Relaciones genómicas y control de calidad del genotipado empleando la noción de identidad por descendencia en el marcador

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

Natalia Soledad Forneris
Ing. Agr. - Universidad de Buenos Aires - 2008

Lugar de trabajo: Universidad de Buenos Aires - Facultad de Agronomía - Departamento de Producción Animal - Cátedra de Mejoramiento Genético Animal

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

COMITÉ CONSEJERO

Director de tesis
Rodolfo Juan Carlos Cantet
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc. (Montana State University, E.E.U.U.)
MSc. (University of Illinois, E.E.U.U.)
Ph.D. (University of Illinois, E.E.U.U.)

Co-director
Alicia Leonor Basso
Lic. Cs. Biol. (Universidad de la República, Uruguay)
Esp. (Universidad de la República, Uruguay)
Doctor (Universidad de Buenos Aires, Argentina)

Consejero de Estudios
Juan Pedro Steibel
Ing. Agr. (Universidad Nacional de La Pampa, Argentina)
MSc. (Universidad de Buenos Aires, Argentina)
Ph.D. (Michigan State University, E.E.U.U.)

JURADO DE TESIS

Director de tesis
Rodolfo Juan Carlos Cantet
Ing. Agr. (Universidad de Buenos Aires, Argentina)
MSc. (Montana State University, E.E.U.U.)
MSc. (University of Illinois, E.E.U.U.)
Ph.D. (University of Illinois, E.E.U.U.)

JURADO
Daniel Omar Maizón
Med. Vet. (Universidad de Buenos Aires, Argentina)
MSc. (Universidad de Buenos Aires, Argentina)
Ph.D. (Cornell University, E.E.U.U.)

JURADO
Guillermo Giovambattista
Lic. Biol. (Universidad Nacional de La Plata, Argentina)
Doctor (Universidad Nacional de La Plata, Argentina)

JURADO
Luis Varona Aguado
Lic. Vet. (Universidad de Zaragoza, España)
MSc. (Universidad de Zaragoza, España)
Doctor (Universidad de Zaragoza, España)

Fecha de defensa de la tesis: 15 de ABRIL de 2016

A mis padres, Lita y Maco, los pilares más grandes de mi vida.

AGRADECIMIENTOS

Quisiera agradecer primero y especialmente a mi comité consejero: Fito, Alicia y Juan Pedro. Fito, guía inagotable, marcó un estilo de trabajo que me acompañará para siempre; confió en mí y me dio alas. Alicia me contagió su pasión por la genética y me señaló el esfuerzo y la perseverancia como el camino para lograr objetivos. Juan Pedro me brindó sabios consejos y su apoyo generoso.

A Zulma y Andrés, a quienes admiro y quienes me guiaron generosamente.

A los jurados, quienes dedicaron su tiempo desinteresadamente y quienes, con sus comentarios constructivos, ayudaron a mejorar esta tesis.

A los miembros de la Escuela para Graduados Alberto Soriano, donde me formé durante esta etapa, especialmente a mis profesores y compañeros. Una mención especial a María del Carmen, quien supo motivarme y desafiarme con sus clases.

A la Facultad de Agronomía de la Universidad de Buenos Aires, por brindarme el soporte institucional, especialmente a la Cátedra de Mejoramiento Genético Animal y a la Cátedra de Genética, donde me he desempeñado como docente todos estos años. Una mención especial a Silvia, por sus enseñanzas y su amistad.

A mis alumnos, por desafiarme constantemente con sus preguntas.

A mis compañeros y amigos del grupo de Mejoramiento Genético Animal de la FAUBA: Anita, Sebas, Andrecito, Yeni, Caro, Juancho, Joselito, Dani, Majo, Laura, Mónica, Valeria, Martín y Oscar, por su apoyo incondicional y sus sabios consejos.

A mis compañeros de oficina y amigos, Lourdes y Hugo, por su alegría cotidiana.

A mi familia y mis amigos, por su cariño, comprensión y constante estímulo.

Finalmente, a todas aquellas personas a quienes tuve la suerte de conocer y aprender de ellas gracias al desarrollo de esta tesis.

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.

Natalia Soledad Forneris

PUBLICACIONES DERIVADAS DE LA TESIS

Forneris, N.S., Legarra, A., Vitezica, Z.G., Tsuruta, S., Aguilar, I., Misztal, I., Cantet, R.J.C. 2015. Quality control of genotypes using heritability estimates of gene content at the marker. *Genetics* 199: 675–681.

Forneris, N.S., Steibel, J.P., Legarra, A., Vitezica, Z.G., Bates, R.O., Ernst, C.W., Basso, A.L., Cantet, R.J.C. 2015. A comparison of methods to estimate relationships using pedigree and markers in livestock populations. *Journal of Animal Breeding and Genetics* (en prensa).

ÍNDICE GENERAL

ÍNDICE DE CUADROS

Cuadro **Página**

ÍNDICE DE FIGURAS

Figura **Página**

ABREVIATURAS

Bg-IBD	*background IBD*, IBD a nivel poblacional
BLUP	*best linear unbiased predictor*, mejor predictor lineal insesgado
bp	*base pairs*, pares de bases
BV	*breeding value*, valor de cría
cM	centimorgan
DG	*descent graph*, grafo de descendencia
DGV	*direct genomic value*, valor genómico directo
DYD	*daughter yield deviations*, desvíos de producción de las hijas de un toro
EBV	*estimated breeding value*, valor de cría predicho
ECM	error cuadrático medio
GC	*gene content*, conteo de alelos
GBLUP	metodología BLUP de predicción del valor de cría genómico
GEBV	*genomic estimated breeding value*, valor de cría genómico predicho
H-W	*Hardy-Weinberg equilibrium*, equilibrio Hardy-Weinberg
HMM	*hidden Markov model*, modelo Markov oculto
IBD	*identical/identity by descent*, idéntico/identidad por descendencia
IBS	*identical/identity by state*, idéntico/identidad por estado
IGF	*inheritance-generating function*, función generadora de herencia
LD	*linkage disequilibrium*, desequilibrio gamético o de ligamiento
LE	*linkage equilibrium*, equilibrio gamético
LRT	*likelihood ratio test*, test de cociente de verosimilitud
MA	modelo animal aditivo
MAF	*minor allele frequency*, frecuencia del alelo menos usual

PA	*parent average*, promedio de los valores de cría de los padres
QC	*quality control*, control de calidad
QTL	*quantitative trait locus*, locus que influye sobre un carácter cuantitativo
REML	*restricted maximum likelihood*, máxima verosimilitud restringida
SNP	*single nucleotide polymorphism*, polimorfismo de un único nucleótido
SG	selección genómica
YD	*yield deviations*, desvíos de producción

Título: Relaciones genómicas y control de calidad del genotipado empleando la noción de identidad por descendencia en el marcador

RESÚMEN

La exactitud de los valores de cría genómicos (GEBVs) depende de la habilidad de la matriz de relaciones genómicas (***G***) para capturar la variabilidad en la proporción de genoma compartido entre individuos con el mismo parentesco genealógico, así como del control de calidad de SNPs. Esta tesis desarrolla métodos para calcular ***G*** considerando: el pedigree, SNPs de calidad y el desequilibrio de ligamiento (LD) entre SNPs. Primero, se presenta un método simple basado en máxima verosimilitud restringida para identificar SNPs con muchos genotipos asignados incorrectamente mediante la estimación de la heredabilidad (h^2) de la "cantidad de alelos" (GC) para cada SNP. El GC es el número de copias de un alelo en particular en el genotipo de un animal (0, 1 ó 2). El método prueba la hipótesis nula: h^2=1 (sin errores en el genotipado). El método es ilustrado con datos reales; su sensibilidad y especificidad se evalúan mediante simulación, mostrando mejor performance que el procedimiento estándar de detección de errores Mendelianos. Luego, se compararon dos enfoques para estimar ***G***: 1) identidad-por-estado ($\boldsymbol{G}_{\mathbf{VR}}$), utilizando las frecuencias alélicas observadas o las de la población base, 2) identidad-por-descendencia ($\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$), usando análisis de ligamiento y LD. Los estimadores fueron evaluados en precisión y en sesgo empírico respecto de la verdadera proporción de genoma compartido simulada ($\boldsymbol{G}_T$). Todos fueron prácticamente insesgados. $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ mostró el menor error cuadrático medio empírico y la mayor correlación con $\boldsymbol{G}_{\mathbf{T}}$. La exactitud de los GEBVs fue mayor con $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ que con $\boldsymbol{G}_{\mathbf{VR}}$. En datos reales, la varianza muestral de $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ se aproximó más al valor teórico en cada relación de parentesco. En resumen, el uso de $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ mejora la precisión de las estimaciones y la exactitud de los GEBVs para los candidatos a la selección. Los métodos propuestos consideran todos los individuos genotipados y su pedigree de manera conjunta.

Palabras clave: proporción de genoma IBD compartido, selección genómica, desequilibrio gamético, cantidad de alelos, REML

Tittle: Identity-by-descent approaches for the calculation of genomic relationships and the quality control of genotypes

ABSTRACT

Accurate prediction of genomic breeding values (GEBVs) with a genomic relationship matrix (***G***) depends on capturing the variability in genome sharing of relatives with the same pedigree relationship. Accuracy can also be increased by quality control (QC) filtering of SNPs. This dissertation develops methods to set up ***G*** using both pedigree and QC-filtered SNPs, while accounting for linkage disequilibrium between SNPs. First, a simple method based on restricted maximum likelihood is presented to identify SNPs with many incorrectly assigned genotypes by estimating the heritability (h^2) of gene content (GC) at each SNP. GC is the number of copies of a particular allele in a genotype of an animal (0, 1 or 2). The method tests the null hypothesis: h^2=1 (no genotyping errors) using a likelihood-ratio test. The method is illustrated with a real data set and its sensitivity and specificity were evaluated using simulated data, performing better than the standard checking of Mendelian errors. Next, two approaches to set up ***G*** are compared: 1) identity-by-state, with either the observed ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) or base population ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$) allele frequencies, 2) identity-by-descent, using linkage disequilibrium and linkage analysis ($\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$). ***G*** estimators were evaluated for precision and empirical bias of genomic relationships with respect to true genome shared values ($\boldsymbol{G}_{\mathbf{T}}$) using simulated data. All estimators were nearly unbiased. $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ displayed the lowest sampling error and the highest correlation with $\boldsymbol{G}_{\mathbf{T}}$. Accuracy of GEBVs for selection candidates was higher when $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ was used, and identical between $\boldsymbol{G}_{\mathbf{VR}}$ matrices. In real data, $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$'s sampling variance was the closest to the theoretical value for each pedigree relationship. Use of $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ improved the precision of estimates and the accuracy of GEBVs. Contrary to other techniques, both the proposed QC method and $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ use all genotype and pedigree data in the population simultaneously.

Key words: proportion of the genome shared IBD, genomic selection, linkage disequilibrium, gene content, REML

CAPÍTULO 1

Introducción general

Introducción general

El objetivo de la evaluación genética animal es la predicción del mérito genético o valor de cría (en inglés *breeding value*, BV) de los animales para los distintos caracteres productivos. En el modelo infinitesimal de herencia poligénica (Falconer y Mackay, 1996), los BVs son controlados por un número muy grande de genes que influyen sobre un carácter cuantitativo denominados poligenes, o bien de QTLs (por *quantitative trait loci* en inglés), que son los segmentos de ADN donde se localizan esos genes. En el modelo infinitesimal convencional los QTLs actúan en forma aditiva, sin ligamiento entre ellos, teniendo cada uno un efecto pequeño sobre la expresión fenotípica del carácter (Bulmer, 1985). Se define, entonces, el BV de un animal como la suma de los efectos genéticos aditivos de todos los genes que gobiernan un carácter cuantitativo, y constituye la variable aleatoria de interés en el modelo animal aditivo (MA) de selección, el cual ha sido utilizado tradicionalmente en las evaluaciones genéticas, y en el que el BV se predice a partir de los registros fenotípicos del animal y de individuos relacionados.

La disponibilidad de una tecnología de evaluación de genotipos ("genotipado") con una alta densidad de polimorfismos de un solo nucleótido o SNPs (en inglés *single nucleotide polymorphism*), cuyo precio disminuye continuamente, ha permitido predecir el BV empleando una metodología conocida como *selección genómica* (SG; Meuwissen *et al.*, 2001). En la SG, la predicción del BV consiste en la estimación simultánea de miles de efectos, cada uno atribuible a un SNP en el genoma, y que tienen un efecto sobre cierto carácter cuantitativo. El supuesto de la SG es que, dada una elevada densidad de marcadores distribuidos a lo largo del genoma, cada QTL se asociará con al menos un SNP en las cercanías y que, al ajustar un efecto para cada marcador, se

capturaría información acerca de los QTLs (Meuwissen *et al.*, 2001). Dicha asociación se conoce como desequilibrio gamético o de ligamiento (LD), y consiste en la ausencia de independencia estadística entre las frecuencias de los alelos en dos o más loci (o segmentos de ADN) en el mismo, o en distinto, cromosoma. En poblaciones de especies pecuarias, el LD es causado por procesos recurrentes de deriva, selección y cruzamiento entre razas (Haley, 1999).

En los programas de mejoramiento que emplean SG, la ganancia genética aumenta debido al incremento de la exactitud de selección. Sin embargo, las evaluaciones genéticas realizadas con el MA ocurren de modo tal que la exactitud de selección en los padres de toros y vacas es de por sí elevada (por ejemplo, en las pruebas de progenie puede alcanzar 0,99). Si bien las pruebas de progenie poseen una alta exactitud, el intervalo entre generaciones es elevado, lo que disminuye la tasa de ganancia genética anual. En este punto, la SG tiene el potencial de aumentar la tasa de ganancia genética, al reducir el tiempo de selección o intervalo generacional debido a la posibilidad de seleccionar toros antes de que se realice la prueba de progenie, y aumentar la exactitud en las predicciones de los BVs a edad temprana (Schaeffer, 2006). Más precisamente, el BV de un animal se puede descomponer en el promedio de los BVs de ambos padres más un residuo de segregación Mendeliana (Bulmer, 1985, página 125). Este residuo es producto de la recombinación de los genes presentes en los cromosomas de cada padre durante la meiosis y representa la configuración génica propia de cada animal. La SG puede captar la información sobre este residuo mucho antes de que un animal cuente con su dato propio y/o hijos con datos (e.g., un toro joven). Esto sólo es posible mediante la utilización de marcadores moleculares en alta densidad cercanos a cada QTL (Goddard, 2008; Cantet *et al.*, 2008). En síntesis, la SG es útil en aquellas situaciones en las que la exactitud de selección es baja; por ejemplo,

en caracteres de baja heredabilidad, o muy costosos de medir, o en aquellos que se miden en etapas tardías de la vida del animal (e.g., a la faena) (Meuwissen, 2003).

Por razones de costo y logística, no es posible genotipar a todos los animales de una población. Coexisten, entonces, diferentes tipos de animales: a) individuos genotipados y b) individuos no genotipados. Por esta razón, actualmente las evaluaciones genómicas utilizan procedimientos de varios pasos (VanRaden, 2008; VanRaden *et al.*, 2009), los cuales son propensos a sesgos y pérdida de información. Estos pasos son: 1) la evaluación mediante el MA, 2) la estimación de los efectos genómicos para un número reducido de animales genotipados, y 3) la predicción de los valores de cría genómicos mediante un índice de selección combinando la información producida en los pasos 1) y 2).

Un modo de simplificar estos pasos es modificar la matriz $\boldsymbol{A}$ de relaciones aditivas dentro de la evaluación genética tradicional de modo de incluir la información genómica y producir una única evaluación genética empleando ambas fuentes de información: pedigree y marcadores moleculares. Varios autores (Legarra *et al.*, 2009; Misztal *et al.*, 2009; Christensen y Lund, 2010; Meuwissen *et al.*, 2011) propusieron modificar la matriz $\boldsymbol{A}$, condicionando el valor genético de los animales no genotipados en el valor genético de animales genotipados vía un índice de selección, y luego utilizando una matriz de relaciones genómicas de parentesco ($\boldsymbol{G}$) para estos últimos. Condicional a los marcadores presentes en cada individuo, las relaciones genómicas son calculadas para cada par de individuos e ignoran la información del pedigree; por lo tanto, se basan en el concepto de identidad por estado (en inglés *identity by state*, IBS). Las relaciones aditivas (Wright, 1922) utilizadas en el MA, en cambio, son condicionales a la información del pedigree y se basan en el concepto de identidad por descendencia (en inglés *identity by descent*, IBD; Malécot, 1948). Asimismo, en el

cálculo de los elementos de $\boldsymbol{G}$ se asume que los efectos de los marcadores son variables aleatorias normales, idénticamente distribuidas e independientes; esto último implica equilibrio de ligamiento o equilibrio gamético (LE) entre ellos. Sin embargo, el supuesto de LE es violado cuando la densidad de los mismos es alta, sumado a que la teoría de la SG descansa en el supuesto de existencia de LD (Gianola *et al.*, 2009).

La utilización de genotipos de alta densidad de SNPs para aplicar la SG se ha convertido en una práctica común. Con la creciente importancia de esta nueva fuente de información, la necesidad de herramientas para editar y comprobar la calidad de este tipo de datos también aumenta. La calidad del genotipado define el éxito de posteriores análisis estadísticos cuando se trabaja con datos genómicos. El control de calidad (en inglés *quality control*, QC) y filtrado de SNPs en SG puede incrementar la precisión, reducir el esfuerzo computacional y mejorar la estabilidad de las estimaciones de los efectos de los marcadores (Wiggans *et al.*, 2009). Es habitual, entonces, que se empleen procedimientos de QC para depurar una base de datos de genotipos antes de que sea utilizada en el cálculo de las relaciones genómicas.

Uno de los pasos habituales en la edición de los genotipos de un SNP, es verificar si hay inconsistencias Mendelianas. Los conflictos Mendelianos son detectados porque los genotipos observados para un SNP son inconsistentes con el patrón de transmisión esperado por la Primera Ley de Mendel dado un pedigree. El procedimiento habitual para identificar y filtrar las inconsistencias es analizando genotipos de tríos (padre-madre-cría) y de pares padre-cría (Wiggans *et al.*, 2009, 2012). Sin embargo, los errores que son consistentes desde un punto de vista Mendeliano a menudo pueden pasar desapercibidos (e.g., un hijo “Aa” de padres “Aa” y “AA” es erróneamente genotipado como “AA”) con este tipo de filtro. Esto ocurre porque al observar la

herencia de un SNP, no se utilizan simultáneamente todos los genotipos de los individuos genotipados y su pedigree.

Sobre la base de todo lo expuesto, el objetivo general de esta tesis es el desarrollo de matrices de relaciones genómicas teniendo en cuenta: a) el pedigree, b) la información de un chip de alta densidad de SNPs de calidad, c) el desequilibrio de ligamiento (LD) entre SNPs, y su implementación algorítmica dentro de la evaluación genética animal. A tal efecto, se proponen métodos eficientes para incluir la información de los SNPs de un modo distinto al que se utiliza actualmente en SG. Cada uno de los capítulos que siguen aborda un objetivo específico de la tesis. En el Capítulo 2, se propone un método para identificar SNPs con genotipos asignados incorrectamente en una gran proporción de los animales de la base de datos, mediante la estimación de la heredabilidad de "la cantidad de alelos" o "conteo de alelos" (en inglés *gene content*, GC) para cada SNP. El GC es el número de copias de un alelo de referencia en particular en el genotipo de un animal (0, 1 ó 2). El método trata al GC como un carácter cuantitativo y asume que la relación entre los GCs es lineal, y que la covarianza entre los GCs es proporcional a la fracción de alelos IBD entre los animales. En el Capítulo 3, se propone una fórmula para estimar las relaciones genómicas entre individuos genotipados introduciendo la noción de fracción de genoma IBD compartido. A tal efecto, se emplea un algoritmo utilizado en el área de la genética humana, que incorpora la información del pedigree además de aquella proveniente de un panel denso de SNPs, teniendo en cuenta el LD entre los marcadores. En el Capítulo 4, se compara la metodología utilizada actualmente en SG para el cálculo de las relaciones genómicas, la cual estima la fracción de genoma compartido para cada par de individuos genotipados basándose únicamente en la información de los marcadores presentes en cada individuo, con aquella propuesta como alternativa en el Capítulo 3. En el Capítulo

5, se evalúa cómo el aumento en la precisión de las estimaciones de las relaciones genómicas se podría traducir en un aumento en la exactitud de las predicciones genómicas de los BVs, dependiendo del estimador de $\boldsymbol{G}$ utilizado. Para ello, se aborda la implementación algorítmica de la matriz de relaciones genómicas propuesta en el Capítulo 3 dentro de la evaluación genética animal. Para su implementación, la nueva matriz genómica es extendida de modo de incluir la información fenotípica de los animales del pedigree no genotipados en el cálculo de la exactitud de las predicciones genómicas.

CAPÍTULO 2

Estimación de la heredabilidad del conteo de alelos en el marcador para el control de calidad del genotipado[1]

[1] Forneris, N.S., Legarra, A., Vitezica, Z.G., Tsuruta, S., Aguilar, I., Misztal, I., y Cantet, R.J.C. 2015. Quality control of genotypes using heritability estimates of gene content at the marker. *Genetics* 199: 675–681.

Estimación de la heredabilidad del conteo de alelos en el marcador para el control de calidad del genotipado

2.1 Introducción

La reciente disponibilidad de plataformas para evaluar genotipos de un gran número ("alta densidad") de marcadores de ADN polimórficos de nucleótido único (SNP) ha revolucionado la evaluación genética animal y vegetal. Además, y concurrentemente, ha habido un gran desarrollo de las técnicas de "imputación", que consisten en predecir los SNPs faltantes de genotipos de "baja densidad", de modo de obtener genotipos de alta densidad. Sin embargo, estas metodologías son susceptibles a fallas técnicas o errores de genotipado, los cuales se deben a: 1) errores de "laboratorio húmedo" (mala calidad de las muestras de ADN, o de las lecturas, etc.), 2) la diferente naturaleza bioquímica de los paneles de marcadores, 3) el cambio de una etiqueta por otra en las muestras de ADN, 4) errores en los registros de pedigree. A modo de ejemplo, Wiggans *et al.* (2012) eliminaron 127 marcadores de un total de 2.886, dentro del panel Bovine3K BeadChip (Illumina Inc., San Diego, CA) por presentar más de un 2% de conflictos Mendelianos. Asimismo, los errores pueden surgir de los procedimientos de imputación; por ejemplo, si la posición de un marcador en el mapa ha sido determinada en forma incorrecta, sus marcadores adyacentes estarán mal asignados y el análisis de imputación sería incorrecto (Hickey *et al.*, 2012; Wang *et al.*, 2013). La calidad de los registros genotípicos en las evaluaciones genómicas ha sido entonces considerada cuidadosamente desde hace algún tiempo, razón por la cual se han desarrollado una serie de procedimientos para el control de calidad (en inglés *quality control*, QC). Estos procedimientos se emplean, por ejemplo, para depurar una base de datos de genotipos antes de que sea utilizada en la estimación de las relaciones

genómicas de parentesco. El filtrado de QC de SNPs en la evaluación genómica puede aumentar la precisión, reducir el esfuerzo computacional y mejorar la estabilidad de las estimaciones de los efectos de los demás SNPs (Wiggans *et al.*, 2009). En este capítulo se propone un método basado en máxima verosimilitud restringida o residual (en inglés *restricted maximum likelihood*, REML) para comprobar la calidad de los genotipos de los marcadores en una población con un pedigree posiblemente complejo, genotipada parcial o totalmente, sobre un conjunto de marcadores bialélicos. Este método apunta, específicamente, a detectar loci para los cuales se genotiparon erróneamente un número importante de individuos. A los efectos de la presentación, primero se describen brevemente los métodos de QC que se utilizan en la actualidad. A continuación, se presenta el método propuesto en esta tesis y, finalmente, se muestran los resultados obtenidos utilizando una base de datos simulada y otros empleando datos reales de disponibilidad pública.

El control de calidad (QC) de genotipos. Los criterios de filtrado de QC más comúnmente utilizados incluyen el QC de los individuos en función de su "call rate" (proporción de SNPs con genotipo asignado), la existencia de duplicados y de conflictos padre(madre)-hijo; y el QC de los SNPs en función de su "call rate" (proporción de individuos con genotipo asignado), la frecuencia del alelo menos usual (inglés *minor allele frequency*, MAF), las desviaciones del equilibrio Hardy-Weinberg y, en particular, las inconsistencias o conflictos Mendelianos (véase el trabajo de Wiggans *et al.* (2009) para una descripción general de estos criterios). Los conflictos Mendelianos son detectados porque los genotipos observados para un SNP son inconsistentes con el patrón de transmisión esperado en un pedigree por la Primera Ley de Mendel. Se puede evaluar la presencia de inconsistencias analizando genotipos de tríos (padre-madre-cría) y de pares padre-cría (Wiggans *et al.*, 2009, 2012). Sin embargo, los errores que son

consistentes desde un punto de vista Mendeliano a menudo pueden pasar desapercibidos (e.g., un hijo “Aa” de padres “Aa” y “AA” es erróneamente genotipado como “AA”). Alternativamente se puede evaluar si un padre heterocigota que se aparea con varias hembras tiene una heterocigosidad promedio de 0,5 en su progenie (LeRoy *et al.*, 2013).

Cheung *et al.* (2014) propusieron un método para detectar inconsistencias Mendelianas y errores ocasionales que parecen consistentes desde un punto de vista Mendeliano. El procedimiento fue diseñado para inferir los patrones de herencia de los marcadores en pedigrees humanos de tamaño pequeño a moderadamente grande (e.g., 100 individuos). En humanos, éste y otros procedimientos de detección similares, son computacionalmente eficientes para detectar errores de genotipado a nivel de marcador, y podrían permitir detectar aquellos errores ocasionalmente consistentes a nivel de individuo en algunos marcadores. Sin embargo, no resulta inmediatamente evidente si este método es igualmente aplicable a datos de la producción animal o vegetal donde el pedigree puede ser de tamaño considerable (e.g., >1.000 individuos), sobretodo cuando la motivación es detectar marcadores sospechosos para los cuales una proporción considerable de los individuos presenta errores de genotipado.

La aplicación de estos filtros de QC (excepto Cheung *et al.*, 2014) no utiliza toda la información disponible del pedigree y de los marcadores de manera global. Por ejemplo, consideremos diez hermanos enteros, cinco con genotipos “AA” y cinco con genotipos “aa”, provenientes de un padre y una madre heterocigotas. Esta distorsión en la segregación no es un conflicto Mendeliano. Sin embargo, la situación es de ocurrencia muy poco probable. El problema se vuelve muy complejo en pedigrees de gran tamaño en los que sólo una fracción de los animales está genotipada. Por ejemplo, VanRaden (2008) utilizó los genotipos de 3.000 toros, todos conectados a través de un pedigree que abarcaba 23.105 individuos.

2.2 Objetivo

El objetivo de este capítulo es presentar un método práctico para identificar SNPs de baja calidad (i.e., SNPs cuyos genotipos han sido asignados erróneamente para una gran proporción de los individuos) considerando el conteo de alelos (en inglés *gene content*, GC) en un SNP particular, como un carácter de naturaleza cuantitativa y testeando la hipótesis nula $h^2 = 1$ en cada SNP. La sensibilidad y la especificidad del método son evaluadas por medio de la simulación de datos en una población núcleo porcina. Asimismo, el método se ilustra con una base real de datos en porcinos.

2.3 Materiales y métodos

2.3.1 Teoría del método

2.3.1.1 *Conteo de alelos visto como un carácter cuantitativo*

El conteo de alelos (z) presentes en un marcador es el número de copias de un alelo de referencia en particular (por ejemplo, $z = 0$, 1 ó 2 para "AA", "AG" y "GG", Falconer y McKay, 1996). Esto permite analizar al conteo de alelos (realizado u observado) como un carácter cuantitativo donde el mapeo de los genotipos codificados a fenotipos $\{0, a, 2a\}$, se realiza sobre la base que el efecto aditivo a del alelo de referencia ("G" en el ejemplo de arriba) es exactamente igual a uno. Por lo tanto, se asume la ausencia de dominancia o epistasis. Adicionalmente, y a menos que exista una mutación en el marcador, este es genotipado con precisión y no existe error asociado con el fenotipo. En tal sentido y por construcción, la heredabilidad del conteo de alelos es igual a 1 y toda la variación genética es estrictamente aditiva. La media de z en la población base es igual a $2p$, donde p es la frecuencia del alelo de referencia en la población base y su varianza es $2p(1 - p)$.

La covarianza entre los conteos de alelos de dos individuos es $\mathrm{cov}\left(z_i, z_j\right) = A_{ij}\, 2p\left(1-p\right)$ (Cockerham, 1969, ecuación (8)) donde A_{ij} es la relación aditiva entre dos individuos, *i* y *j*, la que se calcula generalmente a partir del pedigree (véase Toro *et al.*, 2011 para una explicación más detallada). Este hecho ha sido recientemente resaltado y utilizado por McPeek *et al.* (2004) y Gengler *et al.* (2007) en contextos similares. Por lo tanto, sea $\boldsymbol{z}$ un vector que contiene los conteos de alelos de un conjunto de individuos genotipados, $\mathrm{Var}(\boldsymbol{z}) = \boldsymbol{A}_{22}\, 2p\left(1-p\right)$, donde $\boldsymbol{A}_{22}$ es la matriz de relaciones aditivas entre los individuos genotipados. Por otra parte, $\boldsymbol{A}_{22}$ es una submatriz de la matriz $\boldsymbol{A}$ de relaciones aditivas entre todos los individuos del pedigree. Con cada marcador, analizaremos el conteo de alelos $\boldsymbol{z}$ mediante un modelo lineal: $\boldsymbol{z} = \boldsymbol{1}\left(2p\right) + \boldsymbol{u} + \boldsymbol{e}$, donde $\boldsymbol{u}$ es la desviación de cada individuo de la media (2*p*) y $\boldsymbol{e}$ es el error aleatorio que debe ser igual a 0 en ausencia de errores de genotipado. En ese caso, $\sigma_e^2 = 0$, de modo que $h^2 = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}$ es igual a 1 cuando no existen errores de genotipado. Recordemos que $\mathrm{Var}\left(\boldsymbol{u}\right) = \boldsymbol{A}_{22}\boldsymbol{\sigma}_u^2$ y $\boldsymbol{\sigma}_u^2 = 2p\left(1-p\right)$.

