Fw: TGACCTAGTGCCACCATCAA Rv: TAGCAACCGCACCAATGTAG	1

113	M9861	XM_010910705.2	U05: 84508671-84508856 (186 bp)	endoglucanase 10	Fw: GCAATACCGGGAGTGGTAGA Rv: TGGTGCAAGGACACTTTTCA	2
114	MADS11-1	XM_019851612.1	C07: 16866616-16866465 (152 bp)	MADS-box protein JOINTLESS	Fw: GGCGAGGGAGAAGATTCAG Rv: CGGTGGAGGAGAAGATGATG	3
115	mEg3275	XM_010912620.2	C02: 7612917-7613050 (134 bp)	serine hydroxymethyltransferase 7-like	Fw: GAAGCCTGAGACCGCATAGA Rv: TTCGGTGATGAAGATTGAAG	2
116	MUM4	XM_010917136.2	C03: 8044490-8044684 (195 bp)	trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM1	Fw: CACCGTGGAGAAGTTGGACAT Rv: CTCTGACCACCCCAAATTCT	1
117	O3FAD	XM_010922542.2	C05: 4454772-4454530 (243 bp)	omega-3 fatty acid desaturase	Fw: AAGTAGCGGGGAGGAGAGAG Rv: ACCAAACAAAAAGCCCTCCT	2
118	OLEOSIN	XM_010935827.1	C11: 23477579-23477749 (171 bp)	oleosin 16 kDa-like	Fw: AATCTCCCCTCGCTTCACTT Rv: CAACACTCACGCTACGAGGA	3
119	OLEOYL	XM_010926998.2	C07: 1586207-1586053 (155 bp)	palmitoyl-acyl carrier protein thioesterase	Fw: AAGCAGTGGACCCTTCTTGA Rv: TCTATAGAAGCCGTCCGATCA	1
120	PACT	DQ422858.1	C07: 1586069-1586260 (192 bp)	palmitoyl-ACP thioesterase	Fw: GATCAGCCCCGATCTCATAC Rv: CTGACTGGAGCGTGCTTCTT	1
121	PAT_1	XM_010910419.2	U05: 40088262-40088420 (159 bp)	5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase 1	Fw: GCGAGGGAGTGAAATATGGT Rv: GTCTTGAGCCCACAGTCAGG	1
122	PAT_11	XM_010936324.2	C12: 9940907-9941065	heat shock protein 83-like	Fw: GGTGGATGCTATCGACGAGT Rv: ACCTTGCATAGGCTCTCGAA	1

			(159 bp)			
123	PAT_12	XM_010940376.2	C14: 11183666-11183860 (195 bp)	nascent polypeptide-associated complex subunit alpha-like protein 1	Fw: GCAAAGATCGAGGACCTGAG Rv: CTTGGACCTCGAAACTCCAG	1
124	PAT_13	XM_010940239.2	C14: 14557627-14557473 (155 bp)	coatomer subunit epsilon-1	Fw: CCGACCACCTCTTCAATCTC Rv: GACCTGGTAGGAGCCAAGG	1
125	PAT_14	XM_010943249.2	C07: 36442648-36442806 (159 bp)	60S acidic ribosomal protein P0	Fw: CAAGGTTGGCTCTTCTGAGG Rv: TCCAGCTGCAAACTTCTCAA	2
126	PAT_15	XM_010930830.2	C08: 36345443-36345620 (178 bp)	temperature-induced lipocalin-1	Fw: CCAAGAACGGGGAGAACAC Rv: AAGAAGGGTGGCACGTAGAA	2
127	PAT_2	XM_010932692	C09: 34724992-34725149 (158 bp)	actin-101	Fw: ATTCCGGTGATGGTGTGAGT Rv: CCGTTCTGCAGTGGTAGTGA	2
128	PAT_3	NM_001319906.1	C02: 12398335-12398515 (181 bp)	actin-3-like	Fw: CACTTCCTCATGCCATCCTT Rv: GCAGACTCCAATTCCTGCTC	2
129	PAT_4	XM_010936612.2	C12: 15018134-15017947 (188 bp)	caffeic acid 3-O-methyltransferase	Fw: TGTTCAATGAGGGCATGAAG Rv:AGAGATGACATGAGGAAGATCAA	1
130	PAT_6	XM_010930111.2	C08: 27075380-27075546 (167 bp)	probable plastid-lipid-associated protein 2	Fw: CTCCATCGTTTTACCCGAGA Rv: AGGACTGTGCATTGTCGTTG	2
131	PAT_7	XM_010929278.2	C08: 5260692-5260513 (180 bp)	fructose-bisphosphate aldolase 1	Fw: TCCGTGAGCTCCTCTTTTGT Rv: TAGTGCCAGCAAGTTCGATG	2
132	PAT_8	XM_010919548.2	C03: 57312509-57312669 (161 bp)	5-methyltetrahydropteroyltriglutamate--homocysteine	Fw: GATCCCATCCACAGAGGAGA Rv: GGAGGAGCTTAGCAGCAGAA	2

				methyltransferase 1		
133	PAT_9	XM_010939180.2	C13: 27325349-27325185 (165 bp)	PLAT domain-containing protein 3-like	Fw: ATCAGGACGGGGTCCATCT Rv: GCTGAAGATGTCGAGGTTGC	3
134	PDAT_2	XM_010918834.2	C03: 29958824-29958974 (151 bp)	Phospholipid-Diacylglycerol acyltransferase	Fw: TTCCCTGTAACTGAGAAGCAAA Rv: CAGAATGCAAAATCAGAACAAAA	3
135	PDHB	XM_010942881.2	C01: 51857666-51857866 (201 bp)	pyruvate dehydrogenase E1 component subunit beta	Fw: TCACAGCGGTTGGAATCATA Rv: CTCTGCCTCCTCCAGACAAC	1
136	PDS3_CH13	XM_010938924.2	C13: 22675283-22675467 (185 bp)	pre-mRNA-splicing factor clf1-like	Fw: GAACCAAAGAAGCTCGCCTG Rv: CGCCTTTGCCAGCTCAATTA	1
137	PKP-ALPHA	XM_010937608.1	C01: 40816787-40816570 (218 bp)	pyruvate kinase isozyme A	Fw: ACAAGCCTGTCATTGTAGCT Rv: CTTCTCCTCTCTCCACCACC	1
138	PLT2	XM_010914736.1	C09: 15548387-15548160 (228 bp)	ethylene-responsive transcription factor	Fw: GGGGCATCATGGGGAAATTC Rv: CTGGGTCTGGTTTGCTTCAG	3
139	PO3_5-10	XM_010934991.2	C11: 11856922-11857078 (157 bp)	GEM-like protein 7	Fw: ACGTCGAGTGAGAATCTGGA Rv: TTCCGCAGAAAGGTCATTGT	3
140	PO3_5-13	XM_010941549.1	C15: 19816511-19816313 (199 bp)	proteasome subunit alpha type-5	Fw: TCTGTTTGCCTTCTCCACCA Rv: GCATTGGCGGTAGAACTCTG	3
141	PO3_5-14	XM_010941553.2	C15: 19862793-19862514 (280 bp)	1-phosphatidylinositol-3-phosphate 5-kinase FAB1A-like	Fw: AACACCTAGAGACGTGGGTG Rv: ACCCCAGAAAGATTGGTCGT	3
142	PO3_5-7	XM_010917092.2	C03: 7306161-7306358 (198 bp)	NADP-dependent malic enzyme	Fw: TTGGCTAGTCATCTCCCTCG Rv: CCCAATGATCAAGGGGCTCA	3

143	PO3_5-8	XM_019849509.1	C03: 8074053-8073904 (150 bp)	MADS-box transcription factor 3 isoform X2	Fw: AGATTGCTGAGAATGAGAGAGC Rv: TCGTCTGCTGCTGATGAGAG	3
144	PRT6	XM_010939321.1	C14: 886973-887164 (192 bp)	E3 ubiquitin-protein ligase PRT6-like	Fw: TGCGTTCCATGTTTCCAGAA Rv: ATCCAGAACAGGTCCCACAG	3
145	PSII1	EU016942.1	CT: 75549-75755 (207 bp)	photosystem II phosphoprotein (psbH)	Fw: TGGCTACACAAACCGTTGAG Rv: TTCCATCCAGTAAAACGAAAGAA	2
146	PSII2	EU016919.1	CT: 75289-75438 (150 bp)	photosystem II protein N (psbN)	Fw: TGGAAACAACAACCCTAGTCG Rv:GGGAGACTCATTACTTCAACTAGTCC	2
147	PSII3	EU016895.1	CT: 65546-65717 (172 bp)	photosystem II cytochrome b559 alpha subunit (psbE)	Fw: TTTGTGGAGCTCAGCATGTC Rv: TTGGTCGAGGACTTCCAAAC	2
148	PYRKIN	XM_010942455.2	C01: 50296952-50297101 (150 bp)	pyruvate kinase 1	Fw: GGATACGGTGGGTCCAGAG Rv: GCCTTTGACAATCCACTGAAA	3
149	QM	XM_010939750.2	C14: 5029626-5029785 (160 bp)	60S ribosomal protein L10	Fw: GCTCTTGAGGCTGCTCGTAT Rv: CCCTCATTCCAGTCTGAAGC	2
150	R2r3	XM_019854941	C14: 1636715-1636911 (197 bp)	myb-related protein MYBAS1-like	Fw: TGAGCTTCGAGGGATACAAGA Rv: TGGACAGGGGTAGAAAGAGAA	1
151	RAP2.2_3	XM_019854559.1	C13: 22032414-22032123 (292 bp)	ethylene-responsive transcription factor	Fw: GGCTTCATTCAGAGGACCCT Rv: ACTTGCAAGCTCTCATATCAACT	3
152	RPL10	XM_010939750.2	C14: 5029479-5029721 (243 bp)	60S ribosomal protein L10	Fw: AAGCCATACCCAAAGTCACG Rv: ATGGAAGGGATGCACTCTCA	3
153	RU1	EU016944.1	C02:	ribulose bisophosphate	Fw: CAGGGGGTATTCATGTTTGG	2

			62056727-62056886 (160 bp)	carboxylase	Rv: TTCACGAGCAAGATCACGTC	
154	SEQUI	XM_010906840.2	U02: 19591209-19591378 (170 bp)	alpha-humulene synthase-like, transcript variant X2	Fw: TCCATGGAAAGCCATATGAA Rv: TGAAAATCCAACTTTGCAAGC	2
155	SHELL	XM_010909778.2	C02: 3056550-3056256 (295 bp)	MADS-box transcription factor 21	Fw: GGATCGAGAACACCACAAGC Rv: AATTTGGCTTGGCCATAGAA	1
156	SHELL2	XM_010909778.2	C02: 3056550-3056256 (295 bp)	MADS-box transcription factor 21	Fw: TAGCAGAGAATGAGCGAGCA Rv: TCAGACAAGTCTTCTAACACACCTTT	1
157	SQUA3	AF411842.1	C15: 13726940-13726660 (281 bp)	MADS box transcription factor (SQUA3)	Fw: AGGCACTAGTTTGCCTGCAT Rv: TTTGAGCTCCAAAGCCAACT	1
158	TO1	JN848783.1	U02: 79752127-79752276 (150 bp)	gamma-tocopherol methyltransferase	Fw: GCACCAGGAGCCACCATTAT Rv: CACATAATCACTGGCTGAGCA	1
159	TO2	XM_019848829.1	C03: 13878917-13878801 (117 bp)	tocopherol cyclase	Fw: GGGAATACAGCACACATCCA Rv: CCATGCATACTTGCCAATGA	2
160	TO3	XR_831277.2	C03: 13885380-13885529 (150 bp)	probable tocopherol cyclase	Fw: AAGGTCTCGATCCCTGAATG Rv: ATCATCCGCACCAAGAATTT	2
161	VVuACT	XM_010939417.2	C14: 1766743-1766542 (202 bp)	actin-3	Fw: CCACAACAGCAGAACGAGAA Rv: CCACAACAGCAGAACGAGAA	3
162	Wild-type_VIR	XM_010932909.2	C01: 29321468-29321727 (260 bp)	virescens R2R3-MYB gene	Fw: TGGTCAGAAGATCAGCAATCA Rv: CAAAGCAAGTCATCCCATCC	1
163	WOS104	XM_010926998.2	C07: 1586323-1586513	palmitoyl-acyl carrier protein thioesterase	Fw: TTCCCCACACCATCTTTCTC Rv: GATTCAGCTTTCAGGCCAAC	2

			(191 bp)			
164	WOS6942	XM_010924566.2	C05: 40852751-40852568 (184 bp)	NAC protein 1	Fw: GTTCCCGGACTTTGACGATA Rv: AGCCATGCATGTACTGTGGA	2
165	wri1	XM_010928170.2	C07: 15188796-15188619 (178 bp)	DELLA protein SLR1-like	Fw: TGGTGAAGCAGATCTCGATG Rv: TAGGGGCAGCTCTCGTAGAA	3
166	ZCD	XR_002166117.1	C01: 37600361-37600171 (191 bp)	zeta-carotene desaturase	Fw: GGCACCCTGAGAGATTCAAA Rv: TGAACGACATGGGAAAGACA	1
167	EOCHYB	XM_010920813.1	C04: 37534421-37534541 (121 bp)	beta-carotene 3-hydroxylase 2	Fw: CAGAACCGGAGTTCGGAGAT Rv: GCTTCCTCGCGATCTTCTC	2
168	**PAT_12_ML***	XM_010923558.2	C05: 25085015-25085209 (195 bp)	nascent polypeptide-associated complex subunit alpha-like protein 1	Fw: GCAAAGATCGAGGACCTGAG Rv: CTTGGACCTCGAAACTCCAG	1
169	**HtC2_11412_ML***	XM_010915041	C02: 41981898-41982092 (195 bp)	pyrophosphate-energized vacuolar membrane proton pump-like	Fw: AGGCAGTTCAACACCATTCC Rv: AACAAGAGAGCCTGCAAGGA	1
170	**PAT_2_ML***	XM_010914104	C02: 23775797-23775954 (158 bp)	actin-101	Fw: ATTCCGGTGATGGTGTGAGT Rv: CCGTTCTGCAGTGGTAGTGA	2
171	**ATAGB1_ML***	XM_019854409	C13: 103406-103595 (190 bp)	guanine nucleotide-binding protein subunit beta	Fw: ATTCACTGGATTGGGCTCCT Rv: CATGCACTATCAAGACCACCAC	1

_ML*= Multi Locus CG**. Legend: No**: consecutive number of the CG; **CG Name**: internal name of the CG; **GeneID_NCBI**: identifier in the nucleotide data base of NCBI; **CGpos_MPOB (Amplicon_length)**: CG position according to MPOB´s Oil Palm reference genome on Chromosome (**Ci**), unassigned Scaffold (**Ui**) or chloroplast gene (**CT**) and amplicon length; **CG function**: function of the CG indicated in the nucleotide database; **Amplicon primers**: forward and reverse primers used for producing amplicons from each CG; Library No: library number in which a particular CG was included.

