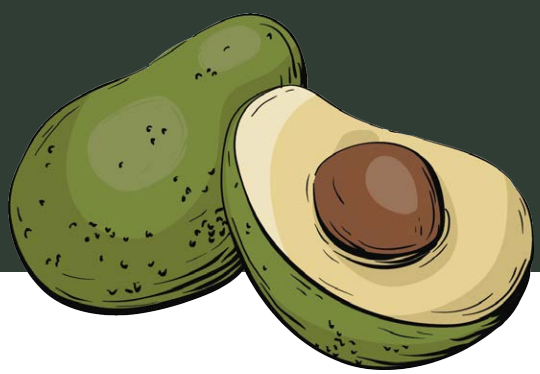
Tesis doctoral

Desarrollo de herramientas genómicas en frutales subtropicales: aguacate y chirimoyo

Alicia Talavera Júdez

Directores de tesis:
Dr. José Ignacio Hormaza Urroz y Dr. Antonio J. Matas Arroyo



Málaga, 2020

Programa de doctorado: Biología Celular y Molecular

AUTOR: Alicia Talavera Júdez

http://orcid.org/0000-0002-1285-7891

EDITA: Publicaciones y Divulgación Científica. Universidad de Málaga

Tesis Doctoral

Desarrollo de herramientas genómicas en frutales subtropicales: aguacate y chirimoyo

Memoria de tesis doctoral presentada por Alicia Talavera Júdez para optar al grado de Doctora por la Universidad de Málaga, programa de Doctorado "Biología Celular y Molecular".

Directores:

Dr. José Ignacio Hormaza Urroz, Profesor de Investigación
Dr. Antonio Javier Matas Arroyo

Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora"
(IHSM-UMA-CSIC)

Málaga, Junio de 2020

Instituto de Hortofruticultura Subtropical y Mediterránea la Mayora
Avda. Dr. Wienberg s/n
29750 Algarrobo Costa (Málaga)
Tel.: 34-952548990 ext. 147
http.//www.eelm.csic.es
email: ihormaza@eelm.csic.es

D. José Ignacio Hormaza Urroz, Profesor de Investigación del Consejo Superior de Investigaciones Científicas en el Instituto de Hortofruticultura Subtropical y Mediterránea “La Mayora” y director del Departamento de Fruticultura,

CERTIFICA

Que Alicia Talavera Júdez, “Licenciada en Biología”, ha realizado en el Departamento de Fruticultura del Instituto de Hortofruticultura Subtropical y Mediterránea “La Mayora” (IHSM-UMA-CSIC), bajo su dirección, el trabajo de investigación que con el título “Desarrollo de herramientas genómicas en frutales subtropicales: aguacate y chirimoyo” presenta para optar al grado de Doctora, y en cumplimiento de la legislación vigente

AUTORIZA

Su lectura en la Universidad de Málaga.

En Algarrobo-Costa a 10 de junio de 2020.

Fdo. Dr. José Ignacio Hormaza

D. Antonio Javier Matas Arroyo, Profesor Titular en el Departamento de Botánica y Fisiología Vegetal en la Facultad de Ciencias de la Universidad de Málaga

CERTIFICA:

Que Alicia Talavera Júdez, "Licenciada en Biología", ha realizado en el Departamento de Fruticultura del Instituto de Hortofruticultura Subtropical y Mediterránea "La Mayora" (IHSM-UMA-CSIC), bajo su dirección, el trabajo de investigación que con el título "Desarrollo de herramientas genómicas en frutales subtropicales: aguacate y chirimoyo" presenta para optar al grado de Doctora, y en cumplimiento de la legislación vigente

AUTORIZA:

Su lectura en la Universidad de Málaga.

Junio de 2020.

Firmado por MATAS ARROYO ANTONIO JAVIER - el día 11/06/2020 con un certificado emitido por AC FNMT Usuarios

Fdo. Dr. Antonio J. Matas Arroyo

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña ALICIA TALAVERA JÚDEZ
Estudiante del programa de doctorado BIOLOGÍA CELULAR Y MOLECULAR de la Universidad de Málaga, autor/a de la tesis, presentada para la obtención del título de doctor por la Universidad de Málaga, titulada: DESARROLLO DE HERRAMIENTAS GENÓMICAS EN FRUTALES SUBTROPICALES: AGUACATE Y CHIRIMOYO.

Realizada bajo la tutorización de JOSÉ IGNACIO HORMAZA URROZ y dirección de ANTONIO JAVIER MATAS ARROYO (si tuviera varios directores deberá hacer constar el nombre de todos)

DECLARO QUE:

La tesis presentada es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, conforme al ordenamiento jurídico vigente (Real Decreto Legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), modificado por la Ley 2/2019, de 1 de marzo.

Igualmente asumo, ante a la Universidad de Málaga y ante cualquier otra instancia, la responsabilidad que pudiera derivarse en caso de plagio de contenidos en la tesis presentada, conforme al ordenamiento jurídico vigente.

En Málaga, a 9 de JUNIO de 2020

Fdo.:

Edificio Pabellón de Gobierno. Campus El Ejido. 29071
Tel.: 952 13 10 28 / 952 13 14 61 / 952 13 71 10
E-mail: doctorado@uma.es

Este trabajo se ha realizado en el Departamento de Fruticultura del Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM-UMA-CSIC) La Mayora, gracias a la financiación de una ayuda predoctoral para la formación de personal investigador (FPI) (BES-2014-068832) del Ministerio de Economía y Competitividad de España y dos ayudas para la realización de estancias breves en centros I+D (EEBB-I-17-12076 y EEBB-I-18-12852), asociadas al proyecto (AGL2013-43732-R).

A mis padres
A Arlo

Agradecimientos

Esta etapa no hubiera sido tan especial sin toda la gente que he tenido a mi alrededor. De verdad, sin vosotros todo esto no hubiese sido posible.

Quisiera agradecer a mi director Iñaki Hormaza, por haberme dado esta gran oportunidad (que sinceramente me ha cambiado la vida). A pesar de estar siempre viajando, y con una agenda repleta de eventos, reuniones y visitas, siempre ha estado ahí. Gracias por todo lo que me ha enseñado, por su apoyo, confianza, paciencia y dedicación.

Gracias a mi director Antonio J. Matas por introducirme en el mundo de R, por su paciencia, consejos y ánimos en numerosas ocasiones. Por mantener y darme acceso al servidor que tanto he utilizado. Por lo que me ha enseñado, por hacer hincapié sobre los pequeños detalles, por resaltar siempre las cosas bien hechas, y por contar conmigo para los eventos de divulgación en los que ha participado.

Otra persona fundamental en esta etapa, y de la que me siento profundamente agradecida es Aureliano Bombarely (mi tercer director). Gracias por todo lo que me ha enseñado, por las reuniones, por su apoyo, paciencia y consejos; por mantener el servidor de Virginia, y ahora el de Milán. Por contestar tan rápido a los correos. ¡Gracias por todas las aventuras relacionadas con el pawpaw!. Y porque sin él, este proyecto no hubiera sido posible.

Aunque en muchas ocasiones ha sido complicado trabajar con tres directores (no lo voy a negar), ¡Me siento muy afortunada por haber contado con ellos!. Y me gustaría agradecerles el entusiasmo que transmiten con la ciencia.

Gracias por todo a los fruitis. Habéis sido un gran pilar en este recorrido. A Vero y Nerea, que a pesar de haber compartido menos tiempo con ellas siempre me han ayudado y apoyado. A Juan Losada por su simpatía y consejos. A Miguel por el ánimo y por esos suministros de comida que tanta energía me han dado en los últimos meses. A Jorge Lora por aguantar mil preguntas en el laboratorio, por sus consejos, aportaciones y por estar siempre ahí. A Tina por escucharme y ser un gran apoyo en los últimos momentos. Pero sin duda, a las que tengo que agradecer muchísimo es a Yolanda y a Librada. Gracias Yoli por ser la alegría del laboratorio, qué sería de nosotros sin sus historias, risas y ayuda. Gracias a Librada, por sus consejos, por su ayuda desde el comienzo hasta el fin, por sus bromas, por escucharme y comprenderme en muchas ocasiones. Desde luego sin su cariño y apoyo este periodo hubiese sido mucho más complicado, ¡Muchas gracias!.

También agradecer a los que todavía no he nombrado, y con los que he compartido tantos momentos en el cuartito. A Isa, por ser "abogada del diablo", por su comprensión, por escucharme, por su compañía, sus consejos, apoyo, cariño y ayuda, ¡Gracias!. A Paco, a Anabel y a Gloria. A Sara (de Zaragoza) por su apoyo y los buenos momentos que hemos vivido. A Alejandra, por el cariño y por ser la mejor compañera de carreras; y a Ángel, por esa alegría que transmitió en los pocos meses que estuvo en "La Mayora".

Gracias al departamento de Fruticultura, en especial a Sonia Cívico por su cariño y apoyo cuando trabajamos con el citómetro de flujo. También gracias a Jose y a Sonia Ruiz, por acompañarme en mis primeros días de campo. A Ruth por estar pendiente de los chirimoyos y, al igual que Sara, ayudarme a localizar aguacates para mi video del CSIC "yo investigo".

Igualmente me gustaría agradecer a todas las personas con las que he compartido muy buenos momentos: Rida, Elisa, Lidia (¡Ya no os queda nada a vosotras!), Olaya, Diego, David (que sigue dando su apoyo desde la distancia), Efren, Dani, Meche, Cristina, Cañi, Reme, Ana Cris e Irene (cuánto me alegro de haberte encontrado de nuevo). ¡Muchas gracias!.
Gracias por la ayuda y el apoyo de Mariola, Juan, Fali, Jesús y Elvira. Por las visitas de la tarde y la alegría de Mari y Adela. Gracias por la ayuda de los conserjes. Gracias al personal de los departamentos de Informática, Virología, Micología y Mejora, a la biblioteca (a Jero y a Juan), al personal de administración, en especial a Antoñita, que tanto se preocupaba por que terminara la tesis. A Antonio Cordón por su disposición y ayuda. Gracias al personal del IHSM-UMA-CSIC de Málaga, sobre todo a Gloria por las veces que me ha echado una mano.

Muchas gracias a todos, porque hacéis que trabajar aquí sea una gozada.

Gracias a las personas que me han acompañado y ayudado en mis estancias en Virginia Tech. Al grupo de "Pawpaw Hunters", especialmente a Lisa, a Silvia, a Elijah y a James. No me olvido de mis compañeros de "laboratorio" (Tomas, Andie, Haidong, Chenming y Ariel) gracias por la ayuda, la amistad y el intercambio cultural; gracias a mi amiga Unnati, por la ayuda, las risas, las nuevas recetas y el cariño.

Me gustaría dar las gracias también al CSIC, a la beca del Ministerio de Economía y Competitividad que me fue concedida. Al igual que a todos los organismos públicos donde he recibido formación previa, y a todas aquellas personas que comparten su conocimiento bioinformático en las redes.

Gracias a mi familia, porque siempre está ahí. Gracias Papá (Abel) por tu apoyo incondicional, por inculcarme la importancia de seguir aprendiendo y de ser valiente. Gracias Mamá (Maribel), por enseñarme que con constancia todo lo que uno se propone se puede conseguir y, por supuesto, por tu apoyo y cariño día tras día. Ambos sois un ejemplo de superación, y sin vuestra ayuda todo lo que he conseguido hasta ahora hubiese sido imposible. No os lo digo demasiado, pero …¡Os quiero!.
Gracias a Arlo, por toda la fuerza, alegría, tranquilidad, apoyo y cariño que me da cada día (tanto en la distancia como en la cercanía) y, sobre todo, gracias por creer siempre en mi. También a mi segunda familia, especialmente a Nieves y a Dani, por el ánimo, apoyo y cariño y, por supuesto, gracias a Pili por su cariño, y por ser la mayor admiradora de la chirimoya.

Por último, me gustaría agradecer a mis amigos por el ánimo y ayuda en la distancia. A mi grupo de "Bioremember", especialmente a David Herrero, Nacho, Sandra, Marta, Alberto Sánchez y Javi. También a mis compis del G-27, mi Iya (Rocio) y Ari (Ariana). A Rumi (Paloma Soler) y Paula (quien me ha ayudado y aconsejado con la portada). A pesar de verlas cada vez menos, sé que puedo contar con ellas. Por otro lado, y no por ello menos importante, gracias a Paloma Goñi, por estar siempre ahí, en lo bueno y en lo malo.

En definitiva, ¡Muchas gracias a todos!.

ÍNDICE

Resumen 1

Summary 3

Introducción general 5

1. El material vegetal. Aguacate y chirimoyo 7

Aguacate (*Persea americana* Mill.) 7

Taxonomía, origen y distribución geográfica 7

Descripción botánica 9

Usos e importancia económica 11

Chirimoyo (*Annona cherimola* Mill.) 12

Taxonomía, origen y distribución geográfica 12

Descripción botánica 13

Usos e importancia económica 15

2. Aproximaciones moleculares y sus necesidades en *Persea americana* y *Annona cherimola* 16

Marcadores moleculares 17

Mapas de ligamiento 19

3. Aproximaciones genómicas 21

Aproximaciones de genotipado 23

Genomas de referencia 24

Objetivos 29

Capítulo 1: Caracterización genómica de aguacate (*Persea americana* Mill.) 33

Resumen 35

Introducción 35

Material y métodos 38
Resultados 46
Discusión 62

Capítulo 2: Secuenciación, ensamblado y anotación del genoma del chirimoyo (*Annona cherimola* Mill.) 67

Resumen 69
Introducción 69
Material y métodos 72
Resultados 77
Discusión 89

Capítulo 3: Elaboración de un mapa genético en el género *Annona* 95

Resumen 97
Introducción 97
Material y métodos 100
Resultados 103
Discusión 108

Discusión general 113

Conclusiones 123

Bibliografía 127

Anexo 1. Genome-Wide SNP discovery and genomic characterization in avocado (*Persea americana* Mill.) 157

Anexo 1.1. Protocolo para la construcción de genotecas para el genotipado por secuenciación (GBS) 173

Resumen

El aguacate (*Persea americana* Mill.) y el chirimoyo (*Annona cherimo*la Mill.) son dos cultivos nativos de los Neotrópicos muy apreciados desde tiempos precolombinos. Además de su interés agrícola, tienen el valor añadido de pertenecer a uno de los grupos más primitivos de las angiospermas, por lo que son de gran importancia para estudios evolutivos en plantas. No obstante, la información molecular generada en estas especies es escasa, lo que dificulta su estudio y la disponibilidad de herramientas para optimizar el proceso de mejora. Con el objetivo de reducir esta brecha e incrementar los recursos genómicos disponibles en estas especies se ha llevado a cabo este estudio, que se divide en tres objetivos: (1) desarrollar marcadores moleculares SNPs mediante GBS para la caracterización de genotipos de aguacate (*Persea americana* Mill.), (2) ensamblar y anotar el primer genoma del chirimoyo (*Annona cherimola* Mill.) y (3) elaborar un mapa de ligamiento en el género *Annona* a partir de una población F2 generada a partir de un cruzamiento interespecífico (*Annona cherimola* 'Fino de Jete' *x Annona squamosa* 'Thai seedless').

El desarrollo de un borrador de genoma de aguacate (cv. Hass) junto a la secuenciación de genotecas mediante GBS han permitido detectar un total de 7.108 SNPs, los cuales han facilitado la caracterización de 71 genotipos que representan las 3 razas botánicas (Mexicana, Guatemalteca y Antillana) de esta especie. Además, estos marcadores moleculares han facilitado la caracterización de la diversidad genética y la estructura poblacional, mostrando claras agrupaciones basadas en el origen racial.

Por otro lado, se ha generado el primer ensamblaje y anotación *de novo* del genoma del chirimoyo (cv. Fino de Jete), siendo el primer genoma dentro de la familia Annonaceae. Este genoma está formado por un total de 15.076 secuencias y un N50 de 171,2 Kb. Aproximadamente, el 67 % de las secuencias son repetitivas, y los genes predichos se encuentran asociados a un evento de hibridación o duplicación reciente, que podría estar generando cierta inestabilidad cromosomal.

Finalmente, se ha desarrollado por primera vez un mapa genético para el género *Annona* a partir de una población F2, generada del autocruzamiento de individuos de una F1 realizada a partir de un cruce interespecífico entre 'Fino de Jete' (*Annona cherimola*) y 'Thai seedless'

(*Annona squamosa*). Se empleó un total de 550 SNPs para la construcción del mapa genético, que se distribuyeron en 8 grupos de ligamiento. El mapa generado cubre aproximadamente 1.388 cM con una distancia media entre marcadores de 2,6 cM, tratándose del primer mapa generado para estas especies (*A. cherimola* y *A. squamosa*).

Los resultados obtenidos en este trabajo aportan nuevos conocimientos sobre aguacate y chirimoyo que son esenciales para la optimización de programas de mejora, pero también abren las puertas a futuros trabajos de estudio de la diversidad genética, conservación o evolución de las angiospermas.

Summary

The avocado (*Persea americana* Mill.) and the cherimoya (*Annona cherimola* Mill.) are two perennial fruit crops native of the Neotropics appreciated since pre-Columbian times. In addition to their agricultural interest, both species belong to a clade of early-divergent angiosperms, thus being very relevant for evolutionary studies. However, no previous significant genomic information is available in these species, complicating the development of tools to optimize breeding processes. In order to fill this gap, the main goal of this work is to increase the genomic resources available for these species. This is accomplished within three objectives: (1) to develop SNP molecular markers using GBS in order to characterize the genetic diversity of avocado (*Persea americana* Mill.), (2) to assemble and annotate the first cherimoya (*Annona cherimola* Mill.) genome, (3) to develop a genetic map from an F2 population developed from the interspecific cross [‘Fino de Jete’ (*Annona cherimola*) x ‘Thai seedless’ (*Annona squamosa*)].

The development of an avocado (cv. Hass) draft genome in addition to the sequencing of a GBS library allowed the production of 7,108 SNPs from 71 accessions, which represent the three traditionally recognized avocado horticultural races (Mexican, Guatemalan and West Indian). These molecular markers have allowed the study of the genetic diversity and the population structure grouping the cultivars studied according to their race.

On the other hand, a first assembly and annotation of a cherimoya genome (cv. Fino de Jete) was developed. The resulting assembly contained 15,076 contigs, with an N50 of 171.2 Kb. Approximately 67 % of the genome was identified as repetitive DNA, and predicted genes revealed a recent whole-genome duplication or hybridization event that could be related with a possible chromosome instability. After this, a genetic map was constructed using a F2 population from self-pollination of F1 individuals from an interspecific cross between ‘Fino de Jete’ (*Annona cherimola* Mill.) and ‘Thai seedless’ (*Annona squamosa* L.). The genetic map predicted eight linkage groups that included 550 SNPs. It covered 1,388 cM with a 2.6 cM average distance between adjacent markers. To date, this is the first genomic map of these species (*A. cherimola* and *A. squamosa*).

The results of this study provide novel information on avocado and cherimoya, that could help to optimize breeding programs, germplasm management and, in general, a better characterization of these crops. Furthermore, these results provide important resources for future works dealing with evolutionary and diversity conservation studies.

INTRODUCCIÓN GENERAL

Introducción general

1. El material vegetal. Aguacate y chirimoyo

Este trabajo se ha llevado a cabo en dos cultivos de reciente introducción comercial en España y originarios de los Neotrópicos, el aguacate (*Persea americana* Mill.) y el chirimoyo (*Annona cherimola* Mill.). El aguacate es miembro de las Lauráceas en el orden Laurales (3.874 especies) y el chirimoyo pertenece a las Annonáceas en el orden Magnoliales (3.140 especies) (APG IV 2016). Ambos órdenes junto a los órdenes Canellales (123 especies) y Piperales (3.190 especies) forman el clado Magnoliid (Zanis *et al.* 2002), dentro de las angiospermas basales (Soltis *et al*. 2005) que, después del clado de eudicotiledóneas y monocotiledóneas, es el clado más numeroso de las angiospermas (Massoni *et al*. 2015).

1.1. Aguacate (*Persea americana* Mill.)

Taxonomía, origen y distribución geográfica

Persea americana Mill. pertenece a una familia pantropical (Lauraceae) incluida en el orden Laurales. Este se compone por 7 familias (Atherospermataceae, Calycanthaceae, Gomortegaceae, Hernandiaceae, Lauraceae, Monimiaceae y Sipuranaceae) siendo la familia Lauraceae la más amplia del orden (Buzgo *et al.* 2007). Incluye aproximadamente 50 géneros y entre 2.500 y 3.000 especies de árboles y lianas, existiendo también arbustos y un género de parras parásitas, que se distribuyen en áreas tropicales y subtropicales (Chanderbali *et al*. 2008). Posiblemente debido a su notable variabilidad fenotípica, la taxonomía de la familia es complicada y sufre frecuentes cambios (Boza *et al.* 2018).

Dentro del orden Laurales se encuentran numerosas especies de interés económico, como el laurel (*Laurus nobilis* L.), la canela (*Cinnamomum zeylanicum* Breyn.), el alcanfor (*Cinnamomum camphora* L. J.Presl) y árboles madereros (*Nectandra* spp., *Ocotea* spp. y *Phoebe* spp.), aunque el aguacate destaca por su importancia económica.

En el género *Persea* se han reconocido dos subgéneros: *Persea* y *Eriodaphne* (Kopp 1966). En el subgénero *Persea* se han diferenciado 3 especies (Chanderbali *et al.* 2013): *P. schiedeana*, *P. parviflora* y *P. americana*. Dentro de esta última, Scora *et al.* (2002) propuso

8 variedades botánicas, de las que 5 carecen de importancia comercial (*P. floccosa, P. nubigena, P. steyermarkiana, P. tolimanensis y P. zentmyerii*). Dentro del aguacate cultivado, *P. americana*, se distinguen tradicionalmente tres razas o tipos botánicos en base a diferencias morfológicas, ecológicas y genéticas: Mexicana (*P. americana* var. *drymifolia*), Guatemalteca (*P.americana* var. *guatemalensis*) y Antillana (*P. americana* var. *americana*) (Boza *et al*. 2018).

Según Chanderbali *et al.* (2008) la mayoría de las especies pertenecientes a la familia Lauraceae, como es el caso del aguacate, presentan un genoma diploide (2n=2x=24) y 12 cromosomas, siendo este número el más común y el más bajo dentro de este grupo. Se ha sugerido que las Lauráceas actuales resultaron de un evento o eventos de poliploidía, hace al menos 100 millones de años, con una segunda poliploidía posterior (Schaffer *et al.* 2013). De hecho, análisis transcriptómicos indican la existencia de dos rondas de duplicación en los ancestros de *Persea* (Chaw *et al.* 2019).

Ciertos registros fósiles evidencian que hace 32 millones de años los ancestros de *Persea americana* migraron desde el Norte de América al sur, hacia áreas subtropicales y tropicales, a causa del cambio de temperatura producido durante el Eoceno-Oligoceno (Bost *et al.* 2013). El modo en el que los ancestros del aguacate se dispersaron es incierto, aunque la teoría de que existe una coevolución de la megafauna como dispersores de *P. americana* cada vez ha sido más aceptada, al igual que ha ocurrido para otras frutas tropicales (Schaffer *et al.* 2013; van Zonneveld *et al.* 2018). En Mesoamérica, asociada a la colonización por parte de los humanos, se produjo una extinción masiva de esta megafauna (mastodontes, perezosos gigantes, mamuts etc.) por lo que los seres humanos fueron los responsables de mantener la diversidad genética de estos frutales (van Zonneveld *et al*. 2018).

Actualmente, el origen del aguacate se sitúa en una región geográfica amplia, que se extiende desde el Este hasta las tierras altas de México (alcanzando 3000 m de altitud), Guatemala y la costa del Pacífico de América Central (Tierras bajas tropicales) (Popenoe 1920; Schaffer *et al*. 2013).

Cada una de las razas comentadas anteriormente (Mexicana, Guatemalteca y Antillana), tiene un origen geográfico distinto y, por tanto, se encuentra adaptada a diferentes condiciones climáticas. Los genotipos de raza Mexicana están adaptados a zonas altas de entre 1400 a 2700 metros, la raza Guatemalteca se encuentra en zonas templadas húmedas (1500-2350

metros) mientras que los genotipos de raza antillana se encuentran adaptados a zonas bajas (100-450 metros), húmedas y con altas temperaturas, estando su cultivo limitado por las bajas temperaturas (Wolstenholme 2013). Los cruces interraciales han sido frecuentes a lo largo de la historia evolutiva y la domesticación de esta especie, de hecho, muchos de los cultivares más relevantes económicamente, son híbridos interraciales.

El aguacate posiblemente se trate de uno de los primeros frutales domesticados en la zona neotropical, lo que hace de esta especie un excelente modelo para el estudio del proceso de domesticación de cultivos en esa región (Galindo-Tovar *et al*. 2008). La primera evidencia de consumo de este frutal proviene de unos cotiledones encontrados en Puebla, México, datados en 8000-7000 a. C. (Bost *et al.* 2013).

A pesar de que el aguacate poco a poco fue abriéndose camino por regiones con climas tropicales y subtropicales (Morton 1987), no fue hasta el siglo XX cuando comenzó a producirse un cultivo comercial con diferentes genotipos en distintos países. En España, su primera introducción se sitúa a comienzos del siglo XVI (Popenoe 1963).

Descripción botánica

Persea americana Mill. es un árbol de hoja perenne. Su altura varía según la raza botánica, pudiendo alcanzar desde los 10-15 m (raza Mexicana) a 30 m (raza Guatemalteca y Antillana) (Scora *et al.* 2002). La madera es bastante esponjosa, debido a que las fibras de las paredes son relativamente delgadas y, por tanto, las ramas se doblan con facilidad por el peso de la fruta. Las raíces son poco profundas, y no se extienden más allá de la cobertura del dosel arbóreo (Chanderbali *et al.* 2013).

Las hojas son alternas, con nerviación regular, algo pubescentes y rojizas en estado juvenil, adoptando durante la madurez una forma coriácea, lisa y su color se torna a verde oscuro, con un tamaño de 10 a 30 cm de longitud y de 3 a 19 cm de ancho (Chanderbali *et al.* 2013).

Cada árbol puede producir cientos de inflorescencias y miles de flores. Sin embargo, solo una pequeña proporción de las flores llega a transformarse en fruto a causa de numerosos factores intrínsecos (como el contenido de almidón y carbohidratos solubles) y extrínsecos a la planta (como la temperatura, disponibilidad de agua, nutrientes o polinizadores) (Alcaraz *et al.* 2013). Las flores son hermafroditas, poco llamativas, pequeñas, de color verde amarillento, actinomorfas (con simetría radiada en torno al eje del pedúnculo floral), y forman

agrupaciones en panículas axilares o terminales. Presentan dicogamia (separación temporal entre el estado femenino y el masculino) protogínica (maduración de la parte femenina antes que la masculina). En base a su comportamiento floral podemos diferenciar cultivares de tipo A o tipo B. En los de tipo A las flores se abren por la mañana en estado funcionalmente femenino, al mediodía se cierran, y se abren por la tarde del día siguiente en estado funcionalmente masculino. En los cultivares de tipo B, la flor se abre funcionalmente en estado femenino por la tarde, se cierra por la noche y se abre en estado funcionalmente masculino al día siguiente por la mañana (Alcaraz & Hormaza 2014).

Figura I.1. Flor, fruto y árbol de aguacate de la variedad Hass. Fuente: Librada Alcaraz.

El fruto es una baya con mesocarpo carnoso de color verde que rodea a una gran semilla. Su desarrollo puede durar desde 6 a más de 12 meses, variando según el genotipo y las condiciones de crecimiento (Chanderbali *et al.* 2013). Desde el comienzo de la época de recolección, los frutos pueden permanecer en el árbol de 3 a 6 meses, permitiendo una cosecha escalonada (Kaiser & Wolstenholme 1994) (Figura I.1).

Las características organolépticas de los frutos varían dependiendo de la raza de origen. Los frutos de origen mexicano y guatemalteco contienen una mayor proporción de ácidos grasos

en el mesocarpo maduro (9-30 %), menor contenido de azúcar, y mayor sabor a nuez que los antillanos, con un 2-5 % de ácidos grasos (Schaffer *et al.* 2013).

Usos e importancia económica

El aguacate tiene unas excelentes propiedades organolépticas y nutritivas, con un elevado contenido en carotenoides, hierro y vitaminas E, C, B2, B12, B1, K y D (Ibarra-Laclette *et al.* 2015). De hecho, se ha consumido desde épocas precolombinas posiblemente como un aporte alimentario básico. Actualmente su pulpa se consume fresca, aunque también se utiliza para la elaboración de sopas, batidos, salsas, helados e incluso licores (Bost *et al.* 2013). Sus cualidades fueron destacadas por los primeros cronistas españoles tras la llegada a América (Fernández de Oviedo 1526; de Cieza de León 1532). De hecho, la primera cita del aguacate corresponde a Martín Fernández de Enciso (1519) quien en su obra "Suma de geografía" lo describe "[…] lo que hay dentro es como mantequilla, tiene un sabor delicioso y deja un gusto tan blando y bueno que es algo maravilloso" (Popenoe 1963).

Actualmente, el aguacate se cultiva en aproximadamente 70 países: Estados Unidos, toda América tropical y las islas más grandes del Caribe, Polinesia, Filipinas, África tropical, Australia, Argelia, China, Grecia, Egipto, España (tanto en la península como en las Islas Canarias), Francia, Israel, Madagascar, Madeira, Mauricio, Marruecos, Nueva Zelanda, Palestina, Portugal, Sicilia, Túnez o Turquía (FAO 2018). La producción mundial de aguacate ha sido estimada en aproximadamente 6 millones de toneladas (2018). La producción está concentrada en algunos países (México, República Dominicana, Perú, Indonesia, Colombia y Brasil), siendo México el mayor productor mundial con aproximadamente el 34 % de la producción mundial (más de 2 millones de toneladas) (FAO 2018).

España se considera un caso particular en el cultivo de esta especie, al ser el único país europeo con una producción comercial significativa. Aunque la introducción de este frutal en España tuvo lugar a comienzos del siglo XVI (Popenoe 1963), no fue hasta la segunda mitad del siglo XX, a partir de los años 70, cuando surgieron las primeras plantaciones y su desarrollo comercial (Farré & Pliego 1987). En los últimos años se ha registrado un notable aumento de producción. Esta producción se concentra en la costa Mediterránea andaluza, especialmente en las provincias de Málaga y Granada, y en las Islas Canarias, aunque ha crecido bastante en el Algarve portugués y, en menor medida, en las provincias de Cádiz,

Huelva y la Comunidad Valenciana. Los últimos datos de MAPA contabilizan 14.104 ha con una producción de 97.727 t (MAPA 2019).

1.2 Chirimoyo *(Annona cherimola* Mill.)

Taxonomía, origen y distribución geográfica

La familia Annonaceae contiene alrededor de 107 géneros y 2.400 especies de árboles, arbustos y lianas (Guo *et al*. 2017), con un elevado número de endemismos en distintos continentes (Doyle & Le Thomas 1997). Dentro de los géneros *Asimina* y *Annona*, en el que ha sido incluido el género *Rollinia* (Rainer 2007), encontramos algunas especies que destacan por generar frutos comestibles, incluso algunos con importancia económica: el chirimoyo (*A. cherimola*), anón (*A. squamosa*), atemoyo (híbrido entre *A. cherimola* y *A. squamosa*), guanábano (*A. muricata*), anona (*A. reticulata*), ilama (*A. macroprophyllata*), soncoya (*A. purpurea*) o pawpaw (*Asimina triloba*), encontrándose este último en sus primeras etapas de comercialización siendo la única especie de la familia adaptada a zonas frías (Hormaza 2014; Losada *et al.* 2017).

La mayoría de las especies del género *Annona* presentan un número cromosómico de 2n=2x=14 o 16, con la excepción de *Annona glabra* que es tetraploide 2n=28. En el caso de *Annona cherimola*, como ocurre en general en el género, la determinación del número cromosómico ha sido objeto de debate. Recientemente se ha indicado que, posiblemente, el número de cromosomas del chirimoyo es 2n=2x=14 (Martín *et al.* 2019; Falistocco & Ferradini 2020) tal y como había sido propuesto por algunos autores (Thakur & Sign 1965; Walker 1972), en contraste con lo establecido por Bowden, que sugirió un número cromosómico de 2n=2x=16 (Bowden 1945, 1948).

Las Anonáceas surgieron en el Cretácico superior, en Laurasia y en el norte de Gondwana (Doyle & Le Thomas 1997; Richardson *et al.* 2004; Scharaschkin & Doyle 2005) comenzándose a diversificar hace 89 millones de años (Couvreur *et al.* 2011). Los géneros *Guatteria*, *Annona* y *Duguetia* se distribuyen en los neotrópicos, aunque *Duguetia* aparece también en el África tropical. *Uvaria* y *Polyalthia* son originarios de África y Asia. *Xylopia* destaca al ser el único género de la familia que no se encuentra en todas la regiones

tropicales, y *Anaxagorea* al ser el único género que aparece en Asia tropical y en los neotrópicos.

El origen geográfico del chirimoyo ha sido motivo de discusión entre los científicos. De hecho, hasta hace unos años, la teoría más apoyada era un origen sudamericano, concretamente en los valles interandinos del sur de Ecuador y del norte de Perú, ya que en esta región se había localizado una mayor variabilidad fenotípica, supuestas poblaciones silvestres así como descubrimientos arqueológicos de semillas y vasijas que asemejan a los frutos de chirimoya (Popenoe 1921; Bonavia *et al.* 2004). Sin embargo, nuevos trabajos moleculares con material vegetal de todo el rango de distribución de la especie en el continente americano y análisis geográficos, apuntan a la existencia de una mayor diversidad genética en América Central, concretamente en Honduras y Guatemala, reforzando la posibilidad de que esta zona sea el origen de la especie (Larranaga *et al.* 2017). Desde esta zona se habría dispersado hacia otras regiones de Centroamérica gracias a los dispersores naturales, fundamentalmente la macrofauna ya extinguida. Por otro lado, existen evidencias de que el fruto se consumía y se domesticaba en tiempos precolombinos en regiones andinas (Popenoe *et al.* 1989; Larranaga *et al.* 2017), por lo que se ha propuesto la hipótesis de que se podría haber producido la dispersión del material vegetal hacia Sudamérica por parte de los seres humanos vía marítima entre México y el Norte de Perú/Ecuador (Larranaga 2016).

La introducción de este frutal en España peninsular tuvo lugar a comienzos del siglo XVIII, desde donde probablemente se llevó a Italia, Madeira (Portugal), Islas Canarias, Argelia, Egipto y posiblemente desde Italia se llevó a Eritrea, Libia y Somalia. A finales del siglo XVIII fue introducido en Hawai (1790), Jamaica (1785) y Haití. Desde México se introdujeron semillas en California y, a comienzos del siglo XX, el departamento de Agricultura de Estados Unidos importó semillas de chirimoya desde Madeira (Morton 1987). Desde entonces este frutal ha ido expandiéndose poco a poco, y en la actualidad se encuentra en un amplio rango de países de clima subtropical, aunque sigue siendo un cultivo infrautilizado.

Descripción botánica

Annona cherimola es un árbol semi-caduco, que puede tolerar levemente las heladas. En su madurez puede alcanzar 7-8 m de altura y su tronco es corto. El sistema radicular es superficial y ramificado, pudiendo formar varios niveles (Rosell *et al.* 1997). Sus hojas

durante el estado juvenil son simples y de color verde blanquecino; sin embargo, durante la madurez son enteras, de color verde, pubescentes dorsalmente, presentando un peciolo corto y nerviación regular (Rosell *et al.* 1997) con disposición alterna, y forma de oval a elíptica, con tamaño de 10 a 25 cm de longitud y de 4 a 10 cm de ancho. Las yemas se encuentran protegidas por el peciolo de las hojas, pero cuando estas se caen, las yemas comienzan a desarrollarse emitiendo entre cuatro o más brotes (Guirado *et al.* 2003) que generaran flores y brotes vegetativos.

Figura I.2. Flor, fruto y árbol de chirimoyo. Fuente (flor): Jorge Lora.

Por lo general las flores se presentan principalmente en madera de un año o más de edad (Guirado *et al.* 2003). Aparecen solitarias o en grupos de hasta 8 o 9. Las flores son hermafroditas, aromáticas, colgantes, de color amarillo verdoso, fragantes y poco llamativas. En general, presentan tres pétalos grandes (de 2,5 a 4 cm), carnosos y tres pétalos rudimentarios. Los estambres son numerosos (150-200) y la pirámide de pistilos contiene alrededor de 150 unidades independientes, con un solo óvulo. Al igual que ocurre en el aguacate, presenta dicogamia protogínica aunque, en el caso del chirimoyo, no se produce cierre de las flores entre los estados sexuales. Estas características dificultan la autofecundación dentro de la misma flor, al estar sincronizado el estado sexual entre las flores de un mismo árbol (Lora *et al.* 2010, 2011a, 2011b). El ciclo de la flor dura alrededor de 2 días (Lora *et al.* 2010). Produciéndose la

antesis (apertura de la flor) generalmente durante las primeras horas del día, como ocurre en España, Israel y Nueva Zelanda. El cuajado varía según las condiciones ambientales y, debido a la escasez de polinizadores naturales y a la poda intensiva que dificulta la supervivencia de dichos polinizadores, la polinización en España suele ser manual. Sin embargo, algunos grupos de nitidúlidos (coleoptera) parece que pueden ser efectivos en la polinización bajo condiciones de sombreado y alta humedad (De la Peña *et al.* 2018).

El fruto es un sincarpo, que se forma a causa de la fusión de los carpelos procedentes de una única flor. Su forma es cónica o en forma de corazón, y un tanto irregular debido a su polinización heterogénea, y el tamaño variable dependiendo del genotipo (Figura I.2). Su pulpa es blanca, cremosa, con un alto contenido de azúcar (20-22 %) y baja acidez (1 %), y tiene un sabor característico que recuerda a una mezcla de piña y banana o fresa y pera.

Usos e importancia económica

La chirimoya es una fruta dulce con notables propiedades organolépticas y nutritivas, como minerales, fibra y vitaminas C y B3 (Galán *et al.* 2014). A esos compuestos se pueden sumar las acetogeninas, que presentan diversas propiedades farmacológicas como antitumorales o insecticidas y que están presentes en todas las especies de la familia, fundamentalmente en semillas y hojas (Morton 1987). Su pulpa se consume fresca, aunque también es utilizada para la elaboración de batidos, helados, yogurt, flan y vino (Popenoe *et al*. 1989; Scheldeman 2002). A pesar de ser un cultivo infrautilizado, suele ser muy apreciado por su agradable sabor. Como anécdota se puede citar a Charles Darwin, quien la probó durante su última parada en Perú en 1835, antes de partir hacia las Islas Galápagos y que remarcó en su diario del Beagle "Hay dos cosas en Lima que todos los viajeros han comentado; las mujeres "tapadas", u ocultas en "saya" y "manta", y la fruta llamada "chilimoya". En mi opinión, las primeras son tan hermosas como delicioso es lo último […]" (Darwin 1835).

A pesar de que el chirimoyo se cultivaba en tiempos precolombinos por civilizaciones americanas, sigue considerándose un frutal infrautilizado. Su cultivo se extiende a numerosas áreas subtropicales de todo el mundo: Perú, Ecuador, Chile, Bolivia en el continente americano, Italia e Israel entre los países productores del Mediterráneo, Australia y Nueva Zelanda en Oceanía, y Sudáfrica en el continente africano (Scheldeman 2002). Sin embargo, el principal productor comercial a nivel mundial es España, con unas aproximadamente 3.000 hectáreas cultivadas y 44.000 t en el año 2019 (MAPA 2019). En el caso de España, a partir

de 1940 y 1950 comenzó a tener importancia comercial, principalmente en la costa de la provincia de Granada. De hecho, en la población granadina de Jete se originó el cultivar 'Fino de Jete', el cual ocupa más del 95 % de la superficie cultivada de chirimoyo en España. Actualmente su cultivo se sigue concentrando en la costa mediterránea andaluza, especialmente en las provincias de Málaga y Granada.

2. Aproximaciones moleculares, y sus necesidades en *Persea americana* y *Annona cherimola*

A pesar del interés biológico y económico del aguacate y del chirimoyo, siguen existiendo numerosas cuestiones sin resolver sobre evolución, origen o diversidad genética, y sus programas de mejora se ven limitados debido a la escasa información molecular disponible, si la comparamos con otras especies. Un ejemplo puede ser con *Amborella trichopoda* (la especie vegetal viva más primitiva evolutivamente) o con el melocotonero (Tabla I.1).

En el caso de estos cultivos, como en la mayoría de los cultivos leñosos, los programas de mejora son más lentos debido a su largo periodo intergeneracional, por lo que la aplicación de herramientas moleculares puede servir para acelerar el proceso significativamente. Ambas especies tienen el valor añadido de ser angiospermas evolutivamente primitivas por lo que presentan un gran interés desde el punto de vista taxonómico y evolutivo, pudiendo emplearse como modelo para el estudio de los genes que han facilitado el éxito reproductivo de las angiospermas.

Tabla I.1. Comparación de la base de datos NCBI de *Amborella trichopoda*, *Annona cherimola, Persea americana* y *Prunus persica.*

Base de datos NCBI	***Amborella trichopoda***	***Prunus persica***	***Annona cherimola***	***Persea americana***
		Genes		
Gene	19.534	27.209	164	133
Popset	236	493	26	256
		Genomas		
Assembly	1	4	0	3
Bioproject	13	137	3	40
Genome	1	1	0	1
SRA	108	2.518	44	202

Popset: base de datos de secuencias de ADN que han sido seleccionadas para el estudio de una población particular. *Unigene*: base de datos de transcriptoma. *Bioproject*: base de datos de información biológica relacionados con un determinado objetivo. SRA: base de datos de secuencias generadas a partir de NGS. Consultado: 2 de marzo del 2020.

Marcadores moleculares

Tradicionalmente, los métodos de identificación de cultivares y patrones se han llevado a cabo mediante observaciones fenotípicas. De hecho, organismos como la UPOV (Unión Internacional para la Protección de las Obtenciones Vegetales) siguen anteponiendo los caracteres morfológicos a los marcadores moleculares a la hora de evaluar y diferenciar distintas variedades para su protección (UPOV 2011). A pesar de la importancia de este tipo de observaciones para asociar fenotipo con genotipo, existen una serie de limitantes de este tipo de metodología: las observaciones son lentas debido al largo periodo intergeneracional de los frutales, y los caracteres pueden verse modificados debido a las influencias ambientales. Por lo tanto, son necesarios procedimientos adicionales como el uso de marcadores moleculares para la caracterización genotípica y la diversidad genética (Wünsch & Hormaza 2002).

Durante las últimas décadas, se han producido importantes progresos en los métodos usados para estudiar las especies a nivel molecular, permitiendo estimaciones precisas y rápidas y, por tanto, favoreciendo el uso de marcadores moleculares en distintos campos de la biología.

Por definición un marcador molecular es una biomolécula cuya variación puede determinarse entre individuos, dicha variación está controlada por un único locus y puede usarse para el

análisis genético. Existen dos grandes grupos: bioquímicos y las secuencias de ADN. Según la información que pueden ofrecer se pueden clasificar como codominantes (como los marcadores bioquímicos, los RFLPs, microsatélites o SNPs), es decir, que permiten distinguir entre genotipos homocigotos o heterocigotos o, por el contrario, dominantes (como RAPDs o AFLPs), cuando los genotipos homocigotos no se pueden diferenciar de los heterocigotos.

En los últimos años las mejoras en las tecnologías de secuenciación de nueva generación (NGS) junto a su abaratamiento han permitido grandes avances en las plataformas de genotipado y generalizar el uso de polimorfismos mononucleótidos (SNPs). Los SNPs pueden ser definidos como variaciones de un único nucleótido (A,T,C, o G). Se localizan tanto en regiones codificantes como en no codificantes (Morin *et al*. 2004), siendo la variación más común a largo del genoma (Scheben *et al.* 2018). Son marcadores bialélicos y codominantes. Tienen bajas tasas de mutación, alta heredabilidad, existen numerosas metodologías para su detección y son altamente reproducibles entre laboratorios (Garrido-Cardenas *et al.* 2018). Su desventaja respecto a los marcadores multialélicos como los SSRs a la hora de detectar la heterocigosidad, se compensa con su relativa abundancia (Edwards *et al.* 2007; Guichoux *et al*. 2011). Dependiendo del tipo de mutación genética que se produzca, podemos clasificar los SNPs en transversiones, con cambios en nucleótidos C/G, A/T, C/A y T/G y transiciones con cambios C/T o G/A.

