# Evaluación de la predicción genómica de un solo paso en plantas

Esta tesis se escribio usando los paquetes de R (R) Markdown, LaTeX , `bookdown` y `amsterdown`.

# Evaluación de la predicción genómica de un solo paso en plantas

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 publico 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. Revisión de literatura** **1**
1.1. Principios de mejora genética en plantas y animales . . . . . 1
1.2. Breve historia hacia la selección genómica . . . . . 1
1.3. La selección genómica . . . . . 4

**2. Titulo** **11**
2.1. Introducción . . . . . 12
2.2. Métodos . . . . . 14
2.3. Resultados . . . . . 18
2.4. Discusión . . . . . 20

**A. Anexos del capitulo 2** **21**
A.1. Función para el calculo de la matriz de relación combinada . 21

**Bibliografía** **23**

**Agradecimientos** **26**

# Capítulo 1

# Revisión de literatura

## 1.1. Principios de mejora genética en plantas y animales

## 1.2. 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.

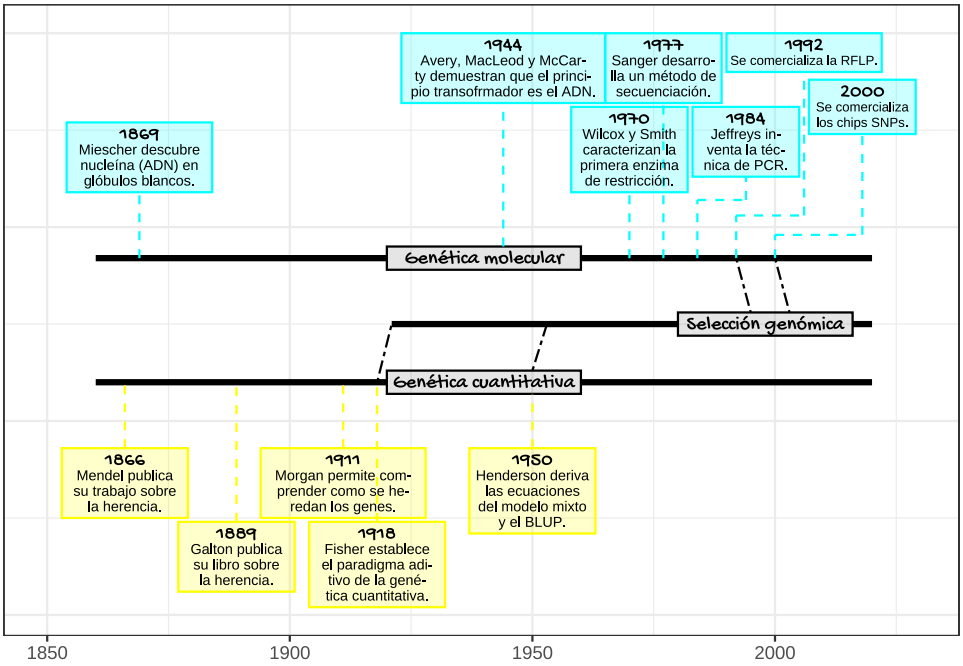


**Figura 1.1:** Cronología de las disciplinas de la genética molecular y la genética cuantitativa. Varios descubrimientos permitieron la conexión entre ambas disciplinas lo que permitió el desarrollo de la selección genómica. Figura adaptada de Nelson, Pettersson, y Carlborg (2012).

La genética cuantitativa se formo hace más de un siglo en ausencia directa de datos genéticamente observables (Nelson, Pettersson, y Carlborg 2012). Esta disciplina se formo 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 esta 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 pedigrí 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; delosCampos et al. 2013), comparado al método BLUP mencionado en párrafos anteriores. Según los mismos autores (delosCampos 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 (delosCampos 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).

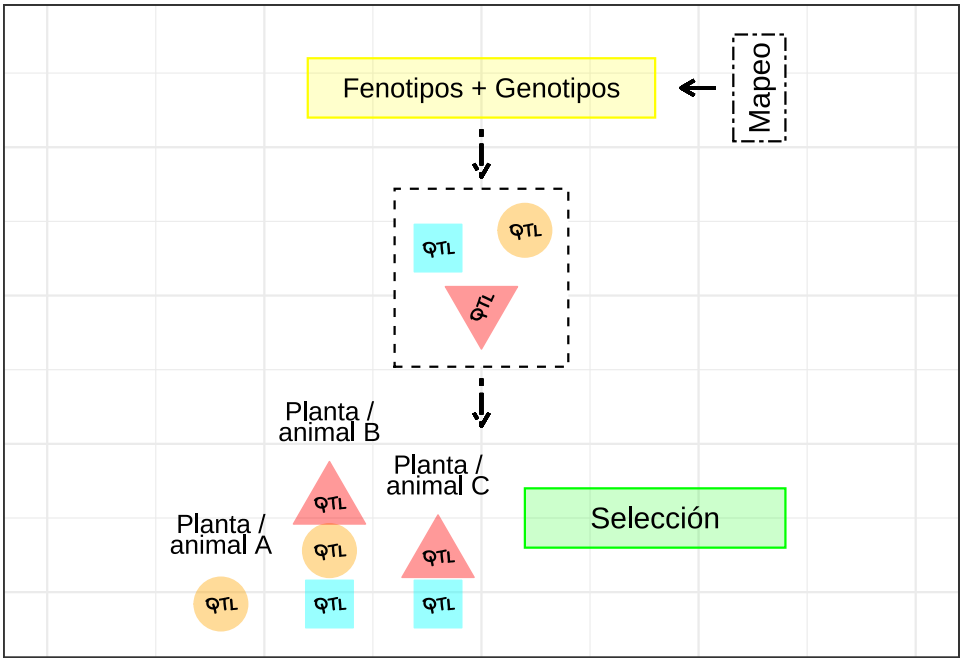



**Figura 1.2:** Esquema de la MAS. En la MAS, los fenotipos y genotipos de la población de mapeo se analizan usando un modelo estadístico, identificando con ello relaciones significativas entre fenotipos y genotipos. Por último, se seleccionan los individuos favorables con base en datos de

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.

