Journal of
Animal Breeding and Genetics

J. Anim. Criar. Gineta. ISSN 0931-2668

ARTÍCULO ORIGINAL

Reducción de dimensiones y selección de variables para la selección genómica: aplicación para predecir la producción de leche en vacas Holstein N. Long1 , D. Gianola1,2,3, GJM Rosa1,3 y KA Weigel2

1 Departamento de Ciencias Animales, Universidad de Wisconsin, Madison, WI, EE. UU.
2 Departamento de Ciencias Lácteas, Universidad de Wisconsin, Madison, WI, EE. UU.
3 Departamento de Bioestadística e Informática Médica, Universidad de Wisconsin, Madison, WI, EE. UU.

Palabras
clave reducción de dimensiones; selección genómica; selección de variables.

Correspondencia N.
Long, Departamento de Ciencias Animales, Universidad de Wisconsin, Madison, WI 53706, EE. UU. Teléfono: +1 (608) 263-3499; Fax: +1 (608) 263-5157; Correo electrónico: nlong@wisc.edu

Recibido: 3 de septiembre de 2010;
aceptado: 5 enero 2011

Resumen

La predicción asistida por el genoma del mérito genético de los individuos para un rasgo cuantitativo requiere la construcción de modelos estadísticos que puedan manejar conjuntos de datos que consisten en una gran cantidad de marcadores y muchas menos observaciones. Se han propuesto numerosos modelos de regresión en los que los efectos de los marcadores se tratan como variables aleatorias. Alternativamente, las técnicas de reducción de dimensiones multivariadas [como la regresión de componentes principales (PCR) y la regresión de mínimos cuadrados parciales (PLS)] modelan una pequeña cantidad de componentes latentes que son combinaciones lineales de variables originales, lo que reduce la dimensionalidad. Además, la selección de marcadores ha llamado cada vez más la atención en la selección genómica. Este estudio evaluó métodos de reducción de dos dimensiones, a saber, PCR supervisada y PLS disperso, para predecir valores genéticos genómicos (BV) de toros lecheros para la producción de leche utilizando polimorfismos de un solo nucleótido (SNP). Estos dos métodos realizan una selección de variables de forma además de reducir la dimensionalidad. La PCR supervisada preselecciona los SNP en función de la fuerza de asociación de cada SNP con el fenotipo. Sparse PLS promueve la escasez al imponer alguna penalización a los coeficientes de las combinaciones lineales de las variables SNP originales. Se examinaron dos tipos de PCR supervisadas (I y II).

El método I se basó en análisis de un solo SNP, mientras que el método II se basó en análisis de múltiples SNP. La PCR II supervisada fue claramente mejor que la PCR I supervisada en capacidad predictiva cuando se evaluó en subconjuntos de SNP de varios tamaños, y el PLS disperso estaba en el medio. La PCR II supervisada y el PLS disperso lograron correlaciones predictivas similares cuando el tamaño del subconjunto de SNP era inferior a 1000. La PCR II supervisada con 300 y 500 SNP logró correlaciones de 0,54 y 0,59, respectivamente, correspondientes al 80 y el 87 % de la correlación (0,68) obtenido con los 32 518 SNP en un modelo de PCR. La correlación predictiva de la PCR II supervisada alcanzó una meseta de 0,68 cuando el número de SNP aumentó a 3500. Nuestros resultados demuestran el potencial de combinar la reducción de dimensiones y la selección de variables para una predicción precisa y rentable de la VB genómica.

Introducción

En modelos de regresión para rasgos cuantitativos que utilizan una gran cantidad de marcadores genéticos [por ejemplo, polimorfismos de un solo nucleótido (SNP)] como variables predictoras, la multicolinealidad ocurre porque los marcadores están intercorrelacionados. Esto es atribuible al desequilibrio de ligamiento, es decir, la asociación no aleatoria de alelos en dos o más loci. Una consecuencia bien conocida de la multicolinealidad en la regresión de mínimos cuadrados es

