Imputación de genotipos y detección de regiones genómicas aditivas para caracteres de crecimiento y deposición de grasa en una población cruza porcina

Tesis presentada para optar al título de Doctor de la Universidad de Buenos Aires, Área Ciencias Agropecuarias

José Luis Gualdrón Duarte
Zootecnista – Universidad Nacional de Colombia (Sede Bogotá)
-2007-

Lugar de trabajo: Facultad de Agronomía, Universidad de Buenos Aires

Escuela para Graduados Ing. Agr. Alberto Soriano
Facultad de Agronomía – Universidad de Buenos Aires

COMITÉ CONSEJERO

Director de tesis
Rodolfo Juan Carlos Cantet
Ingeniero Agrónomo., Universidad de Buenos Aires, Argentina.
MSc. Montana State University, Estados Unidos de América.
MSc. University of Illinois, Estados Unidos de América.
Ph.D., University of Illinois, Estados Unidos de América.

Consejero de Estudios
Daniel Omar Maizón
Médico Veterinario, Universidad de Buenos Aires, Argentina.
MSc., Universidad de Buenos Aires, Argentina.
Ph.D., University of Cornell, Estados Unidos de América.

JURADO DE TESIS

JURADO
Pablo Corva
Ingeniero Agrónomo., Universidad Nacional de la Plata, Argentina.
MSc., Universidad Nacional de la Plata, Argentina.
Ph.D., University of California Davis, Estados Unidos de América.

JURADO
César López
Ingeniero Agrónomo., Universidad Nacional de Lomas de Zamora, Argentina.
MSc., Universidad Nacional de Rosario, Argentina.
Ph.D., Oregon State University, Estados Unidos de América.

JURADO
Miguel Perez Encizo
Biologo., Universidad Complutense de Madrid, España.
Ph.D., Universidad Complutense de Madrid, España.

Fecha de defensa de la tesis: 17 de Febrero de 2016

A mis padres y hermanos, uno de mis grandes pilares en la vida. A Jenny por su amor y paciencia, y a todos aquellos familia y amigos que cerca o lejos me estuvieron apoyando.
A todos gracias.

AGRADECIMIENTOS

La presente tesis fue un camino de aprendizaje y formación de conocimientos a nivel personal y académico, del cual hacen parte un gran grupo de personas. Agradezco principalmente a mi tutor y director de tesis Dr. Rodolfo Juan Carlos Cantet por su amistad, dedicación y el aprecio de enseñarme el proceso de ser investigador.

También, a los docentes de la cátedra de Mejoramiento Animal: Ana Birchmeir, Valeria Schindler, Mónica Santos Cristal, Laura Pruzzo y su personal administrativo Oscar Rhodas y Martín Mosbruker.

Al Dr. Daniel Omar Maizón, por su colaboración siendo mi consejero de estudios en mi proyecto de tesis.

Agradezco igualmente al Dr. Juan Pedro Steibel, por el aprendizaje, la paciencia y el apoyo ofrecido en el desarrollo de mi tesis. De igual manera a sus colaboradores de grupo de Michigan State University: Ron Bates, Cathy Ernst, Nancy Raney. Y a sus estudiantes: María Arceo, Yvonne Badke, Yijian Huang y Pablo Reeb con los que compartí parte de mi proceso formativo. También, a mis compañeros de casa Juan David Muñoz, Eduardo Rico e igualmente a María Arceo que hicieron de mi estadía una experiencia gratificante.

A mis compañeros y amigos del grupo de Mejoramiento Animal y de oficina: Yeni Bernal, Carolina García Baccino, Natalia Forneris, Sebastian Munilla, Juan David Corrales y Andrés Rogberg, por su amistad, anécdotas y compañía que serán parte de la historia de mi vida.

A Ricardo José Hernández Duarte por su hermandad, apoyo y compañía en todos estos años. Igualmente, a Santiago Bernal Zuñiga por su amistad y ayuda con gráficos finales.

A la Escuela para Graduados (EPG) Alberto Soriano - Facultad de Agronomía UBA, personal docente y administrativo. Por toda sus enseñanzas y colaboración.

A las agencias FONCyT - CONICET (Argentina) y COLCIENCIAS (Colombia) por financiar mi doctorado, y por consiguiente la culminación de mi tesis doctoral.

Declaro que el material incluido en esta tesis es, a mi mejor saber y entender, original producto de mi propio trabajo (salvo en la medida en que se identifique explícitamente las contribuciones de otros), y que este material no lo he presentado, en forma parcial o total, como una tesis en ésta u otra institución.

José Luis Gualdrón Duarte

PUBLICACIONES DERIVADAS

1. Gualdrón Duarte JL, Bates RO, Ernst CW, Raney NE, Cantet RJC, Steibel JP. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genetics 2013,14:38.

2. Gualdrón Duarte JL, Cantet RJC, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transforming genomic evaluations. BMC Bioinformatics 2014,15:246.

3. Gualdrón Duarte JL, Cantet RJC, Bernal Rubio YL, Bates RO, Ernst CW, Raney NE, Rogberg-Muñoz A, Steibel JP. Refining genome-wide association for growth and fat deposition traits in an F2 pig population. 2016 (Aceptado: Journal of Animal Science)

ÍNDICE GENERAL

Página

ÍNDICE DE CUADROS

CUADRO **Página**

ÍNDICE DE FIGURAS

FIGURA **Página**

ABREVIATURAS

ADN	Ácido desoxirribonucleico;
BLUP	Mejor Predictor Lineal Insesgado de Mínima Varianza;
e.g.	Acrónimo del latín *exempli gratia* (‘dado como ejemplo’). Se utiliza para indicar ‘véase, e.g.…’;
GEBV	Valor de cría genómico;
HD	Genotipado en alta densidad;
IA	Exactitud de imputación;
i.e.	Acrónimo del latín *id est* (‘esto es’). Se utiliza para indicar ‘es decir, …’;
LD	Desequilibrio de Ligamiento;
LE	Equilibrio de Ligamiento;
LowD	Genotipado en baja densidad;
MAF	Frecuencia Alélica Menor;
MAS	Selección Asistida por Marcadores;
ML	Máxima verosimilitud;
MME	Ecuaciones del modelo mixto;
PEV	Varianza del Error de Predicción;
QTL	Locus de un carácter cuantitativo;
REML	Máxima verosimilitud restringida;
SNP	Polimorfismo de un único nucleótido;
tagSNP	SNP representativo de una región genómica;

Título: Imputación de genotipos y detección de regiones genómicas aditivas para caracteres de crecimiento y deposición de grasa en una población cruza porcina

RESUMEN

Los estudios de asociación genómica (GWAS) llevan consigo un alto costo monetario, y a su vez requieren algoritmos complejos de análisis de información que consumen tiempo y memoria computacional. En este sentido, el objetivo principal de esta tesis es presentar un esquema de genotipado apropiado para poblaciones cruza, junto con un algoritmo eficiente para GWAS de caracteres complejos productivos. Inicialmente, se presenta un esquema de genotipado que maximiza la exactitud de imputación de genotipos en alta densidad (HD) a partir de paneles de baja densidad (LowD), reduciendo el costo de genotipificación. Posteriormente, se propone un algoritmo que facilita identificar regiones genómicas que explican parte de la variabilidad de un carácter, reduciendo la tasa de falsos positivos, el tiempo de cálculo y el requerimiento de memoria RAM. De igual manera, el algoritmo evalúa segmentos candidatos a partir de las posiciones detectadas significativas y calcula la fracción de la varianza aditiva total explicada por cada segmento. Finalmente, se presentan estudios de asociación para características de crecimiento y deposición de grasa, empleando el algoritmo propuesto junto con genotipos imputados en HD. La implementación de dicho algoritmo permite identificar regiones significativas relevantes y genes candidatos que explican parte de la variación de los caracteres evaluados. En conclusión, la tesis propone un enfoque estructurado, práctico y eficiente para la realización de GWAS de caracteres complejos aplicado en poblaciones experimentales con fines productivos.

Palabras claves: cerdo, imputación de genotipos, estudio de asociación, varianza del efecto de marcador, deposición de grasa, crecimiento.

Title: Genotype imputation and detection of additive genomic regions for growth and fat deposition traits in a cross pig population.

ABSTRACT

The genomic wide association studies (GWAS) involve high cost and complicated algorithms that concern a high memory requirements and considerable computing time. In that sense, the main goal of this thesis is to present a specified genotyping scheme for experimental crossbred population, together with an efficient GWAS algorithm for complex traits. Initially, a genotyping scheme is presented. The main advantages of this scheme are that it maximizes the genotype imputation accuracies in high density (HD) from genotypes in low density (LowD) and, consequently, reduces the genotyping cost. Next, an algorithm to identify significant genomic locations associated with the expression of complex traits is proposed. This algorithm reduces the number of false positives, the memory requirements and the computing time. Moreover, a candidate segment approach can be carried out from the significant positions to estimate the proportion of variance explained by each segment. Finally, a GWAS for growth and fat depositions traits is implemented using HD imputed genotypes and the algorithm. As a result, significant genomic regions and candidate genes for trait expressions are identified. In conclusion, the research shows a practical and efficient approach for GWAS applied to cross population with complex animal production traits.

Key words: pig, genotype imputation, genome wide association, marker effect variance, fat deposition, growth.

CAPÍTULO 1

Introducción general

1.1. Introducción

En la última década, y gracias al desarrollo de nuevas técnicas de la biología molecular fue posible secuenciar el genoma de distintas especies. Este avance científico permitió profundizar el conocimiento sobre los mecanismos de herencia de caracteres complejos de modo de poder explorar el nivel de control o expresión explicado por una región genómica específicamente delimitada. Por lo tanto, la búsqueda de genes que codifican para un carácter de interés ó QTLs (Quantitative Trait Loci) actualmente, se realiza mediante una panel de alta densidad de marcadores SNP (del inglés: "Single Nucleotide Polymorphisms") que se encuentran localizados a lo largo del genoma entero. Estos marcadores varían en una posición física conocida dentro del genoma, es decir la secuencia de ADN, diferenciándose en un nucleótido (adenina, timina, guanina o citosina). Dado que se dispersan por todo el genoma, la expectativa es que el marcador SNP está asociado ("ligado") al QTL, siendo el fenómeno descripto por la expresión desequilibrio gamético o, más comúnmente, como "desequilibrio por ligamiento" (LD, por su sigla en inglés: "Linkage disequilibrium"). El LD es la asociación estadística de falta de independencia entre variables aleatorias asociadas con esos SNP y/o QTLs. La intensidad del LD decrece en proporción al número de generaciones por la recombinación, la cual "desarma" los haplotipos ancestrales (Reich *et al.*, 2001). Si bien el LD es medido variadamente, es común el empleo del estadístico r^2 (Hill y Robertson, 1966) que informa sobre el grado de asociación entre dos pares de loci en una escala de 0 a 1, correspondiendo el valor 0 la situación de independencia entre loci. En el otro extremo, el valor 1 refleja la situación en que ambos loci se encuentran en completo LD, lo que indica que son transmitidos conjuntamente entre generaciones sucesivas. La estrategia de detección de QTL es indagar sobre la fracción de la varianza genética aditiva de un carácter cuantitativo que se encuentra asociada con cada SNP.

Inicialmente los mejoradores animales emplearon el ligamiento marcador-QTL en lo que se denominó "selección asistida por marcadores" (MAS, Marker assisted Selectión, Lande y Thompson, 1990) para mapear QTLs dentro del genoma. Sin embargo, este mapeo es muy limitado porque requiere un gran número de familias de medio-hermanos como de individuos por familia, de modo tal que las posiciones de los QTLs putativos dentro del cromosoma posean intervalos de confianza razonables. Además, MAS es problemática cuando el ligamiento entre el QTL y el marcador no sea lo suficientemente intenso para asegurar que la relación persista entre generaciones y no se quiebre por recombinación. Este problema se atenúa con una gran densidad de marcadores para captar más LD y mayores posibilidades de detectar QTL (Hayes, 2007).

Para evitar buscar asociaciones marcadores - QTL y utilizar toda la información disponibles Meuwissen *et al.* (2001) propusieron regresar datos o predicciones del mérito genético sobre todos los marcadores SNP dispersos por todo el genoma disponibles, metodología conocida como "*selección genómica*". Dicho panel de marcadores es conocido como de alta densidad (HD). Los valores de cría genómicos (GEBV) son calculados sumando las soluciones de cada posición en relación con el genotipo observado de cada individuo. Schaeffer *et al.* (2006) y VanRaden *et al.* (2009) observaron que cuanto mayor es el número de SNP incluidos en el análisis, mayor fue la exactitud de predicción del GEBV que se obtiene del carácter (estadístico que depende del carácter analizado), calculada por validación cruzada. Sin embargo, más SNP en el panel implican un costo mayor que económicamente es injustificable salvo para un

número reducido de candidatos a la selección. Si bien los costos de los paneles HD disminuyeron en los últimos años, aún son del orden de los 40-60 u$d por animal haciendo difícil realizar un proyecto de investigación con más de 500 – 1000 individuos, un número mínimo como para obtener un grado de precisión de mapeo razonable. En algunos casos, la relación costo-beneficio es negativa, afectando la implementación de esta útil herramienta. Un modo de abaratar los costos de evaluación es emplear un dispositivo de baja densidad (LowD) que, con un número de SNP menor condense la mayor parte de la información que posee el panel HD. Este planteo requiere de algoritmos que identifiquen grupos de marcadores altamente representativos, es decir no redundantes por el LD o independientes, de modo de poder predecir más precisamente el valor de cría (Habier *et al*, 2009). En adición, se podría utilizar paneles LowD para la "imputación", o predicción de genotipos no observados en LowD que se formen parte de un panel HD, utilizando el LD (Badke et al., 2013) de la población bajo estudio, o alternativamente a través de la información de pedigrí (Daetwyler *et al.*, 2011; Druet y Georges, 2010; Hickey et al., 2011; Hickey *et al.* 2012; Huang et al., 2012)**.** En tal sentido es posible conseguir genotipos HD a partir de dispositivos LowD con una alta exactitud de imputación, siendo más rentable la implementación de esta técnica y, además, permitiendo contar con un número de animales genotipados en HD mayor, aumentando así la potencia de los análisis de asociación en caracteres complejos (Dikmen *et al.*, 2013; Do *et al.*, 2014; B. Fan *et al.*, 2011; Hayes *et al.*, 2010) relevantes a la industria pecuaria y también en estudios de asociación con enfermedades en humanos (Lee *et al.*, 2011). Consecuentemente, en esta tesis se propone 1) desarrollar una metodología para imputar genotipos en LowD y obtener HD de modo de 2) proponer un algoritmo altamente eficiente para estudios de asociación genómica empleando los paneles imputados. Finalmente, 3) emplear los genotipos imputados en HD y el algoritmo propuesto, para realizar un análisis en caracteres de crecimiento y deposición de grasa en una población experimental de cerdos Duroc × Pietrain, en la búsqueda de regiones que significativamente se asocien con la variación aditiva de dichos caracteres.

Esta tesis contiene cinco capítulos siendo primero el introductorio, luego en segundo se evalúa la exactitud de imputación de paneles en LowD, creados a partir de la selección de SNP en LD y sus posiciones físicas (SNP aproximadamente equidistantes), mediante simulación estocástica. Se estima además la exactitud de imputación de diferentes esquemas de genotipado utilizando paneles en HD (60K) y en LowD (Comercial 9K) de las distintas generaciones de la población cruza. En el tercer capítulo se desarrolla un algoritmo de GWAS, cuya novedad radica en la estandarización de los efectos de los SNP empleando la varianza particular de cada marcador en vez de un parámetro de dispersión único para todos los marcadores, hecho que genera una reducción en la tasa de falsos positivos del análisis, siendo computacionalmente eficiente en tiempo y memoria RAM. Posteriormente, en el cuarto capítulo se emplea dicho algoritmo con los genotipos imputados (Capítulo 2) y no imputados en HD, para realizar un análisis de asociación para caracteres de crecimiento y de deposición de grasa en la población cruza, a los efectos de detectar regiones genómicas que significativamente se asocien con la expresión de los caracteres evaluados. Finalmente, en el quinto capítulo se enumeran las conclusiones generales de la tesis.

CAPÍTULO 2

Exactitud de imputación genotípica en una población de cerdos F_2 usando paneles SNP de alta y baja densidad [2]

[2] Gualdrón Duarte JL, Bates RO, Ernst CW, Raney NE, Cantet RJC, Steibel JP. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genetics 2013, 14:38.

2.1. Introducción

La búsqueda de regiones en el genoma que contengan variantes asociadas con caracteres productivos requiere evaluar poblaciones experimentales donde se identifican los QTL segregando "dentro" y "entre" poblaciones parentales (Edwards et al., 2008 (a)). Frecuentemente, dichas poblaciones son las F_2, porque permiten detectar QTL que segregan luego del primer cruzamiento entre líneas divergentes (Choi et al., 2010; Haley y Elsen, 1994). Los análisis de asociación y de evaluación genética empleando información genómica requieren una gran cantidad de individuos con fenotipos más genotipos en alta densidad (HD) dispersos a lo largo del genoma, para poder obtener resultados confiables (Hickey et al., 2012). Sin embargo, el costo es una limitación importante al aplicar el genotipado en HD con una gran cantidad de animales (Anderson et al., 2008; Habier, et al., 2009). Una manera de reducir este costo consiste en genotipar en HD sólo a aquellos individuos de generaciones iniciales, mientras que su descendencia es genotipada con un panel de baja densidad (LowD) (Habier et al., 2009; Huang et al., 2012). Dicho panel está integrado por cierto número de SNP (llamados "tagSNP") selectos que provienen del panel de HD, para luego predecir (o "imputar") los genotipos en HD con una elevada exactitud (Huang et al., 2012). Condicionando en los genotipos parentales y de los abuelos, este método puede generar exactitudes mucho mayores que las obtenidas usando un panel de referencia no relacionado (Hickey et al., 2012; Huang et al., 2012; Zhang y Druet, 2010), porque conociendo la fase de los alelos paternos (Huang et al., 2012) se pueden inferir los genotipos en HD de la progenie genotipada en LowD, empleando probabilidades de segregación o descendencia dentro de familia sobre los genotipos en HD de generaciones base (Habier et al., 2009).

La mayoría de los estudios de imputación de genotipos realizados en especies pecuarias fueron hechos con razas puras (Hayes et al., 2011; Hickey et al., 2012; Huang et al., 2012; Huang et al., 2012; Zhang y Druet, 2010), mientras que en poblaciones cruzas dichos estudios se encuentran ausentes. En plantas se investigó imputando líneas puras y recombinantes (RILs) mediante diseños de estudios de asociación anidados (NAM, Poland et al., 2011; Tian et al., 2011) y en estudios con líneas multi-parentales y generaciones avanzadas de cruzamiento (MAGIC, Kover et al., 2009). A su vez, en humanos se imputaron genotipos para analizar pruebas de asociación por ligamiento de reguladores transcripcionales con expresión génica (Burdick et al., 2006). También, en modelos animales de investigación biomédica *e.g.* ratones, se imputaron genotipos de líneas cruza de modo de poder identificar genes en caracteres complejos (Bouxsein et al., 2004; Leduc et al., 2011). La imputación de genotipos en humanos, animales domésticos, plantas o animales de laboratorio es similar, en el sentido que un número reducido de individuos fundadores son genotipados en HD, en tanto que la mayor parte de la población es evaluada en LowD, para luego utilizar la información del ligamiento de modo de estimar los genotipos faltantes. En este capítulo se realizará un análisis de imputación de genotipos F_2 en baja densidad, dentro de una población con tres generaciones (F_0, F_1 y F_2) resultante de una cruza experimental Duroc × Pietrain, donde los animales F_0 y F_1 fueron todos genotipados en HD (60K). Las poblaciones F_0 utilizadas para mapear QTL en porcinos típicamente cuentan con un pequeño número de fundadores. El presente estudio no es una excepción y la F_0 está compuesta por 4 machos y 15 hembras (Edwards *et al*., 2008 (a)). En tal sentido, se espera que haya ocurrido un limitado número de recombinaciones en el genoma durante las primeras generaciones (Mackay, 2001). Por lo tanto, dichas poblaciones tendrán una baja resolución en la precisión del mapeo de QTLs (Mackay, 2001). Sin embargo, y por la

misma razón, la falta de recombinación induce una exactitud de imputación elevada. Este último efecto puede ser provechoso para imputar genotipos en HD a partir de F_2 genotipada en LowD de muy bajo costo, y que subsecuentemente puede combinarse, junto con varias poblaciones F_2 experimentales existentes, en un estudio de meta-análisis de asociación. Existen varias razones que hacen atractiva esta posibilidad. Primero, las poblaciones experimentales F_2 han sido creadas recientemente (Liu et al., 2007; Sanchez et al., 2007; Yang et al., 2008; Nonneman et al., 2011) y su material se encuentran fácilmente disponible. En segundo lugar, dichas poblaciones fueron fenotipadas para caracteres costosos de difícil disponibilidad, por ejemplo, el contenido de grasa intramuscular y la composición de ácidos grasos (Sanchez *et al*., 2007), la edad a la pubertad en cerdas (Yang *et al*., 2008), la terneza en calidad de carne (Meyers *et al.*, 2007). Un argumento adicional sobre el empleo de poblaciones cruzas es que, siendo generalmente formadas a partir de razas que muestran divergencia en los caracteres de interés zootécnico, *e.g.* porcentaje de carne magra, calidad de carne o eficiencia reproductiva, etc, muestran diversidad para combinar caracteres favorables. Ejemplo de dichas cruzas son Duroc × Pietrain (Liu et al., 2007; D. B. Edwards et al., 2008 (a), (b)), Duroc × Landrace (Nonneman et al., 2011), Duroc × Large-White (Sanchez et al., 2007), White-Duroc × Erhualian (Yang et al., 2008), Meishan × Duroc (Sato et al., 2003), Berkshire × Duroc (Stearns et al., 2005).

Dado todo lo expuesto anteriormente, es posible señalar que la imputación de genotipos en baja densidad (LowD) para que resulten de HD a partir de genotipos F_2 es muy útil y conveniente, porque mejora la relación costo-beneficio de genotipar, y es una estrategia muy empleada en los estudios de asociación. Se han empleado distintos métodos para la selección de tagSNP con paneles de baja densidad (LowD): 1) imponer restricciones sobre el valor mínimo de desequilibrio de ligamiento (LD) o relación entre marcadores (r_t^2) (Xu *et al.*, 2007), 2) selección de tagSNP equidistantes, basándose en la distancia física entre marcadores (Hayes et al., 2011; Hickey *et al.*, 2012; Huang *et al.*, 2012). En adición, existen también chips comerciales de densidad media que contienen SNP que segregan en un grupo importante de poblaciones, tales como el bovino (Wiggans *et al.*, 2012) y el cerdo (Badke *et al.*, 2013). Adicionalmente, es de interés evaluar el número de SNP necesario para obtener una alta exactitud de imputación en una F_2 en particular, y si es justificable diseñar un chip específico, o si se pueden utilizar chips comerciales. Por otra parte, es necesario evaluar si deben genotiparse las generaciones F_0 y F_1 en HD, o si solo genotipando la F_0 en HD puede resultar en altos valores de exactitud de imputación para la generación F_2. El objetivo de este capítulo es estimar la exactitud de imputación (*IA*) a HD (60K), a partir de genotipos F_2 LowD provenientes de una cruza de cerdos Duroc × Pietrain, empleando diferentes esquemas de genotipado. La estrategia considerada consistió en evaluar por simulación de Monte Carlo condicional en los genotipos de los animales de las dos primeras generaciones (F_0 y F_1). En el desarrollo de la simulación se probaron dos métodos de selección de tagSNP, y se compararon los resultados con aquellos obtenidos empleando un chip comercial (9K). Adicional a la simulación, se evaluó la *IA* usando datos experimentales, aprovechando la existencia de un reducido número de animales F_2 genotipados en HD.

2.2. Métodos

2.2.1. Animales

Los datos utilizados para todos los análisis pertenecen a una población experimental cruza porcina desarrollada en Michigan State University Swine Teaching y Research Farm, East Lansing, Michigan, EEUU (D. B. Edwards et al., 2008 (a),(b)). Se seleccionó el semen de 4 F_0 reproductores machos Duroc para ser utilizado en 15 hembras F_0 Pietrain mediante inseminación artificial, técnica también utilizada para producir la F_2. Entre la progenie F_1, 50 hembras y 6 machos generaron 1259 F_2 lechones nacidos vivos en 142 camadas provenientes de 11 grupos de servicio. Las madres F_1 se cruzaron con reproductores F_1 de modo de evitar aparear hermanos enteros o medio-hermanos. Todos los protocolos de manejo de los animales fueron revisados y aprobados por "Michigan State University All University Committee on Animal Use and Care" (AUF# 09/03-114-00).

2.2.2. Genotipado y Control de Calidad de Datos

Se genotiparon un total de N = 411 animales (4 machos F_0 Duroc, 15 hembras F_0 Pietrain, 6 machos F_1, 50 hembras F_1 y 336 animales F_2) con el panel Illumina PorcineSNP60 (62163 SNP) Genotyping beadchip (Illumina Inc., Ramos et al., 2009) en un laboratorio comercial (GeneSeek, a Neogen Company, Lincoln, NE). Del total de marcadores M = 62163 SNP, se eliminaron 6422 SNP por tener posición física desconocida. Las inconsistencias mendelianas (≤ 0.01%) fueron tomadas como genotipos perdidos y, si un animal tenia genotipo faltante en más del 10% de los SNP (MIND > 0.10) fue descartado para el estudio. De la misma manera fueron descartados SNP que no estuvieran resueltos en al menos un 90% (GENO > 0.10) de los individuos genotipados. Se removieron marcadores fijos en la población de estudio, en este sentido, se descartaron aquellos marcadores cuya frecuencia alélica en el alelo menos observado (MAF) fuera inferior a 1% (MAF < 0.01). Después de aplicar los criterios de filtrado antes mencionados, se excluyeron 12 animales (1 F_1 y 11 F_2) por MIND > 0.10 , 3038 SNP por GENO > 0.10 y 10139 SNP por MAF < 0.01. Finalmente, el archivo quedó constituido con datos de 399 cerdos, cada uno con 45003 SNP. Este filtrado fue realizado por Badke *et al*. (2012), empleando el programa PLINKv1.07 (Purcell *et al*., 2007).

2.2.3. Simulación de Genotipos

Se realizó una simulación estocástica para evaluar dos enfoques de selección de tagSNP, y calcular la *IA* de los genotipos F_2 resultantes. Para ello se simularon los genotipos de 932 animales F_2, condicionalmente a los genotipos en alta densidad reales de 55 individuos F_1 (6 machos y 49 hembras). El modelo de la simulación fue el "*gene-dropping*", en el cual a través de un pedigrí real se obtuvieron gametas F_2 recombinantes y no recombinantes. Los haplotipos se estimaron con alta exactitud a partir de los genotipos parentales en la F_1 y de los 19 ancestros F_0 empleando el programa MERLIN (Abecasis *et al*., 2002). El número de recombinaciones en la F_1 fue simulado aleatoriamente muestreando de una distribución *Poisson* con media igual a la distancia del cromosoma en Morgans (M), asumiendo 1Mb = 1cM (Laval y Excoffier, 2004), y considerando una distribución *uniforme* y la función de mapeo de Haldane (Haldane, 1919; Cheema y Dicks, 2009; Leduc *et al*., 2011) de modo de obtener la posición de los "crossing-overs" con las gametas recombinantes. Así por ejemplo, se

simuló el cromosoma 12 con 1405 SNP posicionados a lo largo de 64.2 Mb, con una distancia promedio entre marcadores de 0.04573 Mb. Asumiendo una tasa de recombinación de 1cM por Mb (Ledur *et al.*, 2010), el número de recombinaciones en el cromosoma 12 se muestreó aleatoriamente a partir de una distribución *Poisson* con parámetro igual a 64.2 / 100 = 0.642. Posteriormente, se asignaron las gametas resultantes que llevaron dichas recombinaciones en los genotipos de la F_1, para formar la F_2.

2.2.4. Selección de tagSNP en datos simulados.

Para seleccionar tagSNP se emplearon dos métodos distintos. El primero es una metodología de selección estadística teniendo en cuenta el LD, que fuera propuesta por Carlson et al. (2004) e implementada en el programa FESTA (Qin et al., 2006). En este caso, un SNP puede pertenecer a un grupo de tagSNP o estar en LD con un SNP perteneciente al grupo de tagSNP, si supera un valor igual o mayor a un umbral especifico de r_t^2 (Badke et al., 2013). Se seleccionó un nivel mínimo de r_t^2 (entre 0.1 y 0.5) sobre la base de mediciones de pares de SNP en LD dentro de los haplotipos en la F_1. Por lo tanto, todos los SNP que estuvieron por encima de este umbral fueron seleccionados como tagSNP. El segundo método consistió en escoger SNP igualmente espaciados ("equiespaciados"). En consecuencia, un cromosoma fue dividido en k segmentos de igual tamaño. Acto seguido, el SNP que estuviese más cerca del centro del segmento fue seleccionado como tagSNP. En aquellos casos donde no hubiese SNP dentro del segmento no se seleccionaría ninguno, lo que conduce a tagSNP $\leq k$ en segmentos de aproximadamente igual tamaño.

2.2.5. Imputación de genotipos simulados

Empleando el algoritmo de Lander-Green del programa MERLIN (Abecasis et al., 2002) se imputaron los genotipos F_2 mediante la predicción de los SNP que no constituyen tagSNP, condicional a los marcadores observados. Por motivos computacionales, el pedigrí fue analizado de una camada a la vez con toda la información disponible de las tres generaciones (Haley y Elsen, 1994). Así, cada F_2 contó con datos de los cuatro abuelos F_0, los dos padres F_1 y un máximo de 10 animales emparentados F_2. Cuando el tamaño de camada tenía más de 10 progenies, se formó una nueva "familia" con los cuatro abuelos F_0, los dos padres F_1, y el número restante de animales F_2.

2.2.6. Imputación de genotipos experimentales.

Se imputaron los genotipos F_2 experimentales no observados mediante el programa AlphaImpute (Hickey et al., 2012). El algoritmo implementado en dicho software utiliza la información total de la población y LD dentro de familia, para lo cual se requiere ajustar los valores de los parámetros del programa. Dentro de estos parámetros, el tamaño de núcleo fue de 100, 150, 400 y 600 SNP junto con el parámetro de marcadores adyacentes al núcleo: 300, 400, 600 y 800 SNP respectivamente. A su vez, el parámetro de porcentaje de error de genotipado fue fijado en 0%, de modo de obtener un alto porcentaje de alelos con su fase correctamente calculada (Hickey *et al.*, 2011). En este caso, el algoritmo de AlphaImpute no tuvo restricciones de cómputo y se corrió con el pedigrí completo.

2.2.7. Cálculo exactitud de imputación (IA)

La *IA* en individuos F_2 se calculó para los datos simulados y los experimentales mediante dos estadísticos diferentes. El primero de ellos utiliza la diferencia de medias entre dosis alélica observada e imputada (Weigel et al., 2010; Zhang y Druet, 2010), tal cual se describe en la siguiente fórmula:

$$IA = 1 - \frac{1}{2\,N}\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left|g_{ij} - \hat{g}_{ij}\right| \qquad [2.1]$$

Definiendo como dosis alélica al número de copias del alelo de referencia con valores de 0, 1 y 2 para el homocigota de referencia, el heterocigota y el otro homocigota, respectivamente, en [2.1] N corresponde al número total de animales imputados, M_i representa al número de marcadores con genotipo observado en el animal i, g_{ij} es la dosis alélica simulada o experimental del marcador j observada en el animal i, y $\hat{g}_{ij}$ es la dosis que resulta al imputar.

El segundo método empleado para cuantificar *IA* es el cuadrado de la correlación existente entre los genotipos observados y simulados para cada uno de los alelos, y se calcula con el estadístico R^2 propuesto por Huang *et al.* (2009). Sea $\bar{\hat{g}}_{ij}$ el valor promedio de los genotipos imputados y sea $\bar{g}$ el valor promedio de los genotipos observados. Entonces, el estadístico R^2 es calculado mediante la siguiente expresión:

$$R^2 = \left[\frac{\sum_{i=1}^{N}\left(\hat{g}_{ij} - \bar{\hat{g}}_{ij}\right)\left(g_{ij} - \bar{g}\right)}{\sqrt{\sum_{j=1}^{N}\left(\hat{g}_{ij} - \bar{\hat{g}}_{ij}\right)^2 \sum_{i=1}^{N}\left(g_{ij} - \bar{g}\right)^2}}\right]^2 \qquad [2.2]$$

Al interpretarse como el cuadrado de un coeficiente de correlación, sus valores se encuentran en el intervalo [0, 1].

2.3. Resultados

2.3.1. Genotipos simulados

2.3.1.1. *Selección de tagSNP y exactitud de imputación de los genotipos F_2*

El cuadro 2.1 presenta el número de tagSNP seleccionados para distintos valores de LD en un tamaño de cromosoma simulado intermedio (cromosoma 12), reflejado por la medida de $r^2 = r_t^2$. A medida que aumenta el umbral de r_t^2 se selecciona un mayor número de tagSNP y se obtiene una mayor *IA*. A modo de ejemplo y con un $r_t^2 = 0.2$, se seleccionaron 79 tagSNP con una distancia promedio de 0.79 Mb y una *IA* = 0.970, pero cuando se utilizó un umbral $r_t^2 = 0.5$ se seleccionaron 399 tagSNP, con distancia promedio igual a 0.16 Mb, resultando en *IA* = 0.982.

Cuadro 2.1. Número de tagSNP para cada r^2 y su exactitud de imputación en el cromosoma 12.

r_t^2	[1] tagSNP	[2] Número de SNP en genoma	[3] Distancia promedio entre tagSNP (Mb)	[4] *IA*
0.1	33	1295	1,86	0,960
0.2	79	3100	0,79	0,970
0.3	158	6199	0,40	0,976
0.4	266	10436	0,24	0,980
0.5	399	15654	0,16	0,982

r_t^2: umbral de r^2. [1] Numero de tagSNP seleccionados mediante r_t^2. [2] Número equivalente de SNP que se necesitan para cubrir el genoma total, manteniendo una distancia promedio similar entre marcadores tomando como base la distancia entre los tagSNP seleccionados. [3] Distancia promedio en Mega bases entre los tagSNP seleccionados. [4] Exactitud de Imputación.

