Universidad Nacional de Córdoba

Maestría en Estadística Aplicada

Imputación de genotipos faltantes en datos de secuenciación masiva

Maestrando: Bioing. Gabriela Alejandra Merino

Director
Dr. Julio Alejandro Di Rienzo

Co-Director
Dra. Norma Paniego

Marzo de 2018

Imputación de genotipos faltantes en datos de secuenciación masiva

por

Bioing. Gabriela Alejandra Merino

Dr. Julio Alejandro Di Rienzo
Director

Dra. Norma Paniego
Co-Director

Tribunal Evaluador

Dra. Cecilia Inés Bruno
FCA-UNC

Dr. Máximo Lisandro Rivarola
CICVYA-INTA

Dr. Cristóbal Fresno
INMEGEN

Esta Tesis fue enviada a la Escuela de Graduados de Ciencias Económicas de la Universidad Nacional de Córdoba para cumplimentar los requerimientos de obtención del grado académico de Magister en Estadística Aplicada.

Córdoba, Argentina
Marzo de 2018

Agradecimientos

Este trabajo no habría sido posible sin el apoyo y el estímulo de mi profesor y ahora Director, Dr. Julio Di Rienzo, bajo cuya supervisión escogí este tema y comencé la tesis. La Dra. Norma Paniego, también ha sido generosamente servicial, y me ha ayudado en el desarrollo de este trabajo. No debo dejar de mencionar a la Dra. Carla Filippi quien ha sido una orientadora para mí, guiándome fundamentalmente en los inicios del desarrollo de esta tesis.

Me gustaría agradecer a mis compañeros del Grupo de Minería de Datos en Biociencias, los cuales me han acompañado y apoyado durante mis últimos cinco años de investigación. También quisiera agradecer a todos los docentes y compañeros de la Maestría en Estadística Aplicada Cohorte 2014, con quienes he compartido dos hermosos años. Agrego también a todos aquellos amigos que fui cosechando a lo largo de los años, esos que hoy, pese a la distancia, siempre recuerdo y llevo conmigo.

No puedo terminar sin agradecer a mi familia, en cuyo estímulo constante y amor he confiado a lo largo de mis años de estudio. Mis padres, Luis y Ana, y mis hermanas Vanesa y Agostina quienes han sido pilares fundamentales y fuende de inspiración para seguir siempre hacia adelante. Gracias a mis abuelos, cuñados, sobrinos, primos y tíos que siempre me han acompañado. Por útlimo a mi pequeña gran familia, mis grandes amores, mis cómplices, Diego y Cleta. Es a ellos a quienes dedico este trabajo.

Resumen

Palabras clave: *Genotipado por secuenciación*, *SNP*, *imputación*.

Las estrategias de genotipificación masiva de poblaciones de mejoramiento mediante secuenciación de alto rendimiento son cada vez más utilizadas en el ámbito de las ciencias agrarias. Tales estrategias favorecen la exploración de la diversidad genética propia de una población, aunque, generan matrices de genotipado con un alto porcentaje de datos faltantes. Para resolver esta limitante se recurre a la predicción de los genotipos faltantes mediante la implementación de técnicas estadísticas. No obstante, la mayoría de éstas han sido desarrolladas para trabajar con especies como maíz o soja que disponen de genomas de referencia de alta calidad y matrices de genotipado completo, lo que aporta información valiosa para la imputación. Sin embargo, la mayoría de los cultivos no se encuentra en esta situación en términos de información útil disponible. Esta tesis tiene como objetivo aportar soluciones al problema de imputación en matrices de genotipado obtenidas mediante secuenciación de especies poco estudiadas. Aquí se propuso diseñar una estrategia de imputación basada en la combinación de técnicas estadísticas y evidencias genéticas. Dado que la matriz de trabajo contiene muchos más genotipos incompletos que individuos genotipados, se seleccionó la metodología *Random Forest* para la predicción y posterior imputación de los genotipos faltantes. Adicionalmente, se conoce que las variantes genotípicas, en este caso polimorfismos de nucleótido único (SNPs), están correlacionadas desde el punto de vista genético (grupos de ligamiento) y/o genómico (pseudo-moléculas de ADN), por lo que se incorporó tal información con el fin de obtener resultados más precisos. En base a estos principios, se diseñaron seis alternativas de imputación y se establecieron cuatro métricas de desempeño (exactitud, F-score, sensibilidad y precisión) para su evaluación y comparación. Los algoritmos propuestos inicialmente

se ensayaron usando datos simulados y los resultados obtenidos fueron contrastados con los conseguidos al utilizar estrategias de imputación de uso frecuente, según la literatura, sobre las mismas matrices simuladas. De los seis métodos desarrollados, se encontró que el algoritmo `RFCorOOBLD` que considera la correlación entre un SNP incompleto y los SNPs completos del mismo grupo de ligamiento, y un umbral de error de predicción (OOB), fue la que logró el mejor desempeño. Si bien las estrategias que no consideran el error OOB permitieron recuperar más SNPs incompletos, `RFCorOOBLD` fue superior a todas las alternativas propuestas en términos de sensibilidad y precisión. Se analizó además el impacto de la modificación del umbral del error OOB sobre el desempeño de las estrategias evaluadas, observándose que un umbral de 0,2 permite obtener un óptimo entre el porcentaje de SNPs imputados y el máximo error de estimación admitido. Se encontró además que la metodología `RFCorOOBLD` fue la más robusta ante las variaciones en el porcentaje de genotipos faltantes en la matriz inicial, observándose también que es la que mejor desempeño ofrece en matrices con valores superiores al 20 % de datos faltantes. En cuanto al desempeño como función del porcentaje de SNPs completos, esta metodología fue una de las que más incrementó sus medidas como consecuencia del aumento de datos completos. Se demostró además que la metodología desarrollada resultó superior en desempeño respecto de otras metodologías disponibles y comúnmente utilizadas para la imputación de genotipos faltantes, como son la *imputación por la moda*, *Beagle* y *LinkImputeR*. Adicionalmente, las medidas de desempeño de las estrategias aquí propuestas fueron más robustas con respecto al porcentaje de datos faltantes que las correspondientes a las tres metodologías alternativas contrastadas. Los algoritmos desarrollados que tuvieron los mejores desempeños se aplicaron además a un estudio real basado en una matriz de datos incompletos generada mediante genotipificación por secuenciación de una población de asociación de girasol, llevada a cabo por el Instituto Nacional de Tecnología Agropecuaria. En este caso, la estrategia `RFCorOOBLD` permitió recuperar miles de SNPs incompletos, logrando conservar más del 75 % de todos los SNPs de la matriz de genotipado luego de la imputación. Por lo expuesto, se concluye que la metodología aquí presentada representa un aporte importante al problema de imputación de genotipos faltantes en matrices de genotipicación por secuenciación de individuos no relacionados o poco relacionados genéticamente.

Abstract

Keywords: *Genotyping by sequencing*, *SNP*, *imputation*.

Strategies for massive genotyping of breeding populations by means of high throughput sequencing technologies are commonly used in the agricultural field. Such strategies allow the exploration of the genetic diversity of a population but, they generate genotyping matrices with a high percentage of missing data. In order to overcome this difficulty, statistical strategies for missing genotypes prediction are implemented. However, most popular techniques have been developed to work with species like maize or soybean, which have high-quality reference genome and complete genotype matrices providing helpful information for missing genotypes imputation. Nevertheless, most crops do not have this useful information. This thesis was directed to provide solutions to the problem of imputation in genotyping matrices obtained by sequencing of less-studied species. Here, the design of an imputation strategy based on the combination of statistical techniques and genetic evidence was proposed. Given that the information matrix has more incomplete genotypes than genotyped individuals, the *Random Forest* methodology was selected to predict and impute missing genotypes. In addition, since genotype variants, here single nucleotide polymorphisms (SNPs), are correlated from a genetic (linkage group) and genomic (DNA pseudo-molecules) point of view, this information was incorporated in order to achieve precise results. Based on these principles, six imputation alternatives were designed and four performance measures (accuracy, F-score, sensitivity and precision) were established to evaluate and compare those strategies. The developed algorithms were firstly used to analyze simulated datasets and the obtained results were compared against the results of the application of commonly used methods over the same simulated datasets. Among the six developed methods, `RFCorOOBLD`, a stra-

tegy considering the correlation between an incomplete SNP and the complete SNPs within its linkage group and a threshold of the prediction error (OOB), exhibited the best performance. Although strategies that not consider the OOB error allowed to recover more incomplete SNPs, they achieved lower sensitivity and precision than `RFCorOOBLD`. The impact of the selection of the OOB threshold was also evaluated. We found that a threshold value of 0.2 allows the obtention of an optimal between the percentage of imputed SNPs and the maximal admitted estimation error. We also found that the `RFCorOOBLD` methodology was the least affected by changes in missing genotypes percentages, exhibiting the best performances for matrices with more than 20 % of missing genotypes. The influence of the percentage of complete SNPs over the strategies performances was also evaluated. In particular, `RFCorOOBLD` was one of which improved their performance measures as the number of complete SNPs increased. We also demonstrated that the developed strategy overcame the performance of available and commonly used algorithms like *mode imputation*, *Beagle* y *LinkImputeR*. The robustness of performance measures of developed strategies respect to missing genotypes percentages was higher than the achieved by the three alternatives here considered. The developed strategies which have had higher performance were applied to study a real case. It consisted of a genotype matrix generated from the massive sequencing of a sunflower association population developed by the Instituto Nacional de Tecnología Agropecuaria. In this case, the `RFCorOOBLD` strategy allowed the recovering of thousands of incomplete SNPs, achieving the conservation of more than 75 % of the SNPs in the sunflower genotype matrix. We concluded that the methodology here presented is a valuable contribution to solve the imputation of missing data in genotyping by sequencing matrices of weakly or non genetically related individuals.

Índice de figuras

Índice de tablas

Abreviaturas

ADN : Ácido DesoxirriboNucleico

ARN : Ácido RiboNucleico

CDS : Coding DNA Sequence, secuencia de ADN codificante

ddRAD-seq : double digested RAD-seq

DNA : DesoxiriboNucleic Acid, ácido desoxirribonucleico

EST : Expressed Sequence Tag, marcador de secuencia expresada

FP : Falso Positivo

FN : Falso Negativo

GBS : Genotyping by Sequencing, genotipificación por secuenciación

InDel : Inserción o Deleción de una o más nucleótidos en un fragmento de ADN/ARN

KNN : K-nearest neighbors o K vecinos más cercanos

MAF : Minimum Alternative Frequency, frecuencia alternativa mínima

MAS : Markers Assisted Selection, selección asistida de marcadores

N : Negativo

NGS : Next Generation Sequencing, Secuenciación de Segunda Generación

OOB : Out Of Bag

P : Positivo

pb : pares de bases

PCR : Polimerase Chain Reaction, reacción en cadena de la polimerasa

RAD-seq : Restriction site associated DNA sequencing, secuenciación de ADN asociado a sitios de reconocimiento de enzimas de restricción

RF : Random Forests, bosques aleatorios

RRL : Reduced Representation Library , librería de representación reducida

SNP : Single Nucleotide Polymorphism, polimorfismo de nucleótido único

SSR : Simple Sequences Repeat, secuencias simples repetidas

UTR : UnTranslated Region, región no codificante

VN : Verdadero Negativo

VP : Verdadero Positivo

Índice general

Capítulo 1

Marco Teórico

1.1. Marcadores moleculares

1.1.1. Generalidades

El estudio de las variaciones genéticas ha sido una herramienta fundamental en el ámbito de la medicina humana, como en los programas de mejoramiento de cultivos o en estudios evolutivos (Agarwal et al., 2008; Sidransky, 2002). Con el fin de detectar y monitorear estas variaciones en los individuos de una progenie, se han desarrollado herramientas genéticas llamadas *marcadores moleculares* (Mammadov et al., 2012). Los marcadores moleculares son porciones de ADN[1] que pueden medirse en uno o muchos individuos de una población y proveen información útil acerca de las variaciones que ocurren en un determinado locus o región genómica, por lo que han sido utilizados para diversos análisis genéticos (Schlötterer, 2004).

Entre todas las variaciones alélicas del genoma, las que más se utilizan en el análisis genético de una especie son los microsatélites o secuencias simples repetidas (SSR), las pequeñas inserciones o deleciones de segmentos (InDels) y los polimorfismos de nucleótido único, más conocidos como SNPs por sus siglas del inglés (Mammadov

[1]El Ácido DesoxirriboNucleico (ADN) una molécula helicoidal formada por dos hebras antiparalelas poliméricas, donde cada unidad (mero) es un nucleótido. Cada nucleótido a su vez está formado por un azúcar desoxirribosa, un grupo fosfato y una de las cuatro bases nitrogenadas: A, C, G y T. La molécula de ADN se mantiene unida gracias a la complementariedad que existe entre sus bases (A-T, C-G), de manera que la información genética contenida en una hebra está replicada en la complementaria.

et al., 2012).

A lo largo de los años, el desarrollo de los marcadores moleculares ha evolucionado en base a aspectos tales como el costo económico del método de detección, el rendimiento del mismo y el nivel de reproducibilidad (Bernardo, 2008). Los primeros marcadores, basados en la técnica de hibridación[2], fueron ampliamente utilizados pese a su elevado consumo de tiempo y costo económico. El desarrollo de la tecnología de reacción en cadena de la polimerasa (PCR)[3] favoreció al surgimiento de una segunda generación de marcadores, entre los que se destacan los SSR (Agarwal et al., 2008). Estos marcadores permitieron incrementar la reproduciblidad, posibilitando la automatización de la detección y la disminución en los tiempos de ejecución. Sin embargo, en la última década su uso se ha restringido por la aparición de los marcadores de tercera generación cuya detección es fácil de automatizar (Mammadov et al., 2012). Entre estos marcadores se destacan los marcadores de secuencia expresada o ESTs (siglas del inglés Expressed Sequence Tags) y los SNPs. Específicamente, los EST se han desarrollado para determinar variaciones en las regiones codificantes del genoma (CDS)[4], mientras que los SNPs pueden encontrarse tanto en las CDS, como en las no codificantes (UTRs)[5] (Ekblom and Galindo, 2011).

1.1.2. Polimorfismos de Nucleótido Único

Los SNPs son posiciones genómicas puntuales que presentan diferentes bases nucleotídicas en individuos de una población, con una frecuencia mínima de ocurrencia de 1 % (Brookes, 1999) (Figura 1.1). La gran mayoría de los SNPs son bialélicos, los

[2]La hibridación es un proceso por el cual se combinan dos cadenas de ácidos nucleicos complementarias simples para formar una única molécula de doble cadena por apareamiento de sus bases. Las técnicas basadas en hibridación utilizan sondas, fragmentos cortos de secuencia conocida, de ADN con las cuales hibridarán los fragmentos de ADN de la muestra. El proceso inverso a la hibridación involucra el sometimiento de la molécula de ADN a un aumento de temperatura para romper el apareamiento entre bases y separar las dos hebras. Este proceso es utilizado en distintas estrategias de biología molecular, incluso en la genotipificación.

[3]La enzima ADN polimerasa II es la responsable de la síntesis de una nueva cadena de ADN a partir de una cadena molde. El proceso de síntesis consiste en la incorporación de nucleótidos complementarios a los del ADN molde. Es la enzima encargada de la replicación del ADN.

[4]Una región codificante, también llamada CDS (Coding DNA Sequence) es la porción de un gen que contiene los exones que codifican a proteínas

[5]La sigla UTR refiere a UnTranslated Region, es decir que son regiones que no son transcritas, por lo que no son codificantes aunque colindan con la región CDS.

cuales están representados por una sustitución de una base por otra, incluyendo las transiciones purina-purina (A-G) o pirimidina-pirimidina (C-T) y las transversiones purina-pirimidina o pirimidina-purina (A-C, A-T, G-C o G-T) (Rao and Gu, 2008). El interés en el uso de los SNPs como marcadores moleculares radica en su amplia distribución y abundancia en el genóma y en su elevada fidelidad hereditaria, ya que poseen una baja tasa de mutación ($\sim 10^{-8}$) (González et al., 2015; Mammadov et al., 2012). En particular, se ha determinado que los SNPs pueden ser más abundantes en plantas que en humanos, lo cual los ha hecho realmente atractivos tanto para programas de mejoramiento vegetal como para aplicaciones de selección asistida por marcadores moleculares (MAS, de sus siglas en inglés) (Gupta et al., 2001). Por tal motivo, en los últimos 15 años, el esfuerzo se ha concentrado en el desarrollo de herramientas que permitan automatizar el estudio en paralelo de miles de SNPs, proveyendo un elevado rendimiento al menor costo posible.

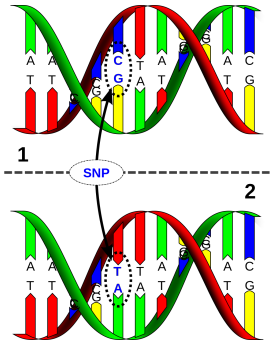

Figura 1.1: Ilustración de un polimorfismo de nucleótido único (SNP)

1.2. Genotipado mediante secuenciación masiva

En el año 2003, (Luikart et al., 2003) expresó: "El enfoque molecular ideal para el estudio de la genómica poblacional debería cubrir cientos de marcadores moleculares polimórficos a lo largo de todo el genoma en un único experimento sencillo y confiable. Desafortunadamente, en la actualidad esto no es posible". Sin embargo, no pasó mucho tiempo para que esto fuese posible, gracias al advenimiento de las técnicas de secuenciación masiva.

1.2.1. Tecnologías de secuenciación masiva

La secuenciación consiste en determinar el orden de las bases A, C, G y T en un fragmento de ADN (Jimenez-Escrig et al., 2012). La primera tecnología de secuenciación, ampliamente utilizada en los últimos 25 años, ha sido el método clásico de terminación de cadena o método de Sanger automatizado en equipos de electroforesis capilar (Lado Lindner, 2012). Si bien esta tecnología se caracteriza por una elevada eficiencia y sensibilidad, la demanda por ampliar el campo de estudio y reducir los costos y tiempos de procesamiento derivaron en el desarrollo de nuevas metodologías. En el año 2005 surgieron las tecnologías de secuenciación de segunda generación (del inglés Next Generation Sequencing, NGS), las cuales tuvieron un tremendo impacto en el desarrollo de la genómica (Morozova and Marra, 2008). Las NGS abrieron las puertas a la exploración genomas y transcriptomas completos, en forma masiva a velocidades sin precedentes, incluso para especies nunca antes estudiadas (Varshney et al., 2009). Brevemente, una corrida de secuenciación genera millones de fragmentos cortos de ADN ($\sim$ 100pb), las lecturas, a partir de fragmentos de ADN extraídos de uno o varios individuos.

Una de las grandes ventajas de las NGS es que permiten una gran diversidad de aplicaciones, dependiendo de cómo se ha obtenido y pre-procesado el ADN (o ARN) de los individuos bajo estudio, y del post-procesamiento a realizar sobre las lecturas de secuenciación. Si bien se han desarrollado varias tecnologías de secuenciación de segunda generación, todas involucran los mismos pasos: fragmentación del ADN extraído, ligación de adaptadores y amplificación mediante PCR (generación de la *librería*) y secuenciación (Shendure and Ji, 2008). La versatilidad de las NGS se debe a que no se requiere ningún conocimiento *a priori* para poder secuenciar todo o parte del genoma de un individuo. Por lo tanto, en aquellas especies que no cuentan con un genoma descrito es posible crear una referencia, esto es un ensamble *de novo*, utilizando las secuencias cortas provenientes de tecnologías NGS. Pese a esto, crear un ensamble *de novo* es uno de los grandes desafíos de las técnicas de secuenciación masiva, especialmente para aquellas especies con genomas complejos como los eucariotas (González et al., 2015).

En el contexto de los marcadores moleculares, las NGS facilitaron el desarrollo de métodos de genotipado simultáneo de miles o decenas de miles de marcadores mo-

leculares en cientos de individuos, revolucionando rápidamente la forma de pensar la genómica poblacional (Davey et al., 2011). Áreas como la medicina, agricultura, ganadería, forestería, medio ambiente, entre otras, se han visto altamente favorecidas por las innumerables aplicaciones desarrolladas a partir de las NGS (Morozova and Marra, 2008; Poland and Rife, 2012; Tucker et al., 2009). El genotipado mediante técnicas de secuenciación de alto rendimiento (GBS, por sus siglas en inglés) es una de las aplicaciones de las NGS, basada en la identificación de SNPs, que se utiliza cada vez más en el ámbito de la genómica funcional en plantas (Elshire et al., 2011; Pegadaraju et al., 2013; Peterson et al., 2012). Utilizando GBS es posible generar un mapa genético de SNPs mediante la secuenciación de genomas completos de múltiples individuos, sin embargo, los costos elevados lo hacen prácticamente imposible más aún para pequeños grupos de investigación. En este contexto, la flexibilidad de las NGS y GBS ha llevado al desarrollo de técnicas alternativas de representación reducida. Éstas permiten explorar sólo las partes del genoma relevantes para estudios específicos, logrando así una reducción en los costos de secuenciación por individuo (Poland and Rife, 2012).

1.2.2. Técnicas de representación reducida

Existen diferentes técnicas de reducción, todas basadas en el uso de enzimas de restricción como herramientas de selección de las regiones genómicas de interés. Entre las técnicas existentes se destacan las librerías de representación reducida (Reduced-representation library, RRL), la secuenciación de ADN asociado a sitios de reconocimiento de enzimas de restricción (Restriction-site-associated DNA sequencing, RAD-seq) y la técnica de secuenciación a baja profundidad conocida como genotipado mediante secuenciación (Genotyping By Sequencing, GBS) (Davey et al., 2011).

