Tesis para optar al grado de
Magíster en Estadística Aplicada

Universidad Nacional de Córdoba

Modelos de Selección Genómica para Caracteres Cuantitativos basados en Marcadores Moleculares aplicados al mejoramiento de maíz

Lic. María Valeria Paccapelo
- 2015 -

Modelos de Selección Genómica para Caracteres Cuantitativos basados en Marcadores Moleculares aplicados al mejoramiento de maíz por María Valeria Paccapelo se distribuye bajo una Licencia Creative Commons Atribución-NoComercial 4.0 Internacional.

COMISIÓN ASESORA DE TESIS

Director
Dr. Julio Alejandro Di Rienzo

Co-Director
Ph. D Shengqiang Zhong

Miembros
Dr. Fernando Casanoves
Dr. Mariano Augusto Córdoba
Dr. Martín Oscar Grondona

Fecha Aprobación:
29 de Diciembre de 2015

AGRADECIMIENTOS

En este momento donde que alcancé el objetivo de finalizar mis estudios de maestría, me siento enormemente agradecida a cada una de las personas que me acompañó de alguna de las infinitas maneras que se puede ser parte del camino que recorrí desde que empecé este trayecto.

Gracias a mis padres, mis hermanos y abuelos por los valores que me transmitieton, por la confianza y el apoyo para que siga creciendo.

Gracias a mis amigos de cada rincón por ayudarme a disfrutar del camino que recorrí. En particular, gracias a esos amigos que me dio la maestría.

Gracias a Martín Grondona, porque desde múltiples roles me enseñó a hacer mi trabajo lo mejor posible, confió en mi y me dio oportunidades para desarrollarme profesionalmente.

Gracias a Guillermo Van Becelaere por la gestión para que pueda trabajar el tema que quería, pero sobre todo, por creer en mi y motivarme con “maestría, autonomía y propósito”.

Gracias a Julio Di Rienzo por inspirarme siendo su estudiante y por dirigirme con tanta dedicación, por darme libertad y trabajar con tanto profesionalismo y calidez.

Gracias a mis directores y miembros del jurado por la guía y las sugerencias constructivas que me permitieron seguir aprendiendo.

Finalmente, quiero agradecer a Advanta Semillas que apostó a mi desarrollo profesional, brindandome el tiempo y los recursos para estudiar; a Monsanto que continuó dando soporte a mi desarrollo, facilitando los materiales para poder realizar esta investigación y apoyandome a que siga creciendo; al cuerpo docente y personal administrativo de la maestría que hicieron que estudiar sea enriquecedor no sólo desde el punto profesional sino también personal.

RESUMEN

En la actualidad, los modelos de selección genómica (SG) han cobrado gran importancia ya que permiten predecir los valores genéticos de los individuos en función de marcadores moleculares (MM). La incorporación de numerosos MM en modelos de regresión conduce a problemas de dimensionalidad y multicolinealidad. Esta tesis tuvo como objetivo evaluar seis métodos de SG que confrontan estas dificultades (selección de variables, estimación penalizada y la combinación de ambos) desde enfoques clásicos o bayesianos y evaluar su habilidad predictiva para tres caracteres fenotípicos observados en 20 poblaciones de maíz (*Zea mays* L.). Los resultados indican que la habilidad predictiva se vio asociada a la heredabilidad del carácter y fue superior para los métodos penalizados, entre los que se recomienda la Regresión de Ridge vía modelos mixtos (RR-BLUP). Este trabajo permitió analizar diferentes técnicas estadísticas aplicadas a la SG en un contexto propio de un programa de mejoramiento genético de maíz.

PALABRAS CLAVE

Valores genéticos, métodos de estimación penalizada, métodos de estimación bayesiana, regresión de Ridge, método LASSO.

ABREVIACIONES

BLR: método de regresión LASSO Bayesiano, del inglés *Bayesian LASSO Regression*

BLUP*:* mejor predictor lineal insesgado (del inglés, *Best Linear Unbiased Predictor*)

BRR: método de Regresión de Ridge con enfoque Bayesiano, del inglés *Bayesian Ridge Regression*

CMEP: Cuadrado Medio del Error de Predicción

CV: Coeficiente de Variación

GH: Grupo Heterótico

LASSO: del inglés *Least Absolute Shrinkage and Selection Operator*

LR: método de Regresión LASSO con enfoque clásico

MM: Marcadores Moleculares

QTL: del inglés *Quantitative Trait Loci*

REML: Máxima Verosimilitud Restringida, del inglés *REstricted Maximum Likelihood*

RR: método de Regresión de Ridge con enfoque clásico

RR-BLUP: método de Regresión de Ridge con enfoque de modelos mixtos

SMC: método de Selección de variables y ajuste por Mínimos Cuadrados

SNP: polimorfismo de nucleótidos simples, del inglés *Single Nucleotide Polymorphism*

INDICE

1. INTRODUCCIÓN

El **mejoramiento genético** vegetal o animal es la ciencia, el arte y el negocio de mejorar los organismos para el beneficio de los seres humanos (Bernardo, 2002). Como una ciencia, el mejoramiento se sustenta sobre el conocimiento teórico y empírico de la genética. Como un arte, requiere juicios subjetivos en el diseño y la implementación de un programa de mejoramiento. Finalmente, como un negocio, necesita de inversiones de tiempo y dinero en distintos recursos tales como: técnicos, equipamiento y materiales.

La importancia relativa del arte y la ciencia en el mejoramiento ha cambiado a lo largo del tiempo. En los comienzos del mejoramiento, la habilidad de una persona para identificar visualmente los individuos más deseados era la única herramienta disponible. Así, la apariencia de un individuo o de un grupo de individuos, denominada **fenotipo**, determinaba si el individuo resultaba elegido (Fehr, 1987). Algunos ejemplos de caracteres fenotípicos son: la altura y el peso de un individuo, el rendimiento en grano de un cultivo, la resistencia a enfermedades, entre otros. A pesar de que la apariencia visual continúa siendo parte del mejoramiento, actualmente no es la única fuente de información y es posible planear un programa de mejoramiento con información basada en la configuración genética, o **genotipo**, de un individuo.

La expresión del fenotipo de un individuo depende de dos clases de factores: ambientales y genéticos. Mientras que el fenotipo hace referencia a la apariencia o medición de un carácter, el genotipo comprende los genes que controlan ese carácter y el ambiente, incluye todos los factores externos que pueden influir en la expresión de esos genes. Entre los factores ambientales se pueden mencionar: temperatura, fertilidad del suelo y el manejo del cultivo (Fehr, 1987).

Los caracteres fenotípicos pueden ser de naturaleza cualitativa o cuantitativa. Los cualitativos son variables fenotípicas que se registran como categorías o clases, por ejemplo: coloración del grano, color de la flor o resistencia a cierta enfermedad. En general, este tipo de carácter suele

estar controlado por uno o unos pocos genes y se los denomina **caracteres simples**. Por otro lado, los caracteres cuantitativos, se distinguen por ser variables que se miden en una escala numérica que eventualmente puede ser continua, por ejemplo, altura de una planta, rendimiento en grano, peso de un animal, entre otros. Muchos caracteres de interés agronómico son de naturaleza cuantitativa y se encuentran controlados por múltiples genes de efectos pequeños y por ello reciben el nombre de **caracteres poligénicos, cuantitativos o complejos** (Collard *et al.*, 2005; Buckler *et al.*, 2009).

Las regiones dentro del genoma que contienen genes asociados a un carácter cuantitativo se conocen con el nombre de **QTL**, del inglés ***Quantitative Trait Loci***. La localización de tales regiones basada únicamente en la evaluación fenotípica no es posible (Falconer y Mackay, 2001). Por lo tanto, uno de los mayores avances en el estudio de la arquitectura genética de estos caracteres fue conducido por el desarrollo de la genómica estructural (Falconer y Mackay, 2001; Bernardo, 2002). En particular, los **marcadores moleculares (MM)** son una herramienta que permite estudiar variaciones genéticas directamente al nivel del **ácido desoxirribonucleico (ADN).** Los MM tienen la ventaja de no variar con el ambiente y pueden ser observados en etapas tempranas del ciclo de vida de un individuo. Por estas razones, desde los años 1980s, los MM han sido ampliamente utilizados y estudiados para responder en qué medida pueden optimizar esquemas de mejoramiento genético (Bernardo y Yu, 2007).

Los MM permiten no sólo caracterizar el genoma de un individuo sino también, la construcción de **mapas de ligamiento** (Collard *et al.*, 2005). Estos mapas pueden ser utilizados para localizar regiones del cromosoma que contienen genes que controlan caracteres simples o complejos. Si los MM están asociados a los QTL, se dice que se encuentran en **desequilibrio de ligamiento**. Por lo tanto, tanto los marcadores localizados sobre genes como los ligados a los QTL se pueden convertir en una herramienta molecular que ayuda en el proceso de selección de los individuos en el mejoramiento genético.

Los primeros intentos de incorporar MM al estudio de caracteres fenotípicos de interés se basaron en la **localización o mapeo de QTL** (Soller y Plotkin-Hazan, 1977; Soller, 1978). El mapeo de QTL asume que existen unas pocas regiones del genoma que contienen genes que afectan a un carácter. El objetivo es localizar el o los QTL en el genoma y estimar la magnitud de sus efectos sobre el carácter. Una vez identificado un QTL, se utilizan MM flanqueantes para efectuar la selección, proceso que se denomina **selección asistida por marcadores**. Esta metodología generó importantes progresos en la selección para muchos caracteres en distintas especies. No obstante, el impacto que ha tenido en el mejoramiento genético se vio acotado puesto que la proporción de varianza explicada por un QTL puede ser pequeña y difícil de detectar (de los Campos *et al.*, 2013). Además, otro factor limitante es que requiere de una gran inversión de tiempo y recursos económicos para generar los datos que permiten emplear esta técnica.

El desarrollo de nuevos enfoques de selección asistida por marcadores ha sido alentado por la combinación de varios hechos. En primer lugar, hay un consenso general en que muchos caracteres suelen ser afectados por un gran número de genes de efectos pequeños y por lo tanto, su estudio requiere la consideración de un gran número de variantes genéticas (Lorenz *et al.*, 2011). Por otro lado, para el mejoramiento genético de caracteres complejos en animales y plantas, es clave predecir los **valores genéticos**, es decir, cuánto se desvía cada individuo respecto del promedio de la población a la que pertenece (Crossa *et al.*, 2010). Finalmente, el advenimiento de tecnologías que permiten identificar grandes cantidades de MM más rápidamente y a menor costo, ha motivado el uso de MM a gran escala en los programas de mejoramiento (Bernardo y Yu, 2007).

Meuwissen *et al.* (2001) fueron los primeros en introducir las nuevas técnicas que se conocen como métodos de **selección genómica (SG)** o en inglés, *Genome-Wide Selection*. Su trabajo, enmarcado dentro del mejoramiento animal, propuso la incorporación de numerosos MM en los modelos estadísticos utilizados para estimar el valor genético de un individuo. Los autores

implementaron una idea simple pero poderosa: expresar los fenotipos sobre todos los marcadores disponibles usando un modelo lineal. La aplicación de SG consiste de dos grandes etapas. La primera comprende el desarrollo de un modelo basado en un conjunto de individuos para los cuales se cuenta tanto con datos fenotípicos como con datos provenientes de una alta densidad de MM ubicados a lo largo del genoma. En la segunda etapa, se utiliza dicho modelo con el fin de predecir los valores genéticos de otros individuos para los cuales sólo se dispone de datos de MM (Thomson, 2014).

En la actualidad, la SG ha ganado terreno no sólo en el mejoramiento animal (VanRaden *et al.*, 2009) sino también en el mejoramiento vegetal (Nakaya e Isobe, 2012). Estudios de SG en especies vegetales fueron reportados a partir del año 2007. En Piyasatian *et al.* (2007) se simuló la eficiencia de la SG en una cruza de líneas endocriadas pero sin especificar ninguna especie. El primer estudio de simulación en una especie en particular fue el de Bernardo y Yu (2007) que consistió en la comparación de la SG y la selección asistida por marcadores en maíz (*Zea mays* L.), dando evidencia de que la primera es más efectiva. Los estudios de simulación continuaron presentándose no sólo en maíz (Mayor y Bernardo, 2009; Bernardo, 2009) sino también en otras especies tales como palma aceitera (*Elaeis guineensis* Jacq.) y cebada (*Hordeu vulgare* L.) (Wong y Bernardo, 2008; Bernardo, 2010; Zhong *et al.*, 2009; Jannink, 2010 y Iwata y Jannikk, 2011). El primer trabajo empírico sobre SG fue en múltiples especies: maíz, cebada y *Arabidopsis thaliana* (Lorenzana y Bernardo, 2009); en todos los casos se trataba de poblaciones biparentales. Posteriormente, se presentaron más trabajos en trigo (*Triticum aestivum L.*) y en maíz (Piepho, 2009; Crossa *et al.*, 2010; Heffner *et al.*, 2011; Guo *et al.*, 2012).

En general, la precisión de los métodos de SG realizados en estudios empíricos para vegetales resultó mayor que la alcanzada en el caso de animales. Adicionalmente, en la mayoría de los trabajos en vegetales se empleó menor cantidad de marcadores. Los estudios empíricos sobre vegetales muestran que la SG es un método con potencial para el mejoramiento vegetal y que

puede desarrollarse con tamaños de poblaciones y cantidad de marcadores que se observan en la práctica del mejoramiento (Nakaya e Isobe, 2013).

Dada la gran cantidad de genes intervinientes en los caracteres cuantitativos, el objetivo principal de los métodos de SG ya no es la localización y estimación de efectos de los QTL. El propósito es incorporar la gran cantidad de datos de MM a modelos estadísticos para lograr predecir los valores genéticos de los individuos. Tales predicciones promueven la selección de los mejores individuos en etapas tempranas de su ciclo de vida ya que, los MM que se pueden observar incluso antes registrar los caracteres fenotípicos. La posibilidad de evitar la observación del fenotipo, se traduce en una significativa reducción de tiempo y costos (de los Campos *et al.*, 2009).

En este sentido, uno de los mayores desafíos del mejoramiento genético basado en MM es lograr buenas predicciones del valor genético y es aquí donde los métodos estadísticos juegan un rol crucial. El conjunto de datos disponibles para estimar el modelo de SG incluye datos fenotípicos de distinto nivel de agrupamiento (múltiples ambientes, repeticiones, entre otros) y datos de MM. Si bien hay distintos tipos de MM, todos ellos pueden codificarse para incorporarlos a los modelos estadísticos como factores (variables cualitativas) o como covariables. Además, es posible contar con datos de pedigrí de los individuos, es decir, con información acerca de las estructuras de parentesco entre los mismos. Se propusieron varias metodologías basadas tanto en enfoques paramétricos como semi-paramétricos, con el fin de incorporar toda la información mencionada en los modelos estadísticos (Gianola *et al.*, 2006).

El presente trabajo se ocupa de la primera etapa de la SG, es decir: el desarrollo del modelo. Así, el objetivo es evaluar distintos métodos estadísticos de SG que permitan expresar un carácter cuantitativo en función del valor genético de los individuos basado en MM. Puntualmente, se estudiarán métodos estadísticos paramétricos que permiten predecir los valores genéticos de los individuos en función de los MM, empleando un modelo de regresión. La ventaja de estos

métodos es que permiten detectar marcadores que afectan significativamente al carácter fenotípico y por lo tanto posibilitan la identificación de regiones del genoma asociadas al carácter de interés (de los Campos *et al.*, 2009).

Sin embargo, la incorporación de una gran cantidad de MM en un modelo de regresión puede tener efectos no deseados. El número de variables explicativas (*p*), correspondientes a los MM del modelo, generalmente llega a ser tan grande que supera al número de individuos (*n*), afectando las estimaciones de los parámetros. Este fenómeno es conocido como **problema de dimensionalidad**.

Otra desventaja del uso de numerosos MM en el modelo de regresión es que pueden dar origen a **problemas de multicolinealidad**. El fenómeno estadístico de multicolinealidad se refiere a la presencia de asociación lineal entre las variables explicativas en un modelo de regresión. Este problema afecta las estimaciones de los parámetros pues, induce la sobreestimación de las varianzas de los estimadores. En particular, la multicolinealidad puede estar presente ya que los MM más cercanos en el mapa de ligamiento se encuentran fuertemente asociados (desequilibrio de ligamiento).

Los modelos de SG requieren la implementación de métodos estadísticos que puedan confrontar los problemas mencionados para obtener buenas predicciones. Entre ellos se pueden mencionar: 1) técnicas de selección de variables, 2) procedimientos de estimación penalizada y 3) combinación de selección de variables y estimación penalizada.

Las técnicas de selección de variables**,** permiten elegir sólo algunos marcadores, reduciendo de manera notable el número de variables explicativas en el modelo. No obstante, dada la gran cantidad de MM no es posible usar métodos automáticos de selección como lo son los métodos de selección hacia adelante y selección paso a paso. Por esta razón, en este trabajo se implementa la selección por marcador individual y luego, se realiza la estimación del modelo por

medio del método clásico: mínimos cuadrados ordinarios. La estimación penalizada se aborda empleando la **regresión de Ridge** introducida por Hoerl y Kennard (1970). Esta metodología penaliza los coeficientes de los marcadores comprimiendo sus estimaciones hacia el valor cero. Finalmente, la combinación de ambas estrategias, es decir: selección de variables conjuntamente con estimación penalizada, es abordada a través del método **LASSO** del inglés ***Least Absolute Shrinkage and Selection Operator*** (Tibshirani, 1996).

La tesis se encuentra estructurada en diferentes secciones. Los objetivos se plantean en la siguiente sección mientras que en la Sección 3 se describen detalladamente los datos, cubriendo conceptos propios de su proceso de generación dentro del programa de mejoramiento, describiendo las poblaciones utilizadas, los datos fenotípicos y los datos moleculares disponibles. En la Sección 4, el foco es estudiar los métodos estadísticos comenzando por aquéllos correspondientes al análisis de datos fenotípicos y de MM separadamente; incluyendo a su vez, conceptos introductorios de los MM para comprender el análisis realizado. Luego, en el Apartado 4.3 se presentan los distintos enfoques de los modelos de SG correspondientes al análisis simultáneo de los datos fenotípicos y de MM; además, en dicho apartado se presentan los métodos utilizados para evaluar la capacidad predictiva de los modelos.. Los resultados de los análisis son presentados en la Sección 5 y finalmente, la Sección 6 resume los resultados hallados contrastándolos con distintos antecedentes bibliográficos y planteando futuras líneas de trabajo.

2. OBJETIVOS

2.1. OBJETIVO GENERAL

Evaluar distintos modelos estadísticos de selección genómica que permiten predecir la expresión fenotípica de un carácter cuantitativo en función del valor genético de los individuos utilizando marcadores moleculares, en el contexto de un programa de mejoramiento de maíz.

2.2. OBJETIVOS ESPECÍFICOS

- Estudiar metodologías estadísticas que utilicen estrategias de selección de variables y/o de estimación penalizada para superar el problema de dimensionalidad presente en los datos de selección genómica.
- Aplicar las metodologías de selección genómica a un carácter fenotípico de una población de maíz y evaluar la habilidad predictiva en cada caso.
- Comparar las metodologías de estimación de modelos de selección genómica en términos de su habilidad predictiva utilizando múltiples poblaciones y caracteres fenotípicos.
- Estudiar si la habilidad predictiva del modelo depende de factores tales como el carácter fenotípico, y características de la población.

3. MATERIALES

Los distintos métodos de SG se aplicaron a un conjunto de datos proveniente de un programa de mejoramiento de maíz establecido en Estados Unidos y perteneciente a la compañía multinacional Monsanto. Un programa de mejoramiento de maíz sigue diferentes esquemas compuestos por procedimientos multietápicos a lo largo de varios años entre los cuales se puede implementar métodos de SG. A continuación se presentan: 1) una breve descripción de los procedimientos involucrados en la generación de los datos a los cuales se aplicaron los métodos de SG en este trabajo, 2) información propia de los datos fenotípicos observados y 3) descripción los datos de MM con que se trabajó.

3.1. Proceso de generación de datos para selección genómica

El principal objetivo de un programa de mejoramiento genético de plantas es la generación y selección de nuevas combinaciones de genes para crear genotipos con un carácter fenotípico que supera a los genotipos ya existentes, en un conjunto objetivo de ambientes (Chapman *et al.*, 2003). En el caso particular del maíz, la base de los programas de mejoramiento genético es el desarrollo de líneas endocriadas y la evaluación del desempeño de los híbridos que se originan al cruzar esas líneas (Hallauer *et al.*, 1988). A continuación se describen ambos procesos.

Una **línea pura o endocriada** de maíz se define como una entidad genéticamente estable, que puede producirse a través de repetidas auto-fecundaciones (Bernardo, 2002). Por otro lado, un **híbrido** puede producirse por medio de: la cruza de dos líneas (híbrido simple), o de una cruza simple con otra línea (híbrido triple) o cruzando dos híbridos simples (hibrido doble) (Bernardo, 2002). La generación originada de la cruza de dos líneas (por ejemplo línea A y línea B) se simboliza como F1. Si cada una de las plantas F1 es auto-fecundada, se obtiene lo que se

denomina F2. Este proceso puede continuarse repetidamente, en términos generales, *n* veces hasta lograr genotipos F*n* altamente homocigotos y de esta manera se logra desarrollar una nueva línea endocriada (Figura 3.1.1, parte A).

El conjunto de individuos de una misma generación se denomina **población.** Por lo tanto, se suele hablar de poblaciones F2, …, Fn para hacer referencia a las poblaciones de distintas generaciones. A su vez, las poblaciones están compuestas por **familias** que representan un conjunto de individuos que tiene el mismo origen en la generación inmediatamente anterior. Por ejemplo, un planta F2 tiene una espiga con semillas que darán origen a una familia de individuos F3, todas esas plantas F3 conforman una familia por provenir de la misma planta F2.

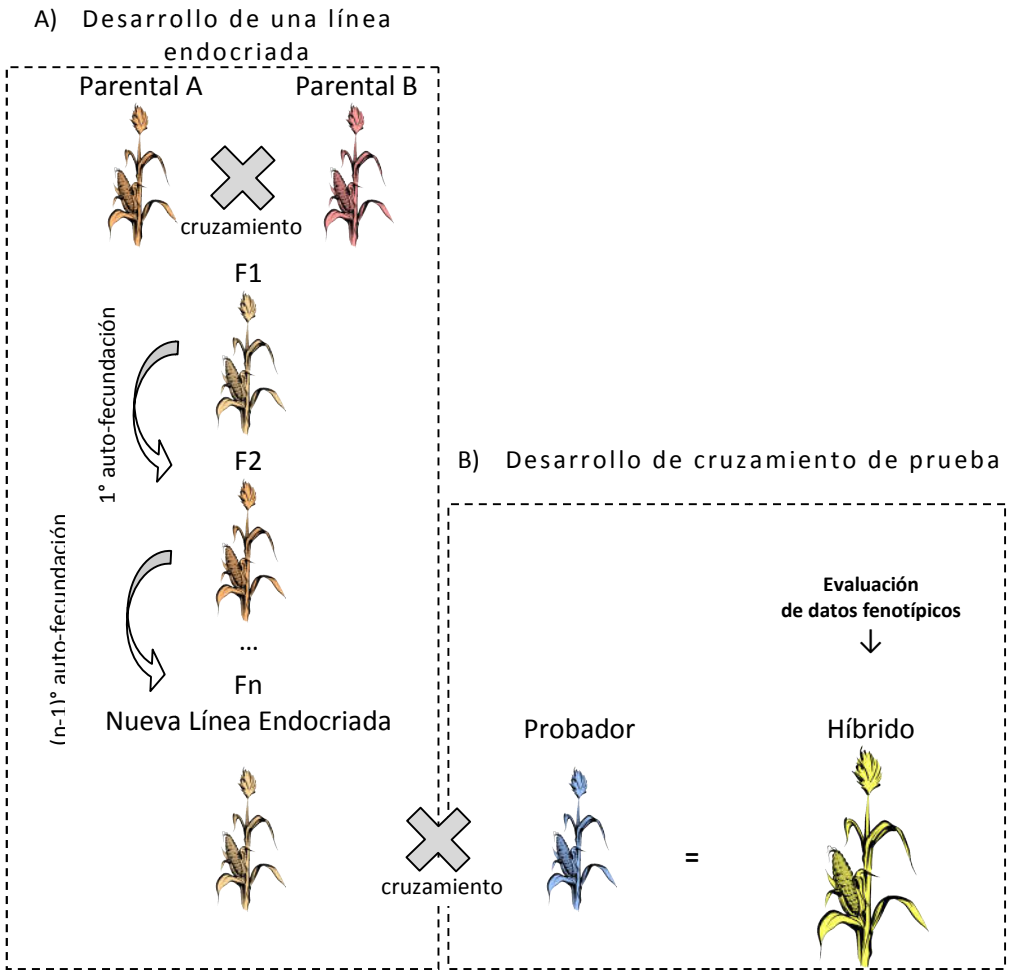

Figura 3.1.1 - Esquema simplificado de mejoramiento genético en maíz

Existe un procedimiento muy común para evaluar los individuos de una población en cultivos de polinización cruzada como el maíz. El mismo consiste en realizar cruzamientos con una línea denominada **probador,** diferente de las líneas parentales que dieron origen a la población (Figura 3.1.1, parte B). De esta manera, se obtiene un híbrido por cada familia de la población original y a cada uno de esos híbridos se los denomina **cruzamiento de prueba** (Bernardo, 2002). Generalmente, las líneas parentales A y B son del mismo grupo heterótico mientras que el probador pertenece a un grupo heterótico diferente. Un **grupo heterótico** es un conjunto de líneas que tienen desempeño similar cuando se cruzan con líneas de otro grupo heterótico (Bernardo, 2002).