2.3.1.2 *Estimación de la heredabilidad*

Comúnmente los componentes de varianza para obtener la heredabilidad con genealogías complejas, se estiman mediante "máxima verosimilitud restringida" (REML; Patterson y Thompson, 1971) por sus propiedades como estimador (Searle *et al.*, 1992). Los estimadores REML son desarrollados asumiendo normalidad, situación que no ocurre con el conteo de alelos debido a que no se trata de un carácter continuo. Sin embargo, el supuesto de normalidad multivariante para el conteo de alelos es bastante común (McPeek *et al.*, 2004; Gengler *et al.*, 2007), y conduce a una

linealización conveniente del problema. En particular, los algoritmos REML tienen dos características interesantes, para nuestros propósitos. La primera es que utilizan las Ecuaciones de Modelos Mixtos de Henderson, que en este caso son

$$\begin{bmatrix} \boldsymbol{1}'\boldsymbol{1} & \boldsymbol{1}'\boldsymbol{W} \\ \boldsymbol{W}'\boldsymbol{1} & \boldsymbol{W}'\boldsymbol{W} + \boldsymbol{A}^{-1}\frac{\sigma_e^2}{\sigma_u^2} \end{bmatrix} \begin{bmatrix} \mu \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{1}'\boldsymbol{z} \\ \boldsymbol{W}'\boldsymbol{z} \end{bmatrix} \qquad [2.1]$$

Este sistema de ecuaciones corresponde al modelo lineal $\boldsymbol{z} = \boldsymbol{1}\mu + \boldsymbol{W}\boldsymbol{u} + \boldsymbol{e}$, donde $\boldsymbol{z}$ es un vector de orden $n_G \times 1$ (n_G = número de individuos con genotipo asignado mediante el proceso de genotipado) que incluye los valores observados del conteo de alelos (0, 1, 2) para un marcador del panel en particular, $\mu = 2p$ es la media de la población para ese marcador, $\boldsymbol{u}$ es un vector de orden $n \times 1$ expandido de modo de incluir a todos los individuos del pedigree (tengan o no genotipo observado, $n > n_G$; Gengler *et al.*, 2007) y $\boldsymbol{W}$ es una matriz de incidencia de orden $n_G \times n$ que relaciona los elementos en $\boldsymbol{u}$ con aquellos en $\boldsymbol{z}$. El hecho de incluir los individuos no genotipados en la formulación permite utilizar toda la matriz $\boldsymbol{A}^{-1}$ (Henderson, 1977), que es rala y puede obtenerse fácilmente utilizando las reglas de Henderson (1976), las cuales son más convenientes computacionalmente con respecto a las ecuaciones de McPeek *et al.* (2004). De la salida final, se obtiene una estimación de p como $\hat{p} = \frac{\hat{\mu}}{2}$. Otra estimación de p, ligeramente diferente a la anterior debido a la maximización numérica, surge a partir de resolver la ecuación $\sigma_u^2 = 2p(1-p)$, i.e. $0.5 \pm \sqrt{1 - 2\hat{\sigma}_u^2}/2$.

2.3.1.3 *Prueba de hipótesis de errores en el genotipado*

La otra característica interesante de REML es que permite calcular la verosimilitud y construir pruebas estadísticas. En nuestro caso existen dos hipótesis: la nula establece que no existen errores de genotipado, y por lo tanto $\sigma_e^2 = 0$ (o $h^2 = 1$) y

la alternativa permite cualquier valor positivo de la varianza del error. Se puede utilizar un test de cociente de verosimilitud (en inglés *likelihood ratio test*, LRT) de modo de rechazar la hipótesis nula, de la siguiente manera. Bajo la hipótesis nula la varianza del error de genotipado es igual a cero, y el estadístico LRT se distribuye asintóticamente como ½χ2(0) + ½χ2(1) (Self y Liang, 1987; Visscher, 2006). El *p*-valor asociado con el valor observado del estadístico LRT puede calcularse asumiendo esta distribución. Estrictamente hablando, para obtener la citada distribución se requiere una función de verosimilitud sobre la base de la normal, pero el LRT es robusto con respecto a las desviaciones de la normalidad (e.g., Almasy y Blangero, 1998).

2.3.1.4 *Implementación*

En la práctica, el método es simple. Para cada marcador, se obtiene una estimación REML de σ_e^2 y de σ_u^2, junto con el valor de máxima verosimilitud – esto es, bajo la hipótesis alternativa. Se realiza también una estimación bajo hipótesis nula igual a $\sigma_e^2 = 0$ (en la práctica, σ_e^2 se fija con un valor muy pequeño). Luego, el *p*-valor del LRT se computa a partir de las dos verosimilitudes, y se establece un umbral de rechazo basado en la distribución asintótica arriba mencionada y el error de tipo I deseado (en este trabajo, 1%). También sugerimos, como un procedimiento menos formal, una inspección de la estimación de la h^2; valores de heredabilidades inferiores a 1 son considerados sospechosos.

2.3.1.5 *Test en ausencia de pedigree*

Si el pedigree no se encuentra disponible pero se considera que la mayoría de los marcadores están *a priori* genotipados correctamente, sugerimos realizar el siguiente procedimiento, el cual no ha sido probado en el presente proyecto:

a) Luego de aplicar los criterios de QC basados en equilibrio Hardy-Weinberg, "call rates" y MAF, construir una matriz de relaciones genómicas ***G*** utilizando los marcadores que pasaron el filtrado, e.g., siguiendo la metodología propuesta por VanRaden (2008).
b) Testear la heredabilidad de cada marcador utilizando estimadores REML tal como se describió anteriormente y la matriz ***G*** (en lugar de la matriz ***A***; este procedimiento es conocido a veces como GREML) para la covarianza del conteo de alelos entre los individuos.
c) Descartar los marcadores rechazados por el test e iterar el procedimiento hasta que no se eliminen más marcadores.

El procedimiento anterior supone que la mayoría de los marcadores tienen el genotipo correctamente asignado.

2.3.2 Pruebas del método

2.3.2.1 *Simulaciones*

Para determinar la sensibilidad (fracción de los marcadores incorrectos que son rechazados) y la especificidad (fracción de los marcadores correctos que *no* son rechazados) de la prueba o test estadístico para la heredabilidad, se simuló una base de datos siguiendo los principios Mendelianos de herencia (es decir, datos sin error) empleando el programa QMSim (Sargolzaei y Schenkel, 2009). La estructura de los datos se asoció con los de un programa de mejoramiento de cerdos, formulado de manera simplificada, con 10 cromosomas autosómicos de igual longitud (160 cM) y 70.000 SNPs. Primero, se alcanzó el equilibrio mutación-deriva en 2.500 generaciones de apareamiento aleatorio en una población de tamaño efectivo igual a 500 y una tasa de mutación igual a 2×10^{-4}, seguido de un "cuello de botella" severo que redujo el tamaño

efectivo poblacional a 75 durante 30 generaciones. Luego se realizó selección al azar durante 5 generaciones, en la cual 20 machos se aparearon con 200 hembras produciendo 2.000 descendientes, con un total de 10.220 animales con información completa de pedigree y 5.000 individuos genotipados (elegidos al azar) con 28.254 marcadores de MAF > 0,01.

Para el test de heredabilidad, el error de tipo I se evaluó como el número de SNPs con *p*-valores por encima del nivel de significancia de 0,01. El error de tipo II se evaluó bajo dos escenarios de errores de genotipado permutando 10% o 5% de los genotipos de cada SNP. La permutación no modificó la MAF ni la condición de equilibrio Hardy-Weinberg de cada marcador, y se realizó al azar para cada SNP. En realidad, un genotipo permutado puede ser reemplazado por el correcto simplemente por azar, por lo que el número de errores efectivos es inferior a los valores de arriba y es función de las frecuencias alélicas, como se explica a continuación. En equilibrio Hardy-Weinberg un genotipo heterocigota tiene una frecuencia igual a $2p(1-p)$, y es permutado por otro genotipo heterocigota con probabilidad de permutación x multiplicada por $2p(1-p)$ (la frecuencia de otro heterocigota). Extendiendo este razonamiento a los tres genotipos posibles, la tasa de error es igual a $x\left[p^2\left(1-p^2\right)+2pq\left(1-2pq\right)+q^2\left(1-q^2\right)\right]$, una función cuártica de las frecuencias alélicas. Por lo tanto, el error real es $0{,}625x$ para una frecuencia alélica igual a 0,5, $0{,}47x$ a través de un espectro de frecuencias uniforme, y toma valores más bajos para distribuciones de las frecuencias en forma de U; en la presente simulación en particular, los errores reales fueron iguales a 0,02 y 0,04 para las tasas de permutación de 5% y 10%, respectivamente.

2.3.2.2 *Base de datos reales*

Se utilizó una base de datos de cerdos de acceso público para la comunidad científica (Cleveland *et al.*, 2012). La base de datos consistía en 3.534 animales de una sola línea genética porcina de un núcleo de la empresa PIC con genotipos para los SNPs incluidos en el chip PorcineSNP60 de Illumina (Ramos *et al.*, 2009) con muy poco control de calidad y pedigree registrando dos generaciones ancestrales de los animales genotipados (N = 6.473). En la práctica, este conjunto de datos debería someterse a controles Mendelianos de los pares padre-hijo de modo de eliminar aquellos animales inconsistentes; hemos preferido no hacerlo, con el fin de utilizar la base de datos tal como está. Se emplearon para el estudio un total de 50.433 SNPs, luego de filtrar los genotipos por MAF (<0,01) y por "call rate" de los SNPs (<90%), excluyéndose además los SNPs ubicados en los cromosomas sexuales. La razón para excluir los SNPs con MAF <0,01 es que la maximización numérica de REML es poco confiable en ese caso. A los efectos comparativos, el mismo análisis se llevó a cabo en dos escenarios extremos permutando aleatoriamente la mitad o todos los genotipos para cada SNP en la base de datos.

2.3.2.3 *Análisis estadístico*

Se calculó la heredabilidad y el estadístico LRT para testear la hipótesis de varianza nula de los errores de genotipado de cada SNP en la base. El valor máximo para el logaritmo de las funciones de verosimilitud restringidas en el modelo completo (hipótesis alternativa, sin información *a priori* para los valores de σ_u^2 y σ_e^2) y en el modelo reducido (hipótesis nula, asumiendo $\sigma_e^2 = 0$) fueron calculados mediante el programa remlf90 (Misztal *et al.*, 2002). Se calculó la heredabilidad del conteo de alelos para cada SNP, bajo el modelo completo. Las estimaciones pueden realizarse en

paralelo de ser necesario. Debido a que el software utiliza las ecuaciones del modelo mixto de Henderson, el modelo reducido tuvo que aproximarse especificando un valor muy pequeño para la varianza residual: $\sigma_e^2 =$ 0,0001. La ventaja de calcular la verosimilitud de esta manera es el poder utilizar un software estándar para la estimación de componentes de varianza vía REML. También se emplearon los filtros de control de calidad sobre la base de los errores Mendelianos incluidos en el programa preGSf90 (Aguilar *et al.*, 2014): se rechazaba un marcador si sus genotipos mostraban incoherencias Mendelianas en más del 1% de los pares padre-hijo o tríos.

Los programas auxiliares ("scripts") completos para el análisis de control de calidad de los SNPs, la base de datos reales de la PIC, y las tres bases de datos simuladas se encuentran disponibles en el sitio http://genoweb.toulouse.inra.fr/~alegarra/qualitycontrol.tar.gz .

2.4 Resultados

2.4.1 Simulaciones

La Figura 2.1A (2.1B) muestra las estimaciones de la heredabilidad del conteo de alelos (*p*-valor del estadístico LRT) en función de la frecuencia alélica menor de cada marcador para el conjunto de datos simulados sin error. La mayoría de las estimaciones de heredabilidad son muy cercanas a 1, incluso para muy baja MAF, si bien hay una tendencia a que los marcadores con muy baja MAF tienen heredabilidades más bajas (por ejemplo, pueden fijarse por deriva). Estableciendo un umbral en un valor nominal para el error de tipo I igual a 0,01, la sensibilidad resultó igual a 0,99 (0,91) cuando el 10% (5%) de los genotipos simulados eran incorrectos (permutados al azar). La especificidad fue igual a 0,95 (i.e., 5% de los SNPs correctamente asignados fueron rechazados). Por otra parte, se delimitaron las estimaciones de heredabilidad para el

rechazo. La Figura 2.1C muestra los errores empíricos de tipo I (ó 1−especificidad) o tipo II (ó 1−sensibilidad) *versus* umbrales posibles de rechazo, sobre la base de la heredabilidad. Para ejemplificar, los marcadores fueron rechazados si la estimación de heredabilidad fue inferior a 0,975; esto resulta en una especificidad de 0,96 (4% de los marcadores correctos son rechazados) y una sensibilidad de 0,99 (para el 10% de los datos permutados; sólo el 1% de todos los marcadores incorrectos son aceptados). En cambio, la elección de un límite inferior de heredabilidad igual a 0,90 resultó en sólo el 0,04% de los marcadores incorrectamente rechazados, pero tanto como el 6,5% de los marcadores incorrectamente aceptados. Estas figuras son distintas con el nivel de calidad de los datos, y el escenario en el que se permutó un 5% de los genotipos arrojó un valor de error de tipo II mayor. La comprobación de errores Mendelianos con el programa preGSf90 se comportó peor que el método propuesto en este capítulo, con un valor para el error de tipo I igual a 0 (como era esperable) y un valor de sensibilidad igual a 0,84 y 0,54, para los escenarios en el que se permutó 10% y 5% de los genotipos, respectivamente.

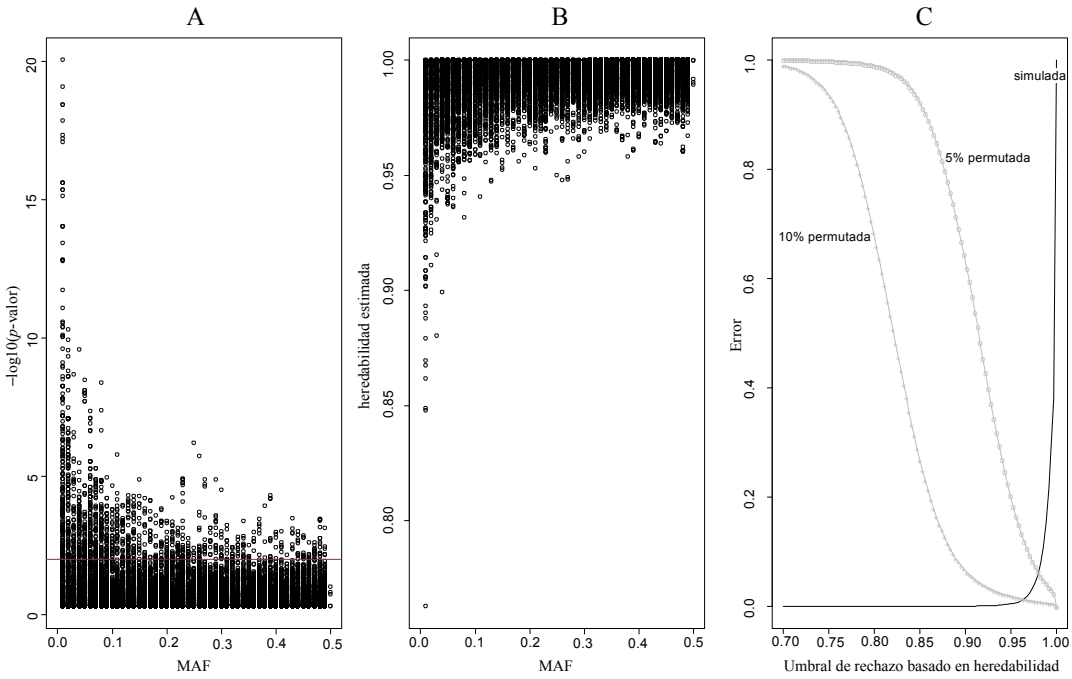


Figura 2.1. Resultados de las simulaciones. **A.** La cifra de menos el logaritmo del *p*-valor de la hipótesis nula *vs.* la MAF del marcador para la base de datos simulada sin error. **B.** Las estimaciones de la heredabilidad del conteo de alelos del marcador *vs.* la MAF del marcador para la base de datos simulada sin error. La línea horizontal en (A) es el umbral de rechazo del 1%. **C.** El error de tipo I (línea continua) y el error de tipo II en función del umbral de rechazo basado en la heredabilidad del conteo de alelos del marcador, para la base de datos simulada sin error (línea continua) o con 5% (círculos) y 10% (cruces) de genotipos permutados.

2.4.2 Base de datos reales

La Figura 2.2A muestra los gráficos de caja (“boxplots”) de las estimaciones de heredabilidad del conteo de alelos de los SNPs, para la base de datos original, la base de datos permutada al 50% y la base completamente permutada. La heredabilidad media para la base de datos original fue igual a 0,99 y el 75% de los SNPs tuvo

heredabilidades por encima de dicho valor, aunque algunos de los marcadores se desviaron mucho respecto de 1. Cuando se permutó aleatoriamente el 50% de los genotipos de cada SNP, la heredabilidad tomó valores entre 0,02 hasta 0,84, la heredabilidad media fue igual a 0,25 y 75% de las estimaciones estuvieron por debajo de 0,27. En la base de datos totalmente permutada, todas las heredabilidades fueron inferiores a 0,07. Nótese que las cajas se desplazan hacia arriba a medida que la calidad general de la base de datos mejora. Al probar la hipótesis nula de error de genotipado igual a cero con un nivel de significancia del 1%, la hipótesis nula fue rechazada en 8% (N = 4.099) de los SNPs de la base de datos original, mientras que todos los p-valores estuvieron por debajo de 10^{-12} en la base de datos permutada al 50%, y por debajo de 10^{-93} en la base completamente permutada. Este último resultado ejemplifica cómo un procedimiento de genotipificación que es en gran medida incorrecto para gran parte de los individuos (> 50%) puede detectar fácilmente errores utilizando nuestra metodología. Los 4.099 marcadores que no pasaron el test en la base de datos original deben ser declarados como incorrectamente genotipados y sus genotipos no deben utilizarse en análisis posteriores. La Figura 2.2B ilustra la relación entre las estimaciones de heredabilidad y los p-valores del estadístico LRT cuando se utiliza REML para estimar los componentes de varianza en la base de datos original. Se puede observar que los SNPs rechazados presentaron los valores más bajos de heredabilidad, si bien el rango fue importante ($0,13 < h^2 < 0,97$, para SNPs con p-valor < 0,01). Por ejemplo, un marcador con (muy) baja MAF, puede resultar en una estimación alta de heredabilidad pero el LRT puede ser poco concluyente, porque hay muy poca información en los datos. Sin embargo y en general, se obtendrán resultados similares utilizando el LRT formal así como estimaciones de heredabilidad. Estas últimas tienen un menor sustento teórico estadístico; sin embargo, resultan fáciles de interpretar por

parte de los genetistas cuantitativos, y pueden agilizar fácilmente el análisis como se discutirá más adelante. En esta base de datos, el control de errores Mendelianos que realiza el programa preGSf90 detectó sólo un marcador incorrecto utilizando un valor umbral de tolerancia para inconsistencias Mendelianas de 1%, 577 utilizando 0,1% de tolerancia, y 2.019 con tolerancia cero. De los 577 marcadores con inconsistencias Mendelianas mayores a 0,1%, 388 estaban incluidos dentro de los 4.099 rechazados por el test de heredabilidad. Los 577 marcadores con inconsistencias Mendelianas mayores a 0,1% mostraron estimaciones de heredabilidad más bajas (0,92 en promedio) que aquellos que no fueron rechazados (0,99).

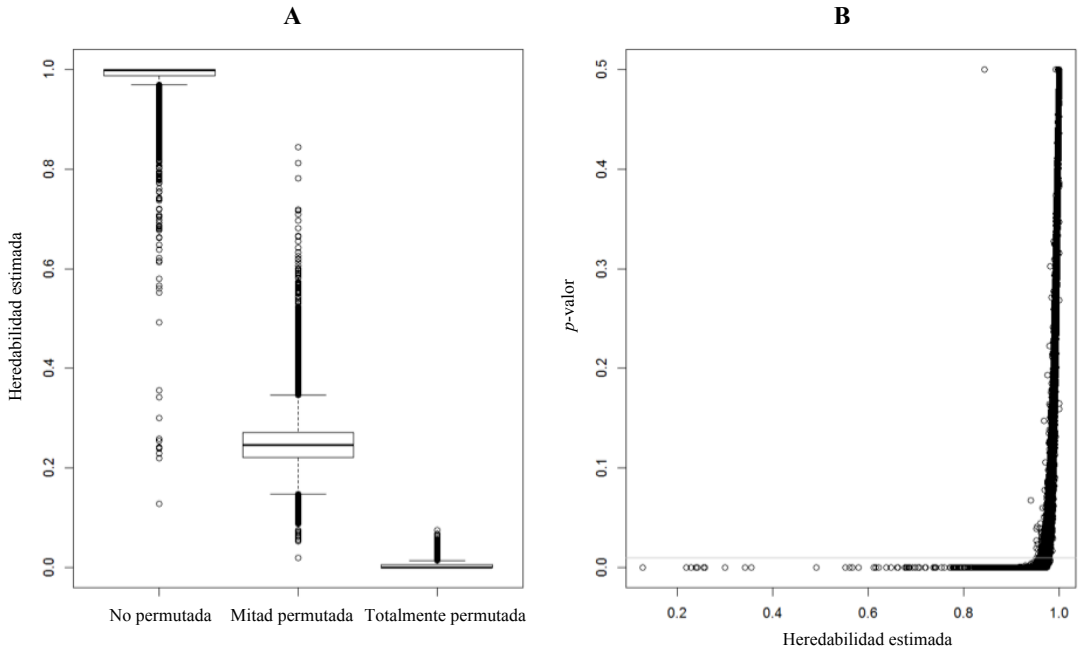


Figura 2.2. Resultados con la base de datos real de la PIC. **A.** Estimaciones de la heredabilidad del conteo de alelos en la base de datos original, con la mitad de los genotipos permutados, o con todos los genotipos permutados. **B.** Estimaciones de la heredabilidad del conteo de alelos en la base de datos original *vs.* los *p*-valores del test del cociente de verosimilitud.

2.5 Discusión

El método original aquí presentado tuvo éxito en identificar marcadores de baja calidad en un pedigree complejo. El método evita la búsqueda de estructuras de pedigree como pares padre-hijo. Si bien el análisis de cada marcador individual toma unos minutos, puede ser paralelizado porque cada marcador es independiente. El procedimiento proporciona un test estadístico, y por lo tanto sus propiedades son conocidas, mientras que para otros procedimientos los umbrales de corte son en gran medida arbitrarios. Como demuestran los resultados presentados en este capítulo, otras pruebas como el análisis de pares padre-hijo tienen una alta especificidad, pero no necesariamente una alta sensibilidad, por ejemplo, si no existen muchos pares padre-hijo. Asimismo, el método aquí propuesto tiene en cuenta las distorsiones en la segregación y parece ser robusto a la presencia de un alelo en baja frecuencia (digamos, superior a 0,05). Sin embargo, para una frecuencia alélica muy baja, las estimaciones de heredabilidad y LRT tienden a ser poco fiables (Figuras 2.1A y 2.1B) aunque no se ha probado la metodología con valores de MAF inferiores a 0,01. El procedimiento propuesto no puede corregir errores Mendelianos en los marcadores que no son rechazados, y en este caso es necesaria la utilización de comparaciones padre-hijo.

El método aquí presentado asume una única población en equilibrio Hardy-Weinberg, hipótesis que parece ser escasamente restrictiva dado que la base de datos simulada incluía selección. Aún cuando la población tuviera diferentes orígenes genéticos, $\sigma_e^2 = 0$. Sin embargo, la hipótesis de media y varianza común no sería sustentable. Un método aproximado consiste en ajustar por los diferentes orígenes utilizando un modelo con grupos genéticos (Quaas, 1988), i.e. permitiendo diferentes frecuencias alélicas en las distintas poblaciones base. En este caso se asume el mismo valor para $\sigma_u^2 = 2p\left(1-p\right)$ entre las distintas poblaciones, hecho verdadero cuando las

frecuencias son similares entre las distintas poblaciones, pero potencialmente falso en caso que existieran diferencias importantes de frecuencias.

Mientras que la estimación REML de la heredabilidad para un solo SNP probablemente sea rápida, el número total de cálculos para un gran número de SNPs puede tomar días. En REML-EM, las operaciones más costosas en una iteración incluyen: a) armado de las ecuaciones del modelo mixto, b) cálculo de las soluciones, y c) cálculo de las trazas. Sin embargo, las ecuaciones de modelos mixtos son las mismas en todos los marcadores para un valor dado de heredabilidad. Por lo tanto, se pueden realizar ahorros importantes de tiempo computacional cuando las factorizaciones y las trazas se calculan previamente para distintos valores de heredabilidad. Si el propósito del control de calidad es seleccionar SNPs con valores de $h^2 > 0{,}98$, se necesitarían sólo tres conjuntos de matrices para $h^2 = 0{,}99$, 0,98 y 0,97.

El procedimiento aquí propuesto no puede identificar errores de pedigree (i.e., el etiquetado incorrecto de muestras de ADN). En este caso, los errores ocurren a través de los marcadores en un individuo, en lugar de ocurrir a través de los individuos para un marcador. Las discordancias padre-hijo pueden marcar tal error si muchos marcadores no siguen las leyes de Mendel para un par padre-hijo determinado. Existen procedimientos para asignar padres (Wiggans, 2009; Hayes, 2011; VanRaden *et al.*, 2013). Sin embargo, no existe aún un procedimiento general para identificar y corregir los errores de pedigree. Un manera práctica de control consiste en comparar relaciones genómicas (VanRaden, 2008) con las relaciones aditivas e inspeccionar las diferencias, las cuales dependen de la misma relación de parentesco y la arquitectura del genoma. Una descripción detallada de estas diferencias se puede encontrar en Wang *et al.* (2014).

Un caso particular es el empleo de genotipos de diferentes paneles o chips, posiblemente con diferente química, por ejemplo, los chips de 50K y 3K en bovinos

(Wiggans *et al.*, 2012). Estos autores encontraron que algunos marcadores se leían correctamente utilizando un panel, pero no con el otro. Con nuestro método dichas diferencias se observarían porque la heredabilidad estimada disminuiría al incluir los genotipos del chip defectuoso, sea solo o combinado con el otro panel. Esto se aplica también a las muestras genotipadas conjuntamente. Por ejemplo, cuando aparece un gran número de muestras individuales con ADN en mal estado, la incorporación de los genotipos de dicha muestra disminuirá la estimación de heredabilidad.

El método propuesto obtiene, como un subproducto, las estimaciones de las frecuencias alélicas de los marcadores en la población base. Esto puede resultar de utilidad cuando se tienen familias grandes con muy pocos animales genotipados y las frecuencias alélicas de la población base no están bien representadas por las frecuencias alélicas de los animales genotipados de la generación fundadora del pedigree. De hecho, para calcular las relaciones genómicas en SG, se deberían utilizar las frecuencias alélicas de una población base sin selección en lugar de las que resultan de los procesos de selección o endogamia. Una población base anticipada o adelantada en el tiempo puede dar lugar a relaciones genómicas de mayor o menor magnitud y a un mayor o menor nivel de consanguinidad (VanRaden, 2008).

En nuestra experiencia, el procedimiento resulta más útil cuando se trata con bases de datos nuevas y completas, en particular, a partir de estudios experimentales. Las evaluaciones genéticas regulares, como en el ganado lechero, mantienen un mejor seguimiento de las muestras de ADN y, debido a la abundancia de tríos y pares padre-hijo, es sencillo detectar marcadores de mala calidad (Wiggans *et al.*, 2009, 2012).

En resumen, en el presente capítulo se propone un procedimiento práctico para realizar un análisis de QC, de modo de identificar SNPs de baja calidad en un número grande de individuos. El filtro QC propuesto es, en esencia, una estimación de la

heredabilidad del conteo de alelos en los SNPs, donde cualquier desviación de 1 es sospechosa, y el *p*-valor sirve para testear la hipótesis nula de "ausencia de errores de genotipado". Este procedimiento de QC puede considerar todos los individuos genotipados y su pedigree de manera conjunta y utiliza procedimientos estándar de pruebas de hipótesis. Debería ser utilizado como un complemento de los procedimientos estándar de QC, y muy posiblemente después de éstos.

CAPÍTULO 3

Medida de identidad por descendencia condicional al pedigree y a la información de un panel denso de marcadores

Medida de identidad por descendencia condicional al pedigree y a la información de un panel denso de marcadores

3.1 Introducción

La estimación de parámetros genéticos, como la heredabilidad y las correlaciones genéticas entre caracteres y su aplicación en los programas de mejoramiento genético animal, se fundamentan en base al grado de parecido entre parientes de origen genético. La covarianza entre parientes depende del número de alelos idénticos por descendencia (IBD) compartidos para cada uno de los loci que afectan el carácter de interés. Se dice que dos alelos presentes en un mismo locus - que son muestreados en dos individuos - son IBD cuando ambos genes son copias de un mismo gen presente en un ancestro en común a los dos individuos (Malécot, 1948). Para una relación de parentesco en particular, se puede obtener la distribución de probabilidad de compartir 0, 1 ó 2 pares de alelos IBD en un locus dentro del genoma. A partir de dicha distribución, se puede calcular el valor esperado de la proporción de alelos IBD entre dos individuos, o relación aditiva (Bulmer, 1985, pág. 42). Este valor constituye cada uno de los elementos de la matriz de relaciones $\boldsymbol{A}$ en el modelo animal (MA) de evaluación genética (Henderson, 1984). Ahora bien, la proporción de alelos IBD es una medida de identidad genética respecto de un locus y, por lo tanto, sumamente variable. Esta variabilidad se puede visualizar comparando la relación padre-hijo con la de un par de hermanos enteros. Mientras que la proporción de alelos IBD entre un padre y su hijo es *exactamente* 0,5, entre hermanos enteros es *en promedio* 0,5; la varianza de esta medida es igual a 0 para el par padre-hijo y 0,125 para hermanos enteros. Asimismo, mientras que para las relaciones tío-sobrino, abuelo-nieto, medio

hermanos y doble primos hermanos, la esperanza de la proporción de alelos IBD es igual a 0,25, el desvío estándar es al menos 25% (Guo, 1996).

La noción de IBD se extiende a múltiples loci, e inclusive al genoma autosómico completo. De hecho, el genoma no se transmite "punto a punto" sino en segmentos (Guo, 1995). En tal sentido, dos haplotipos correspondientes a un segmento del genoma son IBD si son copias de un mismo haplotipo en un ancestro común. Los segmentos IBD se recombinan por el crossing-over durante la meiosis. Por lo tanto, su longitud depende del número de generaciones desde el ancestro común; los segmentos IBD serán más pequeños cuanto más atrás en el tiempo haya vivido el ancestro. Sin embargo, los eventos de recombinación durante la meiosis ocurren relativamente con poca frecuencia, típicamente 1-3 por cada 100 centimorgans (cM) (Visscher, 2009). En términos estadísticos, esto implica que loci adyacentes tienden a no ser independientes respecto de sus estados IBD, violando el supuesto de equilibrio Hardy-Weinberg. En consecuencia, este proceso agrega complejidad al cómputo de las probabilidades de IBD cuando se considera al genoma como un todo. Es posible estimar la proporción "realizada", o verdadera, de alelos IBD entre dos individuos condicional a la información de marcadores moleculares distribuidos a lo largo del genoma, y al pedigree. Los métodos tradicionales de estimación de las probabilidades de IBD se basan en modelos Markov "ocultos" (HMM, *Hidden Markov Models*, Rabiner, 1989) que incorporan dependencia estadística entre los estados de IBD para loci marcadores vecinos pero asumen equilibrio gamético (LE). Dos algoritmos ampliamente utilizados son el de Elston y Stewart (1971) y el de Lander y Green (1987).

