Universidad Nacional de Rosario

FACULTAD DE CIENCIAS ECONÓMICAS Y ESTADÍSTICA

HERRAMIENTAS ESTADÍSTICAS PARA LA IDENTIFICACIÓN DE ASOCIACIONES ENTRE DATOS GENÓMICOS Y CARACTERÍSTICAS FENOTÍPICAS EN GERMOPLASMA DE DURAZNERO (*Prunus persica* L.)

Doctorado en Estadística

Autora:
Lic. Eugenia Belén Bortolotto

Directores:
Dra. Cecilia Bruno
Dr. Gerardo Cervigni

2026

A mi mamá que me marcó el camino

A mi papá que lo ilumina

A Adrián que camina conmigo

Agradecimientos

Quisiera agradecer a mi mamá y mi papá por ser mis grandes soportes en el transcurrir de la vida. A mis hermanos, Julián y Agustín, quienes me acompañan desde la más tierna edad. A Adrián mi compañero de vida, que más de una vez me sostuvo el corazón en sus manos.

A mi abuela Mande, que cada vez que rendía una materia, ella lo celebraba conmigo.

A la República Argentina por brindarme la oportunidad de estudiar y formarme.

A mi directora de tesis, la Dra. Cecilia Bruno por su inmensa paciencia y dedicación. Por darme siempre una palabra de aliento y transmitirme buenas energías.

A mi director de beca, el Dr. Gerardo Cervigni, quién confió en mi y me dio la oportunidad de comenzar en este camino.

A la Dra. Marta Quaglino por dedicarle su tiempo con tanto amor y pasión a este doctorado.

A mi amiga Mari por sus mil consejos y su amistad incondicional.

A Vero y Pau, mi grupo de apoyo doctoral, que me motivó en momentos difíciles

A Barbi por sus mañanas de tesis y por siempre alentarme a seguir.

A Juli y Gabi por acompañarme los primeros años del doctorado en el CEFOBI.

A mis amores perrunos, en especial Auri y Ramón, que me dan felicidad cada día.

A mis amigas de la escuela, a mis amigas de la vida, a mis cuñadas y a mi familia política, gracias por dejar una huella en mi vida.

Finalmente, a mis compañeras y amigas que descubrí estudiando y ejerciendo esta hermosa profesión: Estadística.

A todos MUCHAS GRACIAS.

Índice general

Índice de cuadros

Índice de figuras

Índice de siglas

A. Ambiente

ACP. Análisis de Componentes Principales

Ad. Aditividad

ANOVA. Análisis de la Varianza, del inglés *ANalysis Of VAriance*

ASI. Intervalo de tiempo transcurrido entre antesis y floración, del inglés *Anthesis-Silking Interval*

BD. Días a floración, del inglés *Bloom Date*

BLUE. Mejor estimador linear insesgado, del inglés *Best Linear Unbiased Estimator*

BLUP. Mejor Predicción Lineal Insesgada¸ del inglés *Best Linear Unbiased Prediction*

CMLM. Modelo Lineal Mixto Comprimido

CP. Componentes Principales

cM. CentiMorgan

DAB. Días después de la floración, del inglés *Days After Bloom*

ECMLM. Modelo Lineal Mixto Comprimido Enriquecido

EGP. Estructura Genética Poblacional

FarmCPU. Unificación de probabilidad circulante de modelo fijo y aleatorio, por sus siglas en inglés *Fixed and random model Circulating Probability Unification*

FDR. Tasa de falsos positivos, del inglés *False Discovery Rate*

FFL. Días a floración femenina, del inglés *Female Flowering*

FN. Falsos negativos

FP. Falsos positivos

G. Genotipo

G-BLUP. BLUP del Genotipo

GLM. Modelo lineal general, del inglés *Generalized Linear Model*

GWAS. Estudio de asociación de genoma completo, del inglés *Genome Wide Association Study*

GY. Rendimiento de grano, del inglés *Grain Yield*

HWE. Equilibrio de Hardy–Weinberg, del inglés *Hardy–Weinberg equilibrium*

IGA. Interacción genotipo-por-ambiente

K. Kinship

Kb. Kilobases

LD. Desequilibrio de ligamiento, del inglés *Linkage Disequilibrium*

LOESS. Regresión no paramétrica localmente ponderada, del inglés *Smoothed Scatterplot Smoothing*

MA. Mapeo asociativo

MAF. Frecuencia alélica menor, del inglés *Minor Allele Frequency*

MFL. Días a floración masculina, del inglés *Male Flowering*

M-F. Base de datos de maíz en las variables relacionadas a la floración

M-GY. Base de datos de maíz en las variables relacionadas a producción de granos

MLM. Modelo Lineal Mixto

MLMM. Modelo Lineal Mixto Multilocus

ML. Máxima Verosimilitud, por sus siglas en inglés (*Maximum Likelihood*

Modelo-P. Modelo con efectos fijos y CP como cofactores

NA. Valor ausente, del inglés *Not Available*

NIPALS. Mínimos Cuadrados Parciales Iterativos No Lineales, del inglés *Nonlinear Iterative Partial Least Squares*

PLS. Mínimos Cuadrados Parciales, del inglés *Partial Least Squares*

QTL. Locus de rasgo cuantitativo, del inglés *Quantitative Trait Locus*

Rc. Rango columna

REML. Máxima Verosimilitud restringida, del inglés *Restricted Maximum Likelihood*

Rf. Rango fila

SIMPLS. Implementación sencilla del PLS basado en matrices, por sus siglas en inglés *Straightforward Implementation of Matrix-based PLS*

SNP. Polimorfismo de un solo nucleótido, del inglés *Single-Nucleotide Polymorphism*

SS. Estrés por sequía, del inglés *drought-stress*

SUPER. Liquidación de MLM Bajo Relación Progresivamente Exclusiva, del inglés *Settlement of MLM Under Progressively Exclusive Relationship*)

SVD. Descomposición del Valor Singular, del inglés *Singular Value Decomposition*

VN. Verdaderos negativos

VP. Verdaderos positivos

WW. Bien regado, deel inglés *well-watered*

Resumen

La amplia disponibilidad de herramientas estadísticas representa una oportunidad para analizar datos aplicando diferentes algoritmos y poder obtener mayor información sobre un mismo conjunto de datos. Seleccionar el análisis estadístico según la pregunta de investigación requiere conocer acerca de la metodología que las mismas aplican e interpretar sus resultados. La generación de datos en el ámbito del mejoramiento genético vegetal es constante y de alta dimensión. Este gran volumen de información conlleva el desafío de poder procesarla de manera eficiente para obtener resultados interpretables. Los modelos de asociación en estudios de todo el genoma han sido ampliamente estudiados con el objetivo de encontrar señales que identifiquen los genes que producen determinadas características de interés en las plantas. Esta expresión del fenotipo, no solo se encuentra regida por la información genética si no también por otros factores como las condiciones donde se desarrolla el cultivo y la estructura genética que existe en los datos. En el capítulo 1, se introducen conceptos generales que se desarrollan en cada uno de los capítulos subsiguientes. En el Capítulo 2 se aborda el estudio de la eficiencia de modelos GWAS para detectar verdaderas asociaciones genotipo–fenotipo bajo heterocedasticidad ambiental. Se realizó un estudio de simulación y sus resultados fueron evaluados con bases de datos públicas a través de ocho modelos que contemplaron el análisis loci a loci y multiloci. Los resultados mostraron que la heterocedasticidad puede enmascarar los efectos medios del genotipo. El modelo MLMM permitió distinguir adecuadamente la señal molecular del ruido ambiental, resultando el más parsimonioso. El Capítulo 3 aborda la aplicación de la regresión por PLS, una técnica multivariada que modela simultáneamente marcadores genotípicos y rasgos fenotípicos, siendo especialmente útil frente a multicolinealidad. Además de evaluar la capacidad del modelo para detectar asociaciones relevantes con valor agronómico, se compararon dos algoritmos para la construcción de componentes latentes: SVD y NIPALS. Esta comparación permitió analizar ventajas computacionales y diferencias en la

estabilidad de las soluciones, aportando evidencia sobre la elección metodológica más adecuada según el contexto de los datos. En el Capítulo 4 se utilizó el algoritmo NIPALS dentro del marco PLS para evaluar asociaciones entre caracteres fenotípicos y SNP con datos faltantes. Se evaluó la performance de PLS-NIPALS en escenarios con diferentes niveles de datos faltantes, evaluando la estabilidad del modelo mediante validación cruzada. Los resultados indicaron que, aunque el error de predicción aumenta con la proporción de valores faltantes, PLS mantiene capacidad de identificar asociaciones significativas, lo que evidencia su flexibilidad para estudios GWAS aún en condiciones de datos incompletos. Todos los códigos generados para el presente trabajo de tesis se encuentran disponibles en bortolottoeugenia-gif, 2025 y en el Anexo de esta tesis. En conjunto, la tesis aporta evidencia sobre la influencia de la heterogeneidad ambiental en GWAS, la utilidad de PLS como herramienta multivariada para la asociación genómica y destaca la robustez de sus algoritmos ante escenarios con información faltante, contribuyendo así al avance de metodologías estadísticas aplicadas al mejoramiento genético vegetal.

Abstract

The wide availability of statistical tools represents an opportunity to analyze data by applying different algorithms and to obtain more information from the same dataset. Selecting the statistical analysis according to the research question requires knowledge about the methodology they apply and the interpretation of their results. Data generation in the field of plant genetic improvement is constant and high-dimensional. This large volume of information entails the challenge of processing it efficiently to obtain interpretable results. Association models in genome-wide studies have been widely studied with the aim of finding signals that identify the genes responsible for certain traits of interest in plants. This expression of the phenotype is governed not only by genetic information but also by other factors such as the conditions under which the crop develops and the genetic structure present in the data. In Chapter 1, general concepts that are developed in each of the subsequent chapters are introduced. Chapter 2 addresses the study of the efficiency of GWAS models to detect true genotype–phenotype associations under environmental heteroscedasticity. A simulation study was carried out and its results were evaluated with public databases through eight models that considered both loci-by-loci and multilocus analyses. The results showed that heteroscedasticity can mask the average effects of the genotype. The MLMM model adequately distinguished the molecular signal from environmental noise, proving to be the most parsimonious. Chapter 3 deals with the application of PLS regression, a multivariate technique that simultaneously models genotypic markers and phenotypic traits, being especially useful in the presence of multicollinearity. In addition to evaluating the model's ability to detect relevant associations with agronomic value, two algorithms for constructing latent components were compared: SVD and NIPALS. This comparison made it possible to analyze computational advantages and differences in the stability of the solutions, providing evidence on the most appropriate methodological choice depending on the data context. In Chapter 4, the NIPALS algorithm was used within

the PLS framework to evaluate associations between phenotypic traits and SNP with missing data. The performance of PLS-NIPALS was assessed in scenarios with different levels of missing data, evaluating model stability through cross-validation. The results indicated that, although prediction error increases with the proportion of missing values, PLS maintains the ability to identify significant associations, demonstrating its flexibility for GWAS studies even under incomplete data conditions. All codes generated for this thesis are available in bortolottoeugenia-gif, 2025 and in the Appendix of this thesis. Overall, the thesis provides evidence on the influence of environmental heterogeneity in GWAS, the usefulness of PLS as a multivariate tool for genomic association, and highlights the robustness of its algorithms in scenarios with missing information, thereby contributing to the advancement of statistical methodologies applied to plant genetic improvement.

Capítulo 1

Introducción general

1.1. La importancia del mapeo genético en la agricultura

El mapeo genético en plantas constituye una herramienta esencial en la genética moderna y en los programas de mejoramiento, ya que permite identificar y localizar genes o loci de carácter cuantitativo (QTL, por sus siglas en inglés *Quantitative Trait Loci*) asociados a rasgos de interés agronómico. A través del uso de marcadores moleculares, como los *Single-Nucleotide Polymorphisms* (SNP), es posible establecer relaciones entre la variabilidad genotípica y la expresión fenotípica de características como la resistencia a enfermedades, la tolerancia a estreses bióticos y abióticos, y la productividad (Collard et al., 2005). Este enfoque ha optimizado los procesos de selección mediante la implementación de la selección asistida por marcadores, reduciendo significativamente el tiempo y los costos asociados a los métodos tradicionales de mejoramiento (Y. Xu y Crouch, 2008). Con el advenimiento de las tecnologías de secuenciación de nueva generación, la densidad y resolución de los mapas genéticos se ha incrementado de manera sustancial, posibilitando análisis de asociación de genoma completo (GWAS, por sus siglas en inglés *Genome-wide Association Studies*) y mapeo de ligamiento con mayor precisión (Zhu et al., 2008). En conjunto, estas herramientas han contribuido a una comprensión más profunda de la arquitectura genética de los rasgos cuantitativos y a la utilización eficiente de los recursos genéticos en la generación de cultivares mejorados y adaptados a diferentes condiciones ambientales (Tanksley y McCouch, 1997). Los caracteres de herencia cuantitativa suelen tener distribución normal y están muy influenciados por el ambiente en el cual se desarrollan los individuos. Una metodología utilizada para identificar los caracteres cuantitativos asociados a las características agronómicas de interés es el Mapeo Asociativo (MA) basado en el principio del desequilibrio de ligamiento (LD, por sus siglas en inglés *Linkage Disequilibrium*). Los GWAS

permiten encontrar las regiones genómicas involucradas en el control de la variación de los caracteres agronómicos (Kulwal y Singh, 2021). De esta manera, los GWAS surgen como una alternativa para analizar un conjunto de genotipos que presenten la máxima variabilidad genética para el carácter agronómico de interés, sin necesidad de haber generado una población de mapeo o de conocer su correlación genética. Sin embargo, a pesar de que no se conozcan de manera explícita las correlaciones genéticas, éstas pueden estar presentes generando asociaciones espurias (falsos positivos) si no son tenidas en cuenta en el análisis (Price et al., 2006).

1.2. Bases del mapeo genético

Los dos métodos de mapeo más utilizados para analizar caracteres cuantitativos son el mapeo de ligamiento (mapeo de QTL) y el MA (S. A. Flint-Garcia et al., 2005). Ambos métodos se basan en el LD, que se refiere a la asociación no aleatoria de alelos que se encuentran ubicados en diferentes loci. Sin embargo, la diferencia entre ambos enfoques radica en que en el mapeo de ligamiento, el LD utilizado se genera mediante el diseño de cruzamientos entre individuos obteniendo así lo que se denomina una población de mapeo (Lu et al., 2010), mientras que en el MA el LD está presente en el conjunto de germoplasma en estudio y los métodos de análisis buscan identificar aquellos marcadores moleculares asociados, de manera no aleatoria, con el caracter fenotípico de interés (Brachi et al., 2011). A pesar de que el MA se presenta como un método con menor potencia estadística para detectar alelos poco frecuentes o interacciones epistáticas que el mapeo por QTL, sus ventajas como la capacidad de obtener una mayor resolución de mapeo y la disposición de un número potencialmente mayor de alelos evaluados (Yu et al., 2006), sumado a que no requiere el desarrollo de poblaciones biparentales, reduciendo tiempo y costo, ha dado como resultado que el MA sea ampliamente utilizado sobre el mapeo por QTL.

Diversos factores evolutivos y poblacionales pueden generar LD o mantenerlo a lo largo del tiempo determinando cómo se distribuye en el genoma y su extensión. Los procesos que lo afectan son: (i) tipo de apareamiento, (ii) deriva genética, (iii) selección, (iv) mutación, (v) subestructura y parentesco poblacional, y (vi) sesgo de verificación

(Clark et al., 2005; Stich et al., 2005; Yu et al., 2006). Dado que, para la mayoría de las poblaciones de MA, la importancia de estos factores se desconoce o solo puede estimarse de forma aproximada, la extensión del LD no puede derivarse teóricamente, sino que debe inferirse a partir de estudios empíricos basados en marcadores moleculares para evaluar la aplicabilidad y la resolución de estos enfoques. Sin embargo, para interpretar los resultados de los estudios empíricos del LD, es esencial comprender las propiedades de las medidas utilizadas. El patrón que este presenta en el genoma refleja la historia recombinacional y demográfica de la población, por lo que su extensión y velocidad de decaimiento difieren entre especies, poblaciones y regiones genómicas (Slatkin, 2016).

1.3. Estimación del desequilibrio de ligamiento

Para la estimación del LD, se propusieron varias estadísticas (Hedrick, 1987). Sin embargo, las dos medidas más utilizadas en un contexto de MA son r^2 y D' (Lewontin, 1964; Hill y Robertson, 1968). Las cuales consideran únicamente loci bialélicos. La medida D' (Lewontin, 1964) toma valores entre -1 y 1 y es útil para detectar la ausencia de recombinación histórica (valores de D' cercanos a -1 o 1 indican falta de recombinación entre sitios), pero es sensible a frecuencias alélicas bajas, por lo que puede resultar elevado aun cuando la correlación real entre loci sea pequeña (Zapata, 2000). La medida más usada es r^2. Para loci bialélicos, r^2 se define como la correlación cuadrada entre las variables alélicas y puede ser estimado de la siguiente manera:

$$r^2 = \frac{D^2}{p_A(1-p_A)p_B(1-p_B)} \tag{1.1}$$

donde, $D = p_{AB} - p_A p_B$, representa la diferencia entre la frecuencia de las gametas que llevan el par de alelos A y B para estos dos loci (p_{AB}) y el producto de las frecuencias alélicas de cada uno de estos alelos (p_A y p_B). Los valores de r^2 varían entre 0 y 1 y se interpreta como la proporción de varianza explicada de un locus por otro (es decir, cuánto de la variación alélica en el locus causal queda explicada por el marcador). Un valor de $r^2 = 1$ es posible sólo si la frecuencia alélica de dos loci es idéntica (S. Flint-Garcia et al., 2003). La significancia estadística, a través del valor p, para el LD se estima a través

del test exacto de Fisher (Fisher, 1935) para comparar sitios con dos alelos en cada locus o para un análisis de permutación que compara sitios con más de dos alelos. Los estadísticos r^2 y D', donde $D' = \frac{D}{D_{max}}$, reflejan diferentes aspectos del LD y pueden comportarse de manera diferentes bajo ciertas condiciones, por ejemplo bajo situaciones donde las mutaciones ocurren en diferentes linajes alélicos, lo cual reflejaría la misma historia de recombinación pero con diferentes historias de mutación, provocando un valor de $D' = 1$ pero un valor de r^2 menor (S. Flint-Garcia et al., 2003). Es importante recalcar que en poblaciones con tamaños de muestra pequeños y/o frecuencias alélicas pequeñas, ninguno de estos dos estadísticos son buenos estimadores, sin embargo cada uno tiene sus propias ventajas. Mientras que r^2 resume tanto la historia de mutación como la de recombinación, D' mide solamente la historia de recombinación y podría considerarse, desde un punto de vista estadístico, más preciso para estimar diferentes recombinaciones. Sin embargo, D' es sensible o se ve afectado de manera fuerte por tamaños de muestra pequeños, estimando valores erráticos cuando compara loci con baja frecuencia alélica. Este comportamiento errático se debe a que disminuye la probabilidad de encontrar todas las combinaciones de los alelos de baja frecuencia polimórfica aún si los loci no están ligados. Por esta razón, para estudios de asociación, se recomienda usar el estadístico r^2. Matemáticamente está directamente relacionada con la potencia estadística para detectar asociaciones: cuanto mayor r^2 entre marcador y variante causal, mayor la potencia del estudio para detectar un efecto de tamaño dado (Hill y Robertson, 1968). En términos prácticos, un marcador con un valor de correlación de $r^2 = 0.8$ con la variante causal, explica el 80 % de la varianza genética debida a esa variante y por tanto permite detectar su efecto con casi la misma potencia que si se hubiese genotipado la variante causal directamente. Además, r^2 es directo de interpretar en modelos lineales y logísticos usados en GWAS.

1.4. El desequilibrio de ligamiento en el mapeo por asociación

El primer paso es confirmar la presencia de LD a través de la estimación de la tasa de decaimiento del LD con la distancia física o genética para extrapolar la información

obtenida de un pequeño conjunto de loci muestreados en todo el genoma investigado. Esta extrapolación es esencial para el diseño de estudios de MA, ya que puede utilizarse para determinar la densidad de marcadores necesaria para analizar regiones del genoma previamente inexploradas, así como la resolución máxima que se puede alcanzar para las asociaciones entre los marcadores moleculares y el fenotipo en la población en estudio. El funcionamiento del LD en mapeo genético se interpreta bajo la premisa de que, si un QTL influye en la expresión de un caracter fenotípico, las variantes (marcadores moleculares entre los que se encuentran los SNP) que estén en alto LD con ese QTL, mostrarán asociación estadísticamente significativa con el QTL en la población estudiada. De tal manera que, genotipar un conjunto denso de marcadores y buscar asociaciones marcador-fenotipo permite localizar regiones del genoma que contienen variantes causales (Abdurakhmonov y Abdukarimov, 2008). La extensión del LD determina la resolución del MA: en poblaciones con LD extenso (i.e., poco decaimiento con la distancia), se necesitan menos marcadores pero la región asociada será amplia (baja resolución); por otro lado, en poblaciones con LD corto (i.e., rápido decaimiento), se necesita mayor densidad de marcadores para cubrir el genoma, pero la resolución puede ser muy fina (posibilidad de localizar variantes causales con precisión). En especies autógamas o con historia de cuellos de botella es frecuente que el LD sea más extendido que en especies alógamas con alta recombinación (S. Flint-Garcia et al., 2003). Varios investigadores se han centrado en la distancia a la que los valores r^2 promedio se reducen a $0,10$ como un punto razonable para concluir que existe un LD mínimo para detectar asociaciones con QTL. Un ejemplo de la curva de decaimiento de LD en función de la distancia entre pares de loci puede observarse en la Figura 1.1, una línea horizontal marca el umbral de correlación al cuadrado (r^2) en $0,2$, indicando una distancia de LD de aproximadamente 175 kb de distancia física entre loci. Cabe señalar que la relación entre distancia física (kilobases - kb) y distancia genética (centiMorgan - cM) no es constante entre especies ni a lo largo del genoma. Por ejemplo, en duraznero se ha estimado que 1 cM equivale aproximadamente a entre 100 y 200 kb, mientras que en maíz el promedio es considerablemente mayor, del orden de 500 a 1000 kb por cM. Estas diferencias reflejan variaciones en las tasas de recombinación y deben considerarse al interpretar la extensión del LD y al planificar la densidad de

marcadores.

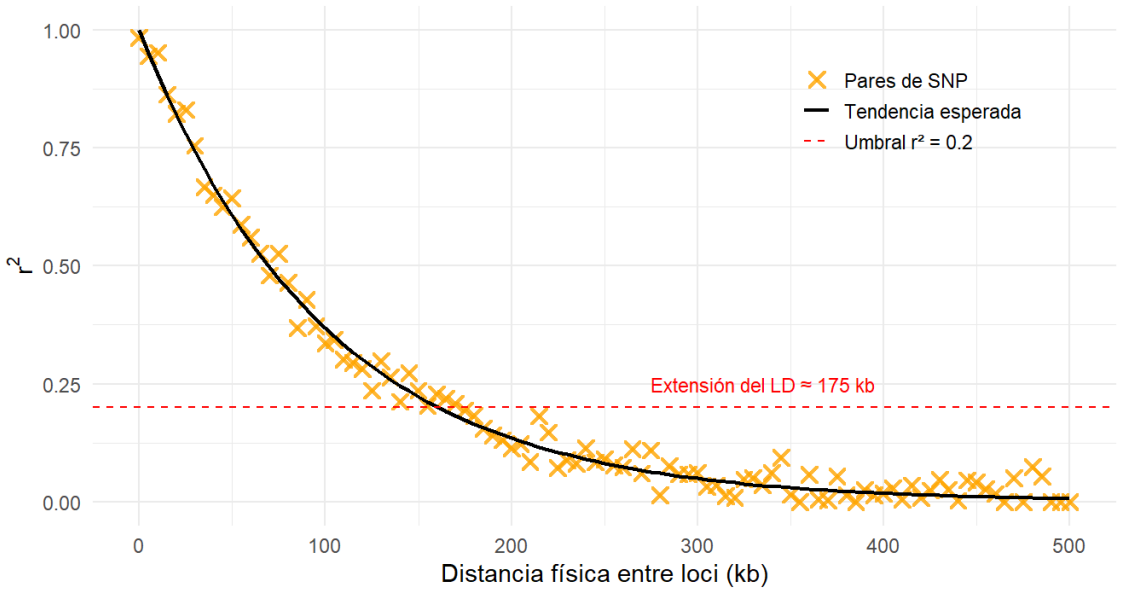

Figura 1.1: Ilustración del decaimiento del desequilibrio de ligamiento (LD) medido como r^2 en función de la distancia física entre loci. Este ejemplo representa cómo el LD disminuye a medida que aumenta la distancia genética entre loci considerando un umbral de $r^2 = 0.2$ como el valor que suele emplearse para estimar la extensión efectiva del LD (175 kb en este ejemplo)

1.5. Impacto de SNP ausentes sobre el desequilibrio de ligamiento

La presencia de datos ausentes o no disponibles, representan un desafío recurrente en el análisis multivariado, especialmente en contextos genómicos y fenotípicos donde la obtención de información suele ser costosa, parcial o técnicamente restringida (B. L. Browning y Browning, 2009). Tradicionalmente, esta problemática se aborda mediante técnicas de imputación, como la sustitución por la media, la regresión múltiple o mediante la eliminación directa de observaciones incompletas (de Souto et al., 2015); sin embargo, estas estrategias pueden introducir sesgos o disminuir la varianza explicada, comprometiendo la calidad del análisis posterior.

Como ya fue mencionado, en los estudios genómicos, el LD constituye una medida fundamental que describe la asociación no aleatoria entre alelos que se encuentran en distintos loci dentro de una población. El LD refleja la historia evolutiva, las fuerzas de selección y la estructura poblacional, y es un componente clave para determinar la

densidad de marcadores necesaria en análisis de asociación, mapeo fino y selección genómica (S. Flint-Garcia et al., 2003; Hill y Robertson, 1968). Sin embargo, la estimación precisa del LD depende críticamente de la calidad y completitud de los datos genotípicos. En este contexto, la presencia de datos faltantes de SNP, ya sea por fallas en la lectura de secuencias, baja cobertura, errores de alineamiento o exclusión de loci con baja calidad, puede generar sesgos significativos en las estimaciones de LD, afectando la interpretación biológica y estadística de la estructura genómica (Laurie et al., 2010; S. Xu, 2013).

Los datos faltantes en matrices de SNP se originan principalmente en las limitaciones técnicas de las plataformas de genotipado y en las características del ADN analizado. En tecnologías como Genotyping-by-Sequencing o RAD-seq, el bajo número de lecturas por sitio y la distribución desigual de la cobertura entre individuos contribuyen a la presencia de una alta proporción de genotipos ausentes (Elshire et al., 2011; Glaubitz et al., 2014). Este patrón de datos faltantes rara vez es completamente aleatorio, pues puede depender del contenido de Guanina y Citocininia, la complejidad de la región o la presencia de repeticiones, generando una estructura no aleatoria que sesga las correlaciones alélicas entre loci (O'Connell et al., 2014). Así, el patrón de ausencia de datos no solo reduce la información disponible, sino que también puede inducir una subestimación o sobrestimación del LD, dependiendo de cómo se manejan los valores ausentes y del método de imputación aplicado (Yao et al., 2020).

El efecto directo de los datos faltantes sobre las estimaciones del LD se manifiesta principalmente en dos aspectos: la reducción del número efectivo de pares de genotipos disponibles para calcular las frecuencias haplotípicas y la alteración de las frecuencias alélicas marginales. Cuando la proporción de datos ausentes es elevada, las estimaciones de r^2 (uno de los parámetros más usados para cuantificar LD) pueden volverse inestables y sesgadas, especialmente si los valores faltantes se concentran en determinados individuos o grupos genéticos (Weir y Hill, 1986; S. Xu, 2013). En estos casos, el desequilibrio observado en lugar de cuantificar la coherencia entre loci, podría reflejar artefactos derivados de un muestreo desigual o de la exclusión de pares con información incompleta. Este sesgo es particularmente relevante en poblaciones con estructura, donde la pérdida de datos puede afectar diferencialmente a subpoblaciones con distinta

frecuencia alélica, amplificando las señales de LD espurias (Mangin et al., 2012).

Para mitigar estos problemas, los estudios de genómica vegetal y animal recurren con frecuencia a estrategias de imputación de datos faltantes, que buscan reconstruir los valores de los genotipos ausentes a partir de la información de loci vecinos y patrones de LD conocidos (B. L. Browning y Browning, 2009). Métodos como Beagle, IMPUTE2 o FImpute utilizan modelos de haplotipos o algoritmos de inferencia bayesiana para predecir los alelos más probables, incrementando así la densidad efectiva de marcadores y la precisión en las estimaciones del LD (B. Howie et al., 2012; Rutkoski et al., 2013). Sin embargo, el éxito de la imputación depende de la calidad del panel de referencia, del nivel de diversidad genética y de la densidad inicial de SNP. En poblaciones con alta diversidad, como muchas especies autógamas o alógamas de plantas cultivadas, los errores de imputación pueden introducir correlaciones artificiales entre loci, generando una inflación artificial del LD que distorsiona las inferencias sobre el tamaño de bloque haplotípico o el decaimiento del LD con la distancia física (Money et al., 2015).

La imputación inadecuada o el exceso de datos faltantes no tratados puede también afectar el análisis de MA y la selección genómica. Un LD subestimado tiende a reducir la capacidad de detectar asociaciones genuinas entre marcadores y loci causales (verdaderos positivos), disminuyendo la potencia de los estudios GWAS; mientras que un LD sobrestimado puede inducir asociaciones espurias (falsos positivos) y dificultar la delimitación de regiones causales (Yu et al., 2006; S. Xu, 2013). Además, la pérdida de precisión en la estimación del decaimiento del LD limita la capacidad para diseñar paneles de marcadores eficientes y optimizar el número de SNP necesarios para cubrir el genoma con suficiente resolución (Vos et al., 2017). En consecuencia, la gestión adecuada de los datos faltantes, incluyendo la evaluación de su patrón de distribución y la selección del método de imputación apropiado, constituye un paso crítico en la construcción de matrices genotípicas confiables y en la correcta interpretación de los patrones de LD.

Estudios recientes en genómica vegetal han demostrado que la proporción y el patrón de datos faltantes pueden modificar significativamente las estimaciones del decaimiento del LD en diferentes especies. Por ejemplo, estudios en maíz, soja y trigo muestran que la imputación con paneles de referencia genéticamente distantes tiende

a sobreestimar la extensión del LD, mientras que la imputación basada en individuos de la misma población conserva con mayor fidelidad la estructura local de haplotipos (Pasam et al., 2012; Y. Zhang et al., 2018; Yao et al., 2020). En este sentido, la elección del enfoque de imputación y la proporción máxima de datos ausentes admisible antes de la imputación deben adaptarse al tipo de especie, sistema de reproducción y objetivo del estudio. Así, comprender y controlar el impacto de los datos faltantes sobre el LD no sólo mejora la calidad de los estudios de asociación, sino que también optimiza el aprovechamiento de la información genómica en programas de mejoramiento asistido por marcadores y selección genómica (Rutkoski et al., 2013; Yao et al., 2020).

1.6. Modelos estadísticos empleados en el mapeo por asociación

El GWAS utiliza modelos estadísticos diseñados para detectar asociaciones entre caracteres fenotípicos cuantitativos o cualitativos y marcadores moleculares, considerando la estructura genética poblacional y las relaciones de parentesco entre los individuos. La presencia de estratificación poblacional o parentesco no controlado puede generar falsos positivos (asociaciones espurias) debido a la correlación entre marcadores y fenotipos que no se debe a efectos genéticos causales, sino a diferencias de frecuencia alélica entre subpoblaciones (Pritchard et al., 2000; Yu et al., 2006). Para mitigar este problema se desarrollaron modelos que incorporan matrices que describen la estructura genética y el parentesco genómico. Las tres más utilizadas son la matriz Q, la matriz P y la matriz K. La matriz Q (estructura poblacional) es derivada de métodos de agrupamiento bayesianos implementados en software como STRUCTURE (Pritchard et al., 2000) y FastSTRUCTURE (Lippert et al., 2011). La matriz Q representa la probabilidad de pertenencia de cada individuo a diferentes subpoblaciones genéticas. Su inclusión como covariable en Modelos Lineales Generales (GLM, por sus siglas en inglés General Linear Models) permite controlar parcialmente la estratificación poblacional, reduciendo la tasa de falsos positivos. Sin embargo, el uso exclusivo de Q no considera las relaciones de parentesco entre individuos dentro de una misma subpoblación. La matriz P (componentes principales) es obtenida mediante Análisis de Componentes

Principales (ACP) y resume la variación genética a través de ejes ortogonales (Price et al., 2006). Cada Componente Principal (CP) representa gradientes de variación genética asociados con estructura poblacional o aislamiento por distancia. La inclusión de los primeros CP como covariables en el modelo tiene un efecto similar al uso de la matriz Q, pero con menor costo computacional y sin asumir un número discreto de subpoblaciones a priori. Finalmente, la matriz K (parentesco genómico), también llamada matriz Kinship, cuantifica el grado de relación genética entre pares de individuos, si se conoce su pedigree o en base a marcadores genómicos en caso de que no se conozca la información de sus padres. Al incluir esta matriz en un Modelo Lineal Mixto (MLM), se modela la covarianza entre individuos atribuible a efectos poligénicos compartidos (Yu et al., 2006). Esto permite distinguir entre asociaciones debidas al efecto real del marcador y las debidas al parentesco, mejorando la precisión de las estimaciones. El MLM que incorpora como covariables la matriz de estructura genética poblacional Q y la matriz de parentesco genético K, también conocido como Modelo Q+K o modelo de asociación mixta, fue propuesto por Yu et al. (2006) como una extensión de los GLM. Su formulación general es:

$$Y = X\beta + Zu + \varepsilon \tag{1.2}$$

donde, Y es el vector de valores fenotípicos observados de dimensión Nx1 con N igual a cantidad de observaciones, X es la matriz de diseño de efectos fijos (que incluye los marcadores y las covariables Q o P), β es el vector de coeficientes fijos, Z es la matriz de diseño para los efectos aleatorios, u es el vector de efectos aleatorios distribuidos como $u \sim \mathcal{N}\left(0, K\sigma_u^2\right)$ y ε es el vector de errores aleatorios, distribuídos como $\varepsilon \sim \mathcal{N}\left(0, I\sigma_e^2\right)$ El modelo MLM-Q+K, y su variante MLM-P+K que utiliza CP en lugar de la matriz Q, se convirtieron en modelos estándares en estudios de asociación en plantas debido a su eficacia para controlar la inflación de la significancia estadística y aumentar la robustez de las asociaciones detectadas (Yu et al., 2006; Stich et al., 2008). En síntesis, el uso conjunto de las matrices K y Q (o P) en modelos MLM representa una estrategia robusta para reducir las asociaciones espurias (falsos positivos) y aumentar la precisión en la detección de loci asociados a rasgos complejos. En especies vegetales con historia evolutiva diversa, autogamia o estructura genética poblacional, estos

modelos son esenciales para asegurar inferencias confiables y replicables. Numerosos métodos de estimación para GWAS basados en MLM han sido propuestos, como EMMA, FaST-MLM, GEMMA, GRAMMAR, EMMAX y P3D. EMMA utiliza la descomposición espectral para simplificar el procedimiento de cálculo, que ha demostrado ser más eficiente que el método MLM (Kang et al., 2008). Zhou y Stephens (2012) propusieron un método mejorado denominado GEMMA cuya velocidad computacional es n veces de magnitud más rápida que EMMA, donde n representa la cantidad de individuos. Sin embargo, requiere datos genotípicos completos o imputados. Lippert et al. (2011) desarrollaron el método FaST-MLM, donde seleccionan un subconjunto de marcadores moleculares para capturar el efecto poligénico en un tiempo computacional reducido (alta velocidad). Los métodos anteriores se conocen como métodos exactos porque la varianza poligénica se vuelve a estimar con cada marcador. GRAMMAR primero estima los residuos bajo el modelo nulo y luego trata los residuos como fenotipos para un análisis de asociación adicional utilizando MLM. En general, los MLM permiten flexibilizar los supuestos distribucionales del GLM al poder incorporar efectos aleatorios, considerar correlaciones y/o permitir varianzas heterogéneas. Nguyen y McLachlan (2016) propusieron un MLM novedoso al suponer que los efectos aleatorios siguen una distribución desconocida y los errores asociados con las respuestas son gaussianos. Si bien, se han propuesto numerosos modelos, aún es necesario continuar investigando el desempeño de los mismos sobre otros cultivos cuya estructura genética poblacional puede cambiar, como es el caso de un cultivo frutal, perenne y leñoso como el duraznero. Además, es importante bajo este escenario, evaluar el desempeño de modelos contemporáneos que permitan analizar la asociación contemplando otras distribuciones en las correlaciones genéticas.

1.7. GWAS con marcadores moleculares que contienen datos faltantes

La presencia de datos faltantes en matrices de SNP constituye un problema frecuente en estudios genómicos de plantas, estos datos ausentes suelen surgir cuando los marcadores moleculares no están cubiertos por la plataforma de genotipado utilizada,

especialmente cuando se emplean técnicas de secuenciación de baja cobertura como Genotyping-by-Sequencing (GBS) o Restriction-site Associated DNA sequencing (RAD-seq). Estas metodologías buscan reducir la complejidad del genoma para genotipar miles de marcadores de manera rentable, pero inherentemente producen un conjunto de datos disperso y con genotipos faltantes que requieren un procesamiento riguroso. Frente a este desafío existen dos estrategias conceptuales: (1) imputar los genotipos faltantes antes del análisis de asociación para reconstruir una matriz completa de marcadores y (2) incorporar la incertidumbre genotípica directamente en el modelo de asociación (por ejemplo, usando probabilidades o dosajes), muchas veces no se cuenta con esta información (Elshire et al., 2011). Sin embargo, a pesar de los avances en las herramientas de imputación la naturaleza incierta de los genotipos imputados introduce un sesgo inevitable en el análisis posterior. Z. Zhang et al. (2021) demostraron que la precisión de la imputación disminuye especialmente para alelos minoritarios (baja frecuencia alélica), y que las discordancias entre los genotipos observados y los imputados pueden inflar los valores p en un GWAS, generando falsos positivos. El uso de imputación dependerá de la proporción de datos ausentes, de la existencia de paneles de referencia apropiados y del objetivo del estudio, es decir, si se busca controlar el Error de Tipo I o aumentar la potencia del estudio. El Error de tipo I se refiere a la probabilidad de identificar asociación estadísticamente significativa entre un marcador molecular y la expresión fenotípica cuando en realidad no lo está. Dicho de otra manera, encontrar asociaciones falsas o espurias (falsos positivos). La potencia del estudio está relacionada con la capacidad de detectar los verdaderos positivos. La potencia es el complemento a uno de la probabilidad de cometer Error de Tipo II que en términos de asociación son identificados como los r falsos negativos, dicho de otra manera, la falta de potencia provoca la no detección de las verdaderas asociaciones y ello se traduce en un aumento del Error de tipo II. En general, la literatura recomienda combinar filtrados de marcadores moleculares usando herramientas como la frecuencia alélica mínima (MAF, por sus siglas en inglés *Minor Allele Frequency*), que indica la frecuencia del alelo menos común en la población y funciona como un criterio clave para evaluar la variabilidad de cada marcador molecular o SNP. Las matrices SNP utilizadas en GWAS suelen incluir marcadores con una amplia gama de MAF, desde variantes casi monomórficas

(MAF < 0,5 %) hasta variantes muy comunes (MAF cercano a 50 %). Sin embargo, los loci con MAF bajao (<10 %) poseen una capacidad menor para detectar asociaciones débiles en comparación con aquellos con MAF alto (>40 %) (Ardlie et al., 2002). Diversos estudios han demostrado que los genotipos raros (bajo MAF) son más propensos a generar resultados espurios o inestables (Lam et al., 2007). Por este motivo, muchos GWAS optan por eliminar SNP con MAF <10 % como parte de sus procedimientos de control de calidad (Florez et al., 2007; Tabangin et al., 2009). Además del filtrado por MAF, también se recomienda aplicar imputación cuando la tasa de genotipos faltantes lo permite (se recomienda que el porcentaje de datos faltantes no supere el 10 %) y el uso de tests que acepten dosajes o probabilidades para preservar la información de incertidumbre cuando sea relevante (S. R. Browning, 2008; B. N. Howie et al., 2009).

Los métodos de imputación basados en haplotipos son los más usados en poblaciones con estructura genética y cuando existe un panel de referencia representativo. Herramientas como Beagle (haplotype clustering), IMPUTE2 y Minimac (modelos de haplotipos y algoritmo de MCMC/EM) reconstruyen segmentos haplotípicos y predicen alelos faltantes con alta precisión cuando la densidad de marcadores y la similitud poblacional son adecuadas; esto reduce el sesgo en las estimaciones de LD y aumenta la potencia de GWAS (Y. Li et al., 2009). Para especies no modelo o conjuntos con ausencia de panel de referencia, han surgido métodos específicos como LinkImpute, que usan algoritmos basados en redes neuronales del tipo kNN y LD local diseñados para datos GBS de baja cobertura; estos métodos son prácticos y suelen ofrecer buen rendimiento cuando la población es homogénea y la tasa de datos faltantes moderada (Money et al., 2015; Swarts et al., 2014).

1.8. Mapeo asociativo en diferentes ambientes: Interacción Genotipo-Ambiente

El MA en ambientes variados considera la interacción genotipo-ambiente (IGA), donde los genotipos se comportan de manera diferente en diversos ambientes, afectando el comportamiento del caracter de interés, como por ejemplo, el rendimiento y la adaptación de los mismos a determinadas características ambientales. Modelos

multiplicativos capturan IGA no cruzada (cambios en escala) y cruzada (cambios en rango), considerando matrices de varianza genética entre ambientes. Análisis multiambientales mejoran el poder de detección de QTL en comparación con estudios de un solo ambiente, identificando QTL estables o específicos. Numerosas propuestas se han realizado para modelar las estructuras varianzas y covarianzas heterogéneas, donde cada ambiente posee su propia varianza (Huang y Han, 2014). Esta formulación mejora la estimación de efectos genéticos y la eficiencia del test de asociación, al ponderar cada observación según su precisión ambiental.

1.9. Heterocedasticidad entre ambientes y su relevancia en estudios GWAS en plantas

En estudios de genética cuantitativa y mejoramiento genético vegetal, la heterogeneidad de varianza entre ambientes constituye un aspecto frecuentemente subestimado que puede tener efectos sustanciales sobre la precisión de los estimadores y la potencia estadística de los análisis GWAS. La heterocedasticidad entre ambientes se refiere a la presencia de varianzas no homogéneas entre localidades, años o diferentes condiciones experimentales, fenómeno común en ensayos multiambientales (MET) debido a diferencias en el control experimental, variabilidad climática o IGA. En tales contextos, asumir homocedasticidad puede conducir a una estimación sesgada de los errores estándar y, por ende, a un incremento en la tasa de falsos negativos o a una reducción de la potencia para detectar loci de efecto real (Huang y Han, 2014;Al Kawam et al., 2018; Shi, 2022).

En los estudios GWAS, la heterocedasticidad adquiere una relevancia particular porque la magnitud del efecto de un marcador puede variar según la expresión fenotípica en cada ambiente. Un mismo SNP puede mostrar una asociación significativa en ambientes con baja variabilidad residual, pero pasar inadvertido en ambientes con alta varianza, enmascarando interacciones ambientales o efectos condicionales (Braz et al., 2021). La incorporación de una función para modelar la estructura de varianzas y covarianzas puede afectar la estimación de componentes genéticos y ambientales y, como consecuencia modificar la interpretación de la arquitectura genética de los

QTL (Rönnegård et al., 2010). Por ello, el modelado explícito de la heterocedasticidad se ha convertido en una estrategia clave para mejorar la robustez de los estudios de asociación genómica en plantas.

La incorporación de estructuras de varianza heterogéneas y de IGA explícitas en los modelos de asociación mejora notablemente la precisión y la interpretabilidad de los resultados de GWAS en plantas. La heterocedasticidad entre ambientes no debe considerarse una simple violación de supuestos estadísticos, sino una fuente de información biológicamente relevante que refleja diferencias en la estabilidad y plasticidad genotípica. El uso de MLM que permiten flexibilizar supuestos distribucionales e incorporar, estructuras de varianza ambiental específicas constituye hoy una estrategia esencial para avanzar en la comprensión de los mecanismos genéticos que determinan la adaptación y la estabilidad de los cultivos (Corty y Valdar, 2018).

1.10. Regresión PLS y su aplicación en la detección de asociaciones fenotipo-genotipo

El análisis de asociación entre variación fenotípica y genotípica en plantas enfrenta el desafío de la naturaleza multivariada de los caracteres agronómicos de interés. La mayoría de los caracteres agronómicos presentan correlaciones fenotípicas y genéticas como resultado de pleiotropía, LD y vías metabólicas compartidas. Los métodos tradicionales de asociación univariada, como los MLM, suelen analizar cada variable fenotípica de manera independiente, ignorando dicha estructura de correlación y reduciendo la potencia para detectar loci que afectan simultáneamente múltiples caracteres. En este contexto, la regresión por Mínimos Cuadrados Parciales (PLS por sus siglas en inglés *Partial Least Squares Regression*) ofrece una alternativa eficiente, ya que permite modelar simultáneamente un conjunto de variables predictoras (SNP) y múltiples respuestas (variables fenotípicas) maximizando la covarianza entre ambos conjuntos de datos (H. Wold, 1985; Abdi, 2010).

El método PLS se fundamenta en la construcción de componentes latentes o variables proyectadas, que son combinaciones lineales de los predictores y de las variables respuestas originales (variables fenotípicas). Estas componentes se calculan de modo

que capturen la mayor covarianza posible entre los espacios genotípicos y fenotípicos, permitiendo extraer la estructura subyacente de asociación multivariante entre SNP y variables fenotípicas. A diferencia del ACP, que busca la máxima varianza dentro del conjunto de predictores sin considerar las respuestas, el PLS está orientado a la relación predictiva entre ambos conjuntos de variables. En el marco del mejoramiento genético y la biología de sistemas, esto se traduce en la posibilidad de identificar combinaciones de marcadores moleculares que explican simultáneamente variaciones coordinadas en múltiples caracteres agronómicos (Boulesteix y Strimmer, 2007; Colombani et al., 2012). Una de las ventajas principales del PLS en estudios genómicos es su capacidad para manejar alta dimensionalidad, es decir, situaciones donde el número de variables predictoras (SNP) excede ampliamente al número de observaciones (individuos). Esto es típico de los estudios de genómica vegetal, donde pueden evaluarse cientos de miles de marcadores en relativamente pocas líneas o genotipos. Los métodos de regresión tradicionales, como la regresión lineal múltiple cuya método de estimación es por mínimos cuadrados ordinarios, colapsan en este escenario debido a la multicolinealidad entre predictores y la falta de grados de libertad. En cambio, el PLS reduce el espacio de predictores a un número pequeño de componentes ortogonales, evitando el sobreajuste y conservando la información estructural de los datos genómicos (Mevik y Wehrens, 2007; Westerhuis et al., 1998).

El algoritmo más comúnmente utilizado para estimar los componentes PLS es el NIPALS (por sus siglas en inglés *Non-linear Iterative Partial Least Squares Algorithm*), desarrollado por Herman Wold en la década de 1960. El NIPALS constituye un procedimiento iterativo que permite calcular los componentes latentes de manera secuencial, sin requerir la descomposición completa de las matrices de predictores y respuestas. En cada iteración, el algoritmo estima un par de vectores de pesos,uno para los predictores (X) y otro para las respuestas (Y), que maximizan la covarianza entre las proyecciones de los datos, denominadas scores (P y Q, respectivamente). Posteriormente, se deflactan las matrices originales eliminando la información explicada por el componente extraído, y el proceso se repite hasta alcanzar el número deseado de componentes (Geladi y Kowalski, 1986). Este número de componentes suele denominarse raíz o solución. La cantidad de raíces o componentes debe ser igual al mínimo número cardinal entre la

cantidad de variables predictoras y variables respuestas (Stocchero et al., 2018). En el Capítulo 3 del presente trabajo de tesis se discute con mayor detalle.

La simplicidad y eficiencia computacional del NIPALS lo hacen particularmente adecuado para conjuntos de datos incompletos o de alta dimensionalidad, como ocurre frecuentemente en matrices genotípicas con SNP faltantes o en experimentos multi-fenotípicos con datos incompletos. A diferencia de otros algoritmos basados en descomposición de valores singulares (SVD, por sus siglas en inglés *Singular Value Decomposition*), el NIPALS puede trabajar directamente con valores faltantes mediante estimación iterativa, lo que evita la imputación previa y conserva la variabilidad original del sistema (Mevik y Wehrens, 2007). Esta propiedad es relevante en el contexto del GWAS multivariado, ya que permite incorporar más individuos y caracteres en el análisis sin sacrificar información. Finalmente, la interpretación biológica de los resultados del PLS puede realizarse a través de las cargas y los scores de los componentes latentes, que indican la contribución relativa de cada SNP y cada caracter fenotípico al patrón de covariación detectado. Esta estructura facilita la identificación de grupos de marcadores que afectan simultáneamente múltiples caracteres fenotípicos, permitiendo inferir posibles mecanismos pleiotrópicos o regiones genómicas con efectos coordinados sobre la fisiología de la planta. De esta manera, la regresión PLS, implementada a través del algoritmo NIPALS, representa un enfoque robusto y flexible para detectar asociaciones fenotipo-genotipo complejas considerando la correlación entre caracteres fenotípicos y la estructura multivariante de los datos (S. Wold et al., 2001; Mevik y Wehrens, 2007; Rohart et al., 2017).

1.11. Tipo de datos fenotípicos en el mapeo asociativo

Los datos fenotípicos pueden clasificarse en cuantitativos y cualitativos, según la naturaleza del carácter medido. Los fenotipos cuantitativos como por ejemplo el rendimiento, la altura de planta y el peso de fruto, se expresan como variables continuas y son influenciadas por múltiples genes y por el ambiente, por ello son generalmente denominadas caracteres de tipo poligénico o cuantitativo (Falconer y Mackay, 1996). En contraste, los datos feontípicos cualitativos presentan una segregación discreta y

suelen estar controlados por uno o pocos loci de efecto mayor (Collard et al., 2005). En el contexto de la genómica moderna, los fenotipos pueden incluir además mediciones de caracteres fisiológicos, bioquímicos o morfoagronómicos, así como información derivada de fenotipado de alto rendimiento (high-throughput phenotyping, HTP). Esta última estrategia, basada en sensores ópticos, imágenes multiespectrales y plataformas automatizadas, permite capturar datos fenotípicos masivos, multidimensionales y de alta resolución temporal (Araus et al., 2018; W. Yang et al., 2020).La calidad, precisión y representatividad de los datos fenotípicos determinan en gran medida la potencia estadística para detectar asociaciones fenotipo-genotipo y la confiabilidad de las inferencias genéticas (Collard et al., 2005).

1.12. Análisis de datos fenotípicos

El análisis de los datos fenotípicos implica una serie de procedimientos estadísticos y en muchos casos provienen de estudios experimentales que responden a un diseño de experimento con el propósito de reducir el "ruido" ambiental y obtener estimaciones robustas del valor genético de los individuos o genotipos evaluados. En los ensayos multiambientales (MET), es habitual aplicar MLM para ajustar efectos fijos y efectos aleatorios), que permiten estimar valores promedios ajustados (BLUE, por sus siglás en inglés *Best Linear Unbiased Estimator*) en el caso de los factores considerados como efecto fijo o estimar el predictor lineal del valor genético (BLUP, por sus siglas en inglés *Best Linear Unbiased Prediction*) (Piepho et al., 2008; Smith et al., 2005). El uso de BLUP es particularmente ventajoso en mapeo por asociación (GWAS) y estudios de selección genómica, ya que permite separar de manera más precisa los efectos genéticos de la variación ambiental y de la estructura experimental (Resende et al., 2012). Asimismo, la homogeneidad de varianzas entre ambientes y la ausencia de heterocedasticidad son requisitos fundamentales para evitar sesgos en la detección de asociaciones, por lo que se emplean procedimientos de análisis de varianza (ANOVA), pruebas de Levene o modelos heterocedásticos mixtos (Garrick y Fernando, 2013).Dado que muchos caracteres agronómicos están correlacionados, el uso de enfoques multivariantes se ha vuelto indispensable. Métodos como el ACP, la regresión PLS y los modelos estruc-

turales multivariantes permiten identificar combinaciones de rasgos fenotípicos que maximizan la variación explicada o que se asocian significativamente con marcadores genómicos (Abdi, 2010; S. Wold et al., 2001). Estas estrategias son especialmente útiles en el mapeo de caracteres complejos y en la detección de QTL pleiotrópicos. Finalmente, la calidad de los datos fenotípicos depende del diseño experimental, la repetibilidad de las mediciones y la gestión de valores perdidos o atípicos. Los outliers o valores atípicos pueden identificarse mediante el análisis de residuos estandarizados o mediante algoritmos robustos (Rousseeuw y Leroy, 1987). Los valores faltantes pueden imputarse usando métodos basados en la media ponderada, la regresión multivariante o técnicas más avanzadas como imputación múltiple o *machine learning* (Buuren y Groothuis-Oudshoorn, 2011).