La Figura 1.2 esquematiza los protocolos de RAD-seq y GBS aplicados a dos muestras tomadas de distintos individuos, uno cuyos fragmentos de ADN se han identificado en azul, y el otro, en celeste. Ambos métodos comienzan con la digestión de las muestras mediante una enzima de restricción. En el protocolo de RAD-seq, los fragmentos generados por la digestión se ligan a etiquetas moleculares, útiles para reconocer los diferentes individuos que se secuenciarán en conjunto, conocidas como

adaptadores P1. En la Figura 1.2, éstos se representan mediante segmentos amarillos para la muestra 1 y violetas para la muestra 2. Los fragmentos modificados se fraccionan al azar usando un método físico. Posteriormente se selecciona un subconjunto de fragmentos dentro de un rango de tamaños entre 300 y 700 pb. Posteriormente

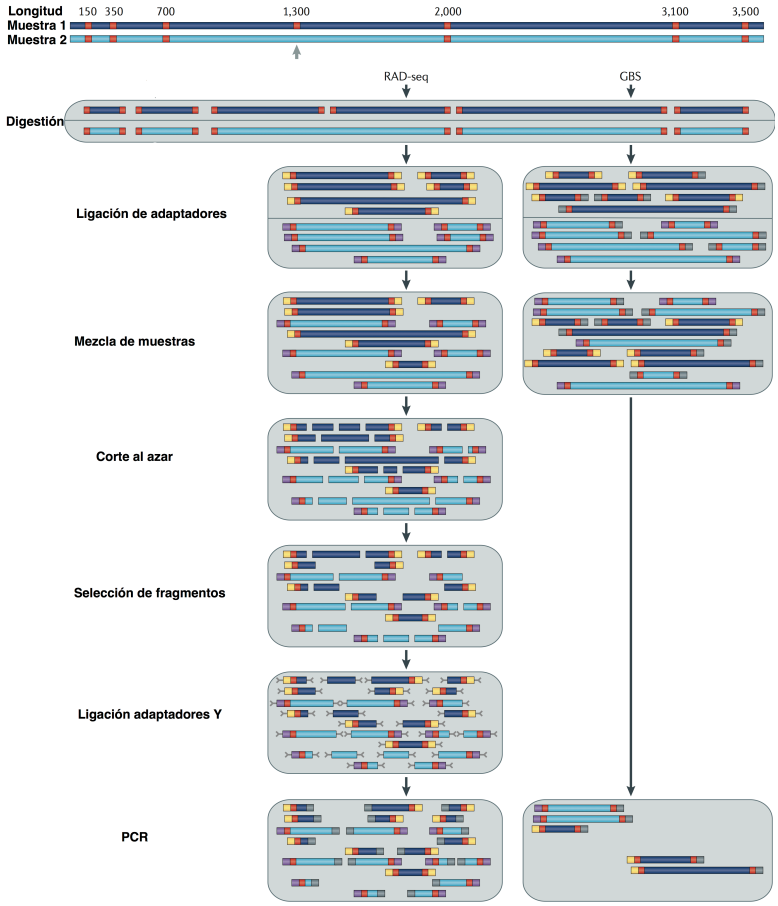

Figura 1.2: Técnicas de genotipado usando enzimas de restricción. A la izquierda, la técnica RAD-seq y a la derecha, GBS. Se ilustra una región genómica y dos muestras de ADN (una en azul, la otra en celeste) conteniendo sitios de restricción (en rojo). La muestra en celeste no contiene el sitio de corte a la altura de la base 1.300 (flecha gris). Ejemplo extraído de Davey et al. (2011).

se añaden los adaptadores P2 (gris, con forma de Y) y se amplifican mediante PCR con cebadores (primers) específicos para P1 y P2. Este procedimiento permite que únicamente los fragmentos que tienen adaptadores P1 y P2 o sea que abarcan sitios de restricción, sean amplificados para su posterior secuenciación. Por otro lado, en el

protocolo de GBS se ligan etiquetas moleculares específicas para cada muestra (segmentos amarillos o violetas) y adaptadores comunes a todas ellas (segmentos grises). Luego se realiza la digestión y se obtienen tres tipos de fragmentos: con dos etiquetas, con etiqueta y adaptador común y sólo con adaptadores comunes. Posteriormente, se mezclan las muestras y se realiza la amplificación de fragmentos mediante PCR conservando sólo aquellos que contienen combinaciones etiqueta y adaptador común.

El protocolo RAD-seq suele conllevar pérdida en la representación de fragmentos durante su proceso. Adicionalmente, carece de un control preciso de los fragmentos a secuenciar lo cual la hace una técnica poco efectiva. Para solventar estas dificultades se ha desarrollado una alternativa conocida como ddRAD-seq (double digest RAD-seq) (Peterson et al., 2012). Este protocolo involucra el uso de una segunda enzima de restricción y un proceso de selección precisa de fragmentos a secuenciar, como se ilustra en la Figura 1.3. Brevemente, la etapa de digestión en la técnica de ddRAD-seq se basa en el uso combinado de dos enzimas de restricción, una de sitio de corte frecuente y la otra de sitio de corte raro. Esto posibilita la captura de las mismas regiones en una población de individuos disminuyendo pérdida en la representación de los fragmentos. En particular, para el estudio de especies con genomas muy grandes y/o poco estudiados, el uso de enzimas de restricción de corte raro sensibles a metilación puede ayudar a crear un mayor nivel de reducción de la complejidad, dirigiendo la secuenciación a un menor número de sitios (Poland and Rife, 2012). En este protocolo, el paso de corte al azar es eliminado y se realiza una selección precisa de fragmentos, lo que reduce el número de duplicados en las regiones a secuenciar y genera profundidades de secuenciación balanceadas en todas las muestras.

1.2.3. El problema de los genotipos faltantes

El genotipado usando ddRAD-seq o GBS permite reducir los costos de la secuenciación por individuo, distribuyendo las lecturas generadas en centenas de individuos, con el fin de aumentar la potencia de las estimaciones estadísticas que se realizarán *a posteriori* (Elshire et al., 2011). La consecuencia de la reducción de costos se paga, en el caso de GBS, en un porcentaje elevado de datos perdidos. En el caso de ddRAD-seq, si bien se espera que éste no genere un porcentaje elevado de genotipos

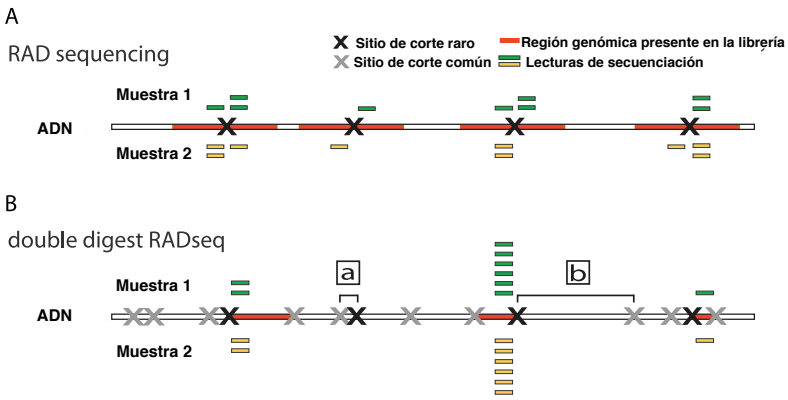

Figura 1.3: Ilustración comparativa de los protocolos RAD-seq y ddRAD-seq. Figura modificada de Peterson et al. (2012).

faltantes, se ha demostrado que éstos ocurren en grandes cantidades cuando la selección de los fragmentos de ADN a secuenciar se realiza en forma manual (Quail et al., 2012). En consecuencia, utilizando tanto GBS como ddRAD-seq, sólo un porcentaje bajo de SNPs resultan completamente representados a lo largo de todos los individuos, mientras que un gran número de SNPs sólo estarán representados en algunos pocos, generando un número importante de SNPs con genotipos faltantes (Li et al., 2011). Tales datos faltantes pueden ser evitados, por ejemplo, secuenciando a muy alta profundidad aunque ésto incrementa significativamente los costos. La alternativa económica es predecir los genotipos faltantes mediante técnicas de imputación (Poland and Rife, 2012).

1.3. Imputación de genotipos faltantes

1.3.1. Datos faltantes en matrices de genotipado

Un flujo de análisis bioinformático de datos obtenidos mediante GBS generalmente se inicia con una etapa de filtrado de lecturas de baja calidad. Posteriormente, se realiza la identificación de loci y alelos *de novo* o, en caso de contar con un genoma anotado, el alineamiento de las lecturas contra su referencia. Ésto permite luego poder descubrir y/o anotar los SNPs identificados e incluso asociarles una medida de escore. En general, las herramientas de procesamiento se suelen dividir en dos

grandes grupos: las que se *basan en una referencia* (por ej. Tassel, Glaubitz et al., 2014) y las *de novo* (ej. Stacks, Catchen et al., 2011) (Torkamaneh et al., 2016). Como resultado del procesamiento, se genera una matriz de genotipos, MG, de dimensión pxn, la cual tendrá en cada celda m_{ij} el genotipo del *i-ésimo* SNP detectado en el *j-ésimo* individuo bajo estudio. El número de SNPs, p, suele ser varios órdenes mayor que el número de individuos, n. Sumado a esto, es común encontrar una baja cantidad de SNPs, l, que estarán genotipados en los n individuos. Mientras tanto, los restantes $p - l$ SNPs tendrán un porcentaje variable de datos faltantes que podrán ser imputados para evitar la pérdida de información.

1.3.2. Imputación de genotipos

La imputación de alelos es una técnica estadística bien establecida que permite inferir los genotipos de los SNPs no observados (Chan et al., 2016). En el caso de individuos que poseen un genoma de referencia bien caracterizado, la tarea de ordenar los marcadores secuenciados e imputar los genotipos faltantes suele ser una tarea sencilla (Poland and Rife, 2012). Actualmente, existen varias herramientas de imputación de SNPs las cuales se han desarrollado y optimizado fundamentalmente para reconstruir haplotipos basados en un genoma de referencia bien establecido, como el de humano (Chan et al., 2016; Poland and Rife, 2012; Rutkoski et al., 2013). Cuando no se tiene una referencia sólida, los mapas genéticos y los paneles de referencia secuenciados a alta profundidad pueden ayudar al proceso de imputación. En su mayoría, estas herramientas se basan en el uso de modelos (por ejemplo, de Markov) que utilizan el conocimiento *a priori* contenido en los paneles de referencia y la información del desequilibrio de ligamiento (propiedad de algunos genes de no segregarse de forma independiente) para inferir simultáneamente haplotipos y genotipos de SNPs (Browning and Browning, 2016; Howie et al., 2009). En las primeras etapas de la aplicación de GBS, éstas fueron suficientes ya que fundamentalmente se focalizaba en el estudio de genomas bien conocidos.

La aplicación de los métodos de imputación es limitada o impracticable cuando se desea trabajar con organismos no modelo sin genoma de referencia o con genoma sin mapas de referencia, ya que no se tiene otra información *a priori* más que la secuenciación por GBS en sí misma (Chan et al., 2016; Halperin and Stephan,

2009). En este contexto, se han propuesto diversas estrategias sencillas basadas en metodologías bien conocidas. Entre ellas, se encuentra la imputación basada en el algoritmo K vecinos más cercanos (KNN) y en *Random Forests*, aunque sólo existen escasos reportes bibliográficos soportando su uso. En particular, Rutkoski et al. (2013) realizó una comparación del desempeño de tales estrategias al ser utilizadas en diferentes conjuntos de datos, sobre los cuales se controló el porcentaje de datos faltantes y donde las conclusiones a las que arribó sólo fueron soportadas por medidas de exactitud (accuracy). No obstante, aunque los métodos estadísticos propuestos son bien conocidos, aún queda poco claro cómo se debe seleccionar el conjunto de haplotipos/genotipos usados para la imputación de los datos faltantes de manera de maximizar las imputaciones correctas.

1.4. Algoritmos Estadísticos

1.4.1. Clasificadores

Sea $(\boldsymbol{X}, Y)$ un vector aleatorio donde $\boldsymbol{X}$ denota el vector de características e Y, la variable de salida o respuesta. Un *clasificador* es un algoritmo que tiene como fin predecir el valor de la variable salida frente a un valor particular, $\boldsymbol{x}$, del vector de características. En el ámbito de la clasificación, Y tomará valores en $\{C_1, C_2, ..., C_K\}$, con K número máximo de categorías posibles para Y. El vector $\boldsymbol{X}$ estará formado por un conjunto de P *características* o *features* que describirán en su conjunto una situación particular, es decir:

$$\boldsymbol{X} = (X_1, X_2, ..., X_P) \tag{1.1}$$

Las características son elementos importantes dado que representan los aspectos más significativos de un objeto. Posteriormente, tales aspectos permitirán reconocer objetos similares con facilidad.

En este trabajo, se consideraron clasificadores construidos desde un enfoque del *aprendizaje supervisado*, el cual consta de dos etapas. Asumiendo que

$$Y = f_\theta(\boldsymbol{X}) \tag{1.2}$$

para cierta función desconocida f_θ, el objetivo del aprendizaje es estimar en una primera etapa de *inferencia* los parámetros θ dado un conjunto de entrenamiento etiquetado. Posteriormente, en la etapa de *decisión*, será posible realizar predicciones para un $\boldsymbol{x}$ dado utilizando las estimaciones, según la Ecuación 1.3.

$$\hat{y} = f_{\hat{\theta}}(\boldsymbol{x}) \tag{1.3}$$

Un clasificador sencillo podría ser uno que asigne para $\boldsymbol{X}$ la clase C_i con probabilidad más alta. Para ello, deberá primero estimarse las probabilidades de las K clases dado $\boldsymbol{x}$, es decir $p(C_k|\boldsymbol{x})$, $k = 1, ..., K$. Un problema que surge a la hora de decidir clasificar un nuevo $\boldsymbol{x}$ en base a una probabilidad, es establecer qué criterio se utilizará para ello. En este contexto, la teoría de decisión provee herramientas para establecer los criterios óptimos.

1.4.2. Árboles de Decisión

Los *árboles de decisión* son algoritmos de aprendizaje supervisado sencillos, fácilmente interpretables que pueden ser utilizados tanto para problemas de regresión como de clasificación. En particular, los arboles de clasificación representan un conjunto de restricciones o condiciones que se organizan de manera jerárquica con el fin de tomar la decisión más acertada al asignar una categoría particular a Y ante un nuevo $\boldsymbol{x}$, para el cual Y se desconoce.

La etapa de *inferencia* de un árbol inicia con un conjunto de entrenamiento, conformado por n pares aleatorios $\{(\boldsymbol{X}_1, Y_1), (\boldsymbol{X}_2, Y_2), ..., (\boldsymbol{X}_n, Y_n)\}$ idénticamente distribuidos. Todo árbol comienza con un nodo raíz al que pertenecen los n pares de dicho conjunto. Este nodo se particiona binaria y recursivamente hasta que se alcanza un criterio preestablecido de parada. La última generación de nodos, se conoce como nodos terminales o nodos hojas. Cada nodo interno está etiquetado con una de las características de entrada. Las ramas que vienen de un nodo etiquetado reciben como nombres los valores posibles de dicha característica, mientras que cada hoja de un árbol es etiquetada con una de las C_k.

Un árbol puede ser linealizado en reglas de decisión, donde el resultado es el contenido del nodo hoja y las condiciones a lo largo del camino forman una conjunción

en la cláusula *if*. En general, las reglas tienen la forma: *if condición1 and condición2 and condición3 then resultado* (Breiman et al., 1984). En la Figura 1.4 se muestra un ejemplo sencillo donde se ha construido un árbol cuyo fin es decidir si se otorga o no un préstamo a un individuo. En este caso, se han seleccionado cinco características categóricas para su construcción.

Las reglas que generarán las particiones sucesivas, involucran las P características, de a una por vez. En el caso de que la característica X_j sea categórica (o discreta), la regla basada en X_j será del tipo $X_j \in B$ con $B \subset \{x_j^1, x_j^2, ..., x_j^L\}$. Si la característica X^j es del tipo continua, entonces las reglas serán del tipo $X^j \leq c$, con $c \in \Re$ y siendo c un punto intermedio entre un par de valores posibles de X^j. Cabe destacar que el conjunto de valores posibles de las características son aprendidos del conjunto de entrenamiento. Independientemente de la naturaleza de cada atributo, existe un número finito de reglas posibles a plantear para generar las particiones (Roche, 2009).

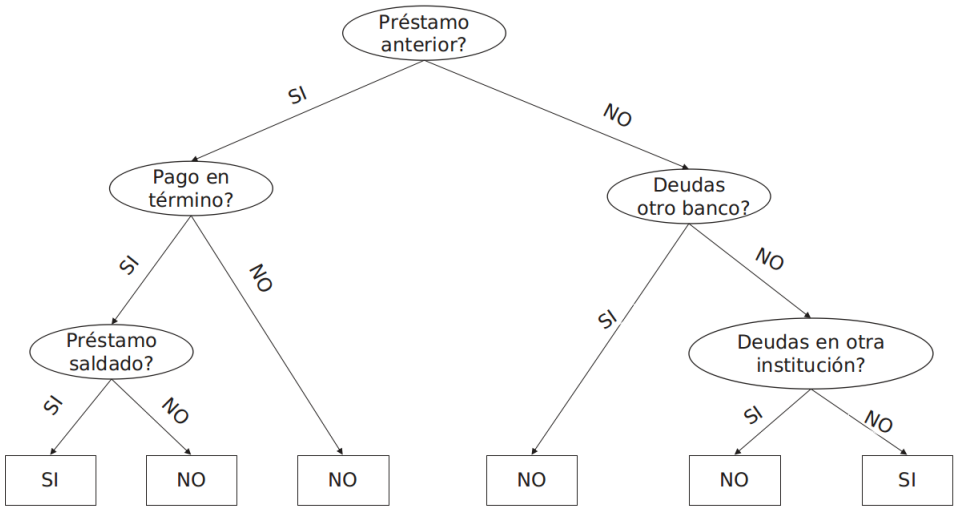

Figura 1.4: Ejemplo sencillo de un árbol de decisión.

La inducción del árbol a partir de un conjunto de entrenamiento requiere de una medida de evaluación de las características con el fin de generar particiones que maximicen la heterogeneidad entre clases y la homogeneidad o pureza de los nodos

(Breiman et al., 1984). Existen distintas medidas de impureza, $i(t)$, para la selección de atributos en el *t-ésimo* nodo. Una medida de impureza se basa en una *función de impureza*, ϕ, definida sobre el conjunto de probabilidades $p_k(t)$ con $k \in \{1, ..., K\}$, donde $p_k(t)$ es la probabilidad de que la clase asociada a un elemento sea C_k dado que pertenece al nodo t, definida según la Ecuación 1.4.

$$p_k(t) = p(Y = C_k|t) \tag{1.4}$$

Esta probabilidad se estima empíricamente a través de la proporción de elementos de la clase C_k en t. En particular, $\phi(t)$ se caracteriza por alcanzar un máximo cuando todas las posibles clases de Y están igualmente representadas en el nodo, mientras que es mínima cuando sólo una de ellas es asociada a dicho nodo. Como ϕ mide la impureza, estamos interesados en los nodos con impureza mínima, es decir, máxima pureza. Dada la *función de impureza*, es posible definir para cada nodo la *medida de impureza* mediante la Ecuación 1.5.

$$i(t) = \phi[p_1(t), ..., p_k(t)] \tag{1.5}$$

Entre las *medidas de impureza* más conocidas se destacan la *entropía* (Ecuación 1.6) y el *índice Gini* (Ecuación 1.7).

$$i_{ent}(t) = -\sum_{k=1}^{K} p_k(t) log_2 p_k(t) \tag{1.6}$$

$$i_{Gini}(t) = 1 - \sum_{k=1}^{K} p_k(t)^2 \tag{1.7}$$

Breiman afirma que la elección de la medida de impureza depende del problema en cuestión, aunque el predictor construido no parece ser muy sensible a dicha elección (Breiman et al., 1984). La medida utilizada deberá determinar el orden en que los atributos serán indagados y la regla que se utilizará para producir las particiones óptimas desde la raíz hacia los nodos hoja. Para ello, se debe determinar la bondad de cada partición generada, de manera de poder luego optar cuál partición es la óptima. Dada una partición s, que divide al nodo t en un nodo derecho, t_D, y un

nodo izquierdo, t_I, y p_D y p_I las proporciones de elementos del nodo t que caen en sus particiones, t_D y t_I, respectivamente, una medida de su bondad se obtiene midiendo la disminución en la impureza debida a la partición, es decir:

$$\Delta i(s,t) = i(t) - p_D(s,t)i(t_D), -p_I(s,t)i(t_I) \tag{1.8}$$

Lógicamente, la partición óptima será aquella que posea la mayor bondad sobre el conjunto de particiones posibles. Una vez que se ha llegado a un nodo hoja, se le debe asignar una de las K clases, lo cual se hace mediante el sistema de *voto mayoritario* y en caso de empate, se sortea entre las clases más frecuentes. Este proceso de construcción genera un árbol *maximal*, el cual puede ser posteriormente *podado* mediante la eliminación de ramas que produzcan escasa reducción de la impureza. La poda permite disminuir la complejidad del árbol, evitar el sobre-ajuste y aumentar la capacidad de generalización (Roche, 2009).

Si bien los árboles de decisión son algoritmos sencillos, aplicables en diversos campos, éstos suelen tener elevadas tasas de error ya que son sensibles a datos ruidosos, por lo que se los llama clasificadores *débiles*.