## 1.3. 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 et al. (2020) 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).

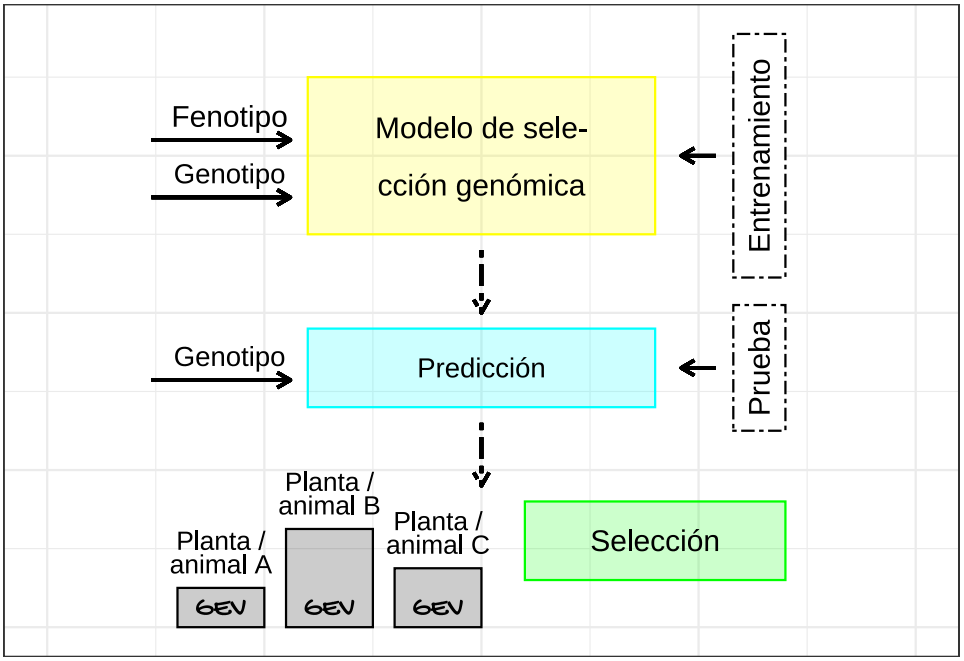



**Figura 1.3:** Esquema de la selección genómica. La selección genómica utiliza un modelo estadístico, diseñado a partir de datos genotípicos y fenotípicos en una población de entrenamiento, para predecir el GEBV de los individuos en una población de prueba con datos genotípicos. Por último, los individuos 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).

### 1.3.1. 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; delosCampos et al. 2013; Crossa et al. 2017).

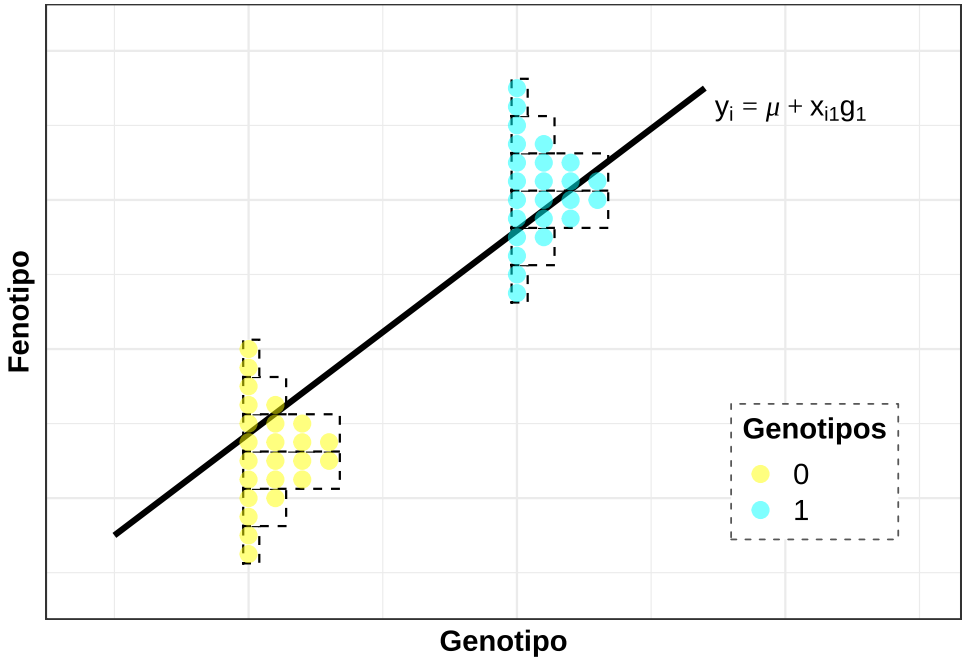



**Figura 1.4:** Relación genotipo-fenotipo de individuos (circulos amarillo y azul) para un solo marcador. $Y_i$ y $x_{i1}$ denotan los fenotipos y genotipos, y $\mu$ y $g_i$ son los parámetros a determinar. Los genotipos bialélicos se codifican como 0 y 1, y los fenotipos se distribuyen de acuerdo a una normal. Figura adaptada de Nakaya y Isobe (2012).

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

$$y_i = \mu + \sum_{j=1}^{p} x_{ij} g_j + e_i, \tag{1.1}$$

donde $i$ $(1, 2, 3, \ldots, n)$ representa a los individuos, $j$ $(1, 2, 3, \ldots, p)$ 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:

$$y = Zg + e, \tag{1.2}$$

donde $y$ es un vector de longitud igual al número de individuos $(1, 2, 3, \ldots, n)$ 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; delosCampos 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).

