UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana

Optimización de la secuenciación del genoma de berenjena (Solanum melongena) a baja cobertura

Trabajo Fin de Máster

Máster Universitario en Mejora Genética Vegetal

AUTOR/A: Baraja Fonseca, Virginia

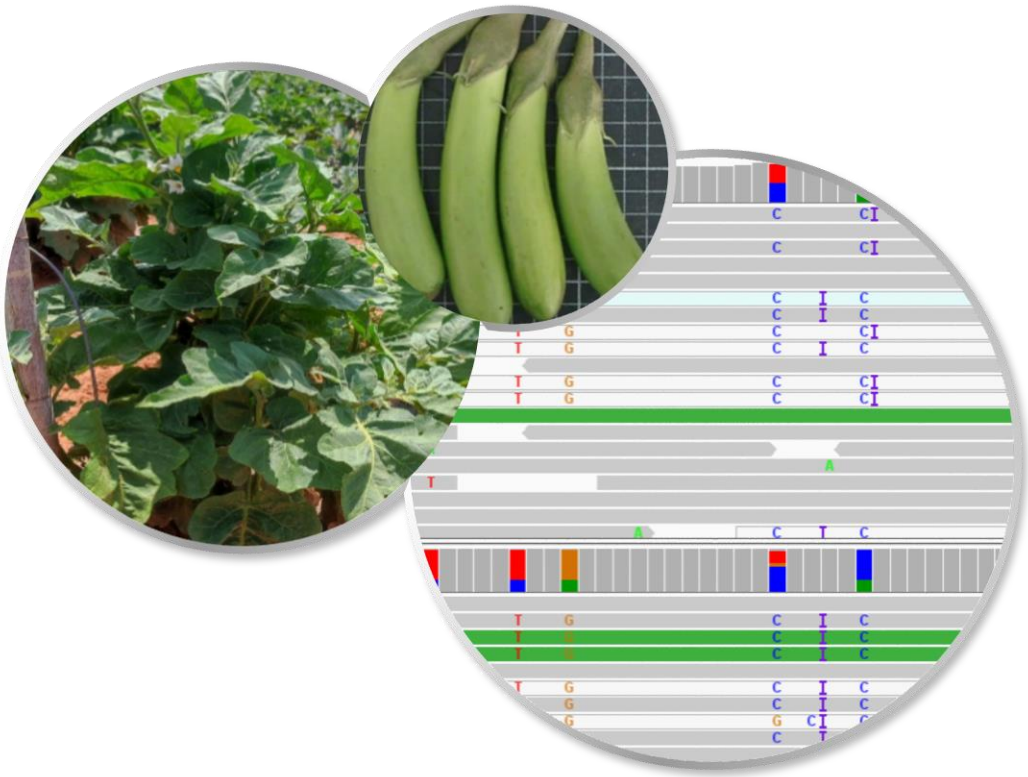
Tutor/a: Gramazio, Pietro

Cotutor/a: Plazas Ávila, María de la O

Director/a Experimental: ARRONES OLMO, ANDREA

CURSO ACADÉMICO: 2022/2023

Optimización de la secuenciación del genoma de berenjena (*Solanum melongena*) a baja cobertura



Autora: Virginia Baraja Fonseca

Tutor: Pietro Gramazio

Cotutora: María de la O Plazas Ávila

Directora experimental: Andrea Arrones Olmo

COMAV
Instituto de Conservación y Mejora de la Agrodiversidad Valenciana

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSIDAD POLITÉCNICA DE VALENCIA

MÁSTER EN MEJORA GENÉTICA VEGETAL

CURSO 2022-2023

RESUMEN

La berenjena común (*Solanum melongena* L., 2n = 2x = 24) es la sexta hortaliza con mayor producción mundial, ocupando el tercer lugar entre las solanáceas, después del tomate y la patata. A pesar de su gran relevancia, en comparación con otros cultivos económicamente importantes y especies modelo, se ha quedado rezagada en el avance de herramientas genómicas. Sin embargo, el progreso tecnológico, la disminución de los costes de secuenciación y los avances logrados en el desarrollo de programas bioinformáticos están permitiendo la reducción de la brecha entre especies en términos de disponibilidad de recursos genómicos. Esto se refleja en el acceso a genomas de referencias, incluso de cultivos minoritarios, así como en la adopción de nuevas estrategias de genotipado masivo, como el skim whole genome resequencing (SWGR), que permite la identificación de un gran número de polimorfismos, incluso a bajas coberturas de secuenciación (<10X).

Para analizar y optimizar la estrategia SWGR en berenjena, se evaluaron cinco coberturas de secuenciación (1-5X) generadas *in silico.* La evaluación consistió en el mapeo de las lecturas contra el genoma de referencia de berenjena 67/3, el marcado de las secuencias duplicadas, la identificación de polimorfismos y su filtrado a diferentes profundidades mínimas de mapeo (1-10FD). Finalmente, cada conjunto de datos se comparó con un estándar de referencia resultante de la resecuenciación a 20X del mismo genotipo. En relación a la identificación de polimorfismos, se observó una reducción en el número de variantes al marcar las secuencias duplicadas con respecto a no marcarlas, salvo para coberturas de secuenciación bajas (1X). El número de polimorfismos identificados aumentó con la cobertura de secuenciación (227.346 y 932.989 polimorfismos a 1X y 5X, respectivamente, filtrando a 1FD), y disminuyó al aplicar filtros de cobertura mínima más altos (3.248 y 62.802 polimorfismos a 1X y 5X, respectivamente, filtrando a 10FD). El uso de coberturas de secuenciación más bajas ofrece una mayor precisión a nivel de polimorfismos y genotipos con respecto al estándar de referencia al utilizar umbrales de profundidad mínima bajos (1-4FD), alcanzando el 69% y 62% a 1X y 5X para 1-2FD, respectivamente, mientras que se alcanza una mayor precisión con el uso de coberturas de secuenciación más altas al utilizar filtros de umbral más alto (5-10FD), siendo del 43% y 54% a 1X y 5X para 10FD, respectivamente.

Este estudio ofrece una valiosa evaluación de parámetros relevantes a considerar en la caracterización genotípica de la berenjena mediante SWGR, tomando en cuenta variables económicas, número de genotipos, abundancia de polimorfismos y tipo de estudio o programa de mejora, entre otras. Es por ello que la mejor combinación de cobertura de secuenciación y filtrado de polimorfismos debe ser evaluada en cada caso concreto. En términos generales, los resultados indican que SWGR es una herramienta efectiva, rentable y pragmática para el análisis genético y genómico de la berenjena, lo que permitirá a los investigadores avanzar en el desarrollo de nuevos recursos genéticos y a los mejoradores implementar programas de mejora más precisos, rápidos y eficientes.

Palabras clave: berenjena; genotipado; skim whole genome resequencing; cobertura de secuenciación; estándar de referencia.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE LISTAS DE COMANDOS

ÍNDICE DE FIGURAS

1. INTRODUCCIÓN

1.1. LA BERENJENA

1.1.1. Clasificación taxonómica

La berenjena común (*Solanum melongena* L., 2n = 2x = 24) es una especie perteneciente a la familia Solanaceae y al género *Solanum*, específicamente al subgénero *Leptostemonum*, que se distingue por agrupar especies espinosas (Bohs, 2005). Esta familia comprende 101 géneros, entre los que se encuentran varios importantes a nivel económico y hortícola, como *Nicotiana* y *Petunia* (Stehmann *et al*., 2000; Lewis, 2011). El género *Solanum* no es únicamente importante por su gran tamaño, llegando a agrupar aproximadamente 1.500 especies (Frodin, 2004), sino que también lo es por agrupar especies de gran importancia económica para los humanos, como la patata (*S. tuberosum*) (Spooner, 1990) y el tomate (*S. lycopersicum*) (Tanksley, 2004), además de la berenjena.

El cultivo de berenjena se encuentra extendido por todos los continentes, a excepción de La Antártida, ocupando un rango diverso de hábitats, desde bosques tropicales hasta desiertos. Sin embargo, se cultiva más ampliamente en los países del Sudeste Asiático, de África y del Mediterráneo (Bean, 2004; Aubriot *et al*., 2016; Vorontsova and Knapp, 2016). Además de la berenjena común, existen otras dos especies de berenjena cultivadas, la berenjena gboma (*S. macrocarpon*) y la berenjena escarlata (*S. aethiopicum*), las cuales se cultivan localmente en África o como cultivos menores en otros lugares (Schippers, 2000).

1.1.2. Descripción botánica y agronómica

Al igual que en muchas otras especies, existe una elevada variación fenotípica y diversidad metabólica entre variedades (Martínez-Ispizua *et al*., 2021). *Solanum melongena* se caracteriza por ser una planta herbácea, a pesar de que sus tallos poseen tejidos que le otorgan una apariencia de arbusto, con una estatura que puede variar de 1,0 a 1,8 m y con un sistema radicular que puede llegar a un metro de profundidad (AVGRIS, 2022). En líneas generales, sus hojas son lobuladas, gruesas y de color verde, presentando nerviaciones con espinas y reversos revestidos con vellosidades grisáceas. Las flores son pentaméricas y con sépalos constantes. El cáliz es espinoso, aunque actualmente se tiende a cultivar variedades de cáliz sin espinas. El color de los pétalos varía desde el blanco hasta el morado oscuro pasando por el azul (AVGRIS, 2022). Tras la etapa de cuajado, se obtiene una baya unida a la planta por un largo pedúnculo. Se trata de un fruto carnoso, por lo general, alargado y ovoide, con carne dura y piel gruesa, en cuyo interior se alberga una gran cantidad de pequeñas semillas. La epidermis es llana y brillante, y su color varía entre genotipos, desde morado oscuro a morado claro, existiendo variedades con frutos blancos, verdes o casi negros (**Figura 1**) (García Fortea *et al*., 2021; Martínez-Ispizua *et al*., 2021; AVGRIS, 2022).

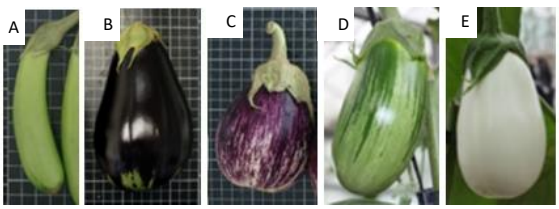


Figura 1. Diferentes fenotipos de frutos de *S. melongena* según la accesión: **(A)** MM1597, **(B)** Black Beauty, **(C)** IVIA371, **(D)** MEL3, **(E)** MEL1 (Adaptación de García Fortea *et al*., 2021).

En cuanto a las condiciones edafoclimáticas óptimas para su desarrollo vegetativo, las temperaturas han de rondar los 27 – 32 ºC a lo largo del día y los 21 – 27 ºC durante la noche. El desarrollo de los frutos se ve favorecido cuando la temperatura oscila los 22 – 26 ºC. La humedad ideal para su cultivo varía entre el 50 al 65%. También se recomienda realizar el cultivo en suelos francos y profundos,

puesto que la raíz es vigorosa y profunda, y evitar los suelos arcillosos (Maroto, 2002; García Fortea *et al*., 2021).

1.1.3. Importancia económica

La berenjena es uno de los cultivos económicamente más importantes, ocupando la tercera posición entre las solanáceas, después de la patata y el tomate, y la sexta entre los cultivos hortícolas, con una producción global de 58,6 M de toneladas en 2021 (FAOSTAT, 2022). Actualmente, el principal país productor es China (37,43 Mt), seguido de India (12,87 Mt) y Egipto (1,29 Mt) (**Figura 2**) (FAOSTAT, 2022). España ocupa el décimo puesto a nivel global; sin embargo, es el segundo país europeo con mayor producción, por detrás de Italia (**Figura 2**). En 2021, en el territorio español se produjeron 265.300 toneladas en 3.600 hectáreas (FAOSTAT, 2022). El cultivo de la berenjena tiene lugar fundamentalmente en Andalucía y, más concretamente, en la provincia de Almería. Esta comunidad concentra más de un 80% de la producción nacional, con 245.760 toneladas en una extensión de cultivo de 2.871 hectáreas, seguida por la Comunidad Valenciana, con 12.702 toneladas en una extensión de 251 hectáreas (MAPA, 2021).

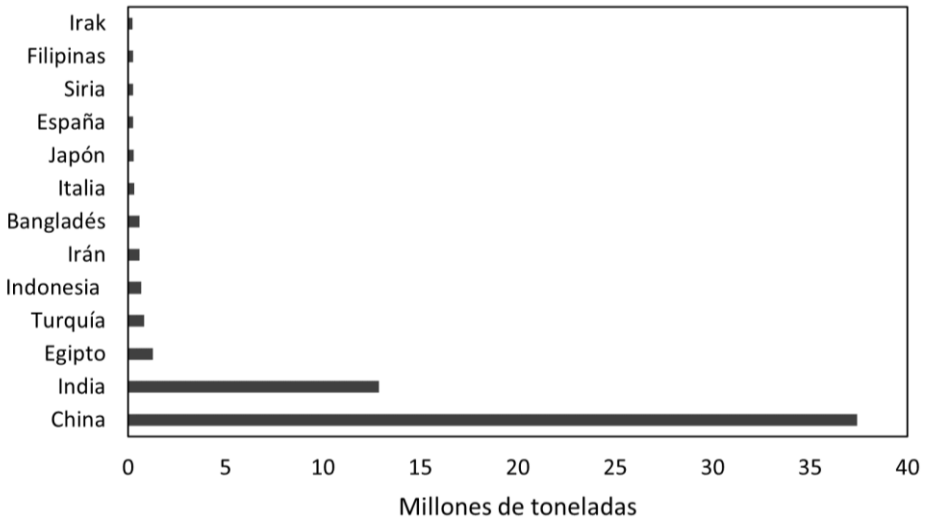


Figura 2. Distribución de la producción mundial de berenjena entre los 13 países más productores. China: 37,43 Mt; India: 12,87 Mt; Egipto: 1,29 Mt; Turquía: 0,83 Mt; Indonesia: 0,68 Mt; Irán: 0,60 Mt; Bangladés: 0,59 Mt; Italia: 0,31 Mt; Japón: 0,28 Mt; España: 0,27 Mt; Siria: 0,25 Mt; Filipinas: 0,24 Mt; Irak: 0,22 Mt (FAOSTAT, 2022).

1.1.4. Origen y domesticación

A pesar de los estudios realizados sobre el origen y la domesticación de la berenjena, aún existe cierto grado de incertidumbre acerca del trayecto evolutivo de esta especie cultivada. Tradicionalmente, India y el sureste de Asia han sido consideradas como las dos regiones de origen más probables de la berenjena común, debido a la presencia de registros escritos igualmente antiguos sobre el uso de berenjenas, además de una amplia diversidad de razas autóctonas y de parientes silvestres cercanos (Vavilov, 1951; Meyer *et al.*, 2012). Sin embargo, estudios genéticos recientes realizados por Page *et al*., (2019) han identificado la región de Malasia, Tailandia e Indonesia como el centro de domesticación de la berenjena.

Uno de los parientes silvestres más cercanos de la berenjena cultivada es *S. incanum*, especie nativa de África, que fue transportada hasta la región Indo-China, donde evolucionó el verdadero ancestro de la berenjena cultivada, *S. insanum* (Weese and Bohs, 2010; Knapp *et al*., 2013). Éste se dividió en dos grupos: uno occidental formado por accesiones indias y otro oriental formado por accesiones de Tailandia, Indonesia y Malasia, siendo el segundo, como bien se ha adelantado anteriormente, el

grupo a partir del cual se domesticó la berenjena común (Page *et al.*, 2019). A pesar de la falta de información acerca de cuándo apareció la especie *S. melongena*, su distribución a lo largo de todos los continentes sí que se puede situar de forma aproximada en el tiempo. El cultivo de berenjena se extendió hacia el este, alcanzando Japón alrededor del siglo VIII, y posteriormente hacia el oste a lo largo de la Ruta de la Seda, llegando a Europa y África durante el s. XIV. Más tarde se introdujo en América, poco después de la llegada de los europeos (Prohens *et al*., 2005).

Como consecuencia del proceso de domesticación, la berenjena cultivada presenta una menor diversidad genética que su ancestro y otras especies silvestres estrechamente relacionadas (Vorontsova *et al.*, 2013; Page *et al.*, 2019), al igual que ocurre en otros cultivos, sobre todo autógamos, como el tomate (Bellucci *et al*., 2014; Gao *et al.*, 2019). Además, los procesos de selección a los que son sometidos continuamente los cultivares reducen aún más la diversidad genética presente en los mismos (Fu, 2015). Por tanto, se presenta la posibilidad de utilizar los recursos genéticos de las especies silvestres para mejorar los cultivares actuales y que puedan adaptarse a nuevas condiciones edafoclimáticas (Swarup *et al*., 2021).

1.1.5. Especies silvestres estrechamente relacionadas

El interés en las especies silvestres de cultivos importantes desde el punto de vista económico y agronómico reside, fundamentalmente, en su uso como fuentes de resistencias y/o tolerancias frente a estreses tanto bióticos como abióticos. Esto se debe a que se pueden encontrar en una amplia gama de condiciones ambientales (Vorontsova and Knapp, 2016) y, por lo tanto, pueden presentar una gran diversidad alélica para características agronómicas importantes (Namisy *et al.*, 2019; Brenes *et al.*, 2020; Kouassi *et al.*, 2021). Estas especies también son variables en cuanto a compuestos del fruto, tales como fenoles y ácidos, por lo que se pueden utilizar para mejorar la calidad de las berenjenas cultivadas (Kaushik, 2020).

En función de la facilidad de hibridación y de las relaciones filogenéticas con la berenjena cultivada, las especies silvestres se pueden clasificar en diferentes acervos genéticos. El acervo primario está constituido únicamente por la propia berenjena común y su ancestro *S. insanum*, con el que se puede cruzar fácilmente y dar lugar a híbridos fértiles (**Figura 3**) (Ranil *et al.*, 2017). Esta especie silvestre se encuentra distribuida por India y el sureste asiático. Se caracteriza por presentar flores pentámeras de estilo largo y fruta verde con pulpa jugosa (**Figura 4**) (Knapp *et al.*, 2013). El acervo secundario incluye especies del "clado berenjena" (*S. campylacanthum*, *S. lichtensteinii* y *S. linnaeanum*), la hermana "grado anguivi", como *S. anguivi* y *S. dasyphyllum*, ancestros silvestres de las berenjenas escarlata y gboma, respectivamente, y el "clado Madagascar" (*S. pyracanthos*) (**Figura 3**) (Vorontsova *et al.*, 2013). Por último, en el acervo terciario podemos encontrar especies más alejadas, como *S. torvum* o *S. sisymbriifolium* (**Figura 3**), con las que en ocasiones es posible obtener híbridos estériles o con baja fertilidad tras emplear técnicas tales como el rescate de embriones o la hibridación somática (Plazas *et al.,* 2016).

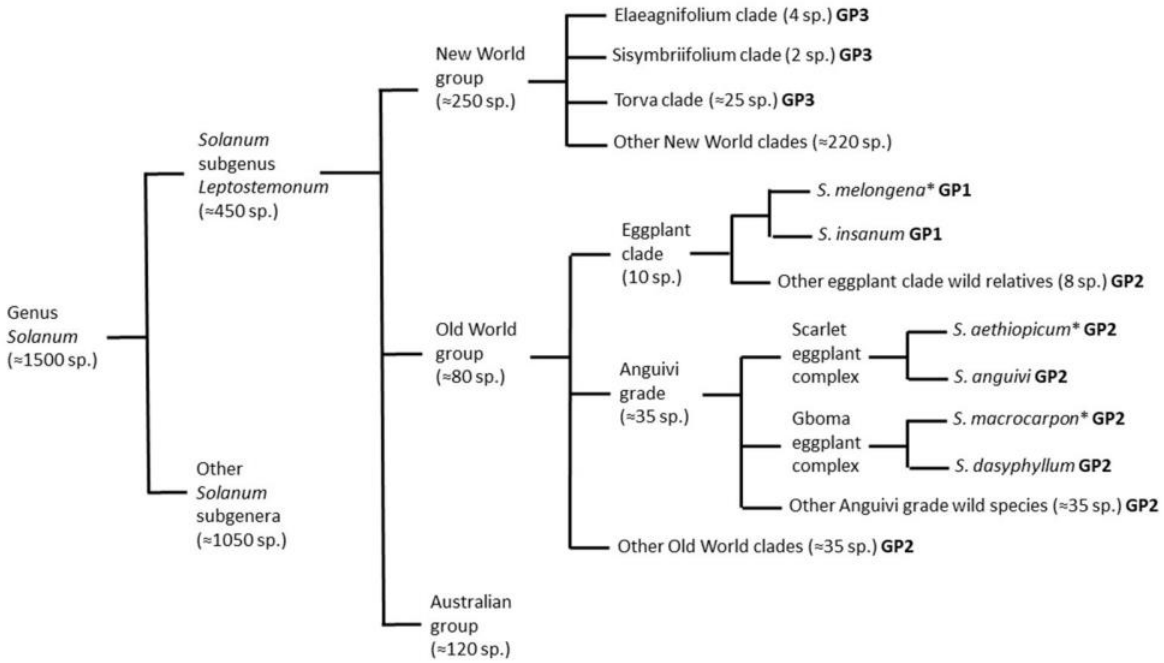


Figura 3. Relación taxonómica de la berenjena cultivada (*S. melongena)* con otras especies cultivadas (marcadas con un asterisco) y silvestres del género *Solanum* junto con el acervo genético al que pertenecen. GP1: acervo primario; GP2: acervo secundario. GP3: acervo terciario (Taher *et al.*, 2017).

Figura 4. Fruto y flor del ancestro *S. insanum*. **A**: fruto verde con el cáliz espinoso. **B**: flor pentamérica de color morado unida a un tallo espinoso (Toppino *et al*., 2021).

1.2. IMPORTANCIA Y DESARROLLO DE LA GENÓMICA

Uno de los principales avances de la genómica tiene que ver con el desarrollo de la tecnología de secuenciación, entendida como el proceso por el cual es posible conocer la estructura primaria o secuencia de nucleótidos de una determinada región genómica o del genoma completo de uno o de un conjunto de individuos. Se trata de una poderosa herramienta en la investigación genética y genómica de plantas que se implementó en los laboratorios a finales del siglo XX, siendo la secuenciación de Sanger la técnica que más ha trascendido (Maxam and Gilbert, 1977; Sanger *et al.*, 1977). En las últimas décadas se han desarrollado las plataformas de secuenciación de segunda generación (NGS), que destacan por su alta velocidad de procesamiento, el bajo costo y su capacidad para secuenciar el genoma completo de un individuo (Metzker, 2010). Las más empleadas en el ámbito de la genómica, tanto con plantas modelo como con plantas no modelo, han sido 454/Roche, ABI/SOLiD, Helicos y Illumina/Solexa (Unamba *et al*., 2015). Sin embargo, existe una mejora continua de dichas plataformas, lo que conlleva la disponibilidad de nuevos y alternativos métodos de secuenciación. Un ejemplo es DNBSeq, una reciente plataforma de secuenciación que se ha desarrollado con el objetivo de servir de alternativa a Illumina utilizando nanopartículas de ADN (DNB) y síntesis combinatoria de sonda-ancla (cPAS) (Patterson *et al.*, 2019).

La propia evolución tecnológica y la reducción de costes que ha experimentado la metodología de secuenciación (Wetterstrand, n.d.), ha abierto una gran ventana para el desarrollo de recursos genéticos que hasta la fecha no habían estado disponibles. Uno de estos recursos son los genomas de referencia, es decir, la secuencia completa y, en el mejor de los casos, anotada del genoma de un determinado individuo de una especie concreta (**Figura 5**). El primer genoma disponible fue el de *Arabidopsis thaliana* en el año 2000 (The Arabidopsis Genome Initiative, 2000). Dos años más tarde se publicó el de arroz, siendo el primer cultivo en tener su genoma secuenciado (Yu *et al.*, 2002). La disponibilidad del genoma de referencia de un cultivo acelera drásticamente la mejora del mismo, pues posibilita acceder a la información de su conjunto completo de genes y de elementos reguladores, así como conocer su estructura genómica básica (Purugganan and Jackson, 2021). Otro recurso sobre el que han tenido un gran impacto las nuevas plataformas de secuenciación han sido los marcadores moleculares, que son fragmentos de ADN asociados con un carácter en concreto que presentan diferencias entre organismos individuales o especies (Kordrostami and Rahimi, 2015). Antes de la era de la secuenciación de segunda generación, los marcadores que se empleaban eran, sobre todo, los polimorfismos de longitud de fragmento de restricción (RFLPs) (Burr *et al.*, 1983) y los polimorfismos de longitud de fragmento amplificados (AFLPs) (Angiolillo *et al.*, 1999). Sin embargo, presentaban una serie de problemas, tales como su limitada cobertura del genoma, la necesidad de elaborar un gran número de electroforesis, lo cual suponía un consumo importante de recursos y tiempo, y su alto coste. Todo ello ha desencadenado que en las últimas décadas hayan sido sustituidos por los marcadores basados en secuencia, tales como las repeticiones de secuencia simple (SSRs) (Liu *et al.*, 2018) y los SNPs (Jing *et al.*, 2019). Ambos marcadores son codominantes, robustos y fáciles de identificar mediante las plataformas de NGS (Mishra *et al.*, 2022). No obstante, los SNPs son más ampliamente utilizados en los diferentes análisis genómicos debido a su mayor abundancia, ubicuidad y facilidad de automatización (Thomson, 2014).

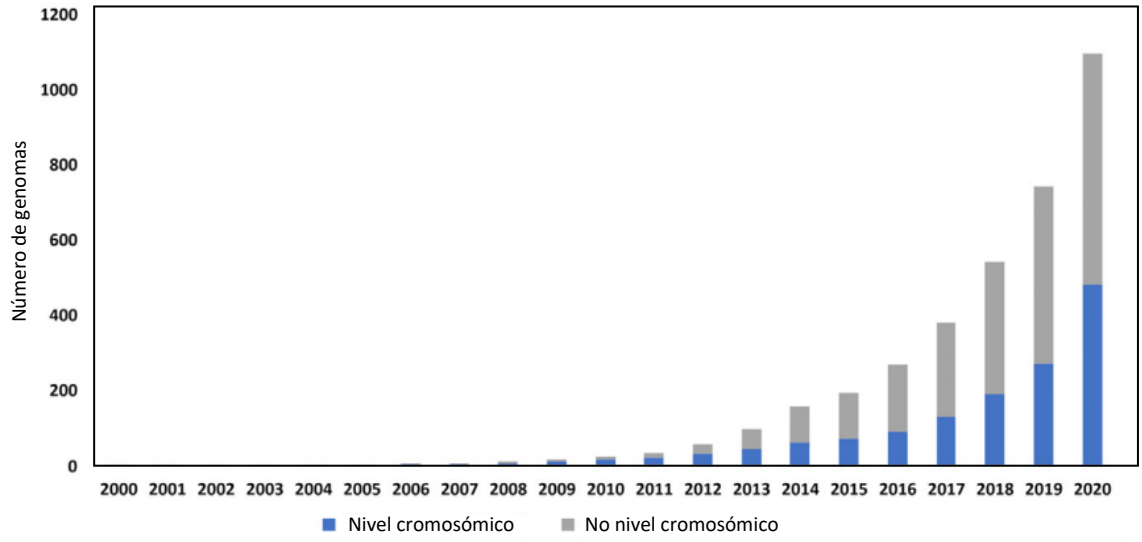


Figura 5. Genomas de especies vegetales publicados. Número de genomas de plantas secuenciados a nivel cromosómico y no cromosómico desde la publicación del genoma de *Arabidopsis thaliana* en el año 2000 hasta el año 2020 (Sun *et al.*, 2022).

Los marcadores moleculares tienen infinidad de aplicaciones en el ámbito de la mejora genética vegetal. Una de las más importantes es la selección asistida por marcadores moleculares, que consiste en genotipar en plántulas los marcadores estrechamente vinculados a un determinado carácter con el fin de predecir su fenotipo para dicho carácter (Nayak *et al.*, 2017). Esto permite seleccionar individuos antes de que el carácter en cuestión se manifieste y pueda realizarse una detección

fenotípica, y ayuda en la selección en base a caracteres cuya evaluación fenotípica es compleja (Moodley *et al.*, 2019; Rana *et al.*, 2019; Wang *et al.*, 2019). Por otra parte, en muchas ocasiones, el objetivo final, más allá de identificar marcadores asociados al carácter, es conocer el gen o el locus de rasgo cuantitativo (QTL) responsable del fenotipo, que se puede clonar y utilizar para la mejora vegetal desde la perspectiva de la transformación genética (Akano *et al*., 2002; Liu *et al*., 2012; Wu *et al.*, 2019a). Los marcadores moleculares también se han empleado en estudios filogenéticos, para entender las relaciones evolutivas dentro y entre especies, para distinguir cultivares estrechamente relacionados y para caracterizar las accesiones de germoplasma vegetal (Sukhotu and Hosaka, 2006; Calonje *et al.,* 2009; Butiuc-Keul *et al.*, 2019; Jamil *et al*., 2021).

1.2.1. Genotipado por secuenciación (GBS)

El proceso por el cual se determinan las diferencias entre los genomas de diferentes individuos o, lo que es lo mismo, se identifican marcadores moleculares, se conoce como genotipado (Davey *et al*., 2011). Hasta la fecha se han desarrollado y utilizado diversas herramientas de genotipado que se basan en diferentes aproximaciones moleculares. Las más destacables, aunque no las únicas, son: (1) extensión de cebadores que hibridan en la región adyacente al polimorfismo, de forma que la primera base que se incorpora a la secuencia bajo la acción de la ADN polimerasa permite la identificación del genotipo, ya sea mediante fluorescencia o espectrometría de masas (Sokolov, 1990); (2) hibridación de una sonda específica de alelo (LGC Biosearch Technologies, 2013); y (3) corte con enzimas de restricción para identificar polimorfismos localizados en sitios de restricción (Lyamichev *et al.*, 1999) (**Figura 6**). Sin embargo, estas estrategias de genotipado requieren la previa identificación del polimorfismo, su validación y el diseño de los cebadores, las sondas o la elección de las enzimas de restricción a emplear. Además, considerando que en la mayoría de los casos se aplican en forma de *arrays* de SNPs, con frecuencia no están disponibles para cultivos de especies no modelo y no posibilitan la identificación de nuevos loci (Rasheed *et al.*, 2017).

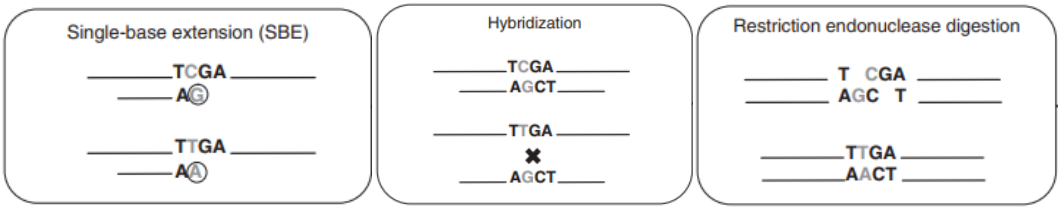


Figura 6. Tecnologías de genotipado de SNP. **A**: Extensión de cebadores. El cebador hibrida en la posición anterior al SNP, de forma que, según el nucleótido que sea añadido por la ADN polimerasa se puede conocer el genotipo del SNP. **B**: Hibridación de sondas específicas de alelo. Conociendo la secuencia del genoma, se pueden diseñar sondas que hibriden de forma específica. Sin embargo, la presencia de un SNP hace que dicha sonda no hibride. **C**: Detección de polimorfismos posicionados en regiones que reconocen enzimas de restricción. La presencia de un SNP en dichas zonas impide que el enzima pueda realizar el corte (Adaptación de Bayés and Gut, 2012).