Gracias a su abundancia, los avances en las tecnologías y el abaratamiento de los procesos para generarlos, su importancia se está incrementando en los análisis genéticos de la última década (Henry & Edwards 2009), siendo uno de los marcadores más empleados para multitud de trabajos.

Como se ha comentado anteriormente, para especies como *Persea americana* y *Annona cherimola* la disponibilidad de marcadores moleculares es limitada. Aunque hasta la fecha se han producido importantes progresos tanto en aguacate (minisatelites [Lavi *et al.* 1991], VNTRs [Mhameed *et al.* 1996], RAPDs [Fiedler *et al.* 1998], RFLPs [Furnier *et al*. 1990; Davis *et al.* 1998], SSRs [Sharon *et al.* 1997; Schnell *et al.* 2003; Ashworth *et al.* 2003,2004; Borrone *et al.* 2007; Alcaraz & Hormaza 2007; Gross-German & Viruel 2013; Guzman *et al.* 2017; Boza *et al.* 2018; Ge *et al.* 2019a], SNPs [Ge *et al.* 2019b; Kuhn *et al.* 2019; Rubinstein *et al.* 2019]) como en chirimoyo (isoenzimas [Ellstrand & Lee 1987; Pascual *et al.* 1993; Perfectti & Pascual 1996,1998a,1998b,2005a,2005b]; RAPDs [Ronning *et al*. 1995]; AFLPs [Rahman *et al.* 1997;1998]; SSRs [Escribano *et al.* 2004,2007,2008a,2008b;

van Zonneveld *et al.* 2012]) todavía el número de accesiones identificado es escaso y faltan estudios de diversidad a gran escala.

Mapas de ligamiento

Los árboles frutales, como la mayoría de las especies leñosas, se caracterizan por tener un largo periodo intergeneracional y necesitar una gran superficie para su cultivo, por lo que la obtención de nuevas variedades requiere de mucho tiempo, dinero y espacio.

La mayoría de los caracteres de interés agronómico como la producción o la resistencia a determinadas plagas son complejos, es decir, caracteres que se encuentran controlados por varios genes (Goddard & Hayes 2009; Mackay *et al.* 2009). Una de las herramientas que permite acelerar la mejora de los genotipos es un mapa genético altamente saturado de marcadores moleculares, pues facilita la identificación de QTLs (*Quantitative Trait Loci*), regiones del genoma que contienen genes asociados a caracteres cuantitativos y, por tanto, la selección asistida por marcadores (MAS). La construcción de estos mapas es una de las aplicaciones más relevantes de los marcadores genéticos (Hayward *et al*. 2015).

Un mapa de ligamiento puede realizarse gracias a que los genes y los marcadores segregan por recombinación cromosómica durante la meiosis. El concepto de ligamiento entre marcadores genéticos en la herencia fue propuesto en 1911 por Morgan, y el primer mapa genético fue publicado dos años más tarde por Sturtevant, quien estableció que la frecuencia de recombinación entre genes podría aplicarse como medida relativa de la distancia que los separaba (Ayala & Kiger 1984). De hecho, aquellos marcadores que se encuentran más próximos en el cromosoma se transfieren juntos a la descendencia con más frecuencia que aquellos marcadores o genes que se encuentran alejados. La frecuencia de recombinación se emplea como indicador de la distancia existente entre dos loci concretos. Esta distancia genética se mide en centiMorgans (cM), equivaliendo un centiMorgan a una frecuencia de recombinación del 1 %. A la hora de distribuir los loci en un grupo de ligamiento el orden debe de ser el mismo que el orden en el que están distribuidos a lo largo del genoma; en cambio las distancias que se basan en la frecuencia de recombinación podrían variar según el fragmento del genoma.

El mapa de ligamiento o genético es una representación en la que se muestra la distribución relativa de los genes y/o de los marcadores moleculares en los grupos de ligamiento o cromosomas. Para su desarrollo es necesario escoger unos progenitores adecuados que se

diferencien en uno o más caracteres de interés para realizar un cruzamiento óptimo. Por otro lado, es necesario calcular la frecuencia de recombinación, realizar un análisis de ligamiento y estimar las distancias genéticas (Staub *et al.* 1996).

En los últimos años se han construido numerosos mapas genéticos con el objetivo de identificar una región del genoma que controle un rasgo de interés (Baxter *et al.* 2011). También se han desarrollado mapas genéticos con el objetivo de generar pseudocromosomas cuando el genoma generado no se encuentra organizado a nivel cromosómico (Hayward *et al.* 2015; Gabay *et al.* 2018). De hecho, los mapas genéticos han sido ampliamente utilizados para generar un ensamblaje de calidad facilitando la reconstrucción de los cromosomas más probables (Tang *et al*. 2015).

La robustez del mapa dependerá de los marcadores moleculares usados y del tamaño de la población, empleándose normalmente poblaciones de entre 50 a 250 individuos (Mohan *et al.* 1997). A mayor número de marcadores moleculares, mayor probabilidad de encontrar algún marcador ligado a genes que afecten al carácter de interés, o de mejorar el ensamblaje de un genoma. Se han utilizado, tanto por separado como combinados, diversos tipos de marcadores (RAPDS, AFLPs, RFLP o SSR) para generar mapas de ligamiento. Sin embargo, en los últimos años los SNPs se han convertido en los marcadores más valiosos a la hora de desarrollar mapas ya que, gracias a las nuevas tecnologías y metodologías, se pueden generar de forma económica un gran número de marcadores.

Tradicionalmente, la construcción de mapas se ha llevado a cabo a partir de una población F2 o BC1 (población obtenida del primer retrocruzamiento) generada a partir de cruzamientos entre dos líneas puras. Sin embargo, este tipo de líneas son complicadas de obtener en la mayoría de especies leñosas por limitaciones de tiempo y espacio. En frutales, aunque en ocasiones se generan poblaciones F2 a partir de la autofecundación de individuos relativamente heterocigotos, el tipo de población más frecuente como base de mapeo es la "F1 segregante" (Grattapaglia & Sederoff 1994) generada a partir de individuos no relacionados y heterocigotos. Algunos ejemplos recientes en los que se generan mapas a partir de poblaciones F1 en frutales incluyen achiote (Romero-López *et al*. 2019), peral (Gabay *et al.* 2018) o manzano (Ban & Choi 2018), aunque también hay recientes ejemplos de mapas genéticos generados a partir de poblaciones F2 como en cerezo (Calle *et al*. 2018) o en melocotonero (Da Silva Linge *et al.* 2018).

3. Aproximaciones genómicas

En la década de 1970 se comenzaron a desarrollar tecnologías de secuenciación del ADN, gracias a la información desarrollada años antes en la caracterización de los ácidos nucleicos por Rosalind Franklin en 1951 y el descubrimiento de la estructura del ADN por James D. Watson y Francis Crick en 1953. Los primeros pasos consistieron en la determinación del orden de los nucleótidos en una molécula de ADN, concretamente la secuenciación química, desarrollada por Gilbert y Maxam, y con la que secuenciaban una base por mes (Shendure *et al.* 2017). No fue hasta 1977, cuando Sanger desarrolló el método de terminación de cadena que, junto al método desarrollado por Gilbert y Maxam, transformó el campo de la secuenciación, apareciendo la llamada secuenciación de primera generación (Metzker 2010). De hecho, en 1987, Smith, Hood y Applied Biosystems (Smith *et al.* 1986) diseñaron secuenciadores basados en la metodología propuesta por Sanger, comenzando a secuenciar 1.000 bases por día. Sin embargo, el coste y la lentitud de esta metodología enfatizó la necesidad de desarrollar nuevas tecnologías.

En 2005 salió a la luz el primer secuenciador comercial de secuenciaciones de nueva generación (*Next Generation Sequencing*, NGS) (con la tecnología 454). Este avance supuso una nueva revolución en el estudio básico y aplicado de la genómica, además de un gran desafío metodológico, un comienzo de una segunda generación de secuenciación y de la secuenciación masiva (*High Throughput Sequencing*, HTS).

Al pirosecuenciador 454, le siguió la secuenciación de Illumina (2005) que se basa en un procedimiento denominado "amplificación por puentes" y el uso de nucleótidos marcados con fluoróforos (detectados por una cámara al ser incorporados) y que bloquean de forma reversible la elongación de la cadena (Bentley *et al.* 2008), la secuenciación por ligamiento (SBL) de SOliD (2007) y la secuenciación por iones semiconductores de Ion Torrent Biosystems (2011).

Para algunas plataformas como Illumina, es necesario la creación de genotecas para llevar a cabo la secuenciación. Este procedimiento se basa en la fragmentación aleatoria de ADN, la ligación de adaptadores, amplificación mediante una PCR, selección de tamaño de los fragmentos, purificación y cuantificación final. Existen distintos tipos. *Single reads* es la más sencilla y permite secuenciar el ADN desde un solo extremo. *Paired ends* permite secuenciar fragmentos a partir de los dos extremos mejorando la calidad de las lecturas generadas, y

mate pairs es parecida a *paired ends*, pero se genera a partir de fragmentos de mayor longitud y proporciona una cobertura más uniforme (Illumina 2010). Todas estas plataformas de segunda generación se caracterizan por generar lecturas cortas (desde 75 pb a 700 pb) (Jung *et al.* 2019) (Tabla I.2), y cada vez son más rápidas. De hecho, actualmente existen secuenciadores como NovaSeq 5000/6000 que secuencian 1-3 trillones de bases por día (Rech 2019, comunicación personal). Estas nuevas tecnologías han revolucionado el análisis de genética de poblaciones, la filogenética, el ensamblaje de nuevas referencias, y la identificación de variaciones como SNPs, inserciones y deleciones (Le Nguyen *et al.* 2018).

Tabla I.2. Características de algunas de las metodologías de secuenciación NGS.

Método	**Longitud de lectura (pb)**	**Precisión**	**Lecturas/corrida**
Applied Bio 3730XL (Sanger)	400-900	99,9 %	384
Roche 454 GS FLX (Pirosequenciación)	700	99,9 %	1.000.000
Illumina HiSeq 2500 (Secuenciacón por síntesis)	1x36 2x50 2x100 2x125 2x250	99 %	5.000.000.000
SoLiD 4 (Secuenciación por ligación)	1x25 2x50	99,9 %	1.400.000.000
ION Torrent (Secuenciación por semiconducción)	200-400	98 %	80.000.000
Pacific Biosciencies RSII (SMRT)	10.000	85 %	750.000
Oxford N.MinION (Nanopore)	5.000-200.000	95 %	7.000-12.000
Pacific Biosciencies Sequel II (CCS)	15.000-20.000	99 %	-

Actualmente Illumina es la plataforma NGS más utilizada pero, aunque los precios de la secuenciación de segunda generación son cada vez más asequibles, existen una serie de limitaciones. Entre ellas destaca que la generación de lecturas cortas presenta dificultades a la hora de trabajar con regiones repetitivas y complejas. Además, hay que tener en cuenta los sesgos producidos por la PCR, el tiempo necesario, y la complicación de construir las genotecas, junto al precio de los instrumentos empleados. Por ello, a pesar de los enormes beneficios de la metodología implementada, se ha generado una tercera generación, donde actualmente dominan dos plataformas: Pacbio y Oxford Nanopore. Estas tecnologías ofrecen lecturas más largas, entre 8 kb y 40 kb (con máximo de > 150 kb para PacBio y >2 Mb para

Nanopore) (Jung *et al.* 2019), por lo que están siendo ampliamente usadas a la hora de generar genomas *de novo,* con una preparación de muestras más simple, y donde se secuencia una sola molécula en tiempo real. Hasta hace poco estas plataformas se caracterizaban por presentar altas tasas de error (~5-10 %). Sin embargo, gracias a los avances en la química empleada y en el secuenciador, ha surgido una nueva metodología de PacBio (*single molecule High-Fidelity* [HiFi]). Esta tecnología ha permitido generar lecturas de gran longitud (aproximadamente 10-20 kb) y precisas (~1 %) (Wenger *et al.* 2019).

Como se ha comentado anteriormente, los sistemas de segunda generación son los más empleados al ser menos costosos, y no solo para el estudio de los polimorfismos de ADN, pues a la hora de generar ensamblajes *de novo*, numerosos estudios han apostado por enfoques híbridos que combinan los datos generados por plataformas de segunda y tercera generación. Sin embargo, en un futuro próximo los sistemas de tercera generación posiblemente reemplazarán a los sistemas anteriores, aunque un aspecto limitante puede ser el gran esfuerzo necesario en el análisis bioinformático para procesar los datos generados con estas tecnologías (Bansal & Boucher 2019).

Aproximaciones de genotipado

El estudio de los polimorfismos del ADN, conocido como genotipado, es esencial para los programas de mejora, estudios de genética de poblaciones, filogenética, conservación, mapeado genético, adaptación y evolución. Dentro de los avances producidos en las últimas décadas para el genotipado, cabe destacar el genotipado por secuenciación, GBS (*Genotyping By Sequencing*) que puede generar de cientos a miles de marcadores tipo SNP en un solo ensayo (Elshire *et al.* 2011; Spindel *et al.* 2013; Poland & Rife 2012), sin la necesidad de tener un genoma de referencia, por lo que es una herramienta muy adecuada para especies no modelo (Scheben *et al.* 2017). Además, las secuencias generadas mediante GBS se pueden analizar repetidamente (Poland & Rife 2012). Sin embargo, posee una serie de limitaciones, como el alto número de datos ausentes que genera, y la tasa de error de la técnica de secuenciación (Glaubitz *et al*. 2014). Esta técnica de alto rendimiento se basa fundamentalmente en la reducción del genoma por el uso de enzimas de restricción (REs) (Elshire *et al.* 2011) para secuenciar posteriormente con tecnología Illumina (Deschamps *et al.* 2012). Se han llevado a cabo numerosos estudios de cultivos empleando esta metodología; algunos ejemplos son: pecanero (Bentley *et al.* 2019), café (Anagbogu *et al.* 2019), centeno (Schreiber *et al.* 2019), garbanzo (Kujur *et al.* 2015) o maíz (Lu *et al.* 2015).

Otra técnica que cada vez es más popular para genotipado es la resecuenciación. Esta metodología genera un genotipado más exhaustivo, y evita la búsqueda de SNPs erróneos (Huang *et al.* 2009). Sin embargo, a diferencia de GBS, es necesario disponer de un genoma de referencia. Algunos ejemplos recientes publicados en cultivos han sido: algodón (Wen *et al.* 2019b), nogal (Stevens *et al.* 2018) o mijo (Varshney *et al.* 2017).

Como metodología alternativa destaca la secuenciación de ARN, del transcriptoma entero para clonación al azar o RNA-seq. Esta técnica se basa en la secuenciación de genotecas de ADN complementario (ADNc) obtenido a partir de un ARN de buena calidad. Permite trabajar en zonas concretas del genoma, y facilita a los investigadores determinar SNPs que alteran secuencias de codificación. Normalmente se utiliza para analizar la expresión génica; de hecho varios estudios han utilizado este tipo de metodología en frutales. Algunos ejemplos son: aguacate (Ibarra-Laclette *et al.* 2015; Xoca-Orozco *et al.* 2017; Xoca-Orozco *et al.* 2019; Chabikwa *et al.* 2020), atemoyo (Li *et al*. 2019a; Chen *et al*. 2019b) o cerezo (Wen *et al*. 2019a). Sin embargo, también puede ser empleada para genotipado, a pesar de que, en numerosas ocasiones, el número de marcadores generados con esta metodología no es suficiente para llevar a cabo estudios de asociación del genoma (GWAS), pues el número de SNPs en las regiones codificantes es menor que en las regiones no codificantes (Scheben *et al*. 2017). Algunos ejemplos reportados en especies leñosas incluyen aguacate (Ge *et al.* 2019b), ginkgo (Wu *et al.* 2019b) o longan (Wang *et al.* 2015).

Las metodologías citadas anteriormente para el genotipado utilizan plataformas de segunda generación, aunque a menudo se han beneficiado de los genomas de referencia generados a través del empleo de la tecnología de PacBio o Oxford NanoPore. Recientemente, esta última tecnología ha comenzado a utilizarse para genotipar, como en el caso de canola (Malmberg *et al.* 2019), y posiblemente su uso para este fin se extienda gracias a las nuevas optimizaciones en la metodología de PacBio (Wenger *et al.* 2019).

Genomas de referencia

El genoma de un organismo contiene todo el material genético codificado en ácidos desoxirribonucleicos (ADN). Posee un tamaño que varía de una especie a otra (Primrose & Twyman 2006), encontrándose en plantas grandes diferencias incluso entre especies próximas.

La secuenciación del genoma es una pieza clave que nos permite comprender con mayor exactitud la biología básica del organismo.

En plantas encontramos tres genomas. El nuclear, el mitocondrial y el cloroplástico, siendo los dos últimos mucho más pequeños. Por ejemplo, en *Arabidopsis thaliana,* el genoma mitocondrial tiene un tamaño de 367 Kb, y el genoma cloroplástico tiene un tamaño de 155 Kb frente al nuclear que tiene un tamaño de 135 Mb (Kersey 2019). Otra característica destacable de los genomas de plantas es que a lo largo de su historia evolutiva han sufrido frecuentes eventos de poliploidización favoreciendo la especiación y adaptación (Martín *et al.* 2019).

La secuenciación de genomas comenzó en 1976, concretamente con el genoma del bacteriófago MS2 (Fiers *et al.* 1976). Pero no fue hasta 1987, cuando el desarrollo de los secuenciadores automáticos supuso un cambio fundamental, y en 1995 se secuenció el primer genoma bacteriano de *H. influenzea* (Fleischmann *et al.* 1995), seguido por el primer genoma de eucariotas de *Saccharomyces cerevisiae* (Goffeau *et al.* 1996). Tres años antes de publicar el genoma humano (Collins *et al*. 2003), salió a la luz la publicación del genoma de *Arabidopsis thaliana* con un tamaño estimado inicialmente de 125 Mb (The Arabidopsis Genome Initiative 2000), marcando una nueva era en el estudio de plantas (Nature Plants 2018), al considerarse una referencia en los estudios de genética molecular. Tras Arabidopsis, se secuenció el genoma del arroz (Goff *et al.* 2002; Yu *et al.*, 2002) a pesar de su gran tamaño y la falta de información molecular previa, siendo el primer genoma generado para una especie de interés económico (Arús & Puigdomènech 2008).

A partir del 2005, el número de genomas disponibles se incrementó debido a la reducción del coste en la secuenciación de plataformas de segunda generación. Sin embargo, aunque la calidad de un genoma de referencia desarrollado no solo depende de la secuenciación usada, el desarrollo de ensamblajes de genomas a partir de lecturas de longitud cortas tiene una serie de limitaciones según el tamaño del genoma, las repeticiones, la heterocigosidad [considerándose una heterocigosidad alta a partir de aproximadamente 0,50 % (Vurture *et al*. 2017)] o la ploidía. De hecho, en los ensamblajes de genomas a partir de este tipo de lecturas, las repeticiones y las duplicaciones suelen causar secuencias quiméricas (secuencias biológicas y fragmentos de ADN no relacionados que son unidos artificialmente mediante bioinformática) y cóntigos (*contigs*, regiones consenso de ADN) fragmentados (Jung *et al.* 2019).

Una de las primeras aproximaciones para superar los problemas causados por el uso de lecturas cortas y la repeticiones en el genoma, es el empleo de una metodología basada en *Bruijn graphs*, un algoritmo ampliamente utilizado en el ensamblaje de genomas que utiliza K-mers (secuencias de nucleótidos de un tamaño concreto) con el objetivo de construir cóntigos (*contigs*, regiones consenso de ADN). Sin embargo, al trabajar con genomas de mayor tamaño, no se desarrollan ensamblajes correctos, completos y con contigüidad. Las tecnologías que generan lecturas más largas (como PacBio y Oxford Nanopore) facilitan la contigüidad del ensamblado aunque, por lo general, presentan altas tasas de error. Una alternativa a la hora de generar genomas de referencia puede ser el uso de un enfoque "híbrido", en el que se combinan los datos generados por plataformas de segunda y tercera generación, corrigiendo las lecturas largas mediante el cartografiado de lecturas generadas con secuenciadores de Illumina. Numerosos trabajos han utilizado esta estrategia. Algunos ejemplos recientes son el genoma del liriodendron (Chen *et al.* 2019c) o del aguacate (Rendón-Anaya *et al.* 2019). Sin embargo, esta aproximación no sería necesaria si se emplea la nueva tecnología de PacBio, que genera lecturas con mayor calidad. De hecho, si comparamos estas lecturas HiFi con las que se generaban empleando las metodologías PacBio anteriores, podemos ver que, aunque son de menor longitud, su alta precisión compensa a la hora de generar un ensamblaje de calidad (Wenger *et al.* 2019).

Recientemente se ha producido un gran progreso en nuevas metodologías que facilitan la mejora del ensamblaje de genomas, destacando la secuenciación de una sola molécula, junto al cartografiado de datos ópticos (BioNano) y Hi-C. Hi-C se basa en la captura de la conformación de los cromosomas, y ha sido empleada en el ensamblaje de genomas de numerosas especies vegetales, facilitando la corrección de la orientación y el orden de los *scaffolds* (unión de *contigs* usando información complementaria sobre la orientación y posición en el genoma). En los últimos años se han generado ensamblajes de alta calidad usando una o la combinación de varias de las herramientas comentadas (Daccord *et al.* 2017; Zhang *et al.* 2018; Hosmani *et al.* 2019). Algunos ejemplos recientes de su uso en especies leñosas son el ensamblaje del genoma del alcanforero (Chaw *et al.* 2019) o el ensamblaje del genoma de durian (Teh *et al.* 2017).

Estas numerosas metodologías junto a los avances bioinformáticos de los últimos años están solventando los problemas que aparecen a la hora de generar el ensamblaje de un genoma de referencia, y facilitan la disponibilidad de genomas de calidad (Shendure *et al.* 2017; Kersey

2019). Como resultado, el ensamblaje de genomas se está convirtiendo en una rutina en biología (Kersey 2019), a pesar de no existir un ensamblaje de genoma de plantas perfecto (Jung *et al*. 2019).

A la hora de evaluar la calidad del ensamblaje de un genoma de referencia se tienen en cuenta una serie de parámetros, entre los que destaca N50, que especifica la contigüidad del genoma (Yandell & Ence 2012), pues es la longitud de la secuencia que representa el 50% del ensamblaje siempre y cuando las secuencias se encuentren ordenadas de mayor a menor tamaño. Aunque también se suelen usar otros parámetros como el tamaño del ensamblaje total, el número total de secuencias (*scaffold*/*contigs*), la longitud más larga del s*caffold* o de los c*ontigs*, el tamaño medio del *scaffold* y *contigs* (Jung *et al*. 2019) o como de completo respecto al espacio de genes se encuentra el genoma (Simão *et al.* 2015). Cuanto menos fragmentado sea el ensamblaje de un genoma, es decir, más largo el N50, mayor contigüidad tendrá el genoma desarrollado, permitiendo una anotación más exhaustiva de este. De hecho, si N50 tiene una longitud igual a la longitud media de un gen, aproximadamente el 50 % de los genes se localizarán en un *scaffold*, lo que facilitará el uso de este para numerosos análisis (Yandell & Ence 2012).

Actualmente, se han publicado al menos 341 genomas de referencia en especies vegetales (https://www.plabipd.de/), incluyendo algas, briofitas, gimnospermas y angiospermas. Debido al éxito en los ensamblajes de genomas de algunas plantas modelo, se ha anunciado un proyecto que se compromete con la secuenciación de 10.000 genomas de plantas (Chen *et al.* 2018). No obstante, la mayoría de los genomas generados hasta ahora pertenecen a especies eudicotiledóneas o monocotiledóneas (Soltis & Soltis 2019), con la excepción de *Amborella trichopoda* en 2013 con un tamaño de genoma de 870 Mb (Albert *et al.* 2013), *Liriodendron chinense* (Chen *et al*. 2019c) con un tamaño de genoma de 1.75 Gb, *Cinnamomum kanehirae* (Chaw *et al.* 2019) con aproximadamente un tamaño de genoma de 823 Mb, *Persea americana* (Rendón-Anaya *et al.* 2019) con un genoma de tamaño aproximado de 1 Gb, *Nymphaea colorata* con un tamaño de genoma aproximado de 400 Mb (Zhang *et al*. 2019c), *Euryale ferox* con un tamaño de genoma de aproximadamente 768 Mb y *Ceratophyllum demersum* con un tamaño de genoma aproximado de 777 Mb (Yang *et al.* 2020b) que proporcionan nuevas visiones sobre la evolución de las angiospermas.

El desarrollo tanto de un genoma de referencia para chirimoyo como la optimización del genoma de aguacate podrían aportar una nueva perspectiva evolutiva en angiospermas

primitivas. Recientemente, se ha publicado un primer ensamblaje del genoma de aguacate (Rendón-Anaya *et al.* 2019). Sin embargo, hasta el momento, no se ha generado ningún genoma que represente a la familia Annonaceae. Por otro lado, los mapas genéticos generados para estos cultivos son limitados y de baja resolución (Sharon *et al*. 1997 y Borrone *et al.* 2009, en aguacate; Escribano 2007 y Martín 2013 en chirimoyo). Hasta el momento no se dispone de ningún estudio en ambas especies en el que se desarrollen mapas genéticos usando marcadores moleculares tipo SNPs. En el caso del chirimoyo, los mapas genéticos disponibles hasta la fecha han sido construidos a partir de poblaciones F1 segregantes, resultado de cruzamientos interespecíficos entre 'Fino de Jete' x 'Thai seedless' a partir de 59 y 82 SSRs (Escribano 2007; Martín 2013), y 'Bonita' x 'Fino de Jete' (Martín 2013) a partir de 66 SSRs. A esta carencia de información en ambas especies pretende dar respuesta los trabajos recogidos en esta tesis doctoral.

OBJETIVOS

Objetivos

El objetivo general de esta tesis es avanzar en el conocimiento genómico del aguacate (*Persea americana* Mill.) y del chirimoyo (*Annona cherimola* Mill.), dos cultivos originarios de los Neotrópicos, con el propósito de posibilitar el desarrollo de herramientas que faciliten los programas mejora y el estudio de su diversidad genética. Este objetivo general se desglosa en tres objetivos específicos:

1. Desarrollar un borrador del genoma (cv. Hass) y marcadores moleculares SNPs mediante genotipado por secuenciación en aguacate (*Persea americana* Mill.), para caracterizar un conjunto de genotipos mediante el estudio de su diversidad genética y estructura poblacional (Capítulo 1).

2. Generar el primer ensamblaje y anotación del genoma del chirimoyo (*Annona cherimola* Mill.) (Capítulo 2).

3. Desarrollar marcadores moleculares SNPs mediante genotipado por secuenciación para una población F2 generada a partir de un cruzamiento interespecífico 'Fino de Jete' (*Annona cherimola* Mill.) x 'Thai seedless' (*Annona squamosa* L.) y elaborar un mapa de ligamiento (Capítulo 3).

CAPÍTULO 1

Caracterización genómica de aguacate (*Persea americana* Mill.)[1]

Resumen

La mejora de cultivos se basa en el uso de material diverso tanto fenotípicamente como genéticamente. Para llevar a cabo un programa de mejora efectivo se necesita un buen conocimiento de la estructura poblacional y de la diversidad genética de la especie a mejorar. En el caso del aguacate (*Persea americana* Mill.), un cultivo nativo de Mesoamérica con un aumento significativo de popularidad a nivel mundial, los programas de mejora son muy lentos debido a la existencia de un claro vacío de información molecular. De hecho, actualmente la producción mundial de aguacate está basada en una única variedad, 'Hass', que surgió por azar en California hace casi 100 años. Con el objetivo de reducir esa brecha e impulsar los programas de mejora, en este estudio se ha generado el ensamblaje de un borrador del genoma de la variedad Hass para usarlo de referencia al estudiar 71 genotipos de aguacate que representan las tres razas o subespecies (Mexicana, Guatemalteca y Antillana). Se generó una media de 5,72 millones de lecturas por individuos, y 7.108 polimorfismos mononucleótidos (SNPs) para los 71 genotipos analizados. Estos marcadores moleculares se han usado para estudiar la diversidad genética y la estructura poblacional. Los resultados mostraron una agrupación de acuerdo a las razas botánicas en cuatro grupos: Mexicanos, Guatemaltecos, Antillanos y un grupo adicional de híbridos entre Guatemaltecos y Mexicanos (GxM). El alto número de marcadores tipo SNPs desarrollados en este estudio será un recurso genómico relevante para la comunidad de aguacate.

Introducción

El aguacate (*Persea americana* Mill.) es un árbol subtropical de hoja perenne nativo de Mesoamérica. Pertenece a las Lauraceae, una familia del orden Laurales que, junto con los órdenes Canellales, Piperales y Magnoliales, está incluida en el clado Magnoliid de las

[1] Talavera A, Soorni A, Bombarely A, Matas A J, Hormaza J I. Genome-Wide SNP discovery and genomic characterization in avocado (*Persea americana* Mill.). Scientific Reports 9, 20137 (2019). (Anexo 1).

angiospermas basales. Esta familia pantropical tiene alrededor de 50 géneros y entre 2.500 a 3.000 especies.

Dentro de esta familia existen otras especies con importancia económica entre las que se incluyen el laurel (*Laurus nobilis* L.), la canela (*Cinnamomum verum* J.Presl), el alcanforero (C. *camphora* (L.) J.Presl) y árboles madereros (*Nectandra* spp., *Ocotea* spp. y *Phoebe* spp.).

Tradicionalmente, los genotipos de aguacate se han clasificado en tres razas o subespecies según sus características botánicas y preferencias ecológicas, Mexicana, Guatemalteca y Antillana. Las accesiones de raza Mexicana y Guatemalteca están adaptadas a zonas altas de América Central (climas fríos), siendo la raza Guatemalteca más susceptible a las bajas temperaturas, mientras que las accesiones de raza Antillana se encuentran adaptadas a áreas de tierras bajas con climas tropicales (Schaffer *et al.* 2013).

La demanda del mercado del aguacate ha aumentado exponencialmente en los últimos años, siendo la producción mundial del aguacate en 2018 superior a 6 millones de toneladas (FAO 2018). La mayor parte de la producción se concentra en unos pocos países, siendo México el mayor productor con el 34 % de la producción mundial (más de 2 millones de toneladas). Sin embargo, a pesar de la creciente importancia de este cultivo, existen significativos cuellos de botella para la mejora de esta especie. El desarrollo de nuevos cultivares de calidad es una necesidad, ya que aproximadamente el 90 % de la producción mundial de aguacate depende de un solo cultivar, 'Hass', que se seleccionó por azar en California en la segunda década del siglo XX.

En aguacate, se han utilizado diversos marcadores moleculares para distintos fines: genotipado, análisis de paternidad, estudios de diversidad y de filogenia, elaboración de mapas de ligamiento y detección de rasgos de interés. Los primeros trabajos estuvieron enfocados en el uso de minisatélites (Lavi *et al*. 1991) (*Variable Number of Tandem Repeats*) (VNTRs) (Mhameed *et al*. 1996), amplificación aleatoria de ADN polimórfico (*Random Amplified Polymorphic*) (RAPDs) (Fiedler *et al.* 1998) y polimorfismos de longitud de fragmentos de restricción (*Restriction Fragment Length Polymorphism*) (RFLPs) (Furnier *et al*. 1990; Davis *et al*. 1998). Posteriormente se comenzó a usar marcadores codominantes y altamente polimórficos, los microsatélites o SSRs (*Single Sequence Repeats*) que facilitan el estudio de relaciones intraespecíficas y de diversidad. Este tipo de marcadores han sido específicamente desarrollados en aguacate y usados para diversos objetivos (Sharon *et al*.

1997; Schnell *et al*. 2003; Ashworth & Clegg 2003; Ashworth *et al*. 2004; Borrone *et al*. 2007; Alcaraz & Hormaza 2007; Gross-German & Viruel 2013; Guzmán *et al.* 2017; Boza *et al*. 2018; Ge *et al*. 2019a). Sin embargo, a pesar de sus numerosas ventajas, tienen el inconveniente de que, al no encontrarse distribuidos uniformemente a lo largo del genoma, dificultan los análisis de asociación (Ching *et al.* 2002). Además, el uso de SSRs en distintas poblaciones es laborioso y costoso si lo comparamos con el empleo de las secuenciaciones de nueva generación (*Next Generation Sequencing*, NGS) (Rasheed *et al.* 2017).

En la última década, los SNPs se han convertido en el tipo de marcador más usado en genética vegetal para distintas aplicaciones: desarrollo de mapas de ligamiento, búsqueda de locus asociados a un carácter cuantitativo, selección genómica (GS) o selección asistida por marcador (MAS) (Scheben *et al.* 2017). Los SNPs presentan una serie de ventajas frente a los marcadores utilizados previamente. Son la mutación más frecuente a lo largo de los genomas de los eucariotas, su naturaleza bialélica ofrece precisión a la hora de hacer el llamamiento de variantes, son altamente reproducibles y su coste reducido hace que sean accesibles para la mayoría de laboratorios (Studer & Kölliker 2013; Chagné *et al*. 2008; Wang *et al*. 2015). Sus ventajas son especialmente relevantes en cultivos perennes leñosos, ya que su uso podría reducir significativamente el tiempo y el coste en programas de mejora.

Hasta el momento el empleo de la secuenciación de nueva generación (NGS) en aguacate se ha centrado en estudios transcriptómicos (Ibarra-Laclette *et al.* 2015; Vergara-Pulgar *et al*. 2019) y el desarrollo de SNPs para caracterizar diversidad genética (Kuhn *et al*. 2019; Ge *et al.* 2019b; Rubinstein *et al.* 2019). Además, recientemente el primer genoma de aguacate de un genotipo criollo mexicano se ha publicado (Rendón-Anaya *et al.* 2019). En este trabajo, con el objetivo de generar SNPs de alta calidad para avanzar en estudios genómicos en aguacates, se han genotipado 71 accesiones que representan las tres razas botánicas. Estos marcadores han sido seleccionados cartografiando contra un borrador de genoma del cultivar más importante a nivel mundial, 'Hass', con el objetivo de aumentar la calidad de los marcadores desarrollados.

Material y métodos

Material vegetal

Se seleccionaron 71 accesiones de aguacate (*Persea americana* Mill.), de los que se recolectaron sus hojas jóvenes a partir de árboles que proceden de tres bancos de germoplasma distintos: IHSM-UMA-CSIC "La Mayora" (Algarrobo Costa, España), Westfalia farms (Tzaneen, Sudáfrica) y el banco de germoplasma nacional de aguacate de Estados Unidos (Miami, FL, USA). Las accesiones escogidas combinan genotipos obtenidos a partir de distintos programas de mejora ('Gem', 'Gwen', 'Iriet' o 'Lamb Hass'), variedades comerciales ('Bacon', 'Choquette', 'Edranol', 'Fuerte', 'Hass' o 'Reed'), portainjertos ('Dusa', 'Thomas' o 'Toro Canyon') y accesiones locales españolas con posible interés para generar nuevos patrones con tolerancia a *Rosellinia necatrix* ('La Piscina' o 'C.A. Bueno'). También se incluyeron en el análisis como control de los resultados obtenidos, dos muestras diferentes de 'Hass' procedentes de dos bancos de germoplasma distintos (Tabla 1.1). Tanto las hojas de las accesiones conservadas en el banco de germoplasma de IHSM-UMA-CSIC "La Mayora", como las del banco de germoplasma nacional de aguacate de Estados Unidos, se conservaron en fresco, mientras que el material procedente de Westfalia Farms se conservó en seco usando bolsas con gel de sílice hasta su uso.

Tabla 1.1. Lista de las 71 accesiones de aguacates estudiadas en este trabajo con SNPs. Las razas están representadas por los siguientes códigos: G = Guatemalteca; M = Mexicana; WI = Antillana. Los híbridos interraciales se indican con un aspa (x).

Accesiones	Identificador de la muestra	Código	Colección de Germoplasma	Asignación previa de raza	Asignación de raza a partir de los resultados de este trabajo
0028(Ardith)	2835	1	Sudáfrica	G×M[1]	G×M
A0.25	A02554	2	Sudáfrica	Desconocida	G×M
A0.68	A06852	3	Sudáfrica	Desconocida	G×M
87.17.1	871728	4	Sudáfrica	Desconocida	G×M
1.14.2	114218	5	Sudáfrica	Desconocida	G×WI
Alcaraz	ALCA74	6	España	Desconocida	G×M
Bacon	BACO39	7	Sudáfrica	G×M[2],M[3,4] o G[8]	G×M
Bernecker	BERN18	8	EE.UU.	WI[5]	WI

Beta	BETA19	9	EE.UU.	G×WI[6]	G×WI
A0.57	A05720	10	Sudáfrica	G×M[2]	G×M
Butler	BUTL16	11	EE.UU	WI[1]	WI
C.A. Bueno	CABU95	12	España	Desconocida	M
Catalina	CATA11	13	EE.UU	WI[1]	WI
Choquette	CHOQ9	14	EE.UU.	G×WI[1]	G×WI
Cilfam	CILF46	15	Sudáfrica	Desconocida	G×M
Colin V-33	COLI31	16	Sudáfrica	G×M[1]	G×M
Collinred B	COLL1	17	EE.UU.	G×WI[1]	G×WI
Collinson	COLL36	18	EE.UU.	G×WI[1]	G×WI
Dusa	DUSA33	19	España	G×M[2]	G×M
Edranol	EDRA63	20	Sudáfrica	Híbrido[4] o G[4]	G×M
Fuchsia	FUCH17	21	EE.UU.	WI[1]	G×M×WI
Fuerte	FUER16	22	Sudáfrica	G×M[2] o M[8]	G×M
G-6	G692	23	España	M[2]	M×WI
Gem	GEM77	24	España	G×M[2] o G[4]	G×M
Gottfried	GOTT04	25	Sudáfrica	M[9]	M×WI
Grace	GRAC26	26	Sudáfrica	Desconocida	G×M
Gwen	GWEN40	27	Sudáfrica	G×M[1] o G[8]	G×M
H287	H28757	28	Sudáfrica	Desconocida	G×M
Hansie	HANS05	29	Sudáfrica	Desconocida	M
Hass	HASS38	30	España	G×M[3,31] o G[2]	G×M
Hass	HASS55	31	Sudáfrica	G×M[3,31] o G[2]	G×M
Iriet	IRIE34	32	España	G×M[3]	G×M
A0.67	A06729	33	Sudáfrica	Desconocida	G×M
Lamb Hass	LAHA24	34	Sudáfrica	G×M[2,3]	G×M
La Piscina	LAPI93	35	España	Desconocida	M
Largo	LARG24	36	EE.UU.	WI[1]	G×WI
Linda	LIND50	37	Sudáfrica	G[1]	G
Lisa	LISA23	38	EE.UU.	M×WI[1]	G×M×WI
Lyon	LYON25	39	Sudáfrica	Híbrido[4] o G[1]	G×M
Maluma	MALU85	40	España	G×M[7]	G×M
Melendez 2	MELE12	41	EE.UU	G×WI[1]	G×WI
Mike	MIKE30	42	Sudáfrica	Desconocida	G
Monroe	MONR10	43	EE.UU	M×WI[1] o G×WI[1]	G×WI
MrsTooley	MRTO08	44	Sudáfrica	Desconocida	G×M×WI
Murrieta Green	MUGR27	45	Sudáfrica	G[4]	G
Nabal	NABA21	46	Sudáfrica	G[1]	G
Negra de la Cruz	NECR31	47	Sudáfrica	M[10]	G×M
Nimlioh	NIML09	48	Sudáfrica	G[1]	G
Nn10	NN1068	49	Sudáfrica	G[4]	G×M
NN63	NN6310	50	Sudáfrica	G[4]	G×M

Pinkerton	PINK45	51	Sudáfrica	G×M[2] o G[8]	G×M
Pollock	POLL6	52	EE.UU	WI[1]	WI
Reed	REED89	53	España	G[4]	G×M
Regal	REGA11	54	Sudáfrica	Desconocida	G×M
Rincon	RINC12	55	Sudáfrica	Desconocida	G×M
RR-86	RR8691	56	España	Desconocida	G×M×WI
Rustenburg Round	RURO36	57	Sudáfrica	Desconocida	G×M×WI
Russell	RUSS22	58	EE.UU.	WI[1]	WI
Ryan	RYAN13	59	Sudáfrica	G×M[1]	G×M
Semil 43	SEMI14	60	EE.UU.	G×WI[5]	G×WI
Shepard	SHEP42	61	Sudáfrica	G[4]	G×M
Teague	TEAG60	62	Sudáfrica	M[1,4]	G×M
Telez	TELE66	63	Sudáfrica	Desconocida	M×WI
Thomas	THOM90	64	Sudáfrica	M[2]	M×WI
Toro Canyon	TOCA96	65	Sudáfrica	M[2] o G×M[11]	G×M
Trapp	TRAP2	66	EE.UU.	WI[1]	WI
TX531	TX5344	67	Sudáfrica	Híbrido[1]o G[1]	G×M
Vero Beach nº 1	VERO4	68	EE.UU.	M×WI[1]	M×WI
Waldin	WALD28	69	EE.UU.	WI[1]	WI
Wester	WEST5	70	EE.UU.	WI[1]	WI
Yon	YON3	71	EE.UU.	G×WI[1]	G×WI

[1] Hofshi, Avocado database, 2019.
[2]Ashworth & Clegg 2003.
[3]Schnell *et al.* 2003.
[4]Variety Database of the University of California at Riverside, http://ucavo.ucr.edu/
[5]U.S. National Plant Germplasm System, https://npgsweb.ars-grin.gov/gringlobal/search. aspx?.
[6]Avocado information database, https://www.myavocadotrees.com/beta-avocado.html.
[7]Crane *et al*. 2013.
[8]Chen *et al*. 2009.
[9]Wolfe *et al.* 1949.
[10]Ben-Ya´cov *et al.* 2003.
[11]Gross-German & Viruel 2013.

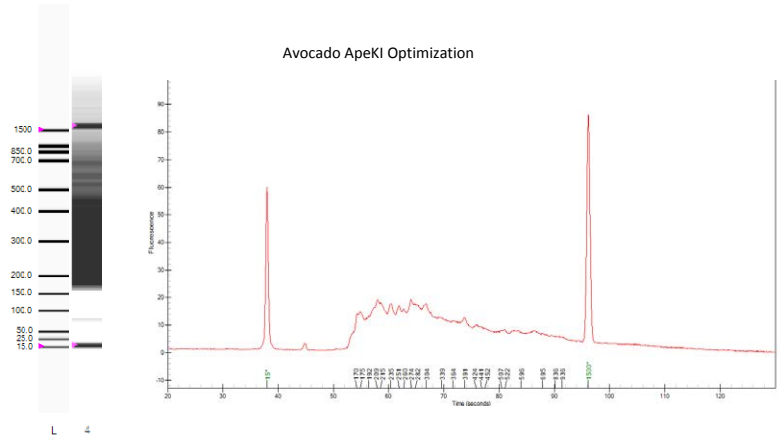
Extracción de ADN, preparación de genotecas, secuenciación y procesamiento de las lecturas en bruto

A partir de las hojas jóvenes previamente lavadas con agua destilada, se aisló ADN usando el Kit "DNesy Plant Mini Kit" de Qiagen siguiendo el protocolo descrito por el fabricante. La pureza (ratio 280/260 y 260/230) y la concentración del ADN (ng/ul) se determinaron usando el espectrofotómetro NanoDrop y el fluorómetro Qubit 2.0 del servicio de genómica de la Universidad de Málaga (SCBI).

