

EVALUACIÓN DE LA SELECCIÓN  
GENÓMICA DE UN SOLO PASO  
EN ARROZ

Esta tesis se escribió usando los paquetes de R (R) Markdown, L<sup>A</sup>T<sub>E</sub>X , bookdown y amsterdown.



Una versión en línea de esta tesis está disponible en [https://github.com/Leo4Luffy/TFM\\_UAB](https://github.com/Leo4Luffy/TFM_UAB), bajo la licencia Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



# Evaluación de la selección genómica de un solo paso en arroz

Tesis académica para obtener  
el grado de Máster en Mejora Genética y  
Biología Molecular de la Reproducción bajo la  
dirección del prof. dr. Miguel Pérez Enciso  
ante una comisión constituida por la Junta del Máster,  
para ser defendido en público el  
Colocar aquí la fecha de la defensa, a las colocar la hora aquí

Jorge Leonardo López Martínez



**Dirección:**

Director: prof. dr. M. Pérez-Enciso Centre for Research in Agricultural Genomics

# Índice general

<b>1. Resumen / Summary</b>	<b>1</b>
<b>2. Revisión de literatura</b>	<b>3</b>
2.1. Breve historia hacia la selección genómica . . . . .	3
2.2. La selección genómica . . . . .	6
2.3. Diferencias de la mejora genética en animales y en plantas . .	16
2.4. La mejora genética en el arroz . . . . .	19
<b>3. Objetivo</b>	<b>23</b>
<b>4. Materiales y métodos</b>	<b>24</b>
4.1. Recurso vegetal y datos fenotípicos . . . . .	24
4.2. Predicción basada en información del pedigrí e información genómica . . . . .	26
4.3. Precisión de la predicción mediante simulación del pedigrí ancestral . . . . .	27
4.4. Precisión de la predicción mediante simulación de descendientes	30
<b>5. Resultados</b>	<b>34</b>
5.1. Estructura poblacional y heredabilidad en la simulación ancestral . . . . .	34
5.2. Precisión de la predicción mediante simulación ancestral . . .	37
5.3. Precisión de la predicción mediante simulación de descendientes	39
<b>6. Discusión</b>	<b>41</b>
<b>7. Conclusiones</b>	<b>43</b>
<b>A. Anexos</b>	<b>44</b>
A.1. Función para el calculo de la matriz de parentesco combinada	44
A.2. Visualización del GWAS . . . . .	46
A.3. Precisión de la predicción mediante simulación de pedigrí ancestral . . . . .	46
A.4. Precisión de la predicción mediante simulación de descendientes	48

*ÍNDICE GENERAL*

VI

**Bibliografía**

**51**

**Agradecimientos**

**57**

# Capítulo 1

## Resumen / Summary

En los modelos de predicción genómica convencionales se utilizan solo los individuos con genotipo. El método del mejor predictor lineal insesgado genómico de un solo paso (ssGBLUP) permite integrar individuos genotipados y no genotipados, siempre que estén conectados por un pedigree. El objetivo de este estudio consistió en evaluar la utilidad de esta metodología de selección en la predicción genómica del arroz. Se usaron datos del tiempo de floración de 451 individuos de arroz de la variedad Indica genotipada con 100.231 SNP. Debido a la falta del pedigree, se llevó a cabo simulaciones de dos tipos, hacia atrás (ancestral) y hacia adelante (descendientes). En la simulación ancestral, se usó el software Molcoanc para crear antepasados virtuales a partir de los 451 individuos genotipados. En la simulación de descendientes, se empleó el paquete de Python SeqBreed para generar descendencia de esos 451 individuos, repartidos en distintas generaciones (F1, F2 y F3). Se evaluaron distintos escenarios de densidad de marcadores (ninguno, 1.000, 10.000 y 100.000) y de individuos genotipados según el tipo de simulación: (i) simulación ancestral con ninguno, 148, 298 y 451 individuos genotipados, y (ii) simulación de descendientes con ningún individuo, solo la generación F2, la generación F2 y F3, y todos los individuos genotipados. Se usó el coeficiente de correlación entre los valores fenotípicos observados y predichos como medida de predictibilidad. Los resultados revelaron que, en ambos tipos de simulación, la precisión de la predicción mediante el ssGBLUP tiende a mejorar comparado con el mejor predictor lineal insesgado convencional (BLUP) que utiliza únicamente la información del pedigree. En la simulación ancestral se observaron ganancias esperadas en la precisión por encima del 1.0 % con respecto al BLUP (al considerar una densidad del marcador de 10.000 en 298 individuos genotipados) y entre el 13 % al 18 % (al considerar una densidad del marcador de 10.000 en 451 individuos genotipados). En cuanto a la simulación de descendientes, también se pudo observar precisiones más altas al realizar la predicción mediante el ssGBLUP. Los análisis sugieren utilizar una cantidad de marcadores en-

tre 1.000 y 10.000 para obtener una precisión equivalente a usar todos los marcadores disponibles. Por lo tanto, el uso combinado de la información genómica y del pedigree mediante el método ssGBLUP puede ser un enfoque interesante para aumentar la precisión de la predicción en los programas de mejora genética en los cultivos de arroz.

**Palabras clave:** matriz de parentesco, parámetros genéticos, precisión de la predicción, predicción genómica.

---

In traditional genomic prediction models, only individuals with the genotype are used. The single-step genomic best linear unbiased predictor (ssGBLUP) method allows integrating genotyped and non-genotyped individuals, provided they are connected by a pedigree. The objective of this study was to evaluate the use of this selection methodology in the genomic prediction of rice. Flowering time data from 451 Indica rice individuals genotyped with 100.231 SNPs were used. Due to the absence of the pedigree, two types of simulations were carried out, backwards (ancestor) and forwards (offspring). In the ancestry simulation, Molcoanc software was used to create virtual ancestors from the 451 genotyped individuals. In the offspring simulation, the SeqBreed Python package was used to generate offspring from these 451 individuals, divided into different generations (F1, F2 and F3). Different marker density (none, 1.000, 10.000 and 100.000) and genotyped individuals scenarios were evaluated according to the type of simulation: (i) ancestors simulation with none, 148, 298 and 451 individuals genotyped, and (ii) offspring simulation with none individuals, only F2 generation, the F2 and F3 generation, and all the individuals genotyped. The correlation coefficient between the observed and predicted phenotypic values was used as a measure of predictive ability. The results revealed that in both types of simulation, the prediction accuracy using the ssGBLUP is better compared to the conventional best linear unbiased prediction (BLUP) that uses only the pedigree information. In the ancestors simulation, expected gains in precision were observed above 1.0 % with respect to the BLUP (considering a marker density of 10.000 in 298 genotyped individuals) and between 13 % and 18 % (considering a marker density of 10.000 in 451 genotyped individuals). In the offspring simulation, higher precisions were also observed when making the prediction using ssGBLUP. The analyzes suggest using a number of markers between 1.000 and 10.000, to obtain an accuracy equivalent to using all available markers. Therefore, the combined use of genomic and pedigree information in the ssGBLUP method may be an interesting approach to improvement prediction accuracy in rice breeding programs.

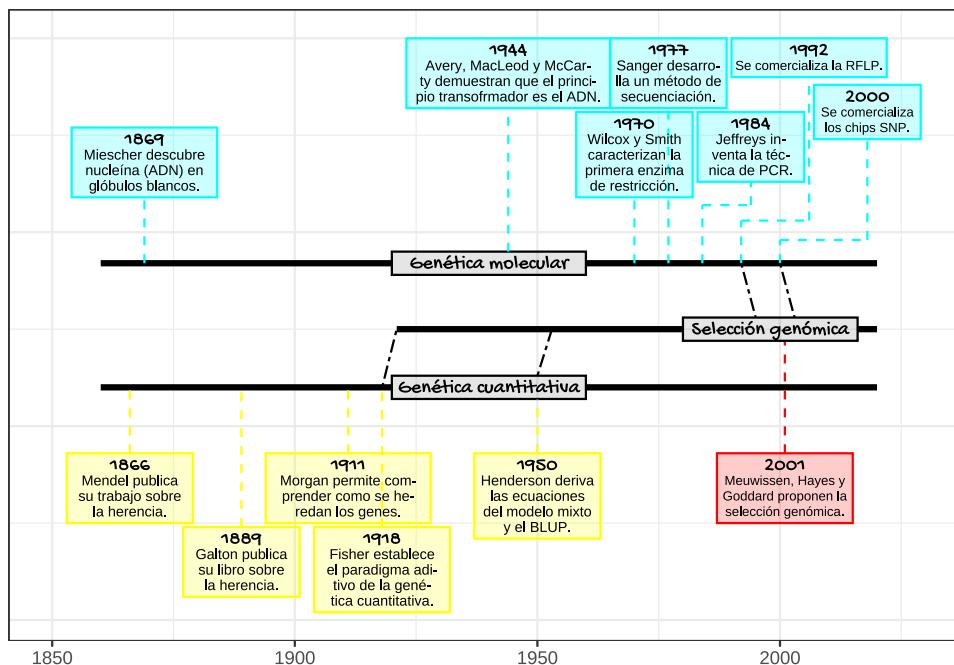
**Key words:** relationship matrix, genetic parameters, prediction accuracy, genomic prediction.

# Capítulo 2

## Revisión de literatura

### 2.1. Breve historia hacia la selección genómica

La historia de la genética tanto cuantitativa como molecular se debe a la contribución de muchas personas (Figura 2.1), hecho que permitió la conexión entre ambas disciplinas y el desarrollo de lo que hoy en día se conoce como selección genómica.



**Figura 2.1:** Cronología de las disciplinas de la genética molecular y la genética cuantitativa. Varios descubrimientos permitieron la conexión entre ambas disciplinas lo que permitió el desarrollo de la selección genómica. Figura adaptada de Nelson, Pettersson, y Carlborg (2012).

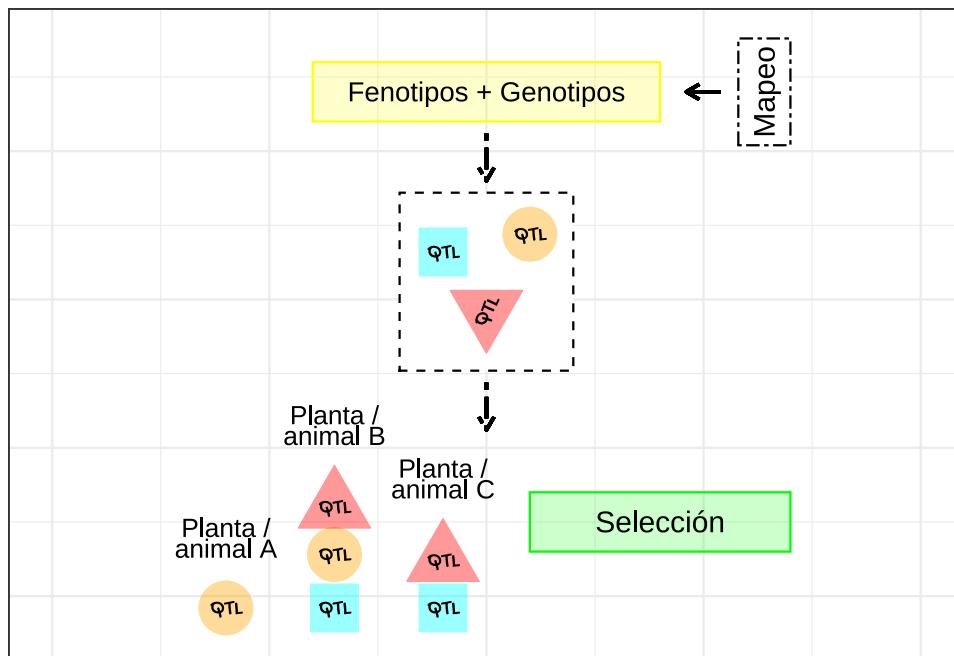
La genética cuantitativa se originó hace más de un siglo en ausencia directa de datos genéticamente observables (Nelson, Pettersson, y Carlborg 2012). Uno de los fundadores más influyentes fue Ronald Fisher, quien proporcionó una teoría que hizo posible interpretar los descubrimientos de la genética biométrica dentro de los principios de herencia Mendeliana, permitiendo con ello unificar las escuelas de pensamiento Mendeliano y biométrico que para ese entonces estaban en constante debate. Dicha teoría, denominada como teoría del modelo infinitesimal, asume que la herencia genética es principalmente aditiva, y que la varianza genética de un carácter está determinada por un gran número de factores ‘Mendelianos’ (hoy en día conocidos como genes), cada uno de los cuales tiene una pequeña contribución al fenotipo del carácter (Nelson, Pettersson, y Carlborg 2012; Turelli 2017).

Otro concepto fundamental, sobre todo en mejora genética, es el valor de cría estimado (EBV), definido como el efecto genético que un individuo posee y que puede transmitir a su descendencia. Este se puede predecir en función de un modelo que relaciona el fenotipo de una población con la información de pedigree mediante el uso del mejor predictor lineal insesgado (BLUP) (Henderson 1986). Este procedimiento fue resultado del esfuerzo de Charles Roy Henderson, quien a inicio de la década de 1950 contribuyó a su desarrollo (Freeman 1991; Searle 1991; Schaeffer 1991). A pesar que desde entonces el BLUP fue el método más utilizado para la mejora genética tanto en animales como en plantas, dicho procedimiento ignora la base física de la herencia (el ADN).

A partir de los años 60 hubo un rápido desarrollo de la genética molecular. Esta disciplina permitió, a diferencia de la genética cuantitativa, estudiar de forma directa el gen, lo que facilitó a finales de la década de 1970 e inicio de 1980 el descubrimiento de secuencias variables de ADN con fenotipos fácilmente observables (Legarra, Lourenco, y Vitezica 2018). Son ejemplo de estas secuencias (denominadas como marcadores de ADN) los microsatélites, los polimorfismos en el tamaño de los fragmentos de restricción (RFLP) y los polimorfismos de un sólo nucleótido (SNP), siendo este último hoy en día el principal marcador utilizado para detectar variaciones en el ADN.

Dichos marcadores de ADN, al representar las diferencias en el ADN heredado por dos individuos, abrieron la posibilidad de obtener una predicción más precisa de los EBV (Misztal, Aggrrey, y Muir 2012; de los Campos et al. 2013), comparado al método BLUP mencionado en párrafos anteriores. Según los mismos autores (de los Campos et al. 2013), los primeros intentos de integrar datos de marcadores de ADN en las predicciones se basaron en el supuesto de que era posible encontrar genes que contribuyeran a la variación genética del carácter, en lo que se denominó como selección asistida por marcadores (MAS) (Blasco y Toro 2014). En la MAS, se supone los individuos portadores de un marcador de ADN deseado podían ser identificados

y seleccionados para aumentar la respuesta genética de caracteres cuantitativos de relevancia económica (Kyselova, Tichý, y Jochová 2021). Blasco y Toro (2014) describen la MAS como un proceso en el cual se detectan genes que afectan directamente un carácter (QTL), que al ser seleccionados, logran una mejora genética al aumentar su frecuencia (Figura 2.2).



**Figura 2.2:** Esquema de la MAS. En la MAS, los fenotipos y genotipos de la población de mapeo se analizan usando un modelo estadístico, identificando con ello relaciones significativas entre fenotipos y genotipos. Por último, se seleccionan los individuos favorables con base en datos de genotipo. Figura adaptada de Nakaya y Isobe (2012).

Si bien la MAS abrió la posibilidad de investigar la variación genética en animales y en plantas, permitiendo también identificar genes que afectaban el desempeño de caracteres económicoimportantes, la literatura científica coincide en afirmar lo limitado que fue esta metodología al no detectar marcadores de ADN con efectos genéticos menores (Blasco y Toro 2014; Desta y Ortiz 2014; Kyselova, Tichý, y Jochová 2021; Tong y Nikoloski 2021). Y es que, como es sabido, la mayoría de los caracteres económicoimportantes son cuantitativos y complejos, lo que quiere decir que son caracteres controlados por muchos genes de pequeño efecto y/o por una combinación de genes mayores y menores, lo que hace de la MAS un método poco adecuado para este tipo de arquitectura genética de caracteres.