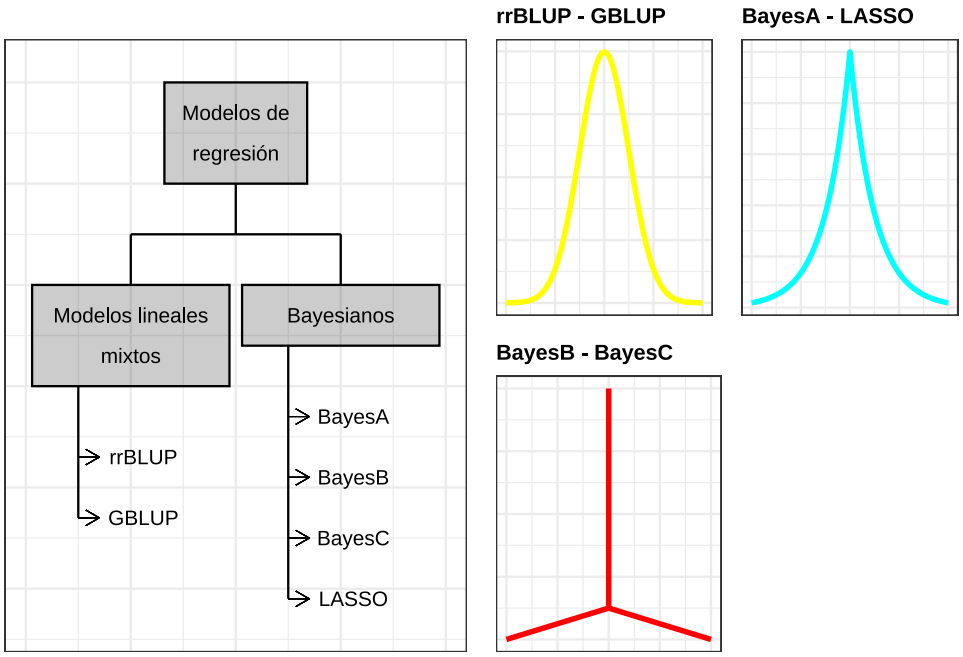



**Figura 1.5:** Enfoques estadísticos de la selección genómica. Cada uno de estos enfoques suponen distintas distribuciones de efectos de los marcadores sobre el carácter. Figura adaptada de delosCampos et al. (2013) y de Tong y Nikoloski (2021).

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:

$$y = Xb + Zg + e, \tag{1.3}$$

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 efector 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 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 relación basada en marcadores de ADN (denominada como matriz G) en lugar de la matriz de relación 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:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + G^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}, \tag{1.4}$$

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, 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).

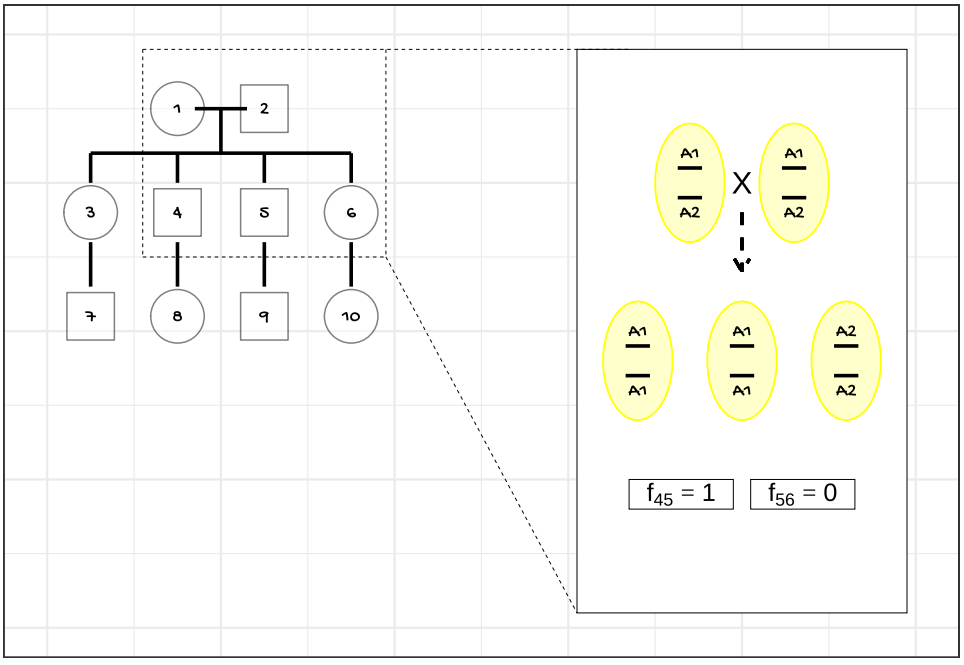



**Figura 1.6:** Parentesco observado entre individuos emparentados.

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 su GEBV, y calcular su precisión (Tan et al. 2017). En cuanto a la velocidad de calculo, el GBLUP es mucho más rápido que los método 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, delosCampos et al. (2013) proporcionan algunos ejemplos sobre que 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 esta en una región del genoma que alberga genes que pueden afectar una carácter de interés.

