

SELECCIÓN GENÓMICA EN
ARROZ: EFECTO DEL NÚMERO
DE MARCADORES Y DE
INDIVIDUOS GENOTIPADOS
SOBRE LA PRECISIÓN DE LA
PREDICCIÓN

Esta tesis se escribió usando los paquetes de R (R) Markdown, L^AT_EX , bookdown y amsterdown.



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



Selección genómica en arroz: efecto del número de marcadores y de individuos genotipados sobre la precisión de la predicción

Tesis académica para obtener
el grado de Máster en Mejora Genética y
Biotecnología 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

1. Resumen	1
2. Revisión de literatura	2
2.1. Breve historia hacia la selección genómica	2
2.2. La selección genómica	5
2.3. Breve descripción de la mejora genética en arroz	16
3. Objetivos	20
4. Materiales y métodos	21
4.1. Recurso vegetal y datos fenotípicos	21
4.2. Predicción basada en información de pedigrí e información genómica	23
4.3. Estimación de la heredabilidad	24
4.4. Generación de pedigríes ancestrales, subconjuntos de datos y habilidad predictiva	24
4.5. Simulación de fenotipos y genotipos, subconjuntos de datos y habilidad predictiva	27
5. Resultados y discusión	30
5.1. Fenotipo y heredabilidad	30
6. Conclusiones	33
A. Anexos	34
A.1. Función ¹ para el calculo de la matriz de parentesco combinada	34
A.2. Visualización del GWAS	36
A.3. Habilidad predictiva	36
A.4. Habilidad predictiva	38
Bibliografía	41

¹http://rstudio-pubs-static.s3.amazonaws.com/378595_edda8cfe948a4786ae6dd74962cf6e94.html

ÍNDICE GENERAL

VI

Agradecimientos

45

Capítulo 1

Resumen

Insert abstract.

Possibly insert citation here.

Capítulo 2

Revisión de literatura

2.1. Breve historia hacia la selección genómica

La historia de la genética tanto cuantitativa como molecular se remonta a la contribución de muchas personas (Figura 1.1), hecho que permitió la conexión entre ambas disciplinas y el desarrollo de lo que hoy en día se conoce como selección genómica.

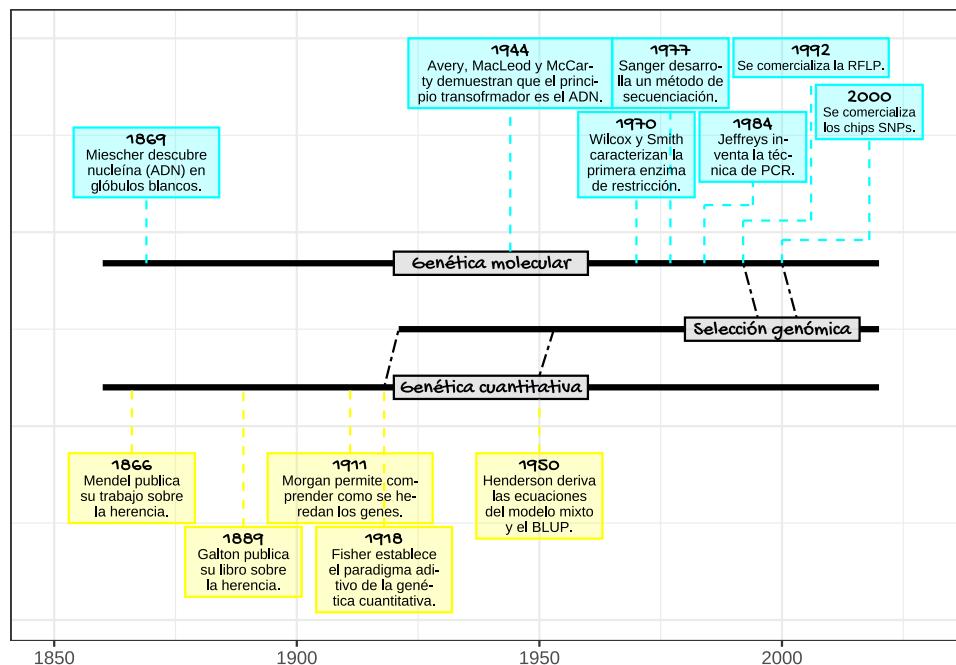


Figura 1.1: Cronología de las disciplinas de la genética molecular y la genética cuantitativa. Varios descubrimientos permitieron la conexión entre ambas disciplinas lo que permitió el desarrollo de la selección genómica.

Figura adaptada de Nelson, Pettersson, y Carlborg (2012).

La genética cuantitativa se formó hace más de un siglo en ausencia directa de datos genéticamente observables (Nelson, Pettersson, y Carlborg 2012). Esta disciplina se formó gracias a los avances teóricos de Ronald Fisher quien proporcionó una teoría que hizo posible interpretar los descubrimientos de la genética biométrica dentro de los estudios 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, supuso 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). A partir de este entonces, la genética cuantitativa fue extremadamente productiva a medida que fue adhiriéndose a la teoría del modelo infinitesimal.

Se denomina como valor de cría estimado (EBV) al 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) (Tong y Nikoloski 2021). 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, hoy en día se reconoce que dicho procedimiento ignora la base física de la herencia (el ADN), y utiliza una representación conceptual elemental de como la información genética es heredada (esto es, ambos progenitores deben aportar la mitad de la información genética a su descendencia) (Legarra et al. 2014).

En otro orden de ideas, el rápido desarrollo de la genética molecular a partir de los años 60 permitió comprender mejor los mecanismos de la herencia. 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 (Legarra et al. 2014), 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. Este enfoque, conocido como etiquetado de genes o mapeo de QTL, permitió identificar la genética subyacente a la variación fenotípica de un carácter (de los Campos et al. 2013; Legarra, Lourenco, y Vitezica 2018; Qanbari 2020).

Tanto en animales como en plantas, el interés principal en el mapeo de QTL consistió en usarse en un método conocido como selección asistida por marcadores (MAS) (Blasco y Toro 2014), proceso en el cual 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 1.2).

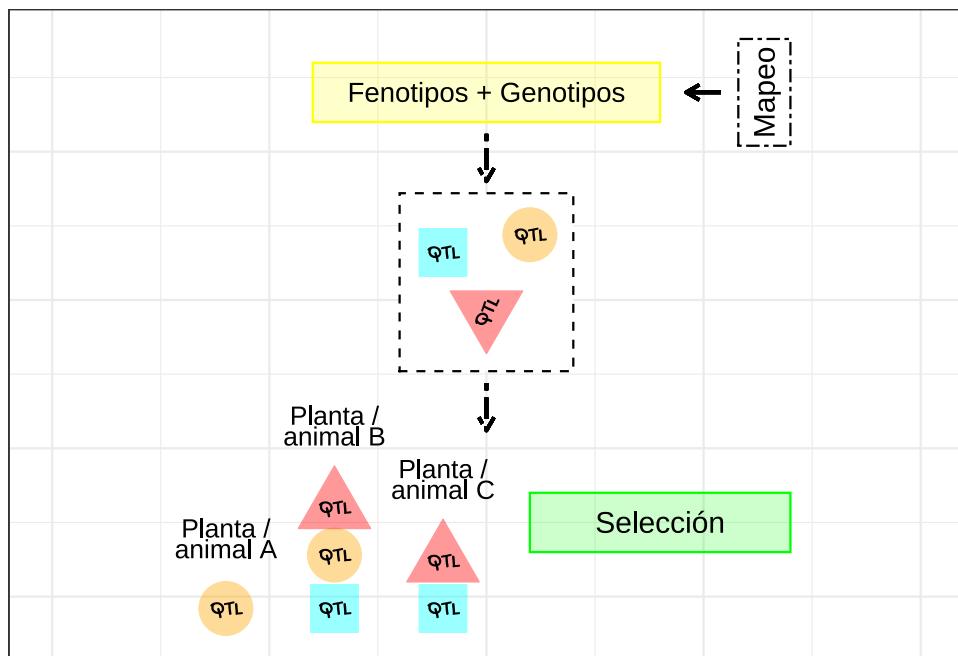


Figura 1.2: Esquema de la MAS. En la MAS, los fenotipos y genotipos de la población de mapeo se analizan usando un modelo estadístico, identificando con ello relaciones significativas entre fenotipos y genotipos. Por último, se seleccionan los individuos favorables con base en datos de genotipo. Figura 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 económicamente importantes, 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 económicamente importantes 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 del 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