Finalmente, en el año 2001, Theodorus Meuwissen, Ben Hayes y Michael Goddard presentaron una alternativa a la MAS, superando con ello las limitaciones que suponía el uso de esta metodología. A esta nueva alternativa se le dio el nombre de selección genómica. Solo fue cuestión de tiempo para que los datos obtenidos de la genética molecular se integraran a los modelos estadísticos de la genética cuantitativa, permitiendo así el análisis de caracteres complejos en el marco de efectos del modelo infinitesimal.

## 2.2. La selección genómica

Se denomina selección genómica a una serie de métodos que usan numerosos marcadores de ADN, principalmente SNP, para realizar la predicción del EBV (aunque en selección genómica es común referirse al EBV como valor de cría basado en marcadores de ADN o GEBV). Blasco y Toro (2014) y Ahmadi, Bartholomé, Cao, et al. (2020a) describen este método como un proceso en el cual se usan grandes cantidades de marcadores de ADN para construir un modelo de relaciones genotipo-fenotipo en una población de entrenamiento. Luego el modelo de selección genómica resultante se utiliza en una población de prueba que solo está genotipada, y se predice en ella el GEBV con el que se lleva a cabo la selección (Figura 2.3).



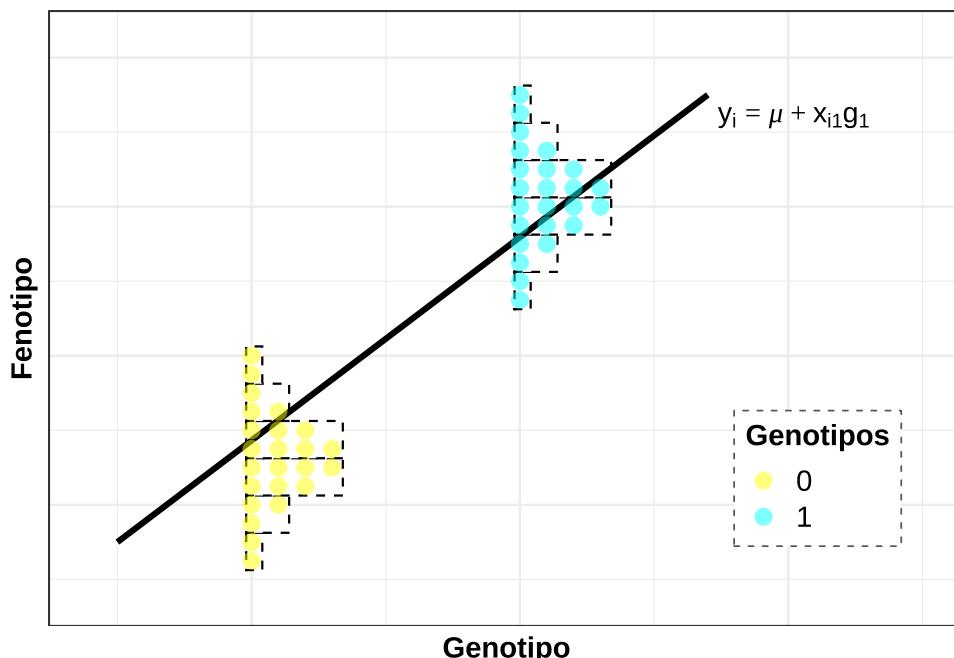
**Figura 2.3:** Esquema de la selección genómica. La selección genómica utiliza un modelo estadístico, diseñado a partir de datos genotípicos y fenotípicos en una población de entrenamiento, para predecir el GEBV de los individuos

en una población de prueba con datos genotípicos. Por último, los individuos se seleccionan de acuerdo a su GEBV. Figura adaptada de Tong y Nikoloski (2021).

El uso de decenas de miles de marcadores de ADN es una de las características fundamentales de la selección genómica (Desta y Ortiz 2014). Al contar con tal cantidad, la probabilidad de que algunos de estos marcadores estén en desequilibrio de ligamiento con el QTL tiende a aumentar (Meuwissen, Hayes, y Goddard 2001), con lo cual, aún cuando dichos marcadores no tienen efecto directo sobre el carácter, sí que se captura una asociación entre el carácter y los marcadores.

### 2.2.1. Métodos estadísticos en la selección genómica

En la selección genómica, la relación genotipo-fenotipo suele ser representada como un modelo lineal (Figura 2.4). Por tanto, el modelo de regresión lineal es un enfoque fundamental en la selección genómica (Nakaya y Isobe 2012; de los Campos et al. 2013; Crossa et al. 2017).



**Figura 2.4:** Relación genotipo-fenotipo de individuos (círculos amarillo y azul) para un solo marcador.  $Y_i$  y  $x_{i1}$  denotan los fenotipos y genotipos, y  $\mu$  y  $g_i$  son los parámetros a determinar. Los genotipos bialélicos se codifican como 0 y 1, y los fenotipos se distribuyen de acuerdo a una normal. Figura adaptada de Nakaya y Isobe (2012).

Dicha relación genotipo-fenotipo se puede expresar de la forma:

$$y_i = \mu + \sum_{j=1}^p x_{ij}g_j + e_i, \quad (2.1)$$

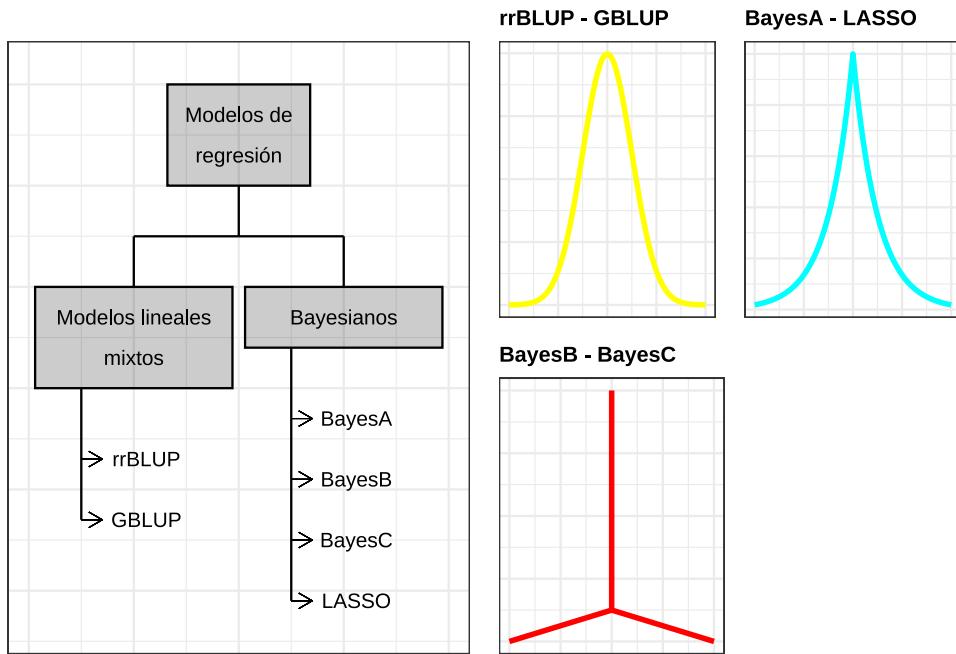
donde  $i$  ( $1, 2, 3, \dots, n$ ) representa a los individuos,  $j$  ( $1, 2, 3, \dots, p$ ) corresponde a los marcadores,  $y_i$  denota el fenotipo para el  $i$ -ésimo individuo,  $\mu$  corresponde a la media de la población,  $x_{ij}$  representa al genotipo del  $j$ -ésimo marcador en el  $i$ -ésimo individuo,  $g_j$  corresponde al efecto del  $j$ -ésimo marcador en el fenotipo, y  $e_i$  es el término del error.

Así mismo, el modelo anterior se puede expresar en notación matricial como:

$$\mathbf{y} = \mu + \mathbf{Z}\mathbf{g} + \mathbf{e}, \quad (2.2)$$

donde  $y$  es un vector de longitud igual al número de individuos ( $1, 2, 3, \dots, n$ ) que representa al fenotipo,  $Z$  es una matriz que indica si el marcador es homocigoto dominante, heterocigoto u homocigoto recesivo (por ejemplo, 2 si es homocigoto dominante, 1 si es heterocigoto y 0 si es homocigoto recesivo),  $g$  es un vector de efectos del marcador en el fenotipo (tratados aquí como efectos fijos), y  $e$  es el término del error. Luego el GEBV se puede predecir como  $\hat{g} = (Z'Z)^{-1}Z'y$  mediante mínimos cuadrados.

Sin embargo, con el uso decenas de miles de marcadores de ADN para predecir el GEBV, al emplear el modelo lineal en selección genómica puede surgir un problema conocido como  $p$  grande y  $n$  pequeño (Nakaya y Isobe 2012; de los Campos et al. 2013; Tong y Nikoloski 2021), hecho que impide el uso de la regresión por mínimos cuadrados, ya que este solo se puede aplicar en situaciones en las que el número de observaciones es mayor al número de variables o predictores. En tal sentido, la selección genómica brinda la oportunidad de enfrentar el problema del  $p$  grande y  $n$  pequeño por medio del uso de modelos de regresión lineal alternativos (Figura 2.5).



**Figura 2.5:** Enfoques estadísticos de la selección genómica. Cada uno de estos enfoques suponen distintas distribuciones de efectos de los marcadores sobre el carácter. Figura adaptada de de los Campos et al. (2013) y de Tong y Nikoloski (2021).

Al proponer la teoría de la selección genómica, Meuwissen, Hayes, y Goddard (2001) proporcionaron también una serie de métodos estadísticos como solución al problema planteado en el párrafo anterior, esto es, el mejor predictor lineal insesgado por ridge-regression (rrBLUP), y los métodos Bayesianos BayesA y BayesB.

En relación al rrBLUP, este se puede expresar como un modelo lineal mixto:

$$y = Xb + Zg + e, \quad (2.3)$$

donde  $y$ ,  $Z$  y  $e$  denotan los mismos términos del modelo (1.2),  $g$  es un vector de efectos del marcador en el fenotipo (tratados aquí como efectos aleatorios),  $X$  es una matriz que indica los efectos fijos y  $b$  es un vector de efectos fijos. Luego la predicción del GEBV se realiza a partir de  $\hat{g} = (Z'Z + I\lambda)^{-1}Z'y$ , donde  $I$  es una matriz identidad y  $\lambda$  es un factor de penalización que se agrega a la diagonal de  $Z'Z$ , y permite estimar cualquier número de efectos de marcador. Dicho factor de penalización se estima por máxima verosimilitud restringida (REML) como  $\frac{\sigma_e^2}{\sigma_g^2}$ , donde  $\sigma_g^2$  es la varianza del efecto del marcador y  $\sigma_e^2$  es la varianza del error residual. En el rrBLUP, se asume que todos los marcadores explican cantidades iguales de variación

genética (esto es, varianza común para el efecto del marcador) y supone que sus efectos son normalmente distribuidos (Tong y Nikoloski 2021).

Con respecto a los métodos Bayesianos, estos, a diferencia del método anterior, no asumen una distribución normal de los efectos de los marcadores, sino que en su lugar permiten que una parte de dichos marcadores tengan efectos importantes sobre el carácter, permitiendo así efectos del marcador diferentes (Medina et al. 2021).

Los modelos de regresión Bayesianos (BayesA y BayesB) propuestos por Meuwissen, Hayes, y Goddard (2001), se diferencian en los distintos supuestos sobre los efectos del marcador y sus distribuciones. De acuerdo a Blasco (2021), en BayesA se supone que los efectos de los marcadores se distribuyen de acuerdo a una distribución t de Student en lugar de una normal, permitiendo así que algunos marcadores tengan mayor efecto que otros. En cuanto a BayesB, éste presenta los mismos supuestos de BayesA, sin embargo, a diferencia de este último, en BayesB se permite que una parte de los marcadores no tengan efecto alguno sobre el carácter (Blasco 2021).

Sobre la base de los dos modelos de regresión Bayesianos propuestos por Meuwissen, Hayes, y Goddard (2001), se han desarrollado una gran variedad de modelos Bayesianos para la predicción del GEBV. Por ejemplo, en BayesC se supone que todos los marcadores se distribuyen de forma normal (como la rrBBLUP), sin embargo, al igual que en BayesB, se permite que un porcentaje de los marcadores no tengan efecto sobre el carácter (Blasco 2021). Por otro lado, en el LASSO Bayesiano se supone que el efecto del marcador obedece a una distribución de Laplace, distribución que comparte las mismas características de la t de Student en BayesA al suponer que todos los marcadores tienen un efecto distinto de cero y con diferentes varianzas (Tan et al. 2017).

Un modelo de regresión equivalente al rrGBLUP, denominado como mejor predictor lineal insesgado genómico (GBLUP), fue propuesto por VanRaden (2007). En el GBLUP, se utiliza una matriz de parentesco basada en marcadores de ADN (denominada como matriz G) en lugar de la matriz de parentesco basado en pedigríes (denominada como matriz A) del BLUP descrito por Henderson (1975). Luego la predicción del GEBV se realiza mediante un modelo lineal mixto, cuyas ecuaciones son:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}, \quad (2.4)$$

donde  $X$ ,  $Z$ ,  $b$ ,  $g$  y  $\lambda$  denotan los mismos términos de los modelos (1.2) y (1.3), y  $G^{-1}$  corresponde a la inversa de la matriz G. El GBLUP, al igual que el rrBLUP, supone una varianza común para el efecto del marcador y que el efecto de los marcadores están normalmente distribuidos.

La idea de VanRaden (2007), al sustituir la matriz A por la matriz G, consistió en precisar el parentesco real entre individuos al momento de

estimar los valores de cría. Según Blasco (2021), el parentesco que proviene al usar la matriz A es un parentesco esperado, lo cual puede no reflejar el porcentaje real de genes idénticos entre dos individuos emparentados, mientras que G refleja mejor el parentesco realizado (Figura 2.6).



**Figura 2.6:** Parentesco observado entre individuos emparentados. El coeficiente de parentesco molecular entre los individuos 4 y 5, y 5 y 6, es igual a 1.0 y 0.0, respectivamente, caso contrario al coeficiente de parentesco genealógico, el cual entre todos los individuos, por tratarse de hermanos completos, sería igual a 0.25.

Es más rápida la velocidad del cálculo del GBLUP, comparado con los métodos Bayesianos descritos anteriormente, por lo que el GBLUP es más adecuado para obtener rápidamente el GEBV. Sin embargo, los métodos Bayesianos permiten incorporar al modelo información previa proveniente de múltiples estudios, siendo esto una ventaja de los mismos sobre otros métodos como el GBLUP. En tal sentido, de los Campos et al. (2013) proporcionan algunos ejemplos sobre qué tipo de información se podría incorporar a los efectos previos asignados a los marcadores, por mencionar algunos de ellos, la ubicación del marcador en el genoma, si dicha ubicación corresponde a una región codificante o no, y si el marcador está en una región del genoma que alberga genes que pueden afectar un carácter de interés.

### 2.2.2. De la selección genómica de múltiples pasos a un solo paso

Para implementar los métodos de selección genómica mencionados anteriormente (rrBLUP, BayesA-B-C, LASSO Bayesiano y GBLUP), es necesario disponer de información genotípica y fenotípica de todos los individuos. Esta situación puede ser desventajosa al implementar la selección genómica, ya que por lo general no todos los individuos pueden genotiparse y en ocasiones (principalmente en animales) no se tienen valores fenotípicos para caracteres de interés (por ejemplo, la producción de leche en machos) (Legarra, Aguilar, y Misztal 2009; de los Campos et al. 2013; Jurcic et al. 2021; Blasco 2021). Como solución al problema de la falta de fenotipos, VanRaden (2007), con la implementación del GBLUP, propuso asignarle pseudo-fenotipos o valores de-regresados (estos son, valores fenotípicos estimados a partir de los EBV) a aquellos individuos con fenotipos faltantes, basándose en la información de sus parientes, permitiendo de esta forma implementar la selección genómica combinando los EBV y los genotipos a través de múltiples pasos (Figura 2.7) (Legarra, Aguilar, y Misztal 2009; Misztal, Legarra, y Aguilar 2009; Misztal, Aggrrey, y Muir 2012; Legarra et al. 2014; Misztal, Lourenco, y Legarra 2020).