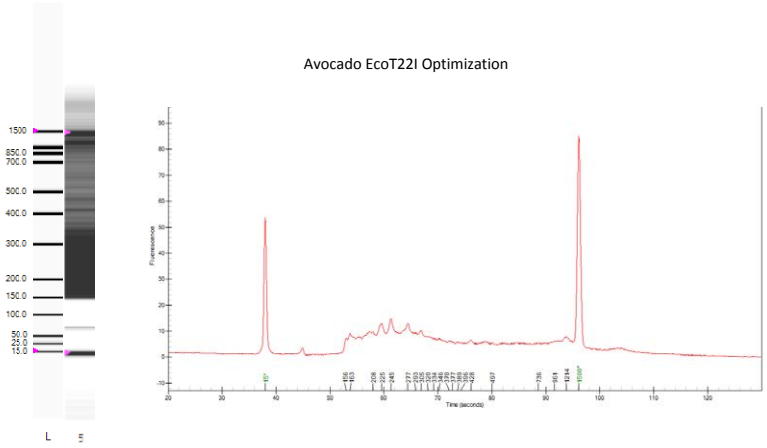
Para realizar la optimización de la enzima necesaria en el proceso de digestión del ADN y la construcción de las genotecas se utilizó ADN genómico de la variedad Hass, el cual se digirió con Pstl, EcoT221 y ApeKI. La distribución del ADN fragmentado se evaluó con Agilent 2100 Bioanalyzer System (Figura1.1).

Figura 1.1. Perfiles de tamaños moleculares de ADN genómico obtenido mediante un Agilent Bioanalyzer 2100 tras su digestión con ApeKI(A), EcoT221(B) y Pstkl(C), con el objetivo de seleccionar la mejor enzima para la reducción del genoma previo a su secuenciación como parte de un análisis de genotipado por secuenciación (GBS) de aguacate.

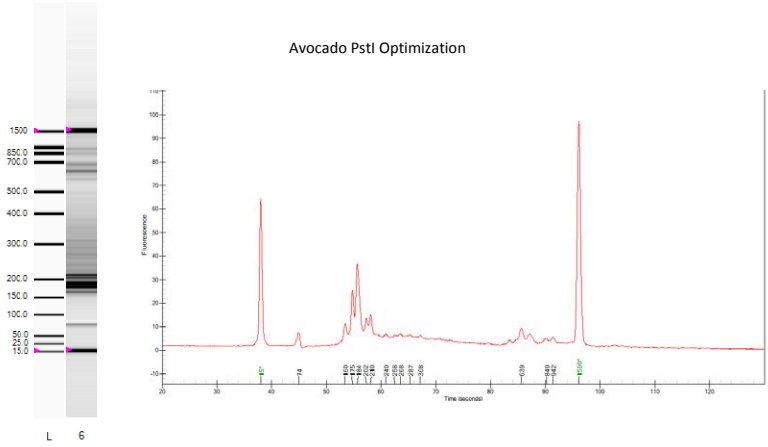
A



B



C



Las genotecas se construyeron siguiendo el protocolo descrito por Sonah *et al*. (2013), digiriendo 100 ng de ADN genómico de cada variedad con ApeKI (Anexo 1.1). Una de las genotecas se secuenció con Illumina HiSeq 2500 (1 × 100) en el "Center for Genomic and Computational Biology" de la Universidad de Duke (https://genome.duke.edu/), y otra se secuenció en la empresa Novogene (https://genome.duke.edu/) con Illumina HiSeq 4000 (2 × 150).

Las lecturas en bruto se demultiplexaron usando el paquete GBSx (Herten *et al.* 2015). Posteriormente, las lecturas se procesaron para eliminar posibles secuencias de adaptadores, lecturas con una longitud menor a 50 bases, y filtrar aquellas con regiones de baja calidad usando el software fastq-mcf versión 1.04.807 (-l 50 y -q 30) (Aronesty 2013).

Ensamblaje del borrador del genoma de aguacate (cv. Hass)

Con el fin de cartografiar las lecturas a un borrador del genoma de aguacate, se secuenció el ADN genómico de la variedad 'Hass' (2 × 150) con una profundidad de 100× usando Illumina. El tamaño del genoma y su heterocigosidad se estimaron usando la distribución de kmer descrita por Liu *et al.* (2013). En resumen, la distribución de Kmers de tamaños 19, 25, 31, 37, 43, 55, 61, 67, 73, y 85 se calculó con Jellyfish, y posteriormente el archivo generado se cargó en la página web GenomeScope (Vurture *et al.* 2017). Se usaron dos ensambladores distintos para ensamblar las lecturas de Illumina: Minia (Chikhi & Rizk 2013) y SOAPdenovo2 (Luo *et al.* 2012). Aunque ambos ensambladores emplean algoritmos para el ensamblaje de lecturas cortas, Minia requiere menos recursos computacionales que SOAPdenovo2 y filtra falsos positivos (Chikhi & Rizk 2013). Con ambos ensambladores se usó un rango de tamaños de Kmers desde 17 a 115. La comparación de los resultados de los cóntigos ensamblados mostró que el ensamblaje producido por Minia, con Kmer de 115, fue el que más contigüidad presentaba. Los cóntigos se agruparon para crear un supercóntigo usando SSPACE v 3.0 (Boetzer *et al*. 2011).

Cartografiado, identificación de SNPs y filtrado

Las lecturas generadas se cartografiaron usando bwa versión 0.7.120-r789 (Li & Durbin 2010) con los parámetros establecidos por defecto. Aquellas lecturas que no fueron cartografiadas se eliminaron usando Samtools versión 1.3.1 (Li *et al*. 2009), mientras que con las lecturas retenidas se crearon archivos BAM. Todos estos archivos BAM se unieron usando Bamaddrg (https://github.com/ekg/bamaddrg), y se usó el paquete Samtools version 1.3.1 (Li *et al.* 2009) para ordenar e indexar los archivos. Para detectar las variaciones, y eliminar SNPs con una calidad de mapeado menor que 20 y con una profundidad menor que 5, se usó el programa Freebayes versión 0.9.20 (Garrison & Marth 2012). Los polimorfismos brutos obtenidos se filtraron con el paquete VCFtools versión 0.1.12 (Danecek *et al*. 2011) eliminando los SNPs no bialélicos, los datos ausentes y SNPs que estuvieran dentro de una distancia de 1.000 pares de bases. Antes y después del filtrado se generaron resúmenes

estadísticos usando vcf-stats versión 0.1.12 (Danecek *et al.* 2011). Su diversidad se estudió usando el paquete Adegenet versión 2.1.1 (Jombart 2008) en Rstudio versión 1.1.453 (R Core Team 2018) y R versión 3.5.1, y se analizó el equilibrio de Hardy-Weinberg usando el paquete pegas versión 0.10 (Paradis 2010).

Todos los análisis de cartografiado, identificación de SNPs y filtrado se realizaron en un servidor del laboratorio del Prof. Aureliano Bombarely en la Universidad de Milán con 160 threads, 3 Tb de RAM y 22 Tb de disco duro (RAID 6).

Análisis molecular de la estructura genética de accesiones de aguacate

Con el objetivo de mostrar la utilidad de los marcadores moleculares tipo SNPs generados, se emplearon distintos enfoques para estudiar las relaciones y la estructura genética, junto a la divergencia entre los grupos establecidos a partir de las 71 accesiones de aguacate. Se realizó un análisis de componentes principales (ACP), seguido de un árbol de unión de vecinos (Neighbor-joining [NJ]), análisis discriminantes de componentes principales (ADCP) y agrupación bayesiana, así como la estimación de propiedades genéticas de las agrupaciones a través de parámetros como Fst, Fis, Pi y Watterson´s Theta.

El análisis de componentes principales (ACP) se llevó a cabo usando el paquete Adegenet versión 2.1.1 (Jombart 2008), y se visualizó usando el paquete ggplot2 version 3 (Wickham 2009) en Rstudio versión 1.1.453 y R versión 3.5.1 (R Core Team 2018).

La matriz de distancia Prevosti

$$DPrevosti(a,b) = \frac{1}{2r}\sum_{k=1}^{\upsilon}\sum_{j=1}^{m(k)}|Pajk - Pbjk| \qquad [1]$$

(donde υ es el número de loci considerado, Pajk la frecuencia de la disposición de alelos k en el locus j en la población a, y Pbjk el valor correspondiente en la población b [Prevosti 1975]) y el árbol de unión de vecinos (Neighbor-joining [NJ]) se generaron usando el paquete Poppr versión 2.8.2 (Kamvar *et al.* 2015; Kamvar *et al.* 2014) con 2.000 replicaciones de bootstrap. Las figuras se representaron usando el programa FigTree versión 1.4.4 (Rambaut 2009). Se implementó ADMIXTURE versión 1.3 (Alexander *et al.* 2009) con una iteración de K desde

1 a 20, con el objetivo de estudiar el grado de hibridación entre los diversos grupos. Este análisis se basa en un enfoque bayesiano, como el que realiza el programa STRUCTURE, y lleva a cabo una estimación de máxima verosimilitud de individuos a partir de los SNPs generados. Además, con el objetivo de elegir el número de óptimo de poblaciones (K), se usó un enfoque de validación cruzada para todos los SNPs. Cada valor de K elegido se representó en RStudio versión 1.1.453 y R versión 3.5.1 (R Core Team 2018). Con el mismo objetivo se usó también el software STRUCTURE: se aplicó 5 veces por cada número de poblaciones (K), en cada uso se estableció un periodo de entrenamiento de 20.000 pasos seguido de 200.000 cadenas de repetición de Markov de Monte Carlo (Larranaga *et al.* 2017; Martín *et al*. 2011; Pritchard *et al*. 2010). El método de Evanno *et al.* (2005) se empleó para determinar el número más probable de poblaciones con el software STRUCTURE HARVESTER (Earl & vonHoldt 2012).

Posteriormente, teniendo en cuenta que enfoques como STRUCTURE asumen que los marcadores no están ligados y que las poblaciones son panmícticas (Pritchard *et al*. 2000), se llevaron a cabo análisis discriminantes de componentes principales (ADCP), con el objetivo de identificar grupos bien definidos usando el paquete de R Adegenet versión 2.1.1 (Jombart 2008). Para realizar este análisis, los datos se tuvieron que transformar previamente usando ACP y se usó la función find.clusters para identificar el número de grupos. Igualmente, se calculó el criterio de información bayesiano (BIC) para asociar el número correcto de subgrupos, y se usó la función de la validación cruzada (XvalDapc) para corroborar el mejor número de ACP retenido. Antes de llevar a cabo estos análisis, los archivos se leyeron con la función read.vcf, y se convirtieron en objetos genind y genlight con las funciones: VcfR2genind y Vcf2genlight. Todos los análisis que se llevaron a cabo usando Rstudio versión 1.1.453 y R versión 3.5.1 (R Core Team 2018) se encuentran disponibles en un repositorio público de Github (https://github.com/IHSMFruitCrops/Hass-genotyping).

Finalmente, para analizar la diferenciación de los grupos establecidos, se estimó el índice de fijación (Fst), que permite diferenciar las poblaciones con un rango entre 0 (no diferenciación) y 1 (diferenciación completa) (Hahn 2018) con el paquete de R PopGenome versión 2.6.1 (Pfeifer *et al.* 2014). Con el fin de observar si existía una reducción de la heterocigosidad debido al cruzamiento no aleatorio dentro de cada grupo, se estimó el coeficiente de endogamia (Fis) usando el paquete hierfstat versión 0.04-22 (Goudet 2005). Además, se estimaron los estadísticos Pi y Watterson´s theta para calcular la diversidad

nucleotídica considerando la agrupación establecida por ADCP K=3, K=4 y K=5 utilizando el mismo paquete de R.

Todos los análisis de R se ejecutaron en un servidor del Prof. Antonio J. Matas en la Universidad de Málaga con Intel i7 3930K (6 cores, 12 threads), 64 GB de memoria RAM y 6 x 2 TB Seagate en RAID 1.

Disponibilidad de datos

Las lecturas brutas del borrador del genoma de 'Hass' se han depositado en NCBI en el BioProjecto PRJNA564097, mientras que el conjunto de datos generados mediante GBS se depositaron en PRJNA564105. La mayoría de los análisis se llevaron a cabo usando R 3.5.1. y todos los scripts se han depositado en un repositorio público de GitHub: https://github.com/IHSMFruitCrops/Hass-genotyping.

Resultados

Desarrollo del borrador del genoma de aguacate para el cartografiado de las lecturas en bruto

En primer lugar, se desarrolló un borrador del genoma de aguacate cv. Hass con el fin de poder cartografiar las lecturas generadas y de conocer la posición de los SNPs que se encuentran a lo largo del genoma. Con la secuenciación usando illumina de ADN genómico del cv. Hass se generaron 487,54 millones de lecturas brutas (73,13 Gb) y 487,21 millones de lecturas procesadas (72,15 Gb). La estimación del tamaño haploide del genoma oscila entre 1,33 Gb y los 1,63 Gb con una estimación de heterocigosidad que varía entre el 1,05 % y el 1,41 %. El ensamblaje representa el 77 % del tamaño estimado del genoma (1,33 Gb). El número de secuencias totales indican que el ensamblaje es altamente fragmentado con una secuencia media de 0,54 Kb y un L50 (número más pequeño de cóntigos cuya longitud total constituye la mitad del tamaño del genoma) de 0,68 Kb. Este valor es inferior a la longitud media de un gen en plantas (por ejemplo: 2,01 Kb para *Arabidopsis thaliana*) y, consecuentemente, no se pudo realizar la anotación estructural de genes (Wortman *et al.* 2003) aunque si facilitó el cartografiado de las lecturas generadas. Un resumen estadístico del borrador del genoma se muestra en la Tabla 1.2.

Tabla 1.2. Resumen del ensamblaje del borrador del genoma de *Persea americana* Mill. cv. Hass.

Estadística de ensamblaje	Cóntigos (*Contigs*)[1]	Supercóntigos (*Scaffolds*)[2]
Tamaño total de ensamblaje (Gb)	1,03	1,01
Total de secuencias ensambladas	2.096.006	1.852.224
Longitud de la secuencia más larga (Kb)	57,80	160,08
Media de la longitud de secuencia (Kb)	0,49	0,54
Índice N50 (secuencias)	475.145	377.224
Longitud L50 (Kb)	0,56	0,68

[1]Cóntigo o *contig*: región consenso de ADN.
[2]Supercóntigo o *scaffold*: unión de cóntigos usando información complementaria sobre la orientación y posición en el genoma.

Genotipado por secuenciación (GBS) y llamamiento de variantes

Las 71 accesiones de aguacate (Tabla 1.1) se genotiparon mediante dos genotecas secuenciadas con Illumina Hiseq 2500 (1x100) e Illumina Hiseq 4000 (2x150) lo que produjo 405,93 millones de lecturas totales o brutas. Tras el procesamiento de dichas lecturas se obtuvieron 345,37 millones de lecturas filtradas (Figura 1.2). Estas se compararon con el genoma de referencia mediante cartografiado y se conservaron solo aquellas localizadas en una posición única, representando estas alrededor del 80 % del total. Finalmente, se detectaron 1.070.902 variaciones entre las que se identificaron 945.064 SNPs, de los cuales 22.321 fueron indeles (inserciones y delecciones), 69.500 MNPs (polimorfismos multinucleóticos) y 6.604 combinaciones entre las variaciones anteriores.

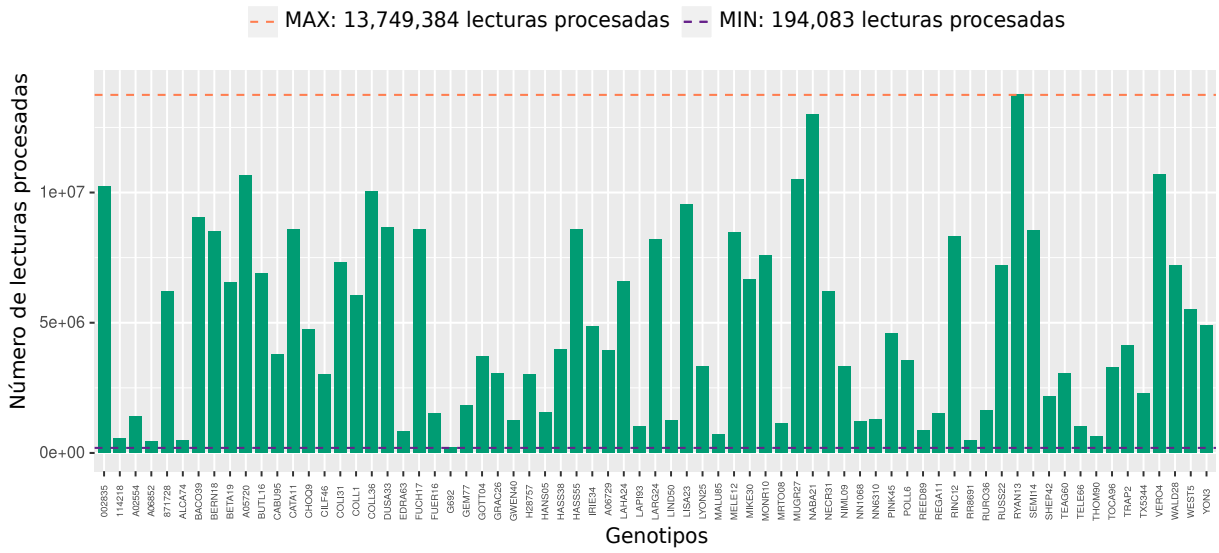


Figura 1.2. Número de lecturas procesas por accesión. La línea naranja y azul representan el máximo y mínimo de lecturas procesadas respectivamente.

Desarrollo de SNPs

Tras el filtrado, se detectaron 7.108 SNPs para las 71 accesiones de aguacate (Tabla 1.1) sin datos ausentes, de los cuales 19,45 % fueron polimorfismos privados (Tabla 1.3), es decir, que solo aparecen en una población y están ausentes en las otras. El total de los polimorfismos se categorizó conforme a las sustituciones nucleotídicas: 61,04 % fueron transiciones (C/T [2.195] o A/G [2144]) y 38,96 % fueron transversiones (A/C [778], C/G [646], A/T [666], G/T [679]). La proporción transición/transversión fue 1,57, similar a resultados previos en otras especies (Soorni *et al*. 2017b; Shearman *et al*. 2015; Pootakham *et al*. 2015). La media de heterocigosidad observada fue 0,16 mientras que la media de la heterocigosidad esperada fue 0,17. La media de la frecuencia del alelo menos común fue 0,11.

Tabla 1.3. SNPs privados de las 71 accesiones de aguacates analizados en este trabajo.

Identificador de las muestras	SNPs privados
2835	6
A06852	1
871728	1
BACO39	17
BETA19	5
BUTL16	5
CABU95	41
CATA11	6
CHOQ9	7
COLL1	6
COLI31	2
COLL36	75
DUSA33	21
FUCH17	6
G692	2
GEM77	1
GOTT04	41
H28757	104
HANS05	246
HASS38	1
IRIE34	8
A06729	8
LAHA24	1
LAPI93	13
LARG24	338
LIND50	9
LISA23	12
MELE12	4
MIKE30	4
MONR10	4
MRTO08	28
MUGR27	8
NABA21	18
NECR31	81
NIML09	1
POLL6	3
REED89	8
RURO36	21
RUSS22	10

RYAN13	11
SEMI14	3
TEAG60	11
TELE66	37
THOM90	42
TOCA96	64
TRAP2	1
TX5344	2
VERO4	28
WALD28	1
WEST5	8
YON3	2
TOTAL	1383

Diversidad y estructura poblacional usando SNPs filtrados

Se llevaron a cabo diversos análisis para estudiar las relaciones entre las accesiones de aguacate estudiadas.

Como primera aproximación se realizó un análisis de componente principales (ACP), teniendo en cuenta todos los SNPs, con el objetivo de estudiar la estructura genética de las accesiones (Figura 1.3). Los dos primeros componentes principales explicaron más del 40% de la variación (26,1 % y 15,1 % respectivamente). El análisis reveló tres grupos principales que representan las tres razas hortícolas definidas para el aguacate y, tal y como se esperaba, las accesiones híbridas se observaron entre los tres grupos principales.

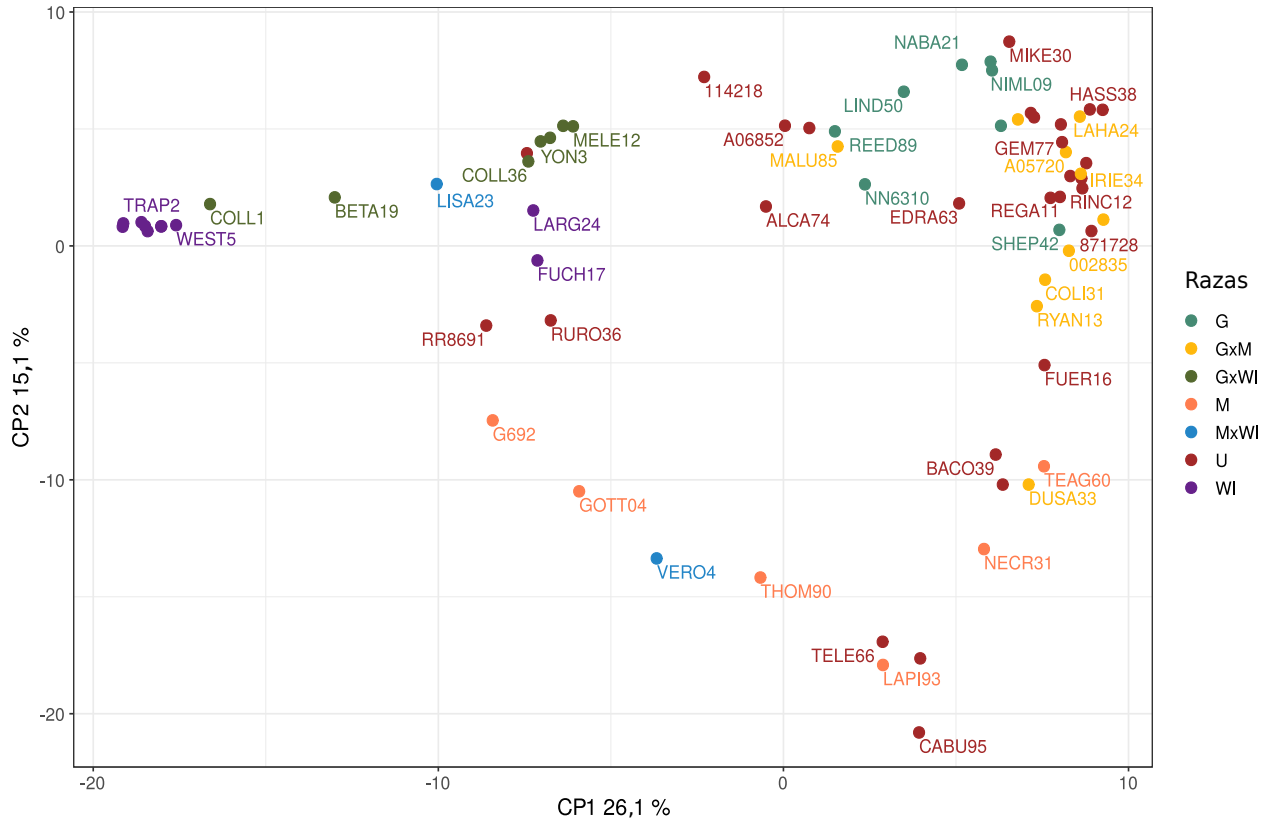


Figura 1.3. Análisis de componentes principales de 71 accesiones de aguacate usando 7.108 SNPs. Cada genotipo se encuentra representado por su identificador (Tabla 1.1). Los colores indican la raza según la literatura: G, GxM, GxWI, M, MxWI, U y WI (G: Guatemaltecos; M: Mexicanos; WI: Antillanos; U: Desconocido).

Como segunda aproximación, se usó la distancia de Prevosti [1] (Prevosti *et al*. 1975) para evaluar la estructura genética de las accesiones y determinar las diferencias entre los genotipos estudiados. Estas diferencias se analizaron con el método de unión de vecinos (Neighbor-Joining [NJ]) para mostrar las relaciones entre los genotipos y generar un dendrograma (Figura. 1.4.a). En el dendrograma se muestran dos grupos principales con un valor de bootstrap poco significativo (27,8) (Figura 1.4.a). Uno de los grupos se encuentra formado por un subgrupo altamente respaldado (con un valor de bootstrap de 71,8) en el que se incluyen principalmente genotipos híbridos Guatemaltecos x Mexicanos (GxM) (‘Pinkerton’, ’Lyon’, ’Iriet’, ’Gem’, ’Hass’, ’Lamb Hass’, entre otros), algunos genotipos descritos como Mexicanos (‘Teague’, ’Negra de la Cruz’), otros genotipos considerados Guatemaltecos (‘Shepard’), y un genotipo de origen desconocido (‘TX531’)(Figura 1.4.a).

En otro subgrupo (con un valor de bootstrap de 38,1) se incluyen principalmente accesiones consideradas Guatemaltecas (’Reed’, ‘Nabal’, ‘Nimlioh’, ‘Linda’, ‘Murietta Green’), junto a varios genotipos de raza desconocida (‘A0-67’, ‘Mike’, ‘Mrs Tooley’) (Figura 1.4.a). Otros dos genotipos descritos como Guatemaltecos (‘NN10’,’NN63’) formaron un grupo altamente respaldado (con un valor de bootstrap de 67,6), y ‘Maluma’ y ‘Alcaraz’, aparecieron separados de estos (Figura 1.4.a).

En el segundo grupo principal, se encuentran dos genotipos de origen desconocido (‘A0.68’ y ‘1.14.2’) y un grupo altamente respaldado (con un valor de bootstrap de 80,5) compuesto por dos subgrupos. En uno de ellos (con un valor de bootstrap de 85,9), se englobaron genotipos considerados Mexicanos (‘G-6’, ‘Thomas’, ‘Gottfried’), un híbrido Mexicano x Antillano (MxWI) (‘Vero Beach nº1’), así como genotipos de raza desconocida (‘RR-86’, ‘Telez’, ‘Rustenburg Round’, ‘C.A. Bueno’ y ‘Hansie’). El otro subgrupo se mostró poco respaldado (con un valor de bootstrap de 26,1) y compuesto por dos subgrupos. En uno de ellos (con un valor de bootstrap de 29,1), se incluyen principalmente genotipos Antillanos (‘Pollock’, ‘Bernecker’, ‘Waldin’, ‘Russel’, ‘Catalina’, ‘Butler’, ‘Wester’, ‘Trapp’,‘Fuchsia’, ’Largo’) junto con algunos híbridos Guacemalteco x Mexicano (GxM) (‘Beta’, ‘Collinred B’) y un híbrido Mexicano x Antillano (MxWI)(‘Lisa’). El otro subgrupo se mostró también poco soportado (con un valor de bootstrap de 52,6), y en él se incluyen híbridos Guatemaltecos x Antillanos (GxWI) (‘Yon’, ‘Choquette’, ‘Collinson’,‘Melendez 2’,‘Semil 43’) junto a un híbrido Mexicano por Antillano (MxWI) (‘Monroe’).

Tras el análisis de componentes principales se llevó a cabo un análisis de mezcla genética usando el software ADMIXTURE (Alexander *et al*. 2009). El número de grupos más probable fue 4, seguido de 3 y 5, aunque se observó una leve diferencia entre los posibles grupos con un valor de validación cruzada entre 0,28 y 0,29. Para K=4, la división mostrada entre los genotipos Mexicanos, Guatemaltecos y Antillanos fue evidente, aunque se formó otro grupo con genotipos híbridos ‘GxM’ (Figura. 1.4.b). Con el objetivo de tener una visión más robusta de la estructura genética poblacional se utilizaron los programas STRUCTURE (Pritchard *et al.* 2000) y STRUCTURE HARVESTER (Earl & vonHoldt 2012). Los resultados obtenidos mostraron que K=4 era el número de poblaciones más probable (Figura 1.5.a y 1.5.b; Figura 1.6b), coincidiendo con el análisis anterior, pero, en este caso, las accesiones consideradas como Guatemaltecas y las accesiones consideradas híbridas entre Guatemaltecas y Mexicanas no se diferenciaron claramente.

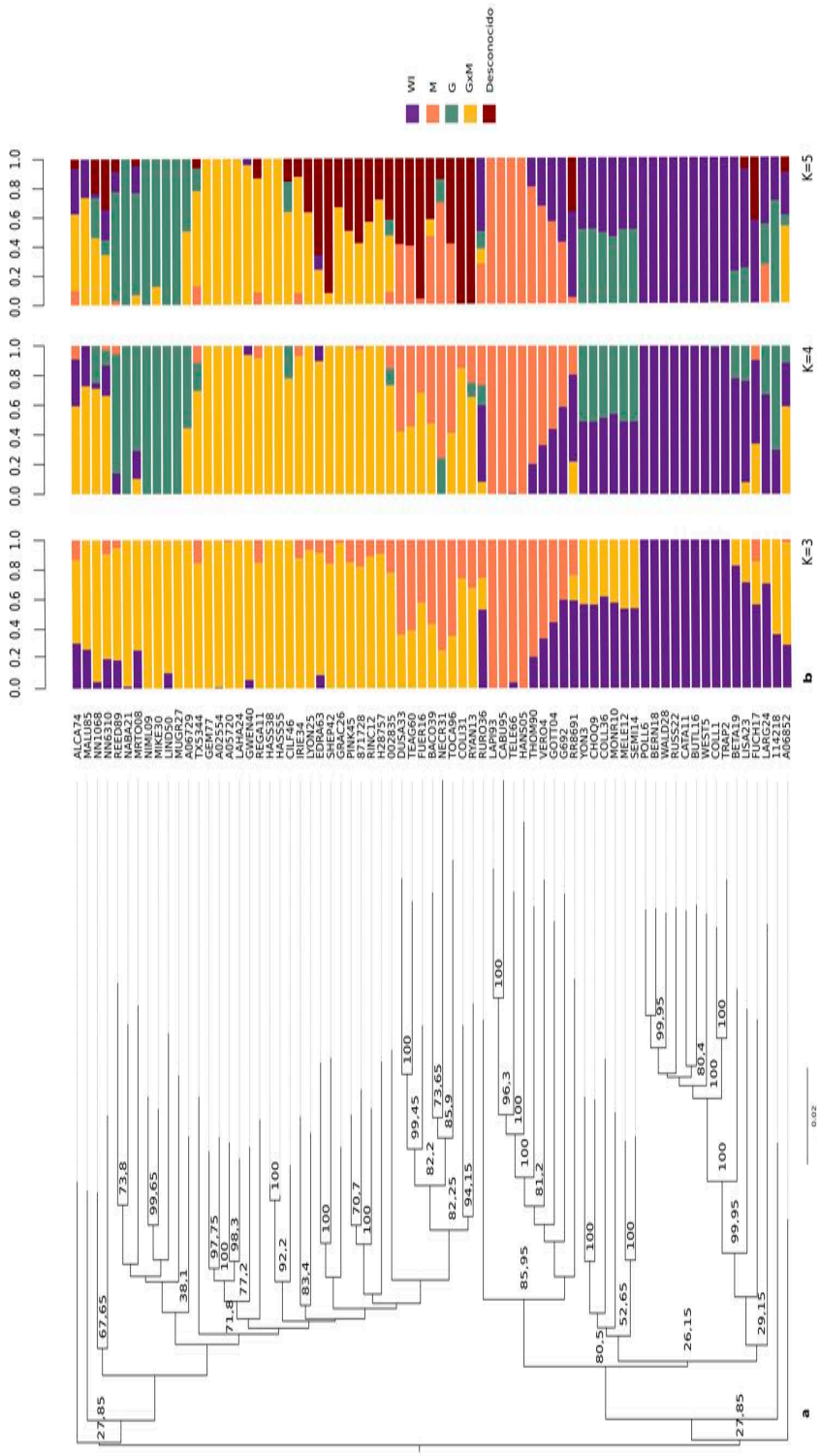


Figura 1.4.(a) Dendrograma basado en neighbor joining (NJ) mostrando las relaciones entre los 71 genotipos de aguacate. En cada nodo se muestra el valor de bootstrap citado en el texto o aquellos valores superiores al 70 % tras 2.000 réplicas. La figura se representó con el software Figtree (Rambaut *et al*. 2009). (b) Gráfico de barras mostrando la estratificación de la población para el número de poblaciones más probable, K=4 seguido de K=3, y K=5 obteniendo con ADMIXTURE (Alexander et al. 2009). Esta figura fue generada usando R versión 3.5.1. (R core team 2018). Las accesiones fueron ordenadas según el orden mostrado en el dendrograma. Para K=4, los colores indican la raza según la literatura: M , G , GxM y WI (G = Guatelmaltecos, M = Mexicanos y WI = Antillanos).

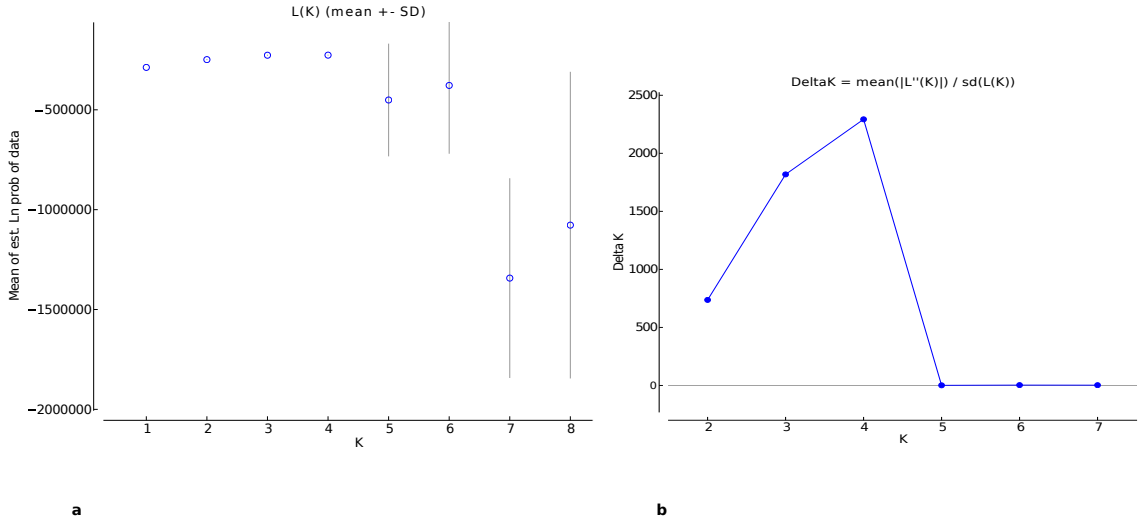


Figura 1.5. (a) Probabilidad para K=1-8. (**b)** Estimación de número de grupos usando el método propuesto por Evanno *et al*. (2005).

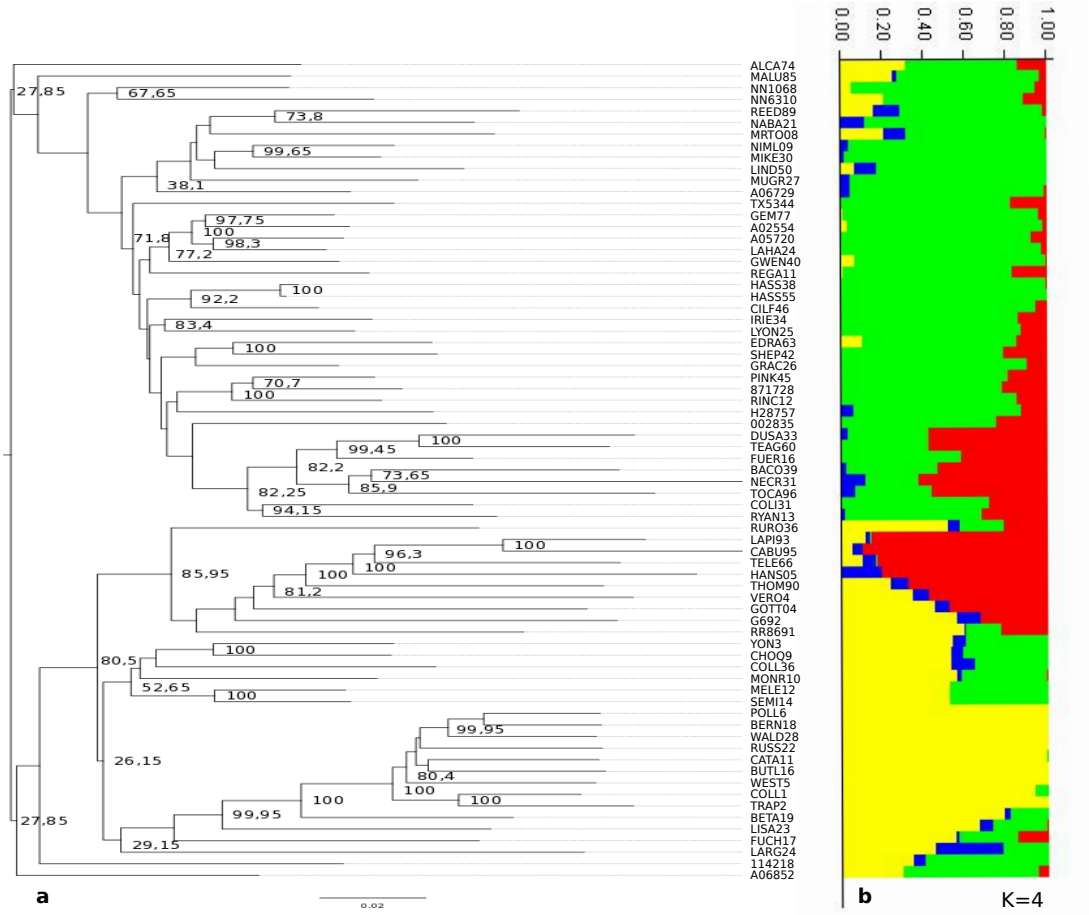


Figura 1.6. (a) Dendrograma de 71 cultivares de aguacates basados en métodos de unión de vecinos (NJ). En cada nodo se muestra el valor de bootstrap citado en el texto o aquellos valores superiores al 70% tras 2.000 réplicas. La figura se representó en el software Figtree (Rambaut *et al.* 2009). **(b)** Gráfico de barras describiendo la estratificación de la población para el número de población más probable, K=4 generado usando STRUCTURE (Pritchard *et al*. 2000).

Con el objetivo de describir la diversidad entre grupos predefinidos se realizó un análisis discriminante de componentes principales, esperando obtener el número de grupos más probable. Estos resultados fueron congruentes con el error de la validación cruzada (ADMIXTURE) y con el algoritmo de Evanno (STRUCTURE) respecto al número de grupos (K), ya que 4 se mostró de nuevo como el número más probable de grupos (Figura. 1.5). Para K=3, las accesiones se dividieron de acuerdo a otros métodos (ADMIXTURE y STRUCTURE). Un primer grupo incluyó principalmente accesiones de raza Guatemalteca e híbridos GxM. Un segundo grupo incluyó accesiones de raza Antillana, híbridos GxWI y MxWI, mientras que el tercer grupo incluyó accesiones de raza Mexicana, híbridos GxM e híbridos MxWI. Para K=4, las accesiones con componente Antillano se separaron en dos

grupos, uno incluyó principalmente genotipos Antillanos, y otro grupo incluyó híbridos entre Guatemaltecos y Antillanos. Para K=5, los genotipos Guatemaltecos e híbridos entre Guatemaltecos y Mexicanos se dividieron en diferentes agrupaciones (Tabla 1.4).

Tabla 1.4. Clasificación de cada muestra por grupo (K=2-4) establecidos a través del ADCP.

K=3		**K=4**		**K=5**	
Grupo 1	COLI31	**Grupo 1**	YON3	**Grupo 1**	YON3
	HASS38		CHOQ9		CHOQ9
	REGA11		MONR10		MONR10
	MALU85		MELE12		MELE12
	RINC12		SEMI14		SEMI14
	RYAN13		FUCH17		FUCH17
	EDRA63		LISA23		LISA23
	GWEN40		LARG24		LARG24
	REED89		COLL36		COLL36
	FUER16		114218		RR8691
	MRTO08		RR8691		G692
	SHEP42		G692		RURO36
	114218		RURO36	**Grupo 2**	COLL1
	NN1068		ALCA74		TRAP2
	TX5344	**Grupo 2**	COLI31		WEST5
	A05720		HASS38		POLL6
	PINK45		REGA11		CATA11
	NABA21		MALU85		BUTL16
	CILF46		RINT12		BERN18
	LAHA24		RYAN13		BETA19
	NIML09		EDRA63		RUSS22
	LYON25		GWEN40		WALD2
	ALCA74		REED89	**Grupo 3**	MALU85
	GRAC26		FUER16		REED89
	LIND50		MRTO08		MRTO08
	MUGR27		SHEP42		114218
	NN6310		NN1068		NN1068
	871728		TX5344		NABA21
	A06852		A05720		NIML09
	GEM77		PINK45		ALCA74
	A06729		NABA21		LIND50

	MIKE30		CILF46		MUGR27
	A02554		LAHA24		NN6310
	H28757		NIML09		A06852
	HASS55		LYON25		MIKE30
	IRIE34		GRAC26	**Grupo 4**	COLI31
	2835		LIND50		HASS38
Grupo 2	COLL1		MUGR27		REGA11
	TRAP2		NN6310		RINT12
	YON3		871728		RYAN13
	WEST5		A06852		EDRA63
	POLL6		GEM77		GWEN40
	CHOQ9		A06729		FUER16
	MONR10		MIKE30		SHEP42
	CATA11		A02554		TX5344
	MELE12		H28757		A05720
	SEMI14		HASS55		PINK45
	BUTL16		IRIE34		CILF46
	FUCH17		2835		LAHA24
	BERN18	**Grupo 3**	COLL1		LYON25
	BETA19		TRAP2		GRAC26
	RUSS22		WEST5		871728
	LISA23		POLL6		GEM77
	LARG24		CATA11		A06729
	WALD28		BUTL16		A02554
	COLL36		BERN18		H28757
	RR8691		BETA19		HASS55
	G692		RUSS22		IRIE34
	RURO36		WALD28		2835
Grupo 3	VERO4	**Grupo 4**	VERO4	**Grupo 5**	VERO4
	DUSA33		DUSA33		DUSA33
	BACO39		BACO39		BACO39
	TELE66		TELE66		TELE66
	THOM90		THOM90		THOM90
	LAPI93		LAPI93		LAPI93
	CABU95		CABU95		CABU95
	TOCA96		TOCA96		TOCA96
	GOTT04		GOTT04		GOTT04
	NECR31		NECR31		NECR31
	HANS05		HANS05		HANS05
	TEAG60		TEAG60		TEAG60

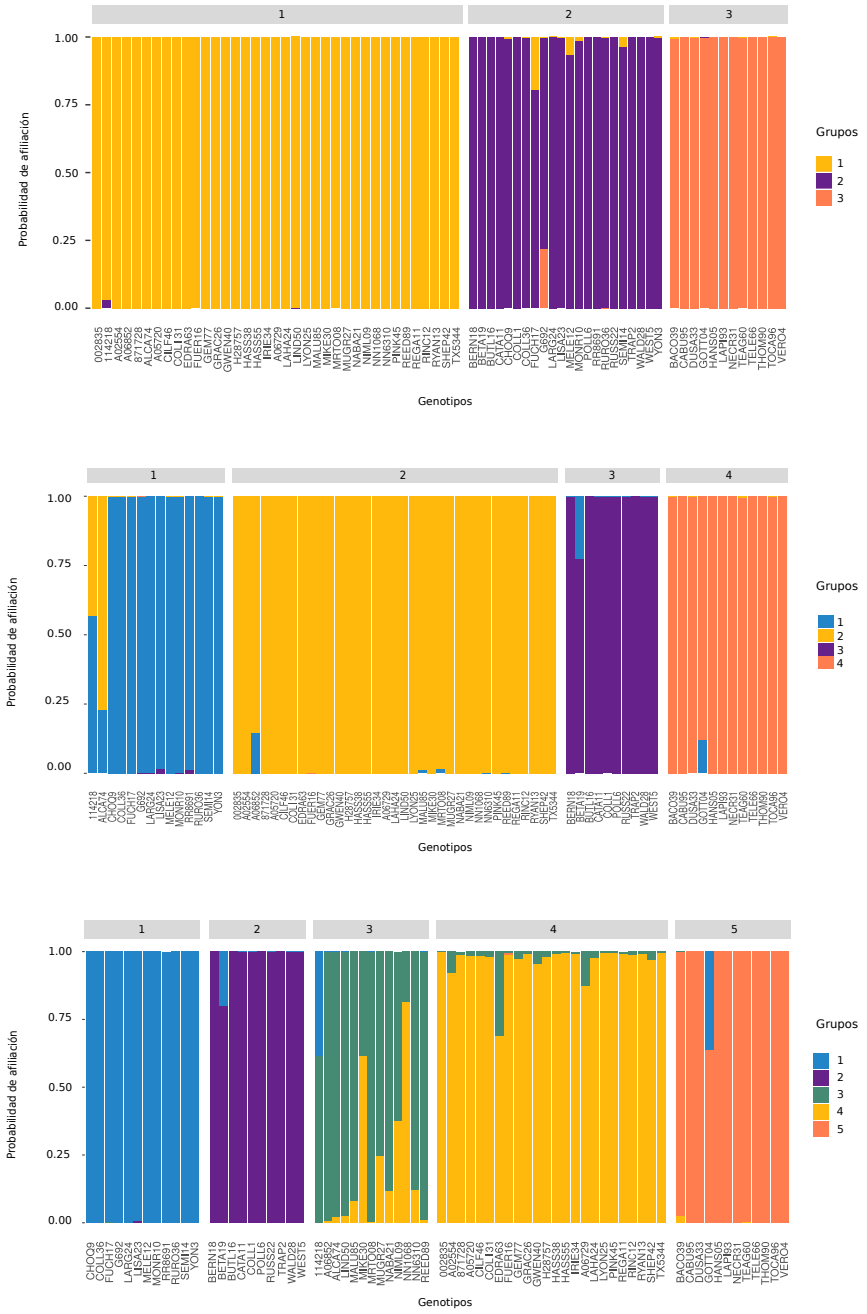


Figura. 1.5. Análisis discriminante de componentes principales (ADCP) mostrando la estructura poblacional para K=3-5 (obtenido con la función find. clusters) (Tabla S3). Cada genotipo es una barra en el eje de abscisas, y la probabilidad de la asignación de un genotipo a una población se muestra en el eje de ordenadas. Cada población se muestra en colores diferentes. En general, para K=3, grupo 1 = GxM, grupo 2 = WI, grupo 3 = M; para K=4, grupo 1 = GxWI y MxWI, grupos 2 = GxM , grupo 3 = WI, grupo 4 = M; para K = 5, grupo 1 = GxWI y MxWI, grupo 2 = WI, grupo 3 = G, grupo 4 = GxM, grupo 5 = M. Figura representada usando el software R versión 3.5.1. (R core team 2018).