### 1.3.2. De un proceso de selección genómica de múltiples pasos a un solo paso

### 1.3.3. SUBTITULO SOBRE ML Y DL

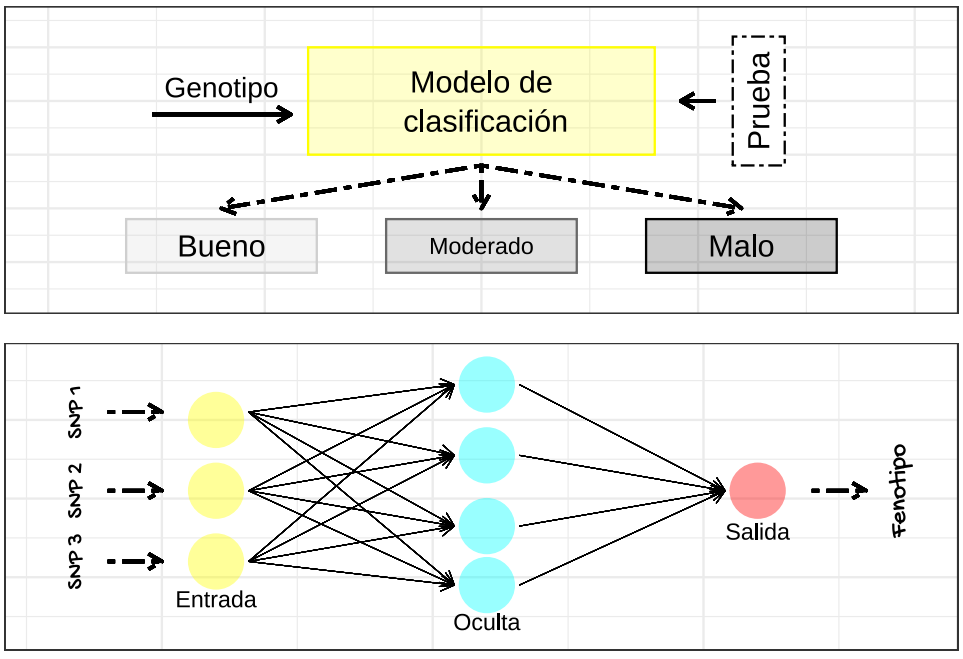



**Figura 1.7:** . Figura adaptada de Tong y Nikoloski (2021).

# Capítulo 2

# Titulo

**Resumen**
Insert abstract.

---

*Possibly insert citation here.*

## 2.1. Introducción

La teoría de la genética en el estudio de caracteres cuantitativos se estableció hace más de un siglo cuando Ronald Fisher presentó un documento (Fisher 1918) donde dio a conocer el desarrollo de la teoría del modelo infinitesimal, permitiendo con ello unificar dos de las escuelas de pensamiento que para ese entonces estaban en constante debate: la escuela de pensamiento Mendeliano, cuyo objetivo consistía en localizar y caracterizar factores de herencia, y la escuela de pensamiento biométrico, cuyo origen se remonta a Galton quien buscaba aplicar modelos biométricos con el fin de estudiar las relaciones entre parientes (Nelson, Pettersson, y Carlborg 2012; Blasco y Toro 2014).

La teoría del modelo infinitesimal desarrollado por Fisher establece que la varianza genética de un carácter esta determinado por un gran número de factores Mendelianos, cada uno de los cuales tiene una pequeña contribución aditiva al fenotipo de dicho carácter (Nelson, Pettersson, y Carlborg 2012; Turelli 2017). Naturalmente, los modelos usados en estudios de mejoramiento genético han sido concebidos en base a esta teoría (Villemereuil et al. 2016; Pérez-Enciso 2017), siendo ejemplo de ello el mejor predictor lineal insesgado (BLUP) y el mejor predictor lineal insesgado genómico (GBLUP).

En las ciencias animales, el valor de cría estimado (EBV) se suele predecir en función de un conjunto de modelos que relacionan el fenotipo de una población con la información del pedigrí, mediante el uso del BLUP. No obstante, este método no es factible para poblaciones sin información de pedigrí o con una estructura poblacional compleja, como suele ser el caso de las plantas (Nakaya y Isobe 2012; Tong y Nikoloski 2021). Para el año 2001, Meuwissen, Hayes y Goddard propusieron un método innovador para predecir los valores de cría basado en marcadores de ADN (GEBV), denominándose tiempo después como selección genómica (Nakaya y Isobe 2012; Blasco y Toro 2014), el cual permitió también superar las limitaciones que suponía el uso del BLUP para predecir los valores de cría en plantas.

Hoy en día, la selección genómica se considera como un método potencial para el mejoramiento genético en plantas (Nakaya y Isobe 2012), ya que sus ciclos reproductivos suelen ser prolongados, por lo cual con el uso de la selección genómica es posible acelerar dichos ciclos reproductivos con el beneficio adicional de mejorar la tasa de ganancia genética anual por unidad de tiempo y costo (Desta y Ortiz 2014; Jurcic et al. 2021). Además, los datos sobre marcadores de ADN en todo el genoma están cada vez más disponibles para cultivos de relevancia agronómica (Tong y Nikoloski 2021).

El GBLUP es uno de los métodos más comunes de selección genómica (Jurcic et al. 2021). De hecho, es el método más popular debido a su simplicidad al sustituir la matriz de relación de parentesco basado en pedigríes (Wright 1922) por una matriz de relación basada en marcadores de ADN (Hayes, Visscher, y Goddard 2009). Así mismo, el GBLUP predice con may-

or precisión los GEBV en comparación a los EBV del BLUP, debido a que con el primero se estima mejor las relaciones entre individuos (Misztal, Aggrrey, y Muir 2012), por lo cual la matriz de las relaciones genómicas suele verse como un estimador mejorado de las relaciones basadas en marcadores en lugar de pedigríes (Legarra et al. 2014).

En términos generales, la selección genómica es un proceso de tres pasos en el que los individuos, sobre la base de su información fenotípica y de pedigrí, son evaluados inicialmente mediante una evaluación genética tradicional por medio del BLUP, y posteriormente a partir de los fenotipos corregidos o pseudo-fenotipos resultantes de esta evaluación genética inicial, es llevado a cabo un análisis genómico de los individuos genotipados mediante el GBLUP. Por último y en base a la información generada, se calculan los GEBV por medio de un índice de selección (Legarra, Aguilar, y Misztal 2009; Misztal, Legarra, y Aguilar 2009; Misztal, Aggrrey, y Muir 2012; Legarra et al. 2014; Misztal, Lourenco, y Legarra 2020).

Como no todos los individuos pueden genotiparse, la selección genómica se lleva a cabo a partir del proceso anterior de tres pasos (Legarra, Aguilar, y Misztal 2009). Sin embargo, este proceso es tendente a cometer errores (Misztal, Aggrrey, y Muir 2012), además de presentar inconvenientes como son la perdida de información y la difícultad 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 sus valores de cría (Imai et al. 2019; Jurcic et al. 2021), método el cual se denomino como mejor predictor lineal insesgado genómico de un solo paso (ssGBLUP).

En el ssGBLUP se dispone de una matriz de parentesco genómica global de individuos genotipados y no genotipados, denominada como matriz de relación combinada o matriz H. Esta matriz se obtiene combinando información de la relación genómica entre individuos genotipados, e información de pedigrí entre individuos genotipados y no genotipados (Imai et al. 2019). Con ello, el proceso anterior de tres 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 calculo posterior de fenotipos corregidos y la construcción del índice de selección mencionado previamente (Misztal, Lourenco, y Legarra 2020).