Todo ello ha llevado a desarrollar otros métodos de genotipado, como el GBS, que aprovecha el potencial de la tecnología de NGS tanto para identificar un gran número de variantes en el genoma como para determinar el genotipo que presenta un determinado individuo o conjunto de individuos en dichas posiciones (Peterson *et al.*, 2014). Existen dos enfoques para llevar a cabo este tipo de genotipado. El primero de ellos es la secuenciación de representación reducida (RRS) (van Tassell *et al.*, 2008; Hegarty *et al.*, 2013), en la que se lleva a cabo un paso de reducción de la complejidad de la biblioteca de secuencias de ADN. La desventaja de este enfoque es la pérdida de una gran cantidad de información acerca de la presencia de variantes en el genoma. Sin embargo, cuenta con la ventaja de ser una estrategia económicamente asequible y de no requerir de un genoma de referencia (**Tabla 1**) (Scheben *et al.*, 2017). El segundo enfoque es la resecuenciación del genoma completo (WGR)

(Ratnaparkhe *et al*., 2020; Ren *et al*., 2021), que, a diferencia del anterior, consiste en conseguir una representación de todo el genoma del individuo gracias a que prescinde de la reducción de la complejidad de la biblioteca de ADN y a la utilización de una cobertura de secuenciación medio/alta. No obstante, implica un mayor costo y que, por tanto, únicamente se utilice con genomas de pequeño tamaño o en cultivos económicamente más importantes (**Tabla 1**) (Scheben *et al.*, 2017). Además, requiere la disponibilidad de un genoma de referencia.

1.2.2. Resecuenciación del genoma completo a bajas coberturas. Herramientas bioinformáticas

Más recientemente, gracias a que los avances tecnológicos han permitido una disminución significativa de los costes de secuenciación, posibilitando, entre otras cosas, el ensamblaje de genomas de referencia incluso de cultivos minoritarios, se ha implementado una nueva estrategia de genotipado, skim whole genome resequencing (SWGR), con la que se consigue una representación imparcial de todo el genoma empleando muy bajas coberturas de secuenciación (<10X) (Golicz *et al*., 2015; Kumar *et al.*, 2021). Se trata de un método de genotipado de alto rendimiento con el que es posible la identificación de una alta densidad de marcadores (**Tabla 1**) (Malmberg *et al*., 2018; Happ *et al.*, 2019).

Tabla 1. Comparación entre diferentes técnicas de genotipado por secuenciación: secuenciación de representación reducida (RRS), resecuenciación del genoma completo (WGR) y skim whole genome resequencing (SWGR).

	RRS	WGR	SWGR
Coste por muestra	Bajo	Alto	Bajo
Tasa de polimorfismos	Bajo	Alto	Alto
Enzimas de restricción	Requeridas	No requeridas	No requeridas
Cobertura de secuenciación	Variada	Alta	Baja
Complejidad de los análisis	Moderado	Alto	Alto
Genoma de referencia	No esencial	Esencial	Esencial
Cobertura del genoma	Baja cobertura del genoma con baja resolución	Alta cobertura del genoma con alta resolución	Alta cobertura del genoma con alta resolución

Además de por los avances tecnológicos en secuenciación, el genotipado, como el SWGR, se ha visto ayudado por el desarrollo de herramientas bioinformáticas, necesarias para el análisis de las secuencias obtenidas tras la secuenciación de las moléculas de ADN (Luscombe *et al.*, 2001; Diniz and Canduri, 2017). De hecho, hoy en día resultaría inviable manejar y analizar la enorme cantidad de datos que se generan a partir de las plataformas de NGS si no fuera por la disponibilidad de equipos informáticos que cuentan con una gran memoria de almacenamiento y de programas y/o softwares especializados, muchos de ellos basados en el sistema operativo LINUX (Batley and Edwards, 2009; Lee *et al*., 2012).

El análisis bioinformático de los datos generados por la técnica de genotipado SWGR consta de cinco pasos principales. El primero de ellos es el filtrado de las lecturas resultantes de la secuenciación para eliminar los adaptadores utilizados en la preparación de las librerías genómicas y en el proceso de secuenciación, las lecturas de una reducida longitud y aquellas que presentan una baja calidad (Ewing *et al*., 1998). Para ello, existen diferentes herramientas bioinformáticas (**Tabla 2**), como fastq-mcf (Aronesty, 2011) y trimmomatic (Bolger *et al.*, 2014). Ambos programas son ejecutables mediante la terminal de LINUX y funcionan con archivos de extensión FASTA, que es la extensión que suelen presentar los archivos generados por NGS, los cuales contienen información de múltiples secuencias

de lecturas (Leonard *et al*., 2006). Una vez llevado a cabo este paso, se pueden emplear herramientas como FASTQC (**Tabla 2**) (Andrews, 2016) para analizar la calidad de los datos en cuanto a longitud de secuencia, calidad de base y de secuencia, y proporción de secuencias repetidas, entre otros. Este paso se puede realizar adicionalmente con los datos sin filtrar con el fin de evaluar el impacto del proceso de filtrado en la calidad de los mismos.

El siguiente paso consiste en alinear las lecturas remanentes contra el genoma de referencia de la especie en estudio o, en caso de no estar disponible, contra el genoma de una especie cercana, para, a continuación, poder llevar a cabo la identificación de las variantes genéticas presentes en las lecturas en comparación con el genoma de referencia (Li *et al*., 2008a; Kim *et al.*, 2014). Para realizar el alineamiento de las secuencias existe una gran variedad de herramientas informáticas (**Tabla 2**) capaces de alinear billones de lecturas cortas de ADN de forma precisa y rápida. Se pueden diferenciar dos grupos de alineadores: los basados en la transformada de Burrows-Wheeler (BWT) (Burrows and Wheeler, 1994), como Bowtie (Langmead and Salzberg, 2012), SOAP2 (Li *et al*., 2008b) y BWA (Li and Durbin, 2009); y los basados en el algoritmo de hash (Ning *et al.*, 2001), como MAQ (Li *et al.*, 2008a) y Stampy (Lunter and Goodson, 2011). Los primeros se caracterizan por ser rápidos y eficientes en memoria, pero menos sensibles que los segundos (Magi *et al*., 2010; Holtgrewe *et al*., 2011). La elección del alineador es importante, puesto que la precisión con que se lleva a cabo el alineamiento de las lecturas repercute en la identificación de las variantes presentes en las muestras de ADN analizadas (Altmann *et al.*, 2012). Además, en algunas situaciones también puede existir cierta incertidumbre acerca de la región del genoma de referencia con la que se alinea la secuencia en cuestión. No obstante, esta cuestión se puede abordar mediante el empleo de lecturas de extremos emparejados. El uso de este tipo de lecturas también puede llegar a resolver otros problemas, como la dificultad de alinear secuencias cuando existe una alta diversidad entre las lecturas secuenciadas y el genoma de referencia, o cuando se trata de una especie poliploide (Lee and Schatz, 2012).

En ocasiones, antes de proceder a identificar las variantes en las secuencias de ADN, es necesario marcar las lecturas que puedan estar duplicadas para evitar problemas de sobrerrepresentación de algunas regiones del genoma (Pabinger *et al.*, 2014). Las duplicaciones pueden surgir como consecuencia de la acción de la ADN polimerasa durante la preparación de la biblioteca por PCR o durante el proceso de secuenciación, ya que las mismas moléculas de ADN pueden secuenciarse varias veces (van der Auwera *et al.*, 2013; Ebbert *et al*., 2016). Por ejemplo, la herramienta MarkDuplicates del conjunto de herramientas de Picard utiliza un algoritmo para diferenciar las lecturas primarias de las duplicadas (**Tabla 2**). Esta herramienta analiza archivos con extensión SAM/BAM, que son archivos que contienen información sobre el alineamiento de las secuencias. El archivo resultante, también de extensión BAM, contiene la misma información que el original, solo que las lecturas duplicadas se encuentran marcadas como tales (van der Auwera *et al.*, 2013).

La identificación de variantes en las secuencias de ADN es un paso importante en el análisis de datos de NGS, que implica la identificación de mutaciones que han de ser respaldadas por varias lecturas. Existen herramientas basadas en métodos heurísticos, como SAMTools (Li *et al.*, 2009) y VarScan (Koboldt *et al.*, 2009), y herramientas basadas en métodos probabilísticos, como GATK (McKenna *et al.*, 2010), FreeBayes (Garrison and Marth, 2012) y SOAPsnp (Li *et al*., 2008b) (**Tabla 2**). GATK engloba un conjunto de herramientas bioinformáticas, entre las que se encuentra HaplotypeCaller, que permite identificar SNPs, inserciones y deleciones simultáneamente haciendo un reensamblaje local y considerando los haplotipos. Por defecto, es muy permisivo, lo cual conlleva un alto número de falsos positivos (van der Auwera *et al.*, 2013). En cuanto a Freebayes, este es más rápido que GATK y emplea las secuencias reales, sin necesidad de hacer reensamblajes (Yao *et al.*, 2020a). Debido a que no existe una herramienta perfecta a la hora de detectar únicamente los polimorfismos reales, resulta

necesario realizar un paso de filtrado de las variantes genéticas identificadas (**Tabla 2**). Se pueden filtrar por calidad, por frecuencia del alelo menor y por porcentaje de datos faltantes entre todos los genotipos, entre otros (Lee *et al*., 2012; Yao *et al.*, 2020a). Este paso es fundamental, sobre todo en el SWGR, puesto que, debido a la baja cobertura de secuenciación, contra una posición concreta del genoma únicamente se alinearán unas pocas lecturas, por lo que es difícil diferenciar entre una verdadera variante genética, un genotipo heterocigoto o un error de la secuenciación. También se utilizan otros métodos para identificar y descartar los falsos positivos, como el empleo de dos programas bioinformáticos diferentes para llevar a cabo la identificación de los polimorfismos (Yeo *et al.,* 2014; Yao *et al.*, 2020a) o su comparación con un estándar de referencia, es decir, con un conjunto de variantes que se asumen que son correctas, identificadas a coberturas de secuenciación superiores, por ejemplo, de 10X o 20X, o contrastadas en experimentos distintos y/o mediante el empleo de distintas técnicas de genotipado (Malmberg *et al.,* 2018).

Una vez identificadas y filtradas las variantes, se puede llevar a cabo, en función de la finalidad del análisis, un paso de imputación con el fin de incrementar la densidad de marcadores, puesto que, debido a la baja cobertura de secuenciación empleada y a los diversos pasos de filtrado, es posible que el número de variantes genéticas identificadas se vea reducido (Pasaniuc *et al*., 2012; Gorjanc *et al.*, 2017). Específicamente, la imputación consiste en estimar el genotipo de un SNP no identificado y se fundamenta en el desequilibrio de ligamiento entre variantes, es decir, en el hecho de que no se producen recombinaciones entre dos polimorfismos muy cercanos en el genoma (Halperin and Stephan, 2009). Para llevar a cabo este proceso, resulta fundamental contar con un conjunto de datos de referencia, como una población debidamente caracterizada, pues la disponibilidad de este tipo de datos facilita en gran medida la imputación de los genotipos faltantes (Chung *et al.*, 2017).

Tabla 2. Herramientas bioinformáticas disponibles para realizar cada uno de los pasos del análisis bioinformático de los datos generados por la técnica de genotipado SWGR.

Análisis bioinformático	Programas	
Filtrado de lecturas	Fastq-mcf	https://wiki.biouml.org/index.php/Fastq_mcf
	Trimmomatic	https://github.com/usadellab/Trimmomatic
	Fastqc	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Alineamiento	Bowtie2	https://bowtie-bio.sourceforge.net/bowtie2/
	SOAP2	https://bio.tools/soap2
	BWA	https://bio-bwa.sourceforge.net/
	MAQ	https://maq.sourceforge.net/
	Stampy	https://bio.tools/stampy
Marcaje de secuencias duplicadas	Picard	https://broadinstitute.github.io/picard/
Identificación de polimorfismos	SamTools	https://www.htslib.org/
	VarScan	http://dkoboldt.github.io/varscan/
	GATK	https://gatk.broadinstitute.org/
	FreeBayes	https://github.com/freebayes/freebayes
	SOAPsnp	https://sourceforge.net/projects/soapsnp/
Filtrado de polimorfismos	VCFTools	http://vcftools.sourceforge.net/

El SWGR se empleó por primera vez en líneas recombinantes consanguíneas de arroz (RILs) secuenciadas a una cobertura de 0,02X con el objetivo de identificar puntos de recombinación en la población, lo cual mejora la precisión de la detección de QTLs y la eficiencia y la tasa de éxito a la hora

de clonar genes (Huang *et al.*, 2009). También se ha empleado en una población de individuos doble haploides (DHs) de colza con coberturas de secuenciación desde 0,1X hasta 7X y en RILs de garbanzo aplicando coberturas de secuenciación de aproximadamente 7X y 9X, para evaluar la frecuencia y distribución de los eventos recombinantes (Bayer *et al.*, 2015). Por su parte, Clot *et al.* (2023) emplearon una cobertura de secuenciación media de 1,5X en una población de patata para identificar QTLs involucrados en el rendimiento de tubérculos y en la producción de polen. Los datos generados mediante el SWGR también se pueden emplear para caracterizar finamente variantes estructurales, las cuales exhiben una asociación enriquecida con rasgos fenotípicos. Se han realizado análisis de este tipo en maíz, utilizando una cobertura de secuenciación media de 0,3X (Lu *et al.*, 2015), y en trigo, con coberturas de secuenciación desde 0,01X hasta 1X (Adhikari *et al*., 2022). En conclusión, los recursos generados gracias a la información obtenida con el SWGR permiten el avance tanto de la investigación básica como de la mejora genética de muchos cultivos.

1.2.3. Poblaciones experimentales

Las poblaciones experimentales, conocidas como *mapping populations,* son poblaciones segregantes desarrolladas generalmente a partir del cruce de dos o más líneas genéticamente distintas (Singh and Singh, 2015). Concretamente, los mejoradores y los genetistas moleculares las utilizan para mapear genes o QTLs involucrados en el control de caracteres de interés (Goncalves-Vidigal *et al.*, 2020; Liu *et al.*, 2020; Mangino *et al.*, 2022). También ayudan en la determinación de distancias genéticas entre pares de loci y en la identificación de marcadores moleculares ligados a éstos para su uso en la selección asistida por marcadores (Singh and Singh, 2015). Cabe destacar que estas aplicaciones, e incluso la propia caracterización genética de las poblaciones experimentales, se han visto facilitadas y apoyadas por los avances y la reducción de costes que han sufrido las tecnologías de secuenciación (Wetterstrand, n.d.; Yan *et al.*, 2020; Purugganan and Jackson, 2021). Se pueden distinguir dos grupos de poblaciones experimentales en función del número de parentales que se empleen para su desarrollo: poblaciones biparentales o poblaciones multiparentales. Entre las primeras podemos encontrar DH, RIL, líneas casi isogénicas (NIL) y líneas de introgresión (ILs), entre otras; y entre las poblaciones multiparentales destacan las poblaciones multiparentales de intercruzamientos avanzados (MAGIC) y las poblaciones anidadas para mapeo por asociación (NAM) (Singh and Singh, 2015). Por ejemplo, las RILs son líneas homocigotas producidas por la continua autofecundación de plantas F2 (Burr and Burr, 1991), y las ILs son un conjunto de líneas homocigóticas, cada una de las cuales lleva en su genoma un solo segmento cromosómico, diferente entre líneas, del parental donante en el fondo genético del parental recurrente (Eshed and Zamir, 1994). Por otro lado, tanto las poblaciones NAM como las poblaciones MAGIC comprenden varias RILs. Mientras que las líneas de la población NAM comparten el genoma de uno de los genotipos parentales (Kitony *et al.*, 2021), las líneas de la población MAGIC comparten el genoma de los múltiples parentales empleados en los cruzamientos (Mackay and Powell, 2007).

Tradicionalmente, se vienen utilizando las poblaciones biparentales para el mapeo de genes o QTLs (Wang *et al*., 1994; Cui *et al.*, 2011; Liu *et al.*, 2020); sin embargo, las principales desventajas de estas poblaciones son que tan sólo se pueden analizar los alelos derivados de los dos parentales y que la recombinación genética suele ser baja, por lo que la resolución en la identificación de genes o QTLs acostumbra ser limitada (**Tabla 3**) (Bergelson and Roux, 2010). Es por todo ello que en los últimos años el empleo de poblaciones multiparentales se ha visto favorecido, ya que presentan una variación genética y diversidad fenotípica mucho mayor (Yan *et al*., 2020; Mangino *et al.*, 2022). Además, al aumentar el número de recombinaciones, aumenta la precisión y resolución del mapeo de genes o QTLs (**Tabla 3**) (Valdar *et al*., 2006). No obstante, es preciso tener en cuenta que, al ser poblaciones en las que participa un mayor número de líneas parentales, se requieren tamaños poblacionales

superiores para poder identificar todos los eventos recombinantes posibles, y que el tiempo y costo de creación también son mayores que los requeridos en la creación de poblaciones biparentales (**Tabla 3**) (Bernardo, 2021). Adicionalmente, se requiere un mayor número de marcadores moleculares para el análisis del genoma de cada una de las líneas resultantes, puesto que la complejidad estructural de éste se incrementa con el número de líneas parentales empleadas en el cruzamiento inicial (Samantara *et al.*, 2021). Sin embargo, ha dejado de ser una limitación gracias al desarrollo de la tecnología NGS, al establecimiento de genomas de referencia de alta calidad y a los avances que han tenido lugar en el campo de la bioinformática.

Tabla 3. Comparación entre varias poblaciones experimentales: líneas recombinantes consanguíneas (RIL), líneas de introgresión (IL), población multiparental de intercruzamientos avanzados (MAGIC) y población anidada para mapeo por asociación (NAM) (Adaptación de Bohra, 2013).

	RIL	IL	MAGIC	NAM
Número de parentales	Dos	Dos	Múltiple	Múltiple
Generaciones requeridas	Seis - ocho	Seis - ocho	Más de ocho	Más de seis
Múltiples alelos	No	No	Permitido	Permitido
Eventos recombinantes	Limitado	Limitado	Alto	Alto
Resolución del mapeo	Medio	Alto	Alto	Alto
Replicabilidad	Posible	Posible	Posible	Posible
Detección de QTLs menores	Baja	Baja	Alto	Alto

En el año 2000, Mott *et al*. desarrollaron la primera población MAGIC con ratones, mientras que la primera población MAGIC en plantas fue desarrollada por Kover *et al.* (2009) con ejemplares de *Arabidopsis.* El proceso por el cual se obtiene una población de este tipo consta de cuatro etapas principales: 1) la selección de las líneas parentales (normalmente 4, 8, 16 o más), 2) el cruzamiento de dichas líneas para el desarrollo de híbridos, 3) el intercruzamiento avanzado, y 4) las generaciones de autofecundación para fijar los alelos (Samantara *et al.,* 2021). Una de las aproximaciones para conseguir desarrollar esta población multiparental consiste en cruzar los parentales de dos en dos para, a continuación, realizar lo mismo con la descendencia hasta conseguir individuos que reúnan fragmentos del genoma de todas las líneas fundadoras empleadas, que serán los que, finalmente, se someterán a varias generaciones de autofecundación (**Figura 7**).

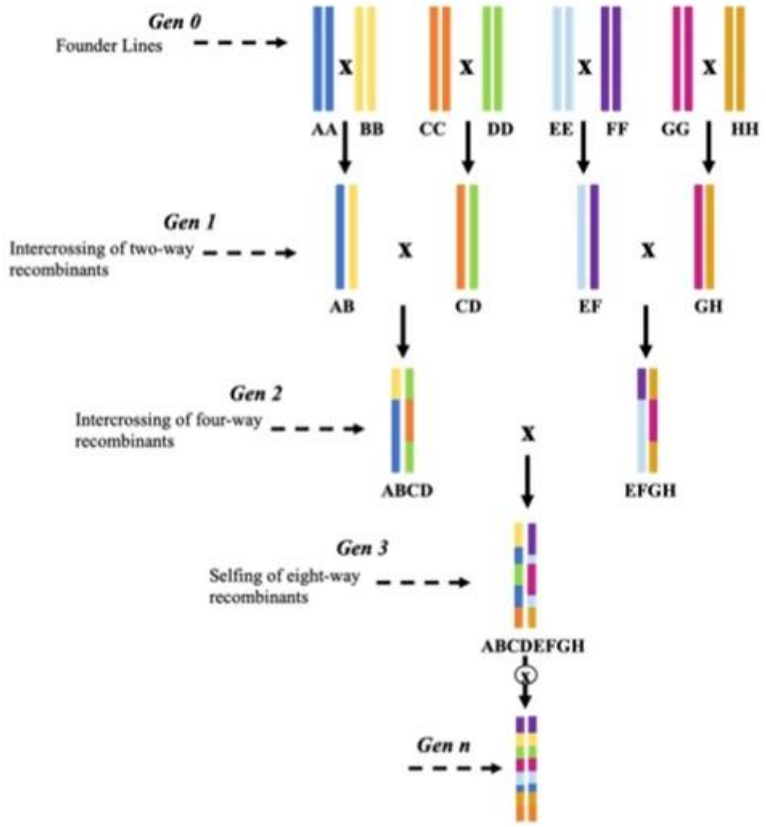


Figura 7. Desarrollo de una población multiparental de intercruzamientos avanzados a partir de ocho parentales (Samantara *et al.*, 2021).

El éxito de esta población reside en que, además de las ventajas comentadas previamente, es inmortal, pues las líneas homocigotas finales se pueden reproducir año tras año, manteniendo la misma composición genómica, y en que carece de estructura poblacional (Singh and Singh, 2015). Hasta la fecha, se han empleado tanto para llevar a cabo estudios de ligamiento y de asociación, como para la construcción de mapas de alta densidad (Singh and Singh, 2015; Gardner *et al.*, 2016). Y, gracias a los resultados que se han ido obteniendo, cada vez son más los cultivos en los que se están desarrollando y empleando poblaciones MAGIC, tales como tomate (Pascual *et al*., 2015), algodón (Islam *et al.*, 2016), fresa (Wada *et al.*, 2017), arroz (Bossa-Castro *et al.*, 2018), cebada (Mathew *et al.*, 2018), sorgo (Ongom and Ejeta, 2018), maíz (Jiménez-Galindo *et al.,* 2019), judía (Diaz *et al.,* 2020) y berenjena (Mangino *et al*., 2022), entre otros.

1.3. AVANCES GENÓMICOS EN BERENJENA

A pesar de que la berenjena ocupa la sexta posición entre los cultivos hortícolas más cultivados a nivel mundial (FAOSTAT, 2022), no ha sido hasta la última década cuando se ha iniciado el desarrollo de diferentes recursos genómicos y se ha acortado la brecha con otros cultivos importantes del género *Solanum*, tales como el tomate y la patata (Gramazio *et al.*, 2018). En 2011 se publicó el primer borrador del genoma de patata (The Potato Genome Sequencing Consortium, 2011) y al año siguiente el de tomate (The Tomato Genome Consortium, 2012). Sin embargo, no fue hasta 2014 cuando se publicó el primer borrador del genoma de berenjena, que se encuentra fragmentado en 33.873 *scaffolds* y que cubre aproximadamente el 69% del tamaño del genoma (Hirakawa *et al.*, 2014). Este primer borrador no presenta una alta calidad, puesto que se encuentra muy fragmentado y las regiones sin ensamblar constituyen el 4,75% de la longitud total (Hirakawa *et al.,* 2014). El hecho de no disponer de un genoma de referencia de alta calidad en berenjena ha atrasado los avances

genéticos en cuanto a la identificación de QTLs o genes, al desarrollo de poblaciones experimentales y a los estudios de resecuenciación. De hecho, para el mapeo de algunos QTLs de berenjena se ha empleado el genoma de referencia disponible de tomate (Cericola *et al*., 2014; Portis *et al*., 2015).

Debido a la necesidad de disponer de un genoma de referencia menos fragmentado y correctamente anotado, Barchi *et al.* (2019a) publicaron la secuencia a nivel de cromosoma de la línea endogámica 67/3 (v3.0). Este genoma, comparado con el de Hirakawa *et al.* (2014), presenta una contigüidad y cobertura mayores. Está formado por 10.383 *scaffolds*, el 77,5% de los cuales se encuentran anclados (Barchi *et al*., 2019a). Sin embargo, la plataforma de secuenciación utilizada fue Illumina, una plataforma de segunda generación de lecturas cortas, las cuales dificultan en parte la realización de un ensamblaje de alta calidad (Pareek *et al*., 2011). Un año después, se publicó un genoma de alta calidad también a nivel cromosómico basado en el ensamblaje de lecturas obtenidas con plataformas de segunda y tercera generación (Wei *et al.*, 2020). Lo conforman un total de 2.263 *scaffolds*, estando el 92,7% anclados. Poco después, se ensambló el genoma de otra variedad cultivada de berenjena, que presenta ciertas ventajas sobre el genoma 67/3 v3.0, como el menor número de *scaffolds*, siendo 319 en total, y el mayor tamaño de los *contigs* (N50 = 5,3 Mb vs. 16,7 kb) (Li *et al*., 2021). Por último, recientemente ha sido publicada la versión 4.0 del genoma de la línea 67/3 (Barchi *et al*., 2021). Esta versión presenta un mayor número de *scaffolds* asignados a los cromosomas correspondientes (95,6%), que, además, en conjunto presentan el mayor tamaño de todos los empleados hasta la fecha (N50 = 92,1 Mb) (Barchi *et al*., 2021). Cabe destacar que dicha versión ha sido actualizada por los mismos autores, los cuales han realizado la reorientación de la secuencia de alguno de los cromosomas (67/3 v4.1) (https://solgenomics.net/).

En cuanto a las poblaciones experimentales desarrolladas en berenjena, la primera población de RILs fue desarrollada en 2013 por Lebeau *et al*. con el objetivo de determinar el control genético de la resistencia a *Ralstonia solanacearum*. Por otro lado, en 2017 se desarrolló la primera población de ILs empleando la especie *S. incanum* como parental donante (Gramazio *et al.*, 2017; Mangino *et al*., 2020). Este proceso se vio favorecido por la disponibilidad del primer genoma de referencia, ya que facilitó la realización de un genotipado por secuenciación. Finalmente, en 2022 se publicó la primera población MAGIC de berenjena derivada del cruce de siete accesiones de *S. melongena* y de la especie silvestre *S. incanum* (Mangino *et al.*, 2022). En este caso, se empleó la tecnología SPET (Single Primer Enrichment Technology) para realizar el genotipado de las casi 350 líneas (Barchi *et al.*, 2019b). Esta técnica se basa en el diseño de sondas que hibridan en las secuencias flanqueantes de cada uno de los SNPs, por lo que se tiene que disponer de un panel de SNPs previamente identificados. El conjunto de SNPs para el desarrollo del panel SPET se obtuvo resecuenciando los parentales de la población MAGIC anteriormente nombrada (Gramazio *et al.*, 2019). Hoy en día, cada vez son más las poblaciones experimentales desarrolladas para ampliar el conocimiento sobre el control genético de los caracteres que manifiestan las diferentes accesiones de berenjena (Barchi *et al.*, 2018; Salgon *et al.*, 2018; Toppino *et al.,* 2020), así como los análisis de resecuenciación gracias a la disponibilidad de un genoma de referencia de alta calidad (Liu *et al.*, 2019; Qian *et al*., 2021; Guan *et al*., 2022).

2. OBJETIVOS

- Objetivo 1. Evaluar el efecto de dos variables claves, la cobertura de secuenciación y la profundidad mínima de mapeo, en la identificación y caracterización de polimorfismos genéticos mediante el SWGR.
- Objetivo 2. Determinar la factibilidad de implementar el SWGR como una herramienta práctica y de uso común para el genotipado rutinario en berenjena.

3. MATERIALES Y MÉTODOS

3.1. SECUENCIACIÓN Y GENERACIÓN *IN SILICO* DE LOS ARCHIVOS FASTQ

El material vegetal utilizado en este estudio corresponde a la accesión MM1597, uno de los parentales empleados en el desarrollo de la primera población MAGIC de berenjena (Gramazio *et al*., 2019). El ADN genómico fue extraído a partir de aproximadamente 100 mg de hoja joven mediante el protocolo de extracción SILEX (Vilanova *et al.*, 2020). La calidad del ADN se determinó mediante electroforesis en gel de agarosa al 1% y midiendo los ratios de absorbancia 260/230 y 260/280 con el espectrofotómetro Nanodrop ND-1000 (Nanodrop Technologies, Wilmington, DE, EEUU), mientras que su concentración se cuantificó utilizando el fluorímetro Qubit (Thermo Fisher Scientific, Waltham, MA, EEUU). Posteriormente, el ADN fue enviado a la compañía BGI Genomics (Yantian, Shenzhen, China) para el desarrollo de una librería genómica de extremos pareados de 150 pb (paired ends, PE150), seguido de su secuenciación a una cobertura media de 5X (12,6 Gb) utilizando la plataforma DNBseq. Las lecturas crudas (raw reads) fueron filtradas por BGI Genomics con el software SOAPnuke (-n 0.001 -l 10 -q 0.4 --adaMR 0.25 --ada_trim) (Chen *et al.*, 2018a) para eliminar: (1) los adaptadores empleados durante la secuenciación, (2) las lecturas con una calidad de base inferior a 10 en más del 40% de su longitud total y, (3) las lecturas con un número de bases desconocidas superior al 0,1% de su longitud (**Figura 8**). El sistema de valoración de la calidad de las lecturas remanentes se estableció en una calidad media de Phred por encima de 33. Estos parámetros, al igual que otros, como el contenido en guaninas y citosinas (GC), la distribución de secuencias por longitud, y los niveles de secuencias duplicadas y de regiones sobrerrepresentadas, fueron visualizados en el informe generado por la herramienta de control de calidad FastQC (Andrews, 2016). Por otro lado, en un proyecto anterior la misma muestra fue enviada al Centro de Genómica y Biología Computacional de Duke (Durham, NC, Estados Unidos) para su secuenciación a una cobertura media de 20X (50,4 Gb) mediante un secuenciador Illumina HiSeq 4000 (Gramazio *et al*., 2019). Las lecturas resultantes fueron sometidas a un paso de filtrado con la herramienta fastq-mcf (versión 1.04.676, https://github.com/ExpressionAnalysis/ea-utils/blob/wiki/FastqMcf.md). Este proceso tuvo como objetivo eliminar los adaptadores, las lecturas de baja calidad (umbral de calidad: -q 30) y las lecturas con una longitud inferior a 50 pb (-l 50) (**Figura 8**). A continuación, se generó también un informe con FastQC (Andrews, 2016) para obtener una descripción general de la calidad de las lecturas procesadas.

Los archivos fastq resultado de la resecuenciación a una cobertura media de 5X se emplearon para generar *in silico* cuatro nuevos archivos fastq correspondientes a las coberturas de secuenciación 1X, 2X, 3X y 4X, respectivamente, con la herramienta seqtk (versión 1.3-r106, https://github.com/lh3/seqtk) (**Lista 1**). Teniendo en cuenta que el tamaño del genoma de berenjena es de aproximadamente 1,2 Gb y que el tamaño de cada uno de los dos archivos fastq a 5X fue de 6,3 Gb, para el subconjunto 1X se tomó una muestra aleatoria del 20% de pares de lecturas del archivo original (que corresponde aproximadamente a 1,26 Gb), del 40% para el subconjunto 2X (2,52 Gb), del 60% para el subconjunto 3X (3,78 Gb) y del 80% para el subconjunto 4X (5,04 Gb) con el propósito de conseguir las coberturas deseadas, respectivamente. Además, por cada pareja de extremos pareados se utilizó el mismo valor del filtro *-s* para no perder el emparejamiento entre los mismos (**Lista 1**). Para cada cobertura de 1X, 2X, 3X y 4X se generaron cinco réplicas con el fin de minimizar el error asociado a la variabilidad aleatoria y de aumentar la robustez de los resultados obtenidos (**Figura 8**).