Con el objetivo de validar los grupos definidos anteriormente se calculó el índice de fijación (Fst). Este índice se aplicó a todos los grupos predefinidos mediante el análisis discriminante de componentes principales (K= 3-5) (Tabla 1.4). En todos los casos se mostraron diferencias entre los grupos corroborando los análisis previos. Para K=4, el valor más bajo fue 0,18 entre el grupo 2 (principalmente formado por genotipos considerados híbridos entre Guatemaltecos y Mexicanos, y algunos considerados Guatemaltecos) y el grupo 1 (compuesto principalmente por híbridos entre accesiones Guatemaltecas y Antillanos). El valor más elevado fue 0,61 entre el grupo 3 (compuesto principalmente por genotipos Antillanos) y grupo 2 (principalmente genotipos considerados híbridos entre Guatemaltecos y Mexicanos, y algunos considerados Guatemaltecos) (Tabla 1.5).

Tabla 1.5. Diferenciación genética (Fst) de 71 accesiones de aguacate agrupadas para K=4. La raza más representada por grupo se muestra entre paréntesis.

	Grupo1 (GxWI)	**Grupo2** (G) + (GxM)	**Grupo3** (WI)	**Grupo4** (M)
Grupo1 (GxWI)	0	0,18	0,39	0,23
Grupo2 (G) + (GxM)	0,18	0	0,61	0,33
Grupo3 (WI)	0,39	0,61	0	0,48
Grupo4 (M)	0,23	0,33	0,48	0

Se estudió la diversidad nucleotídica para cada grupo con diferentes índices (Pi y Watterson´s Theta) (Tabla 1.6). Para K=4, Pi osciló desde 270,14 a 515,27, y Watterson´s Theta varió desde 304,74 a 471,15. La mayor diversidad nucleotídica se obtuvo en el grupo compuesto principalmente por genotipos Mexicanos, seguidos del grupo compuesto por híbridos entre Guatemaltecos y Antillanos (GxWI). En cambio, en el grupo compuesto principalmente por híbridos Guatemaltecos y Mexicanos (GxM), se apreció una menor diversidad.

Tabla 1.6. Diversidad nucleotídica según la estructura poblacional (K=3, K= 4 y K= 5) establecida a través del ADCP. Las accesiones que pertenecen a cada grupo se especifican en la tabla 1.4. La raza más representada por grupo se muestra entre paréntesis.

	Grupos	**Número de accesiones**	**Pi**	**Watterson´s Theta**
K=3	1 (GxM)	37	273,65	307,58
	2 (WI)	22	543,69	521,76
	3 (M)	12	515,27	471,15
K=4	1 (GxWI)	14	419,23	467,9
	2 (GxM)	35	270,14	304,74
	3 (WI)	10	417,75	434,08
	4 (M)	12	515,27	471,15
K=5	1 (GxWI)	12	420,06	458,96
	2 (WI)	10	417,75	434,08
	3 (G)	13	293,23	303,88
	4 (GxM)	24	234,76	264,03
	5 (M)	12	515,27	471,15

También se analizó la diversidad genética de los grupos establecidos por el análisis discriminante de componentes principales, el coeficiente de endogamia (Fis) y la frecuencia del alelo menos común. La mayor heterocigosidad observada y esperada se mostró en el grupo formado por genotipos principalmente Mexicanos y, en el caso de la frecuencia del alelo menos común, también se indicaron los valores más altos en este grupo; sin embargo, se observó poca diferencia respecto al grupo representado principalmente por genotipos híbridos GxWI. Del mismo modo, al considerarse el coeficiente de endogamia se pudo apreciar que el grupo con genotipos principalmente Mexicanos y el grupo compuesto por híbridos GxWI son los que menor grado de endogamia mostraron (Tabla 1.7).

Tabla 1.7. Proporción de heterocigosidad observada (Ho), proporción de heterocigosidad esperada (He), coeficiente de endogamia (Fis) y frecuencia del alelo menos común para K=3, K=4 y K=5. La raza más representada por grupo se muestra entre paréntesis.

	Grupos	Número de accesiones	Proporción de heterocigosidad observada (Ho)	Proporción de heterocigosidad esperada (He)	Coeficiente de endogamia (Fis)	Frecuencia media del alelo menos común
K=3	1(GxM)	37	0,14	0,13	-0,13	0,08
	2(WI)	22	0,15	0,14	-0,09	0,10
	3(M)	12	0,20	0,16	-0,15	0,11
K=4	1(GxWI)	14	0,19	0,16	-0,18	0,11
	2(GxM)	35	0,14	0,12	-0,14	0,08
	3(WI)	10	0,10	0,09	-0,06	0,07
	4(M)	12	0,2	0,16	-0,15	0,11
K=5	1(GxWI)	12	0,19	0,16	-0,19	0,11
	2(WI)	10	0,10	0,09	-0,06	0,07
	3(G)	13	0,14	0,13	-0,09	0,10
	4(GxM)	24	0,14	0,12	-0,23	0,10
	5(M)	12	0,20	0,16	-0,15	0,11

Asignación de genotipos de pedigrí desconocido al grupo establecido

Basado en los análisis anteriores, se pudieron asignar a grupos raciales genotipos de origen confuso o desconocido. Entre los genotipos conocidos con asignaciones raciales ambiguas encontramos 'Bacon', 'Edranol', 'Fuerte', 'Gem', 'Gwen', 'Hass', 'Lyon', 'Pinkerton', 'Toro Canyon' y 'TX531', que han sido considerados por diferentes autores como Mexicanos puros (Chen *et al*. 2009), Guatemaltecos (Ashworth & Clegg 2003; Crane *et al.* 2013) o híbridos GxM (Schnell *et al.* 2003; Ashworth & Clegg 2003; Crane *et al*. 2013)(Tabla 1.1).

Los resultados obtenidos tras realizar los análisis de ADMIXTURE sugieren que todos ellos son híbridos GxM, aunque en 'Edranol' se encontró también un componente Antillano. Algunos genotipos cuyo origen era desconocido ('A0.25', 'A0.68', '87.17.1', '1.14.2' y 'Alcaraz') parecen ser híbridos GxM, aunque algunos posiblemente son híbridos de las tres razas con baja proporción de herencia Antillana. Otras accesiones ('Mike' y 'Mrs Tooley') parecen ser Guatemaltecas puras, mientras que otras ('Hansie' y 'C.A. Bueno') aparecen como Mexicanas puras.

Discusión

Aunque numerosos programas de mejora se han beneficiado de los nuevos enfoques de genotipado, estos avances son más lentos para la mayoría de plantas perennes leñosas, principalmente para cultivos tropicales y subtropicales, debido al gran vacío de información genómica disponible. Respecto al aguacate, a pesar de que varios programas de mejora están activos, y se han desarrollado distintos tipos de marcadores moleculares en las últimas décadas (Furnier *et al.* 1990; Lavi *et al.* 1991, 1994; Sharon *et al*. 1997; Borrone *et al.* 2007; Alcaraz & Hormaza 2007; Chen *et al*. 2008, 2009; Gross-German & Viruel 2013; Guzmán *et al.* 2017; Kuhn *et al.* 2019; Ge *et al.* 2019a,2019b; Rubinstein *et al.* 2019), sigue existiendo la necesidad de generar marcadores adicionales que puedan usarse a gran escala. De este modo, el uso de nuevas metodologías como la secuenciación de alto rendimiento puede llenar ese vacío con el objetivo de acelerar la mejora y el conocimiento del genoma del aguacate como ha ocurrido en otros cultivos.

Borrador del genoma de 'Hass' para el análisis de diversidad

Con el objetivo de llevar a cabo este estudio, se desarrolló un genoma fragmentado de aguacate (cv. Hass) con pequeños cóntigos. Esta fragmentación presenta varias limitaciones para estudios genómicos, como la imposibilidad de anotar el genoma y, consecuentemente, su uso para el descubrimiento de genes. Aun así, este borrador de genoma permite la alineación de lecturas producidas mediante "metodologías de representación reducida" y la obtención de un mayor número de marcadores moleculares. La disponibilidad de programas que no necesitan un genoma de referencia como Stacks (Catchen *et al*. 2011), TASSEL-UNEAK (Lu *et al*. 2013) y GBS-SNP-CROP (Melo *et al.* 2016) permite la identificación de variaciones, pero también aumenta las posibilidades de hacer un llamamiento erróneo (Lu *et al.* 2013; Leggett & MacLean 2014; Berthouly-Salazar *et al*. 2016). Por este motivo, el uso de un genoma de referencia, a pesar de que sea fragmentado, puede reducir este tipo de errores.

Al comienzo de este trabajo no se disponía de ningún genoma de aguacate publicado. Sin embargo, durante su desarrollo, se publicó otro genoma de la variedad Hass por Rendón-Anaya *et al*. (2019) junto al genoma de una variedad Mexicana. En este caso se generaron secuencias de larga longitud que, corregidas, facilitaron la generación de un genoma menos fragmentado (con 2811,28 Kb como secuencia más larga respecto a los 57,80 Kb presentados en el genoma de este trabajo). No obstante, a pesar de los esfuerzos realizados hasta ahora, se

necesita seguir trabajando en la construcción de un genoma de referencia de mayor calidad. De hecho, el borrador del genoma de 'Hass' presentado en este capítulo es un punto de partida para ello.

Aunque algunos estudios previos han desarrollado marcadores SNP en aguacate (Chen *et al.* 2008; Kuhn *et al*. 2019; Ge *et al*. 2019b; Rubinstein *et al*. 2019), este trabajo es el primero en usar un borrador de genoma para facilitar la identificación de variaciones a partir de la "secuenciación de una representación reducida" (*Genotyping by sequencing*, GBS), lo que permite el desarrollo de unos marcadores de mayor calidad.

Análisis de diversidad y estructura poblacional

Se detectaron 7.108 SNPs para las 71 accesiones estudiadas usando el borrador del genoma de 'Hass' para alinear las lecturas. Estos marcadores moleculares mostraron una mayor proporción de transiciones (61,10 %) que transversiones (38,89 %). Este hecho, observado en estudios anteriores (Kuhn *et al.* 2019; Taranto *et al.* 2016; Pootakham *et al*. 2015; Kujur *et al.* 2015), se conoce como 'sesgo de transición', y se atribuye a que las transiciones permiten conservar más la estructura original de la proteína que las transversiones.

El porcentaje de polimorfismos privados (19,45 %) es relativamente bajo, lo que puede deberse a la falta de barreras de esterilidad entre las razas de aguacate, favoreciendo el intercambio de información genética entre ellas. La media de heterocigosidad (0,16) fue menor que la obtenida previamente en otros estudios usando SSRs (Alcaraz & Hormaza 2007; Gross-German & Viruel 2013; Guzmán *et al*. 2017), aunque este resultado era esperable considerando la naturaleza de los SSRs (Taranto *et al.* 2016; Helyar *et al.* 2011). La heterocigosidad es menor que la observada en otros estudios con diferentes cultivos leñosos, como melocotonero, litchi, y olivo (Aranzana *et al.* 2012; Biton *et al.* 2015; Liu *et al.* 2015). Estas diferencias posiblemente se deban a que la estructura varietal del aguacate difiere a la de otros frutales. Así, a nivel mundial, el mercado de aguacate está dominado por un genotipo, 'Hass', y tanto 'Hass' como descendientes de 'Hass', como 'Gwen', son parte del pedigrí de diferentes variedades del grupo híbrido GxM (el más representado en este estudio) y esta selección sesgada puede ocasionar el descenso de la heterocigosidad.

En este trabajo, se han realizado diferentes análisis usando los SNPs detectados (ACP, método de unión de vecinos [NJ], ADMIXTURE, STRUCTURE y ADCP). Todos estos análisis mostraron una clara separación entre las razas, aunque con algunas excepciones en

algunos resultados de STRUCTURE y ADCP, en los cuales no se observó una clara distinción entre los genotipos de raza Guatemalteca y los híbridos GxM. Este último resultado se puede explicar por la preponderancia de genes Guatemaltecos en el germoplasma actual de aguacate (Chanderbali *et al.* 2013). Además, al no existir una barrera de esterilidad entre razas, la hibridación entre genotipos de diferentes razas ha podido ocurrir frecuentemente durante la historia evolutiva y el proceso de domesticación y selección del aguacate (Schaffer *et al*. 2013). En cualquier caso, la agrupación llevada a cabo con ADCP mostró menor hibridación entre las accesiones que con los análisis de STRUCTURE o ADMIXTURE. Resultados similares de baja hibridación genética se han obtenido en otros estudios empleando ADCP (Frosch *et al*. 2014; Söderquist *et al.* 2017) y podría deberse a una sobreestimación de la probabilidad de afiliación a posteriori. Curiosamente para K=5 se obtiene un nuevo subgrupo con ADMIXTURE (Figura. 1.4b) en el grupo GxM. Este nuevo grupo podría representar accesiones con un mayor componente Mexicano.

El grupo compuesto principalmente por accesiones de raza Mexicana mostró una mayor diversidad genética y una mayor proporción de SNPs privados (46,42 %) (Tabla 1.8) junto a una alta heterocigosidad observada, tal y como se obtuvo en trabajos anteriores (Schnell *et al*. 2003; Ashworth *et al.* 2003; Gross-German & Viruel 2013). Respecto a la diversidad genética, cabe destacar que el grupo compuesto principalmente por accesiones Guatemaltecas, y el grupo con accesiones principalmente Mexicanas mostraron mayor diversidad genética que el grupo híbrido de GxM, a pesar de un menor número de muestras. Los resultados obtenidos también mostraron una clara separación entre las accesiones pertenecientes al grupo Antillano de las otras dos razas, tal y como se ha mostrado en estudios anteriores (Davis *et al.* 1998; Gross-German & Viruel 2013; Boza *et al*. 2018; Chen *et al.* 2009) usando un menor número de individuos. Este resultado era esperable, teniendo en cuenta que las razas Mexicana y Guatemalteca tienen en común un nicho ecológico, en las tierras altas tropicales, mientras que los genotipos Antillanos están adaptados a las tierras bajas de América central con climas tropicales (Schaffer *et al.* 2013). Este hecho también podría relacionarse con un valor superior en el coeficiente de endogamia (Fis) en el grupo Antillano respecto al resto de grupos.

Tabla 1.8. SNPs privados por grupos establecidos a través del ADCP. Porcentajes de los 7108 SNPs privados. La raza más representada por grupo es mostrada entre paréntesis.

	Grupos	SNPs Privados	SNPs %
K=3	1 (GxM)	222	16.05
	2 (WI)	519	37.52
	3 (M)	642	46.42
K=4	1(GxWI)	474	34.27
	2 (GxM)	222	16.05
	3 (WI)	45	3.25
	4 (M)	642	46.42
K=5	1(GxWI)	474	34.27
	2 (WI)	45	3.25
	3 (G)	77	5.56
	4 (GxM)	145	10.48
	5 (M)	642	46.42

Asignación de genotipos de pedigrí desconocido al grupo establecido

En aguacate, la agrupación de accesiones en razas se ha llevado a cabo mediante el estudio de caracteres fenotípicos. Además, como la mayoría de cultivares han sido desarrollados a partir de hibridaciones espontáneas el pedigrí es generalmente desconocido. La aproximación llevada a cabo en este trabajo ha permitido asignar algunos cultivares de origen desconocido a los grupos establecidos. De acuerdo con trabajos previos (Chen *et al*. 2009), algunos cultivares parecen ser híbridos complejos entre las 3 razas, aunque la mayoría de los cultivares estudiados son híbridos GxM, siendo estos híbridos los más frecuentes a nivel mundial, especialmente en climas subtropicales.

En este estudio, el desarrollo de un alto número de marcadores moleculares tipo SNP tras el cartografiado contra el borrador del genoma de la cv. Hass, ha permitido el genotipado y la discriminación de accesiones de aguacate revelando claras agrupaciones basadas en el origen racial. Los marcadores tipo SNP desarrollados son un recurso público que será muy útil para futuros estudios sobre el mantenimiento y la caracterización del germoplasma de aguacate, selección genética (GS), selección asistida por marcadores (MAS) o búsqueda de loci asociados a caracteres cuantitativos (QTL) y, consecuentemente, ayudará significativamente reduciendo el coste de la mejora de este cultivo. Sin embargo, se necesitan estudios adicionales con el objetivo de aumentar el número de marcadores moleculares disponibles.

CAPÍTULO 2

Secuenciación, ensamblado y anotación del genoma del chirimoyo (*Annona cherimola* Mill.)

Resumen

Un genoma de referencia es un recurso fundamental para estudios de biología comparativa, genómica de poblaciones, conservación, mejora y evolución. En este capítulo se presenta por primera vez el ensamblaje *de novo* del genoma del chirimoyo (*Annona cherimola* Mill.), el primer genoma secuenciado dentro de la familia Annonaceae, que incluye alrededor de 2.400 especies y el segundo genoma dentro del orden Magnoliales, uno de los grupos más primitivos dentro de las angiospermas. El chirimoyo es un frutal infrautilizado apreciado por sus excelentes cualidades organolépticas, que se cultiva en muchas áreas subtropicales distribuidas en todo el mundo. El tamaño estimado del genoma secuenciado de *Annona cherimola* es de aproximadamente 1,17 Gb, con una heterocigosidad aproximada de 1,05 %. El ensamblaje que se presenta en este trabajo consta de 15.076 cóntigos, con un N50 de 171,3 Kb, siendo aproximadamente el 67 % secuencias repetitivas. La anotación estructural predijo el número de genes de esta especie que se encuentran asociados a un evento de duplicación o hibridación reciente. A pesar de ser la primera versión del ensamblaje del genoma, este recurso es fundamental para la mejora del chirimoyo y para futuros estudios evolutivos de las angiospermas, tratándose del primer genoma disponible de la familia Annonaceae.

Introducción

El chirimoyo pertenece a la familia Annonaceae, una de las familias más ricas y abundantes de los bosques húmedos tropicales (Chatrou *et al*. 2018). Se encuentra incluida en el orden Magnoliales según la clasificación APG IV (APG IV 2016) y junto a los órdenes Canellales, Laurales y Piperales, conforman el complejo Magnoliid dentro de las angiospermas basales. Este clado es uno de los más numerosos dentro de las angiospermas e incluye alrededor de 10.000 especies (Massoni *et al*. 2015) asignadas a 20 familias, algunas con importancia económica como la pimienta negra (Piperaceae, Piperales), el tulípero de Virginia (Magnoliaceae, Magnoliales) o el aguacate (Lauraceae, Laurales).

El chirimoyo ha sido descrito mayormente como 2n=2x=14 (Thakur & Sigh 1965; Walker

1972; Martín *et al.* 2019; Falistocco & Ferradini 2020) con un tamaño de genoma estimado en aproximadamente 1,6 Gb (Martín 2013); sin embargo, existen ciertas discrepancias respecto al número de cromosomas en diferentes trabajos.

Este frutal comenzó a cultivarse por las civilizaciones americanas en tiempos precolombinos (Popenoe 1989) y en los últimos años ha incrementado su importancia, dadas sus excelentes cualidades nutritivas, organolépticas y por la presencia de acetogeninas; compuestos con propiedades antitumorales, antipalúdicas y pesticidas (Alaly *et al.* 1999; Liaw *et al.* 2011). *Annona cherimola* Mill. se conoce en Sudamérica como "chirimoya", palabra que posiblemente tiene su origen en el quechua, y que significa "semilla fría" (Bonavia *et al.* 2004) y en Centroamérica como anona. Aunque su producción es significativa en Perú y Chile, España es el mayor productor a nivel mundial, donde destacan dos cultivares, 'Fino de Jete' y 'Campas', correspondiendo el 95% de la superficie cultivada a 'Fino de Jete'. Este cultivar se distingue por sus excelentes cualidades organolépticas y por su adaptación a la zona de cultivo en la costa subtropical de Granada y Málaga, siendo especialmente interesante para los programas de mejora de la especie.

Dentro del género *Annona* se han producido ciertos avances a nivel molecular, especialmente en estudios de diversidad y en la optimización de la gestión de los recursos genéticos tanto en España como en sus países de origen. Entre estos estudios destacan el aislamiento del gen INO, implicado en el desarrollo del integumento externo del óvulo y que está implicado en la producción de frutos sin semillas (Lora *et al.* 2011b), junto a la determinación del origen de *Annona cherimola* (Larranaga *et al*. 2017) y el movimiento del material vegetal de esta especie en tiempos precolombinos desde Centroamérica a Sudamérica (Larranaga 2016). Sin embargo, como ocurre con la mayoría de los cultivos infrautilizados, sigue existiendo una gran limitación de herramientas moleculares en esta especie.

Gracias a los avances de las nuevas tecnologías de secuenciación (*Next Generation Sequencing*, NGS), el acceso a nuevos conocimientos para fines básicos como aplicados ha permitido avances significativos en numerosos cultivos en los últimos años (Aranzana *et al.* 2019) ya que el ensamblaje de genomas *de novo,* a pesar de su complejidad, se ha convertido en un proceso cada vez más rutinario. En la última década, se han publicado un alto número de genomas de plantas, sobre todo de especies hortícolas (Chen *et al.* 2019a). No obstante, la mayoría pertenecen a unos clados concretos, el de las monocotiledóneas y eucotiledóneas (Soltis & Soltis 2019; Yang *et al*. 2020b), por lo que la diversidad de angiospermas basales

no ha sido representada en suficientes análisis.

El genoma de *Amborella*, la angiosperma viviente más primitiva y que se considera hermana evolutiva del resto de angiospermas, marcó la excepción en 2013 (Albert *et al.* 2013) y, seis años más tarde, se generó el primer genoma del orden Nymphaeales (Zhang *et al.* 2019c), seguido un año más tarde de otro genoma del mismo orden (Yang *et al.* 2020b). Hasta la fecha, a pesar de su valor comercial y su diversidad, se han descifrado solo tres genomas del orden Laurales (Rendón-Anaya *et al.* 2019; Chaw *et al.* 2019; Chen *et al*. 2020), uno del orden Magnoliales (Chen *et al.* 2019c), y otro del orden Piperales (Hu *et al*. 2019), proporcionando algunos de ellos (Chaw *et al.* 2019; Chen *et al.* 2019c; Hu *et al*. 2019; Chen *et al.* 2020) una nueva visión sobre las angiospermas de divergencia temprana, pero con distintos puntos de vista. Esta escasez de genomas en los órdenes más basales de las angiospermas hace que todavía la posición filogenética de algunas de estas familias siga siendo ambigua (Hu *et al.* 2019).

En los últimos años, la metodología NGS ha sido aplicada en la familia Annonaceae, aunque, hasta la fecha, los esfuerzos se han centrado en estudios transcriptómicos. En *Annona squamosa* (anón), Gupta *et al*. (2015) realizaron un análisis transcriptómico del desarrollo del fruto; Liu *et al*. (2016a) identificaron genes asociados a la transición floral y al desarrollo de las flores; Liu *et al.* (2017) analizaron genes diferenciales entre flores malformadas y normales; Li *et al*. (2019a) identificaron genes involucrados en el rajado del fruto en el atemoyo.

Sin embargo, a pesar de la importancia de *Annona cherimola* Mill. a nivel evolutivo, esta especie sigue estando pobremente caracterizada genómicamente como ocurre con el resto de las especies de la familia Annonaceae (Li *et al.* 2019a; Berumen-Varela *et al*. 2019; Liu *et al*. 2016a, 2017; Gupta *et al*. 2015). Los nuevos enfoques basados en 'NGS', por costes decrecientes y alta capacidad de generación de información, brindan una gran oportunidad para solucionar estas limitaciones.

En este capítulo se desarrolla el primer borrador del genoma de la familia Annonaceae, concretamente del chirimoyo, cv. Fino de Jete, usando datos generados mediante la combinación de dos plataformas de secuenciación: Pacific Biosciences (PacBio) e Illumina. Este valioso recurso proporcionará información para el estudio de la evolución de la especie,

podría acelerar los procesos de mejora así como facilitar el desarrollo de nuevas variedades como ya ha ocurrido en otros cultivos.

Material y métodos

Material vegetal y extracción de ADN genómico

Para llevar a cabo los trabajos de genómica, se seleccionó un árbol de chirimoyo (*Annona cherimola* Mill.) del cultivar Fino de Jete procedente de la colección de germoplasma del IHSM-UMA-CSIC "La Mayora", Algarrobo Costa (36° 45' 28.6" N, 4° 02' 37.0" W).

Se recolectaron hojas jóvenes en el campo entre los meses de marzo y mayo. La mayoría de ellas se conservaron a 4 ºC hasta su uso, mientras que una fracción fue congelada usando nitrógeno líquido tras limpiarlas con agua destilada. Estas muestras congeladas fueron enviadas al servicio de secuenciación para su procesamiento según se indica más adelante.

A partir de las hojas guardadas en fresco se aisló ADN de alto peso molecular usando el kit "DNeasy Plant Mini Kit" de Quiagen siguiendo el protocolo establecido por el fabricante. La pureza (ratios: 280/260 y 260/230) y la concentración (ng/ul) del ADN se determinaron usando el espectrofotómetro NanoDrop, electroforesis en gel de agarosa (1%, utilizando TAE como eluyente) y el fluorómetro Qubit 2.0 del servicio de genómica de la Universidad de Málaga (SCBI).

El ADN de las hojas enviadas a la compañía Novogene se aisló, y su calidad se evaluó mediante dos metodologías, electroforesis en gel de agarosa y el uso del fluorómetro Qubit 2.0.

Secuenciación de ADN genómico

La secuenciación de las muestras de ADN genómico se realizó en el "Center for Genomic and Computational Biology" de la Universidad de Duke (https://genome.duke.edu/) y en la compañía Novogene (https://en.novogene.com/).

Se seleccionaron dos plataformas para llevar a cabo la secuenciación: Illumina para generar lecturas pareadas (*Paired-End* [PE]), y la plataforma Pacific Bioscience (PacBio) para generar lecturas largas mediante SMRT (*Single Molecule Real Time*).

Las genotecas DNA-Seq (~300 pb) para Illumina se construyeron mediante los protocolos establecidos para dicha plataforma en el "Center for Genomic and Computational Biology" de la Universidad de Duke. Su secuenciación se llevó a cabo mediante el secuenciador HiSeq 4000, con una longitud de lectura de 2x150 pb. En el caso de PacBio Sequel, en el "Center for Genomic and Computational Biology" de la Universidad de Duke (https://genome.duke.edu/) se secuenciaron 4 SMRTs en la plataforma PacBio Sequel, y 1 SMRTs se secuenció mediante la misma plataforma en la compañía Novogene (https://en.novogene.com/).

Material vegetal y extracción de RNA genómico

A partir de distintos árboles de chirimoyo clonales de la variedad 'Fino de Jete' procedentes de la colección de germoplasma del IHSM-UMA-CSIC "La Mayora", se seleccionaron muestras de hoja joven (en marzo), hoja madura (en febrero), hojas tras 8 días sin riego (en octubre), embrión y pulpa (en noviembre), ovario antes de polinizar y tras 5, 10 y 20 días después de ser polinizado, óvulo, tépalo (entre junio y agosto), semilla, yema hoja y yema floral (en junio) con el objetivo de anotar el genoma. También se germinaron semillas in vitro de un fruto de chirimoyo (cv. Fino de Jete) a 30 ºC y en oscuridad durante aproximadamente 3 semanas, con el fin de mantener tejido en un medio aséptico. Se tomaron muestras de raíz y plántula. Las muestras recolectadas se congelaron inmediatamente en nitrógeno líquido y se guardaron a -80 ºC hasta su uso.

El ARN total se aisló utilizando diferentes protocolos dependiendo del tejido ("RNeasy Plant Mini Kit" de Quiagen y el protocolo adaptado de Gambino, G. *et al*. [2008] con algunas modificaciones). Aproximadamente 300 mg de tejido se homogeneizaron con el buffer de extracción (2 % CTAB, 2.5 % PVP-40, 2 M NaCl, 100 mM Tris-HCl pH 8.0. 25 mM EDTA 8.0) previamente precalentado a 65 ºC más 18 ul de BME (β-mercaptoetanol, añadido inmediatamente antes de su uso). El mismo volumen de SEVAG (cloroformo: alcohol isoamílico; 24:1 v/v) se agregó y se centrifugó a 13.750 rpm durante 10 minutos a 4 ºC. La fase superior acuosa se recuperó y se mezcló con 1/3 del volumen de LiCl 9 M dejándolo precipitar en hielo durante aproximadamente una hora y media. El precipitado se recuperó tras una centrifugación a 14.000 rpm durante 40 minutos. El pellet se resuspendió con cuidado en 500 ul del buffer SSTE (10 mM Tris-HCL pH 8.0, 1 mM EDTA pH 8.0, 1 % SDS, 1M NaCl) precalentado a 65 ºC y sin hielo. Posteriormente, se mezclaron con el mismo volumen de SEVAG (cloroformo: alcohol isoamílico; 24:1 v/v) y se centrifugaron a 14.000

rpm durante aproximadamente 30 minutos a 4 ºC. La fase superior acuosa se recuperó y se mezcló con 0.7 volúmenes de isopropanol frío e inmediatamente, se recuperó el pellet tras una centrifugación a 14.000 rpm durante 30 minutos a 4 ºC. Tras un lavado con etanol al 75 % el pellet se secó y se diluyó en aproximadamente 40 ul de agua miliQ DEPC.

Tras la extracción de ARN se realizó un tratamiento con DNasa I libre de RNasa de Quiagen siguiendo las instrucciones del fabricante. La integridad del ARN se evaluó mediante una electroforesis en gel de agarosa (1 %, utilizando TAE como eluyente) y un Bioanalizador Agilent 2100 (Agient Technologies). La pureza (proporción 260/280 y 260/230) y la concentración (ng/ul) se determinaron usando un espectrofotómetro NanoDrop.

Secuenciación de genotecas RNA-seq

Para la construcción de las genotecas de ADN complementario (ADNc), solo se emplearon aquellas muestras con un número de integridad del ARN (RNA integrity number, RIN) superior o igual a 7. Estas se construyeron siguiendo el protocolo descrito por Hunt & Li (2015) y se secuenciaron mediante Illumina HiSeq 4000 en la compañía Novogene (https://en.novogene.com/). Se obtuvieron 396.098.262 lecturas pareadas (del inglés, *Paired-End* [PE]) con una longitud media de 150 pb.

Estimación del tamaño del genoma y su heterocigosidad

El tamaño, la heterocigosidad y la cantidad de repeticiones del genoma se calcularon empleando un enfoque de distribución de Kmer descrito en Liu *et al*. (2013). La distribución de Kmers (subsecuencias de una determinada longitud (K) que podemos encontrar dentro de una secuencia completa) de tamaños 19, 25, 31 y 37, se estimaron con el software Jellyfish v. 2.3.0 (Marçais & Kingsford 2011) y se cargaron en la página web de GenomeScope (Vurture *et al*. 2017). Gracias al método del análisis de la frecuencia de K-mer podemos estimar estos parámetros genómicos a partir de lecturas sin procesar (Vurture *et al*. 2017).

Ensamblaje del genoma *de novo*

Las lecturas generadas mediante PacBio se cartografiaron contra el genoma cloroplastídico publicado por Blazier *et al*. (2016) (disponible en NCBI en el BioProjecto: PRJNA321881) usando la herramienta BLASR versión 2.2 (Chaisson & Tesler 2012) con el fin de facilitar el ensamblaje del genoma. Aquellas lecturas que no se cartografiaron, se transformaron en

formato FASTA y se unieron en un único archivo mediante el comando cat de Linux. Este archivo se utilizó para generar un primer ensamblaje utilizando el programa Canu versión 1.8 (Koren *et al.* 2017), con el parámetro correctedErrorRate = 0,075, al tratarse de un genoma bastante heterocigoto. Este programa permitió realizar una primera autocorrección y ensamblar el genoma. Aquellas lecturas que no se cartografiaron contra el cloroplasto se alinearon de nuevo contra el primer ensamblaje usando BLASR versión 2.2 (Chaisson & Tesler 2012); además, se ordenaron y se unieron usando Samtools versión 1.3.1 (Li *et al*. 2009). Posteriormente, el archivo generado y el primer ensamblaje realizado con Canu se utilizaron para realizar otra corrección con Arrow versión 5.1.0 (https://github.com/PacificBiosciences/GenomicConsensus) y generar una nueva secuencia consenso.

Con la intención de encontrar el mejor ensamblaje y de comparar distintos ensambladores, se utilizó también Minimap/miniasm 0.3-r179 (Li 2016) siendo posteriormente necesarias dos iteraciones de Racon (Vaser *et al*. 2017) para pulir el ensamblaje. Además, se utilizó Canu versión 1.8 pero esta vez con los parámetros por defecto.

Las lecturas de Illumina procesadas usando fastq-mcf versión 1.04.807 (-q 30 -l 50) (Aronesty 2013) se cartografiaron contra el ensamblaje realizado anteriormente con el programa bwa versión 0.7.120-r789 (Li & Durbin 2010) con los parámetros establecidos por defecto. Posteriormente, se utilizó Pilon versión 1.23 (Walker *et al.* 2014) con el objetivo de generar un genoma consenso de mayor calidad a partir de los ensamblajes generados por Canu versión 1.8. Este corrige bases, mejora el ensamblaje y reduce los huecos (del inglés, *gaps*) (Walker *et al.* 2014). Además, se utilizó de forma iterativa sobre el mismo ensamblaje con el fin de mejorar la secuencia consenso.

Con el fin de conocer cómo de completo es el genoma en término de genes, se utilizó BUSCO versión 3.1.0 (Simão *et al.* 2015) empleando los modelos de *Arabidopsis* para la predicción y GenoToolBox (https://github.com/aubombarely/GenoToolBox/blob/master/SeqTools/FastaSeqStats) para calcular distintos parámetros relacionados con la distribución de tamaños de las secuencias, como N50, que conforman el ensamblaje del genoma.

Anotación estructural y funcional del genoma del chirimoyo

Los elementos repetitivos se identificaron usando primero RepeatModeler (Smit *et al.* 2015b) versión 2.1 con parámetros por defecto, y posteriormente se anotaron mediante

RepeatMasker (Smit *et al*. 2015a) versión 2.1. Asimismo, se utilizó la pipeline de MAKER-P, generando un ensamblaje con un enmascaramiento suave (del inglés, *softmasking*) que facilitaría la anotación del genoma.

Las lecturas generadas mediante RNA-seq se procesaron usando fastq-mcf versión 1.04.807 (l 50 y q 30) (Aronesty 2013), se alinearon contra el genoma empleando Hisat2 (Kim *et al*. 2015) versión 2.1.0 (https://daehwankimlab.github.io/hisat2/) y se ensamblaron utilizando StringTie versión 2 (Pertea *et al.* 2015).

Posteriormente, se utilizaron dos tipos de aproximaciones para la anotación del genoma. Por un lado, se utilizó BRAKER2 (Hoff *et al.* 2018), que usa los archivos GFF generados por Stringtie como evidencia empírica para generar un primer set de genes con GeneMark-ET (Lomsadze *et al.* 2014), siendo refinado posteriormente por Augustus (Stanke *et al.* 2006), mientras que, por otro lado, se empleó MAKER-P, que utiliza como evidencia empírica los archivos GFF y el set de proteínas para el clado Magnoliid de la base de datos Uniprot, y como herramientas de predicción *de novo* se usó SNAP (Korf 2004) y Augustus (Stanke *et al.* 2006). Finalmente, ambas anotaciones se evaluaron mediante BUSCO (Simão *et al.* 2015).

La anotación funcional se realizó utilizando BLASTp para los genes predichos con BRAKER con un valor de corte de $1e^{-10}$, empleando las bases de datos TrEMBL, Swiss-Prot (Bairoch & Apweiler 2000) y Araport11 (Cheng *et al.* 2017). Los dominios proteicos se localizaron durante la anotación funcional con InterProScan (Jones *et al*. 2014) utilizando todas las bases de datos disponibles por defecto (cdd, gene3d, panther, phobius, prints, prosite, smart, tigrfam, freemarker, hamap, pfam, pirsf, prodom, sfld, superfamily, tmhmm). Para visualizar los resultados se utilizó la herramienta IGV (del inglés, *Integrative Genome Viewer*) (Robinson *et al*. 2017).

Todos los análisis se ejecutaron en un servidor del laboratorio del Prof. Aureliano Bombarely en la Universidad de Milán con 160 threads, 3 Tb de RAM y 22 Tb de disco duro (RAID 6).

Análisis de eventos de duplicación en el genoma de *Annona cherimola*

El análisis sobre los posibles eventos de duplicación en *Annona cherimola* y en dos de las especies pertenecientes al clado magnoliid con genomas disponibles (*Liriodendron chinensis* [Chen *et al*. 2019c] y *Cinnamomum kanehirae* [Chaw *et al*. 2019]) se estimaron y se representaron usando la herramienta wgd version 1.1 (Zwaenepoel & Van de Peer 2019).

Disponibilidad de datos

Una vez publicado este estudio, las lecturas brutas serán depositadas en NCBI y el ensamblaje estará disponible para la comunidad científica.

Resultados

Diseño experimental

En un comienzo, con el objetivo de llevar a cabo el ensamblaje del primer borrador del genoma del chirimoyo, se tuvieron en cuenta los resultados obtenidos a través de la citometría de flujo (Martín 2013). Este análisis mostró un tamaño de genoma de chirimoyo aproximado de 1,66 Gb (Martín 2013), por lo que se propuso como primera aproximación un ensamblaje híbrido, secuenciando mediante PacBio Sequel (con una cobertura aproximada de 20x) e Illumina Hiseq 4000 (con una cobertura aproximada de 100x).

Posteriormente, las lecturas de Illumina generadas facilitaron la estimación del tamaño del genoma secuenciado mediante el análisis de K-mers, indicando un tamaño del genoma de aproximadamente 1,17 Gb y, por tanto, una cobertura para el genoma superior a la esperada (de aproximadamente 40x) a partir de los datos generados mediante la plataforma PacBio.

Teniendo en cuenta esta información, junto a los tamaños de genoma para especies del género *Annona* disponibles en la base de datos del Real Jardín Botánico de Kew (https://cvalues.science.kew.org/search/angiosperm) se planteó secuenciar de nuevo mediante PacBio Sequel con el objetivo de incrementar la profundidad de cobertura. De esta forma, se alcanzó aproximadamente una cobertura de 60x (55,54x, tras eliminar todas las secuencias correspondientes al genoma cloroplastídico), lo que permitiría realizar un ensamblaje con Canu y, posteriormente, realizar una corrección mediante las lecturas producidas a través de la plataforma Illumina.

Datos generados

- Genómicos

Tras la secuenciación mediante Illumina Hiseq 4000, se generaron 343.593.016 lecturas pareadas con un rendimiento aproximado de 109 Gb y una puntuación de calidad Q30 del 86,68 %, siendo Q30, el índice de calidad Phred (Q) que refleja la probabilidad de que la llamada de bases incorrecta sea una entre 1.000.

En la primera aproximación, se generaron 7.363.924 lecturas a través de la plataforma PacBio Sequel, con una longitud promedio de 6.714,82 pb y un rendimiento de 49,27 Gb, obteniéndose una cobertura aproximada de 42,20x. Posteriormente, se generaron 2.050.726 lecturas con la misma plataforma, con una longitud media de 10.101 pb, lográndose una profundidad de cobertura de 17,68x y un rendimiento aproximado de 20 Gb (Tabla 2.1).

- Transcriptómicos

Se generaron 396.098.262 lecturas pareadas tras la secuenciación de las librerías de RNA-seq mediante Illumina Hiseq 4000 con una longitud media aproximada de 150 pb (Tabla 2.1).

Tabla 2.1. Resumen de los datos genómicos y transcriptómicos generados.

Fuente	Sistema de secuenciación	Tamaño medio de fragmento secuenciado	Número de secuencias	Bases totales secuenciadas
ADN	Illumina Hiseq 4000	151	343.593.016	51.882.545.416
ADN	PacBio Sequel	6.714,82	7.363.924	49.532.534.122
ADN	PacBio Sequel	10.135	2.050.726	20.785.364.949
ARN	Illumina Hiseq 4000	150	396.098.262	59.215.823.339

Estimación k-mers

Basándonos en la distribución de K-mer (Tabla 2.2), se estimó que el tamaño del genoma del chirimoyo (que se calcula dividiendo el número total de kmers por la profundidad de kmer del pico principal) es de aproximadamente 1,17 Gb. Asimismo, se observó una alta

heterocigosidad (aproximadamente, 1,05 %) mostrándose numerosos sitios polimórficos en el pico izquierdo (a una cobertura de 24,5x) (Figura 2.1).

Tabla 2.2. Estimación del tamaño y heterocigosidad del genoma basado en la distribution de k-mers.

Longitud de k-mers	Tamaño del genoma (pb)	Heterocigosidad (%)
19	1,027,555,199	1,13
25	1,117,888,266	1,10
31	1,171,425,546	1,04
37	1,207,843,004	0,99

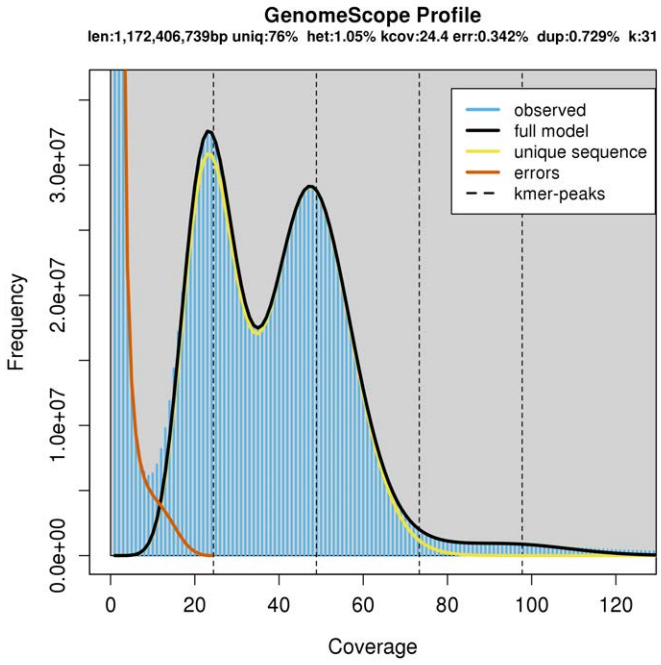


Figura 2.1. Estimación del tamaño, contenido repetitivo, y heterocigosidad del genoma por GenomeScope, basado en 31-mers. En la gráfica se traza la cobertura (eje x) por cada frecuencia k-mer (eje y). El pico homocigoto se observa en una cobertura de 49x, mientras que el pico heterocigoto se observa en una cobertura de 24,5x.