El algoritmo de Elston y Stewart (1971) es apropiado para pedigrees grandes, pero retarda el cálculo de manera exponencial con el número de marcadores. Por su parte, el algoritmo de Lander y Green (1987) es apropiado cuando el número de

marcadores es grande pero el esfuerzo de cálculo es exponencial en relación con el tamaño del pedigree. La complejidad de este método aumenta cuando existen individuos en el pedigree sin genotipar. Este algoritmo es implementado en el software Merlin (Abecasis *et al.*, 2002), uno de los más populares para estimar probabilidades de IBD en pedigrees pequeños. Como se mencionó anteriormente, estos métodos asumen que los haplotipos de los fundadores provienen de una población en equilibrio gamético (LE); sin embargo, para un panel denso de marcadores este supuesto no se cumple. El desequilibrio gamético o de ligamiento (LD) se torna significativo cuando los SNPs están separados a distancias menores a 10^5 bp (pares de bases). En humanos, dichas distancias equivalen – aproximadamente – a un panel de 30K (Thompson, 2008). Estos métodos son adecuados cuando la densidad de marcadores es baja; si se aplican a paneles de marcadores densos con altos niveles de LD, tienden a sobreestimar la proporción de IBD compartida. Abecasis y Wigginton (2005) modificaron Merlin para incorporar el LD al organizar los marcadores en clusters (representando bloques de haplotipos) y estimando las frecuencias poblacionales de los haplotipos en cada cluster. Sin embargo, el método asume ausencia de recombinación dentro de cada cluster y LE entre clusters (que sean independientes). Además, posee un número elevado de parámetros a estimar y tiene requerimientos de tiempo y memoria supralineales con respecto al número de marcadores. Por su parte, Keith *et al.* (2008) incorporaron el LD a la estimación de IBD modelando los haplotipos de los fundadores como una cadena de Markov de primer orden. Si bien éste método es de menor complejidad que el del software Merlin, sus autores admiten que modelar el LD con clusters aumenta la precisión de las estimaciones. El software PLINK (Purcell *et al.*, 2007) elude el problema del LD eliminando marcadores para luego asumir LE; sin embargo, se reduce la información disponible, disminuyendo la potencia para detectar segmentos cortos

IBD. Albrechtsen *et al.* (2009) extendieron el enfoque PLINK e incorporaron LD utilizando probabilidades de haplotipos al tomar marcadores de a pares. Es probable que este enfoque no corrija adecuadamente el LD en regiones donde el fenómeno de ausencia de independencia es intenso (Browning y Browning, 2010). Meuwissen y Goddard (2010) desarrollaron un algoritmo de imputación de genotipos y determinación de fase, y utilizaron la información del LD para imputar los genotipos faltantes, al mismo tiempo que incorporaban el ligamiento generado – únicamente – por la estructura familiar, de modo de estimar las probabilidades IBD entre haplotipos. Más recientemente, Li *et al.* (2010) propusieron un nuevo enfoque para inferir las probabilidades de IBD en pedigrees humanos de gran tamaño con muchos individuos sin genotipar utilizando una alta densidad de SNPs. El enfoque de Li *et al.* (2010) modela directamente la relación entre un par de individuos genotipados, sin necesidad de enumerar cada uno de los posibles genotipos y patrones de herencia de sus ancestros no genotipados. Este método tiene en cuenta el LD generado por ligamiento, así como aquel producido por las relaciones ancestrales que van más allá del pedigree conocido.

Los avances tecnológicos recientes y la reducción de los costos, posibilitan el genotipado con chips de alta densidad a gran escala. Esto permite la detección de segmentos IBD cada vez más pequeños. En esta tesis, se busca detectar el estatus de IBD dentro de la genealogía conocida, es decir asumiendo que todos los alelos ancestrales de los genes que afectan el carácter de interés son diferentes en la población base de la genealogía. Para maximizar la capacidad de detectar tales segmentos, es importante la utilización de herramientas estadísticas adecuadas, lo cual implica contar con algoritmos eficientes que tomen en cuenta el LD entre marcadores.

3.2 Objetivo

El objetivo de este capítulo es proponer y describir un método para estimar la proporción de genoma compartido IBD por un par de individuos genotipados dentro de un pedigree de animales, condicional a las relaciones de parentesco y a la información de un chip de alta densidad de SNPs, considerando al mismo tiempo el LD entre marcadores. Para este proceso de estimación se tiene en cuenta a los animales del pedigree sin genotipar. Dicha estimación representa un coeficiente de relación aditiva genómico y constituirá cada uno de los elementos de la matriz de relaciones genómicas entre animales genotipados. La performance del método será evaluada en los capítulos que siguen, en términos de la precisión de las relaciones estimadas (Capítulo 4) y de la exactitud de la predicción de los valores de cría genómicos (Capítulo 5).

3.3 Metodología

Para calcular la proporción del genoma compartido IBD en un par de individuos se debería conocer el estado de IBD en cada posición del genoma, entendiendo por posición el lugar físico que ocupa cada nucleótido en la cadena de ADN. Un modo de estimar dicha proporción consiste en evaluar el estado IBD en cada locus marcador, para luego promediarlo con respecto al total de marcadores. Ahora bien, teniendo en cuenta la información que brinda actualmente un panel de SNPs de alta densidad (ausencia de fase conocida, presencia de LD entre marcadores) y el pedigree, no siempre es posible conocer con certeza el estado de IBD en un marcador. Por lo tanto, utilizaremos el algoritmo de Li *et al.* (2010) para estimar las probabilidades de compartir 0, 1 ó 2 pares IBD en cada uno de los loci marcadores, entre un par de individuos relacionados. A tal efecto, los citados autores emplean un HMM.

A continuación (sección 3.3.1), se dará una definición formal del proceso HMM. Para una descripción más detallada, véase Rabiner (1989) o Durbin *et al.* (1998). En la sección 3.3.2 se describirá en detalle el marco teórico sobre el que se sustenta el algoritmo de Li *et al.* (2010). A partir de la estimación del estado de IBD en cada locus, obtenida mediante el citado algoritmo, se sugiere una fórmula para estimar la verdadera proporción del genoma compartido IBD por un par de individuos genotipados en un pedigree animal, condicional a las relaciones de parentesco y a la información de un chip de alta densidad de SNPs (sección 3.3.3). Ésta es una medida de identidad genética sobre todo el genoma y constituirá cada elemento de la matriz de relaciones genómicas que será empleada y evaluada en los Capítulos 4 y 5.

3.3.1 Modelo de Markov oculto

Los algoritmos Markov ocultos (*Hidden Markov Model*, HMM) son utilizados para modelar el proceso estocástico por el cual se genera una secuencia de observaciones. El modelo comienza en un estado elegido según un vector de probabilidades de inicio; éste "emite" una observación para luego transitar a un nuevo estado que emite nuevamente una observación, y así sucesivamente hasta alcanzar el final de la secuencia. Como en general se conoce la secuencia de observaciones pero no la de estados, se dice que esta última se encuentra "oculta"; de ahí la denominación de modelo Markov oculto.

Consideraremos aquí el modelo básico, o HMM discreto de primer orden, y que es definido por los siguientes elementos:

- $S = \{S_1, S_2, \ldots, S_N\}$, el conjunto de N estados ocultos del modelo. La variable aleatoria q_i es el estado que realmente ocurrió (*realised*) en la posición i;
- $V = \{v_1, v_2, \ldots, v_M\}$, el conjunto de M símbolos observables de cada estado.

- $\boldsymbol{A} = \{a_{jk}\}$, la matriz de orden $N \times N$ de probabilidades de transición entre estados, donde $a_{jk} = P(q_{i+1} = S_k \mid q_i = S_j)$, $1 \leq j, k \leq N$;
- $\boldsymbol{B} = \{b_k(l)\}$, la matriz $N \times M$ de probabilidades de emisión de símbolos observables en el estado k, donde $b_k(l) = P(v_l \text{ en } i \mid q_i = S_k)$, $1 \leq k \leq N$ y $1 \leq l \leq M$;
- $\boldsymbol{\pi} = \{\pi_j\}$, el vector de probabilidades del estado inicial q_1, donde $\pi_j = P(q_1 = S_j)$.

En un HMM, cada secuencia de observaciones es una cadena de T símbolos, es decir, una sucesión $\{O_i\}$, donde cada O_i es uno de los posibles símbolos de V, $i \in \{1, 2, \ldots, T\}$ y T es el número de observaciones en la secuencia. A su vez, existe una sucesión oculta de estados $\{q_i\}$ correspondiente, donde cada q_i puede tomar uno de los posibles estados S.

Existen dos relaciones de independencia condicional para las sucesiones de estados, $Q = q_1\, q_2 \ldots q_T$, y de observaciones, $O = O_1\, O_2 \ldots O_T$. En primer lugar, dado $q_{i\text{-}1}$, se tiene que q_i es independiente de todos los estados anteriores:

$$P\left(q_i = S_k \,\middle|\, q_{i-1} = S_j, q_{i-2} = S_h, \ldots\right) = P\left(q_i = S_k \,\middle|\, q_{i-1} = S_j\right) = a_{jk} \qquad [3.1]$$

Esta relación indica que la sucesión de estados constituye una cadena de Markov de primer orden. Asimismo, dado q_i, la probabilidad de observar un determinado símbolo en la posición i, O_i, es independiente de los símbolos emitidos en las demás posiciones.

$$P(O|Q) = \prod_{i=1}^{T} P(O_i|q_i) \qquad [3.2]$$

En la Figura 3.1 se representa un HMM de primer orden. Los círculos y las flechas horizontales determinan la cadena de Markov; estas flechas indican la dependencia entre estados consecutivos. La ausencia de flechas entre los cuadrados refleja las relaciones de independencia condicional entre las observaciones.

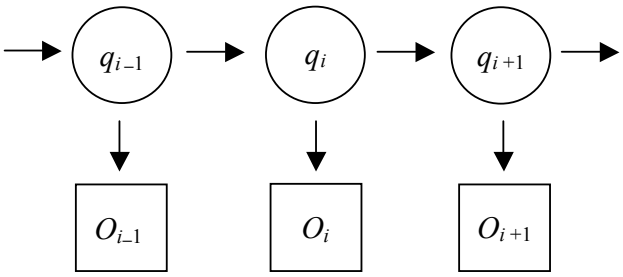


Figura 3.1. Representación de un HMM de primer orden. La ausencia de flecha indica independencia condicional.

3.3.2 Algoritmo de Li *et al.* (2010)

El algoritmo de Li *et al.* (2010) estima el estado IBD en cada uno de los loci marcadores del genoma, para un par de individuos relacionados. Se propone un HMM donde los símbolos observables son el número de pares de alelos en un locus que son IBS (idénticos por estado) y compartidos por ambos individuos. Los estados ocultos son el número de pares de alelos IBD (es decir, copias del mismo gen ancestral) compartidos por los dos individuos. Para derivar su modelo, Li *et al.* (2010) introdujeron primero el concepto de "grafo de descendencia", y definieron una "función generadora de herencia" entre un par de alelos presente en un "grafo de descendencia". Luego elaboraron un HMM para un par de alelos cuyas probabilidades de transición se fueron obtenidas teóricamente a partir de la "función generadora de herencia". Con este modelo básico de referencia, construyeron el HMM para un par de individuos sobre el cual se sustenta finalmente el algoritmo.

3.3.2.1 *"Grafo de descendencia" y "función generadora de herencia"*

Los conceptos que se describen a continuación provienen de la teoría de grafos. Un "grafo de descendencia" (DG, *descent graph*, Sobel y Lange, 1996) es una

estructura definida para un locus y un pedigree, y consiste en el alelo paterno y materno de cada individuo actuando como “nodos” y un enlace entre cada par padre-hijo como “aristas”. Cada arista especifica cuál de los dos alelos de un padre es transmitido a su hijo. Un DG ilustra un posible patrón de herencia dentro de un pedigree (Figura 3.2B).

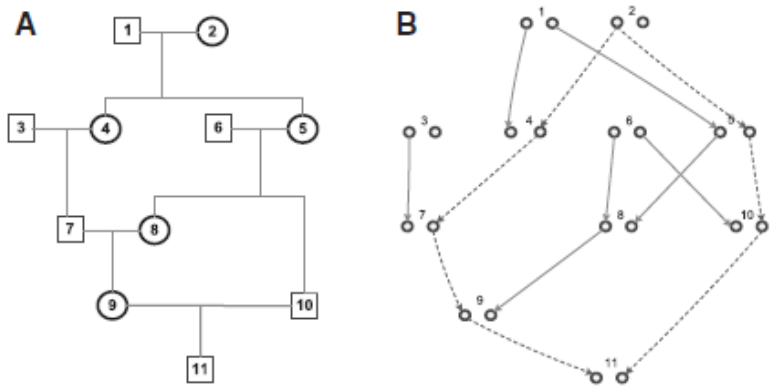


Figura 3.2. A. Estructura de un pedigree. **B.** Un posible DG. Tomado de Li *et al.* (2010).

Para dos nodos (i.e. dos alelos) *a* y *b* en un DG, un “sendero de herencia”, indicado como $p^{a,b}$, es un camino que conecta *a* y *b*. Así, dos alelos son IBD si y sólo si existe un sendero de herencia entre ellos. Para dos alelos cualquiera existe, *a lo sumo*, un sendero de herencia por cada DG. Por ejemplo, si consideramos un locus cualquiera de dos medio hermanos paternos, existen cuatro DGs posibles de los cuales sólo dos resultan en senderos de herencia que conectan los alelos *a* y *b*:

Li *et al.* (2010) definen la “función generadora de herencia” (IGF, *inheritance-generating function*) para dos alelos *a* y *b* en un pedigree como

$$\theta_{a,b}(h) = \sum_{l=0}^{L} \lambda_l \, 0.5^{\,l} \qquad [3.3]$$

donde λ_l es el número de posibles senderos de herencia con longitud *l* entre *a* y *b*. La función describe la probabilidad de transmitir el alelo paterno o el materno. Un sendero de herencia con longitud *l* involucra *l* gametogénesis que conectan *a* con *b*. Por ejemplo, sean *a* y *b* los alelos paterno y materno del individuo 11 en la Figura 3.2A. Teniendo en cuenta el pedigree, existen 8 senderos de herencia posibles que unen *a* con *b*: 4 de ellos involucran 5 meiosis (Figura 3.3B), mientras los otros 4 involucran 7 meiosis (Figura 3.3A). El resultado de cada meiosis es indicado por una flecha que une dos alelos pertenecientes a cada uno de los dos individuos involucrados, que a su vez se encuentran en dos generaciones sucesivas. En el ejemplo, tenemos que $\lambda_0 = \ldots = \lambda_4 = 0$, $\lambda_5 = 4$, $\lambda_6 = 0$ y $\lambda_7 = 4$. Por lo tanto la IGF para *a* y *b* es $\theta_{a,b} = 4\,(0.5)^5 + 4\,(0.5)^7$. Existe sólo un número finito de DGs en cada pedigree; por lo tanto, existe sólo un número finito de senderos de herencia entre dos alelos y la sumatoria tiene un número finito de términos. De este modo, la IGF especifica el número de senderos de cualquier longitud sobre todos los posibles DGs de un pedigree. Eventualmente, la IGF es la probabilidad de que *a* y *b* sean IBD, es decir: $\theta_{a,b} = P(a \equiv b)$. Así por ejemplo en el caso anterior es $\theta_{a,b} = 4\left(\frac{1}{2}\right)^5 + 4\left(\frac{1}{2}\right)^7 = \frac{1}{8} + \frac{1}{32} = \frac{5}{32}$. Esta función se utiliza en la derivación de la distribución de probabilidades de transición del HMM. Valiéndose de la estructura del pedigree, Li *et al.* (2010) obtuvieron una fórmula recursiva para calcular eficientemente las IGF.

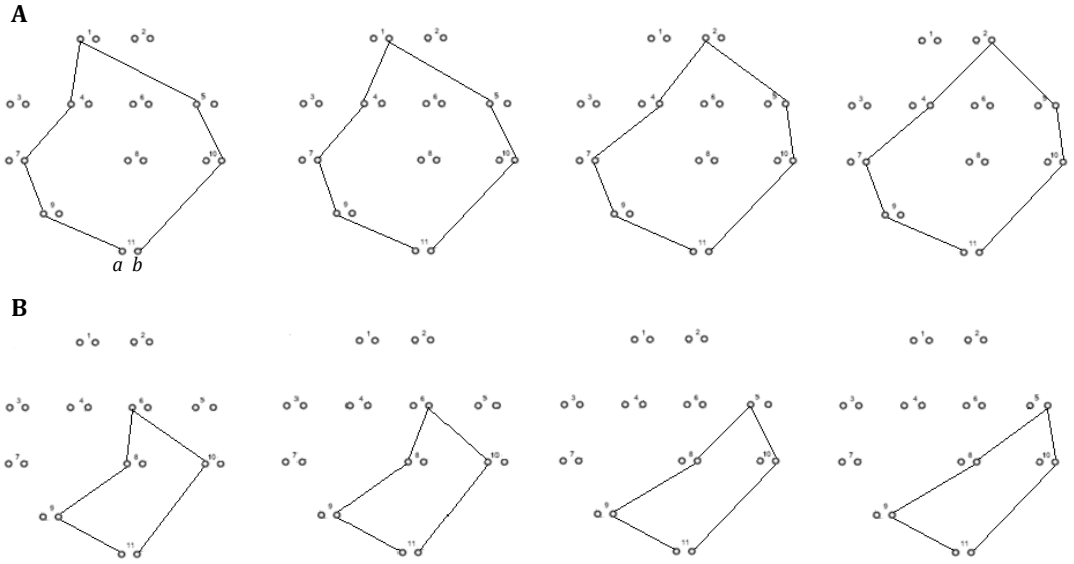


Figura 3.3. A. Posibles senderos de herencia de longitud 7 entre los alelos *a* y *b* del individuo 11. **B.** Posibles senderos de herencia de longitud 5 entre los alelos *a* y *b*.

3.3.2.2 *HMM para un par de alelos*

Imaginemos un par de haplotipos como una cadena de Markov de primer orden, donde cada posición representa la ubicación de un SNP en el genoma. Un par de alelos *a* y *b*, asociados al SNP ubicado en la *i*-ésima posición (locus), tiene tres estados ocultos posibles: IBD, No-IBD y "background IBD" (Bg-IBD) o IBD a nivel poblacional. El estado Bg-IBD describe el grado de parentesco histórico oculto entre individuos, más allá de la relación de parentesco que se observa a partir del pedigree. Esta relación entre individuos "más allá del pedigree" es desconocida y se utiliza para captar el LD entre SNPs. La causa de este parecido adicional entre individuos se encuentra en el parentesco ancestral que va más allá de las relaciones conocidas, y se debe a que tanto la población como el número de alelos iniciales son finitos. No separar el Bg-IBD del IBD que surge exclusivamente del pedigree daría lugar a una inferencia sesgada del verdadero estado de IBD.

Las transiciones posibles entre estados vecinos se ilustran en la Figura 3.4 a través de flechas. La probabilidad de transición del estado IBD a sí mismo se denota

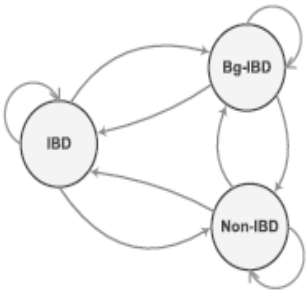


Figura 3.4. Estructura del HMM para un par de alelos, *a* y *b*. Tomado de Li *et al.* (2010).

$\psi(i, i+1) \triangleq P\left(a_{i+1} \equiv b_{i+1} \middle| a_i \equiv b_i\right)$ y se asocia al evento de que dos alelos a_{i+1} y b_{i+1} en el locus $i + 1$ sean IBD, condicional a que dos alelos a_i y b_i en el locus i de los mismos haplotipos sean IBD. Se asume que ψ (i, i +1) es distinta de cero sí y sólo si el sendero de herencia ocurrido en el locus $i + 1$ es el mismo que el que ocurrió en el locus i, lo cual significa que no hay eventos de recombinación a lo largo del camino de herencia entre estos dos loci. Este supuesto ignora la posibilidad de que múltiples eventos de recombinación coincidentes entre estos dos loci resulten en el mismo estado oculto; sin embargo, dicha probabilidad es extremadamente pequeña dada la alta densidad de SNPs que se encuentra disponible en la actualidad.

Para derivar ψ (i, i +1), primero se debe calcular la probabilidad de ocurrencia de un determinado sendero de herencia p de largo l entre dos alelos a y b, en un DG generado al azar dentro de un pedigree, e indicada como $\Phi(p)$. Dado que las l gametogénesis involucradas en el sendero de herencia son independientes, y que un padre transmite a un hijo uno de sus dos alelos con probabilidad igual a ½, $\Phi(p) = (0.5)^l$. Dado que existe a lo sumo un sendero de herencia por cada DG, la probabilidad de que a y b sean IBD es igual a la sumatoria de las probabilidades de todos los senderos de herencia:

$$P(a \equiv b) = \sum_{p} \Phi(p) = \sum_{l=0}^{L} \lambda_l (0.5)^l = \theta(0.5) \qquad [3.4]$$

que es la IGF θ(0.5). Ahora bien, dado que existe un sendero de herencia p^{a_i,b_i} de longitud *l* entre los alelos a_i y b_i en el locus *i*, la probabilidad de que este sendero no se modifique en el locus vecino $i + 1$ para los alelos a_{i+1} y b_{i+1}, e indicada como $\phi(p, i, i+1)$, requiere que no haya recombinación en ninguna de las *l* gametogénesis involucradas en p^{a_i,b_i} y omitiendo la posibilidad de ocurrencia de doble recombinaciones. Al ser las gametogénesis independientes, se tiene que $\phi(p,i,i+1) = (1-r)^l$, donde *r* es la tasa de recombinación y que se puede calcular mediante la función de mapeo de Haldane (1919), basada en la distancia genética medida en cM entre dos marcadores. La probabilidad de transición $\psi(i, i+1)$ es simplemente un promedio ponderado de la probabilidad $\phi(p, i, i+1)$ de cada sendero de herencia posible:

$$\psi(i,i+1) = \frac{\sum_{p} \Phi(p) \cdot \phi(p,i,i+1)}{\sum_{p} \Phi(p)} = \frac{\sum_{l=0}^{L} \lambda_l (0.5)^l \cdot (1-r)^l}{\sum_{l=0}^{L} \lambda_l (0.5)^l} = \frac{\theta\left(0.5(1-r)\right)}{\theta(0.5)} \qquad [3.5]$$

Ahora bien, tanto el estado Bg-IBD como el estado No-IBD implican que los dos alelos no sean IBD *dentro del pedigree*. Esto es,

$$P(a_i \not\equiv b_i) = P(\text{No-IBD}_i) + P(\text{Bg-IBD}_i) = 1 - P(a_i \equiv b_i) = 1 - \theta\left(\tfrac{1}{2}\right) \qquad [3.6]$$

Por lo tanto, las probabilidades de transición desde cada uno de estos dos estados al caso IBD serán las mismas, i.e. $P(\text{IBD}_{i+1} \mid \text{No-IBD}_i) = P(\text{IBD}_{i+1} \mid \text{Bg-IBD}_i)$. Además, sabiendo que $P(a_{i+1} \equiv b_{i+1}) = P(a_i \equiv b_i)$ y condicionando en el estado oculto en el locus anterior, tenemos que

$$P(\text{IBD}_{i+1}) = \theta(0.5) = P(\text{IBD}_{i+1}|\text{IBD}_i)P(\text{IBD}_i) + P(\text{IBD}_{i+1}|\text{No-IBD}_i)P(\text{No-IBD}_i) + P(\text{IBD}_{i+1}|\text{Bg-IBD}_i)P(\text{Bg-IBD}_i) \quad [3.7]$$

Aplicando la restricción anterior y manipulando algebraicamente la expresión 3.7 obtenemos:

$$P(\text{IBD}_{i+1}|\text{No-IBD}_i) = P(\text{IBD}_{i+1}|\text{Bg-IBD}_i) = \frac{[1-P(\text{IBD}_{i+1}|\text{IBD}_i)]P(\text{IBD})}{P(\text{No-IBD})+P(\text{Bg-IBD})} = \frac{[1-\psi(i,i+1)]\theta(0.5)}{1-\theta(0.5)}. \quad [3.8]$$

Partiendo ahora del estado IBD, las transiciones hacia los estados Bg-IBD y No-IBD implican que los eventos de recombinación rompan el sendero de herencia existente entre los dos alelos. Esto es,

$$P(a_{i+1} \not\equiv b_{i+1}|\text{IBD}_i) = P(\text{No-IBD}_{i+1}|\text{IBD}_i) + P(\text{Bg-IBD}_{i+1}|\text{IBD}_i) = 1 - P(\text{IBD}_{i+1}|\text{IBD}_i) = 1-\psi(i,i+1). \quad [3.9]$$

Utilizando la definición de probabilidad condicional, la propiedad de reversibilidad de la cadena de Markov, i.e. $P(\text{Bg-IBD}_{i+1}, \text{IBD}_i) = P(\text{IBD}_{i+1}, \text{Bg-IBD}_i)$, y el resultado de la expresión 3.8 se obtiene,

$$P(\text{Bg-IBD}_{i+1}|\text{IBD}_i) = \frac{P(\text{Bg-IBD}_{i+1}, \text{IBD}_i)}{P(\text{IBD})} = \frac{P(\text{IBD}_{i+1}|\text{Bg-IBD}_i)P(\text{Bg-IBD})}{P(\text{IBD})} = \frac{[1-P(\text{IBD}_{i+1}|\text{IBD}_i)]P(\text{IBD})P(\text{Bg-IBD})}{[P(\text{No-IBD})+P(\text{Bg-IBD})]P(\text{IBD})} = \frac{[1-\psi(i,i+1)]P(\text{Bg-IBD})}{1-\theta(\frac{1}{2})} \quad [3.10]$$

Análogamente,

$$P(\text{No-IBD}_{i+1}|\text{IBD}_i) = \frac{[1-\psi(i,i+1)]P(\text{No-IBD})}{1-\theta(\frac{1}{2})} \quad [3.11]$$

Se asume que las probabilidades de transición hacia Bg-IBD y No-IBD son proporcionales a las probabilidades de observar Bg-IBD o No-IBD.

La probabilidad P(Bg-IBD) representa la coancestría entre dos individuos en exceso del pedigree y se estima directamente del HMM. La probabilidad de transición $P(\text{Bg-IBD}_{i+1}|\ \text{Bg-IBD}_i)$ se aproxima a través de $(1-r)^k$, donde k es el promedio del número de meiosis a partir de cada uno de los posibles senderos de herencia que conectan a estos dos individuos en generaciones previas a las incluidas en el pedigree. Las demás probabilidades de transición se calculan por diferencia en base a las probabilidades de transición derivadas en esta sección, ya que la suma de las probabilidades de transición desde un estado en particular debe ser siempre igual a 1. En el Apéndice I se derivan la demás probabilidades de transición. Este modelo sirve como base para construir el HMM entre un par de individuos sobre el cual, finalmente, se sustenta el algoritmo de Li *et al.* (2010).

3.3.2.3 *HMM para un par de individuos*

Sea $I(a, b)$ una función que, cuando los alelos a y b en un locus asociado a un SNP son IBD, es $I(a, b) = 1$; es igual a 0 si los alelos son No-IBD e igual a –1 si son Bg-IBD. Para simplificar la notación y de ahora en más, se denota como p^A (m^A) al alelo paterno (materno) del individuo A. Entre dos individuos A y B, se define al número de pares de genes IBD en un locus como

$$I(A,B) = \max\ [\ |I(p^A, p^B)| + |I(m^A, m^B)|,\ |I(p^A, m^B)| + |I(m^A, p^B)|\ ], \qquad [3.12]$$

de modo que $I(A, B) = 0$, 1 ó 2. Si se deja de lado, por el momento, la posibilidad de que dos alelos tomados al azar de los individuos A y B se encuentren en el estado Bg-IBD, los cuatro alelos de estos individuos tienen en total 15 configuraciones diferentes para los estados de IBD (Figura 3.5). Estas configuraciones son las 15 clásicas descriptas por

Jacquard (1974). Ahora bien, el HMM para un par de individuos se construye bajo el supuesto de ausencia de consanguinidad. En consecuencia, las combinaciones posibles para los estados de IBD se reducen a siete (Figura 3.5, izquierda). Si bien el modelo asume ausencia de consanguinidad, el evento de que los alelos paterno y materno de un individuo sean IBD no modifica el número de pares de genes IBD que comparte con otro individuo, i.e. el valor de $I(A, B)$. El modelo funciona bien aun en pedigrees con "loops" (Li *et al.*, 2010); de cualquier manera, las probabilidades de transición calculadas bajo este supuesto son una aproximación a la verdadera probabilidad.

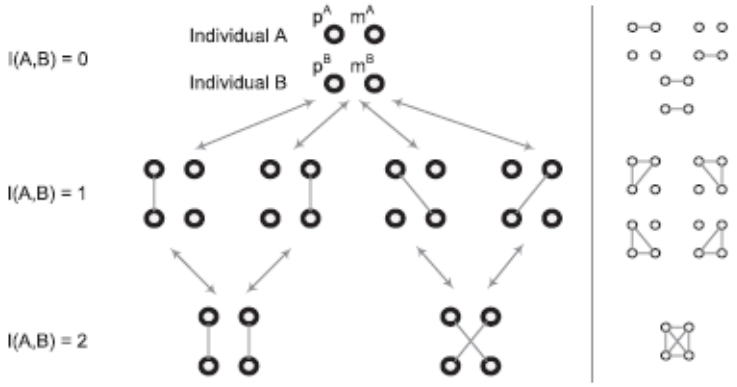
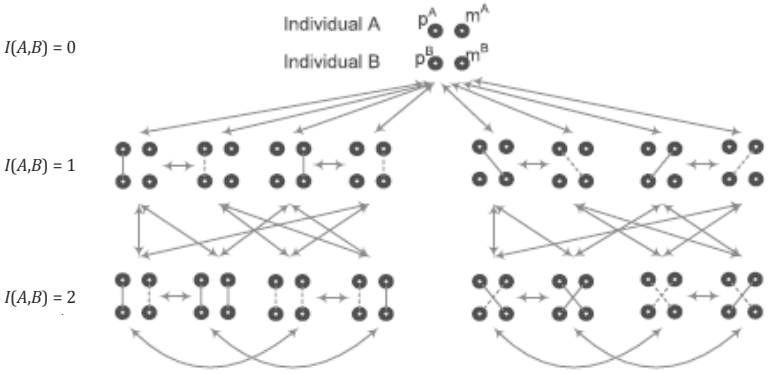


Figura 3.5. Estados de IBD posibles entre dos individuos, sin incorporar el LD. A la derecha se encuentran los estados no considerados por el modelo. Líneas llenas indican IBD. Tomado de Li *et al.* (2010).

Ahora bien, para incorporar el LD o Bg-IBD, se debe modificar la estructura del modelo de la Figura 3.5 (izquierda) agregando un estado más por cada estado original identificado como $I(A,B) = 1$, y adicionando tres estados más por cada estado original identificado como $I(A,B) = 2$. De este modo, los cuatro alelos de un par de individuos tienen en total 17 configuraciones posibles (Figura 3.6A). Estos 17 estados constituyen los estados ocultos del HMM para un par de individuos. Sea el vector de estado $\boldsymbol{s} = (I_1, I_2, I_3, I_4)$, donde $I_1 = I(p^A, p^B)$, $I_2 = I(p^A, m^B)$, $I_3 = I(m^A, p^B)$, $I_4 = I(m^A, m^B)$. Cada estado

es representado por un único vector (Figura 3.6B). Por ejemplo, sea $\boldsymbol{s}$ el estado en el cual los alelos paternos de *A* y *B* son IBD; entonces, $\boldsymbol{s}' = (1, 0, 0, 0)'$.

A



B

$I(A, B)$	$\boldsymbol{s}$	(I_1, I_2, I_3, I_4)
0	$\boldsymbol{s}_1$	(0, 0, 0, 0)
1	$\boldsymbol{s}_2$	(1, 0, 0, 0)
	$\boldsymbol{s}_3$	(-1, 0, 0, 0)
	$\boldsymbol{s}_4$	(0, 0, 0, 1)
	$\boldsymbol{s}_5$	(0, 0, 0, -1)
	$\boldsymbol{s}_6$	(0, 1, 0, 0)
	$\boldsymbol{s}_7$	(0, -1, 0, 0)
	$\boldsymbol{s}_8$	(0, 0, 1, 0)
	$\boldsymbol{s}_9$	(0, 0, -1, 0)
2	$\boldsymbol{s}_{10}$	(1, 0, 0, -1)
	$\boldsymbol{s}_{11}$	(1, 0, 0, 1)
	$\boldsymbol{s}_{12}$	(-1, 0, 0, -1)
	$\boldsymbol{s}_{13}$	(-1, 0, 0, 1)
	$\boldsymbol{s}_{14}$	(0, 1, -1, 0)
	$\boldsymbol{s}_{15}$	(0, 1, 1, 0)
	$\boldsymbol{s}_{16}$	(0, -1, -1, 0)
	$\boldsymbol{s}_{17}$	(0, -1, 1, 0)

Figura 3.6. A. HMM para un par de individuos. Líneas llenas indican IBD y líneas punteadas, Bg-IBD. Las flechas indican las transiciones posibles entre estados (faltan las flechas que representan la probabilidad de transición de un estado hacia sí mismo). **B.** Estados ocultos representados vectorialmente. Modificado de Li *et al.* (2010).

