eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Tesis Doctoral

Mapeo por asociación mediante genes candidatos en Palmera de Aceite Africana (*E. guineensis* Jacq.)

Emma López de Armentia Adan

Dirigida por: Dr. Enrique Ritter Azpitarte

Vitoria-Gasteiz, 2017

(c)2017 EMMA LOPEZ DE ARMENTIA ADAN

eman ta zabal zazu

Universidad del País Vasco

Euskal Herriko Unibertsitatea

Tesis Doctoral

Mapeo por asociación mediante genes candidatos en Palmera de Aceite Africana (*E. guineensis* Jacq.)

Emma López de Armentia Adán

Dirigida por: Dr. Enrique Ritter Azpitarte

neiker tecnalia

Sampoerna Agro

Vitoria-Gasteiz, 2017

A Alaia, Edu, y Mamá

Sin vosotros este trabajo no hubiera sido posible

"Yo soy de los que piensan que la ciencia tiene una gran belleza. Un científico en su laboratorio no es sólo un técnico: también es un niño colocado ante fenómenos naturales que lo impresionan como un cuento de hadas."

Maria Salomea Skłodowska-Curie (1867-1934)

AGRADECIMIENTOS

Tengo que dar las gracias a la Fundación Cándido Iturriaga y Maria Doñabeitia por la concesión de la beca doctoral que me ha permitido desarrollar este trabajo. A Neiker-Tecnalia , y en concreto al Departamento de Producción Vegetal (antes Biotecnología), dirigido por Dra. Sonia Castañón de la Torre, por acogerme y poner los medios necesarios para mi formación y aprendizaje como investigadora.

A mi director, el Dr. Enrique Ritter, que desde un primer momento confió en mí y me aceptó en su equipo. Además de ser él quién me ha transmitido muchos de los conocimientos adquiridos sobre la mejora genética vegetal a nivel molecular, tan desconocida para mí cuando llegué. Gracias por valorar mi trabajo y estar pendiente de mis dudas.

I would like to thank Sampoerna Agro TBK, particularly Dr. Dwi Asmono and Mr. Zulhermana Sembiring, for giving me the opportunity to take part in DAMASO project, and allowing to learn more about the oil palm breeding.

Al equipo del que he formado parte, a Ana Herrán por acogerme como una más desde el día que llegué al laboratorio, y por las horas compartidas en la búsqueda de genes candidatos delante del ordenador. Gracias a su organización hoy puedo presentar parte de este trabajo. A Mónica Hernández porque con su ayuda y colaboración pude llevar a cabo las librerías, y a analizar los datos de la secuenciación, y siempre estuvo dispuesta a echarme una mano.

I am grateful to Pratiwi Erika and Baitha Santika for teaching me new culture and for sharing moments at the lab and outside. I know I have two Indonesian's friends. Always helping me with my doubts by the distance. Both are special workmates.

Gracias a Maite, por preocuparse porque el material estuviera siempre apunto. Por su disponibilidad y porque siempre se ha preocupado por cómo estoy, y se acuerda de todos los que hemos estado en labo.

No me puedo olvidar tampoco de mis compañer@s. A Iratxe por su apoyo y ánimo a cada momento, por las charlas y risas que hemos compartido en el labo y en inglés. Y por esos deberes!! A Itzi, Judith y Zior por compartir conmigo vuestro conocimiento y estar siempre disponibles para resolverme una duda. A Olatz, mi compañera de tesis, por esos momentos de estrés que hemos vivido juntas, por la ayuda que me has dado, y por esas conversaciones trascendentales que hemos tenido tú y yo. A Néstor por esos cafés y charlas compartidas Sé que puedo contar contigo. Y a todos los demás que os habéis preocupado por mí.

A mi cuadri "neikeriana" porque desde el momento que os conocí formáis parte de mis amigos. A Isi por ser mi principal apoyo allí, por estar siempre pendiente de cómo estaba en lo personal y en lo laboral, por preocuparte de mí. Por escucharme y aconsejarme. Son tantas las cosas que tengo que agradecerte… A Javi porque aparte de ser un gran compañero, es un gran amigo. Siempre dispuesto a colaborar y a ayudar. A Marina y a Vanesa porque sois grandes personas, dispuestas a escuhar. Vuestra amistad es uno de los grandes regalos personales que me llevo.

A mis amigas de toda la vida porque a pesar de la desconexión siempre están ahí cuando más las necesito. A Nere, Camino y Vega, por esos momentos que me han ayudado a desahogarme durante este último año. Por animarme, y echarme una mano cuando habéis podido, por

preocuparos por el reto que esto suponía, y por entenderme. A todas os quiero un montón y sois parte de mi familia.

A mi madre porque gracias a todos los esfuerzos que ha hecho he podido llegar hasta aquí. Mamá me has enseñado a luchar por lo que uno quiere, y a levantarme después de caerme. ¡Gracias, Mamá! A mi padre que me enseñó la constancia y a superarme a mi misma, y que el esfuerzo tiene su recompensa. A mi hermano Jon porque siempre puedo contar con él, y a pesar de nuestras diferencias siempre acabamos entendiéndonos. A mis abuelos porque ellos quizás no entendían mi visión de la vida pero siempre tenían los brazos abiertos para mí, para escucharme y ayudarme a entender las cosas desde su punto de vista. A mis suegros y mis cuñados porque os habéis implicados en esta etapa y en el esfuerzo que requería. Por preocuparos de que estemos bien, y esos "tapers" de Angelines que nos han salvado, sobre todo, en estos últimos meses.

A Edu por estar siempre a mi lado, luchando y apoyándome. Compartir este camino ha sido más fácil a tu lado, todo es más fácil contigo,cariño. A Alaia porque es lo más especial de mi vida. Me aportas alegría y tu sonrisa hace que se curen todos mis males. Me recuerdas cada día como se ve la vida siendo niño. Esta tesis es tan vuestra como mía.

ÍNDICE

ÍNDICE

ABREVIATURAS

µl: microlitro

ADNc: Ácido Desoxirribonucleico Complementario

AFLP: Amplified Lenght Polymorphism

AGS: Ácido Graso Saturado.

ARNm: Ácido Ribonucléico Mensajero

ATP: Adenosin Trifosfato

BSA: Bulk Segregant Analysis

CDK: Ciclyn Dependent Kinase

cDNA- AFLP: Complementary DNA - Amplified Fragment Length Polymorphism

CIRAD: Agricultural Research for Development

Cm: Centimetros

cM: centiMorgan

DL: Desequilibrio de Ligamiento

EST: Expressed Sequence Tag

GC: Gen Candidato

GCA: General Combining Ability

GL: Grupo de Ligamiento

GLM: General Linear Model

GO: Gene Onthology

gr: Gramos

GWAS: Genome Wide Association

Ha:Hectárea

Kg: Kilogramo

LD: Linkage Disequilibrium

LTP: Lipid-Transfer Protein

MA: Mapeo por Asociación

MAS: Marker Assisted Selection

MLM: Modelo Linear Mixto o "Mixed Linear Model"

MPOB: Malasian Palm Oil Board

OP: Oil Palm

OPGP: Oil Palm Genome Proyect

PD: Phoenix dactylifera

QTL: Quantitative Trait Loci

RF: Frecuencia de Recombinación o Recombination Frecuency

RGA: Resistance Gene Analogs o Genes Análogos de Resitencia

RRMS: Reciprocal Recurrent Modified Selection

RRS: Reciprocal Recurrent Selection

SAM: Selección Asistida por Marcadores

SNP: Single Nucleotide Polymorphism

SSR: Single Sequence Repeat

TAG: Triacilglicérido

Ton: Tonelada

USDA: United StatesDepartment of Agriculture.

CAPÍTULO 1: INTRODUCCIÓN

1. IMPORTANCIA ECONÓMICA DE LOS CULTIVOS OLEAGINOSOS: PALMERA DE ACEITE

Uno de los principales sectores agrícolas mundiales es el de los cultivos oleaginosos, cuyo crecimiento aumenta anualmente. Sus semillas se utilizan principalmente para consumo humano, pero también en la industria oleo química. La soja, la palmera de aceite, la colza, el girasol y el cacahuete son los cultivos con una participación más activa en el mercado mundial (Figura 1).

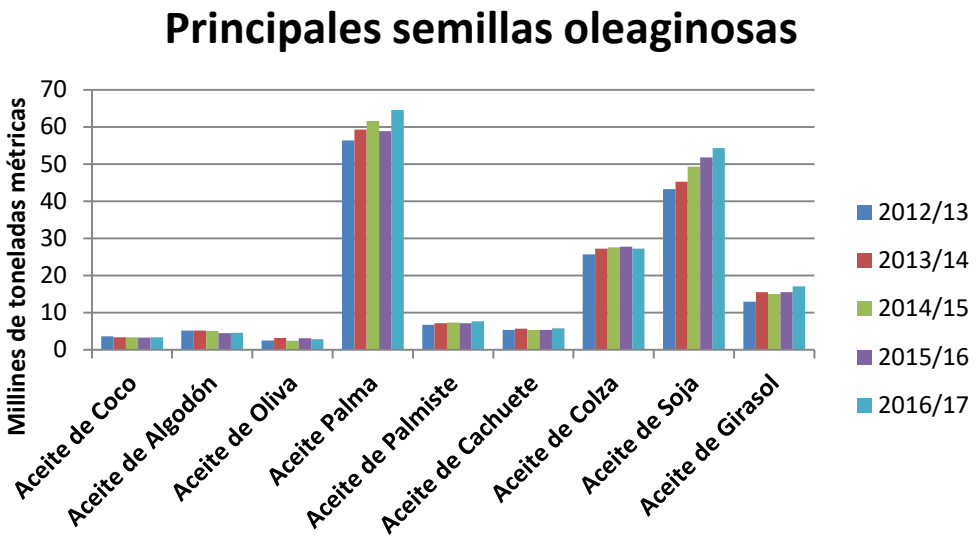


Figura 1: Producción de las diferentes semillas oleaginosas (millones de toneladas métricas ente el año 2011 y 2015)

Estos cultivos crecen en regiones climáticas muy diversas, siendo de vital importancia para los países productores y la economía mundial. El aumento de la producción se debe al crecimiento de la demanda de aceites, grasas y derivados. En consecuencia, la superficie de cultivo aumenta considerablemente, y por tanto la búsqueda de variedades más productivas. El uso de los avances científicos y de las nuevas tecnologías colabora en este aumento de la productividad, principalmente en países donde los niveles de producción agrícola son muy altos (Sharma y col., 2012)

Según los datos proporcionados por USDA (Departamento de Agricultura de Estados Unidos) la producción global del aceite de palma ha aumentado desde 15,2 millones de toneladas en el año 1995 a más de 60 millones en 2014, situando su volumen de producción a la cabeza de los aceites vegetales, superando a la soja, colza o girasol (Figura 3a). Su consumo también se ha triplicado en estos últimos 20 años, y convirtiéndose en el aceite más consumido del mundo (Figura 3b).

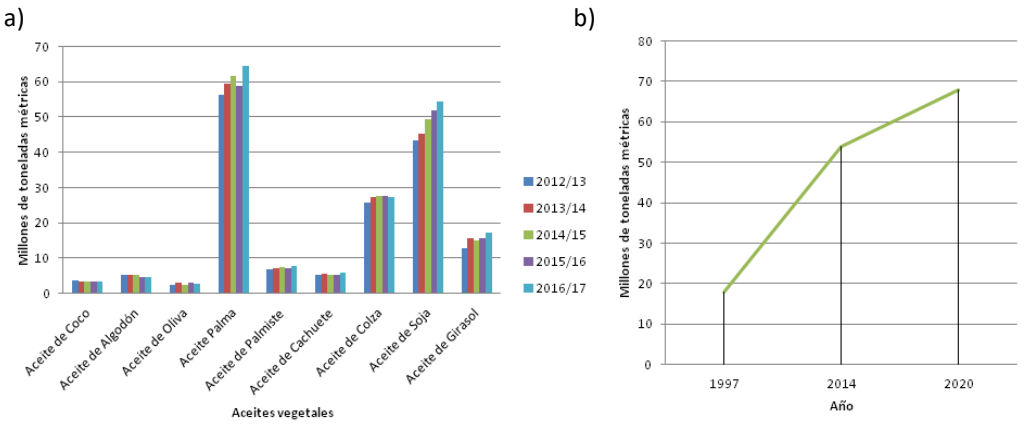


Figura 2: a) Producción de aceites vegetales en el periodo 2012-2017 (USDA, 2017). b) Demanda mundial de aceite de palma (European Palm Oil Alliance, 2016).

Los principales países productores de aceite de palma son Indonesia y Malasia, aunque en los últimos años se han sumado Tailanda, Colombia y Nigeria dónde se ha producido un aumento considerable en la producción. (Figura 4.a). Los principales consumidores de aceite de palma son India, Indonesia y Unión Europea. El 70% de la producción mundial se exporta a otros países, siendo India y la Unión Europea los principales importadores de aceite de palma. Además, la palmera de aceite es el cultivo oleaginoso más eficiente del mundo, con una alta tasa de productividad por superficie de tierra. Ocupa el 5% de la superficie de la tierra, pero produce el 32% del aceite total consumido (Figura 4b).

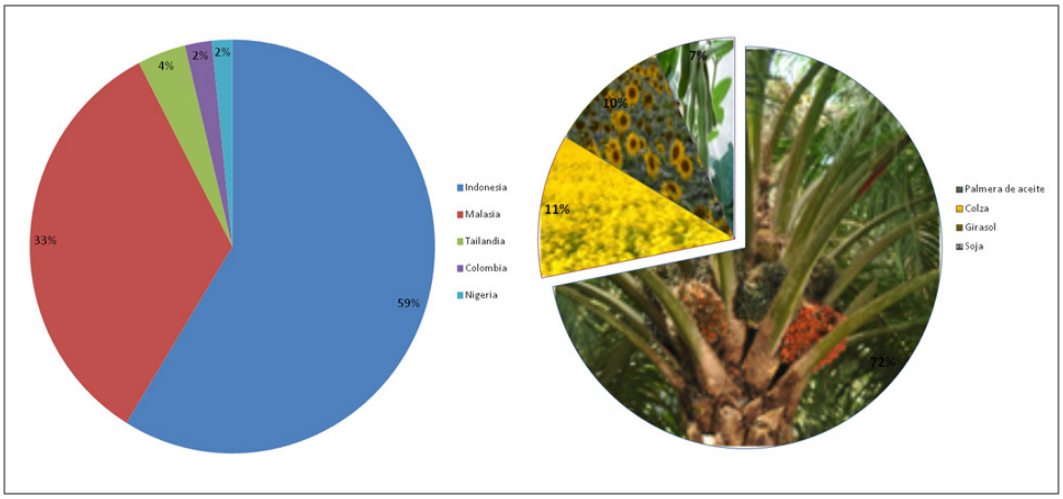


Figura 3: a) Principales países productores de aceite de palma. En el gráfico se muestra como Indonesia (59%) y Malasia (33%) siendo los principales productores mundiales. El resto de paises productores son Tailandia (4%), Colombia (2%) y Nigeria (2%). b). Rendimiento de los cultivos oleaginosos más importantes: palmera de aceite (72%), colza (11%), girasol (10%), y soja (7%) (http://apps.fas.usda.gov/psdonline/psdReport).

1.1. Usos del aceite de palma y sus derivados

El aceite obtenido del fruto se destina a su consumo en crudo como aceite de cocina doméstico en el sudeste asiático, África y algunas partes de Brasil. En Europa y Estados Unidos, el aceite se utiliza en su forma refinada y su grasa sólida como un ingrediente alimentario muy versátil para aportar sabor

y calidad al producto final. Se puede encontrar en numerosos alimentos como margarinas, chocolate, helados, productos de panadería y confitería, y snacks. También se desarrollan nuevos usos como la extracción de micronutrientes, carotenos, vitamina E o esteroles, para su utilización individual.

En la industria no alimentaria se utiliza el palmiste, es la grasa obtenida de la semilla, para la producción de jabones, detergentes, velas, resinas o cosméticos. En los últimos años hay un aumento del desarrollo de nuevos productos para la industria oleoquímica, donde los derivados de la propia palmera pueden utilizarse como fuente de biomasa y energía renovable para sustituir el combustible diesel.

2. LA PALMERA DE ACEITE: EL CULTIVO

2.1. Generalidades

La palmera de aceite (*Elaeis guineensis* Jacq) o palmera africana es una especie diploide (2n=32 cromosomas), y monocotiledónea perteneciente a la orden *Aracales* y familia *Aracaceae* (Dransfield y col., 2005). Está considerada como la planta fósil más antigua conocida. Diferentes indicios sitúan su origen en la Costa de Guinea -Este África- (Corley y Tinker, 2007), dónde se utilizaba con fines culinarios hace mas 5000 años (Zeven, 1967). Su primera descripción botánica fue realizada por Jacquin en 1763. Existe otra especie asociada a este género conocida como palmera americana o noli, *E. oleifera* HBKCortes.Esta especie fue descrita por Cortés oficialmente en 1897, situando su procedencia en Panamá, Costa Rica y Colombia.

El aceite de palma africana se extendió hacia Indonesia en 1482 con fines comerciales, y hacia América en 1562 con el objetivo de alimentar a los esclavos procedentes del continente africano (Corley y Tinker. 2007). Aunque el momento clave para su introducción en Asia como cultivo comercial tuvo lugar en 1848 cuando se plantaron 4 semillas en el jardín botánico de Buitenzorg (Bogor, Java, Indonesia) procedentes de Ámsterdam e islas Mauricio. Estas cuatro palmeras obtenidas eran muy similares, y la uniformidad mostrada en sus progenies sugiere la procedencia de un único parental. Parte de estos descendientes llegaron a Deli -Sumatra, Indonesia- en 1857 (Whitmore, 1973) donde se distribuyeron por la región con fines decorativos. En 1860 surgieron las plantaciones experimentales, y comenzóc su explotación comercial, siendo el origen de las palmeras actuales. El rápido aumento del mercado de aceite durante el siglo XX fue un gran incentivo para la búsqueda de un sistema de producción más eficiente que permitió una rápida expansión hacia el resto de Indonesia y Malasia, donde es el cultivo más importante (Hartley, 1988).

En la actualidad, la mayoría de las plantaciones experimentales y comerciales se concentran en el cinturón húmedo intertropical: Sudeste de Asia (Indonesia y Malasia), región central y oeste del continente africano y Sudamérica (Hartley, 1988). Este emplazamiento proporciona al cultivo las condiciones adecuadas de irrigación, precipitación (entre 200 y 300 cm3/año) y temperatura (entre 22ºC y 33ºC) para su crecimiento y desarrollo (Hartley 1988; Goh, 1999).

2.2. Biología y Morfología

La palmera de aceite tiene un tallo no ramificado cubierto por 35-60 hojas pinnadas que puede crecer hasta alcanzar los 20 metros de longitud. Cada hoja produce cada mes dos o tres palmas compuestas cada una de ellas por un raquis con peciolo y foliolos (Figuras 4a y 4b).

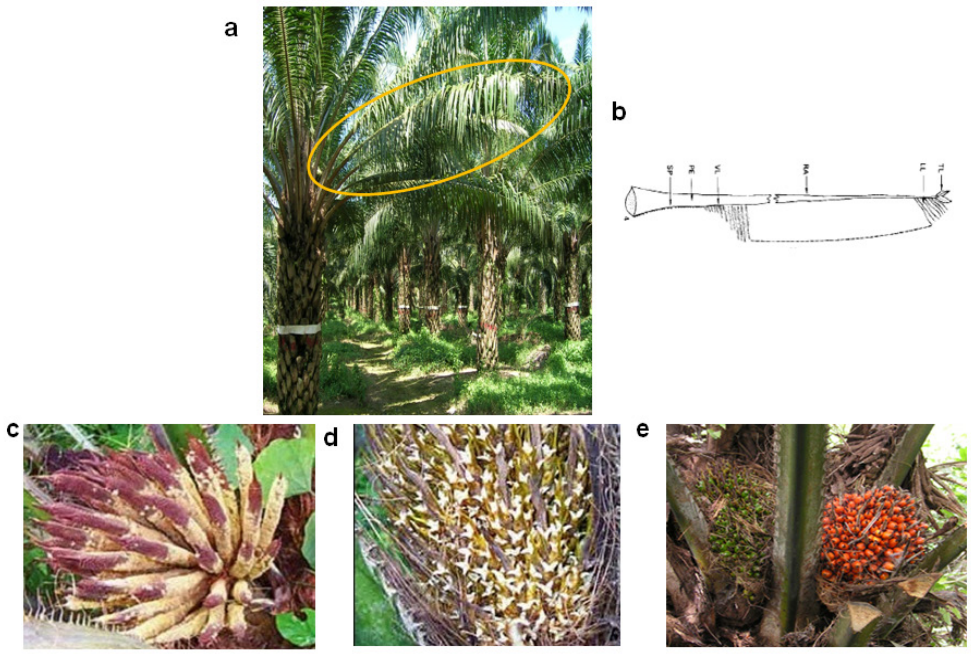


Figura 4: a) Palmera de aceite; b) Hoja de la Palmera de Aceite. TL: parte terminal de los foliolos ovales; LL: foliolos; RA:raquis; VL: foliolos con laminas vestigiales; PE: Peciolo; SP: espinas. (Corley y Tinker, 2003); c) Inflorescencia masculina; d) Inflorescencia femenina;e) Racimo de frutos.

Esta planta, alógama y monoica, combina ciclos de flores masculinas y femeninas que maduran a diferentes tiempos para asegurar la polinización cruzada (Figura 4c; 4d). Su ciclo floral cambia cada 4 o 6 meses, influenciado por factores genéticos y ambientales (Purseglove, 1972).Aunque sus inflorescencias no aparecen hasta los 2 o 3 años de edad, cuando la planta alcanza la madurez (Soh y col, 2003). En algunas ocasiones se desarrollan flores hermafroditas debido a que cada primordio floral posees dos órganos masculinos y dos femeninos. Ambos órganos pueden desarrollarse completamente, siendo más común en plantas jóvenes y durante la transición floral del ciclo (Beirnaert, 1935). Como curiosidad, uno de los indicadores de productividad es el "Sex Ratio", valor obtenido de la proporción de flores femeninas del total de inflorescencias. Es deseable que este valor sea alto para lograr una alta productividad. Las palmeras jóvenes pueden tener un ratio cercano al 98% disminuyendo con la edad (35% en las más antiguas) (Lattif, 2000).

Los frutos de la palmera de aceite son ovoides o alargados con una longitud entre 2-5 cm de largo y un peso entre 5 y 20 g. Están compuestos por tres partes bien diferenciadas la cáscara o epicarpio, la pulpa o mesocarpio y el endocarpio o el núcleo donde está la semilla (Figura 5). Además, forman racimos cuyo peso puede oscilar entre 10 -30 kg, conteniendo aproximadamente cada racimo 1500 frutos (Figura 4e).

a

b

	Aceite de Palma		Aceite de Palmiste
Ácido Palmítico (P) 16:0	44%		9%
Ácido Oléico (O)18:1	40%		15%
Ácido Linoléico (L)18:2	10%		2%
Ácido Caprílico 8:0			3%
Ácido Cáprico 10:0			4%
Ácido Laúrico 12:0			49%
Ácido Mirístico 14:0			16%

Figura 5:a) Fruto de la palmera de aceite y sus principales partes. El aceite de palma se obtiene de la pulpa, y de la semilla se obtiene el aceite de palmiste. Ambos productos se diferencian en su composición de ácidos grasos por lo que su funcionalidad es diferente y por tanto su utilidad es diferente. b) Composición en ácidos grasos de aceite de palma y de palmiste (Basiron, 2005). En el aceite de palma el contenido de ácidos grasos saturados está en la misma proporción que los insaturados siendo mayoritarios el oleico. Además este aceite es único respecto a la composición de sus triglicéridos, ya que posee un número significativo de AGS en 2 posiciones del TAG (POP, PPO) o en una única posición (POO, OPO, y PLO) permite su fácil separación en dos productos, oleína y estearina. El aceite de palmiste, en cambio, es muy similar al aceite de coco en su composición, con una alta composición de ácidos grasos saturados.

Además el fruto puede clasificarse en función de dos características relevantes, como son el color del epicarpio (figura 6) y el grosor del mesocarpo (figura 7). Con respecto al color de su epicarpio los frutos *Nigrescens* son de color morado cuando son inmaduros, a medida que maduran adquieren una tonalidad púrpura. En cambio, los frutos *Virescens* cuando no han madurado todavía presentan una tonalidad verde, transformándose en anaranjados a medida que maduran, por lo que es más fácil identificar la maduración del fruto. En cuanto al grosor del mesocarpo los frutos se clasifican en *Dura*, cuando este es fino, *Pisifera* cuando el grosor es mayor y *Tenera* cuando el grosor es intermedio. El rendimiento del aceite extraído es superior para el fruto tenera, cuyo grosor es intermedio entre *Dura* y *Pisifera*.

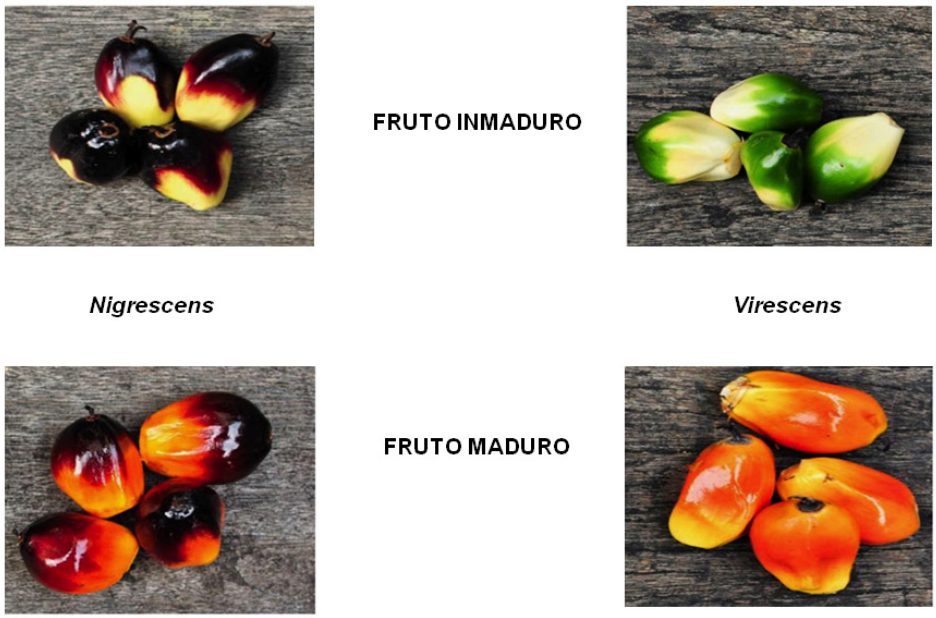


Figura 6: Tipos de frutos en función del color de su epicarpio. (Imágenes de Singh y col. 2014)

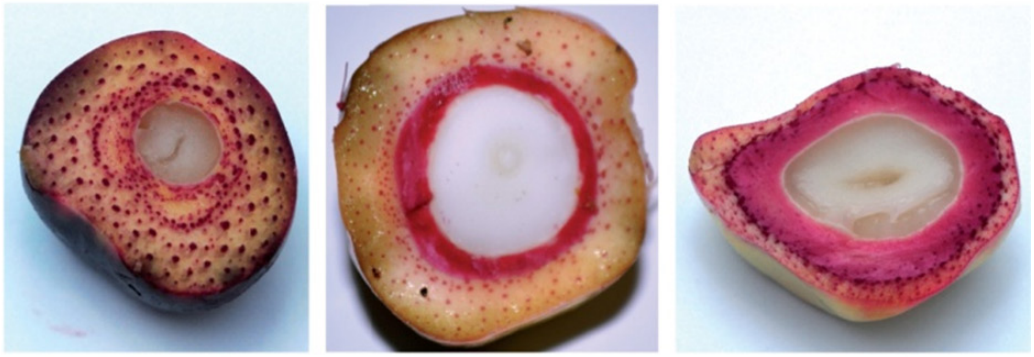


Figura 7: Clasificación de los frutos de la palmera de aceite en función del grosor de su mesocarpo. (Imágenes de Singh y col.,2013b)

2.3. Proceso de mejora del cultivo

Desde la edad antigua se ha producido una mejora clásica en los cultivos adaptando las especies silvestres a las diferentes condiciones de cultivo y creando nuevas variedades. Mediante el proceso de selección fenotípica el hombre mejoró las diferentes especies adaptándolas a sus necesidades.

La selección es un proceso sistemático donde se identifican los "mejores" individuos que serán elegidos como parentales en la siguiente generación mejorando la población inicial. La conservación de los "alelos superiores" permite que el cultivo, bajo determinados ambientes y condiciones, exprese los caracteres deseados desde el punto de vista agronómico, y por tanto, la creación de nuevas variedades.

2.3.1. Mejora tradicional

El objetivo de la mejora tradicional es la creación, identificación y caracterización de nuevas variedades basándose en los caracteres morfológicos y/o agronómicos de interés (Rallo y col., 2002). La selección fenotípica junto con la aplicación posterior de otras técnicas como la hibridación en variedades intrapoblaciones o polinización efectiva, entre otras, permitirán el desarrollo de las nuevas variedades.

En la palmera de aceite este proceso de selección comenzó con las progenies de las cuatro semillas plantadas en el Jardín Botánico de Bogor. Estas progenies se distribuyeron a diferentes plantaciones sitas en Deli -Sumatra- y Malasia, dando origen a la palmeras Deli (Rosenquist, 1986). A principios del siglo XX se establecieron en Indonesia, Malasia y algunas regiones de África diferentes centros de investigación para desarrollar el cultivo de la palmera de aceite. Estos centros crearon el primer banco de germoplasma, y representan la base genética para la mejora del cultivo (Cochard y col., 2009), afianzando el uso de las palmeras Deli *Dura*, las cuales tienen mayor calidad que las *Dura* de origen africano. En 1920, investigadores de la plantación experimental de Yangambi en la República Democrática del Congo - África- descubrieron el carácter monogénico codominante del gen para el espesor de la cáscara o *Shell-tickhness*, *Sh*, y determinaron las variedades de tipo de frutos existentes en las diferentes palmeras (Beirnaert y Vanderweyen, 1941).

El gen *Sh* está controlado por un "locus" o región del cromosoma con dos alelos. La clasificación de los tres tipos de frutos explicada en el apartado 1.2 se debe al descubrimiento de este carácter monogénico. El fruto *Dura* es homocigótico - *ShSh*- y se caracteriza por un grosor del epicarpio entre 2 y 8 mm y un bajo contenido de aceite en su mesocarpo. El fruto *Pisifera* -*shsh*- caracterizado por la ausencia de cáscara parece ser el material ideal para ser cultivado debido a un alto contenido de aceite en el mesocarpo, 95%. Sin embargo, sus flores femeninas suelen ser estériles, y las que no, suelen sufrir abortos durante el desarrollo del fruto, por lo que no se utiliza como parental femenino en el material comercial o de mejora. Por último, el fruto *Tenera*, heterocigótico -*Shsh*- se origina mediante el cruce de palmeras *Dura* como parentales femeninos y *Pisifera* como parental masculino (Figura 9). Estos frutos tienen un epicarpio más delgado y su mesocarpo es un 30% más grueso, por lo que el rendimiento de aceite por rácimo se incrementa en un 10% (Corley y Lee, 1992). Las características de estos tres tipos de fruto fueron descritas por Sambanthamurthi y col. (2009), y se resumen en la tabla 1.

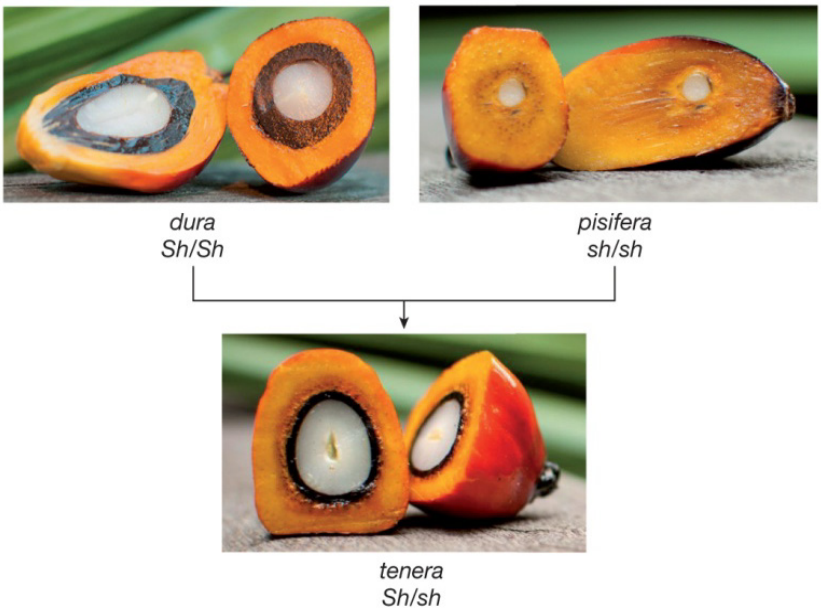


Figura 8: Efectos del gen *Sh*

Tabla 1: Características de los diferentes tipos de fruto: *Dura*, *Pisifera* y *TenerA.* M/F: Proporción tamaño del mesorcarpio en relación al tamaño del fruto; S/F: Proporción del tamaño de la semilla en relación al tamaño de l fruto; O/B: Proporción de aceite extraído en relación al tamaño del racimo.

CARACTERÍSTICA	*Dura*	*Pisifera*	*Tenera*
Grosor Cáscara (mm)	2- 8	0	0,5- 4
Anillo de Fibras	Si	No	Si
M/F (%)	35-55	95	60-96
S/F (%)	7-20	-	3-15
O/B (%)	16	-	26

A partir de este descubrimiento se enfatizó en el desarrollo de nuevos cruces para la obtención de palmeras *Tenera*, para ello en 1946 se creó una experiencia internacional en la que participaron diferentes centros de investigación de África y Malasia con el objetivo de intercambiar sus materiales. Las primeras observaciones mostraron que las palmeras con diferentes orígenes eran diferentes en cuanto a caracteres vegetativos, componentes de rendimiento de racimo y diferente relación aceite/mesocarpo. Por lo que se clasificaron en dos grupos, el primer grupo de palmeras se caracterizaban por tener pocos racimos pero de gran tamaño, en su mayoría eran las procedentes de Deli y Angola. Y, el segundo grupo tenían muchos racimos pero de pequeño tamaño y de procedencia africana como La Mé o NIFOR, entre otros (Gascon y De Berchoux, 1964)

Los cruces entre las palmeras Deli y las procedentes del continente africano obtuvieron mejores rendimientos que los cruces entre las palmeras del mismo origen, por lo que se generalizó el uso de *Dura* Deli y *Pisifera* como parentales femenino y masculino respectivamente para la producción de semillas *Tenera* (Ngando-Ebongue y col., 2012). A partir de este momento se iniciaron diferentes

programas de mejora para la producción de semillas híbridas *DxP* y la mejora de las poblaciones de *Dura* y *Pisifera*. Estos programas se basan en dos técnicas habituales en cultivos de polinización cruzada como son la **selección recíproca recurrente** -RRS- aplicada por el CIRAD (Centro internacional para el Desarrollo Agronómico de la Agricultura) en las plantaciones de Indonesia y en la región este de África y la **selección recíproca recurrente modificada** -RRMS- (Figura 9), aplicada en la mayoría de las plantaciones de Malasia influenciados por el grupo Unilever (Soh y col., 2003; Soh y col., 2011).

En **RRS** se obtienen una buena actitud general combinatoria (GCA), y las palmeras *Dura* y *Pisifera* se utilizan para la producción de semillas a nivel comercial. La principal ventaja de este método es que se pueden obtener más cruces recombinantes en un período de tiempo más corto, y en consecuencia, el espacio y esfuerzos necesarios para obtener las progenies de prueba son menores. Por el contrario, los parentales *Dura* no tienen progenies de prueba y los efectos de GCA de los cruces *DxD* y *TxT* no se reflejan en las características de los híbridos procedentes de los cruces *DxP* (Soh, 1999; Soh y Hor,2000).

En **MRRS** se obtienen buenas actitudes combinatorias generales y específicas , aunque el principal inconveniente es la necesidad de grandes superficies para probar los cruces *DxP*, y sus retrocruzamientos (Soh, 1999).

En la actualidad el material comercial más utilizado es el procedente de los cruces entre Deli x Lame, y Deli x Congo (Corley y Tinker, 2003). Aunque la estrecha base genética debida a las palmeras madre (Deli *Dura*) junto con un número limitado de polen *Pisifera* de diferentes orígenes (Kushairi y Rajanaidu, 2000) hizo necesario aumentar la diversidad genética para asegurar su conservación y continuar con el proceso de mejora durante la década de los 70. Para ello, se realizaron prospecciones en diferentes zonas de África -*E. guineensis*- y América -*E.oleifera*- recogiendo diferente material genético entre poblaciones silvestres y semisilvestres (Rajanaidu y Jalani, 1994). Uno de los centros de investigación que participaron en estas prospecciones, situado en Malasia y conocido como MPOB, "Malasian Palm Oil Board" tiene la mayor colección de germoplasma de palmera de aceite del mundo.

El germoplasma y el material élite *Dura* y *Pisifera* se utilizaron en el desarrollo de nuevos materiales como palmeras enanas de alto rendimiento, y otras con un alto porcentaje de ácidos grasos insaturados. Estos materiales se distribuyeron a la industria para su desarrollo en paralelo mediante cruces con el material utilizado hasta la fecha. Este cultivo no tiene variedades, sino una interpoblación de híbridos, muchas veces mezclados e incluso cruzados entre sí (Soh, 1999).

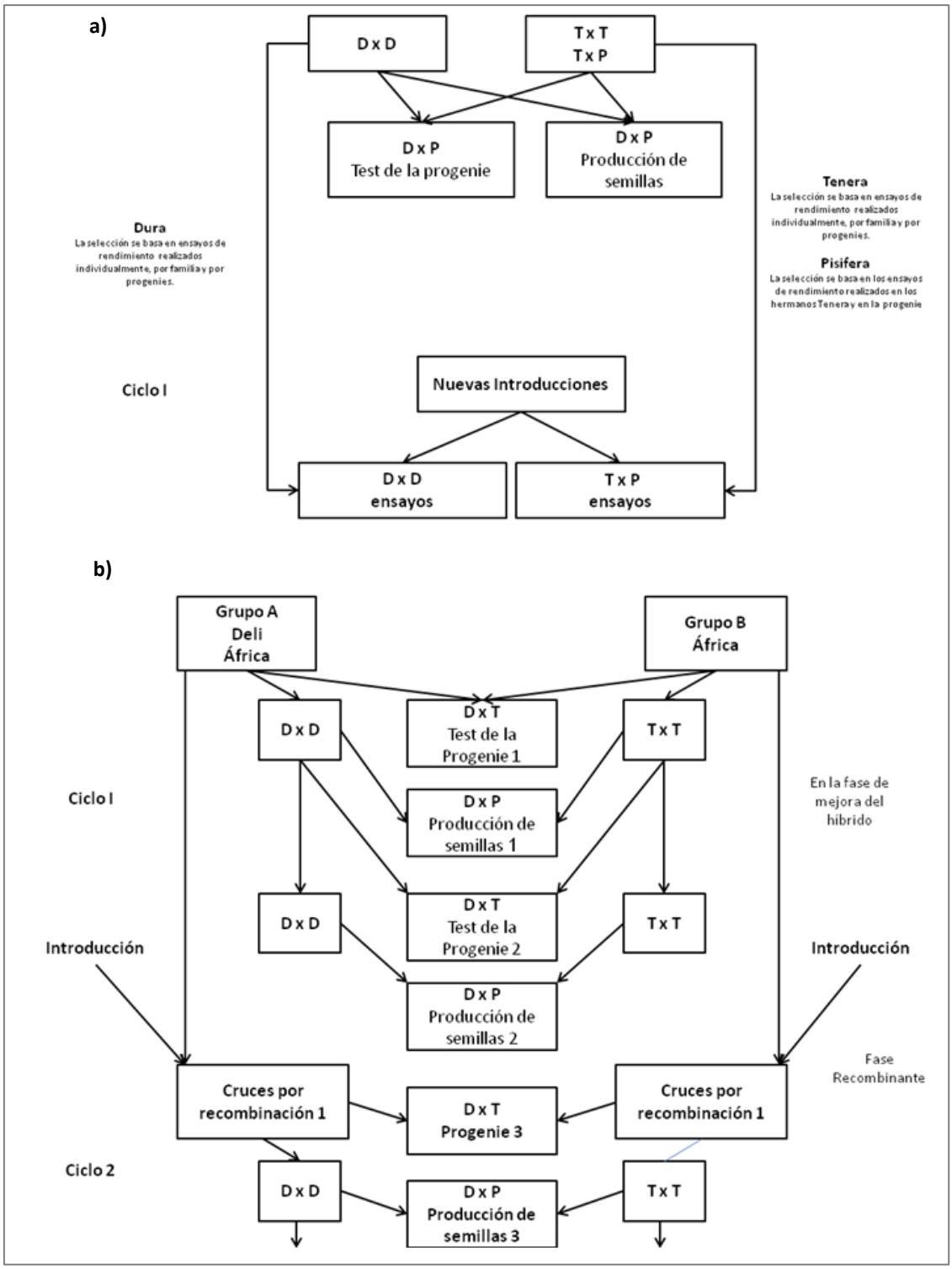


Figura 9: a) Esquema de la selección recurrente modificada. La selección de las palmeras *Dura* se realiza en base a las características individuales y de cada familia (Rosenquist, 1990). Las palmeras *Pisifera* se seleccionan en función de las características de los hermanos *Tenera* en las familias de los cruces TxT/P, ya que las *Pisiferas* suelen ser estériles. Estas *Pisiferas* seleccionadas se cruzan con las *Duras* seleccionadas para obtener la primera progenie de prueba; b) Esquema de la selección recíproca recurrente modificada. En este caso los parentales *D* y *T* se identifica mediante las características de sus progenies. Los parentales de los mejores cruces se autofecundan o se cruzan con sus hermanos. Las palmeras *D* y *P* de la progenie se utilizan para la producción de semillas (Imágenes adaptadas de Rajanaidu y col., 2000).

Estos métodos convencionales de hibridación y selección utilizados para la introducción de nuevos caracteres han demostrado su efectividad aumentando la productividad de los cultivos. Pero todavía existen dificultades que interfieren en este proceso de mejora como: 1) la probabilidad de combinación de caracteres deseables y no deseables, 2) el carácter poligénico de la mayoría de los caracteres de interés agronómico (Rajanaidu y col., 2000),3) incompatibilidades inherentes al propio cultivo de la palmera como son sus largos ciclos de selección, hasta 19 años por ciclo incluyendo la selección fenotípica (Wong y Bernardo, 2008). y la polinización controlada que disminuye la velocidad del proceso, y por último, 4) la necesidad de grandes superficies de plantación (143plantas/Ha) para su uso experimental (Jalani y col., 1993).

2.3.2. Mejora genética molecular

Las dificultades nombradas anteriormente pueden superarse mediante diferentes herramientas biotecnológicas, las cuáles van a permitir el estudio directo de su genotipo, y establecer su relación con el fenotipo (Tester y Langrigde,2010). Este conjunto de herramientas biotecnológicas pueden definirse como un proceso de mejora genética molecular, basado en el conocimiento de la información genética de la especie. El tamaño del genoma de la palmera de aceite es intermedio, aproximadamente 1.8Gb (Singh y col., 2013a) y un número de cromosomas de 2n=32 (Madon y col., 1995).

Rafalsky y Tingey (1993) definen el concepto de mejora molecular como el proceso de mejora convencional asistido por marcadores de ADN. Las principales ventajas de esta metodología es la selección en estados tempranos de la planta, antes de que se manifieste el carácter en cuestión, por lo que el proceso de selección se hace más eficiente, incluso en los cultivos perennes como la palmera de aceite., Además, permite crear nuevas variedades en menos tiempo (Mazur y y Tingey, 1995), por lo que esta metodología puede ayudar en el proceso de mejora para: 1. conseguir mejores rendimientos de aceite y mejorar la calidad del mismo, 2. conseguir algunos caracteres vegetativos deseables como una altura menor del tallo, y 3.para obtener resistencias a estreses bióticos y abióticos.

Por último, la secuenciación y ensamblaje del genoma por Singh y col. en 2013, junto con el análisis del transcriptoma en diferentes tejidos han supuesto uno de los grandes hitos de este cultivo permitiendo predecir cerca de 34800 genes, los cuales son susceptibles de participar en este proceso. En esta tesis se desarrolla una de las herramientas eficaces para la mejora del cultivo, tal y como se describe en los apartados siguientes.

3. SELECCIÓN ASISTIDA POR MARCADORES

Entre los objetivos de la mejora vegetal esta la búsqueda de combinaciones genéticas que sean más ventajosas en términos de producción, calidad y rentabilidad que las ya disponibles. La utilización de técnicas basadas en la biología molecular como los marcadores moleculares, unidas al estudio de la variación fenotípica amplía el rango de variabilidad seleccionable. Los marcadores moleculares son capaces de evidenciar una proporción significativa de variabilidad potencial no determinada por la

selección fenotípica (García-Mas y col., 2000) mejorando la eficacia y precisión de la selección (Collard y Mackill, 2008).

El término de Selección Asistida por Marcadores (SAM), más conocido por su acrónimo en inglés "MAS","Marker Assisted Selection", fue acuñado por primera vez por Beckmann y Soller en 1986, para referirse en los diferentes métodos y técnicas orientadas a la selección de individuos en base a los patrones de sus marcadores -genotipo- más que en los caracteres observables -fenotipo- (Boopathi, 2012).

3.1. Herramientas moleculares para la selección asistida por marcadores

3.1.1. Marcadores moleculares

Como se ha dicho, el objetivo de la mejora vegetal es la búsqueda de nuevas combinaciones génicas que supongan ciertas ventajas en términos de producción y rentabilidad frente las ya disponibles por los agricultores (cultivares, variedades híbridas…). La variación fenotípica de los caracteres agronómicos es el indicador del proceso, aunque hay una proporción significativa de variabilidad genética que el fenotipo no puede evidenciar, además de la influencia medioambiental existente.

Una de las herramientas más utilizadas es en este tipo de mejora son los **marcadores moleculares basados en ADN,** capaces de evidenciar una proporción significativa de esta variabilidad potencial (Capel y col., 2005). Estos marcadores muestran regiones donde hay variaciones en el ADN por lo que pueden representar las diferencias existentes entre organismos o especies (Jones y col 1997; Winter y Kahl, 1995). Al igual que los genes tienen una localización determinada en el cromosoma denominado "locus" (Collard y col., 2005).Son fragmentos de ADN sin función definida que en algunos casos pueden estar ligados a genes con caracteres de interés agronómico, como por ejemplo en el arroz el nivel de producción que está relacionado con el carácter número de brotes o el número de granos en cada espiga. Una mayor cantidad indicará un nivel producción mayor. Estas son las características utilizadas para seleccionar las progenies superiores de una población heterogénea (Boopathi, 2012). Cuando son los propios genes causantes del carácter se denominan marcadores funcionales, cuya clonación y conocimiento de su función favorece su uso como marcador específico. De esta forma permiten genotipar de manera precisa las poblaciones de mejora, e incluso descubrir si las variaciones fenotípicas observadas se deben a la existencia de alelos diferentes en uno o varios genes o a la expresión diferencial de genes reguladores.

Los marcadores moleculares participan también en la genómica estructural y funcional (Soh, 2011). La genómica funcional se encarga del estudio de la función del ADN a lo largo del genoma, -genes y elementos no génicos-, así como de los ácidos nucléicos y de las proteínas codificadas por el ADN por lo que puede aplicarse en todos los niveles de estudio: genoma, transcriptoma y proteoma (Pesvner, 2009). Así los marcadores son útiles en la mejora de plantas y en estudios genéticos para: 1.mostrar

diferencias entre individuos de la misma especie o diferente mediante los polimorfismos presentes en cada individuo, 2.construir mapas de ligamiento, 3.mapeo de genes y su identificación, 4.genotipar y conocer la diversidad del germoplasma, y 5 la identificación varietal, entre otros (Boopathi, 2012) Las principales características que los hace útiles para estas aplicaciones son: 1. su independencia del fenotipo, 2. son polimórficos o segregantes, 3. están libres de efectos epistático, 4. son independientes de las condiciones ambientales, 5. pueden ser evaluados en estadíos iniciales de la planta, y 6. pueden aplicarse a cualquier tipo de material vegetal (Tanksley, 1983; Powell y Moss, 1992; Phillipis Mora y col, 1995; Rallo y col. 2002). En la tabla 2 se resumen los principales marcadores utilizados en SAM.

Como se ha demostrado mediante estudios en numerosos cultivos los marcadores moleculares son herramientas valiosas aplicadas a la mejora vegetal. Collard y col. en 2005 demuestra con numerosos ejemplos este hecho y destaca cultivos como arroz, trigo, maíz, cebada, tubérculos, legumbres, semillas oleaginosas, especies de cultivos hortícolas y especies de pastos. Incluso algunos estudios han sugerido un importante papel para la mejora en la producción de alimentos mediante la mejora de programas convencionales de mejora de cultivos (Ortiz, 1998; Kasha, 1999; Collard y col., 2005). Su aplicación es valiosa en cultivos perennes, como la palmera de aceite, con largos ciclos de selección (Mayes y col., 1997). En este cultivo, la ganancia genética puede ser un proceso largo y tedioso, con una gran necesidad superficie de tierra para dedicarla a plantaciones experimentales (140-160palmeras/hectárea), lo que se traduce en una gran necesidad de mano de obra para gestionar y llevar a cabo los ensayos de mejora. Los marcadores moleculares permiten reducir el número de ciclos de reproducción, evaluar los alelos de "loci" de interés agronómico, y participar en la selección de parentales y material de siembra (Singh y Chea Suan, 2005). Los estudios llevados a cabo en esta especie relacionados con marcadores moleculares son diversos y con diferentes objetivos de aplicación en programas de mejora, conservación y prospección (Rajanaidu y col., 2000). Entre ellos pueden destacarse estudios de diversidad genética mediante SSR (Abdullah y col., 2011; Arias y col 2014), identificación de clones en cultivo “in vitro” (Rival y col, 1998) y/o “fingerprinting” (Mayes y col, 1996), construcción de mapas de ligamiento (Billote y col., 2005; Seng y col., 2011) o para el análisis y mapeo de QTL (Billote y col 2010; Jeennor y Volkaert, 2013; Montoya y col., 2013).

Tabla 2: Marcadores moleculares más utilizados por su abundancia en la Selección Asistida por Marcadores. Surgen de mutuaciones puntuales, reordenamientos como inserciones o delecciones o por errores de replicación (Paterson,1996a; Collard y col., 2005)

Marcador Molecular	Principio	Polimorfismo	Nivel de polimorfismo	Abundancia en el genoma	Dominancia	Información previa de secuencia	Ventajas	Desventajas	Referencias
RFLP ("Restriction Fragment Length Polymorphisms")	Enzimas de Restricción. Análisis Southerm (Southerm, 1975)	Cambio de bases: inserciones, delecciones	Medio	Alta	Codominante	No	Robusto	Coste medio alto	Beckmann & Soller (1986); Tanksley et al (1989); Kochert (1994),
							Exacto	Alta cantidad de DNA (2-10µg)	
							Transferible a traves de poblaciones	Polimorfismo limitado sobre todo en lineas relacionadas	
RAPD (" Random Amplified Polymorphic DNA")	PCR con cebadores aleatorios	Cambio de bases: inserciones, delecciones	Medio	Muy Alta	Dominante	No	Rápido, económico y simple	Problemas con reproducibilidad	Welsh & McClelland (1990); Williams et al. (1990); Penner (1996),
							Múltiples "loci" a partir de un único primer.	No transferible	
							Poca cantidad de ADN (10-25ng)		

SSR o Microsatélites (single sequence repeat)	PCR de secuencias simples repetidas	Cambios en la longitud de las unidades de repetición	Alto	Medio	Codominante	Sí	Simple Robusto y exacto Transferible entre poblaciones Poca cantida de ADN (50-100ng)	Desarrollo de cebadores laborioso Gel de poliacrilamida	McCouch et al. (1997); Powell et al. (1996), Taramino & Tingey (1996)
AFLP ("Amplified Lenght Polymorphism")	Digestión con enzimas y PCR de los fragmentos	Cambios de bases, inserciones, delecciones	Muy alto	Muy alta	Dominante	No	Múltiples "loci" Reproducibles	Metodología complicada	Vos et al. (1995)
SNP ("Single Nucleotide Polymorphism")	PCR alelo específica	Cambio de Bases	Muy alto	Muy alto	Codominante	Si	Detectan marcadores con el minimo nivel de polimorfismo	Caro, necesidad de chips o secuenciación masiva	Wang et al. 1998(libro marcadores)

3.1.2. Mapas de ligamiento y QTL

Los mapas genéticos o de ligamiento indican la posición y la distancia genética relativa entre los marcadores moleculares a lo largo de los cromosomas (Collard y col., 2005). Para entenderlo con facilidad Paterson y col. (1996) lo compara con un mapa de carreteras de los cromosomas de los dos parentales. El mapeo es posible cuando existe una variación genética tenga o no consecuencias en la transcripción, en la traducción, la función de la proteína o en la expresión del fenotipo.

Uno de los puntos más importantes en la creación de un mapa genético es la elección de las líneas parentales que difieran entre sí para los caracteres de interés (Mohan y col., 2007; Jones y col. 2009; Herrero, 2013) o de los genotipos heterocigóticos que se autofecundaran. La población de mapeo, incluidos los parentales, debe ser genotipada para todos los marcadores moleculares elegidos, y posteriormente se construye a partir del cálculo de su frecuencia de recombinación (RF). El proceso de recombinación ocurre entre los marcadores en los cromosomas homólogos durante la meiosis celular. A mayor valor de la FR mayor será la distancia entre los marcadores. Los marcadores estrechamente ligados tienen mayor probabilidad de heredarse juntos en la descendencia.

Estos mapas genéticos son de vital importancia para identificar regiones cromosómicas que contienen genes que controlan caracteres cualitativos y cuantitativos de interés agronómico, facilitando la selección asistida por marcadores. La acción conjunta de las regiones genéticas que afectan al fenotipo se conoce como QTL - "Quantitative Trait Loci"- (Geldermann, 1975) y se puede decir que buscan asociar el fenotipo con los marcadores segregantes de la población (Kearsey y Farquar, 1998). Los marcadores muestran la presencia o ausencia de la región en el cromosoma, y se determina si hay diferencias significativas entre los grupos con respecto a la característica que se mide (Tanksley, 1993; Young, 1996). Las diferencias significativas entre las medias fenotípicas de los grupos, en función del marcador y de la población, indica que el marcador está ligado al QTL que controla el carácter.

Los efectos cuantitativos del QTL puenden ser estudiados mediante un análisis clásico mendeliano, pero cuando un marcador y un QTL estan estrechamente ligados hay una mayor probabilidad que se hereden conjuntamente (Beavis y col., 1998). Por ello se establecen relaciones estadísticas entre la herencia de los caracteres y los marcadores moleculares situados en el mapa. La segregación en la progenie permite localizar el QTL en el cromosoma y determinar en qué medida contribuye al fenotipo (Thoday, 1961). Cuando los experimentos están diseñados correctamente la estimación de la heredabilidad ayuda a comparar los efectos del genotipo frente a los efectos medioambientales. Estos estudios de caracteres cuantitativos han permitido descubrir que los marcadores localizados cerca del QTL con función biológica conocida pueden ser útiles para su utilización como genes candidato. Sin embargo, en la mayoría de los casos, la distancia genética existente entre el marcador y el gen/QTL es insuficiente para que este marcador permita un buen diagnóstico del carácter. Con el objeto de resolver esta situación, es conveniente desarrollar mapas de ligamiento genético de alta densidad, y obtener marcadores moleculares física y estrechamente ligados al gen/QTL que controle el carácter de interés (Collard y col., 2005).

La palmera de aceite es un cultivo con un número reducido de mapas de ligamiento (Mayes y col., 1997; Moretzsoh y col., 2000; Chua y col., 2001; Billote y col., 2005; Seng y col., 2011; Lee y col., 2015). La mayoría de ellos se basan en marcadores de ADN como RFLP, AFLP o SSR y se desconoce si pertenecen a regiones codificantes o no del genoma. Estos marcadores son fuente de información de los niveles de variabilidad genética, pero no hay ninguna información acerca de las posibles funciones biológicas. Estas posibles funciones biológicas pueden ser detectadas mediante el mapeo del transcriptoma. Algunas técnicas como "cDNA-AFLP" o los "microarrays" permiten aislar genes y obtener marcadores con sentido biológico, tal y como se desarrollará en el capítulo 2 de esta tesis doctoral. Singh y col. publicaron en 2008 y 2009 dos estudios realizados en ADN complementario derivado de ARN mensajero y basados en marcadores RFLP, y Tranbarger y col. en 2012 publicaron un mapa basado en EST o etiquetas de secuencia expresadas derivadas de microsatélites (SSR). Los estudios publicados relacionados con el análisis y mapeo de QTL's tampoco son numerosos, y pretenden situar los caracteres de producción en los mapas de ligamiento (Rance y col 2001; Singh y col 2009; Billote y col 2010; JEennor y Volkaert, 2013; Montoya y col., 2013). Estos caracteres están relacionados con el desarrollo de las estrategias de mejora de los híbridos procedentes de los cruces de DxP, para mejorar el rendimiento en la obtención de aceite, mejorar su valor nutricional obteniendo una mayor proporción de ácidos grasos insaturados o un alto contenido en carotenoides, y por último algunos caracteres vegetativos como pueden ser la obtención de palmeras enanas. La mayoría de estos caracteres con variación cuantitativa pertenecen a sistemas poligénicos influenciados por factores medioambientales en su mayoría (Vargas y col. 2006).

El mapeo genético también permite comparar diferentes mapas entre especies relacionadas o mapeo comparativo, anclar mapas físicos y facilita la clonación posicional de genes de interés (Semagn y col., 2006). Estas acciones ayudan a crear mapas funcionales muy útiles para la búsqueda de genes candidato relacionados con los caracteres de interés agronómico, y además es el paso previo a la secuención del genoma.

3.1.3. Genes candidato

El principal objetivo del mapeo de QTL es identificar los genes responsables del QTL concreto y los mecanismos que afectan a la variación del carácter (Remington y Purugganan, 2003). En *Arabidopsis thaliana* se encontraron diferentes factores de transcripción que controlaban el tiempo de floración (Ratcliffe y Richman, 2002), o por ejemplo, algunos genes regulatorios y otros que codificaban enzimas que contribuían a la variación en caracteres metabólicos (Mitchell-Olds, 1998), como la ruta glicolítica que se activa en los procesos de defensa de la planta.

Un **gen candidato (GC)** puede definirse como aquel gen con función biológica conocida que participa directa o indirectamente en la regulación del carácter sujeto a estudio (Zhu y Zhao, 2007). Se proponen genes con función conocida y secuenciados que correspondan con los "*loci*" donde se encuentran los principales caracteres, bien cualitativos o cuantitativos (QTL), de interés. La hipótesis asume que la presencia de polimorfismos moleculares dentro del GC se relaciona con la variación

fenotípica (Pflieger y col., 2001). Esta estrategia fue aplicada en primer lugar en genética humana y animal (Rothschild y Soller, 1997) y desde los años 90 en plantas (Byrne and Mc-Mullen, 1996).

Los GC relacionados con los caracteres agronómicos de importancia pueden identificarse y localizarse mediante el uso de marcadores funcionales estrechamente ligados al QTL de interés. Estos marcadores funcionales pueden obtenerse mediante: 1. el estudio del transcriptoma por las técnicas citadas en el apartado anterior, ya que la variación de los caracteres son consecuencia directa de una variación en el transcriptoma y proteoma, 2.mediante genómica comparativa utilizando especies próximas genéticamente para identificar y caracterizar posible GC, o bien, 3. por su búsqueda "*in-silico*" utilizando diferentes herramientas bioinformáticas para su búsqueda (Zhu y Zhao 2007), tal y como se explicará en el capítulo 1 de esta tesis.

4. MAPEO POR ASOCIACIÓN

Como se ha explicado en el apartado anterior, una parte esencial del programa de mejora es la capacidad de identificar las variaciones de los diferentes genotipos y asociarlas con los genotipos de las poblaciones de estudio. Este proceso se realiza típicamente mediante la creación de mapas de ligamiento genético y la identificación de las regiones donde se encuentran situadas los caracteres cuantitativos agronómicos relevantes del cultivo.

Los actuales avances de la "Era Genómica", entre ellos la secuenciación masiva, han reducido los costes del genotipado de múltiples marcadores, sobre todo SNP, y la aplicación de nuevos métodos de selección asistida por marcadores, como es el **mapeo por asociación**. Este enfoque permite, gracias a estas nuevas tecnologías, explotar la diversidad natural presente en las plantas (Zhu y col., 2008).La aplicación de los estudios de mapeo por asociación implica la interacción de diferentes disciplinas como la genómica, la genética estadística, la biología molecular y la bioinformática para formar la base de la selección, evaluación y asociación de regiones genómicas para la correlación con los caracteres de variación (Oraguzie y col., 2007) constituyendo una herramienta potencial y novedosa en el mejoramiento de cultivos.

Su aplicación en cultivos comenzó a principios del siglo XXI, el primer estudio fue reportado en maíz y asociaba el gen *Dwarf8* con el tiempo de floración (Thornsberry et al., 2001), posteriormente se han ido desarrollando diversos estudios en otras especies de cultivo, como arroz, maíz, trigo, patata, o cebada (Palaisa et al., 2004; Agrama et al. 2007; Breseghello y Sorrells, 2006; Kraakman et al., 2006; Malosetti et al., 2007) y en árboles frutales como manzano (Cevik et al., 2010) y peral (Oraguzie et al., 2010). En palmera de aceite son dos los estudios publicados basados en mapeo por asociación. En uno de ellos se identificaron tres regiones asociadas con un alto contenido de aceite en el mesocarpio (Teh y col., 2016), y Babu y col. (2017) han identificado dos posibles regiones relacionadas con la proporción F/B y O/B.

Como en los mapas de ligamiento y QTL, el objetivo del mapeo por asociación es identificar las variaciones alélicas funcionales ligadas a diferencias fenotípicas para un carácter de interés (Oraguzie,

2007)(Figura 10). Este estudio se aplica a una población de individuos no relacionados entre sí, esto es, que a diferencia de los mapas de ligamiento no proceden de cruzamientos dirigidos (Zhu y col., 2008; Rafalski 2010;).En 2006, Yu y Buckler determinaron tres ventajas principales del mapeo por asociación frente al analisis de ligamiento tradicional: 1. mayor resolución de mapeo,2. menor tiempo y, 3. más número de alelos.

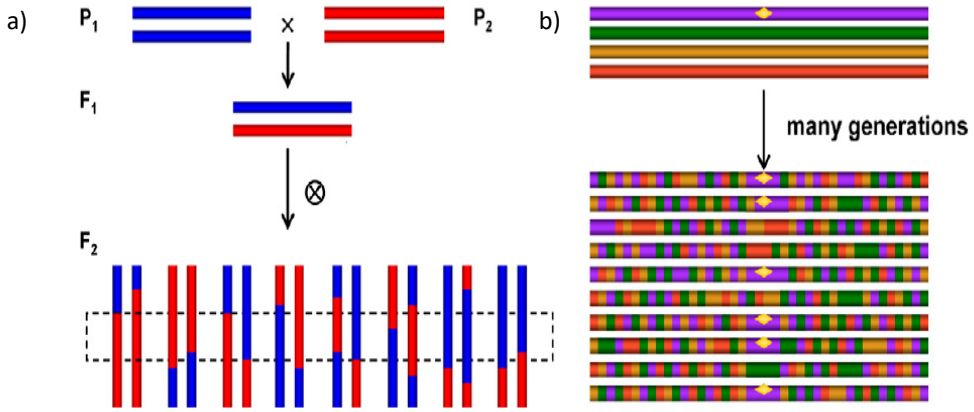


Figura 10 (Zhu y col., 2008): Análisis de ligamiento vs mapeo por asociación. Ambas estrategias permiten detectar asociaciones no aleatorias entre genotipo-fenotipo, basándose en la capacidad de heredar juntos polimorfismos funcionales o variaciones alélicas próximas. a) El análisis de ligamiento, se aplica a familias, en este caso aplicado a una población F2, procedentes de un cruce P1xP2 con fenotipos diferentes para el carácter de estudio (P1=azul; P2=rojo). Las opciones de recombinación, mostrados por los segmentos azules y rojos en F2, son pocas dentro de la propia familia y de sus ascendentes, por lo que la resolución de mapeo es baja; b) En el mapeo por asociación, el objetivo es la detección de la variante alélica o el haplotipo (diamante amarillo) responsable del fenotipo de interés en una población natural, alguno de ellos portador del alelo de interés. Los eventos recombinatorios a lo largo de la historia consiguen una mayor resolución.

4.1. Desequilibrio de ligamiento: Base conceptual del mapeo por asociación

En el **mapeo por asociación** (MA) se determina si existe asociación entre el marcador y el carácter mediante el **desequilibrio de ligamiento** (DL) en una población no relacionada (Flint-García y col., 2003), por lo que MA es una aplicación del DL (Figura 12). El **DL** es la combinación no aleatoria de los alelos en dos "l*oci*" debido a mutaciones y deriva genética ocurridas a lo largo de la historia las poblaciones que no tienen apareamientos controlados.

El DL (D) es la diferencia entre las frecuencias de los haplotipos observados y de los esperados en condiciones de equilibrio.

$$\mathbf{D = P_{AB} - P_A P_B}$$

Ecuación 1:P_{AB} es la frecuencia de los gametos con el alelo A y el B en los 2 "*loci*", P_A la frecuencia del alelo A y P_B la frecuencia del alelo B (Zhu y col. 2008). En la figura 11, se observa la situación en completo equilibrio (a) donde las probabilidades de cada haplotipo son iguales, y en DL (b) donde la probabilidad de los haplotipos es diferente.

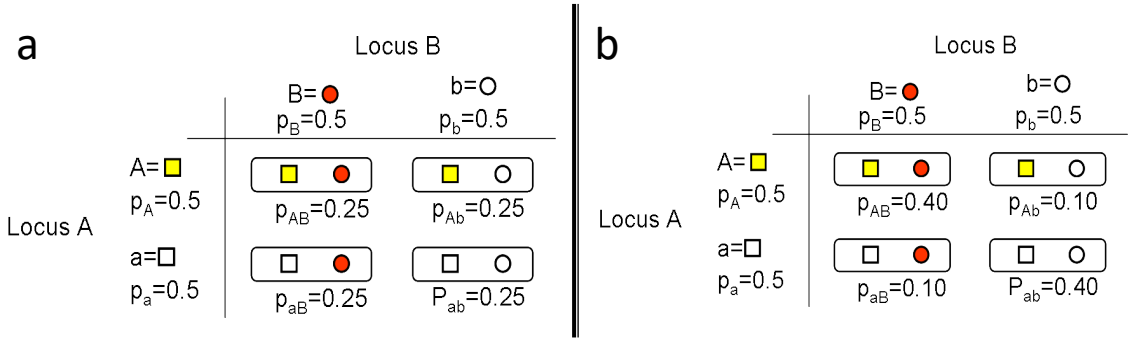


Figura 11: Equilibrio de ligamiento (a) y Desequilibrio de ligamiento (b)

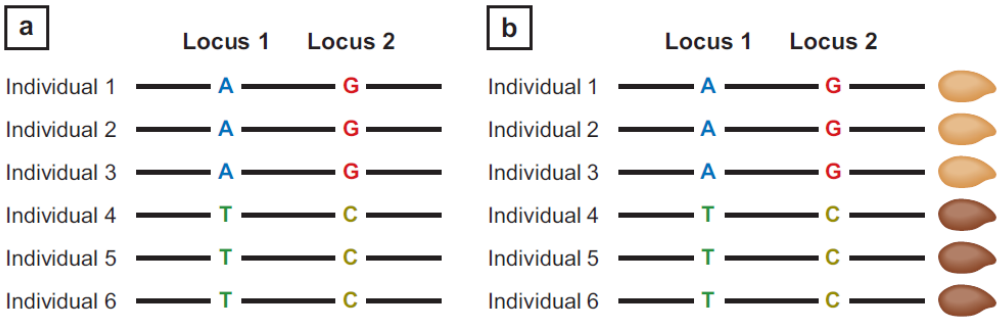


Figura 12 (Soto-Cerdá y Cloutier 2010): Principios del desequilibrio de ligamiento y el mapeo por asociación. a) Desequilibrio de ligamiento. El locus 1 y 2 presentan un patrón inusual de asociación entre los alelos A-G y T-C, aunque no tienen ninguna relación con el fenotipo. b) Mapeo por asociación. Los dos locus están en DL, y además hay una diferencia en el color de la semilla, para el alelo A-G es beige y para T-C marrón, por lo que hay una evidencia de asociación.

Son múltiples los **factores** que pueden afectar al **DL**, entre ellos la recombinación, las mutaciones, el sistema de apareamiento, los procesos de selección, o la estructura de la población (Soto-Cerdá y Couplier, 2010). La mutación crea nuevos sitios polimórficos que estarán en DL, y en cambio los eventos de recombinación romperán este DL, aunque no será uniforme a lo largo del genoma, como tampoco lo es la recombinación que varía en las diferentes regiones a lo largo del genoma. En cuanto, al sistema de apareamiento las especies alógamas tienden a DL más estrecho que las especies autógamas, como se ha demostrado en numerosas especies (Flint-Garcia y col.2003; Abdurakhmonov y Abdukarimov, 2008).

La selección de "*locus*" positivos crea un DL más extenso porque se limita la diversidad genética y se estructura la población, mediante que una selección equilibrada tiende a mantener o aumentar el nivel de polimorfismos (Soto-Cerdá y Couplier, 2010). Esta estructuración críptica de la población puede reducir la fuerza de la asociación (Balding, 2006), ya que se debe a una distribución no homogénea entre los alelos de una subpoblación con diferentes ancestros. Si estos subgrupos se seleccionan para crear un panel de líneas donde aplicar el MA, la mezcla de los individuos con diferentes frecuencias alélicas origina un DL. Si este DL es significante entre los "*loci*" pueden aparecer falsas asociaciones entre el carácter y el marcador.

4.2. Estrategias para abordar un estudio de mapeo por asociación

Un estudio de mapeo por asociación se inicia seleccionando la población, teniendo en cuenta que la resolución y la fuerza del mapeo, la densidad de marcadores y los métodos estadísticos a aplicar dependerán de la diversidad genética, de la extensión del DL a lo largo del genoma y de las relaciones de parentesco existentes en la población. El fenotipado en los caracteres agronómicos sujetos a estudio es otra etapa en la que el diseño de campo y la recogida de los datos experimentales deben ser exhaustivo y metódico. Este proceso consume mucho tiempo y recursos, y es una de los ejes principales del estudio.

En función del objetivo de estudio el mapeo por asociación puede abordarse de dos estrategias diferentes: la genotipificación del genoma -"Genome-wide association-GWAS"- o mediante la caracterización de genes candidato -"Candidate Genes -CG"- (Figura 13).

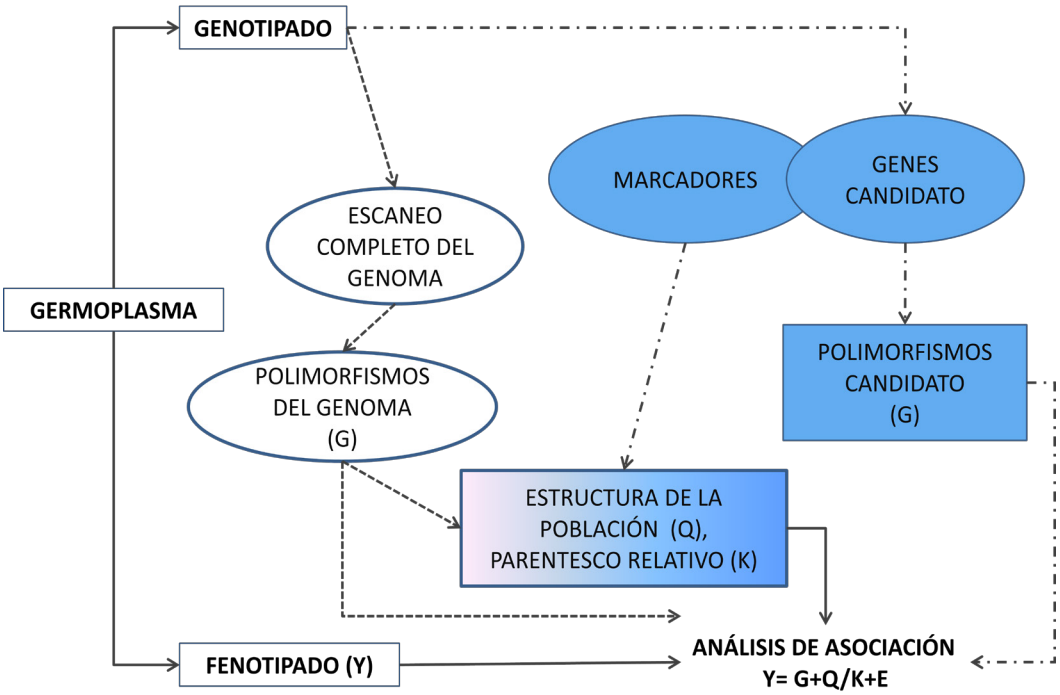


Figura 13 (Adaptación de Zhu y col. 2008): Esquema de las estrategias que abordan el mapeo por asociación (GWAS- blanco- y CG - azul-). En el análisis final se incluirá la estructura de la población (Q), las relaciones de parentesco (K) o ambas en función de la relación genética de la población de mapeo y de las divergencias en el carácter examinado. E es la varianza residual.

La **caracterización del genoma completo** se basa en el genotipado con un número de marcadores suficientes repartidos uniformemente por todo el genoma, probablemente al menos uno de los alelos funcionales en DL este genotipado por el marcador (Myles y col., 2009). Además, no es necesaria información previa de los posibles genes candidato (Zhu y col. 2008). En esta última década se han realizado multitud de estudios en numerosos cultivos como tomate (Sauvage y col., 2014) o arroz (Zhao y col., 2011).

El **mapeo por asociación** mediante **genes candidato** se basa en la selección de genes que participan en el control del carácter de estudio, por lo que se necesita un conocimiento previo de los genes. Esta estrategia tiene probabilidad de éxito si el carácter está bien caracterizado a nivel bioquímico y/o fisiológico (Pfliger y col., 2001). El coste económico es inferior, ya que se necesitan menos marcadores que genotipar, disminuyendo también el tiempo de análisis. En *Arabidopsis thaliana* (Ehrenreich y col., 2009) concluyeron que un tercio de los polimorfismos encontrados estaban asociados al tiempo de floración, partiendo de 51 loci genotipados. Esta estrategia se aborda en la presente tesis doctoral.

4.3. Modelos estadísticos

La elección del modelo estadístico aplicar en el mapeo por asociación depende de numerosos factores relacionados con el DL, la estructura de la población, de los caracteres de estudio y del conocimiento y recursos genómicos disponibles de la especie, entre otros (Abdurakhmonov and Abdukarimov, 2008).

En plantas son varios los métodos estadísticos que pueden aplicarse, siendo el más simple el modelo linear generalizado -"General Linear Model- GLM"-, similar a la regresión lineal donde la hipótesis nula para la variable categórica predictora es que un marcador SNP sin diferencia en el carácter de interés puede pertenecer a cualquier genotipo y por tanto los grupos son independientes (Bush y Moore, 2012).

Cuando las poblaciones son estructuradas se aplica la asociación estructurada (SA) para encontrar grupos relacionados estrechamente mediante el método bayesiano, entre otros y corregir a posteriori las falsas asociaciones mediante matrices de agrupamiento (matriz-Q) por regresión logística (Abdurakhmonov y Abdukarimov, 2008).

Los modelos lineales mixtos (Yu y Buckler, 2006) permiten eliminar los datos espurios debidos a la estructura de la población en el mapeo por asociación, ya que combinan la estructura de la población (Q-matriz) con los diferentes coeficientes establecidos por una relación de pares entre individuos o relación de parentesco, de la población de mapeo (matriz-K).

5. OBJETIVOS GENERALES E HIPÓTESIS

Esta tesis se enmarca dentro de un proyecto colaborativo del centro de investigación NEIKER TECNALIA (España) con SAMPOERNA AGRO (Indonesia) cuyo objetivo principal es el desarrollo de estrategias de mejora moleculares, las cuáles pueden ser implementadas y aplicadas mediante la selección asistida por marcadores para producir semillas de calidad superior y desarrollar nuevas variedades mejoradas en palmera de aceite africana. Además de disminuir los costes de producción y de contribuir a la mejora de la sostenibilidad del cultivo, contribuyendo a reducir las superficies de plantación.

Por tanto el **objetivo principal** de esta tesis es la búsqueda de marcadores genéticos que estén relacionados con los principales caracteres agronómicos relacionados con la productividad y el rendimiento del cultivo.

La **hipótesis** de esta tesis es :

"**El mapeo por asociación mediante genes candidato permite encontrar marcadores genéticos funcionales relacionados con el fenotipo mostrado en la población seleccionada de *E.guineensis* Jacq**."

CAPÍTULO 2: BÚSQUEDA Y SELECCIÓN DE GENES CANDIDATOS

1. INTRODUCCION

Como se ha explicado en la introducción a esta tesis (apartado 4) una de las posibles estrategias para abordar el mapeo por asociación es la búsqueda y selección de genes candidatos en función de los caracteres agronómicos deseados. Se puede definir como **gen candidato** a la región polimórfica con función biológica conocida que participa, directa o indirectamente, en diferentes procesos de regulación del carácter de estudio y susceptible de participar en la variación fenotípica. Sus efectos pueden confirmarse mediante un análisis posterior de asociación con el fenotipo (Pliegfer 2001; Zhu y Zhao, 2007).

El objetivo es buscar e identificar posibles variaciones en regiones próximas o en los mismos genes que puedan ser causa de variación fenotípica, o bien por un cambio a nivel de expresión, o bien por un cambio en la proteína. Si estos cambios se identifican, pueden correlacionarse estadísticamente con los datos fenotípicos del carácter en cuestión (Tabor y col., 2002). Las probabilidades de éxito aumentan cuando este carácter este bien descrito desde el punto de vista bioquímico y fisiológico, ya sea en su especie o en otras (Khan y Korban, 2012).

Estas regiones se seleccionan aplicando tres estrategias funcionales y/o posicionales, que no son independientes entre sí, ya que su aplicación es más efectiva cuando se aplica en la búsqueda una estrategia combinada (Zhu y Zhao, 2007).

Estrategias funcionales

La genómica funcional describe la función biológica de los genes a partir del conocimiento de su actividad en diferentes condiciones o estadios de la planta, entendiendo así la función de éstos en relación a su fenotipo (Morgante y Salamini, 2003). Estas características funcionales se reflejan en sus patrones de transcripción, y el análisis de su expresión proporciona modelos que permiten el estudio de caracteres complejos Por tanto, el análisis de expresión génica y la abundancia de transcriptos son claves para entender los patrones de expresión que aparecen durante diferentes procesos o estadíos biológicos (Xia y col.,2014), y se convierten en un punto de partida para la selección de genes candidatos.

A día de hoy se pueden conocer los niveles de expresión en el genoma a partir de diferentes técnicas. El punto de partida es común a todas ellas, y comienza con la construcción de librerías específicas de ADNc (ADN complementario) a partir de ARNm (ARN mensajero) de diferentes órganos, etapas de desarrollo o de plantas sometidas a estrés abiótico o biótico. La selección diferencial posterior ayuda al aislamiento e identificación de genes candidatos.

Los primeros marcadores utilizados para conocer la expresión del genoma fueron los marcadores de secuencia expresada (EST) (Adams y Kelley,. 1991). Su metodología es eficiente y rápida para entender la expresión génica (Fields, 1994) y el descubrimiento de nuevos genes “per se” (Verdun y col., 1998). Desde que en 2005 Jouannic y col. publicaron 2411 secuencias de EST de colecciones de

diferentes tejidos en palmera de aceite (*E.guineensis* Jacq), se han publicado númerosos estudios utilizando estos marcadores (Lin y col., 2009; Bourgis y col. 2011; Montoya y col. 2013) y el número de secuencias de EST en las bases de datos se ha multiplicado por 16 (41102 EST en enero de 2016: http://www.ncbi.nlm.nih.gov/nucest visitada 8 de enero de 2016).

La selección de genes candidatos puede realizarse mediante “microarrays" aplicados al conocimiento de la expresión diferencial a partir de librerías de ADNc. Aunque su coste es superior, permite conocer la expresión de múltiples genes al mismo tiempo (Schena. y col., 1995).Esta tecnología se ha aplicado en cultivo in vitro de palmera de aceite para conocer la expresión diferencial génica en diferentes estadios de desarrollo durante el proceso de embriogenésis somática (Low y col., 2008) y obtener así genes de referencia para su aplicación posterior (Xia y col., 2014).

Un método alternativo en la búsqueda de GC es la obtención de fragmentos polimórficos amplificados de ADNc (**cDNA-AFLP**) (Bachem y col, .1996). Esta técnica, basada en la reacción en cadena de la polimerasa, ha sido utilizada con éxito para el análisis sistemático de genes que participan en diferentes procesos, ya que se obtienen patrones de expresión diferencial, aunque su análisis de expresión no es válido desde el punto de vista cuantitativo (Breyne y Zebau, 2001). Es una alternativa “*low-cost*” que permite la detección de variaciones de expresión entre diferentes individuos y grupos (Korpelainen y Kostamo, 2010). Los resultados de esta técnica son robustos, sensibles y específicos. Además no necesita de una validación posterior, ya que su cinética de expresión es comparable mediante análisis Northerm (Bachem y col., 1996). Se ha utilizado en numerosas especies de plantas como patata (Bachem y col., 1996), arroz (Ahikiro y col., 2006), mijo (Jayaraman y col., 2008), trigo (Rampino y col., 2011), alubia (Shi y col., 2011) o trigo sarraceno (Gupta et al., 2012) para la identificación de genes candidato en diferentes condiciones o estadios como desarrollo de tubérculo, senescencia o respuestas a estrés, confirmando este hecho. En palmera de aceite los únicos estudios publicados hasta la fecha se relaciona con la identificación de genes expresados durante el proceso de embriogenesis somática (Pattarapimol y col., 2015), y la búsqueda de genes relacionados con una enfermedad relacionada con la sequía denominada "hard bunch phenomena".

Esta técnica aplicada junto con un **análisis de grupos segregantes** o BSA (Bulk Segregant Analysis) (Michelmore y col., 1991) que compara rápidamente mezclas de ADN individuales con diferentes características. En cada mezcla los individuos son idénticos para un determinado rasgo de interés, pero arbitrarios para el resto. La comparación de estas mezclas de individuos con fenotipos extremos permite una rápida identificación de marcadores ligados al carácter de interés. Esta metodología se utiliza con frecuencia para identificar genes próximos a un QTL, ya que cuando los individuos de una única familia se mezclan, se puede identificar algún marcador ligado al carácter de interés si los alelos han segregado en su progenie (Gupta y Rustgi, 2004).

Otra alternativa para la búsqueda de genes candidato funcionales es su búsqueda "***in silico"***. Los últimos avances en los software de extracción de datos facilitan esta búsqued, desde la gran

cantidad de literatura disponible en la red (Pubmed -http://www.ncbi.nlm.nih.gov/pubmed -, WOS - https://apps.webofknowledge.com-, GOOGLE - https://scholar.google.es/-.)(Patnala y col., 2013) como revisiones y/o estudios, entre otros. Las bases de datos públicas también proporcionan información útil relacionada con la función y la estructura (NCBI - http://www.ncbi.nlm.nih.gov/ -, o Gene Ontology - http://geneontology.org/-). Las funciones bien caracterizadas, la ontología y los genes anotados son herramientas que actualizadas son muy útiles en esta aplicación (Krallinger y col., 2008). Por otro lado, si las rutas bioquímicas, como por ejemplo la ruta biosintética de ácidos grasos mostrada en la figura 1, y/o fisiológicas (KEGG, - http://www.genome.jp/kegg/-),relacionadas con el carácter objeto de estudio se conocen bien pueden utilizarse también para identificar los genes relacionados con el mismo como posibles candidatos. En el caso de que el genoma de la especie no esté completo, la genómica comparativa, es la alternativa para encontrar genes homólogos funcionales, mediante métodos de alineamientos como Clustal (Thompson y col.,2002) y/o comparación de secuencias como el algoritmo BLAST (Altschul y col., 1990), en especies relacionadas y/o en especies modelo. Este hecho aumenta la probabilidad de que estos sean relevantes (Zhu y Zhao, 2007; Remington y Purugganan, 2003).

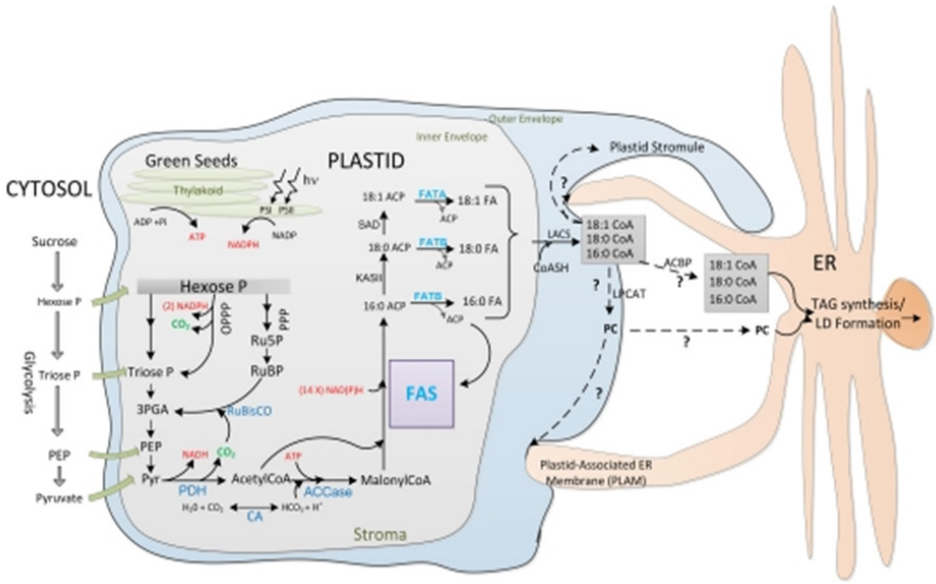


Figura 1: Biosíntesis de triacilgliceroles (TAG) en los diferentes tejidos y semillas (Chapman & Ohlrogge, 2012). Conociendo bien esta ruta bioquímica pueden identificarse genes candidato relacionados con la síntesis de ácidos grasos en palmera de aceite.

Como ejemplo, en 2012 los investigadores Sharma y Chauhan publicaron un estudio donde mediante genómica comparativa *in silico* identificaban las variaciones de 261 genes pertenecientes a la síntesis de ácidos grasos, triacilgliceroles y a la formación de cuerpos grasos en una especie modelo, *Arabidopsis*, y tres cultivos oleaginosos (*Brassica rapa*, *Ricinus communis*, y *Glicine max*). En palmera de

aceite un estudio compara los transcriptos y metabolitos existentes durante la partición del carbono en el mesocarpo entre la palmera de aceite y la datilera (*Phoenix dactilyfera*) (Bourgis y col., 2011).

Estrategias posicionales

Los **GC posicionales** son aquellos identificados a partir de la vinculación física del segmento cromosómico en un mapa con un QTL de interés (Zhu y Zhao, 2007). Los mapas de consenso utilizados son mapas de referencia saturados obtenidos a partir de la integración de mapas procedentes de diferentes genotipos del cultivo. De esta forma pueden situar en una región específica los marcadores próximos al QTL de interés, siendo susceptibles a utilizarse como herramienta para identificar marcadores estrechamente vinculados al carácter de interés (Borevitz y Chory, 2004), como puede observarse en la figura 9 del apartado 3.2.1 de materiales y métodos.

Cuando estos marcadores tienen una función biológica relevante, es decir son genes con función conocida, aumenta la exactitud estadística del mapeo de QTL y facilita un análisis minucioso de los caracteres complejos. Estos genes se consideran marcadores funcionales, se busca en ellos la presencia de polimorfismos como SNPs (polimorfismos de un solo nucleótido) capaces de aumentar la resolución de este tipo de mapeo. Si se encuentran en desequilibrio de ligamiento, el polimorfismo de esta secuencia lo hace susceptible al genotipado en los estudios de asociación (Rafalski, 2002). También se buscan en el mapa otras secuencias anotadas y ligadas al QTL que controla el carácter deseado. El grado de similitud de estas secuencias con otras de especies relacionadas o en sintenia junto con su colinealidad con los segmentos cromosómicos ayudan a identificar y aislar estas regiones hipotéticas como posibles GC (Remington y Purugganan,. 2003). Por ejemplo, para el QTL relacionado con el índice de iodo (I) en el aceite de palma la búsqueda más efectiva puede realizarse buscando genes relacionados con enzimas desaturasas próximos a este QTL. Ya qué este índice muestra el grado de insaturación del aceite y por tanto la cantidad de ácidos grasos insaturados presentes en él y en cuyo proceso de formación participan estas enzimas.

En plantas el mapeo de diferentes poblaciones ha sido exitoso para localizar genes candidatos de caracteres simples (Harjes y col, 2008;. Zheng y col., 2008) y aquellos con gran evidencia de ejercer un rol en el fenotipo de interés (Werner y col., 2005). Aunque esta estrategia se ha aplicado, sobre todo, en la búsqueda de genes de resistencia (RGA) (Pliefger y col.,2001). En palmera de aceite se han encontrado genes candidatos a partir de QTL buscando las homologías de las secuencias disponibles con los genes participantes en el metabolismo de lípidos y expresados en las semillas durante la maduración del fruto que es cuando comienza la síntesis y almacenamiento del aceite en él (Jeennor y Volkaert, 2013).

2. OBJETIVOS

En este capítulo se buscan, se identifican y seleccionan mediante una estrategia combinada, funcional y posicional, genes candidato relacionados con los caracteres agronómicos de interés de relacionados con la producción, el rendimiento y la calidad de aceite.

Esta estrategia combinada se basa en:

I. El análisis de la expresión diferencial mediante BSA cDNA AFLP

II. La búsqueda de genes *in silico*

III. La búsqueda mediante co-localización de secuencias anotadas con QTLs en un mapa integrado de palmera de aceite.

3. MATERIAL Y MÉTODOS

3.1. Análisis de grupos segregantes mediante cDNA-AFLP

3.1.1. Origen del Material Vegetal

El material vegetal utilizado para el análisis de los grupos segregantes desciende de distintos cruces dirigidos entre 225 genotipos *Dura* (parental femenino) y 50 genotipos *Pisifera* (parental masculino) con diferentes orígenes (Figura 2) y utilizados como parentales en distintos procesos de selección recurrente modificada. Se seleccionaron 242 genotipos en total procedentes de 29 progenies (Tabla 1).

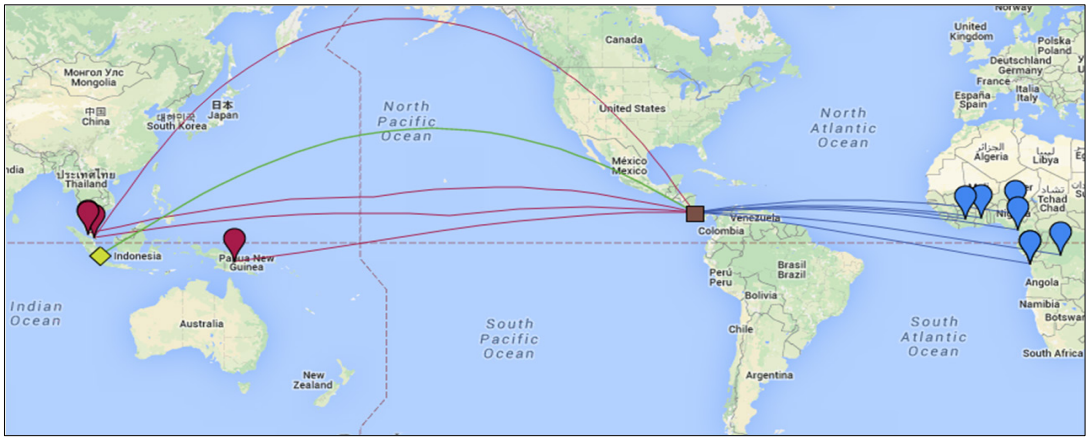


Figura 2: Origen de las líneas parentales femeninas y masculinas seleccionadas para la evaluación de sus progenies. Estos parentales llegaron a la estación experimental de Surya Adi ◊ en 1994 recogidos en Costa Rica ■ . Los parentales femeninos ◦- Dura- , M= Mardi, HC= Harrison and Crossfield, D= Dami, CH= Chemara, eran de origen asiático, aunque todos tienen su origen inicial en África y los masculinos ◦– Pisifera- de origen africano, N=Nigeria, Y=Yangambi, D= Dami Komposit, E= Ekona, A=Avros, G=Ghana, LM= La Me.

Tabla 1:Detalle del germoplasma de palmeras *Tenera* utilizadas en el análisis de grupos segregantes. En la primera fila y columna se muestran los parentales *Pisifera*, y *Dura* respectivamente. El resto de celdas muestran el número total de genotipos *Tenera* procedentes de cada cruce DxP. [a]HC = Harrison Crossfield

Nº DXP	Avros	Dami	Yangambi	Ekona	Ghana	LaMe	New Nigeria	Total
Chemara	10	8	0	10	14	1	22	**65**
Chemara x HC[a]	0	2	8	0	1	0	0	**11**
Dami	30	21	10	30	6	4	19	**120**
HC[a]	12	7	0	8	7	0	10	**44**
Mardi	0	0	0	0	0	0	2	**2**
Total	**52**	**38**	**18**	**48**	**28**	**5**	**53**	**242**

3.1.2. Caracteres de interés agronómico

Estas 242 palmeras (genotipos) de 15 años de edad se fenotiparon durante 10 años en los caracteres de interés agronómico relacionados con la producción, los componentes de racimo y vegetativos mostrados en la tabla 2. Cada carácter estaba formado por 4 o 5 familias de las que se seleccionaron los 5 mejores y 5 peores genotipos de cada familia para el análisis de grupos segregantes como se muestra en la tabla 3.

Tabla 2: Relación de los caracteres de interés agronómico en el cultivo, para los que hay QTL relacionados en el mapa LM2TXDA10D.

Caracteres de Interés Agronómico	**Descripción**
BN	Número medio de racimos/palmera/ año entre 3-5 años y/o 6-9 años (QBn3_5; QBn 6_9)
BW	Peso medio de racimo entre 3-5 años y/o entre 6-9 años (QBwt3_5; QBwt6_9)(kg)
CPO	Rendimiento de aceite/palmera/ año entre 3-5 años y/o entre 6-9 años (QPO3_5; QPO6_9)(ton/ha/año)
FN	Número medio de frutos por racimo (Fn)
FW	Peso medio de fruto (Fwt) (g)
MF	Ratio de pulpa con respecto al fruto (% PF)
OM	Ratio de aceite con respecto a la pulpa (%POP)
HT	Incremento de altura de tallo (Ht) (cm)
IV	Indice de Iodo (IV)

Estos caracteres son poligénicos y tienen diferente heredabilidad (figura 3), esto es, diferente estimación de la contribución relativa de las diferencias en factores genéticos y no-genéticos a la varianza fenotípica total de la población. Por tanto este valor indica en qué grado el rasgo o carácter se debe a causas genéticas o ambientales. Es necesario destacar que estos valores de heredabilidad son diferentes para cada población, dependiendo del tipo de población y del medio ambiente dónde se desarrollen. No pueden extrapolarse los valores a diferentes poblaciones, aunque la revisión de los diferentes estudios poblaciones en tamaños grandes de muestras se puede alcanzar una generalización, como se puede ver en la figura 3 adaptada de los datos aportados por Corley y Tinker (2003).

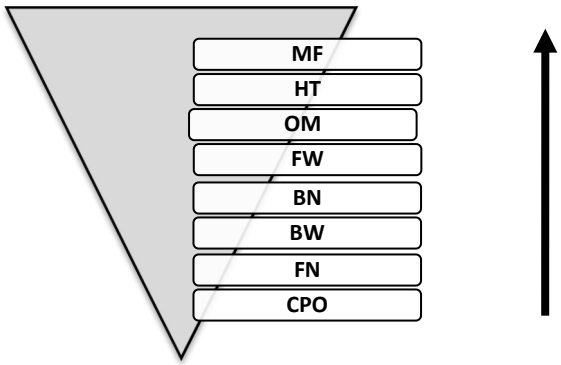


Figura 3: Caracteres de interés agronómico sujetos a estudio ordenados en función de su heredabilidad (Corley y Tinker, 2003) . MF= Mesocarp to fruit (%) o Relación del tamaño del mesocarpo con respecto al fruto; HT= " Height Increment" o Aumento de Altura (cm); OM= Oil to wet mesocarp (%) o relación de contenido de aceite frente a peso seco de mesocarpo ; FW= "Fruit Weight" o Peso de los Frutos (g); BN= "Bunch Number" o Número de Racimos;BW= "Bunch Weight" o Peso de Racimo (kg); FN= "Fruit Number" o Número de Frutos; CPO= "Crude Palm Oil" o Aceite Crudo (ton/ha).

Tabla 3: Familias seleccionadas para el carácter Número de Rácimos (BN). La tabla muestra los genotipos seleccionados para cada familia y los valores fenotípicos promedio para el carácter. En cada familia se obtuvieron dos muestras a partir de la mezclas del ARN extraído de los genotipos buenos para el carácter y de los malos, respectivamente, para el anaálisis de grupos segregantes. M=Mardi; N=Nigeria; CH= Chemara; HC= Harrison Crossfield; Y= Yagambi; D= Dami.

CARÁCTER	CRUCE	FAMILIA	GENOTIPO BUENO	VALOR	GENOTIPO MALO	VALOR
BN	M X N	552	1	13.8	9	7.7
			22	13.8	12	8.2
			23	16.0	41	8.5
			25	16.7	46	8.2
	(CHxHC)x Y	698	20	13.2	1	6.7
			26	13.6	17	6.2
			29	13.2	36	6.4
			30	13.7	45	7.1
	D X D	718	6	18.3	12	10.2
			14	16.8	22	8.4
			15	15.5	41	9.8
			19	16.3	44	8.4
	D X N	773	2	15.3	37	9.6
			4	16.0	41	10.1
			5	15.9	45	7.8
			11	16.3	48	7.4

3.1.3. Material vegetal utilizado en los análisis moleculares

Las hojas de las diferentes palmeras fueron recogidas durante abril de 2011 de la estación experimental de PT Binasawit Makmur, en Surya Adi Estate (latitud: 105°2'0"- 105°4'0" E, longitud: 04°1'0" – 04°2'0"S, elevación: 28m) y perteneciente a la provincia de Sur Sumatra (Indonesia).

3.1.3.1. Método BSA cDNA AFLP

El esquema del método de cDNA AFLP se muestra en la figura 4, y conlleva varios pasos que se describen a continuación.

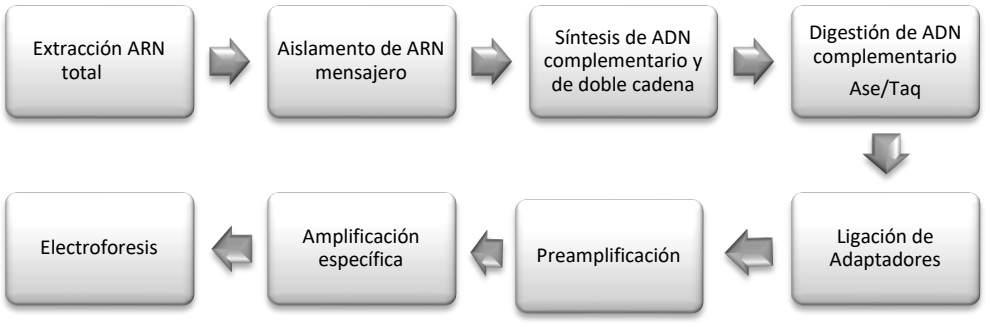


Figura 4: Esquema método cDNA AFLP (Bachem y col. 1996).

3.1.3.1.1. Extracción de ARN total

El material vegetal procedente de hoja se cortó en cuadrados de 2x2mm y se introdujo en microtubos de 1.5ml con RNA later (RNAlaterTM, Ambion Europe LTD.) para su conservación y transporte a España. El material vegetal se conservó a -20ºC hasta su utilización.

La extracción del ARN total se realizó introduciendo algunas modificiones en protocolo utilizado por Saïdi y col. (2009). Se descongelaron los 200mg de material vegetal en hielo y se secaron con papel secante (Figura 5a). Una vez seco los trozos de hoja se volvieron a introducir en microtubos (1,5ml) junto con dos bolas de tungsteno, y se sumergieron en nitrógeno líquido. Así, se congeló el tejido y se evitó la degradación del material genético. La homogenización del tejido se realizó con un molino mezclador (Retsch MM200) durante 90s y a una frecuencia de 1/13s para su correcto triturado (Figura5b).

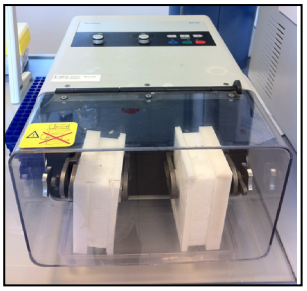
a.  b. 

Figura 5: a) Trozos de hoja en papel secante; b) Molino mezclador utilizado para la homogenización del tejido.

Una vez triturado el tejido vegetal se transfirió a otro microtubo de 1.5ml libre de ribonucleasas, se añadieron 500µl de tampón de extracción conservado a 4ºC (100mM Tris-HCL pH 8.0; 0.5% Nodidet 40; 50mM EDTA; 1% PVP K-30; 5% β-Mercaptoetanol), y se mezclaron con un agitador tipo vortex. A continuación, las muestras se incubaron a temperatura ambiente durante 10 minutos y se

añadieron 500µl de una mezcla de fenol cloroformo (5:1). Ambas soluciones se mezclaron con un vortex, y se centrifugaron durante 10min (12500rpm; 4ºC) para separar las fases, y recuperar la fase superior dónde se encuentra el ARN total. A esta fase se añadió el mismo volumen que lo recuperado de cloroformo, y se centrifugó de nuevo en las mismas condiciones. Las fases volvieron a separarse desechando la fase acuosa, añadiendo a la fase restante 400µl de isopropanol y 100µl de cloruro sódico 5M. Se mezclaron con una inversión manual de los tubos y se centrifugaron a máxima velocidad durante 45min a 4ºC. Las muestras se incubaron durante 10 min en hielo. El pellet se limpió con 1ml de etanol al 70% (v/v) y se centrifugó a 10000rpm durante 10 min a 4ºC. Finalmente, se eliminó el etanol de los tubos, se secaron los pellets a temperatura ambiente y se resuspendieron en 25 µl de agua estéril y desionizada. Todo el proceso se llevó a cabo rápidamente para evitar la degradación del ARN.

La cuantificación del ARN total se realizó mediante espectrofotometría UV/Vis (Nanodrop 2000, ThermoScientific, USA), y se midió el grado de pureza mediante la relación de la absorbancia A260/280. En un gel de agarosa al 1% de tampón TAE (40mM Tris-Acetato; 1mM EDTA, pH 8.0) con una tinción de Gel Red 1X (GelRed 10000X in DMSO, Biotium) se visualizaron las bandas 28S y 18S analizando su integridad.

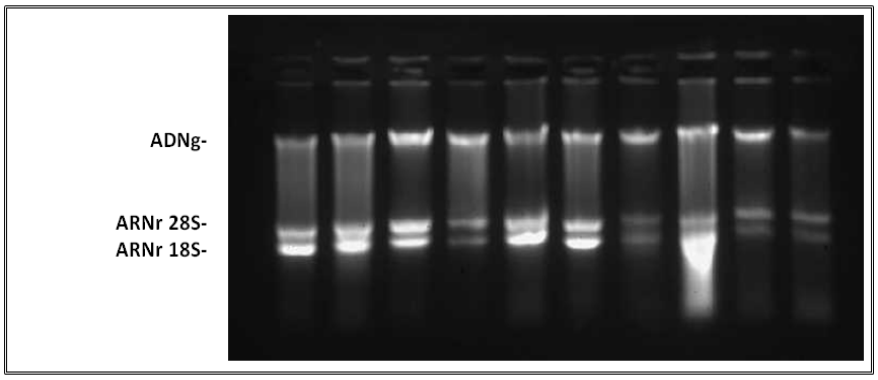


Figura 6: Ejemplo de la extracción de las muestras de ARN para una de las familias correspondientes al carácter número de frutos (FN).

Una vez cuantificado y revisado el grado de pureza se juntaron los ARN de los genotipos "buenos" y de los genotipos "malos" a igual concentración.

3.1.3.1.2. Aislamiento de ARN mensajero

Para aislar el ARN mensajero de cada mezcla se utilizaron bolas magnéticas recubiertas de estreptavidina (Dynabeads M-280 Streptavidin, Dynal A.S., Oslo, Norway) acondicionadas.

El acondicionamiento de las bolas paramagnéticas (0.1mg/µl) se realizó modificando el protocolo de Bachem (2002). Se utilizó 1µl de la disolución de bolas por cada µg de ARN presente en la muestra. El volumen total de bolas para cada muestra se limpió con una disolución tampón, 1X Stex (1M NaCl; 10mM Tris-HCl pH 8; 1mM EDTA; 0.1% (p/v) Triton X-100), se colocaron los diferentes tubos en un imán (MPC-S, Dynal S.A, Oslo, Noruega),y una vez que las bolas se adhirieron a la pared, se retiró la fase

acuosa. Esta operación se repitió tres veces. Las bolas limpias se resuspendieron en dos volúmenes del volumen inicial en una disolución tampón - 2x-Stex -.

A continuación, se añadieron 2 volúmenes de un oligonucleótido biotinado en el extremo 5', d[T]25V (MWG-Biotech Inc,USA), compuesto por 25 timinas y en el extremo 3' una base degenerada (V= A, C o G), y se incubó la mezcla durante 30min a temperatura ambiente. En esta etapa el oligo dT se unió a las bolas paramagnéticas mediante un enlace covalente entre la biotina y la estreptavidina.

En la etapa final el ARN total se desnaturalizó durante 3 minutos a una temperatura de 65ºC, se añadió el mismo volumen que la cantidad de ARN de las bolas unidas al oligo dT, y se incubó la mezcla a temperatura ambiente durante 30 minutos. Durante este periodo de incubación la cadena poliA, propia del ARN mensajero, se unió al oligo dT hibridado en las bolas magnéticas. La mezcla, bolas magnéticas/ ARN, se limpió 3 veces con un volumen de disolución tampón 1x STEX, y se resuspendió en 20µl de agua Mili-Q®. Para eluir el ARN mensajero, la resuspensión se calentó durante 5 minutos a 65ºC, inmediatamente se colocaron las muestras en el imán, se retiró la fase líquida, donde se encontraba el ARN mensajero, y se transfirió a un tubo eppendorf de 1.5ml libre de ribonucleasas.

3.1.3.1.3. Sintésis de ADN complementario

El ADN complementario se sintetizó a partir de ARNm en tres etapas, según el protocolo de Sambrook et al. (1989).

1ª Etapa: Síntesis de ADNc de una hebra

El ARN mensajero se incubó junto a un oligodT durante 2 min a 70ºC para favorecer la su hibridación, después se preparó la reacción para la síntesis ADNc mediante una transcriptasa inversa que incluía el oligodT y los hexámeros aleatorios junto con su disolución tampón. Por cada microgramo de ARN total se obtiene un 3-5% de ARN mensajero.

Tabla 4: Mix de reacción para la síntesis de ADNc de una hebra

Reactivos	Volumen (µl)
mRNA [ ]	20
Oligo $d(T)_{25}$ V (100ng/ µl)	1
5x VILO tampón	6
10x Mezcla de enzima Superscript VILO	3

.

La mezcla se puso en un baño de agua con agitación (Eppendorf Thermomixer® confort, USA) con el programa mostrado en la tabla 5.

Tabla 5: Programa del termociclador para la reacción en cadena de la polimerasa en la síntesis de ADNc de una hebra.

Tiempo (min)	Temperatura (ºC)
10	25
60	42
5	85

2ª Etapa: Síntesis de ADNc de doble hebra

Se preparó la mezcla de reacción como se decribe en la TABLA 4, y se incubó a 16ºC durante 2h..

Tabla 6: : Mix de reacción para la síntesis de ADNc de doble hebra

Reactivos	**Volumen (µl)**
ADNc de hebra única	30
10x ADNc buffer (200mM Tris-HCl pH 7.5; 750mM KCl; 100mM$(NH_4)_2SO_4$; 50mM $MgCl_2$;10mM DTT)	15
ADN polimerasa (10U/µl)	3.5
ARNasa H (2U/µl)	1.5
dNTPs (25mM)	1
H_2O Mili-Q®	98

Pasadas las 2 horas de incubación, se observó la calidad de ADNc mediante electroforesis en un gel de agarosa 1.5% (p/v). La imagen mostró un rastro entre 0.5 y 4 Kilobases.

3ª Etapa: Purificación de ADNc de doble hebra

La muestra final se mezcló con un volumen de fenol:cloroformo:isoaminoalcohol (25:24:1 en volumen) y se vorteó fuertemente. La siguiente etapa fue la centrifugación durante 20 min, a 4ªC y a 12500 rpm. Se recogió el sobrenadante y se introdujo en un microtubo nuevo de 1.5ml para precipitarlo con 2 volúmenes de etanol absoluto a una temperatura de -20ºC durante toda la noche. A la mañana siguiente, se centrifugó a máxima velocidad durante 30min a 4ºC, y el pellet se limpió con etanol frío a 70% (v/v). Por último, el pellet se resuspendió en 40µl de agua Mili-Q®.

3.1.3.1.4. Procesado de ADNc de doble hebra

Se aplicó el protocolo descrito por Bachem y col. (1998) que constaba de las siguientes etapas:

1ª - Digestión de ADNc

Se digirió una muestra de 500ng totales de ADNc sintetizado con dos enzimas de restricción Taq I, de corte frecuente, y Ase I, de corte poco frecuente o raro.

Enzima	Corte
Taq I*	T **C G A** A G C **T**
Ase I*	A T **T A A T** T A A T **T A**

* Enzimas de restricción (NEB Biolabs Inc. New Brunswick, NE, USA)

Esta digestión ocurrió en dos reacciones diferentes, primero se digirió con la enzima TaqI y en otra reacción con Ase I.

Tabla 7: Mix de reacción y condiciones de incubación para la digestión con el enzima de restricción TaqI.

Reactivos	Volumen (µl)	
ADNc de doble hebra	20	
Taq I	1	**Incubación 1h 65ºC**
10x R/L tampón (10mM Tris-HCl pH 7.5; 10mM Mg-Ac; 50mM K-Ac; 5mM DTT; 50ng/µl BSA)	4	
H_2O Mili-Q®	15	

Tabla 8: Mix de reacción y condiciones de incubación para la digestión con el enzima de restricción AseI.

Reactivos	Volumen (µl)	
ADNc de doble hebra	20	
Ase I	1	**Incubación 1h 65ºC**
10x R/L tampón (10mM Tris-HCl pH 7.5; 10mM Mg-Ac; 50mM K-Ac; 5mM DTT; 50ng/µl BSA)	4	
H_2O Mili-Q®	15	

2ª- Ligación

En esta etapa fue necesario preparar las hebras de los oligonucleótidos de anclaje o adaptadores. La hibridación de las parejas de adaptadores se realizó mediante una incubación a temperatura ambiente durante 10 minutos. La concentración final de los pares de adaptadores fue 5µM para TaqI y 50µM para AseI.

Tabla 9: Secuencia de los adaptadores utilizados en la ligación.

Oligonucleótido de anclaje TaqI	Hebra superior	5'-GAC GAT GAG TCC TGA C-3'
	Hebra inferior	5'-CGG TCA GGA CTC AT-3'
Oligonucleótido de anclaje AseI	Hebra superior	5'-CTC GTA GAC TGC GTA CC-3'
	Hebra inferior	5'-TAG GTA CGC AGT C-3'

Los adaptadores hibridados eran complementarios a los fragmentos de restricción obtenidos en la digestión, y se añadieron en la mezcla de reacción para la ligación. Ésta se produjo por la acción de una ligasa T4 (Invitrogen Inc., Barcelona, España).

Tabla 10: Mix de reacción y condiciones de incubación para la ligación del producto de digestión**.**

Reactivos	**Volumen (µl)**	
Producto de digestión	**50**	
Taq I Adaptador (50pM)	**1**	
Ase I Adaptador (5pM)	**1**	**Incubación**
ATP (10mM)	**1**	**2h a 37ºC**
10x R/L tampón	**0.5**	
T4 ADN ligasa	**1**	
H2O Mili-Q®	**0.5**	

El producto de ligación se diluyó 1:10 y fue el molde para la pre-amplificación.

3ª- Pre-amplificación del molde primario.

Los cebadores utilizados fueron universales y complementarios a los adaptadores ya ligados.

Tabla 11: Secuencia de los cebadores de la reacción de preamplificación.

Cebador TaqI +0*	5'-GAC GAT GAG TCC TGA CCG A-3'
Cebador AseI +0*	5'-CTC GTA GAC TGC GTA CCT AAT-3'

* MWG-Biotech Inc, USA

Tabla 12: Mix de reacción y programa utilizado en el termociclador para la preamplificación.

Reactivos	Volumen (µl)	Programa PCR
Producto digestión/ligación (1:10)	10	[94ºC,30 seg;55ªC,30seg; 72ºC,60seg]* 25 ciclos
Ase I (10µM)	1	
Taq I (10µM)	1	
10x tampón PCR	5	
$MgCl_2$ (50mM)	1.5	
dNTPs (25mM)	0.5	
H_2O Mili-Q®	32	

El producto se visualizó en un gel de agarosa al 1% en una disolución tampón de TAE 1x y se detectaron fragmentos entre 50 y 700 pares de bases (pb).

4º- Amplificación selectiva

El producto obtenido en la pre-amplificación fue el segundo molde y la muestra de esta etapa. Este producto se diluyó 10 veces. Los cebadores utilizados en esta amplificación también eran complementarios a los adaptadores, pero incluían 2 o 3 bases selectivas (NN o NNN) en su extremo 3' (MWG-Biotech Inc., USA). El cebador AseI incluía una molécula que emitía en el espectro de infrarrojos (IRD 700/800; LI-COR, Lincoln, Nebraska, USA) para facilitar la detección del producto en el sistema LI-COR.

Tabla 13: Secuencia de los cebadores específicos Taq I y AseI.

Cebador TaqI +2/+3*	5'-GAT GAG TCC TGA CCG ANN/N*-3'
Cebador AseI +2/+3*	5'- GAC TGC GTA CCT AAT NN/N*-3'

*N= BASES ESPECÍFICAS

Tabla 14: Nombre de los cebadores específicos y sus bases nucleotídicas terminales para Ase/Taq +2 y Ase/Taq +3.

Cebadores Ase(A)/Taq(T) +2 utilizados en las reacciones específicas		**Cebadores Ase(A)/Taq(T) +3 utilizados en las reacciones específicas**	
Código	NN	**Código**	NNN
A11/T11	AA	**A51/T51**	CCA
A12/T12	AC	**A52/T52**	CCC
A13/T13	AG	**A53/T53**	CGC
A14/T14	AT	**A54/T54**	CCT

Cebadores Ase(A)/Taq(T) +2 utilizados en las reacciones específicas		Cebadores Ase(A)/Taq(T) +3 utilizados en las reacciones específicas	
A15/T15	CA	**A55/T55**	CGA
A16/T16	CC	**A56/T56**	CGC
A17/T17	CG	**A57/T57**	CGG
A18/T18	CT	**A58/T58**	CGT
A19/T19	GA	**A67/T67**	CGA
A20/T20	GC	**A68/T68**	GCC
A21/T21	GG	**A69/T69**	GCG
A22/T22	GT	**A70/T70**	GCT
A23/T23	TA	**A71/T71**	GGA
A24/T24	TC	**A72/T72**	GGC
A25/T25	TG	**A73/T73**	GGG
A26/T26	TT	**A74/T74**	GGT

Tabla 15: Mix de reacción y programa utilizado en el termociclador para la amplificación selectiva.

Reactivos	Volumen (µl)	Programa PCR
ADNc (5:50)	5	[94ºC,30 seg;65ªC-56ºC (-0.7ºC/ciclo),30seg; 72ºC,60seg]* 53 ciclos [94ºC, 30seg; 56ºC, 30seg; 72ºC, 60seg] *24 ciclos
Ase I +N (50µM)	0.5	
Taq I +N (50µM)	0.3	
50x tampón PCR	5	
$MgCl_2$ (500mM)	0.55	
dNTPs (2.5mM)	0.8	
Taq ADN Polimerasa (5U/µl)	0.04	
H_2O Mili-Q®	2.25	

Todas las reacciones en cadena de la polimerasa se realizaron en termocicladores Primus 96/384 (MWG AG Biotech). Para poder visualizar los amplificados en la plataforma LI-COR (DNA Analyser Gene Reader 4300, LI-COR, MWG-Biotech) se añadieron 7µl de de un tampón de carga compuesto por formamida desionizada 98% (v/v), 50mM EDTA pH 8.0, y azul de bromofenol 0.5% (p/v), se desnaturalizó la reacción en un termociclador durante 5 min a 95ºC y se enfrío rápidamente en hielo.

Los geles de acrilamida para separar los fragmentos derivados de los transcriptos (TDF) se realizaron utilizando un gel de acrilamida al 6% (p/v) y 7M de urea.

3.1.4. Análisis de los fragmentos con expresión diferencial

Los geles se visualizaron buscando la presencia o la ausencia de bandas en cada familia, bien presentes en todas las familias de pool + para un carácter y ausentes en el pool -, o viceversa (Figura 7). Estos fragmentos se aislaron, se reamplificaron y se secuenciaron como se describe a continuación.

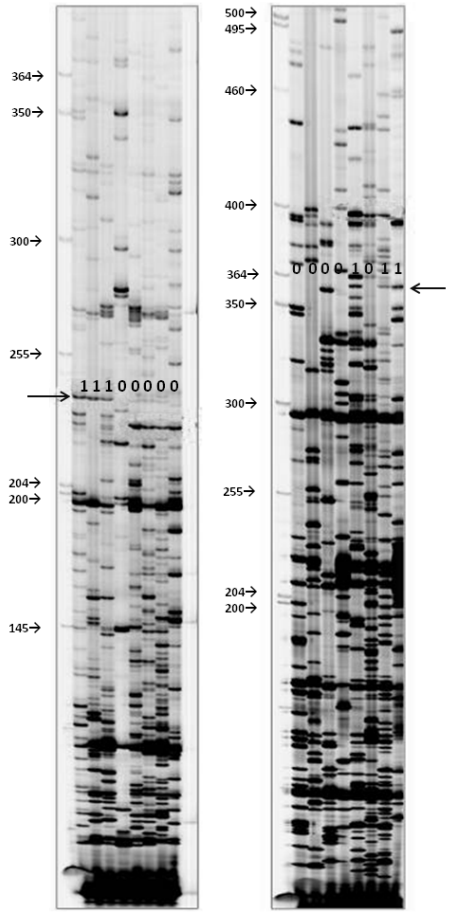


Figura 7: Ejemplo del conteo de bandas en el ánalisis de expresión diferencial en geles de acrilamida. En ambos geles, la primera columna es el marcador de peso molecular (Hyperleader II). El primer gel muestra la presencia del TDF presente en tres de las cuatro familias calificadas como "buenas" a una altura aproximada de 230pb. El segundo gel muestra el TDF en tres de las cuatro familias calificadas como "malas" para el carácter analizado.