1.4.3. Random Forests

Random Forests o *bosques aleatorios* es la marca de un algoritmo que utiliza un conjunto de clasificadores *débiles* para construir un único clasificador *fuerte*. Es una herramienta del aprendizaje maquinal indicada para problemas donde el número de características es muy superior al número de datos del conjunto entrenamiento. El algoritmo *Random Forests* ha mostrado ser más robusto y exacto que los clasificadores simples basados en árboles de decisión. Consiste en una colección de clasificadores con estructura de árbol, $\{h(\boldsymbol{X}, \boldsymbol{\Theta}_r), l = 1, ..., R\}$, donde $h(\boldsymbol{X}, \boldsymbol{\Theta}_r) = h_r(\boldsymbol{X})$ es un clasificador que recibe como entrada el vector $\boldsymbol{X}$, $\boldsymbol{\Theta}_r$ es el vector de parámetros de construcción del r-ésimo árbol y $\{\boldsymbol{\Theta}_r\}$ son vectores aleatorios independientes e idénticamente distribuidos. De esta manera, ante un vector de entrada cada árbol emite un voto para la elección de la clase más popular de la variable respuesta que se debe asignar a dicha entrada (Breiman, 2001).

La unidad básica del *Random Forests* es un árbol de decisión aleatorio. La di-

ferencia fundamental que existe entre *Random Forests* y los árboles de clasificación es que el primero se construye mediante un proceso no determinístico usando un procedimiento de aleatorización en dos etapas. Por un lado, el algoritmo de entrenamiento para los árboles aleatorios del bosque aplica una técnica llamada *bagging* o *bootstrap aggregating*. Ésta se encarga de que cada árbol del bosque se construya con una muestra bootstrap del conjunto total de valores de entrada. La segunda etapa de aleatorización se establece durante el crecimiento de cada árbol, donde la partición de sus nodos estará determinada en una forma no determinista. Para ello, en vez de utilizar todas las características, se selecciona aleatoriamente un conjunto de ellas sobre las cuales se elegirá la que determine la mejor partición para dicho nodo.

Este proceso de aleatorización en dos etapas tiene como objetivo reducir la correlación entre los árboles del bosque, de manera que el conjunto tenga la mayor exactitud posible. Para ilustrar este proceso, supongamos un conjunto de entrenamiento $L = \{(\boldsymbol{X}_n, Y_n), n = 1, ...N\}$ y un clasificador $h_r(\boldsymbol{X}, L)$ tal que $P(h_r(\boldsymbol{X}, L) = Y) = p$. Supongamos además que se tiene una secuencia de conjuntos de aprendizaje, L_j, donde cada uno consiste en un conjunto de N observaciones independientes e idénticamente distribuidas a L. El objetivo es que el predictor que se obtenga usando L_j supere al árbol de decisión que se genera al utilizar L, es decir que $P(h_r(\boldsymbol{X}, L_j) = Y) = p_j < p$. Breiman estableció que un método para agregar los J predictores generados con L_j es a través del voto, es decir el valor de Y que se asigne a $\boldsymbol{X}$ será el más frecuente entre los predichos por los J clasificadores. Notablemente, el primer inconveniente que se encontrará a la hora de poner en práctica el procedimiento descrito es la disponibilidad de J conjuntos de entrenamiento independientes e idénticamente distribuidos a L. Con el fin de solventar este inconveniente, Breiman sugiere que es posible imitar el procedimiento anteriormente explicado generando muestras bootstrap repetidas, $\{L^B\}$ a partir de L. Formalmente, las muestras bootstrap, $\{L^B\}$, de L se obtienen tomando uniformemente b muestras con reemplazo de L. Es notable que cada $L_i \in \{L^B\}$ puede contener pares $(\boldsymbol{X}_n, Y_n)$ repetidos entre sus elementos. Así funciona el algoritmo de *bagging* previamente mencionado, el cual permite generar el conjunto de B predictores, $\{h(\boldsymbol{X}, L^B)\}$, que votará para obtener $h_B(\boldsymbol{X}, L)$. El valor de B es un parámetro libre, el cual se suele tomar en el orden de los cientos o miles de árboles, dependiendo de la naturaleza del problema y del tamaño del conjunto

de entrenamiento. Claramente, en cada muestra bootstrap quedará afuera un grupo de datos que no se utilizarán para ajustar el correspondiente árbol. Estas muestras conforman otro conjunto de datos llamado *out of bag* (OOB) o de validación, el cual se utilizará para estimar el error de clasificación del *Random Forests*. Para ello, cada dato del conjunto OOB será clasificado por aquellos árboles en los que no ha sido utilizado para su construcción. Luego se estimará la proporción de mala clasificación sobre todas las clasificaciones y se promediarán para obtener una estimación del error de OOB que caracterizará al *Random Forests* (Carranza Astrada, 2015; Chen and Ishwaran, 2012; Qi, 2012)

El procedimiento explicado en el párrafo anterior es efectivo en la generación de B conjuntos de entrenamiento, pero puede derivar en predictores altamente correlacionados. Para evitar este inconveniente es que el algoritmo *Random Forests* incorpora la segunda etapa de aleatorización. Luego, el *Random Forests* se construye de la siguiente manera:

- Muestrear B veces b casos de L con reemplazo para formar el conjunto de muestras entrenamiento bootstrap $\{L^B\}$.
- Para cada conjunto de entrenamiento $L_i \in \{L^B\}$ hacer crecer un árbol aleatorio sin poda. Para ello, si M es el número de características de cada vector de entrada, en cada nodo se debe seleccionar aleatoriamente m características tal que $m << M$, entre las cuales se determinará cual genera su mejor partición.
- Agregar la información de los B árboles siguiendo la regla de voto mayoritario.
- Obtener una estimación del error de clasificación calculando el error OOB.

1.5. Selección de Características

Si bien *Random Forests* es una técnica adecuada para los problemas donde el número de casos observados es bastante menor que el número de características, a medida que aumenta el número de características se incrementarán cada vez más los tiempos de confección y la complejidad de los bosques aleatorios. Por lo tanto, la reducción del número de características, previa a la confección del bosque, suele ser

una tarea necesaria. Este proceso que se conoce como *selección de características*, consiste en elegir un subconjunto de las disponibles basándose en una métrica o criterio pre-definido. Así, sólo este subconjunto de características será posteriormente utilizado para confeccionar el *Random Forests*. Los objetivos de esta selección son: **(a)** evitar el sobre-ajuste y mejorar el desempeño del modelo, **(b)** proveer modelos más rápidos y efectivos y **(c)** obtener una visión más profunda de los procesos subyacentes que generaron los datos (Saeys et al., 2007).

Existen numerosas técnicas de selección de características, aunque todas ellas realizan la búsqueda del conjunto óptimo en base a una función *objetivo*. En el contexto de los problemas de clasificación, las diferentes técnicas se agrupan en métodos wrappers y de filtrado. Los métodos wrappers emplean el propio conjunto de reglas generadas por el algoritmo de aprendizaje, que luego se usará en la clasificación, como función *objetivo*; por otro lado, los métodos de filtrado utilizan una función independiente del algoritmo de aprendizaje basado en criterios como distancia o dependencia. Los métodos de filtrado evalúan la relevancia de los atributos sólo en base a características intrínsecas de los datos. Adicionalmente, son métodos sencillos en términos computacionales, rápidos e independientes del algoritmo que se utilizará posteriormente para la clasificación. Esto último resulta atractivo ya que si se desean comparar distintos métodos, no es necesario realizar la selección de características para cada método, como sí lo sería si se emplean los métodos wrappers. Notablemente, la desventaja que presentan las técnicas de filtrado es que no miden la relación entre los datos y el algoritmo de clasificación. Además, si bien la metodología *Random Forests* propone la selección de características categóricas en base al indice Gini, su cálculo es más complejo que una simple medida de distancia o correlación entre variables ya que implica el ajuste del bosque para su evaluación.

Las técnicas de filtrado más conocidas se basan en la ganancia de información, la distancia euclídea y el test de independencia χ^2, los cuales generan una suerte de ranking de los mismos.

1.5.1. Test de independencia

El test de independencia χ^2 se utiliza en estadística para evaluar la independencia que existe entre dos variables categóricas. Supongamos que de n elementos de una

población se han observado dos características X e Y, obteniéndose una muestra aleatoria simple bidimensional $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$. En base a dichas observaciones se desea evaluar si las características X e Y son independientes o no. Supongamos además que ambas variables son del tipo categórico siendo $\{x^1, x^2, ..., x^r\}$ y $\{y^1, y^2, ..., y^s\}$ los posibles valores de X e Y, respectivamente. De esta manera y con los datos observados en la muestra es posible construir una tabla de contingencia de dimensiones $r \times p$ de la forma de la Tabla 1.1, donde n_{ij} representa la cantidad de veces que se observó la combinación (x^i, y^j) en la muestra. Además, $n_{i.}$ indica la cantidad de veces que se observó el valor x^i para X y $n_{.j}$ la cantidad de veces que se observó el valor y^j para Y, siendo $n_{i.}$ y $n_{.j}$ los totales marginales.

Tabla 1.1: Ejemplo de tabla de contingencia. Caso de un conjunto de n datos donde se registraron dos variables categóricas X e Y.

	y^1	y^2	...	y^s	**Total**
x^1	n_{11}	n_{12}		n_{1s}	$n_{1.}$
x^2	n_{21}	n_{22}		n_{2s}	$n_{2.}$
...					
x^r	n_{r1}	n_{r2}		n_{rs}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$		$n_{.s}$	n

Con los datos arreglados de esta manera es posible determinar la independencia entre X e Y mediante la comparación de las frecuencias observadas con las frecuencias esperadas bajo el supuesto de independencia entre variables. En este contexto, las frecuencias observadas están dadas por los n_{ij} y las frecuencias esperadas, e_{ij}, están dadas por el producto de las probabilidades marginales de x^i e y^j, que se estiman a partir de las frecuencias marginales ($x_{.i}$ y $y_{.j}$) divididas por n, como se indica en la Ecuación 1.9.

$$e_{ij} = \frac{n_{i.} n_{j.}}{n} \tag{1.9}$$

Luego, el estadístico del test de independencia se puede obtener según la Ecuación 1.10, y tiene distribución χ^2 con $(r-1)(s-1)$ grados de libertad (Monge Ivars and Perez, 2017).

$$\sum_{i_1}^{r} \sum_{j_1}^{p} \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \tag{1.10}$$

Si el valor p asociado al estadístico resulta menor a un nivel de significancia, α, entonces hay suficiente evidencia para rechazar la hipótesis nula de independencia. De esta manera, si X es uno de los atributos e Y es el vector de respuesta, es posible determinar si existe o no independencia entre dichas variables de manera de descartar aquellas X que sean independientes de Y y así seleccionar solo los atributos correlacionados con la variable en cuestión.

1.6. Objetivos

1.6.1. Objetivo General

Desarrollar procedimientos estadísticos que permitan imputar los valores faltantes en datos de genotipado por secuenciación de organismos no modelo.

1.6.2. Objetivos Específicos

- Diseñar una metodología estadística de imputación de SNPs para organismos no modelo.
- Evaluar el desempeño de la metodología propuesta en base a simulaciones de genotipos faltantes, utilizando matrices completas de datos reales.
- Comparar el desempeño del algoritmo propuesto con metodologías alternativas disponibles.
- Aplicar el algoritmo desarrollado en una matriz de datos genotípicos incompleta obtenida a partir de genotipifación por secuenciación para girasol.

1.7. Hipótesis

En este trabajo se hipotetiza que si se utiliza únicamente la información generada mediante estrategias de genotipificación a escala de genoma completo por tecnologías de secuenciación masiva es posible imputar los valores faltantes de SNPs. En particular, se considera que la clave del análisis se encuentra en los alelos de SNPs que

se han secuenciado y en la dependencia que existe entre los SNPs que pertenecen al mismo grupo de ligamiento.

Capítulo 2

Materiales y Métodos

2.1. Datos

El desarrollo de esta tesis ha sido motivado por la problemática detectada ante la necesidad de analizar datos de genotipado por secuenciación. En particular, los datos con los que se trabajó contienen un alto porcentaje de genotipos faltantes, los cuales debieron ser imputados para su posterior análisis.

2.1.1. Descripción

La base de datos analizada ha sido generada por el genotipado mediante secuenciación de 135 líneas endocriadas de girasol (*Helianthus annuus*). El girasol es una especie diploide (2n = 2x = 34), con un genoma haploide medio de 3.500 Mpb. Las líneas endocriadas o endogámicas son individuos de una especie que son casi idénticos entre sí en su genotipo, lo cual resulta en un individuo prácticamente homocigota.

La genotipificación se realizó mediante un protocolo ddRADseq, donde se utilizaron dos enzimas de restricción: *Sphl* y *EcoRI*. En particular, ambas enzimas tienen un sitio de reconocimiento de 6 pb (GCATGC y GAATTC, respectivamente) y *EcoRI* es una enzima de corte raro sensible a metilación. Luego de la digestión y posterior ligación de adaptadores y etiquetas moleculares para identificar cada muestra, se hizo una selección manual en gel de agarosa de aquellos fragmentos dentro del rango 34-550 pb, correspondientes a fragmentos originales de 260-470 pb. La secuenciación

propiamente dicha de los fragmentos obtenidos se realizó con un equipo Illumina HiSeq 2500, con lecturas paired-end (2x125 pb) generando un total aproximado de 126.325.000 lecturas. Si bien las líneas endocriadas consideradas para el estudio fueron 135, la secuenciación se realizó sobre 140 muestras. Esto se debe a que cinco de las líneas fueron genotipadas en dos oportunidades ya que la cobertura obtenida en una primera secuenciación fue muy baja. El protocolo de ddRADseq y la secuenciación se hicieron en el *Istituto di genomica applicata* (IGA; Udine, Italia).

2.1.2. Pre-procesamiento

Los datos genotípicos han sido generados mediante un pre-procesamiento bioinformático de los datos de secuenciación. En una primer instancia, las lecturas se cortaron a 110 pb, para evitar caída de calidad en las bases del extremo 3', y se eliminaron aquellas lecturas de baja calidad, con errores en la secuencia de reconocimiento de la enzima, y/o con errores en la etiqueta molecular. En total, se obtuvo un promedio de 936.000 lecturas por línea endocriada (muestra). Cada muestra fue luego alineada contra el **genoma de referencia** utilizando el programa `Bowtie2`, con al menos un 95 % de las lecturas mapeando al menos una vez contra la referencia. La posterior identificación de SNPs se realizó con el software `Stacks` (Catchen et al., 2011), usando como referencia la versión del genoma de girasol más actual (Badouin et al., 2017). Brevemente, los pasos que se realizaron para construir la matriz de genotipos fueron:

- `pstacks`: Agrupamiento de las lecturas de cada línea endocriada en loci para formar las pilas y así poder identificar los sitios polimórficos.
- `cstacks`: Agrupamiento de los loci de todas las líneas endocriadas para armar el catálogo de SNPs.
- `sstacks`: Comparación de los loci de cada muestra contra el catálogo para determinar el estado alélico de cada locus en cada individuo.
- `rxstacks`: Corrección de los genotipos y haplotipos en cada muestra.
- Nuevamente `cstacks`, para armar un nuevo catálogo con los marcadores que superaron las correcciones.

Como procesamiento adicional y utilizando la misma herramienta, se eliminaron aquellos SNPs que presentaron frecuencia alternativa mínima (MAF) menor a 0,05 y genotipos faltantes en más del 80 % (112 o más) de las muestras en estudio. Una vez finalizado este procesamiento, se obtuvo una colección de datos contenida en un archivo VCF. Este archivo es comúnmente utilizado para almacenar los datos de variantes genotípicas y está conformado por tres secciones. La primer sección es la de *metadata*, la cual se caracteriza por contener cada una de sus líneas iniciadas con ##. La segunda sección es el *encabezado*, el cual contiene al menos ocho campos fundamentales:

- CHROM, cromosoma: Identificador tomado del genoma de referencia.
- POS, posición: Posición de referencia de la variante en el cromosoma.
- ID, identificador: Nombre del fragmento donde se determinó la variante.
- REF, base/s de referencia: Base nucleotídica o haplotipo encontrado en la referencia.
- ALT, alternativo: Lista de los alelos/haplotipos alternativos separados por coma
- QUAL, calidad: Valor de calidad en escala Phred calculado según la Ecuación 2.1, donde P indica la probabilidad de que el alelo que se haya identificado como alternativo sea falso. Los valores altos, por lo general superiores a 20, corresponden a alelos confiables.

$$Q = -10log_{10}P \tag{2.1}$$

- FILTER, filtro: Indica si la posición correspondiente ha pasado los distintos filtros.
- INFO, información adicional: Contiene una serie de campos separados por ";", entre los cuales se destacan GT y DP que refieren al genotipo identificado en la muestra y la cantidad de lecturas que lo soportan, respectivamente.

En el caso de los datos analizados en esta tesis, el archivo VCF contenía luego de estos ocho campos los nombres de las muestras genotipadas. La tercer sección está organizada en columnas, y contiene la información correspondiente a cada uno de los campos especificados en la sección anterior. A partir de este archivo se obtuvieron dos matrices de datos. La primera de ellas, con información referente a cada SNP identificado, la cual es común para todas las muestras. Esta matriz comprendió las primeras ocho columnas de la sección de información del archivo VCF. La segunda, es la matriz de genotipos propiamente dichos (ver Sección 1.3.1). Claramente, los datos faltantes se encontraron en esta segunda matriz de datos, la cual resultó de dimensiones $p \times n$ donde p =34.730 es el número de SNPs identificados y $n = 140$ es el total de muestras obtenidas para líneas endocriadas genotipificadas.

2.2. Propuesta

2.2.1. Estrategia de imputación

En el contexto de esta tesis, el objetivo es la imputación de genotipos faltantes en cada uno de los SNPs que han sido identificados en algunas de las muestras de un grupo de individuos de una población. Para ello, la única información disponible con la que se contó es la matriz de genotipos incompleta, obtenida mediante genotipado por secuenciación, y el genoma de referencia, el cual permitió establecer una relación entre SNPs mediante su posición.

Dada las características del problema abordado, donde el número de casos (líneas endocriadas, 135) es bastante menor que el número de características (SNPs, $>$ 30.000), se propuso la implementación de una estrategia de imputación basada en el algoritmo de clasificación *Random Forests*. Según lo descrito en la Sección 1.4.3, para ajustar dicho algoritmo y utilizarlo luego como predictor, se necesita un conjunto de entrenamiento que conste de N casos en los que se han observado la variable de interés y P características. Una vez ajustado, el *Random Forests* puede ser utilizado para generar valores de la variable de interés ante nuevos valores de las características. Adicionalmente, dado que al ajustar un *Random Forests* se obtiene una estimación del error de clasificación, mediante el error OOB, es posible considerar esta estimación para determinar si es conveniente o no imputar los valores con el clasificador.

La propuesta consideró que los SNPs identificados en el estudio pueden ser clasificados en dos grupos, los SNPs completos, es decir los genotipados en todos los individuos, y los SNPs incompletos o con genotipos faltantes. Cada SNP se asume como una variable de interés categórica con, a lo sumo, tres posibles valores: *homocigota de referencia* "0/'0'), *heterocigota* ("0/1" o "1/0") y *homocigota alternativo* ("1/1"). Cada uno de los SNPs incompletos se considera una variable respuesta o de salida. Por otro lado, los SNPs completos se asumieron atributos o características conteniendo información útil para la predicción de los genotipos faltantes de los SNPs incompletos. Luego, el conjunto de datos definido por los genotipos conocidos de un SNP incompleto y los genotipos de todos los SNPs completos constituyen un conjunto de entrenamiento para ajustar un *Random Forests*. Éste, una vez ajustado, puede ser utilizado para predecir los genotipos faltantes del SNP incompleto. Es decir, que el problema de imputación de genotipos faltantes se convirtió en un problema de clasificación de una variable multiclase, seguido de la predicción. De esta manera, es posible inferir los genotipos faltantes de matrices incompletas transformándolas en matrices de genotipos completas.

Adicionalmente, en este trabajo se consideró que el número de atributos, SNPs completos, puede ser reducido teniendo en cuenta que algunos SNPs se segregan en conjunto, es decir están correlacionados. Más aún, la dependencia entre SNPs es mucho más fuerte mientras más próximos están unos de otro y compartan grupo de ligamiento. Dado que los SNPs son variables categóricas, es posible utilizar el Test de Independencia χ^2 para determinar la correlación entre SNPs. Considerando lo expuesto, en este trabajo se desarrolló una estrategia basada en *Random Forests* y dependencia entre SNPs con el fin de imputar los genotipos faltantes de aquellos SNPs que no lograron secuenciarse en todos los individuos bajo estudio.

2.2.2. Evaluación mediante simulación

Base de datos

La evaluación del desempeño del algoritmo de imputación planteado en la sección anterior se ha realizado mediante su aplicación en una base de datos genotípicos completa a la cual se le simularon datos faltantes. Dicha base contiene datos de

SNPs de peces de la especie *Gasterosteus aculeatus* (espinoso) (Catchen et al., 2013). En particular, este conjunto de datos se aplicó al entrenamiento del programa de anotación de variantes Stacks, que es el mismo software que se utilizó para procesar los datos de genotipado en girasol. En particular, para este trabajo se seleccionaron los datos de una de las nueve poblaciones de espinoso bajo estudio, que incluye alrededor de 50.000 SNPs genotipados en 105 individuos.

La simulación de los datos faltantes se realizó en dos etapas. En la primera, se construyó una base de datos que permitiese determinar el desempeño del algoritmo bajo resultados análogos a los de la base de datos de secuenciación de girasol. Para ello se analizó la matriz de datos genotípicos de girasol para obtener el número de SNP genotipados en todos los individuos, y el número y la distribución de datos faltantes. La segunda etapa de simulación se realizó con el fin de caracterizar el comportamiento del método ante variaciones de escenarios, caracterizados por cambios en dichos porcentajes.

