UNIVERSITAT POLITÈCNICA DE VALÈNCIA

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Instituto Universitario de Conservación y Mejora de la Agrodiversidad Valenciana

Comparación de estrategias de imputación para mejorar la resolución del genotipado por secuenciación a muy baja cobertura

Trabajo Fin de Máster

Máster Universitario Erasmus Mundus en Mejora Genética Vegetal / Erasmus Mundus Master in Plant Breeding - emPLANT +

AUTOR/A: Tinajero Malta, Ana Carolina

Tutor/a: Gramazio, Pietro

Director/a Experimental: Baraja Fonseca, Virginia

CURSO ACADÉMICO: 2023/2024

UNIVERSIDAD POLITÉCNICA DE VALENCIA

Instituto de Conservación y Mejora de la Agrodiversidad Valenciana (COMAV)

Comparación de estrategias de imputación para mejorar la resolución del genotipado por secuenciación a muy baja cobertura

Ana Carolina Tinajero Malta

Tutor: Pietro Gramazio

Directora experimental: Virginia Baraja Fonseca

Trabajo de Fin de Máster

Máster en Mejora Genética Vegetal

Curso 2023-2024

Valencia, Julio de 2024

RESUMEN

La berenjena (*Solanum melongena*) es una de las especies hortícolas más consumidas a nivel mundial debido, entre otros aspectos, a sus propiedades beneficiosas para la salud, ya que es rica en antioxidantes. La demanda por este cultivo aumenta exponencialmente al igual que el interés por desarrollar nuevas variedades. Sin embargo, en comparación con otros cultivos de importancia económica, el desarrollo de herramientas genéticas y genómicas para la mejora genética y la disección de caracteres morfoagronómicos ha sido menor. Para el estudio de la base genética se emplean poblaciones experimentales, como son las poblaciones MAGIC. Las poblaciones multiparentales MAGIC son una herramienta poderosa para identificar QTLs de caracteres de interés de una manera más certera. Sin embargo, la secuenciación a alta profundidad de estas poblaciones no es viable económicamente debido a los altos costes que conllevaría. Por esta razón, se propone el uso de bajas coberturas de secuenciación junto con la imputación genética para inferir y predecir el genotipo completo de un individuo determinado.

Para la imputación genética se han desarrollado varios programas computacionales basados en el modelo estadístico Hidden Markov Model (HMM). En este estudio se han utilizado los programas Beagle, Impute y Minimac para imputar genotipos en una población S5 MAGIC de berenjena genotipada por secuenciación del genoma completo a baja cobertura (3X), empleando la información de los parentales (20X) como panel de referencia. También se generaron dos mapas genéticos, imprescindible para imputar con Beagle e Impute, a partir de la información tanto de la generación S3 como de la S5 de la MAGIC. Finalmente, se utilizó el mapa generado a partir de la información de genotipado por la plataforma SPET (Single primer enrichment technology) de la población S3 debido a su mayor resolución. Para determinar la precisión de los programas de imputación y su comparación, se hizo uso del estadístico R^2. Minimac fue el programa con el que se obtuvo una mayor precisión en la imputación de los genotipos faltantes, mientras que Impute fue el más complicado en términos de ejecución e interpretación. Los resultados obtenidos no solo beneficiarán directamente a los programas de mejora de la berenjena, sino que también servirán como una alternativa para el avance del estudio genético de otros cultivos de gran importancia económica.

ABSTRACT

Eggplant (*Solanum melongena*) is one of the most consumed horticultural species worldwide due to its beneficial health properties as it is rich in antioxidants. The demand for this crop is increasing exponentially, as is the interest in developing new varieties. However, compared to other crops of economic importance, the development of genetic and genomic tools for genetic improvement and dissection of morphoagronomic characters has been delayed. To study the genetic basis, experimental populations such as MAGIC populations are used. Multiparental populations, also known as MAGIC, are a powerful tool to identify QTLs of traits of interest in a more accurate way. Hiwever, high-depth sequencing of these populations is not economically viable due to the high costs it will entail. For this reason, the use of computational tools, such as genetic imputation, is proposed to infer and predict the complete genotype of a given individual.

For genetic imputation, several computer programs have been developed based on the Hidden Markov Model statistical model (HMM). In this study, Beagle, Impute and Minimac programs have been used to impute genotypes from a S5 MAGIC eggplant population genotyped by low-coverage whole genome sequencing (3X), using founder parents' information (20X) as reference panel. Two genetic maps were also generated, essential for imputing with Beagle and Impute, based on the information from both S3 and S5 generations of the MAGIC population. Finally, the map generated from the genotyping information by the SPET (Single primer enrichment technology) platform of the S3 population was used due to its higher resolution. To determine the precision of the imputation programs and for comparison, the R^2 statistic was used. Minimac was the software with the highest imputation accuracy for the missing genotypes, while Impute was the most complicated in terms of execution and interpretation. The results obtained will not only directly benefit eggplant improvement programs but will also serve as an alternative for the advancement of the genetic study of other crops of great economic importance.

DEDICATORIA

Para você, mamãe.

AGRADECIMIENTOS

A mi familia, por siempre creer en mí y apoyarme, aunque estemos a kilómetros de distancia.

A mis amigos de emPLANT, Hayu, Sol, Bry y Naty, por todo el apoyo tanto en los buenos como en los malos momentos.

A Pietro, por darme la oportunidad de realizar este trabajo con ustedes y abrirme las puertas al mundo de la bioinformática.

A Virginia, por toda la paciencia y apoyo en estos meses de trabajo. Por ayudarme a descifrar todos los inconvenientes que se iban presentando y por guiar este trabajo por el mejor camino posible.

Y principalmente a Santi y Polo, por embarcarse en esta aventura conmigo sin pensarlo dos veces y ser mi apoyo incondicional siempre.

TABLA DE CONTENIDO

1. INTRODUCCIÓN 1

1.1. La berenjena (*Solanum melongena* L.) 1

1.1.1. Descripción de la especie 1

1.1.2. Clasificación taxonómica 2

1.1.3. Especies silvestres 2

1.1.4. Origen y domesticación 3

1.1.5. Importancia económica 4

1.2. Mejora genética 5

1.2.1. Poblaciones experimentales 6

1.3. Imputación genética 7

1.3.1. Métodos de imputación genética 10

1.3.2. Información necesaria para la imputación genética 11

2. OBJETIVOS 13

Objetivo general 13

Objetivos específicos 13

3. MATERIALES Y MÉTODOS 14

3.1. Selección del material a imputar 14

3.2. Instalación de los programas 15

3.3. Elaboración del panel de referencia 16

3.4. Elaboración del mapa genético 16

3.5. Phasing 18

3.6. Imputación de datos 19

3.6.1. Beagle 5.1 19

3.6.2. Impute 2 20

3.6.3. Minimac 4 23

3.7. Evaluación del desempeño de la imputación 24

4. RESULTADOS 26

4.1. Calidad del panel de referencia 26

4.2. Resumen del mapa genético 27

4.3. Imputación de datos 33

4.3.1. Beagle 5.1 33

4.3.2. Impute 2 34
4.3.3. Minimac 4 35
4.4. Evaluación del desempeño de la imputación 35
5. DISCUSIÓN 39
5.1. Calidad del panel de referencia 39
5.2. Resumen del mapa genético 41
5.3. Imputación de datos 43
5.4. Evaluación del desempeño de la imputación 45
6. CONCLUSIONES 48
7. BIBLIOGRAFÍA 49
8. ANEXOS 60

ÍNDICE DE TABLAS

Tabla 1. Archivos finales obtenidos con Impute2 21

Tabla 2. Número total de polimorfismos por parental 26

Tabla 3. Distribución de SNPs por cromosoma para los datos de la generación S3 y S5 de la población MAGIC. 28

Tabla 4. Comparación de la cobertura de SNPs por cromosoma entre la población S3 y S5. 33

Tabla 5. Resumen de los parámetros de imputación de cada programa 33

Tabla 6. ANOVA del valor de R^2 de los datos totales para cada programa de imputación 36

Tabla 7. ANOVA del valor de R^2 de los datos imputados de novo para cada programa de imputación 36

ÍNDICE DE LISTAS DE COMANDOS

Lista 1. Conjunto de comandos empleados para la instalación y ejecución de los programas de imputación. a) Beagle 5.1. b) Impute 2. c) Minimac 4. 15

Lista 2. Conjunto de comandos empleados para la elaboración del panel de referencia. a) División de los archivos VCF por cromosoma. b) Eliminación de marcadores duplicados. 16

Lista 3. Conjunto de comandos empleados para el phasing del panel de referencia. a) Phasing con Beagle 5.1. del cromosoma 1. b) Edición del encabezado. 19

Lista 4. Conjunto de comandos empleados para la transformación del panel de referencia a los formatos necesarios para la imputación. a) Transformación a HAPS y LEGEND para Impute 2. b) Transformación a MSAV para Minimac 4. 19

Lista 5. Conjunto de comandos empleados para la imputación con Beagle 5.1. a) Imputación cromosoma 1. b) Unión de los archivos por cromosoma en uno solo. c) Cálculo del valor MAF por posición. 20

Lista 6. Conjunto de comandos empleados para la imputación con Impute 2. a) Transformación de la población objetivo a formato GENS. b) Imputación del cromosoma 1 por chunks. c) Unión de los chunks imputados en un archivo único. d) Convertir los archivos de output a formato VCF. e) Unión de los archivos por cromosoma en uno solo. f) Calculo del valor MAF por posición. 21

Lista 7. Conjunto de comandos empleados para la imputación con Minimac 4. a) Cromosoma 1. b) Unión de los archivos por cromosoma en uno solo. 23

Lista 8. Obtención de las métricas de calidad. a) Extracción de R^2 y MAF de los resultados obtenidos con Beagle. b) Extracción de InfoScore (R^2) de los resultados obtenidos con Impute. c) Extracción de MAF de los resultados obtenidos con Impute. d) Unión de los valores R^2 y MAF, obtenidos a partir de Impute, en un solo archivo. e) Extracción de R^2 y MAF de los resultados obtenidos con Minimac. 24

ÍNDICE DE FIGURAS

Figura 1. Diversidad fenotípica en frutos de berenjena (Knapp et al., 2013). 1

Figura 2. Representación de la relación taxonómica de la berenjena cultivada (*S. melongena*) con otras especies del género Solanum (Taher et al., 2017). 3

Figura 3. Producción promedio de berenjena entre el año 2010 y 2019 a nivel mundial y en los 10 mayores productores (Solberg et al., 2021). 4

Figura 4. Distribución de la producción de berenjena en España (Ministerio de Agricultura, 2007). 5

Figura 5. Esquema de desarrollo de una población MAGIC a partir de 8 parentales (Stadlmeier et al., 2018). 7

Figura 6. Proceso de imputación genética usando un panel de referencia (Treccani et al., 2023). 9

Figura 7. Desarrollo de la población multiparental de intercruzamientos avanzados de berenjena a partir de 8 líneas fundadoras (Mangino, et al., 2022). 14

Figura 8. Distribución de SNPs por cromosoma con respecto a la longitud total en mega pares de bases (Mb). 27

Figura 9. Mapa genético para los 12 cromosomas. 32

Figura 10. Relación entre MAF y R^2 por marcador. a) Datos generales. b) Datos imputados de novo. c) Datos originales imputados. 34

Figura 11. Relación entre MAF y R^2 por marcador en los datos originales imputados 34

Figura 12. Relación entre MAF y R^2 por marcador. a) Datos generales. b) Datos imputados de novo. c) Datos originales imputados. 35

Figura 13. R^2 promedio por programa. a) Datos generales. b) Datos imputados de novo. 36

Figura 14. Distribución de los marcadores según su valor de R^2 por programa..... 37

ÍNDICE DE ANEXOS

Anexo 1. Script en Python para la preparación de los datos para usarlos en R/mpMap2 60

Anexo 2. Script en R para la construcción del mapa genético 65

Anexo 3. Script en R para la visualización del mapa genético mediante un gráfico de dispersión 66

Anexo 4. Script en R para la transformación del mapa genético a formato PLINK . 66

Anexo 5. Script en R para la transformación del mapa genético a formato .map.... 67

Anexo 6. Script en R para la elaboración de los gráficos a partir de los resultados de los procesos de imputación 67

ANEXO I. RELACIÓN DEL TRABAJO CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE DE LA AGENDA 2030

Anexo al Trabajo de Fin de Grado y Trabajo de Fin de Máster: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**		X		
ODS 3. **Salud y bienestar.**		X		
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				X
ODS 9. **Industria, innovación e infraestructuras.**		X		
ODS 10. **Reducción de las desigualdades**.				X
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**			X	
ODS 13. **Acción por el clima.**		X		
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**				X

Descripción de la alineación del TFG/TFM con los ODS con un grado de relación más alto.

Este estudio se encuentra relacionado en mayor medida con los objetivos 2, 3, 9 y 13, ya que se está contribuyendo para el desarrollo de nuevas variedades de berenjena que se puedan adaptar a las condiciones de cultivo necesarias para combatir el cambio climático, además de ayudar a satisfacer las necesidades de producción de alimento para la población.

1. INTRODUCCIÓN

1.1. La berenjena (*Solanum melongena* L.)

1.1.1. Descripción de la especie

La berenjena común (*Solanum melongena* L., 2n = 2x = 24) es una planta herbácea que desarrolla una estructura arbustiva con una altura que oscila entre 60 y 120 centímetros (Solanke & Tawar, 2019). Sus hojas son simples, grandes y ovaladas, y pueden presentar espinas en la parte media (Dash et al., 2019). La flor es completa, actinomorfa, hermafrodita y de tipo pentámera (Khaleghi et al., 2021). Generalmente, las flores poseen una coloración violeta y se presentan en inflorescencias de entre dos a siete botones florales. Su fruto es una baya carnosa comestible que puede encontrarse en varios colores, como por ejemplo violeta, verde, negro, blanco y amarillo en función de la presencia de clorofilas y antocianos (Figura 1). Además, pueden presentar rayas más claras (Solanke & Tawar, 2019). La carne del fruto es dura y su piel gruesa. La forma y tamaño de la berenjena son muy variables, habiendo desde variedades redondas hasta alargadas, con pesos que van desde unos pocos gramos hasta un kilogramo (Solanke & Tawar, 2019).

Figura 1. Diversidad fenotípica en frutos de berenjena (Knapp et al., 2013).

La berenjena se cultiva principalmente en las regiones tropicales y subtropicales del mundo, ya que requiere de temperaturas cálidas para su óptimo desarrollo vegetativo. Estas temperaturas, no deben de ser menores de 15ºC ni mayores de 35ºC. Existen evidencias de que a temperaturas menores las plantas detienen su desarrollo y

pueden llegar a abortar tanto las flores en formación como los frutos; mientras que temperaturas cercanas a 40ºC pueden causar deformaciones en los frutos y que el polen deje de ser viable (León Pacheco et al., 2019). Bajo las condiciones adecuadas, el ciclo productivo ronda los 9 meses y una planta de berenjena puede dar entre 20 a 30 cosechas (Adarraga Mejía et al., 2022).

1.1.2. Clasificación taxonómica

La especie *S. melongena* L., pertenece a la familia Solanaceae y al género *Solanum*. Esta familia contiene alrededor de 3.000 especies distribuidas en 90 géneros. El género *Solanum* es el más grande de ellos, abarcando el 50% de las especies de la familia (Vorontsova & Knapp, 2012). A su vez, este género se subdivide en 13 clados, con la berenjena ubicada en el clado *Leptostemonum* (Taher et al., 2017). Este subgénero se caracteriza por agrupar 450 especies espinosas, con espinas epidérmicas en sus tallos y hojas (Vorontsova et al., 2013).

1.1.3. Especies silvestres

Como consecuencia de la domesticación, al igual que en muchos de los cultivos que han pasado por este proceso, la berenjena común ha perdido gran parte de diversidad genética y en la actualidad su base genética es estrecha. Sin embargo, las especies silvestres emparentadas con la berenjena suelen presentar mayor diversidad y sirven como recurso genético para la mejora de los cultivares de berenjena actuales (Swarup et al., 2021). Según su capacidad de hibridación con la berenjena común, las especies silvestres relacionadas se pueden clasificar en tres acervos genéticos (Figura 2). Dentro del primer grupo, o acervo primario (GP1), se encuentra la berenjena común y su ancestro *S. insanum*. En el segundo grupo (GP2), o acervo secundario, se encuentran ancestros más silvestres de la berenjena como el clado “berenjena”, conformado por *S. campylacanthum*, *S. lichtensteinii* y *S. linnaeanum*, el clado “Madagascar”, conformado por *S. pyracanthos*, y el grado “anguivi”, conformado por *S. anguivi* y *S. dasyphyllum*. Finalmente, en el tercer grupo (GP3), o acervo terciario, se encuentran especies más alejadas, como por ejemplo *S. torvum* y *S. sisymbriifolium* (Plazas et al., 2016). La obtención de híbridos de berenjena es más sencilla dentro del primer acervo genético. Sin embargo, en los acervos genéticos posteriores, aunque es posible obtener híbridos, a menudo resultan estériles, por lo

que es necesario recurrir a técnicas como el rescate de embriones (Plazas et al., 2016).

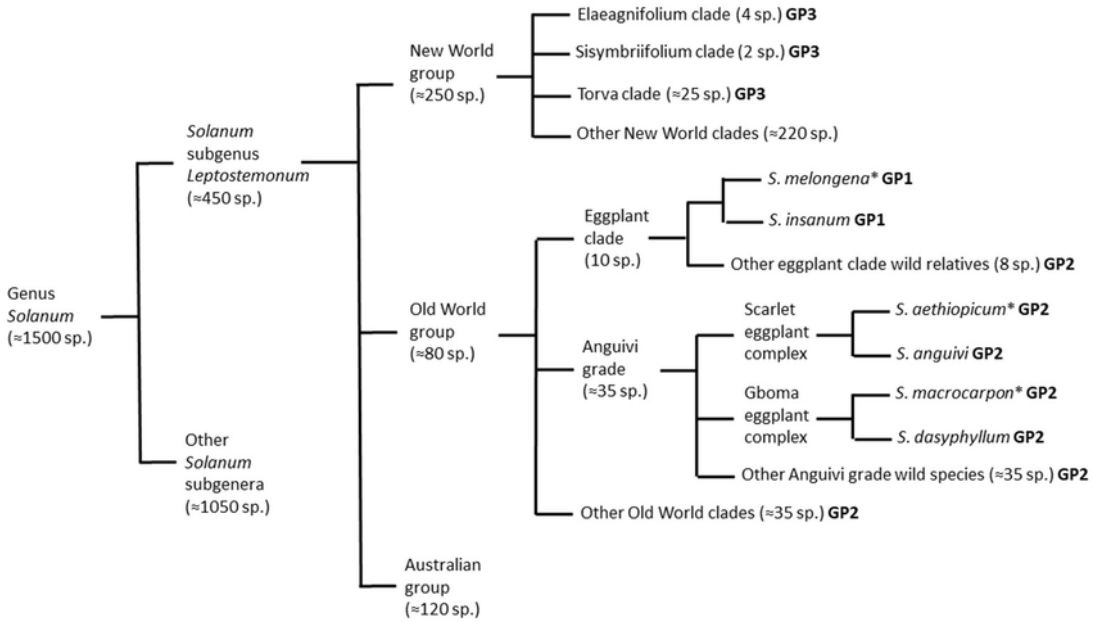


Figura 2. Representación de la relación taxonómica de la berenjena cultivada (*S. melongena*) con otras especies del género *Solanum* (Taher et al., 2017).

1.1.4. Origen y domesticación

Existen varias zonas en Asia propuestas como centros de domesticación de la berenjena, como es el caso de India, China y Tailandia (Page et al., 2019). En estas regiones se han encontrado especies silvestres en forma de malezas, las cuales se cree que pueden pertenecer a los primeros parentales (Doganlar et al., 2002). El registro más antiguo sobre el cultivo y la domesticación de *S. melongena* está documentado en la literatura china antigua, en una obra del 59 a.C. Se ha encontrado que el proceso de domesticación de la berenjena en China implicó tres aspectos principales de la calidad del fruto: (I) el tamaño, que cambió de pequeño a grande; (II) la diversidad de formas; (III) y el sabor de la fruta, que se volvió más apetecible (Solanke & Tawar, 2019).

La berenjena fue domesticada en Asia a partir de *S. insanum*, una especie ampliamente distribuida en Asia tropical (Knapp et al., 2013). El cultivo de la berenjena se difundió hacia el oeste por la ruta de la seda en el siglo VIII, hasta que en el siglo XIV los comerciantes árabes llevaron la especie a África. Posteriormente, los comerciantes europeos la llevaron hasta América (Prohens et al., 2005).

La berenjena común se puede dividir en tres cultivares según su diversidad morfológica y su información geográfica: oriental, occidental y miniatura (Cericola et al., 2013). La berenjena occidental proviene del norte de África y Oriente Medio y generalmente produce frutos grandes de forma ovalada y color oscuro, su piel es fina y su sabor es dulce (Solberg et al., 2021). Por el otro lado, la berenjena oriental proviene tanto del sur como del este de Asia. Sus frutos son más pequeños, alargados y de una gran diversidad de colores (Solberg et al., 2021). Finalmente, las berenjenas miniatura son redondas, de diversos colores, con sabor dulce y textura suave (Solberg et al., 2021).

1.1.5. Importancia económica

La familia Solanaceae abarca algunas de las especies más importantes a nivel mundial, como son la patata (*S. tuberosum*), el tomate (*S. lycopersicum*) y los pimientos (*Capsicum annuum*). La berenjena, específicamente, es la sexta hortaliza más importante a nivel mundial, con un total de 103 millones de toneladas producidas en el año 2022 (FAO, 2024). La mayor parte de su producción se concentra en Asia, donde se genera un 90% del peso total producido (Figura 3; Solberg et al., 2021). Concretamente, China produce más del 50% de la cosecha mundial, seguida por India, con un 30%. En África, la producción se concentra en Egipto, Argelia, Costa de Marfil y Sudán. Por otro lado, los mayores productores de berenjena en Europa son Italia y España; mientras que en América el principal productor es México, seguido de Estados Unidos (Solberg et al., 2021).

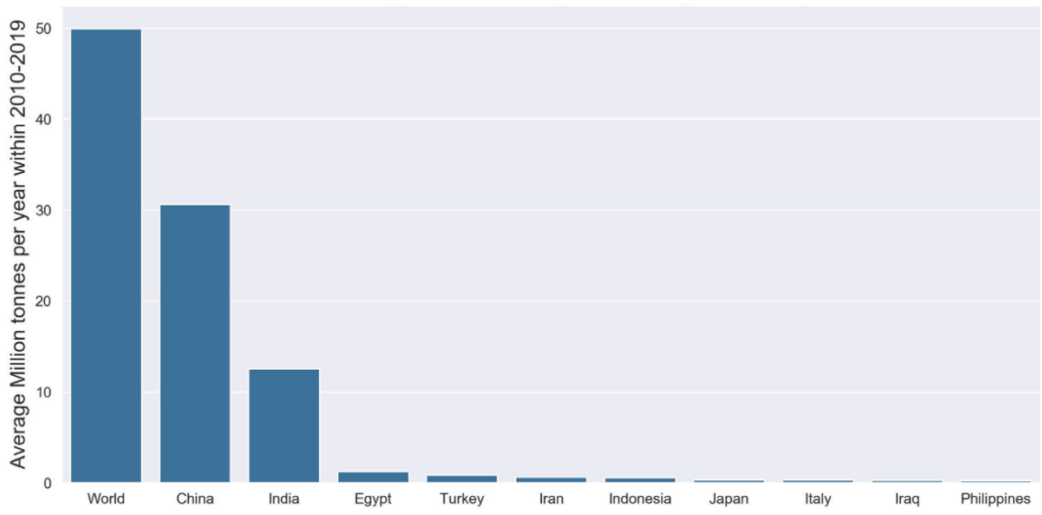


Figura 3. Producción promedio de berenjena entre el año 2010 y 2019 a nivel mundial y en los 10 mayores productores (Solberg et al., 2021).

En España, la producción de berenjena se concentra en las provincias de Almería, Barcelona, Valencia y Murcia (Figura 4); siendo Almería la responsable del 80% de la producción nacional (Hortoinfo, 2024). En el año 2023, España exportó 175 mil toneladas de berenjenas para consumo europeo, generando 220 millones de euros. De esta cantidad, 140 mil toneladas se produjeron en Almería, 11 mil toneladas en Barcelona y 5 mil toneladas en Valencia y Murcia (Hortoinfo, 2024).