Table A 2: Mean values, standard deviations (SD), minimum and maximum values of each analysed trait, and ANOVA significance levels between the different origins of oil palm hybrids.

Production traits	**Mean**	**SD**	**Min**	**Max**	**ANOVA**
BN [nº]*	50.36	20.81	1.00	87.00	***
BW [Kg]	11.28	3.15	2.65	18.30	***
BY [Kg]*	600.14	320.35	5.30	1280.90	***
OilfM [%]	29.90	5.52	15.30	45.18	***
OildM [%]*	54.49	8.41	35.67	77.68	***
OilB [%]	19.10	4.99	5.71	32.16	***
Oil quality traits	**Mean**	**SD**	**Min**	**Max**	**ANOVA**
Sat [%]*	37.30	4.33	20.07	45.26	***
Mono-Un [%]*	48.95	5.92	37.46	85.46	***
Poly-Un [%]	13.53	1.64	9.93	17.33	***
OA [%]*	47.36	5.71	35.00	65.20	***
IV [cg/g]*	64.30	4.17	54.62	81.39	***
SSS [%]*	1.28	0.92	0.10	8.50	
SUS [%]*	23.69	4.39	8.93	33.20	***
SUU [%]	32.28	4.25	21.40	45.22	***
UUU [%]*	12.84	5.80	3.34	31.95	***
Tocph [ppm]*	214.47	96.51	18.30	624.20	***
Alpha [ppm]*	151.10	75.71	18.30	467.40	***
Delta [ppm]*	43.10	16.20	10.50	98.30	
Gamma [ppm]*	45.78	13.52	28.40	131.90	
Toc3 [ppm]	1149.75	367.84	306.80	2096.40	***
Alpha3 [ppm]	324.82	142.99	48.50	743.70	***
Delta3 [ppm]*	108.06	58.54	18.60	272.20	***
Gamma3 [ppm]	720.15	200.23	211.70	1199.70	***
Toc [ppm]	1366.86	423.13	392.90	2361.5	***
Car [ppm]*	795.68	241.13	353.00	1469.00	***

Significance levels: $p < 0.001$***; $p < 0.01$** and $p < 0.05$*. Traits marked with "*" did not follow a normal distributions according to Saphiro-Wilk tests. Production traits: bunch number (BN), bunch weight (BW), bunch yield (BY), oil % in fresh mesocarp (OilfM), oil % in dry mesocarp (OildM) and oil % in bunch (OilB). Quality traits: oleic acid % (OA), saturated fatty acids % (Sat), mono-unsaturated fatty acids % (Mono-Un), poly-unsaturated fatty acids % (Poly-Un), iodine value (IV), carotene contents (Car), different types of triglycerides in % (SSS, SUS, SUU, UUU), tocols (Toc), tocopherols (Tocph) and compounds Alpha, Delta and Gamma, tocotrienols (Toc3) and compounds Alpha3, Delta3 and Gamma3.

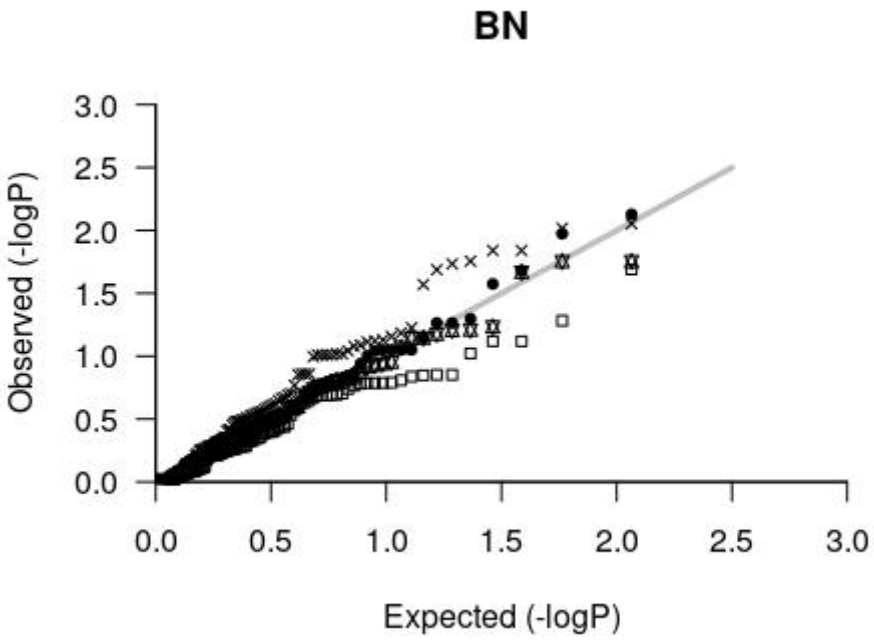
Table A 3: List of the 62 Candidate Genes (CG) targeted by single nucleotide polymorphism (SNP) which were used for the Association Mapping studies in Oil palm hybrids.

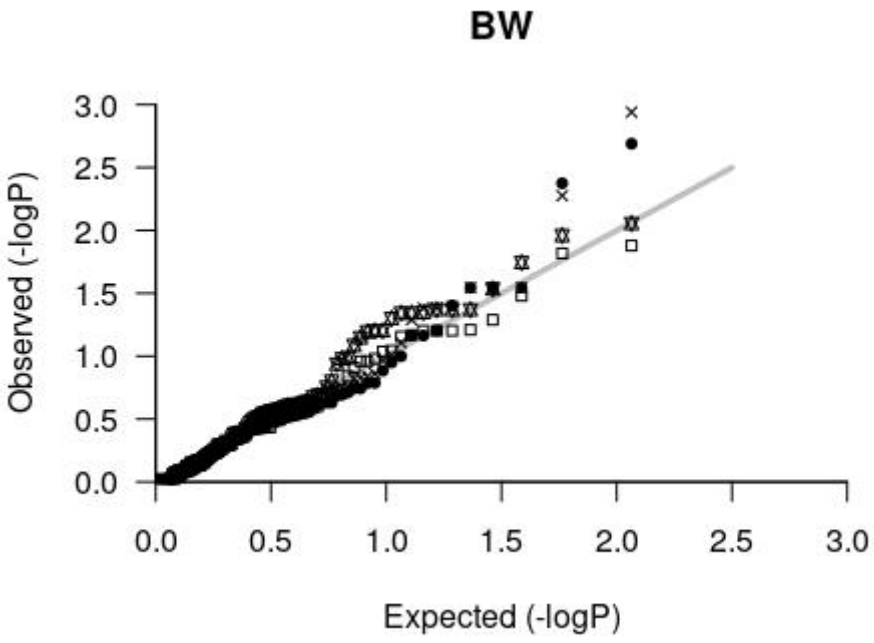
No	CG Name	GeneID_NCBI	CGpos_MPOB	CG function
1	HOLOS	NM_001304427.1	C01: 2788014-2788219	holocarboxylase synthetase
2	JC47	XM_010912944.2	C01: 5753178-5753378	Vacuolar Processing Enzyme
3	M14540	XM_010934428.1	C01: 34062235-34062416	ubiquitin-conjugating enzyme 15-like
4	CA4	XR_002166117.1	C01: 37606005-37605886	zeta-carotene desaturase
5	PKP-ALPHA	XM_010937608.1	C01: 40816787-40816570	pyruvate kinase isozyme A
6	PDHB	XM_010942881.2	C01: 51857666-51857866	pyruvate dehydrogenase E1 component subunit beta
7	SHELL	XM_010909778.2	C02: 3056550- 3056256	MADS-box transcription factor 21
8	PAT_3	NM_001319906.1	C02: 12398335-12398515	actin-3-like
9	PAT_2_ML*	XM_010914104	C02: 23775797- 23775954	actin-101
10	FFB2_C3566_S9	XM_019847934.1	C02: 31308535-31308364	transport protein Sec61 subunit alpha-like
11	CA3	XM_010914762.2	C02: 35978110-35978321	15-cis-zeta-carotene isomerase
12	HtC2_1255C2-411	XR_830848.2	C02: 43975808-43976030	ferredoxin-thioredoxin reductase catalytic chain
13	EgFATB2.2	XM_010916714.2	C03: 1846524-1846322	palmitoyl-acyl carrier protein thioesterase
14	HDAC3	XM_010916856.2	C03: 4083972-4084211	histone deacetylase 19-like
15	PO3_5-7	XM_010917092.2	C03: 7306161-7306358	NADP-dependent malic enzyme
16	MUM4	XM_010917136.2	C03: 8044490-8044684	trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM1
17	PO3_5-8	XM_019849509.1	C03: 8074053- 8073904	MADS-box transcription factor 3 isoform X2
18	TO3	XR_831277.2	C03: 13885380-13885529	probable tocopherol cyclase
19	DWARF7	XM_010918132.2	C03: 21298109-21298302	delta(7)-sterol-C5(6)-desaturase
20	BnC3_792	XM_010918923.1	C03: 32128327-32128119	oryzain gamma chain
21	JC59	XM_010920313.2	C04: 27102747-27102946	serine carboxypeptidase-like
22	EOCHYB	XM_010920813.1	C04: 37534421-37534541	beta-carotene 3-hydroxylase 2
23	EgDSI	AY182168.1	C04: 56149385-56149241	opsc112 protein disulphide isomerase
24	O3FAD	XM_010922542.2	C05: 4454772-4454530	omega-3 fatty acid desaturase
25	EPS3	XM_010923043.2	C05: 10242828-10242630	Peroxirredoxin 1-Cys
26	JC55	XM_010923296.2	C05: 14759213-14759438	probable ethylene response sensor 1
27	EgNAC	DQ267443.1	C05: 40852033-40852228	1 NAC protein
28	WOS6942	XM_010924566.2	C05: 40852751-40852568	NAC protein 1
29	JC8	XM_010926273.2	C06: 36747648-36747487	malonate--CoA ligase-like
30	HtC7_9200	XM_010913888.1	C06: 41269444-41269579	PP2A regulatory subunit TAP46

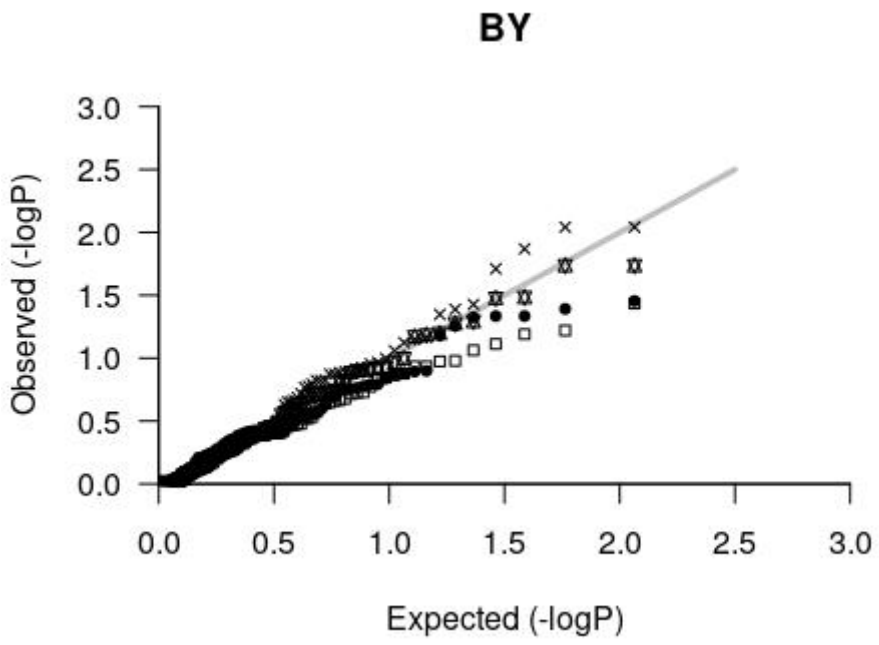
31	M847	XM_010927799.2	C07: 12154005-12153789	microtubule-associated protein 70-1-like
32	M3256	XR_002165148.1	C07: 12405669-12405851	T-complex protein 1 subunit delta
33	LIPOIC	XM_010927965.2	C07: 18431940-18432130	lipoyl synthase
34	FA4	XM_010928432.1	C07: 23768995-23768846	stearoyl-[acyl-carrier-protein] 9-desaturase 5
35	BnC8_761	XM_010929225.2	C08: 4351998-4351827	60S ribosomal protein L23
36	PAT_7	XM_010929278.2	C08: 5260692-5260513	fructose-bisphosphate aldolase 1
37	HtC2_11412	XM_010929998	C08: 25294193-25293999	pyrophosphate-energized vacuolar membrane proton pump
38	PAT_6	XM_010930111.2	C08: 27075380-27075546	probable plastid-lipid-associated protein 2
39	M6ASA	XM_010930131.2	C08: 27391101-27391282	Microsome localized omega-6-desaturase
40	PAT_2	XM_010932692	C09: 34724992-34725149	actin-101
41	BKACPII_1	FJ940767.1	C10: 22949664-22949486	beta-ketoacyl-ACP synthase II
42	PAT_11	XM_010936324.2	C12: 9940907-9941065	heat shock protein 83-like
43	ATAGB1	XR_002165879	C12: 17556353-17556542	GTP binding protein beta 1
44	GLUT1	AF261691	C12: 28135291-28135449	glutelin
45	ATAGB1_ML*	XM_019854409	C13: 103406- 103595	guanine nucleotide-binding protein subunit beta
46	M2200	XM_010938349.2	C13: 12503327-12503494	uncharacterized protein
47	JC35	XM_010938902.2	C13: 22806861-22807087	peptidyl-prolyl cis-trans isomerase nima-interacting 4-like
48	PAT_9	XM_010939180.2	C13: 27325349-27325185	PLAT domain-containing protein 3-like
49	QM	XM_010939750.2	C14: 5029626-5029785	60S ribosomal protein L10
50	DXS2	NM_001303573.1	C14: 21869601-21869411	1-deoxy-D-xylulose-5-phosphate synthas
51	GID1	XM_010940559.2	C14: 22469857-22470046	gibberellin receptor
52	M8373	XM_010940580.2	C14: 23021215-23021047	polyadenylate-binding protein RBP47B
53	ATP2	EU016918.1	CT: 53908-54078	ATP synthase CF1 epsilon subunit
54	atpB	EU016907.1	CT: 54576-54425	ATP synthase beta subunit (atpB)
55	PSII2	EU016919.1	CT: 75289-75438	photosystem II protein N (psbN)
56	SEQUI	XM_010906840.2	U02: 19591209-19591378	alpha-humulene synthase-like, transcript variant X2
57	FA8	XM_010906903.1	U02: 21412859-21413024	stearoyl-[acyl-carrier-protein] 9-desaturase 5
58	TO1	JN848783.1	U02: 79752127-79752276	gamma-tocopherol methyltransferase
59	M6256	XM_019846597.1	U02: 80801398-80801584	DUO pollen 3 (DUO3) gene
60	PAT_1	XM_010910419.2	U05: 40088262-40088420	5-methyltetrahydropteroyltriglutamate--homocysteine