Medidas de desempeño

En el contexto de la imputación, los genotipos reales que fueron sustraídos de la matriz para simular los datos faltantes, se utilizaron posteriormente para determinar el desempeño de la estrategia propuesta mediante medidas objetivas obtenidas a partir de la *matriz de confusión*. Esta matriz resume los resultados de una tarea de clasificación sobre un conjunto de N datos. Dada la variable respuesta Y, cuyo conjunto de valores posibles es $\{C_1, C_2, ..., C_K\}$, es posible construir una matriz de confusión para cada una de las clases C_i, como la ilustrada en la La Tabla 2.1.

Tabla 2.1: Matriz de confusión que resume los resultados de una tarea de clasificación de N datos, para la clase C_i perteneciente al conjunto de valores posibles, $\{C_1, C_2, ..., C_K\}$, de la variable respuesta, Y.

Clase de Y	Clasificado como C_i	Clasificado como C_j, $j \neq i$
C_i	VP	FN
C_j, $j \neq i$	FP	VN

En dicha matriz se resume el agrupamiento de los resultados de la clasificación de N datos en función del valor que el algoritmo de clasificación asignó a Y y el valor

original de dicha variable para cada uno de los N datos. Luego, dado el *n-ésimo* dato ($n \in [1, N]$) si el valor original de Y^n era C_i y el clasificador le asignó el mismo valor, entonces este caso es contado como un *verdadero positivo* (VP), mientras que si el clasificador le asignó una clase C_j distinta de C_i, entonces éste es un *falso negativo* (FN). Por otro lado, si el valor original de Y^n era C_j, C_j distinta de C_i, y el clasificador le asignó la clase C_i, entonces éste es un caso de *falso positivo* (FP), finalmente si el clasificador le asignó una clase C_j distinta de C_i, entonces se considera éste un *verdadero negativo* (VN).

En función del conjunto de las K matrices de confusión es posible definir un conjunto de medidas de desempeño para caracterizar un clasificador según detalla la Tabla 2.2 (Sokolova and Lapalme, 2009). Estas medidas responden a dos formas de evaluación, por un lado es posible definir una medida como el promedio de dicha medida calculada para cada una de las K clasificaciones (*macro*-promedios), por el otro, es posible obtener los valores acumulados de VP, VN, FP y FN y luego calcular la medida sobre estos valores (*micro*-promedios). La diferencia general entre estas medidas es que las de *macro*-evaluación consideran que todas las clases son igualmente probables, mientras que las *micro*-medidas favorecen a las más abundantes. Luego, si se tienen clases menos frecuentes, las *micro*-medidas prácticamente las ignoran.

En el contexto de este trabajo, el uso de las *micro*-medidas en cierta manera oculta la esencia del análisis de SNPs, cuyo objetivo es precisamente conocer el efecto de las pequeñas alteraciones en las frecuencias alélicas. Adicionalmente, las medidas *micro-precisión*, *micro-sensibilidad* y *micro-F-score* resultan iguales entre sí, ya que el número de clases coincide con el número de clasificaciones. Además, dado que en general sólo son dos los genotipos presentes en la población bajo estudio (diploide homocigotas), estas medidas serán prácticamente iguales a la *exactitud promedio*. Por lo tanto, en esta tesis se optó por estudiar las siguientes medidas como indicadoras del desempeño de un método: *exactitud promedio*, *macro-precisión*, *macro-sensibilidad* y *macro-F-score*. Con el objetivo de simplificar la notación, las macro medidas son simplemente referidas como exactitud, precisión, sensibilidad y F-score.

Las medidas de desempeño descritas anteriormente se utilizaron para determinar objetivamente cuál/cuáles de las diferentes alternativas de la estrategia de imputación propuesta en la Sección 2.2.1 fueron superiores con respecto al resto.

Tabla 2.2: Medidas de desempeño de un clasificador obtenidas a partir de la clasificación de N datos de evaluación siendo que la variable respuesta es del tipo categórica con K posibles valores.

Medida	**Fórmula**	**Foco de evaluación**
Exactitud promedio	$\frac{1}{K}\sum_{i=1}^{K}\frac{VP_i+VN_i}{N}$	Promedio de efectividad por clase del clasificador
Micro-precisión	$\frac{\sum_{i=1}^{K} VP_i}{\sum_{i=1}^{K} VP_i+FP_i}$	Acuerdo entre la clase verdadera del dato y la asignada por el clasificador calculado sobre las sumas de las clasificaciones
Micro-sensibilidad	$\frac{\sum_{i=1}^{K} VP_i}{\sum_{i=1}^{K} VP_i+FN_i}$	Eficacia del clasificador para identificar los clases verdaderas de los datos calculada a partir de sumas de las clasificaciones
Micro-F-score	$\frac{2Micro-pre*Micro-rec}{Micro-pre+Micro-rec}$	Balance entre micro-precisión y micro-recall. Resume la eficacia de un clasificador
Macro-precisión	$\frac{1}{K}\sum_{i=1}^{K}\frac{VP_i}{VP_i+FP_i}$	Promedio del acuerdo entre la clase verdadera del dato etas de datos con los de un clasificador
Macro-sensibilidad	$\frac{1}{K}\sum_{i=1}^{K}\frac{VP_i}{VP_i+FN_i}$	Promedio de la eficacia por clase de un clasificador para identificar correctamente las clases
Macro-sensibilidad	$\frac{1}{K}\sum_{i=1}^{K}\frac{VP_i}{VP_i+FN_i}$	Promedio de la eficacia por clase de un clasificador para identificar correctamente las clases
Macro-F-score	$\frac{2Macro-pre*Macro-rec}{Macro-pre+Macro-rec}$	Balance entre macro-precisión y macro-recall. Resume la eficacia de un clasificador

Comparación con herramientas existentes

Los genotipos incompletos de las matrices de datos simuladas fueron también imputados con algunas de las herramientas disponibles para tal fin en especies no modelo. Posteriormente se compararon los resultados obtenidos con dichas herramientas con los de las mejores alternativas de la metodología propuesta. Como se ha mencionado, no se cuenta con demasiadas herramientas para imputación en organismos no modelos. En este caso se optó por contrastar la estrategia de imputación propuesta

contra las basadas en las estrategias *imputación por la moda*, *Beagle*(Browning and Browning, 2016) y *LinkImputeR* (Money et al., 2017). La primera de ella es la opción más sencilla y la que muchas veces se utiliza ya que sólo requiere identificar el genotipo más frecuente para cada SNP y asignarlo a las muestras donde no ha sido identificado. La herramienta *Beagle* es un algoritmo inicialmente desarrollado para la imputación guiada por paneles de referencia que ha sido ampliamente utilizado. Sin embargo, en sus últimas versiones han sido mejoradas para permitir la imputación sin esta referencia. Es un algoritmo iterativo basado en la técnica de modelos ocultos de Markov. El algoritmo alterna entre la construcción de modelos y el muestreo, utilizando la estrategia de maximización de la esperanza para converger hacia las soluciones más probables. Por último, *LinkImputeR* es un software recientemente desarrollado que se basa en la técnica de K vecinos más cercanos y considera el desequilibrio de ligamiento a la hora de elegir los vecinos más próximos.

2.2.3. Aplicación a datos reales

La matriz de genotipos reales de las líneas endocriadas de girasol fue analizada con la alternativa de imputación propuesta que mejor desempeño evidenció. Se exploraron valores tales como el porcentaje de datos imputados y números de SNPs correlacionados con el fin de comparar con los resultados obtenidos para los conjuntos de datos simulados. Finalmente, se determinó el porcentaje de información que logró recuperar el proceso de imputación.

Capítulo 3

Resultados

3.1. Base de datos de girasol

3.1.1. Preparación

La base de datos de genotipado en girasol inicialmente contó con 34.730 SNPs genotipados en 140 muestras provenientes de 135 líneas endocriadas. Previa a su utilización se realizó un análisis exploratorio con el fin de determinar su consistencia e integridad. A continuación se describen los aspectos inspeccionados y el resultado obtenido.

- Muestras con baja cobertura: Aunque las duplicaciones de cinco de las líneas endocriadas fueron incluidas al hacer la anotación de los SNPs con Stacks, éstas fueron eliminadas de la matriz de genotipos aquí estudiada ya que representan información redundante y sólo aportan muy pocos genotipos. Luego de la eliminación las cinco muestras, se calculó nuevamente los valores de frecuencia alélica, para lo cual se implementó la función `calculateAFALT`. Esta función calcula sólo la frecuencia del alelo alternativo. Luego, la frecuencia del alelo de referencia se obtuvo restando a 1 dicha cantidad. Adicionalmente, el número de líneas genotipadas en cada SNP también debió ser re-calculado, para lo que se diseñó la función `getNS`.

- SNPs duplicados: Se determinó que 300 de los SNPs genotipados eran duplicados. Esto significa que la base de datos contenía un mismo SNP (cromosoma +

posición) identificado con dos nombres diferentes. Específicamente, se encontró dos tipos de duplicados: aquellos que comparten alelo de referencia y aquellos que tienen cruzados los alelos de referencia y alternativo. Con el fin de unificar la información referente a los genotipos de los SNPs duplicados se implementó la función `mergeSNPs`. Ésta combina en primer instancia los SNPs duplicados y posteriormente corrige las estimaciones de las frecuencias alélicas utilizando la función `calculateAFALT`. Luego de ejecutar esta función, la cantidad de SNPs de la base de datos se redujo a 37.430.

- SNPs con alelos invertidos: Siguiendo el criterio utilizado por Stacks, de definir como alelo de referencia al alelo más frecuente, se controló que ésto se verifique en cada SNP. Con el fin de corregir los casos donde no se verificaba, se implementó la función `invertGT`. Ésta se encargó de invertir la definición de los alelos de referencia y alternativo así como también de corregir la especificación de los genotipos en cada muestra.

- Filtrado de SNPs: Si bien al ejecutar el programa Stacks, se fijó la opción de filtrado de SNPs con MAF menor a 0,05, como los pasos descritos anteriormente implicaron cálculos que modificaron las frecuencias alélicas, fue necesario controlar que se siguiera cumpliendo dicho criterio. Otro aspecto que se controló fue que los SNPs contuviesen datos en al menos el 20 % de las muestras consideradas, es decir 27 líneas endocriadas. En este paso no se encontraron SNPs que no cumpliesen con estos dos requisitos.

- Filtrado de regiones con abundantes SNPs: Cada SNP tiene asociado el nombre del fragmento en el que ha sido detectado. Estos fragmentos tienen longitudes de unos 150-500 pb. Adicionalmente, se ha estudiado previamente que, en promedio, el genoma de girasol presenta un SNP cada 143pb (Pegadaraju et al., 2013). Por lo tanto, en este trabajo se consideró como atípico que un fragmento tenga más de 4 SNPs. Un total de 1.293 fragmentos superaron este valor y fueron removidos. En consecuencia, la base de datos se redujo a 26.712 SNPs.

3.1.2. Exploración

Como resultado de la etapa de preparación del conjunto de datos, se obtuvo una matriz de genotipos correspondientes a 26.712 SNPs. La exploración de esta matriz confirmó que la MAF es mayor o igual a 0,05 (Figura 3.1). De todos los SNPs, se encontró que sólo 1.700 (6,4 %) fueron genotipados en las 135 líneas endocriadas, entretanto, el resto fueron identificados en al menos 27 de ellas.

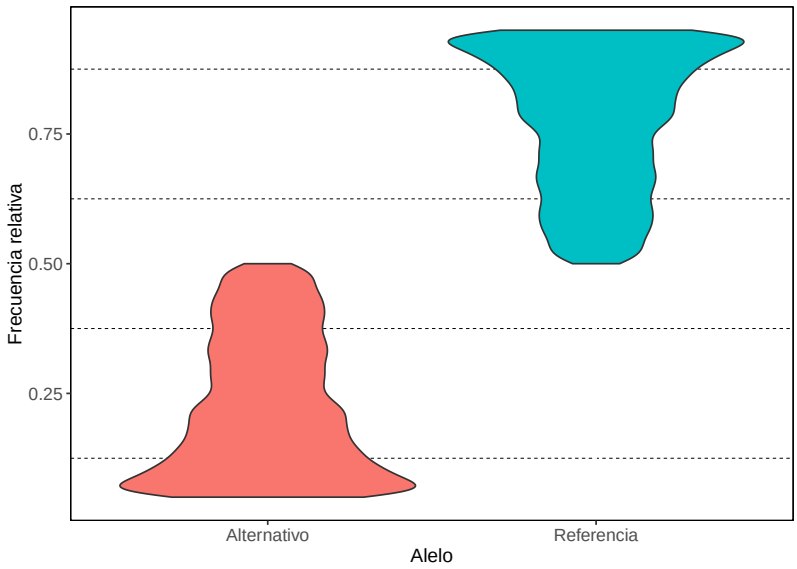

Figura 3.1: Diagramas de violín de las frecuencias alélicas observadas en la matriz de genotipos de girasol.

Las frecuencias observadas de los porcentajes de genotipos faltantes en la matriz de genotipos incompletos se ilustran en la Figura 3.2. Específicamente, se encontró que la mayor cantidad de SNPs se correspondieron con porcentajes de genotipos faltantes entre 0 y 10 % y entre 70 % y 80 %.

La distribución de los SNPs en los cromosomas del girasol también se exploró. Para ello, los SNPs fueron agrupados según el número de éstos identificados en la misma región. De esta manera, la Figura 3.3 refleja que los SNPs se distribuyeron a lo largo de todos los cromosomas del girasol, y lo mismo ocurrió para los fragmentos con distinto número de SNPs. En particular, la mayor cantidad de SNPs se encontró en los cromosomas HanXRQChr05, HanXRQChr10 y HanXRQChr17.

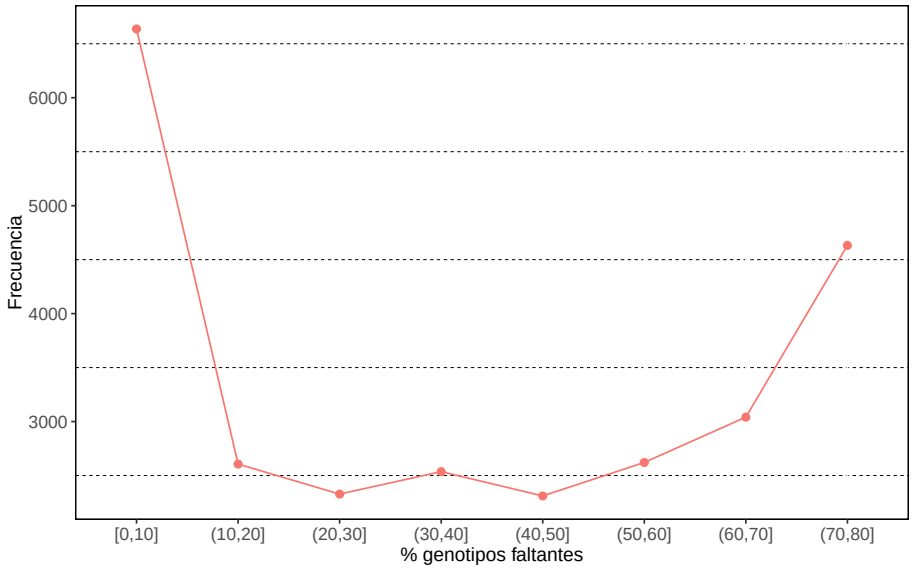

Figura 3.2: Diagrama de frecuencia de genotipos faltantes observadas en la matriz de genotipos de girasol.

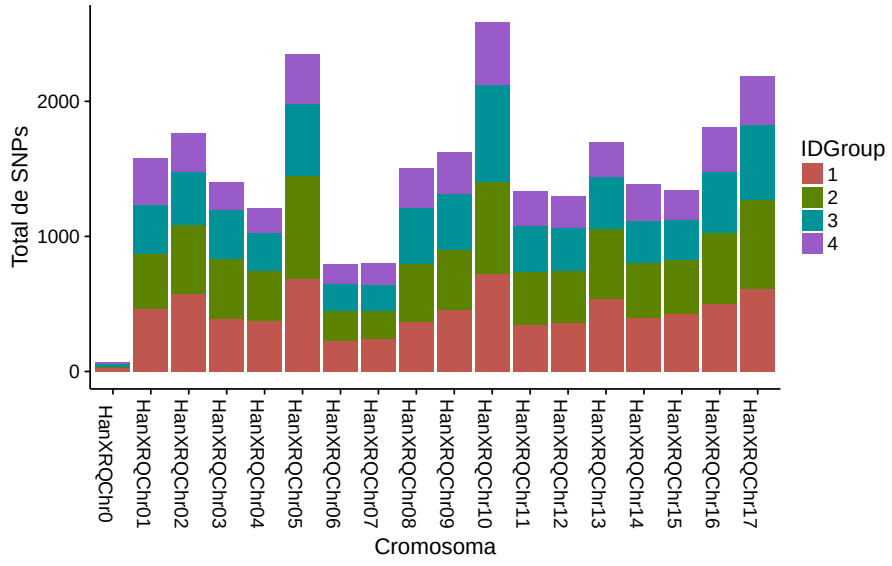

Figura 3.3: Distribución de los SNPs en los cromosomas de girasol. El código de colores se corresponde con el grupo al que pertenece cada SNP, definido en función del número total de SNPs identificados en la región donde se encuentra el mismo.

Posteriormente, se determinó el porcentaje de cambios alélicos observado en el experimento. Dado que los posibles valores de un alelo son cuatro, dos purinas (A y G) y dos pirimidinas (C y T), el número de modificaciones posibles es 12 ($= 4(4-1)$). Estos cambios se dividen según las bases que involucren. Por un lado una transversión implica el cambio entre una purina y una pirimidina, en cambio, una transición indica el cambio entre dos purinas o entre dos pirimidinas. Luego, las posibles transversiones son A/C , A/T, G/C y G/T (incluyendo C/A, T/A, C/G y T/G) y las posibles transiciones son A/G y C/T (incluyendo G/A y T/C). En la Figura 3.4 se ilustra los cambios observados en la matriz de genotipos bajo estudio. Tal y como se puede apreciar, las transiciones fueron las más frecuentes representando un 66,7 % de los cambios observados. Estos resultados coinciden con los reportados por Pegadaraju et al. (2013).

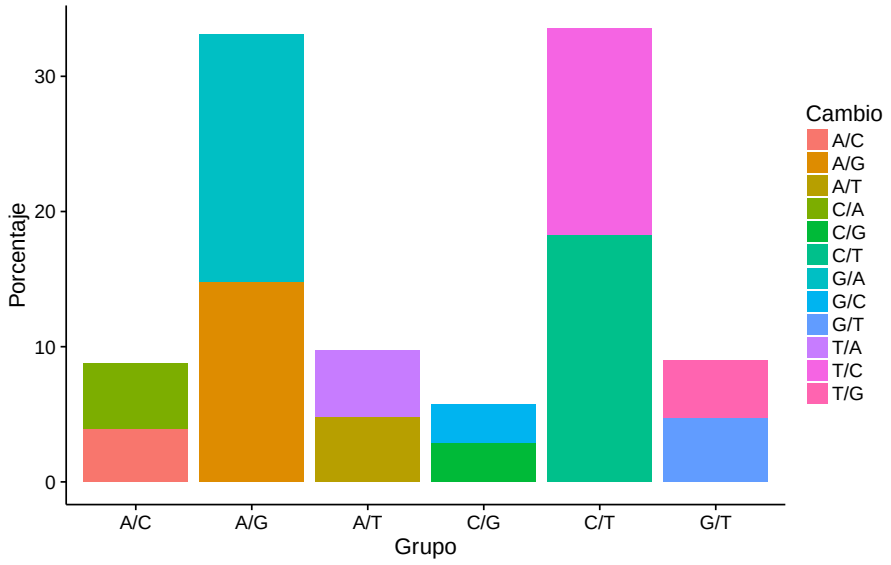

Figura 3.4: Cambios alélicos observados en las 135 líneas genotipificadas.

3.2. Algoritmo de imputación

3.2.1. Notación

Sea:

- La matriz de genotipos (MG), una matriz de $p \times n$ elementos, donde cada fila representa un SNP y cada columna el genotipo observado para dicho SNP en una de las p muestras bajo estudio.
- La matriz de meta-información (MMI), una matriz de $p \times k$ elementos, donde cada fila representa un SNP y cada columna contiene información de cada SNP, común a todas las muestras bajo estudio, como por ejemplo el cromosoma, la posición, el alelo de referencia y el alelo alternativo observados.

Sabiendo que los SNPs que provienen del mismo grupo de ligamiento (cromosoma) tienen elevada probabilidad de ser segregados en forma conjunta, se desarrolló la siguiente estrategia para imputar los valores faltantes de cada SNP de la matriz de genotipos.

3.2.2. Algoritmo

1. Dividir MG en dos matrices, una conteniendo las filas de MG que tienen datos faltantes y la otra con los elementos restantes. Estas matrices serán llamadas MC y $MINC$ en referencia a si tienen datos completos o incompletos, respectivamente. Si l es el número de SNPs que han sido genotipados en todas las líneas endocriadas (LE), entonces MC tendrá dimensión $l \times n$, mientras que $MINC$ será de $(p - l) \times n$. Adicionalmente, el SNP_i en $MINC$ tendrá asociadas m líneas donde éste ha sido genotipado y $n - m$ donde no se lo identificó (Figura 3.5).

2. Para cada SNP incompleto de $MINC$, $minc_i$ con i tomando valores $1, ..., (p-l)$:

   a) Identificar las m líneas ($m < n$) donde $minc_i$ ha sido genotipado y la sub-matriz de MC correspondiente a esas m líneas (Figura 3.6 **a**).