Ensamblado y anotación

Antes de comenzar el ensamblado, aquellas lecturas generadas en la plataforma PacBio se cartografiaron contra el genoma cloroplastídico publicado por Blazier *et al.* (2016) con el fin de facilitar el ensamblaje del genoma. Se obtuvieron 8.736.804 lecturas no cartografiadas (aproximadamente un 92 % de las lecturas originales), con una longitud media de 7.446,76 pb. Estas lecturas se utilizaron en distintas aproximaciones (Canu y Minimap/miniasm) con el fin de encontrar la mejor estrategia de ensamblado.

El software Canu se utilizó para generar dos ensamblajes. En uno de ellos se tuvo en cuenta la elevada heterocigosidad que presenta el genoma (v0.1, Tabla 2.2). El ensamblaje resultante mostró una longitud total de 1.125.206.059 pb, con una longitud de la secuencia más larga de 3.437.413 pb y N50 de 171.117 pb. Este ensamblaje se corrigió mediante el algoritmo Arrow (v0.1.1) incrementando el tamaño de ensamblaje anterior (1.125.713.506 pb), la longitud de la secuencia más larga (3.440.110 pb) y N50 (171.353 pb). Asimismo, se utilizó Pilon con el fin de mejorar la precisión del ensamblaje v0.1.1 (v0.1.2).

El otro ensamblaje producido con el software Canu (v0.2) mostró parámetros muy parecidos a los anteriores. Sin embargo, en esta ocasión solo se corrigió con Pilon, produciendo un ensamblaje de 1.125.345.885 pb, con una longitud de la secuencia más larga de 3.436.411 pb y la longitud de N50 de 171.325 pb (v0.2.1), siendo algo inferior a la mostrada en el ensamblaje v0.1.2 (Tabla 2.2).

Minimap/miniasm es un programa que genera ensamblajes continuos de forma mucho más rápida que el resto de ensambladores para lecturas largas, aunque cuenta con la necesidad de implementar numerosas rondas de corrección (del inglés, *polishing*) para producir una secuencia consenso (Koren *et al.* 2017).

Con el fin de comparar distintos ensambladores se utilizó Minimap/minimasm tras 2 rondas de Racon; sin embargo, este ensamblaje no se incluyó en los análisis posteriores tras evaluarse su tamaño, la longitud de la secuencia más larga, la longitud de N50 y el índice N50 (v0.3, Tabla 2.2) al mostrar peor calidad de ensamblado respecto a los anteriores.

Con el objetivo de conocer cómo de completos son los genomas generados, con respecto al número de genes (v0.1, v0.1.1, v0.1.2, v0.2, v0.2.1), se utilizó el programa BUSCO empleando los modelos de *Arabidopsis* para la predicción, mostrando valores entre el 96 % -

96,5 % del genoma (con duplicación entre el 26,2 % y 30,55 %), con una fragmentación entre 0,8 % - 1,3 % y entre 2,5 % - 2,8 % de datos ausentes. Por lo tanto, se mostraron ligeras diferencias entre los ensamblajes generados por Canu y corregidos con Arrow y Pilon, o los ensamblajes generados mediante Canu y solo corregidos con el programa Pilon (Figura 2.2). Al considerar los parámetros citados anteriormente (tamaño del ensamblaje, la longitud de secuencia más larga y la longitud de N50), se escogió el ensamblaje del genoma de *Annona cherimola* v0.1.2 al mostrar mejores estadísticas. Además, a pesar de no ser el ensamblaje más completo en termino de genes, las diferencias respecto a otros ensamblajes se compensarían con los resultados obtenidos en los otros parámetros.

Tabla 2.3. Resumen estadístico de los distintos ensamblajes del genoma de *Annona cherimola*.

Estadística de ensamblajes	**v0.1**[a]	**v0.1.1**[a,c]	**v0.1.2**[a,c,d]	**v0.2**[a]	**v0.2.1**[a,d]	**v0.3**[b,e]
Tamaño del ensamblaje (Gb)	1.125.206.059	1.125.713.506	1.125.817.265	1.125.115.682	1.125.345.885	1.027.802.131
Longitud de la secuencia más larga (pb)	3.437.413	3.440.110	3.439.214	3.437.413	3.436.411	2341144
Longitud de secuencia media (pb)	74.635,58	74.669,24	74.676,12	74.639,49	74.654,76	83.276,78
Índice N90 (secuencias)[f]	9.346	9.345	9.346	9.346	9.345	7.520
Longitud N90 (pb)[g]	27.587	27.600	27.618	27.585	27.598	33.801
Índice N50 (secuencias)[f]	1.314	1.313	1.314	1.313	1.314	1.613
Longitud N50 (pb)[g]	171.117	171.353	171.333	171.257	171.325	149.784

Tipos de ensamblajes: Canu, a; Minimap/miniasm, b. Corrección: Arrow (1x), c; Pilon (1x),d; Racon(2x),e.

[f]Al ordenar las secuencias de mayor a menor tamaño, el índice N50 o N90 indica el número de secuencias más largas que contienen el 50 % o 90 %, de la secuencia ensamblada total. [g]La longitud de N50 y N90 indica la longitud de la secuencia más corta en el conjunto de los *contigs* (o *scaffolds*) más largos que contienen el 50 % o 90 %, de todas las secuencias del ensamblaje.

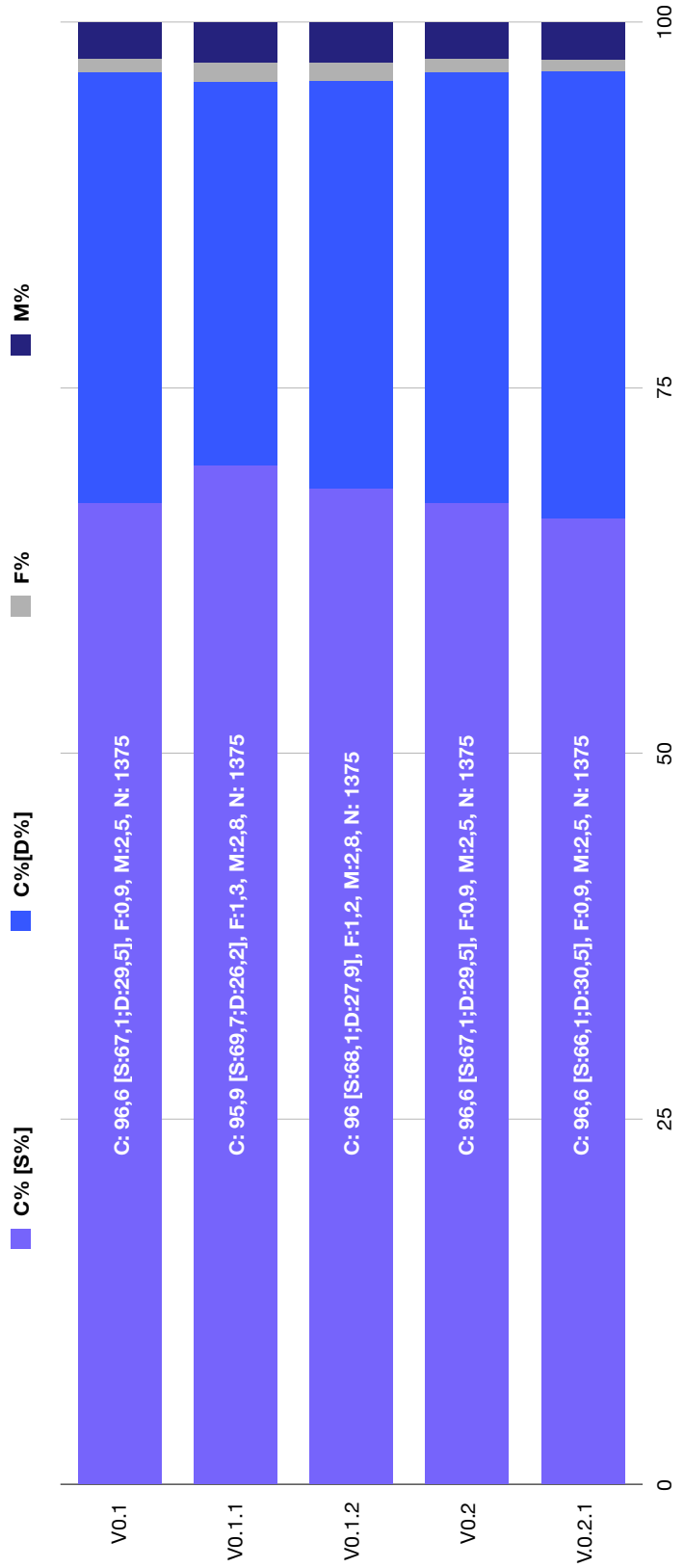


Figura 2.2. Resultados de la evaluación de los ensamblajes por BUSCO. C: integridad del genoma completo; F: fragmentado, M: ausente, N: número de genes usados.

Secuencias repetitivas

Aproximadamente el 67,64 % (761 Mb) del genoma del chirimoyo se identificó en el ensamblaje v0.1.2 como secuencia repetida, mostrándose resultados similares en el genoma de *Ceratophyllum demersum* (63,08 %; Yang *et al.* 2000b), en el de *Liriodendron chinense* (61,64 %; Chen *et al*. 2019c), en el de *Trochodendron aralioides* (64,22 %; Strijk *et al.* 2019), en el de *Nicotiana tabacum* (67 %; Edwards *et al.* 2017) o en el de *Olea europea* (63 %; Cruz *et al*. 2016).

Los elementos de repetición (conocidos) más abundantes fueron los elementos de repetición terminal larga (LTR) representando el 25,54 % del genoma, mientras que los elementos nucleares intercalados (LINE) ocuparon 2,77 % y los transposones de ADN un 1,01 %. Dentro de los LTR, la superfamilia *Copia* se determinó como la más abundante (14,43 %) seguida de la superfamilia *Gypsy* (10,56 %) (Tabla 2.4). Sin embargo, la mayoría de los elementos repetitivos (35,06 %) no se pudieron asociar a ninguna familia conocida.

Tabla 2.4. Elementos repetitivos para el ensamblaje de *Annona cherimola***.**

Clasificación	Número de elementos	Tamaño total (kb)	Tamaño medio (pb)	% del genoma
Orden/Superfamilia				
SINE	**21**	**1,45**	**476,33**	**0**
LINE	**29.151**	**31.177,69**	**2.862,09**	**2,77**
LINE/L1	21.060	25.875,73	1.228,67	2,3
LINE/RTE-BovB	7.506	5.256,74	700,34	0,47
LTR	**126.505**	**287.575,59**	**7.427,41**	**25,54**
LTR/Caulimovirus	3.505	5.923,47	1.690,01	0,53
LTR/Copia	72.913	162.442,43	2.227,89	14,43
LTR/ERVK	466	207,3	444,85	0,02
LTR/Gypsy	48.698	118.932,18	2.442,24	10,56
DNA elements	**13.450**	**11.385,4**	**7.640,81**	**1,01**
DNA	739	70,79	95,79	0,01
DNA/CMC-EnSpm	1.745	438,25	251,15	0,04
DNA/MULE-MuDR	3.566	5.934,38	1.664,16	0,53
DNA/PIF-Harbinger	1.883	741,71	393,9	0,07
DNA/hAT-Ac	2.260	2.736,11	1.210,67	0,24
DNA/hAT-Tag1	1.152	1.235,95	1.072,88	0,11

DNA/hAT-Tip100	390	80,57	206,59	0,01
Unknown	590.738	394.731,12	668,20	35,06
Simple_repeat	579.080	26.126,48	45,12	2,32
Low_complexity	139.511	7.878,57	56,47	0,7
Satellite	1.006	167,46	225,63	0,01
Simple_repeat	579.080	26.126,48	45,12	2,32
RC/Helitron	1.149	837.45	728.85	0,07
rRNA	473	434,55	918,72	0,04
tRNA	678	214,60	316,51	0,02
Total	**1.483.873**	**761.546,40**	**513,22**	**67,64**

Predicción de genes y anotación funcional

El número de genes del ensamblaje del chirimoyo v0.1.2 se estimó con la combinación de los transcriptos generados en este trabajo y en estudios que todavía no se encuentran publicados para la misma especie (Lora *et al.,* comunicación personal), la predicción de los genes *de novo*, junto a las proteínas disponibles para el clado Magnoliid de la base de datos Uniprot al utilizar MAKER.

El número estimado de genes varió entre 31.683, al utilizar el programa MAKER, y 77.015, con el programa BRAKER. Las longitudes medias del gen y exón fueron 5.186 pb y 239 pb respectivamente, a partir de los resultados de MAKER, mientras BRAKER arrojó unas longitudes medias de 3.307 pb y 231 pb para gen y exón (Tabla 2.5).

Tabla 2.5. Resumen estadístico de la anotación del genoma del chirimoyo (*Annona cherimola* Mill.) realizada a partir de dos programas y aproximaciones distintas: MAKER y BRAKER.

	MAKER	**BRAKER**
Número de genes totales	31.683	77.015
Longitud media de gen (pb)	5.186	3.307
Máxima longitud de gen (pb)	266.020	93.324
Longitud media de exón (pb)	239	231
% del genoma cubierto por el gen	14,6	22,6

Tras emplear el programa BUSCO con el objetivo de evaluar las dos aproximaciones utilizadas, se detectó que la predicción de BRAKER fue ligeramente más completa respecto al espacio de genes, con un 96,1 % en comparación con el 84,6 % de la anotación realizada con MAKER (Figura 2.3). Sin embargo, de los 77.015 genes predichos por BRAKER, solo 42.662 (55,39 %) demostraron estar relacionados con proteínas reales según los resultados obtenidos. Además, tras realizarse la anotación funcional de los genes predichos, se observó que de los 77.015 genes, 39.089 se anotaron como desconocidos (50,75 %), siendo la mayoría de estos fragmentos de pequeño tamaño por lo que no mostraron coincidencias fiables con las bases de datos empleadas. A pesar de que la anotación realizada con MAKER fue ligeramente menos completa, la detección de genes pareció ser más precisa, coincidiendo el 92,47 % de los genes predichos con proteínas presentes en *Arabidopsis*.

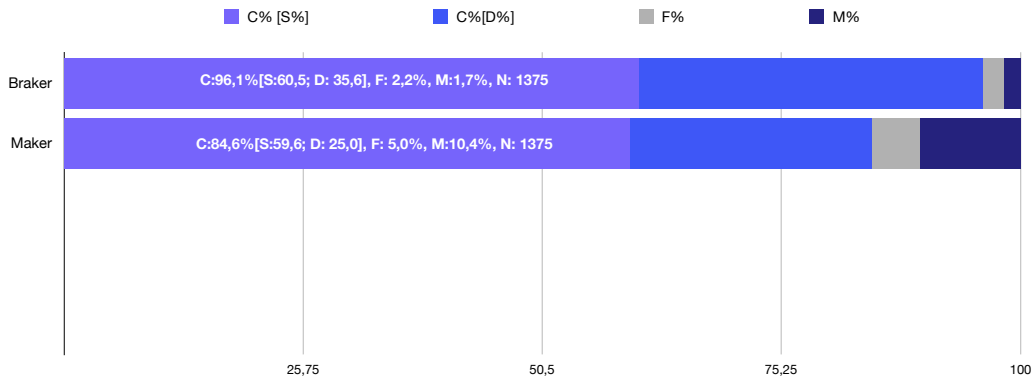


Figura 2.3. Resultados de la evaluación de las anotaciones por BUSCO. C: integridad del genoma completo; F: fragmentado, M: ausente, N: número de genes usados en el análisis.

Por otro lado, tras visualizar algunos de los genes predichos a través del programa IGV, se observó que genes anotados por BRAKER sin anotación funcional, por ejemplo 'g42828.t1' (113.843-114.214 pb), no eran predichos por el programa MAKER; sin embargo, sí se detectaban en la misma zona genes de mayor longitud. En el ejemplo 'Ancher151C09258g0000070' (112.783-115.185 pb), el cual muestra una coincidencia con una hipotética proteína CKAN_01463300 (*Cinnamomum micranthum* f. Kanehirae) (Figura 2.4).

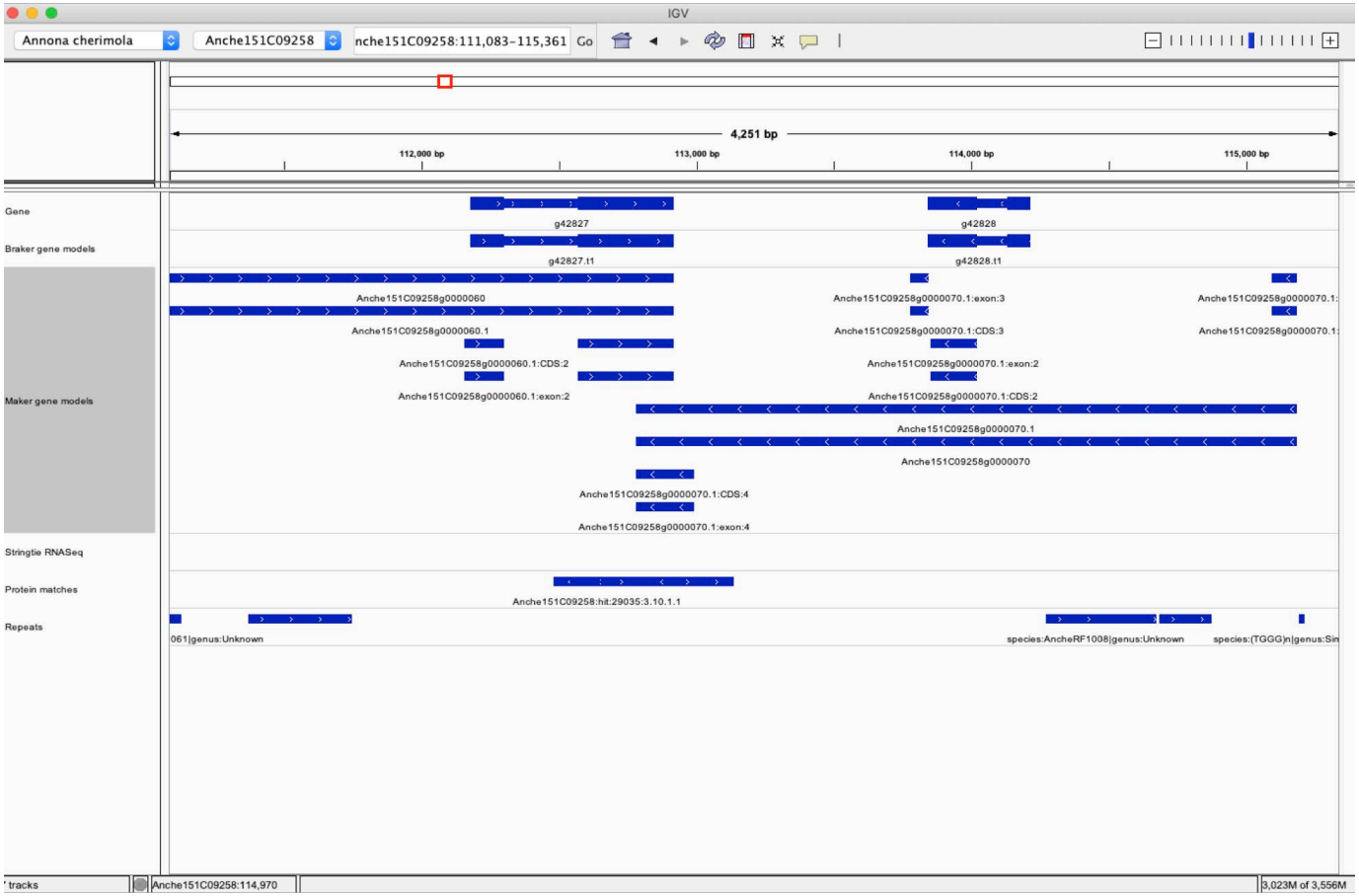


Figura 2.4. Imagen del visualizador IGV tras comparar distintos genes predichos a partir de dos aproximaciones distintas, MAKER y BRAKER.

Análisis de eventos de duplicación en el genoma de *Annona cherimola*

Estudios previos han encontrado una asociación entre los eventos de duplicación del genoma (*Whole Genome Duplication*, WGD) y cambios en la tasa de diversificación de las angiospermas (Landis *et al*. 2018). Los análisis de WGD se basan en la identificación de genes duplicados para conocer el número de sustituciones sinónimas que contienen (Ks). Suponiendo que cada sitio de sustitución sinónima evoluciona de manera neutral, Ks se considera un buen indicador de la edad de duplicación (Kimura 1977; Udall & Wendel 2006). En este estudio, tras analizar la distribución de Ks (sustitución sinónima) en *Annona cherimola*, y otras especies de divergencia temprana como *L. chinensis* y *C. kanehirae,* se pudo observar que el chirimoyo presenta un gran número de genes duplicados en un evento reciente. De hecho, utilizando los genes predichos mediante MAKER se detectaron 23.236 genes de los cuales 18.312 mostraron un Ks entre 0,05 y 0,5 (Figura 2.5).

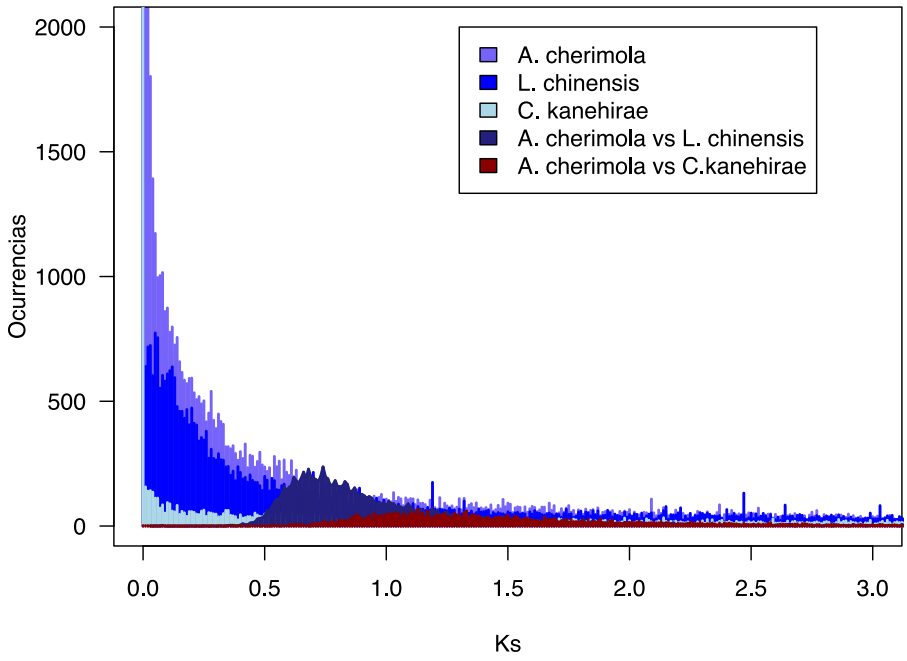


Figura 2.5. Representación de la tasa Ks respecto a las ocurrencias de genes duplicados. En morado aparece el análisis del chirimoyo (*A. cherimola*), en azul oscuro el del liriodendron (*L. chinensis*) y en azul claro el análisis del alcanforero (*C. kanehirae*).

Discusión

La disponibilidad de genomas de referencia ha favorecido el estudio evolutivo y la mejora de un gran número de frutales como el melocotonero (Aranzana *et al*. 2019), el manzano (Chen *et al.* 2019a), el almendro (Sánchez-Pérez *et al*. 2019; Alioto *et al.* 2020), o el naranjo (Xu *et al*. 2013). Sin embargo, actualmente existe un importante sesgo hacia frutales de clima templado respecto frutales tropicales o subtropicales de interés económico, desde los que hasta ahora solo se ha secuenciado el genoma de la papaya (*Carica papaya*) (Ming *et al.* 2008), el cacao (*Theobroma cacao*) (Argout *et al.* 2011), el durian (*Durio zibethinus*) (Teh *et al.* 2017), el aguacate (*Persea americana*) (Rendón-Anaya *et al.* 2019) o el mango (*Mangifera indica*) (Wang *et al.* 2020). Ocurre lo mismo para angiospermas basales respecto a especies pertenecientes a otros linajes, como monocotiledóneas y eucotiledóneas (Soltis & Soltis 2019; Yang *et al*. 2020b). El chirimoyo pertenece a la familia Annonaceae, situándose esta familia en el orden Magnoliales, dentro de las angiospermas primitivas. Actualmente solo existe un genoma secuenciado dentro de este orden, de la especie *Liriodendron chinense*, que pertenece a la familia de las Magnoliáceas, que divergió de la familia Annonaceae hace aproximadamente 100 millones de años (Hedges *et al.* 2015), por lo que los resultados derivados de este trabajo tienen utilidad para estudios filogenéticos y evolutivos en angiospermas primitivas. Además, el ensamblaje que se presenta en este trabajo, proporciona un recurso útil y valioso para progresar en la mejora de este cultivo. *Annona cherimola* es una especie diploide pero con un tamaño de genoma que supera la media de los genomas de especies de angiospermas secuenciados hasta ahora (Chen *et al.* 2019c). Además, es un cultivo infrautilizado con un interés comercial limitado, y posee un largo periodo intergeneracional los que, en conjunto, no son cualidades que faciliten el estudio genómico.

El tamaño del genoma de chirimoya obtenido en este trabajo (aproximadamente 1,17 Gb mediante un análisis de k-mer para el genoma secuenciado) es inferior al descrito previamente en otros trabajos con otras metodologías. Así, se han estimado tamaños de 1,66 Gb (Martín 2013) con citometría de flujo. Sin embargo, a pesar de que la técnica de citometría de flujo se ha convertido en una herramienta fundamental, rápida y eficiente para conocer el tamaño del genoma y poliplidía (Doležel *et al*. 2007), el resultado de 1,66 Gb fue, posiblemente, una sobrestimación del tamaño, lo que también ha ocurrido con otras especies de la familia de las Anonáceas (*Annona squamosa, Annona glabra, Asimina triloba*) (Martín 2013) si se comparan los resultados con los obtenidos por otros autores

(https://cvalues.science.kew.org/), los cuales detectaron tamaños similares al propuesto en nuestro estudio tras el análisis de k-mer. Del mismo modo, se detectó una elevada heterocigosidad (~ 1,05 %) como ocurre en otros cultivos (Hu *et al*. 2019; Teh *et al.* 2017), lo que normalmente favorece la formación de ensamblajes fragmentados (Del Angel *et al*. 2018), disminuyendo la contigüidad.

Ensamblaje del genoma

Antes de comenzar el ensamblaje del genoma de chirimoyo, las lecturas generadas mediante PacBio se cartografiaron contra el genoma cloroplastídico, ya que en plantas podría facilitar el ensamblaje (Soorni *et al*. 2017a) y aumentar su calidad (Jung *et al.* 2019). De las distintas aproximaciones de ensamblado utilizadas, el mejor ensamblaje del genoma se obtuvo con Canu, con un resultado más contiguo. De hecho, Canu ha sido considerado por algunos autores (Jayakumar & Sakakibara 2019) como uno de los mejores ensambladores para lecturas largas, y tiene la ventaja añadida de incluir su propio paso de corrección. Tras el ensamblaje, se llevaron a cabo una serie de mejoras con Arrow y, utilizando las lecturas de Illumina, se incrementó la exactitud del ensamblaje con Pilon, tal y como se ha realizado por ejemplo, en el genoma publicado recientemente de *Nymphaeaceae colorata* (Zhang *et al*. 2019c).

La calidad de los ensamblajes de genomas es bastante desigual y, a menudo, se debe a la complejidad y heterocigosidad de cada especie (Chen *et al.* 2019a), afectando varios factores (principalmente abundancia de las zonas repetitivas y heterocigosidad) a la integridad y contigüidad del ensamblaje (Veeckman *et al.* 2016). Debido a las características del genoma de chirimoyo, la cantidad de datos generados, y las metodologías empleadas, se observó una menor contigüidad en el ensamblaje de esta especie respecto a los ensamblajes de otras especies publicados recientemente (Chen *et al*. 2019c; Hu *et al*. 2019; Zhang *et al.* 2019c; Chaw *et al.* 2019; Yang *et al.* 2020b). Sin embargo, debe tenerse en cuenta que en la mayoría de estos ensamblajes no solo se utilizó un enfoque híbrido (lecturas de corta y larga longitud), sino que emplearon herramientas adicionales para mejorar el ensamblaje, como el mapeo óptico (BioNano) o la asociación de cromatina (Hi-C) con el fin de unir los cóntigos. Este tipo de herramientas (BioNano o Hi-C) disminuyen el número de supercóntigos y aumentan su tamaño desde tres a diez veces, dando lugar a ensamblajes a nivel cromosomal (Jung *et al*. 2019). A pesar de ello, el ensamblaje de *Annona cherimola* se mostró bastante completo respecto al espacio de genes (aproximadamente 96 %) y con un N50 de 171,3 Kb, siendo

superior a la longitud media de un gen en plantas, por lo que se pudo llevar a cabo una anotación estructural y funcional de los genes a partir de este genoma.

Anotación del genoma

En los genomas de las plantas, una proporción considerable del genoma está constituido por secuencias repetitivas. De hecho, estas están implicadas en procesos relevantes como la evolución del genoma, el reordenamiento cromosómico o la creación y regulación de genes (Li *et al*. 2019b). En nuestro ensamblaje, tras la anotación utilizando RepeatMasker, se mostró que aproximadamente el 67,64 % del genoma corresponde a secuencias repetitivas. Proporciones similares de secuencias repetitivas se han descrito en otras especies (Yang *et al.* 2020b; Chen *et al.* 2019c; Strijk *et al*. 2019; Edwards *et al.* 2017; Cruz *et al*. 2016). Tal y como suele ocurrir en especies vegetales, los elementos LTR fueron los más abundantes, (Suguiyama *et al*. 2019).

Predicción de genes y análisis de eventos de duplicación en el genoma del chirimoyo

Los distintos análisis descritos en este trabajo han mostrado discrepancias en cuanto el número de genes detectados en chirimoyo; de los genes predichos con BRAKER solo el 55 % demostraron estar relacionados con proteínas reales, y de los genes predichos por MAKER el 92,47 % coincidieron con los genes de *Arabidopsis*. Por tanto, se optó por estudiar los eventos de duplicación teniendo en cuenta los genes detectados mediante MAKER, mostrándose esta predicción más precisa. El número de genes estimados se mostró entre un mínimo de 37.926 genes mediante el análisis realizado con MAKER y un máximo de 77.015 genes predichos al emplear BRAKER. En general, el número de genes estimado en chirimoyo fue superior al propuesto para *Arabidopsis thailana* (27.416 genes; TAIR 2016), *Prunus persica* (27.852 genes; Verde *et al.* 2013), *Prunus dulcis* (27.969 genes; Alioto *et al*. 2020) o *Citrus sinensis* (29.445 genes; Xu *et al.* 2013), mostrándose un número de genes dentro del rango del chirimoyo en *Olea europea* (56,349 genes; Cruz *et al.* 2016), *Durio zibethinus* (45.335 genes; Teh *et al*. 2017) o *Piper nigrum* (63.466 genes; Hu *et al.* 2019). No obstante, hay que tener en cuenta que serían necesarios análisis complementarios para definir un número de genes concreto, y que la predicción de genes se encuentra sujeta a cambios como ha ocurrido en otros genomas (Sterck *et al*. 2007). Por ejemplo, en el genoma del manzano se predijeron inicialmente 57.386 genes (Velasco *et al.* 2010) y en la última versión

se ha reducido este número hasta 44.677 (Zhang *et al.* 2019b).

La poliploidía es fundamental en la evolución de las angiospermas (Cai *et al.* 2019; Yang *et al.* 2020b), y también tiene una importancia relevante en la mejora de cultivos (Fang & Morrell 2016). La existencia de un evento de duplicación o hibridación reciente en el genoma de chirimoyo estaría apoyada por la detección de un porcentaje relativamente alto de copias duplicadas en el genoma. Estudios anteriores han descrito la existencia de un evento de duplicación común en el clado Magnollid antes de la diversificación de las Lauráceas y Magnoliáceas hace aproximadamente 116 millones de años (Cui *et al*. 2006). Sin embargo, cuando se llevó a cabo ese estudio, todavía no se habían secuenciado los genomas de angiospermas primitivas, que están disponibles actualmente. En este trabajo se ha podido observar un evento de duplicación o hibridación reciente del genoma de *Annona cherimola*, que se habría producido prácticamente a la vez que el evento de duplicación mostrado en *Cinnamomum* (Lauraceae, Laurales) y posterior al observado en *Liriodendron* (Magnoliaceae, Magnoliales).

A pesar de que Martín *et al.* (2019) y Falistocco & Ferradini (2020) propusieron un número de cromosomas haploide de *Annona cherimola* de 7, coincidiendo con estudios anteriores (Kumar & Ranadive 1941; Asana & Adiata 1945; Thakur & Singh 1965, 1969; Tanaka & Okada 1972; Morawetz 1986), sigue existiendo cierta discrepancia respecto a otros trabajos que consideran la existencia de 8 cromosomas (Bowden 1945, 1948). Asimismo, dentro de la tribu Annoneae solo se ha citado el número de 7 cromosomas en el género *Annona*. Tras un evento de duplicación se producen cambios en la expresión de genes, reducción de estos, y una reorganización del genoma (Cui *et al.* 2006; Soltis *et al*. 2015), por lo que el desacuerdo comentado anteriormente podría deberse a cierta inestabilidad cromosómica en *Annona cherimola* tras el evento descrito. Esto podía explicar la observación de un alto número de individuos triploides en cruzamientos intra e interespecíficos de genotipos diploides en *Annona* (Martín *et al*. 2019) Además, no es descartable la hipótesis que considera que *Annona cherimola* podría ser un híbrido interespecífico (H. Rainer, comunicación personal). Por tanto, todavía son necesarios estudios adicionales para conocer en más profundidad la genómica de este cultivo.

En este trabajo se presenta el primer ensamblaje y anotación del genoma de *Annona cherimola*, el primer genoma disponible de la familia Annonaceae. Un recurso público y fundamental para avanzar en la mejora del chirimoyo y para futuros estudios comparativos o

evolutivos en las angiospermas más primitivas.

CAPÍTULO 3

Elaboración de un mapa genético en el género *Annona*

Resumen

Dentro del género *Annona*, encontramos algunas especies de gran interés económico en ciertas regiones del mundo, como el chirimoyo, el anón o el atemoyo. Sin embargo, la ausencia de estudios genéticos, y el bajo número de programas de mejora en estas especies, dificulta el desarrollo de nuevos cultivares. Los mapas genéticos proporcionan una información fundamental para comprender la organización de los genomas y para facilitar la mejora de los cultivos. En este trabajo, se han generado por primera vez SNPs de una población F2 desarrollada a partir del cruce interespecífico entre *Annona cherimola* ('Fino de Jete') y *Annona squamosa* ('Thai seedless'). De la primera generación (F1) se seleccionaron 2 individuos ('JT20' y 'JT07') que se autofecundaron para producir 24 y 27 individuos de cada uno, respectivamente. 'Fino de Jete' es el cultivar de chirimoyo más cultivado en España pero, a pesar de sus cualidades organolépticas, tiene una serie de desventajas, entre las que destaca el alto número de semillas. Sin embargo, 'Thai seedless' debido a una mutación natural, solo produce unas semillas rudimentarias. En este trabajo se generaron 550 SNPs que se distribuyen en 8 grupos de ligamiento. El mapa genético producido utilizando los paquetes de R, qtl y ASMap, cubre aproximadamente 1.388 cM con una distancia media entre marcadores de aproximadamente 2,6 cM, tratándose del primer mapa genético para el género *Annona* generado a partir de una población F2.

Introducción

El chirimoyo (*Annona cherimola* Mill.) es una especie diploide que ha sido mayoritariamente descrita como 2n=2x=14 (Thakur & Sigh 1965; Walker 1972; Martín *et al.* 2019; Falistocco & Ferradini 2020), aunque existe cierta discrepancia puesto que algunos autores han considerado que contiene un mayor número de cromosomas (2n=2x=16) (Bowden 1948) (Ver Capítulo 2). Esta especie forma parte de la familia Annonaceae, incluida dentro del orden Magnoliales. El número de 8 cromosomas ha sido descrito en otros géneros de la familia de las Anonáceas, como *Asimina*, *Disepalum*, *Goniothalamus* y *Neostenanthera,*

siendo, por tanto, el número de cromosomas x=7 único del género *Annona* dentro de la tribu Annoneae (Martín *et al.* 2019).

El chirimoyo se cultiva en numerosas zonas tropicales y subtropicales de África, América, Asia, Australia y Europa. España es el mayor productor de este cultivo a nivel mundial, con 3.048 hectáreas cultivadas y unas 43.645 t en el año 2019 (MAPA 2019), seguido de Perú y Chile (Hormaza *et al.* 2020). Actualmente, el cultivo de chirimoyo en la costa española es prácticamente monovarietal, con más del 95 % de la superficie ocupada por el cultivar Fino de Jete, que posee excelentes propiedades organolépticas y que se encuentra muy bien adaptado a la zona. Sin embargo, hay caracteres que serían de interés mejorar en este cultivar, como la corta vida poscosecha, la susceptibilidad a la mosca del mediterráneo (*Ceratitis capitata*), la calidad de la fruta (ºBrix) especialmente en los meses más fríos, y el alto número de semillas. Hasta el momento son pocos los programas de mejora centrados en este cultivo. El principal es el que se lleva a cabo en el IHSM-UMA-CSIC "La Mayora" como resultado del cual, recientemente, se ha registrado un nuevo cultivar denominado 'Alborán', que posee un menor contenido de semillas en comparación con 'Fino de Jete', y una calidad excelente en los meses de invierno y primavera (Lora *et al*. 2018). No obstante, son necesarios esfuerzos adicionales para poder aumentar el interés por este cultivo.

En el campo de la mejora genética, la función principal de la hibridación entre cultivos y sus parientes, ha sido la introgresión de rasgos de interés de una especie a otra (Warschefsky *et al*. 2014; Polanco *et al.* 2019). En los últimos años, uno de los objetivos del programa de mejora del IHSM-UMA-CSIC "La Mayora" ha sido la búsqueda de genotipos sin semillas a partir de un cruzamiento interespecífico, entre 'Fino de Jete' (*Annona cherimola*) y 'Thai seedless' (*Annona squamosa*), un cultivar con ausencia de semillas. Los híbridos entre ambas especies se denominan atemoyos y son de interés comercial en algunos países de climas tropicales. Trabajos moleculares han permitido conocer que una deleción en el gen INO (inicialmente descrito en *Arabidopsis*) en 'Thai seedless', es la causante de la ausencia de semillas (Lora *et al.* 2011b). Sin embargo, este es un caso puntual de un gen concreto y, por tanto, se necesitan trabajos adicionales para acelerar y optimizar los programas de mejora en esta especie.

Los mapas genéticos representan las posiciones de los marcadores y/o genes a lo largo de los cromosomas (grupos de ligamiento) en función de la frecuencia de recombinación. Son una herramienta básica para llevar a cabo estudios genéticos, permiten la mejora de los

ensamblajes de genomas de novo (Deokar *et al*. 2014; Liu *et al.* 2016b), realizar estudios de genómica comparativa (da Silva Linge *et al*. 2018), y facilitan la mejora de cultivos (da Silva Linge *et al.* 2018; Polanco *et al.* 2019) al poseer numerosas aplicaciones como la identificación de genes ligados a rasgos agronómicos (QTLs) y selección asistida de marcadores (MAS) (Hussain *et al.* 2017; Calle *et al*. 2018).

Los trabajos realizados en este campo en especies del género *Annona* son muy limitados al ser escasos los recursos moleculares existentes (Tabla I.1, Introducción). Durante los últimos 13 años se han generado dos mapas de ligamientos basados en la población F1 del cruce interespecífico 'Fino de Jete' x 'Thai seedless'. El primero fue generado en 2007 resultado del estudio de segregación de 59 SSR de *Annona cherimola*. No obstante, la eficacia de cartografiado fue baja debido al bajo número de marcadores informativos para 'Thai seedless'. 59 marcadores se agruparon en 8 grupos de ligamiento (7 de más de 3 marcadores, y uno con dos marcadores) con un LOD ≥ 6. En conjunto, se generó un mapa con una distancia total de 293 cM (Escribano 2007). Basándose en esta aproximación, seis años más tarde se llevó a cabo otro mapa de ligamiento, en esta ocasión a partir de 82 SSRs (57 de *Annona cherimola* y 25 de *Annona squamosa*), para un total de 8 grupos de ligamiento con LOD ≥ 3 (6 grupos de ligamiento con más de 8 marcadores, y dos grupos con 3 marcadores), con una distancia genética total de 583,7 cM. Sin embargo, el desarrollo de un mapa de ligamiento para el parental 'Thai seedless' se vio obstaculizado, al comprobarse mediante trabajos moleculares (SSRs) que aproximadamente el 90 % de su genoma se encontraba en homocigosis (Martín 2013).

Con el objetivo de dar un salto cualitativo en la generación de mapas de ligamiento en el género *Annona* que permita avanzar en los programas de mejora genética, se describe en este trabajo el primer mapa de ligamiento de atemoya basado en una población F2, realizado a partir de los análisis de los marcadores moleculares (SNPs) generados mediante el uso de metodologías de nueva generación. Este mapa se ha construido con el fin de proporcionar una herramienta básica que facilite el estudio del genoma, así como abrir la puerta a futuras aplicaciones en la mejora del cultivo ya que, siendo una especie con un largo periodo intergeneracional, la posibilidad de realizar una selección temprana con marcadores sería un avance muy relevante.

Material y métodos

Material vegetal para el desarrollo del mapa genético

La construcción del mapa genético de atemoya se ha llevado a cabo utilizando 58 muestras procedentes de una población F2, los individuos de la F1 que se autopolinizaron, junto a sus parentales del banco de germoplasma de Annonaceae del Instituto de Hortofruticultura Subtropical y Mediterránea (IHSM-UMA-CSIC) "La Mayora". Esta población ha sido generada a partir de la autopolinización de dos individuos ('JT20' y 'JT07') de una F1, generados a partir de un cruzamiento interespecífico dirigido entre dos genotipos de dos especies del género *Annona*, 'Fino de Jete' (parental femenino), perteneciente a la especie *Annona cherimola*, y 'Thai seedless' (parental masculino), perteneciente a la especie *Annona squamosa,* que produce frutos sin semilla (Figura 3.1). Este cruzamiento se realizó en 2001 en el marco de un programa de mejora del chirimoyo.

Figura 3.1. Fruto de la variedad Fino de Jete de *Annona cherimola* a la izquierda y fruto de la variedad Thai seedless de *Annona squamosa* a la derecha.

Extracción de ADN, preparación de genotecas, secuenciación y procesamiento de lecturas en bruto

A partir de las hojas jóvenes previamente lavadas con agua destilada, se aisló ADN de 'Fino de Jete', 'Thai seedless', de los individuos que se autopolinizaron, 'JT07' y 'JT20', y de los individuos de la descendencia F2, usando el Kit "DNesy Plant Mini Kit" de Qiagen siguiendo el protocolo descrito por el fabricante. La purificación (ratios: 280/260 y 260/230) y la concentración (ng/ul) del ADN se determinó usando el espectrofotómetro NanoDrop y el fluorómetro Qubit 2.0 del servicio de genómica de la Universidad de Málaga (SCBI).