Específicamente, los materiales con que se contó en este trabajo consistieron de 20 poblaciones F3, 10 poblaciones por cada uno de dos grupos heteróticos utilizados en el programa: GH1 y GH2 (Tabla 3.1.1). Cada población tuvo como origen la cruza de dos parentales (denominados L01 a L32) y el cruzamiento de prueba fue realizado con un determinado probador (denominados P01 a P12). El tamaño de las poblaciones varió entre 76 y 186 familias.

La Figura 3.1.2 presenta un esquema del programa de mejoramiento de maíz que genera los datos utilizados para estimar los modelos de SG en una población F3. El proceso se inició con la cruza de dos parentales (A y B) de un mismo GH, de donde se cosechó la semilla que conforma la F1. Esa semilla se sembró y las plantas F1 fueron auto-fecundadas para obtener semilla que dio origen a la población F2. Nuevamente, se procedió a la auto-fecundación, en esta instancia, de las plantas F2. Cada espiga individual de la población F2 dio origen a semillas F3. Las semillas F3 de una espiga individual conformaron una familia F3, compartiendo el mismo origen: la planta F2 de donde proviene esa espiga. Por lo tanto, se obtuvieron distintas familias F3: familia_1, familia_2, …, familia_n.

Tabla 3.1.1 – Descripción de las poblaciones F3

Población	Grupo Heterótico	Origen de la Población	Probador	No. Familias	No. SNPs Evaluados
1	GH1	L01/L02	P01	181	99
2	GH1	L03/L04	P02	134	86
3	GH1	L05/L06	P03	136	92
4	GH1	L07/L01	P04	143	95
5	GH1	L01/L08	P05	184	94
6	GH1	L09/L10	P04	90	97
7	GH1	L11/L12	P04	186	100
8	GH1	L13/L14	P04	179	102
9	GH1	L15/L09	P04	182	89
10	GH1	L16/L17	P03	181	90
11	GH2	L18/L19	P06	161	86
12	GH2	L20/L21	P07	132	80
13	GH2	L22/L23	P08	177	70
14	GH2	L24/L25	P09	178	101
15	GH2	L26/L27	P06	158	92
16	GH2	L26/L23	P06	157	96
17	GH2	L28/L29	P08	184	103
18	GH2	L19/L30	P10	159	81
19	GH2	L31/L29	P11	182	101
20	GH2	L32/L18	P12	73	100

Las familias F3 fueron la unidad de estudio de este trabajo, es decir, para cada familia se obtuvo un dato fenotípico y otro molecular. Con tal finalidad, de cada familia F3 se extrajeron dos muestras de semilla, una que se envió al laboratorio para obtener los datos de MM y otra que se utilizó para el desarrollo de los cruzamientos de prueba (además, se destinó semilla para continuar con las auto-fecundaciones y así desarrollar líneas endocriadas, pero se trata de una parte del proceso que excede el alcance de este trabajo). En el laboratorio, se realizó la extracción de ADN de la muestra para cada familia y se obtuvieron los genotipos de un cierto número (p) de marcadores.

Por otro lado, para el estimar el modelo de SG, fue necesario contar con datos fenotípicos de las familias a través del desarrollo del cruzamiento de prueba que implicó cruzar cada familia F3 con un probador (P). Los híbridos obtenidos de este cruzamiento de prueba fueron sembrados en

q ambientes y se observaron los caracteres fenotípicos de interés. Los datos fenotípicos fueron analizados a través de un modelo estadístico (Sección 4.1) para resumirlos a un único dato por familia (valor genético, BLUPs). Finalmente, la matriz de datos para la población F3 estuvo compuesta por el genotipo de los *p* marcadores y valor genético de cada una de las *n* familias de la población.

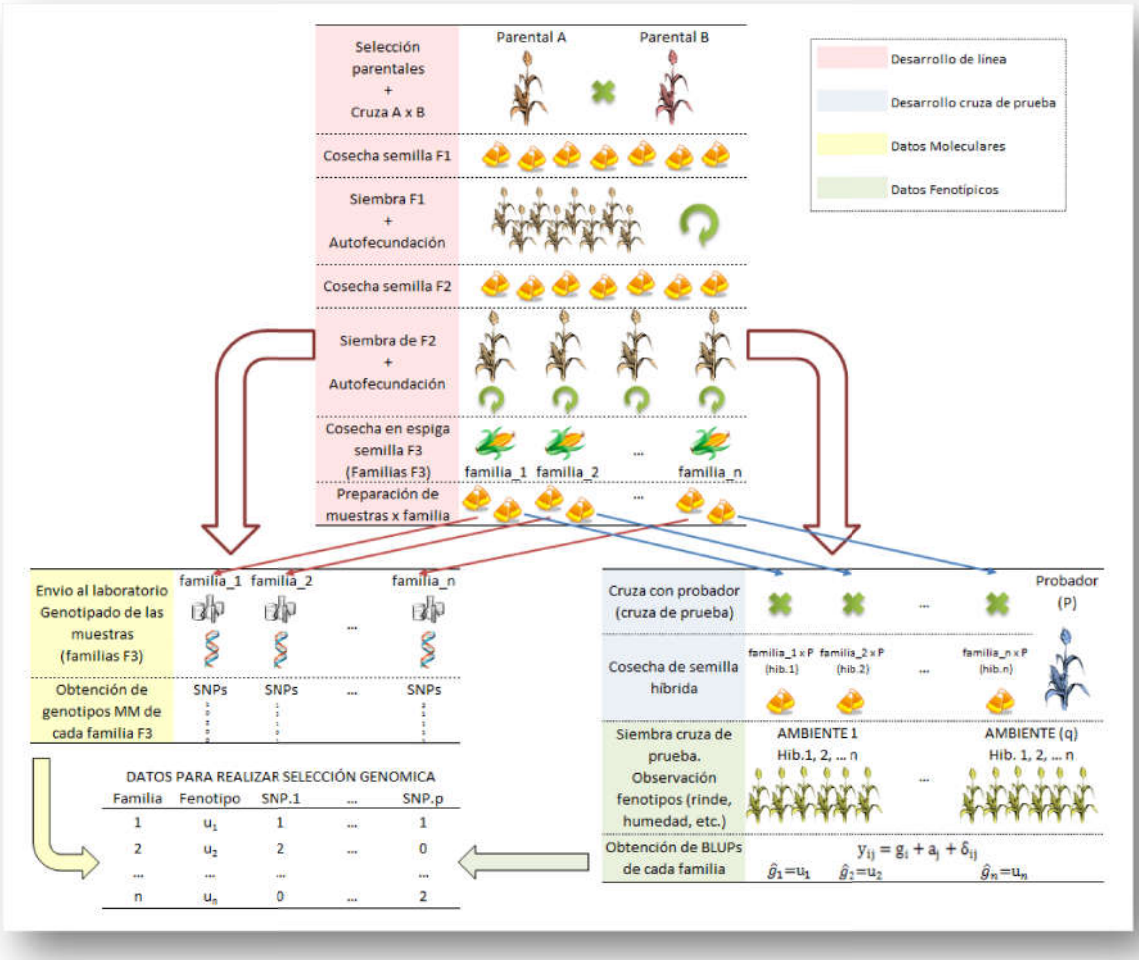

Figura 3.1.2 – Esquema de mejoramiento de maíz y generación de datos fenotípicos y de MM necesarios para estimar el modelo de SG en una población F3.

3.2. Datos fenotípicos

La caracterización fenotípica de las poblaciones se realizó para cada población por separado, siguiendo un diseño de bloques completos aleatorizados. Los bloques se corresponden con ambientes y estos ambientes consistieron en localidades de Estados Unidos, evaluadas en mismo año (que varía entre 2002 y 2008 según la población). Cada población fue evaluada en 4 a 8 ambientes contando con una sola repetición por ambiente. En todos los casos se midieron los siguientes caracteres: rendimiento en grano medido en bushel/acre (Rendimiento, bu/ac), peso hectolítirico medido en libras por bushel (Peso Hectolítrico, lb/bu) y porcentaje de humedad del grano a cosecha (Humedad, %). Hubo casos de familias para las cuales no se encontró disponible la evaluación fenotípica en la totalidad de los ambientes en que se pretendían evaluar.

3.3. Datos de marcadores moleculares

Los datos de MM para cada población fueron el resultado de implementar una estrategia de múltiples etapas para reducir el costo total de los estudios de SG. La misma consistió en obtener genotipos de una alta densidad de MM para un conjunto principal de líneas endocriadas y de menor densidad a todas las poblaciones de individuos derivadas de ese conjunto principal (He *et al.*, 2015). Así, los datos moleculares para cada población, fueron el resultado de varias etapas. En primer lugar, para las líneas parentales (L01-L32) se obtuvieron genotipos de 2.911 marcadores de polimorfismo de nucleótidos simples (**SNP del inglés, *Single Nucleotide Polymorphism***) distribuidos a lo largo de todo el genoma de maíz. Para mantener la confidencialidad de los datos, los cromosomas fueron codificados aleatoriamente con letras y, con base en la localización en el mapa de ligamiento, a los MM se les dio un orden o posición dentro del cromosoma.

Por otro lado, y de acuerdo al esquema descripto en la Figura 3.1.2, para las poblaciones también se obtuvieron datos moleculares pero para un número menor de SNPs, como se detalla en la Tabla 3.1.1. Este subconjunto de SNPs se caracterizó por ser informativo o polimórfico para los parentales de la población, es decir, que los parentales eran diferentes para ese marcador (si los parentales fueran iguales para un MM, toda su descendencia no presentaría variabilidad para ese MM y no arrojaría información para asociar a el carácter fenotípico sobre el cual se desea hacer selección).

El tercer paso realizado por la empresa surge por haber implementado una estrategia para reducir el costo total de los estudios de SG. La estrategia consistió en obtener genotipos de una alta densidad de MM para un conjunto principal de líneas endocriadas y con menor densidad para todas las poblaciones de individuos derivadas de ese conjunto principal (He *et al.*, 2015). Luego, para incrementar la cobertura del genoma alcanzada con los MM se efectuó imputación de MM en las poblaciones. Este proceso de genotipificación e imputación fue recomendado por Jacobson *et al.* (2015) en poblaciones biparentales de maíz provenientes del mismo programa de mejoramiento de Monsanto del cual se obtuvieron los datos del presente trabajo. El método de imputación que empleó la compañía fue el de regresión basada en los marcadores flanqueantes o mapeo por intervalos propuesto por (Haley y Knott, 1992) y condujo a un número total de 2.800 MM, aproximadamente, en cada población.

4. METODOLOGÍA

La aplicación de métodos de SG involucró el análisis de datos fenotípicos y de MM. Así, en esta sección se presentan en primer lugar, las metodologías propias del análisis de los datos fenotípicos. Luego, se desarrollan los métodos para el análisis de los datos de MM. Finalmente, se plantean los modelos de SG que combinaron ambas fuentes de información, se estudian distintos enfoques para llevar a cabo su estimación y se describe la estrategia para evaluar y comparar su habilidad predictiva a través de la aplicación a los datos del programa de mejoramiento de maíz.

4.1. Metodología para el análisis de datos fenotípicos

El mejoramiento genético, tanto animal como vegetal, tiene sus bases en la ciencia de la genética cuyo principal rol es establecer los factores más importantes que afectan a diferentes caracteres fenotípicos. Una vez que los factores principales han sido reconocidos, las investigaciones se enfocan en la identificación de genes que tienen efectos secundarios (Siegmund y Yakir, 2007).

En particular, la genética cuantitativa es la ciencia que se dedica al estudio de la herencia de los caracteres de tipo cuantitativo en los individuos y las diferencias que existen entre ellos (Falconer y Mackay, 2001). El conocimiento de la herencia de esas diferencias es de fundamental importancia para el mejoramiento genético; de hecho esto impulsó su desarrollo. La base teórica de la genética cuantitativa se estableció alrededor de 1920 con los trabajos de Fisher (1919), Wright (1921) y Haldane (1932). La premisa de la cual se parte es que la herencia de las diferencias cuantitativas se deben a causas genéticas y a que la expresión del genotipo en un determinado fenotipo puede modificarse por acciones no genéticas (Falconer y Mackay, 2001). Dicho de otra manera, la variación de los fenotipos es influenciada tanto por una componente

genética propia del individuo como por el ambiente al que el individuo se encuentra expuesto. Este concepto puede verse reflejado en un modelo estadístico que se plantea a continuación. Asumiendo que se observan *n* individuos en *q* ambientes y simbolizando w_{ij} al carácter observado para el *i*-ésimo individuo ($i = 1, 2, \dots, n$) en el *j*-ésimo ambiente ($j = 1, 2, \dots, q$), el modelo tiene la forma:

$$\mathrm{w_{ij}} = \mu + \mathrm{g_i} + \mathrm{a_j} + \delta_{\mathrm{ij}} \tag{4.1.1}$$

donde μ es la media general, $\mathrm{g_i}$ es el efecto genotípico propio del *i*-ésimo individuo, $\mathrm{a_j}$ es el efecto del *j*-ésimo ambiente y δ_{ij} es el error aleatorio. Se asume que los efectos de ambiente ($\mathrm{a_j}$), genotipo ($\mathrm{g_i}$) y el error (δ_{ij}) son aleatorios, independientes y siguen una distribución normal con media cero y varianzas $\sigma^2_{\mathrm{A}}, \sigma^2_{\mathrm{G}}$ y σ^2_{δ}, respectivamente. El modelo puede extenderse, incorporando un término para contemplar la interacción entre genotipo y ambiente; sin embargo, con los materiales disponibles para este trabajo no se puede evaluar dicho efecto.

El modelo de la Ecuación 4.1.1 se puede expresar en notación matricial como sigue:

$$\boldsymbol{w} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\delta} \tag{4.1.2}$$

donde $\mathbf{w}$ es el vector de los datos fenotípicos, $\boldsymbol{\beta}$ es el vector de efectos fijos (en este caso, simplemente la media general), $\mathbf{X}$ es la matriz de incidencia correspondiente a los efectos fijos, $\mathbf{u}$ es un vector de efectos aleatorios (genotipo y ambiente), $\mathbf{Z}$ es la matriz de incidencia correspondiente a los efectos aleatorios y $\boldsymbol{\delta}$ es el vector de errores aleatorios. Además, el supuesto sobre las componentes aleatorias es equivalente a la expresión:

$$\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\delta}\end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{D} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right) \tag{4.1.3}$$

siendo $\mathbf{D} = \begin{bmatrix}\sigma^2_{\mathrm{G}}\mathbf{I} & 0\\ 0 & \sigma^2_{\mathrm{A}}\mathbf{I}\end{bmatrix}$ y $\mathbf{R} = \sigma^2_{\delta}\mathbf{I}$. Así, la varianza del vector de datos resulta:

$$\mathrm{V} = \mathrm{V}(\mathbf{w}) = \mathbf{ZDZ}' + \mathbf{R} \tag{4.1.4}$$

Searle (1992) demuestra que, siendo $(\ \ )^{-}$ la inversa generalizada de una matriz, el mejor estimador lineal insesgado de $\boldsymbol{\beta}$ viene dado por:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{w} \qquad (4.1.5)$$

Mientras que el **mejor predictor lineal insesgado (BLUP, del inglés *Best Linear Unbiased Predictor*)** de u resulta:

$$\widehat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{w} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \qquad (4.1.6)$$

Patterson y Thompson (1971) introdujeron el método de estimación por **máxima verosimilitud restringida (REML del inglés, *REstricted Maximum Likelihood*)** que incluye un ajuste en los grados de libertad usados para estimar los efectos fijos ya que generalmente, las componentes de varianza también son parámetros desconocidos que se desean estimar. En presente trabajo, se aplicó este método de estimación usando el paquete estadístico R (R Core Team, 2013) y empleando la función **lmer** de la librería **lme4** (Bates, *et al.*, 2013).

Para evaluar la bondad del ajuste del modelo, se calculó el R^2 condicional de Nakagawa y Schielzeth (2013) que contempla tanto los efectos fijos como los efectos aleatorios de un modelo. En el caso del modelo de la Ecuación 4.1.1, todos los efectos son aleatorios y los errores se asumen distribuidos normalmente, el R^2 condicional tiene la siguiente forma:

$$R^2 = \frac{\widehat{\sigma}^2_G + \widehat{\sigma}^2_A}{\widehat{\sigma}^2_G + \widehat{\sigma}^2_A + \widehat{\sigma}^2_\delta} \qquad (4.1.7)$$

Del modelo se pueden extraer distintas estimaciones de particular interés e interpretación en la genética cuantitativa. Por un lado, el efecto genotípico (g), conocido como el valor genético de un individuo, representa el valor fenotípico promedio para un cierto genotipo si éste se pudiera estudiar sobre el universo de todos los ambientes posibles a los cuales los individuos podrían estar expuestos. Asimismo, se suele hacer referencia con el término de valor genético al desvío de un individuo respecto la media de todos los individuos. Lograr buenas estimaciones de los valores

genéticos de los individuos para un carácter poligénico observable ayuda a realizar la selección de los mejores individuos o genotipos dentro de una población, camada o grupo determinado (de los Campos *et al.*, 2009).

Por otro lado, la estimación de las componentes de varianza de las distintas fuentes de variabilidad subyacentes en el fenotipo se transforma en otra de las herramientas utilizadas en genética cuantitativa (Lynch y Walsh, 1998). La **heredabilidad** de un carácter fue definida originalmente por Lush (1943) como la proporción de varianza fenotípica entre los individuos de una población explicada por efectos genéticos heredables. En la actualidad, este concepto se conoce como heredabilidad en sentido estricto y se simboliza h^2. Del mismo modo, surgieron distintas definiciones o tipos de heredabilidad; en este trabajo se empleó el concepto de heredabilidad de las medias de las familias a través de ambientes (Holland *et al.*, 2003) cuya estimación viene dada por la siguiente expresión:

$$\hat{\mathrm{h}}^2 = \frac{\hat{\sigma}^2_{\mathrm{G}}}{\hat{\sigma}^2_{\mathrm{G}} + \hat{\sigma}^2_{\delta}/\mathrm{q}} \qquad (4.1.8)$$

siendo σ^2_{G} la componente de varianza correspondiente a los valores genéticos, σ^2_{δ} la componente de varianza propia de los errores experimentales y *q* el número de ambientes.

En el presente trabajo, en cada una de las 20 poblaciones de maíz y para cada uno de los caracteres fenotípicos rendimiento en grano (bu/ac), peso hectolítrico (lb/bu) y humedad del grano a cosecha (%) se ajustó el modelo lineal de la Ecuación 4.1.1 con efectos aleatorios de ambiente y de genotipo. Notar que, los individuos en el modelo de la Ecuación 4.1.1 corresponden a familias de una población. El valor genético ($\hat{\mathrm{g}}$), fue utilizado en las etapas posteriores como variable respuesta en los modelos de SG.

4.2. Metodología para el análisis de datos moleculares

4.2.1. Introducción a los marcadores moleculares

La variación genética en individuos se puede presentar en distintos niveles. En uno de los extremos están las variantes genéticas manifestadas con efectos fácilmente distinguibles en el fenotipo (caracteres cualitativos) como por ejemplo, el color de las flores. El otro extremo ocurre cuando una gran parte de la variación entre individuos no da lugar a clases fenotípicas naturalmente definidas (caracteres cuantitativos) por ejemplo, el número de granos por espiga. El supuesto biológico detrás de la variabilidad de los caracteres cuantitativos es que su expresión está regulada por una gran cantidad de componentes genotípicos y sus interacciones con el ambiente. Luego, para comprender las variaciones genéticas detrás de los caracteres cuantitativos se emplean técnicas como los MM que permiten captar las diferencias genéticas de los individuos en el ADN para luego, estudiar la expresión de estos genotipos en distintos ambientes.

El ADN contiene instrucciones genéticas usadas en el desarrollo y funcionamiento de todos los organismos eucariotas, muchos microorganismos y algunos virus. El papel principal de la molécula de ADN es el almacenamiento a largo plazo de información necesaria para construir los componentes de las células y las proteínas esenciales para la estructura y función de organismos.

En una noción simplificada, un **gen** es una secuencia distintiva de ADN que contiene todas las instrucciones para sintetizar una proteína. En los organismos que se reproducen sexualmente (como plantas y animales), la reproducción involucra la unión de dos gametos femenino y masculino derivados del mismo parental o de diferentes parentales (Fehr, 1987). Por lo tanto, hay al menos dos copias de cada gen (si tienen exactamente dos copias se los conoce como organismos diploides) y cada copia proviene de cada uno de los organismos progenitores.

Las variaciones o secuencias alternativas de ADN son llamadas **alelos** y ellos producen diferencias en la cantidad o tipo de proteína producida por un gen específico. Por ejemplo, si se

asume que existen dos alelos para un gen (organismo diploide), simbolizados como A y B y que un individuo puede heredar exactamente dos alelos, las distintas combinaciones de alelos que se pueden dar en el individuo definen el genotipo para ese gen y sus valores posibles son: AA, BB (**genotipos homocigotos**) o AB/BA (**genotipos heterocigotos**).

El ADN se encuentra organizado en estructuras dentro de la célula llamadas **cromosomas** (Figura 4.2.1). Una posición específica dentro del cromosoma de llama **locus** mientras que múltiples posiciones se denominan **loci**. Watson y Crick (1953) propusieron un modelo para la estructura del ADN de doble hélice que está compuesto por pares de cuatro nucleótidos o bases: adenina (A), citosina (C), guanina (G) y timina (T). Químicamente, se ha demostrado que en cualquier molécula de ADN, el nucleótido A solamente forma par con la base T y la base C solo forma par con la G.

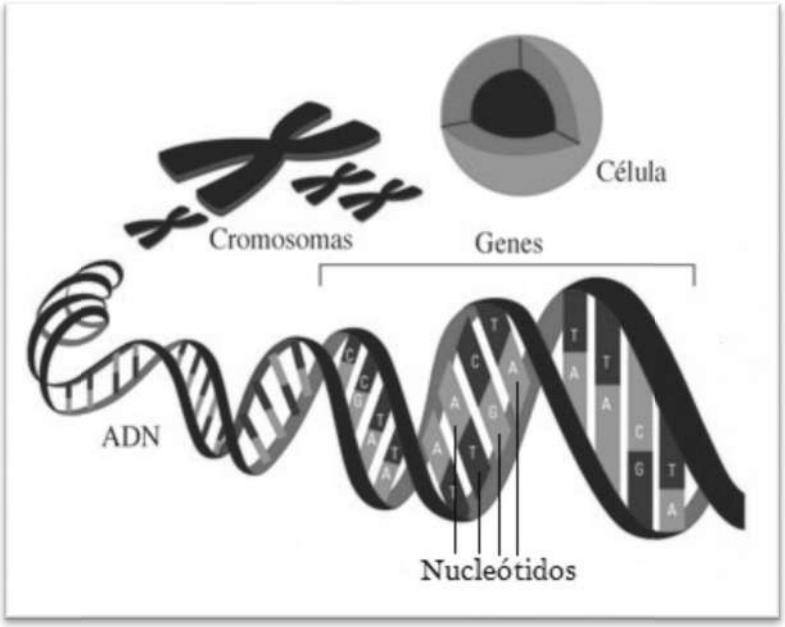

Figura 4.2.1 – Esquema de ADN dentro de una célula (CRI México, s. f.)

Las variaciones de las secuencias de ADN se pueden detectar utilizando una amplia variedad de técnicas (Lynch y Walsh, 1998) desde la secuenciación del ADN, es decir, la determinación de la secuencia de nucleótidos, hasta caracterizaciones menos precisas (más rápidas y menos costosas) logradas a través de distintos tipos de MM.

Uno de los MM más utilizados en la actualidad es denominado SNP. Los SNPs, como sugiere el nombre, identifican la variación de nucleótidos en un locus o posición específica en la secuencia de nucleótidos del genoma. Dada la forma en que se disponen los pares de nucleótidos (A-T,C-G), en la práctica los SNPs son típicamente bialélicos, es decir, tienen dos alelos posibles A o T por un lado o bien, C o G por otro. Por lo tanto, los genotipos de un SNP pueden ser AA, AT/TA y TT o bien CC, CG/GC y GG. Generalmente, estos genotipos se codifican con números, -1, 0 y 1 representando con 0 los casos heterocigotos y con -1/1 los casos homocigotos. La estructura de los datos de MM se ejemplifica a continuación con 10 MM y 5 familias de una población:

Familia	SNP1	SNP2	SNP3	SNP4	SNP5	SNP6	SNP7	SNP8	SNP9	SNP10
1	1	-1	-1	-1	0	0	0	0	1	-1
2	-1	-1	-1	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	-1	1
4	0	-1	-1	-1	-1	1	1	-1	-1	1
5	-1	1	1	1	0	0	0	0	-1	1

Cuando los loci se heredan de manera independiente unos con otros, es decir, no existe asociación entre los loci, se dice que se produjo **recombinación genética**. Sin embargo, cuando dos o más loci están cercanos dentro de un cromosoma, estos pueden heredarse juntos y a este fenómeno de asociación se lo conoce como **ligamiento**.