3.1.4.1. Aislamiento de los fragmentos amplificados.

Los TDF se volvieron a cargar en un gel de acrilamida para visualizar en la plataforma LI-COR en el tamaño de banda deseado. Estos geles se escanearon en un equipo láser (Odyssey®Infrared Imaging System, LI-COR, MWG-Biotech) para identificar las bandas correctamente. Estas bandas se aislaron y cortaron. Las muestras cortadas se purificaron, machacaron y se incubaron con agitación a 65ºC durante 3 min, se recogió el sobrenadante que fue precipitado con dos volúmenes de etanol absoluto a -40ºC durante toda la noche. Después se centrifugaron a 4ºC durante 30 minutos, se secó el pellet y se diluyó en 15 µl de agua desionizada y estéril.

3.1.4.2. Reamplificación y secuenciación de amplicones

Para la reamplificación de los amplicones cortados y purificados se utilizaron las mismas condiciones de PCR que en la amplificación específica.

Tabla 16: Mix de reacción y programa utilizado en el termociclador para la reamplificación de las bandas seleccionadas.

Reactivos	**Volumen (µl)**	**Programa PCR**
Muestra	1	[94ºC,30 seg;65ªC-56ºC (-0.7ºC/ciclo),30seg; 72ºC,60seg]* 53 ciclos + [94ºC, 30seg; 56ºC, 30seg; 72ºC, 60seg] *24 ciclos
Ase I +N (50µM)	1.5	
Taq I +N (50µM)	0.75	
10x tampón PCR	2.5	
$MgCl_2$ (500mM)	0.55	
dNTPs (2.5mM)	2	
Taq ADN Polimerasa (5U/µl)	0.1	
H_2O Mili-Q®	17.5	

El producto de amplificación se diluyó 10 veces en agua, y se visualizó en un gel de agarosa al 2% en disolución tampón TAE 1x. A continuación, estos productos de amplificación se enviaron a secuenciar en un equipo electroforesis capilar (48-capillary DNA Analyzer, Fraunhofer IME, Aachen, Alemania) para secuenciación "SANGER" (Sanger F y col.1977).

3.1.4.3. Análisis de los fragmentos

Las secuencias obtenidas se procesaron mediante el software libre Ridom TraceIT (Rothgänger y col., 2006), y una vez revisadas se buscaron homologías utilizando alineamientos locales en la base de datos no redundantes de NCBI (BLASTn), y en las diferentes librerías disponibles del proyecto de OPGP (Oil Palm Genome Project) como EST ("Expressed Sequence Tag"), librería de "scaffolds" y ARN mensajeros de *Phoenix dactylifera*, ya que cuando se realizó este experimento no había información suficiente relativa al género *Elaeis*. Las secuencias homólogas altamente conservadas se compararon para dilucidar posibles funciones de los genes detectados. Las secuencias con una alta similitud fueron analizadas con Blast 2GO (Conesa et al. 2005) para identificar y mediante una comparativa de los alineamiento conocer los posibles genes con sentido biológico en los caracteres sujetos a estudio.

3.2. Búsqueda de genes candidato co-localizados

A partir de un mapa de referencia en palmera de aceite se buscaron y seleccionaron los genes candidato relacionados con los caracteres de interés agronómico descritos en 3.1.2 del presente capítulo.

3.2.1. Mapa de referencia en *Elaeis guineensis*

Inicialmente el genoma de la palmera de aceite africana era desconocido, por lo que se eligió la palmera datilera (*Phoenix dactylifera*) como especie de referencia. Su alto grado de colinealidad con *E.guineensis* Jacq. y la disposición pública de sus "scaffolds" en la página web de "Weill Cornell Medical College" de Catar (http://qatar-weill.cornell.edu/research/datepalmGenome/) justifica esta elección. En sus "scaffolds" se mapearon los marcadores secuenciados y publicados en NCBI (http://www.ncbi.nlm.nih.gov) de nuestra especie como EST y SSR, así como secuencias disponibles de diferentes autores como Brugis y col. (2011) y Tranbanger y col. (2012). A partir de este mapeo se pudieron anclar los "scaffolds" de *P. dactylifera* al mapa de referencia de *E. guineensis* Jacq. obtenido en el proyecto europeo "Link2Palm" para la población LM2TxDA10D.

RN	Chr	POS	OP_Scaf	PD_Scaf	DT	D-Name	M-Name	MT	DES	Pr	Fr
994	1	107,0			DL	361/6	EAgA/MCTg	AFLP	P1	361	6
Q	1	107,0	-----		QTL	QaBwt_a	---	QTL	P1	--	--
Q	1	107,0	-----		QTL	QLeaf_n_a	---	QTL	P1	--	--
P	1	107,1	-----	-----	PM	*QaBwt_1	ConsMap	---	--	--	--
P	1	108,7	scaffold031	PDK_30s762641	PM	*mEgCIR04	SA569map	---	--	--	--
997	1	110,2	scaffold046	PDK_30s806071	M	mEgCIR384	mEgCIR384	SSR	P1	272	2
Q	1	110,2	-----		QTL	QI_a	---	QTL	P1	--	--
Q	1	114,0	-----		QTL	qFn_e	---	QTL	P2	--	--
P	1	114,1	scaffold013	PDK_30s780861	PM	*mEgEST00	SA569map	---	--	--	--
998	1	114,5			M	401/9	EgAg/MTgT	AFLP	P1	401	9
A	1	114,5			A	310/3	EAAg/MCTT	AFLP	C	310	3
Q	1	114,5	-----		QTL	Q%POP_a	---	QTL	P1	--	--
Q	1	114,5	-----		QTL	QIER_a	---	QTL	P1	--	--
Q	1	114,5	-----		QTL	QPO3_5_a	---	QTL	P1	--	--
P	1	115,8	-----	-----	PM	*QFwt_c	ConsMap	---	--	--	--
P	1	117,1			PM	*IER_1C	SA569map	---	--	--	--
P	1	117,1	-----	-----	PM	*Q%POP_1	ConsMap	---	--	--	--
P	1	117,9	NONE	NONE	PM	*sEgOPGP0	SA569map	---	--	--	--
999	1	117,9	scaffold872	NONE		mEgCIR339	mEgCIR339	SSR	C	193	1
P	1	121,5	scaffold034	PDK_30s1147901	PM	*mEgCIR08	SA569map	---	--	--	--
P	1	121,5	scaffold140	NONE	PM	*sEgOPGP0	SA569map	---	--	--	--
P	1	122,1	-----	-----	PM	*Q%PF_1C	ConsMap	---	--	--	--
Q	1	123,7	-----		QTL	q%KF_g	---	QTL	P2	--	--

Figura 8: Extracto de la proyección del mapa funcional de referencia basado en la población LM2T x DA10D. Cómo se observa en la imagen los "sacffolds" de la palmera datilera (PDK) se han podido integrar en el mapa gracias a los diferentes marcadores secuenciados como SSRs (mEgCIR) y AFLP mostrados en la imagen. Leyenda: RN= Nº de registro; CHR= cromosoma; POS= posición (cM); OP SCAF= scaffold de palmera de aceite africana; PD_SCAF= scaffold de la palmera datilera; DT= ; M- Name= nombre del marcador; MT= tipo de marcador; DES=; Pr=; Fr=:

Con la publicación del genoma de *E. guineesis* Jacq. en 2013a por Singh y col., se obtuvieron las pseudomoléculas, representantes de los 16 cromosomas de la especie, que se anclaron y se alinearon al mapa anterior, como puede observarse en la figura 8. Por último, este mapa se saturó integrando un mapa del transcriptoma basado en marcadores cDNA-AFLP, un mapa de consenso de diferentes cruzamientos que incluyen QTL relacionados con caracteres de productividad (Billotte y col., 2010), un mapa de ligamiento con nuevos QTL relacionados con calidad de aceite en un población de mapeo interespecífica [(*E.oleifera* SA49D x *E. guineensis* LM2466P) x *E.guineensis* PO3228D] (Montoya y col., 2013) y, por último un mapa de ligamiento consensuado para un QTL relacionado con la altura de tallo (Lee y col., 2015). En la figura 8 se muestra un ejemplo del mapa resultante con los "scaffold"s resultantes de palmera datilera y de aceite, junto con los marcadores utilizados para anclar los primeros "scaffolds" en el mapa de referencia de la población. Esta herramienta facilita la búsqueda de genes

candidato ya que puede proyectarse cualquier "EST" o cDNA en estas secuencias y predecir dónde se localizan los genes, tal y como se muestra en el esquema de la figura 9.

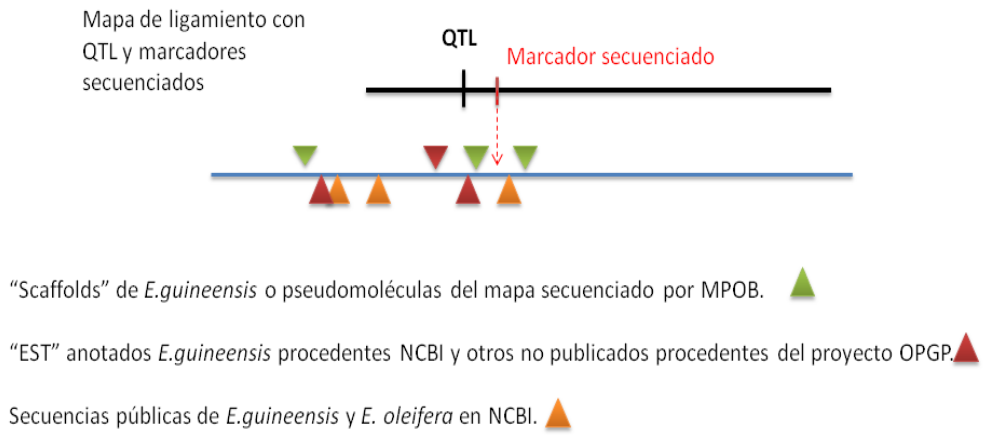


Figura 9: Esquema del análisis de co-localización de genes candidato. La creación del mapa funcional de alta densidad se realizó a partir del anclaje al mapa funcional de referencia de las "pseudomoléculas" procedentes de secuenciación del genoma por el MPOB, los "EST "y las secuencias públicas y no públicas procedentes del proyecto OPGP, y por último la búsqueda de genes candidato con un sentido biológico relacionado con el QTL de interés.

3.2.2. Búsqueda de Genes Candidato

Cuando un marcador se encuentra ligado a la misma posición o próximo a un QTL de interés puede ser un gen candidato que explica el QTL, sobre todo si este marcador es un gen con un significado biológico relacionado con el mismo. Asumiendo esta afirmación la búsqueda se realizó en las regiones QTL de interés relacionados con los caracteres de interés agronómico de nuestra población. Se analizaron las anotaciones de los marcadores que se encontraban a una distancia ± 1cM, minimizando los posibles eventos recombinatorios ("Linkage drag") (Collard y Mackill, 2008). Algunos de los genes colocalizados poseían regiones polimórficas, SNPs, por lo que fueron prioritarios en su selección para su posterior evaluación. Para esta búsqueda en el mapa funcional de alta densidad se utilizó un software PHOENIX (E.Ritter, no publicado) que recoge todos los datos disponibles, públicos o no, como se muestran en las figuras 10a y 10b.

Una vez obtenidas las secuencias, se buscó en las diferentes bases bibliográficas la literatura relacionada con el posible gen co-localizado para referenciar su funcionalidad. A continuación, se buscaron posibles homologías alineando mediante BLASTn con un E-value mínimo de 1e-35 contra la librería de Palmera de Aceite no redundante, y contra cromosomas de PD (*Phoenix dactylifera*) y de OP (Oil Palm). Asi se pudo comprobar la colinealidad de las secuencias con *P. dactylifera*, y además poder observar la posible presencia de intrones. Posteriormente las secuencias funcionales fueron analizadas mediante el software B2GO (Conesa y col., 2005).

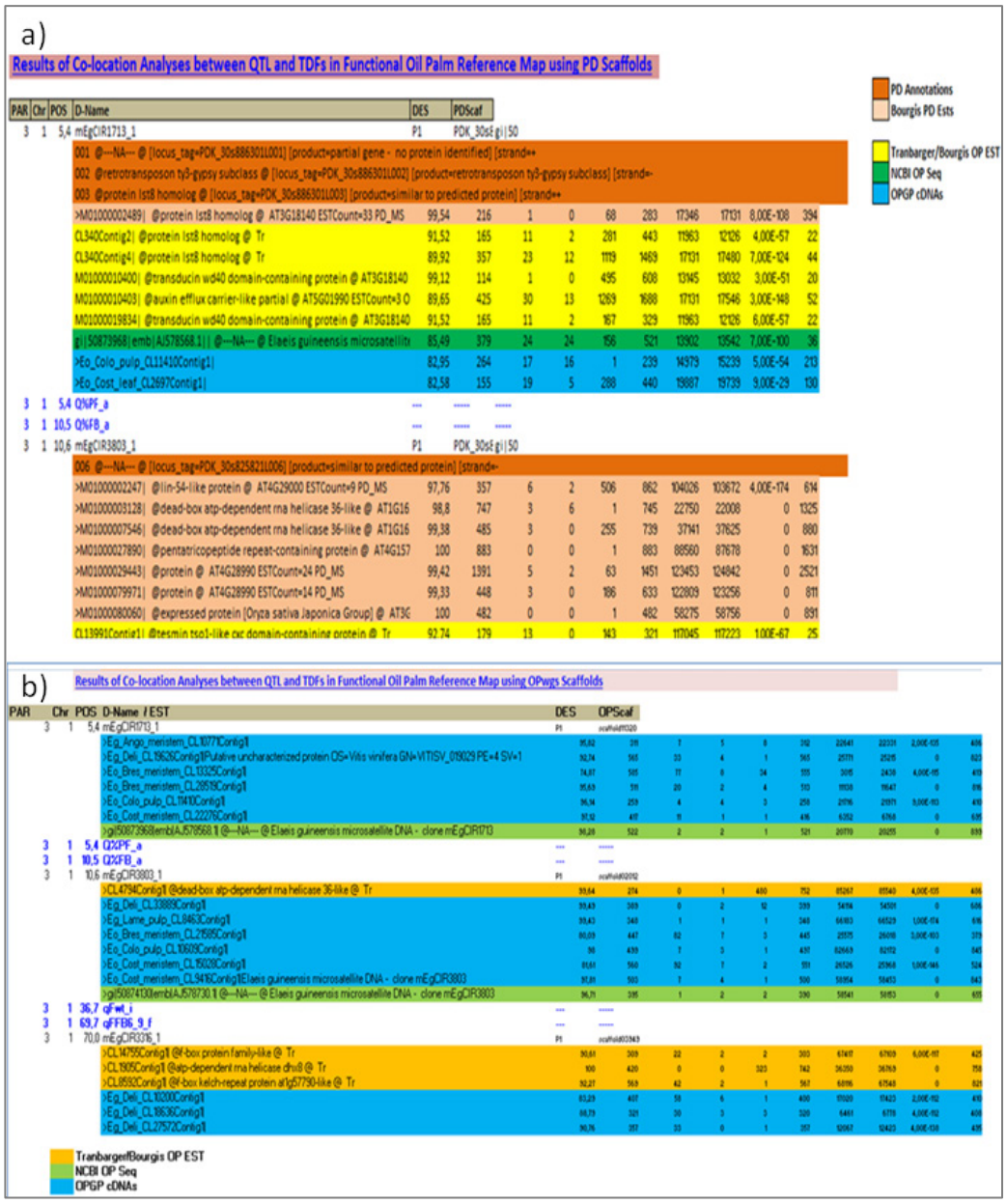


Figura 10: a) Extracto del análisis de co-localización en el mapa funcional de referencia a partir de los scaffolds de palmera datilera. b) Extracto del análisis de co-localización en el mapa de referencia a partir de los "scaffolds" de palmera de aceite africana. En ambas imágenes se muestran las anotaciones de los genes y marcadores que flanquean a los diferentes QTL y sus posiciones. Los orígenes de las secuencias de los genes se muestran en ambas imágenes en diferente color. Leyenda: PAR = parental; Chr = cromosoma; POS= posición (cM); D-Name/EST= anotación y/o nombre del marcador; DES= descendiente; PDK/OP scaf= "scaffold" de palmera datilera (PDK) o palmera de aceite africana (OP); PD Annotations= anotaciones de secuencias de palmera datilera; Bourgis PD EST= anotaciones de EST's secuenciados por Bourgis y col. (2011) de la palmera datilera; Tranberger/Bourgis OP EST's= anotaciones de los EST secuenciados por Bourgis y col (2011) y por Tranberger y col. (); NCBI OP seq= anotaciones de las secuencias publicadas en la página web de NCBI en palmera de aceite; OPGP cDNA= anotaciones de las secuencias correspondientes a las librerías de cDNA procedentes del proyecto OPGP (datos no públicos).

La configuración en el análisis de este software fue en primer lugar BLASTx con e-value mínimo de 1e-30 y un máximo "Hits" o secuencias homólogas de 20. Una vez realizado este paso y revisado las anotaciones obtenidas se procedió al “Mapping “ para asociar los resultados obtenidos en el alineamiento con la terminología ontológica, y por último al “GO Annotation” para conocer su anotación

funcional, y comprobar la posible relación, establecida mediante la bibliografía, de la funcionalidad de la secuencia con el carácter sujeto a estudio.

3.3. Búsqueda de genes candidato conocidos

Los genes candidato se obtuvieron a partir de diferentes búsquedas en bases de datos como NCBI, KEGG , patentes (WIPO -https://patentscope.wipo.int/search/en/search.jsf-, EUROPATENT -https://www.epo.org/searching/free/espacenet.html-) o a partir de publicaciones (PUBMED o WOS) para buscar evidencias experimentales sobre sus funciones.

Las búsquedas se hicieron en base a las características agronómicas deseadas para el proceso de mejora del cultivo, por lo que se relacionaron con las rutas biosintéticas de aceite, ácidos grasos, o triglicéridos, otros genes participantes en el desarrollo y crecimiento, o genes que actúan como factores de transcripción, entre otros. Se priorizó la búsqueda en la misma especie, o género. Pero también se buscaron genes candidato en especies modelo como *Arabidopsis thaliana* o relacionada genéticamente como *Phoenix dactylifera*.

Estos genes clasificados como candidatos se agruparon en diferentes bases de datos, para una mayor evaluación de su funcionalidad. Las secuencias se analizaron con el software Blast2GO, tal y como se describe en el apartado anterior. En estos hits se buscaron las secuencias con similitudes a *Elaeis guineensis*. A continuación se procedió a buscar las anotaciones de todas las secuencias, mediante "Mapping" en B2GO para corroborar su funcionalidad en relación al carácter de interés y finalmente se procesó con "GO Annotation" y "GO Slim" este último en plantas, acotando así las funcionalidades.

Una vez obtenidos estos datos, se utilizó el software PHOENIX donde se encuentra también el mapa de consenso utilizado, para buscar homologías mediante BLASTn en nuestras bases de datos, tal y como se describe en el caso de la búsqueda de genes co-localizados. El objetivo de estos alineamientos es situar los genes encontrados como posibles candidatos en el mapa, y buscar los mismos en nuestra especie.

4. RESULTADOS

4.1. Análisis deltranscriptoma mediante la técnica cDNA-AFLP

4.1.1. Obtención de fragmentos derivados del transcriptoma (TDF)

En cada carácter se estudiaron las familias correspondientes desarrolladas en el anexo 1 (tabla1.1) mezclando los 4 genotipos caracterizados como “buenos” y los 4 “malos” para cada una de las familias como se describe en el apartado de materiales. En la tabla 17 se muestra el resultado de los fragmentos obtenidos derivados del transcriptoma para cada carácter. Se realizaron un total de 640 combinaciones de cebadores de amplificación específica, con 220 combinaciones de cebadores diferentes, aplicando a cada carácter sujeto a estudio entre 40 y 175 combinaciones de cebadores. El rendimiento de las combinaciones de cebadores de amplificación específica fue inferior a la unidad

(106/220=0.48 TDF/CC). Del total de combinaciones, 106 presentaron bandas polimórficas, siendo los caracteres con mayor número de polimorfismos CPO (17),y HT (17) en relación al número del total de combinaciones realizadas.

Tabla 17: Resultados obtenidos de fragmentos derivados del transcriptoma (TDF) para cada carácter de interés y número de CC utilizados. CC Pol: comobinaciones de cebadores polimórficas; CC No Pol: combinaciones de cebadores no polimórifcas.

Carácter	BN	BW	CPO	FW	MF	HT	OM	FN	TOTAL
CC Pol	5	13	17	15	12	17	12	15	**106**
CC No Pol	35	37	38	47	59	46	163	109	**534**
Total CC	**40**	**50**	**55**	**62**	**71**	**63**	**175**	**124**	**640**

Estos fragmentos polimórficos se clasificaron como **positivos** (+) cuando estaban presentes en todas las familias; **positivo/negativo** (+/-) cuando estaban presentes en todas las familias menos en una; y, como **otros** cuando estaban presentes en 2 familias o menos. En la tabla 18, se presentan el número obtenido de los diferentes TDF en cada carácter. El carácter con mayor número de **TDF** positivos fue **HT,** en el cual todos los fragmentos polimórficos estaban presentes en las cuatros familias. Para **BW** y **MF** la mayoría de los TDF obtenidos, correspondían a las cuatro familias (6 y 9 TDF respectivamente). En cambio, en los caracteres **BN**, **FW**, **OM** y **FN** la mayoría de sus transcriptos fueron de la segunda categoría, es decir, estaban presentes en tres de las cuatro familias. Todas las familias, excepto HT, presentaron polimorfismos en dos familias o menos, siendo el carácter con mayor abundancia de estos transcriptos **CPO**. Los resultados correspondientes a las combinaciones de cebadores polimóficas por cada carácter, así como la altura en pares de bases de estos fragmentos se muestran en el anexo 1(Tabla1.2). Puede observarse que los tamaños de banda obtenidos variaron entre 50pb y 800pb, y que las combinaciones de cebadores más efectivas fueron A19T13 y A16T13 con 8 y 6 fragmentos, respectivamente.

Tabla 18: Clasificación y número de TDF para cada carácter

	BN	BW	CPO	FW	MF	HT	OM	FN	TOTAL
TDF +	1	6	1	5	9	17	2	2	**43**
TDF +/-	3	5	3	8	2	0	8	9	**38**
TDF OTROS	1	2	13	2	1	0	2	4	**25**

4.1.2. Análisis de las secuencias

En total 92 amplicones fueron secuenciados y analizados para su posible utilización como genes candidatos, de los que se obtuvieron 56 secuencias válidas nombradas como CDA (Anexo 1; Tabla 1.2). Para el análisis de las secuencias se utilizaron bases de datos locales obtenidas a partir de las secuencias de los cromosomas de palmera de aceite (MPOB), y una librería de secuencias no redundantes de palmera de aceite con secuencias obtenidas en el proyecto OPGP a partir de una colección de EST's

realizados en hoja, pulpa, meristemos y raíz de palmeras *E.guineensis* y *E. oleifera* con diferentes origenes. También, se buscaron homologías de las secuencias mediante el software en su versión libre B2GO, mediante el algortimo BLASTx, y anotaciones funcionales mediante GO y "Plant Slim" y así poder relacionar "in silico" el carácter buscado con un posible sentido biológico.

4.1.2.1. Bases de datos locales

Del total de los 56 transcriptos, 20 obtuvieron homologías superiores a 1e-50 en diferentes grupos de ligamiento (GL), con un e-valor medio de 3,64e-57 (Tabla 19). Estos secuencias de expresión diferencial se situaron en los diferentes cromosomas o grupos de ligamiento a lo largo del genoma de la palmera de aceite, siendo el GL 12 quién presentó una presencia más repetitiva en los fragmentos analizados (CDA43, CDA8, CDA32, CDA44). Los caracteres donde proporcionalmente al número de transcriptos obtenidos se encontraron un mayor número de homologías fueron BN (67%) y CPO (50%)(resultados no mostrados).

Los resultados de las homologías en la librería de secuencias no redundantes de palmera de aceite (Tabla 19) mostraron 8 transcriptos con un e-valor medio 2,55e-58 (no mostrado)..

Tabla 19: Resultados de la búsqueda de homologías mediante Blastn en las bases de datos locales. GL: grupo de ligamiento o cromosoma. Librería NR: librería de secuencias no redundantes en palmera de aceite.

NOMBRE TRANSCRIPTO	CARÁCTER	GL	E-valor	Libreria NR Palmera de Aceite	E-valor
CDA43	BN	12	8e-62	-	
CDA8	BN	12	8e-62	-	
CDA18	BW	7	3e-71	-	
CDA20	BW	6	0.0	-	
CDA27	BW	-		CL1Contig5969| @ "(at3g22142 : 117.0)Codifica una proteasa inhibidora/ almacenamiento en semilla de la familia LTP	4E-59
CDA7	BW	14	4e-144	-	
CDA3	CPO	1	0.0	-	
CDA4	CPO	5	7e-124	-	
CDA42	CPO	2	2e-62	-	
CDA44	CPO	12	4e-67	gi|191204475|gb|EY407406.1|EY407406| NR @ "(loc_os12g27096.1 : 144.0) Relacionado con un poliproteína POL de transposon TNT 1-94 (*Nicotiana tabacum*)	3e-105
CDA5	CPO	4	4E-	-	

			62		
CDA9	CPO	11	2E-71	-	
CDA90	FN	-		Contig2857_S_C2f	2e-134
CDA92	FN	1	2e-95	-	
CDA24	FW	4	2e-71	Eg_Deli_CL35037Contig1	2e-57
CDA26	FW	-		Contig2793_S_C2f	5e-100
CDA31	FW	2	5e-67	CL1Contig1267| @ "(p33278|sui1_orysa : 215.0) Homólogo a factor de traduccion de proteína SUI1 (*Oryza sativa*)	9e-61
CDA32	FW	12	7e-64	-	
CDA40	HT	11	2e-73	-	
CDA41	HT	14	2e-59	-	
CDA33	MF	6	5e-151	-	
CDA37	MF	3	4e-79	-	
CDA75	OM	4	8e-56	-	
CDA76	OM	5	7e-151	CL1Contig209| @ "(loc_os09g28810.1 : 437.0) (at1g47500 : 363.0) Proteína unión RNA 47 (RBP47C)	e-150
CDA78	OM	16	3e-78	Contig3707_S_EoP3-P2	6e-67

De estos transcriptos 4 presentaron homologías con anotaciones funcionales en su descripción, CDA27 para peso de racimo (BW), CDA44 para cantidad de aceite crudo (CPO), CDA31 para peso de fruto (FW) y CDA76 para proporción de aceite en el mesocarpo en relación al tamaño del fruto (MF). El resto de los trancsriptos pertenecían a diferentes "contigs" o fragmentos de ADN superpuestos que forman una región de consenso (www.genome.gov) para la construcción de mapas físicos. En este caso los caracteres dónde se obtuvieron mayor número de homologías fueron Fw (18%) y Bw (11%) (resultados no mostrados).

4.1.2.2. Bases de datos públicas (B2GO)

El análisis mediante el software Blast2Go permitió conocer las homologías de los transcriptos con otras especies de plantas y los procesos en los que estaban implicados mediante su anotación con "GO onthology" y "Plant Slim"(Tabla 20).

Sólo se encontraron **homologías funcionales** para 4 de los 9 caracteres sujetos a estudio - Peso de Racimo (**BW**), contenido de aceite crudo (**CPO**), peso medio de fruto (**FW**) y relación del tamaño del mesocarpio frente al tamaño de fruto(**MF**). El resto de caracteres no mostraron posibles anotaciones funcionales como en altura del tallo (**HI**),o en relación de contenido de aceite frente al peso del mesocarpo (OM), o no hubo ningún resultado mediante los algoritmos Blastx y Blastn.

Las anotaciones funcionales para estos transcriptos se clasificaron en función de "Gene Ontology Cosortium" (www.geneontolgy.org), y en algunos casos pudo aplicarse "GO-Slim" (www.geneontolgy.org) centrando el análisis en plantas, a partir de la base de datos del genoma de Arabidopsis (Figura 11). Estos resultados se resumen en tres categorías diferentes, aunque pueden estar relacionadas entre sí, componente celular (C), proceso biológico (P) y función molecular(F).

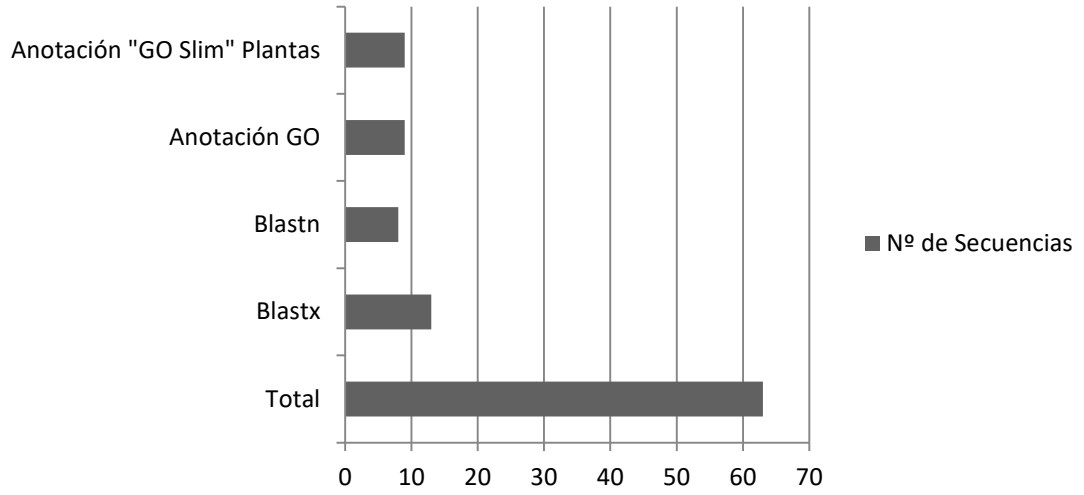


Figura 11: Gráfico de los resultados obtenidos en B2Go mediante BLASTx y BLASTn, con un e-value inferior 1e-10.En total se analizaron 63 amplicones de los que mostraron homología 23 de ellos y anotación funcional en 9 de ellos.

En el carácter **BW** se encontraron 2 transcriptos (CDA13 y CDA18) que mostraba homología funcional en dos de las tres categorías (P y F), y además pudieron mapearse mediante "GO-Slim" basado en plantas. La secuencia CDA 27 cuyo resultado fue una proteína no caracterizada para la misma especie no se mapeo mediante GO, y el transcripto CDA6 mostró homología cuando se aplicó el algoritmo BLASTn con una secuencia de ARN ribosómico en *Rhabdodendron amazonicum.*

El carácter **CPO** obtuvo como resultado 3 transcriptos con homologia (CDA4 ,CDA5, CDA44), y todas ellas tuvieron como resultado anotaciones funcionales. CDA 4 se caracterizó como un enzima metiltransferasa, y CDA5 tuvo como resultado de BLASTx un gen relacionado con el enzima dimetil triptofano sintasa en *Aspergillus*, el cual participa en diferentes procesos metabólicos relacionados con el metabolismo secundario y el de proteínas. Además también participa en la biosíntesis de diferentes compuestos orgánicos cíclicos. Por otro lado CDA44 relacionado con un gen denominado *gag-pop fusion* mostró posibles funciones en plantas relacionadas con la unión de compuestos.

Tres de las secuencias pertenecientes al carácter **FW** (CDA22, CDA26 y CDA31) tuvieron homología funcional, aunque CDA 26 únicamente como componente celular en la mitocondria, durante el proceso de respiración celular y CDA31 como proteína localizada en los plastidios. CDA23 obtuvo homología mediante BLASTn caracterizando la secuencia como una proteína de 26kda en *Cucumis sativus*.

El último carácter que muestra un transcripto con homología funcional (CDA74) es **OM**. El resultado de BLASTx fue una enzima kinasa dependiente de ciclina en una especie de café (*Coffea canephora*) cuyas anotaciones funcionales permiten dilucidar su participación por ejemplo en la división celular.

Tabla 20: Resultados del análisis de TDF mediante B2GO.

NOMBRE TDF	CARÁCTER	DESCRIPCIÓN BLASTx	BLASTn	E-VALOR	ANOTACIÓN GO
CDA6	BW	-	*Rhabdodendron amazonicum* Secuencia parcial ARN ribosómico 26s	2,2E-55	-
CDA13	BW	Parcial péptido no ribosómico		1,6e-18	F:actividad catalítica; F:unión
CDA15	BW		*Phoenix dactylifera* genoma completo	2,6E-95	
CDA18	BW	Similar a receptor de proteína kinasa (Rica en cisteína-rlk 8)		.1E-20	P:proceso metabólico; F:unión; P:proceso celular
CDA27	BW	Proteína no caracterizada LOC101511867		31e-14	
CDA4	CPO	Posible metiltransferasa pmt11		1,22E-23	F:actividad transferasa
CDA5	CPO	dimetialilo triptófano		2,1E-20	P:proceso de metabolismo de proteínas; P:proceso biosintético; F:actividad transferasa; P:proceso celular
CDA44	CPO	Proteína de fusión "gag-pol"		6,3e-13	F:unión
CDA14	FN		*Phoenix dactylifera* RNAm variante transcripta no caracterizada loc103704827	1,7E-28	

CDA90	FN	Subunidad beta de ATP sintasa		58E-27	
CDA22	FW	Isoforma 1 helicasa 39 ARN"Dead-Box" ATP dependiente		2,1e-12	F:unión de ácidos nucléicos; F:unión de nucleótidos; F:actividad hidrolasa
CDA23	FW		*Cucumis sativus* cds completo: proteína 26 kda de floema	4,7E-27	-
CDA26	FW	Proteína hipotética (mitocondria)		58E-32	C:mitocondria
CDA31	FW	Proteína hipotética JCGZ_00471		2,7E-29	C:plastidio
CDA39	HT		*Ttriphyophyllum peltatum* Secuencia completa ARN ribosómico 26s	6,4E-18	-
CDA34	MF	Proteína no caracterizada		8,3e-14	-
CDA37	MF	Quinasa f-1 dependiente de ciclina		1,7e-12	P:desarrollo post-embrionario; P:modificación de proteínas celulares process; P:morfogenésis de estructura anatómica; P:proceso biosintético; P: proceso metabólico ADN; P:ciclo celular; F:unión de nucleótidos; C:núcleo; C:citosol; F:actividad kinasa; F:actividad como enzima regulador
CDA75	OM		*Elaeis guineensis* Variante transcripta de ARNm de proteína bifuncional de degradación de ácido úrico	4,2E-47	
CDA76	OM	Proteína asociada a la senecencia		36E-45	

CDA78	OM		*Phoenix dactylifera* Variante transcripta no caracterizada loc103718754	2,8E-53	
CDA79	OM		*Ooryza punctata* ARN ribosómico 18s, transcrito ARN interno ribosómico espaciador y secuencia completa ARN ribosómico 26s	7,6E-28	

.

4.2. Análisis de secuencias co-localizadas

Tabla 21: Caracteres de interés sujetos a estudio, y número de genes identificados mediante analísis de co-localización y de genes conocidos seleccionados a partir de diferentes referencias como la búsqueda de literatura relacionada y/o patentes, y su hipotética relación con los caracteres de interés. Aquellos genes que pueden estar realcionados con más de un carácter se han contado para carácter, a pesar de ser el mismo gen*..

			GC CO-LOCALIZADOS	GC CONOCIDOS*	TOTAL
CARACTERES DE INTERÉS	PRODUCCIÓN	BN	13	2	**15**
		FFB	17	1	**18**
		PO	11	44	**55**
		BW	6	1	**7**
	COMPONENTES	POP	8	37	**45**
		FW	2	28	**30**
		I	2	8	**10**
		PF	1	26	**27**
	VEGETATIVOS	HI	21	22	**43**
		OTROS	3	17	**20**
		TOTAL	**84**	**186**	**270**

Los genes candidato se buscaron en las regiones próximas a los QTL asociados con los caracteres de interés para la población sujeta a estudio a lo largo de los 16 GL. El número de secuencias denominadas como posibles genes candidatos relacionadas con los **caracteres** de interés relacionados con la **producción** (Anexo 3; Tabla 3.1) fueron **47** secuencias, de las cuales **13** se relacionaron con número de racimos (**BN**). En la tabla 3.1 (Anexo 3) se muestran las secuencias que estaban co-localizadas con QTL relacionados con el carácter como QBn3_5 o QBn6_9. Estos QTL están presentes en los grupos de ligamiento 2, 3, 4, 5, 7, 8, 9, 10, 12, 13 y 15. Hay que destacar también que en el GL7 a una altura de 48,8cM hay dos QTL's colocalizados uno relacionado con BN, QBN6_9, y otro relacionado con Bw, QBWt3_5. Para el peso de los racimos de fruta fresca (**FFB**) fueron **17** las secuencias co-localizadas con los QTL's relacionados. Cómo puede revisarse en el anexo 3 (Tabla 3.1), los QTL's se encontraban en 6 de los 16 cromosomas de la especie, siendo los cromosomas 2 y 11 los que más QTL's relacionados tenían (QFFB3_5 y QFFB6_9). En el cromosoma 2 son 2 los QTL's relacionados con FFB a 107,1cM, y los genes co-localizados que se encontraron están por encima de los mismos. En el cromosoma 6 se detectaron dos genes candidatos que flanquean al QTL situado a 73,5cM por los dos extremos. En relación al contenido de aceite (**PO**) fueron **11** las secuencias co-localizadas. En el anexo 3 (Tabla 3.1) se muestran los QTL's relacionados con este carácter, QPO3_5 y QPO6_9. Estos QTL's estaban presentes en 4 de los 16 grupos de ligamiento, en el GL1 se co-localizaron tres posibles genes candidatos con el QTL situado a 92,9cM, en el GL3 se encontraron dos QTLs relacionados con el carácter, uno a 12,9cM en el que se co-localizaron 2 genes que lo flanqueaban por ambas regiones, y otro a una distancia de 51,9cM

con un único gen co-localizado. Por último, el GL15, dónde se situaba el QTL a una distancia de 78cM, se co-localizaron 4 genes candidato, flanqueando 2 a 2 por su parte superior e inferior. Seis secuencias fueron co-localizadas con peso de rácimo (**BW**) de los cuáles tres estaban co-localizados también con un QTL de BN, dos se co-localizaron en QTL's QBwt6_9 en los GL 2 y 7 y cuyas posiciones pueden revisarse en el la tabla 3.1(Anexo 3), y el último gen se co-localizó con 4 QTl's en la misma posición (125cM), 3 de los cuales pertenecían a Bw y otro a FW, en el cromosoma 5.

Como puede revisarse en el anexo 3 (Tabla 3.2) los **caracteres** relacionados con **componentes de racimo** obtuvieron **13** posibles genes candidatos, de los cuales **8** correspondieron al % de aceite obtenido en relación al tamaño de la pulpa (**POP**), los QTL relacionados en 3 de los 16 GL, posicionandose 2 QTL's en el GL3 a una distancia de10,3cM y12,9cM. Este último presentaba en la misma posición otro QTL relacionado con PO, lo mismo sucede en el GL7. Los genes candidatos se encontraban flanqueando los QTL's por ambas partes en 3 de los 4 QTL's. Siendo el QTL posicionado en el GL7 el que presentaba los dos genes candidatos co-localizados por su parte inferior. En cuanto al índice de iodo (**IV**), indicador de la calidad del aceite obtenido, fueron **3** los co-localizados también con QTL's de diferentes GLs (GL1, GL3 y GL13). Como puede revisarse en la tabla 3.2 (Anexo 3) el QTL qI-j situado en el GL13 se encuentra también co-localizado con un QTL relacionado con componentes vegetativos de la planta, por lo que la funcionalidad del gen candidato KG12 puede estar relacionado con ambos QTL y en consecuencia con sus caracteres , el cual también es un indicador de la calidad del aceite obtenido. **Dos** fueron las secuencias relacionadas con el peso del fruto (**FW**). Estas se co-localizaron con QTL's situados en diferentes GL, siendo el más destacable el QTL QFwt_1 posicionado en el GL5 (125cM) y donde se posicionaban 3 QTL's más relacionados con QBw. y **una** única secuencia al % de pulpa en relación al fruto (**PF**).

Por útimo, el objetivo de la búsqueda de los **caracteres vegetativos** (Anexo 3; Tabla 3.3) se centró en el **tamaño del tallo de la palmera**, para el cual se encontraron **21** posibles genes candidato en diferentes QTL relacionados con el mismo como (St-Gr o Crecimiento de tallo). Estos genes candidato se co-localizaron en QTL's presentes en 9 de los 16 grupos de ligamiento del mapa funcional, siendo el GL 2 el que posee mayor número de QTL's relacionados con el carácter. Como se muestra en el anexo 3; (Tabla 3.3), en el QTL Qht_1C se co-localizaron 3 genes candidatos que flanqueaban al QTL por su parte superior e inferior, 2 de los cuales estaban en la misma posición (93,8cM). En el GL2 el QTL Q-Cs-t posicionado a 136.6cM se co-localizaron dos genes, KG166 y KG167, en la misma posición (136,6cM). Los genes co-localizados KG170, KG147 y KG171 flanquean al QTL QHt_d por ambas partes, con una distancia máxima de -0,1 y 0,8cM, respectivamente. El QTL QHt_e del cromosoma 12 se co-posiciona con otro QTL relacionado también con un carácter vegetativo qP_W_f, y del que únicamente se selecciona un gen candidato que lo flanquea por su parte superior. También se encontraron otras tres secuencias relacionadas con otros caracteres vegetativos de interés, como anchura de hoja y anchura media del peciolo.

En total se seleccionaron **84** secuencias como posibles genes candidatos colocalizadas a una distancia del QTL de interés de ± 1 cM y con una distancia media en pares de bases de ± 51519pb. En la figura 12 se muestra el mapa con los 16GL y dónde se posicionan los genes candidatos co-localizados con los QTLs de interés.

4.2.1. Determinación de la funcionalidad de las secuencias

Caracteres relacionados con la producción (Anexo 4; Tabla 4.1)

Las anotaciones de las secuencias fueron confirmadas por la función obtenida mediante BLASTtx en 11 de las 13 secuencias seleccionadas.

Las funciones y procesos biológicos más relevantes que se obtuvieron para **BN** fueron genes candidatos que participan en el desarrollo del embrión (KG201, TEST), diferenciación celular y en el proceso de reproducción (KG204). Una de las secuencias se caracterizó también con un factor de transcripción MADS-Box, encargado de regular diferentes funciones en plantas, pero sobre todo en el desarrollo floral (Adam y col., 2007).

En el caso de **FFB,** fueron 14 las secuencias que coincidían con la función anotada, aunque en el caso KG192, la anotación GO fue similar para las dos. Las funciones moleculares y procesos biológicos en los que participan estos genes candidatos están relacionados en su mayoría con el metabolismo de lípidos (KG181, KG186, KG187 y KG188). El resto de funciones y procesos se relacionan con el metabolismo secundario (KG179), diferentes procesos enzimáticos y de unión (KG183, KG184, KG185, KG186, KG194, KG195), y procesos en los que participan ácidos nucleicos (KG189, KG192).

Para PO las anotaciones de las secuencias coincidían con las funciones obtenidas en *E.guineensis* Jacq. mediante blastx en 9 de las 11 secuencias calificadas como posibles genes candidatos. Los dos genes candidatos para los que las anotaciones de las secuencias que no coincidían fueron KG257 y KG261, pero para KG257 las anotaciones GO sí correspondian con la funión previa detectada en la anotación de la secuencia, caracterizada como un transportador de lípidos en su anotación inicial. Los genes candidatos (KG256, KG257, KG258, KG260) mostraron funciones relacionadas con el metabolismo de lípidos en las anotaciones GO y el resto se caracterizaron también en procesos enzimáticos y de unión (KG262, KG263 y KG264), y por último, KG258 cumple una función como factor de transcripción y en la unión de ADN.

BW fue el último carácter para el que se buscaron secuencias co-localizadas en el mapa integrado. Las anotaciones de los posibles genes candidatos coinciden en todos los casos con las funciones obtenidas en BLASTn en la especie. Las funciones moleculares y los diferentes procesos biológicos encontrados mediante anotaciones GO en plantas reflejaron que estos genes candidatos participan en el metabolismo de carbohidratos (KG140 y KG143), en la transducción de señales (KG142) y en diferentes procesos de unión de proteínas, entre otros (KG141 y KG146).

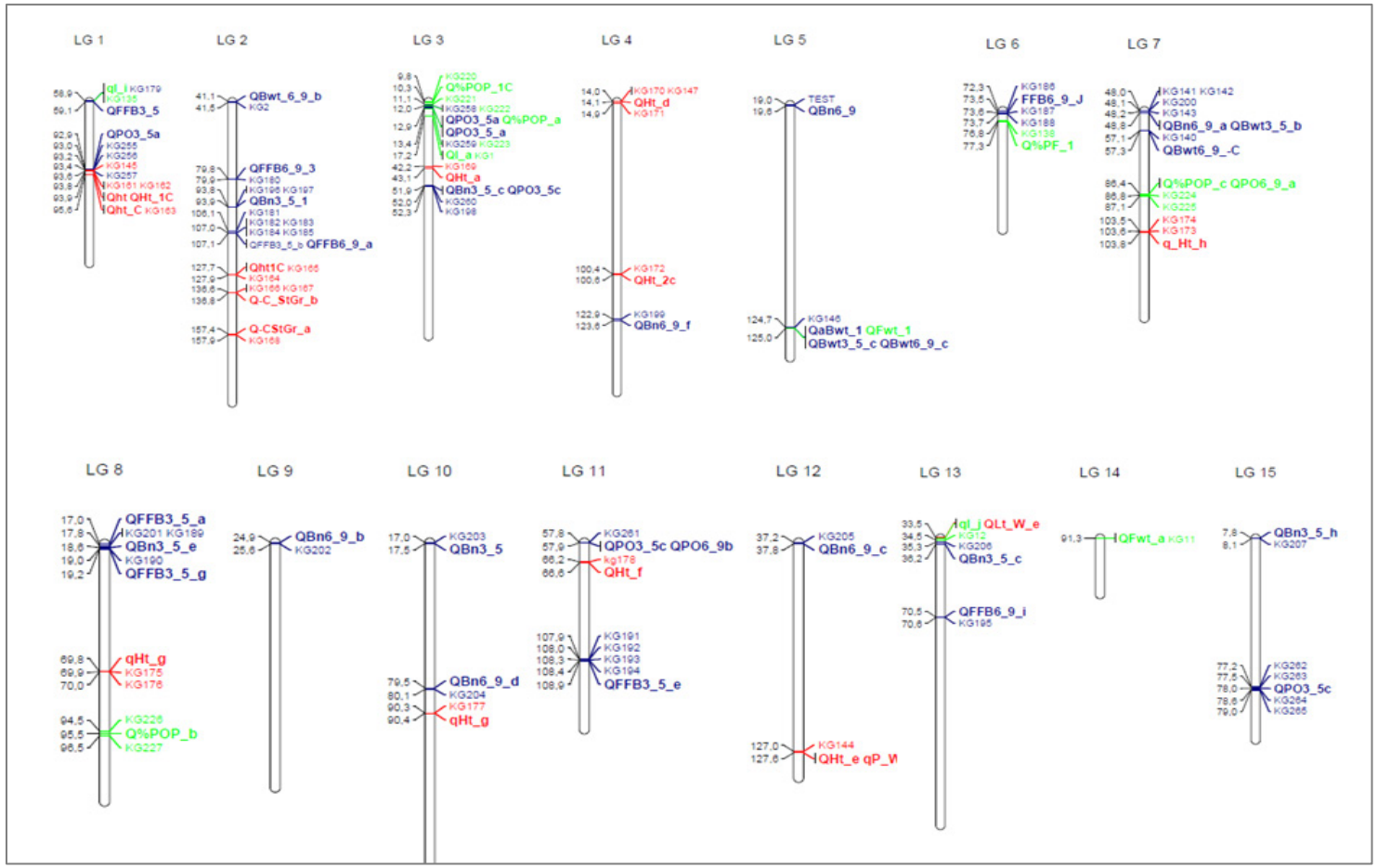


Figura 12: Imagen de los GC co-localizados en los diferentes grupos de ligamiento junto con sus QTLs. El color azul muestra el nombre de los genes candidatos relacionados con los caracteres de producción y su posición en cM. En negrita se muestra el nombre del QTL y su posición. El color verde muestra el nombre de los genes candidatos relacionados con componentes de rácimo y su posición en cM. En negrita verde muestra el nombre del QTL y su posición. En color rojo se muestran los nombres de los genes candidatos relacionados con caracteres vegetativos y su posición en cM. El negrita rojo se muestran el nombre de los QTL y su posición.

Caracteres relacionados con componentes de racimo (Anexo 4; Tabla 4.2)

En el carácter POP todas las anotaciones previas de las secuencias seleccionadas como posibles genes candidatos coincidían con el análisis en BLASTn de las funciones para las que estaban caracterizadas. Las anotaciones mediante "Go Ontology" mostraron la participación de estos genes en diferentes procesos metabólicos relacionados con lípidos y carbohidratos (KG222), afianzaron la presencia de un factor de transcripción MADS BOX cuya importancia se ha destacado anteriormente (KG223). KG224 y KG225 presentan funciones enzimáticas y reguladoras, y por último KG226 parece participar en la generación de precursores de algunos metabolitos y de energía, así como también en procesos fotosintéticos. KG220 y KG221 no obtuvieron resultados para las anotaciones GO, pero la bibliografía encontrada parece referenciar para el primero de ellos una función reguladora (Wirtz y col., 2004) y para el segundo de ellos la bibliografía confirma su participación como enzima reguladora la glucolisis, y participante en el metabolismo primario de las plantas (Nielsen y col.,2004).

Las secuencias co-localizadas próximas a un QTL para peso del fruto (**FW**)y seleccionadas como posibles genes candidatos coincidieron en uno de los dos casos la anotación de la secuencia con la función obtenida mediante BLASTx y fue en KG146. Para KG11 las anotaciones fueron diferentes, pero ambas mostraron que participaban en algún proceso relacionado con el ARN. Esto último fue confirmado por las anotaciones procedentes de GO. KG146 participa en la regulación de la transcripción y en la replicación del ADN.

De los dos genes candidatos seleccionados para **IV**, uno de ellos (KG12) mostraba una proteína no caracterizada en *E.guineensis*, aunque la anotación GO la caracterizó como participante en un proceso metabólico y actividad enzimática. En el caso de KG135, las anotaciones fueron coincidentes, y GO mostró que forman parte del metabolismo secundario de las plantas.

El último gen candidato seleccionado para el carácter relacionado con el tamaño de la pulpa con respecto al fruto (**PF**) fue KG138. La anotación de la secuencia como proteína adiposa regulatoria no coincidió con el resultado de BLASTtx, que correspondía a una proteína no caracterizada en la especie. Por ello se lanzó contra todas las especies, y se obtuvo como resultado la misma proteína para *Medicago trunculata* con un E-value 1e-75.

Caracteres relacionados con componentes vegetativos (Anexo 4; Tabla 4.3)

En este grupo los colocalizados se buscaron principalmente en QTL relacionados con la **altura del tallo**, ya que uno de los objetivos es encontrar palmeras con menor altura para facilitar la recolección de los frutos. Los QTL relacionados con el tallo son Ht y StGr. De las 21 secuencias seleccionadas como posibles genes candidatos, sólo KG161 no coincide la anotación de la secuencia con la función obtenida mediante BLASTx en la misma especie. Se realizaron BLASTn y BLASTx contra el resto de especies mostrando el mismo resultado que para *E. guineensis* ,por lo que se acepto esta función como válida, y no la anotación inicial. Hay que destacar que tanto KG161 y KG162 no obtuvieron ningún

resultado con anotación GO por lo que la selección de la secuencia como posible GC se basa, únicamente, en las referencias bibliográficas.

Las funciones reveladas por las anotaciones procedentes de GO estuvieron relacionadas con la fotosíntesis (KG170), con genes relacionados con precursores de metabolitos y energía (KG165), metabolismo de carbohidratos (KG163, KG168, KG145), participación en el metabolismo secundario y además con actividad catabólica (KG173 y KG174). KG164 participa en el desarrollo de organismos multicelulares, y el resto de las secuencias seleccionadas participan de en reacciones enzimáticas y de unión como KG166, KG169, KG172 o como reguladores de la expresión génica (KG178).

4.3. Genes candidatos conocidos

El uso de la literatura, diferentes rutas biosintéticas y patentes públicas permitió seleccionar 119 posibles genes candidatos de diferentes especies como *E.guinnensis*, *E.oleifera*, *A. thaliana*, *V.vinifera*, *J.Curcas* y *O.europeae*, entre otras (Tabla 21). Estos genes candidato se relacionaron y agruparon con los caracteres de interés para la población de estudio en palmera de aceite mediante un BLASTx en *Elaeis guineensis*, y su posterior anotación con "Gene Onthology" en B2GO, dilucidando así su hipotética participación en el carácter de interés o caracteres de interés

En la tabla 22 se muestran algunos ejemplos de los genes seleccionados como candidatos .El gen candidato (P4) implicado en aumento de rendimiento se encontró en una patente publicada en 2006 por Reuzeau y col. Se relacionó con los caracteres de producción BN y FFB, y con los componentes FN y FW. Las patentes P39 (Puzio y col., 2008), P58 (Abdullah y Kulaveerasingam,2002) y KG269 (Sanz Molinero y col., 2009) relacionadas con la biomasa de la planta se relacionaron con caracteres de componentes vegetativos clasificados como otros. En este grupo de otros se englobaron también los genes implicados en el desarrollo, crecimiento, fotosíntesis y/o en respuesta a estrés. El enzima serina carboxipeptidasa (KG118) participante en el metabolismo secundario y un factor de transcripción relacionado con la síntesis de etileno (KG243) también se incluyeron en este grupo. Los genes candidatos que eran factores de transcripción MADS-Box, relacionados con el desarrollo floral, y aquellos implicados en el desarrollo del fruto se relacionaron con los caracteres relacionados con componentes FN y FW. Con el componente vegetativo HI se relacionaron todos aquellos genes que estaban relacionados con la elongación del tallo. Por último, los genes candidatos implicados en el metabolismo de lípidos (50) se relacionaron con el carácter de producción PO y en su clasificación de componentes con POP y/o I. Estos caracteres tambien fueron relacionados con KG275 -"Oleosin"- relacionado con la acumulación de aceite en la semilla y KG290 una proteína relacionada con la producción de aceite.

Algunos de estos genes candidatos se seleccionaron gracias a la publicación del genoma de la especie por Singh y col. en 2013a, y a la siguente aparición de publicaciones relacionadas con genes de interés para la especie en cuanto a su calidad y rendimiento. Estos genes son KG120, *Shell* (Singh y col.

2013b), KG271, *Vir* (Singh y col. 2014), KG210, *Lipasa* (Morcillo y col. 2013) y KG233 y KG234, *Asparagina sintasa* (Lee y col. 2015).

El resto de genes seleccionados como candidatos se muestran en el anexo 5, junto con sus posibles funciones, homologías y referencias bibliográficas. Entre las funciones moleculares detectadas (Figura 13) fueron mayoritarias las funciones relacionadas con la unión de compuestos heterocíclicos, de proteínas y de lípidos. También fueron relevantes las funciones de enzimas transferasas y factores de transcripción. En cuanto a los procesos biológicos (Figura 14) donde participan los procesos relacionados con diferentes metabolismos y procesos celulares son los más abundantes.

Tabla 22: Algunos de los genes seleccionados como candidatos a partir de bibliografía, rutas biosintéticas relacionadas con el metabolismo de lípidos, y patentes. Nombre CG: señala el nombre dado al gen cadidato; Carácter: indica con el carácter con el que se ha relacionado; Especie: la especie de origen dónde se ha detectado; Secuencia; señala el código del NCBI para acceder a la secuencia de origen; Referencia bibliográfica: muestra de dónde sa ha obtenido la información del gen candidato; Función: la función descrita del gen candidato; Blastx: muestra la homología obtenida en *E.guineensis* mediante este algoritmo y el E-valor correspondiente; Ontología GO: muestra los resultados obtenidos con respecto a la participación del gen candidato en procsesos biológicos, funciones moleculares y dónde se encuentran a nivel celular.

NOMBRE GC	CARÁCTER RELACIONADO	ESPECIE	SECUENCIA	REFERENCIA BIBLIOGRÁFICA	FUNCIÓN	BLASTx	E-VALOR	ONTOLOGÍA GO
P4	Bn, FFB, FN,FW	*Elaeis guineensis*	HC924133	Patente EP2199398 Reuzeau. y col.,2010	Factor de transcripción MADS BOX	NP_001290521.1 factor de transcripcion MADS-box 14	3,0E-138	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Desarrollo floral; P:Diferenciación celular; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
P39	OTROS	*Elaeis guineensis*	FB787669	WO2008034648 (A1) Puzio Piotr y col. 2008)	Proteína 1 NAC	ABB72845.1 NAC protein 1	0,0E+00	C:Núcleo; F:Unión ADN; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG118	OTROS	*Jatropha curcas*	JC000059	Patente EP 2615174 Sanz y Reuzeau, 2013	Serina Carboxipeptidasa	XP_010918615.1 Serina carboxipeptidasa	3,0E-76	P:Proceso metabólico de proteínas; C:Citosol; C:Vacuola; F:Actividad Hidrolasa
KG120	FW,SH	*Arabidopsis thaliana*	NM_001203767.1	Singh y col.2013b	Factor de transcripcion MADS BOX	CAE46181.1 Factor de transcripción MADS-box	2,0E-30	P:Morfogenésis de la estructura anatómica; P:Proceso biosintético;

						AGAMOUS [*Elaeis guineensis*]		P:Proceso metabólico compuesto por una nucleobase; P:Transportador; P:Polinización; F:Unión de Nucleótidos; C:Núcleo; F:Unión ADN; P:Respuesta a estímulos bióticos; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Respuesta a estímulos externos; P:Desarrollo floral; P:Organización de componentes celulares; P:Diferenciación celular; P:Crecimiento celular; F:Actividad Hidrolasa
KG210		*Elaeis guineensis*	JX556215;HF562332	Morcillo y col. 2013	Lipasa	XP_010917338.1 PREDICTED: uncharacterized protein LOC105041961	0,0E+00	P:Proceso Metabolismo de Lípidos; F:Actividad Hidrolasa//P:Proceso Metabolismo de Lípidos; F:Actividad Hidrolasa
KG233	HI	*Elaeis guineensis*	AY556420	Lee y col. 2015	Proteina Asaparagina sintasa	XP_010940408.1 Proteína tallo específica TSJT1	0,0E+00	C:Núcleo; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; P:Transducción de señales; P:Proceso metabólico; C:Citosol; C:Membrana plasmática; P:Respuesta a

								estrés
KG234	HI	*Elaeis guineensis*	AY556423	Lee y col. 2015	Promotor Proteínas Asparagina sintasa	–		
KG243	OTROS	*Elaeis guineensis*	EgEBF*	Preedakoon, 2009	EIN3 (Ethylene insensitive) Unión factor as F-box	XP_010909152.1 PREDICTED: coronatine-insensitive protein 1	0,0E+00	P:Respuesta a estímulos bióticos; P:Proceso metabólico de proteínas; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; P:Transducción de señales; P:Desarrollo floral; P:Proceso catabólico; P:Respuesta a estímulos endógenos; P:Respuesta a estrés
KG271	FW,FN	*Elaeis guineensis*	KJ789862	Singh y col.2014	Gen virescens R2R3-MYB	XP_010931211.1 Factor de transcripción MYB75	2,0E-95	F:Unión ADN
KG275	POP, PO	*Elaeis guineensis*	XM_010935827	NCBI-Nucleótidos	Oleosin	XP_010934129.1 16 kDaOleosin	4,0E-117	C:Membrana; C:Intracelular
KG290	PO, POP	*Elaeis guineensis*	AY182168	NCBI-Nucleótidos	opsc112 protein disulphide isomerase	AAO26314.1 protein disulphide isomerase, partial	0,0E+00	P:Proceso metabólico de proteínas; P:cellular homeostasis; C:Retículo endoplasmático; C:Vacuola; C:Plastidioio; P:Respuesta a estrés; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; C:Membrana; F:Actividad Transferasa

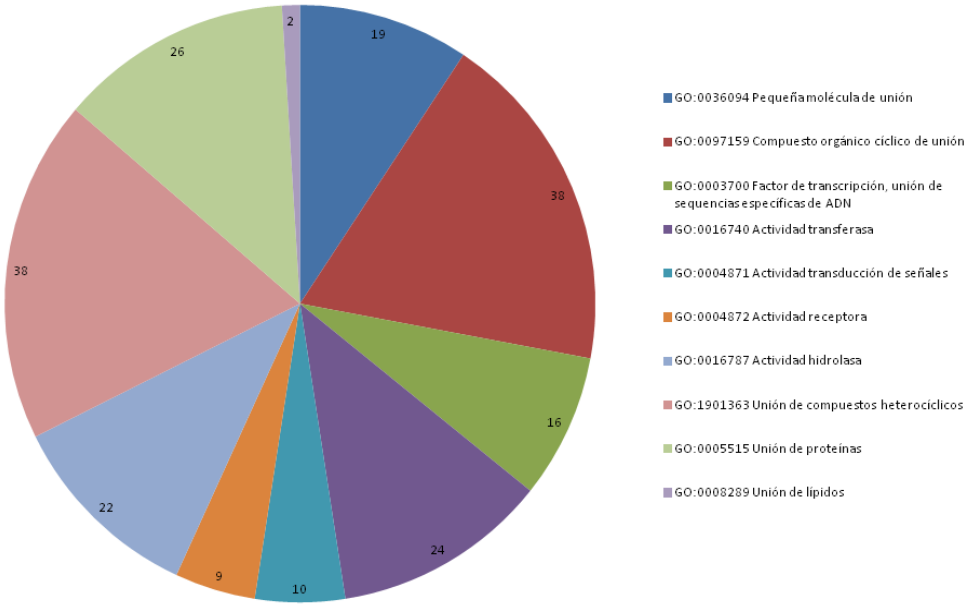


Figura 13: Gráfico de las Funciones moleculares de los genes candidatos seleccionados. GO: código de la ontología. El número de secuencias obtenidas de cada ontología se muestran en el gráfico.

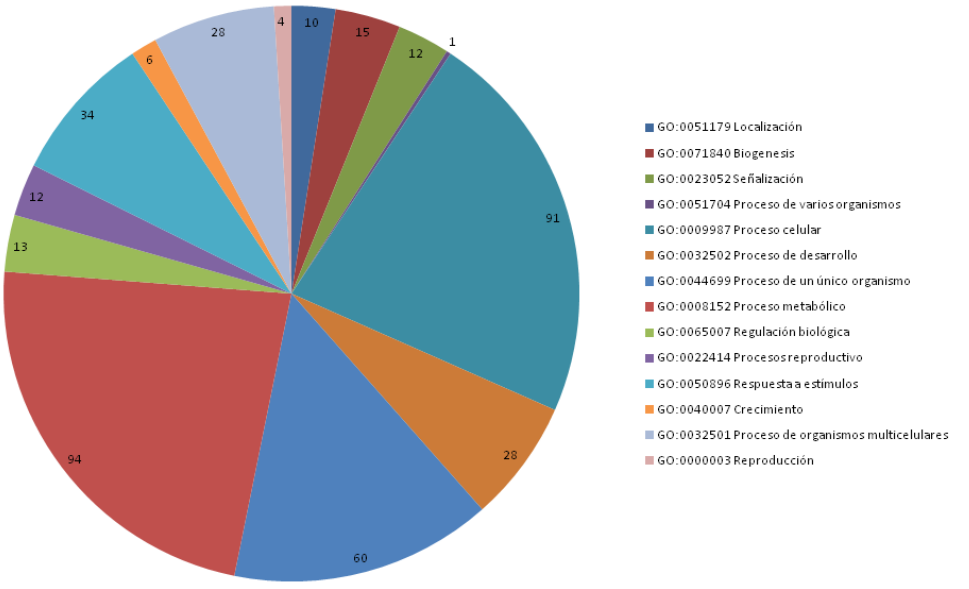


Figura 14: Gráfico de los procesos biológicos donde están implicados los genes candidatos seleccionados.GO: código de la ontología. El número de secuencia obtenidas para cada ontología se encuentra en el gráfico.

5. DISCUSIÓN

Abordar el mapeo por asociación mediante genes candidatos puede convertirse en una tarea ardua y compleja, sobre todo cuando los caracteres estudiados son poligénicos. Esta estrategia es factible para caracteres morfológicos, de crecimiento y genes relacionados con resistencias como se ha demostrado en numerosos estudios (Ingvarsson y Street, 2011). Esto justifica la selección de los caracteres para el desarrollo de esta tesis doctoral, ya que pueden ser susceptibles de una asociación genotipo- fenotipo. En la actualidad, en palmera de aceite africana no hay estudios en los que se aborde el mapeo por asociación mediante la estrategia de genes candidatos, y los caracteres estudiados para la selección y mejora del cultivo en esta tesis se han abordado desde el mapeo de QTL tratando de identificar las regiones involucradas en estos procesos (Lee M. 2015, Montoya y col. 2013, Billote y col. 2010, Singh y col. 2009, entre otros).

Estos caracteres se centran en la mejora del rendimiento y la productividad de la palmera de aceite africana, y son el objetivo principal del proceso de mejora del cultivo. El rendimiento de aceite está formado por dos componentes principales que son: 1. el rendimiento del racimo, caracterizado por el número de racimos (BN) y el peso medio del racimo (BW), y 2. la relación entre aceite/racimo que depende a su vez del número de frutos/racimo (FN), de la relación entre el porcentaje en peso del mesocarpio/porcentaje en peso del fruto (MF), y de la relación entre el volumen de aceite del mesocarpio y el peso del mesocarpio(OM)(Ngando-Ebongue y col., 2012). También se incluye un carácter vegetativo de importancia como es la altura de tallo (HI), ya que las palmeras de la especie *E.guineensis* pueden alcanzar los 25 metros de altura, dificultando la identificación del grado de madurez de los frutos, y su recolección. Además, se incluyo también el índice de iodo (IV) como indicador de calidad del aceite obtenido, buscando un mayor contenido de ácidos grasos insaturados. Estos caracteres variaran en función de las subpoblaciones a la que pertenezcan los individuos (Ngando Ebongue y col.,2012), nuestro material es *Tenera* pero sus parentales pertenecen a diferentes subpoblaciones (Tabla1) y el grado de heredabilidad es diferente en cada carácter como así lo han demostrado numerosos estudios (Rafii y col., 2002; Musa y col., 2004; Okoye y col., 2009; Okwuagwu y col., 1995, 2008), siendo los componentes principales del carácter FFB formado por BN y BW los que menos heredabilidad muestran.

La elección de los genes candidatos en el presente capítulo se realizó mediante tres metodologías diferentes para poder identificar un mayor número de marcadores moleculares funcionales después de su genotipado y búsqueda de haplotipos, tal y como se desarrollará en el siguiente capítulo.

5.1. Detección de genes candidatos mediante BSA cDNA AFLP

La técnica combinada **BSA cDNA-AFLP** permite identificar regiones que diferencian a nivel de genotipo un grupo de individuos de otros, y con un número relativamente bajo de falsos positivos (Yao y col., 2007). Estos conjuntos de individuos se fenotiparon durante 15 años para estos caracteres

agronómicos y se calificaron como favorables o no. Mediante esta técnica se obtuvieron diferentes patrones de expresión en hoja para estos caracteres. En el momento de realizar esta técnica el genoma de la palmera no estaba secuenciado y el conocimiento de secuencias previas era limitado. Esta técnica aportaba rapidez, ya que no necesita que los fragmentos sean clonados como en el caso de creación de librerías de ADNc (Adams y col., 1991; Okubo y col., 1992) y reproducibilidad en los resultados (Polesani y col., 2008). Aunque la técnica es sensible al número de fragmentos detectados por cada combinación de cebadores lo que depende del grado de heterocigosis de la especie de estudio (Herrero, 2013).

En esta tesis se realizaron en total 220 combinaciones de cebadores diferentes de las que se obtuvieron 106 TDF en total de los diferentes caracteres. El número de TDF obtenido por combinación de cebadores es inferior a la unidad (**0.48TDF/CC**). Este rendimiento es inferior al obtenido en otros estudios dónde esta técnica se ha aplicado. Cao y col. en 2013 obtuvieron un rendimiento de 50.5TDF/CC en hojas de arroz sometidas a diferentes estreses térmicos, o el estudio realizado por Gupta y col. (2012) para conocer los diferentes genes implicados en el proceso de biosíntesis y acumulación de flavonoides en diferentes estadios de desarrollo de semillas de dos especies diferentes de *Fagopyrum* los cuáles obtuvieron un rendimiento de 7TDF/CC. En cambio, en otros estudios el rendimiento fue similar como en el caso de la búsqueda de genes relacionados con la tolerancia al frío en hojas y raíces de garbanzo (Dinari y col., 2013) o el estudio llevado a cabo por Yin y col. en 2010 para identificar los genes que participan en la resistencia al hongo *Puccinia graminis* en trigo. Los estudios más recientes en palmera de aceite donde se aplicó esta técnica obtuvieron un rendimiento entre 1-2 TDF/CC, en su mayoría. Roberdi y col. (2015) obtuvo un rendimiento de 1,25TDF/CC cuando aplicó esta técnica para descubrir los genes implicados en un fenómeno que ocurre cuando la planta está sometida a un déficit hídrico en los racimos de frutos conocido como "*hard bunch*" , afectando a la productividad y al rendimiento del cultivo y cuyos patrones de expresión eran hasta ahora desconocidos. También se ha aplicado para conocer los genes que pudieran estar implicados durante la embriogenésis somática en cultivos "*in vitro*" y cuyo rendimiento en el número de transcriptos fue 1,25TDF/CC (Pattarapimol y col., 2015).

Este bajo número de polimorfismos en nuestros genotipos puede ser debido a que nuestras muestras proceden de programas de mejoramiento con cruzamientos dirigidos, y a que el 80% de las secuencias son repetitivas. Este hecho unido al origen inicial único de la especie hace que la heterocigosis disminuya, a favor de conseguir buenos caracteres de interés agronómico mediante la fijación de alelos favorables. Por otro lado, con la disponibilidad de la secuencia completa del genoma de *Elaeis guineensis*, podría optimizarse el proceso mediante una simulación "*in silico*" de la técnica cDNA-AFLP (Stölting y col., 2009) para seleccionar combinaciones de cebadores que muestren mayor nivel de transcriptos previamente, ahorrando en tiempo y costes de laboratorio.

5.1.1. Análisis de los fragmentos de expresión diferencial

En este estudio se identificaron un total de 56 transcriptos diferenciales relacionados con los 8 caracteres de interés agronómico que se expresaban en las hojas de las familias de estudio. Cómo se ha

explicado al inicio de esta discusión, estos caracteres están implicados en el rendimiento y la productividad de la palmera en relación a su valor económico. La productividad depende de diversos factores como el material de siembra, dónde se aprecia la influencia del genotipo del individuo, las necesidades agronómicas del cultivo, la actividad fotosintética y las condiciones climáticas estacionales (Hanniff, 2000). Estos caracteres poligénicos influirán sobre la capacidad de adaptación del cultivo desde el punto de vista fisiológico a los diferentes situaciones que se muestren, de ahí la importancia de descubrir los posibles genes implicados directa o indirectamente en estos procesos.

El análisis bioinformático de de los transcriptos mediante La búsqueda de homologías con la base de datos locales y la base de datos de EST's de E.guineensis y E. oleifera públicos en NCBI mediante el algoritmo BLASTn, junto con la utilización del software B2GO permitieron la caracterización funcional de los los transcriptos (Conesa y Götz, 2008). Las anotaciones obtenidas en GO muestran los resultados en relación al proceso biológico donde participa, su función molecular y el dónde se localiza a nivel celular. Esto junto con la bibliografía relacionada es una herramienta útil para buscar el sentido biológico a cada transcripto. Los resultado mostraron **9 TDF** que pudieran estar **implicados** positivamente en **6** de los 8 **caracteres.**

El carácter **BN** o número de rácimos por año está influenciado por numerosos factores como puede ser la proporción de flores femeninas del total, el número de inflorescencias abortadas y defectos en la formación del mesocarpio, además de factores medioambientales (Kallarackal y col., 2004; Henson y Harun, 2005). Por lo que los genes participantes en los factores implicados de estos procesos fisiológicos son susceptibles de ser seleccionados como genes candidatos. Este carácter es altamente heredable (Meunier y col., 1970), y puede correlacionarse con el rendimiento de aceite (Okoye y col., 2009). Ninguno de los dos transcriptos obtenidos CDA8 Y CDA43 en las familias calificadas como "buenas" para el carácter arrojaron anotaciones funcionales. Aunque ambos se localizaron en el cromosoma 12 y mostraron homología con un EST público (EL685771.1) procedente de una librería relacionada con flores maduras en palmera de aceite publicado por Ho y col. en 2007.

El peso medio de racimo o **BW** depende del peso de cada fruto, y de cada fruto junto con su pequeño tallo. Este peso aumenta con la edad de la palmera y depende también de los asimilados de carbono por parte del fruto, y cuya eficiencia de conversión depende además de los factores medioambientales, de la variación genética existente en cada individuo. El transcripto CDA27 aparece únicamente en las familias calificadas como "buenas" fenotípicamente para el carácter y aunque el mejor resultado de BLASTx muestra una proteína no caracterizada en garbanzo, la comparación con las bases de datos locales muestra homología con un "contig" cuya anotación hace referencia a una proteína relacionada con el almacenamiento en la semilla y el transporte de lípidos en *Arabidopsis thaliana* (AT3G22142)(www.arabidopsis.org). Estas proteinas denomidas LTP,- "lipid-transfer protein" están en las plantas superiores en altas concentraciones y sus funciones se relacionan con la formación de las membranas celulares, la regulación de los ácidos grasos, la formación de cutinas presentes en las cutículas de las plantas y compuestas por ácidos grasos de cadena larga, la embriogenésis y por último

con los mecanismos de defensa ante situaciones de estrés abiótico y biótico (Kader, 1996; Jouannic y col., 2005; Safi y col., 2015). En relación a los mecanismos de defensa Al-Obaidi y col. (2013) caracterizaron en palmera de aceite una LPT responsable de la defensa de la planta contra la enfermedad causada por el hongo *Ganoderma boninensi* durante la infección. También se han caracterizado dos EST's durante los procesos de callogenesis y embriogenesis en cultivos in vitro (Low y col., 2008). Como puede participar en el desarrollo de la semillas y en la acumulación de lípidos en su interior puede estar implicado en el carácter, y seleccionarse como gen candidato.

Los caracteres **FN** y **FW** son componentes de racimo, y se correlacionan negativamente de manera que un mayor número de frutos en el racimo disminuye el peso medio de fruto del racimo (Corley y Tinker, 2003). El número de frutos se relaciona con el número de inflorescencias que se han transformado en fruto, por lo que es importante en este aspecto una polinización eficiente de las inflorescencias femeninas (Corley y Tinker, 2003).En el carácter **FN** se detectó el transcripto CDA90 en 3 de los 4 familias calificadas como "buenas" fenotípicamente que presento homologia con un EST presente en el mesocarpo de *E.oleifera* y cuya descripción funcional hace referencia a la subunidad beta de una enzima ATP (adenosin trifosfato) sintasa. En *Elaeis guineensis* Low y col.(2008) encontraron 22 EST con esta función en tejido embrionario en su estudio de identificación de genes durante la callogenesis y embriogenesis en la propagación vegetativa de la planta. Las enzimas ATP sintasas trabajan bidireccionalmente, bien catalizando la síntesis de ATP o bien su degradación, y lo combinan con el transporte de electrones de un lado a otro de la membrana (Seelert y col., 2000). Están formadas por dos subunidades catalíticas F0 y F1, y es en esta subunidad donde se localizan las subunidades beta que participan en la catálisis de ATP. El ATP es la primera fuente de energía procedente de la fotosíntesis en plantas y de ella dependerá los asimilados de carbono de la planta, y en consecuencia el crecimiento y desarrollo de la planta en todos los niveles influyendo potencialmente sobre FN.

Para el carácter **FW**, el transcripto CDA22 obtenido en todos los fenotipos calificados como "buenos" fue homólogo con un helicasa de ARN con caja "*DEAD*" dependiente de ATP. Esta subfamilia de enzimas es la mayor de la familia de las ARN helicasas que participan en la mayoría de las etapas del metabolismo de ARN como en el inicio de la traducción, en la unión de ARN mensajero, y en el ensamblaje del ribosoma, entre otros (Aubourg y col., 1999; Cordin y col., 2006) por lo que se asocia con numerosas funciones celulares que incluyen el crecimiento y desarrollo de la planta, y parece que también en respuestas a estrés abiótico (Aubourg y col., 19998; Gong y col., 2005; Zhu y col.2015). En palmera de aceite africana un estudio publicado por Ho y col. en 2015 aparece este transcripto únicamente en las inflorescencias masculinas durante el desarrollo floral, pero esto no descarta también su participación en el desarrollo y maduración del fruto como lo afirman algunos estudios de otras especies de cultivos como la manzana (Janssen y col., 2008), el tomate (Zegzouti y col., 1999) o en la pera donde este grupo de enzimas parecen estar presente durante todo el desarrollo del fruto aunque con diferente intensidad en cada etapa (Xie y col., 2013). CDA23 fue otro transcripto obtenido en todas las familias caracterizadas como "buenas" fenotípicamente que obtuvo homología funcional con una

proteína del floema (PP2) cuya anotación ontológica describe su proceso funcional como unión de carbohidratos. Estas proteínas son conocidas como proteínas P y se localizan en el floema, tejido responsable del transporte de nutrientes y fotoasimilados en la planta a través de la savia, aunque también lleva numerosas moléculas estructurales y de información como proteínas y ARNm (Dinant y col., 2003) . En la palmera de aceite africana la savia del floema transporta los azúcares para el desarrollo de los racimos de frutos (Corley y Tinker, 2003). Este mecanismo de transporte permite la comunicación entre los diferentes órganos de la planta, y existe evidencia indirecta de que los eventos que ocurren en el floema pueden controlar la fisiología y desarrollo de la planta (Ruiz-Medrano y col., 2001). La proteína PP2 se transloca en el flujo de asimilados ejerciendo como lectina, cuya principal función se cree es la defensa de la planta frente a los ataques por insectos (Kehr, 2006), u otros factores de estrés como heridas o factores oxidativos (Zhang y col., 2011), pero además tiene capacidad para unirse al ARN (Gomez y col., 2005), y sus efectos pueden verse en otros tejidos de la planta debido a su capacidad de movimiento a través de la savia. La mayoría de las investigaciones sobre la funcionalidad de esta proteína se ha realizado en el género de las curcubitaceas y en *A. thaliana*, aunque se sabe que están presentes en la mayoría de las angiospermas, y en algunas ginnospermas (Dinant y col,. 2003). En nuestra especie no se ha publicado hasta le fecha ningún estudio sobre la funcionalidad de esta proteína ni su posible relación con el fruto. Debido a su participación en el transporte de asimilados y de ARN se puede incluir como un posible gen candidato relacionado con el peso del fruto. El transcripto CDA26 presente también en todos las familias cuyos fenotipos fueron "buenos" es homólogo con una proteína desconocida que se localiza en la mitocondria. Este transcripto fue homólogo con el contig M01000001578 obtenido por Bourgis y col. (2011) en el estudio donde se realizó un análisis comparativo del transcriptoma y perfil de metabolitos entre la palmera de aceite africana y la palmera datilera en el mesocarpo en diferentes estadíos en relación a la partición del carbono. Este contig se anotó como homólogo funcional con una secuencia de *A. thaliana* (TAIR: AT4G21980) cuya función se relaciona con un precursor de una proteína relacionada con el mecanismo de autofagia celular, proceso importante en el reciclado de nutrientes por la célula sobre todo durante la senescencia y en condiciones de crecimiento con déficit de carbono y nitrógeno (Thompson y Vierstra, 2005). Los estudios muestran esta proteína como participante de la senescencia de la hoja, y en *A.thaliana* se ha demostrado que los genotipos mutantes de esta proteína presentan una senescencia temprana de la hoja, por lo que parece que la autofagia tiene un papel determinante en la longevidad de la hoja (Avila-Ospina y col., 2014). Gracias a este mecanismo los componentes celulares como proteínas, lípidos y ácidos nucleicos se degradan y se movilizan hacia otras partes de la planta cuando las hojas se mueren (Shpilka y col., 2011). También parece que participan en la degradación de azúcares acumulados durante el día en los cloroplastos que durante la noche se exportan al citosol para asegurar un aporte óptimo de carbono que permita el crecimiento y desarrollo de la planta (Wang y col., 2013). Por tanto, este transcripto puede ser un posible gen candidato que aunque no muestra una relación aparente con el fruto, sí lo hace con las hojas de donde se ha obtenido el ARN y puede que esté relacionado de manera indirecta en el fruto.

El caracter **CPO** o aceite de palma crudo se define como el volumen de aceite extraído del mesocarpio del fruto en toneladas/hectárea de plantación. Se estima que el valor medio de rendimiento está entre 4,1 y 18 ton/ha (Lee y col., 2015). Un factor determinante para este carácter es el tipo de fruto presente en la población, controlado por el gen de herencia co-dominante relacionado con el grosor de la cáscara o *SH*. Nuestro estudio se centra en una población de variedad *Tene*ra (*SH*sh) por lo que esté carácter presentará un rendimiento del 30% de media superior a las variedades cuyo fruto son *Dura* (*SHSH*) (Corley y Tinker, 2003). Por otro lado, las plantas acumulan aceite en diferentes tejidos de la semilla y el fruto, aunque también se acumula en hojas y tallos (Lersten y col., 2006; Durret y col., 2008;). El proceso de desarrollo y maduración del fruto, cuyas fases claves dependen de las características morfológicas, ceulares, bioquímicas y hormonales, influenciará sobre el carácter CPO y se implicarán otros caracteres como OM, IV Y MF. Ya que es durante la maduración del fruto, entre las 12 y las 24 semanas después de la polinización, cuando comienza la síntesis y elongación de los ácidos grasos para formar los triacilgliceroles y continuar con la biosíntesis de aceite (Bates y col., 2009; Baud y Lepiniec, 2010). Los resultados obtenidos para este carácter estaban relacionados con los fenotipos "malos", y estos TDF obtenidos pueden mostrar la presencia de un alelo de un gen candidato que influye negativamente en el carácter ya que los fragmentos de expresión diferencial son debidos a la abundancia de transcriptos que pueden o no apreciarse o a la presencia de alelos específicos que presentan una región de corte.

En cambio, en el carácter **MF** los transcriptos CDA34 y CDA37 están presentes en todos los fenotipos calificados como "buenos". El transcripto CDA34 parece ser una proteína no caracterizada en una planta con flor *Ambronella trichopoda*, aunque su función es desconocida. El transcripto CDA37 se caracteriza como una enzima quinasa dependiente de ciclina F1 (CDKF1=CDK-acyivating kinases) y cuya anotación "GO" muestra su participación en el desarrollo post-embrionario, en la morfogenesis de la planta, y en múltiples rutas metabólicas. Las diferentes fases del ciclo celular, donde la célula crece, se replica el ADN y se produce la división celular, está regulada por un grupo de enzimas denominado CDK ("Cyclin dependent Kinases") (Inze y De Veylder, 2006). La división celular es crucial para el desarrollo de la planta, por lo que si ocurren cambios en la expresión de los genes implicados en este proceso pueden asociarse con las transiciones que se producen en el ciclo celular durante el crecimiento de un órgano (Malladi y Johnson, 2011), y ser susceptibles a cambios en la arquitectura de la planta y en los caracteres de interés agronómico como el rendimiento (Den Boer y Murray, 2002). La enzima caracterizada por el transcripto CDA37 es propia de plantas y en condiciones "in vitro" parece activar a CDK2 y CDK3 mediante la fosforilación de un residuo de Thr (Simotohno y col., 2004; Umeda y col., 2005) de la CDK ("Cyclin depent protein kinases") cuando se une a ellas. En *Arabidopsis* CDKF:1 actúa como un regulador positivo de la proliferación celular, y cuando pierde esta función se reduce la producción celular y la endorreplicación (Takatsuka y col., 2009). Por último, en manzana Malladi y Johnson (2011) comprobaron que los genes implicados en el ciclo celular facilitan la producción de células durante el desarrollo del fruto. Esto ocurre igualmente en la palmera datilera, especie cercana genéticamente a la palmera de aceite africana (Janssen, 2004), se validó mediante PCR a tiempo real

una CDK tipo A que aparecía en las etapas del desarrollo del fruto donde se produce la multiplicación celular, la expansión celular y en la etapa de maduración del fruto donde hay cambio de color en el fruto (Yin y col., 2012). Todos los indicios existentes hacen a este transcripto susceptible a ser denominado gen candidato para este carácter, aunque es importante resaltar que el material utilizado para realizar la técnica fue hoja, y también estas enzimas participan en el crecimiento celular de las hojas. Por tanto, siempre y cuando este posible gen candidato este asociado al carácter de interés deberían realizarse experimentos que acrediten que su presencia es debida al fruto y no a la hoja.

En el carácter **OM** el único transcripto obtenido en las familias calificadas como "buenas" fue CDA78, y apareció en tres de las cuatro familias fenotipadas. Aunque se encontró homología con un contig de palmera aceitera americana (*E.oleifera*) de las bases de datos propias, no mostró ninguna homología funcional con proteínas y por tanto ninguna anotación ontológica por lo que su función es desconocida. A pesar de ello, el transcripto se seleccionó como posible gen candidato ya que parece estar implicado en el carácter, y su caracterización puede realizarse a posteriori si se encuentra una asociación fenotipo-genotipo relevante durante el desarrollo de la presente tesis.

5.2. Detección de genes candidatos co-localizados con QTL's de interés agronómico en un mapa genético funcional de alta densidad

La utilización de un mapa genético de ligamiento funcional de alta densidad y en el que se integran los QTL de consenso relacionados con los caracteres de interés agronómico es una herramienta útil en los programas de mejora de *E.guineensis* Jacq. debido a su largo ciclo de selección. En esta tesis se utiliza un mapa genético de referencia en el que se han integrado numerosos recursos genéticos disponibles descritos en el apartado de materiales y métodos, para saturar el mapa de diferentes marcadores y genes anotados y crear una herramienta útil de aplicar para conseguir el objetivo de este capítulo de búsqueda de genes candidatos, ya que la identificación de genes posicionales, algunos con función conocida en otras especies, dentro de los intervalos de confianza del QTL puede considerarse que contribuye a la detección de genes implicados en el carácter de interés (Pflieger y col., 2001).

Esta estrategia se ha utilizado en la identificación de posibles genes candidatos en numerosas especies de plantas incluidos árboles en los que los ciclos de selección también son largos. En estos estudios se han mapeado, en primer lugar, los QTL de interés agronómico en la población de estudio y posteriormente los genes asociados al mismo (Liu y col.,2011; Correa y col., 2014; de Miguel y col., 2014; Nuñez-Lillo y col., 2015). En palmera de aceite africana, Jeennor y Volkaert (2014) identificaron mediante esta estrategia genes relacionados con la ruta biosintética de aceite, factores de transcripción y algunos genes expresados durante la acumulación de aceite en el fruto que posteriormente se colocalizaron con algunos de los QTL que habían mapeado. Finalmente, mantienen la hipótesis de que la co-localización de marcadores basados en genes con QTL de interés en la especie sugieren que esos genes son responsables en parte de la variación en los caracteres relacionados con el rendimiento de aceite de la palmera. Esto hace nuestra estrategia una herramienta válida para la búsqueda de genes

candidatos relacionados con los caracteres de interés, explicados al inicio de esta discusión. En 2015, Lee y col. publicaron también un mapa de ligamiento de consenso en el que identificaron el QTL que explicaba el mayor porcentaje de variación en el fenotipo para la altura de tallo, y encontraron colocalizado a él un gen que codificaba un enzima que parece estar implicado en la disminución de esta altura, la asparagina sintetasa. En esta tesis se utilizó como **mapa de referencia** el mapa de ligamiento para la población **LM2TxDA10D** publicado por Billotte y col. (2005) y saturado de marcadores SSR al que se integraron QTL's de consenso en diferentes poblaciones de *E.guineensis* Jacq. (Billotte y col.,2010; Lee y col., 2015), e incluso de una población interespecífica (*E.guineensis* Jacq x *E.oleifera* Cortes HBK) (Montoya y col. 2013). Por último, se integró también en el mapa genético el genoma de *E.guineesis* Jacq publicado por Singh y col. en 2013a que facilita la identificación de genes relacionados con los caracteres de calidad y productividad del cultivo. Estos mapas utilizados incluyen marcadores de anclaje como SSR, RFLP y SNP que han demostrado ser congruentes entre diferentes poblaciones (Wu y col., 2001; Wang y col.,2003; Schneider y col.,2007), por lo que sirven para ser utilizados en estudios comparativos posteriores e integrar diferentes mapas para cubrir ampliamente el genoma (Jeennor y Volkaert, 2014). Es importante mantener presente el objetivo de esta tesis que es buscar la asociación fenotipo-genotipo mediante un mapeo por asociación en la población de estudio que se desarrollará en posteriores capítulos, y dónde sí se tendrá en cuenta nuestra población de palmeras.

La estrategia utilizada para la búsqueda de genes candidatos co-localizados con los QTL que representaban los caracteres agronómicos de interés para el rendimiento y la productividad de aceite fue sondear las secuencias anotadas, y algunas de ellas eran marcadores SNP a una distancia de ± 1cM (centimorgan) en cada QTL correspondiente al carácter, y revisar sus anotaciones y bibliografía relacionada para poder deducir su posible sentido biológico. Se eligió esta distancia de manera arbitraria para asegurarnos que los marcadores secuenciados que flanquean al QTL estén estrechamente ligados a él, disminuyendo las probabilidades de recombinación, tal y como recomiendan Collard y col. (2005) para el mapeo de alta resolución de QTL. En el caso de Jeennor y Volkaert (2014) los genes candidatos encontrados se encontraban a <5 cM del QTL y mostraban polimorfismos en sus secuencias tipo SNP, lo mismo se muestra para el estudio de Lee y col. en 2015. La presencia de marcadores SNP anotados refuerza la probabilidad de posibles variaciones en el fenotipo, dependiendo como el cambio de base afecte a la proteína por lo que son genes muy susceptibles a ser genes candidatos.

5.2.1. Selección de genes candidatos co-localizados con QTL's relacionados con los caracteres agronómicos de interés

En total se seleccionan 86 genes candidatos colocalizados con QTL's para los caracteres agronómicos de interés. Estos genes candidatos se agruparon en función del carácter y del QTL para facilitar su discusión. Los caracteres estudiados junto con los diferentes QTL's analizados se muestran en el anexo 3.

5.2.1.1. Caracteres relacionados con la producción

Los QTL's revisados para **BN** se localizaron en 11 de los 16 grupos de ligamiento del mapa. En el GL 7 a una altura de 48,8 cM se localizo un QTL BN6_9 que estaba colocalizado con otro relacionado con BW lo que puede evidenciar que están funcionalmente relacionados (Thumma y col., 2001), y los genes que los flanquean pueden estar implicados en ambos caracteres. Estos genes fueron KG141, KG142 y KG143 y flanqueaban el QTL en su parte superior, y sus secuencias localizadas en el mesocarpio del fruto de la palmera de aceite proceden del estudio de Bourgis y col. (2011). Aunque no se obtuvieron anotaciones ontológicas para ellos la bibliografía existente ayuda a dilucidar su funcionalidad.

KG141 es una proteína asociada a microtubulo (MAP) que participa en numerosos procesos del desarrollo de plantas y en su morfogenésis regulando la división y expansión celular (Wasteneys, 2004; Hamada, 2007, 2014; Celler y col.,2016).

KG142 se caracterizó como una proteína reguladora de respuesta tipo b (AAR2). Estos reguladores de respuesta que se han caracterizado en *Arabidopsis thaliana* son factores de transcripción que actúan como reguladores positivos en la ruta de señalización de citoquininas que está conservada en plantas (Mason y col., 2005; Argyros y col. 2008). Las citoquininas son hormonas que promueven la división y la diferenciación celular, y son fundamentales en el proceso de organogenesis, así como en la regulación de numerosos procesos fisiológicos. Concretamente, las proteínas ARR2 parece que juegan un papel importante en el desarrollo floral, y en diferentes procesos de desarrollo del raíces y brotes (Kim y col., 2012), así como en la senescencia de las hojas (Hill y col., 2013). En palmera de aceite, Ramli y Abdullah (2010) identificaron durante la caracterización de los promotores de metaloproteínas MT3-A y MT3-B un regulador de respuesta a citoquininas en diferentes regiones de ambos promotores. Las metaloproteínas presentan unos perfiles de expresión específicos en diferentes tejidos y temporales (Cobbett y Goldsbrough, 2002), lo que permite entender mejor sus funciones relacionadas con la absorción de metales por la planta e implicadas en el crecimiento y en respuesta a estrés. MT3-A y MT3-B son metaloprotínas de tipo 3 que aparece en tejidos en fase de maduración (Reid y Ross, 1997), y ambos se expresan en el mesocarpio de la palmera de aceite en diferentes estadios (Abdullah y col., 2002; Omidvar y col., 2008).

La última secuencia co-localizada en esta región ,KG143, mostró una proteína T del complejo 1 subunidad delta llamada TCP1 es un miembro localizado en el citosol de la familia de las chaperoninas, las cuales están implicadas en el plegamiento de las proteínas utilizando la energía procedente de la hidrólisis de ATP y cuyos sustratos son la actina y la tubulina (http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=239454). Este gen está relacionado con KG141 co-localizado en la misma región, ya que los microtubulos están formados por diferentes isoformas y modificaciones de tubulina (Mandelkow y Mandelkow, 1995). Además parece estar implicado en la comunicación célula-célula en las plantas, y en los que participan los factores de transcripción de la familia KNOX esenciales para el establecimiento y mantenimiento de las células madre, para el que uno de ellos necesita para trabajar el complejo de chaperoninas (Xu y col., 2011).

KG196 y KG197 se encuentran co-localizados muy próximos por encima del QTL, QBN3_5, en el GL 2 cuyas funciones se relacionan con el desarrollo floral. KG196 es un gen caracterizados como factor de transcripción de tipo MADS BOX 21, los cuáles controlan diferentes procesos del desarrollo de las plantas con flor que suceden desde la raíz hasta la flor y en el desarrollo del fruto (Becker y Theiβen, 2003). Este factor de transcripción es de tipo II, conocido en plantas como MIKC y dentro del modelo ABC de genes que explican el desarrollo de los órganos florales dirigidos por los diferentes tipos de factores de transcripción MADS BOX en Arabidopsis thaliana se identifica con los genes C o AGAMOUS (AG) (Shore y Sharrocks, 1995; Becker y Theiβen, 2003). En palmera de aceite existen estudios que han confirmado la presencia de estos factores de transcripción en el desarrollo floral y durante la maduración del fruto (Adam y col., 2006; 2007; Ho y col., 2007; Shearman y col., 2013). Tranbarger y col. (2011) encontraron diferentes transcriptos pertenecientes a factores de transcripción tipo MADS BOX en diferentes etapas de la maduración del fruto de la palmera de aceite, tres de los cuales pertenecían a la familia AG. KG197 se caracterizó como una proteína denominada FRIGIDA. Esta proteína es el mayor determinante en las variaciones de tiempos de floración y en respuesta a la vernalización en diferentes ecotipos de *Arabidopsis thaliana*, y parece que forman un complejo de proteínas FRI-C, cuyos componentes tienen funciones especializadas relacionadas con la unión de ADN, con activadores de la transcripción de proteínas y mantenimiento del complejo (Johanson y col., 2000; Choi y col., 2011). La palmera de aceite africana no sufre procesos de vernalización ya que es un cultivo tropical que no requiere frío para su floración, pero sí posee diferentes períodos de floración alternativos de flores masculinas y flores femeninas, aunque los ciclos de floración parecen al azar, incluso en climas estacionales (Hemptinne y Ferwerda, 1961; Corley, 1977). En su mayoría son las flores femeninas las que posteriormente formaran los racimos. Posiblemente este gen actuará conjuntamente al factor de transcripción MADS BOX anterior durante la formación de las inflorescencias, aunque se necesitan estudios donde se caracterice la funcionalidad de KG197, y la influencia de KG196 en el desarrollo floral y concretamente en su participación en el desarrollo de las inflorescencias femeninas.

Como puede observarse en la tabla 3.1 (Anexo 3) el resto de genes se co-localizaron individualmente en diferentes QTLs relacionados con el carácter BN. KG198 se caracterizó como una enzima endopeptidasa de tipo cisteína localizada en los espacios extracelulares que participa en el desarrollo y crecimiento de la planta, en la senescencia, muerte celular, también participan en la acumulación de proteínas de reserva y en su movilización, y por último en respuestas a situaciones de estrés (Grudkowska y Zagdanska, 2004). En *Jatropha curcas* se identificaron tres cisteína proteinasas en el endospermo durante la germinación de la semilla avalando la movilización de las proteínas durante la germinación (Costa y col., 2010). En palmera de aceite africana se han identificado dos cisteínas proteasas durante el proceso de embriogenesis somática y su posterior germinación (Aberlenc-Bertossi y col., 2008), esto es debido a que estas enzimas participan en la generación de reservas de tipo globulina en la semilla, en su acumulación durante el proceso de embriogenesis y en su movilización durante la germinación (Fisher y col., 2000). KG199 se caracterizó como una enzima base de cadena larga esfingoide quinasa (LCBK). Esta enzima cataliza la reacción de fosforilación de los esfingoides de

cadena larga (LCB) para formar LCB-1-fosfato (LCB-P)(Imai y Nushira, 2005). Diferentes autores han evidenciado las funciones de LCB libre y LCBP como mediadores de la respuesta celular en *Arabidopsis* (Ng y col., 2001;Coursol y col., 2003, 2005;Xiongy col., 2008), además participan en la señalización intermedia de la apertura de los estomas en condiciones de sequía (Ng y col., 2001;Coursol y col., 2003), asi como en el control de la germinación (Worral y col., 2008). El gen candidato TEST se caracterizó como un enzima acetil-co-A oxidasa(ACX4) perteneciente a la familia de genes ACX que cataliza la reacción de beta oxidación en los ácidos grasos peroxisomales. Esta familia de genes parece tener especial importancia en el catabolismo de los ácidos grasos de los lípidos para producir un esqueleto de carbono y producir la energía metabólica necesaria para el crecimiento temprano y la post-germinación en las semillas oleaginosas (Kindl, 1987; Rylott y col., 2001). Rylot y col. (2003) mostraron que la doble mutación en ACX3 y ACX4 en *Arabidopsis* ejercían funciones esenciales para el desarrollo temprano del embrión. JEennor y Volkaert (2014) en su mapeo de QTL's para caracteres de rendimiento de aceite co-localizaron genes de esta familia con QTL de BW y FFB, pero no con BN, aunque es necesario tener presente la correlación negativa entre BW y BN.

KG200 se caracterizó como un gen que codifica una proteína trehalosa 6 fosfato sintasa implicado en el metabolismo de carbohidratos. En *Arabidopsis* se ha observado que es necesaria para el crecimiento y la maduración floral (van Dijken y col. 2004; Wahl y col.2013), así como que participa en la regulación de la glucosa y la señalización ABA durante el desarrollo vegetativo (Avonce y col., 2004). Este enzima junto con otros implicados en el metabolismo de carbohidratos parece que juega un papel importante en el desarrollo y maduración del fruto en la palmera datilera (Bourgis y col. 2011; Yin y col., 2012; All-Mssallem y col.,2013; Xin y col. 2015), aunque en su fruto abundan los carbohidratos y no los lípidos como en el caso de la palmera de aceite africana, a pesar de ser ambas especies altamente conservadas. El gen candidato KG201 codifica una proteína L23 ribosómica 60s, localizada en el citosol celular. Estas proteínas encargadas de la síntesis de proteínas a partir ARN mensajero parecen tener un papel importante en el desarrollo de las plantas (Byrne, 2009). Este hecho los sugieren numerosos estudios con mutaciones de proteínas ribosómicas que muestran fenotipos con crecimiento defectuosos, cambios en el desarrollo de las hojas y fenotipos relacionados con auxinas (Horiguchi ycol. 2012). En *E. guineensis* Jacq. se han detectado estas proteínas en los transcriptos generados en librerías de EST's de ápices normales, y anormales (aquellos que muestran "mantled") y en inflorescencias masculinas (Jouannic y col. 2005). KG202 se caracterizó como un gen que codifica un proteína de la familia de enzimas O-glucosyltransferasa localizado en el aparato de Golgi de la célua. Aunque esta familia de proteínas no está lo suficientemente caracterizada en plantas (https://www.ebi.ac.uk/interpro/entry/IPR024709), algunos estudios muestran una posible relación con las células que forman la pared celular en las plantas por su participación en la biosíntesis de xiloglucanos (Keegstra y Raikhel, 2001; Reiter, 2002; Scheible y Pauly, 2004). La principal función de la pared celular es reforzar el cuerpo de la planta, pero también ejerce un papel clave en su crecimiento, en la diferenciación celular, en la comunicación intercelular, el movimiento del agua y en los mecanismo de defensa (Cosgrove, 2005).

Como puede revisarse en el anexo 3, KG203 y KG204 se encontraron en el mismo GL, pero co-localizados en diferentes QTL's relacionados con el carácter. El gen candidato KG203 codifica una proteína relacionada con el enzima 2-3 bifosfoglicerato-independiente fosfoglicerato mutasa, cuya principal proceso funcional es en el metabolismo de carbohidratos, concretamente en la glucolisis para la degradación de los carbohidratos. Participa en la ruta de conversión de sacarosa en piruvato (http://www.uniprot.org/uniprot/P47669). Parece que los genes implicados en la glucolisis se se coordinan con el inicio de la síntesis de lípidos, por lo que la glucolisis podría regularse a nivel transcripcional en los tejidos vegetales que acumulan aceite (Hajduch y col. , 2006;. Troncoso-Ponce y col., 2011), para ello existen dos vías paralelas en el citosol y en el plastidio (Plaxton, 1996). Este enzima se ha detectado en el mesocarpo del fruto de *Cocus nucifera*, también oleaginoso, durante la maduración del mismo en las transiciones de pulpa tierna al inicio del engrosamiento, y del engrosamiento al endurecimiento de la pulpa (Liang y col.,2014). Es en estas fases donde hay una cambio de composición en los ácidos grasos presentes, ya que a medida que avanza la maduración del fruto los ácidos grasos tienden a ser más saturados y por tanto sólidos. Esto sugiere que durante estas transiciones la ruta glucolítica está activa para formar piruvato que suministrará la energía suficiente a la ruta biosintética de ácidos grasos para formarlos.Pero, en un estudio llevado a cabo por Dussert y col. (2013) en *Elaeis guineesis* Jacq. postula además que esta ruta en los plastidios del mesocarpo tenga otra funcionalidad compartida con las semillas verdes fotoherotrófas. KG 204 codifica una proteína que control divisón celular (CDC48B) que probablemente participa en la división y el crecimiento celular (http://www.uniprot.org/uniprot/Q9ZPR1; Feiler y col. 1995; Rancour y col. 2002; Merai y col., 2014). KG205 codifica una proteína de unión GTP nucleolar (NOG2). Esta proteína es una enzima GTPasa que se asocia con la unidad pre-ribosomal 60S en el nucleólo, para que el ribosoma pueda ser exportado y madure el ARN ribosómico (https://www.ebi.ac.uk/interpro/entry/IPR024929; Saveanu y col. 2001). Estudios recientes han mostrado que los genes que codifican factores de unión a los ribosomas, como GTPasas, juegan un papel importante en el desarrollo de la planta (Byrne, 2009; Horiguchi y col., 2012). El gen candidato KG206 codifica una proteín homóloga a TDP1 localizada en la membrana y con actividad catalítica. En *Arabidopsis*, este gen participa en la especialización celular durante la antesis y desarrollo del polen de la flor (Yang y col., 2003, 2005). Además puede servir como un ligando extracelular para un receptor quinasa que señaliza la determinación de células finales ("*cell fate*") durante la reproducción sexual de la planta (Jia y col., 2008). Ho y col. (2007) detectaron 9 EST de este gen en cultivos celulares en suspensión, donde este TDP1 puede ser clave en la disolución callosa, como se ha visto en *Arabidopsis*, y jugar un papel clave en la diferenciación del tapete y su función (Zhu y col. 2008).

El último gen candidato co-localizado con BN fue KG207. Este gen candidato codifica una proteína de la membrana mitocondrial externa porina 5 (VDAC), cuya función es formar canales de aniones dependientes de voltaje en la membrana externa de la mitocondria para difundir pequeñas moléculas hidófilas (http://www.uniprot.org/uniprot/Q84P97). La caracterización y funcionalidad de esats proteínas por diferentes autores no sólo le otorga un importante papel en el transporte de

metabolitos, si no que parece que están involucradas en la muerte celular programada en respuesta a diferentes tipo de estrés (Kusano y col., 2009; Tateda y col., 2011; Homblé y col., 2012; Duncan y col.,2013;)

El carácter **BW** mostró QTL's en 3 de los 16GL, y se encontraron 6 genes candidatos co-localizados con estos QTL's. Tres de estos genes ya han sido motivo de discusión en este apartado (KG141, KG142 y KG143) por compartir co-localización con BN. El gen candidato KG146 se encontró co-localizado por la parte superior con 4 QTL's presentes en el GL5 a 125cM. La secuencia de este gen (M01000069634) fue detactada por Bourgis y col.(2011) en la palmera datilera (*Phoenix dactylifera*) y se caracterizó como una Histona H2B.11 presente en los nucleosomas de las células eucariotas formando parte de la estructura de la cromatina (Bhasin y col., 2006). Esta estructura de la cromatina es dinámica por lo que modula la accesibilidad al ADN, regula los procesos donde el ADN actúa como modelo (replicación y reparación de ADN, recombinación, transposición, y transcripción) y afecta a diferentes procesos como el crecimiento de raíces, el tiempo de floración, la organogenesis floral, la formación del embrión, así como en las respuestas a estrés abiótico y biótico (Nelissen y col., 2005; Shen y Xu, 2009; Álvarez y col., 2010; Berr y Shen, 2010; Berr y col., 2011). Las modificaciones post-traduccionales de las histonas, como la acetilación o la metilación, entre otros, producen cambios estructurales y funcionales en la cromatina (Loidl, 2003; Pfluger y Wargner, 2007). Es en estas modificaciones donde radica su importancia, ya que pueden darse cambios epigenéticos que influyan en la expresión de diferentes genes. Por ejemplo, la monoubiquitinación de las histonas H2B parece regular los niveles de ácido abscinico (ABA) durante el desarrollo de la semilla, por lo que los cambios en las histonas pueden repercutir en los niveles de expresión de ABA, hormona relacionada también con la acumulación de reservas durante la maduración de la semilla, tolerancia a la deshidratación, dormacia y germinación de la semilla (Chinnusamy y col. 2008). Es desde estas modificacines estructurales dónde podría influenciar sobre los caracteres BW y FN con los que esta co-localizado.

La secuencia del gen KG2 procede de un marcador microsatélite *mEgCIR3275* secuenciado por Billotte y col. (2005) en hoja y cuya función se identifica con una enzima serina hidroximetiltransferasa que está implicada en el fotorrespiración de la planta y es inducida por la luz (Wringler y col., 2000; Ros y col., 2014). Este compuesto participa en la biosíntesis de numerosas moléculas necesarias para la proliferación celular, como aminoácidos, bases nitrogenadas, fosfolípidos y esfingolípidos (Ros y col., 2014).

La secuencia (M01000043696) del gen candidato KG140 pertenece a la base de datos de contigs obtenidos en el mesocarpio por Bourgis y col. (2011). Este gen codifica un factor de transcripción ERF014 sensible a etileno que pertenece a la famila de factores de transcripción AP2/ERF. Este factor de transcripción parece actuar como un activador transcripcional que regula la expresión de genes relacionados con las situaciones de estrés de la planta y sus rutas metabólicas correspondientes (http://www.uniprot.org/uniprot/Q9LPE8). Trabanger y col. (2011) determinaron en qué grado la hormona etileno participa en los procesos de maduración del mesocarpio en la palmera de aceite

africana. Estos autores observaron un gran número de transcriptos que codificaban diferentes ERFs cuya máxima expresión fue a los 140 días después de la polinización cuando el fruto ya está maduro y ha acumulado lípidos y carotenos en el mesocarpio. A continuación, hay un aumento de hormonas como etileno y aumenta considerablemente el peso del mesocarpio, acumulando grandes cantidades de lípidos y carotenos. También se ha visto la implicación de los factores de transcripción ERF asociados a la hormona etileno en los procesos de maduración de otros frutos como en el plátano (Xiao y col.2013), manzana (Wang y col., 2007), tomate (Liu y col., 2015) o kiwi (Yin y col., 2010).

El carácter **FFB** íntimamente relacionado con BN y BW mostró en total 17 genes candidatos en 6GL de la especie, destacando que en el GL2 se encontraban los QTLs, QFFB3_5 relacionado con la medida fenotípica en palmeras de entre 3 y 5 años, y QFFB6_9 con palmeras de entre 6 y 9 años de edad en la misma posición, lo que puede indicar que los genes implicados en este carácter no sufran variación con la edad de la palmera durante este periodo. El gen candidato KG181 codifica una proteína 3-oxoacil-ACP reductasa, llamada también β-ketoacyl-[ACP] reductasa, es una enzima que participa en la biosíntesis de ácidos grasos durante la elongación de las cadenas y en la biosíntesis de ácidos grasos poliinsaurados. Bourgis y col. en 2011 detectaron una presencia muy superior de EST's de estas enzimas durante el proceso de maduración del mesocarpio en *Elaeis guineensis* en comparación con los EST's detectados en la palmera datilera. Esto avala su participación en el metabolismo de ácidos grasos de esta enzima, ya que es durante este proceso cuando se acumulan en el fruto, y por tanto aumenta la masa del mesocarpio, sin tener en cuenta el peso de la semilla también rico en lípidos que suponen un 40% del peso seco final (Slabas y Fawcett, 1992). KG182 es un gen que codifica una proteína transportadora sec61 cuyo principal papel parece ser la translocación de proteínas desde el citosol al retículo endoplasmático (Wiertz y col.1996; Osborne y col., 2005), y como retrotranslocón (Nakatsuka y Brodsky, 2008). KG183, KG184 y KG185 son genes que codifican un citocromo P450 ,y como puede verse en la tabla 3.1 (Anexo 3)están infimamente cercanos en pares de bases, lo que puede indicar que o bien los tres "contigs" son un único gen o son una superfamilia de genes situados en la misma región del genoma, lo que es más probable debida al gran número de genes que codifican citocromos P450 y a sus subfamilias. Estos citocromos P450 se han calificado como enzimas monooxigenasas que tienen como sustrato una amplio espectro de compuestos. Además poseen un amplio número de funciones entre las que se encuentran su actividad como precursores esteroles de membrana, su participación en la homeostasis de las fitohormonas o su implicación en la biosíntesis de numerosos compuestos como por ejemplo pigmentos (Bak y col., 2011). Los tres genes candidatos pertenecen a la misma subfamilia, CYP71A. Esta familia se aisló la primera vez en el fruto del aguacate y se relacionó con su proceso de maduración relacionando la implicación de la hormona etileno con la regulación de esta familia de citocromos P450 (Bozak y col., 1990). Umemoto y col. (1993) también relaciona esta familia de genes con el proceso de maduración del fruto en *Solanum melongena*, aunque no lo limita sólo a la maduración del fruto porque supone que puede tener funciones en los diferentes tejidos de la planta como flores o plántulas. En *Arabidopsis thaliana* sólo han mostrado funcionalidad 3 de los 54 transcriptos detectados para estas familias que parecen tener relación con respuestas a estrés, ya que el

sustrato que utilizan para catalizar la reacción está implicado en la biosíntesis del ácido indólico (Bak y col.,2011). En la palmera de aceite africana se seleccionó un clon, detectado en raíz y tallo, que era homólogo a un citocromo P450 que se caracterizó como un posible enzima detoxificante (Phongdara y col., 2012), pero si puede tener función en el fruto y su implicación en el carácter no ha sido determinada todavía.

El QTL situado en el GL6, QFFb6_9, se encuentra flanqueado a nivel superior e inferior por genes que entre sus anotaciones ontológicas de los procesos dónde participan muestran una relación con el metabolismo de lípidos, siendo los co-localizados en el nivel inferior los que presentaron la misma función. KG186, situado a 72,5cM y a una distancia del QTL de -0.8cM, codifica una proteína caracterizada como 3-ketoacil-CoA sintasa 4 (KCS4) la cual participa en la biosíntesis de ácidos grasos de cadena larga (>18C) (http://www.uniprot.org/uniprot/Q9LN49). Su especificidad de sustrato determina la longitud y el grado de insaturación de los productos de sus reacción (Lassner y col., 1996; Millar and Kunst, 1997; Millar y col., 1998; Cahoon y col., 2000). Estos ácidos grasos están presentes en las ceras cuticulares y en el aceite de las semillas de algunas plantas (Todd y col., 1999; Han y col., 2001; Li-Beisson y col., 2013) , como en nuestra especie dónde abundan los ácidos grasos de cadena larga como ácido oléico (C18:1) y linoléico (C18:2). Los genes candidatos co-localizados con el mismo QTL, pero a una distancia de 0.1 y 0.2 cM fueron KG187 y KG188 los cuáles codifican la misma proteína calificada como una enzima orizasin aspártico proteinasa cuyo rol biológico no se conoce muy bien, aunque parece que sus funciones están relacionadas con el procesamiento de las proteínas y su degradación bajo determinadas condiciones y estados de desarrollo de la planta, como la organogénesis, la senescencia, las respuestas a situaciones de estrés, la muerte celular programada y la reproducción, lo que sugiere que sus funciones sean especializadas (Simões y Faro, 2004). Aunque en sus anotaciones ontológicas, mostradas en la tabla 4.1 (Anexo 4), parece que existe una relación con el metabolismo de lípidos, no se ha encontrado ninguna referencia bibliográfica que confirme este hecho directamente.

En el GL11 se encontraron 4 genes co-localizados con el mismo QTL, QFFB3_5, que lo flanqueaban por su parte superior a una distancia máxima de -1cM. El gen co-localizado KG191 codifica una proteína insensible a la vernalización (VIN3), relacionada con los tiempos de floración en *Arabidopsis thaliana* y *Oryza sativa*, que ha sido asociada mediante mapeo por asociación con los tiempos de floración en maíz (Khan y Korban, 2014). Esta proteína es un homeodominio de dedos de zinc ("*PHD zinc finger*") que produce cambios en la estructura de la cromatina en los genes relacionados con el tiempo de floración en *Arabidopsis thaliana* (Sung y Amasino, 2004;Sung y col., 2006; Hu y col., 2010). Es posible, por ello, que en *E.guineensis* Jacq. tenga alguna implicación en el desarrollo de los órganos florales del cultivo, y expresados en diferentes estados de su desarrollo (Low y col. 2014). KG192 es un gen candidato que codifica una proteina calificada como un gen de expresión maternal. Estos genes se relacionan con el mecanismo de impronta genética, por el cual algunos genes se expresan de un modo específico dependiendo del sexo del progenitor debido a un cambio epigenético por la metilación del ADN, que produce el silenciamiento del alelo procedente del otro parental. En las

plantas este fenómeno de impronta genética ocurre en el endospermo durante el desarrollo de la semilla como se ha observado en *Arabidopsis thaliana*, maíz o arroz, de ahí que algunos genes de los progenitores queden silenciados (Henderson y Jacobsen, 2007; Hsieh y col., 2009, 2011; Luo y col, 2011; Waters y col., 2011). Por lo que KG191 y KG192 están relacionados con diferentes mecanismos epigenéticos. El gen KG193 codifica una proteina que en *Elaeis guineensis* no ha sido caracterizada pero el análisis mediante B2GO muestra una alta homología con un dominio repetido ankirina y un dominio KH. Los estudios referentes a estas proteínas en plantas muestran una función relacionada con mecanismos de defensa, en el crecimiento y desarrollo (Vo y col., 2015). En *Medicago trunculata* se ha asociado con procesos relacionados con la iniciación de la transcripción, el transporte de iones y la traducción de señales (Zhang y col. 2013), los cuáles también pueden implicar cambios en las funciones anteriormente señaladas. Por último, el gen candidato KG194 codifica una proteína Citocromo B5 isoforma E que en las plantas superiores, animales y hongos proporciona los electrones necesarios para la reacción de desaturación de acil-CoA de los ácidos grasos, y mediando en la formación de ácidos grasos poliinsaturados o PUFAS (Nappier y col., 1997; Kumar y col., 2012). Además también se ha implicado en las plantas superiores como donante de electrones en la hidroxilación de ácidos grasos, en la formación de triples enlaces, en la hidroxilación y saturación de los esfingolípidos de cadena larga, en la desaturación de los esteroles y en las reacciones mediadas por el citocromo P450 (Smith y col., 1992; Rahier y col., 1997; Lee y col., 1998; de Vetten y col., 1999; Broadwater y col., 2002; Nappier y col., 2003; Kumar y col., 2006; Nam y Kappock, 2007; Nagano y col., 2009). También se ha visto que este complejo interacciona con el retículo endomplasmático de las células aumentando la afinidad en la membrana plasmática de los transportadores de sacarosa y sorbitol por sus sustratos, mecanismo necesario para regular las necesidades de azúcares en las células (Fan y col., 2009), y por tanto el flujo de energía de energía que reciben desde el floema para el crecimiento y desarrollo de la planta (Ainsworth y Bush, 2011). Por tanto, puede estar implicado en el carácter FFB así como también podría estar implicado en otro carácter que es el índice de iodo (IV), ya que en frutos caracterizados como de alto rendimiento se ha observado un aumento de la proteína durante el proceso de maduración, mientras que en los frutos de bajo rendimiento no (Loei y col., 2013).El gen candidato KG179, co-localizado con un QTL del GL1 como puede revisarse en el Anexo 4 (Tabla 4.1), codifica una proteína caracterizada como un enzima 2-hidroxiacil -CoA liasa, el cual participa en las reacciones llevadas a cabo en los peroxisomas de la célula, relacionadas con el metabolismo de lípidos, concretamente en la α oxidación del ácido fitánico (Watkins y Ellis, 2012), aunque su papel no está claramente definido en las plantas.

En el GL2 se encontró colocalizado el gen candidato KG180 que codifica una proteína F- box "*tubby like*". Diferentes autores han estudiado la funcionalidad de estas proteínas en *Arabidopsis thaliana* y en *Oryza sativa*, y parecen desempeñar diferentes roles en el crecimiento y en la adaptación de las plantas a diferentes tipos de estrés (Lai y col.,2004; Kou y col., 2009; Bao y col., 2014).