**Figura 2.7:** Esquema de comparación del BLUP, GBLUP y ssGBLUP. El GBLUP es un proceso de tres pasos en el que los individuos, con base en su información fenotípica y de pedigrí, son evaluados inicialmente mediante el BLUP; luego, a partir de los pseudo-fenotipos resultantes de esta evaluación

inicial, se lleva a cabo un análisis genómico de los individuos genotipados mediante el GBLUP. En el ssGBLUP se simplifica este proceso al incorporar la información genómica (la matriz G) desde el primer paso.

Empero, esta forma de implementar la selección genómica en múltiples pasos no es eficiente (Misztal, Aggrrey, y Muir 2012), además de presentar inconvenientes como son la perdida de información y la dificultad de generalizarse a caracteres múltiples y maternos (Legarra, Aguilar, y Misztal 2009; Legarra et al. 2014). Conscientes de esto, Legarra, Aguilar, y Misztal (2009) simplificaron el proceso de varios pasos al desarrollar un método de selección genómica en el que los fenotipos de los individuos genotipados y no genotipados se analizan conjuntamente para predecir su GEBV (Imai et al. 2019; Jurcic et al. 2021), método el cual se denominó como mejor predictor lineal insesgado genómico de un solo paso (ssGBLUP).

En el ssGBLUP se dispone de una matriz de parentesco genómica de individuos genotipados y no genotipados, denominada como matriz de parentesco combinada o matriz H (Figura 2.7). Esta matriz se obtiene combinando información de parentesco basado en marcadores de ADN (matriz G) entre individuos genotipados, e información de parentesco basado en pedigríes (matriz A) entre individuos genotipados y no genotipados (Imai et al. 2019). Con ello, el proceso anterior de múltiples pasos se simplifica al incorporar la información genómica desde el primer paso (Legarra et al. 2014; Misztal, Legarra, y Aguilar 2009), sin la necesidad del cálculo posterior de pseudo-fenotipos (Misztal, Lourenco, y Legarra 2020).

El proceso de construcción de la matriz H es simple (Recuadro 1.1). De acuerdo a de los Campos et al. (2013), el parentesco genómico de los individuos no genotipados se estima a partir de los que sí lo están, usando un procedimiento de regresión lineal que predice los genotipos no observados como combinaciones lineales de los genotipos observados con coeficientes de regresión derivados de las relaciones basadas en el pedigrí.



#### Recuadro 1.1

Conociendo que la matriz  $H$  equivale a:

$$\begin{bmatrix} var(g_1) & cov(g_1, g_2) \\ cov(g_2, g_1) & var(g_2) \end{bmatrix}$$

El desarrollo de las ecuaciones que conducen a la matriz  $H$  y su posterior uso dentro de las ecuaciones del modelo mixto se explican a continuación.

Partiendo de un modelo GBLUP en el que no se incluyen efectos fijos,  $y = Zg + e$ , un modelo en el cual se incluyen tanto individuos genotipados como no genotipados puede ser de la forma:

$$y = Z \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} + e,$$

donde  $g_1$  corresponde a los individuos no genotipados, y  $g_2$  a los individuos genotipados. Esto es, el vector efectos del marcador en el fenotipo o valores de cría ( $g$ ) es dividido en dos partes, una con los valores de cría de los individuos no genotipados ( $g_1$ ) y otra con los valores de cría de los individuos genotipados ( $g_2$ ).

Para estimar el parentesco genómico de los individuos no genotipados, sus valores de cría ( $g_1$ ) se predicen a partir de los valores de cría de los individuos que si lo están ( $g_2$ ), con base en la expresión:

$$g_1 = cov(g_1, g_2) \times [var(g_2)]^{-1} \times g_2 + e$$

Sabiendo que la matriz de parentesco en base al pedigrí ( $A$ ) puede descomponerse como  $\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ , donde  $A_{11}$  y  $A_{22}$  corresponden, respectivamente, a las matrices de parentesco en base al pedigrí de los individuos no genotipados y genotipados (o bien, la varianza de sus respectivos valores de cría), y  $A_{12} = A_{21}$  corresponden a la matriz de parentesco en base al pedigrí entre individuos genotipados y no genotipados (o bien, la covarianza de sus valores de cría), la expresión anterior puede reescribirse de la siguiente forma:

$$g_1 = A_{12}A_{22}^{-1}g_2 + e$$

Luego la expresión anterior en términos de varianza:

$$var(g_1) = var(A_{12}A_{22}^{-1}g_2) + var(e)$$

Con un poco de álgebra, finalmente la expresión correspondiente a  $var(g_1)$  es:

$$\text{var}(g_1) = A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21}$$

En relación a la expresión  $\text{var}(g_2)$ , ésta sería igual a:

$$\text{var}(g_2) = G$$

Por último, la expresión  $\text{cov}(g_1, g_2)$ , equivaldría a:

$$\text{cov}(g_1, g_2) = \text{cov}(A_{12}A_{22}^{-1}g_2 + e, g_2)$$

Qué con un poco de álgebra, equivaldría a:

$$\text{cov}(g_1, g_2) = A_{12}A_{22}^{-1}G$$

Finalmente, la matriz que contiene las relaciones conjuntas de individuos genotipados y no genotipados sería:

$$H = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A'_{12} & G \end{bmatrix},$$

Una vez obtenida la matriz H, se obtiene su inversa ( $H^{-1}$ ), y el modelo de selección genómica se puede resolver mediante un modelo lineal mixto:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + H^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix},$$

Al ser una forma de BLUP o GBLUP en el que la matriz A y G, respectivamente, es sustituida por la matriz H (Legarra, Aguilar, y Misztal 2009; Legarra et al. 2014; Blasco 2021), el ssGBLUP se puede adecuar con facilidad a caracteres múltiples y maternos (Blasco 2021), además se adapta también a las herramientas informáticas ya desarrolladas en base al BLUP y GBLUP (Lourenco et al. 2020). Este hecho hace del ssGBLUP un método de uso rutinario para la selección genómica, donde ha demostrado que

produce una predicción más precisa en comparación a los métodos BLUP y GBLUP ya mencionados (Misztal, Aggrrey, y Muir 2012; Pérez-Rodríguez et al. 2017; Misztal, Lourenco, y Legarra 2020).

### 2.2.3. Factores que afectan la habilidad predictiva de la selección genómica

La precisión con la que se predice el GEBV es el factor más importante que determina el éxito de la selección genómica (Nakaya y Isobe 2012). Dicha precisión se suele evaluar como la correlación entre el GEBV predicho y el valor fenotípico real u observado, lo que se denomina como habilidad predictiva (Ahmadi, Bartholomé, Cao, et al. 2020b). Asimismo, la habilidad predictiva de la selección genómica depende principalmente del número de individuos genotipados y la densidad de marcadores (Ahmadi, Bartholomé, Cao, et al. 2020b; Tong y Nikoloski 2021).

Con relación a la densidad de marcadores, hoy en día se conoce que la habilidad predictiva de la selección genómica disminuye a menor número de marcadores de ADN en desequilibrio de ligamiento con el QTL (Desta y Ortiz 2014; Ahmadi, Bartholomé, Cao, et al. 2020b). Asimismo, los estudios de simulación indican que a mayor número de marcadores se mejora la habilidad predictiva de la selección genómica (Tong y Nikoloski 2021). Sin embargo, el aumento de la precisión puede alcanzar su límite a medida que se aumenta la densidad del marcador (Blasco y Toro 2014; Crossa et al. 2017; Ahmadi, Bartholomé, Cao, et al. 2020b), por lo cual se recomienda mediante simulación evaluar el efecto del número de marcadores de ADN en la precisión de la predicción, antes de implementar la selección genómica en escenarios prácticos (Pérez-Enciso, Ramírez-Ayala, y Zingaretti 2020).

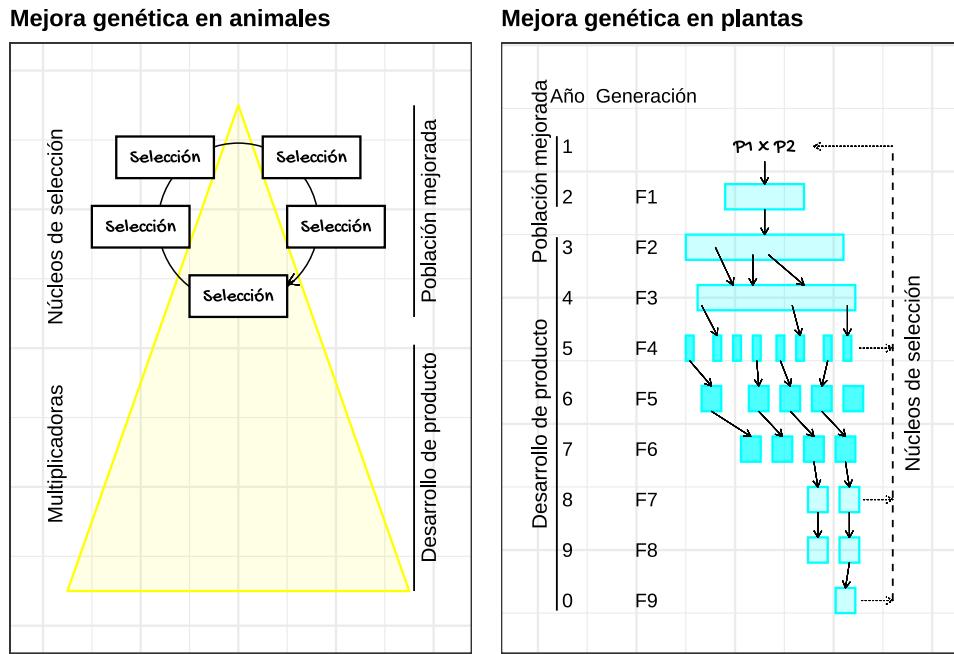
En cuanto al número de individuos genotipados, la habilidad predictiva de la selección genómica es mayor a medida que aumenta su proporción. Esto lo evidencian Nakaya y Isobe (2012), Blasco y Toro (2014), y Desta y Ortiz (2014), quienes coinciden en afirmar que a mayor tamaño de la población de entrenamiento, mayor será la precisión de la predicción del GEBV.

## 2.3. Diferencias de la mejora genética en animales y en plantas

Tras la domesticación de animales y el cultivo de plantas, la especie humana consiguió producir animales y plantas mejoradas gracias a la selección artificial. La clave del éxito de los antiguos domesticadores y cultivadores consistió en seleccionar y cruzar individuos portadores de caracteres deseables, al comprender que los descendientes podrían heredar estas características, aunque se desconocían los mecanismos biológicos de la herencia (Holland 2014). Hoy en día, se conoce como mejora genética al proceso sistemático de

mejorar los caracteres fenotípicos deseables en animales y plantas mediante selección artificial (Tong y Nikoloski 2021).

En términos generales, la mejora genética se puede organizar en tres procesos, siendo éstos la producción de variación genética, la selección entre la variación y la multiplicación para uso comercial (Figura 2.8).



**Figura 2.8:** Esquema de la mejora genética en animales y en plantas. Figura adaptada de Hickey et al. (2017).

El primer requisito para la mejora genética es disponer de variación genética de los caracteres que se desean a mejorar. Tanto en animales como en plantas, la variación genética se da a través del apareamiento de individuos cuya expresión de caracteres es deseable (Caligari y Brown 2017). De esta forma, mediante el proceso natural de reproducción sexual, se puede obtener una descendencia que contiene genes de interés heredados de los dos progenitores.

A partir de una población variable, es necesario seleccionar individuos con mejor expresión de caracteres. Tanto en animales como en plantas, la selección se lleva a cabo de forma recurrente (en los denominados núcleos de selección) con la finalidad de aumentar la frecuencia de alelos favorables. Sin embargo, los métodos de selección usados han sido distintos. Por un lado, en plantas se ha utilizado principalmente la MAS con el fin de identificar e incorporar genes beneficiosos, favoreciendo así que genes con efectos moderados a grandes hayan sido explotados más ampliamente en plantas que en

animales. Por otro lado, en animales la mayoría de caracteres económicamente importantes han sido cuantitativos y complejos, lo cual obligó a los mejoradores genéticos a usar enfoques biométricos para predecir el EBV mediante la combinación de información fenotípica y de pedigrí (propriamente el BLUP), y con ello tomar las decisiones de selección (Hickey et al. 2017).

Finalmente, es necesario multiplicar o difundir el mérito genético medio obtenido de la población mejorada, y así facilitar que en las granjas comerciales los agricultores produzcan los caracteres mejorados (Blasco 2021). Según Hickey et al. (2017), una diferencia importante entre la mejora genética de animales y plantas, es que en animales la mejora difundida a las granjas comerciales no se utiliza en los núcleos de selección, mientras que en plantas, los productos mejorados (en forma de variedades mejoradas) obtenidos en cualquier parte de un ciclo de cultivo, si pueden ser usados como progenitores en un nuevo ciclo.

Por otro lado, desde el momento en que la selección genómica fue propuesta por Meuwissen, Hayes, y Goddard (2001), se adaptó rápidamente a la mejora genética en animales, principalmente al sector ganadero. Sin embargo, el uso de la selección genómica en plantas no se ha extendido tanto (Wang, Crossa, y Gai 2020) y son varias las razones de ello:

1- Los métodos de mejora genética en animales y en plantas han divergido a lo largo de los años, lo cual implica que se requiera de tiempo para que los avances y las contribuciones realizadas en un campo se trasladen al otro (Hickey et al. 2017).

2- El genoma de muchas especies de plantas es más compleja al genoma de los animales. Los animales al ser individuos diploides, aportan a la descendencia solo uno de sus dos alelos, por lo cual es más fácil predecir en ellos cuán efectiva será la selección, al suponer que de los distintos componentes de la varianza genética solo se heredara la varianza debido a efectos aditivos (esto es, la heredabilidad en el sentido estricto). Caso contrario sucede en las plantas cuya respuesta a la selección puede implicar, en caso de que la especie sea poliploide o se haya propagado de forma vegetativa, otros tipos de interacción como la dominancia entre dos alelos (Holland 2014).

3- Es costoso invertir en infraestructura computacional y de registro tanto de datos genotípicos como fenotípicos requeridos para implementar la selección genómica. El tamaño de las poblaciones en la cría de animales son mucho más pequeños que la mayoría de poblaciones en la cría de plantas. Si bien los costos de genotipado por individuo son cada vez más bajos, el costo total de genotipado al considerar todas las plantas es aún hoy en día demasiado alto para la mayoría de programas de mejora genética en plantas (Wang, Crossa, y Gai 2020).

Pese a lo anterior, en la literatura científica (Tabla 2.1) se evidencia el uso potencial de la selección genómica para mejorar el mérito genético medio por selección tanto en animales como en plantas. A pesar de sus diferencias, ambas disciplinas requieren de conceptos y herramientas similares

de selección genómica.

**Tabla 2.1:** .

Especie	Carácter	Habilidad predictiva	Modelo	Referencia

## 2.4. La mejora genética en el arroz

### 2.4.1. Importancia de la mejora genética en el arroz

El arroz es uno de los principales cultivos básicos del mundo, ya que proporciona aproximadamente el 20 % de la energía alimentaria mundial (Wing, Purugganan, y Zhang 2018). En la figura a continuación (Figura 2.9) se muestra el aumento en la producción de arroz a lo largo de los años en respuesta a una población mundial en crecimiento. De igual modo, en los próximos años se prevé que dicho crecimiento en la población mundial continúe, lo cual requerirá del aumento en la producción de arroz en relación con los niveles actuales (Hickey et al. 2017; Kyselova, Tichý, y Jochová 2021).