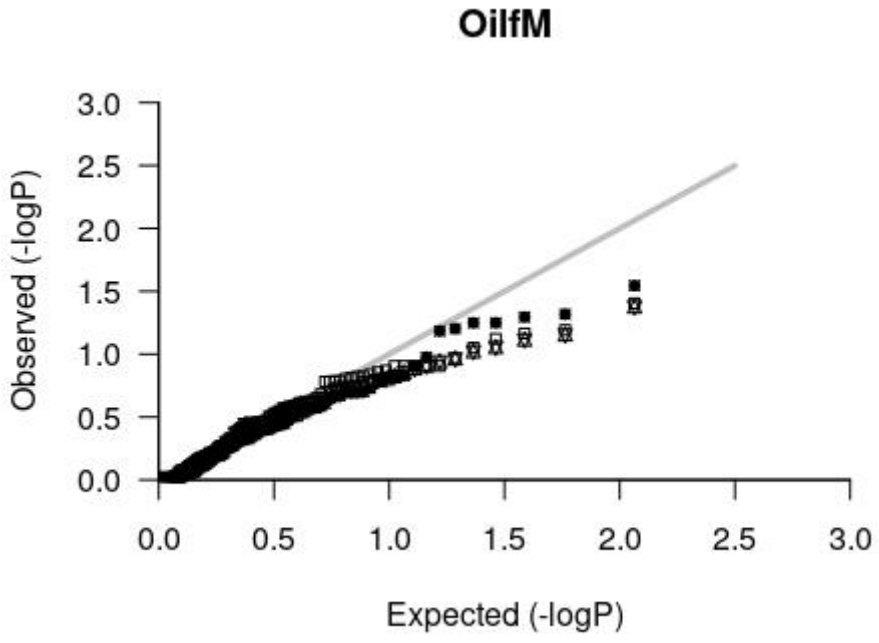
				methyltransferase 1
61	ATP3	EU016883.1	U05: 50035784-50035933	ATP synthase CF0 subunit III
62	M7467	No annotation	U07: 44621406-44621548	pentatricopeptide repeat-containing protein

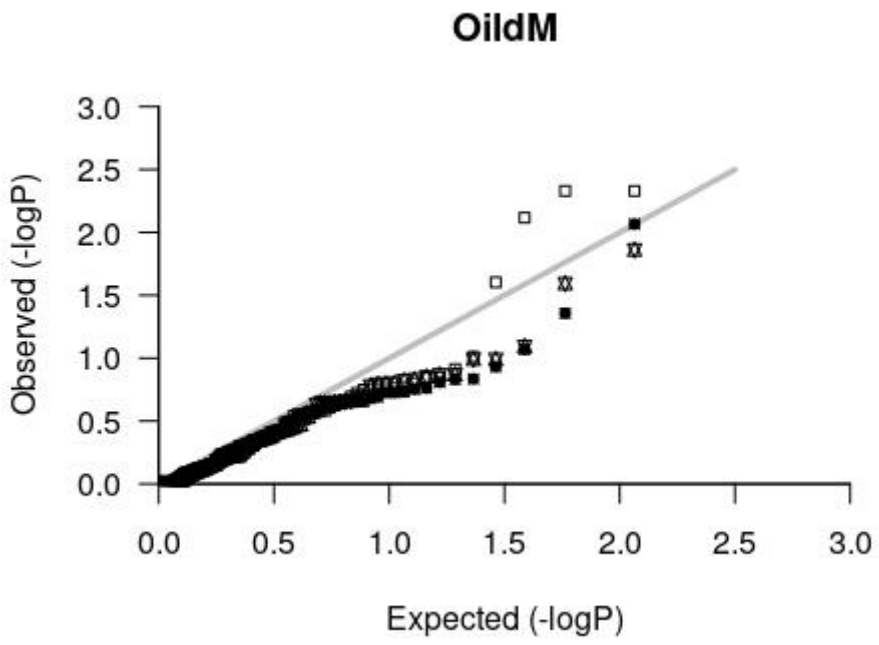
_ML*= Multi Locus CG. No: consecutive number of the CG; CG Name: internal name of the CG; GeneID_NCBI: identifier in the nucleotide data base of NCBI; CGpos_MPOB: CG position according to MPOB´s Oil Palm reference genome on Chromosome (Ci), unassigned Scaffold (Ui) or chloroplast gene (CT); CG function: function of the CG indicated in the nucleotide database; Amplicon primers: forward and reverse primers used for producing amplicons from each CG.