Como puede observarse en la tabla 3.1 (Anexo 3), en el GL8 se encontraron dos QTL's relacionados con el carácter. QFFB3_5 situado a 17 cM presento un gen co-localizado, KG189 que codificaba la misma proteina que el gen candidato KG201 co-localizado con un QTL del carácter BN. Esta coincidencia puede ser debida a la fuerte correlación fenotípica positiva que presentan estos caracteres (Jeennor y Volkaert, 2014), pudiendo estar implicado la misma familia de genes en ambos caracteres. KG190, co-localizado con QFFB3_5 situado a 1,2 cM, es un gen que codifica una proteína que es un enzima denominado ubiquitina E3 ligasa RGLG2, cuyas principales funciones se desempeñan en la regulación (Stone y col., 2005), y señalización celular (Yin y col., 2007). Se ha observado en *Arabidopsis* que los mutantes de estas enzimas pueden alterar la expresión de citoquinas y auxinas, hormonas reguladoras del desarrollo de la planta. Además participan en la regulación de algunos genes relacionados con respuestas a estrés abiótico como estrés salino y sequia (Wang y col.2011; Cheng y col. 2012; Kim y Kim, 2013).

Por último KG195, co-localizado con un QFFB6_9, en el GL13 codifica un enzima peptidil-prolil-cis-trans isomerasa que interacciona con NIMA 4. Estas enzimas pueden modificar la estructura de las proteínas en el núcleo celular, y participan en la regulación génica durante la transcripción del ARNm y a nivel de cromatina (Dilworth y col., 2011). En la revisión realizada por Hanes (2015) de esta familia de enzimas en células eucariotas se muestra como estas enzimas pueden participar en el desarrollo, la floración y en la respuesta a estrés ambiental de la planta.

El último carácter de los componentes de producción para el que se han buscado QTL's en nuestro mapa es producción de aceite en toneladas por hectarea y por año (**PO**). Para él se han encontrado dos QTLs, QPO3_5 y QPO6_9. Billotte y col. (2010) mostraron en su mapa que estos QTL's una fuerte correlación positiva de estos QTL's con los relacionados con FFB, QFFB3_5 Y QFFB3_6. En el anexo 3 (Tabla 3.1) se muestra como el QTL, QPO3_5, está flanqueando por un gen en la parte superior a una distancia de -0,9cM. Este gen KG258 codifica una proteína que es un enzima málico dependiente de NADH que participa en la glucolisis (Planxton,1996; Drincovich y col., 2001). En las plantas con metabolismo fotosintético C3, este enzima localizado en el citosol parece participar en el mecanismo de defensa de la planta , y en los frutos participa en la respiración durante la maduración. En los plastidios de algunos frutos parece implicado en el metabolismo de lípidos (Kilaru y col., 2015), contribuyendo al desarrollo del embrión por ejemplo en colza (Kang y Rawsthorne, 1994) o en el desarrollo del endospermo del fruto del ricino (Smith y col., 1992). Loei y col. (2013) han observado que durante la maduración del fruto de la palmera de aceite muchos enzimas glucolíticos están sobrerregulados para conseguir más cantidad de NADPH y piruvato como fuente alternativa de energía que será utilizada en la síntesis de ácidos grasos. KG259 que se co-localiza a una distancia +0,5cM del mismo QTL codifica un factor de transcripción MADS-BOX3, que participa probablemente en el desarrollo de los orgános florales, actuando dentro del modelo de desarrollo como una proteína de clase C o AGAMOUS (http://www.uniprot.org/uniprot/Q40704), por lo que su función sea probablemente similar al gen candidato KG196, co-localizado con QBN3_5 en el GL2.

Los genes co-localizados en el GL15 flanqueaban al QTL, QPO3_5, por su parte superior con una distancia máxima de -0.8cM y por su parte inferior de +1cM. KG262 codifica una misma proteína que KG195, co-localizado con QFFB6_9 del GL13. Desde el punto de vista de la funcionalidad en relación al contenido de aceite en el fruto no existe ninguna literatura que lo refiera en *E.guineensis*, ni tampoco en ninguna otra especie modelo o no, aunque sí en la respuesta a situaciones de estrés ambiental de la planta dónde este carácter puede presentar una mayor variación del contenido de aceite en función de las condiciones climáticas (Corley y Tinker, 2003). KG263 es un gen que codifica un enzima cistationina ϒ sintasa que participa en la síntesis del aminoácido metionina (Met) precursor de múltiples metabolitos que regulan el crecimiento y la respuesta al medioambiente (Amir y col., 2002). El gen candidato KG264 que codifica una proteína proteosoma subunidad α-5 pertenece al complejo20S y esta a su vez a 26S (Kurepa y col.,2009). Es un enzima del complejo de las proteinasas con actividad dependiente de ATP (http://www.uniprot.org/uniprot/Q42134). Diferentes estudios han confirmado sus funciones en la división y expansión celular, así como en la tolerancia al estrés oxidativo y durante la senescencia de la planta (Kurepa y Samlle, 2008; Kurepa y col ., 2009). En palmera de aceite se expresan en los órganos reproductivos (Xia y col., 2014; Ho y col., 2015) pero no hay una evidencia de su participación en el contenido de aceite de los frutos. El último gen co-localizado que en este GL es KG265 que codifica a una enzima 1 fosfatidilinositol-3-fosfato kinasa 5, FAB1B. Estas enzimas forman parte de la ruta biosintética de fosfoinositidos, los cuales son lípidos con mecanismos de regulación. Se encuentran en el interior de las membranas celulares donde ejercen mecanismos de reclutamiento de señales, regulan las funciones de las proteínas de membrana o son precursores de metabolitos secundarios, por lo que influyen en multitud de procesos celulares fundamentales para el desarrollo y funcionamiento de la planta (Anderson y col., 1999; Falasca y Maffucci, 2009; Heiman, 2016).

Para este mismo carácter en el GL1 se encontró un QTL a una distancia de 92,9cM, del que se seleccionaron 3 genes co-localizados por su parte inferior a una distancia máxima de 0,7cM del QTL. El gen co-localizado KG255 codifica una proteína que no está caracterizada, y por tanto su función es desconocida. KG256 es un gen candidato cuya proteína es cicloartenol-C-24-metiltransferasa 1 implicada en la biosíntesis de esteroles (http://www.uniprot.org/uniprot/Q9LM02), los cuales forman parte de la fracción insaponificable del aceite de palma en un bajo porcentaje (Sambanthamurthi y col., 2000), y son precursores de hormonas esteroideas en las plantas necesarias para el desarrollo y la fertilidad (Shamsi y col., 2012). El gen candidato KG257 codifica una proteína no caracteriza en *E.guineensis*, pero el resultado arrojado por blastx en el software B2GO la caracteriza como una proteína de membrana CP5 que participa en cebada se expresan constitutivamente en los genotipos tolerante a sequía (Guo y col., 2008). Este gen codifica una proteína de membrana con un dominio de unión de lípidos START, relacionada con proteínas reguladoras esteroidogénicas, que regula las vías de señalización intracelular de lípidos (Pointing y Aravind, 1999).

El gen co-localizado KG260 con el QTL QPO3_5 del GL3 a 51,9cM codifica una proteína relacionada con la biosíntesis de bases de cadena larga (LCB2a). Esta proteína es una de las subunidades

del enzima serina palmitoiltransferasa, la cual forma parte del metabolismo de esfingolípidos (http://www.uniprot.org/uniprot/Q2R3K3; Teng y col., 2008). En *A. thaliana*, Teng y col.(2008) concluyeron que los genes LCB1 y LCB2, eran esenciales para el desarrollo y la viabilidad del polen, ese mismo año Dietrich y col. confirmaron la implicación de LCB2 en la viabilidad de los gametofitos en la misma especie. Las revisiones realizadas hasta la fecha confirman que los esfingolípidos ejercen funciones esenciales en las membranas celulares y como moléculas de señalización participantes en la diferenciación y proliferación celular, en los mecanismos de apoptosis y en los mecanismos de respuesta a estrés (Sperling y Heinz, 2003; Markham y col., 2013).

Por último, el gen KG261 co-localizado con dos QTL en la misma posición codifica una proteina GEM8, no se obtuvo ninguna anotación ontológica por lo que su funcionalidad se puede caracterizar unicamente mediante las pocas referencias bibliográficas existentes de este gen en plantas. Estas proteínas son similares a las proteínas *geminin* en mamíferos, y en plantas controlan el paso entre la proliferación y la diferenciación celular, y sus funciones parecen relevantes en cuanto a la adquisición y mantenimiento de la organización multicelular (Caro y Gutierrez, 2007).

5.2.1.2. Caracteres relacionados con componentes de racimo

En el carácter **Fw** o peso del fruto (g) se co-localizaron dos genes candidatos, siendo el de más relavancia KG146, co-localizado a una distancia de -0.3cM del QTL QFwt_1 pero dónde se encuentran en la misma posición 3 QTL relacionados con un componente de la producción que es Bw. Por lo que su funcionalidad ha sido discutida en el apartado anterior. KG11 es un gen candidato, cuya secuencia pertenece a un EST obtenido por Bourgis y col. (2011) en el mesocarpo de la palmera de aceite, lo que es un indicativo de su posible funcionalidad. Codifica una proteína de unión poliadenilato RBP47B', la cual pertenece a un grupo de proteínas de unión al ARNm que se encuentran en el citoplasma de la célula, y posee componentes para el inicio de la traducción de proteínas. Este grupo de proteínas se producen en respuestas a situaciones de estrés celular o a de bloqueo del inicio de la traducción celular (Peal y col., 2011; Muench y col., 2012).

Para el carácter indicador de la calidad del aceite obtenido **IV** o indice de iodo se obtuvieron 3 genes candidatos co-localizados. KG1, es una secuencia cuyo origen es un microsatélite que se encuentra en la mismo posición que el QTL QI_a en el GL3. Aunque los microsatélites normalmente no son genes *per se*, en este caso se obtuvo una homología con una proteína no caracterizada en *E.guineensis* Jacq. de ahí su selección. KG135, procede de un EST obtenido por Bourgis y col. (2011) del mesocarpo de la palmera de aceite, y se encuentra en la misma posición que el QTL (58,9cM, GL1) por lo que puede tener una importancia relevante para el carácter. Este gen codifica la misma proteína que KG179, co-localizado con un QTL de FFB, y cuya secuencia procede de un contig secuenciado y se encuentra en la misma posición, por lo que puede ser que este EST sea parte del mismo gen. Como se ha discutido en el apartado anterior el gen que codifica una enzima hidroxiacil CoA liasa que participa en el metabolismo de lípidos en el peroxisoma y parecen implicados en los mecanismos de fotorrespiración

de las plantas con metabolismo C3. El último gen candidato co-localizado para este carácter KG12, también procede de un EST publicado por Bourgis y col. (2011) del mesocarpo de la palmera de aceite durante su maduración, se encuentra co-localizado con dos QTL's en la misma posición, QI-j y QLt_W_e, éste último relacionado con componentes vegetativos de la planta. El resultado arrojado por BLASTx en B2Go muestra que codifica una proteína de la superfamilia NAD(P) unida a un anillo Rosmman, estos anillos son motivos estructurales a los que se unen diferentes co-factores enzimáticos, en este caso NAD(P)(Bhattacharyya y col., 2012), y los cuáles participan en multitud de procesos metabólicos (http://www.ebi.ac.uk/interpro/entry/IPR016040), como por ejemplo en los que participan las enzimas aldehído deshidrogenasas, sobreexpresadas en condiciones de estrés abiótico (Kirch y col., 2004; Strommer, 2011).

Para el carácter **PF** o proporción de pulpa frente a la proporción de fruto se obtuvo un único gen candidato, KG138, que como puede revisarse en el anexo 3 (Tabla 3.2) estaba co-localizado con PF_1 en el GL6. Su secuencia pertenece también a un EST expresado durante la maduración del mesocarpio de la palmera publicado por Bourgis y col. (2011), y aunque en *E.guinensis* aparece como una proteína no caracterizada, los resultados arrojados por B2GO muestran que este gen codifica una proteína reguladora de la adiposidad (*SEIPIN*). Este gen está bien descrito en humanos, donde algunas mutaciones en él mismo son causantes de una lipodistrofia congénita, en ratones, en *Drosophila melanogaster* y en levaduras dónde juegan un importante papel en la formación de los cuerpo lipídicos (Magré y col., 2001; Cui y col., 2011, 2012; Tian y col., 2011; Cartwright and Goodman, 2012; Bi y col., 2014). Los cuerpos lipídicos son orgánulos celulares presentes en la mayoria de las células eucariotas (Chapman y col., 2012; Murphy , 2012). Su principal función es el almacenamiento de las reservas de carbono de la célula mediante el secuestro de los lípidos neutros del citosol, aunque también se han identificado funcionalidades relacionadas con la señalización de lípidos, entre otras (Goodman, 2008; Murphy y col , 2009; Bozza y col., 2011; van der Schoot y col., 2011; Chapman y col., 2012; Saka y Valdivia 2012; Zechner y col., 2012). En las plantas superiores, las proteínas que forman los cuerpos lipídicos son numerosas, por ejemplo las oleosinas, pero se han caracterizado mayoritariamente en los tejidos de las semillas y órganos florales (Huang, 1992; Tzen and Huang, 1992;Frandsen y col., 2001;Lin y col., 2005;Huang y col., 2013). Pero, como ha revisado Horn y col. (2013) están ausentes en la mayoría de los tejidos vegetativos y en los frutos, incluidos en el mesocarpio de frutos oleaginosos como el de la palmera de aceite o el del olivo. *Seipin* se ha identificado en *A. thaliana* (Chapman y col., 2012) y se ha comprobado que aumenta los niveles de TAG ,y el tamaño de los cuerpos lipídicos en las hojas y semillas, y se obtuvo un aumento del 10% en el contenido de aceite de las semillas de las plantas salvajes. La supresión de este gen mediante ARN interferente produjo el fenotipo contrario (Cai y col, 2015). La presencia de este gen durante el desarrollo del mesocarpio de *E.guineensis* Jacq. detectada por el EST colocalizado con el QTL PF sugiere su implicación en la formación de los cuerpos lipídicos en el mesocarpio del fruto, y en la acumulación de los TAG en él. Aunque se necesitan estudios de expresión y funcionalidad para determinarlo, una vez obtenidos los resultados de nuestro mapeo por asociación.

El último carácter de este grupo es la proporción de aceite obtenida en relación al tamaño del mesocarpio (**POP**) mostró 8 genes candidatos co-localizados. Como puede revisarse en los resultados y en la tabla 3.2 (Anexo 3), K220 y KG221 flanqueaban al QTL, Q%POP_1C por su parte superior en inferior a una distancia máxima de él de -0,5cM y +0,8cM. KG220 codifica un proteína que es una enzima cisteína sintasa que participa en la ruta metabólica del aminoácido cisteína para la asimilación del azufre inorgánico por la planta (Witrz y col., 2004; Droux,.2004). Además, este aminoácido es precursor de moléculas orgánicas que contengan azufre reducido, como el glutation implicado en el control redox y en mecanismos de respuesta a estrés, o en hormonas y co-factores como el acetil CoA (Hoefgen y Hesse, 2008), el cuál es precursor en la síntesis de ácidos grasos en las célula. El gen candidato KG221 codifica una proteína caracterizada como 6 fosfofructoquinasa 3 ATP dependiente (PFK), enzima participante en la glucolisis que tiene lugar en el citosol y en los plastidios celulares. Bourgis y col. (2011) observaron un aumento en la expresión de los enzimas de los plastidios que participan en la glucolisis, lo que implicaba la existencia de mayores tasas de piruvato disponible para la síntesis de ácidos grasos. Los EST's que pertenencian a los genes que codificaban a esta enzima se mostraron elevados durante la maduración del mesocarpio de la palmera de aceite, y concluyeron que una mayor producción de aceite se correlacionaba fuertemente con aumentos temporales de transcriptos relacionados con enzimas participantes en la síntesis de ácidos grasos, transportadores de plastidios y enzimas glucolíticas como PFK, así como un factor de transcripción similar a WRINKLED1. Dussert y col. 2013, confirmaron que la glicolisis via PFK era más activa en el mesocarpio de la palmera que en el ensorpermo y en su semilla, dónde parecía estar más activa la ruta dependiente de fosfato.

Los genes co-localizados con el QTL Q%POP del G 3 a 12,9cM están también co-localizados con un QTL en la misma posición QPO3_5. KG222 y KG223 codifican los mismos genes que KG258 y KG259,respectivamente. Estos se encuentran co-localizados con el carácter PO, en la misma posición.

KG226 Y KG227 flanquean Q%POP_b situado a 86,4 cM en el GL7, con una distancia máxima de -0,4 y +0,7cM respectivamente. El gen co-localizado KG226 codifica una proteína caracterizada como Proteasoma 26S no ATPasa subunidad reguladora 1 homólogo A, unidad que participa en la degradación de proteínas por la ruta de la ubiquitina, siendo la ruta más abundante en plantas (http://www.uniprot.org/uniprot/O48844; Smalle y Viestra, 2004), y está implicada en la regulación de procesos celulares esenciales como el ciclo celular, mecanismos de defensa, transcripción y transducción de señales (Hershko,1998). KG227 se caracterizó como un gen que codifica una proteína rodanesa. Estas proteínas pertenecen a una superfamilia de proteínas versátiles que llevan a cabo múltiples funciones celulares como la resistencia a estrés ambiental y reacciones celulares relacionadas con el metabolismo del azufre y el ciclo celular. En plantas algunas de estas proteínas se han asociado a los mecanismos de senescencia de las hojas (Cipollone y col., 2007), y tienen un papel importante en la biogénesis de los plastidios (Majeran y col., 2012).

KG224 y KG225 se co-localizaban con Q%POP_c en el GL7, que a su vez estaba co-localizado con otro QTL en la misma posición relacionado con el carácter PO correspondiente a componentes de la

producción. KG224 codifica una subunidad de un importador mitocondrial TOM20. TOM20 forma parte de un complejo de translocasas de la membrana mitocondrial externa que facilita el reconocimiento de proteínas precursoras y su translocación a través de la membrana externa de la mitocondria (Taylor y Pfanner, 2004). En *A.thaliana* se ha concluido que TOM20 funciona como receptor que importa proteínas precursoras a la mitocondria y parece que es necesaria para la importación de ARN transferente junto con los canales VDAC (Salinas y col., 2006; Lister y col., 2007; Homblé y col., 2012). KG225 es un gen que codifica una proteína caracterizada como proteína ribosómica 60S L7a, encargada de la síntesis de proteínas a partir de ARN mensajero al igual que KG201 para el carácter BN.

5.2.1.3. Caracteres relacionados con componentes vegetativos

El principal carácter para este grupo es la **altura de tallo**, importante en la palmera de aceite debido a su elevada altura que hace difícil la identificación y recolección de los frutos maduros. Los QTL's revisados en el mapa, como muestran los resultados y la tabla 3.3 (Anexo 3), están relacionados con el incremento de la altura (Ht) y el crecimiento del tallo (StGr). Lee y col. (2015) publicaron un mapa de ligamiento donde identificaban QTL relacionados con el incremento de la altura, siendo un QTL del GL5 el que mayor varianza explicaba, y Billote y col. (2010) en su mapeo de QTLs en un mapa de ligamiento de múltiples parentales también mapeo algunos QTLs relacionados con el incremento de altura en el GL15. En nuestro mapa, no se han detectado genes candidato co-localizados con QTL situados en esos grupos de ligamiento, a pesar de estar integrados ambos mapas en él. Posiblemente, porque si existían QTL's posicionados con el carácter en estos grupos de ligamiento no se encontró sentido biológico en los genes que los flanqueaban.

En el anexo 3 (Tabla 3.3) se muestra que en el grupo de ligamiento 1 fueron 3 los genes que flanqueban al QTL Qht situado a 93,9cM, y todos lo hacían por encima del mismo a una distancia máxima de -0,5cM. KG161 se caracterizó como un gen que codifica una cadena catalítica ferrodoxin-tiorredoxin reductasa (FTR), este gen forma parte del sistema de proteínas de los cloroplastos que detecta los ambos de potenciales redox y transforma la señal electrónica en una señal bioquímica en el estroma (Schürman, 2003; Buchanan y col, 2005). Estos potenciales redox se originan a causa de los cambios de luz que percibe las plantas durante el proceso de fotosíntesis, para adaptarse en las variaciones de los ciclos luz-oscuridad (Dai y col.,2004). Este complejo está presente en las enzimas diana encargadas de la asimilación fotosintética del carbono como el complejo NASP-MDH (EC.1.1.1.82) que en plantas C3, como *E.guineensis* Jacq., no está implicado directamente en la fijación del carbono, y dos enzimas del ciclo de Calvin, la enzima fructosa-1,6-bifosfatasa y la enzima gliceraldehído 3 fosfato deshidrogenasa (GADPH) de los cloroplastos (Dai y col., 2000, 2004). Thormählen y col. (2015) han demostrado que este sistema participa en la regulación del ciclo de Calvin-Benson , el metabolismo del almidón y en la regulación del crecimiento de la planta bajo diferente condiciones de luz y coordinado con otro sistema basado en el NADPH relacionado con el metabolismo de los azúcares en la oscuridad. El gen candidato KG162 codifica una proteína caracterizada como GDP-manosa 3,5 epimerasa. Este gen

altamente conservado en plantas (Wolucka y col., 2003) participa la biosíntesis de ácido ascórbico (vitamina C) que en las plantas actúa como cofactor enzimático y como antioxidante, y en la biosíntesis de los polisacáridos que conforman la pared celular. El ácido ascórbico participa en procesos fisiológicos fundamentales como son la biosíntesis de la pared celular, de fitohormonas, y de metabolitos secundarios. Además también participa en la división celular, el crecimiento, la resitencia a estrés y en mecanismos fotoprotectores (Smirnoff y Wheeler, 2000). La inactivación parcial de este enzima provoca defectos en el crecimiento de la planta que afectan a la división y expansión celular (Gilbert y col., 2009). En tomate, se observó que mutantes con bajos niveles de estas enzimas y por tanto disminución de ácido ascórbico afectaba a la formación de las células y a una disminución del crecimiento (Voxeur y col., 2011). Además, por otro lado, Senn y col. (2016) demuestran que el ácido ascórbico sintetizado por las mitocondrias participa en la regulación de la fotosíntesis mediante procesos específicos. Por tanto ambos genes pueden estar relacionados, aunque no se tiene constancia como puede afectar directa o indirectamente sobre la altura de la planta, únicamente sobre su crecimiento en general. El último gen co-localizado (KG145) en esta posición procede de un EST secuenciado en las hojas y en el mesocarpio de la palmera de aceite por Bourgis y col. (2011), y codifica una proteína que es una subunidad 8 catalítica de una celulosa sintasa (CesA8), la cuál participa en la formación de la pared celular secundaria (Endler y Persson, 2011). En la revisión realizada por Carpita y McCan en 2015 se recoje que en plantas con mutaciones en los genes que codifican estas celulosas sintasas se han caracterizado por un colapso en los vasos del xilema y con menores niveles de celulosa que las plantas salvajes, y una pared celular más delgada. En cambio, los fenotipos enanos se han mostrado en genotipos mutantes de genes relacionados con la síntesis de celulosa en la pared celular primaria.

KG163 co-localizado con QHt (GL1, 95,6cM) codifica un proteína caracterizada como UDP-arabinopiranosa mutasa 1. Las mutasa pertenecen a una pequeña familia de genes que codifican proteínas glicosiladas,RGP1, específicas de plantas que están altamente conservadas en el citosol de la célula y tienden a asociarse con las membranas del aparato de Golgi (Rautengarten y col., 2011). Estos genes están implicados en la biosíntesis de polisacáridos no celulósicos de la pared celular como demostraron Zhao y Liu (2002). También parecen participar en la respuesta a estrés biótico (Shelth y col., 2006) y en el desarrollo del polen durante la mitosis afectando a la divisón celular o a la integridad de las vacuolas (Drakakaki y col., 2006). Se ha demostrado la importancia del gen que codifica este enzima en diferentes experimentos demostrando su participación en el desarrollo de la planta donde diferentes mutaciones han mostrado fenotipos enanos, retrasos en el crecimiento e infertilidad en *Arabidopsis thaliana* y en *Oryza sativa* (Rautengarten y col., 2011; Konishi y col., 2011).

En los resultados se muestra que en el GL2 se han seleccionado 2 QTLs a diferente altura, y para cada uno 2 genes candidatos. KG164 (QHtC_1; 127,7cM) co-localizado a +0,2cM del QTL codifica una enzima pirofosfato de la bomba de protones de la membrana vacuolar(H^+PPasa). Esta enzima participa en el transporte de solutos dependientes de energía a través del tonoplasto o membrana vacuolar mediante el impulso de protones (Sarafian y col., 1992). Esta bomba de protones participa en la

elongación y diferenciación celular, ya que contribuyen al mantenimiento de la presión osmótica y la expansión vacuolar (Maeshima y col., 1996). También, parece participar en el transporte de auxinas, hormonas presentes durante el desarrollo y crecimiento de la planta, y su sobreexpresión en condiciones de estrés abiótico, concretamente sequía y estrés salino, mejoran el crecimiento de la planta (Gaxiola y col., 2007). KG165 posicionado a la misma distancia que el QTL (127,7cM), codifica una proteina caracterizada como una enzima Isoleucina ARNt ligasa. Estas enzimas son las responsables de la unión del ARN transferente, proceso esencial en todas las células eucariotas, después del corte por una endonucleasa del intrón (Englert y Beier, 2005). Estos autores han sugerido que la funcionalidad en plantas de esta enzima no está sujeta únicamente a los procesos de unión y reparación de ARNt, si no que participan en otros procesos celulares no detectados, e incluso pueden utilizarse como una alternativa para la iniciación de la traducción (Englert y col., 2007; Popow y col., 2012). KG166 y KG167 se encuentran posicionados en el QTL QStGr (136,8cM), los dos a una distancia -0,2cM. El gen KG166 codifica una enzima peroxidasa de clase III (http://www.uniprot.org/uniprot/Q9FJR1). Estos enzimas están implicados en multiples procesos biológicos como la detoxificación del peróxido de hidrógeno, el catabolismo de auxinas, la biosíntesis de lignina durante la formaicón de la pared secundaria celular y la respuestas a diferentes tipos de estrés (Tognolli y col., 2002; Cosio y Dunand, 2009). KG167 codifica una enzima del complejo FTR, al igual que KG161 co-localizado con un QTL del mismo carácter en el GL1, y la cúal participa en la fotosíntesis. Ambos genes, co-localizados en la misma posición en cM, pueden estar relacionados entre ellos, ya que tanto el mecanismos de fotosíntesis como la acción de las peroxidasas generan moléculas muy pequeñas altamente reactivas del oxígeno (ROS) con una actividad intrínseca en diferentes respuestas a estímulos ambientales y procesos implicados en la planta como el estrés oxidativo o la elongación celular (Bolwell y col, 2002 ; Schopfer y col., 2002 ; Delannoy y col., 2003; Liszkay y col., 2004 ; Bindschedler y col., 2006).

El co-localizado KG168 (GL2;QStGr;157,4cM) codifica una proteina asociada a estrés con dominio AN1 y A20 de dedos de zinc. Su funcionalidad parece estar relacionada con los mecanismos de respuestas a estrés abióticos (Vij y Tyagi, 2008; Giri y col. 2013). Aunque parece que tambien pueda estar implicada en la elongación celular mediante una regulación negativa de las giberelinas en *O.sativa* (Liu y col., 2011). Estas hormonas regulan el crecimiento y el desarrollo en procesos como la germinación de la semilla, la elongación del tallo, la expansión de las hojas y el desarrollo reproductivo (Sun y Gubler, 2004).Además, en *Medicago trunculata*, la inhibición de la expresión del gen MtSAP1 en la semilla mostró que estas plantas transgénicas eran más pequeñas en peso y altura que las de tipo salvaje (Gimeno-Gilles y col., 2011)

En el GL3 únicamente se seleccionó un gen candidato, KG169, co-localizado con QHt (43,1cM). Este gen codifica una proteína caracteriza como subunidad MU del complejo AP4, la cuál pertene a una familia de proteínas del complejo adaptador de clatrina. El mecanismo de endocitosis de las plantas depende de esta familia de proteínas. Este mecanismo desempeña un papel fundamental en las funciones celulares, en el crecimiento y desarrollo de las plantas, la señalización hormonal y su

interacción con el medio ambiente (Chen y col., 2011). En *E.guineensis* Jacq, Jeennor y Volkaert (2014) co-localizaron un gen que codifica una proteína de este complejo con un QTL en el GL15 de su mapa relacionado con el número de racimos, pero no encontraron ningún efecto que influyera sobre el carácter en cuestión. También hay que destacar que Ho y col. (2015) presentaron un transcripto que sólo aparecía en las flores femeninas de las palmeras (isotig 31484) que codifica esta misma proteína. Aunque estos dos estudios no revelan nada importante en cuanto a su funcionalidad en relación a la altura de la palmera, si lo hacen en cuanto a su presencia en la especie para la cual estamos buscando genes candidatos.

En el GL4, 2 genes candidatos estaban co-localizados en los misma posición (14cM) a una distancia de -0,1cM del QTL QHt. KG170 codifica una proteína caracterizada como APO2 P700 fotosistema I que participa en la fotosíntesis y por tanto en el metabolismo celular de la planta. KG145 posicionado a la misma distancia del QTL se caracteriza como una proteina mejorada de resistencia (EDR2), que participa en los procesos de respuesta a estrés biótico con capacidad de unirse a lípidos de membrana mitocondrial y activar así la muerte celular programada mediada por la mitocondria (Tang y col., 2005). KG171 co-localizado inferiormente al QTL codifica una posible metiltransferasa, cuya anotación la clasifica dentro de la superfamilia de S-adenosil-L metionina dependiente. La literatura muestra una función relacionada con la modificación de la pectina durante la biosíntesis de la pared celular, y parece que es esencial para el la adhesión celular y el desarrollo de co-ordinado de la planta (Krupková y col., 2007), aunque son necesarios más estudios para confirmar está función. En el mismo GL pero a 100,6cM se sitúa otro QTL QHt y en el que se ha seleccionado un gen co-localizado a -0,2cM (Anexo 3; Tabla 3.3). KG172 codifica una proteína caracterizada como membrana de unión de esteroides. Su funcionalidad parece relacionada con el metabolismo de hormonas esteroideas en plantas, concretamente las brasinoesteroides que participan en la regulación de múltiples procesos en el desarrollo como la elongación celular, fertilidad, senescencia y fotomorfogénesis, aunque sus receptores no están bien estudiados en las plantas. En *A.thaliana* se ha caracterizado un gen que codifica una proteína de la familia de unión esteroidea y que actúa como regulador negativo de la elongación celular inhibiendo la elongación del hipocotilo en presencia de luz (Yang y col., 2005). Song y col. (2009) confirmaron que este gen actuaba como regulador negativo de los brasinoesteroides, y Shi y col.(2011) su actividad en relación a la fotomorfogénesis durante los ciclos de luz y oscuridad.

En el GL7, los genes candidatos co-localizados KG173 y KG174 codifican una misma proteína caracteriazada como como subunidad TAP46 reguladora PP2A, ambos genes se encuentran muy próximos entre sí por lo que es probable que en esta región estén posicionados genes de esta misma familia de proteínas, o bien las secuencias codificantes sean regiones diferentes del mismo gen. De hecho las proteínas PP2A pertenecen a la familia de enzimas fosfoproteina fosfatasa y cuyas subunidades se han identificado y conocido su mecanismo regulatorio (Uhrig y col., 2013). El silenciamiento de genes en *Arabidopsis thaliana* ha demostrado que TAP46, subunidad reguladora de PP2A y sustrato de la reacción llevada a cabo por TOR (rapamicina proteina quinasa), es crucial para el

crecimiento y la supervivencia celular, el mecanismo de autofagia y la síntesis de proteínas (Ahn y col., 2011). Ahn y col (2014) demostraron que la sobreexpresión de este gen induce la modulación de la tanscripción de genes implicados en el metabolismo de nitrógeno, la biogenésis de ribosomas y la biosíntesis de lignina, sugiriendo además que TAP46 modula el crecimiento de las plantas, y la vía de regulación mediada por TAP46 es independiente a la regulación del tamaño del órgano.

KG175 y KG176 se co-localizaron a nivel inferior del QTL QHt (Chr8; 69,8cM). El gen co-localizado KG175 codifica una proteina caracterizada como LORELEI anclada a GPI implicadas en la fertilización y presente en el gametofito femenino (http://www.uniprot.org/uniprot/B3GS44; Capron y col., 2008; Tsukamoto y col., 2010). Este gen no parece tener una relación directa o indirecta con la altura de la palmera, y no hay información disponible que lo pueda asociar. Pero se ha seleccionado por su posible implicación en otros caracteres de importancia como FFB, FN o BN que pueden depender del número de inflorescencias femeninas, entre otras. El gen co-localizado KG176 se ha caracterizado como una proteína que contiene un dominio repetitivo WD. Los resultados de anotación no mostraron ninguna anotación ontológica, y según la anotación del gen en NCBI (http://www.ncbi.nlm.nih.gov/gene/?term=xp_010928199) se caracteriza dentro de la familia de dominio WD40. Este dominio que aparece en multitud de proteínas eucariotas con funciones diversas como reguladores de la señal de transducción, el procesamiento del pre-ARNm ARNm y ensamblaje del citoesqueleto (http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=271593).

KG177 (GL10;QHt;90,4cM) se caracteriza por codificar el dominio de una proteína AGENET homóloga a un dominio de Bromo adyacente (http://www.uniprot.org/uniprot/Q8L7Q0). En *E.guineensis* Jacq es una proteína no caracterizada LOC105053217, desconociendo su funcionalidad, aunque este dominio se ha encontrado en reguladores de la cromatina y en motivos de reconocimiento de ARN (RRM) (Cousthman y col., 2014) por lo que puede estar implicado en el control epigenético. El gen co-localizado con QHt (66,6cM) en el GL11 codifica una subunidad del complejo THO que parece desempeñar un papel importante en el sistema inmune innato de la planta (http://www.uniprot.org/uniprot/Q93VM9), en el silenciamiento de ARN (Yelina y col., 2010), así como en la generación de ribonucleoproteínas mensajeras funcionales (Furumizu y col., 2010).

En el GL12 se seleccionó un gen co-localizado con dos QTL's relacionados con componentes vegetativos en la misma posición, uno de los cuales estaba relacionado con un carácter no sujeto a estudio en esta tesis, que es la anchura media del peciolo de la hoja (P_w_F). La secuencia de este gen procede de los EST's secuenciados por Bourgis y col. (2011) en el mesocarpio del fruto (M01000023551) y codifica una proteína que es una enzima manano sintasa que pertenece a la familia de genes celulosa sintasas que participan en la formación de la pared celular en las células vegetales. Las manano sintasas contribuyen al crecimiento de la pared celular primaria mediante la síntesis de las hemicelulosas manano y glucomanano.

5.3. Genes candidatos conocidos

La última herramienta para la búsqueda de genes candidatos fue la genómica comparativa mediante la búsqueda de genes candidatos en la información existente de genes que pudieran estar relacionados con los caracteres de interés agronómico que se quieren estudiar en esta tesis. Asumiendo que está técnica, sobre todo para caracteres complejos, no siempre es efectiva por las diferencias biológicas entre especies, ya que ocasiona una heterogeneidad genética o diferencias relativas al proceso de evolución (Zhu y Zhao, 2007). Esta herramienta es útil cuando los genes proceden de la misma especie o de especies relacionadas filogenéticamente (genes ortólogos), o cuando se seleccionan regiones altamente conservadas de esos genes entre especies no relacionadas. De los 119 genes seleccionados como candidatos 80 son genes presentes en el género Elaeis, 19 de la especie modelo por excelencia en plantas *Arabidopsis thaliana*, y el resto de las secuencias de genes proceden de especies de cultivos oleaginosos como *Olea europeae* o *Jatropha curcas*, o de especies que forman racimos con frutos como *Vitis vinifera*. Para determinar la funcionalidad de estos genes en nuestra especie se utilizó el método BLASTx basado en comparaciones de pares de secuencias para determinar su homología con *E. guineensis* Jacq, asumiendo el posible error que puede arrojar esta metodología de que el resultado obtenido no siempre es el más cercano filogenéticamente hablando (Koski y Golding, 2001). Esto es debido a que en esta metodología sólo se tiene en cuenta el mejor hit en la especie de interés, en este caso *E.guineensis* Jacq., o de una especie de la misma familia, como E. oleifera o cercana filogenéticamente como *P.dactylifera*. Los resultados obtenidos mediante este análisis mostró una alta homología de las secuencias en nuestra especie, no inferior a un e-valor 1E- 20, que junto con las evidencias obtenidas en la bibliografía, en la inclusión en patentes públicas y por las anotaciones ontológicas mostradas, fueron suficientes para su selección. No se han publicados estudios similares a este, dónde los caracteres de interés agronómico sujetos a estudio son tan numerosos. Por ejemplo, Fusari (2010) en su tesis doctoral realizada en Girasol utilizó genes candidatos en *A.thaliana* y otros genes que tenían un rol empírico en procesos de defensa, utilizando para determinar la ortología un análisis filogenético de las secuencias. Sharma y Chauhan en 2012 identificaron "in silico" algunos de los genes que participaban en la biosíntesis y acumulación de aceite en semillas de cultivos oleaginosos de cuatro especies Arabidopsis, Brassica, soja y semillas de ricino y mediante genómica comparativa, utilizando el algoritmo ClustalW y BLAST, identificaron las variaciones en las secuencias de esos genes y las asociaron al contenido y composición de aceite en las semillas.

En la tabla 5.1 (Anexo 5) se muestra la agrupación de los genes candidatos seleccionados en relación a su participación en los diferentes procesos fisiológicos en los que participan y su relación hipotética con los diferentes caracteres agronómicos. Los procesos fisiológicos considerados estaban relacionados con el proceso de fotosíntesis, el desarrollo del fruto, su calidad, procesos de crecimiento y desarrollo, elongación de la tallo, factores de transcripción y metabolismo de lípidos, por la implicación que puedan tener en los caracteres de estudio como se ha discutido con anterioridad.

Los genes que se resaltaron en los resultados de este capítulo, apartado 4.3, se seleccionaron por su gran implicación en los procesos de mejora en la especie y con algunos caracteres de interés. KG120 denominado como gen "*Shell*" e identificado por Singh y col. (2013b) es el responsable de las diferentes formas que pueden presentar los frutos, y por tanto del contenido de aceite del mesocarpio como ya se ha explicado en la introducción de esta tesis. El gen "*VIR*" (Singh y col., 2014) es el responsable de la variación del color del fruto a medida que madura, y facilita por tanto la identificación del grado de madurez del fruto. Otros genes importantes previamente identificados fue el gen que codifica la enzima asparagina sintasa (KG233 y KG234) el cuál es responsable de fenotipos enanos en algunas especies de plantas y, por tanto, puede estar ínfimamente relacionado con el carácter HI y el fenotipo deseado en las palmeras de aceite (Lee y col., 2015). Morcillo y col. (2013) identificaron el gen que controlaba la enzima lipasa, que bien no se relaciona con ningún carácter sujeto a estudio, pero si es interesante porque esta enzima, presente en el mesocarpio, hidroliza los triglicéridos aumentando el contenido de ácidos grasos libres y devaluando la calidad del aceite.

En plantas la ruta biosintética de ácidos grasos de *novo* está conservada aunque hay variaciones en cuanto a contenido y composición de ácidos grasos en cada especie (Sharma y Chauhan, 2012), por lo que es una buena elección para la búsqueda de genes candidatos relacionados con la calidad y el rendimiento de aceite en *E.guineensis* Jacq. De hecho, algunos autores como Bourgis y col. (2011), Tranbarger y col. (2011), Dussert y col. (2013) y Teh y col.(2013) la han estudiado a nivel transcriptómico y metabolómico con anterioridad, aportando una buena herramienta para la selección de posibles genes candidatos para los caracteres relacionados como **IV**, **PO** o **POP**.

Los genes implicados en el desarrollo del fruto pueden aportar información relativa a los caracteres FN y FW. Por ejemplo, KG242 es un gen que codifica una proteína receptora de etileno. Estas hormonas en plantas participan en diferentes procesos como el desarrollo y el crecimiento de la planta, o en la maduración del fruto. De hecho, Tranbarger y col. (2011) lo identifica durante la maduración del fruto de la palmera de aceite africana. Los factores de transcripción MADS-Box también pueden utilizarse como posibles genes candidatos debido a que controlan el desarrollo floral, y pueden estar implicados en el desarrollo del fruto (Seymour y col., 2008; Pabón-Mora y col., 2014). Por tanto pueden estar implicados también en el carácter FFB, en el que influye el sex ratio o proporción de inflorescencias femeninas como se ha discutido anteriormente. Adam y col (2007) caracterizaron los genes MADS box que participan en la determinación de la estructura floral en la palmera de aceite africana aportando una buena herramienta de selección de genes candidato, y algunos de ellos participan en la maduración del fruto (Trabanger y col.,2011). Como ejemplo el gen candidato KG123 codifica una proteína caracterizada como factor de transcripción MADS BOX GLO2. Este gen pertenece a la familia GLO/PI que se expresa en el desarrollo de las flores en E.guineensis Jacq. (Adam y col., 2007), pero que en el caso del desarrollo del mesocarpio sólo se expresa uno de ellos EgGLO1, por lo que Trabanger y col. (2011) postulan que es el mimo gen el que participa en durante los dos estados de

desarrollo, floral y el del fruto, o bien se ha producido una duplicación y es ese otro gen divergido el que ha adquirido una nueva función relacionada con el desarrollo del fruto.

En cuanto al grupo que hace referencia a la **elongación del tallo**, proceso que hemos relacionado con el carácter "altura de tallo"(HI), algunos de los genes seleccionados como candidatos mostraban fenotipos enanos, debido a mutaciones en el gen, justificando así su selección debido a que el objetivo es conseguir una disminución en la altura de la palmera, tal y como se ha discutido en apartados anteriores. Es el caso de KG286 gen que codifica una proteína DELLA reguladora negativa de la señalización de las hormonas giberelinas, y que sus mutaciones dominantes ocasionan fenotipos enanos y han sido responsables del aumento del rendimiento en las cosechas de arroz y maíz (Asano y col., 2009; Lawit y col., 2010).

También se han seleccionado genes participantes en el proceso fotosintético por ser el elemento clave de suministro de energía de la planta, y por tanto del resto de procesos metabólicos que ocurren en las plantas. Además de genes que participan en el desarrollo y crecimiento celular, y por tanto de la planta porque podrían estar implicados indirectamente en alguno de los caracteres de estudio. Como ejemplo el gen candidato KG87 es un gen que codifica una proteína relacionada con el metabolismo de las auxinas, hormonas que regulan el crecimiento en las plantas y que influyen en la división, elongación y diferenciación celular, impactando fuertemente sobre la forma final y función de las células y tejidos en las plantas superiores (Ljung, 2013).

Para finalizar esta amplia discusión sobre los genes candidatos seleccionados, es importante destacar que serán los SNP que se presenten en los genotipos de nuestra especie y dentro de los genes específicos los que los hagan genes candidatos reales y no hipotéticos para este estudio de mapeo por asociación, lo que se desarrolla en el siguiente capítulo de esta tesis.

CAPÍTULO 3: SECUENCIACIÓN DE AMPLICONES E IDENTIFICACIÓN DE PATRONES

1. INTRODUCCIÓN

Después de la selección de los genes candidatos, la siguiente etapa del mapeo por asociación es la búsqueda de variaciones genéticas de la población de mapeo en esos genes candidatos (genotipado), para posteriormente buscar asociaciones entre esos cambios del genotipo con la variación fenotípica de los individuos en los caracteres de interés. La identificación de estas variaciones genéticas en los estudios de mapeo por asociación es crucial para el desarrollo de marcadores genéticos, dónde los SNPs e "Indels", término que engloba las inserciones/delecciones que pueden existir a lo largo del genoma, son los marcadores por excelencia, por su abundancia y amplia distribución en los genomas de las células eucariotas (Rafalski, 2002; Fusari y col., 2008; Riju y Arunachalam, 2009; Grattapaglia y col., 2011; Hayward y col., 2012).

Los **SNPs** se definen como las variaciones que ocurren en la secuencia genómica de los individuos de una población o entre pares de cromosomas de un individuo a nivel de un único nucleótido. Se clasifican en tres categorías diferentes: transversión (C/G, A/T, C/A y T/G), transición (C/T o G/A) e inserción/delección (indels). Estos marcadores codominantes pueden localizarse en las regiones codificantes y no codificantes de los genes, y en las regiones intergénicas. Además, los SNPs son marcadores estables, reproducibles y altamente heredables en comparación con otros marcadores, como los AFLPs y SSRs. Aunque los SNPs pueden presentar cuatro posibilidades diferentes, una por cada nucleótido, los estudios han demostrado que sólo son dos las posibilidades que se encuentran en un mismo loci de una población determinada (Brookes, 1999). Esta condición de bialelismo podría ser una desventaja frente a otros marcadores como los SSRs, pero su abundancia y estabilidad en el genoma lo compensa, ya que se ha demostrado que su frecuencia en el genoma de las plantas es muy superior a los SSRs (Kwok y col., 1996; Fusari y col., 2008), determinado como 1SNP por cada 100-300pb (Gupta y col., 2001), proporcionando por tanto una alta densidad de marcadores por ejemplo cerca de una región de interés.

Los SNPs son marcadores directos, representan la diferencia en un único nucleótido entre dos individuos en una región determinada, y la información proporcionada por su secuencia muestra la naturaleza exacta de la variación alélica (Berkman y col., 2012). Esta variación puede influir en el fenotipo mostrado por la planta, como se muestra en la figura 1, ya que los cambios en el genoma pueden alterar las funciones de genes importantes y su expresión, como lo recoge la revisión realizada por Huq y col. (2016). Por tanto, mediante estos marcadores se pueden detectar las diferencias en diferentes loci del genoma o región seleccionada entre un gran número de individuos, no sólo individualmente, sino también mediante los haplotipos que pueden mostrar. Entendiendo por haplotipo en este caso a la combinación de alelos estrechamente ligados al loci que tienden a heredarse juntos (Rafalski, 2002).

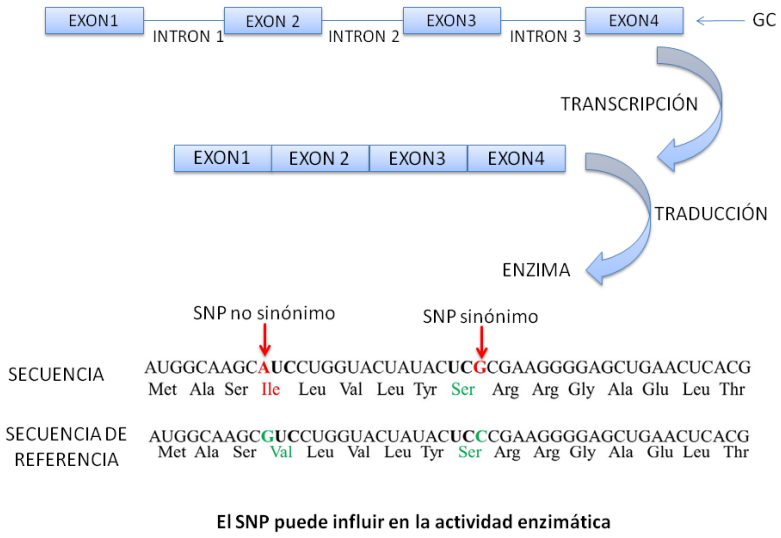


Figura 1: Representación del rol de un SNP sobre la función de un gen que codifica un enzima y como puede modificar un cambio de aminoácidos la actividad del mismo. Met, Metionina; Ala, Alanina; Ser, Serina; Ile, Isoleucina; Leu, Leucina, Val, Valina; Tyr, Tirosina; Arg, Arginina; Gly, Glicina; Glu, Ácido Glutamico y Thr, Treonina. Adaptada de Huq y col. (2015). Los SNPs posicionados en las regiones codificantes no necesariamente cambian la secuencia de aminoácidos de la proteína que se produce, debido a la degeneración del código genético. Se conocen como SNP sinónimos o mutación silenciosa a aquellos SNPs situados en las regiones codificantes que no producen variación en la secuencia polipeptídica, en cambio los que si producen un cambio en la proteína se denominan no-sinónimos. Pero, los SNPs que están fuera de estas regiones codificantes pueden tener consecuencias en las modificaciones post-transcripcionales en las que un único gen puede codificar múltiples proteínas o "gene splicing", en la unión de los factores de transcripción o en la secuencia de ARN no codificante, pudiendo originar cambios en el fenotipo (Riju y Arunachalam, 2009; Patnala y col., 2013).

Los SNPs se aplican también en diferentes cultivos para: 1. la construcción de mapas genéticos de alta densidad y mapeo de QTLs como se presenta en una revisión realizada por Mammadov y col. (2012) en especies como arroz, trigo y soja, entre otras; 2. la integración de mapas físicos y genéticos, ya que permiten ordenar físicamente los "contigs" de los diferentes cromosomas (Rafalski, 2002; Hayward y col. 2012); 3. la caracterización del germoplasma y el estudio de poblaciones (Grattaplagia y col., 2011; Hayward y col., 2012). En el género *Elaeis* que engloba a la palmera de aceite africana y americana son numerosos los estudios realizados con este tipo de marcadores pero con diferentes aplicaciones, como el análisis de la diversidad genética, la identificación de genes como *Shell*, *Virescens*, y *Asparagina sintasa*, el mapeo de QTLs, el mapeo genético de alta densidad y en un estudio de asociación abordado por la genotipificación de marcadores homogéneamente distribuidos en el genoma o "Genome-Wide" (Riju y Arunachalam, 2009; Shing y col., 2013b, 2014; Jeennor y Volkaert, 2013; Pootakham y col.,2013; Ting y col., 2014; Lee y col. 2015; Ong y col., 2015; Teh y col., 2016).

El **genotipado** del conjunto de individuos que componen la población de mapeo puede realizarse mediante la identificación de SNPs *de novo*, bien a lo largo del genoma o bien en regiones específicas, o si estos ya han sido identificados previamente en la población mediantes técnicas de re-secuenciación y análisis bioinformáticos. Las identificación de SNPs *de novo* se lleva a cabo por

diferentes métodos que combinan técnicas de secuenciación masiva con análisis bioinformáticos que permiten la comparación de las secuencias entre diferentes individuos. En la actualidad, las técnicas más utilizadas para el descubrimiento de SNPs *de novo* están basadas en el análisis *in silico* de las numerosas librerías de EST disponibles en diferentes cultivos, junto con su re-secuenciación, o bien en función de la existencia de genoma de referencia en el cultivo o no, en la secuenciación masiva mediante tecnologías de última generación y análisis comparativo con el genoma de referencia o con un genoma en sintenia con la especie (Ganal y col., 2009; Grattapaglia y col., 2011). Las secuencias obtenidas permiten identificar fácilmente los SNPs ya que se obtienen de un pool de ADN de germoplasma diverso con una cobertura del genoma redundante (Grattapaglia y col., 2011). El desarrollo de estos marcadores *de novo* es muy costoso debido a los altos costes de secuenciación y genotipado, por lo que sólo es justificable su aplicación en cultivos con alto valor comercial (Khan y Korban, 2012). En el caso de que los marcadores ya estén desarrollados, se utilizan técnicas como los ensayos Taqman®, ensayos iplex Gold, o el fundido de alta resolución (High Resolution Melt (HRM)) que permiten la identificación multiplex de los SNPs en la población de estudio (Figura 2) (Kumar y col., 2012; Patel y col., 2015).

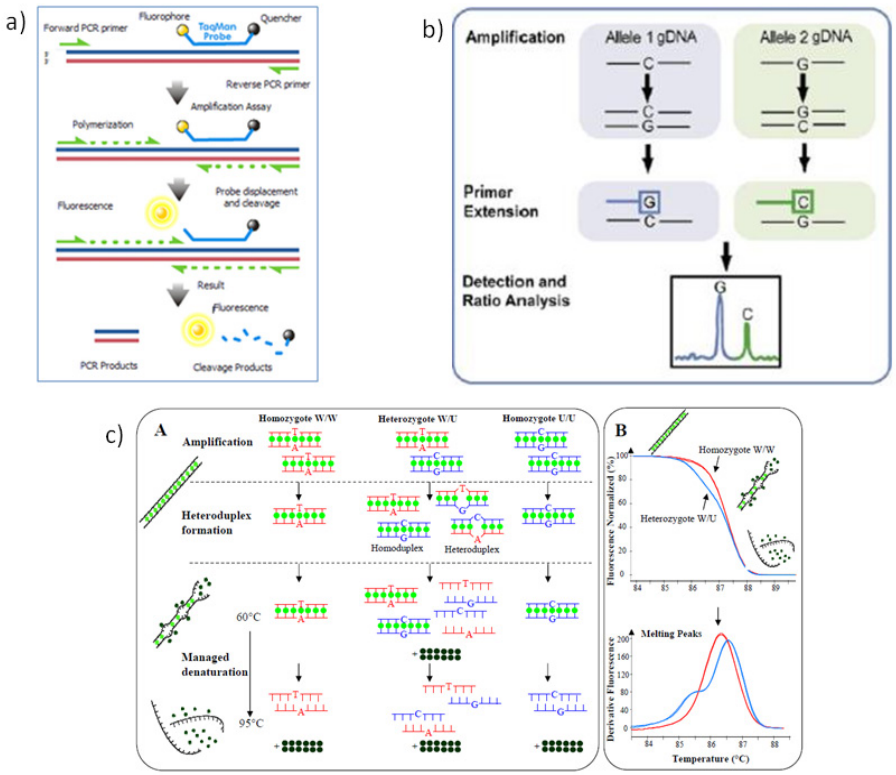


Figura 2: Técnicas para el genotipado de SNP's. a) Ensayos Taqman® basados en sondas alelo específicas marcadas con fluoróforos (recuperado de Braindamaged,2009©) ; b) Ensayos iplex GOLD (Sequenom) basados en la extensión del cebador en una única base y análisis posterior en un espectofotómetro de masas MALDI TOF (Matrix-Assisted Laser Desorption/Ionization Time of Flight) (recuperado de www.biotechniques.com); c) HRM A. Los fragmentos de ADN se amplifican intercalando un fluorescente en el ADN que se desnaturaliza con calor y se enfria, a continuación. Después del enfriamiento se formaron un heterocigoto (W/U), y dos homocigóticos (W/W y U/U). B. Los resultados se muestran en las curvas de fusión durante la desnaturalización. La detección de los SNP se realiza por el cambio de temperatura en la curva de fusión durante la desnaturalización debido a la variación en la temperatura de fusión del cebador por el cambio de base o a la variación de las formas en la curva de fusión cuando hay heteroduplex (recuperado de Meistertzheim y col., 2012).

Una de las técnicas desarrolladas en la última década en cultivos con genomas de gran tamaño y diversidad es el genotipado por secuenciación o GBS (Elshire y col., 2011; He y col., 2014). Mediante esta técnica se descubren nuevos marcadores al mismo tiempo que se genotipa la población, gracias a las tecnologías de secuenciación de segunda y tercera generación o next generation sequencing (NGS). Es un sistema multiplexado para la construcción de librerías de representación reducidas desarrollado por Illumina NGS (Elshire y col., 2011). Sus características como el bajo coste de sus componentes, un manejo reducido de la muestra, un menor número de etapas de PCR y purificación, no necesita fraccionamientos por tamaños ni hay límites de secuencias de referencias, un código de barras o "barcode" eficiente, entre otras (Davey y col., 2011), hacen que sea una de las técnicas más elegidas en los análisis de mapeo por asociación mediante la genotipificación a lo largo del genoma (Genome-wide AM), como postulan Khan y Korban (2012). Así lo demuestran algunas de las publicaciones en diferentes especies como arroz, soja o sorgo (Huang y col., 2010; Morris y col., 2012; Sonah y col., 2015; Iquira y col., 2015).

Los estudio **de mapeo por asociación** basados en **genes candidatos** requiere la identificación de los SNP entre las líneas de estudio y dentro de los genes candidatos (Zhu y col., 2008). En este caso, el método más sencillo es la secuenciación de amplicones por PCR, o su re-secuenciación, si previamente han sido detectados los SNPs (Hayward y col., 2012), en varios individuos genéticamente distintos que forman parte de la población seleccionada para llevar a cabo el análisis de asociación. La amplificación por PCR se utiliza para aislar la secuencia de interés utilizando el ADN de la población, individualmente o agrupado, como molde. Los amplicones de los genes candidatos pueden combinarse mediante multiplexado y secuenciarse juntos, siempre y cuando las lecturas de las secuencias de estos amplicones sean suficientes para poder alinearse con el ADN de referencia (Henry y col., 2012). Mediante comparación con la secuencia original se extraen los diferentes polimorfismos, y los análisis y ensayos posteriores se centran en esos marcadores. Aunque esta estrategia es costosa y requiere numerosa mano de obra, los resultados son exitosos cuando se necesita analizar un número moderado de SNPs.

Los avances en las **técnicas de secuenciación** han aumentado las publicaciones en plantas como cítricos, cacahuete o en colza, entre otras, en las que se aplica la **secuenciación de amplicones** (Durstewitz y col., 2010; Kharabian-Masouleh y col., 2011; Curk y col., 2015; Guo y col., 2015), ya que los niveles de la cobertura de secuenciación son significativamente más altos que en la secuenciación por el método clásico Sanger (Henry y col., 2012). Aunque estas técnicas de secuenciación masiva de amplicones de PCR producen lecturas más cortas y con mayores tasas de error que la secuenciación tradicional Sanger (Sanger y col., 1977), se compensa con las múltiples reacciones de secuenciación en paralelo que realizan generando un gran volumen de datos con un costo relativamente bajo y en menor tiempo de múltiples genotipos (M. Perez de Castro y col., 2012; Berkman y col., 2012; Liu y col., 2012). La gran capacidad que muestran estas técnicas de secuenciación hacen que los amplicones que se

obtienen tengan una cobertura significativamente más alta que en la secuenciación Sanger, asegurando que el cambio de base pueda definirse como SNP (Henry y col., 2012).

Las plataformas de nueva generación más recomendadas para la secuenciación masiva de amplicones son Illumina (www.illumina.com), Roche 454 (www.454.com), Ion torrent PGM (Clarke y col., 2014) y Ion Proton (www.thermofisher.com). Se han utilizado, por ejemplo, en estudios de mutaciones inducidas EcoTILLING para conocer la variación natural del germoplasma en diferentes especies como tomate, arroz, trigo, tabaco o lino (Rigola y col., 2009; Tsai y col.,2011; Reddy y col., 2012; Galindo-Gonzalez y col., 2015). De estas plataformas, la plataforma **Ion torrent PGM** presenta el coste más bajo por secuenciación en comparación con otros sistemas, puede generar un tamaño de amplicón de hasta 400pb aunque la media es de 200pb, el volumen de lecturas por sesión de secuenciación puede alcanzar hasta 1Gb en función del chip utilizado, el tiempo de secuenciación máximo es de 4h, y los datos de salida son fácilmente manejables para poder ser interpretados (Liu y col., 2012; Quail y col., 2012 Grada y Weinbrecht, 2013; Galindo Gonzalez y col., 2015). Aunque su tasa de error es mayor (aprox. 1.8%), si la cobertura de la secuenciación es suficiente presenta menos falsos positivos que otras plataformas (Quail y col., 2012).

La técnica de estas plataformas requiere de la **creación de librerías** en las que se aplican secuencias adaptadoras en los extremos 5' y 3' de las secuencias de ADN que se quieren secuenciar. Además, también mediante secuencias cortas (4-10pb) denominadas MIDS (identificadores multiplex) se identifica individualmente el ADN de cada genotipo, permitiendo crear una única librería para secuenciar (Binladen y col., 2007; Meyer y col., 2010). En el caso de Ion Torrent™ PGM, una vez creada la librería, esta se amplifica por clonación en las partículas Ion Sphere ™ (microesferas) mediante una PCR de emulsión, con el objetivo de que cada partícula contenga múltiples copias del fragmento de ADN adosado a la misma (Figura 3a). Estas partículas junto con la muestra amplificada se introducen en el chip de Ion torrent para ser secuenciado por el equipo Ion PGM™ (Figura 3b).

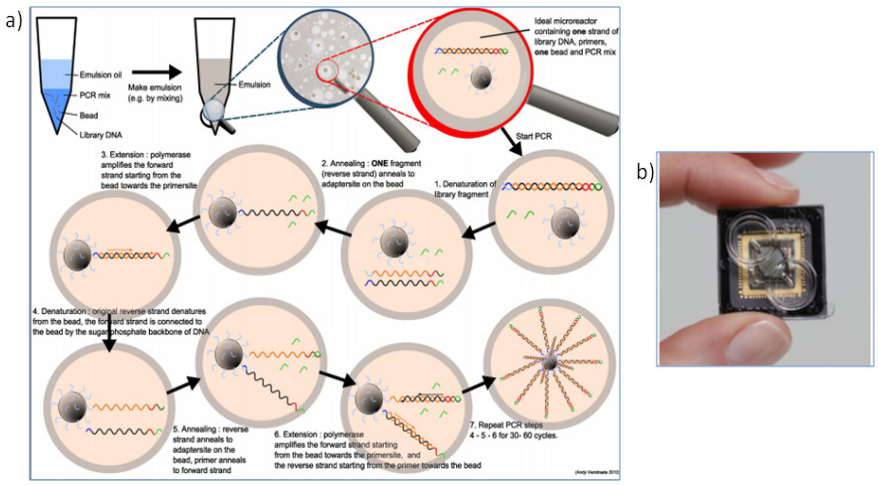


Figura 3: Etapas de la PCR de emulsión y Chip de Ion Torrent (recuperada de Viestraete, 2012).a) Las moléculas individuales de ADN se aislan en microesferas recubiertas con cebadores en burbujas acuosas en una fase oleosa. Posteriormente, mediante una PCR se recubre cada microesfera con copias clonales de la librería y se inmovilizan

para ser secuenciadas; b) Chip Ion Torrent. Estos chips contienen millones de micropocillos con sensores que permiten la detección de los cambios producidos por la liberación de una molécula de H+ cuando los nucleótidos se han incorporado a la cadena de ADN clonal amplificada unidas a la microesferas de cada pocillo (Merriman y col., 2012).Existen tres tipos de chip con capacidades diferentes: Ion 314™ con un millón de pocillos y una capacidad de lectura de hasta 20Mb, Ion 316™ con 6 millones de pocillos y una capacidad de lectura de hasta 200Mb y Ion 318™ con 11 millones de pocillos y una capacidad lectura de hasta 1Gb.

Ion torrent™ PGM utiliza una **tecnología de secuenciación semiconductora**, en la que cuando un nucleótido se incorpora por la polimerasa a la cadena de ADN, se libera un protón, que produce un cambio de pH. El chip recibe un nucleótido tras otro, cuando no es el nucleótido correcto no hay variación de voltaje y es eliminado. En cambio, cuando se incorporan dos nucleótidos seguidos, el voltaje que se detecta es el doble (Flusgberg y col., 2010) (Figura 4).

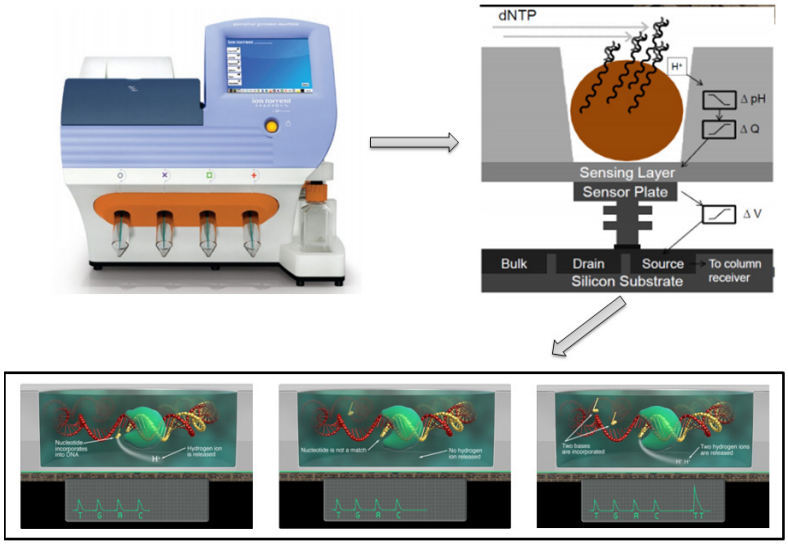


Figura 4: Secuenciador Ion Torrent PGM y mecanismo de secuenciación (recuperada de Vierstraete, 2012).

Junto con el desarrollo de las tecnologías de secuenciación NGS se han desarrollado nuevos algoritmos y **herramientas bioinformáticas** para facilitar el manejo y la interpretación de los datos de salida de estas plataformas, así como para la identificación de SNPs e Indels. La gran cantidad de lecturas que generan estas plataformas requiere de la aplicación de algoritmos que haga posible descubrir estos SNP e indels (Batley y Edwards, 2009), a partir del archivo bruto obtenido de la secuenciación, normalmente FASTQ, dónde se agrupan las lecturas y la calidad de sus bases. Para la identificación de SNP de *novo* es necesario diferenciar entre su descubrimiento a lo largo del genoma o en regiones determinadas como pueden ser en genes concretos, ya que en el primer caso los algoritmos para el mapeo de las secuencias a lo largo del genoma deben ser mucho más potentes.

En el primer caso es necesario algoritmos de mapeo de las secuencias más potentes como BLAT, MAQ, Bowtie, SOAPaligner/SOAP2, BWA y BFast. Estos algoritmos se basan en la alineación de secuencias en dos etapas: 1. Búsqueda heurística de los lugares de alineación de los resultados por indexación con la secuencia original, y 2. Alineamiento real con la misma (Lee y col., 2011). Si no hay un genoma de referencia se utilizan estrategias de ensamblaje de novo que comparan las lecturas entre los

diferentes genotipos. Los archivos generados por los programas de mapeo son los utilizados para la identificación de los SNPs, en los que es necesario aplicar diferentes criterios estadísticos y empíricos como el número de lecturas máximo y mínimo en función del volumen de datos obtenidos de la secuenciación, la calidad y la relación de bases de consenso, entre otros. Pero estos criterios se deben ajustar a la longitud de secuencias obtenidas y a la cobertura de los datos alcanzados por la secuenciación mediante estas plataformas (Kumar y col., 2012). Existen programas que permiten visualizar graficamente los datos generados durante el mapeo de las secuencias para facilitar la interpretación y el manejo como puede ser TABLET, SNP Vista o Savant, así como para la identificación de SNP e Indels como SAMTools, SNVer o SOAPsnp, entre otros, como se muestra en la revisión realizada por Kumar y col., (2012).En el caso de la secuenciación de una región determinada del genoma, o de la secuenciación de genes determinados como es el caso de la secuenciación de amplicones, es suficiente comparar las secuencias obtenidas entre sí y con la secuencia de referencia mediante algoritmos más sencillos como BLAST (Ganal y col., 2009) o ClustalW.

Aunque el principal reto es predecir si estos polimorfismos son debidos a un error de secuenciación, ya que como se ha dicho anteriormente se sacrifica la calidad de las secuencias obtenidas en favor de la generación de un gran número de datos (Lai y col., 2012), lo que puede impedir la identificación de polimorfismos relevantes biológicamente, o asignar falsos positivos. Las confusiones a la hora de determinar si un SNP es verdadero o no dependen de la plataforma de secuenciación utilizada, pero las más generales son la presencia de elementos repetitivos, paralogos y secuencias incompletas (Treangen y Salzberg, 2012). La identificación de SNPs mejora significativamente cuando se utilizan criterios para el filtrado de los datos específicos a las características del genoma y al conjunto de datos. Una opción es comparar el conjunto de datos de un genotipo y eliminar aquellos que presenten grandes diferencias entre sí, ya que esta estrategia identifica la mayor fuente de error como se ha demostrado en el proyecto 1000 genomas que identifica las variaciones en el genoma humano mediante la secuenciación de múltiples poblaciones (Durbin y col., 2010). Otros criterios de filtrado que mejoran la exactitud de esta búsqueda pueden ser el número mínimo de lecturas por genotipo, más del 90% de nucleótidos de un genotipo estén en la misma posición o evitar el enmascaramiento de SNP en homopolímeros con una longitud de la cadena de bases determinada, entre otros (Kumar y col., 2012).

2. OBJETIVOS

En esta tesis doctoral, se utiliza la plataforma Ion torrent PGM para la secuenciación de los genes candidatos seleccionados en el capítulo 2 y el genotipado de la población de mapeo en esos genes candidatos. Los objetivos de este capítulo son:

i. Secuenciación de los genes candidatos y genotipado de la población de mapeo por asociación.

ii. Identificación de SNP e Indels en los genes candidatos.

iii. Determinación de la composición alélica de cada genotipo de la población de palmera de aceite.

3. MATERIAL Y MÉTODOS

3.1. Material Vegetal

El material vegetal utilizado para el mapeo por asociación tiene la misma procedencia que el material utilizado para la búsqueda de posibles genes candidato mediante expresión diferencial, desarrollado en el capítulo 2, apartado 3.1.1 de esta tesis, (Figura 2).

De las 440 familias descendientes de los diferentes cruces DxP, se seleccionaron 242 genotipos que se fenotiparon durante 10 años para los mismos caracteres de interés agronómico sujetos a estudio (Capítulo 3; Apartado 3.1.2; Tabla 2). De estos 242 genotipos, 202 genotipos se seleccionaron considerando el mejor y el peor fenotipo. Las muestras vegetales de hoja se recolectaron durante el mes de mayo de 2012 en la estación experimental de PT Binasawit Makmur, en Surya Adi Estate (latitud: 105°2'0"- 105°4'0" E, longitud: 04°1'0" – 04°2'0"S, elevación: 28m) y perteneciente a la provincia de Sur Sumatra (Indonesia).

Los 40 genotipos restantes pertenecían a las 29 familias utilizadas para el análisis BSA cDNA-AFLP. Para aplicar el mapeo por asociación se seleccionaron los individuos que mostraban el mejor y el peor fenotipo para cada uno de los caracteres de interés, al igual que en los 202 genotipos. En este caso, las muestras vegetales (hoja) se recogieron durante abril de 2011, en la misma localización.

3.2. Métodos

3.2.1. Extracción de adn genómico del material vegetal

Las hojas de estos 40 genotipos se cortaron en cuadrados de 2x2mm y se introdujeron en tubos eppendorf en RNA later (RNAlaterTM, Ambion Europe LTD.) para su conservación y envío al laboratorio de Neiker Tecnalia (Arkaute, Vitoria, España), dónde se extrajó el ADN genómico de cada muestra mediante el kit DNeasy® plant mini de Qiagen. Por otro lado, se extrajo el ADN genómico de las 202 hojas en el laboratorio de Sampoerna Agro TK (Palembang, Sumatra, Indonesia) mediante el mismo kit de extracción de ADN. Este método se basa en una membrana de silica gel que permite una extracción completa de ADN de las muestras de diferentes tejidos de plantas. En primer lugar se rompieron el tejido y los componentes celulares, a continuación el ADN se une a la matriz de gel de la membrana, se purifica el ADN y por último, se eluyó en un volumen final de 30µl de solución tampón AE.

La cuantificación del ADN total de las extracciones se realizó mediante espectrofotometría UV/VIS (Nanodrop 2000, ThermoScientific, USA), y se midió el grado pureza mediante las relaciones de absorbancia A260/280 (>1,8) y A260/230 (1,80-2,20). Además, se visualizó en un gel de agarosa al 0,8% de tampón TAE (40mM Tris-Acetato; 1mM EDTA, pH 8,0) con una tinción de Gel Red 1X (GelRed 10000X in DMSO, Biotium) para corroborar su integridad.

3.2.2. Diseño de cebadores de los genes candidatos

Con el objetivo de amplificar una región específica de los genes candidatos seleccionados en el capítulo 2 se realizó un análisis comparativo mediante Blastn (E- valor < 1E-35) en la página web de NCBI contra la base de datos de EST's de *Elaeis guineensis* y *Phoenix dactilifera,* y se seleccionaron las regiones conservadas del genoma para el diseño de los cebadores.

Los cebadores de los genes candidatos co-localizados seleccionados cuya procedencia eran secuencias con SNP's ya detectados previamente (Proyecto OPGP, no publicado) se diseñaron en la misma secuencia integrando en el amplicón al menos uno de los SNPs (Figura 5).

El diseño de de cebadores se realizó con el software Primer3 (Rozen y Skaletsky 2000), utilizando por defecto los parámetros del mismo y con un longitud entre 150pb y 400pb. Los cebadores se sintetizaron unidos a unos adaptadores comunes (uniA y uniB) necesarios para multiplexar las PCRs. La síntesis de los cebadores fue realizada por una empresa externa (MWG Eurofins Genomics, Alemania). Todos los cebadores diseñados se presentan en el anexo 6.

Figura 5: Ejemplo de diseño de los cebadores de la secuencia del gen candidato KG196 co-localizado con un QTL de BN mediante el software libre Primer3. a) La secuencia KG196 posee dos SNPs en dos posiciones diferentes, la región dónde se encuentra cada una de ellos se muestra en la imagen de la siguiente forma [NNNN]; b) Muestra los cebadores diseñados para la secuencia y sus características, así como el tamaño del amplicón obtenido; c) Muestra el amplicón obtenido, la región dónde se sitúan los cebadores (Fw:>>>>>; Rv: <<<<) y las región donde están situados los polimorfismos de un solo nucleótido (****).

Cada pareja de cebadores se probó individualmente en 5 genotipos al azar del ADN genómico extraído corroborando así su funcionalidad.

3.2.3. Creación de librerías para secuenciación Ion Torrent ™

La técnica utilizada para la creación de las librerías fue "Amplicon Fusion" de Ion Torrent (Life Technologies™) con modificaciones. Se realizaron 6 librerías denominadas OPn, en las que se incluyeron entre 20 y 60 amplicones de los genes candidatos seleccionados para cada una de ellas. A continuación se desarrolla el proceso para la creación de la librería OP4.

1º- Multiplex PCR

Se realizaron reacciones de PCR multiplexadas entre 10 y 15 cebadores por reacción (Figura 9a). Todos los cebadores de los genes candidato incluían una secuencia de oligonucleótido común denominadas UniA, posicionado en 5'→3', y UniB, posicionado en 5'←3' (Figura 6) y otra parte específica de cada gen candidato seleccionado.

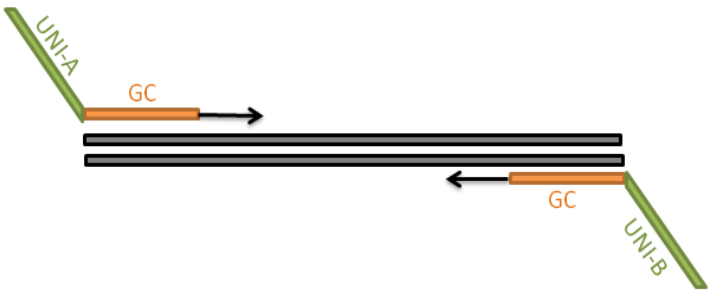


Figura 6: Modificación de los cebadores realizada para la secuenciación. La imagen muestra la orientación de las secuencias moldes específicas utilizando los cebadores de fusión formados por los cebadores específicos del gen candidato (GC) y los cebadores universales UniA y UniB.

La mezcla de la reacción y el programa de amplificación se muestra en la tabla 1.

Tabla 1: Mezcla de reacción para PCR Multiplex y programa PCR.

REACCIÓN	1X	Programa PCR[4]
H20 estéril dd[1]	Variable	
10X tampón KCl PCR[2]	2,5 µl	
dNTP's (2,5mM)[3]	2 µl	
Cebador derecho (10µM)	0.2µl/cebador	**94ºC 5 min, 35ciclos [94ºC 30sg, 58ºC 30seg, 72ºC**
Cebador izquierdo (10µM)	0.2µl/cebador	**45sg], 72ºC 7min**
TAQ ADN Polimerasa 5U/µL[2]	0.1 µl	
ADN (10ng/µl)	2 µl	
Volumen final	25µl	

[1]UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™, USA);2DFS-Taq DNA Polymerase (Bioron ™, Alemania): 3Bioron ™, Alemania;4 Termocicladores Primus 96/384 (MWG AG Biotech)

Estas reacciones multiplexadas se verificaron visualmente en geles de agarosa al 1,5% en una disolución tampón 1X TAE (40mM Tris-Acetato; 1mM EDTA, pH 8,0) con una tinción de Gel Red 1X (GelRed 10000X in DMSO, Biotium), detectándose los fragmentos en un rango entre 150pb y 400pb, en función del tamaño de los amplicones de los genes candidatos (Figura9b). Además, el tamaño y la concentración de las multiplexes se comprobaron en el bioanalizador Agylent 2100 Bioanalizer (Agilent Technologies, USA) con su kit de alta sensibilidad (Agilent High Sensitivity DNA Kit) (Figura 9c y 9d). Finalmente, las multiplexes se diluyeron a una concentración de 5ng/µl, y se mezclaron por genotipo individualmente.

2º- Unión de los adaptadores de secuencias específicas

La identificación individual de cada uno de los 242 genotipos se realizó mediante la asignación de un "barcode" o código de barras denominado MID ("multiplex identifier"). Este código de barras consiste en una secuencia nucleotídica de entre 6 y 10 bases. En nuestras librerías, se utilizaron dos códigos de barras, uno al inicio de la secuencias y otro al final de la secuencia por cada genotipo, permitiendo analizar mediante la combinación de 16 barcodes diferentes hasta 256 genotipos (16x16), por lo que en esta tesis se pudieron identificar los 242 genotipos, tal y como se muestra en la tabla 2. Todos los genotipos y su correspondiente "barcode" se muestran en el anexo 6 (Tabla 6.2)..

Tabla 2: Barcodes (MIDs) utilizados para identificar cada genotipo individualmente.

Codificación	Recodificación	Fw	Rv
11	AA	ACGCTCAG	CTGAGCGT
12	AC	TACATCAT	ATGATGTA
13	AG	CGCGACTA	TAGTCGCG
14	AT	AGCTAGTC	GACTAGCT
21	CA	CGAGATCA	TGATCTCG
22	CC	TCAGTGCTG	CAGCACTGA
23	CG	TATGCTAGA	TCTAGCATA
24	CT	CGCACTGAG	CTCAGTGCG
31	GA	CAGACTCTA	TAGAGTCTG
32	GC	TGTGAGCAC	GTGCTCACA
33	GG	GCGATAGTAC	GTACTATCGC
34	GT	GCTATGACAG	CTGTCATAGC
41	TA	ATACATAGCT	AGCTATGTAT
42	TC	GACAGCGCGT	ACGCGCTGTC
43	TG	CATGTCAGTA	TACTGACATG
44	TT	ACGCACTCGC	GCGAGTGCGT

Para ello, a partir de las multiplexes generadas mezcladas por genotipo, cuantificadas y diluídas a 5 ng/µl, se realizó una segunda reacción de PCR donde en 5' del cebador directo o forward se añadía al adaptador Uni A un barcode y la secuencia necesaria para la secuenciación con Ion Torrent. A este primer le denominamos AK, disponiendo de 16 AKs diferentes, uno para cada barcode (Figura 7).

Igualmente, el cebador Reverso o Reverse consistía en el adaptador utilizado anteriormente en las multiplexes Uni B, unido en 5' a un barcode y al adaptador necesario para la secuenciación con Ion Torrent. De la misma forma tendríamos así 16 cebadores BKs diferentes para cada barcode (Figura 7).

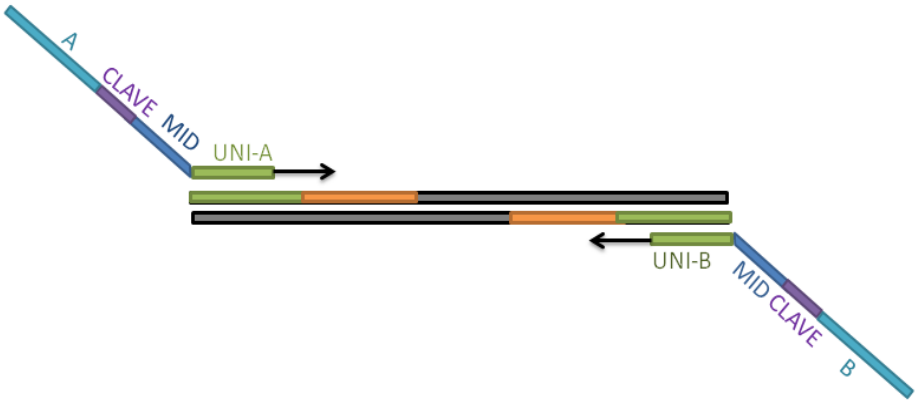


Figura 7 Modificación de los cebadores realizada para la secuenciación. La imagen muestra la orientación de las secuencias molde utilizando un cebador fusión con colas universales (UniA y UniB) y al que se le han añadido las secuencias de identificación del genotipo individual denominadas MIDS. El cebador A y B son los cebadores utilizados en la PCR de emulsión previa a la secuenciación por la tecnología Ion Torrent.

La mezcla de reacción y el programa de amplificación para la unión de los adaptadores en cada genotipo se muestra en la siguiente tabla:

Tabla3: Reacción y programa PCR para la unión de los adaptadores en cada genotipo.

REACCIÓN	**1X**	**Programa PCR[4]**
H20 estéril dd[1]	16,2µl	**94ºC 5 min, 30ciclos [94ºC 30sg, 58ºC 30seg, 72ºC 45sg], 72ºC 7min**
10X tampón KCl PCR[2]	2,5 µl	
dNTP's (2,5mM)[3]	2 µl	
Cebador derecho_AKn(10µM)	0.2µl	
Cebador izquierdo_BKn(1µM)	2µl	
TAQ ADN Polimerasa 5U/µL[2]	0.1 µl	
Producto PCR (5ng/µl)	2 µl	
Volumen final	25µl	

[1]UltraPure™ DNase/RNase-Free Distilled Water (Invitrogen™, USA);[2]DFS-Taq DNA Polymerase (Bioron ™, Alemania): [3]Bioron ™, Alemania;[4] Termocicladores Primus 96/384 (MWG AG Biotech)

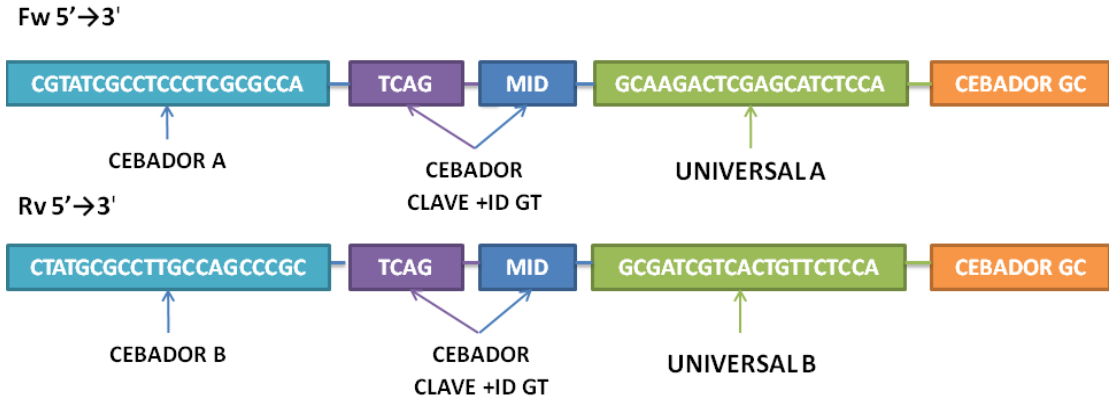


Figura 8: Secuencias de los adaptadores utilizados en la técnica amplicón fusión en ambos sentidos. Cebador A y B= Cebadores utilizados para la PCR de emulsión realizada antes de la secuenciación; CLAVE+ ID GT= Secuencia de adaptores utilizados para la identificación de cada genotipo. CLAVE es la secuencia necesaria para la unión posterior de los cebadores utilizados en la PCR de emulsión. ID GT= se refiere a la secuencia "barcode" que identifica al genotipo y se denomina MID; UNIVERSAL A y B= Secuencias de los cebadores universales utilizados para la unión de los MIDs y unidos también a los cebadores específicos de cada gen candidato; CEBADOR GC= cebador específico de cada gen candidato.

A continuación, las reacciones de amplificación se verificaron en geles de agarosa al 1,5 % en 1 x TAE (Figura 9e) (40mM Tris-Acetato; 1mM EDTA, pH 8,0) con una tinción de Gel Red 1X (GelRed 10000X in DMSO, Biotium), detectándose los fragmentos en un rango entre 190pb y 440pb. El producto final de cada genotipo se mezcló en un tubo eppendorf (Figura 9f) y se purificó mediante columnas GeneRead Size Selection Kit® (QIAGEN GmbH – Germany). La calidad de la librería final se verificó en el bioanalizador Agilent 2100 Bioanalyzer (Agilent Technologie, USA) con el kit de alta sensibilidad (Agilent High Sensitivity DNA Kit) (Figura 9g, 9h, 9i, 9j), y se envió a secuenciar mediante Ion Torrent en un chip de 318G.

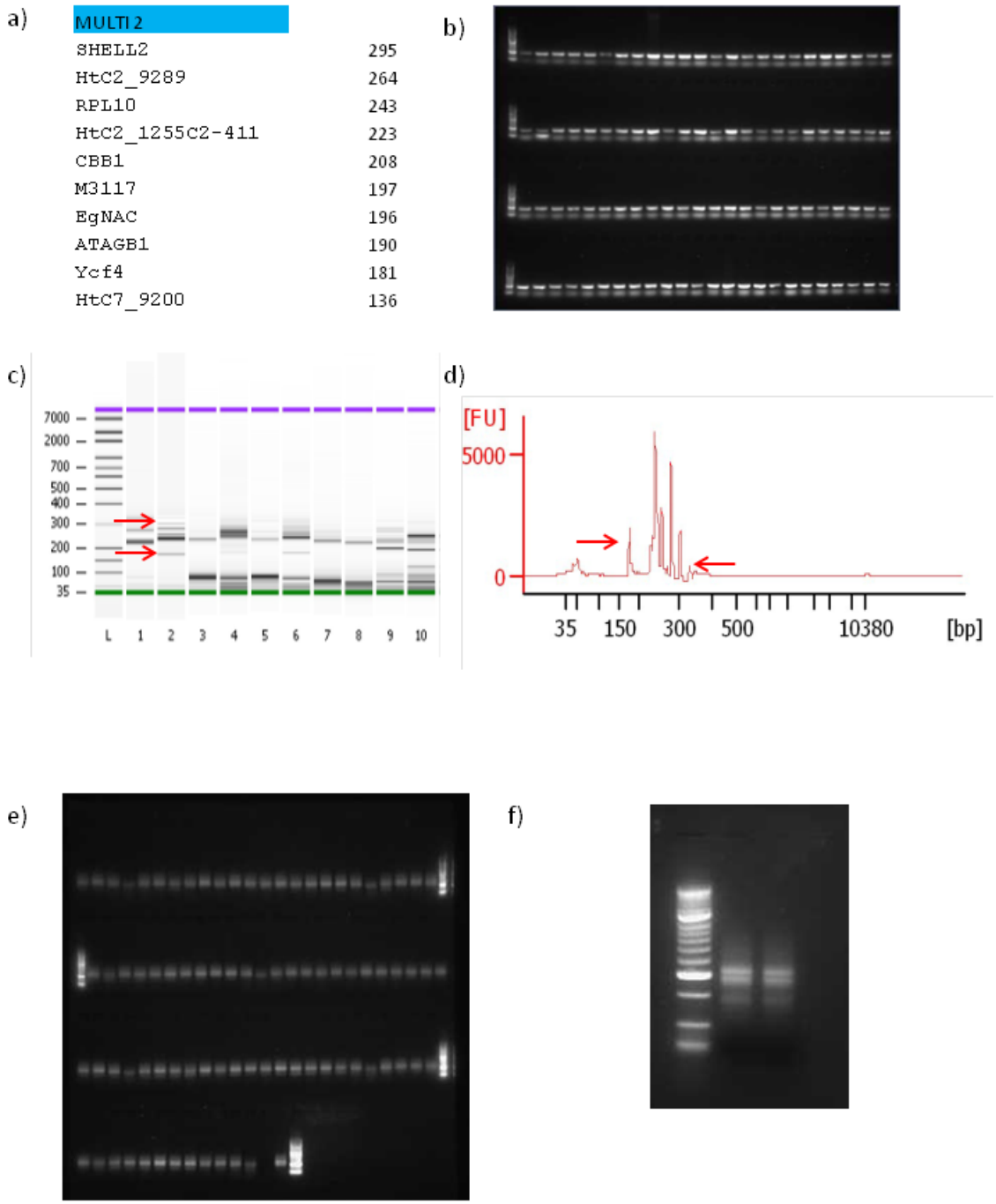


Figura 9: Etapas de la creación de librerías por el método de "amplicon fusion" de Ion Torrent modifado. El ejemplo corresponde a la creación de librería OP4. a) Organización de un grupo de cebadores para la reacción de la PCR multiplexada corresponde a la multiplex 2.b) Imagen en el gel de agarosa de la PCR multiplex correspondiente a la placa 2 multiplex 2; c) Imagen del bionalizador Agilent 2100 de la misma PCR multiplex, y señalando la altura mínima y máxima de los amplicones creados (muestra2); c) Imagen del electroforegrama de la misma multiplex dónde se observan las alturas de los amplicones para la misma multiplex PCR; e) Imagen en gel de agarosa que muestra la unión de los adaptadores y MIDs a los amplicones de cada genotipo; f) Imagen en el gel de agarosa con dos réplicas que muestra la librería creada sin purificar y el marcador.

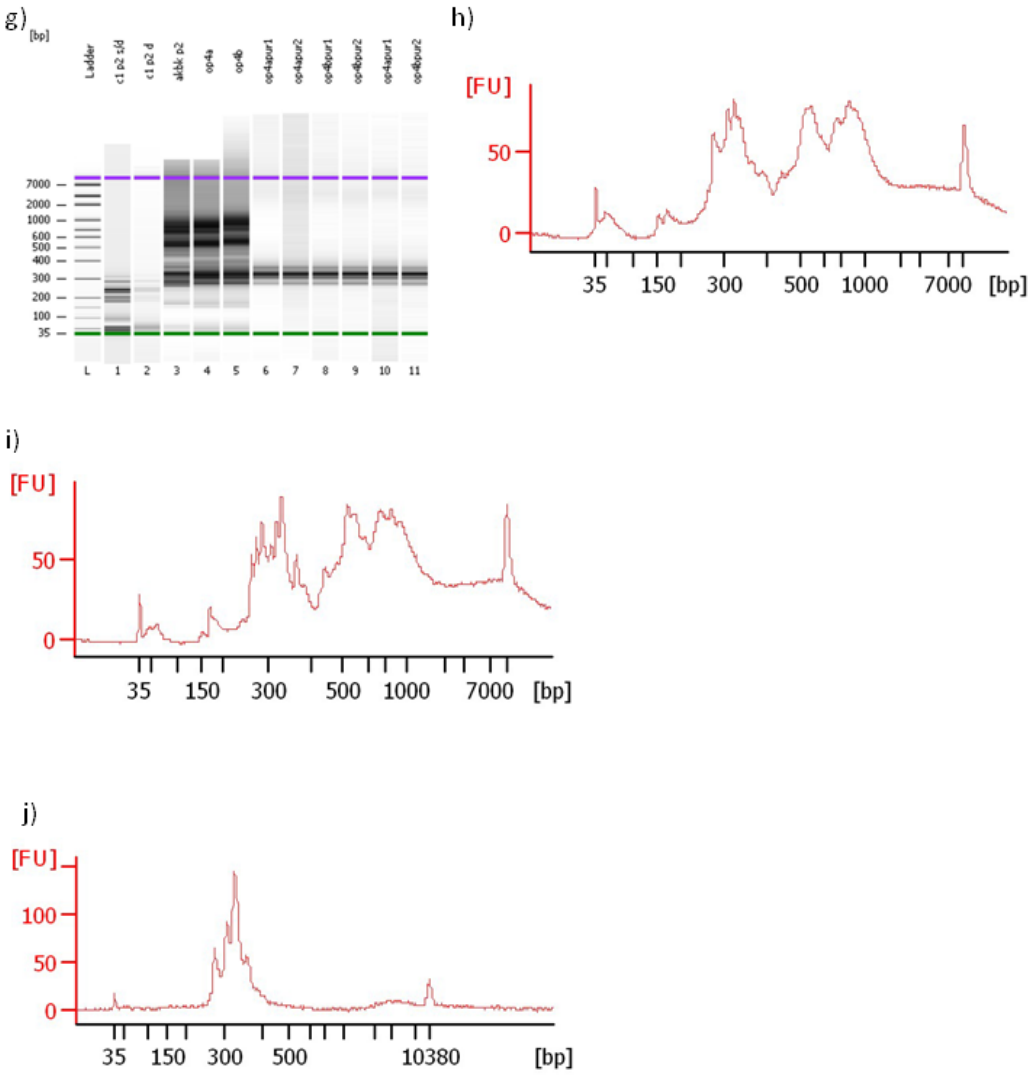


Figura 9 (continuación): Etapas de la creación de librerías por el método de "amplicon fusion" de Ion Torrent modifado. g) Imagen del bioanalizador donde se muestran la unión de los adaptadores y MIDs a la librería (carril 3), la librería OP4 sin purificar (muestras 4 y 5), purificada (6, 7, 8 y 9), purificada y diluida 1:50 (10 y 11); h) Electroforegrama mostrado por el Bioanalizador Agilent 2100 de la unión de adapatadores; i) Electroforegrama mostrado por el Bioanalizador Agilent 2100 de la la librería sin purificar; j) Electroforegrama mostrado por el Bioanalizador Agilent 2100 de la librería purificada y lista para enviar a secuenciar.

3.2.4. Procesado de los resultados de la secuenciación

Los datos moleculares procedentes de la secuenciación se analizaron con un software creado con esta finalidad denominado ASPAM (E. Rittter, comunicación personal, 2012). Este software permitió filtrar los datos de la secuenciación y detectar si existía diversidad alélica en los amplicones de los genes candidatos secuenciados en cada genotipo basándose en el número de secuencias repetitivas. El proceso realizado para llevar a cabo el análisis de cada librería se muestra a continuación.

1º- Preparación de los datos

Antes del inicio del análisis se crearon los archivos necesarios para el filtrado de los datos. Estos archivos contenían: 1. La lista de MIDs con las diferentes combinaciones de secuencias identificativas del genotipo codificadas, tal y como se muestra en el anexo 6 (Tabla 6.2) (Archivo MIDS.prn); 2. Las secuencias de los amplicones correspondientes a cada gen candidato en formato fasta (Figura 10) denominado CGlist (Archivo CGList.txt); 3. Los datos fenotípicos recogidos durante 10 años para los genotipos estudiados (Anexo 10; Tabla 10.1).. Aunque en este capítulo no se utilizaron, fue necesario incluirlos porque ASPAM permite realizar el mapeo por asociación, tal y como se desarrollará en el siguiente capítulo de esta tesis doctoral; 4. El archivo comprimido recibido con los datos de la secuenciación en formato fastaQ, que contiene las lecturas de la secuenciación y los archivos de calidad asociados a esas lecturas y convertido a formato Windows mediante el programa Unix2win (Telcosoft), denominado Crag1.fastaQ.

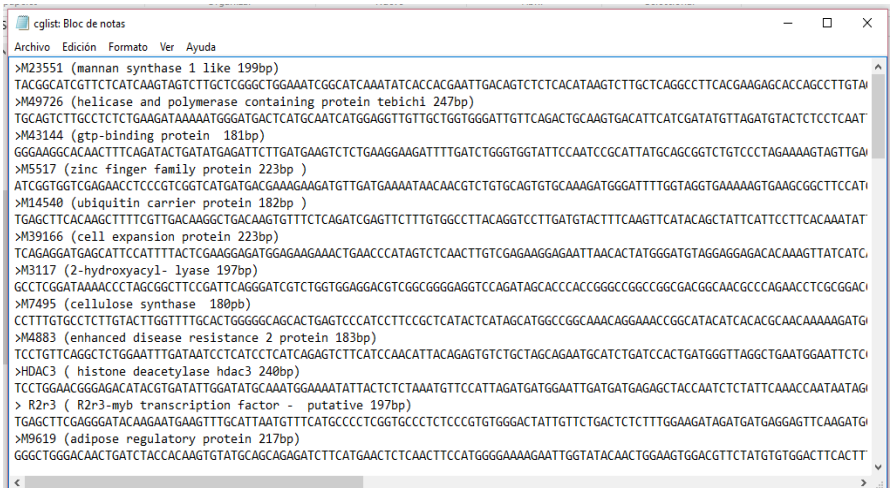


Figura 10: Ejemplo del archivo CGlist.txt. La imagen muestra parte del archivo creado con los amplicones de los genes candidatos de la librería OP4. El archivo en formato fasta muestra en la primera línea> Nombre Cebador CG (función posible, longitud del amplicón en pb) y en la segunda línea la secuencia del amplicón.

A continuación se creó un nuevo proyecto que se llamó como el nombre de la librería secuenciada y se incluyeron estos archivos.

2º- Procesamiento archivo Crag1.FastaQ

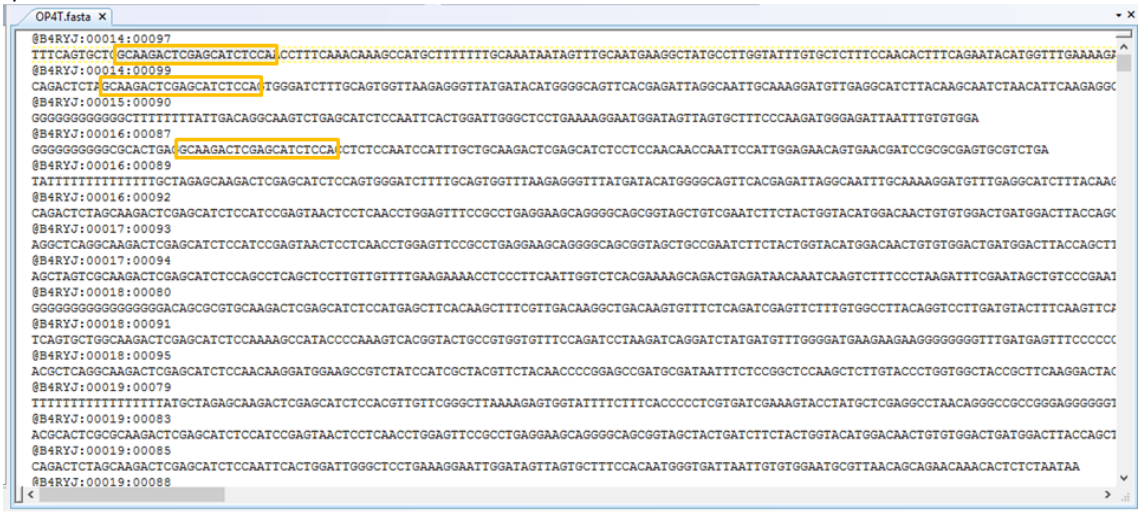
El archivo en formato Windows se procesó para separarlos en dos archivos diferentes: 1. El archivo Crag1T.fasta dónde se mostraron todas las lecturas obtenidas de la secuenciación con una longitud superior a 120pb, ya que además de las secuencias específicas de cada gen candidato las secuencias contenían las secuencias de los cebadores universales A y B, de los MIDS y de los cebadores utilizados en la PCR de emulsión. Las bases que suman estas secuencias extras anexas al amplicón del gen candidato sumaban ±120pb (Figura 11a), y 2. El archivo Crag1T.qual dónde se compilaron los datos de calidad de cada una de las lecturas obtenidas, y cada base tiene un valor que se muestra en ASCII. Este archivo no se utilizó para el análisis de nuestros datos, para evitar perdidas innecesarias debidas a

errores de secuenciación, ya que ASPAM está basado en el nº de repeteciones de secuencias idénticas, considerando que aquellas secuencias que sean más repetitivas sean las verdaderas y no en la calidad de las secuencias obtenidas.

3º- Filtrado de los datos para lecturas completas

El paso siguiente fue el filtrado de las secuencias que en cada extremo contenían los MIDs identificativos del genotipo, y se separaron creando un archivo en formato fasta (Crag1Z.fasta) que contenía cada secuencia específica que se iniciaba y terminaba por los cebadores específicos, y el nombre de cada secuencia al que se añadió los cuatro caracteres procedentes de los MIDS y que eran representantes del genotipo (Imagen 11b).

a)



b)

Figura 11: Archivos generados durante el filtrado de datos de la librería OP4. a) Las secuencias con >120pb se recogen en este archivo, y comienzan con las secuencias correspondientes al MID y al cebador universal UNI_ A. En la imagen se resalta donde está situado el cebador UNI_A en la secuencia (Crag1T.fasta). b) Los MIDs situados al inicio y final de las secuencias se transforman en 2pb de se sitúan en el nombre de la secuencia, se eliminan los cebadores universales, y la secuencia restante es el amplicón del GC (Crag1Z.fasta). En el ejemplo, el rectángulo naranja muestra el inicio del amplicón del GC cuyo cebador es SDP1. El nombre de la secuencia muestra GACC que corresponde al genotipo 3122, los rectángulos rojos el inicio del amplicon del GC cuyo cebador es HARBC. El

primero de ellos corresponde al genotipo TTAC (4412), y el segundo al genotipo GACC (3122); los rectángulos verdes al inicio de los amplicones del GC cuyo cebador es EgACP. En el primero el genotipo es AATC (1142), y el segundo genotipo es ATAG (1413).

4º- Identificación de los genes candidatos

Las secuencias de los genes candidatos se identificaron mediante la detección en todas las secuencias de los 10pb al inicio y al final de las mismas. Estos pb correspondían a los cebadores específicos de cada gen candidato. Para ello, el programa utilizó el archivo procedente del paso anterior (Crag1Z.fasta) y generó un nuevo archivo en formato fasta (Crag1Y.fasta) que contenía las secuencias de los amplicones anteriores, y en el nombre de cada secuencia se incluyó el número de gen candidato (Figura 12a).

El siguiente paso fue generar los archivos individuales para las secuencias obtenidas de cada gen candidato en los diferentes genotipos, es decir un archivo por gen candidato (CGi.fasta). A partir de este archivo, se filtraron estos datos para generar nuevos archivos en los que los genes candidatos se separaron por genotipo, gracias a la abreviatura del MID en el nombre de cada secuencia (Figura 12b). El filtro aplicado exigía que el genotipo tuviera un número mínimo de 3 lecturas del gen candidato para ser incluido en el archivo generado.

Junto con este archivo, el proceso creó otros archivos diferentes para cada gen candidato que incluían la información de la secuencia, datos estadísticos, todas las secuencias de cada genotipo por gen candidato, todas las secuencias no repetitivas de cada genotipo por gen candidato y archivos con las alineaciones múltiples realizadas con Clustal W (Thomposon y col., 1994), para el filtrado de las diferentes secuencias.

5º- Detección de los patrones

La diversidad alélica de cada gen candidato se basó en el número más alto de repeticiones de secuencias idénticas en cada genotipo, buscando sus patrones en cada gen candidato y asignando de esta forma la composición alélica de cada genotipo.

Para llevar a cabo esta detección se filtraron en un primer momento los alelos de los genotipos bajo las siguientes condiciones: diploidía 2x, número mínimo de secuencias totales consideradas (5) y número mínimo de secuencias repetitivas que se consideran para un alelo (2). Este filtrado significa que únicamente se extrajeron de todos las lecturas de los 4 patrones con mayor frecuencia para cada genotipo que tenían un máximo de 5 lecturas con al menos 2 repeticiones de cada secuencia patrón. En este paso se crearon nuevos archivos, uno de ellos con resultados estadísticos, otro con las secuencias más repetitivas y otro dónde se mostraban los alineamientos múltiples realizados (Tabla 4; Figura 13).

a)

b)

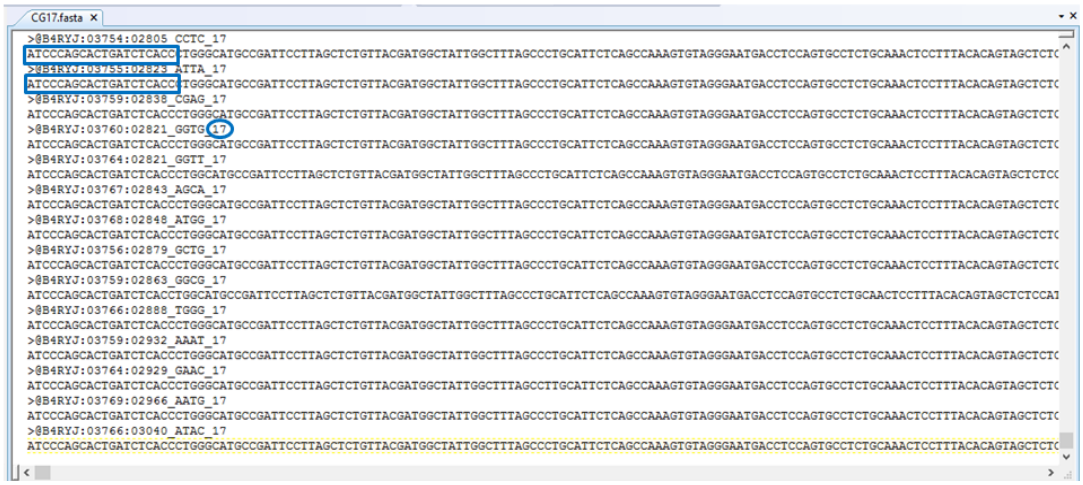


Figura 12: Archivos generados durante la identificación de los genes candidatos de la librería OP4. a) El archivo muestra las secuencias seleccionadas como genes candidato mediante la identificación de las 10pb iniciales del amplicón correspondiente (Crag1Y.fasta). En la imagen se resalta el inicio del amplicón de 3 genes candidatos cuyos cebadores son M3256 (azul) y señalado con GC17 por el programa, EgACP (amarillo) y señalado como GC 47 y HaRBC (verde) y señalado como GC52. b) Ejemplo del archivo generado por ASPAM para el GC cuyo cebador es M3256 y nombrado como GC17 por el programa (CG17.fasta).

Tabla 4: Archivo modificado mostrado por ASPAM para el gen candidato GC17 de la librería OP4. Genotipo= codificación del genotipo; Nº Lecturas= muestra el número de lecturas totales del GC en el GT; Nº Secuencias NR= muestra el número total de secuencias no repetitivas. Las siguientes columnass muestran el número de secuencias repetitivas que caracterizan cada patrón extraído. Por ejemplo, el genotipo AAAA muestra la existencia de un patrón claro de un posible alelo, con 2348 lecturas idénticas. La siguiente columna puede indicar la presencia de otro alelo (38), pero es necesario que sea revisado manualmente porque hay mucha similitud en el número de secuencias repetidas con las siguientes columnas.

GENOTIPO	**Nº LECTURAS**	**Nº SECUENCIAS NR**										
AAAA	3446	566	2348	38	28	25	22	15	12	12	11	9
AAAC	543	104	399	6	6	4	4	3	3	3	3	3
AAAG	1000	206	693	16	10	8	5	4	4	4	4	3
AAAT	332	74	243	6	3	3	2	2	2	2	2	2
AACG	1363	250	982	23	15	7	7	7	6	5	5	4
AGGA	5841	721	4096	190	61	35	29	22	17	15	15	15
AAGA	1515	241	1071	49	25	12	8	7	6	5	5	5
AAGC	2352	353	1668	56	27	17	12	10	9	8	7	7
AAGG	1626	287	1158	17	16	9	9	8	7	6	6	5
AAGT	2098	326	1500	33	31	18	14	13	10	10	8	7
CGTG	4255	588	3039	100	19	17	17	16	13	13	12	12
CTTA	3242	414	2370	118	21	15	11	10	9	9	8	8
CTTG	4845	612	3506	145	34	21	18	18	17	12	12	10

Figura 13: Ejemplo del archivo fasta para el GC17 en el GT AAAA (1111). El título de cada secuencia indica el número del patrón otorgado por el programa, y sus frecuencias o número de repeticiones. Como ejemplo, S271-2348 indica que el patrón S271 se repite 2348 veces en ese genotipo.

A continuación, se llevo a cabo la definición de los patrones determinando el número posible de alelos que existen para cada gen candidato en la población o conjunto de genotipos. En este caso se seleccionaron los cuatro patrones (2+2) con mayor frecuencia en cada genotipo que ocurrió en al menos un genotipo, asumiendo la redundancia de las secuencias. Despues se eliminan todos los patrones duplicados sumando el número de presencias al patrón representativo (Figura 14).

Figura 14: Ejemplo del archivo fasta generado para el GC17 después de determinar el conjunto de alelos para la población. El título de cada secuencia iniciado por > indica el patrón y el número de genotipos dónde se encuentran. Por ejemplo, >P1-9, el patrón 1 se encuentra en 9 genotipos.

Esta detección de patrones permitió en su primer paso reducir el número de alelos presentes en la población, y en su segundo paso obtener los genes candidatos con un menor número de lecturas. Se obtuvo así un conjunto de alelos preliminares para la población de estudio.

6º-Asociación de alelos y composición alélica a la población (conjunto de genotipos)

El conjunto de alelos preliminares obtenido se depuró mediante el alineamiento de las secuencias y la determinación manual del conjunto final de los alelos para cada gen candidato en la población. Las secuencias de cada gen candidato se alinearon en formato ClustalW y se determinaron visualmente los SNPs e Indels (en al menos tres pares de bases o en múltiplos de tres) como variaciones alélicas, que posteriormente se posicionaran en el conjunto final de alelos (Figura 15; Figura 16).

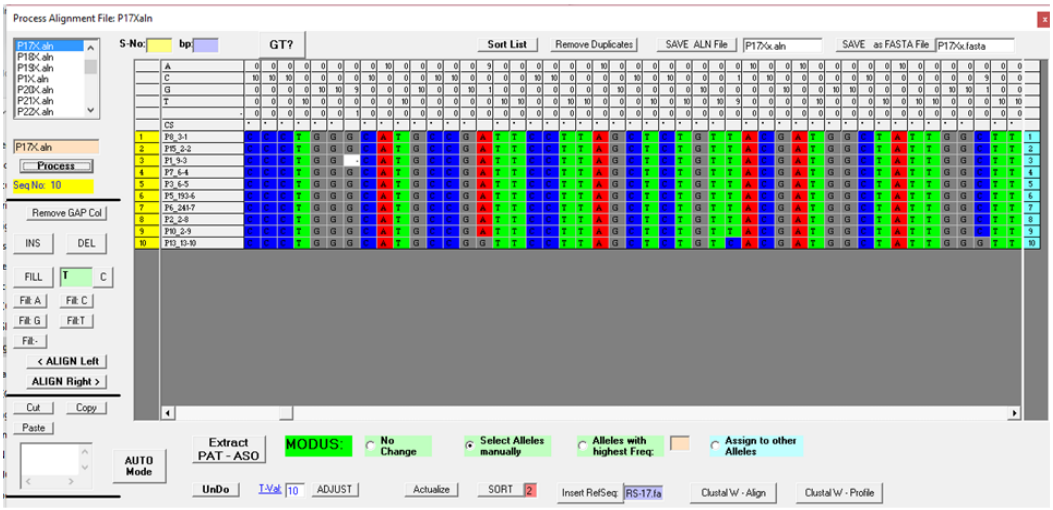


Figura 15: Alineamiento de las secuencias y determinación manual del conjunto final de alelos en la población. La imagen muestra el alineamiento de las secuencias para el GC17 (librería OP4), antes de la determinación manual de los alelos. Por ejemplo el patrón 3 presenta un gap en el homopolímero, debido a un error en la secuenciación, y se eliminará. En cambio, el patrón 10 parece ser un SNP verdadero (T->C).En cada fila de la segunda columna se muestran los patrones generados en el paso anterior con el siguiente formato, Px_y-n. Px= número del patrón inicial; y=número de genotipos dónde está;n= número original del patrón.

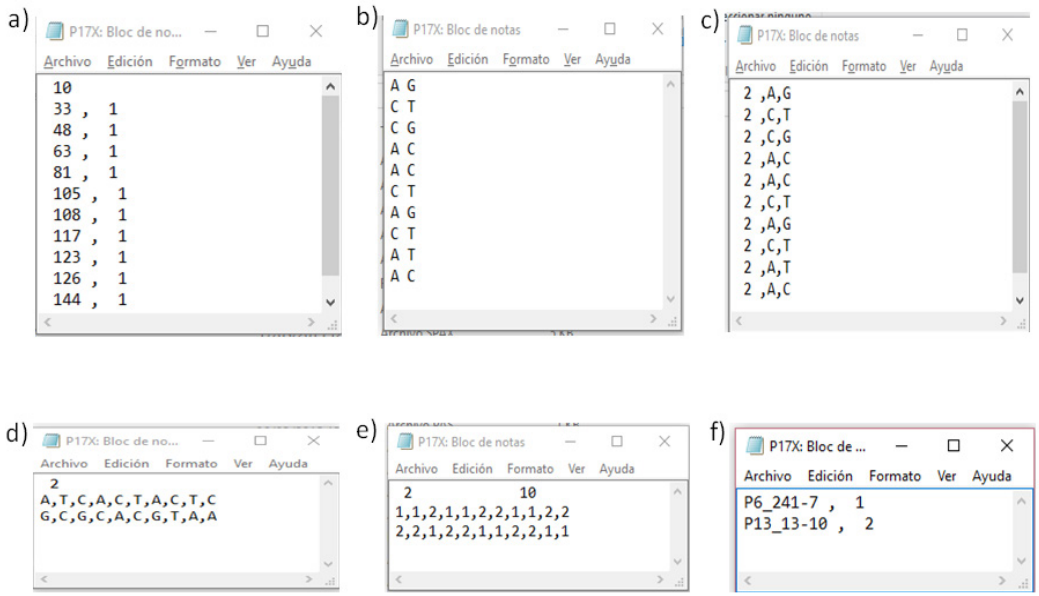


Figura 16: Archivos con información de utilidad generados por ASPAM después de la determinación manual del conjunto final de alelos de la población (Ejemplo GC17, librería OP4). a) Archivo SNP: la primera fila muestra el número inicial total de SNPs e Indels identificados, en este caso 10. Las siguientes muestran filas muestran la posición (pb) y la longitud, es decir el número de bases que están afectadas; b) Archivo SNS: muestra para cada SNP el polimorfismo detectado, en la figura se muestran 10 SNP con sus cambio de base (Fila 1 base A o base G, Fila2 base C o base T, y asi sucesivamente); c) Archivo SNX: muestra para cada SNP el número de polimorfismos detectados y su combinación, indicando el número de niveles. En la figura, se muestra en la primera columna el número de niveles de cada cambio de base, en este caso 2. En la segunda y tercera columna, el cambio de base que ocurre, como el archivo de la figura b.; d) Archivo PAN: muestra el número de patrones detectados definido por sus combinaciones de SNPs. En el ejemplo, 2 indica el número total de patrones, y las 2 filas siguientes la secuencia para cada patrón; e) Archivo PAX: muestra la secuencia correspondiente del patrón y el número de polimorfismos o niveles detectados. En el ejemplo, 2 indica el número de patrones, y 10 número de niveles o polimorfismos detectados. La primera columna 1 es el patrón 1, y 2 el patrón 2. El resto de columnas identifican al polimorfismo o nivel detectado, de la siguiente manera: columna 2 1=A y 2 =G, columna 2 1=T y 2=C y así sucesivamente; f) Archivo PAS: muestra el nombre de los alelos existentes, Alelo 1=P6_241-7 y Alelo 2= P13_13-10.

7º- Composición alélica de cada genotipo

Cada uno de estos alelos seleccionados se asociaron al genotipo corresponidentemediante un paso que determinó en que genotipos se encuentran los alelos generados previamente (Figura 17). Esta asociación se realizó por dos vías diferentes

1- Coincidencia total de la secuencia del genotipo con la secuencia patrón (Perfect pattern match). En esta opción la secuencias de los alelos seleccionados se alinearon con todas las secuencias posibles de todos los genotipos en cada gen candidato, buscando los genotipos que tenían exactamente el mismo patrón. Esta opción es la más adecuada cuando se dispone de un número suficientes de lecturas para cada genotipo.

2- Coincidencia de la región dónde se sitúan los SNP del genotipo con la región dónde se posicionan los SNP en la secuencia patrón (SNP match). Para esta opción únicamente se asociaron las secuencias de los genotipos que coincidían con la región de 2pb anterior y posterior a la variación (SNP e Indel) en el patrón de referencia, siendo por tanto esta opción la menos restrictiva. Sin embargo, es la

más aconsejada porque engloba a la opción 1- y determina adicionalmente los alelos en genotipos con nº de lecturas reducidas.

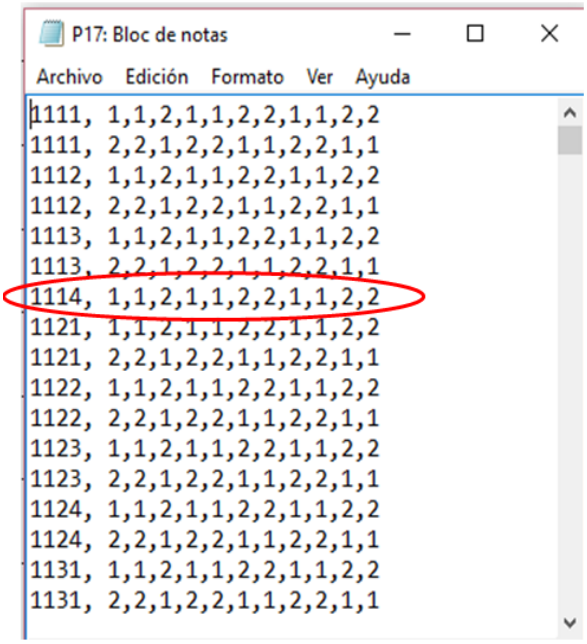


Figura 17 Parte del archivo SPA generado por el programa ASPAM (GC17, librería OP4). Este archivo muestra los patrones generados en los genotipos de estudio. La primera columna es el genotipo, la segunda columna indica el patrón, y las columnas restantes los niveles de SNPs. Como se observa en la figura los genotipos 1111,1112,1113 son heterocigóticos, con 2 alelos, mientras que el genotipo 1114 es homocigótico para el alelo 1.Este archivo se revisó para comprobar que ningún genotipo presentaba más 2 de alelos, ya que *E.guineensis* Jacq. es una especie diploide.

Los resultados de esta asociación se revisaron para identificar de genotipos que presentaban un exceso de alelos de acuerdo al nivel de ploidia de la palmera de aceite (máximo 2 patrones por genotipo), y procesar estos genotipos de nuevo.

Algunos genes candidatos presentaron más de dos alelos considerándose como potenciales genes candidatos multi-locus. Estos se procesaron por separado después de la revisión, ya que no es posible conocer a qué loci pertenece cada alelo. Este procesado consideró la presencia o ausencia de cada alelo como un único gen candidato, generando tantos supuesto genes candidatos como alelos hay.

7º- Análisis de la combinación de alelos

En este paso, previo al mapeo por asociación se determinó la combinación alélica particular en cada genotipo y para cada gen candidato que fue polimórfico (Figura 18a), y su estado homocigótico o heterocigótico en cada genotipo(Figura 189b).