Al ser una forma de BLUP en el que la matriz de relación de parentesco es sustituida por la matriz de relación combinada (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 informaticas ya desarrolladas en base al BLUP (Lourenco et al. 2020). Este hecho hace del ssGBLUP un método de uso rutinario para la evaluación genómica en animales, donde ha demostrado que

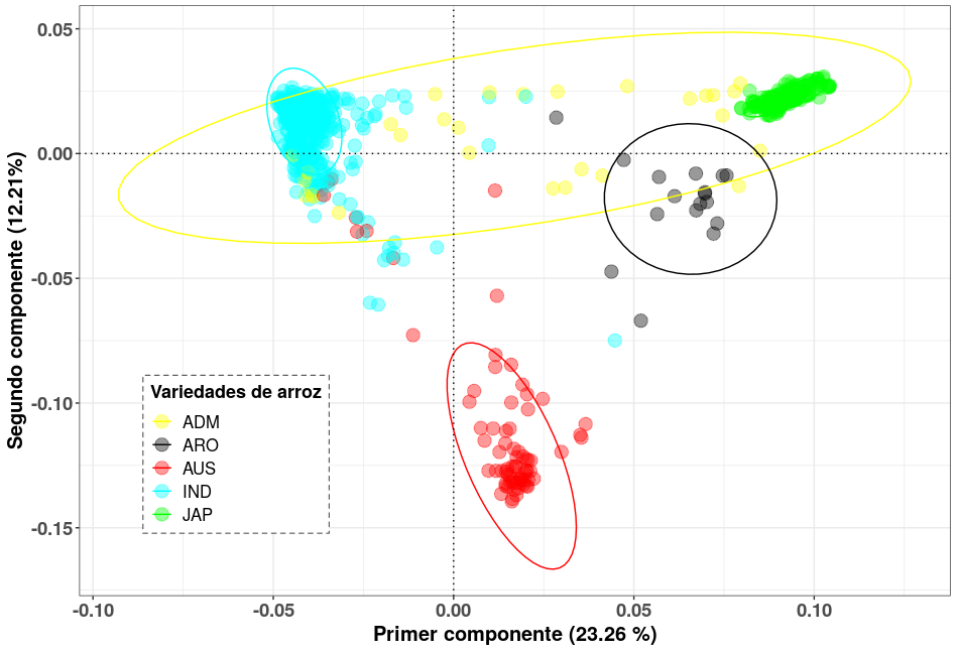
produce una predicción más precisa de los valores de cría en comparación a los métodos BLUP y GBLUP antes mencionados (Misztal, Aggrrey, y Muir 2012; Pérez-Rodríguez et al. 2017; Misztal, Lourenco, y Legarra 2020). No obstante, el uso del ssGBLUP para la selección genómica en plantas es más reciente y escaso (Pérez-Rodríguez et al. 2017; Jurcic et al. 2021). En consecuencia, el **objetivo**

## 2.2. Métodos

### 2.2.1. Recurso vegetal y datos fenotípicos

Los conjuntos de datos se obtuvieron del Rice SNP-Seek Database[1], el cual es un cibersitio con información sobre datos de genotipado de SNP y de fenotipos de distintas variedades de arroz (*Oryza sativa L.*). Posteriormente, dichos conjuntos de datos fueron usados por Vourlaki et al. (s. f.), quienes sometieron los datos de genotipado de SNP 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.

Mediante un análisis de componentes principales realizado sobre los datos de genotipado de SNP (Figura 2.1) se observaron diferentes grupos varietales de arroz, de los cuales la variedad indica fue seleccionada para llevar a cabo este estudio una vez la misma era 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 2.1:** Análisis de componentes principales en datos de arroz. Los puntos y las circuferencias de color representan distintos grupos varietales: 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 proporciono 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 2.2), siendo seleccionada para este estudio el carácter tiempo de floración ya que en este se obervo suficiente variación fenotípica.

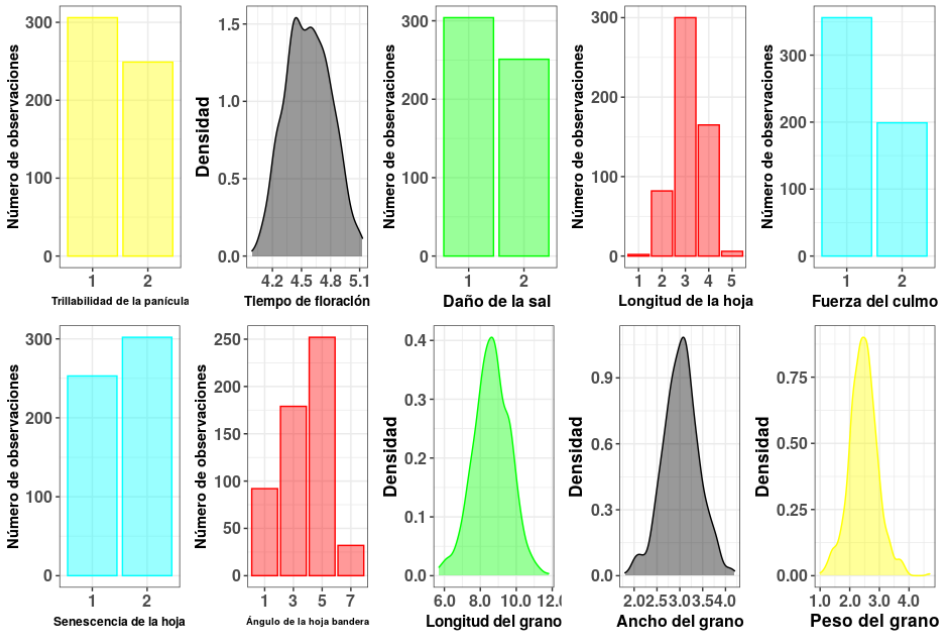



**Figura 2.2:** Distribución de cada uno de los caracteres del conjunto de datos fenotípicos de arroz.

En lo que respecta a la información de pedigrí, esta no estaba disponible. Por ello, se utilizó la metodología implementada en el software MOLCOANC (Fernández y Toro 2006) con el fin de contar con esta información. Este software . (Figura 2.3).