Lista 1. Ejemplo de la generación *in silico* de archivos fastq a una cobertura de secuenciación de 1X mediante el uso de la herramienta seqtk.

```
seqtk sample -s001 V350056001_L01_SOLpogrR146511-657_2.fq 0.2 > Smel_A_PC22_R2_1X1.fq

seqtk sample -s001 V350056001_L01_SOLpogrR146511-657_1.fq 0.2 > Smel_A_PC22_R1_1X1.fq
```

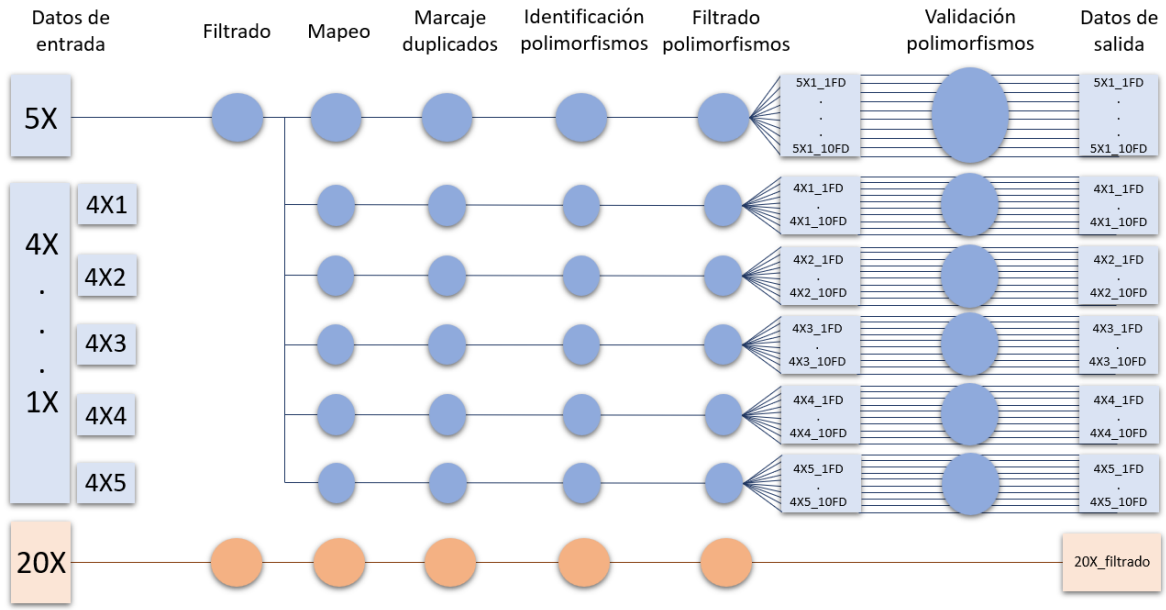


Figura 8. Esquema del análisis bioinformático realizado sobre los datos de resecuenciación a 5X, 20X y los archivos generados *in silico* correspondientes a coberturas de secuenciación de 1-4X. Los datos de entrada (5X y 20X) se sometieron a un paso de filtrado de las lecturas por calidad y longitud. A continuación, se generó *in silico* conjuntos de lecturas correspondientes a coberturas de secuenciación de 1-4X a partir de la muestra de resecuenciación a 5X; concretamente, cinco réplicas por cada cobertura de secuenciación. Todas las muestras resultantes se alinearon contra la versión 4.1 del genoma de referencia de berenjena 67/3. Las lecturas duplicadas fueron marcadas de forma previa a la identificación de los polimorfismos. Tras este paso, los polimorfismos identificados se filtraron con diferentes profundidades de mapeo mínimas (1-10FD), excepto en el caso de la muestra de resecuenciación a 20X, cuyos polimorfismos identificados se sometieron a un filtrado más restrictivo para conformar el estándar de referencia (profundidad de mapeo máxima y mínima de 40FD y 10FD, respectivamente, fracción mínima de alelo alternativo de 0,3, eliminación de los polimorfismos homocigotos para el alelo de referencia y los localizados en el cromosoma 0). Finalmente, los polimorfismos filtrados de las muestras de resecuenciación a 1-5X se validaron mediante su comparación con los polimorfismos resultantes que conformaron el estándar de referencia.

3.2. ALINEAMIENTO CONTRA EL GENOMA DE REFERENCIA

El mapeo de las lecturas procesadas contra la versión 4.1 del genoma de referencia de berenjena 67/3 (Barchi *et al*., 2021; https://solgenomics.net/) se llevó a cabo con la herramienta Bowtie2 (versión 2.4.4), una herramienta eficiente para alinear lecturas de secuenciación frente a secuencias de referencia (**Lista 2.b**) (Langmead and Salzberg, 2012). Se utilizaron los parámetros establecidos por defecto y el genoma previamente indexado con la extensión bowtie2-build (**Lista 2.a**). Para calcular los estadísticos de mapeo se utilizó la aplicación QualiMap (versión 2.2.1) (**Lista 2.c**), que, además, facilita un archivo con gráficos acerca de diferentes parámetros, como la distribución de la calidad del mapeo (MAPQ) a lo largo del genoma de referencia (García-Alcalde *et al*., 2012). El porcentaje de bases del genoma de referencia alineadas con al menos una lectura, conocido como cobertura del genoma, se calculó con el programa samtools coverage, incluido dentro del conjunto de herramientas SAMtools (versión 1.13; Li *et al.*, 2009) (**Lista 2.d**). Por su parte, se utilizó la función genomecov del paquete de herramientas Bedtools (versión 2.30.0; https://bedtools.readthedocs.io/) (**Lista 2.e**) para conocer el porcentaje de bases para cada nivel de profundidad de mapeo (i.e., número de lecturas alineadas contra una posición específica del genoma de referencia) y la profundidad de mapeo máxima para cada cobertura de secuenciación. Finalmente, se determinó la distribución de la profundidad de mapeo a lo largo del primer cromosoma para observar las diferencias entre las distintas coberturas de secuenciación, que, en principio, serán extrapolables al resto de cromosomas. Para ello, se empleó bamCoverage (versión 3.5.1) y un tamaño de ventana de 10 Kpb (Ramírez *et al*., 2014) (**Lista 2.f**).

Lista 2. Conjunto de comandos utilizados en el alineamiento de las lecturas contra el genoma de referencia y en el análisis posterior. **a)** Indexado del genoma de referencia. **b)** Alineamiento de las lecturas contra el genoma de referencia. **c)** Cálculo de estadísticos y visualización de gráficos de diferentes parámetros del mapeo. **d)** Cálculo de la cobertura del genoma de referencia (i.e. porcentaje del genoma soportado por al menos una lectura). **e)** Cálculo del porcentaje de bases para cada nivel de profundidad de mapeo y la profundidad de mapeo máxima. **f)** Distribución de la profundidad de mapeo a lo largo del primer cromosoma (tamaño de ventana: 10 Kpb).

```
a   bowtie2-build Eggplant_V4.1.fa eggplant_genome_V4.1_index

b   bowtie2 -p30 -x eggplant_genome_V4.1_index -1 Smel_A_PC22_R1_1X1.fq.gz -2
    Smel_A_PC22_R2_1X1.fq.gz | samtools view -bS - | samtools sort - | samtools
    addreplacerg -r ID:Smel_A -r SM:Smel_A___MM1597
    - -o Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.bam

c   qualimap bamqc -bam Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.bam -outfile
    Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_qualimap.pdf

d   samtools coverage Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.bam >
    1X1_samtools_coverage.txt

e   bedtools genomecov -ibam Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.bam >
    Smel_A_mapped_V4.1_ID_PC22_1X1_cov_genome.txt

f   bamCoverage --bam Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.bam -of bedgraph -o
    Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_coverage.bedgraph -bs 10000 -r chr1
```

Los dos últimos parámetros, tanto el porcentaje de bases para cada nivel de profundidad de mapeo como la distribución de las distintas profundidades de mapeo a lo largo del primer cromosoma, se representaron en gráficos con la función plot (versión 3.6.2; https://r-coder.com/plot-r/) (**Lista 3**) de R (versión 4.2.2; R Core Team, 2022).

Lista 3. Comandos utilizados en la representación en RStudio. **a)** Porcentaje de bases mapeadas a diferentes profundidades de mapeo. **b)** Distribución de las diferentes profundidades de mapeo a lo largo del primer cromosoma.

```
a   plot(Smel_A_mapped_V4.1_ID_PC22_1X1_cov_genome$V2,
    Smel_A_mapped_V4.1_ID_PC22_1X1_cov_genome$V5, xlab="Profundidad de mapeo",
    ylab="Porcentaje de bases", xlim=c(0,10), type="o", pch=19, cex=1.5)

b   plot(Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_coverage$V2/1000000,
    Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_coverage$V4*150/10000, xlab="Posición (Mpb)",
    ylab="Profundidad de mapeo", ylim=c(0,10), type="h")
```

En el primer escenario, la variable independiente corresponde a la segunda columna del archivo de texto generado con la función genomecov, que representa la profundidad de mapeo, y la variable dependiente al porcentaje de bases mapeadas para cada nivel de profundidad, encontrándose estos datos en la quinta columna del archivo de texto. En el caso del archivo en formato bedgraph obtenido con la herramienta bamCoverage, la variable independiente se define como la posición a lo largo del cromosoma (segunda columna) y la variable dependiente como la profundidad de mapeo media por ventana, que se calculó a partir del número total de lecturas mapeadas por cada ventana (cuarta columna), de la longitud de las lecturas y de la ventana de actuación (**Ecuación 1**) (Lander and Waterman, 1988):

$$Profundidad\ de\ mapeo\ media\ = \frac{longitud\ de\ las\ lecturas\ x\ número\ de\ lecturas}{longitud\ de\ la\ ventana} \qquad (\boldsymbol{Ecuación\ 1})$$

3.3. IDENTIFICACIÓN Y FILTRADO DE VARIANTES GENÉTICAS

Como paso previo a la identificación de polimorfismos en el genoma, resultó necesario marcar las lecturas duplicadas originarias de un solo fragmento de ADN, para lo cual se empleó la herramienta MarkDuplicates del software Picard (versión 1.119; https://broadinstitute.github.io/picard/) (**Lista 4.a**). La identificación de variantes se llevó a cabo con el software Freebayes basado en métodos probabilísticos (versión v1.3.6; Garrison and Marth, 2012). Se emplearon parámetros de calidad mínima de mapeo (-m 20) y de base (-q 20) (**Lista 4.b**) con el fin de excluir las lecturas que tuvieran una calidad de mapeo inferior y los alelos que tuvieran una calidad de base de apoyo inferior, respectivamente. A continuación, los polimorfismos identificados y listados en los archivos VCF generados fueron sometidos a un proceso de filtrado utilizando diferentes umbrales de profundidad mínima de mapeo. La herramienta utilizada fue VCFtools (versión 0.1.16; Danecek *et al*., 2011) con la función—min-meanDP (**Lista 4.c**). Las profundidades de mapeo mínimas evaluadas variaron desde 1X hasta 10X para las coberturas de secuenciación de 1X a 5X, mientras que, para el caso de la cobertura de secuenciación de 20X, se aplicaron filtros de profundidad de mapeo mínima de 1X, 10X y 20X (**Figura 8**). Para los conjuntos de datos en los que las secuencias duplicadas no fueron marcadas, estos filtros se designan como 1F – 10F, mientras que para los conjuntos de datos en los que dichas secuencias sí que fueron marcadas, se designan como 1FD – 10FD. Finalmente, se recopiló información acerca del número de cada tipo de polimorfismo y del número de genotipos homocigotos y heterocigotos con la herramienta vcf-stats (https://pwwang.github.io/vcfstats/) (**Lista 4.d**).

El archivo de resecuenciación a 20X se utilizó como estándar de referencia en pasos posteriores, por lo que los polimorfismos identificados en este conjunto tuvieron que ser lo más fiables posibles. Es por ello que, además de los filtros comentados anteriormente, se emplearon otros cuatro para que el filtrado fuera más restrictivo: (I) una profundidad de mapeo máxima de 40FD para evitar las regiones sobrerrepresentadas; (II) una fracción mínima del alelo alternativo de 0,3 con el fin de distinguir las variantes heterocigotas (**Lista 4.e**); (III) la eliminación de los genotipos homocigotos para el alelo de referencia (0/0); y (IV) la eliminación de las variantes localizadas en el cromosoma 0, constituido por las secuencias nucleotídicas que no pudieron ser asignadas a un cromosoma específico durante el proceso de ensamblaje (**Lista 4.f**).

Lista 4. Conjunto de comandos empleados en la identificación de variantes tanto en el conjunto de muestras como en el estándar de referencia y en el posterior filtrado. **a)** Marcado de lecturas duplicadas en el conjunto de datos. **b)** Identificación de polimorfismos en los conjuntos de datos de resecuenciación de 1X a 5X localizados en loci con una cobertura de mapeo y de base superior a 20 y una cobertura mínima superior a 1. **c)** Filtrado de los polimorfismos de los conjuntos de datos de resecuenciación de 1X a 5X con coberturas de profundidad de mapeo mínimas de 1-10FD. **d)** Recopilación de información acerca del tipo de polimorfismo y del genotipo identificado. **e)** Identificación de polimorfismos en los conjuntos de datos de resecuenciación a 20X localizados en loci con una cobertura de mapeo y de base superior a 20, una cobertura mínima superior a 1/10/20X y máxima inferior a 40X y una fracción mínima de alelo alternativo de 0,3. **f)** Eliminación de los polimorfismos homocigotos para el alelo de referencia (0/0) y de los presentes en el cromosoma 0 de los datos de resecuenciación a 20X.

a
```
java -jar MarkDuplicates.jar I=Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.bam
O=Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.dedup.bam
M=Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_duplicateinfo.txt
```

b
```
freebayes -f Eggplant_V4.1.fa -b Smel_A_mapped_V4.1_ID_PC22_1X1_sorted.dedup.bam -v
Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes.vcf -m 20 -q 20 --min-
coverage 1
```

c
```
vcftools --vcf Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes.vcf --min-
meanDP 2 --recode --recode-INFO-all --out
Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes_meanDP2.vcf
```

```
d  vcf-stats -p out/
   Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes_meanDP2.vcf

e  freebayes -f Eggplant_V4.1.fa -b Smel_A_Q30L50_mapped_V4.1_ID_sorted.dedup.bam -v
   Smel_A_Q30L50_mapped_V4.1_ID_PC22_lim1X40X_alt30_sorted_raw_variants_freebayes.vcf -m
   20 -q 20 --min-coverage 1 --limit-coverage 40 --min-alternate-fraction 0.3

f  grep -v "0/0"
   Smel_A_Q30L50_mapped_V4.1_ID_PC22_lim1X40X_alt30_sorted_raw_variants_freebayes.recode
   .vcf | grep -v "^0" >
   Smel_A_Q30L50_mapped_V4.1_ID_PC22_lim1X40X_alt30_sorted_raw_variants_freebayes_no_00_
   no_chr0.vcf
```

3.4. COMPARACIÓN CON EL ESTÁNDAR DE REFERENCIA

Se realizó una comparación entre los conjuntos de polimorfismos obtenidos a cada combinación de cobertura de secuenciación y profundidad mínima de mapeo, y los datos del estándar de referencia. Esto permitió determinar la precisión y la sensibilidad de la técnica utilizada. La precisión se refiere a la capacidad para descartar de manera correcta los polimorfismos falsos o errores de identificación. Se calcula como la proporción de polimorfismos verdaderos identificados correctamente en la muestra en comparación con el total de polimorfismos identificados (**Ecuación 2**). Por otro lado, la sensibilidad se refiere a la capacidad para detectar correctamente los polimorfismos existentes en el genoma. Este parámetro se calcula como la proporción de polimorfismos verdaderos identificados correctamente en la muestra en comparación con el total de polimorfismos verdaderos presentes en el genoma (**Ecuación 3**).

$$Precisión\ (\%) = \frac{variaciones\ en\ la\ muestra\ compartidas\ con\ el\ estándar}{número\ total\ de\ variaciones\ en\ la\ muestra}\ x\ 100 \qquad (\boldsymbol{Ecuación\ 2})$$

$$Sensibilidad\ (\%) = \frac{variaciones\ en\ la\ muestra\ compartidas\ con\ el\ estándar}{número\ total\ de\ variaciones\ en\ la\ estándar}\ x\ 100 \qquad (\boldsymbol{Ecuación\ 3})$$

Se utilizó el paquete de herramientas BCFtools (versión 1.13; https://samtools.github.io/bcftools/bcftools.html) con el comando *isec* para manipular los archivos VCF generados en el paso de identificación de los SNPs (**Lista 5.a; Lista 5.b y Lista 5.c**). Por defecto, con este comando se exportan únicamente los polimorfismos con idéntico genotipo entre la muestra y el estándar de referencia. Se generaron cuatro documentos por cada grupo de datos: (I) los SNPs exclusivos de la muestra; (II) los SNPs exclusivos del estándar de referencia; (III) los SNPs de la muestra compartidos por ambos; y (IV) los SNPs del estándar compartidos por ambos. Finalmente, se representó el número de variantes compartidas y el número de variantes únicas del estándar de referencia y de las muestras en diagramas de Venn de dos entradas. Se realizó para todas las combinaciones de cobertura de secuenciación y profundidad de mapeo mínima. La función empleada fue draw.pairwise.venn de la librería VennDiagram de R (versión 1.7.3; https://rdocumentation.org/packages/VennDiagram/versions/1.7.3) (**Lista 5.d**).

Lista 5. Conjunto de comandos empleados en la comparación de los polimorfismos presentes en la muestra con el estándar de referencia. **a)** Compresión del archivo vcf tras su filtrado a una determinada profundidad de mapeo mínima. **b)** Indexado del archivo resultante de la compresión. **c)** Comparación de las diferentes muestras con el estándar de referencia. **d)** Representación en RStudio de los polimorfismos únicos y comunes entre las muestras y el estándar de referencia en diagramas de Venn de dos entradas.

```
a  bgzip Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes_meanDP2.recode.vcf
```

```
b  bcftools index
   Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes_meanDP2.recode.vcf.gz

c  bcftools isec -p dir
   Smel_A_mapped_V4.1_ID_PC22_1X1_sorted_raw_variants_freebayes_meanDP2.recode.vcf.gz
   Smel_A_Q30L50_mapped_V4.1_ID_PC22_lim1X40X_alt30_sorted_raw_variants_freebayes.recode
   .vcf.gz

d  draw.pairwise.venn(227347,1784060,202160,    lwd    =    0.08,    fill    =    c("black",
   "grey"),category = c("1X", "20X"), cat.pos = c(270,45), cat.dist = 0.08, cat.cex = 0.7,
   ext.text = TRUE, ext.percent = rep(0.4, 3), ext.pos = 10, ext.dist = rep(0.15, 2), cex
   = 0.8)
```

4. RESULTADOS

4.1. SECUENCIACIÓN Y ANÁLISIS DE LAS LECTURAS OBTENIDAS

La secuenciación de la accesión MM1597 a 5X, tanto en sentido directo como inverso, y el posterior paso de filtrado de las lecturas obtenidas generó un total de 39.570.264 lecturas y aproximadamente 6 miles de millones de bases limpias. La calidad de estos datos, medida como el porcentaje de bases con una calidad de *Phred* superior a 20, que representa una tasa de error de 1 en 100 (Illumina, 2011), fue del 96,31%. Es decir, tan solo el 4% de los datos generados presentaron una calidad de *Phred* inferior a 20 (**Figura 9**). Por otro lado, la secuenciación del mismo genotipo a 20X generó un total de 170.761.552 lecturas y aproximadamente 25,5 miles de millones de bases limpias. En este caso, el 100% de las posiciones de las lecturas presentaron una calidad superior a 20 en la escala de *Phred* (**Figura 10**). Además, como se puede ver en las cuatro representaciones de la calidad de *Phred* a lo largo de las lecturas (**Figuras 9 y 10**), la calidad media en ambos casos fue superior a 33.

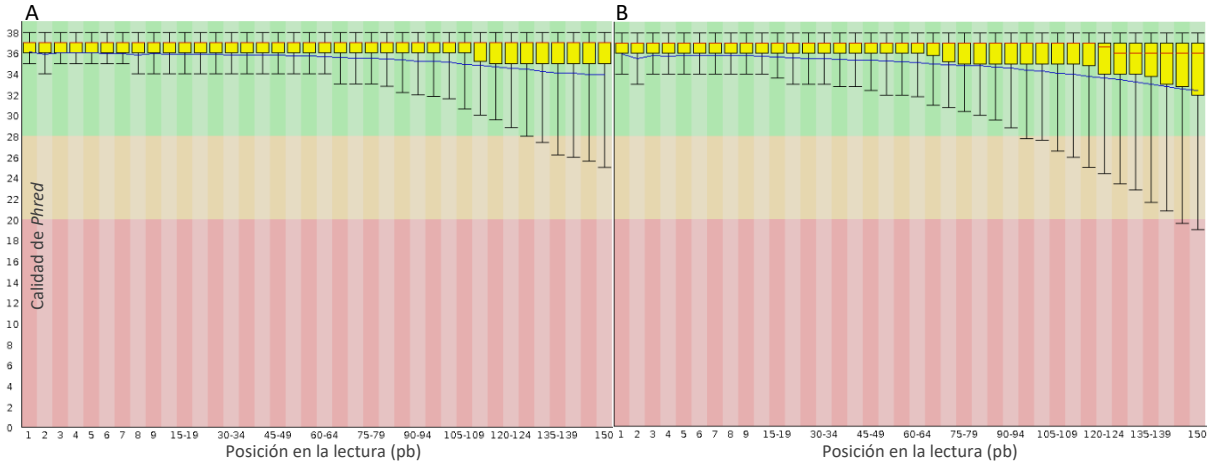


Figura 9. Distribución de la calidad de *Phred* a lo largo de las lecturas obtenidas de la secuenciación a una cobertura media de 5X. **A.** Lecturas de la secuenciación directa o *forward*. **B.** Lecturas de la secuenciación inversa o *reverse*. El color del gráfico denota qué puntuaciones se consideran altas (verde), medias (amarillo) y bajas (rojo) en calidad. La figura ha sido generada vía FastQC (Andrews, 2016).

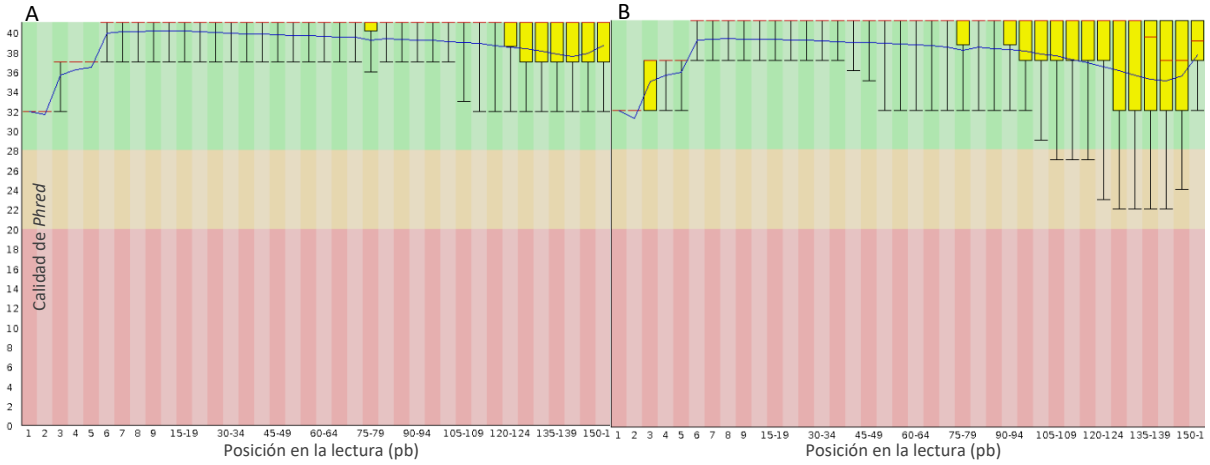


Figura 10. Distribución de la calidad de *Phred* a lo largo de las lecturas obtenidas de la secuenciación a una cobertura media de 20X. **A.** Lecturas de la secuenciación directa o *forward*. **B.** Lecturas de la secuenciación inversa o *reverse*. El color del gráfico denota qué puntuaciones se consideran altas (verde), medias (amarillo) y bajas (rojo) en calidad. La figura ha sido generada vía FastQC (Andrews, 2016).

Otro de los parámetros a tener en cuenta, y representado en el informe generado con FASTQC, es el contenido en GC. En una librería aleatoria, se espera ver una distribución más o menos normal del contenido de GC, donde el pico central corresponde a la media de GC del genoma subyacente. Esta medida resulta útil para conocer si la muestra se encuentra contaminada con ADN de otro organismo o si existen secuencias sobrerrepresentadas, ya que, en este caso, se generaría una distribución con un pico más ancho que el esperado o con picos secundarios, respectivamente (Nederbragt *et al*., 2010; Bérénice *et al.*, 2022). En el caso de la resecuenciación a 5X, el contenido en GC fue del 36,67% (**Figura 11.A**) y en el caso de la secuenciación a 20X, de 36,5% (**Figura 11.B**).

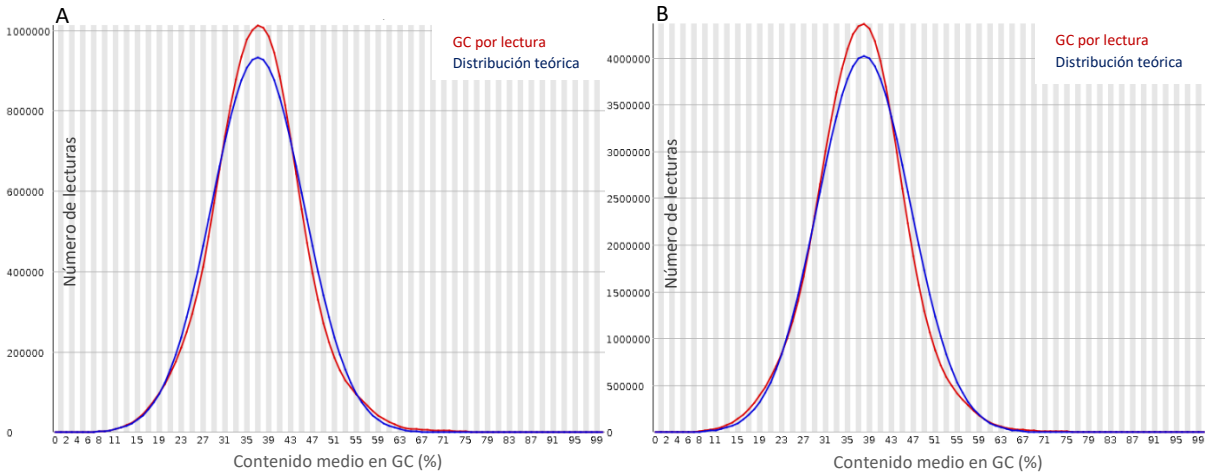


Figura 11. Contenido de guaninas y citosinas en todas las lecturas obtenidas de la secuenciación. **A.** Cobertura media de 5X. **B.** Cobertura media de 20X. La línea azul corresponde con la distribución teórica calculada a partir de los datos observados y la línea roja con la distribución real. La misma distribución se obtuvo tanto para las lecturas resultantes de la secuenciación directa como de la secuenciación inversa. La figura ha sido generada vía FastQC (Andrews, 2016).

Con respecto a la distribución de la longitud de las lecturas, tanto en el caso de la resecuenciación a 5X como a 20X, la mayor parte presentó una longitud de 150 pb, como se muestra en la **Figura 12**. El nivel de secuencias duplicadas, entendido como el porcentaje de lecturas de una secuencia dada presente un número determinado de veces, fue bajo. En el caso de los datos obtenidos mediante la resecuenciación a 5X, si se eliminasen los duplicados, permanecería el 89-90% del total de lecturas en el conjunto de datos. Un porcentaje muy limitado de secuencias, alrededor del 5%, se encontraron duplicadas entre 10 y 50 veces (**Figura 13**). Por otro lado, el 18,50% de lecturas del conjunto obtenido mediante la resecuenciación a 20X fueron lecturas duplicadas. Es decir, permanecería el 81,50% de las lecturas si se eliminaran las duplicadas. Cabe destacar que, en comparación con la resecuenciación a 5X, el número de lecturas duplicadas en un rango de 10 a 50 veces fue ligeramente mayor, constituyendo alrededor del 8% del total de lecturas (**Figura 13**). Finalmente, las secuencias sobrerrepresentadas son aquellas que representan más del 0,1% del total de secuencias (Wright *et al*., 2017; Bérénice *et al*., 2022). Este módulo ayuda en la identificación de contaminantes, como las lecturas procedentes del vector o de los adaptadores, pero no fueron observadas en ningún caso.

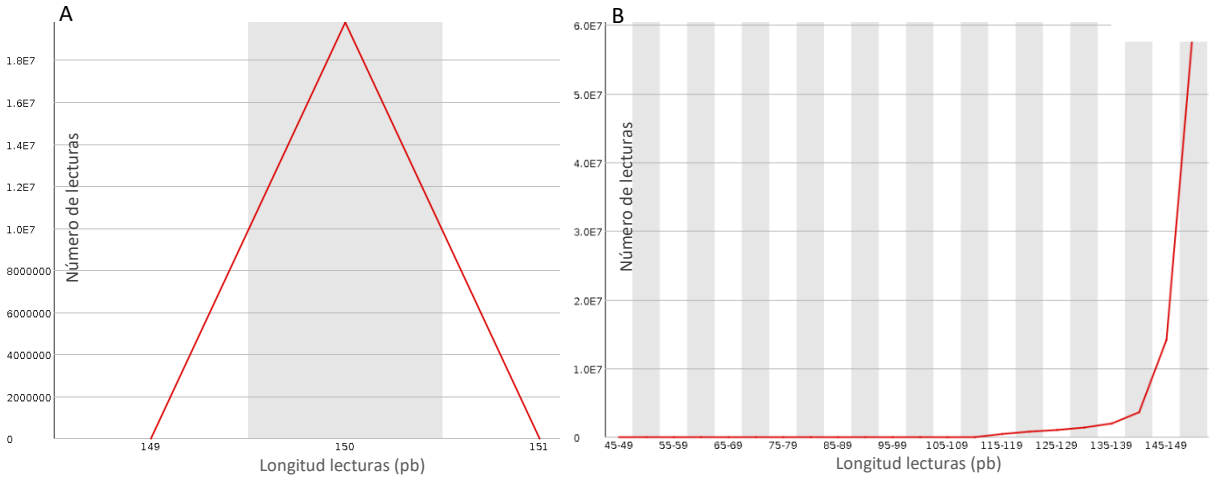


Figura 12. Distribución de la longitud de las lecturas obtenidas tras la secuenciación. **A.** Cobertura media de 5X. **B.** Cobertura media de 20X. La misma distribución se obtuvo tanto para las lecturas resultantes de la secuenciación directa como de la secuenciación inversa. La figura ha sido generada vía FastQC (Andrews, 2016).

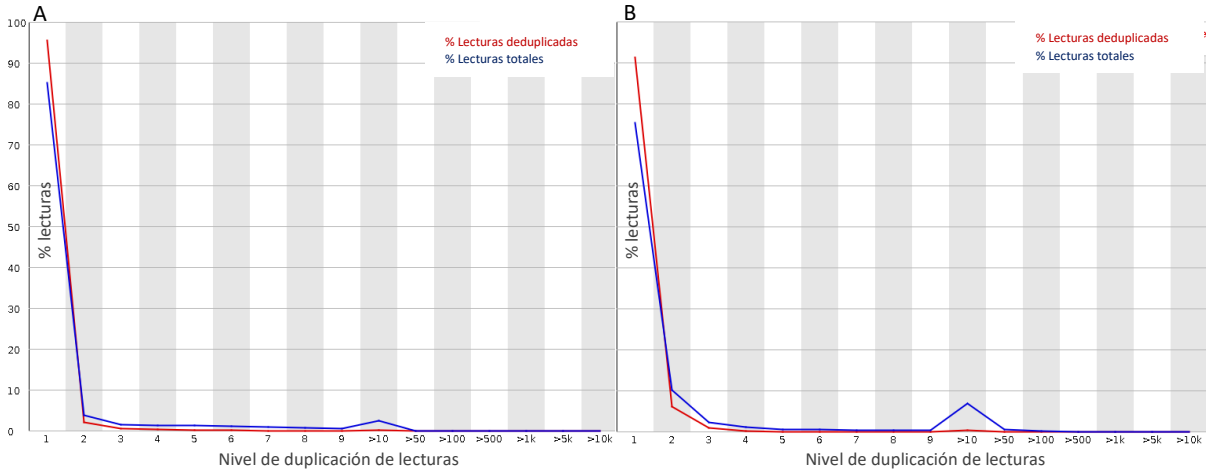


Figura 13. Distribución de lecturas en base a su nivel de duplicación. **A.** Cobertura media de 5X. **B.** Cobertura media de 20X. La línea azul representa el porcentaje de lecturas totales obtenidas tras la secuenciación y la línea roja representa el porcentaje de lecturas remanentes si se eliminasen las secuencias duplicadas. La misma distribución se obtuvo tanto para las lecturas resultantes de la secuenciación directa como de la secuenciación inversa. La figura ha sido generada vía FastQC (Andrews, 2016).