 doi:10.1111/j.1439-0388.2011.00917.x

estimaciones inestables. Infla las varianzas de las estimaciones.
de coeficientes de regresión (por ejemplo, efectos de marcador) debido a la casi singularidad de la matriz de incidencia. Varios Se han propuesto métodos para manejar la multicolinealidad, como la regresión de crestas, componente principal regresión (PCR) (Massy 1965) y parcial mínimo regresión de cuadrados (PLS) (Wold 1985). PCR y PLS son métodos bien documentados en quimiometría (Frank & Friedman 1993), y también han encontrado aplicaciones en bioinformática (por ejemplo, Boulesteix & Strimmer 2007), análisis de neuroimagen (por ejemplo, McIntosh et al. 1996) y recientemente en la predicción asistida por genoma de los valores genéticos (BV) (Solberg et al. 2009; Macciotta et al. 2010). Además de la multicolinealidad, Otro problema desafiante es que los datos típicamente contienen muchas más variables (p, por ejemplo, genes o marcadores) que observaciones (n, por ejemplo, individuos), inhabilitando el uso de la metodología de mínimos cuadrados. Por ejemplo, es común tener genotipos para cientos de miles de marcadores SNP en solo cientos de animales en selección genómica.

Para reducir la dimensión del modelo y superar la problema de multicolinealidad, transformada PCR y PLS el gran número de variables originales en un número relativamente pequeño de componentes latentes ortogonales y luego retroceder la variable de respuesta en esos componentes latentes. Cuando se aplica al análisis de todo el genoma con n<<p, un pequeño número (en relación con n) de los componentes latentes suele ser suficiente para la modelo para lograr un desempeño predictivo competitivo. Por lo tanto, la carga computacional puede ser reducido con PCR y PLS, en comparación con los métodos que estiman todos los efectos de los marcadores tratándolos como variables aleatorias, como Bayes A (Meuwissen et al. 2001) o la contracción absoluta mínima bayesiana y operador de selección (Lasso) (Park & Casella 2008; de los Campos et al. 2009).

En las aplicaciones prácticas de la selección habilitada por el genoma, a menudo se desea genotipar un subconjunto de marcadores SNP de genoma completo para reducir el genotipado costo, especialmente para la evaluación inicial de los animales. Para modelos de regresión basados en SNP comúnmente utilizados, tiene demostrado (Long et al. 2007; Habier et al. 2009; Usai et al. 2009;Weigel et al. 2009) que, con una cuidadosa subconjunto SNP elegido, uno puede lograr un razonablemente alta fiabilidad de la BV predicha. Por ejemplo, Weigel et al. (2009) encontraron que usando los 300 SNPs que tenían los mayores efectos estimados de un bayesiano completo Regresión de Lasso con los 32 518 SNP, la predicción la fiabilidad fue la mitad de la obtenida con todos los SNP. Usai et al. (2009) encontraron en simulaciones que la precisión del BV estimado alcanzó un máximo cuando Se utilizaron 169 marcadores, los cuales fueron seleccionados de 6000 SNP utilizando un algoritmo de regresión de ángulo mínimo de Lasso.

Aunque PCR y PLS pueden reducir eficientemente dimensión del modelo, no ofrecen la posibilidad de seleccionando variables importantes, porque cada latente componente es una combinación lineal de todas las variables originales. Es deseable tener ambas dimensiones del modelo y número de variables reducido. Una forma sencilla es ignorar variables con coeficientes de pequeña magnitud en los componentes latentes, pero esto a menudo conduce a una mala interpretación de las variables originales (Cadima & Jolliffe 1995). En el contexto del análisis de la expresión génica, Bair et al. (2006) propusieron la noción de supervisión análisis de componentes principales (PCA) para PCR, es decir, preseleccionar un subconjunto de genes antes de formar componentes latentes, con la selección basada en la fuerza de asociación de cada gen con el resultado. Porque de este procedimiento de preselección 'supervisado' (es decir, guiado por el resultado), los componentes latentes resultantes tuvo una mejor precisión de predicción para el resultado que que se alcanza con todos los genes. Para PLS, Chun y Keleş (2010) desarrollaron una metodología novedosa (el 'escaso PLS') para realizar reducción de dimensión y variable selección simultáneamente. Brevemente, el PLS disperso promueve la escasez al imponer alguna penalización sobre los coeficientes de las combinaciones lineales de los originales. variables La utilidad de estos métodos ha sido demostrada en aplicaciones a datos de micromatrices (Chen et al. 2008; Chun & Keles (2009).