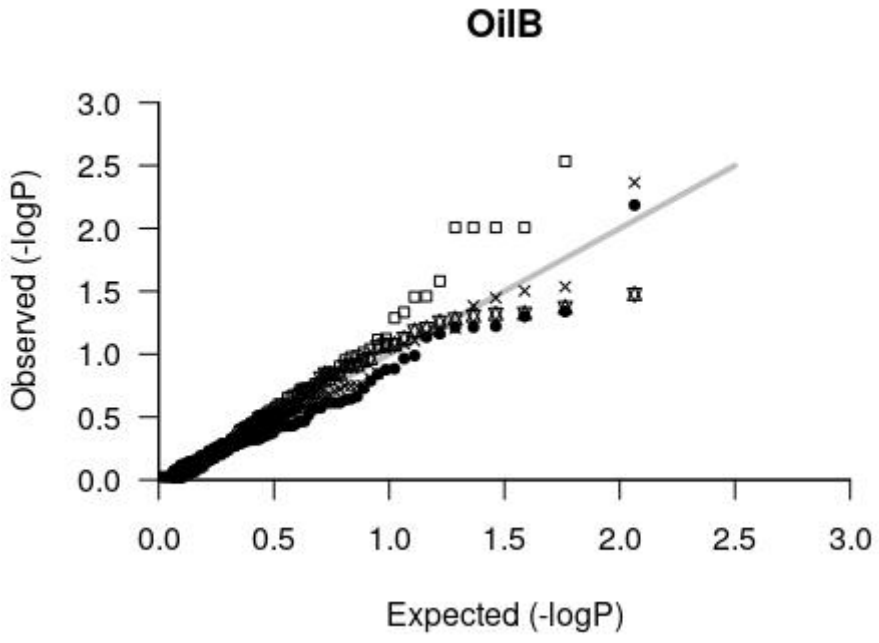

BN
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

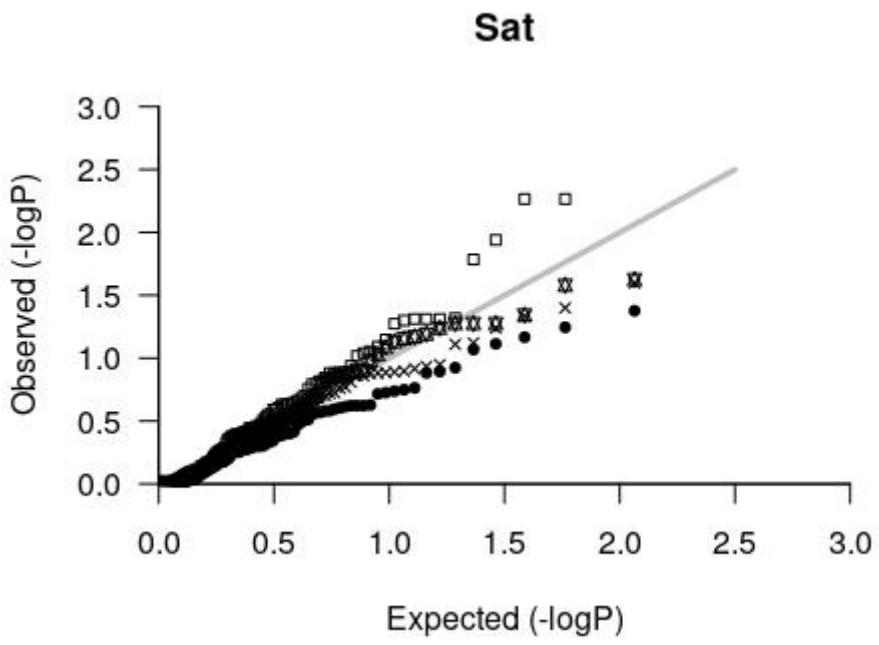

BW
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

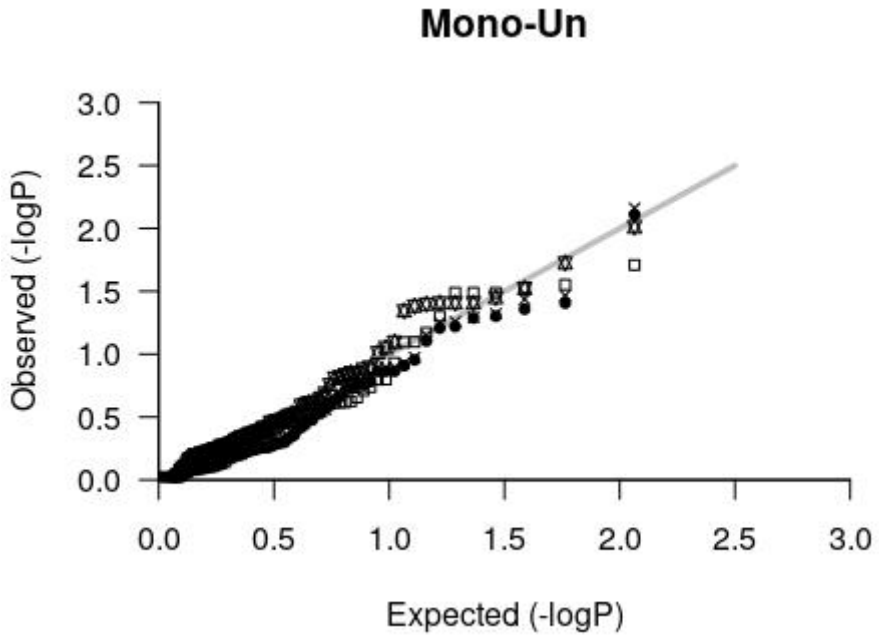

BY
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

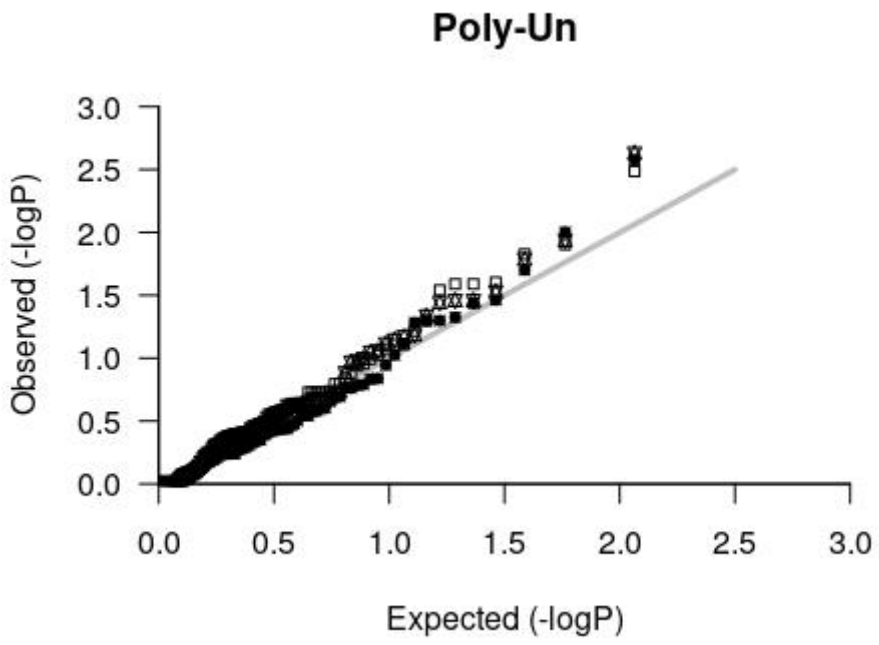

OilfM
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

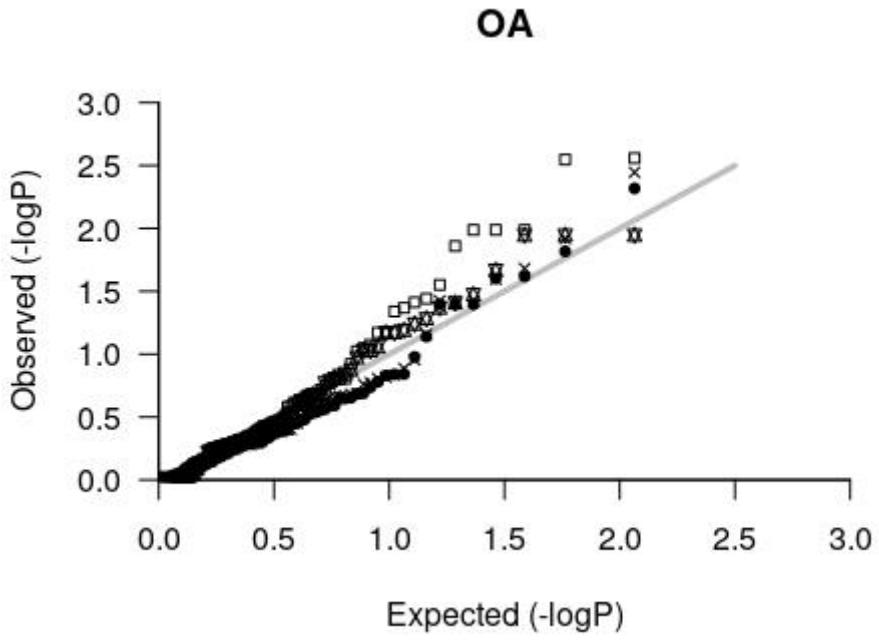

OildM
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

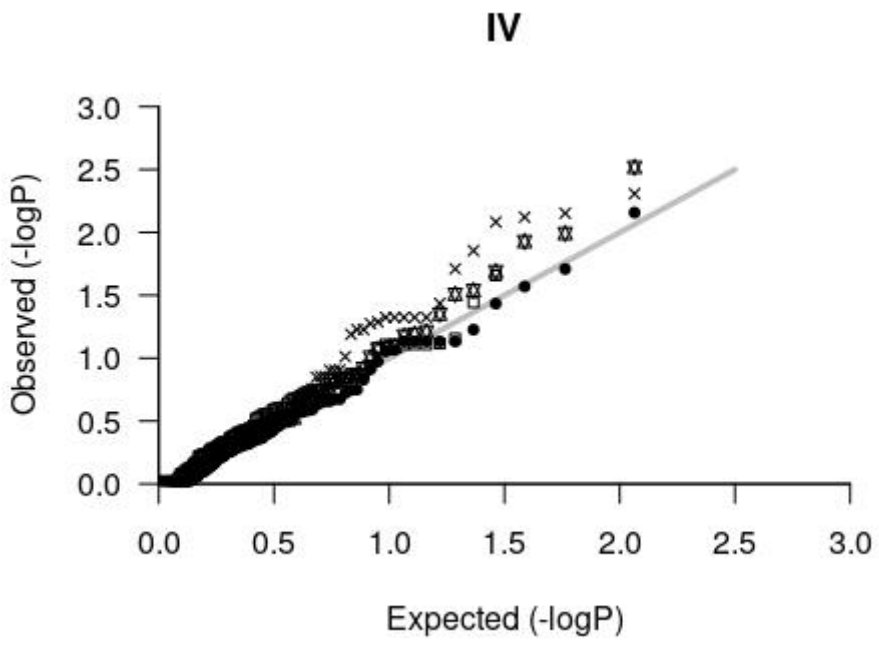

OilB
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

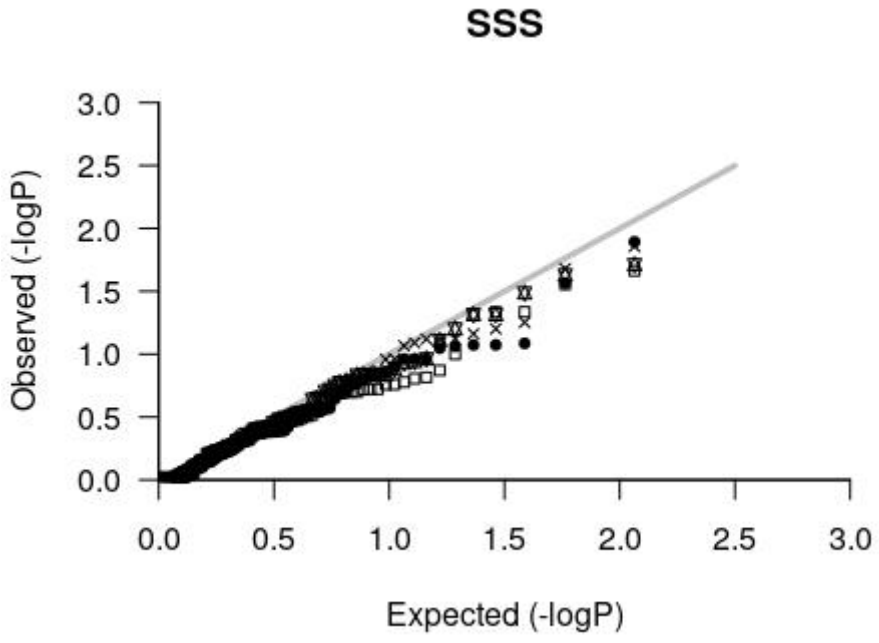

Sat
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

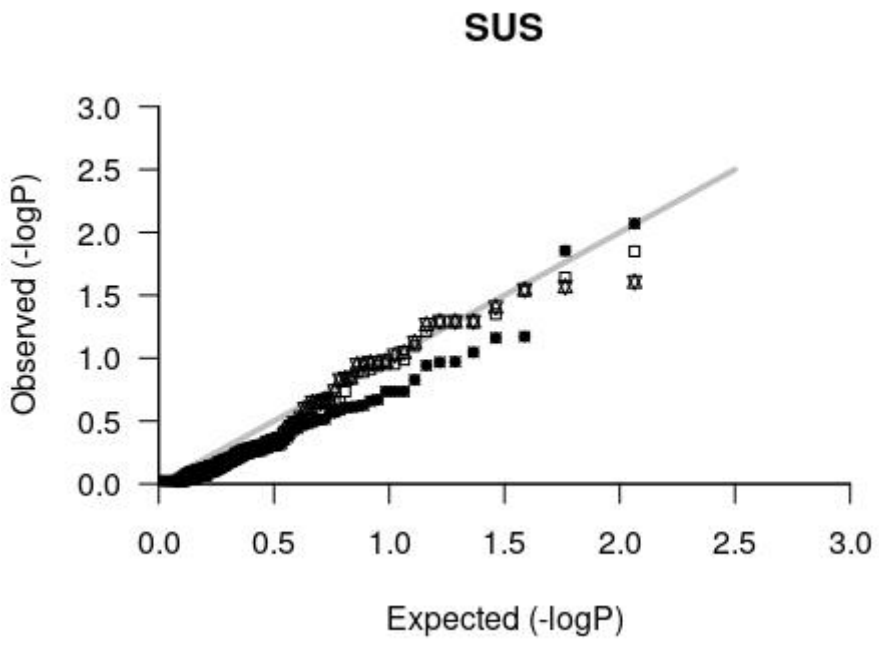

Mono-Un
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

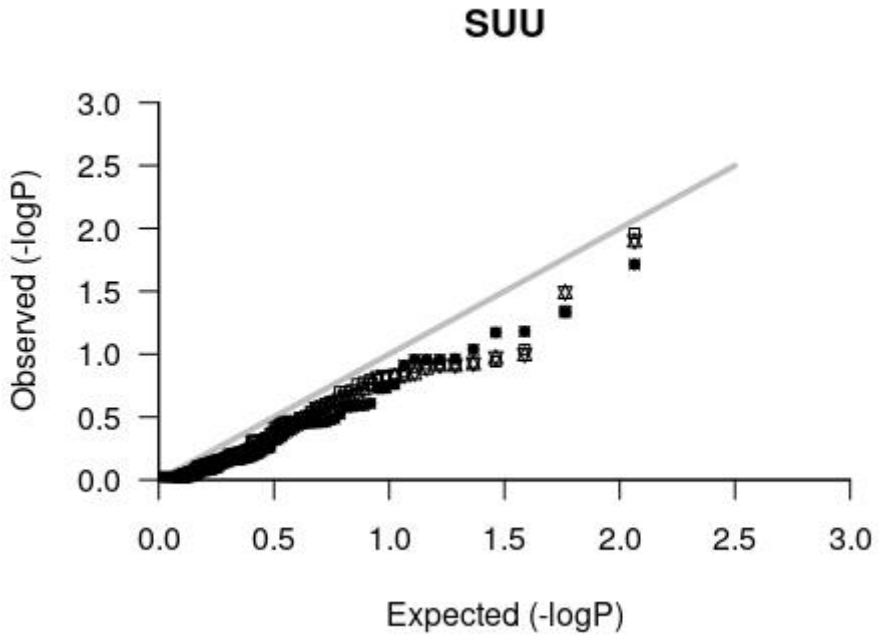

Poly-Un
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

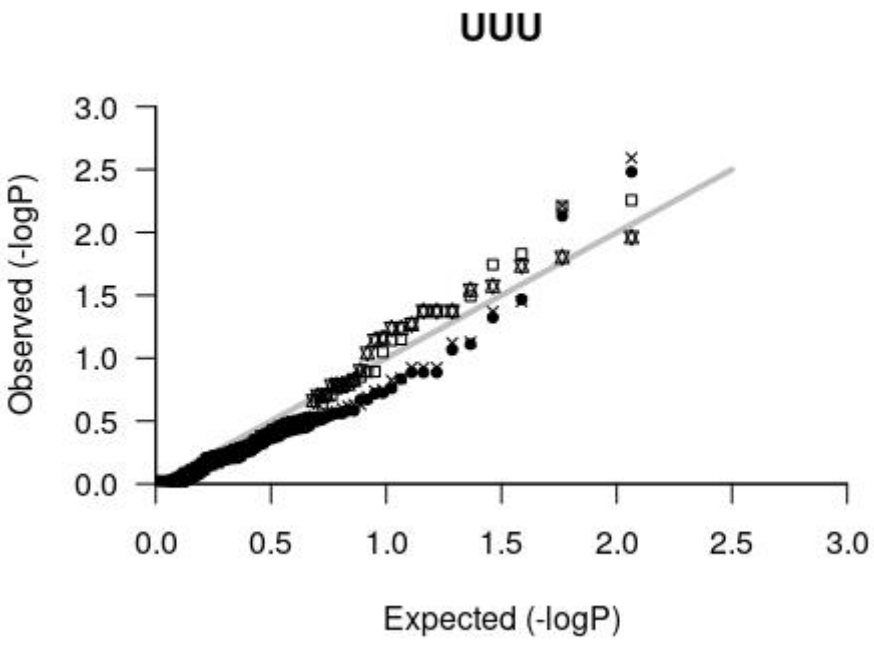

OA
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

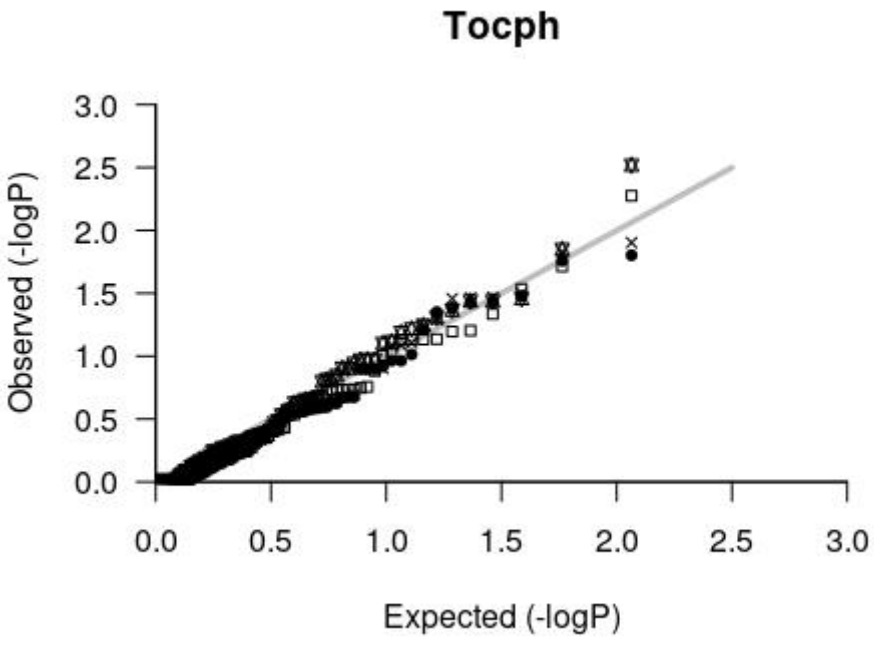

IV
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

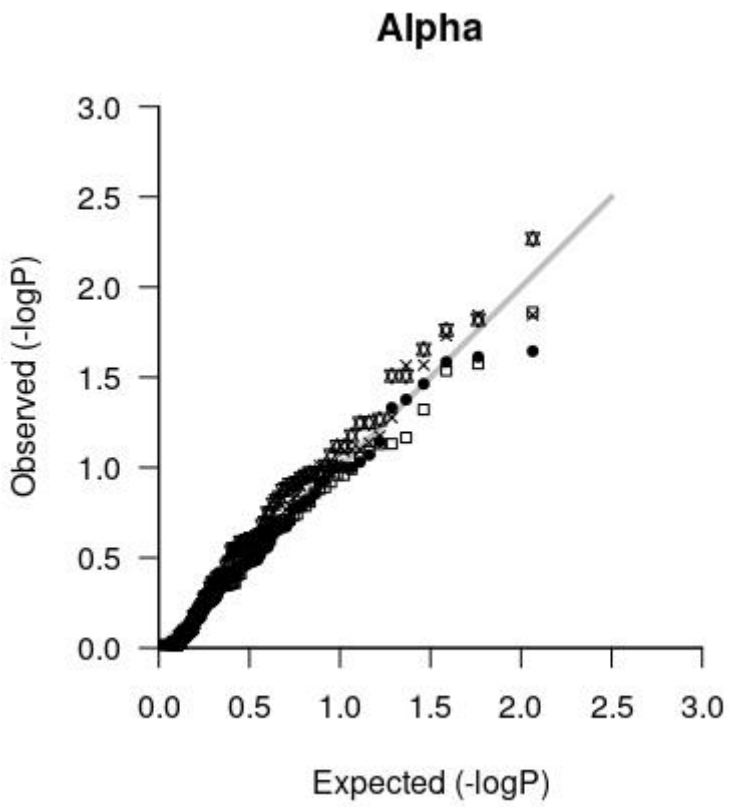

SSS
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

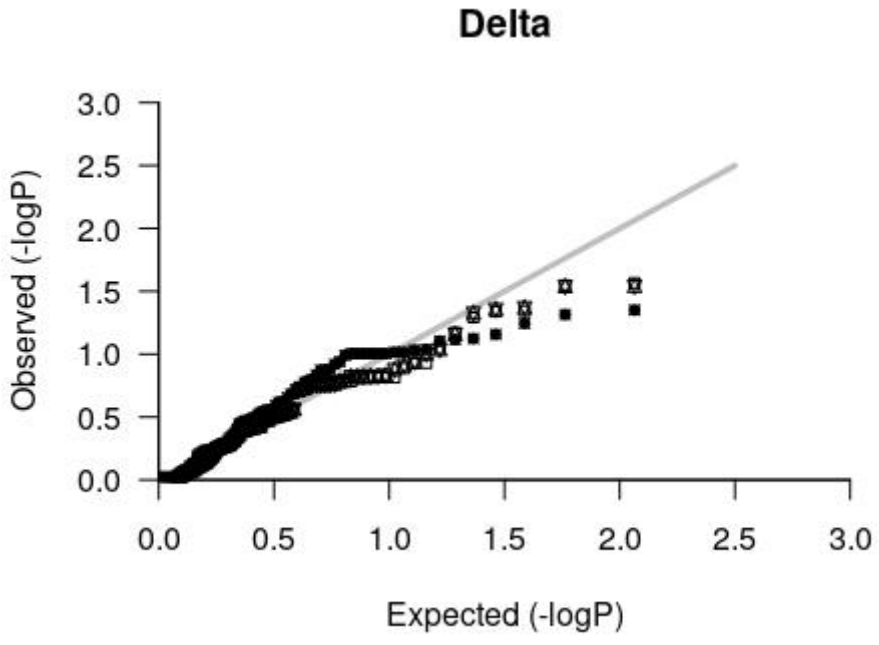

SUS
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

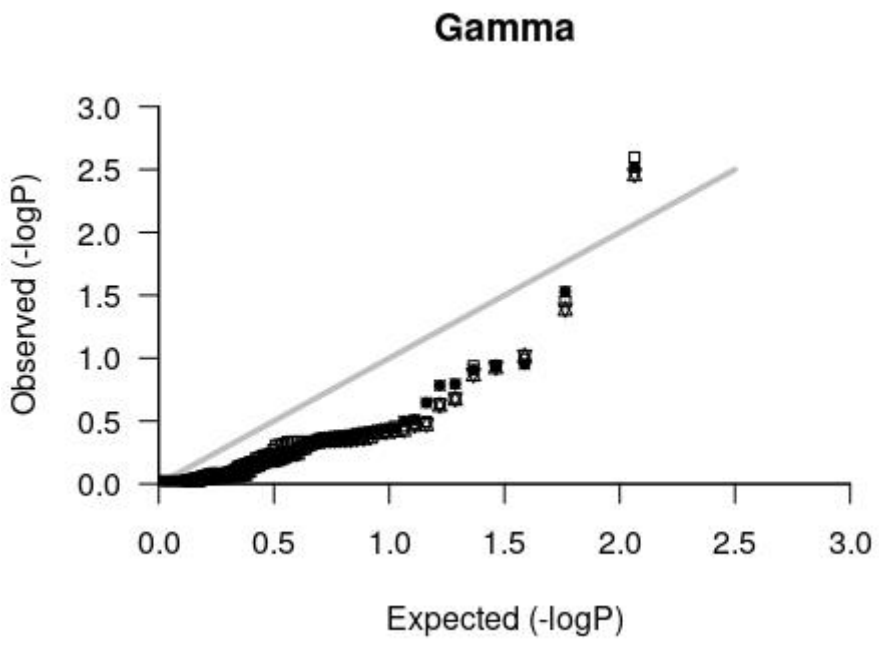

SUU
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

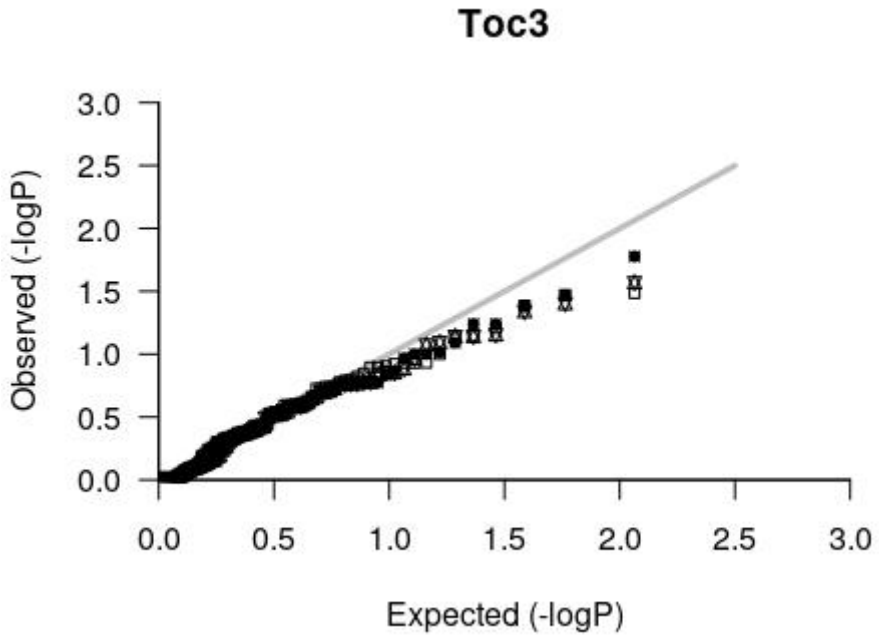

UUU
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

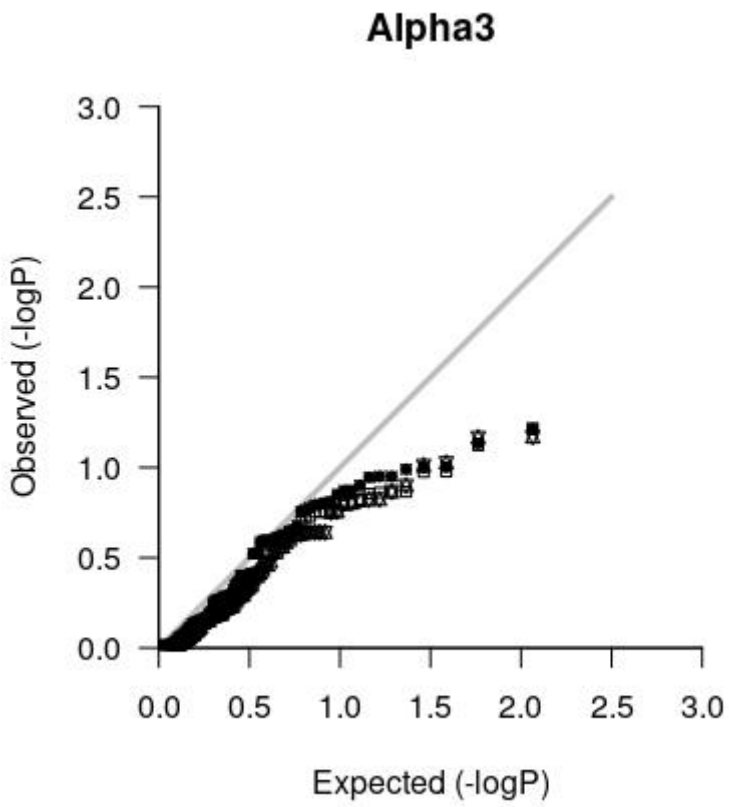

Tocph
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

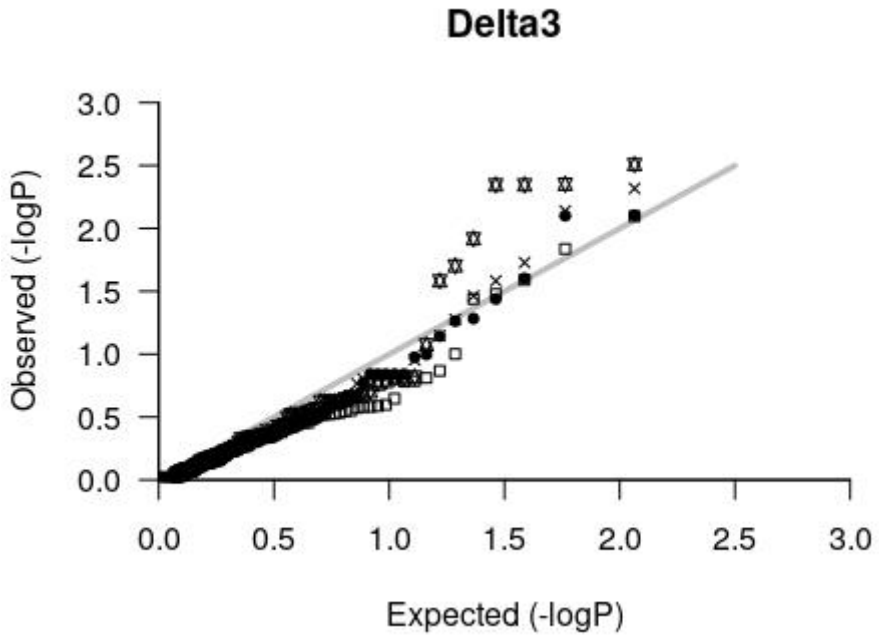

Alpha
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Delta
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Gamma
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Toc3
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Alpha3
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Delta3
Observed (-logP)
Expected (-logP)
0.0
0.5
1.0
1.5
2.0
2.5
3.0

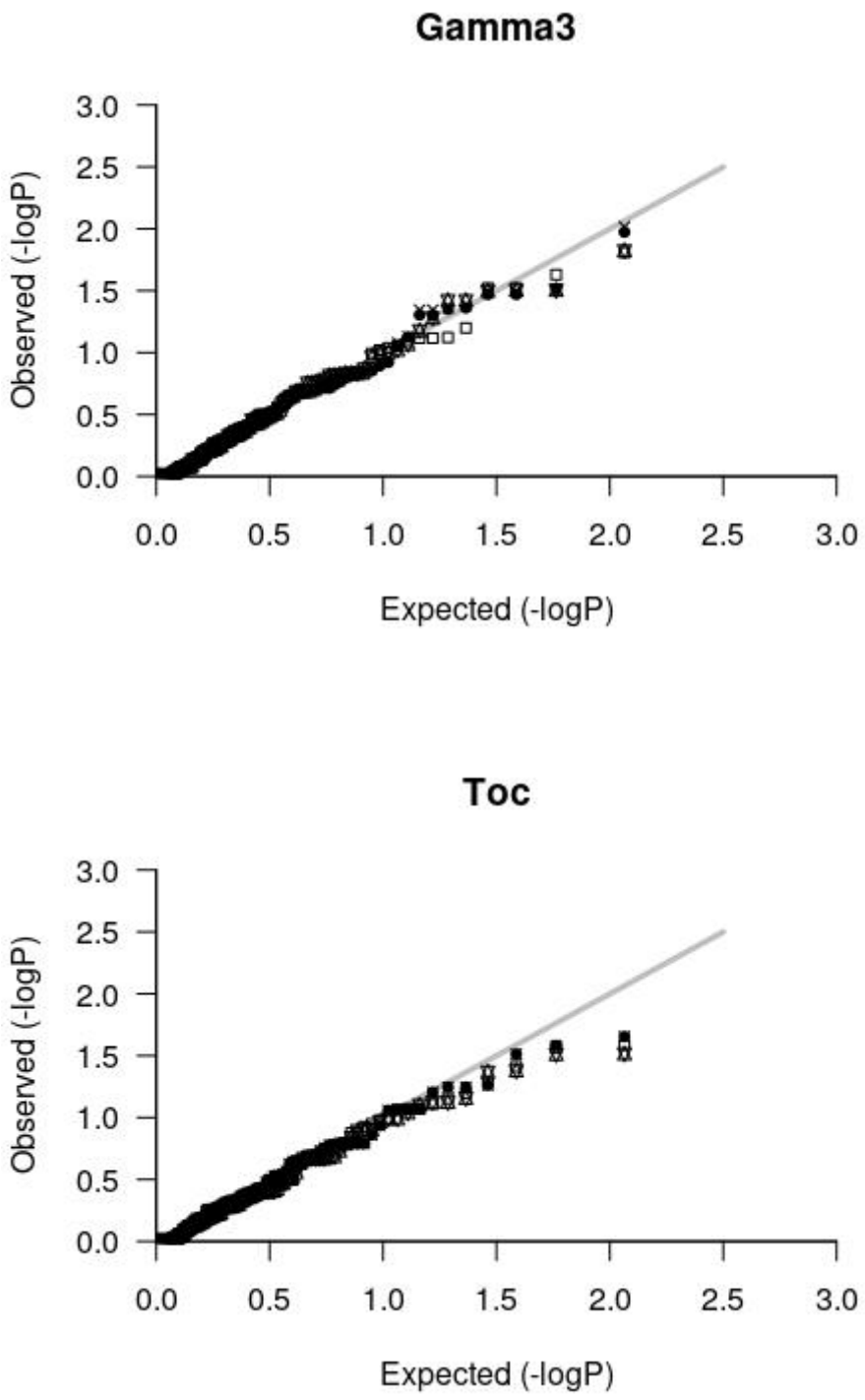


Figure A 1: Quantile-Quantile plots of the different studied traits for the tested models (black circles: MLM_PCA+K; white squares: MLM_Q+K; stars: GLM_Q; crosses: GLM_PCA).

List of publications

Published:

Astorkia M, Hernandez M, Bocs S, Ponce K, León O, Morales S, Quezada N, Orellana F, Wendra F, Sembiring Z, Asmono D, Ritter E (2020). Analysis Of The Allelic Variation In The Shell Gene Homolog Of E. oleifera And Design Of Species Specific Shell Primers. Euphytica (DOI: 10.1007/s10681-019-2538-7).

Published:

Astorkia M, Hernandez M, Bocs S, Lopez de Armentia E, Herran A, Ponce K, León O, Morales S, Quezada N, Orellana F, Wendra F, Sembiring Z, Asmono D, Ritter E (2019). Association Mapping between Candidate Gene SNP and Production and Oil Quality Traits in Interspecific Oil Palm Hybrids. Plants (DOI: 10.3390/plants8100377).

Published:

Astorkia M, Hernandez M, Bocs S, Ponce K, León O, Morales S, Quezada N, Orellana F, Wendra F, Sembiring Z, Asmono D, Ritter E (2020). Detection of significant SNP associated with production and oil quality traits in interspecific oil palm hybrids using RARSeq. Plant Science (DOI: 10.1016/j.plantsci.2019.110366).

REFERENCES

Alabady MS, Rogers WL, Malmberg RL (2015) Development of Transcriptomic Markers for Population Analysis Using Restriction Site Associated RNA Sequencing (RARseq). PLoS One 10:e0134855. doi: 10.1371/journal.pone.0134855

Alipour H, Bihamta MR, Mohammadi V, et al (2017) Genotyping-by-Sequencing (GBS) Revealed Molecular Genetic Diversity of Iranian Wheat Landraces and Cultivars. Front Plant Sci 8:1–14. doi: 10.3389/fpls.2017.01293

Altschul SF, Gish W, Miller W, et al (1990) Basic local alignment search tool. J Mol Biol 215:403–410. doi: 10.1016/S0022-2836(05)80360-2

Álvarez MF, Angarita M, Delgado MC, et al (2017) Identification of Novel Associations of Candidate Genes with Resistance to Late Blight in Solanum tuberosum Group Phureja. Front Plant Sci 8:1040. doi: 10.3389/fpls.2017.01040

Amblard P, Billotte N, Cochard B, et al (2004) El mejoramiento de la palma de aceite Elaeis guineensis y Elaeis oleifera por el Cirad-CP. Palmas 25:306–310

Ames-Sibin AP, Barizão CL, Castro-Ghizoni C V., et al (2018) β-Caryophyllene, the major constituent of copaiba oil, reduces systemic inflammation and oxidative stress in arthritic rats. J Cell Biochem 119:10262–10277. doi: 10.1002/jcb.27369

AOCS (2017a) AOCS Official Method Ce 1h-05. In: Official Methods and Recommended Practices of the AOCS

AOCS (2017b) AOCS Official Method Da 15-48. In: Official Methods and Recommended Practices of the AOCS

AOCS (2017c) AOCS Official Method Ce 5c-93. In: Official Methods and Recommended Practices of the AOCS

AOCS (2017d) AOCS Official Method Ce 8-89. In: Official Methods and Recommended Practices of the AOCS

Aprile MC, Caputo V, Nayga Jr RM (2012) Consumers' valuation of food quality labels: the case of the European geographic indication and organic farming labels. Int J Consum Stud 36:158–165. doi: 10.1111/j.1470-6431.2011.01092.x

Arias D, González M, Prada F, et al (2015) Genetic and phenotypic diversity of natural American oil palm (Elaeis oleifera (H.B.K.) Cortés) accessions. Tree Genet Genomes 11:122. doi: 10.1007/s11295-015-0946-y