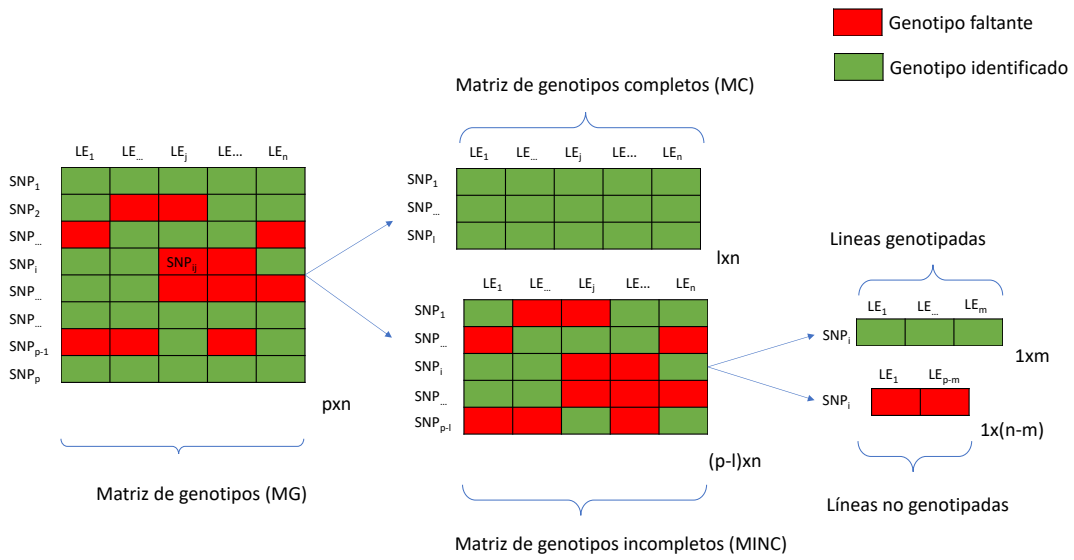

Figura 3.5: Estructuración de la matriz de genotipos previa a la imputación.

b) Definir el subconjunto de tamaño c ($c \leq l$) conteniendo los SNPs correlacionados con el SNP incompleto (Figura 3.6 **b**). Para ello, en primer lugar se selecciona los f SNPs en MC que pertenecen al mismo grupo de ligamiento que $minc_i$. Posteriormente, se determina si existe o no independencia ($\alpha = 0{,}05$) entre $minc_i$ y cada uno de dichos SNPs. Para ello se realiza un Test de Independencia χ^2, donde cada par (X, Y) está dado por los genotipos de los SNPs $minc_i$ y mc_j en las m líneas identificadas en el paso anterior. En caso de que no se encuentre ningún SNP correlacionado con $minc_i$, se considerarán todos los SNPs completos.

c) Confeccionar la matriz de genotipos completos asociada a $minc_i$, MC_i, la cual tendrá dimensiones $c \times n$, conteniendo sólo los genotipos de los SNPs que comparten grupo de ligamiento con $minc_i$ y que además están correlacionados con éste.

d) Dividir MC_i en dos matrices, según las m muestras donde $minc_i$ fue identificado. Así se tendrá una matriz de entrenamiento de dimensiones $c \times m$, y otra para la imputación ($c \times (n - m)$).

e) Entrenar un bosque aleatorio tomando como características los SNPs en

MC_i y como variable respuesta al SNP $minc_i$ (Figura 3.6 **c**). Para ello, formar el conjunto de datos con los valores de $minc_i$ en las m muestras donde fue genotipado y con la matriz de entrenamiento.

f) Utilizar el *Random Forests* entrenado en el paso anterior para estimar los genotipos faltantes en las $(n-m)$ muestras. Utilizar como predictora a la matriz de imputación (Figura 3.6 **d**).

3. Imputar los genotipos faltantes de MG con los predichos mediante *Random Forests*.

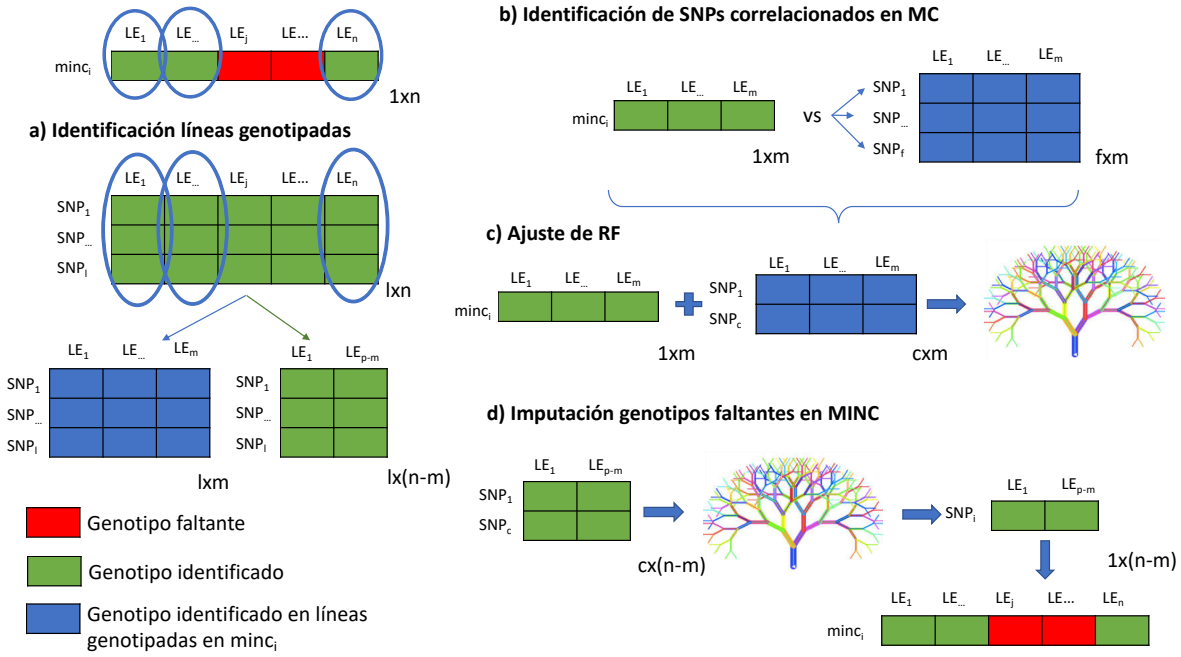

Figura 3.6: Algoritmo de imputación. El algoritmo desarrollado consta de cuatro etapas, las cuales se deben ejecutar para cada SNP que posee genotipos faltantes. **a)** Identificación de los individuos genotipados en el SNP incompleto. **b)** Identificación de los SNPs correlacionados con el SNP incompleto. **c)** Ajuste del *Random Forests* utilizando los genotipos identificados de dicho SNP y los SNPs correlacionados. **d)** Imputación de los genotipos faltantes, predichos con el *Random Forests* previamente ajustado.

3.2.3. Implementación

El algoritmo desarrollado ha sido implementado en el lenguaje `R` utilizando librerías disponibles. Los procesamientos básicos de filtrado y preparación de las matrices se realizaron mediante funcionalidades básicas, incluidas en el `R`. Para realizar el test de independencia, `R` provee la función `chisq.test` del paquete `stats`. Sin embargo, los tiempos de ejecución que ésta implica no están optimizados para grandes volúmenes de datos o grandes cantidades de pruebas. Por lo tanto, se implementó una función alternativa que ejecuta la misma tarea pero utilizando lenguaje `C++`, utilizando la interfaz que provee el paquete `Rcpp` (Eddelbuettel, 2013). La librería `ranger` (Wright and Ziegler, 2017) fue utilizada para el ajuste de los *Random Forests* y predicción de los genotipos faltantes. El código utilizado para realizar todos los procesamientos así como también la implementación del algoritmo propuesto se encuentra disponible en el repositorio GitHub SNPsRFImputation (`https://github.com/gamerino/SNPsRFImputation`).

En este trabajo se evaluaron distintas alternativas en base al algoritmo propuesto, el cual consta de dos etapas fundamentales. La primera de ellas es la de detección de SNPs correlacionados, y la segunda es el ajuste del bosque aleatorio seguido de la imputación. En términos de la primera etapa, se consideraron tres posibilidades

- Ignorar la correlación entre SNPs,
- Considerar la existencia de correlación entre todos los SNPs,
- Considerar la existencia de correlación sólo entre SNPs que comparten grupo de ligamiento.

Por otro lado, en término del ajuste e imputación mediante *Random Forests* se probó

- Ajuste e imputación de todos los genotipos faltantes
- Ajuste e imputación de los genotipos faltantes sólo en los SNPs cuyo error OOB durante el ajuste fue menor a cierto umbral, el cual se tomó en 0.2.

De esta manera seis alternativas del algoritmo propuesto fueron evaluadas, las cuales han sido resumidas en la Figura 3.7

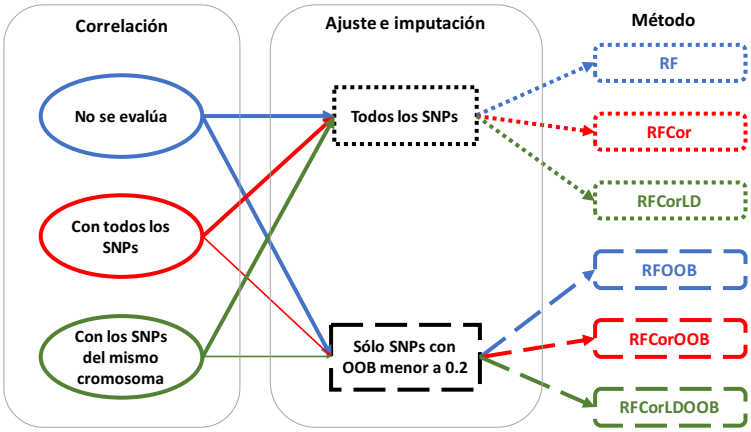

Figura 3.7: Alternativas de imputación evaluadas. Todas ellas se basan en la estrategia de imputación propuesta.

3.3. Evaluación mediante datos simulados

3.3.1. Generalidades

La base de datos de espinoso que se utilizó como punto de partida cuenta con información de 51.497 SNPs genotipados en 105 individuos. Distintas simulaciones se realizaron con el fin de evaluar diversos aspectos del algoritmo propuesto. Cada simulación se replicó diez veces para proveer de potencia estadística a las estimaciones. Los pasos llevados a cabo en cada simulación fueron:

- Imponer a los SNPs la distribución de datos faltantes observada en la base de datos incompleta real de girasol. De esta manera, un porcentaje de los SNPs fueron elegidos para conservar los genotipos en las 105 muestras, mientras que al resto se les "borró" esa información de genotipo en al menos una muestra. Para esto, se diseñó la función `simulateNAs`.

- Calcular para cada SNP el total de muestras donde ha sido conservado el genotipo (`getNS`) y las frecuencias alélicas correspondientes (`calculateAFALT`).

- Filtrar los SNPs con MAF $< 0{,}05$.

- Controlar que el alelo de referencia sea el de mayor frecuencia, sino invertir la

definición de dicho alelo y los genotipos (`invertGTs`).

- Filtrar las regiones con sobre abundancia de SNPs (> 4).
- Guardar los genotipos observados en la base de datos completa, de los SNPs que se conservaron luego de los filtrados.
- Controlar que se respete el porcentaje de SNPs con datos completos, la relación transversiones/transiciones.

3.3.2. Desempeño global

Para evaluar el desempeño del algoritmo de imputación en sus seis alternativas, se construyó una base de datos conformada por diez matrices de genotipos con datos faltantes siguiendo el procedimiento explicado en la sección anterior. En particular, el porcentaje de SNPs con datos completos que se fijó fue el mismo observado en la base de datos de girasol. Las matrices simuladas consistieron, en promedio, de 3.272 SNPs genotipados en al menos 21 individuos. El porcentaje promedio de SNPs con genotipos en todas las muestras fue 6,4 %. Las frecuencias observadas de porcentajes de datos faltantes, en las diez simulaciones, se ilustran en la Figura 3.8 donde también se ha incluido las frecuencias correspondientes a la base de datos de girasol.

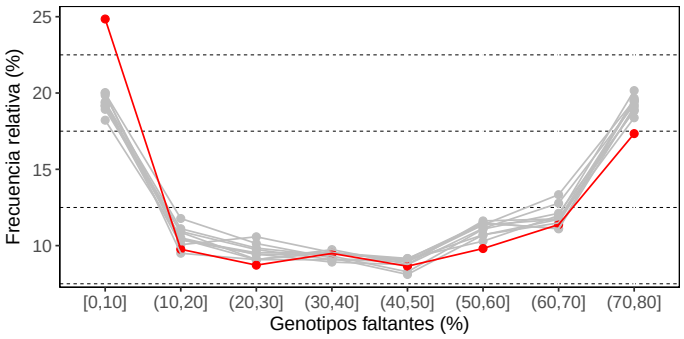

Figura 3.8: Frecuencias observadas de porcentajes de datos faltantes. Cada SNP ha sido clasificado según el porcentaje de individuos en los cuales no ha sido genotipado. En color gris, se muestran los resultados para las diez matrices simuladas y en rojo los resultados de la matriz de genotipos de girasol.

El análisis del conjunto de datos simulados comenzó con la identificación de la matriz de genotipos incompletos. En promedio, el número de SNPs con genotipos faltantes fue de 3.062, con la distribución mostrada en la Figura 3.9**A)**. Luego se procedió al análisis e imputación de los genotipos con las seis alternativas del algoritmo propuesto. En particular, las alternativas que consideran el error de estimación, OOB, imputaron un número menor de SNPs. En la Figura 3.9**B)** se ilustra el diagrama de cajas correspondiente al porcentaje de SNPs que se imputó con cada una de ellas. Como se puede apreciar, la alternativa *RFCorOOB* es la que consiguió imputar el mayor porcentaje de SNPs, mientras que *RFOOB* logró el menor porcentaje.

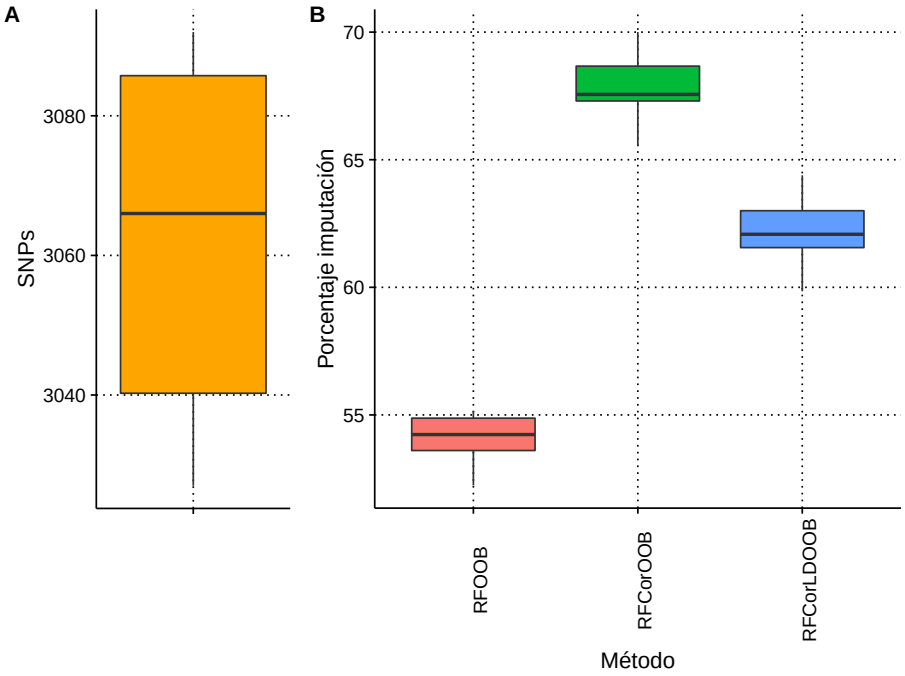

Figura 3.9: Resumen de cantidad de SNPs. **A)** Diagrama de cajas denotando el número SNPs simulados. **B)** Porcentajes de SNPs imputados con los métodos que consideran el error en la estimación (OOB).

Luego de la imputación, la matriz *MG* fue procesada para determinar si algúno de sus SNPs tenía MAF$<$ 0,05, y, en caso de que esto ocurriera, eliminarlo. Así, el número promedio de SNPs en la *MG* final (completos más imputados) fue de 3.272, con valores dentro del rango [3.234, 3.309], según la distribución mostrada en la

Figura 3.10**A)**. El porcentaje de datos completos finalmente obtenido por cada una

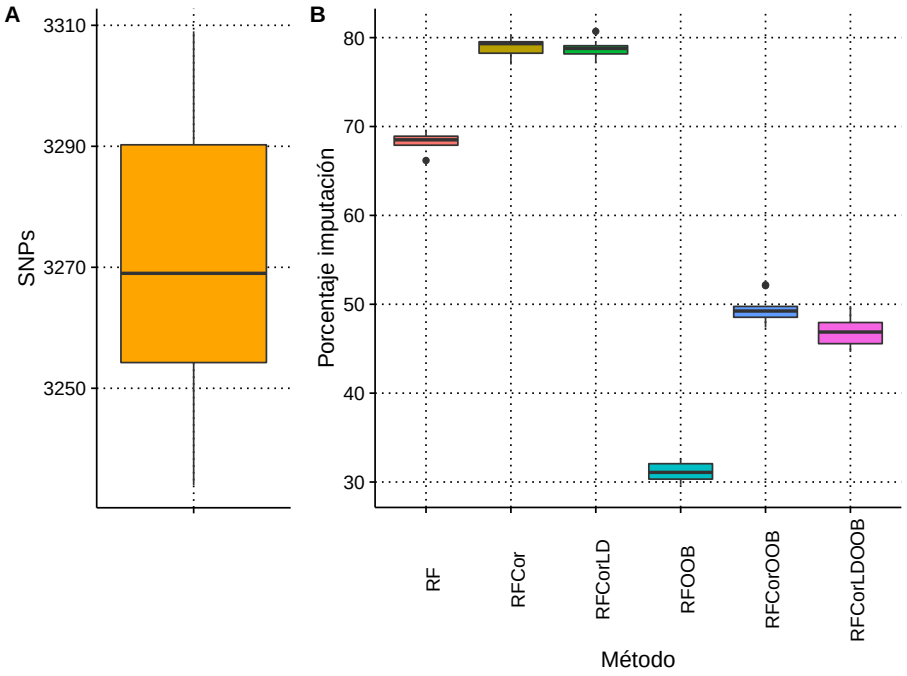

Figura 3.10: Resumen de cantidad de SNPs completos contenidos en la matriz de genotipos final. **A)** Diagrama de cajas denotando el número total de SNPs simulados. **B)** Porcentaje final de SNPs completos, obtenidos con los métodos propuestos.

de las seis alternativas fue también explorado para determinar, de cierta manera, el desempeño de los métodos. La Figura 3.10**B)** ilustra, mediante diagramas de cajas, la variación de los porcentajes de datos completos finalmente obtenido por cada una de las seis alternativas. Claramente se aprecia que los métodos que no consideran el error OOB lograron porcentajes superiores a las que sí lo hacen, lo cual es de esperar ya que imputan todos los SNPs incompletos. En particular, las dos metodologías que consideran correlación entre SNPs lograron porcentajes cercanos al 80 %, superando en términos significativos a la alternativa *RF*, que no alcanzó el 70 %. Además, no se encontraron diferencias significativas entre los porcentajes logrados por *RFCor* y *RFCorLD* (valor-p de la Prueba de Wilcoxon para muestras relacionadas= 0,1934). Por otro lado, las alternativas que consideran correlación entre SNPs y error *OOB* lograron porcentajes superiores al alcanzado por *RFOOB*, el cual apenas superó el

30 %. Particularmente, *RFCorOOB* logró porcentajes cercanos al 50 %, mientras que los alcanzados por *RFCorLDOOB* fueron significativamente más pequeños (valor-p de la Prueba de Wilcoxon para muestras relacionadas = 0,001953), aunque los valores promedios resultaron muy próximos entre sí: 49,5 % y 46,9 %, respectivamente.

Complementariamente, se exploró el número de SNPs determinados como correlacionados por las metodologías *RFCor* y *RFCorLD*. Cabe aclarar que éstas comparten, respectivamente, resultados con las alternativas *RFCorOOB* y *RFCorLDOOB* ya que sólo se diferencian de ellas en procesamientos posteriores. La Figura 3.11 ilustra bandas de confianza obtenidas de las estimaciones de las funciones de densidad del número de SNPs correlacionados en las diez bases de datos simuladas.

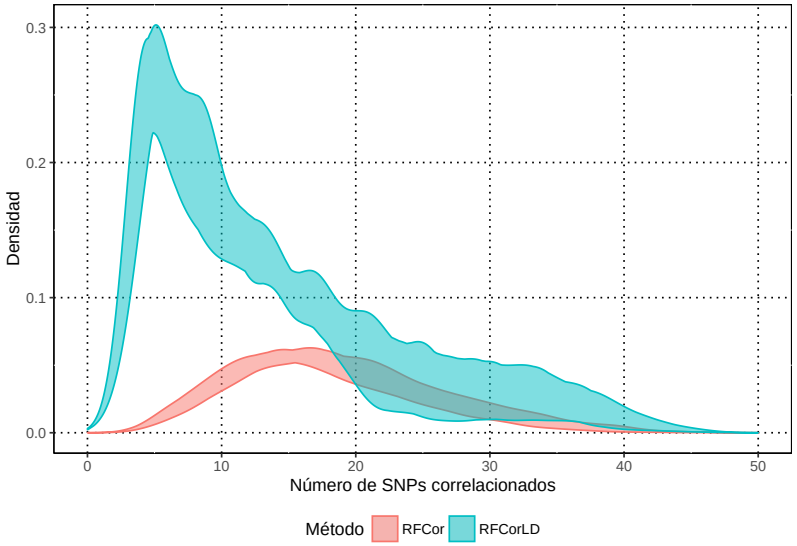

Figura 3.11: Densidad estimada del número de SNPs correlacionados con los SNPs incompletos, encontrados con las metodologías *RFCor* y *RFCorLD*.