2.2.1. Definición de la selección genómica

Se denomina selección genómica a una serie de métodos que usan decenas de miles de 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 1.3). Por tanto, la selección genómica suele ser vista como una forma de MAS en la que se seleccionan individuos según el GEBV en lugar de pocos QTL (Nakaya y Isobe 2012).

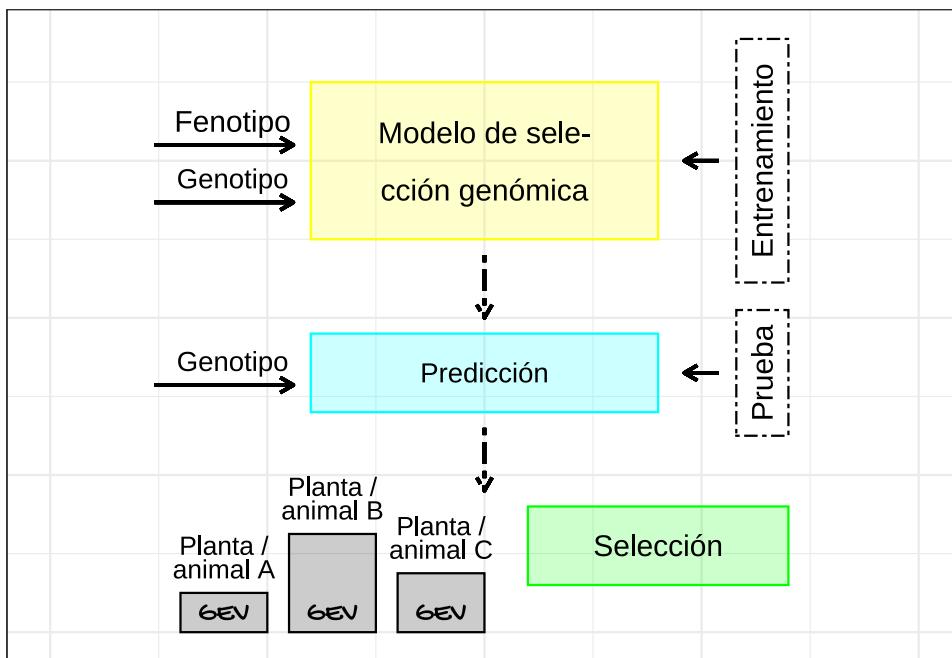


Figura 1.3: Esquema de la selección genómica. La selección genómica utiliza un modelo estadístico, diseñado a partir de datos genotípicos y fenotípicos en una población de entrenamiento, para predecir el GEBV de los individuos en una población de prueba con datos genotípicos. Por último, los individuos 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 biológico sobre el carácter, a partir de este hecho biológico si que se garantizaría una asociación (no observada) entre el QTL y el carácter (Legarra, Lourenco, y Vitezica 2018; Grinberg, Orhobor, y King 2020; Qanbari 2020).

2.2.2. Métodos estadísticos en la selección genómica

En la selección genómica, la relación genotipo-fenotipo puede ser representada como un modelo lineal (Figura 1.4). 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).

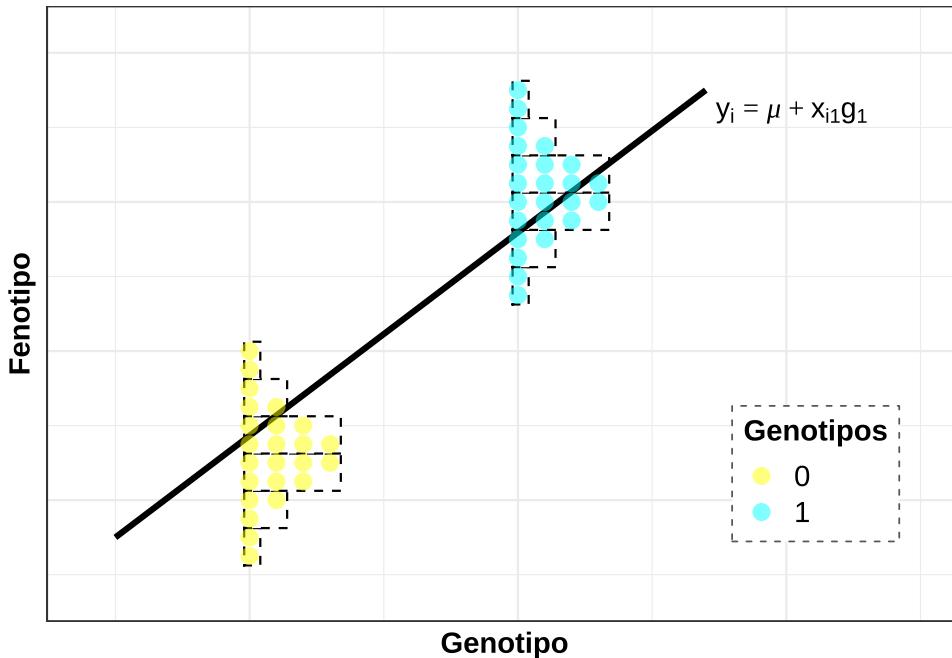


Figura 1.4: Relación genotipo-fenotipo de individuos (círculos amarillo y azul) para un solo marcador. y_i y x_{i1} denotan los fenotipos y genotipos, y μ y g_i son los parámetros a determinar. Los genotipos bialélicos se codifican como 0 y 1, y los fenotipos se distribuyen de acuerdo a una normal. Figura 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, μ 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:

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

donde \mathbf{y} es un vector de longitud igual al número de individuos ($1, 2, 3, \dots, n$) que representa al fenotipo, \mathbf{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), \mathbf{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 puede afectar el uso de la regresión por mínimos cuadrados, ya que esta 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 1.5).