1.13. Simulación de datos que representan el genoma de plantas y su expresión fenotípica

La simulación de genomas de plantas en la biología computacional y el mejoramiento genético vegetal, ha permitido modelar procesos genéticos complejos en un periodo de tiempo más corto respecto al tiempo que llevaría estudiar una alta cantidad de procesos biológicos, que además de tiempo, requieren capacitación de recurso humano y contar con recursos económicos para conducir dichos ensayos. La simulación de datos genotípicos y fenotípicos constituye una herramienta esencial en la genética estadística moderna. Permite explorar escenarios hipotéticos, comparar metodologías de análisis, y optimizar diseños experimentales antes de su implementación real. Además, la creciente disponibilidad de paquetes en R y Python facilita su integración directa en flujos de análisis reproducibles, ampliando las posibilidades para la investigación en mejoramiento genético y modelización estadística en plantas.

En el contexto del mejoramiento genético vegetal, la simulación de datos aborda desafíos relacionados con la complejidad de los genomas, la estructura de las poblaciones y la interacción genotipo-ambiente. Los programas de simulación permiten modelar la herencia de loci causales, la segregación de marcadores moleculares, y la expresión fenotípica resultante de efectos genéticos aditivos, dominantes y epistáticos.

A su vez, posibilitan incorporar procesos como selección, cruzamientos controlados o aleatorios, mutación, migración y deriva genética, reflejando la dinámica evolutiva de las poblaciones de plantas (E. Wang et al., 2019).

Existen múltiples herramientas desarrolladas con estos fines. QMSim (Sargolzaei y Schenkel, 2009) es uno de los programas más utilizados para simular poblaciones a gran escala, permitiendo generar fenotipos y genotipos bajo esquemas complejos de cruzamiento y selección, tanto en poblaciones animales como vegetales. *XSim* (Cheng et al., 2015) y *Xbreed* (Esfandyari y Sørensen, 2019) amplían estas funcionalidades con mayor flexibilidad para definir estructuras poblacionales, múltiples rasgos correlacionados y efectos ambientales. Por su parte, AlphaSimR (Gaynor et al., 2021) proporciona un marco integrado en el software R para simular programas de mejoramiento genético, con control detallado de la arquitectura genética de los rasgos, el diseño experimental y los métodos de selección, siendo ampliamente empleado para validar modelos de selección genómica y GWAS. Más recientemente, se han desarrollado herramientas que incorporan arquitecturas genéticas más realistas y flujos de trabajo reproducibles en entornos de programación estadística. Por ejemplo, sim1000G y HAPGEN2 permiten generar datos de genotipos a partir de haplotipos reales, preservando el desequilibrio de ligamiento observado en poblaciones naturales, mientras que librerías como PhenotypeSimulator (Meyer y Birney, 2018) facilitan la creación de fenotipos simulados con componentes genéticos, ambientales y de ruido controlados. En una revisión reciente, Shang et al. (2024) destacaron la diversidad de enfoques disponibles para la simulación de datos genómicos y comparan programas representativos como ms, msprime, SLiM, SimuPOP, HAPGEN2, GWAsimulator, GENOME y ForSim, resaltando que SLiM y SimuPOP ofrecen el mayor nivel de realismo y flexibilidad para estudios modernos de interacción genómica.

1.14. Hipótesis

1. Los modelos GWAS multilocus pueden detectar asociaciones fenotipo-genotipo con mayor eficiencia (menor tasa de falsos positivos) que los modelos locus a locus, incluso en presencia de heterocedasticidad entre ambientes.

2. La regresión multivariada por PLS permitirían detectar asociaciones fenotipo-genotipo más eficientes al incorporar la correlación entre los caracteres fenotípicos.

3. El algoritmo NIPALS, en el marco del método PLS, tendría un desempeño estable y robusto para detectar asociaciones fenotipo-genotipo frente a la presencia de datos faltantes en los marcadores moleculares.

1.15. Objetivo general

Contribuir al desarrollo y perfeccionamiento de herramientas estadísticas para la identificación de asociaciones entre características fenotípicas de interés agronómico en germoplasma de cultivos vegetales a partir de datos genómicos masivos.

1.15.1. Objetivos específicos

1. Evaluar y comparar el desempeño de distintos procedimientos analíticos destinados a la detección de asociaciones fenotipo-genotipo, incorporando información sobre la estructura genética poblacional y/o el parentesco entre individuos, bajo condiciones de heterocedasticidad entre ambientes.

2. Diseñar y desarrollar herramientas de visualización basadas en técnicas multivariadas de ordenación de genotipos, que permitan la representación gráfica efectiva y la interpretación intuitiva de asociaciones fenotipo-genotipo.

3. Aplicar enfoques multivariados para caracterizar la variabilidad genética y detectar asociaciones fenotipo-genotipo, considerando la robustez de los métodos frente a la presencia de datos faltantes en marcadores moleculares.

4. Implementar y validar rutinas de software de código abierto que faciliten la aplicación de los modelos estadísticos y analíticos propuestos, asegurando su reproducibilidad y utilidad en estudios de mapeo genético y mejoramiento genético vegetal.

Para dar respuesta a los objetivos antes planteados, en el Capítulo 2 se abordó la comparación y evaluación de modelos de asociación GWAS en su eficiencia para detectar verdaderas asociaciones (verdaderos positivos) entre genotipo y fenotipo bajo diferentes niveles de variabilidad entre ambientes. En el Capítulo 3 se implementó el método multivariado de regresión PLS, con los algoritmos algebraicos SVD y NIPALS, y se propuso un método para seleccionar los marcadores estadísticamente significativos asociados a múltiples caracteres agronómicos de manera simultánea. Además, en este Capítulo se presenta un gráfico triplot para visualizar la regresión PLS e identificar los marcadores moleculares asociados a los caracteres agronómicos presentes en los individuos con mejor comportamiento para dichas variables fenotípicas. En Capítulo 4, se consideró el algoritmo multivariado PLS en el contexto de datos faltantes para el estudio de asociación de múltiples caracteres agronómicos con variables explicativas del tipo binarias. La visualización del ordenamiento de los marcadores en un plano factorial es un desafío. Por ello, tanto en el Capítulo 3 como en el Capítulo 4 ilustramos con bases de datos públicas, provenientes de ensayos, el análisis completo, la visualización de sus resultados y la interpretación, en términos prácticos, de los mismos. Todos los códigos de análisis estadísticos desarrollados en la presente Tesis, se encuentran disponibles en el Anexo y en el GitHub **bortolotto2025codigo**.

Capítulo 2

Evaluación del efecto de la heterocedasticidad ambiental en la identificación de asociación fenotipo-genotipo

2.1. Modelos de asociación fenotipo-genotipo

La asociación del fenotipo con el genotipo, también conocida como mapeo genético, se usa en el mejoramiento de cultivos (X. Yang et al., 2017; Racedo et al., 2016; Bouchet et al., 2017). El enfoque más utilizado para el mapeo genético es el MA, también llamados GWAS. Este método tiene una serie de ventajas sobre otras técnicas de mapeo genético, incluido el potencial para una mayor resolución de QTL y un mayor muestreo de la variación molecular a través del polimorfismo genético provisto por los marcadores moleculares (Buckler y Thornsberry, 2002). Sin embargo, la probabilidad de identificar asociaciones espurias es mayor en los estudios de MA en comparación con los estudios tradicionales de mapeo de QTL (Yu y Buckler, 2006). El MA se utiliza ampliamente en estudios sobre plantas (Kulwal y Singh, 2021), especies modelo de animales (Brachi et al., 2010; Al Kawam et al., 2018) y genética humana (Nordborg y Tavaré, 2002; Tak y Farnham, 2015). Los caracteres cuantitativos de las plantas están controlados por muchos genes e influenciados por el medio ambiente. Se encuentran disponibles una vasta cantidad de modelos estadísticos para identificar asociaciones entre loci de marcadores y caracteres fenotípicos medidos. Los caracteres o variables fenotípicas evaluadas pueden responder a caracteres simples o complejos. A medida que los datos genotípicos se encuentran accesibles, decodificar con precisión la complejidad de los rasgos fenotípicos en una población diversa sólo es posible a través del ajuste de modelos estadísticos que pueden distinguir las asociaciones biológicas verdaderas de aquellas asociaciones que no lo son y que se denominan falsos positivos. Estas asociaciones falsas positivas pueden

estar relacionadas con la presencia de estructura genética poblacional causada por la relación de parentesco entre los individuos (Pritchard et al., 2000; Yu et al., 2006, Videla et al., 2021). En otras palabras, la presencia de falsos positivos, significa haber identificado como significativa mayor cantidad de asociaciones entre el fenotipo y el genotipo que las que realmente existen. En los modelos GWAS, los falsos positivos representan una de las principales preocupaciones y pueden ser atribuidos, en parte, a asociaciones espurias causadas por la estructura poblacional (Z. Zhang et al., 2010). Sin embargo, tener en cuenta solamente la estructura poblacional en los modelos puede conducir a un control inadecuado de los falsos positivos o una pérdida de potencia en los modelos debido a las relaciones de parentesco existente entre los individuos (Yu et al., 2006). Por el contrario, una falta de potencia nos llevaría a no identificar asociaciones verdaderas, en este caso, se denominan falsos negativos. La estructura genética poblacional fue abordada usando GLM a través de la incorporación de una matriz de parentesco calculada a partir de un método de agrupamiento bayesiano que estimaba la probabilidad de que dos individuos fueran idénticos por azar (Pritchard et al., 2000). Una propuesta adicional frecuente es el uso de CP de los datos genéticos como covariables del modelo de mapeo (Patterson et al., 2006). En este modelo, conocido como Modelo-P, se sintetiza la estructura genética de la población en el conjunto de CP obtenidas del ACP (He et al., 2008) realizado sobre la matriz de frecuencias alélicas. Con el ACP es posible obtener un conjunto de nuevas variables, que se genera como una combinación lineal de los datos moleculares originales denominadas variables sintéticas o CP. Tales variables tienen la característica de no estar correlacionadas y son óptimas para señalar variabilidad y estructuras entre los genotipos de la población mapeada. Así, las CP formadas a través de la combinación de marcadores moleculares de los genotipos de la población mapeada, nos permiten señalar diferencias entre genotipos causadas por la existencia de estructuración genética. Dado que las matrices de marcadores moleculares son demasiado grandes, generan muchas CP y se sabe que las de mayor orden tienen información espuria. Para determinar el número de CP significativos se utiliza la prueba de Tracy-Widom (Tracy y Widom, 1994). Este es el método más común para detectar CP significativas (Bruno et al., 2019). Posteriormente, los MLM, a través de la incorporación de la información de relación genética entre los

individuos como efectos aleatorios, demostraron una mejora en la estimación de las verdaderas asociaciones entre el fenotipo y el genotipo (Z. Zhang et al., 2010; S.-B. Wang et al., 2016).

Otra estrategia de modelado, también factible en el contexto del ajuste de un MLM, es el uso de la matriz K, que se incorpora a la estructura de varianza y covarianza del MLM calculada a partir de la información de parentesco entre los individuos, la cual define el grado de covarianza genética entre ellos (Yu et al., 2006). El parentesco entre dos individuos es igual al coeficiente de consanguinidad que tendría un descendiente suyo independientemente de si los dos individuos se han apareado realmente. Boyce (1983) revisa los dos principales enfoques de cálculo para los coeficientes de consanguinidad y parentesco con un análisis de senderos (*path analysis*) y con algoritmos recursivos. Cuando no se conoce el verdadero parentesco entre los individuos, se estima una matriz de aditividad (Ad) a partir de los marcadores moleculares, que se denomina matriz de similitud. Los modelos basados en MLM también pueden inducir falsos negativos debido a un ajuste excesivo del modelo en el que se pueden pasar por alto algunas asociaciones potencialmente importantes (Liu et al., 2016). El MLM con CP y K (MLM-PK) extiende la estructura de correlaciones del Modelo-P añadiendo, además, una matriz K para modelar la estructura genética de la población. Estos modelos realizan el análisis de asociación locus a locus, esto significa que comprenden un escaneo del genoma unidimensional al probar un marcador a la vez de forma iterativa para cada marcador en un conjunto de datos. Este enfoque de un solo locus no coincide con el verdadero modelo genético de rasgos complejos que están controlados por numerosos loci simultáneamente. Para hacer frente a este problema, se han recomendado modelos de asociación multilocus porque estos modelos consideran la información de todos los loci simultáneamente (S.-B. Wang et al., 2016). Bajo esta premisa, se han propuesto distintos modelos de asociación multilocus. El MLM comprimido (CMLM) optimiza la eficiencia computacional y estadística mediante la agrupación de individuos genéticamente similares. Esta estrategia se ve mejorada en el CMLM enriquecido (ECMLM), que ajusta dinámicamente los grupos para mejorar la detección de asociaciones. El modelo SUPER (por sus siglas en inglés *Settlement of MLM Under Progressively Exclusive Relationship*) selecciona un subconjunto de marcadores asociados para definir relaciones entre in-

dividuos y aumentar el poder estadístico. El modelo multivariado MMLM permite la detección simultánea de múltiples loci asociados, mientras que el FarmCPU (por sus siglas en inglés *Fixed and Random Model Circulating Probability Unification*) combina modelos de efectos fijos y de efectos aleatorios de manera iterativa para mejorar la separación entre los efectos verdaderos (verdaderos positivos) y los efectos espurios (falsos negativos) causados por la estructura poblacional y el LD.

2.2. Heterocedasticidad a través de los ambientes y su efecto en la identificación de la asociación fenotipo-genotipo

En ensayos conducidos en más de un ambiente con el objetivo de comparar el comportamiento de varios genotipos para algún caracter agronómico de interés (caracter fenotípico), los modelos GWAS se utilizan para identificar los marcadores asociados con un alto valor fenotípico en todos los ambientes. En este tipo de ensayos, se espera que los modelos estadísticos permitan estimar la contribución relativa de los efectos del genotipo (G) y de la interacción genotipo-por-ambiente (IGA) a la varianza total de los caracteres fenotípicos. Los MLM (West et al., 2014) han sido ampliamente utilizados para estimar los componentes de varianza de los efectos G y de la IGA (Balzarini, 2002; R.-C. Yang, 2007). Además de la estrategia utilizada para modelar la correlación entre genotipos, es común proceder con el GWAS utilizando el promedio de valores fenotípicos en todos los ambientes. Sin embargo, un modelo con IGA se puede utilizar para seleccionar marcadores y genotipos cuyos caracteres agronómicos de interés se observan en condiciones ambientales variables y la información del marcador molecular sólo una vez. Aunque este tipo de análisis ha tenido éxito en descubrir una amplia variedad de asociaciones genéticas, la detección de cambios en la media explica solo una pequeña parte de la varianza total. En algunos casos, la heterocedasticidad existente que se origina en las exposiciones de los genotipos a distintas condiciones ambientales conduce a un sesgo en la prueba conjunta de marcadores moleculares y en la interacción marcador-ambiente, pero no afecta la prueba marginal de los marcadores moleculares (Jin y Shi, 2021; Al Kawam et al., 2018). Este problema, si se ignora, puede

llevar a la invalidación de las pruebas de interacción conjunta en los estudios GWAS. La incorporación de funciones que modelan la heterocedasticidad beneficia la reducción de falsos positivos en la detección de asociaciones fenotipo-genotipo (Segura et al., 2012; Malosetti et al., 2013).

Una estrategia efectiva para manejar la heterocedasticidad consiste en emplear un MLM que incorpore la estructura de covarianza entre individuos y ambientes, proporcionando estimaciones más estables y menos sesgadas de los valores fenotípicos ajustados (Möhring y Piepho, 2009). De esta manera, se obtiene una medida del desempeño de cada genotipo que contempla la variabilidad ambiental y la correlación entre observaciones, lo que posibilita utilizar dichos valores como una variable respuesta más robusta para los análisis de asociación (Crossa et al., 2010). Dicha medida es conocida como el mejor predicción lineal insesgada o BLUP presentado por primera vez por Henderson (1950). Un valor predicho del comportamiento del G basado en BLUP representa el valor ajustado del rendimiento observado de un genotipo de acuerdo con la cantidad de información que tenemos sobre ese individuo en particular, que incluye la información de su desempeño a través de los ambientes (A) donde ha sido evaluado (Ekine et al., 2014). Si bien algunos estudios han utilizado el valor genético del individuo como la puntuación del caracter fenotípico para GWAS, asumiendo que abarca la mejor estimación del mérito genético de un individuo (Johnston et al., 2011; Becker et al., 2013; Čepica et al., 2013), generalmente se asume homocedasticidad. La contribución de este capítulo es proponer distintas estrategias de análisis y comparar su desempeño en la estimación de asociación fenotipo-genotipo, en cuanto a la mejora de la eficiencia en la detección de verdaderas asociaciones disminuyendo los falsos negativos, frente a escenarios de heterocedasticidad entre ambientes.

2.3. Objetivo

Evaluar y comparar el desempeño de diferentes procedimientos analíticos para la detección de asociaciones fenotipo-genotipo, incorporando información sobre la estructura genética poblacional y/o el parentesco entre individuos, bajo condiciones de heterocedasticidad entre ambientes.

2.4. Materiales y Métodos

En este capítulo, se simuló un conjunto de datos y se ilustraron sus resultados en dos bases públicas, una de duraznero, cultivo que dio origen al proyecto de este trabajo de investigación, y otra de maíz debido a que es un cultivo de importancia agrícola relevante para Argentina y ampliamente estudiado por otros autores. La simulación del conjunto de datos genómicos permite a los investigadores crear diferentes escenarios posibles que podrían aparecer en la vida real, controlar las características de la población y estudiar el impacto de los diferentes modelos estadísticos utilizados. Este enfoque permite comprender mejor el desempeño de los modelos GWAS bajo los escenarios que se simularon. Para poder emular, a través de la simulación, una base de datos genómica con sus correspondientes valores fenotípicos, se consideraron los valores genéticos de mutación, heredabilidad y cantidad de cromosomas de una base de datos de duraznero [*Prunus persica* (L.) Batsch]. Los hallazgos encontrados sobre los escenarios simulados fueron ilustrados en las dos bases de datos públicas que representan posibles escenarios reales. A continuación se presenta una descripción de cada conjunto de datos utilizados en el presente capítulo de tesis.

2.4.1. Descripción de la simulación de datos genómicas y fenotípicos emulando un genoma vegetal

Simulación de una base de datos genómica

La simulación de los datos, tanto fenotípicos como genotípicos, emularon el genoma de *Prunus persica* (L.) Batsch y se realizó en el paquete *xbreed* de R (Esfandyari y Sørensen, 2019) en dos etapas. En la primera etapa, se simuló la población histórica para crear un nivel deseable de LD. El paquete *xbreed* requiere especificar los siguientes parámetros: número de cromosomas, cantidad de marcadores moleculares y cantidad de QTL, ubicación de los marcadores moleculares y de los QTL, tasa de mutación y frecuencia alélica inicial. Los QTL simulados por *xbreeed* representan los verdaderos loci de caracter cuantitativos asociados a la característica fenotípica. En un contexto de datos reales se pretende identificar dichos QTL a través de los marcadores moleculares.

La ubicación tanto de los marcadores moleculares como de los QTL se distribuyen en cromosomas y se miden en cM, unidad de medida que representa la distancia física relativa entre genes / marcadores en un cromosoma. La flexibilidad de este paquete usado para la simulación, permitió colocar valores que simulan el genoma de referencia correspondiente al duraznero para luego poder ilustrar sobre los escenarios simulados el comportamiento de los datos reales. En el segundo paso se generó la estructura poblacional denominada población reciente que deriva de la población histórica a través de cruzamientos. Se creó una población histórica a partir de 500 individuos independientes entre sí. Luego, el programa realizó, a través de cruzamientos aleatorios, 300 generaciones de individuos. Los parámetros utilizados para el proceso de generación de las poblaciones fueron: heredabilidad en sentido estrecho (h^2): 0,4; relación entre la varianza de dominancia y la varianza fenotípica: 0,1; tasa de mutación: 0,0005; QTL segregante en la última generación histórica: 0,05; marcadores segregantes en la última generación histórica: 0,05; frecuencias alélicas de loci (marcador y QTL) en la primera generación histórica: 0,5. En la simulación de la generación histórica, se consideraron ocho cromosomas con 10.000 marcadores cada uno y una longitud total de 260cM con 15 QTL. Resultando en una base de datos de 240 individuos por 80 mil marcadores moleculares. Los parámetros genéticos seleccionados para la simulación del genoma emulan una base de datos real de variedades de duraznero en Argentina genotipadas con SNP (Romeu et al., 2014).

Desequilibrio de Ligamiento

El LD se refiere a la no independencia entre alelos ubicados en distintos loci, de manera que determinados marcadores, genes o QTL se transmiten de forma conjunta con una probabilidad mayor a la esperada bajo segregación aleatoria (Andrade et al., 2019; Vos et al., 2017; Goddard y Meuwissen, 2005; Wall y Pritchard, 2003). Este fenómeno resulta clave para caracterizar la estructura genética y la historia evolutiva de las poblaciones, así como para la detección de regiones genómicas asociadas a caracteres agronómicos de interés. Al final del proceso de simulación, es necesario confirmar que las características de los datos simulados son similares a las esperadas. Para ello, Esfandyari y Sørensen (2019) proponen estimar el nivel de LD como una manera de informar la tasa de recombinación dentro de una población.

Para validar los datos simulados y comprobar la existencia de LD para los subsiguientes estudios de asociación, a partir de la última población reciente y luego del proceso de depuración de la base de datos donde se eliminaron los loci con frecuencia alélica inferior a $0,05$, se estimaron los valores de r^2 en función de las distancias genéticas (cM) entre pares de marcadores en cada cromosoma. Se calculó el r^2 en cada uno de los ocho cromosomas simulados. Se representó una muestra de todas las combinaciones posibles de pares de marcadores en el mismo cromosoma para visualizar el decaimiento del LD. Las curvas se ajustaron para cada cromosoma del genoma simulado mediante una regresión no paramétrica localmente ponderada (LOESS, del inglés *Smoothed Scatterplot Smoothing*) de segundo grado. Basados en el percentil 95 de los valores de r^2, se estableció un umbral que representa el valor a partir del cual, los marcadores no se encuentran asociados (Figura 2.1).

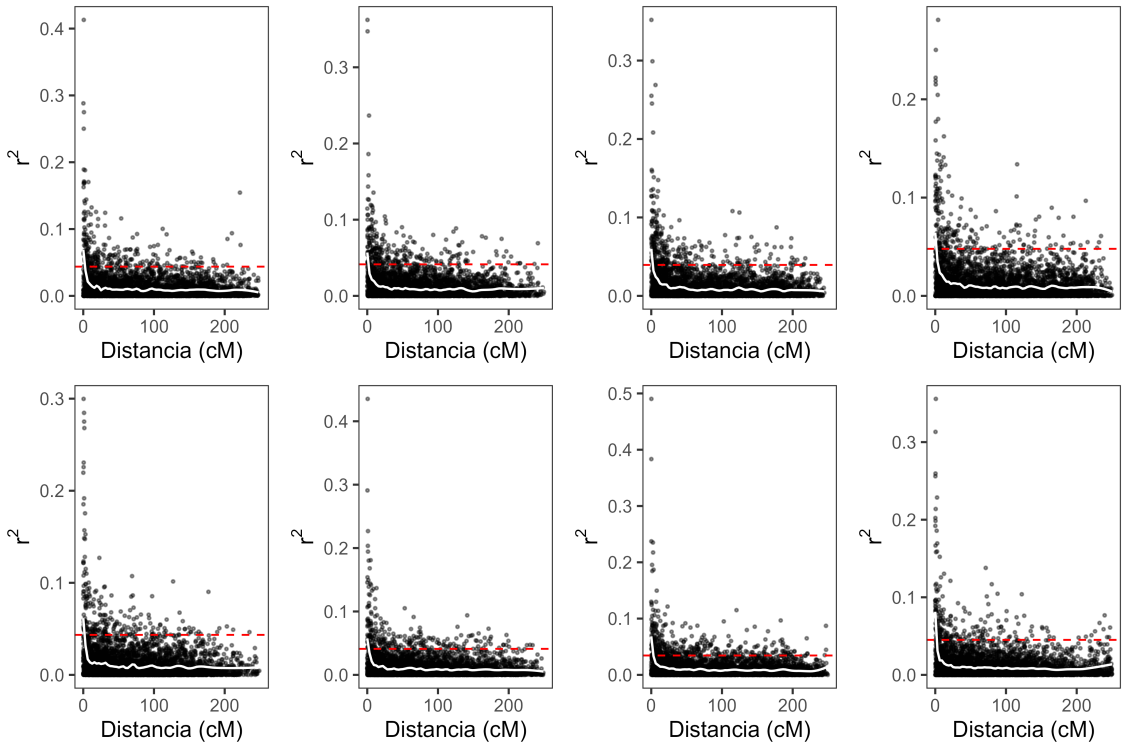

Figura 2.1: Gráficos de los valores de desequilibrio de ligamiento (LD) (r^2) frente a la distancia genética (cM) entre pares de marcadores en cada cromosoma (1-4 arriba; 5-8 abajo). Curvas ajustadas para cada cromosoma mediante LOESS de segundo grado (blanco). Las líneas horizontales indican los umbrales (valores de referencia de r^2) basados en el percentil 95 de la distribución de los valores de r^2, que indican los valores mínimos a partir de los cuales no hay asociación (rojo).

Simulación de variables fenotípicas asociadas el genotipo simulado y obtención de la variabilidad entre ambientes

El paquete *xbreed* en R es una herramienta de simulación tanto de datos genómicos como fenotípicos. En este trabajo de tesis, se obtuvo un caracter fenotípico con distribución normal, sin repeticiones, en un solo ambiente (A_1) asociado al genoma simulado. De esta manera, para cada individuo se generó su perfil molecular y su carácter fenotípico asociado. Con el propósito de evaluar la eficiencia de los modelos para detectar las verdaderas asociaciones entre el fenotipo y el genotipo en contexto de heterocedasticidad entre ambientes, a partir de los datos fenotípicos simulados para el A_1, se creó una nueva variable fenotípica para cada individuo correspondiente a un nuevo ambiente, denominado A_2, de la siguiente manera:

1. Partimos de una variable aleatoria continua $y_i^{A1} \sim \mathcal{N}(\mu_{A_1}, \sigma^2_{A_1})$, donde i es la

observación para el i-ésimo genotipo, simulados por xbreed, con $\sigma^2_{A_1} = 1$, de manera que

$$y_i^{A1} = \mu_{A_1} + \delta_i + \varepsilon_i \tag{2.1}$$

donde δ_i agrupa los efectos de los parámetros genéticos como: tasa de mutación (u), heredabilidad (h), número de alelos (k) y tamaño efectivo de la población (N_e) para lograr una heterocigosidad (h_e) deseada según Kimura y Crow, 1964; y $\varepsilon_i \sim \text{NIID}(0, \sigma^2_\varepsilon)$.

2. Generamos los valores fenotípicos de y^{A2} a partir de y^{A1}, de manera que

$$y_i^{A2} = y_i^{A1} + \varphi_{ip} \tag{2.2}$$

donde φ_{ip} representa un factor de corrimiento de la media y la varianza, para el i-ésimo genotipo según el percentil p ($p = 1, \ldots, 6$) de la variable y^{A1}, tal como se muestra a continuación:

$$y_{ip}^{A2} = \begin{cases} y_{ip}^{A1} + \varphi_{ip}, & \text{donde } \varphi_{ip} \sim \mathcal{N}(2.5, 0.5) \quad \text{si } y_i^{A1} \leq y_{ip_{10}}^{A1}, \\ y_{ip}^{A1} + \varphi_{ip}, & \text{donde } \varphi_{ip} \sim \mathcal{N}(0, 1.5) \quad \text{si } y_{ip_{10}}^{A1} < y_i^{A1} \leq y_{ip_{25}}^{A1}, \\ y_{ip}^{A1} + \varphi_{ip}, & \text{donde } \varphi_{ip} \sim \mathcal{N}(0, 1) \quad \text{si } y_{ip_{25}}^{A1} < y_i^{A1} \leq y_{ip_{50}}^{A1}, \\ y_{ip}^{A1} + \delta_{ip}, & \text{donde } \varphi_{ip} \sim \mathcal{N}(0.5, 1) \quad \text{si } y_{ip_{50}}^{A1} < y_i^{A1} \leq y_{ip_{75}}^{A1}, \\ y_{ip}^{A1} + \delta_{ip}, & \text{donde } \varphi_{ip} \sim \mathcal{N}(0.5, 1.5) \quad \text{si } y_{ip_{75}}^{A1} < y_i^{A1} \leq y_{ip_{90}}^{A1}, \\ y_{ip}^{A1} + \varphi_{ip}, & \text{donde } \varphi_{ip} \sim \mathcal{N}(-2.5, 0.5) \quad \text{si } y_i^{A1} > y_{ip_{90}}^{A1}. \end{cases} \tag{2.3}$$

donde $y_{ip_a}^{A1}$ corresponde al percentil a de la distribución de y_i^{A1}, con $a = 10, 25, 50, 75, 90$.

3. Así, $y_i^{A2} = \sum_{p=1}^{6} y_{ip}^{A2}$, donde $y_{ip}^{A2} \sim \mathcal{N}(\mu_{A2}, \sigma^2_{A2})$.

4. Para estimar el efecto de la IGA se simularon tres repeticiones de la variable fenotípica para cada genotipo en cada ambiente, agregando un número aleatorio

con distribución normal cuya media fue cero y cuya varianza se modificó para generar en el ambiente 2 heterocedasticidad respecto al ambiente 1:

$$y_{ijkl} = y_{ik} + \varphi_{lk} + \varepsilon_{ijkl}, \tag{2.4}$$

donde $\varphi_{lk} \sim \text{NIID}(0, \sigma^2_{lk})$ y $\varepsilon_{ijkl} \sim \text{NIID}(0, \sigma^2_{\varepsilon})$, con $i = 1, \ldots, n, \quad j = 1,2,3, \quad k = 1,2, \quad l = 1,2,3$

Aquí, y_{ijkl} representa el carácter fenotípico y en el i-ésimo genotipo, j-ésima repetición, k-ésimo ambiente y l-ésimo escenario de heterocedasticidad entre ambientes; y ε_{ijkl} es el error aleatorio correspondiente. Las distribuciones de λ_k se muestran en la Tabla 2.1.

Tabla 2.1: Distribución del término φ según el k-ésimo ambiente y el l-ésimo escenario de heterocedasticidad entre ambientes. Notar que la distribución de φ es la misma entre repeticiones de un mismo genotipo por ambiente y escenario

	Ambiente 1			Ambiente 2		
	E1	E2	E3	E1	E2	E3
Repetición 1	$N(0,0.5)$	$N(0,0.5)$	$N(0,0.5)$	$N(0,0.5)$	$N(0,2.5)$	$N(0,12.5)$
Repetición 2	$N(0,0.5)$	$N(0,0.5)$	$N(0,0.5)$	$N(0,0.5)$	$N(0,2.5)$	$N(0,12.5)$
Repetición 3	$N(0,0.5)$	$N(0,0.5)$	$N(0,0.5)$	$N(0,0.5)$	$N(0,2.5)$	$N(0,12.5)$

De esta manera, se crearon seis nuevas variables fenotípicas asociadas a los perfiles moleculares generados en la simulación original por *xbreed*.

Se realizó la prueba de Bartlett para comprobar la heterocedasticidad simulada entre ambientes en cada escenario. La prueba para el escenario 1 que bajo hipótesis nula postula homogeneidad de varianzas entre ambientes, arrojó una probabilidad asociada de $0,86$, no hubo evidencias suficientes para no rechazar la hipótesis nula. Mientras que para el segundo y tercer escenario la probabilidad asociada obtenida de la prueba de Bartlet fue menor a $0,0001$ en los dos últimos casos. A partir de aquí, confirmamos que la simulación contiene los datos genotípicos asociados con el carácter fenotípico y con escenarios de homocedasticidad y heterocedasticidad para evaluar el comportamiento de los modelos de asociación (Figura 2.2).

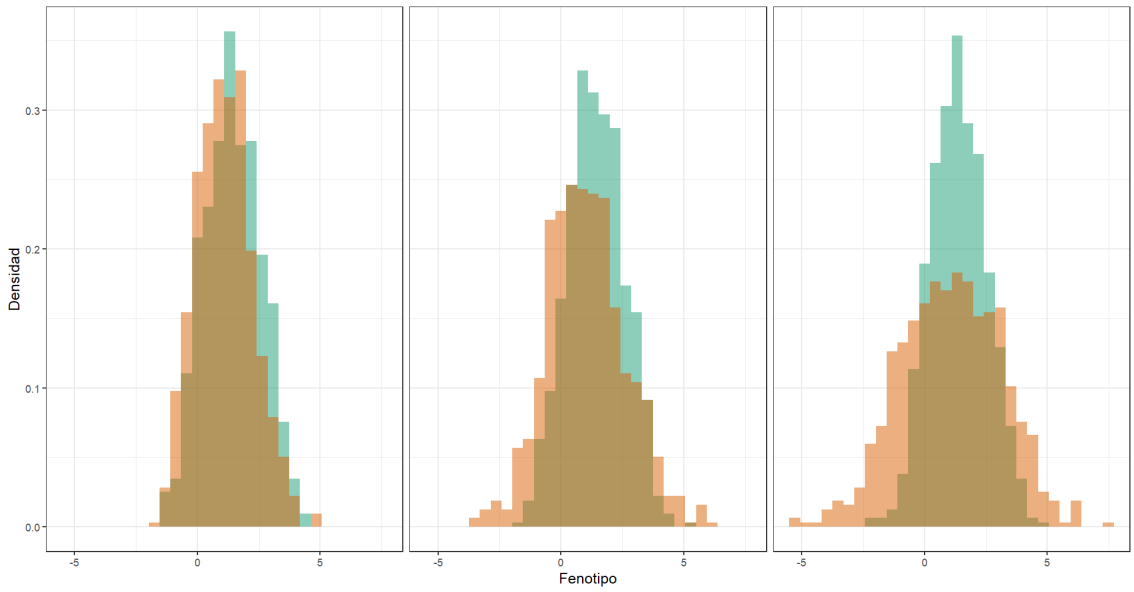

Figura 2.2: Función de distribución del rasgo fenotípico para dos ambientes (y^{A1} verde y y^{A2} naranja) bajo tres escenarios: (izquierda) sin heterocedasticidad, (centro) heterocedasticidad moderada (0,25 y 2,25) y (derecha) alta heterocedasticidad entre ambientes (0,25 y 12,25)

Para controlar la presencia del efecto de la IGA en cada escenario se utilizó la prueba de cociente de verosimilitud entre dos modelos, el primero es el modelo completo que incluye el efecto G, A e IGA y el segundo modelo que sólo incluye G y A (una diferencia en grados de libertad entre los dos modelos), de esta forma es posible probar formalmente si los parámetros adicionales (en este caso, término IGA) son cero bajo hipótesis nula. Esta prueba se definió como:

$$\text{LRT} = -2\ln(\lambda), \tag{2.5}$$

donde $\lambda = \frac{L_0}{L_1}$, L_0 representa la verosimilitud del modelo reducido (sin el término IGA) y L_1 la verosimilitud del modelo completo (con el término IGA). Si el valor p resulta inferior al nivel de significación ($\alpha = 0,05$), se rechaza la hipótesis nula, lo cual indica que existen diferencias entre el modelo completo y el modelo reducido. Dicho de otra manera, el modelo que tiene un término adicional (modelo completo) es estadísticamente diferente del modelo reducido y, por lo tanto, debe preferirse el modelo completo. En este caso, el modelo completo es el que incluye el término correspondiente al efecto de la IGA (Little, 2006).

2.4.2. Descripción de las dos bases de datos públicas utilizadas para ilustración

Base de datos de duraznero

Se trabajó con una base de datos perteneciente a un germoplasma de cultivo de durazno de mercado fresco de América del Norte (da Silva Linge et al., 2021) con 72 cultivares derivados de selecciones avanzadas y 548 genotipos pertenecientes a tres programas de mejoramiento de duraznos del mercado fresco pertenecientes a la División de Agricultura del Sistema de la Universidad de Arkansas (AR), la Universidad de Clemson (SC) y la Universidad de Texas AyM. (TX). Para cada genotipo se registraron dos variables fenotípicas: (1) los días a floración (BD; en días julianos) que corresponden a los días transcurridos desde que el árbol comienza a producir flores hasta visualizar el 60–80 % de las flores abiertas en cada árbol y (2) los días posteriores a la floración (DAB) hasta la cosecha. Los datos fenotípicos se registraron durante dos campañas agrícolas (2011–2012). Para este trabajo de tesis, consideramos cada temporada de crecimiento como un ambiente. Luego, se realizó la prueba de Bartlett para comprobar homogeneidad entre ambientes (temporadas de crecimiento). Se encontraron diferencias estadísticamente significativas entre los ambientes, comprobando de esta forma la presencia de heterocedasticidad entre ellos (valor $p \leq 0,0001$ para BD y valor $p = 0,01$ para DAB).

Base de datos de Maíz

Se utilizó un conjunto de datos de maíz proveniente del proyecto *Drought Tolerance Maize for Africa* del Programa Global de Maíz del CIMMYT (Crossa et al., 2010). El conjunto de datos incluía 300 líneas de maíz tropical genotipadas con 1.148 marcadores moleculares de polimorfismos de un solo nucleótido (SNP, por sus siglas en inglés single nucleotide polimorphism). Para cada SNP, el alelo con la frecuencia más baja se codificó como uno. No se disponía de información sobre el parentesco genético entre estas líneas (pedigrí). La frecuencia media de alelos menores en estos conjuntos de datos fue de 0,20. Se eliminaron los marcadores con frecuencia alélica $< 0,05$ o $> 0,95$. Los valores de marcadores SNP faltantes se imputaron utilizando muestras de

la distribución marginal de marcadores, es decir, $x_{ij} \sim Bernoulli(p^j)$, donde p^j es la frecuencia alélica estimada a partir de las líneas de maíz (genotipos) no faltantes. Las variables fenotípicas medidas para estas líneas fueron el rendimiento de grano (GY), los días a floración femenina (FFL), los días a floración masculina (MFL) y el intervalo de tiempo transcurrido entre antesis y floración (ASI). En lo sucesivo, denominaremos a la variable de rendimiento de grano de maíz como M-GY y a las variables de floración de maíz como M-F para referirnos a FFL, MFL y ASI. El número de líneas en el conjunto de datos M-GY fue 264, mientras que un total de 284 líneas estaban disponibles en M-F. Los números de marcadores disponibles para el análisis fueron 1.135 y 1.148 en M-GY y M-F, respectivamente. Todas las líneas fueron evaluadas en dos condiciones (ambientes), uno bajo estrés por sequía severa (SS) y otro ambiente donde se aplicaba riego y por lo tanto no presentó estrés por sequía (WW).

Las repeticiones para las variables fenotípicas en cada ambiente no se encontraban publicadas y por ello simulamos tres repeticiones fenotípicas para cada genotipo en cada ambiente (SS y WW), agregando un término de error aleatorio normalmente distribuido con media cero y varianza entre ambientes iguales y con un valor de $\sigma^2 = 0,5$. De esta manera, generamos un escenario donde la varianza ente ambientes era homogénea. Consideramos a la situación de homocedasticidad entre ambientes como escenario 1 y generamos dos situaciones de heterocedasticidad entre ambientes, denominadas escenario dos con los siguientes niveles de varianza para cada ambiente: $\sigma^2_{SS} = 0,5$ y $\sigma^2_{ww} = 3,5$ y escenario tres con $\sigma^2_{SS} = 0,5$ y $\sigma^2_{ww} = 9,5$ para cada ambiente. Se utilizó la prueba de Bartlett para corroborar las varianzas simulada en cada ambiente. Para el primer escenario de varianzas homogéneas entre ambientes no se encontraron diferencias estadísticamente significativas entre ellos (valor p=0,79) como era de esperar dado que ambos ambientes fueron simulados con la misma varianza. En el segundos y tercer escenario se rechazó la hipótesis nula de homogeneidad de varianzas, confirmando la heterocedasticidad simulada entre ambientes (valor $p \leq 0,0001$).

2.4.3. Análisis estadístico

Estructura de varianza-covarianza para la estimación G-BLUP

Identificar la función de estructura de varianza-covarianza que mejor capture la estructura de los datos, es posiblemente, la característica más importante de los modelos mixtos. En este capítulo, para cada escenario de heterocedasticidad dentro de cada conjunto de datos, tanto simulado como reales, y para cada carácter fenotípico, se ajustó un MLM usando como variable respuesta cada uno de los caracteres fenotípicos. Se consideró el efecto del ambiente como fijo, mientras que los efectos de G y de la IGA como aleatorios. Se utilizó la forma general de los MLM con distintas estructuras de varianza–covarianza.

$$y_{N\times 1} = X\beta + Zu + \varepsilon \tag{2.6}$$

donde $y_{N\times 1}$ es el vector de la variable de respuesta en el ambiente i (en nuestro caso, $i = 1,2$) y el genotipo j; N es el total de observaciones, es decir, genotipos x ambiente x repeticiones. Se trabajó tres conjuntos de datos que presentaron dimensiones diferentes. Para el conjunto de datos simulados, $N = 1.440$, en el conjunto de datos de maíz $N = 1.584$ en las variables relacionados a producción de granos (M-GY) y $N = 1.704$ para el conjunto de variables relacionadas a floración (M-F). Para el conjunto de datos de durazno $N = 1.037$; X es la matriz de incidencia, también llamada matriz de diseño, para el efecto fijo, β es el vector de parámetros que incluye los efectos del intercepto y del ambiente; Z es la matriz de incidencia de los efectos aleatorios; u es el vector desconocido asociado a los efectos aleatorios que sigue una distribución normal con media 0 y matriz de varianza-covarianza G ($\mu \sim N(0,G)$) cuyos elementos de la diagonal representan la varianza de cada efecto aleatorio en u; i.e,. la varianza genotípica o a varianza del genotipo y los elementos fuera de la diagonal representan las covarianzas entre dos efectos aleatorios, i.e, la covarianza entre los genotipos. Dado que hay q efectos aleatorios en el modelo asociado con el sujeto i -ésimo, G es una matriz cuadrada $q \times q$ simétrica y definida-positiva (West et al., 2014). G puede incorporar la covarianza entre genotipos a través de una matriz de parentesco o matriz K, la cual refleja el grado de similitud genética entre individuos. Esto permite modelar explícitamente

la dependencia entre genotipos derivada de la estructura poblacional, mejorando la estimación de los componentes de varianza y reduciendo la tasa de asociaciones espurias. El paquete *sommer* implementa esta estrategia mediante el uso de matrices de relación genética positivas definidas, facilitando la inclusión de información genómica en la estructura de varianza–covarianza del modelo (Pérez y De Los Campos, 2014). Finalmente, ε es el vector de n_i errores aleatorios con distribución normal, media 0 y matriz de varianza-covarianza simétrica y positiva R ($\varepsilon \sim N(0, R)$). También se asume que los errores son independientes. Luego, la matriz de varianzas-covarianzas de Y queda definida como:

$$V(Y) = ZGZ' + R \tag{2.7}$$

Si el genotipo o el ambiente fue considerado como aleatorio, sus efectos pueden ser estimados a través de los BLUP. El BLUP en un MLM es aquel que logra estimar de manera precisa y no sesgada los valores de la variable dependiente en función de las variables independientes y los efectos aleatorios del modelo. Un MLM combina componentes fijos y aleatorios para capturar tanto la variación sistemática como la variación aleatoria en los datos. En un MLM, los efectos fijos representan las relaciones sistemáticas entre las variables independientes y la variable dependiente, mientras que los efectos aleatorios capturan la variación no explicada de forma sistemática, que puede estar relacionada con características específicas de las unidades de observación o de las muestras. Estos efectos aleatorios se modelan como variables aleatorias y se consideran componentes adicionales en el modelo. El BLUP busca minimizar el sesgo sistemático en las predicciones. Esto significa que, en promedio, el valor predicho por el modelo es igual al valor real de la variable dependiente. Para lograr esto, el predictor lineal insesgado toma en cuenta tanto los efectos fijos como los efectos aleatorios en el modelo y utiliza estimaciones adecuadas de los parámetros del modelo. Es importante destacar que la elección del mejor predictor lineal insesgado puede depender de varios factores, como la estructura de los datos, la distribución de los efectos aleatorios y las suposiciones del modelo. En este trabajo, se denominó a los efectos de cada genotipo como G-BLUP. Se asume que la distribución de los datos observados es $y \sim MVN(X\beta, V)$, donde V representa los efectos aleatorios y el error residual dada por

$V = ZG\hat{Z}' + R$. Los predictores BLUP son estimadores insesgados, que se basan en las varianzas del genotipo para determinar el valor de respuesta, eliminando la confusión por la heterocedasticidad. El uso de BLUP requiere el supuesto de normalidad (West et al., 2014). Los G-BLUP se estiman para cada genotipo (efectos aleatorios) mediante $\hat{u} = \hat{G}Z'\hat{V}^{-1}(y - X\hat{\beta})$. El BLUP tiene el menor error cuadrático medio entre todos los predictores lineales insesgados (Balzarini, 2002). Las técnicas basadas en la verosimilitud implicadas en la estimación de MLM proporcionan un enfoque analítico más flexible para G y R. Por lo tanto, consideramos cuatro modelos diferentes para la estructura de varianza-covarianza de los efectos aleatorios (G) para modelar $Var(u)$:

1. Modelo Diagonal, en el que cada efecto aleatorio en u tiene sus propias varianzas, y todas las covarianzas en G se definen como 0. En nuestro estudio, se consideró como efecto aleatorio los términos G y IGA, por lo que este modelo supone que la variación del genotipo se expresa de forma diferente en cada ambiente, por lo que el modelo diagonal ajusta una componente de varianza para el efecto del genotipo en cada ambiente;

2. Simetría compuesta, el modelo de simetría compuesta se corresponde a una correlación uniforme, es decir, esta estructura se utiliza a menudo cuando se supone una correlación igual. En este estudio, la estructura de covarianza de simetría compuesta supone que existe IGA y que un CP (constante) de varianza-covarianza del genotipo se expresa en todos los ambientes;

3. Autorregresivo de primer orden (AR(1)), se supone que la varianza es constante y que la covarianza que está separada por w unidades es igual a p^w. Esta estructura implica que las observaciones más cercanas en el tiempo muestran una correlación más alta que las observaciones más alejadas en el tiempo (West et al., 2014);

4. No estructurada, es el modelo más flexible, supone que existe IGA y que existe una varianza específica en cada ambiente además de tantas covarianzas para cada combinación de ambiente a ambiente, en este caso la matriz G tiene $q \times q + 1$ número de parámetros.

Índices de bondad de ajuste para la comparación de la estructura de varianza-covarianza

Se utilizaron los Criterios de Información Akaike y Bayesianos (AIC y BIC, respectivamente) (Wolfinger, 1993) para comparar modelos alternativos y estimar las correlaciones genéticas dentro del modelo. El criterio de selección indica que el valor más pequeño de AIC y BIC identifica el modelo de mejor ajuste. Dado que estos criterios están basados en el logaritmos de la verosimilitud y la cantidad de parámetros, ambos penalizan más cuanto máyor sea la cantidad de parámetros a estimar. Ambos criterios de información (AIC y BIC) representan una forma de comparar cualquier dos modelos ajustados que tienen el mismo conjunto de observaciones, es decir, los modelos no necesitan estar anidados en sus parámetros.

La fórmula para estimar AIC fue propuesta por Akaike (1974) como se describe a continuación:

$$AIC = -2\ell_{ML} + 2k \tag{2.8}$$

donde ℓ_{ML} es el logaritmo de la función de máxima verosimilitud (*Maximum Likelihood*) del modelo y k es el número total de parámetros independientes estimados. La estimación por máxima verosimilitud corresponde al conjunto de valores de los parámetros que maximiza la probabilidad de observar los datos. Dado que el logaritmo es una función monótona creciente, maximizar la verosimilitud es equivalente a maximizar su logaritmo, lo que simplifica los cálculos al transformar productos en sumas (Little, 2006). El criterio AIC incorpora el principio de parsimonia, favoreciendo modelos que logran un adecuado ajuste con un número reducido de parámetros. Para modelos lineales, AIC es asintóticamente insesgado y presenta un buen comportamiento para tamaños de muestra moderados.

En el contexto de modelos lineales mixtos (MLM), Vaida y Blanchard (2005) extendieron el uso de AIC a partir de la verosimilitud restringida, dando lugar a un criterio basado en REML. La verosimilitud restringida se obtiene eliminando el efecto de los parámetros fijos mediante una transformación del modelo, y su logaritmo puede expresarse como:

$$\ell_{\text{REML}} = \ell_{\text{ML}} - \frac{1}{2}\log\left|X'V^{-1}X\right| - \frac{p}{2}\log(2\pi) \tag{2.9}$$

donde X es la matriz de diseño de los efectos fijos, V es la matriz de varianza-covarianza del vector de observaciones y p es el número de efectos fijos del modelo. A partir de esta cantidad se define el criterio:

$$\text{AIC}_{\text{REML}} = -2\,\ell_{\text{REML}} + 2k_{\text{REML}} \tag{2.10}$$

donde k_{REML} corresponde al número de parámetros asociados a las estructuras de varianza-covarianza. Cabe destacar que el AIC calculado bajo REML no es directamente comparable con el AIC basado en ML cuando los modelos difieren en la especificación de los efectos fijos (Vaida y Blanchard, 2005).

El criterio de información de Bayes (BIC), propuesto por Schwarz (1978), es también frecuentemente utilizado en la selección de modelos y aplica una penalización más severa a la complejidad del modelo que AIC. Mientras que AIC penaliza con $2k$, BIC penaliza con $k\log(n)$, donde n es el número total de observaciones utilizadas en el ajuste del modelo (Coelho y Rodrigues, 2012). El criterio BIC basado en máxima verosimilitud se define como:

$$\text{BIC}_{\text{ML}} = -2\,\ell_{\text{ML}} + k\log(n), \tag{2.11}$$

donde k es el número de parámetros estimados y n es el número total de observaciones utilizadas en el ajuste del modelo. El paquete `sommer` permite calcular una versión del criterio basada en REML, definida como:

$$\text{BIC}_{\text{REML}} = -2\,\ell_{\text{REML}} + k_{\text{REML}}\log(n), \tag{2.12}$$

donde k_{REML} corresponde al número de parámetros asociados a las estructuras de varianza-covarianza del modelo. En consecuencia, los valores de BIC obtenidos mediante ML y REML no son directamente comparables cuando los modelos difieren en la especificación de los efectos fijos.

Descripción de los modelos GWAS

Los modelos comparados en este trabajo de tesis fueron ocho. El modelo denominado GWAS es el único que se realiza con la variable respuesta (y), en el mismo se ajustó un MLM univariado GWAS por el método REML con el paquete *sommer* (Covarrubias-Pazaran, 2018). Los otros siete modelos fueron ajustados utilizando los G-BLUP estimados como variable respuesta según lo descripto anteriormente con el paquete *GAPIT* (Lipka et al., 2012). Se consideró que al usar el valor genotípico obtenido a través de una matriz de varianza-covarianza adecuada para el conjunto de datos, la estimación de la asociación entre el valor genotípico y la información molecular tendrá menos falsos positivos. Entre los siete modelos ajustados con G-BLUP se incluyeron modelos que contemplan la estructura genética poblacional como Modelo-P, MLM con PCA y K - MLM-PK y modelos multi-locus que surgen como una forma de controlar los falsos positivos, como CMLM, ECMLM, SUPER y FarmCPU.