Una vez definidos los estados ocultos del modelo, se procede a la derivación de las probabilidades de transición entre estados. Dada una distancia pequeña entre marcadores, resulta muy poco probable que tenga lugar más de una recombinación, por lo tanto sólo se permiten transiciones entre cada estado con sí mismo y transiciones entre estados vecinos, i.e. estados que se diferencian en a lo sumo un estado alélico oculto, situación que se cumple cuando $\left\|\boldsymbol{s}-\boldsymbol{s}'\right\| \in \mathbb{N}_{0,1,2}$. Es decir que la norma sólo puede tomar como valor los números naturales 0, 1 ó 2. Por ejemplo, partiendo del estado oculto $\boldsymbol{s}' = (1, 0, 0, 0)$ las transiciones permitidas (flechas) se muestran en la Figura 3.7. Una transición no permitida desde $\boldsymbol{s}$ se produce cuando $\boldsymbol{s}' = (-1, 0, 0, -1)$. En ese caso, el vector diferencia es igual a $\boldsymbol{s} - \boldsymbol{s}' = (2, 0, 0, 1)$ y $\|\boldsymbol{s} - \boldsymbol{s}'\| = \sqrt{5}$.

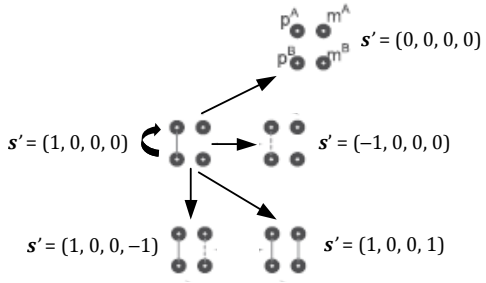


Figura 3.7. Transiciones permitidas desde el estado $\boldsymbol{s}' = (1, 0, 0, 0)$. En todos los casos $||\boldsymbol{s} - \boldsymbol{s}'||$ toma valores iguales a 0, 1 ó 2.

Condicional a que el par de individuos se encuentre en el estado $\boldsymbol{s}$ en el locus i, la probabilidad de que se halle en el estado $\boldsymbol{s}'$ en el locus $i + 1$, indicada como $f(\boldsymbol{s}'|\boldsymbol{s})$, es esencialmente el producto de dos cadenas independientes generadas por HMMs alélicos como el descripto en la sección 3.3.2.2. Esto es, $f(\boldsymbol{s}'|\boldsymbol{s}) = P(I'_j|I_j)\, P(I'_{\boldsymbol{k}}|I_{\boldsymbol{k}})$, donde $1 \leq j, k \leq 4$ son dos coordenadas independientes en los vectores $\boldsymbol{s}$ y $\boldsymbol{s}'$, lo cual significa que los dos alelos en $I_{\boldsymbol{k}}$ son diferentes en origen a los dos alelos en I_j, y los dos alelos en $I_{\boldsymbol{k}}$ (I_j) son los mismos que aquellos en $I'_{\boldsymbol{k}}$ (I'_j). La independencia se deduce a partir del supuesto de la ausencia de consanguinidad. Tomando como ejemplo una de las transiciones posibles de la Figura 3.7,

$$\begin{aligned} & f\left(\boldsymbol{s}' = (1,0,0,1) \middle| \boldsymbol{s} = (1,0,0,0)\right) = P\left(I'_1 \middle| I_1\right) P\left(I'_4 \middle| I_4\right) \\ & = P\left(I'\left(p^A, p^B\right) = 1 \middle| I\left(p^A, p^B\right) = 1\right) P\left(I'\left(m^A, m^B\right) = 1 \middle| I\left(m^A, m^B\right) = 0\right) \\ & = P\left(\mathrm{IBD}_{i+1} \middle| \mathrm{IBD}_i\right)\ P\left(\mathrm{IBD}_{i+1} \middle| \mathrm{No\text{-}IBD}_i\right) \end{aligned} \qquad [3.13]$$

Y, reemplazando en la expresión 3.13 por las expresiones 3.5 y 3.8 de la sección anterior, tenemos finalmente que

$$f\left(\boldsymbol{s}' = (1,0,0,1) \middle| \boldsymbol{s} = (1,0,0,0)\right) = \psi(i, i+1) \cdot \frac{\left(1 - \psi(i, i+1)\right)\theta(0.5)}{1 - \theta(0.5)} \qquad [3.14]$$

Las probabilidades de transición del HMM de un par de individuos $P(\boldsymbol{s}'|\ \boldsymbol{s})$, para todo $\left|\left|\boldsymbol{s} - \boldsymbol{s}'\right|\right| \in \mathbb{N}_{0,1,2}$, son proporcionales a las probabilidades condicionales $f(\boldsymbol{s}'|\ \boldsymbol{s})$, tal que

$$P(\boldsymbol{s}'|\boldsymbol{s}) = \frac{f(\boldsymbol{s}'|\boldsymbol{s})}{\sum_{t:\|t'-s\|\in\mathbb{N}_{0,1,2}} f(\boldsymbol{t}|\boldsymbol{s})} \qquad [3.15]$$

En la expresión 3.15 la notación $\boldsymbol{t}$ indica todos los estados posibles $\boldsymbol{s}'$ para el locus $i + 1$ que cumplan con la condición $\|\boldsymbol{s}-\boldsymbol{s}'\|\in\mathbb{N}_{0,1,2}$, y por lo tanto que puedan ser transitados desde el estado $\boldsymbol{s}$ para el locus i.

Para terminar de construir el modelo resta definir los símbolos observables para cada estado y derivar las probabilidades de emisión. Sea $G(a, b)$ una variable indicadora definida para un par de alelos a y b, i.e. $G(a, b) = 1$ si son IBS y $G(a, b) = 0$ en caso contrario. Entre dos individuos A y B, se define al número de pares de alelos IBS en un locus como

$$G(A, B) = \max [\ G(p^A, p^B) + G(m^A, m^B),\quad G(p^A, m^B) + G(m^A, p^B)\ ], \qquad [3.16]$$

de modo que $G(A, B) = 0$, 1, ó 2. Es decir, $G(A, B)$ es el número de pares de alelos del mismo tipo entre los dos genotipos y constituye los símbolos observables del HMM. Para derivar las probabilidades de emisión $P(G(A, B)|\ \boldsymbol{s})$, se clasifican los 17 estados ocultos en tres clases de acuerdo al valor de $I(A, B)$ (Figura 2.6B). Todos los estados que tengan igual valor de $I(A, B)$ tendrán la misma probabilidad de emitir un valor de $G(A, B)$ (Cuadro 3.1). La derivación detallada de las probabilidades de emisión se encuentra en el Apéndice I. La probabilidad $P(G(A, B)|\ \boldsymbol{s})$ es en esencia el producto de dos probabilidades de emisión para dos pares de haplotipos independientes. En términos generales, dado un par de haplotipos y dado que los dos alelos en un locus son IBD (sea Bg-IBD o IBD), deben ser necesariamente IBS. Si los dos alelos son No-IBD, todavía existe la posibilidad de que sean IBS; esa probabilidad es la de observar dos alelos del mismo tipo, la cual es igual a $p^2 + q^2$, donde p y $q = 1 - p$ son las frecuencias alélicas poblacionales.

Cuadro 3.1. Probabilidades de emisión de un número de pares de alelos IBS, $G(A, B)$, dado un estado IBD, $\boldsymbol{s}$. Modificado de Li *et al.* (2010).

$G(A, B)$	$I(A, B)$	$\boldsymbol{s}$	$P(G(A, B) \mid \boldsymbol{s})$
0	0	$\boldsymbol{s}_1$	$2p^2q^2$
1	0	$\boldsymbol{s}_1$	$4p^3q + 4pq^3$
2	0	$\boldsymbol{s}_1$	$p^4 + q^4 + 4p^2q^2$
0	1	$\boldsymbol{s}_2 - \boldsymbol{s}_9$	0
1	1	$\boldsymbol{s}_2 - \boldsymbol{s}_9$	$2pq$
2	1	$\boldsymbol{s}_2 - \boldsymbol{s}_9$	$p^2 + q^2$
0	2	$\boldsymbol{s}_{10} - \boldsymbol{s}_{17}$	0
1	2	$\boldsymbol{s}_{10} - \boldsymbol{s}_{17}$	0
2	2	$\boldsymbol{s}_{10} - \boldsymbol{s}_{17}$	1

Dadas las probabilidades de transición, las probabilidades de emisión y la secuencia de observaciones $G = G_1(A, B)\ G_2(A, B)\ldots\ G_T(A, B)$ entre dos individuos A y B a partir de sus genotipos en cada locus, es posible obtener dos tipos de soluciones. Una de ellas es encontrar la secuencia de estados de IBD $Q = q_1\ q_2\ldots\ q_T$ más probable, donde q_i denota el valor más probable que toma el vector de estado $\boldsymbol{s}$ en el i-ésimo SNP ($\boldsymbol{s}$ es un estado oculto de los descriptos en la Figura 3.5) y T es el número total de SNPs; en otras palabras, la secuencia de estados de IBD que mejor explica la secuencia de observaciones G. Esta secuencia se obtiene utilizando el algoritmo de Viterbi (Rabiner, 1989). La otra posibilidad es obtener las probabilidades a posteriori de los estados de IBD $P\left(q_i = \boldsymbol{s} \middle| G\right)$ para cada locus, es decir, la probabilidad de ocurrencia del estado $\boldsymbol{s}$ en el locus i para todo $i \in \{1, 2,\ldots, T\}$, dada la secuencia de observaciones G y dado el modelo. Estas probabilidades se calculan utilizando el algoritmo Forward-Backward (Rabiner, 1989).

3.3.3 Estimación de la proporción del genoma IBD

Esta sección constituye la contribución teórica original del capítulo. Para estimar la proporción del genoma IBD entre dos individuos *A* y *B* seguiremos el enfoque de Bulmer (1985). Sea la variable aleatoria *X* el número de pares de alelos IBD entre dos individuos en un locus al azar del genoma. El espacio muestral de *X* es 0, 1 ó 2, con probabilidades P_0, P_1 y P_2, respectivamente. Dichas probabilidades dependen de la relación de parentesco. Sea la variable aleatoria $Z = X/2$ la proporción de alelos IBD entre dos individuos, que toma valores 0, 0,5 ó 1, según *X* sea 0, 1 ó 2, respectivamente. Se define el coeficiente de relación aditiva a_{AB} como el valor esperado de *Z*. De este modo, $a_{AB} = \mathrm{E}(Z) = \frac{1}{2}P_1 + P_2$. Ésta es una medida de identidad genética basada en *un* locus. Generalizando este concepto de modo de obtener una medida de identidad genética global basada en el genoma completo, y condicional a la información que brinda el panel de SNPs, se propone estimar la proporción del genoma compartido IBD entre dos individuos *A* y *B*, $\hat{g}^{AB}_{\text{IBD-LD}}$, a partir de las probabilidades $P(q_i = \boldsymbol{s}|G)$ para cada locus obtenidas mediante el algoritmo de Li *et al.* (2010), dado que

$$P(I_i(A,B)=1|G) = \sum_{\boldsymbol{s}\,\therefore\, I(A,B)=1} P(q_i = \boldsymbol{s}|G), \qquad i = 1, 2, ..., T \qquad [3.17]$$

y del mismo modo,

$$P(I_i(A,B)=2|G) = \sum_{\boldsymbol{s}\,\therefore\, I(A,B)=2} P(q_i = \boldsymbol{s}|G), \qquad i = 1, 2, ..., T \qquad [3.18]$$

donde $P(I(A,B)_i = 1|G)$ y $P(I(A,B)_i = 2|G)$ son, respectivamente, las probabilidades *a posteriori* de compartir 1 o 2 pares de genes IBD en el *i*-ésimo SNP, condicional a la secuencia de observaciones *G*. En función de estas probabilidades, se propone calcular $\hat{g}^{AB}_{\text{IBD-LD}}$ mediante la expresión siguiente

$$\hat{g}^{AB}_{\text{IBD-LD}} = 0.5 \sum_{i=1}^{T} w_i P\left(I_i\left(A,B\right)=1 \middle| G\right) + \sum_{i=1}^{T} w_i P\left(I_i\left(A,B\right)=2 \middle| G\right) \quad [3.19]$$

donde w_i es un coeficiente de ponderación que representa la cobertura del i-ésimo SNP relativa a la longitud física del genoma. Las expresiones $\sum_{i=1}^{T} w_i P\left(I_i\left(A,B\right)=1 \middle| G\right)$ y $\sum_{i=1}^{T} w_i P\left(I_i\left(A,B\right)=2 \middle| G\right)$ representan, respectivamente, una estimación del promedio de la probabilidad de compartir 1 y 2 pares de genes IBD para un par de individuos y reemplazarían a P_1 y P_2 en E(Z) o relación aditiva a_{AB}. Los valores de $\hat{g}^{AB}_{\text{IBD-LD}}$ calculados a partir de la expresión 3.19 constituirán los elementos de la matriz de relaciones genómicas basada en la noción de la fracción de ADN compartido IBD, la cual de aquí en adelante denotaremos como $\boldsymbol{G}_{\textbf{IBD-LD}}$.

CAPÍTULO 4

Comparación de métodos para estimar relaciones genómicas utilizando el pedigree y los marcadores en familias con muchos animales sin genotipar[1]

[1] Forneris, N.S., Steibel, J.P., Legarra, A., Vitezica, Z.G., Bates, R.O., Ernst, C.W., Basso, A.L., Cantet, R.J.C. 2015. A comparison of methods to estimate relationships using pedigree and markers in populations with many ungenotyped animals. Journal of Animal Breeding and Genetics (*En revisión*).

Comparación de métodos para estimar relaciones genómicas utilizando el pedigree y los marcadores en familias con muchos animales sin genotipar

4.1 Introducción

En las evaluaciones genéticas tradicionales que utilizan la información del pedigree, el cálculo de las relaciones de parentesco esperadas (i.e., las relaciones aditivas, que son calculadas a partir de la genealogía), asume implícitamente equilibrio Hardy-Weinberg y un número infinito de loci de segregación independiente distribuidos por todo el genoma. Las relaciones aditivas contienen información sobre el parecido genético que proviene de la herencia en común (identidad por descendencia, IBD). Por lo tanto, es probable que los genes que son copias de un mismo gen en un ancestro común, compartan en promedio los mismos loci causales. Consecuentemente, los datos fenotípicos de individuos emparentados resultan informativos para la predicción del BV de cualquiera de esos animales.

La disponibilidad de paneles de marcadores moleculares en alta densidad del tipo SNP para especies ganaderas permitió estimar las relaciones de parentesco realizadas, las cuales pueden ser derivadas sólo con información de los SNPs. La matriz de relaciones de parentesco genómicas (***G***) calculada con marcadores tiene un papel primordial en la predicción de los BVs a partir de los modelos animales, cuando se utilizan predictores genómicos lineales insesgados de minina varianza (VanRaden, 2008). VanRaden (2007, 2008) propuso calcular las relaciones genómicas sumando los productos cruzados de los alelos de los SNPs, codificados como el número de copias de un alelo que tiene un SNP (0, 1, ó 2), desviados de una función de las frecuencias alélicas de cada locus y escalados por esas mismas frecuencias. Como resultado, la semejanza entre alelos en todos los loci marcadores (identidad por estado, IBS)

constituye el tipo de información en la que el parecido genético entre los animales se traduce a ***G***. La eficacia predictiva del mejor predictor lineal insesgado (BLUP, *best linear unbiased predictor*) o exactitud, depende de la medida en que las relaciones genómicas derivadas de los marcadores capturan los patrones de relaciones genómicas realizadas en los loci causales (VanRaden, 2007, 2008; de los Campos *et al.*, 2013). Así, los elementos de la matriz ***G*** son estimaciones de la proporción "realizada" del genoma compartido entre dos individuos, mientras que los elementos de la matriz ***A*** de relaciones aditivas calculadas sólo con pedigree son los valores "esperados" de esa proporción (Goddard *et al.*, 2011).

El tamaño finito del genoma y el proceso de la recombinación introducen aleatoriedad y variabilidad en la proporción del genoma compartido IBD para cualquier relación de parentesco genealógico en particular (Risch y Lange, 1979; Guo, 1996; Hill y Weir, 2011), hecho que lleva a que las relaciones realizadas suelan diferir de su valor esperado. Esta variabilidad extra es responsable de la ganancia en exactitud cuando se intenta predecir los BVs. Puede demostrarse que esta ganancia en exactitud se debe a la reducción en la varianza de los residuos de segregación Mendelianos de los BVs genómicos (Cantet y Vitezica, 2014). Si bien las estimaciones de las relaciones utilizando tanto el pedigree como los marcadores (IBD) o sólo los marcadores (IBS) son capaces ambas de estimar las relaciones "realizadas", su eficiencia depende de cuán bien estos métodos capten las señales distorsionadas de la distribución "multilocus" (desequilibrio de ligamiento, ligamiento) y de los valores realizados de los modos de identidad (Jacquard, 1974), los que a su vez se ven afectados por la información incompleta del pedigree y la consanguinidad, para cualquier relación de parentesco genealógico y par de individuos en particular. Las estimaciones de las relaciones genómicas de VanRaden (2008) pueden no ser precisas porque involucran las

frecuencias alélicas de la población base, a menos que haya un grupo suficientemente grande de animales genotipados (de la población base). Si este no es el caso, incorporar la información del pedigree en estos cálculos podría ser una estrategia cuando se cuenta con familias grandes con un número pequeño de animales genotipados.

Se podría decir, entonces, que existen dos enfoques para construir matrices de relaciones de parentesco genómicas. Uno es el actual, ampliamente utilizado en el método BLUP de predicción de valores de cría genómicos (GBLUP; VanRaden, 2008), sustentado en la noción de identidad por estado (IBS) y el empleo solamente de marcadores para inferir la proporción de genoma compartido entre los individuos. El otro está basado en la noción de IBD e infiere las relaciones rastreando la transmisión de los marcadores a lo largo de todo el pedigree conocido (análisis de ligamiento). Existen diferentes algoritmos fundamentalmente en el área de la genética humana que podrían utilizarse a priori en la construcción de matrices genómicas basadas en este segundo enfoque. En el Capítulo 3, se hizo una revisión del marco teórico detrás de varios de estos algoritmos y se propuso una metodología que puede aplicarse incluso cuando se trabaja con familias de animales con muchos miembros sin genotipar, y que además tiene en cuenta el efecto del LD poblacional entre marcadores. Se espera, de este modo, que la correcta elección de un algoritmo IBD para calcular las relaciones dentro del segundo enfoque mejore la precisión de las estimaciones.

4.2 Objetivo

El objetivo de este capítulo es evaluar dos metodologías para el cálculo de las relaciones de parentesco genómicas entre los animales genotipados: la primera, utilizada actualmente en SG, estima la fracción de ADN compartido para cada par de individuos basándose únicamente en la información de los marcadores presentes en cada individuo;

la segunda, basada en algoritmos utilizados en el área de la genética humana y propuesta como alternativa en el Capítulo 3, incorpora la información del pedigree además de la información del panel denso de marcadores, e introduce, de este modo, la noción de fracción de ADN compartido IBD. Se evaluará la performance de los métodos en la inferencia de la proporción del genoma compartido en un pedigree animal complejo mediante simulación estocástica y su aplicación en una base de datos real.

4.3 Materiales

4.3.1 Base de datos reales

Los datos provienen de una población experimental de cerdos domésticos (MSUPRP; East Lansing, Michigan, USA; Edwards *et al.*, 2008), disponible mediante un convenio con Michigan State University. La población F_0 (generación fundadora) surgió del apareamiento entre 4 machos de la raza Duroc (no emparentados) con 15 hembras de la raza Pietrain. De los animales resultantes en la F_1, se mantuvieron 50 hembras y 6 machos (hijos de 3 padres F_0) para ser utilizados como padres de la generación F_2, compuesta por 1.259 cerdos pertenecientes a 141 camadas y 11 grupos de destete. Todos los apareamientos se organizaron de modo de evitar el apareamiento de hermanos enteros o medios hermanos (i.e., sin consanguinidad), mediante inseminación artificial. De la población F_2, se seleccionaron 336 cerdos para ser genotipados, en representación de todas las familias de hermanos enteros (Gualdrón Duarte *et al.*, 2013). Un total de 411 animales (4 machos Duroc F_0, 15 hembras F_0 Pietrain, 6 machos F_1, 50 hembras F_1 y 336 cerdos F_2) fueron genotipados con el chip de alta densidad PorcineSNP60 BeadChip (Illumina, Inc.) diseñado por Ramos *et al.* (2009). El genotipado se realizó en un laboratorio comercial (GeneSeek a Neogen Company, Lincoln, NE, EE.UU.). Se descartaron aquellos SNPs que no pertenecían a

los cromosomas autosómicos (15.298), SNPs con una MAF < 0,01, "call rate" < 90% (i.e., más de 10% de genotipos faltantes) y SNPs que mostraban más de 2% de inconsistencias Mendelianas. De un total de 62.163 SNPs, 38.263 pasaron los filtros de control de calidad y fueron empleados en los análisis posteriores.

4.3.2 Base de datos simulados

El objetivo de la simulación estocástica fue comparar la performance del método propuesto para estimar la proporción de IBD inferida con diferentes relaciones, respecto del algoritmo utilizado en la actualidad en las evaluaciones genómicas (VanRaden, 2008). A tal efecto, se generó un conjunto de datos simulados utilizando el programa QMSim (Sargolzaei y Schenkel, 2009). La estructura de los datos se asoció con los de un programa de mejoramiento de cerdos que fue formulado de manera simplificada. La simulación se llevó a cabo en dos etapas. En la primera, se simulan las generaciones históricas para crear el nivel deseado de LD y, en una segunda etapa, se genera la estructura poblacional más reciente que surge por los procesos de selección y apareamiento en especies pecuarias. Se asumió un genoma formado por 10 cromosomas (5 pares de autosomas) de 160 cM de longitud. En la primera generación histórica, se distribuyeron 35.000 marcadores bialélicos aleatoriamente a lo largo del genoma, asignando la posición de cada uno mediante una distribución uniforme, con una frecuencia alélica igual a 0,5. Se aplicó una tasa de mutación igual a 2×10^{-4} por locus por generación, asumiendo un modelo de mutación recurrente. A través de las generaciones, los segmentos cromosómicos maternos y paternos fueron heredados con una media de 1 crossover/cromosoma y simulados a partir de una distribución Poisson, ubicando los crossovers a lo largo del cromosoma a partir de una distribución uniforme.

A fin de establecer un equilibrio entre deriva y mutación y crear LD, la historia de la población involucró: 2.500 generaciones de tamaño constante bajo apareamiento aleatorio, comenzando con una población de tamaño efectivo N_e = 500 e igual número de machos y hembras, seguido por un cuello de botella severo durante 30 generaciones que redujo la variabilidad a un valor de N_e = 75. Para la última generación histórica, se generaron 20 machos y 200 hembras mediante la elección aleatoria de dos gametas de los pooles gaméticos masculino y femenino. Estos animales constituyeron la población fundadora o base (G_0). Entre los loci marcadores con MAF > 0.01 en G_0, se eligieron 16000 SNPs (espaciados cada 0,05 cM en promedio, imitando la densidad del panel de 60K actualmente disponible en cerdos) al azar.

Se consideró un carácter poligénico con heredabilidad (h^2) igual a 0,25 y varianza fenotípica igual a 1. Luego, se aplicó el siguiente esquema de apareamiento y selección por 5 generaciones. En cada generación, 20 verracos se cruzaron con 200 cerdas, minimizando la endogamia, para producir 2.000 crías (la mitad de ellas machos). De la descendencia, se seleccionaron aquellos machos cuyo BV estimado sea mayor en base al BLUP vía un MA, y se seleccionan 200 hembras al azar. Se asumió que los registros del pedigree se encontraban disponibles para las 5 generaciones (10.220 animales), incluyendo los fenotipos de todos los machos. Se asumió que el genotipado se realizó en 140 animales (i.e., los 20 machos de G_0, los machos selectos de cada generación, y 40 candidatos a la selección para G_6). Se generaron 50 réplicas a los efectos de disminuir la varianza de muestreo.

4.4 Métodos

4.4.1 Proporción del genoma compartido IBD por un par de individuos emparentados

Sea g^{ij}_{T} la verdadera proporción del genoma compartido IBD por un par de individuos, *i* y *j*, dentro de un pedigree conocido (i.e., a partir de la generación F_0 en los datos reales y de la generación G_0 en los datos simulados). El valor de g^{ij}_{T} es conocido en los datos simulados, pero desconocido en los datos reales; no obstante, existen fórmulas teóricas que permiten computar su media y su varianza. Para simplificar la notación, de aquí en adelante omitiremos los superíndices *ij*.

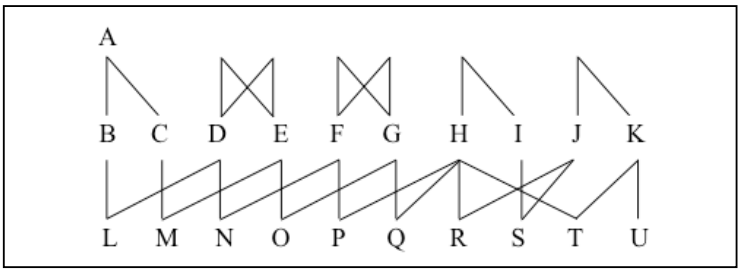
Para el conjunto de datos reales sin consanguinidad, todos los pares de animales (84.254 pares) fueron clasificados en 14 clases diferentes, según la relación de parentesco genealógico (e.g., medios hermanos, hermanos enteros, ver Figura 4.1). Para cada relación, el valor promedio de g_{T} se calculó a partir del pedigree (es igual al coeficiente de relación aditiva) y la varianza fue calculada utilizando las fórmulas teóricas derivadas por Hill y Weir (2011) para individuos no consanguíneos, que dependen del número de cromosomas y su longitud de mapa. La longitud de mapa promedio entre sexos (cM) fue tomada de las tasas de recombinación reportadas por Tortereau *et al.* (2012). La media y la varianza general de g_{T} se pueden derivar utilizando la teoría de distribución de mezclas finitas (Frühwirth-Schnatter, 2006). La proporción de genoma compartido IBD por un par de animales será simbolizada como g_{T}, y proviene de una mezcla de distribuciones con función de densidad igual a:

$$f\left(g_{\mathrm{T}}\right) = \sum_{k=1}^{K} \eta_k \; p\left(g_{\mathrm{T}_k}\right) \qquad [4.1]$$

En la expresión 4.1, $p(g_{T_k})$ denota la función de densidad de probabilidad condicional de g_T dada la relación de parentesco genealógico de la clase k (k = 1, .., 14) (Figura 4.1), y η_k es el coeficiente de mixtura para la clase k tal que $\sum_{k=1}^{K} \eta_k = 1$. Entonces,

$$\mathrm{E}(g_T) = \sum_{k=1}^{K} \eta_k \ \mathrm{E}\left(g_{T_k}\right) \qquad [4.2]$$

$$\mathrm{Var}(g_T) = \sum_{k=1}^{K} \eta_k \left[\mathrm{E}^2\left(g_{T_k}\right) + \mathrm{Var}\left(g_{T_k}\right)\right] - \left[\mathrm{E}(g_T)\right]^2 \qquad [4.3]$$



Parentesco	Ejemplo	Parentesco	Ejemplo
De una vía		*De dos vías*	
Padre-hijo	AB	Hermanos enteros (HE)	DE
Abuelo-Nieto	AL	Doble medio primos	ST
Medio hermanos (MH)	HI	Triple medio primos	LM
Medio tío-sobrino	JU	Doble primos hermanos	NO
Medio primos	SU	MH, madres(padres) MH	RS
Tío-sobrino	DO	MH, madres (padres) HE	PQ
Primos	NQ	No relacionados	BD

Figura 4.1: Ejemplos de las relaciones de parentesco genealógico en los datos reales.

Para los datos simulados, g_T se computó para cada par de animales como:

$$g_T = \frac{\sum_{c=1}^{C} l_c \, g_{T,c}}{L} \qquad \text{y} \qquad L = \sum_{c=1}^{C} l_c \qquad [4.4]$$

La notación $g_{T,c}$ indica el valor de la proporción de genoma compartido en el cromosoma c de longitud l_c (Guo, 1995). A fin de computar $g_{T,c}$, los cromosomas materno y paterno de cada animal fueron divididos en segmentos delimitados por los

crossovers que tuvieron lugar durante los procesos de gametogénesis que interconectan un individuo con sus ancestros en G_0. En un segmento dado, se determinó cuál de los $2n$ alelos fundadores en G_0 heredó un individuo utilizando el archivo de output "crossover" que arroja el programa QMSim. Entonces, $g_{T,c}$ se calculó como

$$g_{T,c} = \frac{2}{l_c} \int_0^{l_c} r(t)\, dt \qquad [4.5]$$

donde $r(t) = \tilde{\Delta}_1 + \frac{1}{2}\left(\tilde{\Delta}_3 + \tilde{\Delta}_5 + \tilde{\Delta}_7\right) + \frac{1}{4}\tilde{\Delta}_8$ es el coeficiente de coancestría realizado u observado en la posición t en el segmento c, y $\tilde{\Delta}_1$, $\tilde{\Delta}_3$, $\tilde{\Delta}_5$, $\tilde{\Delta}_7$ y $\tilde{\Delta}_8$ son los valores realizados de los nueve "coeficientes de identidad condensados" de Jacquard (1974). El valor realizado de cada coeficiente $\left(\tilde{\Delta}_l ; l = 1, \ldots, 9\right)$ es igual a 0 ó 1, dependiendo del patrón IBD observado entre los cuatro alelos presentes en ambos animales (Figura 4.2). También se calcularon la media y varianza de g_T para cada réplica.

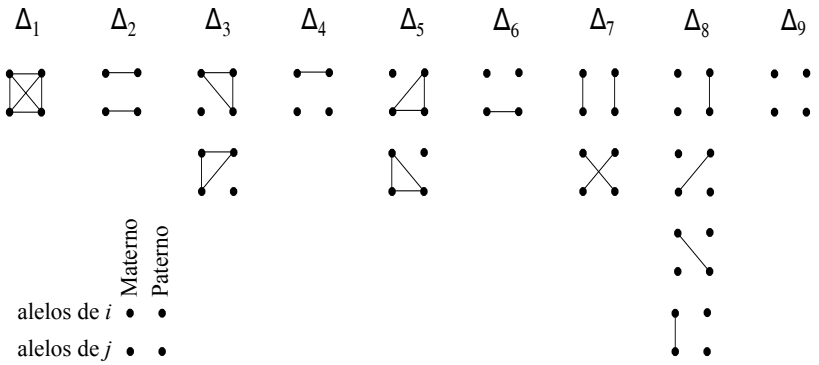


Figura 4.2: Coeficientes de identidad condensados (Δ_1, Δ_2, …, Δ_9) de Jacquard (1974).