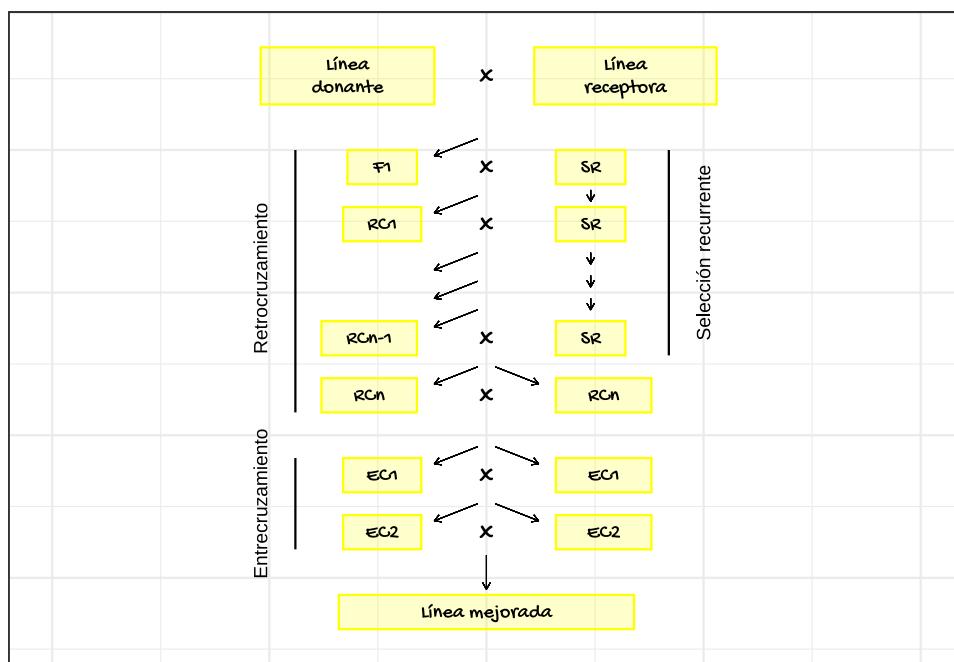


**Figura 2.9:** Evolución de la producción y uso del arroz a nivel mundial. Se utilizaron las bases de datos estadísticos de la Organización para la Alimentación y la Agricultura (FAOSTAT) para desarrollar estas figuras.

#### 2.4.2. Breve descripción de la mejora genética en plantas con ejemplos en el arroz

El método tradicional y más sencillo en la mejora genética en plantas es la selección de individuos de una sola línea de acuerdo a su fenotipo. Este método, conocido como selección masal, se implementó con un conocimiento elemental de la herencia, basado en el supuesto de que los descendientes de los mejores individuos heredarían sus características (Breseghello 2013). En el arroz, Bartholomé, Thathapalli-Prakash, y Cobb (2021) citan distintos estudios donde se muestra la mejora genética lograda en este cultivo a través del método de selección masal.

Posterior al método de selección masal, se propuso un método en plantas en el cual la mejora genética era posible a través del apareamiento controlado de individuos. Se distinguen dos estrategias: la selección fenotípica recurrente y la introgresión (Figura 2.10).



**Figura 2.10:** Esquema de mejora genética de especies agrícolas. Figura adaptada de Dekkers y Hospital (2002).

En la selección fenotípica recurrente, la mejora genética se realiza a través de un esquema de rondas consecutivas de selección. En esta estrategia de apareamiento, se cruzan individuos progenitores (líneas donante y receptora) y se generan poblaciones segregantes a través de generaciones de autopolinización y selección, hasta obtener variedades mejoradas que combinan las

buenas características y los avances logrados en generaciones anteriores. La característica principal de la selección fenotípica recurrente es aumentar las frecuencias de los alelos favorables (Guimarães 2009).

En cuanto a la introgresión, ésta al igual que la anterior, se inicia con el cruce de individuos progenitores, seguido de repetidos ciclos de retrocruzamiento y entrecruzamiento. El propósito de esta estrategia de apareamiento consiste en generar variedades mejoradas (híbridas) reuniendo en una misma planta las mejores características de los distintos cruces (Breseghezzo 2013), aumentando también el rendimiento más allá de las variedades iniciales mediante la explotación de la heterosis (Guimarães 2009).

En relación con las estrategias descritas en el párrafo anterior, Grenier et al. (2015) hacen mención del programa de mejora genética en el arroz para América Latina y el Caribe, desarrollados por el centro internacional de agricultura tropical, y por la organización francesa de investigación agrícola y de cooperación internacional para el desarrollo sostenible. Los autores describen a este programa como un proceso de tres etapas realizadas de manera recurrente: evaluación de individuos derivados de una población inicial, selección con presión suave de los mejores individuos para aumentar gradualmente la frecuencia de alelos favorables, y entrecruzamiento de los individuos seleccionados para formar una nueva población con una media mejorada.

La MAS también fue importante en la mejora genética de plantas (Xu et al. 2021), ya que permitió identificar e incorporar QTL beneficiosos en tiempos relativamente cortos (Fukuoka et al. 2010). Sin embargo, el impacto general al implementar este método para aumentar la eficiencia de la mejora genética en el arroz ha sido limitada (Robertsen, Hjortshøj, y Janss 2019), puesto que la mayoría de caracteres de interés en este cultivo no están controlados por pocos genes con grandes efectos (siendo esta la principal característica para que la MAS sea efectiva), sino por muchos genes de pequeño efecto y/o por una combinación de genes mayores y menores (Kyselova, Tichý, y Jochová 2021; Xu et al. 2021).

La selección genómica brindó nuevas oportunidades para la mejora genética de caracteres con herencia poligénica (Cappetta et al. 2020). Este método se implementó para la mejora de la ganancia genética en plantas, debido al éxito alcanzado en animales (Xu et al. 2021), y hoy en día se ha establecido como un método prometedor (Robertsen, Hjortshøj, y Janss 2019). En el arroz, los primeros estudios que informaron del uso de la selección genómica fueron publicados en el año 2014, siendo esto considerado por Bartholomé, Thathapalli-Prakash, y Cobb (2021) como un hecho donde se evidencia un retraso en la transición del mapeo de QTL a la selección genómica, debido a que años atrás ya se contaba con los recursos genómicos necesarios para su implementación.

En plantas, una de las ventajas con la implementación de la selección genómica es que permite reducir el tiempo requerido para el desarrollo de

variedades mejoradas. Cappetta et al. (2020) aluden esta ventaja en la mejora genética del tomate, donde a partir del uso de la selección genómica es posible reducir el número de generaciones de autopolinización necesarias para generar variedades de tomate de buen rendimiento. De igual modo, Crossa et al. (2017) mencionan que el uso de la selección genómica en la mejora genética del maíz conlleva una reducción de los costos de mejoramiento comparado con el método tradicional de selección fenotípica, al reducir el tiempo requerido para el desarrollo de variedades. En cuanto a la mejora genética del arroz, Bartholomé, Thathapalli-Prakash, y Cobb (2021) hacen mención que el uso de la selección genómica permite aumentar la tasa de ganancia genética en este cultivo al reducir el tiempo necesario en la generación de variedades mejoradas.

La mayor parte de los estudios sobre selección genómica en plantas han ignorado la información de pedigrí, a diferencia de los animales donde el uso del pedigrí en modelos de selección ha sido considerado un factor importante en los programas de mejora genética (Robertsen, Hjortshøj, y Janss 2019). Con el desarrollo del ssGBLUP existe la posibilidad de usar esta información, y hoy en día se pueden encontrar ejemplos del uso de esta metodología en plantas (Pérez-Rodríguez et al. 2017; Jurcic et al. 2021). Sin embargo, todavía se necesitan esquemas de selección genómica optimizados y validados en el arroz. En estos momentos, no se han evaluado los distintos factores que pueden afectar la implementación y precisión del método del ssGBLUP en cultivos de arroz. Contar con estos resultados, ya sea a través de enfoques de simulación o usando datos empíricos, será clave para evaluar el impacto de dichos factores al implementar el ssGBLUP en el arroz.

## **Capítulo 3**

### **Objetivo**

El objetivo de este estudio fue evaluar el interés de la selección genómica en un solo paso en plantas. Para ello, se evaluó el efecto de la cantidad de individuos genotipados y del número de marcadores en un diseño que combinó datos reales de arroz y simulaciones de dos tipos, hacia atrás (ancestral) y hacia delante (descendientes).

## Capítulo 4

# Materiales y métodos

### 4.1. Recurso vegetal y datos fenotípicos

Los conjuntos de datos se obtuvieron del Rice SNP-Seek Database (disponible en la dirección web o URL <http://iric.irri.org/>), que contiene información sobre genotipados de SNP y fenotipos de distintas variedades de arroz (*Oryza sativa L.*). En un estudio previo a éste (Vourlaki et al., [s. f.](#)), los genotipados de SNP fueron sometidos a procedimientos de control de calidad, eliminando loci con una frecuencia del alelo menor inferior a 0.01 y con una tasa de ausencia mayor a 0.01. Para este estudio se eliminaron loci con una frecuencia del alelo menor inferior a 0.05 usando Plink (Purcell et al. [2007](#)).

Luego, el conjunto de datos de genotipado final se transformó de la codificación de genotipo de nucleótidos (esto es, A, C, T y G) a la codificación numérica (0, 1 y 2 para los homocigotos de clase I, heterocigotos y homocigotos de clase II, respectivamente) para facilitar el análisis estadístico posterior.

En la figura Figura 4.1 se presenta un análisis de componentes principales realizado sobre los datos de genotipado de SNP, donde se observan diferentes grupos varietales de arroz. La variedad indica fue seleccionada para llevar a cabo este estudio ya que es el grupo varietal con mayor número de individuos genotipados (451 individuos de un total de 738).



**Figura 4.1:** Análisis de componentes principales en datos de arroz. Los puntos y las circuferencias de color representan los distintos grupos varietales disponibles: tipo intermedio o mezclado (ADM), aromático (ARO), aus (AUS), indica (IND) y japónica (JAP).

De entre los datos de fenotipos disponibles, se eligió el carácter tiempo de floración, ya que en un estudio previo (Vourlaki et al., s. f.) se observó que en este carácter la predicción genómica podría funcionar mejor. Con la figura a continuación (Figura 4.2) se observa que la distribución empírica general de este carácter fue aproximadamente normal, lo que permitió el uso de modelos de distribución normal. Los datos fenotípicos se centraron (restando la media general) y estandarizaron (dividiendo por la desviación estándar) antes de ajustar los modelos que se describen a continuación.



**Figura 4.2:** Distribución y media  $\pm$  desviación estándar del carácter tiempo de floración del conjunto de datos fenotípicos de arroz.

## 4.2. Predicción basada en información del pedigrí e información genómica

Para llevar a cabo la predicción usando el BLUP en individuos no genotipados y el ssGBLUP tanto en individuos genotipados como no genotipados, se aplicó el siguiente modelo:

$$\mathbf{y} = \mu + \mathbf{Zg} + \mathbf{e}, \quad (4.1)$$

donde  $y$  representa el valor del fenotipo (es decir, el tiempo de floración) y  $Z$  es la matriz de incidencia que relaciona  $g$  con  $y$ .  $1_n$  es un vector de unos,  $\mu$  es la media de la población, el vector  $g$  representa los efectos aleatorios genéticos aditivos, y  $e$  es el vector de residuos con una distribución que se asume normal con media igual a 0 y matriz de covarianza  $I\sigma_e^2$ , siendo  $I$  representá la matriz identidad, y  $\sigma_e^2$  la varianza residual.

En la ecuación anterior (4.1), se asume que  $g$  sigue una distribución normal con media igual a 0 y matriz de covarianza  $A\sigma_g^2$  en el modelo BLUP, donde  $A$  representa la matriz de parentesco basada en información del pedigrí, y  $\sigma_g^2$  es la varianza genética aditiva.

En el modelo ssGBLUP, la matriz  $A$  del modelo de la ecuación (4.1) es reemplazada por la matriz  $H$ , de la misma dimensión que la matriz  $A$ . Dicha

matriz H es una función de la matriz A y de la matriz G (Misztal, Legarra, y Aguilar 2009), y se define de la siguiente manera:

$$H = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ GA_{22}^{-1}A'_{12} & G \end{bmatrix}, \quad (4.2)$$

donde  $A_{11}$ ,  $A_{12}$ ,  $A_{21}$  y  $A_{22}$  son submatrices de A, siendo  $A_{11}$  la submatriz de los individuos sin genotipo,  $A_{22}$  la submatriz de los individuos con genotipo, y  $A_{12}$  y  $A_{21}$  las submatrices que contienen las relaciones genéticas aditivas esperadas entre individuos con genotipo y sin genotipo. Para el cálculo de la matriz H, los datos de genotipado de SNP fueron escalados. Luego, a partir de los datos escalados, se obtuvo la matriz G utilizando el método de VanRaden (2007),  $\frac{XX'}{2\sum_{j=1}^{nSNP} p_j(1-p_j)}$ , donde X es una matriz de dimensión  $n \times nSNP$  que contiene los genotipos con la codificación numérica descrita anteriormente (0, 1 y 2),  $p_j$  es la frecuencia del  $j$ -ésimo SNP,  $n$  corresponde al número de individuos y  $nSNP$  representa al número de SNP. Para evitar posibles problemas de singularidad, a los elementos de la diagonal de la matriz G se les sumó un valor de 0.05.

Se usaron diferentes matrices H basados en diferentes matrices G, situación que se describe en las dos secciones metodológicas que se describen a continuación. En el Anexo A.1 se presenta la función R (R Core Team 2020) usada para construir la matriz H.

La heredabilidad en sentido estricto se estimó mediante el método BLUP como:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}, \quad (4.3)$$

donde  $\sigma_g^2$  es la varianza genética aditiva y  $\sigma_e^2$  es la varianza residual.

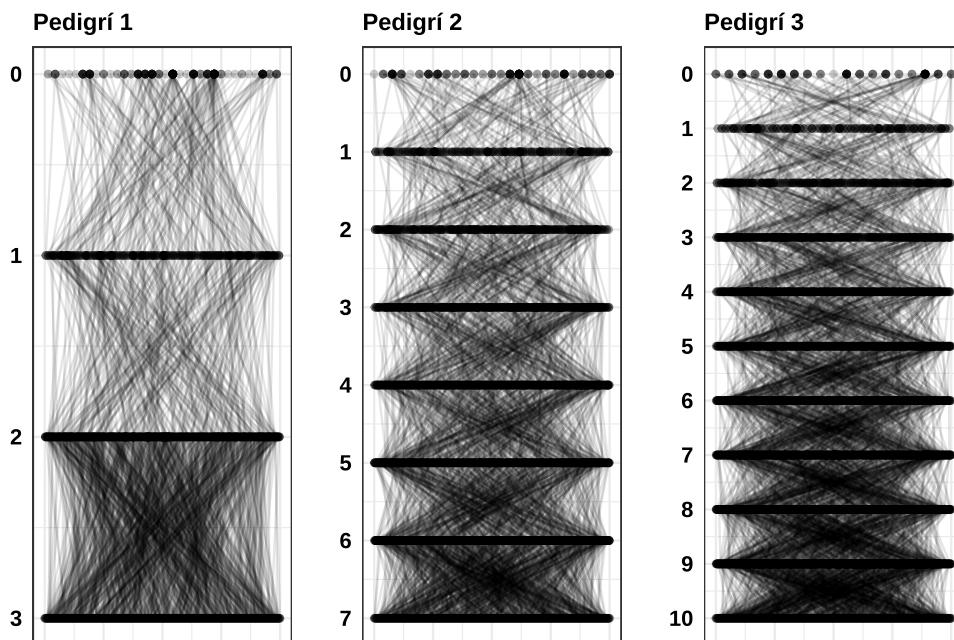
### 4.3. Precisión de la predicción mediante simulación del pedigrí ancestral

En especies de plantas es común que los pedigríes de las poblaciones reproductoras sean completamente, o al menos parcialmente, desconocidos (las razones del porque de esto se describen en Cros et al. (2014)). Este es el caso de los conjuntos de datos del Rice SNP-Seek Database usados en este estudio. Por esta razón, se utilizó la metodología implementada en el software Molcoanc (Fernández y Toro 2006) con el fin de simular esta información.