La Figura 4.2.2 representa gametos recombinantes y no recombinantes originados de auto-fecundar una F1. Específicamente, se parte de dos líneas endocriadas (Parental 1 y 2), que por su estabilidad genética tienen genotipos homocigotos para los dos marcadores observados

(Marcador 1 y 2). Se asume que para los parentales estos marcadores resultan polimórficos, es decir, los genotipos de ambos marcadores difieren entre los parentales. De la cruza de ambos parentales se obtiene la F1, que es heterocigota en ambos marcadores. La auto-fecundación de la F1 produce gametos recombinantes, es decir, aquéllos en donde el marcador 1 y el 2 se heredaron de manera independiente, con una frecuencia *Fr*. Por otro lado, con una frecuencia (1-*Fr*) se producen gametos no recombinantes, donde hubo ligamiento entre los marcadores.

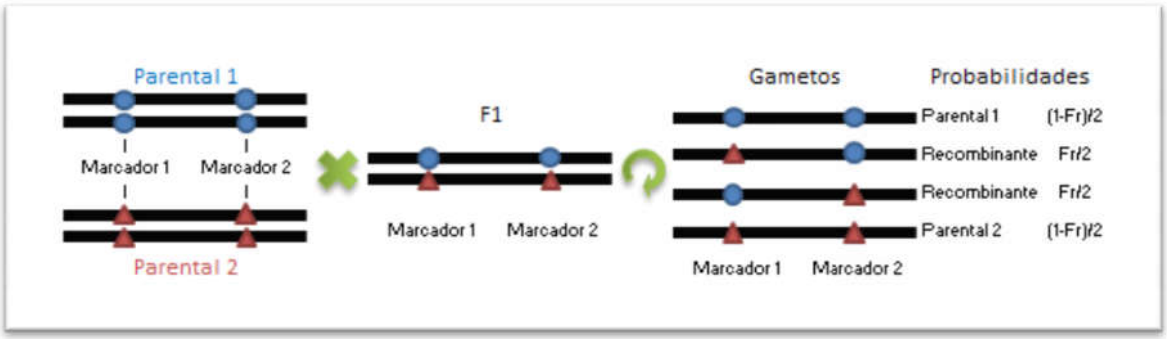

Figura 4.2.2 – Gametos parentales y recombinantes con una frecuencia de recombinación Fr.

Sin embargo, con un marcador no se identifica cada gameto, se registran los genotipos. Así, considerando el Marcador 1 del ejemplo, la F1 presenta genotipos heterocigotos para todos los individuos. Luego del primer ciclo de auto-fecundación, se espera que la población F2 tenga genotipos heterocigotos y homocigotos con frecuencias esperadas: 50% de genotipos heterocigotos, 25% homocigotos para el genotipo correspondiente con el Parental 1 y el 25% restante homocigotos para el genotipo del Parental 2 (Lynch y Walsh, 1998). Luego del segundo ciclo de auto-fecundación, las frecuencias esperadas para la población F3 son: 37.5% para los genotipos homocigotos correspondientes con cada parental y 25% para el genotipo heterocigoto.

Los MM permiten no sólo describir el genoma de un individuo sino también la construcción de **mapas de ligamiento**, con base en el estudio de la ocurrencia de recombinaciones *(Fr*). Estos mapas dan un ordenamiento a los marcadores y pueden ser utilizados para localizar regiones del cromosoma que contienen genes que controlan caracteres de interés. La distancia (*d*) en los

mapas de ligamiento se calcula usando funciones de mapeo y la unidad de medida es el **centiMorgan (cM).** Una de las funciones de mapeo más utilizada es la de Haldane (1919) y viene dada por la expresión: $d = -0{,}5\ ln(1 - 2Fr)$. Para más detalles de construcción de mapas de ligamiento se recomienda Lynch y Walsh (1998) y Siegmund y Yakir (2007).

4.2.2. Análisis de los marcadores moleculares

Una forma de resumir la información de un MM observado en una población, dada su naturaleza cualitativa, es calcular las **frecuencias genotípicas**, es decir, contabilizar cuántas familias de la población F3 tienen cada uno de los tres posibles genotipos para un SNP codificados como: -1, 0, 1. El estudio de estas frecuencias fue ejecutado en el programa estadístico R, permitiendo realizar un control de calidad para luego poder emplear estos datos en los modelos de SG. Si un marcador presenta un genotipo con una frecuencia mayor o igual al 90%, ese marcador es excluido por comportarse como un marcador monomórfico y por lo tanto no informativo. Los marcadores de las líneas parentales también se utilizan para describir cada población en términos de la **similaridad** entre sus parentales. En particular se considera la proporción de alelos compartidos entre las líneas. Si x_{ik} representa el genotipo de una línea *i* para el marcador *k* codificado con -1, 0 o 1, la similaridad entre una línea *i* y otra línea *i'* usando *p* marcadores viene dada por la siguiente expresión:

$$\mathrm{S_{i,i'}} = 1 - \frac{\sum_{k=}^{p} \mathrm{d_k}}{\mathrm{p}} = 1 - \frac{\sum_{k=}^{p} |\mathrm{x_{ik}} - \mathrm{x_{i'k}}|/2}{\mathrm{p}} \qquad (4.2.1)$$

Notar que si se comparan dos genotipos homocigotos diferentes es decir, $\mathrm{x}_{ik} = 1$ y $\mathrm{x}_{i'k} = -1$, el valor $d_k = 2/2 = 1$ significa que la proporción de alelos en que difieren ambas líneas es 1, en otras palabras, no tienen ningún alelo común y la similaridad resulta igual a 0. Si $\mathrm{x}_{ik} = 1$ o -1 y $\mathrm{x}_{i'k} = 0$, $d_k = 1/2 = 0.5$, es decir, el genotipo heterocigoto y un homocigoto difieren en la mitad de los alelos. Por último, si se comparan dos genotipos iguales, $d_k = 0$, con lo cual los individuos

no difieren en ningún alelo para ese marcador. Es importante notar que $S_{i,i'} = 1$, indica completa similaridad mientras que $S_{i,i'} = 0$ sería el extremo en que las líneas no tienen ningún alelo común.

La efectividad de la SG depende fuertemente de la densidad de marcadores requerida para lograr las predicciones deseadas y del costo de obtener genotipos con esa densidad de MM (Riedelsheimer y Melchinger, 2013). Los paneles de MM de alta (disponibles para las líneas parentales) y baja densidad (disponibles para las poblaciones) fueron combinados aplicando un método de imputación. Para cada una de las poblaciones se mide el número de SNPs con que se evaluó genotípicamente cada población, el número de MM final luego de la imputación realizada por Monsanto y se calcula la distribución de MM por cromosoma. Además, se calculó el coeficiente de variación (CV) del número de MM por cromosoma para estudiar la variabilidad en el número de MM por cromosoma.

El método de imputación para los datos de MM fue elegido por Monsanto y se trata del método de regresión de Haley y Knott (1991) basado en la información del mapa de ligamiento y de los marcadores disponibles. Si bien la descripción del método en si mismo excede a los objetivos del presente trabajo, a continuación se presenta un ejemplo que dan los autores en su publicación para comprender los principales conceptos detrás de la imputación.

Se asume que se desea imputar un marcador Z que se encuentra localizado entre dos marcadores A y B codominantes, es decir, se pueden distinguir los genotipos heterocigotos. Además, se asume que se desea realizar tal imputación en la generación F2 de una cruza de dos líneas parentales que llevan alelos diferentes para los tres marcadores cuyos genotipos pueden simbolizarse: $A_1A_1Z_1Z_1B_1B_1$ para el parental 1 y $A_2A_2Z_2Z_2B_2B_2$ para el parental 2.

La frecuencia de recombinación entre A y Z se simboliza Fr_A, entre Z y B, Fr_B y finalmente, entre A y B: *Fr*. La frecuencia de recombinación puede calcularse con base en la distancia que

existe entre los marcadores utilizando el mapa de ligamiento y la función de mapeo de Haldane (1919) que convierte las distancias medidas en cM a frecuencia.

Las frecuencias esperadas de cada genotipo pueden derivarse con base en las frecuencias de recombinación. El gameto $A_1Z_1B_1$ tiene una frecuencia esperada de $(1- Fr_A)(1- Fr_B)/2$ (los marcadores A y Z no recombinan y tampoco lo hacen los marcadores Z y B) mientras que el gameto $A_1Z_2B_1$ tiene una frecuencia esperada de $Fr_A Fr_B/2$ (recombinan A con Z y Z con B). En términos de genotipos, la frecuencia esperada del genotipo homocigoto $A_1A_1B_1B_1$ en la F2 es de $0.25(1- Fr)^2$ y la frecuencia esperada de los tres genotipos posibles para el marcador Z cuando los marcadores flanqueantes tienen ese genotipo $A_1A_1B_1B_1$ son: $0.25*(1- Fr_A)^2 (1- Fr_B)^2$, $0.25*2*(1- Fr_A)(1- Fr_B) Fr_{AF}Fr_B$ y $0.25* Fr_A^2 Fr_B^2$ para los genotipos Z_1Z_1 , Z_1Z_2 y Z_2Z_2, respectivamente. Así, el genotipo más probable será el valor con que se impute el marcador Z.

4.3. Modelos de selección genómica

En esta sección el foco es investigar diferentes estrategias para desarrollar el modelo de SG que incorpore los MM para predecir el valor genético (g) de los individuos. En el marco de la SG se asume que hay múltiples QTL que afectan al carácter y no se persigue el objetivo de identificarlos. En los modelos que se presentan a continuación se habla, en términos generales, de individuos pero al aplicarlos a los materiales de este trabajo se debe notar que se trata en realidad de familias de una población F3.

Generalmente, por razones computacionales no se estiman simultáneamente los efectos ambientales y de genotipo usando los MM sobre un carácter. Por lo tanto, el desarrollo del modelo de SG se realiza en dos etapas. La primera consta del análisis de datos fenotípicos detallado en el Apartado 4.1, ajustando el modelo de la Ecuación 4.1.1, donde se consideran los efectos propios del ambiente. Luego, como resultado de ese análisis se toman las predicciones (BLUP) de los valores genéticos ($\hat{g}$) para ser utilizados como variable respuesta de modelos de SG que consideran los MM. Por simplicidad, las predicciones de los valores genéticos ($\hat{g}$) en el rol de variable respuesta del modelo de SG, son referenciadas simplemente como valor observado y se simbolizan con y.

Partiendo del caso más simple, se modela la relación causal entre el valor genético y un único MM a través de un modelo lineal de regresión simple. Específicamente, se asume que se tienen *n* individuos para los cuales se dispone del genotipo de un marcador bialélico. Para un individuo *i*-ésimo ($i = 1, 2, \dots, n$), se denota como (y_i, x_{1i}) al par de datos valor genético y genotipo del marcador respectivamente. Además, dado que se trata de un marcador bialélico, se supone que los genotipos son codificados en -1, 0 y 1 (siendo 0 el genotipo heterocigoto) y por lo tanto el modelo queda expresado como a continuación:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i \qquad (4.3.1)$$

donde β_0, β_1 son los parámetros a estimar correspondientes a un término constante y al coeficiente de regresión del MM y ε_i es un término de error que usualmente se asume que tiene una distribución normal con media 0 y varianza σ^2.

Una forma clásica de estimar los parámetros de este modelo es a través del método de mínimos cuadrados. El mismo determina las estimaciones de los parámetros de manera que la suma de los residuos al cuadrado sea mínima, es decir:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i}\right)^2 \qquad (4.3.2)$$

El modelo de la Ecuación 4.3.1 corresponde a un modelo para un único marcador, sin embargo, se sabe que los caracteres cuantitativos dependen de múltiples QTL y se suele utilizar un gran número de MM. Así, el modelo puede extenderse para incluir *p* marcadores de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i = \sum_{k=0}^{p} \beta_k x_{ki} + \varepsilon_i = \mathbf{x'_i}\boldsymbol{\beta} + \varepsilon_i \qquad (4.3.3)$$

donde $\mathbf{x'}_i = \left(1, x_{i1}, x_{i2}, \ldots, x_{ip}\right)$ es el vector de variables explicativas que incluye un 1 para incorporar una constante al modelo y los valores x_{ik} para $k = 1, \ldots, p$ reflejan los genotipos para los MM (codificados como -1, 0 o 1) correspondientes al individuo *i*-ésimo, y $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \beta_2, \ldots, \beta_p\right)'$ es el vector de coeficientes que contiene la constante del modelo (β_0) y los coeficientes de regresión de cada marcador ($\beta_k, k = 1, \ldots, p$). En forma matricial, el modelo puede expresarse como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (4.3.4)$$

donde $\mathbf{y} = (y_1, y_2, \ldots, y_n)'$ es el vector correspondiente a la variable respuesta (específicamente, en los análisis de este trabajo se trata de las predicciones de los valores genéticos $y_i = \hat{g}_i$), $\mathbf{X} = \{x_{ik}\}_{n\times(p+1)}$ es la matriz de variables explicativas y $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ es el vector de términos de error.

En este caso, los parámetros también pueden estimarse a través de mínimos cuadrados, es decir, obteniendo estimaciones de los parámetros del modelo de la Ecuación 4.3.3 que logren minimizar la suma de los cuadrados de las diferencias entre el valor observado y el esperado, es decir:

$$\widehat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \arg\min \sum_{i=1}^{n} \left(y_i - \mathbf{x'_i}\,\widehat{\boldsymbol{\beta}}\right)^2 = \arg\min \left\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right\|^2 \quad (4.3.5)$$

siendo $\|\cdot\|$ la norma del vector que viene dada por la expresión $\|\mathbf{z}\| = \sqrt{\sum z_i^2}$. Sobre la base Ecuación 4.3.4 del modelo de regresión, si existe $(\mathbf{X'X})^{-1}$, la estimación del vector de parámetros queda expresada como:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'y} \quad (4.3.6)$$

Se puede demostrar que la varianza del estimador resulta igual a $\mathrm{Var}\left(\widehat{\boldsymbol{\beta}}\right) = (\mathbf{X'X})^{-1}\sigma^2$. Por lo tanto, la varianza del estimador de un coeficiente de regresión tiene la forma $Var\left(\hat{\beta}_k\right) = C_{kk}\sigma_\varepsilon^2, k = 0, 1, \ldots p$, siendo C_{kk} el *k*-ésimo elemento diagonal de la matriz $\mathbf{C} = (\mathbf{X'X})^{-1}$ de dimensión (*p+1)* x (*p+1)*. Es decir, la varianza se encuentra afectada por el número de MM (*p*) y el número de individuos (*n*). La varianza del estimador de mínimos cuadrados aumenta rápidamente a medida que lo hace el número de variables explicativas en un modelo donde el número de observaciones (*n*) permanece constante.

Ahora bien, uno de los supuestos realizados en la estimación es que existe la matriz inversa de $\boldsymbol{X'X}$ lo cual no siempre es cierto. Hay dos hechos que afectan su existencia, la dimensionalidad de los datos y la multicolinealidad de las variables explicativas. En el primero de los casos, con dimensionalidad se hace referencia a la dimensión de la matriz $\boldsymbol{X}_{\boldsymbol{n}\times(\boldsymbol{p}+\mathbf{1})}$. Si $p \geq n$ la matriz $\boldsymbol{X'X}$ no es invertible y como consecuencia, no existe una única solución para vector de coeficientes de regresión.

Por otro lado, si las columnas de $\boldsymbol{X}$ no son linealmente independientes (caso extremo de multicolinealidad), la matriz $\boldsymbol{X}$ no es de rango completo en las columnas y por lo tanto tampoco puede invertirse la matriz $\boldsymbol{X}'\boldsymbol{X}$. En estos casos, dada $(\boldsymbol{X}'\boldsymbol{X})^{-}$ inversa generalizada de $\boldsymbol{X}'\boldsymbol{X}$, es posible encontrar una solución para los coeficientes y, aunque esta solución no es única, las predicciones $\widehat{\boldsymbol{y}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ sí lo son por la propiedad de invarianza de $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ (Searle, 1992).

En el marco de los modelos de SG, los datos utilizados se caracterizan por tener un gran número de marcadores, que superan ampliamente a la cantidad de individuos (*p >> n*) y a su vez, estos marcadores se encuentran asociados entre sí por la existencia de desequilibrio de ligamiento (Lande y Thompson, 1990). Por lo tanto, ambos problemas de dimensionalidad y multicolinealidad están presentes en los datos de SG y la implementación de métodos estadísticos que puedan confrontarlos es una necesidad central para los programas de mejoramiento asistidos por MM.

En este trabajo se presentan tres estrategias para confrontar los problemas de dimensionalidad de la matriz $\mathbf{X}$ y la de presencia de multicolinealidad en el modelo. La primera estrategia consiste en un enfoque clásico de ajuste de modelos de regresión: selección de variables y estimación por mínimos cuadrados ordinarios. La segunda estrategia se sustenta en los desarrollos de los métodos de estimación penalizada propuestos en el marco de la SG. La última estrategia, es una combinación las dos anteriores y por lo tanto no solo selecciona MM sino que además penaliza sus coeficientes. Las metodologías se desarrollan a continuación en los Apartados 4.3.1 a 4.3.3 y asumen que la variable respuesta se encuentra centrada, es decir, no se incluirá el término constante β_0. Por lo tanto, en la Ecuación 4.3.3 el vector de coeficientes resulta $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ y el vector de variables explicativas equivale a $\mathbf{x}'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$.

4.3.1. Selección de variables y ajuste por Mínimos Cuadrados

Generalmente, para los modelos de regresión estimados vía mínimos cuadrados, idealmente se ajustarían todos los modelos posibles con p = 1, 2, ... , n-1 variables explicativas, es decir, con un MM, con dos MM, y así sucesivamente hasta n-1 MM. Luego, se seleccionaría aquel modelo con mejor ajuste de acuerdo a algún criterio de bondad de ajuste (por ejemplo, criterio de Akaike, 1973). Sin embargo, cuando p es grande, ajustar todos los modelos posibles no es factible y se debe recurrir a algún otro procedimiento de selección de MM (Bernardo y Yu, 2007).

En el presente trabajo, se aplicó uno de los algoritmos más sencillos para ajustar el modelo por mínimos cuadrados ordinarios. El mismo consiste en estimar el modelo de regresión para cada marcador individual, es decir, se estima por mínimos cuadrados el modelo de la Ecuación 4.3.1 para cada MM. En el ajuste de cada modelo, se obtiene una medida de asociación entre el MM y la variable respuesta. En particular, se considera la probabilidad asociada a la prueba de hipótesis sobre el coeficiente del MM. El gráfico del logaritmo en base 10 de esa probabilidad asociada para cada marcador ordenado según su posición en el mapa de ligamiento permite, además, identificar posibles regiones del genoma asociadas al carácter.

Una vez realizado el análisis por MM, se llevan a cabo sucesivos ajustes de modelos incrementando el número de MM. Se considera en primer lugar el marcador más asociado al carácter fenotípico, luego se contemplan los dos marcadores más asociados y así sucesivamente, hasta llegar a un máximo número de MM en el modelo equivalente a n - 1.

A este procedimiento se lo denomina en adelante **SMC** (Selección de variables y ajuste por Mínimos Cuadrados). Para cada uno de estos sucesivos ajustes a medida que se incrementa el número de marcadores en el modelo, se obtiene la correlación entre el valor observado y el predicho por el modelo. Esta estadística es la herramienta a utilizar para seleccionar el mejor modelo de este enfoque.

Una de las ventajas del método SMC es que se trata de una metodología estadística clásica y por lo tanto puede aplicarse en cualquier paquete estadístico. Además, el trabajo computacional es inferior comparado con otras métodos de estimación que puedan involucrar algoritmos iterativos. Sin embargo, a pesar de haber realizado selección de variables para poder superar el problema de dimensionalidad y descartar marcadores duplicados, es factible que exista alto grado de multicolinealidad entre los marcadores presentes en el modelo (Whittaker *et al.*, 2000). Un modelo sobreestimado puede exagerar fluctuaciones mínimas en los datos y generalmente tiene una pobre habilidad predictiva (Lorenz *et al.,* 2011).

4.3.2. Estimación Penalizada: regresión de Ridge

Los problemas de dimensionalidad y multicolinealidad son abordados por varios métodos estadísticos que básicamente evitan el proceso de selección de variables manteniendo todos los MM en el modelo (Piepho, 2009). Uno de esos métodos es la regresión de Ridge, introducida por Hoerl y Kennard (1970). La misma fue aplicada por primera vez en el marco de la SG en el trabajo de Whittaker *et al.* (2000).

La regresión de Ridge pretende balancear la bondad de ajuste del modelo con la complejidad del mismo. Con tal objetivo, se define como una medida de la complejidad del modelo a la suma de los cuadrados de los coeficientes de los MM, es decir, $\sum_{k=1}^{p} \beta_k^2$ y esta complejidad se penaliza a través de una constante. Específicamente, siendo $\lambda \geq 0$ el parámetro de penalización (o regularización) que controla la compensación entre la falta de ajuste y la complejidad del modelo, el estimador tiene como objetivo minimizar la suma de cuadrados penalizada dada por la siguiente expresión:

$$\sum_{i=1}^{n}(y_i - \mathbf{x'_i}\,\boldsymbol{\beta})^2 + \lambda \sum_{k=1}^{p} \beta_k^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2 \qquad (4.3.7)$$

Luego, siendo $\mathbf{I}$ una matriz identidad de dimensiones $p \times p$, se puede demostrar que el estimador de la regresión de Ridge resulta:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \qquad (4.3.8)$$

El término de penalización es una estrategia que permite superar el problema de multicolinealidad entre columnas de la matriz $\boldsymbol{X}$ que causa la no existencia de la matriz inversa de $\boldsymbol{X}'\boldsymbol{X}$. Este estimador penalizado involucra "encogimiento", por consiguiente evita sobreajuste del modelo y estabiliza las estimaciones en comparación con la regresión por mínimos cuadrados (Piepho, 2009).

El parámetro de penalización λ que determina la magnitud del encogimiento puede elegirse siguiendo distintos criterios (Ruppert *et al.*, 2003). A continuación se presentan tres alternativas: **Regresión de Ridge clásica (RR)**, **Regresión de Ridge BLUP (RR-BLUP)** y **Regresión de Ridge Bayesiana (BRR)**.

Regresión de Ridge Clásica (RR)

Este enfoque implica como primer paso, determinar cuál es el valor óptimo del parámetro de penalización λ. Para elegirlo se emplea el método de validación cruzada de 10 iteraciones o submuestras. El mismo consiste en dividir el conjunto de datos total con el cual se estima el modelo en 10 submuestras. El modelo se ajusta 10 veces, excluyendo las submuestras de a una por vez; la submuestra excluida se utiliza como conjunto de validación mientras que las restantes 9 se utilizan para la estimación del modelo. En cada iteración se calcula el **Cuadrado Medio del Error de Predicción**: $CMEP = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_{-i})^2$, siendo y_i la observación para el individuo *i*-ésimo de la submuestra de validación, $\hat{y}_{-i}$ la predicción del individuo *i*-ésimo cuando la estimación del modelo lo excluía y *N* el total de individuos en la submuestra utilizada para la

validación. El valor óptimo de λ es aquél que minimiza el promedio de los 10 valores de CMEP. Para llevar a cabo esta búsqueda, se utilizó la función **cv.glmnet** de la librería **glmnet** (Friedman *et al.*, 2010). Una vez encontrado el parámetro de penalización óptimo, se procedió a la estimación de los coeficientes del modelo, utilizando la función **glmnet**.

Regresión de Ridge BLUP (RR-BLUP)

Meuwissen *et al.* (2001) emplearon otra forma de estimar un valor óptimo del parámetro de penalización en el marco de la SG. Este enfoque asume que los coeficientes de los MM son aleatorios, con distribución normal de media 0 y varianza común σ_β^2. Se puede demostrar que el parámetro de regularización λ en la Ecuación 4.3.8 del estimador penalizado equivale al cociente de varianzas: $\sigma_\varepsilon^2/\sigma_\beta^2$ (Ruppert *et al.*, 2003).

A esta metodología se la denomina RR-BLUP y su formulación a través de un modelo mixto tiene como ventaja que se pueden estimar de manera directa no sólo las componentes de varianza sino también, el parámetro de penalización usando REML.

Además, RR-BLUP permite el ajuste de un modelo más complejo que el planteado en la Ecuación 4.3.3 ya que en un modelo mixto se pueden contemplar otras fuentes de variación agregando efectos ya sean fijos o aleatorios.

Para la estimación del modelo RR-BLUP, se empleó el método REML a través de la función **mixed.solve** de la librería de R **rrBLUP** (Endelman, 2011).

Regresión de Ridge Bayesiana (BRR)

La mayoría de los métodos penalizados son equivalentes a modelos bayesianos que asumen cierta distribución *a priori* de los parámetros del modelo (Tibshirani, 1996). Por lo tanto, hay un

análisis bayesiano equivalente a la RR que se denomina Regresión de Ridge Bayesiana (BRR) (de los Campos *et al.*, 2013).