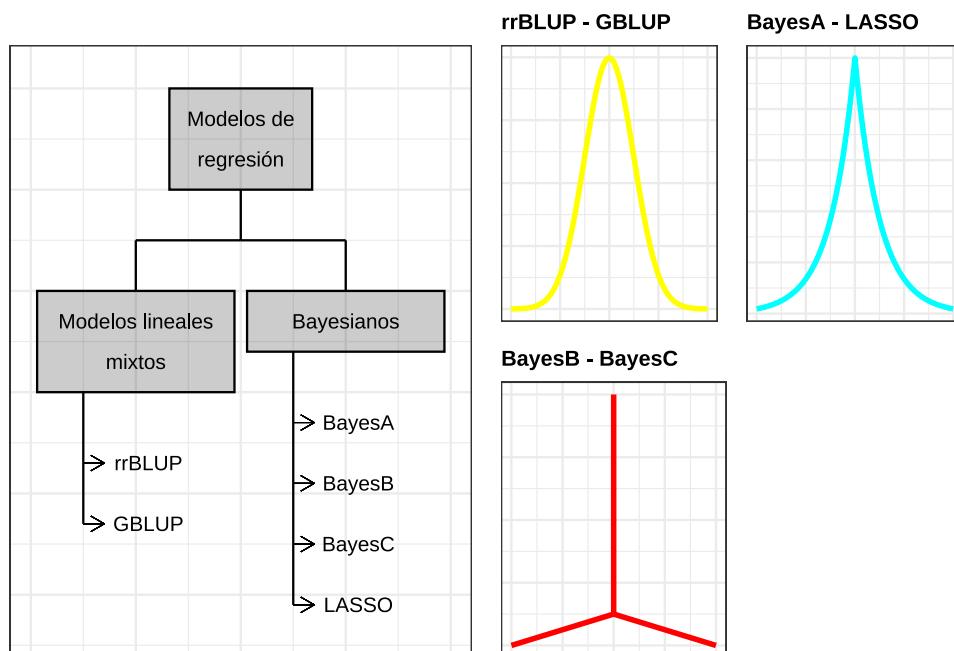


Figura 1.5: Enfoques estadísticos de la selección genómica. Cada uno de estos enfoques suponen distintas distribuciones de efectos de los marcadores sobre el carácter. Figura adaptada de de los Campos et al. (2013) y de Tong y Nikoloski (2021).

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 regresión de crestas (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 λ 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 σ_g^2 es la varianza del efecto del marcador y σ_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 efectos grandes, otros medianos y otros pequeños. En cuanto a BayesB, este 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). De la misma manera, BayesA y BayesB se diferencian en el procedimiento de estimación: BayesA utiliza el método de cadenas de Markov Monte Carlo (MCMC) para ... (Tan et al. 2017).

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 pedigree (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 λ 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, caso contrario al parentesco derivado al usar la matriz G que es observado, debido a lo cual si puede evidenciar de forma precisa el parentesco real entre dos individuos (Figura 1.6).

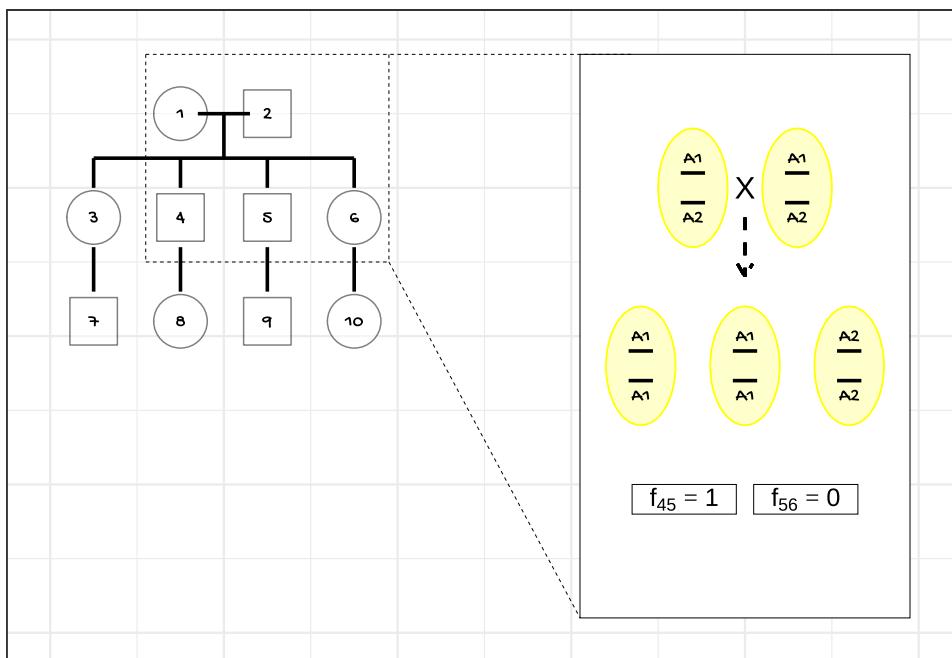


Figura 1.6: Parentesco observado entre individuos emparentados.

En comparación a los métodos Bayesianos descritos anteriormente, en el GBLUP no es necesario usar una población de entrenamiento para estimar el efecto del marcador de ADN y luego predecir el GEBV. En su lugar, en el GBLUP se pueden colocar directamente a los individuos con fenotipo y sin fenotipo en el mismo modelo, y al mismo tiempo predecir el GEBV, y calcular su precisión (Tan et al. 2017). En cuanto a la velocidad de cálculo, el GBLUP es mucho más rápido que los métodos Bayesianos, por lo que 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 asignado 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.

2.2.3. 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 ya que los modelos estadísticos generalmente se construyen en base a esta información. Esta situación puede ser desventajosa al momento de 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 1.7) (Legarra, Aguilar, y Misztal 2009; Misztal, Legarra, y Aguilar 2009; Misztal, Aggrrey, y Muir 2012; Legarra et al. 2014; Misztal, Lourenco, y Legarra 2020).

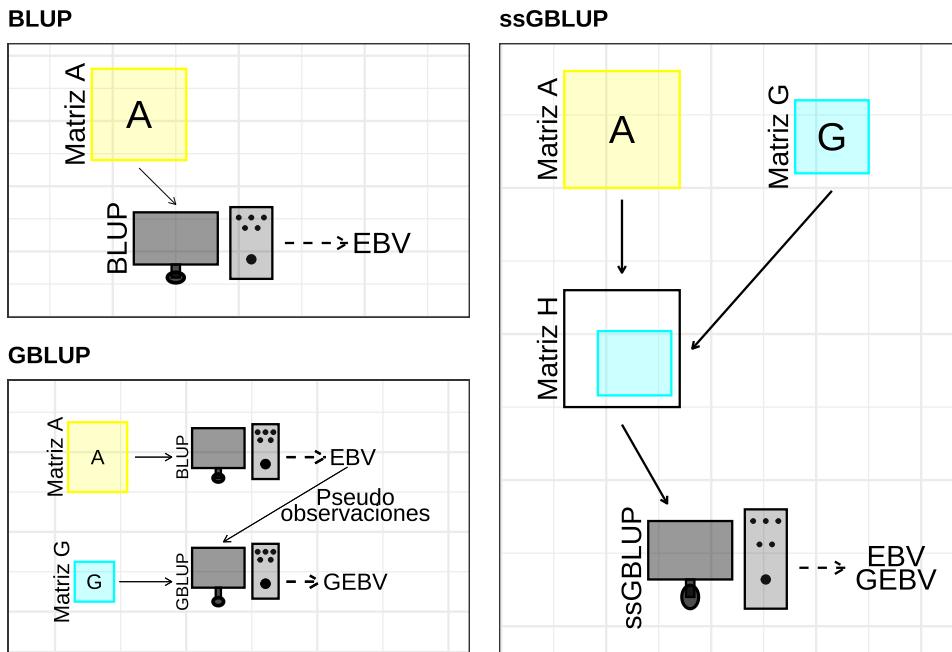


Figura 1.7: Esquema de comparación del BLUP, GBLUP y ssGBLUP. El GBLUP es un proceso de tres pasos en el que los individuos, con base en su información fenotípica y de pedigrí, son evaluados inicialmente mediante el BLUP; luego, a partir de los pseudo-fenotipos resultantes de esta evaluación inicial, 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 es tendente a cometer errores (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 EBV o 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 1.7). 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 tiende a simplificarse 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 pedigree.