Con el software Molcoanc se simulan pedigríes mediante la creación de antepasados virtuales para los individuos genotipados (esto es, la población fundadora), de tal manera que la correlación entre el parentesco genealógico calculado a partir del pedigrí generado tenga la correlación más alta posible

con la matriz de parentesco molecular calculada a partir de los marcadores de ADN proporcionados (Fernández y Toro 2006).

El software Molcoanc se utilizó para construir tres pedigríes, que se diferenciaron en el número de generaciones ancestro de la población fundadora (Figura 4.3). Este proceso de generación de pedigríes se replicó diez veces para cada pedigrí, con el fin de posteriormente usar cada replica para medir la variabilidad de la predicción mediante el BLUP y el ssGBLUP.



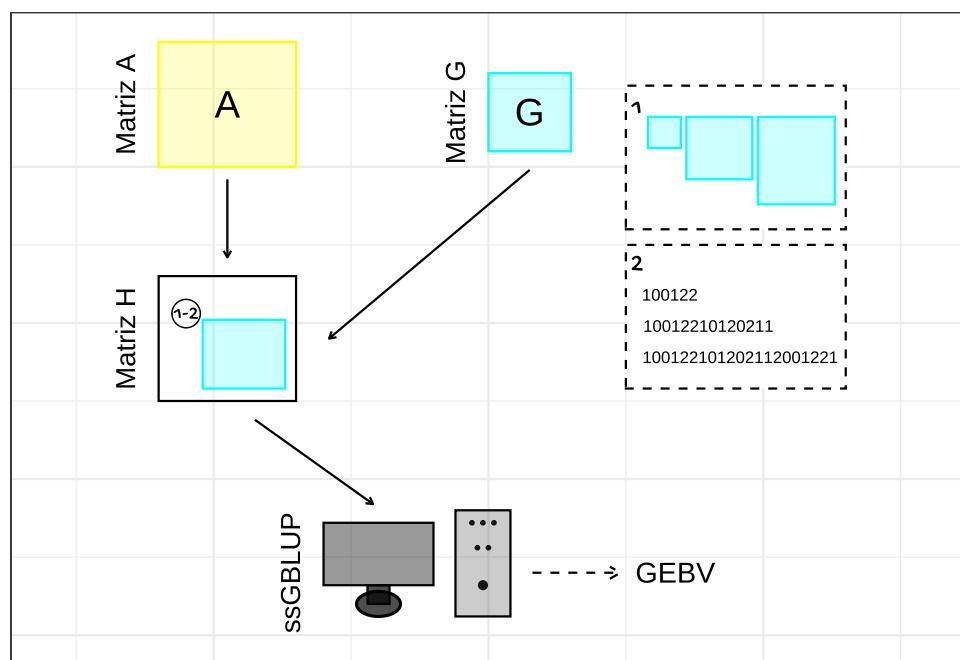
**Figura 4.3:** Ejemplo de pedigríes generados en la primera replica. En cada pedigrí se generaron distintas generaciones por encima de los individuos de la población fundadora. El número de individuos totales en cada pedigrí fueron: 751 individuos en el pedigrí 1, 1661 en el pedigrí 2 y 2451 en el pedigrí 3.

Para estudiar el efecto sobre la predictibilidad del número de datos de genotipado de SNP y del número de individuos genotipados, y así determinar el número de marcadores y de individuos en la población de entrenamiento necesarios para implementar el ssGBLUP en datos de arroz, se usaron diferentes subconjuntos de datos con las siguientes características:

1. Distinta cantidad de individuos genotipados: del conjunto total de individuos genotipados (451), se seleccionaron de forma aleatoria 148 y 298 individuos, generando con ello tres subconjuntos de datos (o diferentes poblaciones de entrenamiento) que incluían 148, 298 y 451 individuos.

2. Diferentes densidades de SNP: del conjunto total con 100231 SNP después del control de calidad, se seleccionaron de forma aleatoria mediante el uso de Plink tres subconjuntos de datos de SNP, de manera que el número de marcadores aproximado fuera igual a 1.000, 10.000 y 100.000 SNP.

A continuación, usando los diferentes subconjuntos de datos descritos anteriormente, se obtuvo un total de nueve matrices H con distintas combinaciones de densidad de marcadores e individuos genotipados (Figura 4.4), usando la función R descrita en el anexo A.1.



**Figura 4.4:** Esquema de construcción de la matriz H a partir de la matriz A y la matriz G, con base en diferentes subconjuntos de datos (o matrices G). El recuadro 1 representa tres matrices G con distinta dimensión dado el número de individuos genotipados y el recuadro 2 representa tres diferentes densidades de SNP, para cada uno de las tres matrices del recuadro 1.

Además de los casos descritos con anterioridad, también se consideró el caso en el que ningún individuo haya sido genotipado y no se usara información de marcadores de ADN.

Los componentes de varianza y los valores fenotípicos predichos se obtuvieron ajustando los modelos BLUP y ssGBLUP descritos anteriormente. Para dicho ajuste se utilizaron los paquetes BGLR (Pérez-Rodríguez y de los

Campos 2014) y **lme4GS** (Caamal-Pat et al. 2021) del lenguaje de programación R, permitiendo así la predicción mediante procedimientos Bayesianos y de máxima verosimilitud restringida (REML), respectivamente. Para la estimación de los componentes de varianza y la predicción de los valores fenotípicos usando el paquete **BGLR**, la Cadena de Markov Monte Carlo (MCMC) se generó con 50000 iteraciones, de las cuales fueron descartadas las primeras 10000 como muestras burn-in. En cuanto al paquete **lme4GS**, no se cambio el número de iteraciones necesarios para alcanzar la convergencia, ya que los autores del paquete (Caamal-Pat et al. 2021) indicaron que el número de iteraciones que se tiene por defecto era suficiente (comunicación personal).

Se uso el coeficiente de correlación entre los valores fenotípicos observados y predichos como medida de predictibilidad. De acuerdo a Xua, Zhub, y Zhang (2014), la predictibilidad debe obtenerse usando una muestra de validación independiente donde los individuos predichos no deben contribuir a la estimación de parámetros. En este estudio, los valores fenotípicos observados de 48 del total de 451 individuos de arroz de la variedad indica (que corresponde a los individuos clasificados como variedades mejoradas en los conjuntos de datos) se consideraron como fenotipos faltantes .

#### 4.4. Precisión de la predicción mediante simulación de descendientes

Se evaluó el efecto sobre la predictibilidad del número de datos de genotipado de SNP y del número de individuos genotipados realizando simulación de descendientes, usando el paquete **SeqBreed** (Pérez-Enciso, Ramírez-Ayala, y Zingaretti 2020) del lenguaje de programación **Python** (Van-Rossum y Drake 1995). En esta sección se especuló sobre la base genética del carácter tiempo de floración, mientras que en la sección metodológica anterior se usaron los fenotipos y genotipos observados. Los pasos llevados a cabo en esta simulación se describen a continuación (Figura 4.5):

1. Uso de genotipos de la población fundadora: para simular los datos de fenotipo y de genotipo se usó el conjunto de datos con los 100.231 SNP resultantes del control de calidad. Dicho conjunto de datos se pasó previamente a formato variant call usando Plink, ya que así lo requería el paquete **SeqBreed**. Luego usando las funciones **gg.GFounder()** y **gg.Genome()**, se obtuvo un archivo que indicaba el número de individuos genotipados o individuos de la población fundadora (451), la ploidía (2) y el número de SNP (100.231).
2. Especificación de la arquitectura genética (SNP causales (QTN) y sus efectos) para el carácter tiempo de floración y heredabilidad deseada: se uso el software GCTA (Yang et al. 2011) para hacer un estudio de

asociación a nivel genómico (GWAS) para identificar las regiones genómicas asociadas al carácter (Anexo A.2). Luego, en base al GWAS, se generaron los datos que indicaban el efecto de los QTN y su localización, seleccionando solo 50 de ellos con un efecto proporcional a la varianza aditiva explicada. Por último, se uso la función `gg.QTNs()` sobre estos datos generados, indicando también la heredabilidad del carácter (0.7) de acuerdo a resultados previamente reportados (Vourlaki et al., s. f.).

3. Generación de pedigríes: se generaron cuatro pedigríes (con diez replicas para medir la variabilidad de la predicción mediante el BLUP y el ssGBLUP), cada uno de ellos con esquemas de cruzamiento diferentes partiendo de la población fundadora con 451 individuos (Tabla 4.1). Usando la función `gg.Population()`, se generó mediante simulación los fenotipos y genotipos de cada uno de los individuos en cada uno de los cuatro pedigríes.

**Tabla 4.1:** Representación de los cuatro pedigríes generados. Cada uno de estos pedigríes tienen esquemas de cruzamientos diferentes, dando lugar a distintos número de individuos en la generación  $F_1$ , pero con mismo número de individuos en las generaciones  $F_2$  y  $F_3$ .

	Pedigrí 1	Pedigrí 2	Pedigrí 3	Pedigrí 4
F0	451	451	451	451
F1	$10^1$	20	40	80
F2	$800 (10 \times 80)^2$	$800 (20 \times 40)$	$800 (40 \times 20)$	$800 (80 \times 10)$
F3	$800 (800 \times 1)^3$	$800 (800 \times 1)$	$800 (800 \times 1)$	$800 (800 \times 1)$
Total	2061	2071	2091	2131

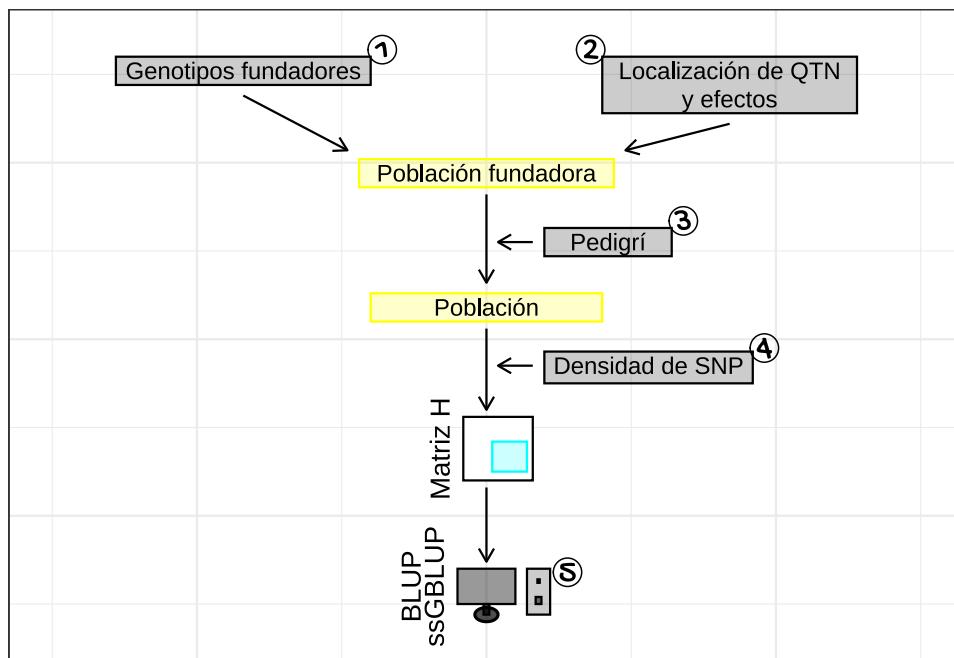
<sup>1</sup>10 indica el número de descendientes que tendrían los 451 individuos de la generación F0 mediante cruzamiento.

<sup>2</sup> $10 \times 80$  indica el número de descendientes (80) que tendrían cada uno de los 10 individuos de la generación F1 por autofecundación, dando un total de 800 individuos en la generación F2.

<sup>3</sup> $800 \times 1$  indica el número de descendientes (1) que tendrían cada uno de los 800 individuos de la generación F2 por autofecundación, dando un total de 800 individuos en la generación F3.

4. Uso de subconjuntos de datos con diferentes densidades de SNP: se especificaron distintas densidades de SNP usando la función `gg.Chip()`, con el fin de determinar el número de marcadores necesarios para implementar el ssGBLUP. Para ello, se usaron los tres subconjuntos de datos de SNP creados en la sección metodológica anterior, con número de marcadores aproximado de 1.000, 10.000 y 100.000 SNP.

5. Implementación de la selección: se usó la función `sel.doEbv()` para llevar a cabo la predicción, usando los modelos BLUP y ssGBLUP descritos anteriormente. En el caso del ssGBLUP, el cual requiere de datos de marcadores de ADN, se usó previamente la función `gg.do_X()` para generar la matriz G. Cabe aclarar que `SeqBreed` generó dicha matriz G de acuerdo a las condiciones descritas anteriormente, como son calcularla utilizando el método de VanRaden (2007), y sumarle a sus elementos de la diagonal un valor de 0.05 para evitar posibles problemas de singularidad.



**Figura 4.5:** Esquema de predicción usando simulación con el paquete `SeqBreed`. Figura adaptada de Pérez-Enciso, Ramírez-Ayala, y Zingaretti (2020).

Para identificar el efecto del número de individuos genotipados sobre la predictibilidad, fueron considerados cuatro casos:

- Ningún individuo genotipado,
- Solo los individuos de la generación  $F_2$  genotipados,
- Los individuos de las generaciones  $F_1$  y  $F_2$  genotipados y,
- Todos los individuos de las distintas generaciones genotipadas.

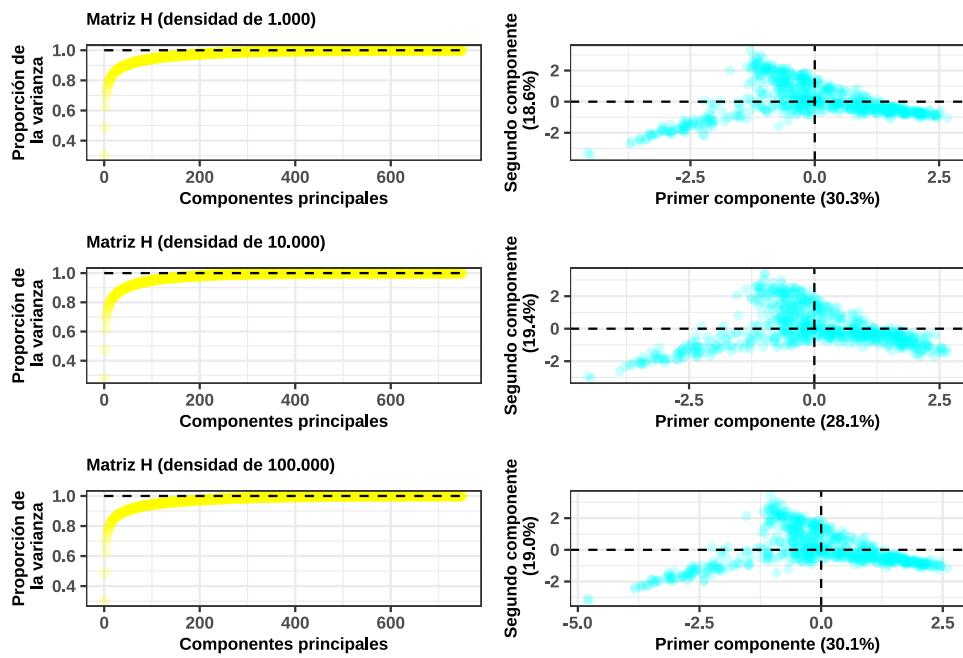
Por último y al igual que en la sección metodológica anteriormente descrita, se uso el coeficiente de correlación entre los valores fenotípicos observados y predichos como medida de predictibilidad. Para esto, los valores fenotípicos observados de los individuos de las generaciones  $F_2$  y  $F_3$  se consideraron como fenotipos faltantes (esto es, como conjuntos de validación).