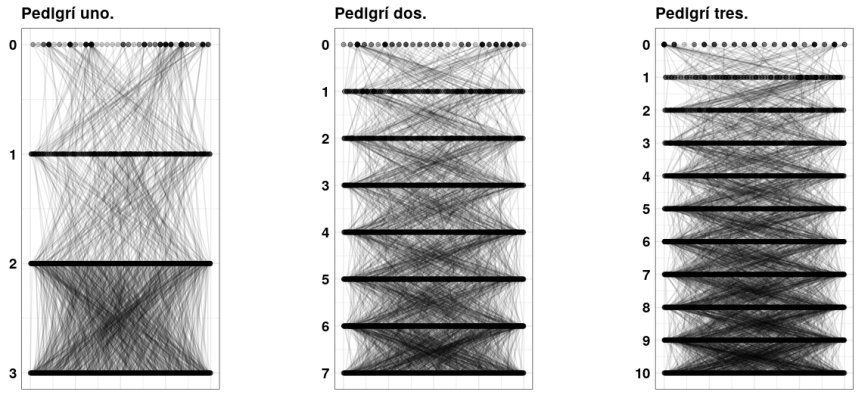



**Figura 2.3:** .

### 2.2.2. Modelo para la predicción genómica y habilidad predictiva

Para llevar a cabo la predicción genómica mediante el mejor predictor lineal insesgado genómico de un solo paso (ssGBLUP), se eliminaron los loci de SNP con una frecuencia del alelo menor de menos de 0.05. La predicción genómica se realizó mediante el siguiente modelo con los datos descritos anteriormente:

$$y = Za + e, \tag{2.1}$$

donde $y$ representa el valor del fenotipo a predecir (tiempo de floración) y $Z$ es la matriz de incidencia que relaciona $a$ con $y$. El vector $a$ representa los valores genotípicos como se describen en el siguiente parrafo, y $e$ es el vector de residuos con una distribución que se asume normal con media igual a 0 y matriz de covarianza $I\sigma_e^2$.

En la ecuación (1), $a$

Para identificar el efecto sobre la predictibilidad del tamaño de la muestra de entrenamiento, el número de datos de genotipado de SNP y el número de individuos genotipados, se usaron diferentes subconjuntos de datos (Figura 2.4) con la siguientes características:

1. Diferente información de pedigrí:
2. Diferentes densidades de SNP:
3. Distinta cantidad de individuos genotipados:

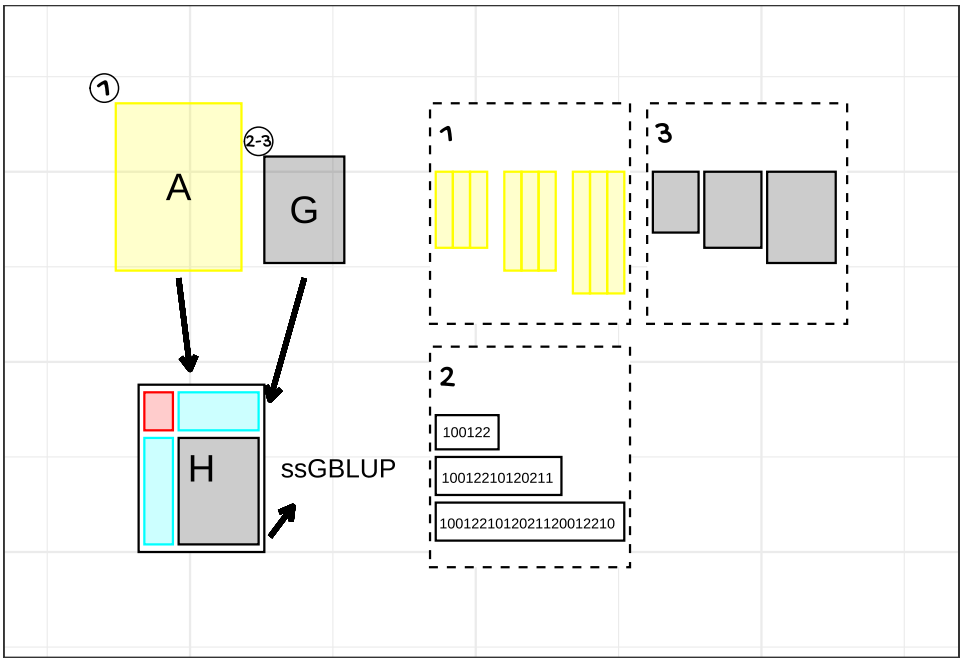
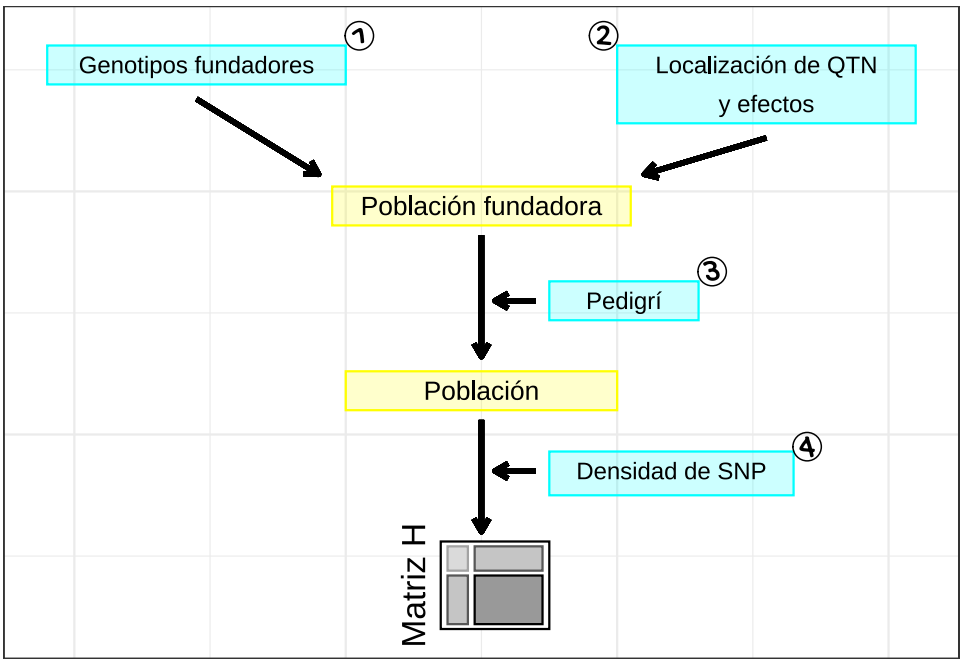