Recuadro 1.1

Conociendo que la matriz H equivale a:

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

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 sí lo están (g_2), con base en la expresión:

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

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:

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

Con un poco de álgebra, finalmente se tendría la expresión correspondiente a $\text{var}(g_1)$:

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

En relación a la expresión $\text{var}(g_2)$, esta 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).

2.2.4. 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). Así mismo, la habilidad predictiva de la selección genómica depende de las características de los datos genotípicos, esto es, el número de individuos genotipados y la densidad del marcador (Ahmadi, Bartholomé, Cao, et al. 2020b; Tong y Nikoloski 2021).

En relación a la densidad del marcador, 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). Así mismo, 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. Breve descripción de la mejora genética en arroz

2.3.1. 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. Esto a partir de la selección artificial. La clave del éxito de los antiguos domesticadores y cultivadores consistió en cruzar individuos portadores de caracteres deseables, al comprender que los descendientes podrían heredar estas características, pese a que para esos tiempos se desconocían los mecanismos biológicos de la herencia (Holland 2014). Hoy en día, se conoce como mejora genética al proceso de mejorar los caracteres fenotípicos deseables en animales y plantas mediante selección artificial (Tong y Nikoloski 2021), siendo la misma reconocida como una intrincada integración de ciencia y practicidad (Caligari y Brown 2017).

En términos generales, la mejora genética se puede organizar en tres procesos, siendo estos la producción de variación genética, la selección entre la variación y la multiplicación para uso comercial (Figura 1.8).

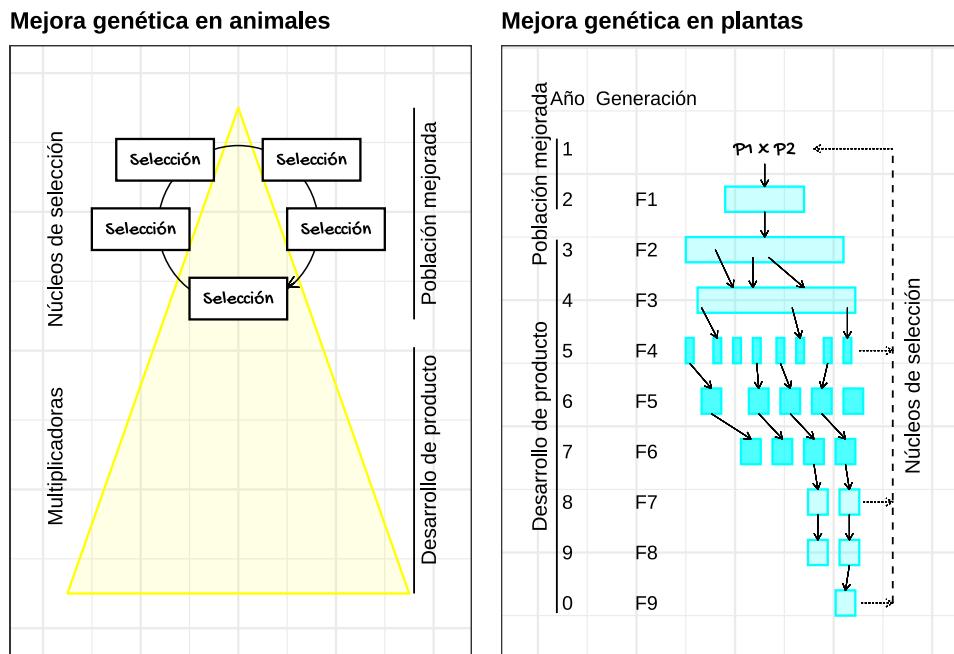


Figura 1.8: Esquema de la mejora genética en animales y en plantas.

Figura adaptada de Hickey et al. (2017).

El primer requisito en la mejora genética, como se mencionó en el párrafo anterior, consiste en producir variación genética de los caracteres que se desean a mejorar. Tanto en animales como en plantas, la producción de 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.

Una vez producida la variación genética, 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 genes favorables. Sin embargo, los métodos de selección históricamente 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 utilizar enfoques biométricos para predecir el EBV mediante la combinación de información fenotípica y de pedigree (propriamente el BLUP), y con ello tomar las decisiones de selección (Hickey et al. 2017).

Justo después de aumentar la frecuencia de genes deseables, 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 recicla a los núcleos de selección, mientras que en plantas, al darle importancia a la selección de productos mejorados en forma de variedades vegetales, dichas variedades pueden ser usadas como progenitores en un ciclo nuevo de cultivo.

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 se ha quedado atrás (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- Varios mejoradores genéticos de plantas argumentan que se pueden obtener algunos de los beneficios esperados de la selección genómica a través del uso de otros métodos (Hickey et al. 2017).

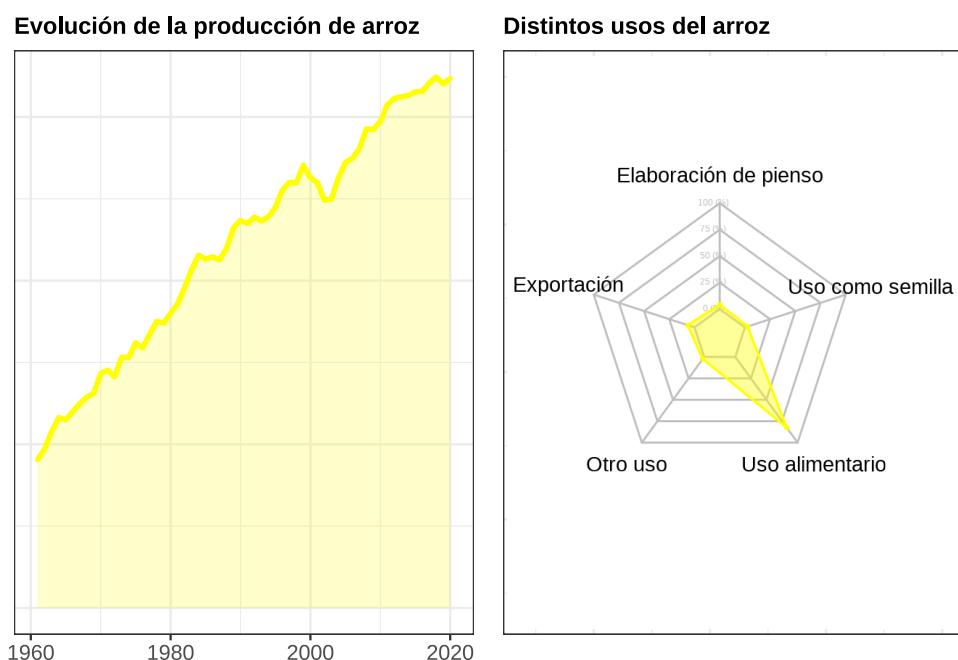
4- Es costoso invertir en infraestructura computacional y de registro tanto de datos genotípicos como fenotípicos requeridos para implementar la selección genómica. El tamaño de las poblaciones en la cría de animales son mucho más pequeños que la mayoría de poblaciones en la cría de plantas. Si bien 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 (Tabla1.1) 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. Por lo tanto, es de esperar que los mejoradores genéticos de animales y de plantas se beneficien del trabajo conjunto para solucionar aquellos problemas que les son comunes.