Las genotecas se construyeron siguiendo el protocolo descrito por Sonah *et al*. (2013), digiriendo 100 ng de ADN genómico de cada individuo con ApeKI (Anexo 1.1). La secuenciación de las genotecas se realizó por la empresa Novogene con Illumina HiSeq 4000 (2x150). Las lecturas en bruto fueron demultiplexadas usando el paquete GBSx (Herten *et al.* 2015) siendo posteriormente procesadas para eliminar posibles secuencias de adaptadores, lecturas con una longitud menor a 50 bases, y filtrar aquellas con regiones de baja calidad usando el software fastq-mcf versión 1.04.807 (-l 50 y -q 30) (Aronesty 2013).

Cartografiado, identificación de SNPs y filtrado

Las lecturas generadas fueron cartografiadas contra uno de los primeros ensamblajes desarrollados durante el Capítulo 2 (Tabla 3.1) usando bwa versión 0.7.120-r789 (Li & Durbin 2010) con los parámetros establecidos por defecto. Aquellas lecturas que no fueron cartografiadas se eliminaron usando Samtools versión 1.3.1 (Li *et al*. 2009), mientras que con las lecturas retenidas se crearon archivos BAM. Todos estos archivos BAM se unieron usando Bamaddrg (https://github.com/ekg/bamaddrg), y se utilizó el paquete Samtools versión 1.3.1 (Li *et al.* 2009) con el fin de ordenarlos e indexarlos. Para detectar las variaciones, y eliminar SNPs con una calidad de mapeado menor que 20 y con una profundidad menor que 5, se usó el programa Freebayes versión 0.9.20 (Garrison & Marth 2012). De las 58 muestras genotipadas, se eliminaron aquellos genotipos repetidos de los que se obtuvieron menor número de lecturas, junto a los individuos de la F1 (‘JT07’ y ‘JT20’). De los 51 individuos restantes, 24 pertenecen a la familia generada a partir de la autopolinización del individuo ‘JT07’, y 27 pertenecen a la familia generada a partir de la autopolinización del individuo ‘JT20’. Los polimorfismos brutos obtenidos se filtraron con el paquete VCFtools versión 0.1.12 (Danecek *et al*. 2011) eliminando los SNPs no bialélicos, SNPs con una proporción de datos faltantes en más del 80% de las muestras, y aquellos que estuvieran dentro de una distancia de 1.000 pares de bases. Antes y después del filtrado, se generó un resumen estadístico usando vcf-stats versión 0.1.12 (Danecek *et al.* 2011).

Tabla 3.1. Resumen del ensamblaje del primer borrador del genoma de *Annona cherimola* Mill. (cv. Fino de Jete) utilizado en este trabajo.

Estadística del ensamblaje	
Secuencias totales	15.076
Longitud total (pb)	1.125.713.506
Secuencia más larga (pb)	3.440.110
Secuencia más corta (pb)	1.039
N50 (pb)	171.353

Finalmente, usando la herramienta Vcf2Mapmaker del paquete GenoToolBox (https://github.com/aubombarely/GenoToolBox/blob/master/SNPTools/Vcf2Mapmaker) se transformó el archivo vcf, y se filtraron datos genómicos para posteriormente realizar el análisis de ligamiento y la construcción del mapa en R. Esta herramienta compara cada SNP de la población con los de los parentales, determinados como A ('Fino de Jete') y B ('Thai seedless'). Igualmente, permitió excluir marcadores que fueran iguales para ambos parentales, aquellos que faltaran en alguno de los parentales, junto a aquellos marcadores que mostraran una distorsión de segregación tras ser analizados mediante Chi-cuadrado ($\chi 2$)(p-valor < 0,01), teniendo en cuenta una segregación mendeliana esperada de 1:2:1. Todos los análisis de cartografiado, identificación de SNPs y filtrado se realizaron en un servidor del Prof. Aureliano Bombarely de la Universidad de Milán con 160 threads, 3 Tb de RAM y 22 Tb de disco duro (RAID 6).

Análisis de ligamiento y construcción del mapa

Los marcadores retenidos fueron utilizados para la elaboración del mapa genético. Para ello se emplearon dos paquetes de R versión 3.6.2, qtl (Broman *et al.* 2003) y ASMap (Taylor & Butler 2017) en Rstudio versión 1.2.1335 (R Core Team 2019).

En un primer lugar, se hizo un nuevo filtrado, eliminando el mayor número de datos ausentes posibles, y se estimó la frecuencia de recombinación, además de comprobarse si existía intercambio de alelos (del inglés, *switching allele*). Para generar el primer borrador de mapa, se agruparon los marcadores en un mismo grupo de ligamiento si la fracción de recombinación era menor o igual a 0,35 y tenían un LOD mayor o igual a 6,5. Además, se

eliminaron aquellos grupos de ligamiento con un bajo número de marcadores, obteniéndose 8 grupos de ligamiento. No obstante, se observaron elevadas distancias genéticas, por lo que se examinaron con más detalle los marcadores e individuos estudiados.

Se eliminaron los marcadores con el mismo patrón genotípico en todos los individuos, ya que se asignarían sobre la misma localización del genoma y, de esta forma, se reduciría la complejidad del mapa de ligamiento. Asimismo, se descartaron marcadores que mostraran doble recombinación cromosómica e individuos con una recombinación cromosómica excesivamente elevada, para evitar problemas debido a muestras que presenten errores de genotipado o que realmente no pertenezcan a la población estudiada (Broman *et al.* 2019).

Al emplear el paquete qtl, deben ordenarse los marcadores de cada grupo de ligamiento manualmente, por lo que se utilizó el algoritmo MSTmap (Wu *et al*. 2008) que organiza los grupos de ligamiento, y ordena los marcadores en poco tiempo, utilizando la función de Kosambi (Kosambi 1944) para el cálculo de distancias genéticas.

Para representar gráficamente el mapa de ligamiento se utilizó el paquete de R LinkageMapView (Ouellette *et al*. 2018). Todos los análisis de R se ejecutaron en un servidor del Prof. Antonio Matas en la Universidad de Málaga con cálculo QHR Lightning TR4 XL con AMD ThreadRipper 2990WX (32 cores, 64 threads), placa chipset AMD X399, tarjeta gráfica AMD R5 230, 128 Gb RAM DDR4, SSD M.2 512 Gb y 4x 8Tb Seagate en RAID 10.

Resultados

Genotipado por secuenciación (GBS) y llamamiento de variantes

Se genotiparon 58 muestras, produciendo 449,52 millones de lecturas brutas. obteniéndose tras el procesado, 390,32 millones de lecturas filtradas. El 98,7 % de estas lecturas fueron cartografiadas en una versión inicial del genoma.

Finalmente, tras eliminar los genotipos repetidos de los que se obtuvieron menor número de lecturas, los individuos de la F1, y filtrar aquellas variaciones con una proporción de datos ausentes en más del 80 % de los individuos, se retuvieron 1.656.918 variaciones (SNPs, indeles (inserciones y deleciones) y polimorfismos multinucleotídicos (MNPs)).

Genotipado y elaboración de mapa genético

Tras eliminar aquellos SNPs no biálelicos, y aquellos que se localizaron dentro de una distancia menor a 1.000 pares de bases, se obtuvieron 82.045 SNPs. De estos, 28.919 (35,25 %) se descartaron al mostrar el mismo genotipo en ambos parentales, al igual que las 15.914 (19,40 %) variantes a las cuales les faltaba el genotipo de 'Fino de Jete', y 19.612 (23,90 %) variantes a las cuales les faltaba el genotipo de 'Thai seedless'. De esta manera, se obtuvieron 17.600 (21,45 %) marcadores. Posteriormente, tras eliminar aquellas variantes que mostraron valores distorsionados respecto a la segregación mendeliana esperada (1:2:1), se obtuvieron 7.609 (9,27 %) SNPs, de los cuales el 23,7 % fueron homocigotos AA ('Fino de Jete'), 50,3 % heterocigotos AB y 26 % homocigotos BB ('Thai seedless') (Figura 3.2), manteniéndose 3.656 marcadores para generar el mapa genético tras eliminar el máximo de datos ausentes por marcador.

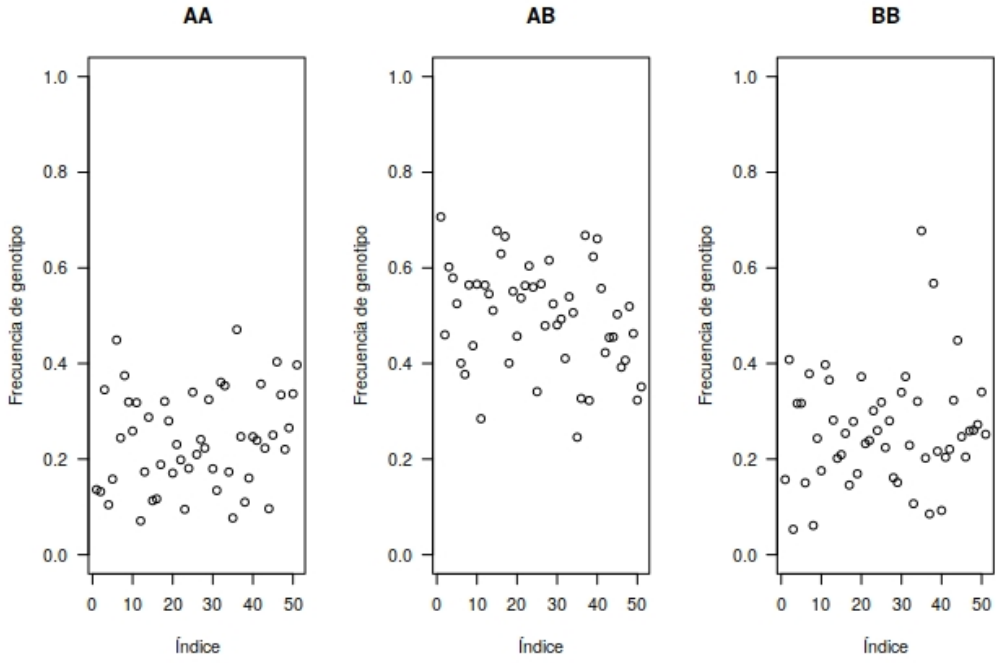


Figura 3.2. Frecuencias de genotipo en cada individuo.

En el primer borrador del mapa generado se obtuvieron 10 grupos de ligamiento, de los cuales 2 fueron eliminados al estar formados por un número de marcadores muy bajo (todos compuestos por un marcador o dos) y considerarse posible ruido de secuenciación. Finalmente se obtuvieron 8 grupos compuestos por 3.653 SNPs. Su distancia varió de 70 a

15.690 cM, mostrando una distancia total del mapa muy sobreestimada. Sin embargo, debe de tenerse en cuenta que los marcadores no estaban ordenados. No obstante, tras un filtrado más estricto, se obtuvieron 656 SNPs por la eliminación de aquellos marcadores que mostraran una doble recombinación elevada, los individuos que mostraron un entrecruzamiento cromosómico elevado respecto al resto de individuos (> 400) (Figura 3.3), y aquellos SNPs monomórficos, pues estos se localizarían en la misma ubicación del genoma. De estos 656, 20 % resultaron ser homocigóticos AA ('Fino de Jete'), 56,8 % heterocigotos AB y 23,3 % homocigóticos BB ('Thai seedless').

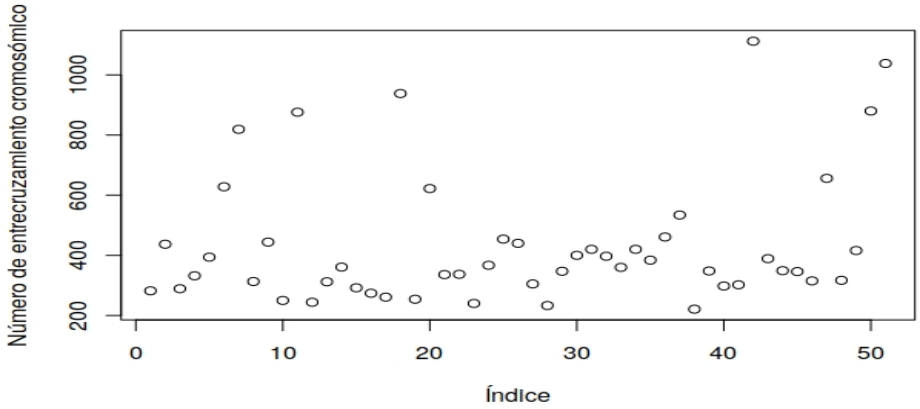


Figura 3.3. Número de entrecruzamientos observados en cada individuo**.**

Posteriormente, tras el filtrado anterior, y la eliminación de aquellos grupos de ligamiento formados por pocos marcadores (todos compuestos por un marcador, y algunos grupos que contenían menos de 10 marcadores de los cuales no todos se encontraban vinculados entre sí), se generó un nuevo mapa genético mediante el uso de datos genotípicos de la población F2 (33 individuos). Este mapa constó de 550 SNPs distribuidos en 8 grupos de ligamiento (Tabla 3.2; Figura 3.4). Los grupos de ligamiento tenían entre 18 y 129 marcadores, y su tamaño varió entre 37,9 cM (LG8) y 326,1 cM (LG2), con una distancia genética total de 1.387,8 cM. La distancia genética media entre marcadores fue de aproximadamente 2,6 cM y la distancia máxima entre los marcadores varió entre 3 cM (LG7) y 6,4 cM (LG2 y LG3).

Tabla 3.2. Tabla resumen de los grupos de ligamiento generados.

Grupos de ligamiento	Nº de marcadores	Longitud (cM)	Espacio medio entre marcadores	Espacio máximo entre marcadores
LG1	96	246,0	2,6	6,3
LG2	119	326,1	2,8	6,4
LG3	129	324,7	2,5	6,4
LG4	66	161,5	2,5	4,9
LG5	46	104,1	2,3	4,9
LG6	39	97,2	2,6	4,9
LG7	37	90,3	2,5	5,2
LG8	18	37,9	2,2	3,0
TOTAL	550	1.387,8		

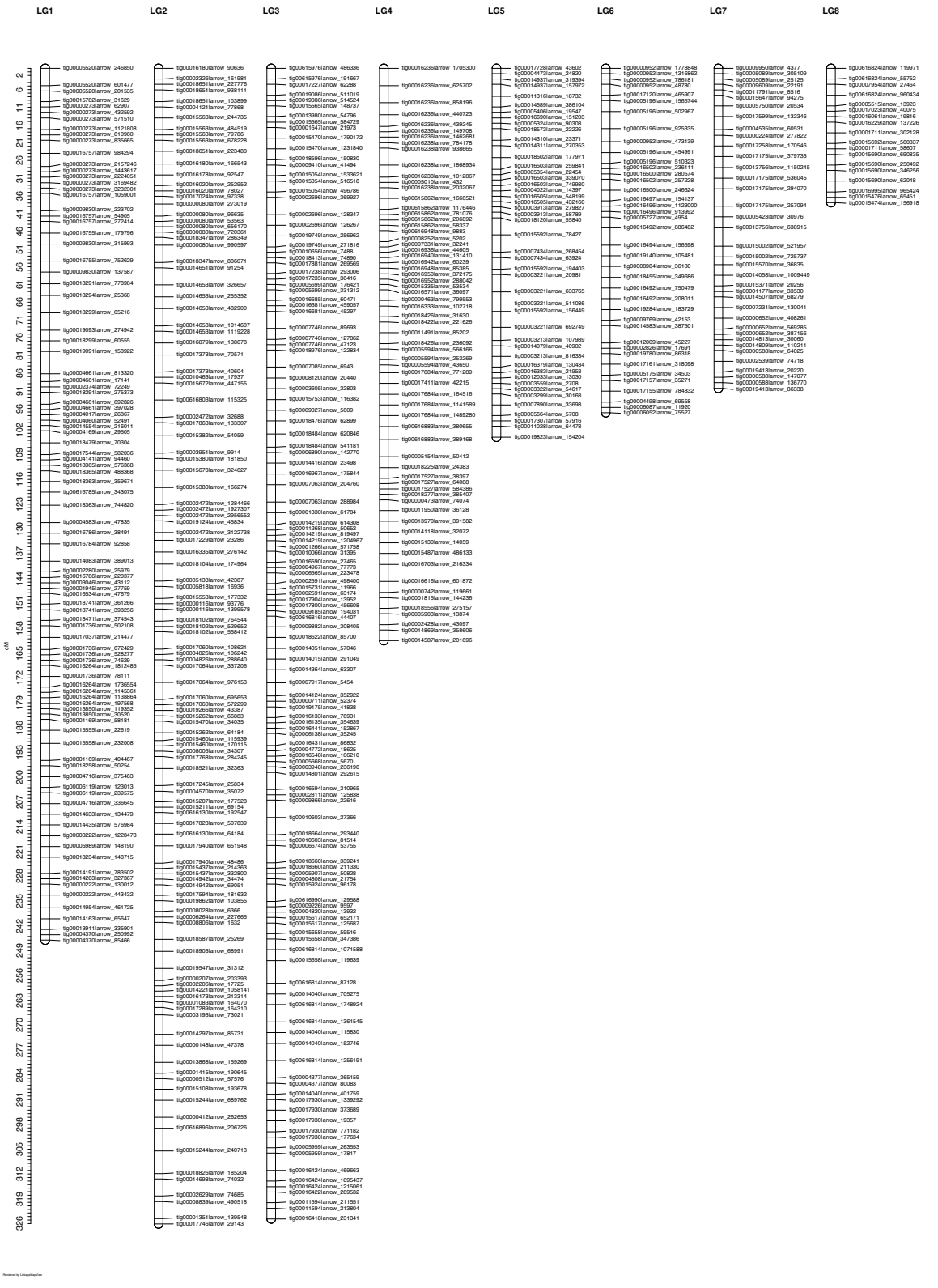


Figura 3.4. Mapa de ligamiento de la población F2 (´Fino de Jete´ x 'Thai seedless'). Cada grupo de ligamiento (LG) se muestra con los marcadores que los conforman. Las distancias genéticas, en centiMorgans (cM), se encuentran indicadas a la izquierda.

Discusión

Los mapas genéticos son una herramienta fundamental para acelerar y optimizar programas de mejora (Tanksley & McCouch 1997; Collard *et al*. 2005; Carrasco *et al.* 2018; Hirao *et al.* 2019) y se han desarrollado en numerosos cultivos.

Un mapa genético nos concede una representación reducida del genoma, esencial para numerosas aplicaciones como la identificación de marcadores asociados a un rasgo, selección asistida por marcadores (MAS), estudios genéticos comparativos (Da Silva Linge *et al*. 2018) y orientación de los cóntigos en el ensamblaje del genoma (Schneider 2005; Varshney *et al*. 2016; Bourke *et al.* 2018). Actualmente, el desarrollo de mapas genéticos se ve facilitado por la gran cantidad de datos que se pueden generar mediante NGS. Así, metodologías como el genotipado por secuenciación (GBS) permiten generar un gran número de marcadores (SNPs), en poco tiempo, y por un coste relativamente bajo, convirtiéndose en una herramienta fundamental para elaborar mapas genéticos (Hussain *et al*. 2017; Gabay *et al.* 2018). Como consecuencia, los SNPs son los marcadores más empleados debido a su gran abundancia y heredabilidad (Zhang *et al.* 2020).

La mayoría de los mapas de ligamiento en frutales se basan en una progenie F1 debido al largo periodo intergeneracional que dificulta la obtención de generaciones más avanzadas. Algunos ejemplos recientes sobre el desarrollo de mapas genéticos basados en poblaciones F1 los encontramos en mango (Kuhn *et al.* 2017), peral (Gabay *et al.* 2018) y nogal (Aradhya *et al.* 2019).

En especies frutales infrautilizadas como el chirimoyo y otras especies del género *Annona*, la escasez de datos moleculares dificulta el desarrollo de mapas de ligamiento saturados. Así, hasta este momento, se han generado dos mapas genéticos dentro del género *Annona*, empleando SSRs y una población F1 ‘Fino de Jete’ x ‘Thai seedless’ desarrollada en el IHSM-UMA-CSIC “La Mayora” (Escribano 2007; Martín 2013). ‘Fino de Jete’ es la principal variedad de chirimoyo en España, y reúne unas excelentes condiciones organolépticas, elevada producción, y presenta una buena adaptación al medio. Sin embargo, presenta un alto número de semillas. Por el contrario, ‘Thai seedless’ debido a una mutación natural produce frutos sin semillas (solo presenta algunas semillas rudimentarias). El cruce producido entre ambas especies genera un híbrido fértil conocido como atemoya, de interés

comercial en algunos países como Australia, Egipto, Israel, Sudáfrica o Taiwan (Hormaza *et al.* 2020).

Genotipado y elaboración de un mapa genético

En este estudio, se genotiparon 55 individuos de una población F2 generada a partir de la autofecundación de dos individuos de la población F1 obtenida mediante el cruzamiento de 'Fino de Jete' x 'Thai seedless' ('JT7' y 'JT20'), y sus parentales. A pesar de que este número de genotipos puede considerarse bajo para generar un mapa genético de calidad, ya que en general se recomienda alrededor de 100 individuos de una población F2 (Schneider 2005), cabe destacar la dificultad de obtener una población F2 en estas especies. La población usada en este estudio es una de las pocas poblaciones F2 existentes en el género *Annona*. Además, en otros trabajos en cultivos leñosos se han desarrollado mapas genéticos basados en un número parecido de individuos de poblaciones F2 como, por ejemplo, en cerezo (n=67) (Calle *et al.* 2018) o en melocotonero (n= 57) (Bielenberg *et al.* 2015).

Del total de polimorfismos mononucleótidos detectados (82.045) la gran mayoría no se utilizaron en la elaboración del mapa genético, de forma similar a lo que se ha descrito en otras especies (Hussain *et al*. 2017; Carrasco *et al*. 2018). De hecho, a pesar de que el genotipado por secuenciación (GBS) se considera una herramienta útil y clave para el desarrollo de mapas genéticos, muestra una serie de limitaciones, como los errores de genotipado o la cantidad de datos ausentes que genera (Hussain *et al*. 2017; Carrasco *et al.* 2018). El uso de enzimas de restricción alternativas a la hora de generar las genotecas podría reducir los datos ausentes. En este trabajo, se utilizó la enzima de restricción ApeKI ya que permite detectar un mayor número de marcadores; sin embargo, a su vez, genera una mayor proporción de datos ausentes (Bastien *et al.* 2018).

De todos los marcadores generados solo 7.609 marcadores moleculares mostraron una segregación mendeliana esperada (1:2:1). La distorsión en la segregación mendeliana es muy común en plantas y animales, especialmente en cruces interespecíficos (Dai *et al.* 2017). Su origen puede ser el resultado de errores en el genotipado, datos ausentes (Liu *et al.* 2016b), elevado grado de divergencia entre los parentales (Ren *et al.* 2012; Jenczewski *et al.,* 1997; Zamir & Tadmor 1986), presencia de alelos letales (Pillen *et al*. 1993), competencia gamética o selección a nivel pre o postzigótico (King *et al.* 1991, Lyttle 1991). En plantas, el porcentaje de distorsión de segregación varía significativamente dependiendo de las especies

(Wu *et al.* 2019a), cruces y tipos de marcadores (Zhang *et al.* 2019a). A la hora de comparar nuestros resultados con otras especies, en este trabajo se muestra una elevada distorsión de segregación, lo que ya ha sido descrito en otros estudios previos en el género *Annona* (Escribano 2007; Martín 2013).

Desafortunadamente, no existen trabajos en los que se hayan generado mapas de ligamiento a partir de una población F2 dentro del género *Annona*. En comparación con los mapas genéticos elaborados a partir de una población interespecífica F1 ('Fino de Jete' x 'Thai seedless'), el mapa que se presenta en este capítulo mostró una distancia genética total más elevada. Sin embargo, dicha distancia debe de ser considerada con cautela, pues podría estar sobreestimada. La distancia genética media entre marcadores fue de aproximadamente 2,6 cM, siendo similar a las encontradas en otras especies (Bielenberg *et al*. 2015), y mucho menor a las distancias medias entre loci reportadas en mapas anteriores (10,6 cM) (Martín 2013). Estas diferencias eran esperables, debido al mayor número de datos genotípicos empleados en la elaboración del mapa.

El número de grupos de ligamiento propuesto, coincidió con los grupos de ligamiento mostrados en trabajos anteriores (Escribano 2007; Martín 2013), que se corresponde con el número haploide de cromosomas para la mayoría de las especies del género *Annona* según Bowden (1945, 1948). Además, en ambos mapas de ligamiento, el grupo de ligamiento de menor tamaño está compuesto por una proporción de marcadores similar a la mostrada en los mapas producidos en trabajos anteriores. A pesar de ello, recientemente, Martín *et al.* (2019) y Falistocco & Ferradini (2020) han propuesto la existencia de 7 cromosomas en *Annona cherimola,* tal como estaba descrito en *Annona squamosa*, apoyando trabajos previos (Kumar & Ranadive 1941; Asana & Adiata 1945; Thakur & Singh, 1965, 1969; Tanaka & Okada, 1972; Morawetz 1986), y justificando que la observación de 8 cromosomas haploides podría deberse a una confusión entre un satélite distante con un cromosoma (Martín *et al.* 2019).

Aunque se observa que un gran número de marcadores generados se encuentran ordenados en grupos de ligamiento coincidiendo con los cóntigos del genoma de referencia, lo que apunta hacia la solidez del mapa, las distancias génicas podrían estar sobrestimadas y deben de considerarse con precaución hasta que se obtengan más datos complementarios. Se necesitan trabajos adicionales combinando mapas de ligamiento saturados con observaciones de cromosomas al microscopio para elucidar este punto. Hay que tener en cuenta que la observación al microscopio de los cromosomas en especies del género *Annona* no es fácil por

el pequeño tamaño de los mismos. La hibridación in situ (FISH) ha sido una herramienta muy importante para la identificación de cromosomas de muchas plantas (Jiang 2019) y permite estimar las distancias espaciales entre loci mediante la visualización de los cromosomas, por lo que sería de interés su aplicación en estudios en el género *Annona*. Estudios adicionales, como la captura de la conformación de los cromosomas, Hi-C, ayudarían a esclarecer la estructura y detalles del genoma incluyendo la frecuencia con la que un par de loci se acerca entre sí en el espacio 3D (Abbas *et al.* 2019).

En este trabajo se presenta por primera vez un mapa de ligamiento de atemoya, generado a partir de una población F2, mediante el uso de marcadores tipo SNPs, ofreciendo información sobre la organización del genoma de *Annona*, y siendo un comienzo fundamental para simplificar el proceso de mejora de estos cultivos en los que, debido al largo periodo juvenil, este tipo de herramientas son muy necesarias.

DISCUSIÓN GENERAL

Discusión general

Los frutales subtropicales, como el aguacate (*Persea americana*) y el chirimoyo (*Annona cherimola*) son esenciales para la alimentación humana en muchos países, y su importancia comercial está incrementando en los últimos años. Actualmente, su producción se centra en un genotipo concreto, en aguacate a nivel mundial cv. Hass y en chirimoyo a nivel nacional cv. Fino de Jete, siendo el desarrollo de nuevas variedades una necesidad en ambos cultivos. Actualmente se están llevando a cabo varios programas de mejora para ambas especies aunque su progreso se ve obstaculizado por la falta de información básica, debido a limitaciones en el desarrollo de recursos genéticos para estos cultivos. De hecho, esta restricción dificulta incluso la identificación de genotipos, siendo esta limitación común en especies tropicales y subtropicales. Por ello, los resultados obtenidos en este trabajo permiten avanzar en el conocimiento genómico de estos dos frutales.

En primer lugar, se ha desarrollado un elevado número de marcadores moleculares para 71 genotipos de aguacate que representan las tres razas botánicas tradicionales (Mexicana, Guatemalteca y Antillana) permitiendo su caracterización y análisis poblacional (Capítulo 1). Por otro lado, se ha generado el primer ensamblaje del genoma de chirimoyo, lo que ha permitido predecir de forma aproximada el número de genes que posee la especie y sus características (Capítulo 2). Finalmente, se ha construido un mapa de ligamiento a partir de una población F2, facilitando una primera caracterización del genoma del atemoyo, un híbrido fértil entre '*Annona cherimola*' x '*Annona squamosa*' con implicaciones para estudios genómicos en chirimoyo (Capítulo 3). Los resultados obtenidos abren las puertas a futuras aplicaciones para la mejora del aguacate y del chirimoyo, junto a su utilidad para avanzar en el conocimiento de la evolución de angiospermas, cuya rápida radiación constituía el "misterio abominable" de Darwin.

Caracterización genómica de aguacate (*Persea americana* Mill.)

Conocer la diversidad genética, la estructura poblacional y el fenotipado y genotipado de una especie es fundamental para el desarrollo de programas de mejora modernos. En aguacate,

aunque se han generado distintos tipos de marcadores en los últimos años (Furnier *et al*. 1990; Lavi *et al*. 1991; Lavi *et al.* 1994; Sharon *et al.* 1997; Borrone *et al*. 2007; Alcaraz & Hormaza 2007; Chen *et al*. 2008; Chen *et al*. 2009; Gross-German & Viruel 2013; Guzmán *et al*. 2017; Kuhn *et al*. 2019; Ge *et al*. 2019a, 2019b; Rubinstein *et al*. 2019) sigue existiendo un vacío de información genómica, y por tanto, la necesidad de incrementar el número de marcadores disponibles que permitan caracterizar genotipos y acelerar los programas de mejora.

En la actualidad, una de las metodologías más importante para los estudios genómicos de plantas es el genotipado por secuenciación (GBS) (Elshire *et al.* 2011; Melo *et al.* 2016; D´Agostino *et al*. 2018) que permite el desarrollo de un alto número de marcadores moleculares tipo SNP. A pesar de que los arrays de SNPs pueden proporcionar un alto número de marcadores y su análisis sea más fácil, GBS es mucho más económico (Scheben *et al*. 2017, 2018) además de ser un sistema abierto a nuevas variantes y, por tanto, ha facilitado la disponibilidad de marcadores moleculares en cultivos (Le Nguyen *et al*. 2019).

En este trabajo, el desarrollo de un borrador de genoma de aguacate (cv. Hass) y la secuenciación de librerías de GBS ha permitido generar 7.108 SNPs que facilitaron la caracterización, el estudio de la diversidad genética y la estructura poblacional de 71 genotipos, entre los que se encuentran accesiones obtenidas en distintos programas de mejora, variedades comerciales y accesiones locales con posible interés como nuevos portainjertos y/o variedades. Tras los análisis realizados (ACP, NJ, ADMIXTURE, STRUCTURE, ADPC) se pudo observar una clara separación según las razas, obteniéndose cuatro grupos (Mexicano, Guatemalteco, Antillano e híbridos Guatemalteco x Mexicano [GxM]). Sin embargo, en ciertos análisis (STRUCTURE y ADPC) no se encontró una clara distinción entre genotipos de raza Guatemalteca e híbridos GxM. Esto puede ser debido a que los genes de raza Guatemalteca predominan en el germoplasma actual de aguacate, ya que ese tipo de genotipos han sido utilizados como parentales en los programas de mejora y selección de la especie debido a varias de sus características de gran interés agronómico (huesos pequeños, una madurez tardía y adaptación a climas subtropicales) (Chanderbali *et al*. 2013). Gracias a estos análisis, se pudo asignar a los genotipos con pedigrí poco claro o desconocido su posible origen y, de acuerdo con trabajos anteriores, se demostró que la mayoría correspondían a híbridos entre la raza Guatemalteca y la raza Mexicana. El grupo compuesto

principalmente por genotipos Antillanos, que son los más adaptados a condiciones tropicales reveló una clara separación respecto al resto.

Los resultados obtenidos han puesto de manifiesto que el alto número de marcadores moleculares generados permite un eficiente genotipado y discriminación según el origen racial de cada accesión. La utilización de estos marcadores en futuros estudios facilitará la caracterización y el manejo de bancos de germoplasma junto a la implementación de protocolos de mejora más eficientes, siendo un recurso público para toda la comunidad de aguacate.

Ensamblaje del genoma del chirimoyo (*Annona cherimola* Mill.)

Los ensamblajes de genomas son una herramienta clave para la mejora y el conocimiento de la estructura genética de los cultivos (Yuan *et al*. 2017), aunque hasta hace pocos años estos estudios estaban restringidos a especies modelo o de alta importancia comercial, por lo que eran muy escasos en especies frutales infrautilizadas, como el chirimoyo.

Generar el primer genoma del chirimoyo presenta cierta complejidad debido a la proporción de elementos repetitivos, alto nivel de heterocigosidad y el tamaño de genoma. A pesar de estos inconvenientes, en este trabajo se presenta el primer ensamblaje y anotación del genoma del chirimoyo cv. Fino de Jete. La disponibilidad de este recurso abre un amplio abanico de oportunidades, como el estudio de genes de la especie o la variación genómica. Además, paralelamente, permitirá avanzar en el conocimiento de la evolución de las angiospermas basales.

El tamaño estimado del genoma secuenciado del chirimoyo es aproximadamente 1,17 Gb, considerándose un tamaño medio dentro de las angiospermas, siendo el tamaño más pequeño de 1C = 0.06 Gb en *Genlisea tuberosa* (Fleischmann *et al*. 2014) y el más grande el genoma de *Paris japonica* con 1C = 148,8 Gb (Pellicer *et al*. 2018). El ensamblaje final se compone de 15.076 secuencias, con un N50 de 171,3 kb, de manera que, a pesar de ser un ensamblaje fragmentado en comparación con otros genomas publicados recientemente (Chen *et al*. 2019c; Hu *et al*. 2019; Zhang *et al*. 2019b, 2019c; Chaw *et al*. 2019; Yang *et al*. 2020b; Chen *et al*. 2020), se estima que se encuentra bastante completo en términos de genes (96%).

Los eventos de duplicación se consideran fundamentales para la adaptación y la especiación (Cui *et al*. 2006; Soltis *et al.* 2015; Van de Peer *et al.* 2017), siendo más abundantes y frecuentes en plantas que en animales, sobre todo en las angiospermas (Soltis *et al*. 2015). De hecho, alrededor del 35 % de las angiospermas se consideran poliploides recientes; sin embargo, eventos pasados de poliploidía también se observan a lo largo de la filogenia de las angiospermas (Godden *et al.* 2019). En este trabajo tras realizar la anotación estructural del genoma de chirimoyo, se pudo observar que parte de los genes de esta especie se encuentran asociados a un evento de duplicación o hibridación reciente. Este fenómeno parece haberse producido prácticamente a la vez que el evento reciente del genoma *Cinnamomum* (Lauraceae, Laurales) (Chaw *et al*. 2019), y posterior al producido en el genoma de *Liriodendron* (Magnoliaceae, Magnoliales) (Chen *et al*. 2019c). Cui *et al.* (2006) demostraron la existencia de un evento de duplicación común en el clado Magnoliid (que incluye los órdenes Canellales, Laurales, Magnoliales, y Piperales) antes de la diversificación de las Lauráceas y las Magnoliáceas; sin embargo, en estudios posteriores, tras el desarrollo de genomas de referencia (Rendón-Anaya *et al.* 2019; Chen *et al.* 2019c; Chaw *et al.* 2019), se han observado eventos de poliploidización compartidos entre el genoma de *Persea* y *Cinnamomun* e independientes entre los clados principales entre los que se incluye el clado Magnoliid (Yang *et al*. 2020b).

Aunque en los últimos años se han realizado algunos estudios filogenéticos con genomas de angiospermas primitivas, siguen existiendo ciertas incongruencias respecto a la relaciones entre las angiospermas más basales (básicamente el clado ANITA), el clado magnoliid, eudicotiledóneas y monocotiledóneas (Chen *et al.* 2019c; Chaw *et al*; 2019; Hu *et al*, 2019; Soltis & Soltis 2019; Yang *et al*. 2020b). Hasta el momento dentro de las angiospermas más basales [el clado ANITA que consta de los órdenes Amborellalles con una única especie (*Amborella trichopoda*), Nymphaeales y Austrobaileyales] se dispone del genoma de *Amborella trichopoda* (Amborellaceae, Amborellales) (Albert *et al.* 2013) y del de dos especies de Nymphaeales (*Nymphahea colorata* (Zhang *et al.* 2019c) y *Euryale ferox* (Yang *et al.* 2020b). Dentro del clado Magnoliid, del genoma del aguacate (Lauraceae, Laurales) (Rendón-Anaya *et al.* 2019), del liriodendron (*Liriodendron chinense*) (Magnoliaceae, Magnoliaes) (Chen *et al.* 2019c), del alcanforero (*Cinnamomum kanehirae*) (Lauraceae, Laurales) (Chaw *et al.* 2019), de la pimienta común (*Piper nigrum*) Piperaceae, Piperales) (Hu *et al*. 2019), y de *Litsea cubeba* (Chen *et al.* 2020). Por lo tanto, son necesarios estudios genómicos adicionales en las familias que se encuentran dentro del clade Magnollid y, en este

sentido, el primer genoma aquí presentado en la familia Annonaceae puede ayudar a entender la historia evolutiva de las angiospermas más primitivas (APG IV 2016). (Figura D.1).

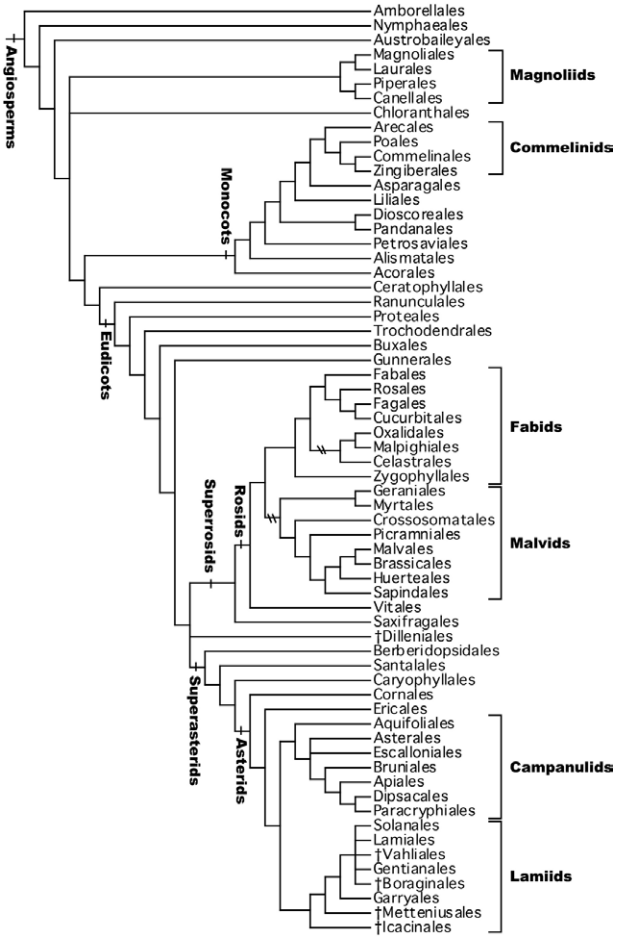


Figura D.1. Relaciones filogenéticas de los ordenes APG IV y algunas familias respaldadas por porcentajes jackknife / bootstrap> 50 o probabilidad posterior bayesiana > 0.95 en angiospermas. Las topologias alternativas que representan incongruencia entre los resultados nucleares / mitocondriales y plastídicos para el clado Celastrales / Oxalidales / Malpighiales (COM) se indican con (\\). † Ordenes recientemente reconocidos en APG. Fuente: APG IV (2016).

Aunque el desarrollo del genoma de chirimoyo es un recurso esencial para el estudio de la especie, presenta la limitación de que únicamente se ha secuenciado un genotipo por lo que no representa la diversidad de la especie (Rasheed & Xia 2019). Por ello, la resecuenciación de baja cobertura con el objetivo de detectar la diversidad se ha convertido en una estrategia bastante común en cultivos tras secuenciar un genoma de referencia (Kersey 2019). De hecho, aunque las "metodologías de representación reducida" como GBS (Elshire *et al*. 2011), RAD-seq (Baird *et al*. 2008) o ddRADseq (Peterson *et al*. 2012) son herramientas excepcionales, la resecuenciación permite una mayor detección de variaciones genómicas (Ekblom & Wolf 2014).

En el género *Annona* y en la familia de las Anonáceas en general, los estudios de diversidad genética son muy limitados, por lo que este genoma podría abrir nuevas oportunidades para su estudio. En otros cultivos el desarrollo de nuevos marcadores o la secuenciación de genomas han acelerado los programas de mejora e incrementado la información disponible a nivel genómico. Algunos ejemplos en cultivos leñosos son los siguientes: en manzano el desarrollo de un genoma de referencia de calidad permitió localizar el alelo responsable de la coloración roja de la piel (Zhang *et al*. 2019b); el borrador del genoma del cocotero facilitó la identificación de rasgos de interés agronómico como la resistencia a enfermedades (Lantican *et al*. 2019); la resecuenciación de 60 accesiones de cítricos y la disponibilidad de un genoma de referencia proporcionó información sobre el origen y la evolución del género *Citrus,* cuya relación filogenética era confusa debido a su alto nivel de hibridación interespecífica (Wu *et al*. 2018); el genotipado de distintas accesiones y la disponibilidad de un genoma de referencia reveló un locus candidato relacionado con el sabor del arándano (Ferrão *et al.* 2020); la resecuenciación de 44 accesiones de melocotonero proporcionó genes candidatos asociados a rasgos de interés agronómico como el sabor o la textura (Yu *et al.* 2018) y el desarrollo del genoma de referencia de almendro ha permitido analizar las diferencias existentes con el melocotonero (Alioto *et al*. 2020) o detectar el factor de transcripción responsable del amargor de algunas variedades (Sánchez-Pérez *et al*. 2019).

Como se ha mencionado previamente, los avances en las herramientas y la tecnología de secuenciación y computación aumentan la calidad de los genomas de referencia en plantas (Jung *et al*. 2019). Una estrategia que hasta la fecha ha sido bastante empleada para mejorar el ensamblaje de genomas, es el uso de mapas de ligamiento (Xu *et al.* 2013; Shirasawa *et al*. 2017; Varshney *et al.* 2017). En el caso del chirimoyo, como ocurre con la mayoría de los cultivos infrautilizados, no se encuentran disponibles mapas de ligamiento saturados generados a partir de poblaciones intraespecíficas; sin embargo, el uso del mapa generado en el Capítulo 3 podría ayudar a mejorar el ensamblaje generado.

El genoma de chirimoyo que se presenta en este trabajo es una primera aproximación, por lo que en un futuro próximo posiblemente se disponga de un genoma de referencia de mayor calidad o incluso de un pan-genoma, que aglutine información de la resecuenciación de los genomas de un gran número individuos (Chen *et al*. 2019a; Rasheed & Xia 2019), que facilita el entendimiento de las variaciones de una especie o género (Chen *et al*. 2019a). Lo mismo podría ocurrir con el borrador del genoma de aguacate (Capítulo 1). De hecho, ya se

está trabajando para generar un genoma de referencia a partir del borrador generado en este estudio.

Elaboración de un mapa de ligamiento en el género *Annona*

La elaboración de los mapas genéticos es una de las numerosas aplicaciones de los marcadores moleculares. Estos mapas son esenciales para la mejora de frutales, pues facilitan la asociación marcador-fenotipo a través del mapeo cuantitativo (QTL), el descubrimiento de genes asociados a un carácter de interés o la selección asistida por marcadores (Da Silva Linge *et al.* 2018). El progreso en las metodologías de genotipado ha facilitado y mejorado notablemente su construcción en numerosas especies (Jaganathan *et al.* 2020).

En el caso del género *Annona*, los mapas genéticos disponibles hasta la fecha han sido limitados y de baja resolución. Los mapas existentes se desarrollaron a partir de poblaciones F1 segregantes, como resultado de cruzamientos interespecíficos entre 'Fino de Jete' x 'Thai seedless' (Escribano 2007; Martín 2013) e intraespecífico entre 'Bonita' x 'Fino de Jete' (Martín 2013). No obstante, la baja heterocigosidad observada en el parental 'Thai seedless' imposibilitó la elaboración de un mapa genético para este genotipo (Martín 2013).

En este trabajo, con el fin de generar un mapa genético con mayor eficiencia de cartografiado para los genotipos 'Fino de Jete' y 'Thai seedless', se empleó una población F2, generada a partir de la autofecundación de dos individuos de la primera generación (F1) empleada en trabajos anteriores (Escribano 2007; Martín 2013). El interés de esta población se basa en la introgresión del carácter de ausencia de semillas del genotipo 'Thai seedless' en el genotipo 'Fino de Jete'. Para la construcción de este mapa, se genotiparon 58 individuos mediante la metodología GBS (al igual que en el Capítulo 1) y se utilizó una de las primeras versiones del genoma de chirimoyo generadas en el Capítulo 2, con el fin de reducir el número de marcadores erróneos. En comparación con otras especies, se observó una elevada distorsión de segregación. No obstante, resultados similares han sido descritos en otros trabajos del género *Annona* (Perfectii & Pascual, 1996; Escribano 2007; Martín 2013).