**Figura 2.4:** Esquema del calculo de la matriz H a partir de las matrices A y G, con base en diferentes subconjuntos de datos. El recuadro 1 representa los tres pedigríes con diferentes número de individuos y que posteriormente se usaron para el calculo de la matriz A. El recuadro 2 representa diferentes densidades de SNP. El recuadro 3 representa matrices G con distinta dimensión dado el número de individuos genotipados.

Se uso el coeficiente de correlación entre los valores fenotípicos observados y predichos como medida de la predictibilidad. De acuerdo a Xua, Zhub, y Zhang (2014), la predictibilidad debe obtenerse usando una muestra de validación independiente o mediante validación cruzada 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 la variedad indica (que corresponde a los individuos clasificados como variedades mejoradas) se considero como faltante.

### 2.2.3. Estudio de simulación




## 2.3. Resultados

### 2.3.1. Fenotipo y heredabilidad

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

| | reml | | |
|---|---|---|---|
| Parámetros | Pedigrí 1 | Pedigrí 2 | Pedigrí 3 |
| Varianza aditiva | 0.49 | 0.46 | 0.57 |
| Varianza ambiental | 0.11 | 0.17 | 0.13 |
| Heredabilidad | 0.82 | 0.73 | 0.81 |

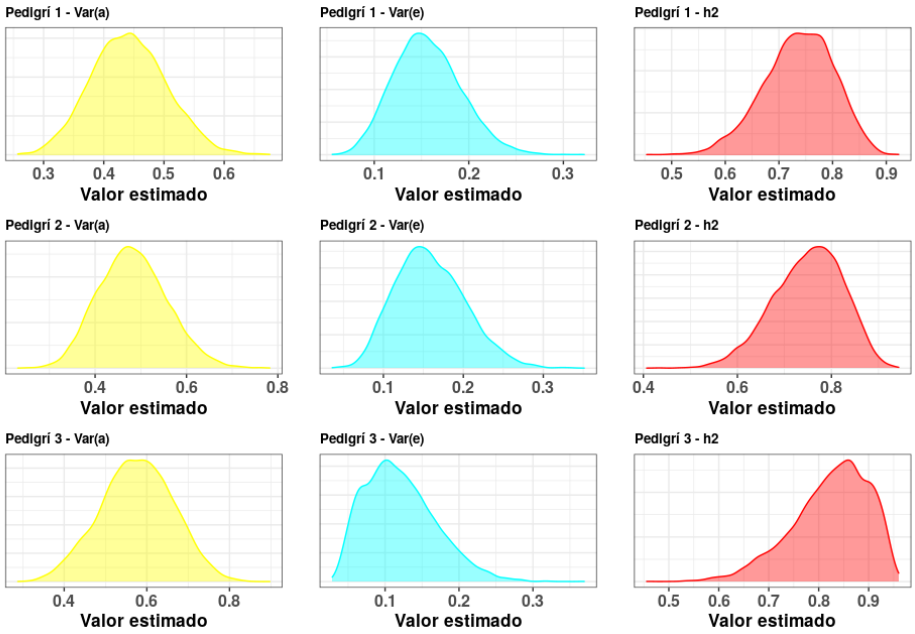
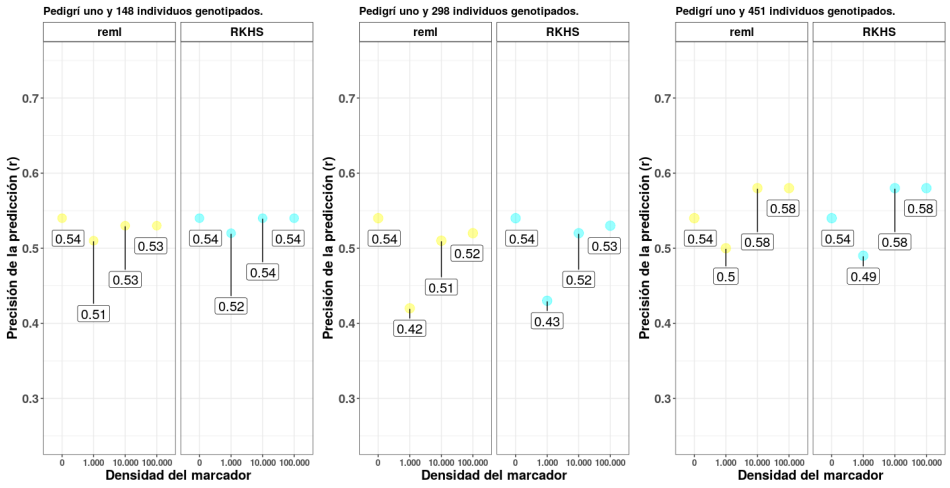



**Figura 2.5:** .

Los resultados del análisis de máxima verosimilitud restringida (REML) y... (RKHS) bajo el modelo aditivo se observan en la Figura 2.5.