En la estadística bayesiana todos los valores desconocidos (parámetros, efectos aleatorios, entre otros) son tratados como variables aleatorias, reflejando la incertidumbre respecto de esos valores en una distribución de probabilidad (Sorensen y Gianola, 2002). Todo aquel conocimiento que se tenga sobre los parámetros antes de observar los datos es representado en términos de una **distribución de probabilidad *a priori***, representada como $p(\boldsymbol{\beta}|\boldsymbol{\omega})$, siendo $\boldsymbol{\omega}$ valores desconocidos de la distribución que se suelen denominar **hiper-parámetros**. Si por ejemplo, un parámetro puede tomar cualquier valor entre 0 y 1, pero no se tiene más información al respecto, la distribución uniforme entre 0 y 1 podría ser una propuesta para distribución *a priori*, siendo 0 y 1 los correspondientes hiper-parámetros. Sin embargo, si se supiera que es más probable que ese parámetro sea próximo a 0.5, se puede asumir una distribución a *priori* beta con ambos hiper-parámetros iguales a 2. La distribución beta es más informativa respecto del parámetro que la distribución uniforme.

Además, toda la información proveniente de los datos se ve reflejada en la verosimilitud, $p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\omega}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta},\boldsymbol{\omega})$. El concepto detrás de la estimación bayesiana es combinar el conocimiento *a priori* con la información proveniente de los datos, para obtener la **distribución de probabilidad *a posteriori,*** $p(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{\omega})$, desde donde se realizan las inferencias.

En particular, la BRR plantea el modelo en dos niveles, el primero es al nivel de los datos y se basa en el modelo de la Ecuación 4.3.3 pero asumiendo además que los coeficientes de los MM provienen de una distribución normal con media 0 y varianza σ_β^2, es decir, $p(\boldsymbol{\beta}|\boldsymbol{\omega}) = \prod_{k=1}^{p} N(\beta_k|0,\sigma_\beta^2)$, donde $N(\beta_k|0,\sigma_\beta^2)$ indica la función de densidad normal de la variable β_k con los hiper-parámetros de la media y varianza iguales a 0 y σ_β^2, respectivamente. El segundo nivel del modelo BRR se refiere a las varianzas de los coeficientes de los MM y de los errores aleatorios;

se asume que las varianzas tienen distribución *a priori* Chi-cuadrado invertida escalada con grados de libertad (gl_β, gl_ε) y parámetros de escala (S_ε, S_β). A estos últimos parámetros también se los conoce con el nombre de hiper-parámetros pero de mayor grado (de los Campos, 2012).

La función **BGLR** de la librería **BGLR** (de los Campos y Perez-Rodríguez, 2014) se utiliza para aplicar el método BRR con los hiper-parámetros (gl y parámetros de escala de las distribuciones chi-cuadrado invertidas escaladas) definidos por defecto. La función BGLR emplea reglas para asignar los valores de hiper-parámetros cuando estos no son provistos por el usuario; estas reglas permiten obtener valores apropiados aunque no siempre resultan muy informativos (de los Campos y Perez-Rodríguez, 2014).

Las inferencias en el marco de análisis bayesiano se basan en la distribución *a posteriori* de los parámetros dados los datos $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{\omega})$. Generalmente, no es posible obtener una expresión analítica para dicha distribución *a posteriori* por lo tanto se emplea algún algoritmo de aproximación. Tanto la función **BLR** de la librería con el mismo nombre (de los Campos y Perez-Rodriguez, 2012) como su nueva versión BGLR utilizan el algoritmo de Monte Carlo vía Cadenas de Markov (MCMC del inglés Markov Chain Monte Carlo) denominado *Gibbs Sampler* (Geman y Geman, 1984; Casella y George, 1992) que muestrea repetidamente y calcula estadísticas resúmenes de las distribuciones *a posteriori* lo cual lo convierte en un método muy demandante computacionalmente.

Para utilizar la función BGLR se realizan 150.000 iteraciones para cada una de las poblaciones y caracteres fenotípicos, descartando las primeras 50.000 iteraciones (valor conocido en el área bayesiana como *"burn-in"*) por posibles problemas de ruido en las estimaciones iniciales. Además, se guardan las estimaciones de los parámetros cada 10 iteraciones (valor conocido como *"thin"*). Dichas estimaciones se utilizan para estudiar la convergencia del algoritmo a través de los

llamados **gráficos de convergencia** que consisten en representar las sucesivas estimaciones versus el número de iteración.

4.3.3. Selección de variables y estimación penalizada: Regresión LASSO

La SG generalmente se basa en un gran número de MM no obstante, es probable que muchos de ellos estén localizados en regiones que no afecten al carácter de interés. Por otro lado, hay marcadores que están en desequilibrio de ligamiento con los QTL o en regiones que limitan con genes involucrados. Esto sugiere que el modelo de la Ecuación 4.3.3 debería ofrecer diferentes niveles de "encogimiento" a los coeficientes de los MM.

Tibshirani (1996) propuso el método de **Regresión LASSO** (del inglés *Least Absolute Shrinkage and Selection Operator*) que combina la selección de variables y la penalización de los parámetros. La regresión de Ridge utiliza un parámetro de penalización λ ligado a la función $J(\boldsymbol{\beta}) = \sum_{k=1}^{p} \beta_k^2 = \|\boldsymbol{\beta}\|^2$ que mide la complejidad del modelo. La estimación LASSO propone la función $J(\boldsymbol{\beta}) = \sum_{k=1}^{p} |\beta_k|$ para reflejar la complejidad del modelo y el estimador de $\boldsymbol{\beta}$ viene dado por la siguiente expresión:

$$\widehat{\boldsymbol{\beta}} = \arg\min \left[\sum_{i=1}^{n} (y_i - \mathbf{x}'_i \boldsymbol{\beta})^2 + \lambda \sum_{k=1}^{p} |\beta_k| \right] \qquad (4.3.9)$$

El parámetro λ se denomina parámetro de regularización que controla el impacto de la penalización, cuanto más próximo a cero es, más se parece a la estimación por mínimos cuadrados ordinarios mientras que cuanto mayor es λ, mayor es la penalización. Sin embargo, no debe compararse con el parámetro de penalización en la RR, pues, la función $J(\boldsymbol{\beta})$ es diferente.

Ahora bien, el ajuste del modelo LASSO se lleva a cabo desde dos alternativas: **Regresión LASSO clásica (LR)** y **Regresión LASSO bayesiana (BLR)**; ambas se describen a continuación.

Regresión LASSO clásica (LR)

La solución a la Ecuación 4.3.9 admite a lo sumo tantos coeficientes de regresión no nulos como *n*-1 (Park y Casella, 2008). Por ello, el enfoque clásico de la Regresión LASSO, es una combinación de la estrategia de selección de variables y de la estrategia de penalización. En este trabajo este enfoque clásico se implementó utilizando nuevamente la función **glmnet** que utiliza un método optimización de coordenadas cíclicas descendentes (Friedman *et al.*, 2010). La elección del parámetro de penalización se hizo a través de validación cruzada con 10 iteraciones utilizando la función **cv.glmnet** que busca minimizar el CMEP (también implementado en el caso de RR).

Regresión LASSO Bayesiana (BLR)

De la misma manera que para la regresión de Ridge, existe una contrapartida bayesiana para el método LASSO denominada Regresión LASSO Bayesiana (BLR). En el marco de la SG, *a priori*, no hay razón por la cual el número de marcadores con coeficientes no nulos deba ser menor que el número de individuos, hecho que ocurre en LR. Sin embargo, este problema no surge en el enfoque bayesiano (de los Campos *et al.*, 2009): La distribución *a priori* en el caso BLR es el producto de *p* densidades centradas doble-exponenciales independientes, es decir, $p(\boldsymbol{\beta}|\,\lambda^2/\sigma_\varepsilon^2) = \prod_{k=1}^{p}(\lambda^2/2\sigma_\varepsilon^2)\exp(-|\beta_k|\,\lambda^2/\sigma_\varepsilon^2)$.

A modo de ejemplo, se representan la distribución normal estándar, densidad utilizada como *priori* para BRR, y la distribución doble exponencial con parámetro de escala igual a 1, densidad a *priori* para BLR (Figura 4.3.1). Se puede observar que la distribución doble exponencial tiene mayor masa de probabilidad en el entorno del cero (es decir, hay mayor encogimiento de los coeficientes hacia el cero) y colas más pesadas que la distribución normal. No obstante, la

probabilidad de que los coeficientes sean iguales a cero es nula y por ello BLR se debe considerar como un método de penalización (no de selección de variables como es el enfoque clásico LR).

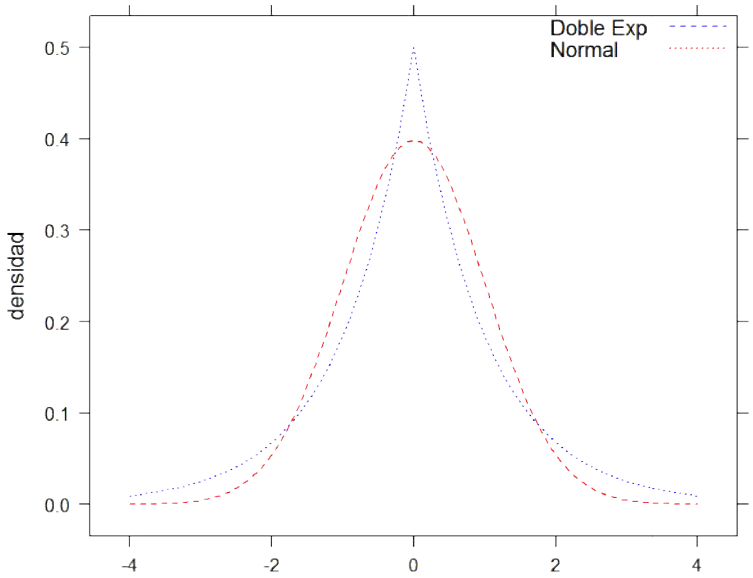

Figura 4.3.1 – Función de densidad normal (---) y doble exponencial (…) correspondientes a las distribuciones *a priori* de los coeficientes de los MM en las metodologías BRR y BLR.

Para la implementación de la BLR, se utilizó la función **BGLR** de la librería con el mismo nombre (de los Campos y Perez-Rodriguez, 2012) que tienen en cuenta una propiedad de las distribuciones de colas pesadas. Estas distribuciones pueden ser representadas como mezcla de densidades normales escaladas de la forma $p(\beta_k|\boldsymbol{\omega}) = \int N\left(\beta_k\middle|0,\sigma^2_{\beta_k}\right) p\left(\sigma^2_{\beta_k}\middle|\boldsymbol{\omega}\right)\partial\sigma^2_{\beta_k}$ donde $p\left(\sigma^2_{\beta_k}\middle|\boldsymbol{\omega}\right)$ es la distribución *a priori* de la varianza específica de cada marcador y $\boldsymbol{\omega}$ son los hiper-parámetros de esta distribución. En particular, cuando $p\left(\sigma^2_{\beta_k}\middle|\lambda^2/2\right)$ es la densidad exponencial, la distribución marginal de los coeficientes es la distribución doble exponencial (de los Campos *et al.*, 2013).

Un desafío de la BLR es la elección del valor λ, para lo cual hay diferentes enfoques. Uno de ellos es elegir un valor fijo basado y el otro es darle un enfoque completamente bayesiano, es

decir, asignar una distribución *a priori* al parámetro λ. Dado que, el enfoque bayesiano presentó reiteradas veces problemas de convergencia, en el presente trabajo se aplica únicamente el primer enfoque. Siguiendo los métodos para selección de hiper-parámetros propuestos por (Pérez *et al.*, 2010), el valor de λ se determina con base a la heredabilidad como: $\lambda = \sqrt{2h^2 \sum_{k=1}^{p} \bar{x}_k^2}$ donde h^2 es la heredabilidad que en el presente trabajo se estimó siguiendo la Ecuación 4.1.1 y $\sum_{k=1}^{p} \bar{x}_k^2 = \sum_{k=1}^{p} \left(\frac{1}{n}\sum_{i=1}^{n} x_{ik}\right)^2$.

A modo de resumen de todo lo desarrollado en el Apartado 4.3, la Tabla 4.3.1 describe de manera concisa todos los métodos de SG que se emplean en este trabajo. Notar que el parámetro de penalización es llamado de manera general λ, pero no son comparables ya que, difiere la función que refleja la complejidad del modelo por ejemplo, la suma de valores absolutos de los coeficientes en el caso de estimación LASSO o suma de los coeficientes al cuadrado en estimación RR.

Tabla 4.3.1 – Resumen de los métodos de SG estudiados

Método de SG	Estrategia incorporación de MM en el modelo	Método Estadístico de Estimación	Tratamiento coeficientes de MM	Tratamiento de parámetros desconocidos
SMC	Selección de variables	Mínimos Cuadrados Ordinarios	Fijos	Estimación clásica de $\boldsymbol{\sigma_\varepsilon^2}$
RR	Penalización enfoque clásico	Mínimos Cuadrados Penalizados	Fijos	Estimación clásica de $\boldsymbol{\sigma_\varepsilon^2}$ $\boldsymbol{\lambda}$: valor (dentro de un rango) que minimice una medida de bondad de ajuste
RR-BLUP	Penalización enfoque modelos mixtos	Máxima Verosimilitud Restringida	Aleatorios	Estimación de $\boldsymbol{\sigma_\varepsilon^2}$ y $\boldsymbol{\sigma_\beta^2}$ por REML $(\boldsymbol{\lambda = \sigma_\varepsilon^2/\sigma_\beta^2})$
BRR	Penalización enfoque bayesiano	Algoritmo Bayesiano *Gibbs Sampler*	*Priori* N(0, $\boldsymbol{\sigma_\beta^2}$)	$\boldsymbol{\sigma_\varepsilon^2 \sim \chi^{-2}(gl_\varepsilon, S_\varepsilon)}$ $\boldsymbol{\sigma_\beta^2 \sim \chi^{-2}(gl_\beta, S_\beta)}$
BLR	Penalización enfoque bayesiano	Algoritmo Bayesiano *Gibbs Sampler*	*Priori* DE($\boldsymbol{\lambda^2}$)	$\boldsymbol{\lambda^2}$: valor definido por el usuario
LR	Selección de variables y penalización	Mínimos Cuadrados Penalizados	Fijos	Estimación clásica de $\boldsymbol{\sigma_\varepsilon^2}$

4.3.4. Evaluación de la habilidad predictiva de los modelos

Con los fines de evaluar la habilidad predictiva de los distintos métodos de SG se realiza validación cruzada, definiendo un conjunto de datos de entrenamiento que permite la estimación de los parámetros del modelo y otro de validación con el cual se contrastan los valores predichos y los observados.

El conjunto de entrenamiento debe ser representativo de la población que se desea evaluar para obtener buenas predicciones. En teoría, la mejor población de entrenamiento para una población es un subconjunto de la misma población para la cual se dispone de información de los caracteres fenotípicos y de los MM. Por lo tanto, cada población F3 se dividió al azar en dos conjuntos: uno de entrenamiento (75%) y otro de validación (25%). El conjunto de entrenamiento fue utilizado para estimar el modelo de SG siguiendo las metodologías desarrolladas en los Apartados 4.3.1 a 4.3.3. El conjunto de validación quedó conformado por individuos que no participaron del proceso de estimación del modelo y por lo tanto se consideraron como individuos nuevos.

La habilidad predictiva de los distintos métodos de SG se midió a través de la correlación entre el valor observado y el valor predicho para el conjunto de validación en cada una de las 20 poblaciones y los tres caracteres: rendimiento, humedad y peso hectolítrico.

Con los fines de estudiar qué factores influyeron en la habilidad predictiva, se empleó una estrategia similar a la aplicada en el trabajo de Asoro *et al.* (2011). Se ajustó un modelo donde la variable respuesta fue la correlación entre el valor observado y el valor predicho por el modelo de SG pero dentro del conjunto de validación y como variables explicativas se consideraron: grupo heterótico (GH1, GH2), método de SG aplicado (SMC, RR, RR-BLUP, BRR, BLR, LR) y carácter (Humedad, Peso Hectolítrico, Rendimiento). Además, se incorpora un efecto aleatorio por población. Es decir, el modelo lineal mixto planteado resulta:

$$r_{tuvw} = \mu + G_t + M_u + F_v + (GM)_{tu} + (GF)_{tv} + (MF)_{uv} + (GMF)_{tuv} + \delta_w + \varepsilon_{tuvw} \qquad (4.3.10)$$

donde r_{tuvw} es la habilidad predictiva (correlación) en el GH *t*-ésimo ($t = 1,2$), correspondiente al método de SG *u*-ésimo ($u = 1,2, \dots, 6$), del carácter fenotípico *v*-ésimo ($v = 1,2,3$), en la población número *w* ($w = 1,2, \dots 20$); μ es la media general de la habilidad predictiva, G_t representa el efecto del GH, M_u representa el efecto del método de selección gnómica empleado, F_v es el efecto debido al carácter fenotípico analizado, luego se incorporan las interacciones dobles de los efectos principales, la interacción de tercer orden y el efecto aleatorio de la población (δ_w) y un error aleatorio ε_{tuvw}. En el caso de encontrar algún factor significativo, se procede a realizar comparaciones múltiples de las medias de los factores en cuestión utilizando el método de Tukey (1949). Se trabaja con un nivel de significación $\alpha = 0.05$.

Se reconoce que el supuesto de independencia y normalidad de los residuales del modelo modelo lineal mixto en la Ecuación 4.3.10 puede ser violado por lo tanto, las probabilidades asociadas no resultan exactas bajo la hipótesis nula. Sin embargo, el propósito de este ajuste no fue probar específicamente las magnitudes de cada uno de los efectos bajo estudio sino que el objetivo es dar una aproximación que ayude a cuantificar los factores que pueden estar afectando a la habilidad predictiva de los modelos de SG.

5. RESULTADOS

5.1. Resultados del análisis de datos fenotípicos

Los resultados del análisis de datos fenotípicos de las 20 poblaciones de maíz para los tres caracteres (Humedad, Peso Hectolítrico y Rendimiento) se resumen en la Tabla 5.1.1. En la misma, se presenta para cada población F3, el número de familias y por cada carácter: número de localidades, media, R2 del modelo mixto de la Ecuación 4.1.1 y heredabilidad.

Se observó variabilidad entre las poblaciones tanto en medidas propias de los caracteres como en el número de familias. El número de familias por población en el GH1 varió entre 90 y 186 familias mientras que, en el GH2 varió entre 73 y 184 familias.

La humedad del grano a cosecha se midió entre 5 y 8 localidades por población. Los valores medios de las poblaciones variaron entre 14,5% y 24,5%. En cuanto al ajuste del modelo, el R^2 fue superior a 80% en todas las poblaciones. Finalmente, la heredabilidad varió entre 54,1% y 90,9%.

El peso hectolítrico se midió entre 5 y 8 localidades por población. Los valores medios de las poblaciones variaron entre 53,2 lb/bu y 59,3 lb/bu. En cuanto al ajuste del modelo, el R^2 varió entre 39% y 97%. Finalmente, la heredabilidad varió entre 49,9% a 81,2%.

El rendimiento en grano se midió entre 5 y 8 localidades por población. Los valores medios de las poblaciones variaron entre 176,9 y 220,0 bu/ac. En cuanto al ajuste del modelo, el R^2 varió entre 32% y 93%. Finalmente, la heredabilidad varió entre 40,0% a 71,3%.

Para todas las poblaciones y caracteres fenotípicos, se predijeron los valores genéticos de las familias que se emplearon luego como variable respuesta en los modelos de SG siguiendo la Ecuación 4.3.3.

Tabla 5.1.1 – Número de familias, número de localidades, media, R2 y heredabilidad para los caracteres fenotípicos: humedad, peso hectolítrico y rendimiento en las 20 poblaciones F3 de maíz.

Grupo	Población	No. Familias	Humedad (%)				Peso Hectolítrico (lb/bu)				Rendimiento en Grano (bu/ac)			
			# Loc	Media	R^2	h^2	# Loc	Media	R^2	h^2	# Loc	Media	R^2	h^2
GH1	1	181	8,0	18,2	96%	77,8	5,0	59,3	76%	65,3	8,0	197,6	93%	53,0
GH1	2	134	7,0	23,0	90%	83,7	7,0	57,7	48%	66,7	7,0	179,1	82%	71,3
GH1	3	136	5,0	15,7	89%	65,0	5,0	59,1	86%	70,9	5,0	189,0	64%	54,8
GH1	4	143	8,0	20,5	95%	76,7	8,0	56,0	81%	66,3	8,0	214,8	73%	62,9
GH1	5	184	7,0	21,7	97%	77,1	6,0	56,0	82%	58,0	7,0	189,5	68%	53,9
GH1	6	90	6,0	17,2	94%	73,5	6,0	56,8	92%	79,7	6,0	184,8	85%	54,5
GH1	7	186	8,0	19,0	81%	79,3	7,0	57,7	84%	81,2	8,0	200,6	86%	59,3
GH1	8	179	8,0	22,8	84%	72,8	8,0	55,8	89%	79,0	8,0	189,7	59%	66,5
GH1	9	181	8,0	19,4	93%	84,7	8,0	56,7	91%	52,7	8,0	201,9	68%	40,0
GH1	10	184	7,0	17,7	89%	64,9	7,0	58,9	86%	51,6	7,0	185,5	60%	58,2
GH2	11	161	7,0	14,5	91%	65,6	6,0	55,4	86%	74,8	7,0	176,9	75%	55,3
GH2	12	132	7,0	22,8	88%	90,9	7,0	56,4	39%	54,4	7,0	209,5	32%	54,5
GH2	13	177	8,0	17,3	92%	54,1	8,0	57,0	78%	56,1	8,0	200,9	53%	53,0
GH2	14	178	7,0	24,5	91%	82,3	7,0	53,2	66%	64,0	7,0	199,0	82%	52,5
GH2	15	158	7,0	17,3	91%	78,7	7,0	56,9	62%	63,4	7,0	220,0	64%	47,4
GH2	16	157	7,0	16,2	93%	66,8	7,0	58,0	63%	55,2	7,0	210,8	67%	56,8
GH2	17	184	8,0	17,8	95%	75,3	8,0	57,9	92%	62,0	8,0	204,0	74%	41,8
GH2	18	159	7,0	17,3	96%	66,7	7,0	55,9	84%	64,6	7,0	191,9	80%	48,0
GH2	19	182	8,0	20,2	94%	71,6	6,0	56,3	90%	49,9	8,0	205,9	80%	63,1
GH2	20	73	5,0	21,4	96%	64,9	5,0	54,9	97%	69,0	5,0	219,3	82%	51,4

5.2. Resultados del análisis de los datos de marcadores moleculares

El análisis de la información de tipo molecular se llevó a cabo para cada una de las 20 poblaciones de maíz y los resultados se resumen en la Tabla 5.2.1. Notar que, para mantener la confidencialidad de los datos, los cromosomas fueron codificados aleatoriamente con letras.

El número de marcadores con que se evaluaron inicialmente las familias F3 de cada población varió entre 71 y 105 SNPs a lo largo de todo el genoma. En las poblaciones, se incrementó la densidad incorporando marcadores imputados y alcanzando un total de 2.886 MM. Sobre este conjunto total de MM, se realizaron los controles de calidad de datos moleculares. El número final de MM por población se incrementó a más de 5 veces el número inicial de SNPs evaluados. El número final de MM para el GH1 varió entre 848 y 1.178 a lo largo de todo el genoma mientras que, en las poblaciones del GH2, varió entre 512 y 872. Se calculó la distribución de MM a lo largo del genoma para cada población, es decir, se contabilizó el número de MM disponibles en cada uno de los cromosomas (A-J). La distribución del número de MM por cromosoma no resultó uniforme dentro de las poblaciones, por ejemplo, la población 17 presentó un CV superior al 5%: en un sólo cromosoma se tuvieron 15 MM mientras que en cinco cromosomas se tuvieron más de 100 MM. En general, los cromosomas con menor cantidad de MM por población resultaron los cromosomas B y E, mientras que los cromosomas de mayor cantidad de MM fueron: J, H e I. La similaridad de los parentales de cada una de las poblaciones varió entre 0,60 y 0,70 en el GH1 y entre 0,70 y 0,82 en el GH2.

Los MM finales, es decir, aquéllos evaluados más los imputados, luego de haber superado los criterios de control de calidad, fueron la base de los modelos de SG cuyos resultados se presentan en los dos apartados siguientes.

Tabla 5.2.1 – Número de marcadores genotipados y finales (luego de la imputación y control de calidad) por población, distribución de los marcadores finales a lo largo del genoma y similaridad entre las líneas parentales basada en SNPs.