2.3.1.1. *tagSNP equiespaciados*

Se halló que las *IA* obtenidas a partir de tagSNP equiespaciados o seleccionados por LD mostraron similares resultados. Así, se observó que con un panel de 80 SNP aproximadamente equiespaciados en el cromosoma 12 simulado se obtuvo una *IA* de 0.973, la cual es comparable con el valor de *IA* = 0.970 a partir de 79 tagSNP selectos por el nivel de LD (r_t^2 = 0.2), equivalencia en *IA* que se mantuvo para otras densidades de tagSNP (Figura 2.1.).

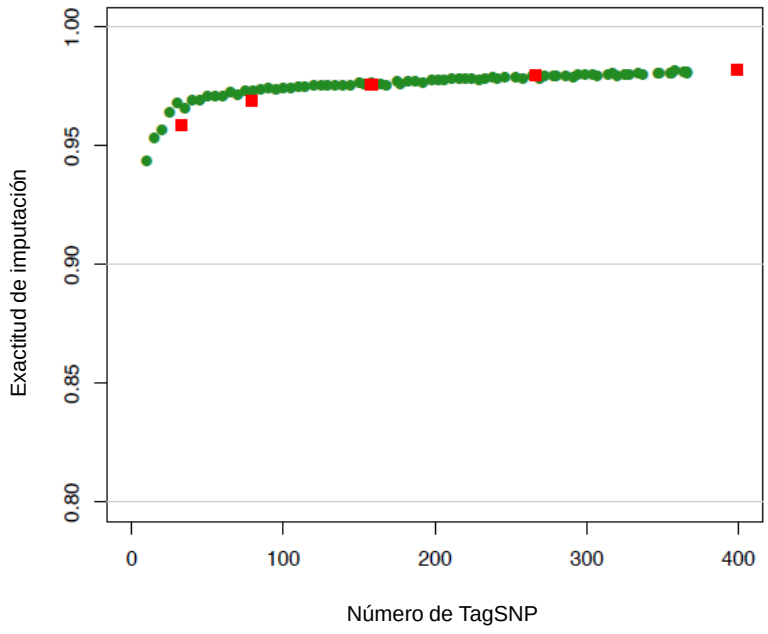


Figura 2.1. Exactitud de imputación mediante tagSNP selectos equidistantes ó por LD, en el cromosoma 12. Exactitud de imputación en función del número de tagSNP selectos equidistantes (puntos verdes) ó con una metodología estadística que emplea el LD (cuadrados rojos).

Asimismo, la *IA* empleando tagSNP equiespaciados fue similar para todos los cromosomas, asumiendo que se mantiene una distancia promedio entre tagSNP. Entonces, en los cromosomas 1 (140 tagSNP) y 12 (30 tagSNP) y para una distancia promedio entre marcadores consecutivos igual a 2.1 Mb, se obtuvieron *IA* iguales a 0.969 y 0.968, respectivamente (Figura 2.2).

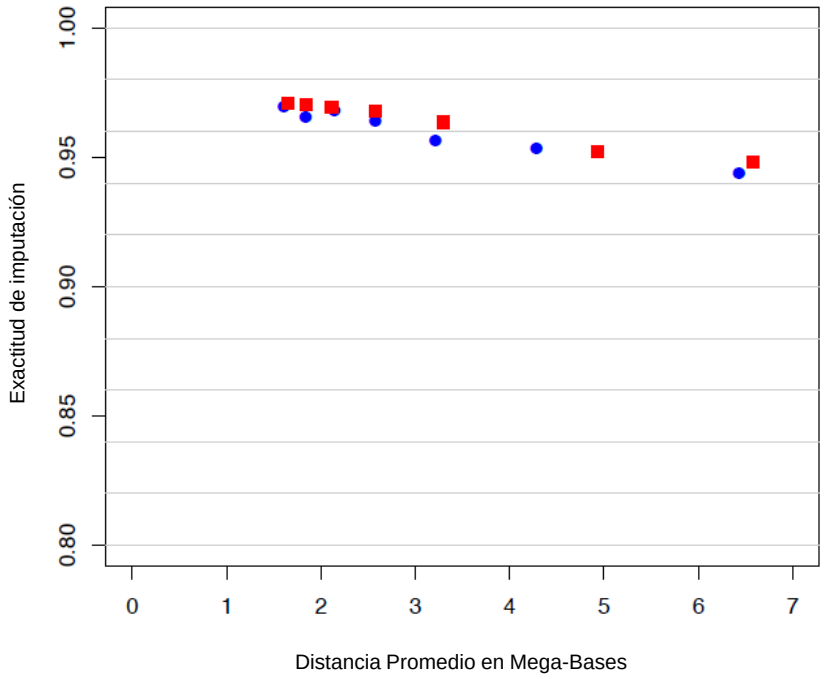


Figura 2.2. Exactitud de imputación a distancias promedio, para los cromosomas 1 y 12. La exactitud de imputación en función de la distancia promedio entre tagSNP en Mega-bases (Mb), para los cromosomas 1 (cuadrados rojos) y cromosoma 12 (círculos azules).

Consecuentemente, la distancia promedio entre SNP consecutivos es un buen parámetro para predecir la *IA* dentro de un cromosoma. Para el ejemplo anterior con una distancia promedio entre SNP de 2.1 Mb se necesitaría un mínimo de 1200 SNP aproximadamente equidistantes en la generación F_2, de modo de obtener un valor de *IA* de 0.97 en todo el genoma, cuando las generaciones F_0 y F_1 se encuentran genotipadas con un panel de 60K.

2.3.2. Imputación de genotipos experimentales F_2

***2.3.2.1.** Panel comercial de 9K*

Se calcularon los valores de *IA* para dos esquemas y para cada cromosoma, empleando como tagSNP el listado de marcadores del panel comercial 9K desarrollado en LD (Geneseek, Inc., Lincoln, NE, USA; descripto en Badke *et al*., (2013)). El esquema donde la F_1 fue genotipada en LowD produjo aproximadamente *IA* = 0.9 (90%) para todos los cromosomas. Sin embargo, cuando la F_1 fue genotipada en HD, la *IA* promedio fue 0.99 (99%, Figura 2.3).

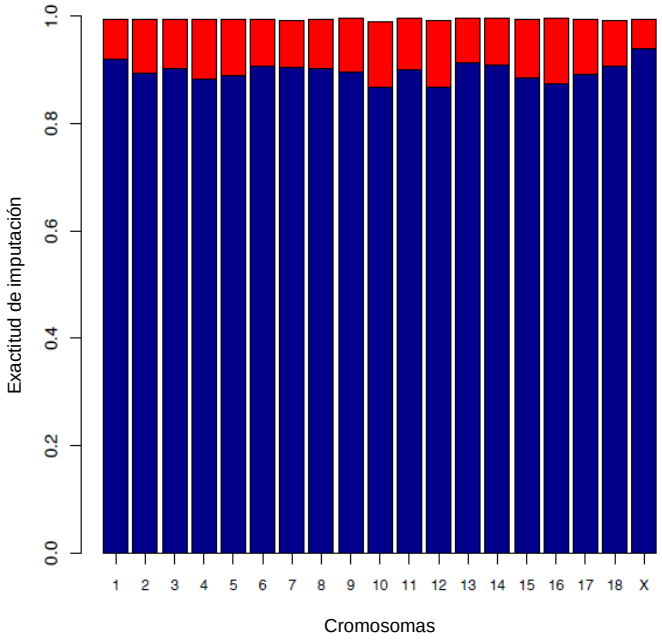


Figura 2.3. Exactitud de imputación para el panel 60K empleando el panel de 9K como tagSNP. Exactitud de imputación promedio para cada cromosoma: Barras azules esquema F_0 en alta densidad (HD), F_1 y F_2 en baja densidad (LowD). Barras rojas ganancia en exactitud con el esquema donde la F_1 es genotipada en HD.

Si bien la exactitud promedio de imputación fue cercana a 99% para todos los cromosomas, en algunas regiones de ciertos cromosomas se imputaron SNP con menor exactitud, tal como muestra la Figura 2.4 en los cromosomas 1 y 12. Los gráficos de las exactitudes de los cromosomas 1 al 18 se muestran en el **Apéndice 1**. Asimismo, como se describe en la parte superior de la Figura 2.4, se obtuvo alta *IA* en todos los SNP de la en la F_2 cuando sus padres F_1 fueron genotipados en HD (Figura 2.4 a,b). Sin embargo, al genotipar la F_1 en LowD, la *IA* en la F_2 disminuyó en todo el cromosoma (Figura 2.4 c,d), excepto por algunos SNP en los extremos del cromosoma 1.

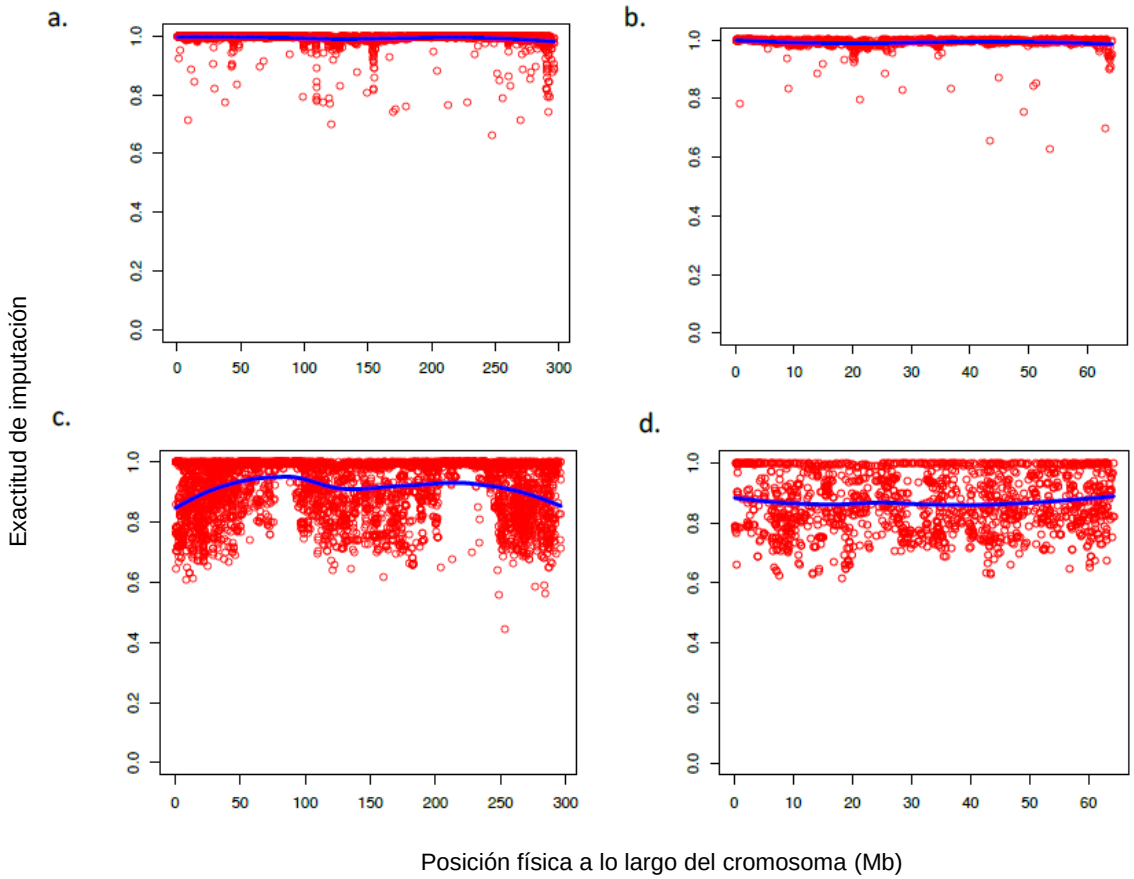


Figura 2.4. Exactitud de imputación en cromosomas 1 y 12 bajo dos escenarios de genotipado. Exactitud de imputación como función de la posición en el cromosoma bajo dos escenarios (columna izquierda 1, columna derecha 12): a) Generación F_0 y F_1 alta densidad, F_2 baja densidad, b) Generación F_0 en alta densidad, F_1 y F_2 en baja densidad. Círculos rojos representan cada SNP a lo largo del cromosoma. Línea azul es una regresión local. Resultados obtenidos usando datos experimentales.

Un interrogante adicional radica en cuál es la ganancia en exactitud al incluir la información de pedigrí, en relación con utilizar sólo el LD como única fuente de información. Se repitió entonces la imputación empleando las generaciones F_0 y F_1 en HD como panel de referencia y la generación F_2 en LowD, pero sin especificar el pedigrí de la F_2. En otras palabras, se asumió que todos los animales F_2 no estaban emparentados y sus padres eran desconocidos. En la Figura 2.5 se puede observar que la *IA* promedio en la F_2 fue igual a 0.90, valor que cuando se lo compara con *IA* = 0.99 en promedio para el cromosoma 1 considerando el pedigrí de los F_2 (Figura 2.4 a.), lo que sugiere que la inclusión de genotipos de HD de animales emparentados y se especifican las paternidades, produce un aumento considerable de la *IA*.

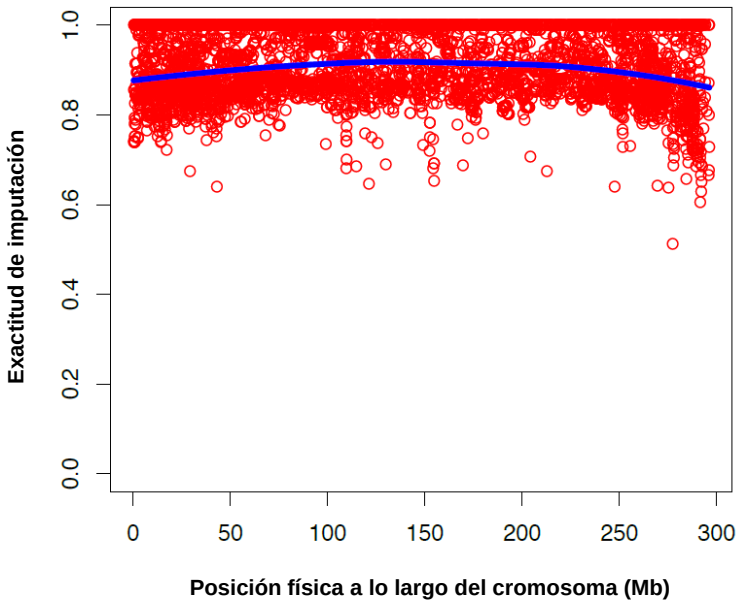


Figura 2.5. Exactitud de imputación de SNP utilizando LD, en el cromosoma 1. Exactitud de imputación como función del SNP en el cromosoma, utilizando solamente información de desequilibrio de ligamiento (LD. Generación F_0 y F_1 en alta densidad, F_2 en baja densidad pero se ignora la relación de parentesco de la F_2. Círculos rojos representan cada SNP a lo largo del cromosoma. Línea azul es una regresión local.

Las exactitudes de imputación en los diferentes esquemas (Figuras 2.3 y 2.4) reflejaron una caída promedio de 0.1, cuando la F_1 se hallaba genotipada en LowD. Para comprender mejor los resultados obtenidos, se evaluaron los haplotipos simulados de dos familias dentro de esta población y se calculó la exactitud de imputación para cada esquema. Los resultados que se obtuvieron mostraron que con F_1 en LowD, el error de fase aumentó entre los marcadores que no son tagSNP. El origen de esta imprecisión se debería a que aumenta la incertidumbre en la F_0 sobre los marcadores que no son tagSNP. Es así que cuando la F_1 estuvo genotipada en HD, la proporción de marcadores no selectos para tagSNP con fase incierta en la F_0 fue 4% y la exactitud de haplotipado en la F_1 fue 0.97. Sin embargo, cuando la F_1 estuvo genotipada en LowD la proporción de marcadores en F_0 que no son tagSNP y poseen fase incierta aumentó hasta un 30%, y la exactitud de haplotipado de la F_1 disminuyó, siendo igual a 0.85. En un segundo análisis se utilizaron los genotipos F_1 en HD como población de referencia (ignorando los genotipos de la generación F_0), situación que resultó en 43% de marcadores que no eran tagSNP con fase incierta en la F_1 con una exactitud de haplotipado aún menor a 0.78. Los anteriores resultados mostraron que la correcta imputación de marcadores que no son tagSNP en genotipos F_2 se requiere que la fase en F_1 sea cierta. En adición, de modo de garantizar una alta certeza de fase en F_1 es necesario que las dos generaciones previas, F_0 y F_1, se encuentren genotipadas en HD.

Finalmente, los resultados en las Figuras 2.4 y 2.5 muestran que, dependiendo de su posición, algunos marcadores sufrieron una exactitud de imputación variable. Consecuentemente y para conocer que otras características influyen sobre la exactitud de imputación de un marcador, se tomó como referencia la exactitud de imputación en la F_2 cuando los genotipos F_1 son de HD, y se calcularon: MAF, distancia más cercana a

un tagSNP y diferencia alélica entre las razas fundadoras, situación que se describe a continuación.

2.3.2.2. *Frecuencia alélica del alelo menos observado (MAF)*

Es bien sabido que la medida de exactitud donde se cuentan los alelos imputados correctamente es sensible al valor de la frecuencia alélica (Lin et al., 2010; Hayes et al., 2011; John M. Hickey et al., 2012). En el presente capitulo, se utilizó el cuadrado de la correlación R^2 entre el genotipo imputado y el observado como una medida robusta de la exactitud. Se debe señalar que la escala es distinta de la empleada con la medida de exactitud derivada del conteo alélico en *AI* (Cuadro 2.2).

Cuadro 2.2. Exactitud de imputación mediante *IA* y R^2 en cromosoma 12.

Escenario	Diseño de genotipado			Exactitud de imputación	
	Abuelos	**Padres**	**Progenie**	***AI***	R^2
1	HD	HD	LowD	0.962	0.884
2	HD	LowD	LowD	0.833	0.408

Comparación entre las exactitudes de imputación calculadas, sea por *conteo alélico* (*AI*), o como el *cuadrado de la correlacion* (R^2) en los datos experimentales (panel de tagSNP = 30 SNP, distancia promedio entre tagSNP de 2.1 Mb) bajo dos esquemas de genotipado: 1) F_0 y F_1 en alta densidad (HD), F_2 baja densidad (LowD), 2) F_0 en HD, F_1 y F_2 LowD

2.3.2.3. *MAF utilizando el panel de 9K en la F2*

La Figura 2.6 muestra que MAF de los SNP imputados no está asociada con la exactitud de imputación R^2 en los datos. Además, alelos con frecuencias muy extremas (MAF < 0.1) pueden ser imputados con una exactitud similar a aquellos cuyas frecuencias son intermedias (MAF > 0.3).

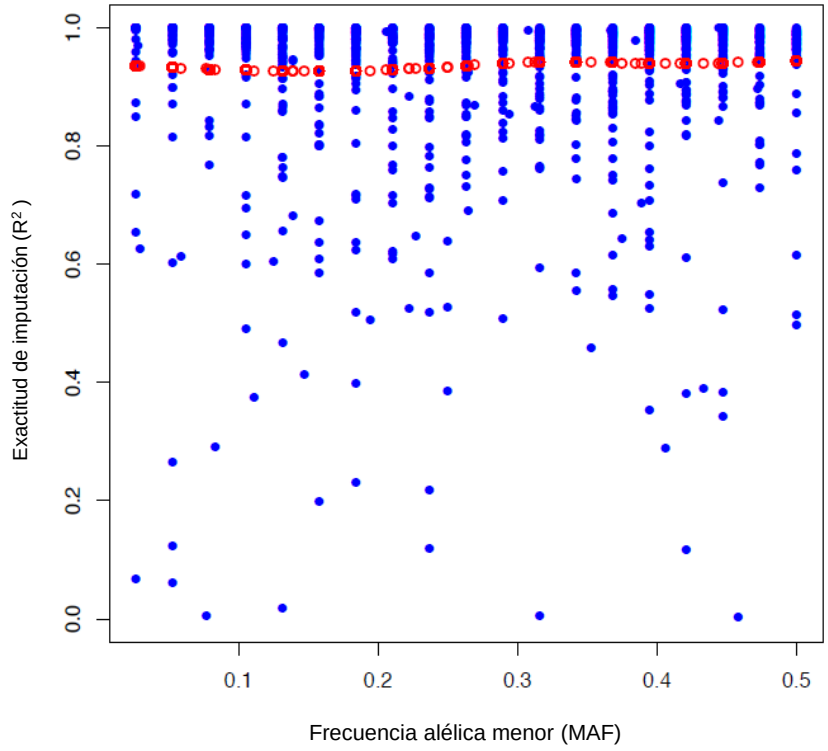


Figura 2.6. Exactitud de imputación en el cromosoma 12 como función de la frecuencia alélica F_0 menos observada. Exactitud de imputación (R^2) de los datos experimentales como función de la frecuencia alélica menos observada (MAF) con SNP (Puntos azules). Regresión local (puntos rojos).

2.3.2.4 *Distancia al tagSNP más cercano*

La Figura 2.7 muestra que no hubo diferencias en *IA* calculada por R^2 para el rango de distancias entre marcadores que no eran tagSNP y aquellos que lo eran, siendo en promedio igual a 0.936 Mb. Por lo tanto, para una densidad promedio de 0.26 Mb entre tagSNP, imputar un SNP que se encuentre localizado hacia el centro del intervalo entre dos tagSNP o muy cercano a un tagSNP, produce un efecto similar. Esto sugiere que la densidad de los marcadores de referencia (tagSNP) fue suficiente para calcular un R^2 similar entre todos los SNP que se encontraban en el intervalo.

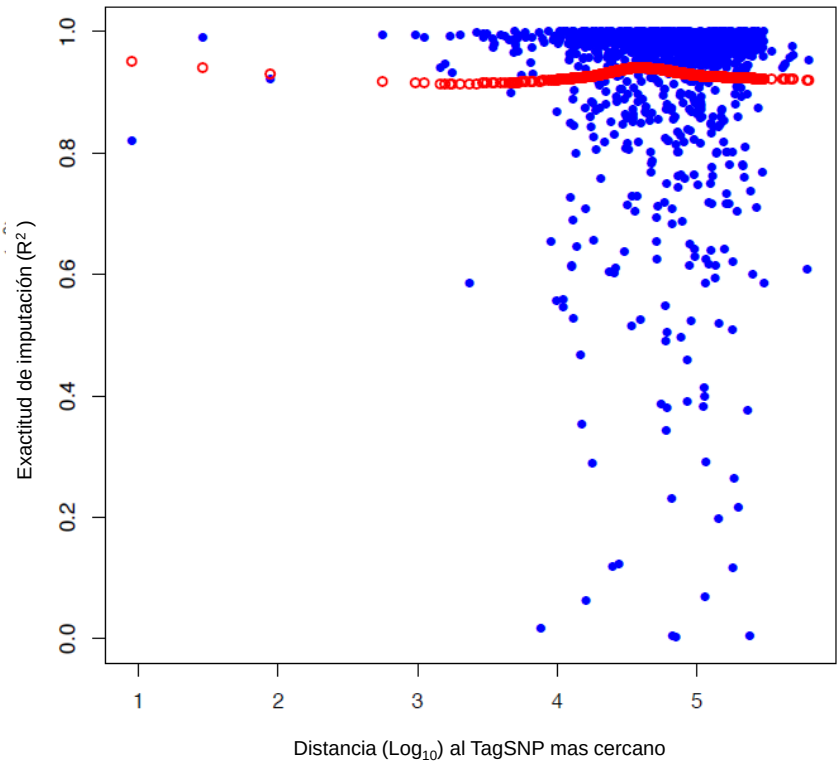


Figura 2.7. Exactitud de imputación R^2 como función de la distancia al tagSNP más cercano en el cromosoma 12. Exactitud de imputación (R^2) como función de la distancia más cercana a un tagSNP, medida en pares de bases (Log10). Los puntos azules representan no-tagSNP y los rojos a la regresión local.

2.3.2.5. *Efecto de las diferencias de frecuencias alélicas en la F_0*

Las diferencias en frecuencias alélicas entre la población fundadora no tuvo efecto apreciable sobre la exactitud de imputación (R^2), indicando que SNP segregando en los fundadores con frecuencias muy distintas, pueden imputarse con alta exactitud (Figura 2.8). Además, la aparente caída en R^2 para diferencias de MAF por encima de 0.75 que muestra la Figura 2.8, es en gran medida el resultado de que se utiliza un pequeño número de SNP para calcular la regresión local (puntos rojos).

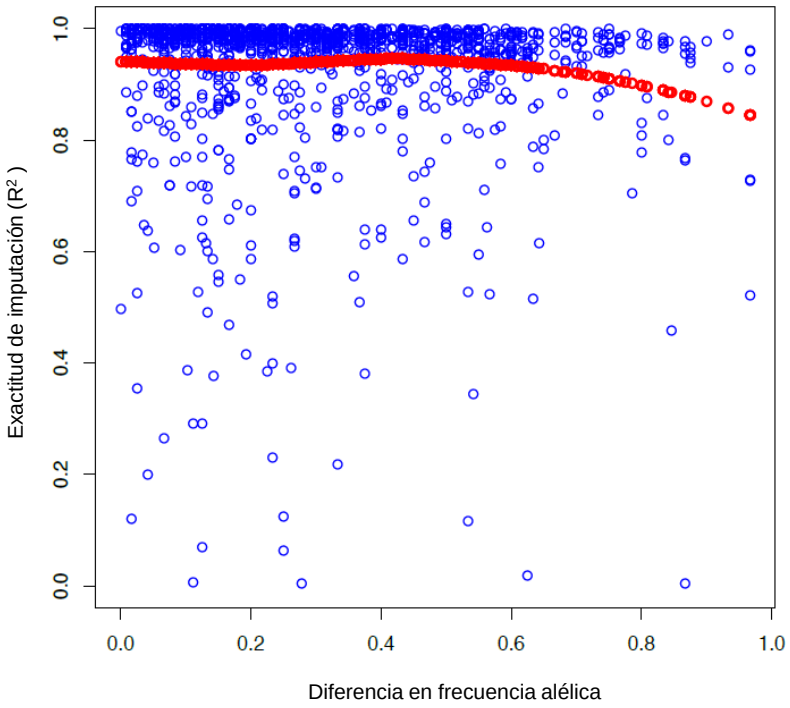


Figura 2.8. Exactitud de imputación (R^2) como función de la diferencia de la frecuencia alélicas en cromosoma 12. Exactitud de imputación (R^2) como función de la frecuencia alélicas. (puntos azules) entre las razas fundadoras (Pietrain y Duroc). Regresión local (puntos rojos).

2.4. Discusión

2.4.1. Métodos de selección de SNP y exactitud de imputación

Uno de los principales objetivos en el presente estudio fue determinar la exactitud de imputación en la población cruza F_2 Duroc × Pietrain, empleando distintos escenarios de genotipado. En una primera etapa se calculó la *IA* en una F_2 simulada. A priori parece ser ideal para aprovechar el LD para la imputación, mediante la selección de SNP igualmente espaciados (equidistantes) utilizando la posición física genética, de modo tal que la posibilidad de recombinación entre el SNP imputado y el tagSNP podría ser mínima. Sin embargo, esto no es posible en ausencia de un mapeo de alta resolución. Por consiguiente, para posicionar tagSNP se utilizaron dos procedimientos: a) espaciado físico, y b) selección de SNP basada en LD. Se obtuvieron los mismos resultados con ambos procedimientos, probablemente porque se asumió una tasa de recombinación uniforme e igual a 1cM = 1Mb. Por lo tanto, la distancia promedio entre tagSNP mostró ser un buen indicador de la *IA* en estos datos simulados, dado que se obtuvieron exactitudes iguales o mayores a 0.97 cuando la distancia promedio entre tagSNP fue menor o igual a 2.1Mb. Luego se comparó la selección de tagSNP empleando LD respecto de utilizar tagSNP más o menos equidistantes. Para cualquier marcador que no se constituya en tagSNP, se utilizó su valor de r_t^2 (medida de LD) con respecto a - al menos - un tagSNP y se lo contrastó con un valor umbral mínimo. Se observó entonces que cuando r_t^2 aumentó, el número de tagSNP y la *IA* también lo hicieron: la exactitud aumentó de 0.960 (r_t^2 = 0.1) a 0.982 (r_t^2 = 0.5), con distancias

promedio entre tagSNP de 1.86 Mb y 0.16 Mb, respectivamente. Xu *et al.* (2007) emplearon un umbral de $r_t^2 = 0.8$, para seleccionar tagSNP en un estudio de asociación genómico en seres humanos. El propósito de dicha investigación fue distinto a la presente dado que los citados autores intentaban encontrar un set de tagSNP informativos para detectar variantes causales en estudios de asociación genómica, empleando como única fuente de información el nivel de LD poblacional. Por otra parte, en este capítulo se desea utilizar este método para seleccionar SNP que se encuentren aproximadamente equidistantes, como se ha hecho en otros análisis con poblaciones de cerdos puras (Badke et al., 2013), empleando simultáneamente el LD dentro y entre familias. Con lo cual se seleccionaron menores umbrales de r_t^2, dado que con valores de $r_t^2 \geq 0.6$ se seleccionaban demasiados tagSNP y la ganancia de *IA* aumentaba marginalmente. Un segundo método fue seleccionar tagSNP equiespaciados, para lo cual se dividió el cromosoma en segmentos de igual tamaño y se seleccionó el SNP más cercano al centro del segmento. Otros estudios utilizaron tagSNP equiespaciados escogiendo un SNP por cada cierto número de marcadores (Hayes et al., 2011), o seleccionando el SNP de mayor MAF en un determinado segmento (Hayes et al., 2011; Hickey et al., 2012 ; Huang et al., 2012). El hecho que se tiene una gran disponibilidad de SNP (*e.g.* 60K) en todo el genoma hizo posible seleccionar SNP equiespaciados - o aproximadamente equiespaciados - con un amplio rango de MAF, siempre y cuando los SNP se encontrasen segregando (es decir que no eran los mismos en todos los genotipos; no eran mono-mórficos). Las exactitudes de imputación calculadas con SNP seleccionados por ambos métodos fueron muy similares. En síntesis, no hubo diferencias en *IA* usando tagSNP seleccionados por estos dos métodos. Esta similitud de resultados, podría deberse a que el método con SNP "equiespaciados" asume que el LD se distribuye relativamente uniforme a lo largo del genoma y no existen grandes bloques de LD. En nuestro estudio, los haplotipos de la F_1 provienen de dos poblaciones: Duroc y Pietrain. El LD resultante mostró ser relativamente alto y uniformemente distribuido a excepción de unos pocos bloques de alto LD, los cuales estuvieron conformados por 7 SNP consecutivos con LD $\geq$ 0.8. Por esta razón, seleccionar tagSNP mediante LD, o por equiespaciamiento, produjo similar *IA* cuando las densidades fueron equivalentes. Si bien se simuló asumiendo tasas de recombinación uniforme, los resultados concordaron con datos experimentales donde se emplearon ambos métodos de selección de tagSNP y se obtuvo virtualmente la misma *IA* que en una población porcina pura (Badke et al., 2013). Esto sugiere que diseñar un panel de SNP apropiado para cada población no sería rentable. Consecuentemente, se evaluó la *IA* de un panel comercial cuyos marcadores fueron seleccionados sobre la base de su posición física en el genoma y la frecuencia del alelo menos usual o MAF (Badke et al., 2013).

2.4.2. Imputación mediante un panel de 9K y escenarios de genotipado

En esta etapa se utilizó la información de un panel de 9K, donde la distancia promedio entre SNP fue de 0.30 Mb, para imputar un panel de 60K (HD) en los datos experimentales F_2 bajo distintos escenarios de genotipado. En el primer escenario se contó con los genotipos F_0 y F_1 en HD y la F_2 en LowD, produciendo como resultado una *IA* promedio de 0.99. Weigel et al. (2010) imputaron genotipos de 8K a 43K empleando información de padre, madre y de abuelos paterno y materno, y obtuvieron valores *IA* > 0.95. En el segundo escenario los genotipos F_0 estuvieron en HD pero los genotipos F_1 fueron de 9K, y se observó una caída de *IA* a 0.9. Finalmente, en el tercer

escenario se emplearon genotipos F_0 y F_1 en HD y F_2 en LowD, pero se ignoraron las relaciones de parentesco entre la F_2 con el panel de referencia paterno, empleándose el LD poblacional exclusivamente, para producir un *IA* promedio de 0.9. Badke *et al.*, (2013) utilizaron genotipos de una población de referencia formada por tríos para imputar en una población no relacionada, y obtuvieron *IA* del orden de 0.90 a 0.95, empleando grupos de referencia de 16 y 64 animales, respectivamente.

Al reflexionar sobre la disminución de la exactitud de imputación bajo los distintos escenarios podemos citar a Habier *et al*. (2009), quienes observaron que la imprecisión de la imputación puede deberse fundamentalmente a dos factores: 1) La exactitud del haplotipado de los tagSNP que flanquean a los marcadores que no son tagSNP, y 2) la exactitud de generar los haplotipos durante la imputación de los marcadores distintos a los tagSNP, condicional a que los marcadores tagSNP hayan sido "haplotipados" correctamente. Para estos dos factores se evaluó entonces, mediante simulación estocástica, las exactitudes al generar haplotipos dentro de los dos primeros escenarios, cuando se considera la relación entre la F_2 y el panel de referencia. La exactitud de haplotipado se calculó como el número de inferencias de fase erróneas entre marcadores consecutivos heterocigotas, de manera similar a la efectuada por Druet y Georges (2010). Se observó en todos los escenarios que la fase de los tagSNP fue correcta, con lo cual la incertidumbre se debió al origen, dentro de los abuelos, para los marcadores que no son tagSNP y que son flanqueados por tagSNP.