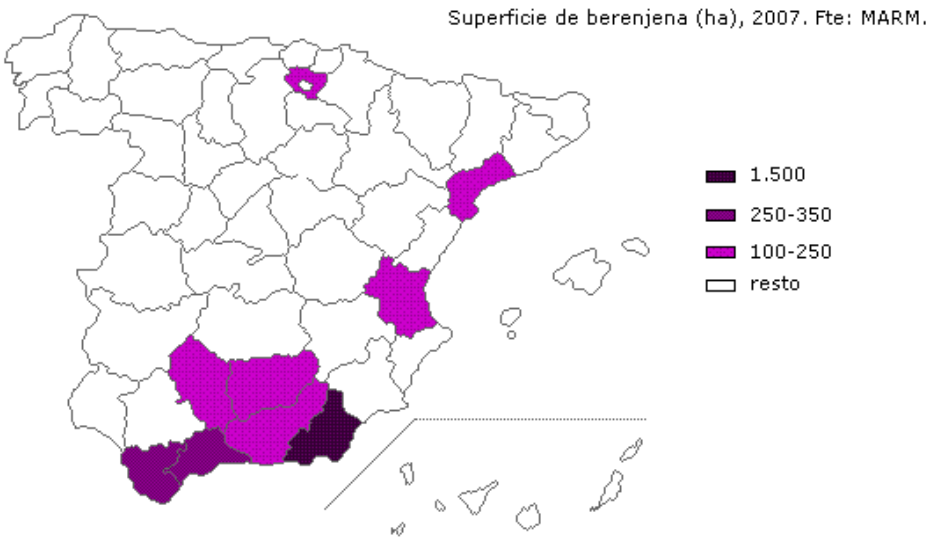


Figura 4. Distribución de la producción de berenjena en España (Ministerio de Agricultura, 2007).

La producción de este cultivo ha aumentado con los años debido a su gran demanda e interés, ya que la berenjena es rica en ácidos fenólicos y antocianinas, los cuales le confieren una actividad antioxidante que otorga múltiples beneficios para la salud (Solberg et al., 2021). Además, se conoce que tiene una alta capacidad para absorber radicales de oxígeno y neutralizar radicales libres (Sawa et al., 1998).

1.2. Mejora genética

El proceso de domesticación de las especies, de manera indirecta, causa una reducción en su variabilidad genética, lo cual conlleva a una disminución de su diversidad. Hoy en día, el proceso de mejora genética es prácticamente indispensable para cualquier especie destinada al consumo humano, ya que es necesario desarrollar variedades que tengan una producción sostenible y que, a su vez, se encuentren adaptadas a diferentes escenarios climáticos. Sin embargo, para poder realizar esta mejora, es necesario tener diversidad genética que permita adquirir ciertos caracteres de interés.

Generalmente, los programas actuales de mejora vegetal tienen como objetivo principal el desarrollo de variedades con elevados rendimientos, alta calidad, larga vida en postcosecha, resistencia a plagas y tolerancia a estreses ambientales (Taher et al., 2017).

1.2.1. Poblaciones experimentales

Existen varias estrategias que pueden ser empleadas en los programas de mejora genética de especies, dependiendo principalmente de la especie o especies con las que se van a trabajar, de los rasgos a mejorar y los recursos a disposición. Un enfoque de gran interés es el uso de poblaciones experimentales con el objetivo de estudiar rasgos complejos de un cultivo mediante la combinación de los genomas de parentales específicos (Scott et al., 2020). Dentro de los tipos de poblaciones experimentales tradicionales se encuentran las líneas recombinantes consanguíneas (RILs), las líneas de introgresión (ILs), las líneas casi isogénicas (NILs) y las líneas endogámicas de retrocruzamientos (BILs), entre otras (Keurentjes et al., 2007). En los últimos años, se han desarrollado poblaciones experimentales multiparentales, como es el caso de las poblaciones multiparentales de entrecruzamientos de generación avanzada, conocidas también como MAGIC.

El diseño de una población MAGIC es complejo (Figura 5). Generalmente, se emplean 4, 8 o 16 parentales que se cruzan en pares para posteriormente auto fecundarse y generar líneas puras (B. E. Huang et al., 2015). De esta manera, los cromosomas de la población son mosaicos aleatorios de los haplotipos de los parentales fundadores (Scott et al., 2020). El beneficio de las poblaciones MAGIC es que permite mezclar genomas de varios parentales con el objetivo de estudiar la diversidad genética de ese cultivo. En varios cultivos se han desarrollado poblaciones MAGIC, como por ejemplo el arroz (Bandillo et al., 2013), el trigo (Mackay et al., 2014), el maíz (Holland, 2015), el tomate (Pascual et al., 2015), la canola (Zhao et al., 2017), la soja (Shivakumar et al., 2018), el fréjol (Diaz et al., 2020), el cacahuete (Wankhade et al., 2021) y el algodón (M. Wang et al., 2022).

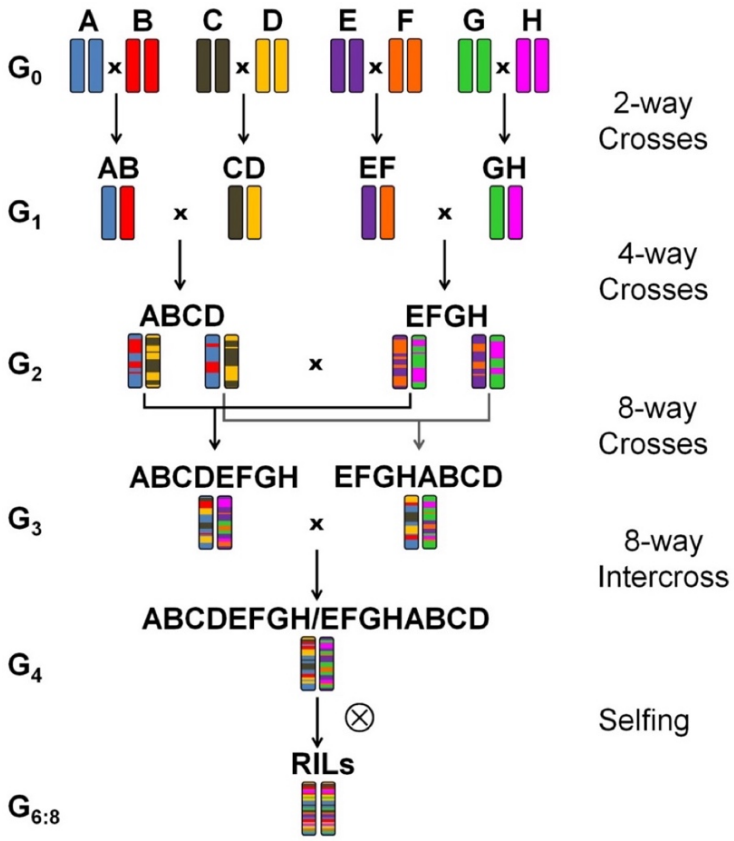


Figura 5. Esquema de desarrollo de una población MAGIC a partir de 8 parentales (Stadlmeier et al., 2018).

En el caso de la berenjena, previamente fueron desarrolladas poblaciones tanto RILs, en el año 2013 (Lebeau et al.), como ILs, en el año 2017 (Gramazio et al.), las cuales han permitido el estudio de varios caracteres de interés de este cultivo, como el tamaño y forma del fruto (Sulli et al., 2021), el color del fruto y su composición bioquímica (Toppino et al., 2020). La primera población MAGIC fue desarrollada en 2022 a partir de siete accesiones de *S. melongena* y una accesión de la especie *S. incanum* (Mangino et al., 2022). La generación S3 está compuesta por 350 líneas y se encuentra genotipada mediante la tecnología SPET, lo cual ha permitido el estudio del control génico de caracteres de interés, como por ejemplo la pigmentación de antocianos (Mangino et al., 2022) y los patrones de netting en los frutos (Arrones et al., 2022).

1.3. Imputación genética

En los programas de mejora genética, herramientas como la secuenciación son cada día más necesarias para acelerar los procesos y reducir recursos. Las técnicas de secuenciación más empleadas son: genotyping by sequencing (GBS), whole-genome

sequencing (WGS) y low-coverage whole genome sequencing (lcWGS) (Adhikari et al., 2022). La tecnología GBS emplea enzimas de restricción con la finalidad de simplificar y reducir la representación del genoma para su secuenciación siendo una técnica de menor coste respecto a WGS (Wickland et al., 2017). Sin embargo, al simplificar el genoma, la cobertura es menor, volviendo su aplicación menos atractiva (N. Wang et al., 2020). La tecnología WGS se basa en la secuenciación directa de todo el genoma, dando un resultado más preciso, pero a un costo más alto (Ng & Kirkness, 2010). Por último, la tecnología de lcWGS permite la secuenciación del genoma completo a baja profundidad para obtener un genotipado más completo de GBS y a un coste menor que WGS (Kumar et al., 2021).

Para el estudio de caracteres específicos, la secuenciación de alta cobertura, o WGS, es necesaria debido a su alta precisión. Sin embargo, esto puede llegar a ser muy costoso. Una solución a esto es la imputación de genotipos en datos de genotipado por lcWGS. La imputación es una técnica computacional que se puede emplear para inferir alelos de SNPs faltantes o no genotipados (Schurz et al., 2019). Esto es posible basándose en patrones del desequilibrio de ligamiento derivados de marcadores previamente secuenciados y comparando con un panel de referencia (Malhotra et al., 2014). El principio fundamental de la imputación genética es identificar secuencias de ADN similares entre la información objetivo y la de referencia (Stahl et al., 2021). A partir de esta información, se infiere la información faltante a partir del panel de referencia. Esto implica que ambos datos, referencia y objetivo, deben tener un origen similar para que los resultados sean precisos.

El realizar lcWGS complementado con la imputación genética presenta un gran beneficio a nivel económico con respecto a la secuenciación a elevadas coberturas en dos puntos principales. En primer lugar, el coste de secuenciación es menor a bajas coberturas que a elevadas coberturas, ya que se emplea menos mano de obra y menos insumos de laboratorio (Chat et al., 2022). Y, en segundo lugar, existe un mejor aprovechamiento de los recursos permitiendo estudios con poblaciones más grandes y cubriendo una mayor parte de estas (Y. Li et al., 2011).

La imputación genética funciona combinando un panel de referencia de individuos secuenciados a alta densidad con un alto número de sitios polimórficos (generalmente SNPs) con una muestra de estudio genéticamente similar y secuenciada a baja

cobertura (B. N. Howie et al., 2009). Esto se debe a que la imputación se basa en la identidad por descendencia, lo que quiere decir que dos individuos van a heredar una porción idéntica del genoma de sus parentales, lo que permite inferir información faltante de una población (Sticca et al., 2021). Dado que existe incertidumbre al momento de inferir esta información, se emplean modelos probabilísticos que permiten tener en cuenta esta incertidumbre y predecir la probabilidad para cada alelo posible en un marcador imputado en la población (Browning et al., 2018). De esta manera, los genotipos empleados como referencia son clave para obtener mayor precisión en la imputación.

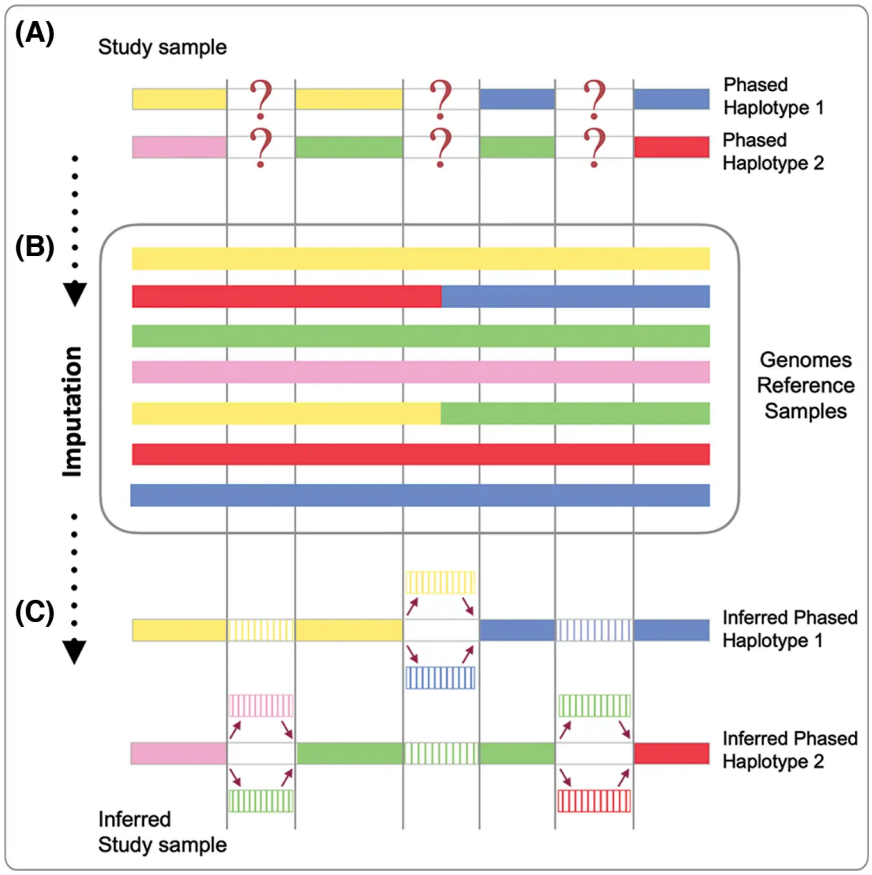


Figura 6. Proceso de imputación genética usando un panel de referencia (Treccani et al., 2023).

El proceso de imputación tiene dos pasos principales: deducción de los haplotipos e imputación de los genotipos faltantes (Figura 6; Stahl et al., 2021). Un haplotipo es la combinación de alelos en un solo cromosoma que se hereda de un solo padre (S. S. Li, 2003). Los haplotipos se deben deducir y reconstruir de manera estadística mediante la utilización de los datos observados y sus probabilidades de encontrarse en uno u otro de los dos cromosomas paternos. Este proceso se conoce como phasing (Baldrighi et al., 2022). Como segundo paso en el proceso de imputación, se

comparan los haplotipos previamente estimados con un grupo de haplotipos de referencia con el objetivo de obtener los genotipos de las variantes no observadas en la muestra del estudio (Marchini, 2019). Al existir muchos posibles escenarios, el resultado de la imputación se presenta como una probabilidad.

Se ha observado que la imputación genética de SNPs es un método poderoso para obtener marcadores genéticos en una muestra sin la necesidad de genotiparlos, por lo que es de gran utilidad para estudios de asociación. La imputación también podría servir como medio de control de calidad al resaltar posibles errores de genotipado y, además, podría ayudar en la reconstrucción de genotipos faltantes en miembros de la familia no tipificados en datos genealógicos (Ellinghaus et al., 2009).

1.3.1. Métodos de imputación genética

Hasta hoy, se han desarrollado e implementado varios algoritmos y modelos de imputación, la mayoría de los cuales se basan en el modelo Hidden Markov Model (HMM) desarrollado por Li y Stephens (De Marino et al., 2022). Un HMM es un tipo de modelo estadístico que permite describir la evolución de eventos observables que dependen de factores que no son directamente observables y son más bien internos o invisibles (Yoon, 2009). De esta manera, un HMM está compuesto por dos procesos: uno invisible de estados ocultos, que forman una cadena de Markov, y otro visible de símbolos observables, que posee una distribución de probabilidad que depende del estado anterior (Yoon, 2009). En otras palabras, este modelo propone que la secuencia del genoma de un individuo puede representarse mediante eventos de recombinación y en parte por mutaciones provenientes de otros individuos (Naito & Okada, 2024). Actualmente, la mayoría de los softwares disponibles para la imputación están basados en este modelo, como Beagle, Impute y Minimac (De Marino et al., 2022).

El software Beagle fue desarrollado por la Universidad de Washington en el año 2007, actualmente, la última versión es Beagle 5.4. Este programa está escrito en Java y, por lo tanto, se ejecuta en todas las plataformas que tengan un intérprete de Java adecuado (Naito & Okada, 2024). Beagle permite realizar tanto el phasing del panel de referencia como la imputación de genotipos. El modelo de este programa asume que el haplotipo objetivo corresponde a una ruta no observada y que cada marcador

corresponde a un espacio de estado HMM (Browning et al., 2018). Por lo tanto, el algoritmo usado calcula la probabilidad de que la ruta no observada pase por un estado HMM. En cada marcador, la suma de las probabilidades de los estados marcados con un alelo es la probabilidad imputada para ese alelo (Browning et al., 2018).

El software Impute fue desarrollado por la Universidad de Oxford en el año 2009, actualmente la última versión es Impute 5. Este programa funciona en las plataformas Linux, MacOS y Windows (Ellinghaus et al., 2009). El algoritmo de este programa combina un modelo HMM con un algoritmo Markov Chain Monte Carlo (MCMC) con la finalidad de separar el phasing de los haplotipos de la imputación y ser computacionalmente más eficiente (Ellinghaus et al., 2009). El algoritmo MCMC se encarga de realizar iteraciones durante el proceso de phasing de la población con la finalidad de obtener resultados más precisos.

El software Minimac fue desarrollado por la Universidad de Michigan en el año 2012, actualmente la última versión es Minimac 4. Al igual que el software Impute, el proceso se realiza en dos pasos, phasing e imputación de los genotipos faltantes de la población, usando tanto un modelo HMM como un modelo MCMC de manera simultánea (B. Howie et al., 2012).

1.3.2. Información necesaria para la imputación genética

Para realizar un estudio de imputación es imprescindible contar con una muestra de estudio a imputar. Dependiendo del algoritmo que se emplee en la imputación, se requerirá de información adicional. En la mayoría de los programas, la imputación genética funciona al combinar un panel de referencia con sitios polimórficos genotipados a alta profundidad con una muestra de estudio de una población genéticamente similar que se encuentra genotipada a menor profundidad para estos mismos sitios polimórficos (B. N. Howie et al., 2009). De esta manera, los sitios faltantes en la muestra de estudio se infieren a partir de este panel de referencia (Y. Li et al., 2009). En estos casos, es necesario contar con un panel de referencia obtenido de una población emparentada a esta muestra de estudio. Existen determinados programas que realizan la inferencia a partir de las lecturas obtenidas mediante la secuenciación de la muestra de estudio (Davies et al., 2016). Además de

estos datos, el análisis puede requerir de un mapa genético para la especie en estudio. El mapa genético documenta la forma en que varían las tasas de recombinación en un genoma, lo cual permite que la estimación de los haplotipos sea más precisa (Myers et al., 2005). Algunos modelos de imputación dependen de la tasa de recombinación estimada a partir de un mapa genético para inferir como los haplotipos van a cambiar a lo largo de una secuencia determinada; además, con esta información también se estima la tasa de mutación de la población (Marchini & Howie, 2010).

2. OBJETIVOS

Objetivo general

- Evaluar el uso de la imputación genética como herramienta complementaria a la secuenciación mediante lcWGS.

Objetivos específicos

- Imputar los genotipos faltantes de las 325 líneas de la generación S5 de una población MAGIC de berenjena resecuenciadas a baja cobertura.
- Comparar la precisión de la imputación genética de tres programas computacionales: Beagle 5.1, Impute 2 y Minimac 4.

3. MATERIALES Y MÉTODOS

3.1. Selección del material a imputar

El material vegetal utilizado en este estudio corresponde a la generación S5 de una población MAGIC de berenjena compuesta por 325 líneas obtenidas después de cruzar ocho parentales, en la forma en la que se ha explicado anteriormente, con un esquema de tipo 'funnel' y 5 ciclos de autofecundación (Figura 7).

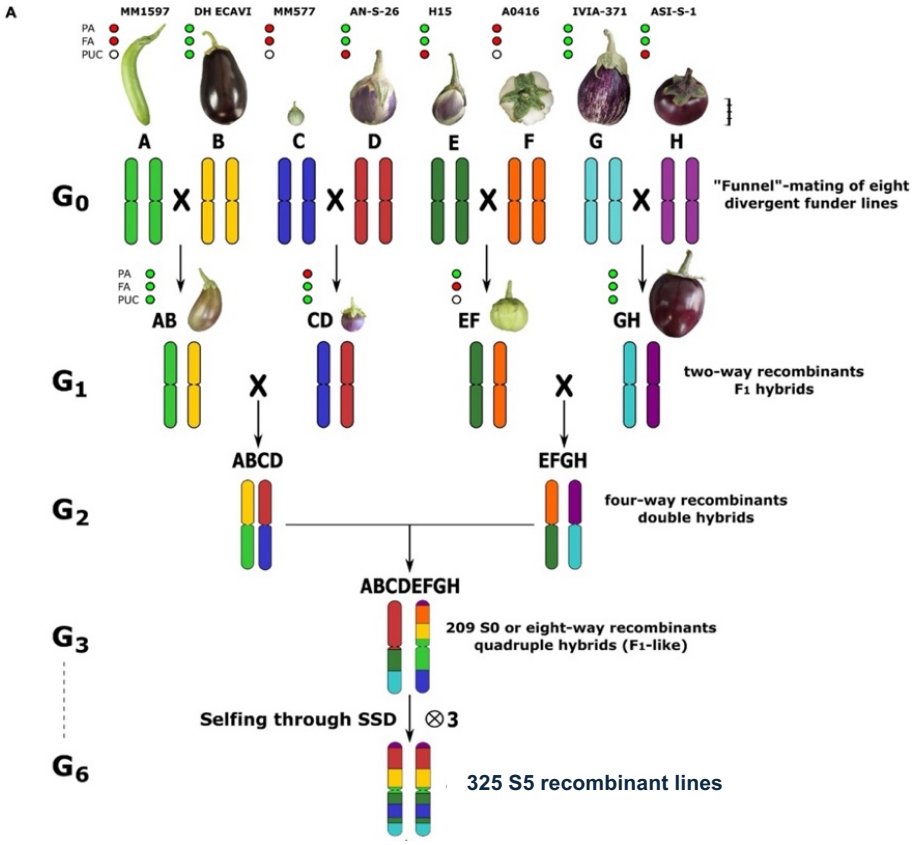


Figura 7. Desarrollo de la población multiparental de intercruzamientos avanzados de berenjena a partir de 8 líneas fundadoras (Mangino, et al., 2022).

Los parentales de la población fueron previamente secuenciados a una profundidad de 20X (Gramazio et al., 2019), mientras que las líneas S5 fueron resecuenciadas a una profundidad de 3X. Las lecturas limpias (clean reads) se mapearon contra el genoma de referencia de berenjena de alta calidad v3.0 "67/3" (Barchi, Pietrella, et al., 2019) utilizando la herramienta BWA-mem (versión v.0.7.17–r1188; Li, 2013). La identificación de variantes a nivel poblacional se llevó a cabo con Freebayes (versión 1.3.6; Garrison & Marth, 2012). Se utilizaron los parámetros por defecto del software, excepto los requisitos de calidad mínima de mapeo y de base, que se fijaron en un umbral de 20.

El post-filtrado de las variantes genéticas varió entre los fundadores y las líneas S5. De los archivos VCF generados para cada uno de los parentales se seleccionaron las variantes genéticas: (I) bialélicas, (II) soportadas por un máximo de 40 lecturas y un mínimo de 10 lecturas, (III) con una frecuencia del alelo alternativo superior a 0.3, (IV) polimórficas con respecto al genoma de referencia. Finalmente, la información se agrupó en un único archivo VCF. Por otro lado, del archivo VCF obtenido para la población, se seleccionaron los SNPs bialélicos compartidos con los parentales. A continuación, se filtraron con una profundidad de mapeo mínima de 3X y con un MAF (minor allele frequency) de 0,05. Finalmente, se eliminaron los sitios monomórficos y aquellos donde el porcentaje de heterocigotos superaba el 20% del total de los individuos.

3.2. Instalación de los programas

Los programas para la manipulación e imputación de los datos se instalaron en un servidor con sistema operativo Ubuntu 22.04.4. Para la manipulación de los archivos VCF, se utilizó el paquete de herramientas Bcftools (versión 1.13; Danecek et al., 2021). Para la imputación de datos, se utilizaron los programas Beagle 5.1 (versión 18May20.d20; Browning & Browning, 2016), Impute 2 (versión 2.3.2.; Howie et al., 2009) y Minimac 4 (versión 4.1.6.; Fuchsberger et al., 2015) ya que hoy en día son los programas de licencia libre más utilizados para la imputación genética y se han empleado previamente en otras investigaciones similares (Liu et al., 2015; Korkuć et al., 2019; De Marino et al., 2022). La lista con los comandos empleados para la instalación de todos los programas se presenta a continuación (Lista 1).

Lista 1. Conjunto de comandos empleados para la instalación y ejecución de los programas de imputación. **a)** Beagle 5.1. **b)** Impute 2. **c)** Minimac 4.

a	`wget https://faculty.washington.edu/browning/beagle/beagle.18May20.d20.jar` `java -jar beagle.18May20.d20.jar`
b	`wget https://mathgen.stats.ox.ac.uk/impute/impute_v2.3.2_x86_64_static.tgz` `tar -zxvf impute_v2.3.2_x86_64_static.tgz`

c	`git clone https://github.com/statgen/Minimac4.git`
	`cd Minimac4`
	`bash minimac4-4.1.6-Linux-x86_64.sh`

3.3. Elaboración del panel de referencia

Para la elaboración del panel de referencia se emplearon los datos de la resecuenciación de los ocho parentales que dan origen a la población MAGIC. Los parentales corresponden a las siguientes accesiones: MM1597, DH-ECAVI, MM577, AN-S-26, H15, A0416, IVIA-371 y ASI-S-1. El archivo inicial contiene la información en conjunto de todos los parentales y se encuentra previamente filtrado, proporcionando solo la información de SNPs bialélicos.