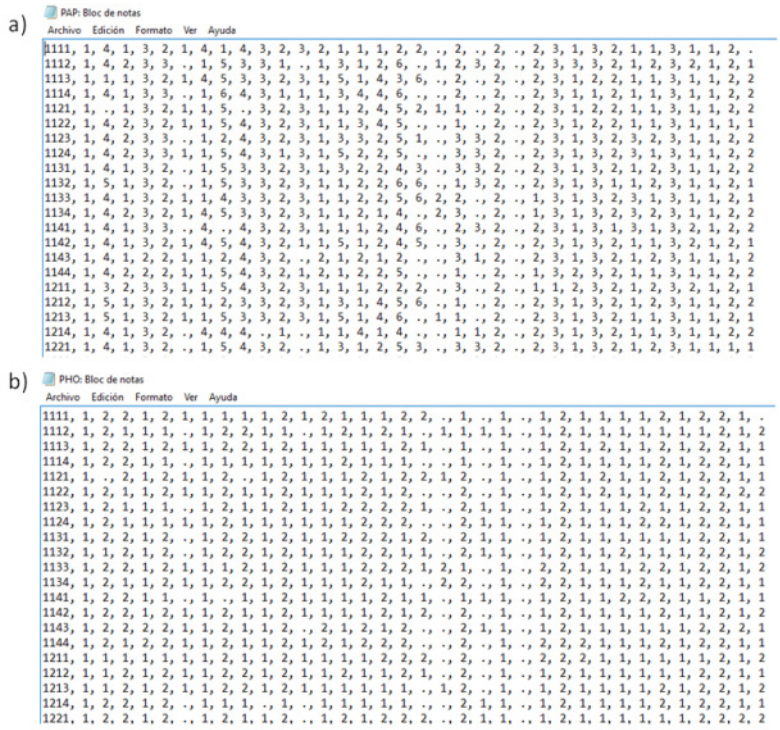


Figura 18: Parte de los archivos generados por ASPAM después del análisis de las combinaciones alélicas. A) Archivo PAP que muestra la composición alélica para cada genotipo en aquello genes candidatos que fueron polimórficos. La primera columna muestra el genotipo, la segunda columna representa un patrón virtual incluido para la compatibilidad en procesamientos posteriores (1), y el resto de columnas muestran los diferentes patrones generados para cada genotipo; b) Archivo PHO que muestra el estado homocigótico (1) o heterocigótico (2) en la combinación de los alelos en cada genotipo.

3.2.5. Estudio básico de la diversidad genética

Los parametros relacionados con la diversidad genética de cada loci y/o haplotipo polimórfico de los genes candidatos bialélicos,como la **frecuencia del mayor alelo**, la **heterocigosidad esperada** (He) y la **observada** (Ho), así como el **contenido de información polimórfica** (PIC; $PIC_i = 1 - \sum p_i^2$; p_i es la frecuencia de los alelos *i* de cada locus) se calcularón utilizando el programa Cervus versión 3.0.7 (Kalinowski y col., 2007). Las frecuencias alélicas y genotípicas se calcularon para cada locus utilizando el programa programa POPGENE v 1.3.1 (Yeh y col., 1997). GeneAlEx version 6.5 (Peakall and Smouse, 2006) se utilizó para calcular el test Chi-cuadrado en cada loci para determinar su desviación del **equilibrio de Hardy-Weinberg** (HWE). En este caso, la hipótesis nula (H_0=0; P valor> 0,005) indica que la población esta en equilibrio, y su apareamiento es al azar. Los valores P obtenidos inferiores a 0,005, después de aplicar la corrección de Bonferroni se consideraron estadísticamente significativos, rechazando la H_0.

La población de estudio es un conjunto de genotipos "*Tenera*" procedentes de 107 familias diferentes cuyos parentales tienes diferentes origenes (7). En este análisis fueron agrupadas las familias por el origen de su parental masculino "*Pisifera*", responsable de la mayor parte de la variabilidad genética en la especie, ya que la base genética del parental femenino "*Dura*" es muy estrecha (Anexo 10;Tabla 10.1). Por ello se realizó un análisis interpoblacional dónde se estimaron los **estadísticos F** (Fit,

Fis, Fst) (Wright, 1949), el **flujo génico** (Nm) (Wright, 1969), y la **matriz de distancia genética** entre las familias (Nei's, 1972) mediante POPGENE V.1.3.1. El dendograma se genéro a partir de esta matriz de distancias genéticas y con el método clúster UPGMA (promedio no ponderado de pares de grupos) utilizando POPGENE V 1.3.1 y editado en http://itol.embl.de/ (Letunic y Bork, 2016). Por último, se ejecutó un **análisis molecular de la varianza** (AMOVA) basado en 999 permutaciones calculando la variación total dentro de la población, entre las diferentes poblaciones agrupadas por el origen del parental "*Pisifera*", y entre los propios individuos mediante el programa GENEAlex v 6.5.

4. RESULTADOS

4.1. Diseño de cebadores

En total se diseñaron 224 parejas de cebadores para los 234 genes seleccionados como candidatos en el capítulo 2. Después de probar estos cebadores en muestras aleatorias de la población fueron válidos 201 parejas de cebadores que se correspondían con 198GC. Los cebadores que no funcionaron correspondían a los transcriptos CDA41 y CDA44, a los genes candidatos co-localizados KG1, KG146, KG163, KG165, KG16, KG170, KG178, TEST, KG203,KG220, KG221, KG222, KG223, KG224, KG225, KG226 Y KG227, y a los genes candidatos conocidos P12, P46, y P75. Estos genes candidatos se eliminaron del genotipado, y por tanto del estudio de mapeo por asociación.

Los genes candidatos que continuaron en el estudio, sus parejas de cebadores, y la longitud de sus amplicones se muestran en la tabla 6.1 del anexo 6.

4.2. Secuenciación de amplicones

Los resultados de la secuenciación de las librerías mediante la plataforma Ion Torrent con el chip 318 y recibidos en el archivo comprimido FASTQ arrojaron un resultado medio de 1,28Gb de secuencias que representaban 6.390.106 lecturas después de filtrar las secuencias policlonales, los dímeros y las secuencias de baja calidad. La figura 19 muestra el gráfico con los valores correspondientes a cada librería en cuanto al número de secuencias obtenidas y sus lecturas después de filtrar los datos iniciales. El número de parejas de cebadores de cada librería fue variable, entre 31 y 61 parejas de cebadores para 242 genotipos. El análisis de correlación de las variables número total de lecturas, nº total de bases y nº de amplicones muestra que hay una correlación fuertemente positiva entre el número total de lecturas y el nº total de bases secuenciados (0,98149 p=0,0005), pero no es significativa la correlación con el número de amplicones incluído en cada librería frente a ninguna de las variables anteriores (Amplicon Vs MPB= 0,3684 p=0,4727 ; Amplicon VS Nº Lecturas =0,36074 p=0,4824).

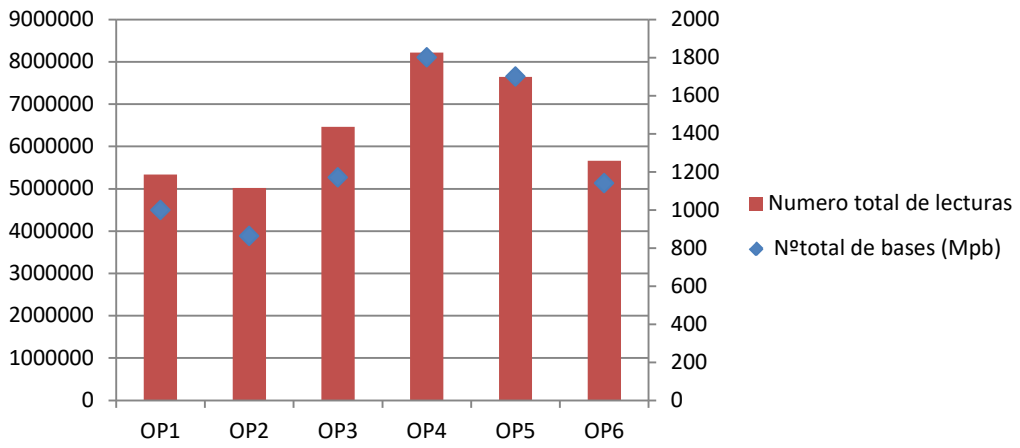


Figura 19: Gráfico de los resultados de cada librería secuenciada (OP1 a OP6).

4.3. Filtrado de las secuencias

Las secuencias de cada librería se filtraron tal y como se describe en el apartado 3.2.4 de materiales y métodos. En la figura 20 puede observarse como este proceso de filtrado disminuyó el número de secuencias disponibles para analizar posteriormente. En las tres primeras librerías el número de secuencias finales disminuyó hasta la mitad de las inicialmente secuenciadas, en cambio la efectividad de la secuenciación mejoró en las librerias OP4, OP5 y OP6, en las que el número de lecturas disminuyó entre un 30 y 40%.

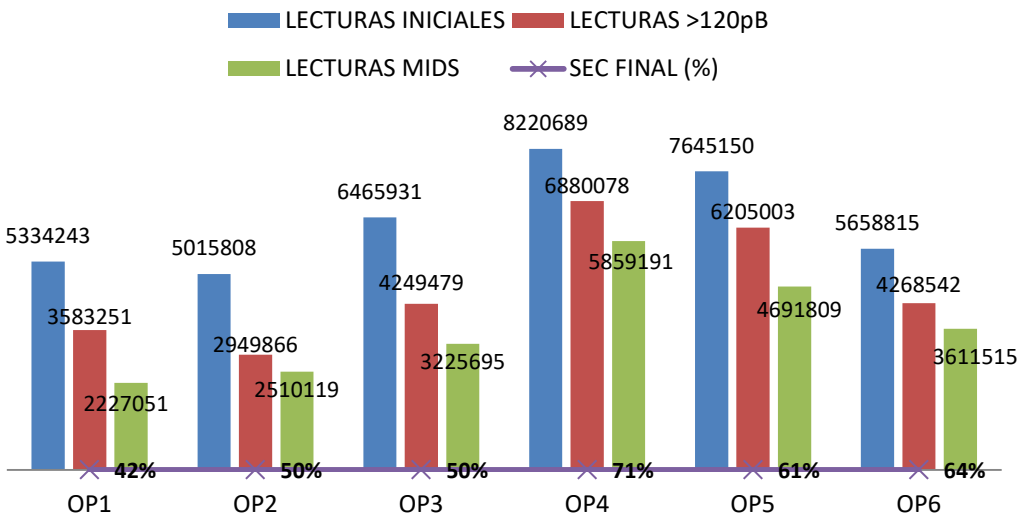


Figura 20: Gráfico de resultados del filtrado de datos en las diferentes librerías. El gráfico muestra el número de lecturas iniciales correspondientes a la secuenciación, el número de lecturas después de aplicar el filtrado para seleccionar aquellas que obtuvieron más de >120pb de longitud, y las lecturas de las secuencias completas que incluyeron los MIDs (identificadores de multiplex). Además también muestra el % del número de secuencias disponibles después de todo el proceso de filtrado en ambos lados.

4.4. Identificación de los genes candidatos

El filtrado por genes candidatos redujo el número de lecturas o las frecuencias para cada GC, ya que se tuvieron en cuenta las 10pb iniciales que fueron exactas en sus cebadores específicos "Forward" y "Reverse". Como se observa en el gráfico de la figura 21, después de la selección de los genes candidatos el número de lecturas disponibles totales disminuyeron como era lo esperado en cada librería. Únicamente en la primera librería (OP1) quedaron disponibles menos del 50% de las lecturas. En cambio, las librerías OP2, OP3 y OP6 conservaron más del 90% de las lecturas. La media del total de las lecturas disponibles después de los primeros pasos de filtrado fue de 3.687.563, y después de la selección de las lecturas correspondientes a cada GC fue de 3.066.992 lecturas. En términos medios las lecturas disponibles después del proceso de filtrado correspondía a un 83%.

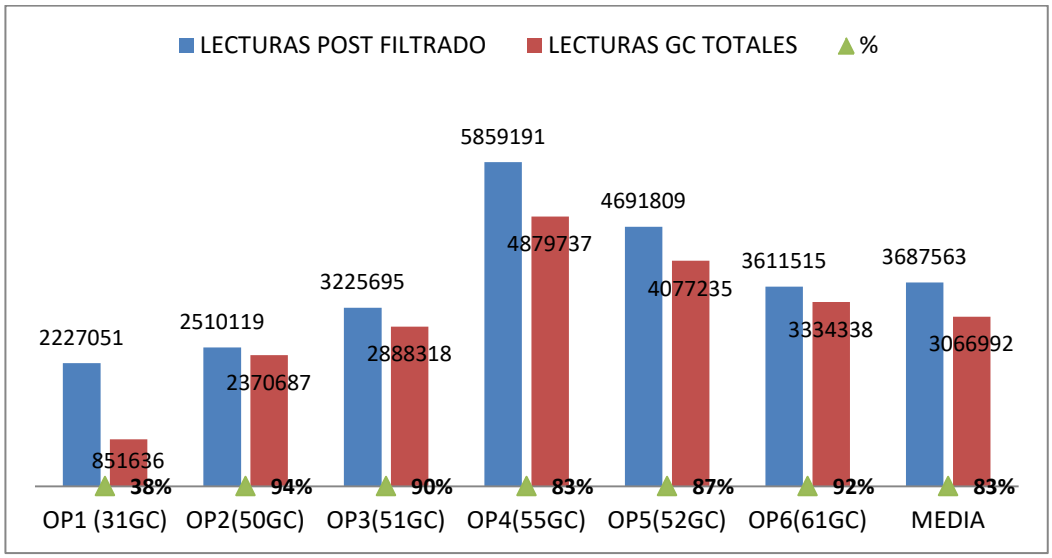


Figura 21: Gráfico de lecturas de genes candidatos en cada OP antes y después del filtrado por gen candidato, así como el % de secuencias que se conservaron.

En relación a los genes candidatos, el número medio de lecturas por cada gen candidato fue diferente en cada librería, siendo la librería OP5 la que más número de lecturas por gen candidato obtuvo, y OP4 la que menos lectura por gen candidato obtuvo. El tamaño medio de los amplicones de cada gen candidato también fue diferente en cada librería siendo la librería OP5 la que mayor tamaño medio de amplicón presentaba y OP1 la que menos (Figura 22).

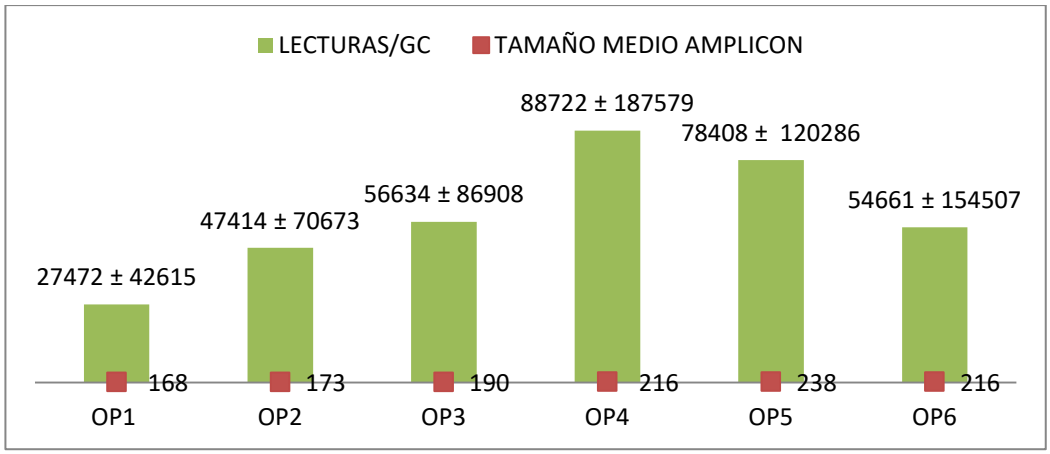


Figura 22: Gráfico de lecturas medias y su desviación estándar por gen candidato en cada librería y tamaño medio de los amplicones de la librería en pares de bases.

Se aplicó un análisis de correlación de Pearson en cada librería para averiguar si había una relación entre el tamaño de los amplicones y el número de lecturas obtenidos. Los resultados (Tabla 5) mostraron una débil correlación negativa, pero significativa ($p<0,05$) en las librerías OP4, OP5 y OP6, lo que indica que el número de lecturas disminuye a medida que aumenta el tamaño de los amplicones.

Tabla 5: Análisis de correlación para cada una de las librerías entre las variables número de lecturas del gen candidato y la longitud de su amplicón.

	Lecturas Vs Tamaño Amplicon
OP1	-0,10479 (p=0,578)
OP2	0,06178 (p=0,6699)
OP3	-0,04028 (p=0,7790)
OP4	-0,30683 (p=0,0227)*
OP5	-0,39645 (p=0,0036)*
OP6	-0,26668 (p=0,0378)*

* p< 0,05 = estadísticamente sifjnificativo.

En el anexo 7, en las tablas 7.1 y 7.2, se muestran las frecuencias o el número de lecturas para cada gen candidato obtenido en la secuenciación de su librería. El número de lecturas presenta una fuerte variabilidad, como puede observarse en la tabla 6, hay una gran diferencia entre el número de lecturas máximo y mínimo en cada librería. El origen de estas diferencias puede ser debido al proceso de multiplexado de la librería, y a las diferentes temperaturas de fusión de los cebadores. O bien por errores de la secuenciación que han hecho que durante el filtrado de datos se pierdan esas lecturas, por no coincidir exactamente los cebadores flanqueantes de los amplicones con la secuencia de origen.

Tabla 6: Genes candidato con el máximo y mínimo de lecturas en cada librería.

LIBRERÍA	GC	MAXIMO	GC	MÍNIMO
OP1	CDA26	151303	KG35	23
OP2	CDA40	267778	KG64	4
OP3	KG126	371435	KG82,KG87, KG64*	1
OP4	KG153	615787	KG271	0
OP5	KG212	326380	KG174*	384
OP6	KG242	852988	KG120	0

* Gen candidato incluído en más de una librería.

4.5. Determinación del conjunto preliminar de patrones en los genotipos

Para llevar a cabo esta detección se filtraron en un primer momento los alelos de los genotipos bajo las siguientes condiciones: diploidía 2x, número mínimo de secuencias totales consideradas (5) y número mínimo de secuencias repetitivas que se consideran para un alelo (2). Este filtrado significa que únicamente se extrajeron de todos las lecturas de los 4 patrones con mayor frecuencia para cada genotipo que tenían un máximo de 5 lecturas con al menos 2 repeticiones de cada secuencia patrón, llamándose a continuación las secuencias repetitivas. En este paso se crearon nuevos archivos, uno de ellos con resultados estadísticos, otro con las secuencias más repetitivas y otro dónde se mostraban los alineamientos múltiples realizados (Tabla 4; Figura 13).

La tabla 7.1 y 7.2 del anexo 7 muestra la frecuencia de los genotipos en los genes candidatos de cada librería. Como puede observarse en los gráficos de la figura 23 no todos los genes candidatos estaban presentes en todos los genotipos, e incluso algunos genes candidato no estaban en ninguno de los genotipos del conjunto de la población como CDA22 (OP1), KG35 (OP1), o KG282 (OP2), KG64 (OP2), CDA6 (OP2), P63 (OP2), P64 (OP2), KG69 (OP2), KG64 (OP3),KG103 (OP3), KG68 (OP3), KG94 (OP3), KG82 (OP3), KG159 (OP4), KG213 (OP5), KG215 (OP5), KG185 (OP5) y KG120 (OP6).

El número medio de genotipos obtenido por gen candidato fue de 171 genotipos, lo que implica su presencia en el 70,6% de la población. La librería con más genes candidatos en todos los genotipos fue la librería 2, con un total de 16 genes candidatos, y la que obtuvo menos presencia en los genotipos de los genes candidatos fue OP6, como también puede observarse en el gráfico de barras de la figura 23.

Figura 23: Gráficos de las frecuencias del conjunto de genotipos que componen la población en los genes candidatos de cada librería. En el eje vertical de cada gráfico se representa el número de genotipos, y el eje horizontal cada gen candidato en cada librería.

Lo más destacado del número de patrones obtenidos (resultados no mostrados) después de realizar la búsqueda de los alelos en el conjunto de la población fue:1. la desaparición del estudio de algunos genes candidatos porque después del filtrado no generaron patrones en la población(KG47, CDA34, CDA37, KG278, CDA3, KG35, KG86, KG93, KG113 y KG115), y 2. algunos genes candidatos mostraron un único patrón (KG29, P39, KG24, P6, KG128, KG69 KG152, KG268, KG270, KG213, KG248, KG249, KG251 y KG257). Todos los patrones de los genes candidatos se revisaron en el siguiente análisis manualmente, tal y cómo se describe en el apartado 3.2.4 de materiales y metodos de este capítulo.

4.6. Determinación final de patrones y composición alélica de la población

El total de amplicones que llegaron a esta fase final fueron 213(71%), teniendo en cuenta los GC candidatos secuenciados en más de una librería por errores en su anterior librería.. Estos resultados mostraron 19 genes candidatos sin ningún patrón (Tabla 8.1, Anexo 8) en la población, 66 poseían un único patrón (Tabla 8.2; Anexo 8), 81 genes candidatos tenían 2 patrones o alelos (Tabla 8.3; Anexo 8), y 36 de ellos más de dos patrones en la población (Tabla 8.4; Anexo 8) (Figura 24a).

El número total de SNP e Indels obtenidos en los genes candidatos con 2 o más patrones fue de 401, de los que 391 fueron SNP y 10 fueron Indels, de los cuáles 4 se obtuvieron en los GC con 2 patrones y 6 en los de más de 2 patrones. La tabla 7 muestra el total de SNP obtenidos en los GC que mostraron 2 patrones, con un ratio de 2SNP por GC, y en los GC con más de 2 patrones con 6SNP de media por GC. Como puede revisarse en el anexo 8, tabla 8.3, los genes candidatos que obtuviero mayor número de SNPs fueron los genes conocidos P74 y KG286 (2patrones) con 18 SNP. 25, 24 y 23 SNP se detectaron en los transcritpos CDA4, CDA9 y en el gen co-localizado KG189, respectivamente para aquellos GC con más de dos patrones.

Cómo se observa en los gráficos de la figura 25, el 59% de los SNP fueron transiciones con el cambio de base más frecuente C/T (Gráfico a y b), siendo más frecuente esta transición en los genes candidatos con 2 patrones que en los genes candidatos con más de 2 patrones, como puede observarse en la tabla 7. La transversión ocurrió en el 41% de los SNPs, siendo el cambio de base más frecuente A/T (29,81%) (Gráfico a y c), y observando en la tabla 7 puede comprobarse que el cambio más frecuente en los GC con dos patrones fue G/C (25), mientras que para A/T ocurrió en 18 ocasiones. El ratio de Tr/Tv total fue de 1,43.

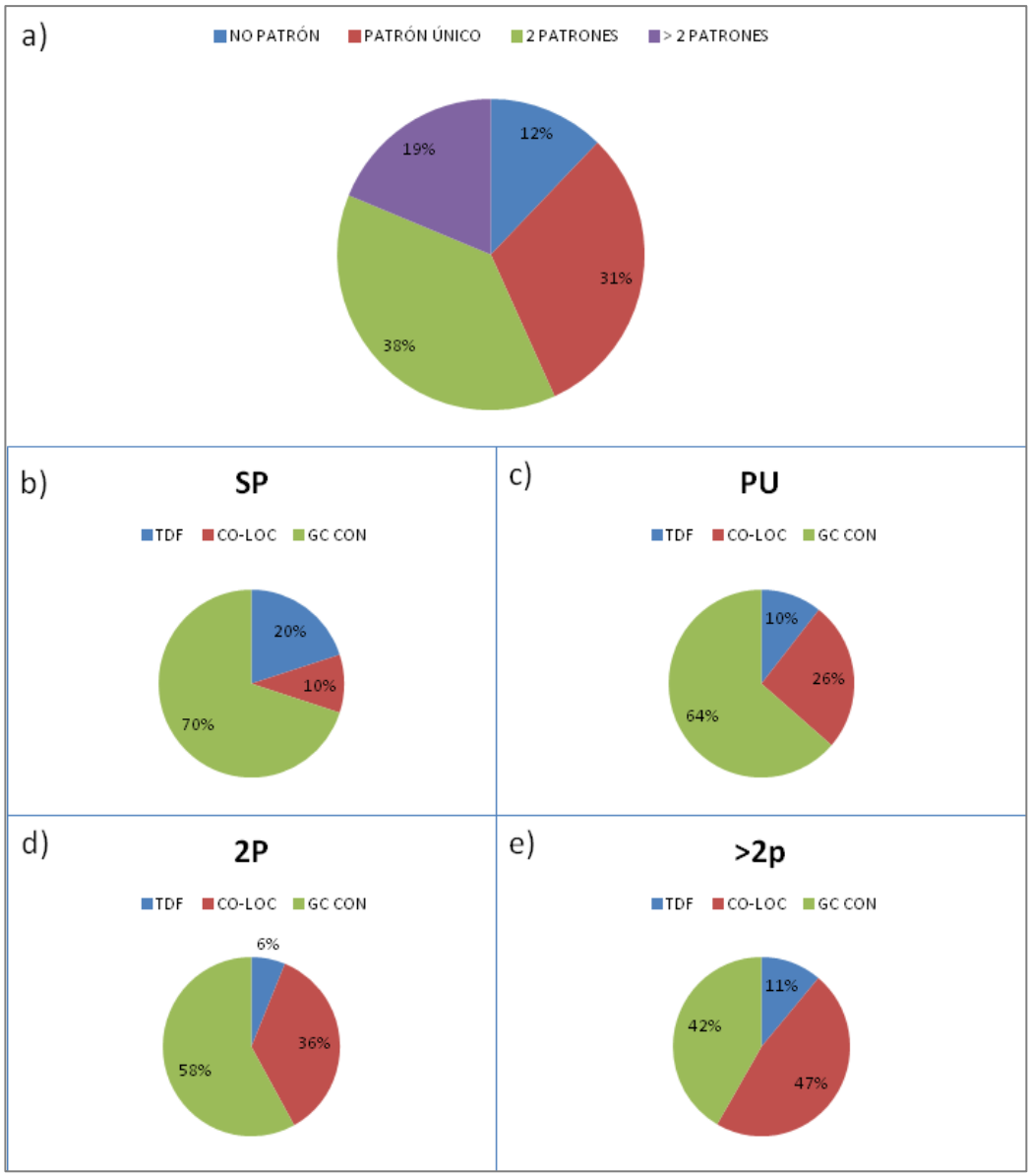


Figura 24: Gráficos de resultados de la determinación manual de los patrones de los genes candidatos.a) Conjunto total de los resultados obtenidos. El 12% de los GC no mostraron ningún patrón, el 31% mostró un único patrón, 2 patrones mostraron el 38% de los genes candidatos, y el 19% mostró más de 2 patrones; b) Gráfico en función del origen de los GC que no mostraron ningún patrón (SP). El 20% de estos GC correspondía a los TDF obtenidos en el experimento de cDNA-AFLP, el 10% a genes co-localizados y el 70% a genes conocidos; c) Gráfico en función del origen de los GC que obtuvieron un patrón único (PU). El 10% de los GC procedía de transcriptos, el 26% de genes co-localizados y el 64% a genes conocidos.d) Gráfico en función del origen de los GC que obtuvieron dos patrones (2P). El 6% de los GC que obtuvieron 2 patrones correspondían a transcriptos, el 36% a genes co-localizados en el mapa y el 58% a genes conocidos; e) Gráfico en función del origen de los GC que obtuvieron más de dos patrones (>2P). El 11% de los GC que obtuvieron más de dos patrones tenían su origen en los transcriptos, el 47% eran genes co-localizados y el resto tenían su origen en los genes conocidos.

Tabla 7: SNP detectados y sus tipos.

	SNP	Tr		Tv				INDEL
		A/G	C/T	G/T	G/C	A/C	A/T	
2 PAT	179	42	55	20	25	19	18	4
>2 PAT	212	71	62	17	17	15	30	6
TOTAL	**391**	**113**	**117**	**37**	**42**	**34**	**48**	**10**
	TOTAL Tr	**230**		**TOTAL Tv**	**161**		**Ratio Tr/Tv**	**1,43**

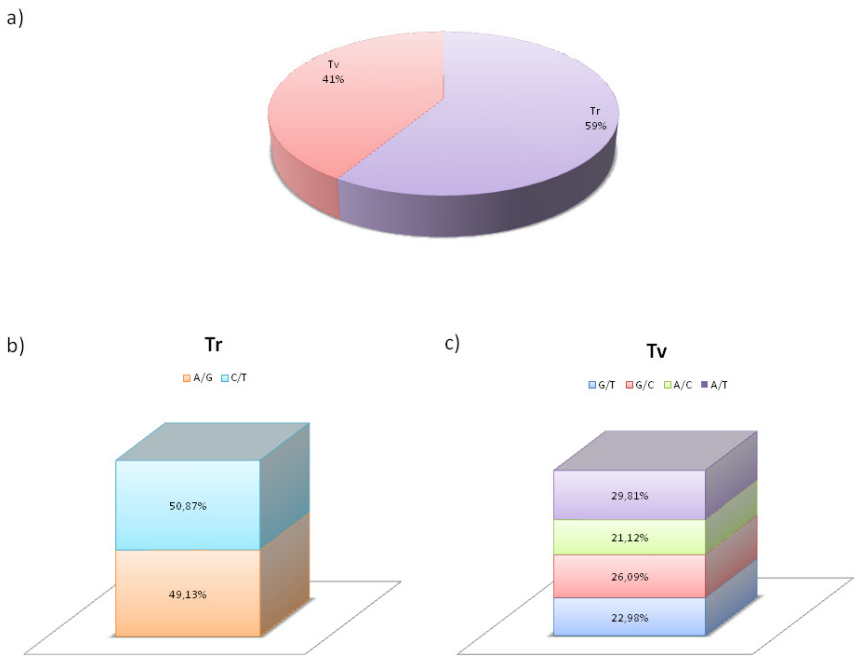


Figura 25: Clasificación de los SNPs; a) Porcentaje de transiciones y transversiones totales;b) Transiciones totales (A/G o C/T); c) Transversos totales (G/T, G/C, A/C y A/T).

El conjunto total de bases de los amplicones secuenciados correspondía a 48510, y se obtuvó 1SNP por cada 122 pares de bases. Respecto al total de genes candidatos, tomando como dato el número de amplicones secuenciados que obtuvieron patrón (203), se obtuvieron 1,95 SNPs por amplicón.

4.7. Asociación de los patrones a los genotipos de la población: valores perdidos y frecuencias genotípicas.

Los patrones de cada gen candidato fueron asociados con cada genotipo individual mediante la opción de ASPAM de coincidencia de la región dónde se sitúan los SNPs e Indels, cuya longitud son el SNP/Indel y las 3pb adyacentes, en sus secuencias. Los resultados de la asociación de los patrones con los genotipos individuales mostraron la presencia de valores perdidos en algunos genotipos de determinados genes candidatos.

En total se obtuvieron 26632 combinaciones de alelos para 117 genes candidatos que obtuvieron 2 o más patrones en el conjunto de los 242 genotipos, y los valores perdidos fueron un 5,94% (1650 valores perdidos en diferentes genotipos y genes candidatos), con una media de 15 GT/GC, lo que supone 6,19% de los genotipos de cada gen candidato. Estas combinaciones de alelos fueron los snp o conjuntos de snp (haplotipos) característicos de la población (conjunto de 242 genotipos) detectados en cada gen como se presenta en el anexo 9, tablas 9.1, 9.2, y 9.3.

Aquellos genotipos que no mostraron combinaciones alélicas se consideraron como valores perdidos. Como se muestra en el gráfico de la figura 26 de los 117 GC, 91GC obtuvieron menos del 5% de valores perdidos en el conjunto de genotipos, siendo de ellos 29 los que mostraron diferentes combinaciones de alelos en todos los genotipos, como puede revisarse en la tabla 9.4 del anexo 9. Los GC KG191 (2 patrones), KG233 (>2 patrones) y KG140 (2 patrones), fueron los que mayor número de valores perdidos en sus genotipos con un porcentaje superior al 50% (Figura 26; Anexo 9, tabla 9.4).

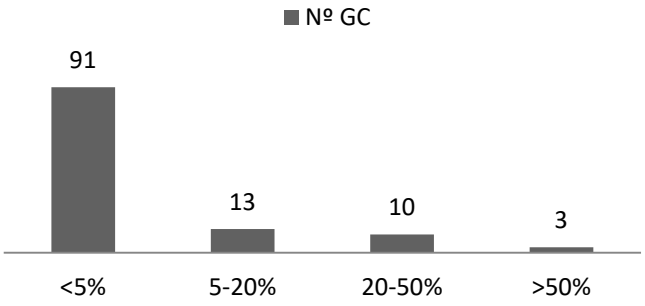


Figura 26: Gráfico del número de GC agrupados por su % de valores perdidos en el conjunto de la población agrupados en diferentes rangos.

Una vez revisados los genotipos y su composición alélica, un total de 34 genes candidatos (33%) mostraron la presencia de más de dos alelos (multi-locus) en diversos genotipos. De estos genes candidatos el 43,6% pertenecían a los genes candidatos co-localizados, 44,1% a genes candidatos conocidos, y 5,9% a genes candidatos cuyo origen fueron los transcriptos. Estos genes candidatos fueron desacartados del estudio de asociación que se desarrolla en el siguiente capítulo (Tabla 8).

El estudio continúo con los genes candidatos bialélicos que estaban presentes en al menos el 95% de la población, para los cuales se muestran los resultados de sus frecuencias genotípicas en función del origen de la selección del gen candidato (Figura 27). Los resultados muestran que de los 65 genes candidatos, 45 genes candidatos presentaban más del 50% de sus genotipos en estado heterocigótico. Los datos de los genotipos observados se representaron en el gráfico de la figura 38, agrupándolos en función de su origen. En los genes candidatos co-localizados se observa que 4 de ellos (KG187, KG11, KG143 y KG195) tenían más del 90% de sus genotipos en estado heterocigótico, y sólo KG179 presentaba menos 10% de sus genotipos en el mismo estado (Figura 27). En el caso de los genes candidatos cuyo origen fueron transcriptos todos los genes candidatos todos presentaron un valor superior al 90% de heterocigótos, siendo CDA26 y CDA40, los que mostraron un 100% de hetrocigosis

(Figura 38). Once de los genes candidatos conocidos mostraron un 90% de individuos hetrocigóticos, siendo KG243 y KG290, los que estaban en completo estado de heterocigosis en el conjunto de la población de estudio. En cambio 3 genes candidatos, KG155, KG272 y KG282 presentaron más del 90% en homocigosis, siendo para KG155 la presencia mayoritaria de alelo A2, y para KG272 y KG282 su alelo A1 (las combinaciones de sus haplotipos se muestran en la tabla 9.1 del anexo 9) (Figura 27).

Tabla 8: Genes candidatos multi-locus (GC_CEB= Nombre del gen candidato y de sus cebadores; ORIGEN= procedencia de la selección del gen candidato; CO-LOC=colocalizado; GCon=Gen candidato conocido; TDF= transcripto)

GC_CEB	ORIGEN	GC_CEB	ORIGEN
KG144_M23551	CO-LOC	**KG153_ATAGB1**	GCon
KG138_M9619	CO-LOC	**P44_M6ASA**	GCon
KG141_M847	CO-LOC	**KG114_JC41**	GCon
KG162_HtC1_5925	CO-LOC	**KG217_PDAT_2**	GCon
KG166_HtC2_7081	CO-LOC	**KG212_EgFATB2.2**	GCon
KG167_HtC2_1255C2-411	CO-LOC	**KG233_ASP1**	GCon
KG173_HtC7_9200	CO-LOC	**KG234_ASP2**	GCon
KG185_FFB2_C6_S1.2	CO-LOC	**KG242_EgETR_F2R2**	GCon
KG189_FFB8_C545_S1	CO-LOC	**KG244_EgMAX4_F1R1**	GCon
KG190_FFB8_C1455_S3.4.5.6	CO-LOC	**KG246_EgARF1**	GCon
KG197_BnC2_1289	CO-LOC	**KG250_EgPINF3-6_PIN1**	GCon
KG204_BnC10_7131	CO-LOC	**KG252_AIL5**	GCon
KG205_BnC12_2975	CO-LOC	**KG271_VIR**	GCon
KG264_PO3_5-13	CO-LOC	**KG21_PACT**	GCon
KG258_PO3_5-7	CO-LOC	**KG31_atpB**	GCon
KG164_ HTC2_11412	CO-LOC	**CDA4_B3/B12**	TDF
KG177_HTC10_11102	CO-LOC	**CDA9_B25**	TDF

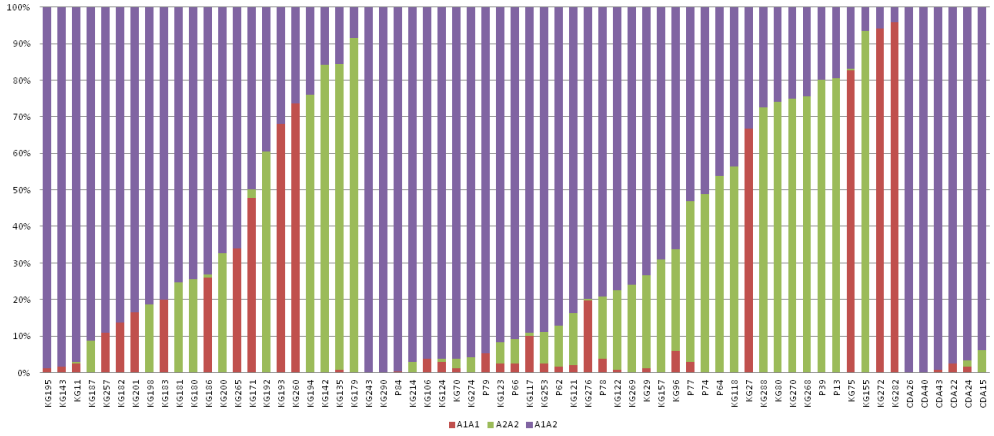


Figura 27: Gráficos de las frecuencias de las combinaciones alélicas mostradas mediante sus frecuencias genotípicas (%) para los genes candidatos bialélicos agrupados en función de su procedencia. Los GC se ordenande la siguiente manera: GC co-localizados (KG195-KG179), GC conocidos (KG243-KG282) y GC procentes de TDF (CDA26-CDA15).

4.8. Estudio básico de la diversidad genética

El resumen de los los estadisticos utilizados para el estudio de la diversidad genética en los genes candidatos bialélicos que continuaron en el estudio se muestran en la tabla 8. Es de destacar que de los 136 SNP presentes en estos 64 genes candidatos, 45 SNP correspondían a genes candidatos individuales, es decir a un único loci. Los SNPs situados en regiones muy cercanas constituyen un haplotipo. En este estudio, hay 19 genes candidatos (loci) en los que hay más de una posición polimórfica cercana (Tabla 8, genes candidato señalizados con superindice 9), mostrando la presencia de diferentes haplotipos, como puede revisarse en la tabla 9.1, anexo 9 dónde se muestran los patrones para esos genes candidato. Por ejemplo, el gen candidato conocido KG243, presenta un haplotipo constituido por los SNPs posicionados en las bases 30, 42, 46,y 93.

Los locus que mostraron alelos raros, clasificados como aquellos alelos con una frecuencia <0,05 fueron los correspondientes a los GC KG282, KG272, KG155 y KG179. La media de la heterocigosidad esperada y de la observada fue 0,398 (±0,317) y 0,666 (±0,143), respectivamente, siendo los intervalos de Ho de 0,042(KG282) y 1 (KG243, KG290, CDA26 y CDA40) y de He entre 0,041 (KG282) y 0,501 (P79 a CDA40). El contenido de información polimórfica o PIC varió entre el 0,040 (KG282_ ACYL-ACPF) y 0,375 (CDA40_B64), con una media de 0,306 (±0,098). Todos los loci, excepto 9, mostraron desviaciones significativas del equilibrio de Hardy- Weinberg (p<0,05).

Tabla 9: Resultados de los estadísticos de diversidad genética basados en los SNPs y haplotipos de lo genes candidatos bialélicos (N=242). MAF=Alelo de Menor Frecuencia; Ho= Heterocigosidad Observada; He= Heterocigosidad Esperada; PIC=Contenido de Información Polimórfica; HW (Equilibrio Hardy-Weinberg) (P-Valor)= P valor obtenido después de aplicar el test chi-cuadrado (Pearson K., 1900); Signif= Significancia (ns =no significativo; * P<0.05, ** P<0.01, *** P<0.001;. DE= desviación estándar.

Locus	MAF	Ho	He	PIC	HW P_valor	Signif
KG282_ACYL-ACPF	0,021	0,042	0,041	0,04	0,741	ns
KG272_OLEOYL	0,029	0,058	0,057	0,055	0,642	ns
KG155_ELO2	0,033	0,066	0,064	0,062	0,595	ns
KG179_FFB1_CL1016_S1.2	0,042	0,085	0,081	0,078	0,497	ns
KG142_M2552	0,079	0,158	0,146	0,135	0,183	ns
KG135_M3117a	0,087	0,156	0,158	0,146	0,852	ns
KG75_GID1	0,088	0,168	0,161	0,148	0,508	ns
P13_QM	0,098	0,195	0,176	0,161	0,093	ns
P39_WOS6942	0,099	0,198	0,179	0,163	0,087	Ns
KG194_FFB11_C3877_S4	0,120	0,24	0,212	0,189	0,037	*
KG268_EgMBAGL2-3	0,122	0,244	0,215	0,191	0,031	*
KG270_EgPPGLa	0,125	0,25	0,219	0,195	0,027	*
KG80_EIN4	0,130	0,26	0,227	0,201	0,020	*
KG288_EgTPase	0,137	0,274	0,237	0,208	0,014	*
KG193_FFB11_C1741_S3,4	0,160	0,32	0,269	0,232	0,003	**
KG27_BKACPII_1	0,166	0,332	0,277	0,239	0,002	**
KG192_FFB11_C1_S1	0,197	0,394	0,317	**0,266**	<0,001	***
KG118_JC59a	0,218	0,437	0,342	**0,283**	<0,001	***

P64_PAT4	0,231	0,462	0,356	**0,292**	<0,001	***
P74_PAT9a	0,256	0,512	0,382	**0,308**	<0,001	***
KG171_HtC4_4489	0,274	0,498	0,399	**0,319**	<0,001	***
P77_PAT11	0,295	0,53	0,417	**0,33**	<0,001	***
KG265_PO3_5-14	0,330	0,661	0,443	**0,345**	<0,001	***
KG200_BnC7_3962	0,336	0,672	0,447	**0,347**	<0,001	***
KG157_RPL10	0,345	0,69	0,453	**0,350**	<0,001	***
KG180_FFB2_C4663_S1,2	0,372	0,744	0,468	**0,358**	<0,001	***
KG181_FFB2_C4741_S3a	0,376	0,752	0,47	**0,359**	<0,001	***
KG186_FFB6_C2082_S1[a]	0,374	0,731	0,469	**0,359**	<0,001	***
KG269_EgNAC	0,380	0,759	0,472	**0,360**	<0,001	***
KG29_GLUT1[a]	0,380	0,734	0,472	**0,360**	<0,001	***
KG96_FA8	0,391	0,662	0,477	**0,363**	<0,001	***
KG122_DEF1	0,396	0,775	0,479	**0,364**	<0,001	***
KG183_FFB2_C2_S1	0,400	0,801	0,481	**0,365**	<0,001	***
KG276_PYRKIN	0,403	0,798	0,482	**0,365**	<0,001	***
KG198_BnC3_792	0,406	0,813	0,483	**0,366**	<0,001	***
KG201_BnC8_761	0,417	0,834	0,487	**0,368**	<0,001	***
KG182_FFB2_C3566_S9	0,432	0,863	0,492	**0,370**	<0,001	***
KG121_MADS11-1	0,440	0,838	0,494	**0,371**	<0,001	***
P78_PAT12a	0,434	0,793	0,492	**0,371**	<0,001	***
KG257_PO3_5-5	0,446	0,891	0,495	**0,372**	<0,001	***
KG117_JC55	0,454	0,891	0,497	**0,373**	<0,001	***
KG187_FFB6_C3684_S1	0,456	0,913	0,497	**0,373**	<0,001	****
P62_PAT2a	0,452	0,871	0,496	**0,373**	<0,001	***
CDA15_B33	0,469	0,938	0,499	**0,374**	<0,001	***
KG253_ANT	0,470	0,89	0,499	**0,374**	<0,001	***
P79_PAT13	0,473	0,946	0,500	**0,374**	<0,001	***
CDA24_B44a	0,500	0,967	0,501	**0,375**	<0,001	***
CDA43_B67	0,496	0,992	0,501	**0,375**	<0,001	***
KG106_PSII2a	0,481	0,963	0,500	**0,375**	<0,001	****
KG11_M8373a	0,490	0,971	0,501	**0,375**	<0,001	***
KG123_GLO2a	0,483	0,917	0,500	**0,375**	<0,001	***
KG124_AG1	0,490	0,963	0,501	**0,375**	<0,001	***
KG143_M3256a	0,492	0,983	0,501	**0,375**	<0,001	****
KG195_FFB13_C2168_S1	0,494	0,988	0,501	**0,375**	<0,001	***
KG214_EgWRI1,1	0,485	0,971	0,501	**0,375**	<0,001	***
KG243_EgEBFa	0,500	1,000	0,501	**0,375**	<0,001	***
KG274_LIPOIC	0,479	0,959	0,500	**0,375**	<0,001	***
KG70_BAK1a	0,494	0,962	0,501	**0,375**	<0,001	***
KG290_EgDSI	0,500	1,000	0,501	**0,375**	<0,001	***
P66_PAT6	0,479	0,909	0,500	**0,375**	<0,001	***
P84_PAT14	0,498	0,996	0,501	**0,375**	<0,001	***
CDA22_B42a	**0,488**	**0,975**	**0,501**	**0,375**	**<0,001**	****
CDA26_B46a	0,500	1,000	0,501	**0,375**	<0,001	***

CDA40_B64a	**0,500**	**1,000**	**0,501**	**0,375**	**<0,001**	*******
	Media	**0,666**	**0,398**	**0,309**		
	DE	**±0,317**	**±0,143**	**±0,098**		

[a]Haplotipos

Tabla 10: Resultados de los estadísticos utilizados a nivel poblacional para la diversidad genética (N=242). Fis= coeficiente de endogamia del individuo respecto a la subpoblación; Fit= coeficiente de endogamia de un individuo respecto a la población total; Fst= efecto de las subpoblaciones comparando con la población total.Nm= Flujo génico

ORIGEN PARENTAL "*Pisifera*"	Fis	Fit	Fst	Nm
AVROS	-0,8800	-0,7569	0,0654	3,5699
DAMI	-0,8733	-0,7576	0,0617	3,7997
EKONA	-0,8763	-0,7417	0,0717	3,2438
GHANA	-0,9250	-0,7723	0,0793	2,9008
LAME	-0,9356	-0,7640	0,0887	2,5697
NIGERIA	-0,8581	-0,7214	0,0735	3,1492
YANGAMBI	-0,8486	-0,7393	0,0591	3,9800
MEDIA	**-0,8868**	**-0,7509**	**0,0719**	**3,3162**
DE	**0,0325**	**0,0179**	**0,0105**	**0,5003**
CONJUNTO POBLACIONAL	**-0,8875**	**-0,7326**	**0,0782**	**2,9469**

En cuanto a los estadísticos utilizados a entre familias de un mismo origen y a nivel poblacional, los resultados para los estadísticos F (Tabla 10) mostraron valores negativos en todas las familias y en el conjunto de la población en Fis y Fit, por lo que la endogamia no aparece a nivel de subpoblaciones dentro de cada familia ni a nivel de cada familia, como tampoco lo hace a nivel poblacional. Todos los valores obtenidos a este respecto son muy similares con un media para el estadístico Fis de -0,8868 ±0,0325 y para el estadístico Fit de -0,7509 ±0,0179. Para el estadístico Fis el menor valor se obtuvo para las familias cuyo parental es LaMe (-0,9356) y el mayor para las familias Yangambi (-0,8486). En cambio, para el estadístico Fit el menor valor correspondió a las familias Ghana (-0,7723) y el mayor para Nigeria (-0,7214). A nivel del conjunto de la población los valores fueron de -0,8875 (Fis) y -0,7326 (Fit). Para el estadístico Fst que compara cada subpoblación frente a la población total midiendo la disminución de la heterocigosis debida a la diferenciación genética el valor medio obtenido para las familias fue de 0,0719±0,0105, siendo el valor mínimo 0,0591 (Yangambi) y el máximo para LaMe (0,0887). En el conjunto total de la población el valor obtenido también fue próximo a 0 (0,0782).

El resultado del flujo génico (Nm) (Tabla 10) a nivel de familias mostró valores entre 2,5697 (LaME) y 3,9800 (Yangambi), siendo el valor medio obtenido de 3,3162±0,5003. Nm para el conjunto poblacional fue inferior (2,9469).

El dendrograma obtenido mediante el algoritmo UPGMA, tomando cada familia (106 familias) como unidad de análisis mostró 3 agrupamientos principales bien diferenciados (Figura 28a). El primer y

segundo agrupamiento correspondia a dos únicas familias, Fam_668 NIG y Fam_697 GH, y el tercer cluster agrupó las 104 familias restantes, diferenciando 8 subgrupos, donde se mezclan las familias con diferentes origenes en su parental masculino "*Pisifera*". El dendograma obtenido a partir de las distancias genéticas para el conjunto de la población clasificada por el origen en su parental masculino (Figura 28b), mostró tambien 3 agrupamiento diferenciados. Las familias de origen Ghana y LaMe se agrupaban en el cluster I, en el cluster II, Nigeria y Yangambi, y en el último cluster Ekona, Avros y Dami. En este último, las distancias genéticas entre estos grupos (Nei, 1978) y la identidad genética entre los mismos se muestran en la tabla 11. Como puede observarse la máxima identidad genética se da entre las familias Yagambi y Nigeria (1,0019) y Dami y Avros (1,0016), y la menor identidad genética entre Nigeria y Ghana (0,9944), y Ghana y Ekona (0,9946).

Tabla 11: Matriz de identidad genética Nei's (1978) (encima de la diagonal) y distancia genética (debajo de la diagonal) entre los grupos de diferentes origenes en el conjunto de la población subdividida por el origen de su parental.

Pob ID	**AVROS**	**DAMI**	**EKONA**	**GHANA**	**LAME**	**NIGERIA**	**YAGAMBI**
AVROS		1,0016	1,0008	0,9981	1,0009	0,9971	1,0000
DAMI	-0,0016		1,0010	0,9948	1,0029	0,9981	0,9990
EKONA	-0,0008	-0,0010		0,9946	0,9990	0,9974	0,9990
GHANA	0,0019	0,0053	0,0054		1,0007	0,9944	0,9969
LAME	-0,0009	-0,0029	0,0010	-0,0007		0,9980	0,9992
NIGERIA	0,0029	0,0019	0,0026	0,0056	0,0020		1,0019
YANGAMBI	0,0000	0,0010	0,0010	0,0031	0,0008	-0,0019	

Además, el estudio de la AMOVA (Análisis Molecular de la Varianza) explicó que únicamente el 1% del total de la variación genética es debida a la variación ocurrida entre las diferentes poblaciones, atribuyendo el resto de la variación genética a la variación ocurrida dentro de los individuos (Tabla 12).

Tabla 12: AMOVA de las 7 poblaciones basadas en el origen de los parentales Pisifera basadas en los las combinaciones alélicas obtenidas a partir de los marcadores SNP bialélicos.

Origen	**g.l**	**Suma de los cuadrados**	**Media de los cuadrados**	**Variacion estimada**	**%**
Entre poblaciones	6	69,469	11,578	0,103	1%
Entre individuos dentro de la población	235	1120,230	4,767	0,000	0%
Dentro de los individuos	242	4885,000	20,186	20,186	99%
Total	483	6074,698		20,289	

*P ≤0,05

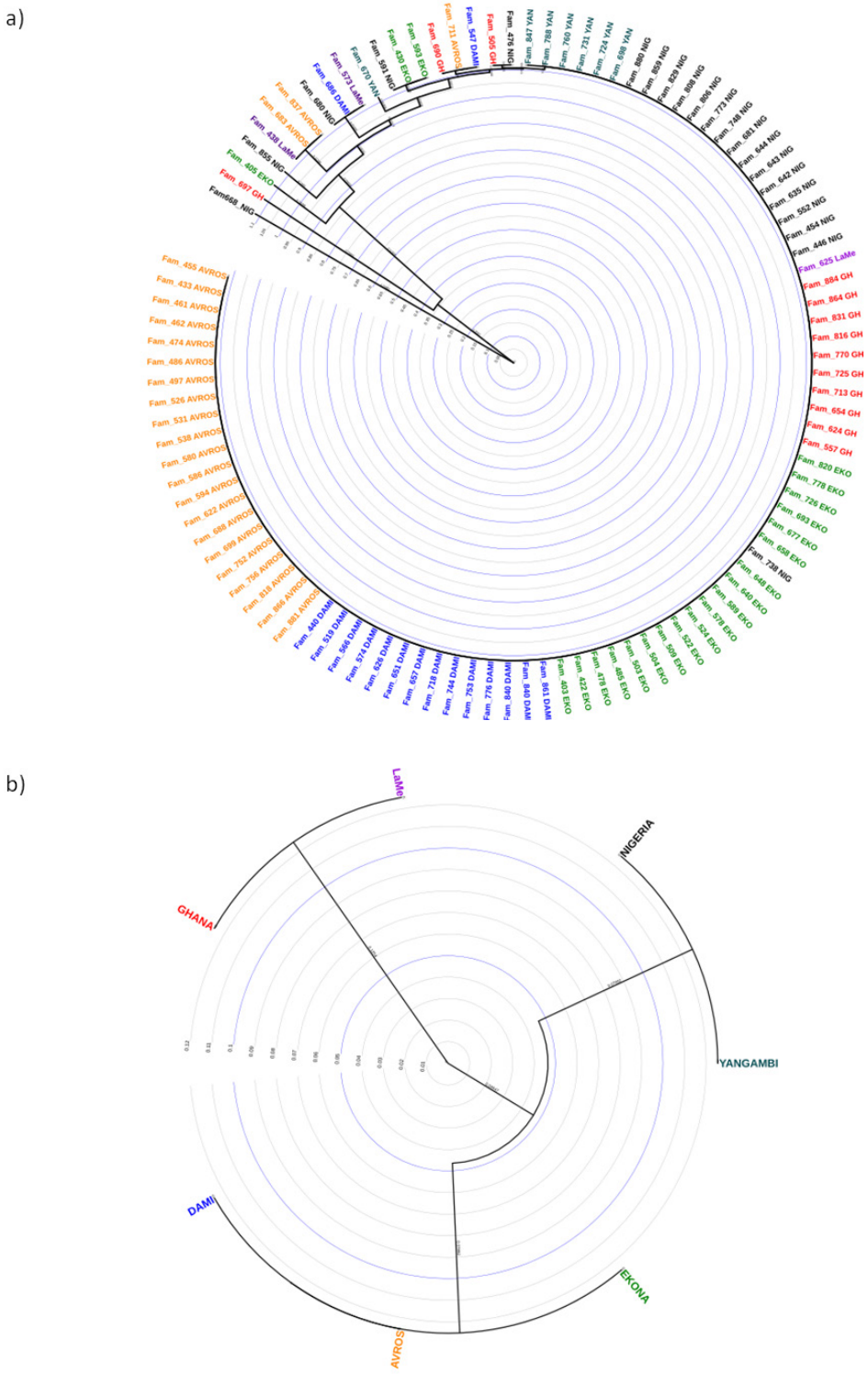


Figura 28: Dendograma realizado mediante la matriz de distancia genética, utilizando el método de agrupamiento UPGMA para la población agrupada por familias (a), y para la población agrupada por origen de su parental masculino "*Pisifera*"(b).

5. DISCUSIÓN

En el presente capítulo se ha desarrollado la identificación de los patrones y SNPs en regiones parciales de los genes candidatos y el genotipado de esos polimorfismos en la población de palmera de aceite africana sujeta a estudio, ya que es necesario identificar los patrones y SNPs presentes entre las líneas de estudio y dentro de cada gen candidato (Zhu y col., 2008) antes de realizar el estudio por asociación mediante genes candidatos.

5.1. Diseño de cebadores

Las regiones de los genes candidatos elegidos en el capítulo 2 por su posible implicación en los caracteres de interés agronómico fueron seleccionadas mediante **dos estrategias** diferentes dependiendo de su origen para el diseño de los cebadores. Estas estrategias fueron: 1. La selección de **regiones conservadas** en los genes candidatos mediante comparación de sus secuencias por BLASTn de los ESTs *de E. guineensis* y *P. dactylifera*, para los genes candidatos conocidos, co-localizados y procedentes del análisis del transcriptoma mediante cDNA-AFLP, y 2. La **re-secuenciación** de las regiones de los genes candidatos co-localizados, cuyos polimorfismos habían sido previamente identificados en otros proyectos para poblaciones diferentes de palmera de aceite africana, incluyendo en los diferentes amplicones las regiones donde ya se habían identificado polimorfismos previamente. Estas estrategias fueron exitosas ya que permitieron diseñar 224 parejas de cebadores (234 genes candidatos iniciales) en las regiones exónicas de los genes candidatos, de las cuales amplificaron correctamente en el ADN genómico 201, correspondientes a 198 genes candidatos. El no funcionamiento de alguno de los cebadores de los genes candidatos pudo tener su causa en errores en el diseño de los cebadores que no permitió una hibridación correcta de los mismos, o aquellos que mostraron bandas de tamaño superior a lo esperado en la posible presencia de intrones en la región del gen candidato seleccionada, como ocurrió en el gen co-localizado KG1, y aquellos que mostraron una amplificación inespecífica podría indicar la presencia de múltiples regiones en el genoma de la palmera de aceite africana similares a los cebadores diseñados ya que pueden pertenecer a familias multigénicas, como por ejemplo KG146, caracterizado como una histona H2B, que pertenece a una familia de genes (Chabeouté y col., 1993).

5.2. Resultados de la secuenciación

La **plataforma** de secuenciación **Ion Torrent** utilizada durante el desarrollo de esta tesis para la secuenciación masiva de los amplicones de los genes candidatos ofrece numerosas ventajas frente a otras plataformas como un bajo coste por base y una rápida salida de datos (Galindo-Gonzalez y col., 2015). Estudios que comparan diferentes tecnologías de secuenciación han confirmado que a pesar de tener una tasa de error superior a otras plataformas, cuando la **cobertura de la secuenciación** es suficiente, Ion Torrent posee una capacidad y representatividad similar a otras plataformas, como por ejemplo la plataforma Illumina para detectar SNPs verdaderos mostrando incluso un menor ratio de falsos positivos (Quail y col., 2012; Galindo-Gonzalez y col., 2015). Para garantizar la cobertura de secuenciación y minimizar los posibles errores se utilizó el **Chip Ion 318™** cuya capacidad de salida de

bases puede alcanzar los 2Gb para tamaños de lecturas de 400pb (www.thermofisher.com), ya que como postularon Galindo-González y col. (2015) cuando se aumenta la intensidad de lecturas disminuye el ratio de falsos positivos en la identificación de variaciones de un solo nucleótido.

Los **resultados** para **cada librería** secuenciada fueron **diferentes** tal y como se esperaba, ya que como agrupó Cronn y col. (2012) son numerosos los factores que pueden influir en las estrategias de secuenciación basadas en el enriquecimiento de secuencias mediante PCR. Estos **factores** que pueden producir **errores** son debidos a: 1. amplificaciones de regiones no diana (pseudogenes o parálogos), eventos recombinatorios y sesgos debidos a PCR. 2. dificultades en la precisión de la cuantificación y en las mezclas equimolares de los amplicones generados; 3. fallos en las reacciones individuales por baja calidad en la muestra o a la presencia de contaminantes inhibitorios. También, influye en esta estrategia la utilización de reacciones multiplexadas. Para **minimizar** estos **errores** a medida que se desarrolló el experimento el procesó se optimizó. En todas las librerías se partió de la misma cantidad de ADN (20ng) para evitar diferentes concentraciones de amplificados, y obtener mezclas de las reacciones multiplex lo más similares posibles como postularon Walsh y col. (1992), pero con el objetivo de estandarizar la concentración final de cada librería las reacciones fueron cuantificadas antes de su mezcla y después de la ligación de los adaptadores e identificadores del genotipos (MIDS) como se describe en el apartado 3.2.3 de Materiales y métodos. La disminución de la presencia de estructuras secundarias o dímeros de cebadores formados durante la reacción a causa de la longitud de los cebadores se mejoró mediante la purificación en columnas por selección de tamaño, y los artefactos originados por las múltiples reacciones de PCR mediante diluciones de la mezcla de la muestra de amplificados(Materiales y métodos).

Un **factor** que no se tuvo en cuenta durante las preparaciones de las librerías, y fue **relevante** en otros estudios de secuenciación dónde se utilizó esta plataforma, fue el **contenido de GC** (Guanina-Citosina) de los amplicones o del genoma de estudio. Un alto o bajo contenido de GC disminuye la cobertura de la secuenciación (Merriman y col., 2012; Ross y col., 2013). Quail y col. (2012) compararon los resultados de tres tipos de plataformas de secuenciación, e Ion torrent mostró una cobertura más irregular que el resto de plataformas en el genoma rico en GC (67,7%, algunas regiones incluso 90%) de la bacteria *Bordetella pertusis* y en *Plasmodium falciporum* (19,3% GC, algunas regiones incluso 0%). Esto también lo postuló Galindo-Gonzalez y col. (2015) en su estudio realizado en lino como factor que influencia en la cobertura de la secuenciación. En **futuros experimentos** es una **variable** a tener en cuenta para mejorar la calidad de la secuenciación.

El análisis de correlación de Pearson (Lawrence y Link, 1989) (Resultados) entre las variables número de amplicones de la librería, número de lecturas y mega pares de bases obtenidas en la secuenciación mostró que no hay una relación lineal entre el número de amplicones incluidos en la librería y el número de lecturas y pares de bases obtenidas en la secuenciación, por lo que esta es una variable que no es necesaria optimizar, aunque nos indicó que el tamaño de los amplicones si podría

estar relacionado con el número final de lecturas, debido a la relación existente entre el número de lecturas totales y megabases secuenciadas.

5.3. Procesado de los datos

El procesado de los resultados de la secuenciación es una de las partes más complejas de este capítulo debido al gran volumen de datos obtenidos. En la actualidad existen numerosos programas para el procesado de los datos iniciales resultantes de la secuenciación (**archivo FastQ**), como se ha desarrollado en la introducción al capítulo. En este archivo se agrupan los datos de calidad de la secuenciación (codificados en ASCII o American Estándar Code for Information Interchange) además de las secuencias obtenidas, y puede procesarse mediante software de distribución libre como FASTQC (Andrews, 2010) FASTX-TOOLKIT (Gordon y Hannon, 2010) o con una "pipeline" de la plataforma Galaxy (Blackenberg y col., 2010). Este tipo de programas no servirían para el propósito de este capítulo, aunque el procesado inicial de los datos filtrando por calidad de la secuenciación disminuiría las lecturas totales perdiendo información, debido a que este tipo de programas están diseñados para el análisis de "sheared DNA" (ADN fragmentado), que incluyen un ensamblaje de las secuencias solapadas, ya que se obtienen secuencias cortas con ADN compartido para construir el genoma o una región de él, comparando con un genoma de referencia o construyéndolo de novo (Grada y Weinbrecht, 2013). En cambio, en esta tesis, los amplicones de pequeños tamaño, sólo poseen pequeñas variaciones frente a la secuencia de referencia, siendo este el principal motivo para utilizar el **programa ASPAM**, creado por E.Ritter, en el que se procesaron todas las secuencias obtenidas mediante **el filtrado de los datos brutos** basándose en el número de repeticiones de secuencias idénticas existentes, asumiendo que las secuencias idénticas de mayor frecuencia fueron las verdaderas.

El **proceso de filtrado**, denominado como "pipeline", para la determinación de las variaciones alélicas en las secuencias se realizó por GC en dos etapas. En la primera de ellas, durante la preparación de los datos, se seleccionaron aquellas secuencias que tenían un número superior a 120pb, lo que evitó la aparición de pequeños amplicones, causados por hibridación de cebadores en regiones no dianas ("mispriming") y la eliminación de los cebadores corrrespondientes a la PCR de emulsión realizada previamente a la secuenciación y en la segunda etapa, durante el proceso de filtrado, para seleccionar todas las secuencias que incluyeran los códigos identificativos del genotipo (MIDs), incluyendo 5 pares de bases adyacentes de cada secuencias adaptadoras. La presencia de estos códigos de barras en cada amplicón permitió buscar mediante el algoritmo BLAST, en el total de las lecturas, el número de copias de cada molécula de ADN original, como ya postularon Peng y col. (2015), ya que las lecturas se separan gracias a la combinación de MIDs al principio y final de cada secuencia (Clarke y col., 2014). La principal limitación, en este punto, es la capacidad de distinguir los MIDs originales de posibles MIDs "mutantes" generados durante las reacciones de PCR o por errores de secuenciación, pero la búsqueda de coincidencias perfectas mediante la comparación de las secuencias de los MIDS en ambos extremos unidos a las 5pb adyacentes que corresponden con las secuencias universales adaptadoras, y bajo el supuesto de que las secuencias con mayor número de repeticiones son las verdaderas, minimiza esta

limitación descartando aquellas secuencias no coincidentes en esas regiones. Los resultados muestran un aumento en el número de lecturas conservadas a medida que se crearon las librerías después de este filtrado, en la librería OP4 se conservaron un 71% de las lecturas totales iniciales (Figura 21), debido a las mejoras realizadas durante su creación, optimizando la técnica, hecho que también puede referenciarse en el aumento de las secuencias conservadas tras el filtro inicial (Figura 21).

A continuación, se identificaron las secuencias correspondientes a los genes candidato de cada librería comparando únicamente entre las 10pb correpondientes a cada pareja de cebadores de cada gen candidato. Esta selección evitó descartar del estudio secuencias que no fueran idénticas en su totalidad y la aplicación del filtro para incluir cada secuencia de gen candidato en su genotipo correspondiente, aseguró 3 lecturas mínimas de la secuencia en cada genotipo. Las diferencias en los resultados totales de cada librería tiene una explicación similar a la discusión sobre la preparación de los datos y el proceso de filtrado (Figura 22). Con respecto al sesgo mostrado por la desviación estándar de cada librería, está puede ser debido a la mayor afinidad de unos cebadores sobre otros con respecto a la muestra, u otras condiciones como la influencia de la temperatura de hibridación de los cebadores sobre los amplificados obtenidos. La explicación está en que cuando se amplifican regiones diferentes en una única reacción la temperaturade hibriddación debe ser 4ºC o 5ºC menor que la del cebador individual, y para la afinidad de los cebadores es que las regiones que fueron amplificadas más eficientemente pudieron influir sobre el rendimiento de la reacción en otras regiones menos eficientes a causa de que el contenido de la enzima Taq polimerasa y de los nucleótidos fue limitado y todos los cebadores compitieron por él, disminuyendo la amplificación de unas regiones frente al aumento de otras y por tanto, afectando a la eficiencia de la reacción de PCR multiplexada (Henegairu y col., 1997; Markoulatos y col., 2012).

El mayor número de lecturas en genes candidatos con amplicones de menor tamaño fue confirmado en las librerías OP4, OP5 y OP6, mediante un análisis de correlación evidenciando lo postulado por otros autores en cuanto a la preferencia de la PCR de emulsión por los amplicones de menor tamaño, aumentando así la cobertura de la secuenciación en estos amplicones (Cronn y col., 2013; Galindo-Gonzalez, 2015). Una alternativa para minimizar estos efectos, es el diseño de amplicones de tamaños similares, para ser incluidos en una misma sesión de secuenciación.

5.4. Detección de los patrones a nivel poblacional (conjunto de genotipos)

En este estudio el concepto de patrón es definido como la combinación de SNPs e Indels presentes en cada gen candidato. La extracción de estos patrones a nivel poblacional, considerados como los alelos de cada gen candidato, se realizó en tres etapas (Materiales y métodos) en las que los criterios de filtrado sobre las secuencias que restaban influyeron sobre los resultados para determinar los patrones que representaban los alelos de la población. La selección de los 4 alelos idénticos con mayor frecuencia, a pesar de la diploidía de la especie, se realizó con el objeto de aumentar el número de secuencias disponibles, asumiendo la presencia de rebundancias, que serían corregidas

posteriormente mediante el alineamiento de las secuencias y la inspección visual de cada gen candidato para determinar los patrones finales en la población. Esta detección de patrones puede compararse con las 3 etapas para la identificación de SNPs de otros programas definidas por Azam y col. (2012), conocido como "SNP calling". En primer lugar se alinean las secuencias de los genotipos con la secuencia de referencia (mapeo), a continuación, se genera una secuencia consenso para cada genotipo individual, y por último se identifican los SNPs por comparación con la secuencia de referencia. En el procedimiento realizado, en cambio, en primer lugar se extraen las secuencias de cada genotipo, a continuación se establecen los patrones en el conjunto de genotipos eliminándose los duplicados, y por último las secuencias alineadas fueron seleccionadas por el operador. Los criterios de filtrado en las dos primeras etapas influyen en el número de patrones mostrados en la última etapa, mejoran la identificación de los marcadores SNP e Indels como postulan Kumar y col. (2012). Unas condiciones poco restrictivas aumentan los patrones, y las posibilidades de falsos positivos, en cambio criterios muy exigentes disminuyen los patrones, pudiendo darse situaciones de pérdidas patrones y por ende de genes candidatos para el mapeo posterior. Los criterios seleccionados para esta tesis no fueron excesivamente restrictivos, pero como se muestran en los resultados algunos genes candidatos no continuaron en el estudio porque no mostraron patrones de consenso en la población, y los que mostraron un único patrón fueron definidos como monomórficos, y se excluyeron del estudio, al no mostrar ningún polimorfismo en las regiones seleccionadas del gen candidato. Los patrones o alelos preliminares se alinearon y se revisaron por el operador para seleccionar el conjunto final de alelos de la población, donde la principales limitaciones se encontraron en la selección de los patrones con frecuencias similares que presentaban un Indel o gap en una región homopolimérica o región dónde hay un número determinado de nucleótidos iguales o del mismo tipo, ya que uno de los principales errores de la secuenciación en la plataforma Ion torrent es la introducción de gaps, en este tipo de regiones porque no es capaz de distinguir la longitud real de los homopolímeros, que algunos estudios sugieren en homopolímeros de longitud de más de 8 bases y otros en homopolímero de 2 o 3 bases (Quail y col., 2012; Zeng y col., 2013; Bragg y col., 2013), incluso cuando el Indel se detecta en un gran número de lecturas. Además, la subjetividad del operador en la toma de decisiones sobre el patrón verdadero en estos casos también influye, por lo que en trabajos futuros es necesario validar los SNPs e Indels detectados mediante su análisis en los parentales de la población y un mayor número de descendientes convirtiendo nuestra población en una población segregante, a pesar de proceder de un conjunto de líneas no relacionadas. Esta validación en poblaciones segregantes es más informativa, ya que como postula Mammadov y col. (2010), permiten al investigador revisar la segregación de los patrones de los marcadores dilucidando si se trata de un locus con herencia mendeliana o de una secuencia duplicada o repetitiva que pasó el proceso de filtrado, dónde no pueden existir más de 4 patrones en especies diploides, a excepción de aquellos genes candidatos multi-locus.

A falta de esta **validación**, se utilizaron dos **estrategias** de control después de la determinación de los patrones que mostraron la acertada selección de los patrones en los amplicones de los genes candidatos. La primera de ellas relacionada con los genes candidatos co-localizados que procedían de

secuencias con SNP ya conocidos, los cuales mostraron que el 70,6% de estos genes candidatos co-localizados presentaron 2 o más patrones (resultados no mostrados), y la segunda la comprobación en dos genes conocidos en palmera de aceite africana que sirvieron de guía para dilucidar si los patrones que se obtuvieron eran correctos o no. Estos genes candidato fueron el gen *SH* o *Shell* (KG120) y el gen *Vir* o *Virescens* (KG271), y mostraban polimorfismos responsables de la forma y el color de la fruta, respectivamente, y dónde se seleccionaron para el diseño de los cebadores las regiones del gen que los mostraban. La secuenciación del amplicón del gen SH correspondiente al exón 1 realizada por Singh y col. 2013 mostró diferencias alélicas que produjeron cambios en los aminoácidos codificados por este gen entre *Sh Dura*, *Sh Pisifera* de origen Nigeria y *Sh Pisifera* de origen Congo. El amplicón para este gen se diseño en esa región, y pudieron observarse tres patrones diferentes (Figura 29), correspondientes a los dos alelos que deben estar presentes en nuestra población de genotipos *Tenera*, uno otorgado por el parental femenino *Dura* (Nº2) y otro por el parental masculina *Pisifera*, que en función de su origen será un patrón u otro (Nº1 y Nº3).

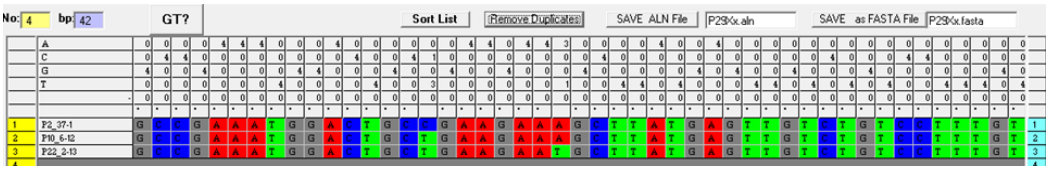


Figura 29: Patrones mostrados por el gen candidato KG120 correspondiente con el gen SH de E.guineensis. Los patrones mostrados coinciden con la secuenciación llevada a cabo por Singh y col. (2013) de los amplicones del exón 1 del gen, dónde los cambios de bases (SNP) producen cambios en las proteínas, identificando por tanto el tipo de fruto presente en la población. Nº1 corresponde a Pisifera de origen Nigeria, Nº2 a Dura y nº3 a Pisifera de origen Congo.

En el caso del gen candidato KG271, correspondiente al gen *Virescens*, Singh y col. (2014) detectaron 5 mutaciones dominantes a partir de secuencias de cDNA de la proteína R2R3-MYB que codifica para color del fruto violáceo-negruzco (*Nigrescens*)(Figura 30). Estas mutaciones truncan la proteína y originan el color de fruto verde-anaranjado (Virescens) por una inhibición en la síntesis de antocianinas, acumulando carotenoides y degradando la clorofila. De los 5 eventos obtenidos por estos autores, los resultados de la secuenciación sólo arroja el evento 1 (A/T), observado en la colección de germoplasma de origen Ghana y Nigeria, en las poblaciones de mejora y en la población de mapeo de la investigación de su investigación. En nuestra población hay 9 genotipos cuyos parentales masculinos tienen procedencia Ghana y 6 genotipos cuyos parentales son de origen Nigeria, por lo que al menos de sus alelos debe estar presente.

En ambos casos, se identificaron en los algoritmos emplados los alelos correctamente, lo que valida la estrategia empleada.

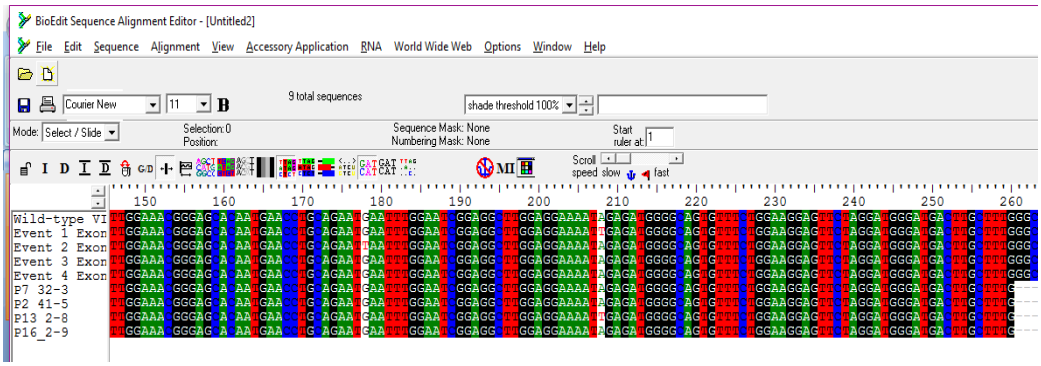


Figura 30: Alineamiento comparativo entre los patrones obtenidos en la secuenciación para el gen KG271 (Virescens) (P7 32-3; P2 41-5; P13 2-8; P16 2-9) y 4 de los patrones obtenidos por Singh y col. (2014)(Wild type; Event1; Event-2; Event-3; Event-4) en la misma región del gen. El evento 1 localizado en la base 208, es el mismo evento que el obtenido en el patrón P13 2-8 para nuestra población. Alineamiento tipo ClustalW realizado en Bioedit 7.2.2 (Hall, 1999).

Los resultados finales de la determinación de los patrones definitivos en el conjunto de la población arrojaron 19 genes candidatos sin patrón, 66 genes candidatos monomórficos, y 117 genes candidatos con 2 o más patrones, siendo siempre los genes candidatos conocidos los más abundantes de cada grupo debido al mayor número de genes conocidos incluidos en el estudio. Si revisamos los resultados mostrados en el anexo 7 y anexo 8 (Tabla 8.1), la causa de la desaparición de algunos de los genes candidatos del estudio por no generar ningún patrón puede estar en el bajo número de lecturas obtenido para estos genes candidatos (<1000 lecturas en 16 de los 19 genes candidatos), con lo que aumentan las posibilidades de que se eliminaran sus lecturas en algunas de las etapas del proceso de filtrado. En cambio, el gen candidato KG215 con más de 10000 lecturas no obtuvo ningún patrón, esto puede ser debido a errores de secuenciación en la región de los MIDS, ya que no se encontró en ningún genotipo (Anexo 7, Tabla 7.2), o bien errores durante la creación de las librerías.

Por otro lado, de los 66 genes candidatos que fueron monomórficos, 9 correspondieron con genes candidato co-localizados que fueron re-secuenciados por mostrar polimorfismos en esas regiones. La causa de que la re-secuenciación de estos amplicones no mostrará ningún cambio de base en las secuencias obtenidas puede ser debido a que el conjunto de genotipos utilizados en este estudio fue diferente a la población utilizada en el descubrimiento de esos SNPs. Puede ser un indicio de un proceso de deriva génica originado por los diferentes procesos de sufridos en nuestra población primando que esta región del gen sea homocigótica y no muestre ningún polimorfismo. Este indicio podría ser el origen de una nueva hipótesis que debería investigarse para ser confirmada.

En definitiva, de los genes candidatos que presentaron uno o más patrones un 63,93% fueron polimórficos (resultados no mostrados), es decir, que presentaron uno o más SNPs e Indels en sus secuencias. Este porcentaje de SNPs secuenciados comparado con los resultados obtenidos en otras especies de cultivos como maíz (80% de los amplicones secuenciados) (Rafalski, 2002) es inferior, a pesar de ser *E.guineensis* una especie alógama con mayor diversidad nucleotídica que las especies autógamas, como señalan Anderson y Lübberstedt (2003). Este porcentaje de polimorfismos se podría calificar como medio, destacando que los amplicones seleccionados son parte de exones de regiones

génicas dónde normalmente los SNPs e Indels presentan una menor frecuencia, y puede subestimarse el número verdadero de SNP, reduciendo la resolución de los estudios de diversidad genética, como muestra la revisión realizada por Edwards y col. (2007). Esto es debido a la limitación que se genera al seleccionar únicamente regiones de transcriptos o ricas en genes del genoma (Patel y col., 2015).

En plantas se establece una **frecuencia** de 1**SNP** cada 100-500pb (Pootakham y col., 2015), aunque los estudios realizados en cultivos como el maíz o la patata muestran una densidad muy superior. En maíz, se han estimado frecuencias de 1SNP/30-45pb (Jones y col., 2009; Kumar y col., 2014), y en patata, 1SNP/24pb (Uitdewilligen y col., 2013). Los resultados de este estudio mostraron una frecuencias de **1SNP/125pb** en la población del experimento acorde con lo postulado por Edwars y col. (2007) y similar al estudio realizado en colza (1SNP/130pb) de mapeo por asociación mediante genes candidatos relacionados con el contenido de tocoferol (Fritsche y col., 2012). En palmera de aceite, los estudios muestran diferencias en cuanto a la densidad de los polimorfismos de un solo nucleótido. Por ejemplo, Riju y col. (2007) establecieron un ratio inferior (1SNP/74pb), mientras que Low y col. (2014) y Poothakam y col. (2015) postularon densidades muy diferentes en sus estudios, 2,30SNP/100pb y 1SNP/665pb, respectivamente. Esta diferencia en las frecuencias de SNP/pb tiene su explicación en las poblaciones y los métodos utilizadas para los experimentos, ya que que en el caso de Riju y col. (2007) correspondía una mezcla de individuos no relacionados entre sí, al igual que los materiales utilizados en esta tesis, y los SNPs fueron detectados a partir de una análisis *in silico* de las bases de datos de ESTs, mientras que en el caso de Pootakham y col. (2015) procedía de una población F2 procedente de cruces de clones de *E.guineensis* y se utilizó el método de genotipado por secuenciación a lo largo del genoma ("Genotyping by sequencing"), para detectarlos.

Los SNPs fueron agrupados en función del **tipo de substitución de nucleótidos**, como se explica en la introducción a este capítulo. Las transiciones fueron superiores (230) a las transversiones (161), siendo los dos tipos de transición muy similares (A↔G; C↔T), y la transversión más prevalente fue A↔T, ambos resultados son similares a los mostrados por Low y col. (2014) en su análisis de las regiones hipometiladas del genoma *E.guineensis* Jacq y *E.oleifera.* El ratio resultante de **transición:transversión** fue de **1,43**, indicando más transiciones que transversiones. Este ratio fue similar al obtenido en otros cultivos como en vid (1,56),y en patata (1,50) (Salmaso y col., 2005; Simko y col., 2006), y también similar a los ratios obtenidos en diferentes estudios de palmera de aceite africana (1,67 (Poothakam y col., 2015) , 1,77 (Poothakam y col., 2013) y 1,55 (Ting y col., 2014).

5.5. Composición alélica de cada genotipo y análisis de la variación alélica

En el apartado de materiales y métodos se describe como fueron asociados los diferentes alelos en cada genotipo, y de las dos opciones que presentaba el programa ASPAM para realizarlos se eligió la opción "**SNP match**" en la que era determinante la región dónde se encontraba el SNP y las dos bases adyacentes a cada lado del mismo. Como se explica en el apartado es la opción menos restrictiva, pero es la más aconsejable porque incluye alelos que puedan estar presentes en genotipos con un número de

lecturas reducidas. Además, como muestra Santika (2015) (no publicado)en los resultados de su tesis de máster, hay pequeñas diferencias entre una y otra opción, pero al ser más restrictiva, pequeños errores en las secuencias completas de los amplicones, como por ejemplo, en una base hace que el genotipo que los contienen se filtren y desaparezca del estudio. Aunque lo más recomendable es utilizar ambas opciones en el análisis, esta combinación exige más recursos en cuanto a tiempo que la opción aplicada.

Por otro lado, la opción de la **asignación de la combinación de alelos a cada genotipo** acorde al nivel de ploidía de la especie (2n), permitió identificar los genotipos que poseían un exceso de alelos, y revisarlos para corregir los posibles errores. Estos errores pudieron originarse principalmente durante el proceso de asignación de patrones a la población, por lo que fue necesario revisar los patrones de esos genes candidatos. Uno de los problemas detectados en esta etapa fue la detección de genotipos con más de dos alelos , después de la revisión de sus patrones, esto puede ser debido a que esos genes candidatos muestren loci duplicados o son miembros de familias de genes altamente conservadas, que como explica Patel y col. (2015) pueden comprometer la aplicabilidad de los SNPs posicionados en los exones de los genes candidatos para futuras aplicaciones, como el mapeo por asociación, por lo que se descartaron del estudio en esta tesis doctoral.

Un ejemplo de uno de estos genes es el gen co-localizado KG162, que como se discutió en el capítulo 2 codifica una proteína GDP-manosa3,5 epimerasa que participa en la ruta biosíntetica del ácido ascórbico y está altamente conservado en plantas (78% a nivel de ADN)(Wolucka y col., 2003). Watabe y col. (2006) apuntan a qué es un único gen el que codifica esta enzima, aunque en tomate encontraron otro gen homólogo los cuales fueron mapeados en tres poblaciones de tomate por Stevens y col (2007), pero tienen motivos de consenso en su secuencia de proteínas con la familia de las enzimas epimerasa/dehidratasa, por lo que si el cebador fue diseñado en una de esas regiones puede dar lugar a la presencia de múltiples alelos, perteneciendo el amplicón a cualquiera de estos genes que codifican este tipo de enzimas, o bien que en *Elaeis guineensis* exista un gen homólogo. El gen candidato KG217 es otro ejemplo dónde han aparecido más de dos alelos. Este gen conocido codifica a una enzima fosfolípido:diacilglicerol aciltransfersa (PDAT) participante en la sintesis de ácidos grasos en la planta. Los estudios sobre este enzima recogidos por Pan y col. (2015) han confirmado que existen multiples copias de este gen en el genoma de las plantas, que los diferentes paralogos pueden codificar enzimas con diferentes hablidades para la síntesis de TAG, y que algunas de estas enzimas son sustrato selectivos. Estos mismos autores han confirmado mediante un análisis filogenético que las diferentes PDAT se agrupan en 7 subtipos diferentes soportando su conclusión en el grado de conservación y variación en la estructura del gen, las propiedades de la proteína, la recurrencia de los motivos y la divergencia funcional entre los clados. Confirma esto la aparición de múltiples alelos para este gen candidato.

Como se revisa en los resultados, algunos genotipos no mostraron alelos en algunos genes candidatos, y se consideraron como valores perdidos. Aquellos genes candidatos en los que el más del 5% de sus genotipos eran valores perdidos (26GC), fueron descartados del estudio para evitar errores en

el siguiente capítulo que se desarrollará posteriormente. En estudios de asociación en otras especies como tomate y girasol, se eliminaron aquellos marcadores que tenían 10% o más de valores perdidos en sus genotipos (Ruggieri y col., 2014; Filippi y col., 2015). El objetivo de esta mayor restricción en los valores perdidos fue evitar desequilibrios y complicaciones en los análisis posteriores a analizar, así como la creación de posibles sesgos. Balding y col. (2006) postula que la pérdida de unos pocos genotipos en análisis individuales de SNPs no generaría excesivos problemas, pero si el análisis combina múltiples SNP sí, ya que se combinarían los valores pérdidos de múltiples genotipos para varios marcadores generando un sesgo y aumentado la probabilidad de error. Una alternativa a la eliminación de estos marcadores del estudio, hubiera sido aplicar un método de imputación, como el método de estimación de máxima verosimilitud (imputaciones simples) o la selección al azar desde una distribución de probabilidad (imputaciones múltiples) reemplazando los genotipos perdidos con la predicción de sus valores basándose en el número de genotipos observados con los SNPs adyacentes, mejorando los datos observados finales (Balding y col., 2006).(Beagle_ revisar bibliografia que hace para introducirlo aki)

Los resultados de las **frecuencias genotípicas** agrupados por las combinaciones de SNPs e Indels en cada patrón, es decir, de los haplotipos presentes en cada gen candidato resultaron acordes con lo que se espera en la especie *E.guineensis*, ya que la palmera de aceite es una especie alógama en la que la polinización es cruzada, lo que lleva a una alta proporción de loci heterozigotos (Hanley y col., 2002). Este razonamiento también puede estar respaldado por el tipo de población seleccionada para este estudio, ya que es una mezcla de individuos "*Tenera*" no relacionados entre sí, y que han sufrido diferentes programas de mejora. Sus parentales masculinos ("*Pisifera*") tienen diferentes orígenes (Tabla XX; Capítulo 4) pudiendo ser los responsables de esta variabilidad existente en un 69,23% de los genes candidatos que continuaron en el estudio. Pero hay que resaltar que el análisis posterior de diversidad genética es el que dará una mayor certeza a este nivel de heterocigosidad existente en la población. Por otro lado, el resultado de estas frecuencias genotípicas confirma que la utilización de la comparación de secuencias entre los diferentes individuos de palmera de aceite africana es una metodología adecuada para la detección de marcadores SNPs debido a que el genoma de cada individuo es altamente heterocigótico, como postularon Pootakham y col. (2013), en su discusión sobre la validación de los SNPs detectados en su estudio.

5.6. Estudio básico de diversidad genética en el conjunto de la población

Los estadísticos utilizados para el cálculo de la diversidad genética mostraron valores similares a otros autores en estudios recientes de diversidad genética mediante marcadores SNP en palmera de aceite africana. Uno de estos valores es el **contenido de información polimórfica** (PIC) utilizado para medir el grado de información que aporta cada marcador genético, calculado a partir de sus frecuencias alélicas. En el presente estudio, el PIC medio en el total de los 242 genotipos de palmera de aceite africana de la población y para el conjunto de los 64 genes candidato que continuaron en el estudio fue de 0,306, con un intervalo entre 0,040 y 0,375. Este valor medio es comparable con los valores

obtenidos en palmera de aceite africana por Pootakham y col. (2013) y Ong y col. (2015) que fue de 0,293 y 0,315, respectivamente. La clasificación del contenido de información polimórfica (Hayden y col., 2010) indica en este estudio que el valor medio es razonablemente informativo (0,5>0,306>0,25), pero señala que algunos marcadores son poco informativos, como se muestra en la tabla de resultados (tabla 8; valores sin resaltar en negrita). Hay que destacar que el PIC esperado en los estudios de diversidad genética a partir de marcadores SNPs es inferior al esperado en estudios realizados a partir de SSRs debido a su condición bialélica (Kruglyak, 1997), y esto se refleja al comparar los resultados con algunos estudios de diversidad genética en palmera de aceite africana con marcadores SSR (PIC medio= 0,65, Ting y col., 2010 y Taeprayoon y col., 2015; o PIC medio= 0,546 en Arias y col., 2014) con los realizados mediante marcadores SNPs en palmera de aceite.

La **heterocigosidad observada** (Ho) media fue superior a la **heterocigosidad esperada** (He) media mostrando una gran proporción de genotipos heterocigotos en el conjunto de la población. Esta proporción de heterocigotos tan alta puede ser debida al carácter híbrido otorgado entre los parentales femeninos "*Dura*" y los parentales masculinos "*Pisifera*", ya que estos últimos tienen diferentes alelos en muchos casos. Como defienden Arias y col. (2014) en su estudio de caracterización molecular de diferentes líneas utilizadas en programas de mejora de palmera de aceite africana, esto es debido a los distintos orígenes de los parentales de la población. En nuestro estudio, los parentales de las familias de la población de "*Tenera*" utilizadas también tienen diferentes procedencias como Nigeria, Yangambi, Avros, Avros, Ekona, Ghana y LaMe para el parental masculina "*Pisifera*". El valor de He (0,398 ± 0,317), que caracteriza a la diversidad genética, fue algo superior y/o similar al obtenido en otros cultivos mediantes marcadores SNPs como en girasol (0,29; Filipi y col., 2015), en guisantes (0,297; Diapari y col., 2015) o en melocotonero (0,39; Michelenetti y col. 2015). Con respecto a los estudios más recientes en *E.guineensis* Jacq. el resultado obtenido fue igual al obtenido por Ong y col. (2015), y ligeramente superior al obtenido por Poothakam y col. en 2013 (0,372). Al igual que ocurre con el PIC, los valores He mediante marcadores SNPs son inferiores a los obtenidos mediante marcadores SSRs debido a la naturaleza multialélica de estos últimos (He= 0,449, Arias y col., 2014; He=0,70, Taeprayoon y col., 2015).

Por otro lado, el **equilibrio de Hardy- Weinberg** que postula que cuando el apareamiento es al azar las frecuencias alélicas se mantienen estables en el tiempo, generación tras generación. En este estudio esta ley se incumple en 55 (48 de ellos, P<0,001 muy significativamente) de los 64 loci estudiados, esto puede ser debido a que no se cumpla el apareamiento por azar en la población. Estas desviaciones del equilibrio pueden ser debidas a procesos de endogamia, a la estratificación de la población o a procesos de selección, y deberían aumentar la homocigosis en la población, lo que en este estudio no sucede. Además, Filippi y col. (2015) afirma que esta violación ocurre en la mayoría de las poblaciones de mejora, líneas puras incluidas. También, las poblaciones naturales se encuentran en desequilibrio para esta ley a causa de los diferentes eventos genéticos que ocurren a lo largo de la historia de una población como mutaciones, procesos de selección artificial o poblaciones

estructuradas, entre otros (Abdurakhmonov y Abdukarimov, 2008). Por ejemplo, en un estudio de asociación a lo largo del genoma en melocotonero (Micheletti y col., 2015), el 98,5% de los loci mostraban desvíos significativos de esta ley, pero seleccionaron aquellos que a pesar de desviarse significativamente mostraron un valor P> 0,0001 para el estudio de asociación. Por lo que no interfirió este desequilibrio en el análisis. Este dato será importante posteriormente ya que en estudios para conocer la estructura de la población, previos al mapeo, se asume que la población se encuentra bajo este equilibrio, como se desarrollará en el siguiente capítulo de esta tesis doctoral.

Este desequilibrio en la ley de HW es justificado por los valores negativos obtenidos en los resultados mostrados a nivel de familias con un mismo origen y a nivel del conjunto total de la población en el **coeficiente de endogamia** (Fis). El coeficiente de endogamia muestra la desviación de las frecuencias genotípicas del equilibrio HW dentro de cada población. Los valores negativos vienen dados por un exceso de heterocigóticos entre las diferentes familias y la población y por tanto una bajo nivel de endogamia. Este exceso de alelos heterocigóticos puede ser debido a un proceso de heterosis (Waples y col., 2015), ya que la población seleccionada para el estudio procede de diferentes familias que ha sufrido diferentes procesos de mejoras durante múltiples generaciones. Arias y col. (2014) en su estudio de diversidad genética en poblaciones de mejora de *E.guineensis* Jacq. también obtuvo valores negativos en este estadístico a causa de un alto número de heterocigóticos en su conjunto poblacional de diferentes familias con diferentes orígenes, de las que coinciden 3 parentales masculinos con el mismo origen a los incluidos en esta tesis (Ekona, Yangambi y Avros). En cambio, Taeprayoon y col. (2015) obtuvo valores positivos, próximos a 0, entre la media de las poblaciones estudiadas y a nivel multipoblacional (0,20 y 0,22), pero sus poblaciones de estudio fueron diferentes. Este exceso de heterocigotos en la población es también visible por los resultados también negativos y próximos a -1 del **estadístico Fit** que estiman un exceso de heterocigóticos a nivel global entre las familias y en la población total, por lo que la deriva genética no parece existir en nuestra población, evitando la diferenciación genética de la población.

El **estadístico Fst** (índice de fijación) que mide la diferenciación entre las poblaciones por efecto de la deriva génica, mostró valores muy próximos a cero que indican una baja diferenciación entre las familias con parental masculino común y, también, entre el conjunto total de la población. En cambio, en estudios más amplio de diversidad genética este valor fue superior a 0,1, sugiriendo una diferenciación entre las diferentes familias de estudio y entre la población total (0,2492, Singh y col., 2008; 0,174, Arias y col. 2014; 0,33, Taeprayoon y col., 2015; 0,177, Okoye y col., 2016). Es relevante destacar que estos estudios fueron realizados mediante marcadores microsatélites que poseen mayor capacidad informativa como se ha discutido anteriormente, y aunque el origen de algunos de los genotipos pueden ser comunes, las familias estudiadas en la presente tesis son diferentes ya que pertenecen a diferentes programas y se han desarrollado en regiones diferentes, por lo que puede existir una clara diferenciación de los resultados. Este estadístico está profundamente relacionado con el **flujo génico** (Nm), que en los resultados muestra una hipotética no diferenciación entre las diferentes

familias y a nivel completo de la población. Por tanto, existen indicios que la población de estudio evolucionan como un conjunto y no existe una población estructurada. En cambio, Hayati y col., (2004), Bakoume (2006) y Taeprayoon (2015) supusieron en cambio una estructura genética entre sus poblaciones de estudio, debido a que este parámetro junto con los resultados del índice de fijación (Fst) evidenciaban poblaciones estructuradas que divergían genéticamente entre ellas.

Los dos dendrogramas generados, uno a partir de las familias agrupadas por el origen del parental (Figura 28a), y el otro a partir del agrupamiento de la población por su parental (Figura 28b) mostraron 3 agrupamientos diferentes. Los resultados mostraron que los dos primeros agrupamientos del dendrograma a, correspondían a dos familias cuyos parentales son de origen Nigeria y Ghana, la causa de la separación de estas familias, y su mayor distancia genética venga dada por su proceso de mejora en la que se han utilizado cruces controlados. En cambio, para el tercer agrupamiento las familias con diferentes parentales masculinos se encuentran mezclados, lo que implica una gran similitud entre las diferentes familias y origenes. Cuando se analizan el conjunto de la población por origen de su parental masculino, sin tener en cuenta la familia a la que pertenecen también mostraron 3 agrupamientos bien diferenciados, pero se observa un relación más estrecha entre los parentales de diferentes orígenes. El más destacado es la estrecha relación entre Avros y Dami, cuya procedencia original es la República Democrática del Congo (África) (Corley y Tinker 2003), pero en cambio Yangambi cuya procedencia es la misma ha evolucionado obteniendo una mayor similitud con Nigeria, lo que puede ser debido a los procesos de mejora a los que estas plantas han sido sometidas. Es importante destacar que no se estudio el mismo número de familias para cada agrupamiento, por lo estos datos son meramente indicativos y pueden presentar sesgos, ya que por ejemplo en el caso del parental LaMe se incluyeron 5 genotipo que lo contenían pertenecientes a 3 familias diferentes, y en cambio, fueron 52 los genotipos de estudio en 25 familias. Para un estudio de mayor exactitud sería necesario aumentar el número de familias con parentales del mismo origen con los mismos marcadores.

Todos estos datos fueron confirmados con los resultados obtenidos por el **análisis molecular de la varianza** (AMOVA) dónde se mostró que el 99% de la diferenciación genética es debido a la diferenciación dentro de los individuos del conjunto de la población, y sólo 1% se atribuye a la variación entre cada población. Si se relaciona los datos obtenidos mediante el estadístico Fst que mide la diferenciación entre las poblaciones puede decirse que el total de la variación genética del conjunto de la población total entre el 1 y 7% es debida a la diferenciación entre las diferentes poblaciones que la conforman y entre el 93 y 99% es debida a los individuos. Los resultados obtenidos en palmera de aceite africana por otros autores también muestra que más del 60% de la diferenciación genética existente en la especie es debida a los individuos dentro de la población (Teaprayoon y col. 2015; Ong y col., 2015; Arias y col, 2014; Cochard y col., 2009). Como postula Ong y col. (2015) estos niveles de variación genética en plantas está asociado directamente con el sistema de mejora (Olmstead, 1999), y las especies perennes con un ciclo de vida largo y de polinización cruzada presentan una mayor variabilidad genética entre los individuos dentro de una misma población.

Para finalizar esta discusión cabe destacar que este estudio de diversidad genética en la población utilizada en esta tesis parece indicar que la población no esta estructurada, por lo que puede ser adecuada para el estudio de mapeo por asociación a través de genes candidatos que se desarrolla en el siguiente capítulo. Aunque esta hipótesis debe ser confirmada con un estudio de la estructura de la población.

CAPÍTULO 4: ASOCIACIÓN GENOTIPO-FENOTIPO

1. INTRODUCCIÓN

El mapeo por asociación es un análisis estadístico de asociación entre los genotipos, normalmente sus SNPs o sus haplotipos, determinados en un conjunto de individuos, y los caracteres (fenotipos) de los mismos individuos (Rafalski, 2010). Como ya se ha explicado en el apartado 4.2 de la introducción de esta tesis doctoral, este análisis puede abordarse por dos métodos diferentes: 1. mediante la búsqueda de marcadores distribuidos homogéneamente a lo largo del genoma, o 2. mediante la búsqueda de marcadores en los genes candidatos que pudieran estar relacionados con el o los caracteres que quieran estudiarse. En la primera metodología es necesario conocimiento previo del genoma de la especie, ya que las regiones polimórficas se buscan en toda su longitud, para luego buscar la causalidad con el carácter, y en la segunda metodología se busca la correlación de los polimorfismo detectados en los genes candidatos con el carácter o los caracteres de estudio, adaptándose mejor al estudio en especies donde la información genómica es pobre (Fusari, 2010) como ocurría en *E.guineensis* Jacq en el momento de iniciar esta investigación.

La población seleccionada es uno de los factores críticos en los estudios de mapeo por asociación (Flint-García y col., 2003; Breseghello y Sorrells, 2006; Yu y Buckler.,2006), ya que a comparación con el mapeo de ligamiento y de QTLs dónde se utilizan poblaciones biparentales con sus respectivas progenies, en los estudios de MA la población procede de diferentes individuos que pueden formar parte de un banco de germoplasma o de una población natural, y según muestran Abdurakhmonov y Abdukarimov (2008) aportan una mayor variabilidad genética porque permite el análisis simultáneo de gran número de marcadores simultáneamente, y en consecuencia hay una mayor probabilidad de obtener mayor resolución en el mapeo debido a la utilización de mayores eventos recombinatorios durante la evolución del germoplasma, entre otros. Por otro lado, la diversidad genética, la extensión del desequilibrio de ligamiento a lo largo del genoma y las relaciones dentro de la población determinaran la resolución del mapeo, la densidad de marcadores, los métodos estadísticos y la fuerza del mapeo (Zhu y col., 2008). Las poblaciones más recomendadas (Breseghello y Sorrells, 2006) serían: 1. Las colecciones procedentes de un banco de germoplasma o "core collection" representan la diversidad genética porque poseen un amplia diversidad alélica, y son útiles en estudios dónde se buscan caracteres cualitativos como resistencia a enfermedades, o características de calidad. En cambio, para la búsqueda de asociaciones con caracteres cuantitativos no son las más recomendadas ya que la mayoría de sus individuos no están adaptados a las condiciones de crecimiento y a la prevalencia de enfermedades obteniendo una baja precisión en las medidas de los caracteres. Los marcadores obtenidos pueden utilizarse para la introgresión de nuevas variaciones en líneas élite mediante retrocruzamientos asistidos por marcadores (Frish y Melchinger, 2005). 2. La población formada por líneas élite o cultivares es adecuada cuando los caracteres de estudio poseen una baja heredabilidad, como pueden ser el rendimiento, o la tolerancia a estrés abiótico, ya que estas líneas son genéticamente estables y están adaptadas a las condiciones normales de crecimiento. En este caso, el mapeo por asociación produce mayor número de polimorfismos y la detección de alelos favorables en la población

objetivo, que necesitarán ser validados posteriormente mediante cruzamientos para ser utilizados como marcadores de las progenies en futuros cruces. 3. Por último, las poblaciones sintéticas estarán formadas por un conjunto de individuos con una buena aproximación a la asunción de apareamientos al azar, ya que están diseñadas para mantener y minimizar la endocria, aunque cuando esta población se somete continuamente a procesos de selección será necesario genotipificar los individuos de generaciones sucesivas para reflejar adecuadamente la constitución genética de la población (Fusari, 2010). Los marcadores obtenidos a partir de este tipo de población son útiles para el desarrollo de índices de selección de los individuos (Breseghello y Sorrells, 2006).

La especie y el tipo de población seleccionada condiciona el resto del estudio, ya que el desequilibrio de ligamiento, base conceptual del mapeo por asocciación, difiere entre y dentro de las especies, e incluso entre diferentes regiones de un único genoma. Además, el tipo de población seleccionada puede presentar estructura poblacional, siendo ésta la primera causa de asociaciones espurias (Tabla 1).

Tabla 1: Tipos de población y sus características más relevantes para el mapeo por asociación. Adaptada de Breseghello y Sorrels, 2006. DL= Desequilibrio de ligamiento.

Tipo de Población	Grado de DL	Estructura Poblacional
Germoplasma o colecciones núcleo	Bajo	Medio
Líneas élite y cultivares	Alto	Alto
Población sintética (individuos o progenies)	Intermedio	Bajo

El desequilibrio de ligamiento puede definirse como el grado de asociación no aleatoria entre los alelos de diferentes loci (Zhu y col., 2008), tal y como se ha explicado en el apartado 4.1 de la introducción de esta tesis doctoral. Un fuerte ligamiento entre los "*loci*" es la principal causa de desequilibrio de ligamiento, por tanto únicamente aquellos polimorfismos que estén estrechamente ligados a un locus con efectos fenotípicos tendrán una probabilidad significativa de asociación con el carácter de interés en una población cuyos apareamientos hayan sido al azar (Remington y col., 2001). La explicación está en que cuando existen dos SNPs muy próximos, con pocas bases de diferencia, éstos están sometidos a la misma presión por selección y deriva genética en el tiempo. La recombinación entre dos bases muy próximas es poco frecuente, por lo que están altamente ligados, traduciéndose en un alto desequilibrio de ligamiento. En cambio, cuando los SNPs están situados en cromosomas diferentes están sometidos a la selección y a una segregación independiente, por lo que la recombinación entre ambos es más probable, y su DL será menor (Flint-García y col., 2003).

Los estadísticos para medir el DL se basan en las diferencias entre las frecuencias de los haplotipos observados y esperados (Capítulo 1; Apartado 4.1). Estos estadísticos son numerosos y se adaptan a las necesidades del estudio en función de la ploidía de la especie o de la presencia de alelos multilocus (Gupta y col.,2005). Para "*loci*" bialélicos, como es el caso de los genes candidatos

seleccionados para este estudio, y especies diploides como *E.guineensis* Jacq. los estadísticos más utilizados son **r^2** y **D'** (Figura 1)(Gupta y col., 2005; Ersoz y col., 2007; Orazugie y col., 2007). La principal diferencia entre estos estadísticos es que **r^2** resume los eventos recombinatorios y la historia de las mutaciones, mientras que **D'** mide únicamente las diferencias debidas a la recombinación (Figura 2), teniendo en cuenta estas diferencias los autores señalan a **r^2** como el mejor estadístico para utilizar en estudios de mapeo por asociación, actuando como un indicador de las correlaciones entre marcador-carácter (Abdallah y col., 2003; Gupta y col.,2005; Oraguzie y col.,2007; Abdurakhmonov y Abdurakimov, 2008).

Otro factor a tener en cuenta es la extensión del DL porque determina la distancia física en la que existe asociación a carácter mediante ese marcador, determinando la densidad de marcadores necesarios para el mapeo y su posible abordaje (Ingvarsson y Street, 2011). Esta extensión puede visualizarse bien en todo el genoma, en cada cromosoma o en un único gen, ya que como se explicó en la introducción el DL no es uniforme en todas las regiones del genoma dentro de una misma población, y por ende en la especie. Aunque inicialmente el estudio del DL se fue aplicado en humanos, en la actualidad acompañado por el desarrollo de la era genómica estos estudios se han incrementado en plantas como *Arabidopsis* y en diferentes cultivos como arroz, caña de azúcar, cebada, maíz, trigo o soja, entre otros, como muestra la revisión realizada por Maskri y col. (2012).

a)

$$r^2 = \frac{(\pi_{AB}\,\pi_{ab} - \pi_{Ab}\pi_{aB})^2}{\pi_A \pi_B \pi_a \pi_b}$$

b)

$$D' = \begin{cases} \dfrac{(\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB})}{min(\pi_A\pi_b, \pi_a\pi_B)} \; si\; \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB} > 0 \\ \dfrac{(\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB})}{min(\pi_A\pi_B, \pi_a\pi_b)} \; si\; \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB} < 0 \end{cases}$$

Figura 1: Estadísticos más utilizados para el estudio de DL. a) El estadístico r^2 es una medida del grado de correlación. Valores altos de r^2 indican que ambos SNP transmiten información similar, y sólo es necesario genotipar una de los dos SNPs (Bush y Moore,2012); b) D' es una medida utilizada en genética de poblaciones relacionada con los eventos recombinatorios entre los marcadores. El valor de D' se encuentra entre 0 y 1. D'=0 indica equilibrio total de ligamiento, por tanto una mayor frecuencia de recombinación entre los dos marcadores. En cambio, D'=1 indica un desequilibrio de ligamiento completo, y por tanto la frecuencia de recombinación será nula(Bush y Moore, 2012). π_{AB}=frecuencia del haplotipo AB; π_{ab}= frecuencia del haplotipo ab; π_{Ab}= frecuencia del haplotipo Ab; π_{aB}= frecuencia del haplotipo aB; π_A=frecuencia del alelo A; π_B= frecuencia del alelo B; π_a=frecuencia del alelo a; π_b= frecuencia del alelo b.

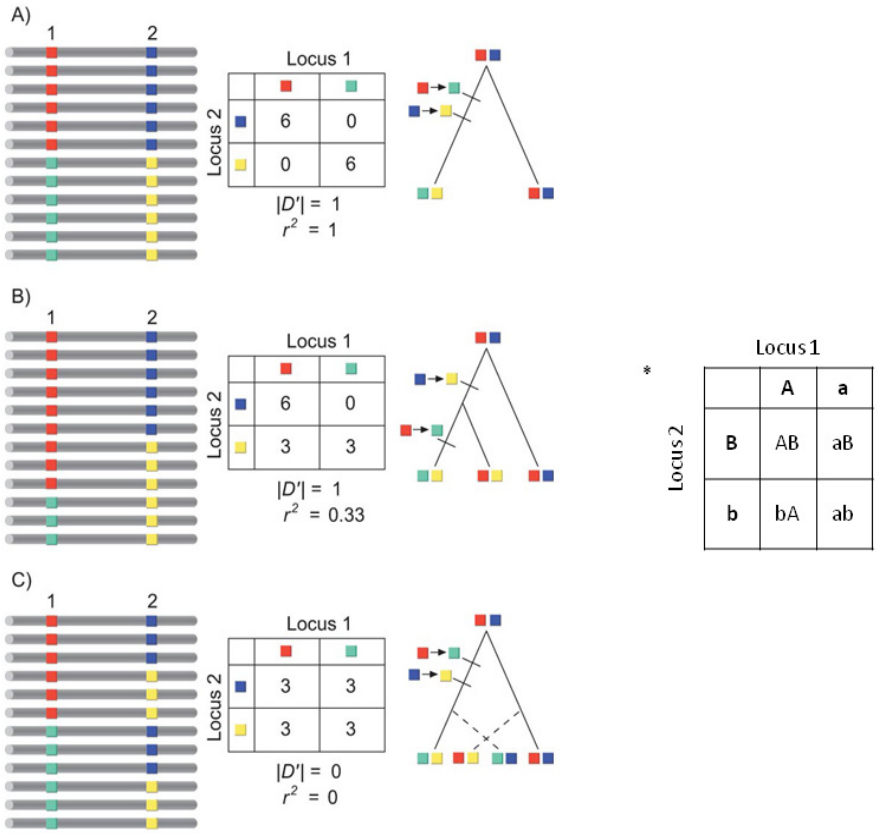


Figura 2: Demostración de los estadísticos r2 y D' en diferentes situaciones de DL entre polimorfismos ligados causados por mutaciones y eventos recombinatorios. Las imágenes de la izquierda representan los alelos de dos loci. En el medio, las tablas de contingencia de los haplotipos* y los resultados de ambos estadísticos. Las imágenes de la derecha muestran los árboles para observar el DL. **A**. DL completo (1) cuando los dos loci comparten una historia de mutaciones similares sin recombinación. **B**. DL resultante cuando existen mutaciones en diferentes linajes sin recombinación entre loci D'>r2. **C**. Estado de equilibrio completo debido a eventos recombinatorios entre los loci, con independecia de sus mutaciones (r2 y D'=0). Imagen de Flint-Garcia y col., 2003.

Los cultivos que han sufrido largos procesos de domesticación y se han sometido a procesos de selección continuados en un ambiente específico y para un carácter determinado modifican la arquitectura de su genoma reduciendo su diversidad genética y creando poblaciones estructuradas. Se entiende por estructura poblacional a la división de la población en diferentes subgrupos con una relación de parentesco (Rafalsky, 2010), como consecuencia se crea una distribución desigual de los alelos entre estos subgrupos que al muestrearlos para crear un población de mapeo la mezcla de estos individuos con diferentes frecuencias alélicas crea DL (Soto-Cerdá y col. 2012). Esta estructuración de la población es una de las fuentes más importantes de error tipo I en los estudios de asociación, ya que originan falsas asociaciones debidas al significativo DL que generan entre los polimorfismos situados en diferentes cromosomas (Flint-García y col., 2003).

La estimación de la estructura de la población en los estudios de mapeo por asociación se aborda mediante metodologías estadísticas basadas en el uso de marcadores no ligados que detectan la estratificación de la población y la corrigen (Pritchard y Rosenberg, 1999). Estas metodologías se aplican mediante el uso de marcadores SSR y/o SNP, debido a su codominancia y a la posibilidad de automatizar el proceso de genotipado. Estos marcadores también permiten estimar los niveles de parentesco, otro

factor influyente sobre el DL (Capítulo1; Apartado 4.1). La principal diferencia en la elección de los marcadores radica en la necesidad de un mayor número de marcadores tipo SNPs por su naturaleza bialélica para llevar a cabo el estudio (Zhu y col., 2008; Fillipi y col., 2015). El método estadístico más utilizado en la estimación de la estructura poblacional es el programa **STRUCTURE** desarrollado por Pritchard y col. (2000), y basado en la inferencia bayesiana. Mediante este programa se busca atribuir el genoma de cada individuo del estudio a poblaciones hipotéticas y el análisis de asociación estará condicionado a estas sub-poblaciones mediante la matriz de estructura poblacional denominada Q. Este método asume que los marcadores no están ligados y la panximia de la población, es decir, el apareamiento aleatorio. El **análisis de componentes principales (ACP)** (Price y col., 2006) es otro de los métodos utilizados para detectar e inferir la estructura poblacional en el mapeo por asociación. Este análisis resume la variación observada en los marcadores a lo largo de la población en un número menor de variables componentes. Estos componentes principales pueden interpretarse como sub-poblaciones no observadas originadas a partir de los individuos de origen (Zhu y col. 2008). La elección de uno de estos métodos dependerá del modelo estadístico elegido para buscar la asociación genotipo-fenotipo (Capítulo 1; Apartado 4.3) que se aplica en el mapeo por asociación.

El **modelo lineal generalizado** conocido por sus siglas en ingles "General Linear Model" (GLM) es uno de los modelos más sencillos utilizado en plantas para el mapeo por asociación, que para cada combinación de marcador-carácter la aplicación de este análisis proporciona una solución de mínimos cuadrados (Searle, 1987) mediante múltiples comparaciones entre las medias de los niveles de los factores para hallar diferencias significativas. Otros análisis estadísticos sencillos que pueden utilizarse son la regresión linear, el análisis de la varianza (ANOVA), el test t o el test chi-cuadrado (Zhu y col., 2008). El principal inconveniente de estos métodos es que no tienen en cuenta la presencia de estructura poblacional y/o las relaciones de parentesco. En el caso de GLM, en presencia de estructura poblacional pueden darse dos opciones: 1. excluir las muestras con similitud a una población no objetivo, o 2. incluirlo en el estudio de asociación como una co-variable (Bush y Moore, 2012), lo que se asemeja al método desarrollado por Pritchard y col. (2000) para poblaciones estructuradas en el que se estima la estructura poblacional (Matriz Q) y esta estimación se incorpora al GLM para corregir falsas asociaciones (Sotó-Cerda y Cloutier, 2012). Otro método muy aplicado es el **modelo lineal mixto** (MLM) (Yu y col., 2006) en el que se incorporan la estructura poblacional (Matriz Q) y las relaciones de parentesco (Matriz K) (Soto-Cerdá y Cloutier, 2012). En caso de aplicar un análisis de componentes principales para estimar la estructura poblacional, es la matriz P la que se incorpora en vez de la matriz Q, para ambos modelos estadísticos. A estos análisis pueden aplicarse test de correcciones múltiples como la corrección de Bonferroni (Dunn, 1961) - muy restrictivo-, la tasa de descubrimiento de falsos positivos o FDR (Benjamini y Hochberg, 1995) - ampliamente utilizada-, el test de permutaciones o el concepto de significancia de GWA (Dudbrigde y Gusnanto, 2008), este último sólo en el abordaje del estudio de asociación a nivel de genoma completo, con el objetivo de controlar la acumulación de falsos positivos (Balding 2006; Bush y Moore, 2012). Estas correcciones deben realizarse con cierta precaución

porque pueden originar o aumentar los errores de tipo II (falsos negativos) por aceptación de la hipótesis nula, y por tanto disminuye el poder detección de asociaciones (Fusari, 2010).

Los programas existentes para llevar a cabo los análisis estadísticos para el mapeo por asociación son numerosos. Desde Proc GLM y Proc Mixed en SAS (SAS Institute, 1999) o R scripts (Ihaka y Gentleman, 1996) en los que son necesarios conocimientos de programación, a paquetes de programas como TASSEL (Bradbury y col., 2007) que permite también calcular y mostrar visualmente el desequilibrio de ligamiento, en el que también pueden incorporarse los datos de la estructura poblacional bien a partir de la matriz de datos Q, o bien a partir de la matriz P, y las relaciones de parentesco K, aplicables a los modelos lineales mixtos, siendo este el más ampliamente seleccionado para llevar a cabo este tipo de análisis. En el último año por ejemplo Tello y col. (2015) han publicado un estudio realizado en vid de mapeo por asociación mediante genes candidatos relacionados con el carácter complejo densidad de racimo en el que la asociación fue abordada por 4 modelos estadísticos diferentes en **TASSEL**: GLM, GLM+Q, MLM+K, y MLM+Q+K obteniendo como resultado un conjunto de SNP asociados a dos componentes complejos de este carácter, y posicionados en genes candidatos que previamente no habían sido relacionados con el carácter en cuestión. Mariette y col. (2015) publicaron la identificación de genes relacionados con la resistencia al virus *Plum Pox* en albaricoque mediante la asociación por el escaneo del genoma completo utilizando con método estadístico MLM+Q+K mediante un paquete estadístico aplicado en R.

Desde los primeros estudios de DL en plantas realizados en cultivos como maíz, arroz y avena mediante marcadores AFLP, RFLP e isoenzimas (Soto- Cerdá y col., 2012) en la década de los 90, el número de publicaciones también en otros cultivos como trigo, cebada, o soja, ha ido en aumento como lo lo muestran las revisiones realizadas por Gupta y col. (2005), Abdurakhmonov y Abdukarimov (2008), Zhu y col. (2008), o Soto-Cerdá y col. (2012). En especies forestales y cultivos frutales con largo ciclo de vida como eucalipto, pino, vid, o manzana también se han desarrollado numerosos estudios de asociación bien mediante el abordaje de genes candidato bien por el escaneo del genoma completo (Khan y Korban, 2012). En palmera de aceite africana el único estudio publicado de mapeo por asociación ha sido abordado mediante la estrategia del escaneo completo del genoma (GWAS) para el carácter complejo contenido de aceite en el mesocarpo (O/DM) identificando polimorfismos asociados con diferencias para este carácter (Teh y col., 2016).