A continuación se cuantificó la incertidumbre de fase de los marcadores que no eran tagSNP. Para el escenario F_0 y F_1 en HD, F_2 en LowD la fracción de marcadores distintos a tagSNP con fase incierta fue 4%, mientras que alcanzó el 30% en el escenario con F_0 en HD, F_1 y F_2 en LowD, cantidades que se reflejaron en exactitudes de haplotipado de 0.97 y 0.85, respectivamente. Los resultados sugieren que las diferencias encontradas entre las exactitudes de imputación fueron afectadas por el conocimiento de la fase de los marcadores distintos a los tagSNP. Además, incorporando el genotipo en HD de más individuos emparentados (por ejemplo, F_0) aumenta la exactitud de haplotipado y, por lo tanto, la *IA* aumenta. Nótese que estos resultados aplican para un diseño con un número pequeño de fundadores genotipados en HD y una gran cantidad de progenie genotipada en LowD. Si la fase de los fundadores es conocida, es sencillo seguir la transmisión de segmentos cromosómicos hacia las generaciones posteriores empleando la información del ligamiento. Sin embargo en la práctica, la fase debe ser estimada mediante la información que provee el LD, la cual fue muy exigua en la población bajo estudio donde el número de genotipos F_0 fue muy reducido. Hay dos aproximaciones para lidiar con este problema. En primer lugar, y como ocurre con grandes pedigríes, se pueden genotipar más animales de la(s) misma(s) población(es) de fundadores de modo de emplear el LD para evaluar la fase en los animales base con mayor precisión. Segundo, como se discutió en este capítulo, es de utilidad contar con dos generaciones consecutivas genotipadas en HD para utilizar la información de los abuelos (F_0) y evaluar con exactitud la fase de los padres (F_1), de modo tal de emplear la información de ligamiento al imputar los genotipos de la progenie F_2. Para que este enfoque sea funcional, es necesario pedigrí completo intergeneracional (3 generaciones) y genotipos en HD de los animales de las generaciones base. A pesar del esfuerzo en número de individuos, genotipado y duración, el enfoque es económicamente efectivo para una población F_2 típica (Druet y Farnir, 2011; Habier et al., 2009), dado que con esta estructura, se pueden obtener genotipos F2 en HD imputados a partir de chips de baja densidad.

2.4.3. Efecto de la frecuencia alélica menor (MAF)

Las medidas de exactitud de imputación que tienen en cuenta únicamente conteos alélicos carecen de utilidad para comparar SNP con distinto valor de MAF, debido a que los errores de imputación son muy sensibles al valor de las frecuencias alélicas (Hayes et al., 2011; Hickey et al., 2012; Lin et al., 2010). Se han propuesto dos mediciones de la exactitud de imputación que intentan evitar esta dependencia: 1) la correlación entre el valor del genotipo imputado y el observado (Hickey et al., 2012); y 2) la exactitud de imputación corregida por su valor esperado (Lin et al., 2010; Hayes et al., 2011). El segundo método consiste en ajustar la exactitud de imputación, calculada mediante la diferencia entre la exactitud observada y una estimación del valor esperado bajo muestreo aleatorio, situación que genera varias posibilidades de cálculo. Sin embargo, para distintas medidas, se observó una tendencia a decrecer en exactitud de imputación cuando MAF < 0.15. Por ejemplo, Hickey *et al.* (2012) observaron en que genotipos de maíz mostraban un R^2 decreciente cuando MAF < 0.10, siendo la caída mayor cuando los genotipos se encontraban "ocultos" en una proporción > 84% del total de SNP. De igual manera, Lin *et al.* (2010) analizaron genotipos humanos corregidos por la exactitud esperada y observaron una marcada caída en la exactitud de imputación con MAF < 0.15. Hayes *et al.* (2011) realizaron la misma corrección que Lin *et al.* (2010), pero en genotipos de ovejas y, si bien encontraron altas exactitudes de imputación, observaron una tendencia de estas a decrecer cuando MAF < 0.10. En la presente investigación se utilizó la correlación entre el genotipo observado y el imputado (R^2) para evaluar el efecto del MAF sobre la exactitud de imputación. Los resultados muestran que los marcadores con MAF < 0.10 en la generación F_0 fueron imputados en genotipos F_2 con exactitudes razonablemente importantes (Figura 2.6), un resultado distinto a los discutidos previamente. Sin embargo, esto es esperable si se considera que se utilizó, tanto el LD como el ligamiento físico (información del pedigrí), como fuentes de información para la población cruza. Por lo tanto, la frecuencia alélica en la generación F_0 no es relevante siempre y cuando ambos alelos se encuentren segregando. En adición, al ser la F_1 genotipada en HD, los SNP con bajo valor de MAF fueron observados en la F_0 y en la F_1, hecho que sumado a que todas las paternidades y los abuelos fueron conocidos, simplificó la imputación de los animales F_2.

2.4.4. Posibles efectos en asociación

En la investigación que se describe en este capítulo, se compararon las dosis alélicas de genotipos observados e imputados, para poder contar con una metodología aceptable de diseño de genotipado e imputación empleando genotipos en BD de animales F_2. Zheng *et al.,* (2011) reportaron que la regresión del fenotipo en la dosis alélica constituye un método acertado para evaluar los efectos de QTL. Además, observaron que cuando las exactitudes de imputación son altas, la potencia del test de asociación es elevada. Por ejemplo: exactitudes de imputación > 0.95 estuvieron asociadas con valores de potencia > 0.85. La obtención de una exactitud de imputación con el panel de 9K cercana a $R^2 = 0.94$, sugiere que la potencia de un test de asociación es alta. Otros estudios también encontraron que la imputación de genotipos mejoró la potencia de las pruebas de asociación. Hao *et al.* (2009) compararon la potencia en un análisis GWAS para cuatro diferentes estrategias de imputación de genotipos humanos:

(1) prueba por asociación directa con un panel Illumina de 317K SNP, (2) asociación sobre un panel imputado completamente del panel entero respecto del chip HapMap SNP, proveniente del Illumina 317K SNP, (3) análisis de asociación directo empleando el panel Illumina 650Y, y (4) analizaban para asociación del panel entero imputado sobre el HapMap SNP proveniente del Illumina 650Y. Se observó que con las estrategias de imputación completa del genoma (métodos 2 y 4), se mejoró la potencia en 5.5% para el Ilumina 317K, o 3.3% para el Illumina 650Y, comparado con los análisis usando solamente los SNP (estrategias 1 y 3, respectivamente). Anderson *et al*., (2008) obtuvieron resultados similares con las plataformas de 300K y 550K SNP. Cabe anotar que siempre se tiene un margen de error de imputación (entre un 2 a 5%), que este podría estar en igual medida al error de genotipado. Además, el imputar correctamente gran parte del genoma hace que las pruebas tengan un mejor grado de potencia al poseer mas genomas con que contrastar la prueba, mejorando la potencia del test.

El costo de genotipado constituye una consideración importante. En la actualidad, el costo comercial promedio de un panel de HD (60K) en cerdos es más de dos veces mayor que el de un panel LowD (9K). En este sentido, asumiendo estas proporciones fijas y una población similar a la del presente estudio: 20 animales F_0, 56 F_1 y 1000 F_2, se podrían genotipar aproximadamente 1.9 veces la misma población empleando el escenario donde la F_0 y la F_1 se hallan en HD y la F_2 en LowD (que luego será imputada a HD) y , con el mismo costo, se obtendría la misma información que cuando se genotipan las tres generaciones F_0, F1 y F_2 en HD. La metodología de imputación puede entonces utilizarse en estudios de asociación o de meta-análisis.

2.5. Conclusión

El diseño de paneles de SNP particulares para cada población F_2 es una práctica costosa, dado que es necesario tener un gran número de SNP para obtener exactitudes de imputación razonables y tiene un muy elevado costo de desarrollo. En el caso particular de la población bajo estudio se necesitaría un número mínimo de 1200 marcadores con una distancia promedio entre cada uno de 2.1 Mb, de modo de poder obtener un valor de *IA* en la F_2 superior a 0.97. Por otra parte, el empleo de un panel de 9K como tagSNP (es decir, un panel LowD) tuvo como resultado un valor de *IA* igual a 0.99 cuando la F_0 y F_1 fueron genotipadas en HD y la F_2 en BD. El costo de este esquema de genotipado podría ser menor que la mitad del costo de utilizar el panel de HD en todos los animales. La correlación entre los genotipos observados e imputados fue alta (R^2 = 0.94), lo que sugiere que la potencia de los estudios de asociación podría ser alta. Entonces, bajo una estrategia eficiente de genotipado con una alta exactitud de imputación (ejemplo: F_0 y F_1 en HD y F_2 en LowD), se pueden obtener la información de genotipos imputados de más animales en HD con un relativamente bajo costo. Aplicados a imputar marcadores en poblaciones donde el número de fundadores genotipados en HD es pequeño, permite evaluar la fase de los padres de animales imputados con bastante certeza. Por otra parte, los resultados obtenidos empleando LD poblacional se encuentran restringidos a poblaciones de cerdos que muestren niveles de LD similares a los de las poblaciones fundadoras aquí empleadas (Badke et al., 2012).

CAPÍTULO 3

Análisis de asociación genética transformando linealmente las evaluaciones genómicas [3]

[3] Gualdrón Duarte JL, Cantet RJC, Bates RO, Ernst CW, Raney NE, Steibel JP. Rapid screening for phenotype-genotype associations by linear transformation of genomic evaluatiations. BMC *Bioinformatics* 2014,15:246.

3.1. Introducción

Existe una gran disponibilidad de genotipos en alta densidad con marcadores SNPs en plantas y animales domésticos en la actualidad. Al ser utilizados en datos fenotípicos de caracteres complejos, dichos marcadores permiten: 1) la predicción de los valores de cría genómicos (GEBVs) (Crossa *et al*., 2011; Goddard y Hayes, 2009) para evaluaciones genómicas (Hayes *et al*., 2009), y 2) la estimación de los efectos de regiones genómicas asociadas con la variabilidad genética del carácter en los análisis de asociación en todo el genoma conocidos como GWAS (Goddard y Hayes, 2009; Hayes *et al*., 2010; Kumar *et al*., 2013). En este tipo de análisis se utilizan modelos mixtos y procedimientos de múltiples pruebas de hipótesis (MPH, Zhou y Stephens, 2012), ajustando todos los efectos individuales de regiones genómicas en el modelo (Hayes *et al*., 2010). Este ajuste es necesariamente complejo cuando el número de individuos y el número de efectos de marcador son muy grandes. En este capítulo se propone transformar linealmente los GEBVs en efectos de marcador mediante un modelo mixto equivalente, para posteriormente evaluarlos mediante una prueba estadística de estandarización que emplea la varianza de los efectos, en vez de utilizar la varianza del error de predicción.

El primer paso del método de *selección genómica* propuesto por Meuwissen *et al.* (2001) consiste en estimar GEBVs ajustando los efectos de marcador en una cierta muestra de datos o población de referencia. En una segunda etapa, se predicen los GEBV sumando las soluciones de los "SNP" estimadas anteriormente según el genotipo particular de cada animal. El modelo mixto empleado contiene los vectores de efectos fijos y aleatorios de marcadores SNP $\left(g\right)$. Estos últimos son asumidos provenir de una distribución normal con esperanza igual a cero y una matriz de covarianzas igual al producto de la matriz identidad por la varianza de los efectos de SNP $\left(\boldsymbol{I}\,\sigma_g^2\right)$. Para los errores aleatorios se asume una distribución similar pero con matriz de covarianzas $\boldsymbol{I}\,\sigma_e^2$. Garrick (2007) y Stranden *et al.* (2009) propusieron un modelo mixto equivalente a partir de la transformación lineal de $\boldsymbol{a} = \boldsymbol{Z}\,\boldsymbol{g}$, siendo $\boldsymbol{a}$ el vector aleatorio de los valores de cría, y $\boldsymbol{Z}$ la matriz de incidencia que relaciona los elementos en $\boldsymbol{a}$ con aquellos elementos en $\boldsymbol{g}$, o vector de los efectos de SNP. Cada columna de $\boldsymbol{Z}$ esta asociada con un marcador y los elementos se encuentran estandarizados por funciones de las frecuencias alélicas de los SNP y por el número total de SNP. Esta transformación utiliza la misma matriz $\boldsymbol{Z}$ que el modelo de Meuwissen *et al.* (2001), la cual relaciona el vector de efectos de marcador en $\boldsymbol{g}$ con los datos fenotípicos. Además, los GEBVs en el modelo equivalente poseen una matriz de covarianzas $\boldsymbol{G}\,\sigma_a^2 = \boldsymbol{Z}\,\boldsymbol{Z}'\,\sigma_g^2$. El procedimiento requiere que las varianzas (escalares) sean iguales, *e.g.* $\sigma_A^2 = \sigma_g^2$. Una vez ajustado el modelo equivalente, los efectos de los SNP se calculan mediante la transformación $\boldsymbol{g} = \boldsymbol{Z}'\boldsymbol{G}^{-1}\boldsymbol{a}$, y los efectos individuales en $\boldsymbol{g}$ son divididos por la raíz cuadrada de su varianza individual $\left(\mathrm{Var}\left(g_j\right)\right)$ para obtener un test estadístico, denominado SNP_{ej}. En la presente investigación se desarrolló una fórmula para calcular la $\mathrm{Var}\left(g_j\right)$ sin tener que ajustar los efectos de SNP en el modelo. El paso siguiente que será seleccionar en cada cromosoma segmentos genómicos potencialmente asociados con la variabilidad genética del carácter. Para ello, se escoge el SNP con mayor valor de menos el logaritmo del *p*-valor a lo largo del cromosoma. Una vez que el SNP es localizado, se define un segmento de una Mb hacia la derecha y

otro hacia la izquierda (teniendo en cuenta la distancia de caída de LD en la población de estudio), sobre la base de la posición del SNP selecto, con una matriz de relaciones calculada empleando solamente la información de los SNP que se encuentran dentro del segmento. Dicha matriz de relaciones es proporcional a la matriz de covarianzas de los efectos del segmento, dentro de un modelo que también incluye efectos fijos y las variables aleatorias de los GEBVs.

Finalmente, se calculó la verosimilitud para evaluar la significancia de aquellos efectos de segmento que explican una mayor parte de la variabilidad dentro de cada cromosoma, y se la comparará con la verosimilitud de un modelo reducido con efectos fijos y GEBVs. El valor crítico (o "tamaño" del test) será ajustado mediante la corrección de Bonferroni. El algoritmo presentado en este capítulo permite calcular predicciones genómicas y realizar con ellas GWAS, en un tiempo mínimo de cómputo y con bajos requerimientos de memoria. Como además, la varianza específica de los efectos de SNP tiene en cuenta la cantidad de información de cada marcador, será utilizada para calcular la prueba de asociación, lo que no ocurre con otras pruebas que emplean una varianza a priori, o una constante estimada a partir de la varianza aditiva.

3.2. Métodos

3.2.1. Animales

Los animales utilizados en esta investigación provienen de la misma población empleada en el Capítulo 2. (ver sección *2.2.1. Animales*). Se recolectaron datos fenotípicos para crecimiento, res y calidad de carne para 950 cerdos F_2 (para más detalles ver Edwards *et al.*, 2008 (a); (b)). Se utilizó el carácter *grasa dorsal en la decima costilla* (*mm*) medida durante la semana de vida 13 (bf10_13wk). El carácter fue seleccionado porque muestra una heredabilidad alta (0.42) y los datos se distribuyen normalmente.

3.2.2. Genotipado y control de calidad

Se evaluaron los genotipos empleando dos paneles de marcadores SNP: 1) tal como se describe en el Capítulo 2 (Sección Métodos), se genotiparon 411 animales (4 machos F_0 Duroc, 15 hembras F_0 Pietrain, 6 machos F_1, 50 hembras F_1 y 336 animales F_2) de la población mediante el chip de Illumina PorcineSNP60 (62163 SNP) Genotyping beadchip (Illumina Inc.) (Ramos et al., 2009), y 2) 612 animales con un segundo panel compuesto de 9K TSNP como el GeneSeek Genomic Profiler para Porcine LD (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE) (Badke et al., 2013). Un grupo de 5350 SNP se eliminaron para todos los análisis debido a que su posición física era desconocida. Las inconsistencias mendelianas ($\leq$ 0.01%) fueron consideradas como genotipos perdidos y no se tuvieron en cuenta para el estudio los genotipos de 21 animales (1 F_1 y 20 F_2) que perdieron información en más del 10% de los SNP (MIND > 0.10). De manera similar, se eliminaron 2978 SNP por tener más del 10% de datos faltantes. En adición, se eliminaron 9877 SNP cuya frecuencia alélica del alelo menos observado (MAF) estuvo debajo del 1% (MAF < 0.01). Este proceso de edición fue similar al realizado Badke *et al*. (2012) y Gualdrón Duarte *et al*. (2013), y fue realizado con el programa PLINKv1.07 (Purcell *et al.*, 2007). Los animales F_2 genotipados con el panel de 9K fueron imputados a 60K siguiendo los procedimientos descritos por Gualdron *et al*. (2013), mediante el programa AlphaImpute (Hickey *et al*., 2012), dando como resultado exactitudes de imputación cercanas a 0.99 (Gualdrón

Duarte *et al*., 2013). Los genotipos imputados en la F_2 tuvieron un segundo procedimiento de edición por MAF<0.01, el cual excluyó 759 SNP virtualmente monomórficos. Como resultado del procedimiento de edición e imputación, se obtuvieron 1002 animales de las tres generaciones (F_0, F_1 y F_2) con información de 44055 SNP por individuo.

3.2.2. Estimación de la matriz genómica de relaciones

La matriz de relaciones genómicas se estimó a partir de los genotipos observados e imputados en alta densidad (aproximadamente 44K), los que fueron expresados en dosis alélicas (Badke *et al*., 2013; Gualdrón Duarte *et al.*, 2013). Para ello se calculó una matriz $\boldsymbol{M}$ de dimensión $n \times m$; n es el número de animales y m el de marcadores SNP. Los elementos resultantes se encuentran en el intervalo [0,2], dado que se obtienen contando el número de alelos del referente. Utilizando el sub-índice i para referirse al individuo, se estandarizó la matriz $\boldsymbol{M}$ para calcular $\boldsymbol{Z}$ del modo siguiente:

$$\boldsymbol{Z}_{ij} = \frac{\boldsymbol{M}_{ij} - 2p_j}{\sqrt{m\left(2p_j\left(1-p_j\right)\right)}} \qquad [3.1]$$

(VanRaden, 2008). La frecuencia alélica p_j se calculó a partir de la generación F_0 (19 animales) y la matriz de relaciones genómicas $\boldsymbol{G}$ de dimensión ($n \times n$) es entonces igual a

$$\boldsymbol{G} = \boldsymbol{Z}\ \boldsymbol{Z}' \qquad [3.2]$$

3.2.3. Modelo predictivo

La ecuación del modelo animal centrado para evaluación genómica es:

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{a} + \boldsymbol{e} \qquad [3.3]$$

Donde $\boldsymbol{y}$ es el vector de observaciones para bf10_13wk; $\boldsymbol{X}$ es la matriz de incidencia que relaciona los datos con el efecto de sexo en $\boldsymbol{\beta}$; $\boldsymbol{a}$ es el vector de los valores de cría cuya distribución es $\boldsymbol{a} \sim N\left(0, \boldsymbol{G}\sigma^2_{\mathrm{A}}\right)$ y el vector de errores $\boldsymbol{e}$ tiene distribución $\boldsymbol{e} \sim N\left(0, \boldsymbol{I}\sigma^2_e\right)$. Se utilizó el programa *regress* versión 1.3-10 R (Clifford y McCullagh, 2006) para estimar los componentes de varianza, empleando el método de Máxima Verosimilitud Restringida (REML). Stranden y Garrick (2009) propusieron el siguiente modelo equivalente a [3.3]:

$$\boldsymbol{y} = \boldsymbol{X}\ \boldsymbol{\beta} + \boldsymbol{Z}\ \boldsymbol{g} + \boldsymbol{e} \qquad [3.4]$$

Con excepción del vector de efectos de SNP $\boldsymbol{g}$, los elementos en [3.4] son definidos tal como se lo hizo previamente para el modelo [3.3]. Debemos mostrar ahora que [3.3] y [3.4] son modelos equivalentes. Dado que $\boldsymbol{a} = \boldsymbol{Z}\ \boldsymbol{g}$, las varianzas de $\boldsymbol{a}$ y $\boldsymbol{g}$ están relacionadas de la siguiente manera:

$$\boldsymbol{G}\,\sigma^2_{\mathrm{A}} = \mathrm{Var}\left(\boldsymbol{a}\right) = \mathrm{Var}\left(\boldsymbol{Z}\ \boldsymbol{g}\right) = \boldsymbol{Z}\ \mathrm{Var}\left(\boldsymbol{g}\right)\boldsymbol{Z}' = \boldsymbol{Z}\,\boldsymbol{Z}'\,\sigma^2_g \qquad [3.5]$$

Las condiciones necesarias para que los modelos [3.3] y [3.4] sean equivalentes (Henderson, 1984) son a) $\boldsymbol{G} = \boldsymbol{Z}\,\boldsymbol{Z}'$ y b) $\sigma^2_{\mathrm{A}} = \sigma^2_{\mathrm{g}}$.

3.2.4. Varianza de los efectos de los SNP

En esta sección se describe el algoritmo para calcular la varianza de los efectos de los SNP en $\boldsymbol{g}$ $\left(\mathrm{Var}\left(\hat{\boldsymbol{g}}\right)\right)$. En primer lugar, $\boldsymbol{g}$ y $\boldsymbol{a}$ se asocian por la transformación lineal $\boldsymbol{a} = \boldsymbol{Z}\,\boldsymbol{g}$ (Hayes *et al.*, 2010; McClure *et al*., 2012; Strandén y Garrick, 2009; Wang *et al*., 2012). Consecuentemente:

$$\begin{aligned}\mathrm{BLUP}\left(\boldsymbol{g}\right) &= \hat{\boldsymbol{g}} \\ &= \mathrm{cov}\left(\boldsymbol{g},\boldsymbol{a}'\right)\left[\mathrm{Var}\left(\boldsymbol{a}\right)\right]^{-1}\hat{\boldsymbol{a}} \\ &= \mathrm{cov}\left(\boldsymbol{g},\boldsymbol{g}'\right)\boldsymbol{Z}'\boldsymbol{G}^{-1}\left(\sigma_{\mathrm{A}}^{2}\right)^{-1}\hat{\boldsymbol{a}} \\ &= \left(\frac{\sigma_{\mathrm{g}}^{2}}{\sigma_{\mathrm{A}}^{2}}\right)\boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{a}} \\ &= \boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{a}}\end{aligned} \qquad [3.6]$$

El último paso resulta de la condición b) de equivalencia entre [3.3] y [3.4] que requiere que $\sigma_a^2 = \sigma_g^2$. Ahora bien, la varianza de los efectos de los SNP, $\mathrm{Var}\left(\hat{\boldsymbol{g}}\right)$, en [3.6] es igual a:

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{g}}\right) &= \mathrm{Var}\left(\boldsymbol{Z}'\boldsymbol{G}^{-1}\hat{\boldsymbol{a}}\right) \\ &= \boldsymbol{Z}'\boldsymbol{G}^{-1}\,\mathrm{Var}\left(\hat{\boldsymbol{a}}\right)\boldsymbol{G}^{-1}\boldsymbol{Z}\end{aligned} \qquad [3.7]$$

A su vez, la varianza del error de predicción (PEV) de $\hat{\boldsymbol{a}}$ en [3.3] es:

$$\mathrm{PEV}\left(\hat{\boldsymbol{a}}\right) = \mathrm{Var}\left(\boldsymbol{a}-\hat{\boldsymbol{a}}\right) \qquad \boldsymbol{C}^{\boldsymbol{aa}} = \mathrm{Var}\left(\boldsymbol{a}\right) - \mathrm{Var}\left(\hat{\boldsymbol{a}}\right) \qquad [3.8]$$

Tal que,

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{a}}\right) &= \mathrm{Var}\left(\boldsymbol{a}\right) - \boldsymbol{C}^{\boldsymbol{aa}} \\ &= \boldsymbol{G}\,\sigma_a^2 - \boldsymbol{C}^{\boldsymbol{aa}}\end{aligned} \qquad [3.9]$$

La matriz $\boldsymbol{C}^{\boldsymbol{aa}}$ es la fracción de la inversa de la matriz de los coeficientes correspondiente a los efectos aleatorios (Henderson, 1984) o de animal, e igual a:

$$\boldsymbol{C}^{\boldsymbol{aa}} = \sigma_e^2\left(\boldsymbol{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \boldsymbol{G}^{-1}\lambda\right)^{-1} \qquad \lambda = \frac{\sigma_e^2}{\sigma_a^2} \qquad [3.10]$$

Ahora se reemplaza en [3.7] con el resultado [3.9], y la varianza de los efectos de SNP $\hat{\boldsymbol{g}}$ puede expresarse como:

$$\begin{aligned}\mathrm{Var}\left(\hat{\boldsymbol{g}}\right) &= \boldsymbol{Z}'\boldsymbol{G}^{-1}\left(\boldsymbol{G}\sigma_{\mathrm{A}}^{2} - \boldsymbol{C}^{\boldsymbol{aa}}\right)\boldsymbol{G}^{-1}\boldsymbol{Z} \\ &= \boldsymbol{Z}'\boldsymbol{G}^{-1}\boldsymbol{Z}\sigma_{\mathrm{A}}^{2} - \boldsymbol{Z}'\boldsymbol{G}^{-1}\boldsymbol{C}^{\boldsymbol{aa}}\boldsymbol{G}^{-1}\boldsymbol{Z}\end{aligned} \qquad [3.11]$$

La expresión [3.11] produce una matriz de orden $m \times m$, siendo m el número de SNP. Sin embargo, solo se necesitan los elementos de la diagonal. Además, nótese que en el primer termino de la ecuación [3.11]: $\boldsymbol{Z}'\boldsymbol{G}^{-1}\boldsymbol{Z}$, puede computarse y almacenarse para ser empleado en otros caracteres, mientras que $\boldsymbol{C}^{\boldsymbol{aa}}$ debe ser calculada para cada carácter.

3.2.5. Estandarización de los efectos de los SNP

Los efectos de cada *j*-ésimo SNP $\hat{\boldsymbol{g}}_j$ estimado en [3.6] se estandarizaron (SNP_{ej}) mediante la división correspondiente de la varianza para cada marcador, $\text{Var}\left(\hat{\boldsymbol{g}}_j\right)$, que se observa en [3.11]. La estandarización se realizó como se muestra a continuación:

$$SNP_{ej} = \frac{\hat{\boldsymbol{g}}_j}{\sqrt{\text{Var}\left(\hat{\boldsymbol{g}}_j\right)}} \qquad [3.12]$$

3.2.6. P-valores y rastreo genómico

El *p*-valor para el *j*-ésimo SNP fue calculado como se muestra a continuación:

$$p\text{-valor}_j = 2\left[1 - \Phi\left(\left|SNP_{ej}\right|\right)\right] \qquad [3.13]$$

La expresión $\Phi\left(x\right)$ es la función de distribución acumulada para la densidad normal absoluta de la variable *x*. Para analizar bf10_13wk con un Manhattan plot, los *p*-valores de cada SNP se ubicaron a lo largo del genoma como $-\text{Log}_{10}(p\text{-valor}_j)$, teniendo en cuenta la posición física de cada SNP en el genoma en términos de Mega-bases (Mb).

3.2.7. Estandarización de los efectos de los SNP mediante el PEV del marcador

La segunda estandarización del SNP $\hat{\boldsymbol{g}}_j$ en [3.6] empleó el $\text{PEV}\left(\hat{\boldsymbol{g}}_j\right)$:

$$SNP_{epj} = \frac{\hat{\boldsymbol{g}}_j}{\sqrt{\text{Var}\left(\boldsymbol{g}_j\right) - \text{Var}\left(\hat{\boldsymbol{g}}_j\right)}} \qquad [3.14]$$

Como se indicó anteriormente $\sigma^2_{\text{A}} = \sigma^2_{\text{g}}$. Asimismo, los *p*-valores y el barrido genómico para el test SNP_{epj} fueron calculados y graficados de la misma manera que para SNP_{ej}.

3.2.8. Simulación

Se realizó una simulación para comparar el efecto de la estandarización (SNP_{ej} y SNP_{epj}) sobre el cálculo de los *p*-valores bajo hipótesis nula. Para ello se utilizaron los datos de los 928 animales, cada uno con 44055 SNP. De dicho total de SNP, aquellos marcadores correspondientes al cromosoma 18 (1018 SNP) fueron reorganizados en dos escenarios: 1) escenario de "*Dependencia*": las filas de la matriz de genotipos se permutaron entre ellas; de este modo se mantuvo el Desequilibrio de ligamiento (LD) dentro de cromosoma pero se rompió la relación genotipo-fenotipo para los 1018 SNP.
2) Escenario de "*Independencia*": los genotipos de los animales se permutaron independientemente del marcador (resultando en Equilibrio de ligamiento (LE) entre marcadores) en el cromosoma. Del mismo modo se rompió la relación genotipo - fenotipo.

Se ajustó el modelo [3.3] para los dos escenarios y se calcularon ambas pruebas de hipótesis en cada uno de ellos (test1= SNP_{ej} y test2= SNP_{epj}). Se realizaron 200 permutaciones por escenario, y en cada permutación se calculó la matriz $\boldsymbol{G}$ ajustando [3.3]. Como resultado, la heredabilidad del carácter fue similar a la heredabilidad original debido a la relación existente con los otros 17 cromosomas que se conservaron intactos. A su vez, se calcularon los *p*-valores para los SNPs en el cromosoma 18 (que

ahora no se encontraban asociados) mediante las dos pruebas. Asumiendo hipótesis nula e independencia de los marcadores (es decir, SNP que no se encuentran ligados a un polimorfismo que controle la variación del carácter), una estrategia que controla el error tipo I apropiadamente (Yu *et al*., 2006), los 1018 *p*-valores siguen una distribución uniforme. Consecuentemente, para estimar los cuantiles empíricos de la distribución de la hipótesis nula, se utilizó una densidad uniforme U∼ (0, 1) generando grupos de 200 replicas para los 1018 *p*-valores.

3.2.9. Efectos de SNP mediante un modelo de marcador único

Se realizaron las pruebas de hipótesis para los efectos de marcador incluyendo de a un SNP a la vez. A tal efecto, se utilizó el enfoque "modelo mixto de asociación eficiente" ó (EMMA, por sus siglas en inglés de: Efficient Mixed-model Association") (Kang *et al*., 2008) contenido en el programa rrBLUP (Endelman, 2011) que utiliza el lenguaje de uso libre *R*. El modelo incluyó los efectos fijos de sexo y de un marcador a la vez; la variable aleatoria consistió en el efecto animal con matriz de covarianzas igual a la de relaciones genómicas (calculada con todos los marcadores, tal como se describe en [3.2]).

3.2.10. Proporción de la varianza explicada por los segmentos con gran efecto

Los SNP con los *p*-valores más bajos en cada cromosoma fueron selectos de modo de formar segmentos genómicos de largo igual a 2 Mb: una Mb a la izquierda y otra a la derecha del SNP. La elección de 2 Mb fue hecha con el mismo criterio que Hayes *et al*., (2010): el cambio en la caída de la tasa de LD de la población bajo estudio, valor cercano a r^2 = 0.2 en 1 Mb. Esta referencia implica que los marcadores que se encuentran más allá de 1 Mb contribuyen en forma mínima o nula a explicar la contribución del segmento a la varianza aditiva, y los efectos contenidos en ese segmento son independientes de las restantes contribuciones a la varianza aditiva. En adición, se han reportado segmentos de aproximadamente 2 Mb con contribución significativa en varios estudios de asociación (McClure et al., 2012; Fan et al., 2013; Rangkasenee et al., 2013; Do et al., 2014).

La proporción de la varianza asociada con cada segmento fue estimada calculando una matriz genómica $\boldsymbol{G}_1$ (tal como se describe en [3.1] y [3.2]), empleando todos los SNP pertenecientes al segmento, mientras que se calculó una matriz genómica $\boldsymbol{G}_2$ con los marcadores restantes que no pertenecían al segmento. A continuación se presenta la ecuación del modelo utilizado:

$$\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e} \qquad [3.15]$$

El vector aleatorio $\boldsymbol{a}_1$ contiene los efectos aditivos asociados con los SNP localizados en el segmento, y se distribuye tal que $\boldsymbol{a}_1 \sim N\left(0, \boldsymbol{G}_1\sigma^2_{A_1}\right)$. Además, $\boldsymbol{a}_2$ es el vector aleatorio de efectos aditivos asociados con todos los SNP exceptuando aquellos que se encuentran en $\boldsymbol{a}_1$, con distribución $\boldsymbol{a}_2 \sim N\left(0, \boldsymbol{G}_2\sigma^2_{A_2}\right)$. El modelo [3.15] permite calcular la proporción de la varianza explicada por el segmento de interés (varianza local) a partir de la varianza genómica explicada para todos los marcadores (varianza global). Las varianzas estimadas en el modelo [3.15] se compararon con aquellas estimadas bajo el modelo [3.3].

Hayes *et al*., (2010) emplearon un modelo similar para calcular la varianza de los segmentos. Sea empleando el modelo [3.15] o utilizando el enfoque de Hayes *et al*., (2010), resultó en similares estimaciones de los componentes de varianza. En la práctica, la ventaja de utilizar [3.15] consiste en que $\boldsymbol{G}_2$ es calculada a partir de $\boldsymbol{G}$ sustrayendo las columnas de $\boldsymbol{Z}$ relacionadas con el segmento que está siendo evaluado, como se describe a continuación:

$$\boldsymbol{G}_2 = \boldsymbol{G} - \boldsymbol{Z}_s \boldsymbol{Z}_s' \qquad [3.16]$$

La matriz $\boldsymbol{Z}_s$ contiene la columna relacionada con el segmento que se está probando. Por el contrario, en el modelo de Hayes et al. (2010), la matriz $\boldsymbol{G}$ es diferente de un segmento al otro. Además, el cálculo de $\boldsymbol{G}_1$ y $\boldsymbol{Z}_s \boldsymbol{Z}_s'$ es veloz y tiene en cuenta sólo aquellos SNP localizados en el segmento.

Se utilizó la Corrección de Bonferroni (BC) para ajustar el nivel de significancia de las pruebas de comparaciones múltiples (TCM). Si el genoma del cerdo es de aproximadamente ~2800 Mb de largo y el tamaño del segmento fue de 2 Mb, existirían 1400 segmentos a lo largo del genoma que correspondería al número de TCM. Entonces, para un valor de α = 0.05, el BC fue igual a 0.05/1400 = $3.571429e^{-05}$ (α ajustado o valor crítico). Por lo tanto, en orden de evaluar la significancia de los segmentos, se calculó un segundo *p*-valor para el test del Cociente de Verosimilitud ($p-\text{valor}_{\text{LRT}}$) de modo de compararlo con BC. Este $p-\text{valor}_{\text{LRT}}$ fue calculado por uno menos la función de distribución chi-cuadrado (χ^2) de una variable aleatoria con 0.5 grados de libertad (Self y Liang, 1987; Liang y Self, 1996), tal como se describe a continuación:

$$p-\text{valor}_{\text{LRT}} = 1-\chi^2(\text{LRT}) \qquad [3.16]$$

La notación $\chi^2(x)$ corresponde a la función de distribución de una variable aleatoria ($\boldsymbol{x}$) cuya función de densidad sigue una χ^2 y LRT es el test del cociente de verosimilitud que se obtiene contrastando los modelos apropiadamente elegidos.