Con esta información, primero se dividió cada archivo por cromosomas empleando la función split del paquete Bcftools (Lista 2.a, versión 1.13; Danecek et al., 2021). Posteriormente, se indexó el archivo y se realizó un nuevo filtrado de los datos donde se eliminaron las posiciones duplicadas de los marcadores usando el comando awk (Lista 2.b). Esto se realizó porque, al identificar un SNP, Freebayes en algunas posiciones reportó tanto el SNP aislado, como el SNP acompañado por algunas bases contiguas. Como resultado, se obtuvieron 12 archivos en formato VCF, uno por cada cromosoma.

Lista 2. Conjunto de comandos empleados para la elaboración del panel de referencia. **a)** División de los archivos VCF por cromosoma. **b)** Eliminación de marcadores duplicados.

a
```
bcftools index -s
merge_all_parents_mapped_BWA_v3_20X_dedup_final_biallelic_snps_mindp20_nomono
.vcf.gz | cut -f 1 | while read C; do bcftools view -O z -o split.${C}.vcf.gz
merge_all_parents_mapped_BWA_v3_20X_dedup_final_biallelic_snps_mindp20_nomono
.vcf.gz "${C}" ; done
```

b
```
awk 'BEGIN{FS=OFS="\t"} /^#/ {print; next} (length($4) == 1 && length($5) ==
1 && $4 !~ /N/ && $5 !~ /N/) {print}' split.SMEL3_ABCDEFGH_Ch01.vcf >
SMEL3_ABCDEFGH_Ch01.vcf
```

3.4. Elaboración del mapa genético

Para obtener resultados más precisos en la imputación, ciertos modelos requieren un mapa genético con la finalidad de estimar la tasa de mutación de la población y la tasa de recombinación entre los marcadores que serán empleados. Por esta razón, previo al proceso de imputación, se elaboró un mapa genético para los doce cromosomas de berenjena empleando el paquete mpMap2 de R (versión v0.0.6; Shah et al., 2020). Se generó un script en Python (Anexo 1) para la preparación de los archivos de entrada, que fueron tres: pedigree (información de los cruces realizados

para obtener la población), founders (información de los SNPs de los parentales fundadores de la población) y finals (información de los SNPs de la población).

Se usó dos sets de datos de dos poblaciones MAGIC distintas: (I) datos de SPET de la generación S3 y (II) datos de resecuenciación de la generación S5. Esto se realizó con la finalidad de comparar los resultados de ambos sets de datos y elegir aquel que arrojase un mapa más preciso. La generación S3 está compuesta por 420 individuos, que fueron secuenciados con la tecnología de secuenciación SPET (Single primer enrichment technology) (Mangino et al., 2022). Por el otro lado, la generación S5 está compuesta por 325 individuos, que fueron secuenciados mediante lcWGS a una profundidad de 3x. Mediante el software Tassel se eliminaron los SNPs monomórficos (minimum frequency < 0,01) como mínimo para el 90% del total de las secuencias (minimum count) de los datos de la generación S3. El conjunto de variantes identificadas en las líneas de la generación S5 fueron filtradas como se ha indicado previamente.

El script preparado constaba de varios pasos. Primero se convirtieron los archivos VCF con la información de los SNPs de la población y de los parentales a archivos de tipo CSV, que es el formato requerido por R/mpMap2. A continuación, de los archivos generados se seleccionó la información relativa al cromosoma, a la posición de la variante genética y a los alelos. Además, se traspuso la tabla, de forma que cada marcador se encontrara en una columna y cada individuo en una fila. Los datos de los alelos de los SNPs de ambos archivos se transformaron a formato binario (0: homocigoto para el alelo de referencia; 1: heterocigoto; 2: homocigoto para el alelo alternativo). Mientras que los de los SNPs multialélicos se reemplazaron por “NA”. En el archivo correspondiente a la información de los parentales, se eliminaron los SNPs: (I) heterocigotos, (II) con información faltante y (III) homocigotos para el alelo de referencia para todos los individuos, ya que son situaciones de alta complejidad que el programa no las va a tomar en cuenta de la manera correcta (Shah et al., 2020). El objetivo fue mantener únicamente los SNPs bialélicos. Por último, se comparó el archivo de los parentales con el archivo de la población para seleccionar en cada archivo los polimorfismos presentes en ambos. Con los archivos ya en el formato necesario, se hizo un script en R para la construcción del mapa genético (Anexo 2). Con el paquete R/mpMap2 se creó un objeto mpcross con toda esta información. Este

objeto de tipo mpcross representa la población multiparental, ya que recibe los datos genéticos sobre las líneas fundadoras y la población final, un pedigree e información sobre cómo se han codificado los marcadores heterocigotos. Posteriormente, se procedió a estimar la fracción de recombinación entre los marcadores con la función estimateRF y se estimó las distancias genéticas con la función estimateMap usando el algoritmo Haldane y los parámetros establecidos por defecto. Finalmente, esta información se exportó en formato CSV. Para la visualización de los mapas finales, se elaboró un gráfico de dispersión usando R para cada cromosoma (Anexo 3).

El formato del mapa genético es diferente para cada programa de imputación. En el caso de Beagle 5.1 (versión 18May20.d20; Browning & Browning, 2016), se redactó un script de R (Anexo 4) para transformar el mapa genético a formato PLINK, que consiste en cuatro columnas sin encabezado (cromosoma, ID marcador, distancia genética, y posición física). Por otro lado, para Impute 2 (versión 2.3.2.; Howie et al., 2009) se transformó a formato MAP con un script de R (Anexo 5). Este formato consiste en tres columnas (posición física, tasa de recombinación, y distancia genética). La tasa de recombinación se calculó con la siguiente fórmula:

$$tasa\ de\ recombinación = \frac{distancia_f - distancia_o}{posición_f - posición_o} \times 10^6$$

Al final, tanto el archivo para Beagle como el archivo para Impute tienen la misma extensión (MAP), pero los formatos y la información que contienen son distintos entre sí. Minimac (versión 4.1.6.; Fuchsberger et al., 2015) no requiere del mapa genético, por lo que no se realizó ninguna adecuación adicional.

3.5. Phasing

Con el panel de referencia elaborado, se realizó el phasing de la información. Para este proceso se empleó el programa Beagle 5.1 (versión 18May20.d20; Browning & Browning, 2016) con el comando que se presenta en la Lista 3.a. Posteriormente, con Bcftools (versión 1.13; Danecek et al., 2021), se editó el encabezado (header) del archivo, ya que al hacer el phasing, Beagle elimina la información sobre los contigs, la cual es necesaria para la imputación. Para este proceso, en un archivo de texto se insertó la totalidad del encabezado deseado para, a continuación, con la función reheader de Bcftools añadirlo al archivo VCF resultante del phasing (Lista 3.b).

Lista 3. Conjunto de comandos empleados para el phasing del panel de referencia. **a)** Phasing con Beagle 5.1. del cromosoma 1. **b)** Edición del encabezado.

a	`java -Xmx200g -jar beagle.18May20.d20.jar gt=SMEL3_ABCDEFGH_Ch01.vcf out=phased_SMEL3_ABCDEFGH_Ch01 map=chr01_beagle.map impute=false`
b	`bcftools reheader -h contig_header.txt phased_SMEL3_ABCDEFGH_Ch01.vcf.gz -o phased_SMEL3_ABCDEFGH_Ch01.vcf.gz`

Finalmente, el archivo completo se transformó a los formatos deseados por cada programa de imputación: VCF para Beagle, HAPS y LEGEND para Impute y MSAV para Minimac. En el caso de Beagle, no fue necesario realizar ningún cambio porque el archivo final del phasing ya se encontraba en formato VCF. Para conseguir el formato de entrada de Impute se empleó la función convert de Bcftools (Lista 4.a), mientras que para transformar los datos al formato de Minimac se empleó una función propia del programa (Lista 4.b).

Lista 4. Conjunto de comandos empleados para la transformación del panel de referencia a los formatos necesarios para la imputación. **a)** Transformación a HAPS y LEGEND para Impute 2. **b)** Transformación a MSAV para Minimac 4.

a	`bcftools convert --haplegendsample --vcf-ids phased_SMEL3_ABCDEFGH_Ch01.vcf.gz -O z -o phased_SMEL3_ABCDEFGH_Ch01`
b	`./minimac4 --compress-reference phased_SMEL3_ABCDEFGH_Ch01.vcf.gz > phased_SMEL3_ABCDEFGH_Ch01.msav`

3.6. Imputación de datos

3.6.1. Beagle 5.1

Para la imputación de los genotipos faltantes con el programa Beagle 5.1 (versión 18May20.d20; Browning & Browning, 2016), se utilizaron como datos de entrada: el VCF con los SNPs filtrados de la población a imputar (gt), el panel de referencia faseado (ref) y el mapa genético (map). Beagle se lanzó con los parámetros establecidos por defecto (Lista 5.a). El archivo VCF resultante de la imputación para cada cromosoma almacenó los valores de la correlación estimada entre el alelo estimado y el verdadero (DR^2) y la frecuencia alélica (AF). Se utilizó la función concat de Bcftools (versión 1.13; Danecek et al., 2021) para combinar la información de cada archivo de salida en un único archivo (Lista 5.b). Para el análisis de los resultados, se calculó el valor MAF para cada marcador usando la función fill-tags de Bcftools (Lista 5.c, versión 1.13; Danecek et al., 2021).

Lista 5. Conjunto de comandos empleados para la imputación con Beagle 5.1. **a)** Imputación cromosoma 1. **b)** Unión de los archivos por cromosoma en uno solo. **c)** Cálculo del valor MAF por posición.

```
a  java -Xmx50g -jar beagle.18May20.d20.jar gt=split.SMEL3Ch01.vcf.gz
   ref=phased_SMEL3_ABCDEFGH_Ch01.vcf.gz
   out=imputed_Eggplant_MAGIC_2023_S5_Ch01_gp_nthreads20_beagle
   map=chr01_beagle.map nthreads=20
```

```
b  bcftools concat
   imputed_Eggplant_MAGIC_2023_S5_Ch01_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch02_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch03_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch04_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch05_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch06_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch07_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch08_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch09_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch10_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch11_gp_nthreads20_beagle.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch12_gp_nthreads20_beagle.vcf.gz >
   imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_beagle.vcf
```

```
c  bcftools +fill-tags imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_beagle.vcf >
   imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_MAF_beagle.vcf -- -t MAF
```

3.6.2. Impute 2

El programa Impute 2 (versión 2.3.2.; Howie et al., 2009) requiere que el archivo de la población a imputar (-g) se encuentre en formato GEN, para lo que se utilizó la función convert –gensample de Bcftools (Lista 6.a, versión 1.13; Danecek et al., 2021). El resto de los archivos de entrada requeridos son el panel de referencia faseado en formato HAP (-h) y LEGEND (-l) y el mapa genético en formato MAP (-m). La información del panel de referencia se divide en dos archivos distintos, uno con la información de los haplotipos (HAP) y otro con la información de los SNPs (LEGEND). El programa Impute se ejecutó con los parámetros establecidos por defecto, excepto por el parámetro -int, que indica los límites para realizar la imputación, y el parámetro -phase, que arroja como resultado un archivo extra con la información de los haplotipos de la población imputada. Por demanda del programa, la imputación se realizó en ventanas de 5 Mb (Lista 6.b). Posteriormente, se unieron todos los archivos resultantes en uno general por cromosoma con la función cat (Lista 6.c).

Como resultado de la imputación, se obtuvieron cinco archivos (Tabla 1). Para los análisis posteriores se empleó únicamente dos archivos (Archivo 2 y 4). La información de los haplotipos en formato HAPS se transformó a formato VCF con la función convert –hapsample2vcf de Bcftools (Lista 6.d). Al igual que con Beagle (versión 18May20.d20; Browning & Browning, 2016), se calculó el valor MAF para cada marcador usando la función fill-tags de Bcftools (Lista 6.f).

Tabla 1. Archivos finales obtenidos con Impute2

Archivo	Extensión	Información
1	.impute2_summary	Resumen de los parámetros usados para el análisis
2	.impute2_haps	Genotipos originales imputados
3	.impute2	Información de los marcadores y un conjunto de tres probabilidades del genotipo (AA, AB, BB) de cada individuo
4	.impute2_info	Tabla informativa con: (I) campo INFO, que corresponde a la medida de la información estadística observada asociada con la estimación de la frecuencia alélica para cada marcador la cual toma valores entre 0 y 1, donde los valores cercanos a 1 indican que un SNP se ha imputado con alta certeza, (II) campo CERTAINTY, que corresponde a la certeza promedio de los genotipos mejor estimados, y (III) campo TYPE, donde un valor igual a 0 o 1 indica que el marcador fue imputado, y un valor igual a 2 indica que el marcador no fue imputado y ya existía previamente
5	.impute2_warnings	Alertas generadas por el programa durante el proceso de imputación

Lista 6. Conjunto de comandos empleados para la imputación con Impute 2. **a)** Transformación de la población objetivo a formato GENS. **b)** Imputación del cromosoma 1 por chunks. **c)** Unión de los chunks imputados en un archivo único. **d)** Convertir los archivos de output a formato VCF. **e)** Unión de los archivos por cromosoma en uno solo. **f)** Calculo del valor MAF por posición.

```
a  bcftools convert --gensample --vcf-ids split.SMEL3Ch01.vcf -O z -o SMEL3Ch01
```

b	./impute2 -m chr01_impute.map -h phased_SMEL3_ABCDEFGH_Ch01.hap.gz -l phased_SMEL3_ABCDEFGH_Ch01.legend.gz -g SMEL3Ch01.gen.gz -int 1 5000000 -o imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk1.impute2 -phase
c	cat imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk1.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk2.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk3.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk4.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk5.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk6.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk7.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk8.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk9.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk10.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk11.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk12.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk13.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk14.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk15.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk16.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk17.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk18.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk19.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk20.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk21.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk22.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk23.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk24.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk25.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk26.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk27.impute2_info imputed_Eggplant_MAGIC_2023_S5_Ch01_impute_chunk28.impute2_info > imputed_Eggplant_MAGIC_2023_S5_Ch01_impute.impute2_info
d	bcftools convert --hapsample2vcf imputed_Eggplant_MAGIC_2023_S5_Ch01_impute.hap,SMEL3Ch01.samples -Oz -o imputed_Eggplant_MAGIC_2023_S5_Ch01_impute.vcf.gz
e	bcftools concat imputed_Eggplant_MAGIC_2023_S5_Ch01_impute.vcf.gz imputed_Eggplant_MAGIC_2023_S5_Ch02_impute.vcf.gz imputed_Eggplant_MAGIC_2023_S5_Ch03_impute.vcf.gz imputed_Eggplant_MAGIC_2023_S5_Ch04_impute.vcf.gz

```
   imputed_Eggplant_MAGIC_2023_S5_Ch05_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch06_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch07_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch08_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch09_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch10_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch11_impute.vcf.gz
   imputed_Eggplant_MAGIC_2023_S5_Ch12_impute.vcf.gz -Oz -o
   imputed_Eggplant_MAGIC_2023_S5_impute.vcf.gz
f  bcftools +fill-tags imputed_Eggplant_MAGIC_2023_S5_impute.vcf.gz >
   imputed_Eggplant_MAGIC_2023_S5_MAF_impute.vcf -- -t MAF
```

3.6.3. Minimac 4

Los archivos de entrada para el programa Minimac 4 (versión 4.1.6.; Fuchsberger et al., 2015) fueron el VCF con los SNPs filtrados de la población a imputar y el panel de referencia faseado, previamente transformado a formato MSAV. A diferencia de Beagle 5.1 e Impute 2, no requiere de un mapa genético para realizar la imputación. Minimac 4 fue ejecutado con los parámetros establecidos por defecto (Lista 7.a). La imputación se realizó para cada cromosoma de forma separada, por lo que, posteriormente, mediante la función concat de Bcftools (versión 1.13; Danecek et al., 2021) se combinó toda la información en un único archivo final (Lista 7.b). Este archivo VCF arrojó el valor de correlación entre los genotipos imputados y los genotipos verdaderos no observados (R^2), el MAF para cada posición original e imputada y un campo IMPUTED, donde un valor igual a 1 indica que el marcador fue imputado.

Lista 7. Conjunto de comandos empleados para la imputación con Minimac 4. **a)** Cromosoma 1. **b)** Unión de los archivos por cromosoma en uno solo.

```
a  ./minimac4 phased_SMEL3_ABCDEFGH_Ch01.msav split.SMEL3Ch01.vcf.gz -o
   imputed_Eggplant_MAGIC_2023_S5_Ch01_minimac.vcf
b  bcftools concat imputed_Eggplant_MAGIC_2023_S5_Ch01_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch02_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch03_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch04_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch05_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch06_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch07_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch08_minimac.vcf
   imputed_Eggplant_MAGIC_2023_S5_Ch09_minimac.vcf
```

```
    imputed_Eggplant_MAGIC_2023_S5_Ch10_minimac.vcf
    imputed_Eggplant_MAGIC_2023_S5_Ch11_minimac.vcf
    imputed_Eggplant_MAGIC_2023_S5_Ch12_minimac.vcf >
    imputed_Eggplant_MAGIC_2023_S5_minimac.vcf
```

3.7. Evaluación del desempeño de la imputación

Posterior a la imputación, es importante determinar la calidad de los resultados. Se extrajeron las métricas de calidad obtenidas por cada programa empleado: DR^2 para Beagle 5.1 (Lista 8.a, versión 18May20.d20; Browning & Browning, 2016), InfoScore para Impute 2 (Lista 8.b, versión 2.3.2.; Howie et al., 2009) y R^2 para Minimac 4 (Lista 8.e, versión 4.1.6.; Fuchsberger et al., 2015), además del MAF para todos ellos. Se empleó la función query de Bcftools (versión 1.13; Danecek et al., 2021) para los archivos VCF de Beagle y Minimac, (Lista 8.a, Lista 8.e) y la función awk para el archivo IMPUTE2 de Impute (Lista 8.c).

Lista 8. Obtención de las métricas de calidad. **a)** Extracción de R^2 y MAF de los resultados obtenidos con Beagle. **b)** Extracción de InfoScore (R^2) de los resultados obtenidos con Impute. **c)** Extracción de MAF de los resultados obtenidos con Impute. **d)** Unión de los valores R^2 y MAF, obtenidos a partir de Impute, en un solo archivo. **e)** Extracción de R^2 y MAF de los resultados obtenidos con Minimac.

a
```
bcftools query -H -f '%CHROM\t%POS\t%ID\t%REF\t%ALT\t%FILTER\t%DR2\t%MAF\n'
imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_MAF_beagle.vcf.gz >
imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_R2_MAF_beagle.r2
```

b
```
awk '{ print $3,  $7 }' imputed_Eggplant_MAGIC_2023_S5_impute.impute2_info >
imputed_Eggplant_MAGIC_2023_S5_impute_info.txt
```

c
```
bcftools query -H -f '%POS\t%MAF\n'
imputed_Eggplant_MAGIC_2023_S5_MAF_impute.vcf >
imputed_Eggplant_MAGIC_2023_MAF_impute.txt
```

d
```
echo 'BEGIN { FS = OFS = " " } FNR == NR { a[$1] = $2; next } $1 in a { print
$1, a[$1], $2 }' > compare.awk

awk -f compare.awk imputed_Eggplant_MAGIC_2023_S5_impute_info.txt
imputed_Eggplant_MAGIC_2023_MAF_impute.txt >
imputed_Eggplant_MAGIC_2023_S5_R2_MAF_impute.txt
```

e
```
bcftools query -H -f '%CHROM\t%POS\t%ID\t%REF\t%ALT\t%FILTER\t%R2\t%MAF\n'
imputed_Eggplant_MAGIC_2023_S5_minimac.vcf >
imputed_Eggplant_MAGIC_2023_S5_minimac.r2
```

Con el paquete ggplot2 de R (Versión 3.5.0., Wickham, 2009), se representaron los resultados de los tres procesos de imputación (Anexo 6). Primero, se comparó el valor R^2 medio entre los tres programas mediante boxplots. Segundo, se determinó la distribución de los SNPs imputados con cierto nivel de confianza mediante histogramas. Y, por último, usando un scatterplot, se comparó la relación entre el MAF y el R^2 para cada marcador.

Para el análisis estadístico se empleó el software R (Versión 4.3.2., R Core Team, 2023). Primero, se realizó un ANOVA unidireccional para comparar el valor medio de R^2 para cada uno de los tres programas. Posteriormente, se revisó la homogeneidad de las varianzas mediante la prueba de Bartlett. Finalmente, los resultados para cada programa se separaron mediante la prueba post hoc de Games-Howell.

4. RESULTADOS

4.1. Calidad del panel de referencia

El panel de referencia resultante se generó a partir de los ocho parentales fundadores de la primera población MAGIC de berenjena secuenciados a una profundidad de 20X (Gramazio et al., 2019). Si bien es posible realizar la imputación sin emplear un panel de referencia, el uso de esta información permitirá obtener resultados más precisos. En trabajos previos, tras el mapeo de las lecturas limpias contra el genoma de referencia v3.0 “67/3” de berenjena, el SNP calling y la selección de SNPs bialélicos y polimórficos soportados por más de 20 lecturas, se retuvo un total de 17.069.371 polimorfismos. Cada parental presentó un número distinto de polimorfismos (Tabla 2). Una vez eliminadas las posiciones duplicadas, se obtuvo un total de 15.790.792 SNPs, equivalente al 92,51% del conjunto original. Los marcadores se encontraron repartidos en los doce cromosomas, como se muestra en la Figura 8.

Tabla 2. Número total de polimorfismos por parental

Parental	Número de polimorfismos
MM1597 **(A)**	2.596.066
DH-ECAVI **(B)**	2.545.239
MM577 **(C)**	11.087.421
AN-S-26 **(D)**	2.548.723
H15 **(E)**	2.432.997
A0416 **(F)**	2.639.242
IVIA-371 **(G)**	2.636.417
ASI-S-1 **(H)**	1.387.747

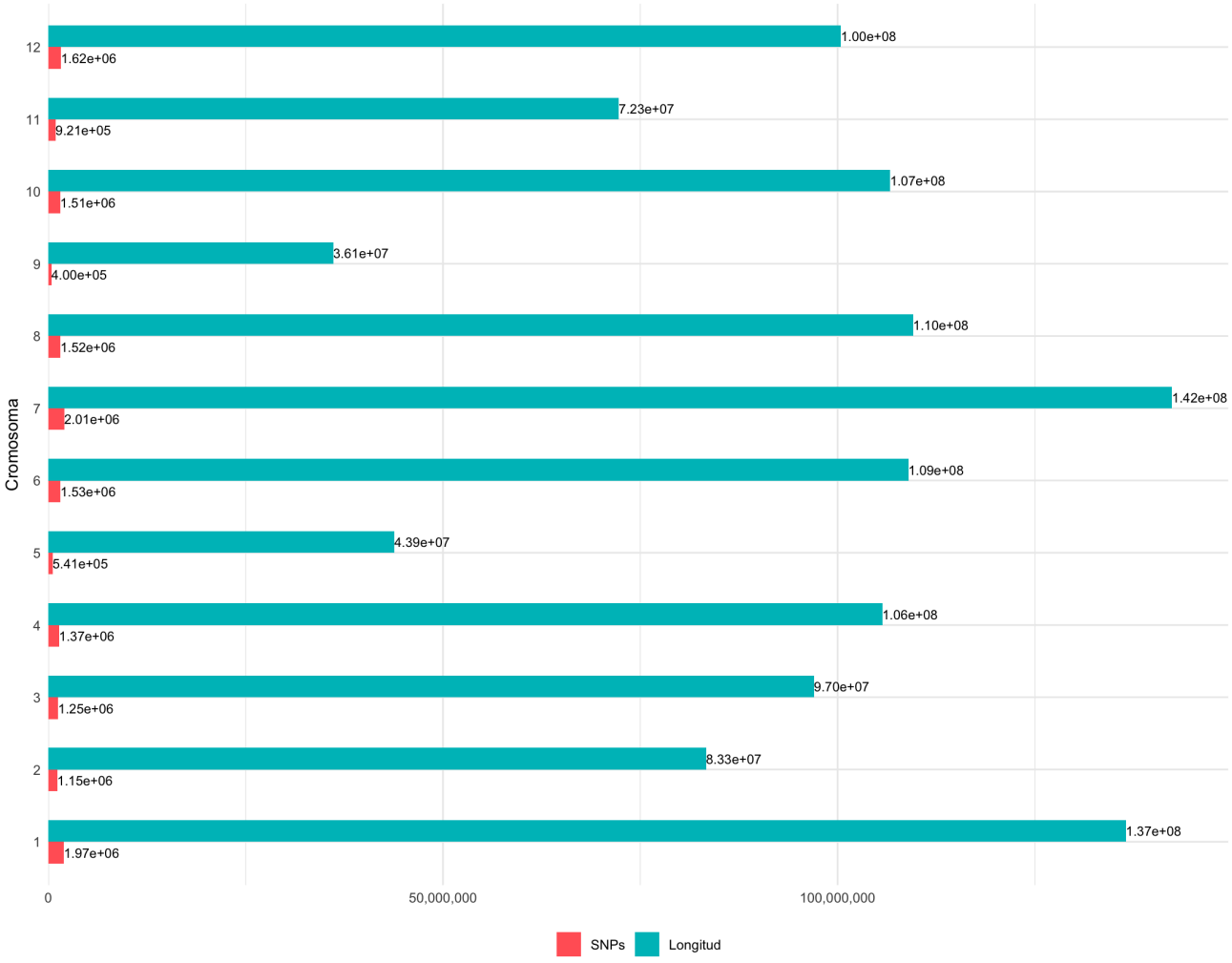


Figura 8. Distribución de SNPs por cromosoma con respecto a la longitud total en mega pares de bases (Mb).