GH	Población	No. SNPs genotipados	No. MM finales	No. MM finales por cromosoma										CV No. de MM finales por cromosoma (%)	Similaridad entre líneas parentales de la población
				A	B	C	D	E	F	G	H	I	J		
GH1	1	102	898	89	53	62	66	70	94	117	115	110	122	2,86	0,69
GH1	2	86	1116	97	68	117	84	67	109	144	132	133	165	2,94	0,61
GH1	3	93	848	50	76	31	85	27	115	90	120	92	162	4,93	0,70
GH1	4	96	1029	63	56	119	78	68	101	84	133	156	171	3,92	0,65
GH1	5	94	1031	100	82	108	60	58	104	117	106	132	164	3,09	0,64
GH1	6	98	1134	103	57	124	93	88	118	104	141	165	141	2,75	0,61
GH1	7	102	1178	95	91	146	80	98	128	83	125	140	192	3,00	0,60
GH1	8	105	1006	83	87	89	89	67	102	114	114	102	159	2,49	0,66
GH1	9	92	1053	66	72	115	94	28	112	121	139	133	173	3,96	0,64
GH1	10	93	1118	99	69	124	90	65	102	120	135	145	169	2,96	0,61
GH2	11	92	624	61	44	54	64	44	77	62	80	55	83	2,24	0,78
GH2	12	80	872	65	54	92	89	57	103	92	107	83	130	2,72	0,70
GH2	13	71	583	67	52	22	69	51	59	32	102	87	42	4,16	0,80
GH2	14	104	774	83	64	66	81	49	96	64	91	67	113	2,45	0,72
GH2	15	94	812	57	46	85	103	47	103	83	55	116	117	3,47	0,71
GH2	16	98	568	63	39	83	51	28	47	85	39	38	95	4,12	0,80
GH2	17	103	817	79	40	48	47	15	119	100	117	111	141	5,14	0,72
GH2	18	87	702	64	48	80	69	45	83	81	69	61	102	2,44	0,76
GH2	19	102	655	82	44	53	69	51	47	87	72	62	88	2,53	0,77
GH2	20	100	512	60	30	65	40	36	54	60	63	52	52	2,35	0,82

5.3. Aplicación de métodos de selección genómica a una población y carácter

En una población elegida al azar, se estimó el modelo de regresión de los valores genéticos para la humedad en términos de los MM, empleando los datos de 100 familias que conformaron el conjunto de entrenamiento. A continuación, se presentan los resultados de cada uno de los seis métodos de SG, se representan los coeficientes estimados por cromosoma y se estudia la habilidad predictiva del modelo a través de la estimación de la correlación entre los valores observados y predichos para el conjunto de validación, compuesto por otras 34 familias de la misma población.

5.3.1. Aplicación de Selección de variables y ajuste por Mínimos Cuadrados (SMC).

Se estimó un modelo de regresión por cada MM (Ecuación 4.3.1). Como consecuencia, se ajustaron tantos modelos como marcadores disponibles, en este caso, resultaron 1.116 modelos. De cada uno de esos ajustes por MM individual se tomó la probabilidad asociada (p-valor) de la prueba de hipótesis correspondiente al coeficiente del MM. A continuación, se representa la asociación entre cada MM y la variable respuesta a través del valor de $-\log_{10}$(p-valor) versus el orden de los MM dentro de cada cromosoma (Figura 5.3.1). Valores altos $-\log_{10}$(p-valor) ocurren cuando un MM resultó altamente significativo para el carácter en estudio. Así, esta información permitió identificar regiones del genoma que, se encontrarían asociadas a la humedad del grano. Por ejemplo, en el cromosoma B, los MM que están próximos a 100 (en el ordenamiento dentro de este cromosoma) muestran la mayor asociación con los valores genéticos correspondientes a la humedad de grano.

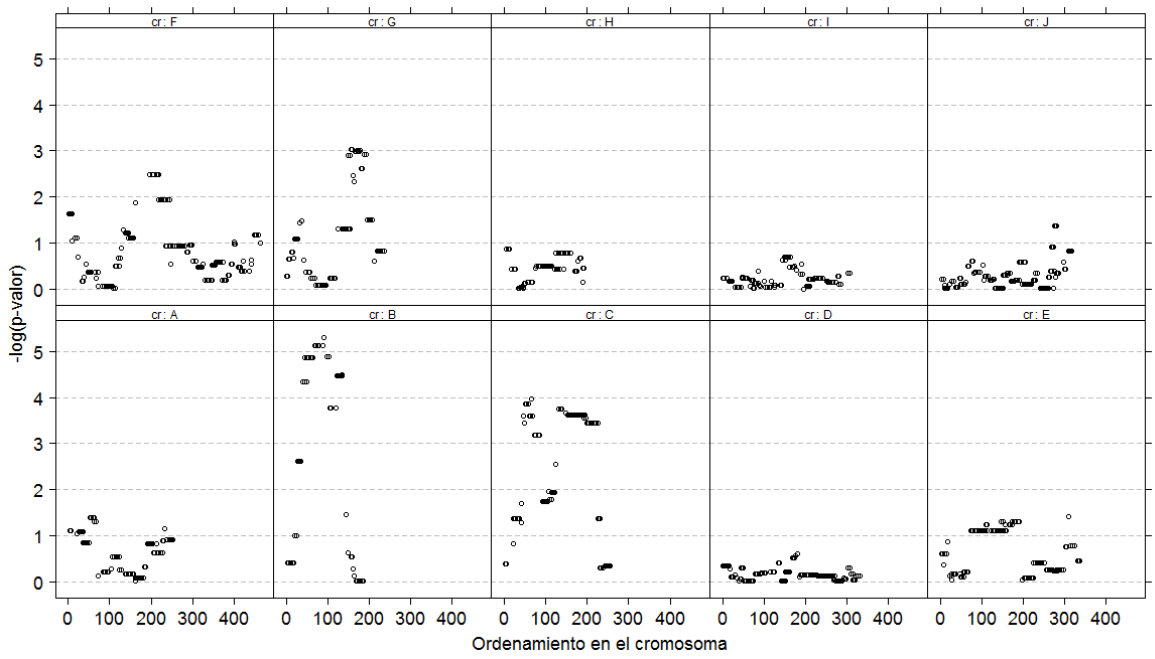

Figura 5.3.1 – Asociación entre cada MM y los valores genéticos para la humedad de una población, representada por el - logaritmo en base 10 de la probabilidad asociada de los coeficientes de MM versus el ordenamiento de los MM dentro de cada cromosoma con base en el mapa de ligamiento.

La selección de variables se basó en el ordenamiento de los MM según la probabilidad asociada (p-valor) obtenida para cada uno de sus coeficientes. Siendo que la posición 1 la ocupó el MM con menor p-valor, a continuación se representa un ejemplo del ordenamiento para un subconjunto de MM:

MM	MM.A	MM.B	MM.C	MM.D	MM.E	MM.F	MM.B
p-valor	0,1012	0,06859	0,2419	0,2943	0,4099	0,2545	0,4060
Posicionamiento	934	929	935	936	937	920	921

Para aplicar el método SMC, se realizaron ajustes consecutivos comenzando por el modelo con el marcador más significativo (es decir, de posicionamiento igual a 1), luego un ajuste con los dos MM más significativos (posicionamientos 1 y 2) y así sucesivamente, se fue agregando el MM que seguía según el posicionamiento, hasta que el número de MM en el modelo alcanzó el número de familias en el conjunto de entrenamiento menos 1, es decir, 99. A continuación se

representa la correlación entre el valor observado y el predicho por cada uno de esos modelos (Figura 5.3.2). Se distinguen con distintos símbolos a las correlaciones en el conjunto de entrenamiento (círculo) y en el conjunto de validación (triángulo). La correlación en el conjunto de entrenamiento, nunca disminuyó a medida que se incrementó el número de MM (la correlación se mantuvo igual o aumentó). Sin embargo, ese patrón no se reflejó en el conjunto de validación, pues a partir del modelo con 38 MM la correlación mostró una tendencia decreciente. Con menos de 40 MM, las correlaciones se aproximaron a 0,60 en ambos conjuntos.

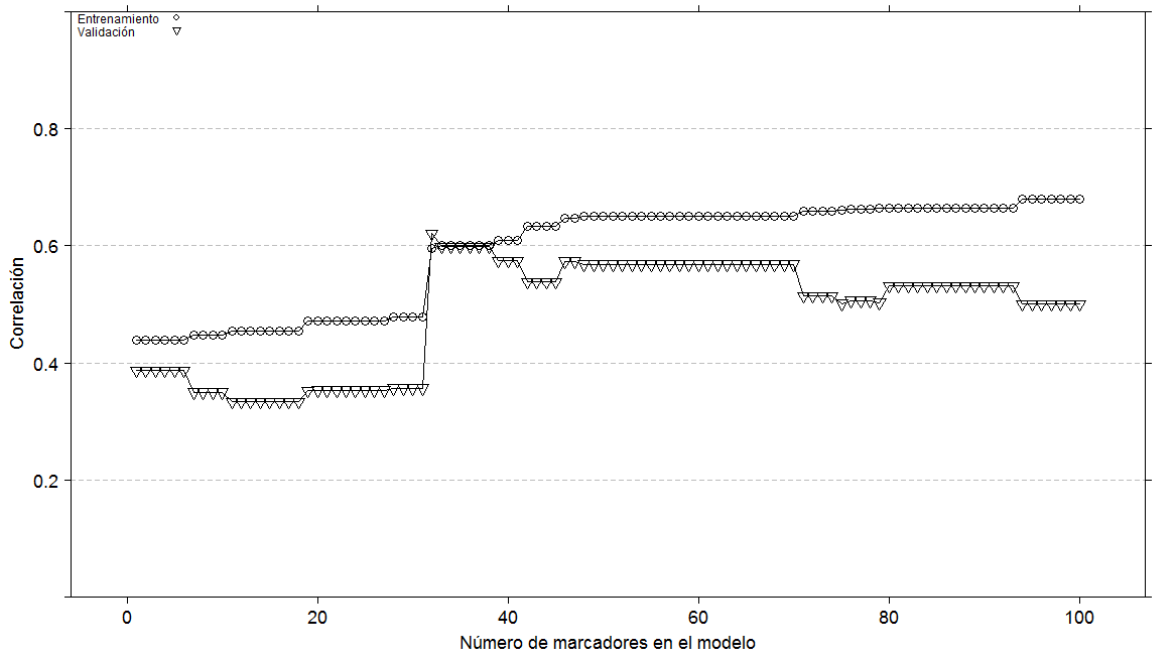

Figura 5.3.2 – Correlación entre valor observado y el predicho por el modelo en una población versus el número de MM selectos para incluir en el modelo de regresión múltiple. Las correlaciones para el conjunto de entrenamiento se representan con el círculo y para el conjunto de validación con triángulo.

En la aplicación de SMC, la máxima correlación en el conjunto de entrenamiento resultó de 0,6799 con el modelo de los 94 MM que resultaron más significativos en el modelo de marcador individual. Por otro lado, la correlación en el conjunto de validación se redujo a 0,5004.

5.3.2. Aplicación de Regresión de Ridge clásica (RR).

En la aplicación del enfoque clásico de la regresión de Ridge se utilizó un valor de penalización (lambda) definido de antemano. El efecto que tiene el valor de lambda sobre las estimaciones del modelo se puede visualizar en la Figura 5.3.3. La misma representa las estimaciones penalizadas de los coeficientes de todos los MM para dos valores diferentes del parámetro de penalización (10 y 200). Las estimaciones, en el caso del mayor valor del parámetro de penalización, variaron en un rango aproximado de -0,02 a 0,02 mientras que, para el menor valor de penalización, el rango se amplió de -0.,07 a 0,09. En otras palabras, a mayor parámetro de penalización, mayor encogimiento de las estimaciones de los coeficientes de los MM hacia el cero.

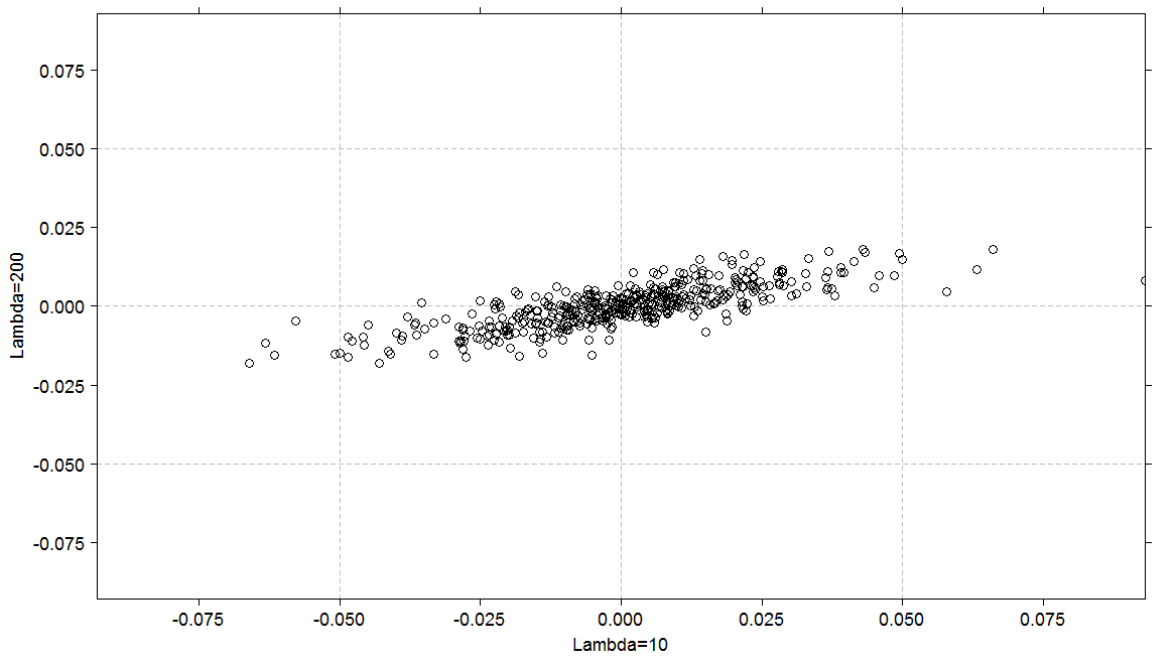

Figura 5.3.3 – Visualización de la estimación penalizada de los coeficientes de todos los MM para dos valores distintos del parámetro de penalización (10 y 200) en el marco del método RR.

La selección del valor de penalización en el enfoque RR se realiza con base en el CMEP obtenido para distintos valores del parámetro de penalización (Figura 5.3.4). En la misma se

representa la media de los CMEP en las 10 submuestras (con sus respectivas barras de error) versus el logaritmo de lambda. El valor de lambda que minimizó el CMEP fue de 2,672 (correspondiente a 0,9828 en la escala logarítmica). Para ese valor óptimo de penalización, se ajustó el modelo con todas las familias del conjunto de entrenamiento.

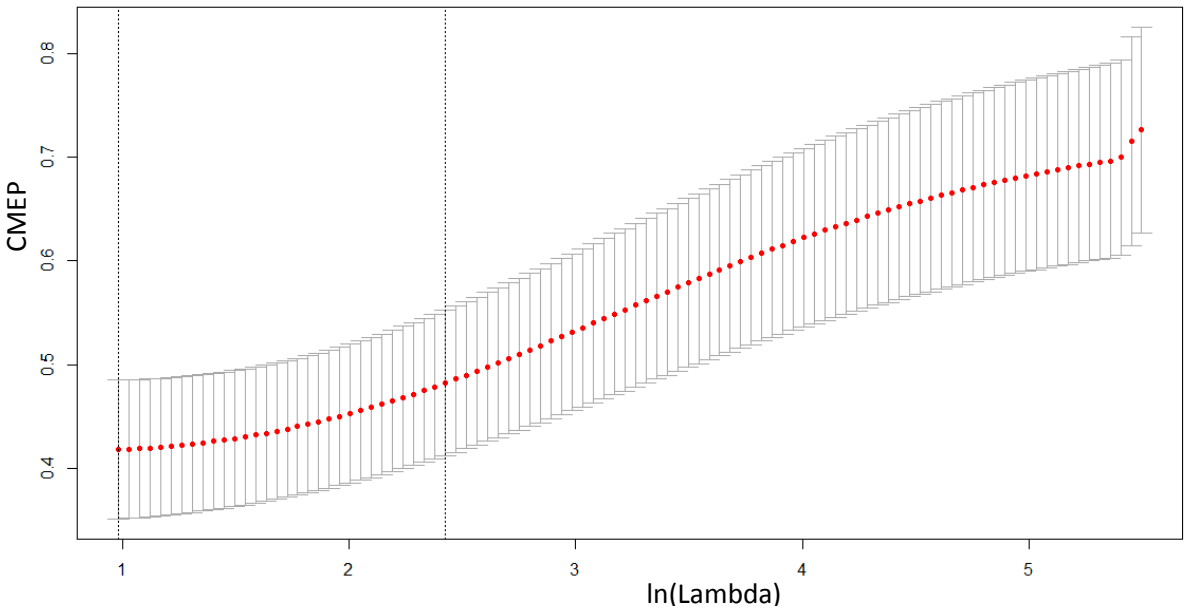

Figura 5.3.4 – CMEP obtenidos en la validación cruzada con 10 submuestras (media ± desvío estándar) cuando se empleó el modelo RR para distintos valores del parámetro de penalización (lambda).

Se obtuvieron las estimaciones de los coeficientes de cada uno de los MM, agrupadas por cromosoma y ordenadas según la posición dentro de cada uno de ellos (Figura 5.3.5). Este gráfico es la contrapartida del análisis por MM individual (Figura 5.3.1), permitiendo identificar posibles regiones asociadas al carácter (por ejemplo, en los cromosoma B, C y G donde se observan ciertos picos).

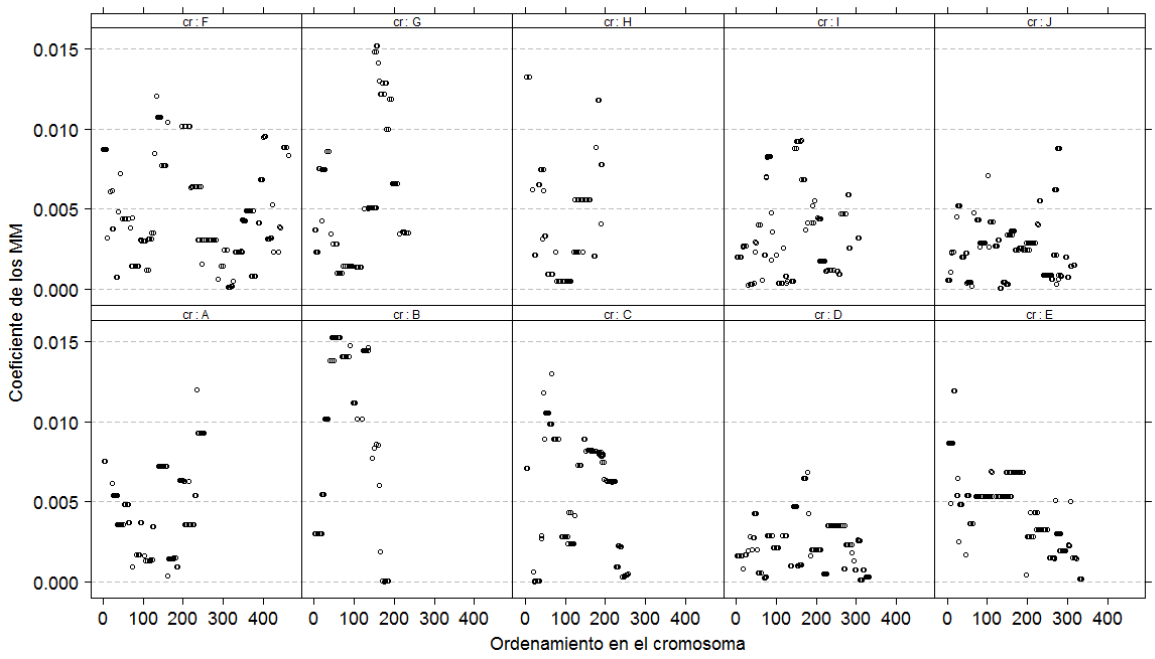

Figura 5.3.5 – Estimación de los coeficientes para todos los MM del modelo correspondiente al análisis RR para la humedad del grano en una población. Los coeficientes se presentan de acuerdo al ordenamiento de los MM dentro de cada cromosoma con base en el mapa de ligamiento.

Del ajuste se extrajeron los valores predichos y la correlación con los datos observados resultó de 0,9007 en el conjunto de entrenamiento. Por otro lado, con las estimaciones de los coeficientes se obtuvieron las predicciones para las familias en el conjunto de validación y la correlación con el valor observado resultó de 0,7325.

5.3.3. Aplicación de la Regresión de Ridge BLUP (RR-BLUP).

Con los fines de aplicar el método RR-BLUP a una población, se estimó el modelo mixto de la Ecuación 4.3.3 con coeficientes aleatorios de los MM, se extrajeron los valores predichos de cada uno de sus coeficientes y su correspondiente error estándar; la estandarización de los coeficientes de los MM se ordenó según la localización dentro de cada cromosoma (Figura 5.3.6). Esta información permitió identificar posibles regiones del genoma que se encontrarían asociadas a la humedad del grano. Por ejemplo, las regiones delimitadas por los MM próximos a 100 (en el

ordenamiento dentro de este cromosoma) en el cromosoma B y para el cromosoma G, entre 150 y 200.

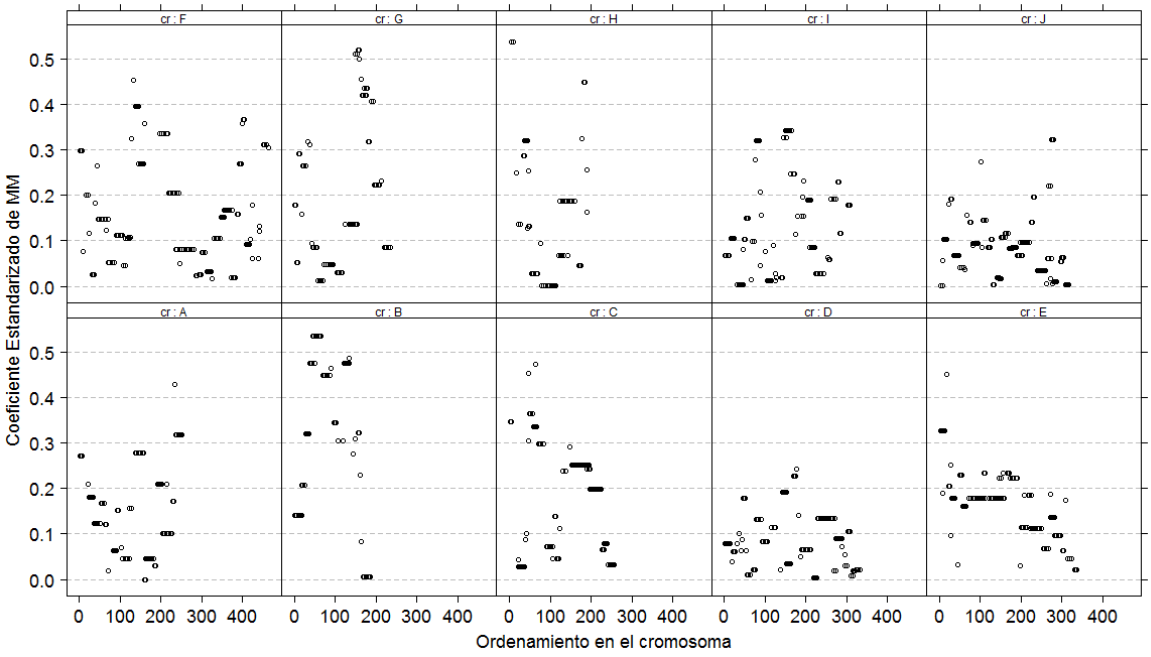

Figura 5.3.6 – Predicción de los coeficientes estandarizados para todos los MM del modelo correspondiente al análisis RR-BLUP para la humedad del grano en una población. Los coeficientes se presentan de acuerdo al ordenamiento de los MM dentro de cada cromosoma con base en el mapa de ligamiento.

Las estimaciones de las componentes de varianza tanto de los coeficientes de los MM como de los residuos del modelo se encuentran en la Tabla 5.3.1 así como también el valor del parámetro de penalización de la regresión de Ridge equivalente a este ajuste.

Tabla 5.3.1 – Estimación de las componentes de varianza de los coeficientes de los MM y de los residuos al emplear el método RR-BLUP para la humedad de grano en una población. Determinación del parámetro de penalización para emplear RR en forma equivalente al RR-BLUP

Estimación de σ_β^2	Estimación de σ_ε^2	$\lambda = \sigma_\varepsilon^2/\sigma_\beta^2$	$\ln(\lambda)$
0.001152	0.2379	206.6	5.33

Finalmente, el enfoque RR-BLUP arrojó una correlación entre los valores predichos y los observados de 0,9162 en el conjunto de entrenamiento mientras que, en el conjunto de validación esta correlación fue de 0,7307.

5.3.4. Aplicación de la Regresión de Ridge Bayesiana (BRR).

La contrapartida bayesiana de la regresión de Ridge (BRR) fue aplicada con 150.000 iteraciones, de las cuales se descartaron las primeras 50.000 (*burn-in*). Cada 10 iteraciones (*thin*) se guardó el valor de la varianza del error y de los MM (Figura 5.3.7). A partir de la iteración 50.000, la serie se observó estable, sin presencia de saltos o patrones de tendencia, con lo cual, en lo que respecta a esta estimación de componente de varianza no fue necesario aumentar la cantidad de iteraciones realizadas. Finalmente, con todas las iteraciones posteriores a 50.000, se obtuvo la estimación de este parámetro: $\hat{\sigma}_{\varepsilon}^{2} = 0{,}1966$.