Arias D, Ochoa I, Castro F, Romero H (2014) Molecular characterization of oil palm Elaeis guineensis Jacq. of different origins for their utilization in breeding programmes. Plant Genet Resour Charact Util 341–348. doi: 10.1017/S1479262114000148

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol 22:3179–3190. doi: 10.1111/mec.12276

Augusto A, Garcia F (2001) Genetic Architecture of Quantitative Traits. Architecture 35:303–39

Avila-Diazgranados RA, Daza ES, Navia E, Romero HM (2016) Response of various oil palm materials (Elaeis guineensis and Elaeis oleifera × Elaeis guineensis interspecific hybrids) to bud rot disease in the southwestern oil palm-growing area of Colombia. Agron Colomb 34:74–81. doi: 10.15446/agron.colomb.v34n1.53760

Babu BK, Mathur RK, Kumar PN, et al (2017) Development, identification and validation of CAPS marker for SHELL trait which governs dura, pisifera and tenera fruit forms in oil palm (Elaeis guineensis Jacq.). PLoS One 12:e0171933. doi: 10.1371/journal.pone.0171933

Babu BK, Mathur RK, Ravichandran G, Venu MVB (2019) Genome-wide association study (GWAS) for stem height increment in oil palm (Elaeis guineensis) germplasm using SNP markers. Tree Genet Genomes 15:40. doi: 10.1007/s11295-019-1349-2

Babu MK (2008) Bunch Analysis of Oil Palm. Pedavegi, India

Bachem CWB, Oomen RJFJ, Visser RGF (1998) Transcript Imaging with cDNA-AFLP: A Step-by-Step Protocol. Plant Mol Biol Report 16:157–173

Bai B, Wang L, Lee M, et al (2017) Genome-wide identification of markers for selecting higher oil content in oil palm. BMC Plant Biol 17:93. doi: 10.1186/s12870-017-1045-z

Bai B, Wang L, Zhang YJ, et al (2018) Developing genome-wide SNPs and constructing an ultrahigh-density linkage map in oil palm. Sci Rep 8:691. doi: 10.1038/s41598-017-18613-2

Barba J (2019) Oleiferas ecuatorianas alternativa de manejo Agronomico para compensar las perdidas ocasionadas por la pudricion del cogollo en America Latina. Orellana, Ecuador

Barba J (2016) Introgresión de genes E. guineensis en híbridos interespecíficos OxG para recuperar la fertilidad del polen y otras características deseables en palma de aceite. Palmas 37:285–293

Barcelos E, Amblard P, Berthaud J, Seguin M (2002) Genetic diversity and relationship in American and African oil palm as revealed by RFLP and AFLP molecular markers. Pesq agropec bras 37:1105–1114

Basiouni S, Fuhrmann H, Schumann J (2013) The influence of polyunsaturated fatty acids on the phospholipase D isoforms trafficking and activity in mast cells. Int J Mol Sci 14:9005–9017. doi: 10.3390/ijms14059005

Basiron Y, Weng CK (2004) The Oil Palm and its sustainability. J Oil Palm Res 16:1–10

Beirnaert A (1941) Contribution à l'étude genetique et biometrique des variétés d'Elaeis Guineensis Jacquin. Inst Natl pour l'étude Agron du Congo Série scie:1–101

Benelli G, Pavela R, Petrelli R, et al (2018) The essential oil from industrial hemp (Cannabis sativa L.) by-products as an effective tool for insect pest management in organic crops. Ind Crops Prod 122:308–315. doi: 10.1016/j.indcrop.2018.05.032

Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. B 57:289–300

Billotte N, Marseillac N, Risterucci A-M, et al (2005) Microsatellite-based high density linkage map in oil palm (Elaeis guineensis Jacq.). Theor Appl Genet 110:754–765. doi: 10.1007/s00122-004-1901-8

Bradbury PJ, Zhang Z, Kroon DE, et al (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinforma Appl 23:2633–2635. doi: 10.1093/bioinformatics/btm308

Branham SE, Wright SJ, Reba A, et al (2016) Genome-Wide Association Study in Arabidopsis thaliana of Natural Variation in Seed Oil Melting Point: A Widespread Adaptive Trait in Plants. J Hered 107:257–265. doi: 10.1093/jhered/esw008

Broad Institue (2015) Picard Tools - By Broad Institute.

https://broadinstitute.github.io/picard/index.html. Accessed 30 Aug 2018

Brouard J-S, Boyle B, Ibeagha-Awemu EM, Bissonnette N (2017) Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genet 18:32. doi: 10.1186/s12863-017-0501-y

Buckler ES, Thornsberry JM (2002) Plant molecular diversity and applications to genomics. Curr Opin Plant Biol 5:107–111. doi: 10.1016/S1369-5266(02)00238-8

Cadena T, Prada F, Perea A, Romero HM (2013) Lipase activity, mesocarp oil content, and iodine value in oil palm fruits of Elaeis guineensis, Elaeis oleifera, and the interspecific hybrid O×G (E. oleifera × E. guineensis). J Sci Food Agric 93:674–680. doi: 10.1002/jsfa.5940

Camillo J, Leão AP, Alves AA, et al (2014) Reassessment of the Genome Size in Elaeis guineensis and Elaeis oleifera , and Its Interspecific Hybrid. Genomics Insights 7:13–22. doi: 10.4137/GEI.S15522

Carlson KM, Heilmayr R, Gibbs HK, et al (2018) Effect of oil palm sustainability certification on deforestation and fire in Indonesia. PNAS 115:121–126. doi: 10.1073/pnas.1704728114

Carrasco B, González M, Gebauer M, et al (2018) Construction of a highly saturated linkage map in Japanese plum (Prunus salicina L.) using GBS for SNP marker calling. PLoS One 13:e0208032. doi: 10.1371/journal.pone.0208032

Castle WE (1903) MENDEL'S LAW OF HEREDITY. Science (80-) 18:396–406. doi: 10.1126/SCIENCE.18.456.396

Chafin TK (2016) fragmatic. https://github.com/tkchafin/fragmatic. Accessed 7 Jun 2019

Chan EKF, Rowe HC, Kliebenstein DJ (2010) Understanding the evolution of defense metabolites in Arabidopsis thaliana using genome-wide association mapping. Genetics 185:991–1007. doi: 10.1534/genetics.109.108522

Cheng P, Holdsworth W, Ma Y, et al (2015) Association mapping of agronomic and quality traits in USDA pea single-plant collection. Mol Breed 35:75. doi: 10.1007/s11032-015-0277-6

Cingolani P, Platts A, Wang LL, et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila

melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. doi: 10.4161/fly.19695

Corley, R.H.V. and Tinker PB, Corley RH V., Tinker PBH (2016) The Oil Palm, Fifth Edit. Blacwell Science Ltd., John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ , UK

Danecek P, Auton A, Abecasis G, et al (2011) The variant call format and VCFtools. Bioinforma Appl NOTE 27:2156–2158. doi: 10.1093/bioinformatics/btr330

Davey JW, Blaxter ML, Blaxter ML, Blaxter MW (2010) RADSeq: next-generation population genetics. Brief Funct Genomics 9:416. doi: 10.1093/BFGP/ELQ031

Davey JW, Hohenlohe PA, Etter PD, et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510. doi: 10.1038/nrg3012

de Carvalho LM, Borelli G, Camargo AP, et al (2019) Bioinformatics applied to biotechnology: A review towards bioenergy research. Biomass and Bioenergy 123:195–224. doi: 10.1016/j.biombioe.2019.02.016

de Wit N, Derrien M, Bosch-Vermeulen H, et al (2012) Saturated fat stimulates obesity and hepatic steatosis and affects gut microbiota composition by an enhanced overflow of dietary fat to the distal intestine. Am J Physiol Liver Physiol 303:G589–G599. doi: 10.1152/ajpgi.00488.2011

Din M (2000) Performance of Elaeis oleifera from Panama, Costa Rica, Colombia and Honduras in Malasya. J Oil Palm Res 12:71.80

Donini P, Cooke RJRJ, Reeves JCJC (2000) Molecular markers in variety and seed testing. Dev Plant Genet Breed 5:27–34. doi: 10.1016/S0168-7972(00)80005-5

Dunford NT (2012) Advancements in Oil and Oilseed Processing. In: Food and Industrial Bioproducts and Bioprocessing. Wiley-Blackwell, Oxford, UK, pp 115–143

Ehret GB (2010) Genome-wide association studies: contribution of genomics to understanding blood pressure and essential hypertension. Curr Hypertens Rep 12:17–25. doi: 10.1007/s11906-009-0086-6

Ewald R, Hoffmann C, Florian A, et al (2014) Lipoate-Protein Ligase and Octanoyltransferase Are Essential for Protein Lipoylation in Mitochondria of Arabidopsis 1[W][OPEN]. Plant

Physiol 165:978–990. doi: 10.1104/pp.114.238311

FAO (2019) Food and Agriculture Organization of the United Nations. http://www.fao.org/home/en. Accessed 20 Nov 2019

Fernandes ES, Passos GF, Medeiros R, et al (2007) Anti-inflammatory effects of compounds alpha-humulene and (−)-trans-caryophyllene isolated from the essential oil of Cordia verbenacea. Eur J Pharmacol 569:228–236. doi: 10.1016/j.ejphar.2007.04.059

Flutre T, Gay L, Rode N (2017) GitHub. https://github.com/timflutre/quantgen/blob/master/demultiplex.py. Accessed 22 Jul 2019

Food and Agriculture Organization of the United Nations (2019) FAOSTAT. In: Prod. Oil, palm top 10 Prod. http://www.fao.org/faostat/en/#data/QC/visualize. Accessed 9 Jan 2019

French GC, Ennos RA, Silverside AJ, Hollingsworth PM (2005) The relationship between flower size, inbreeding coefficient and inferred selfing rate in British Euphrasia species. Heredity (Edinb) 94:44–51. doi: 10.1038/sj.hdy.6800553

Friedline CJ, Lind BM, Hobson EM, et al (2015) The genetic architecture of local adaptation I: the genomic landscape of foxtail pine (Pinus balfouriana Grev. & Balf.) as revealed from a high-density linkage map. Tree Genet Genomes 11:49. doi: 10.1007/s11295-015-0866-x

Gamazon ER, Wheeler HE, Shah KP, et al (2015) A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 47:1091–1098. doi: 10.1038/ng.3367

Gao L, Turner MK, Chao S, et al (2016) Genome Wide Association Study of Seedling and Adult Plant Leaf Rust Resistance in Elite Spring Wheat Breeding Lines. PLoS One 11:e0148671. doi: 10.1371/journal.pone.0148671

García JA, Yáñez EE (2000) Aplicación de la metodología alterna para análisis de racimos y muestreo de racimos en tolva. PALMAS 21:303–311

Gariani K, Ryu D, Menzies KJ, et al (2017) Inhibiting poly ADP-ribosylation increases fatty acid oxidation and protects against fatty liver disease. J Hepatol 66:132–141. doi: 10.1016/j.jhep.2016.08.024

Geer LY, Marchler-Bauer A, Geer RC, et al (2009) The NCBI BioSystems database. Nucleic Acids Res 38:. doi: 10.1093/nar/gkp858

Ghelis T, Bolbach G, Clodic G, et al (2008) Protein Tyrosine Kinases and Protein Tyrosine Phosphatases Are Involved in Abscisic Acid-Dependent Processes in Arabidopsis Seeds and Suspension Cells. Plant Physiol 148:1668. doi: 10.1104/PP.108.124594

GitHub (2019) The world's leading software development platform GitHub

Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351. doi: 10.1038/nrg.2016.49

Goudet J, Jombart T (2015) Package "hierfstat": Estimation and Tests of Hierarchical F-Statistics

Grover A, Sharma PC (2016) Development and use of molecular markers: past and present. Crit Rev Biotechnol 36:290–302. doi: 10.3109/07388551.2014.959891

Gunstone FD (2011) Vegetable oils in food technology : composition, properties and uses. Wiley-Blackwell

Guo L, Xia J, Yang S, et al (2015) GHRH, PRP-PACAP and GHRHR Target Sequencing via an Ion Torrent Personal Genome Machine Reveals an Association with Growth in Orange-Spotted Grouper (Epinephelus coioides). Int J Mol Sci 16:26137–26150. doi: 10.3390/ijms161125940

Hardon JJ (1969) Interspecific hybrids in the genus Elaeis II. vegetative growth and yield of F1 hybrids E. guineensis x E. oleifera. Euphytica 18:380–388. doi: 10.1007/BF00397785

He J, Zhao X, Laroche A, et al (2014) Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci 5:484. doi: 10.3389/fpls.2014.00484

Henson IE (2012) A Brief History of the Oil Palm. AOCS Press

Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6:95–108. doi: 10.1038/nrg1521

Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70. doi: 10.2307/4615733

Hu X-G, Wu B-H, Liu D-C, et al (2013) Variation and their relationship of NAM-G1 gene and grain protein content in Triticum timopheevii Zhuk. J Plant Physiol 170:330–337. doi: 10.1016/j.jplph.2012.10.009

Ismail A, Mamat M (2002) The optimal age of oil palm replanting. Oil Palm Ind Econ J (Malasyan Palm Oil Board) 11–18

J. B, Cloutier S (2012) Association Mapping in Plant Genomes. In: Genetic Diversity in Plants. InTech

Jenkins GM, Frohman MA (2005) Phospholipase D: A lipid centric review. Cell. Mol. Life Sci. 62:2305–2316

Jeong KJ, Kim GW, Chung SH (2014) Amp-activated protein kinase: An emerging target for ginseng. J. Ginseng Res. 38:83–88

Jo B-S, Choi SS (2015) Introns: The Functional Benefits of Introns in Genomes. Genomics Inform 13:112–8. doi: 10.5808/GI.2015.13.4.112

Johnson MG, Shaw AJ (2015) Genetic diversity, sexual condition, and microhabitat preference determine mating patterns in Sphagnum (Sphagnaceae) peat-mosses. J Linn Soc 115:96–113. doi: https://doi.org/10.1111/bij.12497

Jombart T (2008) adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24:1403–1405. doi: 10.1093/bioinformatics/btn129

Jung J, Fan R, Jin L (2005) Combined linkage and association mapping of quantitative trait loci by multiple markers. Genetics 170:881–98. doi: 10.1534/genetics.104.035147

Kizer L, Pitera DJ, Pfleger BF, Keasling JD (2008) Application of functional genomics to pathway optimization for increased isoprenoid production. Appl Environ Microbiol 74:3229–3241. doi: 10.1128/AEM.02750-07

Kochikuzhyil BM, Devi K, Fattepur SR (2010) Effect of saturated fatty acid-rich dietary vegetable oils on lipid profile, antioxidant enzymes and glucose tolerance in diabetic rats. Indian J Pharmacol 42:142–5. doi: 10.4103/0253-7613.66835