3.3. Resultados

3.3.1 Exploración del genoma

Los *p*-valores de los 44055 SNP, calculados para SNP_{ej} estandarizando con la raíz cuadrada de la $\text{Var}(\hat{\boldsymbol{g}}_j)$ para el carácter bf10_13wk, se muestran en el Manhattan-plot de la Figura 3.1. Se pueden observar picos elevados con valores de $-\text{Log}_{10}(p\text{-valor}_j) > 5$ en los cromosomas 6 y 3, sugiriendo la existencia de regiones genómicas que afectan la variación del carácter en dichos cromosomas. Por el contrario, los *p*-valores obtenidos mediante $SNP_{ep\,j}$ (es decir, estandarizados con la PEV) mostraron varios picos, con un valor máximo de $-\text{Log}_{10}$(*p*-valor) de 0.20 (Figura 3.2). En esta última figura se observa el patrón que resulta de dividir los efectos de SNP por una constante. La normalización empleada fue la de $\text{Var}(\boldsymbol{g}_j) - \text{Var}(\hat{\boldsymbol{g}}_j)$. Dado que $\text{Var}(\boldsymbol{g}_j) = 2.6768$, la diferencia produjo un denominador constante e igual a 2.66, al que se le aplicó el operador raíz cuadrada de modo de obtener estadístico para la prueba de hipótesis. La Figura 3.2, mostró además señales en los cromosomas 1, 12, 14 y 18, las cuales no se observan en

la Figura 3.1. Sin embargo, esto se debería a que el estadístico tiene un denominador constante que sobreestimaría la verdadera variabilidad para algunos marcadores, hecho que lleva a reportar falsos positivos.

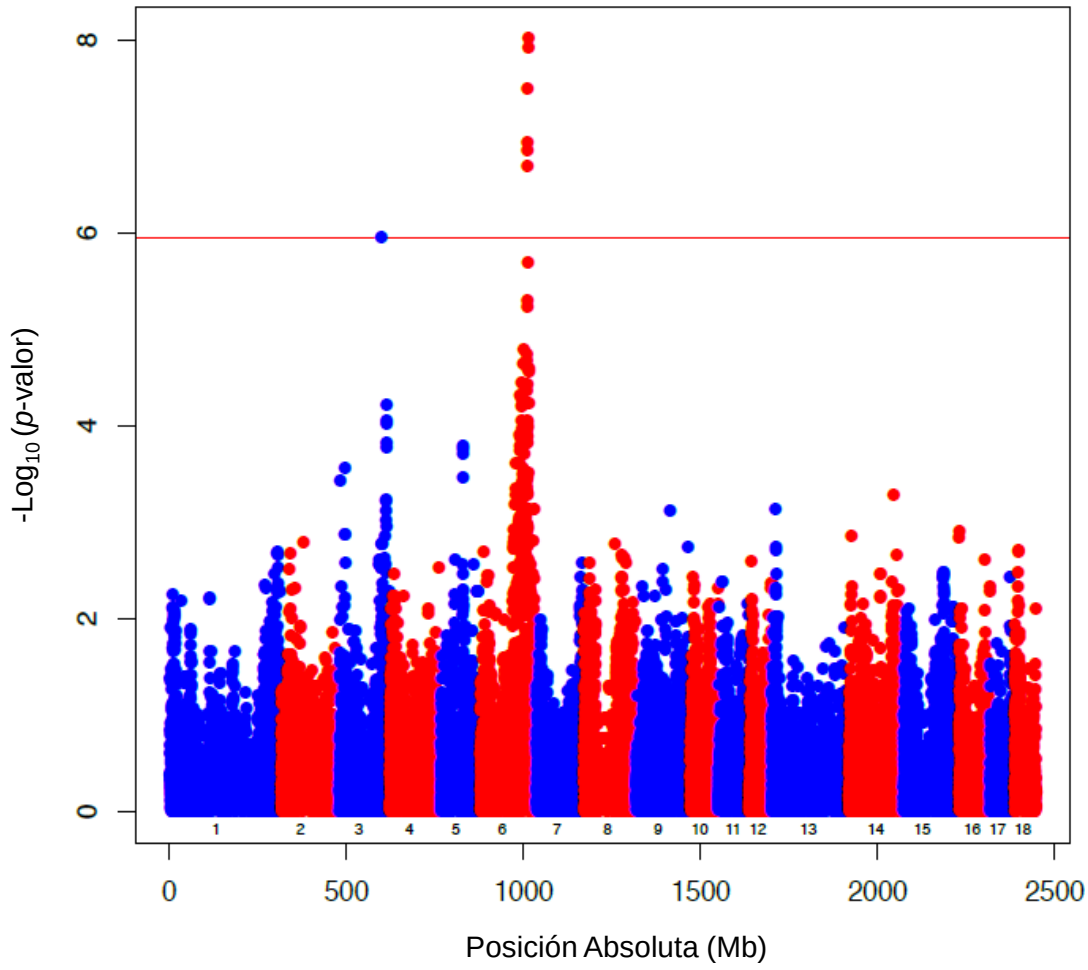


Figura 3.1. Manhattan-plot para el carácter grasa dorsal en la decima costilla (mm) medida a la semana 13 mediante la estandarización SNP_{ej}. Barrido genómico para 44055 SNP mediante estandarización $\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)$. Se grafica el valor de –Log(*p*-valor) (eje y) respecto de la posición absoluta de los SNP (en Mb, eje x). La línea roja representa un umbral de significancia del *p*-valor $< 1.1349\times10^{-6}$ a lo largo del genoma. Los números 1 a 18 identifican al cromosoma.

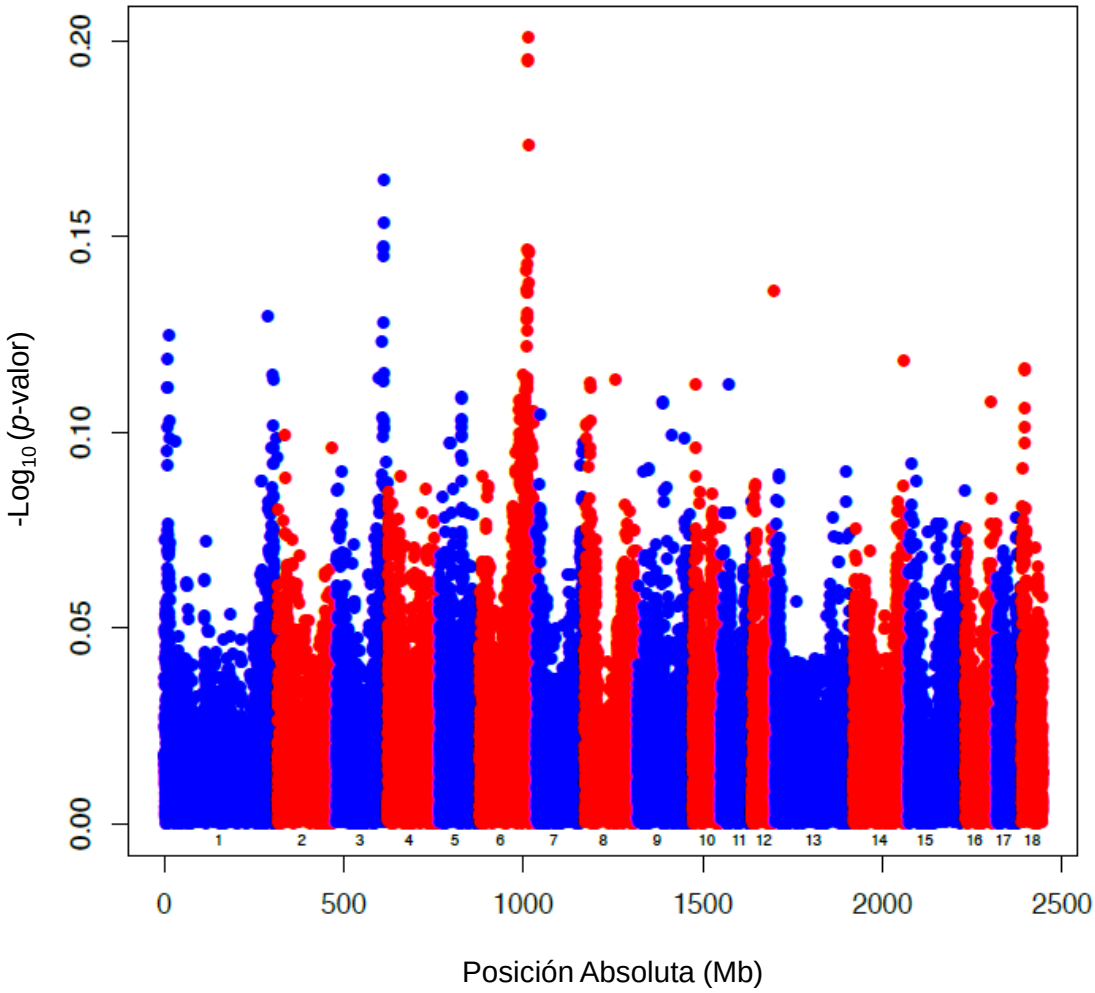


Figura 3.2. Manhattan-plot para el carácter grasa dorsal en la decima costilla (mm) medida a la semana 13 mediante estandarización SNP_{ejp}. Barrido genómico para 44055 SNP mediante estandarización $PEV = \mathrm{Var}\left(\boldsymbol{g}_j\right) - \mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)$. Se grafica el valor de –Log(*p*-valor) (eje y) respecto de la posición absoluta de los SNP (en Mb, eje x). Los números 1 a 18 identifican al cromosoma.

Para evaluar el error tipo I de las dos pruebas propuestas, se desarrolló una simulación empleando un "plasmode" o simulación estructurada (Vaughan *et al*., 2009). Un "plasmode" es un conjunto de datos simulado, desarrollado a partir de un grupo de datos reales, que genera una restricción en la estructura del muestreo. Brevemente, el "plasmode" utilizado en la simulación es una reorganización de una porción de los datos como se explica en la sección 3.2 Métodos, en el cual se trabajaron dos escenarios: 1) manteniendo la dependencia entre marcadores (estructura LD) y 2) asumiendo independencia del SNP (estructura LE). Los resultados se graficaron en un gráfico de quantiles o "Quantil-Quantil-Plot" (QQPlot) (Figura 3.3) utilizando –Log(*p*-valor) del SNP para cada estandarización. Se observó que los *p*-valores para el test1: SNP_{ej} obtenidos en el escenario bajo independencia (escenario 2: LE), presentaron una distribución idéntica de *p*-valores a los obtenidos mediante la distribución de referencia $U \sim (0, 1)$. En contraste, bajo el enfoque de dependencia (escenario 1: LD) se obtuvieron menos *p*-valores extremos cuando se los compara con la distribución uniforme. Este fenómeno es bien conocido en epidemiología humana (Klein *et al*.,

2006), donde la implementación de la corrección de Bonferroni para los *p*-valores de SNP asociados bajo el supuesto de independencia son demasiados conservadores en relación con el tamaño empírico (α) de las pruebas de hipótesis. Por otra parte, para la prueba de $SNP_{ep\ j}$ los *p*-valores obtenidos bajo independencia (escenario 2, LE) mostraron ser más conservadores. Asimismo, los resultados para el escenario de ausencia de independencia (LD) con este mismo test fueron aun más conservadores que aquellos mostrados bajo independencia (LE), resultados que no son graficados en el QQ-plot. Esto indicó que la utilización de la raíz cuadrada de $\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)$ como denominador para el test estadístico resultó ser más potente ni demasiado liberal, al compararlo con el error estándar de predicción $\left(\sqrt{\mathrm{Var}\left(\boldsymbol{g}_j\right)-\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)}\right)$.

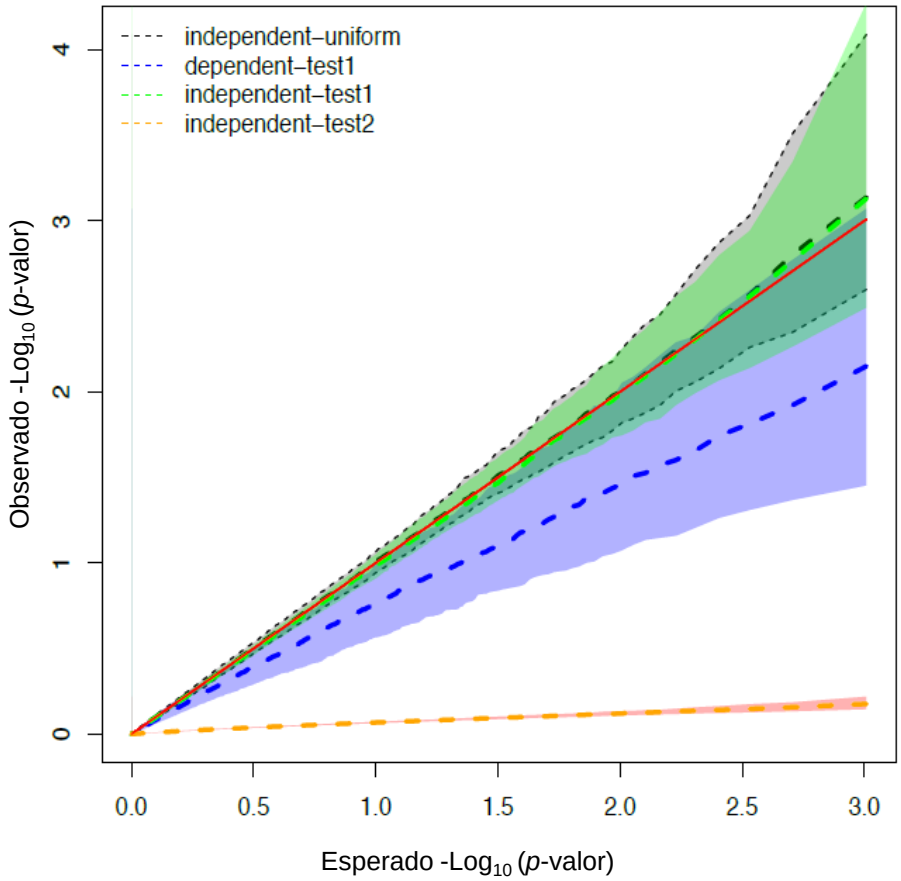


Figura 3.3. Gráfico QQ-plot para –Log(p-valores) observados y esperados obtenidos mediante simulación. Se utilizó como referencia una distribución Uniforme independiente U ~ (0, 1) para 1018 *p*-valores simulados (línea punteada negra). Test1(escenario1) = bajo dependencia (LD) y estandarización mediante $\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)$ (línea punteada azul). Test1(escenario2) = bajo independencia (LE) y estandarización mediante $\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)$ (línea punteada verde). Test2 (escenario2) = bajo independencia (LE) y estandarización mediante PEV (línea punteada naranja). Cada escenario posee 1018 *p*-valores permutados 200 veces. Las bandas de color representan intervalos de confianza del 95% (banda azul = test1(secenario1), banda verde = test1(escenario2), banda rosada = test2(escenario2).

3.3.2 Efectos de marcador obtenidos mediante modelo marcador (EMMA)

La estimación de un marcador a la vez mediante el algoritmo EMMA (Kang *et al.*, 2008), resultó en *p*-valores altamente similares (**Apéndice 2**) en relación con aquellos obtenidos con la prueba SNP_{ej} (**Apéndice 3**). Al implementar el algoritmo EMMA, el tiempo de cálculo requerido para estimar los efectos de los 44055 SNP fue de 84 minutos. Comparativamente, el modelo [3.3] junto con la estandarización de los efectos de marcador requirió de 29 minutos en total (CPU y memoria: Quad-core 2.7GHz AMD Opteron 8384, 256 GB). Los 29 minutos incluyeron el cálculo de la matriz $\boldsymbol{G}$, implementación del modelo animal, transformación de los GEBVs en efectos de SNP, y el cálculo del errores estándares necesarios para computar el estadístico de la prueba.

3.3.3 Evaluación de segmentos genómicos

En la mejora genética animal se evaluó la significancia de los segmentos genómicos sobre la base de las pruebas del cociente de verosimilitud (Hayes et al., 2010). Se utilizaron las predicciones de los efectos de marcador, calculadas con el algoritmo aquí propuesto, en la búsqueda de segmentos genómicos que expliquen significativamente la variabilidad aditiva del carácter, pero sólo se calculó el LRT para ciertos segmentos debido al esfuerzo computacional requerido para evaluar cada fracción del genoma. Así se escogió el SNP con el mayor *p*-valor (ó el –Log(*p*-valor) más alto) en cada cromosoma (Cuadro 3.1). Los tres segmentos de cromosomas con los *p*-valores más bajos (cromosomas 3, 5 y 6) se describen en la Cuadro 3.2. Los 15 cromosomas restantes se refieren al **Apéndice 4**. Todos los segmentos tuvieron una longitud de 2 Mb (1Mb hacia cada lado del SNP en relación con el *p*-valor más bajo ó SNP de referencia). Las estimaciones de los componentes de varianza y el Logaritmo de la verosimilitud (LogLikehood) obtenidos para [3.3] fueron comparados con el modelo [3.15] (Cuadro 3.2).

Cuadro 3.1. SNP seleccionados por el *p*-valor valor más bajo por cromosoma

Nombre-SNP	Cromosoma	Posición Mb	-Log_{10}(p-valor)	$\left\|\hat{\boldsymbol{g}}\right\|$
ALGA0104402	6	136.08	8.02	0.77
H3GA0010564	3	119.34	5.95	0.48
ALGA0032063	5	61.37	3.78	0.42
ALGA0081287	14	125.98	3.28	0.33
DRGA0011971	13	10.47	3.12	0.36
MARC0022304	9	94.99	3.12	0.42
ALGA0106422	16	111.82	2.90	0.28
ASGA0010464	2	62.15	2.79	0.30
ALGA0111088	8	88.01	2.77	0.48
ASGA0078865	18	10.72	2.70	0.49
ALGA0010607	1	302.88	2.69	0.43
MARC0082230	12	6.14	2.59	0.31
ALGA0045724	7	129.47	2.57	0.41
ASGA0092331	4	138.29	2.52	0.27
ASGA0070227	15	111.82	2.48	0.29
ASGA0077393	17	55.27	2.43	0.32
ASGA0045992	10	7.00	2.42	0.30
ALGA0060793	11	10.50	2.38	0.34

Nombre-SNP = nombre del marcador SNP, Posición Mb = posición física del marcador en Mega-Bases, $-Log_{10}$(*p*-valor) = −Logaritmo en base 10 del *p*-valor, $\left\|\hat{\boldsymbol{g}}\right\|$ = Valor absoluto del efecto del SNP estimado para el carácter *grasa dorsal en la decima costilla (mm) medida a la semana 13*.

Cuadro 3.2. Componentes de varianza y verosimilitud para modelos con y sin segmento.

Seg-cromosoma	6	3	5
SNP $-\mathbf{Log_{10}}$(p-valor)	8.02	5.94	3.78
Lk_m$_1$	−1227.938	−1227.938	−1227.938
Lk_m$_2$	−1210.800	−1223.178	−1224.540
LRT	34.28	9.52	6.80
***p*-valor$_{\mathbf{LRT}}$**	1.1×10^{-9}	6.5×10^{-4}	3.1×10^{-3}
VarE_m$_1$	3.70	3.70	3.70
VarA_m$_1$	2.68	2.68	2.68
VarE_m$_2$	3.73	3.67	3.69
VarA_m$_2$	1.95	2.42	2.55
segmVA	0.70	0.63	0.15
%segmVA	0.11	0.09	0.02

Seg-cromosoma = Número del cromosoma donde el segmento esta localizado, **m$_1$** = modelo [3.3] sin segmento: $y = X\beta + a + e$, **m$_2$** = modelo [3.15] con segmento $\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e}$, **SNP $-\mathbf{Log_{10}}$(p-valor)** = −Logaritmo en base 10 de los *p*-valore del SNP seleccionado para crear el segmento, **Lk_m$_1$** = −LogLikehood para m$_1$, **Lk_m$_2$** = −LogLikehood para m$_2$, **LRT** = test de la tasa de verosimilitud para m$_1$ and m$_2$, ***p*-valor$_{\mathbf{LRT}}$** = *p*-valor para LRT, **VarE_m$_1$** = Varianza del error $\left(\sigma_e^2\right)$ para m$_1$, **VarA_m$_1$** = Varianza aditiva $\left(\sigma_{\mathrm{A}}^2\right)$ para m$_1$,**VarE_m$_2$** = Varianza del error $\left(\sigma_e^2\right)$ para m$_2$, **VarA_m$_2$** = Varianza aditiva $\left(\sigma_{\mathrm{A}}^2\right)$ para m$_2$, **segmVA** = Varianza aditiva para el segmento $\left(\sigma_{\mathrm{A}_1}^2\right)$ para m$_2$, %**segmVa** = Proporción en porcentaje (%) del total de la varianza aditiva explicada por el segmento.

Los resultados del LRT muestran que el segmento en el cromosoma 6 tuvo un efecto significativo sobre la variabilidad en B13: $p\text{-valor}_{\mathrm{LRT}} = 1.133459e^{-0.9}$, valor demasiado bajo en comparación con el umbral de Bonferroni para 1400 segmentos (BC= 0.05/1400 = $3.571429e^{-05}$: α ajustado o valor crítico). Por el contrario, los segmentos localizados en los demás cromosomas no fueron significativos ($p\text{-valor}_{\mathrm{LRT}} > \mathrm{BC}$). La proporción de la varianza total explicada por el segmento en el cromosoma 6 **(**$-\mathrm{Log}_{10}$(*p*-valor = 8.02) fue de 11% $\left(\sigma_{\mathrm{A}_1}^2 = 0.698\right)$, un hecho que se reflejó en la disminución de la varianza aditiva estimada $\left(\sigma_{\mathrm{A}_1}^2\right)$ en el modelo [3.15]: 1.952 + 0.698 = 2.650. Este valor es muy cercano a 2.678, *i.e.* al valor de la varianza aditiva σ_{A}^2 estimado mediante el modelo [3.3] (ver Cuadro 3.2). Las estimaciones de σ_{A}^2 para los restantes segmentos genómicos de los demás cromosomas no disminuyeron en una proporción significativa.

3.4. Discusión

En este capítulo el objetivo se centró en desarrollar un algoritmo para un análisis de asociación genómica (GWAS) computacionalmente eficiente, a partir de las predicciones calculadas durante una evaluación genómica en una sola etapa (Legarra *et al*.,2009). La metodología emplea el estadístico suficiente del Predictor Lineal Insesgado de Mínima Varianza (BLUP) de los valores de cría, calculados con un

modelo animal. También, utiliza a la matriz $\boldsymbol{G}$ de (co)varianzas (o $\boldsymbol{H}$ en la evaluación en una sola etapa, Legarra *et al*., 2009)), a la matriz $\boldsymbol{Z}$ (para estandarizar los efectos de marcador), a los componentes de varianza y a la matriz $\boldsymbol{C}^{aa}$. Este conjunto de características hace que la implementación del algoritmo sea factible.

3.4.1. Varianza de los efectos de SNP

Los efectos de SNP $\hat{\boldsymbol{g}}_j$ fueron calculados transformando linealmente $\hat{\boldsymbol{a}}$ mediante la expresión [3.6]. Luego se calcularon las varianzas del efecto de cada marcador $\text{Var}\left(\hat{\boldsymbol{g}}_j\right)$ de modo de estandarizar los efectos de SNP, y el estadístico resultante se denominó SNP_{ej}**.** En consecuencia, se obtuvieron efectos de marcador estandarizados $-\text{Log}_{10}$(*p*-valores) mayores a 5. Adicionalmente, para cada marcador se calculó una segunda prueba de hipótesis estandarizada por la varianza de error de predicción, prueba que fue denominada $SNP_{ep\,j}$. Como resultado de la estandarización se obtuvieron valores máximos de $-\text{Log}_{10}$(*p*-valor) = 0.20 y señales importantes a lo largo del genoma (picos elevados en el Manhattan-plot) comparado con los obtenidos por el estadístico SNP_{ej} .

En adición se realizó una simulación con la misma estructura de la población analizada (el mimo número de animales y de marcadores SNP), con el fin de comparar el desempeño de ambas estandarizaciones para las pruebas de hipótesis (SNP_{ej} y SNP_{epj}), en cuanto hace a los valores empíricos del *p*-valor. Así, los SNP del cromosoma 18 se reordenaron bajo dos escenarios: 1) Dependencia entre genotipos (LD), y 2) Independencia entre genotipos (LE). No hubo asociación alguna entre el escenario y el fenotipo.

Se utilizó una distribución *uniforme* como referencia para los *p*-valores. Se observó para el escenario de independencia (LE), que la estandarización de los efectos de SNP mediante $\text{Var}\left(\hat{\boldsymbol{g}}_j\right)$ mostró una distribución empírica de los *p*-valores similar a los *p*-valores que se generan con la distribución *uniforme*. Sin embargo, bajo el escenario de dependencia (LD), la estandarización con $\text{Var}\left(\hat{\boldsymbol{g}}_j\right)$ ó prueba SNP_{ej} mostró un comportamiento conservador. A su vez, la estandarización por $\left[\text{Var}\left(\boldsymbol{g}_j\right) - \text{Var}\left(\hat{\boldsymbol{g}}_j\right)\right]$ de la prueba $SNP_{ep\,j}$, produjo resultados conservadores bajo independencia (LE) y marcadamente conservadores bajo dependencia (LD). En tal sentido, estandarizar los efectos de los SNP con $\text{Var}\left(\hat{\boldsymbol{g}}_j\right)$ en la prueba SNP_{ej} resultó en *p*-valores más similares a los valores simulados con la distribución uniforme. Asimismo, el desempeño del estadístico SNP_{ej} bajo el escenario de LD no se manifestó demasiado conservador, situación que podría extrapolarse a los genotipos estudiados en el presente capítulo. En adición, los *p*-valores calculados mediante el procedimiento EMMA (Kang *et al*., 2008) fueron similares a aquellos obtenidos mediante SNP_{ej}. Estos resultados sugieren que el SNP_{ej} controla razonablemente la tasa de error tipo I, o los falsos positivos. Asimismo, el tiempo de cálculo para ajustar el modelo animal [3.3] y obtener la varianza de los efectos de marcador [3.12] utilizando las expresiones [3.6 - 3.11], fue entre 2.5 a 3 veces menor comparado con la aplicación del modelo EMMA.

La estimación de los efectos de los marcadores SNP ($\hat{\boldsymbol{g}}_j$) ha sido calculada a partir de los valores de cría genómicos $\hat{\boldsymbol{a}}$ en distintos análisis de asociación (Garrick, 2007; Strandén y Garrick, 2009; Sun *et al*., 2011; McClure *et al*., 2012; Wang *et al*., 2012; Kumar *et al*., 2013). En alguno de estos trabajos, la varianza de los efectos de

SNP fue estimada empleando distintos enfoques. Wang *et al*. (2012) utilizaron la definición clásica de la varianza de los efectos aditivos descripta en genética cuantitativa (Falconer y Mackay, 1996), en la cual la varianza del *j*-esimo marcador se calcula mediante $\sigma^2_{A,j} = \hat{\boldsymbol{g}}^2_j 2p_j\left(1-p_j\right)$. Mientras que McClure *et al*. (2012) propusieron igualar la varianza de los efectos de SNP a $\left(2\sum p_j q_j\right)^{-1}\sigma^2_A$, para luego calcular la raíz cuadrada y normalizar el resultado. Esta última expresión se comporta de manera similar a la prueba de hipótesis $SNP_{ep\,j}$ [3.14], donde la estimación de los efectos de los SNP $\hat{\boldsymbol{g}}_j$ es dividida por un valor constante y cercano a la varianza a priori de 2.67, resultando en una prueba muy conservadora.

Por el contrario, la ventaja de utilizar el test de estandarización SNP_{ej} fue que cada efecto de SNP es escalado por su propia (y distinta) desviación estándar, en vez de utilizar una varianza a priori (McClure *et al*., 2012) o mediante el efecto de cada SNP elevado al cuadrado $\hat{\boldsymbol{g}}^2_j$ (Wang *et al*., 2012) tal como en una varianza. Además, el cálculo de SNP_{ej}, involucra a la misma varianza para marcadores y animales, *i.e.* $\sigma^2_g = \sigma^2_A$, y la matriz de incidencia estandarizada *Z* que es función de $2p_j\left(1-p_j\right)$. Adicionalmente, la matriz *Z* fue calculada a partir las frecuencias de animales no emparentados (generación F_0), empleando la varianza esperada particular de cada marcador (ver sección 3.2 Métodos). Más aún, el estadístico SNP_{ej} produce *p*-valores, probabilidades que son familiares para aquellos que utilizan el modelo de estimación de un marcador a la vez. Finalmente, otra ventaja de la prueba de hipótesis SNP_{ej} es el detectar y descartar falsos positivos, resaltando las posiciones con mayor efecto sobre todo el genoma.

3.4.2. El enfoque de los posible segmentos ("candidatos") responsables de la variación

En la etapa final de la investigación realizada en éste capítulo, se detectaron segmentos genómicos que expresaban señales importantes en la expresión del carácter. Para este propósito, se seleccionaron los SNP con los *p*-valores mínimos de la prueba SNP_{ej} [3.12], para extenderlos luego a un segmento con longitud igual a 2 Mb (1 Mb hacia cada lado desde el SNP selecto). Posteriormente se estimaron los componentes de varianza y el logaritmo de la función de verosimilitud para los modelos animales centrados [3.3] y [3.15]. Este último incluyó al vector aleatorio de los SNP ubicados dentro del segmento (a_1). Por último, se comparó el desempeño de ambos modelos. Hayes *et al.* (2010) utilizaron un modelo similar a [3.15], aunque los efectos aleatorios de marcador se calcularon a partir de los valores de cría y ajustados separadamente como un efecto de segmento. Hemos encontrado resultados similares usando cualquiera de ambos métodos. La ventaja de ajustar el modelo [3.15] radica en que la matriz *G* es la misma para todos los segmentos. Consecuentemente, se la calcula sólo una sola vez y se la almacena en memoria, mientras que con el modelo de Hayes et al., (2010) se requiere calcular una matriz *G* diferente para cada segmento. Consecuentemente, para obtener resultados similares a los encontrados con el modelo [3.15] se necesita más tiempo de cálculo y mayores requerimientos de memoria CPU.

Se utilizó el test del cociente de verosimilitud (LRT) entre los modelos [3.3] y [3.15] para evaluar el efecto de los segmentos en cada cromosoma, prueba que fue ajustada por la corrección de Bonferroni (BC). Se observó como resultado que el segmento localizado en el cromosoma 6 (posición física dentro del cromosoma 135 Mb

– 137 Mb) explicó significativamente un 11% de la varianza total del carácter. En análisis previos realizados en esta misma población experimental por Edwards et al. (2008 (a)) y Choi et al. (2010) con microsatellites y un número pequeño de marcadores SNP, se encontró una efecto significativo de una región en el cromosoma 6 (posición física dentro del cromosoma entre 135 Mb – 139 Mb) para el carácter grasa dorsal en la décima costilla, medida durante la semana 13 de edad.

En el sitio www.animalgenome.org/QTLdb/pig.html se puede encontrar información de 48 marcadores situados entre las 128 Mb y 139 Mb del cromosoma 6, que se encuentran asociados con la variabilidad genética de B13. Asimismo, estudios recientes de Choi *et al.* (2010), Fan *et al.* (2011) y Switonski *et al.* (2010) mostraron el protagonismo del cromosoma 6 en la expresión del citado carácter. En consecuencia, los resultados de nuestros análisis en B13 confirman la presencia de variabilidad genética aditiva en el cromosoma 6.

3.5. Conclusión

La presente investigación ha mostrado que él análisis de asociación realizado a partir de una transformación lineal de los valores de cría genómicos de la evaluación genética BLUP en una etapa, es un método muy eficiente para detectar segmentos responsables de la variabilidad genética de un carácter de herencia aditiva. Asimismo, la prueba de hipótesis que es función de la varianza de los efectos de cada marcador $\left(\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)\right)$ desarrollada en este capítulo, permite detectar regiones genómicas asociadas con la variabilidad reduciendo el número de falsos positivos y empleando menor tiempo de cómputo. Estos procedimientos permitieron detectar segmentos genómicos de aproximadamente de 2 Mb, formados alrededor del SNP con el *p*-valor más bajo en cada cromosoma, ajustando la prueba por la corrección de Bonferroni. La metodología de detección de asociación en segmentos “candidatos” es potencialmente generalizable al meta-análisis empleando varias poblaciones.

CAPÍTULO 4

Refinamiento de asociación genómica para caracteres de crecimiento y de deposición de grasa en una población experimental de cerdos [(4)]

[4] Gualdrón Duarte JL, Cantet RJC, Bernal Rubio YL, Bates RO, Ernst CW, Raney NE, Rogberg-Muñoz, Steibel JP. Refining genome-wide association for growth and fat deposition traits in an F2 pig population. 2016 (Aceptado Journal of Animal Science).