Asimismo, se guardó la serie de valores para la varianza del MM cada 10 iteraciones (Figura 5.3.8). A partir de la iteración 50.000, se observa que la serie se estabilizó, con lo cual no fue necesario aumentar la cantidad de iteraciones realizadas. Finalmente, con todas las iteraciones a posteriores a 50.000, se obtuvo la estimación de este parámetro: $\hat{\sigma}_{\beta}^{2} = 0{,}01574$.

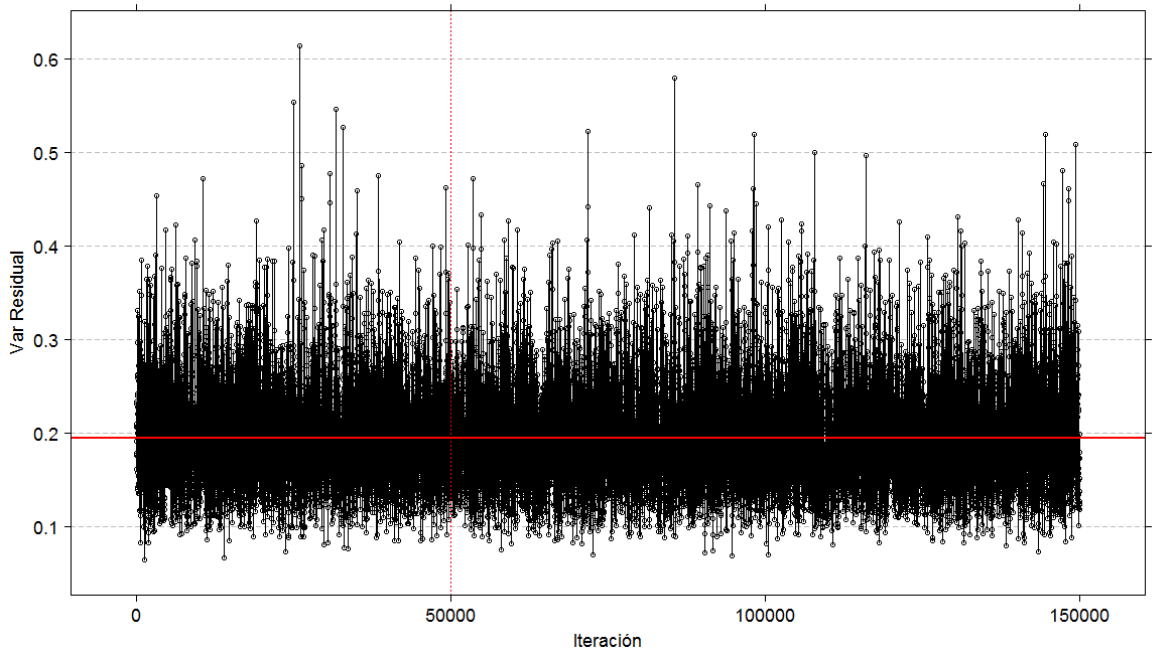

Figura 5.3.7 – Gráfico de convergencia correspondiente a la varianza residual en el método de BRR aplicado a la humedad del grano en una población. Se representan los valores de la varianza residual muestreados cada 10 iteraciones entre las 150.000 realizadas. La línea horizontal representa la estimación final del parámetro habiendo descartado las primeras 50.000 iteraciones (*burn-in*).

La Tabla 5.3.2 presenta la estimación de cada una de las componentes de varianza y el cociente entre las mismas que corresponde al parámetro de penalización si se quisiera emplear el enfoque RR en forma equivalente a BRR; el valor de lambda en este caso, es 12,39.

Se obtuvieron las estimaciones de los coeficientes correspondientes a los MM (Figura 5.3.9). Como posibles regiones del genoma que se encontrarían asociadas a la humedad del grano se pueden mencionar los primeros MM del cromosoma H (posicionamiento entre 0 y 50).

La correlación entre los valores predichos y observados en el conjunto de entrenamiento resultó de 0,9895 y en el conjunto de validación fue de 0,7040.

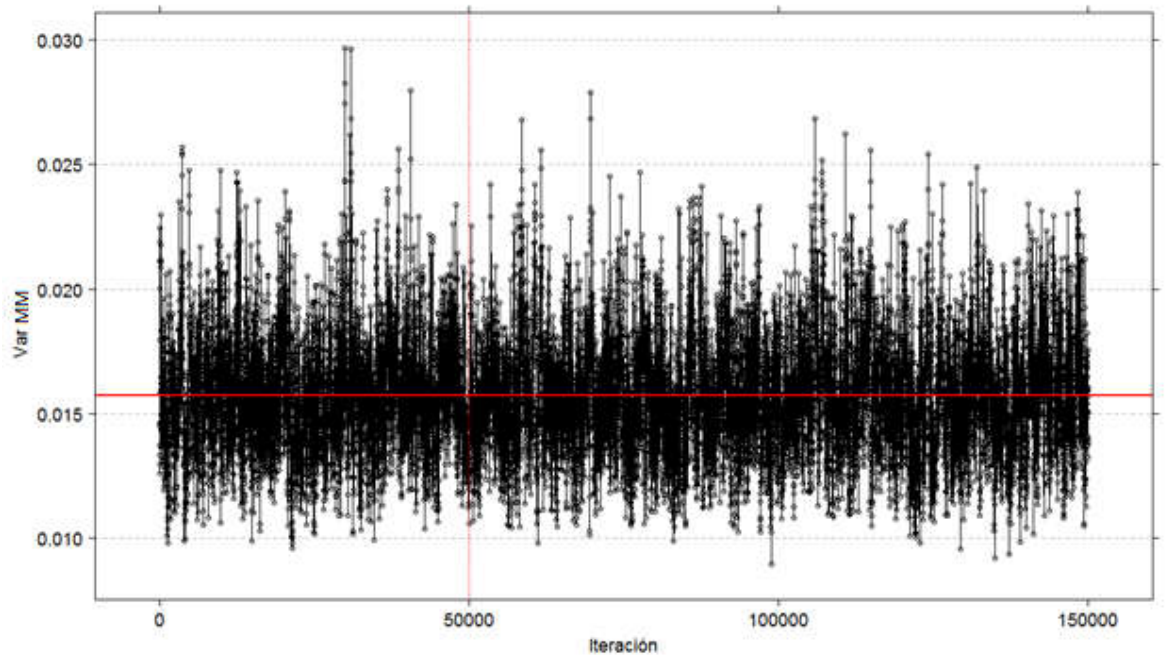

Figura 5.3.8 – Gráfico de convergencia correspondiente a la varianza de los MM en el método de BRR aplicado a la humedad del grano en una población. Se representan los valores de la varianza de MM muestreados cada 10 iteraciones entre las 150.000 realizadas. La línea horizontal representa la estimación final del parámetro habiendo descartado las primeras 50.000 iteraciones (*burn-in*).

Tabla 5.3.2 – Estimación de la componente de varianza de los coeficientes de MM y de los residuos al emplear el método BRR para la humedad de una población. Determinación del parámetro de penalización para emplear RR en forma equivalente a BRR.

Estimación de σ_β^2	Estimación de σ_ε^2	$\lambda = \sigma_\varepsilon^2/\sigma_\beta^2$	$\ln(\lambda)$
0,01574	0,1966	12,39	2,51

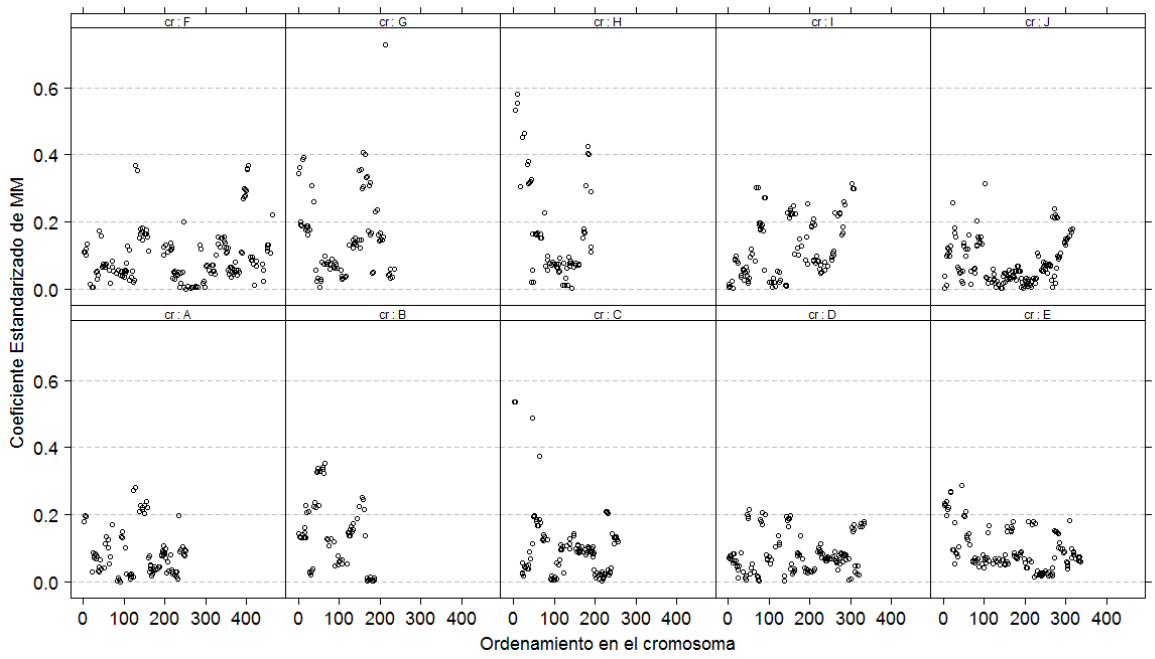

Figura 5.3.9 – Estimación de los coeficientes estandarizados para todos los MM del modelo correspondiente al análisis BRR para la humedad del grano en una población. Los coeficientes se presentan de acuerdo al ordenamiento de los MM dentro del cromosoma con base en el mapa de ligamiento.

5.3.5. Aplicación de la regresión LASSO Bayesiana (BLR).

Para la aplicación de la regresión LASSO con el enfoque bayesiano, se determinó el parámetro de regularización con base en la heredabilidad y resultó igual a 12,94. La Figura 5.3.10 representa la serie de los valores de la varianza del error versus el número de iteración. A partir de la iteración 50.000 (*burn-in*), la serie se estabilizó y con estas iteraciones, se obtuvo la estimación del parámetro: $\hat{\sigma}_{\varepsilon}^{2} = 0{,}1825$.

Las estimaciones de los coeficientes correspondientes a los MM se representan en la Figura 5.3.11. Con base en esta figura, se identificaron posibles regiones del genoma que se encontrarían asociadas a la humedad del grano. Por ejemplo, en el cromosoma H, la región delimitada por los MM que están entre las posiciones 0-50 (en el ordenamiento dentro de este cromosoma).

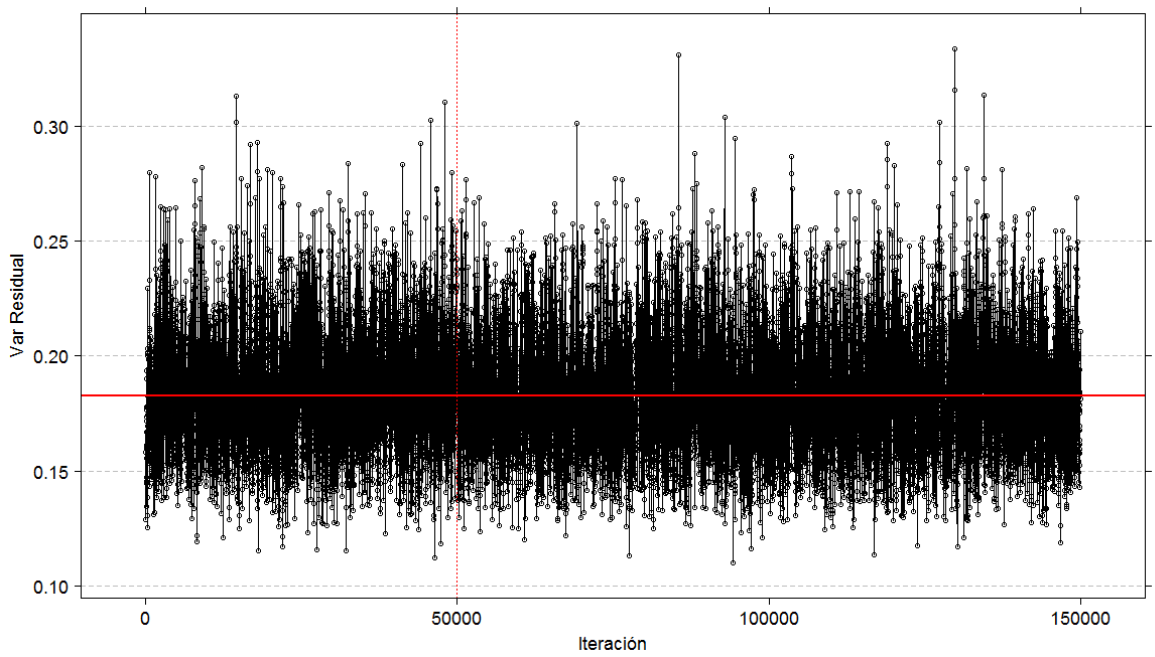

Figura 5.3.10 – Gráfico de convergencia correspondiente a la varianza residual en el método de BLR aplicado a la humedad del grano en una población. Se representan los valores de la varianza residual muestreados cada 10 iteraciones entre las 150.000 realizadas. La línea horizontal representa la estimación final del parámetro habiendo descartado las primeras 50.000 iteraciones (*burn-in*).

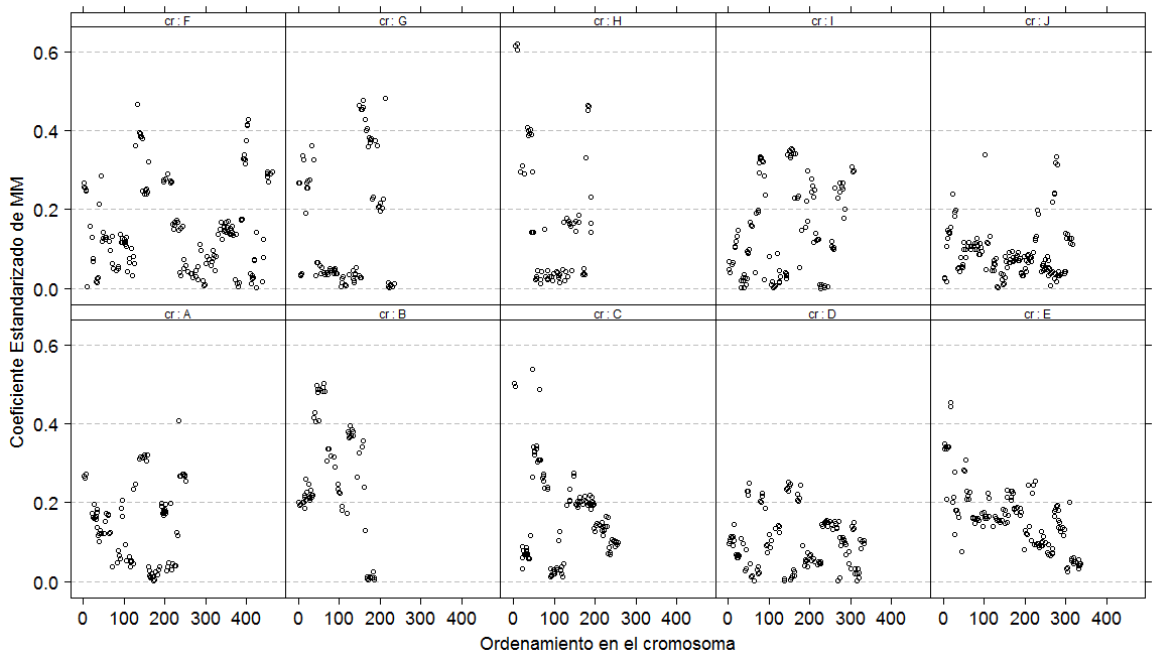

Figura 5.3.11 – Estimación de los coeficientes estandarizados para todos los MM del modelo correspondiente al análisis BLR para la humedad del grano en una población. Los coeficientes se presentan de acuerdo al ordenamiento de los MM dentro de cada cromosoma con base en el mapa de ligamiento.

Finalmente, la correlación entre los valores predichos y los observados en el conjunto de entrenamiento resultó de 0,9462 y en el conjunto de validación 0,7190.

5.3.6. Aplicación de la regresión LASSO (LR).

La elección del valor del parámetro de penalización (lambda) se basó en la Figura 5.3.12 que representa la media de los CMEP (de las 10 submuestras del conjunto de entrenamiento) con sus respectivas barras de error versus el logaritmo de lambda. El valor de lambda que minimizó el CMEP fue de: 0,03145 (correspondiente a -3,459 en la escala logarítmica). Para ese valor de penalización, se ajustó el modelo con todas las familias del conjunto de entrenamiento.

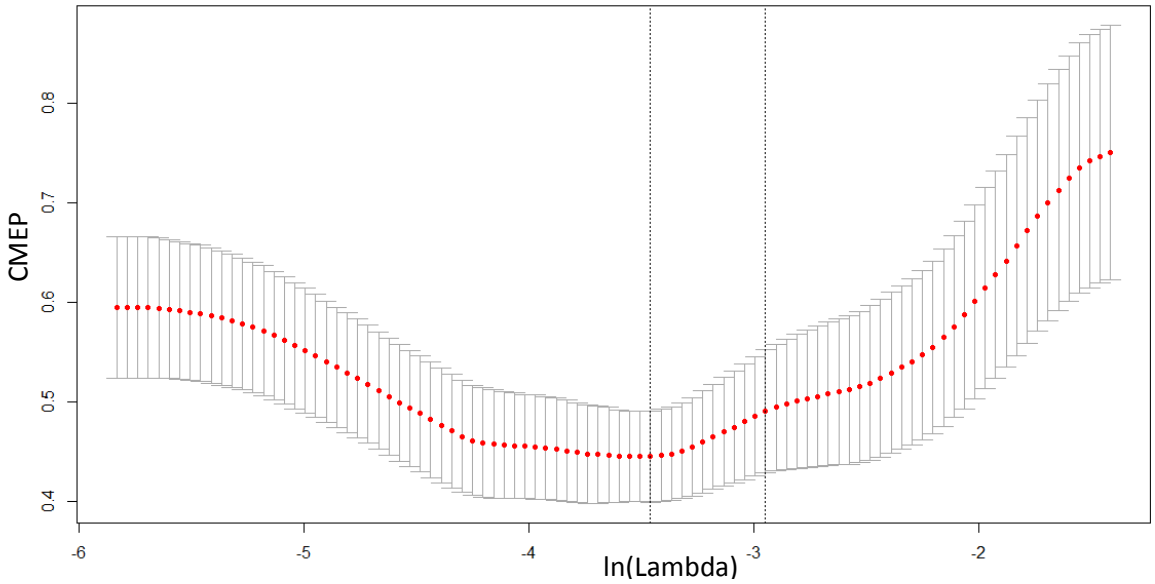

Figura 5.3.12 – CMEP obtenidos en la validación cruzada con 10 submuestras (media ± desvío estándar) cuando se empleó el modelo LR para distintos valores del parámetro de penalización (lambda).

Las estimaciones de los coeficientes de cada uno de los MM se encuentran representadas en la Figura 5.3.13, agrupadas por cromosoma y ordenadas según la posición dentro de cada uno de ellos. La selección de variables realizada se puede ver claramente en el gráfico: sólo algunos MM tuvieron estimaciones no nulas para sus coeficientes. Esta figura muestra que en los cromosomas

A, D, E, H y J fueron muy pocos los MM seleccionados mientras que los cromosomas con mayor número de MM seleccionados son, por ejemplo B, C, F y G, dando indicios de posibles QTL para la humedad.

Del ajuste para el conjunto de entrenamiento se extrajeron los valores predichos y su correlación con los datos observados resultó de 0,8904. Por otro lado, el conjunto de validación la correlación resultó de 0,6568.

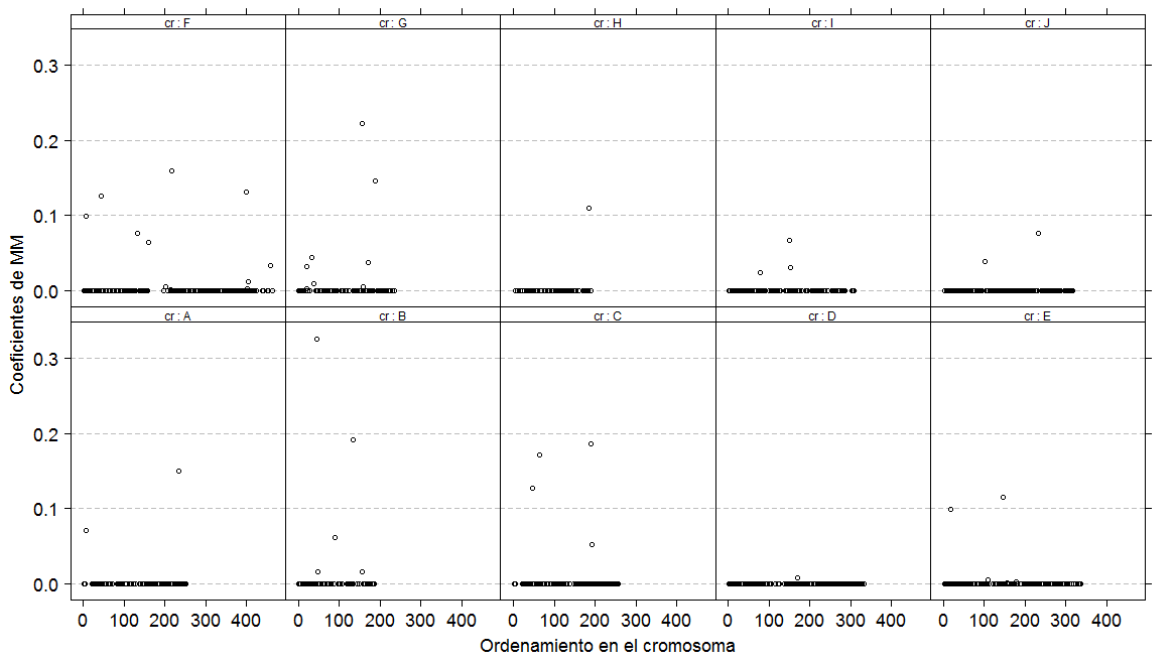

Figura 5.3.13 – Estimación de los coeficientes para todos los MM del modelo correspondiente al análisis LR para la humedad del grano en una población. Los coeficientes se presentan de acuerdo al ordenamiento de los MM dentro de cada cromosoma con base en el mapa de ligamiento.

5.4. Evaluación de la habilidad predictiva de los modelos de selección genómica

Los seis métodos aplicados detalladamente en la Sección 5.3 a una sola población y carácter, se extendieron no sólo a los tres caracteres fenotípicos observados sino también a las 20 poblaciones de maíz; esto resultó en un total de 360 (6x3x20) análisis de SG.

Para cada población, los enfoques SMC, RR y LR requirieron de la selección de un ajuste (entre múltiples ajustes realizados) que optimizaron algún criterio. En el caso de SMC, de los ajustes con distinto número de MM en el modelo, se seleccionó aquel modelo que maximizó la correlación entre el valor predicho y el observado en el conjunto de entrenamiento. En los enfoques RR y LR, de los múltiples ajustes correspondientes a distintos valores del parámetro de penalización se tomó aquél que minimizó el CMEP.

Las correlaciones entre los valores predichos y los observados fueron estimadas para el conjunto de validación y resumidas a través de la correlación media ± desvío estándar según método de SG, carácter y GH (Tabla 5.4.1). Para la humedad del grano a cosecha, el valor medio de correlación más bajo resultó de 0,39 correspondiente al método SMC en el GH1 mientras que, el valor máximo fue de 0,64 observado para los métodos RR, RR-BLUP y BLR también en el GH1. En el caso del peso hectolítrico, el valor medio de correlación mínimo fue de 0,30 para el método de SMC en el GH1 y el máximo fue de 0,55 que correspondió a los métodos RR, RR-BLUP y BLR, en todos los casos en GH1. Finalmente, para el rendimiento en grano, el valor de correlación media más bajo se presentó para el método SMC en el GH2 y resultó de 0,23 mientras que, el máximo valor alcanzado fue de 0,44 al aplicar RR y RR-BLUP en el GH1. Combinando ambos GH, la estrategia de selección de variables (SMC) arrojó el mínimo valor de correlación mientras que, el máximo valor de capacidad predictiva fue alcanzado con las técnicas de penalización: RR, RR-BLUP y BLR.

Tabla 5.4.1 – Media y desvío estándar de la correlaciones entre valor predicho por los modelo de SG y el valor observado en los conjuntos de validación de cada una de las 20 poblaciones de maíz, para los tres caracteres fenotípicos observados.