4.4.2 Estimación de la proporción del genoma compartido IBD por un par de animales genotipados

Se compararon dos enfoques para estimar las relaciones en base a la información que brinda actualmente un panel de SNPs de alta densidad (ausencia de fase conocida, presencia de LD entre marcadores), incorporando o no la información del pedigree y de la herencia. El primero se asocia con la noción de identidad por estado (IBS), ampliamente utilizada en los métodos BLUP de predicción del valor de cría genómico (GBLUP) (VanRaden, 2008), y utiliza los genotipos centrados para cuantificar el número de alelos compartidos entre individuos, suma sobre todos los SNPs y escala por la varianza. Por lo tanto, la relación estimada entre un par de animales *i* y *j* ($\hat{g}_{VR}$) es igual a:

$$\hat{g}_{VR} = \frac{\sum_{m=1}^{M}\left(x_{im} - \mu_m\right)\left(x_{jm} - \mu_m\right)}{2\sum_{m=1}^{M} p_m\left(1-p_m\right)}, \qquad [4.6]$$

En [4.6] la variable aleatoria x_m es codificada como −1, 0, y 1 para el genotipo homocigota, heterocigota, y el otro homocigota, respectivamente, $\mu_m = 2(p_m - 0.5)$ es la media poblacional de los valores genotípicos, y p_m es la frecuencia poblacional del segundo alelo en el locus *m*. Estas relaciones constituyen cada uno de los elementos de la matriz de relaciones genómicas $\boldsymbol{G}$. Estas fueron calculadas con el programa PREGSF90 (Aguilar *et al*., 2011), sea utilizando las frecuencias alélicas observadas para cada SNP ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) o las frecuencias alélicas de los animales de la población base (F_0/G_0) ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$). Para evitar problemas de singularidad, la matriz $\boldsymbol{G}_{\mathbf{VR}}$ se calculó como $\boldsymbol{G} = w\,\boldsymbol{G}^* + (1 - w)\,\boldsymbol{A}_{\mathbf{22}}$, donde $w = 0.95$, $\boldsymbol{G}^*$ es la matriz de relaciones genómicas antes de la ponderación y $\boldsymbol{A}_{\mathbf{22}}$ es la matriz de relaciones aditivas para los animales genotipados.

La matriz $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ fue además escalada basada en $\boldsymbol{A}_{\mathbf{22}}$ para controlar el sesgo (Vitezica *et al.*, 2011).

El segundo enfoque infiere las relaciones genómicas rastreando la transmisión de los marcadores a través del pedigree conocido (análisis de ligamiento). Para ello, se utilizó el modelo Markov oculto (*Hidden Markov Model*, HMM) propuesto por Li *et al.* (2010) e implementado en el algoritmo Forward-Backward disponible en el programa PEDIBD (Li *et al.*, 2010). Para una definición formal de un HMM y una descripción detallada del marco teórico sobre el cual se construye el algoritmo de Li *et al.* (2010), véase la sección 3.3.1 y 3.3.2 del Capítulo 3. El citado algoritmo incorpora la información del pedigree, para cualquier individuo esté o no genotipado (como es el caso en nuestras bases de datos). A continuación describiremos las ideas centrales. Para un par de individuos genotipados en particular, el estado oculto (q_m) del HMM es el número (0, 1 ó 2) de pares de alelos IBD en el SNP de la posición m. El estado observable, O_m, es el número de pares de alelos IBS en la misma posición. Primero, se construye el HMM para un par de alelos con tres estados ocultos posibles: 1) No-IBD, 2) IBD dentro del pedigree conocido y 3) "background IBD" o IBD a nivel poblacional, para ajustar por el grado de parentesco histórico "oculto" entre individuos, más allá del parentesco que se observa a partir del pedigree disponible. Este "background IBD" es un factor de "ruido" (*nuisance*), dado que nuestro objetivo es estimar IBD a partir de los fundadores del pedigree pero no más atrás en el tiempo. Las probabilidades de transición entre estados dependen, no sólo del intervalo entre marcadores, sino también de todos los senderos de herencia posibles dentro del pedigree que conectan dos alelos marcadores. A partir de este modelo básico, se construye el HMM para un par de individuos asumiendo independencia entre los cromosomas homólogos de un individuo,

que es una aproximación en un pedigree con "loops". Así, para los individuos *i* y *j*, la proporción estimada de genoma compartido IBD ($\hat{g}_{\text{IBD-LD}}$) se puede calcular como

$$\hat{g}_{\text{IBD-LD}} = \sum_{m=1}^{M} w_m \left[\tfrac{1}{2} P(q_m = 1 \left| O_1, ..., O_M \right.) + P(q_m = 2 \left| O_1, ..., O_M \right.) \right] \qquad [4.7]$$

donde w_m es el coeficiente de ponderación del *m*-ésimo SNP y $P(q_m = 1 | O_1,...,O_M)$ ($P(q_m = 2 | O_1,...,O_M)$ es la probabilidad *a posteriori* de compartir 1(2) par(pares) de alelos IBD en la posición *m*, condicional a la información de todos los loci marcadores. Cada coeficiente w_m se calculó como la cobertura del *m*-ésimo SNP relativa a la longitud física del genoma. La expresión 4.7 es equivalente a la expresión 3.19 propuesta en la sección 3.3.3 del Capítulo 3. Ésta es una medida de identidad genética basada en todo el genoma y constituye cada uno de los elementos de la matriz de relaciones genómicas $\boldsymbol{G}_{\textbf{IBD-LD}}$.

La matriz $\boldsymbol{G}_{\textbf{IBD-LD}}$ puede ser indefinida (i.e., puede tener autovalores negativos y pequeños). Esto se debe a que los elementos de $\boldsymbol{G}_{\textbf{IBD-LD}}$ (las relaciones genómicas) se estiman "de a pares" en lugar de hacerlo de manera global. Por lo tanto, se utilizó la función "nearPD" del paquete "Matrix" de R para computar la matriz positiva definida más cercana a la matriz $\boldsymbol{G}_{\textbf{IBD-LD}}$ original (Cheng y Higham, 1998; Higham, 2002). Estas estimaciones fueron las retenidas para los análisis estadísticos posteriores.

4.4.3 Análisis estadístico

En la base de datos reales, la media y la varianza muestral de la proporción estimada de genoma compartido IBD ($\hat{g}_{\text{VR-O}}$, $\hat{g}_{\text{VR-B}}$ o $\hat{g}_{\text{IBD-LD}}$) fueron calculadas dentro de cada clase de parentesco genealógico (Figura 4.1), para el pedigree completo y contrastadas con los valores teóricos de g_T. También se calcularon las correlaciones

entre los valores estimados de las relaciones o de la proporción de genoma compartido y sus correspondientes coeficientes de relación aditiva, a_{ij}, obtenidos a partir del pedigree.

A cada réplica de la simulación, se calcularon como estimadores de precisión el error cuadrático medio (ECM), el coeficiente de correlación de Pearson (ρ) entre los valores estimados ($\hat{g}_{\text{VR-O}}$, $\hat{g}_{\text{VR-B}}$ ó $\hat{g}_{\text{IBD-LD}}$) y los valores verdaderos de la proporción de genoma compartido (g_T). Los estimadores también fueron evaluados respecto del sesgo empírico, calculando la diferencia $\hat{g} - g_T$ para cada par de animales y promediándola sobre todos los pares. Por último, se calculó la regresión de los valores verdaderos de la proporción de genoma compartido en los valores estimados, como una medida de la proximidad entre los estimadores y las verdaderas relaciones.

4.5 Resultados

4.5.1 Datos reales

Se compararon la media y el desvío estándar (DE) general de la proporción estimada de genoma compartido en los datos reales, con sus valores teóricos (Cuadro 4.1), basados en los mapas genéticos porcinos. La media de las relaciones genómicas fue idéntica al verdadero valor cuando se utilizó la matriz $\boldsymbol{G}_{\text{VR-O}}$, dado que este estimador es escalado por construcción, de manera que las medias de los elementos diagonales y de los no-diagonales de $\boldsymbol{G}_{\text{VR-O}}$ sean iguales a las de la matriz $\boldsymbol{A}$ de relaciones aditivas (Vitezica *et al.*, 2011). La media general de los elementos de la matriz $\boldsymbol{G}_{\text{IBD-LD}}$ fue muy próxima al valor verdadero. El estimador que difirió más de la media teórica fue $\boldsymbol{G}_{\text{VR-B}}$. Con respecto al DE general de la proporción estimada de genoma compartido, el valor para $\boldsymbol{G}_{\text{IBD-LD}}$ estuvo más cerca del teórico que $\boldsymbol{G}_{\text{VR-O}}$ o $\boldsymbol{G}_{\text{VR-B}}$.

Cuadro 4.1: Media y desvío estándar (DE) general de las relaciones genómicas estimadas (N = 84.254) en una base de datos reales de cerdos.

	Teórica	$\boldsymbol{G}_{\textbf{IBD-LD}}$	$\boldsymbol{G}_{\textbf{VR-O}}$	$\boldsymbol{G}_{\textbf{VR-B}}$
Media	0,1062	0,1087	0,1062	0,1416
DE	0,1100	0,1090	0,0985	0,1273

$\boldsymbol{G}_{\textbf{IBD-LD}}$: matriz de relaciones genómicas basada en la noción de IBD; $\boldsymbol{G}_{\textbf{VR}}$: matriz genómica basada en la noción de IBS y construida con las frecuencias alélicas observadas ($\boldsymbol{G}_{\textbf{VR-O}}$) o con las frecuencias alélicas de los animales de la población base (F_0) ($\boldsymbol{G}_{\textbf{VR-B}}$).

Para la base de datos reales con registros genotípicos de cerdos, el patrón observado de la proporción de genoma compartido dentro de cada clase de parentesco genealógico fue similar en los tres estimadores: la media estimada disminuyó a medida que el parentesco genealógico fue más distante (Cuadro 4.2). Sin embargo, la media de $\boldsymbol{G}_{\textbf{IBD-LD}}$ fue más cercana al valor teórico en nueve de las catorce clases de parentesco genealógico evaluadas; $\boldsymbol{G}_{\textbf{VR-O}}$ fue el estimador más cercano al valor teórico promedio en las relaciones abuelo-nieto y medio-primos. Esta última relación involucra la anterior, ya que los medio-primos tienen un abuelo en común. Además, la media de la relación estimada entre medio-primos siguió el mismo patrón que la media general, y fue calculada con el mayor número de pares de animales. El estimador $\boldsymbol{G}_{\textbf{VR-B}}$ fue el que estuvo más cerca de las medias verdaderas en las relaciones tío-sobrino, medio tío-sobrino y doble medio primos. Nótese que la relación tío-sobrino puede pensarse como una relación medio tío-sobrino por dos vías, mientras que los dobles medio primos se pueden ver como los descendientes de cuatro pares de medio tío-sobrinos.

Cuadro 4.2: Tamaño y media muestral de la proporción estimada de genoma compartido en una base de datos reales de cerdos para diferentes relaciones de parentesco genealógico (N = 84.254).

Parentesco	N	$\boldsymbol{G}_{\mathbf{T}}$	$\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}O}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$
Padre-hijo	784	0,5000	0,5000	0,4299	0,4824
Hermanos enteros (HE)	639	0,5000	0,5046	0,4286	0,4886
MH, madres(padres) HE	816	0,3750	0,3730	0,3126	0,3588
MH, madres(padres) MH	2848	0,3125	0,3231	0,2522	0,2997
Abuelo-nieto	1344	0,2500	0,2067	0,2299	0,2709
Medio hermanos (MH)	7061	0,2500	0,2537	0,2185	0,2811
Tío-sobrino	1716	0,2500	0,2282	0,2279	0,2468
Doble primos hermanos	544	0,2500	0,2343	0,2193	0,3150
Triple medio primos	2912	0,1875	0,1754	0,1533	0,2197
Doble medio primos	5408	0,1250	0,1313	0,1076	0,1229
Medio tío-sobrino	6800	0,1250	0,1344	0,1216	0,1266
Primos	6960	0,1250	0,1169	0,1097	0,1780
Medio primos	22944	0,0625	0,0735	0,0585	0,1019
No relacionados	23478	0,0000	0,0000	0,0444	0,0599

$\boldsymbol{G}_{\mathbf{T}}$: verdadera matriz de relaciones genómicas (media teórica); $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$: matriz de relaciones genómicas basada en la noción de IBD; $\boldsymbol{G}_{\mathbf{VR}}$: matriz genómica basada en la noción de IBS y construida con las frecuencias alélicas observadas ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) o con las frecuencias alélicas de los animales de la población base (F_0) ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$).

El Cuadro 4.3 presenta los desvíos estándar muestrales (DE) de la proporción de genoma compartido y los valores teóricos para cada clase de parentesco genealógico. Para cada clase, el DE de las estimaciones con $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ fue siempre menor que el de las estimaciones obtenidas con $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ o $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$. En promedio, los DE utilizando $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ y $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ fueron 7,50%, 60,37% y 174,07% más elevados que el DE teórico para cada clase de parentesco genealógico, respectivamente. En consecuencia, el

solapamiento de las curvas de distribución de la proporción estimada de genoma compartido provenientes de clases de parentesco genealógico distintas fue mayor para las estimaciones basadas en IBS.

Cuadro 4.3: Desvío estándar muestral de la proporción estimada de genoma compartido en una base de datos reales de cerdos para diferentes relaciones de parentesco genealógico.

Parentesco	$\boldsymbol{G}_{\mathbf{T}}$	$\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}O}}$	$\boldsymbol{G}_{\mathbf{VR\text{-}B}}$
Padre-hijo	0,0000	0,0000	0,0573	0,1188
Hermanos enteros (HE)	0,0527	0,0578	0,0826	0,1317
MH, madres(padres) HE	0,0476	0,0478	0,0711	0,1180
MH, madres(padres) MH	0,0447	0,0438	0,0641	0,1086
Abuelo-nieto	0,0456	0,0465	0,0993	0,1454
Doble primos hermanos	0,0419	0,0472	0,0581	0,1017
Medio hermanos, MH	0,0373	0,0344	0,0609	0,0895
Tío-sobrino	0,0348	0,0361	0,0512	0,1204
Triple medio primos	0,0386	0,0420	0,0560	0,1038
Doble medio primos	0,0350	0,0385	0,0504	0,0862
Medio tío-sobrino	0,0335	0,0375	0,0465	0,1110
Primos	0,0297	0,0321	0,0535	0,0793
Medio primos	0,0248	0,0279	0,0495	0,0709
No relacionados	0,0000	0,0000	0,0651	0,0854

Los coeficientes de correlación de Pearson entre los valores estimados de la proporción de genoma compartido y su correspondiente coeficiente de relación aditiva basado en el pedigree fueron 0,959, 0,797 y 0,702 para $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ y $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$, respectivamente.

4.5.2 Datos simulados

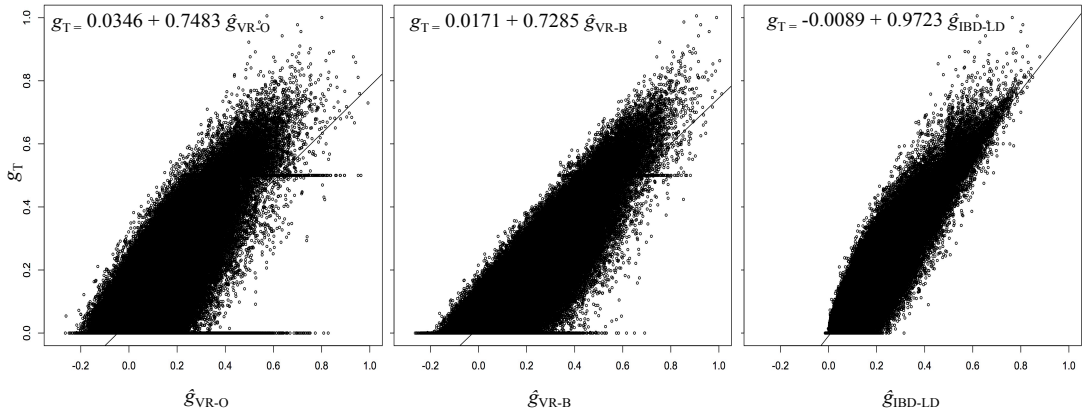
El Cuadro 4.4 resume la precisión y el sesgo empírico, promediados sobre el total de las réplicas, alcanzados por los tres estimadores ($\hat{g}_{\text{VR-O}}$, $\hat{g}_{\text{VR-B}}$ y $\hat{g}_{\text{IBD-LD}}$) de la verdadera proporción de genoma compartido IBD entre los genotipos simulados (g_T). Los tres estimadores tuvieron muy bajo sesgo empírico, siendo la matriz $\boldsymbol{G}_{\textbf{VR-B}}$ la menos insesgada de todas. La matriz $\boldsymbol{G}_{\textbf{IBD-LD}}$ presentó el valor más bajo de ECM empírico y el más alto de correlación con los valores verdaderos de la proporción de genoma compartido. Cuando se utilizaron las frecuencias alélicas de la población base, $\boldsymbol{G}_{\textbf{VR-B}}$ se aproximó a los valores de correlación de $\boldsymbol{G}_{\textbf{IBD-LD}}$, y tuvo mayores valores de correlación y menor ECM que $\boldsymbol{G}_{\textbf{VR-O}}$. El estimador $\boldsymbol{G}_{\textbf{VR-B}}$, aunque no siempre es posible calcularlo (dado que las frecuencias de la población base no siempre están disponibles), aseguró un mejor escenario. Finalmente, la última columna del Cuadro 4.4 y la Figura 4.3 muestran la regresión de las relaciones genómicas verdaderas sobre las relaciones genómicas estimadas. El coeficiente de regresión estuvo cerca de 1 para $\boldsymbol{G}_{\textbf{IBD-LD}}$, siendo significativamente menor para ambos estimadores $\boldsymbol{G}_{\textbf{VR}}$.

Cuadro 4.4: Performance de los estimadores de las relaciones genómicas sobre el total de las réplicas (elementos no-diagonales de la matriz) en la base de datos simulados.

	ECM(×100)	**Correlación de Pearson**	**Sesgo empírico**	$\boldsymbol{b_1}$*
$\boldsymbol{G}_{\textbf{VR-O}}$	0,9352 ± 0,2847	0,6779 ± 0,0482	-0,0086 ± 0,0095	0,7483
$\boldsymbol{G}_{\textbf{VR-B}}$	0,5703 ± 0,2059	0,8762 ± 0,0219	0,0180 ± 0,0185	0,7285
$\boldsymbol{G}_{\textbf{IBD-LD}}$	0,1886 ± 0,0535	0,9458 ± 0,0082	0,0122 ± 0,0091	0,9723

*b_1 es el coeficiente de regresión lineal de las relaciones genómicas verdaderas en las relaciones genómicas estimadas.

Figura 4.3: Regresión de los valores verdaderos (g_T) en los valores estimados ($\hat{g}_{VR\text{-}O}$, $\hat{g}_{VR\text{-}B}$ y $\hat{g}_{IBD\text{-}LD}$) de las relaciones genómicas en la base de datos simulados.



4.6 Discusión

Utilizando el enfoque basado en el estimador $\hat{g}_{VR}$, de los Campos *et al.* (2013) encontraron que la eficacia de la metodología BLUP de predicción del valor de cría genómico (GBLUP) depende fundamentalmente de la medida en que las relaciones genómicas derivadas de los marcadores reflejan los patrones de las relaciones genéticas realizadas en los loci causales. La investigación actual llevada a cabo en este capítulo ha intentado comparar el enfoque propuesto por VanRaden (2008) para estimar las relaciones de parentesco utilizando sólo marcadores ($\hat{g}_{VR}$), el cual ha sido y es ampliamente utilizado en SG, con el enfoque propuesto en el Capítulo 3 de esta disertación, que incorpora la información del pedigree y de la herencia además de los marcadores a la estimación de las relaciones de parentesco ($\hat{g}_{IBD\text{-}LD}$).

El conjunto de datos reales permitió comparar la variación empírica que existe en la proporción de genoma compartido entre pares de animales con la misma relación de parentesco genealógico, ya sea a través de estimadores basados en la noción de IBS o IBD. El DE de la proporción estimada de genoma compartido fue notablemente más cercano al valor teórico para $\hat{g}_{IBD\text{-}LD}$ que para $\hat{g}_{VR\text{-}O}$ o $\hat{g}_{VR\text{-}B}$. Por otro lado, resultó

extremadamente difícil distinguir las diferentes relaciones de parentesco genealógico a partir de los valores realizados de la fracción del genoma compartido estimados mediante el enfoque $\hat{g}_{VR}$. Si bien $\hat{g}_{VR}$ es una estimación del valor realizado de la proporción de genoma compartido IBD, no sólo no tiene en cuenta la transmisión de padres a hijos, sino que además no tiene en consideración la naturaleza segmentaria de la herencia del ADN (Thompsom, 2013). De hecho, si se permutaran los genotipos de cada SNP, no se modificarían los elementos de la matriz de relaciones genómicas $\boldsymbol{G}_{VR}$. El promedio de $\hat{g}_{IBD\text{-}LD}$ fue muy cercano al valor teórico para la mayoría de las relaciones de parentesco genealógico. El estimador $\hat{g}_{VR\text{-}O}$ resultó insesgado con respecto a la media general, sin embargo, no se comportó tan bien como el estimador $\hat{g}_{IBD\text{-}LD}$ cuando las comparaciones se realizaron dentro de una clase o parentesco genealógico particular. El estimador más sesgado fue $\hat{g}_{VR\text{-}B}$ (Cuadro 4.1), que tendió a sobreestimar la proporción de genoma compartido entre los animales del pedigree. En parte, esto puede explicarse por el hecho de que las frecuencias alélicas de base fueron calculadas partir de un número muy pequeño de animales que, además, pertenecían a dos razas distintas (4 machos Duroc y 15 hembras Pietrain), por lo que las estimaciones de las frecuencias alélicas de la población base sufrieron una falta de precisión. De hecho, $\hat{g}_{VR\text{-}B}$ fue el estimador más sesgado para las clases de los medio primos y de los pares de animales no relacionados, que representan el 27,2% y el 27,9% de las relaciones genómicas estimadas, respectivamente, y se espera que tengan el valor más bajo (o nulo) de la media teórica de la proporción de genoma compartido de una clase (Cuadro 4.2).

Los resultados de la simulación permitieron comparar la precisión y el sesgo empírico alcanzado por los diferentes estimadores de la verdadera proporción de genoma compartido IBD entre los animales genotipados en un pedigree complejo. El

estimador $\hat{g}_{\text{IBD-LD}}$ mostró una mayor precisión que $\hat{g}_{\text{VR-O}}$. Esto puede deberse a que $\hat{g}_{\text{VR-O}}$ no logró capturar correctamente las relaciones de parentesco históricas no observables que se dan (por lo general) en poblaciones pequeñas de ganado porcino, como la simulada en este capítulo, cuando se cuenta con una pequeña proporción de animales genotipados. Cuando se utilizaron las frecuencias alélicas en la población base (conocidas), se aseguró un mejor escenario, permitiendo que el estimador $\hat{g}_{\text{VR-B}}$ se aproximara en precisión al estimador $\hat{g}_{\text{IBD-LD}}$. Este resultado también concuerda con el hecho de que $\hat{g}_{\text{VR-B}}$ fue prácticamente insesgado en la simulación, en contraste con los resultados de los datos reales, donde las frecuencias alélicas de la población base no estuvieron bien representadas por las frecuencias alélicas de los animales genotipados en la generación F_0. Para remediar este problema VanRaden (2008) propuso estimar las frecuencias alélicas de la población mediante un modelo lineal que predice el conteo de alelos en los antepasados y descendientes no genotipados de los animales genotipados utilizando el pedigree (Gengler *et al.* 2007). Sobre las metodologías de estimación de las frecuencias alélicas se discutió en la sección 2.5 del Capítulo 2.

Los resultados de este capítulo apuntan a que la incorporación de la información de los registros genealógicos en el cálculo de las relaciones genómicas mejora la precisión de las estimaciones, sobre todo cuando se trabaja con familias grandes con muy pocos animales genotipados. En este escenario, las estimaciones de las relaciones genómicas basadas en la noción de IBS sólo se acercarían en precisión a las basadas en la noción de IBD si se dispone de estimaciones precisas de las frecuencias alélicas en la población base.

CAPÍTULO 5

Consecuencias de utilizar diferentes matrices de parentesco genómico en la exactitud de las predicciones de los valores de cría

Consecuencias de utilizar diferentes matrices de parentesco genómico en la exactitud de las predicciones de los valores de cría

5.1 Introducción

La disponibilidad de una tecnología de genotipado de alta densidad de SNPs, cuyo precio disminuye continuamente, ha impulsado la aplicación de técnicas de SG en animales domésticos (Meuwissen *et al.*, 2001). En general, las poblaciones de especies pecuarias poseen altos niveles de LD causados por procesos recurrentes de deriva, selección y cruzamiento entre razas (Haley, 1999). Se espera, entonces, que, dada una alta densidad de marcadores a lo largo del genoma, algunos de ellos estén en LD con al menos uno de los loci que influyen sobre un carácter cuantitativo (QTLs), y que, al ajustar un efecto para cada marcador, se pueda capturar información acerca de los QTLs (Meuwissen *et al.*, 2001). Los modelos de evaluación utilizados para la SG consisten en la predicción de los BVs de los animales a través de la predicción simultánea de miles de efectos, cada uno atribuible a un SNP. Dichas predicciones pueden computarse tan pronto se obtiene la información del ADN; de este modo, la SG tiene el potencial de aumentar la tasa de ganancia genética, reducir el intervalo generacional, y aumentar la exactitud en las predicciones de los BVs a edad temprana (Schaeffer, 2006). Esta técnica ha sido adoptada en forma masiva en el mejoramiento genético de ganado lechero (VanRaden *et al.*, 2009).

Por razones de costo y logística, es imposible que todos los animales de una población sean genotipados. Los individuos con genotipo evaluado son, en general, toros de alta difusión; también promisorios toros jóvenes y, posiblemente, hembras potenciales candidatas a la selección (por ejemplo, madres para transferencia embrionaria, Legarra *et al.*, 2009). Actualmente las evaluaciones genómicas se realizan

por un procedimiento en varias etapas (en ingles *multiple step procedure*; VanRaden, 2008; VanRaden *et al.*, 2009): 1) evaluación genética tradicional de todos los animales del pedigree mediante el MA, 2) cálculo de las predicciones de los valores genómicos directos (en inglés *direct genomic values*, DGVs) para un número reducido de animales genotipados a partir de sus fenotipos corregidos, i.e. las DYDs o YDs (*daughter yield deviations* o *yield deviations*; VanRaden y Wiggans, 1991) y 3) predicción de los valores de cría genómicos (en inglés *genomic estimated breeding values*, GEBVs) para todos los animales mediante un índice de selección. Este índice combina las predicciones genómicas del paso 2) con las predicciones de los BVs obtenidas con el MA (*estimated breeding values*, EBVs). Los DGVs pueden obtenerse estimando los efectos individuales de cada SNP primero y luego sumándolos (Meuwissen *et al.*, 2001), o mediante la utilización de ecuaciones de modelo mixto con una matriz de relaciones genómicas en lugar de la matriz de relaciones aditivas (VanRaden, 2008). Ambos modelos son equivalentes (Hayes *et al.*, 2007; Strandén y Garrick, 2009) y pueden diferir en su exactitud por sus distintos métodos de cálculo. Si bien este procedimiento tiene la ventaja de no producir modificaciones en la evaluación genética tradicional que se realiza en forma periódica, la principal desventaja es que sus etapas son propensas a sesgos y pérdida de información (Legarra *et al.*, 2009). La información de los animales con registro fenotípico pero que no han sido genotipados no es aprovechada de manera óptima debido a que la misma se utiliza como parte del DYD en el paso 2) y/o como parte del EBV en el paso 3). Esto hace que, o bien el uso de esta información resulte fragmentado o, si sólo se utiliza en el paso 3), los fenotipos de los animales no genotipados no contribuyen a las predicciones de los valores genómicos directos (Meuwissen *et al.*, 2011). Asimismo, en el uso de las DYDs y YDs, existen problemas de ponderación (causados por la diferente cantidad de información en la base

de datos original), sesgos (causados por la selección, por ejemplo), disminución de la exactitud (en los animales de rodeos pequeños), y colinealidad (por ejemplo, el YD de dos vacas en el mismo rodeo, Legarra *et al.*, 2009).

Una forma de simplificar estos pasos es modificar la matriz $\boldsymbol{A}$ de relaciones aditivas entre los animales (genotipados y no genotipados) dentro de la evaluación genética tradicional de modo de incluir la información genómica y producir una única evaluación genética empleando ambas fuentes de información: pedigree y marcadores moleculares. De este modo, se utiliza la misma metodología de la evaluación genética tradicional, es decir las ecuaciones de modelos mixtos de Henderson (1988), y se incorpora toda la información fenotípica, de pedigree y genómica disponible en un solo paso (*single-step procedure*). Para ello, Misztal *et al.* (2009), Legarra *et al.* (2009), Aguilar *et al.* (2010), e independientemente Christensen y Lund (2010) propusieron modificaciones a la matriz $\boldsymbol{A}$, condicionando el valor genético de los animales no genotipados en el valor genético de animales genotipados vía un índice de selección, y utilizando una matriz de relaciones genómicas $\boldsymbol{G}$ para estos últimos. La matriz $\boldsymbol{G}$ se obtiene según la metodología de VanRaden (2008). La distribución de los valores de cría de los animales no genotipados, condicional a los BVs de los animales genotipados, es calculada por medio de una regresión cuyos coeficientes están basados en el pedigree y en las propiedades de la distribución normal (Sorensen y Gianola, 2002, p.254; Legarra *et al.*, 2009).

La extraordinaria velocidad con la que se han dado los avances tecnológicos en SG dificulta la elaboración de un marco teórico sólido para aplicar estas nuevas técnicas, así como la comprensión de los métodos estadísticos propuestos hasta el momento (Gianola *et al.*, 2009). En este punto, una cuestión de importancia es cómo impactan en las predicciones genómicas los supuestos elaborados a nivel genético sobre

las distribuciones a priori de los efectos de los marcadores. Las fórmulas empleadas, por ejemplo en Meuwissen *et al.* (2001) y VanRaden (2008), asumen que los efectos de los marcadores son independientes y que aquellos son variables aleatorias normales e idénticamente distribuidas, lo cual implica LE entre los marcadores. Sin embargo, el supuesto de LE es violado cuando la densidad de los mismos es alta, sumado a que la teoría SG descansa en el supuesto de existencia de LD (Gianola *et al.*, 2009). Asimismo, la matriz ***G*** descripta por VanRaden (2007, 2008), asume a priori una población base genotipada sin selección; sin embargo, en la práctica, los individuos genotipados pertenecen a un subgrupo de la población que ha sido sujeto a un proceso de selección. Varios autores han propuesto distintas modificaciones de la matriz original de VanRaden (2007, 2008) con el objeto de corregir el sesgo y aumentar la exactitud en las predicciones; por ejemplo, asumiendo diferentes valores para las frecuencias alélicas de los marcadores, desplazándola y reescalándola, o eliminando los SNP cuyas frecuencias sean cercanas a uno en la población (VanRaden, 2008; Aguilar *et al.*, 2010; Forni *et al.*, 2011; Vitezica *et al.*, 2011; Chen *et al.*, 2011).

Un abordaje alternativo sería calcular una matriz ***G*** entre individuos genotipados que tuviera en cuenta el efecto de la selección previa mediante cuantificar el pasaje de genes (o marcadores) entre generaciones, considerando todas las relaciones de parentesco posibles entre los individuos de un pedigree. A tal efecto, los marcadores pueden ser utilizados para estimar los valores de la proporción de alelos IBD entre pares de individuos que realmente ocurrieron, promediados sobre todas las posiciones del genoma (Villanueva *et al.*, 2005; Visscher *et al.*, 2006). Esta propuesta resulta impráctica a priori, dado que sería necesario contar con genotipos para familias enteras (Legarra *et al.*, 2009). Sin embargo, muy recientemente se han desarrollado algoritmos que modelan directamente la relación de parentesco entre un par de individuos

genotipados, sin necesidad de enumerar cada uno de los posibles genotipos y patrones de herencia de sus ancestros en el pedigree no genotipados (por ejemplo, Li *et al.*, 2010). Asimismo, estos algoritmos consideran el LD entre marcadores que surge como consecuencia de las relaciones de parentesco lejanas y desconocidas. En el Capítulo 3, se propuso utilizar una matriz genómica alternativa para los animales genotipados ($\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$), cuyos elementos representan los valores estimados de la proporción realizada de alelos IBD entre pares de individuos, promediados sobre todas las posiciones del genoma. En el Capítulo 4, esta matriz demostró ser más precisa que aquella calculada en base al concepto de IBS. Este capítulo aborda la implementación algorítmica de la matriz genómica de parentesco $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$, propuesta en el Capítulo 3, dentro de la evaluación genética animal y en un escenario en el cual no todos los animales que entran en la evaluación tienen genotipo.