Como se puede apreciar, la metodología *RFCorLD* resultó en un menor número de SNPs correlacionados con los SNPs a imputar con una moda aproximadamente en 5, lo cual permitió reducir el tiempo de confección de cada *Random Forests*. Por otro lado, la metodología *RFCor* evidenció una función de densidad más suave, con un pico aproximadamente en 15. En este punto cabe recordar que para aquellos SNPs que no correlacionaron con ninguno de los SNPs completos, ambas metodo-

logías consideraron el total de los SNPs completos para hacer la imputación. En el caso de *RFCor*, se encontró que en promedio el 0,24 % (DE = 0,06) de los SNPs incompletos no correlacionaron con los completos, en cambio, este porcentaje fue aproximadamente cien veces superior para el caso de *RFCorLD* (28,6 %, DE=1,54). Este resultado es esperable ya que cuando se restringe los SNPs completos al subconjunto de éstos que comparten grupo de ligamiento con el SNP incompleto, el número de SNPs a indagar vía la correlación disminuye, e incluso puede llegar a ser nulo. Sin embargo, resultó interesante indagar qué sucede con aquellos SNPs incompletos que han sido imputados con todos los SNPs completos, por no haber podido establecer correlación sólo con un subconjunto de ellos. En este sentido, la Figura 3.12 ilustra los diagramas de cajas correspondientes a los porcentajes de SNPs que fueron imputados utilizando la información de todos los SNPs completos y que además lograron un valor de *OOB* menor al requerido para ser imputados por las metodologías *RFCorOOB* y *RFCorLDOOB*. Al comparar las dos distribuciones, se encontraron

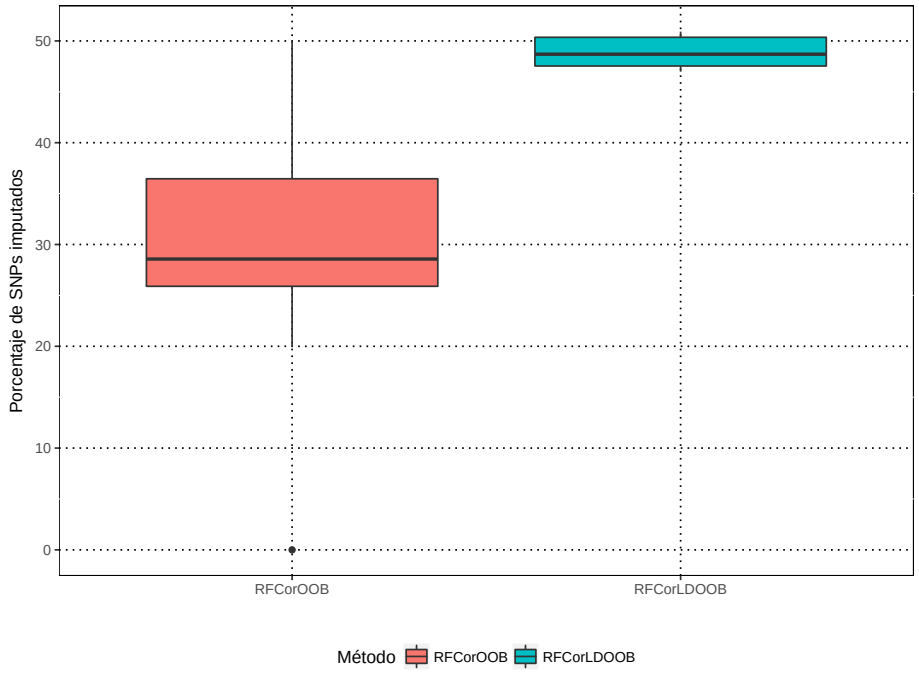

Figura 3.12: Diagrama de cajas denotando el número SNPs incompletos que fueron imputados con todos los SNPs completos logrando un *OOB* menor al requerido.

diferencias significativas (valor-p de la Prueba de Wilcoxon = 0,000986), indicando que el porcentaje de los SNPs conservados luego de la consideración del error *OOB*, es mayor para la metodología *RFCorLDOOB*. En particular, el porcentaje promedio para esta metodología fue cercano a 50 %, entretanto, para *RFCorOOB* fue cercano a 30 %.

Las medidas de desempeño obtenidas para las seis alternativas en las diez matrices de genotipos simuladas se han resumido en los diagramas de puntos de la Figura 3.13. En términos de exactitud (Figura 3.13**A**) se aprecia que todas las alternativas lograron valores entre 0,77 y 0,84, lo cual refleja su bondad a la hora de identificar los verdaderos genotipos (VN y VP). Asimismo, es fácil notar dos agrupamientos entre las alternativas, las que lograron mayores valores y las que lograron los valores más bajos lo cual coincide con el hecho de considerar o no el OOB a la hora de determinar que SNPs imputar, respectivamente.

Por otro lado, en términos del F-score (Figura 3.13**B**) no se aprecia dicho agrupamiento aunque los valores más pequeños estuvieron asociados a las alternativas que no consideran el OOB y los más altos a las que sí lo tienen en cuenta, con la misma tendencia observada para la sensibilidad (Figura 3.13**C**). A diferencia de la exactitud, la sensibilidad observada resultó ser menor a 0,65 y sólo supero el valor 0,6 para la alternativa *RFCorLDOOB*, mientras que el resto de las metodologías evidenció valores inferiores. Esto indica que la probabilidad promedio de clasificar correctamente los genotipos es inferior a 0,65, un valor que claramente en el caso ideal debería ser igual a 1. En términos de la precisión (Figura 3.13**D**) se identificaron tres agrupamientos entre las metodologías.

Por un lado, la metodología más sencilla, *RF*, evidenció los valores más bajos ($< 0{,}45$), seguida de *RFCor*, *RFCorLD* y *RFOOB*, que lograron valores de precisión en torno a 0,5; mientras, *RFCorOOB* y *RFCorLDOOB* obtuvieron los valores más elevados, con valor medio 0,55 y 0,58, respectivamente. Sin embargo, teniendo en cuenta que la precisión refleja la probabilidad de que el genotipo asignado a un SNP sea realmente el que se debería identificar, los valores observados no son los suficientemente altos. Los peores resultados se obtuvieron para la alternativa más sencilla, *RF*, por el contrario, el mejor desempeño fue el de *RFCorLDOOB*. Si bien la alternativa *RFOOB* evidenció una mayor exactitud (Prueba de Wilcoxon para muestras

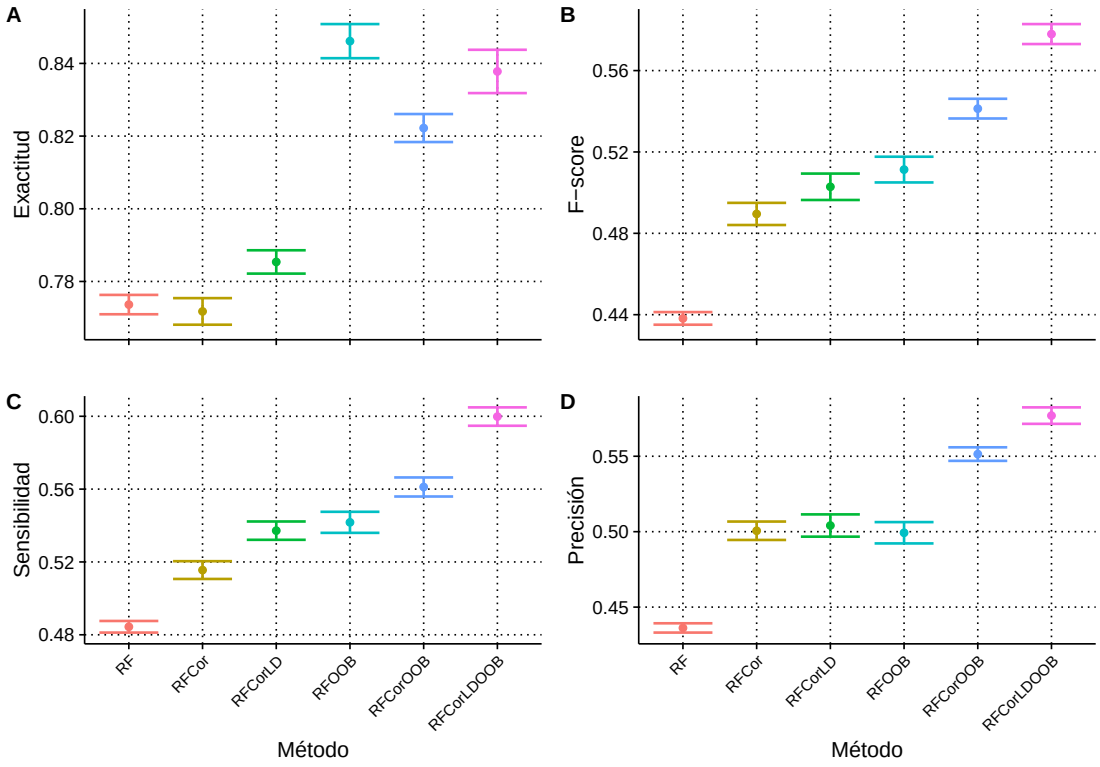

Figura 3.13: Diagrama de puntos para las medidas de desempeño de los métodos evaluados. Se muestra el valor medio (± desvío estándar) para **A)** Exactitud, **B)** F-score, **C)** Sensibilidad y **D)** Precisión.

relacionadas, valor-p= 0,003906) que *RFCorLDOOB*, el porcentaje de datos faltantes imputados con ésta fue inferior (valor-p= 0,0009766, Prueba de Wilcoxon para muestras relacionadas) al imputado con *RFCorLDOOB*.

3.3.3. Efecto de la selección de umbral de OOB

En la sección anterior se ha determinado que las metodologías que consideran el OOB a la hora de llevar a cabo la imputación evidenciaron un mejor desempeño que las que no lo consideran. Como es de esperar, variaciones en el umbral de OOB producen variaciones en el número de SNPs a imputar, por lo que es posible que afecten el desempeño de los algoritmos. La Figura 3.14 ilustra y resume el efecto de la selección del umbral de OOB sobre el número de SNPs a imputar y sobre las

medidas de desempeño de las metodologías evaluadas.

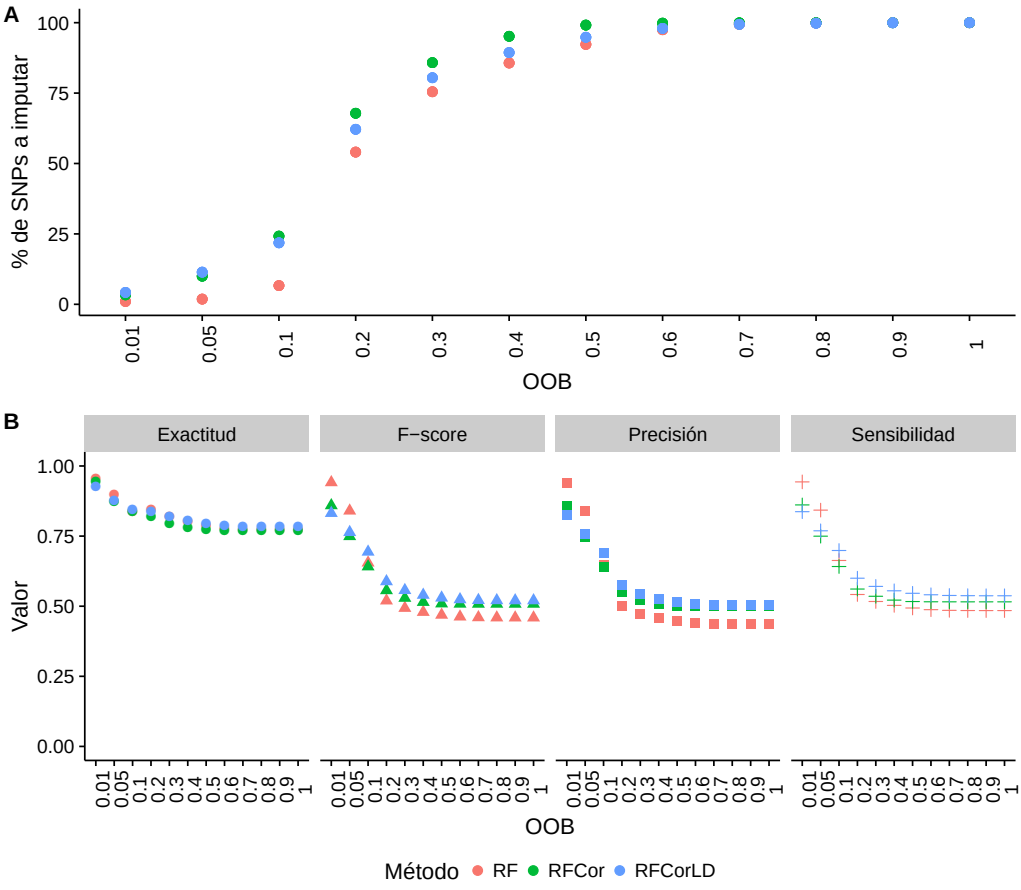

Figura 3.14: Resultados a diferentes umbrales de error de estimación (OOB). El panel **A)** muestra el porcentaje de SNPs a imputar por cada metodología, determinado por el valor de OOB seleccionado. En **B)** se muestran las medidas de desempeño obtenidas al imputar sólo los SNPs que superan el OOB correspondiente.

El análisis de la Figura 3.14**A**, reveló que a medida que se relajó el umbral de OOB, se pudo estimar un mayor porcentaje de SNPs. Tal y como se aprecia en la gráfica, la región de mayor cambio en el porcentaje de estimación es la que abarca los OOB entre 0,1 y 0,3. Valores de OOB menores o iguales a 0,1 determinaron la imputación de tan sólo el 25 % de los SNPs con genotipos faltantes, lo cual representa un elevado porcentaje de pérdida de SNPs. Considerando un umbral de 0,2, todos

los métodos lograron imputar al rededor del 50 % de los genotipos faltantes. Por otra parte, valores de OOB iguales o superiores a 0,3 permitieron la imputación de más del 75 % de los SNPs, por supuesto a costa de un mayor error en la estimación.

La relación entre el umbral de OOB y las medidas de desempeño ha sido ilustrada en la Figura 3.14**B**. Allí se aprecia que a medida que se incrementó el umbral de error aceptado, el desempeño del algoritmo disminuyó ya que todas las medidas exhibieron valores más pequeños. En particular, la medida menos influenciada resultó ser la exactitud promedio, cuyo rango de variación fue entre 0,75 y 1. La precisión, la sensibilidad y por ende el F-score fueron las más afectadas, con rango de variación entre 0,4 y 1. Es apreciable además que todas las medidas de desempeño exhibieron la región máxima de variación para valores de OOB entre 0 y 0,2. Estos resultados sustentan la selección de umbral OOB igual a 0,2 ya que con este valor es posible imputar el 50 % de los SNPs con genotipos faltantes, optimizando el desempeño de los algoritmos. Adicionalmente, se destaca que la metodología *RFCorLD* ha evidenciado mejores medidas de desempeño que las otras dos alternativas permitiendo imputar, en promedio, el 62 % de los genotipos faltantes.

3.3.4. Efecto de la selección del valor de significancia

La prueba de independencia que se utilizó para determinar cuáles de los SNPs completos están correlacionados con los SNPs incompletos requirió de la utilización de un valor de significancia para el rechazo de la hipótesis nula. Si bien el valor 0,05 es ampliamente utilizado, en este trabajo se evaluó el efecto de cambiar ese valor por uno más flexible: 0,1. Luego, se compararon las alternativas propuestas anteriormente con las mismas estrategias pero considerando otro umbral de significancia para determinar la correlación entre SNPs. Es así que se definió un conjunto nuevo de alternativas llamadas *RFCor01*, *RFCorLD01*, *RFCorOOB01* y *RFCorLDOOB01*. En primer instancia, se estudió el efecto del cambio en el umbral de significancia sobre el porcentaje de SNPs imputados por las metodologías que consideran el OOB (Figura 3.15). Las pruebas de diferencia entre distribuciones de los porcentajes obtenidos por cada alternativa, en sus dos variantes, revelaron que no se encontró diferencias significativas (valor-p de la prueba de Wilcoxon mayores a 0,1) entre ellas. Es decir, que usar la alternativa originalmente propuesta o modificando el umbral de signifi-

cancia a 0,1 conllevaría a los mismos resultados en términos del porcentaje de SNPs imputados.

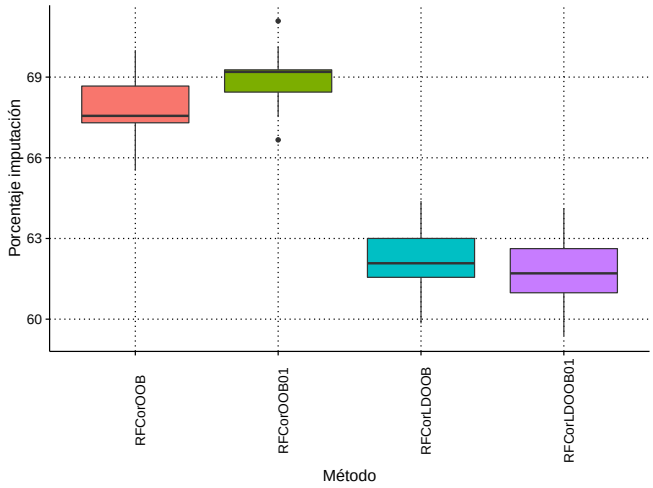

Figura 3.15: Porcentajes de SNPs imputados con los métodos que consideran el error en la estimación (OOB). Los métodos *RFCorOOB* y *RFCorLDOOB* han sido utilizados con un valor de significancia 0,05, mientras que para los dos restantes se utilizó un valor igual a 0,1.

En forma consistente con los descrito anteriormente, se encontró que las alternativas que evalúan correlación con todos los SNPs completos permiten imputar más SNPs que las que consideran sólo la posible correlación con SNPs del mismo grupo de ligamiento.

El efecto de la variación en el valor de significancia sobre las medidas de desempeñó también ha sido evaluado. La Figura 3.16 ilustra los resultados obtenidos para las alternativas que consideraron correlación con umbral de significancia 0,05 (*RFCor*, *RFCorLD*, *RFCorOOB* y *RFCorLDOOB*) y las que utilizaron 0,1 como dicho umbral (*RFCor01*, *RFCorLD01*, *RFCorOOB01* y *RFCorLDOOB01*). Las estimaciones ilustradas en dicha figura han sido comparadas de a pares, utilizando la Prueba de Wilcoxon para muestras relacionadas, para determinar si existen o no diferencias significativas como consecuencia de la variación en dicho parámetro. En cuanto a la exactitud de las alternativas (Figura 3.16**A**), no se evidenciaron cambios significativos en las metodologías analizadas (valor p> 0,05), excepto para *RFCor* (valor p

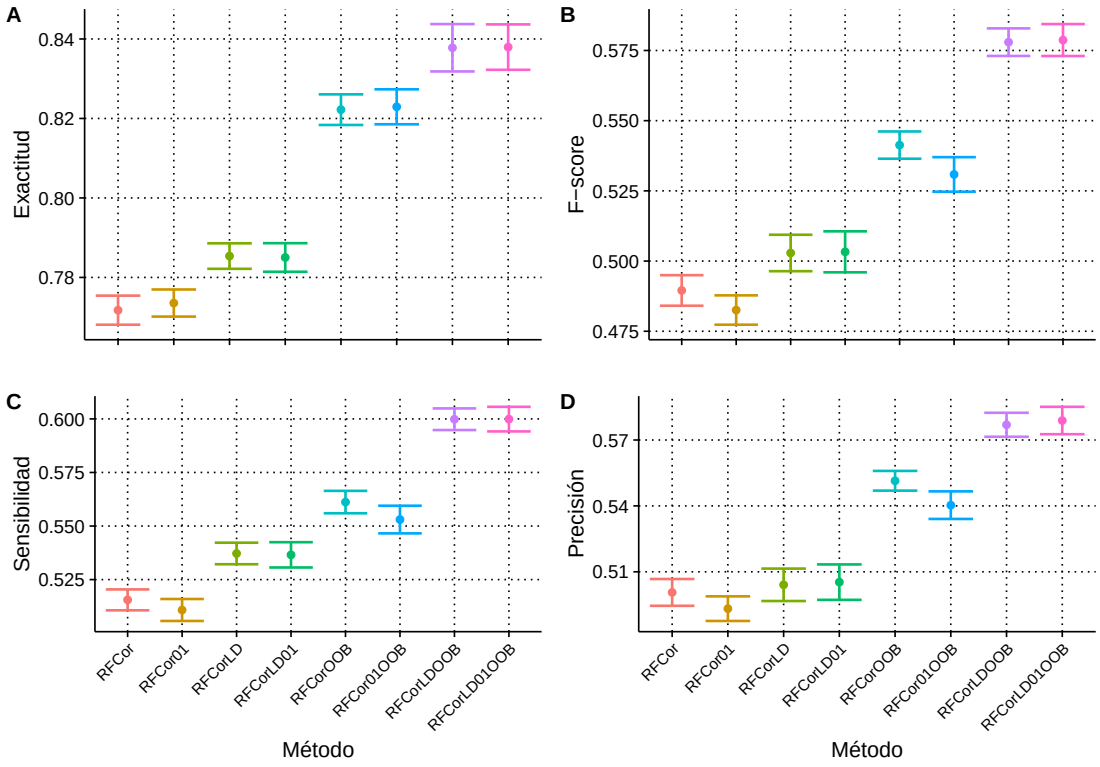

Figura 3.16: Diagrama de puntos para las medidas de desempeño de los métodos evaluados, considerando dos umbrales de significancia. Se muestra el valor medio (± desvío estándar) para **A)** Exactitud, **B)** F-score, **C)** Sensibilidad y **D)** Precisión.