# Capítulo 5

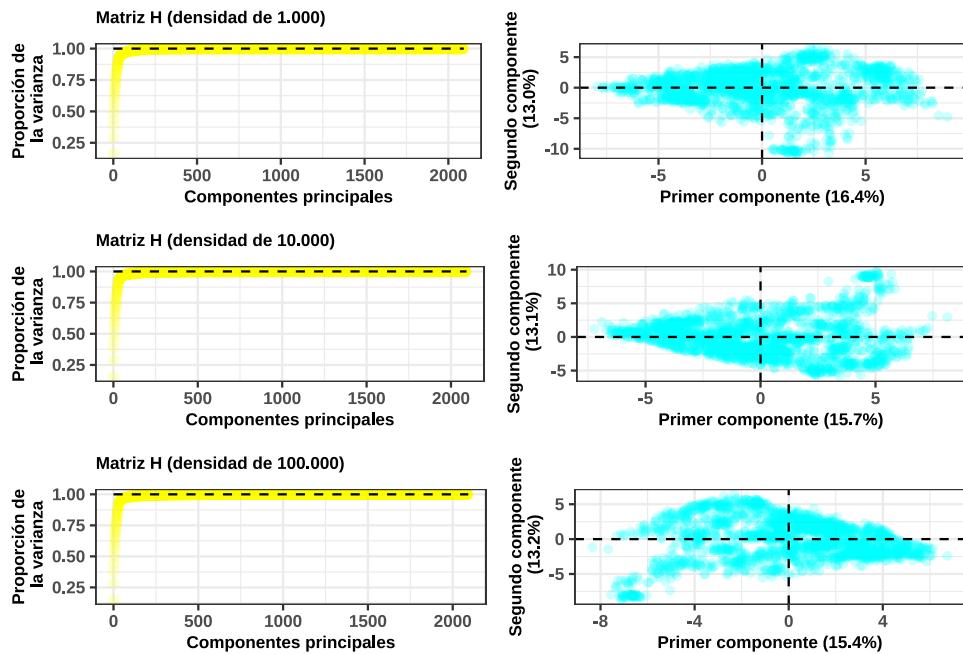
## Resultados

### 5.1. Estructura poblacional y heredabilidad en la simulación ancestral

Se examinó si había alguna evidencia de estructura poblacional a partir de la matriz H resultante tanto de la simulación ancestral como de descendientes, de la cual no se observó una estructura poblacional clara según los primeros dos componentes principales (Figuras 5.1 y 5.2). De acuerdo a Morais-Júnior, Duarte, et al. (2018), el hecho de que la mayor parte de la variación total sea explicada por unos pocos valores propios (en este estudio, los primeros diez a veinte valores propios explicaron el 80.0 % de la variación total), sugiere que la población no es especialmente variable, lo cual era de esperar dado el método implementado de generación de individuos a través de la simulación del pedigree ancestral y de descendientes. Con esta estrategia se trató de replicar los métodos que se emplean en la mejora de los cultivos de arroz, esto es, generación inicial de híbridos a partir del cruzamiento entre distintos individuos, y posterior recombinación en distintas generaciones y por autopolinización. La ausencia de estructura poblacional, según Guo et al. (2014), es un escenario favorable para obtener estimaciones confiables de predicción genómica.



**Figura 5.1:** Análisis de componentes principales a partir de la matriz H resultante de la simulación ancestral. Aquí, la matriz H se estimó a partir de 451 individuos genotipados con densidad de marcador aproximada de 1.000, 10.000 y 100.000, y en el pedigrí 1 (primera réplica) mediante simulación ancestral.



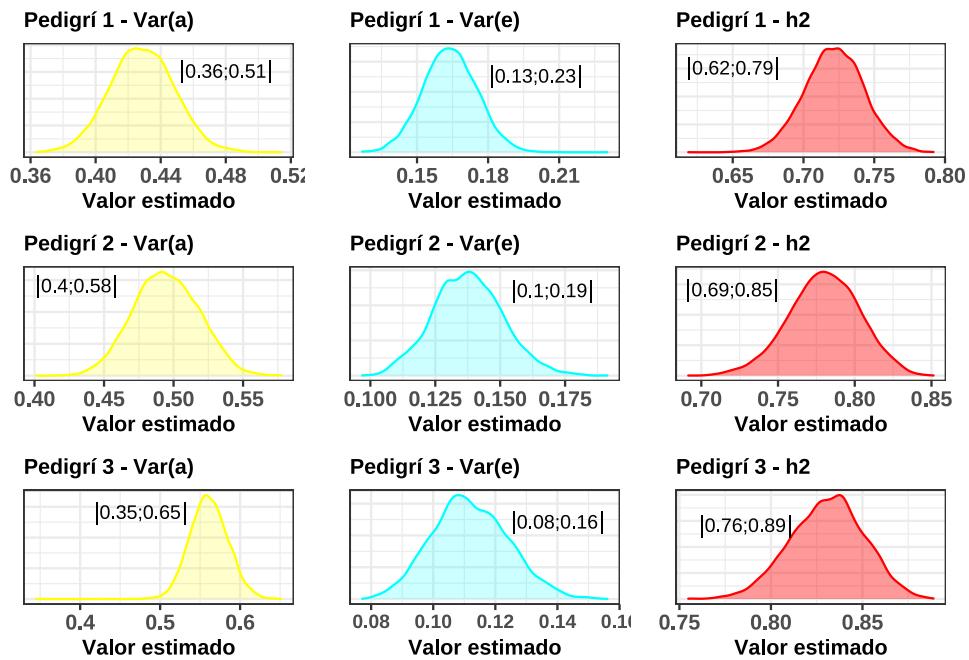
**Figura 5.2:** Análisis de componentes principales a partir de la matriz H resultante de la simulación de descendientes. Aquí, la matriz H se estimó a partir de 2091 individuos genotipados con densidad de marcador aproximada de 1.000, 10.000 y 100.000, y en el pedigrí 3 (primera réplica) mediante simulación de descendientes.

Los resultados de la estima de heredabilidad (en sentido estricto) para el carácter tiempo de floración, mediante procedimientos Bayesanos y de máxima verosimilitud restringida (REML) en los tres pedigríes generados a partir de la simulación ancestral, se resumen en la Tabla 5.1 y en la Figura 5.3. Esta estima, definida como la relación entre la varianza genética aditiva y la varianza fenotípica, varió de 0.72 a 0.84 mediante procedimientos Bayesanos, y de 0.79 a 0.84 mediante REML, valores que indican que el tiempo de floración es un carácter muy heredable y puede responder a la selección. En ambos enfoques (Bayesianos y REML), el BLUP proporcionó estimaciones de heredabilidad muy similares.

**Tabla 5.1:** Estimaciones de componentes de varianza y de heredabilidad para el carácter tiempo de floración mediante el método BLUP en los tres pedigríes simulados mediante simulación ancestral.

Parámetro	Bayesiano			REML		
	Ped. 1 <sup>1</sup>	Ped. 2	Ped. 3	Ped. 1	Ped. 2	Ped. 3
Varianza aditiva	0.43	0.50	0.56	0.47	0.50	0.56
Varianza ambiental	0.16	0.14	0.11	0.12	0.13	0.11
Heredabilidad	0.72	0.78	0.84	0.80	0.79	0.84

<sup>1</sup>Ped. 1 indica Pedigrí 1



**Figura 5.3:** Distribuciones posteriores de componentes de varianza y de heredabilidad obtenidas mediante enfoques Bayesianos. Los valores presentados en la Tabla 5.1 (para el enfoque Bayesiano), corresponden a la media calculada a partir de estas distribuciones.

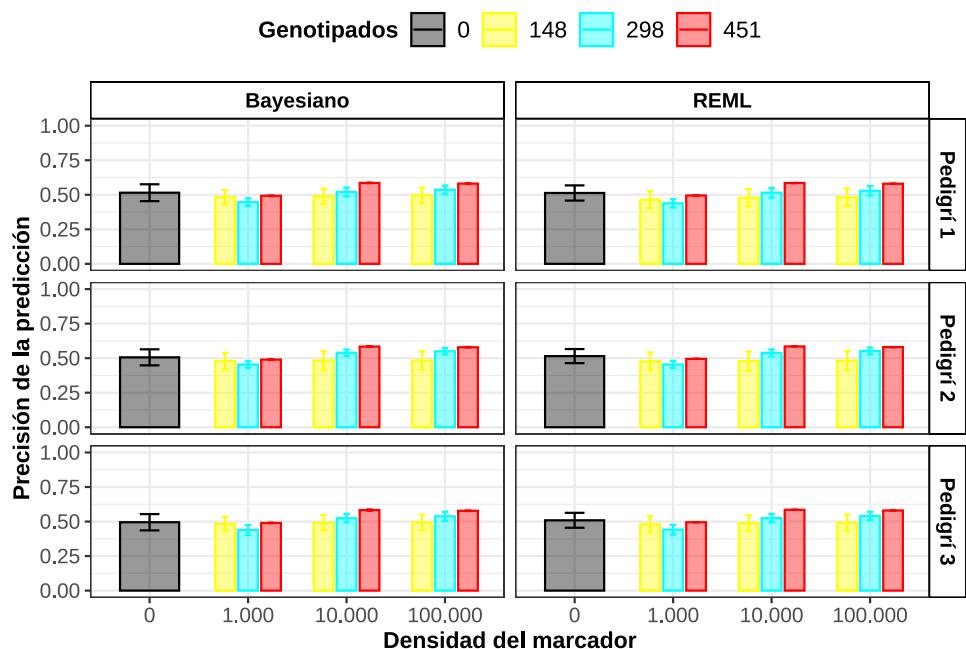
## 5.2. Precisión de la predicción mediante simulación ancestral

El efecto de la densidad del marcador y de la cantidad de individuos genotipados sobre la precisión de la predicción para el carácter tiempo de floración, se evaluó a través del coeficiente de correlación entre los valores fenotípicos observados y predichos. Los resultados se resumen en la Figura 5.4 y en el Anexo A.3. En general, y como era de esperar, la precisión de la predicción aumentó a medida que aumentaban tanto la densidad del marcador como la cantidad de individuos genotipados.

Con relación a la precisión de la predicción debido a la densidad del marcador en la simulación ancestral (Figura 5.4 y Anexo A.3), ésta fue mayor cuando no se utilizó información genómica en comparación a una densidad aproximada de 1.000 (en todos los casos de individuos genotipados). Al considerar una densidad del marcador por encima de 1.000, dicha precisión tendió a aumentar hasta alcanzar la precisión de la predicción más alta cuando se usaron todos los marcadores disponibles (aunque al considerar el genotipado de 148 individuos, la precisión de la predicción usando 10.000 y 100.000 marcadores no pudo superar el caso donde no se utilizó información

genómica). Sin embargo, dado el aumento marginal en la precisión de la predicción al aumentar de 10.000 a 100.000 marcadores, se puede afirmar que el límite en el aumento de la precisión se alcanzó en aproximadamente 10.000 marcadores. Por otra parte, el desvío estándar de las precisiones de predicción fue similar, independientemente del número de marcadores. Este comportamiento de la densidad del marcador sobre la precisión de la predicción curiosamente se observó por igual en los tres pedigríes simulados, tanto en el enfoque Bayesiano como en REML.

En cuanto a la cantidad de individuos genotipados, la precisión de la predicción aumentó sustancialmente al pasar de 148 a 451 individuos genotipados (en todos los casos de densidades de SNP). A diferencia de lo mencionado en el párrafo anterior para la densidad del marcador, no se pudo observar un punto en el que la precisión de la predicción se estabilizara en respuesta al aumento en la cantidad de individuos genotipados, situación que demuestra la importancia de un tamaño adecuado del conjunto de entrenamiento para construir modelos de predicción genómica. Adicionalmente, el desvío estándar de las precisiones disminuyó a medida que aumentaba la cantidad de individuos genotipados, hecho que pone de manifiesto que la diferencia en la precisión de la predicción es más alta en conjuntos de entrenamiento más pequeños. Al igual que lo mencionado en el párrafo anterior para la densidad del marcador, se observó un comportamiento muy similar en la precisión de la predicción debido a la cantidad de individuos genotipados en los tres pedigríes simulados y en ambos enfoques (Bayesiano y REML).

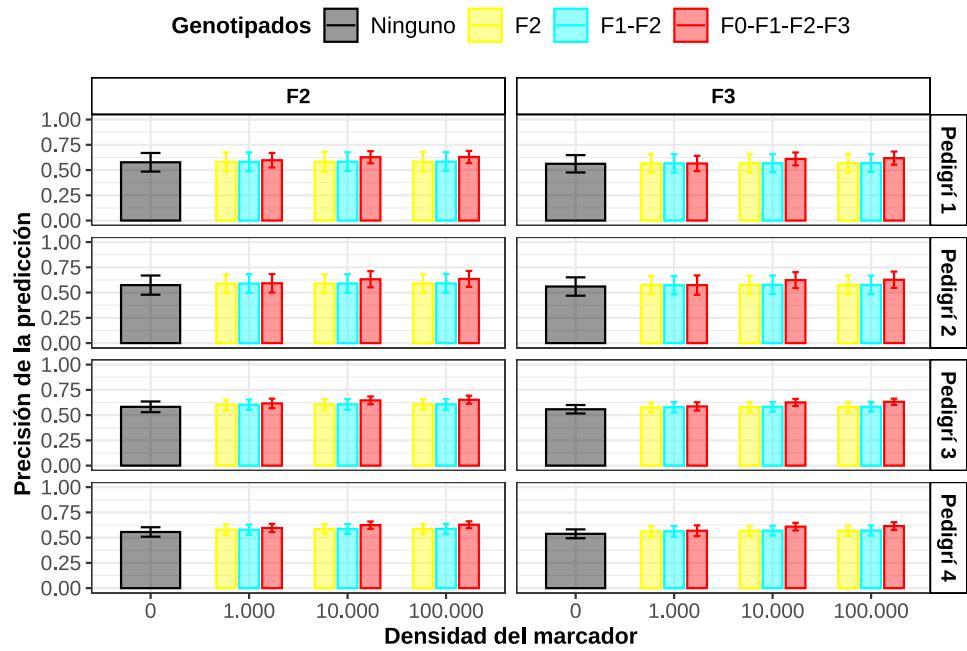


**Figura 5.4:** Precisión de la predicción con tamaños crecientes en el número de marcadores y de individuos genotipados en la simulación ancestral para los tres pedigríes simulados. Las barras representan la media de la precisión de la predicción de las 10 réplicas de cada pedigrí y las barras de error representan el desvío estándar como medida de variabilidad.

### 5.3. Precisión de la predicción mediante simulación de descendientes

El efecto de la densidad del marcador sobre la precisión de la predicción también se evaluó en la simulación de descendientes (Figura 5.5 y Anexo A.4), observándose en éste un comportamiento similar al observado con la simulación ancestral, en el sentido de que fue necesario de una cantidad mínima de marcadores para obtener una precisión equivalente al usar la cantidad máxima, aunque en esta simulación dicha cantidad mínima de marcadores fue solo de aproximadamente 1.000. Se observó un aumento irrelevante en la precisión de la predicción al pasar de 1.000 a 10.000 marcadores y de 10.000 a 100.000 marcadores (~ 0.0 %), para todos los casos de individuos genotipados. La diferencia con el escenario anterior (simulación ancestral) es que en este caso el desequilibrio fue mucho mayor, ya que los descendientes se generaron a partir de los parentales, mientras que en el caso anterior el desequilibrio es el observado en la población. Las mayores ganancias esperadas en la precisión de la predicción se observaron al pasar de 0 a 1.000, 10.000 y 100.000 marcadores, demostrando con esto que las predicciones basadas en información genómica son más precisas que las predicciones basadas solo en información del pedigrí. Este comportamiento de la densidad del marcador sobre la precisión de la predicción se observó por igual en los cuatro pedigríes simulados y considerando los individuos de las generaciones F<sub>2</sub> y F<sub>3</sub> como conjuntos de validación.