Kou X, Liu C, Han L, et al (2016) NAC transcription factors play an important role in ethylene biosynthesis, reception and signaling of tomato fruit ripening. Mol Genet Genomics 291:1205–1217. doi: 10.1007/s00438-016-1177-0

Kozaki A, Kamada K, Nagano Y, et al (2000) Recombinant carboxyltransferase responsive to redox of pea plastidic acetyl-CoA carboxylase. J Biol Chem 275:10702–10708. doi: 10.1074/jbc.275.14.10702

Krück NC, Innes DI, Ovenden JR (2013) New SNPs for population genetic analysis reveal possible cryptic speciation of eastern Australian sea mullet (*Mugil cephalus*). Mol Ecol Resour 13:715–725. doi: 10.1111/1755-0998.12112

Kwong Q Bin, Teh CK, Ong AL, et al (2016) Development and Validation of a High-Density SNP Genotyping Array for African Oil Palm. Mol Plant 9:1132–1141. doi: 10.1016/j.molp.2016.04.010

Lal R (1996) Deforestation and land-use effects on soil degradation and rehabilitation in western Nigeria. II. Soil chemical properties. L Degrad Dev 7:87–98. doi: 10.1002/(SICI)1099-145X(199606)7:2<87::AID-LDR219>3.3.CO;2-O

LaO M, Arencibia AD, Carmona ER, et al (2008) Differential expression analysis by cDNA-AFLP of Saccharum spp. after inoculation with the host pathogen Sporisorium scitamineum. Plant Cell Rep 27:1103–1111. doi: 10.1007/s00299-008-0524-y

Lei X, Xiao Y, Xia W, et al (2014) RNA-seq analysis of oil palm under cold stress reveals a different C-repeat binding factor (CBF) mediated gene expression pattern in Elaeis guineensis compared to other species. PLoS One 9:e114482. doi: 10.1371/journal.pone.0114482

Lemaire SD, Guillont B, Le Maréchal P, et al (2004) New thioredoxin targets in the unicellular photosynthetic eukaryote Chlamydomonas reinhardtii. Proc Natl Acad Sci U S A 101:7475–7480. doi: 10.1073/pnas.0402221101

León-Del-Río A, Valadez-Graham V, Gravel RA (2017) Holocarboxylase Synthetase: A Moonlighting Transcriptional Coregulator of Gene Expression and a Cytosolic Regulator of Biotin Utilization. Annu Rev Nutr 37:207–223. doi: 10.1146/annurev-nutr-042617-104653

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

Li H, Handsaker B, Wysoker A, et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–9. doi: 10.1093/bioinformatics/btp352

Li R, Sun R, Hicks GR, Raikhel N V. (2015) Arabidopsis ribosomal proteins control vacuole

trafficking and developmental programs through the regulation of lipid metabolism. Proc Natl Acad Sci U S A 112:E89–E98. doi: 10.1073/pnas.1422656112

Li Y, Wilcox P, Telfer E, et al (2016) Association of single nucleotide polymorphisms with form traits in three New Zealand populations of radiata pine in the presence of genotype by environment interactions. Tree Genet Genomes 12:63. doi: 10.1007/s11295-016-1019-6

Liang C, Wang Y, Zhu Y, et al (2014) OsNAP connects abscisic acid and leaf senescence by fine-tuning abscisic acid biosynthesis and directly targeting senescence-associated genes in rice. Proc Natl Acad Sci 111:10013–10018. doi: 10.1073/pnas.1321568111

Lin Y, Liu S, Liu Y, et al (2017) Genome-wide association study of pre-harvest sprouting resistance in Chinese wheat founder parents. Genet Mol Biol 40:620–629. doi: 10.1590/1678-4685-gmb-2016-0207

Liu P, Lu Y, Liu H, et al (2012) Cancer Genes and Genomics Genome-Wide Association and Fine Mapping of Genetic Loci Predisposing to Colon Carcinogenesis in Mice. Mol Cancer Res 10:66–74. doi: 10.1158/1541-7786.MCR-10-0540

Liu Y, He Y, Jin A, et al (2014) Ribosomal protein-Mdm2-p53 pathway coordinates nutrient stress with lipid metabolism by regulating MCD and promoting fatty acid oxidation. Proc Natl Acad Sci U S A 111:. doi: 10.1073/pnas.1315605111

López de Armentia E (2017) Mapeo por asociación mediante genes candidatos en Palmera de Aceite Africana (E. guineensis Jacq.). UPV/EHU

Lu W, Tang X, Huo Y, et al (2012) Identification and characterization of fructose 1,6-bisphosphate aldolase genes in Arabidopsis reveal a gene family with diverse responses to abiotic stresses. Gene 503:65–74. doi: 10.1016/j.gene.2012.04.042

Madsen A, Bozickovic O, Bjune JI, et al (2015) Metformin inhibits hepatocellular glucose, lipid and cholesterol biosynthetic pathways by transcriptionally suppressing steroid receptor coactivator 2 (SRC-2). Sci Rep 5:. doi: 10.1038/srep16430

Mancini A, Imperlini E, Nigro E, et al (2015) Biological and nutritional properties of palm oil and palmitic acid: Effects on health. Molecules 20:17339–17361. doi: 10.3390/molecules200917339

Marangoni F, Galli C, Ghiselli A, et al (2017) Palm oil and human health. Meeting report of NFI: Nutrition Foundation of Italy symposium. Int J Food Sci Nutr 68:643–655. doi:

10.1080/09637486.2016.1278431

Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human population structure on large genetic association studies. Nat Genet 36:512–517. doi: 10.1038/ng1337

Marrano A, Birolo G, Prazzoli ML, et al (2017) SNP-Discovery by RAD-Sequencing in a Germplasm Collection of Wild and Cultivated Grapevines (V. vinifera L.). PLoS One 12:e0170655. doi: 10.1371/journal.pone.0170655

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10. doi: 10.14806/ej.17.1.200

Matsumura H, Miyagi N, Taniai N, et al (2014) Mapping of the gynoecy in bitter gourd (Momordica charantia) using RAD-seq analysis. PLoS One 9:e87138. doi: 10.1371/journal.pone.0087138

McKenna A, Hanna M, Banks E, et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1–7. doi: 10.1101/gr.107524.110

Meunier J, Boutin D (1975) L' Elaeis melanococca et l'hybride Elaeis melanococca x Elaeis guineensis. Oléagineux 30:5–8

Money D, Gardner K, Migicovsky Z, et al (2015) LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. G3 (Bethesda) 5:2383–90. doi: 10.1534/g3.115.021667

Money D, Migicovsky Z, Gardner K, Myles S (2017) LinkImputeR: user-guided genotype calling and imputation for non-model organisms. BMC Genomics 18:523. doi: 10.1186/s12864-017-3873-5

Montoya C, Cochard B, Flori A, et al (2014) Genetic architecture of palm oil fatty acid composition in cultivated oil palm (Elaeis guineensis Jacq.) compared to its wild relative E. oleifera (H.B.K) Cortés. PLoS One 9:e95412. doi: 10.1371/journal.pone.0095412

Montoya C, Lopes R, Flori A, et al (2013) Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross from Elaeis oleifera (H.B.K.) Cortés and oil palm (Elaeis guineensis Jacq.). Tree Genet Genomes 9:1207–1225. doi: 10.1007/s11295-013-0629-5

Montúfar R, Louise C, Trabarger T (2018) Elaeis oleifera (Kunth) Cortés: A neglected palm from the Ecuadorian Amazon. Rev Ecuat Med Cienc Biol 39:11–18. doi: 10.26807/remcb.v39i1.584

Morcillo F, Cros D, Billotte N, et al (2013) Improving palm oil quality through identification and mapping of the lipase gene causing oil deterioration. Nat Commun 4:2160. doi: 10.1038/ncomms3160

Moreno-Chacón AL, Camperos-Reyes JE, Ávila Diazgranados RA, Romero HM (2013) Biochemical and physiological responses of oil palm to bud rot caused by Phytophthora palmivora. Plant Physiol Biochem 70:246–251. doi: 10.1016/j.plaphy.2013.05.026

Mosquera T, Alvarez MF, Jiménez-Gómez JM, et al (2016) Targeted and untargeted approaches unravel novel candidate genes and diagnostic SNPs for quantitative resistance of the potato (Solanum tuberosum L.) to Phytophthora infestans causing the late blight disease. PLoS One 11:e0156254. doi: 10.1371/journal.pone.0156254

Mozzon M, Pacetti D, Lucci P, et al (2013) Crude palm oil from interspecific hybrid Elaeis oleifera x Elaeis guineensis: Fatty acid regiodistribution and molecular species of glycerides. Food Chem 141:245–252. doi: 10.1016/j.foodchem.2013.03.016

MPOB (2010) The Official Portal of Malaysian Palm Oil Board. http://www.mpob.gov.my/. Accessed 6 Sep 2019

Nadeem MA, Nawaz MA, Shahid MQ, et al (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32:261–285. doi: 10.1080/13102818.2017.1400401

Narum SR (2006) Beyond Bonferroni: Less conservative analyses for conservation genetics. Conserv Genet 7:783–787. doi: 10.1007/s10592-005-9056-y

Nazzicari N, Biscarini F, Cozzi P, et al (2016) Marker imputation efficiency for genotyping-by-sequencing data in rice (Oryza sativa) and alfalfa (Medicago sativa). Mol Breed 36:1–16. doi: 10.1007/s11032-016-0490-y

Nei M (1987) Molecular evolutionary genetics. Columbia University Press

Nesaretnam K, Guthrie N, Chambers AF, Carroll KK (1995) Effect of Tocotrienols on the Growth of a Human Breast Cancer Cell Line in Culture 1. Lipids 30:1139–1143

Nguyen CC, Nakaminami K, Matsui A, et al (2016) Oligouridylate Binding Protein 1b Plays an Integral Role in Plant Heat Stress Tolerance. Front Plant Sci 7:. doi: 10.3389/fpls.2016.00853

Nguyen CC, Nakaminami K, Matsui A, et al (2017) Overexpression of oligouridylate binding protein 1b results in ABA hypersensitivity. Plant Signal Behav 12:e1282591. doi: 10.1080/15592324.2017.1282591

Nigro D, Gadaleta A, Mangini G, et al (2019) Candidate genes and genome-wide association study of grain protein content and protein deviation in durum wheat. Planta 249:1157–1175. doi: 10.1007/s00425-018-03075-1

Noble WS (2009) How does multiple testing correction work? Nat Biotechnol 27:1135–1137. doi: 10.1038/nbt1209-1135

Nuruzzaman M, Sharoni AM, Kikuchi S (2013) Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. Front Microbiol 4:248. doi: 10.3389/fmicb.2013.00248

Ooi LC-L, Low E-TL, Abdullah MO, et al (2016) Non-tenera Contamination and the Economic Impact of SHELL Genetic Testing in the Malaysian Independent Oil Palm Industry. Front Plant Sci 7:1–13. doi: 10.3389/fpls.2016.00771

Paradis E, Jombart T, Brian K, et al (2018) Package “pegas” Title Population and Evolutionary Genetics Analysis System. 1–27

Pasam RK, Sharma R, Malosetti M, et al (2012) Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biol 12:16. doi: 10.1186/1471-2229-12-16

Paterson AH, Lander ES, Hewitt JD, et al (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721–726. doi: 10.1038/335721a0

Pelaez E, Ramirez D, Gerardo C (2010) Fisiología comparada de palmas africana (Elaeis guineensis Jacq .), americana (Elaeis oleifera hbk Cortes) e híbridos (E . oleifera x E . guineensis) en Hacienda La Cabaña. Palmas 31:29–38

Perrier, X., Jacquemoud-Collet J. (2006) DARwin software

Preciado CA, Bastidas S, Betancourth C, et al (2011) Predicción y control de la cosecha en el híbrido interespecífico Elaeis oleifera x Elaeis guineensis en la zona palmera occidental de Colombia. I. Determinación del periodo de madurez para obtener racimos con alto contenido de aceite. Corpoica Cienc Tecnol Agropecu 12:5–12

Pritchard JK, Rosenberg NA (1999) Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. Am J Hum Genet 65:220–228. doi: 10.1086/302449

Puyaubert J, Denis L, Alban C (2008) Dual Targeting of Arabidopsis HOLOCARBOXYLASE SYNTHETASE1: A Small Upstream Open Reading Frame Regulates Translation Initiation and Protein Targeting. Plant Physiol 146:478–491. doi: 10.1104/pp.107.111534

Pyne R, Honig J, Vaiciunas J, et al (2017) A first linkage map and downy mildew resistance QTL discovery for sweet basil (Ocimum basilicum) facilitated by double digestion restriction site associated DNA sequencing (ddRADseq). PLoS One 12:e0184319. doi: 10.1371/journal.pone.0184319

Python Software Foundation (2001) Python Software Foundation | Python Software Foundation. https://www.python.org/psf/. Accessed 23 Aug 2019

Qian F, Korat AA, Malik V, Hu FB (2016) Metabolic Effects of Monounsaturated Fatty Acid-Enriched Diets Compared With Carbohydrate or Polyunsaturated Fatty Acid-Enriched Diets in Patients With Type 2 Diabetes: A Systematic Review and Meta-analysis of Randomized Controlled Trials. Diabetes Care 39:1448–1457. doi: 10.2337/dc16-0513

Quail M, Smith ME, Coupland P, et al (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13:341. doi: 10.1186/1471-2164-13-341

Rafalski JA (2010) Association genetics in crop improvement. Curr Opin Plant Biol 13:174–180. doi: 10.1016/j.pbi.2009.12.004

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573–89. doi: 10.1534/genetics.114.164350

Ramachandiran I, Vijayakumar A, Ramya V, Rajasekharan R (2018) Arabidopsis serine/threonine/tyrosine protein kinase phosphorylates oil body proteins that regulate oil content in the seeds. Sci Rep 8:1154. doi: 10.1038/s41598-018-19311-3

Raquel Meléndez M, Ponce WP (2016) Pollination in the oil palms Elaeis guineensis, E. oleifera

and their hybrids (OxG), in tropical America. Agropec Trop 46:102–110. doi: 10.1590/1983-40632016v4638196

Reiman M, Laan M, Rull K, Sõber S (2019) Effects of RNA integrity on transcript quantification by total RNA sequencing of clinically collected human placental samples. FASEB J 3298–3308. doi: 10.1096/fj.201601031RR

Reyes PA, Ochoa JC, Montoya C, et al (2015) Development and validation of a bi-directional allele-specific PCR tool for differentiation in nurseries of dura, tenera and pisifera oil palms. Agron Colomb 33:5–10. doi: 10.15446/agron.colomb.v33n1.47988

Riedelsheimer C, Lisec J, Czedik-Eysenberg A, et al (2012) Genome-wide association mapping of leaf metabolic profiles for dissecting complex traits in maize. PNAS 109:8872–8877. doi: 10.1073/pnas.1120813109