5.2 Objetivo

El objetivo general de este capítulo es evaluar cómo el aumento en la precisión de las estimaciones de las relaciones genómicas de parentesco se podría traducir en un aumento en la exactitud de la predicción de los BVs genómicos, dependiendo del estimador de las relaciones utilizado. El supuesto detrás de este objetivo es que, en el "verdadero modelo", la matriz $\boldsymbol{G}$ de varianzas y covarianzas de los BVs genómicos está dada por los valores realizados de la proporción del genoma IBD compartida por cada par de individuos. Los objetivos específicos son:

1. Revisar los aspectos fundamentales que hacen a la derivación de $\boldsymbol{G}$ en los modelos de evaluación genética vigentes y derivar una matriz $\boldsymbol{G}$ a partir de las relaciones genómicas de parentesco estimadas mediante el método desarrollado en el Capítulo 3, que tiene en cuenta el pedigree y el LD entre marcadores ($\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$);

2. Incorporar la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ en un modelo de evaluación genética y emplear la simulación estocástica del Capítulo 4 para cuantificar su impacto sobre las exactitudes de las predicciones genómicas de los BVs de los animales candidatos a la selección. Las exactitudes calculadas empleando el modelo consistente con la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ serán comparadas con aquellas que se obtienen con la matriz propuesta por VanRaden (2008) ($\boldsymbol{G}_{VR}$).

5.3 Materiales y métodos

Este capítulo es de naturaleza esencialmente teórica y los resultados empíricos son obtenidos a partir de los datos simulados en el Capítulo 4. La base de datos simulada corresponde una estructura simplificada de una población núcleo porcina, compuesta por 10.220 animales con genotipos de alta densidad (imitando el panel comercial de 60K SNPs) distribuidos en cinco generaciones. La descripción detallada de la simulación se encuentra en la sección 4.3.2 del Capítulo 4 y los parámetros se resumen en el Cuadro 5.1.

Cuadro 5.1 Parámetros de la simulación

Número de réplicas	50	**Estructura de la población**	
Genoma		***Paso 1:*** *Generaciones históricas (GH)*	
Nro. de pares de cromosomas	5	Nro. de generaciones[tamaño] - etapa 1	2500[500]
Largo total	800 cM	Nro. de generaciones[tamaño] - etapa 2	30[75]
Nro. de SNPs	16.000	Nro. de generaciones[tamaño] - etapa 3	1[220]
Distribución de los SNPs	Aleatoria	***Paso 2:*** *Generaciones recientes*	
Distancia media entre SNPs	0,05 cM	Nro. de machos[hembras] fundadores de la última GH	20[220]
MAF mínimo	0,01	Nro. de generaciones (discretas)	5
Tasa de mutación	0,0002	Tamaño de camada [machos : hembras]	10[1:1]

Consideraremos primero un escenario hipotético sencillo en el cual cada animal es medido a lo sumo una vez para un carácter de interés. Asumiremos, además, que cada animal ha sido genotipado utilizando un chip de alta densidad de SNPs. Las predicciones de los BVs pueden computarse mediante el siguiente modelo lineal mixto:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e} \qquad [5.1]$$

donde $\boldsymbol{y}$ es un vector de orden $n \times 1$ de registros fenotípicos para un carácter, $\boldsymbol{\beta}$ es un vector $p \times 1$ de efectos fijos, $\boldsymbol{X}$ es una matriz de incidencia, $\boldsymbol{a} = \{a_i\}$ es un vector aleatorio de orden $q \times 1$ que contiene los BVs de los animales, $\boldsymbol{Z}$ es una matriz de incidencia que relaciona los elementos de $\boldsymbol{a}$ con los de $\boldsymbol{y}$ ($\boldsymbol{Z} = \boldsymbol{I}$ cuando todos los animales en $\boldsymbol{a}$ tienen registro), y $\boldsymbol{e}$ es el vector $n \times 1$ de errores. Se asume que

$$\text{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta} \qquad \text{Var}(\boldsymbol{a}) = \boldsymbol{G} \qquad \text{Var}(\boldsymbol{e}) = \boldsymbol{I}\sigma^2_e \qquad \text{Cov}(\boldsymbol{a}, \boldsymbol{e}') = \boldsymbol{0}$$

La matriz de varianzas y covarianzas de los BVs, $\boldsymbol{G}$, se construye en función de una serie de supuestos genéticos que dependen del modelo de herencia que se adopte. Actualmente, en las evaluaciones genéticas se utilizan dos modelos de herencia: el infinitesimal de herencia poligénica, que es la base del MA utilizado en las evaluaciones genéticas tradicionales, y el modelo que propuso VanRaden (2007, 2008) en SG. En este capítulo proponemos derivar $\boldsymbol{G}$ utilizando la información del pedigree y de los marcadores de un modo diferente. Los argumentos teóricos que respaldan esta propuesta serán desarrollados en las siguientes secciones. Siguiendo el artículo de Gianola *et al.* (2009), en las sección 5.3.1 y 5.3.2, se revisarán los aspectos fundamentales que hacen a la derivación de $\boldsymbol{G}$ en los dos modelos vigentes. Quedarán expuestos los supuestos detrás de los modelos de predicción de los BVs en el contexto clásico y en el contexto de la SG a partir de paneles densos de marcadores. En la sección 5.3.3, se derivarán los elementos de $\boldsymbol{G}$ a partir del método propuesto en la

sección 3.3.3 del Capítulo 3 de esta tesis. En la sección 5.3.4 se extenderá la matriz ***G*** propuesta a los animales no genotipados. Finalmente, en la sección 5.3.5 se plantearán las ecuaciones de modelos mixtos y se evaluará, de manera analítica y mediante simulación estocástica, el impacto de la matriz ***G*** propuesta en la exactitud de las predicciones genómicas de los BVs.

5.3.1 Modelo infinitesimal de herencia poligénica

En esta sección describiremos brevemente el modelo infinitesimal, siguiendo el enfoque de Gianola *et al.* (2009) quienes parten del modelo de "un locus" presentado por Falconer y Mackay (1996). Considérese un carácter afectado por un número muy grande ($K \to \infty$) de loci autosómicos bialélicos. Se asume que las frecuencias génicas, y genotípicas se encuentran en equilibrio Hardy-Weinberg (H-W) y ausencia de epistasis. Se define el BV del i-ésimo animal, a_i, como la suma de los efectos aditivos de todos los loci que gobiernan el carácter. En ausencia de dominancia,

$$a_i = \sum_{k=1}^{K} w_{ik}\, u_k \; - \; \sum_{k=1}^{K} \left(p_k - q_k \right) u_k , \qquad [5.2]$$

El primer término a la derecha de la igualdad corresponde al valor genotípico del animal i y el segundo a una constante asociada con la media poblacional del valor genotípico. La variable w_{ik} toma valores de 1, 0 ó −1, con probabilidad p_k^2, $2p_kq_k$ y q_k^2, respectivamente, según el genotipo en el k-ésimo locus del animal i sea A_1A_1, A_1A_2 ó A_2A_2; $E(w_{ik}) = p_k - q_k$, donde p_k y $q_k = 1 - p_k$ son las frecuencias alélicas poblacionales para ese locus; w_{ik} es una variable no observable; u_k es el efecto de sustitución para el k-ésimo locus y es considerado un efecto fijo.

En una población bajo equilibrio H-W y LE, los distintos loci segregan en forma independiente, dentro y entre individuos, i.e. cov (w_{ik} , w_{jl}) = 0, $\forall$ i, j = 1, 2, ..., q ,

donde *k* y *l* son dos loci distintos. De este modo, teniendo en cuenta la expresión 5.2,

$$\mathrm{Var}\left(a_i \middle| u_1,...u_K\right) = \mathrm{Var}\left(\sum_{k=1}^{K} w_{ik}u_k \middle| u_1,...u_K\right) = \sum_{k=1}^{K} \mathrm{Var}\left(w_{ik}\right)u_k^2 = \sum_{k=1}^{K} 2p_kq_ku_k^2 = \sigma_A^2 \qquad [5.3]$$

donde σ_A^2 es la varianza genética aditiva. Siguiendo el mismo razonamiento para la covarianza entre los BVs de dos animales, *i* y *j*, de la misma población,

$$\mathrm{cov}\left(a_i, a_j \middle| u_1,...,u_K\right) = \mathrm{cov}\left(\sum_{k=1}^{K} w_{ik}u_k, \sum_{k=1}^{K} w_{jk}u_k \middle| u_1,...,u_K\right) = \sum_{k=1}^{K} \mathrm{cov}\left(w_{ik}, w_{jk}\right)u_k^2 \qquad [5.4]$$

Analicemos el término cov (w_{ik}, w_{jk}), correspondiente al locus *k*. Sea i_S el alelo que recibió el animal *i* de su padre e i_D el alelo que recibió *i* de su madre. La covarianza entre los genotipos de los animales *i* y *j* será distinta de 0 si poseen al menos un par de genes (i_S y j_S, i_S e i_D, i_D y j_S, i_D y j_D) originados en algún ancestro en común, y por lo tanto IBD. Esto es

$$\begin{aligned} \mathrm{cov}\left(w_{ik}, w_{jk}\right) &= \mathrm{cov}\left(w_{ik}, w_{jk} \middle| i \equiv j\right)P\left(i \equiv j\right) + \mathrm{cov}\left(w_{ik}, w_{jk} \middle| i \not\equiv j\right)\left[1 - P\left(i \equiv j\right)\right] \\ &= \mathrm{cov}\left(w_{ik}, w_{jk} \middle| i \equiv j\right)r_{ij} + 0\left(1 - r_{ij}\right) \end{aligned} \qquad [5.5]$$

donde $P(i \equiv j) = r_{ij}$ es el coeficiente de coancestría entre *i* y *j* (Malécot, 1948), definido como la probabilidad de que un gen muestreado al azar en *i* sea IBD a un gen muestreado al azar en *j*. Se puede demostrar (ver Apéndice II, sección II.1) que

$$\mathrm{cov}\left(w_{ik}, w_{jk} \middle| i \equiv j\right) = 4p_kq_k \qquad [5.6]$$

Reemplazando en la ecuación 5.4 con las expresiones en 5.5 y 5.6, se llega a la expresión conocida para la covarianza entre los BVs de dos animales

$$\mathrm{cov}\left(a_i, a_j \middle| u_1,...,u_K\right) = \sum_{k=1}^{K} 4p_kq_kr_{ij}u_k^2 = \boldsymbol{A}_{ij}\sum_{k=1}^{K} 2p_kq_ku_k^2 = \boldsymbol{A}_{ij}\sigma_A^2 \qquad [5.7]$$

donde $\boldsymbol{A}_{ij} = 2r_{ij}$ es la relación aditiva entre los animales *i* y *j*. En forma matricial,

$$\mathrm{Var}(\boldsymbol{a}) = \boldsymbol{A}\,\sigma^2_{\mathrm{A}} \qquad [5.8]$$

Por lo tanto, la matriz de varianzas y covarianzas de los BVs involucra a los efectos de sustitución alélica, u_k, si bien éstos quedan absorbidos en σ^2_{A}.

5.3.2 Modelo de VanRaden (2008)

Para describir el modelo de herencia propuesto por VanRaden (2008) para realizar SG consideraremos un número *K* finito de marcadores de tipo SNP distribuidos a lo largo del genoma. En cada posición, cada SNP tiene dos alelos posibles: 1 ó 2, dando como resultado 3 genotipos: 11, 12 ó 22. A diferencia del modelo de un locus, estos genotipos son observables, dada la información que brinda el genotipado de los animales. La teoría de la SG asume que cada marcador está en fuerte LD con al menos un QTL (Meuwissen *et al.*, 2001). Siguiendo esta idea, VanRaden (2008) adoptó la aproximación de que cada locus marcador es un QTL. Utilizando una notación ligeramente diferente a la ecuación 5.2 del modelo anterior, el BV del *i*-ésimo animal se aproxima mediante la expresión

$$a_i^* = \sum_{k=1}^{K}\left[w_{ik} - (p_k - q_k)\right]u_{mk} = \sum_{k=1}^{K} m_{ik}\,u_{mk} \qquad [5.9]$$

Los elementos w_{ik} son conocidos e iguales a -1, 0 ó 1, según el genotipo observado para el marcador *k* sea 11, 12 ó 22; u_{mk} es el efecto de sustitución alélica para el *k*-ésimo marcador. Matricialmente, el vector $q \times 1$ de BVs se puede escribir como

$$\boldsymbol{a}^* = (\boldsymbol{W} - \boldsymbol{P})_{(q \times K)}\,\boldsymbol{u_m}_{(K \times 1)} = \boldsymbol{M}\,\boldsymbol{u_m} \qquad [5.10]$$

$\boldsymbol{P}$ es una matriz de orden $q \times K$ en la cual, cada elemento de la columna *k*, es igual $p_k - q_k$. Los valores p_k y q_k son las frecuencias alélicas para el *k*-ésimo marcador en la población base, en ausencia de selección. Es decir, $\boldsymbol{P} = E(\boldsymbol{W})$ bajo H-W. De este modo,

M es una matriz centrada por las frecuencias alélicas, que fija la esperanza de los BVs en 0.

A diferencia del modelo infinitesimal, los efectos u_{mk} (k = 1, 2, …, K) constituyen una muestra aleatoria de tamaño K, tal que $u_{mk} \sim (\theta, \sigma^2_u)$. Dando al enfoque un sentido Bayesiano, el parámetro σ^2_u representa la incertidumbre acerca del verdadero, pero desconocido, valor del efecto de sustitución de un marcador específico. Por ejemplo, $\sigma^2_u = 0$ significa en un sentido Bayesiano, que $u_m = \theta$ con total certeza. No significa, sin embargo, que el locus no posee un efecto, dado que θ puede (o no) ser igual a 0 (Gianola *et al.*, 2009). Concretamente, este modelo asume que $\theta = 0$.

La matriz de varianzas y covarianzas de los BVs es condicional a los genotipos observados para los marcadores, w_{ik}, y es igual a

$$\mathrm{Var}\left(\boldsymbol{a}^* \middle| \boldsymbol{W}\right) = \mathrm{Var}\left(\boldsymbol{M}\,\boldsymbol{u_m}\right) = \boldsymbol{M}\,\boldsymbol{M}'\,\sigma^2_u = \boldsymbol{M}\,\boldsymbol{M}'\,\frac{\sigma^2_{\mathrm{A}}}{\sum_{k=1}^{K} 2p_k q_k} = \boldsymbol{G}_{\mathbf{VR}}\,\sigma^2_{\mathrm{A}} \qquad [5.11]$$

La relación entre la varianza de los efectos de sustitución de los marcadores, σ^2_u, y la varianza genética aditiva en el modelo infinitesimal, σ^2_{A}, ha sido presentada por Habier *et al.* (2007) y Gianola *et al.* (2009) y depende de los supuestos del modelo. Algunos años antes, Broman (2001) encontró que la covarianza entre las frecuencias alélicas de individuos emparentados depende la relación IBD entre ellos. Asumiendo el modelo de un locus para un número finito K de loci marcadores, la varianza genética aditiva es igual a

$$\mathrm{Var}\left(a_i^* \middle| u_{m1}, u_{m2}, \ldots, u_{mK}\right) = \sum_{k=1}^{K} 2p_k q_k u^2_{mk} \qquad [5.12]$$

análoga a la expresión 5.3. Si tomamos $\sigma^2_{\mathrm{A}}{}'$ como la varianza aditiva promedio, que surge de tomar el valor esperado de la varianza respecto de la distribución de los u_{mk},

tenemos que

$$\sigma_A^2{}' = E_{u_m}\left(\sum_{k=1}^{K} 2p_k q_k u_{mk}^2\right) = \left(\sum_{k=1}^{K} 2p_k q_k \; E_{u_m}\left(u_{mk}^2\right)\right) = \left(\sigma_u^2 + \theta^2\right)\sum_{k=1}^{K} 2p_k q_k \qquad [5.13]$$

Además, si θ = 0, entonces

$$\sigma_u^2 = \frac{\sigma_A^2{}'}{\sum_{k=1}^{K} 2p_k q_k} \qquad [5.14]$$

La fórmula 5.14 es la relación utilizada para llegar a la expresión 5.11. Analicemos ahora la conexión entre la matriz ***A*** de relaciones aditivas y la matriz que acompaña a σ_A^2 en la expresión 5.11. Si tomamos el valor esperado de la matriz ***MM′*** se tiene que

$$\begin{aligned} \mathrm{E}(\boldsymbol{MM'}) &= \mathrm{E}\left[(\boldsymbol{W}-\boldsymbol{P})(\boldsymbol{W}-\boldsymbol{P})'\right] \\ &= \mathrm{E}(\boldsymbol{WW'}) - \mathrm{E}(\boldsymbol{W})\boldsymbol{P'} - \boldsymbol{P}\,\mathrm{E}(\boldsymbol{W'}) + \boldsymbol{PP'} \\ &= \mathrm{E}(\boldsymbol{WW'}) - \boldsymbol{PP'} \end{aligned} \qquad [5.15]$$

El elemento *ij* de esta matriz es igual a

$$\begin{aligned} E\left(\boldsymbol{m}_i'\boldsymbol{m}_j\right) &= E\left(\boldsymbol{w}_i'\boldsymbol{w}_j\right) - \boldsymbol{p}_i'\boldsymbol{p}_j = \sum_{k=1}^{K} E\left(w_{ik}w_{jk}\right) - \sum_{k=1}^{K}\left(p_k - q_k\right)^2 \\ &= \sum_{k=1}^{K}\left[E\left(w_{ik}w_{jk}\right) - E\left(w_{ik}\right)^2\right] = \sum_{k=1}^{K}\mathrm{cov}\left(w_{ik}, w_{jk}\right) = \boldsymbol{A}_{ij}\sum_{k=1}^{K} 2p_k q_k \end{aligned} \qquad [5.16]$$

donde *i*, *j* = 1, 2, …, *n*, tal como se describe en Broman (2001) para un locus autosómico, y por Habier *et al.* (2007) y Gianola *et al.* (2009) para *K* loci. De las expresiones 5.15 y 5.16 se desprende que $E\left(\frac{\boldsymbol{MM}'}{\sum_{k=1}^{K} 2p_k q_k}\right) = \boldsymbol{A}$. Por lo tanto, $\frac{\boldsymbol{MM}'}{\sum_{k=1}^{K} 2p_k q_k}$ es un estimador insesgado de la matriz de relaciones aditivas y refleja las relaciones observadas u ocurridas, antes que las relaciones esperadas, teniendo en cuenta el residuo de segregación mendeliano (i.e., puede distinguir entre hermanos enteros) y relaciones desconocidas o muy lejanas (Legarra *et al.*, 2009).

Ahora bien, tal como explicaron Gianola *et al.* (2009), a menos que todos los marcadores sean verdaderamente los QTLs, $\sigma^2_A{}'$ representa sólo la fracción de la varianza genética aditiva capturada por los marcadores. Asimismo, la expresión 5.12 surge de muestrear conjuntamente genotipos (no sus efectos) en *K* loci marcadores en LE; esto es una condición necesaria para llegar a la expresión 5.14. Cuando se considera el LD entre marcadores, se generan covarianzas entre los genotipos de diferentes loci y la varianza genética aditiva (asumiendo H-W en cada marcador) se transforma en

$$\mathrm{Var}_D\left(a_i^* \middle| u_{m1},...,u_{mK}\right) = \mathrm{Var}\left(\sum_{k=1}^{K} w_{ik}u_{mk}\right) = \sum_{k=1}^{K}\mathrm{Var}\left(w_{ik}\right)u_{mk}^2 + 2\sum_{k=1}^{K}\sum_{l>k}^{K} 2D_{kl}u_{mk}u_{ml} \quad [5.17]$$

donde $D_{kl} = P(\mathrm{AB})_{kl} - P(\mathrm{A})_k P(\mathrm{B})_l$ es el estadístico relacionado con el desequilibrio de ligamiento involucrando dos loci, *k* y *l*, y AB es el haplotipo de la gameta del individuo *i*. El primer término en la expresión 5.17 es la varianza aditiva bajo LE; el segundo término es una contribución a la varianza del LD, y puede ser negativo o positivo. La media marcada para la varianza aditiva bajo LD, $\sigma^2_A{}'_{(D)}$, se transforma en

$$\sigma^2_A{}'_{(D)} = E_{u_m}\left[\mathrm{Var}\left(a_i^* \middle| \boldsymbol{u_m}\right)_{(D)}\right] = \left(\sigma_u^2 + \theta^2\right)\sum_{k=1}^{K} 2p_k q_k + 2\theta^2\sum_{k=1}^{K}\sum_{l>k}^{K} 2D_{kl} \quad [5.18]$$

Únicamente cuando $\theta = 0, \sigma^2_A{}'_{(D)} = \sigma^2_A{}'$. En ese caso, el LD no afectaría la relación de la expresión 5.14. Sin embargo, tal como explicaron Gianola *et al.* (2009) no existe una base mecanística para esperar que todos los marcadores tengan los mismos efectos, ni que estén idénticamente distribuidos o sean mutuamente independientes. Es decir, es esperable que algunos de estos efectos sean iguales a 0, sobre todo si u_{mk} representa el efecto de sustitución de un marcador ubicado en una región del genoma sin QTLs para el carácter; contrariamente, si un marcador se encuentra en fuerte LD con un QTL, no es esperable que su efecto sea nulo. Por lo tanto, es necesario incorporar parámetros relacionados al LD entre marcadores a la hora de estimar una matriz de varianzas y

covarianzas de los BVs utilizando la información de chips de alta densidad de SNPs.

En resumen, en esta sección, hemos revisado la teoría detrás de la matriz de varianzas y covarianzas de los BVs de los animales genotipados utilizada por VanRaden (2008) . En su derivación se asumen los mismos supuestos que los de un modelo infinitesimal (donde se separa un efecto aleatorio asociado con la incertidumbre del genotipo del QTL desconocido y su efecto de sustitución), i.e., equilibrio H-W para cada loci marcador y LE entre marcadores. De cualquier manera, se debe tener presente que se trata de un número finito de marcadores, y no de genes o QTLs.

5.3.3 Modelo incorporando la proporción de genes IBD en el genoma

En esta sección proponemos derivar la covarianza entre los BVs de los animales, incorporando la información de los marcadores de modo de no ignorar el LD. Para ello, volveremos sobre algunas ideas presentadas en las secciones anteriores y en los Capítulos 3 y 4. En la sección 5.3.1, se derivó la covarianza genética aditiva para el modelo infinitesimal, de forma tal que

$$\mathrm{cov}\left(a_i, a_j \,|\, u_1, u_2, ..., u_K\right) = \boldsymbol{A}_{ij}\,\sigma^2_{\mathrm{A}}$$

donde $\boldsymbol{A}_{ij}$ es el coeficiente de relación aditiva entre los individuos i y j, y σ^2_{A} es la varianza genética aditiva. Siguiendo el enfoque de Bulmer (1985), la relación aditiva es el valor esperado de la proporción de alelos IBD entre un par de individuos. Sea Z la proporción de alelos IBD entre dos individuos en un locus al azar en el genoma, asumiendo ausencia de consanguinidad. Debido a la segregación mendeliana, Z vale 0, 0,5 ó 1, según el número de pares de alelos IBD sea 0, 1 ó 2, con probabilidad P_0, P_1 y P_2, respectivamente. Las probabilidades de identidad dependen, incondicionalmente, de la relación de parentesco y se derivan a partir del pedigree, utilizando teoría de probabilidad. Luego, $E(Z) = \frac{1}{2}P_1 + P_2 = A_{ij}$. En presencia de consanguinidad,

$E(Z) = 2\Delta_1 + (\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{2}\Delta_8 = A_{ij}$, donde Δ_1, Δ_3, Δ_5, Δ_7 y Δ_8 son los valores de algunos de los nueve "coeficientes de identidad condensados" de Jacquard (1974, ver Capítulo 4, Figura 4.2). De lo anterior se desprende que *Z* es una medida de identidad genética basada en un solo locus.

Consideremos ahora una medida de identidad genética basada en todo el genoma. Sea *K* el número de loci influyendo sobre un carácter cuantitativo, se define la variable aleatoria R_{ij} como la proporción de genes IBD para un par de individuos, *i* y *j*, tal que

$$R_{ij} = \frac{1}{K}\sum_{k=1}^{K} Z_k \qquad 0 < R_{ij} < 1 \qquad [5.19]$$

Al igual que *Z*, el valor esperado de R_{ij} es

$$E(R_{ij}) = \frac{1}{K}\sum_{k=1}^{K} E(Z_k) = \boldsymbol{A}_{ij} \qquad [5.20]$$

Si el número de loci fuera infinitamente grande y éstos segregaran independientemente, entonces, no habría diferencia entre la proporción de genes IBD, R_{ij}, y su valor esperado, $\boldsymbol{A}_{ij}$, para un par de individuos. Esto se debe a que

$$\begin{aligned} \lim_{K\to\infty} \operatorname{Var}(R_{ij}) &= \lim_{K\to\infty} \frac{1}{K^2}\left[\sum_{k=1}^{K}\operatorname{Var}(Z_k) + 2\sum_{k<l}^{K}\operatorname{cov}(Z_k, Z_l)\right] \\ &= \lim_{K\to\infty} \frac{1}{K^2}\left[K\operatorname{Var}(Z_k) + 0\right] = 0 \end{aligned} \qquad [5.21]$$

Sin embargo, los genes que influyen sobre un carácter se sitúan en un número finito, relativamente pequeño de cromosomas y no segregan independientemente debido al ligamiento. Por lo cual, uno esperaría encontrar variación en la proporción de genes IBD en el genoma entre pares de individuos con el mismo valor de relación aditiva $\boldsymbol{A}_{ij}$ (Visscher, 2009; Weir y Hill, 2011). En efecto, a igual valor de $\boldsymbol{A}_{ij}$, uno esperaría que aquellos pares de individuos que compartan más alelos IBD en los QTLs tengan valores

genotípicos más parecidos, lo que se traduciría en una covarianza genética mayor. Luego, R_{ij} podría visualizarse como el valor que ocurrió (*realised*) para la relación aditiva entre un par de individuos (Visscher, 2009).

La varianza de R_{ij} ha sido obtenida por varios autores (Hill, 1993 a, b; Guo, 1994, 1995, 1996; Visscher *et al.*, 2006; Hill y Weir, 2011) para diferentes tipos de relaciones de parentesco. Si bien no entraremos en los detalles de la derivación, asumiendo que los diferentes cromosomas segregan independientemente, y cuando el número de loci se hace muy grande, la varianza teórica de R_{ij} puede expresarse en función del largo total del genoma (L) e igual a $L = \sum_{i=1}^{m} l_i$, del número de cromosomas (m) y del largo de cada cromosoma (l_i). El cuadro 5.2 muestra algunas de las fórmulas obtenidas por Guo (1996), coincidiendo con las de los demás autores antes mencionados. Nótese que dada su concepción, R_{ij} constituye una medida de identidad genética de menor variabilidad que Z y, por lo tanto, mucho más precisa.

Cuadro 5.2 Varianza de la proporción del genoma IBD para varios tipos de relaciones de parentesco. Modificado de Guo (1996).

Parentesco	Un locus		Todo el genoma
	E (Z)	Var(Z)	Var (R_{ij})
Hermanos enteros	1/2	1/8	$\frac{1}{128\mathrm{L}^2}\left(4\mathrm{L}-\mathrm{m}+\sum_{i=1}^{\mathrm{m}} e^{-4l_i}\right)$
Medio hermanos	1/4	1/16	$\frac{1}{64\mathrm{L}^2}\left(4\mathrm{L}-\mathrm{m}+\sum_{i=1}^{\mathrm{m}} e^{-4l_i}\right)$
Abuelo-nieto	1/4	1/16	$\frac{1}{32\mathrm{L}^2}\left(2\mathrm{L}-\mathrm{m}+\sum_{i=1}^{\mathrm{m}} e^{-2l_i}\right)$
Tío-sobrino	1/4	1/16	$\frac{1}{64\mathrm{L}^2}\left(\frac{5}{3}\mathrm{L}-\frac{13}{36}\mathrm{m}+\sum_{i=1}^{\mathrm{m}} e^{-4l_i}+e^{-6l_i}\right)$
Primos hermanos	1/8	3/64	$\frac{1}{256\mathrm{L}^2}\left[\frac{29}{6}\mathrm{L}-\frac{149}{144}\mathrm{m}+\sum_{i=1}^{\mathrm{m}}\left(\frac{3}{4}e^{-4l_i}+\frac{2}{9}e^{-6l_i}+\frac{1}{16}e^{-8l_i}\right)\right]$
Doble primos hermanos	1/4	3/32	$\frac{1}{128\mathrm{L}^2}\left[\frac{29}{6}\mathrm{L}-\frac{149}{144}\mathrm{m}+\sum_{i=1}^{\mathrm{m}}\left(\frac{3}{4}e^{-4l_i}+\frac{2}{9}e^{-6l_i}+\frac{1}{16}e^{-8l_i}\right)\right]$

Los argumentos expuestos anteriormente favorecen el empleo de R_{ij} en lugar de Z como medida de identidad genética. Al igual que Z, R_{ij} es una variable aleatoria no observable. No obstante, podemos calcular su esperanza, condicional a la información de los marcadores moleculares $\mathcal{M}$, i.e. $E(R_{ij}|\ \mathcal{M})$. El genotipado con chips de alta densidad de SNPs permite contar con información genotípica para un número importante de posiciones (actualmente del orden de las 770.000 en el bovino), distribuidas a lo largo de todo el genoma. Los métodos desarrollados para calcular $E(R_{ij}|\mathcal{M})$ (Goldgar, 1990; Guo, 1994; Ball y Stefanov, 2005) asumen que la fase de los marcadores es conocida, es decir que se conoce cuál entre ambos alelos de un individuo proviene de su padre y cuál de su madre; por lo tanto, se puede determinar el número de pares de alelos IBD en cada locus marcador. Estos métodos se basan en cadenas de Markov que modelan el proceso de gametogénesis a lo largo del genoma. Cuando la fase de los marcadores es desconocida, situación común en los SNPs, se debe recurrir a algoritmos que permitan estimar las probabilidades de identidad en cada posición del genoma, condicionales a los genotipos de la secuencia de marcadores. Algunos de estos algoritmos fueron revisados en el Capítulo 3 y se basan en HMMs. A partir de la expresión 5.19, la $E(R_{ij}\ |\ \mathcal{M})$, se podría calcular como

$$E(R_{ij}\ |\ \mathcal{M}) = \tfrac{1}{K}\sum_{k=1}^{K} E\left(Z_k \middle|\ \mathcal{M}\right) = \tfrac{1}{K}\sum_{k=1}^{K}\left[\tfrac{1}{2}P\left(I_k = 1\middle|\ \mathcal{M}\right) + P\left(I_k = 2\middle|\ \mathcal{M}\right)\right] \qquad [5.22]$$

donde $E(Z_k\ |\ \mathcal{M})$ es la proporción esperada de alelos IBD en la posición k del genoma, condicional a la información de todos los marcadores, para un par de individuos genotipados, i y j; $P\left(I_t = 1\middle|\ \mathcal{M}\right)$ y $P\left(I_t = 2\middle|\ \mathcal{M}\right)$ son, respectivamente, las probabilidades de identidad de compartir 1 y 2 pares de genes IBD en el i-ésimo SNP, condicional a $\mathcal{M}$. Estas probabilidades se pueden estimar en base al algoritmo de Li *et al.* (2010), utilizando las expresiones 3.17 y 3.18 de la sección 3.3.3 del Capítulo 3. Este algoritmo

considera el ligamiento y el LD entre marcadores para el cálculo de las probabilidades de identidad. Nótese que al utilizar [5.22] estaríamos sumando sobre la secuencia de SNPs, y no sobre la secuencia de QTLs; por lo tanto, estariamos asumiendo, tal como ocurre en SG, que los SNPs están en fuerte LD con al menos un QTL. De modo de independizarnos de este supuesto en el Capítulo 3 se propone reemplazar el término $1/K$ en [5.22] por un coeficiente de ponderación w_k que representa la cobertura del k-ésimo SNP relativa a la longitud física del genoma. De este modo se llega a la expresión 3.19 derivada en la sección 3.3.3 del Capítulo 3

$$\hat{g}_{\text{IBD-LD}} = \sum_{k=1}^{K} w_k \left[\tfrac{1}{2} P\left(I_k = 1 \middle| \mathcal{M}\right) + P\left(I_k = 2 \middle| \mathcal{M}\right) \right] \qquad [5.23]$$

donde $\hat{g}_{\text{IBD-LD}}$ es la $\text{E}(R_{ij}|\mathcal{M})$ calculada utilizando el algoritmo de Li *et al.* (2010). Al igual que el coeficiente que utiliza VanRaden (2008), $\hat{g}_{\text{IBD-LD}}$ constituye un estimador insesgado de $\boldsymbol{A}_{ij}$.