Tabla 1.1: .

Especie	Carácter	Habilidad predictiva	Modelo	Referencia

2.3.2. Relevancia de la mejora genética en arroz

**Figura 1.9:** 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 - FAOSTAT¹ para desarrollar estas figuras.

¹<http://faostat.fao.org/>

Capítulo 3

Objetivos

Capítulo 4

Materiales y métodos

4.1. Recurso vegetal y datos fenotípicos

Los conjuntos de datos se obtuvieron del Rice SNP-Seek Database¹, el cual es un cibersitio con información sobre datos de genotipado de SNP y de fenotipos de distintas variedades de arroz (*Oryza sativa L.*). En un estudio previo a este (Vourlaki et al., s. f.), los datos de genotipado de SNP fueron sometidos a procedimientos de control de calidad, en los que fueron eliminados loci de SNP con una frecuencia del alelo menor de menos de 0.01 y con una tasa de ausencia mayor a 0.01. Para este estudio se consideró eliminar loci de SNP con una frecuencia del alelo menor de menos de 0.05.

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.

Mediante un análisis de componentes principales realizado sobre los datos de genotipado de SNP (Figura 4.1) se observaron diferentes grupos varietales de arroz, de los cuales la variedad indicada fue seleccionada para llevar a cabo el estudio una vez la misma fue el grupo varietal con mayor número de individuos genotipados (451 individuos de un total de 738).

¹<https://snp-seek.irri.org/index.zul;jsessionid=DD991975FDC4F320BE3C33ED056D0363>

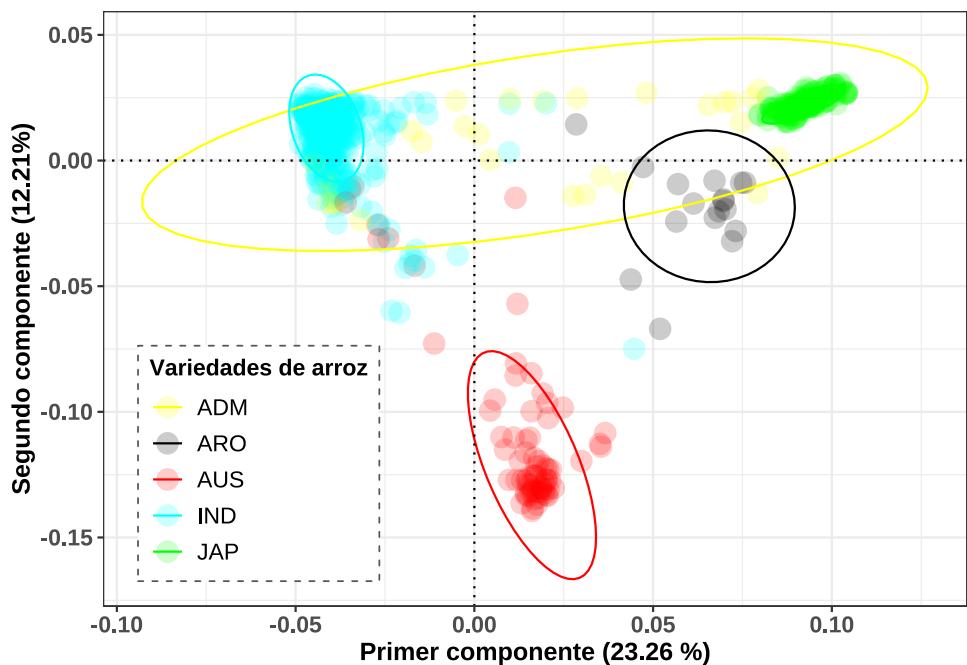


Figura 4.1: Análisis de componentes principales en datos de arroz. Los puntos y las circunferencias de color representan los distintos grupos varietales disponibles: tipo intermedio o mezclado (ADM), aromático (ARO), aus (AUS), indica (IND) y japónica (JAP).

En relación a los datos de fenotipo, el conjunto de datos proporcionó información sobre distintos caracteres fenotípicos de relevancia agronómica como son la trillabilidad de la panícula, el peso del grano, la fuerza del culmo, entre otros (Figura 4.2), siendo seleccionada para la predicción posterior utilizando los modelos de selección genómica el carácter tiempo de floración, ya que en este se observó suficiente variación fenotípica.

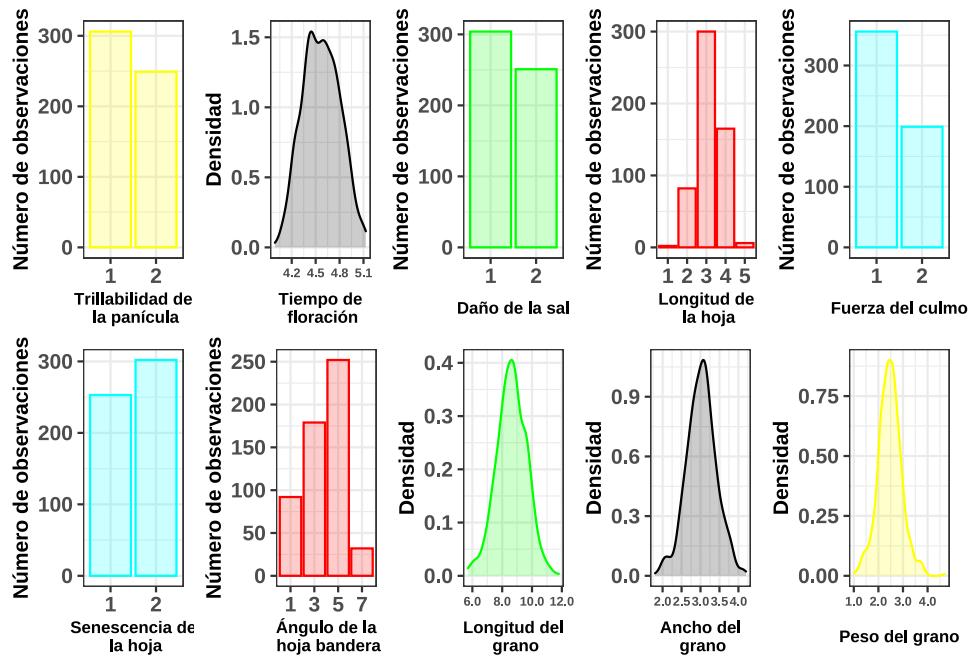


Figura 4.2: Distribución de cada uno de los caracteres del conjunto de datos fenotípicos de arroz.

4.2. Predicción basada en información de pedigree e información genómica

Para llevar a cabo la predicción usando el mejor predictor lineal insesgado (BLUP) en individuos no genotipados y el mejor predictor lineal insesgado genómico de un solo paso (ssGBLUP) tanto en individuos genotipados como no genotipados, se aplicó el siguiente modelo:

$$y = 1_n \mu + Zg + e, \quad (4.1)$$

donde y representa el valor del fenotipo a predecir (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, μ 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 representar la matriz identidad, y σ_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 de pedigree, y σ_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 descrita anteriormente y de la matriz de parentesco basado en marcadores de ADN (o 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 la matriz A, y los subíndices 1 y 2 representan los individuos genotipados y no genotipados, respectivamente. Inicialmente 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 metodologías 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.

4.3. Estimación de la heredabilidad

Se estimó la heredabilidad mediante el método BLUP con base en el modelo descrito anteriormente. La heredabilidad se calculó como:

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

donde σ_g^2 es la varianza genética aditiva y σ_e^2 es la varianza residual.