El mapa desarrollado agrupó un total de 550 marcadores, con una distancia media entre marcadores de 2,6 cM, siendo bastante inferior a la distancia media entre loci mostrada en los estudios previos (10,6 cM) (Martín 2013). Se generaron 8 grupos de ligamiento, tal y como

ocurrió en los mapas construidos anteriormente (Escribano 2007; Martín 2013). Sin embargo, otros estudios (Martín *et al.* 2019) han propuesto la existencia de 7 cromosomas en *Annona cherimola* y en *Annona squamosa*, indicando que x = 7 podría haber surgido tras la pérdida de un cromosoma. Estas observaciones contradictorias, podrían estar asociadas con la duplicación reciente del genoma (Capítulo 2) y una posterior diploidización, lo que produciría cierta inestabilidad cromosomal. La disponibilidad de los estudios genómicos realizados en este trabajo abre la puerta a utilizar otras herramientas como Hi-C para poder conocer con exactitud el número real de cromosomas en estas especies.

La coincidencia de los grupos de ligamiento con los cóntigos del genoma utilizado (Capítulo 2) sugiere la solidez del mapa; sin embargo, la distancia total mostrada es bastante elevada, por lo que debería de ser considerada con cautela hasta la obtención de datos complementarios.

Los resultados obtenidos en este trabajo han permitido diferenciar genotipos de aguacate, estudiar su diversidad, elaborar el primer mapa de ligamiento a partir de una población F2 en el género *Annona* y generar un genoma de referencia para el chirimoyo. Estos estudios son fundamentales para comprender la estructura del genoma y para llevar a cabo futuros programas de mejora en ambas especies. En definitiva se puede concluir que, gracias a las aproximaciones genómicas de este estudio, se han podido incrementar los recursos genéticos de aguacate y chirimoyo, acortando las diferencias respecto a otros cultivos (Chen *et al*. 2019a; Benevenuto *et al*. 2019). No obstante, es necesario tener en cuenta que para llevar a cabo estudios de asociación también es esencial la información fenotípica (Hayward *et al*. 2015; Shakoor *et al*. 2017; Scheben *et al*. 2018). De hecho, actualmente se sigue trabajando en el fenotipado de estas especies, con el objetivo de que en un futuro próximo esta información junto a la genómica permita facilitar la mejora de estos cultivos e incrementar el conocimiento sobre ellos. Los avances exponenciales que han tenido lugar en los últimos años con el desarrollo de herramientas genómicas, hacen que el fenotipado de calidad sea actualmente el mayor cuello de botella (Furbank & Tester 2011; Araus & Cairns 2014; Minervini *et al*. 2015; Zhao *et al.* 2019; Yang *et al.* 2020a) para aplicar los avances genómicos en avances prácticos, poniendo en valor la necesidad de disponer de colecciones de germoplasma diversas y bien caracterizadas.

CONCLUSIONES

Conclusiones

1. El desarrollo del borrador del genoma del cultivar Hass y los marcadores moleculares SNPs desarrollados mediante metodología de representación reducida han facilitado la caracterización de un conjunto de genotipos de aguacate (*Persea americana* Mill.), a través del estudio de su diversidad y estructura poblacional.

2. El borrador del genoma del cultivar Hass, a pesar de ser un ensamblaje altamente fragmentado, ha favorecido la identificación de variantes a partir de las secuencias producidas mediante genotipado por secuenciación (GBS).

3. Las relaciones genéticas entre los cultivares de aguacate estudiados determinaron una agrupación entre genotipos que, en general, se corresponde con las razas previamente descritas (Antillana, Guatemalteca y Mexicana), a excepción de un grupo híbrido entre Guatemaltecos x Mexicanos (GxM).

4. Se ha llevado a cabo con éxito un primer ensamblaje y anotación del genoma del chirimoyo (*Annona cherimola* Mill.) mediante la combinación de los datos generados a través de dos plataformas de secuenciación (Illumina y PacBio). Este es el primer genoma desarrollado en la familia de las Anonáceas.

5. El análisis de duplicación del genoma detectó la existencia de un evento reciente de hibridación o duplicación en chirimoyo, lo que podría estar ocasionando cierta inestabilidad cromosomal en este cultivo.

6. La combinación de marcadores moleculares tipo SNP y del genoma de chirimoyo han facilitado el desarrollo del primer mapa de ligamiento en el género *Annona,* mediante el genotipado con GBS de una población F2 obtenida a partir de una población F1 interespecífica entre dos especies del género (‘Fino de Jete’ (*A.cherimola*) x ‘Thai seedless’ (*A. squamosa*).

BIBLIOGRAFÍA

Bibliografía

Abbas A, He X, Niu J, Zhou B, Zhu G, Ma T, Song J, Gao J, Zhang M Q, Zeng J. (2019) Integrating Hi-C and FISH data for modeling of the 3D organization of chromosomes. Nature Communications 10(1), 1-14.

Alaly, F Q, Liu X X, McLaughlin J L. (1999) Annonaceous acetogenins: recent progress. Journal of Natural Products 62, 504540.

Albert V A, Barbazuk W B, de Pamphilis C W, Der J P, Leebens-Mack J, Ma H, Palmer J D, Rounsley S, Sankoff D, e*t al.* (2013) The *Amborella* genome and the evolution of flowering plants. Science 342(6165), 1241089.

Alcaraz M L, Hormaza, J I. (2007) Molecular characterization and genetic diversity in an avocado collection of cultivars and local Spanish genotypes using SSRs. Hereditas 144, 244-253.

Alcaraz M L, Hormaza J I, Rodrigo J. (2013) Pistil starch reserves at anthesis correlate with final flower fate in avocado (*Persea americana*). PLoS ONE 8(10), e78467.

Alcaraz M L, Hormaza J I. (2014) Optimization of controlled pollination in avocado (*Persea americana* Mill., Lauraceae). Scientia Horticulturae 180, 79-85.

Alexander D H, Novembre J, Lange K. (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Research 19, 1655-1664.

Alioto T, Alexiou K G, Bardil A, Barteri F, Castanera R, Cruz F, Dhingra A, Duval H, Fernández i Martí A, Frias L, Galán B, García J L, Howad W, Gómez-Garrido J, Gut M, Julca I, Morata J, Puigdomènech P, Ribeca P, Rubio Cabetas M J, Vlasova A, Wirthensohn M, Garcia-Mas J, Gabaldón T, Casacuberta J M, Arús P. (2020) Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. The Plant Journal 101(2), 455-472.

Anagbogu C F, Bhattacharjee R, Ilori C, Tongyoo P, Dada K E, Muyiwa A A, Gepts P, Beckles D M. (2019) Genetic diversity and re-classification of coffee (*Coffea canephora* Pierre ex A. Froehner) from South Western Nigeria through genotyping-by-sequencing-single nucleotide polymorphism analysis. Genetic Resources and Crop Evolution 66(3), 685-696.

APG IV (2016) An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG IV. Botanical Journal of the Linnean Society 181(1), 1-20.

Aradhya M K, Velasco D, Wang J R, Ramasamy R, You F M, Leslie C, Dandekar A, Luo M C, Dvorak J. (2019) A fine-scale genetic linkage map reveals genomic regions associated with economic traits in walnut (*Juglans regia*). Plant Breeding 138(5), 635-646.

Aranzana M J, Illa E, Howad W, Arús P. (2012) A first insight into peach [*Prunus persica* (L.) Batsch] SNP variability. Tree Genetics & Genomes 8, 1359-1369.

Aranzana M J, Decroocq V, Dirlewanger E, Eduardo I, Gao Z S, Gasic K, Lezzoni A, Jung S, Peace C, Prieto H, Tao R, Verde I, Abbott A G, Arús P. (2019) *Prunus* genetics and applications after de novo genome sequencing: achievements and prospects. Horticulture Research 6(1), 1-25.

Araus J L, Cairns J E. (2014) Field high-throughput phenotyping: the new crop breeding frontier. Trends in Plant Science 19(1), 52-61.

Argout X, Salse J, Aury J, Guiltinan M J, Droc G, Gouzy J, Allegre M, Chaparro C, Legavre T, Maximova S N, Abrouk M, Murat F, Fouet O, Poulain J, Ruiz M, Roguet Y, Rodier-Goud M, Fernandes Barbosa-Neto J, Sabot F, Kudrna D, Ammiraju J S S, Schuster S C, Carlson J E, Sallet E, Schiex T, Dievart A, Kramer M, Gelley L, Shi Z, Bérard A, Viot C, *et al.* (2011) The genome of *Theobroma cacao*. Nature Genetics 43(2), 101.

Aronesty E. (2013) Comparison of sequencing utility programs. The Open Bioinformatics Journal 7, 1-8; 10.2174/1875036201307010001.

Arús P, Puigdomènech P. (2008) La genómica de plantas: una oportunidad para España. Laboratorio alternativas.

Asana J, Adiata R. (1945) The chromosome numbers in the family Annonaceae. Current Science 14, 74–75.

Ashworth V E T M, Clegg M T. (2003) Microsatellite markers in avocado (*Persea americana* Mill.): genealogical relationships among cultivated avocado genotypes. Journal of Heredity 94, 407-415.

Ashworth V E T M, Kobayashi, M C, De La Cruz M, Clegg M T. (2004) Microsatellite markers in avocado (*Persea americana* Mill.): development of dinucleotide and trinucleotide markers. Scientia Horticulturae 101, 255-267.

Avocado information database. Disponible en: https://www.myavocadotrees.com/beta-avocado.html (último acceso: 13 de septiembre del 2019).

Ayala F J, Kiger J A. (1984) Genética moderna. (Omega, S.A. Barcelona).

Baird N A, Etter P D, Atwood T S, Currey M C, Shiver A L, Lewis Z A, Selker E U, Cresko W A, Johnson E A. (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3(10).

Bairoch A, Apweiler R. (2000) The Swiss-Prot protein sequence database and its supplement trembl in 2000. Nucleic Acids Research 28, 45–48.

Ban S H, Choi C. (2018) Development of an apple F1 segregating population genetic linkage map using genotyping-by-sequencing. Plant Breeding and Biotechnology 6(4), 434-443.

Bansal V, Boucher C. (2019) Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going?. iScience 18, 37-41.

Bastien M, Boudhrioua C, Fortin G, Belzile F. (2018) Exploring the potential and limitations of genotyping-by-sequencing for SNP discovery and genotyping in tetraploid potato. Genome 61(6), 449-456.

Baxter S W, Davey J W, Johnston J S, Shelton A M, Heckel D G, Jiggins C, Blaxter M L. (2011) Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS ONE 6(4), e19315.

Ben-Ya'cov A, Zilberstaine M, Goren M, Tomer E. (2003) The Israeli avocado germplasm bank: where and why the items had been collected. En Proceedings V World Avocado Congress. Spain. October 19-24.

Benevenuto J, Ferrão, L F V, Amadeu R R, Munoz P. (2019) How can a high-quality genome assembly help plant breeders? GigaScience 8(6), giz068.

Bentley D R, Balasubramanian S, Swerdlow H P, Smith G P, Milton J, Brown C G, Hall K P, Evers D J, Barnes C L, Bignell H R, Boutell J M, Bryant J, Carter R J, Cheetham R K, Cox A J, Ellis D J, Flatbush M R, Gormley N A, Humphray S J, Irving L J, Karbelashvili M S, Kirk S M, Li H, Liu X, Maisinger K S, Murray L J, Obradovic B, Ost T, Parkinson M L, Pratt M R, *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218), 53.

Bentley N, Grauke L J, Klein P. (2019) Genotyping by sequencing (GBS) and SNP marker analysis of diverse accessions of pecan (*Carya illinoinensis*). Tree Genetics & Genomes 15(1), 8.

Berthouly-Salazar C, Mariac C, Couderc M, Pouzadoux J, Floc'h J B, Vigouroux Y. (2016) Genotyping-by-Sequencing SNP identification for crops without a reference genome: using transcriptome based mapping as an alternative strategy. Frontiers in Plant Science 7, 777; 10.3389/fpls.2016.00777.

Berumen-Varela G, Hernández-Oñate M A, Tiznado-Hernández M E. (2019) Utilization of biotechnological tools in soursop (*Annona muricata* L.). Scientia Horticulturae 245, 269-273.

Bielenberg D G, Rauh B, Fan S, Gasic K, Abbott A G, Reighard G L, Okie W R, Wells C E. (2015) Genotyping by sequencing for SNP-based linkage map construction and QTL analysis of chilling requirement and bloom date in peach [*Prunus persica* (L.) Batsch]. PLoS ONE 10(10).

Biton I, Doron-Faigenboim A, Jamwal M, Mani Y, Eshed R, Rosen A, Sherman A, Ophir R, Lavee S, Avidan B, Ben-Ari G. (2015) Development of a large set of SNP markers for assessing phylogenetic relationships between the olive cultivars composing the Israel olive germplasm collection. Molecular Breeding 35, 107.

Blazier J C, Ruhlman T A, Weng M L, Rehman S K, Sabir J S M, Jansen R K. (2016) Divergence of RNA polymerase α subunits in angiosperm plastid genomes is mediated by genomic rearrangement. Scientific Reports 6, 24595.

Boetzer M, Henkel C V, Jansen H J, Butler D, Pirovano W. (2011) Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27, 578-9.

Bonavia D, Ochoa C M, Tovar S O, Palomino R C. (2004) Archaeological evidence of

cherimoya (*Annona cherimola* Mill.) and guanabana (*Annona muricata* L.) in ancient Peru. Economic Botany 58, 509522.

Borrone J W, Schnell R J, Viola H A, Ploetz R C. (2007) Seventy microsatellite markers from *Persea americana* Miller (avocado) expressed sequences tags. Molecular Ecology Notes 7, 439-444.

Borrone J W, Brown J S, Tondo C L, Mauro-Herrera M, Kuhn D N, Violi H A, Sautter R T, Schnell R J. (2009) An EST-SSR-based linkage map for *Persea americana* Mill. (avocado). Tree Genetics & Genomes 5(4), 553-560.

Bost J B, Smith, N J H, Crane J H. (2013) History, Distribution and Uses. The Avocado: Botany, Production and Uses. (eds: Schaffer B, Wolstenholme N, Whiley A W) 10-15 (CABI, Wallingford).

Bourke P M, Van Geest G v, Voorrips R E, Jansen J, Kranenburg T, Shahin A, Visser R G F, Arens P, Smulders M J M, Maliepaard C. (2018) PolymapR–linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics 34(20), 3496-3502.

Bowden W M. (1945) A List of chromosome numbers in higher plants. I. Acanthaceae to Myrtaceae. American Journal of Botany 32, 81–92; 10.2307/2437114.

Bowden W M. (1948) Chromosome numbers in the Annonaceae. American Journal of Botany 35, 377-381.

Boza E J, Tondo C L, Ledesma N, Campbell R J, Bost J, Schnell R J, Gutiérrez O A. (2018) Genetic differentiation, races and interracial admixture in avocado (*Persea americana* Mill.), and Persea spp. evaluated using SSR markers. Genetic Resources and Crop Evolution 65(4), 1195-1215.

Broman K W, Wu H, Sen Ś, Churchill G A. (2003) R/qtl: QTL mapping in experimental crosses. Bioinformatics 19, 889-89.

Broman K W, Gatti D M, Svenson K L, Sen Ś, Churchill G A. (2019) Cleaning genotype data from Diversity Outbred mice. G3: Genes, Genomes, Genetics 9(5), 1571-1579.

Buzgo M, Chanderbali A S, Kim S, Zheng Z, Oppenheimer D G, Soltis P S, Soltis D E. (2007) Floral developmental morphology of *Persea americana* (avocado, Lauraceae): the oddities of male organ identity. International Journal of Plant Sciences 168(3), 261-284.

Cai L, Xi Z, Amorim A M, Sugumaran M, Rest J S, Liu L, Davis C C. (2019) Widespread ancient whole-genome duplications in Malpighiales coincide with Eocene global climatic upheaval. New Phytologist 221(1), 565-576.

Chabikwa T G, Barbier F F, Tanurdzic M, Beveridge C A. (2020) De novo transcriptome assembly and annotation for gene discovery in avocado, macadamia and mango. Scientific Data 7(1), 1-7.

Chaisson M J, Tesler G. (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics 13(1), 238.

Chatrou L W, Turner I M, Klitgaard B B, Maas P J, Utteridge T M. (2018) A linear sequence to facilitate curation of herbarium specimens of Annonaceae. Kew Bulletin 73(3), 39.

Calle A, Cai L, Iezzoni A, Wünsch A. (2018) High-density linkage maps constructed in sweet cherry (*Prunus avium* L.) using cross-and self-pollination populations reveal chromosomal homozygosity in inbred families and non-syntenic regions with the peach genome. Tree Genetics & Genomes 14(3), 37.

Carrasco B, González M, Gebauer M, García-González R, Maldonado J, Silva H. (2018) Construction of a highly saturated linkage map in Japanese plum (*Prunus salicina* L.) using GBS for SNP marker calling. PLoS ONE 13(12).

Catchen J M, Amores A, Hohenlohe P, Cresko W, Postlethwait J H. (2011) Stacks: Building and genotyping loci de novo from short-read sequences. G3: Genes, Genomes, Genetics 1, 171–182.

Chagné D, Gasic K, Crowhurst R N, Han Y, Bassett H C, Bowatte, D R, Lawrence T J, Rikkerink E H A, Gardiner S E, Korban S S. (2008) Development of a set of SNP markers present in expressed genes of the apple. Genomics 92, 353–358.

Chanderbali A S, Albert V A, Ashworth V E, Clegg M T, Litz R E, Soltis D E, Soltis P S. (2008) *Persea americana* (avocado): bringing ancient flowers to fruit in the genomics era. BioEssays 30(4), 386-396.

Chanderbali A S, Soltis D E, Soltis P S, Wolstenholme B N. (2013) Taxonomy and botany. The Avocado: Botany, Production, and Uses. (eds: Schaffer B, Wolstenholme B N, Whiley A W) 32-50 (CABI, Wallingford, UK).

Chaw S, Liu Y, Wu Y, Wang H, Lin C I, Wu C, Ke H, Chang L, Hsu C, Yang H, Sudianto E, Hsu M, Wu K, Wang L, Leebens-Mack J H, Tsai I J. (2019) Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. Nature Plants 5(1), 63.

Chen H, Morrell P L, de la Cruz M. (2008) Nucleotide diversity and linkage disequilibrium in wild avocado (*Persea americana* Mill.). Journal of Heredity 99, 382-389.

Chen H, Morrell P L, Ashworth V E T M, Clegg M T. (2009) Tracing the geographic origins of major avocado cultivars. Journal of Heredity 100, 56-65.

Chen F, Dong W, Zhang J, Guo X, Chen J, Zhengjia W, Lin Z, Tang H, Zhang L. (2018) The sequenced angiosperm genomes and genome databases. Frontiers in Plant Science 9, 418.

Chen F, Song Y, Li X, Chen J, Mo L, Zhang X, Lin Z, Zhang L. (2019a) Genome sequences of horticultural plants: past, present, and future. Horticulture Research 6(1), 1-23.

Chen J, Duan Y, Hu Y, Li W, Sun D, Hu H, Xie J. (2019b) Transcriptome analysis of atemoya pericarp elucidates the role of polysaccharide metabolism in fruit ripening and cracking after harvest. BMC Plant Biology 19(1), 219.

Chen J, Hao Z, Guang X, Zhao C, Wang P, Xue L, Zhu Q, Yang L, Sheng Y, Zhou Y, Xu H, Xie H, Long X, Zhang J, Wang Z, Shi M, Lu Y, Liu S, Guan L, Zhu Q, Yang L, Ge S, Chen T, Laux T, Gao Q, Peng Y, Liu N, Yang S, Shi J. (2019c) Liriodendron genome sheds light on angiosperm phylogeny and species–pair differentiation. Nature Plants 5(1),18.

Chen Y, Li Z, Zhao Y, Gao M, Wang J, Liu K, Wang X, Wu L, Jiao Y, Xu Z, He W, Zhang Q, Liang C, Hsiao Y, Zhang D, Lan S, Huang L, Xu W, Tsai W, Liu Z, Peer Y V, Wang Y. (2020) The *Litsea* genome and the evolution of the laurel family. Nature Communications 11(1), 1-14.

Cheng C Y, Krishnakumar V, Chan A P, Thibaud-Nissen F, Schobel S, Town C D. (2017) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. The Plant Journal 89(4), 789-804.

Chikhi R, Rizk G. (2013) Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms for Molecular Biology 8, 22; 10.1186/1748-7188-8-22.

Ching A, Caldwell K S, Jung M, Dolan M, Smith O S H, Tingey S, Morgante M, Rafalski A. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genetics 3, 19; 10.1186/1471-2156-3-19.

Collard B C, 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. Euphytica142(1-2), 169-196.

Collins F S, Morgan M, Patrinos A. (2003) The Human Genome Project: lessons from large-scale biology. Science 300(5617), 286-290.

Couvreur T L P, Pirie M D, Chatrou L W, Saunders R M K, Su Y C F, Richardson J E. (2011) Early evolutionary history of the flowering plant family Annonaceae: Steady diversification and boreotropical geodispersal. Journal of Biogeography 38, 664–680.

Crane J H, Douhan G, Faber B A, Arpaia M L, Bender G S, Balerdi C F, Barrientos-Priego A F. (2013) Cultivars and rootstocks. The Avocado: Botany, Production, and Uses (eds: Schaffer B, Wolstenholme B N, Whiley A W) 1-9 (CABI, Wallingford, UK).

Cruz F, Julca I, Gómez-Garrido J, Loska D, Marcet-Houben M, Cano E, Galán B, Frias L, Ribeca P, Derdak S, Gut M, Sánchez-Fernández M, García J L, Gut I G, Vargas P, Alioto T S, Gabaldón T. (2016) Genome sequence of the olive tree, *Olea europaea*. Gigascience 5(1), s13742-016.

Cui L, Wall P K, Leebens-Mack J H, Lindsay B G, Soltis D E, Doyle J J, Soltis P S, Carlson J E, Arumuganathan K, Barakat A, Albert V A, Ma H, dePamphilis C W. (2006) Widespread genome duplications throughout the history of flowering plants. Genome Research 16(6), 738-749.

Da Silva Linge C, Antanaviciute L, Abdelghafar A, Arús P, Bassi D, Rossini L, Ficklin S, Gasic K. (2018) High-density multi-population consensus genetic linkage map for peach. PLoS ONE 13(11), e0207724; https://doi.org/10.1371/journal. pone.0207724.

Daccord N, Celton J M, Linsmith G, Becker C, Choisne N, Schijlen E, van de Geest H, Bianco L, Micheletti D, Velasco R, Di Pierro E A, Gouzy J, Rees D J G, Guérif P, Muranty H, Durel C, Laurens F, Lespinasse Y, Gaillard S, Aubourg S, Quesneville H, Weigel D, van de Weg E, Troggio M, Bucher E. (2017) High-quality de novo assembly of the apple genome and methylome dynamics of early fruit development. Nature Genetics 49(7), 1099.

D'Agostino N, Taranto F, Camposeo S, Mangini G, Fanelli V, Gadaleta S, Miazzi M M, Pavan S, di Rienzo V, Sabetta W, Lombardo L, Zelasco S, Perri E, Lotti C, Ciani E, Montemurro C. (2018) GBS-derived SNP catalogue unveiled wide genetic variability and geographical relationships of Italian olive cultivars. Scientific Reports 8, 15877.

Dai B, Guo H, Huang C, Ahmed M M, Lin Z. (2017) Identification and characterization of segregation distortion loci on cotton chromosome 18. Frontiers in Plant Science 7, 2037.

Danecek P, Auton A, Abecasis G, Albers C A, Banks E, DePristo M A, Handsaker R E, Lunter G, Marth G T, Sherry S T, McVean G, Durbin R, 1000 Genomes Project Analysis Group. (2011) The variant call format and VCFtools. Bioinformatics 27, 2156–2158.

Darwin C R. (1835) Beagle diary (1831-1836). Disponible en: http://darwinonline.org.uk/ content/frameset?itemID=F1925&viewtype=text&pageseq=1 (último acceso 1 de octubre del 2019).

Davis J, Henderson D, Kobayashi M, Clegg M T, Clegg M T. (1998) Genealogical relationships among cultivated avocado as revealed through RFLP analysis. Journal of Heredity 89, 319-323.

De la Peña-Alonso E, Méndez V P, Alcaraz L, Lora J, Larrañaga N, Hormaza I. (2018) Polinizadores y polinización en frutales subtropicales: implicaciones en manejo, conservación y seguridad alimentaria. Revista Ecosistemas 27(2), 91-101.

Del Angel V D, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Pettersson O V, Amselem J, Bouri L, Bocs S, Klopp C, Gibrat J, Vlasova A, Leskosek B L, Soler L, Binzer-Panchal M, Lantz H. (2018) Ten steps to get started in Genome Assembly and Annotation. F1000Research 7.

Deokar A A, Ramsay L, Sharpe A G, Diapari M, Sindhu A, Bett K, Warkentin T D, Tar´an B. (2014) Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. BMC Genomics 15(1), 708.

Deschamps S, Llaca V, May G D. (2012) Genotyping-by-sequencing in plants. Biology 1(3),460-483.

Doležel J, Greilhuber J, Suda J. (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nature Protocols 2(9), 2233.

Doyle J A, Le Thomas A. (1997) Phylogeny and geographic history of Annonaceae. Geographie Physique et Quaternaire 51, 353-361.

Earl D A, vonHoldt B M. (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources 4, 359-361.

Edwards D, Forster J W, Cogan N O I, Batley J, Chagne D. (2007) Single Nucleotide Polymorphism Discovery. Association Mapping in Plants. (eds: Oraguzie N C, Rikkerink E H A, Gardiner S E, De Silva H N) (Springer, New York, NY).

Edwards K D, Fernandez-Pozo N, Drake-Stowe K, Humphry M, Evans A D, Bombarely A, Allen F, Hurst R, White B, Kernodle S P, Bromley J R, Sanchez-Tamburrino J P, Lewis R S, Mueller L A. (2017) A reference genome for *Nicotiana tabacum* enables map-based cloning of homeologous loci implicated in nitrogen utilization efficiency. BMC Genomics 18(1), 448.

Ekblom R, Wolf J B. (2014) A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 7(9), 1026-1042.

Ellstrand N, Lee J M. (1987) Cultivar identification of cherimoya (*Annona cherimola* Mill.) using isozyme markers. Scientia Horticulturae 32, 25-31.

Elshire R J, Glaubitz J C, Sun Q, Poland J A, Kawamoto K, Buckler E S, Mitchell S E. (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5), e19379.

Escribano P, Viruel M A, Hormaza J I. (2004) Characterization and cross-species amplification of microsatellite markers in cherimoya (*Annona cherimola* Mill., Annonaceae). Molecular Ecology Notes 4, 746-748.

Escribano P. (2007) Desarrollo de marcadores moleculares para la identificación de genotipos, estudios de diversidad y mejora de chirimoyo (*Annona cherimola* Mill.). Tesis doctoral, Universidad de Málaga.

Escribano P, Viruel M.A, Hormaza J I. (2007) Molecular analysis of genetic diversity and geographic origin within an ex situ germplasm collection of cherimoya by using SSRs. Journal of the American Society for Horticultural Science 132, 357-367.

Escribano P, Viruel M A, Hormaza J I. (2008a) Comparison of different methods to construct a core germplasm collection in woody perennial species with simple sequence repeat markers. A case study in cherimoya (*Annona cherimola*, Annonaceae), an underutilised subtropical fruit tree species. Annals of Applied Biology 153, 25-32.

Escribano P, Viruel M A, Hormaza J I. (2008b) Development of 52 new polymorphic SSR markers from cherimoya (*Annona cherimola* Mill.): transferability to related taxa and selection of a reduced set for DNA fingerprinting and diversity studies. Molecular Ecology Notes 8, 317-321.

Evanno G, Regnaut S, Goudet J. (2005) Detecting the number of clusters of individuals using the software: STRUCTURE: a simulation study. Molecular Ecology 14, 2611-2620.

Falistocco E, Ferradini N. (2020) Advances in the cytogenetics of Annonaceae, the case of *Annona cherimola* L. Genome 999, 1-8.

Fang Z, Morrell P L. (2016) Domestication: Polyploidy boots domestication. Nature Plants 2(8), 1-2.

Farré J M, Pliego F. (1987) Avocado in Spain. South African Avocado Growers Association Yearbook 10, 27-28.

Ferrão L F V, Johnson T S, Benevenuto J, Edger P P, Colquhoun T A, Munoz P R. (2020) Genome-wide association of volatiles reveals candidate loci for blueberry flavor. New Phytologist; https://doi.org/10.1111/nph.16459.

Fiedler J, Bufler G, Bangerth, F. (1998) Genetic relationships of avocado (*Persea americana* Mill.) using RAPD markers. Euphytica 101, 249-255.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M. (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260(5551), 500.

Fleischmann A, Michael T P, Rivadavia F, Sousa A, Wang W, Temsch E M, Greilhuber J, Müller K F, Heubl G. (2014) Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. Annals of Botany 114(8), 1651–1663.

Fleischmann R D, Adams M D, White O, Clayton R A, Kirkness E F, Kerlavage A R, Bult C J, Tomb J, Dougherty B A, Merrick J M, McKenney K, Sutton G, FitzHugh W, Fileds C, Gocayne J D, Scott J, Shirley R, Liu L, Glodek A, Kelley J M, Weidman J F, Phillips C A, Spriggs T, Hedblom E, Cotton M D, Utterback T R, Hanna M C, Nguyen D T, Saudek D M, Brandon R C, *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269(5223), 496-512.

Food and Agriculture Organization (FAO) (2018) Statistics Division of Food and Agriculture. Organization of the United Nations (FAOSTAT). Disponible en: http://www.fao.org/faostat/es/#data/QC (último acceso: 1 de octubre del 2019).

Frosch C, Kraus R H S, Angst C, Allgöwer R, Michaux J, Teubner J, Nowak C. (2014) The genetic legacy of multiple beaver reintroductions in Central Europe. PLoS ONE 9, e97619; 10.1371/journal.pone.0097619.

Furbank R T, Tester M. (2011) Phenomics-technologies to relieve the phenotyping bottleneck. Trends in Plant Science 16(12), 635-644.

Furnier G R, Cummings M P, Clegg M T. (1990) Evolution of the avocados as revealed by DNA restriction site variation. Journal of Heredity 81, 183-188.

Gabay G, Dahan Y, Izhaki Y, Faigenboim A, Ben-Ari G, Elkind Y, Flaishman M A. (2018) High-resolution genetic linkage map of European pear (*Pyrus communis*) and QTL fine-mapping of vegetative budbreak time. BMC Plant Biology 18(1), 175.

Galán Sauco V, Herrero M, Hormaza J I. (2014) Tropical and subtropical fruits. Horticulturae: Plants for People and Places. (eds: Dixon G R, Aldous D E) 123-157 (Springer, Dordrecht, Holland).

Galindo-Tovar M E, Ogata-Aguilar N, Arzate-Fernández A M. (2008) Some aspects of avocado (*Persea americana* Mill.) diversity and domestication in Mesoamerica. Genetic Resources and Crop Evolution 55, 441-450.

Gambino G, Perrone I, Gribaudo I. (2008) A rapid and effective method for RNA extraction from different tissues of grapevine and other woody plants. Phytochemical Analysis 19(6), 520-525.

Garrido-Cardenas J A, Mesa-Valle C, Manzano-Agugliaro F. (2018) Trends in plant research using molecular markers. Planta 247(3), 543-557.

Garrison E, Marth G. (2012) Haplotype-based variant detection from short-read sequencing. Pre-impresa: http://arxiv.org/abs/1207.3907.

Ge Y, Tan L, Wu B, Wang T, Zhang T, Chen H, Zou M, Ma F, Xu Z, Zhan R. (2019a) Transcriptome Sequencing of Different Avocado Ecotypes: de novo Transcriptome Assembly, Annotation, Identification and Validation of EST-SSR Markers. Forests 10, 411.

Ge Y, Zhang T, Wu B, Tan L, Ma F, Zou M, Chen H, Pei J, Liu Y, Chen Z, Xu Z, Wang T. (2019b) Genome-wide assessment of avocado germplasm determined from Specific Length Amplified Fragment sequencing and transcriptomes: population structure, genetic diversity, identification, and application of race-specific markers. Genes 10, 215; 10.3390/genes10030215.

Glaubitz J C, Casstevens T M, Lu F, Harriman J, Elshire R J, Sun Q, Buckler E S. (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline. PLoS ONE 9(2), e90346.

Goddard M E, Hayes B J. (2009) Mapping genes for complex traits in domestic animals and their ise in breeding programmes. Nature Reviews Genetics 10(6), 381-391.

Godden G T, Kinser T J, Soltis P S, Soltis D E. (2019) Phylotranscriptomic analyses reveal asymmetrical gene duplication dynamics and signatures of ancient polyploidy in mints. Genome Biology and Evolution 11(12), 3393-3408.

Goff S A, Ricke D, Lan T, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H, Hadley D, Hutchinson D, Martin C, Katagiri F, Lange B M, Moughamer T, Xia Y, Budworth P, Zhong J, Miquel T, Paszkowski U, Zhang S, Colbert M, Sun W, Chen L, Cooper B, Park S, Wood T C, Mao L, Quail P, *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa L. ssp. japonica*). Science 296(5565), 92-100.

Goffeau A, Barrell B G, Bussey H, Davis R W, Dujon B, Feldmann H, Galibert F, Hoheisel J D, Jacq C, Johnston M, Louis E J, Mewes H W, Murakami Y, Philippsen P, Tettelin H, Oliver S G. (1996) Life with 6000 genes. Science 274(5287), 546-567.

Goudet J. (2005) Hierfstat, a package for R to compute and test hierarchical F-statistics. Molecular Ecology Notes 5(1), 184-186.

Grattapaglia D, Sederoff R. (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137, 1121-1137.

Gross-German E, Viruel M A. (2013) Molecular characterization of avocado germplasm with a new set of SSR and EST-SSR markers: genetic diversity, population structure, and identification of race-specific markers in a group of cultivated genotypes. Tree Genetics & Genomes 9, 539-555.

Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, Lepoittevin C, Malausa T, Revardel E, Salin F, Petit R J. (2011) Current trends in microsatellite genotyping. Molecular Ecology Resources 11(4), 591-611.

Guirado E, Hermoso J M, Pérez M, Farré J. (2003) Introducción al cultivo del chirimoyo. (eds: Caja Rural de Granada, Granada).

Guo X, Tang C C, Thomas D C, Couvreur T L, Saunders R M. (2017) A mega-phylogeny of the Annonaceae: taxonomic placement of five enigmatic genera and support for a new tribe, Phoenicantheae. Scientific Reports 7(1), 7323.

Gupta Y, Pathak A K, Singh K, Mantri S S, Singh S P, Tuli R. (2015) De novo assembly and characterization of transcriptomes of early-stage fruit from two genotypes of *Annona squamosa* L. with contrast in seed number. BMC Genomics 16(1), 86.

Guzmán L F, Machida-Hirano R, Borrayo E, Cortés-Cruz M, Espíndola-Barquera M D C, Heredia García E. (2017) Genetic structure and selection of a core collection for long term conservation of avocado in Mexico. Frontiers in Plant Science 8, 243; 10.3389/fpls.2017.00243.

Hahn M W. (2018) Population structure in Molecular Population Genetics. (eds: Sinauer Associates) 81-83 (Oxford University Press. U.S.A.).

Hayward A C, Tollenaere R, Dalton-Morgan J, Batley J. (2015) Molecular marker applications in plants. Plant Genotyping: Methods and Protocols in Molecular Biology. (eds: Barley J) 13-27 (Springer Science Business Media New York).

Hedges S B, Marin J, Suleski M, Paymer M, Kumar S. (2015) Tree of life reveals clock-like speciation and diversification. Molecular Biology and Evolution 32(4), 835-845.

Helyar S J, Hemmer-Hansen J, Bekkevold D, Taylor M I, Ogden R, Limborg M T, Cariani A, Maes G E, Diopere E, Carvalho G R, Nielsen E E. (2011) Application of SNPs for population genetics of nonmodel organisms: new opportunities and challenges. Molecular Ecology Resources 1, 123-36.

Henry R, Edwards K. (2009) New tools for single nucleotide polymorphism (SNP) discovery and analysis accelerating plant biotechnology. Plant Biotechnology Journal 7(4), 311-311.

Herten K, Hestand M S, Vermeesch J R, Van Houdt J K J. (2015) GBSX: a toolkit for experimental design and demultiplexing genotyping by sequencing experiments. BMC Bioinformatics 16, 73; 10.1186/s12859-015-0514-3.

Hirao T, Matsunaga K, Hirakawa H, Shirasawa K, Isoda K, Mishima K, Tamura M, Watanabe A. (2019) Construction of genetic linkage map and identification of a novel major locus for resistance to pine wood nematode in Japanese black pine (*Pinus thunbergii*). BMC Plant Biology 19(1), 424.

Hoff K J, Lomsadze A, Stanke M, Borodovsky M. (2018) BRAKER2: incorporating protein homology information into gene prediction with GeneMark-EP and AUGUSTUS. Plant and Animal Genomes XXVI, January 14th 2018.

Hofshi R. Avocado database. Disponible en: http://www.avocadosource.com/Avocado Varieties/QueryDB.asp (último acceso: 13 de septiembre del 2019).

Hormaza J I. (2014) The pawpaw, a forgotten North American fruit tree. Arnoldia 72(1), 13-23.

Hormaza J I, Carmona E, González-Padilla I M, Larranaga N, Lora J, Talavera A, López-Encina C. (2020) Annonaceae: *Annona* spp. (Atemoya, Cherimoya, Soursop and Sugar Apple) and *Asimina* spp. (Pawpaw). Biotechnology of Fruit and Nut Crops. (eds: Litz R E, Pliego-Alfaro F, Hormaza J I) (CABI, Wallingford, UK).

Hosmani P S, Flores-Gonzalez M, van de Geest H, Maumus F, Bakker L V, Schijlen E, van Haarst J, Cordewener J, Sanchez-Perez G, Peters S, Fei Z, Giovannoni J J, Mueller L A, Saha S. (2019) An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. bioRxiv. 767764.

Hu L, Xu Z, Wang M, Fan R, Yuan D, Wu B, Wu H, Qin X, Yan L, Tan L, Sim S, Li W, Saski C A, Daniell H, Wendel J F, Lindsey K, Zhang X, Hao C, Jin S. (2019) The chromosome-scale reference genome of black pepper provides insight into piperine biosynthesis. Nature Communications 10(1), 1-11.

Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, Guan J, Fan D, Weng Q, Huang T, Dong G, Sang T, Han B. (2009) High-throughput genotyping by whole-genome resequencing. Genome Research 19(6), 1068-1076.

Hunt A G, Li Q Q. (eds) (2015), Polyadenylation in Plants: Methods and Protocols. Methods in Molecular Biology, 1255 (Springer Science+Business Media New York).

Hussain W, Baenziger P S, Belamkar V, Guttieri M J, Venegas J P, Easterly A, Sallam A, Poland J. (2017) Genotyping-by-sequencing derived high-density linkage map and its application to QTL mapping of flag leaf traits in bread wheat. Scientific Reports 7(1), 1-15.

Ibarra-Laclette E, Méndez-Bravo A, Pérez-Torres C A, Albert V A, Mockaitis K, Kilaru A, López-Gómez R, Cervantes-Luevano J I, Herrera-Estrella L. (2015) Deep sequencing of the Mexican avocado transcriptome, an ancient angiosperm with a high content of fatty acids. BMC Genomics 16, 599; 10.1186/s12864-015-1775-y.

Illumina. (2010) Genomic sequencing. Disponible en: https://www.illumina.com/documents/products/datasheets/datasheet_genomic_sequence.pdf (último acceso: 1 de octubre del 2019).

Jaganathan D, Bohra A, Thudi M, Varshney R K. (2020) Fine mapping and gene cloning in the post-NGS era: advances and prospects. Theoretical and Applied Genetics.

Jayakumar V, Sakakibara Y. (2019) Comprehensive evaluation of non-hybrid genome assembly tools for third-generation PacBio long-read sequence data. Briefings in Bioinformatics 20(3), 866-876.

Jenczewski E, Gherardi M, Bonnin I, Prosperi J M, Olivieri I, Huguet T. (1997) Insight on segregation distortions in two intraspecific crosses between annual species of *Medicago* (Leguminosae). Theoretical and Applied Genetics 94, 682-691.

Jiang J. (2019) Fluorescence in situ hybridization in plants: recent developments and future applications. Chromosome Research 27(3), 153-165.

Jombart T. (2008) Adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics 24, 1403–1405.

Jones P, Binns D, Chang H, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn A F, Sangrador-Vegas A, Scheremetjew M, Yong S, Lopez R ,Hunter S. (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.

Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P. (2019) Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. Trends in Plant Science 24.

Kaiser C, Wolstenholme B N. (1994) Aspects of delayed harvest of 'Hass' avocado (*Persea americana* Mill.) fruit in a cool subtropical climate. II. Fruit size, yield, phenology and whole-tree starch cycling. Journal of Horticultural Science 69, 447-457.

Kamvar Z N, Tabina J F, Grünwald N J. (2014) Poppr: an R package for genetic analysis of populations with clonal, partially clonal , and/or sexual reproduction. PeerJ 2, e281; 10.7717/peerj.281.

Kamvar Z N, Brooks J C, Grünwald N J. (2015) Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. Frontiers in Genetics 6, 208; 10.3389/fgene.2015.00208.

Kersey P J. (2019) Plant genome sequences: past, present, future. Current Opinion in Plant Biology 48, 1-8.

Kim D, Langmead B, Salzberg S L. (2015) HISAT: a fast spliced aligner with low memory requirements. Nature Methods.

Kimura M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature 267(5608), 275-276.

King I P, Koebner R M D, Reader S M, Miller T E. (1991) Induction of a mutation in the male fertility gene of the preferentially transmitted *Aegilops sharonensis* chromosome 4S 1 and its application for hybrid wheat production. Euphytica 54(1), 33-39.

Kopp L E. (1966) A taxonomic revision of the genus Persea in the Western Hemisphere (Persea: Lauraceae). Revisión taxonómica del género Persea en el hemisferio occidental (Persea: Lauraceae). Memoirs of the New York Botanical Garden 14(1), 1-120.

Koren S, Walenz B P, Berlin K, Miller J R, Bergman N H, Phillippy A M. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research 27(5), 722-736.

Korf I. (2004) Gene finding in novel genomes. BMC Bioinformatics 5, 59.

Kosambi D D. (1944) The estimation of map distances from recombination values. Annals of Eugenics 12, 172–175.

Kuhn D N, Bally I S E, Dillon N L, Innes D, Groh A M, Rahaman J, Ophir R, Cohen Y, Sherman A. (2017) Genetic Map of Mango: A Tool for Mango Breeding. Frontiers in Plant Science 8, 577; 10.3389/fpls.2017.00577.

Kuhn D N, Livingstone III D S, Richards J H, Manosalva P, Van den Berg N, Chambers A H. (2019) Application of genomic tools to avocado (*Persea americana*) breeding: SNP discovery for genotyping and germplasm characterization. Scientia Horticulturae 246, 1-11.

Kujur A, Bajaj D, Upadhyaya H D, Das S, Ranjan R, Shree T, Saxena M S, Badoni S, Kumar V, Tripathi S, Gowda C L L, Sharma S, Singh S, Tyagi A K, Parida S K. (2015) Employing genome-wide SNP discovery and genotyping strategy to extrapolate the natural allelic diversity and domestication patterns in chickpea. Frontiers in Plant Science 6, 162; 10.3389/fpls.2015.00162.

Kumar L S S, Ranadive K. (1941) A cytological study of the genus *Annona*. Journal of University of Bombay 10B, 1-8; 10.3732/ajb.1400312.