4.1. Introducción

Los estudios de asociación genómica que emplean una gran cantidad de marcadores SNP y mediciones fenotípicas de caracteres complejos (como por ejemplo, el crecimiento, la deposición de grasa, etc.; *Choi et al.*, 2010; Fontanesi *et al*., 2012; Gualdrón Duarte *et al*., 2014; Lee *et al*., 2011; Okumura *et al.*, 2013), ofrecen una buena oportunidad para detectar genes (o grupos de genes) determinantes de la expresión de caracteres económicamente relevantes. Un modo eficiente de realizar estos estudios de asociación, es mediante una transformación lineal de los valores de cría genómicos predichos (GEBVs) en los efectos de marcador SNP (Garrick, 2007; Gualdrón Duarte *et al*., 2014; Strandén y Garrick, 2009; Sun *et al*., 2011). Dichas estimaciones son luego graficadas (Manhattan-plot) para identificar las posiciones genómicas sugerentemente asociadas con la variabilidad genética del carácter. Más aún, dada la presencia de *desequilibrio gamético* (linkage disequilibrium, LD), los efectos de SNP no son estadísticamente independientes, con lo cual es mayor la potencia de detección testando por asociación para segmentos genómicos formados a partir de las posiciones de los SNP de mayor efecto (Gualdrón Duarte *et al*., 2014; Hayes *et al*., 2010) sobre la expresión del carácter. Asimismo, si la población experimental posee mediciones para varios caracteres, es posible encontrar una región genómica significativa asociada con la expresión común entre dichos caracteres. Consecuentemente, el objetivo de este capítulo es identificar regiones genómicas asociadas con la varianza aditiva de caracteres de crecimiento y de deposición de grasa, medidos en el transcurso de distintas semanas, para una población F_2 cruza. A tal efecto se empleó una transformación lineal de los GEBVs en efectos de marcador, para posteriormente seleccionar las regiones genómicas utilizando un esquema de "segmentos candidatos" (presentado en el capítulo 3).

4.2. Métodos

4.2.1. Animales y caracteres

La población analizada es la misma que se describió en el Capítulo 3 (Ver sección: Capítulo 3: Sección 3.3.1. Animales). Se recolectaron datos fenotípicos para los caracteres de crecimiento y deposición de grasa de 950 cerdos F_2 que se describen a continuación: grasa dorsal en la décima costilla (**bf10**), grasa dorsal en la última costilla (**lrf**), y área del músculo longissimus (**lma**) medidos en las semanas 10, 13, 16, 19 y 22 de edad. El Peso (**wt**) se registró durante las semanas 3, 6, 10, 13, 16, 19 y 22 de edad. Las mediciones de carne magra total libre de grasa (**fftoln**), tejido graso total (**tofat**), proteína del animal eviscerado (**mtfat**), y lípidos del animal eviscerado (**mtpro**) fueron registradas a la semana 22 de edad. El promedio de ganancia diaria (**ADG**) fue registrado entre las semanas 10 y 22 de edad, y el número de días (**days**) para alcanzar los 105 kg se calculó a partir de los caracteres **ADG** y **days.** Para más detalles se pueden consultar los trabajos de Edwards *et al.* (2008 (a)) y Choi *et al.* (2010).

4.2.2. Genotipado y control de calidad

Los genotipos y el control de calidad en la presente investigación fue el mismo que se presentó en el Capítulo 3 (Gualdrón Duarte *et al*., 2014). A manera de resumen, se empleó una genotipificación mediante dos paneles de marcadores SNP: 1) 411

animales (4 machos F_0 Duroc, 15 hembras F_0 Pietrain, 6 machos F_1, 50 hembras F_1 y 336 animales F_2) se genotiparon mediante el chip de Illumina PorcineSNP60 (62163-SNP) Genotyping beadchip (Illumina Inc.) (Ramos *et al*., 2009), y 2) 612 animales se genotiparon con un segundo panel compuesto de 9K tagSNP (GGP-Porcine, GeneSeek a Neogen Company, Lincoln, NE) (Badke *et al.*, 2013), he imputados posteriormente para el chip de 60K. Posteriormente, se realizó un control de calidad de los genotipos imputados el cual se describió en el Capítulo 3: Sección: *"3.2.2. Genotipado y control de calidad"* y se utilizó un MAF<0.05. Como resultado se obtuvo 1002 animales (F_0, F_1 y F_2) con 40569 marcadores SNP por animal.

4.2.3. Estimación de la matriz genómica de relaciones

Se calculó la matriz de relaciones genómicas a partir de los genotipos observados e imputados en alta densidad (aproximadamente 44K) utilizando la expresión [3.1] como se describió en el Capítulo 3. A manera de resumen, los genotipos fueron expresados en valores de dosis alélicas (Badke et al., 2013; Gualdrón Duarte et al., 2013; Gualdrón Duarte et al., 2014), dentro de una matriz cuadrada $\boldsymbol{M}$ con dimensión igual al número de animales (n), con elementos en en el intervalo [0, 2]. Se estandarizo luego $\boldsymbol{M}$ mediante la expresión [3.1], para producir $\boldsymbol{Z}$. Finalmente, el producto $\boldsymbol{Z}\,\boldsymbol{Z}'$ (ver [3.2]) es igual a la matriz de relaciones genómicas $\boldsymbol{G}$.

4.2.4. Modelo predictivo para caracteres

La ecuación del modelo animal centrado para la evaluación genómica de los caracteres evaluados es descripto por:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{a} + \boldsymbol{e}$$

El vector $\boldsymbol{y}$ contiene las observaciones para cada uno de los caracteres de crecimiento y de deposición de grasa descriptos en la sección 4.2.1; $\boldsymbol{X}$ es la matriz de incidencia que relaciona el sexo de los animales con el vector de efectos fijos $\boldsymbol{\beta}$; $\boldsymbol{a}$ es el vector aleatorio de los valores de cría con distribución $\boldsymbol{a} \sim N\left(0, \boldsymbol{G}_{\sigma^2_A}\right)$; $\boldsymbol{e}$ es el vector de efectos residuales con distribución $\boldsymbol{e} \sim N\left(0, \boldsymbol{I}_{\sigma^2_e}\right)$, e $\boldsymbol{I}$ es la matriz identidad. Se estimaron los componentes de varianza mediante REML, empleando el programa *regress* version 1.3-10 R (Clifford y McCullagh, 2006),

4.2.5. P-valores y rastreo genómico mediante Manhattan-plot

Como se describió en el Capítulo 3, el *p*-valor para cada *j*-ésimo SNP, se calculó restándole a 1 la probabilidad del valor absoluto del efecto del marcador estandarizado (SNP_{ej}, expresión [3.12]), para luego multiplicar por 2 el valor resultante de modo de obtener el *p*-valor$_j$ [3.13]. Posteriormente, se realizó el gráfico Manhattan (Manhattan-plot) para analizar cada carácter, y los *p*-valores de cada SNP a lo largo del genoma fueron evaluados como $-\text{Log}_{10}$(*p*-valor$_j$), teniendo en cuenta la posición física de cada SNP expresado en términos de Mega-bases (Mb) en el genoma.

4.2.6. Proporción de la varianza explicada por segmentos de gran efecto

Después de una primera exploración del genoma mediante los efectos de marcador, se seleccionaron los SNP con los *p*-valores más bajos (peak-SNP) y una tasa

de falsos descubrimientos (FDR, Storey y Tibshirani, 2003) menor a 0.05, para formar segmentos genómicos. Los segmentos se definieron bajo dos metodologías: 1) tomando todos los SNP dentro de 1 Mb a la izquierda y 1 Mb a la derecha del pico en SNP seleccionado en cada cromosoma (Gualdrón Duarte *et al*., 2014), y 2) si dos o más picos en SNP se encuentran localizados dentro de una distancia pequeña, entonces, se consideró un segmento que incluyó el SNP pico, y aquellos SNPs situados a una distancia de 2 Mb a la derecha y otra a la izquierda.

Se estimó la proporción de varianza asociada con cada segmento mediante una matriz genómica $\boldsymbol{G}_1$ (como se describe en el Capítulo 3, fórmulas: [3.1] y [3.2]) utilizando todos los SNP que pertenecieran al segmento, mientras que con los marcadores restantes (que no pertenecían al segmento) se construyó la matriz genómica $\boldsymbol{G}_2$. El modelo ajustado es similar al descripto en [3.15]:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e}$$

El vector $\boldsymbol{a}_1$ contiene los efectos aleatorios aditivos asociados con los SNP localizados en el segmento, y se distribuye cómo $\boldsymbol{a}_1 \sim N\left(0, \boldsymbol{G}_1\sigma^2_{A_1}\right)$, mientras que $\boldsymbol{a}_2$ es el vector de efectos aleatorios aditivos, asociados con todos los SNP exceptuando aquellos que se encuentran en $\boldsymbol{a}_1$ y con distribución $\boldsymbol{a}_2 \sim N\left(0, \boldsymbol{G}_2\sigma^2_{A_2}\right)$. El modelo [3.15] calcula la proporción de la varianza explicada por el segmento de interés (varianza local) a partir de la varianza genómica explicada para todos los marcadores (varianza global). Como se describió anteriormente, se compararon las varianzas estimadas en el modelo [3.15] que incluyen el efecto del segmento, con aquellas estimadas en el modelo reducido [3.3].

4.2.7. Prueba de significancia de segmentos

El nivel de significancia para las comparaciones múltiples y simultáneas fue ajustado mediante la corrección de Bonferroni (BC). Tal como se planteó en el Capítulo 3, si el genoma del cerdo es aproximadamente ~2800 Mb de longitud y el tamaño del segmento es de λ Mb, existirían 2800/ λ segmentos a lo largo del genoma, lo que correspondería al número de pruebas múltiples y simultáneas. Entonces, para un tamaño de prueba $\alpha = 0.05$, BC es igual a $0.05/(2800/\lambda) = \alpha^*$ (α ajustado o valor crítico). A los efectos de evaluar la significancia de los segmentos, se calculó un segundo *p*-valor para la prueba del cociente de verosimilitud ($p-\text{valor}_{\text{LRT}}$) de modo de poder comparar contra BC, tal como se describe en [3.16].

4.2.8. Rastreo de genes candidatos

Para aquellos segmentos candidatos que resultaron significativos, se implementó un método de indagación de genes potencialmente causales, para lo cual se empleó la base de datos Ensembl (http://ensembl.org/Sus_scrofa/Info/IndexGeneCards)

4.3. Resultados

4.3.1 Exploración del genoma

Se calcularon los *p*-valores para los 40569 efectos de SNP en cada carácter, tal como se describe en la sección de 4.2 *Métodos*, para luego graficarlos respecto de todo el largo del genoma (**Apéndice 6**), e identificar así las posiciones genómicas asociadas

con la variación del carácter. Se observaron picos importantes ($-Log_{10}$(*p*-valor) mayor a 5) en el gráfico de Manhattan-plot para los caracteres: **wt_birth**, **wt_13wk**, **ADG**, **mtpro**, **tofat**, **bf10** en las semanas 10, 13, 19, y 22, y para **lrf** en las semanas 10, 13, 16, 19 y 22, hecho sugerente de que dichas posiciones se asocien notoriamente con la variación genética aditiva (dado un valor de corte. La evaluación posterior de estos caracteres empleando FDR > 0.05, mostró significancia sólo para **wt_birth**, **bf10** y **lrf** en las semanas 10, 13, 16, 19 y 22. Sin embargo, **wt_13wk**, **ADG**, **mtpro** y **tofat** evidenciaron picos altos (**Apéndice 5**) dentro del Manhattan-plot, los cuales no pueden descartarse totalmente como posibles regiones conteniendo QTLs.

El SNP ALGA0045948 (en el cromosoma 7) resultó ser el SNP con el mayor efecto para **mtfat**, y mostró el valor $-Log_{10}$(*p*-valor) más alto para **Days** y **fftoln**, mientras que ALGA0045724 (en el cromosoma 7) fue el SNP con mayor efecto para **wt_13** y tuvo el valor más alto de $-Log_{10}$(*p*-valor) para **wt_16, 19wk** (**Apéndice 5**). Por tal motivo, la influencia de estas regiones en el cromosoma 7 en **Days**, **fftoln**, **wt_16**,**19wk** evidenciarían cierto indicio de señal con lo cual tampoco deberían descartarse como potenciales candidatos.

4.3.2 Selección de SNP por p-valor y FDR

El número de marcadores SNP con FDR < 0.05 fue 117, los cuales incluyeron a los caracteres: **wt_birth** (en el cromosoma 3), **bf10** en las semanas 10, 13, 19, y 22 (en los cromosomas 2, 3 y 6), y **lrf** en las semanas 10, 13, 16, 19 y 22 semanas (en los cromosomas 2, 5 y 6). Entre los 117 SNP se seleccionó por carácter y cromosoma, el SNP con el *p*-valor más reducido (peak-SNP). Aplicando este filtro, se seleccionaron ocho peak-SNP para formar segmentos de 2 Mb (1 Mb a cada lado del peak-SNP).

Cuadro 4.1. Marcadores SNP significativos por carácter

SNP-ID	Cromosoma	Posición (Mb)	Carácter	*p*-valor
MARC0087200	2	146.7230	bf10_16wk	$4{,}73^{-06}$
			lrf_19wk	$7{,}27^{-06}$
ALGA0075667	3	19.1643	wt_birth	$5{,}33^{-08}$
H3GA0010564	3	119.3397	bf10_13wk	$1{,}07^{-06}$
ALGA0031990	5	58.3026	lrf_16wk	$9{,}03^{-07}$
M1GA0008917	6	133.8855	bf10_22wk	$6{,}42^{-07}$
			lrf_16wk	$4{,}84^{-08}$
			lrf_22wk	$5{,}22^{-07}$
ASGA0029651	6	133.9292	bf10_10wk	$9{,}00^{-07}$
			lrf_10wk	$4{,}28^{-10}$
ALGA0122657	6	136.078566	lrf_13wk	$3{,}04^{-09}$
ALGA0104402	6	136.0844	bf10_13wk	$1{,}01^{-08}$
			bf10_16wk	$1{,}42^{-07}$
			bf10_19wk	$9{,}16^{-07}$
			lrf_19wk	$7{,}20^{-08}$

SNP-ID: Nombre del marcador SNP pico**, Posición (Mb):** Posición física del marcador SNP en Mega-Bases dentro del cromosoma, **Carácter:** bf10_10 (13,16,19wk): Grasa dorsal en la decima costilla (mm) en la semana 10 (13,16,19), lrf_10(13,16,19)wk: Grasa dorsal en la última costilla (mm) durante la semana 10 (13,16,19) (mm). ***p*-valor:** *p*-valor del marcador SNP por carácter.

En el cromosoma 6, cuatro de los ocho peak-SNP seleccionados que fueron significativos para 10 caracteres en total, se localizaron en una región de 2.2 Mb (entre las 133.8 Mb y 136 Mb). Para estos cuatro SNP se observa además lo siguiente: M1GA0008917 y ASGA0029651 se localizan consecutivamente y muestran LD (r^2) de 1, al igual que los dos SNP picos restantes ALGA0122657 y ALGA0104402 (**Apéndice 8**). Gráfico de LD para el cromosoma 6). Dado que estos cuatro peak-SNP se hallaron dentro de una región de tamaño 2.2 Mb flanqueada por un par de estos mismos marcadores, y a los efectos de tener en cuenta el LD, se adicionaron segmentos de genoma de 2Mb en cada lado de dicha región para producir una segmentación de aproximadamente 6 Mb que fue evaluada para los caracteres en los que los cuatro peak-SNP fueron significativos.

4.3.3 Significancia de los segmentos

Se compararon las estimaciones de los componentes de varianza y el logaritmo de la verosimilitud para el modelo animal centrado con el de segmentos separados, mediante el LRT. El cromosoma 3 fue significativo para **wt_birth**, explicando un 30% de la varianza aditiva total. De modo similar, se encontró significancia en el cromosoma 6 para **bf10_22wk, lrf_16wk, lrf_22wk, bf10_10wk, lrf_10wk, lrf_13wk, bf10_13wk, bf10_16wk, bf10_19wk, lrf_19wk,** explicando entre 4 a 10 % de la varianza aditiva total (Cuadro 4.2). Estos caracteres mostraron p–valores$_{LRT}$ más reducidos en comparación con la corrección de Bonferroni (BC, Cuadro 4.2), la cual tuvo un valor

crítico de $3.571429e^{-05}$. Este último valor proviene del cálculo siguiente: 2800 Mb/ 2 Mb = 1400 segmentos, $P_{critical} = \alpha^* = 0.05/ 1400 = 3.571429e^{-05}$.

En cuanto a **wt_birth**, la proporción de la varianza total explicada por el segmento de 2 Mb fue notablemente elevada en el orden de 30%, lo que sugiere que el efecto del segmento haya sido posiblemente sobreestimado. El valor del SNP pico "ALGA0075667" fue considerable ($-Log_{10}$(*p*-values) = 7.27) en relación con los de todos los marcadores restantes en la región genómica (**Apéndice 6**). Más aún, este marcador no mostró estar correlacionada y en LD con los SNP adyacentes dentro del segmento (**Apéndice 7**). Inversamente, los segmentos de 2 Mb en los cromosomas 2 (**bf10_16wk** y **lrf_19wk**), 3 (**bf10_13wk**) y 5 (**lrf_16wk**) no alcanzaron significancia. Sin embargo, un mayor número de observaciones o una estructura de población distinta a la aquí utilizada podrían mejorar la significancia y por lo tanto la presencia de posibles genes candidatos en estas regiones de los cromosomas 2, 3 y 5.

Al evaluar por el LRT el segmento de 6 Mb (en el cromosoma 6), se obtuvo significancia para los 10 caracteres estudiados puesto que los $p-valores_{LRT}$ fueron menores que el umbral BC = 0.0001073 (2800 Mb / 6 Mb = 4600 segmentos, con lo cual $\alpha^* = 0.05/ 466 = 0.0001073$, Cuadro 4.3). Sin embargo, y a pesar de alcanzar el nivel de significancia, cuando el tamaño del segmento aumentó de 2 Mb a 6 Mb, la proporción total de la varianza explicada cayó algo para **bf10_10wk** (de 9.59% a 7.6%), **lrf_13wk** (de 8.33% a 5.7%), **bf10_16wk** (de 8.1% a 7.2%) y **lrf_19wk** (de 8.3% a 7.5%) (ver Cuadro 4.3).

Cuadro 4.2. Componentes de varianza y logaritmo de la verosimilitud para modelos con y sin el segmento de 2 Mega-bases.

Seg-cromosoma	**2**	**2**	**3**	**3**	**5**	**6**	**6**	**6**	**6**
Carácter	bf10_16wk	lrf_19wk	wk_birth	bf10_13wk	lrf_16wk	bf10_22wk	lrf_16wk	lrf_22wk	bf10_10wk
SNP-ID	MARC0087200	MARC0087200	ALGA0075667	H3GA0010564	ALGA0031990	M1GA0008917	M1GA0008917	M1GA0008917	ASGA0029651
-Log$_{10}$(*p*-valor)	5.33	5.14	7.27	5.97	6.04	6.19	7.32	6.28	6.05
Lk_m$_1$	-1434.968	-1376.411	629.4213	-1228.382	-1100.176	-1969.69	-1100.176	-1581.85	-888.2347
Lk_m$_2$	-1431.301	-1370.919	618.1091	-1222.47	-1093.996	-1960.819	-1088.799	-1573.335	-875.8269
LRT	7.334517	10.98302	22.62452	11.82256	12.36101	17.74152	22.754	17.02978	24.81554
p-valor$_{LRT}$	0.002272237	0.000283139	5.15^-06	0.000177285	0.00013151	7.00^-06	4.81^-07	1.03^-05	1.61^-07
VarE_m$_1$	5.57	4.49	0.082	3.73	2.79	17.80	2.79	7.65	1.82
VarA_m$_1$	4.70	5.25	0.022	2.61	2.07	14.64	2.07	6.45	1.18
VarE_m$_2$	5.53	4.51	0.088	3.69	2.79	18.36	2.82	7.84	1.85
VarA_m$_2$	4.46	4.77	0.014	2.33	1.88	11.56	1.68	5.27	0.88
segmVA	0.38	0.53	0.039	0.87	0.35	1.47	0.26	0.60	0.29
%segmVA	0.04	0.05	0.30	0.13	0.07	0.05	0.06	0.04	0.10

Seg-cromosoma= Número del cromosoma donde esta localizado el segmento de 2 Mega-bases, **Carácter:** bf10_10 (13,16,19wk): Grasa dorsal en la decimal costilla (mm) en la semana 10 (13,16,19) de edad, lrf_10(13,16,19)wk: Grasa dorsal en la última costilla (mm) en la semana 10 (13,16,19) de edad, **SNP-ID** = nombre del marcador SNP **m$_1$**= modelo sin el segmento: $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{a} + \boldsymbol{e}$, **m$_2$** = modelo con el segmento $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e}$, **SNP –Log$_{10}$(p-valor)**= –Logaritmo en base 10 del *p*-valor del SNP seleccionado para formar el segmento, **Lk_m$_1$**= –Logaritmo de la verosimilitud para le modelo m$_1$, **Lk_m$_2$**= –Logaritmo de la verosimilitud para el modelo m$_2$, **LRT**= El test de la tasa de la verosimilitud para m$_1$ y m$_2$, ***p*-valor$_{LRT}$**= *p*-valor para LRT, **VarE_m$_1$**= Varianza del error (σ_e^2) de m$_1$, **VarA_m$_1$**= Varianza Aditiva (σ_A^2) de m$_1$, **VarE_m$_2$**= Varianza del Error (σ_e^2) de m$_2$, **VarA_m$_2$**= Varianza Aditiva (σ_A^2) de m$_2$, **segmVA**= Varianza aditiva del segmento $(\sigma_{A_1}^2)$ of m$_2$, **%segmVA**= Proporción en % de la varianza total explicada por el segmento.

Cuadro 4.2 Componentes de varianza y logaritmo de la verosimilitud para modelos con y sin el segmento de 2 Mega-bases

Seg-cromosoma	**6**	**6**	**6**	**6**	**6**	**6**
Carácter	lrf_10wk	lrf_13wk	bf10_13wk	bf10_16wk	bf10_19wk	lrf_19wk
SNP-ID	ASGA0029651	ALGA0122657	ALGA0104402	ALGA0104402	ALGA0104402	ALGA0104402
-Log$_{10}$(p-valor)	9.37	8.52	7.96	6.85	6.04	7.14
Lk_m$_1$	-454.6966	-651.4428	-1228.382	-1434.968	-1764.37	-1376.411
Lk_m$_2$	-440.0969	-636.8158	-1214.745	-1422.7	-1755.351	-1363.176
LRT	29.19952	29.2539	27.27313	24.53514	18.03773	26.46939
p-valor$_{LRT}$	1.61^-08	1.56^-08	4.42^-08	1.87^-07	5.97^-06	6.75^-08
VarE_m$_1$	0.75	1.07	3.73	5.57	11.16	4.49
VarA_m$_1$	0.38	0.76	2.61	4.70	10.06	5.25
VarE_m$_2$	0.76	1.07	3.78	5.64	11.25	4.56
VarA_m$_2$	0.27	0.60	1.98	3.73	8.55	4.25
segmVA	0.08	0.15	0.53	0.83	1.14	0.80
%segmVA	0.08	0.08	0.08	0.08	0.05	0.08

Seg-cromosoma= Número del cromosoma donde esta localizado el segmento de 2 Mega-bases, **Carácter:** lrf_10(13,16,19)wk: Grasa dorsal en la última costilla (mm) en la semana 10 (13,16,19) de edad, **SNP-ID** = nombre del marcador SNP **m$_1$**= modelo sin el segmento: $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{a} + \boldsymbol{e}$, **m$_2$** = modelo con el segmento $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e}$, **SNP –Log$_{10}$(p-valor)**= –Logaritmo en base 10 del *p*-valor del SNP seleccionado para formar el segmento, **Lk_m$_1$**= –Logaritmo de la verosimilitud para le modelo m$_1$, **Lk_m$_2$**= –Logaritmo de la verosimilitud para el modelo m$_2$, **LRT**= El test de la tasa de la verosimilitud para m$_1$ y m$_2$, ***p*-valor$_{LRT}$**= *p*-valor para LRT, **VarE_m$_1$**= Varianza del error (σ_e^2) de m$_1$, **VarA_m$_1$**= Varianza Aditiva (σ_A^2) of m$_1$, **VarE_m$_2$**= Varianza del Error (σ_e^2) of m$_2$, **VarA_m$_2$**= Varianza Aditiva (σ_A^2) of m$_2$, **segmVA**= Varianza aditiva del segmento $(\sigma_{A_1}^2)$ of m$_2$, **%segmVA**= Proporción en % de la varianza total explicada por el segmento.

Cuadro 4.3. Componentes de varianza y logaritmo de la verosimilitud para modelos con y sin el segmento de 6 Mega-bases en cromosoma 6

SNP: M1GA0008917, ASGA0029651, ALGA0122657, ALGA0104402										
Carácter	bf10_22wk	lrf_16wk	lrf_22wk	bf10_10wk	lrf_10wk	lrf_13wk	bf10_13wk	bf10_16wk	bf10_19wk	lrf_19wk
Lk_m$_1$	-1969.69	-1100.176	-1581.85	-888.2347	-454.6966	-651.4428	-1228.382	-1434.968	-1764.37	-1376.411
Lk_m$_2$	-1961.178	-1087.883	-1570.727	-876.1334	-438.8021	-638.4396	-1211.147	-1423.301	-1755.646	-1364.51
LRT	17.02325	24.58635	22.24448	24.2026	31.78912	26.00623	34.46962	23.33323	17.44731	23.80281
p-valor$_{LRT}$	1.03^-05	1.82^-07	6.30^-07	2.23^-07	4.15^-09	8.61^-08	1.03^-09	3.54^-07	8.20^-06	2.76^-07
VarE_m$_1$	17.80	2.79	7.65	1.82	0.75	1.07	3.73	5.57	11.16	4.49
VarA_m$_1$	14.64	2.07	6.45	1.18	0.38	0.76	2.61	4.70	10.06	5.25
VarE_m$_2$	18.33	2.80	7.84	1.85	0.75	1.08	3.79	5.67	11.30	4.60
VarA_m$_2$	11.40	1.64	5.00	0.87	0.27	0.59	1.79	3.65	8.33	4.15
segmVA	1.68	0.40	0.73	0.22	0.08	0.10	0.68	0.72	1.25	0.71
%segmVA	0.05	0.08	0.05	0.08	0.08	0.06	0.11	0.07	0.06	0.08

SNP= nombre de los peak-SNP que se encuentran en el segmento de 6 Mb , **Carácter:** bf10_10 (13,16,19,22wk): Grasa dorsal en la decima costilla (mm) en la semana 10 (13,16,19) de edad, lrf_10(13,16,19,22)wk: Grasa dorsal en la última costilla (mm) en la semana 10 (13,16,19,22) de edad, **SNP-ID** = nombre del marcador SNP **m$_1$**= modelo sin el segmento: $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{a} + \boldsymbol{e}$, **m$_2$** = modelo con el segmento $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e}$, **SNP –Log$_{10}$(p-valor)**= –Logaritmo en base 10 del *p*-valor del SNP seleccionado para formar el segmento, **Lk_m$_1$**= –Logaritmo de la verosimilitud para le modelo m$_1$, **Lk_m$_2$**= –Logaritmo de la verosimilitud para el modelo m$_2$, **LRT**= El test de la tasa de la verosimilitud para m$_1$ y m$_2$, ***p*-valor$_{LRT}$**= *p*-valor para LRT, **VarE_m$_1$**= Varianza del error (σ_e^2) de m$_1$, **VarA_m1**= Varianza Aditiva (σ_A^2) of m$_1$, **VarE_m$_2$**= Varianza del Error (σ_e^2) de m$_2$, **VarA_m$_2$**= Varianza Aditiva (σ_A^2) de m$_2$, **segmVA**= Varianza aditiva del segmento $(\sigma_{A_1}^2)$ de m$_2$, **%segmVA**= Proporción en % de la varianza total explicada por el segmento.

4.3.4 Indagación de genes en segmentos que podrían explicar variación aditiva.

Se evaluó la funcionalidad de genes potencialmente candidatos en un segmento del cromosoma 3, y se dedujo que dichos genes podrían involucrarse en la determinación del peso al nacimiento (**wt_birth**): gen de la fibrosina (localizado en 18187484 -18193499 bp) y el gen de la miosina de cadenas ligeras (localizado en 18376798 - 18379771 bp). Ambos se relacionan con el tejido cartilaginoso y el desarrollo del musculo esquelético. Asimismo, la evaluación del segmento candidato de 6 Mb (131.9 Mb -137.9 Mb) en el cromosoma 6, sugirió una relación con la deposición de grasa. Como resultado, y en adición a lo informado para los genes *PDE4B*, *C1orf141* (Lee et al., 2011) y *LEPROT* (Óvilo *et al.,* 2005), el gen *SERBP1* fue observado responsable potencial de la señal significativa para los caracteres de deposición de grasa. El gen *SERBP1* está localizado en 134.068.990 -134.081.998 bp, y se asocia con la regulación de mRNA y el metabolismo lipídico (http://www.ensembl.org/).

4.4. Discusión

El principal objetivo de este capítulo fue el mejorar la identificación de regiones genómicas (segmentos) asociados con la variación aditiva en caracteres de crecimiento y de deposición de grasa, y la observación sobre cuáles de estas regiones determinan la expresión de varios caracteres evaluados simultáneamente.

4.4.1 Asociación genómica

En una primera etapa, se realizaron GWAS para cada carácter en la búsqueda de regiones genómicas relevantes a su expresión. Luego, entre las regiones halladas con mayor relevancia, se eligieron SNP con el menor *p*-valor (peak-SNP) y FDR < 0.05, por cada cromosoma y cada carácter. Finalmente, se seleccionaron 8 peak-SNP localizados en los cromosomas 2, 3, 5 y 6 que fueron significativos para 11 caracteres (**wt_birth**, **bf10** en la semana 10, 13, 16, 19, y 22, y **lrf** en las semanas 10, 13, 16, 19 y 22). Se selección especialmente una región en el cromosoma 6 de 2.2 Mb localizada entre las posiciones 133.8 Mb a 136.2 Mb que contó con 4 peak-SNP.

Los resultados obtenidos para **wt_birth** mostraron una posición relevante en el cromosoma 3 (posición 19.1 Mb). Estudios previos informaron un QTL en la misma región del cromosoma 3. Así, Liu *et al.*(2007) encontraron un supuesto QTL con pico en la posición 17.8 Mb para una población similar (Duroc × Pietrain), y Malek *et al.* (2001) detectaron un supuesto QTL en la posición 19 Mb para una población cruza Berkshire × Yorkshire. Además, Edwards *et al*. (2008 (a)), observaron mediante micro-satélites en la misma población con la cual se realizó la presente investigación, un QTL en el cromosoma 5 para **wt_birth**. Estas dos regiones en los cromosomas 3 y 5 podrían considerarse candidatas a alojar QTL para el carácter.

Empleando micro-satélites y un pequeño número de marcadores en la población objeto de nuestro análisis, Edwards *et al.* (2008 (a)) y Choi *et al.*(2010) observaron picos de significancia en el cromosoma 6, entre las posiciones 134 a 143 Mb y entre 108 a 143 Mb, las que se asociaron con **bf10** y **lrf** medidos en la semanas 10, 13, 16, 19, y 22. En adición, en un estudio más reciente con animales de la raza Duroc (Okumura *et al.,* 2013) se detectó una región en el cromosoma 6 localizada entre las posiciones 135.1 a 136.2 Mb que fue relevante para grosor de grasa dorsal, en acuerdo

con estudios previos de mapeo de QTLs y de asociación (Fontanesi *et al*., 2012; Lee *et al*., 2011). Más aún, Lee *et al*. (2011) reportaron una significativa asociación para dos marcadores SNP (MARC0083918 y ASGA0029677) con espesor de grasa dorsal, los cuales están localizados dentro del segmento de 2.2 Mb (133.8 Mb a 136 Mb) en el cromosoma 6 descrito en la sección de *Resultados*, y se encuentran cercanos a los genes PDE4B y C1orf141, que participan en el metabolismo graso. Finalmente, Óvilo *et al.* (2005) han evaluado el gen LEPROT por su gran significancia con caracteres de deposición grasa, y encontrado una región reducida (130-132 cM) incluida en el segmento del cromosoma 6 que fuera descripta previamente. Finalmente, Muñoz *et al.* (2009) evaluaron el efecto de este gen, encontrando otra señal para deposición de grasa entre 60 – 100 cM.

Respecto a los resultados en los cromosomas 2 (**bf10_16wk** y **lrf_19wk**), 3 (**bf10_13wk**) y 5 (**lrf_16wk**), en estudios previos mediante micro-satélites se encontraron evidencia para los cromosomas 2 (Koning et al., 1999; Lee et al., 2003; Kim et al., 2006) y 5 (Kim *et al.*, 2006) en espesor de grasa, y en el cromosoma 3 para el espesor de grasa lateral (Liu *et al*., 2007). Sin embargo, las regiones genómicas donde se localizaron los peak-SNP en la presente investigación no se encuentran incluidas en las descriptas dentro de los trabajos citados. A pesar de ello, la significancia alcanzada por los SNP en los cromosomas 2, 3 y 5 sugiere la necesidad de una revisión más detallada en un futuro estudio de asociación.

4.4.2 Segmento significativo

Una vez detectados, a los 8 SNP de significancia relevante se los prolongó en un segmento con 1 Mb a cada lado del peak-SNP, para luego ajustar modelos que incluían o no el efecto del segmento correspondiente y estimarse los componentes de varianza y el logaritmo de la verosimilitud tal como se propusiera en el capítulo anterior (Gualdrón Duarte *et al*., 2014; Hayes *et al*., 2010). A su vez, se evaluó la significancia del efecto del segmento en cada cromosoma mediante el test del cociente de verosimilitud (*p*–valor$_{LRT}$) y se lo comparó con un *p* – valor ajustado mediante la corrección de Bonferroni. La misma metodología fue aplicada para un segmento de 6 Mb de longitud que afectó significativamente diez caracteres. Se observaron entonces segmentos de 2 Mb localizados en el cromosoma 3 y 6 con efecto significativo (*p*-valor$_{LRT}$ < BC) para el peso al nacer (**wt_birth**), la grasa dorsal en la décima costilla (**bf10**) y la grasa dorsal en la última costilla (**lrf**), caracteres medidos a diferentes semanas de edad, que explicaron entre 4 y 10 % de la varianza aditiva total de cada carácter. Se halló un resultado inusual con peso al nacer (**wt_birth**) al observarse un segmento de 2 Mb explicando 30% de la variación total. Se evalúa como potencial causa de la sobreestimación en la contribución de dicho segmento a la ausencia de correlación o LD entre el SNP pico y los marcadores SNP adyacentes (**Apéndice 7**).