En conclusión, el tipo de población presente en el estudio, el DL, la estructura de la población, los caracteres de estudio y el conocimiento y recursos genómicos disponibles en la especie y el objetivo del estudio son los que influirán sobre la decisión de la estrategia mediante la que se aborde el mapeo por asociación y el modelo estadístico a aplicar. Normalmente cuando el DL es bajo el abordaje mediante el genoma completo se hace impracticable por el gran número de marcadores que se necesitan para llevarlo a cabo, siendo más útil la identificación de polimorfismos a nivel de genes candidatos relacionados con los caracteres de interés.

2. OBJETIVO

El objetivo de este capítulo por tanto es la realización de un estudio de mapeo por asociación para los principales caracteres agronómicos relacionados con los componentes de producción, de racimo y componentes vegetativos, todos ellos relacionados con el rendimiento del cultivo *E.guineensis* Jacq. utilizando los SNPs detectados en algunos de los genes candidatos identificados como partícipes de estos caracteres en el capítulo 2 de la presente tesis doctoral.

3. MATERIALES Y MÉTODOS

3.1. Fenotipado

Los datos fenotípicos para los caracteres de estudio (**BN**= Número de medio de racimos/palmera/año; **BW**=Peso medio del racimo/año (Kg); **CPO**=Rendimiento de aceite/hectarea/año (ton/ha/año); **FN**=Número medio de frutos por racimo; **FW**=Peso medio del fruto (g); **HT**= Incremento de altura de tallo/año (cm); **MF**=Ratio de pulpa o mesocarpio húmedo con respecto al fruto (%); **OWM**= Ratio de aceite con respecto al peso húmedo de pulpa o mesocarpio(%) se tomaron durante 10 años (2000-2010), en la estación experimental de PT Binasawit Makmur, en Surya Adi Estate (latitud: 105°2'0"- 105°4'0" E, longitud: 04°1'0" – 04°2'0"S, elevación: 28m) y perteneciente a la provincia de Sur Sumatra (Indonesia). Los resultados de esta caracterización fenotípica fueron mostrados como el valor medio de cada carácter en cada palmera (Anexo 10; Tabla 10.1).

Para cada caracter de la población se calculo el valor medio, el máximo y el mínimo, y su desviación estándar. Además se realizó una análisis de correlación de Pearson para conocer las posibles relaciones existentes entre los caracteres. Estos análisis estadísticos se aplicaron mediante el programa SAS v 8 (SAS Institute, 1999) utilizando el algoritmo PROC CORR.

3.2. Genotipado

Los datos de genotípicos correspondían a las 238 palmeras utilizadas para el genotipado mediante marcadores SNPs utilizadas en el capítulo 3 (Apartado 3.2.1). En total se seleccionaron 142 SNPs localizados en 61 genes candidatos de los 198 seleccionados inicialmente para el estudio y procedentes del análisis de colocalización, del transcriptoma y conocidos que pudieran participar en los caracteres elegidos en el capítulo 2 de esta tesis doctoral.

3.3. Desequilibrio de ligamiento

En primer lugar se seleccionaron los marcadores SNPs bialélicos genotipados en el capítulo 3 con un 20% máximo de valores pérdidos. Los valores perdidos de aquellos marcadores que los presentaron se imputaron por un método desarrollado por E.Ritter (comunicación personal) en R (R Core Team, 2013), basado en valores medios de clases de marcadores. A continuación, todos los marcadores se mapearon en el mapa "in silico" utilizado en el capítulo 2 de esta tesis doctoral con el objetivo de conocer su posición y la distancia entre los marcadores.

Los valores $\mathbf{r^2}$ para el desequilibrio de ligamiento y sus valores p correspondientes para todos los pares de loci entre los marcadores bialélicos se calcularon utilizando los programas **SNPanalyzer** (Yoo y col., 2008) y **Haploview** (Barret y col., 2005), pero aquellos SNPs con una frecuencia inferior al 5% fueron excluidos del análisis. Este cálculo se realizo a nivel intracromosómico e interloci, entre los amplicones de los genes candidatos que se agrupaban en un mismo cromosoma,y a nivel intra gen candidato en aquellos genes candidatos que presentaron más de un SNP en su amplicón.

3.4. Analisis de componentes principales (ACP) y estructura poblacional

El análisis de componentes principale (ACP) se realizó para analizar la presencia de estructura genética en la población de estudio mediante el programa **TASSEL 5.0** (Bradbury y col, 2007; Disponible en http://www.maizegenetics.net) utilizando el conjunto de SNPs previamente detectados en la población, Los parámetros utilizados para el ACP fueron los siguientes: covarianza para obtener los eigenvalores procedentes de la descomposición de la matriz de covarianza, número máximo de componentes 3, y mostrando sus eigenvalores (Disponible en https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual). Estos eigenvalores fueron revisados para determinar el número de ejes independientes de diferenciación genética en los genotipos, y con el objeto de poder incorporar la matriz P al modelo estadístico posterior para la búsqueda de asociación

La estructura poblacional se determinó mediante un modelo de agrupamiento bayesiano con el programa **STRUCTURE v 2.3** (Pritchard y col., 2000), sobre la base de los 29 SNP detectados en los 8 genes candidato que cumplían el supuesto de equilibrio HW (Capítulo 3, Tabla 8). El número de grupos analizados osciló entre 1 y 10, con 20 repeticiones para predefinir el número de grupos presentes en la población. El período de "burn-in length" se estableció en 10000, seguido de 50000 repeticiones de la cadena de Markov. La definición de los grupos no fue predefinida, y las frecuencias alélicas fueron correlacionadas entre los grupos (Pritchard y col., 2009). El número óptimo de grupos se determinó como K=2, mediante el programa **STRUCTURE HARVESTER** (Earl y col., 2012) que aplica el método de Evanno y col., (2005) basado en la estadística ad hoc para estimar el número de grupos presentes. Con el objetivo de confirmar este resultado, se ejecuto de nuevo el programa con 10 repeticiones de cada k entre 1 y 5, con el período de "burn-in length de 50000 y 100000 repeticiones de la cadena de Markov.

3.5. Análisis de asociación de marcadores con el fenotipo

El ánalisis de asociación de los diferentes marcadores (Total 129 macadores SNP) con los 8 caracteres de estudio (BN, BW, CPO, FN, FW, HI, OWM, MF) se testó en tres modelos diferentes de asociación basados en el modelo linear generalizado (MLG): 1. MLG simple, 2. MLG+Q que incluye como covarianza la estructura poblacional obtenida en STRUCTURE, y 3. MLG+P que incluye la matriz P del análisis de componentes principales obtenido en TASSEL. Estos análisis se realizaron en **TASSEL v. 5.0** (Bradbury y col., 2007). La elección del modelo se realizó mediante una gráfica de cuartiles del conjunto de valores p obtenidos del análisis de asociación para cada marcador en el conjunto de los fenotipos y en cada modelo realizada en EXCEL 2007.

Finalmente, una vez seleccionado el modelo se realizó en SAS V 8.0 el análisis de asociación mediante PROC GLM y aplicando posteriormente el test de DUNCAN para la comparación de las medias observadas con un valor de error de probabilidad p <0,05.

4. RESULTADOS

4.1. Fenotipado

La población fue fenotipada para siete caracteres relacionados con la producción y el rendimiento para los cuáles se calcularon los valores medios en cada genotipo (Anexo 10, Tabla 10.1). Además se calcularon los valores medios, máximo y mínimo para cada carácter como también el coeficiente de variación (Tabla 2). La variación estimada por el coeficiente de variación varió entre un 8,10% (MF) y 41,92% (FN).

Tabla 2: Variación fenotípica en el conjunto de la población para los caracteres agronómicos analizados. BN= Número de medio de racimos/palmera/año; BW=Peso medio del racimo/año (Kg); CPO=Rendimiento de aceite/hectarea/año (ton/ha/año); FN=Número medio de frutos por racimo; FW=Peso medio del fruto (g); HT=Incremento de altura de tallo/año (cm); MF=Ratio de pulpa o mesocarpio humedo con respecto al fruto (%); OWM= Ratio de aceite con respecto al peso húmedo de pulpa o mesocarpio (%).

	Media	Máximo	Mínimo	%CV
BN	10,46	18,30	4,80	23,35
BW	15,04	23,80	10,30	14,43
CPO	3,97	9,30	1,00	38,04
FN	1039	2619	221	41,92
FW	10,33	30,70	3,80	41,34
HT	62,55	105,10	31,00	23,82
MF	79,02	93,00	63,30	8,10
OWM	49,53	69,70	25,00	19,16

Los coeficientes de correlación de Pearson (Tabla 3) fueron estadísticamente muy significativos (p valor < 0.01) en 12 de las 28 parejas, y tan sólo 1 significativa (valor p< 0.05). En cuanto a los valores de los coeficientes es de destacar una correlación r >0.5 para las parejas BN y CPO, CPO y OWM y, por último, FW y FN. Por otro lado, algunas de las parejas de marcadores se correlacionaron con signo negativo como BN y BW, FN y FW, y FN y MF, indicando el sentido inverso de su relación.

Tabla 3: Coeficientes de correlación de Pearson entre todos los caracteres agronómicos analizados. Se muestran además los p valores correspondientes a cada coeficiente. CAR= Carácter.

CAR.	BN	BW (Kg)	CPO (ton/ha/año)	FN	FW (g)	HI (cm)	MF (%)
BW (Kg)	-0.24**	-					
	0.0002						
CPO (ton/ha/año)	0.63**	0.11	-				
	<.0001	0.0893					
FN	-0.07	0.31**	0.12	-			
	0.273	<.0001	0.0598				
FW (g)	-0.06	0.04	-0.02	-0.72**	-		
	0.3552	0.4964	0.783	<.0001			
HI (cm)	0.19**	0.25**	0.27**	0.01	0.11	-	
	0.0033	0.0001	<.0001	0.9256	0.0941		
MF (%)	0.05	-0.05	0.31**	-0.20**	0.20**	0.08	-
	0.4792	0.4839	<.0001	0.0021	0.0016	0.1979	
OWM (%)	0.16*	0.02	0.66**	0.02	-0.03	0.10	0.17**
	0.0147	0.7724	<.0001	0.7333	0.6028	0.1414	0.007

4.2. Posicionamiento en el mapa y desequilibrio de ligamiento

En total fueron 142 los SNPs pertenecientes a 65 genes candidatos bialélicos los que continuaron en el estudio de asociación. Estos genes candidatos se localizaron en el mapa a lo largo de los 16 cromosomas de la especie *E.guineensis* Jacq., siendo el cromosoma 13 el que presentaba el mayor número de SNPs (27 SNP) contenidos en 3 genes candidatos (KG195, KG70 y P74), y el cromosoma 16 el que menos SNPs tenía (2 SNPs) pertenecientes a un único gen candidato (KG106) (Anexo 11, Tabla 11.1), en la tabla inferior (Tabla 4) se muestran los genes candidatos y sus SNPs mapeados en el cromosoma 2.

Tabla 4: Posicionamiento en el mapa de los SNPs pertenecientes a los genes candidatos bialélicos posicionados en el cromosoma 2 . GC= Gen candidato; SNP= Variación encontrada; POS_MAP= posición en el mapa en pares de bases (pb).

	GC	SNP	POS_MAP
GL-2	**GLO2**	A/G	6792342
		C/T	6792272
	KG288_EgTPase	G/T	8992884
	KG154_DDB1CUL4	C/T	9118625
	KG180_FFB2_C4663_S1.2	A/G	13183483
	KG196_BnC2_10C3-629	C/T	28517614
		T/G	28517750
		A/C	28517852

	KG181_FFB2_C4741_S3	G/T	31047861
		A/G	31047886
		T/C	31047899
		G/A	31047996
	KG182_FFB2_C3566_S9	A/G	31308475
	KG183_FFB2_C2_S1	A/G	31455617
	KG268_EgMBAGL2-3	C/G	33901095
	KG105_PSII1	A/G	59000784
		G/T	59000929

El **desequilibrio de ligamiento (DL) inter-locus** fue significativo en los genes candidatos posicionados en 10 grupos de ligamiento (GL2, GL3, GL5, GL6, GL7, GL8, GL10, GL12, GL13 y GL14) como pueden observarse en los gráficos de las figura 11.1 y figura 11.2 (Anexo 11). Como puede revisarse en la tabla 5, el GL con un valor de DL medio más fuerte fue el GL13 (r^2>0,8; DE=±0,009) con una distancia media de 3Mpb, y los GL11 (r^2>0,019; DE=±0,014) y 15 (r^2>0,007; DE=±0,000) obtuvieron un valor del DL prácticamente inexistente y su distancia media 6Mpb y 5 Mpb, respectivamente. En cuanto a las desviaciones estandar que muestran la diversidad de los DL a nivel inter-locus en cada cromosoma, el GL7 presentó una mayor desviación respecto a la media (DE=±0,305) y los GL13 (DE=±0,009) y GL6 (DE=±0,064) fueron los que menor desviación respecto a la media tenían.

Tabla 5: DL medio (r2) inter-locus en cada grupo de ligamiento (GL), y su distancia media.

	DL Medio (r^2)	**DE (±)**	**Distancia Media (Mpb)**	**DE (±)**
GL2	0,281	0,239	20	14
GL3	0,787	0,168	15	13
GL5	0,326	0,168	13	14
GL6	0,437	0,064	1	0
GL7	0,490	0,305	3	3
GL10	0,139	0,037	10	2
GL11	0,019	0,014	6	5
GL12	0,108	0,083	11	6
GL13	0,877	0,009	3	0
GL14	0,269	0,234	10	5
GL15	0,007	0,000	5	0

El **DL intra-locus** se presentó en todos los genes candidatos que poseían más de un SNP distribuidos a lo largo de los 16 grupos de ligamiento, con valores de r^2 superiores a 0,8 y con una

distancia media de 38pb (DE=33). Por tanto estos SNPs posicionados dentro de cada gen candidato conformaron diferentes haplotipos, como puede observarse en las figuras 11.1 y 11.2 del anexo 11.

En la figura 2, se muestra el mapa de DL para el GL2, observándose cuatro bloques bien diferenciados correspondientes al DL máximo significativo entre los SNPs de los genes candidatos GLO2, KG196, KG181 Y KG185, abarcando una distancia máxima de 238pb entre el primer y tercer SNP del gen candidato KG196. Entre estos 4 bloques los valores D' varían entre 0,4 y 0,91, entre el bloque 2 y 3 (KG196; KG181) y entre el bloque 3 y 4 (KG181 y KG185) (Anexo 11, Figura 11.2). También, se muestra la presencia de DL inter-locus, siendo estadísticamente significativo, por ejemplo, el DL entre los SNPs de los genes candidatos GLO2 y KG105 (r^2=0,78; *p<0,000*), y cuya distancia es superior a 50Mpb (Figura 2).

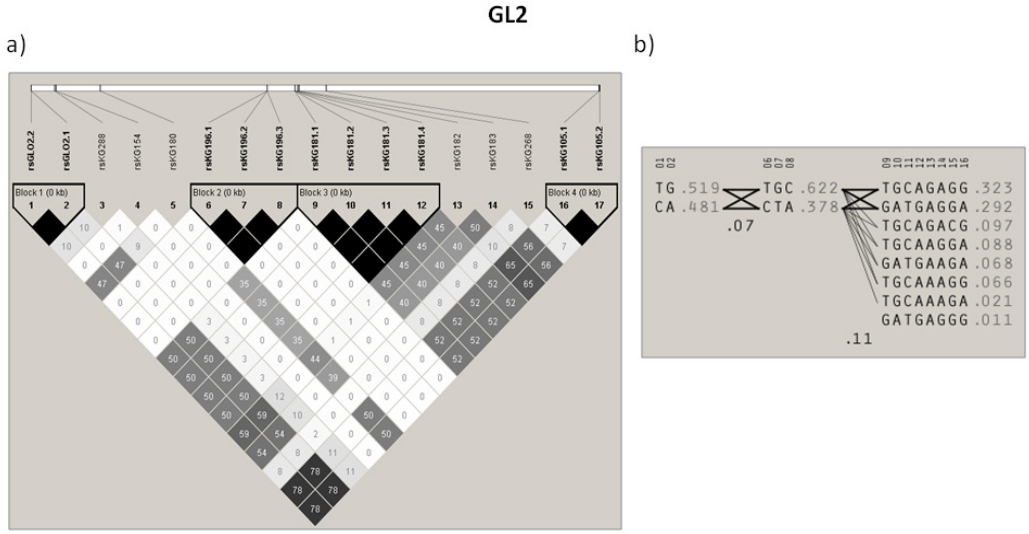


Figura 2: a) Mapa de desequilibrio de ligamiento entre los marcadores SNP posicionados en el cromosoma 2. Mapa de desequilibrio de ligamiento entre los marcadores SNPs posicionados en el cromosoma 2. El mapa muestra como los SNPs de un mismo gen candidato se encuentran entre ellos en DL mostrando el cuadro en negro e indicando que r^2>0,8, y formando un haplotipo. También se muestra el DL inter-locus, como es el caso de los SNPs de los genes candidatos GLO2 (KG)y KG105 están en DL siendo el valor R^2=0.78, y por lo tanto mostrando un fuerte DL, a pesar de la distancia que los separa indicada a modo orientativo por la barra horizontal superior del mapa. b) Haplotipos formados por los marcadores SNP de los genes candidato posicionados en el GL 2 a nivel inter-locus e intra-locus. En la imagen se muestran cada haplotipo del bloque con su frecuencia en la población y las conexiones entre los diferentes haplotipos. El valor D' se muestra en la parte inferior entre los bloques e indica el grado de recombinación entre ambos bloques.

A continuación, el desequilibrio de ligamiento (DL) se calculó para cada cromosoma en un valor de 100.000kb con el objeto de incorporar todos los marcadores presentes en cada cromosoma y hacer el análisis inter locus de cada cromosoma e intra locus para aquellos genes candidatos que presentaron más de un polimorfismo, con el objetivo de identificar los posibles haplotipos presentes. Los mapas de DL de cada uno de los cromosomas se muestran en el Anexo 11, Figura 11.1 a y 11.1b.

En total 142 regiones polimórficas formaron 591 pares de r^2 calculados, de los cuales 446 (75%) fueron significativos (p<0.05). Cuando se enfrentan los valores de r2 frente a su distancia genética (cM)

entre las regiones SNP que fueron significativosse muestra como el DL desciende muy lentamente sin alcanzar el valor de r^2 =0,2 a los 140cM (Figura 3).

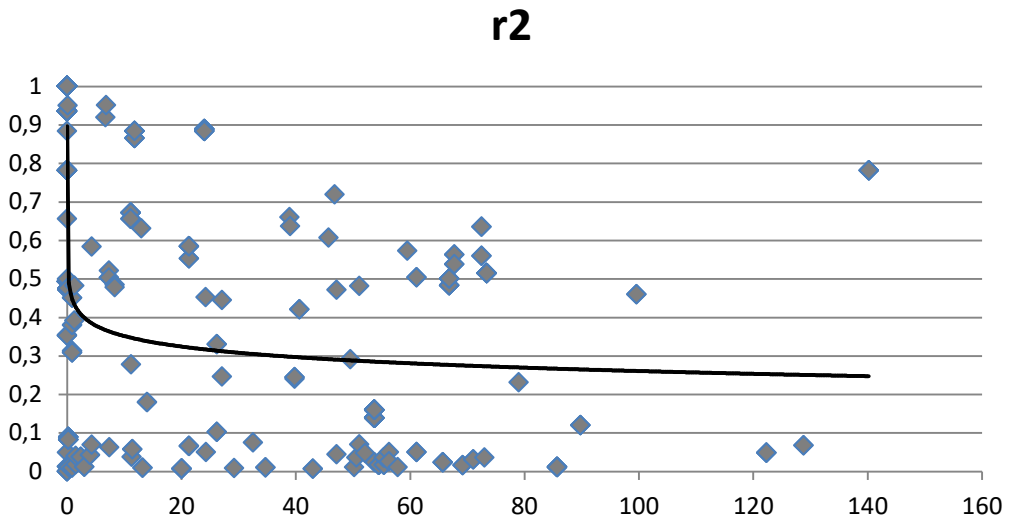


Figura 3: Disminución del DL (r^2) como función de la distancia genética (cM)entre dos SNPs. La mezcla de los datos de todos los marcadores a lo largo de los 16 cromosomas se utilizó para estimar el el DL entre los locus. La línea muestra la curva logarítmica que ajusta los datos de los 238 individuos que representan el conjunto de individuos.

4.3. Análisis de componentes principales (ACP) y estructura poblacional

Los resultados de ACP mostraron que los tres primeros componentes principales resumían un 22% de la variación total observada en la población de estudio, mientras que el 50% de la variabilidad se alcanzaba en CP=10 y el 100% CP=62, aunque a partir de CP=25 el aumento del valor de varianza explicado era muy bajo (Figura 4). Los datos están mostrados en la tabla 12.2 del anexo 12.

Por otro lado, en el gráfico de la figura 5, se muestran los 2 primeros ejes de los componentes principales, en los que puede observarse la presencia de tres hipotéticas subpoblaciones. Estas subpoblaciones representan el el 9,6% y el 6%, respectivamente de la variabilidad total del conjunto de los genotipos.

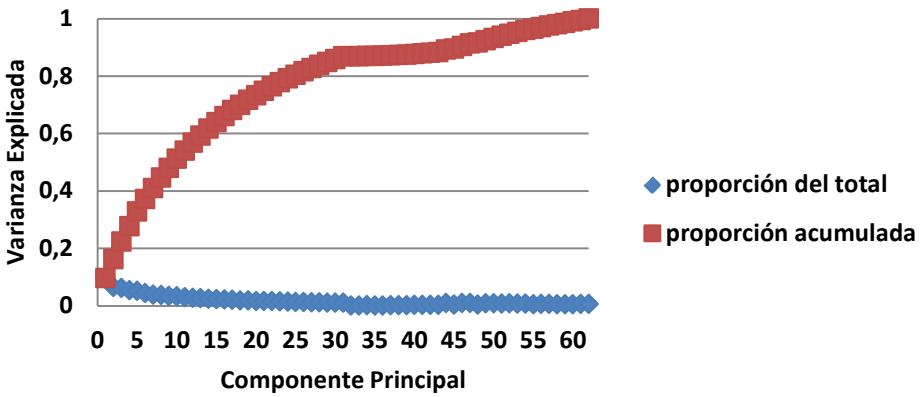


Figura 4: Gráfico de la varianza explicada por cada eje obtenido como componente principal.

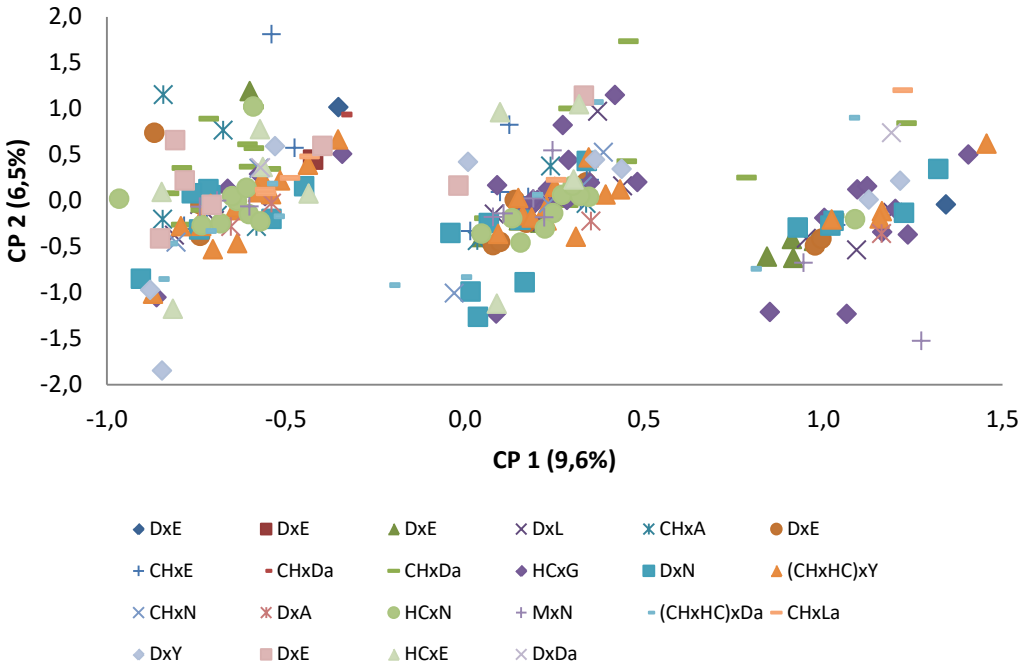


Figura 5: Gráfico de los ejes correspondientes a los dos primeros componentes principales. El CP1 explica el 9,6% de la variabilidad observada y el CP2 el 6,5%. A= Avros, D=Dura, Da=Dami, CH= Chemara, E = Ekona, HC= Harrison-Crossfield, L= LaMe, M= Mardi,N= Nigeria, Y= Yangambi.

Por otro lado, se analizó la presencia de estructura poblacional mediante el programa STRUCTURE (Pritchard y col., 2000). Los resultados mostraron que la probabilidad más alta para la presencia de subpoblaciones fue con K=2 (Figura 6) y el valor de máxima probabilidad la subpoblación Ln p(D)=-6057,41 con una desviación estándar de 0,733. Además, en la figura 8b se observa asimetría en las proporciones asignadas a cada grupo, y una fuerte asignación (>95%) a una de las subpoblaciones señalando una verdadera estructura poblacional (Pritchard y col., 2009).

Tabla 5: Estimación de la presencia de subpoblaciones en el conjunto de la población (238 genotipos) evaluados con STRUCTURE mediante 29 SNPs El período largo de "burn-in " seleccionado fue de 10000 repeticiones y la cadena de Markov 50000 en un modelo de población mezcla y frecuencias alélicas correlacionadas. Se testaron entre 1 y 10 subpoblaciones. K= número de subpoblaciones. Ln p(D) estimación del logaritmo de la probabilidad de los datos, DE= desviación estandar.

K	Ln p(D)	DE
1	-6440,865	0,099
2	**-6057,41**	**0,733**
3	-6068,265	27,987
4	-6071,765	56,330
5	-6152,26	234,408
6	-6087,445	97,371
7	-6224,415	188,743
8	-6274,67	204,281
9	-6332,885	202,247
10	-6468,305	343,528

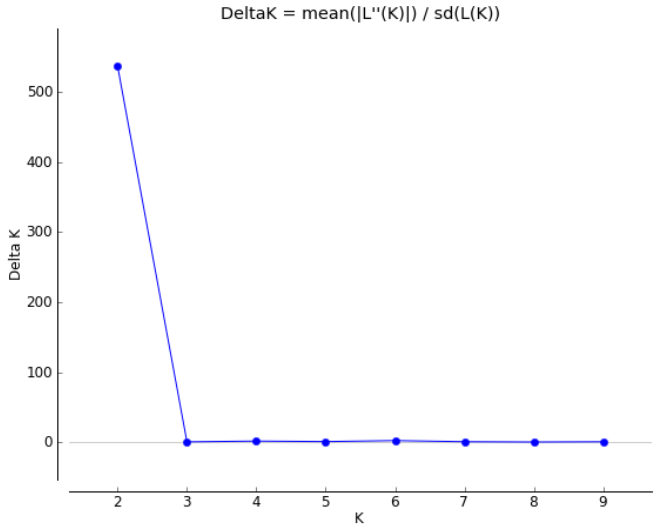


Figura 6: Estimación del número de subpoblaciones mediante el valor Delta K utilizando el método propuesto por Evanno y col. (2005)

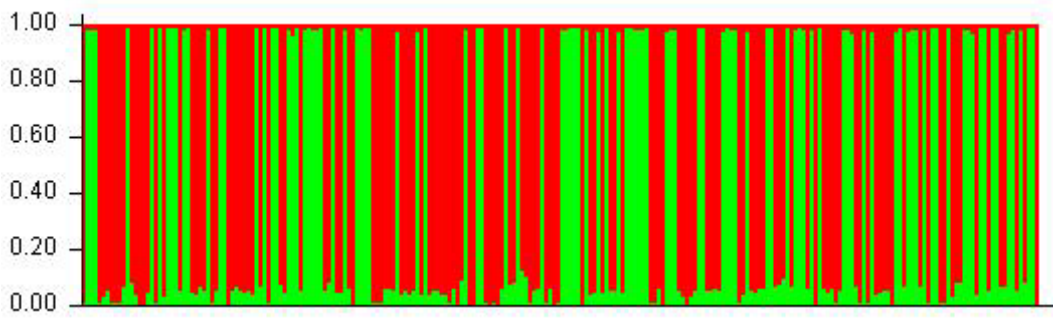


Figura 7: Estructura poblacional de los 238 genotipos basada en 29 marcadores SNPs bajo la asunción de una sub-población K=2. Cada genotipo se representa por una barra divida en dos colores diferentes indicando la fracción estimada de cada individuo que corresponde a cada sub-población. El eje X muestra el número de genotipos y los el eje Y el porcentaje de pertenencia a cada grupo.

Por otro lado para conocer como la estructura poblacional puede afectar a la varianza fenotípica se realizo un análisis de regresión múltiple en el que la variable dependiente fue cada carácter y las variables independientes estaban formadas por la matriz Q. Todos los caracteres se vieron afectados por la estructura poblacional, explicando cada modelo valores superiores al 80% de la varianza fenotípica.

Tabla 6: Análisis de regresión múltiple entre el carácter fenotípico y la matriz Q de estructura poblacional. R^2 es la proporción de varianza representada por la estructura poblacional; Valor P es la significancia estadística del modelo.

Carácter	R^2	Valor p	Carácter	R^2	Valor p
BN	0.9484	<0.0001	**FW**	0.8560	<0.0001
BW	0.9797	<0.0001	**HT**	0.9465	<0.0001
CPO	0.8750	<0.0001	**MF**	0.9935	<0.0001
FN	0.8511	<0.0001	**OWM**	0.9648	<0.0001

4.4. Asociación fenotipo-genotipo

Se analizaron 3 modelos estadísticos para la búsqueda de asociaciones entre el fenotipo y el genotipo de la población seleccionada. Estos modelos fueron comparados mediante un gráfico de cuartiles o "QQ plot" de los valores p obtenidos como resultados en la prueba de Fisher para la asociación realizada en TASSEL v.5.0 (Bradbury y col.2007) (Figura 8), en el que para su representación se unieron todos los caracteres y marcadores. En el gráfico inferior se observa que en el análisis de asociación incluyendo la matriz del análisis de componentes principales (GLM+P) hay un mayor desvío de los valores p observados en los 149 marcadores con respecto a los valores esperados. En cambio, el modelo simple y el modelo de población estructurada muestran valores muy próximos de la desviación de la recta, por lo que debido al tipo de población y al bajo nivel de estructura poblacional analizado en STRUCTURE mostrada se seleccionó el modelo lineal generalizado (GLM) para la búsqueda de asociaciones significativas entre los SNPs y los diferentes caracteres. Los resultados a partir del modelo simple (GLM) asociaron 15 SNPs de 11 genes candidatos con 6 de los 8 caracteres de estudio (Tabla 7).

El carácter **BN** tuvo dos SNP asociados significativamente que correspondían a dos genes candidatos diferentes. El gen candidato KG254_PLT2 cuyo marcador SNP (A→C; Transversión) se posiciona en la base 110 de su amplicón y explica 1,5% de la variación fenotípica. Los genotipos en los que el alelo C estuvó presente fueron los que mostraron un mayor número de racimos (10,95) (Figura 9.a, gráfico a). El SNP del gen candidato P64_PAT4, a 104pb del inicio del amplicón, explica un 1,1% de la variación fenotípica, siendo los genotipos que presentaron el alelo T los que mostraron una variación ligeramente superior a la media mostrada por la población (Figura 9.a, gráfico a).

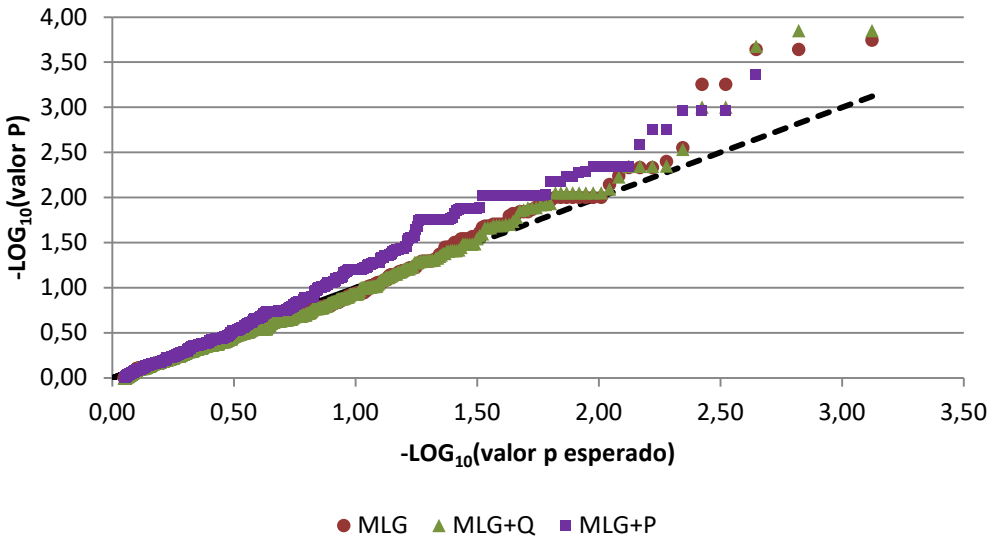


Figura 8: Gráfico de cuartiles de los valores p, para la comparación de cada modelo de asociación elejido. La línea diagonal negra muestra los valores p acordes al valor esperado y al obtenido (x=y), cuando la diagonal se transforma en una diagonal intermitente indica los valores de p<0,05, los cuáles son significativos en cuanto a la asociación estadística. El resto de puntos muestra los valores p ordenados para cada asociación analizada mediante el modelo lineal general (GLM), el modelo lineal general incluyendo como covariable la matriz de estructura poblacional (k=2) (GLM+Q), y el modelo lineal general incluyendo como covariable la matriz de análisis de componentes principales (GLM+P).

Cuatro son los SNPs significativamente asociados al carácter **CPO**. Estos SNP se correspondían con los genes candidato KG135_M3117, KG261_PO3_5-10, KG194_FFB11_C3877_S4. Todos los SNPs explican un porcentaje de varianza similar sobre el fenotipo (entre 1,1% y 1,3%), y es de destacar que los genes candidato KG261 y KG194 mapean en el cromosoma 11, pero no se encuentran en DL entre ellos (Anexo 11, Figura 11.1b). En cambio, los 2 SNPs pertenecientes a KG135 (C/G, 22; A/C,127) y localizado en el cromosoma 1 están en total DL (R^2=1) y conforman un haplotipo. En el gráfico f de la figura 9.b pueden revisarse los alelos que presentan la mayor variación con respecto a la media del fenotipo para cada gen candidato.

FN está asociado significativamente con el SNP (G→T; Transversión) del gen candidato KG288. Los genotipos con el alelo G son los que muestran mayor número de frutos con respecto a la media de la población, y aproximadamente 162 frutos más que aquellos genotipos que poseían el alelo T (Figura 9, gráfico b).

Tabla 7: Resultado de la asociación de los genes candidatos con los caracteres de interes agronómico en *E.guineensis* Jacq.

Carácter	GC	SNP	Crom	Posición	R2	Valor p
BN	KG254_PLT2	A/C	9	15548277	0,015	0,0108
	P64_PAT4	C/T	12	15018030	0,011	0,0197
CPO	KG135_M3117	C/G	1	21688462	0,013	0,0153
		A/C		21688567	0,013	0,0153
	KG194__ FFB11_C3877_S4	A/G	11	20390248	0,012	0,02
	KG261__ PO3_5-10	G/T	11	11857044	0,011	0,0293
FN	KG288_EgTPase	G/T	2	8992884	0,016	0,0053
FW	KG171_HtC4_4489	A/T	4	4462591	0,008	0,0494
	KG186_FFB6_C2082_S1	G/T	6	33052707	0,008	0,0486
		A/C		33052739	0,008	0,0486
	KG27_BKACPII_1	C/G	10	22949603	0,022	0,0012
		A/C		22949604	0,022	0,0012
MF	KG148_MUM4	C/T	3	8044652	0,013	0,0196
	KG125_SQUA3	C/T	15	13726804	0,02	0,0027
OWM	KG135_M3117	C/G	1	21688462	0,016	0,0056
		A/C		21688567	0,016	0,0056
	KG148_MUM4	C/T	3	8044652	0,009	0,0446

En el carácter **FW** se encuentran el gen candidato que explican mayor porcentaje de varianza del marcador sobre el carácter y los genes candidatos que menos porcentaje de varianza explican sobre

el fenotipo. El gen candidato cuyos SNPs explican el mayor efecto para la variación fenotípica es KG27 (r^2=2,2%) en el carácter peso de los frutos. Este gen candidato presenta 2 SNPs en las posiciones 60 y 61 de su amplicón en desequilibrio de ligamiento (Apartado 4.1 de este capítulo), ambos transversos (C→G; A→C), y siendo la base C y la base A las que mostraron mayor peso de los frutos (Figura 9, gráfico e), con una diferencia respecto a las bases G y C de 1,71 g y a la media fenotípica de 1,48 g. En cambio, los genes candidatos KG171 y KG 186 explican únicamente el 0,8% de la variación fenotípica. El alelo A del amplicón de KG171, cuyo cambio de base también es una transversión (A→T; posición en amplicón 220pb) es la que muestra un mayor peso de los frutos (Figura 9; Gráfico e), siendo 0,21g superior a la media fenotípica en la población. En KG186, los genotipos que poseían los alelos T y A son los que poseen un mayor peso de los frutos con 0,53 gr sobre la media, y 0,79 gr más sobre los genotipos que poseían el alelo G y T.

Los marcadores de los genes candidatos KG125 y KG148, ambos transiciones (C→T), se asocian significativamente con **MF** explicando el 1,3% y 2% respectivamente, de la variación fenotípica del carácter. En ambos casos los genotipos con el alelo T presente son los que obtienen una mayor proporción de pulpa con respecto al tamaño del fruto, siendo 2,12% (KG125) y 0,84% (KG148) superior a la media mostrada por la población (Figura 9, gráfico c).

Por último, el carácter OWM está significativamente asociado con los marcadores de los genes candidato KG135 y KG148. Ambos genes candidato están asociados con los caracteres CPO y MF, respectivamente. En el caso de KG135, los marcadores explican un 1,6% de la variación del carácter, y KG148 un 0,9%. Y en ambos casos, los genotipos que poseen los mismos alelos con una variación del fenotipo superior a la media para el carácter CPO (KG135) y MF (KG148) son los que obtienen una variación superior a la media de la población en OWM (Figura 9, Gráfico d).

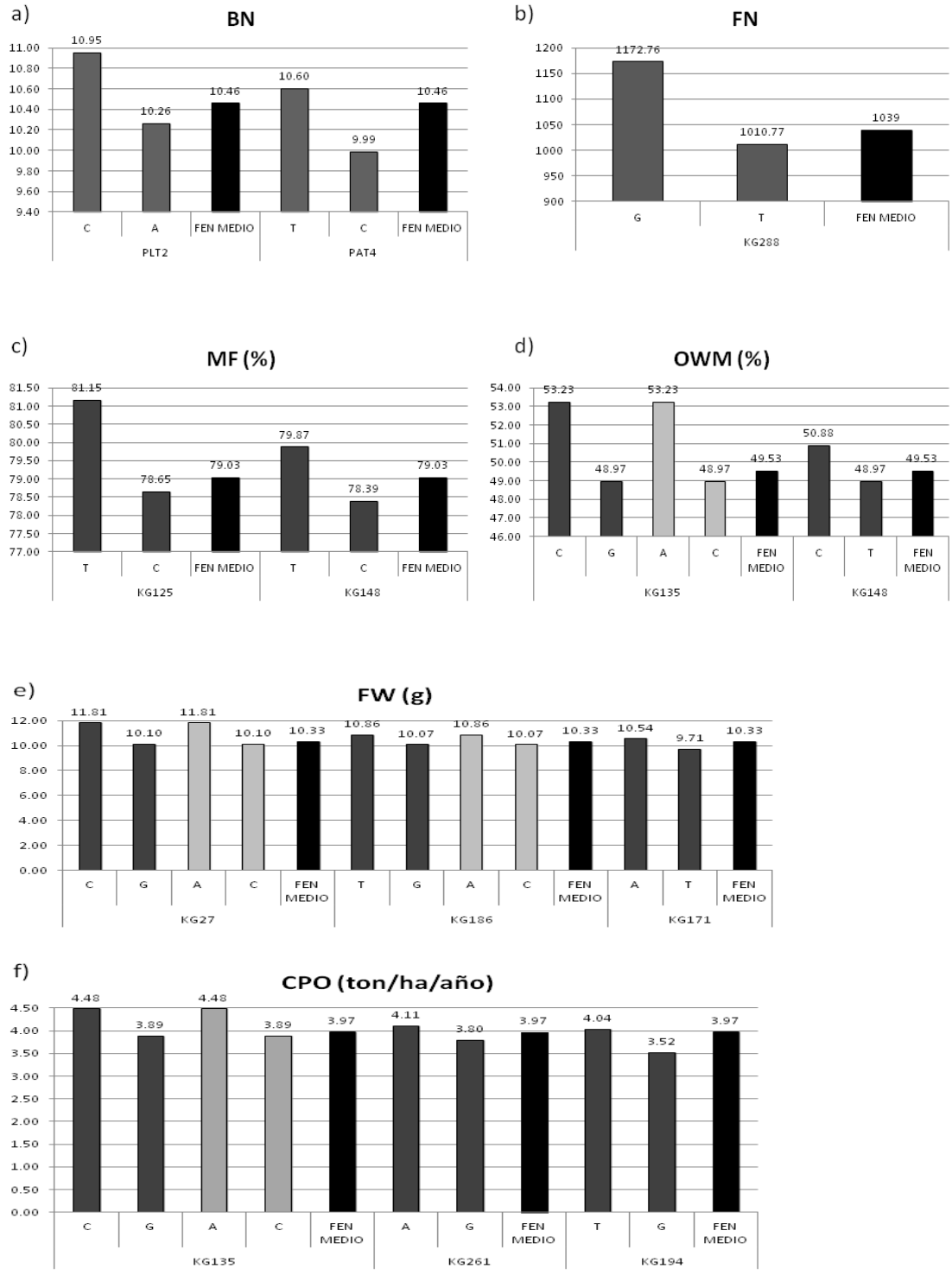


Figura 9: Gráficos de la asociación de los efectos de cada alelo en cada gen candidato. a) Efecto de los alelos de los genes candidatos KG254 y P64 sobre el carácter BN, y sus valor fenotípico medio en la población; b) Efecto de los alelos del gen candidato KG288 y su valor fenotípico medio en la población para FN; c) Efecto de los alelos de los genes candidatos KG125 y KG148 sobre el carácter MF (%) y su valor fenotípico medio en la población; d) Efecto de los alelos de los genes candidatos KG135 y KG148 sobre el carácter OWM (%) y su valor fenotípico medio en la población. e) Efecto de los alelos de los genes candidatos KG27, KG186 Y KG171 sobre el carácter FW (g), y sus valor fenotípico medio en la población; f) Efecto de los alelos de los genes candidatos KG135, KG261 y KG194 y su valor fenotípico medio en la población para CPO (ton/ha/año).

5. DISCUSIÓN

Este capítulo aborda la última etapa de un estudio de mapeo por asociación que es la búsqueda de asociaciones estadísticamente significativas entre el fenotipo, en nuestro caso algunos de los caracteres de interés agronómico más importantes de *E.guineensis* Jacq., y el genotipo determinado por los polimorfismos de un solo nucleótido detectados en los genes candidatos seleccionados.

5.1. Fenotipado de la población

El fenotipado de la población es elemental en el mapeo por asociación y la elección de un gran número de individuos que conformen la población de estudio permitirá obtener datos fenotípicos más fiables y precisos, y en consecuencia los resultados de la asociación tendrán mayor potencia (Ingvarsson y Street, 2011). Rafalski (2010) recomienda entre 100 y 500 individuos para conformar la **población** de estudio, en este estudio se seleccionaron 238 individuos, con fenotipos extremos aumentando así la variación. Esta característica se puede observar en los coeficientes de variación obtenidos en los resultados (Tabla 2) para los caracteres CPO, FN y FW, dónde este valor fue superior al 35%.

Por otro lado, las **correlaciones entre caracteres** arrojaron resultados similares a otros estudios en *E.guineensis* Jacq como en el caso de la correlación negativa establecida entre BN y BW (Billote y col., 2010; Montoya y col., 2014). Corley y Tinker (2015) explica que esta correlación negativa es debida al aumento del peso de los racimos con la edad de la palmera y la tendencia a disminuir el número de racimos, pero sin ver perjudicado el rendimiento de la planta. Montoya y col. (2014) también apreció una correlación significativa entre BW y FN al igual que en nuestro estudio. También, MF y FN están correlacionados, ya que parece que los frutos de menor tamaño tienen menos mesocarpio manteniéndose el peso de la semilla constante, como postulan Henson y Dolman (2004). Jalani (1994) comparó frutos de peso extremo y encontraron que los racimos cuyos frutos eran más pesados obtenían un mayor ratio MF, confirmando también esta correlación. Aunque algunos autores postulan que el rendimiento de aceite en *E. guineensis* está fuertemente influenciado por BW, en este estudio no se ha encontrado una correlación con significancia estadística suficiente que refuerce esta afirmación con ninguno de los caracteres relacionados con el rendimiento de aceite como CPO y OWM, pero sí CPO se correlacionó con BN. Está correlación sustenta que la variación en las palmeras maduras en el número de racimos contribuye más en los ciclos de rendimiento que BW (Broekmans, 1957; Brédas y Scuvie, 1960).

A pesar de las diferencias de las poblaciones de los estudios con los que se han comparado la significancia y el sentido de la correlación mostrado por los resultados de este estudio aprecia diferencias sugiriendo una relación inherente a la especie entre estos caracteres y no al tipo de población, aunque hay que reseñar que se necesitan más estudios para confirmar este hecho.

5.2. Desequilibrio de ligamiento

El DL desempeña un papel fundamental en el mapeo por asociación, ya que el estadístico ($\mathbf{r^2}$) con el que se correlacionan los alelos de la población definirá el rango predictivo del marcador genético

con el fenotipo. Los resultados obtenidos en esta tesis muestran como el DL intra-locus, considerando como un locus cada gen candidato, están altamente correlacionados, y conforman haplotipos. Estos haplotipos señalan que los marcadores SNPs de un mismo gen candidato se heredaran conjuntamente, apoyado también por la distancia media en pb (38) a la que se encuentran los diferentes polimorfismos dentro de cada gen candidato (Anexo 11; Figura 11.1). Como postulan Ardlie y col. (2002) y recogen en su artículo Patnala y col. (2013), un "locus" del gen que posee los SNPs y están de DL son más propensos a conservarse en la descendencia a pesar de los eventos recombinatorios que puedan darse en regiones próximas. Por tanto, el análisis de **DL dentro del gen candidato** es una manera de reducir la región genómica susceptible de estar afectada por el carácter agronómico, y todos los SNPs se heredaran conjuntamente mostrando frecuencias similares en los individuos afectados en la población. Estos SNPs muy próximos entre sí y en DL pueden ser representados por un único SNP seleccionado entre ellos denominado SNP diana o "tagSNP", aunque en esta tesis no fueron determinados. En cambio en otros estudios de mapeo por asociación mediante el abordaje de genes candidato es habitual la determinación de los haplotipos y/o el SNPdiana que posteriormente será el lanzado en el análisis de asociación. Este es el caso de Jourdan y col. (2015) que identificaron un total de 470 SNPs que conformaron 169 haplotipos en 17 genes candidato relacionados con el contenido de carotenoides y el color de la raíz en zanahoria, o de Simko y col. (2009) que mediante la identificación de los SNPs diana en los haplotipos que conformaban los SNPs posicionados en diferentes regiones de un gen de resistencia a una enfermedad causada por dos virus de la familia *Tombusviridae* en lechuga.

En cuanto al **DL inter-locus** (Tabla 5)**,** los resultados muestran como el DL medio varía entre los diferentes cromosomas, esto es debido a la diferencia de marcadores SNPs existentes en cada cromosoma y a la distancia a la que se encuentran posicionados entre ellos. En general, aquellos cromosomas donde el DL medio es más pequeño existe una mayor distancia entre los marcadores, cumpliendo así lo que muestra la teoría (Myles y col. 2009). Aunque, por ejemplo, entre los marcadores de los loci KG123 y KG105 el DL fue fuerte (r^2= 0,78) encontrándose a una distancia aproximada de 52Mpb (Resultado no mostrado), esta alta correlación a pesar de la gran distancia se debe a las frecuencias álelicas presentes en cada marcador y puede tener su origen en alguno de los factores interdependientes que afectan al DL y a sus patrones como son la recombinación, la deriva genética, la endogamia, las mutaciones o el flujo genético.

Por último, la **lenta disminución del DL** (Figura 3) en la población de estudio indica que la distancia a la que el DL se conserva es alta, y por tanto la probabilidad de recombinación a esa distancia será baja. Este alto DL a lo largo del genoma puede deberse a que la población seleccionada para el estudio procede de un proceso de selección recíproca recurrente. El proceso de mejora de la población puede crear regiones en DL, ya que favorece las combinaciones alélicas y/o promueve la deriva genética (Palaisa y col., 2003). Además, los marcadores SNPs se posicionan en genes candidatos que pueden estar relacionados con los caracteres agronómicos de interés, lo que ha podido influir indirectamente

mediante la elección de posibles combinaciones alélicas en diferentes genes candidatos que hayan participado de este proceso de mejora.

Una lenta disminución del DL en *E.guineensis* Jacq. también fue obtenido por Jin y col. 2016, los cuales observaron que el DL se reducía a la mitad a una distancia mayor de 200Kb en todas las palmeras estudiadas, incluídas *Tenera*, las cúales son coincidentes con nuestra población. En cambio, en la población de estudio de Teh y col. (2016) el DL obtuvó una disminución más rápida en su población *Tenera* que en esta tesis y que lo obtenido por Jin y col. (2016). Estas comparaciones ponen de manifiesto como el DL varía no sólo entre las distintas especies, o dentro de la misma especie, sino también en cada población debido a su propia historia de selección, en el caso de los cultivos.

5.3. Estructura poblacional

La estructura poblacional es uno de los principales factores que pueden originar asociaciones espurias durante el mapeo por asociación ya que afecta al DL, y puede generar una distribución desigual de los alelos dentro de los posibles subgrupos (Knowler y col., 1988). En esta tesis se han aplicado los dos métodos más utilizados para desengranar la presencia de una posible estratificación o estructura en la población del estudio. Estos métodos son el ánalisis de componentes principales (**ACP**) (Price y col., 2006) y el método de inferencia bayesiana desarrollado por Pritchard y col. (2000), mediante el programa **STRUCTURE** (Pritchard y col., 2000)**.** En el caso de **ACP** se utilizaron los 142 SNPs definitivos para los que se había genotipado la población cuyo MAF > 5%. En cambio, para el análisis mediante **STRUCTURE** se seleccionaron 29 SNPs (MAF>5%) que se encontraban bajo equilibrio de Hardy-Weinberg, ya que el modelo matemático desarrollado así lo requiere (Pritchard y col., 2000).

En cuanto a los resultados obtenidos, en el **ACP** las posibles subpoblaciones representadas en el gráfico (Figura 5) como el cruce de procedencia de cada individuo, no poseen una clara definición para establecer una población estratificada, ya que en cada uno de los conjuntos los individuos se encuentran mezclados. Además la varianza explicada por los dos primeros componentes principales sólo alcanza el 16% de la varianza total (PC1=9,6% y PC2=6,5%),y siendo necesarios más de 25 CP (80%) para poder resumir la estructura poblacional. Estos valores se desvían del objetivo del método que es resumir la variación observada en todos los marcadores en un número menor de observaciones, y pudiendo interpretar estos componentes principales como posibles subpoblaciones no observadas de las que se originaron los individuos del conjunto de datos analizados o sus ancestros (Zhu y col., 2008). En cambio, en el análisis llevado a cabo en **STRUCTURE** y utilizando el método "ad hoc" propuesto por Evanno y col. (2005) para su interpretación, reveló una mayor probabilidad para K=2, es decir, la presencia de dos posibles subpoblaciones, y por tanto una baja estratificación poblacional. Además el análisis de regresión entre los subgrupos y los caracteres de estudio mostró fuertes relaciones entre los diferentes subgrupos y el carácter como puede revisarse en los resultados, lo que indica que el origen de los subgrupos no es en sí una subpoblación real, sino un sesgo introducido por efectos de la selección del material para el estudio, en el que se eligieron genotipos de una misma familia con diferentes valores fenotípicos, y por tanto un efecto debido a los ciclos de selección a los han sido sometidos las familias

que componen la población de estudio, favoreciendo la presencia de unos alelos sobre otros. También es necesario destacar que este resultado es una aproximación a la posible estructura poblacional ya que el número de SNPs con el que se realizó fue muy bajo, y al ser marcadores bialélicos son necesarios un mayor número de marcadores distribuidos a lo largo del genoma. Este análisis debe validarse o bien aumentando el número de marcadores SNP o utilizando marcadores SSR que al ser multialélicos tienen una mayor resolución en este método basado en la inferencia bayesiana. Zhu y col. (2008) recomienda un número de SSRs que cuadruplique el número de cromosomas de la especie, para al menos incorporar dos marcadores por cromosoma. Los resultados esperados en esta tesis en cuanto a estructura poblacional eran que debía ser nula o baja, debido a que la población seleccionada puede clasificarse como sintética. Este tipo de población, como postulan Breseghello y Sorrells (2006) es la que mejor se aproxima a la suposición de apareamientos al azar, por lo que la presencia de la estructura poblacional será nula o baja.

5.4. Elección del modelo de asociación

La asociación estadística genotipo-fenotipo puede realizarse mediante diferentes modelos como se explicó en la introducción a este capítulo. Como los resultados arrojados por la estructura poblacional no son reveladores, se realizó la asociación bajo el modelo linear generalizado simple, el modelo linear generalizado incluyendo como covariable la matriz P procedente de ACP, y el modelo linear generalizado adjuntando como covariable la matriz de estructura poblacional Q procedente de STRUCTURE. La elección del modelo que mejor se ajustaba se realizó como propone Balding (2006) mediante una gráfica logarítmica QQ en el que se compararon el logaritmo de los p valores de asociación obtenidos en cada modelo frente a los p valores esperados, pudiendo seleccionar el modelo que menos se desviaba del valor esperado para evitar asociaciones espurias. Este método de selección es aplicado por diferentes autores para la selección del modelo que mejor se ajusta como Fritsche y col. (2012) en su estudio de asociación mediante genes candidatos en colza, el estudio de Mariette ycol. (2015) en melocotonero de asociación mediante el escaneo completo del genoma o el estudio realizado en vid también mediante el escaneo completo del genoma en el que se comparan el modelo GLM, MLM+K, MLM+Q+K (Tello y col., 2015).

En esta tesis el modelo que mejor se ajustó a la recta fue el modelo basado en GLM simple como se muestra en los resultados (Figura 8), aunque fue bastante similar al modelo en el que la covariable era Q (estructura poblacional). Como puede observarse en la gráfica de resultados GLM presenta dos marcadores asociados que no los presenta el modelo GLM+Q, lo que puede indicar la presencia de dos falsos positivos, pero en cambio los dos marcadores superiores presenta una ligera desviación superior en GLM+Q. Esta controversia puede deberse a una sobreparametización del modelo aumentando la tasa de falsos positivos en situaciones de baja estructuración genética, al igual que ocurre al incluir los CP al modelo en estas situaciones como postulan Peña Malavera y col., (2014), aunque en su estudio la introducción de los CP disminuyó la tasa de falsos positivos.

5.5. Asociación genotipo-fenotipo

GLM fue el modelo seleccionado para el análisis de asociación que arrojó 15 SNPs de 11 genes candidatos asociados con los caracteres relacionados con la producción BN, CPO, y OWM, y con los componentes de racimo FN, FW yMF (Tabla 7).

Los SNPs de los genes candidatos **KG135** y **KG148** se asociaron a dos caracteres diferentes, como puede observarse en los resultados. **KG 135** se asoció a los caracteres de producción **CPO** y **OWM**. Estos caracteres presentan una significativa correlación, lo que sugiere la participación del gen candidato sobre ambos caracteres. A esta evidencia se añade en que son los mismos SNPs los que aparecen con un incremento del valor fenotípico (C-A). Además el DL entre ambos SNPs es máximo, por lo que conforman un haplotipo, aunque observando los resultados de DL (Anexo 11;Tabla 11.1, GL1) es el haplotipo con menor frecuencia (0,088) en la población el que tiene un mayor efecto sobre el fenotipo, lo que no es muy habitual. En algunos estudios de asociación en *A. thaliana*, por ejemplo, esto sucede en genes candidatos relacionados con el tiempo de floración, en los que casi la mitad de los polimorfismos que alteran el tiempo de floración en estos son raros (MAF<10%) (Ehrenreich y col., 2009). También en los polimorfismos detectados en genes candidato relacionados con caracteres nutricionales y de calidad en tomate mediante el escaneo del genoma completo (Ruggieri y col., 2014). En cuanto a la funcionalidad del gen candidato en el capítulo 2 se detectó como un gen codificante de una enzima que participa en la α-oxidación de ácidos grasos en la ruta del ácido fitánico, aunque su funcionalidad en las plantas es desconocida. Este gen candidato correspondia a un EST secuenciado por Bourgis y col., (2011) procedente del mesocarpo del fruto de *E.guineensis* Jacq. y se encontraba próximo a un QTL relacionado con un indicador de calidad como es el índice de iodo (Anexo 3; Tabla 3.2). La presencia en el mesocarpo del fruto y su proximidad a un QTL de calidad de aceite sugiere una posible actuación sobre estos caracteres relacionados con el contenido de aceite del fruto.

Por otro lado, el SNP de **KG148 (GL·3)** mostró asociación con el carácter de producción **OWM** y el componente de rácimo **MF**. Estos caracteres mostraron una correlación significativa inferior al 50%, pero en sentido positivo, aunque en este caso el **SNP T** tiene un efecto positivo sobre el tamaño del mesocarpo en relación al fruto y el **C** sobre el peso húmedo de aceite en el mesocarpo. Este gen candidato codifica una enzima que cataboliza la reacción de conversión de UDP-D-glucosa en UDP-L-Ramanosa, necesario para la síntesis de la pared celular primaria en plantas y otros compuestos naturales como flavonoides, terpenoides y saponinas (Ikan, 1999; Ridley y col., 2001). En *A. thaliana* este gen (MUM4) participa en la formación del mesocarpo mucilaginoso y se ha observado que la pérdida de su funcionalidad produce semillas con mayor contenido oleaginoso, debido a que el mesocarpo mucilaginoso y la semilla comparten la misma fuente de carbono, la sacarosa, y la principal hipótesis que se baraja es que ambos compiten por esta fuente de carbono, por lo que en ausencia de la formación de mucílago la sacarosa se utilizará como fuente de carbono para la biosíntesis de aceite en la semilla (Shi y col., 2012). En *E.guineensis* no hay estudios que muestren cuál es la funcionalidad del gen ni sus mecanismos de actuación, y las diferencias del fruto son evidentes frente al fruto de *Arabidopsis*.

Aunque si este SNP es funcional y no sinómino podría causar la pérdida de funcionalidad del gen, al igual que ocurre en *A.thaliana*, aumentando la producción de aceite, en este caso en el mesocarpo. Para ello será necesario estudiar en profundidad el gen candidato seleccionando y secuenciando más amplicones en su región codificante y buscando nuevos polimorfismos que puedan afectar a su funcionalidad en *E.guineensis* Jacq., así como un estudio de los mecanismos de actuación en la especie. En cuanto a su asociación con la relación del tamaño del mesocarpio frente al tamaño del fruto, su actuación puede deberse al período de maduración del fruto en el que se sintetizan los ácidos grasos y el aceite y se producen numerosos cambios en las paredes celulares, aunque no hay literatura suficiente publicada que pueda ayudar a dilucidar su actuación, pero puede estar relacionado con la pérdida de funcionalidad del gen y su participación en la formación de la pared celular primaria.

El carácter **BN** es un componente del rendimiento de la palmera de aceite africana de control poligénico y cuya heredabilidad normalmente es alta, lo que implica que está bajo un fuerte control genético (Corley, 2015), sugiriendo que las asociaciones encontradas para este carácter pueden tener un alto grado de heredabilidad. Los resultados mostraron la asociación de dos SNPs pertenecientes a dos genes candidatos (**KG254** y **P64**) posicionados en GL9 y GL12, respectivamente. El gen candidato **KG254** se identificó como un factor de transcripción AP2 sensible a etileno participante en la regulación transcripcional y post-transcripcional durante el crecimiento y desarrollo de la planta, y también responsable de la evolución (Doebley y col., 1998; Riechman y col., 2000; Licausi y col., 2013). Trabanger y col., (2012) identificaron también un EST-SSR que formaba parte de un factor de transcripción AP. Estos autores postulan que los polimorfismos encontrados en los transcriptos codificantes de proteínas implicadas en la regulación transcripcional puede deberse al esquema de selección recíproca recurrente al que se la ha sometido a la palmera africana, aunque es necesario comprobar esta afirmación con futuros estudios para comprobar esta diversidad funcional a nivel fenotípico. Además, Licausi y col. (2013) los califica como ideales para la mejora tradicional o asistida con caracteres específicos como la tolerancia a estrés o mejora del rendimiento, como es el carácter al que se asocia. Por ejemplo, en arroz se asoció un factor de transcripción de este tipo con el índice de tolerancia a estrés por sequía (Yu y col., 2012), y en guisante un único SNP de un región intrónica de un factor de transcripción de este tipo se asoció con el número de vainas (Kujur y col., 2015).

P64, por su parte, codifica una enzima metiltransferasa COMT que participa en en el metabolismos de compuestos fenólicos, en la biosíntesis de lignina (Kim y col., 2009), implicada en el metabolismo de la pared celular, y que además posee un papel relevante en el sistema defensivo de la planta (Zhang y Erickson, 2012). Teh y col. (2014) obtuvieron actividad de esta proteína durante el desarrollo y maduración del fruto de la palmera aceitera africana, el cuál es uno de los principales componentes del racimo. La planta posee también ciertas partes en las que la síntesis de lignina es crucial, como las raíces y las hojas, elementos fundamentales para el desarrollo posterior de los racimos y sus componentes, sugiriendo una sútil implicación en el carácter.

El carácter relacionado con la productividad del cultivo, **CPO** que está relacionado con el contenido de aceite en el mesocarpio del fruto tuvo en total 4 SNPs que estuvieron asociados al carácter, dos pertenecientes a KG135, y discutida su implicación con anterioridad. Los otros dos SNPs pertenecen a dos genes candidatos detectados por co-localización en el capítulo 2, **KG194** y **KG261**, y posicionados ambos en el GL11, pero a una gran distancia entre ellos (8533204pb) por lo que su DL es muy bajo (r^2=0,03), aunque su valor D' no mostrado en los resultados es próximo a 0,5 indicando una posibilidad media de heredarse conjuntamente del 50%. **KG194** como puede revisarse en la tabla 4.1, del anexo 3, se posicionó a menos de 1cM de un QTL relacionado con el carácter FFB (ton/ha/año), el cuál es uno de los principales determinantes de la productividad en palmera de aceite (Jeenor y Volkaert. 2013). Este gen candidato como se discutió en el capítulo 2, codifica la isoforma E de una proteína del citocromo B5 participante en las reacciones de desaturación de acil-CoA de los ácidos grasos, y mediando en la formación de ácidos grasos poliinsaturados o PUFAS, que en el aceite de palma procedente del mesocarpo contiene menos del 10% de acidos grasos poliinsaturados, concretamente de ácido linoleico (C18:2) y linolénico (C18:3), como revisa Montoya y col. (2014). Esto sugiere a que el aumento debido al SNP T podría incrementar también la proporción de alguno de estos ácidos grasos, porque Loei y col. (2013) comprobó en sus estudio el aumento de esta proteína en los frutos caracterizados como de alto rendimiento en *E.guineensis* Jacq. Por otro lado, ninguno de los QTL que se relacionan con ácidos grasos se han posicionado en el mismo grupo de ligamiento, pero en el estudio de asociación de Teh y col.(2016) encontraron que 14 SNPs agrupados en el cromosoma 11 en su mapa de referencia que mostraban asociaciones significativas para el carácter O/DM (Proporción de Aceite frente al peso seco del mesocarpio) relacionado con CPO. **KG261** también es un gen candidato co-localizado en el mismo grupo de ligamiento (GL11) y próximo a un QTL PO3_5 y PO6_9. Ambos QTL explican la variación fenotípica respecto al rendimiento medio de aceite (ton)por palmera por año entre los 3 y 5 años de edad y entre los 6 y los 9 años de edad, lo que coincide con la hipótesis inicial de influencia sobre el carácter, aunque su funcionalidad no ha sido ampliamente estudiada en plantas y sólo indican relación con la proliferación y diferenciación celular, tal y como se discutió en el capítulo 2 de esta tesis.

El polimorfismo encontrado en **KG288** se asoció al carácter número medio de frutos por palmera y por año (**FN)** posicionado en el GL2. Su SNP mostró un bajo DL con otros polimorfismos encontrados en genes candidatos próximos como KG181, KG182 y KG183 (Anexo 11, Figura 11.1a), para los que no se obtuvó asociación. Este gen candidato caracterizado como un *pseudogen TPasa* es un fragmento de un transposón *En/Spm* que ha sido estudiado en palmera de aceite (Kubis y col., 2002), con el objetivo de identificar posibles cambios en este elemento relacionados con el fenotipo "mantled". Este transposón está vinculado a los genes y a su expresión, y la transposición es controlada por la interactuación de mecanismos de autorregulación y epigenéticos (Fedoro, 1999), actúa principalmente durante la meiosis y puede desencadenar cambios cromosómicos que produzcan alteraciones en el fenotipo. Además como postulan Gbadegesin y col. (2008) en su estudio de caracterización de estos elementos en yuca, estos elementos contribuyen a la evolución y a la estructura del genoma de las plantas, y son herramientas muy útiles en la genómica funcional de plantas modelo y en cultivos.

Cuando estos elementos están bien caracterizados pueden utilizarse como marcadores en el análisis de ligamiento para la evaluación de la progenie, (Ramachandran y Sundaresan, 2001; Queen y col., 2004; Grzebelus y col., 2007) y pueden ser de gran utilidad en la caracterización genética y mejora del cultivo.

El **peso medio del fruto** (g) se asoció con 5 SNPs de 3 genes candidatos, de los cuáles dos (KG171 y KG186) de ellos pertenecían a genes co-localizados en diferentes QTLs, uno relacionado con incremento de altura (KG171) y el otro relacionado con el rendimiento del rácimo (Kg/Ha) (KG 186) y posicionados a menos de 1cM de distancia (Capítulo2). El tercer gen candidato se identificó por ser un gen conocido (KG27). El SNP de **KG171** no estuvo en DL con ninguno de los genes candidatos que se situaban en el mismo grupo de ligamiento, posiblemente a consecuencia de la distancia que le separa con el siguiente gen candidato (KG181). Como se discutió en el capítulo 2, su funcionalidad la caracterizó como una enzima metiltransferasa que puede estar relacionada con posibles modificaciones de la pectina durante la biosíntesis de la pared celular, y el desarrollo coordinado de la planta. Los dos SNPs de **KG186** se encontraban en un fuerte DL entre ellos, y además también lo hacían con un valor inferior a 50% con el gen candidato KG187. Ambos genes candidatos tienen funciones similares relacionadas con el metabolismo de lípidos (Capítulo 2, Discusión) y en consecuencia con el peso del fruto, ya que algunos estudios mostraron un mayor contenido de aceite en aquellos racimos con frutos más pesados en cuanto a su contenido de aceite del mesocarpo, y no de la semilla (Rajanaidu y Jalani, 1994; Sharma y Tan, 1997). Por último, los polimorfismos de **KG27**, también se encontraban en un fuerte DL. El gen candidato es conocido por ser un enzima relacionado con la biosíntesis de ácidos grasos de cadena larga (β-Ketoacil ACP sintasa II), concretamente la elongación entre 16:0 ACP y 18:0 ACP, e influye en la composición de los ácidos grasos del aceite de palma procedente del mesocarpo (Pidkowich y col., 2007), de hecho la manipulación genética de esta enzima junto con el silenciamiento de la enzima Palmitoil ACP tioesterasa ayuda incrementar el contenido de ácido oléico (C18:1) a expensas del ácido palmítico (C18:0) (Sambanthamurthi y col., 2009; Barcelos y col., 2015). La asociación de los polimorfismos de los genes candidatos (KG186, KG187 y KG27) implicados en la ruta biosíntética de la síntesis de ácidos grasos confiere fuerza a lo postulado por Sharma y Tan, (1997) y Rajanaidu y Jalani (2004) de la influencia del contenido de aceite y grasa en el sobre el peso del fruto, aunque en esta tesis no se ha estudiado la composición en ácidos grasos del material y no hay ninguna evidencia de la influencia de la composición en ácidos grasos sobre el peso del fruto, pero Montoya y col. (2014) encontraron independencia estadística entre los caracteres de producción entre los que estaba FW y la composición en ácidos grasos.

El último carácter en el que se encontraron asociaciones fue **MF**, y en el que la mayor parte de la variación del carácter se debe a un factor hereditario (Corley y Tinker, 2015.). A este carácter se le asociaron 2 SNPs de 2 genes candidatos, uno de ellos **KG148**, discutido con anterioridad por estar también asociado a CPO. El otro SNP corresponde a un gen conocido (**KG12**5) posicionado en el GL15, el cual no se encontró en DL inter-génico. KG125 es un factor de transcripción *SQUA3* perteneciente a la

subfamilia *SQUAMOSA,* implicado en procesos vegetativos, y no existe evidencia de funciones en los órganos reproductivos ni en el fruto.

Por otro lado, es necesario conocer el tipo de SNP y su posible funcionalidad, ya que aquellos SNPs no sinónimos pueden producir cambios sobre algún aminoácido que altere el codón y cambie la funcionalidad del gen, como se ha discutido para el gen candidato KG148. Este análisis es realizado en la mayoría de los estudios de mapeo por asociación basados en genes candidatos, debido a que el tamaño de los amplicones utilizados es mayor, y se seleccionan diferentes regiones del gen dónde el polimorfismo puede implicar un cambio susceptible en la funcionalidad como las regiones 5'- UTR, 3'- UTR, o regiones codificantes, está última ha sido la utilizada en esta tesis para seleccionar el amplicón (Alencar Figuereido y col., 2010; Fritsche y col., 2012; Kumar y col., 2014; Tello y col., 2015)

En cuanto a los efectos debidos a la asociación los Los resultados mostraron una variación fenotípica entre el 0,8% y el 1,5% en los diferentes caracteres, este rango de explicación de la varianza fenotípica es bajo pero se encuentra dentro de lo esperado, ya que los marcadores SNPs explican sólo una proporción de varianza fenotípica muy pequeña (0,5%- 5%) debido a la complejidad de los caracteres de estudio (Wegrzyn y col., 2010) y posiblemente a su naturaleza poligénica. Por otro lado, es conveniente estudiar los efectos en los genotipos heterocigotos y homocigotos para saber cómo afecta la variación alélica sobre el fenotipo, así como la combinación de los diferentes alelos en los que se ha encontrado asociación, buscando la posibilidad de que puedan actuar conjuntamente sobre el fenotipo, ya que algunos de los genes candidatos comparten funcionalidad, y los caracteres son poligénicos.

La interpretación biológica a partir de las funciones determinadas para los genes candidatos cuyos polimorfismos tuvieron asociación significativa con alguno de los caracteres agronómicos de estudio, y el DL presente en la región intragénica en aquellos genes candidatos con más de un polimorfismo, y en la región inter-génica de algunos de los genes candidatos, hacen de estos genes candidatos herramientas útiles para aplicarlos en una selección asistida por marcadores en los individuos de la población o próximos a ellos. El tipo de población seleccionada hace de estos marcadores también adecuados para incorporarlos por los mejoradores a los índices de selección, como postula Breshegello y Sorrells, (2006). De todas formas, es necesario la confirmación de estos polimorfismos en otras poblaciones de mapeo.

CHAPTER 5: GENERAL DISCUSSION AND FINAL CONCLUSIONS

1. GENERAL DISCUSSION

The African Oil Palm is the most important oil crop around the world as shown by the data in the introduction of this thesis. Its classical breeding history started in 1860, but the molecular breeding began to develop at the end of XX century. Marked assisted selection is a useful tool for this crop because of its long selection cycle; enabling selection at the nursery stage and reducing cost, time and areas of trial plantation. One of these methods is QTL mapping which have proved their utility with the publication of several studies and their enforcements in *E.guineensis* Jacq. (Mayes et al., 1997; Rance et al., 2001; Singh et al.,2009; Billote et al., 2010; Montoya et al., 2013; Lee et al., 2015). QTL mapping is an efficient tool to dissect traits, but it needs a linkage mapping population in order to be applied, which has some limitations as the poor resolution for quantitative traits due to the number of recombination, and two alleles at any given locus can be studied (Flint -Garcia et al., 2003; Gupta et al., 2005). During the two last decades an alternative has been developed in plants, which is association mapping (Thornsberry et al., 2001). This method relies on linkage disequilibrium and shows advantages over traditional linkage mapping such as increased mapping resolution, reduced research time and greater allele number evaluation (Yu and Blucker, 2006; Zhu et al., 2008). The association mapping searches for associations between the phenotype and the genotype, and it can be applied by two approaches: randomly genome-wide and candidate gene driven. This doctoral thesis is the first study on African oil palm where association mapping by candidate gene approach has been applied, and the first in which eight traits and several putative candidate gene has been included related to these traits. So far, candidate gene approach is only applied in one or two agronomic traits and less number of candidate genes. As an example Hill et al. (2012) studied the differences in starch phenotype in *sorghum* using three known genes involved in starch byosinthesis, or three traits related to bunch compactness in grapevine were studied by Tello et al. (2015) using 183 candidate genes.

On the one hand, the **traits** were chosen based on their importance with respect to the **yield** in African oil palm (Chapter 2), being the main goal in crop breeding. That is why they have been studied largely. On the other hand, the **candidate genes** were selected by three different approaches: 1. Transcriptome analysis by BSA cDNA-AFLP methodology, 2. Co-located candidate genes, and 3. Known candidate genes by "in silico" mining. The aim was to identify a greater number of putative molecular functional markers associated with the traits of interest.

BSA cDNA-AFLP is an effective technique to screen differentially expressed genes (Yao et al., 2007), and it allowed the identification of expression patterns from leaves in a set of 242 genotypes. The results were 0,48 TDF/PC, a low number compared to other studies published in different crops as well as in previous assays in the oil palm, Roberdi et al. (2015) and Pattarapimol et al. (2015). The unique origin of the species and the large background from a reciprocal recurrent selection of used genotypes could be the cause of the low level of polymorphisms. That is,because the level of heterozygosity decreases in order to get good agronomical traits through the fixation of positive alleles. The publication of the whole genome sequence of *E.guineensis* Jacq. in 2013a by Singh et al. was a useful tool to

improve these results because it makes possible the selection of primer combinations with more level of transcripts applying a cDNA AFLP "in silico" simulation (Stölting et al.,2009). Finally, the study obtained 56 TDFs related to traits BN, BW, FN, FW, CPO, MF and OM, but the bioinformatics analysis against different database showed only 9 TDF with GO annotations that could be significantly considered as putative candidate genes involved in 6 out of the 8 traits of interest as explained in chapter 2.

The identification of putative candidate genes via **co-location analysis** was a more successful approach, although it needed a hard work by computational analysis. We identified a total of 86 putative positional candidate genes in a high density functional linkage map in which there were integrated the consensus QTL's and anchor molecular markers such as SSR, RFLP and SNP shared between populations in *E.guineensis* Jacq. (Chapter 2). The approach looked for annotated sequences at a close range (±1cM) of interesting QTL's, checking their annotations and specialised literature, for finding the biological sense. This short distance of QTL allowed to ensure a low chance of recombination as Collard et al. (2005) recommended, while other studies of co-location analysis in African oil palm the range is larger than our (± 5cM) (Jeennor and Volkaert, 2014; Lee et al., 2015).

The **comparative genomics** to identify putative candidate genes is not too much useful in complex trait because of the biological differences between species. But when the known genes are from the same species, orthologous genes and/or the sequences have a high degree of conservation, the strategy may be useful as this thesis has demostrated. Also, our study is the first identifying a greater number of putative candidate genes via "in-silico" with biological sense and involved in 8 traits of interest. We identified 119 putative candidate genes related to our agronomic traits, most of them (80) present in *E.guineensis* Jacq., searching homology sequences by BlastX algorithm in order to establish the putative function of the selected sequence. The sum of putative candidate genes identified by the three approaches resulted on 224 genes.

Association mapping requests the **genotyping** a population in order to identify the allelic variants that may be positive in the association with the phenotypic expression. The kind of **molecular markers** selected were **SNPs**. These are friendly markers for use in genotyping a population because they are large and steady in the genome of the plants and can easily adapted to high throughput sequencing techniques (Gupta et al., 2001). In this thesis, the exonic regions of the 198 putative candidate genes were selected to genotype 238 individuals via amplicon sequencing with an Ion Torrent Personal Genome Machine platform through the development of candidate gene libraries by a modified Fusion method (Chapter 3). This method enriches these regions of genome (exonic regions) before sequencing, and DNA from many individuals (238 genotypes) could be analized in the same pool thanks to barcodes. This kind of techniques are preferred for most genomic applications including evolutionary biology, association mapping and biodiversity conservation (Kirkness, 2009; Kilian and Graner, 2012). The process needed to be optimized due to several factors affect the PCR-based enrichment strategies (Cronn et al., 2012), and also, this sequencing platform is usually applied in studies based on the shotgun sequencing of microbial genomes like Egan et al. (2012) showed. Therefore, there are not enough

reference studies for comparison so far. **FastaQ** is the raw data file fasta where it groupes the nucleotide sequences and the quality data, but the raw data could not be processed by the available software. One of the main reason for this is that normally this technique is used in combination randomly sheared DNA and an assembly analysis, which requires to consider quality data. Our approach deals with amplicon and as based on a mayority principle, "the more repetitive, the more ", so that quality data are omitted.. Then, ASPAM was developed by Ritter (unpublished) rely on sequence alignment, and assuming the identical sequences were right.

After the first steps of filter process we could check the preference of emulsion PCR for short amplicons, and increasing the depth of sequencing for this sequences as evidenced by some authors (Cronn et al., 2013; Galindo-González et al., 2015).

Consecutively, the **patterns** were **detected**, considering the combination of SNP's and Indel's in each candidate gene. The filter criterion was not restrictive; therefore the patterns need to be **validated**. The best option it is to analyze the parental population and more number of progenies. In this way our population is transformed in a segregant population which allows us to check the pattern as a locus with Mendelian heredity or as a redundant sequence that the filtered process ignored (Mammadov et al., 2010). This option was not possible, so for checking our results we applied two approaches. The first one was checking the re-sequencing of co-located candidate genes that as we knew they had SNP's markers in their sequences. The 70,6% of these candidate genes showed two or more pattern, as we expected. The second one was checking the patterns of two well known candidate genes in African oil palm. These genes were *Shell* gene (Shing et al., 2013b) and *Virescens* gene (Shing et al., 2014), and both are dominant genes having different patterns depending on the phenotype that the plant shows. As it sees in the discussion of chapter 3, for both genes the alleles were correct, confirming our strategy. Moreover in the on going research work for different populations have been evaluated with this methodology using in part identical CG. Always the sequences were found shearing the same alleles besides some additional new alleles, in part from *E.oleifera,* validating in this way the discriminative algorithms employees (E. Ritter, personnal communication).

With respect to the number of detected polymorphisms, we obtained a less number of polymorphisms than expected in cross-pollinated species assuming they have higher nucleotide diversity than self-pollinated species (Anderson and Lübberstedt, 2003). However, this study is focused on exonic regions, where the SNP's and Indel's are less abundant (Edwards et al., 2007; Patel et al., 2015). The **frequency** of **SNP's** was 1SNP/125bp within the target range expected in plants (1SNP/100-500bp) (Pootakham et al., 2015), but there are marked differences in the frequency of SNP's in the studies of oil palm (Riju et al., 2007; Low et al., 2014; Pootakham et al., 2015). Related to the type of the nucleotide substitutions our results were similar to Low et al's. (2014) results, and the Ts/Tv ratio (1,43) was similar to studies in other plant species such as grapevine or potato, and also in *E.guineensis* Jacq.(Salmaso et al., 2005; Simko et al., 2006; Pootakham et al., 2013, 2015; Ting et al., 2014).

Then, the patterns were associated to genotypes. There were two options to perform the association: 1. SNP match and 2. Association by pattern. In this thesis, it was selected the first option. Since frequently many small random errors with this sequencing thecnology, unlike in sheared DNA, we have few sequences with many small variations and number of reads per CG and genotype varies considerably probably to annealing preference and DNA quality. Although less restrictive than the second, the results previously showed by Santika (2015, unpublished) did not show a great differences between them. Also, the SNP match option avoided the loss of genotypes with small changes because they did not match in an exact way. In addition, checking GT with allele assignment allows to identify genotypes with excess of alleles according to their ploidity, in this case diploid. The presence of genotypes with more than two alleles after checking their patterns could be determined by candidate genes with duplicate loci or these candidate genes belong to gene families with high degree of conservation. In this case the feasibility of these markers could obstruct future applications as association mapping (Patel et al., 2015), and for this reason these candidate genes were excluded from this study, for instance, KG162 and KG217, as it was explained in the discussion in chapter 3.

The **genotypic frequencies** showed high level of heterozygosity as it was expected in cross-pollinated species (Hanley et al., 2002). This argument is supported by the kind of selected population. Our genotypes are a set of *Tenera* palms coming from different breeding selections. Their male parental (*Pisifera*) have different origins and they may be reliable of variability. Furthermore, the genotype frequencies proved the methodology of sequence comparison between genotypes of African oil palm is suitable for SNP detection due to greater heterozygosity found in each genotype (Poothakam et al. 2013).

The study of **genetic diversity** is an important step for plant breeding and the creation of new plant varieties (FAO, 1996). This thesis contains a study of genetic diversity with the main aim of knowing how the population has selected (Chapter 3). The results showed a reasonably **PIC** (0,306), similar to obtained by Poothakam et al. (2013) and Ong et al. (2015) in oil palm. The high level of heterozygosity was confirmed by a greater value of Ho than the value of He, and as Arias et al. (2014) advocated the main reison could be the different origins of the parents in the population. For example, some Tenera male parent (*Pisifera*) are from Nigeria, Yagambi, or Ghana, among others. The **Hardy-Weinberg equilibrium** infringed in 55 out of 64 loci and the reason could be the non random mating of populations. As we know the genotypes were submitte to a reciprocal recurrent selection and this is one of possible reason for deviations from HW equilibrium (Abdurakhmonov and Abdukarimov, 2008).

The **Wright statistics** or **F statistics** (Wright, 1951) allows to estimate the proportion of genetic variation found within and between populations by F statisitcs (Fis, Fit, Fst). In our study the population set was divided in *Pisifera* parental origin in order to calculate these statistics and elucidate if the population structure was present in the genotype set. The results demonstrated a high degree of heterozygosity between families and within the comprehensive set of population and a low level of genetic drift. Thus, there are first signs of unstructured populations and this was confirmed by the **gene**

flow statistic (Nm). Furthermore, two dendrograms based on genetic distances were made. In the first one, the population was clustered by families in 3 groups (Chapter 3, Figure 28). and most of families clustered in the third arm in a mixed way. On the other hand, the genotypes in the second dendogram were grouped by the origin of their male parental. In this case, there were also 3 clusters, but there was a narrow relationship between different parents. For instance, Avros and Dami are from Democratic Republic of the Congo developed all together (Corley and Tinker, 2003).However, Yangambi which is from the same origin, evolved close together to Nigeria, due to the breeding process. Finally, the **AMOVA** confirmed the results, because the 99% of the total genetic variation is due to the differentiation within the population themselves, and only the 1% was attributed to the variation between populations. In this sense, the population sampled acts like a panmictic unit, which means that the individuals do not show a tendency to choose partners with particular trait. Perennial cross-pollinated species with a large life cycle show more genetic variation within the population, such as oil palm (Ong et al.,2015). In addition, when association mapping is one the aims of the study, it is necessary to control the population structure, since the presence of population stratification and an unequal distribution of alleles within these groups can result in spurious associations (Knowler et al.,1988; Flint-García et al., 2003). For this reason, in chapter 4 the population structure was analyzed by two different methods: PCA and Structure. Although the results were inconclusive they were in agreement with the AMOVA result of 1% of the variation occurring between populations.This is supported by Odong et al. (2011) when they predicted that populations with low differentiation levels will have an optimum of two clusters. This fact was expected and observed in a synthetic population like ours.

Another relevant step in association mapping is the **phenotyping** of the population, because the power of the method is strongly dependent on the quality of phenotypic data (Rafalski, 2010). Then the number of plants is also relevant in order to obtain more power in the association studies (Ingvarsson and Street, 2011). Our study was formed by a population of 238 individual plants, phenotyped during 15 years and they were selected for opposite field performance (positive and negative for each trait) with the aim to increase the variability (Chapter 4). With respect to trait correlations the results were similar those of other authors such as Billote et al.(2010) or Montoya et al. (2014) suggesting that agronomical traits such as BN or BW are intrinsic within *E. guineensis* Jacq., although it will need a greater number of studies to ensure this point.

Another point to note is the **linkage disequilibrium** which is the conceptual basis of association mapping (Flint-Garcia et al., 2003). In a study relying on candidate gene approach it is necessary to meet intra-locus LD, and if the combination of the intra-locus SNPs when they are in LD correspond to a haplotype. Our results showed that all candidate gene with two or more SNPs within the candidate gene were in high LD and each of them formed an haplotype, confirming the patterns detected in the population (Chapter 3). We also observed a slowly decrease of LD along the genome, suggesting LD is maintained in high distances. Probably, this is due to the reciprocal recurrent selection that our

population has suffered for some years as Palaisa et al. (2003) defends that the intense selection may build up LD regions because it facilitates the good allelic combination and/or promotes the genetic drift.

The **genotype-phenotype association** was done by a **GLM** approach. This method showed for our data the best performance in the QQ plots and this confirmed our low or none level of population stratification. Fifteen SNP's from eleven candidate genes were associated to 6 traits of interest, half of them related to yield (BN, CPO, and OWM) and the rest related to bunch components (FN, FW, and MF). About half of candidate genes belonged to co-located candidate genes (KG135, KG194, KG201, KG171 and KG186) and the rest of them belonged to known candidate genes (KG27, KG125, KG148, KG254, KG288 and P64). However, none of the candidate genes from TDF showed associations. This establishes the identification by co-located and *in silico* candidate genes as the best approach to look for putative candidate genes. The transcriptome strategy could have failed, among other causes, because of the selection of leaves as the tissue template when the main traits are related to the fruit or the bunch. So, we could improve this approach using the tissue from these parts of the palms. The likelihood of identifying transcripts linked to the trait will increase if we obtain RNA from this kind of tissue because some genes are tissue-specific. On the other hand, in some cases, the co-located candidate genes that showed polymorphisms in their amplicons were associated indirectly with other traits. For instance, KG194 was co-located with FFB (kg/palm/year) QTL in 11 LG, but the association analysis showed that it was linked to CPO trait. The oil yield of the oil palm may be regarded as a sum of characteristic which its final expression depends on the number of different components, fresh fruit bunch yield, bunch and quality traits (Sparnaaij et al., 1963). Also, CPO trait is correlated with BN (Chapter 3, Results) which in turn is a component of FFB as Jeenor and Volkaert (2012) explained. So, these results could reveal that the natural variation for this trait (CPO) could be controlled by genes involved in related traits, and they could be interdependent. Direct associations have also been found in the same trait. KG261 co-located near PO QTL in the LG could be one of the gene controlling the QTL of interest, but it is necessary to confirm its biological function.. Some *in silico* known candidate genes were close to QTLs in the map (Chapter 2; Ritter, personal comunication). This candidate genes are: 1. KG288 (LG2) is co-located with ±5cM of two QTL's, BN and FFB, and it associated to FN, 2. KG27 (LG10) is co-located near BN and their association was FW and 3. KG148 (LG3) is co-located close to two QTL's, PO and FN. The last one is the most interesting gene because KG148 is associated to MF and OWM. Both traits are correlated in a positive way (Chapter 3, Results) but the allele associated with each trait is different. The allele T had a positive effect on MF whereas the allele C had a positive effect on OWM. Furthemore, two QTL's were positioned close to them, PO and FN. The trait FN was correlated with MF in a negative way. Then the allele T could have a good effect on MF but not on FN. In addition, OWM and PO are related traits because the percentage of oil wet mesocarp is part of the total amount of palm oil. So if the allele T would be present in the genotype the consequent phenotype could have a negative effect on both traits. As we discuss in chapter 3, how this gene could affect the traits will be an important question to rsolve in the future.

The knowledge contribution of this thesis with respect to molecular breeding in oil palm consist of a large list of putative candidate genes related to agronomical traits. Even though some of 224 putative candidate genes did not show any polymorphism, the GO annotation and the literature revision evidence their putative relation with the traits (Chapter 2). In addition, the set of SNP's markers detected could be integrated in different linkage maps, helping to saturate the map with this kind of markers. They can be applied in diversity screening studies will provide information for the breeder about the strategy for collecting samples for genetic conservation or in molecular characterization of new germplasm by allele mining (Kilian and Graner, 2012). On the other hand, the candidate genes strongly associated with the phenotypic traits like *Shell Thickness* or *Virescens* characteristics could be use for marker assisted selection by allele specific probe approaches (Ritter et al., 2015). The main and very important advantage is the early selection of genotypes (even at nursery stage), saving time, resources and planting area increasing in tihis way sustainable oil palm cultivation.

2. FINAL CONCLUSIONS

This section shows the main conclusions as a result of the study and agree objective in each developed chapter under the assumption: "Association mapping by candidate gene approach allows to find functional genetic markers relate to phenotypic values in a population of oil palm (*E.guinnensis* Jacq.)".

1. A total of 224 putative candidate genes were selected by three different strategies: 1. Transcriptome analysis by BSA cDNA AFLP, 2. Co-location of candidate genes with QTL's using an integrated genetic map, and 3. In silico search through functional and comparative genomics.

2. Positional and functional approaches were more fruitful than transcriptome analysis. We obtained more number of candidate genes (86 co-located candidate genes and 119 known candidate genes) with these two approach rather than by transcriptome analysis (9 apparent putative candidate genes). In addition, the costs and the time of the study is reduced when we apply bioinformatics techniques.

3. Ion Torrent PGM was a useful method to apply high-throughput amplicon sequencing strategy in oil palm, but the process to create the libraries affected the number of sequencing reads. It was necessary to optimize the process as the experiments progressed and when the size of the amplicons sre simila, the sequencing sequencing will be more efficient and homogeneous. A total of 198 putative candidate genes were sequenced. The average raw reads obtained were 6.390.106 ± 1.301.556 and 1,28 ± 0,38 Mpb were sequenced per library.

5. The ASPAM software was useful to identify the patterns in our population, although it took time to analyze the data. After the filtering process from the set of initial patterns in the population, 12% of the candidate genes did not show any consistent pattern, 31% of them were monomorphic, 38% had two patterns and 19% had three or more patterns.

6. We obtained 1SNP/125pb within the range expected in plants (1SNP/100pb-1SNP/500pb) and the ratio Tr:Tv was 1,43 such as in other studies of oil palm.

7. After checking genotypes and allelic composition, 65 candidate genes were bi-allelic and 34 candidate genes were classified as "multi-locus". This were excluded of the association studies.

8. The genetic diversity study showed Ho>He. That is a high ratio of heterozygote genotypes due to the "hybrid" character of the *Tenera* population and the origin of their parents. Only 9 candidate genes were in HW equilibrium, 7 of them deviated slightly from HW equilibrium, and 48 of them deviated strongly of HW equilibrium.

9. The results of Wright statistics and AMOVA reflected that the genetic variation is occurring between population themselves, and there is not any genetic drift.

10. The LD intra-locus was strong, confirming the candidate genes as haplotypes. The LD decrease was slow along the genome, withouth getting a value of r^2 at 140cM. As a consequence the likelihood of recombination was low at that distance.

11. The level of stratification in the population is low or null as expected.

12. The GLM statistical method was the best method in order to associate molecular data with phenotypic data.

13. Finally we found 11 candidate genes associated with 3 traits related to production and with 3 traits for bunch component. The allelic effect on the phenotypic variance explained less than 5% as expected for complex and polygenic traits.

In summary, this thesis is the first study about association mapping using the candidate gene approach in African oil palm. The results presented and the candidate genes studied should be validated in a deeper way. In the future these markers could be used to apply molecular assisted selection in the crop. Furthemore, interesting putative candidate genes could be widely studied by selecting different regions of them in order to find new polymorphisms that could be associated to production and yield traits.

ANEXOS

Anexo 1: Análisis del transcriptoma

Tabla 1.1: Familias seleccionadas para la detección de genes candidato mediante BSA cDNA-AFLP. Parental femenino: CH= Chemara; D= Dami; HC= Harrison - Crossfield; M= Mardi. Parental masculino (Pisifera): A= Avros; Da= Dami Komposit; E= Ekona; G= Ghana; LM= LaMé; N= Nigeria; Y= Yangambi. BN=número medio de racimos/palmera/año; BW=peso medio de racimo/palmera/año (kg); CPO= rendimiento de aceite (ton/ha/año); FN= número medio de fruto/palmera/año; FW= peso medio de fruto/palmera/año (g); HT=incremento de altura de tallo (cm); MF= ratio tamaño de mesocarpo vs tamaño de fruto (%); OM= ratio de aceite de mesocarpo vs tamaño de mesocarpo (%).

FAMILIA	CRUCE	CARÁCTER
552	MxN	BN
698	(CH x HC)xY	BN
718	DxDa	BN
773	DxN	BN
519	CHxDa	BW
522	DxE	BW +FN
711	DxA	BW
804	(CH x HC)xD	BW
476	CHxN	CPO
505	CHxG	CPO + HT
580	HCxA	CPO
864	HCxG	CPO
422	CHxE	FN
625	CHxLM	FN
651	DxDa	FN
670	(CH x HC)xY	FW
731	DxY	FW
756	DxA	FW
818	HCxA	FW
658	DxE	HT
831	DxG	HT
837	DxA	HT
440	DxDa	MF
578	HCxE	MF
681	DxN	MF
840	HCxDa	MF
626	HCxDa	OM
776	DxDa	OM

Tabla 1.2:Fragmentos del transcriptoma polimórficos obtenidos mediante cDNA-AFLP en cada carácter. El tamaño en pares de bases (pb), la combinación de cebadores polimórfica (CC), el nombre de la familia a partir del cual se realizó la mezcla de genotipos, siendo B los clasificados como buenos y M los clasificados como malos, y por último en que clase se enmarcaron (positivos, +, positivos/negativos,±, y otros). La columna nombre de la secuencia muestra cómo se nombró el amplicón obtenido.

					BN					**CLASE**	**NOMBRE SECUENCIA**
TAMAÑO(pb)	**CC**	**519B**	**522B**	**711B**	**804B**	**519M**	**522M**	**711M**	**804M**		
255	A16/T13	1	1	1	0	0	0	0	0	±	
160	A16/T13	1	1	1	0	0	0	0	0	±	
185	A16/T14	1	1	0	0	0	0	0	0	Otros	CDA8
190	A16/T14	1	1	1	0	0	0	0	0	±	CDA43
430	A19/T12	1	1	1	1	0	0	0	0	+	
					BW					**CLASE**	**NOMBRE SECUENCIA**
TAMAÑO(pb)	**CC**	**519B**	**522B**	**711B**	**804B**	**519M**	**522M**	**711M**	**804M**		
400	A12/T11	0	0	0	0	1	1	1	0	±	
420	A12/T11	0	0	0	0	1	1	1	1	+	
340	A+G/T11	0	0	0	0	1	1	1	1	+	CDA20
490	A+G/T12	0	1	1	1	0	0	0	0	±	CDA2
480	A+G/T13	1	1	1	0	0	0	0	0	±	CDA6
540	A+G/T13	0	0	0	0	0	0	1	1	Otros	
550	A+G/T13	1	1	0	0	0	0	0	0	Otros	
495	A+G/T14	0	0	0	0	0	1	1	1	±	CDA7
140	A19/T12	0	0	0	0	1	1	1	1	+	CDA13
260	A21/T11	1	0	1	1	0	0	0	0	±	
230	A19/T13	1	1	1	1	0	0	0	0	+	
160	A19/T13	0	0	0	0	1	1	1	1	+	
220	A19/T13	0	0	0	0	1	1	1	1	+	
					CPO					**CLASE**	**NOMBRE SECUENCIA**
TAMAÑO(pb)	**CC**	**476B**	**505B**	**580B**	**864B**	**476M**	**505M**	**580M**	**864M**		
350	A12/T14	0	0	0	0	1	1	1	1	+	CDA19
160	A12/T14	0	0	0	0	1	1	1	1	±	
155	A12/T14	0	0	0	0	0	1	1	1	±	CDA10
350	A12/T14	0	0	0	0	1	1	1	1	±	
220	A14/T11	0	0	0	0	0	1	1	1	±	CDA17
300	A14/T12	0	0	0	0	1	1	1	1	±	
410	A+G/T12	0	0	0	0	0	0	1	1	Otros	CDA3
400	A+G/T12	0	0	1	0	0	1	1	0	Otros	CDA4
300	A+G/T13	0	0	0	0	1	0	1	1	±	CDA16
500	A+G/T13	0	0	0	0	0	0	1	1	±	
420	A+G/T14	0	0	0	0	1	1	0	0	Otros	CDA5
440	A+GT/14	0	0	0	0	1	1	0	0	Otros	CDA42
340	A19/T11	2	0	0	2	1	1	1	2	±	
150	A19/T11	1	1	1	0	0	0	0	2	±	CDA12
350	A19/T13	0	0	0	0	0	0	1	1	Otros	CDA44
270	A19/T13	0	0	0	0	1	1	1	0	±	CDA9

230	A19/T13	0	0	0	0	0	0	1	1	Otros	CDA11

Tabla 1.2: Continuación

				FW						CLASE	NOMBRE SECUENCIA
TAMAÑO(pb)	**CC**	**670B**	**731B**	**756B**	**818B**	**670M**	**731M**	**756M**	**818M**		
200	A12/T13	1	1	1	1	0	0	0	0	+	CDA32
350	A12/T14	1	1	1	1	2	2	0	0	±	CDA22
270	A16/T13	1	0	1	1	0	0	0	0	±	CDA29
340	A16/T13	1	1	1	1	0	0	0	0	+	
305	A16/T15	0	1	1	1	0	0	1	0	Otros	
145	A17/T13	1	1	1	0	0	0	0	0	±	
220	A17/T14	1	1	1	1	0	0	0	0	+	CDA23
240	A17/T15	0	1	0	1	0	0	0	0	Otros	CDA24
145	A17/T24	1	1	1	0	0	0	0	0	±	CDA21
440	A18/T15	0	1	1	1	0	0	0	0	±	CDA30
470	A19/T12	1	1	1	0	0	0	0	0	±	
180	A19/T13	1	1	1	1	0	0	0	0	+	CDA26
210	A19/T18	0	0	0	0	1	1	1	0	±	CDA25
300	A19/T18	0	0	0	0	1	1	1	2	±	CDA31
465	A23/T13	0	0	0	1	1	1	1	2	+	
				MF						**CLASE**	**NOMBRE SECUENCIA**
TAMAÑO(pb)	**CC**	**440B**	**578B**	**681B**	**840B**	**440M**	**578M**	**681M**	**840M**		
220	A17/T24	1	1	1	1	0	0	0	0	+	
325	A18/T11	1	1	1	1	0	0	0	0	+	CDA37
205	A19/T13	1	1	1	1	0	0	0	0	+	CDA34
350	A67/T72	1	1	1	1	0	0	0	0	+	
200	A67/T72	1	1	1	1	0	0	0	0	+	
360	A67/T73	1	1	1	1	0	0	0	0	+	
255	A69/T69	1	1	1	1	0	0	0	0	+	CDA35
130	A69/T70	1	1	1	1	0	0	0	0	+	
135	A69/T72	1	1	0	1	0	0	0	0	±	
135	A71/T72	0	0	0	0	1	1	1	1	+	
245	A71/T74	0	0	1	1	0	0	0	0	Otros	CDA36
220	A70/T69	1	0	1	1	0	0	0	0	±	

Tabla 1.2: Continuación

				HT						CLASE	NOMBRE SECUENCIA
TAMAÑO(pb)	**CC**	**505B**	**658B**	**831B**	**837B**	**505M**	**658M**	**831M**	**837M**		
140	A17/T19	1	1	1	1	0	0	0	0	+	CDA39
225	A17/T19	1	1	1	1	0	0	0	0	+	
160	A16/T11	0	0	0	0	1	1	1	1	+	
185	A16/T14	0	0	0	0	1	1	1	1	+	
265	A16/T15	0	0	0	0	1	1	1	1	+	CDA41
305	A16/T17	1	1	1	1	0	0	0	0	+	
200	A16/T19	0	0	0	0	1	1	1	1	+	
185	A16/T20	0	0	0	0	1	1	1	1	+	
245	A16/T20	0	0	0	0	1	1	1	1	+	
300	A16/T24	0	0	0	0	1	1	1	1	+	CDA40
275	A16/T24	0	0	0	0	1	1	1	1	+	
170	A18/T18	1	1	1	1	0	0	0	0	+	
145	A18/T19	1	1	1	1	0	0	0	0	+	
145	A14/T13	1	1	1	1	0	0	0	0	+	
135	A21/T14	0	0	0	0	1	1	1	1	+	
105	A21/T15	1	1	1	1	0	0	0	0	+	
270	A21/T16	0	0	0	0	1	1	1	1	+	
				OM						**CLASE**	**NOMBRE SECUENCIA**
TAMAÑO(pb)	**CC**	**557B**	**626B**	**776B**	**855B**	**557M**	**626M**	**776M**	**855M**		
190	A17/T19	0	0	1	1	0	0	0	0	Otros	CDA74
200	A23/T12	0	0	0	0	0	1	1	1	±	CDA75
440	A17/T24	1	1	1	0	0	0	0	0	±	CDA77
490	A19/T24	0	0	0	0	1	0	1	1	±	CDA76
350	A12/T20	1	1	1	0	0	0	0	0	±	CDA78
380	A12/T12	0	0	0	0	0	0	1	1	Otros	CDA79
255	A52/T56	0	0	0	0	1	1	0	1	±	
206	A56/T54	0	0	0	0	1	1	1	0	±	
200	A58/T55	1	1	0	1	0	0	0	0	±	
132	A57/T57	0	1	1	1	0	0	0	0	±	
460	A19+C/T20	1	1	1	1	0	0	0	0	+	
350	A21+C/T20	0	0	0	0	1	1	1	1	+	

Tabla 1.2: Continuación.

		FN								CLASE	NOMBRE SECUENCIA
TAMAÑO(pb)	CC	422B	522B	625B	651B	422M	522M	625M	651M		
158	A12/T11	0	0	0	0	0	1	1	1	±	CDA85
140	A17/T11	0	1	1	1	0	0	0	0	±	CDA89
130	A16/T13	0	1	1	1	0	0	0	0	±	CDA87
118	A16T13	1	1	1	0	0	0	0	0	+	
270	A18/T13	0	0	0	0	1	1	1	1	+	
205	A23/T16	1	1	1	1	0	0	0	0	+	
198	A12/T17	1	1	0	1	0	0	0	1	±	CDA84
215	A16/T18	0	0	0	0	1	1	1	0	±	
100	A56/T52	1	0	1	1	0	0	0	0	±	
200	A56/T52	0	0	0	0	0	1	1	0	Otros	CDA86
202	A56/T54	0	0	0	0	0	1	1	0	Otros	
157	A18/T20	0	0	0	0	0	1	1	1	±	CDA88
120	A17/T12	1	0	1	1	0	0	0	0	±	CDA90
156	A23/T20	0	0	0	0	0	0	1	1	Otros	CDA91
245	A19/T22	0	1	1	0	0	0	0	0	Otros	CDA92

Tabla 1.3: Número de polimorfismos observados en función de CC aplicada para el análisis de cDNA-AFLP.

Nº	CC	Nº TDF	Carácter	Total TDF	Nº	CC	Nº TDF	Carácter	Total TDF
1	A16/T13	2	BN	6	**25**	A67/T72	2	MF	2
		2	FW		**26**	A67/T73	1	MF	1
		2	FN		**27**	A69/T70	1	MF	1
2	A16/T14	2	BN	3	**28**	A69/T72	1	MF	1
		1	HT		**29**	A71/T72	1	MF	1
3	A19/T12	1	BN	3	**30**	A71/T74	1	MF	1
		1	BW		**31**	A70/T69	1	MF	1
		1	FW		**32**	A16/T11	1	HT	1
4	A12/T11	2	BW	3	**33**	A16/T17	1	HT	1
		1	FN		**34**	A16/T19	1	HT	1
5	A+G /T11	1	BW	1	**35**	A16/T20	2	HT	2
6	A+G /T13	4	BW	6	**36**	A16/T24	2	HT	2
		2	CPO		**37**	A18/T18	1	HT	1
7	A+G /T14	1	BW	3	**38**	A18/T19	1	HT	1
		2	CPO		**39**	A14/T13	1	HT	1
8	A21/T11	1	BW	1	**40**	A21/T14	1	HT	1
9	A19/T13	3	BW	8	**41**	A21/T15	1	HT	1
		3	CPO		**42**	A21/T16	1	HT	1
		1	FW		**43**	A23/T12	1	OM	1
		1	MF		**44**	A19/T24	1	OM	1
10	A12/T14	4	CPO	5	**45**	A12/T20	1	OM	1
		1	FW		**46**	A12/T12	1	OM	1
11	A14/T11	1	CPO	1	**47**	A52/T56	1	OM	1
12	A14/T12	1	CPO	1	**48**	A56/T54	1	OM	1
13	A+G/T12	2	CPO	2	**49**	A58/T55	1	OM	1
14	A19/T11	2	CPO	2	**50**	A57/T57	1	OM	1
15	A12/T13	1	FW	1	**51**	A19+C/T20	1	OM	1
16	A16/T15	1	FW	2	**52**	A21+C/T20	1	OM	1
		1	HT		**53**	A17/T11	1	FN	1
17	A17/T13	1	FW	1	**54**	A18/T13	1	FN	1
18	A17/T14	1	FW	1	**55**	A23/T16	1	FN	1
19	A17/T15	1	FW	1	**56**	A12/T17	1	FN	1
20	A17/T24	1	FW	3	**57**	A16/T18	1	FN	1
		1	MF		**58**	A56/T52	2	FN	2
		1	OM		**59**	A56/T54	1	FN	1
21	A18/T15	1	FW	1	**60**	A18/T20	1	FN	1
22	A19/T18	2	FW	2	**61**	A17/T12	1	FN	1
23	A23/T13	1	FW	1	**62**	A23/T20	1	FN	1
24	A18/T11	1	MF	1	**63**	A19/T22	1	FN	1

Anexo 2: Resultados de los alineamientos comparativos entre los transcriptos obtenidos (TDF) y las diferentes bases de datos analizadas. Los resultados se muestran por caracteres de interés junto con la calificación de las familas cómo buenas o malas. Leyenda: TDF: Nombre del transcripto; Carácter: el carácter para el que se ha fenotipado la familia; Familia: Calificación de la familia y sus transcriptos."B"= calificada como buena, esto es, fenotipo positivo para ese carácter, "M"= calificada como mala o fenotipo negativo para el carácter; "+"= TDF presente en todas las familias; "±"= TDF presente en todas las familias menos en una; "OTROS"= TDF presente en dos familias o menos.; GL= grupo de ligamiento o cromosoma dónde se mapeó; E-valor= valor resultante de la homología para el grupo de ligamiento, para la librerías o en el análisis con B2GO; Librería NR= anotación mostrada como resultado de la búsqueda de homologías en las librerías de secuencias no redundantes en las bases de datos locales; Librería EST's Públicas (NCBI) = anotación mostrada como resultado de la búsqueda de homologías en las librería de EST's en la página web de NCBI mediante Blastn y acotando con el género Elaeis; BLAST2GO = muestra los resultados arrojados por el software B2GO en cuanto a las homologías "BLASTx" o "BLASTn" y la anotación ontológica obtenida "Anotación GO".

Tabla 2.1: Resultados para el carácter BN.

TDF	CARÁCTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PÚBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E- Valor	ANOTACION "GO"
CDA8	BN	B OTROS	12	8e-62	-		gi|161973648|gb|EL685771.1|EL685771| @---NA--- @ OPMF02012 Elaeis guineensis mature flower 26 cm Elaeis guineensis cDNA clone OPN6SG_S692.seq T3 - mRNA sequence		-	-		-
CDA43	BN	B ±	12	8e-62	-							

Tabla 2.2: Resultados para el carácter BW.

TDF	CARÁCTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PÚBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E-Valor	ANOTACION "GO"
CDA6	BW	B±	-		-		-		-	*Rhabdodendron amazonicum* Secuencia parcial ARN ribosómico 26s	2.2E-55	-
CDA7	BW	M±	14	4e-144	-		-		-	-		-
CDA13	BW	M+	-		-		-		Parcial péptido no ribosómico		1.6e-18	F:actividad catalítica; F:unión
CDA15	BW	B±	-		-		-		-	*Phoenix dactylifera* genoma completo	2.6E-95	-
CDA20	BW	M+	6	0.0	-		-		-			-
CDA27	BW	B+	-		CL1Contig5969| @ "(at3g22142 : 117.0)Codifica una proteasa inhibidora/ almacenamiento en semilla de la familia LTP	4E-59	-		Proteína no caracterizada LOC101511867		3.1e-14	-

Tabla 2.3: Resultados para el carácter FN

TDF	CARACTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PUBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E-Valor	ANOTACION "GO"
CDA14	FN	M±	-		-		-		-	*Phoenix dactylifera* RNAm variante transcripta no caracterizada loc103704827	1.7E-28	-
CDA90	FN	B±	-		Contig2857_S_C2f	2e-134	gi|133925368|gb|EL930609.1|EL930609| @subunidad beta atp sintasa @ EoEST1446 Oil palm mesocarp-tissue cDNA Entry Library *Elaeis oleifera* cDNA clone Mo17-1497 5' similar to Unknown protein - mRNA sequence	1E-144	Subunidad beta de ATP sintasa		5.8E-27	-
CDA92	FN	B OTROS	1	2e-95	-		-		-	-		-

Tabla 2.4: Resultados para el carácter FW

TDF	CARACTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PUBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E-Valor	ANOTACION "GO"
CDA22	FW	B+	-		-		-		Isoforma 1 helicasa 39 ARN"Dead-Box" ATP dependiente	-	2.1e-12	F:unión de ácidos nucléicos; F:unión de nucleótidos; F:actividad hidrolasa
CDA23	FW	B+	-		-		-		AF527536_1 26kDa floema		5,2e-05	F: unión de carbohidratos
CDA24	FW	B OTROS	4	2e-71	Eg_Deli_CL35037Contig1	2e-57	-		-	-		-
CDA26	FW	B+	-		Contig2793_S_C2f	5e-100	M01000001578| @proteína precursora 8 relacionada con la autofagia@ AT4G21980	3e-127	proteína hipotética (mitocondria)	-	5.8E-32	C:mitocondria
CDA31	FW	M±	2	5e-67	CL1Contig1267| @ "(p33278|sui1_orysa : 215.0) Homólogo a factor de traduccion de proteína SUI1 (*Oryza sativa*)	9e-61	-		proteína hipotética JCGZ_00471		2.7E-29	C:plastidio
CDA32	FW	B+	12	7e-64	-		-		-	-		-

Tabla 2.5: Resultados para el carácter CPO

TDF	CARACTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PUBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E-Valor	ANOTACION "GO"
CDA3	CPO	M OTROS	1	0.0	-		-		-	-		-
CDA4	CPO	M OTROS	5	7e-124	-		-		Posible metiltransferasa pmt11		1.2E-23	F:actividad transferasa;P: proceso metabólico; C: endosoma; C:aparato de Golgi
CDA5	CPO	M OTROS	4	4E-62	-		-		dimetialilo triptófano		2.1E-20	P:proceso de metabolismo de proteínas; P:proceso biosintético; F:actividad transferasa; P:proceso celular
CDA9	CPO	M±	11	2E-71	-		-		-	-		-
CDA18	CPO	M+	7	3e-71	-				Similar a receptor de proteína kinasa (Rica en cisteína-rlk 8)		2.1E-20	P:proceso metabólico; F:unión; P:proceso celular
CDA42	CPO	M OTROS	2	2e-62	-		-		-	-		-
CDA44	CPO	M OTROS	12	4e-67	gi|191204475|gb|EY407406.1|EY407406| NR @ "(loc_os12g27096.1 : 144.0) Relacionado con un poliproteína POL de transposon TNT 1-94 (*Nicotiana tabacum*)	3e-105	gi|191204475|gb|EY407406.1|EY407406| @retrotransposon ty1-copia subclass @ pOP-EAP01683_EST_C_1_pBSK_SK EA (Oil Palm Embryoid) *Elaeis guineensis* cDNA clone pOP-EAP01683 5' - mRNA sequence	1e-107	Proteína de fusión "gag-pol"		6.3e-13	F:unión

Tabla 2.6: Resultados para el carácter MF.

TDF	CARACTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PUBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E-Valor	ANOTACION "GO"
CDA33	MF		6	5e-151	-		M01000011073| @hypothetical protein OsI_06902 [*Oryza sativa* Indica Group] @ AT2G15420	2E-113	-	-		-
CDA34	MF	B+	-		-		-		Proteína no caracterizada	-	8.3e-14	-
CDA37	MF	B+	3	4e-79	-		-		Quinasa f-1 dependiente de ciclina		1.7e-12	P:desarrollo post-embrionario; P:modificación de proteínas celulares process; P:morfogenésis de estructura anatómica; P:proceso biosintético; P: proceso metabólico ADN; P:ciclo celular; F:unión de nucleótidos; C:núcleo; C:citosol; F:actividad kinasa; F:actividad como enzima regulador

Tabla 2.7: Resultados para el carácter OM.

TDF	CARACTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PUBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E-Valor	ANOTACION "GO"
CDA76	OM	M±	5	7e-151	CL1Contig209| @ "(loc_os09g28810.1 : 437.0) (at1g47500 : 363.0) Proteína unión RNA 47 (RBP47C)	7e-150	gi|112819434|gb|EE593335.1|EE593335| @rrna intron-encoded homing endonuclease @ EgTEST0049 *Elaeis guineensis* Lambda ZAP-II Express Library *Elaeis guineensis* cDNA clone TE-144 5' similar to Unknown protein - mRNA sequence	0.0	Proteína asociada a la senecencia	-	3.6E-45	-
CDA78	OM	B±	16	3e-78	Contig3707_S_EoP3-P2	6e-67	-		-	*Phoenix dactylifera* Variante transcripta no caracterizada loc103718754	2.8E-53	-
CDA79	OM	M OTROS	-		-		-		-	*Oryza punctata* ARN ribosómico 18s, transcrito ARN interno ribosómico espaciador y secuencia completa ARN ribosómico 26s	7.6E-28	-

Tabla 2.8: Resultados para el carácter HT.

TDF	CARACTER	FAMILIA	GL	E-valor	LIBRERIA NR	E-valor	LIBRERIA EST'S PUBLICAS (NCBI)	E-valor	BLAST2GO			
									BLASTX	BLASTN	E- Valor	ANOTACION "GO"
CDA39	HT	B+	-		-		-		-	*Ttriphyophyllum peltatum* Secuencia completa ARN ribosómico 26s	6.4E-18	-
CDA40	HT	M+	11	2e-73	--		-		-	-		-
CDA41	HT	M+	14	2e-59	-		-		-	-		-

Anexo 3: Resultados de la búsqueda de genes candidato co-localizados en el mapa integrado. ID QTL: indica el nombre del QTL dónde se encuentra co-localizada la secuencia; Posición QTL: GL es el grupo de ligamiento dónde esta el QTL, cM en qué centimorgan (cM) y en qué número de pares de bases está; Locus Candidato: es la secuencia seleccionada como posible gen candidato; Longitud en pb es el número de pares de bases que tiene; Posición LC candidato: cM indica la distancia a la que se encuentra el Lc candidato del QTL, Inicio en qué pares de bases empieza y Fin en qué pares de base termina; Distancia: muestra la distancia en pares de bases del LC candidato con respecto al QTL, si es negativo se encuentra antes y si es positivo después.

Tabla 3.1: Caracteres relacionados con la producción.BN= Número medio de racimos por palmera/año; QBn3_5= QTL Bn en palmera entre 3 y 5 años de edad; QBn6_9= QTL Bn en palmera entre 6 y 9 años de edad; Bw= Peso medio de racimos/palmera/año (kg/palmera/año); QBw3_5= QTL Bw en palmeras entre 3 y 5 años de edad/año; QBw6_9= QTL Bw en palmeras entre 6 y 9 años de edad/año; FFB= Rendimiento medio de frutos frecos por racimo/ palmera/año (kg/palmera/año); QFF3_5= QTL FFB en palmeras entre 3 y 5 años de edad); QFF6_9= QTL FFB en palmeras entre 6 y 9 años de edad/año (kg/palmera/año); PO= rendimiento medio de aceite/palmera/año(ton/ha/año); QPO3_5= QTL PO en palmera entre 3 y 5 años de edad; QPO6_9= QTL PO palmeras entre 6 y 9 años de edad.