1. Modelo GWAS con fenotipo como variable de entrada

   El modelo denominado GWAS considera como variable respuesta a los caracteres fenotípicos evaluados. Su función considera efectos fijos y efectos aleatorios además de considerar todos los marcadores moleculares.

$$y_i = X\beta + Zg + Mt + \varepsilon_i \tag{2.13}$$

   donde y_i es el vector que contiene la información de la variable de respuesta (característica fenotípica de los individuos evaluados) para la i-ésima observación, β es un vector de efectos fijos que puede modelar tanto los factores ambientales como la estructura de la población. El vector de coeficientes g modela el trasfondo genético de cada genotípico como un efecto aleatorio con $Var(g) = K\sigma^2$. Si no se conoce la matriz K; puede ser reemplazada por una matriz de Aditividad (A) estimada a partir de la similitud molecular (D. Gianola y De Los Campos, 2008). El vector t modela el efecto aditivo del marcador del tipo SNP como un efecto fijo. M es la matriz con marcadores moleculares SNP. El término ε_i representa el error experimental cuya distribución se asume normal, independiente y con varianzas

homogéneas $Var(\varepsilon) = \sigma^2$. Para ajustar este modelo se utilizó la función *GWAS* del paquete *sommer* (Covarrubias-Pazaran, 2018) en R (R core, 2023). Se analizó marcador por marcador mediante un MLM (un modelo por marcador) para obtener el efecto del marcador y un estadístico que refleja el nivel de asociación para cada rasgo como variable de respuesta (univariante).

2. Modelos de asociación fenotipo-genotipo considerando los efectos de los genotipos (G-BLUP) como variable respuesta

   Los modelos incluyen GLM considerando la estructura genética poblacional a través de las componentes principales obtenidas a través del ACP sobre los marcadores moleculares, este modelo es denominado Modelo-P, MLM considerando las Componentes Principales del ACP y la matriz K, a este modelo se lo denominó-MLM-PK. También se consideraron modelos multilocus:, MLM comprimido - CMLM, CMLM enriquecido -ECMLM, MLM de locus múltiple - MMLM, Liquidación de MLM Bajo Relación Progresivamente Exclusiva - SUPER, Modelo fijo y aleatorio Unificación de Probabilidad Circulante - FarmCPU.

   a) Modelo lineal general considerando estructura genética poblacional a partir del ACP (Modelo-P)

   Este modelo fue propuesto por Price et al., 2006 para detectar y corregir la estructura genética de la población (EGP) y evitar asociaciones espurias causadas por la estratificación de la población. La ecuación utilizada en la cartografía de asociación, también denominada modelo P, es un modelo de regresión lineal:

$$y_i = X\beta + Sv + \varepsilon_i \tag{2.14}$$

   donde y_i es el vector G-BLUP estimado para la i-ésima observación; β es un vector que contiene los coeficientes de regresión que permiten asociar la variable fenotípica con cada marcador molecular, X es la matriz de incidencia de efectos fijos; S es la matriz de estructura genética obtenida con componentes principales a partir de la matriz de marcadores moleculares e,

ε_i es el vector no observado de residuos aleatorios. Por otra parte, podríamos presentar el modelo de esta forma:

$$y_i = \mu + M + P + \varepsilon_i \tag{2.15}$$

donde y_i es el vector de G-BLUP estimado para la i-ésima observación, M es la matriz de marcadores. P es la matriz de componentes principales incluida en el modelo como cofactor. Se espera que los CP obtenidos del ACP describan la estructura genética caracterizada por los marcadores moleculares. Donde el número de columnas de P se selecciona mediante el método de Tracy-Widom (Tracy y Widom, 1994). La ventaja es que los CP tienen la característica de no estar correlacionados y permiten señalar diferencias entre genotipos provocadas por la existencia de estructuración genética entre los genotipos de la población mapeada. (Pritchard et al., 2000; Peña-Malavera et al., 2014). ε_i es el vector no observado de residuos.

b) Modelo Lineal Mixto con P y K (MLM-PK)

La información sobre las relaciones entre individuos, genotipos o líneas puede incorporarse al MLM mediante matriz de parentesco (K), ya sea a partir de pedigríes o marcadores moleculares, basándose en la matriz aditiva de relaciones (Ad) entre genotipos:

$$y_i = X\beta + Zu + \varepsilon_i \tag{2.16}$$

donde y_i es el vector de G-BLUP estimado para la i-ésima observación; β es el vector de parámetros de los efectos fijos, incluido los SNP, las CP y el intercepto; u es un vector desconocido de efectos genéticos aditivos aleatorios de múltiples QTL para individuos; X y Z son las matrices de diseño conocidas; y ε_i es el residuo de la i-ésima observación. Se supone que los vectores u y ε se distribuyen normalmente con una media 0 y una varianza de:

$$Var\begin{pmatrix} u \\ \varepsilon \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \tag{2.17}$$

donde $G = \sigma_a^2 K$ con σ_a^2 como la varianza genética aditiva y K como la matriz de parentesco. Se asume una varianza homogénea para el efecto residual, es decir, $R = \sigma_\varepsilon^2 I$, donde σ_ε^2 es la varianza residual. Usualmente, cuando se ajusta este tipo de modelo donde existe información respecto al pedigrí a través de la matriz K, la proporción de la varianza total explicada por la varianza genética se define como heredabilidad (h^2):

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\varepsilon^2} \tag{2.18}$$

c) MLM comprimido (CMLM)

Los estudios de asociación entre genotipos y fenotipos en muestras de gran tamaño se vuelven computacionalmente intensivos porque cada iteración requiere una cantidad de tiempo que es proporcional al cubo del número de individuos en el efecto aleatorio (Z. Zhang et al., 2010). En este escenario, se propuso el modelo lineal mixto comprimido (CMLM, por sus siglas en inglés *compressed linear mixed model*) para reducir el tamaño del efecto aleatorio. Para lograr este objetivo, en el CMLM, los efectos genéticos individuales se sustituyen por un grupo de efectos genéticos. En consecuencia, este menor número de efectos aleatorios en un MLM reduce el tiempo de cálculo. El término comprimido se utiliza para referirse a cómo el efecto aleatorio en un MLM se comprime de individuos a grupos. Así, los individuos se comprimen en grupos o conglomerados basados en el parentesco entre pares de grupos en lugar del parentesco entre pares de individuos.

d) CMLM Enriquecido (ECMLM)

La optimización del parentesco entre grupos mejora aún más el poder estadístico en el CMLM enriquecido (ECMLM, por sus siglas en inglés *enriched* CLMM). Este método encuentra la combinación óptima de algoritmos de nivel de compresión, parentesco y clúster, lo que da como resultado una

mayor potencia estadística y un mejor ajuste del modelo. Los valores de los parámetros optimizados se pueden usar para las pruebas de asociación de SNP, que es el paso que consume más tiempo en GWAS. El ECMLM calcula el parentesco utilizando varios algoritmos diferentes y luego elige la mejor combinación entre algoritmos de parentesco y algoritmos de agrupación (M. Li et al., 2014). Los ocho algoritmos de agrupamiento fueron: UPGMA; centroide de grupo de pares no ponderado (UPGMC); enlace completo (COM); método beta flexible de Lance-Williams (FLE); el análisis de similitud de McQuitty, también llamado método de grupos de pares ponderados usando promedios aritméticos (WPGMA); método de grupos de pares ponderados usando centroide (WPGMC); enlace único (SIN), que también se denomina vecino más cercano; y el método de Ward (WAR).

e) Multi-locus Mixed Model (MMLM)

Este método propuesto por Segura et al., 2012 utiliza un modelo de regresión lineal mixto paso a paso adelante-atrás,también conocido como *stepwise forward* y *backward*, donde los componentes de la varianza $\hat{\sigma}_g^2$ y $\hat{\sigma}_\varepsilon^2$ se estiman antes de cada paso. Las estimaciones de varianza se utilizan para obtener estimaciones del tamaño del efecto de mínimos cuadrados generalizados y valores p de la prueba F para cada SNP. Luego, el SNP más significativo se agrega al modelo como cofactor para el siguiente paso, y los valores p para todos los cofactores se vuelven a estimar junto con los componentes de la varianza. Después de detener la regresión por pasos hacia adelante, se realiza una regresión por pasos hacia atrás descartando el cofactor menos significativo en el modelo en cada paso. El componente de varianza y los valores de p de todos los cofactores se volvieron a estimar en cada paso. Para la estimación de los componentes de la varianza en cada paso hacia adelante y hacia atrás, los marcadores incluidos como cofactores en el modelo pueden excluirse del cálculo de la matriz de parentesco, aunque no lo hicimos porque su efecto sobre el parentesco es posiblemente insignificante.

f) Modelo SUPER

En el método SUPER, la matriz K se deriva de los marcadores asociados y se

ajusta en consecuencia mediante las pruebas de marcadores. Como la K se deriva de un número menor de marcadores que la K en MLM y MMLM que se derivan de todos los marcadores, la confusión entre K y algunos de los marcadores se vuelve más severa. SUPER elimina la confusión mediante el uso del parentesco complementario derivado de los marcadores asociados, excepto aquellos que se encuentran en un fuerte LD con los marcadores de prueba por debajo de un umbral definido por el usuario.

Se utilizó el algoritmo eficiente de VanRaden (2008) (implementado en *GAPIT*) para calcular la matriz de parentesco.

El método SUPER añade el efecto del marcador (v) de uno en uno en el marco de un MLM estándar para realizar un GWAS:

$$y_i = X\beta + Zu + Wv + \varepsilon_i \tag{2.19}$$

donde W es la matriz de incidencia para v.

g) Modelo FarmCPU

Para resolver el problema del control de falsos positivos y la confusión entre probar marcadores y cofactores simultáneamente, en 2016 se desarrolló un método iterativo, denominado FarmCPU (Liu et al., 2016). Los marcadores asociados detectados a partir de las iteraciones se ajustan como cofactores para controlar los falsos positivos y así probar el resto de los marcadores moleculares en un modelo considerando a los mismos como efectos fijos. Para evitar el problema de sobreajuste del modelo en la regresión por pasos, se usa un modelo de efectos aleatorios para seleccionar los marcadores asociados usando el método de máxima verosimilitud. Como FarmCPU prueba los marcadores en un modelo de efectos fijos, es computacionalmente más eficiente que los métodos que prueban los marcadores en un modelo de efectos aleatorios. La estructura de la población está representada por tres covariables, que se obtuvieron utilizando la herramienta integrada de predicción y asociación del genoma (Lipka et al., 2012). FarmCPU es considerado el mejor modelo GWAS multi locus ya que controla tanto los falsos positivos

como los falsos negativos (Kaler et al., 2020).

Evaluación de modelos e interpretación de gráficos Q-Q

Los siete modelos de mapeo por asociación evaluados se compararon mediante gráficos cuantil-cuantil (Q-Q). El gráfico Q-Q permite determinar si los modelos controlan adecuadamente los falsos positivos y los falsos negativos utilizando el logaritmo de los valores p observados en las pruebas de los marcadores (eje Y) frente a los valores p esperados (eje X). Si el gráfico Q-Q muestra que los valores p obtenidos del modelo ajustado (valores p observados) se ubican sobre la línea recta de 45° y no presentan una cola desviada, significa que siguen una distribución uniforme. Esto indica que la hipótesis nula no se rechaza y que no existe una asociación estadísticamente significativa entre el genotipo y el fenotipo. Por el contrario, cualquier desviación respecto de la línea recta cercana a la relación 1:1 indicaría que la hipótesis nula no es verdadera (es decir, se rechaza) y que existen asociaciones significativas. Cuando los valores p del modelo ajustado se sitúan por debajo de la línea recta, el gráfico Q-Q indica la presencia de falsos negativos; esto significa que existen marcadores SNP asociados con el fenotipo, pero sus valores p no alcanzan significación estadística. Si la línea del gráfico Q-Q se ubica por encima de la línea recta, ello indica la presencia de falsos positivos, es decir, asociaciones espurias. Cuando los valores p obtenidos se distribuyen sobre la línea recta, es decir, cerca de la línea 1:1, pero presentan una cola claramente desviada hacia arriba, el gráfico Q-Q indica la existencia de asociaciones verdaderas y polimorfismos causales. En este caso, tanto los falsos positivos como los falsos negativos han sido adecuadamente controlados por el modelo ajustado (Kaler et al., 2020).

Método de comparaciones múltiple

Cuando se prueba un único marcador, el valor p objetivo para rechazar la hipótesis nula y considerar una asociación significativa debe ser igual o inferior a $\alpha = 0,05$. Dicho de otra manera, un nivel de significación del 5 % está asociado a la probabilidad de cometer un Error de tipo I, es decir, la probabilidad de identificar asociaciones espurias

(falsos positivos) el 5 % de las veces (Gangurde et al., 2023). Sin embargo, en la investigación genética moderna es habitual realizar un gran número de pruebas estadísticas correlacionadas. Cuando no se controla adecuadamente este efecto de multiplicidad, aumenta la probabilidad de obtener resultados falsos positivos y, simultáneamente, disminuye la potencia para detectar asociaciones verdaderas. Para hacer frente a este problema se han propuesto diversos métodos de corrección por comparaciones múltiples. Los procedimientos más simples son los denominados métodos tipo Bonferroni, donde el nivel de significación se divide entre el número total de pruebas realizadas (Bonferroni, 1936; Dunn, 1961). Aunque este enfoque controla estrictamente el error familiar de tipo I, suele considerarse demasiado conservador, especialmente cuando las pruebas no son independientes, como ocurre frecuentemente en marcadores genéticos ligados. Para reducir el exceso de conservadurismo, se han desarrollado métodos secuenciales como el de Holm–Bonferroni, que ajusta los valores p de manera progresiva sin asumir independencia estricta entre pruebas (Holm, 1979). Otra familia de métodos, ampliamente usada en estudios de asociación genética a gran escala, es la corrección del *FDR* (*False Discovery Rate*) propuesta por Benjamini y Hochberg (1995), que controla la proporción esperada de falsos positivos entre los resultados declarados significativos y ofrece mayor potencia en contextos de alta dimensionalidad. No obstante, muchas de estas aproximaciones asumen independencia entre las pruebas o ignoran la estructura de correlación debida al desequilibrio de ligamiento. Para estudios donde los marcadores presentan correlación moderada o alta, se han planteado métodos que estiman un número efectivo de pruebas independientes. En esta tesis se empleó el método de J. Li y Ji (2005), el cual utiliza los valores propios de la matriz de correlación entre marcadores para calcular dicho número efectivo. De esta manera, se obtiene un umbral de significación más realista, menos conservador que Bonferroni y más adecuado para datos genéticos correlacionados.

Para implementar el método de corrección propuesto por J. Li y Ji (2005), basado en la idea originalmente planteada por Cheverud (2001) para ajustar pruebas de hipótesis correlacionadas, se siguieron los siguientes pasos:

1. Se calculó la matriz de correlación entre todos los loci analizados.

2. Se estimó el número efectivo de pruebas independientes (M_{eff}) a partir de los valores propios de la matriz de correlación. Sea M el número total de pruebas y λ_i ($i = 1, \ldots, M$) los valores propios de dicha matriz. El número efectivo de pruebas se definió como:

$$M_{\text{eff}} = \sum_{i=1}^{M} f(\lambda_i), \tag{2.20}$$

donde la función $f(\cdot)$ se define como:

$$f(x) = \begin{cases} 1, & x \geq 1, \\ x - \lfloor x \rfloor, & 0 < x < 1, \end{cases} \tag{2.21}$$

siendo $\lfloor x \rfloor$ la función parte entera, que devuelve el mayor entero menor o igual que x, e $I(x \geq 1)$ una función indicadora que toma el valor 1 si $x \geq 1$ y 0 en caso contrario.

3. Se ajustó el nivel de significación global considerando M_{eff} pruebas independientes mediante la corrección de Šidák (Sidak, 1967):

$$\alpha^* = 1 - (1 - \alpha)^{1/M_{\text{eff}}}, \tag{2.22}$$

donde α se fijó en $0,05$.

4. Finalmente, se realizaron las M pruebas de hipótesis locus por locus y se rechazó la hipótesis nula de no asociación cuando el valor p de la prueba fue menor que α^*.

Resumen del capítulo

En síntesis, la estrategia metodológica seguida en este capítulo comprendió la generación y selección de los conjuntos de datos, la definición y verificación de escenarios con diferente grado de heterocedasticidad, la selección de la estructura de varianza-covarianza mediante modelos lineales mixtos, la obtención de los valores G-BLUP, el

ajuste de los modelos de asociación fenotipo-genotipo, la corrección por multiplicidad y la evaluación de su desempeño. La Figura 2.3 resume esta secuencia metodológica.

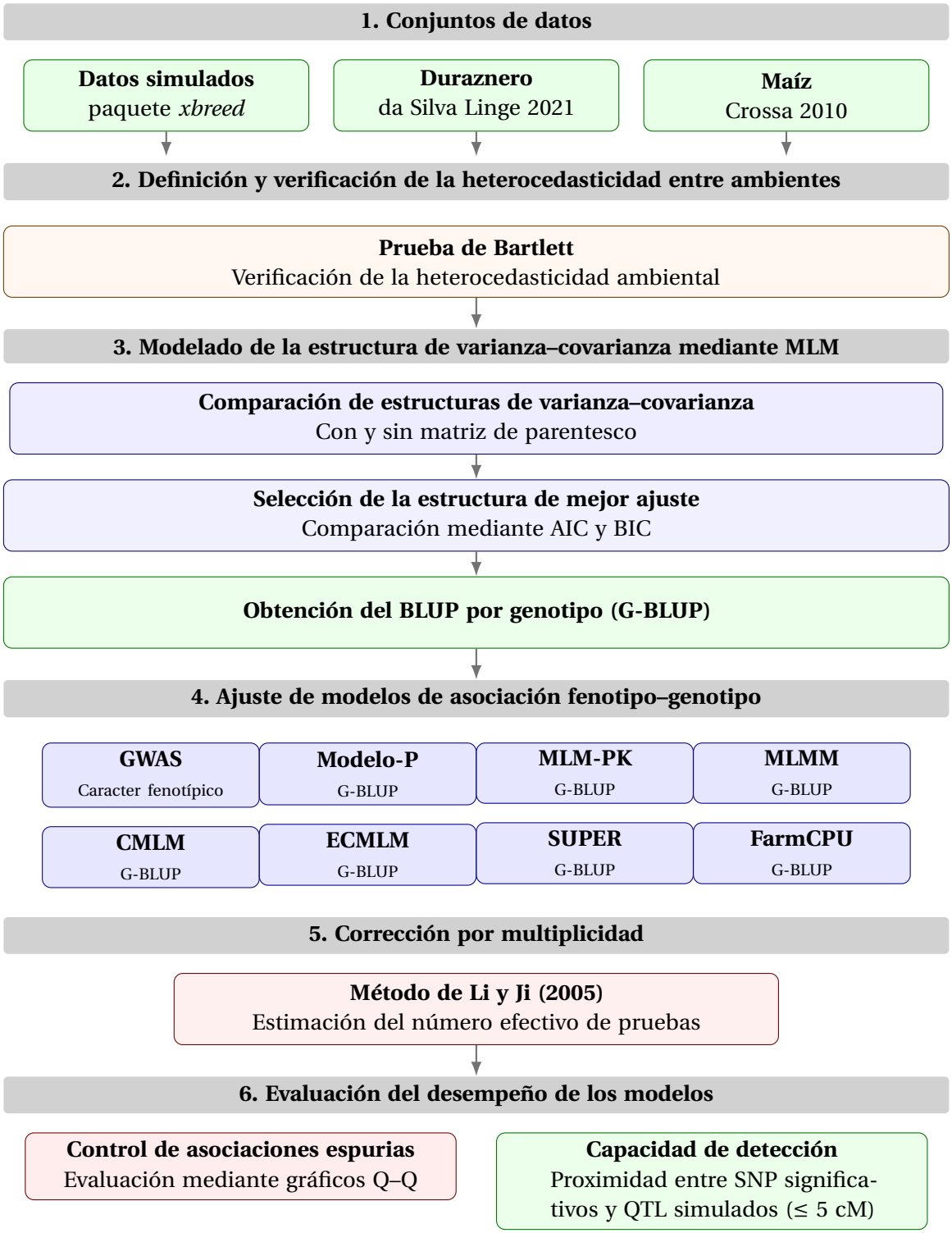

Figura 2.3: Secuencia metodológica seguida en el Capítulo 2 para evaluar el efecto de la heterocedasticidad ambiental sobre el desempeño de distintos modelos de asociación fenotipo–genotipo.

2.5. Resultados

2.5.1. Rendimiento relativo de las estructuras de varianza-covarianza en la estimación G-BLUP utilizando un conjunto de datos de simulación

La comparación de las estructuras de varianza–covarianza utilizando los criterios AIC y BIC mostró que el mejor ajuste se obtuvo al estimar el efecto de ambiente con varianza específica, es decir, sin estructura (Tabla 2.2). Los modelos de estimación con distintas estructuras de varianza-covarianza incluyen un efecto de interacción, se realizó una verificación de que el efecto IGA fuese estadísticamente significativo en nuestro conjunto de datos de simulación. Este valor resultó significativo según el test de cociente de verosimilitud con un valor p menor que $0,0001$ en el conjunto de datos cuyo escenario de simulación sin heterocedasticidad, un valor p= $0,028$ en el escenario de heterocedasticidad moderada y valor p= $0,001$ en el escenario de heterocedasticidad alta.

Tabla 2.2: Criterios AIC y BIC para la comparación de modelos según la estructura de varianza-covarianza propuesta en cada uno de los tres escenarios de variabilidad a través de los ambientes (SH - Sin heterocedasticidad; MH - Moderada heterocedasticidad; AH - Alta Heterocedasticidad)

Estructura de varianza-covarianza	**Kinship**	**Escenarios de niveles de varianza a través de los ambientes**					
		SH		MH		AH	
		AIC	BIC	AIC	BIC	AIC	BIC
Simetría Compuesta	NO	837,42	847,97	1142,63	1153,18	1316,94	1327,49
Simetría Compuesta	SI	801,30	811,85	1172,70	1183,25	1298,15	1308,70
Autorregresiva de Primer Orden	NO	1351,66	1356,93	1386,37	1394,64	1424,22	1429,49
Modelo Diagonal	NO	801,29	811,84	1136,58	1147,13	1265,53	1276,07
Modelo Diagonal	SI	836,04	846,58	1161,52	1172,06	1280,31	1290,85
Varianza no estructurada	NO	788,14	798,69	926,46	937,00	900,31	910,85
Varianza no estructurada	SI	819,47	830,02	1146,76	1157,30	1263,62	1274,17

Los valores predichos para cada genotipo (G-BLUP) de los efectos aleatorios de los

modelos lineales mixtos ajustados se utilizaron como variable respuesta en los modelos de mapeo asociativo.

2.5.2. Rendimiento relativo de los modelos GWAS utilizando un conjunto de datos de simulación

Para el escenario de homocedasticidad, en el modelo SUPER los valores p obtenidos del modelo ajustado se situaron por encima de la línea recta a 45° sin ninguna cola. Si la curva se hubiera ubicado sobre la línea, significa que la hipótesis nula no se rechaza y que no existe asociación estadísticamente significativa entre el fenotipo y los genotipos o que el modelo no tenía suficiente potencia para detectar posiciones de polimorfismo causal. Pero en el caso del modelo SUPER se ubicó por encima. Este resultado revela un elevado número de falsos positivos en el modelo SUPER (Figura 2.4). La cantidad de marcadores moleculares estadísticamente significativos en el escenario SH para el modelo SUPER fue de 87 (Tabla 2.3). Sin embargo, el número de falsos positivos en el modelo SUPER disminuyó en un escenario de mediana heterocedasticidad a 72 marcadores estadísticamente significativos y el mayor número marcadores estadísticamente significativos se detectó con alta heterocedasticidad, alcanzando un valor de 121 marcadores (Tabla 2.3). Esto se corresponde con el comportamiento observado en la Figura 2.4 donde para el modelo SUPER se observa una curva por encima de la recta de 45° y con la proporción de marcadores moleculares estadísticamente significativos que se encuentran a una distancia de menos de 5 cM, la cual representa solo el 3 % del total de marcadores identificados como estadísticamente significativos por el modelo (Tabla 2.3).

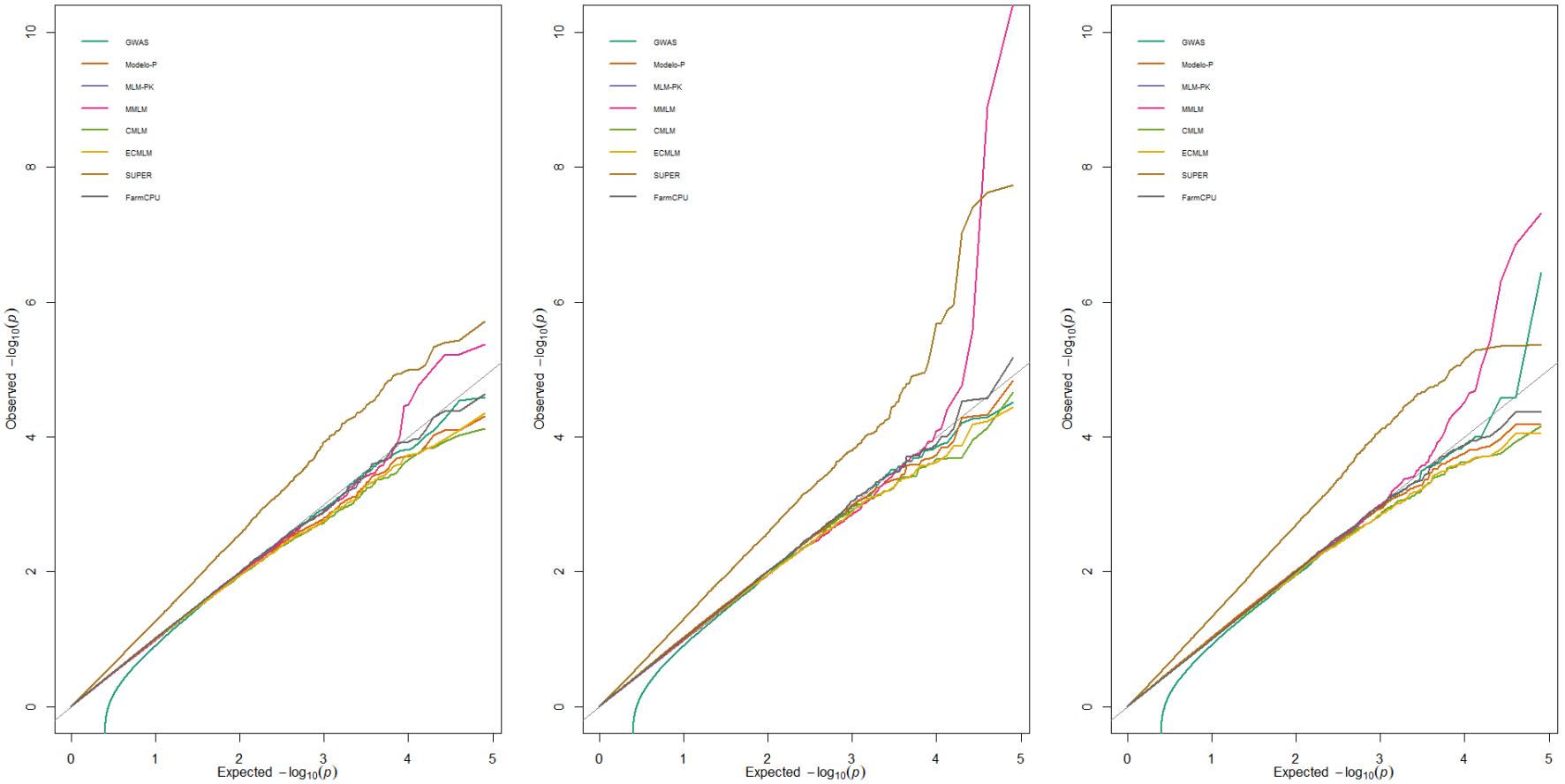

Figura 2.4: Distribución de la probabilidad uniforme del falso positivo y el falso negativo de ocho modelos (GWAS, Modelo-P, MLM-PK, MMLM, CMLM, ECMLM, SUPER y FarmCPU) en un conjunto de datos simulados por SNP para diferentes niveles de heterocedasticidad entre ambientes: Sin heterocedasticidad (izquierda); Moderada heterocedasticidad (centro) y Alta Heterocedasticidad (derecha).

En el escenario de homocedasticidad, para el modelo MMLM el comportamiento fue el esperado, es decir, los valores p del modelo ajustado se situaron en la línea recta cercana a la línea 1:1 con una cola aguda desviada hacia arriba cuando los valores p observados fueron superiores a los valores p esperados. En este modelo, el gráfico Q-Q indica que existen polimorfismos causales y de asociación verdaderos, y tanto los falsos positivos como los falsos negativos fueron controlados, se encontraron 11 marcadores moleculares estadísticamente significativos (Tabla 2.3). Los modelos CMLM, MLM-PK, ECMLM, Modelo-P y FarmCPU mostraron una desviación del valor p por debajo de la línea recta, indicando que había falsos negativos (Figura 2.4). El número de marcadores moleculares detectados como falsos negativos para estos modelos fue de 2, 4, 4, 5 y 11, respectivamente (Tabla 2.3).

Tabla 2.3: Marcadores moleculares estadísticamente significativos luego de la corrección por multiplicidad identificados por el método de Li y Ji en distintos modelos de asociación (SH - Sin heterocedasticidad; MH - Moderada heterocedasticidad; AH - Alta Heterocedasticidad).

Modelos de asociación	Niveles de heterocedasticidad entre ambientes		
	SH	MH	AH
GWAS	6 (33 %)	8 (25 %)	8 (75 %)
Modelo-P	5 (60 %)	5 (40 %)	4 (100 %)
MLM-PK	4 (75 %)	5 (60 %)	2 (50 %)
MMLM	11 (63 %)	11 (100 %)	17 (100 %)
CMLM	3 (33 %)	3 (33 %)	2 (50 %)
ECMLM	4 (75 %)	5 (60 %)	2 (50 %)
SUPER	87 (56 %)	72 (64 %)	121 (58 %)
FarmCPU	11 (54 %)	8 (50 %)	9 (78 %)

Para cada modelo evaluado, en cada escenario de heterocedasticidad, se midió el número de marcadores moleculares estadísticamente significativos (valor p $\leq 0,000143$; valor obtenido con el método de Li y Ji). Dado que se conocía la posición de los QTL simulados, se registró el número de marcadores dentro de intervalos de distancia de 5 cM, 10 cM, 15 cM y 20 cM (Tabla 2.3).

Definimos como Error de tipo I aquellos casos en los que un modelo identifica marcadores como significativos que no se encuentran próximos a ningún QTL, es decir, corresponden al complemento del porcentaje reportado en la Tabla 2.3 para

cada modelo. En cuanto al Error de tipo II, este se refiere a los marcadores que no son detectados como significativos pese a estar ubicados cerca de un QTL. Los valores estimados de este error fueron comparables entre modelos, lo cual puede atribuirse a la elevada densidad de marcadores evaluados por cromosoma (aproximadamente 10.000), de los cuales aproximadamente el 50 % se encuentra en proximidad física a un QTL. En contraste, cada modelo y escenario detectó, en promedio, únicamente 17 marcadores significativos

Cuanto más cerca está un marcador molecular de un QTL simulado, más probable es que identifique con precisión un verdadero positivo, o una región del genoma asociada a un rasgo específico. Esto permitió identificar, para cada modelo en cada escenario, si los marcadores identificados como estadísticamente significativos podrían considerarse verdaderos positivos, es decir, asociados a un QTL simulado. Es importante señalar que en la simulación *Xbreed*, ninguna de las posiciones de los QTL simulados coincidía con la ubicación exacta de un marcador molecular (Esfandyari y Sørensen, 2019). En el proceso de simulación, tanto la posición de los marcadores moleculares como la del QTL se simularon aleatoriamente. En el escenario homocedástico, el modelo GWAS identificó seis marcadores moleculares estadísticamente significativos, de los cuales el 33 % estaban a menos de 5cM de distancia de algunos de los 120 QTL simulados. Por el contrario, en el escenario de moderada heterocedasticidad sólo el 25 % de los 8 marcadores moleculares identificados como estadísticamente significativos se encontraba dentro de los 5cM de un QTL simulado. Esta situación podría reflejar un descubrimiento falso generado por la homocedasticidad. El 75 % de los marcadores moleculares identificados como estadísticamente significativos por los modelos MLM-PK y ECMLM se encontraban dentro de los 5cM del QTL simulado. Este registro sugiere que esos marcadores moleculares podrían considerarse verdaderos positivos, aunque el gráfico Q-Q indicaba que su distribución por encima de la línea recta 1:1 podría sugerir una falta de asociación causal entre el fenotipo y el genotipo. El Modelo-P mostró un rendimiento similar. Se identificó el 60 % de los marcadores moleculares estadísticamente significativos a una distancia inferior a 5cM. Según el gráfico Q-Q, el modelo SUPER presentaba un mayor número de falsos positivos. Sin embargo, solo el 56 % de los marcadores moleculares estadísticamente significativos se encontraba dentro de

los 5cM. Esta cantidad es superior al número de marcadores moleculares verdaderos identificados por MLM-PK y ECMLM. En el escenario de heterocedasticidad moderada, MMLM fue el modelo que presentó el mayor porcentaje (100 %) de marcadores moleculares significativos a una distancia inferior a 5cM. Para este escenario, todos los modelos, excepto SUPER, encontraron una cantidad similar o menor de marcadores moleculares significativos, después de corregir la multiplicidad. La proporción de marcadores moleculares a una distancia inferior a 5cM en cada modelo representa la misma cantidad de QTL para los modelos MLM-PK y ECMLM, se detectaron 5 marcadores moleculares y, en ambos modelos, esta cantidad de marcadores moleculares representa el 60 % de ellos a una distancia inferior a 5cM. En el escenario de alta heterocedasticidad, los modelos MMLM y Modelo-P fueron los que tuvieron el porcentaje más bajo de falsos positivos (100 %), ya que todos los marcadores significativos se detectaron a una distancia inferior a 5cM del QTL simulado. El modelo SUPER, al igual que en otros escenarios, indicó el mayor número de marcadores significativos. También se evaluaron las distancias de los marcadores moleculares significativos a más de 5cM de un QTL simulado (Figura 2.5). Se observó que el modelo SUPER presentó el mayor número de marcadores moleculares significativos, aunque localizados a más de 5 cM de distancia, e incluso a más de 20 cM de un QTL simulado (Figura 2.5).

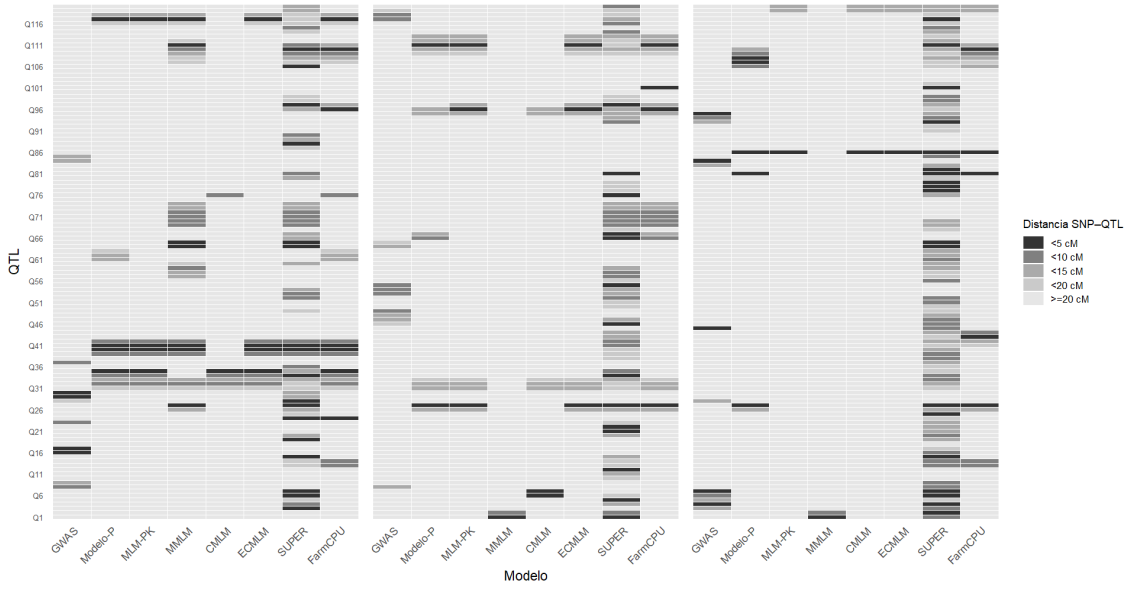

Figura 2.5: Mapa de calor de los marcadores estadísticamente significativos en cada modelo ajustado luego de la corrección por multiplicidad por Li y Ji. Los colores indican la distancia desde el marcador hasta un QTL simulado, los colores más oscuros indican una mayor proximidad Sin heterocedasticidad (izquierda); Moderada heterocedasticidad (centro) y Alta Heterocedasticidad (derecha).

2.5.3. Aplicación de los modelos GWAS en una base de datos de duraznero

En esta base de datos la evaluación del efecto de la interacción genotipo-ambiente fue estadísticamente significativa para los dos caracteres fenotípicos evaluados , BD y DAB. El resultado del valor p fue $< 0,0001$ y $0,359$ respectivamente. En DAB, todos los modelos de asociación fenotipo-genotipo presentaron curvas cercanas a la distribución uniforme teórica. Por otro lado, el BD presentó falsos negativos en los modelos SUPER y GWAS, y los mejores resultados, es decir cuyas distribuciones se ajustaron a la recta 1:1 y se despegó en las puntas, se obtuvieron en los modelos Farm-CPU y MMLM (Figura 2.6). El modelo SUPER identificó marcadores moleculares estadísticamente significativos para ambos caracteres fenotípicos (Tabla 2.4).

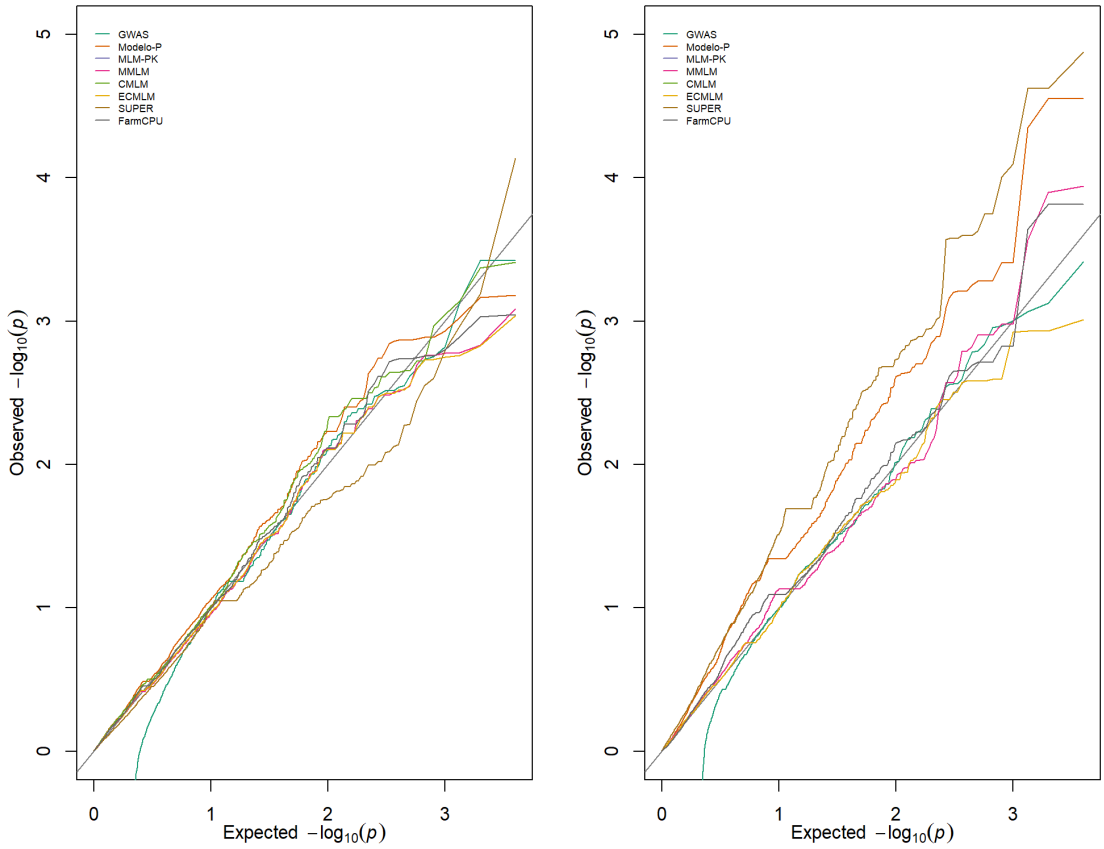

Figura 2.6: Distribución de la probabilidad uniforme del falso positivo y el falso negativo en ocho modelos (GWAS, Modelo-P, MLM-PK, MMLM, CMLM, ECMLM, SUPER y FarmCPU) en duraznos genotipados por SNP para diferentes niveles de heterocedasticidad las variables DAB (izquierda) y BD (derecha)

Tabla 2.4: Cantidad de marcadores moleculares estadísticamente significativos identificados por el método Li y Ji en ocho modelos de asociación, que incluyen: GWAS, Modelo-P, MLM-PK, MMLM, CMLM, ECMLM, SUPER y FarmCPU para las variables DAB (izquierda) y BD (derecha)

Modelos de asociación	Variables	
	DAB	BD
GWAS	0	0
Modelo-P	0	3
MLM-PK	0	0
MMLM	0	1
CMLM	0	0
ECMLM	0	0
SUPER	1	5
FarmCPU	0	0

Nota: DAB: *Days After Bloom* (días después de la floración) y BD: *Bloom Date* (Días a floración).

2.5.4. Aplicación de los modelos GWAS en una base de datos de maíz

En esta sección se muestran los resultados obtenidos al ajustar los modelos sobre un conjunto de datos públicos de maíz. Previo al ajuste del modelo que incluía los marcadores moleculares, se evaluó el efecto de interacción estadísticamente significativo para cada caracter fenotípico (GY, ASI, FFL y MFL) y para cada escenario (homocedasticidad, heterocedasticidad media y heterocedasticidad alta). Los resultados indicaron un efecto estadísticamente significativo (valor $p < 0,0001$) del término IGA para todos los rasgos fenotípicos evaluados bajo los tres escenarios de variabilidad entre ambientes.

Los ocho modelos identificaron diferentes números de marcadores moleculares estadísticamente significativos asociados a cada caracter fenotípico en los dos escenarios de heterocedasticidad (Figura 2.7 y Tabla 2.5). Para el primer caracter fenotípico GY, los gráficos Q-Q de los modelos MMLM y ECMLM presentaron una línea recta con una cola ligeramente desviada por debajo de la modelo uniforme en los tres escenarios, lo que indicaba que estos modelos encontraban falsos positivos (Figura 2.7). En el escenario sin heterocedasticidad, los gráficos Q-Q de los modelos SUPER y GWAS mostraron que

los valores p obtenidos del modelo ajustado se situaban por encima de la línea recta a 45° sin ninguna cola y siguiendo una distribución uniforme. El gráfico Q-Q de los modelos FarmCPU y Modelo-P mostró una desviación de la línea recta cercana a 1:1, lo que rechazó la hipótesis nula (Figura 2.7).

Sin embargo, ningún marcador molecular fue estadísticamente significativo en el escenario de homocedasticidad (Tabla 2.5). El modelo GWAS se ajustó a partir de los datos GY en lugar de G-BLUP, en el escenario de heterocedasticidad moderada y alta, y encontró un marcador molecular significativo. Por otro lado, los modelos identificaron una asociación significativa. Si es cierto que no hay asociación, los marcadores moleculares significativos encontrados en heterocedasticidad son falsos negativos. El mismo comportamiento se observó para la variable fenotípica FFL, donde se asume que no hay una asociación significativa presente porque, en homocedasticidad, ningún marcador molecular fue estadísticamente significativo. Sin embargo, el modelo GWAS detectó dos marcadores moleculares estadísticamente significativos (Tabla 2.5). El gráfico Q-Q mostró que los valores p obtenidos del modelo ajustado se situaban por encima de la línea recta a 45° sin ninguna cola lo que sugiere una distribución uniforme. Esta distribución uniforme significa que la hipótesis nula es cierta y que no existe una asociación estadísticamente significativa entre el genotipo y el fenotipo, y que los dos marcadores moleculares son falsos positivos, al igual que el marcador molecular encontrado por MMLM. Es posible que la heterocedasticidad se haya confundido con la correlación genética. Para la variable ASI, los modelos GWAS y MMLM arrojaron marcadores moleculares estadísticamente significativos en el primer escenario, es decir, con varianzas homogéneas. Sin embargo, sólo el modelo MMLM detectó una asociación estadísticamente significativa con alta heterocedasticidad. En este escenario, el Modelo-P con la incorporación de la estructura poblacional como covariable y los falsos positivos controlados que esperábamos, detectó dos marcadores moleculares estadísticamente significativos (Tabla 2.5). En este caso, el modelo de locus único tuvo el mismo resultado que el enfoque del MMLM. Se observó una visualización similar en el caracter fenotípico MFL, donde los modelos Modelo-P, MMLM y SUPER presentan marcadores moleculares significativos en el escenario de homocedasticidad, pero solo el Modelo-P identificó marcadores moleculares estadísticamente significativos en el

escenario de alta heterocedasticidad. Esto puede estar relacionado a que la estructura genética poblacional podría ser confundida con la heterocedasticidad (Crossa et al., 2010).

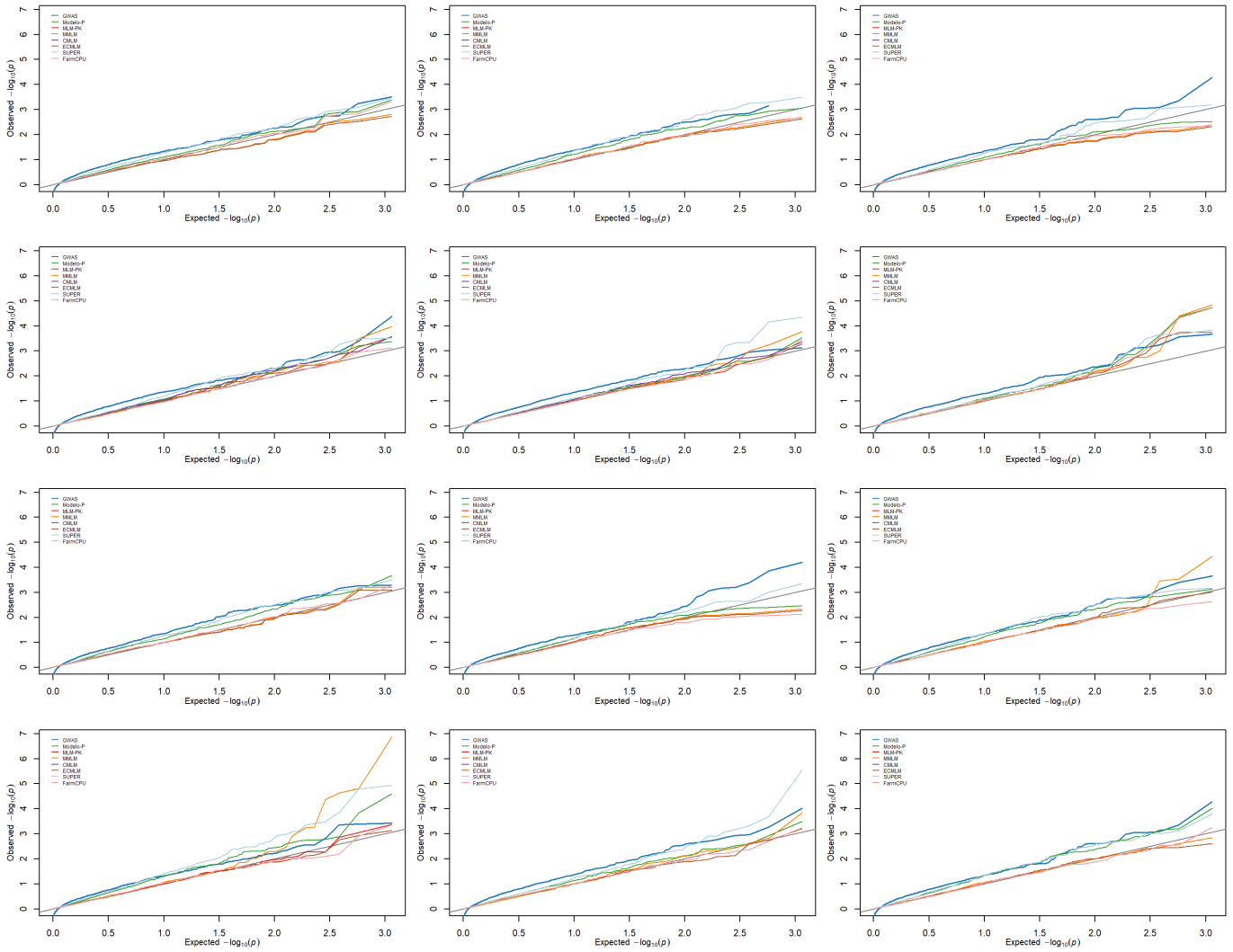

Figura 2.7: Distribución de la probabilidad de ocho modelos (GWAS, Modelo-P, MLM-PK, MMLM, CMLM, ECMLM, SUPER y FarmCPU) en maíz genotipado por SNP para diferentes niveles de heterocedasticidad en GY (primera fila), ASI (segunda fila), FFL (tercera fila) y MFL (cuarta fila) rasgos. Sin heterocedasticidad (izquierda); Moderada heterocedasticidad (centro) y Alta Heterocedasticidad (derecha)

Tabla 2.5: Cantidad de marcadores moleculares estadísticamente significativos identificados luego de la corrección por multiplicidad por el método Li y Ji en ocho modelos de asociación que incluyen a GWAS, Modelo-P, MLM-PK, MMLM, CMLM, ECMLM, SUPER y FarmCPU para rasgos GY, ASI, FFL y MFL (SH - Sin heterocedasticidad; MH - Moderada heterocedasticidad; AH - Alta Heterocedasticidad)

Modelos de asociación	GY			ASI			FFL			MFL		
	SH	MH	AH	SH	MH	AH	SH	MH	AH	SH	MH	AH
GWAS	0	1	1	1	0	0	0	2	0	0	0	0
Modelo-P	0	0	0	0	0	2	0	0	0	1	0	1
MLM-PK	0	0	0	0	0	0	0	0	0	0	0	0
MMLM	0	0	0	1	0	2	0	0	1	4	0	0
CMLM	0	0	0	0	0	0	0	0	0	0	0	0
ECMLM	0	0	0	0	0	0	0	0	0	0	0	0
SUPER	0	0	0	0	2	0	0	0	0	3	1	0
FarmCPU	0	0	0	0	0	2	0	0	0	0	0	0

Nota: GY: *grain yield* (rendimiento de grano), ASI: *anthesis-silking interval* (intervalo de tiempo transcurrido entre antesis y floración), FFL: *female flowering* (días a floración femenina) y MFL: *male flowering* (días a floración masculina).

2.6. Discusión

Los estudios de asociación genética de plantas están interesados en identificar variantes genéticas que actúan junto con otros modificadores ambientales para influir en la expresión del fenotipo. Cuando se detecta IGA usando un modelo estadístico lineal de clasificación considerando solamente el efecto fenotípico y no incluyendo la información genómica en el mismo, es común interpretar la interacción en términos de medias genéticas y ambientales específicas incluidas en el modelo de interacción. Sin embargo, la IGA puede surgir de manera más general cuando el predictor genético o ambiental modera la cantidad total de variación en el fenotipo. Cuando la cantidad total de variación en un fenotipo varía a lo largo del rango de una medida ambiental, los modelos multiplicativos de IGA que asumen homocedasticidad detectan interacciones entre esa medida de variabilidad genética y todos los demás correlaciones genotípicas, tanto genéticas como no genéticas (J. Yang et al., 2012). En este trabajo, ilustramos el efecto de la heterocedasticidad en la detección de la asociación de fenotipo y genotipo

cuando la interacción IGA está presente. En esta tesis se utilizaron los G-BLUP estimados bajo escenarios de heterocedasticidad entre ambientes debido a que el efecto fenotípico está ponderado por la varianza ambiental. Debido a que los tres escenarios no son iguales, el G-BLUP podrían no estar capturando toda la variabilidad debido a los genotipos. Esto estaría provocando que la variabilidad no explicada aumente en la varianza residual aún más en los escenarios con mediana y alta heterocedasticiad que en el modelo homocedástico. Por eso, el modelo GLM cuando se cumplen los supuestos de homocedasticidad, normalidad e independencia entre los errores experimentales, funciona mejor que bajo escenarios de heterocedasticidad.