= 0,001953). Por el contrario, tanto en el F-score (Figura 3.16**B**), la sensibilidad (Figura 3.16**C**) y la precisión (Figura 3.16**C**), se encontraron diferencias significativas (valor p< 0,05) como consecuencia del cambio en el umbral de significancia, con excepción de la precisión y el F-score en las metodologías *RFCorLD* y *RFCorLDOOB*. En particular se determinó que existen diferencias en tales medidas de desempeño para las metodologías que evalúan correlación entre los SNPs incompletos y todos los SNPs completos. Tal y como se puede apreciar en los paneles correspondientes, el desempeño de tales metodologías con umbral de significancia de 0,05, *RFCor* y *RFCorOOB*, fue superior al de las respectivas alternativas que consideraron umbral igual a 0,1, *RFCor01*, *RFCorOOB01*. Estos resultados demuestran que la alternativa de la estrategia propuesta que evalúa independencia entre SNPs de mismo grupo de

ligamiento resultó ser más robusta a los cambios en el umbral de significancia de dicho test.

3.3.5. Efecto del porcentaje de genotipos faltantes

La relación entre el porcentaje de genotipos faltantes por SNP y el desempeño de las alternativas del algoritmo propuesto también fueron evaluadas. Para ello se determinó para cada SNP el porcentaje de muestras que no tenían genotipo registrado. En una primera instancia, se determinó la asociación entre las medidas de desempeño y el porcentaje de datos faltantes para cada SNP mediante el coeficiente de correlación de Pearson (ρ) (Tabla 3.1). Como primer observación es notable que excepto la correlación con la exactitud de la metodología *RFOOB*, el resto de los valores fueron todos negativos, indicando que a medida que el porcentaje de genotipos faltantes aumentó el desempeño de las alternativas disminuyó. En términos medios, la correlación con la exactitud fue prácticamente nula ($|\rho| < 0{,}05$) para todos los métodos. Un comportamiento diferente se encontró para el resto de las medidas de desempeño evaluadas, donde el coeficiente ρ (promedio) varío entre $-0{,}299$ y $-0{,}459$.

Tabla 3.1: Coeficiente de correlación de Pearson entre el porcentaje de genotipos faltantes y las medidas de desempeño de las alternativas propuestas. En cada cuadro se muestra el valor promedio seguido del desvío estándar.

Método	***Exactitud***	***Sensibilidad***	***Precisión***	***F-score***
RF	-0,033 (0,015)	-0,425 (0,012)	-0,362 (0,019)	-0,408 (0,013)
RFCor	-0,034 (0,016)	-0,362 (0,015)	-0,306 (0,016)	-0,349 (0,017)
RFCorLD	-0,023 (0,011)	-0,326 (0,013)	-0,292 (0,017)	-0,317 (0,015)
RFOOB	0,011 (0,02)	-0,459 (0,02)	-0,412 (0,022)	-0,438 (0,02)
RFCorOOB	-0,044 (0,025)	-0,397 (0,026)	-0,331 (0,025)	-0,382 (0,026)
RFCorLDOOB	-0,014 (0,019)	-0,326 (0,026)	-0,299 (0,025)	-0,320 (0,026)

Coincidentemente, tanto para la sensibilidad como para la precisión, se encontró que las metodologías cuyo desempeño resultó menos correlacionado con el porcentaje de genotipos faltantes fueron las que consideran la dependencia entre SNPs del mismo cromosoma (*RFCorLD* y *RFCorLDOOB*), en cambio, las más influenciadas resulta-

ron ser las metodologías que no toman en cuenta la correlación entre SNPs (*RF* y *RFOOB*).

Complementariamente, el efecto del porcentaje de genotipos faltantes sobre las medidas de desempeño se exploró también gráficamente (Figura 3.17). Para ello, los SNPs fueron agrupados en cuatro categorías: $[0, 20]$, $(20, 40]$, $(40, 60]$ y $(60, 80]$ de acuerdo al porcentaje de muestras sin genotipos. Sobre cada grupo se determinó las medidas. En concordancia con lo descrito anteriormente, en la Figura 3.17**A** se

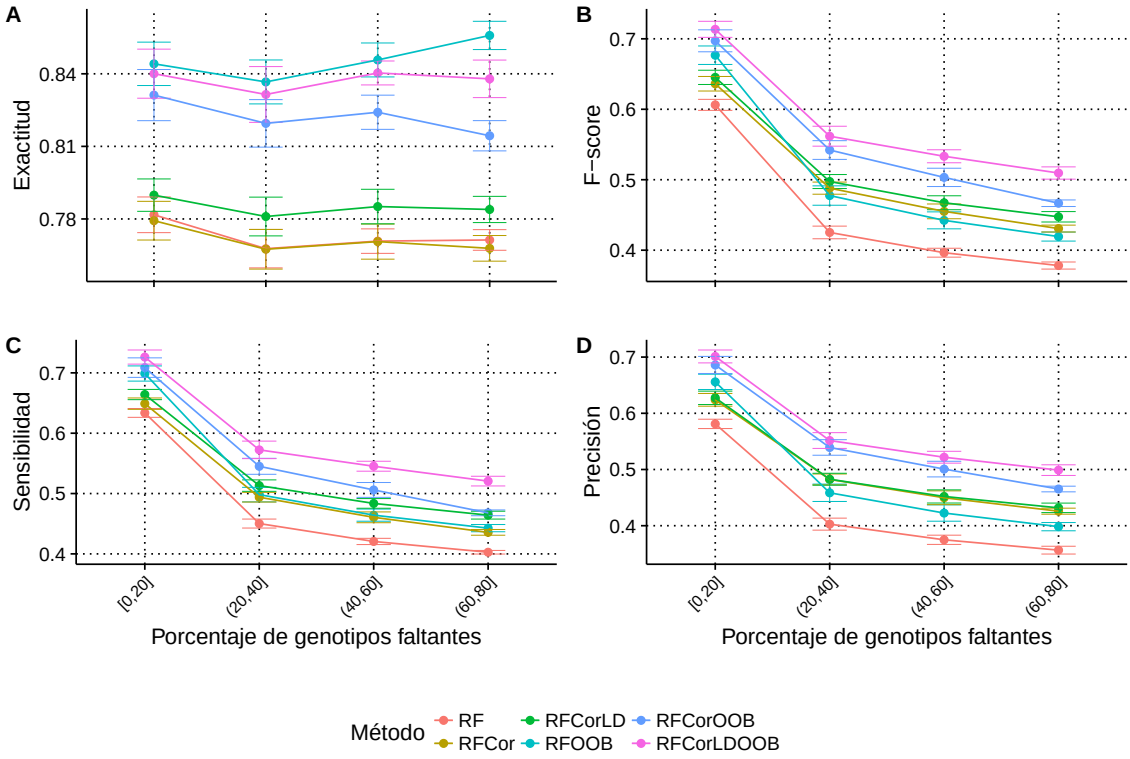

Figura 3.17: Diagrama de puntos y líneas para las medidas de desempeño de los métodos evaluados computados en grupos de SNPs definidos según el porcentaje de genotipos faltantes que éstos presentaron. Se muestra el valor medio (± desvío estándar) para **A)** Exactitud, **B)** F-score, **C)** Sensibilidad y **D)** Precisión.

aprecia que prácticamente no hubo cambios en la exactitud cuando se modificó el porcentaje de muestras sin genotipar. Un comportamiento diferente se observa en los paneles **B**, **C** y **D** de la misma figura, donde es evidente el decaimiento en el F-

score, la sensibilidad y la precision a medida que el porcentaje de SNPs sin genotipar aumentó, es decir la correlación negativa entre estas cantidades. A su vez, es notable que la mayor pérdida de desempeño se produjo cuando el porcentaje de SNPs sin genotipar supera el 20 %.

3.3.6. Efecto del porcentaje de SNPs completos

Otro efecto interesante de estudiar fue el que ejerció el porcentaje de SNPs completos sobre el desempeño de los algoritmos. En este trabajo se tomó como porcentaje base el observado en la base de datos de girasol. A partir de allí, se obtuvieron dos nuevos conjuntos de datos uno en el cual se duplicó el porcentaje de SNPs con genotipos completos y otro en el cual este porcentaje se cuadruplicó. Sobre estas dos nuevas bases de datos se corrieron las alternativas propuestas y posteriormente determinaron las medidas de desempeño.

Con el fin de distinguir la aplicación de los algoritmos en cada uno de los nuevos conjuntos de datos, a la denominación definida en la Figura 3.7 se le añadió un 2 o un 4 según se refiere a la aplicación de los mismos en la base con el doble o cuádruple de SNPs completos, respectivamente. Así, por ejemplo, la metodología denominada *RF* aplicada sobre el primero de los conjuntos se denominó *RF2* y sobre el segundo, *RF4*.

La Figura 3.18 muestra el porcentaje de datos imputados por las metodologías que consideran el OOB. Se encontró que el porcentaje de SNPs a imputar aumentó a medida que el porcentaje de SNPs completos, utilizados para la predicción de los genotipos faltantes, fue mayor. En particular, se observaron similares razones de aumentos cuando se compararon los porcentajes de imputación de la base de datos con el doble de SNPs completos con la de referencia y al comparar la base de datos con el cuádruple de SNPs completos con la duplicada. Consistentemente con lo descrito anteriormente, el porcentaje de datos a imputar resultó más alto para la metodología *RFOOB* en los tres casos, alcanzando valores superiores al 80 % de los SNPs imputados en el caso en que se cuadruplicó el porcentaje de SNPs completos (*RFCorOOB4*).

La Figura 3.19 presenta los gráficos de puntos, resumiendo las medidas de desempeño obtenidas para las seis alternativas en las diez matrices de genotipos simuladas

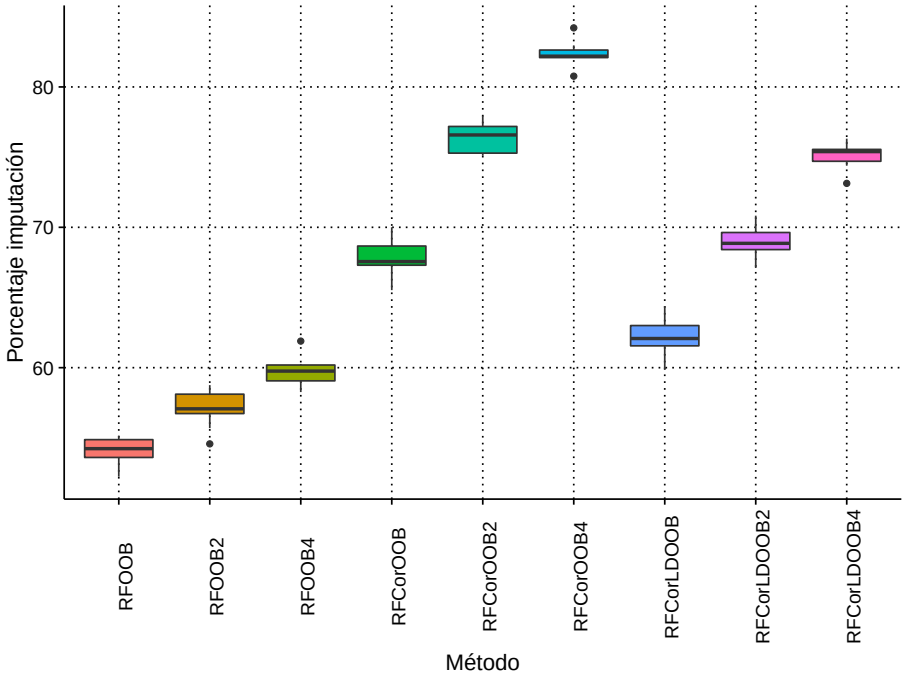

Figura 3.18: Porcentajes de SNPs imputados con los métodos que consideran el error en la estimación (OOB) en tres conjuntos de datos, uno con un porcentaje aproximado de SNPs completos de 6, el otro de 12 y el tercero de 24.

para cada uno de los tres conjuntos de datos. En términos de exactitud (Figura 3.19**A**) se aprecia una tendencia de aumento de dicha medida con el porcentaje de SNPs completos fundamentalmente para las alternativas que no consideran correlación. Mientras tanto, la metodología *RFOOB* es la única que exhibió disminución de la exactitud al aumentar el número de SNPs disponibles para ajustar los *Random Forests*. Por otro lado, las alternativas que consideran correlación entre SNPs fueron las más estables en términos de exactitud.

La Figura 3.19**B** ilustra los valores de F-score observados. En cada una de las metodologías evaluadas, se encontró que el F-score, en general, se incrementó con el aumento de SNPs completos. En particular, los métodos que consideran correlación entre SNPs del mismo grupo de ligamiento, con o sin consideración del OOB, son los que exhibieron mayores incrementos. Los resultados de sensibilidad (Figura 3.19**C**) y precisión (Figura 3.19**D**) exhibieron el mismo comportamiento observado para el

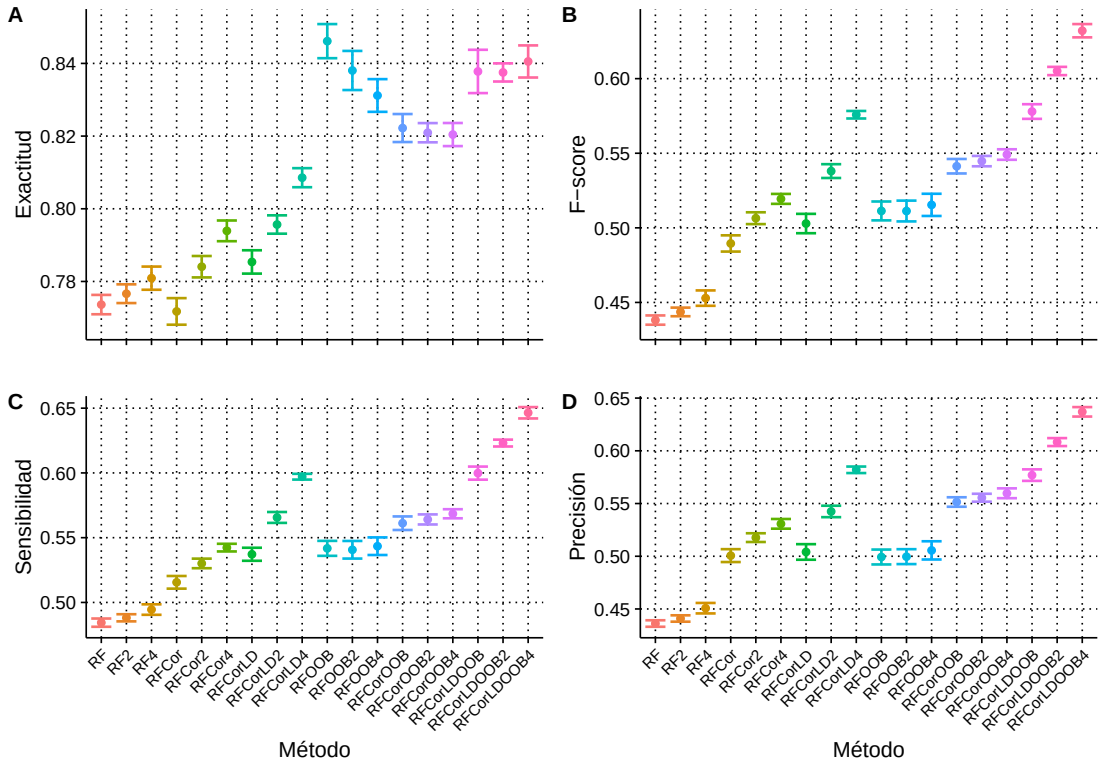

Figura 3.19: Diagrama de puntos para las medidas de desempeño de los métodos evaluados en tres conjuntos de datos, uno con un porcentaje aproximado de SNPs completos de 6, el otro de 12 y el tercero de 24. Se muestra el valor medio (± desvío estándar) para **A)** Exactitud, **B)** F-score, **C)** Sensibilidad y **D)** Precisión.

F-score. En términos generales, es notable ver que todas las metodologías mejoran su desempeño con el aumento de la proporción de SNPs completos.

3.4. Comparación con otras herramientas

Los resultados encontrados en la sección anterior indicaron que las mejores alternativas de la estrategia propuesta fueron *RFCorLD* y *RFCorLDOOB*. Estos algoritmos fueron comparados con tres técnicas de imputación alternativas, *imputación por la moda*, *Beagle* y *LinkImputeR*. La comparación se basó en la evaluación de las medidas de desempeño y la correlación entre éstas y el porcentaje de datos faltantes.

La Figura 3.20 presenta las medidas de desempeño computadas en las cinco metodologías de imputación. El desempeño de los algoritmos *RFCorLD* y *RFCorLDOOB* fue superior al de las alternativas ensayadas en término de las cuatro medidas de desempeño consideradas (valores p de la prueba unilateral de Wilcoxon para datos apareados iguales a 1). Específicamente, en términos de exactitud (Figura 3.20**A**) la

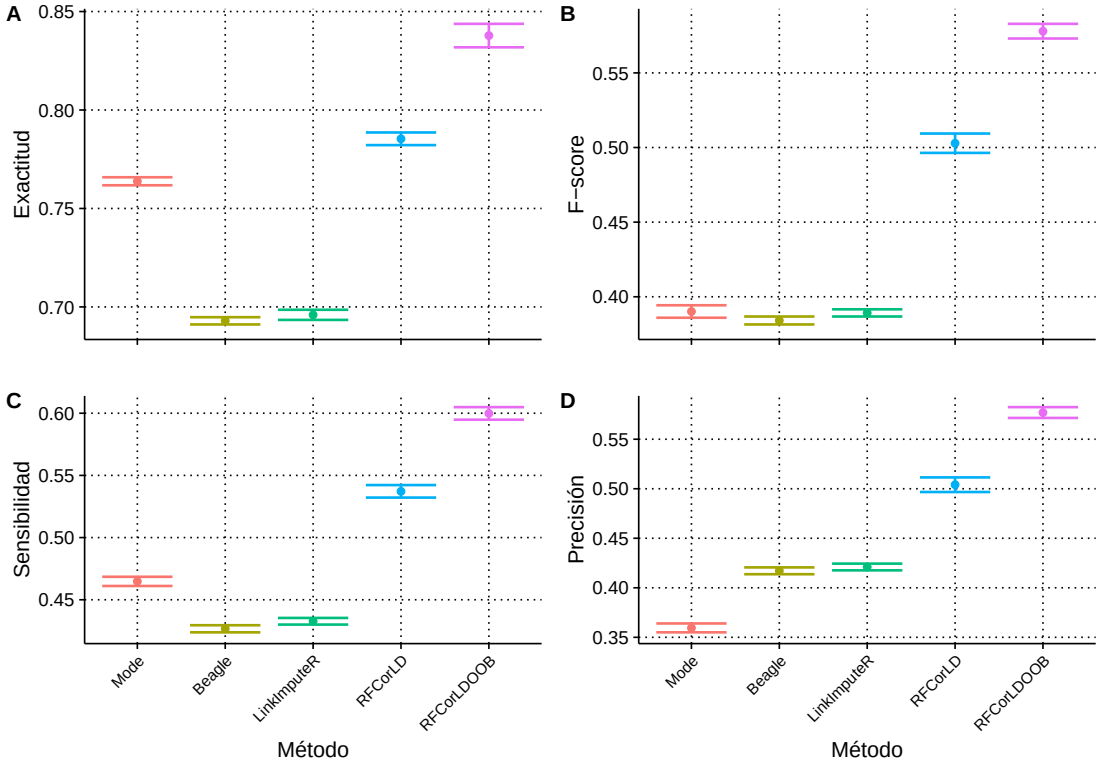

Figura 3.20: Diagrama de puntos para las medidas de desempeño de las dos alternativas desarrolladas en esta tesis y tres métodos alternativos de uso frecuente en la literatura. Se muestra el valor medio (± desvío estándar) para **A)** Exactitud, **B)** F-score, **C)** Sensibilidad y **D)** Precisión.

imputación por la moda fue superior a las otras dos alternativas. A su vez, entre *Beagle* y *LinkImputeR* se encontraron diferencias significativas en la exactitud alcanzada (valores p de la prueba de Wilcoxon para datos apareados de 0,005859). La evaluación del F-score (Figura 3.20**B**) en las tres estrategias alternativas de imputación reveló que los valores más altos fueron para *imputación por la moda* y *LinkImputeR*,

entre los cuales no se encontraron diferencias significativas (valores p de la prueba de Wilcoxon para datos apareados de 0,492). Sin embargo, cuando se inspeccionó separadamente la sensibilidad (Figura 3.20**C**) y la precisión (Figura 3.20**D**) se encontró que si habían diferencias significativas entre estos dos métodos para tales cantidades. En particular, la *imputación por la moda* logró un valor superior de sensibilidad, mientras que *LinkImputeR* fue más preciso. Pese a esto, ninguna de las tres estrategias comparadas con la metodología desarrollada en esta tesis alcanzó valores de sensibilidad, precisión y F-score superiores a 0,47, mientras que éste fue el valor mínimo para las medidas de desempeño obtenidas para *RFCorLD* y *RFCorLDOOB*.