NOMBRE CG	CARÁCTER	ID QTL	POSICIÓN QTL			LOCUS (Lc) CANDIDATO	LONGITUD (pb)	POSICIÓN LC CANDIDATO			DISTANCIA (pb)
			GL	cM	Pb			cM	Inicio	Fin	
KG196	BN	QBn3_5_1	2	93,9	30.116.706	8P_CL10Contig3-629	964	93,8	28.517.546	28.517.885	-1.599.160
KG197	BN	QBn3_5_1	2	93,9	30.116.706	8P_1_CL1Contig1289-1057	2123	93,8	29.220.105	29.221.162	-896.601
KG198	BN	QBn3_5_c	3	51,9	31.675.329	8P_1_CL1Contig792-1055	1646	52,3	32.126.446	32.127.020	451.117
KG199	BN	QBn6_9_f	4	123,6	42.028.269	8P_1_CL1Contig6604-110	813	122,9	41.828.888	41.829.420	-199.381
TEST	BN	QBn6_9	5	19,6	5.118.398	8P_1_CL1Contig8347-1111	1854	19	4.531.301	4.531.780	-587.097
KG200	BN	QBn6_9_a	7	48,8	13.558.256	8P_1_CL1Contig3962-587	1533	48,1	12.214.495	12.215.209	-1.343.761
KG201	BN	QBn3_5_e	8	18,6	4.530.420	8P_1_CL1Contig761-515	780	17,8	4.351.751	4.352.152	-178.669
KG202	BN	QBn6_9_b	9	24,9	5.647.129	CL7954Contig1	778	25,6	5.711.225	5.711.952	64.096
KG203	BN	QBn3_5	10	17,5	8.602.540	8P_1_CL1Contig4936-1113	2174	17	8.440.533	8.441.020	-162.007
KG204	BN	QBn6_9_d	10	79,5	22.539.548	8P_1_CL1Contig7131-127	659	80,1	22.721.734	22.722.334	182.186
KG205	BN	QBn6_9_c	12	37,8	13.363.071	8P_1_CL1Contig2975-1374	2196	37,2	13.275.298	13.275.929	-87.773
KG206	BN	QBn3_5_c	13	36,2	12.843.766	gi|191204957	889	35,3	12.657.333		-186.433
KG207	BN	QBn3_5_h	15	7,8	3.163.351	8P_1_CL1Contig3178-1019	1331	8,1	3.281.969	3.282.560	118.618
KG141	BN/BW	QBn6_9_a//QBwt3_5_b	7	48,8	13.558.256	M01000000847	1965	48,0	12.159.883	12.152.887	-1.398.373
KG142	BN/BW	QBn6_9_a//QBwt3_5_b	7	48,8	13.558.256	M01000002252	1084	48,0	12.193.633	12.192.205	-1.364.623
KG143	BN/BW	QBn6_9_a//QBwt3_5_b	7	48,8	13.558.256	M01000003256	830	48,2	12.405.324	12.405.866	-1.152.932
KG179	FFB	QFFB3_5	1	59,1	21.730.249	8P_CL1016Contig4-126	936	58,9	21.685.403	21.686.028	-44.846
KG180	FFB	QFFB6_9_3	2	79,8	13.172.615	8P_1_CL1Contig4663-310	1446	79,9	13.183.203	13.184.055	10.588
KG181	FFB	QFFB3_5_b//QFFB6_9_a	2	107,1	31.816.341	8P_1_CL1Contig4741-246	1156	106,1	31.047.746	31.048.104	-768.595
KG182	FFB	QFFB3_5_b//QFFB6_9_a	2	107,1	31.816.341	8P_1_CL1Contig3566-1138	2152	107	31.307.882	31.308.237	-508.459
KG183	FFB	QFFB3_5_b//QFFB6_9_a	2	107,1	31.816.341	8P_CL108Contig2-1119	1835	107	31.454.012	31.455.082	-362.329
KG184	FFB	QFFB3_5_b//QFFB6_9_a	2	107,1	31.816.341	8P_CL108Contig8-657	1765	107	31.455.286	31.456.127	-361.055
KG185	FFB	QFFB3_5_b//QFFB6_9_a	2	107,1	31.816.341	8P_CL108Contig6-301	1548	107	31.455.286	31.456.539	-361.055

KG186	FFB	FFB6_9_J	6	73,5	33.606.711	8P_1_CL1Contig2083-445	835	72,3	33.052.152	33.052.988	-554.559
KG187	FFB	FFB6_9_J	6	73,5	33.606.711	8P_1_CL1Contig3684-142	575	73,6	33.629.933	33.630.297	23.222
KG188	FFB	FFB6_9_J	6	73,5	33.606.711	8P_1_CL1Contig596-1136	1908	73,7	33.638.029	33.638.398	31.318
KG189	FFB	QFFB3_5_a	8	17	4.168.434	8P_1_CL1Contig545-259	701	17,8	4.354.602	4.354.888	186.168
KG190	FFB	QFFB3_5_g	8	19,2	4.666.164	8P_1_CL1Contig1455-145	706	19	4.631.331	4.631.783	-34.833
KG191	FFB	QFFB3_5_e	11	108,9	20.573.328	8P_1_CL1Contig6530-161	629	107,9	20.075.557	20.076.184	-497.771
KG192	FFB	QFFB3_5_e	11	108,9	20.573.328	8P_CL1256Contig1-100	991	108	20.181.455	20.181.960	-391.873
KG193	FFB	QFFB3_5_e	11	108,9	20.573.328	8P_1_CL1Contig1174-546	1371	108,3	20.333.359	20.333.975	-239.969
KG194	FFB	QFFB3_5_e	11	108,9	20.573.328	8P_1_CL1Contig3877-158	872	108,4	20.389.980	20.390.570	-183.348
KG195	FFB	QFFB6_9_i	13	70,5	22.794.962	8P_1_CL1Contig2168-309	578	70,6	22.806.305	22.806.663	11.343
KG255	PO	QPO3_5a	1	92,9	36.432.378	8P_CL1247Contig1-857	1696	93	36.835.906	36.836.463	403.528
KG256	PO	QPO3_5a	1	92,9	36.432.378	8P_CL1199Contig1	1115	93,2	38.126.848	38.127.137	1.694.470
KG257	PO	QPO3_5a	1	92,9	36.432.378	8P_CL1Contig7603	1003	93,6	39.988.513	39.988.691	3.556.135
KG258	PO	QPO3_5a	3	12,9	7.813.607	8P_1_CL1Contig1285	1892	12	7.306.158	7.306.432	-507.449
KG259	PO	QPO3_5a	3	12,9	7.813.607	8P_1_CL1Contig894	883	13,4	8.073.356	8.073.714	259.749
KG260	PO	QPO3_5c	3	51,9	31.675.329	8P_CL118Contig7	1951	52	31.732.383	31.732.550	57.054
KG261	PO	QPO3_5c//QPO6_9b	11	57,9	11.871.114	8P_1_CL1Contig5998	1195	57,8	11.856.887	11.857.371	-14.227
KG262	PO	QPO3_5c	15	78	19.667.860	8P_1_CL1Contig7802	846	77,2	19.640.431	19.640.566	-27.429
KG263	PO	QPO3_5c	15	78	19.667.860	8P_1_CL1Contig7659	1135	77,5	19.654.737	19.655.485	-13.123
KG264	PO	QPO3_5c	15	78	19.667.860	8P_1_CL1Contig2722	1094	78,6	19.816.253	19.816.629	148.393
KG265	PO	QPO3_5c	15	78	19.667.860	8P_1_CL1Contig4430	799	79	19.862.334	19.862.897	194.474
KG2	BW	QBwt_6_9_b	2	41,1	7.612.857	mEgCIR3275_xm_02529485,1	510	41,5	7.674.706	7.681.536	61.849
KG140	BW	QBwt6_9_-C	7	57,3	16.185.204	M01000043696	532	57,1	16.119.197	16.118.678	-66.007
KG146	BW	QaBwt_1//QFwt_1//QBwt3_5_c//QBwt6_9_c	5	125	48.016.649	M01000069634	811	124,7	47.962.206	47.757.422	-54.443

Tabla 3.2: Caracteres relacionados con los componentes de racimo.POP= contenido medio de aceite en relación al peso medio de mesocarpo (%);QPOP= QTL POP; QPOP3_5= QTL POP palmeras entre 3 y 5 años; QPOP6_9=QTL POP palmeras entre 6 y 9 años; Fw= Peso medio de fruto/palmera/año (g); QFw=QTL Fw; IV= Indice de iodo; QI= QTL Indice de iodo; PF= ratio de pulpa respecto al fruto (%); QPF= QTL_PF; QLt_W_e= QTL tamaño medio de anchura de foliolo hoja L17 (cm).

NOMBRE CG	**CARÁCTER**	**ID QTL**	**POSICIÓN QTL**			**LOCUS (Lc) CANDIDATO**	**LONGITUD (pb)**	**POSICIÓN LC CANDIDATO**			**DISTANCIA (pb)**
			GL	cM	Pb			cM	Inicio	Fin	
KG220	POP	Q%POP_1C	3	10,3	6.317.682	8P_CL1123Contig4-785	1350	9,8	6.044.124	6.044.435	-273.558
KG221	POP	Q%POP_1C	3	10,3	6.317.682	8P_1_CL1Contig2289-139	2106	11,1	6.776.479	6.776.936	458.797
KG222	POP	Q%POP_a//QPO3_5_a	3	12,9	7.813.607	8P_1_CL1Contig1285-121	1892	12	7.306.158	7.306.432	-507.449
KG223	POP	Q%POP_a//QPO3_5_a	3	12,9	7.813.607	8P_1_CL1Contig894-166	883	13,4	8.073.356	8.073.714	259.749
KG224	POP	Q%POP_c//QPO6_9_a	7	86,4	24.931.432	8P_1_CL1Contig5969-102	958	86,8	25.147.904	25.148.557	216.472
KG225	POP	Q%POP_c//QPO6_9_a	7	86,4	24.931.432	8P_1_CL1Contig4786-186	1179	87,1	25.345.243	25.345.718	413.811
KG226	POP	Q%POP_b	8	95,5	32.463.588	8P_1_CL1Contig6665-1022	1859	94,5	32.127.859	32.128.523	-335.729
KG227	POP	Q%POP_b	8	95,5	32.463.588	8P_1_CL1Contig3765-525	1008	96,5	32.802.067	32.802.329	338.479
KG11	Fw	QFwt_a	14	91,3	23.040.196	M01000008373	1663	91,3	23.034.842	23.015.230	-5.354

KG146[a]	Fw	QaBwt_1//QFwt_1//QBwt3_5_c//QBwt6_9_c	5	125	48.016.649	M01000069634	811	124,7	47.962.206	47.757.422	-54.443
KG1	IV	QI_a	3	17,2	9.666.280	mEgCIR3847	421	17,2	9.684.871	9.683.290	18.591
KG12[b]	IV	qI_j//QLt_W_e	13	33,5	12.270.783	M01000002200	1201	34,5	12.491.918	12.503.583	221.135
KG135	IV	qI_i	1	58,9	21.665.616	M01000003117	2273	58,9	21.685.513	21.688.998	19.897
KG138	PF	Q%PF_1	6	77,3	34.230.310	M01000009619	1410	76,8	34.149.414	5.297.645	-80.896

[a] El mismo gen co-localizado en la mismo posición para el crácter BW de caracteres relacionados con la producción

.[b]El mismo gen co-localizado con un carácter vegetativo

Tabla 3.3: Caracteres relacionados con componentes vegetativos. HI= incremento de altura (cm); Qht= QTL HI; Q-C_StGr= QTL crecimiento en altura (cm); qP_W_f= anchura media del peciolo de hoja L17 (cm).

NOMBRE CG	CARÁCTER	ID QTL	POSICIÓN QTL			LOCUS (Lc) CANDIDATO	LONGITUD (pb)	POSICIÓN LC CANDIDATO			DISTANCIA (pb)
			GL	cM	Pb			cM	Inicio	Fin	
KG161	HI	Qht	1	93,9	41.260.746	8P_1_CL1Contig3885-852	1165	93,8	40.814.814	40.815.410	-445.932
KG162	HI	Qht	1	93,9	41.260.746	8P_1_CL1Contig5925-865	1764	93,8	40.667.392	40.668.097	-593.354
KG163	HI	Qht_C	1	95,6	43.465.831	8P_1_CL1Contig5227-359	910	95,6	43.437.641	43.438.073	-28.190
KG164	HI	Qht1C	2	127,7	41.964.781	8P_1_CL1Contig11412-1053	1665	127,9	41.981.692	41.982.104	16.911
KG165	HI	Qht1C	2	127,7	41.964.781	8P_1_CL1Contig4381-1139	3304	127,7	41.982.390	41.982.905	17.609
KG166	HI	Q-C_StGr_b	2	136,8	44.514.722	8P_1_CL1Contig7081-337	1021	136,6	44.067.642	44.068.634	-447.080
KG167	HI	Q-CSt	2	136,8	44.514.722	8P_CL1255Contig2-411	610	136,6	43.975.576	43.976.083	-539.146
KG168	HI	Q-CSt Gr_a	2	157,4	50.738.342	8P_1_CL1Contig9289-286	1123	157,9	50.794.970	50.795.923	56.628
KG169	HI	QHt_a	3	43,1	24.437.929	8P_1_CL1Contig1336-961	1608	42,2	24.600.281	24.600.536	162.352
KG170	HI	QHt_d	4	14,1	4.219.661	8P_1_CL1Contig1032-823	3053	14	4.196.283	4.196.897	-23.378
KG171	HI	QHt_d	4	14,1	4.219.661	8P_1_CL1Contig4489-1006	2055	14,9	4.462.097	4.462.851	242.436
KG172	HI	QHt_2c	4	100,6	24.597.133	8P_1_CL1Contig240-608	1328	100,4	24.572.939	24.573.908	-24.194
KG173	HI	q_Ht_h	7	103,8	33.778.092	8P_1_CL1Contig9200-521	1656	103,6	33.540.080	33.540.361	-238.012
KG174	HI	q_Ht_h	7	103,8	33.778.092	8P_1_CL1Contig1247-233	464	103,5	33.492.282	33.492.494	-285.810
KG175	HI	qHt_g	8	69,8	23.185.343	8P_CL1026Contig1-144	705	69,9	23.228.331	23.228.829	42.988
KG176	HI	qHt_g	8	69,8	23.185.343	8P_1_CL1Contig11217-507	1456	70	23.297.141	23.297.916	111.798
KG177	HI	qHt_g	10	90,4	26.502.355	8P_1_CL1Contig11102-982	1561	90,3	26.438.811	26.439.938	-63.544
kg178	HI	QHt_f	11	66,6	12.985.034	8P_1_CL1Contig5710-120	1170	66,2	13.064.347	13.065.119	79.313
KG144		QHt_e//qP_W_f	12	127,6	25.726.657	M01000023551	1105	127,0	25.645.651	25.647.388	-81.006
KG145		QHt_1C	1	93,9	41.168.544	M01000007495	551	93,4	38.779.467	38.780.297	-2.389.077
KG147		QHt_d	4	14,1	4.219.661	M01000004883	2567	14,0	4.183.689	4.193.852	-35.972
KG12		qI_j//QLt_W_e	13	33,5	12.270.783	M01000002200	1201	34,5	12.491.918	12.503.583	221.135

Anexo 4: Anotaciones funcionales de los genes co-localizados en el mapa integrado de palmera de aceite. Blastx muestra las anotaciones dónde se ha determinado un mayor homología con la secuencia candidata en *E.guineensis*. La anotación GO muestra dónde se puede localizar a nivel celular (C), el proceso biológico donde participa (P) y su función a nivel molecular (F).

Tabla 4.1: Caracteres relacionados con la producción.

NOMBRE CG	CARÁCTER	BLASTx	E-VALOR	ANOTACIÓN GO
KG196	BN	XP_010912706.1 Factor de transcripción 21 MADS-box	1,00E-172	C:núcleo; F:Unión ADN; F:Actvidad factor de transcripción, unión secuencias específica ADN; F:unión de proteínas; P:proceso biosintético; P:contenedor de bases nucleotídicas componente de proceso metabólicos
KG197	BN	XP_010912749.1 Proteina FRIGIDA 3	0,00E+00	F:función molecular; P:proceso biológico
KG198	BN	ABR19829.1 Cisteína proteinasa	0,00E+00	P:proceso metabólico de proteínas; C:vacuola; F:actividad hidrolasa
KG199	BN	XP_010919353.1 Bases de cadena larga esfingoide kinasa-1	5,00E-58	F:actividad kinasa; C:plastidio
TEST	BN	XP_010920826.1 Acil- coenzima A oxidasa 4, en peroxisoma	0,00E+00	F:unión de nucleótidos; P:desarrollo del embrión; C:peroxisoma; F:actividad catalítica; P:proceso catabólico; P:proceso de metabolismo de lípidos; P:desarrollo post-embrionario; P:organización de componentes celulares; P:transportador
KG200	BN	XP_010926098.1 alfa, alfa-trealosa-fosfato sintasa [UDP-formador]6	0,00E+00	P:proceso de metabolismo de carbohidratos; P:proceso de modificación de proteínas celulares; P:proceso biosintético; F:actividad transferasa; F:actividad hidrolasa
KG201	BN	XP_010927527.1 PREDICTED:Proteina L23 ribosomal 60S	9,00E-88	P:desarrollo del embrión; C:mitocondria; P:desarrollo post-embrionario; F:actividad de molecula estructural; C:citosol; C:nucleolo; C:ribosoma; P:contenedor de bases nucleotídicas componente de proceso metabólicos; C:plastidio; P:traducción
KG202	BN	XP_010929742.1 Proteína no caracterizada At1g04910	2,00E-130	C:endosoma; C:aparato de Golgi; F:actividad transferasa
KG203	BN	XP_010931416.1 2,3 bifosfoglicerato independiente fosfoglicerato mutasa	0,00E+00	P:proceso de metabolismo de carbohidratos; P:generación de metabolitos precursores y energía; C:citoplasma; F:actividad catalítica; P:proceso catabólico; F:unión; P:contenedor de bases nucleotídicas componente de proceso metabólicos
KG204	BN	XP_010932341.1 Proteína 48 control división celular homologa B	2,00E-41	F:unión de nucleótidos; C:núcleo; C:citoesQueleto; C:citoplasma; P:desarrollo organismo multicelular; P:diferenciación celular; P:transportador; C:membrana plasmática; P:ciclo celular; P:reproducción; F:actividad hidrolasa
KG205	BN	XP_010911234.1 Proteína 2 unión GTP nucleolar	0,00E+00	F:unión de nucleótidos; C:nucleolo; F:actividad hidrolasa
KG206	BN	ACI23376.1 desconocido	3,00E-103	F:actividad catalítica
KG207	BN	XP_010938962.1 Proteína Porina 5 membrana mitocondrial externa	0,00E+00	C:membrana; C:mitocondria; F:actividad transportadora; P:proceso celular
KG141	BN	XP_010926101.1 Proteína 71 asociada a microtubulo	0,00E+00	-
KG142	BN	XP_010926099.1 AAR2 regulador de respuesta de dos componentes	0,00E+00	-
KG143	BN	XP_010916258.1 Proteína T complejo 1 subunidad delta	6,00E-71	-

KG179	FFB	XP_010924269.1 2-hidroxiacil-CoA liasa	3,00E-131	P:proceso metabolismo secundario; F:actividad catalítica; P:proceso catabólico; C:citosol; F:unión; P:proceso celular
KG180	FFB	XP_010911991.1 Proteína F8box "tubby-like"	0,00E+00	-
KG181	FFB	XP_010912851.1Proteina baja calidad: 3-oxoacil-ACP-reductasa 4	4,00E-179	F:unión de nucleótidos; P:respuesta a estímulos abióticos; P:desarrollo organismo multicelular; P:proceso de metabolismo de lípidos; P:proceso biosintético; F:actividad transferasa; C:plastidio; P:proceso celular; P:respuesta a estres
KG182	FFB	XP_010930036.1 Proteína transportadora Sec61 subunidad alfa	0,00E+00	C:mitocondria; C:membrana; P:transportador
KG183	FFB	XP_010912890.1 Citocromo P450 71A9	0,00E+00	F:actividad catalítica; F:unión
KG184	FFB	XP_010911404.1 Citocromo P450 71A1	0,00E+00	F:actividad catalítica; F:unión
KG185	FFB	XP_010912890.1 Citocromo P450 71A9	0,00E+00	F:actividad catalítica; F:unión
KG186	FFB	XP_010924276.1 3-ketoacil-CoA sintasa 4	7,00E-93	C:membrana; P:proceso de metabolismo de lípidos; P:proceso biosintético; F:actividad transferasa; P:proceso celular
KG187	FFB	XP_010924299.1 Proteína de baja calidad: aspartico proteinasa orizasin-1	1,00E-57	P:proceso metabólico de proteínas; P:proceso de metabolismo de lípidos; F:actividad hidrolasa
KG188	FFB	XP_010926197.1 aspartico proteinasa orizasin-1	0,00E+00	P:proceso metabólico de proteínas; P:proceso de metabolismo de lípidos; C:vacuola; F:actividad hidrolasa
KG189	FFB	XP_010927527.1 Proteina L23 60S ribosomica	1,00E-68	C:mitocondria; F:actividad de molecula estructural; C:ribosoma; P:traducción
KG190	FFB	XP_010927544.1Proteína ubiQuitina -E3 ligasa RGLG2	4,00E-63	P:proceso metabólico; F:unión; P:proceso celular
KG191	FFB	XP_010933823.1 Proteina VIN3	3,00E-81	-
KG192	FFB	XP_010933806.1 Proteina expresadamaternalmente Gen 5 xpresadoprotein MATERNALLY EXPRESSED GENE 5-like isoform X3	1,00E-102	F:unión de nucleótidos; F:ARN vinculante; P:contenedor de bases nucleotídicas componente de proceso metabólicos
KG193	FFB	XP_010933790.1 Proteína no caracteriazada LOC105054084	0,00E+00	C:citosol
KG194	FFB	XP_010933787.1 Isoforma E Citocromo b5	3,00E-94	F:actividad catalítica; C:retículo endoplasmático; F:unión; C:vacuola; C:membrana plasmática; C:plastidio; C:tilacoide
KG195	FFB	XP_010937204.1 peptidil-prolil cis-trans isomerasa interaccionando NIMA 4	6,00E-53	F:actividad catalítica; P:proceso de modificación de proteínas celulares
KG255	PO	XP_010934393.1 Proteína no caracterizada LOC105054551	0	C:plastidio
KG256	PO	XP_010935215.1 cicloartenol-C-24-metiltransferasa 1	2,00E-176	P:desarrollo del embrión; P:proceso metabolismo secundario; P:proceso catabólico; P:desarrollo post-embrionario; C:retículo endoplasmático; P:proceso biosintético; C:vacuola; P:proceso celular; P:respuesta a estres; P:proceso de metabolismo de lípidos; F:actividad transferasa
KG257	PO	XP_010923463.1 Proteína no caracterizada LOC105046548	2,00E-168	F:unión de lípidos
KG258	PO	XP_010915394.1 Enzima malico NADP dependiente	0	F:unión de nucleótidos; P:proceso de metabolismo de carbohidratos; P:proceso metabólico de proteínas; F:unión de proteínas; F:actividad catalítica; P:proceso catabólico; P:proceso de metabolismo de lípidos; P:organización de componentes celulares; P:proceso biosintético
KG259	PO	XP_010915446.1 Factor de transcripción 3 MADS-box	1,00E-113	C:núcleo; F:Unión ADN; F:Actvidad factor de transcripción, unión secuencias específica ADN; F:unión de proteínas; P:proceso biosintético; P:contenedor de bases nucleotídicas componente de proceso metabólicos
KG260	PO	XP_010907607.1 Proteína 2a bisintesis de bases de larga cadena	0,00E+00	C:membrana; P:proceso de metabolismo de lípidos; C:retículo endoplasmático; P:proceso biosintético; F:actividad transferasa; F:unión; P:proceso celular
KG261	PO	XP_010933293.1 Proteína 7 GEM	1,00E-158	-
KG262	PO	XP_010937204.1 peptidil-prolil cis-trans isomarasa interaccionando NIMA4	2,00E-73	F:actividad catalítica; P:proceso de modificación de proteínas celulares

KG263	PO	XP_010939832.1 cistationina gamma sintasa, cloroplástica	4,00E-43	F:actividad catalítica; F:unión
KG264	PO	XP_010939851.1 proteosoma tipo 5 subunidad alfa	1,00E-174	C:núcleo; P:proceso metabólico de proteínas; C:citoplasma; P:proceso catabólico; P:proceso celular; F:actividad hidrolasa
KG265	PO	XP_010937258.1 1-fosfatidilinositol-3-fosfato-5 kinasa FAB1B	5,00E-121	F:unión de nucleótidos; P:proceso de metabolismo de lípidos; F:actividad kinasa
KG2	BW	XP_010912533.1 serina hidroximetiltransferasa 7	0,00E+00	F:actividad transferasa; F:unión; P:proceso celular
KG140	BW	XP_010926413.1 Factor de transcripción ERF014 sensible a etileno	5,00E-68	C:núcleo; F:Unión ADN; P:proceso de metabolismo de carbohidratos; F:Actvidad factor de transcripción, unión secuencias específica ADN; P:contenedor de bases nucleotídicas componente de proceso metabólicos; F:actividad hidrolasa; P:traducción
KG141	BW	XP_010926101.1Proteína 70-1 asociada a microtúbulo	0,00E+00	F:unión de proteínas; P:organización de componentes celulares
KG142	BW	XP_010926099.1 regulador de la respuesta de dos comoponentes ARR2	0,00E+00	F:Unión ADN; P:transducción de señales; C:intracelular
KG143	BW	XP_010916258.1 PREDICTED: Proteína complejo T subunidad 1 delta	6,00E-71	P:proceso metabólico de proteínas; P:generación de metabolitos precursores y energía; P:proceso catabólico; P:proceso biosintético; P:contenedor de bases nucleotídicas componente de proceso metabólicos; P:respuesta a estres; F:unión de nucleótidos; P:proceso de metabolismo de carbohidratos; P:respuesta a estímulos abióticos; F:unión de proteínas; P:organización de componentes celulares; C:citosol
KG146	BW	XP_010937676.1 Histona H2B.11	4,00E-60	C:núcleo; F:Unión ADN; F:unión de proteínas

Tabla 4.2: Caracteres relacionados con componentes de racimo.

NOMBRE CG	CARÁCTER	BLASTx	E-VALOR	ANOTACIÓN GO
KG220	POP	XP_010915295.1 cisteína sintasa	0,00E+00	-
KG221	POP	XP_010915352.1 6-fosfofructokinasa 3 ATP dependiente	0,00E+00	-
KG222	POP	XP_010915394.1 enzima málico dependite deNADP	0,00E+00	F:unión de nucleótidos; P:proceso de metabolismo de carbohidratos; P:proceso de metabolismo de proteínas; F:unión de proteínas; F:actividad catalítica; P:proceso catabólico; P:proceso metabolismo de lípidos; P:organización componentes celulares; P:proceso biosintético
KG223	POP	XP_010915446.1Factor de transcripción 3 "MADS- box"	1,00E-113	C:núcleo; F:unión ADN; F:actividad como factor de transcripción, unión ADN secuencia específica; F:unión de proteínas; P:proceso biosintético; P:proceso metabólico compuesto por una nucleobase
KG224	POP	XP_010926698.1 subunidad del receptor importador mitocondrial TOM20	3,00E-73	C:membrana; C:mitocondria; P:organización componentes celulares; C:citosol; C:ribosoma; F:unión; P:transporte
KG225	POP	XP_010926689.1 Proteína L7a ribosómica 60S	2,00E-132	F:actividad molécula estructural; C:ribosoma; P:traducción
KG226	POP	XP_010928773.1 Proteosoma 26S subunidad 1 homologa A no ATPasa reguladora	0,00E+00	P:proceso de metabolismo de proteínas; P:proceso catabólico; F:actividad reguladora de enzimas; C:intracelular
KG227	POP	XP_010928835.1 Dominio rodanasa continente proteína ,cloroplasto	8,00E-170	P:igeneración de metabolitos precursores y energía; C:membrana; P:desarrollo post- embrionario; P:morfogenesis de estructuras anatómicas; P:fotosíntesis; P:proceso biosintético; C:plástidos; C:tilacoides

KG11	Fw	XP_010938882.1 Proteína de unión poliadenilato RBP47B'P	0	F:unión de nucleótidos; C:núcleo; P:respuesta a estímulos abióticos; F:ARN vinculante; C:citoplasma; P:respuesta a estrés; P:proceso celular
KG146	Fw	XP_010937676.1 Histona H2B.11	4,00E-60	C:núcleo; F:unión ADN; F:unión de proteínas
KG1	IV	XP_010932283.1 Proteína no caracterizada LOC105052991	0,00E+00	_
KG12	IV	XP_010936651.1 Proteína no caracterizada LOC105056228	1,00E-79	C:cloroplasto; P:proceso metabólico; F:actividad catalítica; F:coenzyme unión
KG135	IV	XP_010924269.1 2-hidroxiacil-CoA liasa	0,00E+00	P:secondary proceso metabólico; F:actividad catalítica; P:proceso catabólico; C:citosol; F:unión; P:proceso celular
KG138	PF	XP_010924347.1 Proteína no caracterizada LOC105047228	0,00E+00	P:proceso biológico; C:componente celular

Tabla 4.3 Caracteres relacionados con componentes vegetativos.

NOMBRE CG	CARÁCTER	BLASTx	E-VALOR	ANOTACIÓN GO
KG161	HI	XP_010913468.1 cadena catalítica ferredoxin-tiorredoxin reductasa	1,00E-105	-
KG162	HI	XP_010935952.1 GDP-manosa 3,5-epimerasa 2	0,00E+00	-
KG163	HI	XP_010937949.1 UDP-arabinopiranosa mutasa 1	6,00E-143	P:proceso metabolismo de carbohidratos; F:unión de proteínas; P:organización de componentes celulares; C:citosol; C:aparato de Golgi; P:Proceso biosintético; F:actividad transferasa; P:proceso metabólico compuesto contenedor de nucleobase
KG164	HI	XP_010928300.1 bomba de protones de membrana vacuolar pirofosfato	0,00E+00	P:desarrollo de organismo multicelulares; C:aparato de Golgi; C:vacuola; C:membrana plasmática; C:plastidios; F:actividad transporteadora; P:respuesta a estrés; P:proceso celular; P:respuesta a estímulos abióticos; C:endosoma; C:mitocondria; F:actividad hidrolasa
KG165	HI	XP_010921969.1 posible isoleucina-ARNt-ligasa	0,00E+00	P:generación de metabolitos precursores y energía; P:proceso catabólico; P:proceso metabólico compuesto contenedor de nucleobase; P:respuesta a estrés; P:traducción; F:unión de nucleótidos; P:proceso metabolismo de carbohidratos; P:respuesta a estímulos abióticos; P:organización de componentes celulares; C:citosol; F:actividad hidrolasa
KG166	HI	XP_010913485.1 peroxidasa 63	1,00E-104	F:actividad catalítica; P:proceso catabólico; C:región extracelular; F:unión; P:respuesta a estrés; P:proceso celular
KG167	HI	XP_010913468.1 cadena catalítica ferredoxin-tiorredoxin reductasa	1,00E-27	P:proceso metabolismo de carbohidratos; F:actividad catalítica; P:Proceso biosintético; F:unión; C:plastidios; P:proceso celular
KG168	HI	XP_010913940.1 proteína 8 dedo de zinc A20 y dominio AN1 asociado a estrés	9,00E-108	F:unión ADN
KG169	HI	XP_010916665.1 subunidad mu compleho AP-4	0,00E+00	C:membrana; C:citosol; P:transporte
KG170	HI	YP_006073103.1 apoproteina A2 de fotosistema I P700, cloroplasto	0,00E+00	F:actividad catalítica; C:membrana; P:proceso de modificación de proteínas celulares; P:fotosíntesis; F:unión; C:plastidios; C:tilacoides
KG171	HI	XP_010918079.1 posible metiltransferasa PMT21	0,00E+00	F:actividad transferasa
KG172	HI	XP_010918518.1 Proteína 2 de unión de esteroides a membrana	8,00E-131	C:núcleo; C:membrana; C:retículo endoplasmático; C:plastidios; C:tilacoides
KG173	HI	XP_010912190.1 TAP46 subunidad	0,00E+00	P:proceso de metabolismo secundario; P:proceso catabólico; P:transducción de señales; P:proceso de modificación de

		reguladora PP2A		proteínas celulares; C:citosol; F:regulador de la actividad enzimática
KG174	HI	XP_010924912.1 TAP46 subunidad reguladora PP2A	0,00E+00	P:proceso de metabolismo secundario; P:proceso catabólico; P:transducción de señales; P:proceso de modificación de proteínas celulares; C:citosol; F:regulador de la actividad enzimática
KG175	HI	XP_010928200.1 proteína LORELEI anclada a GPI	2,00E-120	C:célula; P:polinización
KG176	HI	XP_010928199.1 Proteina " WD repeat-containing protein C2A9.03" no caracterizada	0,00E+00	-
KG177	HI	XP_010932610.1 proteína no caracterizada LOC105053217	0,00E+00	F:cromatina vinculante; C:plasmodesmo
kg178	HI	XP_010933275.1 subunidad 1 complejo THO	1,00E-158	C:núcleo; P:respuesta a estímulos bióticos; P:respuesta a estímulos externos; P:transporte; P:proceso metabólico compuesto contenedor de nucleobase; P:regulación de la expresión génica, epigenética; P:respuesta a estrés
KG144		XP_010935702.1 manano sintasa 1	0,00E+00	C:membrana; P:organización de componentes celulares; C:aparato de Golgi; F:actividad transferasa
KG145		XP_010937172.1 posible subunidad 8 [UDP-formadora]catalítica de celulosa sintasa A	3,00E-68	P:proceso metabolismo de carbohidratos; P:organización de componentes celulares; P:Proceso biosintético; F:unión; F:actividad transferasa; C:membrana plasmática
KG147		XP_010906231.1 PREDICTED: Proteína "ENHANCED DISEASE RESISTANCE 2"	0,00E+00	F:unión de lipidos
KG12		XP_010936651.1 proteína no caracterizada LOC105056228 isoform X1	1,00E-79	C:cloroplasto; P:proceso metabólico; F:actividad catalítica; F:unión de coenzimas

Anexo 5: Genes candidatos seleccionados a partir de diferentes referencias como literatura relacionada, rutas biosintéticas y patentes.

Tabla 5.1: Genes candidatos y su función. Nombre: se refiere al nombre dado al gen cadidato, y si procede de patente -P- o de otras vías -KG. La relación con el carácter de interés se determinó a partir de su función referenciada mediante la clasificación hecha inicialmente. Especie, se refiere a la especie de origen de la secuencia, secuencia es la referencia de NCBI dónde puede buscarse la secuencia; referencia bibliográfica se refiere al origen de la secuencia y a la bibliografía relacionada con la función, y por último la función con su función.

NOMBRE	RELACIÓN CON CARÁCTER DE INTERÉS		ESPECIE	SECUENCIA	REFERENCIA BIBLIOGRÁFICA	FUNCIÓN
	Clasificación	Carácter/es				
KG275	Acumulación de aceite en semilla	POP, PO	*Elaeis guineensis*	XM_010935827	NCBI- Nucleótidos	Oleosin
P4	Aumento del rendimiento	Bn, FFB, FN,FW	*Elaeis guineensis*	HC924133	Patente EP2199398 (Reuzeau y col., 2010)	Factor de transcripción MADS BOX
P39	Biomasa	OTROS	*Elaeis guineensis*	FB787669	WO2008034648 (A1) (Puzio y col. 2008)	Proteína 1 NAC
P58	Biomasa	OTROS	*Elaeis guineensis*	AX463424	EP1217068 (A2) (Abdullah y col. 2002)	Peroxirredoxin 1-Cys
KG269	Biomasa	OTROS	*Elaeis guineensis*	FB669059	Patente WO2009153208 (Sanz Molinero y col.2009)	Proteína 1 NAC
KG210	Calidad de Fruto		*Elaeis guineensis*	JX556215;HF562332	Morcillo y col. 2013	Lipasa
KG212	Metabolismo de lípidos	POP,PO	*Elaeis guineensis*	XM_010916712	Dussert S. y col.2013	Palmitoil-ACP- tioesterasa
KG213	Metabolismo de lípidos	POP,PO	*Elaeis guineensis*	XM_010926808	Dussert S. y col.2013	Palmitoil-ACP-tioesterasa
KG274	Calidad de Fruto	FW,FN	*Elaeis guineensis*	XM_010927965	NCBI- Nucleótidos	Lipoil sintasa
KG153	Desarrollo de fruto	FW,FN	*Arabidopsis thaliana*	NM_119611	NCBI-Nucleótidos; Lease KA y col. 2001	Proteína de unión nucleótido guanina subunidad BETA
KG154	Desarrollo de fruto	FW,FN	*Arabidopsis thaliana*	NM_119266	NCBI-Nucleótidos; Zhang Y. y col. 2008	DDB1 y CUL4 factor asociado 1
KG155	Desarrollo de fruto	FW,FN	*Arabidopsis thaliana*	NM_121371	NCBI- Nucleótidos	Proteína complejo elongador 1
KG242	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	EgETRF2R3*	Preedakoon, P (2009)	Ethylene receptor protein
KG244	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	EgMAX4(F1R1)*	Preedakoon, P (2009)	More axillary branches 4, members of carotenoid cleavage dioxygenase (CCDs) family of enzyme
KG252	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	XM_010915369	NCBI- Nucleótidos	Familia AP-2 factor de transcripción sensible a etileno
KG253	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	XM_010912847	NCBI- Nucleótidos	Familia AP-2 factor de transcripción sensible a etileno
KG254	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	NC_026001	NCBI- Nucleótidos	Familia AP-2 factor de transcripción sensible a etileno
KG270	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	AY182169	NCBI- Nucleótidos	opsc155 fosfogluconolactonasa-6 putativa
KG271	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	KJ789862	Singh y col.2014	Gen virescens R2R3-MYB
KG288	Desarrollo de fruto	FW,FN	*Elaeis guineensis*	AJ507416	NCBI- Nucleótidos	Pseudogene Tpasa
KG114	Desarrollo de la semilla y crecimiento	FW,FN	*Jatropha curcas*	JC000041	NCBI- Nucleótidos; Foidl N. y col. 1996	Alcohol deshidrogensa
KG87	Desarrollo y crecimiento	OTROS	*V. vinifera*	XM_002268016	NCBI- Nucleótidos	Posible Indole-3- piruvato monooxigenasa YUCCA8
KG113	Desarrollo y crecimiento	OTROS	*Jatropha curcas*	JC000035	NCBI- Nucleótidos; Foidl N. y col. 1996	Peptidil-propil cis-trans-isomerasa

KG106	Desarrollo, crecimiento y fotosíntesis	OTROS	*O.europeae*	NC_013707	NCBI- Nucleótidos	PSII Proteína baja MW
KG107	Desarrollo, crecimiento y fotosíntesis	OTROS	*O.europeae*	NC_013707	NCBI- Nucleótidos	PSII citocromo b559
KG108	Desarrollo, crecimiento y fotosíntesis	OTROS	*O.europeae*	NC_015401	NCBI- Nucleótidos	ribulosa-1,5-bifosfato carboxilasa/oxigenasa
KG278	Desarrollo, crecimiento y fotosíntesis	OTROS	*Elaeis guineensis*	XM_010920538.1	NCBI- Nucleótidos	Subunidad IV B centro de reacción Fotosistema I
KG115	Desarrollo, crecimiento y respuesta a estrés	OTROS	*Jatropha curcas*	JC000047	NCBI- Nucleótidos; Foidl N. y col. 1996	Enzima de procesamiento vacuolar
KG117	Desarrollo, crecimiento y respuesta a estrés	OTROS	*Jatropha curcas*	JC000055	NCBI- Nucleótidos; Foidl N. y col. 1996	Receptor de Etileno
KG101	Desarrollo, crecimiento, fotosíntesis, respuesta a estrés	OTROS	*O.europeae*	NC_013707	NCBI- Nucleótidos	ATPasa subunidad IV
KG102	Desarrollo, crecimiento, fotosíntesis, respuesta a estrés	OTROS	*O.europeae*	NC_013707	NCBI- Nucleótidos	ATPasa subunidad epsilon
KG103	Desarrollo, crecimiento, fotosíntesis, respuesta a estrés	OTROS	*O.europeae*	NC_013707	NCBI- Nucleótidos	ATPasa subunidad III
KG105	Desarrollo, crecimiento, fotosíntesis, respuesta a estrés	OTROS	*O.europeae*	NC_013707	NCBI- Nucleótidos	PSII
KG156	Elongación de tallo	HI	*Arabidopsis thaliana*	NM_180285	NCBI- Nucleótidos	Delta(24)-esterol reductasa
KG157	Elongación de tallo	HI	*Arabidopsis thaliana*	NM_101298	NCBI- Nucleótidos	Proteína ribosómica L10-1 60S
KG159	Elongación de tallo	HI	*Arabidopsis thaliana*	NM_116578	NCBI- Nucleótidos	Familia de proteínas GH3 sensibles a auxina
KG160	Elongación de tallo	HI	*Arabidopsis thaliana*	NM_121421	NCBI- Nucleótidos	Proteína complejo CHC1: remodelación de cromodominio
KG233	Elongación de tallo	HI	*Elaeis guineensis*	AY556420	Lee y col. 2015	Proteina Asaparagina sintasa
KG234	Elongación de tallo	HI	*Elaeis guineensis*	AY556423	Lee y col. 2015	Promotor Proteínas Asparagina sintasa
KG245	Elongación de tallo	HI	*Elaeis guineensis*	EgBRX*	Preedakoon, P (2009)	Brevis radix for regulating brassinosteroid-biosynthesis
KG246	Elongación de tallo	HI	*Elaeis guineensis*	EgARF1*	Preedakoon, P (2009)	Auxin response factor as transcription factor
KG247	Elongación de tallo	HI	*Elaeis guineensis*	EgPINF3-9*	Preedakoon, P (2009)	polar auxin Transportador
KG248	Elongación de tallo	HI	*Elaeis guineensis*	EgPINF3-5*	Preedakoon, P (2009)	polar auxin Transportador
KG249	Elongación de tallo	HI	*Elaeis guineensis*	EgPINF3-4*	Preedakoon, P (2009)	polar auxin Transportador
KG250	Elongación de tallo	HI	*Elaeis guineensis*	EgPINF3-6*	Preedakoon, P (2009)	polar auxin Transportador
KG251	Elongación de tallo	HI	*Arabidopsis thaliana*	NM_111270	NCBI- Nucleótidos	Estrigolactona esterasa D14
KG286	Elongación de tallo	HI	*Elaeis guineensis*	XM_010928170.1	NCBI- Nucleótidos	Proteína SLR1 DELLA
KG289	Elongación de tallo	HI	*Elaeis guineensis*	AY254310	NCBI- Nucleótidos	ácido carboxilico-1-aminociclopropano 1 oxidasa
P13	Elongación de tallo	HI	*Elaeis guineensis*	HB772392	EP2090662 (A2) (Puzio Piotr y col. 2009)	Posible Proteína QM
KG64	Elongación de tallo	HI	*A.thaliana*	NM_106017	NCBI-Nucleótidos; Zhou J y col. 2011	Auxina portadora de componente de flujo 1 (PIN1)
KG69	Elongación de tallo	HI	*A.thaliana*	NM_120100	NCBI-Nucleótidos; Zhou J y col. 2011	Proteína Brasinoesteroidea insensitiva 1 (BRI1)
KG70	Elongación de tallo	HI	*A.thaliana*	NM_001203975	NCBI-Nucleótidos; Zhou J y col. 2011	Kinasa receptora asociada a Brasinoesteroide insensitivo 1 (BAK1)
KG75	Elongación de tallo	HI	*S.lycopersicum*	NM_001247838	NCBI-Nucleótidos; Chandler PM y	Receptor Giberellina 1

					col. 2008	
KG78	Elongación de tallo	HI	*A.thaliana*	NM_105305	NCBI-Nucleótidos; Tranbarger T. y col. 2011	Receptor Etileno 1
KG80	Elongación de tallo	HI	*A.thaliana*	NM_111329	NCBI-Nucleótidos; Tranbarger T. y col. 2011	Proteína EIN4: receptor etileno
KG81	Elongación de tallo	HI	*A.thaliana*	NM_113216	NCBI-Nucleótidos; Tranbarger T. y col. 2011	Receptor Etileno 2
KG243	Factor de transcripcion	OTROS	*Elaeis guineensis*	EgEBF*	Preedakoon, P (2009)	EIN3 (Ethylene insensitive) Unión factor as F-box
KG268	Factor de transcripcion	FW,FN	*Elaeis guineensis*	AF411845	NCBI- Nucleótidos	Factor de transcripción MADS box(AGL2-3)
KG119	Factor de transcripción	FW,FN	*Elaeis guineensis*	AJ581467	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG120	Factor de transcripción	FW,SH	*Arabidopsis thaliana*	NM_001203767.1	Singh y col.2013B	Factor de transcripcion MADS BOX
KG121	Factor de transcripción	FW,FN	*Elaeis guineensis*	AY739702	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG122	Factor de transcripción	FW,FN	*Elaeis guineensis*	AY739700	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG123	Factor de transcripción	FW,FN	*Elaeis guineensis*	AF411848	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG124	Factor de transcripción	FW,FN	*Elaeis guineensis*	AY739698	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG125	Factor de transcripción	FW,FN	*Elaeis guineensis*	AF411842	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG126	Factor de transcripción	FW,FN	*Elaeis guineensis*	DQ090962	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG127	Factor de transcripción	FW,FN	*Elaeis guineensis*	AJ581470	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG128	Factor de transcripción	FW,FN	*Elaeis guineensis*	AF411840	NCBI-Nucleótidos	Factor de transcripcion MADS BOX
KG24	Metabolismo de lípidos	POP,PO	*E.guineensis*	AF424808	NCBI- Nucleotidos	Palmitoil ACP tioestersa
KG27	Metabolismo de lípidos	POP,PO	*E.guineensis*	FJ940767	NCBI- Nucleotidos;Bourgis y col. 2011	Beta Ketoacil ACP Sintasa II
KG29	Metabolismo de lípidos	PO, BW	*E.guineensis*	AF261691	NCBI- Nucleotidos; Cha TS y col.2001	Glutelin
KG31	Metabolismo de lípidos	POP,PO	*E.oleifera*	AY012452	NCBI-Nucleótidos	ATP sintasa subunidad beta
KG35	Metabolismo de lípidos	PO, IV	*E.oleifera*	EU057620	NCBI-Nucleótidos; Ho CL y col.2007	Acido graso omega-3 desaturasa
KG38	Metabolismo de lípidos	PO	*E.guineensis*	EU285005	NCBI-Nucleótidos; Low ET y col. 2008	Holocarboxilasa sintetasa III
KG39	Metabolismo de lípidos	PO, IV	*E.oleifera*	AF169015	NCBI- Nucleótidos	Beta Ketoacil ACP Sintasa III
KG47	Metabolismo de lípidos	PO, IV	*E.oleifera*	FJ796069	NCBI- Nucleótidos	Acyl CoA dehidrogenasa
KG57	Metabolismo de lípidos	FW, PO	*A.thaliana/E.guineensis*	At1g01090//JN203210	NCBI- Nucleótidos; Bourgis y col.2011	Piruvato Deshidrogenasa Beta(PDHB)
KG92	Metabolismo de lípidos	PO,POP	*V. vinifera*	XM_002270031	NCBI- Nucleótidos	Proteina multifuncional MFP-a glioxisomal Beta oxidación de ácidos grasos
KG93	Metabolismo de lípidos	IV	*V. vinifera*	XR_785901.1	NCBI- Nucleótidos	Proteina acil ACP desaturasa LOC100252677
KG94	Metabolismo de lípidos	IV	*V. vinifera*	XM_010652471.1	NCBI- Nucleótidos	Proteina acil ACP desaturasa, cloroplasto
KG95	Metabolismo de lípidos	IV	*V. vinifera*	XM_002274616.2	NCBI- Nucleótidos	Proteina acil ACP desaturasa, cloroplasto
KG96	Metabolismo de lípidos	IV	*V. vinifera*	XM_002274672.3	NCBI- Nucleótidos	Proteina acil ACP desaturasa, cloroplasto
KG109	Metabolismo de lípidos	PO,POP	*Jatropha curcas*	JC000008	NCBI- Nucleótidos; Foidl N. y col. 1996	Acyl CoA sintetasa
KG148	Metabolismo de lípidos	PO,POP	*Arabidopsis thaliana*	AY328518	NCBI-Nucleótidos; Shi L. y col.2011	UDP-L-rhamnose synthase
KG149	Metabolismo de lípidos	PO,POP	*Arabidopsis thaliana*	NM_120486	NCBI-Nucleótidos	Triacilglicerol lipasa
KG150	Metabolismo de lípidos	PO,POP	*Arabidopsis thaliana*	NM_124670	NCBI-Nucleótidos; Baud S. y col. 2007	Plastidioial piruvato kinasa 2
KG151	Metabolismo de lípidos	POP, PO, FN,	*Arabidopsis thaliana*	NM_113196	NCBI-Nucleótidos; Baud S. y col.	Plastidioial piruvato kinasa 1

		FW			2007	
KG152	Metabolismo de lípidos	POP, PO, FN, FW	*Arabidopsis thaliana*	NM_120309	NCBI-Nucleótidos; Holman TJ y col., 2009	Proteolisis 6
KG211	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	XM_010926998	Dussert S. y col.2013	Palmitoil-ACP- tioesterasa
KG214	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	XM_010924626	Dussert S. y col.2013	Factor de transcripción WRI1 sensible a etileno
KG215	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	XM_010924633	Dussert S. y col.2013	Factor de transcripción WRI2 sensible a etileno
KG217	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	JN003473	Dussert S. y col.2013	Fosfolípido: diacilglicerol aciltransfersa (PDAT)
KG272	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	XM_010926998	NCBI- Nucleótidos	Palmitoil ACP tioesterasa
KG276	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	XM_010910815	NCBI- Nucleótidos	Piruvato kinasa
KG282	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	XM_010928444.1	NCBI- Nucleótidos	Oleoil-ACP- tioesterasa
P11	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	HC302343	Patente WO2009153208 (Sanz Molinero A.I. y col.2009)	Palmitoil ACP tioesterasa
P12	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	HB775649	EP2090662 (A2) (Puzio Piotr y col. 2009)	Beta ketoacil acp sinthasa III
P20	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	HB785098	EP2090662 (A2) (Puzio Piotr y col. 2009)	Beta ketoacil acp sinthasa II
P23	Metabolismo de lípidos	IV	*Elaeis guineensis*	HB701773	EP2090662 (A2) (Puzio Piotr y col. 2009)	Esteroil-ACP- desaturasa
P44	Metabolismo de lípidos	IV	*Elaeis guineensis*	E38843	JP2000270868 (A) (Murase M. y col., 2000)	Ácido Delta 12 desaturasa(fad 2)
P46	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	BD082755	JP2001314189 (A) (Murase M. y col. 2001)	Oleoil ACP tioesterasa
P60	Metabolismo de lípidos	PO,POP	*Oryza sativa*	JN944357	US20130066088 (A) (Ooi Eng KT y col. 2013)	metil tetrahidro terol-5- tri glutamato homocisteina metil transferasa
P62	Metabolismo de lípidos	PO,POP	*Populues trichocarpa*	XM_002317612	US20130066088 (A) (Ooi Eng KT y col. 2013)	Actina 6
P63	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	AY550991	US20130066088 (A) (Ooi Eng KT y col. 2013)	Actina E
P64	Metabolismo de lípidos	PO,POP	*Vanilla planifolia*	AY550991	US20130066088 (A) (Ooi Eng KT y col. 2013)	ácido cafeíco O-metiltransferasa
P65	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	JX438716	US20130066088 (A) (Ooi Eng KT y col. 2013)	Catalasa 2
P66	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	AY305853	US20130066088 (A) (Ooi Eng KT y col. 2013)	Proteína Fibrilina
P68	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	EU284941	US20130066088 (A) (Ooi Eng KT y col. 2013)	Frructosa- bifosfato aldolasa
P72	Metabolismo de lípidos	PO,POP	*Sorghum bicolor*	AF466201	US20130066088 (A) (Ooi Eng KT y col. 2013)	Proteína Metionina Sintasa
P74	Metabolismo de lípidos	PO,POP	*Oryza sativa*	NM_001054680	US20130066088 (A) (Ooi Eng KT y col. 2013)	Proteína asociada a lípido
P75	Metabolismo de lípidos	PO,POP	*Oryza sativa*	NM_001062403	US20130066088 (A) (Ooi Eng KT y col. 2013)	Desconocido
P77	Metabolismo de lípidos	PO,POP	*Oryza sativa*	AL606620	US20130066088 (A) (Ooi Eng KT y	Proteína Shock Térmico

					col. 2013)	
P78	Metabolismo de lípidos	PO,POP	*Vitis vinifera*	XM_002266561	US20130066088 (A) (Ooi Eng KT y col. 2013)	Proteína 1 complejo alfa asociado a polipeptido
P79	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	EU284905	US20130066088 (A) (Ooi Eng KT y col. 2013)	Transportadorador de proteinas
P84	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	EU284829	US20130066088 (A) (Ooi Eng KT y col. 2013)	Proteína ribosómica L10
P86	Metabolismo de lípidos	PO,POP	*Elaeis guineensis*	EU284955	US20130066088 (A) (Ooi Eng KT y col. 2013)	Lipocalina inducida por la temperatura
KG118	Metabolismo secundario	OTROS	*Jatropha curcas*	JC000059	NCBI- Nucleótidos; Foidl N. y col. 1996	Serina Carboxipeptidasa
KG82	Desarrollo de fruto	BN, BW,FN,FW	*V. vinifera*	XM_002272524	NCBI-Nucleótidos; Bötter C. y col. 2010	indole-3-acetic acid amido synthetase activity
KG290	Producción de aceite	PO, POP	*Elaeis guineensis*	AY182168	NCBI- Nucleótidos	opsc112 protein disulphide isomerase

Tabla 5.2: Homologías de las secuencias de origen mediante Blastx en *Elaeis guineensis* y su anotación GO.

NOMBRE	CARÁCTER	BLASTX	E-VALOR	ANOTACIÓN GO
KG275	POP, PO	XP_010934129.1 16 kDaOleosin	4,00E-117	C:Membrana; C:Intracelular
P4	Bn, FFB, FN,FW	NP_001290521.1 factor de transcripcion MADS-box 14	3,00E-138	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Desarrollo floral; P:Diferenciación celular; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
P39	OTROS	ABB72845.1 NAC protein 1	0,00E+00	C:Núcleo; F:Unión ADN; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
P58	OTROS	XP_010921345.1 Peroxirredoxin 1-Cys	3,00E-136	C:Núcleo; P:Respuesta a estímulos abióticos; F:Actividad catalítica; P:Desarrollo post-embrionario; P:Respuesta a estrés; P:Reproducción
KG269	OTROS	XP_010922868.1 Proteína 68 contiene dominio NAC	0,00E+00	C:Núcleo; F:Unión ADN; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG210		XP_010917338.1 PREDICTED: uncharacterized protein LOC105041961	0,00E+00	P:Proceso Metabolismo de Lípidos; F:Actividad Hidrolasa//P:Proceso Metabolismo de Lípidos; F:Actividad Hidrolasa
KG212	POP,PO	XP_010915014.1 Proteína Palmitoil-ACP-tioesterasa	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; C:Plastidio; P:Proceso celular; F:Actividad Hidrolasa
KG213	POP,PO	XP_010925110.1 Proteína Palmitoil-ACP-tioesterasa	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; P:Proceso celular; F:Actividad Hidrolasa
KG274	FW,FN	XP_010926267.1 Lipoil sintasa, mitocondria	0,00E+00	P:Proceso metabólico secundario; P:Proceso catabólico; P:Desarrollo de organismos multicelulares; P:Proceso de modificación de proteínas celularess; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; F:Unión; P:Proceso metabólico compuesto por una nucleobase; C:Mitocondria; C:Membrana; P:Proceso Metabolismo de Lípidos; P:Diferenciación celular; F:Actividad Transferasa
KG153	FW,FN	XP_010942026.1 Proteína de unión nucleótido guanina subunidad BETA	0,00E+00	P:Respuesta a estímulos extracelulares; P:Proceso catabólico; P:Desarrollo post-embrionario; P:Proceso de modificación de proteínas celularess; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; F:Unión de Nucleótidos; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Muerte celular; C:Retículo endoplasmático; F:Actividad transductora de la señal; P:Proceso metabólico compuesto por una nucleobase; P:Transportador; C:Intracelular; C:Membrana plasmática; P:Respuesta a estrés; P:Proceso de

				metabolismo de carbohidratos; F:Unión de proteínas; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; F:Actividad Hidrolasa
KG154	FW,FN	XP_010910305.1 Homólogo a DDB1 y CUL4 factor asociado 1	0,00E+00	P:Desarrollo embrión; P:Transducción de señales; P:Proceso de modificación de proteínas celularess; P:Morfogenésis de la estructura anatómica; P:Ciclo celular; P:Respuesta a estrés; F:Unión de Nucleótidos; C:Núcleo; P:Respuesta a estímulos abióticos; F:Unión de proteínas; P:Desarrollo floral; P:Organización de componentes celulares; P:Diferenciación celular
KG155	FW,FN	XP_010937674.1 PROTEINA BAJA CALIDAD: complejo elongador 1	0,00E+00	P:Transducción de señales; P:Desarrollo post-embrionario; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; P:Proceso metabólico ADN; P:Respuesta a estrés; P:Reproducción; P:Crecimiento; C:Núcleo; P:Respuesta a estímulos endógenos; P:Diferenciación celular; C:Citosol; F:Actividad Transferasa
KG242	FW,FN	XP_010921598.1 PREDICTED: ethylene receptor-like isoform X3	1,00E-57	C:Citoplasma; C:Membrana; P:Respuesta a estímulos endógenos; P:Proceso de modificación de proteínas celularess; C:Retículo endoplasmático; F:Actividad receptora; F:Unión; F:Actividad Kinasa; F:Actividad transductora de la señal
KG244	FW,FN	XP_010931693.1 PREDICTED: ethylene receptor-like	3,00E-59	C:Citoplasma; C:Membrana; P:Respuesta a estímulos endógenos; P:Proceso de modificación de proteínas celularess; C:Retículo endoplasmático; F:Actividad receptora; F:Unión; F:Actividad Kinasa; F:Actividad transductora de la señal
KG252	FW,FN	XP_010913671.1 AP2 Factor de transcripción sensible a etileno AIL5	0,00E+00	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; P:Desarrollo post-embrionario; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG253	FW,FN	XP_010911149.1 AP2 Factor de transcripción sensible a etileno ANT	0,00E+00	P:Regulación de la transcripción, muestra de ADN; P:Transcripción, muestra de ADN; P:Desarrollo de organismos multicelulares; C:Núcleo; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión ADN; F:Actividad [NAD(P)] 2 alquenal reductasa ; P:Proceso óxido-reducción; F:Actividad oxidoreductasa
KG254	FW,FN	XP_010913038.1 PREDICTED:Factor de transcripción sensible a etileno ESR2	3,00E-94	P:Desarrollo embrión; C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; P:Desarrollo post-embrionario; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; P:Ciclo celular
KG270	FW,FN	XP_010934580.1 Posible fofogluconolactonasa-6 4	0,00E+00	P:Proceso de metabolismo de carbohidratos; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio; F:Actividad Hidrolasa
KG271	FW,FN	XP_010931211.1 Factor de transcripción MYB75	2,00E-95	F:Unión ADN
KG288	FW,FN	-		F:Unión de ácidos nucléicos; P:Integración ADN
KG114	FW,FN	XP_010922188.1 Zerumbone sintasa	3,00E-62	P:Proceso metabólico secundario; P:Respuesta a estímulos abióticos; P:Transducción de señales; P:Proceso Metabolismo de Lípidos; C:Citosol; P:Proceso biosintético; F:Actividad Transferasa; P:Respuesta a estrés
KG87	OTROS	XP_010918863.1 Posible Indol-3- piruvato monooxigenasa YUCCA5	4,00E-122	F:Unión de Nucleótidos; F:Actividad catalítica
KG113	OTROS	XP_010939830.1 Peptidil-propil cis-trans-isomerasa NIMA-4	3,00E-42	F:Actividad catalítica; P:Proceso de modificación de proteínas celularess
KG106	OTROS	YP_006073132.1 Proteína N fotosistema II	7,00E-24	C:Membrana; P:Fotosíntesis; C:Plastidio; C:Tilacoide
KG107	OTROS	YP_006073121.1 Subunidad alfa citocromo b559 fotosistema II (cloroplasto)	3,00E-53	P:Generación de metabolitos precursores y energía; C:Membrana; P:Fotosíntesis; F:Unión; C:Plastidio; C:Tilacoide
KG108	OTROS	YP_006073112.1 Ribulosa bifosfato carboxilasa (cloroplasto)	0,00E+00	P:Proceso de metabolismo de carbohidratos; F:Actividad catalítica; P:Fotosíntesis; P:Proceso biosintético; F:Unión; C:Plastidio
KG278	OTROS	XP_010918840.1 Subunidad IV B centro de reacción Fotosistema I, cloroplasto	1,00E-40	C:Membrana; P:Fotosíntesis; C:Plastidio; C:Tilacoide
KG115	OTROS	XP_010911246.1 Enzima de procesamiento vacuolar	4,00E-59	P:Proceso metabólico de proteínas; F:Actividad Hidrolasa
KG117	OTROS	XP_010914501.1 Receptor de etileno	5,00E-61	P:Desarrollo post-embrionario; P:Proceso de modificación de proteínas celularess; C:Retículo endoplasmático; P:Proceso biosintético; F:Actividad receptora; F:Actividad transductora de la señal; P:Respuesta a estrés;

				F:Unión de Nucleótidos; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; F:Unión de proteínas; P:Respuesta a estímulos externos; C:Citoplasma; C:Membrana; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; F:Actividad Kinasa
KG101	OTROS	YP_006073092.1 ATP sintasa CF0 subunidad IV (cloroplasto)	2,00E-150	P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; C:Membrana plasmática; C:Plastidio; F:Actividad Transportadoradora; C:Tilacoide; F:Actividad Hidrolasa
KG102	OTROS	YP_006073110.1 ATP sintasa CF1 subunidad epsilon (cloroplasto)	2,00E-79	F:Unión de Nucleótidos; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; C:Membrana plasmática; C:Plastidio; F:Actividad Transportadoradora; C:Tilacoide; F:Actividad Hidrolasa
KG103	OTROS	YP_006073091.1 ATP sintasa CF0 subunidad III (cloroplasto)	3,00E-34	C:Membrana; F:Unión de lípidos; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio; F:Actividad Transportadoradora; C:Tilacoide
KG105	OTROS	YP_006073133.1 Fosfoproteína fotosistema II	2,00E-44	C:Membrana; P:Fotosíntesis; F:Unión; C:Plastidio; C:Tilacoide
KG156	HI	XP_010940324.1delta(24)-esterol reductasa	0,00E+00	P:Proceso metabólico secundario; P:Proceso catabólico; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; C:Vacuola; C:Membrana plasmática; F:Unión de Nucleótidos; C:Núcleo; P:Proceso de metabolismo de carbohidratos; P:Respuesta a estímulos abióticos; F:Unión de proteínas; P:Respuesta a estímulos externos; F:Actividad catalítica; P:Desarrollo floral; P:Proceso Metabolismo de Lípidos; P:Organización de componentes celulares; P:Diferenciación celular; P:Crecimiento celular
KG157	HI	XP_010921202.1 Proteína ribosómica L10 60S	2,00E-141	F:Actividad molecular estructural; C:Aparato de Golgi; C:Nucleolo; C:Ribosoma; C:Vacuola; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio; P:Translación; P:Respuesta a estímulos abióticos; F:Unión de proteínas; F:RNA Unión; C:Membrana; P:Organización de componentes celulares; C:Citosol
KG159	HI	XP_010924059.1 Probable Amido sintetasa de ácido indol-3-acético GH3.5	0,00E+00	P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; F:Actividad catalítica; P:Transducción de señales; P:Respuesta a estímulos endógenos; C:Plastidio; P:Respuesta a estrés
KG160	HI	XP_010929776.1 Homologo a componente SNF12 Complejo SWI/SNF	0,00E+00	P:Proceso catabólico; F:Actividad molecular estructural; C:Aparato de Golgi; P:Proceso metabólico ADN; C:Vacuola; P:Transportador; C:Membrana plasmática; P:Respuesta a estrés; C:Núcleo; P:Respuesta a estímulos abióticos; F:Unión de proteínas; P:Desarrollo floral; C:Citosol
KG233	HI	XP_010940408.1 Proteína tallo específica TSJT1	0,00E+00	C:Núcleo; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; P:Transducción de señales; P:Proceso metabólico; C:Citosol; C:Membrana plasmática; P:Respuesta a estrés
KG234	HI	_		-
KG245	HI	XP_010907279.1 PREDICTED: EIN3-Unión F-box protein 1-like	3,00E-108	-
KG246	HI	XP_010917603.1 PREDICTED: auxin-induced protein 22D-like	8,00E-50	C:Núcleo; P:Transducción de señales; P:Respuesta a estímulos endógenos; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG247	HI	AEQ94168.1 ubiquitin-protein ligase 3 HECT domain	5,00E-77	F:Actividad catalítica
KG248	HI	XP_010922067.1 PREDICTED: probable auxin efflux carrier component 1b	7,00E-71	C:Componente integral de membrana; P:Transportador transmembrana
KG249	HI	XP_010938861.1 PREDICTED: gibberellin receptor GID1C-like	1,00E-64	P:Transducción de señales; P:Proceso de modificación de proteínas celularess; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; P:Transportador; C:Intracelular; P:Respuesta a estrés; P:Respuesta a estímulos bióticos; P:Proceso de metabolismo de carbohidratos; P:Respuesta a estímulos abióticos; P:Muerte celular; P:Respuesta a estímulos externos; P:Desarrollo floral; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; F:Actividad Hidrolasa
KG250	HI	XP_010932634.1 PREDICTED: probable auxin efflux carrier component 1c isoform X1	2,00E-65	C:Membrana; P:Transportador; P:Proceso celular
KG251	HI	XP_010942800.1 estrigolactona esterasa DAD2	1,00E-	P:Proceso metabólico secundario; C:Citoplasma; P:Desarrollo de organismos multicelulares; P:Proceso

			137	Metabolismo de Lípidos; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; P:Transportador; P:Respuesta a estrés; P:Proceso celular; F:Actividad Hidrolasa
KG286	HI	XP_010926472.1 Proteína SLR1 DELLA	0,00E+00	C:Núcleo; F:Unión de proteínas; C:Plastidio; P:Proceso celular
KG289	HI	XP_010925034.1 ácido carboxilico-1-aminociclopropano 1 oxidasa	0,00E+00	F:Actividad catalítica; F:Unión
P13	HI	XP_010938052.1Proteína ribosómica 60S L10	1,00E-167	F:Actividad molecular estructural; C:Ribosoma; P:Translación
KG64	HI	XP_010909244.1 Posible auxina portadora de componente de flujo 1c	0,00E+00	P:Desarrollo embrión; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; F:Actividad Transportadoradora; P:Respuesta a estímulos bióticos; P:tropism; P:Respuesta a estímulos abióticos; P:Desarrollo floral; P:Transducción de señales; C:Retículo endoplasmático; P:Proceso metabólico compuesto por una nucleobase; C:Membrana plasmática; F:Unión de proteínas; C:Pared celular; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares
KG69	HI	XP_010936898.1 Proteína serina/treonina kinasa BRI1	0,00E+00	P:Proceso de modificación de proteínas celularess; F:Unión de lípidos; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; F:Unión de Nucleótidos; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; C:Endosoma; P:Respuesta a estímulos externos; P:Muerte celular; P:Desarrollo floral; F:Actividad receptora; F:Actividad transductora de la señal; C:Membrana plasmática; P:Respuesta a estrés; P:Proceso de metabolismo de carbohidratos; F:Unión de proteínas; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; P:Diferenciación celular; F:Actividad Kinasa; P:Crecimiento celular
KG70	HI	AIC09100.1 Proteina Kinasa receptora de regiones ricas en leucina	0,00E+00	P:Proceso de modificación de proteínas celularess; F:Actividad receptora; F:Actividad transductora de la señal; F:Unión de receptores; P:Transportador; C:Membrana plasmática; P:Respuesta a estrés; F:Unión de Nucleótidos; P:Respuesta a estímulos bióticos; C:Endosoma; P:Respuesta a estímulos externos; P:Muerte celular; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; F:Actividad Kinasa; P:Crecimiento celular
KG75	HI	XP_010904836.1 Receptor Giberelina GIDC	2,00E-173	P:Transducción de señales; P:Proceso de modificación de proteínas celularess; P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; P:Transportador; C:Intracelular; P:Respuesta a estrés; P:Respuesta a estímulos bióticos; P:Proceso de metabolismo de carbohidratos; P:Respuesta a estímulos abióticos; P:Muerte celular; P:Respuesta a estímulos externos; P:Desarrollo floral; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; F:Actividad Hidrolasa
KG78	HI	XP_010921594.1 Receptor etileno probable	0,00E+00	P:Desarrollo post-embrionario; P:Proceso de modificación de proteínas celularess; C:Retículo endoplasmático; P:Proceso biosintético; F:Actividad receptora; F:Actividad transductora de la señal; P:Respuesta a estrés; F:Unión de Nucleótidos; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; F:Unión de proteínas; P:Respuesta a estímulos externos; C:Citoplasma; C:Membrana; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares; F:Actividad Kinasa
KG80	HI	XP_010918767.1 Posible Proteína EIN-4	0,00E+00	F:Unión de Nucleótidos; C:Citoplasma; C:Membrana; P:Respuesta a estímulos endógenos; P:Proceso de modificación de proteínas celularess; C:Retículo endoplasmático; F:Actividad receptora; F:Actividad Kinasa; F:Actividad transductora de la señal; P:Respuesta a estrés
KG81	HI	XP_010933099.1 Proteína EIN-4	0,00E+00	P:Proceso de modificación de proteínas celularess; C:Retículo endoplasmático; P:Proceso biosintético; F:Actividad receptora; F:Actividad transductora de la señal; P:Proceso metabólico compuesto por una nucleobase; F:Unión de Nucleótidos; F:Unión de proteínas; C:Citoplasma; C:Membrana; P:Respuesta a estímulos endógenos; F:Actividad Kinasa
KG243	OTROS	XP_010909152.1 PREDICTED: coronatine-insensitive protein 1	0,00E+00	P:Respuesta a estímulos bióticos; P:Proceso metabólico de proteínas; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; P:Transducción de señales; P:Desarrollo floral; P:Proceso catabólico; P:Respuesta a estímulos endógenos; P:Respuesta a estrés

KG268	FW,FN	XP_010913017.1 Homólogo a proteína AGL9 MADS-box agamous	2,00E-172	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG119	FW,FN	CAE46185.1 Factor de transcripción MADS-box AP1	2,00E-130	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG120	FW,SH	CAE46181.1 Factor de transcripción MADS-box AGAMOUS [Elaeis guineensis]	2,00E-30	P:Morfogenésis de la estructura anatómica; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; P:Transportador; P:Polinización; F:Unión de Nucleótidos; C:Núcleo; F:Unión ADN; P:Respuesta a estímulos bióticos; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Respuesta a estímulos externos; P:Desarrollo floral; P:Organización de componentes celulares; P:Diferenciación celular; P:Crecimiento celular; F:Actividad Hidrolasa
KG121	FW,FN	AAW66885.1 Factor de transcripcion MADS BOX	1,00E-162	P:Proceso de modificación de proteínas celularess; P:Morfogenésis de la estructura anatómica; P:Respuesta a estímulos bióticos; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; P:Muerte celular; P:Desarrollo floral; P:Transducción de señales; P:Proceso metabólico compuesto por una nucleobase; P:Transportador; P:Ciclo celular; P:Respuesta a estrés; P:Translación; C:Núcleo; F:Unión ADN; F:Actividad reguladora de la translación; F:Unión de proteínas; P:Respuesta a estímulos endógenos; P:Organización de componentes celulares
KG122	FW,FN	NP_001290512.1 Factor de transcripcion MADS-box 16	2,00E-155	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG123	FW,FN	XP_010905324.1 Factor de transcripcion MADS-box 2	4,00E-131	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG124	FW,FN	XP_010915441.1 Factor de transcripcion MADS-box 3	1,00E-174	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG125	FW,FN	AAQ03223.1Factor de transcripcion MADS BOX	3,00E-142	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG126	FW,FN	XP_010917559.1 Proteína homóloga AGL-9 AGAMOUS MADS-Box	2,00E-173	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG127	FW,FN	CAE46188.1 Factor de transcripción MADS-Box AGL2	1,00E-142	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG128	FW,FN	NP_001290521.1 Factor de transcripcion MADS-box 14	4,00E-131	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; F:Unión de proteínas; P:Desarrollo floral; P:Diferenciación celular; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG24	POP,PO	XP_010915014.1 Proteina Palmitoil-acil tioesterasa	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; P:Proceso celular; F:Actividad Hidrolasa
KG27	POP,PO	XP_010932373.1 3-oxoacyl-[ACP] sintasa II	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Actividad Transferasa; P:Proceso celular
KG29	PO, BW	AAF69015.1 glutelin	0,00E+00	F:Función Molecular
KG31	POP,PO	YP_006073111.1 ATP sintasa CF1 subunidad beta	0,00E+00	F:Unión de Nucleótidos; C:Membrana; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio; F:Actividad Transportadoradora; C:Tilacoide; F:Actividad Hidrolasa
KG35	PO, IV	XP_010920844.1 Ácido graso omega 3 desaturasa	0,00E+00	F:Actividad catalítica; P:Proceso Metabolismo de Lípidos
KG38	PO	NP_001291356.1 proteina no caracteriza LOC105049617	1,00E-150	P:Respuesta a estímulos abióticos; P:Desarrollo de organismos multicelulares; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Actividad Transferasa; C:Plastidio; P:Proceso celular; P:Respuesta a estrés
KG39	PO, IV	AAD33904.1 ketoacyl sintasa III	2,00E-131	P:Respuesta a estímulos abióticos; P:Desarrollo de organismos multicelulares; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Actividad Transferasa; C:Plastidio; P:Proceso celular; P:Respuesta a estrés
KG47	PO, IV	XP_010941056.1 PREDICTED: peroxisomal acil-coenzima A oxidasa 1	3,00E-149	F:Unión de Nucleótidos; C:Peroxisoma; F:Actividad catalítica; P:Proceso catabólico; P:Proceso Metabolismo de Lípidos; P:Proceso celular
KG57	FW, PO	AEW31333.1 Proteína Plastidioio Piruvato deshidrogenasa	0,00E+00	F:Actividad catalítica

		Beta E1		
KG92	PO,POP	XP_010931678.1 Proteina multifuncional MFP-a glioxisomal Beta oxidación de ácidos grasos	1,00E-110	P:Proceso metabólico secundario; P:Proceso metabólico de proteínas; F:Actividad catalítica; P:Proceso catabólico; C:Pared celular; P:Proceso Metabolismo de Lípidos; P:Organización de componentes celulares; C:Nucleolo; P:Respuesta a estrés
KG93	IV	XP_010927705.1 Proteína acil-(ACP)-desaturasa,cloroplasto	1,00E-137	F:Actividad catalítica; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; C:Plastidio; P:Proceso celular
KG94	IV	XP_010926734.1PROTEÍNA BAJA CALIDAD: acil-(ACP) -desaturasa, cloroplasto	2,00E-45	F:Actividad catalítica; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Unión; C:Plastidio; P:Proceso celular
KG95	IV	XP_010929790.1 Proteína acil-(ACP) desaturasa 5, cloroplasto	2,00E-103	F:Actividad catalítica; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; C:Plastidio; P:Proceso celular
KG96	IV	XP_010927705.1 Proteína acil-(ACP)-desaturasa,cloroplasto	8,00E-107	F:Actividad catalítica; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Unión; C:Plastidio; P:Proceso celular
KG109	PO,POP	XP_010924575.1 malonato--CoA ligasa	4,00E-54	C:Núcleo; F:Actividad catalítica; P:Proceso catabólico; C:Membrana; C:Citosol; P:Proceso celular
KG148	PO,POP	XP_010915435.1 PREDICTED: trifunctional UDP-glucose 4,6-dehydratase/UDP-4-keto-6-deoxy-D-glucose 3,5-epimerase/UDP-4-keto-L-rhamnose-reductase RHM1	0,00E+00	P:Proceso de metabolismo de carbohidratos; F:Actividad catalítica; P:Desarrollo post-embrionario; C:Citosol; P:Proceso biosintético; F:Unión; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio
KG149	PO,POP	XP_010922190.1 triacilglicerol lipasa SDP1	0,00E+00	P:Proceso catabólico; F:Actividad receptora; F:Actividad transductora de la señal; C:Intracelular; F:Actividad Transportadoradora; P:Proceso de metabolismo de carbohidratos; C:Membrana; P:Proceso Metabolismo de Lípidos; P:Organización de componentes celulares; F:Actividad Hidrolasa
KG150	PO,POP	XP_010924790.1 Plastidioial piruvato kinasa 2	0,00E+00	P:Generación de metabolitos precursores y energía; P:Proceso catabólico; P:Desarrollo post-embrionario; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio; P:Reproducción; F:Unión de Nucleótidos; P:Proceso de metabolismo de carbohidratos; C:Mitocondria; P:Proceso Metabolismo de Lípidos; C:Citosol; F:Actividad Kinasa
KG151	POP, PO, FN, FW	XP_010923501.1 Isozima A Piruvato kinasa	0,00E+00	P:Generación de metabolitos precursores y energía; P:Proceso catabólico; P:Desarrollo post-embrionario; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase; C:Plastidio; P:Reproducción; F:Unión de Nucleótidos; P:Proceso de metabolismo de carbohidratos; P:Proceso Metabolismo de Lípidos; C:Citosol; F:Actividad Kinasa; P:Crecimiento celular
KG152	POP, PO, FN, FW	XP_010937621.1 Proteína no caracterizada LOC105056937 isoform X1	0,00E+00	P:Proceso catabólico; P:Desarrollo post-embrionario; P:Proceso de modificación de proteínas celularess; P:Proceso biosintético; F:Unión; C:Núcleo; C:Citoplasma; P:Respuesta a estímulos endógenos; P:Proceso Metabolismo de Lípidos; F:Actividad Transferasa
KG211	PO,POP	XP_010925300.1 Proteína Palmitoil-ACP-tioesterasa	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; C:Plastidio; P:Proceso celular; F:Actividad Hidrolasa
KG214	PO,POP	XP_010922928.1 Factor de transcripción WRI1 sensible a etileno	0,00E+00	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; P:Desarrollo de organismos multicelulares; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG215	PO,POP	XP_010922935.1 Factor de transcripción WRI1 sensible a etileno	0,00E+00	C:Núcleo; F:Unión ADN; F:Actividad factor de transcripción, unión de secuencias específicas de ADN; P:Desarrollo de organismos multicelulares; P:Proceso biosintético; P:Proceso metabólico compuesto por una nucleobase
KG217	PO,POP	AEQ94116.1 Aciltransferasa fosfolípido- diacilglicerol Putativa	2,00E-115	P:Proceso Metabolismo de Lípidos; C:Retículo endoplasmático; F:Actividad Transferasa; C:Vacuola
KG272	PO,POP	XP_010925300.1 Palmitoil ACP tioesterasa	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; C:Plastidio; P:Proceso celular; F:Actividad Hidrolasa
KG276	PO,POP	XP_010909117.1Piruvato kinasa,isozima citosólico	0,00E+00	P:Proceso de metabolismo de carbohidratos; P:Generación de metabolitos precursores y energía; P:Proceso catabólico; F:Unión; F:Actividad Kinasa; P:Proceso metabólico compuesto por una nucleobase
KG282	PO,POP	XP_010926746.1 Oleoil-ACP-tioesterasa, cloroplasto	0,00E+00	C:Mitocondria; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; P:Proceso celular; F:Actividad

				Hidrolasa
P11	PO,POP	AAD42220.2 Proteína palmitoil ACP tioesterasa	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; C:Plastidio; P:Proceso celular; F:Actividad Hidrolasa
P12	PO,POP	AAF26738.2 beta-ketoacil-ACP sintasa II	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Actividad Transferasa; P:Proceso celular
P20	PO,POP	AAF26738.2 beta-ketoacyl-ACP sintasa II	0,00E+00	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Actividad Transferasa; P:Proceso celular
P23	IV	AAB41041.1 Esteroil-Acp-desaturasa	0,00E+00	F:Actividad catalítica; P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; F:Unión; C:Plastidio; P:Proceso celular
P44	IV	XP_010928433.1 omega-6 ácido graso desaturasa, isozima 2 reticulo endoplasmático	0,00E+00	F:Actividad catalítica; C:Membrana; P:Proceso Metabolismo de Lípidos
P46	PO,POP	XP_010926746.1 oleoil-ACP tioesterasa, cloroplástica	1,00E-135	P:Proceso Metabolismo de Lípidos; P:Proceso biosintético; P:Proceso celular; F:Actividad Hidrolasa
P60	PO,POP	XP_010908721.1 metiltetrahidroterol-5 triglutamato metiltransferasa homocisteína	0,00E+00	C:Citosol; P:Proceso biosintético; F:Actividad Transferasa; F:Unión; P:Proceso celular
P62	PO,POP	XP_010930993.1Actina-101	0,00E+00	F:Unión de Nucleótidos
P63	PO,POP	XP_010911969.1 Actina 3	0,00E+00	F:Unión de Nucleótidos
P64	PO,POP	XP_010934914.1 PREDICTED: 3-ácido cafeíco O-metiltransferasa	4,00E-173	P:Proceso metabólico secundario; F:Unión de proteínas; P:Proceso biosintético; F:Actividad Transferasa; P:Proceso celular
P65	PO,POP	XP_010934843.1 isozima Catalasa 2	0,00E+00	C:Peroxisoma; F:Actividad catalítica; P:Proceso catabólico; C:Mitocondria; F:Unión; P:Respuesta a estrés; P:Proceso celular
P66	PO,POP	AAP74338.1 Proteína Fibrilina	0,00E+00	F:Actividad catalítica; C:Plastidio
P68	PO,POP	XP_010927580.1 Fructosa-bifosfato aldolasa isozima citoplasmático	0,00E+00	P:Proceso de metabolismo de carbohidratos; P:Generación de metabolitos precursores y energía; C:Citoplasma; F:Actividad catalítica; P:Proceso catabólico; P:Proceso metabólico compuesto por una nucleobase
P72	PO,POP	XP_010908721.1 Homocisteína metiltransferasa 1, 5-metiltetrahidropteoiltriglutamato	0,00E+00	C:Citosol; P:Proceso biosintético; F:Actividad Transferasa; F:Unión; P:Proceso celular
P74	PO,POP	XP_010937482.1 Proteína no caracterizada LOC105056845	0,00E+00	C:Citoplasma
P75	PO,POP	XP_010922826.1 Fosfoglicerato mutasa 2,3 - difosfoglicerato - independient	0,00E+00	P:Proceso de metabolismo de carbohidratos; P:Generación de metabolitos precursores y energía; C:Citoplasma; F:Actividad catalítica; P:Proceso catabólico; F:Unión; P:Proceso metabólico compuesto por una nucleobase
P77	PO,POP	XP_010934626.1Proteína Shock térmcio 83	0,00E+00	F:Unión de Nucleótidos; P:Proceso metabólico de proteínas; P:Respuesta a estímulos bióticos; F:Unión de proteínas; P:Respuesta a estímulos externos; C:Citoplasma; P:Proceso celular; P:Respuesta a estrés
P78	PO,POP	XP_010921860.1 proteína naciente -alfa como subunidad complejo polipéptido - 1 asociado	4,00E-95	-
P79	PO,POP	XP_010938541.1 Coatomero subunidad epsilon-1	0,00E+00	F:Actividad molecular estructural; P:Transportador; C:Intracelular
P84	PO,POP	XP_010941551.1Proteína P0 ácida ribosómica 60S	0,00E+00	P:Proceso biológico; C:Ribosoma
P86	PO,POP	XP_010929132.1 Apolipoproteína D	1,00E-129	P:Respuesta a estímulos abióticos; C:Mitocondria; C:Retículo endoplasmático; C:Aparato de Golgi; C:Vacuola; C:Membrana plasmática; F:Actividad Transportadoradora; P:Respuesta a estrés
KG118	OTROS	XP_010918615.1 Serina carboxipeptidasa	3,00E-76	P:Proceso metabólico de proteínas; C:Citosol; C:Vacuola; F:Actividad Hidrolasa
KG82	BN, BW,FN,FW	XP_010924059.1 PREDICTED: probable indole-3-acetic acid-amido synthetase GH3.5	3,00E-139	P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; F:Actividad catalítica; P:Transducción de señales; P:Respuesta a estímulos endógenos; C:Plastidio; P:Respuesta a estrés
KG290	PO, POP	AAO26314.1 protein disulphide isomerase, partial	0,00E+00	P:Proceso metabólico de proteínas; P:cellular homeostasis; C:Retículo endoplasmático; C:Vacuola; C:Plastidio; P:Respuesta a estrés; P:Respuesta a estímulos bióticos; P:Respuesta a estímulos abióticos; P:Respuesta a estímulos externos; C:Membrana; F:Actividad Transferasa

Anexo 6: Cebadores de los GC secuenciados.

Tabla 6.1: Cebadores de los GC secuenciados. GC=nombre del gen candidato; ORIGEN: muestra la procedencia del GC. TDF=transcripto obtenido del experimento de cDNA-AFLP.CO-LOC= genes candidatos co-localizados en el mapa integrado de alta densidad. GEN CONOCIDO= genes con funciones conocidas procedentes de bibliografía publicada, rutas biosintéticas y patentes; CARÁCTER: muestra el carácter con el que se ha relacionado en origen. COPIAR; NOMBRE CEBADOR=nombre dado a la pareja de cebadores; CEBADOR Fw= cebador derecho o forward; CEBADOR Rv= cebador izquierdo o reverse; TAMAÑO=tamaño del amplicon en pares de bases. La tabla no muestra la temperatura de fusión de cada cebador (Tm) ya que fueron sintetizados junto con las secuencias universales UNIA el Fw y UNIB el Rv.

GC	ORIGEN	CARÁCTER	NOMBRE CEBADOR	CEBADOR Fw	CEBADOR Rv	TAMAÑO
CDA39	TDF	HT	63	GGAATCGAGAGCTCCAAGTG	GACTTCCATGACACCGTCCT	152
CDA40	TDF	HT	64	CATCACCTTTTGCGGGTAGT	TTCGATCAACATCTGCCATC	182
CDA76	TDF	OM	B100	CATTACTCCGATCCCGAAGG	TGCTCTCCTGTCCTTAACTGG	157
CDA78	TDF	OM	B102	ACCAGGTAGCACATCCTCCA	GACCGATCTCTCGTGGATTC	194
CDA34	TDF	MF	B57b	GGACGGGTGAGTAATGCCTA	GGCCTTTACCCCACCAACTA	162
CDA37	TDF	MF	61	CTGCAAAAATCGTAGAAACAAA	GGTCCACCGTGATCTCAAACT	152
CDA3	TDF	CPO	B11	AGCCTTTTGAGGTGTGGTGT	CAATAGTGCCCGAAAAGACC	210
CDA4	TDF	CPO	B3/B12	CAAGAGTTGCCCTGGATGTT	CGTGTTGCAAAAGCTGCTAC	166
CDA9	TDF	CPO	B25	GATGTGCATGGTCCATTTGA	TGATCGAAGTACCTTTAGGATTTTT	185
CDA22	TDF	FW	B42	TCAGCTAATGAGGCGTGATG	TAAACGGCCACCTCCACTTA	175
CDA24	TDF	FW	B44	TCTCTCCTTTCCACCGAAAA	GAAGGGGGAGAAGAGAGGAA	157
CDA26	TDF	FW	B46	ACGAGGGACTGCCAGTGA	ACAATCCGAACTGAGGCAAT	160
CDA31	TDF	FW	B54	GCTAAGCGATCTGCCGAAG	GTTTTCGGGGCATTGGAT	160
CDA32	TDF	FW	B56	CCCCAGGTCCCTCAGAGTA	GGGAGGATCTAAAGGACCCTAA	157
CDA6	TDF	BW	B18	CTCAAGCTATGCATCCAACG	AACAGTCGGATTCCCCTTG	137
CDA15	TDF	BW	B33	TCCGATTCACCTTTTCAAGC	CTGCCCCTATGGTACATGCT	160
CDA27	TDF	BW	B47	AGATTACCCGGGCCTGTC	AACGAGACCTCAGCCTGCTA	199
CDA43	TDF	BN	67	TGTTCATCAACAACAAATCACG	TTGTTTACTGATCTCTCGAATGG	164
KG144	CO-LOC	HI	M23551	TACGGCATCGTTCTCATCAA	TGGTGGAGTTGGAGTGTCAG	199

KG145	CO-LOC	HI	M7495	CCTTTGTGCCTCTTGTACTTGG	CTGCCAGATCTGCGGTGA	180
KG147	CO-LOC	HI	M4883	TCCTGTTCAGGCTCTGGAAT	CAGGGTTGCGTGAATGGTT	183
KG12	CO-LOC	IV	M2200	AGGATCACCAATGCCTACAA	AGCAAGTATGTTGGCACTTCA	169
KG11	CO-LOC	FW	M8373	CGCCCAGTATTTGGATCAGT	ATTGGGCCTCTTTTGGTCTT	169
KG135	CO-LOC	IV	M3117	GCCTCGGATAAAACCCTAGC	GATCAGATGGCTGCTGTGAA	197
KG138	CO-LOC	PF	M9619	GGGCTGGGACAACTGATCTA	TCCGAAGGTTGGTCCACTAC	217
KG141	CO-LOC	BN/BW	M847	TGACCTAGTGCCACCATCAA	TAGCAACCGCACCAATGTAG	217
KG142	CO-LOC	BN/BW	M2252	CTTGCATTGAGGCTTGTTGA	AATTCTTTGTCTGGCGTTGG	196
KG143	CO-LOC	BN/BW	M3256	ATCCCAGCACTGATCTCACC	TGTCAAGGCAACTAGGAGCA	184
KG2	CO-LOC	BW	mEg3275	GAAGCCTGAGACCGCATAGA	TTCGGTGATGAAGATTGAAG	146
KG140	CO-LOC	BW	M43696	TATGATCGGGAGGGTGGTAG	GGAGTGAGGATGAGGAGCTG	236
KG161*	CO-LOC SNP	HI	HtC1_3885	ATGGCCTCTCTTCTGTCACG	ACGTTGGAGGCAAACGATAC	202
KG162*	CO-LOC	HI	HtC1_5925	AGGATGGGCTGAGGATCAC	CCTGCTCTCACAAGGCAAAT	291
KG164*	CO-LOC	HI	HtC2_11412	AGGCAGTTCAACACCATTCC	AACAAGAGAGCCTGCAAGGA	195
KG166*	CO-LOC	HI	HtC2_7081	GGAGGTGCGACAACTTCAAT	TGTGGTGACAAACTGCAGAA	230
KG167*	CO-LOC	HI	HtC2_1255C2_411	AGGCAAACTTGGAGCACAGA	TCAGAAATGTTTCCCGCATC	223
KG168*	CO-LOC	HI	HtC2_9289	TGGAACACGATGAGACTGGA	CCAAAACGGCTTTTGTCTCT	264
KG171*	CO-LOC	HI	HtC4_4489	ATCCTGCGACCAAATGGATA	TGCACTGAATTGCATATTTCC	282
KG172*	CO-LOC	HI	HtC4_240	AGACCATGGCGCTAAATCAG	CCTCCCCTCATCTGAAAAGA	283
KG173*	CO-LOC	HI	HtC7_9200	TGCAACTTTTGCTCAGGATG	TGCCAATCTTTCCCTCTCAC	136
KG174*	CO-LOC	HI	HtC7_1247	TTTCCTTCGATCGGATTCAC	CCTCGATTCCCTTCATCAAT	158
KG175*	CO-LOC	HI	HtC8_1026C1-144	CTCTGCTGTGCTGCTCTCAC	ATGCCAGAACCATTCCAGAT	277
KG176*	CO-LOC	HI	HtC8_11217	CCTCTCCAAATCCATTGCTG	GCGGCTGTATTGAAGGAGAC	259
KG177*	CO-LOC	HI	HtC10_11102	CCATTGATACCTAAAGCTGGAGA	GGACATGACCAAGCTTGAAA	300
KG196*	CO-LOC	BN	BnC2_10C3-629	TGAAGCAGGACATGAGTTTGA	TGCAACAAAAGTCATGCAAT	300
KG197*	CO-LOC	BN	BnC2_1289	TAAGGTCACAGAGGCCCAAG	CATTTTCAGCAGGCTTCACA	254

KG198*	CO-LOC	BN	BnC3_792	GACTGGGAAAGGCATCTCTCT	CGACAACCTTAACACCCACA	209
KG199*	CO-LOC	BN	BnC4_6604	CGGGGAACTTCTTCGTGTAA	ATGACGGTCAATGCACTCAA	270
KG200*	CO-LOC	BN	BnC7_3962	TGTCTGTGATGCAGGAGAGG	TGCGAGCTTCTTACTGCTTG	256
KG201*	CO-LOC	BN	BnC8_761	CAACTTGTTAGTTTGTGTTTTGGAA	ACAGGTTGTTCCCGAATCAG	172
KG202*	CO-LOC	BN	BnC9_CL7954Ctg1	TGCTGCTGGTGAACTTTTTG	ATATTGGAGCCAGGGCTTTT	280
KG204*	CO-LOC	BN	BnC10_7131	GGCAGGCATGGTAGCTCTTA	TGGCTGAAGGTGTTGTCAAG	283
KG205*	CO-LOC	BN	BnC12_2975	ATCATCCCAGCAGCTAATGC	GCAAAACATTTCCCACCTT	274
KG206*	CO-LOC	BN	BnC13_gi191204957	AGGACGACGTGGTGGTGTA	CAGGAGACGGAGGAGACGTA	290
KG207*	CO-LOC	BN	BnC15_3178	AGTTTGGTCTGGCACTTGCT	AAGTGCAAATTCGGCAAGAC	279
KG179*	CO-LOC	FFB	FFB1_CL1016	CCCCAACTTCTTTTGTTCCA	TGGAATGGAGAAACTGCAAA	244
KG180*	CO-LOC	FFB	FFB2_C4663_S1.2	GCACTGCATCATGTACTCCA	TCAGAGAAGCGAGAACTGCT	174
KG181*	CO-LOC	FFB	FFB2_C4741_S3	TAATGTAAGCTCGGGTGGCT	TCCCGCGTAACATAGATCGG	251
KG182*	CO-LOC	FFB	FFB2_C3566_S9	GGTGCACAAAAGTTACTTGGC	TCTGCAGAAGTTCATCCAAACA	172
KG183*	CO-LOC	FFB	FFB2_C2_S1	CTGCTTCCTCGAGAGTCCAT	CCCCAAATTCATTCCAGGGC	234
KG184*	CO-LOC	FFB	FFB2_C8_S1.2	TATCCTGCCAAACACGAGAG	ATTCCTGGTGCCTCGTTCAT	294
KG185*	CO-LOC	FFB	FFB2_C6_S3.4.5	GAAGAATCTGCCCTGGAATGA	TGCACATATTCCTTCCCTTGG	278
			FFB2_C6_S1.2	AGGAGATCTTCACCAGCTTCA	TTCCAGGGCAGATTCTTCGA	299
KG186*	CO-LOC	FFB	FFB6_C2082_S1	TCCTTCTTCACAGCTGCAGA	ATGAAGCACACATGAGGGCA	223
KG187*	CO-LOC	FFB	FFB6_C3684_S1	ACCTGAGAGTTGGATTTGCAG	AACGGGCAGAACAACATACA	211
KG188*	CO-LOC	FFB	FFB6_C596_S2	CTGATGGCCGGCAAGTATG	GCACTTGGAAGATGGAACCC	186
KG189*	CO-LOC	FFB	FFB8_C545_S1	GGGGCGAAGAACCTCTACAT	TCAAAGTACATATAGACGCCGTC	200
KG190*	CO-LOC	FFB	FFB8_C1455_S3.4.5.6	AGGGCCTTGAGTTATGTCCC	TCCAGATTACCGCAACACCA	257
KG191*	CO-LOC	FFB	FFB11_C6530_S1.2	ACCGATGCGATTTTCACGAG	TGAGAAGGAATCCAGCAGCT	235
KG192*	CO-LOC	FFB	FFB11_C1_S1	GAGCATGACCGAGATTCAGC	CCTCCTGATCTGACCGTGTT	253
KG193*	CO-LOC	FFB	FFB11_C1174_S3.4	ATGGTTAGATGCAATCCGGC	CCCCACCTTCACAATCTCCT	258
KG194*	CO-LOC	FFB	FFB11_C3877_S4	GAAGCTCCCTCCAATCCTGA	GTGCGAACTGAAAGGAAGCA	293

KG195*	CO-LOC	FFB	FFB13_C2168_S1	GCGAATGGAAAGCATGTGTT	TGGTGTCTCTTCATTACCCACA	187
KG255*	CO-LOC	PO	PO3_5-22	CTCGCTGTGCATTGTCCTTA	TGTACCACTCTGATCTGCAACT	291
KG256*	CO-LOC	PO	PO3_5-3	TTTGCGGCCTATGAGTTGTG	ACGCATCGAAGGGCAAATAC	217
KG257*	CO-LOC	PO	PO3_5-5	CGTCCAATGCATCAGAAGCA	CCCTCAAAACACATTCCGCA	206
KG258*	CO-LOC	PO	PO3_5-7	TTGGCTAGTCATCTCCCTCG	CCCAATGATCAAGGGGCTCA	198
KG259*	CO-LOC	PO	PO3_5-8	AGATTGCTGAGAATGAGAGAGC	TCGTCTGCTGCTGATGAGAG	150
KG260*	CO-LOC	PO	PO3_5-9	AGAAACTTGCACGCATCCAC	ATCTGAGACACTCCCTGGAT	156
KG261*	CO-LOC	PO	PO3_5-10	ACGTCGAGTGAGAATCTGGA	TTCCGCAGAAAGGTCATTGT	157
KG263*	CO-LOC	PO	PO3_5-12	ATTGTAGACCAGCCAGCCAT	TACGTGGCTCTCTCAGATGC	285
KG264*	CO-LOC	PO	PO3_5-13	TCTGTTTGCCTTCTCCACCA	GCATTGGCGGTAGAACTCTG	195
KG265*	CO-LOC	PO	PO3_5-14	AACACCTAGAGACGTGGGTG	ACCCCAGAAAGATTGGTCGT	280
KG275	GEN CONOCIDO	POP, PO	OLEOSIN	AATCTCCCCTCGCTTCACTT	CAACACTCACGCTACGAGGA	171
P4	GEN CONOCIDO	Bn, FFB, FN,FW	EPS 168	GGCGGATCGAGAACAAGATA	GGCGTACTCGTAGAGCTTGC	149
P39	GEN CONOCIDO	OTROS	WOS6942	GTTCCCGGACTTTGACGATA	AGCCATGCATGTACTGTGGA	184
P58	GEN CONOCIDO	OTROS	EPS3	AACATGGAGGAGGTGGTCAG	TTGTGAAGCGGAGGTAATCC	200
KG269	GEN CONOCIDO	OTROS	EgNAC	CGTTGTTCGGGCTAAAAGAG	ACTTGGTGCCTTCCCAGAGTA	196
KG210	GEN CONOCIDO	OTROS	EgLIP1	CCGGCATGAGAACATTAGTCT	AAGCCTCGTACTCCTCCATT	255
			preEgLIP	TGACTTCTGAGTTGGTTCATCA	TGCAGCAACCCTAAGGAAAT	250
KG212	GEN CONOCIDO	POP,PO	EgFATB2.2	GGAATGTGGGACCGAACTTG	CTCCATCCATCCTAGCACGT	203
KG213	GEN CONOCIDO	POP,PO	EgFATB3.2	TGTGTGGATGTTGGAGAGCT	GCATTCGTCTCATGGCTTCC	276
KG274	GEN CONOCIDO	FW,FN	LIPOIC	AATATGCTCCTCCGGGAACT	GTCGAATGCTTCTGGGGTAA	191
KG153	GEN CONOCIDO	FW,FN	ATAGB1	ATTCACTGGATTGGGCTCCT	CATGCACTATCAAGACCACCAC	190
KG154	GEN CONOCIDO	FW,FN	DDB1-CUL4	ATTCTCAGGCGAAATCCAGA	TTCCTCCCAACCTCAAACAG	242
KG155	GEN CONOCIDO	FW,FN	ELO2	ATCCAGCTGAGGCTGCTAAA	TTCGCCACTTTCTCTGTTCC	214
KG242	GEN CONOCIDO	FW,FN	EgETR_F2R2 (F2R2)	CAACCACCTCAACAAGCTCC	TGTGCCACCAGAAGTTGTTG	173
KG244	GEN CONOCIDO	FW,FN	EgMAX4_F1R1(F1R1)	CCGATGACCTTCCTCTCGTT	CATCCAGATCGATCCCCACA	192

KG252	GEN CONOCIDO	FW,FN	AIL5	TCATGAACAGTGACCTCCCC	CTTCAAAACCCAACGCCAGA	210
KG253	GEN CONOCIDO	FW,FN	ANT	TTGGCTCTTGGTTCATTGGC	AGGATTCAAGTCACGGCTGA	209
KG254	GEN CONOCIDO	FW,FN	PLT2	GGGGCATCATGGGGAAATTC	CTGGGTCTGGTTTGCTTCAG	228
KG270	GEN CONOCIDO	FW,FN	EgPPGL	GTGCTGTGCACAAGGCTCTA	GCTAGTCCCGCAGAAAGTTG	206
KG271	GEN CONOCIDO	FW,FN	VIR	TGGTCAGAAGATCAGCAATCA	CAAAGCAAGTCATCCCATCC	260
KG288	GEN CONOCIDO	FW,FN	EgTPase	TAGGCCTCATTTGCACTCGA	GGCACCATACCTCGAAAAGC	249
KG114	GEN CONOCIDO	FW,FN	JC41	AGAAAGCACGGTGCAAAGAT	ATTCATTGAAGTCGGCATCC	244
KG87	GEN CONOCIDO	OTROS	AUX2	CAGGGTGTTCCTTTCGTGAT	ACTGGTTGAATCTCGGGTTG	223
KG113	GEN CONOCIDO	OTROS	JC35	CCATGTGGGGTAAATCCTTG	GAATGCCCATCAGGAAAGAA	227
KG106	GEN CONOCIDO	OTROS	PSII2	TGGAAACAACAACCCTAGTCG	GGGAGACTCATTACTTCAACTAGTCC	150
KG107	GEN CONOCIDO	OTROS	PSII3	TTTGTGGAGCTCAGCATGTC	TTGGTCGAGGACTTCCAAAC	172
KG108	GEN CONOCIDO	OTROS	RU1	CAGGGGGTATTCATGTTTGG	TTCACGAGCAAGATCACGTC	185
KG278	GEN CONOCIDO	OTROS	PSI	AGCTTCGCATTGGATGAGAT	TGACACTCATTTGGGACTGG	173
KG115	GEN CONOCIDO	OTROS	JC47	CATCAGGGCCACTATCAACC	TTGTGCCCCCTTTTTGTTAG	201
KG117	GEN CONOCIDO	OTROS	JC55	CTTtGTGGAGCAACCCATCT	GTGTCCGTATCAAGCCCATT	226
KG101	GEN CONOCIDO	OTROS	ATP1	TTCGGAATCCACAAACCATT	TCGTAGGCGCAGCTAACTCT	222
KG102	GEN CONOCIDO	OTROS	ATP2	CAATGGTTAACGGTGGCTCT	TGCATGTCTCTTACCCTCAGC	171
KG103	GEN CONOCIDO	OTROS	ATP3	TTGCTGTAGGCCAAGCTGTA	AAAAGGATTCGCCAACAAAA	152
KG105	GEN CONOCIDO	OTROS	PSII1	TGGCTACACAAACCGTTGAG	TTCCATCCAGTAAAACGAAAGAA	207
KG156	GEN CONOCIDO	HI	CBB1	GGTGCTTGTACTGGGAAGGA	TGCACATACTCAAGGGCATC	208
KG157	GEN CONOCIDO	HI	RPL10	AAGCCATACCCAAAGTCACG	ATGGAAGGGATGCACTCTCA	234
KG159	GEN CONOCIDO	HI	GH3	AAGCGAGGAGTTCAAGACCA	CCTTCACCTCCTTTGAAGAGC	197
KG160	GEN CONOCIDO	HI	MUM	TGCATTGATGGAGGTTCTTG	ATGGATTGGTTGTGGAGGAG	220
KG233	GEN CONOCIDO	HI	asn1	CCAACCATTCTCCTCAGCA	CCGGCCTTGCTATCGTAGAG	328
KG234	GEN CONOCIDO	HI	asn2	GGGAAAATAATATCTTGAGTTCCACA	TGGATCATAGATCGGATGGA	211
KG245	GEN CONOCIDO	HI	EgBRX (BRX)	ATGTGTAACCCATCGCTCCA	ACTCTTTCGGGCCTTGGTTA	160

KG246	GEN CONOCIDO	HI	EgARF1 (ARF1)	AGCCCTTGATTTGTCGATGC	ACTGGACTTATCTTGGGGTTGT	282
KG247	GEN CONOCIDO	HI	EgPINF3-9_PIN4 (PINF3-9)	ATGGAGCGAGAGGACTTCAG	TGAGGCTGGAGTAGGTGTTG	208
KG248	GEN CONOCIDO	HI	EgPINF3-5_PIN3 (PINF3-5)	TGAGGCTGGAGTAGGTGTTG	AGTAGATGAGCGGAAGGAGC	242
KG249	GEN CONOCIDO	HI	EgPINF3-4_PIN2 (PINF3-4)	CATTCCAAGCCCTGCATCTG	GCTGAACTCCACCCGAAATC	324
KG250	GEN CONOCIDO	HI	EgPINF3-6_PIN1 (PINF3-6)	CTACCGAGCTCCTCCCAAAA	GCAGGCATTTCAACATCCCA	270
KG251	GEN CONOCIDO	HI	dwarf14	GAGGATTCGAGGAGGGTGAG	AAGATCGCTGTTGAACACCG	197
KG286	GEN CONOCIDO	HI	wri1	TGGTGAAGCAGATCTCGATG	TAGGGGCAGCTCTCGTAGAA	178
KG289	GEN CONOCIDO	HI	EgACP	ACAAGGATGGAAGCCGTCTA	CGTCGTCGTCACTTGAACAT	196
P13	GEN CONOCIDO	HI	QM	GCTCTTGAGGCTGCTCGTAT	CCCTCATTCCAGTCTGAAGC	160
KG64	GEN CONOCIDO	HI	PIN1	TCTCCACCAACGATCCCTAC	GCGGTACTCGAAGAGGAACA	170
KG69	GEN CONOCIDO	HI	BRI1	CTGCAAGGTTGGAGAAGAGC	GTGAATTATGTGGGGAATGC	187
KG70	GEN CONOCIDO	HI	BAK1	ACAGCAGTCCGTGGAACAAT	ACCCAGTCAAGCAACATCAC	179
KG75	GEN CONOCIDO	HI	GID1	CGAGTCGGAGAAGAGATTGG	AAGATCCAGTCCTGCCACTG	190
KG78	GEN CONOCIDO	HI	ETR1	ACAATGGCCTGCTGAAGAAC	AGATGGGTTGCTCCACAAAG	183
KG80	GEN CONOCIDO	HI	EIN4	AGGTGGGACACCAGAGATTG	CTGCGATTCCTCCAAAACT	166
KG81	GEN CONOCIDO	HI	ETR2	GATCCCGCAACTTCTGAGAG	AAGATGGTATGCCGGTCAAG	165
KG243	GEN CONOCIDO	OTROS	EgEBF (EBF)	TGCTGCCCAAAGTTGAAGTC	AGCAACAGCCTTCCCATAGT	174
KG268	GEN CONOCIDO	FW,FN	EgMBAGL2-3	CTGGGCCTAGCGTGAGTAAT	AACAGTTGCCAATTACAGACCA	157
KG119	GEN CONOCIDO	FW,FN	AP1-2MADS	TTCGCCCTGTAAGTGGGTAG	GAGAACAATGGAGCAGATTCAA	209
KG120	GEN CONOCIDO	FW,SH	SHELL	GGATCGAGAACACCACAAGC	AATTTGGCTTGGCCATAGAA	237
KG121	GEN CONOCIDO	FW,FN	MADS11-1	GGCGAGGGAGAAGATTCAG	CGGTGGAGGAGAAGATGATG	152
KG122	GEN CONOCIDO	FW,FN	DEF1	CGAGCACCCAGTTTATGGTT	GCGGATAGAGAGGCTTACCA	255
KG123	GEN CONOCIDO	FW,FN	GLO2	TTGCATGCCAGATTCCAATA	CAGCCATCTATTAGCCCATCA	214
KG124	GEN CONOCIDO	FW,FN	AG1	AGGAGGAGCCAAGAGGTAGC	TTGTTGGCGTATTCGTAGAGG	221
KG125	GEN CONOCIDO	FW,FN	SQUA3	AGGCACTAGTTTGCCTGCAT	TTTGAGCTCCAAAGCCAACT	281
KG126	GEN CONOCIDO	FW,FN	MADS1	CAGGGGGTATTCATGTTTGG	GTTCAGGGCTCCATTTGCTA	228

KG127	GEN CONOCIDO	FW,FN	AGL-2	CACGATTCCAATGGCTACCT	AAAACGGACGCACACTAAGC	272
KG128	GEN CONOCIDO	FW,FN	SQUA1	ACAGCCTGTTACCACCTTGG	CAGCCCCATCTACTTCCAGA	225
KG24	GEN CONOCIDO	POP,PO	PACT	CTGACTGGAGCGTGCTTCTT	GATCAGCCCCGATCTCATAC	192
KG27	GEN CONOCIDO	POP,PO	BKACPII_1	TGACCTTTGCATCATTCAGC	ACGCAGCTTCTTTGTTGGT	178
KG29	GEN CONOCIDO	PO, BW	GLUT1	TTCCAATCCTTCCAGCAATC	AGTATCGGCATCGGTGTAGC	159
KG31	GEN CONOCIDO	POP,PO	atpB	TGCGCAAAGAGTTAAGCAAA	CCACGAAGAAGGGTTGTGAT	152
KG35	GEN CONOCIDO	PO, IV	O3FAD	AAGTAGCGGGGAGGAGAGAG	ACCAAACAAAAAGCCCTCCT	209
KG38	GEN CONOCIDO	PO	HOLOS	GCATCCACATGTCCAAACTG	CATAATATCCGCTCCCCTGA	206
KG39	GEN CONOCIDO	PO, IV	BKACPIII	TGCAACAGTCAAGGATGAGG	AGCCAGTCAATGCTGGAACT	195
KG47	GEN CONOCIDO	PO, IV	ACAD	TCCAGCTATAAAAGGACAAGGAA	TTTGGGATCAAATGTTGCAG	154
KG57	GEN CONOCIDO	FW, PO	PDHB	TCACAGCGGTTGGAATCATA	CTCTGCCTCCTCCAGACAAC	201
KG92	GEN CONOCIDO	PO,POP	FA1	TTGTCCCTCCAACTGATGTG	ATCCGATTCACAGCAAAACC	181
KG93	GEN CONOCIDO	IV	FA2	ATGAGAAGCGCCATGAGACT	GTTCCACCTTCGGACAAGAA	245
KG94	GEN CONOCIDO	IV	FA4	GTGCATCTGAAGCCTGTTGA	CATATCTCCAACCAAGCAAACA	151
KG95	GEN CONOCIDO	IV	FA6	TGCATTTGAAGCCTGTTGAG	AGTAAGGCTTGCCCCTGTCT	233
KG96	GEN CONOCIDO	IV	FA8	GGGATGAAACAGGTGCAAAC	CCATTCCCGAACCAATTAGA	168
KG109	GEN CONOCIDO	PO,POP	JC8	AGGGCACTGATGGAATGAGA	TCATCACATGTAGAAGCTCTGC	162
KG148	GEN CONOCIDO	PO,POP	MUM4	CACCGTGGAGAAGTTGGACAT	CTCTGACCACCCCAAATTCT	195
KG149	GEN CONOCIDO	PO,POP	SDP1	GTGGGATCTTTGCAGTGGTT	GGAAATGCACAGGAAGCAGT	247
KG150	GEN CONOCIDO	PO,POP	PKP-BETA1	GCAACTTGGAGAGGTATCGC	TGACCTTCTGATGGGATGCA	185
KG151	GEN CONOCIDO	POP, PO, FN, FW	PKP-ALPHA	ACAAGCCTGTCATTGTAGCT	CTTCTCCTCTCTCCACCACC	218
KG152	GEN CONOCIDO	POP, PO, FN, FW	PRT6	TGCGTTCCATGTTTCCAGAA	ATCCAGAACAGGTCCCACAG	192
KG211	GEN CONOCIDO	PO,POP	EgFATB1.2	CTAATGACTGCACTGGTGGC	CCTGCCCCAATTGAGAATGG	194
KG214	GEN CONOCIDO	PO,POP	EgWRI1.1	TTCCTCTTCCGCTTCACCAT	ACCTCGTGACTCCTCTGTAG	158
KG215	GEN CONOCIDO	PO,POP	EgWRI1-2.2	ACCAACTTTGCTAGCAGTGC	GGCCATCATCAATTGGGACA	235
KG217	GEN CONOCIDO	PO,POP	PDAT_1	AGACGGGAGCTGTTTGAAAG	CACCTGCTGCCACTCTGATA	263

			PDAT_2	TTCCCTGTAACTGAGAAGCAAA	CAGAATGCAAAATCAGAACAAAA	185
KG272	GEN CONOCIDO	PO,POP	OLEOYL	AAGCAGTGGACCCTTCTTGA	TCTATAGAAGCCGTCCGATCA	155
KG276	GEN CONOCIDO	PO,POP	PYRKIN	GGATACGGTGGGTCCAGAG	GCCTTTGACAATCCACTGAAA	138
KG282	GEN CONOCIDO	PO,POP	ACYL-ACPF	AAGCAGTGGACCCTTCTTGA	TCTATAGAAGCCGTCCGATCA	155
P11	GEN CONOCIDO	PO,POP	WOS 104	TTCCCCACACCATCTTTCTC	GATTCAGCTTTCAGGCCAAC	191
P20	GEN CONOCIDO	PO,POP	EPS134312	AACGAATGGACAAATTCATGC	TATGAGCACTCCGCATTTTG	119
P23	GEN CONOCIDO	IV	EPS50987A	CCAGGATCAGGAGGATTGAA	TTGGTCATGCTCAGAGTTGC	108
P44	GEN CONOCIDO	IV	M6ASA	AAGACGCCCTTCTTCTCTCC	GCGCACCACCTCTTCTCTAC	182
P60	GEN CONOCIDO	PO,POP	PAT_1	GCGAGGGAGTGAAATATGGT	GTCTTGAGCCCACAGTCAGG	160
P62	GEN CONOCIDO	PO,POP	PAT_2	ATTCCGGTGATGGTGTGAGT	CCGTTCTGCAGTGGTAGTGA	158
P63	GEN CONOCIDO	PO,POP	PAT_3	CACTTCCTCATGCCATCCTT	GCAGACTCCAATTCCTGCTC	181
P64	GEN CONOCIDO	PO,POP	PAT_4	TGTTCAATGAGGGCATGAAG	AGAGATGACATGAGGAAGATCAA	188
P65	GEN CONOCIDO	PO,POP	PAT_5	ACATGGAAGGCTTTGGTGTC	AGAGATCCTGAGTGGCATGG	165
P66	GEN CONOCIDO	PO,POP	PAT_6	CTCCATCGTTTTACCCGAGA	AGGACTGTGCATTGTCGTTG	167
P68	GEN CONOCIDO	PO,POP	PAT_7	TCCGTGAGCTCCTCTTTTGT	TAGTGCCAGCAAGTTCGATG	180
P72	GEN CONOCIDO	PO,POP	PAT_8	GATCCCATCCACAGAGGAGA	GGAGGAGCTTAGCAGCAGAA	161
P74	GEN CONOCIDO	PO,POP	PAT_9	ATCAGGACGGGGTCCATCT	GCTGAAGATGTCGAGGTTGC	165
P77	GEN CONOCIDO	PO,POP	PAT_11	GGTGGATGCTATCGACGAGT	ACCTTGCATAGGCTCTCGAA	159
P78	GEN CONOCIDO	PO,POP	PAT_12	GCAAAGATCGAGGACCTGAG	CTTGGACCTCGAAACTCCAG	195
P79	GEN CONOCIDO	PO,POP	PAT_13	CCGACCACCTCTTCAATCTC	GACCTGGTAGGAGCCAAGG	155
P84	GEN CONOCIDO	PO,POP	PAT_14	CAAGGTTGGCTCTTCTGAGG	TCCAGCTGCAAACTTCTCAA	160
P86	GEN CONOCIDO	PO,POP	PAT_15	CCAAGAACGGGGAGAACAC	AAGAAGGGTGGCACGTAGAA	178
KG118	GEN CONOCIDO	OTROS	JC59	GACATTAGGAAGCAGTGCGAAG	CCATCCTCAAGAAGAGCAGGA	200
KG82	GEN CONOCIDO	BN, BW,FN,FW	IA2	GCAGGCGGAAGCTAATACTG	GATCACGTAGTGGCCTGGAT	164
KG290	GEN CONOCIDO	PO, POP	EgDSI	GTGCGGATAGTCAGGAGCTT	GCCTCCTTCTTCAGCAACAC	145

* Estos genes co-localizados fuero resecuenciados en la región dónde estaban situados los polimorfismos de un solo nucleótido (SNPs).

Tabla 6.2: Codificación de los genotipos por los MIDS. COD= número de MIDs; GT= codigo de genotipo (familia/genotipo); CRUCE= Parentales de genotipo. Parental femenino (*Dura*): CH = Chemara; D= Dami; HC= Harrison Crossfield. Parental masculino (*Pisifera*): A= Avros; D= Dami; E= Ekona; G=Ghana; LM= LaMé; N= Nigeria; Y= Yangambi.