Caracter Fenotípico	GH	Método de SG					
		SMC	RR	RR-BLUP	BRR	BLR	LR
Humedad	GH1	0,39 ± 0,16	0,64 ± 0,11	0,64 ± 0,11	0,54 ± 0,12	0,64 ± 0,11	0,57 ± 0,11
	GH2	0,47 ± 0,15	0,61 ± 0,10	0,61 ± 0,11	0,47 ± 0,17	0,60 ± 0,11	0,55 ± 0,16
	Todos	0,43 ± 0,16	0,63 ± 0,11	0,63 ± 0,11	0,50 ± 0,15	0,62 ± 0,11	0,56 ± 0,13
Peso Hectolítrico	GH1	0,30 ± 0,21	0,55 ± 0,12	0,55 ± 0,11	0,42 ± 0,19	0,55 ± 0,09	0,44 ± 0,22
	GH2	0,32 ± 0,09	0,46 ± 0,16	0,46 ± 0,16	0,35 ± 0,18	0,46 ± 0,15	0,39 ± 0,10
	Todos	0,31 ± 0,16	0,50 ± 0,14	0,50 ± 0,14	0,39 ± 0,18	0,51 ± 0,13	0,42 ± 0,16
Rendimiento	GH1	0,31 ± 0,16	0,44 ± 0,17	0,44 ± 0,17	0,41 ± 0,20	0,42 ± 0,19	0,41 ± 0,16
	GH2	0,23 ± 0,24	0,39 ± 0,13	0,39 ± 0,13	0,36 ± 0,16	0,38 ± 0,14	0,30 ± 0,20
	Todos	0,27 ± 0,20	0,42 ± 0,15	0,41 ± 0,15	0,38 ± 0,18	0,40 ± 0,17	0,36 ± 0,18
Todos	Todos	0,34 ± 0,19	0,51 ± 0,16	0,51 ± 0,16	0,42 ± 0,18	0,51 ± 0,16	0,44 ± 0,18

Se representó la correlación para el conjunto de validación versus la heredabilidad en cada población y carácter fenotípico analizado, según los métodos de SG estudiados (Figura 5.4.1). SI bien los R2 resultaron inferiores a 0.5, se observó una tendencia: a mayor heredabilidad, mayor habilidad predictiva del modelo, cualquiera sea el método empleado (p-valores < 0.01).

Por otro lado, se representa la correlación en el conjunto de validación versus la similaridad entre los parentales de origen de cada una de las 20 poblaciones de maíz, según el método de SG estudiado (Figura 5.4.2). No se observó ninguna tendencia clara ni asociación significativa entre la similaridad y la habilidad predictiva del modelo.

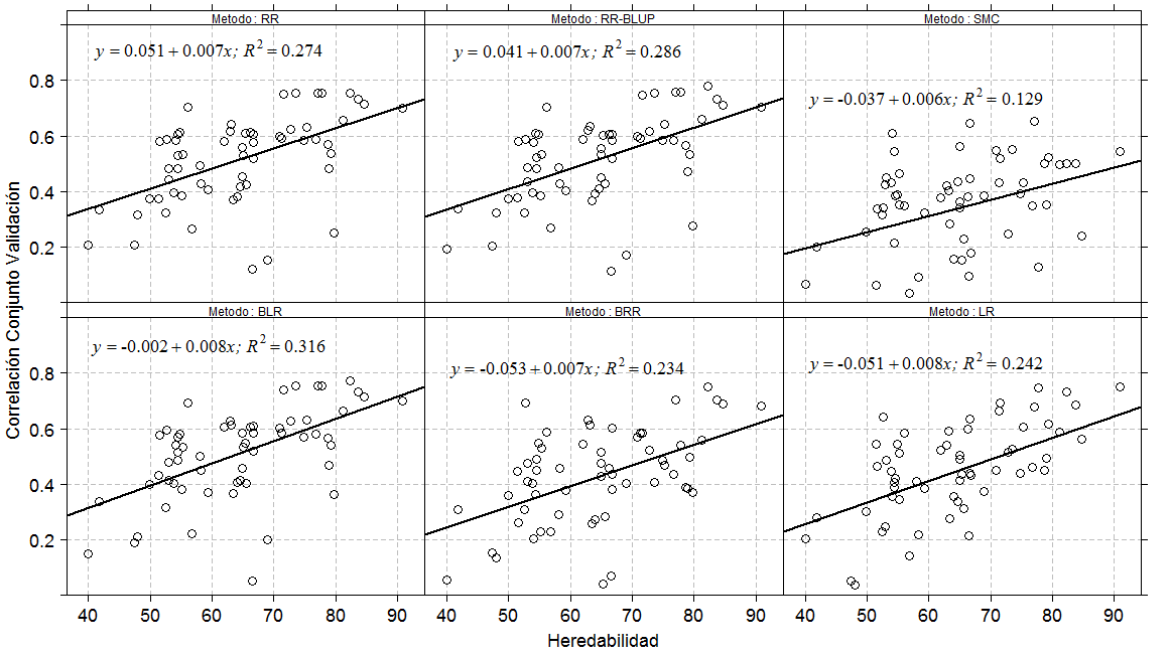

Figura 5.4.1 – Correlación entre el valor predicho y el valor observado en el conjunto de validación de las poblaciones versus la heredabilidad del carácter por cada uno de los enfoques de SG empleados.

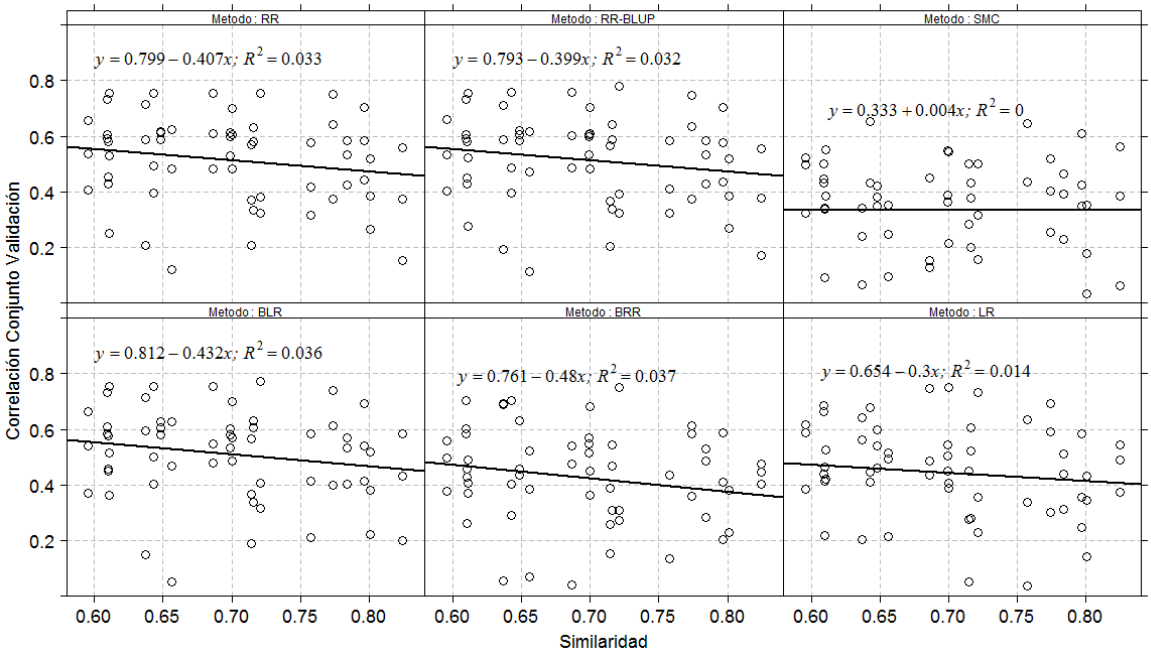

Figura 5.4.2 – Correlación entre el valor predicho y el observado para las familias en el conjunto de validación versus la similaridad de los parentales de las poblaciones por cada uno de los enfoques de SG empleados.

Se presenta el análisis de varianza (Tabla 5.4.2) que resume los resultados de ajustar el modelo de la Ecuación 4.3.10 para estudiar los efectos sobre la habilidad predictiva de los modelos de SG, del carácter y del GH, con un nivel de significación $\alpha = 0.05$. En primer lugar, se encontró que no existe efecto significativo de la interacción triple. Además, tampoco resultaron significativas las interacciones de segundo orden. Entre los efectos principales, los resultados indicaron que la habilidad predictiva de los modelos de SG resultó afectada significativamente por el método y el carácter fenotípico mientras que, no se encontró evidencia significativa de efectos del GH al que pertenecen las poblaciones.

Tabla 5.4.2 – Análisis de la varianza del modelo mixto para estudiar efectos sobre la habilidad predictiva de los modelos de SG.

Fuente de Variación	gl.num	gl.den	F.obs	P(F>F.obs)
Método	5	306	5,33	0,0001
Caracter	2	306	6,75	0,0014
GH	1	18	0,35	0,5608
Método:Caracter	10	306	0,63	0,7904
Método:GH	5	306	0,81	0,5451
Carácter:GH	2	306	0,21	0,8092
Método: Carácter:GH	10	306	0,46	0,9171

Se realizaron comparaciones múltiples de para los dos efectos principales significativos: carácter fenotípico y método de SG. Los resultados se presentan en la Tabla 5.4.3 y se trabajó con un nivel de significación $\alpha = 0.05$. La humedad a cosecha tuvo, en promedio, una correlación de 0,56, significativamente superior tanto a la correlación media del peso hectolítrico (0,44) y como a la de rendimiento en grano (0,37). A su vez, estos dos últimos caracteres fenotípicos también se diferenciaron significativamente entre sí, con lo cual, el rendimiento en grano fue el fenotipo con peor habilidad predictiva.

Las correlaciones medias para los métodos BLR, RR-BLUP y RR no se diferenciaron entre sí significativamente, así como tampoco lo hicieron las correlaciones medias de los métodos BRR y LR. Los métodos BLR, RR-BLUP y RR fueron los de mejor habilidad predictiva (aproximadamente

0,51 en cada uno de los casos), luego siguieron los métodos LR y BRR en ese orden, con correlaciones medias de 0,44 y 0,42, respectivamente y por último, el método de peor habilidad predictiva resultó el de selección de variables (SMC) con una correlación media de 0,34.

Tabla 5.4.3 – Comparaciones múltiples de las correlaciones medias en el conjunto de validación para los distintos caracteres fenotípicos y métodos de SG empleados.

Factor	Nivel A	Nivel B	Media A	Media B	Dif. (A-B)	E.E.	Z	p-valor
Caracter Fenotípico	Humedad	Peso	0,56	0,44	0,12	0,02	7,47	<0,0001
	Humedad	Rendimiento	0,56	0,37	0,19	0,02	11,29	<0,0001
	Peso	Rendimiento	0,44	0,37	0,06	0,02	3,82	0,0004
Método de SG	BLR	BRR	0,51	0,42	0,09	0,02	3,61	0,0042
	BLR	LR	0,51	0,44	0,07	0,02	2,78	0,0609
	BLR	RR	0,51	0,51	0,00	0,02	-0,19	1,0000
	BLR	RR-BLUP	0,51	0,51	0,00	0,02	-0,20	1,0000
	BLR	SMC	0,51	0,34	0,17	0,02	7,39	0,0000
	BRR	LR	0,42	0,44	-0,02	0,02	-0,83	0,9625
	BRR	RR	0,42	0,51	-0,09	0,02	-3,79	0,0021
	BRR	RR-BLUP	0,42	0,51	-0,09	0,02	-3,80	0,0020
	BRR	SMC	0,42	0,34	0,09	0,02	3,78	0,0021
	LR	RR	0,44	0,51	-0,07	0,02	-2,97	0,0357
	LR	RR-BLUP	0,44	0,51	-0,07	0,02	-2,98	0,0348
	LR	SMC	0,44	0,34	0,11	0,02	4,61	0,0001
	RR	RR-BLUP	0,51	0,51	0,00	0,02	-0,01	1,0000
	RR	SMC	0,51	0,34	0,18	0,02	7,58	0,0000
	RR-BLUP	SMC	0,51	0,34	0,18	0,02	7,59	0,0000

6. CONCLUSIONES

En el presente trabajo se presentaron métodos estadísticos de SG que permiten predecir el valor genético de los individuos con base en los MM, superando las dos principales dificultades de los modelos de SG: la dimensionalidad de los datos y la multicolinealidad del modelo. Se abarcaron tres estrategias: selección de variables, estimación penalizada o la combinación de ambos, desde enfoques clásicos o bayesianos. Específicamente, se estudiaron seis métodos de SG: selección de variables y estimación del modelo de regresión por mínimos cuadrados (SMC), Regresión de Ridge clásica (RR), Regresión de Ridge con enfoque de modelos mixtos (BLUP-RR), Regresión de Ridge Bayesiana (BRR), Regresión LASSO Bayesiana (BLR) y Regresión LASSO clásica (LR).

Asimismo, los métodos se aplicaron a datos empíricos se evaluaron en 20 poblaciones biparentales F3 correspondientes a dos GH utilizados en un programa de mejoramiento de maíz de la compañía multinacional Monsanto. Para cada población se contaba tanto con datos correspondientes a los caracteres fenotípicos: humedad del grano a cosecha, el peso hectolítrico y el rendimiento en grano como con datos de MM. Las poblaciones presentaron variabilidad en distintos aspectos: número de familias, número de SNPs evaluados, número de ambientes en que se evaluó cada carácter, distribución de MM a lo largo del genoma.

La aplicación los distintos métodos de SG fue presentada detalladamente para una población y un carácter en particular a modo de ejemplo. En primer lugar, la población de dividió al azar en dos conjuntos, uno de entrenamiento (75%) y otro de validación (25%). Para cada modelo, con base en los valores genéticos y a los datos de MM del entrenamiento, se obtuvo la estimación de los coeficientes de MM. Con las estimaciones de los coeficientes y los datos de MM de las familias en el conjunto de validación, se predijo su desempeño. La habilidad predictiva de cada modelo se

evaluó estimando la correlación entre el valor predicho y el observado en el conjunto de validación.

Siguiendo un proceso similar al descripto para una población y un carácter fenotípico, la aplicación se extendió a todas las poblaciones y caracteres cuantitativos disponibles con los fines de evaluar y comparar la habilidad predictiva de cada método. Esta última etapa, consistió de la estimación de un total de 360 (20 x 3 x 6) modelos. Tales estimaciones permitieron el cálculo de las correlaciones entre valor predicho y observado de las familias en el conjunto de validación de cada población y así se obtuvo una medida de la habilidad predictiva en cada uno de los 360 análisis.

El método SMC se basó en emplear selección de variables previo al ajuste por mínimos cuadrados del modelo de SG con múltiples marcadores. La técnica de selección de variables empleada fue la más sencilla, ya que se basó en la probabilidad asociada al MM en un modelo por marcador individual. SMC resultó de aplicación simple ya que utilizó el método de mínimos cuadrados disponible en cualquier paquete estadístico. Sin embargo, el desempeño en términos de habilidad predictiva de SMC fue el más pobre de todos los métodos estudiados, independientemente del carácter fenotípico o GH. La correlación media (de todos los análisis) entre el valor predicho y el observado dentro del conjunto de validación fue de 0,34 ± 0,19, al menos un 20% menor que los otros métodos. El método SMC no evitó que se generen problemas de multicolinealidad en el modelo; además, para poder llevar a cabo la estimación por mínimos cuadrados, se debió incorporar un número limitado de MM en el modelo (inferior al número de familias) y la baja habilidad predictiva podría explicarse por el sobreajuste del modelo. Estos resultados coinciden con las conclusiones a las que arriban Bernardo y Yu (2009), a través de datos simulados de cruzamientos de prueba. Los autores indican que, para distintos niveles de heredabilidad, distinta cantidad de QTL y tamaños de muestra, los métodos de SG que no involucran la identificación de los marcadores asociados a los caracteres de interés (es decir, que

no hacen selección de variables) superan a los métodos que involucran la búsqueda de un subconjunto de marcadores que afecten significativamente al carácter. Estos resultados presentan a SMC como un método de baja habilidad predictiva y concuerdan con los primeros estudios comparativos presentados por Meuwissen *et al.* (2001) en el marco de mejoramiento animal y específicamente en el mejoramiento de maíz, en los trabajos de Bernardo y Yu (2007), Mayor y Bernardo (2009) y Lorenzana y Bernardo (2009).

Los métodos de regresión de Ridge emplean el concepto de estimación penalizada. El mismo pretende balancear la bondad del ajuste con la complejidad del modelo a través del uso de un parámetro de regularización o penalización (λ). Dentro de este enfoque, se estudiaron tres metodologías alternativas: clásica (RR), basada en modelos mixtos (RR-BLUP) y bayesiana (BRR). Cada uno de ellas aborda el problema de selección del parámetro de penalización desde distintos enfoques.

El método RR, requirió la búsqueda del valor óptimo de λ que consistió en evaluar una grilla de valores posibles de λ, realizar validación cruzada con 10 submuestras para calcular el CMEP y elegir aquél valor de λ que lo haga mínimo. Este procedimiento no se encuentra disponible en cualquier paquete estadístico, en este caso se utilizaron las funciones cv.glmnet y glmnet de la librería glmnet (Friedman *et al.*, 2010) en R; resultando de rápida y fácil implementación. La media de la correlación entre el valor observado y el predicho por el modelo de SG en los conjuntos de validación resultó: 0,51 ± 0,16.

El método basado en modelos mixtos, fue implementado a través de la función mixed.solved de la librería rrBLUP (Endelman, 2011) en R y también resultó de fácil implementación y más rápido que RR ya que no requiere la búsqueda del valor de penalización. RR-BLUP, no mostró diferencias significativas en la habilidad predictiva comparado con RR. En particular para la humedad, la correlación media fue de 0,63 ± 0,11, valor del orden de la máxima correlación

encontrada por Lorenzana y Bernardo (2009) al aplicar RR-BLUP en poblaciones biparentales de maíz también provenientes de Monsanto. En cuanto al rendimiento en grano, en el presente trabajo se obtuvo una correlación media de 0,41 ± 0,15 similar al máximo valor (0,43) encontrado por Lorenzana y Bernardo (2009).

El enfoque BRR, se implementó utilizando la función BGLR de la librería de R con el mismo nombre (de los Campos y Perez-Rodriguez, 2014). Su ejecución tiene mayores desafíos, en primer lugar, por la definición de los hiper-parámetros de las distribuciones *a priori* propias de la estimación bayesiana, en segundo lugar por el tiempo computacional que demanda el método iterativo y en tercer lugar pero no menos importante, requiere el estudio de la convergencia de las estimaciones de las varianzas. Se encontró que la habilidad predictiva para la contrapartida bayesiana de la regresión de Ridge fue significativamente inferior a la habilidad predictiva de RR y RR-BLUP: 0,42 ± 0,18. El desempeño más pobre de BRR respecto de RR o RR-BLUP no era de esperar ya que, como señalan de los Campos *et al.* (2013), tanto RR como BRR realizan una penalización que es homogénea a través de todos los MM, esto puede deberse a efectos en la sdefinición de los hiper-parámetros de las distribuciones *a priori* poco informativas.

De los tres enfoques de la regresión de Ridge, el RR-BLUP es en cierta forma más fácil y directo de implementar ya que no realiza una búsqueda del parámetro de penalización (este valor se relaciona al cociente entre las varianzas del error y de los coeficientes de los MM) y se puede estimar utilizando cualquier paquete estadístico que permita llevar a cabo la estimación de modelos mixtos. De hecho, en el marco de la SG, el método RR-BLUP es más ampliamente aplicado (de los Campos *et al.*, 2013). Su habilidad predictiva se destaca en distintos estudios en el marco del mejoramiento animal (Meuwissen *et al.,* 2001; Habier *et al*., 2007), con datos simulados (Bernardo y Yu, 2007; Lorenz, 2013) y con datos reales de mejoramiento vegetal en distintos cultivos como Heffner *et al.* (2011) en trigo, Zhong *et al.* (2009) en avena, Lorenzana y

Bernardo (2009) en *arabidopsis*, avena y maíz y además, específicamente para este último cultivo Riedelshelmer *et al.* (2013), Jacobson *et al.* (2014), entre otros.

La regresión LASSO en su enfoque bayesiano (BLR) se considera como un método de estimación penalizada ya que no asigna valores nulos a los coeficientes, es decir, no realiza selección de variables. Se encontró que la habilidad predictiva del método BLR, no difiere significativamente de la habilidad predictiva hallada para RR y RR-BLUP (0,51 ± 0,16). Este resultado se contrapone con la conclusión a la que arriban Crossa *et al.* (2010) en líneas de trigo donde encuentran que el método BLR muestra mayor habilidad predictiva que el RR-BLUP, al igual que en el trabajo de Pérez *et al.* (2010) con datos simulados. Ambos trabajos establecen que el motivo por el cual se espera que BLR se desempeñe mejor que RR-BLUP es porque BLR no produce una penalización homogénea a través de todos los MM, sino que permite cierta flexibilidad. Por otro lado, Endelman (2011) establece que la capacidad predictiva del método BLR (implementado con la función BLR de R) es equivalente a la encontrada con RR-BLUP (usando la función mixed.solve) en 7 fenotipos de datos de líneas de trigo; lo cual soporta no solo la metodología empleada en el presente trabajo sino también los resultados hallados. Es importante notar que el enfoque BLR tiene similar complejidad que BRR, en términos de selección de parámetro de penalización, tiempo computacional y estudio de convergencia de la componente de varianza del residual, y que entre estos dos métodos la habilidad predictiva de BLR fue significativamente superior.

Finalmente, la aplicación del método de regresión LASSO clásico (LR), combina estrategias de selección de variables con métodos de penalización sobre los coeficientes. La estimación vía LR es más rápida que BLR (similar al caso de RR versus BRR) pero, al igual que BLR, no está disponible en cualquier paquete estadístico básico. De hecho, es el método menos aplicado en SG pues restringe el número de coeficientes no nulos en el modelo hasta un máximo equivalente al número de familias menos 1. Este método, asume que sólo unos pocos MM afectan la variable

respuesta, lo cual puede o no ser válido para el carácter fenotípico en estudio. Con los datos analizados en el presente trabajo, se encontró que la habilidad predictiva media para LR fue de 0,44 ± 0,18, significativamente superior a SMC, de similar desempeño que BRR y BLR pero de inferior desempeño que RR y RR-BLUP en los tres caracteres fenotípicos analizados.

Uno de los objetivos fue estudiar si la habilidad predictiva de un modelo de SG depende de factores tales como el carácter fenotípico que se observa o características propias de las poblaciones sobre los cuales se desea hacer la SG. Los resultados de la aplicación indicaron que para la habilidad predictiva no existen efectos de interacción de tercer ni segundo orden entre GH, carácter fenotípico y método de SG empleado. Asimismo, en este trabajo se encontró que la habilidad predictiva de un modelo de SG para poblaciones de maíz dependió significativamente de la metodología aplicada, siendo los métodos BLR, RR-BLUP y RR los de mejor desempeño, luego BRR y LR y en último lugar el método SMC. Además, se encontraron evidencias significativas de que la habilidad predictiva del modelo depende de la carácter observado; los modelos de SG presentaron mejor desempeño para la humedad (correlación media de 0.56), en segundo lugar para peso hectolítrico (correlación media de 0.44) y por último para rendimiento (correlación media de 0.37). Este ordenamiento es acorde a los valores de heredabilidad estimados, siendo la humedad el carácter que presentó los valores más altos de heredabilidad en todas las poblaciones (54.1% a 90.9%), peso hectolítrico con valores intermedios (49.9% a 81.2%) y por último el rendimiento, con los valores más bajos de heredabilidad (40% a 71.3%). Estos resultados mantienen la misma tendencia que presentan en el trabajo Jacobson *et al.* (2014) utilizando únicamente el método RR-BLUP un conjunto mayor de poblaciones de maíz de Monsanto. Por otro lado, no se encontraron evidencias significativas de dependencia entre la habilidad predictiva del modelo y el GH al cual pertenece la población ni con la similaridad entre los parentales que dan origen a la población.

7. DISCUSIÓN

La aplicación de SG consiste en primer lugar, del desarrollo de un modelo basado en un conjunto de individuos para los cuales se cuenta tanto con datos fenotípicos como con datos provenientes de una alta densidad de MM ubicados a lo largo del genoma. Este modelo se utiliza luego para realizar selección de otros individuos para los cuales sólo se dispone de datos de MM (Thomson, 2014). Por razones computacionales, no fue factible el ajuste un modelo para un carácter que contemplara simultáneamente efectos ambientales y genéticos expresados en función de los MM. Por lo tanto, la estimación del modelo se realizó en dos etapas. La primera se concentró en estimar los valores genéticos a través de un modelo mixto que contemplaba efectos ambientales y genéticos, estos últimos sin incluir MM). En la segunda etapa, se estimó el modelo de SG propiamente dicho, donde los valores genéticos fueron expresados en función de los MM a través de un modelo de regresión lineal. Sin embargo, el método RR-BLUP, que estuvo entre los de mejor desempeño en este trabajo, permitiría superar esta limitante ya que emplea un modelo mixto donde se pueden incorporar además los efectos de ambiente.