Respecto a la cantidad de individuos genotipados, también se observó una tendencia similar al observado con la simulación ancestral. La precisión de la predicción aumentó al aumentar la cantidad de individuos genotipados, observándose una estabilización al considerar un tamaño de genotipado entre solo los individuos de la generación F<sub>1</sub> y los individuos de las generaciones F<sub>1</sub> y F<sub>2</sub>. La mayor precisión de la predicción se observó al genotipar los individuos de todas las generaciones (F<sub>0</sub>, F<sub>1</sub>, F<sub>2</sub> y F<sub>3</sub>), logrando una ganancia esperada en la precisión mínima del 3.1 % (en el pedigrí 2) y máxima del 13.0 % (en el pedigrí 4), al considerar una densidad de 1.000 y 100.000 marcadores, respectivamente.



**Figura 5.5:** Precisión de la predicción con tamaños crecientes en el número de marcadores y de individuos genotipados en la simulación de descendientes para los cuatro pedigríes simulados. Las barras representan la media de la precisión de la predicción de las 10 réplicas de cada pedigrí y las barras de error representan el desvío estándar como medida de variabilidad.

# Capítulo 6

## Discusión

Los análisis genéticos cuantitativos se han basado tradicionalmente en la metodología de los modelos mixtos. Desde el desarrollo de esta metodología, el progreso genético logrado en los programas de mejora genética ha sido en gran parte al uso del pedigrí (Legarra, Lourenco, y Vitezica 2018), y a la matriz A que se obtiene a partir de este tipo información. La matriz A se ha utilizado ampliamente para estimar el parentesco entre individuos a través del BLUP. Sin embargo, en los últimos años, se han propuesto estimadores del parentesco distintos al parentesco basado en el pedigrí, ya que en este último se asume una relación promedio entre individuos fundadores igual a cero (Toro, García-Cortés, y Legarra 2011), además de que ignora el término de muestreo Mendeliano (Ratcliffe et al. 2017). Por otro lado, a pesar del interés que en su momento hubo en el uso del pedigrí para la mejora genética de animales, el uso de este tipo de información en plantas no fue relevante (Robertsen, Hjortshøj, y Janss 2019).

A partir del estudio de Fernando y Grossman (1989), en el que se propuso incorporar información de genotipos en el BLUP, Meuwissen, Hayes, y Goddard (2001) plantearon los métodos de selección genómica. Siguiendo las consideraciones precedentes, VanRaden (2007) desarrollo el GBLUP, el cual es un método de selección genómica que consiste en resolver un modelo mixto utilizando la matriz G en lugar de la matriz A. En relación con lo anterior, algunos estudios han informado la superioridad de los métodos de selección genómica sobre el BLUP en animales (Karimi et al. 2019) y en plantas (Morais-Júnior, Bresegheello, et al. 2018; Imai et al. 2019; Jurcic et al. 2021). Esto último debido a que los métodos de selección genómica basados en información de genotipos tienen en cuenta la variación genética no explicada por el pedigrí debido a la variación del muestreo Mendeliano (Daetwyler et al. 2007; Morais-Júnior, Duarte, et al. 2018), además de que revelan relaciones entre individuos que aparentemente no están relacionados en el pedigrí (Pszczola et al. 2012). Se ha demostrado también el uso potencial de los métodos de selección genómica para la mejora genética de los

cultivos de arroz a través de estudios de simulación (Xu et al. 2021).

Sin embargo, el costo para obtener un genotipado completo de individuos puede ser una limitante en el uso de los métodos anteriormente mencionados de selección genómica (Pérez-Rodríguez et al. 2017; Ratcliffe et al. 2017; Karimi et al. 2018), además que para su implementación se recomienda utilizar información de individuos sin genotipo y con genotipo en el mismo análisis de datos como forma de eliminar el sesgo asociado con la dependencia en el uso de individuos genotipados (Ratcliffe et al. 2017). Como solución a estos problemas se propuso el desarrollo del ssGBLUP, el cual es un procedimiento de evaluación genética donde la información genómica se extiende a los individuos sin genotipo a partir de una matriz de parentesco combinada con información de genotipos y del pedigrí (Legarra, Aguilar, y Misztal 2009; Misztal, Legarra, y Aguilar 2009). Por lo tanto, la selección genómica con base en el ssGBLUP puede considerarse una opción interesante para la mejora genética de los cultivos de arroz, debido a que permite obtener (de manera rentable) predicciones más precisas a través de una mejor comprensión del parentesco entre individuos.

Por otro lado, el factor más importante que determina el éxito de la selección genómica es la predicción precisa del GEBV (Nakaya y Isobe 2012). Según Morais-Júnior, Duarte, et al. (2018), la evaluación de la precisión se considera un requisito previo que permite comprender el potencial de un conjunto de datos para la predicción genómica. En el tiempo actual, se ha logrado grandes avances en la comprensión de los principales factores que afectan la precisión del GEBV, entre ellos la heredabilidad y la arquitectura genética del carácter, el modelo de predicción genómica, y el número de marcadores y de individuos genotipados (Nakaya y Isobe 2012; Desta y Ortiz 2014; Morais-Júnior, Breseghello, et al. 2018; Bartholomé, Thathapalli-Prakash, y Cobb 2021).

En relación al número de marcadores, de acuerdo con de los Campos et al. (2013) y Imai et al. (2019), la precisión de la predicción es más baja con un número insuficiente de marcadores debido a que la probabilidad de asociación entre dichos marcadores con el QTL tiende a disminuir. Sin embargo, aunque la teoría indica que cuantos más marcadores haya mejor será la predicción, también es cierto que a determinada densidad del marcador sera difícil mejorar significativamente dicha precisión (Wang et al. 2018). Esto último se observó en los resultados obtenidos en este estudio, ya que al utilizar una densidad aproximada entre 1.000 y 10.000 marcadores se evidenció una precisión similar a usar la mayor cantidad de marcadores disponibles. En este sentido, varios estudios han analizado la posibilidad de usar menor cantidad de marcadores sin afectar notablemente la precisión de la predicción genómica.

# **Capítulo 7**

# **Conclusiones**

# Apéndice A

## Anexos

### A.1. Función para el calculo de la matriz de parentesco combinada

```
fn.mH <- function(ped, mG) { # Esta función recibe como argumentos los datos con estructura (id / sire / dam / Gen (TRUE /FALSE)) y la matriz de relaciones genómicas.

  # 1. Se calcula la matriz de relaciones aditivas con base en
  # el pedigrí (A)

  ped_edit <- pedigreeMM::editPed( # Esta función ordena el pedigree.
                                    # digri.

    sire = ped$sire,
    dam = ped$dam,
    label = ped$id
  )
  pedi <- pedigreeMM::pedigree( # Aquí se usa la salida anterior
                                # (ya ordenado) y se crea un objeto de clase pedigree.

    sire = ped_edit$sire,
    dam = ped_edit$dam,
    label = ped_edit$label
  )
  Matrix_A <- pedigreeMM::getA(ped = pedi) # Esto dara la matriz
                                             # de relaciones aditivas A.
```

```

# 2. De lo anterior (Matriz_A) se extraen las partes correspondientes a individuos no genotipados (1) y genotipados (2)

# Individuos no genotipados:
A_11 <- Matrix_A[ped$Genotyped != 1, ped$Genotyped != 1]
# Individuos genotipados:
A_22 <- Matrix_A[ped$Genotiped == 1, ped$Genotiped == 1]
# Individuos no genotipados (en filas) y genotipados (en # columnas):
A_12 <- Matrix_A[ped$Genotiped != 1, ped$Genotiped == 1]
# Transpuesta de la anterior (individuos no genotipados en # columnas y genotipados en filas):
A_21 <- t(A_12)

# 3. Se coloca el nombre de las filas y y de las columnas de la matriz G según los individuos genotipados

rownames(mG) <- ped$id[ped$Genotiped == 1]
colnames(mG) <- ped$id[ped$Genotiped == 1]

# 4. Teniendo todos los componentes de la matriz H, se procede a su construcción

H_11 <- A_11 -
  (A_12 %*% solve(A_22) %*% A_21) +
  (A_12 %*% solve(A_22) %*% mG %*% solve(A_22) %*% A_21)
H_12 <- A_12 %*% solve(A_22) %*% mG
H_21 <- t(H_12)
H_22 <- mG

H_11_H_12 <- cbind(H_11, H_12)
H_21_H_22 <- cbind(H_21, H_22)
mH <- rbind(H_11_H_12, H_21_H_22)

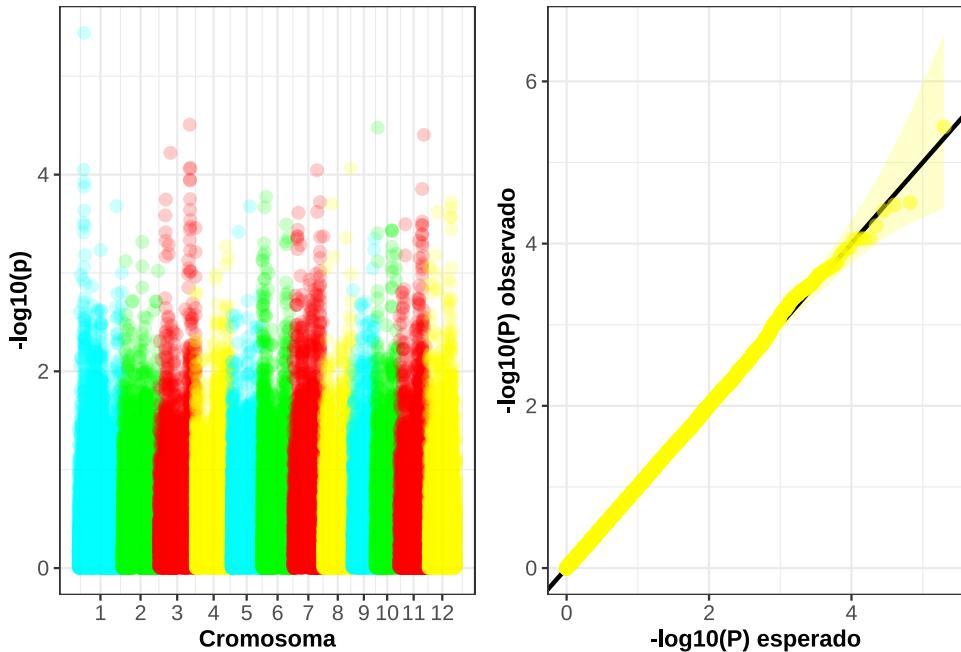
mH <- mH[order(as.numeric(rownames(mH))),
          order(as.numeric(colnames(mH)))]
mH <- Matrix(mH)

# 5. Finalmente se indica retornar la matriz H (mH)

return(mH)
}

```

## A.2. Visualización del GWAS



## A.3. Precisión de la predicción mediante simulación de pedigrí ancestral

Pedirí 1  
751 individuos en total

Genotipados	Bayesiano		REML	
	Media	Desvío	Media	Desvío
<b>Densidad 0</b>				
0	0.515	0.061	0.513	0.055
<b>Densidad 1.000</b>				
148	0.483	0.051	0.464	0.062
298	0.447	0.028	0.438	0.030
451	0.493	0.002	0.495	0.000
<b>Densidad 10.000</b>				
148	0.490	0.054	0.479	0.063
298	0.520	0.032	0.515	0.035
451	0.586	0.002	0.585	0.000

## Densidad 100.000

148	0.496	0.056	0.482	0.064
298	0.536	0.031	0.529	0.035
451	0.581	0.002	0.580	0.000

## Pedirí 2

1661 individuos en total

Genotipados	Bayesiano		REML	
	Media	Desvío	Media	Desvío
<b>Densidad 0</b>				
0	0.506	0.058	0.515	0.051
<b>Densidad 1.000</b>				
148	0.479	0.060	0.477	0.065
298	0.454	0.024	0.454	0.026
451	0.490	0.003	0.495	0.000
<b>Densidad 10.000</b>				
148	0.483	0.067	0.479	0.069
298	0.539	0.024	0.538	0.025
451	0.584	0.003	0.585	0.000
<b>Densidad 100.000</b>				
148	0.483	0.067	0.482	0.070
298	0.550	0.023	0.552	0.025
451	0.579	0.002	0.580	0.000

## Pedirí 3

2451 individuos en total

Genotipados	Bayesiano		REML	
	Media	Desvío	Media	Desvío
<b>Densidad 0</b>				
0	0.495	0.059	0.509	0.054
<b>Densidad 1.000</b>				
148	0.484	0.051	0.480	0.058
298	0.439	0.036	0.441	0.036
451	0.490	0.002	0.495	0.000

Densidad 10.000				
148	0.491	0.055	0.488	0.057
298	0.524	0.031	0.525	0.031
451	0.583	0.005	0.585	0.000
Densidad 100.000				
148	0.494	0.055	0.492	0.058
298	0.538	0.033	0.540	0.031
451	0.578	0.002	0.580	0.000

#### A.4. Precisión de la predicción mediante simulación de descendientes

Pedirí 1				
2061 individuos en total				
Genotipados	F2		F3	
	Media	Desvío	Media	Desvío
Densidad 0				
Ninguno	0.577	0.092	0.562	0.086
Densidad 1.000				
F0-F1-F2-F3	0.597	0.072	0.565	0.076
F1-F2	0.582	0.093	0.566	0.091
F2	0.580	0.094	0.565	0.091
Densidad 10.000				
F0-F1-F2-F3	0.627	0.060	0.610	0.064
F1-F2	0.583	0.093	0.568	0.089
F2	0.582	0.098	0.567	0.090
Densidad 100.000				
F0-F1-F2-F3	0.629	0.061	0.618	0.065
F1-F2	0.583	0.093	0.569	0.088
F2	0.581	0.097	0.568	0.090

Pedirí 2				
2071 individuos en total				
Genotipados	F2		F3	
	Media	Desvío	Media	Desvío

Densidad 0				
Ninguno	0.574	0.095	0.560	0.091
Densidad 1.000				
F0-F1-F2-F3	0.592	0.092	0.574	0.096
F1-F2	0.589	0.093	0.573	0.091
F2	0.589	0.094	0.574	0.091
Densidad 10.000				
F0-F1-F2-F3	0.632	0.080	0.623	0.079
F1-F2	0.590	0.092	0.577	0.092
F2	0.590	0.092	0.576	0.092
Densidad 100.000				
F0-F1-F2-F3	0.635	0.079	0.627	0.081
F1-F2	0.591	0.094	0.576	0.092
F2	0.590	0.093	0.576	0.092

Pedirí 3				
2091 individuos en total				
Genotipados	F2		F3	
	Media	Desvío	Media	Desvío
Densidad 0				
Ninguno	0.582	0.053	0.558	0.042
Densidad 1.000				
F0-F1-F2-F3	0.616	0.047	0.586	0.041
F1-F2	0.603	0.051	0.578	0.051
F2	0.601	0.051	0.576	0.050
Densidad 10.000				
F0-F1-F2-F3	0.646	0.039	0.626	0.035
F1-F2	0.607	0.053	0.583	0.048
F2	0.605	0.053	0.579	0.049
Densidad 100.000				
F0-F1-F2-F3	0.652	0.040	0.632	0.031
F1-F2	0.607	0.052	0.583	0.048
F2	0.607	0.052	0.581	0.049

Pedirí 4  
2131 individuos en total

Genotipados	F2		F3	
	Media	Desvío	Media	Desvío
<b>Densidad 0</b>				
Ninguno	0.556	0.047	0.538	0.044
<b>Densidad 1.000</b>				
F0-F1-F2-F3	0.596	0.041	0.569	0.053
F1-F2	0.579	0.050	0.563	0.053
F2	0.580	0.052	0.561	0.054
<b>Densidad 10.000</b>				
F0-F1-F2-F3	0.624	0.036	0.609	0.038
F1-F2	0.586	0.049	0.570	0.049
F2	0.586	0.047	0.568	0.050
<b>Densidad 100.000</b>				
F0-F1-F2-F3	0.628	0.034	0.615	0.038
F1-F2	0.586	0.050	0.571	0.050
F2	0.585	0.050	0.569	0.051