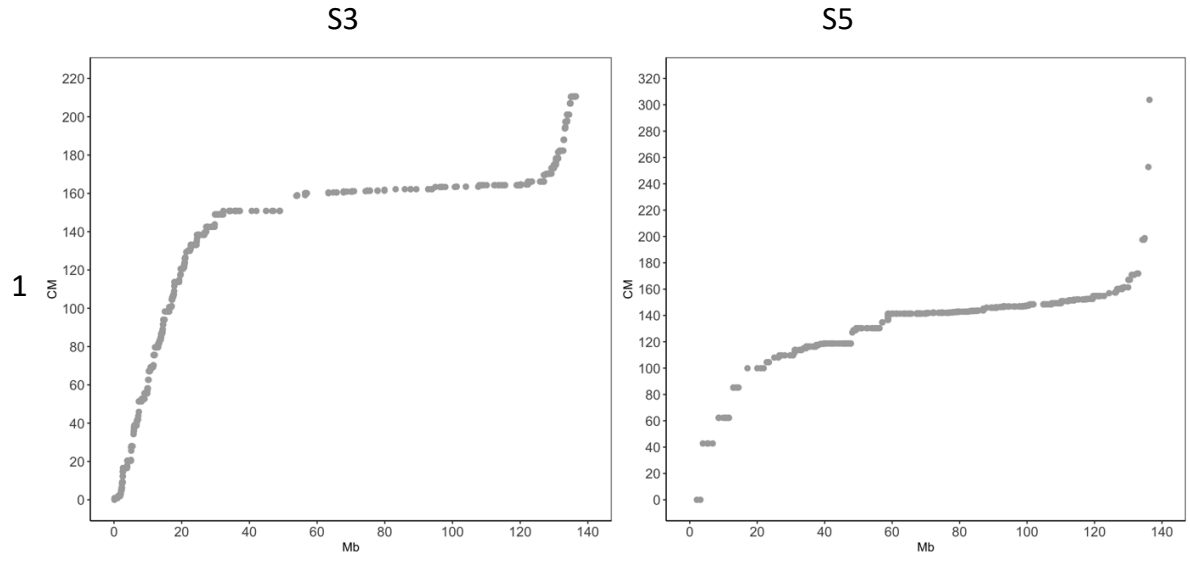
4.2. Resumen del mapa genético

Se construyeron dos mapas genéticos, el primero con la información de la generación S3 de la población MAGIC y el segundo con la generación S5 de la misma población. En el caso de la población S3, se eliminaron los SNPs multialélicos, los cuales ya fueron previamente eliminados para los datos de la población S5. Y se compararon con el panel de referencia con la finalidad de mantener únicamente los marcadores presentes tanto en las generaciones, S3 o S5, como en el panel de referencia. Una vez realizado esto, para la generación S3 se obtuvieron como resultado 6.070 marcadores distribuidos en los 12 cromosomas, mientras que para la generación S5 se obtuvieron 3.375 marcadores (Tabla 3).

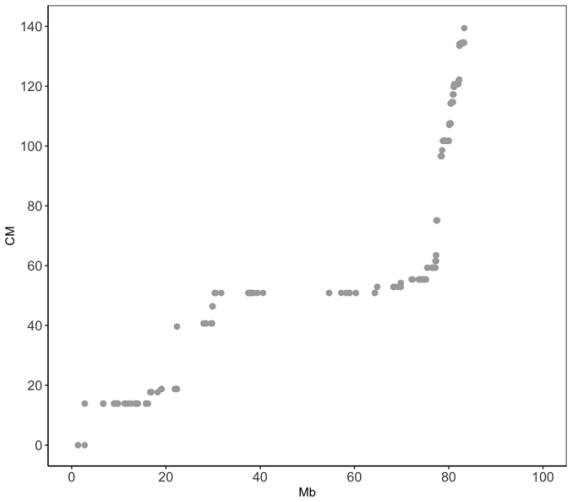
Tabla 3. Distribución de SNPs por cromosoma para los datos de la generación S3 y S5 de la población MAGIC.

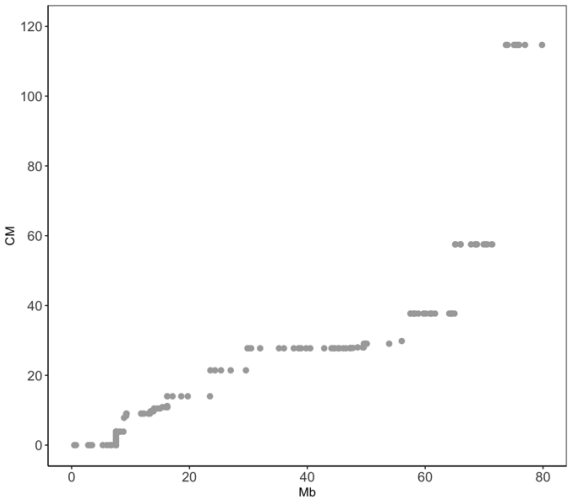
Cromosoma	SNPs S3	SNPs S5
1	999	1.026
2	323	200
3	633	96
4	481	210
5	455	84
6	627	434
7	208	262
8	427	96
9	492	114
10	634	83
11	333	452
12	458	318
TOTAL	6.070	3.375

La comparativa entre ambos resultados se realizó también de manera gráfica, representando las distancias genéticas en función de las distancias físicas (Figura 9). Una separación de un centimorgan (cM) indica una probabilidad del uno por ciento de que dos genes se separen debido a una recombinación (Stapley et al., 2017).

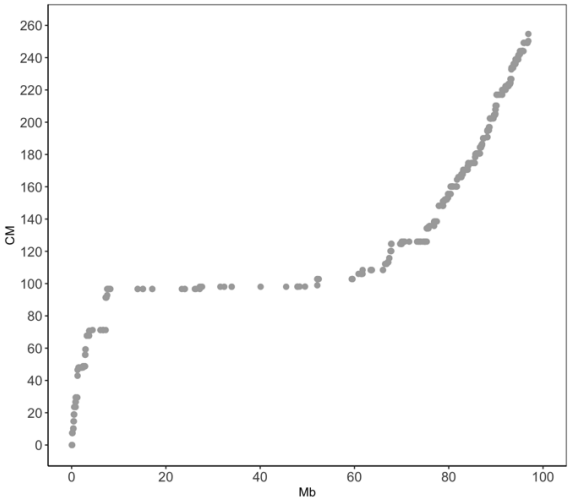


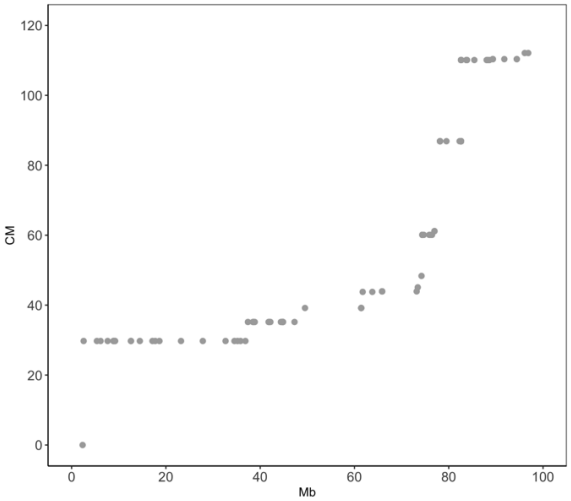
2


CM
Mb
0
20
40
60
80
100
120
140

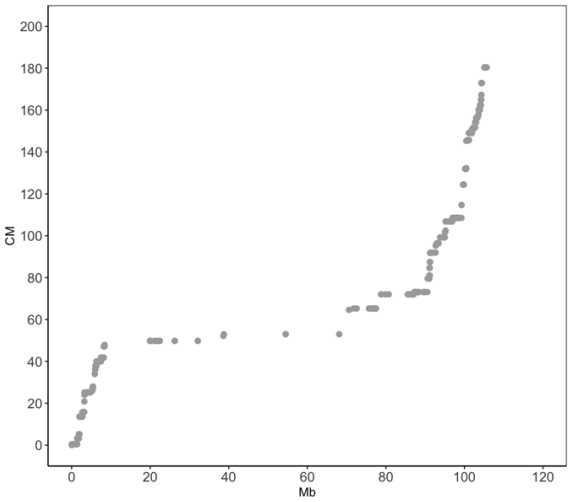

CM
Mb
0
20
40
60
80
100
120

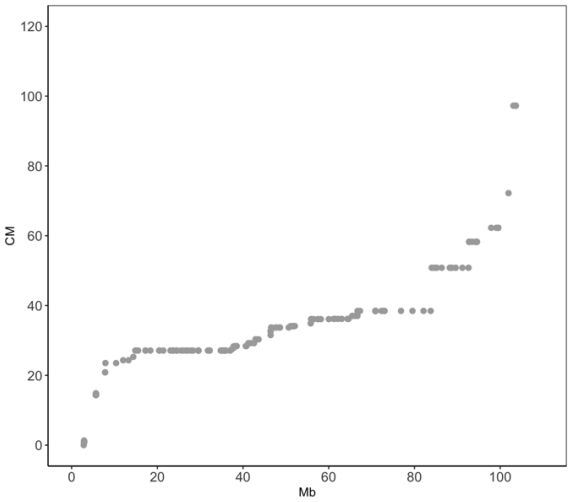
3


CM
Mb
0
20
40
60
80
100
120
140
160
180
200
220
240
260

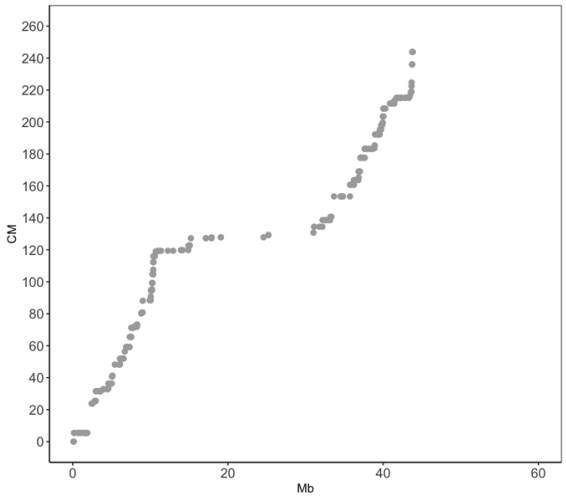

CM
Mb
0
20
40
60
80
100
120

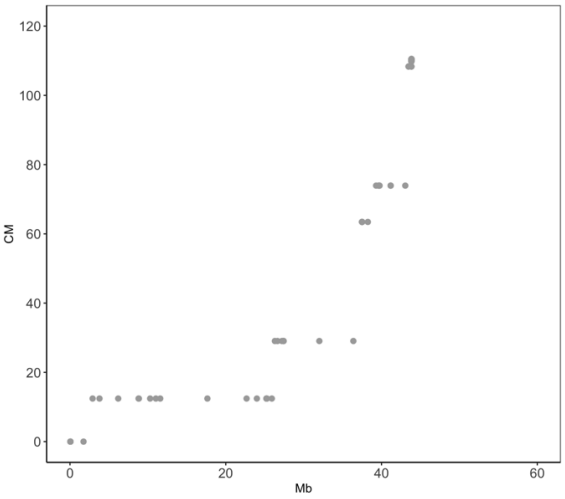
4


CM
Mb
0
20
40
60
80
100
120
140
160
180
200

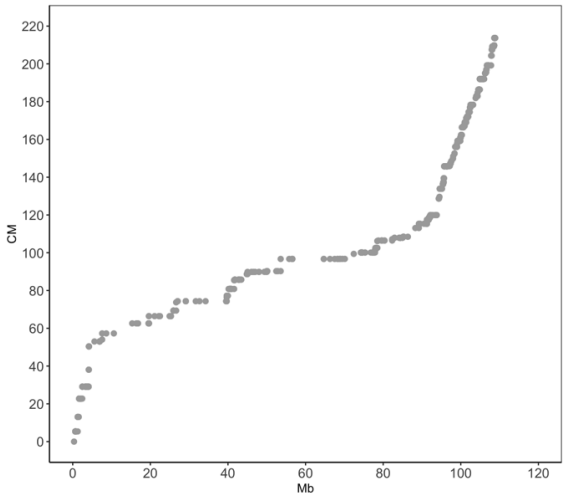

CM
Mb
0
20
40
60
80
100
120

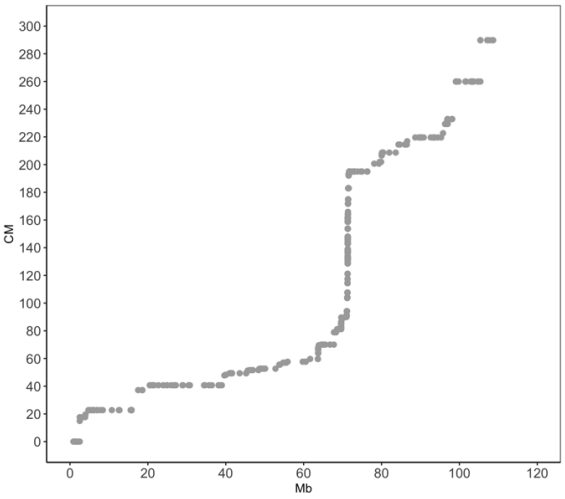
5


cM
Mb
0
20
40
60
0
20
40
60
80
100
120
140
160
180
200
220
240
260

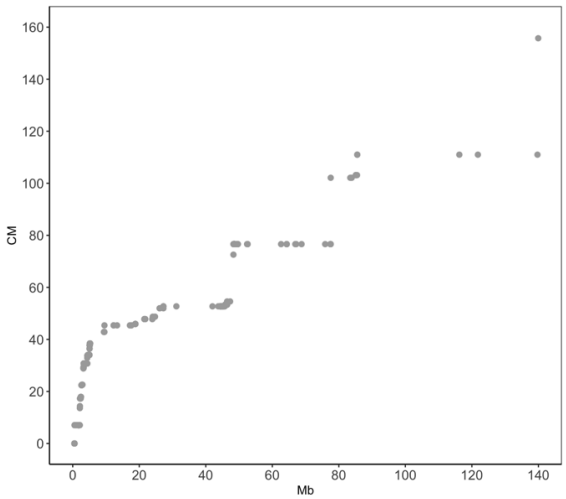

cM
Mb
0
20
40
60
0
20
40
60
80
100
120

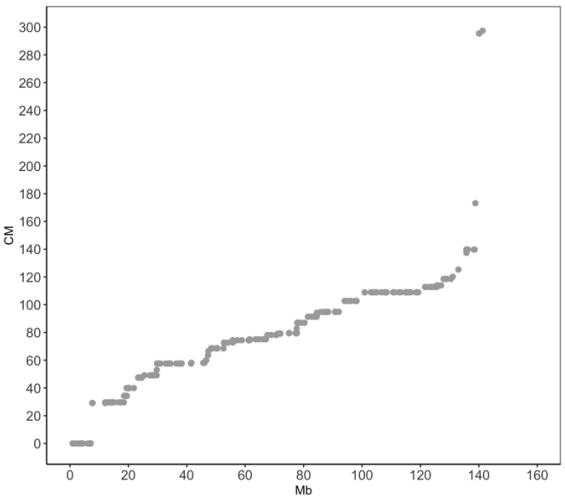
6


cM
Mb
0
20
40
60
80
100
120
0
20
40
60
80
100
120
140
160
180
200
220

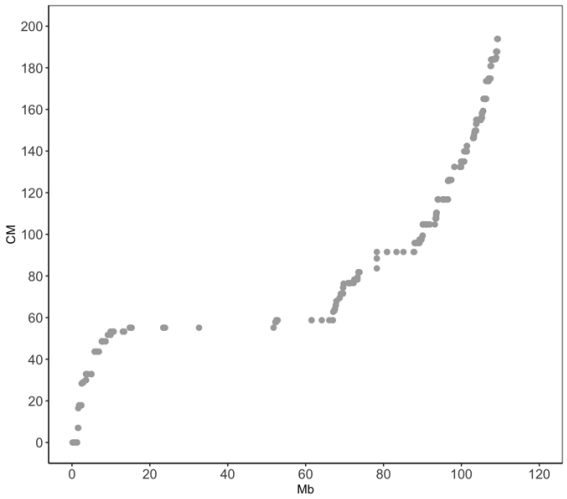

cM
Mb
0
20
40
60
80
100
120
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300

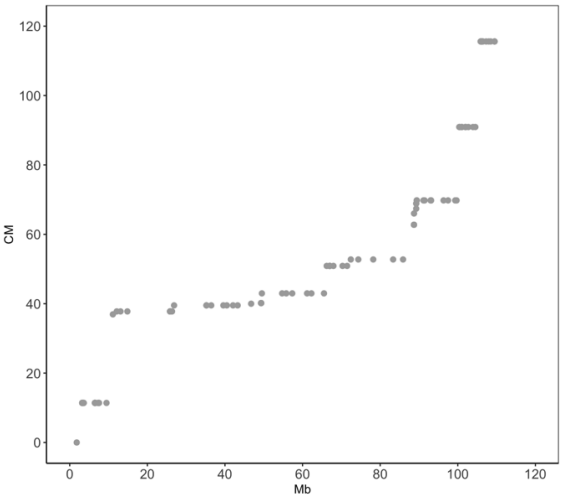
7


cM
Mb
0
20
40
60
80
100
120
140
0
20
40
60
80
100
120
140
160

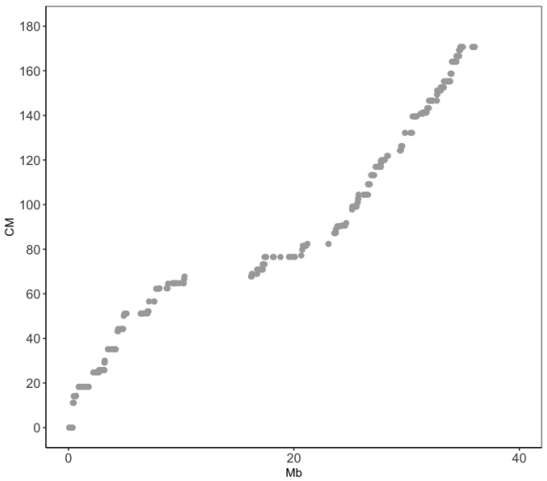

cM
Mb
0
20
40
60
80
100
120
140
160
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300

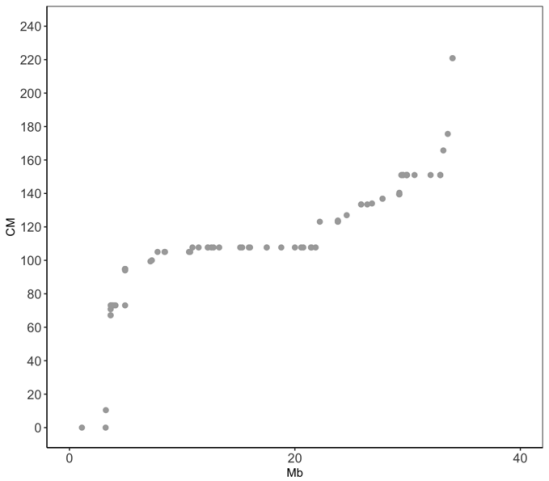
8


CM
Mb
0
20
40
60
80
100
120
140
160
180
200

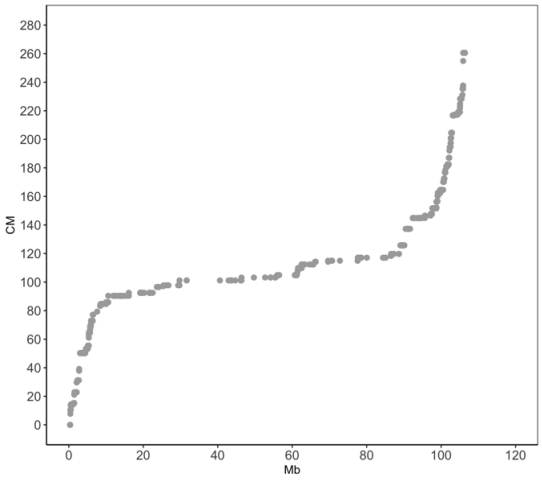

CM
Mb
0
20
40
60
80
100
120

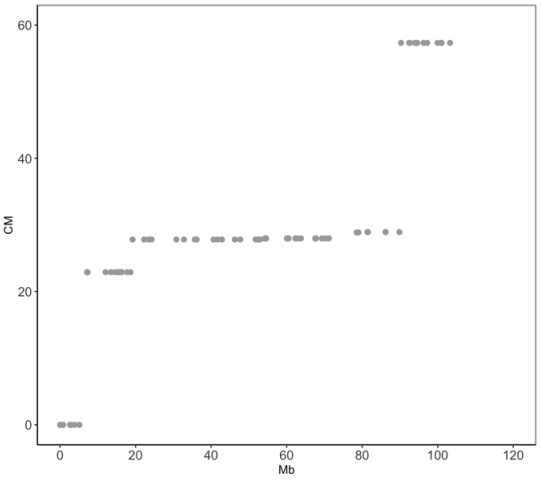
9


CM
Mb
0
20
40
60
80
100
120
140
160
180


CM
Mb
0
20
40
60
80
100
120
140
160
180
200
220
240

10


CM
Mb
0
20
40
60
80
100
120
140
160
180
200
220
240
260
280


CM
Mb
0
20
40
60

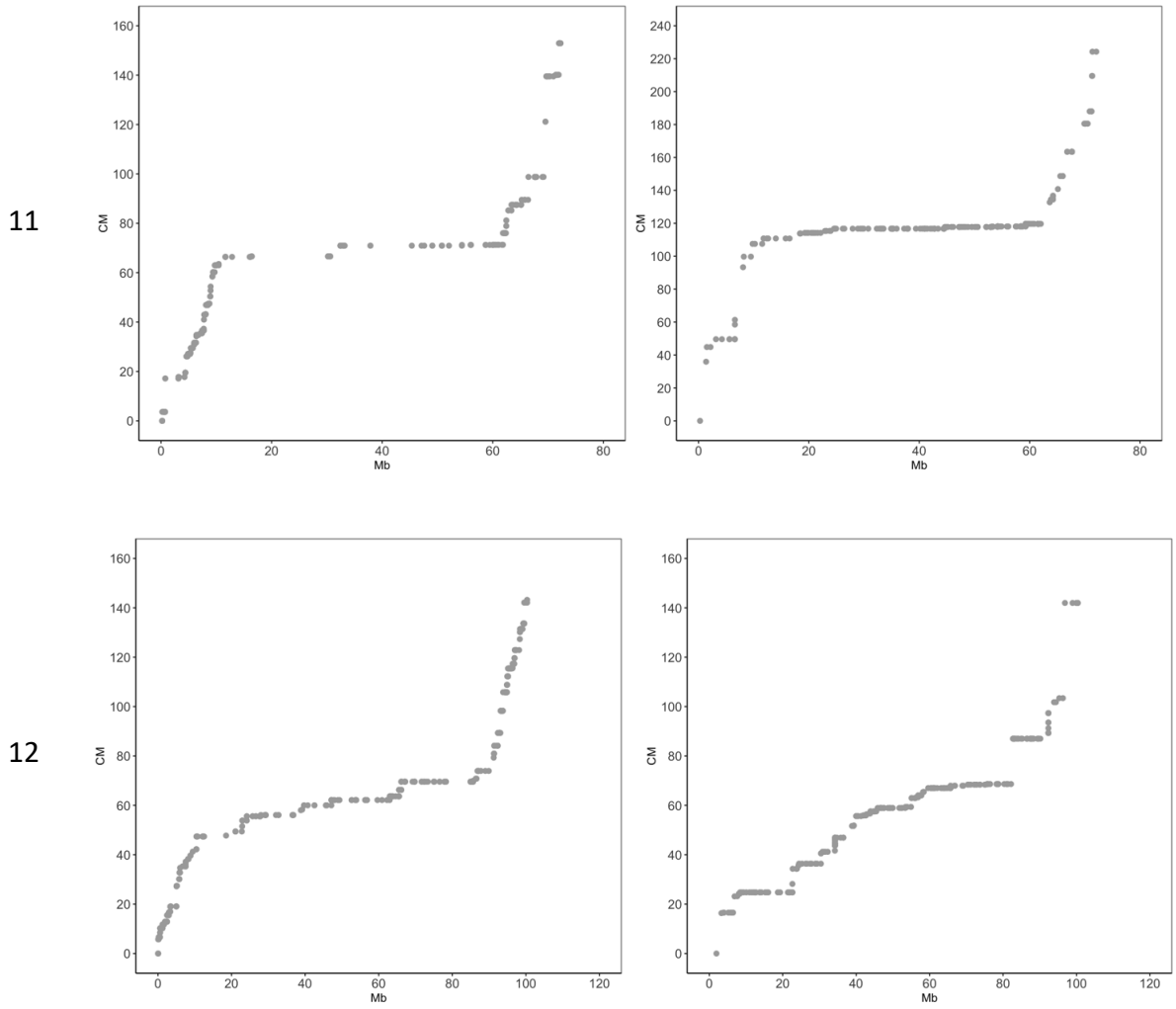


Figura 9. Mapa genético para los 12 cromosomas.
Eje X corresponde a la posición física del marcador en mega pares de bases (Mb) y el eje Y es la posición genética en centimorgans (cM).