Finalmente, $\hat{g}_{\text{IBD-LD}}$ será incluida en el modelo aditivo siguiendo el enfoque de Goldgar (1990), Guo (1994) y Visscher *et al.* (2006). Según estos autores, el BV (a_i) de un individuo con genotipo denso es la suma de los efectos aditivos de un número infinito de loci distribuidos a lo largo del genoma, que a su vez está cubierto por una alta densidad de marcadores, con $\text{E}\left(a_i\right) = 0$ y $\text{Var}\left(a_i\right) = \sigma^2_{\text{A}}$. El parámetro σ^2_{A} representa la varianza aditiva debido a los genes distribuidos en el genoma marcado y $h^2 = \sigma^2_{\text{A}} / \left(\sigma^2_{\text{A}} + \sigma^2_e\right)$ es la heredabilidad del carácter. La covarianza entre los BVs de dos individuos es

$$\text{cov}\left(a_i, a_j \middle| \mathcal{M}\right) = R_{ij}\, \sigma^2_{\text{A}} \qquad [5.24]$$

donde R_{ij} es el valor realizado del coeficiente de relación aditiva para los individuos i y j. Dado que R_{ij} es una variable no observable, utilizaremos $\hat{g}_{\text{IBD-LD}}$ para predecir R_{ij}. La matriz de varianzas de los BVs de los individuos genotipados es expresada como

$$\mathrm{Var}\left(\boldsymbol{a}\middle|\,\mathcal{M}\right)=\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}\ \sigma_{\mathrm{A}}^{2} \qquad [5.25]$$

De este modo, los SNPs se utilizan *sólo* para estimar las probabilidades de identidad en posiciones específicas del genoma, teniendo en cuenta su distribución más o menos uniforme en el genoma y no se realizan supuestos sobre los efectos de sustitución de los SNPs (se mantiene el modelo infinitesimal).

5.3.4 Extensión de la matriz de varianzas y covarianzas de los valores de cría a los animales no genotipados

Las evaluaciones genéticas tradicionales basadas en el MA incluyen a todos los animales del pedigree. Ahora bien, sólo una pequeña fracción de la población de animales es genotipada en SG. Sea $\boldsymbol{a}$ un vector $q \times 1$ de BVs genómicos para todos los animales. Bajo el modelo infinitesimal de herencia poligénica, $\mathrm{Var}(\boldsymbol{a}) = \boldsymbol{A}\,\sigma_{\mathrm{A}}^{2}$, donde $\boldsymbol{A}$ es la matriz de relaciones aditivas entre los animales construida sobre la base del pedigree. Considérese dos tipos de animales en $\boldsymbol{a}$: 1) q_{NG} animales no genotipados (fenotipados y no fenotipados), con vector de BVs $\boldsymbol{a}_1$; 2) q_{G} animales genotipados (fenotipados y no fenotipados), con vector de BVs $\boldsymbol{a}_2$. No hay una distinción entre ancestros y descendientes de los animales genotipados. Entonces, $\boldsymbol{A}$ puede ser particionada de la siguiente manera:

$$\boldsymbol{A}=\begin{bmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{bmatrix}.\ \text{Su inversa es}\ \ \boldsymbol{A}^{-1}=\begin{bmatrix}\boldsymbol{A}^{11} & \boldsymbol{A}^{12}\\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22}\end{bmatrix} \qquad [5.26]$$

La matriz $\boldsymbol{A}^{-1}$ se obtiene siguiendo las reglas de Henderson (1976).

El procedimiento *single-step* desarrollado por Misztal *et al.* (2009), Legarra *et al.* (2009) e independientemente por Christensen y Lund (2010) para SG, asume que

$$\mathrm{Var}(\boldsymbol{a}_2) = \boldsymbol{G}_{\mathrm{VR}}\,\sigma^2_{\mathrm{A}} = \frac{\boldsymbol{MM'}}{\sum_{k=1}^{K} 2p_k q_k}\,\sigma^2_{\mathrm{A}} \qquad [5.27]$$

donde $\boldsymbol{G}_{\mathrm{VR}}$ es la matriz de relaciones genómicas propuesta por VanRaden (2008), cuyos detalles de construcción han sido desarrollados en la sección 5.3.2 (expresiones 5.9 - 5.11). Brevemente, $\boldsymbol{M}$ es una matriz de incidencia de orden $q_{\mathrm{G}} \times K$ que contiene los genotipos observados de los K loci marcadores y está centrada por las frecuencias alélicas; p_k y q_k son las frecuencias alélicas para el k-esimo marcador, y σ^2_{A} es la varianza genética aditiva. De este modo, al condicionar el valor de cría de los animales genotipados en la información genómica, la matriz $\boldsymbol{A}_{22}$ es reemplazada por $\boldsymbol{G}_{\mathrm{VR}}$. En base a propiedades de la distribución normal y al pedigree, esta información es expandida a los animales no genotipados utilizando la "regresión" $\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}$ (Legarra *et al.*, 2009). De este modo, la matriz de varianzas y covarianzas de los valores de cría, $\boldsymbol{A}$, es reemplazada en el procedimiento *single-step* por una nueva matriz de relaciones que incluye la información genómica, $\boldsymbol{H}$, igual a

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_{11} + \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\left(\boldsymbol{G}_{\mathbf{VR}} - \boldsymbol{A}_{22}\right)\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{G}_{\mathbf{VR}} \\ \boldsymbol{G}_{\mathbf{VR}}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21} & \boldsymbol{G}_{\mathbf{VR}} \end{bmatrix} \qquad [5.28]$$

La inversa de $\boldsymbol{H}$ fue obtenida por Aguilar et al. (2010) y es igual a:

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_{\mathbf{VR}}^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \qquad [5.29]$$

Para incorporar la estimación de la proporción IBD compartida a nivel genómico entre animales genotipados, $\hat{g}_{\mathrm{IBD\text{-}LD}}$, en las predicciones de los BVs se debería reemplazar la matriz $\boldsymbol{G}_{\mathbf{VR}}$ en las expresiones 5.27 - 5.29 por la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ construida a partir de la expresión 5.23.

5.3.5 Predicciones genómicas y exactitudes

El objetivo final del capítulo es evaluar de forma analítica y empírica el impacto de reemplazar la matriz $\boldsymbol{G}_{\mathbf{VR}}$ por la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$, sobre la exactitud de las predicciones genómicas de los BVs (GEBVs). Se utilizó la simulación (Cuadro 5.1) para probar empíricamente si la mejora en la precisión de las estimaciones de los elementos de $\boldsymbol{G}$ podría resultar en un aumento significativo de la exactitud de las GEBVs para los animales candidatos a la selección. A tal fin, se asumió un MA unicarácter sin registros fenotípicos para los animales de la última generación

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{Z}\,\boldsymbol{a} + \boldsymbol{e} \qquad [5.30]$$

donde $\boldsymbol{y}$ es un vector de orden $n \times 1$ de registros fenotípicos para un carácter, m es una media general, $\boldsymbol{a} = \{a_i\}$ es un vector aleatorio de orden $q \times 1$ que contiene los BVs de los animales ($q = q_{\mathrm{NG}} + q_{\mathrm{G}}$), $\boldsymbol{Z}$ es una matriz de incidencia de orden $n \times q$ ($n < q$) que relaciona los elementos de $\boldsymbol{a}$ con los de $\boldsymbol{y}$, y $\boldsymbol{e}$ es el vector $n \times 1$ de errores. Se asume que

$$\mathrm{E}(\boldsymbol{y}) = \boldsymbol{I}_n\mu \qquad \mathrm{Var}(\boldsymbol{a}) = \boldsymbol{H}\,\sigma^2_a \qquad \mathrm{Var}(\boldsymbol{e}) = \boldsymbol{I}\,\sigma^2_e \qquad \mathrm{Cov}(\boldsymbol{a}, \boldsymbol{e}') = \boldsymbol{0}$$

donde $\boldsymbol{H}$ es la matriz de varianzas y covarianzas de los BVs genómicos de todos los animales, σ^2_A es la varianza genética aditiva, and σ^2_e es la varianza de los errores residuales. El Cuadro 5.3 muestra la dimensión que toma la base de datos simulados con fines predictivos.

Cuadro 5.3. Dimensión de la base de datos simulados con fines predictivos

Nro. de animales en el pedigree (q)	10.220
Nro. de animales no genotipados (q_{NG})	10.080
Nro. de animales (machos) genotipados (q_{G})	140
Nro. de animales candidatos a la selección de la última generación	2.000
Nro. de candidatos a la selección genotipados	40

Por lo tanto, la matriz de coeficientes a la izquierda de las ecuaciones de modelos mixtos (*left-hand side*, **LHS**) es igual a:

$$\mathbf{LHS} = \begin{bmatrix} \boldsymbol{1}_n'\boldsymbol{1}_n\sigma_e^{-2} & \boldsymbol{1}_n'\boldsymbol{Z}\sigma_e^{-2} \\ \boldsymbol{Z}'\boldsymbol{1}_n\sigma_e^{-2} & \boldsymbol{Z}'\boldsymbol{Z}\,\sigma_e^{-2} + \boldsymbol{H}^{-1}\sigma_{\mathrm{A}}^{-2} \end{bmatrix} \quad [5.31]$$

En [5.31], $\boldsymbol{Z}'\boldsymbol{Z}$ es una matriz diagonal con d_{ii} = 1 cuando el animal *i* tiene registro fenotípico y cero en caso contrario, $\boldsymbol{H}^{-1}$ es la inversa de la matriz de varianzas y covarianzas de los BVs genómicos. Empleando teoría BLUP (Henderson, 1984), la exactitud de la GEBV del animal *i* es:

$$\mathrm{r}_i = \sqrt{1 - \frac{\mathrm{PEV}(\boldsymbol{a}_i)}{\sigma_{\mathrm{A}}^2}} \quad [5.32]$$

donde PEV_i es la varianza del error de predicción del animal *i*. A fin de comparar las diferentes matrices genómicas de parentesco, se asumió que la matriz de varianzas y covarianzas de los BVs ($\boldsymbol{H}$) correcta era $\boldsymbol{H}_{\mathrm{T}} = \boldsymbol{G}_{\mathrm{T}}$, que es una matriz conocida en los datos simulados y cuyos elementos fueron calculados en el Capítulo 4 utilizando las expresiones 4.4 y 4.5. En el modelo "verdadero" (i.e., $\boldsymbol{H} = \boldsymbol{H}_{\mathrm{T}} = \boldsymbol{G}_{\mathrm{T}}$), la matriz **PEV**($\boldsymbol{a}$) puede ser calculada como la inversa de **LHS**. Ahora bien, cuando la matriz de (co)varianzas de los BVs ($\boldsymbol{H}$) está mal especificada, es factible calcular **PEV**($\boldsymbol{a}$) como en Henderson (1975):

$$\mathbf{PEV}(\boldsymbol{a}) = \boldsymbol{C}^{aa} + \boldsymbol{C}^{aa}\,\boldsymbol{H}^{-1}\sigma_{\mathrm{A}}^{-2}\left(\boldsymbol{H}_{\mathrm{T}} - \boldsymbol{H}\right)\boldsymbol{H}^{-1}\,\boldsymbol{C}^{aa} \quad [5.33]$$

con:

$$\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1} + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} - \boldsymbol{A}_{22}^{-1} \end{bmatrix} \quad [5.34]$$

En [5.33] $\boldsymbol{C}^{aa}$ es la inversa de la matriz **LHS**, obtenida utilizando [5.34] en [5.31], $\boldsymbol{H}_T$ es la matriz de (co)varianzas de los BVs bajo el modelo "verdadero", $\boldsymbol{H}^{-1}$ es la inversa de la matriz de (co)varianzas de los BVs que combina la información genómica y del pedigree (Aguilar et al., 2010). En [5.34], $\boldsymbol{G}^{-1}$ es la inversa de la matriz genómica de parentesco estimada por alguno de los métodos propuestos en el Capítulo 4 y discutidos en el contexto de la SG en este capítulo ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ o $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$) o por $\boldsymbol{A}_{\mathbf{22}}$ (i.e., $\boldsymbol{H}^{-1} = \boldsymbol{A}^{-1}$) y $\boldsymbol{A}_{\mathbf{22}}^{-1}$ es la inversa de la matriz de relaciones genéticas aditivas para los animales genotipados basada sólo en la información del pedigree. Se computó la exactitud de los GEBVs bajo dos escenarios de heredabilidad: $h^2 = 0{,}25$ y $h^2 = 0{,}15$.

La significancia de las diferencias en exactitud de los GEBVs entre metodologías de estimación de $\boldsymbol{G}$ fue testeada con el procedimiento de comparaciones múltiples de Scheffé. Las exactitudes (exa_{ijk}) de los candidatos a la selección ($k = 1, \ldots, 2.000$) se analizaron utilizando el modelo lineal mixto (Proc Mixed SAS version 9.3.1, SAS Institute, Cary, NC, USA) $exa_{ijk} = t_i + r_j + e_{ijk}$, donde el estimador de la matriz genómica de parentesco utilizado para calcular la exactitud se trató como un efecto fijo (t_i, $i = 1, \ldots, 4$ para $\boldsymbol{A}_{\mathbf{22}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}O}}$, $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$ y $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$, respectivamente) y la réplica (r_j, $j = 1, \ldots, 50$) se trató como un efecto aleatorio. Se utilizó una matriz de covarianzas $\boldsymbol{R}$ diagonal en bloques para los errores e_{ijk}, en la cual todas las observaciones con el mismo nivel del efecto fijo t_i tienen el mismo parámetro o componente de varianza.

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}\sigma^2_{\varepsilon_1} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}\sigma^2_{\varepsilon_2} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}\sigma^2_{\varepsilon_3} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}\sigma^2_{\varepsilon_4} \end{bmatrix} \qquad [5.35]$$

5.4 Resultados

Para analizar las consecuencias de utilizar diferentes matrices de parentesco genómico $\boldsymbol{G}$ en la exactitud de las predicciones de los BVs, se calculó la exactitud de los GEBVs para los candidatos a la selección bajo dos escenarios: $h^2 = 0{,}25$ y $h^2 = 0{,}15$. El Cuadro 5.4 muestra la media (estimada por mínimos cuadrados) y el error estándar de las exactitudes de los GEBVs para los animales candidatos a la selección utilizando diferentes matrices de parentesco, sobre 50 réplicas. Tal como se esperaba, la utilización de cualquiera de las matrices genómicas resultó en una mayor exactitud de los GEBVs para los candidatos a la selección que con la matriz de relaciones aditivas. El empleo de esta última fue similar a la exactitud alcanzada por el promedio parental (*parent average*, PA) en una evaluación genética tradicional. La exactitud de los GEBVs para los candidatos a la selección fue estadísticamente superior cuando se utilizó $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$. De hecho, las diferencias fueron más marcadas para los animales genotipados. La diferencia entre estimadores de la matriz genómica basados en la noción de IBS ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$ y $\boldsymbol{G}_{\mathbf{VR\text{-}B}}$) en la exactitud de los GEBVs no fue estadísticamente significativa. Las exactitudes cayeron en la misma magnitud en el escenario de $h^2 = 0{,}15$ para los tres estimadores.

Cuadro 5.4. Media (Error estándar) de las exactitudes de los GEBVs para los animales candidatos a la selección utilizando diferentes matrices de parentesco.

h^2		$\boldsymbol{A}$	$\boldsymbol{G}_{\text{VR-O}}$	$\boldsymbol{G}_{\text{VR-B}}$	$\boldsymbol{G}_{\text{IBD-LD}}$
0.25	Genotipados*	0,4984a (0,00235)	0,5379b (0,00244)	0,5382b (0,00245)	0,5592c (0,00235)
	Todos*	0,4970a (0,00143)	0,5176b (0,00143)	0,5176b (0,00143)	0,5211c (0,00143)
0.15	Genotipados*	0,4599a (0,00255)	0,5008b (0,00263)	0,5013b (0,00263)	0,5282c (0,00252)
	Todos*	0,4580a (0,00161)	0,4811b (0,00161)	0,4812b (0,00161)	0,4863c (0,00161)

*Las letras diferentes en la misma fila indican una diferencia estadísticamente significativa entre matrices de covarianzas (p-valor < 0,0001). Se emplearon 50 réplicas.

$\boldsymbol{A}$: matriz de relaciones aditivas; $\boldsymbol{G}_{\text{VR}}$: matriz de relaciones genómicas basada en la noción de IBS y construida con las frecuencias alélicas observadas ($\boldsymbol{G}_{\text{VR-O}}$) o con las frecuencias alélicas de los animales de la población base ($\boldsymbol{G}_{\text{VR-B}}$); $\boldsymbol{G}_{\text{IBD-LD}}$: matriz genómica basada en la noción de IBD.

El Cuadro 5.5 muestra las estimaciones REML de los parámetros de covarianza del modelo utilizado para analizar las exactitudes de los GEBVs de los candidatos a la selección mediante el procedimiento de comparaciones múltiples de Scheffé. Tal como se esperaba, el valor del estimador REML de la varianza del efecto de la réplica ($\hat{\boldsymbol{\sigma}}^2_{rep}$) fue muy inferior a los valores estimados de las varianzas de los efectos residuales asociados a cada matriz de parentesco ($\hat{\sigma}^2_{\varepsilon_A}$, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-O}}$, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-B}}$ y $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{IBD-LD}}$), más aun cuando aumenta el número de observaciones en el modelo (i.e., cuando se incluyen las

exactitudes de todos los candidatos a la selección, tengan o no genotipo). El valor de $\hat{\sigma}^2_{rep}$ no se modificó sustancialmente con el cambio de heredabilidad; no ocurrió lo mismo con $\hat{\sigma}^2_{\varepsilon_A}$, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-O}}$, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-B}}$ y $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{IBD-LD}}$, cuyos valores aumentaron al disminuir la heredabilidad a 0,15. La magnitud de $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{IBD-LD}}$ fue inferior a $\hat{\sigma}^2_{\varepsilon_A}$, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-O}}$ y $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-B}}$ en todos los escenarios y las diferencias fueron más marcadas para los animales genotipados.

Cuadro 5.5. Estimaciones REML de los parámetros de covarianza del modelo utilizado para analizar las exactitudes de los GEBVs de los candidatos a la selección mediante el procedimiento de comparaciones múltiples de Scheffé.

h^2		$\hat{\sigma}^2_{rep}$	$\hat{\sigma}^2_{\varepsilon_A}$	$\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-O}}$	$\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-B}}$	$\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{IBD-LD}}$
0.25	Genotipados*	0,00022	0,00210	0,00300	0,00300	0,00210
	Todos*	0,00010	0,00206	0,00200	0,00204	0,00188
0.15	Genotipados*	0,00025	0,00285	0,00363	0,00363	0,00246
	Todos*	0,00013	0,00290	0,00262	0,00269	0,00246

*Las letras diferentes en la misma fila indican una diferencia estadísticamente significativa entre matrices de covarianzas ($P < 0,0001$)

$\boldsymbol{A}$: matriz de relaciones aditivas; $\boldsymbol{G}_{\mathbf{VR}}$: matriz de relaciones genómicas basada en la noción de IBS y construida con las frecuencias alélicas observadas ($\boldsymbol{G}_{\mathbf{VR\text{-}O}}$) o con las frecuencias alélicas de los animales de la población base ($\boldsymbol{G}_{\mathbf{VR\text{-}B}}$); $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$: matriz genómica calculada sobre el concepto de IBD.

5.5 Discusión

En este capítulo se ha intentado evaluar analíticamente cómo el aumento en la precisión de las estimaciones de las relaciones genómicas de parentesco se podría traducir en un aumento en la exactitud de la predicción de los valores de cría genómicos, dependiendo del estimador de la matriz $\boldsymbol{G}$ utilizado. El supuesto es que, en el "verdadero modelo", la matriz de las covarianzas entre los valores de cría genómicos es función de los valores realizados para la proporción del genoma IBD compartida por cada par de individuos, que es desconocida (salvo en una simulación). La eficiencia de GBLUP depende fundamentalmente de la medida en que las relaciones genómicas derivadas de los marcadores reflejan los patrones de las verdaderas relaciones genéticas realizadas en los loci que afectan el carácter (de los Campos *et al.*, 2013). La investigación actual ha intentado comparar la exactitud de las GEBVs alcanzadas con el enfoque $\boldsymbol{G}_{\mathbf{VR}}$ propuesto por VanRaden (2008), ampliamente utilizado para estimar relaciones genómicas utilizando marcadores, con la alcanzada utilizando el enfoque $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$.

Vela-Avitúa *et al.* (2015) simularon un esquema de SG en peces y mostraron que las diferencias en la precisión de las GEBVs entre distintos estimadores de $\boldsymbol{G}$ dependen de la densidad de los marcadores: las GEBVs calculados en base a la noción de IBS fueron ligeramente más precisos que sus contrapartes basados en la noción de IBD cuando se utilizaron una alta densidad de marcadores, pero también fueron mucho más sensibles a un reducción en la densidad. Sin embargo, estos autores encontraron que la precisión de las GEBVs basados en IBD era estable a través de las diferentes densidades de marcadores y, de hecho, mayor a densidades bajas (≤100 SNP / m) que la obtenida usando la matriz $\boldsymbol{G}$ basada en IBS. En nuestra simulación utilizando marcadores densos, la exactitud de las GEBVs para los candidatos de selección fue

estadísticamente superior cuando se utilizó la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$. Esta ligera superioridad en la precisión podría explicarse por el hecho de que nuestro enfoque basado en la noción de IBD difiere de la utilizada en el artículo antes mencionado en el hecho de que modela la información del LD. Asimismo, debe señalarse que el estimador IBD empleado por Vela-Avitúa *et al.* (2015) no tenía en cuenta el LD, calculando las matrices IBD para cada marcador en particular de manera independiente. En nuestro caso, el algoritmo de Li *et al.* 2010 (ver Capítulo 3) incorpora el estado oculto de IBD poblacional o "Background" IBD para ajustar por la relación de parentesco oculta más allá de la relación que se observa a través de la estructura de pedigree disponible y considerando el LD. Sin embargo, esto se produce a expensas de la utilización de métodos HMM que son computacionalmente intensivos (~ 4 horas por cromosoma).

Las diferencias en la exactitud de GEBVs entre los estimadores basados en IBS no fueron estadísticamente significativas. Strandén y Christensen (2011) mostraron que los cambios en el numerador de $\boldsymbol{G}_{\text{VR}}$ (al igual que las frecuencias alélicas utilizadas para centrar genotipos) no cambian las diferencias relativas entre las GEBVs predichas, ya que están simplemente desplazadas por una constante. Sin embargo, la modificación del denominador que escala $\boldsymbol{G}_{\text{VR}}$ es como dividir o multiplicar $\boldsymbol{G}$ por una constante y podría, en principio, cambiar los resultados, aunque en nuestro caso esto no afectó los resultados en gran medida.

Al estimar vía REML los componentes de varianza del modelo mixto utilizado en el test de comparaciones múltiples de las diferencias de las exactitudes de los GEBVs para los candidatos a la selección con diferentes matrices de parentesco, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{IBD-LD}}$ fue inferior a $\hat{\sigma}^2_{\varepsilon_A}$, $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-O}}$ y $\hat{\sigma}^2_{\varepsilon \boldsymbol{G}\text{VR-B}}$ en todos los escenarios. Esto significa que las exactitudes calculadas con la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ fueron más uniformes (Cuadro 5.5), sobretodo en los candidatos genotipados; y si a este resultado le agregamos el hecho de

que la media del efecto de usar esta matriz fue superior (Cuadro 5.4), se puede concluir que este es el mejor estimador.

Podemos concluir que el incorporar la información de los registros genealógicos en el cálculo de las relaciones genómicas de parentesco entre animales genotipados mejora la exactitud de las predicciones genómicas de los valores de cría para los animales candidatos a la selección.

CAPÍTULO 6

Discusión general

Discusión general

La mayor productividad en los sistemas de producción animal es en parte el resultado de una intensa tarea de selección genética para los caracteres de importancia pecuaria. Los métodos de selección clásicos, que utilizan las predicciones de los BVs calculadas a partir de modelos lineales mixtos y una importante base de registros fenotípicos, han sido hasta la actualidad el motor que impulsó el progreso genético en el sector ganadero. Los avances tecnológicos recientes han hecho posible el genotipado con chips de alta densidad a gran escala, disponiéndose de genotipos para más de 700.000 posiciones en el genoma de cada animal. Por lo tanto, la oportunidad de desarrollar programas de selección empleando información genómica se ha convertido en una realidad. En efecto, la SG ha sido adoptada rápidamente en los programas de mejoramiento de ganado lechero vacuno y muy pronto muchas más especies seguirán ese camino.

La motivación fundamental para el desarrollo de esta tesis fue su orientación hacia los últimos avances en genética poblacional y molecular. Con los paneles de genotipado denso, se introduce el concepto del LD entre SNPs. Este escenario, será cada vez más frecuente en el futuro de la mejora genética animal, dado que año a año tendremos un chip de mayor densidad de SNPs con un costo reducido. El LD puede verse como una fuerza que aumenta el parecido entre individuos, o parentesco ancestral que va más allá de las relaciones conocidas, y se debe a que tanto la población como el número de alelos iniciales son finitos. Por lo tanto comprender e interpretar mejor la información del LD cobrará cada vez mayor importancia, sobre todo si se quiere implementar la SG en poblaciones pequeñas y en aquellas donde la proporción de animales genotipados es pequeña en relación con la dimensión del pedigree.

En líneas generales, esta tesis se centró en el desarrollo de matrices genómicas de parentesco ***G*** teniendo en cuenta: a) el pedigree, b) la información (de calidad) de un chip de alta densidad de SNPs, c) el LD entre SNPs, y su implementación algorítmica dentro de la evaluación genética animal. A tal efecto, se propusieron métodos para utilizar la información de los SNPs y del pedigree de un modo eficiente y distinto al que se utiliza actualmente en SG, ya sea para efectuar análisis previos al cálculo de ***G*** (i.e., filtrado de los SNPs de baja calidad) y/o para generar ***G*** (i.e., calcular las relaciones genómicas).

La principal contribución del Capítulo 2 es de naturaleza metodológica. En el mismo, se abordó el análisis del QC de un chip de alta densidad de SNPs, dado que la calidad del genotipado define el éxito de posteriores análisis estadísticos cuando se trabaja con información genómica. En este capítulo, se elaboró un criterio de filtrado basado en el conteo de alelos (GC) para cada SNP y en su herencia. Al observar la herencia de un SNP, el método utiliza simultáneamente todos los genotipos evaluados y su pedigree. Para ello, asume que la covarianza entre los GCs de dos animales para un SNP es proporcional a la fracción de alelos IBD en el marcador (Gengler *et al.*, 2007), que es una medida de identidad genética basada en un solo locus. El método utiliza REML para estimar la heredabilidad del GC para cada SNP y además proporciona un test estadístico del cociente de verosimilitud que pone a prueba la hipótesis nula de "ausencia de errores en el genotipado". De este modo, permite eliminar los SNPs de baja calidad teniendo en cuenta simultáneamente varios de los criterios de filtrado que se utilizan actualmente en forma independiente (e.g., errores mendelianos, frecuencias alélicas mínimas; Wiggans *et al.* 2009, 2012). El método propuesto en esta tesis superó en performance al chequeo estándar de errores Mendelianos, para datos simulados. A diferencia de otras técnicas, el método propuesto puede ser utilizado en cualquier

población con marcadores y registros genealógicos, y es sencillo de implementar utilizando programas estándar de estimación vía REML.

El criterio de filtrado de QC propuesto en esta tesis podría emplearse, por ejemplo, para depurar una base de datos de genotipos antes de ser utilizada en la estimación de las relaciones de parentesco genómicas. Como sugerencia de una futura investigación, se podría evaluar si la utilización de este criterio de filtrado (en lugar de los procedimientos estándar de QC) aumenta o reduce las diferencias en precisión entre las estimaciones de las relaciones genómicas calculadas empleando los estimadores presentados en el Capítulo 4 (i.e., $\hat{g}_{\text{IBD-LD}}$ y $\hat{g}_{\text{VR}}$). Asimismo, otro interrogante que plantea la investigación sería evaluar si la utilización de las estimaciones de las frecuencias alélicas de los marcadores en la población base (p_m, m = 1, …., M marcadores) que se obtienen como subproducto del método de filtrado de QC aquí propuesto, mejora la precisión de $\hat{g}_{\text{VR-B}}$. Esto podría resultar de utilidad porque las frecuencias alélicas de la población base no siempre están bien representadas por las frecuencias alélicas de los animales genotipados en la generación fundadora del pedigree (este fue el caso de la base de datos reales utilizada en el Capítulo 4).

La incorporación de nuevas tecnologías bioinformáticas en la evaluación genética animal, constituye uno de los aspectos más relevantes de esta tesis. Tal es el caso del algoritmo de Li *et al.* (2010), aplicado en el área de la genética humana, cuyo potencial para el análisis de toda la información genotípica y del pedigree disponible de un animal hacen de esta metodología una excelente alternativa para estimar las relaciones de parentesco genómicas empleando la noción de IBD extendida a todo el genoma. Hasta el momento, en las aplicaciones prácticas de SG, la información de los marcadores ha sido utilizada para estimar las relaciones genómicas empleando la noción de identidad por estado (VanRaden, 2008), como una aproximación a la de identidad

por descendencia. En esta tesis se ha propuesto el algoritmo de Li *et al.* (2010) como base para la estimación de la proporción del genoma IBD (Capítulo 3). Mediante un marco teórico basado en modelos de Markov ocultos se puede modelar el proceso de identidad por descendencia a lo largo de un cromosoma y calcular las probabilidades de IBD en cada SNP a partir de datos genotípicos para pares de individuos en un pedigree complejo. El modelo tiene en cuenta el LD entre marcadores y se puede aplicar a datos de chips de alta densidad. Si bien en la literatura existen otros antecedentes en la estimación de los estados de IBD en posiciones específicas del genoma (Abecasis *et al.*, 2002; Abecasis y Wigginton, 2005; Keith *et al.*, 2008; Purcell *et al.*, 2007; Albrechtsen *et al.*, 2009; Browning y Browning, 2010), todos ellos son en el área de la genética humana por lo cual, en este punto, la utilización de este algoritmo para estimar relaciones genómicas en pedigrees refinando el verdadero parecido entre animales resulta original.