Dado que los archivos fastq correspondientes a las coberturas 1X, 2X, 3X y 4X se generaron *in silico* a partir de los datos de la resecuenciación del genoma completo a 5X, los parámetros aportados en el informe generado por FASTQC fueron idénticos a los presentados anteriormente para la cobertura de secuenciación 5X. La única diferencia notable se encontró en el nivel de secuencias duplicadas, que desciende conforme la cobertura de secuenciación disminuye. Para la cobertura de 1X, el porcentaje de lecturas remanentes en el conjunto de datos sería del 94-95% del total de lecturas si se eliminaran las duplicadas.

4.2. ALINEAMIENTO Y DISTRIBUCIÓN DE LA PROFUNDIDAD DE MAPEO

El alineamiento de las lecturas de alta calidad se realizó contra la versión 4.1 del genoma de referencia de berenjena 67/3 (Barchi *et al.*, 2021; https://solgenomics.net/). Este genoma de referencia se caracteriza por estar formado por 13 super*scaffolds* o pseudocromosomas, uno por cada cromosoma y uno adicional correspondiente al cromosoma 0. La tasa de mapeo fue idéntica entre las diferentes coberturas de secuenciación (1X-5X), variando entre el 98,54 y el 98,55% del total de lecturas (**Tabla 4**). Es importante prestar especial atención a las secuencias pareadas adecuadamente mapeadas, que

son los pares de lecturas que cumplen con los criterios de orientación y distancia establecidos durante la preparación de la librería de secuenciación. El porcentaje de estas secuencias también es alto en todos los casos, rondando el 98,10% del total de lecturas (**Tabla 4**). En el caso del alineamiento realizado con los datos de la secuenciación a 20X, los ratios de mapeo y de lecturas pareadas adecuadamente mapeadas fueron menores, del 88,72% y 76,68% del total de lecturas, respectivamente (**Tabla 4**).

Tabla 4. Resumen de los estadísticos de la secuenciación y del mapeo de las lecturas obtenidas con las coberturas de secuenciación de 1X, 2X, 3X, 4X, 5X y 20X, respectivamente. Para los casos de la resecuenciación *in silico* (1-4X), los valores representan la media de las 5 réplicas ± desviación estándar.

	1X	2X	3X	4X	5X	20X
Lecturas totales (x 10^3)	7.915,29 ± 3,63	15.828,36 ± 2,18	23.741,65 ± 3,99	31.657,00 ± 4,03	39.570,26	170.761,55
Lecturas alineadas (x 10^3)	7.800 ± 3,63	15.598,22 ± 2,25	23.395,48 ± 3,99	31.195,91 ± 4,30	38.993,78	151.497,52
% lecturas alineadas	98,54 ± 0,00	98,55 ± 0,00	98,54 ± 0,00	98,54 ± 0,00	98,54	88,72
Lecturas pareadas mapeadas (x 10^3)	7.764,66 ± 3,73	15.527,55 ± 2,40	23.289,46 ± 4,04	31.054,66 ± 4,43	38.817,08	134.363.128,00
% lecturas pareadas mapeadas	98,10 ± 0,01	98,10 ± 0,00	98,10 ± 0,00	98,10 ± 0,00	98,10	78,68
Lecturas sin pareja (x 10^3)	35,34 ± 0,16	70,67 ± 0,18	106,02 ± 0,18	141,26 ± 0,15	176,69	17.134,39
% lecturas sin pareja	0,45 ± 0,00	0,45 ± 0,00	0,45 ± 0,00	0,45 ± 0,00	0,45	10,03
Lecturas no alineadas (x 10^3)	115,29 ± 0,36	230,14 ± 0,31	346,17 ± 0,49	461,09 ± 0,46	576,49	19.264,03
% lecturas no alineadas	1,46 ± 0,00	1,45 ± 0,00	1,46 ± 0,00	1,46 ± 0,00	1,46	11,28
Ratio de duplicación (%)	2,89 ± 0,01	4,53 ± 0,01	5,64 ± 0,00	6,46 ± 0,01	7,10	15,92
Calidad media del mapeo	27,68 ± 0,01	27,68 ± 0,00	27,68 ± 0,00	27,68 ± 0,00	27,68	28,04
Cobertura del genoma (%)	46,11 ± 0,03	67,24 ± 0,01	77,56 ± 0,01	82,79 ± 0,00	85,54	90,07
Profundidad de lectura máxima	443,20 ± 7,36	893,4 ± 20,19	1.298,00 ± 19,51	1.719,00 ± 16,12	2.147,00	33.380,00
Profundidad de lectura media	1,00 ± 0,00	2,01 ± 0,00	3,01 ± 0,00	4,02 ± 0,00	5,02	18,45

En cuanto al ratio estimado de secuencias duplicadas tras el alineamiento y la profundidad de lectura máxima alcanzada en algunas regiones del genoma, estos dos parámetros aumentaron con la cobertura de secuenciación (**Tabla 4**), llegando a alcanzar, respectivamente, valores de 7,10% y 2.147X a una cobertura de secuenciación de 5X, y de 15,92% y 33.380X a una cobertura de 20X. En la **Tabla 4** también se indica la MAPQ, que en todos los casos fue de 27,68, excepto para el caso de la secuenciación a 20X, que fue algo mayor, de 28,04. Una visión más detallada de cómo varía la calidad del mapeo a lo largo del genoma de referencia para el caso de la resecuenciación a 5X se muestra en la **Figura 14**. Se puede apreciar una calidad de mapeo inferior en las primeras posiciones del genoma de referencia y una calidad relativamente estable en el resto de alineamientos, exceptuando algunos casos en los que se alcanza una calidad inferior a 10. Para el resto de coberturas de secuenciación, incluida la cobertura 20X, el patrón fue el mismo.

Figura 14. Calidad del mapeo de las lecturas obtenidas de la resecuenciación a 5X a lo largo del genoma de referencia. La figura ha sido generada vía QualiMap (García-Alcalde *et al.*, 2012).

La secuenciación a una cobertura de 1X significa que, de media, todas las posiciones del genoma se secuencian una vez o, lo que es lo mismo, de media, todas las posiciones del genoma de referencia se alinean con una única lectura. A pesar de ser lo esperado desde una perspectiva teórica, es esencial llevar a cabo comprobaciones prácticas. Por esta razón, se realizó un análisis de los datos con el fin de determinar el porcentaje de bases del genoma de referencia cubiertas por diferentes profundidades de mapeo para cada una de las coberturas de secuenciación empleadas en el estudio (**Figura 15**).

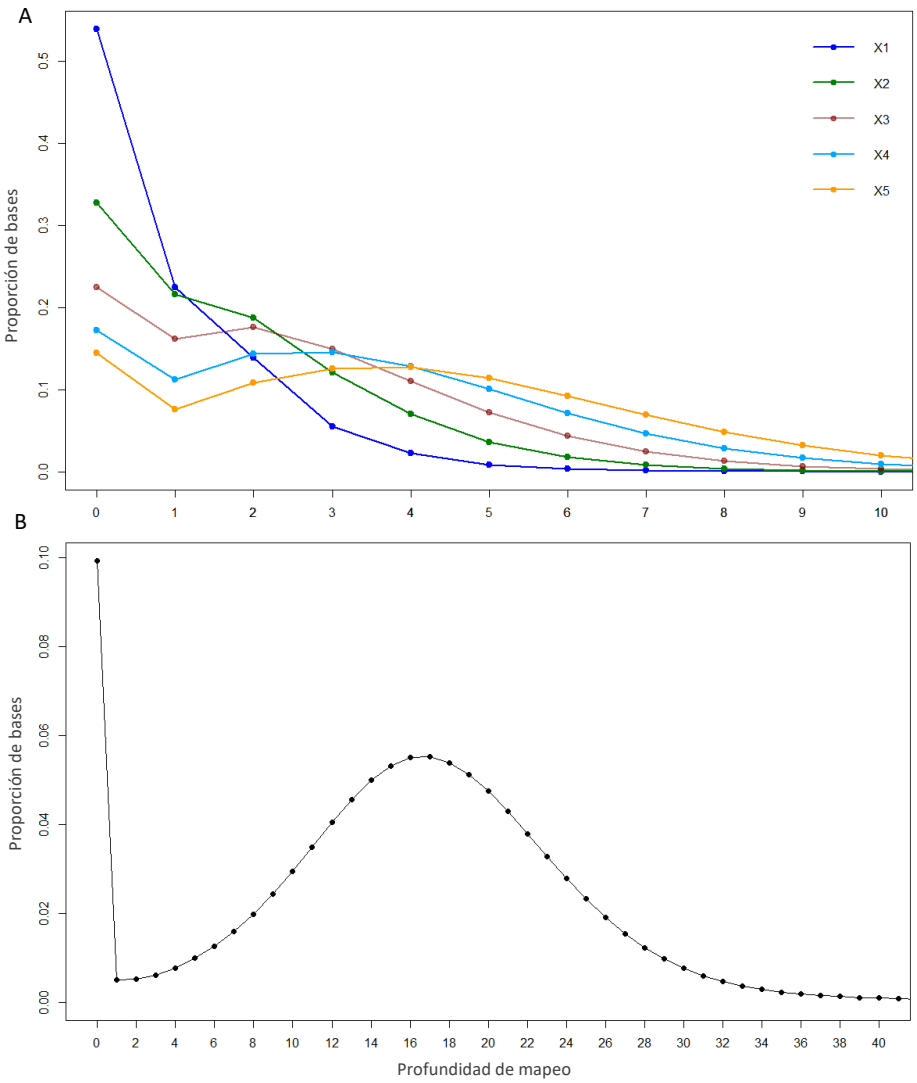


Figura 15. Relación entre la profundidad de mapeo y el porcentaje de bases secuenciadas para diferentes niveles de cobertura de secuenciación. **A.** Coberturas de secuenciación 1X, 2X, 3X, 4X, 5X. Las cinco réplicas mostraron la misma distribución. **B.** Cobertura de secuenciación 20X.

En la **Figura 15** se observa que a bajas coberturas de secuenciación el número de posiciones no secuenciadas del genoma es mayor que al emplear altas coberturas de secuenciación. Por otro lado, a medida que aumenta la cobertura media de secuenciación, se incrementa el porcentaje de bases mapeadas a profundidades más altas, mientras que se reduce el porcentaje de bases mapeadas a profundidades más bajas. Esto conduce a un desplazamiento del pico más alto de porcentaje de bases secuenciadas hacia la derecha.

También es importante conocer cómo se distribuyen las lecturas a lo largo de los cromosomas, pues es probable que existan regiones del genoma que se encuentren mapeadas a mayores profundidades que otras. Si este es el caso, si las lecturas se distribuyen de forma inapropiada, los siguientes pasos

del análisis bioinformático de los datos generados mediante SWGR se pueden ver afectados negativamente, como, por ejemplo, la precisión en la identificación de variantes genéticas. En la **Figura 16** se muestra la distribución de la profundidad de mapeo a lo largo del cromosoma 1 para el conjunto de datos obtenidos con cada una de las coberturas de secuenciación empleadas en este estudio. Las regiones del genoma que presentan una alta concentración de lecturas en una ventana de tamaño de 10 Kpb se encuentran representadas por picos. En general, las lecturas se distribuyen a lo largo de todo el cromosoma y se pueden distinguir ciertos patrones idénticos en los seis casos, como la región sobrerrepresentada - probablemente debido a secuencias repetitivas o errores de ensamblaje - alrededor de la posición 40 Mbp. Aunque inicialmente pueda parecer que la secuenciación a bajas coberturas proporciona una representación completa del genoma de referencia (**Figura 16**), los resultados obtenidos mediante el uso de la herramienta SAMtools revelan lo contrario. Con una cobertura de secuenciación de 1X, tan solo se logra una representación del 46% del genoma (**Tabla 4**), lo cual es consistente con el porcentaje de bases que presentan una profundidad de mapeo igual a 0 (**Figura 15**). Con un incremento de la cobertura de secuenciación a 5X, la cobertura del genoma aumenta casi al doble, alcanzando aproximadamente el 86% (**Tabla 4**). Sin embargo, el incremento en la cobertura del genoma es menor al aumentar la cobertura de secuenciación hasta 20X, donde aproximadamente el 90% del genoma se alinea con lecturas (**Tabla 4**).

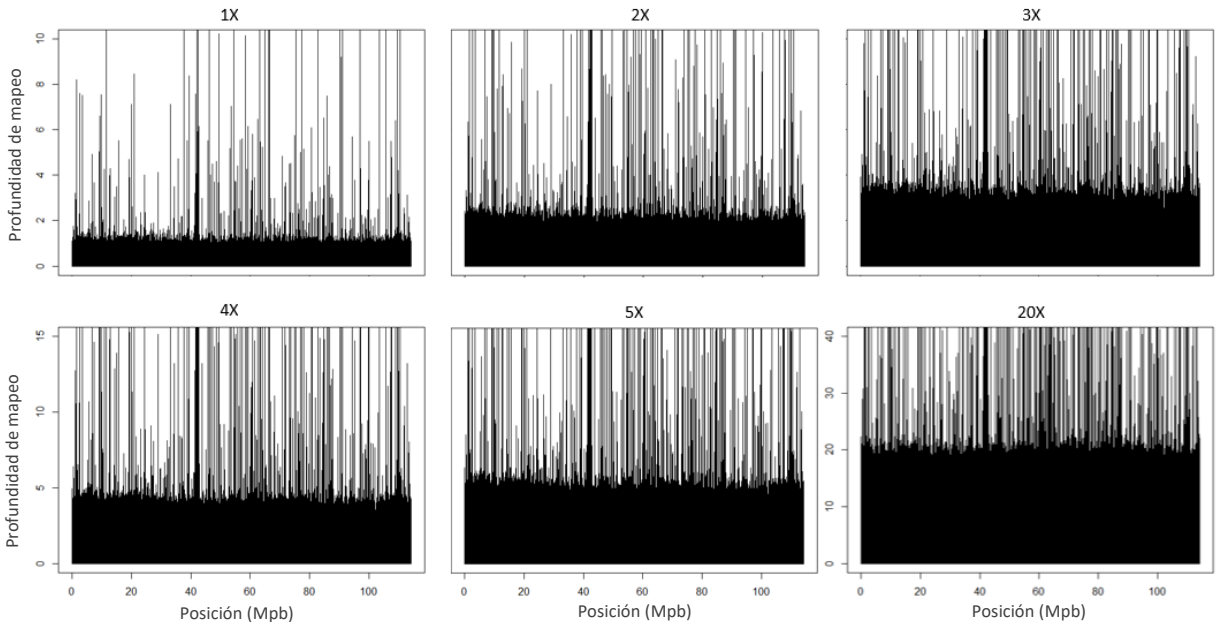


Figura 16. Distribución de la profundidad de mapeo a lo largo del cromosoma 1 para las diferentes coberturas de secuenciación empleadas en el estudio. Los picos representan altas profundidades de mapeo en un tamaño de ventana de 10 Kbp.

4.3. IDENTIFICACIÓN Y DISTRIBUCIÓN DE VARIANTES GENÉTICAS

Ante la presencia de secuencias duplicadas en los conjuntos de datos analizados, se consideró necesario realizar un marcado de las mismas con el fin de evitar que fueran consideradas una evidencia adicional a favor o en contra de la identificación de variantes putativas. En la **Figura 17** se pueden apreciar las diferencias en cuanto al número de variantes identificadas teniendo en cuenta o no las secuencias duplicadas. A mayores coberturas de secuenciación empleadas, estas diferencias aumentaron. En concreto, la omisión del marcado de las secuencias duplicadas conllevó la identificación de un mayor número de variantes. Sin embargo, a coberturas de secuenciación bajas (1X) y utilizando filtros de profundidad mínima de mapeo restrictivos (≥ 4F/FD) no se encontraron

diferencias significativas entre el número de variantes encontradas al realizar o no el marcado de los duplicados.

Una vez realizada la corrección de los duplicados en los datos de secuenciación, se observó un incremento en el número de variantes identificadas al aumentar la cobertura de secuenciación y al disminuir la profundidad mínima de mapeo (**Figura 17**). Cabe destacar también que no se encontraron diferencias significativas entre el número de variantes encontradas filtrando a una profundidad mínima de 1FD y las encontradas filtrando a una profundidad mínima de 2FD para las coberturas de secuenciación de 1X a 4X (**Figura 17**).

Para todas las combinaciones de cobertura de secuenciación y filtro de profundidad mínima de mapeo empleadas, el número de SNPs fue superior al número de indels identificados (**Tabla 5**). Este último, a su vez, fue superior al número de otras variantes identificadas en el genoma, tales como polimorfismos de sitio múltiple y variantes complejas, para profundidades mínimas de mapeo bajas. En el caso de la secuenciación a una cobertura de 1X, el número de indels fue superior al de otras variantes diferentes a SNPs para profundidades mínimas de mapeo inferiores a 4FD, incluida. A una cobertura de secuenciación de 2X, esto ocurrió hasta una profundidad de mapeo de 6FD. Con los datos de resecuenciación a una cobertura 3X, se obtuvo un mayor número de indels que de otras variantes, a excepción de los SNPs, cuando se filtró hasta una profundidad mínima de 7FD. Finalmente, con coberturas de secuenciación de 4X y 5X el número de polimorfismos de sitio múltiple y variantes complejas fue superior al número de indels para profundidades de mapeo mínimas de 9FD y 10FD (**Tabla 5**). Esto quiere decir que, a medida que aumenta el grado de restricción a la hora de aplicar el filtro de la profundidad mínima de mapeo, el número de indels disminuye más rápido que el número de variantes de tipo polimorfismos de sitio múltiple y de variantes complejas.

Puede haber cuatro genotipos diferentes para cada polimorfismo encontrado en el genoma en estudio: homocigoto para un alelo alternativo, heterocigoto para alelos alternativos, homocigoto para el alelo de referencia y heterocigoto con un alelo de referencia y otro alternativo (**Tabla 5**). Aunque los genotipos más abundantes fueron el primero y el último de la lista anterior, la predominancia de uno sobre el otro varió con la cobertura de secuenciación y profundidad mínima de mapeo empleada. Cuando se empleó una cobertura de secuenciación de 1X, predominó el genotipo homocigoto para el alelo alternativo cuando se filtró hasta una profundidad mínima de 5FD, incluida. Con una cobertura de secuenciación de 2X, este genotipo predominó hasta utilizar una profundidad mínima de mapeo de 7FD. Para una cobertura de secuenciación de 3X, predominó hasta utilizar una profundidad mínima de 8FD. Para una cobertura de secuenciación de 4X, hasta utilizar una profundidad mínima de 9FD. Finalmente, con una cobertura de secuenciación de 5X, el genotipo homocigoto para el alelo alternativo fue siempre predominante, independientemente de la profundidad mínima de mapeo utilizada (**Tabla 5**). En contraste, el genotipo que se encontró con menor frecuencia fue el heterocigoto para alelos alternativos (A1A2) (**Tabla 5**).

Los polimorfismos encontrados secuenciando a una cobertura de 20X fueron sometidos a tres pasos de filtrado paralelos con profundidades mínimas diferentes (1FD, 10FD y 20FD) tras haber realizado también el paso de marcado de las secuencias duplicadas. Con el primero de ellos se obtuvo un total de 1.784.060 variantes, mientras que empleando el filtro de 10FD y 20FD se obtuvieron 1.187.895 y 309.642 variantes, respectivamente. En base a estos resultados, se decidió emplear el archivo VCF resultante de la secuenciación a 20X y del filtrado a 10FD como estándar de referencia. La principal razón fue que, a esta profundidad mínima de mapeo, no se perdieron tantas variantes como filtrando a una profundidad mínima de 20FD, y que, además, es una profundidad mínima suficiente como para poder distinguir entre genotipos homocigotos y heterocigotos. Mediante la aplicación del filtro

utilizado para limitar la profundidad de mapeo máxima a 40FD y la fracción mínima del alelo alternativo a 0,3, el número de polimorfismos disminuyó hasta 723.420. Finalmente, éste se redujo hasta 705.954 cuando se eliminaron los genotipos homocigotos para el alelo de referencia y las variantes localizadas en el cromosoma 0 (**Tabla 6**). Del total de 705.954 variantes genéticas, tan sólo 51.477 (7,3%) fueron indels y 56.198 (8%) fueron genotipos heterocigotos (**Tabla 6**).

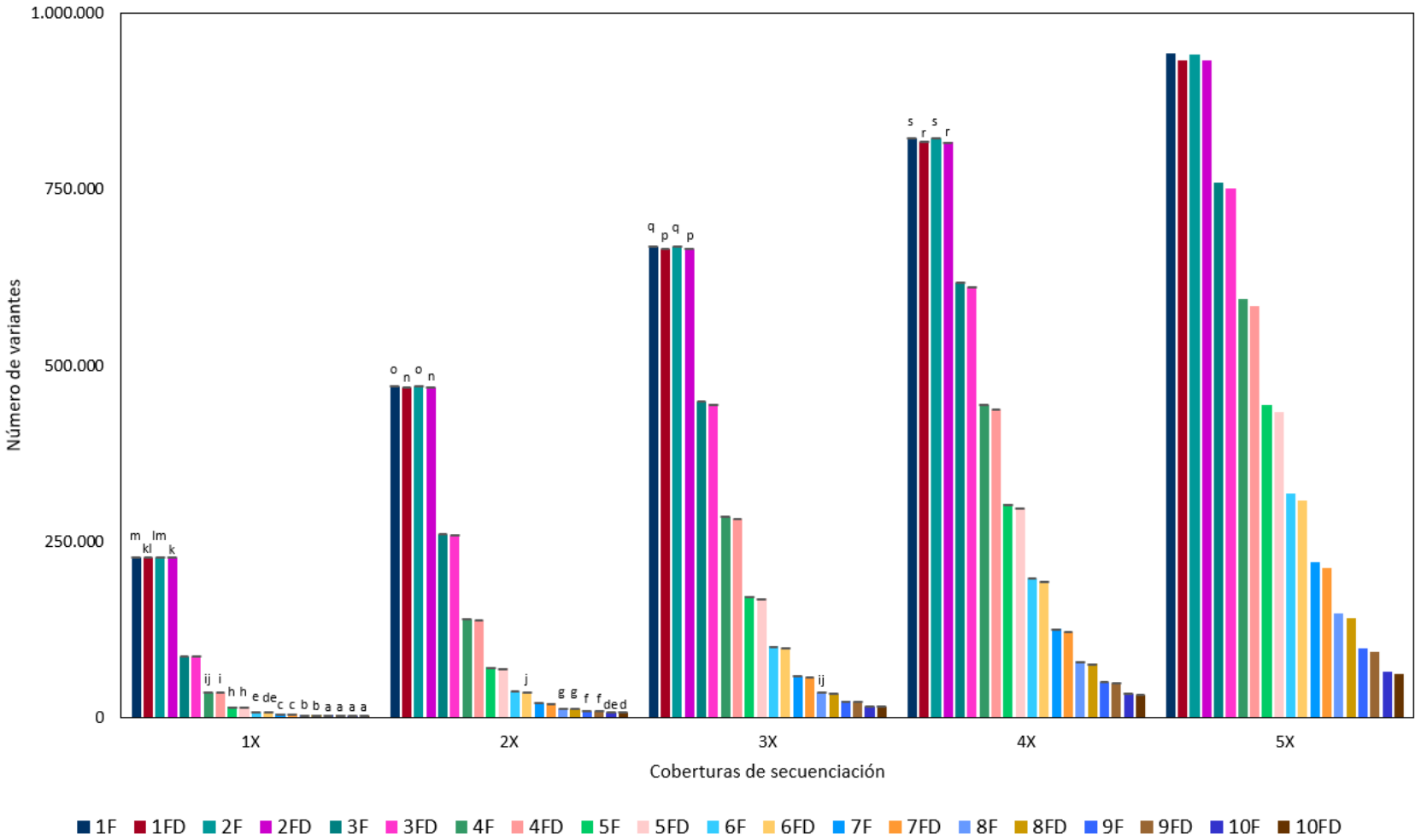


Figura 17. Número de variantes genéticas identificadas con cada cobertura de secuenciación (1-5X) después del filtrado a diferentes profundidades mínimas de mapeo (1-10X). El filtrado de los polimorfismos identificados sin haber realizado previamente el marcado de las secuencias duplicadas se encuentra designado como "F", mientras que el realizado sobre los polimorfismos identificados habiendo marcado las secuencias duplicadas, como "FD". Los datos correspondientes a las coberturas de secuenciación 1-4X llevan su desviación estándar adjunta en forma de barra de error (n=5). Diferentes letras indican diferencias significativas para $p < 0,05$ con el método LSD. Las columnas que no presentan letra para las coberturas de secuenciación 1-4X no comparten grupo homogéneo con ningún otro conjunto de datos.

Tabla 5. Resumen de las variantes genéticas identificadas a cada combinación de cobertura de secuenciación (1-5X) y filtro de profundidad mínima de mapeo (1-10FD). Homocigoto para el alelo alternativo: A1A1. Heterocigoto para el alelo alternativo: A1A2. Homocigoto para el alelo de referencia: RR. Heterocigoto con alelo de referencia y alternativo: A1R. Para los casos de la resecuenciación *in silico* (1-4X), los valores representan la media de las 5 réplicas ± desviación estándar. En cada columna, letras diferentes indican diferencias significativas para $p < 0,05$ con el método LSD.

	1X									
	1FD	2FD	3FD	4FD	5FD	6FD	7FD	8FD	9FD	10FD
Variantes totales	227.346,60 ± 475,51 a	227.232,00 ± 461,87 a	87.101,80 ± 215,77 a	35.757,00 ± 174,97 a	15.241,00 ± 97,95 a	8.150,80 ± 102,24 a	5.419,60 ± 102,35 a	4.279,00 ± 80,66 a	3.662,20 ± 65,33 a	3.248,20 ± 60,54 a
Número de SNPs	208.131,00 ± 420,01 d	208.130,40 ± 420,10 d	79.374,40 ± 166,39 b	32.440,80 ± 133,57 b	13.551,40 ± 91,06 b	6.965,80 ± 90,02 b	4.406,60 ± 71,07 b	3.339,80 ± 52,14 b	2.764,20 ± 39,61 b	2.374,00 ± 36,81 b
Número de indels	16.612,80 ± 106,81 e	16.513,00 ± 111,24 e	6.257,00 ± 74,70 f	2.275,40 ± 43,80 f	812,80 ± 35,29 fg	366,80 ± 13,48 h	217,20 ± 9,60 g	156,80 ± 7,92 g	126,00 ± 8,54 g	106,80 ± 5,67 g
Otras variaciones	2.602,80 ± 21,57 g	2.588,60 ± 22,73 g	1.470,40 ± 36,09 g	1.040,80 ± 33,59 g	876,80 ± 47,77 f	818,20 ± 53,92 f	795,80 ± 52,05 f	782,40 ± 46,60 f	772,00 ± 44,06 f	767,40 ± 44,69 f
Homocigoto para alelo alternativo	212.654,60 ± 426,67 c	212.540,00 ± 413,96 c	72.411,20 ± 236,09 d	25.751,60 ± 104,65 d	8.805,60 ± 58,49 d	3.658,40 ± 27,74 e	2.033,60 ± 40,05 e	1.526,60 ± 34,45 e	1.309,40 ± 19,19 e	1.184,40 ± 20,07 e
Heterocigoto para alelos alternativos	115,40 ± 21,01 i	115,40 ± 21,01 i	114,00 ± 19,57 i	114,00 ± 19,57 i	68,00 ± 9,85 h	47,20 ± 4,55 i	36,00 ± 5,29 h	29,20 ± 4,38 h	25,40 ± 3,91 h	22,00 ± 3,61 h
Homocigoto para alelo de referencia	738,40 ± 46,24 h	738,40 ± 46,24 h	738,40 ± 46,24 h	738,40 ± 46,24 h	738,40 ± 46,24 g	738,40 ± 46,24 g	738,40 ± 46,24 f	738,40 ± 46,24 f	738,40 ± 46,24 f	738,40 ± 46,24 f
Heterocigoto con alelo de referencia y alternativo	13.838,20 ± 152,48 f	13.838,20 ± 152,48 f	13.838,20 ± 152,48 e	9.153,00 ± 154,24 e	5.629,00 ± 141,70 e	3.706,80 ± 78,56 de	2.611,60 ± 69,33 d	1.984,80 ± 66,74 d	1.589,00 ± 50,70 d	1.303,40 ± 45,38 d
Total homocigoto	213.393,00 ± 462,69 b	213.278,40 ± 450,09 b	73.149,60 ± 274,95 c	26.490,00 ± 140,91 c	9.544,00 ± 98,17 c	4.396,80 ± 46,21 c	2.772,00 ± 82,35 c	2.265,00 ± 77,04 c	2.047,80 ± 60,06 c	1.922,80 ± 58,12 c
Total heterocigoto	13.953,60 ± 153,69 f	13.953,60 ± 153,69 f	13.952,20 ± 153,69 e	9.267,00 ± 151,74 e	5.697,00 ± 142,16 e	3.754,00 ± 75,49 d	2.647,60 ± 64,72 d	2.014,00 ± 63,15 d	1.614,40 ± 47,91 d	1.325,40 ± 41,99 d

	2X									
	1FD	2FD	3FD	4FD	5FD	6FD	7FD	8FD	9FD	10FD
Variantes totales	468.128,60 ± 863,72 a	467.896,20 ± 869,23 a	259.197,00 ± 528,09 a	137.970,20 ± 389,76 a	70.217,00 ± 188,75 a	36.574,80 ± 154,28 a	20.594,20 ± 60,98 a	13.124,00 ± 62,72 a	9.580,20 ± 75,88 a	7.756,60 ± 90,39 a
Número de SNPs	426.715,4,00 ± 678,68 c	426.714,00 ± 677,85 d	235.738,00 ± 529,99 b	125.631,80 ± 312,21 b	63.715,60 ± 119,02 b	32.760,40 ± 143,08 b	17.917,00 ± 68,57 b	10.953,00 ± 86,95 b	7.633,40 ± 100,37 b	5.918,40 ± 110,93 b
Número de indels	35.910,00 ± 207,90 f	35.707,00 ± 213,56 f	19.783,80 ± 120,18 f	9.769,40 ± 88,91 g	4.485,40 ± 44,55 g	2.040,00 ± 36,19 g	1.022,40 ± 23,31 h	572,60 ± 15,06 i	381,20 ± 17,25 g	290,20 ± 14,70 g
Otras variaciones	5.503,20 ± 71,21 g	5.475,20 ± 72,90 g	3.675,20 ± 57,49 g	2.569,00 ± 44,51 h	2.016,00 ± 49,18 h	1.774,40 ± 53,93 h	1.654,80 ± 54,62 f	1.598,40 ± 56,36 g	1.565,60 ± 54,29 f	1.548,00 ± 55,44 f
Homocigoto para alelo alternativo	428.199,60 ± 784,52 c	427.967,20 ± 789,89 c	219.268,00 ± 537,34 d	106.931,60 ± 393,15 d	48.198,40 ± 169,89 d	21.142,20 ± 128,81 d	9.533,60 ± 51,67 de	4.814,20 ± 48,79 f	2.960,20 ± 31,42 e	2.227,80 ± 35,65 e
Heterocigoto para alelos alternativos	360,20 ± 32,75 i	360,20 ± 32,75 i	360,20 ± 32,75 i	360,20 ± 32,75 j	264,00 ± 22,86 j	176,80 ± 15,58 j	123,40 ± 11,76 i	89,80 ± 6,98 j	74,40 ± 3,97 h	66,40 ± 4,34 h
Homocigoto para alelo de referencia	1.474,00 ± 52,87 h	1.474,00 ± 52,87 h	1.474,00 ± 52,87 h	1.474,00 ± 52,87 i	1.474,00 ± 52,87 i	1.474,00 ± 52,87 i	1.474,00 ± 52,87 g	1.474,00 ± 52,87 h	1.474,00 ± 52,87 f	1.474,00 ± 52,87 f
Heterocigoto con alelo de referencia y alternativo	38.094,80 ± 216,44 e	38.094,80 ± 216,44 e	38.094,80 ± 216,44 e	29.204,40 ± 182,88 f	20.280,60 ± 129,16 f	13.781,80 ± 108,72 f	9.463,20 ± 89,64 e	6.746,00 ± 80,50 d	5.071,60 ± 109,79 c	3.988,40 ± 103,28 c
Total homocigoto	429.673,60 ± 824,61 b	429.441,20 ± 829,99 b	220.742,00 ± 576,27 c	108.405,60 ± 442,45 c	49.672,40 ± 217,01 c	22.616,20 ± 166,81 c	11.007,60 ± 90,94 c	6.288,20 ± 83,15 e	4.434,20 ± 73,73 d	3.701,80 ± 68,16 d
Total heterocigoto	38.455,00 ± 227,86 e	38.455,00 ± 227,86 e	38.455,00 ± 227,86 e	29.564,60 ± 174,90 e	20.544,60 ± 134,88 e	13.958,60 ± 114,35 e	9.586,60 ± 92,13 d	6.835,80 ± 85,51 c	5.146,00 ± 113,04 c	4.054,80 ± 105,44 c