Como se observa, en los datos de la población S3, la densidad de marcadores, al igual que la distribución a lo largo del cromosoma, es mayor que en el caso de los datos de la generación S5. Además, en cuanto a cobertura, los datos de la población S3 logran ocupar en promedio un 99,46% de la extensión de cada cromosoma. La menor cobertura se alcanza en el cromosoma 7, siendo de un 97,95%. En cambio, los datos de la población S5 se distribuyen, de media, en un 97,47% de la extensión de los cromosomas, ocupando la menor extensión en el cromosoma 9, con un 91,10% (Tabla 4). En cuanto a la distancia media entre marcadores, se observa que es mayor para los datos de la población S5, lo cual refleja la menor saturación en ciertas regiones. Es por ello que el mapa final que se empleó en el proceso de imputación fue el realizado a partir de los datos de la generación S3.

Tabla 4. Comparación de la cobertura de SNPs por cromosoma entre la población S3 y S5.

Cromosoma	Longitud (Mb)	Posición mínima S3 (Mb)	Posición mínima S5 (Mb)	Posición máxima S3 (Mb)	Posición máxima S5 (Mb)	% cobertura S3	% cobertura S5	Distancia media S3 (cM)	Distancia media S5 (cM)
1	136,53	0,13	2,12	136,44	136,28	99,83	98,27	0,21	0,30
2	83,34	1,36	0,46	83,30	80,63	98,32	96,20	0,43	0,58
3	97,01	0,04	2,33	96,88	96,86	99,82	97,44	0,40	1,18
4	105,67	0,01	2,82	105,63	105,22	99,96	96,90	0,38	0,63
5	43,85	0,09	0,05	43,81	43,84	99,70	99,86	0,54	1,33
6	108,97	0,34	0,88	108,86	108,67	99,59	98,92	0,34	0,67
7	142,38	0,48	0,88	139,95	141,28	97,95	98,61	0,75	1,14
8	109,58	0,14	1,80	109,36	109,48	99,67	98,27	0,46	1,22
9	36,10	0,07	1,11	36,05	33,99	99,67	91,10	0,35	1,95
10	106,64	0,31	0,05	106,37	103,27	99,45	96,79	0,41	0,70
11	72,29	0,21	0,28	72,25	72,08	99,66	99,32	0,46	0,50
12	100,42	0,04	1,95	100,34	100,35	99,88	97,99	0,31	0,45

4.3. Imputación de datos

Como se mencionó anteriormente, los mismos datos se imputaron usando tres programas diferentes. A pesar de estar basados en el mismo modelo computacional, cada programa requirió un diferente formato para los datos y tuvo un desempeño computacional diferente (Tabla 5).

Tabla 5. Resumen de los parámetros de imputación de cada programa

Programa	Modelo computacional	Formato de entrada	Formato de salida	Tiempo de ejecución (por cromosoma)	Exigencia de memoria (por cromosoma)
Beagle 5.1	HMM	VCF, MAP	VCF	8 minutos	100 GB
Impute 2	HMM	HAPS, GENS, LEGEND, MAP	IMPUTE2	17 horas	1 GB
Minimac 4	HMM	VCF, MSAV	VCF	15 minutos	1 GB

4.3.1. Beagle 5.1

Tras la imputación con Beagle, los genotipos originales junto con los imputados se almacenaron en un archivo VCF, en el que también se recoge información sobre la calidad en forma de valores DR^2 y sobre la frecuencia del alelo alternativo. Para llevar a cabo la evaluación de la imputación, se extrajo tanto el valor R^2 como el valor MAF, calculado previamente con BCFtools, y se graficó la correlación entre ambos valores. Se realizó un gráfico para la totalidad de los datos, un gráfico únicamente con las posiciones imputadas de novo y otro gráfico únicamente con las posiciones originales con los datos faltantes imputados. Como se observa en la Figura 10, no existe una correlación entre los valores de R^2 y MAF en ninguno de los tres casos. Tanto las variantes con una imputación más confiable (R^2 alto) como las variantes imputadas con baja confianza (R^2 bajo) tienen valores de MAF desde 0 a 0,5. En el caso de las

posiciones originales en las que se ha imputado los genotipos faltantes, su valor de R^2 es igual a uno (Figura 10.c) Beagle solo otorga diferentes valores de R^2, desde cero hasta uno, si los loci imputados no se encuentran en el archivo de entrada, sino que los genotipos son imputados para todos los individuos.

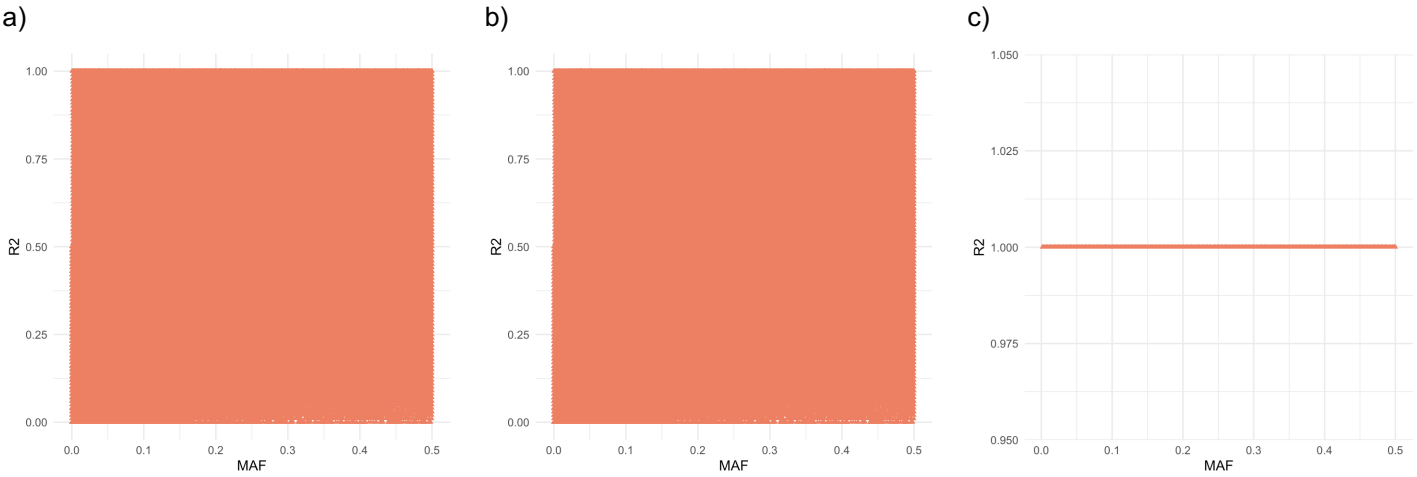


Figura 10. Relación entre MAF y R^2 por marcador. **a)** Datos generales. **b)** Datos imputados de novo. **c)** Datos originales imputados.

4.3.2. Impute 2

Tras la obtención de los cinco archivos al imputar con Impute 2, se calculó el valor de MAF del archivo .impute2_haps y se utilizó el valor de InfoScore (R^2) del archivo .impute2_info para evaluar la calidad del proceso. Como se mencionó anteriormente, el programa no provee la información de los haplotipos de las posiciones imputadas de novo, por lo que sólo se pudo calcular el MAF de las posiciones originales mediante BCFtools. Al igual que con Beagle, no existe una correlación entre los valores de InfoScore (equivalente a R^2) y MAF (Figura 11).

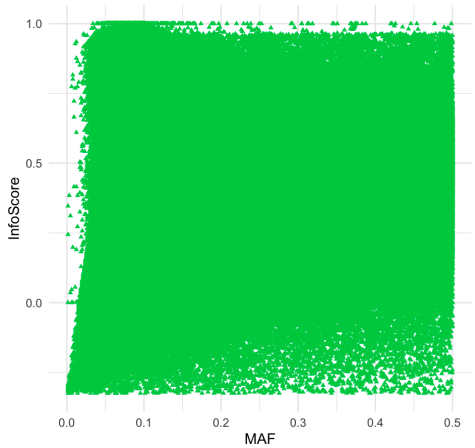


Figura 11. Relación entre MAF y R^2 por marcador en los datos originales imputados

4.3.3. Minimac 4

Como resultado de la imputación con Minimac, se obtuvo un archivo VCF con los genotipos imputados junto con el campo R^2, el campo MAF, y el campo IMPUTED. Para evaluar la calidad de la imputación, se evaluó la correlación de R^2 y MAF. Al igual que con Beagle, se realizó un gráfico para la totalidad de los datos, un gráfico con las posiciones imputadas de novo y otro gráfico únicamente con las posiciones originales imputadas. En los tres casos existe una correlación entre los valores de R^2 y MAF (Figura 12). Las variantes con una imputación más confiable presentaron valores de MAF mayores. Los loci originales imputados presentaron valores mayores de R^2 a medida que mayor era su MAF en comparación con los genotipos imputados de novo.

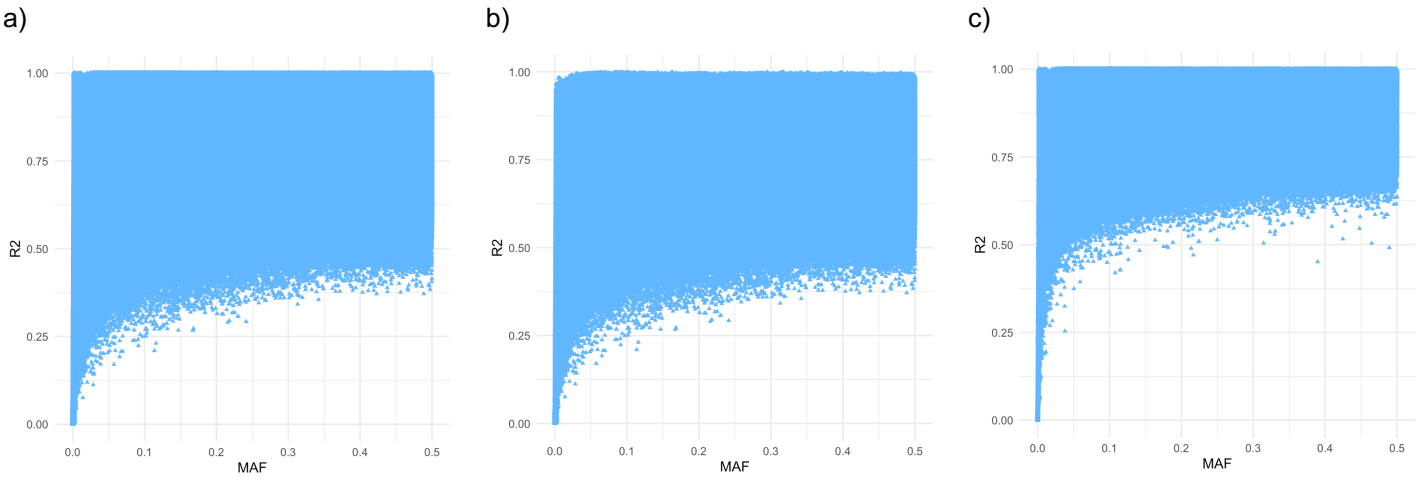


Figura 12. Relación entre MAF y R^2 por marcador. **a)** Datos generales. **b)** Datos imputados de novo. **c)** Datos originales imputados.

4.4. Evaluación del desempeño de la imputación

Para la comparación de los tres softwares se evaluó la precisión de la imputación usando R^2. Tanto para los datos generales como para los datos imputados de novo, Minimac presenta datos de imputación más confiables. En cuanto a Beagle, el R^2 promedio es mayor que Impute cuando se analizan los datos completos, a diferencia de lo que ocurre cuando se analizan únicamente los datos imputados de novo (Figura 13). Sin embargo, Impute no imputa realmente las posiciones de novo, únicamente rellena las posiciones originales que tenían missing data por lo que este resultado no debería ser tomado en cuenta. Para determinar si las medias son estadísticamente

iguales, ya que se encuentran próximas entre sí, se realizó un análisis de varianza (Tabla 6 y 7).

Tabla 6. ANOVA del valor de R^2 de los datos totales para cada programa de imputación

Fuente	Grados de libertad	Suma de cuadrados	Media cuadrática	Valor F	Valor P
Programa	2	434.798	217.399	3.516.739	<2e-16
Residual	47.372.862	2.928.509	0		

Tabla 7. ANOVA del valor de R^2 de los datos imputados de novo para cada programa de imputación

Fuente	Grados de libertad	Suma de cuadrados	Media cuadrática	Valor F	Valor P
Programa	2	282262	141131	2425674	<2e-16
Residual	38754738	2254836	0		

Según el ANOVA unidireccional, se detectó diferencias significativas entre el valor R^2 medio de los tres programas de imputación, tanto para los datos completos como para los datos imputados de novo ($F(2;47372862)= 3516739$, $p<0,001$; $F(2; 38754738)= 2425674$, $p<0,001$). Posteriormente, con la prueba de Bartlett se revisó la homogeneidad de las varianzas ($K= 12158724$, $p<0,001$; $K= 11838216$, $p<0,001$). En ambos casos, las medias son significativamente diferentes entre sí.

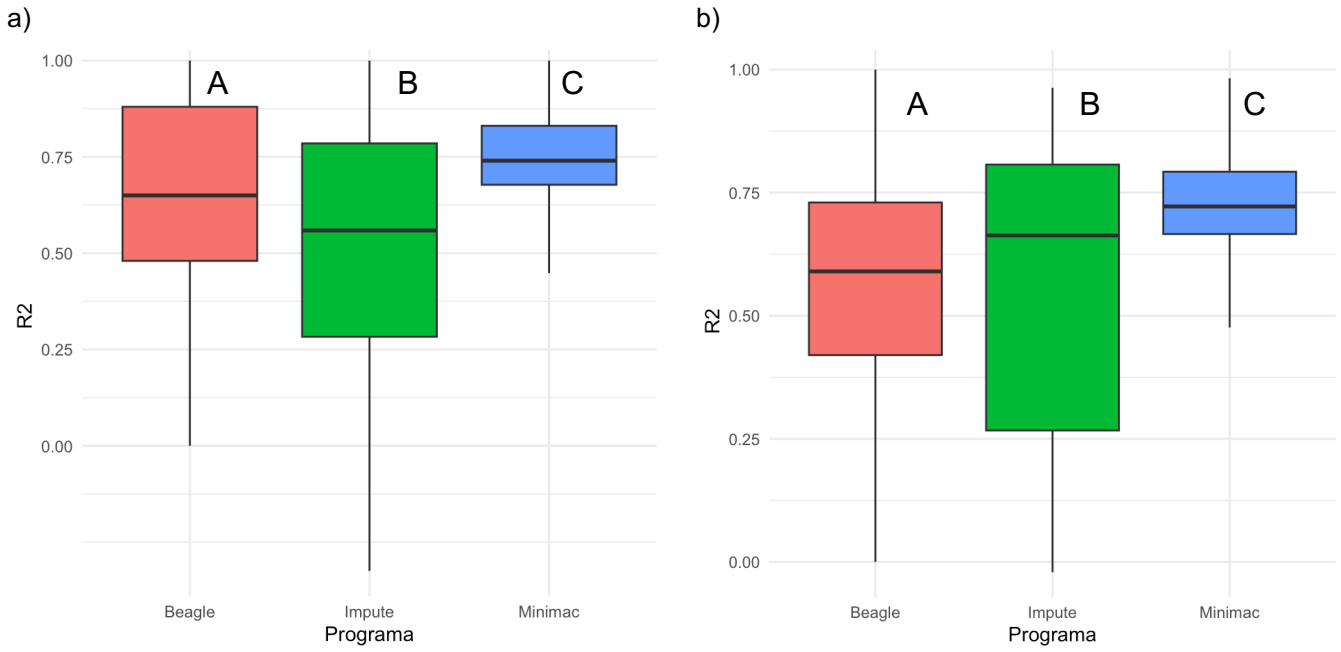


Figura 13. R^2 promedio por programa. **a)** Datos generales. **b)** Datos imputados de novo.
Se presentan las medias y la desviación estándar. Letras diferentes corresponden a grupos significativamente diferentes (Games-Howell, $p<0,001$).

Por otro lado, se evalúo la eficiencia de los marcadores seleccionados para la imputación. En la Figura 14 se observa la distribución de los marcadores en función de su valor de R^2 para cada programa. Tanto Beagle como Impute presentan una distribución equitativa de la confianza de imputación de los marcadores. Sin embargo, Minimac presenta un nivel de confianza superior al 70% para más del 50% de los marcadores en ambos casos (Figura 14). Además, de los tres programas, Minimac es el único que casi no presenta loci imputados con una confianza menor a 10%, mientras que Impute es el programa que presenta la menor cantidad de posiciones imputadas con un R^2 de 1.

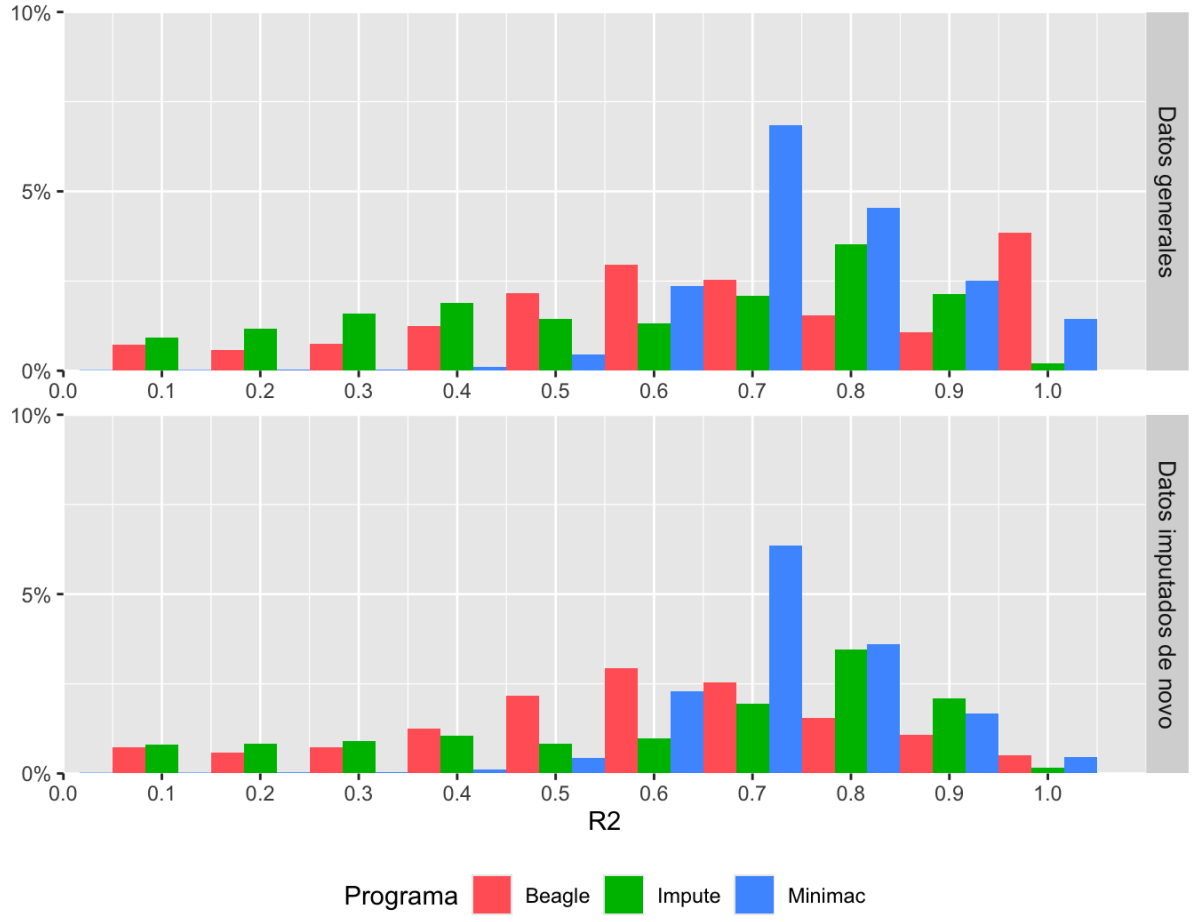


Figura 14. Distribución de los marcadores según su valor de R^2 por programa.
Eje X corresponde al porcentaje de SNPs del total con determinado valor de R^2.

Para Beagle, ambos sets de datos tienen un R^2 mínimo de 0 y máximo de 1. El valor medio es de 0,64 para los datos generales y de 0,56 para los datos imputados. Para Impute, en los datos generales el valor medio es de 0,52, el valor mínimo es 0 y el valor máximo es 1, mientras que para los datos imputados el valor medio es de 0,54, el valor mínimo es 0 y el valor máximo es 0,96. Para Minimac, los datos generales

tienen un R^2 mínimo de 0, máximo de 1 y medio de 0,75; mientras que los datos imputados tienen un R^2 mínimo de 0,001, máximo de 0,99 y medio de 0,73.

5. DISCUSIÓN

En este trabajo se comparó el desempeño de tres programas de imputación con el objetivo de mejorar la resolución del genotipado por secuenciación a muy baja cobertura. Primero, se elaboró el panel de referencia a partir los ocho parentales fundadores de la población MAGIC y un mapa genético a partir de los datos de la población S3, ambos secuenciados a una profundidad de 20X. Con esta información, se imputó los genotipos faltantes de la población S5 secuenciada a baja cobertura (3X) empleando tres programas: Beagle 5.1, Impute 2 y Minimac 4. Estos programas fueron elegidos ya que hoy en día son los programas de licencia libre más utilizados para la imputación genética y se han empleado previamente en otras investigaciones similares (Liu et al., 2015; Korkuć et al., 2019; De Marino et al., 2022).

5.1. Calidad del panel de referencia

El uso de un panel de referencia durante el proceso de imputación ayuda a mejorar la precisión de los genotipos imputados (Sengupta et al., 2023). Se conoce que tanto la diversidad como el tamaño del panel de referencia afectan la precisión de la imputación genética. De hecho, en muchos casos, el aumentar la diversidad del panel conlleva un aumento en su tamaño (Bai et al., 2020). La calidad de la imputación es directamente proporcional al tamaño del panel de referencia, por lo que, mientras mayor sea la cantidad de haplotipos en el panel de referencia, mayor será la precisión de la imputación (Bai et al., 2020). En cuanto a la diversidad, hay estudios en los que se ha observado que el uso de paneles de referencia mixtos, desarrollados a partir de diferentes poblaciones, resulta en una elevada precisión de imputación (L. Huang et al., 2009; Chou et al., 2016; Xu et al., 2023; French et al., 2024). Además del tamaño y la diversidad del panel de referencia, la cobertura de secuenciación también afecta la calidad de la imputación. Usar datos de secuenciación de alta cobertura genera un panel de referencia con genotipos más precisos y, por ende, una inferencia de haplotipos más precisa (O'Connell et al., 2021).