Por otra parte, los segmentos de 2 Mb localizados en los cromosomas 2, 3 y 5 fueron encontrados no significativos (*p*-valor$_{LRT}$ > BC), aún cuando algunos de ellos explicaron una proporción importante de la varianza aditiva. Así por ejemplo para **bf10_13wk** en el cromosoma 3, alrededor de 13% de la varianza aditiva fue explicada por el segmento pero de modo no significativo (Tabla 2). Cabe señalar que el número de genotipos incluidos en el segmento de 2 Mb en el cromosoma 3 fue de 171. Alternativamente, para el mismo carácter y en el cromosoma 6 el segmento de 2 Mb contuvo 356 genotipos. Asimismo, fue mayor el número de replicas por genotipo en el segmento de 2 Mb del cromosoma 6. Se hallaron resultados similares con **lrf_16wk** en el cromosoma 5 (segmento no significativo) y 6 (segmento significativo), con 69 y 223

genotipos, respectivamente. Se puede inferir entonces que la escasa cantidad de genotipos por segmento (en términos estadísticos, un menor número de grados de libertad), disminuyeron la potencia de las pruebas de hipótesis (Christensen, 2011).

El segmento de 6 Mb en el cromosoma 6 afectó significativamente diez caracteres. Sin embargo la proporción de la varianza aditiva explicada en **bf10_10wk**, **bf10_16wk**, **lrf_13wk** y **lrf_13wk** decreció levemente, en comparación con la variación observad en el segmento de 2 Mb utilizando el SNP pico por carácter. En estos casos la inclusión de más marcadores SNP (que aumentan el tamaño de segmento) podría estar agregando regiones sin efecto alguno en la expresión del carácter y, consecuentemente, aumentando la varianza del error. En forma opuesta, para **bf10_13wk**, **bf10_19wk**, **bf10_22wk** y **lrf_10wk**, **lrf_16wk**, **lrf_19wk** y **lrf_22wk**, la proporción de la varianza aumentó y el agregado de SNP estaría abarcando regiones donde se encuentran genes asociados con la variación aditiva de dichos caracteres. Nótese, además, que en esta región se encontraron señales altamente significativas de cuatro marcadores en diez caracteres. Como estos marcadores mostraron un alto grado de LD tomados de a pares, se podría reflexionar que los cuatro detectaron una señal en común que podría estar localizada entre 133.8 a 136 Mb. (ver Apéndice, gráfico de LD). Los resultados de la presente investigación confirman la presencia de variabilidad genética para los caracteres de deposición de grasa en áreas específicas del cromosoma 6.

4.4.3 Rastreo genes candidatos en segmentos significativos

Malek *et al.* (2001) y Liu *et al*. (2007) reportaron evidencias de QTL segregando para peso al nacer en la región del cromosoma 3 detectada en la presente investigación. Del análisis de regiones ortólogas se observan QTL con cierto efecto en hacienda de carne para caracteres al parto (Sahana et al., 2011) y para densidad ósea en ovinos (Campbell et al., 2003). Estos posibles QTLs podrían relacionarse con el peso al nacer, debido a que en humanos se halló una relación negativa entre la densidad ósea y el peso al nacer (Steer et al., 2014). A su vez, genes ubicados en la cercanía de la región en el cromosoma 3 podrían relacionarse con el crecimiento embrionario y, por lo tanto, ser potenciales candidatos. Tal es el caso de la fibrosina (FBRS) que tiene su acción regulando el desarrollo de miofibroblastos, los cuales juegan un rol importante en el desarrollo embrionario (Prakash *et al*., 2007), y de la miosina de cadena liviana (*HUMMLC2B*), gen relacionado con el desarrollo de fibras musculares que se expresó diferencialmente en el musculo esquelético del cerdo durante los días prenatales 33 al 65 (Mei *et al*., 2008).

La búsqueda de genes en el segmento del cromosoma 6 para caracteres de deposición de grasa, resultó en el candidato *SERBP1*. Este gen produce RNAm PAI1 de fijación de proteínas, que juega un rol importante en la regulación y estabilidad de RNAm PAI1 (ver figura Heberlein *et al.*, 2012:
http://atvb.ahajournals.org/content/32/5/1271/F6.large.jpg),
y consecuentemente favorece el traslado de proteína PAI1 (Heaton *et al*., 2001; Heberlein *et al*., 2012). Sobre la expresión de la proteína PAI1 (Codificada mediante el gene *SERPINE1*) en humanos podría estar relacionada con el síndrome metabólico ligado a la obesidad (Alessi y Juhan-Vague, 2006). En bovinos de carne, PAI1 estuvo fuertemente expresado en animales con un espesor importante de la grasa dorsal (Jin *et al.*, 2012). Por otra parte, se encontró en ratones obesos niveles enriquecidos de proteína SERBP1 (Heberlein *et al*., 2012), mientras que en pollos el RNAm SERBP1 estuvo fuertemente expresado en la grasa abdominal, dentro de líneas selectas para alto

contenido graso (Resnyk *et al*., 2013). Estos resultados sugieren que el gen SERBP1 podría estar involucrado en la regulación de la deposición de grasa, así como la proteína SERBP1 estabiliza RNAm PAI1, y PAI1 se relaciona con el metabolismo de los lípidos.

4.5. Conclusión

En el presente capítulo se describe una exploración genómica para localizar efectos de marcador relevantes y, posteriormente, a partir de las posiciones detectadas significativas más relevantes por cada cromosoma, producir segmentos candidatos que permitan identificar mejor las regiones involucradas en la expresión de caracteres complejos. Como resultado, se detectaron dos regiones que afectarían la variación genética aditiva: una en el cromosoma 3 para peso al nacer, y otra localizada en el cromosoma 6 para caracteres de deposición grasa. Estos resultados podrían ser agregados, junto con los de otras poblaciones experimentales con similar estructura y caracteres evaluados, dentro de un meta-análisis para refinar la búsqueda de QTLs de importancia económica en la industria cárnica porcina. Además, y considerando que el cerdo es un razonable modelo biomédico para identificar factores genéticos causales de predisposición de formas comunes en obesidad en humanos (Lee *et al.*, 2011), la localización e identificación de estas regiones y posibles genes candidatos afectando la deposición de grasa ayudaría en los estudios de salud pública.

CAPÍTULO 5

Discusión General

La presente tesis proporciona una contribución teórica y metodológica, para capitalizar la información de unos pocos genotipos en HD y un grupo mucho mayor en LowD. El fin último es analizar datos fenotípicos de crecimiento y deposición de grasa en búsqueda de asociaciones con regiones genómicas, empleando una población experimental de cerdos cruza Duroc × Pietrain. Para lograr este objetivo, se definió un esquema de genotipado en la población cruza usando dos paneles marcadamente distintos (en número de SNP y en precio), el desarrollo de un algoritmo para estimar los efectos de los marcadores de manera eficiente y, la aplicación de dicho algoritmo en el análisis de datos fenotípicos experimentales para indagar sobre la asociación entre regiones genómicas significativas y la variación genético aditiva de caracteres económicamente relevantes para la industria.

Un aspecto importante en esta investigación es la obtención de genotipos imputados en HD mediante paneles de genotipos LowD, explotando toda la información de pedigrí de la población (Daetwyler *et al.*, 2011; Druet y Georges, 2010; Hickey et al., 2011; Hickey *et al.* 2012; Huang et al., 2012) y de LD entre marcadores (Badke et al., 2013). En este sentido, mediante simulación estocástica, se encontró que se requieren 1200 marcadores distanciados en promedio cada 2.1 Mb, para obtener en la generación F_2 una exactitud de imputación igual o superior a 0.97. Sin embargo, se observó también que el diseño de un panel con estas características no produce una mejora en la relación costo-beneficio del análisis, cuando se la comparó con el empleo de un panel comercial en LowD de 9K: con el dinero invertido en diseñar el panel, se puede adquirir un panel comercial (por ejemplo, el 9K porcino) con una mayor densidad de SNP (ó mayor información genómica) y mayores exactitudes de imputación (Cuadro 2.1). Por tal motivo, se utilizó el panel comercial de 9K en LowD para imputar en alta densidad y realizar el análisis de asociación con la población cruza: las generaciones F_0 y F_1 en HD (60K) y la F_2 en LowD (9K). La exactitud de imputación de los genotipos F_2 fue 0.99, cuando se utilizó la información del pedigrí (Figura 2.3), sugiriendo que la estrategia de genotipado es muy efectiva en cuanto a relación costo-beneficio para realizar GWAS en las poblaciones experimentales existentes.

Un aporte novedoso para el rastreo genómico de posiciones altamente significativas en la expresión de caracteres, se concentra en el algoritmo propuesto en esta investigación. El algoritmo selecciona regiones genómicas formadas a partir de los efectos SNP de mayor –Log(*p*-valor) ó "SNP pico" dentro de un cromosoma. Los efectos de los SNP fueron estimados mediante una transformación lineal de los GEBVs, una metodología altamente implementada en estudios de GWAS (Garrick, 2007; Strandén y Garrick, 2009; Sun *et al*., 2011; McClure *et al*., 2012; Wang *et al*., 2012; Kumar *et al*., 2013). En el aspecto estadístico, la originalidad de este algoritmo radica en la ganancia en potencia de prueba al estandarizar dichos efectos de SNP empleando la varianza propia del efecto, $\mathrm{Var}\left(\hat{g}_j\right)$, y en la reducción del número de falsos positivos y menor tiempo de cálculo. A su vez, se obtienen resultados similares a los obtenidos mediante el conocido procedimiento EMMA (Kang *et al*., 2008), el cual estima los efectos de SNP fijando un marcador a la vez, pero su tiempo de computa resulta entre 2 a 3 mayor comparado por nuestro algoritmo. También, se evaluó la fracción de la varianza genética aditiva asociada a estos segmentos genómicos seleccionados (Hayes et al., 2010) de longitud de 2 Mb formados alrededor del SNP con el mayor –Log(*p*-valor) en cada cromosoma. Es dable señalar que el tamaño de la "ventana" (número de

Mb a derecha e izquierda del SNP pico) fue seleccionado teniendo en cuenta la caída del LD promedio en la población cruza analizada, tal como lo describe Hayes et al., (2010), de modo que los distintos segmentos constituyan aportes independientes a la varianza aditiva total como requiere la teoría dentro de la genética cuantitativa. La prueba de significancia de los segmentos utilizó la corrección de Bonferroni, ampliamente utilizado en estudios genómicos de epidemiología humana (Klein *et al*., 2006). Este enfoque de detección de posiciones con mayor efecto y posterior análisis de significancia mediante un segmento candidato es altamente atractivo y apropiado para el meta-análisis de asociación empleando varias poblaciones independientes.

La funcionalidad tanto de los genotipos imputados como del algoritmo, se ponen a prueba en un GWAS para una población cruza Duroc × Pietrain, junto con caracteres de crecimiento y de deposición de grasa. Como resultado, la región entre 133.8 y 136.2 Mb del cromosoma 6 se asoció con la expresión de la deposición de grasa dorsal en la decima y última costilla en las semanas 10, 13, 16, 19 y 22 de vida, confirmando resultados de anteriores investigaciones (Edwards *et al.*, 2008 (a); Choi *et al.*, 2010; Okumura *et al.*, 2013; Gualdrón Duarte *et al*., 2014), y a su vez se proponiendo genes candidatos como *SERBP1* para la expresión de caracteres de deposición de grasa.

Finalmente, la metodología empleada en el desarrollo de esta tesis es una herramienta práctica aplicada a estudios de mejoramiento animal, que involucra una optimización de los recursos para el desarrollo de investigaciones en poblaciones experimentales con fines productivos.

CONCLUSIONES

Conclusiones

Las principales conclusiones y aportes que se desprenden de esta tesis son:

1) la definición de un esquema de genotipado que maximiza la imputación de genotipos en HD en una población cruza usando dos paneles marcadamente distintos (en número de SNP y en precio), de una manera mas eficiente en costo/beneficio, lo cual juega un papel muy importante dentro de proyectos de investigación.

2) El desarrollo de un algoritmo para estimar los efectos de los marcadores de manera eficiente. Desde el aspecto estadístico la originalidad radica en la ganancia en potencia de prueba al estandarizar dichos efectos de SNP empleando la varianza propia del efecto, y en la reducción del número de falsos positivos y menor tiempo de cálculo.

3) la aplicación de dicho algoritmo en el análisis de datos experimentales para indagar sobre la asociación entre regiones genómicas significativas y la variación genético aditiva de caracteres económicamente relevantes en la industria pecuaria.

BIBLIOGRAFÍA

Abecasis, G. R., S. S. Cherny, W. O. Cookson, and L. R. Cardon. 2002. Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nat. Genet. 30:97–101.

Alessi, M.-C., and I. Juhan-Vague. 2006. PAI-1 and the metabolic syndrome: links, causes, and consequences. Arterioscler. Thromb. Vasc. Biol. 26:2200–7.

Anderson, C. A., F. H. Pettersson, J. C. Barrett, J. J. Zhuang, J. Ragoussis, L. R. Cardon, and A. P. Morris. 2008. Evaluating the Effects of Imputation on the Power , Coverage , and Cost Efficiency of Genome-wide SNP Platforms. Am. J. Hum. Genet. 83:112–119.

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, J. Fix, C. P. Van Tassell, and J. P. Steibel. 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14:8.

Badke, Y. M., R. O. Bates, C. W. Ernst, C. Schwab, and J. P. Steibel. 2012. Estimation of linkage disequilibrium in four US pig breeds. BMC Genomics 13:24.

Bouxsein, M. L., T. Uchiyama, C. J. Rosen, K. L. Shultz, L. R. Donahue, C. H. Turner, S. Sen, G. a Churchill, R. Müller, and W. G. Beamer. 2004. Mapping quantitative trait loci for vertebral trabecular bone volume fraction and microarchitecture in mice. J. Bone Miner. Res. 19:587–99.

Burdick, J. T., W.-M. Chen, G. R. Abecasis, and V. G. Cheung. 2006. In silico method for inferring genotypes in pedigrees. Nat. Genet. 38:1002–4.

Campbell, a. ., W. . Bain, a. . McRae, T. . Broad, P. . Johnstone, K. . Dodds, B. . Veenvliet, G. . Greer, B. . Glass, a. . Beattie, N. . Jopson, and J. . McEwan. 2003. Bone density in sheep: genetic variation and quantitative trait loci localisation. Bone 33:540–548.

Carlson, C. S., M. a Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. a Nickerson. 2004. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am. J. Hum. Genet. 74:106–20.

Cheema, J., and J. Dicks. 2009. Computational approaches and software tools for genetic linkage map estimation in plants. Brief. Bioinform. 10:595–608.

Choi, I., J. P. Steibel, R. O. Bates, N. E. Raney, J. M. Rumph, and C. W. Ernst. 2010. Application of alternative models to identify QTL for growth traits in an F2 Duroc x Pietrain pig resource population. BMC Genet. 11:97.

Christensen, R. 2011. Plane Answer to Complex Questions. The Theory of Linear Models. Fourth edi. Springer New York, New York.

Clifford, D., and P. McCullagh. 2006. The regress function. R News 6:6–10.

Crossa, J., P. Pérez, G. de los Campos, G. Mahuku, S. Dreisigacker, and C. Magorokosho. 2011. Genomic Selection and Prediction in Plant Breeding. J. Crop Improv. 25:239–261.

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. a Woolliams, and M. E. Goddard. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189:317–27.

Dikmen, S., J. B. Cole, D. J. Null, and P. J. Hansen. 2013. Genome-Wide Association Mapping for Identification of Quantitative Trait Loci for Rectal Temperature during Heat Stress in Holstein Cattle. PLoS One 8:e69202.

Do, D. N., T. Ostersen, A. B. Strathe, T. Mark, J. Jensen, and H. N. Kadarmideen. 2014. Genome-wide association and systems genetic analyses of residual feed intake, daily feed consumption, backfat and weight gain in pigs. BMC Genet. 15:27.

Druet, T., and F. P. Farnir. 2011. Modeling of identity-by-descent processes along a chromosome between haplotypes and their genotyped ancestors. Genetics 188:409–19.

Druet, T., and M. Georges. 2010. A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics 184:789–798.

Edwards, D. B., C. W. Ernst, R. J. Tempelman, G. J. M. Rosa, N. E. Raney, M. D. Hoge, and R. O. Bates. 2008. Quantitative trait loci mapping in an F2 Duroc x Pietrain resource population: I. Growth traits. (a) J. Anim. Sci. 86:241–253.

Edwards, D B, C. W. Ernst, N. E. Raney, M. E. Doumit, M. D. Hoge, and R. O. Bates. 2008. Quantitative trait locus mapping in an F2 Duroc x Pietrain resource population: II. Carcass and meat quality traits. (b) J. Anim. Sci. 86:254–266.

Endelman, J. B. 2011. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. Plant Genome J. 4:250.

Falconer, D., and T. Mackay. 1996. Introduction to Quantitative Genetics. Longman, New York.

Fan, B., S. K. Onteru, Z.-Q. Du, D. J. Garrick, K. J. Stalder, and M. F. Rothschild. 2011. Genome-wide association study identifies Loci for body composition and structural soundness traits in pigs. PLoS One 6:e14726.

Fan, Y., Y. Xing, Z. Zhang, H. Ai, Z. Ouyang, J. Ouyang, M. Yang, P. Li, Y. Chen, J. Gao, L. Li, L. Huang, and J. Ren. 2013. A further look at porcine chromosome 7 reveals VRTN variants associated with vertebral number in Chinese and Western pigs. PLoS One 8:e62534.

Fontanesi, L., G. Schiavo, G. Galimberti, D. G. Calò, E. Scotti, P. L. Martelli, L. Buttazzoni, R. Casadio, and V. Russo. 2012. A genome wide association study for

backfat thickness in Italian Large White pigs highlights new regions affecting fat deposition including neuronal genes. BMC Genomics 13:583.

Garrick, D. J. 2007. Equivalent mixed model equations for genomic selection. J. Bone Miner. Res. 90(Suppl.):376(Abstr.).

Goddard, M. E., and B. J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat. Rev. Genet. 10:381–91.

Gualdrón Duarte, J. L., R. O. Bates, C. W. Ernst, N. E. Raney, R. J. C. Cantet, and J. P. Steibel. 2013. Genotype imputation accuracy in a F2 pig population using high density and low density SNP panels. BMC Genet. 14:38.

Gualdrón Duarte, J. L., R. J. Cantet, R. O. Bates, C. W. Ernst, N. E. Raney, and J. P. Steibel. 2014. Rapid screening for phenotype-genotype associations by linear transformations of genomic evaluations. BMC Bioinformatics 15:246.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343–353.

Haldane, J. B. S. 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. 8:299–309.

Haley, S., and J. Elsen. 1994. Mapping Quantitative Trait Loci in Crosses Between Outbred Lines Using Least Squares. Genetics 136:1195–1207.

Hao, K., E. Chudin, J. McElwee, and E. E. Schadt. 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genet. 10:27.

Hayes, B. J., P. J. Bowman, a J. Chamberlain, and M. E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92:433–43.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. van der Werf. 2011. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72–80.

Hayes, B. J., J. Pryce, A. J. Chamberlain, P. J. Bowman, and M. E. Goddard. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6:e1001139.

Hayes, B. 2007. QTL Mapping , MAS , and Genomic Selection. In: A short course. Animal Breeding and Genetics - Departement of Animal Science. Iowa State University, Iowa. p. 116.

Heaton, J. H., W. M. Dlakic, M. Dlakic, and T. D. Gelehrter. 2001. Identification and cDNA cloning of a novel RNA-binding protein that interacts with the cyclic nucleotide-responsive sequence in the Type-1 plasminogen activator inhibitor mRNA. J. Biol. Chem. 276:3341–7.

Heberlein, K. R., J. Han, A. C. Straub, A. K. Best, C. Kaun, J. Wojta, and B. E. Isakson. 2012. A novel mRNA binding protein complex promotes localized plasminogen activator inhibitor-1 accumulation at the myoendothelial junction. Arterioscler. Thromb. Vasc. Biol. 32:1271–9.

Henderson, C. 1984. Applications of linear models in animal breeding. (U. Guelph., editor.). Guelph, Canada.

Hickey, John M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. Crop Sci. 52:654–663.

Hickey, John M, B. P. Kinghorn, B. Tier, J. H. van der Werf, and M. a Cleveland. 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44:9.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. F. Wilson, N. Dunstan, and J. H. J. van der Werf. 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 43:12.

Hill, W. G., and A. Robertson. 1966. Linkage Disequilibrium in Finite Populations. Theor. Appl. Genet. 38:226–231.

Huang, L., Y. Li, A. B. Singleton, J. a Hardy, G. Abecasis, N. a Rosenberg, and P. Scheet. 2009. Genotype-imputation accuracy across worldwide human populations. Am. J. Hum. Genet. 84:235–50.

Huang, Yijian, J. M. Hickey, M. a Cleveland, and C. Maltecca. 2012. Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet. Sel. Evol. 44:25.

Huang, Y, C. Maltecca, J. P. Cassady, L. J. Alexander, W. M. Snelling, and M. D. Macneil. 2012. Effects of reduced panel, reference origin, and genetic relationship on imputation of genotypes in Hereford cattle. J. Anim. Sci. 59301:1–17.

Jin, W., E. N. Olson, S. S. Moore, J. a Basarab, U. Basu, and L. L. Guan. 2012. Transcriptome analysis of subcutaneous adipose tissues in beef cattle using 3’ digital gene expression-tag profiling. J. Anim. Sci. 90:171–83.

Kang, H. M., N. a Zaitlen, C. M. Wade, A. Kirby, D. Heckerman, M. J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genet. Soc. Am. 178:1709–23.

Kim, C. W., Y. H. Hong, S.-I. Yun, S.-R. Lee, Y. H. Kim, M.-S. Kim, K. H. Chung, W. Y. Jung, E. J. Kwon, S. S. Hwang, D. H. Park, K. K. Cho, J. G. Lee, B. W. Kim, J. W. Kim, Y. S. Kang, J. S. Yeo, and K.-T. Chang. 2006. Use of Microsatellite Markers to Detect Quantitative Trait Loci in Yorkshire Pigs. J. Reprod. Dev. 52:229–237.

Klein, R. J., C. Zeiss, E. Y. Chew, J. Tsai, R. S. Sackler, C. Haynes, A. K. Henning, J. P. Sangiovanni, S. M. Mane, T. Susan, M. B. Bracken, F. L. Ferris, J. Ott, C.

Barnstable, and J. Hoh. 2006. Complement Factor H Polymorphism in Age-Related Macular Degeneration. Science (80-.). 308:385–389.

Koning, D. J. De, L. L. G. Janss, A. P. Rattink, P. A. M. Van Oers, B. J. De Vries, M. A. M. Groenen, J. J. Van Der Poel, P. N. De Groot, E. W. P. Brascamp, and J. A. M. Van Arendonk. 1999. Detection of Quantitative Trait Loci for Backfat Thickness and Intramuscular Fat Content in Pigs (Sus scrofa). 1690:1679–1690.

Kover, P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott. 2009. A Multiparent Advanced Generation Inter-Cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genet. 5:e1000551.

Kumar, S., D. J. Garrick, M. C. Bink, C. Whitworth, D. Chagné, and R. K. Volz. 2013. Novel genomic approaches unravel genetic architecture of complex traits in apple. BMC Genomics 14:393.

Lande, R., and R. Thompsont. 1990. Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics 124:743–756.

Laval, G., and L. Excoffier. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20:2485–2487.

Leduc, M. S., R. S. Hageman, R. a Verdugo, S.-W. Tsaih, K. Walsh, G. a Churchill, and B. Paigen. 2011. Integration of QTL and bioinformatic tools to identify candidate genes for triglycerides in mice. J. Lipid Res. 52:1672–82.

Ledur, M. C., N. Navarro, and M. Pérez-Enciso. 2010. Large-scale SNP genotyping in crosses between outbred lines: how useful is it? Heredity (Edinb). 105:173–182.

Lee, K.-T., M.-J. Byun, K.-S. Kang, E.-W. Park, S.-H. Lee, S. Cho, Hyoyoung Kim, K.-W. Kim, T. Lee, J.-E. Park, W. Park, D. Shin, Hong-Seog Park, J.-T. Jeon, B.-H. Choi, G.-W. Jang, S.-H. Choi, D.-W. Kim, D. Lim, Hae-Suk Park, M.-R. Park, J. Ott, L. B. Schook, T.-H. Kim, and Heebal Kim. 2011. Neuronal genes for subcutaneous fat thickness in human and pig are identified by local genomic sequencing and combined SNP association study. PLoS One 6:e16356.

Lee, S. S., Y. Chen, C. Moran, S. Cepica, G. Reiner, H. Bartenschlager, G. Moser, and H. Geldermann. 2003. Linkage and QTL mapping for Sus scrofa chromosome 2. J. Anim. Breed. Genet. 120:11–19.

Legarra, a, I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92:4656–63.

Liang, K.-Y., and S. G. Self. 1996. On the Asymptotic Behaviour of the Pseudolikelihood Ratio Test Statistic. J. R. Stat. Soc. Ser. B 58:785–796.

Lin, P., S. M. Hartz, Z. Zhang, S. F. Saccone, J. Wang, J. a Tischfield, H. J. Edenberg, J. R. Kramer, A. M Goate, L. J. Bierut, and J. P. Rice. 2010. A new statistic to evaluate imputation reliability. PLoS One 5:e9697.

Liu, G., D. G. J. Jennen, E. Tholen, H. Juengst, T. Kleinwächter, M. Hölker, D. Tesfaye, G. Ün, H.-J. Schreinemachers, E. Murani, S. Ponsuksili, J.-J. Kim, K. Schellander, and K. Wimmers. 2007. A genome scan reveals QTL for growth, fatness, leanness and meat quality in a Duroc-Pietrain resource population. Anim. Genet. 38:241–252.

Mackay, T. F. C. 2001. The genetic architecture of quantitative traits. Annu. Rev. Genet. 35:303–339.

Malek, M., J. C. Dekkers, H. K. Lee, T. J. Baas, and M. F. Rothschild. 2001. A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. I. Growth and body composition. Mamm. Genome 12:630–6.

McClure, M. C., H. R. Ramey, M. M. Rolf, S. D. McKay, J. E. Decker, R. H. Chapple, J. W. Kim, T. M. Taxis, R. L. Weaber, R. D. Schnabel, and J. F. Taylor. 2012. Genome-wide association analysis for quantitative trait loci influencing Warner-Bratzler shear force in five taurine cattle breeds. Anim. Genet. 43:662–73.

Mei, Y., Y. Chen, J. Li, P. Gao, C. Wang, H. Zhang, F. Ling, Y. Li, S. Xie, S. Li, and G. Zhang. 2008. Sequence identification, tissue distribution and polymorphism of the porcine cathepsin D (CTSD) gene. Anim. Biotechnol. 19:144–58.

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–29.

Meyers, S. N., S. L. Rodriguez-Zas, and J. E. Beever. 2007. Fine-mapping of a QTL influencing pork tenderness on porcine chromosome 2. BMC Genet. 8:69.

Muñoz, G., Ovilo, C., Silió, Tomás, A., Noguera, J., Rodríguez M. 2009. Single- and joint-population analyses of two experimental pig crosses to confirm quantitative trait loci on Sus scrofa chromosome 6 and leptin receptor effects on fatness and growth traits. J. Anim. Sci. 87:459–468.

Nonneman, D., A. K. Lindholm-Perry, S. D. Shackelford, D. A. King, T. L. Wheeler, G. A. Rohrer, C. D. Bierman, J. F. Schneider, R. K. Miller, H. Zerby, and S. J. Moeller. 2011. Predictive markers in calpastatin for tenderness in commercial pig populations. J. Anim. Sci. 89:2663–2672.

Okumura, N., T. Matsumoto, T. Hayashi, K. Hirose, K. Fukawa, T. Itou, H. Uenishi, S. Mikawa, and T. Awata. 2013. Genomic regions affecting backfat thickness and cannon bone circumference identified by genome-wide association study in a Duroc pig population. Anim. Genet. 44:454–7.

Óvilo, C., a. Fernández, J. L. Noguera, C. Barragán, R. Letón, C. Rodríguez, a. Mercadé, E. Alves, J. M. Folch, L. Varona, and M. Toro. 2005. Fine mapping of porcine chromosome 6 QTL and LEPR effects on body composition in multiple generations of an Iberian by Landrace intercross. Genet. Res. 85:57–67.

Poland, J. a, P. J. Bradbury, E. S. Buckler, and R. J. Nelson. 2011. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc. Natl. Acad. Sci. U. S. A. 108:6893–8.

Prakash, S., W. E. Paul, and P. W. Robbins. 2007. Fibrosin, a novel fibrogenic cytokine, modulates expression of myofibroblasts. Exp. Mol. Pathol. 82:42–8.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. a R. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. W. de Bakker, M. J. Daly, and P. C. Sham. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Qin, Z. S., S. Gopalakrishnan, and G. R. Abecasis. 2006. An efficient comprehensive search algorithm for tagSNP selection using linkage disequilibrium criteria. Bioinformatics 22:220–225.

Ramos, A. M., R. P. M. a Crooijmans, N. a Affara, A. J. Amaral, A. L. Archibald, J. E. Beever, C. Bendixen, C. Churcher, R. Clark, P. Dehais, M. S. Hansen, J. Hedegaard, Z.-L. Hu, H. H. Kerstens, A. S. Law, H.-J. Megens, D. Milan, D. J. Nonneman, G. a Rohrer, M. F. Rothschild, T. P. L. Smith, R. D. Schnabel, C. P. Van Tassell, J. F. Taylor, R. T. Wiedmann, L. B. Schook, and M. a M. Groenen. 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PLoS One 4:e6524.

Rangkasenee, N., E. Murani, R. M. Brunner, K. Schellander, M. U. Cinar, H. Luther, A. Hofer, M. Stoll, A. Witten, S. Ponsuksili, and K. Wimmers. 2013. Genome-Wide Association Identifies TBX5 as Candidate Gene for Osteochondrosis Providing a Functional Link to Cartilage Perfusion as Initial Factor. Front. Genet. 4:78.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander. 2001. Linkage disequilibrium in the human genome. Nature 411:199–204.

Resnyk, C. W., W. Carré, X. Wang, T. E. Porter, J. Simon, E. Le Bihan-Duval, M. J. Duclos, S. E. Aggrey, and L. a Cogburn. 2013. Transcriptional analysis of abdominal fat in genetically fat and lean chickens reveals adipokines, lipogenic genes and a link between hemostasis and leanness. BMC Genomics 14:557.

Sahana, G., B. Guldbrandtsen, and M. S. Lund. 2011. Genome-wide association study for calving traits in Danish and Swedish Holstein cattle. J. Dairy Sci. 94:479–86.

Sanchez, M.-P., N. Iannuccelli, B. Basso, J.-P. Bidanel, Y. Billon, G. Gandemer, H. Gilbert, C. Larzul, C. Legault, J. Riquet, D. Milan, and P. Le Roy. 2007. Identification of QTL with effects on intramuscular fat content and fatty acid composition in a Duroc x Large White cross. BMC Genet. 8:55.

Sato, S, Y. Oyamada, K. Atsuji, T. Nade, Shin-ichi Sato, E. Kobayashi, T. Mitsuhashi, A. Nirasawa, Y. Komatsuda, S. Saito, T. Terai, T. Hayashi, and Y. Sugimoto. 2003. Quantitative trait loci analysis for growth and carcass traits in a Meishan × Duroc F2 resource population. J. Anim. Sci. 81:2938–2949.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123:218–223.

Self, S. G., and K.-Y. Liang. 1987. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. J. Am. Stat. Assoc. 82:605–610.

Stearns, T. M., J. E. Beever, B. R. Southey, M. Ellis, F. K. Mckeith, and S. L. Rodriguez-Zas. 2005. Evaluation of approaches to detect quantitative trait loci for growth , carcass , and meat analyses The online version of this article , along with updated information and services , is located on the World Wide Web at : Evaluation of approaches to detect. J. Anim. Sci. 83:1481–1493.

Steer, C. D., A. Sayers, J. Kemp, W. D. Fraser, and J. H. Tobias. 2014. Birth weight is positively related to bone size in adolescents but inversely related to cortical bone mineral density: findings from a large prospective cohort study. Bone 65:77–82.
Storey, J. D., and R. Tibshirani. 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. 2003.

Strandén, I., and D. J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92:2971–5.

Sun, X., R. L. Fernando, D. J. Garrick, and J. C. M. Dekkers. 2011. An iterative approach for efficient calculation of breed- ing values and genome-wide association analysis using weighted genomic BLUP. J. Anim. Sci. 89:(E–Suppl 2)e11.

Switonski, M., M. Stachowiak, J. Cieslak, M. Bartz, and M. Grzes. 2010. Genetics of fat tissue accumulation in pigs: a comparative approach. J. Appl. Genet. 51:153–68.

Tian, F., P. J. Bradbury, P. J. Brown, H. Hung, Q. Sun, S. Flint-Garcia, T. R. Rocheford, M. D. McMullen, J. B. Holland, and E. S. Buckler. 2011. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat. Genet. 43:159–62.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci. 92:16–24.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–23.

Vaughan, L. K., J. Divers, M. Padilla, D. T. Redden, K. Hemant, D. Pomp, and D. B. Allison. 2009. The use of plasmodes as a supplement to simulations: A simple example evaluating individual admixture estimation methodologies. Comput Stat Data Anal. 53:1755–1766.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb). 94:73–83.

Weigel, K. A., C. P. Van Tassell, J. R. O'Connell, P. M. VanRaden, and G. R. Wiggans. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93:2229–2238.

Wiggans, G. R., T. a Cooper, P. M. Vanraden, K. M. Olson, and M. E. Tooker. 2012. Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. J. Dairy Sci. 95:1552–1558.

Xu, Z., N. L. Kaplan, and J. a Taylor. 2007. TAGster: efficient selection of LD tag SNPs in single or multiple populations. Bioinformatics 23:3254–3255.

Yang, G., J. Ren, S. Li, H. Mao, Y. Guo, Z. Zou, D. Ren, J. Ma, and L. Huang. 2008. Genome-wide identification of QTL for age at puberty in gilts using a large intercross F2 population between White Duroc x Erhualian. Genetics 40:529–539.

Yu, J., G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38:203–8.

Zhang, F., Z. Zhang, X. Yan, H. Chen, W. Zhang, Y. Hong, and L. Huang. 2014. Genome-wide association studies for hematological traits in Chinese Sutai pigs. BMC Genet. 15:41.