4.4. Generación de pedigríes ancestrales, subconjuntos de datos y habilidad predictiva

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 contar con esta información.

De acuerdo a Fernández y Toro (2006), la idea del software Molcoanc consiste en construir un pedigrí 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 con la matriz de parentesco molecular calculada a partir de los marcadores de ADN proporcionados.

En este sentido, el software Molcoanc se utilizó para construir tres pedigríes, los cuales se diferenciaron en el número de generaciones generadas por encima de la población fundadora (Figura 4.3). Este proceso de generación de pedigríes se realizó diez veces (diez replicas para cada pedigrí) con el fin de posteriormente usar cada replica para medir la variabilidad de la predicción mediante el BLUP y el ssGBLUP.

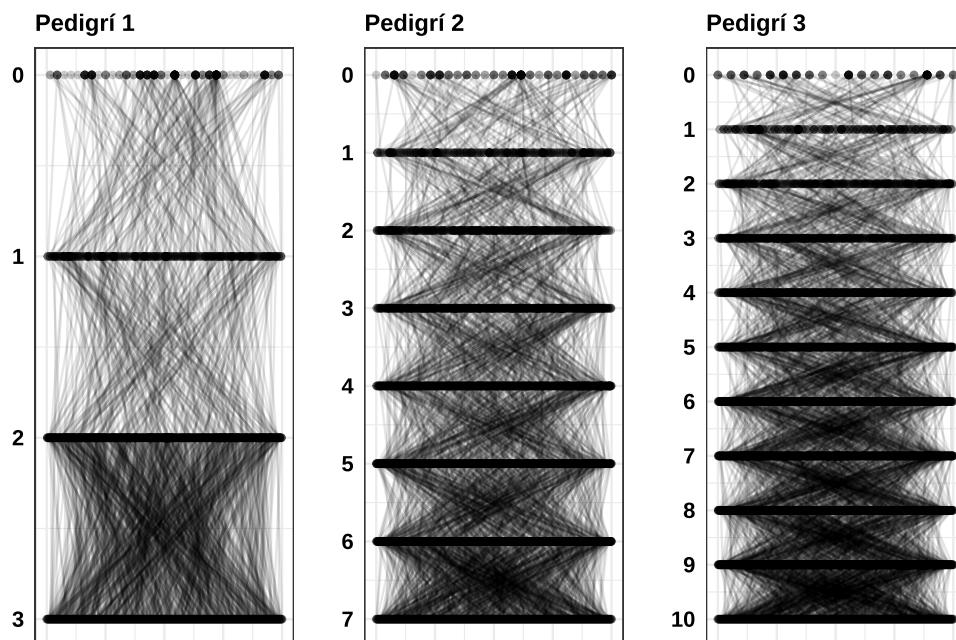


Figura 4.3: Ejemplo de pedigríes generados en la primera replica. En cada pedigrí se generaron distintas generaciones por encima de los individuos de la población fundadora. El número de individuos totales en cada pedigrí fueron: 751 individuos en el pedigrí 1, 1661 en el pedigrí 2 y 2451 en el pedigrí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

Para identificar 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 luego del control de calidad, se seleccionaron de forma aleatoria mediante el uso de Plink (Purcell et al. 2007) tres subconjuntos de datos de SNP, de manera que el número de marcadores aproximado fuera igual a 1000, 10000 y 100000 SNP.

Luego, 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.

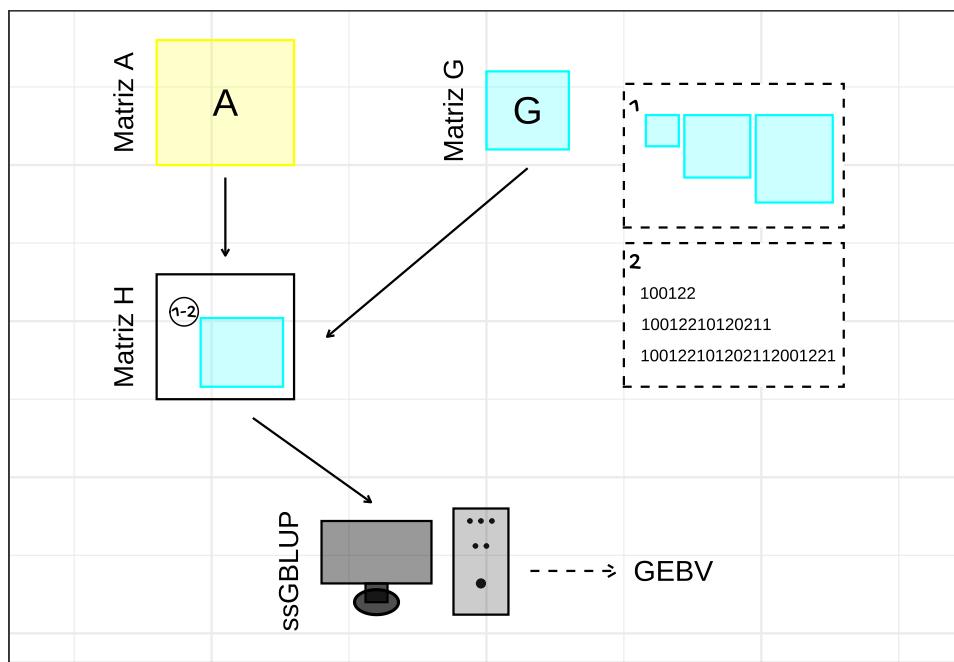


Figura 4.4: Esquema de construcción de la matriz H a partir de la matriz A y la matriz G, con base en diferentes subconjuntos de datos (o matrices G). El recuadro 1 representa tres matrices G con distinta dimensión dado el número de individuos genotipados y el recuadro 2 representa tres diferentes densidades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

Cabe aclarar que además de los casos descritos con anterioridad, también se consideraron los casos en los cuales 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 penalizados, respectivamente. Para la estimación de los componentes de varianza y 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 sentido, el valor fenotípico observado de 48 del total de 451 individuos de arroz de la variedad indica (que corresponde a los individuos clasificados como variedades mejoradas) se consideró como fenotipo faltante.

4.5. Simulación de fenotipos y genotipos, subconjuntos de datos y habilidad predictiva

Se evaluó el efecto sobre la predictibilidad del número de datos de genotipado de SNP y del número de individuos genotipados realizando simulación por ordenador, usando el paquete **SeqBreed** (Pérez-Enciso, Ramírez-Ayala, y Zingaretti 2020) del lenguaje de programación **Python** (Van-Rossum y Drake 1995). Los pasos llevados a cabo en la 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 uso el conjunto de datos con los 100231 SNP resultantes del control de calidad. Dicho conjunto de datos se paso 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 (100231).
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 GWAS e

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 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 ya reportados en la literatura científica.

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). Luego usando la función `gg.Population()`, se generó mediante simulación los genotipos y fenotipos 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 (10x80) ²	800 (20x40)	800 (40x20)	800 (80x10)
F3	800 (800x1) ³	800 (800x1)	800 (800x1)	800 (800x1)
Total	2061	2071	2091	2131

¹10 indica el número de descendientes que tendrían los 451 individuos de la generación F0 mediante cruzamiento.

²10x80 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.

³800x1 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 el apartado anterior, con número de marcadores aproximado de 1000, 10000 y 100000 SNP.

5. Implementación de la selección: se uso la función `sel.doEbv()` para llevar a cabo la predicción, usando los modelos BLUP y ssGBLUP descritos anteriormente. En caso del GBLUP, el cual requiere de datos de marcadores de ADN, se uso previamente la función `gg.do_X()` para generar la matriz G. Cabe aclarar que `SeqBreed` genero 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.