Cuando no se conoce la verdadera correlación genética entre los individuos, en general estimada a través de la matriz K, se estima una matriz de aditividad entre los marcadores moleculares, que finalmente es una matriz de similitud molecular (Ad). Dado que los marcadores moleculares son herramientas biotecnológicas cada vez más disponibles para el uso complementario en el mejoramiento genético vegetal y que en algunos cultivos no es posible conocer la información del pedigré, necesaria para estimar K, entonces utilizar una matriz Ad basada en el genoma construida a partir de similitudes entre pares de individuos representó un avance en los estudios de asociación fenotipo-genotipo en plantas. Otro punto de vista fue planteado por Astle y Balding (2009) donde postulan que la matriz de información genotípica, G, transmite información sobre la estructura genética de la población y por lo tanto información respecto a la relación parental que existe entre pares de individuos, por lo que podría no ser útil considerar la información genética provista desde distintas matrices, es decir, la provista por la matriz Ad y la provista por la matriz K que suele usarse en los Modelos-PK, como covariables de efectos fijos en el modelo. La discusión anterior refleja una falta de consenso en el campo GWAS sobre como incorporar la información molecular para poder estimar los efectos del genotipo y su asociación con el fenotipo observado. S. Gianola et al. (2016), en el contexto del mejoramiento genético animal, abordaron en profundidad la construcción e interpretación de la matriz G, discutiendo su uso como reemplazo de la información de parentesco tradicional y los alcances y limitaciones en problemas de predicción genética y asociación. Si bien su trabajo se desarrolla en genética animal, los fundamentos estadísticos y conceptuales de la

utilización de matrices genómicas resultan directamente transferibles a los estudios GWAS en plantas.

Otros estudios han demostrado que en contextos de bajo LD y baja estructura genética, la incorporación de CP como efecto fijo (P) puede producir una sobre parametrización del modelo que conduce a un aumento de la tasa de falsos positivos (Peña-Malavera et al., 2014). Lo mismo se repite en los resultados de esta tesis, donde se encontraron más casos de falsos positivos en los escenarios de homocedasticidad. En la base simulada se puede observar que de los marcadores moleculares significativos encontrados por el Modelo-P, un bajo porcentaje se acerca al verdadero QTL. Mientras, al ajustar el Modelo-P en bases de datos públicas, donde no conocemos donde se encuentran ubicados verdaderamente los marcadores moleculares a la expresión fenotípica, vemos que este modelo es el que encuentra mayor cantidad de marcadores moleculares estadísticamente significativos. Estos resultados son similares a los obtenidos para Kaler et al. (2020).

Se considera que la probabilidad de identificar un verdadero positivo aumenta a medida que disminuye la distancia física entre el marcador molecular y el QTL. En este sentido, la proximidad del SNP al QTL es un factor determinante: cuanto menor es la distancia entre ambos, mayor es la probabilidad de que el marcador se encuentre en fuerte desequilibrio de ligamiento con el locus causal. Idealmente, un SNP localizado dentro del gen candidato o coincidente con la variante funcional proporciona la evidencia más robusta de asociación. Por esta razón, los estudios de simulación, como los realizados en este trabajo, han permitido identificar el QTL asociado a valores diferenciales del fenotipo para luego poder evaluar qué modelos detectan dichos QTL a través de la significancia de los marcadores moleculares. Cuanto mayor sea el efecto de QTL aún los SNP más distantes podrían ser identificados como marcadores estadísticamente significativos. Si bien es importante detectar la estructura genética de la población para reducir los falsos positivos, en el caso de CMLM, donde los individuos se comprimen en grupos y la similitud está basada entre pares de grupos y no entre pares de individuos, el método de agrupamiento UPGMA no resultó ser eficiente para identificar similitud entre grupos de individuos. Este resultado coincide con los encontrados por Videla et al. (2021). donde el método de agrupamiento entre individuos resultó ser menos

eficiente que los métodos k-means y basados en modelos bayesianos para detectar alta cantidad de grupos de individuos genéticamente similares. Si no existe una Estructura Genética en la Población, el CMLM debería comportarse u obtener los mismos resultado o al menos similares a los obtenidos con los modelos Modelo-P y MLM-KP. CMLM detectando GPS debería funcionar mejor que MLM-KP o Modelo-P. UPGMA agrupa a individuos similares en grandes grupos y separa a diferentes individuos solos, generando así muchos grupos desequilibrados. Aunque la idea de reducir la confusión de la matriz K o P, UPGMA, al generar grupos con pocos individuos, pierde fuerza para el modelo (Videla et al., 2021).

El método MMLM selecciona marcadores SNP a través de un proceso de regresión lineal paso a paso (hacia adelante y hacia atrás) e incluye marcadores significativos como cofactores. Este método no realiza un ACP ni estima una matriz de parentesco K con los marcadores moleculares estadísticamente significativos. Vilhjálmsson y Nordborg (2013) sugiere que se podría estimar una matriz K excluyendo los SNP que no fueran significativos. Sin embargo, consideraron que esto podría ser simplemente insignificante en cuanto al ajuste del modelo, ya que la regresión paso a paso ayuda a filtrar, limpiar o preseleccionar marcadores potencialmente significativos para explicar la asociación fenotipo-genotipo. De esta manera, el modelo MMLM selecciona marcadores moleculares, es decir, variables regresoras (SNP) que explican la asociación y de esta manera espera ganar precisión al disminuir falsos positivos. En este sentido, en este estudio de simulación el método MMLM fue el modelo, después de SUPER, que detectó el mayor número de marcadores moleculares estadísticamente significativos, detectando la misma cantidad de marcadores moleculares estadísticamente significativos que el modelo FarmCPU en el escenario de homocedasticidad, pero más que este último modelo en los escenarios de mediana y alta heterocedasticidad.

Finalmente, el método FarmCPU implementa un ciclo de iteraciones considerando los efectos fijos, en el cual los marcadores se contrastan con aquellos los marcadores moleculares previamente asociados, es decir, que en el ciclo anterior de iteración fueron estadísticamente significativos. Este contraste entre los marcadores moleculares significativos en un ciclo de iteración respecto a los estadísticamente significativos en otro ciclo de iteración es usado en lugar de la matriz de parentesco como lo hacen los

modelos MLM, CMLM, ECMLM, SUPER y MMLM. En el ciclo del modelo de iteraciones considerando los marcadores estadísticamente significativos como efectos aleatorios, los marcadores se seleccionan entre un pequeño número de marcadores asociados utilizando el método de máxima verosimilitud para evitar el problema de sobreajuste del modelo. Esta es una propuesta para resolver el problema del modelo MMLM, que selecciona el o los marcadores estadísticamente asociados al fentoipo entre todos los marcadores disponibles. En consecuencia, FarmCPU exhibe una mayor potencia estadística que MMLM. FarmCPU reduce el tiempo de cálculo computacional respecto a MMLM, brinda resultados confiables al eliminar de manera eficiente la confusión entre la estructura de la población y del parentesco, evitando el ajuste excesivo del modelo y controlando los falsos positivos (Liu et al., 2016). Como FarmCPU prueba los marcadores en un modelo de efectos fijos, es computacionalmente más eficiente que los métodos que prueban los marcadores en el modelo de efectos aleatorios, como MLM, CMLM, ECMLM, SUPER y MMLM. En nuestro estudio de simulación FarmCPU presentó un comportamiento similar a MMLM en los tres escenarios. Sin embargo, en los conjuntos de datos reales usados como ilustración, no detectó ningún marcador estadísticamente significativo excepto para una sola variable fenotípica (ASI) en maíz donde encontró dos marcadores estadísticamente significativos en el escenario de alta heterocedasticidad. En dicho escenario, para esa variable, el modelo MMLM también detectó un solo marcador estadísticamente significativo. Lo cual asumimos como falso positivo dado que bajo el escenario de homocedasticidad no se detectaron marcadores moleculares estadísticamente significativos. Los resultados aquí presentados coinciden con estudios previos de Crossa et al. (2010) en maíz y da Silva Linge et al. (2021) en durazno. En cuanto al rendimiento de grano (GY), casi todos los modelos identificaron al menos un marcador significativo, tal y como informaron Crossa et al. (2007), quienes identificaron cuatro marcadores. Sin embargo, es importante aclarar que las repeticiones para este conjunto de datos fueron simuladas por no encontrarse disponibles.

2.7. Conclusión

Se compararon ocho modelos estadísticos bajo distintos niveles de heterocedasticidad. La evaluación incluyó un carácter simulado junto con marcadores moleculares y QTL, considerado como la "verdad" a detectar, cuatro variables empíricas en maíz y dos variables fenotípicas en duraznero, lo que permitió contrastar el desempeño de los modelos en un cultivo anual y en uno perenne.

De acuerdo con los gráficos Q–Q y el número de marcadores significativos, el modelo SUPER presentó una elevada proporción de falsos positivos, evidenciando inflación del error tipo I. El modelo GWAS aplicado directamente sobre el valor fenotípico observado, en lugar de utilizar los G-BLUP, no mostró mayor eficiencia y, bajo alta heterocedasticidad, incrementó las asociaciones espurias. Los restantes modelos exhibieron desempeños similares entre sí.

La heterocedasticidad tendió a enmascarar los efectos genotípicos medios. Si bien el uso de matrices de covarianza no estructuradas permitió modelar el ruido introducido por la variabilidad residual, ello no se tradujo en una mayor detección de marcadores asociados a los QTL simulados.

En los datos reales, SUPER aplicó una selección estricta de SNP en cada iteración, lo que en algunos casos generó problemas de estimación por insuficiencia de información. En contraste, el modelo MMLM, al considerar los marcadores como efectos fijos, mostró un proceso de estimación más estable y eficiente, permitiendo discriminar con mayor precisión la señal genética del ruido asociado a la heterocedasticidad. La coherencia de los resultados entre datos simulados y empíricos sugiere que estos enfoques pueden aplicarse de manera consistente en distintas especies y contextos experimentales.

Capítulo 3

Nuevo enfoque para representar la asociación entre fenotipos y marcadores moleculares mediante Mínimos Cuadrados Parciales

Los estudios GWAS se han convertido en una herramienta central para identificar relaciones entre marcadores moleculares y características fenotípicas de interés en el mejoramiento genético de cultivos (C. Wang et al., 2025; Dai et al., 2024). No obstante, el análisis estadístico en estudios GWAS plantea desafíos significativos, como la alta dimensionalidad de los datos genómicos, donde se evalúa una gran cantidad de SNP en comparación con el número de observaciones disponibles. También es habitual observar multicolinealidad entre los rasgos fenotípicos evaluados que asumen independencia entre las variables de respuesta (P. Vatcheva y Lee, 2016). Por ejemplo, en caracteres agronómicos relacionados con el crecimiento, la madurez o el rendimiento, las variables medidas suelen estar altamente correlacionadas debido a que responden a procesos biológicos comunes o compartidos. Esta redundancia en la información puede afectar negativamente a los métodos estadísticos que asumen independencia entre las variables respuesta. Estas características comprometen la eficacia de los modelos tradicionales de GWAS y hacen necesario recurrir a métodos multivariados que puedan capturar la estructura latente en los datos y mejorar la detección de asociaciones relevantes (Korte y Farlow, 2013).

3.1. Modelos de Regresión por Mínimo Cuadrados Parciales en estudios GWAS

Los métodos estadísticos multivariados para el análisis de datos genómicos de alta dimensión han sido objeto de numerosas publicaciones en estadística, aprendizaje automático, bioinformática y biología (Vershynin, 2018; Hastie et al., 2009; Elad, 2010). Entre estos métodos, el ACP ha sido ampliamente utilizado para reducir la dimensionalidad de los datos preservando la mayor parte de la variabilidad original (I. Jolliffe, 2002). No obstante, su naturaleza no supervisada puede limitar su capacidad para identificar relaciones específicas entre genotipos y fenotipos. En cambio, la regresión por PLS (H. Wold, 1966) ofrece una alternativa poderosa al modelar la relación entre dos matrices X y Y, donde X representa una matriz de predictores (genotipos) y Y representa la matriz con las variables respuesta (fenotipos). Esta técnica es especialmente útil en contextos con alta multicolinealidad, ya que construye componentes latentes ortogonales que maximizan la covarianza entre ambos conjuntos de variables (Wondola et al., 2020; Ortiz et al., 2023). Las variables latentes se construyen de manera iterativa para cada matriz y representan las verdaderas regresoras para X y Y (Montesinos-López et al., 2022).

La obtención de estas variables latentes surgen de algoritmos algebraicos sobre las matrices. Se han propuesto diversos algoritmos para realizar la descomposición de matrices de datos multivariados, entre ellos SVD (Lange, 2010), , es frecuentemente aplicada a diversas técnicas de análisis multivariado de reducción de dimensionalidad como el ACP (Shen, 2009; I. Jolliffe, 2002), análisis discriminante lineal (Fisher, 1936) y análisis de conglomerados(Kaufman y Rousseeuw, 1990), entre otros. Otro algoritmo propuesto como alternativa a estimaciones no lineales es el algoritmo iterativo no lineal NIPALS (H. Wold, 1973). Existen otras aproximaciones para datos multivariados como SIMPLS (Implementación sencilla del PLS basado en matrices, por sus siglas en inglés *Straightforward Implementation of Matrix-based PLS*) (De Jong, 1993), que ha sido utilizado por Boulesteix y Strimmer, 2006 en un contexto de análisis de datos genómicos con variables respuesta de tipo categóricas. A pesar de que PLS no fue concebido inicialmente para estudios de asociación, su utilidad en GWAS para la selección de

variables y la identificación de genes asociados a caracteres fenotípicos de interés han sido propuestos por Broc et al., 2021. Los resultados de la regresión PLS pueden visualizarse mediante gráficos biplot o triplot (Vicente-Gonzalez y Vicente-Villardon, 2022), que permiten representar simultáneamente, en un espacio de proyección óptima, es decir, sin pérdida de generalidad, los individuos (filas de la matriz), las variables genotípicas (matriz X) y las variables fenotípicas (matriz Y), facilitando la interpretación conjunta de los datos (Schillemans et al., 2019) y permitiendo esclarecer las relaciones subyacentes entre X y Y. Esta representación resulta especialmente útil para identificar patrones de asociación en estudios multivariados complejos.

En este capítulo se propone un enfoque multivariado de reducción de dimensión basado en PLS para identificar asociaciones entre caracteres fenotípicos y marcadores moleculares. Para ello se comparan dos algoritmos para la construcción de variables latentes: SVD y NIPALS. Además, se propone una metodología para identificar los marcadores estadísticamente asociados a los caracteres agronómicos de interés de manera multivariada. Esta contribución busca fortalecer el desarrollo de herramientas multivariadas para el análisis integrado de datos fenotípicos y genotípicos en programas de mejoramiento genético vegetal.

3.2. Objetivo

Evaluar el desempeño de la regresión PLS para estudios de asociación fenotipo-genotipo, aplicando tanto el algoritmo SVD como el algoritmo NIPALS, en su capacidad para detectar marcadores moleculares estadísticamente significativos.

3.3. Materiales y Métodos

La metodología propuesta se aplica tanto a un conjunto de datos simulados como a dos conjuntos de datos reales públicos usados como ilustración. En el caso de los datos simulados y la base de datos pública de maíz, coinciden con las descritas detalladamente en la sección de Materiales y Métodos del Capítulo 2. Con la finalidad de facilitar la

lectura, se realiza una breve descripción de las mismas. Debido a que la motivación de esta tesis es el cultivo de duraznero, se utilizó para ilustración una base de datos argentina de esta especie (Sanchez, 2013; Valentini y Arroyo, 2007).

3.3.1. Descripción de la simulación

Los datos fueron simulados con el paquete *xbreed* de R. Este paquete, simuló en una primera etapa una generación histórica a partir de la cual crea un nivel deseable de LD y luego en una segunda instancia, generó la estructura de una población reciente (Esfandyari y Sørensen, 2019). Los parámetros utilizados en la simulación fueron: 500 individuos como tamaño de la población histórica, 300 como número de generaciones, $0,4$ como heredabilidad en sentido estricto, $0,1$ como relación entre la varianza de dominancia y la varianza fenotípica, $0,0005$ como tasa de mutación, $0,05$ como QTL segregantes en la última generación histórica, $0,05$ como marcadores segregantes en la última generación histórica y $0,5$ como frecuencias alélicas de loci (marcador y QTL) en la primera generación histórica. Para la conformación del genoma, emulando la especie *Prunus persica* (L.) Batsch, se consideraron ocho cromosomas con 10.000 marcadores en cada uno. Cada cromosoma fue constituido con una longitud de 260cM y 15 QTL. Luego de finalizar la simulación, se eliminaron los loci con frecuencia alélica inferior a $0,05$. El paquete xbreed simula tanto matrices de datos genotípicos como su caracter fenotípico asociado a la información molecular. En este Capítulo se utilizaron y^{A1} y y^{A2} como variables fenotípicas cuyos detalles de su obtención se explicaron en el Capítulo 2.

3.3.2. Ilustración a través de un ejemplo de una base de datos real de duraznero

Se trabajó con un conjunto de datos con información proveniente de 74 variedades de *Prunus persica* (L.) Batsch evaluadas durante los periodos 2010/2011, 2011/2012 para tres caracteres agronómicos: (1) rendimiento (Kg/árbol), (2) número de frutos por árbol, (3) peso medio de los frutos (Kg) y (4) calibre ponderado. Las variedades fueron genotipos con marcadores moleculares del tipo SNP. Para aplicar la regresión PLS propuesta en este capítulo de tesis se trabajó con 619 SNP sin datos faltantes.

3.3.3. Ilustración a través de un ejemplo de una base de datos real de maíz

El conjunto de datos de maíz procede del programa mundial de mejora genética de este cultivo del CIMMYT y contiene información de 284 variedades a las que se le midieron tres caracteres fenotípicos (1) floración femenina, días hasta la floración (FFL); (2) floración masculina, días hasta la antesis (MFL); y (3) el intervalo entre la antesis y la floración (ASI). Estos caracteres son evaluados en dos ambientes: WW - bien regado (*well-watered*) y SS - estrés por sequía (*drought-stress*). Cada una de las líneas se genotipificó para 1147 marcadores de la tecnología Diversity Array (DArT). En cada marcador, dos posibles estados y se codificaron como 0/1. El conjunto de datos fue puesto a disposición del público por Crossa et al., 2010. Mayor detalle sobre esta base de datos, puede encontrarse en la sección Materiales y Métodos del Capítulo 2.

3.3.4. Análisis Multivariado: Análisis de Componentes Principales

El ACP es un método estadístico ampliamente utilizado para reducir la dimensionalidad de un conjunto de datos, preservando al máximo la variabilidad original (M. Greenacre et al., 2022).Los CP son combinaciones lineales de las variables originales ponderadas según la importancia de cada variable para explicar la mayor parte de la varianza total de los datos. Sean los datos organizados en una matriz X de dimensión $I \times J$, con I observaciones (filas) y J variables (columnas), donde cada variable X_j puede considerarse un vector en el espacio de I dimensiones. Las CP se obtienen a partir de la descomposición espectral de la matriz de varianzas y covarianzas (o de correlaciones) de $X_1, \ldots, X_J$, o equivalentemente mediante la descomposición en valores singulares de la matriz de datos centrada. Estas componentes definen nuevas variables ortogonales que maximizan la varianza explicada y preservan la estructura geométrica de las distancias euclidianas entre observaciones en el espacio original. Cabe destacar que el desarrollo del ACP no requiere asumir normalidad univariada ni multivariada en la distribución de las variables, sin embargo, dichos supuestos resultan útiles cuando se desea realizar inferencia estadística sobre las componentes principales (I. Jolliffe, 2002).

El procedimiento clásico de ACP se basa en la descomposición espectral de la matriz

de varianzas y covarianzas Σ, de dimensión $J \times J$:

$$\Sigma = V D_\lambda V' \tag{3.1}$$

donde V es la matriz de autovectores ortonormales (columnas) y D_λ es la matriz diagonal de autovalores no negativos ordenados de forma decreciente. En la práctica, cuando se trabaja con datos muestrales, Σ es reemplazada por S, la matriz muestral de varianzas y covarianzas o por la matriz de correlaciones que es la matriz de varianzas y covarianzas estandarizadas (R), donde S es definida como:

$$S = \left(\frac{n}{n-1}\right) S_n = \frac{1}{n-1} \sum_{j=1}^{n} \left(X_j - \overline{X}\right)\left(X_j - \overline{X}\right)' \tag{3.2}$$

y R como:

$$R = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1J}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{JJ}}} \\ \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \frac{\sigma_{22}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{2J}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{JJ}}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\sigma_{1J}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{JJ}}} & \frac{\sigma_{2J}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{JJ}}} & \cdots & \frac{\sigma_{JJ}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{JJ}}} \end{bmatrix} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1J} \\ \rho_{12} & 1 & \cdots & \rho_{2J} \\ \cdots & \cdots & \cdots & \cdots \\ \rho_{1J} & \rho_{2J} & \cdots & 1 \end{bmatrix} \tag{3.3}$$

donde σ_{ij} es la varianza y ρ_{ij} es la correlación entre el i-ésimo y el j-ésimo unidad.

Sea el vector aleatorio $X = [X_1, X_2, ..., X_J]$ con matriz de varianza-covarianza Σ, los autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_J \geq 0$, y los autovectores $a_1, ..., a_J$, se define cada CP como:

$$\begin{aligned} CP_1 &= a_1' X = a_{11} X_1 + a_{12} X_2 + \cdots + a_{1J} X_J \\ CP_2 &= a_2' X = a_{21} X_1 + a_{22} X_2 + \cdots + a_{2J} X_J \\ &\vdots \\ CP_J &= a_J' X = a_{J1} X_1 + a_{J2} X_2 + \cdots + a_{JJ} X_J \end{aligned} \tag{3.4}$$

sujetas a las condiciones:

$$\begin{aligned} \text{Var}(CP_j) &= a_j' \Sigma a_j \\ \text{Cov}(CP_j, CP_k) &= a_j' \Sigma a_k \quad \text{para } j \neq k \end{aligned} \tag{3.5}$$

Es decir, las CP son combinaciones lineales no correlacionadas entre sí que maximi-

zan secuencialmente la varianza. La primera CP maximiza la varianza total explicada bajo la restricción $a_1' a_1 = 1$. Las siguientes se obtienen de forma recursiva imponiendo ortogonalidad respecto a las anteriores:

$$\begin{aligned} &\text{máx}\,\text{Var}(a_1' X) \quad \text{sujeto a } a_1' a_1 = 1 \\ &\text{Cov}(a_1' X, a_2' X) = 0 \end{aligned} \tag{3.6}$$

Este proceso se repite hasta obtener las J CP.

Si la matriz X es previamente estandarizada (centrada y escalada), se utiliza la matriz de correlaciones en lugar de la matriz de varianzas-covarianzas. Es importante notar que las CP obtenidas mediante estas dos matrices pueden diferir sustancialmente, es decir, producir ordenamientos de las observaciones diferentes, como también modificar el peso o autovector de cada variable. La elección entre utilizar la matriz de varianza-covarianza o la matriz de correlación, dependerá del contexto del análisis. Sin embargo, se recomienda estandarizar previamente los datos en un contexto de variables originales de la matriz X inconmensurables entre ellas. También, puede ser que la unidad de medida de las variables que conforman la matriz X sea la misma, pero la varianza entre ellas muy diferentes, en ese caso, también se recomiendo estandarizar y trabajar con la matriz de correlación en lugar de la matriz de varianza-covarianza.

El ACP constituye una herramienta útil para generar nuevas variables sintéticas que resumen la información contenida en las variables originales. Estas nuevas variables no sólo permiten reducir la dimensionalidad del problema, sino también explorar estructuras internas en los datos, siendo especialmente valiosas cuando existe correlación entre variables, como ocurre frecuentemente en estudios genómicos y fenotípicos (I. Jolliffe, 2002).

3.3.5. Análisis de regresión entre dos conjuntos de datos multivariados: PLS

La regresión por PLS es una técnica estadística que permite modelar la relación entre un conjunto de variables predictoras (X) y una o más variables respuesta (Y), maximizando la covarianza entre ambos conjuntos. Está relacionada con el ACP, pero

a diferencia de este, PLS es un método supervisado, ya que incorpora la matriz de respuesta en la construcción de los componentes. El método resulta especialmente útil en contextos donde el número de variables predictoras es alto, existe multicolinealidad entre ellas, o cuando el número de observaciones es menor que el número de predictores ($n < p$), casos en los que los métodos de regresión clásicos, como la regresión lineal múltiple, no pueden ser aplicados directamente debido a la singularidad de la matriz de covarianzas. PLS se basa en la extracción de un conjunto reducido de componentes latentes que capturan información relevante tanto de las variables independientes como de las dependientes.

La ecuación de PLS puede denotarse como:

$$Y = XB + \epsilon \tag{3.7}$$

donde Y es la matriz de variables independientes, X es la matriz de variables explicativas, B la matriz de coeficientes de regresión y ϵ es la matriz residual.

En el caso multivariado, donde se desea predecir q variables respuesta continuas $Y_1, \ldots, Y_q$ a partir de p variables predictoras $X_1, \ldots, X_p$, donde $n < p$ y la matriz de covarianza $p \times p$ es singular, PLS se basa en la descomposición en variables latentes:

$$\begin{aligned} X &= TP' + \epsilon_1 \\ Y &= TQ' + \epsilon_2 \end{aligned} \tag{3.8}$$

donde T es una matriz $n \times c$ dada por las componentes latentes de las n observaciones, P ($p \times c$) y Q ($q \times c$) son matrices de coeficientes, ϵ_1 ($n \times p$) y ϵ_2 ($n \times q$) son matrices de errores aleatorios.

Al igual que en ACP, se construyen componentes latentes T como combinaciones lineales de X:

$$T = XW \tag{3.9}$$

donde W es la matriz de pesos ($p \times c$) para cada variable de la matriz X.

Las variables latentes son usadas luego para predecir en lugar de las variables originales.

Una vez obtenido la componente latente T, se estima Q mediante mínimos cuadrados:

$$Q' = (T'T)^{-1}T'Y \tag{3.10}$$

Finalmente, se obtiene la matriz de coeficientes de regresión B estimados:

$$\hat{B} = WQ' = W(T'T)^{-1}T'Y \tag{3.11}$$

y la matriz de respuesta estimada:

$$\hat{Y} = X\hat{B} = TQ' \tag{3.12}$$

Un problema potencial es que los algoritmos algebraicos que usa PLS para maximizar la correlación entre X e Y pueden tener diferentes resultados dependiendo del punto de partida. El algoritmo produce una secuencia de valores decrecientes de la suma de cuadrados residuales y luego converge al menos al mínimo local (Vicente-Gonzalez y Vicente-Villardon, 2022). La formulación de PLS se basa en la construcción de variables latentes, cuya definición y cantidad óptima serán abordados en las secciones siguientes, ya que ambos aspectos resultan centrales para explicar la relación entre X y Y.

Elección de la cantidad de variables latentes

La determinación del número adecuado de variables latentes en un modelo PLS constituye una etapa fundamental del proceso de modelado, dado que un número excesivo de componentes puede conducir a una sobreparametrización del modelo, generando estimaciones inestables y relaciones de escasa relevancia interpretativa. Por el contrario, un número insuficiente de componentes puede implicar una pérdida de información relevante, afectando la capacidad predictiva del modelo. Desde un punto de vista teórico, el número máximo de variables latentes que puede estimarse está determinado por el rango de las matrices X e Y, y corresponde al mínimo entre ambos. Este límite estructural se debe a que cada componente latente captura la máxima

covarianza posible entre los espacios de X e Y; por lo tanto, una vez agotada dicha covarianza, no es posible estimar componentes adicionales con significado estadístico (Abdi, 2010; Rosipal y Krämer, 2006).

En este contexto, resulta relevante distinguir entre el concepto de variable latente y el de variable sintética. Las variables latentes en PLS se definen como combinaciones lineales de las variables originales que maximizan la covarianza entre los conjuntos explicativos y de respuesta, permitiendo modelar la estructura común subyacente entre ambos. En cambio, las variables sintéticas suelen derivarse de métodos de reducción de la dimensionalidad en los que el criterio de construcción se basa únicamente en la varianza interna de las variables explicativas, sin considerar explícitamente su relación con las variables respuesta (Hardin y Marcoulides, 2011; Yuan et al., 2020).

En esta tesis se utilizó, como umbral superior, el valor mínimo entre los rangos de las matrices X e Y, lo cual impone una restricción natural al número máximo de componentes posibles, alineada con las limitaciones algebraicas del modelo y el principio de parsimonia. Además se aplicaron dos enfoques complementarios para dicha elección (Nengsih et al., 2019). En primer lugar, se utilizó un test de permutaciones para evaluar la significancia estadística de cada componente latente (Dobriban, 2017). Este método consiste en comparar la covarianza observada entre las variables proyectadas y la respuesta con una distribución nula generada a partir de permutaciones aleatorias de las observaciones en la matriz Y. Así, es posible identificar qué componentes presentan una asociación significativa con la respuesta, descartando aquellas cuya relación pueda atribuirse al azar. En segundo lugar, se analizó la covarianza acumulada explicada por los componentes extraídos (Erny et al., 2021). Este criterio busca retener únicamente aquellas variables latentes que contribuyen sustancialmente a explicar la estructura conjunta X e Y, analizando el porcentaje de ganancia en la explicabilidad de cada variable latente sobre la covarianza total entre X e Y. Sin embargo, los resultados obtenidos no presentaron una estructura estrictamente jerárquica en términos de significancia estadística. Este comportamiento es esperable en PLS, dado que, a diferencia de ACP, los componentes no están ordenados por una medida de variabilidad explicada. Por lo tanto, cada componente evalúa una porción distinta y condicional de la relación entre ambos conjuntos de variables, lo que puede dar lugar a que un componente

intermedio capture variabilidad mayormente ruidosa mientras que un componente posterior recupere una estructura sistemática real. En consecuencia, la determinación del número óptimo de componentes no puede basarse únicamente en la significancia individual, sino en la combinación de estos criterios y en el desempeño predictivo global del modelo.

Algoritmos de descomposición Algebraica

Para ajustar los modelos PLS, en este trabajo de tesis, se emplearon dos algoritmos algebraicos para la extracción de componentes, uno es el procedimiento NIPALS implementado en la librería *mixOmics* del software R (R core, 2026) y la descomposición por valor singular SVD implementada en la librería *data4PCCAR* del software estadístico R (R core, 2026). Ambos algoritmos comparten el objetivo de encontrar variables latentes que maximicen la covarianza entre la matriz de marcadores X y la matriz de observaciones fenotípicas Y, pero difieren en su estrategia computacional y en su comportamiento frente a datos faltantes. Mientras que el algoritmo SVD es una descomposición determinista basada en álgebra lineal, que permite obtener directamente todos los componentes latentes de manera simultánea, NIPALS, es un algoritmo iterativo que estima los componentes de forma secuencial, especialmente útil cuando se trabaja con grandes volúmenes de datos o con valores faltantes. A continuación se describe el procedimiento algebraico usado en este trabajo de tesis.

1. Descomposición del valor singular

   El algoritmo SVD es una técnica fundamental del álgebra lineal que permite descomponer cualquier matriz X de dimensión $n \times p$, donde n representa la cantidad de filas de la matriz o las observaciones y p la cantidad de columnas de la matriz o las variables, en el producto de tres matrices:

   $$X = U\Sigma V' \tag{3.13}$$

   donde:

   - U es una matriz ortonormal de dimensión $n \times r$ (autovectores de XX'),

- Σ es una matriz diagonal de dimensión $r \times r$ con los valores singulares en orden decreciente,
- V es una matriz ortonormal de dimensión $p \times r$ (autovectores de $X'X$),
- r es el rango de X.

2. Mínimos cuadrados parciales iterativos no lineares

   El algoritmo NIPALS fue propuesto originalmente por H. Wold (1966) en el contexto de modelos econométricos, y luego adaptado para el cálculo de CP y modelos PLS (Christoffersson, 1970). Este método permite estimar los componentes de manera secuencial, sin necesidad de descomponer la matriz completa, lo que resulta útil en presencia de datos faltantes o cuando el número de observaciones y variables es elevado. Sin embargo, su naturaleza iterativa puede generar soluciones menos estables cuando el número de predictores es muy elevado o cuando los datos presentan alto nivel de colinealidad. El objetivo de este algoritmo es aproximar la matriz X mediante el modelo:

$$X = TP' \tag{3.14}$$

   donde,

   a) las columnas de T se denominan scores,

   b) las columnas de P se denominan cargas o loadings.

   El procedimiento iterativo para calcular cada componente h es:

   a) Inicializar $h = 1$ y definir $X_h = X$.

   b) Seleccionar una columna cualquiera de X_h como vector inicial t_h.

   c) Calcular la carga: $p_h = \frac{X_h' t_h}{t_h' t_h}$.

   d) Normalizar la carga: $p_h = \frac{p_h}{\sqrt{p_h' p_h}}$.

   e) Calcular el nuevo score: $t_h^* = \frac{X_h p_h}{p_h' p_h}$.

   f) Repetir los pasos 3–5 hasta la convergencia del par (t_h, p_h).

Luego, se actualiza la matriz residual:

$$X_{h+1} = X_h - t_h p_h'$$

y se calcula el autovalor correspondiente: $\lambda_h = t_h' t_h$.

Este proceso se repite para $h = 1, ..., H$, donde H es el número de componentes deseado. Las columnas t_h y p_h se ensamblan en las matrices T y P, respectivamente.

3.3.6. Comparación de los algoritmos SVD y NIPALS para la estimación de componentes latentes

La SVD constituye una solución algebraica exacta basada en la factorización matricial simultánea de las matrices de predictores y respuestas, lo cual garantiza convergencia en un número finito de pasos, pero requiere la ausencia de valores faltantes y un alto costo computacional (medido en tiempo que demora la ejecución del modelo) cuando el número de variables supera al de observaciones.

El algoritmo NIPALS, en cambio, estima las componentes de manera secuencial mediante iteraciones alternadas entre predictores y respuestas. Este procedimiento reduce el almacenamiento necesario, acelera la convergencia cuando la dimensionalidad es elevada, y permite el manejo directo de datos faltantes al omitir celdas con valores ausentes en los productos escalares.

Ejemplo ilustrativo de PLS mediante SVD y NIPALS

Consideremos matrices centradas de predictores X y respuesta Y:

$$X = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ -2 & -1 \end{bmatrix}, \qquad Y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

El objetivo del método PLS es construir componentes latentes

$$T = XW$$

de modo que la covarianza entre T y Y sea máxima.

En el caso de una respuesta univariada ($q = 1$), el primer vector de pesos de X es proporcional a $X^\top Y$.

PLS mediante SVD El enfoque basado en SVD utiliza la matriz de covarianza cruzada entre X y Y:

$$\Sigma = X^\top Y$$

Dado que

$$X^\top = \begin{bmatrix} 2 & 0 & -2 \\ 0 & 1 & -1 \end{bmatrix}, \qquad Y = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix},$$

se obtiene

$$\Sigma = \begin{bmatrix} 4 \\ 1 \end{bmatrix}.$$

La descomposición en valores singulares de Σ es

$$\Sigma = UDV^\top$$

y el primer vector singular izquierdo determina el vector de pesos de X. Normalizando,

$$W = \frac{1}{\sqrt{4^2 + 1^2}} \begin{bmatrix} 4 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{17}} \begin{bmatrix} 4 \\ 1 \end{bmatrix}.$$

Las variables latentes se obtienen como

$$T = XW = \frac{1}{\sqrt{17}} \begin{bmatrix} 8 \\ 1 \\ -9 \end{bmatrix}.$$

PLS mediante NIPALS El algoritmo NIPALS obtiene la misma solución mediante un procedimiento iterativo. Se comienza utilizando Y como estimación inicial del score de la respuesta.

Luego, el vector de pesos de X se calcula como

$$W = \frac{X^\top Y}{\|X^\top Y\|} = \frac{1}{\sqrt{17}} \begin{bmatrix} 4 \\ 1 \end{bmatrix},$$

que coincide exactamente con el obtenido mediante SVD.

Las variables latentes se obtienen nuevamente como

$$T = XW = \frac{1}{\sqrt{17}} \begin{bmatrix} 8 \\ 1 \\ -9 \end{bmatrix}.$$

En este caso, dado que la matriz Y es univariada, el algoritmo converge en una sola iteración.

En el contexto de asociación fenotipo-genotipo, la comparación entre ambos algoritmos resulta relevante para evaluar no sólo la coincidencia en las asociaciones fenotipo–genotipo detectadas, sino también la estabilidad numérica, la eficiencia computacional y la sensibilidad frente a valores faltantes. Estas características son especialmente importantes en estudios de asociación genómica, donde las matrices genotípicas suelen contener miles de marcadores y la completitud de los datos rara vez está garantizada. Ambas implementaciones se realizaron en R utilizando las funciones del paquete *mixOmics* (Rohart et al., 2017), manteniendo constante la cantidad de componentes latentes para asegurar comparabilidad entre los resultados.

3.3.7. Representación gráfica

Para ilustrar el ordenamiento de las observaciones en un plano factorial óptimo sin perdida de generalidad, a partir de los algoritmos algebraicos para PLS propuestos anteriormente, se utilizan representaciones gráficas que permiten una interpretación conjunta de las variables, tanto fenotípicas como genotípicas, y de las observaciones. A continuación, se describen las técnicas empleadas.

Gráfico Biplot

El Biplot es una herramienta gráfica desarrollada por Gabriel (1971) para representar simultáneamente observaciones y variables de un conjunto de datos multivariado en un espacio de baja dimensión ($k = 2$ o $k = 3$). En este gráfico, las observaciones se representan como puntos y las variables como vectores. Las dimensiones seleccionadas para el Biplot corresponden a los componentes principales que mejor explican la variabilidad de los datos. En trabajos posteriores, Gabriel (1981) y Gabriel (1982) analizaron las ventajas del Biplot respecto a otros métodos exploratorios, destacando su carácter intuitivo. M. J. Greenacre (1993) estableció las condiciones bajo las cuales un análisis de correspondencias puede interpretarse como un Biplot, y Gabriel (2002) lo propuso como alternativa al análisis de correspondencias múltiples. Dado que los datos suelen presentar alta dimensionalidad, un Biplot no puede reconstruir exactamente la matriz original, pero busca una buena aproximación. El modelo básico de Biplot se expresa como:

$$t_{ij} = x_i' y_j + \varepsilon_{ij} \tag{3.15}$$

donde x_i y y_j son vectores de baja dimensión que representan a la observación i y la variable j, respectivamente, y ε_{ij} es un término de error.

El cálculo de los Biplots se realiza minimizando la suma de los errores cuadráticos $\sum_i \sum_j \varepsilon_{ij}^2$.

Una expresión alternativa, útil para comprender el vínculo con el ACP, es la siguiente fórmula de reconstitución:

$$p_{ij} = r_i c_j \left(1 + \sum_k \sqrt{\lambda_k} \phi_{ik} \gamma_{jk}\right) \tag{3.16}$$

donde:

- p_{ij} son proporciones relativas (n_{ij}/n), con $n = \sum_i \sum_j n_{ij}$,
- r_i y c_j son las masas (pesos) de las filas y columnas,
- λ_k es la inercia (autovalor) del k-ésimo eje,
- ϕ_{ik} y γ_{jk} son las coordenadas estándar de las filas y columnas sobre el eje k.

En esta representación, las direcciones de los vectores de las variables permiten interpretar sus contribuciones a los ejes principales. Los ejes principales derivados de un ACP son denominados componentes principales. Luego, la primera CP asociada al autovalor de mayor magnitud es denominada CP1 y la CP2 es la combinación lineal asociada al segundo autovalor. De esta manera, al graficar las observaciones y las variables sobre el espacio generado por las CP1 y la CP2, el biplot explica la máxima varianza de los datos. En el biplot, la longitud de los vectores asociados a las variables refleja la calidad de su representación en el espacio reducido y su contribución relativa a los ejes principales. El ángulo entre dos vectores permite interpretar la relación entre variables, donde ángulos pequeños indican correlación positiva, ángulos cercanos a 90° ausencia de correlación y ángulos obtusos correlación negativa. Asimismo, la proyección de una observación sobre un vector de variable permite interpretar el valor relativo de dicha variable para esa observación, mientras que la proximidad entre observaciones en el plano indica similitud multivariada en términos de las variables originales (Galindo-Villardón, 1986).

Gráfico Triplot

El Triplot es una extensión del Biplot adaptada al contexto de modelos PLS, en el que se desea representar simultáneamente tres conjuntos de información: variables predictoras (X), variables respuesta (Y) y observaciones. Esta visualización permite evaluar de manera conjunta la estructura latente del modelo, identificar relaciones

entre genotipos y fenotipos, y comparar la influencia relativa de distintas variables. A diferencia de otras representaciones como el análisis de redundancia tradicional (Ter Braak, 1994), el Triplot construye dimensiones latentes tanto para los predictores como para las respuestas. Además, puede incorporar información categórica mediante técnicas híbridas como el Biplot Logístico (Galindo-Villardón y Blázquez-Zaballos, s.f.) o extensiones multivariadas del análisis de redundancia para respuestas binarias.

Polígono de contención

La vista tipo polígono de contención, también conocida como *which-won-where* (Verma et al., 2015), se emplea sobre representaciones biplot o triplot para identificar relaciones dominantes entre genotipos y ambientes (o variables). En este enfoque, los vértices del polígono están determinados por las observaciones más alejadas del centro (en nuestro caso, los marcadores moleculares), y se construyen líneas perpendiculares que dividen el espacio gráfico en sectores. Cada sector representa un subconjunto del espacio latente asociado a un ambiente hipotético. Si una variable fenotípica se encuentra dentro de un sector delimitado por ciertos marcadores, se interpreta que estos últimos tienen mayor asociación con dicha variable. En esta tesis, esta técnica se aplica sobre los Triplots que grafican variedades, caracteres fenotípicos y SNP. El objetivo es identificar visualmente qué marcadores están más estrechamente relacionados con determinados fenotipos. Para ello, se analiza en qué sector cae cada variable fenotípica y qué marcadores se ubican en los vértices correspondientes, lo que facilita la interpretación de las asociaciones.

3.3.8. Métricas de validación de los métodos PLS

Simulación por Monte Carlo

Las simulaciones de Monte Carlo constituyen una herramienta fundamental en estadística computacional para la evaluación de modelos en contextos donde intervienen componentes de aleatoriedad y estructuras complejas difíciles de abordar mediante métodos analíticos clásicos. El método fue desarrollado originalmente en la década

de 1940 por von Neumann y Stanislaw Ulam en el marco del Proyecto Manhattan, un programa de investigación militar durante la Segunda Guerra Mundial, donde surgió la necesidad de abordar problemas físicos de alta complejidad que no admitían soluciones analíticas directas, y recibió su nombre en alusión al casino de Monte Carlo, por su vinculación con procesos aleatorios similares al juego de la ruleta (Metropolis y Ulam, 1949). Las simulaciones de Monte Carlo consisten en repetir muchas veces, de forma artificial y controlada, un proceso aleatorio para observar cómo varía un estimador y así aproximar su distribución. Estas simulaciones resultan útiles para evaluar la robustez de modelos, estimar errores estándar, construir intervalos de confianza y determinar la significancia de parámetros. En esta tesis, se utilizó un enfoque de simulación de Monte Carlo para analizar el comportamiento de los coeficientes de regresión obtenidos por PLS, bajo distintos niveles de correlación entre variables y distintos tamaños muestrales. Cada iteración consistió en la generación de matrices X y Y con estructuras de correlación predefinidas, a partir de las cuales se ajustaron modelos PLS utilizando los algoritmos SVD y NIPALS. Posteriormente, se evaluó la estabilidad y la precisión de los coeficientes obtenidos, así como la capacidad predictiva del modelo sobre conjuntos de prueba independientes.

Para evaluar la significancia estadística de los coeficientes obtenidos mediante PLS, se estimaron valores p utilizando un enfoque de simulación por permutación. Bajo la hipótesis nula de ausencia de asociación entre las variables predictoras y la respuesta, se generó una distribución empírica del estadístico de interés mediante la permutación aleatoria de las observaciones de la matriz Y, manteniendo fija la matriz X. Para cada permutación se ajustó nuevamente el modelo PLS y se calculó el estadístico correspondiente. La repetición de este procedimiento un gran número de veces permitió aproximar la distribución nula del estadístico.

El valor p se estimó como la proporción de estadísticos simulados que resultaron iguales o más extremos que el valor observado en los datos originales. En términos formales,

$$p = \frac{1}{B} \sum_{b=1}^{B} I(T_b \geq T_{obs}),$$

donde T_{obs} es el estadístico calculado con los datos originales, T_b corresponde al

estadístico obtenido en la b-ésima permutación, B es el número total de permutaciones realizadas e $I(\cdot)$ es la función indicadora.

Evaluación del modelo de regresión por PLS mediante validación cruzada

Con el fin de comparar el ajuste de la técnica PLS bajo ambos algoritmos: SVD y NIPALS, se utilizó la validación cruzada k-fold (Refaeilzadeh et al., 2009; Stone, 1974). En la validación cruzada k-fold, los datos se dividen primero en k segmentos y se realizan k iteraciones de entrenamiento y validación, de modo que en cada iteración se retiene un segmento diferente de los datos para la validación, mientras que los restantes se utilizan para el aprendizaje del modelo. Para esta tesis, según los datos utilizados para la validación, se evalúa si es necesario realizar un sobremuestreo, dependiendo del tamaño de la misma. En la minería de datos y el aprendizaje automático, la validación cruzada de 10 veces (k=10) es la más común. En cada iteración, la base de datos completa se dividió aleatoriamente en dos subconjuntos: uno denominado de entrenamiento que contenía el 85% de la base de datos original y otro subconjunto utilizado para la validación del modelo, con el 15%. El subconjunto de validación no participó en el ajuste del modelo, por ello se denomina validación cruzada y pretende evaluar el modelo con un nuevo conjunto de datos. Para evaluar la capacidad de PLS de predecir, se calculó la raíz del error cuadrático medio (RECM), como:

$$\text{RECM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.17}$$

donde, y_i es el valor observado, $\hat{y}_i$ es el valor predicho por el modelo de regresión por PLS y n es el número de observaciones. De esta manera, mientras más grande es el valor de RECM, indica que más se alejaron los valores predichos del valor observado y menor precisión tiene el modelo.

Medidas diagnóstico para evaluar la capacidad predictiva de los modelos de regresión

En el caso de la base de datos proveniente de la simulación, se cuenta con las información acerca de las posiciones de los QTL en el genoma simulado. Por lo tanto es

posible evaluar si los marcadores moleculares que resultaron estadísticamente significativos según el modelo PLS se encuentran cercanos a algún QTL (valor verdadero) y por lo tanto se espera que el modelo PLS los identifique. Debido a que se conoce el valor verdadero y se obtiene el valor estimado por el modelo, luego pueden armarse tablas de clasificación, también denominadas tablas de doble entrada, donde se cuenta la cantidad de marcadores moleculares identificados como estadísticamente significativos y la cantidad de marcadores moleculares cercanos (<5cM) a los QTL. A partir de la tabla de clasificación se derivan distintas métricas diagnósticas que permiten evaluar el desempeño predictivo del modelo PLS en la identificación de marcadores moleculares cercanos a los QTL simulados. En particular, se consideran la sensibilidad, la especificidad, la precisión, la exactitud y el F1-score. La sensibilidad (o tasa de verdaderos positivos) mide la capacidad del modelo para identificar correctamente los marcadores moleculares cercanos a un QTL, y se define como la proporción de verdaderos positivos respecto del total de marcadores verdaderamente asociados. Valores elevados de sensibilidad indican una mayor capacidad del modelo para detectar señales verdaderas, reduciendo la omisión de asociaciones relevantes. La especificidad (o tasa de verdaderos negativos) cuantifica la capacidad del modelo para identificar correctamente los marcadores no asociados a QTL, y se define como la proporción de verdaderos negativos respecto del total de marcadores no asociados. Una alta especificidad refleja una baja tasa de falsos positivos, lo cual resulta especialmente importante en estudios genómicos con alta dimensionalidad. La precisión (o valor predictivo positivo) se define como la proporción de verdaderos positivos entre todos los marcadores identificados como significativos por el modelo. Esta métrica evalúa cuán confiables son las asociaciones detectadas y toma valores altos cuando la mayoría de los marcadores seleccionados corresponden efectivamente a regiones cercanas a QTL. La exactitud mide la proporción total de clasificaciones correctas, considerando tanto verdaderos positivos como verdaderos negativos, respecto del total de marcadores analizados. Si bien proporciona una medida global del desempeño del modelo, su interpretación debe realizarse con cautela en contextos donde existe un fuerte desbalance entre clases, como suele ocurrir en estudios de asociación genética. Finalmente, el F1-score combina sensibilidad y precisión en una única medida, definida como su media armónica, y resulta particularmente útil

cuando se busca un equilibrio entre la detección de asociaciones verdaderas y el control de falsos positivos. Valores elevados de F1-score indican un buen compromiso entre ambas métricas. En general, para todas las métricas consideradas, valores más altos indican un mejor desempeño del modelo; sin embargo, la interpretación conjunta de las mismas permite una evaluación más completa y robusta de la capacidad predictiva del método PLS en escenarios simulados. Las ecuaciones de cada una de las métricas utilizadas son:

$$\begin{aligned}
\text{Precisión} &= \frac{VP}{VP+FP} \\
\text{Sensibilidad} &= \frac{VP}{VP+FN} \\
\text{F1-score} &= 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}} \\
\text{Especificidad} &= \frac{VN}{VN+FP} \\
\text{Exactitud} &= \frac{VP+VN}{VP+FP+FN+VN}
\end{aligned}$$

donde, VP corresponde a los marcadores moleculares verdaderos positivos, es decir, los que fueron estadísticamente significativos y verdaderamente corresponden con un QTL; FP a los falsos positivos son los marcadores moleculares que fueron estadísticamente significativos pero no estaban asociados a un QTL simulado; FN a los falsos negativos que representan los marcadores moleculares que están verdaderamente asociados a un QTL pero el modelo no los detectó como estadísticamente significativos y VN a los verdaderos negativos son los marcadores moleculares que no están asociados a un QTL y no resultaron estadísticamente significativos por el modelo.

3.4. Resultados

3.4.1. Resultados obtenidos sobre la aplicación de PLS en una base de datos simulada

El número de variables latentes considerados para la regresión PLS fue determinado considerando como límite superior el mínimo número cardinal entre los rangos de X e Y. En este caso, la dimensión de la matriz X es 240 × 79.905 y dado que por definición el $Rango(X) = Rf(X) = Rc(X)$, donde Rf es Rango fila de X y Rc es rango columna de X, el $Rango(X) = 240$. En el caso de la matriz Y, su dimensión es 240 × 2, por lo tanto $Rango(Y) = 2$. Luego, con esta base de datos se evaluaron dos variables latentes. De acuerdo con los resultados del test de permutaciones que se muestran en la Tabla 3.1, la técnica SVD identificó como estadísticamente significativas ambas variables latentes, mientras que en NIPALS solo una de ellas resultó estadísticamente significativa. No obstante, según el criterio de covarianza entre las variables regresoras (X) y las variables respuestas (Y) acumulada una única variable latente explica el 55 % de la covarianza en SVD y el 56 % en NIPALS (Tabla 3.1). Por lo tanto se decidió trabajar con ambas componentes bajo los dos algoritmos.