En este estudio, se empleó un panel de referencia que cumple con los factores anteriormente mencionados para que la imputación sea de alta calidad (Tabla 2). El tener un amplio número de parentales fundadores de una población otorga una gran diversidad alélica (Yuan et al., 2023). En este caso, el panel de referencia se

encuentra conformado por los ocho parentales de la población MAGIC de berenjena, incluyendo una especie silvestre (parental C). Este parental es el que otorga la mayor parte de la diversidad genética, lo cual se evidencia en el alto número de polimorfismos identificados en esta accesión (Tabla 2). Por otro lado, como el objetivo fue imputar los genotipos faltantes de individuos de la población MAGIC, el uso de un panel de referencia conformado únicamente por los parentales de esta población es suficiente ya que, en teoría, no se deberían encontrar alelos o haplotipos diferentes a los que presentan los parentales. En segundo lugar, el panel de referencia está formado por más de 17 millones de polimorfismos (Tabla 2), tratándose de un panel de gran tamaño. Por ejemplo, uno de los paneles de referencia más comúnmente usados para análisis de imputación genética es del genoma humano, creado en el proyecto HapMap y formado por un total de tres millones de polimorfismos (Sabeti et al., 2007). En tercer lugar, los datos del panel de referencia provienen de una resecuenciación a 20X; es decir, de media cada posición se encuentra soportada por un elevado número de lecturas. En el estudio llevado a cabo por Luo et al., (2021), se ha observado que se requiere que la profundidad de secuenciación para elaborar el panel de referencia no sea menor a 20X para que sea considerado de alta calidad.

La mayoría de los programas de imputación existentes se basan en la idea de que los individuos que provienen del mismo ancestro o de uno similar podrían compartir pequeñas secciones de ADN entre ellos (Stahl et al., 2021). Por lo tanto, mediante las recombinaciones de haplotipos que se presentan en el panel de referencia, es posible predecir los genotipos de variantes no observadas (Chi Duong et al., 2023). Sin embargo, generar un panel de referencia es costoso, ya que es necesario contar con datos secuenciados a altas profundidades. Es por esto que durante los últimos años se han desarrollado programas que permitan la imputación genética prescindiendo del panel de referencia (Davies et al., 2016; Chi Duong et al., 2023). Estos métodos generan la imputación basándose únicamente en los datos de la secuenciación de la población objetivo. Un ejemplo es el software STITCH, cuyo modelo de tipo HMM busca reconstruir los haplotipos de los parentales fundadores mediante cálculos de probabilidades de manera iterada y así determinar qué par de haplotipos pertenece a cada locus (Davies et al., 2016). Para evaluar el desempeño de este programa, se realizaron dos procesos de imputación con dos conjuntos de datos distintos. Se empleó una población de 2.073 ratones CFW secuenciados a baja cobertura (0,15X)

como primer conjunto de datos y el segundo conjunto de datos se obtuvo a partir de la secuenciación a baja cobertura (1,7X) de una población de 11.670 mujeres chinas pertenecientes a la etnia Han (Davies et al., 2016). Posterior a la imputación, los resultados para ambos datos se compararon con datos de secuenciación a alta cobertura para las mismas especies y se observó que la correlación obtenida entre la posición imputada y la posición real fue mayor a 0,92. Estos resultados sugieren que STITCH consigue imputar con mucha precisión incluso sin panel de referencia (Davies et al., 2016).

Otras alternativas que se han desarrollado emplean modelos de machine learning, como es el caso de las redes neuronales (Chi Duong et al., 2023). Las redes neuronales son modelos de aprendizaje automático que simulan el cerebro humano, usando procesos que imitan la forma en que las neuronas biológicas trabajan en la toma de decisiones para identificar fenómenos y llegar a conclusiones (Han et al., 2018). La Universidad de Tohoku, en Japón, propuso una técnica de imputación basada en una red neuronal recurrente bidireccional (RNN), que utiliza los genotipos faseados como datos de entrada para estimar las probabilidades de alelos para variantes no observadas (Kojima et al., 2020). Sin embargo, se ha demostrado que este modelo tarda el doble de tiempo que Impute 2 (Chi Duong et al., 2023). Con el fin de mitigar esta desventaja, el Vingroup Big Data Institute de Vietnam desarrolló el método GRUD, que consiste en una RNN capaz de aprender combinaciones no lineales y transformar los datos de entrada a una distribución normal, facilitando el proceso de aprendizaje del modelo, con el objetivo de reducir el tiempo computacional (Chi Duong et al., 2023).

5.2. Resumen del mapa genético

Como se mencionó anteriormente, para la imputación se empleó el mapa genético generado a partir de los datos de la generación S3. Para obtener un mapa genético de alta confiabilidad es importante que los marcadores empleados logren una alta cobertura de los cromosomas y que estén repartidos de manera más o menos homogénea. Uno de los requisitos que presenta mpMap2, el programa empleado para la elaboración del mapa, es la ausencia de missing data en los parentales (C. Zheng et al., 2019). Una vez eliminados estos sitios, el número de polimorfismos en los parentales disminuyó mucho. Además, esto hizo que, al volver a realizar la

comparación con los datos de ambas generaciones para seleccionar aquellos sitios presentes tanto en los parentales como en la población, el número de sitios compartidos con la generación S3 fuese mucho mayor que el compartido con la generación S5. El hecho de que el número de polimorfismos útiles de la S3 fuese mayor y que se encontrasen distribuidos a lo largo de los doce cromosomas hizo que la cobertura y la resolución del mapa genético también fueran mayores (Figura 9).

Por otro lado, en el mapa final, se observa una tendencia en la curva de formar un plateau para todos los cromosomas. Estas regiones horizontales corresponden a la región centromérica naturalmente caracterizada por poseer una baja frecuencia de recombinación (Limborg et al., 2016). Como norma general, los organismos muestran recombinación suprimida en el centrómero, lo cual se cree que se debe a la estructura condensada de la cromatina que causa que las roturas de doble cadena sean menos comunes (Stapley et al., 2017).

Los mapas genéticos son de utilidad para la imputación genética al igual que para otros estudios genéticos porque permiten analizar la arquitectura genética de rasgos complejos (Qu et al., 2021). Empleando técnicas de secuenciación de alto rendimiento se han construido gran cantidad de mapas de ligamiento genético de alta resolución que abren paso a realizar otros estudios como podría ser la identificación de QTLs (Yu et al., 2019). Los mapas pueden ser generados tanto a partir de poblaciones individuales como a partir de poblaciones múltiples proporcionando una alternativa eficaz para mejorar la cobertura del genoma y así la densidad de marcadores (Maccaferri et al., 2015). Un mapa con una alta densidad de marcadores permite el mapeo de QTLs en intervalos menores e identificar marcadores vinculados entre sí, lo cual puede ser de gran utilidad para la selección asistida por marcadores en la mejora genética (Qu et al., 2021).

Tradicionalmente, los mapas genéticos se han generado a partir de poblaciones biparentales, por lo que la mayoría de los programas para el desarrollo de mapas genéticos están diseñados para trabajar con este tipo de poblaciones, como por ejemplo JoinMAP (Stam, 1993), CarthaGène (de Givry et al., 2005), MAPMAKER (Lander et al., 1987) y R/qtl (Arends et al., 2010). Sin embargo, el creciente desarrollo de poblaciones multiparentales ha impulsado el desarrollo de programas que permiten construir mapas genéticos a partir de estas poblaciones, como por ejemplo R/happy

(Mott et al., 2000), R/mpMap (B. E. Huang & George, 2011) y GAPL (L. Zhang et al., 2019). Normalmente, estos programas requieren información de pedigree, como es el caso de mpMap (Zheng et al., 2019). El uso de poblaciones multiparentales representa una ventaja sobre las poblaciones biparentales, ya que en ellas tiene lugar una mayor cantidad de eventos de recombinación y, por ende, presentan una mayor diversidad genética. Las poblaciones multiparentales tienen una alta diversidad genética, por lo que tienen una mayor capacidad para la detección de alelos de menor frecuencia (Novakazi et al., 2020). Además, tienen la capacidad de detectar QTLs de menor efecto (Odell et al., 2022).

5.3. Imputación de datos

Existen diversos factores que pueden afectar la precisión de la imputación genética, además de la calidad del panel de referencia, como el método de imputación, la precisión del proceso de phasing, y la población de estudio empleada, entre otros (Das et al., 2018). En cuanto al método de imputación, los tres programas empleados parten del mismo modelo HMM, por lo que las diferencias observadas no se deben a este factor (Stahl et al., 2021). Desde el punto de vista computacional, Beagle fue el más exigente en términos de memoria, pero a su vez también fue el más rápido. En el extremo opuesto se encuentra Impute, que, por su parte, fue el que más tiempo empleó de los tres (Tabla 5). En otros estudios, Beagle 5.1 también fue identificado como el programa más rápido en realizar la imputación en comparación con Impute 2 y Minimac 3 (Stahl et al., 2021; De Marino et al., 2022). Impute fue considerado el menos amigable con el usuario debido a que requiere de archivos de entrada en formatos exclusivos, como son HAPS, GENS y LEGEND, y a que, en lugar de generar un único archivo de salida con toda la información, genera cinco archivos independientes en formato IMPUTE (G.-H. Huang & Tseng, 2014). Si bien Bcftools ha desarrollado comandos para convertir los archivos IMPUTE a formato VCF y viceversa, el proceso es tedioso y complicado para el usuario. Otra desventaja que presenta Impute es que no imputa realmente las posiciones de novo, ya que no presenta los haplotipos resultantes para las nuevas posiciones y únicamente rellena el missing data de las posiciones originales.

En segundo lugar, el phasing de haplotipos es esencial, puesto que es necesario saber qué alelo se encuentra en cada cromosoma para poder realizar la imputación

posteriormente. Existen varios programas disponibles para realizar el proceso de phasing, como por ejemplo Beagle, Eagle y SHAPEIT, los cuales utilizan también un modelo HMM (Stahl et al., 2021). La elección del programa depende del método de imputación que se emplee, ya que son procesos complementarios (De Marino et al., 2022). Estudios comparativos han demostrado que, por sí solo, Beagle 5.1 presenta una alta precisión y un tiempo de ejecución rápido, por lo que es considerada la mejor opción para realizar este proceso (Stahl et al., 2021).

Finalmente, el último factor a tomar en cuenta es la población en la que se quieren imputar los genotipos faltantes. La relación entre el panel de referencia y la población de estudio es esencial, ya que cuanto más emparentados se encuentren, más precisos serán los resultados (Das et al., 2018). No existe evidencia de que el proceso de imputación se vea afectado específicamente por el origen de los datos empleados, es decir por la tecnología de secuenciación seleccionada (Deng et al., 2022). Sin embargo, es importante la selección del tipo de datos según los recursos disponibles y los objetivos del estudio que se tengan. En este estudio había dos posibilidades: imputar los genotipos faltantes del genotipado por SPET de la población S3 MAGIC o de la resecuenciación a baja cobertura de la población S5 MAGIC. La plataforma SPET posibilita el genotipado de un número determinado de posiciones en todos los individuos, de forma que hay un menor número de genotipos faltantes en las posiciones interrogadas (Barchi, Acquadro, et al., 2019). Por otro lado, la resecuenciación a baja cobertura identifica genotipos en posiciones al azar, que pueden o no ser las mismas entre los individuos, por lo que el porcentaje de genotipos faltantes va a ser muy superior (Torkamaneh & Belzile, 2015). Otra diferencia entre las líneas de ambas generaciones es la proporción de genotipos heterocigotos. La generación S5 resulta de someter a la generación S3 a dos ciclos de autofecundación más, por lo que su heterocigosidad será menor. Esto es beneficioso para la imputación, ya que la presencia de genotipos heterocigotos puede introducir errores en los genotipos imputados (Buckley et al., 2022). Fueron estas razones las que llevaron a realizar la imputación sobre las líneas de la generación S5, además de que son los datos con los que se trabajará en estudios posteriores.

5.4. Evaluación del desempeño de la imputación

La evaluación de la calidad de los procesos de imputación depende de si se conocen o no los genotipos verdaderos de los genotipos imputados (Verma et al., 2014). Cuando no se conoce los genotipos verdaderos, la métrica usada de manera más común es R^2 (Browning & Browning, 2009; Van Leeuwen et al., 2015). En cambio, cuando se conocen los genotipos verdaderos, la calidad de la imputación se evalúa mediante la identificación de falsos positivos y falsos negativos; donde un falso positivo es aquella posición imputada erróneamente con un valor de R^2 alto, mientras que un falso negativo es aquella posición con un R^2 bajo, pero correctamente imputada (Deelen et al., 2014). Estudios han demostrado que los valores de R^2 presentados por los diferentes programas de imputación se encuentran altamente correlacionados, por lo que se pueden emplear para comparar la calidad de imputación realizada por cada programa (Van Leeuwen et al., 2015).

Tanto Beagle como Minimac otorgan un valor de R^2 a cada posición imputada. Impute utiliza una métrica llamada InfoScore, equivalente a R^2. Al igual que el valor de R^2, la métrica InfoScore oscila entre 0 y 1, donde los valores cercanos a 1 indican que un SNP ha sido imputado con una alta confianza (H.-F. Zheng et al., 2015). El criterio para definir qué valor de R^2 indica que el genotipo se encuentra bien imputado depende del investigador. En muchos estudios se emplea un valor R^2 igual o superior a 0,3 para definir que el marcador se encuentra bien imputado (Y. Li et al., 2010; H.-F. Zheng et al., 2015; Kreiner-Møller et al., 2015). Sin embargo, es un valor muy bajo y estadísticamente se recomienda utilizar uno superior a 0,7 (Rüeger et al., 2018).

La precisión de la imputación depende también del MAF para cada SNP. Se conoce que el valor R^2 se ve afectado de manera directa por el valor MAF, ya que mientras menor sea el MAF, la precisión de la imputación y, por ende, el R^2 será menor debido a que se trata de variantes raras que son difíciles de predecir (Marchini & Howie, 2010; Nelson et al., 2013; Shi et al., 2018; Z. Zhang et al., 2021). En los resultados de la imputación realizada tanto con Beagle (Figura 10) como con Impute (Figura 11) se observa que la precisión de los datos imputados no se ve influenciada por el MAF. Sin embargo, cuando las variantes son menos frecuentes, la precisión de imputación es menor con Minimac (Figura 12). Estudios de imputación realizados con Minimac e Impute han mostrado que la precisión de la imputación es menor en regiones de baja

frecuencia de alelos menores (Gao et al., 2012; H.-F. Zheng et al., 2012). Principalmente, se ha observado que la precisión disminuye significativamente cuando el valor de MAF es menor a 0,05 (Stahl et al., 2021). Se cree que el bajo desempeño de los programas de imputación en los SNPs con bajo MAF se debe a un sesgo propio del programa con el panel de referencia, ya que, si la información entre individuos del panel se encuentra en discordancia para un determinado SNP, el programa tiende a elegir un genotipo al azar para todos los individuos con dicho SNP (Shi et al., 2018). Por esta razón, una buena solución es aumentar el tamaño del panel de referencia y así disminuir el sesgo que podría existir y a su vez aumentar la precisión de la imputación (Y. Li et al., 2011). Sin embargo, se recomienda que los SNP con MAF bajo deben secuenciarse y no imputarse si son de particular interés en el estudio (Stahl et al., 2021). Por otro lado, también es importante resaltar que los valores de MAF tanto para Beagle como para Impute fueron calculados mediante Bcftools, mientras que los valores de MAF de Minimac fueron proporcionados por el mismo programa. El hecho de que los resultados de Beagle e Impute no presentan la relación esperada entre el MAF y el R^2 se puede deber a que quizás el método empleado por Bcftools no calcula este valor de la misma manera que Minimac.

En este estudio, al no disponer de información acerca de los genotipos faltantes, se empleó la métrica R^2 para evaluar la calidad de la imputación realizada por cada uno de los tres programas. Se observó que el valor de R^2 promedio de Beagle es alto, a pesar de tener muchas posiciones con valores inferiores a 0,2 (Figura 13). Una de las posibles razones es que las posiciones ya existentes en el archivo de entrada reciben el valor de 1. Como se mencionó anteriormente, el valor de R2 representa la relación existente entre el genotipo estimado y el genotipo verdadero. En este caso, el genotipo verdadero y el genotipo estimado son el mismo, por lo que la correlación es igual a 1, aumentando así el valor promedio de R^2 (Browning & Browning, 2009; Marchini & Howie, 2010). En Impute y Minimac esto no ocurre, sino que el valor de R^2 para las posiciones ya existentes en el archivo de entrada se recalcula mediante la estimación de un valor R^2 alélico que corresponde a la correlación entre la dosis del alelo imputada y la verdadera para un marcador (Browning & Browning, 2009).

En términos generales, la precisión de la imputación realizada con Minimac fue superior a la imputación realizada tanto con Impute como con Beagle. Con Minimac

se obtuvo un valor medio de R^2 superior al obtenido con los otros dos programas, una distribución de los valores de R^2 menos dispersa, y se alcanzaron valores mínimos aceptables. Impute presentó la distribución más dispersa entre los tres programas, lo cual puede indicar que su proceso de imputación no es el más confiable de todos. Cabe resaltar que tanto Beagle como Impute generaron posiciones donde la confianza de imputación fue nula, lo que obliga a llevar a cabo un paso para eliminar estos genotipos imputados.

Una posible razón que explique el mal desempeño de Impute en este estudio es el tamaño del panel de referencia. Si bien Impute se considera un software que imputa genotipos con una alta confianza, se ha observado que es más preciso cuando trabaja con paneles de referencia medianos o un subset de un panel de referencia grande (Das et al., 2018). En contraposición, Beagle alcanza mejores resultados cuando utiliza paneles de referencia mayores (Browning & Browning, 2009). Por su parte, en otros estudios se ha observado que Minimac es más preciso que Impute y Beagle cuando se trabaja con un panel de referencia pequeño (Ghoreishifar et al., 2018). La definición de si un panel de referencia es pequeño o grande depende de la especie, del número de individuos y de la cantidad de marcadores presentes. Finalmente, cabe destacar que desde que se realizaron los estudios y hasta la fecha, se han lanzado tres versiones adicionales del software Impute, las cuales afirman ser más precisas y rápidas en sus resultados. Sin embargo, son de licencia pagada (Stahl et al., 2021). Por el contrario, se utilizaron las versiones más recientes tanto de Beagle como de Minimac, por lo que los resultados obtenidos son el reflejo de la calidad de imputación que alcanzan ambos programas hasta la fecha.

6. CONCLUSIONES

La imputación genética se presenta como una técnica complementaria a la secuenciación a baja cobertura, técnica en auge debido a la identificación de un elevado número de polimorfismos a un bajo coste, aunque con un importante porcentaje de genotipos faltantes. La combinación de ambas técnicas se presenta como una propuesta interesante para competir con la secuenciación a alta cobertura en términos de cantidad de polimorfismos identificados, con la ventaja de hacerlo a un coste reducido. Sin embargo, la fiabilidad de los resultados se ve afectada por la calidad del panel de referencia y del genotipado de la población en estudio, así como del programa seleccionado para realizar la imputación.

En este trabajo se evaluó la actuación de tres programas de imputación genética, Beagle 5.1., Impute 2 y Minimac 4, sobre datos de genotipado de baja cobertura de berenjena empleando un panel de referencia. En general, Minimac 4 fue el programa con el que se obtuvieron genotipos imputados con una mayor precisión y confianza. Beagle 5.1 fue el menos demandante a nivel de tiempo y el único que permitió realizar el phasing del panel de referencia, aunque también el que más memoria requirió. Por su parte, Impute 2 fue más complejo en su ejecución y el que más complicaciones presentó a la hora de preparar los archivos de entrada y de interpretar los de salida.

Finalmente, la aplicación de la imputación genética debe ser analizada caso por caso, ya que cada programa puede ser más o menos preciso según los datos de entrada y el panel de referencia. Es aconsejable partir de un panel de referencia formado por individuos emparentados con la población de estudio y de secuenciado a alta cobertura para que la inferencia de datos sea lo más precisa posible.

7. BIBLIOGRAFÍA

Adarraga Mejía, J. E., Padilla González, F., & Ariza Molina, F. M. (2022). *Trazabilidad del proceso productivo de cultivos de berenjena en el departamento de Atlántico*. Universidad Nacional Abierta y a Distancia. https://doi.org/10.22490/ecacen.5787

Adhikari, L., Shrestha, S., Wu, S., Crain, J., Gao, L., Evers, B., Wilson, D., Ju, Y., Koo, D.-H., Hucl, P., Pozniak, C., Walkowiak, S., Wang, X., Wu, J., Glaubitz, J. C., DeHaan, L., Friebe, B., & Poland, J. (2022). A high-throughput skim-sequencing approach for genotyping, dosage estimation and identifying translocations. *Scientific Reports*, *12*(1), 17583. https://doi.org/10.1038/s41598-022-19858-2

Arends, D., Prins, P., Jansen, R. C., & Broman, K. W. (2010). R/qtl: high-throughput multiple QTL mapping. *Bioinformatics*, *26*(23), 2990-2992. https://doi.org/10.1093/bioinformatics/btq565

Arrones, A., Mangino, G., Alonso, D., Plazas, M., Prohens, J., Portis, E., Barchi, L., Giuliano, G., Vilanova, S., & Gramazio, P. (2022). Mutations in the SmAPRR2 transcription factor suppressing chlorophyll pigmentation in the eggplant fruit peel are key drivers of a diversified colour palette. *Frontiers in Plant Science*, *13*, 1025951. https://doi.org/10.3389/fpls.2022.1025951

Bai, W.-Y., Zhu, X.-W., Cong, P.-K., Zhang, X.-J., Richards, J. B., & Zheng, H.-F. (2020). Genotype imputation and reference panel: a systematic evaluation on haplotype size and diversity. *Briefings in Bioinformatics*, *21*(5), 1806-1817. https://doi.org/10.1093/bib/bbz108

Baldrighi, G. N., Nova, A., Bernardinelli, L., & Fazia, T. (2022). A Pipeline for Phasing and Genotype Imputation on Mixed Human Data (Parents-Offspring Trios and Unrelated Subjects) by Reviewing Current Methods and Software. *Life (Basel, Switzerland)*, *12*(12), 2030. https://doi.org/10.3390/life12122030

Bandillo, N., Raghavan, C., Muyco, P. A., Sevilla, M. A. L., Lobina, I. T., Dilla-Ermita, C. J., Tung, C.-W., McCouch, S., Thomson, M., Mauleon, R., Singh, R. K., Gregorio, G., Redoña, E., & Leung, H. (2013). Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice*, *6*(1), 11. https://doi.org/10.1186/1939-8433-6-11

Barchi, L., Acquadro, A., Alonso, D., Aprea, G., Bassolino, L., Demurtas, O., Ferrante, P., Gramazio, P., Mini, P., Portis, E., Scaglione, D., Toppino, L., Vilanova, S., Díez, M. J., Rotino, G. L., Lanteri, S., Prohens, J., & Giuliano, G. (2019). Single Primer Enrichment Technology (SPET) for High-Throughput Genotyping in Tomato and Eggplant Germplasm. *Frontiers in Plant Science*, *10*. https://doi.org/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., Comino, C., Molin, A. D., Ferrarini, A., Maor, L. C., Portis, E., Reyes-Chin-Wo, S., Rinaldi, R., Sala, T., Scaglione, D., … Rotino, G. L. (2019). A chromosome-anchored eggplant genome sequence reveals key events in Solanaceae evolution. *Scientific Reports*, *9*(1), 11769. https://doi.org/10.1038/s41598-019-47985-w

Browning, B. L., & Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics*, *84*(2), 210-223. https://doi.org/10.1016/j.ajhg.2009.01.005

Browning, B. L., & Browning, S. R. (2016). Genotype Imputation with Millions of Reference Samples. *The American Journal of Human Genetics*, *98*(1), 116-126. https://doi.org/10.1016/j.ajhg.2015.11.020

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. *American Journal of Human Genetics*, *103*(3), 338-348. https://doi.org/10.1016/j.ajhg.2018.07.015

Buckley, R. M., Harris, A. C., Wang, G.-D., Whitaker, D. T., Zhang, Y.-P., & Ostrander, E. A. (2022). Best practices for analyzing imputed genotypes from low-pass sequencing in dogs. *Mammalian Genome*, *33*(1), 213-229. https://doi.org/10.1007/s00335-021-09914-z

Cericola, F., Portis, E., Toppino, L., Barchi, L., Acciarri, N., Ciriaci, T., Sala, T., Rotino, G. L., & Lanteri, S. (2013). The population structure and diversity of eggplant from Asia and the Mediterranean Basin. *PloS One*, *8*(9), e73702-e73702. https://doi.org/10.1371/journal.pone.0073702

Chat, V., Ferguson, R., Morales, L., & Kirchhoff, T. (2022). Ultra Low-Coverage Whole-Genome Sequencing as an Alternative to Genotyping Arrays in Genome-Wide Association Studies. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.790445

Chi Duong, V., Minh Vu, G., Khac Nguyen, T., Tran The Nguyen, H., Luong Pham, T., S. Vo, N., & Hong Hoang, T. (2023). A rapid and reference-free imputation method for low-cost genotyping platforms. *Scientific Reports*, *13*(1), 23083. https://doi.org/10.1038/s41598-023-50086-4

Chou, W.-C., Zheng, H.-F., Cheng, C.-H., Yan, H., Wang, L., Han, F., Richards, J. B., Karasik, D., Kiel, D. P., & Hsu, Y.-H. (2016). A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Scientific Reports*, *6*(1), 39313. https://doi.org/10.1038/srep39313

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Das, S., Abecasis, G. R., & Browning, B. L. (2018). Genotype Imputation from Large Reference Panels. *Annual Review of Genomics and Human Genetics*, *19*(1), 73-96. https://doi.org/10.1146/annurev-genom-083117-021602

Dash, S. P., Singh, J., Gandhi, I., Vishwavidyalaya, K., Sharma, I. D., & Sharma, D. (2019). Morphological characterization of brinjal (Solanum melongena L.) germplasm. *Journal of Pharmacognosy and Phytochemistry*, *8*(2).