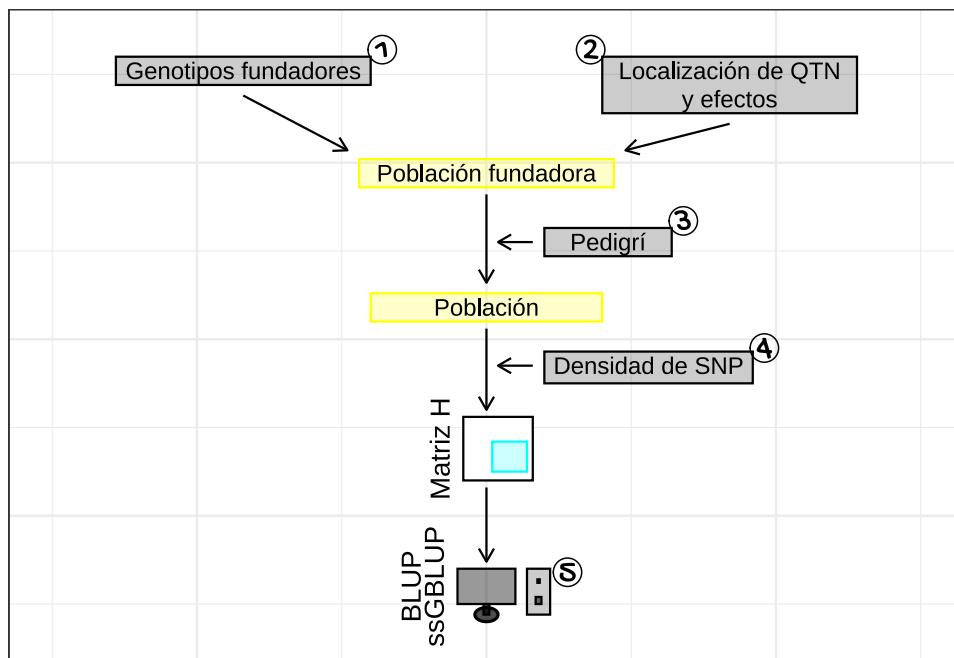


Figura 4.5: Esquema de predicción usando simulación con el paquete `SeqBreed`. Figura adaptada de Pérez-Enciso, Ramírez-Ayala, y Zingaretti (2020).

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 la generación F_1 y F_2 genotipados y todos los individuos de las distintas generaciones genotipadas.

Por último y al igual que en el apartado anteriormente descrito, 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.

Capítulo 5

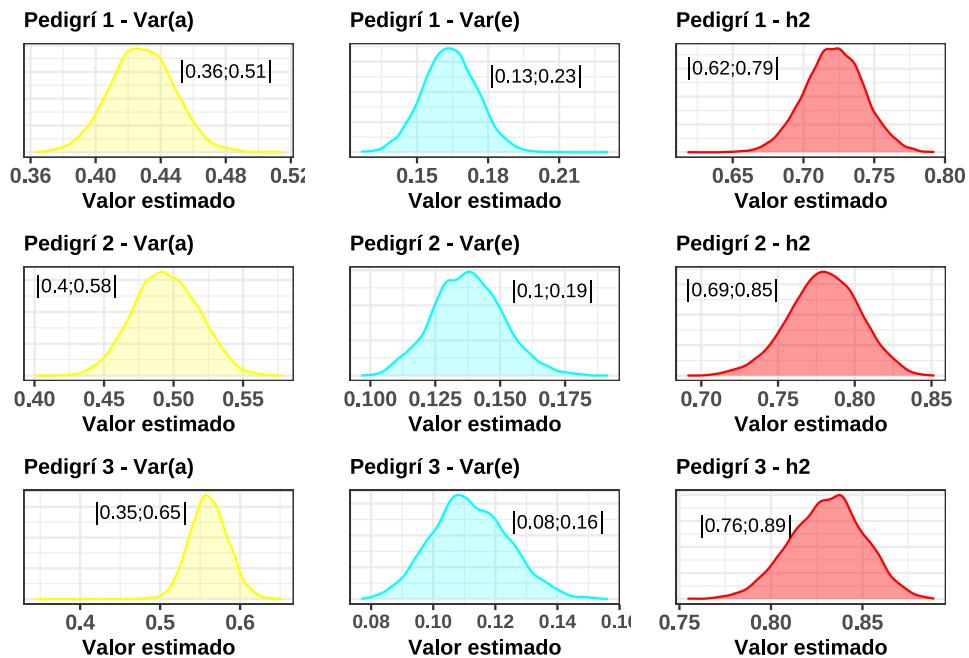
Resultados y discusión

5.1. Fenotipo y heredabilidad

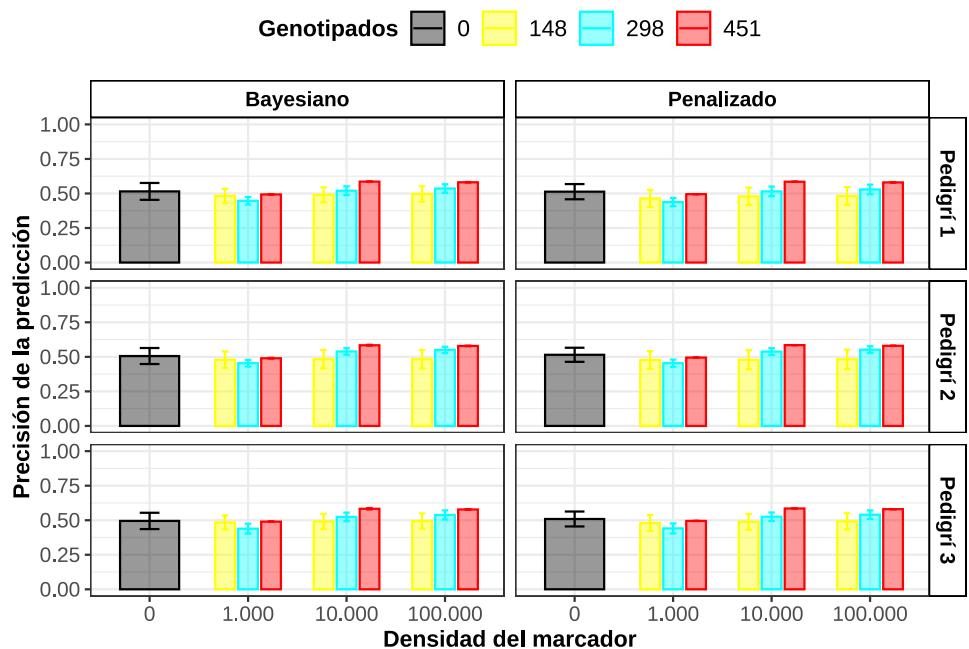
Tabla 2.1: Estimaciones de heredabilidad para el carácter tiempo de floración estimado por BLUP basado en el pedigrí.

Parámetros	Bayesiano			Penalizado		
	Ped. 1 ¹	Ped. 2	Ped. 3	Ped. 1	Ped. 2	Ped. 3
Varianza aditiva	0.43	0.50	0.56	NA	NA	NA
Varianza ambiental	0.16	0.14	0.11	NA	NA	NA
Heredabilidad	0.72	0.78	0.83	NA	NA	NA

¹Ped. 1 indica Pedigrí 1

**Figura 2.5:** .

Los resultados del análisis de máxima verosimilitud restringida (REML) y... (RKHS) bajo el modelo aditivo se observan en la Figura 2.6.