Tabla 3.1: Valores de probabilidad asociada al test de permutaciones para cada variable latente obtenida como resultado de cada uno de los algoritmos (SVD y NIPALS) de PLS, porcentaje de covarianza explicada y porcentaje de covarianza acumulada en cada variable latente para cada algoritmo (SVD y NIPALS)

Cantidad de variables latentes	Valores p		Covarianza Explicada (%)		Covarianza Acumulada (%)	
	SVD	NIPALS	SVD	NIPALS	SVD	NIPALS
1	<0,0001	0,0400	55,05	56,28	55,05	56,28
2	0,0100	0,9800	44,95	43,72	100,00	100,00

*valores p obtenidos del test de permutaciones

La cantidad de marcadores moleculares estadísticamente significativos detectados por ambos algoritmos en los dos ambientes del carácter fenotípico simulado, fue de 738 marcadores moleculares en el primer ambiente y 774 en el segundo. En ambos casos, es decir, tanto con SVD como con NIPALS, coincidieron los marcadores moleculares

identificados, lo que indicó un desempeño similar entre los dos algoritmos de PLS. No obstante, el algoritmo SVD presentó una desventaja computacional considerable, requiriendo 37,82 minutos frente a solo 1,57 segundos en NIPALS. Los resultados de la validación cruzada de tipo k-fold con k=10, mostraron que NIPALS alcanzó menores valores de RECM en ambos caracteres fenotípicos, aunque este error fue mayor en el Ambiente 2 para ambos algoritmos (Figura 3.1).

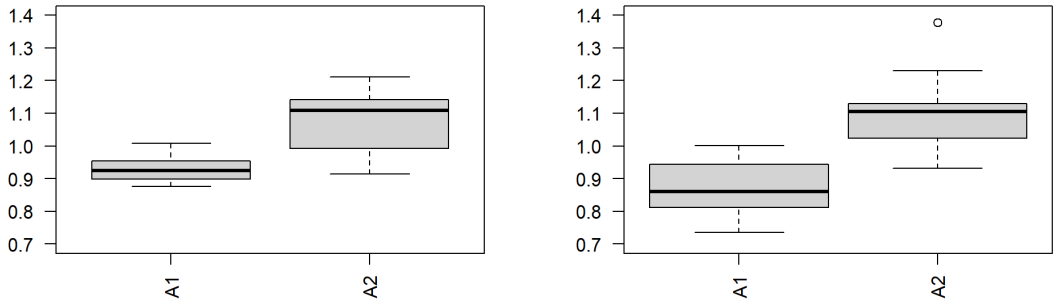

Figura 3.1: Gráficos de caja de los coeficientes de variación de la Raíz del Error Cuadrático Medio (RECM) en ambos ambientes (A_1 y A_2) obtenidos para la descomposición SVD (izquierda) y para NIPALS (derecha) en la regresión por mínimos parciales (PLS)

Los triplots correspondientes a cada algoritmo algebraico utilizado, es decir, SVD y NIPALS mostraron todos los marcadores de la matriz Y, en color (distinto al negro) se destacaron los marcadores moleculares que fueron estadísticamente significativos (Figura 3.2). Para este caso no se consideró realizar el polígono de contención debido al alto número de marcadores moleculares que no permiten definir los extremos. Del total de 79.917 marcadores moleculares que participaron en el análisis, solamente el 0,92 % resultaron estadísticamente significativos. Es importante destacar que se identificaron zonas de marcadores moleculares estadísticamente significativos bien delimitadas, a la izquierda los marcadores significativos asociados al Ambiente 1, mientras que a la derecha se encontraron los marcadores significativos asociados al Ambiente 2. Este mismo patrón se observa con ambos algoritmos algebraicos (Figura 3.2), confirmando que en el contexto de datos completos la conformación de los nuevos ejes son iguales con ambos algoritmos.

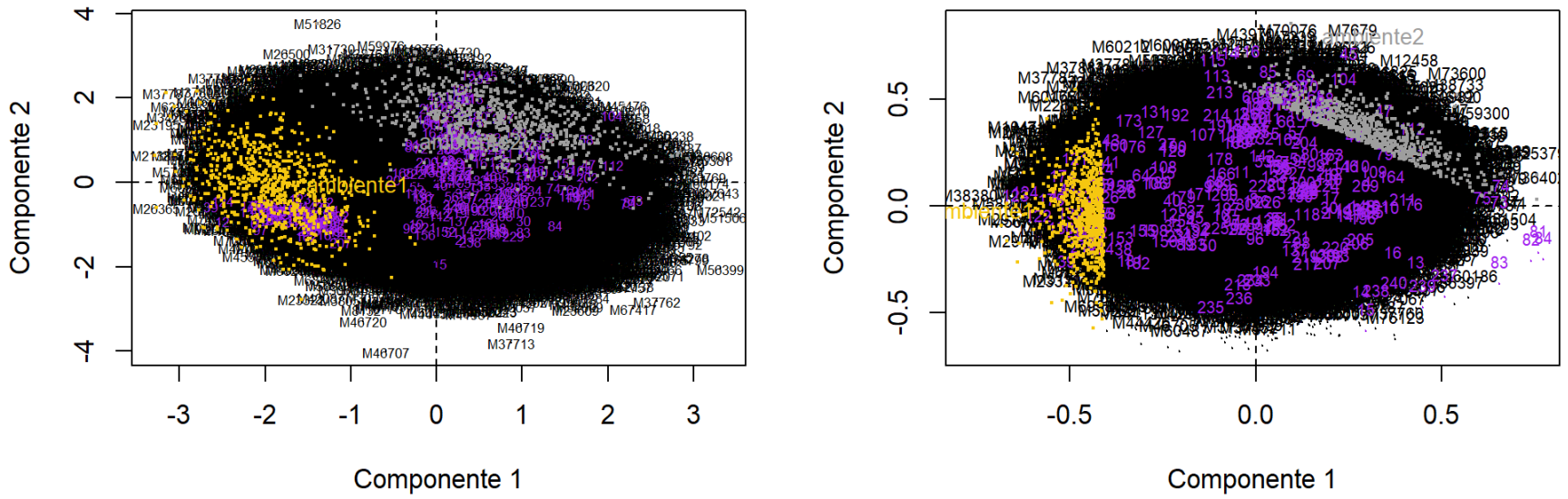

Figura 3.2: Gráficos Triplots obtenidos a partir del PLS sobre la base de datos simulada con ambos algoritmos algebraicos SVD (izquierda) y NIPALS (derecha). Se destacan los SNP significativos para ambos ambientes: en amarillo el Ambiente 1 (A_1) y en gris el Ambiente 2 (A_2)

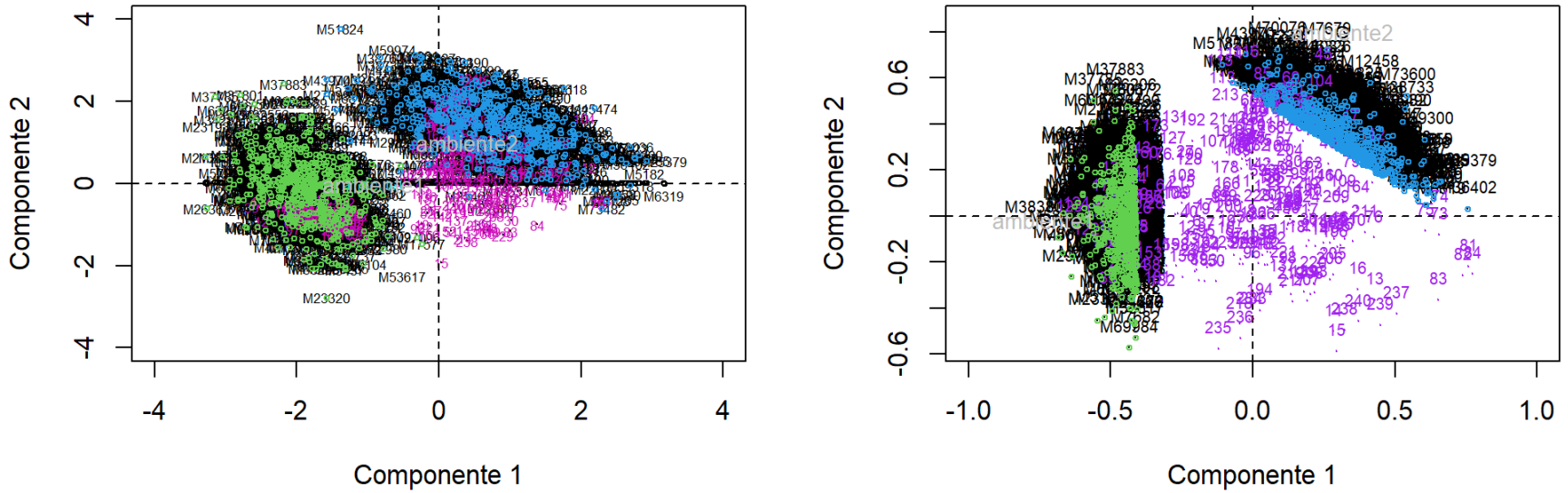

Figura 3.3: Gráfico Triplot obtenido por PLS a través de dos algorimos algebraicos aplicados sobre la base de datos simulada. Se representan los SNP estadísticamente significativos y se destacan los marcadores cercanos a algún QTL (<5cM) en verde (A1) y en azul (A2). Para la descomposición SVD (izquierda) y para NIPALS (derecha)

Al graficar el triplot únicamente con los marcadores moleculares estadísticamente significativos se destacaron visualmente aquellos marcadores moleculares que resultan cercanos a un QTL (Figura 3.3). De los marcadores encontrados como estadísticamente significativamente asociados al caracter fenotípico medido en alguno de los ambientes, los marcadores moleculares cuya ubicación es central en el plano factorial conformado por las dos primeras componentes son los que se destacaron por encontrarse cercano a un QTL, a una distancia menor a 5cM.

El conocimiento de la ubicación de los QTL, debido a que los mismos fueron identificados en el proceso de simulación, permitió evaluar el desempeño de ambos algoritmos algebraicos comparando la proximidad de los SNP detectados respecto de los loci causales de asociación fenotipo-genotipo. A partir de éste conocimiento, se construyeron tablas de doble entrada que relacionan los SNP identificados como estadísticamente significativos y aquellos identificados como estadísticamente no significativos según su cercanía a algún QTL. De esta manera, los SNP cercanos a algún QTL se consideran verdaderas asociaciones entre el fenotipo y el genotipo, mientras que aquellos QTL que no están a una distancia menor a 5cM de ningún QTL se consideran falsos positivos, también denominados Error de tipo I. Dado que ambos algoritmos algebraicos (SVD y NIPALS) produjeron exactamente el mismo conjunto de SNP estadísticamente significativos, los resultados se reportan para uno de los algoritmos, en este caso NIPALS.

Del total de marcadores moleculares que participaron en este análisis (79.917), el 0,92 % resultaron estadísticamente significativos (valor $p \leq 0,05$), pero considerando la cercanía a un QTL simulado, se identificó una tasa de verdaderos positivos, es decir, SNP que se encontraban asociados al caracter fenotípico y fueron detectados estadísticamente significativos, de 49,6 % para el Ambiente 1 y de 46,1 % en el Ambiente 2. Si bien, podría haberse esperado un mayor porcentaje de verdaderos positivos, es importante considerar que la distancia establecida en este trabajo de 5cM puede resultar en este conjunto de datos una distancia física exigente al momento de identificar los verdaderos positivos y la tasa de Error de tipo I, es decir, haber rechazado la hipótesis nula diciendo que hay asociación cuando en realidad no la hay, también considerados falsos positivos. Considerando la misma distancia física de los QTL simulados a los marcadores

moleculares de 5cM que no fueron declarados como estadísticamente significativos, el 54 % resultó en verdaderos negativos por encontrarse a una distancia superior a 5cM (Tabla 3.2).

Tabla 3.2: Matriz de confusión entre los SNP estadísticamente significativos detectados por el modelo PLS sobre el carácter fenotípico simulado en los Ambientes 1 y 2 y los marcadores moleculares ubicados a una distancia menor a 5cM de algún QTL, donde los Errores I y II representan los Errores de tipo I y II respectivamente del modelo

	Ambiente 1			Ambiente 2		
	SNP<5cM de un QTL	SNP>5cM de un QTL	Total	SNP<5cM de un QTL	SNP>5cM de un QTL	Total
SNP sign	366 (49,59 %) (VP)	372 (50,40 %) (FP, Error I)	738 (100 %)	357 (46,12 %) (VP)	417 (53,88 %) (FP, Error I)	774 (100 %)
SNP no sign	36.290 (45,83 %) (FN, Error II)	42.889 (54,17 %) (VN)	79.179 (100 %)	36.299 (45,86 %) (FN, Error II)	42.844 (54,14 %) (VN)	79.143 (100 %)
Total	36.656	43.261	79.917	36.656	43.261	79.917

Nota: Tanto el algoritmo algebraico SVD como NIPALS produjeron el mismo conjunto de marcadores moleculares estadísticamente significativos para PLS, por ello se reportan los resultados de SVD que son los mismos que para NIPALS.

Además de la estimación de los verdaderos positivos, falsos positivos (Error de tipo I), falsos negativos (Error de tipo II) y verdaderos negativos, las métricas de precisión, sensibilidad, F1-score, especificidad y exactitud fueron estimadas a partir de esta tabla de clasificación cruzada como se muestra en la Tabla 3.2. Las métricas de evaluación muestran un comportamiento muy similar en ambos ambientes. La especificidad fue elevada, alcanzando valores del 99 % en ambos ambientes. Este resultado coincide con la capacidad de la técnica en identificar los verdaderos negativos, mientras que la sensibilidad, asociada a la capacidad de detectar los verdaderos positivos resultó extremadamente baja, menor al 1 % lo que indica una baja capacidad del modelo para identificar los marcadores moleculares verdaderamente asociados al caracter fenotípico en estudio. Sin embargo, esta baja sensibilidad puede estar explicada por el desbalance que existe entre las diferentes clases que miden estas métricas. Dicho de otra manera, de la alta cantidad de marcadores moleculares que formaron parte de este estudio, la proporción de marcadores estadísticamente signficativos es considerablemente menor,

menos del 1 %, por lo tanto es de esperar que el modelo sea más específico y menos sensible. La precisión, que tiene en cuenta del total de marcadores considerados estadísticamente significativos, fueron verdaderos positivos, coincidió con lo indicado anteriormente, siendo de 49 % y 46 % para los Ambientes 1 y 2, respectivamente. En concordancia, el F1-score fue bajo porque está directamente ligado a la sensibilidad. Dado que la sensibilidad fue muy baja, es de esperar que al F1 sea bajo ya que fue directamente proporcional a la sensibilidad. La exactitud sería una medida de no error, es decir, la capacidad de identificar los verdaderos positivos asociado con la confianza y los verdaderos negativos, asociados con la potencia de la prueba. En este caso, superaron el 50 % en ambos ambientes (Tabla 3.3).

Tabla 3.3: Comparación de métricas de evaluación del modelo PLS para los Ambientes 1 y 2

Métrica	**Ambiente 1**	**Ambiente 2**
Precisión	0,4960	0,4610
Sensibilidad	0,0099	0,0097
F1-score	0,0196	0,0190
Especificidad	0,9914	0,9904
Exactitud	0,5410	0,5400

3.4.2. Resultados obtenidos sobre la aplicación de PLS en una base de datos de duraznero

La base de datos original de dimensión 6.029, fue depurada eliminando los SNP con datos faltantes y aquellos con un MAF menor a 0,05 resultando un total de 619 marcadores moleculares sin valores faltantes. La distribución univariada para cada una de las ocho variables fenotípicas (combinación de las cuatro variables medidas en dos campañas agrícolas) se presentan en la Figura 3.4. Para las variables rendimiento, peso promedio del fruto y cantidad de frutos por planta, se observó una distribución asimétrica hacia a la derecha, en cambio para el caracter calibre ponderado la distribución es simétrica en ambas campañas agrícolas. Los valores del rendimiento y de la cantidad de frutos por planta se incrementaron de la campaña 2010-2011 a la campaña 2011-2012, para la variable peso promedio del fruto los valores disminuyeron y la distribución se presentó con una mayor asimetría hacia a la derecha en la campaña 2011-2012 y

finalmente para el calibre ponderado se presentaron variedades con calibre mayor en la campaña 2010-2011, pero en general la distribución se centró en valores mayores en la campaña 2011-2012.

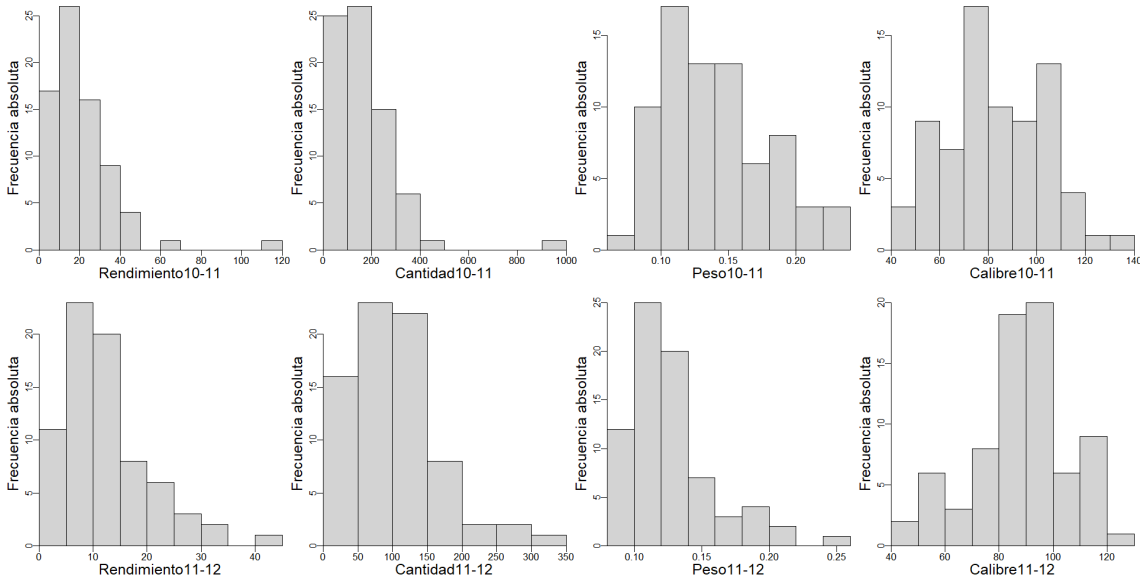

Figura 3.4: Frecuencia absoluta de los valores de Rendimiento obtenidos en la campaña 2010-2011 (izquierda arriba), en la campaña 2011-2012 (izquierda abajo); valores de Cantidad de frutos por planta en la campaña 2010-2011 (centro izquierda arriba), en la campaña 2011-2012 (centro izquierda abajo); valores del peso en la campaña 2010-2011 (centro derecha arriba), en la campaña 2011-2012 (centro derecha abajo); y valores de Calibre promedio en la campaña 2010-2011 (derecha arriba) y en la campaña 2011-2012 (derecha abajo)

El número de variables latentes máxima requeridas fue de ocho variables, estimadas como el número máximo a utilizar (mínimo de los rangos de $X = 74 \times 619$ e $Y = 74 \times 8$). El test de hipótesis realizado a través de permutaciones y generación de la distribución arrojó que para el algoritmos algebraico SVD serían necesarias tres componentes mientras que para NIPALS sólo dos, con un nivel de significación de 0,05. Pero en el caso de la covarianza acumulada, con tres componentes se logró explicar menos del 50 % de la misma en el caso de SVD y aproximadamente 55 % en el caso de NIPALS, por lo tanto en ambas técnicas (SVD y NIPALS) se seleccionaron tres variables latente (Tabla 3.4).

Tabla 3.4: Valores de probabilidad asociada al test de permutaciones para cada variable latente obtenida como resultado de cada uno de los algoritmos (SVD y NIPALS) de PLS, porcentaje de covarianza explicada y porcentaje de covarianza acumulada en cada variable latente para cada algoritmo (SVD y NIPALS)

Cantidad de variables latentes	Valores p*		Covarianza Explicada (%)		Covarianza Acumulada (%)	
	SVD	NIPALS	SVD	NIPALS	SVD	NIPALS
1	0,075	0,005	16,254	23,366	16,254	23,366
2	0,017	0,039	16,788	19,604	33,042	42,970
3	0,024	0,968	15,750	12,706	48,792	55,676
4	0,623	0,934	11,274	12,843	60,066	68,519
5	0,379	1,000	11,583	10,140	71,649	78,659
6	0,604	1,000	9,965	8,755	81,614	87,413
7	0,737	1,000	9,161	6,456	90,775	93,869
8	0,565	1,000	9,225	6,131	100,00	100,00

*valores p obtenidos del test de permutaciones

Se obtuvieron los resultados de la validación cruzada con 10 iteraciones, donde en cada caso se seleccionó el 85 % de las variedades (filas de la base de datos) para entrenar el modelo PLS. A diferencia de la base de datos simulada, donde se contó con una mayor cantidad de variedades, en este caso se debió realizar un sobremuestreo sobre el 15 % de los valores restantes. Luego, se utilizó el RECM entre los valores reales de Y y los valores predichos para ambos casos, SVD y NIPALS (Figura 3.5).

En general, los valores de RECM obtenidos utilizando el algoritmo algebraico NIPALS presentaron menor variabilidad en las distribuciones que los obtenidos utilizando el algoritmo algebraico SVD. Para el carácter rendimiento en la campaña agrícola 2010–2011, las medianas de las distribuciones de RECM fueron similares (1,00 y 1,03 para SVD y NIPALS, respectivamente), y los valores del rango intercuartílico también fueron iguales (0,07 en ambos casos). Sin embargo, el desvío estándar de los valores de RECM obtenidos utilizando el algoritmo NIPALS fue menor (0,08) que el observado utilizando el algoritmo SVD (0,12).

Este mismo carácter, medido en la campaña agrícola 2011–2012, presentó una mediana de los valores de RECM mayor utilizando el algoritmo NIPALS (1,09) que en el

caso de SVD (1,04), pero un valor menor del rango intercuartílico (0,17 y 0,10 para SVD y NIPALS, respectivamente) y también un valor menor del desvío estándar (0,12 y 0,09 para SVD y NIPALS, respectivamente).

En el caso del carácter cantidad de frutos por planta, el comportamiento fue similar al observado para el rendimiento. En ambas campañas, la mediana de la distribución de RECM obtenida utilizando el algoritmo NIPALS (0,99 y 1,06 para las campañas agrícolas 2010–2011 y 2011–2012, respectivamente) presentó valores levemente superiores a los observados en la distribución obtenida utilizando el algoritmo SVD (0,94 y 1,05 para las campañas agrícolas 2010–2011 y 2011–2012, respectivamente). Sin embargo, al analizar la variabilidad de las distribuciones, los valores obtenidos con NIPALS fueron menores tanto para el rango intercuartílico (0,06 y 0,02 para las campañas agrícolas 2010–2011 y 2011–2012, respectivamente) como para el desvío estándar (0,08 y 0,11 para las campañas agrícolas 2010–2011 y 2011–2012, respectivamente), en comparación con los valores obtenidos utilizando SVD (rango intercuartílico de 0,09 y 0,07 y desvío estándar de 0,16 y 0,12 para las campañas agrícolas 2010–2011 y 2011–2012, respectivamente).

Para el carácter peso promedio del fruto medido en la campaña agrícola 2010–2011, la variabilidad de las distribuciones de RECM obtenidas utilizando NIPALS (0,11 y 0,08 para el rango intercuartílico y el desvío estándar, respectivamente) resultó menor que la observada para los valores de RECM obtenidos utilizando SVD (0,20 y 0,13 para el rango intercuartílico y el desvío estándar, respectivamente), mientras que el valor de la mediana fue mayor en la distribución de RECM obtenida utilizando el algoritmo NIPALS (0,86) que en la obtenida utilizando SVD (0,79).

Por el contrario, para este mismo carácter medido en la campaña agrícola 2011–2012, la variabilidad fue mayor en la distribución de RECM obtenida con el algoritmo NIPALS que en la obtenida con el algoritmo SVD, con valores de rango intercuartílico (0,07) y desvío estándar (0,08) inferiores utilizando SVD que los valores de rango intercuartílico (0,15) y desvío estándar (0,12) obtenidos utilizando el algoritmo NIPALS. Sin embargo, en este caso se observó un valor de la media mayor para la distribución de RECM obtenida utilizando el algoritmo NIPALS que para la obtenida utilizando el algoritmo SVD (0,91 y 0,92 para NIPALS y SVD, respectivamente).

Finalmente, el carácter calibre ponderado medido en la campaña agrícola 2010–2011

presentó valores de variabilidad similares y valores de la mediana del RECM mayores utilizando el algoritmo NIPALS en comparación con SVD (0,13, 0,07 y 0,90 para rango intercuartílico, desvío estándar y mediana utilizando NIPALS frente a 0,13, 0,09 y 0,82 utilizando SVD). Mientras que para el mismo carácter medido en la campaña agrícola 2011–2012 se observó un comportamiento similar al del carácter peso promedio del fruto medido en la misma campaña, presentando una mayor variabilidad en la distribución obtenida con NIPALS respecto de SVD, pero una mediana similar (0,15, 0,13 y 0,92 para rango intercuartílico, desvío estándar y mediana utilizando NIPALS frente a 0,11, 0,07 y 0,92 utilizando SVD).

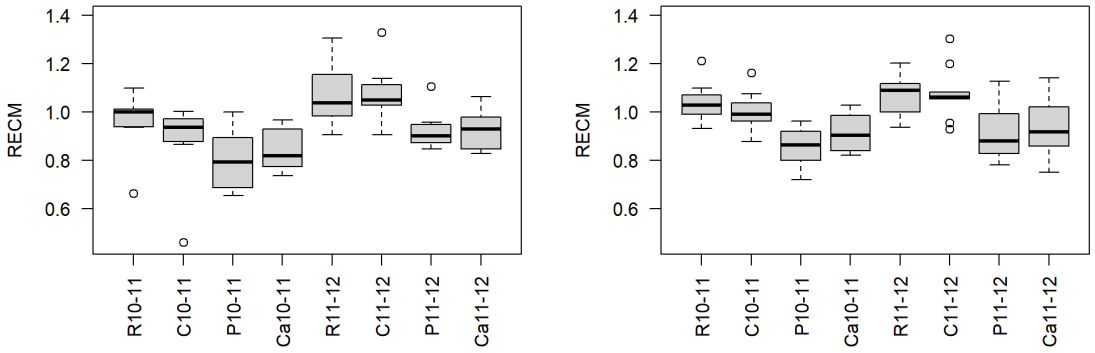

Figura 3.5: Boxplots de los coeficientes de variación de la Raíz del Error Cuadrático Medio (RECM) en las cuatro variables. Para la descomposición SVD (izquierda) y para NIPALS (derecha)

Los gráficos triplot permitieron visualizar los marcadores moleculares estadísticamente significativos identificados por cada algoritmo (SVD – Figura 3.6 y NIPALS – Figura 3.7) en relación con los caracteres fenotípicos evaluados. En los resultados obtenidos mediante el algoritmo algebraico SVD se observó que las variables fenotípicas se ubicaron cercanas al centro del gráfico, presentando valores bajos en las tres componentes, mientras que los marcadores moleculares significativos se localizaron principalmente en los extremos del polígono.

Para el rendimiento medido en la campaña agrícola 2010–2011 (representado en rojo), los marcadores '23', '24' y '26' se ubicaron en posiciones similares, con valores negativos en la componente 1, positivos pero pequeños en la componente 2 y elevados en la componente 3. El marcador '543' se encontró cercano al extremo superior izquierdo del gráfico entre las componentes 1 y 2, con un valor negativo y pequeño en la componente 3. El resto de los marcadores asociados a este carácter se ubicaron en el subpolígono del cuadrante superior derecho (Figura 3.6), presentando valores moderados en las componentes 1 y 2 y altos en la componente 3. Estos marcadores se localizaron en el mismo cuadrante que la variedad 24, la cual presentó el mayor rendimiento en dicha campaña.

El carácter cantidad de frutos por planta, representado en verde, presentó cuatro marcadores coincidentes con los asociados al rendimiento en la misma campaña ('128', '129', '530' y '541'). Estos marcadores también se ubicaron en el cuadrante superior

derecho entre las componentes 1 y 2 (Figura 3.6, izquierda), con valores elevados en la componente 2, mientras que los valores correspondientes a la componente 3 variaron entre marcadores (Figura 3.6, derecha). Este cuadrante coincidió nuevamente con la ubicación de la variedad 24, que presentó los mayores valores de cantidad de frutos por planta en esa campaña.

Para el peso promedio del fruto en la campaña agrícola 2010–2011 (representado en azul), los marcadores significativos se ubicaron en una región diferente del gráfico, con valores negativos en la componente 1, positivos y bajos en la componente 2 y altos en la componente 3. Esta ubicación se encontró próxima a la variedad 54, que presentó los mayores valores de peso promedio del fruto en dicha campaña (Figura 3.6, izquierda). Por su parte, el carácter calibre ponderado (representado en naranja) mostró marcadores asociados en el mismo subpolígono que los observados para rendimiento y cantidad de frutos por planta, destacándose la variedad 34, que presentó valores elevados de calibre ponderado. En este caso, los valores fueron similares para las componentes 1 y 3, pero considerablemente menores para la componente 2.

En la campaña agrícola 2011–2012, el rendimiento (representado en rosa) mostró marcadores asociados distribuidos entre los cuadrantes superior izquierdo y superior derecho en el gráfico formado por las componentes 1 y 2 (Figura 3.6, izquierda). Sin embargo, en el gráfico correspondiente a las componentes 2 y 3 (Figura 3.6, derecha), estos marcadores se concentraron presentando valores positivos pequeños para la componente 2 y elevados para la componente 3. Para el carácter cantidad de frutos por planta (representado en amarillo) se observaron coincidencias con los marcadores asociados al rendimiento en la misma campaña (‘23’, ‘24’, ‘26’, ‘128’, ‘129’ y ‘297’), ubicándose la mayoría en el subpolígono del cuadrante superior izquierdo y presentando valores positivos y elevados en la componente 3.

El peso promedio del fruto en esta campaña (representado en celeste) presentó marcadores coincidentes con los observados para cantidad de frutos por planta (‘50’, ‘51’, ‘57’, ‘92’ y ‘128’). Estos se caracterizaron por presentar valores elevados en las componentes 1 y 2 (Figura 3.6, izquierda), mientras que los valores en la componente 3 fueron variables (Figura 3.6, derecha). En el eje correspondiente a la componente 2, estos valores resultaron similares a los de la variedad 30, que presentó el mayor peso

promedio del fruto en esta campaña. Finalmente, el calibre ponderado (representado en púrpura) presentó dos marcadores asociados ubicados en el subpolígono del cuadrante inferior derecho, con valores negativos en las componentes 2 y 3 (Figura 3.6, derecha) y valores variables en la componente 1 (Figura 3.6, izquierda). En esta región también se ubicó la variedad 14, que presentó los valores más negativos para la componente 2.

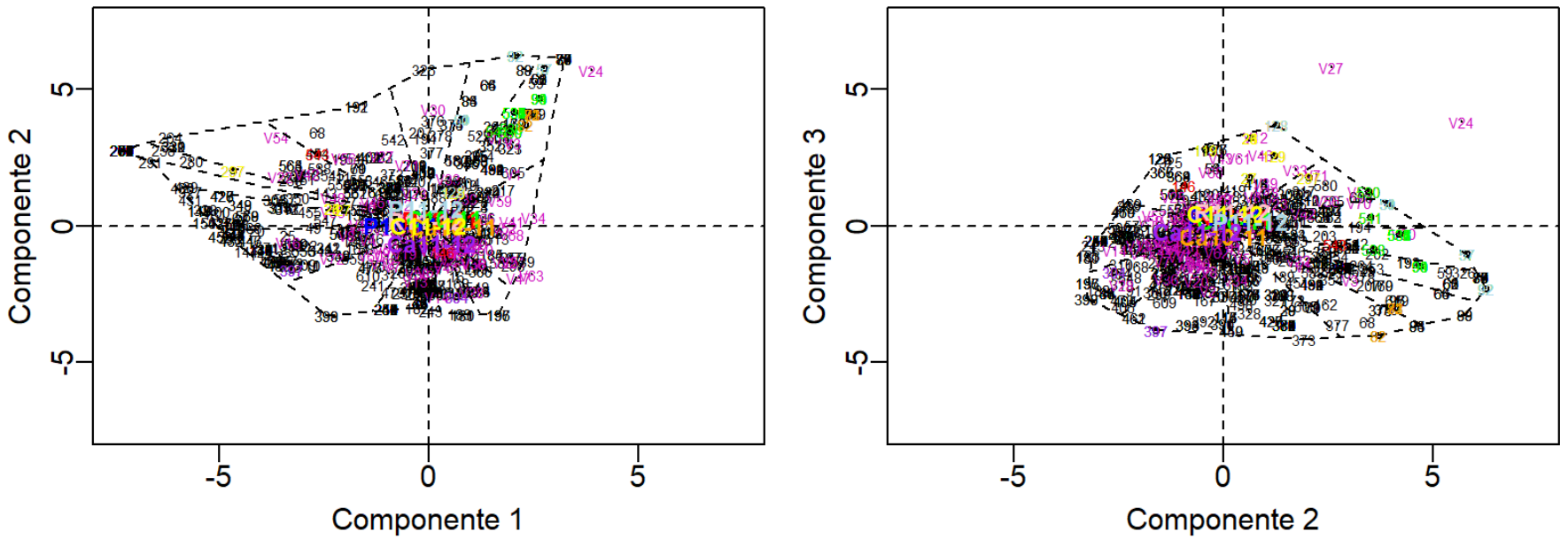

Figura 3.6: Gráficos Triplots obtenidos por PLS con el algoritmo algebraico SVD para las componentes 1 y 2 (izquierda) y para las componentes 2 y 3 (derecha). Las variedades de duraznero se representan en color violeta y los marcadores moleculares no significativos en color negro. Se muestran las variables fenotípicas y sus marcadores moleculares estadísticamente significativos: rendimiento en 2010-2011 (rojo), cantidad de frutos por planta en 2010-2011 (verde), peso en 2010-2011 (azul), rendimiento en 2011-2012 (naranja), cantidad de frutos por planta en 2011-2012 (rosa), peso en 2011-2012 (amarillo), calibre en 2010-2011 (celeste) y calibre en 2011-2012 (purpura). Sobre los marcadores extremos se traza el polígono de contención en líneas punteadas negras y rectas desde el centro de coordenadas al punto medio de cada segmento

En el triplot correspondiente al algoritmo NIPALS (Figura 3.7) se observó que, a diferencia del caso anterior, los valores de las variables fenotípicas se alejaron del centro de coordenadas, excepto para el carácter cantidad de frutos por planta representado en amarillo.

Para el primer carácter fenotípico, rendimiento medido en la campaña agrícola 2010–2011 (representado en rojo), los marcadores significativos se ubicaron en el extremo del cuadrante inferior izquierdo en el gráfico formado por las componentes 1 y 2 (Figura 3.7, izquierda), presentando además valores bajos en la componente 3. Estos marcadores se encontraron cercanos a las coordenadas de las variedades 24, 60 y 70, que presentaron valores elevados de este carácter fenotípico.

Para el carácter cantidad de frutos por planta en la campaña 2010–2011, representado en verde, los marcadores significativos coincidieron casi en su totalidad con los asociados al carácter rendimiento medido en la misma campaña. Por este motivo, los colores se observaron superpuestos y no se distinguen claramente los valores correspondientes al rendimiento en rojo.

Los caracteres fenotípicos peso promedio del fruto (azul) y calibre ponderado (naranja) en la campaña agrícola 2010–2011 no presentaron marcadores moleculares estadísticamente significativos. Estos puntos se ubicaron en extremos opuestos del gráfico entre las componentes 1 y 2 (Figura 3.7, izquierda). En particular, el peso promedio del fruto presentó valores positivos en la componente 1 y valores cercanos al origen en las componentes 2 y 3, mientras que el calibre ponderado presentó valores negativos en la componente 1 y valores cercanos al origen en las componentes 2 y 3.

Para el rendimiento medido en la campaña agrícola 2011–2012, los marcadores se ubicaron en la región inferior central del gráfico entre las componentes 1 y 2 (Figura 3.7, izquierda), mientras que en el gráfico formado por las componentes 2 y 3 se localizaron en la esquina superior izquierda (Figura 3.7, derecha). En ambos casos se encontraron cercanos a las variedades 27 y 33, que presentaron los valores más altos de este carácter.

Los marcadores moleculares asociados al carácter cantidad de frutos por planta para la campaña agrícola 2011–2012 coincidieron en gran medida con los encontrados para el rendimiento y el peso promedio del fruto en la misma campaña, por lo que también se observaron superpuestos. Este carácter se representó en amarillo y su comportamiento

fue consistente con la interpretación anterior, al igual que el peso promedio del fruto representado en celeste.

Finalmente, para el calibre ponderado en la campaña agrícola 2011–2012 (representado en púrpura), tanto el carácter fenotípico como los marcadores moleculares estadísticamente significativos se ubicaron en el extremo superior central del gráfico entre las componentes 1 y 2 (Figura 3.7, izquierda) y en el extremo derecho del gráfico entre las componentes 2 y 3 (Figura 3.7, derecha). En esta región se destacaron las variedades 14, 13 y 47, que presentaron valores elevados para este carácter.

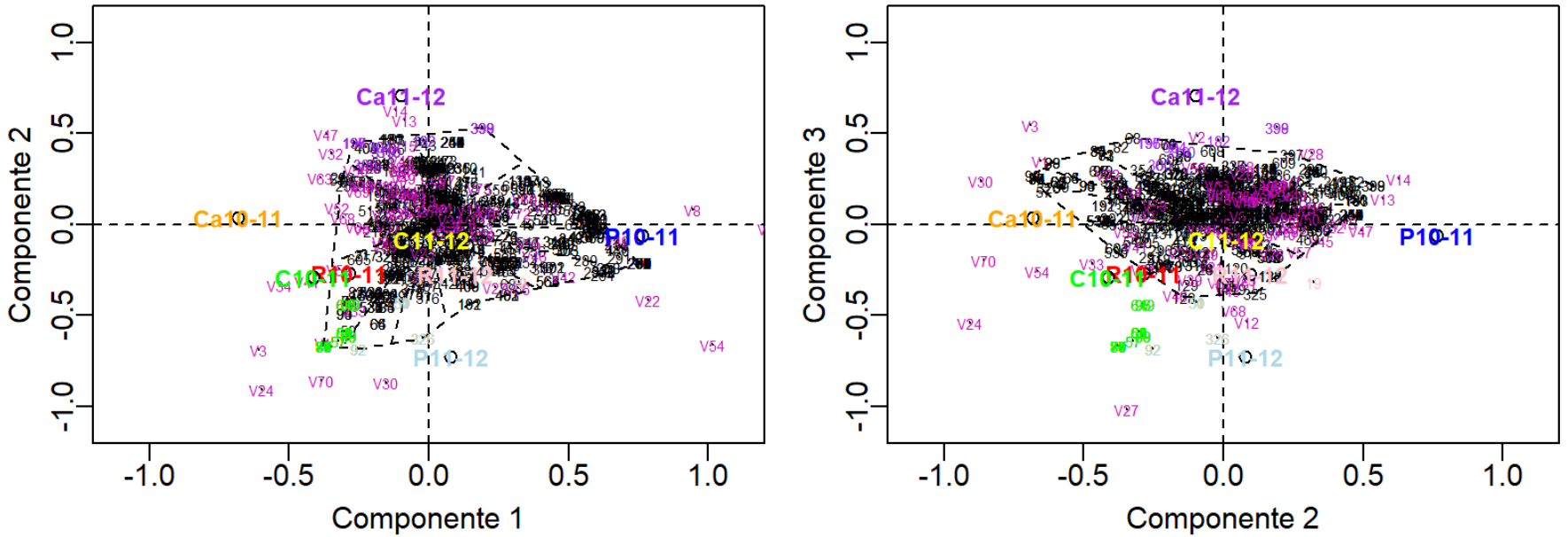

Figura 3.7: Gráficos Triplots obtenidos por PLS con el algoritmo algebraico NIPALS para las componentes 1 y 2 (izquierda) y para las componentes 2 y 3 (derecha). Las variedades de duraznero se representan en color violeta y los marcadores moleculares no significativos en color negro. Se muestran las variables fenotípicas y sus marcadores moleculares estadísticamente significativos: rendimiento en 2010-2011 (rojo), cantidad de frutos por planta en 2010-2011 (verde), peso en 2010-2011 (azul), rendimiento en 2011-2012 (naranja), cantidad de frutos por planta en 2011-2012 (rosa), peso en 2011-2012 (amarillo), calibre en 2010-2011 (celeste) y calibre en 2011-2012 (purpura). Sobre los marcadores extremos se traza el polígono de contención en líneas punteadas negras y rectas desde el centro de coordenadas al punto medio de cada segmento

En la primer campaña agrícola evaluada, NIPALS fue el algoritmo algebraico que detectó un número mayor de marcadores moleculares asociados al caracter rendimiento, con 25 marcadores moleculares estadísticamente significativos sin ninguna coincidencia entre ambos algoritmos algebraicos; para el carácter de cantidad de frutos por planta, sólo dos marcadores coincidieron entre SVD y NIPALS, siendo nuevamente este último el que detectó mayor cantidad (22 marcadores moleculares estadísticamente significativos); en cambio, para el carácter peso promedio del fruto, únicamente SVD identificó asociaciones significativas, encontrando cuatro marcadores moleculares asociados al caracter; finalmente para el calibre ponderado NIPALS tampoco encontró marcadores significativos asociados, mientras que SVD encontró cuatro. En la segunda campaña agrícola evaluada (2011–2012), el patrón difirió: para el rendimiento no se observaron coincidencias en los marcadores, siendo SVD el que detectó un número mayor (seis marcadores moleculares estadísticamente significativos); en cantidad de frutos por planta, tampoco hubo coincidencias entre ambos algoritmos, siendo SVD el que detectó un número mayor de cinco marcadores moleculares; para el Peso promedio por fruto la coincidencia fue de tres marcadores moleculares detectados como significativos encontrando mayor cantidad en SVD (cinco marcadores moleculares estadísticamente significativos); y por último el calibre ponderado que coincidió en un marcador, encontrando NIPALS 10 marcadores moleculares significativos (Tabla 3.5).

Tabla 3.5: Marcadores moleculares estadísticamente significativos asociados a las variables: Rendimiento en la campaña agrícola 2010-2011 (R10-11), Rendimiento en la campaña agrícola 2011-2012 (R11-12), Cantidad de frutos por planta en la campaña agrícola 2010-2011 (C10-11), Cantidad de frutos por planta en la campaña agrícola 2011-2012 (C11-12), Peso en la campaña agrícola 2010-2011 (P10-11), Peso en la campaña agrícola 2011-2012 (P11-12), Calibre en la campaña agrícola 2010-2011 (Ca10-11), Calibre en la campaña 2011-2012 (Ca11-12); según PLS con las descomposiciones SVD (primero) y NIPALS (segundo)

Algoritmo algebraico-variable fenotípica	**Marcadores moleculares estadísticamente significativos**
SVD - R10-11	**23–24–26–128–129–146–530–541–543**
SVD - R11-12	**23–24–26–128–129–297**
SVD - C10-11	**50–51–57–90–92–93–94–128–129–530–531–532 533–534–535–536–537–538–541**
SVD - C11-12	**23–24–26–27–119–128–129–297**
SVD - P10-11	**23–24–26–297**
SVD - P11-12	**50–51–57–92–128**
SVD - Ca10-11	**72–73–81–82**
SVD - Ca11-12	**304–397**
NIPALS - R10-11	**50–51–55–57–60–61–62–63–74–75–76–77–78 79–80–86–87–88–89–92–95–96–97–326–619**
NIPALS - R11-12	**19–50–51–92–326**
NIPALS - C10-11	**55–57–60–61–62–63–74–75–76–77–78–79 80–86–87–88–89–92–95–96–97–619**
NIPALS - C11-12	**50–51–92–93–326**
NIPALS - P10-11	-
NIPALS - P11-12	**50–51–92–326**
NIPALS - Ca10-11	-
NIPALS - Ca11-12	**182–195–196–197–300–304–310–319–398–399**

3.4.3. Resultados obtenidos sobre la aplicación de PLS en una base de datos de maíz

La distribución de las seis variables fenotípicas evaluadas para este conjunto de datos indicaron que para el Ambiente WW los caracteres fenotípicos presentaron mayor variabilidad y distribuciones más simétricas que para el Ambiente SS. La distribución de la variable ASI toma valores entre -4 y 4 y los caracteres FFL y MFL toman valores entre 10 y 20 presentando mayor variabilidad (Figura 3.8).

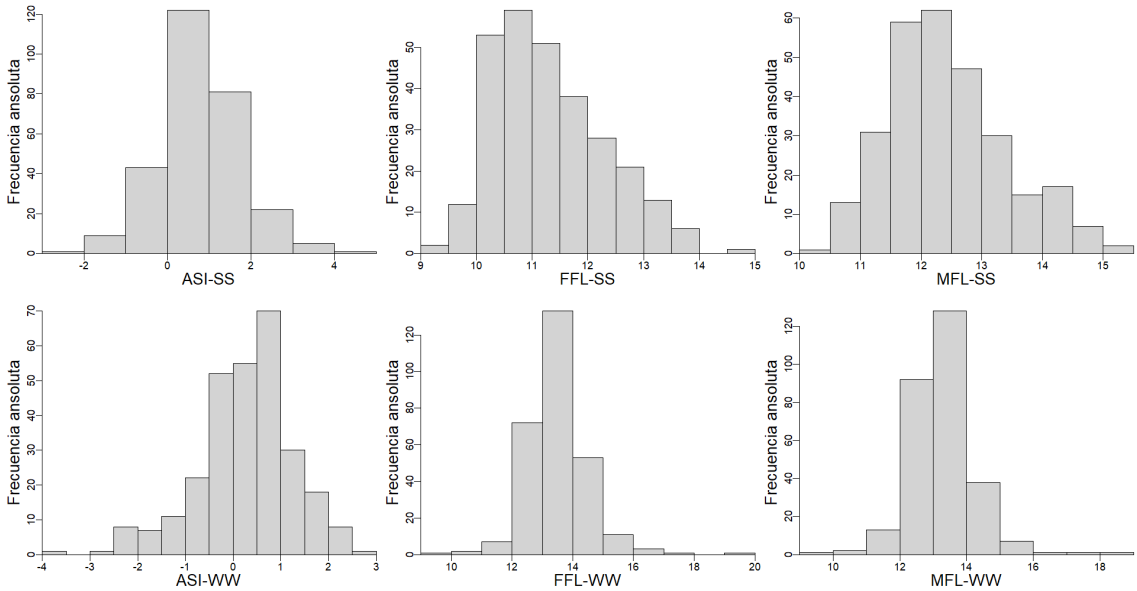

Figura 3.8: Frecuencia absoluta de los valores de ASI en el Ambiente SS (izquierda arriba) en el Ambiente WW (izquierda abajo), valores de FFL en el Ambiente SS (centro arriba), en el Ambiente WW (centro abajo), valores de MFL (derecha arriba) y en el Ambiente WW (derecha abajo)

La cantidad de variables latentes determinadas a través del test de permutaciones y la covarianza acumulada fue de 6 variables latentes (mínimo valor cardinal entre los rangos de X e Y). Dos variables latentes resultaron estadísticamente significativas con el test de permutaciones pero la covarianza explicada por estos dos componentes alcanzan el 45 % en el caso de SVD y el 52 % en el caso de NIPALS. Considerando tres variables latentes la covarianza acumulada explica el 62 % para SVD y el 68 % para NIPALS, razón por la cual en este estudio se consideró utilizar tres (Tabla 3.6).

Tabla 3.6: Valores de probabilidad asociada al test de permutaciones para cada variable latente obtenida como resultado de cada uno de los algoritmos (SVD y NIPALS) de PLS, porcentaje de covarianza explicada y porcentaje de covarianza acumulada en cada variable latente para cada algoritmo (SVD y NIPALS)

Cantidad de variables latentes	valores p*		Covarianza acumulada (%)		Covarianza acumulada (%)	
	SVD	NIPALS	SVD	NIPALS	SVD	NIPALS
1	0,9530	0,0001	19,60	28,08	19,60	28,08
2	0,0001	0,0001	25,09	24,46	44,69	52,54
3	0,6000	0,9450	17,72	15,99	62,40	68,53
4	0,8920	1,0000	11,90	12,52	74,30	81,05
5	0,1490	1,0000	13,64	10,53	87,94	91,58
6	0,5170	1,0000	12,06	8,42	100,00	100,00

*valores p obtenidos del test de permutaciones

La mediana de la distribución de la RECM presentó valores más bajos utilizando el algoritmo NIPALS en comparación con SVD. Esta diferencia entre los valores de la mediana de las distribuciones fue aproximadamente de 0.04 con todos los caracteres agrícolas. También se observaron valores mayores de la media de la distribución de la RECM utilizando SVD en comparación con NIPALS, en este caso la diferencia fue mayor a 0,044 en todos los casos, siendo el caracter ASI medido en la condición SS el que presentó la diferencia mayor (0,80 con NIPALS vs 0,86 con SVD). Sin embargo la variabilidad de las distribuciones no siempre fue mayor utilizando el algoritmo algebraico NIPALS, mientras que para el caracter ASI medido en ambas condiciones (SS y WW) y el caracter MFL medido en la condición SS la variabilidad es menor para la distribución de la RECM calculado utilizando NIPALS con respecto a la misma distribución utilizando el algoritmo SVD, para los caracteres MFL y FFL medidos en condiciones WW el desvío estándar fue mayor, y para el caracter agrícola FFL medio en condiciones SS, la variabilidad resultó similar (Figura 3.9).

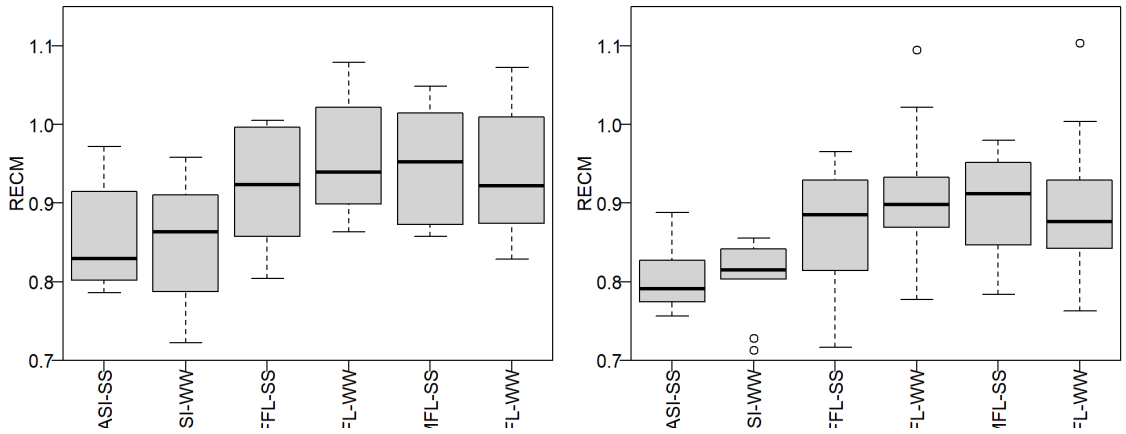

Figura 3.9: Boxplots de los coeficientes de variación de la Raíz del Error Cuadrático Medio (RECM) en los cuatro ambientes. Para la descomposición SVD (derecha) y para NIPALS (izquierda)

Para el modelo PLS estimado mediante el algoritmo algebraico SVD, los marcadores moleculares destacados en rojo, correspondientes a los asociados al carácter ASI medido en la condición SS, se ubicaron en el cuadrante inferior izquierdo en la Figura 3.10-izquierda, presentando valores negativos tanto en la componente 1 como en la componente 2. Asimismo, estos marcadores también se localizaron en la misma región en la Figura 3.10-derecha, mostrando valores negativos en la componente 3. Para este carácter se destacó la variedad 65, que presentó los valores más altos y se ubicó en el extremo inferior tanto con respecto a la componente 2 como a la componente 3. Sin embargo, estos marcadores no se distinguen en color rojo en el gráfico debido a que el 60 % coincidió con los marcadores asociados al carácter ASI en la condición WW (Tabla 3.7), representados en verde. Estos marcadores se localizaron en el extremo izquierdo de la Figura 3.10-izquierda y en el centro de la distribución en la Figura 3.10-derecha, mostrando principalmente valores altos y negativos en la componente 1. En esta región se destacó la variedad '255', ubicada en el extremo izquierdo de la Figura 3.10-izquierda y que presentó valores elevados de este carácter.

Para la variable fenotípica FFL medida en la condición SS, los marcadores asociados se representaron en azul y en su mayoría se solaparon con los marcadores representados en amarillo, correspondientes al carácter MFL medido en la condición WW. Estos marcadores se ubicaron en el extremo inferior de la Figura 3.10-izquierda y presentaron además valores negativos en la componente 3. Cercanas a estos marcadores se destaca-

ron las variedades ‘65’ y ‘84’, que se encuentran entre las que presentaron los valores más altos de este carácter.