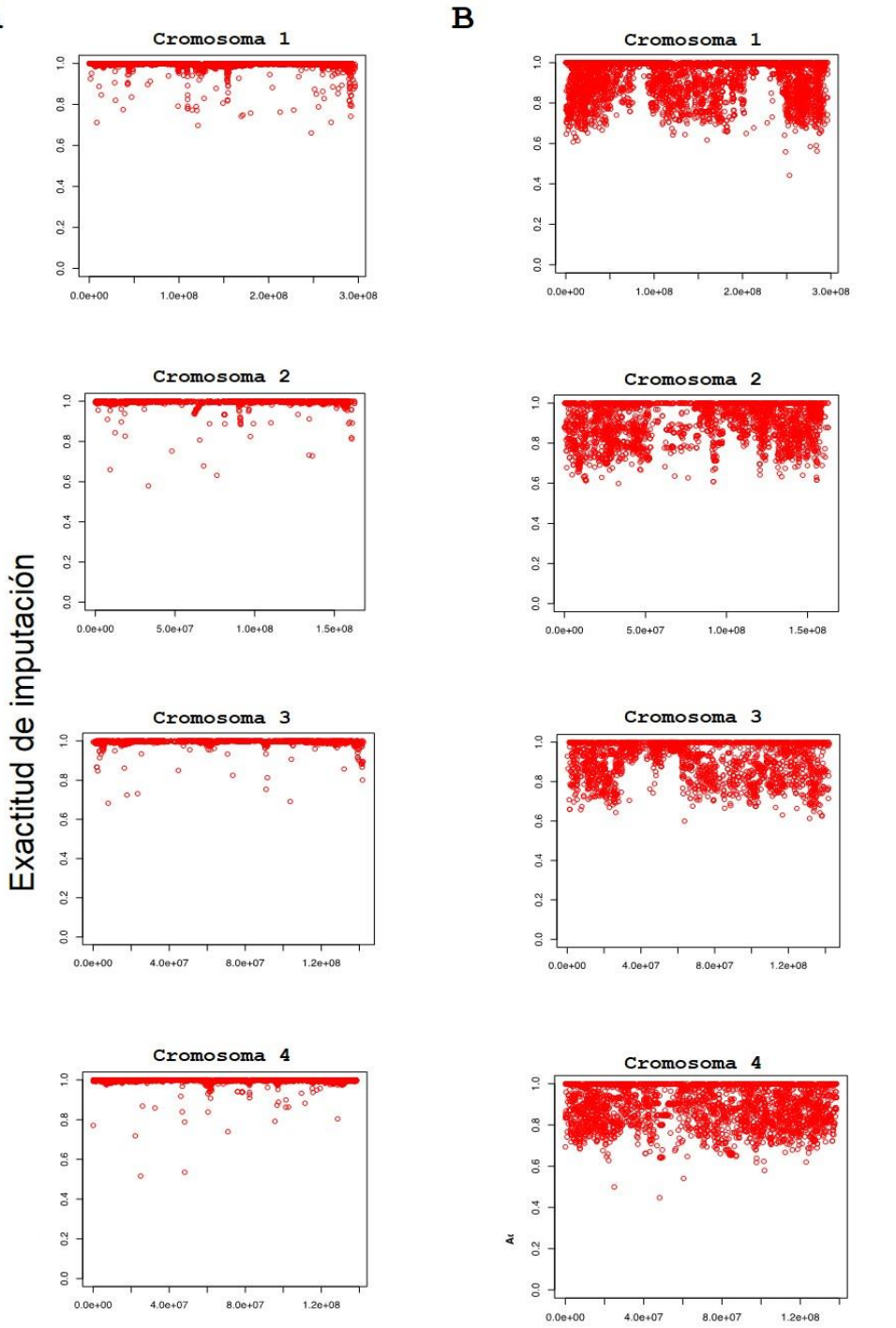
Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93:5487–5494.

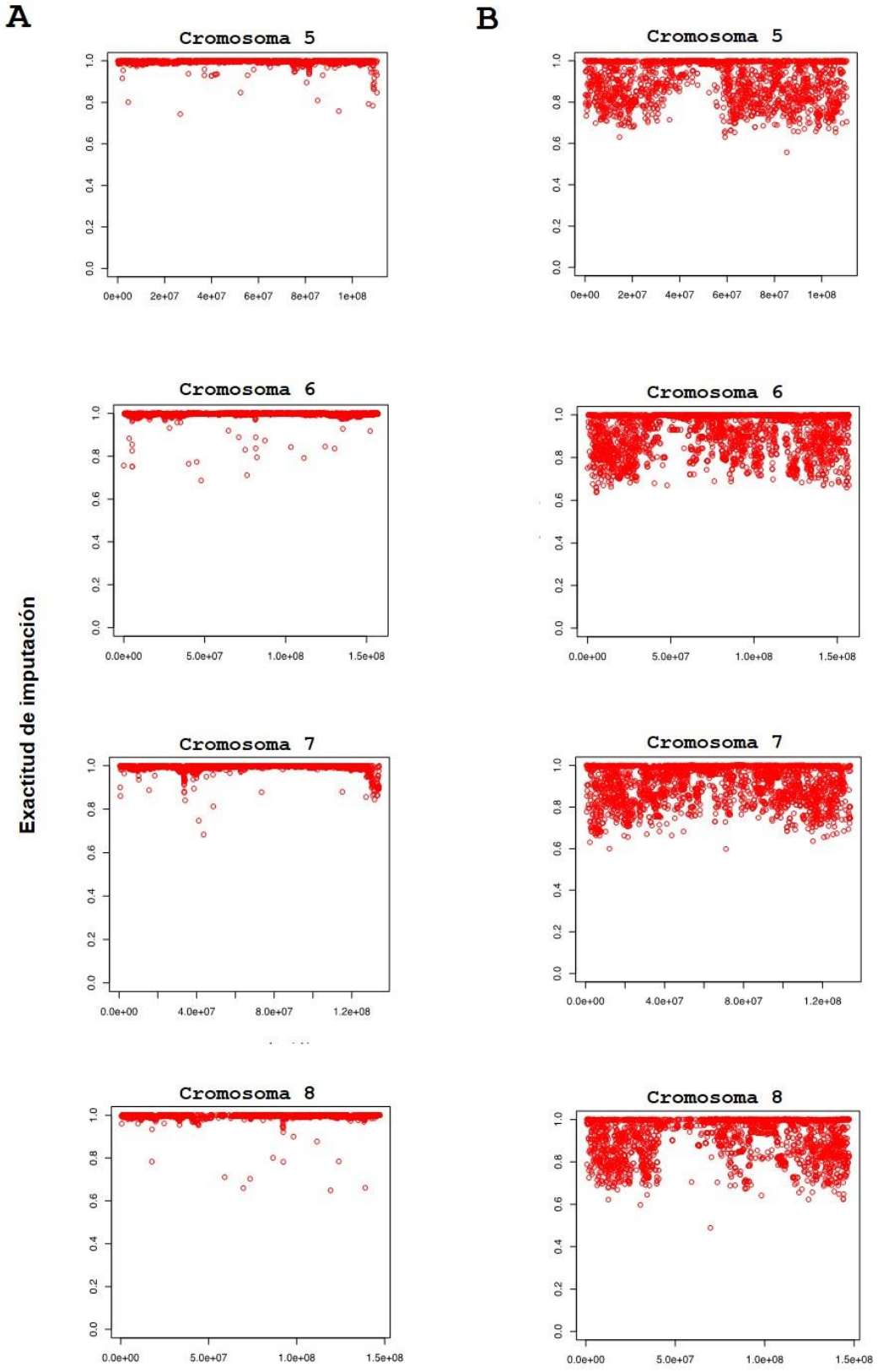
Zheng, J., Y. Li, G. R. Abecasis, and P. Scheet. 2011. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. Genet. Epidemiol. 35:102–110.

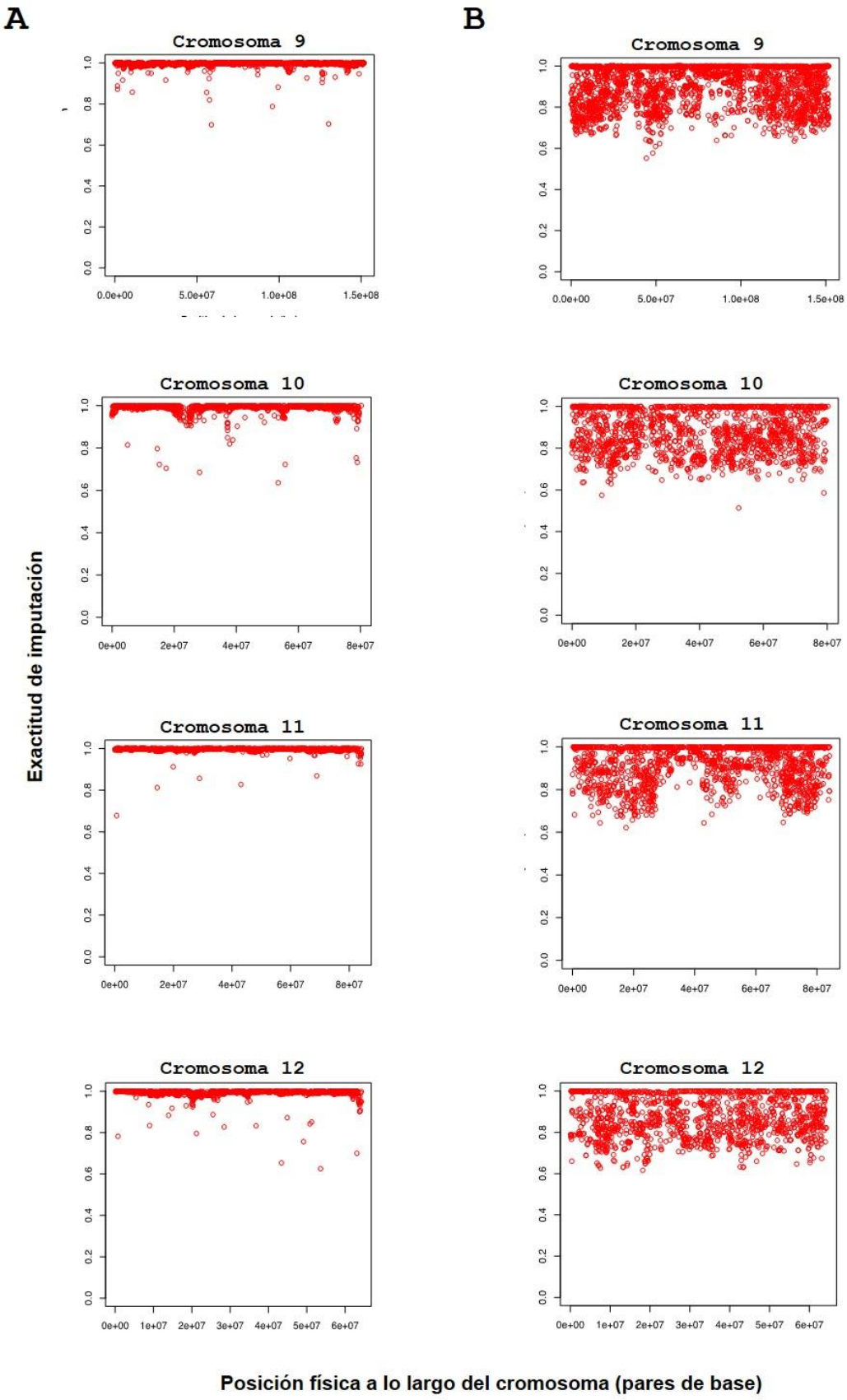
Zhou, X., and M. Stephens. 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44:821–4.

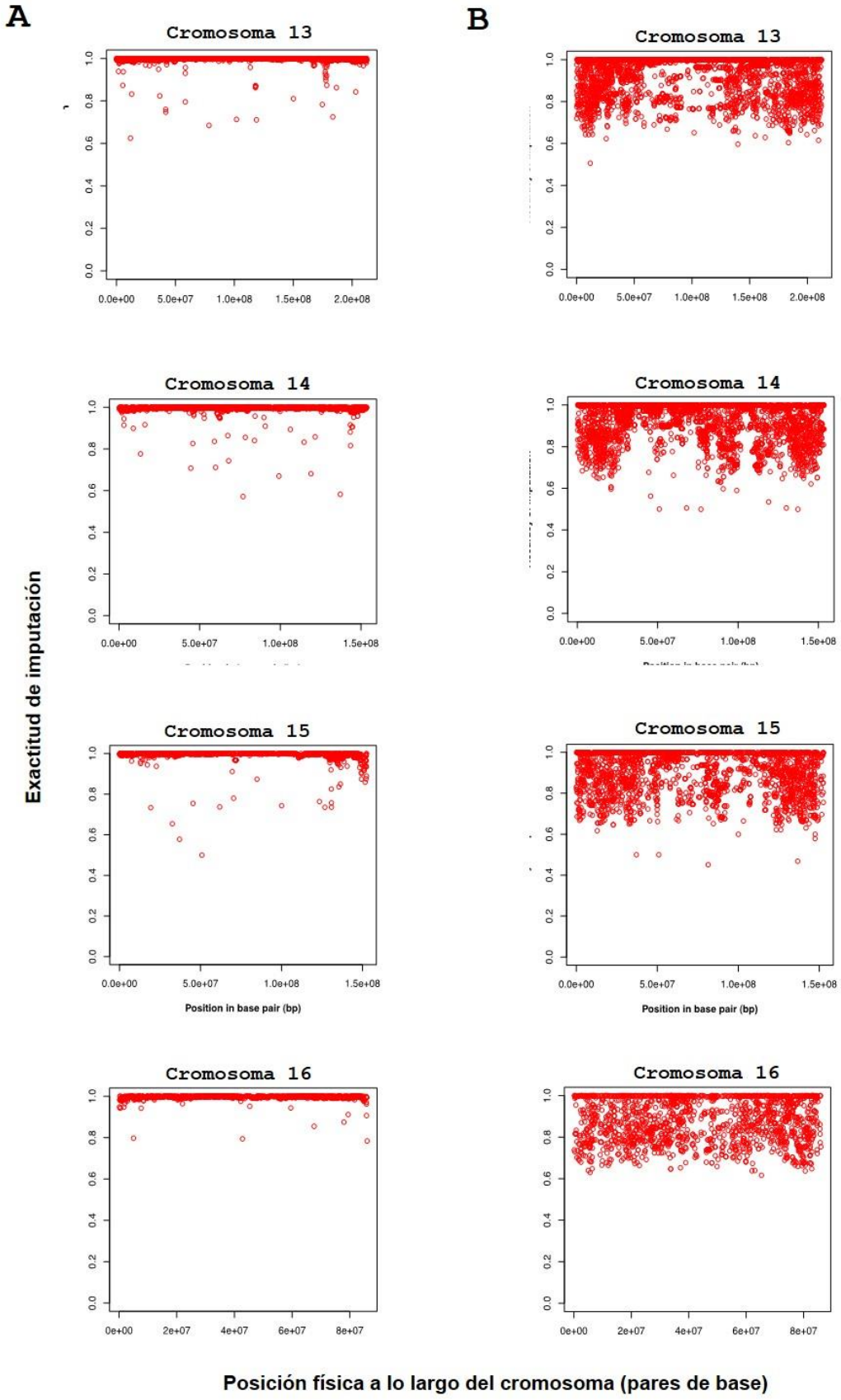
APÉNDICES

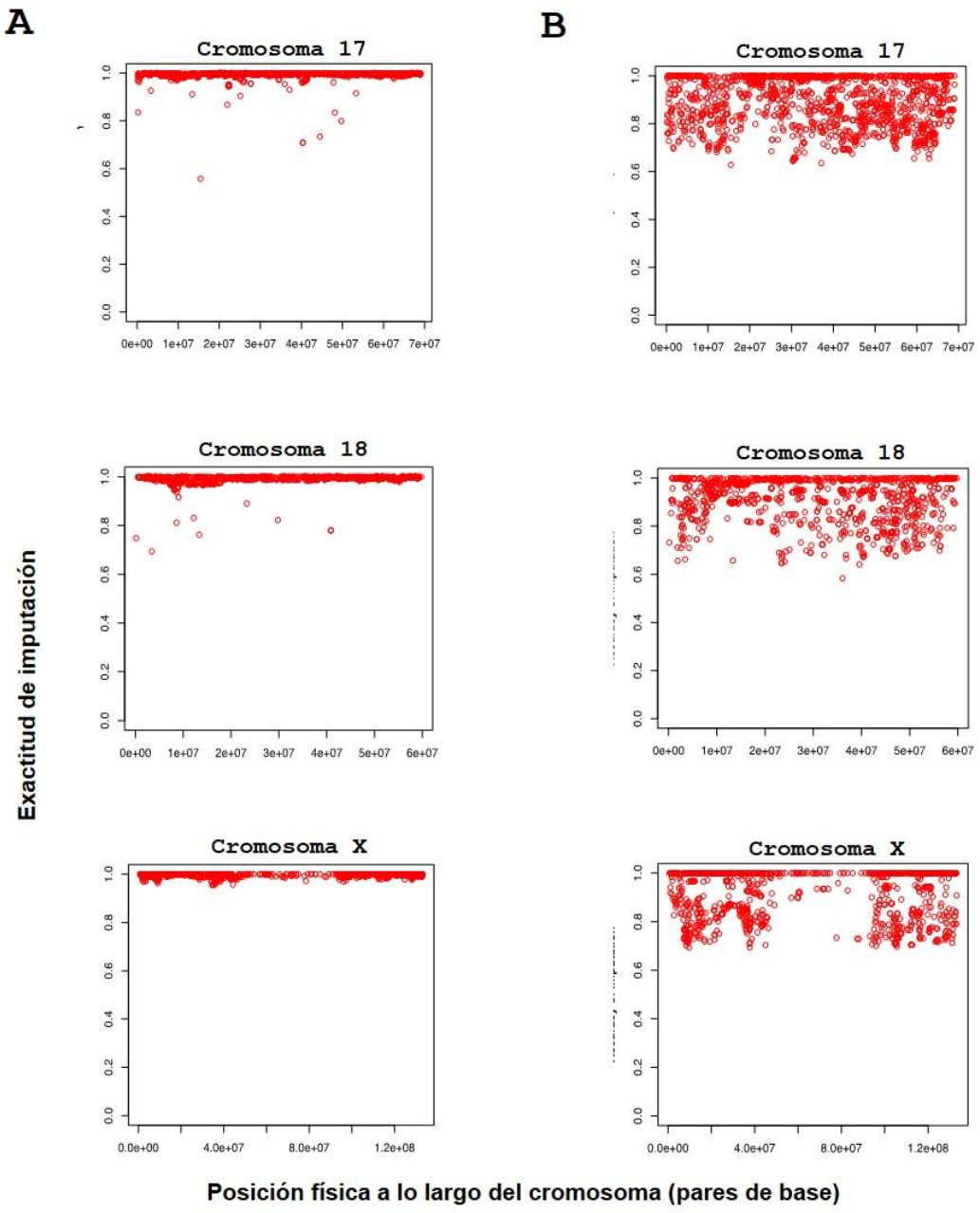
Apéndice 1. Exactitud de imputación en cromosomas 1-18 y X bajo dos escenarios de genotipado




A
B
Cromosoma 5
Cromosoma 6
Cromosoma 7
Cromosoma 8
Exactitud de imputación
Posición física a lo largo del cromosoma (pares de base)

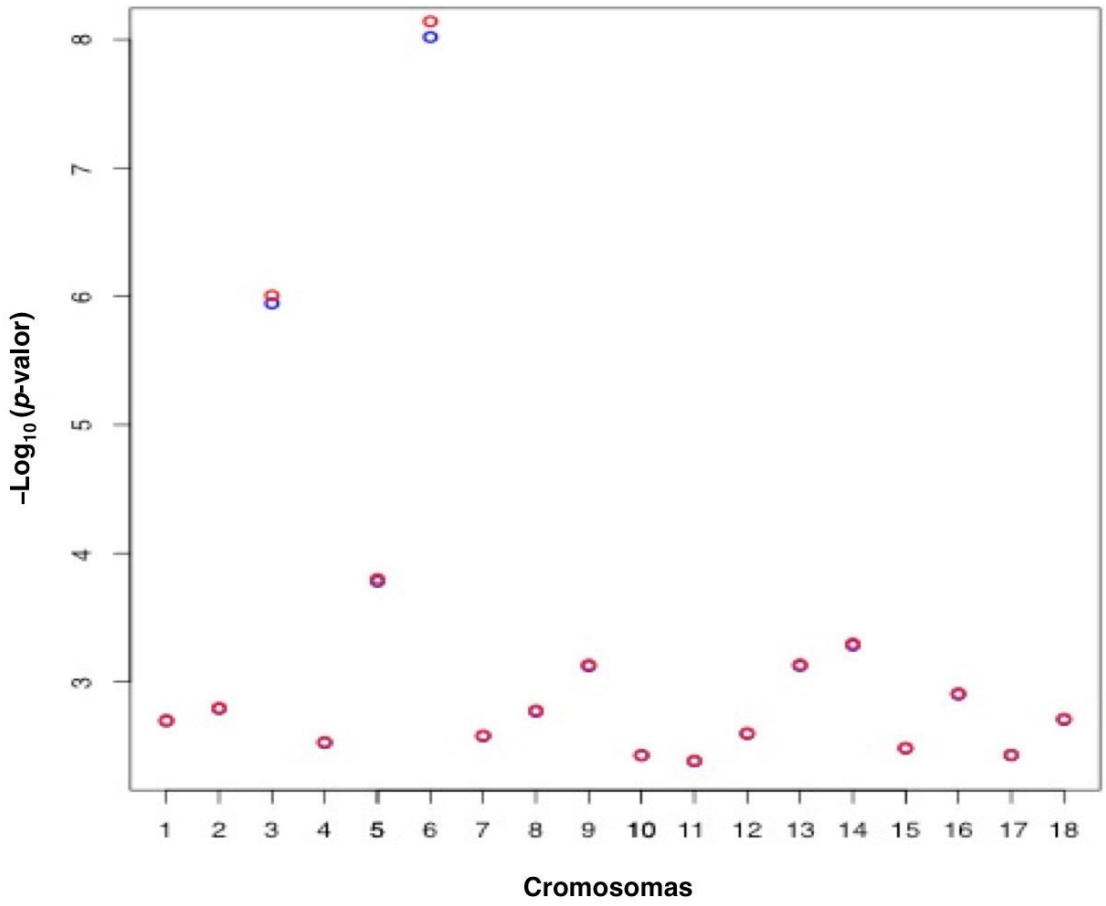

A
B
Cromosoma 9
Cromosoma 10
Cromosoma 11
Cromosoma 12
Exactitud de imputación
Posición física a lo largo del cromosoma (pares de base)


A
B
Cromosoma 13
Cromosoma 14
Cromosoma 15
Cromosoma 16
Exactitud de imputación
Position in base pair (bp)
Posición física a lo largo del cromosoma (pares de base)



Exactitud de imputación como función de la posición en el cromosoma bajo dos escenarios: **a)** Generación F_0 y F_1 alta densidad, F_2 baja densidad (columna izquierda **A**), **b)** Generación F_0 en alta densidad, F_1 y F_2 en baja densidad (columna derecha **B**). Círculos rojos representan cada SNP a lo largo del cromosoma.

Apéndice 2. Valor mas alto de − Log_{10}(p-valores) en cada cromosoma para el carácter grasa dorsal en la decima costilla (mm) en la semana 13 mediante SNP_{ej} y EMMA.



Valor de − Log_{10}(p-valor) mas alto en cada cromosoma para el carácter *grasa dorsal en la decima costilla (mm) en la semana 13* estandarizado por SNP_{ej} (círculos azules) y por el modelo mixto eficiente de asociación (EMMA) usando rrBLUP (círculos rojos).

Apéndice 3. Gráfico de dispersión de − Log_{10}(p-valores) para el carácter grasa dorsal en la decima costilla (mm) en la semana 13 EMMA y SNP_{ej}

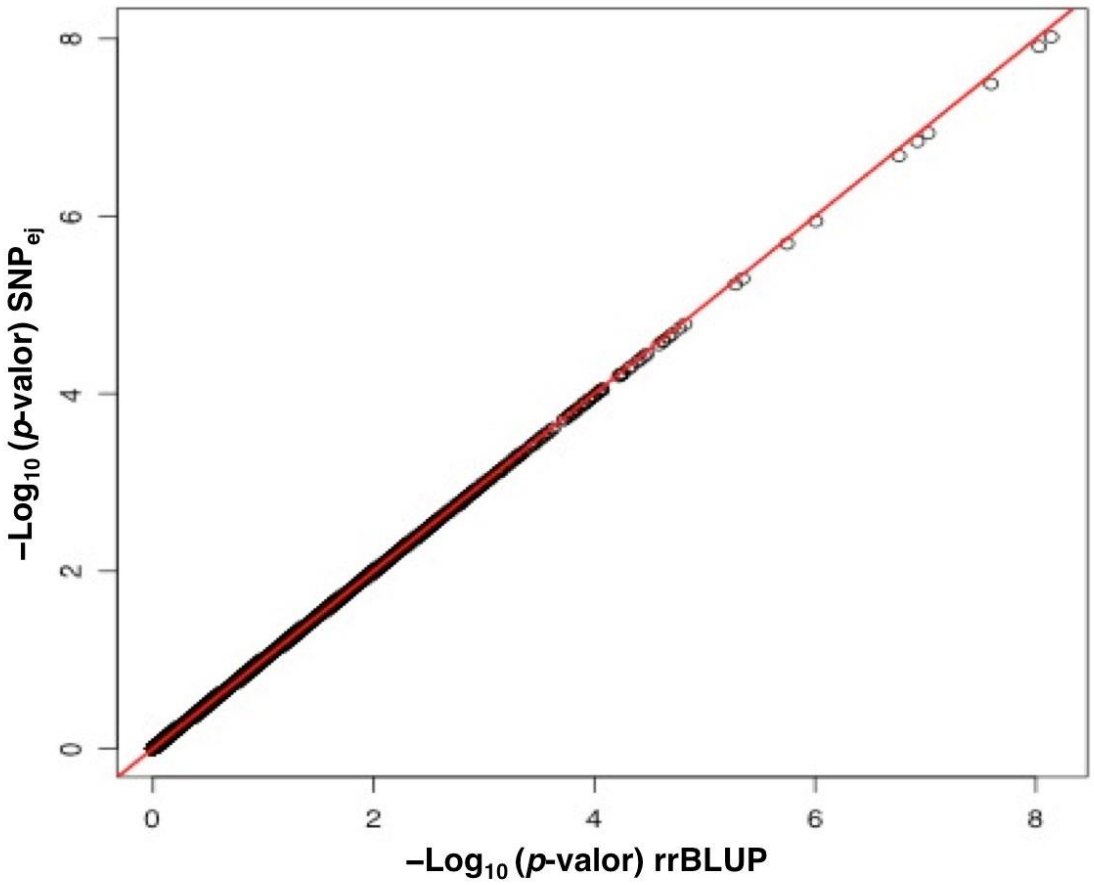


Gráfico de dispersión para 44055 –Log_{10} (*p*-valores) mediante el modelo mixto eficiente de asociación (EMMA) utilizando el programa rrBLUP en el eje de las *x* y estandarizado por $SNP_{ej.}$ en el eje de las *y*. Línea roja recta de 0-1 usada como referencia.

Apéndice 4. Componentes de varianza y verosimilitud para modelos con y sin segmento para cromosomas 1 a 18

Seg-cromosoma	**6**	**3**	**5**	**14**	**13**	**9**	**16**	**2**
-Log$_{10}$(p-valor)	8.02	5.94	3.78	3.28	3.12	3.12	2.90	2.79
Lk_m1	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938
Lk_m2	−1210.800	−1223.178	−1224.540	−1227.746	−1226.184	−1228.144	−1226.223	−1226.625
LRT	34.28	9.52	6.80	0.38	3.51	-0.41	3.43	2.63
p-valor$_{LRT}$	1.1×10^{-9}	6.5×10^{-4}	3.1×10^{-3}	0.3	2.4×10^{-2}	1.0	2.5×10^{-2}	4.3×10^{-2}
VarE_m1	3.70	3.70	3.70	3.70	3.70	3.70	3.70	3.70
VarA_m1	2.68	2.68	2.68	2.68	2.68	2.68	2.68	2.68
VarE_m2	3.73	3.67	3.69	3.70	3.71	3.69	3.65	3.68
VarA_m2	1.95	2.42	2.55	2.64	2.56	2.63	2.59	2.64
segmVA	0.70	0.63	0.15	0.06	0.10	0.27	0.26	0.14
%segmVA	0.11	0.09	0.02	0.01	0.02	0.04	0.04	0.02

Seg-cromosoma = Número del cromosoma donde el segmento esta localizado, **m1** = modelo [3.3] sin segmento: $y = X\beta + a + e$, **m2** = modelo [3.15] con segmento $y = X\beta + a_1 + a_2 + e$, **SNP** −**Log$_{10}$(p-valor)** = −Logaritmo en base 10 de los *p*-valore del SNP seleccionado para crear el segmento, **Lk_m1** = −LogLikehood para m1, **Lk_m2** = −LogLikehood para m2, **LRT** = test de la tasa de verosimilitud para m1 and m2, ***p*-value$_{LRT}$** = *p*-valor para LRT, **VarE_m1** = Varianza del error $\left(\sigma^2_e\right)$ para m1, **VarA_m1** = Varianza aditiva $\left(\sigma^2_A\right)$ para m1,**VarE_m2** = Varianza del error $\left(\sigma^2_e\right)$ para m2, **VarA_m2** = Varianza aditiva $\left(\sigma^2_A\right)$ para m2, **segmVA** = Varianza aditiva para el segmento $\left(\sigma^2_{A_1}\right)$ para m2, %**segmVa** = Proporción en porcentaje (%) del total de la varianza aditiva explicada por el segmento.

Componentes de varianza y verosimilitud para modelos con y sin segmento para el cromosomas 1 a 18.

Seg-cromosoma	**8**	**18**	**1**	**12**	**7**	**4**	**15**	**17**
SNP $-\log_{10}$(p-valor)	2.79	2.70	2.69	2.59	2.57	2.52	2.48	2.43
Lk_m_1	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938	−1227.938
Lk_m_2	−1227.887	−1226.018	−1225.725	−1227.612	−1226.235	−1227.240	−1226.524	−1227.020
LRT	0.10	3.84	4.43	0.65	3.41	1.40	2.83	1.84
p-valor$_{LRT}$	0.48	0.02	0.01	0.22	0.03	0.11	0.04	0.08
VarE_m_1	3.70	3.70	3.70	3.70	3.70	3.70	3.70	3.70
VarA_m_1	2.68	2.68	2.68	2.68	2.68	2.68	2.68	2.68
VarE_m_2	3.69	3.71	3.68	3.71	3.67	3.70	3.69	3.71
VarA_m_2	2.65	2.56	2.58	2.60	2.64	2.71	2.64	2.57
segmVA	0.10	0.07	0.13	0.06	0.11	-0.02	0.08	0.07
%segmVA	0.02	0.01	0.02	0.01	0.02	0.00	0.01	0.01

Seg-cromosoma = Número del cromosoma donde el segmento esta localizado, **m_1** = modelo [3.3] sin segmento: $y = X\beta + a + e$, **m_2** = modelo [3.15] con segmento $\boldsymbol{y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{a}_1 + \boldsymbol{a}_2 + \boldsymbol{e}$, **SNP $-\log_{10}$(p-valor)** = −Logaritmo en base 10 de los *p*-valore del SNP seleccionado para crear el segmento, **Lk_m_1** = −LogLikehood para m_1, **Lk_m_2** = −LogLikehood para m_2, **LRT** = test de la tasa de verosimilitud para m_1 and m_2, ***p*-value$_{LRT}$** = *p*-valor para LRT, **VarE_m_1** = Varianza del error $\left(\sigma^2_e\right)$ para m_1, **VarA_m_1** = Varianza aditiva $\left(\sigma^2_A\right)$ para m_1, **VarE_m_2** = Varianza del error $\left(\sigma^2_e\right)$ para m_2, **VarA_m_2** = Varianza aditiva $\left(\sigma^2_A\right)$ para m_2, **segmVA** = Varianza aditiva para el segmento $\left(\sigma^2_{A_1}\right)$ para m_2, %**segmVA** = Proporción en porcentaje (%) del total de la varianza aditiva explicada por el segmento.

Componentes de varianza y verosimilitud para modelos con y sin segmento para el cromosomas 10 y 11.