	3X									
	1FD	2FD	3FD	4FD	5FD	6FD	7FD	8FD	9FD	10FD
Variantes totales	665.479,80 ± 747,24 a	665.223,00 ± 753,57 a	444.651,40 ± 585,71 a	282.010,20 ± 134,70 a	169.310,80 ± 168,74 a	98.626,60 ± 336,80 a	57.603,60 ± 327,78 a	34.947,80 ± 186,28 a	22.704,80 ± 109,99 a	16.226,60 ± 100,57 a
Número de SNPs	604.108,40 ± 717,25 b	604.105,40 ± 718,21 b	403.043,40 ± 471,28 b	256.237,40 ± 219,94 b	153.959,80 ± 155,14 b	89.276,00 ± 313,80 b	51.442,60 ± 328,25 b	30.418,60 ± 173,99 b	19.022,60 ± 80,43 b	12.968,80 ± 71,99 b
Número de indels	52.997,80 ± 143,59 g	52.774,80 ± 146,72 g	35.356,20 ± 160,12 g	21.080,00 ± 104,11 g	11.634,20 ± 47,81 g	6.178,80 ± 8,26 g	3.281,00 ± 30,41 g	1.811,00 ± 27,78 h	1.048,80 ± 17,77 i	678,00 ± 15,25 i
Otras variaciones	8.373,60 ± 27,89 h	8.342,80 ± 24,13 h	6.251,80 ± 49,78 h	4.692,80 ± 55,01 h	3.716,80 ± 50,94 h	3.171,80 ± 49,08 h	2.880,00 ± 52,23 h	2.718,20 ± 45,33 f	2.633,40 ± 40,97 g	2.579,80 ± 43,68 g
Homocigoto para alelo alternativo	597.167,20 ± 791,83 d	596.910,40 ± 800,23 d	376.338,80 ± 631,10 d	224.216,40 ± 252,19 d	124.167,00 ± 179,03 d	64.820,60 ± 166,61 d	32.631,80 ± 187,67 d	16.314,60 ± 112,10 d	8.412,60 ± 15,85 f	4.856,60 ± 14,88 f
Heterocigoto para alelos alternativos	690,00 ± 22,25 j	690,00 ± 22,25 j	690,00 ± 22,25 j	6.90,00 ± 22,25 j	537,20 ± 19,28 j	372,60 ± 17,50 j	258,40 ± 7,86 j	180,20 ± 7,12 i	138,60 ± 8,38 j	116,00 ± 5,83 j
Homocigoto para alelo de referencia	2.431,40 ± 39,63 i	2.431,40 ± 39,63 i	2.431,40 ± 39,63 i	2.431,40 ± 39,63 i	2.431,40 ± 39,63 i	2.431,40 ± 39,63 i	2.431,40 ± 39,63 i	2.431,40 ± 39,63 g	2.431,40 ± 39,63 h	2.431,40 ± 39,63 h
Heterocigoto con alelo de referencia y alternativo	65.191,20 ± 139,82 f	65.191,20 ± 139,82 f	65.191,20 ± 139,82 f	54.672,40 ± 163,78 f	42.175,20 ± 123,25 f	31.002,00 ± 159,49 f	22.282,00 ± 214,84 f	16.021,60 ± 98,22 e	11.722,20 ± 88,70 d	8.822,60 ± 72,83 d
Total homocigoto	599.598,60 ± 759,02 c	599.341,80 ± 767,37 c	378.770,20 ± 613,80 c	226.647,80 ± 227,79 c	126.598,40 ± 160,85 c	67.252,00 ± 178,95 c	35.063,20 ± 178,39 c	18.746,00 ± 115,98 c	10.844,00 ± 46,23 e	7.288,00 ± 50,29 e
Total heterocigoto	65.881,20 ± 152,37 e	65.881,20 ± 152,37 e	65.881,20 ± 152,37 e	55.362,40 ± 179,15 e	42.712,40 ± 142,04 e	31.374,60 ± 166,23 e	22.540,40 ± 217,67 e	16.201,80 ± 97,90 d	11.860,80 ± 95,59 c	8.938,60 ± 76,00 c

	4X									
	1FD	2FD	3FD	4FD	5FD	6FD	7FD	8FD	9FD	10FD
Variantes totales	816.763,80 ± 197,07 a	816.494,20 ± 192,62 a	611.187,00 ± 424,15 a	437.496,00 ± 404,19 a	296.929,00 ± 260,63 a	193.095,40 ± 302,11 a	122.000,60 ± 223,40 a	76.750,40 ± 135,76 a	49.032,00 ± 91,24 a	32.889,00 ± 167,78 a
Número de SNPs	739.019,60 ± 163,80 b	739.015,20 ± 163,78 b	552.300,60 ± 453,77 b	396.468,00 ± 303,16 b	269.715,60 ± 231,29 b	175.310,60 ± 225,19 b	110.241,60 ± 162,59 b	68.507,20 ± 112,16 b	42.770,80 ± 104,62 b	27.700,80 ± 137,67 b
Número de indels	66.817,20 ± 93,36 g	66.583,80 ± 87,55 g	50.097,40 ± 96,29 g	34.047,00 ± 165,44 g	21.535,40 ± 33,34 g	12.934,80 ± 71,88 g	7.472,60 ± 72,14 g	4.278,20 ± 36,84 g	2.479,20 ± 47,42 i	1.519,00 ± 19,71 i
Otras variaciones	10.927,00 ± 38,84 h	10.895,20 ± 36,55 h	8.789,00 ± 37,18 h	6.981,00 ± 42,70 h	5.678,00 ± 55,39 h	4.850,00 ± 52,42 h	4.286,40 ± 38,15 h	3.965,00 ± 31,52 h	3.782,00 ± 29,53 g	3.669,20 ± 36,56 g
Homocigoto para alelo alternativo	719.404,60 ± 147,12 d	719.135,00 ± 142,94 d	513.827,80 ± 362,20 d	350.555,00 ± 223,54 d	224.222,60 ± 255,63 d	135.033,20 ± 192,08 d	77.205,00 ± 159,67 d	42.670,80 ± 124,71 d	22.943,00 ± 75,50 d	12.516,40 ± 84,22 f
Heterocigoto para alelos alternativos	1.063,60 ± 20,03 j	1.063,60 ± 20,03 j	1.063,60 ± 20,03 j	1.063,60 ± 20,03 j	867,80 ± 17,89 j	631,60 ± 21,14 j	439,80 ± 14,92 j	305,80 ± 13,88 j	229,40 ± 11,24 j	191,60 ± 13,30 j
Homocigoto para alelo de referencia	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 i	3.401,20 ± 29,79 h	3.401,20 ± 29,79 h
Heterocigoto con alelo de referencia y alternativo	92.894,40 ± 112,12 f	92.894,40 ± 112,12 f	92.894,40 ± 112,12 f	82.476,20 ± 177,50 f	68.437,40 ± 135,13 f	54.029,40 ± 143,82 f	40.954,60 ± 120,89 f	30.372,60 ± 96,66 f	22.458,40 ± 56,47 f	16.779,80 ± 76,30 d
Total homocigoto	722.805,80 ± 162,82 c	722.536,20 ± 159,38 c	517.229,00 ± 365,61 c	353.956,20 ± 236,75 c	227.623,80 ± 252,34 c	138.434,40 ± 208,01 c	80.606,20 ± 171,95 c	46.072,00 ± 128,75 c	26.344,20 ± 81,46 c	15.917,60 ± 105,86 e
Total heterocigoto	93.958,00 ± 110,56 e	93.958,00 ± 110,56 e	93.958,00 ± 110,56 e	83.539,80 ± 172,24 e	69.305,20 ± 129,35 e	54.661,00 ± 155,78 e	41.394,40 ± 127,49 e	30.678,40 ± 92,76 e	22.687,80 ± 58,77 e	16.971,40 ± 84,20 c

	5X									
	1FD	2FD	3FD	4FD	5FD	6FD	7FD	8FD	9FD	10FD
Variantes totales	932.989	932.725	750.381	584.277	434.406	308.981	212.175	141.837	93.838	62.802
Número de SNPs	842.049	842.043	676.271	527.913	393.766	280.358	192.286	127.707	83.434	54.582
Número de indels	77.583	77.358	62.794	46.904	32.640	21.712	13.757	8.512	5.132	3.163
Otras variaciones	13.357	13.324	11.316	9.460	8.000	6.911	6.132	5.618	5.272	5.057
Homocigoto para alelo alternativo	806.936	806.672	624.328	467.778	332.112	223.038	142.442	86.831	50.819	29.188
Heterocigoto para alelos alternativos	1.435	1.435	1.435	1.435	1.225	932	680	475	366	299
Homocigoto para alelo de referencia	4.612	4.612	4.612	4.612	4.612	4.612	4.612	4.612	4.612	4.612
Heterocigoto con alelo de referencia y alternativo	120.006	120.006	120.006	110.452	96.457	80.399	64.441	49.919	38.041	28.703
Total homocigoto	811.548	811.284	628.940	472.390	336.724	227.650	147.054	91.443	55.431	33.800
Total heterocigoto	121.441	121.441	121.441	111.887	97.682	81.331	65.121	50.394	38.407	29.002

Tabla 6. Resumen de las variantes genéticas identificadas con una cobertura de secuenciación 20X después de haber filtrado a una profundidad máxima y mínima de mapeo de 40FD y 10FD, una fracción mínima de alelo alternativo de 0,3, eliminado los genotipos homocigotos para el alelo de referencia y los polimorfismos localizados en el cromosoma 0. Homocigoto para el alelo alternativo: A1A1. Heterocigoto para el alelo alternativo: A1A2. Homocigoto para el alelo de referencia: RR. Heterocigoto con alelo de referencia y alternativo: A1R.

Variantes totales	Número de SNPs	Número de indels	Otras variaciones	Homocigoto para alelo alternativo	Heterocigoto para alelos alternativos	Homocigoto para alelo de referencia	Heterocigoto con alelo de referencia y alternativo	Total homocigoto	Total heterocigoto
705.954	649.888	51.477	4.589	649.756	1.492	0	54.706	649.756	56.198

4.4. EVALUACIÓN DE LA PRECISIÓN Y SENSIBILIDAD DE LA TÉCNICA

Para identificar y descartar falsos positivos presentes en el conjunto de datos de la resecuenciación a bajas coberturas (≤ 5X), se empleó el estándar de referencia puesto a punto a partir de los datos de la resecuenciación a 20X. Debido a la suficiente profundidad de lectura de estos polimorfismos, se espera una inferencia precisa de los genotipos correspondientes. En la **Figura 18** se puede ver que, incluso empleando una cobertura de secuenciación y una profundidad mínima de mapeo bajas (≤ 5X y < 7FD), el número de variantes compartidas con el estándar de referencia fue mayor que el número de variantes únicas para el conjunto de datos en estudio. Por otro lado, el número de variantes del estándar de referencia no identificadas como tales en las muestras fue mayor al emplear coberturas de secuenciación más bajas y un filtrado más restrictivo (1X10FD) (**Figura 18**). De la misma manera, es interesante destacar que el número total de variantes identificadas en las lecturas obtenidas de la resecuenciación a 4X y 5X, filtradas a una profundidad mínima inferior a 2FD y 3FD, respectivamente, fue mayor que el número de variantes identificadas en el estándar de referencia debido a la diferencia de filtros aplicados (**Figura 18**).

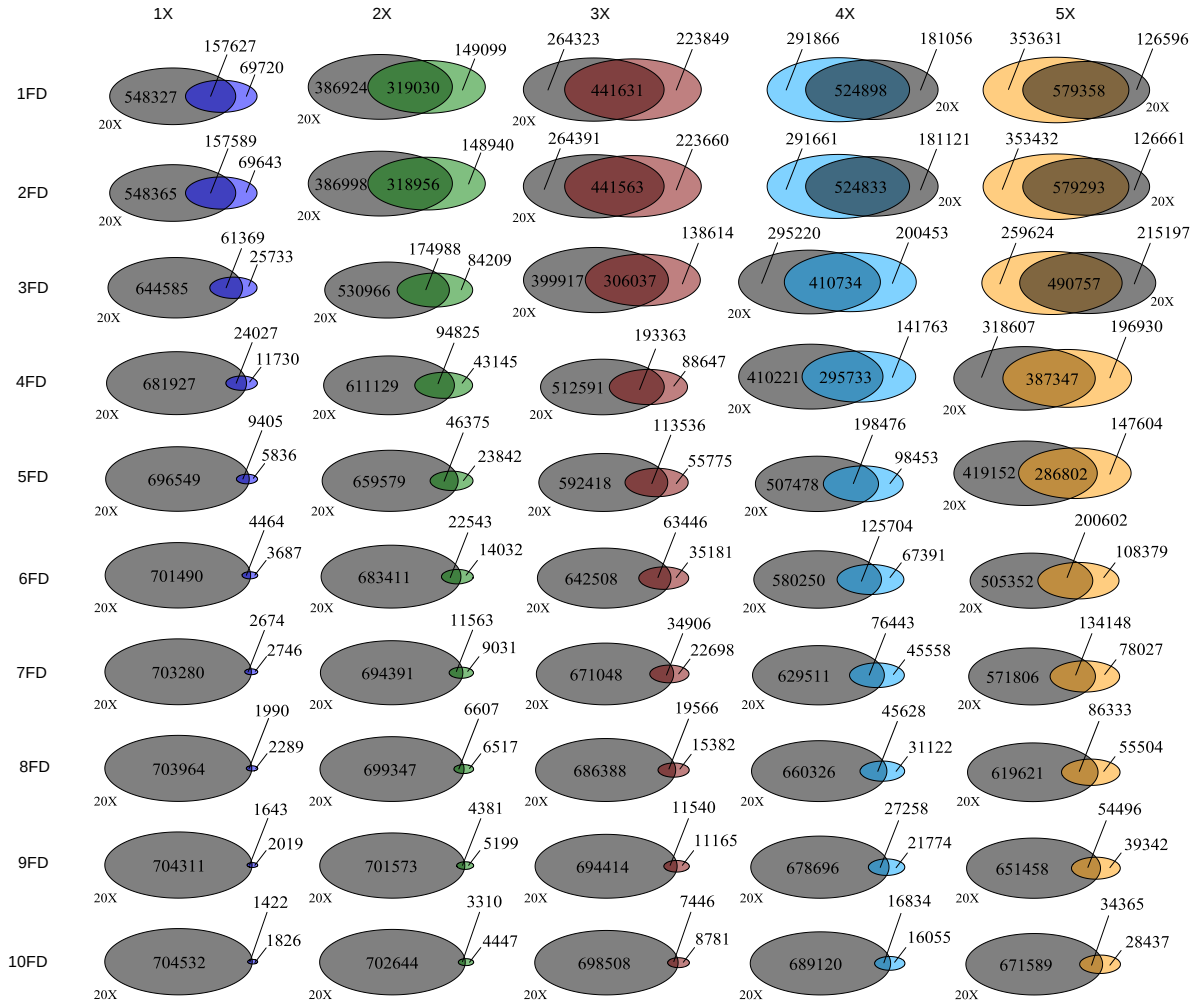


Figura 18. Comparación de cada conjunto de datos (color) con el estándar de referencia a 20X (gris) filtrado a una profundidad máxima de 40FD, a una mínima de 10FD y a una fracción mínima del alelo alternativo de 0,3, sin genotipos homocigotos para el alelo de referencia y sin variantes en el cromosoma 0. En cada columna se encuentran las diferentes coberturas de secuenciación (1-5X) y en cada fila las diferentes profundidades mínimas de mapeo (1-10FD). Los datos que aparecen representados para las coberturas de secuenciación 1-4X se corresponden con la media de las cinco réplicas.

Las **Figuras 19** y **20** presentan la proporción de polimorfismos verdaderos identificados en las muestras en relación con el total identificado (precisión), y la proporción de polimorfismos verdaderos identificados en comparación con el total existente en el genoma (sensibilidad) para diferentes coberturas de secuenciación (1-5X) y profundidades mínimas de mapeo (1-10FD), respectivamente. En los cinco casos, la precisión alcanzó un máximo a diferentes profundidades mínimas de mapeo a partir del cual comenzó a disminuir (**Figura 19**). Mientras que para las coberturas de secuenciación 1X y 3X, el máximo se logró a una profundidad mínima de 3FD, para el resto se alcanzó a una profundidad mínima de 4FD (**Figura 19**). Cabe destacar, por otro lado, que la cobertura de secuenciación requerida para alcanzar la máxima precisión varió en función de la profundidad mínima de mapeo. Para profundidades inferiores a 3FD, la máxima precisión se consiguió con una cobertura de secuenciación 1X. A una profundidad de mapeo de 4FD, se alcanzó con coberturas de secuenciación de 2X y 3X; a 5FD, con coberturas de 3X y 4X; a 6FD, con coberturas de 3X, 4X y 5X; y a 7FD, con coberturas de 4X y 5X. Para profundidades mínimas de mapeo superiores a 7FD, la máxima precisión se logró con una cobertura de secuenciación de 5X (**Figura 19**). En cuanto a la sensibilidad de la técnica, mostró un patrón más claro: fue mayor al utilizar mayores coberturas de secuenciación y menores profundidades mínimas de mapeo (**Figura 20**).

Teniendo en cuenta conjuntamente tanto la precisión como la sensibilidad de la técnica a la hora de identificar polimorfismos, se pudo observar que con una cobertura de secuenciación de 2X y una profundidad mínima de mapeo de 10FD se alcanzó la menor precisión (~44%) (**Figura 19**) y una sensibilidad cercana al 0% (**Figura 20**). Por otro lado, se logró la mayor precisión (~69%) con una cobertura de secuenciación de 1X y una cobertura mínima de mapeo de 3FD (**Figura 19**). Sin embargo, esta combinación resultó en una sensibilidad relativamente baja (~10%) (**Figura 20**). La máxima sensibilidad se alcanzó con una cobertura de secuenciación de 5X y una profundidad mínima de mapeo de 1FD (~81%) (**Figura 20**). Esta combinación de parámetros (5X y 1FD), además, mostró una precisión relativamente alta de aproximadamente el 63% (**Figura 19**).

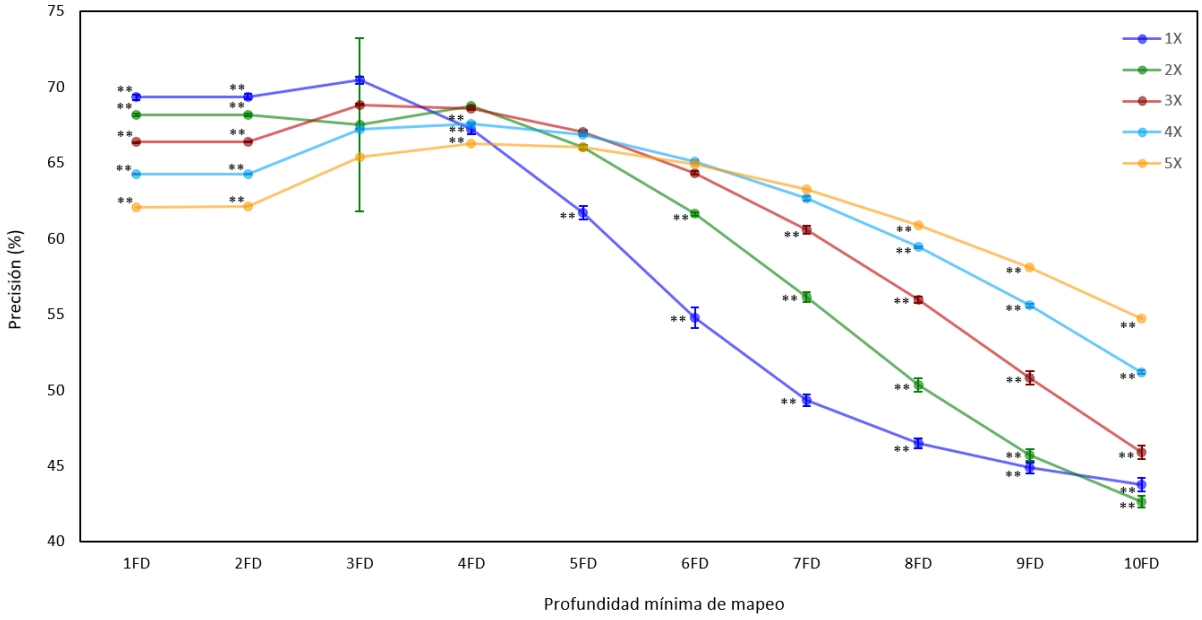


Figura 19. Precisión alcanzada con cada cobertura de secuenciación (1-5X) a la hora de identificar polimorfismos para cada profundidad mínima de mapeo (1-10FD). Las diferencias estadísticamente significativas entre las coberturas de secuenciación para cada profundidad mínima se encuentran marcadas con "**" para $p < 0,01$ con el método LSD.

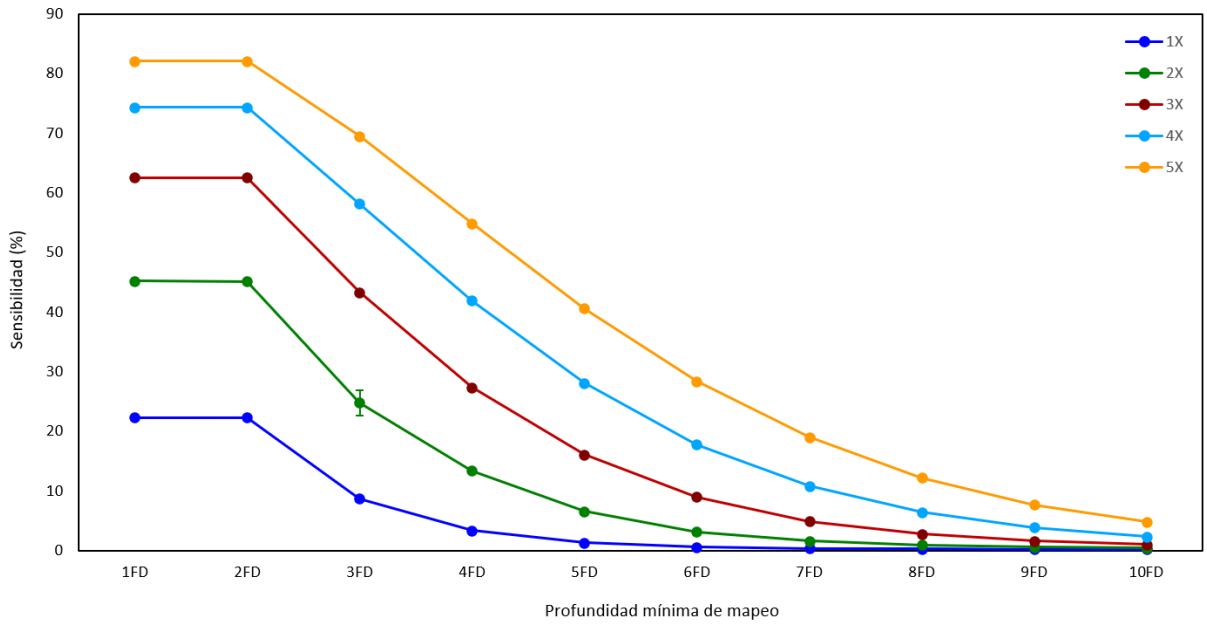


Figura 20. Sensibilidad alcanzada con cada cobertura de secuenciación (1-5X) a la hora de identificar polimorfismos para cada profundidad mínima de mapeo (1-10FD). Existen diferencias significativas entre las coberturas de secuenciación para cada profundidad mínima para $p < 0,01$ con el método LSD.

5. DISCUSIÓN

En el presente estudio se ha demostrado la viabilidad de la secuenciación del genoma completo a bajas coberturas en berenjena, una planta con un genoma diploide de tamaño considerable. Inicialmente, se examinó cómo diferentes coberturas de secuenciación (1-5X) influyen en parámetros de mapeo tales como el porcentaje de lecturas mapeadas, el ratio de duplicados, la cobertura del genoma y la profundidad máxima de las lecturas mapeadas, entre otros (**Tabla 4**). A continuación, se realizó una evaluación del efecto de diferentes combinaciones de coberturas de secuenciación (1-5X) y profundidades mínimas de mapeo (1-10FD) en la tasa de detección de polimorfismos (**Figura 18**). Utilizando la comparación con un estándar de referencia resultado de la resecuenciación del mismo genotipo a 20X como método de validación de los polimorfismos identificados, se encontró que una cobertura de secuenciación de 1X junto con un filtrado a una profundidad de mapeo de 3FD puede llegar a ser adecuada para genotipar un número suficiente de polimorfismos con alta precisión dependiendo del objetivo del estudio (**Figura 19**).

5.1. CALIDAD DE LAS LECTURAS SECUENCIADAS

Los resultados sugieren que tanto la resecuenciación a 5X como a 20X proporcionaron datos de alta calidad para el análisis genómico de la berenjena (**Figura 9 y 10**), lo cual es imprescindible a la hora de generar información reproducible y sólida. Uno de los indicadores de calidad a tener en cuenta en este tipo de estudios es el ratio de duplicación, ya que un alto porcentaje de lecturas duplicadas puede dar lugar a resultados sesgados o inexactos en la identificación de polimorfismos (Díaz-Arce and Rodríguez-Ezpeleta, 2019; Rochette *et al*., 2023). En el estudio llevado a cabo por Andrews *et al.* (2016) se comprobó que tanto el número de ciclos empleados en la PCR durante la preparación de la librería como el uso de diferentes cantidades de material genético de partida tienen un efecto sobre la presencia de un mayor o menor número de secuencias duplicadas. Como se muestra en la **Figura 13** y se resume en la **Tabla 4**, el ratio de duplicación de las lecturas obtenidas en la resecuenciación a 20X es superior al observado en la resecuenciación a 5X. Esto se debe a que existe un número finito de lecturas de secuencias completamente únicas que se pueden obtener de una muestra de ADN a partir del cual, cuanto mayor sea la cobertura de secuenciación, mayor será el ratio de lecturas duplicadas (Schweyen *et al*., 2014). Además, el porcentaje de lecturas duplicadas identificadas con FastQC (**Figura 13**) no coincide con el ratio de duplicación calculado por QualiMap (**Tabla 4**), puesto que cada programa utiliza diferentes métodos y criterios para la identificación de este tipo de secuencias. FastQC considera exclusivamente datos de secuenciación de extremo único, siendo el análisis de una sola lectura insuficiente para determinar duplicados. Además, únicamente analiza los primeros 50 nucleótidos de las primeras 100.000 lecturas de cada archivo, y extrapola las tasas de duplicación a partir de este número limitado de lecturas (https://www.bioinformatics.babraham.ac.uk/projects/fastqc). Por su parte, QualiMap estima niveles de duplicación de lecturas mucho más realistas al incorporar lecturas de extremos pareados tras haber sido mapeadas contra el genoma de referencia (García-Alcalde *et al.*, 2012). Es por ello que se recomienda utilizar QualiMap, al igual que otras herramientas que utilicen lecturas de extremos pareados, como HTSream (https://github.com/s4hts/HTStream) y FASTP (Chen *et al.*, 2018b), para estimar con un mayor nivel de confianza el porcentaje de lecturas duplicadas y otros parámetros de calidad (Therkildsen and Palumbi, 2017; Chaudhary *et al.*, 2023; Hooker *et al.*, 2023).

A la hora de analizar las lecturas generadas en el proceso de secuenciación, es también importante analizar el contenido en GC de cada muestra. El contenido medio en GC en el conjunto de lecturas procedentes tanto de la resecuenciación a 5x como de la resecuenciación a 20X fue de aproximadamente el 37% (**Figura 11**). En base al ensamblaje del genoma de berenjena a nivel cromosómico de alta calidad realizado por Wei *et al.* (2020), el contenido medio en GC del genoma es del 35,94%, similar al de *Arabidopsis* (36,06%) (The Arabidopsis Genome Initiative, 2000) y al del

tomate (34,05%) (The Tomato Genome Consortium, 2012). Los resultados obtenidos sugieren, por tanto, que no se produjo ningún tipo de sesgo o artefacto durante la preparación de la muestra ni durante el proceso de secuenciación, lo que indica que las lecturas resultantes eran de alta calidad y podían ser utilizadas en los análisis bioinformáticos posteriores sin preocupación por posibles errores en los datos. Estos resultados aseguran la integridad y fiabilidad de los resultados obtenidos en el estudio.

5.2. IMPLICACIONES DE LA COBERTURA DE SECUENCIACIÓN EN EL ALINEAMIENTO CONTRA EL GENOMA DE REFERENCIA

La disponibilidad del genoma de referencia de berenjena 67/3 (Barchi *et al.*, 2021) ha permitido llevar a cabo el alineamiento de las lecturas de secuenciación contra el mismo. Es importante tener en cuenta que, en los experimentos de secuenciación, una parte de las lecturas no se mapea contra el genoma de referencia (unmapped reads), ni siquiera después de realizar el paso previo de filtrado para eliminar las lecturas que presentan una baja calidad de bases o un alto número de bases desconocidas (Gouin *et al.,* 2015; Sim *et al.,* 2018; Liang *et al.*, 2019). Por ejemplo, en el caso de la resecuenciación a 5X, se mapeó el 98,54% de las lecturas, mientras que en el caso de la resecuenciación a 20X, lo hizo el 88,72% del total de las lecturas (**Tabla 4**). Una de las causas que podría justificar esta diferencia es el procesamiento previo de las lecturas. La compañía encargada de realizar la resecuenciación a 5X también llevó a cabo el paso de filtrado de las secuencias. Sin embargo, en el caso de la resecuenciación a 20X, las lecturas no fueron procesadas por la empresa contratada y llegaron como lecturas crudas, lo que condujo a la realización del filtrado en nuestro laboratorio. De este modo, es posible que las diferencias en el uso de distintos programas y en la configuración de parámetros hayan contribuido a estas variaciones.

Otras razones que pueden justificar que ciertas lecturas no se alineen con el genoma de referencia son: (1) variantes estructurales y sustituciones debidas a la variación genética individual, o a la distancia genética entre la muestra y las posiciones homólogas de la referencia que exceden el límite de huecos y desajustes permitidos por el algoritmo del software empleado en el mapeo, (2) errores de secuenciación, o (3) contaminación de la muestra (Barchi *et al.,* 2019a; Hasan *et al*., 2019; Valiente-Mullor *et al.,* 2021). Por ejemplo, Gramazio *et al.* (2019) atribuyeron la baja proporción de mapeo observada en la accesión ASI-S-1 de *S. melongena* a la presencia de contaminación en forma de lecturas mitocondriales, debido posiblemente a una eliminación insuficiente de mitocondrias durante la extracción de ADN genómico. En cuanto al algoritmo de Bowtie2, de forma predeterminada realiza un alineamiento de lecturas de extremo a extremo buscando alineamientos que involucren todos los caracteres de la lectura (Langmead and Salzberg, 2012; Langmead *et al.*, 2019). Se considera que un alineamiento es válido en base a una puntuación mínima función de la longitud de la lectura. La función de puntuación mínima (f) por defecto es: f(x) = -0,6 + (-0,6) * x, donde 'x' es la longitud de la lectura (https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml). Por tanto, la puntuación que han de superar los alineamientos realizados en el presente estudio es de -90,6, siendo 0 la mejor puntuación de alineamiento posible en el modo de extremo a extremo, que ocurre cuando no hay diferencias entre la lectura y la referencia. Cuando se producen desajustes entre las lecturas y el genoma de referencia, por defecto, si la base se encuentra en una posición de alta calidad en la lectura (Q40), recibe una penalización de -6 de manera predeterminada. En el caso de un indel de dos nucleótidos consecutivos, este recibe una penalización de -11 (-5 por la apertura del hueco, -3 por la primera extensión y -3 por la segunda) (https://bowtie-bio.sourceforge.net/bowtie2/manual.shtml; Musich, 2020). Por tanto, un alineamiento que presente un SNP y un indel de longitud dos nucleótidos, recibirá una puntuación general de -17, la cual se encuentra por encima de la puntuación mínima permitida para considerar un alineamiento válido entre el genoma de referencia y las lecturas con una

longitud de 150 pb. Sin embargo, si el número de desajustes es mucho mayor y, por tanto, la puntuación general cae por debajo de -90,6, el alineamiento no será considerado válido ni suficientemente bueno para ser informado.