El carácter FFL medido en la condición WW presentó marcadores dispersos tanto en la Figura 3.10-izquierda como en la Figura 3.10-derecha, también solapados con los marcadores representados en amarillo y en rosa, que correspondieron a los marcadores moleculares asociados al carácter MFL medido en la condición SS. Finalmente, este último carácter mostró marcadores con valores negativos para la componente 2, pero más dispersos en las otras dos componentes. En esta región se destacaron variedades cercanas en la Figura 3.10-derecha, como ‘176’, ‘173’ y ‘90’, que correspondieron a las que presentaron los valores más elevados.

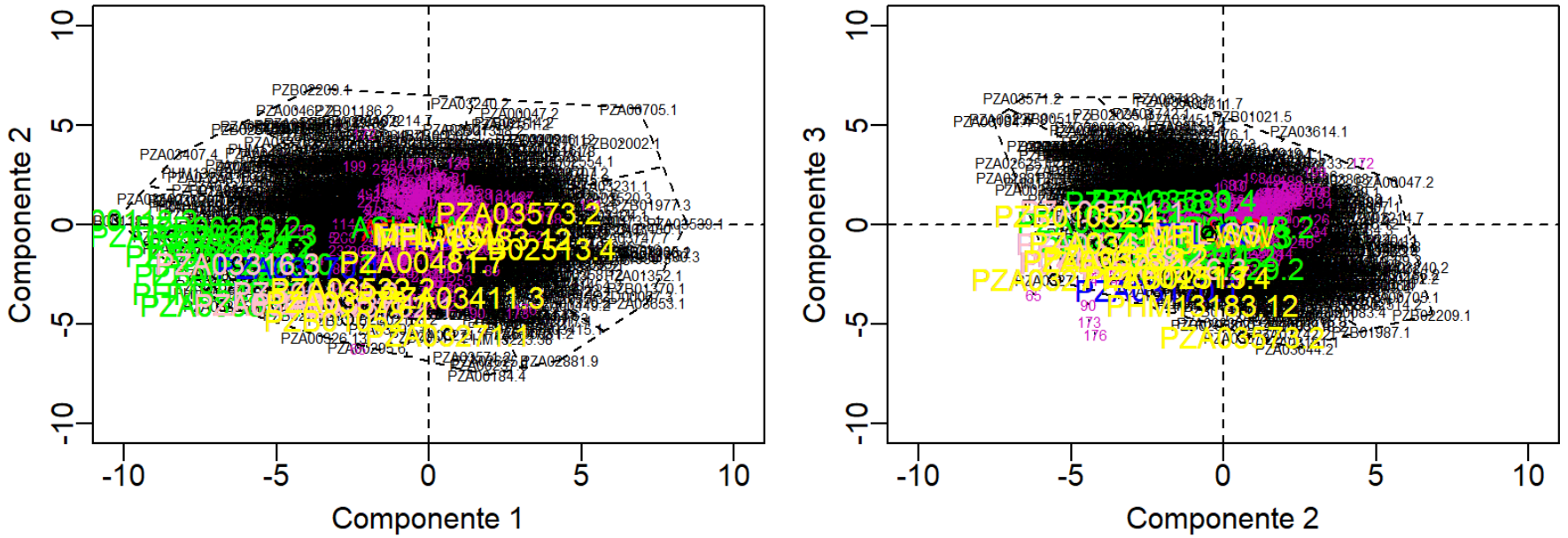

Figura 3.10: Triplots obtenidos mediante el algoritmo algebraico SVD entra la Componente 1 y la Componente 2 (izquierda) y entre la Componente 2 y la Componente 3 (derecha). Las variedades se representan en violeta y los marcadores moleculares no significativos en negro. Las variables fenotípicas y sus marcadores moleculares estadísticamente significativos se grafican por color: ASI en Ambiente SS (rojo), ASI en Ambiente WW (verde), FFL en SS (azul), FFL en WW (naranja), MFL en SS (rosa) y MFL en WW (amarillo). Sobre los marcadores extremos se traza el polígono de contención y una recta del centro de coordenadas al punto medio del segmento

Para el modelo PLS estimado mediante el algoritmo algebraico NIPALS, se observó que el carácter fenotípico ASI en la condición SS, representado en rojo, presentó marcadores asociados ubicados en el cuadrante inferior izquierdo de la Figura 3.11-izquierda, al igual que en el caso del modelo estimado mediante SVD. Estos marcadores coincidieron en un 60 % con los asociados al carácter ASI en la condición WW (Tabla 3.7). Las variedades cercanas a estos marcadores fueron '233', '254' y '5', destacándose la variedad '254' por presentar el valor más alto para este carácter.

En el caso de la variable fenotípica FFL medida en la condición SS, tanto el valor de la variable como sus marcadores asociados se representaron en azul, destacándose como el carácter con el valor negativo más extremo en la componente 2, cercano a las variedades '65' y '209', que presentaron valores elevados para este carácter (Figura 3.11-izquierda). De estos marcadores, el 46 % coincidió con los asociados al carácter FFL medido en la condición WW (Tabla 3.7), representados en los triplots en color naranja. Esta variable fenotípica presentó valores similares en la componente 2, pero valores positivos en la componente 1 (Figura 3.11-izquierda), y valores negativos en la componente 3. Las variedades '173', '65' y '90' se encontraron próximas en la Figura 3.11-derecha y presentaron valores altos para este carácter.

En cuanto al carácter MFL medido en la condición SS, el 100 % de los marcadores asociados coincidieron con los marcadores asociados al carácter FFL medido en la condición WW. Esta variable fenotípica y sus marcadores se representaron en rosa. Finalmente, en amarillo se presenta el carácter MFL y sus marcadores moleculares asociados, los cuales coincidieron casi en su totalidad con los marcadores asociados al mismo carácter pero medido en la condición SS (Tabla 3.7).

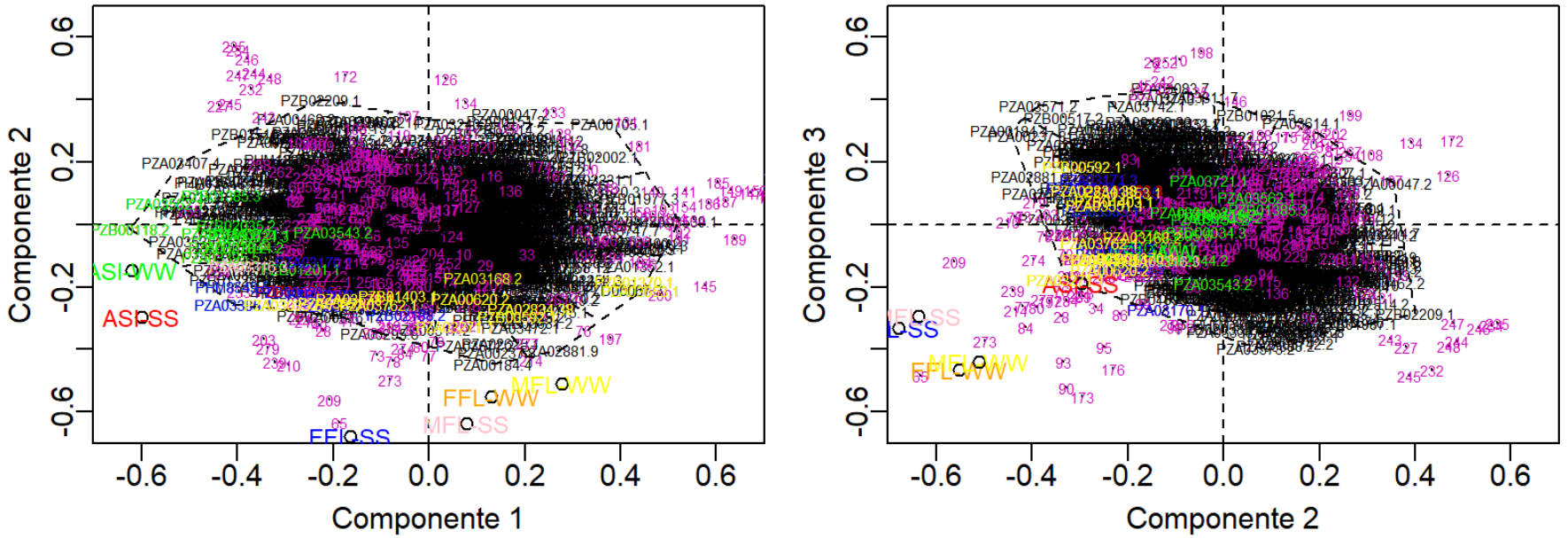

Figura 3.11: Triplots obtenidos mediante el algoritmo algebraico NIPALS entra la Componente 1 y la Componente 2 (izquierda) y entre la Componente 2 y la Componente 3 (derecha). Las variedades se representan en violeta y los marcadores moleculares no significativos en negro. Las variables fenotípicas y sus marcadores moleculares estadísticamente significativos se grafican por color: ASI en Ambiente SS (rojo), ASI en Ambiente WW (verde), FFL en SS (azul), FFL en WW (naranja), MFL en SS (rosa) y MFL en WW (amarillo). Sobre los marcadores extremos se traza el polígono de contención y una recta del centro de coordenadas al punto medio del segmento

Tabla 3.7: Marcadores moleculares estadísticamente significativos asociados a las variables: ASI en el Ambiente SS (ASI-SS), ASI en el Ambiente WW (ASI-WW), FFL en el Ambiente SS (FFL-SS), FFL en el Ambiente WW (FFL-WW), MFL en el Ambiente SS (MFL-SS), MFL en el Ambiente WW (MFL-WW); según PLS con las descomposiciones SVD (primero) y NIPALS (segundo)

Algoritmo algebraico-variable fenotípica	**Marcadores moleculares estadísticamente significativos**
SVD - ASI-SS	PZA03170.1-PHM8549.5-PZA03578.1-PZA03035.5-PZA03271.1 PZA03395.5-PZB00229.2-PZB01052.4-PZD00034.3 PZA03385.1-PZB02044.2-PZB00109.2 PZB01201.1-PZB00118.2-PZA03316.3-PZA03541.4
SVD - ASI-WW	PHM8549.5-PZA03535.3-PZA03395.5-PZB00229.2 PZD00034.3-PZA03385.1-PZB02044.2-PZB01201.1 PZB00118.2-PZA03316.3-PZB01358.3-PZA03541.4-PZA03580.4
SVD - FFL-SS	PZA03551.1-PZA03762.1-PZA03170.1-PZA03578.1-PZA03035.5 PZA03271.1-PZA00481.7-PZA03411.3-PZB01052.4-PHM13183.12 PZB00109.2-PZA03532.2-PZA03316.3
SVD - FFL-WW	PZB02513.4-PZA03578.1-PZA03035.5-PZA03271.1-PZA00481.7 PZA03411.3-PZB01052.4-PHM13183.12-PZA03573.2-PZA03532.2
SVD - MFL-SS	PZA03551.1-PZA03762.1-PZA03578.1-PZA03035.5-PZA03271.1 PZA00481.7-PZA03411.3-PZB01052.4-PHM13183.12-PZB00109.2 PZA03532.2-PZA03316.3
SVD - MFL-WW	PZB02513.4-PZA03578.1-PZA03271.1-PZA00481.7-PZA03411.3 PZB01052.4-PHM13183.12-PZA03573.2-PZA03532.2
NIPALS - ASI-SS	PZA03762.1-PZA03170.1-PHM8549.5-PZA03578.1-PZB01186.3 PZA03035.5-PZA03395.5-PZA03543.2-PZA03171.3-PZD00022.5 PZD00034.3-PZA03385.1-PZB02044.2-PZB01201.1-PZB00118.2 PZA03316.3-PZA03218.2
NIPALS - ASI-WW	PZA03721.1-PZA03154.2-PZA03562.1-PHM8549.5-PZB01385.3 PZA03395.5-PZA03543.2-PZB02035.2-PZD00034.3-PZA03385.1 PZB02044.2-PZB01201.1-PZB00118.2-PZA03316.3-PZA03218.2
NIPALS - FFL-SS	PZA03551.1-PZB02188.2-PZA03762.1-PZA03170.1-PHM8549.5 PZA03578.1-PZA03035.5-PZA03271.1-PZA03171.3-PZD00022.5 PZA03385.1-PZB00109.2-PZA03316.3
NIPALS - FFL-WW	PZA03551.1-PZA03762.1-PZA00620.2-PZA03578.1-PZB01370.1 PZA03035.5-PZA03271.1-PZA02834.38-PZB00592.1-PZB01403.1 PZA03316.3

Algoritmo algebraico-variable fenotípica	Marcadores moleculares estadísticamente significativos
NIPALS - MFL-SS	PZA03551.1-PZA03762.1-PZA00620.2-PZA03578.1-PZA03035.5 PZA03271.1-PZA02834.38-PZB00592.1-PZB01403.1-PZA03316.3
NIPALS - MFL-WW	PZA03653.1-PZA03551.1-PZA03762.1-PZA00620.2-PZA03578.1 PZB01370.1-PZA03035.5-PZA03271.1-PZA02834.38-PZA03168.2 PZB00592.1-PZB01403.1

3.5. Discusión

El ACP constituye una herramienta útil para la reducción de dimensionalidad, sin embargo, su naturaleza no supervisada limita su utilidad en contextos predictivos, ya que las componentes se construyen únicamente en función de la variabilidad de la matriz X que contiene todas las variables, es decir, no hay un conjunto de variables consideradas explicativas o regresoras de un conjunto de variables consideradas variables respuesta, también denominada matriz Y en el contexto del análisis PLS. El método PLS incorpora la información de las variables respuesta (matriz Y) en la construcción de las componentes latentes, orientando la reducción de dimensión hacia la conformación de combinaciones lineales que maximicen la correlación entre el conjunto de variables de la matriz Y con el conjunto de variables de la matriz X, las cuales pueden ser utilizadas con fines de predicción (Sun et al., 2019). Montesinos-López et al. (2022) evaluaron modelos PLS multicaracter en el ámbito de la predicción genómica y reportaron, en varios de los escenarios analizados, mejoras de precisión respecto de enfoques basados en G-BLUP. Si bien dichos resultados se obtuvieron en un marco de predicción y no específicamente de asociación, sugieren que los métodos PLS pueden representar una alternativa útil para explotar la estructura conjunta de marcadores y fenotipos en estudios genómicos complejos.

La implementación del algoritmo NIPALS o del algoritmo SVD como paso previo a la estimación de la regresión PLS revela diferencias relevantes tanto desde el punto de vista computacional como estadístico. Aunque ambos métodos conducen a soluciones PLS equivalentes en escenarios ideales, su comportamiento difiere cuando se

consideran aspectos prácticos como datos faltantes, estabilidad numérica y escalabilidad. El algoritmo NIPALS presenta la ventaja de estimar las componentes latentes de forma iterativa, lo que permite trabajar naturalmente con datos incompletos sin necesidad de imputación previa. Esta característica resulta especialmente valiosa en contextos genómicos, donde la presencia de valores faltantes es frecuente. Aunque el método depende de criterios de convergencia, su estrategia secuencial de extracción de componentes evita descomposiciones matriciales completas, que en contextos de alta dimensionalidad se traduce en una reducción del costo computacional, tal como evidenciaron los resultados obtenidos en este capítulo y ha sido señalado en aplicaciones de PLS en datos de alta dimensión (Mevik y Wehrens, 2007). Por su parte, la implementación basada en SVD proporciona una solución algebraica directa, asociada a buenas propiedades de estabilidad numérica y reproducibilidad (Golub y Van Loan, 2013). Sin embargo, su costo computacional depende de la descomposición de la matriz completa, lo que puede resultar demandante en escenarios de alta dimensionalidad (Hastie et al., 2009). En particular, cuando el número de predictores supera ampliamente al de observaciones y solo se requiere un número reducido de variables latentes, este enfoque puede implicar cálculos innecesarios. Este comportamiento fue consistente con los resultados observados en el presente estudio (Halko et al., 2011).

Desde el punto de vista visual, el triplot obtenido mediante la implementación basada en NIPALS presentó una estructura más clara y estable que la derivada de SVD. La comparación entre las implementaciones basadas en NIPALS y SVD muestra que, si bien ambas recuperan estructuras globales similares en el espacio latente, se observan diferencias en la dispersión relativa de los genotipos y en la definición de las direcciones asociadas a los marcadores. Estas discrepancias se atribuyen a la naturaleza iterativa de NIPALS frente a la solución algebraica directa provista por SVD, lo cual puede impactar en la estabilidad numérica y en la calidad visual de la representación. En consecuencia, para fines exploratorios y de interpretación gráfica en el contexto analizado, la implementación basada en NIPALS resultó más adecuada para la construcción del triplot. Sin embargo, debe tenerse en cuenta que el triplot constituye una proyección bidimensional del espacio latente, por lo que las relaciones de proximidad entre variables y genotipos deben interpretarse de manera aproximada, considerando la proporción de

variabilidad explicada por las componentes representadas. Además, cabe señalar que la alta concentración de individuos en la región central del triplot sugiere que la capacidad discriminante en las primeras componentes es moderada.

En relación con el desempeño predictivo, los valores de la RECM obtenidos para los modelos ajustados se mantuvieron en magnitudes reducidas, lo que indica una adecuada capacidad de ajuste y predicción en los datos analizados. En términos comparativos, las diferencias observadas entre las implementaciones evaluadas fueron pequeñas, lo que sugiere que ambas logran capturar de manera similar la estructura de covarianza relevante entre los conjuntos X e Y. Debido a la estandarización de las variables respuesta, el RECM puede interpretarse en unidades de desvío estándar. Los resultados muestran que, para la mayoría de los fenotipos, el error predictivo se mantiene por debajo de una desviación estándar, evidenciando que los modelos PLS logran capturar señal predictiva aunque con precisión moderada, esperable en rasgos cuantitativos complejos. En términos comparativos, los algoritmos SVD y NIPALS exhibieron desempeños predictivos muy similares.

3.6. Conclusión

En este capítulo se aplicó la técnica de regresión PLS utilizando dos algoritmos distintos: SVD y NIPALS, con el objetivo de identificar marcadores moleculares significativamente asociados a caracteres fenotípicos medidos en más de un ambiente. Los resultados obtenidos indicaron que ambos algoritmos tienden a encontrar números similares de marcadores significativos. Sin embargo, el algoritmo NIPALS presenta en la mayoría de casos, valores de RECM más chicos al realizar validación cruzada, indicando que la diferencia entre la predicción de las variedades y los valores reales es menor con esta técnica. Finalmente NIPALS cuenta con ventajas en tiempos de implementación y se permite en casos de valores faltantes como se analizará en el Capítulo 4. En conjunto, los hallazgos de este capítulo refuerzan la utilidad de PLS como herramienta multivariada para estudios de asociación genómica.

Capítulo 4

Regresión por Mínimos Cuadrados Parciales en el estudio de asociación fenotipo-genotipo en presencia de datos faltantes

4.1. Métodos multivarariados en escenarios de datos faltantes

La rápida mejora de la tecnología de genotipado de alto rendimiento y la drástica reducción de los costes del genotipado han propiciado el uso generalizado de los estudios GWAS. Dado que las plataformas de genotipado actuales solo analizan una pequeña fracción de los SNP del genoma, muchos loci quedarán inevitablemente sin tipificar (es decir, sin genotipar). Por lo tanto, es muy conveniente realizar análisis de asociación en SNP no tipificados. Dichos análisis pueden ayudar a localizar variantes causales y facilitar la selección de SNP para estudios de seguimiento (Hu y Lin, 2010).

La presencia de datos ausentes o no disponibles (NA) es un desafío recurrente en el análisis multivariado, particularmente en estudios genómicos y fenotípicos donde la recolección de información puede ser costosa, parcial o técnicamente limitada (S. R. Browning, 2008). Tradicionalmente, esta problemática se aborda mediante técnicas de imputación, como la sustitución por la media o la regresión múltiple (de Souto et al., 2015), o mediante la eliminación directa de observaciones incompletas. Sin embargo, estas estrategias pueden introducir sesgos o reducir la varianza explicada. Esta problemática ha recibido mucha atención en los últimos años, dado que se espera maximizar el uso de la información disponible sin introducir sesgos significativos.

El ACP ha sido ampliamente utilizado como herramienta de reducción de dimensionalidad en contextos multivariantes (I. T. Jolliffe y Cadima, 2016). No obstante, su

aplicación directa requiere matrices completas, lo que limita su uso en presencia de valores faltantes. En este escenario, el algoritmo NIPALS representa una alternativa robusta, ya que permite estimar componentes principales sin necesidad de imputar previamente los datos (De Souza et al., 2024). Su carácter iterativo permite aprovechar toda la estructura disponible de los datos observados, manteniendo la integridad de las relaciones entre variables.

La técnica PLS pertenece a la familia de métodos multivariados que modela las relaciones entre conjuntos de variables observadas mediante variables latentes o componentes. La suposición subyacente es que los datos observados son generados por un sistema impulsado por un pequeño número de variables latentes que no se observan directamente. En su forma más general, PLS crea vectores de scores ortogonales (variables latentes) maximizando la covarianza entre diferentes conjuntos de variables. El método PLS, en su forma clásica, se basa en el algoritmo iterativo NIPALS (Rosipal y Krämer, 2006).

El algoritmo NIPALS ha demostrado ser especialmente útil en conjuntos de datos de alta dimensionalidad y complejidad estructural, como los encontrados en genética (S. Wold et al., 2001), sensometría, espectroscopía (Perez-Guaita et al., 2013) y química analítica. Su eficiencia computacional y su capacidad para operar con matrices incompletas lo convierten en una herramienta clave para extender el uso de métodos multivariados como ACP y PLS a contextos donde otras metodologías resultan ineficaces o imprecisas.

Bajo este escenario, incluso considerando los avances, aún existe la necesidad de mayor exploración y evaluación del método PLS estimado a través de algoritmo NIPALS en contextos GWAS. Este capítulo busca continuar esta investigación mediante el estudio del desempeño de método bajo distintos niveles de valores faltantes, con el objetivo de evaluar su aplicabilidad y robustez para identificar asociaciones genotipo-fenotipo en condiciones realistas de calidad de datos.

4.2. Objetivo

Evaluar la eficiencia del algoritmo NIPALS frente a distintos niveles de datos faltantes en la matriz de marcadores moleculares en la identificación de asociación fenotipo-

genotipo.

4.3. Materiales y Métodos

4.3.1. Conjuntos de datos utilizados

Las bases de datos utilizadas en este Capítulo 4 se corresponden con los tres conjuntos de datos descriptos en el Capítulo 3, en la sección de Materiales y Métodos, con el objetivo de evaluar la capacidad del modelo de regresión PLS para detectar asociaciones fenotipo-genotipo bajo distintos niveles de datos faltantes. Con el propósito de facilitar la lectura, se recuerda que las bases de datos empleadas fueron: (i) un conjunto de datos simulados para un caracter fenotípico de distribución normal en dos ambientes y sus correspondientes marcadores moleculares asociados, (ii) una base de datos pública de maíz con registros fenotípicos en dos ambientes (SS y WW), y (iii) una base de datos real de duraznero con variables medidas en dos campañas agrícolas (2010-2011 y 2011-2012).

4.3.2. Generación de datos faltantes: Estudio por simulación

Los distintos niveles de datos faltantes se generaron a partir de una matriz de datos completos X. Luego, se realizó una eliminación aleatoria de valores siguiendo un muestreo aleatorio simple, lo que implica que un valor de la matriz de marcadores moleculares esté ausente es independiente de la presencia/ausencia de otro valor. Se consideraron cuatro niveles de valores faltantes, 5 %, 10 %, 25 % y 35 % de la matriz de marcadores moleculares. Con este criterio de eliminación de la matriz de marcadores moleculares, la dimensión de la matriz X se mantiene constante, es decir, tiene la misma dimensión que la matriz de datos original completa.

4.3.3. Algoritmo NIPALS para PLS

El algoritmo NIPALS, introducido por H. Wold (1966) en el contexto del ACP, constituye un método iterativo para la extracción de componentes latentes a partir de matrices

de datos. Posteriormente, fue extendido al marco de la regresión PLS (S. Wold et al., 1983), donde su principal atractivo radica en la capacidad de optimizar simultáneamente la reducción de dimensionalidad y la predicción de las variables respuesta. En este sentido, mientras que en ACP los componentes se construyen únicamente en función de la varianza de X, en PLS los componentes se definen para maximizar la covarianza entre X e Y, lo que convierte a NIPALS en una herramienta particularmente útil en escenarios de predicción multivariada.

En el caso de matrices completas, el algoritmo procede de manera iterativa alternando entre el cálculo de pesos y cargas hasta alcanzar convergencia, seguido de una deflación de la matriz de datos. En el caso de matrices con datos faltantes, NIPALS presenta una ventaja clave respecto a otros enfoques, ya que puede extraer CP o construir variables latentes sin necesidad de realizar imputación previa. Esto se logra al ignorar las celdas ausentes en los productos escalares y normalizaciones, estimando provisionalmente los valores faltantes en cada iteración a partir de los componentes obtenidos, lo que puede interpretarse como una forma de imputación implícita dependiente del modelo.

Si bien este enfoque aprovecha toda la estructura observada en los datos y evita la pérdida de información asociada a la eliminación de registros incompletos, presenta algunas limitaciones: su desempeño puede degradarse cuando el porcentaje de datos faltantes es elevado o cuando los valores ausentes no cumplen el supuesto de encontrarse completamente al azar Nelson et al., 1996. Sin embargo, en contextos de alta dimensionalidad y fuerte correlación entre variables, como ocurre en estudios genómicos, sensometría o espectroscopía, NIPALS sigue siendo una alternativa robusta y computacionalmente eficiente frente a métodos de imputación fija como la media o la regresión múltiple S. Wold et al., 2001.

Los pasos de NIPALS son:

1. Inicializar $E = X$ y $F = Y$

2. Centrar y escalar las matrices.

3. Elegir una columna inicial de Y como u.

4. Iterar hasta la convergencia:

$$w \propto E'u$$

$$t \propto Ew$$

$$c \propto F't$$

$$u = Fc$$

5. Calcular el peso de regresión:

$$b = t'u$$

6. Actualizar las matrices E y F mediante deflación:

$$E = E - tp', \quad F = F - btc'$$

7. Repetir hasta alcanzar el número deseado de componentes.

NIPALS con Datos Completos

El algoritmo NIPALS se utiliza para calcular los componentes latentes en regresión PLS y puede manejar conjuntos de datos con o sin valores faltantes. Para datos completos, el procedimiento sigue un ciclo iterativo (Figura 4.1) que se repite para cada componente latente (L). En cada componente (h), el algoritmo itera entre el cálculo de puntuaciones (t) y cargas (p y c), seguido por la deflación de las matrices X e Y una vez que se alcanza la convergencia.

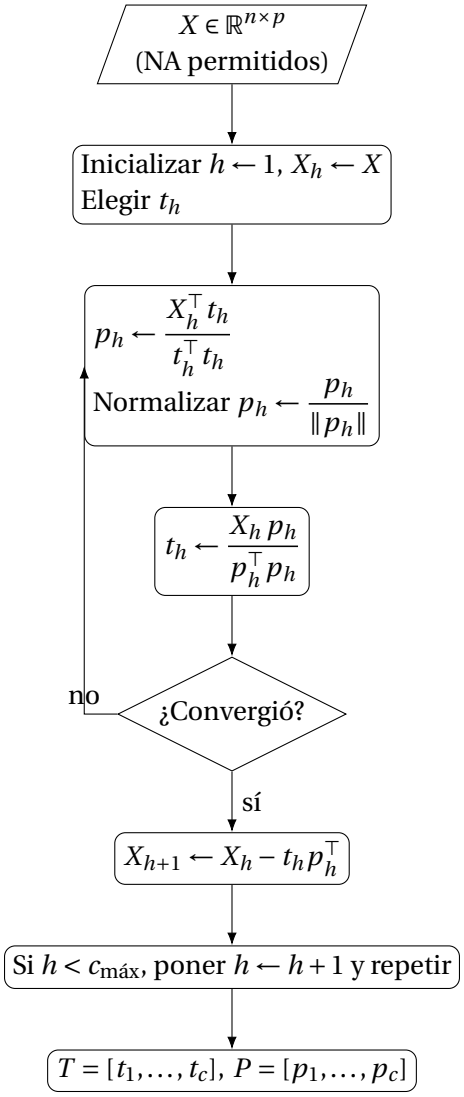

Figura 4.1: Esquema del flujo de NIPALS para datos completos

NIPALS con Datos Faltantes

Cuando existen valores faltantes en X y/o Y, el algoritmo NIPALS se adapta (Figura 4.2), modificando los cálculos de productos matriciales y normalizaciones para ignorar las celdas ausentes. La clave de este enfoque es la imputación dinámica: el algoritmo estima provisionalmente los valores faltantes en cada iteración a partir de la estructura latente descubierta. Formalmente, los valores ausentes en X se actualizan como $x_{ij} = t_i p_j$ si x_{ij} es un dato faltante. Este mecanismo evita la necesidad de imputación previa, pero introduce un desafío en la convergencia; las estimaciones pueden oscilar si el porcentaje de datos faltantes es muy elevado o si el patrón de datos faltantes no es completamente aleatorio debido a la dependencia recursiva entre la estimación de componentes y la actualización de valores faltantes.

En el algoritmo NIPALS, la presencia de datos faltantes se maneja mediante una imputación dinámica durante el proceso iterativo. En cada iteración, los pesos y las cargas se estiman utilizando únicamente las observaciones disponibles para cada variable, y los valores faltantes se actualizan de manera implícita a partir de las componentes calculadas hasta ese momento. Este mecanismo permite que el algoritmo aproveche toda la información disponible sin necesidad de imputar los valores antes de ajustar el modelo. Sin embargo, este procedimiento afecta directamente la convergencia del algoritmo: cuando el porcentaje de datos faltantes es bajo y su distribución es aleatoria, la imputación dinámica tiende a estabilizar las estimaciones y a reducir el número de iteraciones necesarias. En cambio, cuando los faltantes son numerosos o estructurados (por ejemplo, concentrados en un subconjunto de variables), las actualizaciones sucesivas pueden introducir oscilaciones en los valores estimados y retrasar la convergencia. Estas diferencias se reflejan en la Figura 4.2, donde se observa que la velocidad de convergencia depende tanto de la proporción de datos faltantes como de su patrón de distribución.

4.3.4. Evaluación del modelo mediante validación cruzada

Para la validación de los modelos se utilizó el mismo estimador calculado como se describió en el capítulo 3. A modo de repasar, se muestra la ecuación utilizada:

$$\text{RECM} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.1}$$

donde, y_i es el valor observado, $\hat{y}_i$ es el valor predicho por el modelo de regresión por PLS y n es el número de observaciones.

4.3.5. Análisis de concordancia mediante Procrustes

Se emplea un análisis de Procrustes para evaluar el grado de concordancia entre las matrices de valores predichos obtenidas a partir del modelo PLS completo (sin datos faltantes) y aquellas generadas por modelos construidos con distintos niveles de datos faltantes (5 %, 10 %, 25 % y 35 %). El análisis de Procrustes, originalmente introducido en el contexto de la psicometría y el análisis multivariado para comparar configuraciones de puntos mediante transformaciones rígidas, tales como rotaciones y escalados (Hurley y Cattell, 2007; Schönemann, 1966). Este enfoque permite cuantificar el nivel de similitud estructural entre las configuraciones multivariadas de predicción bajo diferentes condiciones de completitud de los datos, evaluando cómo las predicciones se distorsionan o se mantienen frente a la pérdida de información.

La comparación se realiza mediante una rotación, traslación y escalamiento óptimos de las matrices involucradas, de modo de minimizar la distancia entre ellas en el espacio euclídeo, de acuerdo con la formulación clásica del problema ortogonal de Procrustes (Schönemann, 1966). El estadístico resultante, conocido como discrepancia Procrustes, proporciona una medida global de desajuste entre configuraciones y ha sido ampliamente utilizado en estudios multivariados para evaluar estabilidad y concordancia entre representaciones obtenidas bajo distintos escenarios experimentales (Gower, 1975). El estadístico se define como:

$$D^2 = \min_{R,s,\mathbf{t}} \left\| \mathbf{C} - s\mathbf{F}R - \mathbf{1}\mathbf{t}' \right\|^2 \tag{4.2}$$

donde C es la configuración de referencia (modelo sin datos faltantes), F es la configuración ajustada (modelo con datos faltantes), R es una matriz de rotación ortogonal, s es un factor de escala y t es un vector de traslación.

Valores de D^2 cercanos a cero indican un alto grado de similitud entre ambas configuraciones, mientras que valores más elevados reflejan una mayor distorsión debido a la presencia de valores faltantes.

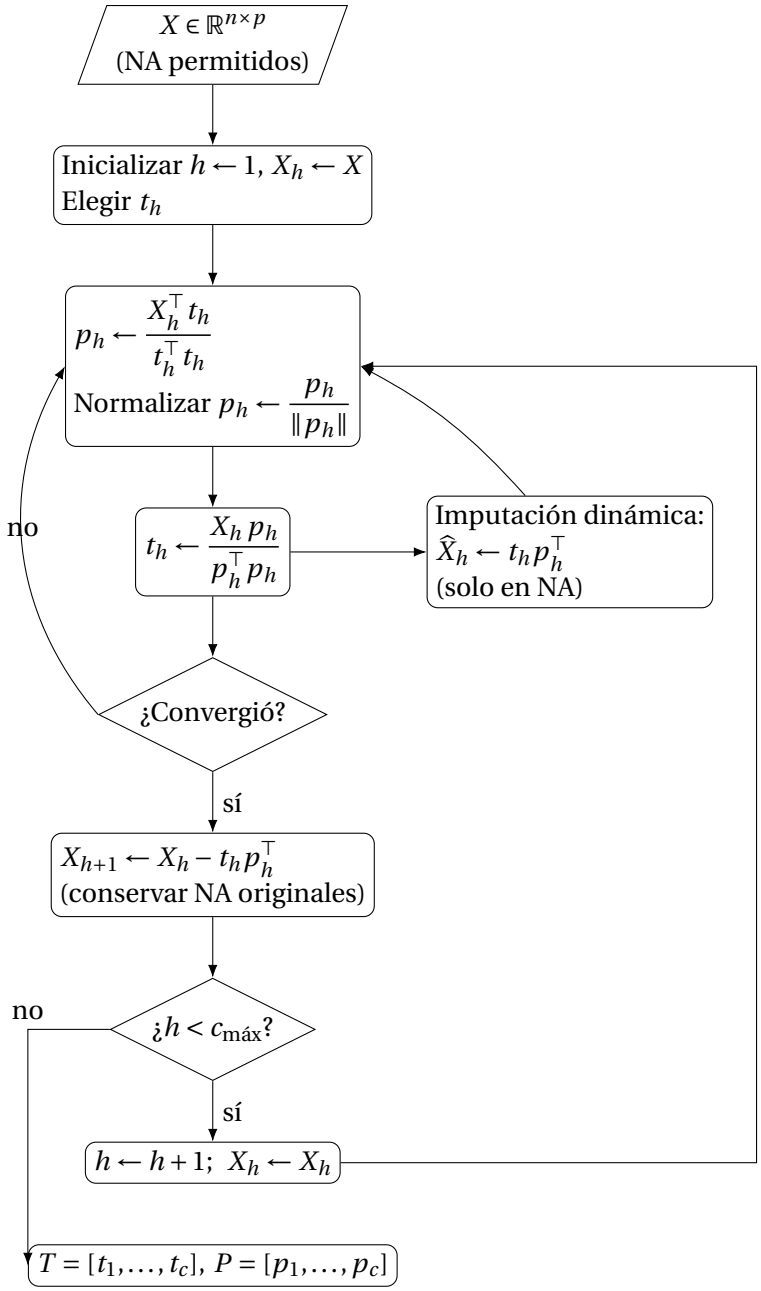

Figura 4.2: Esquema del flujo de NIPALS con datos faltantes. Los productos/normalizaciones omiten celdas NA; la imputación dinámica $\hat{X}_h$ puede usarse para estabilizar la convergencia

4.4. Resultados

4.4.1. Resultados obtenidos en los diferentes escenarios de datos faltantes para la base de datos simulada

El carácter fenotípico (variable Y) simulado en el Ambiente 2 presentó valores de la distribución del error mayores que en el caso del Ambiente 1 (Figura 4.3). Asimismo, se observó que el menor error se obtiene bajo el escenario sin datos faltantes, incrementándose progresivamente a medida que aumenta el porcentaje de valores ausentes (Figura 4.3). Este comportamiento resulta esperable, sin embargo es importante destacar que el aumento en la media de los valores del error es leve, encontrándose los valores de la mediana entre 0,9 y 1 para el caracter medido en el Ambiente 1 y entre 1 y 1,1 para el caracter medido en el Ambiente 2. En el caso de la variabilidad, no se observaron grandes diferencias entre los diferentes niveles de valores faltantes.

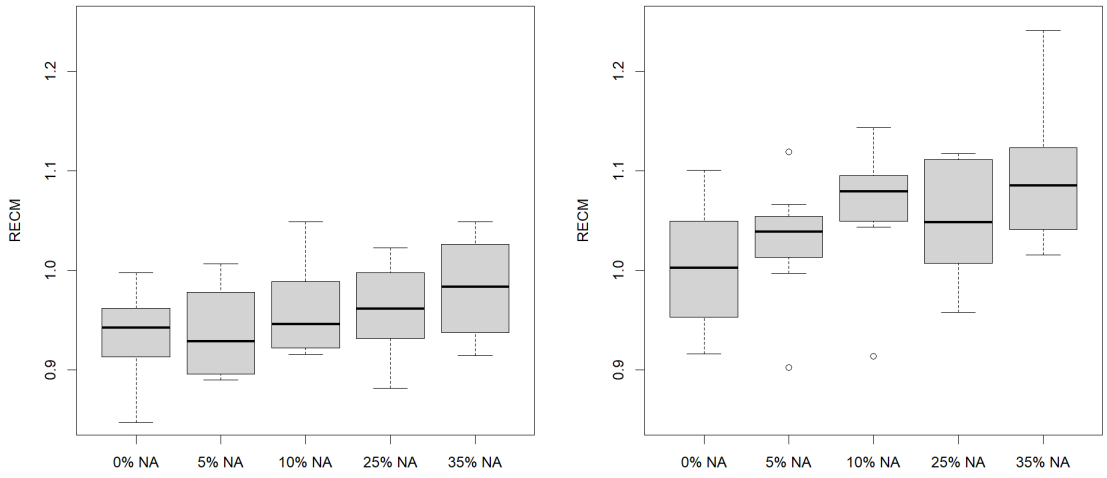

Figura 4.3: Distribución de la raíz del error cuadrático medio (RECM) entre los valores observados y los predichos según una validación cruzada para los caracteres: rendimiento en 2010-2011 (izquierda arriba), rendimiento en 2011-2012 (izquierda abajo), cantidad de frutos por planta en 2010-2011 (centro arriba) y en 2011-2012 (centro abajo) y, para el peso en 2010-2011 (derecha arriba) y en 2011-2012 (derecha abajo)

La cantidad de marcadores moleculares significativos encontrados en cada escenario es similar, siendo para el caracter medido en el primer ambiente, el escenario de 25 % de valores faltantes el que encontró mayor cantidad, mientras que en el caso del caracter

medido en el segundo ambiente el escenario con mayor cantidad de marcadores moleculares significativos se encontró en 35 % de valores faltantes (Tabla 4.1). Las métricas observadas en la matriz de confusión entre los marcadores significativos y los cercanos a algún QTL (a una distancia menor de 5cM), también capturan el mismo patrón que el observado anteriormente con la distribución del RECM, es decir, las métricas tienden a empeorar gradual y levemente entre el escenario sin valores faltantes y los escenarios de 5, 10, 25 y 35 % de valores faltantes (Tablas 4.2 y 4.3).

Tabla 4.1: Cantidad de marcadores moleculares significaivos en cada escenario: sin valores faltantes, y con 5 %, 10 %, 25 % y 35 % de valores faltantes

	0 %	5 %	10 %	25 %	35 %
F1	738	754	714	799	756
F2	774	794	779	773	832

Tabla 4.2: Métricas sobre las matrices de confusión entre los SNP estadísticamente significativos para el caracter fenotipico en el Ambiente 1

Métricas	0 %	5 %	10 %	25 %	35 %
Precisión	0,8401	0,8183	0,9132	0,7860	0,8399
Sensibilidad	0,0093	0,0093	0,0098	0,0095	0,0096
F1-score	0,0183	0,0184	0,0194	0,0187	0,0189
Especificidad	0,9913	0,9898	0,9954	0,9873	0,9910
Exactitud	0,1751	0,1748	0,1762	0,1745	0,1752

Tabla 4.3: Métricas sobre las matrices de confusión entre los SNP estadísticamente significativos para el caracter fenotipico en el Ambiente 2

Métricas	0%	5%	10%	25%	35%
Precisión	0,8140	0,7796	0,8074	0,8137	0,7452
Sensibilidad	0,0095	0,0093	0,0095	0,0095	0,0093
F1-score	0,0187	0,0184	0,0187	0,0187	0,0184
Especificidad	0,9893	0,9870	0,9889	0,9893	0,9843
Exactitud	0,1749	0,1744	0,1748	0,1749	0,1739

El análisis Procrustes mostró una alta concordancia entre la configuración obtenida con datos completos y aquellas derivadas de conjuntos con valores faltantes. La superposición entre ambas configuraciones sugiere que la estructura estimada por PLS se mantiene estable aun en presencia de datos incompletos. Si bien se observa un leve aumento en las discrepancias a medida que se incrementa el porcentaje de valores faltantes, la forma global de la configuración se conserva, indicando una adecuada robustez del método frente a la pérdida de información (Figura 4.4). Además, los valores p obtenidos entre los consensos de todos los escenarios resultó <0,0001, es decir, que existe consenso entre los predichos por el modelo con datos completos, y los modelos predichos en cada escenarios, simulando valores faltantes.

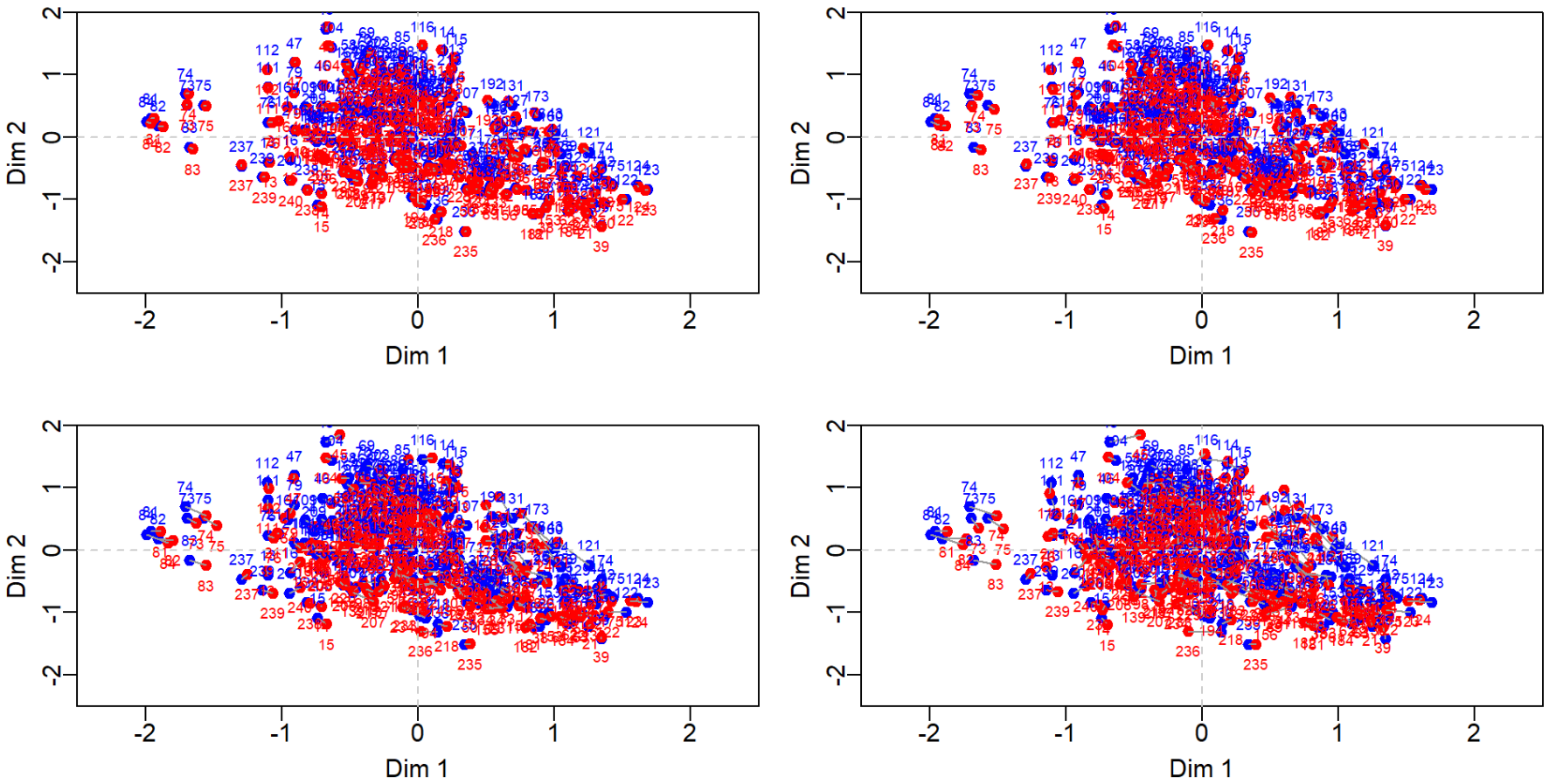

Figura 4.4: Gráfico de dispersión a partir del Análisis de Procrustes del consenso entre el escenario completo y los cuatro niveles de valores faltantes: 5 % (izquierda arriba), 10 % (izquierda abajo), 25 % (derecha arriba) y 35 % (derecha abajo). Predichos por PLS sin datos faltantes (azul) y predichos por PLS con datos faltantes (rojo)

4.4.2. Resultados obtenidos en los diferentes escenarios de datos faltantes para la base de datos de Duraznero

A partir de la matriz X de datos completos se generaron los cuatro escenarios de valores faltantes (5 %, 10 %, 25 % y 35 %). En cada caso la regresión PLS se calculó con tres variables latentes como se determinó en la sección Resultados del Capítulo 3. Al observar los valores de RECM (Figura 4.5), se encontró que los valores más chicos de la mediana se encuentran para el caso de datos completos, excepto en el caso del peso, que pareciera observarse un bajo valor de mediana con 25 % de valores faltantes. Siguiendo con el análisis, en la técnica de Procrustes (Figura 4.6) se observó que a medida que aumenta el porcentaje de valores faltantes se aleja el consenso entre ambas técnicas, es decir, el escenario de datos completos y los escenarios de valores faltantes. Sin embargo, para todos los casos, el valor p del consenso resultó bajo (0,0001), es decir, que el consenso es significativo en todos los escenarios de datos faltantes.

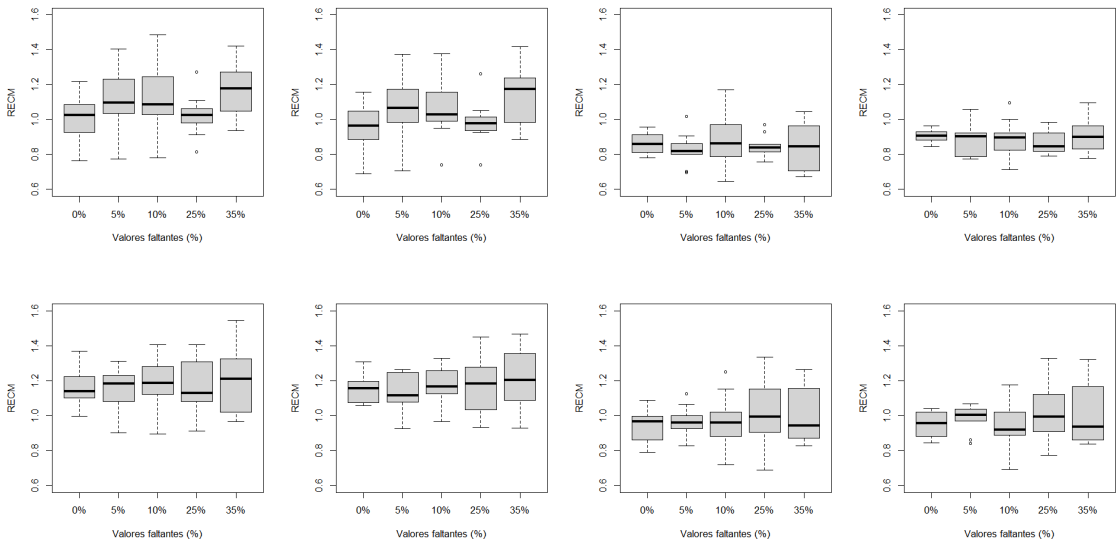

Figura 4.5: Distribución de la raíz del error cuadrático medio (RECM) entre los caracteres fenotípicos observados y los predichos según una validación cruzada para el rendimiento en el ambiente 2010-2011 (izquierda arriba), en el ambiente 2011-2012 (izquierda abajo), para la cantidad de frutos por planta en el ambiente 2010-2011 (centro izquierda arriba), en el ambiente 2011-2012 (centro izquierda abajo), para el peso en el ambiente 2010-2011 (centro derecha arriba) y en el ambiente 2011-2012 (centro derecha abajo) y para el calibre promedio en el ambiente 2010-2011 (derecha arriba) y en el ambiente 2011-2012 (derecha abajo)

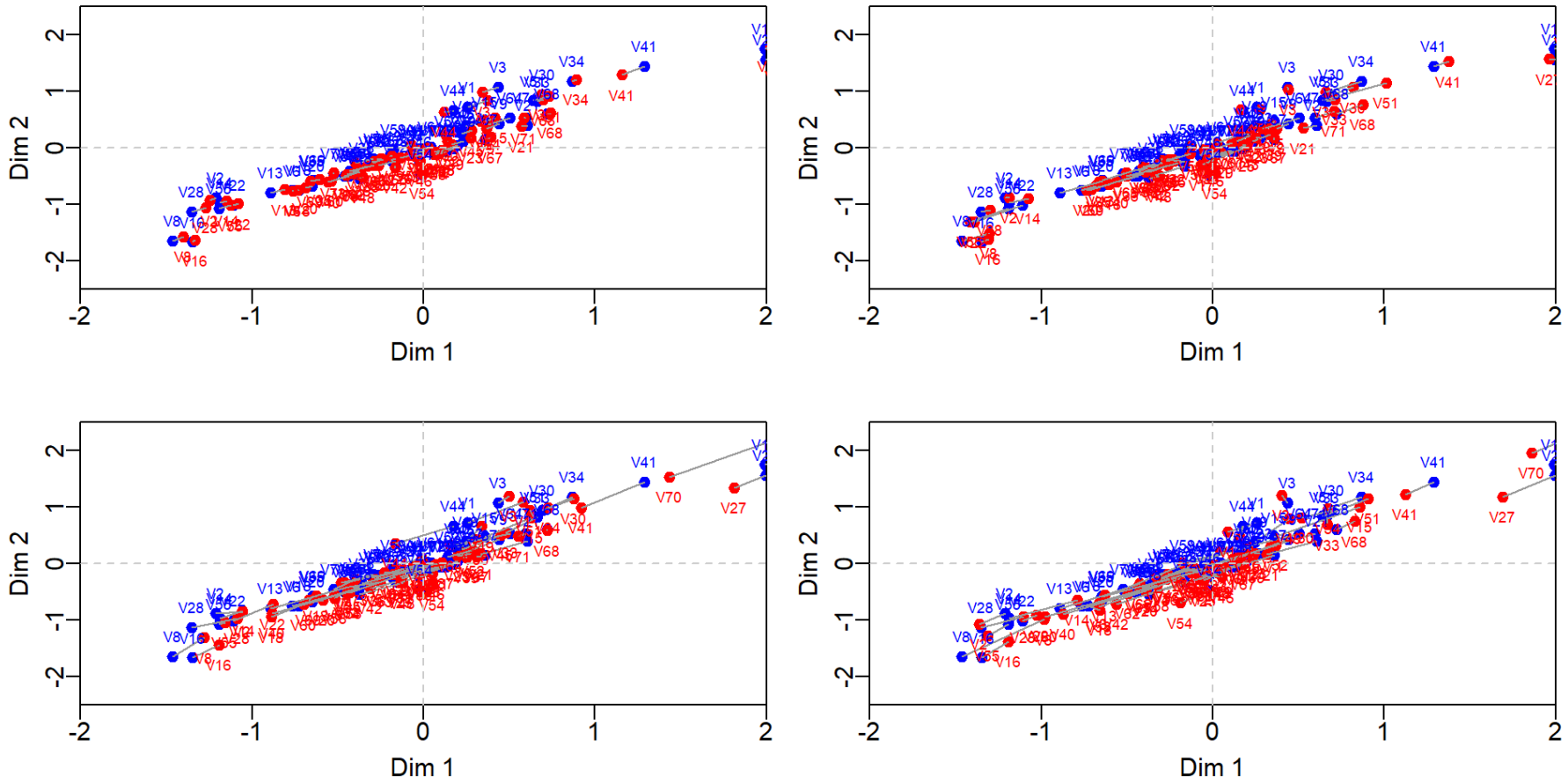

Figura 4.6: Gráfico de dispersión a partir del Análisis de Procrustes del consenso entre el escenario completo y los cuatro niveles de valores faltantes: 5 % (izquierda arriba), 10 % (izquierda abajo), 25 % (derecha arriba) y 35 % (derecha abajo). Predichos por PLS sin datos faltantes (azul) y predichos por PLS con datos faltantes (rojo)

Al comparar los marcadores significativos detectados en el escenario sin datos faltantes (0 %) con aquellos obtenidos bajo distintos niveles de datos faltantes (5 %, 10 %, 25 % y 35 %), se observó una disminución progresiva en el porcentaje de coincidencias a medida que aumenta la proporción de datos faltantes (Tabla 4.4). No obstante, la magnitud de esta disminución varía entre caracteres. En particular, el rendimiento y la cantidad de frutos por plantas en ambas campañas analizadas y el calibre ponderado en la campaña 2011-2012, muestran una mayor estabilidad, manteniendo porcentajes de coincidencia incluso en escenarios con mayor proporción de datos faltantes. En contraste, el peso promedio del fruto presenta una mayor sensibilidad marcada por una caída en el porcentaje de coincidencia de marcadores en los escenarios con 25 % y 35 % de datos faltantes, esto se debe a que fue el caracter con menor cantidad de marcadores moleculares significativos encontrados en el escenario sin datos faltantes. Finalmente el calibre ponderado en la campaña 2010-2011 no mostró ningún marcador molecular significativo por lo tanto no fue posible calcular el porcentaje de coincidencias.