COD	GT	CRUCE	COD	GT	CRUCE	COD	GT	CRUCE	COD	GT	CRUCE	COD	GT	CRUCE
1111	403/14	DxE	**1412**	538/33	CHxA	**2313**	670/5	(CHxHC)xY	**3214**	760/48	DxY	**4121**	718/22	DxD
1112	403/16	DxE	**1413**	538/43	CHxA	**2314**	670/45	(CHxHC)xY	**3221**	770/4	DxG	**4122**	476/28	CHxN
1113	405/5	DxE	**1414**	547/4	CHxD	**2321**	677/3	DxE	**3222**	770/45	DxG	**4124**	505/45	CHxG
1114	405/16	DxE	**1421**	547/17	CHxD	**2322**	677/15	DxE	**3223**	773/15	DxN	**4131**	580/33	HCxA
1121	405/20	DxE	**1422**	547/24	CHxD	**2323**	677/24	DxE	**3224**	773/45	DxN	**4132**	864/16	HCxG
1122	405/33	DxE	**1423**	557/7	CHxG	**2324**	677/42	DxE	**3231**	778/4	HCxE	**4133**	864/36	HCxG
1123	430/19	DxE	**1424**	557/28	CHxG	**2331**	680/6	DxN	**3232**	778/232	HCxE	**4134**	505/17	CHxG
1124	430/33	DxE	**1431**	566/48	CHxD	**2332**	680/32	DxN	**3233**	788/40	DxY	**4141**	422/13	CHxE
1131	433/10	DxA	**1432**	573/9	DxLM	**2333**	680/37	DxN	**3234**	806/431	CHxN	**4142**	422/41	CHxE
1132	433/13	DxA	**1433**	573/39	DxLM	**2334**	683/16	DxA	**3241**	816/192	HCxG	**4143**	522/22	DxE
1133	438/2	DxL	**1434**	574/31	DxD	**2341**	683/42	DxA	**3242**	818/32	HCxA	**4144**	625/36	CHxLM
1134	438/36	DxL	**1441**	574/39	DxD	**2342**	686/12	DxA	**3243**	820/15	CHxE	**4211**	651/20	DxD
1141	440/30	DxD	**1442**	586/36	CHxA	**2343**	686/37	DxD	**3244**	820/32	CHxE	**4212**	651/18	DxD
1142	440/36	DxD	**1443**	589/1	HCxE	**2344**	688/4	DxA	**3311**	837/3	DxA	**4213**	670/36	(CHxHC)xY
1143	446/8	DxN	**1444**	586/47	CHxA	**2411**	688/7	DxA	**3312**	837/32	DxA	**4214**	670/19	(CHxHC)xY
1144	446/20	DxN	**2111**	589/32	HCxE	**2412**	690/4	CHxG	**3313**	847/9	DxY	**4221**	731/44	DxY
1211	454/6	DxN	**2112**	591/11	CHxN	**2413**	690/9	CHxG	**3314**	847/35	DxY	**4222**	731/22	DxY
1212	455/13	DxA	**2113**	591/48	CHxN	**2414**	690/15	CHxG	**3321**	855/19	HCxN	**4223**	756/35	DxA
1213	455/45	DxA	**2114**	593/38	CHxE	**2421**	690/27	CHxG	**3322**	855/30	HCxN	**4224**	756/42	DxA
1214	461/26	DxA	**2121**	593/42	CHxE	**2422**	693/2	DxE	**3323**	855/35	HCxN	**4231**	818/28	HCxA
1221	461/31	DxA	**2122**	594/14	CHxA	**2423**	693/12	DxE	**3324**	855/44	HCxN	**4232**	818/42	HCxA
1222	462/29	CHxA	**2123**	594/27	CHxA	**2424**	697/8	(CHxHC)xG	**3331**	859/24	CHxN	**4233**	505/39	CHxG
1223	462/30	CHxA	**2124**	622/28	DxA	**2431**	698/34	(CHxHC)xY	**3332**	859/7	CHxN	**4234**	505/26	CHxG
1224	474/3	DxA	**2131**	622/40	DxA	**2432**	698/38	(CHxHC)xY	**3333**	861/14	DxDa	**4241**	658/43	DxE

1231	474/39	DxA	**2132**	624/24	HCxG	**2433**	699/17	DxA	**3334**	861/39	DxD	**4242**	658/2	DxE
1232	476/14	CHxN	**2133**	624/36	HCxG	**2434**	699/39	DxA	**3341**	861/40	DxD	**4243**	831/43	DxG
1233	476/16	CHxN	**2134**	626/6	HCxD	**2441**	713/19	HCxG	**3342**	866/39	HCxA	**4244**	831/13	DxG
1234	476/33	CHxN	**2141**	626/9	HCxD	**2442**	724/27	(CHxHC)xY	**3343**	866/44	HCxA	**4311**	837/43	DxA
1241	476/45	CHxN	**2142**	626/35	HCxD	**2443**	724/30	(CHxHC)xY	**3344**	880/1	HCxN	**4312**	837/21	DxA
1242	478/39	DxE	**2143**	635/17	CHxN	**2444**	725/7	HCxG	**3411**	880/26	HCxN	**4313**	440/46	DxD
1243	485/45	DxE	**2144**	635/48	CHxN	**3111**	726/28	HCxE	**3412**	881/5	HCxA	**4314**	440/42	DxD
1244	485/48	DxE	**2211**	640/6	DxE	**3112**	726/44	HCxE	**3413**	881/44	HCxA	**4321**	578/3	HCxE
1311	486/1	DxA	**2212**	640/7	DxE	**3113**	731/11	DxY	**3414**	884/281	CHxG	**4322**	578/1	HCxE
1312	486/12	DxA	**2213**	642/14	CHxN	**3114**	738/20	CHxN	**3421**	808/8	DxN	**4323**	681/5	DxN
1313	497/3	HCxA	**2214**	642/27	CHxN	**3121**	738/38	CHxN	**3422**	808/12	DxN	**4324**	681/31	DxN
1314	497/26	HCxA	**2221**	643/14	HCxN	**3122**	744/28	CHxD	**3423**	829/13	CHxN	**4331**	840/34	HCxD
1321	503/25	DxE	**2222**	643/28	HCxN	**3123**	744/29	CHxD	**3424**	829/3	CHxN	**4332**	840/8	HCxD
1322	503/43	DxE	**2223**	644/18	DxN	**3124**	748/11	CHxN	**3431**	552/25	MxN	**4333**	557/31	CHxG
1323	504/10	CHxE	**2224**	644/32	DxN	**3131**	748/22	CHxN	**3432**	552/9	MxN	**4342**	626/41	HCxD
1324	504/24	CHxE	**2231**	644/38	DxN	**3132**	748/47	CHxN	**3433**	718/6	DxDa	**4343**	776/6	DxD
1331	505/42	CHxG	**2232**	648/1	DxE	**3133**	752/8	DxA	**3434**	773/11	DxN	**4344**	776/42	DxD
1332	509/8	CHxE	**2233**	648/4	DxE	**3134**	752/18	DxA	**3441**	773/48	DxN	**4411**	855/36	HCxN
1333	509/41	CHxE	**2234**	648/47	DxE	**3141**	752/34	DxA	**3442**	519/44	CHxD			
1334	524/7	DxE	**2241**	654/13	DxG	**3142**	752/36	DxA	**3443**	519/24	CHxD			
1341	524/24	DxE	**2242**	654/23	DxG	**3143**	753/11	DxD	**3444**	711/10	DxA			
1342	526/6	CHxA	**2243**	657/13	DxD	**3144**	753/42	DxD	**4111**	711/19	DxA			
1343	526/34	CHxA	**2244**	657/31	DxD	**3211**	760/8	DxY	**4112**	804/46	(CHxHC)xDa			
1344	531/30	HCxA	**2311**	668/21	DxN	**3212**	760/15	DxY	**4113**	804/34	(CHxHC)xDa			
1411	531/47	HCxA	**2312**	668/42	DxN	**3213**	760/45	DxY	**4114**	522/18	DxE			

Anexo 7: Resultados de la secuenciación de los genes candidatos por librería

Tabla 7.1: Nº de lecturas de los genes candidatos de las librerías OP1, OP2 y OP3 y sus frecuencias en los genotipos de la población. GC= nombre del gen candidato; CEBADOR=nombre de la pareja de cebadores; AMPLICÓN= tamaño del amplicón del gen candidato en pares de bases; LECTURAS= frecuencia o número de lecturas del gen candidato; GT= frecuencia en los genotipos de la población (N=238).

OP1					**OP2**					**OP3**				
GC	**CEBADOR**	**AMPLICON**	**LECTURAS**	**GT**	**GC**	**CEBADOR**	**AMPLICON**	**LECTURAS**	**GT**	**GC**	**CEBADOR**	**AMPLICON**	**LECTURAS**	**GT**
CDA15	B33	160	52341	**243**	**CDA22***	B42	175	182123	242	**CDA39**	B63	152	85892	242
CDA22	B42(FW)	175	57	0	**CDA24**	B44	157	83221	242	**CDA76**	B100	157	279146	242
CDA26	B46	160	151303	**249**	**CDA27***	B47	199	169385	242	**CDA78**	B102	194	287087	242
CDA27	B47	199	62213	225	**CDA3**	B11	210	344	46	**KG101**	ATP1	222	133984	234
CDA31	B54	160	146753	**247**	**CDA32**	B56	157	7712	227	**KG102**	ATP2	171	340878	242
CDA34	B57b(MF)	162	509	23	**CDA34***	B57	162	13052	242	**KG103**	ATP3	152	49	0
CDA4	B3/B12	166	20468	242	**CDA37**	B61	152	258	29	**KG105**	PSII1	207	21351	223
KG2	mEg3275	146	3530	108	**CDA40**	B64	182	267778	242	**KG106**	PSII2	150	106408	239
KG24	PACT	192	2427	101	**CDA43**	B67	164	120292	242	**KG107**	PSII3	172	44863	242
KG27	BKACPII_1	178	3566	228	**CDA6**	B18	137	21	0	**KG108**	RU1	185	83027	242
KG29	GLUT1	159	1002	44	**CDA9**	B25	185	8546	240	**KG109**	JC8	162	25629	232
KG31	atpB	152	134810	**249**	**KG11**	M8373	169	66901	241	**KG113**	JC35	227	425	64
KG35	O3FAD	209	23	0	**KG12**	M2200	169	94795	242	**KG114**	JC41	244	17283	231
KG38	HOLOS	206	4551	94	**KG24***	PACT	192	22172	242	**KG115**	JC47	201	84	6
KG39	BKACPIII	195	3730	189	**KG272**	OLEOYL	155	9140	241	**KG117**	JC55	226	43280	229
KG47	ACAD	154	1142	152	**KG274**	LIPOIC	191	41640	242	**KG118**	JC59	200	36349	230
KG57	PDHB	201	7333	175	**KG275**	OLEOSIN	171	261920	242	**KG119**	AP1-2MADS	209	37557	227
P11	WOS104	191	11402	214	**KG276**	PYRKIN	138	127363	242	**KG120**	SHELL	237	28019	231

P13	QM	160	17477	239	**KG278**	PSI	173	402	65	**KG121**	MADS11-1	152	23719	236
P20	EPS134312	119	34992	242	**KG282**	ACYL-ACPF	155	8	0	**KG122**	DEF1	255	14702	228
P23	EPS50987A	108	18178	161	**KG29***	GLUT1	159	13859	237	**KG123**	GLO2	214	56375	234
P39	WOS6942	184	25943	241	**KG35***	O3FAD	209	142	14	**KG124**	AG1	221	77957	239
P4	EPS168	149	32790	241	**KG47***	ACAD	154	2629	206	**KG125**	SQUA3	281	2612	214
P44	M6ASA	182	37250	242	**KG64**	ATPIN1	170	4	0	**KG126**	MADS1	228	371435	242
P58	EPS3	200	6342	74	**KG69**	BRI1	187	49	1	**KG127**	AGL-2	272	26138	223
					KG75	GID1	190	2497	216	**KG128**	SQUA1	225	88883	237
					KG78	ETR1	183	581	79	**KG282***	ACYL-ACPF	155	15844	230
					KG80	EIN4	166	24307	242	**KG286**	wri1	178	2964	159
					KG81	ETR2	165	4050	238	**KG64***	ATPIN1	170	1	0
					P60	PAT_1	160	11722	240	**KG69***	BRI1	187	2689	132
					P62	PAT_2	158	5213	226	**KG70**	BAK1	179	17851	230
					P63	PAT_3	181	15	0	**KG82**	IA2	164	1	0
					P64	PAT_4	188	15	0	**KG87**	AUX2	223	1	0
					P66	PAT_6	167	30815	242	**KG92**	FA1	181	4434	207
					P68	PAT_7	180	5898	241	**KG93**	FA2	245	135	13
					P72	PAT_8	161	18912	242	**KG94**	FA4	151	22	0
					P77	PAT_11	159	5741	224	**KG95**	FA6	233	50721	231
					P78	PAT_12	195	9271	238	**KG96**	FA8	168	7521	224
					P79	PAT_13	155	50121	242	**P64***	PAT_4	188	37763	231
					P84	PAT_14	160	79974	242	**P65**	PAT_5	165	22697	235
					P86	PAT_15	178	1126	162	**P74**	PAT_9	165	54925	242

Tabla 7.2: : Nº de lecturas de los genes candidatos de las librerías OP4, OP5 y OP6 y sus frecuencias en los genotipos de la población. GC= nombre del gen candidato; CEBADOR=nombre de la pareja de cebadores; AMPLICÓN= tamaño del amplicón del gen candidato en pares de bases; LECTURAS= frecuencia o número de lecturas del gen candidato; GT= frecuencia en los genotipos de la población (N=238).

OP4					**OP5**					**OP6**				
GC	**CEBADOR**	**AMPLICON**	**LECTURAS**	**GT**	**GC**	**CEBADOR**	**AMPLICON**	**LECTURAS**	**GT**	**GC**	**CEBADOR**	**AMPLICON**	**LECTURAS**	**GT**
KG135	M3117	197	10008	229	**KG159***	GH3	197	639	51	**KG105***	PSII1	207	145	19
KG138	M9619	217	126858	242	**KG161***	HtC1_3885	202	10995	220	**KG114***	JC41	244	615	27
KG140	M43696	236	2017	188	**KG164***	HtC2_11412	195	53737	242	**KG117***	JC55	226	432	31
KG141	M847	217	6107	219	**KG171***	HtC4_4489	282	5677	219	**KG118***	JC59	200	106	24
KG142	M2252	196	16792	232	**KG172***	HtC4_240	283	8816	231	**KG120***	Shell	237	0	0
KG143	M3256	184	520244	242	**KG174***	HtC7_1247	158	384	27	**KG121***	MADS11-1	152	4	3
KG144	M23551	199	9358	161	**KG177***	HtC10_11102	300	6468	220	**KG122***	DEF1	255	536	33
KG145	M7495	180	123135	242	**KG179**	FFB1_CL1016_S1.2	244	7984	237	**KG123***	GLO2	214	29	12
KG147	M4883	183	1246	143	**KG180**	FFB2_C4663_S1.2	174	85067	**249**	**KG124***	AG1	221	11939	78
KG148	MUM4	195	1468	155	**KG181**	FFB2_C4741_S3	251	136301	**250**	**KG125***	SQUA3	281	13	9
KG149	SDP1	247	159747	242	**KG182**	FFB2_C3566_S9	172	695844	0	**KG167***	HtC2_1255C2_411	223	300	27
KG150	PKP-BETA1	185	335927	242	**KG183**	FFB2_C2_S1	234	132716	242	**KG233**	ASP1	328	374	91
KG151	PKP-ALPHA	218	13922	240	**KG184**	FFB2_C8_S1.2	294	1673	147	**KG234**	ASP2	211	3555	191
KG152	PRT6	192	6011	221	**KG185**	FFB2_C6_S3.4.5	278	1093	0	**KG242**	EgETR_F2R2	173	852988	**254**
KG153	ATAGB1	190	615787	242	**kg185**	FFB2_C6_S1.2	299	21012	183	**KG243**	EgEBF	174	466333	**255**
KG154	DDB1_CUL4	242	2437	189	**KG186**	FFB6_C2082_S1	223	25852	236	**KG244**	EgMAX4_F1R1	192	16168	237
KG155	ELO2	214	9094	234	**KG187**	FFB6_C3684_S1	211	42863	241	**KG245**	EgBRX	160	6334	213
KG156	CBB1	208	21776	234	**KG188**	FFB6_C596_S2	186	276807	242	**KG246**	EgARF1	282	485	144
KG157	RPL10	234	202574	241	**KG189**	FFB8_C545_S1	200	90849	242	**KG247**	EgPINF3-9_PIN4	208	4110	215
KG159	GH3	197	6	0	**KG190**	FFB8_C1455_S3.4.5.6	257	110428	240	**KG248**	EgPINF3-5_PIN3	242	24438	242

KG160	MUM	220	3581	219	**KG191**	FFB11_C6530_S1.2	235	649	52	**KG249**	EgPINF3-4_PIN2	324	2	2
KG161	HtC1_3885	202	431	25	**KG192**	FFB11_C1_S1	253	141886	242	**KG250**	EgPINF3-6_PIN1	270	722	165
KG162	HtC1_5925	171	10881	238	**KG193**	FFB11_C1741_S3.4	258	92564	242	**KG251**	dwarf14	197	23025	236
KG164	HtC2_11412	195	1324	155	**KG194**	FFB11_C3877_S4	293	14386	234	**KG252**	AIL5	210	130483	**252**
KG166	HtC2_7081	230	4000	215	**KG195**	FFB13_C2168_S1	187	75291	242	**KG253**	ANT	209	28542	241
KG167	HtC2_1255C2-411	223	4220	192	**KG196**	BnC2_10C3-629	300	1898	154	**KG254**	PLT2	228	2903	215
KG168	HtC2_9289	264	18963	235	**KG197**	BnC2_1289	254	147036	241	**KG255**	PO3_5-22	291	53	41
KG171	HtC4_4489	282	664	46	**KG198**	BnC3_792	209	61721	241	**KG256**	PO3_5-3	217	85	59
KG172	HtC4_240	283	126	1	**KG199**	BnC4_6604	270	111042	241	**KG257**	PO3_5-5	206	145999	**248**
KG173	HtC7_9200	136	25501	238	**KG200**	BnC7_3962	256	46494	239	**KG258**	PO3_5-7	198	340678	**254**
KG174	HtC7_1247	158	54	3	**KG201**	BnC8_761	172	46989	242	**KG259**	PO3_5-8	150	560788	**253**
KG175	HtC8_1026C1-144	277	2878	199	**KG202**	BnC9_CL7954Ctg1	280	15455	216	**KG260**	PO3_5-9	156	53491	238
KG176	HtC8_11217	259	11491	232	**KG204**	BnC10_7131	283	10877	222	**KG261**	PO3_5-10	157	7617	224
KG177	HtC10_11102	300	365	16	**KG205**	BnC12_2975	274	22723	231	**KG263**	PO3_5-12	285	14	10
KG268	EgMBAGL2-3	157	57815	234	**KG206**	BnC13_gi191204957	290	7249	129	**KG264**	PO3_5-13	195	32243	144
KG269	EgNAC	196	50834	238	**KG207**	BnC15_3178	279	93969	242	**KG265**	PO3_5-14	280	21266	237
KG270	EgPPGL	206	15708	237	**kg210**	preEGLIP	250	1451	242	**KG286***	wri1	178	2	2
KG271	VIR	260	0	0	**KG210**	EgLIP1	255	199572	64	**KG47***	ACAD	154	6158	49
KG288	EgTPase	249	9948	233	**KG211**	EgFATB1.2	194	151886	242	**KG57***	PDHB	201	435	29
KG289	EgAcp	196	197555	242	**KG212**	EgFATB2.2	203	326380	242	**KG75***	GID1	190	569	31
KG290	EgDSI	145	177953	242	**KG213**	EgFATB3.2	276	87657	0	**KG96***	FA8	168	91	24
					KG214	EgWRI1.1	158	188903	242	**P62***	PAT_2	158	16959	86
					KG215	EgWRI1-2.2	235	13631	0	**P64***	PAT_4	188	785	31
					kg217	PDAT_2	185	3147	237	**P74***	PAT_9	165	9158	63
					KG217	PDAT_1	263	29394	131					

KG271*	VIR	260	14905	231

Anexo 8: Patrones definitivos

Tabla 8.1: Genes candidato que no obtuvieron ningún patrón.

SIN PATRÓN			
GC	**LIBRERÍA**	**ORIGEN**	**CEBADOR**
KG278	OP2	GEN CONOCIDO	PSI
KG64	OP2-OP3	GEN CONOCIDO	ATP1N1
KG69	OP2	GEN CONOCIDO	BRI1
P63	OP2	GEN CONOCIDO	PAT3
CDA3	OP2	TDF	B11
CDA37	OP2	TDF	B61
CDA6	OP2	TDF	B18
KG103	OP3	GEN CONOCIDO	ATP3
KG113	OP3	GEN CONOCIDO	JC35
KG115	OP3	GEN CONOCIDO	JC47
KG82	OP3	GEN CONOCIDO	IA2
KG87	OP3	GEN CONOCIDO	AUX2
KG93	OP3	GEN CONOCIDO	FA2
KG94	OP3	GEN CONOCIDO	FA4
KG172	OP4_OP5	CO-LOC*	HtC4_240
KG174	OP4_OP5	CO-LOC*	HtC7_1247
KG159	OP4_OP5	GEN CONOCIDO	GH3
KG215	OP5	GEN CONOCIDO	EgWRI1-2.2
KG249	OP6	GEN CONOCIDO	EgPINF3-4_PIN2

Tabla 8.2: Genes candidato con un patrón.

PATRÓN ÚNICO			
GC	**LIBRERÍA**	**ORIGEN**	**CEBADOR**
KG2	OP1	CO-LOC	mEg3275
KG24	OP1	GEN CONOCIDO	PACT
KG38	OP1	GEN CONOCIDO	HOLOS
P11	OP1	GEN CONOCIDO	WOS104
P20	OP1	GEN CONOCIDO	EPS134312
P23	OP1	GEN CONOCIDO	EPS50987A
P4	OP1	GEN CONOCIDO	EPS168
P58	OP1	GEN CONOCIDO	EPS3
CDA27	OP1	TDF	B47
CDA31	OP1	TDF	B54
CDA34	OP1	TDF	B57b
KG12	OP2	CO-LOC	M2200
KG275	OP2	GEN CONOCIDO	OLEOSIN
KG47*	OP2	GEN CONOCIDO	ACAD
KG81	OP2	GEN CONOCIDO	ETR2
P60	OP2	GEN CONOCIDO	PAT1
P68	OP2	GEN CONOCIDO	PAT7
P72	OP2	GEN CONOCIDO	PAT8
P86	OP2	GEN CONOCIDO	PAT15
CDA27*	OP2	TDF	B47
KG101	OP3	GEN CONOCIDO	ATP1
KG102	OP3	GEN CONOCIDO	ATP2
KG107	OP3	GEN CONOCIDO	PSII3

KG108	OP3	GEN CONOCIDO	RU1
KG109	OP3	GEN CONOCIDO	JC8
KG119	OP3	GEN CONOCIDO	AP1-2MADS
KG126	OP3	GEN CONOCIDO	MADS1
KG127	OP3	GEN CONOCIDO	AGL-2
KG128	OP3	GEN CONOCIDO	SQUA1
KG69*	OP3	GEN CONOCIDO	BRI1
KG92	OP3	GEN CONOCIDO	FA1
KG95	OP3	GEN CONOCIDO	FA6
P65	OP3	GEN CONOCIDO	PAT5
CDA39	OP3	TDF	B63
CDA76	OP3	TDF	B100
CDA78	OP3	TDF	B102
KG145	OP4	CO-LOC	M7495
KG147	OP4	CO-LOC	M4883
KG161	OP4	CO-LOC*	HtC1_3885
KG168	OP4	CO-LOC*	HtC2_9289
KG176	OP4	CO-LOC*	HtC8_11217
KG149	OP4	GEN CONOCIDO	SDP1
KG150	OP4	GEN CONOCIDO	PKP-BETA1
KG151	OP4	GEN CONOCIDO	PKP-ALPHA
KG152	OP4	GEN CONOCIDO	PRT6
KG156	OP4	GEN CONOCIDO	CBB1
KG160	OP4	GEN CONOCIDO	MUM
KG289	OP4	GEN CONOCIDO	EgAcp
KG161*	OP5	CO-LOC*	HtC1_3885
KG185	OP5	CO-LOC*	FFB2_C6_S3.4.5
KG188	OP5	CO-LOC*	FFB6_C596_S2
KG199	OP5	CO-LOC*	BnC4_6604
KG202	OP5	CO-LOC*	BnC9_CL7954Ctg1
KG207	OP5	CO-LOC*	BnC15_3178
KG210	OP5	GEN CONOCIDO	EgLIP1
kg210	OP5	GEN CONOCIDO	preEGLIP
KG211	OP5	GEN CONOCIDO	EgFATB1.2
KG213	OP5	GEN CONOCIDO	EgFATB3.2
KG217	OP5	GEN CONOCIDO	PDAT_1
KG255	OP6	CO-LOC*	PO3_5-22
KG256	OP6	CO-LOC*	PO3_5-3
KG259	OP6	CO-LOC*	PO3_5-8
KG263	OP6	CO-LOC*	PO3_5-12
KG247	OP6	GEN CONOCIDO	EgPINF3-9_PIN4
KG248	OP6	GEN CONOCIDO	EgPINF3-5_PIN3
KG251	OP6	GEN CONOCIDO	dwarf14

Tabla 8.3: Genes candidatos con 2 patrones. La tabla muestra el número y tipo de SNP e INDEL en cada gen candidato. GC= nombre del gen candidato; LIBRERÍA= número de la librería dónde se ha secuenciado; ORIGEN= procedencia del gen candidato; CEBADOR= nombre de la pareja de cebadores; SNP= número de polimorfismos detectados en la secuencia; TR= tipos transición (A/G o C/T); TV= tipo transversión (G/T; G/C; A/C o A/T); INDEL= número de INDEL. El número total de SNP/INDEL , así como el total de cada tipo se muestra al final de la tabla.

GC	**LIBRERIA**	**ORIGEN**	**CEBADOR**	**SNP**	**TR**		**TV**				**INDEL**
					A/G	C/T	G/T	G/C	A/C	A/T	
CDA15	OP1	TDF	B33	**1**	1	0	0	0	0	0	**0**
KG27	OP1	GEN CONOCIDO	BKACPII_1	**2**	0	0	0	1	1	0	**0**
KG39	OP1	GEN CONOCIDO	BKACPIII	**4**	0	0	1	1	1	1	**0**
KG57	OP1	GEN CONOCIDO	PDHB	**1**	1	0	0	0	0	0	**0**
P13	OP1	GEN CONOCIDO	QM	**1**	1	0	0	0	0	0	**0**
P39	OP1	GEN CONOCIDO	WOS6942	**1**	0	1	0	0	0	0	**0**
CDA22*	OP2	TDF	B42	**2**	1	1	0	0	0	0	**0**
CDA24	OP2	TDF	B44	**13**	5	4	0	0	2	2	**0**
CDA32	OP2	TDF	B56	**5**	0	2	0	2	0	1	**0**
CDA43	OP2	TDF	B67	**1**	0	1	0	0	0	0	**0**
KG11	OP2	CO-LOC	M8373	**2**	0	2	0	0	0	0	**0**
KG272	OP2	GEN CONOCIDO	OLEOYL	**1**	0	0	1	0	0	0	**0**
KG274	OP2	GEN CONOCIDO	LIPOIC	**1**	1	0	0	0	0	0	**0**
KG276	OP2	GEN CONOCIDO	PYRKIN	**2**	0	0	0	0	0	2	**0**
KG29*	OP2	GEN CONOCIDO	GLUT1	**3**	1	0	1	0	1	0	**0**
KG75	OP2	GEN CONOCIDO	GID1	**1**	0	0	0	1	0	0	**0**
KG78	OP2	GEN CONOCIDO	ETR1	**3**	2	1	0	0	0	0	**0**
KG80	OP2	GEN CONOCIDO	EIN4	**1**	0	0	0	1	0	0	**0**
P62	OP2	GEN CONOCIDO	PAT2	**8**	1	3	2	1	0	1	**0**
P66	OP2	GEN CONOCIDO	PAT6	**1**	0	0	0	0	1	0	**0**
P77	OP2	GEN CONOCIDO	PAT11	**1**	0	0	0	1	0	0	**0**
P78	OP2	GEN CONOCIDO	PAT12	**3**	1	1	0	0	0	1	**0**

P79	OP2	GEN CONOCIDO	PAT13	**1**	0	0	1	0	0	0	**0**
P84	OP2	GEN CONOCIDO	PAT14	**1**	1	0	0	0	0	0	**0**
KG105	OP3	GEN CONOCIDO	PSII1	**2**	1	0	1	0	0	0	**0**
KG106	OP3	GEN CONOCIDO	PSII2	**2**	0	2	0	0	0	0	**0**
KG117	OP3	GEN CONOCIDO	JC55	**1**	0	0	0	0	1	0	**0**
KG118	OP3	GEN CONOCIDO	JC59	**3**	0	1	0	0	1	1	**0**
KG121	OP3	GEN CONOCIDO	MADS11-1	**1**	0	0	1	0	0	0	**0**
KG122	OP3	GEN CONOCIDO	DEF1	**1**	1	0	0	0	0	0	**0**
KG123	OP3	GEN CONOCIDO	GLO2	**2**	1	1	0	0	0	0	**0**
KG124	OP3	GEN CONOCIDO	AG1	**1**	0	1	0	0	0	0	**0**
KG125	OP3	GEN CONOCIDO	SQUA3	**1**	0	1	0	0	0	0	**0**
KG282*	OP3	GEN CONOCIDO	ACYL-ACPF	**1**	0	0	1	0	0	0	**0**
KG286	OP3	GEN CONOCIDO	wri1	**18**	1	7	2	5	2	1	**0**
KG70	OP3	GEN CONOCIDO	BAK1	**8**	2	4	0	1	0	1	**0**
KG96	OP3	GEN CONOCIDO	FA8	**1**	0	1	0	0	0	0	**0**
P64*	OP3	GEN CONOCIDO	PAT4	**1**	0	1	0	0	0	0	**0**
P74	OP3	GEN CONOCIDO	PAT9	**16**	4	3	2	5	1	1	**2**
KG135	OP4	CO-LOC	M3117	**2**	0	0	0	1	1	0	**0**
KG140	OP4	CO-LOC	M43696	**1**	1	0	0	0	0	0	**0**
KG142	OP4	CO-LOC	M2252	**1**	1	0	0	0	0	0	**0**
KG143	OP4	CO-LOC	M3256	**10**	2	3	0	1	3	1	**0**
KG148	OP4	GEN CONOCIDO	MUM4	**1**	0	1	0	0	0	0	**0**
KG154	OP4	GEN CONOCIDO	DDB1CUL4	**1**	0	1	0	0	0	0	**0**
KG155	OP4	GEN CONOCIDO	ELO2	**1**	0	1	0	0	0	0	**0**
KG157	OP4	GEN CONOCIDO	RPL10	**1**	1	0	0	0	0	0	**0**
KG175	OP4	CO-LOC*	HtC8_1026C1-144	**1**	0	0	0	1	0	0	**0**
KG268	OP4	GEN CONOCIDO	EgMBAGL2-3	**1**	0	0	0	1	0	0	**0**

KG269	OP4	GEN CONOCIDO	EgNAC	**1**	0	0	0	1	0	0	**0**
KG270	OP4	GEN CONOCIDO	EgPPGL	**2**	0	1	0	0	0	1	**0**
KG288	OP4	GEN CONOCIDO	EgTPase	**1**	0	0	1	0	0	0	**0**
KG290	OP4	GEN CONOCIDO	EgDSI	**1**	1	0	0	0	0	0	**0**
KG171*	OP5	CO-LOC*	HtC4_4489	**1**	0	0	0	0	0	1	**0**
KG179	OP5	CO-LOC*	FFB1_CL1016_S1.2	**1**	0	0	0	0	1	0	**0**
KG180	OP5	CO-LOC*	FFB2_C4663_S1.2	**1**	1	0	0	0	0	0	**0**
KG181	OP5	CO-LOC*	FFB2_C4741_S3	**4**	2	1	1	0	0	0	**0**
KG182	OP5	CO-LOC*	FFB2_C3566_S9	**1**	1	0	0	0	0	0	**0**
KG183	OP5	CO-LOC*	FFB2_C2_S1	**1**	1	0	0	0	0	0	**0**
KG184	OP5	CO-LOC*	FFB2_C8_S1.2	**1**	1	0	0	0	0	0	**0**
KG186	OP5	CO-LOC*	FFB6_C2082_S1	**2**	0	0	1	0	1	0	**0**
KG187	OP5	CO-LOC*	FFB6_C3684_S1	**1**	0	1	0	0	0	0	**0**
KG191	OP5	CO-LOC*	FFB11_C6530_S1.2	**1**	1	0	0	0	0	0	**0**
KG192	OP5	CO-LOC*	FFB11_C1_S1	**1**	0	1	0	0	0	0	**0**
KG193	OP5	CO-LOC*	FFB11_C1741_S3.4	**1**	0	0	0	0	0	1	**0**
KG194	OP5	CO-LOC*	FFB11_C3877_S4	**1**	0	0	1	0	0	0	**0**
KG195	OP5	CO-LOC*	FFB13_C2168_S1	**1**	0	0	0	0	0	1	**0**
KG196	OP5	CO-LOC*	BnC2_10C3-629	**3**	0	1	1	0	1	0	**0**
KG198	OP5	CO-LOC*	BnC3_792	**1**	0	0	0	0	0	1	**0**
KG200	OP5	CO-LOC*	BnC7_3962	**1**	0	1	0	0	0	0	**0**
KG201	OP5	CO-LOC*	BnC8_761	**1**	0	0	1	0	0	0	**0**
KG206	OP5	CO-LOC*	BnC13_gi191204957	**1**	0	1	0	0	0	0	**0**
KG214	OP5	GEN CONOCIDO	EgWRI1.1	**1**	0	0	1	0	0	0	**0**
KG243	OP6	GEN CONOCIDO	EgEBF	**4**	2	2	0	0	0	0	**0**
KG245	OP6	GEN CONOCIDO	EgBRX	**1**	0	1	0	0	0	0	**0**
KG253	OP6	GEN CONOCIDO	ANT	**1**	0	0	0	1	0	0	**0**
KG254	OP6	GEN CONOCIDO	PLT2	**1**	0	0	0	0	1	0	**0**

KG257	OP6	CO-LOC*	PO3_5-5	**0**	0	0	0	0	0	0	**1**
KG260	OP6	CO-LOC*	PO3_5-9	**1**	0	1	0	0	0	0	**0**
KG261	OP6	CO-LOC*	PO3_5-10	**1**	1	0	0	0	0	0	**0**
KG265	OP6	CO-LOC*	PO3_5-14	**0**	0	0	0	0	0	0	**1**
			TOTAL	**177**	**42**	**55**	**20**	**25**	**19**	**18**	**4**

Anexo 9: Asociación de los patrones de los genes candidatos a los genotipos de la población.

Tabla 9.1: Alelos y haplotipos observados para cada gen candidato con 2 patrones.

GC-CEBADOR	ORIGEN	POSICIÓN	A1	A2
CDA15_B33	TDF	135	A	G
CDA22_B42	TDF	156, 160	A,T	G,C
CDA24_B44	TDF	36, 37, 46, 53, 54, 57, 59, 64, 65, 66, 72, 111, 115	A,T,T,T,T,A,G,A,C,A,C,A	G,C,C,A,A,C,G,A,G,A,C,T,G
CDA32_B56	TDF	18, 70, 86, 109, 129	C,T,C,G,T	T,A,G,C,C
CDA43_B67	TDF	28	C	T
KG105_PSII1	GC	37, 182	A,G	G,T
KG106_PSII2	GC	55, 73	C,C	T,T
KG11_M8373	CO-LOC	99, 108	C,C	T,T
KG117_JC55	GC	81	A	C
KG118_JC59	GC	30, 61, 94	A,T,T	C,C,A
KG121_MADS11-1	GC	84	G	T
KG122_DEF1	GC	58	A	G
KG123_GLO2	GC	94, 164	A,C	G,T
KG124_AG1	GC	120	C	T
KG125_SQUA3	GC	136	C	T
KG135_M3117	CO-LOC	22, 127	C,A	G,C
KG140_M43696	CO-LOC	21	G	A
KG142_M2252	CO-LOC	160	A	G
KG143_M3256	CO-LOC	33, 48, 63, 81, 105, 108, 117, 123, 123, 144	A,T,C,A,C,T,A,C,T,C	G,C,G,C,A,C,G,T,A,A
KG148_MUM4	GC	162	C	T
KG154_DDB1CUL4	GC	135	C	T
KG155_ELO2	GC	38	C	T

KG157_RPL10	GC	189	A	G
KG171_HtC4_4489	CO-LOC	220	A	T
KG175_HTC8_1026C1-144	CO-LOC	117	C	G
KG179_FFB1_CL1016_S1.2	CO-LOC	37	A	C
KG180_FFB2_C4663_S1.2	CO-LOC	77	A	G
KG181_FFB2_C4741_S3	CO-LOC	86, 111, 124, 221	G,A,T,G	T,G,C,A
KG182_FFB2_C3566_S9	CO-LOC	60	A	G
KG183_FFB2_C2_S1	CO-LOC	108	A	G
KG184_FFB2_C8_S1.2	CO-LOC	61	A	G
KG186_FFB6_C2082_S1	CO-LOC	174, 206	G,C	T,A
KG187_FFB6_C3684_S1	CO-LOC	160	C	T
KG191_FFB11_C6530_S1.2	CO-LOC	163	A	G
KG192_FFB11_C1_S1	CO-LOC	203	C	T
KG193_FFB11_C1741_S3.4	CO-LOC	145	A	T
KG194_FFB11_C3877_S4	CO-LOC	136	G	T
KG195_FFB13_C2168_S1	CO-LOC	31	A	T
KG196_BnC2_10C3-629	CO-LOC	28, 164, 266	C,T,A	T,G,C
KG198_BnC3_792	CO-LOC	184	A	T
KG200_BnC7_3962	CO-LOC	50	T	C
KG201_BnC8_761	CO-LOC	87	G	T
KG206_BnC13_gi191204957	CO-LOC	169	C	T
KG214_EgWRI1.1	GC	79	G	T
KG243_EgEBF	GC	30, 42, 46, 93	C,G,G,C	T,A,A,T
KG245_EgBRX	GC	57	C	T
KG253_ANT	GC	23	C	G
KG254_PLT2	GC	110	A	C
KG257_PO3_5-5	CO-LOC	68	-	A
KG260_PO3_5-9	CO-LOC	73	T	C
KG261_PO3_5-10	CO-LOC	122	A	G
KG265_PO3_5-14	CO-LOC	98	-	T
KG268_EgMBAGL2-3	GC	114	C	G
KG269_EgNAC	GC	10	C	G
KG27_BKACPII_1	GC	60, 61	G,C	C,A
KG270_EgPPGL	GC	180, 181	AC	TT
KG272_OLEOYL	GC	81	G	T
KG247_LIPOIC	GC	158	A	G
KG276_PYRKIN	GC	125, 126	A,T	T,A
KG282_ACYL-ACPF	GC	81	G	T

KG286_WRI1	GC	27, 32, 50, 56, 59, 62, 64, 71, 78, 89, 95, 98, 101, 110, 129, 146, 149, 150	A,C,C,G,T,T,T,C,T,C,C,C,T,G,T,G,C,C	G,A,G,C,C,C,A,T,C,G,T,A,C,C,G,T,T,G
KG288_EgTPase	GC	193	G	T
KG29_GLUT1	GC	40, 71, 89	T,C,G	G,A,A
KG290_EgDSI	GC	75	A	G
KG39_BKACPIII	GC	121, 122, 123, 124	G,C,A,T	T,G,C,A
KG57_PDHB	GC	169	A	G
KG70_BAK1	GC	51, 63, 84, 108, 117, 135, 141, 159	A,A,G,T,A,C,T,C	T,G,C,C,G,T,C,T
KG75_GID1	GC	84	C	G
KG78_ETR1		134, 160, 161	A, T, G	G, C, A
KG80_EIN4	GC	112	C	G
KG96_FA8	GC	147	C	T
P13_QM	GC	42	A	G
P39_WOS6942	GC	158	C	T
P62_PAT2	GC	44, 50, 53, 56, 81, 96, 99, 132	A,C,C,A,G,C,T,T	G,T,T,T,T,G,C,G
P64_PAT4	GC	104	C	T
P66_PAT6	GC	142	A	C
P74_PAT9	GC	27, 35, 49, 52, 64, 66, 67, 70, 74, 79, 81, 84, 97, 102, 108, 121, 122, 130	G,G,-,C, A,C, A, A, C,T,C,G,A,C,T,G,T,C	T, A, C, T, T, G, G, G,G,G, G, C, G, G,C, -, C, A
P77_PAT11	GC	143	C	G
P78_PAT12	GC	76, 129, 147	A,T,A	G,C,T
P79_PAT13	GC	96	G	T
P84_PAT14	GC	109	A	G

Tabla 9.2: Alelos y haplotipos observados para cada gen candidato con 3 patrones.

GC_CEBADOR	ORIGEN	POSICION	A1	A2	A3
CDA26_B46	TDF	52, 124	C,A	C,G	T,G
CDA4_B3/B12	TDF	23, 29, 47, 51, 54, 56, 62, 68, 70, 71, 73, 79, 80, 81, 83, 86, 113, 122, 129, 137, 140, 143, 147	A,C,A,C,C,T,A,G,T,T,C,C,A,G,A,C,A,T,C,A,T,A,G	G,T,T,A,T,G,C,C,C,A,G,T,T,A,T,T,T,A,T,G,A,A,A	G,T,T,C,C,G,T,C,T,C,C,C,C,A,T,T,T,T,C,G,A,C,G
CDA9_B25	TDF	28, 33, 34, 46,	A,G,A,A,T,T,T,G,C,A,A,G,A,T,G,A,G,T,G,G	A,G,G,A,C,T,C,G,T,A,G,A,G,C,A,G,G,T,T,A	G,A,G,G,T,C,T,A,A,T,A,G,G,C,A,G,A,A,G,G

		55, 57, 62, 68, 69, 70, 71, 76, 77, 81, 82, 93, 112, 116, 122, 123, 141, 142, 160, 162	,T,G,T,G	,G,A,T,A	,T,G,C,A
KG114_JC41	GC	30, 153, 224	C,C,G	C,T,C	G,C,G
KG120_SHELL	GC	58, 65, 200, 201, 202, 203, 204, 205, 206	C,A,-,-,-,-,T,-	T,A,-,-,-,-,-,T	T,T,A,A,T,T,T,-
KG141_M847	Gc	123, 134	A, G	G, A	G, G
KG144_M23551	CO-LOC	57, 63	C,A	C,G	T,G
KG153_ATAGB1	GC	30, 47, 57, 60, 83, 89, 135, 142, 147	C,C,A,C,G,A,G,C,T	C,C,A,T,G,A,G,C,T	T,T,T,T,A,G,A,G,A
KG162_HtC1_5925	CO-LOC	89, 113	C,G	T,A	T,G
KG164_ HTC2_11412	CO-LOC	24, 38, 42, 57, 87, 108, 114, 135, 141, 147	C,C,T,A,C,C,C,G,C,T	C,C,T,A,C,T,T,G,C,T	G,G,G,C,T,T,T,A,T,C
KG166_HtC2_7081	CO-LOC	75, 78, 184	A,C,G	C,T,G	C,T,T
KG167_HtC2_1255C2-411	CO-LOC	49, 127, 144, 175	G,G,C,A	G,G,T,A	T,A,T,G
KG173_HtC7_9200	CO-LOC	25, 40, 41, 64, 88, 97, 106, 115, 116	A,C,A,A,C,A,G,A,G	C,T,G,G,T,T,A,T,A	T,C,A,A,C,A,G,A,G
KG177_HTC10_11102	CO-LOC	118, 273	C,A	T,A	C,C
KG189_FFB8_C545_S1	CO-LOC	21, 27, 30, 36, 39, 42, 45, 48, 57, 66, 69, 72, 81, 96, 105, 111, 114, 117, 123, 147, 150, 151, 159, 172	A,T,T,T,T,G,C,C,T,T,T,T,C,T,G,A,T,T,A,T,G,A,G,A	T,G,G,G,C,A,G,G,G,G,C,C,T,C,A,G,C,C,G,C ,A,C,C,C	T,G,G,G,C,A,G,G,G,T,C,C,T,C,A,G,C,C,G,C, A,C,C,C
KG190_FFB8_C1455_S3. 4.5.6	CO-LOC	44, 140, 184, 216	A,T,C,A	G,C,C,A	T,C,G,T
KG204_BnC10_7131	CO-LOC	34, 70, 112, 122, 145, 191, 230	A,T,G,A,C,G,T	G,G,C,G,T,A,C	G,G,G,G,T,A,C
KG205_BnC12_2975	CO-LOC	128, 217	G,T	T,T	G,C

KG212_EgFATB2.2	GC	121, 135	G,A	T,A	T,G
KG217_PDAT_2	GC	104, 107	A,T	T,T	A,A
KG233_ASP1	GC	52, 254	T,G	T,C	G,C
KG242_EgETR_F2R2	GC	26, 28, 29, 31, 51, 52, 53, 59, 62, 86, 95, 115, 129, 140, 146	G,A,A,C,C,C,T,G,T,C,C,T,T,T,T	G,A,A,C,T,C,T,G,T,C,C,T,T,T,T	A,G,G,A,T,T,C,A,A,T,T,C,A,C,A
KG244_EgMAX4_F1R1	GC	54, 78	A,C	G,C	A,A
KG246_EgARF1	GC	21, 38, 119, 136, 162, 197, 206, 213	C,A,T,G,G,G,G,T	C,G,A,A,A,A,A,A,A	T,A,T,G,G,G,G,T
KG252_AIL5	GC	101, 153	T,C	T,G	C,G
KG258_PO3_5-7	CO-LOC	94, 121	G,T	G,C	A,C
KG264_PO3_5-13	CO-LOC	40, 80	A,C	A,G	G,G
KG31_atpB	GC	21,79	A,C	A,T	G,C

Tabla 9.3: Alelos y haplotipos observados para cada gen candidato con 4, 5 y 6 patrones.

GC_CEBADOR	**ORIGEN**	**POSICION**	**A1**	**A2**	**A3**	**A4**		
CDA40_B64_CG46	TDF	60, 100	G,G	G,A	A,G	A,A		
KG138_M9619	CO-LOC	29, 38, 107, 154	A,A,C,G	A,A,T,A	G,A,T,A	G,G,T,A		
KG185_FFB2_C6_S1.2	CO-LOC	69,139,184,235	C,T,G,C	T,G,A,C	T,T,A,A	T,T,A,C		
KG197_BnC2_1289	CO-LOC	49, 76, 102	C,C,C	T,C,C	C,G,C	C,G,T		
KG271_VIR	GC	46, 54, 63, 103, 209	A,G,A,T,A	A,C,A,T,A	G,C,A,T,T	G,C,G,A,A		
P44_M6ASA	GC	30,49,73,123	C,G,C,G	C,T,C,G	C,T,T,G	T,T,T,A		
GC_CEBADOR	**ORIGEN**	**POSICION**	**A1**	**A2**	**A3**	**A4**	**A5**	
KG250_EgPINF3-6_PIN1	CO-LOC	66, 68, 72, 235	A,T,G,T	G,C,A,T	G,T,A,A	G,T,A,T	G,T,G,T	
GC_CEBADOR	**ORIGEN**	**POSICION**	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**
KG234_ASP2	GC	39, 40, 50, 84, 87, 127, 169	C,G,T,A,A,C,A	C,G,T,A,A,T,A	C,G,T,G,A,C,A	T,A,C,A,G,C,A	T,G,C,A,G,C,A	T,G,C,A,G,C,G

Tabla 9.4: Genes candidatos y número de valores perdidos en el conjunto de genotipos. GC_CEBADOR= Nombre del gen candidato y su cebador correspondiente; VP=número de valores pérdidos enel conjunto de los 242 genotipos para ese GC.

GC_CEBADOR	VP	GC_CEBADOR	VP	GC_CEBADOR	VP	GC_CEBADOR	VP	GC_CEBADOR	VP
CDA15_B33	0	CDA26_B46	0	KG189_FFB8_C545_S1	1	KG171_HtC4_4489	5	KG196_BnC2_10C3-629	21
CDA43_B67	0	CDA40_B64	0	KG205_BnC12_2975	1	KG114_JC41	5	KG154_DDB1CUL4	22
KG106_PSII2	0	KG138_M9619	0	KG212_EgFATB2.2	1	KG141_M847	5	KG148_MUM4	26
KG11_M8373	0	KG153_ATAGB1	0	KG242_EgETR_F2R2	1	KG185_FFB2_C6_S1.2	5	KG254_PLT2	27
KG124_AG1	0	KG31_atpB	0	KG187_FFB6_C3684_S1	2	KG204_BnC10_7131	5	KG217_PDAT_2	28
KG143_M3256	0	CDA24_B44	1	KG198_BnC3_792	2	KG177_HTC10_11102	5	KG245_EgBRX	30
KG155_ELO2	0	KG121_MADS11-1	1	KG270_EgPPGL	2	KG122_DEF1	6	KG39_BKACPIII	31
KG157_RPL10	0	KG123_GLO2	1	P62_PAT2	2	KG179_FFB1_CL1016_S1.2	6	KG164_ HTC2_11412	31
KG180_FFB2_C4663_S1.2	0	KG182_FFB2_C3566_S9	1	KG142_M2252	2	KG253_ANT	6	CDA32_B56	49
KG181_FFB2_C4741_S3	0	KG183_FFB2_C2_S1	1	KG173_HtC7_9200	2	KG120_SHELL	6	KG234_ASP2	54
KG186_FFB6_C2082_S1	0	KG192_FFB11_C1_S1	1	KG197_BnC2_1289	2	KG260_PO3_5-9	7	KG286_WRI1	60
KG274_LIPOIC	0	KG193_FFB11_C1741_S3.4	1	KG252_AIL5	2	CDA9_B25	7	KG57_PDHB	66
KG268_EgMBAGL2-3	0	KG195_FFB13_C2168_S1	1	KG258_PO3_5-7	2	KG96_FA8	8	KG206_BnC13_gi191204957	68
KG276_PYRKIN_CG40	0	KG201_BnC8_761	1	KG257_PO3_5-5	3	KG194_FFB11_C3877_S4	9	KG250_EgPINF3-6_PIN1	78
KG290_EgDSI_CG154	0	KG214_EgWRI1.1	1	KG282_ACYL-ACPF	3	KG265_PO3_5-14	9	KG78_ETR1	94
KG80_EIN4_CG33	0	KG243_EgEBF	1	KG144_M23551	3	KG244_EgMAX4_F1R1	9	KG184_FFB2_C8_S1.2	97

P39_WOS6942	0	KG269_EgNAC	1	KG166_HtC2_7081	3	KG75_GID1	10	KG246_EgARF1	103
P66_PAT6	0	KG27_BKACPII_1	1	KG190_FFB8_C1455_S3.4.5.6	3	P77_PAT11	10	KG264_PO3_5-13	105
P79_PAT13	0	KG272_OLEOYL	1	KG117_JC55	4	KG271_VIR	12	KG140_M43696	133
P84_PAT14	0	KG288_EgTPase	1	LG118_JC59	4	KG105_PSII1	13	KG233_ASP1	153
P74_PAT9	0	KG29_GLUT1	1	KG200_BnC7_3962	4	KG125_SQUA3	16	KG191_FFB11_C6530_S.2	191
CDA22_B42	0	P13_QM	1	KG70_BAK1	4	KG261_PO3_5-10	18		
P44_M6ASA	0	P78_PAT12	1	P64_PAT4	4	KG167_HtC2_1255C2-411	18		
CDA4_B3/B12	0	KG162_HtC1_5925	1	KG135_M3117	5	KG175_HTC8_1026C1-144	19		

Anexo 10: Datos fenotípicos

Tabla 10.1: Valores fenotípico medios para cada carácter en cada genotipo de estudio. Estos valores fueron tomados durante 10 años continuados hasta que las palmeras cumplieron los 15 años de edad CRUCES: Parentales femeninos (Dura): CH= Chemara, D=Dami , HC=Harrison Crossfield y M=Mardi. Parentales masculinos (Pisifera): A= Avros,Da= Dami Komposit, E= Ekona, G= Ghana, L= La Mé, N= Nigeria y Y= Yangambi.BN= Número de medio de racimos/palmera/año; BW=Peso medio del racimo/año (Kg); CPO=Rendimiento de aceite/hectarea/año (ton/ha/año); FN=Número medio de frutos por racimo; FW=Peso medio del fruto (g); HT=Altura de tallo (cm); MF=Ratio de pulpa o mesocarpio humedo con respecto al fruto (%); OWM= Ratio de aceite con respecto al peso seco de pulpa o mesocarpio (%).

GENOTIPO	CRUCE	BN	BW (Kg)	CPO (Ton/ha/año	FN	FW (g)	HT (cm)	MF (%)	OWM (%)
403/14	DxE	8,5	15,6	3,8	1001	9	47	92,9	49,5
403/16	DxE	10,9	14,5	3,4	1230	7,8	58	71,2	45,2
405/5	DxE	11	15,2	5,6	2515	4,5	49	76,6	61
405/16	DxE	11,3	14,7	3,7	1196	7,1	49	68,8	43,8
405/20	DxE	9	15,5	5,1	1200	8,7	45	88,7	54,6
405/33	DxE	11,4	12,2	3	792	8,6	45	68,9	53,4
433/10	DxA	11	13,9	2	1046	8,3	52	76,4	28,6
433/13	DxA	12,1	14,1	4,9	1114	7	45	74,6	64,1
438/2	DxL	8,8	21,9	4,8	1414	8	78	84,8	52,5
438/36	DxL	10,2	13,5	3,5	966	9,8	65	73,6	50,5
440/30	DxDa	11,2	11,8	2,7	664	6,7	100	82,9	54,1
440/36	DxDa	9,8	14,4	3,7	1026	8	44	82,4	50,4
446/8	DxN	12,2	15	4,2	1082	9,4	88	66,1	50,4
446/20	DxN	8,9	11,2	2,9	1169	6,2	37	74,6	54,8
455/45	DxA	17,5	14,3	5,3	709	11,4	96	82	44
461/26	DxA	7,1	17,8	3,8	1345	8,2	61	84,7	58,2
461/31	DxA	8,1	16,1	1,8	977	10,7	63	79,7	27,7
462/29	ChxA	9,5	16,1	3,6	506	19,9	94,8	81,6	42,2
462/30	ChxA	8,5	14,6	3,4	1448	6,6	60	78,5	59,8
474/3	DxA	9,6	13,8	2,8	441	19,4	77	77,6	47,7
474/39	DxA	14,5	14,6	4,9	1722	5,8	62	71,6	47,8
476/14	ChxN	8,7	14,7	3,4	498	15,5	77	86,1	46,8
476/16	ChxN	9,1	17,6	5,3	1963	6,4	58	84,7	58,4
476/33	ChxN	11,4	15,1	6,8	522	20,6	70	87	54,3
476/45	ChxN	9,2	14,5	3,7	1325	8,2	65	74,1	49,9
478/39	DxE	9,4	15,3	5,1	1285	8	94	83,7	62,4
485/45	DxE	14,3	13,6	6,4	812	10,3	81	87,8	60,9
485/48	DxE	6,8	13,6	3	1365	6,5	58	81,7	60,2
486/1	DxA	6,8	12,4	1,5	588	10,8	39	77	50,8
497/3	HCxA	9,7	13,9	4,2	2070	4,9	61	77,2	49,6
497/26	HCxA	13,4	14	7,5	1120	9	63	78,3	61,2
503/25	DxE	10,6	14,7	2,6	1353	7,2	67	72,4	31,3
503/43	DxE	9,8	20,4	4,9	1042	12,2	71	85,1	52,4
504/10	ChxE	9,5	12	2,7	785	7,8	45	85,3	49,1
504/24	ChxE	12,8	15,9	5,8	825	13,3	91	80,8	50,2
505/42	ChxG	9,2	14,5	2,9	589	12,5	43	86,8	43,6

509/8	ChxE	9,9	14,8	4,9	801	12,7	48	74,4	63,4
509/41	ChxE	6,6	14,2	2,4	1273	7,1	52	77,4	52,2
524/7	DxE	8,7	13,4	2,8	987	9,7	49	75	50,1
524/24	DxE	10,4	13,7	3,1	1021	9,2	102	80,4	43,1
526/6	ChxA	12,2	12,7	3,7	361	11,1	75	81,9	49
538/43	ChxA	10,2	13,6	4,7	891	10,9	71	88,3	53,3
547/4	ChxDa	11,4	12,6	2,2	805	8,2	71	65,9	43,6
566/48	ChxDa	15,6	13,5	5,4	1195	7,2	56	72,4	55,4
573/9	DxL	8	15,7	3,6	1061	9,3	54	67,6	55,4
573/39	DxL	7,9	22,3	4,1	1321	10,1	77	80,5	56,4
574/31	DxDa	12,7	15	4,3	920	10,2	31	84,9	44,2
574/39	DxDa	7,3	15,4	1,9	954	9,8	34	73,4	40,6
586/36	ChxA	7,7	16,7	3,3	895	11	84	76,1	49
589/1	HCxE	8,2	12,6	2,2	239	26,7	55	86,9	46
586/47	ChxA	8,5	13,7	3,3	1688	5,5	61	82,1	47,5
589/32	HCxE	12,8	16,3	2,8	1385	6,6	61	80,2	31,8
591/11	CHxN	8,3	16,5	5	2619	4,5	57	80,8	66,7
591/48	CHxN	11,6	12,1	4,6	716	11,2	49	78,2	57,8
594/14	CHxA	10	15,2	2,6	937	12	69	83,6	31
594/27	CHxA	17	15,1	7,5	1134	9	80	78,2	57,9
622/40	DxA	6,8	15,7	2,6	735	13,5	65	82,4	47,8
624/24	HCxG	11,8	12,8	4	1021	8,7	87	81,3	46,4
624/36	HCxG	10	11,9	3	805	10,2	38	77,2	45,6
626/6	HCxDa	13,5	14	5,5	764	12,9	98	80,7	54,2
626/9	HCxDa	10,1	11,7	2,7	1012	8,7	39	72,5	42,4
640/6	DxE	10,3	14	6,9	1656	6,6	49	88	62
642/14	CHxN	6,6	13,6	1,6	1308	7,7	61	65,4	35,1
640/7	DxE	9	14,8	2,4	1311	7,4	46	81,2	36,6
642/27	CHxN	10,3	15,1	5,7	1905	6,2	71	89,3	47,7
643/14	HCxN	15,2	14,3	7	709	11,3	51	77,1	62,7
643/28	HCxN	11,3	10,7	2,5	741	9	45	80,9	36,7
644/18	DxN	10,2	15,5	3,2	726	13,4	47	72,8	48,2
644/32	DxN	10,3	15	4,1	2121	4,6	65	81,3	51,1
644/38	DxN	9,9	12,6	3,3	580	10,1	64	91,6	51,9
648/1	DxE	12	13,9	2,5	1013	6,8	72	63,7	46,3
648/4	DxE	12,8	14,8	4	651	12	92	80,7	45
648/47	DxE	4,8	13,3	1,7	516	15,7	43	90,1	44,3
654/13	DxG	9,3	15,4	1,9	895	8,3	57	76,1	33,4
654/23	DxG	13,1	15	7,6	1086	9,7	71	79,9	57,6
657/31	DxDa	16,7	14,1	4,9	968	9,6	59	81	38,6
668/21	DxN	8,6	16,1	4,4	1147	9,3	64	78	60,4
668/42	DxN	13,5	14,1	7,2	864	11,2	64	80,8	61,7
670/45	(CHxHC)xY	9,5	15,2	3,3	492	9,6	75	78,7	47,6
677/3	DxE	11,3	16,3	5,5	2221	5,4	86	76,2	56,2
677/15	DxE	12,5	15,4	2,8	824	12,6	70	76,7	30,9
677/24	DxE	12,1	15,8	6,3	1353	8	66	81,9	62,4

677/42	DxE	15,1	13,1	6,5	818	11,2	55	92,8	51,6
680/6	DxN	12,3	14,3	3,8	1326	6,8	53	69,2	48,2
680/32	DxN	8	20,6	4,6	1594	9,1	89	81,4	58,3
680/37	DxN	10,4	13,7	3	742	10,7	65	79,4	40,7
683/16	DxA	10,6	15,5	4,8	981	10,2	60	71,6	65,6
683/42	DxA	9,4	17,1	3,9	1177	9,9	58	68,8	53,4
686/12	DxA	11,4	13,9	2,1	1418	6	51	68,9	30,8
686/37	DxDa	16,3	13,2	7,4	1098	8,5	68	77,2	60,8
688/4	DxA	15,6	13,9	7,9	975	10,4	74	77,9	61,6
688/7	DxA	8,9	12,4	3	1002	8,9	44	72,6	53,2
690/15	CHxG	8,5	20	4,9	881	15,1	69	83,5	57,6
690/27	CHxG	8,9	11,7	2,2	513	15,4	63	79,6	38,3
693/2	DxE	7,5	15,9	3,2	599	16,7	54	82,6	56,1
693/12	DxE	11,2	17,6	4,1	1962	6,1	75	80,9	39,9
698/38	(CHxHC)xY	15,5	14,2	5,4	925	9,3	55	82,4	48,4
699/17	DxA	8,3	13,9	2,9	823	9,2	48	78	54,3
699/39	DxA	8,7	16,7	2,5	1155	10,4	70	79,4	27,9
713/19	HCxG	12	16,2	4,1	571	12,7	63	71,6	60,6
725/7	HCxG	9,8	17,2	4,2	2268	4,9	50	85,4	48,8
726/28	HCxE	8,7	13,6	3,8	1319	7,1	49	92,6	55,3
726/44	HCxE	10,4	17,2	3,1	722	10	76	78,7	56,8
731/11	DxY	9,9	15,3	3,5	1386	7,9	55	69,9	45,3
738/20	CHxN	9,1	20,8	2,5	2092	6,5	42	67,5	36,2
738/38	CHxN	11,4	16,4	3,4	337	19,5	52	93	45,9
744/28	CHxDa	11,8	13	3,7	855	10	49	78,5	42,2
748/22	CHxN	8,3	17,6	4,5	502	19,9	82	88,5	55,5
748/47	CHxN	8,7	13,7	3,3	1392	6,9	64	76,6	50,6
752/8	DxA	9,5	17,4	6	1343	8,4	97	84,3	63,1
752/18	DxA	7,3	14,8	3,1	1284	8,2	42	83,9	45,9
752/34	DxA	9,7	16,3	4,9	1263	8,2	72	79	69,7
752/36	DxA	9,9	15,2	2,6	833	11,3	63	81,4	37,2
753/11	DxDa	10	18,3	3,1	1008	12,5	78	86	30,6
753/42	DxDa	10,3	15,2	4,9	933	12,2	63	73,4	62,2
760/8	DxY	8,1	23,8	3,2	1599	8,4	70	73,2	50
760/15	DxY	9,1	14,2	3,3	1073	7,5	45	77,2	58,6
760/45	DxY	10,1	14,2	3,3	252	30,7	84	77,6	56,4
760/48	DxY	11,5	12,3	4	1065	7,2	41	77,6	55,5
770/4	DxG	12,7	16,9	3,2	1308	9	70	76	28
770/45	DxG	10,3	15,4	3,7	1753	6,1	53	71,8	48,4
773/15	DxN	9,2	13,3	3,5	548	18,4	46	78,8	47,7
773/45	DxN	8,4	16,4	3,9	577	18,7	59	85,4	57,6
778/4	HCxE	8,2	12,9	1,6	1187	4,6	49	85,6	39,4
806/431	CHxN	10,9	12,2	4	493	13,3	57	81,2	53,9
820/15	CHxE	7	20,6	3,5	1429	9,7	57	70,8	53,3
820/32	CHxE	12,8	14	4,9	927	10,4	62	75,3	54,3
837/3	DxA	15,1	14	5,6	907	9,3	58	83,2	51,7

837/32	DxA	16,6	12,7	4,9	548	10,9	63	73,3	60,3
847/9	DxY	7,9	12,7	2,2	976	8,4	51	75	48,1
847/35	DxY	9,1	18,3	2,2	848	12,6	61	72,7	29,1
855/35	HCxN	11,6	15,6	2,8	437	9,6	56	76,8	67,3
855/44	HCxN	8,1	15,6	1,5	965	7,6	59	72,1	31,7
859/24	CHxN	9,5	14	4	955	10	51	77,9	52,9
859/7	CHxN	9,5	13,8	4,4	1232	8,6	105	76	53,7
861/14	DxDa	14,8	16,7	7	1399	8,2	68	74,8	54,3
861/39	DxDa	15,7	12,5	4,2	504	11,5	60	79,7	53,2
861/40	DxDa	6,7	15,2	1,8	1175	8,6	45	72,7	34,5
866/39	HCxA	11,1	16,1	4	1066	10,5	79	74,6	42,4
866/44	HCxA	12,4	15,3	5,4	538	19,7	69	85,9	49,9
880/1	HCxN	9,6	13	2,4	1228	7,2	61	82,1	29,5
880/26	HCxN	9,7	15,6	5,1	857	10,8	61	89,3	61,2
881/5	HCxA	5,3	12,7	2,1	728	9,6	68	83,7	58,3
881/44	HCxA	13,2	14,2	3,8	588	12	56	85,1	43,1
808/12	DxN	9,8	21,1	5,7	1598	9,9	70	72,3	56,5
829/13	CHxN	9,1	15,4	2,4	1733	5,7	58	83,1	36,8
829/3	CHxN	11,4	13,2	4,8	632	12,2	60	82,7	54,9
552/25	MxN	9,8	14,9	2	221	13,7	73	82,5	46,2
552/9	MxN	9,1	14,4	2,8	1230	8	53	63,3	49
718/6	DxDa	9	14,8	3	1016	9,5	54	84,1	41,6
773/11	DxN	13,9	15,4	7,3	1045	10,1	73	84,1	58,5
773/48	DxN	7,8	16,5	2,7	1203	9,8	56	63,8	44,5
804/46	(CHxHC)xDa	17,3	13,4	7,1	1256	7,5	57	85,2	52,3
804/34	(CHxHC)xDa	9,9	14,5	2,9	893	7,5	55	79,4	47
522/18	DxE	9,1	15,9	2,2	1173	9,9	63	72,9	30,2
505/45	CHxG	10,6	15,8	4,9	568	18	73	80,3	51
864/36	HCxG	10,2	15,6	2,4	974	10,9	65	64,7	37,3
505/17	CHxG	8,7	19,2	4,5	1072	12,7	73	79,8	50,2
522/22	DxE	11,6	13,1	3,6	1132	7,1	63	77	51,8
625/36	CHxLa	9,4	19,5	4,9	1098	11,9	56	78,4	55,8
731/44	DxY	11,6	16	6,8	909	12,6	81	82,8	63
731/22	DxY	8,3	13,6	2,2	444	14,1	74	89,6	45,3
505/39	CHxG	8	19,8	4,3	1170	12,1	66	76,5	58,8
505/26	CHxG	11,5	12,6	5,4	688	14,3	42	73,8	61,2
658/43	DxE	11,6	16,1	2,3	377	16,7	67	78	32,9
658/2	DxE	14,4	14,4	6,4	474	19,1	56	84,5	53,5
831/43	DxG	14,2	14,6	6,7	823	10,5	72	77,5	62,3
831/13	DxG	11,5	15,9	4,1	1224	7,8	56	78,7	46,7
837/21	DxA	11	13,6	4,5	1326	6,6	99	81,4	52
440/46	DxDa	8,5	12,6	1,9	784	8,1	36	78,2	49,5
440/42	DxDa	12,8	15,5	7,4	431	19,2	55	82,3	63,3
578/3	HCxE	8,6	16,1	3	2171	4,7	57	84,7	40,8
578/1	HCxE	8,7	15,4	2,6	1079	9,5	54	73,5	37,6
840/34	HCxDa	11,9	14,6	3,7	447	22,3	69	80,4	40,1

840/8	HCxDa	9,7	16,3	3,4	1450	6,9	65	78,6	43,9
557/31	CHxG	12,4	16	5,2	1115	7,9	62	77,3	67,3
626/41	HCxDa	9,8	16,8	2,8	837	11,9	62	77,7	36,1
776/6	DxDa	11,9	15,3	5,2	586	19	68	90,9	48,4
776/42	DxDa	7,1	14	2,5	1409	6,8	63	73,5	44,2
855/36	HCxN	12,4	16,5	4,6	1881	5,4	68	73,3	48,2
580/33	HCxA	12,2	15	5,1	1047	7,4	43	83,3	57,4
864/16	HCxG	12,2	14,6	6,3	874	9,1	52	83	60,9
651/20	DxDa	8,8	16,7	3,1	1164	7,7	77	69,1	51,8
651/18	DxDa	8,4	15,2	3,7	946	9,7	40	76,5	60,8
593/38	CHxE	16,7	13,5	6	773	9,3	77	85,1	51,3
593/42	CHxE	7,7	14,9	3,6	1101	8,7	90	82,1	60,9
816/192	HCxG	18,3	12,3	6,5	920	8,5	61	87,5	48,4
519/44	CHxDa	16,3	15,1	4,4	905	9,2	68	73	43,3
519/24	CHxDa	7,4	14,1	3,1	889	10,2	49	71,2	54,2
547/17	ChxDa	7,5	22	3,5	1264	12,9	72	75,1	40,4
547/24	ChxDa	7,4	14,5	2,7	919	10,9	78	77,4	44,4
778/232	HCxE	8,3	21,6	3,1	1158	10,5	74	77,5	40,9
788/40	DxY	9,2	11,3	2,6	419	14,6	67	83,6	50,4
718/22	DxDa	7,8	21,9	3,4	1294	10,8	51	78,4	48
476/28	CHxN	10,4	13	4	1320	6,9	41	80,2	46,9
557/7	ChxG	12	14,3	4	850,8	10,6	66	77,2	50,7
818/32	HCxA	8,4	13,9	2,4	913	9,7	39	81,3	35
486/12	DxA	11,4	14,9	6,8	1380	8,2	70	92,2	65,9
526/34	ChxA	7,1	13,8	2,1	1009	8,9	64	78,9	44,3
635/48	CHxN	7	13,4	1,7	936	9,9	46	72,1	39,7
681/5	DxN	12,7	17,6	9,3	1273	8,9	72	86,2	65,9
681/31	DxN	7,4	15,3	2,5	1265	7,5	62	80,7	47,2
531/30	HCxA	12	15,2	7	975	10,8	101,8	88,7	60,2
430/19	DxE	13,6	13,2	5,4	2376	4,1	66	76,4	55,7
430/33	DxE	13,2	12	3,6	390,8	10,2	61	84,2	49,6
557/28	ChxG	8,3	18,2	4,1	674,4	13,2	65	84,4	60
657/13	DxDa	11,8	14,9	2,8	2473,9	3,8	60	66,2	39,3
690/4	CHxG	10,1	16,1	3,8	2090	5,3	68	73,2	48,3
690/9	CHxG	10,5	14,4	2,5	565,9	11,7	48	75,9	44,7
724/27	(CHxHC)xY	7,6	17,3	3,7	479	24,5	61	79,3	47,2
724/30	(CHxHC)xY	10,3	15,5	4,4	1225	8,2	70	76	50,5
855/19	HCxN	12,1	16,3	4,2	460	25,2	62	75,1	37,1
855/30	HCxN	8,2	17,1	5,4	1126	8,7	50	86,3	64,4
884/28	CHxG	5,7	14,1	1	226	26,1	77	67,4	42,3
808/8	DxN	10,4	15,8	5	1392	6,7	75	80,2	59,1
422/13	CHxE	9,7	13,6	3,4	279	28,3	52	88,3	47
422/41	CHxE	9,5	15,9	2,4	1440	6,5	54	73,6	33
531/47	HCxA	9,8	15	3,1	898	11,4	48,1	78,5	44,8
538/33	ChxA	9,7	16,2	3,8	1311	8	105,1	78,8	47,8
697/8	(CHxHC)xG	9	13,4	2,6	827	10,8	31,6	76,8	44,3

698/34	(CHxHC)xY	13,5	15,5	6,1	1288	7,6	84,6	82,5	56,6
670/36	(CHxHC)xY	9,3	14,9	3,2	769	11,7	44,4	82,4	43,7
670/19	(CHxHC)xY	12,9	16,3	6,1	1087	10	98,4	84,7	55,8
756/35	DxA	9,1	13,6	3,6	890	10,7	48,6	84,2	45,8
756/42	DxA	11,1	15,6	4,2	1113	9,9	96,5	80,4	37,8
454/6	DxN	11,2	10,3	3,7	729	8,2	44	89,6	50,5
455/13	DxA	12,4	11,3	2,9	846	8,9	51	65,9	41,1
626/35	HCxDa	8,2	14,9	4,7	1091	9,1	50	86,1	62,6
635/17	CHxN	10,4	14	3	1154	7,3	56	64	51,6
744/29	CHxDa	9,5	12,1	3,1	718	10,1	48	88,9	46,1
748/11	CHxN	12,8	14,9	4,3	1012	9,4	68	65	55,2
818/28	HCxA	10	14,5	5	959	9,9	58	89,2	51
818/42	HCxA	9	16,6	3,4	1140	9,6	60	63,7	53,3
622/28	DxA	12	12,5	5,9	1019	7,3	62	81,8	65
670/5	(CHxHC)xY	11,3	13,5	2,6	1093	8,3	68	73,8	29,6
711/10	DxA	9,8	15,3	5,1	532	13,4	51	86	63,5
711/19	DxA	9,3	16,6	2,4	1013	10,9	54	88,7	25,6
837/43	DxA	13	15,2	5,7	955	9	64	82	64,8

Anexo 11: Posicionamiento en el mapa y desequilibrio de ligamiento.

Tabla 12.1: Posicionamiento en el mapa de los 142 SNPs e Indels pertenecientes a los genes candidatos bialélicos que continuaron en el estudio para su asociación con los caracteres de interés. GC= Gen candidato; SNP= Variación encontrada; POS_MAP= posición en el mapa en pares de bases (pb).

	GC	SNP	POS_MAP
GL-1	KG179_FFB1_CL1016_S1.2	A/C	21685874
		C/G	21688462
	KG135_M3117	A/C	21688567
	KG257_PO3_5-5	-/A	39990125
		A/T	50297077
	KG276_PYRKIN	T/A	50297078
GL-2		A/G	6792342
	GLO2	C/T	6792272
	KG288_EgTPase	G/T	8992884
	KG154_DDB1CUL4	C/T	9118625
	KG180_FFB2_C4663_S1.2	A/G	13183483
		C/T	28517614
		T/G	28517750
	KG196_BnC2_10C3-629	A/C	28517852
		G/T	31047861
		A/G	31047886
		T/C	31047899
	KG181_FFB2_C4741_S3	G/A	31047996
	KG182_FFB2_C3566_S9	A/G	31308475

	GC	SNP	POS_MAP
GL-5	KG117_JC55	A/C	14759294
	KG214_EgWRI1.1	T/G	39792450
	KG269_EgNAC	C/G	40852043
	P39_WOS6942	C/T	40852593
GL-6		G/T	33052707
	KG186_FFB6_C2082_S1	C/A	33052739
	KG187_FFB6_C3684_S1	C/T	33630083
GL-7	KG272_OLEOYL	G/T	1586126
	KG142_M2252	A/G	12193246
	KG200_BnC7_3962	T/C	12213973
		A/G	12405701
		T/C	12405716
		C/G	12405731
		A/C	12405749
		C/A	12405773
		T/C	12405776
		A/G	12405785
		C/T	12405791
	KG143_M3256	T/A	12405794

	GC	SNP	POS_MAP
GL-10		A/G	12588990
		T/C	12588991
		T/C	12589000
		T/A	12589007
		T/A	12589008
		T/C	12589011
		A/G	12589013
		G/A	12589018
		A/G	12589019
		C/A	12589020
		A/C	12589026
		C/T	12589065
	CDA24_B44	A/G	12589069
		G/C	22949604
	KG27_BKACPII_1	C/A	22949603
	KG245_EgBRX	C/T	23849092
GL-11	KG261_PO3_5-10	A/G	11857044
	KG192_FFB11_C1_S1	C/T	20181679
	KG193_FFB11_C1741_S3.4	A/T	20333800

	GC	SNP	POS_MAP
GL-13	KG70_BAK1	T/C	25407491
		C/T	25407509
		T/G	27325322
		A/G	27325314
		C/-	27325300
		T/C	27325297
		T/A	27325285
		G/C	27325283
		G/A	27325282
		G/A	27325279
		G/C	27325275
		G/T	27325270
		G/C	27325268
		C/G	27325251
		G/A	27325252
		G/C	27325247
		C/T	27325241
		-/G	27325228
	P74_PAT9	T/C	27325227

	KG183_FFB2_C2_S1	A/G	31455617			C/A	12405812		**KG194_FFB11_C3877_S4**	G/T	20390248			C/A	27325219
	KG268_EgMBAGL2-3	C/G	33901095		**KG140_M43696**	G/A	16118918		**KG270_EgPPGL**	A/T	8353663		**KG155_ELO2**	C/T	1363229
		A/G	59000784		**KG121_MADS11-1**	G/T	16866532			C/T	8353662		**KG157_RPL10**	A/G	5029668
	KG105_PSII1	G/T	59000929		**KG274_LIPOIC**	A/G	18432098		**P77_PAT11**	C/G	9941050		**P13_QM**	A/G	5029668
	CDA22_B42	A/G	3072707	**GL-8**	**KG201_BnC8_761**	G/T	4351911	**GL-12**	**P64_PAT4**	C/T	15018030			A/G	11183745
		T/C	3072703		**KG175_HtC81026C1-144**	C/G	23228525		**CDA43_B67**	C/T	22093559	**GL-14**		T/C	11183795
GL-3	**KG148_MUM4**	C/T	8044652		**P66_PAT6**	A/C	27075522			T/G	28135331		**P78_PAT12**	A/T	11183813
	KG124_AG1	C/T	8101287	**GL-9**	**KG254_PLT2**	A/C	15548277			C/A	28135362		**P79_PAT13**	G/T	14557531
	KG260_PO3_5-9	T/C	31732461			A/G	34725105		**KG29_GLUT1**	G/A	28135380		**KG75_GID1**	C/G	22469941
	KG198_BnC3_792	A/T	32128143			C/T	34725099		**KG195_FFB13_C2168_S1**	A/T	22806533			C/T	23021116
	KG171_HtC4_4489	A/T	4462591			C/T	34725096		**KG70_BAK1**	A/T	25407401		**KG11_M8373**	C/T	23021107
		A/C	27102777			A/T	34725093	**GL-13**		A/G	25407413		**KG125_SQUA3**	C/T	13726804
GL-4		T/C	27102808			G/T	34725068			G/C	25407434	**GL-15**		G/T	19066802
	KG118_JC59	T/A	27102841			C/G	34725053			T/C	25407458			C/G	19066803
	KG80_EIN4	C/G	31379073			T/C	34725050			A/G	25407467		**KG39_BKACPIII**	A/C	19066804
	KG290_EgDSI	A/G	56149310		**P62_PAT2**	T/G	34725017			C/T	25407485			C/T	21166223
												GL-16	**KG106_PSII2**	C/T	21166205

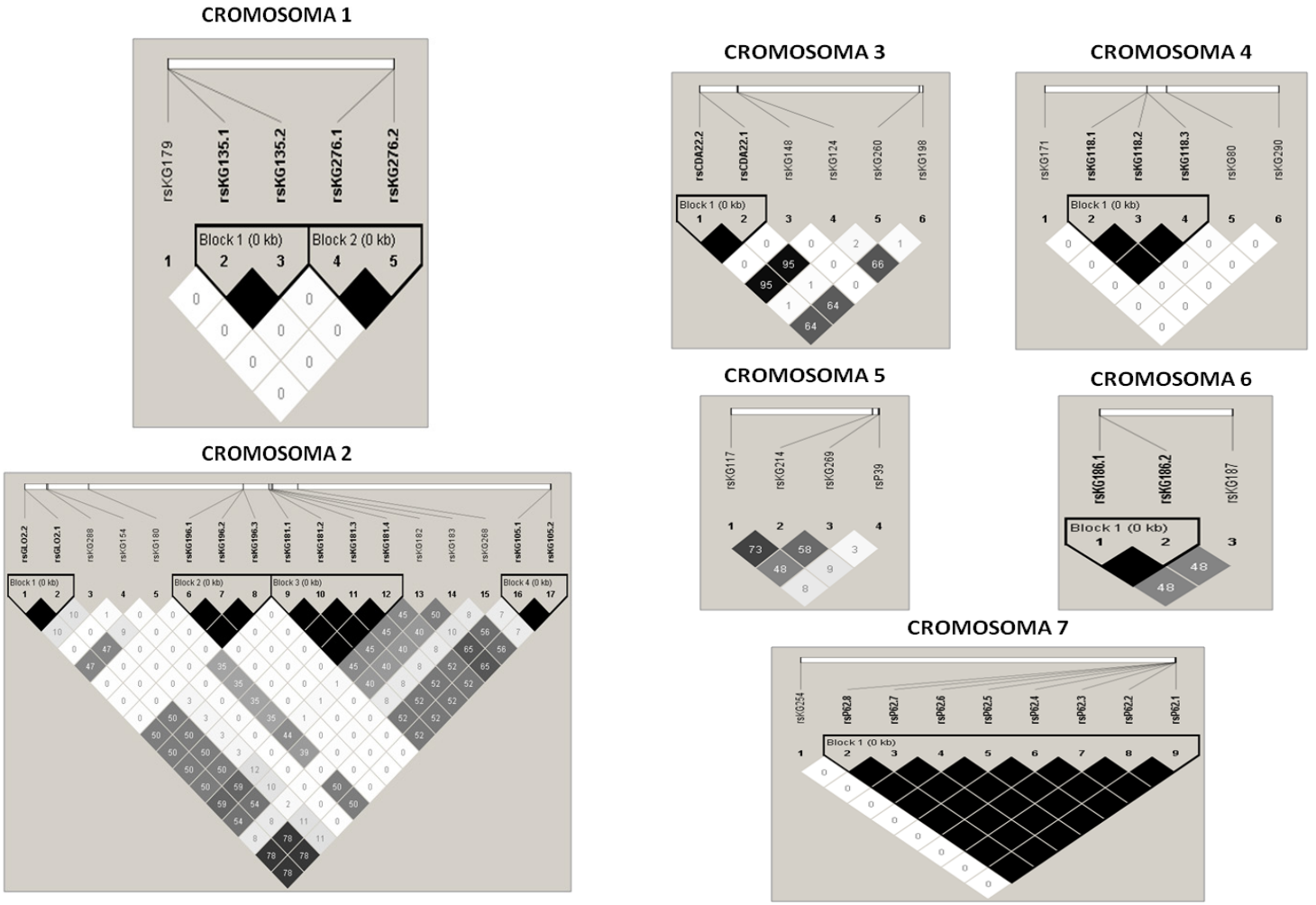


Figura 11.1a: Mapas de DL desde el cromosoma 1 al cromosoma 7.

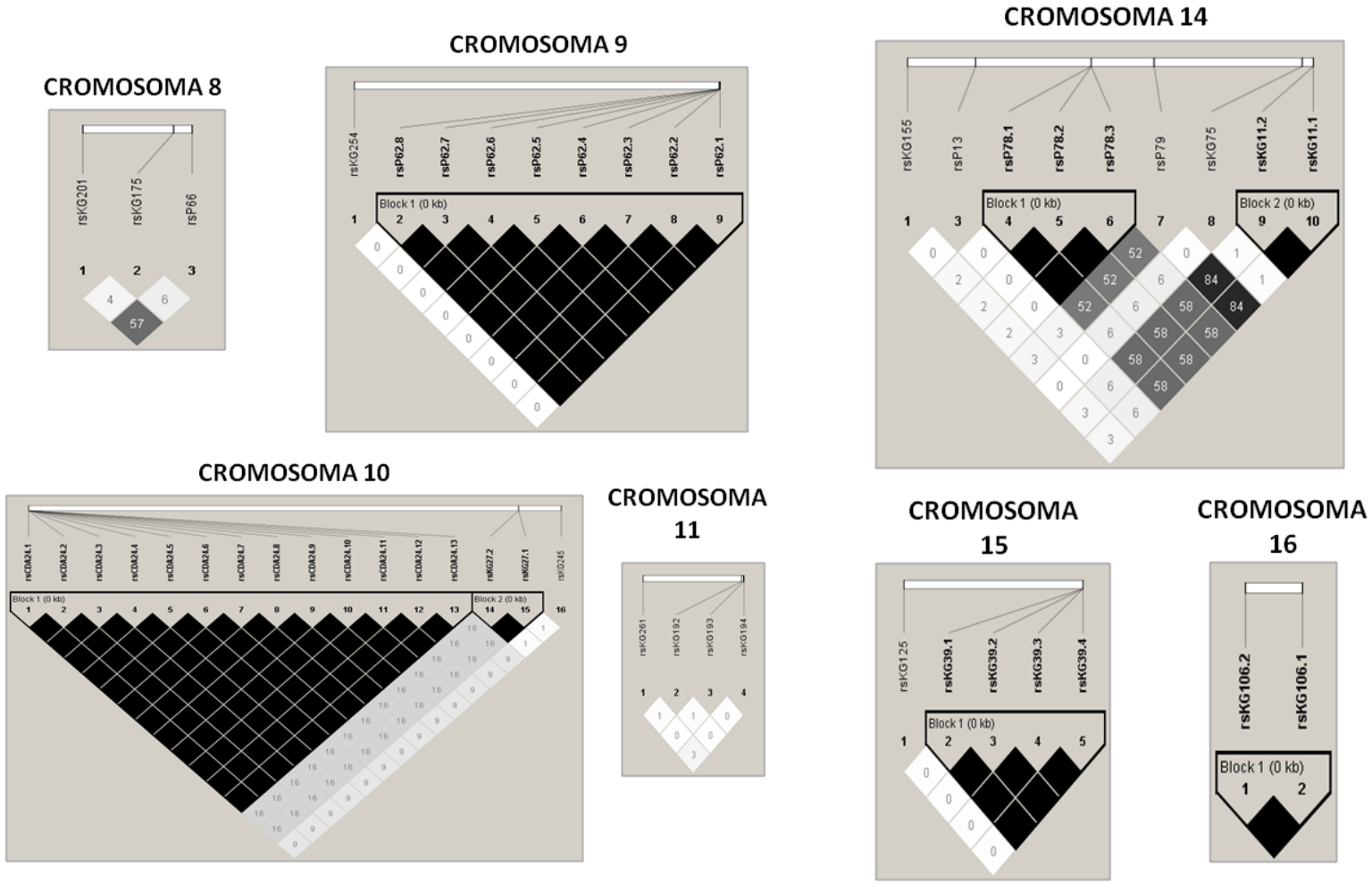


Figura 11.1b: Mapas de DL desde el cromosoma 8 al cromosoma 16.

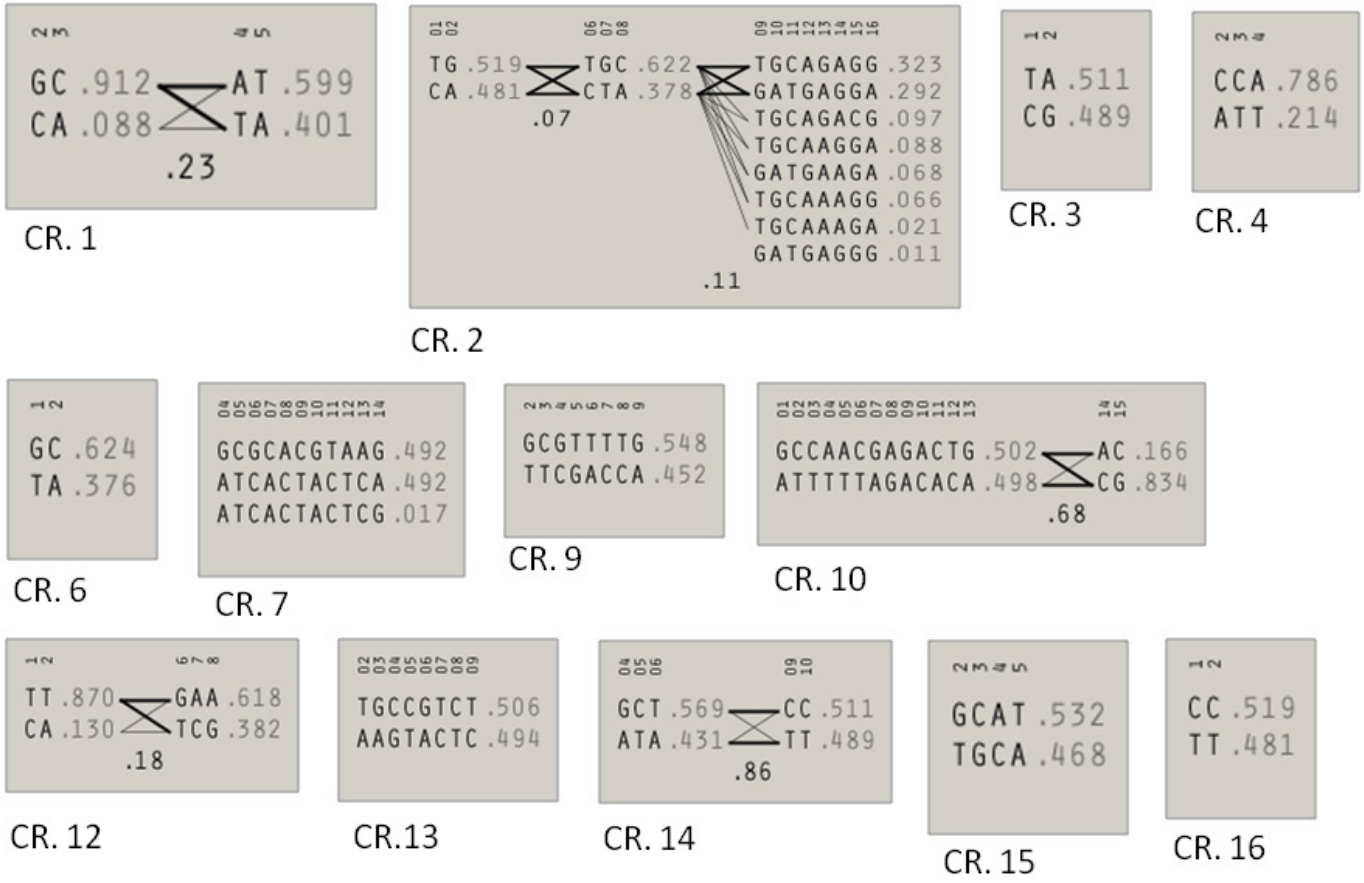


Figura 11.2: Haplotipos formados por los marcadores SNP de algunos genes candidatos a nivel inter-locus y/o intra-locus. En cada cromosoma (CR.n) se muestra los alelos de cada gen candidato sobre 1 (valor total) presentes en la población, y el valor D' (abajo) cuando existe probabilidad de heredarse conjuntamente.

Anexo 12: Análisis de Componentes Principales (ACP) y estructura poblacional.

Tabla 12.1:Matriz P de componentes principales. Esta matriz muestra la correlación de cada componente principal con cada variable (genotipo). COD= genotipo; CRUCE: procedencia de sus parentales. CPn=Componente principal.

<COD>	CRUCE	CP1	CP2	CP3
1111	DxE	1.342	-0.041	0.693
1112	DxE	-0.353	1.017	0.643
1113	DxE	-0.424	0.444	0.841
1114	DxE	-0.601	1.188	-0.627
1121	DxE	0.913	-0.418	0.131
1122	DxE	0.209	-0.240	0.259
1123	DxE	0.151	-0.217	0.003
1124	DxE	0.974	-0.438	0.006
1131	DxA	0.916	-0.622	0.048
1132	DxA	0.842	-0.610	0.040
1133	DxL	0.046	-0.396	-0.176
1134	DxL	-0.740	-0.043	0.005
1141	DxDa	0.082	-0.148	0.098
1142	DxDa	0.370	0.970	-0.807
1143	DxN	1.093	-0.537	0.065
1144	DxN	1.194	-0.129	-0.079
1211	DxN	0.215	-0.235	-0.077
1212	DxA	-0.699	-0.215	0.265
1213	DxA	0.953	-0.409	0.057
1214	DxA	-0.811	-0.377	0.337
1221	DxA	0.440	0.159	-0.069
1222	CHxA	-0.844	-0.204	0.228
1223	CHxA	-0.675	0.766	-0.768
1224	DxA	-0.692	-0.161	0.059
1231	DxA	0.034	-0.435	0.227
1232	CHxN	-0.842	1.153	-0.607
1233	CHxN	-0.582	-0.281	0.135
1234	CHxN	0.338	-0.031	-0.185
1241	CHxN	0.239	0.375	-0.267
1242	DxE	0.078	-0.483	0.216
1243	DxE	0.171	-0.242	0.206
1244	DxE	-0.739	-0.385	0.118
1311	DxA	0.977	-0.486	-0.005
1312	DxA	0.338	0.197	-0.124
1313	HCxA	-0.867	0.736	-0.592
1314	HCxA	-0.739	-0.041	-0.071
1321	DxE	0.994	-0.415	0.051
1322	DxE	0.098	-0.448	0.286
1323	CHxE	0.138	0.008	0.178
1324	CHxE	0.098	0.094	-0.122
1331	CHxG	0.348	0.225	-0.180
1332	CHxE	0.323	0.006	-0.099
1333	CHxE	0.177	0.036	0.036
1334	DxE	-0.476	0.572	0.792
1341	DxE	0.124	0.824	-0.710
1342	CHxA	-0.784	-0.277	0.103
1343	CHxA	1.036	-0.265	-0.087
1344	HCxA	-0.721	-0.237	0.225
1411	HCxA	-0.753	-0.311	0.178
1412	CHxA	0.015	-0.332	0.149
1413	CHxA	0.340	0.346	0.915
1414	CHxDa	-0.540	1.807	-0.161
1421	CHxDa	-0.343	0.936	0.830
1422	CHxDa	-0.715	0.889	-0.459
1423	CHxG	0.456	1.733	-0.242
1424	CHxG	-0.790	0.354	-0.042
1431	CHxDa	-0.675	0.108	0.168
1432	DxL	-0.589	0.570	1.057
1433	DxL	-0.738	-0.102	0.377
1434	DxDa	-0.826	0.078	0.248
1441	DxDa	0.198	-0.016	-0.210
1442	CHxA	0.054	-0.191	0.119
1443	HCxE	-0.608	0.611	-0.580
1444	CHxA	0.452	0.428	0.828
2111	HCxE	0.288	-0.048	-0.111
2112	CHxN	-0.542	0.344	0.015
2113	CHxN	0.289	1.003	-0.894
2114	CHxE	1.233	0.840	-1.105
2121	CHxE	-1.043	-0.536	-0.650
2122	CHxA	-0.590	1.019	-0.700
2123	CHxA	-0.603	0.368	0.708
2124	DxA	-0.792	-0.261	0.137
2131	DxA	0.786	0.249	-0.514
2132	HCxG	1.031	-0.165	-0.286
2133	HCxG	1.095	0.120	-0.247
2134	HCxDa	0.177	-0.095	-0.026
2141	HCxDa	0.289	0.443	0.720
2142	HCxDa	0.087	-1.230	0.973
2143	CHxN	-0.567	0.118	0.235
2144	CHxN	0.273	0.822	-0.964
2211	DxE	0.174	-0.104	-0.134
2212	DxE	0.294	0.040	-0.168

<COD>	CRUCE	CP1	CP2	CP3
2213	CHxN	0.202	0.007	-0.050
2214	CHxN	-0.575	0.290	0.116
2221	HCxN	0.241	-0.069	-0.071
2222	HCxN	-0.662	0.128	0.196
2223	DxN	0.284	0.007	-0.124
2224	DxN	0.419	1.147	-1.176
2231	DxN	0.349	0.194	-0.234
2232	DxE	0.090	0.166	-0.259
2233	DxE	0.232	0.127	-0.209
2234	DxE	1.123	0.155	-0.226
2241	DxG	0.192	-0.036	-0.139
2242	DxG	1.003	-0.190	-0.245
2243	DxDa	0.190	0.010	-0.186
2244	DxDa	-0.558	0.047	-0.138
2311	DxN	1.236	-0.370	0.087
2312	DxN	1.065	-1.232	-0.725
2313	(CHxHC)xY	-0.861	-1.052	-0.534
2314	(CHxHC)xY	-0.343	0.509	0.973
2321	DxE	1.201	-0.089	0.979
2322	DxE	1.165	-0.343	-0.025
2323	DxE	0.851	-1.210	-0.752
2324	DxE	1.405	0.500	0.677
2331	DxN	0.481	0.202	0.648
2332	DxN	-0.904	-0.848	-0.512
2333	DxN	0.016	-0.987	-0.615
2334	DxA	0.167	-0.886	-0.076
2341	DxA	-1.034	-0.975	-0.419
2342	DxA	0.036	-1.264	-0.795
2343	DxDa	-0.041	-0.352	0.242
2344	DxA	1.321	0.346	0.784
2411	DxA	0.155	-0.208	-0.153
2412	CHxG	0.340	0.429	0.660
2413	CHxG	-0.711	0.058	-0.095
2414	CHxG	0.928	-0.294	-0.333
2421	CHxG	0.067	-0.245	0.158
2422	DxE	1.226	-0.133	-0.156
2423	DxE	1.028	-0.218	-0.078
2424	(CHxHC)xG	-0.716	0.125	-0.169
2431	(CHxHC)xY	-0.540	-0.198	-0.173
2432	(CHxHC)xY	-0.449	0.158	-0.024
2433	DxA	-0.742	-0.314	0.018
2434	DxA	-0.763	0.081	-0.049
2441	HCxG	1.020	-0.271	-0.086
2442	(CHxHC)xY	-0.566	0.243	-0.087
2443	(CHxHC)xY	0.392	0.067	-0.172
2444	HCxG	0.161	-0.162	-0.015
3111	HCxE	-0.541	0.068	0.279
3112	HCxE	0.186	-0.186	0.186
3113	DxY	-0.736	-0.196	0.120
3114	CHxN	0.251	0.151	-0.200
3121	CHxN	0.149	0.029	-0.018
3122	CHxDa	-0.631	-0.103	0.235
3123	CHxDa	0.228	-0.215	0.276
3124	CHxN	-0.640	-0.073	0.023
3131	CHxN	-0.710	-0.037	0.197
3132	CHxN	-0.578	0.089	-0.097
3133	DxA	-0.704	-0.529	0.256
3134	DxA	-0.516	0.220	-0.092
3141	DxA	-0.736	-0.275	0.264
3142	DxA	1.163	-0.117	0.717
3143	DxDa	1.023	-0.207	0.044
3144	DxDa	0.093	-0.364	0.049
3211	DxY	1.157	-0.196	-0.156
3212	DxY	-0.636	-0.463	-0.712
3213	DxY	-0.355	0.664	0.679
3214	DxY	-0.794	-0.279	-0.547
3221	DxG	0.345	0.469	0.771
3222	DxG	0.310	-0.394	-0.085
3223	DxN	1.455	0.618	0.454
3224	DxN	0.433	0.121	-0.091
3231	HCxE	0.228	-0.065	0.029
3232	HCxE	-0.871	-1.011	-0.626
3233	DxY	-0.439	0.389	0.947
3234	CHxN	0.327	0.159	-0.230
3241	HCxG	0.387	0.524	0.769
3242	HCxA	-0.030	-1.007	-0.815
3243	CHxE	0.233	-0.033	0.057
3244	CHxE	-0.575	0.350	1.043
3311	DxA	-0.807	-0.453	0.316
3312	DxA	-0.654	-0.283	0.274
3313	DxY	-0.541	-0.029	0.225
3314	DxY	1.163	-0.359	-0.072
3321	HCxN	0.351	-0.223	-0.035
3322	HCxN	-0.604	-0.146	0.180
3323	HCxN	0.271	0.063	0.080
3324	HCxN	0.346	0.043	-0.009
3331	CHxN	0.132	-0.193	0.140
3332	CHxN	0.321	0.053	-0.265
3333	DxDa	-0.966	0.020	0.128
3334	DxDa	-0.637	-0.003	0.127

<COD>	CRUCE	CP1	CP2	CP3
3341	DxDa	0.300	0.164	-0.222
3342	HCxA	0.154	-0.455	0.221
3343	HCxA	0.046	-0.357	0.255
3344	HCxN	-0.572	-0.225	0.098
3411	HCxN	0.246	-0.137	0.049
3412	HCxA	-0.682	-0.246	0.193
3413	HCxA	-0.735	-0.269	0.310
3414	CHxG	-0.651	0.050	-0.257
3421	DxN	0.223	-0.301	0.045
3422	DxN	-0.591	1.021	-0.642
3423	CHxN	1.089	-0.201	-0.064
3424	CHxN	-0.610	0.138	-0.019
3431	MxN	0.191	0.025	-0.094
3432	MxN	0.244	0.547	-0.749
3433	DxDa	0.078	-0.178	0.073
3434	DxN	1.273	-1.525	0.975
3441	DxN	0.221	-0.180	0.080
3442	CHxDa	-0.602	-0.063	0.126
3443	CHxDa	-0.740	-0.067	0.152
3444	DxA	-0.692	0.001	-0.052
4111	DxA	0.108	-0.140	-0.120
4112	(CHxHC)xDa	0.944	-0.674	-1.149
4113	(CHxHC)xDa	-0.830	-0.466	-0.831
4114	DxE	1.074	0.900	0.013
4121	DxDa	-0.722	-0.332	-0.960
4122	CHxN	0.190	0.061	-0.145
4124	CHxG	0.359	1.073	0.454
4131	HCxA	0.296	0.164	0.901
4132	HCxG	-0.008	-0.831	-0.879
4133	HCxG	0.800	-0.742	-0.842
4134	CHxG	-0.532	-0.172	-1.102
4141	CHxE	-0.853	-0.852	-0.699
4142	CHxE	-0.209	-0.918	-0.521
4143	DxE	-0.550	0.185	0.119
4144	CHxLa	-0.434	0.479	0.922

<COD>	CRUCE	CP1	CP2	CP3
4211	DxDa	-0.555	0.080	0.127
4212	DxDa	0.254	0.226	1.009
4213	(CHxHC)xY	-0.558	0.117	-0.126
4214	(CHxHC)xY	1.222	1.201	-0.070
4221	DxY	-0.488	0.245	1.163
4222	DxY	-0.531	0.590	1.078
4223	DxA	1.214	0.216	0.487
4224	DxA	1.127	0.009	0.779
4231	HCxA	-0.879	-0.972	-0.711
4232	HCxA	0.438	0.343	0.888
4233	CHxG	0.363	0.446	0.518
4234	CHxG	0.009	0.420	-1.671
4241	DxE	-0.846	-1.848	0.191
4242	DxE	-0.782	0.220	-1.506
4243	DxG	-0.398	0.596	1.086
4244	DxG	-0.018	0.164	-1.662
4311	DxA	-0.809	0.657	-0.541
4312	DxA	-0.851	-0.410	0.347
4313	DxDa	0.332	1.140	-0.881
4314	DxDa	-0.708	-0.045	0.218
4321	HCxE	-0.815	-1.177	0.993
4322	HCxE	0.318	1.047	-0.882
4323	DxN	0.304	0.229	-0.198
4324	DxN	-0.565	0.363	-0.130
4331	HCxDa	-0.437	0.075	0.045
4332	HCxDa	0.088	-1.121	1.029
4333	CHxG	-0.573	0.777	0.757
4342	HCxDa	0.098	0.959	-0.770
4343	DxDa	-0.847	0.092	0.374
4344	DxDa	-0.571	0.358	1.188
4411	HCxN	1.190	0.736	-0.886

Tabla 12.2: Eigenvalor explicado por cada componente principal y la varianza individual y acumulada explicada para cada componente.

PC	Eigenvalor	Proporción del total	Proporción acumulada
1	0.43661	0.09638	0.09638
2	0.29481	0.06508	0.16146
3	0.27867	0.06151	0.22297
4	0.24222	0.05347	0.27644
5	0.2341	0.05168	0.32811
6	0.19729	0.04355	0.37166
7	0.17147	0.03785	0.40951
8	0.16443	0.0363	0.44581
9	0.15139	0.03342	0.47922
10	0.14546	0.03211	0.51133
11	0.1298	0.02865	0.53999
12	0.12384	0.02734	0.56732
13	0.11411	0.02519	0.59251
14	0.10676	0.02357	0.61608
15	0.1036	0.02287	0.63895
16	0.09789	0.02161	0.66056
17	0.09177	0.02026	0.68081
18	0.08233	0.01817	0.69899
19	0.08015	0.01769	0.71668
20	0.07321	0.01616	0.73284
21	0.0704	0.01554	0.74838
22	0.06912	0.01526	0.76364
23	0.06685	0.01476	0.7784
24	0.06009	0.01326	0.79166
25	0.05526	0.0122	0.80386
26	0.05394	0.01191	0.81576
27	0.05139	0.01134	0.82711
28	0.04869	0.01075	0.83786
29	0.04805	0.01061	0.84846
30	0.04434	0.00979	0.85825
31	0.04546	0.01003	0.86829
32	7.12E-04	1.57E-04	0.86844
33	0.00147	3.26E-04	0.86877
34	0.00267	5.89E-04	0.86936
35	0.00189	4.18E-04	0.86978
36	0.002	4.42E-04	0.87022
37	0.00513	0.00113	0.87135
38	0.0064	0.00141	0.87276
39	0.00723	0.00159	0.87436
40	0.00882	0.00195	0.8763
41	0.00854	0.00189	0.87819
42	0.01056	0.00233	0.88052

43	0.01184	0.00261	0.88314
44	0.04186	0.00924	0.89238
45	0.01436	0.00317	0.89555
46	0.03924	0.00866	0.90421
47	0.0384	0.00848	0.91268
48	0.01714	0.00378	0.91647
49	0.03565	0.00787	0.92434
50	0.03492	0.00771	0.93205
51	0.03292	0.00727	0.93931
52	0.03405	0.00752	0.94683
53	0.0304	0.00671	0.95354
54	0.02972	0.00656	0.9601
55	0.01848	0.00408	0.96418
56	0.02666	0.00588	0.97007
57	0.02578	0.00569	0.97576
58	0.01982	0.00437	0.98013
59	0.02092	0.00462	0.98475
60	0.02186	0.00482	0.98957
61	0.02353	0.00519	0.99477
62	0.02369	0.00523	1

Tabla 12.3: **Matriz Q** de estructura poblacional.

<COD>	Q1	Q2
1111	0.992	0.008
1112	0.02	0.98
1113	0.017	0.983
1114	0.017	0.983
1121	0.992	0.008
1122	0.967	0.033
1123	0.946	0.054
1124	0.991	0.009
1131	0.991	0.009
1132	0.992	0.008
1133	0.932	0.068
1134	0.013	0.987
1141	0.917	0.083
1142	0.963	0.037
1143	0.994	0.006
1144	0.994	0.006
1211	0.952	0.048
1212	0.014	0.986
1213	0.992	0.008
1214	0.011	0.989
1221	0.968	0.032
1222	0.014	0.986
1223	0.013	0.987
1224	0.013	0.987
1231	0.946	0.054
1232	0.016	0.984
1233	0.012	0.988
1234	0.956	0.044
1241	0.958	0.042
1242	0.935	0.065
1243	0.946	0.054
1244	0.017	0.983
1311	0.992	0.008
1312	0.948	0.052
1313	0.008	0.992
1314	0.012	0.988
1321	0.993	0.007
1322	0.946	0.054
1323	0.934	0.066
1324	0.949	0.051
1331	0.952	0.048
1332	0.948	0.052

1333	0.959	0.041
1334	0.012	0.988
1341	0.934	0.066
1342	0.013	0.987
1343	0.991	0.009
1344	0.01	0.99
1411	0.012	0.988
1412	0.926	0.074
1413	0.953	0.047
1414	0.015	0.985
1421	0.041	0.959
1422	0.013	0.987
1423	0.945	0.055
1424	0.02	0.98
1431	0.014	0.986
1432	0.015	0.985
1433	0.015	0.985
1434	0.013	0.987
1441	0.946	0.054
1442	0.919	0.081
1443	0.014	0.986
1444	0.952	0.048
2111	0.954	0.046
2112	0.02	0.98
2113	0.938	0.062
2114	0.993	0.007
2121	0.009	0.991
2122	0.018	0.982
2123	0.014	0.986
2124	0.012	0.988
2131	0.992	0.008
2132	0.992	0.008
2133	0.991	0.009
2134	0.938	0.062
2141	0.937	0.063
2142	0.946	0.054
2143	0.027	0.973
2144	0.961	0.039
2211	0.947	0.053
2212	0.958	0.042
2213	0.95	0.05
2214	0.024	0.976
2221	0.959	0.041
2222	0.01	0.99
2223	0.958	0.042
2224	0.949	0.051
2231	0.947	0.053
2232	0.963	0.037
2233	0.964	0.036
2234	0.992	0.008
2241	0.936	0.064
2242	0.994	0.006
2243	0.912	0.088
2244	0.017	0.983
2311	0.993	0.007
2312	0.993	0.007
2313	0.01	0.99
2314	0.01	0.99
2321	0.992	0.008
2322	0.993	0.007
2323	0.992	0.008
2324	0.993	0.007
2331	0.941	0.059
2332	0.01	0.99
2333	0.926	0.074
2334	0.916	0.084
2341	0.012	0.988
2342	0.872	0.128
2343	0.896	0.104
2344	0.992	0.008
2411	0.945	0.055
2412	0.94	0.06
2413	0.013	0.987
2414	0.991	0.009
2421	0.939	0.061
2422	0.993	0.007
2423	0.992	0.008
2424	0.018	0.982
2431	0.015	0.985
2432	0.014	0.986
2433	0.011	0.989
2434	0.014	0.986
2441	0.993	0.007
2442	0.018	0.982
2443	0.958	0.042
2444	0.955	0.045
3111	0.023	0.977
3112	0.955	0.045
3113	0.01	0.99
3114	0.95	0.05
3121	0.944	0.056
3122	0.025	0.975

3123	0.954	0.046
3124	0.012	0.988
3131	0.011	0.989
3132	0.015	0.985
3133	0.015	0.985
3134	0.02	0.98
3141	0.01	0.99
3142	0.99	0.01
3143	0.991	0.009
3144	0.937	0.063
3211	0.993	0.007
3212	0.022	0.978
3213	0.016	0.984
3214	0.017	0.983
3221	0.949	0.051
3222	0.968	0.032
3223	0.994	0.006
3224	0.967	0.033
3231	0.946	0.054
3232	0.01	0.99
3233	0.011	0.989
3234	0.946	0.054
3241	0.948	0.052
3242	0.936	0.064
3243	0.946	0.054
3244	0.024	0.976
3311	0.012	0.988
3312	0.019	0.981
3313	0.016	0.984
3314	0.992	0.008
3321	0.962	0.038
3322	0.024	0.976
3323	0.95	0.05
3324	0.952	0.048
3331	0.937	0.063
3332	0.938	0.062
3333	0.012	0.988
3334	0.01	0.99
3341	0.935	0.065
3342	0.924	0.076
3343	0.905	0.095
3344	0.014	0.986
3411	0.93	0.07
3412	0.021	0.979
3413	0.011	0.989
3414	0.02	0.98

3421	0.937	0.063
3422	0.021	0.979
3423	0.993	0.007
3424	0.014	0.986
3431	0.936	0.064
3432	0.957	0.043
3433	0.939	0.061
3434	0.99	0.01
3441	0.948	0.052
3442	0.018	0.982
3443	0.021	0.979
3444	0.032	0.968
4111	0.933	0.067
4112	0.992	0.008
4113	0.016	0.984
4114	0.992	0.008
4121	0.025	0.975
4122	0.962	0.038
4124	0.953	0.047
4131	0.948	0.052
4132	0.946	0.054
4133	0.993	0.007
4134	0.022	0.978
4141	0.014	0.986
4142	0.931	0.069
4143	0.023	0.977
4144	0.019	0.981
4211	0.018	0.982
4212	0.934	0.066
4213	0.02	0.98
4214	0.992	0.008
4221	0.014	0.986
4222	0.01	0.99
4223	0.992	0.008
4224	0.992	0.008
4231	0.012	0.988
4232	0.965	0.035
4233	0.917	0.083
4234	0.92	0.08
4241	0.019	0.981
4242	0.015	0.985
4243	0.032	0.968
4244	0.958	0.042
4311	0.01	0.99
4312	0.012	0.988
4313	0.948	0.052

4314	0.013	0.987
4321	0.013	0.987
4322	0.935	0.065
4323	0.933	0.067
4324	0.032	0.968
4331	0.015	0.985
4332	0.944	0.056
4333	0.015	0.985
4342	0.918	0.082
4343	0.011	0.989
4344	0.013	0.987
4411	0.994	0.006

BIBLIOGRAFÍA

Bibliografía

- Abdullah, N., Yusop, M. R., Ithnin, M., Saleh, G., & Latif, M. A. (2011). Genetic variability of oil palm parental genotypes and performance of its' progenies as revealed by molecular markers and quantitative traits. Comptes Rendus Biologies, 334(4), 290-299.
- Abdullah, S. N. A., Cheah, S. C., & Murphy, D. J. (2002). Isolation and characterisation of two divergent type 3 metallothioneins from oil palm, *Elaeis guineensis*. Plant Physiology and Biochemistry, 40(3), 255-263.
- Abdullah,M.O. and Kulaveerasingam,H. (2002). Oil palm (genus Elaeis) peroxiredoxin gene and uses thereof. EP 1217068. MPOB
- Abdurakhmonov, I. Y., & Abdukarimov, A. (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. International Journal of Plant Genomics, 2008.
- Aberlenc-Bertossi, F., Chabrillange, N., Duval, Y., & Tregear, J. (2008). Contrasting globulin and cysteine proteinase gene expression patterns reveal fundamental developmental differences between zygotic and somatic embryos of oil palm. Tree physiology,28(8), 1157-1167.
- Adam, H., Jouannic, S., Morcillo, F., Richaud, F., Duval, Y., & Tregear, J. W. (2006). MADS box genes in oil palm (*Elaeis guineensis*): patterns in the evolution of the SQUAMOSA, DEFICIENS, GLOBOSA, AGAMOUS, and SEPALLATA subfamilies. Journal of Molecular Evolution, 62(1), 15-31.
- Adam, H., Jouannic, S., Morcillo, F., Verdeil, J. L., Duval, Y., & Tregear, J. W. (2007). Determination of flower structure in *Elaeis guineensis*: do palms use the same homeotic genes as other species?.Annals of botany, 100(1), 1-12.
- Adam, H., Jouannic, S., Orieux, Y., Morcillo, F., Richaud, F., Duval, Y., & Tregear, J. W. (2007). Functional characterization of MADS box genes involved in the determination of oil palm flower structure. Journal of experimental Botany, 58(6), 1245-1259.
- Adams, M. D., & Kelley, J. M. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252(5013), 1651.
- Agrama, H. A., & Eizenga, G. C. (2008). Molecular diversity and genome-wide linkage disequilibrium patterns in a worldwide collection of *Oryza sativa* and its wild relatives. Euphytica, 160(3), 339-355.
- Ahn, C. S., Ahn, H. K., & Pai, H. S. (2014). Overexpression of the PP2A regulatory subunit Tap46 leads to enhanced plant growth through stimulation of the TOR signalling pathway. Journal of experimental botany, eru438.
- Ainsworth, E. A., & Bush, D. R. (2011). Carbohydrate export from the leaf: a highly regulated process and target to enhance photosynthesis and productivity. Plant physiology, 155(1), 64-69.
- Akihiro, T., Umezawa, T., Ueki, C., Lobna, B. M., Mizuno, K., Ohta, M., & Fujimura, T. (2006). Genome wide cDNA-AFLP analysis of genes rapidly induced by combined sucrose and ABA treatment in rice cultured cells. FEBS letters, 580(25), 5947-5952.
- Al-Mssallem, I. S., Hu, S., Zhang, X., Lin, Q., Liu, W., Tan, J., Yu, X., Liu, J., Pan, L.[,[...]] & Yin, Y. (2013). Genome sequence of the date palm *Phoenix dactylifera* L. Nature communications, 4.
- Al-Obaidi, J. R., Mohd–Yusuf, Y., & Othman, R. Y. (2013). Characterization and isolation of oil palm lipid transfer protein (EgLTP) gene putatively responsible for defense against fungal infection (*Ganoderma boninense*) during basal stem rot infection. Journal of Medicinal Plants Research, 7(38), 2833-2840.

- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.
- Alvarez, M. E., Nota, F., & Cambiagno, D. A. (2010). Epigenetic control of plant immunity. Molecular Plant Pathology, 11(4), 563-576.
- Amir, R., Hacham, Y., & Galili, G. (2002). Cystathionine γ-synthase and threonine synthase operate in concert to regulate carbon flow towards methionine in plants. Trends in plant science, 7(4), 153-156.
- Andersen, J. R., & Lübberstedt, T. (2003). Functional markers in plants. Trends in plant science, 8(11), 554-560.
- Anderson, R. A., Boronenkov, I. V., Doughman, S. D., Kunz, J., & Loijens, J. C. (1999). Phosphatidylinositol phosphate kinases, a multifaceted family of signaling enzymes. Journal of Biological Chemistry, 274(15), 9907-9910.
- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Reference Source.
- Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics, 3(4), 299-309.
- Argyros, R. D., Mathews, D. E., Chiang, Y. H., Palmer, C. M., Thibault, D. M., Etheridge, N., [[...]] & Schaller, G. E. (2008). Type B response regulators of Arabidopsis play key roles in cytokinin signaling and plant development. The Plant Cell, 20(8), 2102-2116.
- 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 Genetic Resources, 12(03), 341-348.
- Asano, K., Hirano, K., Ueguchi-Tanaka, M., Angeles-Shim, R. B., Komura, T., Satoh, H., [[...]] & Ashikari, M. (2009). Isolation and characterization of dominant dwarf mutants, Slr1-d, in rice. Molecular Genetics and Genomics, 281(2), 223-231.
- Aubourg, S., Kreis, M., & Lecharny, A. (1999). The DEAD box RNA helicase family in *Arabidopsis thaliana*. Nucleic Acids Research, 27(2), 628-636
- Avila-Ospina, L., Moison, M., Yoshimoto, K., & Masclaux-Daubresse, C. (2014). Autophagy, plant senescence, and nutrient recycling. Journal of experimental botany, eru039.
- Avonce, N., Leyman, B., Mascorro-Gallardo, J. O., Van Dijck, P., Thevelein, J. M., & Iturriaga, G. (2004). The Arabidopsis trehalose-6-P synthase AtTPS1 gene is a regulator of glucose, abscisic acid, and stress signaling. Plant Physiology, 136(3), 3649-3659.
- Azam, S., Thakur, V., Ruperao, P., Shah, T., Balaji, J., Amindala, B., [...] & Jones, J. D. (2012). Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (*Cicer arietinum*; Fabaceae), a crop species without a reference genome. American journal of botany, 99(2), 186-192.
- Babu, B. K., Mathur, R. K., Kumar, P. N., Ramajayam, D., Ravichandran, G., Venu, M. V. B., & Babu, S. S. (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(2), e0171933.
- Bachem, C. W., Hoeven, R. S., Bruijn, S. M., Vreugdenhil, D., Zabeau, M., & Visser, R. G. (1996). Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. The plant journal, 9(5), 745-753.
- Bachem, C. W., Oomen, R. J., & Visser, R. G. (1998). Transcript imaging with cDNA-AFLP: a step-by-step protocol. Plant Molecular Biology Reporter, 16(2), 157-157.

- Bak, S., Beisson, F., Bishop, G., Hamberger, B., Höfer, R., Paquette, S., & Werck-Reichhart, D. (2011). Cytochromes P450. The Arabidopsis Book, e0144.
- Bakoume, C. (2006). Genetic diversity of natural oil palm (*Elaeis guineensis* Jacq.) populations using microsatellite markers.
- Balding, D. J. (2006). A tutorial on statistical methods for population association studies. Nature Reviews Genetics, 7(10), 781-791.
- Bao, Y., Song, W. M., Jin, Y. L., Jiang, C. M., Yang, Y., Li, B., [...] & Zhang, H. X. (2014). Characterization of Arabidopsis Tubby-like proteins and redundant function of AtTLP3 and AtTLP9 in plant response to ABA and osmotic stress. Plant molecular biology, 86(4-5), 471-483.
- Barcelos, E., Rios, S. D. A., Cunha, R. N., Lopes, R., Motoike, S. Y., Babiychuk, E., [...] & Kushnir, S. (2015). Oil palm natural diversity and the potential for yield improvement. Frontiers in plant science, 6, 190.
- Barrett, J. C., Fry, B., Maller, J. D. M. J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics, 21(2), 263-265.
- Basiron, Y. (2005). Palm oil. Bailey's industrial oil and fat products.
- Bates, P.D., Durrett,T.P., Ohlrogge, J.B., Pollard, M. (2009) Analysis of acyl fluxes through multiple pathways of triacylglycerol synthesis in developing soybean embryos. Plant Physiol. 150: 55–72
- Baud S, Lepiniec L (2010) Physiological and developmental regulation of seed oil production. Prog Lipid Res 49: 235–249
- Beavis, W. D. (1998). QTL analyses: power, precision, and accuracy. Molecular dissection of complex traits, 1998, 145-162.
- Becker, A., & Theißen, G. (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Molecular phylogenetics and evolution, 29(3), 464-489.
- Beckmann, J. S., & Soller, M. (1986). Restriction fragment length polymorphisms in plant genetic improvement. Oxford surveys of plant molecular and cell biology.
- Beirnaert, A. & Vanderweyen, R. Contribution a` l'étude genétique et biométrique des variétiés d'Elaeis guineensis Jacq. Publ. Inst.Nat. Etude Agron. Congo Belge. Ser. Sci. 27, 1–101 (1941).
- Beirnaert, A. (1935). Introduction à la Biologie florale du Palmier Elaeis. Organisation de l'inflorescence chez le Palmier à huile. Revue de botanique appliquée et d'agriculture coloniale, 15(172), 1091-1108.
- Berr, A., & Shen, W. H. (2010). Molecular mechanisms in epigenetic regulation of plant growth and development. In Plant Developmental Biology-Biotechnological Perspectives (pp. 325-344). Springer Berlin Heidelberg.
- Berr, A., Shafiq, S., & Shen, W. H. (2011). Histone modifications in transcriptional activation during plant development. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1809(10), 567-576.
- Bhasin, M., Reinherz, E. L., & Reche, P. A. (2006). Recognition and classification of histones using support vector machine. Journal of Computational Biology, 13(1), 102-112.
- Bhattacharyya, M., Upadhyay, R., & Vishveshwara, S. (2012). Interaction signatures stabilizing the nad (p)-binding rossmann fold: a structure network approach. PloS one, 7(12), e51676.