Risch N, Merikangas K (1996) The future of genetic studies of complex diseases. Science (80-) 273:1516–1517

Ritter E, Lopez de Armentia E, Erika P, et al (2016) Development of a molecular marker system to distinguish shell thickness in oil palm genotypes. Euphytica 207:367–376. doi: 10.1007/s10681-015-1553-6

Robinson JT, Thorvaldsdóttir H, Winckler W, et al (2011) Integrative Genomics Viewer. Nat Biotechnol 29:24–26. doi: 10.1038/nbt.1754

Romero IG, Pai AA, Tung J, Gilad Y (2014) RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol 12:1–13. doi: 10.1186/1741-7007-12-42

RSPO (2019) RSPO - Roundtable on Sustainable Palm Oil

RSPO (2018) Principles and Criteria for the Production of Sustainable Palm Oil 2018. 1–208

Saadatian-Elahi M, Norat T, Goudable J, Riboli E (2004) Biomarkers of dietary fatty acid intake and the risk of breast cancer: A meta-analysis. Int J Cancer 111:584–591. doi: 10.1002/ijc.20284

Salas-Muñoz S, Rodríguez-Hernández AA, Ortega-Amaro MA, et al (2016) Arabidopsis AtDjA3 Null Mutant Shows Increased Sensitivity to Abscisic Acid, Salt, and Osmotic Stress in Germination and Post-germination Stages. Front Plant Sci 7:220. doi: 10.3389/fpls.2016.00220

Salavarrieta R, Jesús P (2004) Cenipalma Oil Palm Elaeis guineensis (Jacq.) and Elaeis oleifera (H.B.K.) Genetic Collection: Characteristics of Importance for the Oil Palm Sector. Palmas 25 No. Esp:39–48

Sant'Ana GC, Pereira LFP, Pot D, et al (2018) Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. Sci Rep 8:1–12. doi: 10.1038/s41598-017-18800-1

SAS Institute Inc. Analytics, Business Intelligence and Data Management | SAS. https://www.sas.com/en_us/home.html. Accessed 17 Oct 2019

Schoen H, Thimm O, Ritte G, et al (2010) Plants with increased yield (NUE)

Schrier-Uijl AP, Silvius M, Parish F, et al (2013) Environmental and Social Impacts of Oil Palm Cultivation on Tropical Peat-A Scientific Review Kalimantan Wetland and Climate Change Studies (KWACS) View project

Seminar on Worldwide performance of DXP oil palm planting materials clones and interspecific hybrids. P junio 5-6 BC, Amblard P, Kouame B, et al (1999) Comparative performance of interespecific hybrids and commercial E. guineensis material

Seng T-Y, Mohamed Saad SH, Chin C-W, et al (2011) Genetic Linkage Map of a High Yielding FELDA Deli×Yangambi Oil Palm Cross. PLoS One 6:e26593. doi: 10.1371/journal.pone.0026593

Seto KC, Reenberg A (2014) Rethinking global land use in an urban era. MIT Press

Shintani D, DellaPenna D (1998) Elevating the vitamin E content of plants through metabolic engineering. Science 282:2098–100. doi: 10.1126/SCIENCE.282.5396.2098

Siew WL, Tang TS (1995) PORIM p2.6 Method. In: PORIM Test Methods. Malaysian Oil Palm Board (MPOB), Kuala Lumpur, Malaysia, p 181

Singh D, Singh B, Mishra S, et al (2019) Candidate gene based association analysis of salt tolerance in traditional and improved varieties of rice (Oryza sativa L.). J Plant Biochem Biotechnol 28:76–83. doi: 10.1007/s13562-018-0464-8

Singh R, Low E-TL, Ooi LC-L, et al (2013a) The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. Nature 500:340–344. doi: 10.1038/nature12356

Singh R, Low ETL, Ooi LCL, et al (2014) The oil palm VIRESCENS gene controls fruit colour and

encodes a R2R3-MYB. Nat Commun 5:1–8. doi: 10.1038/ncomms5106

Singh R, Ong-Abdullah M, Low E-TL, et al (2013b) Oil palm genome sequence reveals divergence of interfertile species in Old and New Worlds. Nature 500:335–341. doi: 10.1038/nature12309

Singh R, Tan SG, Panandam JM, et al (2009) Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. BMC Plant Biol 9:1–19. doi: 10.1186/1471-2229-9-114

Šmarda P, Bureš P, Horová L, et al (2014) Ecological and evolutionary significance of genomic GC content diversity in monocots. PNAS 111:E4096–E4102. doi: 10.1073/pnas.1321152111

Soh AC, Mayes S, Roberts JA (2017) Oil palm breeding : genetics and genomics. CRC Press

Solmonson A, Deberardinis RJ (2017) Lipoic acid and mitochondrial redox regulation 1 Lipoic acid metabolism and mitochondrial redox regulation. JBC 293:7522–7530. doi: 10.1074/jbc.TM117.000259

Somyong S, Walayaporn K, Jomchai N, et al (2018) Transcriptome analysis of oil palm inflorescences revealed candidate genes for an auxin signaling pathway involved in parthenocarpy. PeerJ 6:e5975. doi: 10.7717/peerj.5975

Soriano A, Guitton P, Mournet P (2018) Workflow-snakemake-capture. https://github.com/SouthGreenPlatform/Workflow-snakemake-capture

South Green Collaborators (2016) The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics. Curr Plant Biol 7–8:6–9. doi: 10.1016/j.cpb.2016.12.002

Su C, Wang W, Gong S, et al (2017) High Density Linkage Map Construction and Mapping of Yield Trait QTLs in Maize (Zea mays) Using the Genotyping-by-Sequencing (GBS) Technology. Front Plant Sci 8:706. doi: 10.3389/fpls.2017.00706

Sumathi S, Chai SP, Mohamed AR (2008) Utilization of oil palm as a source of renewable energy in Malaysia. Renew Sustain Energy Rev 12:2404–2421. doi: 10.1016/j.rser.2007.06.006

Sundram K, Sambanthamurthi R, Tan Y-A (2003) Palm fruit chemistry and nutrition. Asia Pacific J Clin NutrDOI 12:355–62

Sundram S, Intan-Nur AMA (2017) South American Bud rot: A biosecurity threat to South East Asian oil palm. Crop Prot 101:58–67. doi: 10.1016/j.cropro.2017.07.010

Sved JA, Hill WG (2018) One hundred years of linkage disequilibrium. Genetics 209:629–636. doi: 10.1534/genetics.118.300642

Swamy BPM, Shamsudin NAA, Rahman SNA, et al (2017) Association Mapping of Yield and Yield-related Traits Under Reproductive Stage Drought Stress in Rice (Oryza sativa L.). Rice 10:21. doi: 10.1186/s12284-017-0161-6

Swarts K, Li H, Romero Navarro JA, et al (2014) Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. Plant Genome 7:0. doi: 10.3835/plantgenome2014.05.0023

Taeprayoon P, Tanya P, Lee S-H, Srinives P (2015) Genetic background of three commercial oil palm breeding populations in Thailand revealed by SSR markers. AJCS 9:281–288

Tanksley D, Medina-Filho H, Rick CM (1981) The effect of isozyme selection on metric characters in an interspecific backcross of tomato — basis of an early screening procedure. Theor Appl Genet 60:291–296. doi: 10.1007/BF00263721

Tanksley SD (1993) Mapping Polygenes. Annu Rev Genet 27:205–233. doi: 10.1146/annurev.ge.27.120193.001225

Tatematsu K, Kamiya Y, Nambara E (2008) Co-regulation of ribosomal protein genes as an indicator of growth status: Comparative transcriptome analysis on axillary shoots and seeds in Arabidopsis. Plant Signal Behav 3:450–452. doi: 10.4161/psb.3.7.5577

Teh C-K, Ong A-L, Kwong Q-B, et al (2016a) Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. Sci Rep 6:19075. doi: 10.1038/srep19075

Teh CK, Ong AL, Kwong Q Bin, et al (2016b) Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. Sci Rep 6:1–7. doi: 10.1038/srep19075

The R Development Core Team (2008) R: A Language and Environment for Statistical Computing. 2.6.2:1–2673

Thermo Fisher Scientific (2016) Multiple Primer Analyzer.

https://www.thermofisher.com/es/es/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html. Accessed 28 Aug 2019

Tierney AC, Roche HM (2007) The potential role of olive oil-derived MUFA in insulin sensitivity. Mol Nutr Food Res 51:1235–1248. doi: 10.1002/mnfr.200700143

Ting N-C, Jansen J, Mayes S, et al (2014) High density SNP and SSR-based genetic maps of two independent oil palm hybrids. BMC Genomics 15:1–11

Ting N-C, Yaakub Z, Kamaruddin K, et al (2016) Fine-mapping and cross-validation of QTLs linked to fatty acid composition in multiple independent interspecific crosses of oil palm. BMC Genomics 17:1–17. doi: 10.1186/s12864-016-2607-4

Torres GA, Sarria GA, Varon F, et al (2010) First Report of Bud Rot Caused by *Phytophthora palmivora* on African Oil Palm in Colombia. Plant Dis 94:1163–1163. doi: 10.1094/PDIS-94-9-1163A

Torres M, Rey L, Gelves F, Santacruz L (2004) Evaluación del comportamiento de los híbridos interespecíficos Elaeis oleifera x Elaeis guineensis , en la plantación de Guaicaramo S.A. Evaluation of the Behavior of Elaeis Oleifera x Elaeis Guineensis Hybrids in Guaicaramo Plantation. Palmas 25:350–357

Turner PD, Incorporated Society of Planters. (1981) Oil palm diseases and disorders. Published for the Inc. Society of Planters [by] Oxford University Press

Untergasser A, Cutcutache I, Koressaar T, et al (2012) Primer3-new capabilities and interfaces. Nucleic Acid Res. doi: 10.1093/nar/gks596

USDA (2019) Oilseeds: World Markets and Trades. 1–68

USDA (2018) Oilseeds: World Markets and Trade. 1–67

Verma S, Gupta S, Bandhiwal N, et al (2015) High-density linkage map construction and mapping of seed trait QTLs in chickpea (Cicer arietinum L.) using Genotyping-by-Sequencing (GBS). Sci Rep 5:1–14. doi: 10.1038/srep17512

Vida A, Márton J, Mikó E, Bai P (2017) Metabolic roles of poly(ADP-ribose) polymerases. Semin. Cell Dev. Biol. 63:135–143

Vijay V, Pimm SL, Jenkins CN, Smith SJ (2016) The Impacts of Oil Palm on Recent Deforestation

and Biodiversity Loss. PLoS One 11:e0159668. doi: 10.1371/journal.pone.0159668

Wang J, Zhang Z (2018) GAPIT Version 3: An Interactive Analytical Tool for Genomic Association and Prediction. Bioinformatics Draft:7

Wang M, Yan J, Zhao J, et al (2012) Genome-wide association study (GWAS) of resistance to head smut in maize. Plant Sci 196:125–131. doi: 10.1016/j.plantsci.2012.08.004

Wen Z, Boyse JF, Song Q, et al (2015) Genomic consequences of selection and genome-wide association mapping in soybean. BMC Genomics 16:1–14. doi: 10.1186/s12864-015-1872-y

Wen Z, Tan R, Yuan J, et al (2014) Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. BMC Genomics 15:1–11. doi: 10.1186/1471-2164-15-809

Xia W, Luo T, Zhang W, et al (2018) Identification of genes affecting saturated fat acid content in Elaeis guineensis by genome-wide association analysis. bioRxiv 1–32. doi: 10.1101/341347

Xiao J-P, Zhang L-L, Zhang H-Q, Miao L-X (2017) Identification of Genes Involved in the Responses of Tangor (C. reticulata × C. sinensis) to Drought Stress . Biomed Res Int 2017:1–15. doi: 10.1155/2017/8068725

Xu P, Xu S, Wu X, et al (2014) Population genomic analyses from low-coverage RAD-Seq data: a case study on the non-model cucurbit bottle gourd. Plant J 77:430–442. doi: 10.1111/tpj.12370

Yu F, Okamto S, Nakasone K, et al (2008) Molecular cloning and functional characterization of-humulene synthase, a possible key enzyme of zerumbone biosynthesis in shampoo ginger (Zingiber zerumbet Smith). Planta 227:1291–1299. doi: 10.1007/s00425-008-0700-x

Yu J, Buckler ES (2006) Genetic association mapping and genome organization of maize. Curr Opin Biotechnol 17:155–160. doi: 10.1016/j.copbio.2006.02.003

Yu Y, Zhang X, Yuan J, et al (2017) Identification of Sex-determining Loci in Pacific White Shrimp Litopeneaus vannamei Using Linkage and Association Analysis. Mar Biotechnol 19:277–286. doi: 10.1007/s10126-017-9749-5

Yuen May C, Nesaretnam K (2014) Research advancements in palm oil nutrition. Eur J Lipid Sci

Technol 116:1301–1315. doi: 10.1002/ejlt.201400076

Zaki NM, Ismail I, Rosli R (2010) Development and characterization of Elaeis oleifera micorsatellite markers. Sains Malaysiana 39:909–912

Zegeye H, Rasheed A, Makdis F, et al (2014) Genome-Wide Association Mapping for Seedling and Adult Plant Resistance to Stripe Rust in Synthetic Hexaploid Wheat. PLoS One 9:e105593. doi: 10.1371/journal.pone.0105593

Zeng Y, Tan X, Zhang L, et al (2014) Identification and Expression of Fructose-1,6-Bisphosphate Aldolase Genes and Their Relations to Oil Content in Developing Seeds of Tea Oil Tree (Camellia oleifera). PLoS One 9:e107422. doi: doi:10.1371/journal.pone.0107422

Zhang L, Chen K, Xu C (2008) Identification and Characterization of Transcripts Differentially Expressed in Peel and Juice Vesicles of Immature and Ripe Orange (Citrus sinensis) Fruit. Plant Mol Biol Report 26:121–132. doi: 10.1007/s11105-008-0030-y

Zhao J, Huang L, Ren X, et al (2017) Genetic Variation and Association Mapping of Seed-Related Traits in Cultivated Peanut (Arachis hypogaea L.) Using Single-Locus Simple Sequence Repeat Markers. Front Plant Sci 8:2105. doi: 10.3389/fpls.2017.02105

Zheng Y, Deng X, Qu A, et al (2018) Regulation of pollen lipid body biogenesis by MAP kinases and downstream WRKY transcription factors in Arabidopsis. PLOS Genet 14:e1007880. doi: 10.1371/journal.pgen.1007880

Zhu C, Gore M, Buckler ES, Yu J (2008) Status and Prospects of Association Mapping in Plants. Plant Genome J 1:5. doi: 10.3835/plantgenome2008.02.0089