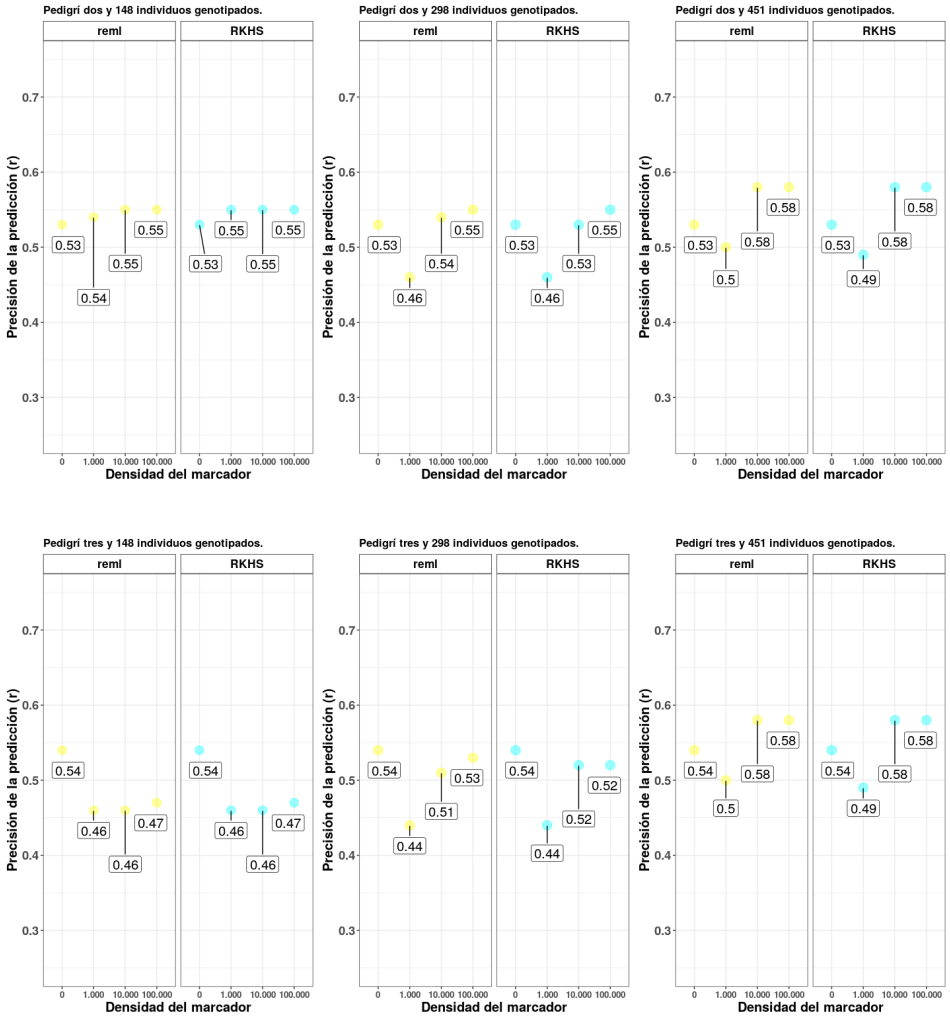



**Figura 2.6:** .

## 2.4. Discusión

# Apéndice A

# Anexos del capitulo 2

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

```
fn.mH <- function(ped, mG) { # Esta función recibe como argu-
                             # mentos los datos con estructu-
                             # ra (id | sire | dam | Gen (TRUE
                             # /FALSE)) y la matriz de relacio-
                             # nes genómicas.

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

  ped_edit <- pedigreemm::editPed( # Esta función ordena el pe-
                                   # digrí.
    sire = ped$sire,
    dam = ped$dam,
    label = ped$id
    )
  pedi <- pedigreemm::pedigree( # Aquí se usa la salida anterior
                                # (ya ordenado) y se crea un ob-
                                # jeto de clase pedigree.
    sire = ped_edit$sire,
    dam = ped_edit$dam,
    label = ped_edit$label
    )
  Matrix_A <- pedigreemm::getA(ped = pedi) # Esto dara la matriz
                                           # de relaciones adi-
                                           # tivas A.
```

```
# 2. De lo anterior (Matriz_A) se extraen las partes correspon-
# dientes a individuos no genotipados (1) y genotipados (2)

# Individuos no genotipados:
A_11 <- Matrix_A[ped$Genotiped != 1, ped$Genotiped != 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 y a calcular su inversa

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)
}
```

# Bibliografía


Ahmadi, N., J. Bartholomé, T. V. Cao, y C. Grenier. 2020. *Quantitative genetics, genomics and plant breeding*. 2nd edition. https://doi.org/10.1079/9781789240214.0243.

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.

Crossa, J., P. Pérez-Rodríguez, J. Cuevas, O. Montesinos-López, D. Jarquín, G. delosCampos, 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.

delosCampos, 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.

Fisher, R. A. 1918. «The correlaction between relatives under the supposition of Mendelian inheritance». *Transactions of the Royal Society of Edinburgh* 52: 399-433.

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.

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.

Hayes, B. J., P. M. Visscher, y M. E. Goddard. 2009. «Increased accuracy of artificial selection by using the realized relationship matrix». *Genetics Research* 91: 47-60.

Henderson, C. R. 1975. «Best linear unbiased estimation and prediction

under a selection model». *Biometrics* 31: 423-47.

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. «Genomic selection: genome-wide prediction in plant improvement». *Trends in Plant Science* 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/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.

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. Carlborg. 2012. «A century after

Fisher: time for a new paradigm in quantitative genetics». *Trends in Genetics* 29 (9): 669-76.

Pérez-Enciso, M. 2017. «Animal breeding learning from machine learning». *Journal of Animal Breeding and Genetics* 134: 85-86.

Pérez-Rodríguez, P., J. Crossa, J. Rutkoski, J. Poland, R. Singh, A. Legarra, E. Autrique, J. Burgueño G. delosCampos, 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.

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.

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.

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.

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.

Villemereuil, P. de, H. Schielzeth, S. Nakagawa, y M. Morrissey. 2016. «General methods for evolutionary quantitative genetic inference from generalized mixed models». *Genetics* 204: 1281-94.

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*.

Wright, S. 1922. «Coefficients of inbreeding and relationship». *The American Naturalist* 56: 330-38.

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.

# Agradecimientos