Davies, R. W., Flint, J., Myers, S., & Mott, R. (2016). Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, *48*(8), 965-969. https://doi.org/10.1038/ng.3594

de Givry, S., Bouchez, M., Chabrier, P., Milan, D., & Schiex, T. (2005). CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics*, *21*(8), 1703-1704. https://doi.org/10.1093/bioinformatics/bti222

De Marino, A., Mahmoud, A. A., Bose, M., Bircan, K. O., Terpolovsky, A., Bamunusinghe, V., Bohn, S., Khan, U., Novković, B., & Yazdi, P. G. (2022). A comparative analysis of current phasing and imputation software. *PloS One*, *17*(10), e0260177-e0260177. https://doi.org/10.1371/journal.pone.0260177

Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., Francioli, L. C., Hottenga, J. J., Karssen, L. C., Estrada, K., Kreiner-Møller, E., Rivadeneira, F., van Setten, J., Gutierrez-Achury, J., Westra, H.-J., Franke, L., van Enckevort, D., Dijkstra, M., Byelas, H., … Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European Journal of Human Genetics*, *22*(11), 1321-1326. https://doi.org/10.1038/ejhg.2014.19

Deng, T., Zhang, P., Garrick, D., Gao, H., Wang, L., & Zhao, F. (2022). Comparison of Genotype Imputation for SNP Array and Low-Coverage Whole-Genome Sequencing Data. *Frontiers in Genetics*, *12*. https://doi.org/10.3389/fgene.2021.704118

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., Beebe, S. E., & Raatz, B. (2020). Genetic mapping for agronomic traits in a MAGIC population of common bean (Phaseolus vulgaris L.) under drought conditions. *BMC Genomics*, *21*(1), 799. https://doi.org/10.1186/s12864-020-07213-6

Doganlar, S., Frary, A., Daunay, M.-C., Lester, R. N., & Tanksley, S. D. (2002). A comparative genetic linkage map of eggplant (Solanum melongena) and its implications for genome evolution in the solanaceae. *Genetics*, *161*(4), 1697-1711. https://doi.org/10.1093/genetics/161.4.1697

Ellinghaus, D., Schreiber, S., Franke, A., & Nothnagel, M. (2009). Current software for genotype imputation. *Human Genomics*, *3*(4), 371-380. https://doi.org/10.1186/1479-7364-3-4-371

FAO. (2024). Overview of Global Eggplant Market. En *FAO*. https://www.tridge.com/intelligences/eggplant

French, J. N., Pua, V. B., Laboulaye, R., Leal, T. P., Olivas, M. C., Lima-Costa, M. F., Horta, B. L., Barreto, M. L., Tarazona-Santos, E., Mata, I., & O'Connor, T. D. (2024). Comparing the effect of imputation reference panel composition in four distinct Latin American cohorts. *bioRxiv : the preprint server for biology*. https://doi.org/10.1101/2024.04.11.589057

Fuchsberger, C., Abecasis, G. R., & Hinds, D. A. (2015). minimac2: faster genotype imputation. *Bioinformatics*, *31*(5), 782-784. https://doi.org/10.1093/bioinformatics/btu704

Gao, X., Marjoram, P., Mckean-Cowdin, R., Torres, M., Gauderman, W. J., & Varma, R. (2012). Genotype Imputation for Latinos Using the HapMap and 1000 Genomes Project Reference Panels. *Frontiers in Genetics*, *3*. https://doi.org/10.3389/fgene.2012.00117

Ghoreishifar, S. M., Moradi-Shahrbabak, H., Moradi-Shahrbabak, M., Nicolazzi, E. L., Williams, J. L., Iamartino, D., & Nejati-Javaremi, A. (2018). Accuracy of imputation of single-nucleotide polymorphism marker genotypes for water buffaloes (Bubalus bubalis) using different reference population sizes and imputation tools. *Livestock Science*, *216*, 174-182. https://doi.org/10.1016/j.livsci.2018.08.009

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. *Frontiers in Plant Science*, *8*, 1477. https://doi.org/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. *Frontiers in Plant Science*, *10*. https://doi.org/10.3389/fpls.2019.01220

Han, S.-H., Kim, K. W., Kim, S., & Youn, Y. C. (2018). Artificial Neural Network: Understanding the Basic Concepts without Mathematics. *Dementia and Neurocognitive Disorders*, *17*(3), 83. https://doi.org/10.12779/dnd.2018.17.3.83

Holland, J. B. (2015). MAGIC maize: a new resource for plant genetics. *Genome Biology*, *16*(1), 163. https://doi.org/10.1186/s13059-015-0713-2

Hortoinfo. (2024, abril 3). *Más de 8 de cada 10 berenjenas que exporta España salen de Almería, que bate el récord histórico de ingresos*. Hortoinfo. https://hortoinfo.es/mas-de-8-de-cada-10-berenjenas-que-exporta-espana-salen-de-almeria-que-bate-el-record-historico-de-ingresos/

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, *44*(8), 955-959. https://doi.org/10.1038/ng.2354

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, *5*(6), e1000529-e1000529. https://doi.org/10.1371/journal.pgen.1000529

Huang, B. E., & George, A. W. (2011). R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines. *Bioinformatics*, *27*(5), 727-729. https://doi.org/10.1093/bioinformatics/btq719

Huang, B. E., Verbyla, K. L., Verbyla, A. P., Raghavan, C., Singh, V. K., Gaur, P., Leung, H., Varshney, R. K., & Cavanagh, C. R. (2015). MAGIC populations in crops: current status and future prospects. *Theoretical and Applied Genetics*, *128*(6), 999-1017. https://doi.org/10.1007/s00122-015-2506-0

Huang, G.-H., & Tseng, Y.-C. (2014). Genotype imputation accuracy with different reference panels in admixed populations. *BMC Proceedings*, *8*(S1), S64. https://doi.org/10.1186/1753-6561-8-S1-S64

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., & Scheet, P. (2009). Genotype-Imputation Accuracy across Worldwide Human

Populations. *The American Journal of Human Genetics*, *84*(2), 235-250. https://doi.org/10.1016/j.ajhg.2009.01.013

Keurentjes, J. J. B., Bentsink, L., Alonso-Blanco, C., Hanhart, C. J., Blankestijn-De Vries, H., Effgen, S., Vreugdenhil, D., & Koornneef, M. (2007). Development of a near-isogenic line population of Arabidopsis thaliana and comparison of mapping power with a recombinant inbred line population. *Genetics*, *175*(2), 891-905. https://doi.org/10.1534/genetics.106.066423

Khaleghi, S., Baninasab, B., & Mobli, M. (2021). RELATIONSHIP BETWEEN FLORAL MORPHOLOGY, FRUIT SETTING BEHAVIOR AND FINAL YIELD IN SOME EGGPLANT (Solanum melongena) GENOTYPES FROM IRAN. *Chilean journal of agricultural & animal sciences*, *37*(2), 128-135. https://doi.org/10.29393/chjaas37-15rbsm30015

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-e57039. https://doi.org/10.1371/journal.pone.0057039

Kojima, K., Tadaka, S., Katsuoka, F., Tamiya, G., Yamamoto, M., & Kinoshita, K. (2020). A genotype imputation method for de-identified haplotype reference information by using recurrent neural network. *PLOS Computational Biology*, *16*(10), e1008207. https://doi.org/10.1371/journal.pcbi.1008207

Korkuć, P., Arends, D., & Brockmann, G. A. (2019). Finding the Optimal Imputation Strategy for Small Cattle Populations. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.00052

Kreiner-Møller, E., Medina-Gomez, C., Uitterlinden, A. G., Rivadeneira, F., & Estrada, K. (2015). Improving accuracy of rare variant imputation with a two-step imputation approach. *European Journal of Human Genetics*, *23*(3), 395-400. https://doi.org/10.1038/ejhg.2014.91

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. *Journal of Genetics*, *100*(2), 38. https://doi.org/10.1007/s12041-021-01285-3

Lander, E. S., Green, P., Abrahamson, J., Barlow, A., Daly, M. J., Lincoln, S. E., & Newberg, L. A. (1987). MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, *1*(2), 174-181. https://doi.org/10.1016/0888-7543(87)90010-3

Lebeau, A., Gouy, M., Daunay, M. C., Wicker, E., Chiroleu, F., Prior, P., Frary, A., & Dintinger, J. (2012). Genetic mapping of a major dominant gene for resistance to Ralstonia solanacearum in eggplant. *Theoretical and Applied Genetics*, *126*(1), 143-158. https://doi.org/10.1007/s00122-012-1969-5

León Pacheco, R. I., Correa Álvarez, E. M., Romero Ferrer, J. L., Arias Bonilla, H., Gómez-Correa, J. C., Yacomelo Hernández, M. J., & Pérez Artiles, L. (2019). Accumulation of degree days and their effect on the potential yield of 15 eggplant (Solanum melongena L.) accessions in the Colombian Caribbean. *Revista*

Facultad Nacional de Agronomía Medellín, *72*(3), 8917-8926. https://doi.org/10.15446/rfnam.v72n3.77112

Li, S. S. (2003). Estimating haplotype frequencies and standard errors for multiple single nucleotide polymorphisms. *Biostatistics*, *4*(4), 513-522. https://doi.org/10.1093/biostatistics/4.4.513

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 Research*, *21*(6), 940-951. https://doi.org/10.1101/gr.117259.110

Li, Y., Willer, C. J., Ding, J., Scheet, P., & Abecasis, G. R. (2010). MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, *34*(8), 816-834. https://doi.org/10.1002/gepi.20533

Limborg, M. T., McKinney, G. J., Seeb, L. W., & Seeb, J. E. (2016). Recombination patterns reveal information about centromere location on linkage maps. *Molecular Ecology Resources*, *16*(3), 655-661. https://doi.org/10.1111/1755-0998.12484

Liu, Q., Cirulli, E. T., Han, Y., Yao, S., Liu, S., & Zhu, Q. (2015). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Briefings in Bioinformatics*, *16*(4), 549-562. https://doi.org/10.1093/bib/bbu035

Luo, Y., Kanai, M., Choi, W., Li, X., Sakaue, S., Yamamoto, K., Ogawa, K., Gutierrez-Arcelus, M., Gregersen, P. K., Stuart, P. E., Elder, J. T., Forer, L., Schönherr, S., Fuchsberger, C., Smith, A. V., Fellay, J., Carrington, M., Haas, D. W., Guo, X., … Raychaudhuri, S. (2021). A high-resolution HLA reference panel capturing global population diversity enables multi-ancestry fine-mapping in HIV host response. *Nature Genetics*, *53*(10), 1504-1516. https://doi.org/10.1038/s41588-021-00935-7

Maccaferri, M., Ricci, A., Salvi, S., Milner, S. G., Noli, E., Martelli, P. L., Casadio, R., Akhunov, E., Scalabrin, S., Vendramin, V., Ammar, K., Blanco, A., Desiderio, F., Distelfeld, A., Dubcovsky, J., Fahima, T., Faris, J., Korol, A., Massi, A., … Tuberosa, R. (2015). A high-density, <scp>SNP</scp> -based consensus map of tetraploid wheat as a bridge to integrate durum and bread wheat genomics and breeding. *Plant Biotechnology Journal*, *13*(5), 648-663. https://doi.org/10.1111/pbi.12288

Mackay, I. J., Bansept-Basler, P., Barber, T., Bentley, A. R., Cockram, J., Gosman, N., Greenland, A. J., Horsnell, R., Howells, R., O'Sullivan, D. M., Rose, G. A., & Howell, P. J. (2014). An Eight-Parent Multiparent Advanced Generation Inter-Cross Population for Winter-Sown Wheat: Creation, Properties, and Validation. *G3 Genes|Genomes|Genetics*, *4*(9), 1603-1610. https://doi.org/10.1534/g3.114.012963

Malhotra, A., Kobes, S., Bogardus, C., Knowler, W. C., Baier, L. J., & Hanson, R. L. (2014). Assessing accuracy of genotype imputation in American Indians. *PloS One*, *9*(7), e102544-e102544. https://doi.org/10.1371/journal.pone.0102544

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. *Frontiers in Plant Science*, *13*, 847789. https://doi.org/10.3389/fpls.2022.847789

Marchini, J. (2019). Haplotype Estimation and Genotype Imputation. En *Handbook of Statistical Genomics* (pp. 87-114). Wiley. https://doi.org/10.1002/9781119487845.ch3

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 499-511. https://doi.org/10.1038/nrg2796

Ministerio de Agricultura, P. y A. (2007). *Berenjena: Distribución geográfica e importancia económica*. Ministerio de Agricultura, Pesca y Alimentación . https://www.mapa.gob.es/app/MaterialVegetal/fichaMaterialVegetal.aspx?idFicha=3995

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. *Proceedings of the National Academy of Sciences*, *97*(23), 12649-12654. https://doi.org/10.1073/pnas.230304397

Naito, T., & Okada, Y. (2024). Genotype imputation methods for whole and complex genomic regions utilizing deep learning technology. *Journal of Human Genetics*. https://doi.org/10.1038/s10038-023-01213-6

Nelson, S. C., Doheny, K. F., Pugh, E. W., Romm, J. M., Ling, H., Laurie, C. A., Browning, S. R., Weir, B. S., & Laurie, C. C. (2013). Imputation-Based Genomic Coverage Assessments of Current Human Genotyping Arrays. *G3 Genes|Genomes|Genetics*, *3*(10), 1795-1807. https://doi.org/10.1534/g3.113.007161

Ng, P. C., & Kirkness, E. F. (2010). *Whole Genome Sequencing* (pp. 215-226). https://doi.org/10.1007/978-1-60327-367-1_12

Novakazi, F., Krusell, L., Jensen, J., Orabi, J., Jahoor, A., & Bengtsson, T. (2020). You Had Me at “MAGIC”!: Four Barley MAGIC Populations Reveal Novel Resistance QTL for Powdery Mildew. *Genes*, *11*(12), 1512. https://doi.org/10.3390/genes11121512

O’Connell, J., Yun, T., Moreno, M., Li, H., Litterman, N., Kolesnikov, A., Noblin, E., Chang, P.-C., Shastri, A., Dorfman, E. H., Shringarpure, S., Aslibekyan, S., Babalola, E., Bell, R. K., Bielenberg, J., Bryc, K., Bullis, E., Coker, D., Partida, G. C., … McLean, C. Y. (2021). A population-specific reference panel for improved genotype imputation in African Americans. *Communications Biology*, *4*(1), 1269. https://doi.org/10.1038/s42003-021-02777-9

Odell, S. G., Hudson, A. I., Praud, S., Dubreuil, P., Tixier, M.-H., Ross-Ibarra, J., & Runcie, D. E. (2022). Modeling allelic diversity of multiparent mapping populations affects detection of quantitative trait loci. *G3 Genes|Genomes|Genetics*, *12*(3). https://doi.org/10.1093/g3journal/jkac011

Page, A., Gibson, J., Meyer, R. S., & Chapman, M. A. (2019). Eggplant Domestication: Pervasive Gene Flow, Feralization, and Transcriptomic Divergence. *Molecular Biology and Evolution*, *36*(7), 1359-1372. https://doi.org/10.1093/molbev/msz062

Pascual, L., Desplat, N., Huang, B. E., Desgroux, A., Bruguier, L., Bouchet, J., Le, Q. H., Chauchard, B., Verschave, P., & Causse, M. (2015). Potential of a tomato <scp>MAGIC</scp> population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. *Plant Biotechnology Journal*, *13*(4), 565-577. https://doi.org/10.1111/pbi.12282

Plazas, M., Vilanova, S., Gramazio, P., Rodríguez-Burruezo, A., Fita, A., Herraiz, F. J., Ranil, R., Fonseka, R., Niran, L., Fonseka, H., Kouassi, B., Kouassi, A., Kouassi, A., & Prohens, J. (2016). Interspecific Hybridization between Eggplant and Wild Relatives from Different Genepools. *Journal of the American Society for Horticultural Science*, *141*(1), 34-44. https://doi.org/10.21273/jashs.141.1.34

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. *Journal of the American Society for Horticultural Science*, *130*(1), 54-63. https://doi.org/10.21273/jashs.130.1.54

Qu, P., Wang, J., Wen, W., Gao, F., Liu, J., Xia, X., Peng, H., & Zhang, L. (2021). Construction of Consensus Genetic Map With Applications in Gene Mapping of Wheat (Triticum aestivum L.) Using 90K SNP Array. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.727077

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rüeger, S., McDaid, A., & Kutalik, Z. (2018). Evaluation and application of summary statistic imputation to discover new height-associated loci. *PLOS Genetics*, *14*(5), e1007371. https://doi.org/10.1371/journal.pgen.1007371

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., & Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, *449*(7164), 913-918. https://doi.org/10.1038/nature06250

Sawa, T., Nakao, M., Akaike, T., Ono, K., & Maeda, H. (1998). Alkylperoxyl Radical-Scavenging Activity of Various Flavonoids and Other Phenolic Compounds: Implications for the Anti-Tumor-Promoter Effect of Vegetables. *Journal of Agricultural and Food Chemistry*, *47*(2), 397-402. https://doi.org/10.1021/jf980765e

Schurz, H., Müller, S. J., van Helden, P. D., Tromp, G., Hoal, E. G., Kinnear, C. J., & Möller, M. (2019). Evaluating the Accuracy of Imputation Methods in a Five-Way Admixed Population. *Frontiers in Genetics*, *10*, 34. https://doi.org/10.3389/fgene.2019.00034

Scott, M. F., Ladejobi, O., Amer, S., Bentley, A. R., Biernaskie, J., Boden, S. A., Clark, M., Dell'Acqua, M., Dixon, L. E., Filippi, C. V, Fradgley, N., Gardner, K. A., Mackay, I. J., O'Sullivan, D., Percival-Alwyn, L., Roorkiwal, M., Singh, R. K., Thudi, M., Varshney, R. K., … Mott, R. (2020). Multi-parent populations in crops: a toolbox integrating genomics and genetic mapping with breeding. *Heredity*, *125*(6), 396-416. https://doi.org/10.1038/s41437-020-0336-6

Sengupta, D., Botha, G., Meintjes, A., Mbiyavanga, M., Hazelhurst, S., Mulder, N., Ramsay, M., & Choudhury, A. (2023). Performance and accuracy evaluation of

reference panels for genotype imputation in sub-Saharan African populations. *Cell Genomics*, *3*(6), 100332. https://doi.org/10.1016/j.xgen.2023.100332

Shah, R., Keeble-Gagnère, G., & Whan, A. (2020). Accurate calling of homeoallelic genotypes of iSelect markers using inbred structured populations. *Bioinformatics*, *36*(15), 4240-4247. https://doi.org/10.1093/bioinformatics/btaa295

Shi, S., Yuan, N., Yang, M., Du, Z., Wang, J., Sheng, X., Wu, J., & Xiao, J. (2018). Comprehensive Assessment of Genotype Imputation Performance. *Human Heredity*, *83*(3), 107-116. https://doi.org/10.1159/000489758

Shivakumar, M., Kumawat, G., Gireesh, C., Ramesh, S. V., & Husain, S. M. (2018). Soybean MAGIC Population:A Novel Resource for Genetics and Plant Breeding. *Current Science*, *114*(04), 906. https://doi.org/10.18520/cs/v114/i04/906-908

Solanke, S., & Tawar, M. (2019). Phytochemical Information and Pharmacological Activities of Eggplant (Solanum Melongena L.): A Comprehensive Review. *EAS Journal of Pharmacy and Pharmacology* , *1*(5). https://doi.org/10.36349/EASJPP.2019.v01i05.001

Solberg, S. , van Zonneveld, M. , Rakha, M. T. , Taher, D. I. , Prohens, J. , Jarret, R., van Dooijeweert, W. , & Peter Giovannini. (2021). Global strategy for the conservation and use of eggplants. *Global Crop Diversity Trust*. https://www.croptrust.org

Stadlmeier, M., Hartl, L., & Mohler, V. (2018). Usefulness of a Multiparent Advanced Generation Intercross Population With a Greatly Reduced Mating Design for Genetic Studies in Winter Wheat. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.01825

Stahl, K., Gola, D., & König, I. R. (2021). Assessment of Imputation Quality: Comparison of Phasing and Imputation Algorithms in Real Data. *Frontiers in Genetics*, *12*, 724037. https://doi.org/10.3389/fgene.2021.724037

Stam, P. (1993). Construction of integrated genetic linkage maps by means of a new computer package: Join Map. *The Plant Journal*, *3*(5), 739-744. https://doi.org/10.1111/j.1365-313X.1993.00739.x

Sticca, E. L., Belbin, G. M., & Gignoux, C. R. (2021). Current Developments in Detection of Identity-by-Descent Methods and Applications. *Frontiers in Genetics*, *12*, 722602. https://doi.org/10.3389/fgene.2021.722602

Sulli, M., Barchi, L., Toppino, L., Diretto, G., Sala, T., Lanteri, S., Rotino, G. L., & Giuliano, G. (2021). An Eggplant Recombinant Inbred Population Allows the Discovery of Metabolic QTLs Controlling Fruit Nutritional Quality. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.638195

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 Science*, *61*(2), 839-852. https://doi.org/10.1002/csc2.20377

Taher, D., Solberg, S. Ø., Prohens, J., Chou, Y.-Y., Rakha, M., & Wu, T.-H. (2017). World Vegetable Center Eggplant Collection: Origin, Composition, Seed Dissemination and Utilization in Breeding. *Frontiers in Plant Science*, *8*, 1484. https://doi.org/10.3389/fpls.2017.01484

Toppino, L., Barchi, L., Mercati, F., Acciarri, N., Perrone, D., Martina, M., Gattolin, S., Sala, T., Fadda, S., Mauceri, A., Ciriaci, T., Carimi, F., Portis, E., Sunseri, F., Lanteri, S., & Rotino, G. L. (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. https://doi.org/10.3390/genes11070745

Torkamaneh, D., & Belzile, F. (2015). Scanning and Filling: Ultra-Dense SNP Genotyping Combining Genotyping-By-Sequencing, SNP Array and Whole-Genome Resequencing Data. *PLOS ONE*, *10*(7), e0131533. https://doi.org/10.1371/journal.pone.0131533

Treccani, M., Locatelli, E., Patuzzo, C., & Malerba, G. (2023). A broad overview of genotype imputation: Standard guidelines, approaches, and future investigations in genomic association studies. *BIOCELL*, *47*(6), 1225-1241. https://doi.org/10.32604/biocell.2023.027884

Van Leeuwen, E. M., Kanterakis, A., Deelen, P., Kattenberg, M. V., Slagboom, P. E., De Bakker, P. I. W., Wijmenga, C., Swertz, M. A., Boomsma, D. I., Van Duijn, C. M., Karssen, L. C., & Hottenga, J. J. (2015). Population-specific genotype imputations using minimac or IMPUTE2. *Nature Protocols*, *10*(9), 1285-1296. https://doi.org/10.1038/nprot.2015.077

Verma, S. S., de Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., Mukherjee, S., Jarvik, G. P., Kottyan, L. C., Burt, A., Bradford, Y., Armstrong, G. D., Derr, K., Crawford, D. C., Haines, J. L., Li, R., Crosslin, D., & Ritchie, M. D. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics*, *5*. https://doi.org/10.3389/fgene.2014.00370

Vorontsova, M. S., & Knapp, S. (2012). A new species of Solanum (Solanaceae) from South Africa related to the cultivated eggplant. *PhytoKeys*, *8*, 1-11. https://doi.org/10.3897/phytokeys.8.2462

Vorontsova, Maria. S., Stern, S., Bohs, L., & Knapp, S. (2013). African spiny*Solanum*(subgenus*Leptostemonum*, Solanaceae): a thorny phylogenetic tangle. *Botanical Journal of the Linnean Society*, *173*(2), 176-193. https://doi.org/10.1111/boj.12053

Wang, M., Qi, Z., Thyssen, G. N., Naoumkina, M., Jenkins, J. N., McCarty, J. C., Xiao, Y., Li, J., Zhang, X., & Fang, D. D. (2022). Genomic interrogation of a MAGIC population highlights genetic factors controlling fiber quality traits in cotton. *Communications Biology*, *5*(1), 60. https://doi.org/10.1038/s42003-022-03022-7

Wang, N., Yuan, Y., Wang, H., Yu, D., Liu, Y., Zhang, A., Gowda, M., Nair, S. K., Hao, Z., Lu, Y., San Vicente, F., Prasanna, B. M., Li, X., & Zhang, X. (2020). Applications of genotyping-by-sequencing (GBS) in maize genetics and breeding. *Scientific Reports*, *10*(1), 16308. https://doi.org/10.1038/s41598-020-73321-8

Wankhade, A. P., Kadirimangalam, S. R., Viswanatha, K. P., Deshmukh, M. P., Shinde, V. S., Deshmukh, D. B., & Pasupuleti, J. (2021). Variability and Trait Association Studies for Late Leaf Spot Resistance in a Groundnut MAGIC Population. *Agronomy*, *11*(11), 2193. https://doi.org/10.3390/agronomy11112193

Wickham, H. (2009). *ggplot2*. Springer New York. https://doi.org/10.1007/978-0-387-98141-3

Wickland, D. P., Battu, G., Hudson, K. A., Diers, B. W., & Hudson, M. E. (2017). A comparison of genotyping-by-sequencing analysis methods on low-coverage crop datasets shows advantages of a new workflow, GB-eaSy. *BMC Bioinformatics*, *18*(1), 586. https://doi.org/10.1186/s12859-017-2000-6

Xu, J., Liu, D., Hassan, A., Genovese, G., Cote, A. C., Fennessy, B., Cheng, E., Charney, A. W., Knowles, J. A., Ayub, M., Peterson, R. E., Bigdeli, T. B., & Huckins, L. M. (2023). Evaluation of imputation performance of multiple reference panels in a Pakistani population. *medRxiv : the preprint server for health sciences*. https://doi.org/10.1101/2023.12.22.23300448

Yoon, B.-J. (2009). Hidden Markov Models and their Applications in Biological Sequence Analysis. *Current Genomics*, *10*(6), 402-415. https://doi.org/10.2174/138920209789177575

Yu, A., Li, F., Xu, W., Wang, Z., Sun, C., Han, B., Wang, Y., Wang, B., Cheng, X., & Liu, A. (2019). Application of a high-resolution genetic map for chromosome-scale genome assembly and fine QTLs mapping of seed size and weight traits in castor bean. *Scientific Reports*, *9*(1), 11950. https://doi.org/10.1038/s41598-019-48492-8

Yuan, G., Sun, K., Yu, W., Jiang, Z., Jiang, C., Liu, D., Wen, L., Si, H., Wu, F., Meng, H., Cheng, L., Yang, A., & Wang, Y. (2023). Development of a MAGIC population and high-resolution quantitative trait mapping for nicotine content in tobacco. *Frontiers in Plant Science*, *13*. https://doi.org/10.3389/fpls.2022.1086950

Zhang, L., Meng, L., & Wang, J. (2019). Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses. *The Crop Journal*, *7*(3), 283-293. https://doi.org/10.1016/j.cj.2018.10.006

Zhang, Z., Xiao, X., Zhou, W., Zhu, D., & Amos, C. I. (2021). False positive findings during genome-wide association studies with imputation: influence of allele frequency and imputation accuracy. *Human Molecular Genetics*, *31*(1), 146-155. https://doi.org/10.1093/hmg/ddab203

Zhao, F., Zhao, H., Wang, X., & Li, X. (2017). Construction and application potential of MAGIC population on genetic breeding of rapeseed (Brassica napus L.). *Chinese Journal of Oil Crop Sciences*, *39*(145).