Tabla 4.4: Cantidad de marcadores moleculares (MM) significativos en la base de datos completa y porcentaje de coincidencias entre los marcadores encontrados en cada escenario y el escenario original

Variable fenotípica y campaña agrícola	MM sign	5 %	10 %	25 %	35 %
R10–11	25	25 88,00 %	24 72,00 %	24 72,00 %	21 52,00 %
C10–11	22	23 86,36 %	21 72,73 %	23 77,27 %	21 54,55 %
P10–11	1	1 100.00 %	1 100,00 %	5 100,00 %	4 0,00 %
Ca10–11	0	1 –	2 –	3 –	1 –
R11–12	5	3 60,00 %	4 80,00 %	5 60,00 %	6 60,00 %
C11–12	4	4 100,00 %	8 75,00 %	7 75,00 %	7 75,00 %
P11–12	4	4 100.00 %	5 75,00 %	7 75,00 %	8 75,00 %
Ca11–12	10	10 90,00 %	15 90,00 %	12 80,00 %	6 50,00 %

4.4.3. Resultados obtenidos en los diferentes escenarios de datos faltantes para la base de datos de Maíz

La regresión PLS se calculó con tres variables latentes como se determinó en la sección Resultados del Capítulo 3. Se visualizaron los diferentes valores de RECM para la validación cruzada según variable y ambiente (Figura 4.7). El incremento en la proporción de valores faltantes produce, en general, un aumento de la RECM, aunque este efecto es gradual y depende del carácter y del ambiente. En particular, hasta niveles intermedios de datos faltantes, el método PLS conserva una capacidad predictiva relativamente estable, lo que sugiere una adecuada robustez frente a información incompleta. En el caso de las variables FFL y MFL en el Ambiente WW, la variabilidad de los datos tiende a incrementarse para todos los escenarios, siendo los porcentajes 10 y 35 % los

que presentan mayor variabilidad.

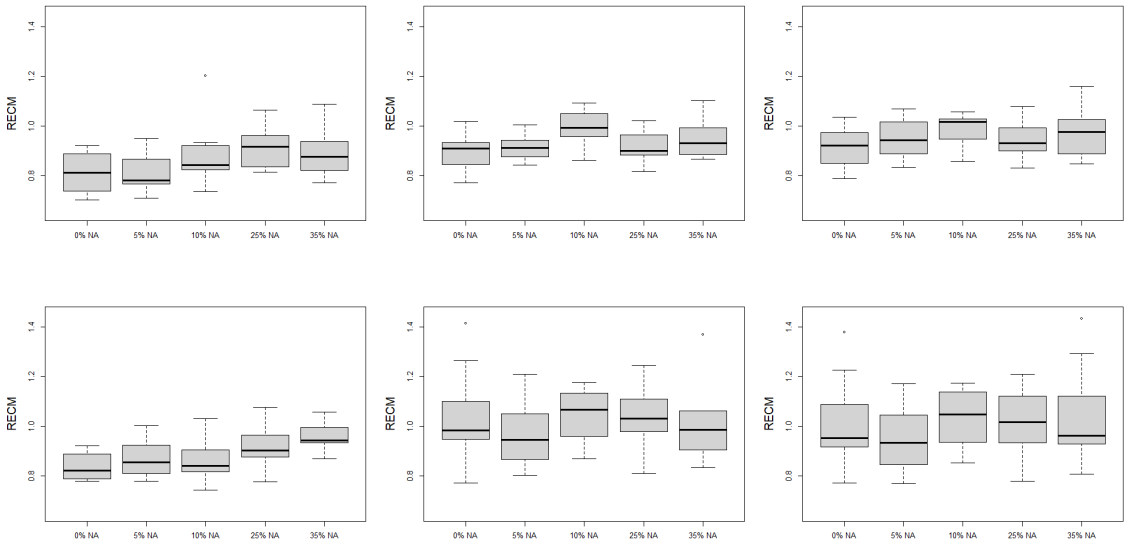

Figura 4.7: Distribución de la raíz del error cuadrático medio (RECM) de los valores de los caracteres fenotípicos reales vs los predichos por el análisis PLS según validación cruzada en cada escenario de valores faltantes para las variables ASI en el Ambiente SS (izquierda arriba), en el Ambiente WW (izquierda abajo), para la variable FFL en el Ambiente SS (centro arriba), en el Ambiente WW (centro abajo), para la variable MFL en el Ambiente SS (derecha arriba) y en el Ambiente WW (derecha abajo)

Con el Análisis de Procrustes (Figura 4.8) y con las coincidencias de marcadores entre la base de datos completa y los escenarios con datos faltantes (Tabla 4.5) se observó que a medida que aumentan los valores faltantes se pierde el consenso entre los marcadores significativos que se encuentran en cada caso.

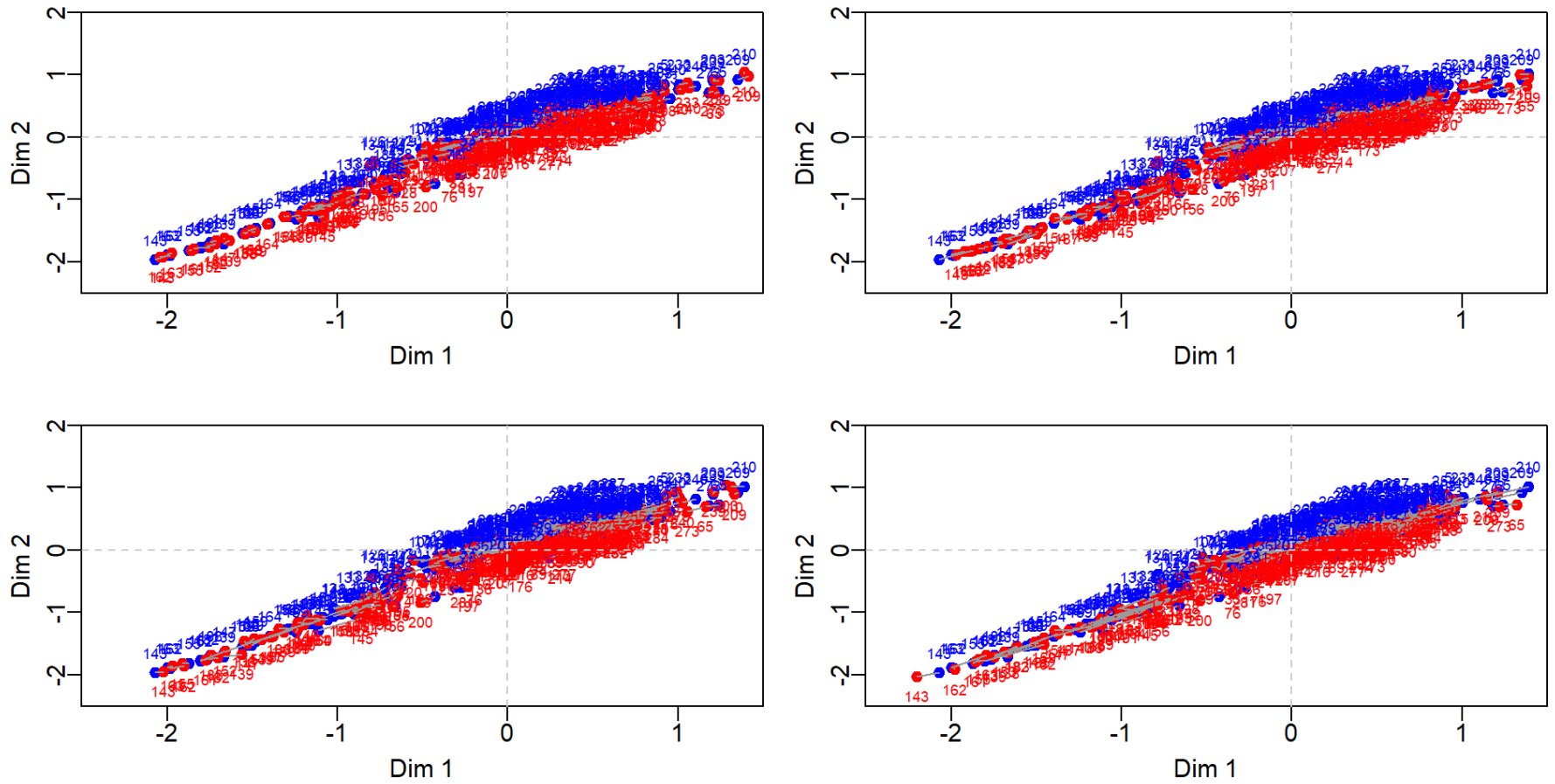

Figura 4.8: Gráfico de dispersión a partir del Análisis de Procrustes del consenso entre el escenario completo y los cuatro niveles de valores faltantes: 5 % (izquierda arriba), 10 % (izquierda abajo), 25 % (derecha arriba) y 35 % (derecha abajo). Predichos por PLS sin datos faltantes (azul) y predichos por PLS con datos faltantes (rojo)

El patrón resultó consistente con lo encontrado previamente en la base de datos de durazno, donde el incremento en la proporción de valores faltantes conduce a una reducción progresiva en la coincidencia de marcadores significativos respecto del escenario completo. Sin embargo, la magnitud de esta reducción varía entre variables y ambientes, lo que sugiere que la estabilidad de la detección de asociaciones depende tanto del método de análisis como del rasgo evaluado. En general, los resultados indican que, si bien la presencia de datos faltantes afecta la selección de marcadores, ciertos conjuntos de variables muestran una mayor robustez frente a la incompletitud de los datos. Los valores p entre los consensos en todos los escenarios de datos faltantes, resultaron $\leq 0,05$.

Tabla 4.5: Cantidad de marcadores moleculares significativos para el escenario de datos completos y el porcentaje de coincidencias entre estos marcadores y cada escenario de valores faltantes: 5 %, 10 %, 25 % y 35 %

Variable fenotípica y condición ambiental	MM Sign	5 %	10 %	25 %	35 %
ASI SS	17	14 76,47 %	14 76,47 %	15 70,59 %	12 52,94 %
ASI WW	15	15 93,33 %	16 93,33 %	14 80,00 %	14 53,33 %
FFL SS	13	12 92,31 %	13 84,63 %	14 69,23 %	10 38,46 %
FFL WW	11	12 81,82 %	14 72,73 %	13 81,82 %	9 36,36 %
MFL SS	10	12 80,00 %	14 80,00 %	13 60,00 %	9 50,00 %
MFL WW	12	11 83,33 %	11 66,67 %	17 50,00 %	11 41,67 %

4.5. Discusión

Los resultados presentados en este capítulo ponen de manifiesto la robustez del algoritmo NIPALS frente a la presencia de datos faltantes en el contexto de modelos PLS

aplicados a estudios de asociación fenotipo-genotipo. A diferencia de otras estrategias que requieren imputación o eliminación de observaciones, NIPALS permite mantener la integridad de la matriz de datos y aprovechar la totalidad de la información disponible.

Se observó que, a medida que aumenta la proporción de datos faltantes, el error de predicción (medido mediante RECM) también se incrementa, lo cual era esperable. Sin embargo, el aumento fue gradual y no condujo a una degradación drástica del rendimiento predictivo. Este comportamiento sugiere que el modelo PLS-NIPALS conserva una buena capacidad de generalización incluso en condiciones de datos incompletos, una propiedad deseable en estudios genómicos donde la información puede estar fragmentada por cuestiones técnicas o logísticas. En los tres conjuntos de datos analizados (simulados, durazno y maíz), se mantuvo una alta concordancia entre las estructuras latentes obtenidas con y sin datos faltantes, como se evidenció mediante los análisis Procrustes. Este hallazgo refuerza la hipótesis de que NIPALS logra preservar las relaciones subyacentes en los datos, incluso en contextos de pérdida parcial de información. Asimismo, la identificación de SNP relevantes se mantuvo estable hasta niveles de 25 %–35 % de datos ausentes, con tasas de coincidencia aceptables respecto al modelo completo.

Los resultados del análisis Procrustes indican que la estructura del espacio latente estimado mediante NIPALS se preserva en mayor medida frente a la introducción de datos faltantes. En particular, los menores valores de discrepancia observados sugieren que las relaciones geométricas entre variedades y variables se mantienen relativamente estables. Desde el punto de vista aplicado, esta propiedad es especialmente relevante en estudios GWAS, donde la interpretación biológica depende fuertemente de la estabilidad de la configuración multivariada.

No obstante, es importante señalar que, aunque NIPALS mostró un rendimiento sólido, su desempeño podría depender de la estructura de correlaciones en los datos y de la naturaleza del mecanismo de ausencia. En este estudio se asumió que los datos faltantes se presentan de forma aleatoria en la base de datos, pero en aplicaciones reales este supuesto puede no cumplirse, lo cual podría impactar la estabilidad del modelo.

En conjunto, los resultados sugieren que el algoritmo NIPALS representa una alternativa metodológica eficiente, precisa y computacionalmente viable para el análisis

multivariado en contextos genómicos con datos faltantes. Su implementación permite evitar pasos intermedios de imputación y preservar la estructura de los datos, facilitando análisis robustos incluso en presencia de información incompleta.

En este estudio no se utilizó imputación previa, pero es importante considerar que en otros enfoques los datos faltantes suelen completarse mediante métodos como la imputación por kNN (*k-nearest neighbors*) o el algoritmo EM (*Expectation-Maximizatio*n). Si bien estas estrategias permiten aplicar modelos convencionales, presentan limitaciones. Por ejemplo, la imputación por la media tiende a subestimar la variabilidad y puede distorsionar relaciones entre variables, mientras que kNN y EM requieren mayores recursos computacionales y pueden ser sensibles a la estructura del conjunto de datos. NIPALS, al evitar la imputación explícita, reduce estos riesgos y ofrece un tratamiento más directo de la información disponible.

La utilidad práctica de NIPALS ha sido ampliamente reportada en la literatura científica. En estudios genéticos, ha sido aplicado en contextos de predicción genómica y análisis multibloque, como los realizados por Montesinos-López et al., 2022 en predicción genómica en trigo y por Mehmood et al., 2012 en espectroscopía. En ambos casos, se destacaron sus capacidades de manejo de alta dimensionalidad y datos incompletos, además de su flexibilidad para integrarse en flujos de análisis supervisados. Esta evidencia respalda su aplicación en la presente tesis y sugiere que puede ser considerada una herramienta confiable en futuros análisis genómicos con datos reales.

4.6. Conclusión

Los resultados obtenidos a lo largo de este capítulo confirman la utilidad del algoritmo NIPALS como una herramienta eficaz para realizar regresión PLS en contextos con datos faltantes. A pesar del aumento progresivo del error de predicción (RECM) con niveles crecientes de datos ausentes, el modelo mantiene una capacidad aceptable para identificar marcadores moleculares asociados a los caracteres fenotípicos analizados.

El análisis de las tres bases de datos permitió verificar que la pérdida de información no afecta de manera crítica la estabilidad de las componentes latentes extraídas, ni la estructura general de relaciones entre predictores y respuestas. La comparación me-

diante análisis Procrustes reforzó esta conclusión, mostrando altos grados de similitud entre los modelos completos y los incompletos.

Estos hallazgos sugieren que, en estudios genómicos donde los datos faltantes son inevitables, el uso de NIPALS como alternativa a métodos de imputación o eliminación de observaciones permite conservar información relevante y obtener resultados robustos. Este enfoque resulta particularmente valioso en entornos con alta dimensionalidad, como los que se presentan en el análisis de SNP en cultivos agrícolas.

En conjunto, la evidencia empírica presentada en este capítulo indica que el algoritmo NIPALS ofrece ventajas claras en escenarios con datos faltantes, tanto en términos de estabilidad predictiva como de preservación de la estructura latente. Estos resultados respaldan su utilización como estrategia preferente para la estimación de modelos PLS en estudios GWAS con matrices incompletas, aportando evidencia sistemática que complementa y amplía la comparación algorítmica desarrollada en el capítulo anterior.

Capítulo 5

Conclusiones Finales

5.1. Síntesis de resultados principales

En esta tesis se abordó la aplicación de la técnica de PLS en el contexto de estudios GWAS en plantas, con foco en el análisis de diferentes algoritmos y en la evaluación de escenarios con datos faltantes.

En el Capítulo 2, se evaluaron diferentes modelos en contexto GWAS, que buscaban una relación entre fenotipos y genotipos en contextos donde existe heterocedasticidad entre ambientes, destacándose el modelo MLMM por su capacidad para separar señal genética del ruido ambiental. También se encontró evidencia de la importancia del modelado de la variabilidad ambiental previo al ajuste de los modelos GWAS.

En el Capítulo 3, se compararon los algoritmos SVD y NIPALS para la implementación de PLS, tanto en datos simulados como en bases reales de duraznero y maíz. Los resultados indican que, aunque ambos algoritmos son capaces de detectar asociaciones relevantes, NIPALS presenta ventajas operativas en contextos realistas donde la presencia de datos incompletos es frecuente en estudios genómicos.

En el Capítulo 4, se evaluó específicamente la robustez del algoritmo NIPALS frente a distintos porcentajes de datos faltantes (5 %, 10 %, 25 % y 35 %). Los resultados indican que, si bien el error de predicción tiende a incrementarse a medida que aumenta el porcentaje de datos ausentes, el método mantiene la capacidad de identificar asociaciones fenotipo–genotipo consistentes, incluso en escenarios con información incompleta.

5.2. Contribuciones de la tesis

Las principales contribuciones de este trabajo pueden resumirse en tres niveles:

- Metodológico: se generó evidencia empírica sobre el desempeño de diversos

modelos GWAS bajo condiciones de heterocedasticidad y sobre la técnica PLS con datos faltantes; escenarios poco explorados de manera conjunta en la literatura GWAS.

- Aplicado: se validó el uso de PLS en bases de datos reales de duraznero y maíz, mostrando su utilidad en la identificación de marcadores moleculares, en contextos de alta dimensionalidad y correlación entre marcadores.

- Práctico: se generaron e implementaron rutinas en software libre y herramientas de visualización que garantizan la transparencia y facilitan la aplicación de estos métodos en estudios de fitomejoramiento, promoviendo prácticas de investigación reproducible en estudios de genética cuantitativa.

5.3. Trabajos futuros

Incorporar técnicas avanzadas de imputación de datos o algoritmos que manejen datos faltantes de manera más eficiente. Explorar otras variantes de PLS, como *sparse PLS* o *kernel PLS*, que podrían mejorar la detección de asociaciones en escenarios de alta dimensionalidad. Extender el análisis a bases de datos de mayor escala, tanto en cantidad de individuos como en número de caracteres fenotípicos y ambientes.

5.4. Conclusión final

En conjunto, esta tesis aporta evidencia de que los métodos de Mínimos Cuadrados Parciales constituyen una alternativa metodológica sólida para estudios GWAS en plantas, particularmente en escenarios donde la alta dimensionalidad, la correlación entre marcadores y la presencia de datos faltantes representan desafíos para enfoques tradicionales.

Los resultados obtenidos sugieren que el algoritmo NIPALS no solo ofrece ventajas computacionales, sino también una mayor adaptabilidad a estructuras de datos incompletas, situación frecuente en estudios genómicos aplicados al fitomejoramiento. En

este sentido, el uso de PLS puede contribuir a mejorar la identificación de asociaciones fenotipo-genotipo sin requerir supuestos estrictos sobre la estructura de los datos.

De este modo, esta tesis contribuye a fortalecer el uso de enfoques multivariados en genética cuantitativa vegetal y aporta evidencia que puede guiar la selección de metodologías en estudios futuros.

El código completo utilizado en los análisis está disponible en GitHub **bortolotto2025codigo**.

Bibliografía

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *WIREs Computational Statistics*, *2*(1), 97-106. https://doi.org/10.1002/wics.51

Abdurakhmonov, I. Y., & Abdukarimov, A. (2008). Application of Association Mapping to Understanding the Genetic Diversity of Plant Germplasm Resources [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2008/574927]. *International Journal of Plant Genomics*, *2008*(1), 574927. https://doi.org/10.1155/2008/574927

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Al Kawam, A., Alshawaqfeh, M., Cai, J. J., Serpedin, E., & Datta, A. (2018). Simulating variance heterogeneity in quantitative genome wide association studies. *BMC bioinformatics*, *19*(Suppl 3), 72. https://doi.org/10.1186/s12859-018-2061-1

Andrade, A. C. B., Viana, J. M. S., Pereira, H. D., Pinto, V. B., & Fonseca E Silva, F. (2019). Linkage disequilibrium and haplotype block patterns in popcorn populations (A. Kumar, Ed.). *PLOS ONE*, *14*(9), e0219417. https://doi.org/10.1371/journal.pone.0219417

Araus, J. L., Kefauver, S. C., Zaman-Allah, M., Olsen, M. S., & Cairns, J. E. (2018). Translating High-Throughput Phenotyping into Genetic Gain. *Trends in Plant Science*, *23*(5), 451-466. https://doi.org/10.1016/j.tplants.2018.02.001

Ardlie, K. G., Lunetta, K. L., & Seielstad, M. (2002). Testing for population subdivision and association in four case-control studies. *American Journal of Human Genetics*, *71*(2), 304-311. https://doi.org/10.1086/341719

Astle, W., & Balding, D. J. (2009). Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science*, *24*(4). https://doi.org/10.1214/09-STS307

Balzarini, M. (2002, enero). Applications of mixed models in plant breeding. En M. S. Kang (Ed.), *Quantitative genetics, genomics and plant breeding* (1.ª ed., pp. 353-363). CABI Publishing. https://doi.org/10.1079/9780851996011.0353

Becker, D., Wimmers, K., Luther, H., Hofer, A., & Leeb, T. (2013). A Genome-Wide Association Study to Detect QTL for Commercially Important Traits in Swiss Large White Boars (S. Moore, Ed.). *PLoS ONE*, *8*(2), e55951. https://doi.org/10.1371/journal.pone.0055951

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289-300. Consultado el 21 de diciembre de 2025, desde http://www.jstor.org/stable/2346101

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, *8*, 3-62.

bortolottoeugenia-gif. (2025, diciembre). bortolottoeugenia-gif/codigo-tesis-bortolotto. Consultado el 10 de diciembre de 2025, desde https://github.com/bortolottoeugenia-gif/codigo-tesis-bortolotto

Bouchet, S., Bertin, P., Presterl, T., Jamin, P., Coubriche, D., Gouesnard, B., Laborde, J., & Charcosset, A. (2017). Association mapping for phenology and plant architecture in maize shows higher power for developmental traits compared with growth influenced traits. *Heredity*, *118*(3), 249-259. https://doi.org/10.1038/hdy.2016.88

Boulesteix, A.-L., & Strimmer, K. (2006). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, *8*(1), 32-44. https://doi.org/10.1093/bib/bbl016

Boulesteix, A.-L., & Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, *8*(1), 32-44. https://doi.org/10.1093/bib/bbl016

Boyce, A. J. (1983). Computation of inbreeding and kinship coefficients on extended pedigrees. *Journal of Heredity*, *74*(6), 400-404. https://doi.org/10.1093/oxfordjournals.jhered.a109825

Brachi, B., Faure, N., Horton, M., Flahauw, E., Vazquez, A., Nordborg, M., Bergelson, J., Cuguen, J., & Roux, F. (2010). Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature (T. F. C. Mackay, Ed.). *PLoS Genetics*, *6*(5), e1000940. https://doi.org/10.1371/journal.pgen.1000940

Brachi, B., Morris, G. P., & Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, *12*(10), 232. https://doi.org/10.1186/gb-2011-12-10-232

Braz, C. U., Rowan, T. N., Schnabel, R. D., & Decker, J. E. (2021). Genome-wide association analyses identify genotype-by-environment interactions of growth traits in Simmental cattle. *Scientific Reports*, *11*(1), 13335. https://doi.org/10.1038/s41598-021-92455-x

Broc, C., Truong, T., & Liquet, B. (2021). Penalized partial least squares for pleiotropy. *BMC Bioinformatics*, *22*(1), 86. https://doi.org/10.1186/s12859-021-03968-1

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

Browning, S. R. (2008). Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, *124*(5), 439-450. https://doi.org/10.1007/s00439-008-0568-7

Bruno, C., Videla, M., & Balzarini, M. (2019). TEST OF INTERACTION IN THE ANALYSIS OF MOLECULAR VARIANCE. *Journal of Basic and Applied Genetics*, *30*(1), 17-23. https://doi.org/10.35407/bag.2019.XXX.01.03

Buckler, E. S., & Thornsberry, J. M. (2002). Plant molecular diversity and applications to genomics. *Current Opinion in Plant Biology*, *5*(2), 107-111. https://doi.org/10.1016/s1369-5266(02)00238-8

Buuren, S. V., & Groothuis-Oudshoorn, K. (2011). **mice** : Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, *45*(3). https://doi.org/10.18637/jss.v045.i03

Čepica, S., Zambonelli, P., Weisz, F., Bigi, M., Knoll, A., Vykoukalová, Z., Masopust, M., Gallo, M., Buttazzoni, L., & Davoli, R. (2013). Association mapping of quantitative trait loci for carcass and meat quality traits at the central part of chromosome 2 in Italian Large White pigs. *Meat Science*, *95*(2), 368-375. https://doi.org/10.1016/j.meatsci.2013.05.002

Cheng, H., Garrick, D., & Fernando, R. (2015). XSim: Simulation of Descendants from Ancestors with Sequence Data. *G3 Genes|Genomes|Genetics*, *5*(7), 1415-1417. https://doi.org/10.1534/g3.115.016683

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, *87*(1), 52-58. https://doi.org/10.1046/j.1365-2540.2001.00901.x

Christoffersson, A. (1970). *The One Component Model with Incomplete Data* [Tesis doctoral, University of Uppsala].

Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H., & Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism [Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press]. *Genome Research*, *15*(11), 1496-1502. https://doi.org/10.1101/gr.4107905

Coelho, C. A., & Rodrigues, A. M. (2012). *Linear Mixed Models: A Practical Guide Using Statistical Software* by Brady T. West, Kathleen B. Welch, and Andrzej T. Galecki: Chapmann & Hall/CRC Press (2006). *Journal of Statistical Theory and Practice*, *6*(3), 590-595. https://doi.org/10.1080/15598608.2012.695708

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, *142*(1), 169-196. https://doi.org/10.1007/s10681-005-1681-5

Colombani, C., Croiseau, P., Fritz, S., Guillaume, F., Legarra, A., Ducrocq, V., & Robert-Granié, C. (2012). A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. *Journal of Dairy Science*, *95*(4), 2120-2131. https://doi.org/10.3168/jds.2011-4647

Corty, R. W., & Valdar, W. (2018). QTL Mapping on a Background of Variance Heterogeneity. *G3 (Bethesda, Md.)*, *8*(12), 3767-3782. https://doi.org/10.1534/g3.118.200790

Covarrubias-Pazaran, G. (2018, junio). Software update: Moving the R package *sommer* to multivariate mixed models for genome-assisted prediction. https://doi.org/10.1101/354639

Crossa, J., Burgueño, J., Dreisigacker, S., Vargas, M., Herrera-Foessel, S. A., Lillemo, M., Singh, R. P., Trethowan, R., Warburton, M., Franco, J., Reynolds, M., Crouch, J. H., & Ortiz, R. (2007). Association Analysis of Historical Bread Wheat Germplasm Using Additive Genetic Covariance of Relatives and Population Structure. *Genetics*, *177*(3), 1889-1913. https://doi.org/10.1534/genetics.107.078659

Crossa, J., Campos, G. d. L., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., Makumbi, D., Singh, R. P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., & Braun, H.-J. (2010). Prediction of genetic values of quantitative traits in plant breeding using

pedigree and molecular markers. *Genetics*, *186*(2), 713-724. https://doi.org/10.1534/genetics.110.118521

Dai, T., Ban, S., Han, L., Li, L., Zhang, Y., Zhang, Y., & Zhu, W. (2024). Effects of exogenous glycine betaine on growth and development of tomato seedlings under cold stress. *Frontiers in Plant Science*, *15*, 1332583. https://doi.org/10.3389/fpls.2024.1332583

da Silva Linge, C., Cai, L., Fu, W., Clark, J., Worthington, M., Rawandoozi, Z., Byrne, D. H., & Gasic, K. (2021). Multi-Locus Genome-Wide Association Studies Reveal Fruit Quality Hotspots in Peach Genome. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.644799

De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, *18*(3), 251-263. https://doi.org/10.1016/0169-7439(93)85002-X

De Souza, V. C., Rodrigues, S. A., & Filho, L. R. A. G. (2024). Comparison of principal component analysis algorithms for imputation in agrometeorological data in high dimension and reduced sample size (S. Heddam, Ed.). *PLOS ONE*, *19*(12), e0315574. https://doi.org/10.1371/journal.pone.0315574

de Souto, M. C. P., Jaskowiak, P. A., & Costa, I. G. (2015). Impact of missing data imputation methods on gene expression clustering and classification. *BMC bioinformatics*, *16*, 64. https://doi.org/10.1186/s12859-015-0494-3

Dobriban, E. (2017). Permutation methods for factor analysis and PCA [Version Number: 3]. https://doi.org/10.48550/ARXIV.1710.00479
Other To appear in the Annals of Statistics.

Dunn, H. (1961). High-level wellness; a collection of twenty-nine short talks on different aspects of the theme "High-level wellness for man and society. *Scotty Book:Arlington, Va., R. W. Beatty Co*.

Ekine, C. C., Rowe, S. J., Bishop, S. C., & De Koning, D.-J. (2014). Why Breeding Values Estimated Using Familial Data Should Not Be Used for Genome-Wide Association Studies. *G3 Genes|Genomes|Genetics*, *4*(2), 341-347. https://doi.org/10.1534/g3.113.008706

Elad, M. (2010). *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer New York. https://doi.org/10.1007/978-1-4419-7011-4

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., & Mitchell, S. E. (2011). A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PloS One*, *6*(5), e19379. https://doi.org/10.1371/journal.pone.0019379

Erny, G. L., Brito, E., Pereira, A. B., Bento-Silva, A., Vaz Patto, M. C., & Bronze, M. R. (2021). Projection to latent correlative structures, a dimension reduction strategy for spectral-based classification. *RSC Advances*, *11*(47), 29124-29129. https://doi.org/10.1039/D1RA03359J

Esfandyari, H., & Sørensen, A. (2019). xbreed: An R package for genomic simulation of purebred and crossbred populations. Consultado el 9 de diciembre de 2025, desde https://rdrr.io/cran/xbreed/f/vignettes/xbreedvignette.Rmd

Falconer, D. S., & Mackay, T. F. (1996). *Introduction to Quantitative Genetics* (4th edition).

Fisher, R. A. (1935). *The design of experiments* [Pages: xi, 251]. Oliver & Boyd.

Fisher, R. A. (1936). THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. *Annals of Eugenics*, *7*(2), 179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Flint-Garcia, S., Thornsberry, J., & Buckler, E. (2003). Structure of Linkage Disequilibrium in Plants* | Annual Reviews. https://doi.org/https://doi.org/10.1146/annurev.arplant.54.031902.134907

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., & Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-313X.2005.02591.x]. *The Plant Journal*, *44*(6), 1054-1064. https://doi.org/10.1111/j.1365-313X.2005.02591.x

Florez, J. C., Manning, A. K., Dupuis, J., McAteer, J., Irenze, K., Gianniny, L., Mirel, D. B., Fox, C. S., Cupples, L. A., & Meigs, J. B. (2007). A 100K genome-wide association scan for diabetes and related traits in the Framingham Heart Study: replication and integration with other genome-wide datasets. *Diabetes*, *56*(12), 3063-3074. https://doi.org/10.2337/db07-0451

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*(3), 453-467. https://doi.org/10.1093/biomet/58.3.453

Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. *In V. Barnett (Ed.), Interpreting Multivariate Data.Chichester, UK: John Wiley and Sons*.

Gabriel, K. R. (1982). Encyclopedia of statistical sciences. 1: A to Circular probable error [Num Pages: 480]. En *Biplot*. Wiley.

Gabriel, K. R. (2002). Goodness of Fit of Biplots and Correspondence Analysis. *Biometrika*, *89*(2), 423-436. Consultado el 28 de diciembre de 2025, desde http://www.jstor.org/stable/4140587

Galindo-Villardón, M. P., & Blázquez-Zaballos, A. (s.f.). Logistic Biplots.

Gangurde, S. S., Khan, A. W., Janila, P., Variath, M. T., Manohar, S. S., Singam, P., Chitikineni, A., Varshney, R. K., & Pandey, M. K. (2023). Whole-genome sequencing based discovery of candidate genes and diagnostic markers for seed weight in groundnut. *The Plant Genome*, *16*(4), e20265. https://doi.org/10.1002/tpg2.20265

Garrick, D. J., & Fernando, R. L. (2013). Implementing a QTL Detection Study (GWAS) Using Genomic Prediction Methodology. En C. Gondro, J. van der Werf & B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction* (pp. 275-298). Humana Press. https://doi.org/10.1007/978-1-62703-447-0_11

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: an R package for breeding program simulations. *G3 Genes|Genomes|Genetics*, *11*(2), jkaa017. https://doi.org/10.1093/g3journal/jkaa017

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*, 1-17. https://doi.org/10.1016/0003-2670(86)80028-9

Gianola, D., & De Los Campos, G. (2008). Inferring genetic values for quantitative traits non-parametrically. *Genetics Research*, *90*(6), 525-540. https://doi.org/10.1017/S0016672308009890

Gianola, S., Frigerio, P., Agostini, M., Bolotta, R., Castellini, G., Corbetta, D., Gasparini, M., Gozzer, P., Guariento, E., Li, L. C., Pecoraro, V., Sirtori, V., Turolla, A., Andreano, A., & Moja, L. (2016). Completeness of Outcomes Description Reported in Low Back Pain Rehabilitation Interventions: A Survey of 185 Randomized Trials. *Physiotherapy Canada*, *68*(3), 267-274. https://doi.org/10.3138/ptc.2015-30

Glaubitz, J. C., Casstevens, T. M., Lu, F., Harriman, J., Elshire, R. J., Sun, Q., & Buckler, E. S. (2014). TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline. *PLOS ONE*, *9*(2), e90346. https://doi.org/10.1371/journal.pone.0090346

Goddard, M. E., & Meuwissen, T. H. E. (2005). The use of linkage disequilibrium to map quantitative trait loci. *Australian Journal of Experimental Agriculture*, *45*(8), 837. https://doi.org/10.1071/EA05066

Gower, J. C. (1975). Generalized Procrustes Analysis. *Psychometrika*, *40*(1), 33-51. https://doi.org/10.1007/BF02291478

Greenacre, M., Groenen, P. J. F., Hastie, T., D'Enza, A. I., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*(1), 100. https://doi.org/10.1038/s43586-022-00184-w

Greenacre, M. J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics*, *20*(2), 251-269. https://doi.org/10.1080/02664769300000021

Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Review*, *53*(2), 217-288. https://doi.org/10.1137/090771806

Hardin, A., & Marcoulides, G. A. (2011). A Commentary on the Use of Formative Measurement. *Educational and Psychological Measurement*, *71*(5), 753-764. https://doi.org/10.1177/0013164411414270

Hastie, T. J., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed). Springer.

He, L.-N., Liu, Y.-J., Xiao, P., Zhang, L., Guo, Y., Yang, T.-L., Zhao, L.-J., Drees, B., Hamilton, J., Deng, H.-Y., Recker, R. R., & Deng, H.-W. (2008). Genomewide Linkage Scan for Combined Obesity Phenotypes using Principal Component Analysis. *Annals of Human Genetics*, *72*(3), 319-326. https://doi.org/10.1111/j.1469-1809.2007.00423.x

Hedrick, P. W. (1987). Gametic disequilibrium measures: proceed with caution. *Genetics*, *117*(2), 331-341. https://doi.org/10.1093/genetics/117.2.331

Henderson, C. (1950). Estimation of genetic parameters. *Ann Math Stat*, *21*(309-310).

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, *38*(6), 226-231. https://doi.org/10.1007/BF01245622

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65-70. Consultado el 21 de diciembre de 2025, desde http://www.jstor.org/stable/4615733

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

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

Hu, Y., & Lin, D. (2010). Analysis of untyped SNPs: maximum likelihood and imputation methods. *Genetic Epidemiology*, *34*(8), 803-815. https://doi.org/10.1002/gepi.20527

Huang, X., & Han, B. (2014). Natural variations and genome-wide association studies in crop plants. *Annual Review of Plant Biology*, *65*, 531-551. https://doi.org/10.1146/annurev-arplant-050213-035715

Hurley, J. R., & Cattell, R. B. (2007). The procrustes program: Producing direct rotation to test a hypothesized factor structure. *Behavioral Science*, *7*(2), 258-262. https://doi.org/10.1002/bs.3830070216

Jin, Q., & Shi, G. (2021). Meta-Analysis of Joint Test of SNP and SNP-Environment Interaction with Heterogeneity. *Human Heredity*, *86*(1-4), 1-9. https://doi.org/10.1159/000519098

Johnston, S. E., McEwan, J. C., Pickering, N. K., Kijas, J. W., Beraldi, D., Pilkington, J. G., Pemberton, J. M., & Slate, J. (2011). Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. *Molecular Ecology*, *20*(12), 2555-2566. https://doi.org/10.1111/j.1365-294X.2011.05076.x

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Jolliffe, I. (2002). *Principal Component Analysis*. Springer-Verlag. https://doi.org/10.1007/b98835

Kaler, A. S., Gillman, J. D., Beissinger, T., & Purcell, L. C. (2020). Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. *Frontiers in Plant Science*, *10*, 1794. https://doi.org/10.3389/fpls.2019.01794

Kang, H. M., Zaitlen, N. A., Wade, C. M., Kirby, A., Heckerman, D., Daly, M. J., & Eskin, E. (2008). Efficient control of population structure in model organism association mapping. *Genetics*, *178*(3), 1709-1723. https://doi.org/10.1534/genetics.107.080101

Kaufman, L., & Rousseeuw, P. J. (1990, marzo). *Finding Groups in Data: An Introduction to Cluster Analysis* (1.ª ed.). Wiley. https://doi.org/10.1002/9780470316801

Kimura, M., & Crow, J. F. (1964). THE NUMBER OF ALLELES THAT CAN BE MAINTAINED IN A FINITE POPULATION. *Genetics*, *49*(4), 725-738. https://doi.org/10.1093/genetics/49.4.725

Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, *9*(1), 29. https://doi.org/10.1186/1746-4811-9-29

Kulwal, K., & Singh, S. (2021). Association Mapping in Plants. En *Crop Breeding*.

Lam, A. C., Schouten, M., Aulchenko, Y. S., Haley, C. S., & de Koning, D.-J. (2007). Rapid and robust association mapping of expression quantitative trait loci. *BMC proceedings*, *1 Suppl 1*(Suppl 1), S144. https://doi.org/10.1186/1753-6561-1-s1-s144

Lange, K. (2010). Singular Value Decomposition [Series Title: Statistics and Computing]. En *Numerical Analysis for Statisticians* (pp. 129-142). Springer New York. https://doi.org/10.1007/978-1-4419-5945-4_9

Laurie, C. C., Doheny, K. F., Mirel, D. B., Pugh, E. W., Bierut, L. J., Bhangale, T., Boehm, F., Caporaso, N. E., Cornelis, M. C., Edenberg, H. J., Gabriel, S. B., Harris, E. L., Hu, F. B., Jacobs, K. B., Kraft, P., Landi, M. T., Lumley, T., Manolio, T. A., McHugh, C., … GENEVA Investigators. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, *34*(6), 591-602. https://doi.org/10.1002/gepi.20516

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, *49*(1), 49-67. https://doi.org/10.1093/genetics/49.1.49

Li, J., & Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, *95*(3), 221-227. https://doi.org/10.1038/sj.hdy.6800717

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.-M., Todhunter, R. J., Buckler, E. S., & Zhang, Z. (2014). Enrichment of statistical power for genome-wide association studies. *BMC Biology*, *12*(1), 73. https://doi.org/10.1186/s12915-014-0073-5

Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype Imputation. *Annual review of genomics and human genetics*, *10*, 387-406. https://doi.org/10.1146/annurev.genom.9.081307.164242

Lipka, A. E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P. J., Gore, M. A., Buckler, E. S., & Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. *Bioinformatics*, *28*(18), 2397-2399. https://doi.org/10.1093/bioinformatics/bts444

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., & Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833-835. https://doi.org/10.1038/nmeth.1681

Little, R. J. (2006). Calibrated Bayes: A Bayes/Frequentist Roadmap. *The American Statistician*, *60*(3), 213-223. https://doi.org/10.1198/000313006X117837

Liu, X., Huang, M., Fan, B., Buckler, E. S., & Zhang, Z. (2016). Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies (J. Listgarten, Ed.). *PLOS Genetics*, *12*(2), e1005767. https://doi.org/10.1371/journal.pgen.1005767

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., Farkhari, M., Ribaut, J.-M., Cao, M., Rong, T., & Xu, Y. (2010). Joint linkage-linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(45), 19585-19590. https://doi.org/10.1073/pnas.1006105107

Malosetti, M., Ribaut, J.-M., & van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, *4*. https://doi.org/10.3389/fphys.2013.00044

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., & Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, *108*(3), 285-291. https://doi.org/10.1038/hdy.2011.73

Mehmood, T., Liland, K. H., Snipen, L., & Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, *118*, 62-69. https://doi.org/10.1016/j.chemolab.2012.07.010

Metropolis, N., & Ulam, S. (1949). The Monte Carlo Method. *Journal of the American Statistical Association*, *44*(247), 335-341. https://doi.org/10.1080/01621459.1949.10483310

Mevik, B.-H., & Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, *18*, 1-23. https://doi.org/10.18637/jss.v018.i02

Meyer, H. V., & Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics (Oxford, England)*, *34*(17), 2951-2956. https://doi.org/10.1093/bioinformatics/bty197

Möhring, J., & Piepho, H.-P. (2009). Comparison of Weighting in Two-Stage Analysis of Plant Breeding Trials. *Crop Science*, *49*(6), 1977-1988. https://doi.org/10.2135/cropsci2009.02.0083

Money, D., Gardner, K., Migicovsky, Z., Schwaninger, H., Zhong, G.-Y., & Myles, S. (2015). LinkImpute: Fast and Accurate Genotype Imputation for Nonmodel Organisms. *G3 (Bethesda, Md.)*, *5*(11), 2383-2390. https://doi.org/10.1534/g3.115.021667

Montesinos-López, O. A., Montesinos-López, A., Kismiantini, Roman-Gallardo, A., Gardner, K., Lillemo, M., Fritsche-Neto, R., & Crossa, J. (2022). Partial Least Squares Enhances Genomic Prediction of New Environments. *Frontiers in Genetics*, *13*, 920689. https://doi.org/10.3389/fgene.2022.920689

Nelson, P. R., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*(1), 45-65. https://doi.org/10.1016/S0169-7439(96)00007-X

Nengsih, T. A., Bertrand, F., Maumy-Bertrand, M., & Meyer, N. (2019). Determining the Number of Components in PLS Regression on Incomplete Data [arXiv:1810.08104 [stat]]. *Statistical Applications in Genetics and Molecular Biology*, *18*(6), 20180059. https://doi.org/10.1515/sagmb-2018-0059
Comment: 16 pages, 3 figures.

Nguyen, H. D., & McLachlan, G. J. (2016). Linear mixed models with marginally symmetric nonparametric random effects. *Computational Statistics & Data Analysis*, *103*, 151-169. https://doi.org/10.1016/j.csda.2016.05.005

Nordborg, M., & Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics*, *18*(2), 83-90. https://doi.org/10.1016/S0168-9525(02)02557-X

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., … Marchini, J. (2014). A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genetics*, *10*(4), e1004234. https://doi.org/10.1371/journal.pgen.1004234

Ortiz, R., Reslow, F., Montesinos-López, A., Huicho, J., Pérez-Rodríguez, P., Montesinos-López, O. A., & Crossa, J. (2023). Partial least squares enhance multi-trait genomic prediction of potato cultivars in new environments. *Scientific Reports*, *13*(1), 9947. https://doi.org/10.1038/s41598-023-37169-y

P. Vatcheva, K., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, *06*(02). https://doi.org/10.4172/2161-1165.1000227

Pasam, R. K., Sharma, R., Malosetti, M., van Eeuwijk, F. A., Haseneyer, G., Kilian, B., & Graner, A. (2012). Genome-wide association studies for agronomical traits in a world wide spring barley collection. *BMC plant biology*, *12*, 16. https://doi.org/10.1186/1471-2229-12-16

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genetics*, *2*(12), e190. https://doi.org/10.1371/journal.pgen.0020190

Peña-Malavera, A., Bruno, C., Fernandez, E., & Balzarini, M. (2014). Comparison of algorithms to infer genetic population structure from unlinked molecular markers. *Statistical Applications in Genetics and Molecular Biology*, *13*(4). https://doi.org/10.1515/sagmb-2013-0006

Pérez, P., & De Los Campos, G. (2014). Genome-Wide Regression and Prediction with the BGLR Statistical Package. *Genetics*, *198*(2), 483-495. https://doi.org/10.1534/genetics.114.164442

Perez-Guaita, D., Kuligowski, J., Quintás, G., Garrigues, S., & Guardia, M. D. L. (2013). Modified locally weighted—Partial least squares regression improving clinical predictions from infrared spectra of human serum samples. *Talanta*, *107*, 368-375. https://doi.org/10.1016/j.talanta.2013.01.035

Piepho, H. P., Möhring, J., Melchinger, A. E., & Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, *161*(1), 209-228. https://doi.org/10.1007/s10681-007-9449-8

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909. https://doi.org/10.1038/ng1847

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959. https://doi.org/10.1093/genetics/155.2.945

Racedo, J., Gutiérrez, L., Perera, M. F., Ostengo, S., Pardo, E. M., Cuenya, M. I., Welin, B., & Castagnaro, A. P. (2016). Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biology*, *16*(1), 142. https://doi.org/10.1186/s12870-016-0829-x

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-Validation. En L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 532-538). Springer US. https://doi.org/10.1007/978-0-387-39940-9_565

Resende, M. F. R., Jr, Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., & Kirst, M. (2012). Accuracy of Genomic Selection Methods in a Standard Data Set of Loblolly Pine (Pinus taeda L.) *Genetics*, *190*(4), 1503-1510. https://doi.org/10.1534/genetics.111.137026

Rohart, F., Gautier, B., Singh, A., & Lê Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, *13*(11), e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Romeu, J. F., Monforte, A. J., Sánchez, G., Granell, A., García-Brunton, J., Badenes, M. L., & Ríos, G. (2014). Quantitative trait loci affecting reproductive phenology in peach. *BMC Plant Biology*, *14*(1), 52. https://doi.org/10.1186/1471-2229-14-52

Rönnegård, L., Felleki, M., Fikse, F., Mulder, H. A., & Strandberg, E. (2010). Genetic heterogeneity of residual variance - estimation of variance components using double hierarchical generalized linear models. *Genetics Selection Evolution*, *42*(1), 8. https://doi.org/10.1186/1297-9686-42-8

Rosipal, R., & Krämer, N. (2006). Overview and Recent Advances in Partial Least Squares [Series Title: Lecture Notes in Computer Science]. En C. Saunders, M. Grobelnik, S. Gunn & J. Shawe-Taylor (Eds.), *Subspace, Latent Structure and Feature Selection* (pp. 34-51, Vol. 3940). Springer Berlin Heidelberg. https://doi.org/10.1007/11752790_2

Rousseeuw, P. J., & Leroy, A. M. (1987, octubre). *Robust Regression and Outlier Detection* (1.ª ed.). Wiley. https://doi.org/10.1002/0471725382

Rutkoski, J. E., Poland, J., Jannink, J.-L., & Sorrells, M. E. (2013). Imputation of Unordered Markers and the Impact on Genomic Selection Accuracy. *G3 Genes|Genomes|Genetics*, *3*(3), 427-439. https://doi.org/10.1534/g3.112.005363

Sanchez, G. (2013, noviembre). *Cartografiado de QTL y genes candidatos asociados a metabolitos determinantes de la calidad de fruto en melocotón* [Tesis doctoral, Universitat Politècnica de València]. https://doi.org/10.4995/Thesis/10251/34511

Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: a large-scale genome simulator for livestock. *Bioinformatics*, *25*(5), 680-681. https://doi.org/10.1093/bioinformatics/btp045

Schillemans, T., Shi, L., Liu, X., Åkesson, A., Landberg, R., & Brunius, C. (2019). Visualization and Interpretation of Multivariate Associations with Disease Risk Markers and Disease Risk—The Triplot. *Metabolites*, *9*(7), 133. https://doi.org/10.3390/metabo9070133

Schönemann, P. H. (1966). A Generalized Solution of the Orthogonal Procrustes Problem. *Psychometrika*, *31*(1), 1-10. https://doi.org/10.1007/BF02289451

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461-464. Consultado el 10 de enero de 2026, desde http://www.jstor.org/stable/2958889

Segura, V., Vilhjálmsson, B. J., Platt, A., Korte, A., Seren, Ü., Long, Q., & Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. *Nature Genetics*, *44*(7), 825-830. https://doi.org/10.1038/ng.2314

Shang, J., Xu, A., Bi, M., Zhang, Y., Li, F., & Liu, J.-X. (2024). A review: simulation tools for genome-wide interaction studies. *Briefings in Functional Genomics*, *23*(6), 745-753. https://doi.org/10.1093/bfgp/elae034

Shen, H. T. (2009). Principal Component Analysis. En L. Liu & M. T. Özsu (Eds.), *Encyclopedia of Database Systems* (pp. 2136-2136). Springer US. https://doi.org/10.1007/978-0-387-39940-9_540

Shi, G. (2022). Genome-wide variance quantitative trait locus analysis suggests small interaction effects in blood pressure traits. *Scientific Reports*, *12*(1), 12649. https://doi.org/10.1038/s41598-022-16908-7

Sidak, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association*, *62*(318), 626. https://doi.org/10.2307/2283989

Slatkin, M. (2016). Statistical methods for analyzing ancient DNA from hominins. *Current Opinion in Genetics & Development*, *41*, 72-76. https://doi.org/10.1016/j.gde.2016.08.004

Smith, A. B., Cullis, B. R., & Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science*, *143*(6), 449-462. https://doi.org/10.1017/S0021859605005587

Stich, B., Melchinger, A., Frisch, M., Maurer, H., Heckenberger, M., & Reif, J. (2005). Linkage disequilibrium in European elite maize germplasm investigated with SSRs | Theoretical and Applied Genetics. *111*, 723-730. Consultado el 5 de diciembre de 2025, desde https://link.springer.com/article/10.1007/S00122-005-2057-X

Stich, B., Möhring, J., Piepho, H.-P., Heckenberger, M., Buckler, E. S., & Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics*, *178*(3), 1745-1754. https://doi.org/10.1534/genetics.107.079707

Stocchero, M., Riccadonna, S., & Franceschi, P. (2018). Projection to latent structures with orthogonal constraints for metabolomics data [_eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/cem.2987]. *Journal of Chemometrics*, *32*(5), e2987. https://doi.org/10.1002/cem.2987 e2987 CEM-17-0118.R1.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *36*(2), 111-133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Sun, H., Zhang, Z., Olasege, B. S., Xu, Z., Zhao, Q., Ma, P., Wang, Q., & Pan, Y. (2019). Application of partial least squares in exploring the genome selection signatures between populations. *Heredity*, *122*(3), 288-293. https://doi.org/10.1038/s41437-018-0121-y

Swarts, K., Li, H., Romero Navarro, J. A., An, D., Romay, M. C., Hearne, S., Acharya, C., Glaubitz, J. C., Mitchell, S., Elshire, R. J., Buckler, E. S., & Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The plant genome*, *7*(3). Consultado el 7 de diciembre de 2025, desde https://agris.fao.org/search/en/providers/122535/records/65df0d890f3e94b9e5d46b91?utm_source=chatgpt.com

Tabangin, M. E., Woo, J. G., & Martin, L. J. (2009). The effect of minor allele frequency on the likelihood of obtaining false positives. *BMC Proceedings*, *3*(7), S41. https://doi.org/10.1186/1753-6561-3-S7-S41

Tak, Y. G., & Farnham, P. J. (2015). Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. *Epigenetics & Chromatin*, *8*(1), 57. https://doi.org/10.1186/s13072-015-0050-4

Tanksley, S., & McCouch, S. (1997). Seed banks and molecular maps: unlocking genetic potential from the wild. https://doi.org/10.1126/science.277.5329.1063

Ter Braak, C. J. F. (1994). Canonical community ordination. Part I: Basic theory and linear methods. *Écoscience*, *1*(2), 127-140. https://doi.org/10.1080/11956860.1994.11682237

Tracy, C. A., & Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, *159*(1), 151-174. https://doi.org/10.1007/BF02100489

Vaida, F., & Blanchard, S. (2005). Conditional Akaike Information for Mixed-Effects Models. *Biometrika*, *92*(2), 351-370. Consultado el 8 de enero de 2026, desde http://www.jstor.org/stable/20441193

Valentini, G. H., & Arroyo, L. E. (2007). 'Don Carlos INTA' Peach. *HortScience*, *42*(5), 1295-1296. https://doi.org/10.21273/HORTSCI.42.5.1295

VanRaden, P. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, *91*(11), 4414-4423. https://doi.org/10.3168/jds.2007-0980

Verma, A., Chatrath, R., & Sharma, I. (2015). AMMI and GGE biplots for G×E analysis of wheat genotypes under rain fed conditions in central zone of India. *Journal of Applied and Natural Science*, *7*(2), 656-661. https://doi.org/10.31018/jans.v7i2.662

Vershynin, R. (2018, septiembre). *High-Dimensional Probability: An Introduction with Applications in Data Science* (1.ª ed.). Cambridge University Press. https://doi.org/10.1017/9781108231596

Vicente-Gonzalez, L., & Vicente-Villardon, J. L. (2022). Partial Least Squares Regression for Binary Responses and Its Associated Biplot Representation. *Mathematics*, *10*(15), 2580. https://doi.org/10.3390/math10152580

Videla, M. E., Iglesias, J., & Bruno, C. (2021). Relative performance of cluster algorithms and validation indices in maize genome-wide structure patterns. *Euphytica*, *217*(10), 195. https://doi.org/10.1007/s10681-021-02926-5

Vilhjálmsson, B. J., & Nordborg, M. (2013). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, *14*(1), 1-2. https://doi.org/10.1038/nrg3382

Vos, P. G., Paulo, M. J., Voorrips, R. E., Visser, R. G. F., van Eck, H. J., & van Eeuwijk, F. A. (2017). Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *TAG. Theoretical and Applied Genetics. Theoretische Und Angewandte Genetik*, *130*(1), 123-135. https://doi.org/10.1007/s00122-016-2798-8

Wall, J. D., & Pritchard, J. K. (2003). Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics*, *4*(8), 587-597. https://doi.org/10.1038/nrg1123

Wang, C., He, W., Li, K., Yu, Y., Zhang, X., Yang, S., Wang, Y., Yu, L., Huang, W., Yu, H., Chen, L., & Cheng, X. (2025). Genetic Diversity Analysis and GWAS of Plant Height and Ear Height in Maize Inbred Lines from South-East China. *Plants*, *14*(3), 481. https://doi.org/10.3390/plants14030481

Wang, E., Brown, H. E., Rebetzke, G. J., Zhao, Z., Zheng, B., & Chapman, S. C. (2019). Improving process-based crop models to better capture genotype×environment×management interactions. *Journal of Experimental Botany*, *70*(9), 2389-2401. https://doi.org/10.1093/jxb/erz092

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., Zhang, J., Dunwell, J. M., Xu, S., & Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports*, *6*(1), 19444. https://doi.org/10.1038/srep19444

Weir, B. S., & Hill, W. G. (1986). Nonuniform recombination within the human beta-globin gene cluster. *American Journal of Human Genetics*, *38*(5), 776-781.