- Bi, J., Wang, W., Liu, Z., Huang, X., Jiang, Q., Liu, G., [...] & Huang, X. (2014). Seipin promotes adipose tissue fat storage through the ER Ca 2+-ATPase SERCA. Cell metabolism, 19(5), 861-871.
- Billotte, N., Jourjon, M. F., Marseillac, N., Berger, A., Flori, A., Asmady, H., [...] & Cheah, S. C. (2010). QTL detection by multi-parent linkage mapping in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 120(8), 1673-1687.
- Billotte, N., Marseillac, N., Risterucci, A. M., Adon, B., Brottier, P., Baurens, F. C., [...] & Amblard, P. (2005). Microsatellite-based high density linkage map in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 110(4), 754-765.
- Bindschedler, L. V., Dewdney, J., Blee, K. A., Stone, J. M., Asai, T., Plotnikov, J., [...] & Ausubel, F. M. (2006). Peroxidase-dependent apoplastic oxidative burst in Arabidopsis required for pathogen resistance. The Plant Journal, 47(6), 851-863.
- Blankenberg, D., Gordon, A., Von Kuster, G., Coraor, N., Taylor, J., Nekrutenko, A., & Galaxy Team. (2010). Manipulation of FASTQ data with Galaxy. Bioinformatics, 26(14), 1783-1785.
- Bolwell, G. P., Bindschedler, L. V., Blee, K. A., Butt, V. S., Davies, D. R., Gardner, S. L., [...] & Minibayeva, F. (2002). The apoplastic oxidative burst in response to biotic stress in plants: a three-component system. Journal of Experimental Botany, 53(372), 1367-1376.
- Boopathi, N. M. (2012). Genetic mapping and marker assisted selection: basics, practice and benefits. Springer Science & Business Media.
- Borevitz, J. O., & Chory, J. (2004). Genomics tools for QTL analysis and gene discovery. Current Opinion in Plant Biology, 7(2), 132-136.
- Bourgis, F., Kilaru, A., Cao, X., Ngando-Ebongue, G. F., Drira, N., Ohlrogge, J. B., & Arondel, V. (2011). Comparative transcriptome and metabolite analysis of oil palm and date palm mesocarp that differ dramatically in carbon partitioning. Proceedings of the National Academy of Sciences, 108(30), 12527-12532.
- Bozak, K. R., Yu, H. O. N. G., Sirevåg, R., & Christoffersen, R. E. (1990). Sequence analysis of ripening-related cytochrome P-450 cDNAs from avocado fruit. Proceedings of the National Academy of Sciences, 87(10), 3904-3908.
- Bozza, P. T., Bakker-Abreu, I., Navarro-Xavier, R. A., & Bandeira-Melo, C. (2011). Lipid body function in eicosanoid synthesis: an update. Prostaglandins, Leukotrienes and Essential Fatty Acids (PLEFA),85(5), 205-213.
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics, 23(19), 2633-2635.
- Bragg, L. M., Stone, G., Butler, M. K., Hugenholtz, P., & Tyson, G. W. (2013). Shining a light on dark sequencing: characterising errors in Ion Torrent PGM data. PLoS Comput Biol, 9(4), e1003031.
- Bredas, J., & Scuvie, L. (1960). Aperçu des influences climatiques sur les cycles de production du palmier à huile. Oléagineux, 15(4), 211-222.
- Breseghello, F., & Sorrells, M. E. (2006). Association analysis as a strategy for improvement of quantitative traits in plants. Crop Science, 46(3), 1323-1330.
- Breyne, P., & Zabeau, M. (2001). Genome-wide expression analysis of plant cell cycle modulated genes. Current opinion in plant biology, 4(2), 136-142.
- Broadwater, J. A., Whittle, E., & Shanklin, J. (2002). Desaturation and Hydroxylation residues 148 and 324 of arabidopsis fad2, in addition to substrate chain length, exert a major influence in partitioning of catalytic specificity. Journal of Biological Chemistry,277(18), 15613-15620.

- Broekmans, A. F. M. (1957). Growth, flowering and yield of the oil palm in Nigeria. JW Afric. Inst. for Oil Palm Res, 2(7), 187-220.
- Buchanan, B. B., Schürmann, P., Wolosiuk, R. A., & Jacquot, J. P. (2005). The ferredoxin/thioredoxin system: from discovery to molecular structures and beyond. Discoveries in Photosynthesis (pp. 859-866). Springer Netherlands.
- Bush, W. S., & Moore, J. H. (2012). Genome-wide association studies. PLoS Comput Biol, 8(12), e1002822.
- Byrne, M. E. (2009). A role for the ribosome in development. Trends in plant science, 14(9), 512-519.
- Byrne, P. F., & McMullen, M. D. (1996). Defining genes for agricultural traits: QTL analysis and the candidate gene approach. Probe, 7(1), 24-27.
- Cahoon, E. B., Marillia, E. F., Stecca, K. L., Hall, S. E., Taylor, D. C., & Kinney, A. J. (2000). Production of fatty acid components of meadowfoam oil in somatic soybean embryos. Plant Physiology, 124(1), 243-252.
- Cai, Y., Goodman, J. M., Pyc, M., Mullen, R. T., Dyer, J. M., & Chapman, K. D. (2015). Arabidopsis SEIPIN proteins modulate triacylglycerol accumulation and influence lipid droplet proliferation. The Plant Cell, 27(9), 2616-2636.
- Cao, Y., Zhang, Q., Chen, Y., Zhao, H., Lang, Y., Yu, C., & Yang, J. (2013). Identification of differential expression genes in leaves of rice (*Oryza sativa* L.) in response to heat stress by cDNA-AFLP analysis. BioMed research international, 2013.
- Capron, A., Gourgues, M., Neiva, L. S., Faure, J. E., Berger, F., Pagnussat, G., [...] & Liu, B. (2008). Maternal control of male-gamete delivery in Arabidopsis involves a putative GPI-anchored protein encoded by the LORELEI gene. The Plant Cell, 20(11), 3038-3049.
- Caro, E., & Gutierrez, C. (2007). A green GEM: intriguing analogies with animal geminin. Trends in cell biology, 17(12), 580-585.
- Carpita, N. C., & McCann, M. C. (2015). Characterizing visible and invisible cell wall mutant phenotypes. Journal of experimental botany, 66(14), 4145-4163.
- Cartwright, B. R., & Goodman, J. M. (2012). Seipin: from human disease to molecular mechanism. Journal of lipid research, 53(6), 1042-1055.
- Celler, K., Fujita, M., Kawamura, E., Ambrose, C., Herburger, K., Holzinger, A., & Wasteneys, G. O. (2016). Microtubules in Plant Cells: Strategies and Methods for Immunofluorescence, Transmission Electron Microscopy, and Live Cell Imaging. Cytoskeleton Methods and Protocols: Methods and Protocols, 155-184.
- Cevik, V., Ryder, C. D., Popovich, A., Manning, K., King, G. J., & Seymour, G. B. (2010). A FRUITFULL-like gene is associated with genetic variation for fruit flesh firmness in apple (*Malus domestica* Borkh.). Tree Genetics & Genomes, 6(2), 271-279.
- Chabouté, M. E., Chaubet, N., Gigot, C., & Philipps, G. (1993). Histones and histone genes in higher plants: structure and genomic organization. Biochimie, 75(7), 523-531.
- Chapman, K. D., & Ohlrogge, J. B. (2012). Compartmentation of triacylglycerol accumulation in plants. Journal of Biological Chemistry, 287(4), 2288-2294.
- Chapman, K. D., Dyer, J. M., & Mullen, R. T. (2012). Biogenesis and functions of lipid droplets in plants. Thematic Review Series: Lipid Droplet Synthesis and Metabolism: from Yeast to Man. Journal of lipid research, 53(2), 215-226.
- Chen, X., Irani, N. G., & Friml, J. (2011). Clathrin-mediated endocytosis: the gateway into plant cells. Current opinion in plant biology, 14(6), 674-682.

- Cheng, M. C., Hsieh, E. J., Chen, J. H., Chen, H. Y., & Lin, T. P. (2012). Arabidopsis RGLG2, functioning as a RING E3 ligase, interacts with AtERF53 and negatively regulates the plant drought stress response. Plant physiology, 158(1), 363-375.
- Chinnusamy, V., Gong, Z., & Zhu, J. K. (2008). Abscisic acid-mediated epigenetic processes in plant development and stress responses. Journal of integrative plant biology, 50(10), 1187-1195.
- Choi, K., Kim, J., Hwang, H. J., Kim, S., Park, C., Kim, S. Y., & Lee, I. (2011). The FRIGIDA complex activates transcription of FLC, a strong flowering repressor in Arabidopsis, by recruiting chromatin modification factors. The Plant Cell, 23(1), 289-303.
- Cipollone, R., Ascenzi, P., & Visca, P. (2007). Common themes and variations in the rhodanese superfamily. IUBMB life, 59(2), 51-59.
- Clarke, L. J., Czechowski, P., Soubrier, J., Stevens, M. I., & Cooper, A. (2014). Modular tagging of amplicons using a single PCR for high-throughput sequencing. Molecular ecology resources, 14(1), 117-121.
- Cobbett, C., & Goldsbrough, P. (2002). Phytochelatins and metallothioneins: roles in heavy metal detoxification and homeostasis. Annual review of plant biology, 53(1), 159-182.
- Cochard, B., Adon, B., Rekima, S., Billotte, N., de Chenon, R. D., Koutou, A., [...] & Noyer, J. L. (2009). Geographic and genetic structure of African oil palm diversity suggests new approaches to breeding. Tree genetics & genomes, 5(3), 493-504.
- Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica, 142(1-2), 169-196.
- Collard, B. C., & Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1491), 557-572.
- Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics, 21(18), 3674-3676.
- Cordin, O., Banroques, J., Tanner, N. K., & Linder, P. (2006). The DEAD-box protein family of RNA helicases. Gene, 367, 17-37
- Corley R.H.V. (1977) Oil palm yield components and yield cycles. In: International developments in oil palm (Ed. by D.A.Earp & W. Newall), pp. 116–129, Incorp. Soc. Planters, Kuala Lumpur [4.4; 4.4.8]
- Corley RHVand Tinker PB (2003) The Oil Palm. 562 pp. Oxford, UK: Blackwell Publishing.
- Corley, R. H. V., & Lee, C. H. (1992). The physiological basis for genetic improvement of oil palm in Malaysia. Euphytica, 60(3), 179-184.
- Corley, R. H. V., & Tinker, P. B. (2003). The classification and morphology of the oil palm. The Oil Palm, 27-51.
- Corley, R. H. V., & Tinker, P. B. (2003). World agriculture. The Oil Palm, 126
- Corley, R. H. V., & Tinker, P. B. (2007). The origin and development of the oil palm industry. The Oil Palm, 1-26.
- Corley, R. H. V., & Tinker, P. B (2016). Selection and Breeding. The Oil Palm, Fifth edition, Oxford: Blackwell Science Ltd Blackwell Publishing 138-207.
- Corley, R., & Tinker, P. (2003). Selection and breeding. The oil palm. 4th ed. Oxford: Blackwell Science Ltd Blackwell Publishing, 133-200.
- Correa, J., Mamani, M., Muñoz-Espinoza, C., Laborie, D., Muñoz, C., Pinto, M., & Hinrichsen, P. (2014). Heritability and identification of QTLs and underlying candidate

genes associated with the architecture of the grapevine cluster (*Vitis vinifera* L.). Theoretical and applied genetics, 127(5), 1143-1162.

- Cosgrove, D. J. (2005). Growth of the plant cell wall. Nature reviews molecular cell biology, 6(11), 850-861.
- Cosio, C., & Dunand, C. (2009). Specific functions of individual class III peroxidase genes. Journal of experimental botany, 60(2), 391-408.
- Costa, G. G., Cardoso, K. C., Del Bem, L. E., Lima, A. C., Cunha, M. A., Campos-Leite, L., [...] & Campos, F. A. (2010). Transcriptome analysis of the oil-rich seed of the bioenergy crop *Jatropha curcas* L. BMC genomics, 11(1), 1.
- Coursol S, Fan L-M, Le Stunff H, Spiegel S, Gilroy S, Assmann SM. (2003). Sphingolipid signalling in Arabidopsis guard cells involves heterotrimeric G proteins. Nature 423:651–654.
- Coursol S, Le Stunff H,Lynch DV, Gilroy S,Assmann SM, Spiegel S.2005. Arabidopsis sphingosine kinase and the effects of phytosphingosine-1-phosphate on stomatal aperture. Plant Physiology, 137: 724–737.
- Coustham, V., Vlad, D., Deremetz, A., Gy, I., Cubillos, F. A., Kerdaffrec, E., [...] & Bouché, N. (2014). SHOOT GROWTH1 maintains Arabidopsis epigenomes by regulating IBM1. PloS one, 9(1), e84687.
- Cronn, R., Knaus, B. J., Liston, A., Maughan, P. J., Parks, M., Syring, J. V., & Udall, J. (2012). Targeted enrichment strategies for next-generation plant biology. American Journal of Botany, 99(2), 291-311.
- Cui, X., Wang, Y., Meng, L., Fei, W., Deng, J., Xu, G., [...] & Zhao, L. (2012). Overexpression of a short human seipin/BSCL2 isoform in mouse adipose tissue results in mild lipodystrophy. American Journal of Physiology-Endocrinology and Metabolism, 302(6), E705-E713.
- Cui, X., Wang, Y., Tang, Y., Liu, Y., Zhao, L., Deng, J., [...] & Yang, H. (2011). Seipin ablation in mice results in severe generalized lipodystrophy. Human molecular genetics, ddr205.
- Dai, S., Johansson, K., Miginiac-Maslow, M., Schürmann, P., & Eklund, H. (2004). Structural basis of redox signaling in photosynthesis: structure and function of ferredoxin: thioredoxin reductase and target enzymes. Photosynthesis research, 79(3), 233-248.
- Dai, S., Schwendtmayer, C., Johansson, K., Ramaswamy, S., Schürmann, P., & Eklund, H. (2000). How does light regulate chloroplast enzymes? Structure–function studies of the ferredoxin/thioredoxin system. Quarterly Reviews of Biophysics, 33(01), 67-108.
- De Alencar Figueiredo, L. F., Sine, B., Chantereau, J., Mestres, C., Fliedel, G., Rami, J. F., [...] & Courtois, B. (2010). Variability of grain quality in sorghum: association with polymorphism in Sh2, Bt2, SssI, Ae1, Wx and O2. Theoretical and applied genetics, 121(6), 1171-1185.
- de Miguel, M., Cabezas, J. A., de María, N., Sánchez-Gómez, D., Guevara, M. Á., Vélez, M. D., [...] & Collada, C. (2014). Genetic control of functional traits related to photosynthesis and water use efficiency in *Pinus pinaster* Ait. drought response: integration of genome annotation, allele association and QTL detection for candidate gene identification. BMC genomics, 15(1), 1.
- de Vetten, N., Ter Horst, J., van Schaik, H. P., de Boer, A., Mol, J., & Koes, R. (1999). A cytochrome b5 is required for full activity of flavonoid 3', 5'-hydroxylase, a cytochrome P450 involved in the formation of blue flower colors. Proceedings of the National Academy of Sciences, 96(2), 778-783.

- Delannoy, E., Jalloul, A., Assigbetsé, K., Marmey, P., Geiger, J. P., Lherminier, J., [...] & Nicole, M. (2003). Activity of class III peroxidases in the defense of cotton to bacterial blight. Molecular plant-microbe interactions, 16(11), 1030-1038.
- Den Boer, B. G., & Murray, J. A. (2000). Control of plant growth and development through manipulation of cell-cycle genes. Current opinion in biotechnology, 11(2), 138-145.
- Diapari, M., Sindhu, A., Warkentin, T. D., Bett, K., & Tar'an, B. (2015). Population structure and marker-trait association studies of iron, zinc and selenium concentrations in seed of field pea (*Pisum sativum* L.). Molecular Breeding, 35(1), 1-14.
- Dietrich, C. R., Han, G., Chen, M., Berg, R. H., Dunn, T. M., & Cahoon, E. B. (2008). Loss-of-function mutations and inducible RNAi suppression of Arabidopsis LCB2 genes reveal the critical role of sphingolipids in gametophytic and sporophytic cell viability. The Plant Journal, 54(2), 284-298.
- Dilworth, D., Gudavicius, G., Leung, A., & Nelson, C. J. (2011). The roles of peptidyl-proline isomerases in gene regulation 1 1. This review is part of Special Issue entitled Asilomar Chromatin and has undergone the Journal's usual peer review process. Biochemistry and Cell Biology, 90(1), 55-69.
- Dinant, S., Clark, A. M., Zhu, Y., Vilaine, F., Palauqui, J. C., Kusiak, C., & Thompson, G. A. (2003). Diversity of the superfamily of phloem lectins (phloem protein 2) in angiosperms. Plant Physiology, 131(1), 114-128.
- Dinari, A., Niazi, A., Afsharifar, A. R., & Ramezani, A. (2013). Identification of upregulated genes under cold stress in cold-tolerant chickpea using the cDNA-AFLP approach. PLoS One, 8(1), e52757.
- Doebley, J., & Lukens, L. (1998). Transcriptional regulators and the evolution of plant form. The Plant Cell, 10(7), 1075-1082.
- Drakakaki, G., Zabotina, O., Delgado, I., Robert, S., Keegstra, K., & Raikhel, N. (2006). Arabidopsis reversibly glycosylated polypeptides 1 and 2 are essential for pollen development. Plant Physiology, 142(4), 1480-1492.
- Drincovich, M. F., Casati, P., & Andreo, C. S. (2001). NADP-malic enzyme from plants: a ubiquitous enzyme involved in different metabolic pathways. FEBS letters, 490(1), 1-6.
- Droux, M. (2004). Sulfur assimilation and the role of sulfur in plant metabolism: a survey. Photosynthesis Research, 79(3), 331-348.
- Duncan, O., van der Merwe, M. J., Daley, D. O., & Whelan, J. (2013). The outer mitochondrial membrane in higher plants. Trends in plant science, 18(4), 207-217.
- Durrett TP, Benning C, Ohlrogge J (2008) Plant triacylglycerols as feedstocks for the production of biofuels. Plant Plant J 54: 593-607
- Durstewitz, G., Polley, A., Plieske, J., Luerssen, H., Graner, E. M., Wieseke, R., & Ganal, M. W. (2010). SNP discovery by amplicon sequencing and multiplex SNP genotyping in the allopolyploid species *Brassica napus*. Genome, 53(11), 948-956.
- Dussert, S., Guerin, C., Andersson, M., Joët, T., Tranbarger, T. J., Pizot, M., [...] & Morcillo, F. (2013). Comparative transcriptome analysis of three oil palm fruit and seed tissues that differ in oil content and fatty acid composition. Plant physiology, 162(3), 1337-1358.
- Earl, D. A. (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation genetics resources, 4(2), 359-361.

- Edwards, D., J.W. Forster, D. Chagne, and J. Batley. 2007. What is SNPs? In: N.C. Or- aguzie, E.H.A. Rikkerink, S.E. Gardiner and H.N. de Silva (eds.), Association Mapping in Plants. Springer, Berlin, pp 41-52
- Egan, A. N., Schlueter, J., & Spooner, D. M. (2012). Applications of next-generation sequencing in plant biology.
- Ehrenreich, I. M., Hanzawa, Y., Chou, L., Roe, J. L., Kover, P. X., & Purugganan, M. D. (2009). Candidate gene association mapping of Arabidopsis flowering time. Genetics, 183(1), 325-335.
- Endler, A., & Persson, S. (2011). Cellulose synthases and synthesis in Arabidopsis. Molecular Plant, 4(2), 199-211.
- Englert, M., & Beier, H. (2005). Plant tRNA ligases are multifunctional enzymes that have diverged in sequence and substrate specificity from RNA ligases of other phylogenetic origins. Nucleic acids research, 33(1), 388-399.
- Englert, M., Latz, A., Becker, D., Gimple, O., Beier, H., & Akama, K. (2007). Plant pre-tRNA splicing enzymes are targeted to multiple cellular compartments. Biochimie, 89(11), 1351-1365.
- Escobar, R., & Alvarado, A. (2004). Mejoramiento genético de palma aceitera y producción de alto rendimiento. XXVII Curso internacional de palma aceitera. ASD de Costa Rica y ACUPALMA, 1-25.
- Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular ecology, 14(8), 2611-2620.
- Falasca, M., & Maffucci, T. (2009). Rethinking phosphatidylinositol 3-monophosphate. Molecular Cell Research, 1793(12), 1795-1803.
- Fan, R. C., Peng, C. C., Xu, Y. H., Wang, X. F., Li, Y., Shang, Y., [...] & Zhang, D. P. (2009). Apple sucrose transporter SUT1 and sorbitol transporter SOT6 interact with cytochrome b5 to regulate their affinity for substrate sugars. Plant physiology, 150(4), 1880-1901.
- Steinfeld, H., de Haan, C., & Blackburn, H. Livestock - environment interactions. Issues and options, FAO. 1996.
- Fedoroff, N. V. (1999). The Suppressor-mutator element and the evolutionary riddle of transposons. Genes to Cells, 4(1), 11-19.
- Feiler, H. S., Desprez, T., Santoni, V., Kronenberger, J., Caboche, M., & Traas, J. (1995). The higher plant *Arabidopsis thaliana* encodes a functional CDC48 homologue which is highly expressed in dividing and expanding cells. The EMBO journal, 14(22), 5626.
- Fields, C. (1994). Analysis of gene expression by tissue and developmental stage. Current opinion in Biotechnology, 5(6), 595-598.
- Filippi, C. V., Aguirre, N., Rivas, J. G., Zubrzycki, J., Puebla, A., Cordes, D., [...] & Hopp, H. E. (2015). Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. BMC plant biology, 15(1), 1.
- Fischer, J., Becker, C., Hillmer, S., Horstmann, C., Neubohn, B., Schlereth, A., [...] & Müntz, K. (2000). The families of papain-and legumain-like cysteine proteinases from embryonic axes and cotyledons of *Vicia* seeds: developmental patterns, intracellular localization and functions in globulin proteolysis. Plant molecular biology, 43(1), 83-101.
- Flint-Garcia, S. A., Thornsberry, J. M., & Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. Annual review of plant biology, 54(1), 357-374.
- Frandsen, G. I., Mundy, J., & Tzen, J. T. (2001). Oil bodies and their associated proteins, oleosin and caleosin. Physiologia plantarum, 112(3), 301-307.

- Fritsche, S., Wang, X., Li, J., Stich, B., Kopisch-Obuch, F. J., Endrigkeit, J., [...] & Jung, C. (2012). A candidate gene-based association study of tocopherol content and composition in rapeseed (*Brassica napus*). The Brassica Genome, 84.
- Furumizu, C., Tsukaya, H., & Komeda, Y. (2010). Characterization of EMU, the Arabidopsis homolog of the yeast THO complex member HPR1. Rna, 16(9), 1809-1817.
- Fusari, C. M. (2010). Mapeo por asociación en girasol: diversidad nucleotídica, desequilibrio de ligamiento e identificación de genes involucrados en la resistencia a la podredumbre húmeda del capítulo (Doctoral dissertation, Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires).
- Galindo-González, L., Pinzón-Latorre, D., Bergen, E. A., Jensen, D. C., & Deyholos, M. K. (2015). Ion Torrent sequencing as a tool for mutation discovery in the flax (*Linum usitatissimum* L.) genome. Plant methods, 11(1), 1.
- Ganal, M. W., Altmann, T., & Röder, M. S. (2009). SNP identification in crop plants. Current opinion in plant biology, 12(2), 211-217.
- García-Mas, J., Graziano E., Aranzana, M. J., Monforte, A., Oliver, M., Ballester, J., Viruel, M. A., Arús, P. (2000). Marcadores de ADN: conceptos, tipos, protocolos. Nuez, F., & Carrillo, J. M. (2000). Los marcadores genéticos en la mejora vegetal.(pp: 90-151) Universidad Politécnica de Valencia.
- Gascon, J. P., & De Berchoux, C. (1964). Caractéristiques de la production de quelques origines d'*Elaeis guineensis* (Jacq.) et de leurs croisements: application à la sélection du palmier à huile. Oléagineux, 19(2), 75-84.
- Gaxiola, R. A., Palmgren, M. G., & Schumacher, K. (2007). Plant proton pumps. Febs Letters, 581(12), 2204-2214.
- Gbadegesin, M. A., Wills, M. A., & Beeching, J. R. (2008). Diversity of LTR-retrotransposons and Enhancer/SuppressorMutator-like transposons in cassava (*Manihot esculenta* Crantz). Molecular Genetics and Genomics, 280(4), 305.
- Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers I. Methods. Theoretical and Applied Genetics, 46(7), 319-330.
- Gilbert, L., Alhagdow, M., Nunes-Nesi, A., Quemener, B., Guillon, F., Bouchet, B., [...] & Petit, J. (2009). GDP-d-mannose 3, 5-epimerase (GME) plays a key role at the intersection of ascorbate and non-cellulosic cell-wall biosynthesis in tomato. The Plant Journal, 60(3), 499-508.
- Gimeno-Gilles, C., Gervais, M. L., Planchet, E., Satour, P., Limami, A. M., & Lelievre, E. (2011). A stress-associated protein containing A20/AN1 zing-finger domains expressed in *Medicago truncatula* seeds. Plant Physiology and Biochemistry, 49(3), 303-310.
- Giri, J., Dansana, P. K., Kothari, K. S., Sharma, G., Vij, S., & Tyagi, A. K. (2013). SAPs as novel regulators of abiotic stress response in plants. Bioessays, 35(7), 639-648.
- Goh, K. J., Teo, C. B., Chew, P. S., & Chiu, S. B. (1999). Fertiliser management in oil palm: Agronomic principles and field practices. Fertiliser management for oil palm plantations, 20-21.
- Gomez, G., Torres, H., & Pallas, V. (2005). Identification of translocatable RNA-binding phloem proteins from melon, potential components of the long-distance RNA transport system. The Plant Journal, 41(1), 107-116.
- Gong, Z., Dong, C. H., Lee, H., Zhu, J., Xiong, L., Gong, D., [...] & Zhu, J. K. (2005). A DEAD box RNA helicase is essential for mRNA export and important for development and stress responses in Arabidopsis. The Plant Cell, 17(1), 256-267.

- Goodman, J. M. (2008). The gregarious lipid droplet. Journal of Biological Chemistry, 283(42), 28005-28009.
- Gordon, A., & Hannon, G. J. (2010). Fastx-toolkit.FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit.
- Grada, A., & Weinbrecht, K. (2013). Next-generation sequencing: methodology and application. Journal of Investigative Dermatology, 133(8), 1-4.
- Grudkowska, M., & Zagdanska, B. (2004). Multifunctional role of plant cysteine proteinases. Acta biochimica polonica-english edition-, 609-624.
- Grzebelus, D., Jagosz, B., & Simon, P. W. (2007). The DcMaster transposon display maps polymorphic insertion sites in the carrot (*Daucus carota* L.) genome. Gene, 390(1), 67-74.
- Guo, P., Baum, M., Grando, S., Ceccarelli, S., Bai, G., Li, R., [...] & Valkoun, J. (2009). Differentially expressed genes between drought-tolerant and drought-sensitive barley genotypes in response to drought stress during the reproductive stage. Journal of experimental botany, 60(12), 3531-3544.
- Gupta, N., Naik, P. K., & Chauhan, R. S. (2012). Differential transcript profiling through cDNA-AFLP showed complexity of rutin biosynthesis and accumulation in seeds of a nutraceutical food crop (*Fagopyrum* spp.). BMC genomics, 13(1), 231.
- Gupta, P. K., & Rustgi, S. (2004). Molecular markers from the transcribed/expressed region of the genome in higher plants. Functional & integrative genomics, 4(3), 139-162.
- Gupta, P. K., Roy, J. K., & Prasad, M. (2001). Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci, 80(4), 524-535.
- Gupta, P. K., Rustgi, S., & Kulwal, P. L. (2005). Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant molecular biology, 57(4), 461-485.
- Hajduch, M., Casteel, J. E., Hurrelmeyer, K. E., Song, Z., Agrawal, G. K., & Thelen, J. J. (2006). Proteomic analysis of seed filling in *Brassica napus*. Developmental characterization of metabolic isozymes using high-resolution two-dimensional gel electrophoresis. Plant Physiology, 141(1), 32-46.
- Hall, T. A. (1999, January). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic acids symposium series (Vol. 41, pp. 95-98).
- Hamada, T. (2007). Microtubule-associated proteins in higher plants. Journal of plant research, 120(1), 79-98.
- Hamada, T. (2014). Microtubule organization and microtubule-associated proteins in plant cells. International Review of Cell and Molecular Biology, 312, 1-52.
- Han, J., Lühs, W., Sonntag, K., Zähringer, U., Borchardt, D. S., Wolter, F. P., [...] & Frentzen, M. (2001). Functional characterization of β-ketoacyl-CoA synthase genes from *Brassica napus* L. Plant molecular biology, 46(2), 229-239.
- Hanes, S. D. (2015). Prolyl isomerases in gene transcription. Biochimica et Biophysica Acta (BBA)-General Subjects, 1850(10), 2017-2034.
- Hanley, S., Barker, J., Van Ooijen, J., Aldam, C., Harris, S., Åhman, I., [...] & Karp, A. (2002). A genetic linkage map of willow (*Salix viminalis*) based on AFLP and microsatellite markers. Theoretical and Applied Genetics, 105(6-7), 1087-1096.
- Hanniff Harun, M.(2000). Yield and Yield components and their physioloy. In Basiron Y, Jalani BS y Chan KW, Advances in oil palm research (pp146-170). Malasia: Malasian Palm Oil Board

- Harjes, C. E., Rocheford, T. R., Bai, L., Brutnell, T. P., Kandianis, C. B., Sowinski, S. G., [...] & Yan, J. (2008). Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. Science, 319(5861), 330-333.
- Hartley, C. W. S. The oil palm. Essex: Longman, 1988. 761 p.
- Hayati, A., Wickneswari, R., Maizura, I., & Rajanaidu, N. (2004). Genetic diversity of oil palm (*Elaeis guineensis* Jacq.) germplasm collections from Africa: implications for improvement and conservation of genetic resources. Theoretical and Applied Genetics, 108(7), 1274-1284.
- Hayden, M. J., Tabone, T. L., Nguyen, T. M., Coventry, S., Keiper, F. J., Fox, R. L., [...] & Eglinton, J. K. (2010). An informative set of SNP markers for molecular characterisation of Australian barley germplasm. Crop and Pasture Science, 61(1), 70-83.
- Heilmann, I. (2016). Plant phosphoinositide signaling-dynamics on demand. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids.
- Hemptinne J. & Ferwerda J.D. (1961) Influence des précipitations sur les productions du palmier a huile (*Elaeis guineensis* Jacq.). Oléagineux,16, 431–437
- Henderson, I. R., & Jacobsen, S. E. (2007). Epigenetic inheritance in plants. Nature, 447(7143), 418-424.
- Henegariu, O., Heerema, N. A., Dlouhy, S. R., Vance, G. H., & Vogt, P. H. (1997). Multiplex PCR: critical parameters and step-by-step protocol. Biotechniques, 23(3), 504-511.
- Henry, R. J., Edwards, M., Waters, D. L., Bundock, P., Sexton, T. R., Masouleh, A. K., [...] & Pattemore, J. (2012). Application of large-scale sequencing to marker discovery in plants. Journal of biosciences, 37(5), 829-841.
- Henson, I. E., & Dolmat, M. T. (2004). Seasonal variation in yield and developmental processes in an oil palm density trial on a peat soil: 2. Bunch weight components. Journal of Oil Palm Research, 16(2), 106-120.
- Henson, I. E., & Harun, M. H. (2005). The influence of climatic conditions on gas and energy exchanges above a young oil palm stand in north Kedah, Malaysia. Journal of oil palm Research, 17(C), 73.
- Herrero J. (2013). Construction of a functional genetic map and detection of candidate genes for useful traits in Oil Palm, *Elaeis guineensis* Jacq.(Doctoral dissertation). Basque Country University, Leioa, Spain
- Hershko, A. (1998). The ubiquitin system (pp. 1-17). Springer US.
- Hill, H., Lee, L. S., & Henry, R. J. (2012). Variation in sorghum starch synthesis genes associated with differences in starch phenotype. Food chemistry, 131(1), 175-183.
- Hill, K., Mathews, D. E., Kim, H. J., Street, I. H., Wildes, S. L., Chiang, Y. H., [...] & Schaller, G. E. (2013). Functional characterization of type-B response regulators in the Arabidopsis cytokinin response. Plant physiology, 162(1), 212-224.
- Ho, C. L., Kwan, Y. Y., Choi, M. C., Tee, S. S., Ng, W. H., Lim, K. A., [...] & Tan, S. H. (2007). Analysis and functional annotation of expressed sequence tags (ESTs) from multiple tissues of oil palm (*Elaeis guineensis* Jacq.). BMC genomics, 8(1), 381.
- Ho, H., Low, J. Z., Gudimella, R., Tammi, M. T., & Harikrishna, J. A. (2015). Expression patterns of inflorescence-and sex-specific transcripts in male and female inflorescences of African oil palm (*Elaeis guineensis*). Annals of Applied Biology.
- Hoefgen, R., & Hesse, H. (2008). Sulfur and cysteine metabolism. Sulfur: A Missing Link between Soils, Crops, and Nutrition, 83-104.
- Homblé, F., Krammer, E. M., & Prévost, M. (2012). Plant VDAC: facts and speculations. Biochimica et Biophysica Acta (BBA)-Biomembranes, 1818(6), 1486-1501.

- Horiguchi, G., Van Lijsebettens, M., Candela, H., Micol, J. L., & Tsukaya, H. (2012). Ribosomes and translation in plant developmental control. Plant science, 191, 24-34.
- Horn, P. J., James, C. N., Gidda, S. K., Kilaru, A., Dyer, J. M., Mullen, R. T., [...] & Chapman, K. D. (2013). Identification of a new class of lipid droplet-associated proteins in plants. Plant physiology, 162(4), 1926-1936.
- Hsieh, T. F., Ibarra, C. A., Silva, P., Zemach, A., Eshed-Williams, L., Fischer, R. L., & Zilberman, D. (2009). Genome-wide demethylation of Arabidopsis endosperm. Science, 324(5933), 1451-1454.
- Hsieh, T. F., Shin, J., Uzawa, R., Silva, P., Cohen, S., Bauer, M. J., [...] & Fischer, R. L. (2011). Regulation of imprinted gene expression in Arabidopsis endosperm. Proceedings of the National Academy of Sciences, 108(5), 1755-1762.
- Hu, G., Chen, G., Hu, Z., Gu, F., & Li, Y. (2010). Function of plant homeodomain-finger proteins in vernalization pathway in Arabidopsis and other cruciferous plants. Chinese journal of biotechnology, 26(1), 1-8.
- Huang, A. H. (1992). Oil bodies and oleosins in seeds. Annual review of plant biology, 43(1), 177-200.
- Huang, N. L., Huang, M. D., Chen, T. L. L., & Huang, A. H. (2013). Oleosin of subcellular lipid droplets evolved in green algae. Plant physiology, 161(4), 1862-1874.
- Ingvarsson, P. K., & Street, N. R. (2011). Association genetics of complex traits in plants. New Phytologist, 189(4), 909-922.
- Inze D, De Veylder L. 2006. Cell cycle regulation in plant development. Annual Review of Genetics 40, 77–105.
- Jacquin, N. J. (1763). *Selectarum stirpium Americanarum historia: in qua ad Linnaeanum systema determinatae descriptaeque sistuntur plantae illae, quas in insulis Martinica, Jamaica, Domingo, aliisque, et in vicinae continentis parte observavit rariores; adjectis iconibus in solo natali delineatis* (Vol. 1).
- Jalani, B. (1994). Influence of planting material on oil extraction ratio (OER). In National Seminar on Palm oil extraction rate: problemas and issues. December 21-22 Kuala Lumpur (No. L-0302). PORIM.
- Jalani, B. S., Rajanaidu, N., & Darus, A. (1993). Perspectivas para el siglo XXI: la palma y la calidad de aceite ideales para el futuro. Revista Palmas, 14(especial), 12-25.
- Janssen TBK (2004) The age of major monocot groups inferred from 800+ rbcl sequences. Botanical Journal of the Linnean Society, 146:385–398
- Janssen, B. J., Thodey, K., Schaffer, R. J., Alba, R., Balakrishnan, L., Bishop, R., [...] & McArtney, S. (2008). Global gene expression analysis of apple fruit development from the floral bud to ripe fruit. BMC Plant Biology, 8(1), 1.
- Jayaraman, A., Puranik, S., Rai, N. K., Vidapu, S., Sahu, P. P., Lata, C., & Prasad, M. (2008). cDNA-AFLP analysis reveals differential gene expression in response to salt stress in foxtail millet (*Setaria italica* L.). Molecular biotechnology, 40(3), 241-251.
- Jeennor, S., & Volkaert, H. (2014). Mapping of quantitative trait loci (QTLs) for oil yield using SSRs and gene-based markers in African oil palm (*Elaeis guineensis* Jacq.). Tree genetics & genomes, 10(1), 1-14.
- Jia, G., Liu, X., Owen, H. A., & Zhao, D. (2008). Signaling of cell fate determination by the TPD1 small protein and EMS1 receptor kinase. Proceedings of the National Academy of Sciences, 105(6), 2220-2225.
- Jin, J., Lee, M., Bai, B., Sun, Y., Qu, J., Alfiko, Y., [...] & Ye, J. (2016). Draft genome sequence of an elite Dura palm and whole-genome patterns of DNA variation in oil palm. Dna Research, dsw036.

- Johanson, U., West, J., Lister, C., Michaels, S., Amasino, R., & Dean, C. (2000). Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. Science, 290(5490), 344-347.
- Jones, E., Chu, W. C., Ayele, M., Ho, J., Bruggeman, E., Yourstone, K., [...] & Warren, J. (2009). Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm. Molecular Breeding, 24(2), 165-176.
- Jones, N., Ougham, H., & Thomas, H. (1997). Markers and mapping: we are all geneticists now. New Phytologist, 137(1), 165-177.
- Jones, N., Ougham, H., Thomas, H., & Pašakinskienė, I. (2009). Markers and mapping revisited: finding your gene. New Phytologist, 183(4), 935-966.
- Jouannic, S., Argout, X., Lechauve, F., Fizames, C., Borgel, A., Morcillo, F., [...] & Tregear, J. (2005). Analysis of expressed sequence tags from oil palm (*Elaeis guineensis*). FEBS letters, 579(12), 2709-2714.
- Kader, J. C. (1996). Lipid-transfer proteins in plants. Annual review of plant biology, 47(1), 627-654.
- Kallarackal, J., Jeyakumar, P., & George, S. J. (2004). Water use of irrigated oil palm at three different arid locations in Peninsular India. Journal of Oil Palm Research, 16, 45-53.
- Kang, F., & Rawsthorne, S. (1994). Starch and fatty acid synthesis in plastids from developing embryos of oilseed rape (*Brassica napus* L.). The Plant Journal, 6(6), 795-805.
- Kasha, K. J. (1999). Biotechnology and world food supply. Genome, 42(4), 642-645.
- Kearsey, M. J., & Farquhar, A. G. L. (1998). QTL analysis in plants; where are we now? Heredity, 80(2), 137-142.
- Keegstra, K., & Raikhel, N. (2001). Plant glycosyltransferases. Current opinion in plant biology, 4(3), 219-224.
- Kehr, J. (2006). Phloem sap proteins: their identities and potential roles in the interaction between plants and phloem-feeding insects. Journal of Experimental Botany, 57(4), 767-774.
- Khan, M. A., & Korban, S. S. (2012). Association mapping in forest trees and fruit crops. Journal of experimental botany, ers105.
- Kilaru, A., Cao, X., Dabbs, P. B., Sung, H. J., Rahman, M. M., Thrower, N., [...] & Mockaitis, K. (2015). Oil biosynthesis in a basal angiosperm: transcriptome analysis of *Persea Americana* mesocarp. BMC plant biology, 15(1), 203.
- Kilian, B., & Graner, A. (2012). NGS technologies for analyzing germplasm diversity in genebanks. Briefings in functional genomics, elr046.
- Kim, J. H., & Kim, W. T. (2013). The Arabidopsis RING E3 ubiquitin ligase AtAIRP3/LOG2 participates in positive regulation of high-salt and drought stress responses. Plant physiology, 162(3), 1733-1749.
- Kim, J. Y., Park, S. C., Hwang, I., Cheong, H., Nah, J. W., Hahm, K. S., & Park, Y. (2009). Protease inhibitors from plants with antimicrobial activity. International journal of molecular sciences, 10(6), 2860-2872.
- Kim, K., Ryu, H., Cho, Y. H., Scacchi, E., Sabatini, S., & Hwang, I. (2012). Cytokinin-facilitated proteolysis of ARABIDOPSIS RESPONSE REGULATOR 2 attenuates signaling output in two-component circuitry. The Plant Journal, 69(6), 934-945.
- Kindl, H. (1987). β-Oxidation of fatty acids by specific organelles. The Biochemistry of plants: a comprehensive treatise (USA).

- Kirch, H. H., Bartels, D., Wei, Y., Schnable, P. S., & Wood, A. J. (2004). The ALDH gene superfamily of Arabidopsis. Trends in plant science, 9(8), 371-377.
- Kirkness EF. Targeted sequencing with microfluidics. Nature Biotechnology 2009;27:998–9.
- Knowler, W. C., Williams, R. C., Pettitt, D. J., & Steinberg, A. G. (1988). Gm3; 5, 13, 14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. American journal of human genetics, 43(4), 520.
- Kochert, G., 1994. RFLP technology, p. 8–38. In: R.L. Phillips & I.K. Vasil (Eds.), DNA-based markers in plants. Kluwer Academic Publishers, Dordrecht.
- Konishi, T., Aohara, T., Igasaki, T., Hayashi, N., Miyazaki, Y., Takahashi, A., [...] & Ishii, T. (2011). Down-regulation of UDP-arabinopyranose mutase reduces the proportion of arabinofuranose present in rice cell walls. Phytochemistry, 72(16), 1962-1968.
- Korpelainen, H., & Kostamo, K. (2010). An improved and cost-effective cDNA-AFLP method to investigate transcription-derived products when high throughput sequencing is not available. Journal of biotechnology, 145(1), 43-46.
- Koski, L. B., & Golding, G. B. (2001). The closest BLAST hit is often not the nearest neighbor. Journal of Molecular Evolution, 52(6), 540-542.
- Kou, Y., Qiu, D., Wang, L., Li, X., & Wang, S. (2009). Molecular analyses of the rice tubby-like protein gene family and their response to bacterial infection. Plant cell reports, 28(1), 113-121.
- Kraakman, A. T., Martinez, F., Mussiraliev, B., Van Eeuwijk, F. A., & Niks, R. E. (2006). Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Molecular Breeding, 17(1), 41-58.
- Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome biology, 9(2), S8.
- Kruglyak, L. (1997). The use of a genetic map of biallelic markers in linkage studies. Nature genetics, 17(1), 21-24.
- Krupková, E., Immerzeel, P., Pauly, M., & Schmülling, T. (2007). The TUMOROUS SHOOT DEVELOPMENT2 gene of Arabidopsis encoding a putative methyltransferase is required for cell adhesion and co-ordinated plant development. The Plant Journal, 50(4), 735-750.
- Kubis, S. E., Castilho, A. M., Vershinin, A. V., & Heslop-Harrison, J. S. P. (2003). Retroelements, transposons and methylation status in the genome of oil palm (*Elaeis guineensis*) and the relationship to somaclonal variation. Plant molecular biology, 52(1), 69-79.
- Kujur, A., Bajaj, D., Upadhyaya, H. D., Das, S., Ranjan, R., Shree, T., [...] & Gowda, C. L. L. (2015). A genome-wide SNP scan accelerates trait-regulatory genomic loci identification in chickpea. Scientific reports, 5, 11166.
- Kumar, B., Abdel-Ghani, A. H., Pace, J., Reyes-Matamoros, J., Hochholdinger, F., & Lübberstedt, T. (2014). Association analysis of single nucleotide polymorphisms in candidate genes with root traits in maize (*Zea mays* L.) seedlings. Plant Science, 224, 9-19.
- Kumar, R., Tran, L. S. P., Neelakandan, A. K., & Nguyen, H. T. (2012). Higher plant cytochrome b 5 polypeptides modulate fatty acid desaturation. PloS one, 7(2), e31370.
- Kumar, R., Wallis, J. G., Skidmore, C., & Browse, J. (2006). A mutation in Arabidopsis cytochrome b5 reductase identified by high-throughput screening differentially affects hydroxylation and desaturation. The Plant Journal, 48(6), 920-932.

- Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP discovery through next-generation sequencing and its applications. International journal of plant genomics. ID 831460.
- Kurepa, J., & Smalle, J. A. (2008). Structure, function and regulation of plant proteasomes. Biochimie, 90(2), 324-335.
- Kurepa, J., Wang, S., Li, Y., & Smalle, J. (2009). Proteasome regulation, plant growth and stress tolerance. Plant signaling & behavior, 4(10), 924-927.
- Kusano, T., Tateda, C., Berberich, T., & Takahashi, Y. (2009). Voltage-dependent anion channels: their roles in plant defense and cell death. Plant cell reports, 28(9), 1301-1308.
- Kushairi, A., & Rajanaidu, N. (2000). Breeding populations, seed production and nursery management.
- Lai, C. P., Lee, C. L., Chen, P. H., Wu, S. H., Yang, C. C., & Shaw, J. F. (2004). Molecular analyses of the Arabidopsis TUBBY-like protein gene family. Plant Physiology, 134(4), 1586-1597.
- Lassner, M. W., Lardizabal, K., & Metz, J. G. (1996). A jojoba beta-Ketoacyl-CoA synthase cDNA complements the canola fatty acid elongation mutation in transgenic plants. The Plant Cell, 8(2), 281-292.
- Latiff, A. (2000) The Biology of the Genus Elaeis. pp. 19–38 In: Advances in Oil Palm Research, Volume 1, (Y. Basiron, B.S. Jalani, and K.W. Chan Eds.), Malaysian Palm Oil Board (MPOB)
- Lawit, S. J., Wych, H. M., Xu, D., Kundu, S., & Tomes, D. T. (2010). Maize DELLA proteins dwarf plant8 and dwarf plant9 as modulators of plant development. Plant and cell physiology, 51(11), 1854-1868.
- Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics, 255-268.
- Lee, M., Lenman, M., Banaś, A., Bafor, M., Singh, S., Schweizer, M., [...] & Sjödahl, S. (1998). Identification of non-heme diiron proteins that catalyze triple bond and epoxy group formation. Science, 280(5365), 915-918.
- Lee, M., Xia, J. H., Zou, Z., Ye, J., Alfiko, Y., Jin, J., [...] & Wong, L. (2015). A consensus linkage map of oil palm and a major QTL for stem height. Scientific reports, 5, 8232.
- Lersten NR, Czlapinski AR, Curtis JD, Freckmann R, Horner HT (2006)Oil bodies in leaf mesophyll cells of angiosperms: overview and a selected survey. Am J Bot 93: 1731-1739
- Liang, Y., Yuan, Y., Liu, T., Mao, W., Zheng, Y., & Li, D. (2014). Identification and computational annotation of genes differentially expressed in pulp development of *Cocus nucifera* L. by suppression subtractive hybridization. BMC plant biology, 14(1), 1.
- Li-Beisson, Y., Shorrosh, B., Beisson, F., Andersson, M. X., Arondel, V., Bates, P. D., [...] & Franke, R. B. (2013). Acyl-lipid metabolism. The Arabidopsis Book, e0161.
- Licausi, F., Ohme-Takagi, M., & Perata, P. (2013). APETALA2/Ethylene Responsive Factor (AP2/ERF) transcription factors: mediators of stress responses and developmental programs. New Phytologist, 199(3), 639-649.
- Lin, H. C., Morcillo, F., Dussert, S., Tranchant-Dubreuil, C., Tregear, J. W., & Tranbarger, T. J. (2009). Transcriptome analysis during somatic embryogenesis of the tropical monocot *Elaeis guineensis*: evidence for conserved gene functions in early development. Plant Molecular Biology, 70(1-2), 173-192.

- Lin, L. J., Liao, P. C., Yang, H. H., & Tzen, J. T. (2005). Determination and analyses of the N-termini of oil-body proteins, steroleosin, caleosin and oleosin. Plant Physiology and Biochemistry, 43(8), 770-776.
- Lister, R., Carrie, C., Duncan, O., Ho, L. H., Howell, K. A., Murcha, M. W., & Whelan, J. (2007). Functional definition of outer membrane proteins involved in preprotein import into mitochondria. The Plant Cell, 19(11), 3739-3759.
- Liszkay, A., van der Zalm, E., & Schopfer, P. (2004). Production of reactive oxygen intermediates (O2−, H2O2, and OH) by maize roots and their role in wall loosening and elongation growth. Plant Physiology, 136(2), 3114-3123.
- Liu, M., Pirrello, J., Chervin, C., Roustan, J. P., & Bouzayen, M. (2015). Ethylene control of fruit ripening: revisiting the complex network of transcriptional regulation. Plant physiology, 169(4), 2380-2390.
- Liu, P., Wang, C. M., Li, L., Sun, F., & Yue, G. H. (2011). Mapping QTLs for oil traits and eQTLs for oleosin genes in jatropha. BMC plant biology, 11(1), 1.
- Liu, Y., Xu, Y., Xiao, J., Ma, Q., Li, D., Xue, Z., & Chong, K. (2011). OsDOG, a gibberellin-induced A20/AN1 zinc-finger protein, negatively regulates gibberellin-mediated cell elongation in rice. Journal of plant physiology, 168(10), 1098-1105.
- Ljung, K. (2013). Auxin metabolism and homeostasis during plant development. Development, 140(5), 943-950.
- Loei, H., Lim, J., Tan, M., Lim, T. K., Lin, Q. S., Chew, F. T., [...] & Chung, M. C. (2013). Proteomic analysis of the oil palm fruit mesocarp reveals elevated oxidative phosphorylation activity is critical for increased storage oil production. Journal of proteome research, 12(11), 5096-5109.
- Loidl, P. (2004). A plant dialect of the histone language. Trends in plant science, 9(2), 84-90.
- Low, E. T. L., Alias, H., Boon, S. H., Shariff, E. M., Tan, C. Y. A., Ooi, L. C., [...] & Singh, R. (2008). Oil palm (*Elaeis guineensis* Jacq.) tissue culture ESTs: identifying genes associated with callogenesis and embryogenesis. BMC Plant Biology, 8(1), 62.
- Low, E. T. L., Rosli, R., Jayanthi, N., Azizi, N., Chan, K. L., Maqbool, N. J., [...] & Ong-Abdullah, M. (2014). Analyses of hypomethylated oil palm gene space. PloS one, 9(1), e86728.
- Luo, M., Taylor, J. M., Spriggs, A., Zhang, H., Wu, X., Russell, S., [...] & Koltunow, A. (2011). A genome-wide survey of imprinted genes in rice seeds reveals imprinting primarily occurs in the endosperm. PLoS Genetics, 7(6), e1002125.
- Madon, M., Clyde, M. M., & Cheah, S. C. (1995). Cytological analysis of *Elaeis guineensis* (tenera) chromosomes. Elaeis, 7(2), 122-131.
- Maeshima, M., Nakanishi, Y., Matsuura-Endo, C., & Tanaka, Y. (1996). Proton pumps of the vacuolar membrane in growing plant cells. Journal of Plant Research, 109(1), 119-125.
- Magré, J., Delépine, M., Khallouf, E., Gedde-Dahl, T., Van Maldergem, L., Sobel, E., [...] & Capeau, J. (2001). Identification of the gene altered in Berardinelli–Seip congenital lipodystrophy on chromosome 11q13. Nature genetics, 28(4), 365-370.
- Majeran, W., Friso, G., Asakura, Y., Qu, X., Huang, M., Ponnala, L., [...] & Van Wijk, K. J. (2012). Nucleoid-enriched proteomes in developing plastids and chloroplasts from maize leaves: a new conceptual framework for nucleoid functions. Plant Physiology, 158(1), 156-189.
- Malladi, A., & Johnson, L. K. (2011). Expression profiling of cell cycle genes reveals key facilitators of cell production during carpel development, fruit set, and fruit

growth in apple (*Malus*× *domestica* Borkh.). Journal of experimental botany, 62(1), 205-219.

- Malosetti, M., van der Linden, C. G., Vosman, B., & van Eeuwijk, F. A. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics, 175(2), 879-889.
- Mammadov, J., Aggarwal, R., Buyyarapu, R., & Kumpatla, S. (2012). SNP markers and their impact on plant breeding. International journal of plant genomics, 2012.
- Mandelkow, E., & Mandelkow, E. M. (1995). Microtubules and microtubule-associated proteins. Current opinion in cell biology, 7(1), 72-81.
- Mariette, S., Wong Jun Tai, F., Roch, G., Barre, A., Chague, A., Decroocq, S., [...] & Nikolski, M. (2016). Genome-wide association links candidate genes to resistance to *Plum Pox* Virus in apricot (*Prunus armeniaca*). New Phytologist, 209(2), 773-784.
- Markham, J. E., Lynch, D. V., Napier, J. A., Dunn, T. M., & Cahoon, E. B. (2013). Plant sphingolipids: function follows form. Current opinion in plant biology, 16(3), 350-357.
- Markoulatos, P., Siafakas, N., & Moncany, M. (2002). Multiplex polymerase chain reaction: a practical approach. Journal of clinical laboratory analysis, 16(1), 47-51.
- Martín, A. (2002). Los marcadores genéticos en la Mejora Vegetal. En Genómica y mejora vegetal (pp. 39-64). Consejería de Agricultura y Pesca.
- Mason, M. G., Mathews, D. E., Argyros, D. A., Maxwell, B. B., Kieber, J. J., Alonso, J. M., [...] & Schaller, G. E. (2005). Multiple type-B response regulators mediate cytokinin signal transduction in Arabidopsis. The Plant Cell, 17(11), 3007-3018.
- Mayes, S., James, C. M., Horner, S. F., Jack, P. L., & Corley, R. H. V. (1996). The application of restriction fragment length polymorphism for the genetic fingerprinting of oil palm (*E. guineensis* Jacq.). Molecular Breeding, 2(2), 175-180.
- Mayes, S., Jack, P. L., Corley, R. H. V., & Marshall, D. F. (1997). Construction of a RFLP genetic linkage map for oil palm (*Elaeis guineensis* Jacq.). Genome, 40(1), 116-122.
- Mazur, B. J., & Tingey, S. V. (1995). Genetic mapping and introgression of genes of agronomic importance. Current Opinion in Biotechnology, 6(2), 175-182.
- McCouch, S. R., Chen, X., Panaud, O., Temnykh, S., Xu, Y., Cho, Y. G., [...] & Blair, M. (1997). Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant molecular biology, 35(1-2), 89-99.
- Mérai, Z., Chumak, N., García-Aguilar, M., Hsieh, T. F., Nishimura, T., Schoft, V. K., [...] & Mechtler, K. (2014). The AAA-ATPase molecular chaperone Cdc48/p97 disassembles sumoylated centromeres, decondenses heterochromatin, and activates ribosomal RNA genes. Proceedings of the National Academy of Sciences, 111(45), 16166-16171.
- Merriman, B., Torrent, I., Rothberg, J. M., & R&D Team. (2012). Progress in ion torrent semiconductor chip based sequencing. Electrophoresis, 33(23), 3397-3417.
- Mertes, F., ElSharawy, A., Sauer, S., van Helvoort, J. M., Van Der Zaag, P. J., Franke, A., [...] & Brookes, A. J. (2011). Targeted enrichment of genomic DNA regions for next-generation sequencing. Briefings in functional genomics, elr033.
- Meunier, J., Gascon, J. P., & Noiret, J. M. (1970). Hérédité des caractéristiques du régime d' *Elaeis guineensis* Jacq. en Côte-d'Ivoire. Héritabilité. Aptitude à la combinaison. Oléagineux, 25(7), 377-382.
- Micheletti, D., Dettori, M. T., Micali, S., Aramini, V., Pacheco, I., Linge, C. D. S., [...] & Lambert, P. (2015). Whole-genome analysis of diversity and SNP-major gene association in peach germplasm. PloS one, 10(9), e0136803.
- Michelmore, R. W., Paran, I., & Kesseli, R. V. (1991). Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect

markers in specific genomic regions by using segregating populations. Proceedings of the national academy of sciences, 88(21), 9828-9832.

- Millar, A. A., & Kunst, L. (1997). Very-long-chain fatty acid biosynthesis is controlled through the expression and specificity of the condensing enzyme. The Plant Journal, 12(1), 121-131.
- Millar, A. A., Wrischer, M., & Kunst, L. (1998). Accumulation of very-long-chain fatty acids in membrane glycerolipids is associated with dramatic alterations in plant morphology. The Plant Cell, 10(11), 1889-1902.
- Mitchell-Olds, T., & Pedersen, D. (1998). The molecular basis of quantitative genetic variation in central and secondary metabolism in Arabidopsis. Genetics, 149(2), 739-747.
- Mittler, R. (2002). Oxidative stress, antioxidants and stress tolerance. Trends in plant science, 7(9), 405-410.
- Mohan, A., Goyal, A., Singh, R., Balyan, H. S., & Gupta, P. K. (2007). Physical mapping of wheat and rye expressed sequence tag–simple sequence repeats on wheat chromosomes. Crop science, 47(Supplement_1), S-3.
- Montoya, C., Lopes, R., Flori, A., Cros, D., Cuellar, T., Summo, M., [...] & Zambrano, J. R. (2013). Quantitative trait loci (QTLs) analysis of palm oil fatty acid composition in an interspecific pseudo-backcross from *Elaeis oleifera* (HBK) Cortés and oil palm (*Elaeis guineensis* Jacq.). Tree genetics & genomes, 9(5), 1207-1225.
- Montoya, C., Cochard, B., Flori, A., Cros, D., Lopes, R., Cuellar, T., [...] & Ritter, E. (2014). Genetic architecture of palm oil fatty acid composition in cultivated oil palm (*Elaeis guineensis* Jacq.) compared to its wild relative *E. oleifera* (HBK) Cortés. PloS one, 9(5), e95412.
- Morcillo, F., Cros, D., Billotte, N., Ngando-Ebongue, G. F., Domonhédo, H., Pizot, M., [...] & Claverol, S. (2013). Improving palm oil quality through identification and mapping of the lipase gene causing oil deterioration. Nature communications, 4.
- Morgante, M., & Salamini, F. (2003). From plant genomics to breeding practice. Current Opinion in Biotechnology, 14(2), 214-219.
- Muench, D. G., Zhang, C., & Dahodwala, M. (2012). Control of cytoplasmic translation in plants. Wiley Interdisciplinary Reviews: RNA, 3(2), 178-194.
- Murphy, D. J. (2012). The dynamic roles of intracellular lipid droplets: from archaea to mammals. Protoplasma, 249(3), 541-585.
- Murphy, S., Martin, S., & Parton, R. G. (2009). Lipid droplet-organelle interactions; sharing the fats. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1791(6), 441-447.
- Musa, B. B., Saleh, G. B., & Loong, S. G. (2004). Genetic variability and broad-sense heritability in two Deli-AVROS D× P breeding populations of the oil palm (*Elaeis guineensis* Jacq.). SABRAO Journal of Breeding and Genetics, 36(1), 13-22.
- Myles, S., Peiffer, J., Brown, P. J., Ersoz, E. S., Zhang, Z., Costich, D. E., & Buckler, E. S. (2009). Association mapping: critical considerations shift from genotyping to experimental design. The Plant Cell, 21(8), 2194-2202.
- Nagano, M., Ihara-Ohori, Y., Imai, H., Inada, N., Fujimoto, M., Tsutsumi, N., [...] & Kawai-Yamada, M. (2009). Functional association of cell death suppressor, Arabidopsis Bax inhibitor-1, with fatty acid 2-hydroxylation through cytochrome b5. The Plant Journal, 58(1), 122-134.
- Nakatsukasa, K., & Brodsky, J. L. (2008). The recognition and retrotranslocation of misfolded proteins from the endoplasmic reticulum. Traffic,9(6), 861-870.

- Nam, J. W., & Kappock, T. J. (2007). Cloning and transcriptional analysis of *Crepis alpina* fatty acid desaturases affecting the biosynthesis of crepenynic acid. Journal of experimental botany, 58(6), 1421-1432.
- Napier, J. A., Michaelson, L. V., & Sayanova, O. (2003). The role of cytochrome b 5 fusion desaturases in the synthesis of polyunsaturated fatty acids. Prostaglandins, leukotrienes and essential fatty acids, 68(2), 135-143.
- Napier, J. A., Sayanova, O., Stobart, A. K., & Shewry, P. R. (1997). A new class of cytochrome b5 fusion proteins. Biochemical Journal, 328(Pt 2), 717.
- Nelissen, H., Fleury, D., Bruno, L., Robles, P., De Veylder, L., Traas, J., [...] & Van Lijsebettens, M. (2005). The elongata mutants identify a functional Elongator complex in plants with a role in cell proliferation during organ growth. Proceedings of the National Academy of Sciences of the United States of America, 102(21), 7754-7759.
- Ng CKY, Carr K, McAinsh MR, Powell B,Hetherington AM. 2001.Drought-induced guard cell signal transduction involves sphingosine-1-phosphate. Nature 410:596–599.
- Ngando-Ebongue, G. F. , Ajambang , W. N.,Koona, P, Lalu Firman, B., and Arondel, V. (2012). Oil Palm. Gupta, S. In: Technological Innovations in Major World Oil Crops, Vol I: Breeding (pp. 165-200). Springer New York
- Nielsen, T. H., Rung, J. H., & Villadsen, D. (2004). Fructose-2, 6-bisphosphate: a traffic signal in plant metabolism. Trends in plant science, 9(11), 556-563.
- Nuñez-Lillo, G., Cifuentes-Esquivel, A., Troggio, M., Micheletti, D., Infante, R., Campos-Vargas, R., [...] & Meneses, C. (2015). Identification of candidate genes associated with mealiness and maturity date in peach [*Prunus persica* (L.) Batsch] using QTL analysis and deep sequencing. Tree Genetics & Genomes, 11(4), 1-13.
- Odong, T. L., Van Heerwaarden, J., Jansen, J., van Hintum, T. J., & Van Eeuwijk, F. A. (2011). Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data?. Theoretical and applied genetics, 123(2), 195-205.
- Okoye, M. N., Okwuagwu, C. O., & Uguru, M. I. (2009). Population improvement for fresh fruit bunch yield and yield components in oil palm (*Elaeis guineensis* Jacq.). American-Eurasian Journal of Scientific Research, 4(2), 59-63.
- Okoye, M. N., Bakoumé, C., Uguru, M. I., Singh, R., & Okwuagwu, C. O. (2016). Genetic Relationships between Elite Oil Palms from Nigeria and Selected Breeding and Germplasm Materials from Malaysia via Simple Sequence Repeat (SSR) Markers. Journal of Agricultural Science, 8(2), 159.
- Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., & Matsubara, K. (1992). Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. Nature genetics, 2(3), 173-179.
- Okwuagwu, C. O., & Tai, G. C. C. (1995). Estimation of variance components and heritability of bunch yield and yield components in the oil palm (*Elaeis guineensis* Jacq.). Plant breeding, 114(5), 463-465.
- Okwuagwu, C. O., Okoye, M. N., Okolo, E. C., Ataga, C. D., & Uguru, M. I. (2008). Genetic variability of fresh fruit bunch yield in Deli/dura x tenera breeding populations of oil palm (*Elaeis guineensis* Jacq.) in Nigeria. Journal of Tropical Agriculture, 46, 52-57.
- Olmstead, R. G. (1990). Biological and historical factors influencing genetic diversity in the *Scutellaria angustifolia* complex (Labiatae). Evolution, 54-70.

- Omidvar, V., Akmar, A. S. N., Marziah, M., & Maheran, A. A. (2008). A transient assay to evaluate the expression of polyhydroxybutyrate genes regulated by oil palm mesocarp-specific promoter. Plant cell reports, 27(9), 1451-1459.
- Ong, P. W., Maizura, I., Abdullah, N. A. P., Rafii, M. Y., Ooi, L. C. L., Low, E. T. L., & Singh, R. (2015). Development of SNP markers and their application for genetic diversity analysis in the oil palm (*Elaeis guineensis*). Genetics and Molecular Research,14(4), 12205-12216.
- Oraguzie, N. C., Whitworth, C. J., Brewer, L., Hall, A., Volz, R. K., Bassett, H., & Gardiner, S. E. (2010). Relationships of PpACS1 and PpACS2 genotypes, internal ethylene concentration and fruit softening in European (*Pyrus communis*) and Japanese (*Pyrus pyrifolia*) pears during cold air storage. Plant breeding, 129(2), 219-226.
- Oraguzie, N.C., P.L. Wilcox, E.H.A. Rikkerink and H.N. De Silva, 2007. Linkage disequilibrium". In: Oraguzie, N.C., E.H.A. Rikkerink, S.E. Gardiner and H.N. De Silva (eds.), Association Mapping in Plants, pp: 11–39. Springer, New York, USA
- Ortiz, R. (1998). Critical role of plant biotechnology for the genetic improvement of food crops: perspectives for the next millennium. Electronic Journal of Biotechnology, 1(3), 16-17.
- Osborne, A. R., Rapoport, T. A., & van den Berg, B. (2005). Protein translocation by the Sec61/SecY channel. Annu. Rev. Cell Dev. Biol., 21, 529-550.
- Pabón-Mora, N., Wong, G. K. S., & Ambrose, B. A. (2014). Evolution of fruit development genes in flowering plants. Molecular basis of fruit development, 116.
- Palaisa, K. A., Morgante, M., Williams, M., & Rafalski, A. (2003). Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. The Plant Cell, 15(8), 1795-1806.
- Palaisa, K., Morgante, M., Tingey, S., & Rafalski, A. (2004). Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an asymmetric selective sweep. Proceedings of the National Academy of Sciences of the United States of America, 101(26), 9885-9890.
- Pan, X., Peng, F. Y., & Weselake, R. J. (2015). Genome-wide analysis of PHOSPHOLIPID: DIACYLGLYCEROL ACYLTRANSFERASE (PDAT) genes in plants reveals the eudicot-wide PDAT gene expansion and altered selective pressures acting on the core eudicot PDAT paralogs. Plant physiology, 167(3), 887-904.
- Patel, D. A., Zander, M., Dalton-Morgan, J., & Batley, J. (2015). Advances in Plant Genotyping: Where the Future Will Take Us. Plant Genotyping: Methods and Protocols, 1-11.
- Paterson, A. H. (1996). Making genetic maps. Genome mapping in plants, 194.
- Patnala, R., Clements, J., & Batra, J. (2013). Candidate gene association studies: a comprehensive guide to useful in silico tools. BMC genetics, 14(1), 39.
- Pattarapimol, T., Thuzar, M., Vanavichit, A., Tragoonrung, S., Roytrakul, S., & Jantasuriyarat, C. (2015). Identification of genes involved in somatic embryogenesis development in oil palm (*Elaeis guineensis* Jacq.) using cDNA AFLP. Journal of Oil Palm Research, 27(1), 1-11.
- Peal, L., Jambunathan, N., & Mahalingam, R. (2011). Phylogenetic and expression analysis of RNA-binding proteins with triple RNA recognition motifs in plants. Molecules and cells, 31(1), 55-64.
- Peña Malavera, A., Gutierrez, L., & Balzarini, M. (2014). Componentes principales en mapeo asociativo. BAG. Journal of basic and applied genetics, 25(2), 32-40.

- Peng, Q., Satya, R. V., Lewis, M., Randad, P., & Wang, Y. (2015). Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC genomics, 16(1), 1.
- Penner, G., 1996. RAPD analysis of plant genomes, In: P.P. Jauhar (Ed.), Methods of Genome Analysis in Plants, pp. 251–268.CRC Press, Boca Raton.
- Pevsner, J. (2009). Functional Genomics. Bioinformatics and Functional Genomics, Second Edition, 460-514.
- Pflieger, S., Lefebvre, V., & Causse, M. (2001). The candidate gene approach in plant genetics: a review. Molecular breeding, 7(4), 275-291.
- Pfluger, J., & Wagner, D. (2007). Histone modifications and dynamic regulation of genome accessibility in plants. Current opinion in plant biology, 10(6), 645-652.
- Phillips Mora, W., Rodríguez, R., & Fritz, P. J. (1995). Marcadores de ADN: teoría, aplicaciones y protocolos de trabajo con ejemplos de investigaciones en cacao (*Theobroma cacao*).
- Phongdara, A., Nakkaew, A., & Nualkaew, S. (2012). Isolation of the detoxification enzyme EgP450 from an oil palm EST library. Pharmaceutical biology, 50(1), 120-127.
- Pidkowich, M. S., Nguyen, H. T., Heilmann, I., Ischebeck, T., & Shanklin, J. (2007). Modulating seed β-ketoacyl-acyl carrier protein synthase II level converts the composition of a temperate seed oil to that of a palm-like tropical oil. Proceedings of the National Academy of Sciences, 104(11), 4742-4747.
- Plaxton, W. C. (1996). The organization and regulation of plant glycolysis. Annual review of plant biology, 47(1), 185-214.
- Polesani, M., Desario, F., Ferrarini, A., Zamboni, A., Pezzotti, M., Kortekamp, A., & Polverari, A. (2008). cDNA-AFLP analysis of plant and pathogen genes expressed in grapevine infected with *Plasmopara viticola*. Bmc Genomics, 9(1), 142.
- Ponting, C. P., & Aravind, L. (1999). START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. Trends in biochemical sciences, 24(4), 130-132.
- Pootakham, W., Jomchai, N., Ruang-areerate, P., Shearman, J. R., Sonthirod, C., Sangsrakru, D., [...] & Tangphatsornruang, S. (2015). Genome-wide SNP discovery and identification of QTL associated with agronomic traits in oil palm using genotyping-by-sequencing (GBS). Genomics, 105(5), 288-295.
- Pootakham, W., Uthaipaisanwong, P., Sangsrakru, D., Yoocha, T., Tragoonrung, S., & Tangphatsornruang, S. (2013). Development and characterization of single-nucleotide polymorphism markers from 454 transcriptome sequences in oil palm (*Elaeis guineensis*). Plant Breeding, 132(6), 711-717.
- Popow, J., Schleiffer, A., & Martinez, J. (2012). Diversity and roles of (t) RNA ligases. Cellular and Molecular Life Sciences, 69(16), 2657-2670.
- Powell, W., & Moss, J. P. (1992). Plant genomes, gene markers, and linkage maps. Biotechnology and crop improvement in Asia., 297-322
- Powell, W., Machray, G. C., & Provan, J. (1996). Polymorphism revealed by simple sequence repeats. Trends in plant science, 1(7), 215-222.
- Preedakoon, P. (2009). DISCOVERY OF PLANT HORMONE SIGNAL TRANSDUCTION HOMOLOGS IN OIL PALM (*ELAEIS GUINEENSIS* JACQ.) (Doctoral dissertation, KASETSART UNIVERSITY).
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics, 38(8), 904-909.

- Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics, 155(2), 945-959.
- Pritchard, J. K., Wen, X., & Falush, D. (2009). Documentation for structure software: Version 2.3.
- Purseglove, J. W. (1972). Tropical crops: monocotyledons, vols 1 and 2 (p. 334). Longman, London.
- Puzio,P., Blau,A., Walk,T.B., Gipmans,M., Haake,V., Weig,A.,Plesch,G. and Ebneth,M (2009). Process for the production of a fine chemical. WO 2008034648
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., [...] & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics, 13(1), 1.
- Queen, R. A., Gribbon, B. M., James, C., Jack, P., & Flavell, A. J. (2004). Retrotransposon-based molecular markers for linkage and genetic diversity analysis in wheat. Molecular Genetics and Genomics, 271(1), 91-97.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rafalski, J. A., & Tingey, S. V. (1993). Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends in Genetics, 9(8), 275-280.
- Rafalski, A. (2002). Applications of single nucleotide polymorphisms in crop genetics. Current opinion in plant biology, 5(2), 94-100.
- Rafalski, J. A. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant science, 162(3), 329-333.
- Rafalski, J. A. (2010). Association genetics in crop improvement. Current opinion in plant biology, 13(2), 174-180.
- Rafii, M. Y., Rajanaidu, N., Jalani, B. S., & Kushairi, A. (2002). Performance and heritability estimations on oil palm progenies tested in different environments. Journal of Oil Palm Research, 14(1), 15-24.
- Rahier, A., Smith, M., & Taton, M. (1997). The role of cytochrome b 5 in 4α-methyl-oxidation and C5 (6) desaturation of plant sterol precursors. Biochemical and biophysical research communications, 236(2), 434-437.
- Rajanaidu , N. & Jalani, B.(1994). Influence of planting material on oil extraction ratio (OER). In National Seminar on Palm oil extraction rate: problemas and issues. December 21-22, Kuala Lumpur (No. L-0302). PORIM.
- Rajanaidu, N., & Jalani, B. S. (1994). Prospects for breeding for kernels in oil palm (*Elaeis guineensis*). Planter, 70(820), 307-318.
- Rajanaidu, N., Kushairi, A., Rafii, M., Mohd Din, M., Maizura, I., and Jalani, B. S. Oil Palm Breeding and Genetic Resources (2000). Basiron, Y., Jalani, B.S. and Chan, K.W. In: Advances in oil palm research. Vol.I (pp. 171-237). MPOB
- Rallo, P., Belaj, A., De La Rosa, R., & Trujillo, I. (2002). Marcadores moleculares (en línea). Córdoba, España.
- Ramachandran, S., & Sundaresan, V. (2001). Transposons as tools for functional genomics. Plant Physiology and Biochemistry, 39(3), 243-252.
- Ramli, Z., & Abdullah, S. N. A. (2010). Functional characterisation of the oil palm type 3 metallothionein-like gene (MT3-B) promoter. Plant molecular biology reporter, 28(3), 531-541.
- Rampino, P., Pataleo, S., Falco, V., Mita, G., & Perrotta, C. (2011). Identification of candidate genes associated with senescence in durum wheat (*Triticum turgidum* subsp. durum) using cDNA-AFLP. Molecular biology reports, 38(8), 5219-5229.

- Rance, K. A., Mayes, S., Price, Z., Jack, P. L., & Corley, R. H. V. (2001). Quantitative trait loci for yield components in oil palm (*Elaeis guineensis* Jacq.). Theoretical and Applied Genetics, 103(8), 1302-1310.
- Rancour, D. M., Dickey, C. E., Park, S., & Bednarek, S. Y. (2002). Characterization of AtCDC48. Evidence for multiple membrane fusion mechanisms at the plane of cell division in plants. Plant physiology, 130(3), 1241-1253.
- Ratcliffe, O. J., & Riechmann, J. L. (2002). Arabidopsis transcription factors and the regulation of flowering time: a genomic perspective. Current issues in molecular biology, 4, 77-92.
- Rautengarten, C., Ebert, B., Herter, T., Petzold, C. J., Ishii, T., Mukhopadhyay, A., [...] & Scheller, H. V. (2011). The interconversion of UDP-arabinopyranose and UDP-arabinofuranose is indispensable for plant development in Arabidopsis. The Plant Cell, 23(4), 1373-1390.
- Reid, S. J., & Ross, G. S. (1997). Up-regulation of two cDNA clones encoding metallothionein-like proteins in apple fruit during cool storage. Physiologia Plantarum, 100(1), 183-189.
- Reiter, W. D. (2002). Biosynthesis and properties of the plant cell wall. Current opinion in plant biology, 5(6), 536-542.
- Remington, D. L., & Purugganan, M. D. (2003). Candidate genes, quantitative trait loci, and functional trait evolution in plants. International Journal of Plant Sciences, 164(S3), S7-S20.
- Reuzeau, C., Frankard, V., Sanz Molinero, A.I. (2010). Plants having enhanced yield-related traits and a method for making the same. EP2199398
- Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C. Z., Keddie, J., [...] & Creelman, R. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science, 290(5499), 2105-2110.
- Riju, A., Chandrasekar, A., & Arunachalam, V. (2007). Mining for single nucleotide polymorphisms and insertions/deletions in expressed sequence tag libraries of oil palm. Bioinformation, 2(4), 128-131.
- Rival, A., Bertrand, L., Beulé, T., Combes, M. C., Trouslot, P., & Lashermes, P. (1998). Suitability of RAPD analysis for the detection of somaclonal variants in oil palm (*Elaeis guineensis* Jacq). Plant Breeding, 117(1), 73-76.
- Roberdi, R, Sobir, Yahya, S., & Toruan-Mathius, N. (2015). Differential analysis of gene related to hard bunch phenomena in oil palm (*Elaeis guineensis* Jacq) fruits. Emirates Journal of Food and Agriculture, 27(8), 596.
- Ros, R., Muñoz-Bertomeu, J., & Krueger, S. (2014). Serine in plants: biosynthesis, metabolism, and functions. Trends in plant science, 19(9), 564-569.
- Rosenquist, E. A. (1986). The genetic base of oil palm breeding populations. In International Workshop on Oil Palm Germplasm and Utilisation, Bangi, Selangor (Malaysia), 26-27 Mar 1985. IPMKSM.
- Rosenquist, E. A. (1990). An overview of breeding technology and selection in *Elaeis guineensis*. In Proceedings of the 1989 International Palm Oil Development Conference–Agriculture (pp. 5-26).
- Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., [...] & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data.Genome biology, 14(5), 1.
- Rothgänger, J., Weniger, M., Weniger, T., Mellmann, A., & Harmsen, D. (2006). Ridom TraceEdit: a DNA trace editor and viewer. Bioinformatics, 22(4), 493-494.

- Rothschild, M. F., & Soller, M. (1997). Candidate gene analysis to detect genes controlling traits of economic importance in domestic livestock. Probe, 8(2), 13-20.
- Ruggieri, V., Francese, G., Sacco, A., D'Alessandro, A., Rigano, M. M., Parisi, M., [...] & Barone, A. (2014). An association mapping approach to identify favourable alleles for tomato fruit quality breeding. BMC plant biology, 14(1), 337.
- Ruiz-Medrano R, Xoconostle-Cazares B, Lucas WJ. (2001) The phloem as a conduit for inter-organ communication. Current Opinion Plant Biology, 4:202–209.
- Rylott, E. L., Hooks, M. A., & Graham, I. A. (2001). Co-ordinate regulation of genes involved in storage lipid mobilization in *Arabidopsis thaliana*. Biochemical Society Transactions, 29(2), 283-286.
- Rylott, E. L., Rogers, C. A., Gilday, A. D., Edgell, T., Larson, T. R., & Graham, I. A. (2003). Arabidopsis mutants in short-and medium-chain acyl-CoA oxidase activities accumulate acyl-CoAs and reveal that fatty acid β-oxidation is essential for embryo development. Journal of Biological Chemistry, 278(24), 21370-21377.
- Safi, H., Saibi, W., Alaoui, M. M., Hmyene, A., Masmoudi, K., Hanin, M., & Brini, F. (2015). A wheat lipid transfer protein (TdLTP4) promotes tolerance to abiotic and biotic stress in *Arabidopsis thaliana*. Plant Physiology and Biochemistry, 89, 64-75.
- Saka, H. A., & Valdivia, R. (2012). Emerging roles for lipid droplets in immunity and host-pathogen interactions. Annual review of cell and developmental biology, 28, 411-437.
- Salinas, T., Duchêne, A. M., Delage, L., Nilsson, S., Glaser, E., Zaepfel, M., & Maréchal-Drouard, L. (2006). The voltage-dependent anion channel, a major component of the tRNA import machinery in plant mitochondria. Proceedings of the National Academy of Sciences, 103(48), 18362-18367.
- Salmaso, M., Faes, G., Segala, C., Stefanini, M., Salakhutdinov, I., Zyprian, E., [...] & Velasco, R. (2004). Genome diversity and gene haplotypes in the grapevine (*Vitis vinifera* L.), as revealed by single nucleotide polymorphisms. Molecular Breeding,14(4), 385-395.
- Sambanthamurthi, R., Singh, R., Kadir, A. P. G., Abdullah, M. O., & Kushairi, A. (2009). Opportunities for the oil palm via breeding and biotechnology. In Breeding plantation tree crops: Tropical species (pp. 377-421). Springer New York.
- Sambanthamurthi, R., Sundram, K., & Tan, Y. A. (2000). Chemistry and biochemistry of palm oil. Progress in lipid research, 39(6), 507-558.
- Sambrook, J., Fritsch, E.F., and Maniatis, T. (1989). Molecular Cloning: A Laboratory Manual, 2nd ed. (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press).
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences, 74(12), 5463-5467.
- Sanz,M.A. and Reuzeau,C. (2013). Plants having enhanced yield-related traits and/or increased abiotic stress resistance, and a method for making the same. EP 2615174
- Sarafian, V., Kim, Y., Poole, R. J., & Rea, P. A. (1992). Molecular cloning and sequence of cDNA encoding the pyrophosphate-energized vacuolar membrane proton pump of *Arabidopsis thaliana.* Proceedings of the National Academy of Sciences,89(5), 1775-1779.
- Sauvage, C., Segura, V., Bauchet, G., Stevens, R., Do, P. T., Nikoloski, Z., [...] & Causse, M. (2014). Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant physiology, 165(3), 1120-1132.

- Saveanu, C., Bienvenu, D., Namane, A., Gleizes, P. E., Gas, N., Jacquier, A., & Fromont-Racine, M. (2001). Nog2p, a putative GTPase associated with pre-60S subunits and required for late 60S maturation steps. The EMBO Journal, 20(22), 6475-6484.
- Scheible, W. R., & Pauly, M. (2004). Glycosyltransferases and cell wall biosynthesis: novel players and insights. Current opinion in plant biology, 7(3), 285-295.
- Scheller, H. V., Jensen, P. E., Haldrup, A., Lunde, C., & Knoetzel, J. (2001). Role of subunits in eukaryotic Photosystem I. Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1507(1), 41-60.
- Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270(5235), 467.
- Schneider, K., Kulosa, D., Soerensen, T. R., Möhring, S., Heine, M., Durstewitz, G., [...] & Tahiro, E. (2007). Analysis of DNA polymorphisms in sugar beet (*Beta vulgaris* L.) and development of an SNP-based map of expressed genes. Theoretical and Applied Genetics, 115(5), 601-615.
- Schopfer, P., Liszkay, A., Bechtold, M., Frahry, G. & Wagner, A. Evidence that hydroxyl radicals mediate auxin- induced extension growth. Planta 214, 821–828 (2002).
- Schürmann, P. (2003). Redox signaling in the chloroplast: the ferredoxin/thioredoxin system. Antioxidants and Redox Signaling, 5(1), 69-78.
- Seelert, H., Poetsch, A., Dencher, N. A., Engel, A., Stahlberg, H., & Müller, D. J. (2000). Structural biology: proton-powered turbine of a plant motor. Nature, 405(6785), 418-419.
- Selth, L. A., Dogra, S. C., Rasheed, M. S., Randles, J. W., & Rezaian, M. A. (2006). Identification and characterization of a host reversibly glycosylated peptide that interacts with the Tomato leaf curl virus V1 protein. Plant molecular biology, 61(1-2), 297-310.
- Semagn, K., & Ndjiondjop, M. N. (2006). Progress and prospects of marker assisted backcrossing as a tool in crop breeding programs. African Journal of Biotechnology, 5(25), 2588-2603.
- Seng, T. Y., Saad, S. H. M., Chin, C. W., Ting, N. C., Singh, R. S. H., Zaman, F. Q., [...] & Alwee, S. S. R. S. (2011). Genetic linkage map of a high yielding FELDA Deli× Yangambi oil palm cross. PLoS One, 6(11), e26593.
- Senn, M. E., Grozeff, G. G., Alegre, M. L., Barrile, F., De Tullio, M. C., & Bartoli, C. G. (2016). Effect of mitochondrial ascorbic acid synthesis on photosynthesis. Plant Physiology and Biochemistry.
- Sexton, T. R., Henry, R. J., Harwood, C. E., Thomas, D. S., McManus, L. J., Raymond, C., [...] & Shepherd, M. (2012). Pectin methylesterase genes influence solid wood properties of *Eucalyptus pilularis*. Plant physiology, 158(1), 531-541.
- Seymour, G., Poole, M., Manning, K., & King, G. J. (2008). Genetics and epigenetics of fruit development and ripening. Current opinion in plant biology, 11(1), 58-63.
- Shamsi, I. H., Shamsi, B. H., & Jiang, L. (2012). Biochemistry of fatty acids. In Technological Innovations in Major World Oil Crops, Volume 2 (pp. 123-150). Springer New York.
- Sharma, M. & Tan, Y.P. (1997). Oil Palm Breeding Programme and the Performance of DxP Planting Materials at United Plantations Berhad. Planter, 73, 591-610.
- Sharma, M., Gupta, S. K., & Mondal, A. K. (2012). Production and trade of major world oil crops. In Technological Innovations in Major World Oil Crops, Volume 1 (pp. 1-15). Springer New York.

- Sharma, A., & Chauhan, R. S. (2012). In silico identification and comparative genomics of candidate genes involved in biosynthesis and accumulation of seed oil in plants. Comparative and functional genomics, 2012.
- Shearman, J. R., Jantasuriyarat, C., Sangsrakru, D., Yoocha, T., Vannavichit, A., Tragoonrung, S., & Tangphatsornruang, S. (2013). Transcriptome analysis of normal and mantled developing oil palm flower and fruit. Genomics, 101(5), 306-312.
- Shen, W. H., & Xu, L. (2009). Chromatin remodeling in stem cell maintenance in *Arabidopsis* thaliana. Molecular plant, 2(4), 600-609.
- Shi, C., Chaudhary, S., Yu, K., Park, S. J., Navabi, A., & McClean, P. E. (2011). Identification of candidate genes associated with CBB resistance in common bean HR45 (*Phaseolus vulgaris* L.) using cDNA-AFLP. Molecular biology reports, 38(1), 75-81.
- Shi, L., Katavic, V., Yu, Y., Kunst, L., & Haughn, G. (2012). Arabidopsis glabra2 mutant seeds deficient in mucilage biosynthesis produce more oil. The Plant Journal, 69(1), 37-46.
- Shi, Q. M., Yang, X., Song, L., & Xue, H. W. (2011). Arabidopsis MSBP1 is activated by HY5 and HYH and is involved in photomorphogenesis and brassinosteroid sensitivity regulation. Molecular plant, 4(6), 1092-1104.
- Shimotohno, A., Umeda-Hara, C., Bisova, K., Uchimiya, H., & Umeda, M. (2004). The plant-specific kinase CDKF; 1 is involved in activating phosphorylation of cyclin-dependent kinase-activating kinases in Arabidopsis. The Plant Cell, 16(11), 2954-2966.
- Shore, P., & Sharrocks, A. D. (1995). The MADS-box family of transcription factors. European Journal of Biochemistry, 229(1), 1-13.
- Shpilka, T., Weidberg, H., Pietrokovski, S., & Elazar, Z. (2011). Atg8: an autophagy-related ubiquitin-like protein family. Genome Biology, 12(7), 226.
- Simko, I., Haynes, K. G., & Jones, R. W. (2006). Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics, 173(4), 2237-2245.
- Simko, I., Pechenick, D. A., McHale, L. K., Truco, M. J., Ochoa, O. E., Michelmore, R. W., & Scheffler, B. E. (2009). Association mapping and marker-assisted selection of the lettuce dieback resistance gene Tvr1. BMC Plant Biology, 9(1), 135.
- Simões, I., & Faro, C. (2004). Structure and function of plant aspartic proteinases. European journal of biochemistry, 271(11), 2067-2075.
- Singh, R., & Cheah Suan, C. (2005). Potential application of marker-assisted selection (MAS) in oil palm. Oil Palm Bulletin, (51), 1-9.
- Singh, R., Tan, S. G., Panandam, J. M., Rahman, R. A., & Cheah, S. C. (2008). Identification of cDNA-RFLP markers and their use for molecular mapping in oil palm (*Elaeis guineensis*). Asia-Pacific Journal of Molecular Biology and Biotechnology, 16(3), 53-63.
- Singh, R., Zaki, N. M., Ting, N. C., Rosli, R., Tan, S. G., Low, E. T. L., [...] & Cheah, S. C. (2008). Exploiting an oil palm EST database for the development of gene-derived SSR markers and their exploitation for assessment of genetic diversity. Biologia, 63(2), 227-235.
- Singh, R., Tan, S. G., Panandam, J. M., Rahman, R. A., Ooi, L. C., Low, E. T. L., [...] & Cheah, S. C. (2009). Mapping quantitative trait loci (QTLs) for fatty acid composition in an interspecific cross of oil palm. BMC Plant Biology, 9(1), 114.
- Singh, R., Ong-Abdullah, M., Low, E. T. L., Manaf, M. A. A., Rosli, R., Nookiah, R., [...] & Azizi, N. (2013a). Oil palm genome sequence reveals divergence of interfertile species in Old and New worlds. Nature, 500(7462), 335-339.
- Singh, R., Low, E. T. L., Ooi, L. C. L., Ong-Abdullah, M., Ting, N. C., Nagappan, J., [...] & Chan, K. L. (2013b). The oil palm SHELL gene controls oil yield and encodes a homologue of SEEDSTICK. Nature, 500(7462), 340-344.

- . Singh, R., Low, E. T. L., Ooi, L. C. L., Ong-Abdullah, M., Nookiah, R., Ting, N. C., [...] & Nagappan, J. (2014). The oil palm VIRESCENS gene controls fruit colour and encodes a R2R3-MYB. Nature communications, 5.
- Slabas, A. R., & Fawcett, T. (1992). The biochemistry and molecular biology of plant lipid biosynthesis. Plant molecular biology, 19(1), 169-191.
- Smalle, J., & Vierstra, R. D. (2004). The ubiquitin 26S proteasome proteolytic pathway. Annual Review Plant Biology, 55, 555-590.
- Smirnoff, N., & Wheeler, G. L. (2000). Ascorbic acid in plants: biosynthesis and function. Critical Reviews in Biochemistry and Molecular Biology, 35(4), 291-314.
- Smith, M. A., Jonsson, L., Stymne, S., & Stobart, K. (1992). Evidence for cytochrome b5 as an electron donor in ricinoleic acid biosynthesis in microsomal preparations from developing castor bean (*Ricinus communis* L.). Biochemical journal, 287(1), 141-144.
- Smith, R. G., Gauthier, D. A., Dennis, D. T., & Turpin, D. H. (1992). Malate-and pyruvate-dependent fatty acid synthesis in leucoplasts from developing castor endosperm. Plant Physiology, 98(4), 1233-1238.
- Soh, A. C. (1999). Breeding plans and selection methods in oil palm. In Symposium on the science of oil palm breeding. Proceedings,1992. Montpellier, Francia (No. L-0435). PORIM.
- Soh, A. C., & Hor, T. Y. (2000). Combining ability correlations for bunch yield and its components in outcrossed populations of oil palm. In Proceedings of the International Symposium on Oil Palm Genetic Resources and Utilization. Malaysian Palm Oil Board, Kuala Lumpur (pp. M1-M14).
- Soh, A. C., Wong, G., Hor, T. Y., Tan, C. C., & Chew, P. S. (2003). Oil palm genetic improvement. Plant Breeding Reviews, 22, 165-220.
- Soh A.C. (2011) Genomics and plant breeding. J. Oil Palm Res., 23, 1019-1028.
- Song, L., Shi, Q. M., Yang, X. H., Xu, Z. H., & Xue, H. W. (2009). Membrane steroid-binding protein 1 (MSBP1) negatively regulates brassinosteroid signaling by enhancing the endocytosis of BAK1. Cell research, 19(7), 864-876.
- Soto-Cerda, B. J., & Cloutier, S. (2012). Association mapping in plant genomes. INTECH Open Access Publisher.
- Sparnaaij L.D., Rees A.R., Chapas L.C. (1963) Annual yield variation in the oil palm. Journal of West African Institute for Oil Palm Research, 4, 11–125.
- Sperling, P., & Heinz, E. (2003). Plant sphingolipids: structural diversity, biosynthesis, first genes and functions. Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids, 1632(1), 1-15.
- Stevens, R., Buret, M., Duffé, P., Garchery, C., Baldet, P., Rothan, C., & Causse, M. (2007). Candidate genes and quantitative trait loci affecting fruit ascorbic acid content in three tomato populations. Plant Physiology, 143(4), 1943-1953.
- Stölting, K. N., Gort, G., Wüst, C., & Wilson, A. B. (2009). Eukaryotic transcriptomics in silico: optimizing cDNA-AFLP efficiency. BMC genomics, 10(1), 565.
- Stone, S. L., Hauksdóttir, H., Troy, A., Herschleb, J., Kraft, E., & Callis, J. (2005). Functional analysis of the RING-type ubiquitin ligase family of Arabidopsis. Plant physiology, 137(1), 13-30.
- Strommer, J. (2011). The plant ADH gene family. The Plant Journal, 66(1), 128-142.
- Sun, T. P., & Gubler, F. (2004). Molecular mechanism of gibberellin signaling in plants. Annual Review Plant Biology, 55, 197-223.

- Sung, S., & Amasino, R. M. (2004). Vernalization in *Arabidopsis thaliana* is mediated by the PHD finger protein VIN3. Nature, 427(6970), 159-164.
- Sung, S., Schmitz, R. J., & Amasino, R. M. (2006). A PHD finger protein involved in both the vernalization and photoperiod pathways in Arabidopsis. Genes & development, 20(23), 3244-3248.
- Tabor, H. K., Risch, N. J., & Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics, 3(5), 391-397.
- 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. Australian Journal of Crop Science,9(4), 281.
- Takatsuka, H., Ohno, R., & Umeda, M. (2009). The Arabidopsis cyclin-dependent kinase-activating kinase CDKF; 1 is a major regulator of cell proliferation and cell expansion but is dispensable for CDKA activation. The Plant Journal, 59(3), 475-487.
- Tang, D., Ade, J., Frye, C. A., & Innes, R. W. (2005). Regulation of plant defense responses in Arabidopsis by EDR2, a PH and START domain-containing protein. The Plant Journal, 44(2), 245-257.
- Tanksley, S. D. (1983). Molecular markers in plant breeding. Plant Molecular Biology Reporter, 1(1), 3-8.
- Tanksley, S. D., Young, N. D., Paterson, A. H., & Bonierbale, M. W. (1989). RFLP mapping in plant breeding: new tools for an old science. Nature Biotechnology, 7(3), 257-264.
- Tanksley, S. D. (1993). Mapping polygenes. Annual review of genetics, 27(1), 205-233.
- Taramino, G., & Tingey, S. (1996). Simple sequence repeats for germplasm analysis and mapping in maize. Genome, 39(2), 277-287.
- Tateda, C., Watanabe, K., Kusano, T., & Takahashi, Y. (2011). Molecular and genetic characterization of the gene family encoding the voltage-dependent anion channel in Arabidopsis. Journal of experimental botany, 62(14), 4773-4785.
- Taylor, R. D., & Pfanner, N. (2004). The protein import and assembly machinery of the mitochondrial outer membrane. Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1658(1), 37-43.
- Teh, C. K., Ong, A. L., Kwong, Q. B., Apparow, S., Chew, F. T., Mayes, S., [...] & Kulaveerasingam, H. (2016). Genome-wide association study identifies three key loci for high mesocarp oil content in perennial crop oil palm. Scientific reports, 6.
- Teh, H. F., Neoh, B. K., Hong, M. P. L., Low, J. Y. S., Ng, T. L. M., Ithnin, N., [...] & Kulaveerasingam, H. (2013). Differential metabolite profiles during fruit development in high-yielding oil palm mesocarp. PloS one, 8(4), e61344.
- Tello, J., Torres-Pérez, R., Grimplet, J., & Ibáñez, J. (2016). Association analysis of grapevine bunch traits using a comprehensive approach. Theoretical and Applied Genetics, 129(2), 227-242.
- Teng, C., Dong, H., Shi, L., Deng, Y., Mu, J., Zhang, J., [...] & Zuo, J. (2008). Serine palmitoyltransferase, a key enzyme for de novo synthesis of sphingolipids, is essential for male gametophyte development in Arabidopsis. Plant physiology, 146(3), 1322-1332.
- Tester, M., & Langridge, P. (2010). Breeding technologies to increase crop production in a changing world. Science, 327(5967), 818-822.
- Thoday, J.M. 1961. Location of polygenes. Nature 191:291-296

- Thompson, A. R., & Vierstra, R. D. (2005). Autophagic recycling: lessons from yeast help define the process in plants. Current opinion in plant biology, 8(2), 165-173.
- Thompson, J. D., Gibson, T., & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Current protocols in bioinformatics, 2-3.
- Thormählen, I., Meitzel, T., Groysman, J., Öchsner, A. B., von Roepenack-Lahaye, E., Naranjo, B., [...] & Geigenberger, P. (2015). Thioredoxin f1 and NADPH-dependent thioredoxin reductase C have overlapping functions in regulating photosynthetic metabolism and plant growth in response to varying light conditions. Plant physiology, 169(3), 1766-1786.
- Thornsberry, J. M., Goodman, M. M., Doebley, J., Kresovich, S., Nielsen, D., & Buckler, E. S. (2001). Dwarf8 polymorphisms associate with variation in flowering time. Nature genetics, 28(3), 286-289.
- Thumma BR, Naidu BP, Chandra A, Cameron DF, Bahnisch LM, Liu C (2001). Identification of causal relationships among traits related to drought resistance in *Stylosanthes scabra* using QTL analysis. Journal of Experimental Botany, 52:203
- Ting, N. C., Zaki, N. M., Rosli, R., Low, E. T. L., Ithnin, M., Cheah, S. C., [...] & Singh, R. (2010). SSR mining in oil palm EST database: application in oil palm germplasm diversity studies. Journal of genetics, 89(2), 135-145.
- Ting, N. C., Jansen, J., Mayes, S., Massawe, F., Sambanthamurthi, R., Ooi, L. C. L., [...] & Ithnin, M. (2014). High density SNP and SSR-based genetic maps of two independent oil palm hybrids. BMC genomics, 15(1), 1.
- Todd, J., Post-Beittenmiller, D., & Jaworski, J. G. (1999). KCS1 encodes a fatty acid elongase 3-ketoacyl-CoA synthase affecting wax biosynthesis in *Arabidopsis thaliana*. The Plant Journal, 17(2), 119-130.
- Tognolli, M., Penel, C., Greppin, H., & Simon, P. (2002). Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*. Gene, 288(1), 129-138.
- Tranbarger, T. J., Dussert, S., Joët, T., Argout, X., Summo, M., Champion, A., [...] & Morcillo, F. (2011). Regulatory mechanisms underlying oil palm fruit mesocarp maturation, ripening, and functional specialization in lipid and carotenoid metabolism. Plant Physiology, 156(2), 564-584.
- Tranbarger, T. J., Kluabmongkol, W., Sangsrakru, D., Morcillo, F., Tregear, W. J., Tragoonrung, S., & Billotte, N. (2012). SSR markers in transcripts of genes linked to post-transcriptional and transcriptional regulatory functions during vegetative and reproductive development of *Elaeis guineensis*. BMC plant biology, 12(1), 1.
- Troncoso-Ponce, M. A., Kilaru, A., Cao, X., Durrett, T. P., Fan, J., Jensen, J. K., [...] & Ohlrogge, J. B. (2011). Comparative deep transcriptional profiling of four developing oilseeds. The Plant Journal, 68(6), 1014-1027.
- Tsukamoto, T., Qin, Y., Huang, Y., Dunatunga, D., & Palanivelu, R. (2010). A role for LORELEI, a putative glycosylphosphatidylinositol-anchored protein, in *Arabidopsis thaliana* double fertilization and early seed development. The Plant Journal, 62(4), 571-588.
- Tzen, J. T., & Huang, A. H. (1992). Surface structure and properties of plant seed oil bodies. The Journal of cell biology, 117(2), 327-335.
- Uhrig, R. G., Labandera, A. M., & Moorhead, G. B. (2013). Arabidopsis PPP family of serine/threonine protein phosphatases: many targets but few engines. Trends in plant science, 18(9), 505-513.

- Uitdewilligen, J. G., Wolters, A. M. A., Bjorn, B., Borm, T. J., Visser, R. G., & van Eck, H. J. (2013). A next-generation sequencing method for genotyping-by-sequencing of highly heterozygous autotetraploid potato. PLoS One, 8(5), e62355.
- Umeda, M., Shimotohno, A., & Yamaguchi, M. (2005). Control of cell division and transcription by cyclin-dependent kinase-activating kinases in plants. Plant and Cell Physiology, 46(9), 1437-1442.
- Umemoto, N., Kobayashi, O., Ishizaki-Nishizawa, O., & Toguri, T. (1993). cDNAs sequences encoding cytochrome P450 (CYP71 family) from eggplant seedlings. FEBS letters, 330(2), 169-173.
- van der Schoot, C., Paul, L. K., Paul, S. B., & Rinne, P. L. (2011). Plant lipid bodies and cell-cell signaling: a new role for an old organelle?. Plant signaling & behavior, 6(11), 1732-1738.
- van Dijken, A. J., Schluepmann, H., & Smeekens, S. C. (2004). Arabidopsis trehalose-6-phosphate synthase 1 is essential for normal vegetative growth and transition to flowering. Plant physiology, 135(2), 969-977.
- Vargas, M., van Eeuwijk, F. A., Crossa, J., & Ribaut, J. M. (2006). Mapping QTLs and QTL× environment interaction for CIMMYT maize drought stress program using factorial regression and partial least squares methods. Theoretical and Applied Genetics, 112(6), 1009-1023.
- Verdun, R. E., Di Paolo, N., Urmenyi, T. P., Rondinelli, E., Frasch, A. C., & Sanchez, D. O. (1998). Gene discovery through expressed sequence tag sequencing in *Trypanosoma cruzi*. Infection and immunity, 66(11), 5393-5398.
- Vij, S., & Tyagi, A. K. (2008). A20/AN1 zinc-finger domain-containing proteins in plants and animals represent common elements in stress response. Functional & integrative genomics, 8(3), 301-307.
- Vo, K. T. X., Kim, C. Y., Chandran, A. K. N., Jung, K. H., An, G., & Jeon, J. S. (2015). Molecular insights into the function of ankyrin proteins in plants. Journal of Plant Biology, 58(5), 271-284.
- Vos, P., Hogers, R., Bleeker, M., Reijans, M., Van de Lee, T., Hornes, M., [...] & Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. Nucleic acids research, 23(21), 4407-4414.
- Voxeur, A., Gilbert, L., Rihouey, C., Driouich, A., Rothan, C., Baldet, P., & Lerouge, P. (2011). Silencing of the GDP-D-mannose 3, 5-epimerase affects the structure and cross-linking of the pectic polysaccharide rhamnogalacturonan II and plant growth in tomato. Journal of Biological Chemistry, 286(10), 8014-8020.
- Wahl, V., Ponnu, J., Schlereth, A., Arrivault, S., Langenecker, T., Franke, A., [...] & Schmid, M. (2013). Regulation of flowering by trehalose-6-phosphate signaling in Arabidopsis thaliana. Science, 339(6120), 704-707.
- Walsh, P. S., Erlich, H. A., & Higuchi, R. (1992). Preferential PCR amplification of alleles: mechanisms and solutions. Genome Research, 1(4), 241-250.
- Wang, A., Tan, D., Takahashi, A., Li, T. Z., & Harada, T. (2007). MdERFs, two ethylene-response factors involved in apple fruit ripening. Journal of Experimental Botany, 58(13), 3743-3748.
- Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., [...] & Kruglyak, L. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science, 280(5366), 1077-1082.
- Wang, L., Zhou, B., Wu, L., Guo, B., & Jiang, T. (2011). Differentially expressed genes in *Populus simonii*× *Populus nigra* in response to NaCl stress using cDNA-AFLP. Plant science, 180(6), 796-801.

- Wang, Y., Wu, X., He, C., Zhang, J., Chen, S., & Gai, J. (2002). A soybean genetic linkage map constructed after the mapping population being tested and adjusted. Zhong guo nongye kexue,36(11), 1254-1260.
- Wang, Y., Yu, B., Zhao, J., Guo, J., Li, Y., Han, S., [...] & Liu, Y. (2013). Autophagy contributes to leaf starch degradation. The Plant Cell, 25(4), 1383-1399.
- Waples, R. S. (2014). Testing for Hardy–Weinberg proportions: have we lost the plot?. Journal of Heredity, esu062.
- Wasteneys, G. O. (2004). Progress in understanding the role of microtubules in plant cells. Current opinion in plant biology, 7(6), 651-660.
- Watanabe, K., Suzuki, K., & Kitamura, S. (2006). Characterization of a GDP-d-mannose 3 ", 5 "-epimerase from rice. Phytochemistry, 67(4), 338-346.
- Waters, A. J., Makarevitch, I., Eichten, S. R., Swanson-Wagner, R. A., Yeh, C. T., Xu, W., [...] & Springer, N. M. (2011). Parent-of-origin effects on gene expression and DNA methylation in the maize endosperm. The Plant Cell,23(12), 4221-4233.
- Watkins, P. A., & Ellis, J. M. (2012). Peroxisomal acyl-CoA synthetases. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1822(9), 1411-1420.
- Wegrzyn, J. L., Eckert, A. J., Choi, M., Lee, J. M., Stanton, B. J., Sykes, R., [...] & Neale, D. B. (2010). Association genetics of traits controlling lignin and cellulose biosynthesis in black cottonwood (*Populus trichocarpa*, Salicaceae) secondary xylem. New Phytologist, 188(2), 515-532.
- Welsh, J., & McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. Nucleic acids research, 18(24), 7213-7218.
- Werner, J. D., Borevitz, J. O., Warthmann, N., Trainer, G. T., Ecker, J. R., Chory, J., & Weigel, D. (2005). Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proceedings of the National Academy of Sciences of the United States of America, 102(7), 2460-2465.
- Whitmore, T.C. (1973) The palms of Malaya. Longmans, Malaysia, pp 56–58
- Wiertz, E. J. H. J., Tortorella, D., Bogyo, M., Yu, J., Mothes, W., Jones, T. R., [...] & Ploegh, H. L. (1996). Sec61-mediated transfer of a membrane protein from the endoplasmic reticulum to the proteasome for destruction. Nature,384(6608), 432-438.
- Williams, J. G., Kubelik, A. R., Livak, K. J., Rafalski, J. A., & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic acids research, 18(22), 6531-6535.
- Wingler, A., Lea, P. J., Quick, W. P., & Leegood, R. C. (2000). Photorespiration: metabolic pathways and their role in stress protection. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 355(1402), 1517-1529.
- Winter, P., & Kahl, G. (1995). Molecular marker technologies for plant improvement. World Journal of Microbiology and Biotechnology, 11(4), 438-448.
- Wirtz, M., Droux, M., & Hell, R. (2004). O-acetylserine (thiol) lyase: an enigmatic enzyme of plant cysteine biosynthesis revisited in *Arabidopsis thaliana*. Journal of Experimental Botany, 55(404), 1785-1798.
- Wolucka, B. A., & Van Montagu, M. (2003). GDP-mannose 3', 5'-epimerase forms GDP-L-gulose, a putative intermediate for the de novo biosynthesis of vitamin C in plants. Journal of Biological Chemistry,278(48), 47483-47490.
- Wong, C. K., & Bernardo, R. (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics, 116(6), 815-824.

- Worrall, D., Liang, Y. K. ,Alvarez, S., Holroyd, G. H.,Spiegel, S., Panagopulos, M.,Gray, J. E., Hetherington, A. M.(2008). Involvement of sphingosine kinase in plant cell signalling. Plant Journal 56: 64–72.
- Wright, S. (1951). The genetical structure of populations. Annals of eugenics, 15(1), 323-354.
- Wu, X. L., He, C. Y., Wang, Y. J., Zhang, Z. Y., Dongfang, Y., Zhang, J. S., [...] & Gai, J. Y. (2001). Construction and analysis of a genetic linkage map of soybean. Yi chuan xue bao= Acta genetica Sinica, 28(11), 1051-1061.
- Xia, W., Mason, A. S., Xiao, Y., Liu, Z., Yang, Y., Lei, X., [...] & Peng, M. (2014). Analysis of multiple transcriptomes of the African oil palm (*Elaeis guineensis*) to identify reference genes for RT-qPCR. Journal of biotechnology, 184, 63-73.
- Xiao, Y. Y., Chen, J. Y., Kuang, J. F., Shan, W., Xie, H., Jiang, Y. M., & Lu, W. J. (2013). Banana ethylene response factors are involved in fruit ripening through their interactions with ethylene biosynthesis genes. Journal of experimental botany, ert108.
- Xie, M., Huang, Y., Zhang, Y., Wang, X., Yang, H., Yu, O., [...] & Fang, C. (2013). Transcriptome profiling of fruit development and maturation in Chinese white pear (*Pyrus bretschneideri* Rehd). BMC genomics,14(1), 1..
- Xin, C., Liu, W., Lin, Q., Zhang, X., Cui, P., Li, F., [...] & Al-Mssallem, I. S. (2015). Profiling microRNA expression during multi-staged date palm (*Phoenix dactylifera* L.) fruit development. Genomics, 105(4), 242-251.
- Xiong, T. C., Coursol, S., Grat, S., Ranjeva, R., Mazars, C.(2008). Sphingolipid metabolites selectively elicit increases in nuclear calcium concentration in cell suspension cultures and in isolated nuclei of tobacco. Cell Calcium 43:29–37.
- Xu, X. M., Wang, J., Xuan, Z., Goldshmidt, A., Borrill, P. G., Hariharan, N., [...] & Jackson, D. (2011). Chaperonins facilitate KNOTTED1 cell-to-cell trafficking and stem cell function. Science, 333(6046), 1141-1144.
- Yang, S. L., Xie, L. F., Mao, H. Z., San Puah, C., Yang, W. C., Jiang, L., [...] & Ye, D. (2003). Tapetum determinant1 is required for cell specialization in the Arabidopsis anther. The Plant Cell, 15(12), 2792-2804.
- Yang, S. L., Jiang, L., San Puah, C., Xie, L. F., Zhang, X. Q., Chen, L. Q., [...] & Ye, D. (2005). Overexpression of TAPETUM DETERMINANT1 alters the cell fates in the Arabidopsis carpel and tapetum via genetic interaction with EXCESS MICROSPOROCYTES1/EXTRA SPOROGENOUS CELLS. Plant Physiology, 139(1), 186-191.
- Yang, X. H., Xu, Z. H., & Xue, H. W. (2005). Arabidopsis membrane steroid binding protein 1 is involved in inhibition of cell elongation. The Plant Cell, 17(1), 116-131.
- Yao, Y. X., Li, M., Liu, Z., Hao, Y. J., & Zhai, H. (2007). A novel gene, screened by cDNA-AFLP approach, contributes to lowering the acidity of fruit in apple. Plant Physiology and Biochemistry, 45(2), 139-145.
- Yelina, N. E., Smith, L. M., Jones, A. M., Patel, K., Kelly, K. A., & Baulcombe, D. C. (2010). Putative Arabidopsis THO/TREX mRNA export complex is involved in transgene and endogenous siRNA biosynthesis. Proceedings of the National Academy of Sciences, 107(31), 13948-13953.
- Yin, J., Wang, G., Xiao, J., Ma, F., Zhang, H., Sun, Y., [...] & Liu, D. (2010). Identification of genes involved in stem rust resistance from wheat mutant D51 with the cDNA-AFLP technique. Molecular biology reports, 37(2), 1111.
- Yin, X. J., Volk, S., Ljung, K., Mehlmer, N., Dolezal, K., Ditengou, F., [...] & Teige, M. (2007). Ubiquitin Lysine 63 Chain–Forming Ligases Regulate Apical Dominance in Arabidopsis. The Plant Cell, 19(6), 1898-1911.

- Yin, X. R., Allan, A. C., Chen, K. S., & Ferguson, I. B. (2010). Kiwifruit EIL and ERF genes involved in regulating fruit ripening. Plant Physiology, 153(3), 1280-1292.
- Yin, Y., Zhang, X., Fang, Y., Pan, L., Sun, G., Xin, C., [...] & Yu, J. (2012). High-throughput sequencing-based gene profiling on multi-staged fruit development of date palm (*Phoenix dactylifera*, L.). Plant molecular biology,78(6), 617-626.
- Yoo, J., Lee, Y., Kim, Y., Rha, S. Y., & Kim, Y. (2008). SNPAnalyzer 2.0: a web-based integrated workbench for linkage disequilibrium analysis and association analysis. BMC bioinformatics, 9(1), 290.
- Young, N. D. (1996). QTL mapping and quantitative disease resistance in plants. Annual review of phytopathology, 34(1), 479-501.
- Yu, J., & Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. Current Opinion in Biotechnology, 17(2), 155-160.
- Yu, S., Liao, F., Wang, F., Wen, W., Li, J., Mei, H., & Luo, L. (2012). Identification of rice transcription factors associated with drought tolerance using the ecotilling method. PLoS One, 7(2), e30765.
- Zechner, R., Zimmermann, R., Eichmann, T. O., Kohlwein, S. D., Haemmerle, G., Lass, A., & Madeo, F. (2012). FAT SIGNALS-lipases and lipolysis in lipid metabolism and signaling. Cell metabolism, 15(3), 279-291.
- Zegzouti, H., Jones, B., Frasse, P., Marty, C., Maitre, B., Latché, A., [...] & Bouzayen, M. (1999). Ethylene-regulated gene expression in tomato fruit: characterization of novel ethylene-responsive and ripening-related genes isolated by differential display. The Plant Journal, 18(6), 589-600.
- Zeng, F., Jiang, R., & Chen, T. (2013). PyroHMMsnp: an SNP caller for Ion Torrent and 454 sequencing data. Nucleic acids research, 41(13), e136-e136.
- Zeven, A. C. (1967). The semi-wild oil palm and its industry in Africa. Pudoc, Centre for Agricultural Publications and Documentation.
- Zhang, C., Shi, H., Chen, L., Wang, X., Lü, B., Zhang, S., [...] & You, Z. (2011). Harpin-induced expression and transgenic overexpression of the phloem protein gene AtPP2-A1 in Arabidopsis repress phloem feeding of the green peach aphid Myzus persicae. BMC plant biology, 11(1), 1.
- Zhang, C., Zhang, H., Zhao, Y., Jiang, H., Zhu, S., Cheng, B., & Xiang, Y. (2013). Genome-wide analysis of the CCCH zinc finger gene family in *Medicago truncatula*. Plant cell reports, 32(10), 1543-1555.
- Zhang, J., & Erickson, L. R. (2012). Harvest-inducibility of the promoter of alfalfa S-adenosyl-L-methionine: trans-Caffeoyl-CoA3-O-methyltransferase gene. Molecular biology reports, 39(3), 2489-2495.
- Zhao, G. R., & Liu, J. Y. (2002). Isolation of a cotton RGP gene: a homolog of reversibly glycosylated polypeptide highly expressed during fiber development. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression, 1574(3), 370-374.
- Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., [...] & McClung, A. M. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nature communications, 2, 467.
- Zheng, P., Allen, W. B., Roesler, K., Williams, M. E., Zhang, S., Li, J., [...] & Bhattramakki, D. (2008). A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nature genetics, 40(3), 367-372.
- Zhu, C., Gore, M., Buckler, E. S., & Yu, J. (2008). Status and prospects of association mapping in plants. The plant genome, 1(1), 5-20.
- Zhu, J., Chen, H., Li, H., Gao, J. F., Jiang, H., Wang, C., [...] & Yang, Z. N. (2008). Defective in Tapetal development and function 1 is essential for anther development and

tapetal function for microspore maturation in Arabidopsis. The Plant Journal, 55(2), 266-277.

- Zhu, M., & Zhao, S. (2007). Candidate gene identification approach: progress and challenges. International Journal of Biological Sciences, 3(7), 420-427.
- Zhu, M., Chen, G., Dong, T., Wang, L., Zhang, J., Zhao, Z., & Hu, Z. (2015). SlDEAD31, a Putative DEAD-Box RNA Helicase Gene, Regulates Salt and Drought Tolerance and Stress-Related Genes in Tomato. PloS one, 10(8), e0133849.

Páginas web visitadas

- http://www.ncbi.nlm.nih.gov/nucest visitada 8 de enero de 2016
- http://www.uniprot.org/uniprot/Q9LPE8 visitada 15 de febrero de 2016
- http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=239454 visitada 23 de febrero de 2016
- https://www.ebi.ac.uk/interpro/entry/IPR024709 visitada 26 de febrero de 2016
- http://www.uniprot.org/uniprot/P47669 visitada 26 de febrero de 2016
- http://www.uniprot.org/uniprot/Q9ZPR1 visitada 26 de febrero de 2016
- http://www.uniprot.org/uniprot/Q84P97 visitada 29 de febrero de 2016
- https://www.ebi.ac.uk/interpro/entry/IPR024929 visitada 29 de febrero de 2016
- http://www.uniprot.org/uniprot/Q9LN49 visitada 3 de marzo de 2016
- http://www.uniprot.org/uniprot/Q40704 visitada 8 de marzo de 2016
- http://www.uniprot.org/uniprot/Q2R3K3 visitada 9 de marzo de 2016
- http://www.uniprot.org/uniprot/Q42134 visitada 9 de marzo de 2016
- http://www.uniprot.org/uniprot/Q9LM02 visitada 9 de marzo de 2016
- http://www.ebi.ac.uk/interpro/entry/IPR016040 visitado 10 de marzo de 2016
- http://www.uniprot.org/uniprot/O48844 visitada 14 de marzo de 2016
- http://www.uniprot.org/uniprot/Q9FJR1 visitada 16 de marzo de 2016
- http://www.ncbi.nlm.nih.gov/gene/?term=xp_010928199 visitada 17 de marzo de 2016
- http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=271593 visitada 17 de marzo de 2016
- http://www.uniprot.org/uniprot/Q8L7Q0 visitada 17 de marzo de 2016
- http://www.uniprot.org/uniprot/Q93VM9 visitada 17 de marzo de 2016
- http://www.uniprot.org/uniprot/B3GS44 visitada 18 de marzo de 2016
- https://bitbucket.org/tasseladmin/tassel-5-source/wiki/UserManual visitada 5 diciembre 2016
- https://www.thermofisher.com/es/es/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-ion-proton-system-chips.html visitada 11 agosto 2016
- https://www.fas.usda.gov. United StatesDepartment of Agriculture (2017). Oilseeds: World Markets and Trade. Visitado 20 de febrero de 2017
- http://www.targetmap.com/viewer.aspx?reportId=17662.Targetmap (s.f).[Top five oil crops en targetmap]. Visitada 25 de febrero de 2017.