Debe señalarse que en esta tesis no se pretende proponer el algoritmo de Li *et al.* (2010) como el mejor estimador de las relaciones genómicas, sino demostrar que utilizar un algoritmo que incorpore la información del pedigree y de los marcadores, teniendo en cuenta los procesos de gametogénesis y el LD (en lugar de sólo utilizar marcadores) mejora las precisiones de las relaciones genómicas (Capítulo 4). En la presente tesis, se trabajó con una base de datos reales sin consanguinidad y una simulación estocástica en donde los apareamientos se diseñaron de manera de minimizar consanguinidad (la consanguinidad promedio en la simulación fue nula en las primeras tres generaciones, y menor a 2 y 4% en las últimas dos generaciones). El algoritmo de Li *et al.* (2010) considera independencia entre los cromosomas homólogos de un individuo; sin embargo, ha demostrado comportarse razonablemente en pedigrees con loops. En una investigación futura podría ser la de comparar la performance de este

algoritmo con la de otros como el propuesto por Han y Abney (2011), quienes modelaron los estados de identidad con consanguinidad.

En el Capítulo 5 se revisó el marco teórico de la evaluación genética tradicional y de la SG, para luego proponer y evaluar la implementación de la matriz genómica propuesta en el Capítulo 3 ($\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$). En este capítulo, la exactitud de las predicciones genómicas de los BVs para los candidatos a la selección calculadas a partir $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$, fue mayor en promedio que las obtenidas con las metodologías utilizadas actualmente en SG y, además, dicha medida presentó menor variabilidad en un número grande de réplicas y de animales (Cuadros 5.4 y 5.5). Recientemente, Vela-Avitúa *et al.* (2015) evaluaron la exactitud de los GEBVs en un esquema acuícola simulado, obtenidos a partir de una matriz $\boldsymbol{G}_{\mathbf{IBD}}$ con soporte teórico en el análisis de ligamiento (Fernando y Grossman, 1989; Luan *et. al*, 2012), y que a su vez utiliza un programa desarrollado por Meuwissen y Goddard (2010) para estimar las probabilidades de identidad. Estos autores encontraron que los GEBVs calculados en base a la noción de IBS (VanRaden, 2008) fueron ligeramente más precisos que sus contrapartes basadas en la noción de IBD, cuando se utilizó una alta densidad de marcadores. La matriz $\boldsymbol{G}_{\mathbf{IBD}}$ empleada por Vela-Avitúa *et al.* (2015), a diferencia de la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ propuesta en esta tesis, no modela la información del LD; en consecuencia esto podría explicar las diferencias en exactitud con respecto a los resultados de esta tesis. Por otra parte, la extensión de la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ a la situación en la cual no todos los individuos están genotipados (Capítulo 5), intenta de alguna manera, unir la teoría clásica de la covarianza entre parientes con las metodologías más recientes en SG de predicción de los BVs, esto es, el procedimiento *single-step* desarrollado por Misztal *et al.* (2009), Legarra *et al.* (2009) e independientemente por Christensen y Lund (2010). Esto resulta beneficioso a la hora de implementar masivamente cualquier alternativa para construir $\boldsymbol{G}$, dado que pueden

utilizarse programas como el PREGSF90 (Aguilar *et al.*, 2014) que utilizan algoritmos optimizados para la inversión de matrices.

La información que brindan los SNPs, y más recientemente la secuenciación, es cada vez mayor, a pesar de la gran variabilidad de procesos involucrados y del cuidado que se debe tener al realizar una asociación demasiado estrecha entre identidad alélica (o por estado) e identidad por descendencia. Existen varias formas de modelar los procesos y las probabilidades de IBD. Son muchas las maneras de utilizar la noción de IBD para analizar la variación fenotípica entre un par de individuos relacionados dentro de un pedigree, i.e., de estimar la relaciones genómicas y/o detectar segmentos IBD. Pero es importante remarcar que la noción de IBD unifica el mapeo genético en pedigrees y poblaciones desde un SNP ubicado en un solo par de bases hasta un haplotipo que se extiende por varios millones de pares de bases, proporcionando así un marco teórico a partir del cual se puede abordar la heredabilidad y la variación cuantitativa en las poblaciones, como también la historia demográfica y evolutiva de las distintas especies (Thompson, 2013).

CAPÍTULO 7

Conclusiones

Conclusiones

Sobre la base de toda la investigación realizada, las principales conclusiones de esta tesis pueden ser sintetizadas como sigue:

1. Se propuso una metodología para estimar la proporción de genoma compartido IBD por un par de individuos genotipados dentro de un pedigree de animales, condicional a las relaciones de parentesco y a la información de un chip de alta densidad de SNPs, considerando al mismo tiempo el LD entre marcadores. Dicha estimación representa un coeficiente de relación aditiva genómico y puede constituir cada uno de los elementos de la matriz de relaciones genómicas entre animales genotipados ($\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$).
2. Se evaluó y comparó esta metodología con aquella utilizada actualmente en SG ($\boldsymbol{G}_{\mathbf{VR}}$) para el cálculo de las relaciones genómicas en términos de la precisión de las relaciones estimadas. La incorporación de la información de los registros genealógicos en el cálculo de las relaciones genómicas mejoró la precisión de las estimaciones, sobre todo cuando se trabajó con familias grandes con muy pocos animales genotipados.
3. Esta tesis presentó contribuciones teóricas y metodológicas sustanciales para la implementación algorítmica de la matriz $\boldsymbol{G}_{\mathbf{IBD\text{-}LD}}$ dentro de la evaluación genómica. Una vez más, el incorporar la información de los registros genealógicos en el cálculo de las relaciones genómicas de parentesco entre animales genotipados mejoró la exactitud de las predicciones genómicas de los valores de cría para los animales candidatos a la selección.
4. De modo original, se desarrolló un procedimiento simple para realizar un análisis de control de calidad (QC), de modo de identificar SNPs de baja calidad

en un número grande de individuos. El filtro QC propuesto es, en esencia, una estimación de la heredabilidad del conteo de alelos en los SNPs, donde cualquier desviación de 1 es sospechosa y el *p*-valor sirve para testear la hipótesis nula de "ausencia de errores de genotipado". Este método de QC tiene la capacidad de considerar todos los individuos genotipados y su pedigree de manera conjunta y utiliza procedimientos estándar de pruebas de hipótesis.

BIBLIOGRAFÍA

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

Abecasis, G.R., Wigginton, J.E. 2005. Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. Am. J. Hum. Genet. 77: 754−767.

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., Lawlor, T. J. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93: 743−752.

Aguilar, I., Misztal, I., Legarra, A., Tsuruta, S. 2011. Efficient computations of genomic relationship matrix and other matrices used in the single-step evaluation. J. Anim. Breed. Genet. 128: 422−428.

Aguilar, I., Misztal, I., Tsuruta, S., Legarra, A., Wang, H. 2014. PREGSF90–POSTGSF90: Computational Tools for the Implementation of Single-step Genomic Selection and Genome-wide Association with Ungenotyped Individuals in BLUPF90 Programs. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada, poster 680.

Albrechtsen, A., Sand Korneliussen, T., Moltke, I., van Overseem Hansen, T., Nielsen, F.C., and Nielsen, R. 2009. Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. Genet. Epidemiol. 33: 266−274.

Almasy, L., Blangero, J. 1998. Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. 62: 1198−1211.

Ball, F., Stefanov, V.T. 2005. Evaluation of identity-by-descent probabilities for half-sibs on continuous genome. Mathematical Biosciences 196: 215–225.

Broman, K.W. 2001. Estimation of allele frequencies with data on sibships. Genetic Epidemiology 20: 307−315.

Browning, S.R., Browning, B.L. 2010. High-resolution detection of identity by descent in unrelated individuals. Am. J. Hum. Genet. 86: 526−539.

Bulmer, M.G. 1985. The mathematical theory of quantitative genetics. Clarendon Press, Oxford.

Cantet, R.J.C., Gualdrón Duarte, J.L., Munilla Leguizamón, S. 2008. Selección Genómica. Revista Argentina de Producción Animal 28: 133−136.

Cantet, R.J.C., Vitezica, Z.G. 2014. Properties of Mendelian residuals when regressing breeding values using a genomic covariance matrix. 10th World Congress on Genetics Applied to Livestock Production, Vancouver, Canada, poster 687.

Chen, C. Y., Misztal, I., Aguilar, I., Legarra, A., Muir, W. M. 2011. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89: 2673−2679.

Cheng S.H., Higham N.J. 1998. A modified Cholesky algorithm based on a symmetric indefinite factorization. SIAM J. on Matrix Anal. and Appl. 19: 1097–1110.

Cheung, C.Y., Thompson, E.A., Wijsman, E.M. 2014. Detection of Mendelian Consistent Genotyping Errors in Pedigrees. Genet. Epidemiol. 38: 291−299.

Christensen, O.F., Lund, M.S. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42: 2.

Cleveland, M.A., Hickey, J.M., Forni, S. 2012. A common dataset for genomic analysis of livestock populations. G3: Genes|Genomes|Genetics 2: 429−435.

Cockerham, C.C. 1969. Variance of gene frequencies. Evolution 23: 72−84.

De los Campos, G., Vazquez, A.I., Fernando, R., Klimentidis, Y.C., Sorensen, D. 2013. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. PLoS Genet. 9, e1003608.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press. UK.

Edwards, D. B., Ernst, C. W., Tempelman, R. J., Rosa, G. J. M., Raney, N. E., Hoge, M. D., Bates, R. O. 2008. Quantitative trait loci mapping in an F_2 Duroc × Pietrain resource population: I. Growth traits. J. Anim. Sci. 86: 241–253.

Elston, R.C., Stewart, J. 1971. A general model for the genetic analysis of pedigree data. Hum. Hered. 21: 523−542.

Falconer, D.S., Mackay, T.F.C. 1996 Introduction to quantitative genetics. Longman New York.

Fernando, R.L., Grossman, M. 1989. Marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. 21: 467−477.

Forni, S. Aguilar, I., Misztal, I. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43: 1.

Frühwirth-Schnatter, S. 2006. Finite Mixture Modeling. En: Finite Mixture and Markov Switching Models, Springer Series in Statistics, Springer, New York, pp. 1–23.

Gengler, N., Mayeres, P., Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1: 21−28.

Gianola, D., de los Campos, G., Hill, W.G., Manfredi, E., Fernando, R. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183: 347−363.

Goddard, M. 2008. Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136: 245−257.

Goddard, M.E., Hayes, B.J., Meuwissen, T.H.E. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. J. Anim. Breed. Genet. 128: 409–421.

Goldgar, D. E. 1990. Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. 47: 957−967.

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

Guo, S.W. 1994. Computation of identity-by-descent proportions shared by two siblings. Am. J. Hum. Genet. 54: 1104−1109.

Guo, S.W. 1995. Proportion of genome shared identical by descent by relatives: concept, computation, and applications. Am. J. Hum. Genet. 56: 1468–1476.

Guo, S.W. 1996. Variation in genetic identity among relatives. Hum. Hered. 46: 61–70.

Han, L., Abney, M. 2011. Identity by descent estimation with dense genome-wide genotype data. Genet. Epidemiol. 35: 557–567.

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

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

Haley, C. 1999. Advances in quantitative trait locus mapping. In: Dekkers, J.C.M., Lamont, S.J., Rothschild, M.F. (eds) From Jay Lush to Genomics: Visions for Animal Breeding and Genetics. Animal Breeding and Genetics Group, Department of Animal Science, Iowa State University, Ames, Iowa. pp. 47–59.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., Goddard, M. E. 2009. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92: 433–443.

Hayes, B. 2011. Technical note: Efficient parentage assignment and pedigree reconstruction with dense single nucleotide polymorphism data. J. Dairy Sci. 94: 2114−2117.

Henderson, C.R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31: 423–447.

Henderson, C.R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32: 69−83.

Henderson, C.R. 1977. Best linear unbiased prediction of breeding values not in the model for records. J. Dairy Sci. 60: 783−787.

Henderson, C.R. 1984. Applications of linear models in animal breeding. Univ. Guelph, Guelph, Ontario, Canada.

Henderson, C. R. 1985. Best linear unbiased prediction using relationship matrices derived from selected base populations. J. Dairy Sci. 68: 443–448.

Henderson, C. R. 1988. Theoretical basis and computational methods for a number of different animal models. J. Dairy Sci. 71: 1–16

Hickey, J.M., Crossa, J., Babu, R., de los Campos, G. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52: 654−663.

Higham N.J. 2002. Computing the nearest correlation matrix–a problem from finance. IMA J. Numer. Anal. 22: 329–343

Hill, W.G. 1993a. Variation in genetic composition in backcrossing programs. J. Hered. 84: 212–213.

Hill, W.G. 1993b. Variation in genetic identity within kinships. Heredity 71: 652–653.

Hill, W.G., Weir, B.S. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet. Res. 93: 47–64.

Jacquard, A. 1974. The genetic structure of populations. Springer-Verlag, New York.

Keith, J.M., McRae, A., Duffy, D., Mengersen, K., Visscher, P.M. 2008. Calculation of IBD probabilities with dense SNP or sequence data. Genet. Epidemiol., 32: 513–519.

Lander, E.S., Green, P. 1987. Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. 84: 2363–2367.

Legarra, A., Misztal, I. 2008. Technical note: Computing strategies in genome-wide selection. J. Dairy Sci. 91: 360−366.

Legarra, A., Aguilar, I., Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92: 4656−4663.

LeRoy, P., Elsen, J.M., Gilbert, H., Moreno, C.R., Legarra, A. et al. 2013. QTLMap 0.9.6 user's manual. Disponible en: http://www.inra.fr/qtlmap

Li, X., Yin, X., Li, J. 2010. Efficient identification of identical-by-descent status in pedigrees with many untyped individuals. Bioinformatics 26: 191−198.

Luan, T., Woolliams, J.A., Ødegård, J., Dolezal, M., Roman-Ponze, S.I., Bagnato, A. *et al*. 2012. The importance of identity-by-state information for the accuracy of genomic selection. Genet Sel Evol. 44: 28.

Malécot, G. 1948. Les mathématiques de l'hérédité. Masson et Cie, Paris.

McPeek, M.S., Wu, X., Ober, C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60: 359−367.

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

Meuwissen, T.H.E. 2003. Genomic selection: the future of marker assisted selection and animal breeding. In: Marker assisted selection: a fast track to increase genetic gain in plant and animal breeding? Session II: MAS in animals. FAO Eletronic Forum on Biotechnology in Food and Agriculture: Conference 10.

Meuwissen T.H.E., Goddard M.E. 2010. The use of family relationships and linkage disequilibrium to impute phase and missing genotypes in up to whole-genome sequence density genotypic data. Genetics 185: 1441−1449.

Meuwissen, T.H.E., Luan, T., Woolliams, J.A. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. J. Anim. Breed. Genet. 128: 429–439.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. et al. 2002. BLUPF90 and related programs (BGF90). CD-ROM, Communication No. 28−07, 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

Misztal, I., Legarra, A., Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92: 4648–4655.

Patterson, H., Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58: 545.

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

Quaas, R.L. 1988. Additive genetic model with groups and relationships. J. Dairy Sci. 71: 1338−1345.

Rabiner, L.R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77: 257–286.

Ramos, A.M., Crooijmans, R.P., Affara, N.A., Amaral, A.J., Archibald, A.L. *et al.* 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.

Risch, N., Lange, K. 1979. An alternative model of recombination and interference. Annals of Human Genetics 43: 61−70.

Sargolzaei, M., Schenkel, F.S. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680−681.

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

Searle, S.R., Casella, G., McCulloch, C.E. 1992. Variance components. John Wiley & Sons, New York.

Self, S.G., Liang, K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. J. Am. Stat. Assoc. 82: 605−610.

Sobel, E., Lange, K. 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. 58: 1323.

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

Strandén, I., Christensen, O.F. 2011. Allele coding in genomic evaluation. Genetics Selection Evolution, 43, 25.VanRaden, P. M., Wiggans, G. R. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74: 2737–2746.

Strandén, I., Garrick, D.J. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92: 2971–2975.

Thompson, E.A. 2008. Analysis of data on related individuals through inference of identity by descent. Technical Report Number 539. Department of Statistics, University of Washington.

Thompsom, E.A. 2013. Identity by descent: variation in meiosis, across genomes, and in populations. Genetics 194: 301–326.

Toro, M.Á., García-Cortés, L.A., Legarra, A. 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. Genet. Sel. Evol. 43: 27.

Tortereau, F., Servin, B., Frantz, L., Megens, H.J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A.L., Schook, L.B., Groenen, M. 2012. A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC genomics 13: 586.

VanRaden, P. M., Wiggans, G. R. 1991. Derivation, calculation, and use of national animal model information. J. Dairy Sci. 74: 2737–2746.

VanRaden, P.M. 2007. Genomic measures of relationship and inbreeding. Interbull Bull. 37: 33–36.

VanRaden, P.M. 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91: 4414−4423.

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

VanRaden, P., Cooper, T., Wiggans, G., O'Connell, J., Bacheller, L. 2013. Confirmation and discovery of maternal grandsires and great-grandsires in dairy cattle. J. Dairy Sci. 96: 1874−1879.

Vela-Avitúa, S., Meuwissen, T.H.E., Luan, T., Ødegård, J. 2015. Accuracy of genomic selection for a sib-evaluated trait using identity-by-state and identity-by-descent relationships. Genetics Selection Evolution, 47, 9.Villanueva, B., Pong-Wong, R., Fernández, J., Toro, M. A. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. 83: 1747−1752.

Villanueva, B., Pong-Wong, R., Fernández, J., Toro, M. A. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. 83: 1747−1752.

Visscher, P.M. 2006. A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Res. Hum. Genet. 9: 490−495.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W., Martin, N. G. 2006. Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. PLoS Genet. 2:e41.

Visscher, P.M. 2009. Whole genome approaches to quantitative genetics. Genetica 136: 351−358.

Vitezica, Z. G., Aguilar, I., Misztal, I., Legarra, A. 2011. Bias in genomic predictions for populations under selection. Genet. Res. 93: 357−366.

Wang, C., Habier, D., Peiris, B., Wolc, A., Kranis, A. et al. 2013. Accuracy of genomic prediction using an evenly spaced, low-density single nucleotide polymorphism panel in broiler chickens. Poult. Sci. 92: 1712−1723.

Wang, H., Misztal, I., Legarra, A. 2014. Differences between genomic‐based and pedigree‐based relationships in a chicken population, as a function of quality control and pedigree links among individuals. J. Anim. Breed. Genet. 131: 445−451.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D. et al. 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. J. Dairy Sci. 92: 3431−3436.

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

Wright, S. 1922. Coefficients of inbreeding and relationship. The American Naturalist 56: 330−338.

APÉNDICE I

I.1. Notación utilizada en el algoritmo de Li et al. (2010)

$p^{a,b}$	Un sendero de herencia en un grafo de descendencia que conecta *a* y *b*.
$\theta_{a,b}(h)$	Función generadora de herencia para dos alelos *a* y *b*.
$\Phi(p)$	Probabilidad de ocurrencia de un sendero de herencia *p*.
$\phi(p, i, i+1)$	Probabilidad de que *p* no se modifique del locus *i* al locus *i* + 1.
$\Psi(i, i+1)$	$P(a_{i+1} \equiv b_{i+1} \mid a_i \equiv b_i)$

I.2. Derivación de las probabilidades de transición del HMM alélico no derivadas en la sección 3.3.2.2.

En las expresiones 3.5, 3.8, 3.10 y 3.11 se derivaron 5 de las 9 probabilidades de transición del HMM para un par de alelos (Figura 3.4). En esa misma sección se determinó, además, que $P\left(\text{Bg - IBD}_{i+1} \mid \text{Bg - IBD}_i\right) = (1-r)^k$. Resta, entonces, derivar $P\left(\text{No - IBD}_{i+1} \mid \text{Bg - IBD}_i\right)$, $P\left(\text{Bg - IBD}_{i+1} \mid \text{No - IBD}_i\right)$ y $P\left(\text{No - IBD}_{i+1} \mid \text{No - IBD}_i\right)$.

Por teoría de probabilidad condicional, sabemos que

$$P\left(\text{Bg - IBD}_{i+1} \mid \text{Bg - IBD}_i\right) + P\left(\text{IBD}_{i+1} \mid \text{Bg - IBD}_i\right) + P\left(\text{No - IBD}_{i+1} \mid \text{Bg - IBD}_i\right) = 1 \qquad \text{[ii.1]}$$

Despejando en la expresión i.1 se obtiene

$$P\left(\text{No - IBD}_{i+1} \mid \text{Bg - IBD}_i\right) = 1 - P\left(\text{Bg - IBD}_{i+1} \mid \text{Bg - IBD}_i\right) - P\left(\text{IBD}_{i+1} \mid \text{Bg - IBD}_i\right) \qquad \text{[ii.2]}$$

La probabilidad $P\left(\text{Bg - IBD}_{i+1} \mid \text{No - IBD}_i\right)$ se deriva utilizando la definición de probabilidad condicional y la propiedad de reversibilidad de la cadena de Markov, i.e. $P\left(\text{Bg - IBD}_{i+1}, \text{No - IBD}_i\right) = P\left(\text{No - IBD}_{i+1}, \text{Bg - IBD}_i\right)$. Esto es,

$$P\left(\text{Bg - IBD}_{i+1}\middle|\text{No -IBD}_i\right)=\frac{P\left(\text{Bg - IBD}_{i+1},\text{No - IBD}_i\right)}{P\left(\text{No - IBD}\right)}$$
$$=\frac{P\left(\text{No - IBD}_{i+1}\middle|\text{Bg - IBD}_i\right)P\left(\text{Bg - IBD}\right)}{1-P\left(\text{IBD}\right)-P\left(\text{Bg - IBD}\right)} \qquad \text{[ii.3]}$$

Finalmente, la probabilidad $P\left(\text{No - IBD}_{i+1}\middle|\text{No - IBD}_i\right)$ se obtiene por diferencia, siendo

$$P\left(\text{No - IBD}_{i+1}\middle|\text{No - IBD}_i\right)=1-P\left(\text{IBD}_{i+1}\middle|\text{No - IBD}_i\right)-P\left(\text{Bg - IBD}_{i+1}\middle|\text{No - IBD}_i\right) \qquad \text{[ii.4]}$$

De este modo, las probabilidades de transición $P\left(\text{No - IBD}_{i+1}\middle|\text{Bg - IBD}_i\right)$, $P\left(\text{Bg - IBD}_{i+1}\middle|\text{No - IBD}_i\right)$ y $P\left(\text{No - IBD}_{i+1}\middle|\text{No - IBD}_i\right)$ quedan expresadas en función de las probabilidades derivadas en la sección 3.3.2.2.

II.3. Derivación de las probabilidades de emisión del HMM para un par de individuos.

Considérese el muestreo ordenado de cuatro alelos (p^A, m^A, p^B, m^B) en un locus asociado a un SNP en un par de individuos *A* y *B* no consanguíneos. Cada alelo puede tomar dos valores: 1 ó 2 con probabilidad *p* y *q*, respectivamente. Existen 16 resultados posibles del muestreo (Cuadro I.1, primera columna). La probabilidad de cada resultado se calcula condicionando en los estados de IBD, $\boldsymbol{s}$ (Figura 3.6B), teniendo en cuenta que los alelos que son No-IBD son condicionalmente independientes. Por ejemplo, la probabilidad de muestrear 4 alelos condicional al estado $\boldsymbol{s}$ = (0, 0, 0, 0) es el producto de las frecuencias alélicas poblacionales (Cuadro I.1, tercera columna). Ahora bien, las probabilidades de emisión $P(G(A, B)|\ \boldsymbol{s})$ se obtienen a partir de las probabilidades condicionales $P((p^A, m^A, p^B, m^B)|\ \boldsymbol{s})$, sumando sobre todos los resultados que tengan el mismo valor de G(*A*, *B*):

$$P\left(G\left(A,B\right)=x\middle|\boldsymbol{s}\right)=\sum_{\left(p^A,m^A,p^B,m^B\right)\therefore G(A,B)=x}P\left(\left(p^A,m^A,p^B,m^B\right)\middle|\boldsymbol{s}\right) \qquad x=0,1,2 \quad \text{[ii.5]}$$

Por ejemplo, las probabilidades de emisión de $G(A, B)$ dado el estado $\boldsymbol{s}$ = (0, 0, 0, 0) se muestran en la Cuadro I.1.

Considérese ahora el estado $\boldsymbol{s}$ = (1, 0, 0, 0) en el cual los alelos paternos de ambos individuos son IBD. Condicional dicho estado, hay un par de alelos que no segrega independientemente y ese par debe ser necesariamente IBS. De ello se deduce que $P(G(A, B) = 0 \mid \boldsymbol{s} = (1, 0, 0, 0)) = 0$. Ahora bien, utilizando la expresión ii.5,

$$P\left(G(A,B)=1\middle|\boldsymbol{s}=(1,0,0,0)\right)=$$
$$P\left((1,1,1,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((1,1,2,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((1,2,1,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((2,1,1,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)+$$
$$P\left((1,2,2,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((2,1,2,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((2,2,1,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((2,2,2,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)$$
$$=p^2q+0+p^2q+0+0+q^2p+0+q^2p=2p^2q+2q^2p=2pq(p+q)=2pq$$

Y del mismo modo,

$$P\left(G(A,B)=2\middle|\boldsymbol{s}=(1,0,0,0)\right)=$$
$$P\left((1,1,1,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((2,2,2,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((1,2,1,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((1,2,2,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)+$$
$$P\left((2,1,1,2)\middle|\boldsymbol{s}=(1,0,0,0)\right)+P\left((2,1,2,1)\middle|\boldsymbol{s}=(1,0,0,0)\right)$$
$$=p^3+q^3+pq^2+0+0+p^2q=p^2(p+q)+q^2(p+q)=p^2+q^2$$

Cuadro I.1. Distribución de probabilidad del muestreo ordenado de cuatro alelos condicional al estado de IBD $\boldsymbol{s}$ = (0, 0, 0, 0) y probabilidades de emisión de $G(A, B)$ dado el estado $\boldsymbol{s}$ = (0, 0, 0, 0).

(p^A, m^A, p^B, m^B)	$G(A, B)$	$P((p^A, m^A, p^B, m^B) \mid \boldsymbol{s} = (0,0,0,0))$	$P(G(A, B) \mid \boldsymbol{s} = (0,0,0,0))$
1, 1, 2, 2	0	p^2q^2	$2p^2q^2$
2, 2, 1, 1	0	p^2q^2	
1, 1, 1, 2	1	p^3q	$4p^3q + 4pq^3$
1, 1, 2, 1	1	p^3q	
1, 2, 1, 1	1	p^3q	
2, 1, 1, 1	1	p^3q	
1, 2, 2, 2	1	pq^3	
2, 1, 2, 2	1	pq^3	
2, 2, 1, 2	1	pq^3	
2, 2, 2, 1	1	pq^3	
1, 1, 1, 1	2	p^4	$p^4 + 4p^2q^2 + q^4$
2, 2, 2, 2	2	q^4	
1, 2, 1, 2	2	p^2q^2	
1, 2, 2, 1	2	p^2q^2	
2, 1, 1, 2	2	p^2q^2	
2, 1, 2, 1	2	p^2q^2	

Las fórmulas obtenidas para las probabilidades de emisión dado el estado $\boldsymbol{s} = (1, 0, 0, 0)$ son iguales a las que se obtienen partiendo de cualquier estado que cumpla con la condición de que $I(A, B) = 1$ (estados $s_2 - s_9$, Figura 3.6B).

Considérese ahora el estado $\boldsymbol{s} = (1, 0, 0, 1)$ en el cual los alelos paternos y los alelos maternos de ambos individuos son IBD. Condicional a dicho estado, necesariamente debe haber 2 pares de alelos IBS. De ello se deduce que $P(G(A, B) = 0| \boldsymbol{s} = (1, 0, 0, 1)) = P(G(A, B) = 1| \boldsymbol{s} = (1, 0, 0, 1)) = 0$ y $P(G(A, B) = 2| \boldsymbol{s} = (1, 0, 0, 1)) = 1$. Nuevamente, las fórmulas obtenidas para las probabilidades de emisión dado el estado $\boldsymbol{s} = (1, 0, 0, 1)$ son iguales a las que se obtienen partiendo de cualquier estado que cumpla con la condición de que $I(A, B) = 2$ (estados $s_{10} - s_{17}$, Figura 3.6B).

Si bien no se explicarán aquí los detalles, el algoritmo de Li *et al.* (2010) puede calcular las probabilidades de emisión teniendo en cuenta el efecto de genotipos perdidos en algunos SNPs y errores de genotipado.

APÉNDICE II

II. 1. Covarianza entre los valores de w_{ik} y w_{jk} correspondientes a los genotipos de dos animales en un locus, condicional a la probabilidad de IBD.

Partiendo de la expresión 5.5 obtenida en la sección 5.3.1 y expandiendo sobre los posibles resultados del muestreo de un par de genes en ambos individuos.

$$\begin{aligned}\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i\equiv j\right)=&\Big[\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i\equiv j,i=i_S,j=j_S\right)+\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i\equiv j,i=i_S,j=j_D\right)\\&+\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i\equiv j,i=i_D,j=j_S\right)+\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i\equiv j,i=i_D,j=j_D\right)\Big]\\=&\Big[\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i_S\equiv j_S\right)+\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i_S\equiv j_D\right)+\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i_D\equiv j_S\right)\\&+\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i_D\equiv j_D\right)\Big]\end{aligned}$$

[ii.1]

Tomemos, por ejemplo el término Cov(w_{ik}, w_{jk} | $i_S \equiv j_S$). Por definición de covarianza,

$$\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i_S\equiv j_S\right)=E\left(w_{ik},w_{jk}\,|\,i_S\equiv j_S\right)-E\left(w_{ik}\right)E\left(w_{jk}\right) \qquad \text{[ii.2]}$$

A partir de la distribución de probabilidad conjunta de los genotipos, condicional a que los alelos paternos de ambos individuos son IBD, i.e. P(w_{ik}, w_{jk} | $i_S \equiv j_S$),

	w_{jk}	1	0	−1
w_{ik}		A_1A_1	A_1A_2	A_2A_2
1	A_1A_1	p^3	p^2q	0
0	A_1A_2	p^2q	pq	q^2p
−1	A_2A_2	0	q^2p	q^3

se obtiene $E\left(w_{ik},w_{jk}\,|\,i_S\equiv j_S\right)=p_k^3+q_k^3$. Utilizado esta expresión y recordando que $E\left(w_{ik}\right)=p_k-q_k$, remplazamos en la ecuación ii.2, y luego de cierta algebra se obtiene

$$\mathrm{Cov}\left(w_{ik},w_{jk}\,|\,i_S\equiv j_S\right)=p_k^3+q_k^3-\left(p_k-q_k\right)^2=pq \qquad \text{[ii.3]}$$

Se puede demostrar que las probabilidades condicionales son las mismas cuando se muestrea otro par de alelos, i.e.

$$P\left(w_{ik}, w_{jk} \middle| i_S \equiv j_S\right) = P\left(w_{ik}, w_{jk} \middle| i_S \equiv j_D\right) = P\left(w_{ik}, w_{jk} \middle| i_D \equiv j_S\right) = P\left(w_{ik}, w_{jk} \middle| i_D \equiv j_D\right) \quad \text{[ii.4]}$$

Por lo tanto, la expresión ii.3 vale para los otros tres términos de la ecuación ii.1.

$$\mathrm{Cov}\left(w_{ik}, w_{jk} \middle| i \equiv j\right) = 4\mathrm{Cov}\left(w_{ik}, w_{jk} \middle| i_S \equiv j_S\right) = 4pq \quad \text{[ii.5]}$$