West, B. T., Welch, K. B., & Galecki, A. T. (2014, julio). *Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition* (0.ª ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b17198

Westerhuis, J. A., Kourti, T., & MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models [_eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi 128X%28199809/10%2912%3A5%3C301%3A%3AAID-CEM515%3E3.0.CO%3B2-S]. *Journal of Chemometrics*, *12*(5), 301-321. https://doi.org/10.1002/(SICI)1099-128X(199809/10)12:5<301::AID-CEM515>3.0.CO;2-S

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. *In: Krishnaiah, P.R., Ed., Multivariate Analysis, Academic Press, New York*.

Wold, H. (1985). Partial Least Squares. En *Encyclopedia of Statistical Sciences* (pp. 581-591, Vol. 6).

Wold, H. (1973). Nonlinear Iterative Partial Least Squares (NIPALS) Modelling: Some Current Developments. https://doi.org/https://api.semanticscholar.org/CorpusID:118962244

Wold, S., Martens, H., & Wold, H. (1983). The multivariate calibration problem in chemistry solved by the PLS method [Series Title: Lecture Notes in Mathematics]. En B. Kågström & A. Ruhe (Eds.), *Matrix Pencils* (pp. 286-293, Vol. 973). Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0062108

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109-130. https://doi.org/10.1016/S0169-7439(01)00155-1

Wolfinger, R. (1993). Covariance structure selection in general mixed models. *Communications in Statistics - Simulation and Computation*, *22*(4), 1079-1106. https://doi.org/10.1080/03610919308813143

Wondola, D. W., Aulele, S. N., & Lembang, F. K. (2020). Partial Least Square (PLS) Method of Addressing Multicollinearity Problems in Multiple Linear Regressions (Case Studies: Cost of electricity bills and factors affecting it). *Journal of Physics: Conference Series*, *1463*(1), 012006. https://doi.org/10.1088/1742-6596/1463/1/012006

Xu, S. (2013). Genetic Mapping and Genomic Selection Using Recombination Breakpoint Data. *Genetics*, *195*(3), 1103-1115. https://doi.org/10.1534/genetics.113.155309

Xu, Y., & Crouch, J. H. (2008). Marker-Assisted Selection in Plant Breeding: From Publications to Practice [_eprint: https://acsess.onlinelibrary.wiley.com/doi/pdf/10.2135/cropsci2007.04.0 *Crop Science*, *48*(2), 391-407. https://doi.org/10.2135/cropsci2007.04.0191

Yang, J., Loos, R. J. F., Powell, J. E., Medland, S. E., Speliotes, E. K., Chasman, D. I., Rose, L. M., Thorleifsson, G., Steinthorsdottir, V., Mägi, R., Waite, L., Vernon Smith, A., Yerges-Armstrong, L. M., Monda, K. L., Hadley, D., Mahajan, A., Li, G., Kapur, K., Vitart, V., … Visscher, P. M. (2012). FTO genotype is associated

with phenotypic variability of body mass index. *Nature*, *490*(7419), 267-272. https://doi.org/10.1038/nature11401

Yang, R.-C. (2007). Mixed-Model Analysis of Crossover Genotype–Environment Interactions. *Crop Science*, *47*(3), 1051-1062. https://doi.org/10.2135/cropsci2006.09.0611

Yang, W., Feng, H., Zhang, X., Zhang, J., Doonan, J. H., Batchelor, W. D., Xiong, L., & Yan, J. (2020). Crop Phenomics and High-Throughput Phenotyping: Past Decades, Current Challenges, and Future Perspectives. *Molecular Plant*, *13*(2), 187-214. https://doi.org/10.1016/j.molp.2020.01.008

Yang, X., Sood, S., Glynn, N., Islam, M. S., Comstock, J., & Wang, J. (2017). Constructing high-density genetic maps for polyploid sugarcane (Saccharum spp.) and identifying quantitative trait loci controlling brown rust resistance. *Molecular Breeding*, *37*(10), 116. https://doi.org/10.1007/s11032-017-0716-7

Yao, H., Song, Y., Chen, Y., Wu, N., Xu, J., Sun, C., Zhang, J., Weng, T., Zhang, Z., Wu, Z., Cheng, L., Shi, D., Lu, X., Lei, J., Crispin, M., Shi, Y., Li, L., & Li, S. (2020). Molecular Architecture of the SARS-CoV-2 Virus. *Cell*, *183*(3), 730-738.e13. https://doi.org/10.1016/j.cell.2020.09.018

Yu, J., & Buckler, E. S. (2006). Genetic association mapping and genome organization of maize. *Current Opinion in Biotechnology*, *17*(2), 155-160. https://doi.org/10.1016/j.copbio.2006.02.003

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203-208. https://doi.org/10.1038/ng1702

Yuan, Z., Druzhinina, I. S., Wang, X., Zhang, X., Peng, L., & Labbé, J. (2020). Insight into a highly polymorphic endophyte isolated from the roots of the halophytic seepweed Suaeda salsa: Laburnicola rhizohalophila sp. nov. (Didymosphaeriaceae, Pleosporales). *Fungal Biology*, *124*(5), 327-337. https://doi.org/10.1016/j.funbio.2019.10.001

Zapata, C. (2000). THE D' MEASURE OF OVERALL GAMETIC DISEQUILIBRIUM BETWEEN PAIRS OF MULTIALLELIC LOCI | Evolution | Oxford Academic. https://doi.org/https://doi.org/10.1111/j.0014-3820.2000.tb00724.x

Zhang, Y., Liu, P., Zhang, X., Zheng, Q., Chen, M., Ge, F., Li, Z., Sun, W., Guan, Z., Liang, T., Zheng, Y., Tan, X., Zou, C., Peng, H., Pan, G., & Shen, Y. (2018). Multi-Locus Genome-Wide Association Study Reveals the Genetic Architecture of Stalk Lodging Resistance-Related Traits in Maize. *Frontiers in Plant Science*, *9*, 611. https://doi.org/10.3389/fpls.2018.00611

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

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M., & Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics*, *42*(4), 355-360. https://doi.org/10.1038/ng.546

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, *44*(7), 821-824. https://doi.org/10.1038/ng.2310

Zhu, Q., Sarkis, J., & Lai, K.-h. (2008). Confirmation of a measurement model for green supply chain management practices implementation. *International Journal of Production Economics*, *111*(2), 261-273. https://doi.org/10.1016/j.ijpe.2006.11.029

Anexo

Anexo 1

Código de programación usado para obtener la bases de datos simulada con el programa R

```
##Cargo base de durazno en que se basa la simulacion

tmp_vcf<-readLines("./populations.snps.191snps.vcf")
tmp_vcf_data<-read.table("./populations.snps.191snps.vcf",
stringsAsFactors = FALSE)

tmp_vcf<-tmp_vcf[-(grep("#CHROM",tmp_vcf)+1):-(length(tmp_vcf))]
vcf_names<-unlist(strsplit(tmp_vcf[length(tmp_vcf)],"\t"))
names(tmp_vcf_data)<-vcf_names

cromosoma=tmp_vcf_data[,1]
posicion=tmp_vcf_data$POS

#Defino el genoma para la simulacion
genoma=data.frame(cbind("cromosoma"=as.character(cromosoma),
"posicion"=as.numeric(posicion)))

genoma$posicion=as.character(genoma$posicion)
genoma$posicion=as.numeric(genoma$posicion)

length_cro_1=genoma[genoma$cromosoma=="Pp01",]
length_cro_2=genoma[genoma$cromosoma=="Pp02",]
length_cro_3=genoma[genoma$cromosoma=="Pp03",]
length_cro_4=genoma[genoma$cromosoma=="Pp04",]
length_cro_5=genoma[genoma$cromosoma=="Pp05",]
length_cro_6=genoma[genoma$cromosoma=="Pp06",]
length_cro_7=genoma[genoma$cromosoma=="Pp07",]
length_cro_8=genoma[genoma$cromosoma=="Pp08",]

#Creacion de la poblacion historica

genome<-data.frame(matrix(NA, nrow=8, ncol=6))
names(genome)<-c("chr","len","nmrk","mpos","nqtl","qpos")
genome$chr<-c(1:8)
genome$nmrk<-rep(10000,8)

genome$len<-rep(260,8)
genome$mpos<-rep('rnd',8)
genome$nqtl<-rep(15,8)
```

```
genome$qpos<-rep('rnd',8)
genome
hp<-make_hp(hpsize=500
,ng=300,h2=0.4,d2=0.1,phen_var=1
,genome=genome,mutr=5*10**-4,
sel_seq_qtl=0.05,sel_seq_mrk=0.05,laf=0.5)

# Creo una poblacion reciente

Male_founders<-data.frame(number=50,select='rnd')
Female_founders<-data.frame(number=50,select='rnd')

Selection<-data.frame(matrix(NA, nrow=2, ncol=2))
names(Selection)<-c('Number','type')
Selection$Number[1:2]<-c(60,60)
Selection$type[1:2]<-c('rnd','rnd')
Selection
RP_1<-sample_hp(hp_out=hp,Male_founders=
Male_founders,Female_founders=Female_founders,
ng=5,Selection=Selection,litter_size=4)

for(i in 1:nrow(RP_1$output[[6]]$freqMRK)){
        if(RP_1$output[[6]]$freqMRK$Freq.Allele1[i]<0.05 |
        RP_1$output[[6]]$freqMRK$Freq.Allele1[i]>0.95){
                print(RP_1$output[[6]]$freqMRK$Freq.Allele1[i])
        }
}


#Matriz de marcadores moleculares
MRK_SIMULATION_0=data.frame(RP_1$output[[6]]$mrk[,-c(1,2)])

MRK_SIMULATION=data.frame(matrix(0,nrow(MRK_SIMULATION_0),
(ncol(MRK_SIMULATION_0)/2)))

n=seq(1,ncol(MRK_SIMULATION_0),2)
for(j in n){
        MRK_SIMULATION[,(j+1)/2]<-MRK_SIMULATION_0[,j]+MRK_
           SIMULATION_0[,j+1]
}

MRK_SIMULATION=MRK_SIMULATION-2

#Simulacion de dos ambientes con distribucion normal y con
   interaccion

phen=cbind(RP_1$output[[6]]$data$id,RP_1$output[[6]]$data$phen)
colnames(phen)=c("genotipo","ambiente1")
phen=as.data.frame(phen)
phen$estructura=rep(0,nrow(phen))

for(i in 1:nrow(phen)){
        if(phen$ambiente1[i]<quantile(phen$ambiente1, prob=c(0.10))
           ){ phen$estructura[i]=1
        } else if(phen$ambiente1[i]<quantile(phen$ambiente1, prob=c
           (0.25))){ phen$estructura[i]=2
```

```
        } else if(phen$ambiente1[i]<quantile(phen$ambiente1, prob=c
            (0.50))) {phen$estructura[i]=3
        } else if(phen$ambiente1[i]<quantile(phen$ambiente1, prob=c
            (0.75))) {phen$estructura[i]=4
        } else if(phen$ambiente1[i]<quantile(phen$ambiente1, prob=c
            (0.90))) {phen$estructura[i]=5
        } else{phen$estructura[i]=6}
}

phen$ambiente2=rep(0,nrow(phen))

for(i in 1:nrow(phen)){
        if(phen$estructura[i]==1){phen$ambiente2[i]=phen$ambiente1[
            i]+rnorm(1,2.5,0.5)
        }else if(phen$estructura[i]==2){phen$ambiente2[i]=phen$
            ambiente1[i]+rnorm(1,0,1.5)
        }else if(phen$estructura[i]==3){phen$ambiente2[i]=phen$
            ambiente1[i]+rnorm(1,0,1)
        }else if(phen$estructura[i]==4){phen$ambiente2[i]=phen$
            ambiente1[i]+rnorm(1,0.5,1)
        }else if(phen$estructura[i]==5){phen$ambiente2[i]=phen$
            ambiente1[i]+rnorm(1,0.5,1.5)
        }else if(phen$estructura[i]==6){phen$ambiente2[i]=phen$
            ambiente1[i]+rnorm(1,-2.5,0.5)}
}

#Simulacion de las repeticiones

sigma=t(cbind(c(0.5,0.5),
c(0.5,1.5),
c(0.5,3.5)))

pheno=phen[,-c(1,3)]
pheno_cr=matrix(0,nrow(pheno)*3,ncol(pheno)*3)


for(j in 1:ncol(pheno_cr)){
        for(i in 1:nrow(pheno_cr)){
                if(j%%2!=0){
                        pheno_cr[i,j]=pheno[ceiling(i/3),1]+rnorm
                            (1,0,sqrt(sigma[ceiling(j/2),1]))
                }else{pheno_cr[i,j]=pheno[ceiling(i/3),2]+rnorm
                    (1,0,sqrt(sigma[ceiling(j/2),2]))}
        }
}

#Armar base fenotipo

base_pheno=data.frame(matrix(0,(nrow(pheno_cr)*2),5))

colnames(base_pheno)=c("pheno_1","pheno_2","pheno_3",
"Genotype","Enviroment")

base_pheno["pheno_1"]=c(pheno_cr[,1],pheno_cr[,2])

base_pheno["pheno_2"]=c(pheno_cr[,3],pheno_cr[,4])
```

```
base_pheno["pheno_3"]=c(pheno_cr[,5],pheno_cr[,6])


base_pheno["Genotype"]=rep(sort(as.numeric(rep(RP_1$output[[6]]$
    data$id,3))),2)

base_pheno["Enviroment"]=c(rep(1,nrow(pheno_cr)),rep(2,nrow(pheno_
    cr)))

base_pheno$Enviroment=as.factor(base_pheno$Enviroment)
base_pheno$Genotype=as.factor(base_pheno$Genotype)
```

Código de programación usado para ajustar modelos lineales mixtos para obtener los G-BLUP

```
#Modelo de simetria compuesta (SC) con kinship
E <- diag(length(unique(base_pheno$Enviroment)))
rownames(E) <- colnames(E) <- unique(base_pheno$Enviroment)
EK <- kronecker(E,K, make.dimnames = TRUE)
MLM_sv_1 <- mmer(pheno_1~Enviroment,
random= ~ vs(Genotype, Gu=K) + vs(Enviroment:Genotype, Gu=EK),
rcov= ~ units,
data=base_pheno, verbose = FALSE)
MLM_sv_1$AIC
MLM_sv_1$BIC

#Modelo de simetria compuesta (SC) sin kinship
MLM_sv_2 <- mmer(pheno_1 ~Enviroment,
random= ~ Genotype + Enviroment:Genotype,
rcov= ~ units,
data=base_pheno, verbose = FALSE)
MLM_sv_2$AIC
MLM_sv_2$BIC

#Autorregresivo de primer orden
MLM_sv_3 <- mmer(pheno_1~1,
random=~ vs(Enviroment, Gu=AR1(Enviroment, rho=0.3)) + Genotype,
rcov=~units,
data=base_pheno, verbose = FALSE)
MLM_sv_3$AIC
MLM_sv_3$BIC

#Modelo Diagonal
MLM_sv_4 <- mmer(pheno_1~Enviroment,
random= ~ vs(ds(Enviroment),Genotype),
rcov= ~ units,
data=base_pheno, verbose = FALSE)
MLM_sv_4$AIC
MLM_sv_4$BIC

#Modelo Diagonal con kinship
MLM_sv_5 <- mmer(pheno_1~Enviroment,
```

```
random= ~ vs(ds(Enviroment),Genotype, Gu=K),
rcov= ~ units,
data=base_pheno, verbose = FALSE)
MLM_sv_5$AIC
MLM_sv_5$BIC

#Sin estructura para los ambientes

MLM_sv_6 <- mmer(pheno_1~Enviroment,
random=~ vs(us(Enviroment),Genotype),
rcov=~vs(us(Enviroment),units),
data=base_pheno, verbose = FALSE)
MLM_sv_6$AIC
MLM_sv_6$BIC

#Sin estructura con matriz kinship
MLM_sv_7 <- mmer(pheno_1~Enviroment,
random= ~ vs(us(Enviroment),Genotype, Gu=K),
rcov= ~ units,
data=base_pheno, verbose = FALSE)
MLM_sv_7$AIC
MLM_sv_7$BIC
```

Código de programación usado para ajustar los modelos GWAS, ejemplo del escenario sin heterocedasticidad entre ambientes

```
feno=matrix(0,length(BLUP_sv$pheno_1),2)
colnames(feno)=c("Taxa","pheno_1")
feno[,"Taxa"]=RP_1$output[[6]]$data$id
feno[,"pheno_1"]=as.numeric(BLUP_sv$pheno_1)


Modelos=c("GLM", "MLM","MLMM" ,"CMLM", "ECMLM", "SUPER","FarmCPU")

pvalor_GLM=matrix(0,ncol(X_2)-1,ncol(feno)-1)
pvalor_MLM=matrix(0,ncol(X_2)-1,ncol(feno)-1)
pvalor_MLMM=matrix(0,ncol(X_2)-1,ncol(feno)-1)
pvalor_CMLM=matrix(0,ncol(X_2)-1,ncol(feno)-1)
pvalor_ECMLM=matrix(0,ncol(X_2)-1,ncol(feno)-1)
pvalor_SUPER=matrix(0,ncol(X_2)-1,ncol(feno)-1)
pvalor_FarmCPU=matrix(0,ncol(X_2)-1,ncol(feno)-1)


pvalor_sv_1=list(pvalor_GLM,pvalor_MLM, pvalor_MLMM,
pvalor_CMLM, pvalor_ECMLM,pvalor_SUPER,
pvalor_FarmCPU)



source("http://zzlab.net/GAPIT/GAPIT.library.R")
source("http://zzlab.net/GAPIT/gapit_functions.txt")

map=RP_1$linkage_map_mrk
colnames(map)=c("Name","Chromosome","Position")
```

```
Mod=c(1,2,3,4,5,6,7)
for(j in Mod){
        myGAPIT <- GAPIT(
        Y=feno,
        GD=X_2,
        GM=map,
        model=Modelos[j],
        PCA.total=ncomp,
        #CV=Q_m,
        #kinship.algorithm="VanRaden",
        SNP.MAF=0,
        Inter.Plot=F,
        PCA.3d=F,
        file.output=F
        )
        pvalor_sv_1[[j]]=myGAPIT$GWAS$P.value
}
```

Anexo 2

Código en R de la aplicación de la técnica PLS sobre la base de datos simulada para SVD

```
Y_predicha_S=list()
Y_real_S=list()
k_fold=10
for(i in 1:k_fold){
        n_train <- round(0.85 * n_total)

        indices_train <- sample(1:n_total, n_train, replace = FALSE
           )

        X_train <- Cuanti_molec[indices_train,3:ncol(Cuanti_molec)]
        Y_train <- Cuanti_molec[indices_train, 1:2]
        X_test <- Cuanti_molec[-indices_train,3:ncol(Cuanti_molec)]
        Y_test <- Cuanti_molec[-indices_train, 1:2]

        X_train=as.matrix(X_train)
        Y_train=as.matrix(Y_train)
        X_test=as.matrix(X_test)
        Y_test=as.matrix(Y_test)


        # Build the model on training set

        model <- PLSR_SVD(X=X_train, Y=Y_train, 2)

        ExpectedY=model$Bpls_star[1,] +X_test%*%model$Bpls_star[2:
           nrow(model$Bpls_star),]

        Y_predicha_S=list.append(Y_predicha_S, ExpectedY)
```

```
        Y_real_S=list.append(Y_real_S, Y_test)
        print(i)
}


cv_error_S=matrix(0,k_fold,ncol(Y_train))
for(i in 1:k_fold){
        cv_error_S[i,]=sqrt(colMeans((Y_real_S[[i]]-Y_predicha_S[[i
           ]])**2))
}
```

Código en R de la aplicación de la técnica PLS sobre la base de datos simulada para NIPALS

```
Y_predicha_N=list()
Y_real_N=list()
k_fold=10
for(i in 1:k_fold){
        n_train <- round(0.85 * n_total)

        indices_train <- sample(1:n_total, n_train, replace = FALSE
           )

        X_train <- Cuanti_molec_cn[indices_train,4:ncol(Cuanti_
           molec_cn)]
        Y_train <- Cuanti_molec_cn[indices_train, 2:3]
        X_test <- Cuanti_molec_cn[-indices_train,4:ncol(Cuanti_
           molec_cn)]
        Y_test <- Cuanti_molec_cn[-indices_train, 2:3]

        X_train=as.matrix(X_train)
        Y_train=as.matrix(Y_train)
        X_test=as.matrix(X_test)
        Y_test=as.matrix(Y_test)

        # Build the model on training set

        model <- pls(X_train,Y_train, scale=TRUE, ncomp=2, mode="
           regression")

        # Predecir Y esperado
        Y_hat <- predict(model, newdata = X_test)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_N=list.append(Y_predicha_N, ExpectedY)
        Y_real_N=list.append(Y_real_N, Y_test)
        print(i)
}


cv_error_Ni=matrix(0,k_fold,ncol(Y_train))
```

```
for(i in 1:10){
        cv_error_Ni[i,]=sqrt(colMeans((Y_real_N[[i]]-Y_predicha_N[[
           i]])**2))
        #colMeans(Y_real_N[[i]])
}
```

Código en R para la visualización de la técnica PLS con SVD a través del triplot

```
model_svd <- PLSR_SVD(X=X, Y=Y, 2)
tiempo2 <- system.time(PLSR_SVD(X=X, Y=Y, 2))

# Triplot del pls
plot(model_svd$P[,1]*0.5,model_svd$P[,2]*0.5,
xlab='Componente␣1', ylab='Componente␣2', cex=0.01
)
text(model_svd$P[,1]*0.5,model_svd$P[,2]*0.5,
colnames(X)[3:ncol(X)],
cex = .5)

points(model_svd$U[,1],model_svd$U[,2], cex=0.01)
text(model_svd$U[,1],model_svd$U[,2],
row.names(model_svd$U),
cex = .5, col='Purple')

abline(h=0,v=0, lty=c(2,2), lwd=c(1, 1))

s = 100000
# components
for(i in 1:2){
        Q_ms=qnorm(0.99, mean(model_svd$Bpls[,i]), sd(model_svd$
           Bpls[,i]))
        #model$RegParameters[model$RegParameters[,'R10-11']>Q_R10,'
           R10-11']

        print(length(model_svd$Bpls[model_svd$Bpls[,i]>Q_ms,i]))
}

for(i in 1:2){
        Q_ms=qnorm(0.99, mean(model_svd$Bpls[,i]), sd(model_svd$
           Bpls[,i]))
        Marc_sign=names(model_svd$Bpls[model_svd$Bpls[,i]>Q_ms,i])

        if(length(Marc_sign)!=0){
                points(model_svd$P[Marc_sign,1]*0.5,
                model_svd$P[Marc_sign,2]*0.5,
                col = i+6,
                pch = 19,
                cex=0.3)

                points(model_svd$C[i,1],
                model_svd$C[i,2],cex=0.1)
```

```
            text(model_svd$C[i,1],
            model_svd$C[i,2],
            row.names(model_svd$C)[i],
            cex = 0.8, col=i+6)
        }
}
```

Código en R para la visualización de la técnica PLS con NIPALS a través del triplot

```
model <- pls(X,Y, scale=TRUE, ncomp=2, mode="regression")
tiempo1 <- system.time(pls(X,Y, scale=TRUE, ncomp=2, mode="
    regression"))

cargas=plotVar(model, comp = c(1,2),
var.names = TRUE,
legend = TRUE)

indiv=plotIndiv(model, comp = c(1,2),
rep.space = "XY-variate",
title = "Biplot de PLS (mixOmics)",
legend = TRUE)


# Extraer cargas
load_X <- as.data.frame(model$loadings$X[, 1:2])
load_X$type <- "X"
load_X$varname <- rownames(load_X)

load_Y <- as.data.frame(model$loadings$Y[, 1:2])
load_Y$type <- "Y"
load_Y$varname <- rownames(load_Y)

colnames(load_X)[1:2] <- colnames(load_Y)[1:2] <- c("comp1"
    , "comp2")

# Extraer scores
scores <- as.data.frame(model$variates$X[, 1:2])
scores$type <- "Variedad"
scores$varname <- rownames(scores)
colnames(scores)[1:2] <- c("comp1", "comp2")

# Unir todo en un solo dataframe
df_all <- rbind(load_X, load_Y, scores)

# Crear paleta de colores por tipo
colores <- c("X" = "black", "Y" = "gray", "Variedad" = "
    purple")
df_all$col <- colores[df_all$type]

plot(df_all$comp1[1:ncol(X)], df_all$comp2[1:ncol(X)],
col = df_all$col[1:ncol(X)],
xlab = "Componente 1",
```

```
ylab = "Componente␣2",
cex = 0.1)

points(df_all$comp1[(ncol(X)+3):nrow(df_all)],
df_all$comp2[(ncol(X)+3):nrow(df_all)],
col = df_all$col[(ncol(X)+3):nrow(df_all)],
xlab = "Componente␣1",
ylab = "Componente␣2",
cex = 0.1,
xlim=c(-1.1,1.1),
ylim=c(-1.1,1.1))

# Ejes de referencia
abline(h = 0, lty = 2, col = "black")
abline(v = 0, lty = 2, col = "black")

# Agregar etiquetas
text(df_all$comp1[1:ncol(X)], df_all$comp2[1:ncol(X)],
labels = df_all$varname[1:ncol(X)],
cex = 0.6, pos = 3, col = df_all$col)

text(df_all$comp1[(ncol(X)+3):nrow(df_all)],
df_all$comp2[(ncol(X)+3):nrow(df_all)],
labels = df_all$varname[(ncol(X)+3):nrow(df_all)],
cex = 0.6, pos = 3, col = df_all$col[(ncol(X)+3):nrow(df_
   all)])

# Predicciones de Y
B=predict(model, X)$B.hat[,,2]

# configuration
s = 100000 # number of samples
# components
for(i in 1:2){
        Q_ms=qnorm(0.99, mean(B[,i]), sd(B[,i]))

        print(length(B[B[,i]>Q_ms,i]))
}

for(i in 1:2){
        Q_ms=qnorm(0.99, mean(B[,i]), sd(B[,i]))
        Marc_sign=names(B[B[,i]>Q_ms,i])
        if(length(Marc_sign)!=0){
                df_sign <- df_all[df_all$varname %in% Marc_
                   sign & df_all$type == "X", ]

                points(df_sign$comp1, df_sign$comp2,
                col = i+6,
                pch = 19,
                cex = 0.3)

                points(df_all[ncol(X)+i,1]*0.8,
                df_all[ncol(X)+i,2]*0.8,cex=0.1)
                text(df_all[ncol(X)+i,1]*0.8,
                df_all[ncol(X)+i,2]*0.8,
                row.names(df_all)[ncol(X)+i],
```

```
                    cex = 0.8, col=i+6)
            }
        }
```

Anexo 3

5.4.1. Código de programación en R para obtener los valores de RECM en cada escenario de simulación de datos faltantes a través de validación cruzada en la base de datos simulada

```
Y_predicha_N_test=list()
Y_real_N_test=list()

Y_predicha_N_train=list()
Y_real_N_train=list()

k_fold=10
for(i in 1:k_fold){

        n_total=nrow(SNP_completo)
        n_train <- round(0.85 * n_total)

        indices_train <- sample(1:n_total, n_train, replace = T)

        X_train <- SNP_completo[indices_train,]
        Y_train <- Y[indices_train,]
        X_test <- SNP_completo[-indices_train,]
        Y_test <- Y[-indices_train,]

        rownames(X_train) <- make.unique(rownames(X_train))
        rownames(Y_train) <- make.unique(rownames(Y_train))
        rownames(X_test) <- make.unique(rownames(X_test))
        rownames(Y_test) <- make.unique(rownames(Y_test))

        X_train=as.matrix(X_train)
        Y_train=as.matrix(Y_train)
        X_test=as.matrix(X_test)
        Y_test=as.matrix(Y_test)

        # Build the model on training set

        model <- mixOmics::pls(X_train,Y_train, scale=TRUE, ncomp
           =2,
        mode="regression")

        # Predecir Y esperado test
        Y_hat <- predict(model, newdata = X_test)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred
```

```
        Y_predicha_N_test=list.append(Y_predicha_N_test, ExpectedY)
        Y_real_N_test=list.append(Y_real_N_test, Y_test)

        # Predecir Y esperado train
        Y_hat <- predict(model, newdata = X_train)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_N_train=list.append(Y_predicha_N_train,
           ExpectedY)
        Y_real_N_train=list.append(Y_real_N_train, Y_train)

}

Y_predicha_T_r <- Reduce("+", Y_predicha_N_train) / length(Y_
   predicha_N_train)
Y_real_T_r <- Reduce("+", Y_real_N_train) / length(Y_real_N_train)


RECM_T=matrix(0,k_fold,ncol(Y))

for(i in 1:10){
        RECM_T[i,]=sqrt(colMeans((Y_real_N_test[[i]] - Y_predicha_N
           _test[[i]])^2))
}


#PRIMER ESCENARIO: ELIMINO 5% DE LOS DATOS ALEATORIAMENTE EN LA
   BASE#

Y_predicha_05_test=list()
Y_real_05_test=list()

Y_predicha_05_train=list()
Y_real_05_train=list()

k_fold=10
for(i in 1:k_fold){
        n_train <- round(0.85 * n_total)

        indices_train <- sample(1:n_total, n_train, replace = TRUE)

        X_train <- SNP_completo[indices_train,]
        Y_train <- Y[indices_train,]
        X_test <- SNP_completo[-indices_train,]
        Y_test <- Y[-indices_train,]

        X_train=as.matrix(X_train)
        Y_train=as.matrix(Y_train)
        X_test=as.matrix(X_test)
        Y_test=as.matrix(Y_test)

        rownames(X_train) <- make.unique(rownames(X_train))
        rownames(Y_train) <- make.unique(rownames(Y_train))
```

```
        rownames(X_test) <- make.unique(rownames(X_test))
        rownames(Y_test) <- make.unique(rownames(Y_test))

        total_elements <- nrow(X_train) * ncol(X_train)
        sample_size <- round(0.05 * total_elements)


        random_indices <- sample(1:total_elements, sample_size)

        mat <- as.matrix(X_train)

        mat[random_indices] <- NaN

        # Convertir la matriz de vuelta a un data frame
        SNP_na_1_5 <- as.data.frame(mat)

        # Build the model on training set

        model <- pls(SNP_na_1_5,Y_train, scale=T,
        ncomp=2, mode="regression")


        # Predecir Y esperado test
        Y_hat <- predict(model, newdata = X_test)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_05_test=list.append(Y_predicha_05_test,
           ExpectedY)
        Y_real_05_test=list.append(Y_real_05_test, Y_test)

        # Predecir Y esperado train
        Y_hat <- predict(model, newdata = SNP_na_1_5)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_05_train=list.append(Y_predicha_05_train,
           ExpectedY)
        Y_real_05_train=list.append(Y_real_05_train, Y_train)
}

Y_predicha_5_r <- Reduce("+", Y_predicha_05_train) / length(Y_
   predicha_05_train)
Y_real_5_r <- Reduce("+", Y_real_05_train) / length(Y_real_05_train
   )

#SEGUNDO ESCENARIO: ELIMINO 10% DE LOS DATOS ALEATORIAMENTE EN LA
   BASE#

Y_predicha_10_train=list()
Y_real_10_train=list()

Y_predicha_10_test=list()
Y_real_10_test=list()
```

```
k_fold=10
for(i in 1:k_fold){
        n_train <- round(0.85 * n_total)

        indices_train <- sample(1:n_total, n_train, replace = TRUE)

        X_train <- SNP_completo[indices_train,]
        Y_train <- Y[indices_train,]
        X_test <- SNP_completo[-indices_train,]
        Y_test <- Y[-indices_train,]

        X_train=as.matrix(X_train)
        Y_train=as.matrix(Y_train)

        rownames(X_train) <- make.unique(rownames(X_train))
        rownames(Y_train) <- make.unique(rownames(Y_train))

        total_elements <- nrow(X_train) * ncol(X_train)
        sample_size <- round(0.10 * total_elements)


        random_indices <- sample(1:total_elements, sample_size)

        mat <- as.matrix(X_train)

        mat[random_indices] <- NaN

        # Convertir la matriz de vuelta a un data frame
        SNP_na_1_10 <- as.data.frame(mat)

        # Build the model on training set

        model <- pls(SNP_na_1_10,Y_train, scale=T,
        ncomp=2, mode="regression")


        X_test=as.matrix(X_test)
        Y_test=as.matrix(Y_test)

        rownames(X_test) <- make.unique(rownames(X_test))
        rownames(Y_test) <- make.unique(rownames(Y_test))

        # Predecir Y esperado test
        Y_hat <- predict(model, newdata = X_test)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_10_test=list.append(Y_predicha_10_test,
           ExpectedY)
        Y_real_10_test=list.append(Y_real_10_test, Y_test)

        # Predecir Y esperado train
        Y_hat <- predict(model, newdata = SNP_na_1_10)
        Y_pred <- Y_hat$predict[, , model$ncomp]
```

```
        ExpectedY=Y_pred

        Y_predicha_10_train=list.append(Y_predicha_10_train,
           ExpectedY)
        Y_real_10_train=list.append(Y_real_10_train, Y_train)

}

Y_predicha_10_r <- Reduce("+", Y_predicha_10_train) / length(Y_
   predicha_10_train)
Y_real_10_r <- Reduce("+", Y_real_10_train) / length(Y_real_10_
   train)

#TERCER ESCENARIO: ELIMINO 25% DE LOS DATOS ALEATORIAMENTE EN LA
   BASE#

Y_predicha_25_test=list()
Y_real_25_test=list()

Y_predicha_25_train=list()
Y_real_25_train=list()

k_fold=10
for(i in 1:k_fold){
        n_train <- round(0.85 * n_total)

        indices_train <- sample(1:n_total, n_train, replace = TRUE)

        X_train <- SNP_completo[indices_train,]
        Y_train <- Y[indices_train,]
        X_test <- SNP_completo[-indices_train,]
        Y_test <- Y[-indices_train,]

        X_train=as.matrix(X_train)
        Y_train=as.matrix(Y_train)

        rownames(X_train) <- make.unique(rownames(X_train))
        rownames(Y_train) <- make.unique(rownames(Y_train))

        total_elements <- nrow(X_train) * ncol(X_train)
        sample_size <- round(0.25 * total_elements)


        random_indices <- sample(1:total_elements, sample_size)

        mat <- as.matrix(X_train)

        mat[random_indices] <- NaN

        # Convertir la matriz de vuelta a un data frame
        SNP_na_1_25 <- as.data.frame(mat)

        # Build the model on training set

        model <- pls(SNP_na_1_25,Y_train, scale=T,
```

```
        ncomp=2, mode="regression")


        X_test=as.matrix(X_test)
        Y_test=as.matrix(Y_test)

        rownames(X_test) <- make.unique(rownames(X_test))
        rownames(Y_test) <- make.unique(rownames(Y_test))

        # Predecir Y esperado test
        Y_hat <- predict(model, newdata = X_test)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_25_test=list.append(Y_predicha_25_test,
           ExpectedY)
        Y_real_25_test=list.append(Y_real_25_test, Y_test)

        # Predecir Y esperado train
        Y_hat <- predict(model, newdata = SNP_na_1_25)
        Y_pred <- Y_hat$predict[, , model$ncomp]

        ExpectedY=Y_pred

        Y_predicha_25_train=list.append(Y_predicha_25_train,
           ExpectedY)
        Y_real_25_train=list.append(Y_real_25_train, Y_train)


}

Y_predicha_25_r <- Reduce("+", Y_predicha_25_train) / length(Y_
   predicha_25_train)
Y_real_25_r <- Reduce("+", Y_real_25_train) / length(Y_real_25_
   train)

RECM_25=matrix(0,k_fold,ncol(Y))

for(i in 1:10){
        RECM_25[i,]=sqrt(colMeans((Y_real_25_test[[i]] - Y_predicha
           _25_test[[i]])^2))
}

#CUARTO ESCENARIO: ELIMINO 35% DE LOS DATOS ALEATORIAMENTE EN LA
   BASE#

Y_predicha_35_train=list()
Y_real_35_train=list()

Y_predicha_35_test=list()
Y_real_35_test=list()

k_fold=10
for(i in 1:k_fold){
        n_train <- round(0.85 * n_total)
```

```
indices_train <- sample(1:n_total, n_train, replace = TRUE)

X_train <- SNP_completo[indices_train,]
Y_train <- Y[indices_train,]
X_test <- SNP_completo[-indices_train,]
Y_test <- Y[-indices_train,]

X_train=as.matrix(X_train)
Y_train=as.matrix(Y_train)

rownames(X_train) <- make.unique(rownames(X_train))
rownames(Y_train) <- make.unique(rownames(Y_train))

total_elements <- nrow(X_train) * ncol(X_train)
sample_size <- round(0.35 * total_elements)


random_indices <- sample(1:total_elements, sample_size)

mat <- as.matrix(X_train)

mat[random_indices] <- NaN

# Convertir la matriz de vuelta a un data frame
SNP_na_1_35 <- as.data.frame(mat)

# Build the model on training set

model <- pls(SNP_na_1_35,Y_train, scale=T,
ncomp=2, mode="regression")


X_test=as.matrix(X_test)
Y_test=as.matrix(Y_test)

rownames(X_test) <- make.unique(rownames(X_test))
rownames(Y_test) <- make.unique(rownames(Y_test))

# Predecir Y esperado test
Y_hat <- predict(model, newdata = X_test)
Y_pred <- Y_hat$predict[, , model$ncomp]

ExpectedY=Y_pred

Y_predicha_35_test=list.append(Y_predicha_35_test,
   ExpectedY)
Y_real_35_test=list.append(Y_real_35_test, Y_test)

# Predecir Y esperado train
Y_hat <- predict(model, newdata = SNP_na_1_35)
Y_pred <- Y_hat$predict[, , model$ncomp]

ExpectedY=Y_pred

Y_predicha_35_train=list.append(Y_predicha_35_train,
```

```
            ExpectedY)
        Y_real_35_train=list.append(Y_real_35_train, Y_train)

}

Y_predicha_35_r <- Reduce("+", Y_predicha_35_train) / length(Y_
    predicha_35_train)
Y_real_35_r <- Reduce("+", Y_real_35_train) / length(Y_real_35_
    train)


RECM_35=matrix(0,k_fold,ncol(Y))

for(i in 1:10){
        RECM_35[i,]=sqrt(colMeans((Y_real_35_test[[i]] - Y_predicha
            _35_test[[i]])^2))
}


par(mfcol=c(1,2))
for(i in 1:2){
        boxplot(RECM_T[,i], RECM_5[,i],RECM_10[,i],
        RECM_25[,i],RECM_35[,i],
        ylim=c(0.85,1.25),
        names = c("0% NA",
        "5% NA",
        "10% NA",
        "25% NA",
        "35% NA"),
        main = '',ylab='RECM')
}



# Marcadores significativos #

model <- pls(SNP_completo ,Y_scale,
scale=TRUE, ncomp=2, mode="regression")

# Predicciones de Y
B=predict(model, SNP_completo)$B.hat[,,2]

Y_hat <- predict(model, newdata = SNP_completo)
Y_predicha_T <- Y_hat$predict[, , model$ncomp]

Q_ms=qnorm(0.99, mean(B[,1]), sd(B[,1]))
MS_1=names(B[B[,1]>Q_ms,1])
Q_ms=qnorm(0.99, mean(B[,2]), sd(B[,2]))
MS_2=names(B[B[,2]>Q_ms,2])


# 5%

total_elements <- nrow(SNP_completo) * ncol(SNP_completo)
sample_size <- round(0.05 * total_elements)
```

```r
random_indices <- sample(1:total_elements, sample_size)

mat <- as.matrix(SNP_completo)

mat[random_indices] <- NaN

# Convertir la matriz de vuelta a un data frame
SNP_na_1_5 <- as.data.frame(mat)

# Build the model on training set

model <- pls(SNP_na_1_5,Y_scale,
scale=TRUE, ncomp=2, mode="regression")

# Predicciones de Y
B=predict(model, SNP_na_1_5)$B.hat[,,2]

Y_hat <- predict(model, newdata = SNP_na_1_5)
Y_predicha_05 <- Y_hat$predict[, , model$ncomp]


s = 100000 # number of samples
# components
for(i in 1:2){
        Q_ms=qnorm(0.99, mean(B[,i]), sd(B[,i]))
        print(colnames(Y)[i])
        #print(B[B[,i]>Q_ms,i])
        print(length(B[B[,i]>Q_ms,i]))
}

Q_ms=qnorm(0.99, mean(B[,1]), sd(B[,1]))
MS_1_0.05=names(B[B[,1]>Q_ms,1])
Q_ms=qnorm(0.99, mean(B[,2]), sd(B[,2]))
MS_2_0.05=names(B[B[,2]>Q_ms,2])


# 10%

total_elements <- nrow(SNP_completo) * ncol(SNP_completo)
sample_size <- round(0.10 * total_elements)

random_indices <- sample(1:total_elements, sample_size)

mat <- as.matrix(SNP_completo)

mat[random_indices] <- NaN

# Convertir la matriz de vuelta a un data frame
SNP_na_1_10 <- as.data.frame(mat)

# Build the model on training set

model <- pls(SNP_na_1_10,Y_scale,
scale=TRUE, ncomp=2, mode="regression")

# Predicciones de Y
```

```
B=predict(model, SNP_na_1_10)$B.hat[,,2]

Y_hat <- predict(model, newdata = SNP_na_1_10)
Y_predicha_10 <- Y_hat$predict[, , model$ncomp]

s = 100000 # number of samples
# components
for(i in 1:2){
        Q_ms=qnorm(0.99, mean(B[,i]), sd(B[,i]))
        print(colnames(Y)[i])
        #print(B[B[,i]>Q_ms,i])
        print(length(B[B[,i]>Q_ms,i]))
}

Q_ms=qnorm(0.99, mean(B[,1]), sd(B[,1]))
MS_1_0.10=names(B[B[,1]>Q_ms,1])
Q_ms=qnorm(0.99, mean(B[,2]), sd(B[,2]))
MS_2_0.10=names(B[B[,2]>Q_ms,2])


# 25%

total_elements <- nrow(SNP_completo) * ncol(SNP_completo)
sample_size <- round(0.25 * total_elements)

random_indices <- sample(1:total_elements, sample_size)

mat <- as.matrix(SNP_completo)

mat[random_indices] <- NaN

# Convertir la matriz de vuelta a un data frame
SNP_na_1_25 <- as.data.frame(mat)

# Build the model on training set

model <- pls(SNP_na_1_25,Y_scale,
scale=TRUE, ncomp=2, mode="regression")

# Predicciones de Y
B=predict(model, SNP_na_1_25)$B.hat[,,2]

Y_hat <- predict(model, newdata = SNP_na_1_25)
Y_predicha_25 <- Y_hat$predict[, , model$ncomp]

s = 100000 # number of samples
# components
for(i in 1:2){
        Q_ms=qnorm(0.99, mean(B[,i]), sd(B[,i]))
        print(colnames(Y)[i])
        #print(B[B[,i]>Q_ms,i])
        print(length(B[B[,i]>Q_ms,i]))
}

Q_ms=qnorm(0.99, mean(B[,1]), sd(B[,1]))
MS_1_0.25=names(B[B[,1]>Q_ms,1])
```

```r
Q_ms=qnorm(0.99, mean(B[,2]), sd(B[,2]))
MS_2_0.25=names(B[B[,2]>Q_ms,2])


# 35

total_elements <- nrow(SNP_completo) * ncol(SNP_completo)
sample_size <- round(0.35 * total_elements)

random_indices <- sample(1:total_elements, sample_size)

mat <- as.matrix(SNP_completo)

mat[random_indices] <- NaN

# Convertir la matriz de vuelta a un data frame
SNP_na_1_35 <- as.data.frame(mat)

# Build the model on training set

model <- pls(SNP_na_1_35,Y_scale,
scale=TRUE, ncomp=2, mode="regression")

# Predicciones de Y
B=predict(model, SNP_na_1_35)$B.hat[,,2]

Y_hat <- predict(model, newdata = SNP_na_1_35)
Y_predicha_35 <- Y_hat$predict[, , model$ncomp]

s = 100000 # number of samples
# components
for(i in 1:2){
        Q_ms=qnorm(0.99, mean(B[,i]), sd(B[,i]))
        print(colnames(Y)[i])
        #print(B[B[,i]>Q_ms,i])
        print(length(B[B[,i]>Q_ms,i]))
}

Q_ms=qnorm(0.99, mean(B[,1]), sd(B[,1]))
MS_1_0.35=names(B[B[,1]>Q_ms,1])
Q_ms=qnorm(0.99, mean(B[,2]), sd(B[,2]))
MS_2_0.35=names(B[B[,2]>Q_ms,2])
```

5.4.2. Código en R para el Análisis Procrustes

```r
#procrustes
par(mfcol=c(2,2))
procrustes.results <- vegan::procrustes(Y_predicha_T,
Y_predicha_05)
#plot(procrustes.results, kind=1)

# Extraemos las coordenadas alineadas
X_proc <- procrustes.results$X
Y_proc <- procrustes.results$Yrot
```

```
plot(X_proc,
xlim = c(-2.5, 2.5), ylim = c(-2.5, 2),
pch = 19, col = "blue",
xlab = "Dim␣1", ylab = "Dim␣2")
#     main = "5% valores faltantes")

# Agregamos los puntos rotados
points(Y_proc, pch = 19, col = "red")

segments(X_proc[,1], X_proc[,2], Y_proc[,1], Y_proc[,2],
col = "gray60")

# Etiquetas (ajustables)
text(X_proc, labels = rownames(X_proc), col = "blue", cex = 0.6,
   pos = 3)
text(Y_proc, labels = rownames(Y_proc), col = "red", cex = 0.6, pos
    = 1)

abline(h = 0, v = 0, col = "gray80", lty = 2)

procrustes.results <- vegan::procrustes(Y_predicha_T,
Y_predicha_10)
#plot(procrustes.results, kind=1)

# Extraemos las coordenadas alineadas
X_proc <- procrustes.results$X
Y_proc <- procrustes.results$Yrot

plot(X_proc,
xlim = c(-2.5, 2.5), ylim = c(-2.5, 2),
pch = 19, col = "blue",
xlab = "Dim␣1", ylab = "Dim␣2")
#     main = "10% valores faltantes")

# Agregamos los puntos rotados
points(Y_proc, pch = 19, col = "red")

segments(X_proc[,1], X_proc[,2], Y_proc[,1], Y_proc[,2],
col = "gray60")

# Etiquetas (ajustables)
text(X_proc, labels = rownames(X_proc), col = "blue", cex = 0.6,
   pos = 3)
text(Y_proc, labels = rownames(Y_proc), col = "red", cex = 0.6, pos
    = 1)

abline(h = 0, v = 0, col = "gray80", lty = 2)
```