Landis J B, Soltis D E, Li Z, Marx H E, Barker M S, Tank D C, Soltis P S. (2018) Impact of whole-genome duplication events on diversification rates in angiosperms. American Journal of Botany 105(3), 348-363.

Lantican D V, Strickler S R, Canama A O, Gardoce R R, Mueller L A, Galvez H F. (2019) De novo genome sequence assembly of dwarf coconut *(Cocos nucifera* L.'Catigan Green Dwarf') provides insights into genomic variation between coconut types and related palm species. G3: Genes, Genomes, Genetics 9(8), 2377-2393.

Larrañaga N. (2016) Origen, dispersion y diversidad del chirimoyo (*Annona cherimola* Mill.) en el continente americano. Tesis doctoral, Universidad Politécnica de Madrid.

Larranaga N, Albertazzi F J, Fontecha G, Palmieri M, Rainer H, van Zonneveld M, Hormaza J I. (2017) A Mesoamerican origin of cherimoya (*Annona cherimola* Mill.): Implications for conservation of plant genetic resources. Molecular Ecology 26(16), 4116-4130.

Lavi U, Hillel J, Vainstein A. (1991) Application of DNA fingerprints for identification and genetic analysis of avocado. Journal of the American Society for Horticultural Science 116, 1078-1081.

Lavi U, Cregan P B, Hillel J. (1994) Application of DNA markers for identification and breeding of fruit trees. Plant Breeding Reviews 12, 195-226.

Le Nguyen K, Grondin A, Courtois B, Gantet P. (2018) Next-generation sequencing accelerates crop gene discovery. Trends in Plant Science.

Leggett R M, MacLean D. (2014) Reference-free SNP detection: dealing with the data deluge. BMC Genomics 15, S10; 10.1186/1471-2164-15-S4-S10.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078–2079.

Li H, Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transformation. Bioinformatics 26, 589-595.

Li H. (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics 32(14), 2103-2110.

Li H, Li W, Zhang T, Zhong J, Liu J, Yuan C, Liu K. (2019a) Comparative transcriptomic analysis of split and non-split atemoya (*Annona cherimola* Mill.× *Annona squamosa* L.) fruit to identify potential genes involved in the fruit splitting process. Scientia Horticulturae 248, 216-224.

Li S, Guo Y, Li J, Zhang D, Wang B, Li N, Deng C, Gao W. (2019b) The landscape of transposable elements and satellite DNAs in the genome of a dioecious plant spinach (*Spinacia oleracea* L.). Mobile DNA 10(1), 1-15.

Liaw C C, Wu T Y, Chang F R, Wu Y C. (2011) Historic perspectives on annonaceous acetogenins from the chemical bench to preclinical trials. Planta Medica 76, 1390-1404.

Liu B, Shi Y, Yuan J, Hu X, Zhang H, Li N, Li Z, Chen Y, Mu D, Fan W. (2013) Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. Pre-impresa: https://arxiv.org/abs/1308.2012

Liu W, Xiao Z, Bao X, Yang X, Fang J, Xiang X. (2015) Identifying litchi (*Litchi chinensis* Sonn.) cultivars and their genetic relationships using single nucleotide polymorphism (SNP) markers. PLoS ONE 10, e0135390; 10.1371/journal.pone.0135390.

Liu K, Feng S, Pan Y, Zhong J, Chen Y, Yuan C, Li H. (2016a) Transcriptome Analysis and Identification of Genes Associated with Floral Transition and Flower Development in Sugar Apple (*Annona squamosa* L.). Frontiers Plant Science 7, 1695; 10.3389/fpls.2016.01695.

Liu S, Li Y, Qin Z, Geng X, Bao L, Kaltenboeck L, Kucuktas H, Dunham R, Liu Z. (2016b) High-density interspecific genetic linkage mapping provides insights into genomic incompatibility between channel catfish and blue catfish. Animal genetics 47(1), 81-90.

Liu K, Li H, Li W, Zhong J, Chen Y, Shen C, Yuan C. (2017) Comparative transcriptomic analyses of normal and malformed flowers in sugar apple (*Annona squamosa* L.) to identify the differential expressed genes between normal and malformed flowers. BMC Plant Biology 17(1), 170.

Lomsadze A, Burns P, Borodovsky M. (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Research 42(15), e119; 10.1093/nar/gku557.

Lora J, Hormaza J I, Herrero M. (2010) The progamic phase of an early-divergent angiosperm, *Annona cherimola* (Annonaceae). Annals of Botany 105(2), 221-31.

Lora J, Herrero M, Hormaza J I. (2011a) Stigmatic receptivity in a dichogamous early-divergent angiosperm species, *Annona cherimola* (Annonaceae): influence of temperature and humidity. American Journal of Botany 98(2), 265-74.

Lora J, Hormaza J I, Herrero M, Gasser C S. (2011b) Seedless fruits and the disruption of a conserved genetic pathway in angiosperm ovule development. Proceedings of the National Academy of Sciences 108(13), 54615465.

Lora J, Larranaga N, Hormaza I H. (2018) Genetics and Breeding of Fruit Crops in the Annonaceae Family: *Annona* spp. and *Asimina* spp. Advances in Plant Breeding Strategies: Fruits. (eds: Al-Khayri J M, Johnson D V, Jain S M) (Springer International Publishing AG, part of Springer Nature).

Lyttle T W. (1991) Segregation distorters. Annual Review of Genetics 25, 511-557.

Losada J M, Hormaza, J I, Lora J. (2017) Pollen–pistil interaction in pawpaw (*Asimina triloba*), the northernmost species of the mainly tropical family Annonaceae. American Journal of Botany 104(12), 1891-1903.

Lu F, Lipka A E, Glaubitz J, Elshire R, Cherney J H, Casler M D, Buckler E S, Costich D E. (2013) Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. PLoS Genetics 9, e1003215; 10.1371/journal.pgen.1003215 (2013).

Lu F, Romay M C, Glaubitz J C, Bradbury P J, Elshire R J, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez A G, Mikel M A, Soifer I, Barad O, Buckler E S. (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nature Communications 6(1), 1-8.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung D W, Yiu S, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T, Wang J. (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18; 10.1186/2047-217x-1-18.

Mackay T F, Stone E A, Ayroles J F. (2009) The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics 10(8), 565-577.

Malmberg M M, Spangenberg G C, Daetwyler H D, Cogan N O I. (2019) Assessment of low-coverage nanopore long read sequencing for SNP genotyping in doubled haploid canola (*Brassica napus* L.). Scientific Reports 9(1), 8688.

MAPA (2019) Superficies y producciones anuales de cultivos. Disponible en: https://www.mapa.gob.es/es/estadistica/temas/estadisticas-agrarias/agricultura/superficies-producciones-anuales-cultivos/ (último acceso: 22 de abril del 2020).

Marçais G, Kingsford C. (2011) A fast, lock-free approach for effi- cient parallel counting of occurrences of k-mers. Bioinformatics 27(6), 764–70.

Martín C. (2013) Desarrollo de un mapa genético variación de ploidía y anomalías meióticas en el género *Annona*. Tesis doctoral, Universidad de Málaga.

Martín C, Herrero M, Hormaza J I. (2011) Molecular characterization of apricot germplasm from an old stone collection. PLoS ONE 6, e23979; 10.1371/journal.pone.0023979.

Martín C, Viruel M A, Lora J, Hormaza J I. (2019) Polyploidy in fruit tree crops of the genus *Annona* (Annonaceae). Frontiers in Plant Science 10.

Massoni J, Couvreur T L, Sauquet H. (2015) Five major shifts of diversification through the long evolutionary history of Magnoliidae (angiosperms). BMC Evolutionary Biology 15 (1), 49.

Melo A T O, Bartaula R, Hale L. (2016) GBS-SNP-CROP: a reference-optional pipeline for SNP discovery and plant germplasm characterization using variable length, paired-end genotyping-by-sequencing data. BMC Bioinformatics 17, 29; 10.1186/s12859-016-0879-y.

Metzker M L. (2010) Sequencing technologies-the next generation. Nature Reviews Genetics 11(1), 31.

Mhameed S, Sharon D, Hillel J, Lahav E, Kaufman D, Lavi U. (1996) Level of heterozygosity and mode of inheritance of variable number of tandem repeat loci in avocado. Journal of American Society for Horticultural Science 121, 778-782.

Micheletti D, Dettori M T, Micali S, Aramini V, Pacheco I, Da Silva Linge C, Foschi S, Banchi E, Barreneche T, Quilot-Turion B, Lambert P, Pascal T, Iglesias I, Carbó J, Wang L, Ma R, Li X, Gao Z, Nazzicari N, Troggio M, Bassi D, Rossini L, Verde I, Laurens F, Arús P, Aranzana M J. (2015) Whole-Genome Analysis of diversity and SNP-major gene association in peach germplasm. Plant Genome 5, 92-102.

Minervini M, Scharr H, Tsaftaris S A. (2015) Image analysis: the new bottleneck in plant phenotyping [applications corner]. IEEE Signal Processing Magazine 32(4), 126-131.

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw J H, Senin P, Wang W, Ly B V, Lewis K L T, Salzberg S L, Feng L, Jones M R, Skelton R L, Murray J E, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull R E, Michael T P, Wall K, Rice D W, Albert H, Wang M, Zhu Y J, *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452(7190), 991.

Mohan M, Nair S, Bhagwat A, Krishna T G, Yano M, Bhatia C R, Sasaki T. (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. Molecular Breeding 3(2), 87-103.

Morawetz W. (1986) Remarks on karyological differentiation patterns in tropical woody plants.

Plant Systematics and Evolution 152, 49–100; 10.1007/BF009 85351.

Morin P A, Luikart G, Wayne R K. (2004) SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution 19(4), 208-216.

Morton J F. (1987) Fruits of Warm Climates. 91–102 (Creative Resources Systems, Winterville, NC, USA).

Nature Plants. (2018) From genes to networks. Nature Plants 4, 55; 10.1038/s41477-018-0109-x.

Ouellette L A, Reid R W, Blanchard S G, Brouwer C R. (2018) LinkageMapView- rendering high-resolution linkage and QTL maps. Bioinformatics 34(2), 306-307.

Paradis E. (2010) Pegas: an R package for population genetics with an integrated–modular approach. Bioinformatics 26, 419-420.

Pascual L, Perfectti F, Gutierrez M, Vargas A. (1993) Characterizing isozymes of Spanish cherimoya cultivars. HortScience 28, 845-847.

Pellicer J, Hidalgo O, Dodsworth S, Leitch I J. (2018) Genome size diversity and its impact on the evolution of land plants. Genes 9(2), 88.

Perfectti F, Pascual L. (1996) Segregation distortion of isozyme loci in cherimoya (*Annona cherimola* Mill). Theoretical and Applied Genetics 93(3), 440-446.

Perfectti F, Pascual L. (1998a) Characterization of cherimoya germplasm by isozyme markers. Fruit Varieties Journal 52, 53-62.

Perfectti F, Pascual L. (1998b) Genetic linkage of isozyme loci in *Annona cherimola*. Hereditas 128, 87-90.

Perfectti F, Pascual L. (2005a) Genetic diversity in a worldwide collection of cherimoya cultivars. Genetic Resources and Crop Evolution 52, 959-966.

Perfectti F, Pascual L. (2005b) Geographic variation for isozymes in cherimoya (*Annona cherimola* Mill.). Genetic Resources and Crop Evolution 51, 837-843.

Pertea M, Pertea G M, Antonescu C M, Chang T, Mendell J T, Salzberg S L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology 33, 290–295.

Peterson B K, Weber J N, Kay E H, Fisher H S, Hoekstra H E. (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7(5).

Pfeifer B, Wittelsbürger U, Ramos-Onsins S E, Lercher M J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. Molecular Biology and Evolution 31, 1929-36; 10.1093/molbev/msu136.

Pillen K, Sleinrücken G, Herrmann R G, Jung C. (1993) An extended linkage map of sugar beet (*Beta vulgaris* L.) including nine putative lethal genes and the restorer gene X. Plant Breeding 111(4), 265-272.

Polanco C, Sáenz de Miera L E, González A I, García P, Fratini R, Vaquero F, Vences F J, Pérez de la Vega M. (2019) Construction of a high-density interspecific (*Lens culinaris* x *L.odemensis*) genetic map based on functional markers for mapping morphological and agronomical traits, and QTLs affecting resistance to *Ascochyta* in lentil. PLoS ONE 14(3).

Poland J A, Rife T W. (2012) Genotyping-by-sequencing for plant breeding and genetics. The Plant Genome 5(3), 92-102.

Pootakham W, Jomchai N, Ruang-areerate P, Shearman J R, Sonthirod C, Sangsrakru D, Tragoonrung S, 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, 288-295.

Popenoe W. (1920) Manual of Tropical and Subtropical Fruits. 524 (Macmillan, London).

Popenoe W. (1921) The native home of the cherimoya. Journal of Heredity 12, 331-336.

Popenoe W. (1963) Early history of the avocado. 19–24 (California Avocado Society Yearbook. 47).

Popenoe H, King S R, Leon J, Kalinowski L S, Vietmeyer N D, Dafforn M. (1989) Lost crops of the Incas. Little-known plant of the Andes with promise for worldwide cultivation. National Academy Press, Washington, D.C.

Prevosti A, Ocaña J, Alonso G. (1975) Distance between populations of *Drosophila subobscura* based on chromosome arrangement frequencies. Theoretical and Applied Genetics 45, 231-241.

Primrose S B, Twyman R M. (2006) Principles of Gene Manipulation and Genomics.7th edition. (Blackwell Publishing).

Pritchard J K, Stephens M, Donnelly P. (2000) Inference of population structure using multilocus genotype data. Genetics 155, 945-959.

Pritchard J K, Wen X, Falush D. (2010) Documentation for structure software: version 2.3. Disponible en: http://burfordreiskind.com/wp-content/uploads/Structure_Manual_doc.pdf (último acceso: 13 de septiembre del 2019).

R core Team (2018) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna. Disponible en: https://www.R-project.org (último acceso: 13 de septiembre del 2019).

Rahman M, Yamada M, Yoshida M. (1997) Relationship of *Annona* species as revealed by PCR-RFLP analysis. Breeding Science 47, 335-339.

Rahman M, Shimada T, Yamamoto T, Yonemoto J, Yoshida M. (1998) Genetical diversity of cherimoya cultivars revealed by amplified fragment length polymorphism (AFLP) analysis.

Breeding Science 48, 5-10.

Rainer H. (2007) Monographic studies in the genus *Annona* L. (Annonaceae): inclusion of the genus *Rollinia* A. St.-Hil. Annalen des Naturhistorischen Museums in Wien. Serie B, Botanik und Zoologie. Vienna 108,191-205.

Rambaut A. (2009) FigTree version 1.4.4. Disponible en: http://tree.bio.ed.ac.uk/software /figtree/ (último acceso: 13 de septiembre del 2019).

Rasheed A, Hao Y, Xia X, Khan A, Xu Y, Varshney R K, He Z. (2017) Crop breeding chips and genotyping platforms: progress, challenges, and perspectives. Molecular Plant 10, 1047-1064.

Rasheed A, Xia X. (2019) From markers to genome-based breeding in wheat. Theoretical and Applied Genetics 132(3), 767-784.

Ren Y, Zhao H, Kou Q, Jiang J, Guo S, Zhang H, Hou W, Zou X, Sun H, Gong G, Levi A, Xu Y. (2012) A High Resolution Genetic Map Anchoring Scaffolds of the Sequenced Watermelon Genome. PLoS ONE 7(1), e29453;10.1371/journal.pone.0029453.

Rendón-Anaya M, Ibarra-Laclette E, Méndez-Bravo A, Lan T, Zheng C, Carretero-Paulet L, Perez-Torres C A, Chacón-López A, Hernandez-Guzmán G, Chang T, Farr K M, Barbazuk W B, Chamala S, Mutwil M, Shivhare D, Alvarez-Ponce D, Mitter N, Hayward A, Fletcher S, Rozas J, Sánchez Gracia A, Kuhn D, Barrientos-Priego A F, Salojärvi J, Librado P, Sankoff D, Herrera-Estrella A, Albert V A, Herrera-Estrella L. (2019) The avocado genome informs deep angiosperm phylogeny, highlights introgressive hybridization, and reveals pathogen-influenced gene space adaptation. Proceedings of the National Academy of Sciences 116 (34) 17081-17089; 10.1073/pnas.1822129116.

Richardson J E, Chatrou L W, Mols J B, Erkens R H J, Pirie M D. (2004) Historical biogeography of two cosmopolitan families of flowering plants: Annonaceae and Rhamnaceae. Philosophical Transactions of the Royal Society of London Series B-Biological Sciences 359, 1495-1508.

Robinson J T, Thorvaldsdóttir H, Wenger A M, Zehir A, Mesirov J P. (2017) Variant review with the integrative genomics viewer. Cancer Research 77(21), e31-e34.

Romero-López N, Luna-Martínez F, Gallegos-Brito C, Escobar-Turriza P, Aguilar-Espinosa M, Valdez-Ojeda R, Simpson J, Rivera-Madrid R. (2019) An integrated genetic linkage map of *Bixa orellana* L. Tree Genetics & Genomes 15(4), 65.

Ronning C, Schnell R, Gazit S. (1995) Using randomly amplified polymorphic DNA (RAPD) markers to identify *Annona* cultivars. Journal of the American Society for Horticultural Science 120, 726-729.

Rosell P, Galán V, Hernández P M. (1997) Cultivo del chirimoyo en Canarias. Cuadernos de divulgación. (eds: Gobierno de Canarias. Consejería de Agricultura, Pesca y Alimentación).

Rubinstein M, Eshed R, Rozen A, Zviran T, Kuhn D N, Irihimovitch V, Sherman A, Ophir R. (2019) Genetic diversity of avocado (*Persea americana* Mill.) germplasm using pooled sequencing. BMC Genomics 20 (1), 379, 10.1186/s12864-019-5672-7.

Sánchez-Pérez R, Pavan S, Mazzeo R, Moldovan C, Cigliano R A, Del Cueto J, Ricciardi F, Lotti C, Ricciardi L, Dicenta F, López-Marqués R L, Lindberg Møller B. (2019) Mutation of a bHLH transcription factor allowed almond domestication. Science 364(6445), 1095-1098.

Sauquet H, Doyle J A, Scharaschkin T, Borsch T, Hilu K W. (2003) Phylogenetic analysis of Magnoliales and Myristicaceae based on multiple data sets: implications for character evolution. Botanical Journal of the Linnean Society 142, 125-186.

Schaffer B, Wolstenholme B N, Whiley A W. (2013) Introduction. The Avocado: Botany, Production, and Uses. (eds: Schaffer B, Wolstenholme B N, Whiley A W) 1-9 (CABI, Wallingford, UK).

Scharaschkin T, Doyle J A. (2005) Phylogeny and historical biogeography of *Anaxagorea* (Annonaceae) using morphology and non-coding chloroplast DNA sequence data. Systematic Botany 30, 712-735.

Scheben A, Batley J, Edwards D. (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnology Journal 15(2), 149-161.

Scheben A, Batley J, Edwards D. (2018) Revolution in Genotyping Platforms for Crop Improvement. Plant Genetics and Molecular Biology. Advances in Biochemical Engineering/Biotechnology. (eds: Varshney R, Pandey M, Chitikineni A) 164 (Springer, Cham.).

Scheldeman X. (2002) Distribution and potential of cherimoya (*Annona cherimola* Mill.) and highland papayas (*Vasconcellea* spp.) in Ecuador. Tesis doctoral, Universidad de Ghent.

Schneider K. (2005) Mapping Populations and Principles of Genetic Mapping. The Handbook of Plant Genome Mapping. Genetic and Physical Mapping. (eds: Khalid Meksem, Günter Kahl)(WILEY-VCH).

Schnell R J, Brown J S, Olano C T, Power E J, Krol C A, Kuhn D N, Motamayor J C. (2003) Evaluation of avocado germplasm using microsatellite markers. Journal of the American Society for Horticultural Sciences 128, 881-889.

Schreiber M, Himmelbach A, Börner A, Mascher M. (2019) Genetic diversity and relationship between domesticated rye and its wild relatives as revealed through genotyping-by-sequencing. Evolutionary Applications 12(1), 66-77.

Scora R W, Wolstenholme B N, Lavi U. (2002) Taxonomy and botany. The Avocado: Botany, Production and Uses. (eds: Whiley A W, Schaffer B, Wolstenholme B N)15-37(CAB International, Wallingford, UK).

Shakoor N, Lee S, Mockler T C. (2017) High throughput phenotyping to accelerate crop breeding and monitoring of diseases in the field. Current Opinion in Plant Biology 38, 184-192.

Sharon D, Cregan P B, Mhameed S, Kusharska M, Hillel J, Lahav E, Lavi U. (1997) An integrated genetic linkage map of avocado. Theoretical and Applied Genetics 95(5-6), 911-921.

Shearman J R, Sangsrakru D, Jomchai N, Ruang-areerate P, Sonthirod C, Naktang C, Theerawattanasuk K, Tragoonrung S, Tangphatsornruang S. (2015) SNP identification from RNA sequencing and linkage map construction of rubber tree for anchoring the draft genome. PLoS ONE 10, e0121961; 10.1371/journal.pone.0121961.

Shendure J, Balasubramanian S, Church G M, Gilbert W, Rogers J, Schloss J A, Waterston R H. (2017) DNA sequencing at 40: past, present and future. Nature 550 (7676), 345.

Simão F A, Waterhouse R M, Ioannidis P, Kriventseva E V, Zdobnov E M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19), 3210-3212.

Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, Isobe S. (2017) The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. DNA Research 24(5), 499-508.

Smit A, Hubley R, Green P. (2015a) RepeatMasker Open-4.0. Institute for Systems Biology.

Smit A F A, Hubley R, Green P. (2015b) RepeatModeler Open-1.0. 2008–2015. Seattle, USA: Institute for Systems Biology. Disponible en: www. repeatmasker.org (último acceso el 22 de marzo del 2020).

Smith L M, Sanders J Z, Kaiser R J, Hughes P, Dodd C, Connell C R, Heiner C, Kent S B H, Hood L E. (1986) Fluorescence detection in automated DNA sequence analysis. Nature 321, 674–679.

Söderquist P, Elmberg J, Gunnarsson G, Thulin C G, Champagnon J, Guillemain M, Kreisinger J, Prins H H T, Crooijmans R P M A, Kraus R H S. (2017) Admixture between released and wild game birds: a changing genetic landscape in European mallards (*Anas platyrhynchos*). European Journal of Wildlife Research 63, 98; 10.1007/s10344-017-1156-8.

Soltis D E, Soltis P S, Endress P K, Chase M W. (2005) Angiosperm phylogeny and evolution (Sinauer Associates, Sunderland).

Soltis P S, Marchant D B, Van de Peer Y, Soltis D E. (2015) Polyploidy and genome evolution in plants. Current Opinion in Genetics & Development 35, 119-125.

Soltis D E, Soltis P S. (2019) Nuclear genomes of two magnoliids. Nature Plants 5(1), 6.

Sonah H, Bastien M, Iquira E, Tardivel A, Légaré G, Boyle B, Normandeau E, Laroche J, Larose S, Jean M, Belzile F. (2013) An improved genotyping by sequencing (GBS) approach offering increased versatility and efficiency of SNP discovery and genotyping. PLoS ONE 8, e54603; 10.1371/journal.pone.0054603.

Soorni A, Haak D, Zaitlin D, Bombarely A. (2017a) Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. BMC Genomics 18(1), 49.

Soorni A, Fatahi R, Salami S A, Haak D C, Bombarely A. (2017b) Assessment of genetic diversity and population structure in Iranian cannabis germplasm. Scientific Reports 7, 15668; 10.1038/s41598-017-15816-5.

Spindel J, Wright M, Chen C, Cobb J, Gage J, Harrington S, Lorieux M, Ahmadi N, McCouch S. (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations. Theoretical and Applied Genetics 126(11), 2699-2716.

Stanke M, Tzvetkova A, Morgenstern B. (2006) AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology 7, S11.

Staub J E, Serquen F C, Gupta M. (1996) Genetic markers, map construction, and their application in plant breeding. HortScience 31, 729-741.

Sterck L, Rombauts S, Vandepoele K, Rouze P, Van de Peer Y. (2007) How many genes are there in plants (… and why are they there)?. Current Opinion in Plant Biology 10(2), 199-203.

Stevens P F. (2001 en adelante) Angiosperm Phylogeny Website. Version 12, July 2012. Disponible en: http://www.mobot.org/MOBOT/research/APweb/ (último acceso: el 14 de enero del 2020).

Stevens K A, Woeste K, Chakraborty S, Crepeau M W, Leslie C A, Martínez-García P J, Puiu D, Romero-Severson J, Coggeshall M, Dandekar A M, Kluepfel D, Neale D B, Salzberg S L, Langley C H. (2018) Genomic variation among and within six Juglans species. G3: Genes, Genomes, Genetics 8 (7), 2153-2165.

Strijk J S, Hinsinger D D, Zhang F, Cao K. (2019) *Trochodendron aralioides*, the first chromosome-level draft genome in Trochodendrales and a valuable resource for basal eudicot research. GigaScience 8(11), giz136.

Studer B, Kölliker R. (2013) SNP Genotyping Technologies. Diagnostics in Plant Breeding (eds: Lübberstedt T, Varshney R K)(Springer Science+Business Media Dordrecht).

Suguiyama V F, Vasconcelos L A B, Rossi M M, Biondo C, De Setta N. (2019) The population genetic structure approach adds new insights into the evolution of plant LTR retrotransposon lineages. PLoS ONE 14(5).

Tanaka R, Okada H. (1972) Karyological studies in four species of Annonaceae, a primitive angiosperm. Journal of Science of Hiroshima University 14, 85–105.

Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable J C, Schnable P S ,Lyons E, Lu J. (2015) ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biology 16(1), 3.

Tangphatsornruang S. (2015) Construction of high-density integrated genetic linkage map of rubber tree (*Hevea brasiliensis*) using genotyping-by-sequencing (GBS). Genomics 6, 367; 10.3389/fpls.2015.00367.

Tanksley S D, McCouch S R. (1997) Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277(5329), 1063-1066.

Taranto F, D´Agostino N, Greco B, Cardi T, Tripoli P. (2016) Genome-wide SNP discovery and population structure analysis in pepper (*Capsicum annum*) using genotyping by sequencing. BMC Genomics 17, 943; 10.1186/s12864-016-3297-7.

Taylor J, Butler D R. (2017) Package ASMap: Efficient genetic linkage map construction and diagnosis. Journal of Statistical Software 79(6); doi.org/10.18637/jss.v079.i06.

Teh B T, Lim K, Yong C H, Ng C C Y, Rao S R, Rajasegaran V, Lim W K, Ong C K, Chan K, Cheng V K Y, Soh P S, Swarup S, Rozen S G, Nagarajan N, Tan P. (2017) The draft genome of tropical fruit durian (*Durio zibethinus*). Nature Genetics 49(11), 1633.

Thakur D R, Singh, R N. (1965) Meiosis in *Annona*. Indian Journal of Genetic Plant Breeding 25, 367–371.

Thakur D, Singh R. (1969) Karyomorphological studies in some *Annona* species. Indian Journal of Genetic Plant Breeding 29, 285–290.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796-815.

The Arabidopsis Information Resource (TAIR) (2016) Disponible en: www.arabidopsis.org (ultimo acceso: el 31 de marzo del 2020).

Udall J A, Wendel J F. (2006) Polyploidy and crop improvement. Crop Science 46(Supplement_1), S-3.

U.S. National Plant Germplasm System. Disponible en: https://npgsweb.ars-grin.gov/gringlobal/search.aspx? (último acceso: 13 de septiembre del 2019).

UPOV (2011) Disponible en: https://www.upov.int/about/es/faq.html (último acceso: 1 de octubre del 2019).

Van de Peer Y, Mizrachi E, Marchal K. (2017) The evolutionary significance of polyploidy. Nature Reviews Genetics 18(7), 411.

van Zonneveld M, Scheldeman X, Escribano P, Viruel M A, Van Damme P, Garcia W, Tapia C, Romero J, Sigueñas M, Hormaza J I. (2012) Mapping genetic diversity of cherimoya (*Annona cherimola* Mill.): Application of spatial analysis for conservation and use of plant genetic resources. PLoS ONE 7, e29845.

van Zonneveld M, Larranaga N, Blonder B, Coradin L, Hormaza J I, Hunter D. (2018) Human diets drive range expansion of megafauna-dispersed fruit species. Proceedings of the National Academy of Sciences 115(13), 3326-3331.

Variety Database of the University of California at Riverside. Disponible en: http://ucavo.ucr.edu/ (último acceso el 13 de septiembre del 2019).

Vaser R, Sović I, Nagarajan N, Šikić M. (2017) Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research 27(5), 737-746.

Varshney R K, Singh V K, Hickey J M, Xun X, Marshall D F, Wang J, Edwards D, Ribaut J. (2016) Analytical and decision support tools for genomics-assisted breeding. Trends in Plant Science 21(4) 354-363.

Varshney R K, Shi C, Thudi M, Mariac C, Wallace J, Qi P, Zhang H, Zhao Y, Wang X, Rathore A, Srivastava R K, Chitikineni A, Fan G, Bajaj P, Punnuri S, Gupta S K, Wang H, Jiang Y, Couderc M, Katta M A V S K, Paudel D R, Mungra K D, Chen W, Harris-Shultz K R, Garg V, Desai N, Doddamani D, Kane N A, Conner J A, Ghatak A, *et al.* (2017) Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. Nature Biotechnology 35(10), 969.

Veeckman E, Ruttink T, Vandepoele K. (2016) Are we there yet? Reliably estimating the completeness of plant genome sequences. The Plant Cell 28(8), 1759-1768.

Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, Fontana P, Bhatnagar S K, Troggio M, Pruss D, Salvi S, Pindo M, Baldi P, Castelletti S, Cavaiuolo M, Coppola G, Costa F, Cova V, Dal Ri A, Goremykin V, Komjanc M, Longhi S, Magnago P, Malacarne G, Malnoy M, Micheletti D, Moretto M, Perazzolli M, Si-Ammour A, Vezzulli S *et al.* (2010) The genome of domesticated apple (*Malus* x *domestica* Borkh.). Nature Genetics 42 (10) 833-839.

Verde I, Abbott A G, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori M T, Grimwood J, Cattonaro F, Zuccolo A, Rossini L, Jenkins J, Vendramin E, Meisel L A, Decroocq V, Sosinski B, Prochnik S, Mitros T, Policriti A, Cipriani G, Dondini L, Ficklin S, Goodstein D M, Xuan P, Del Fabbro C, Aramini V, Copetti D, Gonzalez S, Horner D S *et al.* (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nature Genetics 45(5), 487.

Vergara-Pulgar C, Rothkegel K, González-Agüero M, Pedreschi R, Campos-Vargas R, Defilippi B G, Meneses C. (2019) De novo assembly of *Persea americana* cv. "Hass" transcriptome during fruit development. BCM Genomics 20, 108; 10.1186/s12864-019-5486-7.

Vurture G W, Sedlazeck F J, Nattestad M, Underwood C J, Fang H, Gurtowski J, Schatz M C. (2017) GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics 33, 2202–2204; 10.1093/bioinformatics/btx153.

Walker J W. (1972) Chromosome numbers, phylogeny, phytogeography of the Annonaceae and their bearing on the (original) basic chromosome number of Angiosperms. Taxon 21(1), 57-65.

Walker B J, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo C A, Zeng Q, Wortman J, Young S K, Earl A M. (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9(11).

Wang B, Tan H W, Fang W. (2015) Developing single nucleotide polymorphism (SNP) markers from transcriptome sequences for identification of longan (*Dimocarpus longan*) germplasm. Horticulture Research 2, 14065; 10.1038/hortres.2014.65.

Wang P, Luo Y, Huang J, Gao S, Zhu G, Dang Z, Gai J, Yang M, Zhu M, Zhang H, Ye X, Gao A, Tan X, Wang S, Wu S, Cahoon E B, Bai B, Zhao Z, Li Q, Wei J, Chen H, Luo R, Gong D, Tang K, Zhang B, Ni Z, Huang G, Hu S, Chen Y. (2020) The genome evolution and domestication of tropical fruit mango. Genome Biology 21(1), 1-17.

Warschefsky E, Penmetsa R V, Cook D R, Von Wettberg E J. (2014) Back to the wilds: tapping evolutionary adaptations for resilient crops through systematic hybridization with crop wild relatives. American Journal of Botany 101(10), 1791-1800.

Wen B, Song W, Sun M, Chen M, Mu Q, Zhang X, Wu Q, Chen X, Gao D, Wu H. (2019a) Identification and characterization of cherry (*Cerasus pseudocerasus* G. Don) genes responding to parthenocarpy induced by GA3 through transcriptome analysis. BMC Genetics 20(1), 65.

Wen T, Dai B, Wang T, Liu X, You C, Lin Z. (2019b) Genetic variations in plant architecture traits in cotton (*Gossypium hirsutum*) revealed by a genome-wide association study. The Crop Journal 7(2), 209-216.

Wenger A M, Peluso P, Rowell W J, Chang P, Hall R J, Concepcion G T, Ebler J, Fungtammasan A, Kolesnikov A, Olson N D, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin C, Phillippy A M, Schatz M C, Myers G, DePristo M A, Ruan J, Marschall T, Sedlazeck F J, Zook J M, Li H, Koren S, Carroll A, Rank D R, Hunkapiller M W. (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nature Biotechnology 1-8.

Wickham H. (2009) Ggplot2: Elegant Graphics for Data Analysis. (Springer-Verlag New York).

Wolfe H S, Toy L R, Stahl A L. (1949) Avocado production in Florida. Bulletin Agricultural Extension Service 141.

Wolstenholme B N. (2013) Ecology: Climate and Soils. The Avocado: Botany, Production and Uses. (eds: Schaffer B, Wolstenholme N, Whiley A W)(CAB International, Wallingford).

Wortman J R, Haas B J, Hannick L I, Smith R K, Maiti R, Ronning C M, Chan A P, Yu C, Ayele M, Whitelaw C A, White O R, Town C D. (2003) Annotation of the Arabidopsis genome. Plant Physiology 132, 461-468.

Wu Y, Bhat P R, Close T J, Lonardi S. (2008) Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. PLoS Genetics 4(10).

Wu G A, Terol J, Ibanez V, López-García A, Pérez-Román E, Borredá C, Domingo C, Tadeo F R, Carbonell-Caballero J, Alonso R, Curk F, Du D, Ollitrault P, Roose M L, Dopazo J, Gmitter F G, Rokhsar D S, Talon M. (2018) Genomics of the origin and evolution of *Citrus*. Nature 554(7692), 311-316.

Wu D, Koch J, Coggeshall M, Carlson J. (2019a) The first genetic linkage map for *Fraxinus pennsylvanica* and syntenic relationships with four related species. Plant Molecular Biology 99(3), 251-264.

Wu Y, Zhou Q, Huang S, Wang G, Xu L A. (2019b) SNP development and diversity analysis for *Ginkgo biloba* based on transcriptome sequencing. Trees 33(2), 587-597.

Wünsch A, Hormaza J I. (2002) Cultivar identification and genetic fingerprinting of temperate fruit tree species using DNA markers. Euphytica 125(1), 59.

Xoca-Orozco L, Cuellar-Torres E A, González-Morales S, Gutiérrez-Martínez P, López-García U, Herrera-Estrella L, Vega-Arreguín J, Chacón-López A. (2017) Transcriptomic analysis of avocado hass (*Persea americana* Mill) in the interaction system fruit-chitosan-Colletotrichum. Frontiers in Plant Science 8, 956.

Xoca-Orozco L, Aguilera-Aguirre S, Vega-Arreguín J, Acevedo-Hernández G, Tovar-Pérez E, Stoll A, Herrera-Estrella L, Chacón-López A. (2019) Activation of the phenylpropanoid biosynthesis pathway reveals a novel action mechanism of the elicitor effect of chitosan on avocado fruit epicarp. Food Research International 121, 586-592.

Xu Q, Chen L, Ruan X, Chen D, Zhu A, Chen C, Bertrand D, Jiao W, Hao B, Lyon M P, Chen J, Gao S, Xing F, Lan H, Chang J, Ge X, Lei Y, Hu Q, Miao Y, Wang L, Xiao S, Kumar Biswas M, Zeng W, Guo F, Cao H, Yang X, Xu X, Cheng Y, Xu J, Liu J *et al.* (2013) The draft genome of sweet orange (*Citrus sinensis*). Nature Genetics 45(1), 59.

Yandell M, Ence D A. (2012) beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13(5), 329.

Yang W, Feng H, Zhang X, Zhang J, Doonan J H, Batchelor W D, Xiong L, Yan J. (2020a) Crop Phenomics and High-throughput Phenotyping: Past Decades, Current Challenges and Future Perspectives. Molecular Plant 13 (2), 187-214.

Yang Y, Sun P, Lv L, Wang D, Ru D, Li Y, Ma T, Zhang L, Shen X, Meng F, Jiao B, Shan L, Liu M, Wang Q, Qin Z, Xi Z, Wang X, Davis C C, Liu J. (2020b) Prickly waterlily and rigid hornwort genomes shed light on early angiosperm evolution. Nature Plants, 1-8.

Yu J, Hu S, Wang J, Wong G K, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X, Cao M, Liu J, Sun J, Tang J, Chen Y, Huang X, Lin W, Ye C, Tong W, Cong L, Geng J, Han Y, Li L, Li W, Hu G, Huang X, Li W, Li J, Liu Z, Li L *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296(5565), 79-92.

Yu Y, Fu J, Xu Y, Zhang J, Ren F, Zhao H, Tian S, Guo W, Tu X, Zhao J, Jiang D, Zhao J, Wu W, Wang G, Ma R, Jiang Q, Wei J, Xie H. (2018) Genome re-sequencing reveals the evolutionary history of peach fruit edibility. Nature Communications 9(1), 1-13.

Yuan Y, Bayer P E, Batley J, Edwards D. (2017) Improvements in genomic technologies: application to crop genomics. Trends in Biotechnology 35(6), 547-558.

Zamir D, Tadmor Y. (1986) Unequal segregation of nuclear genes in plants. Botanical Gazette 147, 355-358.

Zanis M J, Soltis D E, Soltis P S, Mathews S, Donoghue M J. (2002) The root of the angiosperms revisited. Proceedings of the National Academy of Sciences 99(10), 6848-6853.

Zeng D, Tian Z, Rao Y, Dong G, Yang Y, Huang L, Leng Y, Xu J, Sun C, Zhang G, Hu J, Zhu L, Gao Z, Hu X, Guo L, Xiong G, Wang Y, Li J, Qian Q. (2017) Rational design of high-yield and superior-quality rice. Nature Plants 3(4), 17031.

Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, Liang J, Cai C, Liu Z, Liu B, Wang F, Li S, Liu F, Li X, Cheng L, Yang W, Li M, Grossniklaus U, Zheng H, Wang X. (2018) Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Horticulture Research 5(1), 50.

Zhang L, Guo D, Guo L, Guo Q, Wang H, Hou X. (2019a) Construction of a high-density genetic map and QTLs mapping with GBS from the interspecific F1 population of *P. ostii* 'Fengdan Bai'and *P. suffruticosa* 'Xin Riyuejin'. Scientia Horticulturae 246, 190-200.

Zhang L, Hu J, Han X, Li J, Gao Y, Richards C M, Zhang C, Tian Y, Liu G, Gul H, Wang D, Tian Y, Yang C, Meng M, Yuan G, Kang G, Wu Y, Wang K, Zhang H, Wang D, Cong P. (2019b) A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nature Communications 10(1), 1-13.

Zhang L, Chen F, Zhang X, Li Z, Zhao Y, Lohaus R, Chang X, Dong W, Ho S Y W, Liu X, Song A, Chen J, Guo W, Wang Z, Zhuang Y, Wang H, Chen X, Hu J, Liu Y, Qin Y, Wang K, Dong S, Liu Y, Zhang S, Yu X, Wu Q, Wang L, Yan X, Jiao Y, Kong H. *et al.* (2019c) The water lily genome and the early evolution of flowering plants. Nature 1-6.

Zhang Q, Wei X, Liu N, Zhang Y, Xu M, Zhang Y, Ma X, Liu W. (2020) Construction of an SNP-based high-density genetic map for Japanese plum in a Chinese population using specific length fragment sequencing. Tree Genetics & Genomes 16(1), 1-10.

Zhao C, Zhang Y, Du J, Guo X, Wen W, Gu S, Wang J, Fan J. (2019) Crop Phenomics: Current Status and Perspectives. Frontiers in Plant Science 10,714.

Zwaenepoel A, Van de Peer Y. (2019) wgd—simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics 35(12), 2153-2155.

ANEXO 1. Talavera A, Soorni A, Bombarely A, Matas A J & Hormaza J I. Genome-Wide SNP discovery and genomic characterization in avocado (Persea americana Mill.). Scientific Reports 9, 20137 (2019).

ANEXO 1.1. Protocolo para la construcción de genotecas para el genotipado por secuenciación (GBS).

Protocolo para la construcción de genotecas para el genotipado por secuenciación (GBS).

Usando ADN genómico de alta calidad se llevan a cabo los siguientes pasos:

Digestión:

- Alicuotar 10 ul de 10 ng/ul de ADN por muestra en tubos de PCR de 0,2 ml.
- Hacer una mezcla para el número de muestras que necesites. Manteniendo todo en hielo:

Reactivo	**Microlitros (ul) por muestra**
NED buffer 3	2,0
ApeKI1	1,0
H_20	7,0
Total	10,0

- Transferir los 10 ul de la mezcla a cada tubo de la muestra de ADN. Obteniendo un volumen total de 20 ul.
- Incubar durante 2 horas a 75 ^oC en el termociclador.
- Mantener a 4 ^oC.

Ligación:

- Alicuotar 6 ul de cada adaptador (previamente descongelado) en las muestras digeridas.

- Hacer una mezcla para el número de muestras que se necesite. Manteniendo todo en hielo:

Reactivo	Microlitros (ul) por muestra
10X T4 DNA Ligase Reaction Buffer	5,0
T4 DNA Ligase (NEB, MK0202L)	1,6
H_20	17,4
TOTAL	24,0

- Transferir 24 ul de la mezcla a cada tubo.

- Centrifugar (En este paso, el volumen final debería de ser 50 ul (20ul de la digestión + 6 ul de los adaptadores + 24 ul de la mezcla).

- Incubar en el termociclador durante 2 horas a 22 ^oC aproximadamente (Temperatura ambiente).

- Incubar en el termociclador durante 30 minutos a 65 ^oC.

- Mantener a 4 ^oC.

- Chequear los pasos anteriores mediante una PCR.

Agrupación y limpieza:

- En este paso se lleva a cabo una limpieza usando el kit "Monarch PCR and DNA Cleanup" de BioLabs.

- Combinar 5 ul de cada muestra ligada en un tubo de 1,5 ml conteniendo 5 volúmenes de buffer de enlace por 1 volumen de muestra. (Recomendación: usar 2 tubos de 1,5 ml para 98 muestras).

- Siguiendo las indicaciones del kit Monarch PCR and DNA Cleanup" de BioLabs, traspasar las muestras a columnas de unión.

- Eluir las muestras en 50 ul de buffer NE.

Amplificación PCR:

- Preparar la siguiente mezcla para la PCR para cada muestra:

Reactivos	**Microlitros (ul) por muestra**
Agrupación de ADN limpio y ligado	2
NEB 2X Taq Master Mix (NEB #M0270S)	25
Primers 1 y 2 (25 uM total, 12 uM de cada	2
H_20	21
Total	50

En el termociclador:

- 30 segundos a 95^{o}C.

- 18 ciclos a:

30 segundos a 95 ^{o}C.

20 segundos a 62 ^{o}C.

30 segundos a 68 ^{o}C.

- 5 minutos a 68 ^{o}C.

- Mantener a 4 ^{o}C.

Limpieza

- En este paso se utiliza el kit "Monarch PCR and DNA Cleanup" de Biolabs, siguiendo el protocolo descrito por el fabricante.

- Eluir en 35 ul de buffer NE.

- Chequear la concentración con el fluorómetro Qubit 2.0.

Pasos finales

- Usar BluePippin para seleccionar el tamaño de fragmentos (250-600 bp)
- Chequear la concentración con el fluorómetro Qubit 2.0.
- Chequear el tamaño de la librería con Agilent 2100 Bioanalyzer System.