El número de desajustes entre una lectura y la referencia tiene un impacto significativo en la MAPQ, así como el hecho de que la lectura se alinee contra una única región del genoma o lo haga contra múltiples regiones. Las lecturas que se mapean exclusivamente contra una región del genoma (unireads) pueden alcanzar una alta MAPQ, siempre que sean completamente homólogas a la referencia (MAPQ = 42). De lo contrario, la calidad de mapeo puede disminuir hasta MAPQ = 0 (John, 2014). Por otro lado, las lecturas que se alinean contra múltiples regiones, pero obtienen una puntuación más alta cuando se alinean contra una región específica (maxireads), pueden lograr una MAPQ de hasta 39, si no presentan ningún tipo de desajuste con respecto a la referencia (John, 2014). En contraste, las lecturas que se alinean contra más de una región del genoma con la misma puntuación de alineamiento (multireads), solo pueden recibir MAPQ de 0 o 1, dependiendo del número de desajustes (John, 2014). En base a esta información, la menor calidad de mapeo observada en los alineamientos contra las primeras posiciones del genoma (**Figura 14**) puede explicarse debido a que éstas forman parte del cromosoma 0, que consiste principalmente en secuencias repetitivas que no se pudieron ensamblar durante la construcción del genoma de referencia. Por tanto, las lecturas que se hayan mapeado contra este cromosoma, lo habrán hecho contra varias regiones, lo que contribuye a la baja calidad de los mapeos. De manera similar, el resto de mínimos de calidad de mapeo alcanzados a lo largo del genoma (**Figura 14**) podrían corresponderse con alineamientos de lecturas contra secuencias de los centrómeros y/o de los telómeros de los cromosomas, ya que estas regiones se caracterizan por presentar altos porcentajes de secuencias repetitivas.

En lo que respecta a la cobertura del genoma mapeado, esta se incrementa con la cobertura de secuenciación (**Tabla 4**). El porcentaje del genoma cubierto a una cobertura de 1X es de tan sólo el 45,11%, mientras que aumenta al 85,54% y 90% para las coberturas de 5X y 20X, respectivamente (**Tabla 4**). Estos resultados respaldan la evidencia de que, debido al proceso de secuenciación aleatoria, algunos sitios son secuenciados repetidamente (**Figura 16**), mientras que otros no son secuenciados en absoluto, lo que resulta en la presencia de múltiples puntos de datos faltantes en los datos generados por SWGR. En 2015, Lu *et al*. ya observaron que, al utilizar una cobertura media de 0,3 lecturas por sitio y por muestra en la secuenciación de múltiples líneas de maíz con el propósito de construir un pangenoma mediante el ensamblaje de las lecturas no alineadas con el genoma de referencia, la profundidad de mapeo seguía una distribución de Poisson (**Figura 15**), lo que implica que las lecturas no se distribuyen uniformemente en todo el genoma. Como resultado, llegaron a la conclusión de que el SWGR no puede garantizar una cobertura completa del genoma. No obstante, a pesar de la falta de genotipos en los datos resultantes del SWGR, éstos pueden ser empleados en una amplia gama de análisis (Lou *et al*., 2021), ya que la aceptabilidad de los datos faltantes varía entre estudios según la aplicación prevista. Yao *et al.* (2020b) demostraron que la secuenciación a bajas coberturas (~0,019X) de ADN nuclear de especímenes de *Macaca fascicularis* proporciona una cobertura suficiente del genoma (~1,7%) para realizar análisis genéticos de poblaciones y examinar patrones filogeográficos amplios en una muestra de gran tamaño. Sin embargo, también apuntan que, en el caso de querer distinguir entre poblaciones muy estrechamente relacionadas, haría falta una mayor cobertura del genoma y profundidad de secuenciación. Por su parte, Adhikari *et al.* (2022) utilizaron coberturas de secuenciación tan bajas como en el anterior estudio, de 0,01X a 0,03X, para identificar el tamaño de los segmentos de la cebada (*Hordeum vulgare*) y del triticale salvaje (*Thinopyrum intermedium*) introgresados en trigo, y determinar la dosis cromosómica. Además, comprobaron que una cobertura tan baja como 0,01X es suficiente para genotipar poblaciones DH o RIL de trigo.

La proporción de lecturas mapeadas y la cobertura del genoma de referencia también muestran variabilidad dependiendo de la referencia utilizada para el mapeo (Valiente-Mullor *et al.*, 2021). Hasta finales de 2020, se habían generado y publicado 1.031 genomas de referencia o borradores para un total de 788 especies de plantas (Sun *et al.*, 2022). La disponibilidad de estas secuencias de genomas de plantas, particularmente aquellas de alta calidad, ha proporcionado una valiosa herramienta para llevar a cabo estudios en áreas como la genómica funcional y la genética de poblaciones. Los individuos seleccionados para la secuenciación del genoma de referencia son frecuentemente elegidos por razones históricas. En berenjena, se seleccionó como referencia la línea endogámica '67/3', desarrollada a partir de un cruce intraespecífico entre el cultivar 'Purpura' y 'CIN2', seguido de nueve ciclos de autofecundación, y utilizada como progenitor masculino de una población de mapeo RIL formada por 167 líneas F6 RIL, cuyo progenitor femenino es la línea endogámica '305E40' (Barchi *et al.,* 2019a; Lanteri and Barchi, 2019). Sin embargo, utilizar únicamente un genoma de referencia no es suficiente para abarcar la diversidad genética de una especie y resulta inadecuado para muchos propósitos. De hecho, durante el proceso de alineamiento puede surgir un sesgo de referencia, que se refiere a la tendencia de algunas lecturas a ser mapeadas con mayor facilidad contra los alelos de referencia, mientras que las lecturas con alelos alternativos pueden no ser mapeadas o serlo a tasas más bajas (Ballouz *et al.*, 2019). Este problema se evita mediante el empleo de pangenomas, que es como se denomina a las colecciones de todas las secuencias de ADN que se encuentran en una especie (Della Coletta *et al.*, 2021; Sun *et al.*, 2022). El primer pangenoma de plantas publicado se obtuvo mediante la comparación de ensamblajes completos del genoma de siete individuos de soja silvestre, lo cual reveló la presencia de genes variables asociados con características como la composición de semillas, el tiempo de floración y madurez, entre otros (Li *et al.*, 2014). En berenjena se construyó el primer pangenoma a partir de la resecuenciación de 24 accesiones de *S. melongena*, una de *S. insanum* y una última de *S. incanum*. Finalizado el proceso de ensamblaje, se obtuvieron alrededor de 53 Mb de secuencias adicionales, incluyendo 816 genes codificantes de proteínas ausentes en el genoma de referencia 67/3 (Barchi *et al*., 2021). En resumen, el empleo de pangenomas como referencias representa un avance significativo que beneficiará diversos análisis genómicos. Específicamente, el uso de esta colección de genomas mejora notablemente la precisión del mapeo de secuencias cortas en comparación con una única referencia, lo que se traduce en una identificación de polimorfismos de mayor calidad (Gage *et al*., 2019; Gao *et al*., 2019; Qin *et al.*, 2021).

Al igual que se puede considerar cambiar los parámetros empleados en el alineamiento, por ejemplo, el número de desajustes entre las lecturas y la referencia, y seleccionar un genoma de referencia diferente, también se puede elegir el software de mapeo de lecturas a utilizar. Hasta la fecha, se han realizado diferentes estudios comparativos de múltiples alineadores de datos de secuenciación: Bowtie, Bowtie2, BBMAP, BWA, BWA-MEM, GEM, HISAT2, MimiMap2, Novoalign, TopHat2 (Donato *et al*., 2021; Musich *et al.*, 2021). En la mayoría de estos estudios, el software BWA-MEM mostró el mayor porcentaje de lecturas de extremos pareados alineadas contra el genoma de referencia, capaz de mapear aproximadamente 4,22 veces más lecturas por muestra que Bowtie2 (Wu *et al.*, 2019b; Yan *et al.*, 2021). No obstante, si se ejecuta Bowtie2 en el modo local, el cual permite que algunos de los caracteres finales de la lectura no participen en el alineamiento, se pueden llegar a obtener ratios de alineamiento tan altos como con BWA (Musich, 2020; Musich *et al.*, 2021). En cuanto a la precisión en el mapeo, el porcentaje de lecturas mapeadas correctamente es ligeramente superior al obtenido con Bowtie2; sin embargo, con el software Subread se consigue una precisión aún mayor (Yan *et al.*, 2021). La cobertura del genoma de referencia que se consigue con BWA también es superior, aunque solo ligeramente superior a la que se consigue con Bowtie2 (Musich *et al*., 2021). Finalmente, en cuanto al tiempo requerido para llevar a cabo el alineamiento, recurso muy importante a tener también en cuenta, SOAP2 es considerado el alineador más rápido en comparación con BWA-MEM y

Bowtie2 (Wu *et al.*, 2019b), al igual que Minimap2 y Accel-Align (Yan *et al.*, 2021), y que HISAT2 (Keel and Snelling, 2018). Normalmente, decidir qué alineador presenta un mayor rendimiento se reduce a elegir la herramienta que logra un mejor equilibrio entre velocidad y precisión. En el estudio llevado a cabo por Wu *et al.* (2019b), se llegó a la conclusión de que, BWA-MEM, frente a SOAP2 y Bowtie2, presenta una sensibilidad y precisión mayor, lo que hace que sea, probablemente, el algoritmo más adecuado para el mapeo de lecturas procedentes del genoma de plantas en diferentes estudios genómicos. Así mismo, Musich (2020) concluyó que tanto Bowtie2 como BWA son opciones adecuadas como mapeadores gracias a sus altos ratios de mapeo y cobertura del genoma. No obstante, la notable ventaja en términos de ejecución de BWA hace que esta herramienta sea preferible.

5.3. INFLUENCIA DE LA COBERTURA DE SECUENCIACIÓN Y DE LA PROFUNDIDAD MÍNIMA DE MAPEO EN LA IDENTIFICACIÓN DE POLIMORFISMOS

La mejora de la precisión en la detección de polimorfismos mediante la eliminación computacional de duplicados de lecturas al reducir el sesgo y el ruido de la PCR, o su posible disminución al eliminar datos informativos, aún es objeto de debate. Varios estudios han evaluado el impacto de la supresión de duplicados en la identificación de variantes genómicas. Algunos resultados sugieren que eliminar los clones de PCR no afecta significativamente a la precisión de la detección de polimorfismos, siempre y cuando se alcance una alta profundidad de cobertura de mapeo (Ebbert *et al*., 2016; Euclide *et al*., 2020). Sin embrago, existe una concepción generalizada de que la eliminación de duplicados mejora la precisión del genotipo identificado, especialmente cuando la frecuencia de los clones es alta, ya que los errores de PCR podrían ser confundidos con verdaderos alelos (Andrews *et al.*, 2016, 2014). Investigaciones recientes han validado una correlación positiva entre la presencia de secuencias duplicadas y la identificación de falsos polimorfismos (Díaz-Arce and Rodríguez-Ezpeleta, 2019; Rochette *et al.*, 2023), y nuestro estudio ha confirmado esta asociación. La omisión del marcado o eliminación de las lecturas duplicadas resultó en una sobreestimación significativa en la detección de variantes (**Figura 17**). Estos hallazgos respaldan la hipótesis planteada a partir de los datos derivados del análisis de calidad de las lecturas secuenciadas (**Figura 13 y Tabla 4**), que sugieren que un alto porcentaje de estas lecturas puede generar sesgos en la identificación de polimorfismos. Sin embargo, a diferencia de lo postulado por Euclide *et al.* (2020), quienes afirmaron que, asegurando altas coberturas de mapeo, se minimizan los efectos de las secuencias duplicadas y la aparición de sesgos en los estudios genómicos, nuestros resultados mostraron diferencias significativas entre el número de polimorfismos al considerar o no las secuencias duplicadas en combinaciones que involucran una cobertura de secuenciación superior a 2X para todas las profundidades mínimas de mapeo evaluadas (**Figura 17**).

En este estudio, durante el procedimiento de detección de variantes genéticas, se identificaron los polimorfismos presentes en posiciones donde la calidad de base superaba el umbral de 20 en la escala de Phred y en lecturas con una calidad de mapeo superior también a 20. Como medida adicional, se aplicó un único filtro basado en la profundidad mínima de mapeo (1-10FD) para evitar un filtrado redundante de los polimorfismos teniendo en cuenta que en un paso posterior se iban a comparar con el estándar de referencia disponible de la resecuenciación del mismo genotipo a una cobertura de 20X. La relación directa ente la cobertura de secuenciación y el número de polimorfismos identificados en el genoma, representada en la **Figura 17**, ha sido documentada en numerosas investigaciones (Happ *et al*., 2019; Lou *et al*., 2021; Adhikari *et al*., 2022). En el estudio de Song *et al.* (2016), se registró un notable incremento en la proporción de polimorfismos detectados en el genoma de especies altamente heterocigotas a medida que la cobertura de secuenciación aumentaba de 5X a 18X. Por otra parte, en un estudio de secuenciación e identificación de polimorfismos en el genoma

de gallinas, se observó que el incremento de la cobertura de secuenciación hasta 20X aumentó significativamente el número de polimorfismos identificados por diferentes softwares, entre los que se encontraban Freebayes y GATK (Liu *et al.*, 2022). Sin embargo, a medida que la cobertura superaba los 20X, la tasa de identificación de polimorfismos se reducía. En particular, Freebayes alcanzó el máximo de polimorfismos detectados a 20X (Liu *et al.*, 2022). Deng *et al*. (2022a) también demostraron que la baja cobertura de secuenciación afectaba a las estimaciones de diversidad nucleotídica, ya que tuvo lugar una reducción del número de sitios informativos en comparación a los registrados con altas coberturas de secuenciación. Por ejemplo, con una cobertura de secuenciación de 3,5X se identificaron alrededor de 150.000 polimorfismos en el genoma de *Himalopsyche tibetana*, mientras que, con una cobertura de secuenciación de 12,5X se identificaron alrededor de 10 millones de polimorfismos. Un estudio similar se llevó a cabo en canola (*Brassica napus*) (Malmberg *et al.,* 2018). Además de las coberturas utilizadas en nuestro estudio, estudiaron el efecto de dos coberturas adicionales, 0,25X y 05X, junto con cuatro filtros de profundidad mínima (2-5DP). Los resultados de este análisis fueron concluyentes: se evidencia que a medida que se aumenta la cobertura de secuenciación utilizada y se disminuye la profundidad mínima de mapeo, se produce un incremento significativo en la cantidad de polimorfismos identificados en el genoma (Malmberg *et al*., 2018). Estos hallazgos también evidencian la importancia de considerar la profundidad mínima de mapeo al realizar análisis de secuenciación y genotipado, ya que el número de lecturas que respaldan un polimorfismo influye directamente en la exactitud con que éste es identificado (Yu and Sun, 2013; Mun *et al.*, 2015; Gramazio *et al.*, 2019). De hecho, cuanto mayor es el umbral de la profundidad de mapeo, menor es la tasa de error en el genotipado (Fountain *et al.*, 2016).

Tanto la cobertura de secuenciación como la profundidad mínima de mapeo desempeñan un papel crucial en la detección de polimorfismos heterocigotos. Para cada nivel de profundidad mínima de mapeo, se observa un incremento en el número de polimorfismos heterocigotos a medida que se aumenta la cobertura de secuenciación (**Tabla 5**). Esto se debe a que, por ejemplo, secuenciar un locus heterocigoto a una cobertura de 1X proporciona únicamente una lectura de secuencia, revelando uno de los alelos y, consecuentemente, el polimorfismo heterocigoto se clasificaría incorrectamente como homocigoto para el alelo de referencia o para el alelo alternativo. Por tanto, para que un polimorfismo heterocigoto sea considerado como tal, se requiere que al menos dos lecturas se alineen con el locus y, además, tiene que coincidir que una de las lecturas provenga de la secuenciación de un alelo y la otra del otro (Gorjanc *et al*., 2017). Estos resultados son consistentes con los informados por varios expertos, incluido Bayer *et al.* (2015), quienes notificaron una correlación de 0,81 entre el número de lecturas mapeadas y el número de SNPs heterocigotos por individuo. Esta correlación afecta, sobre todo, a las poblaciones altamente heterocigotas. En el caso de la berenjena, una especie autógama, la heterocigosidad promedio observada en su genoma es extremadamente baja. De hecho, se ha registrado una heterocigosidad inferior a 5,46% en varias accesiones de *S. melongena* (Barchi *et al*., 2021). Esto resulta en una ventaja, ya que secuenciar cada sitio una vez sería suficiente y mantendría los costos bajos (Gorjanc *et al.*, 2017). Lo mismo afirma Malmberg *et al.* (2018) para el caso de la canola, una especie altamente homocigota debido al uso de dobles haploides en los programas de mejora.

Las diferencias entre la heterocigosidad observada en las accesiones de *S. melongena* por Barchi *et al.* (2021) y la observada en nuestro estudio y en el realizado por Gramazio *et al.* (2019) podrían atribuirse a errores durante la secuenciación considerados variantes genéticas debido a la información limitada proporcionada por la baja cobertura empleada en los análisis, así como a la ausencia de filtros para discriminar entre genotipos. También se encontraron diferencias entre la heterocigosidad detectada en la muestra secuenciada a una cobertura de 20x (8%) (**Tabla 6**) y la observada secuenciando a coberturas de 1-5X, que varió desde el 6,14% hasta el 55,09% (**Tabla 5**). Estos resultados son una

evidencia de que el filtro de fracción mínima del alelo alternativo es crucial para evitar la clasificación errónea de los loci homocigotos para el alelo de referencia como polimorfismos heterocigotos. Este filtro se emplea para establecer un umbral mínimo de lecturas de secuenciación que respalden la presencia de una variante genética en relación con la secuencia de referencia, con el fin de determinar si un locus debe considerarse polimórfico. Adicionalmente, se consideraron otros criterios para identificar polimorfismos en las lecturas obtenidas a una cobertura de secuenciación de 20X. El genoma de berenjena se encuentra formado por una alta proporción de elementos repetitivos, representando aproximadamente el 73% de su longitud total (Hirakawa *et al.*, 2014; Li *et al.*, 2019). Desde una perspectiva computacional, estas repeticiones introducen ambigüedades tanto en el ensamblaje del genoma como en el proceso de alineamiento de las lecturas, lo que puede ocasionar errores en la interpretación de los resultados (Treangen and Salzberg, 2012). Una de las principales ambigüedades que introducen durante el proceso de alineamiento es la acumulación considerable de lecturas en regiones repetitivas. Por ejemplo, al alinear lecturas cortas obtenidas con Illumina y lecturas largas obtenidas con PacBio contra el genoma cloroplástico de *Potentilla micrantha*, la profundidad de mapeo promedio fue de 9.111X para las lecturas de Illumina y de tan sólo 320X para las lecturas de PacBio (Ferrarini *et al*., 2013). Esto se debe a que el uso de lecturas largas facilita el alineamiento de lecturas procedentes de regiones repetitivas y evita la acumulación de lecturas en regiones redundantes del genoma, disminuyendo, por tanto, la profundidad del mapeo, a diferencia de lo que ocurre con las lecturas cortas procedentes de Illumina (Sealfon *et al.*, 2012; Du and Liang, 2019). En este estudio, como consecuencia de lo comentado previamente, se han alcanzado profundidades de mapeo máximas de 443,20 y 33.380 lecturas/posición a una cobertura de secuenciación de 1X y 20X, respectivamente (**Tabla 4**). Según Ravinet and Meier (2021), una buena regla es filtrar por una cobertura máxima que sea igual al doble de la cobertura media. Es por esta razón que, para abordar este fenómeno, se implementó un filtro durante el proceso de identificación de SNPs que limitó la detección de polimorfismos a aquellos respaldados por un máximo de 40 lecturas en los datos de la resecuenciación a 20X.

Existen numerosos softwares disponibles para realizar el análisis de datos de secuenciación, y al igual que se han observado discrepancias entre las herramientas utilizadas para el alineamiento de las lecturas con el genoma de referencia, también se han encontrado diferencias entre los softwares empleados en la identificación de variantes genéticas. En un estudio en el que se llevó a cabo una comparación entre los softwares 16GT, GATK, BCFtools-single, BCFtools-múltiple, VarS-can2-single, VarScan2-multiple y Freebayes, este último identificó el menor número de polimorfismos a altas coberturas de secuenciación (> 20X), mientras que BCFtools-múltiple identificó el mayor número de variantes genéticas a coberturas de 5X y 10X, y a coberturas superiores de 20X (Liu *et al.*, 2022). Por otra parte, SAMtools-mpileup y GATK-HC generaron resultados similares en cuanto al número total de SNPs identificados en el conjunto de datos genómicos de tomate (Wu *et al.*, 2019b). Según la simulación realizada, GATK-HC demostró una mayor capacidad para identificar variantes verdaderas con una mayor precisión. No obstante, en poblaciones con baja diversidad y cobertura de secuenciación (1X), se observó que SAMtools-mpileup identificó un mayor número de SNPs verdaderos (Wu *et al*., 2019b). Sobre otro conjunto de datos genómicos de partida, Freebayes generó la mayor cantidad de SNPs candidatos en los diferentes tipos de datos, aunque GATK pareció ser más sensible para la detección de indels (Ksouri *et al.*, 2022). La proporción de SNPs identificados por Freebayes que superaron el filtro de calidad fue del 9%, lo cual llevó a los autores a considerar a GATK-HaplotypeCaller como la herramienta más adecuada para detectar variantes genéticas con alta calidad en comparación con BCFtools y Freebayes (Ksouri *et al.,* 2022). En cuanto a la sensibilidad de las diferentes herramientas informáticas, 16GT mostró una mayor sensibilidad cuando las profundidades de lectura fueron iguales o superiores a 20X, mientras que Freebayes mostró una sensibilidad

moderada a profundidades de secuenciación más bajas en comparación con BCFtools, VarScan2 y GATK (Liu *et al.*, 2022). Entre estos cuatro softwares, BCFtools tuvo una mayor precisión a cualquier profundidad de lectura y Freebayes mostró la menor precisión a profundidades de lectura altas (> 30X) (Liu *et al.*, 2022). Los resultados de estas comparaciones revelan que no existe un software universal para la identificación de variantes genómicas, resaltando la importancia de seleccionar una herramienta apropiada acorde a los objetivos del experimento.

5.4. EVALUACIÓN DE LA PRECISIÓN Y SENSIBILIDAD DE LA TÉCNICA EN EL GENOTIPADO DE LA BERENJENA

Las bajas coberturas de secuenciación pueden ocasionar desviaciones en la identificación de polimorfismos debido a los errores introducidos y amplificados durante la PCR en la etapa de preparación de la biblioteca de secuenciación. Por consiguiente, resulta fundamental llevar a cabo una validación de los polimorfismos identificados. En este contexto, el empleo de un estándar de referencia permite reducir el número de falsos positivos y preservar de manera óptima las variantes genéticas verdaderas. Su uso se encuentra muy extendido, sobre todo, en genética clínica, donde se requiere una identificación precisa de polimorfismos asociados a enfermedades para alcanzar altos rendimientos diagnósticos o pronósticos (Gargis *et al.*, 2016; Roy *et al.*, 2018). En cambio, en genética vegetal es una técnica que se está empezando a implementar como medida de confianza debido al éxito del SWGR. En nuestro estudio se ha utilizado un estándar de referencia compuesto por un total de 705.954 polimorfismos. Los resultados mostraron que a profundidades mínimas de mapeo inferiores a 4FD, se produce un aumento en la precisión (**Figura 19**) a expensas de una menor sensibilidad de la técnica (**Figura 20**) a medida que se reduce la cobertura de secuenciación. Por otro lado, a profundidades mínimas de mapeo superiores a 6FD, tanto la precisión como la sensibilidad de la técnica disminuyen conforme lo hace la cobertura de secuenciación (**Figura 19 y 20**). Deng *et al.,* 2022a en su estudio sobre dos especies de tricópteros, Song *et al*., 2016 en su estudio sobre datos genómicos de *Crassostrea gigas,* y Hardwick *et al.,* 2017 en su revisión sobre el uso de estándares de referencia en la evaluación de los datos generados por NGS, también destacaron una disminución en la sensibilidad a la hora de detectar polimorfismos al reducir la cobertura de secuenciación. Además, resaltaron que la reducción en la cobertura de secuenciación también afecta negativamente la precisión con que las variantes genéticas son identificadas (Song *et al.,* 2016). No obstante, es posible lograr una alta sensibilidad y precisión empleando bajas profundidades de secuenciación al combinar información de un amplio número de individuos (Li *et al.,* 2011; Han *et al*., 2014; Deng *et al.*, 2022a).

A la luz de los datos y las observaciones recopiladas, si el objetivo del estudio exige la identificación de un elevado número de marcadores moleculares, habría que secuenciar a coberturas superiores a 4X y aplicar un filtrado a profundidades mínimas de mapeo lo más bajas posible, siempre y cuando se pueda mantener la capacidad de discriminación entre polimorfismos homocigotos y heterocigotos (**Figura 20**). En cambio, si el estudio requiere una elevada proporción de polimorfismos verdaderos a costa de un menor número de variantes genéticas, habría que secuenciar a coberturas de 4-5X y filtrar a coberturas mínimas de mapeo intermedias (2-6FD) (**Figura 19**). En el caso de utilizar coberturas bajas (1-3X), aunque se obtenga una mayor precisión al filtrar con profundidades mínimas de mapeo bajas (**Figura 19**), sería conveniente aplicar también un filtrado con profundidades intermedias (> 5FD) para poder distinguir entre polimorfismos homocigotos, heterocigotos o errores de secuenciación.

Nuestros resultados difieren parcialmente de los publicados por Malmberg *et al.* (2018). En primer lugar, aseguraron que, al aplicar un filtro de profundidad estricto, se observa un aumento en la precisión de la técnica a coberturas de secuenciación inferiores a 2X. En segundo lugar, afirmaron que, cuando la cobertura de secuenciación es superior a 2X, la precisión disminuye conforme aumenta el umbral de profundidad mínima (Malmberg *et al.*, 2018). En nuestro estudio, observamos un aumento

en la precisión de la técnica para todas las coberturas de secuenciación, hasta alcanzar un umbral específico de profundidad mínima en cada caso, a partir del cual comienza a disminuir (**Figura 19**). Estas diferencias pueden ser atribuidas a varias razones. En primer lugar, el estudio realizado por Malmberg *et al.* (2018) se enfocó en datos genómicos de canola, lo cual puede influir en los resultados obtenidos al tratarse de especies distintas. En segundo lugar, como estándar de referencia se empleó un conjunto de variantes genéticas identificadas en datos de secuenciación a 10X sometidos a un filtrado con diferentes profundidades mínimas, coincidiendo con la profundidad de filtrado de la muestra. Por el contrario, en nuestro estudio se utilizó un único estándar de referencia que fue filtrado a una profundidad mínima de mapeo de 10X, umbral que consideramos necesario y suficiente para lograr una alta precisión en la identificación de polimorfismos (Song *et al.,* 2016).

6. INVESTIGACIONES FUTURAS

En este estudio, se ha desarrollado un método de evaluación del impacto de diferentes combinaciones de parámetros, tales como la cobertura de secuenciación y la profundidad mínima de mapeo, en la detección de variantes genéticas en el genoma de berenjena. Para poner a punto el método, hemos llevado a cabo un ensayo preliminar con un genotipo de berenjena empleado como parental de una población MAGIC (Gramazio *et al*., 2019). El siguiente paso consistiría en utilizar este mismo protocolo de evaluación con el resto de los progenitores utilizados en el desarrollo de la población. En base a los resultados obtenidos, se seleccionaría la cobertura de secuenciación óptima para genotipar los diferentes individuos que conforman la población final y poder así obtener una representación genómica de la misma.

Debido a la baja sensibilidad del SWGR, a menudo se requiere un paso adicional para inferir los genotipos faltantes o de baja confianza (Rubinacci *et al.*, 2021). Este proceso, conocido como imputación, se ha vuelto fundamental en los estudios de asociación a nivel genómico para mejorar la capacidad de detección de QTL, y en la mayoría de los casos implica el uso de un panel de referencia de genotipos de alta densidad (Marchini and Howie, 2010). Se han desarrollado paneles de referencia para *Arabidopsis*, maíz, arroz y soja (Bukowski *et al*., 2018; Cao *et al*., 2011; The 3.000 rice genomes project, 2014; Torkamaneh *et al*., 2018a). Sin embargo, no se encuentran disponibles para la mayoría de especies no modelo. Un enfoque para mitigar esta limitación, cuando se trabaja con poblaciones experimentales, es genotipar los parentales empleando altas coberturas de secuenciación y utilizarlos como paneles de referencia, mejorando la resolución del mosaico genómico de la progenie *(Bayer et al*., 2015; Cericola *et al.*, 2018).

Uno de los factores más críticos que afecta la precisión y fiabilidad de la imputación es la proporción de SNPs identificados en la muestra o marcadores objetivo y la densidad de polimorfismos presentes en el panel de referencia (Deng *et al*., 2022b). Por un lado, una población de referencia más grande puede proporcionar más haplotipos de referencia, lo que facilita la coincidencia con los marcadores objetivo. Del mismo modo, una mayor densidad de polimorfismos identificados en la muestra también contribuye a una mejor imputación (Deng *et al.*, 2022b; Torkamaneh *et al*., 2018b). Sin embargo, en los últimos años se ha observado un creciente interés en evaluar la precisión de la imputación en muestras secuenciadas a bajas coberturas. Zan *et al*. (2019) observaron que no hubo mejoras evidentes en la precisión de la imputación de genotipos a partir de coberturas de secuenciación superiores a 0,5X. Así mismo, la resecuenciación con coberturas de 0,5X y 1X, seguida de la imputación, representa una alternativa más eficiente al uso de matrices de SNPs con muestras humanas (Rubinacci *et al*., 2021). En el caso de la soja, se ha estudiado el impacto de coberturas de secuenciación inferiores a 1X sobre la especificidad de la técnica en 114 líneas. Se observó que la secuenciación a baja cobertura acompañada de la imputación a partir de un panel de referencia puede extenderse por debajo de una profundidad de 1X hasta una de 0,3X para identificar polimorfismos con una alta precisión (Happ *et al*., 2019). Por lo tanto, queda por determinar el potencial de la secuenciación a baja cobertura con imputación como un enfoque económico para obtener información de polimorfismos en berenjena.

7. CONCLUSIONES

El genotipado por secuenciación del genoma completo a bajas coberturas se perfila como una técnica prometedora para obtener información genómica a gran escala, ofreciendo una alternativa económica a los métodos tradicionales de secuenciación. Sin embargo, su aplicación se encuentra limitada debido a la necesidad de contar con un genoma de referencia y a la incertidumbre asociada a su capacidad para distinguir entre falsos positivos, falsos negativos y verdaderos polimorfismos, y entre genotipos homocigotos y heterocigotos.