La robustez de las metodologías respecto del porcentaje de genotipos faltantes de los SNPs incompletos se caracterizó a través del coeficiente de correlación de Pearson entre dicho porcentaje y las medidas de desempeño aquí consideradas. La Tabla 3.2 resume los valores medios de dicho coeficiente para las cinco metodologías comparadas. Se destaca que todos las correlaciones fueron negativas. En particular, en términos de exactitud, los métodos resultaron robustos ya que las correlaciones medias observadas resultaron menores a 0,06 (en términos absolutos). En el caso

Tabla 3.2: Coeficiente de correlación de Pearson entre el porcentaje de genotipos faltantes y las medidas de desempeño de las alternativas propuestas. En cada cuadro se muestra el valor promedio seguido del desvío estándar.

Método	***Exactitud***	***Sensibilidad***	***Precisión***	***F-score***
Mode	-0,058 (0,01)	-0,433 (0,011)	-0,406 (0,011)	-0,412 (0,01)
Beagle	-0,033 (0,011)	-0,335 (0,014)	-0,329 (0,014)	-0,354 (0,011)
LinkImputeR	-0,035 (0,01)	-0,336 (0,014)	-0,331 (0,012)	-0,356 (0,011)
RFCorLD	-0,023 (0,011)	-0,326 (0,013)	-0,292 (0,017)	-0,317 (0,015)
RFCorLDOOB	-0,014 (0,019)	-0,326 (0,026)	-0,299 (0,025)	-0,320 (0,026)

de las otras tres medidas, los valores medios absolutos del coeficiente de correlación se encontraron entre 0,2992 y 0,412. La sensibilidad, precisión y el F-score estuvieron más correlacionados con el porcentaje de genotipos faltantes en el caso de las herramientas *Beagle*, *LinkImputeR* e *imputación por la moda*. Particularmente, el desempeño de esta última resultó ser el más influenciado por el porcentaje de datos faltantes. En cambio, las estrategias *RFCorLD* y *RFCorLDOOB* fueron las más robustas ante cambios en dicho porcentaje.

En forma complementaria al análisis anterior, se exploraron las medidas de desempeño de cada una de las cinco metodologías en los grupos de SNPs, categorizados según el porcentaje de genotipos faltantes, como se hizo en la Sección 3.3.5. Los correspondientes diagramas de punto que ilustran estas medidas se muestran en la Figura 3.21. La robustez de los métodos en términos de exactitud y con respecto al porcentaje de dato faltantes es evidente en la Figura 3.21**A**, donde se aprecia valores medios similares entre los distintos grupos de cada metodología. Por el contrario el

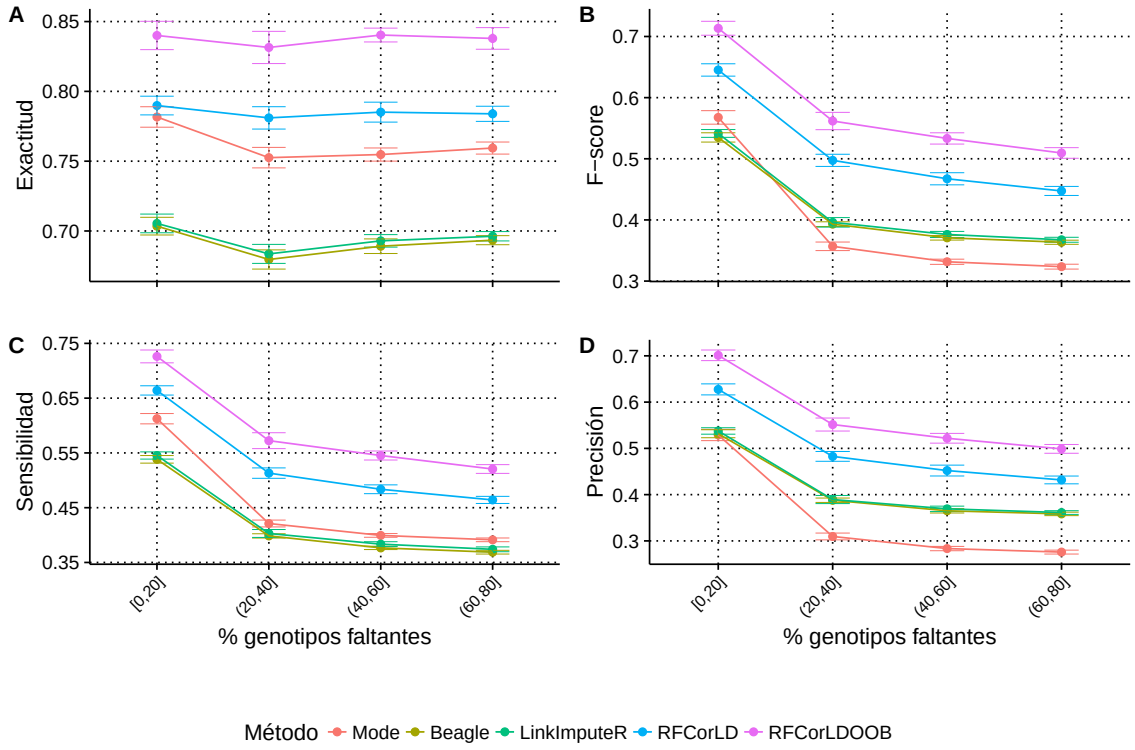

Figura 3.21: Diagrama de puntos y líneas para las medidas de desempeño de los métodos comparados computados en grupos de SNPs definidos según el porcentaje de genotipos faltantes que éstos presentaron. Se muestra el valor medio (± desvío estándar) para **A)** Exactitud, **B)** F-score, **C)** Sensibilidad y **D)** Precisión.

efecto de dicho porcentaje sobre el F-score de la *imputación por la moda* es evidente fundamentalmente al comparar los dos primeros grupos de SNPs (Figura3.21**B**), evidenciando una disminución aproximadamente del 50 % en el valor promedio con

el aumento del porcentaje de SNPs. Incluso, en los grupos de SNPs con más de 20 % de genotipos faltantes, la *imputación por la moda* logró los valores medios de F-score más bajos respecto de *Beagle* y *LinkImputeR* siendo que en el primer grupo ($[0, 20]$ % de genotipos faltantes) se evidenciaron comportamientos opuestos. El análisis de la sensibilidad (3.21**C**) y precisión (3.21**D**) reveló que las medidas de desempeño promedio en las cinco metodologías comparadas mostraron las mismas tendencias a lo largo de los grupos de SNPs.

3.5. Aplicación a datos reales

Los resultados obtenidos a lo largo de las secciones anteriores indican que la metodología más adecuada para la imputación de genotipos faltantes en SNPs incompletos es *RFCorLDOOB*, considerando un valor de significancia igual a 0,05 a la hora de determinar correlación entre SNPs y un umbral de OOB igual a 0,2. Luego, con esta metodología debería ser posible la imputación de al menos el 50 % de los SNPs incompletos. Con el fin de no sólo imputar los genotipos faltantes sino además corroborar que los resultados observados en los datos simulados se correspondan con datos reales, se optó también por imputar la *MG* de girasol con otra de las metodologías de buen desempeño. *RFCorOOB*.

El primer aspecto explorado fue el porcentaje de SNPs incompletos imputados por las metodologías *RFCorOOB* y *RFCorOOBLD*. A diferencia de lo observado en los datos simulados, los porcentajes obtenidos sobre la base de datos de girasol fueron superiores a 75 %. En particular, para el total de SNPs incompletos (25.012) se obtuvieron los siguientes porcentajes de imputación: *RFCorOOB*, 88,5; *RFCorLDOOB*, 76. Al igual que con los datos simulados, sobre esta base de datos también *RFCorLDOOB* obtuvo menor porcentaje que *RFCorOOB*.

La Figura 3.22 ilustra las funciones de densidad empírica estimadas para el número de SNPs completos correlacionados con cada SNP incompleto para las dos metodologías. Consistentemente con lo observado sobre los datos simulados, la función de densidad de *RFCorOOBLD* mostró una moda en valores bajos de número de SNPs correlacionados, con un pico bien pronunciado. Mientras tanto, la metodología *RFCorOOB* evidenció una función de densidad más suave y plana con un moda mayor

que la de *RFCorOOBLD*. Por otro lado, el porcentaje de SNPs incompletos que no correlacionaron con ningún SNP completo fue de 0,024 (6 SNPs) para *RFCorOOB*, mientras que alcanzó el 3,854 % (694 SNPs) para *RFCorOOBLD*. Estos porcentajes fueron ambos inferiores a los respectivos encontrados para los datos simulados, aunque se mantuvo la relación de dos órdenes de magnitud entre ellos. Con *RFCorOOB* todos los SNPs no correlacionados (6) fueron conservados luego del filtrado por *OOB* y MAF, por el contrario, el 25,6 % de los SNPs no correlacionados con *RFCorLDOOB* se encontraron en la MG final.

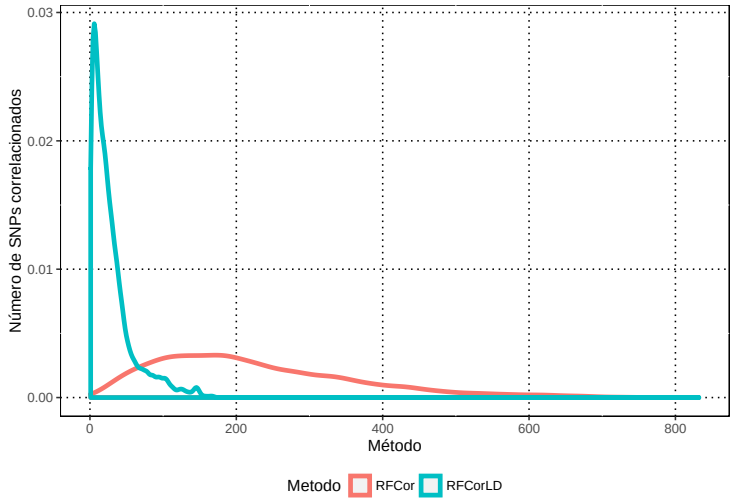

Figura 3.22: Función de densidad del número de SNPs completos correlacionados con cada SNP incompleto.

Finalmente, luego de la imputación y el filtrado por MAF, de los 26.712 SNPs que originalmente contenía la *MG*, el 68 % fueron conservados en la *MG* final obtenida mediante *RFCorLDOOB*, mientras que con *RFCorOOB*, se pudieron conservar el 78,8 % de los mismos. Estos porcentajes representan 18.161 y 21.025 SNPs respectivamente, lo cual claramente supera los 1.700 SNPs completos que se conservarían si simplemente se removieran los SNPs incompletos en vez de optar por una técnica de imputación.

Capítulo 4

Discusión y Conclusiones

En el presente documento de tesis se han aplicado diferentes conocimientos adquiridos durante el cursado de la **Maestría en Estadística Aplicada** de la *Universidad Nacional de Córdoba*. En este sentido, se ha considerado la problemática de la imputación de datos faltantes en el contexto de los genotipificación mediante secuenciación masiva. Ésta, es una dificultad comúnmente encontrada en los planes de mejoramiento vegetal. El experimento que dio origen a esta tesis se ha llevado a cabo en el ámbito del *Instituto Nacional de Tecnología Agropecuaria* (INTA). Durante la exploración de una matriz de genotipos de líneas endocriadas de girasol se determinó la existencia de un elevado porcentaje de marcadores del tipo SNP que no habían sido genotipados en el total de los individuos estudiados. Si bien es posible remover los marcadores con datos incompletos éstos representaron el 93,6 % del total de los SNPs, por lo que su simple eliminación implicaría un pérdida de información relevante. Por tal motivo, se consideró la necesidad de imputar los genotipos de estos SNPs incompletos mediante técnicas estadísticas. A la fecha, existen diferentes técnicas para poder realizar tal tarea. Sin embargo, la mayoría de ellas han sido desarrolladas para trabajar con especies que disponen de genomas de referencia de alta calidad como el de humanos, arroz, maíz, entre otros. La ventaja de estos genomas es que existen muchos estudios previos que han permitido crear paneles de marcadores útiles como referencia a la hora de imputar genotipos faltantes. No obstante, este no es el caso para el genoma de girasol ni para la mayoría de las especies cultivada. Con el fin de aportar soluciones al problema de genotipos faltantes en las matrices

de datos obtenidos por secuenciación de alto desempeño en este tipo de especies, se decidió diseñar una estrategia de imputación basada en técnicas estadísticas. En particular, dada la naturaleza del problema, la matriz de datos sobre la que se debe trabajar contiene muchos más SNPs incompletos que individuos genotipados. Por tal motivo, se decidió tomar la metodología *Random Forest* para la predicción y posterior imputación de los genotipos faltantes. Adicionalmente, dado que se conoce que los SNPs en un genoma están correlacionados, se incorporó tal información con el fin de obtener resultados más precisos. En base a estos principios, se diseñaron seis alternativas de imputación, las cuales fueron evaluadas en datos simulados y luego comparadas con algunas de las estrategias disponibles.

El desarrollo de esta tesis ha derivado en el diseño e implementación de una estrategia de imputación basada en correlación de SNPs y *RandomForests*. En base a esta estrategia, se confeccionaron seis métodos alternativos: *RF*, *RFCor*, *RFCorLD*, *RFOOB*, *RFCorOOB* y *RFCorOOBLD*. Entre ellos, se encontró que aquel que considera correlación entre un SNP incompleto y los SNPs completos del mismo grupo de ligamiento, *RFCorOOBLD*, combinado con un umbral de error *OOB* menor a 0,2 y un nivel de significancia de 0,05 en el test de independencia de SNPs, fue el que mostró mejor desempeño. Si bien las alternativas que no consideran el error *OOB* permitieron recuperar más SNPs incompletos, *RFCorOOBLD* fue superior a todas alternativas propuestas en términos de sensibilidad y precisión. Esta estrategia permitió recuperar miles de SNPs incompletos, logrando que la matriz de genotipos de girasol conserve más del 75 % de SNPs luego de la imputación.

El análisis del impacto de la modificación del umbral de OOB considerado reveló que el umbral elegido, 0,2, permitió obtener el mejor desempeño, logrando un balance entre el porcentaje de SNPs imputados y el máximo error de estimación admitido. En este contexto, si bien la metodología *RFCorOOBLD* logró imputar la mitad de los SNPs que su par *RFCorLD*, lo hizo con mayor sensibilidad y precisión. El efecto de la selección del umbral de significancia del test de independencia de SNPs sobre el desempeño de los métodos evaluados mostró que la estrategia basada en correlación de SNPs del mismo grupo de ligamiento es más robusta que el resto de las metodologías, ya que no evidenció diferencias significativas en el desempeño cuando se modificó dicho umbral. Se encontró además que la metodología *RFCorOOBLD*

fue la menos afectada por las variaciones en el porcentaje de genotipos faltantes. En cuanto la desempeño como función del porcentaje de SNPs completos, esta metodología fue una de las que más incrementó sus medidas como consecuencia del aumento de datos completos. Finalmente, se demostró que la metodología desarrollada resultó superior en desempeño respecto de algunas de las metodologías implementadas en la literatura, comúnmente utilizadas para la imputación de genotipos faltantes, como lo son la *imputación por la moda*, *Beagle* y *LinkImputeR*. Adicionalmente, las medidas de desempeño de las estrategias propuestas en esta tesis fueron más robustas con respecto al porcentaje de SNPs faltantes que las correspondientes a las tres metodologías alternativas de la literatura.

En conclusión, en esta tesis se ha descrito una metodología de imputación de genotipos faltantes que ha mostrado un buen desempeño y robustez ante distintos escenario, superando incluso a los métodos comúnmente utilizados. La herramienta desarrollada permite recuperar miles de SNPs cuyos genotipos en algunos individuos no fueron identificados. Para ello, dichos genotipos faltantes son predichos utilizando la información del resto de los SNPs y focalizando la atención en aquellos que presentan correlación y comparten grupo de ligamiento con el SNP a imputar. La metodología aquí presentada representa un aporte importante al problema de genotipos faltantes en matrices de genotipificación por secuenciación, comúnmente encontrado en programas de mejoramiento vegetal.

Bibliografía

Agarwal, M., Shrivastava, N., and Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant cell reports*, 27(4):617–631.

Badouin, H., Gouzy, J., Grassa, C. J., Murat, F., Staton, S. E., Cottret, L., Lelandais-Brière, C., Owens, G. L., Carrère, S., Mayjonade, B., et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and asterid evolution. *Nature*, 546(7656):148–152.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop science*, 48(5):1649–1664.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, 234(2):177–186.

Browning, B. L. and Browning, S. R. (2016). Genotype imputation with millions of reference samples. *The American Journal of Human Genetics*, 98(1):116–126.

Carranza Astrada, R. P. (2015). *Reconocimiento de caracteres en imágenes no estructuradas*. PhD thesis, Universidad Nacional de Córdoba. Facultad de Matamática, Astronomía y Física.

Catchen, J., Bassham, S., Wilson, T., Currey, M., O'Brien, C., Yeates, Q., and Cresko, W. (2013). Data from: The population structure and recent colonization his-

tory of Oregon threespine stickleback determined using restriction-site associated dna-sequencing. http://datadryad.org/resource/doi:10.5061/dryad.62hb0.

Catchen, J. M., Amores, A., Hohenlohe, P., Cresko, W., and Postlethwait, J. H. (2011). Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, genomes, genetics*, 1(3):171–182.

Chan, A. W., Hamblin, M. T., and Jannink, J.-L. (2016). Evaluating imputation algorithms for low-depth genotyping-by-sequencing (GBS) data. *PloS one*, 11(8):e0160733.

Chen, X. and Ishwaran, H. (2012). Random forests for genomic data analysis. *Genomics*, 99(6):323–329.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, 12(7):499–510.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. Springer, New York. ISBN 978-1-4614-6867-7.

Ekblom, R. and Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity*, 107(1):1–15.

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

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., and Buckler, E. S. (2014). Tassel-gbs: a high capacity genotyping by sequencing analysis pipeline. *PloS one*, 9(2):e90346.

González, M., Andrea, N., et al. (2015). *Identificación y validación de Single Nucleotide Polymorphism (SNPs) distribuidos en el genoma de Eucalyptus globulus*. PhD thesis, Universidad de Concepción. Facultad de Ciencias Forestales.

Gupta, P., Roy, J., and Prasad, M. (2001). Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr Sci*, 80(4):524–535.

Halperin, E. and Stephan, D. A. (2009). SNP imputation in association studies. *Nature biotechnology*, 27(4):349–351.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529.

Jimenez-Escrig, A., Gobernado, I., and Sanchez-Herranz, A. (2012). Whole genome sequencing: a qualitative leap forward in genetic studies. *Revista de neurologia*, 54(11):692–698.

Lado Lindner, B. (2012). *Identificación de SNPs mediante genotipado por secuenciación para el mejoramiento genético de trigo (Triticum aestivum L.)*. PhD thesis, Universidad de la República. Facultad de Ciencias.

Li, Y., Sidore, C., Kang, H. M., Boehnke, M., and Abecasis, G. R. (2011). Low-coverage sequencing: implications for design of complex trait association studies. *Genome research*.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., and Taberlet, P. (2003). The power and promise of population genomics: from genotyping to genome typing. *Nature reviews genetics*, 4(12):981–994.

Mammadov, J., Aggarwal, R., Buyyarapu, R., and Kumpatla, S. (2012). SNP markers and their impact on plant breeding. *International journal of plant genomics*, 2012.

Money, D., Migicovsky, Z., Gardner, K., and Myles, S. (2017). LinkImputeR: user-guided genotype calling and imputation for non-model organisms. *BMC genomics*, 18(1):523.

Monge Ivars, J. F. and Perez, J. A. A. (2017). Estadística no paramétrica; prueba chi cuadrado. [Web; accedido el 22-07-2017].

Morozova, O. and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264.

Pegadaraju, V., Nipper, R., Hulke, B., Qi, L., and Schultz, Q. (2013). De novo sequencing of sunflower genome for SNP discovery using RAD (restriction site associated DNA) approach. *BMC genomics*, 14(1):556.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., and Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, 7(5):e37135.

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

Qi, Y. (2012). Random forest for bioinformatics. In *Ensemble machine learning*, pages 307–323. Springer.

Quail, M. A., Gu, Y., Swerdlow, H., and Mayho, M. (2012). Evaluation and optimisation of preparative semi-automated electrophoresis systems for illumina library preparation. *Electrophoresis*, 33(23):3521–3528.

Rao, D. C. and Gu, C. C. (2008). *Genetic dissection of complex traits*, volume 60. Academic Press.

Roche, A. (2009). *Árboles de decisión y Series de tiempo*. PhD thesis, Universidad de la República. Facultad de Ingeniería.

Rutkoski, J. E., Poland, J., Jannink, J.-L., and Sorrells, M. E. (2013). Imputation of unordered markers and the impact on genomic selection accuracy. *G3: Genes—Genomes— Genetics*, 3(3):427–439.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517.

Schlötterer, C. (2004). The evolution of molecular markers—just a matter of fashion? *Nature Reviews Genetics*, 5(1):63–69.

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145.

Sidransky, D. (2002). Emerging molecular markers of cancer. *Nature Reviews Cancer*, 2(3):210–219.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Torkamaneh, D., Laroche, J., and Belzile, F. (2016). Genome-wide SNP calling from genotyping by sequencing (GBS) data: A comparison of seven pipelines and two sequencing technologies. *PloS one*, 11(8):e0161333.

Tucker, T., Marra, M., and Friedman, J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics*, 85(2):142–154.

Varshney, R. K., Nayak, S. N., May, G. D., and Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in biotechnology*, 27(9):522–530.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.