Seg-cromosoma	10	11
SNP $-\log_{10}$(p-valor)	2.42	2.38
Lk_m$_2$	−1226.179	−1227.640
Lk_m$_1$	−1227.938	−1227.938
LRT	3.52	0.60
p-valor$_{LRT}$	0.02	0.02
VarE_m$_2$	3.70	3.70
VarA_m$_1$	2.68	2.68
VarE_m$_2$	3.73	3.70
VarA_m$_2$	2.66	2.64
segmVA	−0.03	0.04
%segmVA	0.00	0.01

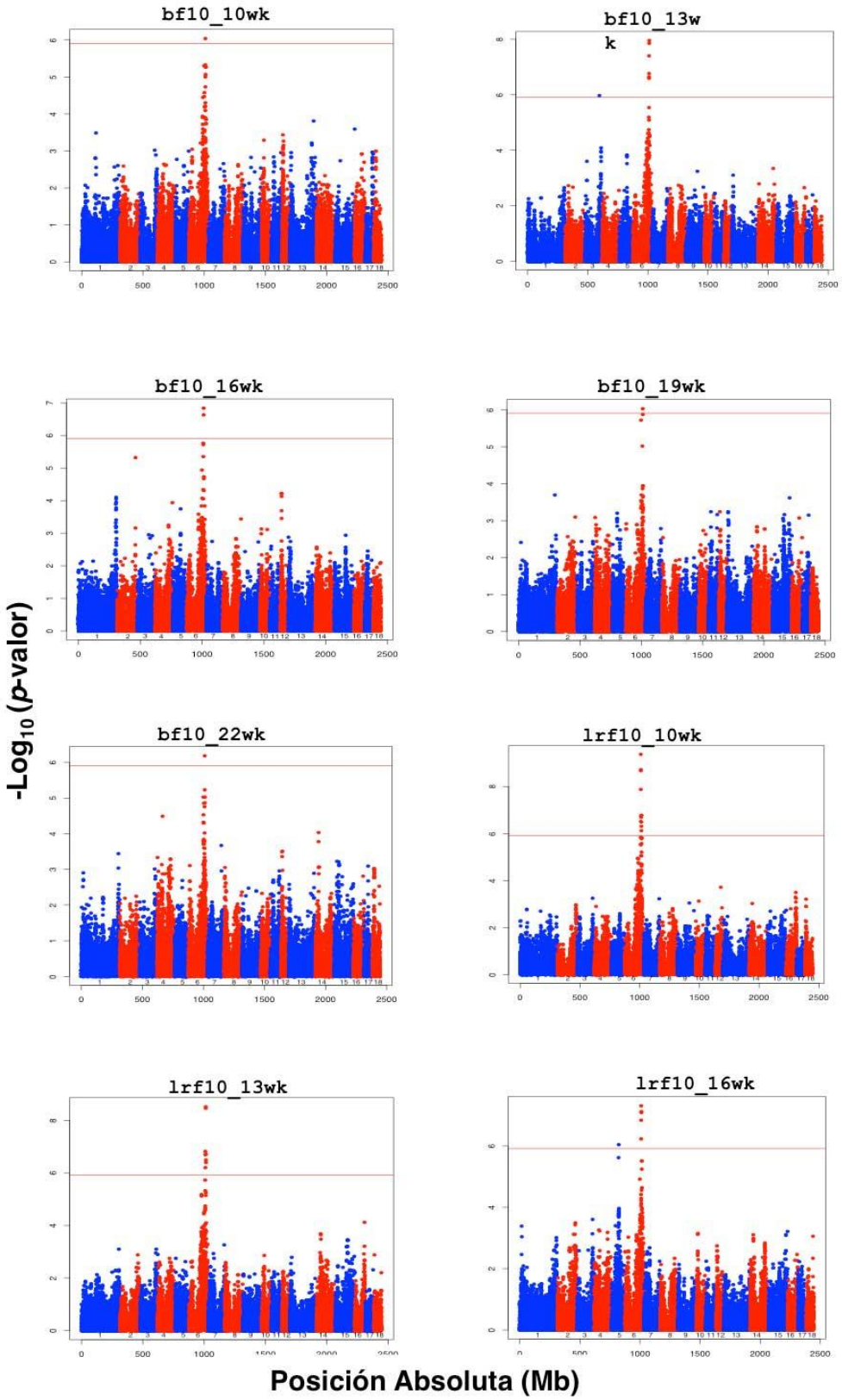
Seg-cromosoma = Número del cromosoma donde el segmento esta localizado, **m$_1$** = modelo [3.3] sin segmento: $y = X\beta + a + e$, **m$_2$** = modelo [3.15] con segmento $y = X\beta + a_1 + a_2 + e$, **SNP** $-$**Log$_{10}$(p-valor)** = −Logaritmo en base 10 de los *p*-valore del SNP seleccionado para crear el segmento, **Lk_m$_1$** = −LogLikehood para m$_1$, **Lk_m$_2$** = −LogLikehood para m$_2$, **LRT** = test de la tasa de verosimilitud para m$_1$ and m$_2$, ***p*-value$_{LRT}$** = *p*-valor para LRT, **VarE_m1** = Varianza del error $\left(\sigma^2_e\right)$ para m$_1$, **VarA_m1** = Varianza aditiva $\left(\sigma^2_A\right)$ para m1,**VarE_m$_2$** = Varianza del error $\left(\sigma^2_e\right)$ para m$_2$, **VarA_m$_2$** = Varianza aditiva $\left(\sigma^2_A\right)$ para m$_2$, **segmVA** = Varianza aditiva para el segmento $\left(\sigma^2_{A_1}\right)$ para m$_2$, %**segmVa** = Proporción en porcentaje (%) del total de la varianza aditiva explicada por el segmento.

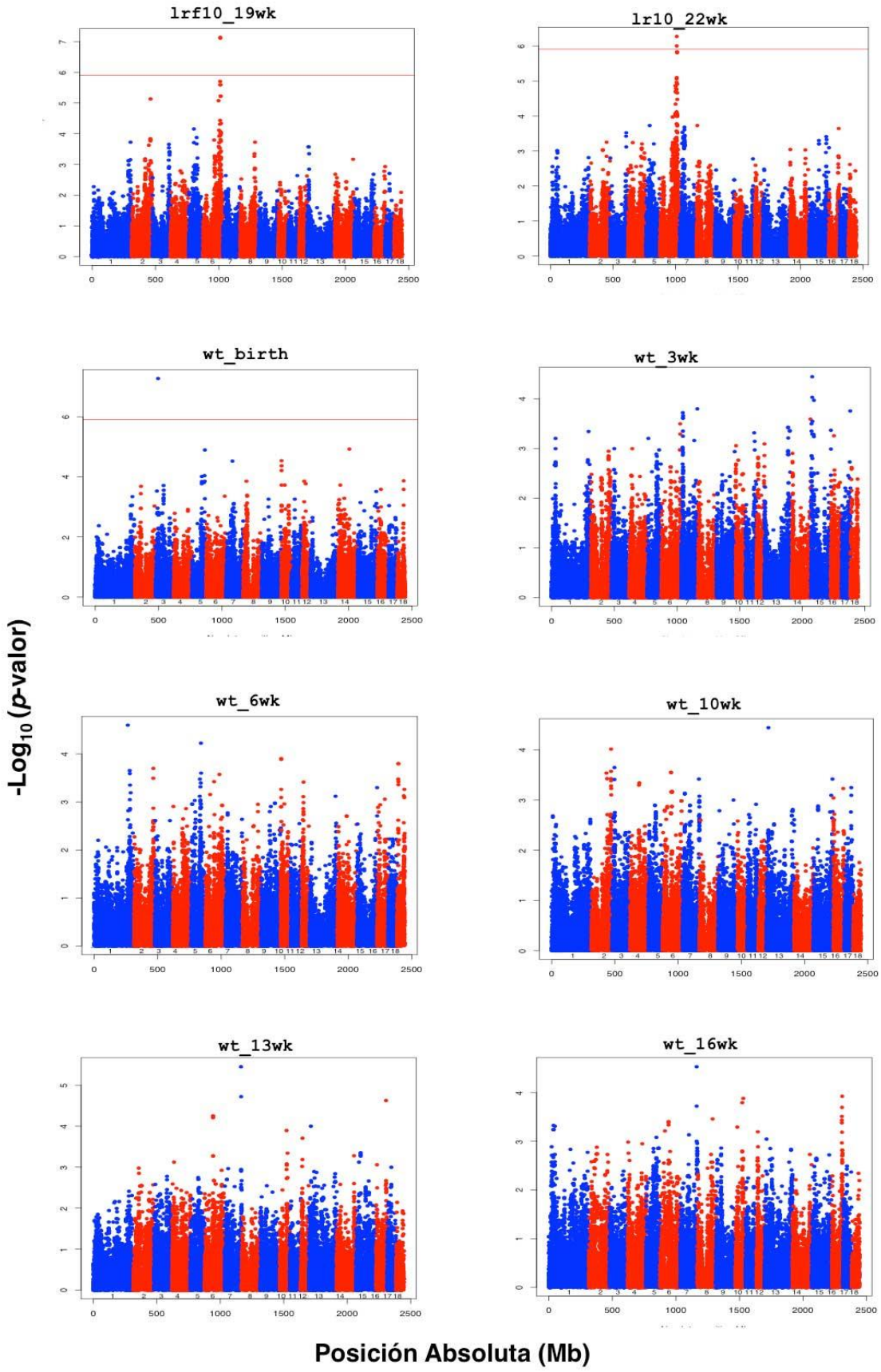
Apéndice 5. Marcador SNP con el –Log(*p*-valor) más alto por carácter

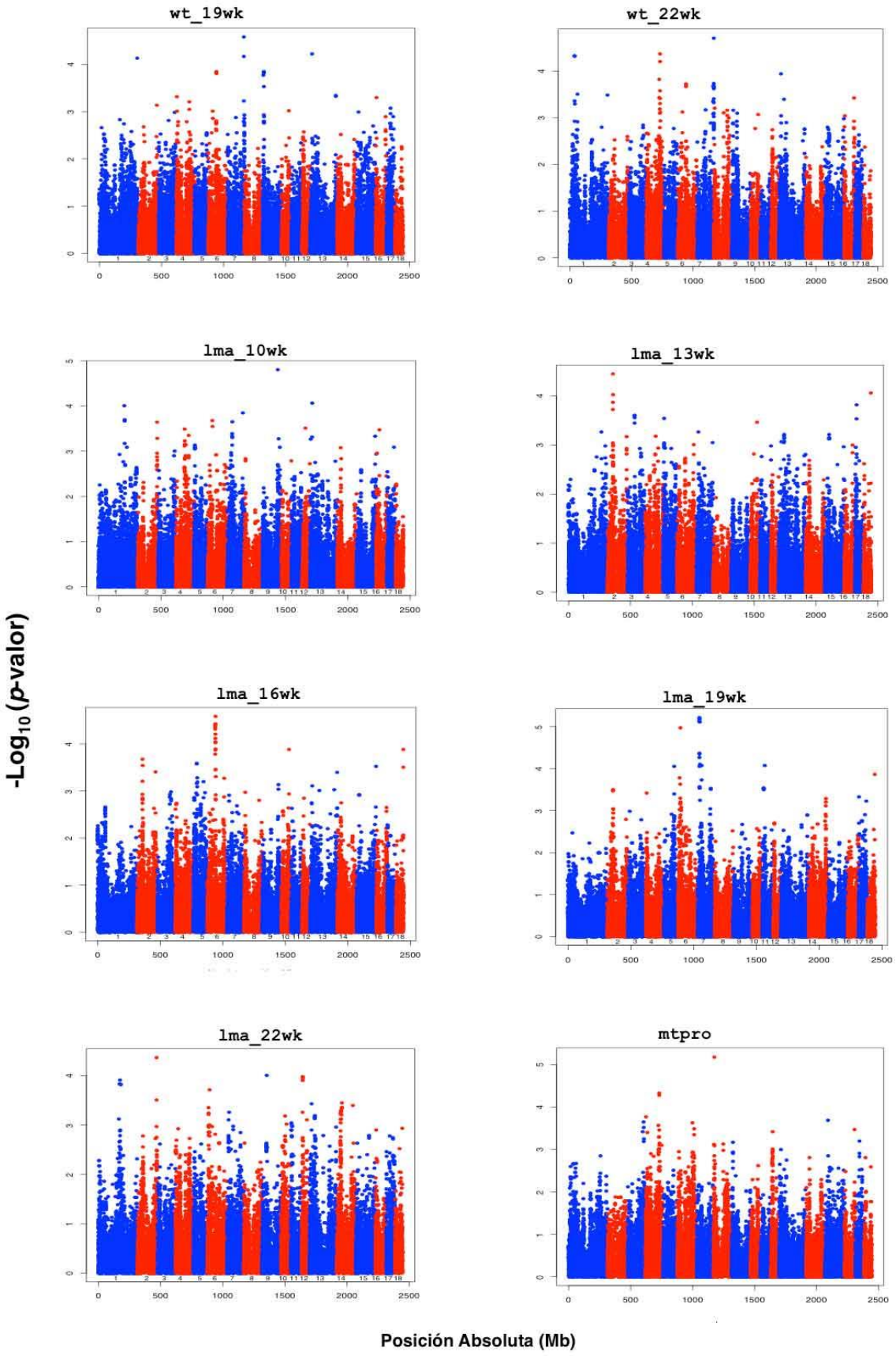
Caracter	SNP-ID	Cromosoma	Posición(Mb)	*p*-valor	$-Log_{10}$(*p*-valor)
ADG	ASGA0021485	4	112.476343	2.73^-06	5.56
Days	ALGA0045948	7	134.035057	2.24^-05	4.65
fftoln	ALGA0045948	7	134.035057	2.31^-05	4.64
mtfat	ALGA0045948	7	134.035057	1.13^-05	4.95
mtpro	ALGA0046300	8	6.406557	6.56^-06	5.18
tofat	DIAS0001383	4	109.781341	7.36^-06	5.13
wt_3wk	ALGA0103594	15	8.142759	3.55^-05	4.45
wt_6wk	H3GA0052708	1	264.895764	2.52^-05	4.60
wt_10wk	ALGA0068161	13	14.664853	3.63^-05	4.44
wt_13wk	ALGA0045724	7	129.466347	3.52^-06	5.45
wt_16wk	ALGA0045724	7	129.466347	2.90^-05	4.54
wt_19wk	ALGA0045724	7	129.466347	2.60^-05	4.58
wt_22wk	ALGA0045948	7	134.035057	1.94^-05	4.71
lma_10wk	H3GA0028038	9	120.964291	1.55^-05	4.81
lma_13wk	ALGA0013323	2	43.071793	3.56^-05	4.45
lma_16wk	ASGA0094600	6	65.297051	2.57^-05	4.59
lma_19wk	M1GA0008394	6	18.653852	1.05^-05	4.98
lma_22wk	ALGA0114651	2	152.416813	4.29^-05	4.37

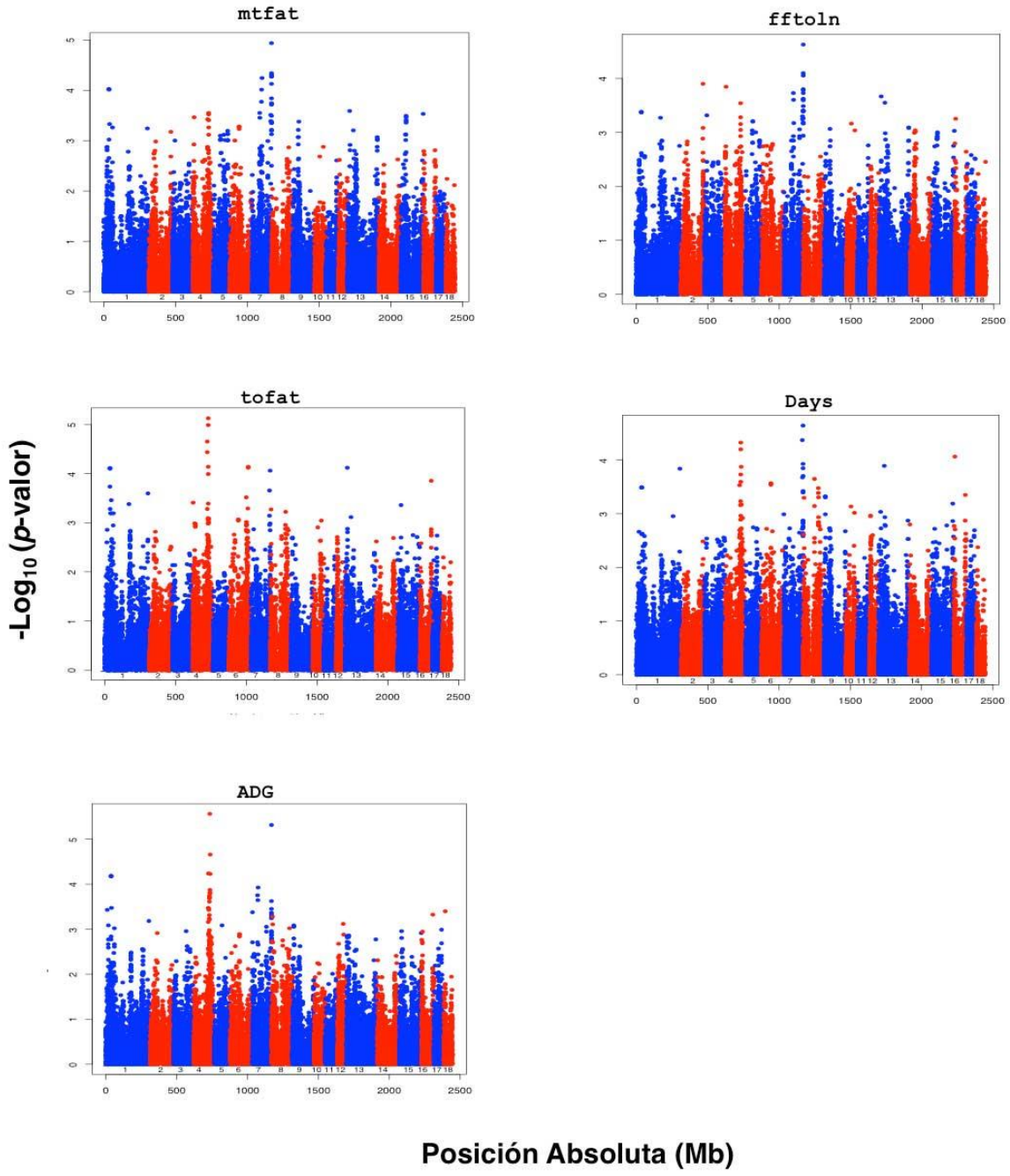
Carácter = ADG: Promedio de ganancia diaria entre las semanas 10 y 22 de edad, Days: número de días para alcanzar los 105 kg. fftoln: carne magra total libre de grasa, mtfat: lípidos del animal eviscerado, mtpro: proteína del animal eviscerado y tofat: tejido graso total registrado a la semana 22 de edad. wt_3(6, 10, 13, 16. 19 and 22)wk: peso registrado en las semanas 3, 6, 10, 13, 16,19 y 22 de edad, lma_10(13, 16, 19 y 22): área del musculo longissimus registrado en las semanas 10, 13, 16, 19 y 22 de edad. SNP-ID= nombre del marcador SNP. Posición (Mb) = Posición física del marcador SNP en Mega-bases dentro del cromosoma. *p*-valor = *p*-valor para el SNP por carácter, $-Log_{10}$(*p*-valor) = –Log10(*p*-valor) para el SNP por carácter.

Apéndice 6. Manhattan-plot para caracteres de crecimiento y de deposición de grasa



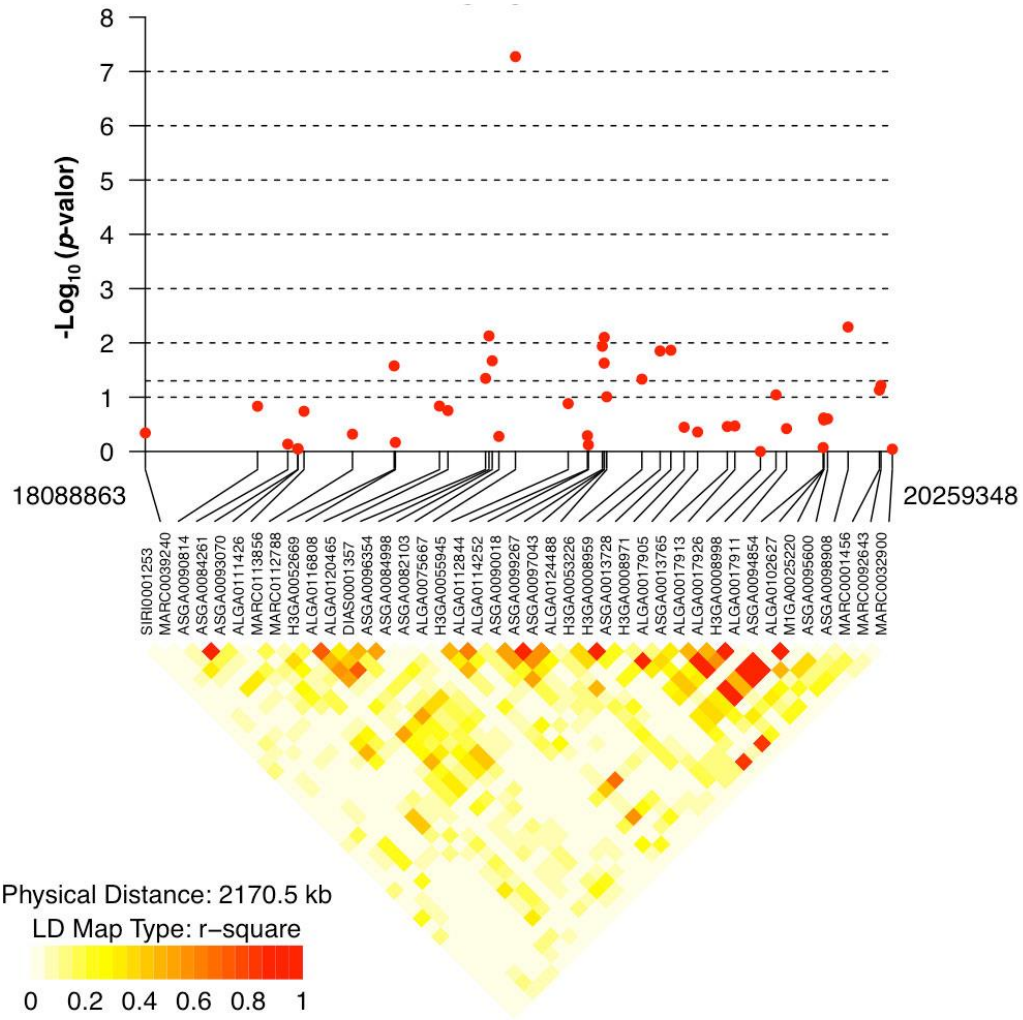

lrf10_19wk
lr10_22wk
wt_birth
wt_3wk
wt_6wk
wt_10wk
wt_13wk
wt_16wk
-Log10 (p-valor)
Posición Absoluta (Mb)


wt_19wk
wt_22wk
lma_10wk
lma_13wk
lma_16wk
lma_19wk
lma_22wk
mtpro
-Log10 (p-valor)
Posición Absoluta (Mb)



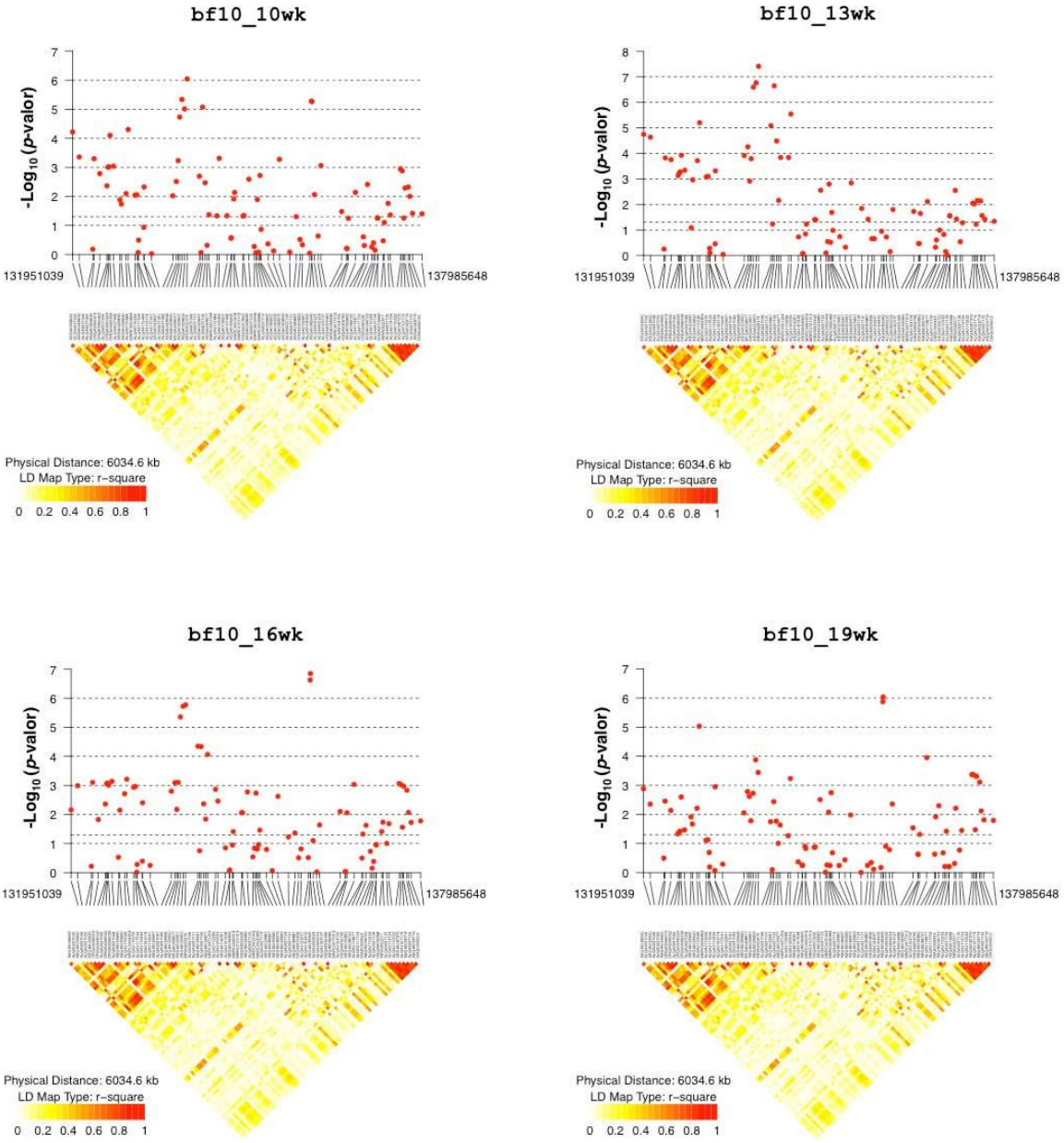
Barrido genómico para 40,569 SNP mediante estandarización $\mathrm{Var}\left(\hat{\boldsymbol{g}}_j\right)$. –Log(*p*-valor) (eje y) contra la posición absoluta de los SNP en Mb (eje x). La línea roja representa un umbral de significancia de *p*-valor $<1.1349\times10^{-6}$ a lo largo del genoma. Números del 1 al 18 representan la identificación del cromosoma. Caracteres: Grasa dorsal en la décima costilla (**bf10_10,13,16,19 y 22wk**), grasa dorsal en la última costilla (**lrf_10,13,16,19 y 22wk**), y área del músculo longissimus (**lma**) registradas en las semanas 10, 13, 16, 19 y 22 de edad. El Peso registrado al nacimiento y en las semanas 3, 6, 10, 13, 16, 19 y 22 de edad (**wt_10,13,16,19 y 22wk**). Mediciones para carne magra total libre de grasa (**fftoln**), tejido graso total (**tofat**), proteína del animal eviscerado (**mtpro**), y lípidos del animal eviscerado (**mtfat**) se registraron en la semana 22 de edad. Promedio de ganancia diaria (**ADG**) se registró entre las semanas 10 y 22 de edad, y el número de días (**days**) para alcanzar los 105 kg se calculó desde los caracteres **ADG** y **days.**

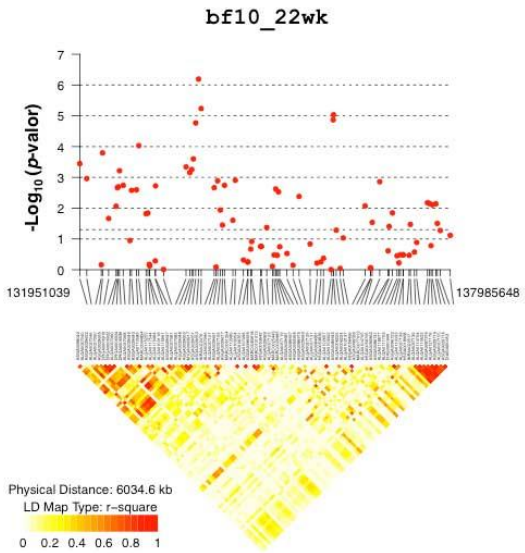
Apéndice 7. Gráfico de Desequilibrio de Ligamiento (LD) para peso al destete en el cromosoma 3.

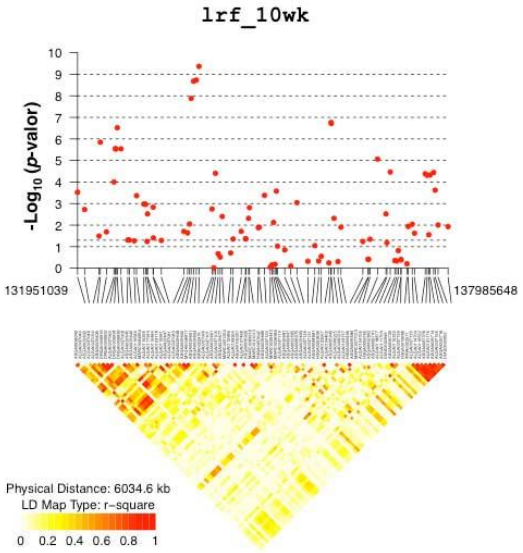


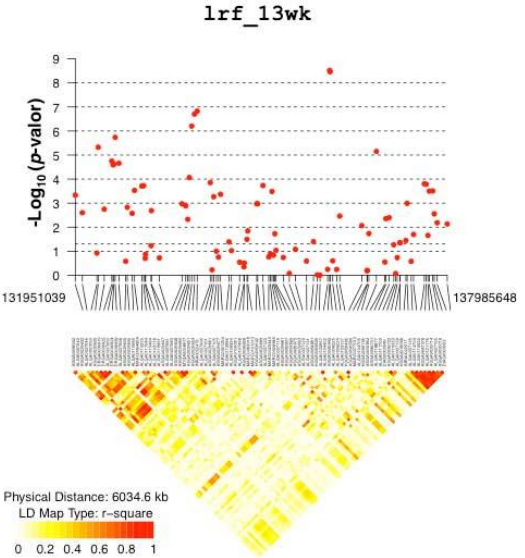
Mapa de Desequilibrio de Ligamiento (LD) para el segmento de 2 Mega-Bases localizadas en el cromosoma 3 para peso al nacimiento (wt_birth). El gráfico de LD muestra un patrón de LD entre los 41 SNP que conforman el segmento en el cromosoma 3. El LD entre los marcadores está medido como r^2. $r^2 = 0$ se muestra de color blanco, $0 < r^2 < 1$ se muestra entre la escala de color amarillo – naranja, y $r^2 = 1$ se muestra en color rojo. En la parte superior se muestra el nombre de los SNP de acuerdo a su posición física dentro del cromosoma.

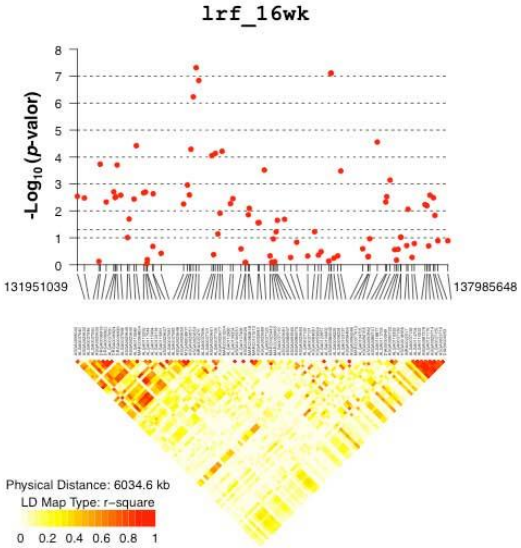
Apéndice 8. Gráficos de Desequilibrio de ligamiento (LD) para caracteres en cromosoma 6

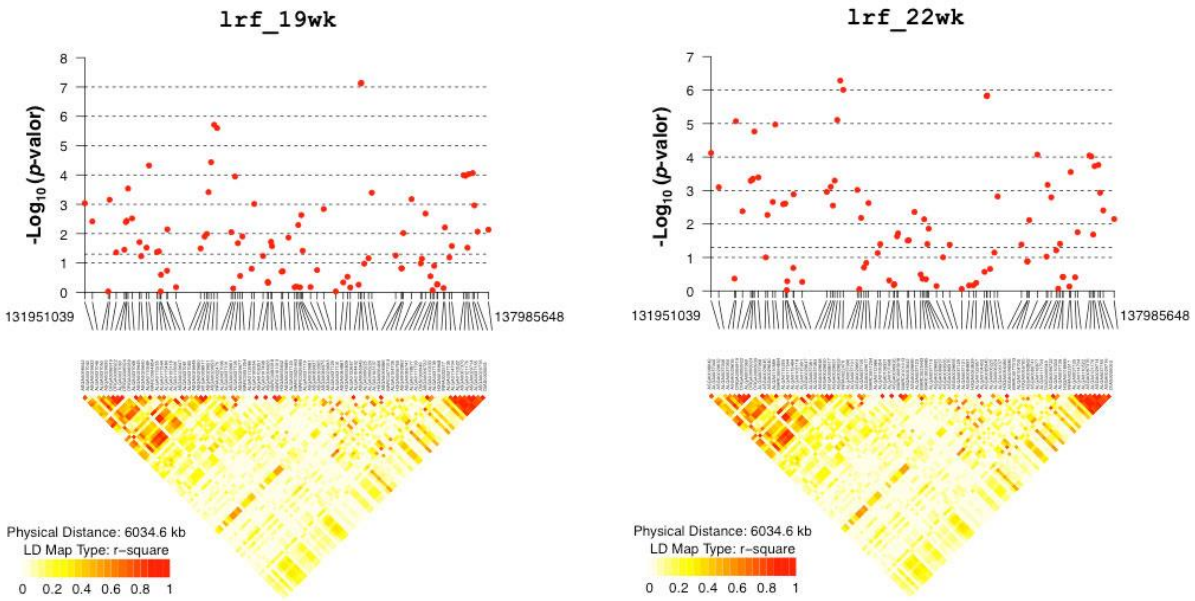



bf10_22wk
-Log10 (p-valor)
131951039
137985648
Physical Distance: 6034.6 kb
LD Map Type: r-square
0 0.2 0.4 0.6 0.8 1


lrf_10wk
-Log10 (p-valor)
131951039
137985648
Physical Distance: 6034.6 kb
LD Map Type: r-square
0 0.2 0.4 0.6 0.8 1


lrf_13wk
-Log10 (p-valor)
131951039
137985648
Physical Distance: 6034.6 kb
LD Map Type: r-square
0 0.2 0.4 0.6 0.8 1


lrf_16wk
-Log10 (p-valor)
131951039
137985648
Physical Distance: 6034.6 kb
LD Map Type: r-square
0 0.2 0.4 0.6 0.8 1



Mapa de Desequilibrio de Ligamiento (LD) para el segmento de 6 Mega-Bases localizadas en el cromosoma 6 para gras dorsal en la décima costilla (bf10_10,13,16,19 y 22wk) y grasa dorsal en la última costilla (lrf_10,13,16,19 y 22wk) registradas en las semanas 10, 13, 16, 19 y 22 de edad. El gráfico de LD muestra un patrón de LD entre los 89 SNP que conforman el segmento en el cromosoma 6. El LD entre los marcadores está medido como r^2. $r^2 = 0$ se muestra de color blanco, $0 < r^2 < 1$ se muestra entre la escala de color amarillo – naranja, y $r^2 = 1$ se muestra en color rojo. En la parte superior se muestra el nombre de los SNP de acuerdo a su posición física dentro del cromosoma.