Zheng, C., Boer, M. P., & van Eeuwijk, F. A. (2019). Construction of Genetic Linkage Maps in Multiparental Populations. *Genetics*, *212*(4), 1031-1044. https://doi.org/10.1534/genetics.119.302229

Zheng, H.-F., Ladouceur, M., Greenwood, C. M. T., & Richards, J. B. (2012). Effect of Genome-Wide Genotyping and Reference Panels on Rare Variants Imputation. *Journal of Genetics and Genomics*, *39*(10), 545-550. https://doi.org/10.1016/j.jgg.2012.07.002

Zheng, H.-F., Rong, J.-J., Liu, M., Han, F., Zhang, X.-W., Richards, J. B., & Wang, L. (2015). Performance of Genotype Imputation for Low Frequency and Rare Variants from the 1000 Genomes. *PLOS ONE*, *10*(1), e0116487. https://doi.org/10.1371/journal.pone.0116487

8. ANEXOS

Anexo 1. Script en Python para la preparación de los datos para usarlos en R/mpMap2

```
import csv
import sys
from io import StringIO

import pandas as pd


class DataProcessing:
    def __init__(self, founders_vcf: str, finals_vcf: str, final_path:
str):
        """
        Constructor de la clase. Esta clase se encarga de leer los archivos
vcf, limpiar y seleccionar la data
        necesaria, y comparar los archivos founders y finals para mantener
los alelos que hayan en ambos archivos.
        :param founders_vcf: Path en donde se encuentra el archivo de
founders
        :param finals_vcf: Path en donde se encuentra el archivo de finals
        :param final_path: Path donde se desea guardar el resultado final
        """

        # Se carga la data de los archivos
        founders_data, founders_columns = self.__load_vcf(founders_vcf,
True)
        finals_data, finals_columns = self.__load_vcf(finals_vcf, False)

        # Se determinan las columnas que son iguales en ambos archivos
        common_columns = self.__get_common_columns(founders_columns,
finals_columns)

        # Se guarda la data transpuesta de founders con su data filtrada
        self._save_csv('transpose/founders_data', founders_data,
common_columns)

        # Uncomment para guardar la data filtrada de finals
        # self._save_csv(finals_vcf, finals_data, common_columns)

        # Se comparan los dos archivos
        self.__compare_files(common_columns, founders_columns, finals_data,
final_path)

    @staticmethod
    def __get_common_columns(founders_columns: list, finals_columns: list):
        """
        Método para obtener las columnas en común
        :param founders_columns: Columnas del archivo de founders
        :param finals_columns: Columnas del archivo de finals
        :return: Las columnas comunes entre las dos tablas
        """

        common_columns = []
        for column in founders_columns:
            if column in finals_columns:
                common_columns.append(column)

        return common_columns
```

```
    def __load_vcf(self, vcf_file: str, delete_rows: bool):
        """
        Método para leer un archivo vcf y es transformado a un JSON
        :param vcf_file: Path del archivo vcf
        :param delete_rows: Booleano para saber si se deben de eliminar las
filas
        :return: Data filtrada, columnas
        """

        print(f'Reading: {vcf_file}')
        df = pd.read_csv(vcf_file, delimiter='\t', comment='#',
header=None)
        with open(vcf_file, 'r') as f:
            for line in f:
                if line.startswith('#CHROM'):
                    columns = line.strip().split('\t')
                    break

        df.columns = columns

        # Se transforma el VCF a CSV
        csv_buffer = StringIO()
        df.to_csv(csv_buffer, index=False)
        csv_data = csv_buffer.getvalue()
        csv_buffer.close()

        # Se transforma el CSV a JSON para mayor facilidad de manejo de
datos
        csv_buffer = StringIO(csv_data)
        csv_reader = csv.DictReader(csv_buffer)
        data = list(csv_reader)
        columns = list(csv_reader.fieldnames)
        csv_buffer.close()

        print(f'Processing: {vcf_file}')
        return self.__process_data(data, columns[0:2] + columns[9:],
delete_rows)

    def __process_data(self, file_dict: list, columns: list, delete_rows:
bool):
        """
        Método donde se procesa la data del archivo
        :param file_dict: Data del archivo en JSON
        :param columns: Columnas del archivo
        :param delete_rows: Booleano para saber si se deben de eliminar las
filas
        :return:
        """

        data_list = []
        new_columns = []
        for row in file_dict:
            data = {}

            just_zeros = True
            invalid_value = False

            chrom_pos = None

            for column in columns:
                value = row[column]
```

```
                # Si son las dos primeras columnas se realiza la siguiente
logica
                if column in columns[:2]:
                    if column == '#CHROM':
                        chrom_pos = value.replace('CH', 'Ch')
                        continue

                    elif column == 'POS':
                        if chrom_pos is None:
                            raise Exception('Pos after ?')

                        value = f'{chrom_pos}_{value}'
                        column = '#CHROM_POS'

                else:  # Para el resto de columnas se asigna el formato
necesario
                    value = value[:3]

                    if value == '0/0':
                        value = '0'

                    elif value == '1/0':
                        value = '1'
                        invalid_value = True

                    elif value == '1/1':
                        value = '2'
                        just_zeros = False

                    else:
                        value = 'NA'
                        invalid_value = True

                data[column] = value

                if column not in new_columns:
                    new_columns.append(column)

            if delete_rows:
                if just_zeros or invalid_value:
                    continue

            data_list.append(data)

        return self.__transpose_data(new_columns, data_list)

    def __transpose_data(self, columns: list, data_list: list):
        """
        Método donde se transpone la data
        :param columns: Columnas de la data
        :param data_list: Lista con la data
        :return:
        """
        # Se guarda CSV filtado en memoria
        csv_buffer = StringIO()
        csv_writer = csv.DictWriter(csv_buffer, fieldnames=columns)
        csv_writer.writeheader()
        csv_writer.writerows(data_list)
        csv_data = csv_buffer.getvalue()
        csv_buffer.close()
```

```
        # Se realiza en transpose de la data
        csv_buffer = StringIO(csv_data)
        csv_reader = csv.reader(csv_buffer)
        transposed_data = list(map(list, zip(*list(csv_reader))))
        csv_buffer.close()

        # Se guarda la data transpuesta en memoria
        csv_buffer = StringIO()
        csv_writer = csv.writer(csv_buffer)
        csv_writer.writerows(transposed_data)
        csv_data = csv_buffer.getvalue()
        csv_buffer.close()

        # Se lee laa data como JSON
        csv_buffer = StringIO(csv_data)
        csv_reader = csv.DictReader(csv_buffer)
        data = list(csv_reader)
        columns = list(csv_reader.fieldnames)
        csv_buffer.close()

        return data, columns

    @staticmethod
    def _save_csv(file_name: str, data: list, columns: list):
        """
        Método para guardar un archivo CSV
        :param file_name: Nombre del archivo
        :param data: Lista donde se encuentra la data
        :param columns: Columnas del CSV
        :return:
        """
        print(f'Saving file: {file_name}')
        with open(f'{file_name}.csv', 'w', newline='') as file:
            writer = csv.DictWriter(file, fieldnames=columns)

            writer.writeheader()

            for item in data:
                new = {}
                for column in columns:
                    new[column] = item[column]

                writer.writerow(new)

    @staticmethod
    def __compare_files(common_columns: list, founders_columns: list,
finals_data: list, output_filename: str):
        """
        Método que compara los archivos founders y finals y lo guarda
        :param common_columns: Columnas en común entre los dos archivos
        :param founders_columns: Columnas de founders
        :param finals_data: Data del archivo de founders
        :param output_filename: Path donde se desea guardar el archivo
final
        :return:
        """
        print(f'Common columns:
{len(common_columns)}/{len(founders_columns)}')

        with open(f'{output_filename}.csv', 'w', newline='') as file:
```

```
            writer = csv.DictWriter(file, fieldnames=common_columns)
            writer.writeheader()

            count = 0
            total = len(finals_data)
            for row in finals_data:
                if count % 50 == 0:
                    print(f'{count}/{total}')

                new = {}
                for column in common_columns:
                    new[column] = row[column]

                writer.writerow(new)
                count += 1


def __get_arguments():
    """
    Método para leer los argumentos cuando se corre.
    :return: Los valores de los argumentos
    """
    arguments = {}
    for arg in sys.argv[1:]:
        name, value = arg.split('=')
        arguments[name] = value

    if 'founders' not in arguments:
        raise Exception('Please add argument: "founders"')

    elif 'finals' not in arguments:
        raise Exception('Please add argument: "finals"')

    elif 'filename' not in arguments:
        raise Exception('Please add argument: "filename"')

    return arguments['founders'], arguments['finals'],
arguments['filename']


if __name__ == '__main__':
    # Comando para correr: python data_processing.py founders=path_founders
finals=path_finals filename=path_final_result
    founders_vcf_path, finals_vcf_path, output_path = __get_arguments()
    data_processing = DataProcessing(founders_vcf_path, finals_vcf_path,
output_path)
```

Anexo 2. Script en R para la construcción del mapa genético

```
setwd("/Users/carolinatinajero/Desktop")

library(mpMap2)

##lectura de los archivos

pedi <- read.csv("pedigree.csv", header=TRUE)
id <- as.character(pedi$ID) ##sacar solo los nombres de los individuos
mother <- as.numeric(pedi$MOTHER) ##sacar el indice de la madre
father <- as.numeric(pedi$FATHER) ##sacar el indice del padre

##cromosoma 1
finals <- read.csv("finals_chr09.csv", header=TRUE, row.names = 1)
##archivo con los datos de la S3(filas) y los SNPs(columnas)
founders <- read.csv("founders_chr09.csv", header=TRUE, row.names = 1)
##archivo con los datos de los parentales (NA en 1 y sin los SNPs con 0 en
todos los individuos)

pedigree <- pedigree(lineNames = id, mother, father, selfing = "finite")
##es finito porque sabemos que son 3 rondas de selfing
plot(pedigreeToGraph(pedigree)) ##confirmar que está bien el pedigree

nMarkers <- length(colnames(finals))

##codificación de cada genotipo
hetData <- replicate(nMarkers, rbind(c(0,0,0),
                                     c(2,2,2),
                                     c(0,2,1),
                                     c(2,0,1)),
                     simplify = FALSE)

names(hetData) <- colnames(finals)
hetData <- new("hetData", hetData) ##crear el nuevo archivo con los datos
finales

##importar datos a un objeto mpcross
mpgenome <- mpcross(founders, finals, pedigree, hetData = hetData)

##estimar fracción de recombinación
rf <- estimateRF(mpgenome)

##linkage groups
groups <- formGroups(rf, groups = 1, clusterBy = "theta", method =
"average")

##estimar las distancias geneticas
estimatedmapH <- estimateMap (groups, mapFunction = rfToHaldane, maxOffset
= 100)

##mapa genetico
mapH <- new("mpcrossMapped", groups, map = estimatedmapH)

centimorgansH <- as.data.frame(estimatedmapH[["1"]])
colnames(centimorgansH) <- c("Genetic_Map(cM)")
write.csv(centimorgansH, "mapH_chr01.csv")
```

Anexo 3. Script en R para la visualización del mapa genético mediante un gráfico de dispersión

```
setwd("/Users/carolinatinajero/Documents/Educación/emPLANT/UPV/TFM/Mapa
genetico/S3/Haldane")

##leer el archivo
mapa < - read.csv("mapH_chr01.csv")

##definir las variables
mb < - mapa$Physical_position..bp. / 1000000
cm < - mapa$Genetic_Map.cM.

# unificar la información en un data frame
mapa_genetico < - data.frame(mb, cm)

library(ggplot2)

summary(mapa_genetico)

##scatterplot
cromosoma < - ggplot(mapa_genetico, aes(x=mb, y=cm)) +
geom_point(size=2, color="darkgrey") +
labs(x="Mb", y="CM") +
scale_x_continuous(breaks=seq(0, 120, by=20), limits=c(0, 120)) +
scale_y_continuous(breaks=seq(0, 160, by=20), limits=c(0, 160)) +
theme_classic() +
theme(panel.background = element_rect(fill="white", colour="black"),
axis.text = element_text(size=12))


print(cromosoma)
```

Anexo 4. Script en R para la transformación del mapa genético a formato PLINK

```
setwd("/Users/carolinatinajero/Documents/Educación/emPLANT/UPV/TFM/Mapa
genetico/Beagle")

##lectura del archivo
map_data <- read.csv("mapH_chr02_beagle.csv", header = TRUE)

##extraer los datos necesarios y definir las columnas
plink_map <- map_data[, c("chr", "markerID", "geneticDistance",
"physicalPosition")]

##guardar el archivo final
write.table(plink_map,
"/Users/carolinatinajero/Documents/Educación/emPLANT/UPV/TFM/Mapa
genetico/Beagle/chr01_beagle.map", quote = FALSE, sep = "\t", row.names =
FALSE, col.names = FALSE)
```

Anexo 5. Script en R para la transformación del mapa genético a formato .map

```
setwd("/Users/carolinatinajero/Documents/Educación/emPLANT/UPV/TFM/Mapa
genetico/Impute")

##lectura del archivo
map_data <- read.csv("mapH_chr01_impute.csv", header = TRUE)

##extraer los datos necesarios y definir las columnas
map <- map_data[, c("Physical_position(bp)",    "COMBINED_rate(cM/Mb)",
"Genetic_Map(cM)")]

##guardar el archivo final
write.table(map,
"/Users/carolinatinajero/Documents/Educación/emPLANT/UPV/TFM/Mapa
genetico/Impute/mapH_chr01_impute.map", quote = FALSE, sep = "\t",
row.names = FALSE, col.names = FALSE)
```

Anexo 6. Script en R para la elaboración de los gráficos a partir de los resultados de los procesos de imputación

```
setwd("/Users/carolinatinajero/Desktop")
library(ggplot2) ##librería para hacer los gráficos

##lectura de los archivos de los datos generales
beagle <-
read.table("imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_R2_MAF_beagle.r2")
R2_beagle <- as.numeric(beagle$V7)
MAF_beagle <- as.numeric(beagle$V8)
head(beagle)

impute <- read.table("imputed_Eggplant_MAGIC_2023_S5_impute_info.txt",
header=T)
R2_impute <- as.numeric(impute$info)
MAF_impute <- as.numeric(impute$V8)
head(impute)

minimac <- read.table("imputed_Eggplant_MAGIC_2023_S5_minimac.r2")
R2_minimac <- as.numeric(minimac$V7)
MAF_minimac <- as.numeric(minimac$V8)
head(minimac)

##lectura de los archivos de solo los datos imputados
beagle_IMP <-
read.table("imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_R2_MAF_IMP_beagle.
r2")
R2_beagle_IMP <- as.numeric(beagle_IMP$V7)
MAF_beagle_IMP <- as.numeric(beagle_IMP$V8)

impute_IMP <-
read.table("imputed_Eggplant_MAGIC_2023_S5_impute_info_IMP.txt")
R2_impute_IMP <- as.numeric(impute_IMP$V2)

minimac_IMP <- read.table("imputed_Eggplant_MAGIC_2023_S5_IMP_minimac.r2")
R2_minimac_IMP <- as.numeric(minimac_IMP$V7)
MAF_minimac_IMP <- as.numeric(minimac_IMP$V8)
```

```
##lectura de los archivos de solo los datos iniciales
impute_0 <- read.table("imputed_Eggplant_MAGIC_2023_S5_SNPs_impute.txt")
R2_impute_0 <- as.numeric(impute_0$V2)
MAF_impute_0 <- as.numeric(impute_0$V3)

minimac_0 <- read.table("imputed_Eggplant_MAGIC_2023_S5_SNPs_minimac.r2")
R2_minimac_0 <- as.numeric(minimac_0$V7)
MAF_minimac_0 <- as.numeric(minimac_0$V8)

beagle_0 <-
read.table("imputed_Eggplant_MAGIC_2023_S5_gp_nthreads20_R2_MAF_SNPs_beagle
.r2")
R2_beagle_0 <- as.numeric(beagle_0$V7)
MAF_beagle_0 <- as.numeric(beagle_0$V8)

##1)scatterplot
##unir toda la información de R2 y MAF en un solo archivo
inicial_impute <- data.frame(MAF_impute_0, R2_impute_0)
inicial_minimac <- data.frame(MAF_minimac_0, R2_minimac_0)
inicial_beagle <- data.frame(MAF_beagle_0, R2_beagle_0)

##gráficos
p1 <- ggplot(completo_beagle, aes(x = MAF_beagle, y = R2_beagle)) +
  geom_point(size = 1, shape = 17, color = "salmon2") +
  labs(x = "MAF", y = "R2", title = "Beagle") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p1)

p2 <- ggplot(completo_minimac, aes(x = MAF_minimac, y = R2_minimac)) +
  geom_point(size = 1, shape = 17, color = "steelblue1") +
  labs(x = "MAF", y = "R2", title = "Minimac") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p2)

p3 <- ggplot(imp_beagle, aes(x = MAF_beagle_IMP, y = R2_beagle_IMP)) +
  geom_point(size = 1, shape = 17, color = "salmon2") +
  labs(x = "MAF", y = "R2", title = "Beagle") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p3)

p4 <- ggplot(imp_minimac, aes(x = MAF_minimac_IMP, y = R2_minimac_IMP)) +
```

```
  geom_point(size = 1, shape = 17, color = "steelblue1") +
  labs(x = "MAF", y = "R2", title = "Minimac") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p4)

p5 <- ggplot(inicial_impute, aes(x = MAF_impute_0, y = R2_impute_0)) +
  geom_point(size = 1, shape = 17, color = "springgreen3") +
  labs(x = "MAF", y = "InfoScore", title = "Impute") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p5)

p6 <- ggplot(inicial_minimac, aes(x = MAF_minimac_0, y = R2_minimac_0)) +
  geom_point(size = 1, shape = 17, color = "steelblue1") +
  labs(x = "MAF", y = "R2", title = "Minimac") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p6)

p7 <- ggplot(inicial_beagle, aes(x = MAF_beagle_0, y = R2_beagle_0)) +
  geom_point(size = 1, shape = 17, color = "salmon2") +
  labs(x = "MAF", y = "R2", title = "Beagle") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5)
  )

print(p7)

##2)boxplot
##unir toda la información de R2 en un solo archivo
R2_complete <- data.frame(value = c(R2_beagle, R2_impute, R2_minimac),
  group = factor(rep(c("Beagle", "Impute", "Minimac"), times =
c(length(R2_beagle), length(R2_impute), length(R2_minimac))))
  )
R2_IMP <- data.frame(value = c(R2_beagle_IMP, R2_impute_IMP,
R2_minimac_IMP),
                     group = factor(rep(c("Beagle", "Impute", "Minimac"),
times = c(length(R2_beagle_IMP), length(R2_impute_IMP),
length(R2_minimac_IMP))))
)

##gráficos
```

```r
p8 <- ggplot(R2_complete, aes(x = group, y = value, fill = group)) +
  geom_boxplot(outlier.shape = NA) +
  labs(x = "Programa", y = "R2", title = "Box Plot") +
  scale_y_continuous(breaks = seq(0, 1, by = 0.25)) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "none"
  )

print(p8)

p9 <- ggplot(R2_IMP, aes(x = group, y = value, fill = group)) +
  geom_boxplot(outlier.shape = NA) +
  labs(x = "Programa", y = "R2", title = "Box Plot") +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    legend.position = "none"
  )

print(p9)

##ANOVA
model_imp <- aov (R2_IMP$value ~ R2_IMP$group)
summary(model_imp)

##prueba de las suposiciones
bartlett.test(R2_IMP$value ~ R2_IMP$group)
qqnorm(model_completo$residuals); qqline(model_completo$residuals)
shapiro.test(model_completo$residuals)

##prueba posthoc
library(rstatix)
R2_IMP %>%
  games_howell_test(value ~ group)

##3)histograma
##unir toda la información de R2 en un solo archivo
data_imputada <- data.frame(value = c(R2_beagle_IMP, R2_impute_IMP,
R2_minimac_IMP),group = factor(rep(c("Beagle", "Impute", "Minimac"), times
= c(length(R2_beagle_IMP), length(R2_impute_IMP), length(R2_minimac_IMP))))
,dataset = "Datos generales")

data_completa <- data.frame(value = c(R2_beagle, R2_impute,
R2_minimac),group = factor(rep(c("Beagle", "Impute", "Minimac"), times =
c(length(R2_beagle), length(R2_impute), length(R2_minimac))))
, dataset = "Datos imputados")

##combinar ambos datos en un solo archivo
library(dplyr)
combined_data <- bind_rows(data_completa, data_imputada)
```

```r
##gráfico
p10 <- ggplot(combined_data, aes(x = value, fill = group)) +
  geom_histogram(width=0.5, position = "dodge",binwidth=0.1, aes(y =
(..count..) / sum(..count..) * 100)) +
  facet_grid(rows=vars(dataset), axes = "all", axis.labels = "all_x") +
  scale_y_continuous(labels = scales::percent_format(scale = 1), breaks =
seq(0, 100, by = 5), limits = c(0, 10), expand = c(0, 0)) +
  scale_x_continuous(breaks = seq(0, 1, by = 0.1), expand = c(0, 0),
limits= c(0,1.1)) +
  theme_grey() +
  labs(x = "R2", y = NULL, fill = "Programa") +
  theme(legend.position = "bottom", strip.text = element_text(size = 10))


print(p10)
```