**Figura 2.6:** .

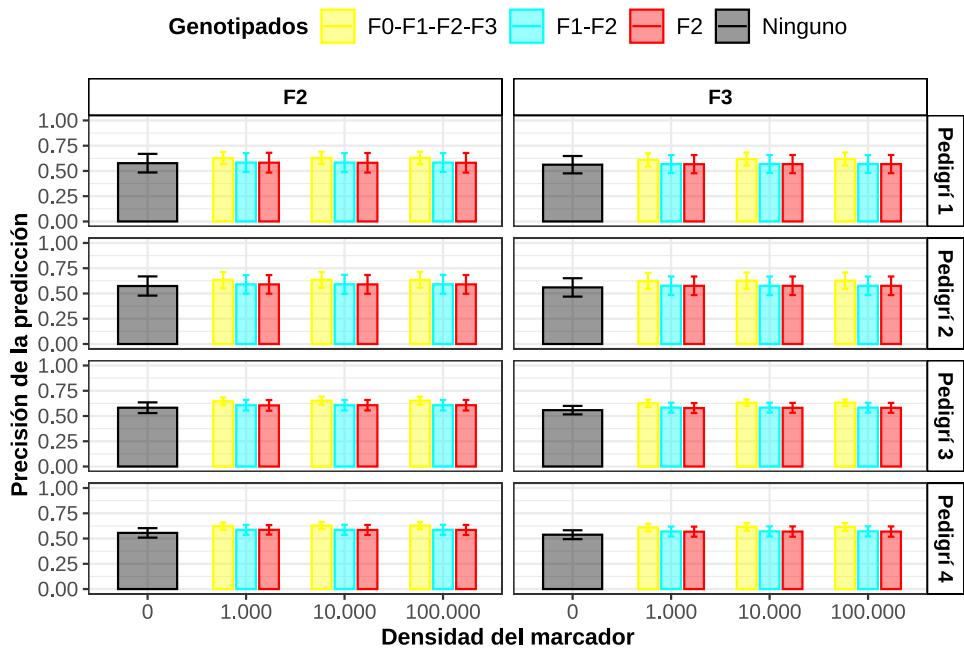


Figura 2.7: .

Capítulo 6

Conclusiones

Apéndice A

Anexos

A.1. Función¹ 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
```

¹http://rstudio-pubs-static.s3.amazonaws.com/378595_edda8cfe948a4786ae6dd74962cf6e94.html

```

# de relaciones adi-
# tivas A.

# 2. De lo anterior (Matrix_A) se extraen las partes correspon-
# dientes 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 pro-
# cede 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)

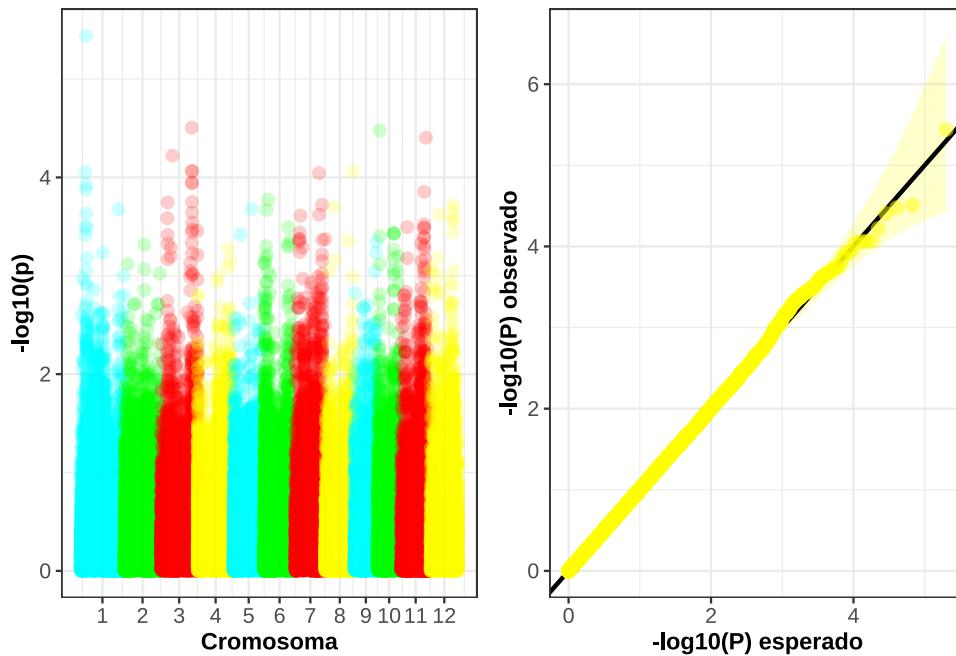
mH <- mH[order(as.numeric(rownames(mH))),
          order(as.numeric(colnames(mH)))]
mH <- Matrix(mH)

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

```

```
    return(mH)
}
```

A.2. Visualización del GWAS



A.3. Habilidad predictiva

Pedirí 1
751 individuos en total

	Bayesiano		Penalizado		
	Genotipados	Media	Desvío	Media	Desvío
Densidad 0					
0		0.515	0.061	0.513	0.055
Densidad 1.000					
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
Densidad 10.000					

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
<hr/>				
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
<hr/>				
Pedirí 2				
1661 individuos en total				
<hr/>				
Genotipados		Bayesiano	Penalizado	
		Media	Desvío	Media
				Desvío
<hr/>				
Densidad 0				
0	0.506	0.058	0.515	0.051
<hr/>				
Densidad 1.000				
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
<hr/>				
Densidad 10.000				
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
<hr/>				
Densidad 100.000				
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
<hr/>				

Pedirí 3				
2451 individuos en total				
<hr/>				
Genotipados		Bayesiano	Penalizado	
		Media	Desvío	Media
				Desvío
<hr/>				
Densidad 0				
0	0.495	0.059	0.509	0.054
<hr/>				
Densidad 1.000				

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. Habilidad predictiva

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	0.627	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 10.000				
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
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	0.632	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 10.000				
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
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	0.646	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 10.000				
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
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
Densidad 0				
Ninguno	0.556	0.047	0.538	0.044
Densidad 1.000				
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
Densidad 10.000				
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
Densidad 100.000				
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.
- Blasco, A. 2021. *Mejora genética animal*. 1st edition. EDITORIAL SÍNTESIS, S. A.
- Blasco, A., y M. A. Toro. 2014. «A short critical history of the application of genomics to animal breeding». *Livestock Science* 166: 4-9.
- 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>.
- 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>.
- 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.
- Fernández, J., y M. 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).
- Grinberg, N. F., O. I. Orhobor, y R. D. King. 2020. «An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat». *Machine Learning* 109: 251-77. <https://doi.org/10.1007/s10994-019-05848-5>.
- Henderson, C. R. 1975. «Best linear unbiased estimation and prediction under a selection model». *Biometrics* 31: 423-47.
- 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-setp genomic prediction of Eucalyptus dunni using different identity-by-descent and identity-by-state relationship matrices». *Heredity* 127: 176-89.
- 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/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>.
- 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/>.
- 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/>.
- Qanbari, S. 2020. «On the extent of linkage disequilibrium in the genome of farm animals». *Frontiers in Genetics* 10. <https://doi.org/10.3389/fgene.2019.01304>.
- R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- 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>.
- 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>.
- 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