En este trabajo se ha realizado una evaluación exhaustiva del impacto que tiene tanto la cobertura de secuenciación como la profundidad mínima de mapeo en la precisión y sensibilidad del genotipado por secuenciación a bajas coberturas en un genotipo de berenjena. La precisión de la técnica puede verse afectada por la calidad de los datos y la presencia de errores de secuenciación. Es por ello que resulta crucial garantizar la alta calidad de las secuencias, incluyendo una calidad de Phred alta, una distribución normal del contenido en GC, y una baja o nula tasa de lecturas duplicadas para continuar con el análisis. En el proceso de alineamiento de las lecturas con el genoma de referencia, uno de los parámetros más críticos que afecta los pasos posteriores es la tasa de mapeo. En este estudio, aunque la tasa de lecturas mapeadas no alcanzó el 100%, se logró una aproximación cercana, lo cual respalda la alta calidad de las lecturas obtenidas. Por otro lado, para mitigar la influencia de la sobrerrepresentación de ciertas regiones del genoma en la detección de polimorfismos, resulta necesario restringir la profundidad máxima, siguiendo el enfoque empleado en el estándar de referencia utilizado. Los resultados del mapeo ponen de manifiesto que con una cobertura de secuenciación de 1X se puede alcanzar una cobertura del genoma de aproximadamente el 50%, aunque con una pérdida alta de sitios informativos. Antes de proceder a realizar la identificación de variantes genéticas, es importante, tal y como se ha demostrado, la eliminación de las secuencias duplicadas, ya que su presencia puede generar artefactos y sobreestimar el número de polimorfismos. Por último, el incremento en la cobertura de secuenciación y la relajación del filtro de profundidad resultaron en un aumento en el número de variantes genéticas identificadas, tal y como se podía sospechar. Sin embargo, se ha enfatizado la necesidad de realizar un paso de validación de los polimorfismos. En este trabajo, la comparación con un estándar de referencia altamente fiable demostró que, si bien la sensibilidad disminuyó hasta en un 60% al reducir la cobertura de secuenciación de 5X a 1X con bajas profundidades de mapeo, esta diferencia se redujo al aplicar un filtro más restrictivo. En cuanto a la precisión, no se observaron diferencias significativas, obteniendo la mayor precisión con una cobertura de 1X y un filtro de profundidad mínima de mapeo de 3FD.

En base a estos resultados, se puede concluir que es posible emplear bajas coberturas de secuenciación (1-5X) en estudios genómicos de berenjena, siempre y cuando se aplique un filtrado estricto para evitar errores en la asignación de genotipos. Este ensayo piloto *in silico* tan sólo es un primer paso en la puesta a punto de la técnica de genotipado por secuenciación a bajas coberturas en berenjena. De hecho, proporciona información valiosa para el diseño de ensayos de genotipado de alto rendimiento, promoviendo así los estudios genéticos y de mejora. Sin embargo, la mejor combinación de cobertura de secuenciación y filtrado de variantes debe ser evaluada caso por caso, teniendo en cuenta los recursos disponibles y los objetivos del estudio.

8. BIBLIOGRAFÍA

Adhikari, L., Shrestha, S., Wu, S., Crain, J., Gao, L., Evers, B., Wilson, D., Ju, Y., Koo, D. H., Hucl, P. *et al*. 2022. A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. *Sci. Rep*., 12, 17583. doi: 10.1038/s41598-022-19858-2.

Akano, A.O., Dixon, A.G.O., Mba, C., Barrera, E., Fregene, M. 2002. Genetic mapping of a dominant gene conferring resistance to cassava mosaic disease. *Theor. Appl. Genet.*, 105, 521-525. doi: 10.1007/s00122-002-0891-7.

Altmann, A., Weber, P., Bader, D., Preuß, M., Binder, E.B., Müller-Myhsok, B. 2012. A beginners guide to SNP calling from high-Throughput DNA-sequencing data. *Hum. Genet.*, 131, 1541-1554. doi: 10.1007/s00439-012-1213-z.

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A. 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. *Nat. Rev. Genet*., 17(2), 81-92. doi: 10.1038/nrg.2015.28.

Andrews, K.R., Hohenlohe, P.A., Miller, M.R., Hand, B.K., Seeb, J.E., Luikart, G. 2014. Trade-offs and utility of alternative RADseq methods: Reply to Puritz *et al*. *Mol. Ecol*., 23(24), 5943-5946. doi: 10.1111/mec.12964.

Andrews, S. 2016. Quality Control tool for High Throughput Sequence Data [WWW Document]. URL http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/ (Acceso: 28/01/23).

Angiolillo, A., Mencuccini, M., Baldoni, L. 1999. Olive genetic diversity assessed using amplified fragment length polymorphisms. *Theor. Appl. Genet*., 98, 411-421. doi: 10.1007/s001220051087.

Aronesty, E. 2011. ea-utils: "Command-line tools for processing biological sequencing data" [WWW Document]. URL https://github.com/ExpressionAnalysis/ea-utils (Acceso: 28/01/23).

Aubriot, X., Singh, P., Knapp, S. 2016. Tropical Asian species show that the Old World clade of "spiny solanums" (*Solanum* subgenus *Leptostemonum* pro parte: Solanaceae) is not monophyletic. *Bot. J. Linn. Soc*., 181(2), 199-223. doi: 10.1111/boj.12412.

AVGRIS. 2022. The AVRDC Vegetable Genetic Resources Information System [WWW Document]. URL http://seed.worldveg.org/search/characterization/solanum_eggplant (Acceso: 22/12/22).

Ballouz, S., Dobin, A., Gillis, J.A. 2019. Is it time to change the reference genome? *Genome Biol.* 20(1), 1–9. doi: 10.1186/s13059-019-1774-4.

Barchi, L., Acquadro, A., Alonso, D., Aprea, G., Bassolino, L., Demurtas, O., Ferrante, P., Gramazio, P., Mini, P., Portis, E. *et al*. 2019b. Single Primer Enrichment Technology (SPET) for High-Throughput Genotyping in Tomato and Eggplant Germplasm. *Front. Plant Sci*., 10, 1005. doi: 10.3389/fpls.2019.01005.

Barchi, L., Pietrella, M., Venturini, L., Minio, A., Toppino, L., Acquadro, A., Andolfo, G., Aprea, G., Avanzato, C., Bassolino, L. *et al.* 2019a. A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Sci. Rep*., 9(1), 11769. doi: 10.1038/s41598-019-47985-w.

Barchi, L., Rabanus-Wallace, M.T., Prohens, J., Toppino, L., Padmarasu, S., Portis, E., Rotino, G.L., Stein, N., Lanteri, S., Giuliano, G. 2021. Improved genome assembly and pan-genome provide key insights into eggplant domestication and breeding. *Plant J.*, 107(2), 579-596. doi: 10.1111/tpj.15313.

Barchi, L., Toppino, L., Valentino, D., Bassolino, L., Portis, E., Lanteri, S., Rotino, G.L. 2018. QTL analysis reveals new eggplant loci involved in resistance to fungal wilts. *Euphytica*, 214, 1-15. doi: 10.1007/s10681-017-2102-2.

Batley, J., Edwards, D. 2009. Genome sequence data: Management, storage, and visualization. *Biotechniques*, 46(5), 333-336. doi: 10.2144/000113134.

Bayer, P.E., Ruperao, P., Mason, A.S., Stiller, J., Chan, C.K.K., Hayashi, S., Long, Y., Meng, J., Sutton, T., Visendi, P. *et al*. 2015. High-resolution skim genotyping by sequencing reveals the distribution of crossovers and gene conversions in *Cicer arietinum* and *Brassica napus*. *Theor. Appl. Genet*., 128, 1039-1047. doi: 10.1007/s00122-015-2488-y.

Bayés, M., Gut, I.G. 2011. Overview of Genotyping, in: Rapley, R., Harbron, S. (eds) *Molecular Analysis and Genome Discovery*, 2nd edn., Wiley-Blackwell, UK, pp 1-23. doi: 10.1002/9781119977438.ch1.

Bean, A.R. 2004. The taxonomy and ecology of *Solanum* subg. *Leptostemonum* (Dunal) Bitter (Solanaceae) in Queensland and far north-eastern New South Wales, Australia. *Austrobaileya*, 6(4), 639-816.

Bellucci, E., Bitocchi, E., Ferrarini, A., Benazzo, A., Biagetti, E., Klie, S., Minio, A., Rau, D., Rodriguez, M., Panziera, A., *et al*. 2014. Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. *Plant Cell*., 26(5), 1901-1912. doi: 10.1105/tpc.114.124040.

Bérénice, B., Maria, D., Alexandre, C., Anthony, B., Laura, L., Erwan, C., Stéphanie, R. 2022. Quality Control (Galaxy Training Materials) [WWW Document]. URL https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html (Acceso: 28/03/23).

Bergelson, J., Roux, F. 2010. Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet*, 11(12), 867-879. doi: 10.1038/nrg2896.

Bernardo, R. 2021. Multiparental populations in line development: Genetic gain, diversity, and practical limitations. *Crop Sci*, 61(6), 4139-4150. doi: 10.1002/csc2.20632.

Bohra, A. 2013. Emerging paradigms in genomics-based crop improvement. *Sci. World J.*, 585467. doi: 10.1155/2013/585467.

Bohs, L. 2005. Major clades in *Solanum* based on ndhF sequence data, in: Keating, R.C., Hollowell, V.C, Croat, T.B. (eds) *A festschrift for William G. D'Arcy: the legacy of a taxonomist, Monographs in Systematic Botany from the Missouri Botanical Garden*, Missouri Botanical Garden Press, St. Louis, pp 27–49.

Bolger, A.M., Lohse, M., Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics,* 30(15), 2114-2120. doi: 10.1093/bioinformatics/btu170.

Bossa-Castro, A.M., Tekete, C., Raghavan, C., Delorean, E.E., Dereeper, A., Dagno, K., Koita, O., Mosquera, G., Leung, H., Verdier, V., *et al*. 2018. Allelic variation for broad-spectrum resistance and susceptibility to bacterial pathogens identified in a rice MAGIC population. *Plant Biotechnol. J*., 16(9), 1559-1568. doi: 10.1111/pbi.12895.

Brenes, M., Solana, A., Boscaiu, M., Fita, A., Vicente, O., Calatayud, Á., Prohens, J., Plazas, M. 2020. Physiological and biochemical responses to salt stress in cultivated eggplant (*Solanum*

melongena L.) and in S. *insanum* L., a close wild relative. *Agronomy*, 10(5), 651. doi: 10.3390/agronomy10050651.

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., Wang, B., Xu, D., Yang, B., Xie, C., *et al.* 2018. Construction of the third-generation *Zea mays* haplotype map. *Gigascience*, 7(4), gix134. doi: 10.1093/gigascience/gix134.

Burr, B., Burr, F.A. 1991. Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. *Trends Genet.*, 7(2), 55-60. doi: 10.1016/0168-9525(91)90232-F.

Burr, B., Evola, S. V., Burr, F.A., Beckmann, J.S. 1983. The Application of Restriction Fragment Length Polymorphism to Plant Breeding, in: Setlow, J.K, Hollaender, A. (eds) *Genetic Engineering*, Springer, Boston, MA, pp 45-59. doi: 10.1007/978-1-4684-4556-5_4.

Burrows, M., Wheeler, D J.. 1994. A block-sorting lossless data compression algorithm. Technical Report 124, Digital Equipment Corporation.

Butiuc-Keul, A., Coste, A., Farkas, A., Cristea, V., Isac, V., Halmagyi, A. 2019. Molecular characterization of apple (*Malus* × *domestica* borkh.) genotypes originating from three complementary conservation strategies. *Turkish J. Agric. For*., 43(5), 464-477. doi: 10.3906/tar-1803-3.

Calonje, M., Martín-Bravo, S., Dobeš, C., Gong, W., Jordon-Thaden, I., Kiefer, C., Kiefer, M., Paule, J., Schmickl, R., Koch, M.A. 2009. Non-coding nuclear DNA markers in phylogenetic reconstruction. *Plant Syst. Evol.,* 282, 257-280. doi: 10.1007/s00606-008-0031-1.

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C. *et al*. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.,* 43(10), 956–963. doi: 10.1038/ng.911.

Cericola, F., Lenk, I., Fè, D., Byrne, S., Jensen, C.S., Pedersen, M.G., Asp, T., Jensen, J., Janss, L. 2018. Optimized use of low-depth genotyping-by-sequencing for genomic prediction among multi-parental family pools and single plants in perennial ryegrass (*Lolium perenne* L.). *Front. Plant Sci.*, 9, 369. doi: 10.3389/fpls.2018.00369.

Cericola, F., Portis, E., Lanteri, S., Toppino, L., Barchi, L., Acciarri, N., Pulcini, L., Sala, T., Rotino, G.L. 2014. Linkage disequilibrium and genome-wide association analysis for anthocyanin pigmentation and fruit color in eggplant. *BMC Genomics*, 15(1), 1-15. doi: 10.1186/1471-2164-15-896.

Chaudhary, R., Koh, C.S., Perumal, S., Jin, L., Higgins, E.E., Kagale, S., Smith, M.A., Sharpe, A.G., Parkin, I.A.P. 2023. Sequencing of *Camelina neglecta*, a diploid progenitor of the hexaploid oilseed *Camelina sativa*. *Plant Biotechnol. J.,* 21(3), 521–535. doi: 10.1111/pbi.13968.

Chen, S., Zhou, Y., Chen, Y., Gu, J. 2018b. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics,* 34(17), i884–i890. doi: 10.1093/bioinformatics/bty560.

Chen, Y., Chen, Yongsheng, Shi, C., Huang, Z., Zhang, Y., Li, S., Li, Y., Ye, J., Yu, C., Li, Z., *et al.* 2018a. SOAPnuke: A MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience*, 7(1), gix120. doi: 10.1093/gigascience/gix120.

Chung, Y.S., Choi, S.C., Jun, T.H., Kim, C. 2017. Genotyping-by-sequencing: a promising tool for plant genetics research and breeding. *Hortic. Environ. Biotechnol,* 58, 425-431. doi: 10.1007/s13580-017-0297-8.

Clot, C. R., Wang, X., Koopman, J., Navarro, A. T., Bucher, J., Visser, R. G. Finkers, R., van Eck, H. 2023. High-density linkage map constructed from a skim sequenced diploid potato population reveals transmission distortion and QTLs for tuber and pollen shed. *Potato Res.,* 1-25. doi: 10.21203/rs.3.rs-2302091/v1.

Cui, F., Li, J., Ding, A., Zhao, C., Wang, L., Wang, X., Li, S., Bao, Y., Li, X., Feng, D., *et al.* 2011. Conditional QTL mapping for plant height with respect to the length of the spike and internode in two mapping populations of wheat. *Theor. Appl. Genet*., 122, 1517-1536. doi: 10.1007/s00122-011-1551-6.

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. *et al*. 2011. The variant call format and VCFtools. *Bioinformatics,* 27, 2156–2158. doi: 10.1093/bioinformatics/btr330.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L. 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.*, 12(7), 499-510. doi: 10.1038/nrg3012.

Della Coletta, R., Qiu, Y., Ou, S., Hufford, M.B., Hirsch, C.N. 2021. How the pan-genome is changing crop genomics and improvement. *Genome Biol*., 22(1), 1-19. doi: 10.1186/s13059-020-02224-8.

Deng, T., Zhang, P., Garrick, D., Gao, H., Wang, L., Zhao, F. 2022b. Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. *Front. Genet*., 12, 2652. doi: 10.3389/fgene.2021.704118.

Deng, X.L., Frandsen, P.B., Dikow, R.B., Favre, A., Shah, D.N., Shah, R.D.T., Schneider, J. V., Heckenhauer, J., Pauls, S.U. 2022a. The impact of sequencing depth and relatedness of the reference genome in population genomic studies: A case study with two caddisfly species (*Trichoptera, Rhyacophilidae, Himalopsyche*). *Ecol. Evol*., 12(12), e9583. doi: 10.1002/ece3.9583.

Díaz-Arce, N., Rodríguez-Ezpeleta, N. 2019. Selecting RAD-Seq data analysis parameters for population genetics: The more the better? *Front. Genet*., 10, 533. doi: 10.3389/fgene.2019.00533.

Diaz, S., Ariza-Suarez, D., Izquierdo, P., Lobaton, J.D., de la Hoz, J.F., Acevedo, F., Duitama, J., Guerrero, A.F., Cajiao, C., Mayor, V., *et al.* 2020. Genetic mapping for agronomic traits in a MAGIC population of common bean (*Phaseolus vulgaris* L.) under drought conditions. *BMC Genomics,* 21(1), 1-20. doi: 10.1186/s12864-020-07213-6.

Diniz, W.J.S., Canduri, F. 2017. Bioinformatics: An overview and its applications. *Genet. Mol. Res*., 16(1), 17. doi: 10.4238/gmr16019645.

Donato, L., Scimone, C., Rinaldi, C., D'Angelo, R., Sidoti, A. 2021. New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies. *Neural Comput. Appl*., 33(22), 15669–15692. doi: 10.1007/s00521-021-06188-z.

Du, H., Liang, C. 2019. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nat. Commun*., 10(1), 5360. doi: 10.1038/s41467-019-13355-3.

Ebbert, M.T.W., Wadsworth, M.E., Staley, L.A., Hoyt, K.L., Pickett, B., Miller, J., Duce, J., Kauwe, J.S.K., Ridge, P.G. 2016. Evaluating the necessity of PCR duplicate removal from next-generation

sequencing data and a comparison of approaches. *BMC Bioinformatics,* 17, 491-500. doi: 10.1186/s12859-016-1097-3.

Eshed, Y., Zamir, D. 1994. A genomic library of *Lycopersicon pennellii* in *L. esculentum*: A tool for fine mapping of genes. *Euphytica*, 79, 175-179. doi: 10.1007/BF00022516.

Euclide, P.T., McKinney, G.J., Bootsma, M., Tarsa, C., Meek, M.H., Larson, W.A. 2020. Attack of the PCR clones: Rates of clonality have little effect on RAD-seq genotype calls. *Mol. Ecol. Resour*., 20(1), 66–78. doi: 10.1111/1755-0998.13087.

Ewing, B., Hillier, L.D., Wendl, M.C., Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*., 8(3), 175-185. doi: 10.1101/gr.8.3.175.

FAOSTAT. 2022. Database Collections [WWW Document]. Food Agric. Organ. United Nations. Roma, Ital. URL http://www.fao.org/faostat/en/#data/QCL (Acceso: 09/02/23).

Ferrarini, M., Moretto, M., Ward, J.A., Šurbanovski, N., Stevanović, V., Giongo, L., Viola, R., Cavalieri, D., Velasco, R., Cestaro, A., *et al*. 2013. An evaluation of the PacBio RS platform for sequencing and de novo assembly of a chloroplast genome. *BMC Genomics*, 14(1), 1-12. doi: 10.1186/1471-2164-14-670.

Fountain, E.D., Pauli, J.N., Reid, B.N., Palsbøll, P.J., Peery, M.Z. 2016. Finding the right coverage: the impact of coverage and sequence quality on single nucleotide polymorphism genotyping error rates. *Mol. Ecol. Resour.*, 16(4), 966–978. doi: 10.1111/1755-0998.12519.

Frodin, D.G. 2004. History and concepts of big plant genera. *Taxon,* 53(3), 753-776. doi: 10.2307/4135449.

Fu, Y.B. 2015. Understanding crop genetic diversity under modern plant breeding. *Theor. Appl. Genet*., 128, 2131-2142. doi: 10.1007/s00122-015-2585-y.

Gage, J.L., Vaillancourt, B., Hamilton, J.P., Manrique-Carpintero, N.C., Gustafson, T.J., Barry, K., Lipzen, A., Tracy, W.F., Mikel, M.A., Kaeppler, S.M., *et al.* 2019. Multiple Maize Reference Genomes Impact the Identification of Variants by Genome-Wide Association Study in a Diverse Inbred Panel. *Plant Genome*, 12(2), 180069. doi: 10.3835/plantgenome2018.09.0069.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D.M., Burzynski-Chang, E.A., Fish, T.L., Stromberg, K.A., Sacks, G.L., *et al.* 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet*., 51(6), 1044–1051. doi: 10.1038/s41588-019-0410-2.

García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L.M., Götz, S., Tarazona, S., Dopazo, J., Meyer, T.F., Conesa, A. 2012. Qualimap: Evaluating next-generation sequencing alignment data. *Bioinformatics*, 28(20), 2678–2679. doi: 10.1093/bioinformatics/bts503.

García Fortea, E., Gramazio, P., Vilanova, S., Prohens, J. 2021. Development of pre-breeding materials and biotechnological tools for adaptation of eggplant to climate change [Tesis Doctoral]. Universidad Politécnica de Valencia.

Gardner, K.A., Wittern, L.M., Mackay, I.J. 2016. A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. *Plant Biotechnol. J*., 14(6), 1406-1417. doi: 10.1111/pbi.12504.

Gargis, A.S., Kalman, L., Lubin, I.M. 2016. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J. Clin. Microbiol*., 54(12), 2857–2865. doi: 10.1128/JCM.00949-16.

Garrison, E., Marth, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr.* arXiv1207.3907.

Golicz, A.A., Bayer, P.E., Edwards, D. 2015. Skim-based genotyping by sequencing, in: Batley, J. (eds) *Plant Genotyping. Methods in Molecular Biology*, Humana Press, New York, pp. 257-270. doi: 10.1007/978-1-4939-1966-6_19.

Goncalves-Vidigal, M.C., Gilio, T.A.S., Valentini, G., Vaz-Bisneta, M., Vidigal Filho, P.S., Song, Q., Oblessuc, P.R., Melotto, M. 2020. New Andean source of resistance to anthracnose and angular leaf spot: Fine-mapping of disease-resistance genes in California Dark Red Kidney common bean cultivar. *PLoS One*, 15(6), e0235215. doi: 10.1371/journal.pone.0235215.

Gorjanc, G., Dumasy, J.F., Gonen, S., Gaynor, R.C., Antolin, R., Hickey, J.M. 2017. Potential of low-coverage genotyping-by-sequencing and imputation for cost-effective genomic selection in biparental segregating populations. *Crop Sci*., 57(3), 1404-1420. doi: 10.2135/cropsci2016.08.0675.

Gouin, A., Legeai, F., Nouhaud, P., Whibley, A., Simon, J.C., Lemaitre, C. 2015. Whole-genome re-sequencing of non-model organisms: Lessons from unmapped reads. *Heredity*, 114(5), 494–501. doi: 10.1038/hdy.2014.85.

Gramazio, P., Prohens, J., Plazas, M., Mangino, G., Herraiz, F.J., García-Fortea, E., Vilanova, S. 2018. Genomic tools for the enhancement of vegetable crops: A case in eggplant. *Not. Bot. Horti Agrobot. Cluj-Napoca*, 46(1), 1-13. doi: 10.15835/nbha46110936.

Gramazio, P., Prohens, J., Plazas, M., Mangino, G., Herraiz, F.J., Vilanova, S. 2017. Development and genetic characterization of advanced backcross materials and an introgression line population of *Solanum incanum* in a *S. melongena* background. *Front. Plant Sci*., 8, 1477. doi: 10.3389/fpls.2017.01477.

Gramazio, P., Yan, H., Hasing, T., Vilanova, S., Prohens, J., Bombarely, A. 2019. Whole-Genome Resequencing of Seven Eggplant (*Solanum melongena*) and One Wild Relative (*S. incanum*) Accessions Provides New Insights and Breeding Tools for Eggplant Enhancement. *Front. Plant Sci.*, 1220. doi: 10.3389/fpls.2019.01220.

Guan, W., Ke, C., Tang, W., Jiang, J., Xia, J., Xie, X., Yang, M., Duan, C., Wu, W., Zheng, Y. 2022. Construction of a High-Density Recombination Bin-Based Genetic Map Facilitates High-Resolution Mapping of a Major QTL Underlying Anthocyanin Pigmentation in Eggplant. *Int. J. Mol. Sci*., 23(18), 10258. doi: 10.3390/ijms231810258.

Halperin, E., Stephan, D.A. 2009. SNP imputation in association studies. *Nat. Biotechnol*., 27(4), 349-351. doi: 10.1038/nbt0409-349.

Han, E., Sinsheimer, J.S., Novembre, J. 2014. Characterizing bias in population genetic inferences from low-coverage sequencing data. *Mol. Biol. Evol*., 31(3), 723–735. doi: 10.1093/molbev/mst229.

Happ, M.M., Wang, H., Graef, G.L., Hyten, D.L. 2019. Generating high density, low cost genotype data in Soybean [*Glycine max* (L.) Merr.]. G3 Genes, Genomes, Genet., 9(7), 2153-2160. doi: 10.1534/g3.119.400093.

Hardwick, S.A., Deveson, I.W., Mercer, T.R. 2017. Reference standards for next-generation sequencing. *Nat. Rev. Genet.*, 18(8), 473-484. doi: 10.1038/nrg.2017.44.

Hasan, M.S., Wu, X., Zhang, L. 2019. Uncovering missed indels by leveraging unmapped reads. *Sci. Rep*. 9(1), 11093. doi: 10.1038/s41598-019-47405-z.

Hegarty, M., Yadav, R., Lee, M., Armstead, I., Sanderson, R., Scollan, N., Powell, W., Skøt, L. 2013. Genotyping by RAD sequencing enables mapping of fatty acid composition traits in perennial ryegrass (*Lolium perenne* (L.)). *Plant Biotechnol. J.*, 11(5), 572-581. doi: 10.1111/pbi.12045.

Hirakawa, H., Shirasawa, K., Miyatake, K., Nunome, T., Negoro, S., Ohyama, A., Yamaguchi, H., Sato, S., Isobe, S., Tabata, S., *et al*. 2014. Draft genome sequence of eggplant (*Solanum melongena* L.): The representative *Solanum* species indigenous to the old world. *DNA Res*., 21(6), 649-660. doi: 10.1093/dnares/dsu027.

Holtgrewe, M., Emde, A.K., Weese, D., Reinert, K. 2011. A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinformatics*, 12(1), 210. doi: 10.1186/1471-2105-12-210.

Hooker, J.C., Nissan, N., Luckert, D., Charette, M., Zapata, G., Lefebvre, F., Mohr, R.M., Daba, K.A., Warkentin, T.D., Hadinezhad, M., *et al.* 2023. A Multi-Year, Multi-Cultivar Approach to Differential Expression Analysis of High- and Low-Protein Soybean (*Glycine max*). *Int. J. Mol. Sci*., 24(1), 222. doi: 10.3390/ijms24010222.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T., *et al*. 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res.*, 19(6), 1068-1076. doi: 10.1101/gr.089516.108.

Illumina. 2011. Quality Scores for Next-Generation Sequencing. URL http://Res.Illumina.Com/Documents/Products/Technotes/Technote_Q-Scores.Pdf (Acceso: 20/02/2023).

Islam, M.S., Thyssen, G.N., Jenkins, J.N., Zeng, L., Delhom, C.D., McCarty, J.C., Deng, D.D., Hinchliffe, D.J., Jones, D.C., Fang, D.D. 2016. A MAGIC population-based genome-wide association study reveals functional association of GhRBB1_A07 gene with superior fiber quality in cotton. *BMC Genomics*, 17(1), 903. doi: 10.1186/s12864-016-3249-2.

Jamil, S., Shahzad, R., Iqbal, M.Z., Yasmeen, E., Rahman, S.U. 2021. DNA Fingerprinting and Genetic Diversity Assessment of GM Cotton Genotypes for Protection of Plant Breeder Rights. *Int. J. Agric. Biol.*, 25(4), 768-776. doi: 10.17957/IJAB/15.1728.

Jiménez-Galindo, J.C., Malvar, R.A., Butrón, A., Santiago, R., Samayoa, L.F., Caicedo, M., Ordás, B. 2019. Mapping of resistance to corn borers in a MAGIC population of maize. *BMC Plant Biol.*, 19, 431. doi: 10.1186/s12870-019-2052-z.

Jing, Y., Zhao, X., Wang, J., Lian, M., Teng, W., Qiu, L., Han, Y., Li, W. 2019. Identification of loci and candidate genes for plant height in soybean (*Glycine max*) via genome-wide association study. *Plant Breed.*, 138(6), 721-732. doi: 10.1111/pbr.12735.

John, U. 2014. How does bowtie2 assign MAPQ scores? *Biofinysics*. URL http://biofinysics.blogspot.com/2014/05/how-does-bowtie2-assign-mapq-scores.html (Acceso: 03/08/2023).

Kaushik, P. 2020. Characterization of cultivated eggplant and its wild relatives based on important fruit biochemical traits. *Pakistan J. Biol. Sci*., 23(9), 1220-1226. doi: 10.3923/pjbs.2020.1220.1226.

Keel, B.N., Snelling, W.M. 2018. Comparison of Burrows-Wheeler transform-based mapping algorithms used in high-throughput whole-genome sequencing: Application to illumina data for livestock genomes. *Front. Genet*., 9, 35. doi: 10.3389/fgene.2018.00035.

Kim, J.E., Oh, S.K., Lee, J.H., Lee, B.M., Jo, S.H. 2014. Genome-wide SNP calling using next generation sequencing data in tomato. *Mol. Cells*, 37(1), 36. doi: 10.14348/molcells.2014.2241.

Kitony, J.K., Sunohara, H., Tasaki, M., Mori, J.I., Shimazu, A., Reyes, V.P., Yasui, H., Yamagata, Y., Yoshimura, A., Yamasaki, M. 2021. Development of an *Aus*-derived nested association mapping (*Aus*-NAM) population in rice. *Plants*, 10(6), 1255. doi: 10.3390/plants10061255.

Knapp, S., Vorontsova, M.S., Prohens, J. 2013. Wild Relatives of the Eggplant (*Solanum melongena* L.: Solanaceae): New Understanding of Species Names in a Complex Group. *PLoS One*, 8(2), e57039. doi: 10.1371/journal.pone.0057039.

Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K., Ding, L. 2009. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17), 2283-2285. doi: 10.1093/bioinformatics/btp373.

Kordrostami, M., Rahimi, M. 2015. Molecular markers in plants: Concepts and applications. *Genet. Third Millenn*., 13, 4024-4031.

Kouassi, A.B., Kouassi, K.B.A., Sylla, Z., Plazas, M., Fonseka, R.M., Kouassi, A., Fonseka, H., N'guetta, A.S.P., Prohens, J. 2021. Genetic parameters of drought tolerance for agromorphological traits in eggplant, wild relatives, and interspecific hybrids. *Crop Sci*., 61(1), 55-68. doi: 10.1002/csc2.20250.

Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C., Mott, R. 2009. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. *PLoS Genet*., 5(7), e1000551. doi: 10.1371/journal.pgen.1000551.

Ksouri, N., Benítez, M.M., Aballay, M.M., Sanchez, G., Contreras-Moreira, B., Gogorcena, Y. 2022. ddRAD-seq variant calling in peach and the effect of removing PCR duplicates. *Acta Hortic.,* 1352, 405–412. doi: 10.17660/ActaHortic.2022.1352.56.

Kumar, P., Choudhary, M., Jat, B.S., Kumar, B., Singh, V., Kumar, V., Singla, D., Rakshit, S. 2021. Skim sequencing: an advanced NGS technology for crop improvement. *J. Genet*., 100, 38. doi: 10.1007/s12041-021-01285-3.

Lander, E.S., Waterman, M.S. 1988. Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics*, 2(3), 231–239. doi: 10.1016/0888-7543(88)90007-9.

Langmead, B., Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4), 357-359. doi: 10.1038/nmeth.1923.

Langmead, B., Wilks, C., Antonescu, V., Charles, R. 2019. Scaling read aligners to hundreds of threads on general-purpose processors. *Bioinformatics*, 35(3), 421–432. doi: 10.1093/bioinformatics/bty648.

Lanteri, S., Barchi, L., 2019. Advances in Eggplant Genome Sequencing, in: Chapman, M. (eds) *The Eggplant Genome*, Springer, Cham, pp. 65–70. doi: 10.1007/978-3-319-99208-2_7.

Lebeau, A., Gouy, M., Daunay, M.C., Wicker, E., Chiroleu, F., Prior, P., Frary, A., Dintinger, J. 2013. Genetic mapping of a major dominant gene for resistance to *Ralstonia solanacearum* in eggplant. *Theor. Appl. Genet*., 126, 143-158. doi: 10.1007/s00122-012-1969-5.

Lee, H., Schatz, M.C. 2012. Genomic dark matter: The reliability of short read mapping illustrated by the genome mappability score. *Bioinformatics*, 28(16), 2097-2105. doi: 10.1093/bioinformatics/bts330.

Lee, H.C., Lai, K., Lorenc, M.T., Imelfort, M., Duran, C., Edwards, D. 2012. Bioinformatics tools and databases for analysis of next-generation sequence data. *Brief. Funct. Genomics*, 11(1), 12-24. doi: 10.1093/bfgp/elr037.