# Bibliografía

- Ahmadi, N., J. Bartholomé, T. V. Cao, y C. Grenier. 2020a. *Quantitative genetics, genomics and plant breeding*. 2nd edition. <https://doi.org/10.1079/9781789240214.0243>.
- Ahmadi, N., J. Bartholomé, T V. Cao, y C. Grenier. 2020b. *Genomic selection in rice: empirical results and implications for breeding*. 2nd edition. CAB International.
- Bartholomé, J., P. Thathapalli-Prakash, y J. N. Cobb. 2021. *Genomic prediction: progress and perspectives for rice improvement*. <https://doi.org/10.48550/arXiv.2109.14781>.
- Blasco, A. 2021. *Mejora genética animal*. 1st edition. EDITORIAL SÍNTESIS, S. A.
- Blasco, A., y M. Á. Toro. 2014. «A short critical history of the application of genomics to animal breeding». *Livestock Science* 166: 4-9.
- Breseghello, F. 2013. «Traditional and modern plant breeding methods with examples in rice (*Oryza sativa L.*)». *Journal of Agricultural and Food Chemistry* 61: 8277-86. <https://doi.org/10.1021/jf305531j>.
- Caamal-Pat, D., P. Pérez-Rodríguez, J. Crossa, C. Velasco-Cruz, S. Pérez-Elizalde, y M. Vázquez-Peña. 2021. «lme4GS: An R-package for genomic selection». *Genetics* 12. <https://doi.org/10.3389/fgene.2021.680569>.
- Caligari, P. D. S., y J. Brown. 2017. *Plant breeding, practice*. 2nd edition. Vol. 2. Elsevier Ltd. <https://doi.org/10.1016/B978-0-12-394807-6.00195-7>.
- Cappetta, E., G. Andolfo, A. Di Matteo, A. Barone, L. Fruscianente, y M. R. Ercolano. 2020. «Accelerating tomato breeding by exploiting genomic selection approaches». *Plants* 9 (9). <https://doi.org/10.3390/plants9091236>.
- Cros, D., L. Sánchez, B. Cochard, P. Samper, M. Denis, J. M. Bouvet, y J. Fernández. 2014. «Estimation of genealogical coancestry in plant species using a pedigree reconstruction algorithm and application to an oil palm breeding population». *Theoretical and Applied Genetics* 127: 981-94. <https://doi.org/10.1007/s00122-014-2273-3>.
- Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. de los Campos, J. Burgueño, et al. 2017. «Genomic selection in plant breeding: methods, models, and perspectives». *Trends in Plant Science*,

- 961-75. <https://doi.org/10.1016/j.tplants.2017.08.011>.
- Daetwyler, H. D., B. Villanueva, P. Bijma, y J. A. Woolliams. 2007. «Inbreeding in genome-wide selection». *Journal of Animal Breeding and Genetics* 124: 369-76.
- Dekkers, J., y F. Hospital. 2002. «The use of molecular genetics in the improvement of agricultural populations». *Nature Reviews Genetics* 3: 22-32. <https://doi.org/10.1038/nrg701>.
- de los Campos, G., J. H. Hickey, R. Pong-Wong, H. D. Daetwyler, y M. P. L. Calus. 2013. «Whole-genome regression and prediction methods applied to plant and animal breeding». *Genetics* 193: 327-45. <https://doi.org/10.1534/genetics.112.143313>.
- Desta, Z. A., y R. Ortiz. 2014. «Genomic selection: genome-wide prediction in plant improvement». *Trends in Plant Science* 19 (9): 592-601.
- Fernando, R., y M. Grossman. 1989. «Marker assisted selection using best linear unbiased prediction». *Genetics Selection Evolution* 21: 467. <https://doi.org/10.1186/1297-9686-21-4-467>.
- Fernández, J., y M. Á. Toro. 2006. «A new method to estimate relatedness from molecular markers». *Molecular Ecology* 15: 1657-67.
- Freeman, A. E. 1991. «C. R. Henderson: contributions to the dairy industry». *Journal of Dairy Science* 74 (11): 4045-51. [https://doi.org/10.3168/jds.S0022-0302\(91\)78600-1](https://doi.org/10.3168/jds.S0022-0302(91)78600-1).
- Fukuoka, S., K. Ebana, T. Yamamoto, y M. Yano. 2010. «Integration of genomics into rice breeding». *Rice* 3: 131-37. <https://doi.org/10.1007/s12284-010-9044-9>.
- Grenier, C., T. V. Cao, Y. Ospina, C. Quintero, M. H. Châtel, J. Tohme, B. Courtois, y N. Ahmadi. 2015. «Accuracy of genomic selection in a rice synthetic population developed for recurrent selection breeding». *PLoS ONE* 10 (8). <https://doi.org/10.1371/journal.pone.0136594>.
- Guimarães, E. P. 2009. *Rice Breeding*. [https://doi.org/10.1007/978-0-387-72297-9\\_2](https://doi.org/10.1007/978-0-387-72297-9_2).
- Guo, Z., D. M. Tucker, C. J. Basten, H. Gandhi, E. Ersoz, B. Guo, Z. Xu, D. Wang, y G. Gay. 2014. «The impact of population structure on genomic prediction in stratified populations». *Theoretical and Applied Genetics* 127: 749-62. <https://doi.org/10.1007/s00122-013-2255-x>.
- Henderson, C. R. 1975. «Best linear unbiased estimation and prediction under a selection model». *Biometrics* 31: 423-47.
- . 1986. *Applications of linear models in animal breeding*. 2nd edition.
- Hickey, J. M., T. Chiurugwi, I. Mackay, W. Powell, y Implementing Genomic Selection in CGIAR Breeding Programs Workshop Participants. 2017. «Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery». *Nature Genetics* 49 (9): 1297-1303. <https://doi.org/10.1038/ng.3920>.
- Holland, J. B. 2014. *Breeding: plants, modern*. Vol. 2. Elsevier Inc. <https://doi.org/10.1016/B978-0-444-52512-3.00226-6>.

- Imai, A., T. Kuniga, T. Yoshioka, K. Nonaka, N. Mitani, H. Fukamachi, N. Hiehata, M. Yamamoto, y T. Hayashi. 2019. «Single-step genomic prediction of fruit-quality traits using phenotypic records of non-genotyped relatives in citrus». *PLoS ONE* 14 (8). <https://doi.org/10.1371/journal.pone.0221880>.
- Jurcic, E. J., P. V. Villalba, P. S. Pathauer, D. A. Palazzini, G. P. J. Oberschelp, L. Harrand, M. N. Garcia, et al. 2021. «Single-step genomic prediction of Eucalyptus dunni using different identity-by-descent and identity-by-state relationship matrices». *Heredity* 127: 176-89.
- Karimi, K., M. Sargolzaei, G. S. Plastow, Z. Wang, y Y. Miar. 2018. «Effect of hidden relatedness on single-step genetic evaluation in an advanced open-pollinated breeding program». *Journal of Heredity*, 802-10. <https://doi.org/10.1093/jhered/esy051>.
- . 2019. «Opportunities for genomic selection in american mink: a simulation study». *PLoS ONE* 14 (3). <https://doi.org/10.1371/journal.pone.0213873>.
- Kyselova, J., L. Tichý, y K. Jochová. 2021. «The role of molecular genetics in animal breeding: a minireview». *Czech Journal of Animal Science* 66 (4): 107-11. <https://doi.org/10.17221/251/2020-CJAS>.
- Legarra, A., I. Aguilar, y I. Misztal. 2009. «A relationship matrix including full pedigree and genomic information». *Journal of Dairy Science* 92: 4656-63. <https://doi.org/10.3168/jds.2009-2061>.
- Legarra, A., O. F. Christensen, I. Aguilar, y I. Misztal. 2014. «Single Step, a general approach for genomic selection». *Livestock Science*. <https://doi.org/10.1016/j.livsci.2014.04.029>.
- Legarra, A., D. Lourenco, y Z. G. Vitezica. 2018. *Bases for genomic prediction*.
- Lourenco, D., A. Legarra, S. Tsuruta, Y. Masuda, I. Aguilar, y I. Misztal. 2020. «Single-step genomic evaluations from theory to practice: using SNP chips and sequence data in BLUPF90». *Genes* 11: 790. <https://doi.org/doi:10.3390/genes11070790>.
- Medina, C. A., H. Kaur, I. Ray, y L. X. Yu. 2021. «Strategies to Increase Prediction Accuracy in Genomic Selection of Complex Traits in Alfalfa (*Medicago sativa L.*)». *Cells* 10 (12). <https://doi.org/10.3390/cells10123372>.
- Meuwissen, T. H. E., B. J. Hayes, y M. E. Goddard. 2001. «Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps». *Genetics* 157: 1819-29.
- Misztal, I., S. E. Aggrrey, y W. M. Muir. 2012. «Experiences with a single-step genome evaluation». *Poultry Science* 92: 2530-4.
- Misztal, I., A. Legarra, y I. Aguilar. 2009. «Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information». *Journal of Dairy Science* 92: 4648-55. <https://doi.org/10.3168/jds.2009-2064>.
- Misztal, I., D. Lourenco, y A. Legarra. 2020. «Current status of genomic

- evaluation». *Journal of Animal Science* 98 (4): 1-14. <https://doi.org/10.1093/jas/skaa101>.
- Morais-Júnior, O. P., F. Breseghello, J. Batista-Duarte, A. S. G. Coelho, T. C. O. Borba, J. T. Aguiar, P. C. F. Neves, y O. P. Morais. 2018. «Assessing prediction models for different traits in a rice population derived from a recurrent selection program». *Crop Science* 58: 1-13. <https://doi.org/10.2135/cropsci2018.02.0087>.
- Morais-Júnior, O. P., J. B. Duarte, F. Breseghello, A. S. G. Coelho, O. P. Morais, y A. M. Magalhães-Júnior. 2018. «Single-step reaction norm models for genomic prediction in multienvironment recurrent selection trials». *Crop Science* 58: 592-607. <https://doi.org/10.2135/cropsci2017.06.0366>.
- Nakaya, A., y S. N. Isobe. 2012. «Will genomic selection be a practical method for plant breeding?» *Annals of Botany* 110: 1303-16.
- Nelson, R. M., M. E. Pettersson, y Ö. Carlberg. 2012. «A century after Fisher: time for a new paradigm in quantitative genetics». *Trends in Genetics* 29 (9): 669-76.
- Pérez-Enciso, M., L. Ramírez-Ayala, y L. M. Zingaretti. 2020. «SeqBreed: a python tool to evaluate genomic prediction in complex scenarios». *Genetion Selection Evolution* 52 (7). <https://doi.org/10.1186/s12711-020-0530-2>.
- Pérez-Rodríguez, P., J. Crossa, J. Rutkoski, J. Poland, R. Singh, A. Legarra, E. Autrique, J. Burgueño G. de los Campos, y S. Dreisigacker. 2017. «Single-step genomic and pedigree genotype x environment interaction models for predicting wheat lines in international environments». *Plant Genome* 10 (2). <https://doi.org/10.3835/plantgenome2016.09.0089>.
- Pérez-Rodríguez, P., y G. de los Campos. 2014. «Genome-wide regression and prediction with the BGLR statistical package». *Genetics* 198 (2): 483-95. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4196607/>.
- Pszczola, M., T. Strabel, J. A. M. van Arendonk, y M. P. L. Calus. 2012. «The impact of genotyping different groups of animals on accuracy when moving from traditional to genomic selection». *Journal of Dairy Science* 95: 5412-21. <https://doi.org/10.3168/jds.2012-5550>.
- Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, D. Bender, J. Maller, et al. 2007. «Plink: a toolset for whole-genome association and population-based linkage analysis». *American Journal of Human Genetics* 81: 981-94. <http://pngu.mgh.harvard.edu/purcell/plink/>.
- Ratcliffe, B., O. G. El-Dien, E. P. Cappa, I. P., J. Klápková, C. Chena, y Y. A. El-Kassaby. 2017. «Single-step BLUP with varying genotyping effort in open-pollinated *Picea glauca*». *G3 Genes/Genomes/Genetics* 7: 935-42. <https://doi.org/10.1534/g3.116.037895>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

- Robertsen, C. D., R. L. Hjortshøj, y L. L. Janss. 2019. «Genomic selection in cereal breeding». *Agronomy* 9: 95. <https://doi.org/10.3390/agronomy9020095>.
- Schaeffer, L. R. 1991. «C. R. Henderson: contributions to predicting genetic merit». *Journal of Dairy Science* 74 (11): 4052-66. [https://doi.org/10.3168/jds.S0022-0302\(91\)78601-3](https://doi.org/10.3168/jds.S0022-0302(91)78601-3).
- Searle, S. R. 1991. «C. R. Henderson, the statistician; and his contributions to variance components estimation». *Journal of Dairy Science* 74 (11): 4035-44. [https://doi.org/10.3168/jds.S0022-0302\(91\)78599-8](https://doi.org/10.3168/jds.S0022-0302(91)78599-8).
- Tan, C., C. Bian, D. Yang, N. Li, Z. Wu, y X. Hu. 2017. «Application of genomic selection in farm animal breeding». *Hereditas* 39 (11): 1033-45. <https://doi.org/10.16288/j.yczz.17-286>.
- Tong, H., y Z. Nikoloski. 2021. «Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data». *Journal of Plant Physiology* 257: 153354. <https://doi.org/10.1016/j.jplph.2020.153354>.
- Toro, M. Á., L. A. García-Cortés, y A. Legarra. 2011. «A note on the rationale for estimating genealogical coancestry from molecular markers». *Genetics Selection Evolution* 43: 27. <https://doi.org/10.1186/1297-9686-43-27>.
- Turelli, M. 2017. «Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps». *Theoretical Population Biology* 118: 46-49.
- VanRaden, P. M. 2007. «Efficient methods to compute genomic predictions». *Journal of Dairy Science* 91: 4414-23.
- Van-Rossum, G., y Jr. F. L Drake. 1995. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.
- Vourlaki, I., R. Castanera, S. Ramos-Onsins, J. Casacuberta, y M. Pérez-Enciso. s. f. «Transposable element polymorphisms improve prediction of complex agronomic traits in rice». *Frontiers in Plant Science*.
- Wang, J., J. Crossa, y J. Gai. 2020. «Quantitative genetic studies with applications in plant breeding in the omics era». *The Crop Journal* 8: 683-87. <https://doi.org/10.1016/j.cj.2020.09.001>.
- Wang, X., Y. Xu, Z. Hu, y C. Xu. 2018. «Genomic selection methods for crop improvement: Current status and prospects». *The Crop Journal* 6: 330-40. <https://doi.org/10.1016/j.cj.2018.03.001>.
- Wing, R. A., M. D. Purugganan, y Q. Zhang. 2018. «The rice genome revolution: from an ancient grain to Green Super Rice». *Nature Reviews Genetics* 19: 505-17. <https://doi.org/10.1038/s41576-018-0024-z>.
- Xu, Y., K. Ma, Y. Zhao, X. Wang, K. Zhou, G. Yu, C. Li, et al. 2021. «Genomic selection: a breakthrough technology in rice breeding». *Nature Reviews Genetics* 9: 669-77. <https://doi.org/10.1016/j.cj.2021.03.008>.
- Xua, S., D. Zhub, y Q. Zhang. 2014. «Predicting hybrid performance in rice using genomic best linear unbiased prediction». *Proceedings of the National Academy of Sciences of the United States of America* 111 (34): 12456-61. <https://doi.org/10.1073/pnas.1413750111>.

Yang, J., S. H. Lee, M. E. Goddard, y P. M. Visscher. 2011. «GCTA: A Tool for Genome-wide Complex Trait Analysis». *American Journal of Human Genetics* 88 (1): 76-82. <https://doi.org/10.1016/j.ajhg.2010.11.011>.

## Agradecimientos