Los distintos métodos aplicados en el presente trabajo son sólo algunos de los múltiples métodos que pueden aplicarse para llevar a cabo SG. La versión del método de SMC fue aplicada como el escenario más desfavorable que no contempla los problemas de multicolinealidad presente en los datos, sin embargo, debe reconocerse que esta versión puede ser optimizada aplicando algún método de selección de variables y luego realizar ajuste por mínimos cuadrados. En cuanto a los enfoques penalizados, existen otros métodos que resultan un compromiso entre la penalización de Ridge y la LASSO (disponibles en la librería de R glmnet). Además, se han implementado otros métodos en el marco de la SG tales como, BayesA, BayesC, y métodos no paramétricos (redes neuronales, “random forest”) entre otros (de los Campos *et al.*, 2013).

En la aplicación de los métodos de SG a una población y carácter particular, se observó que los coeficientes de regresión estimados en cada método variaron. Por un lado, pueden variar en la escala debido a que las funciones de las distintas librerías de R pueden centrar y/o estandarizar tanto las variables explicativas como la respuesta. Sin embargo, los patrones dentro de cada uno de los cromosomas variaron de un método a otro. Si bien la identificación de regiones del genoma no es el objetivo principal de los métodos de SG, no se encontró bibliografía que estudie la dependencia de los distintos métodos de SG y la posibilidad de identificar regiones del genoma asociadas con el carácter de interés que señala de los Campos *et al.* (2009).

En la actualidad, uno de los mayores desafíos del mejoramiento genético basado en MM es lograr buenas predicciones de los valores genéticos y para ello, la aplicación de métodos estadísticos juega un rol crucial. La implementación de SG en la práctica implica tomar múltiples decisiones que resultan determinantes por ejemplo, definir sobre qué caracteres fenotípicos se va a llevar a cabo la SG, estudiando la precisión de los métodos de medición, la heredabilidad y patrones de genes intervinientes. Otros aspectos a considerar son: la distribución eficiente de los recursos para garantizar calidad en el proceso de SG, el tamaño (por ejemplo, número de poblaciones, años y ambientes), la composición del conjunto de entrenamiento del modelo de SG (para los cuales se realiza la observación delos caracteres fenotípicos y de los genotipos de los MM para los individuos), la densidad de los marcadores, la elección del modelo, entre otros. Todos estos factores afectan la capacidad predictiva de los métodos de SG (de los Campos *et al.*, 2013). Algunos factores que afectan la habilidad predictiva del modelo claramente no son controlables. Por ejemplo, la heredabilidad del carácter sobre el cual se realiza selección, la estructura genética, la expansión del desequilibrio de ligamiento.

Este trabajo se concentró en estudiar un aspecto controlable que afecta la habilidad predictiva: la definición del modelo de SG a emplear Los estudios de simulación indican que la elección del modelo depende de la arquitectura genética, es decir, del número de QTL

intervinientes, entre otros factores. En el presente trabajo se encontró que los métodos de penalización RR, RR-BLUP y BLR tienen la mayor habilidad predictiva entre los seis métodos estudiados para las poblaciones biparentales de maíz analizadas. Estudios empíricos recientes concluyen que todos los resultados anticipados por medio de estudios de simulación no son totalmente confirmados al trabajar sobre datos reales (de los Campos *et al.*, 2013). Por lo tanto, el mejor modelo de SG en términos de habilidad predictiva depende de cada conjunto de datos; de allí la importancia de realizar estudios comparativos cubriendo diferentes enfoques como los presentados en esta tesis.

Todos los restantes factores controlables deberían ser materia de estudio a la hora de aplicar SG. La densidad de los MM es un factor que puede incrementar la habilidad predictiva de un modelo (Bernardo y Yu, 2007; Lorenzana y Bernardo, 2009; Heffner *et al.,* 2011; Combs y Bernardo, 2013). En este trabajo se disponía de un número limitado de SNPs por población y además se tenía información de MM imputados para incrementar la densidad a través del genoma. Podría ser de interés estudiar otros métodos de imputación como una estrategia para mejorar no sólo la habilidad predictiva de los métodos de SG sino también como una forma de hacer uso más eficiente de recursos Jacobson *et al.* (2015).

El tamaño de la muestra afecta la habilidad predictiva de cualquier modelo de regresión pues los errores estándares de los coeficientes son menores a medida que el tamaño muestral aumenta (de los Campos *et al.*, 2013). En la aplicación de SG, los coeficientes de los MM son estimados dentro del conjunto de entrenamiento. Luego, esas estimaciones son utilizadas para predecir el desempeño de los individuos en el conjunto de validación para los cuales se dispone de MM pero no de datos fenotípicos. En este trabajo, el conjunto de entrenamiento consistió de una submuestra de cada población, lo cual le da la característica de ser representativo de la población objetivo; sin embargo, el tamaño del conjunto de entrenamiento se encuentra acotado. Una forma de incrementar el tamaño muestral podría abordarse combinando múltiples

poblaciones biparentales de acuerdo con la propuesta de múltiples autores (Schulz-Streeck *et al.*, 2012; Zhao *et al.*, 2012; Riedelsheimer *et al*., 2013; Jacobson *et al.*, 2014).

8. REFERENCIAS

Akaike, H. (1981). Likelihood of a model and information criteria. Journal of Econometrics, 16(1), 3-14.

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., & Jannink, J. L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. The Plant Genome, 4(2), 132-144.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version, 1(4).

Bernardo, R. (2002). Breeding for quantitative traits in plants. Woodbury, Minn.: Stemma Press.

Bernardo, R. (2009). Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Science, 49(2), 419-425.

Bernardo, R. (2010). Genomewide selection with minimal crossing in self-pollinated crops. Crop Science, 50(2), 624-627.

Bernardo, R., & Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. Crop Science, 47(3), 1082-1090.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., ... & McMullen, M. D. (2009). The genetic architecture of maize flowering time. Science, 325(5941), 714-718.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. The American Statistician, 46(3), 167-174.

Chapman, S., Cooper, M., Podlich, D., & Hammer, G. (2003). Evaluating plant breeding strategies by simulating gene action and dryland environment effects. Agronomy Journal, 95(1), 99-113.

Collard, B. C. Y., Jahufer, M. Z. Z., Brouwer, J. B., & Pang, E. C. K. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica, 142(1-2), 169-196.

Combs, E., & Bernardo, R. (2013). Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. The Plant Genome, 6(1) :1–7.

CRI México (s. f.). Selección Genómica una aplicación práctica. Departamento Técnico de Reproduccuón Animal. URL http://www.reproduccionanimal.com.mx/

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueno, J., Araus, J. L., ... & Braun, H. J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics, 186(2), 713-724.

de los Campos, G. (2012). Comunicación Personal. Curso: “Statistical Methods for Genome-Enabled Predictions”. Rosario, Argentina.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., & Calus, M. P. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics, 193(2), 327-345.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., ... & Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182(1), 375-385.

de los Campos, G., & Perez-Rodriguez, P. (2012). BLR: Bayesian Linear Regression. R package version 1.3.

de los Campos, G., & Perez Rodriguez, P. (2014). BGLR: Bayesian Generalized Linear Regression. R package version 1.0.3.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome, 4(3), 250-255.

Falconer, S., & Mackay, T. (2001). Introducción a la genética cuantitativa. Cuarta edición. Edición en español. Editorial Acribia, Zaragoza, España.

Fehr, W. (1987). Principles of cultivar development. New York: Macmillan Pub. Co., c1987.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52(02), 399-433.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, 33(1), 1-22.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. Pattern Analysis and Machine Intelligence, IEEE Transactions on, (6), 721-741.

Gianola, D., Fernando, R. L., & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics, 173(3), 1761-1776.

Guo, Z., Tucker, D. M., Lu, J., Kishore, V., & Gay, G. (2012). Evaluation of genome-wide selection efficiency in maize nested association mapping populations. Theoretical and Applied Genetics, 124(2), 261-275.

Habier, D., Fernando, R. L., & Dekkers, J. C. M. (2007). The impact of genetic relationship information on genome-assisted breeding values. Genetics, 177(4), 2389-2397.

Haldane, J. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. Journal of Genetics, 8(29), 299-309.

Haldane, J. (1932). The causes of evolution. Princeton: Princeton University Press.

Haley, C. S., & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69(4), 315-324.

Hallauer, A., Russell, W., & Lamkey, K. (1988). Corn Breeding. In: Sprague and Dudley (eds) Corn and corn improvements. 3rd Edition. Am. Soc. Agron., Madison, Wisconsin.

He, S., Zhao, Y., Mette, M. F., Bothe, R., Ebmeyer, E., Sharbel, T. F., ... & Jiang, Y. (2015). Prospects and limits of marker imputation in quantitative genetic studies in European elite wheat (Triticum aestivum L.). BMC genomics,16(1), 168.

Heffner, E. L., Jannink, J. L., Iwata, H., Souza, E., & Sorrells, M. E. (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. Crop Science, 51(6), 2597-2606.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67.

Holland, J. B., Nyquist, W. E., & Cervantes-Martínez, C. T. (2003). Estimating and interpreting heritability for plant breeding: An update. Plant Breeding Reviews, 22, 9-112.

Iwata, H., & Jannink, J. L. (2011). Accuracy of genomic selection prediction in barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Science, 51(5), 1915-1927.

Jacobson, A., Lian, L., Zhong, S., & Bernardo, R. (2014). General combining ability model for genomewide selection in a biparental cross. Crop Science, 54(3), 895-905.

Jacobson, A., Lian, L., Zhong, S., & Bernardo, R. (2015). Marker Imputation Prior to Genomewide Selection in Biparental Maize Populations.

Jannink, J. L. (2010). Dynamics of long-term genomic selection. Genetics Selection Evolution, 42(1), 35.

Lande, R., & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124(3), 743-756.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., ... & Jannink, J. L. (2011). 2 Genomic Selection in Plant Breeding: Knowledge and Prospects. Advances in agronomy, 110, 77.

Lorenzana, R. E., & Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. Theoretical and Applied Genetics, 120(1), 151-161.

Lush, J. L. (1943). Animal breeding plans. Animal breeding plans., (Edn 2).

Lynch, M., & Walsh, B. (1998). Genetics and analysis of quantitative traits (Vol. 1). Sunderland, MA: Sinauer.

Mayor, P. J., & Bernardo, R. (2009). Genomewide Selection and Marker-Assisted Recurrent Selection in Doubled Haploid versus F Populations. Crop Science, 49(5), 1719-1725.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics, 157(4), 1819-1829.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. Methods in Ecology and Evolution, 4(2), 133-142.

Nakaya, A., & Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding?. Annals of Botany, mcs109.

Park, T., & Casella, G. (2008). The bayesian lasso. Journal of the American Statistical Association, 103(482), 681-686.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. Biometrika, 58(3), 545-554.

Pérez, P., de los Campos, G., Crossa, J., & Gianola, D. (2010). Genomic-enabled prediction based on molecular markers and pedigree using the Bayesian linear regression package in R. The Plant Genome, 3(2), 106-116.

Piepho, H. P. (2009). Ridge regression and extensions for genomewide selection in maize. Crop Science, 49(4), 1165-1176.

Piyasatian, N., Fernando, R. L., & Dekkers, J. C. M. (2007). Genomic selection for marker-assisted improvement in line crosses. Theoretical and Applied Genetics, 115(5), 665-674.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Riedelsheimer, C., Endelman, J. B., Stange, M., Sorrells, M. E., Jannink, J. L., & Melchinger, A. E. (2013). Genomic predictability of interconnected biparental maize populations. Genetics, 194(2), 493-503.

Riedelsheimer, C., & Melchinger, A. E. (2013). Optimizing the allocation of resources for genomic selection in one breeding cycle. Theoretical and Applied Genetics, 126(11), 2835-2848.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). Semiparametric regression (No. 12). Cambridge university press, UK.

Schulz-Streeck, T., Ogutu, J. O., Karaman, Z., Knaak, C., & Piepho, H. P. (2012). Genomic selection using multiple populations. Crop Science, 52(6), 2453-2461.

Searle, S. R. (1971). Linear Models. John Wiley & Sons.

Siegmund, D., & Yakir, B. (2007). The statistics of gene mapping. Springer Science & Business Media.

Soller, M. (1978). The use of loci associated with quantitative effects in dairy cattle improvement. Animal Production, 27(02), 133-139.

Soller, M., & Plotkin-Hazan, J. (1977). The use marker alleles for the introgression of linked quantitative alleles. Theoretical and Applied Genetics, 51(3), 133-137.

Sorensen, D., & Gianola, D. (2002). Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media.

Thomson, M. (2014). High-throughput SNP genotyping to accelerate crop improvement. Plant Breeding and Biotechnology, 2(3), 195-212.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Tukey, J. W. (1949). Comparing individual means in the analysis of variance. Biometrics 5 (2), 99-114.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science, 92(1), 16-24.

Watson, J. D., & Crick, F. H. (1953). Molecular structure of nucleic acids. Nature, 171(4356), 737-738

Whittaker, J., Thompson, R., & Denham, M. (2000). Marker-assisted selection using ridge regression. Genetical Research, 75(02), 249-252.

Wong, C. K., & Bernardo, R. (2008). Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics, 116(6), 815-824.

Wright, S. (1921). Systems of mating. I. The biometric relations between parent and offspring. Genetics, 6(2), 111.

Zhao, Y., Gowda, M., Longin, F., Würschum, T., Ranc, N., & Reif, J. (2012). Impact of selective genotyping in the training population on accuracy and bias of genomic selection. Theoretical and Applied Genetics, 125(4), 707-713.

Zhong, S., Dekkers, J. C., Fernando, R. L., & Jannink, J. L. (2009). Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. Genetics, 182(1), 355-364.

9. ANEXO

PROGRAMAS

```
##############################################################################
# Aplicación de métodos de selección a una población y carácter particular
##############################################################################


# Definiciones previas ------------------------------------------------------

# Carga de librerias:
library(rrBLUP) # RR-BLUP
library(BLR) # métodos de selección genómica bayesianos
library(BGLR) # métodos de selección genómica bayesianos
library(glmnet) # RR y BL clásicos

# Objetos Cargados:
> impGeno[1:5,1:5]
            M00009437448 M00000202568 M00009439062 M00000139270 M00000213079
00000000001            1            1            1           -1            1
00000000002            0            0            0            0            0
00000000003            0            0            0            0            0
00000000004            0            0            0            0            0
00000000005            0            0            0            0            0

> Pheno_pred[1:5,]
         Hum    Peso  Rinde
[1,] -1.1043  0.4073  5.243
[2,] -0.8293  0.2987  2.717
[3,] -0.2435 -0.1347 -3.077
[4,]  0.5337  0.3968 -6.324
[5,] -1.4391  0.5760  2.055


# Definiciones para el ajuste modelos SG ------------------------------------

# Definiciones Datos
X=as.matrix(impGeno)
Y=as.matrix(Pheno_pred[,Traits])
N<-nrow(X)
p<-ncol(X)

# Conjunto Entrenamiento y Validación
set.seed(1235)
tst<-sample(1:N,size=round(N*.25),replace=FALSE)
XTRN<-X[-tst,];dim(XTRN)
XTST<-X[tst,];dim(XTRN)
t=1
y<-Y[,t]
yTRN<-y[-tst]
yTST<-y[tst]


# SMC: Selección de Variables y ajuste por mínimos cuadrados ----------------

# Regresión por marcador individual
SNPestimate=numeric()
SNPes=numeric()
SNPt=numeric()
pValues=numeric()
for(i in 1:p){
  fm<-lm(yTRN~XTRN[,i])
  SNPestimate[i]<-summary(fm)$coef[2,1]
  SNPes[i]<-summary(fm)$coef[2,2]
  SNPt[i]<-summary(fm)$coef[2,3]
  pValues[i]<-summary(fm)$coef[2,4]
  print(paste('Fitting Marker ',i,'.',sep=''))
}
  names(pValues)=colnames(XTRN)
```

```
One_MM_results=merge(MAP,data.frame(marker=names(pValues),pVal=pValues),by="marke
r")
  One_MM_results=One_MM_results[order(One_MM_results$"cr",One_MM_results$"pos"),]

  myRanking<-order(pValues) # da el ranking del MM (1= menor p-value)
  rbind(pValues[1:14],myRanking[1:14])
  CorTRN<-numeric()
  CorTST<-numeric()
  for(i in 1:(min(p,round(.75*N)))){
    tmpIndex<- myRanking[1:i]
    fm<-lm(yTRN~XTRN[,tmpIndex])
    CorTRN[i]<-cor(yTRN,predict(fm))
    bHat<-coef(fm)[-1] ; bHat<-ifelse(is.na(bHat),0,bHat)
    yHat<-as.matrix(XTST[,tmpIndex])%*%bHat
    CorTST[i]<-cor(yTST,yHat)
    print(paste('Fitting Model with ',i,' markers!',sep=''))
  }

ResultsSMC=cbind(Project=ID,Trait=Traits[t],
                      maxCor_TRN=max(CorTRN),NoMkr=which.max(CorTRN),
                      Cor_TST=CorTST[which.max(CorTRN)],
                      meanCor_TRN=mean(CorTRN),meanCor_TST=mean(CorTST))


# RR: Regresión de Ridge Clásica ------------------------------------------------

# funcion f_lambda con TST incluido
f_lambda = function(lambda, C,rhs,XTRN,yTRN,XTST,yTST) {
  # adds lambda to the diagonal of C (starts at 2)
  for(j in 2:ncol(C)){C[j,j]<-C[j,j]+lambda}
  CInv<-chol2inv(chol(C))
  sol<-crossprod(CInv, rhs)
  yHatTRN<-XTRN%*%sol
  CorTRN<-cor(yTRN,yHatTRN)
  yHatTST<-XTST%*%sol
  CorTST=cor(yTST,yHatTST)
  print(c(lambda,CorTST))
  print(c(lambda,CorTRN)); CorTRN
}

  CorTRN<-numeric(); CorTST<-numeric()
  C0<-crossprod(XTRN)
  rhs<-crossprod(XTRN,yTRN)
  Lambda_Opt=optimize(f_lambda,c(5,5000),C=C0,rhs=rhs,XTRN=XTRN,yTRN=yTRN,
                      yTST=yTST,XTST=XTST,tol=10,maximum=T)
  Lambda_Opt

    C<-C0
    # adds lambda to the diagonal of C (starts at 2)
    for(j in 2:ncol(C)){   C[j,j]<-C[j,j]+Lambda_Opt[[1]]   }
    CInv<-chol2inv(chol(C))
    sol<-crossprod(CInv, rhs)
    yHatTRN<-XTRN%*%sol
    CorTRN<-cor(yTRN,yHatTRN)
    if (round(CorTRN,4)!=round(Lambda_Opt[[2]],4)) stop("issue with Optimize!" )
    yHatTST<-XTST%*%sol
    CorTST<- cor(yTST,yHatTST);CorTST

ResultsRR=cbind(Project=ID,Trait=Traits[t],maxLambda=Lambda_Opt[[1]],
                      maxCorr_TRN=CorTRN[1],Corr_TST=CorTST[1])
```

```
# RR-BLUP: Regresión de Ridge BLUP ---------------------------------------

#predict marker effects
model=mixed.solve(y=yTRN,Z=XTRN,method="REML", SE=TRUE)
model$Vu;model$Ve;model$Ve/model$Vu;log(model$Ve/model$Vu,10)
lambda_RRblup=model$Ve/model$Vu;

ResultsRR_BLUP=cbind(Project=ID,Trait=Traits[t],lambda_RRblup=lambda_RRblup,
                     Corr_TRN=cor(XTRN%*%model$u,yTRN)[1],
                    Corr_TST=cor(XTST%*%model$u,yTST)[1])

# BRR: Regresión de Ridge Bayesiana ----------------------------------------
nIter=150000
burnIn=50000
thin=10
ETA=list(MRK=list(X=XTRN, model="BRR",S0=1,df0=4)) # dfb=5 by default

fm_BRR<-BGLR(y=yTRN,response_type = "gaussian",
             ETA=ETA,nIter=nIter,burnIn=burnIn,
             saveAt=paste("BRR",Traits[t],pop,sep="-"),thin=thin)
fm_BRR$varE

### Gráfico convergencia varE
varE_serie<-scan(file=paste(paste("BRR",Traits[t],pop,sep="-
"),"varE.dat",sep=""))
xyplot(varE_serie~seq(1,nIter,by=thin),type="o",
       xlab=list(cex=1.5,label="Iteración"),
       ylab=list(cex=1.5,label="Var Residual"),
       scales=list(cex=1.5,alternating=F),
       panel=function(...){
         panel.grid(...,h=-1,v=0,lty=2,col.line="grey")
         panel.xyplot(...,col=1,ylim=c(0,0.7))
         panel.abline(v=50000,col=2,lty=3)
         panel.abline(h=fm_BRR$varE,col=2,lwd=2)})

### Gráfico convergencia varB
varB_serie<-scan(file=paste(paste("BRR",Traits[t],pop,sep="-
"),"ETA_MRK_varB.dat",sep=""))
varB=mean(varB_serie[(burnIn/thin+1):length(varB_serie)])
xyplot(varB_serie~seq(1,nIter,by=thin),type="o",
       xlab=list(cex=1.5,label="Iteración"),
       ylab=list(cex=1.5,label="Var MM"),
       scales=list(cex=1.5,alternating=F),
       panel=function(...){
         panel.grid(...,h=-1,v=0,lty=2,col.line="grey")
         panel.xyplot(...,col=1,ylim=c(0,0.4))
         panel.abline(v=50000,col=2,lty=3)
         panel.abline(h=varB,col=2,lwd=2)})

z_coef=fm_BRR$ETA$MRK$b/fm_BRR$ETA$MRK$SD.b
names(z_coef)=fm_BRR$ETA$MRK$colNames
lambda_BRR=fm_BRR$varE/varB;log(lambda_BRR,10)

ResultsBRR=cbind(Project=ID,Trait=Traits[t],
                 lambda_BRR_BGLR=lambda_BRR,
                 corTRN=cor(yTRN,XTRN%*%fm_BRR$ETA$MRK$b)[1],
                 corTST=cor(yTST,XTST%*%fm_BRR$ETA$MRK$b)[1],
                 varE=fm_BRR$varE,varBR=varBR))


# BLR: Regresión LASSO Bayesiana lambda fijo -------------------------------------
h2<-0.8369; df0<-4; K=0
for(i in 1:ncol(X)){ K<-K+var(X[,i])}
lambda2_BL<-2*K*(1-h2)/h2

ETA=list(MRK=list(X=XTRN,
model="BL",lambda=sqrt(lambda2_BL),type="FIXED",S0=1,df0=4))
fm_BLf<-BGLR(y=yTRN,response_type = "gaussian",
             ETA=ETA,nIter=nIter,burnIn=burnIn,
             saveAt=paste("BLf",Traits[t],pop,sep="-"),thin=10)
### Gráfico convergencia varE
varE_serie<-scan(file=paste("BLf",Traits[t],pop,"varE.dat",sep="-"))
xyplot(varE_serie~seq(1,nIter,by=thin),type="o",
       xlab=list(cex=1.5,label="Iteración"),
```

```
       ylab=list(cex=1.5,label="Var Residual"),
       scales=list(cex=1.5,alternating=F),
       panel=function(...){
         panel.grid(...,h=-1,v=0,lty=2,col.line="grey")
         panel.xyplot(...,col=1,ylim=c(0,0.4))
         panel.abline(v=50000,col=2,lty=3)
         panel.abline(h=fm_BLf$varE,col=2,lwd=2)})

z_coef=fm_BLf$ETA$MRK$b/fm_BLf$ETA$MRK$SD.b
names(z_coef)=fm_BLf$ETA$MRK$colNames

varB=2*fm_BLf$varE/lambda2_BL
par_DE=lambda2_BL/fm_BLf$varE;par_DE

ResultsBLR=cbind(Project=ID,Trait=Traits[t],
                    lambda2_BLf=lambda2_BL,
                    corTRN=cor(yTRN,XTRN%*%fm_BLf$ETA$MRK$b)[1],
                    corTST=cor(yTST,XTST%*%fm_BLf$ETA$MRK$b)[1],
                    varE=fm_BLf$varE,varB=varB))


# BL: Regresión LASSO clásica  ------------------------------------------------
-

# seleccion de lambda para LASSO
cv.glmmod <- cv.glmnet(x=XTRN,y=yTRN,family = "gaussian",alpha=1, nfolds=10,
                       standardize=FALSE,intercept=TRUE,type.gaussian="naive")
plot(cv.glmmod)
best_lambda <- cv.glmmod$lambda.min

fm_LASSO<-glmnet(x=XTRN,y=yTRN,family = "gaussian",alpha=1, # LASSO (alpha=0 es
RR)
                 standardize=FALSE,intercept=TRUE,lambda=best_lambda)
corTRN=cor(predict(fm_LASSO,XTRN),yTRN)
corTST=cor(predict(fm_LASSO,XTST),yTST)

coef=coef(fm_LASSO)[-1]

ResultsLASSO=cbind(Project=ID,Trait=Traits[t],
                    lambda_LASSO=best_lambda,
                    corTRN=corTRN,
                    corTST=corTST))
```