Leonard, S.A., Littlejohn, T.G., Baxevanis, A.D. 2006. Common File Formats. *Curr. Protoc. Bioinforma*., 16(1), A-1B. doi: 10.1002/0471250953.bia01bs16.

Lewis, R.S. 2011. *Nicotiana*, in: Kole, C. (eds) *Wild Crop Relatives: Genomic and Breeding Resources,* Springer, Berlin, Heidelberg, pp. 185–208. doi: 10.1007/978-3-642-21201-7_10.

LGC Biosearch Technologies. 2013. A primer on Probe – based SNP genotyping [WWW Document]. URL https://bitesizebio.com/ (Acceso: 27/01/23).

Li, D., Qian, J., Li, W., Yu, N., Gan, G., Jiang, Y., Li, W., Liang, X., Chen, R., Mo, Y., *et al.* 2021. A high-quality genome assembly of the eggplant provides insights into the molecular basis of disease resistance and chlorogenic acid synthesis. *Mol. Ecol. Resour.*, 21(4), 1274-1286. doi: 10.1111/1755-0998.13321.

Li, D., Qian, J., Li, W., Jiang, Y., Gan, G., Li, W., Chen, R., Yu, N., Li, Y., Wu, Y., *et al.* 2019. Genome sequence and analysis of the eggplant (*Solanum melongena* L.). *BioRxiv,* 824540. doi: 10.1101/824540.

Li, H., Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25(14), 1754-1760. doi: 10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25(16), 2078-2079. doi: 10.1093/bioinformatics/btp352.

Li, H., Ruan, J., Durbin, R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*., 18(11), 1851-1858. doi: 10.1101/gr.078212.108.

Li, R., Li, Y., Kristiansen, K., Wang, J. 2008b. SOAP: Short oligonucleotide alignment program. *Bioinformatics*, 24(5), 713-714. doi: 10.1093/bioinformatics/btn025.

Li, Y., Sidore, C., Kang, H.M., Boehnke, M., Abecasis, G.R. 2011. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res*., 21(6), 940–951. doi: 10.1101/gr.117259.110.

Li, Y.H., Zhou, G., Ma, J., Jiang, W., Jin, L.G., Zhang, Z., Guo, Y., Zhang, J., Sui, Y., Zheng, L., *et al.* 2014. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol*., 32(10), 1045–1052. doi: 10.1038/nbt.2979.

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., Liu, C., Nick, P., Du, F., Fan, P., *et al*. 2019. Whole-genome resequencing of 472 *Vitis* accessions for grapevine diversity and demographic history analyses. *Nat. Commun*., 10(1), 1190. doi: 10.1038/s41467-019-09135-8.

Liu, J., Shen, Q., Bao, H. 2022. Comparison of seven SNP calling pipelines for the next-generation sequencing data of chickens. *PLoS One*, 17(1), e0262574. doi: 10.1371/journal.pone.0262574.

Liu, R., Xiao, X., Gong, J., Li, J., Zhang, Z., Liu, A., Lu, Q., Shang, H., Shi, Y., Ge, Q., *et al*. 2020. QTL mapping for plant height and fruit branch number based on RIL population of upland cotton. *J. Cott. Res*., 3(1), 5. doi: 10.1186/s42397-020-0046-x.

Liu, S., An, Y., Li, F., Li, S., Liu, L., Zhou, Q., Zhao, S., Wei, C. 2018. Genome-wide identification of simple sequence repeats and development of polymorphic SSR markers for genetic studies in tea plant (*Camellia sinensis*). *Mol. Breed.,* 38, 59. doi: 10.1007/s11032-018-0824-z.

Liu, S., Yeh, C.T., Tang, H.M., Nettleton, D., Schnable, P.S. 2012. Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS One*, 7(5), e36406. doi: 10.1371/journal.pone.0036406.

Liu, W., Qian, Z., Zhang, J., Yang, J., Wu, M., Barchi, L., Zhao, H., Sun, H., Cui, Y., Wen, C. 2019. Impact of fruit shape selection on genetic structure and diversity uncovered from genome-wide perfect SNPs genotyping in eggplant. *Mol. Breed*., 39, 140. doi: 10.1007/s11032-019-1051-y.

Lou, R.N., Jacobs, A., Wilder, A.P., Therkildsen, N.O. 2021. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Mol. Ecol.*, 30, 5966–5993. doi: 10.1111/mec.16077.

Lu, F., Romay, M.C., Glaubitz, J.C., Bradbury, P.J., Elshire, R.J., Wang, T., Li, Y., Li, Y., Semagn, K., Zhang, X., *et al.* 2015. High-resolution genetic mapping of maize pan-genome sequence anchors. *Nat. Commun*., 6(1), 6914. doi: 10.1038/ncomms7914.

Lunter, G., Goodson, M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res*., 21(6), 936-939. doi: 10.1101/gr.111120.110.

Luscombe, N.M., Greenbaum, D., Gerstein, M. 2001. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med*., 40(4), 346-358. doi: 10.1055/s-0038-1634431.

Lyamichev, V., Mast, A.L., Hall, J.G., Prudent, J.R., Kaiser, M.W., Takova, T., Kwiatkowski, R.W., Sander, T.J., De Arruda, M., Arco, D.A., *et al.* 1999. Polymorphism identification and quantitative detection of genomic DNA by invasive cleavage of oligonucleotide probes. *Nat. Biotechnol*., 17(3), 292-296. doi: 10.1038/7044.

Mackay, I., Powell, W. 2007. Methods for linkage disequilibrium mapping in crops. *Trends Plant Sci.,* 12(2), 57-63. doi: 10.1016/j.tplants.2006.12.001.

Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F., Brandi, M.L. 2010. Bioinformatics for next generation sequencing data. *Genes*, 1(2), 294-307. doi: 10.3390/genes1020294.

Malmberg, M.M., Barbulescu, D.M., Drayton, M.C., Shinozuka, M., Thakur, P., Ogaji, Y.O., Spangenberg, G.C., Daetwyler, H.D., Cogan, N.O.I. 2018. Evaluation and recommendations for

routine genotyping using skim whole genome re-sequencing in canola. *Front. Plant Sci*., 9, 1809. doi: 10.3389/fpls.2018.01809.

Mangino, G., Arrones, A., Plazas, M., Pook, T., Prohens, J., Gramazio, P., Vilanova, S. 2022. Newly Developed MAGIC Population Allows Identification of Strong Associations and Candidate Genes for Anthocyanin Pigmentation in Eggplant. *Front. Plant Sci*., 13, 847789. doi: 10.3389/fpls.2022.847789.

Mangino, G., Plazas, M., Vilanova, S., Prohens, J., Gramazio, P. 2020. Performance of a set of eggplant (*Solanum melongena*) lines with introgressions from its wild relative *S. incanum* under open field and screenhouse conditions and detection of QTLs. *Agronomy*, 10(4), 467. doi: 10.3390/agronomy10040467

Marchini, J., Howie, B. 2010. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet*., 11(7), 499-511. doi: 10.1038/nrg2796.

Maroto, J. V. 2002. Horticultura herbácea especial, 2ª edición, Mundi-Prensa, Madrid, España.

Martínez-Ispizua, E., Calatayud, Á., Marsal, J.I., Mateos-Fernández, R., Díez, M.J., Soler, S., Valcárcel, J.V., Martínez-Cuenca, M.R. 2021. Phenotyping Local Eggplant Varieties: Commitment to Biodiversity and Nutritional Quality Preservation. *Front. Plant Sci*., 12, 696272. doi: 10.3389/fpls.2021.696272.

Mathew, B., Léon, J., Sannemann, W., Sillanpää, M.J. 2018. Detection of epistasis for flowering time using bayesian multilocus estimation in a barley MAGIC population. *Genetics*, 208(2), 525-536. doi: 10.1534/genetics.117.300546.

Maxam, A.M., Gilbert, W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci*., 74(2), 560-564. doi: 10.1073/pnas.74.2.560.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al*. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 20(9), 1297-1303. doi: 10.1101/gr.107524.110.

Metzker, M.L. 2010. Sequencing technologies - the next generation. *Nat. Rev. Genet*., 11(1), 31-46. doi: 10.1038/nrg2626.

Meyer, R.S., Karol, K.G., Little, D.P., Nee, M.H., Litt, A. 2012. Phylogeographic relationships among Asian eggplants and new perspectives on eggplant domestication. *Mol. Phylogenet. Evol.*, 63(3), 685-701. doi: 10.1016/j.ympev.2012.02.006.

Ministerio de Agricultura, Pesca y Alimentación (MAPA). 2021. Anuario de estadística 2021. Superficies y producciones de cultivos [WWW Document]. Madrid, España. URL https://www.mapa.gob.es/es/estadistica/temas/publicaciones/anuario-de-estadistica/2021/ (Acceso: 22/12/22).

Mishra, A., Singh, P. K., Bhandawat, A., Sharma, V., Sharma, V., Singh, P., Roy, J., Sharma, H. 2022. Analysis of SSR and SNP markers, in: Singh, D. B., Pathak, R.K. (eds) *Bioinformatics*, Academic Press, pp. 131–144. doi: 10.1016/B978-0-323-89775-4.00017-1.

Moodley, V., Naidoo, R., Gubba, A., Mafongoya, P.L. 2019. Development of Potato virus Y (PVY) resistant pepper (*Capsicum annuum* L.) lines using marker-assisted selection (MAS). *Physiol. Mol. Plant Pathol.,* 105, 96-101. doi: 10.1016/j.pmpp.2018.12.002.

Mott, R., Talbot, C.J., Turri, M.G., Collins, A.C., Flint, J. 2000. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci*., 97(23), 12649-12654. doi: 10.1073/pnas.230304397.

Mun, J.H., Chung, H., Chung, W.H., Oh, M., Jeong, Y.M., Kim, N., Ahn, B.O., Park, B.S., Park, S., Lim, K.B., *et al.* 2015. Construction of a reference genetic map of Raphanus sativus based on genotyping by whole-genome resequencing. *Theor. Appl. Genet.,* 128, 259–272. doi: 10.1007/s00122-014-2426-4.

Musich, R. 2020. A Recent (2020) Comparative Analysis of Genome Aligners Shows HISAT2 and BWA are Among the Best Tools [Tesis doctoral]. Rochester Institute of Technology.

Musich, R., Cadle-Davidson, L., Osier, M. V. 2021. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Front. Plant Sci.*, 12, 657240. doi: 10.3389/fpls.2021.657240.

Namisy, A., Chen, J.R., Prohens, J., Metwally, E., Elmahrouk, M., Rakha, M. 2019. Screening cultivated eggplant and wild relatives for resistance to bacterial wilt (*Ralstonia solanacearum*). *Agric*., 9(7), 157. doi: 10.3390/agriculture9070157.

Nayak, S. N., Singh, V. K., Varshney, R.K. 2017. Marker-assisted selection, in: Thomas, B., Murray, B.G., Murphy, D.J. (eds) *Encyclopedia of Applied Plant Sciences*, Academic Press, Waltham, MA, pp. 183–197. doi: 10.1016/B978-0-12-394807-6.00192-1.

Nederbragt, A.J., Rounge, T.B., Kausrud, K.L., Jakobsen, K.S. 2010. Identification and Quantification of Genomic Repeats and Sample Contamination in Assemblies of 454 Pyrosequencing Reads. *Sequencing,* 2010, 782465. doi: 10.1155/2010/782465.

Ning, Z., Cox, A.J., Mullikin, J.C. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res*., 11(10), 1725-1729. doi: 10.1101/gr.194201.

Ongom, P.O., Ejeta, G. 2018. Mating design and genetic structure of a multi-parent advanced generation intercross (MAGIC) population of sorghum (*Sorghum bicolor* (L.) moench). *G3 Genes, Genomes, Genet*., 8(1), 331-341. doi: 10.1534/g3.117.300248.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M.R., Zschocke, J., Trajanoski, Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.*, 15(2), 256-278. doi: 10.1093/bib/bbs086.

Page, A., Gibson, J., Meyer, R.S., Chapman, M.A. 2019. Eggplant domestication: Pervasive gene flow, feralization, and transcriptomic divergence. *Mol. Biol. Evol.*, 36(7), 1359-1375. doi: 10.1093/molbev/msz062.

Pareek, C.S., Smoczynski, R., Tretyn, A. 2011. Sequencing technologies and genome sequencing. *J. Appl. Genet*., 52, 413-435. doi: 10.1007/s13353-011-0057-x.

Pasaniuc, B., Rohland, N., McLaren, P.J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B.M., Daly, M.J., Sklar, P., *et al*. 2012. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.,* 44(6), 631-635. doi: 10.1038/ng.2283.

Pascual, L., Desplat, N., Huang, B.E., Desgroux, A., Bruguier, L., Bouchet, J.P., Le, Q.H., Chauchard, B., Verschave, P., Causse, M. 2015. Potential of a tomato MAGIC population to decipher the

genetic control of quantitative traits and detect causal variants in the resequencing era. P*lant Biotechnol. J.*, 13(4), 565-577. doi: 10.1111/pbi.12282.

Patterson, J., Carpenter, E.J., Zhu, Z., An, D., Liang, X., Geng, C., Drmanac, R., Wong, G.K.S. 2019. Impact of sequencing depth and technology on de novo RNA-Seq assembly. *BMC Genomics*, 20, 604. doi: 10.1186/s12864-019-5965-x.

Peterson, G.W., Dong, Y., Horbach, C., Fu, Y.B. 2014. Genotyping-by-sequencing for plant genetic diversity analysis: A lab guide for SNP genotyping. *Diversity*, 6(4), 665-680. doi: 10.3390/d6040665.

Plazas, M., Vilanova, S., Gramazio, P., Rodríguez-Burruezo, A., Fita, A., Herraiz, F.J., Ranil, R., Fonseka, R., Niran, L., Fonseka, H., *et al*. 2016. Interspecific hybridization between eggplant and wild relatives from different genepools. *J. Am. Soc. Hortic. Sci.,* 141(1), 34-44. doi: 10.21273/jashs.141.1.34.

Portis, E., Cericola, F., Barchi, L., Toppino, L., Acciarri, N., Pulcini, L., Sala, T., Lanteri, S., Rotino, G.L. 2015. Association mapping for fruit, plant and leaf morphology traits in eggplant. *PLoS One*, 10(8), e0135200. doi: 10.1371/journal.pone.0135200.

Prohens, J., Blanca, J.M., Nuez, F. 2005. Morphological and molecular variation in a collection of eggplants from a secondary center of diversity: Implications for conservation and breeding. *J. Am. Soc. Hortic. Sci*., 130(1), 54-63. doi: 10.21273/jashs.130.1.54.

Purugganan, M.D., Jackson, S.A. 2021. Advancing crop genomics from lab to field. *Nat. Genet.,* 53(5), 595-601. doi: 10.1038/s41588-021-00866-3.

Qian, Z., Zhang, B., Chen, H., Lu, L., Duan, M., Zhou, J., Cui, Y., Li, D. 2021. Identification of Quantitative Trait Loci Controlling the Development of Prickles in Eggplant by Genome Re-sequencing Analysis. *Front. Plant Sci*., 12, 731079. doi: 10.3389/fpls.2021.731079.

Qin, P., Lu, H., Du, H., Wang, H., Chen, W., Chen, Z., He, Q., Ou, S., Zhang, H., Li, X., *et al*. 2021. Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell,* 184(13), 3542-3558.e16. doi: 10.1016/j.cell.2021.04.046.

R Core Team. 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Ramírez, F., Dündar, F., Diehl, S., Grüning, B.A., Manke, T. 2014. DeepTools: A flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*., 42(W1), W187-W191. doi: 10.1093/nar/gku365.

Rana, M.M., Takamatsu, T., Baslam, M., Kaneko, K., Itoh, K., Harada, N., Sugiyama, T., Ohnishi, T., Kinoshita, T., Takagi, H., *et al*. 2019. Salt tolerance improvement in rice through efficient SNP marker-assisted selection coupled with speed-breeding. *Int. J. Mol. Sci*., 20(10), 2585. doi: 10.3390/ijms20102585.

Ranil, R.H.G., Prohens, J., Aubriot, X., Niran, H.M.L., Plazas, M., Fonseka, R.M., Vilanova, S., Fonseka, H.H., Gramazio, P., Knapp, S. 2017. *Solanum insanum* L. (subgenus *Leptostemonum* Bitter, Solanaceae), the neglected wild progenitor of eggplant (S. *melongena* L.): a review of taxonomy, characteristics and uses aimed at its enhancement for improved eggplant breeding. *Genet. Resour. Crop Evol.,* 64, 1707-1722. doi: 10.1007/s10722-016-0467-z.

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. *Mol. Plant*., 10(8), 1047-1064. doi: 10.1016/j.molp.2017.06.008.

Ratnaparkhe, M.B., Marmat, N., Kumawat, G., Shivakumar, M., Kamble, V.G., Nataraj, V., Ramesh, S.V., Deshmukh, M.P., Singh, A.K., Sonah, H., *et al.* 2020. Whole Genome Re-sequencing of Soybean Accession EC241780 Providing Genomic Landscape of Candidate Genes Involved in Rust Resistance. *Curr. Genomics*, 21(7), 504-511. doi: 10.2174/1389202921999200601142258.

Ravinet, M., Meier, J. 2021. Speciation & Population Genomics: a how-to-guide [WWW Document]. Handl. NGS DATA. URL https://speciationgenomics.github.io/ (Acceso: 20/05/23).

Ren, G., Zhang, X., Li, Y., Ridout, K., Serrano-Serrano, M.L., Yang, Y., Liu, A., Ravikanth, G., Nawaz, M.A., Mumtaz, A.S., *et al.* 2021. Large-scale whole-genome resequencing unravels the domestication history of *Cannabis sativa*. *Sci. Adv*., 7(29), eabg2286. doi: 10.1126/sciadv.abg2286.

Rochette, N.C., Rivera-Colón, A.G., Walsh, J., Sanger, T.J., Campbell-Staton, S.C., Catchen, J.M. 2023. On the causes, consequences, and avoidance of PCR duplicates: towards a theory of library complexity. *Mol. Ecol. Resour*., 00, 1– 20. doi: 10.1111/1755-0998.13800.

Roy, S., Coldren, C., Karunamurthy, A., Kip, N.S., Klee, E.W., Lincoln, S.E., Leon, A., Pullambhatla, M., Temple-Smolkin, R.L., Voelkerding, K. V., *et al*. 2018. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagnostics*, 20(1), 4-27. doi: 10.1016/j.jmoldx.2017.11.003.

Rubinacci, S., Ribeiro, D.M., Hofmeister, R.J., Delaneau, O. 2021. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet*., 53(1), 120–126. doi: 10.1038/s41588-020-00756-0.

Salgon, S., Raynal, M., Lebon, S., Baptiste, J.M., Daunay, M.C., Dintinger, J., Jourda, C. 2018. Genotyping by sequencing highlights a polygenic resistance to *Ralstonia pseudosolanacearum* in eggplant (*Solanum melongena* L.). *Int. J. Mol. Sci*., 19(2), 357. doi: 10.3390/ijms19020357.

Samantara, K., Reyes, V.P., Agrawal, N., Mohapatra, S.R., Jena, K.K. 2021. Advances and trends on the utilization of multi-parent advanced generation intercross (MAGIC) for crop improvement. *Euphytica*, 217(10), 189. doi: 10.1007/s10681-021-02925-6.

Sanger, F., Nicklen, S., Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.*, 74(12), 5463-5467. doi: 10.1073/pnas.74.12.5463.

Scheben, A., Batley, J., Edwards, D. 2017. Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. *Plant Biotechnol. J.*, 15(2), 149-161. doi: 10.1111/pbi.12645.

Schippers, R.R. 2000. African Indigenous Vegetables: An Overview of the Cultivated Species. *Nat. Resour. Int*., London, Uk.

Schweyen, H., Rozenberg, A., Leese, F. 2014. Detection and removal of PCR duplicates in population genomic ddRAD studies by addition of a degenerate base region (DBR) in sequencing adapters. *Biol. Bull*. 227(2), 146–160. doi: 10.1086/BBLv227n2p146.

Sealfon, R., Gire, S., Ellis, C., Calderwood, S., Qadri, F., Hensley, L., Kellis, M., Ryan, E.T., LaRocque, R.C., Harris, J.B., *et al*. 2012. High depth, whole-genome sequencing of cholera isolates from Haiti and the Dominican Republic. *BMC Genomics,* 13(1), 468. doi: 10.1186/1471-2164-13-468.

Sim, S.C., Nguyen, N.N., Kim, N., Kim, J., Park, Y. 2018. Whole-genome resequencing reveals genome-wide single nucleotide polymorphisms between orange-fleshed and green-fleshed melons. *Hortic. Environ. Biotechnol*., 59, 275–283. doi: 10.1007/s13580-018-0030-2.

Singh, B. D., Singh, A.K. 2015. Mapping Populations, in: Singh, B. D., Singh, A.K. (eds) Marker-Assisted Plant Breeding: Principles and Practices. Springer, New Delhi, pp. 125–150. doi: 10.1007/978-81-322-2316-0_5.

Sokolov, B.P. 1990. Primer extension technique for the detection of single nucleotide in genomic DNA. *Nucleic Acids Res*., 18(12), 3671. doi: 10.1093/nar/18.12.3671.

Song, K., Li, L., Zhang, G. 2016. Coverage recommendation for genotyping analysis of highly heterologous species using next-generation sequencing technology. *Sci. Rep*., 6(1), 35736. doi: 10.1038/srep35736.

Spooner, D.M. 1990. The potato: Evolution, biodiversity and genetic resources. J.G. Hawkes. *Am. J. Pot. Res.*, 67, 733-735. doi: 10.1007/bf03044023.

Stehmann, J. R., Lorenz-Lemke, A. P., Freitas, L. B., Semir, J. 2000. The genus Petunia, in: Gerats, T., Strommer, J. (eds) Petunia: Evolutionary, Developmental, and Physiological Genetics, Springer, New York, pp. 1–28. doi: 10.1007/978-0-387-84796-2_1.

Sukhotu, T., Hosaka, K. 2006. Origin and evolution of Andigena potatoes revealed by chloroplast and nuclear DNA markers. *Genome*, 49(6), 636-647. doi: 10.1139/G06-014.

Sun, Y., Shang, L., Zhu, Q.H., Fan, L., Guo, L. 2022. Twenty years of plant genome sequencing: achievements and challenges. *Trends Plant Sci.*, 27(4), 391-401. doi: 10.1016/j.tplants.2021.10.006.

Swarup, S., Cargill, E.J., Crosby, K., Flagel, L., Kniskern, J., Glenn, K.C. 2021. Genetic diversity is indispensable for plant breeding to improve crops. *Crop Sci*., 61(2), 839-852. doi: 10.1002/csc2.20377.

Taher, D., Solberg, S., Prohens, J., Chou, Y.Y., Rakha, M., Wu, T.H. 2017. World vegetable center eggplant collection: Origin, composition, seed dissemination and utilization in breeding. *Front. Plant Sci.*, 8, 1484. doi: 10.3389/fpls.2017.01484.

Tanksley, S.D. 2004. The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. *Plant Cell*, 16(suppl_1), S181-S189. doi: 10.1105/tpc.018119.

The 3,000 rice genomes project. 2014. The 3,000 rice genomes project. *Gigascience*, 3(1), 7. doi: 10.1186/2047-217X-3-7.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408, 796-815. doi: 10.1038/35048692.

The Potato Genome Sequencing Consortium. 2011. Genome sequence and analysis of the tuber crop potato. *Nature,* 475(7355), 189-195. doi: 10.1038/nature10158.

The Tomato Genome Consortium. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485, 635-641. doi: 10.1038/nature11119.

Therkildsen, N.O., Palumbi, S.R. 2017. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Mol. Ecol. Resour*., 17, 194–208. doi: 10.1111/1755-0998.12593.

Thomson, M.J. 2014. High-Throughput SNP Genotyping to Accelerate Crop Improvement. *Plant Breed. Biotechnol*., 2(3), 195-212. doi: 10.9787/pbb.2014.2.3.195.

Toppino, L., Barchi, L., Mercati, F., Acciarri, N., Perrone, D., Martina, M., Gattolin, S., Sala, T., Fadda, S., Mauceri, A., *et al*. 2020. A new intra-specific and high-resolution genetic map of eggplant based on a ril population, and location of QTLS related to plant anthocyanin pigmentation and seed vigour. *Genes, 11(7), 745*. doi: 10.3390/genes11070745.

Toppino, L., Prohens, J., Rotino, G.L., Plazas, M., Parisi, M., Carrizo García, C., Tripodi, P. 2021. Pepper and Eggplant Genetic Resources, in: Carputo, D., Aversano, R., Ercolano, M.R. (eds) *The Wild Solanums Genomes. Compendium of Plant Genomes*, Springer, Cham, pp. 119-154. doi: 10.1007/978-3-030-30343-3_6.

Torkamaneh, D., Boyle, B., Belzile, F. 2018b. Efficient genome-wide genotyping strategies and data integration in crop plants. *Theor. Appl. Genet*., 131, 499-511. doi: 10.1007/s00122-018-3056-z.

Torkamaneh, D., Laroche, J., Tardivel, A., O'Donoughue, L., Cober, E., Rajcan, I., Belzile, F. 2018a. Comprehensive description of genomewide nucleotide and structural variation in short-season soya bean. *Plant Biotechnol. J*., 16(3), 749–759. doi: 10.1111/pbi.12825.

Treangen, T.J., Salzberg, S.L. 2012. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet*., 13(1), 36–46. doi: 10.1038/nrg3117.

Unamba, C.I.N., Nag, A., Sharma, R.K. 2015. Next generation sequencing technologies: The doorway to the unexplored genomics of non-model plants. *Front. Plant Sci*., 6, 1074. doi: 10.3389/fpls.2015.01074.

Valdar, W., Flint, J., Mott, R. 2006. Simulating the collaborative cross: Power of quantitative trait loci detection and mapping resolution in large sets of recombinant inbred strains of mice. *Genetics*, 172(3), 1783-1797. doi: 10.1534/genetics.104.039313.

Valiente-Mullor, C., Beamud, B., Ansari, I., Frances-Cuesta, C., Garcia-Gonzalez, N., Mejia, L., Ruiz-Hueso, P., Gonzalez-Candelas, F. 2021. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. *PLoS Comput. Biol*., 17(1), e1008678. doi: 10.1371/journal.pcbi.1008678.

van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., *et al*. 2013. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma*., 43(1110), 11.10.1-11.10.33. doi: 10.1002/0471250953.bi1110s43.

van Tassell, C.P., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat. Methods*, 5(3), 247-252. doi: 10.1038/nmeth.1185.

Vavilov, N.I. 1951. The Origin, Variation, Immunity and Breeding of Cultivated Plants. Transl. by K. Start. Waltham, Mass.: Chronica Botanica, Nueva York. doi: 10.1097/00010694-195112000-00018.

Vilanova, S., Alonso, D., Gramazio, P., Plazas, M., García-Fortea, E., Ferrante, P., Schmidt, M., Díez, M.J., Usadel, B., Giuliano, G., *et al*. 2020. SILEX: A fast and inexpensive high-quality DNA extraction method suitable for multiple sequencing platforms and recalcitrant plant species. *Plant Methods,* 16, 110. doi: 10.1186/s13007-020-00652-y.

Vorontsova, M.S., Knapp, S. 2016. A Revision of the "Spiny Solanums," *Solanum* subgenus *Leptostemonum* (Solanaceae), in Africa and Madagascar. *Syst. Bot. Monogr*., 99, 1-436.

Vorontsova, M.S., Stern, S., Bohs, L., Knapp, S. 2013. African spiny *Solanum* (subgenus *Leptostemonum*, Solanaceae): A thorny phylogenetic tangle. *Bot. J. Linn. Soc*., 173(2), 176-193. doi: 10.1111/boj.12053.

Wada, T., Oku, K., Nagano, S., Isobe, S., Suzuki, H., Mori, M., Takata, K., Hirata, C., Shimomura, K., Tsubone, M., *et al.* 2017. Development and characterization of a strawberry MAGIC population derived from crosses with six strawberry cultivars. *Breed. Sci.,* 67(4), 370-381. doi: 10.1270/jsbbs.17009.

Wang, C., Qiao, A., Fang, X., Sun, L., Gao, P., Davis, A.R., Liu, S., Luan, F. 2019. Fine Mapping of Lycopene Content and Flesh Color Related Gene and Development of Molecular Marker–Assisted Selection for Flesh Color in Watermelon (*Citrullus lanatus*). *Front. Plant Sci.*, 10, 1240. doi: 10.3389/fpls.2019.01240.

Wang, G.L., Mackill, D.J., Bonman, J.M., McCouch, S.R., Champoux, M.C., Nelson, R.J. 1994. RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. *Genetics*, 136(4), 1424-1434. doi: 10.1093/genetics/136.4.1421.

Weese, T.L., Bohs, L. 2010. Eggplant origins: Out of Africa, into the Orient. *Taxon*, 59(1), 49-56. doi: 10.1002/tax.591006.

Wei, Q., Wang, J., Wang, W., Hu, T., Hu, H., Bao, C. 2020. A high-quality chromosome-level genome assembly reveals genetics for important traits in eggplant. *Hortic. Res*., 7(1), 153. doi: 10.1038/s41438-020-00391-0.

Wetterstrand, K.A. n.d. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). [WWW Document]. URL www.genome.gov/sequencingcostsdata (Acceso: 14/01/23).

Wright, M.N., Gola, D., Ziegler, A. 2017. Preprocessing and quality control for whole-genome sequences from the Illumina HiSeq X platform, in: Elston, R. (eds) *Statistical Human Genetics. Methods in Molecular Biology*, Humana Press, New York, pp. 629-647. doi: 10.1007/978-1-4939-7274-6_30.

Wu, J., Jiang, Y., Liang, Y., Chen, L., Chen, W., Cheng, B. 2019a. Expression of the maize MYB transcription factor ZmMYB3R enhances drought and salt stress tolerance in transgenic plants. *Plant Physiol. Biochem*., 137, 179-188. doi: 10.1016/j.plaphy.2019.02.010.

Wu, X., Heffelfinger, C., Zhao, H., Dellaporta, S.L. 2019b. Benchmarking variant identification tools for plant diversity discovery. *BMC Genomics,* 20, 701. doi: 10.1186/s12864-019-6057-7.

Yan, W., Zhao, H., Yu, K., Wang, T., Khattak, A.N., Tian, E. 2020. Development of a multiparent advanced generation intercross (MAGIC) population for genetic exploitation of complex traits

in *Brassica juncea*: Glucosinolate content as an example. *Plant Breed*., 139(4), 779-789. doi: 10.1111/pbr.12820.

Yan, Y., Chaturvedi, N., Appuswamy, R. 2021. Accel-Align: a fast sequence mapper and aligner based on the seed–embed–extend method. *BMC Bioinformatics*, 22(1), 257. doi: 10.1186/s12859-021-04162-z.

Yao, L., Witt, K., Li, H., Rice, J., Salinas, N.R., Martin, R.D., Huerta-Sánchez, E., Malhi, R.S. 2020b. Population genetics of wild *Macaca fascicularis* with low-coverage shotgun sequencing of museum specimens. *Am. J. Phys. Anthropol*., 173(1), 21–33. doi: 10.1002/ajpa.24099.

Yao, Z., You, F.M., N'Diaye, A., Knox, R.E., McCartney, C., Hiebert, C.W., Pozniak, C., Xu, W. 2020a. Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics*, 21(1), 360. doi: 10.1186/s12859-020-03704-1.

Yeo, Z.X., Wong, J.C.L., Rozen, S.G., Lee, A.S.G. 2014. Evaluation and optimisation of indel detection workflows for ion torrent sequencing of the BRCA1 and BRCA2 genes. *BMC Genomics*, 15(1), 516. doi: 10.1186/1471-2164-15-516.

Yu, J., Hu, S., Wang, J., Wong, G.K.S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., *et al*. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 296(5565), 79-92. doi: 10.1126/science.1068037.

Yu, X., Sun, S. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics*, 14, 274. doi: 10.1186/1471-2105-14-274.

Zan, Y., Payen, T., Lillie, M., Honaker, C.F., Siegel, P.B., Carlborg, Ö. 2019. Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: A cost-efficient approach. *Genet. Sel. Evol*., 51, 44. doi: 10.1186/s12711-019-0487-1.