Para la predicción de BV asistida por genoma, el problema de La selección de subconjuntos de SNP no ha sido investigada en el marco de métodos de variables latentes para la dimensión reducción. Por lo tanto, el objetivo de este estudio fue aprovechar los avances en los métodos que combinan la reducción de dimensiones y la selección de variables para analizar datos genómicos de alta dimensión presentados por chips SNP modernos. Dos variantes de PCR supervisada y el modelo de regresión PLS disperso de Chun & Keleş (2010) fueron examinados usando datos de lácteos vacas. El objetivo era predecir la predicción de los toros capacidad de transmisión (PTA, la mitad de la BV) para la leche rendimiento derivado de pruebas de progenie, utilizando marcadores SNP. Este estudio evaluó el rendimiento predictivo entregado por cada uno de los tres modelos instalados en SNP subconjuntos de varios tamaños para juzgar la eficacia del uso de subconjuntos de SNP para la selección genómica.

Materiales y métodos

Datos

Genotipos SNP de alta densidad y PTA para la producción de leche (derivado de registros de rendimiento de crías hembras

a través de pruebas de progenie) de toros Holstein se obtuvieron del Laboratorio de Genómica Funcional Bovina y del Laboratorio de Programas de Mejoramiento Animal, respectivamente, en el Centro de Investigación Agrícola Beltsville del USDA-ARS (Beltsville, MD). Después de la edición, los datos finales utilizados en el análisis consistieron en 32 518 SNP y 4703 toros; Se utilizaron 3305 toros nacidos entre 1952 y 1998 como conjunto de entrenamiento para construir modelos, y los 1398 toros restantes nacidos entre 1999 y 2002 se designaron como conjunto de prueba para la evaluación del modelo.
Los PTA de prueba de progenie de agosto de 2003 y los PTA de prueba de progenie de abril de 2008 estaban disponibles para los toros en los conjuntos de entrenamiento y prueba, respectivamente. Weigel et al. (2009) utilizaron los mismos datos genotípicos con un rasgo diferente (mérito neto de por vida), y allí se puede encontrar una descripción detallada del preprocesamiento de datos.

Métodos

Los modelos de PCR supervisada y PLS disperso se ajustaron a los datos de los 3305 toros de entrenamiento utilizando diferentes números de SNP seleccionados. La comparación entre los métodos se basó en su capacidad predictiva, evaluada por las correlaciones de Pearson entre los PTA de pruebas de progenie realizados y los PTA predichos por PCR supervisada o PLS dispersa en los toros de prueba 1398 (en lo sucesivo se utiliza "correlación predictiva").

También se ajustaron PCR y PLS ordinarios utilizando los 32ÿ518 SNP, de modo que se pudo evaluar el potencial de los enfoques de selección de SNP que se estaban examinando.
En esta sección, se revisan primero los PCR y PLS ordinarios. Luego, se dan los elementos esenciales de la PCR supervisada y el PLS disperso. Por último, se describe brevemente la regresión bayesiana de Lasso, ya que se utilizó para estimar los coeficientes de regresión en todos los modelos de PCR en este estudio.

Siguiendo las convenciones de la literatura, en lo que sigue, los componentes latentes en PCR se denominarán "componentes principales" o PC; para PLS, se utilizará el término "componentes latentes". La estrategia para encontrar PC/componentes latentes difiere entre PCR y PLS. PCR considera maximizar las varianzas de las PC, mientras que PLS busca componentes latentes que tengan grandes varianzas así como altas correlaciones con la variable respuesta.

En el contexto de los datos de marcadores genéticos, defina una matriz n · p, X, que consta de genotipos de p marcadores SNP para cada uno de los n individuos. Cada elemento de X toma el valor 0, 1 o 2, según el número de copias del alelo observado. Las respuestas (ACP de pruebas de progenie realizadas) se almacenan en un vector de columna n · 1 y. Sea r el rango de columna de X.

Derivación de componentes principales en PCR

El primer paso de PCR es aplicar PCA a X (asumiendo que se ha centrado en la columna de antemano) para extraer un número (K, K < r) de PC, siendo cada una de ellas una combinación lineal de las columnas de X. La matriz T(n·K), que contiene estos PC, se forma como T ¼ XP, donde P(p · K) es la matriz de carga y se puede obtener a partir de la descomposición espectral de X¢ X/(n) 1) (covarianza muestral de X) o descomposición de valor singular de X. El enfoque anterior da X0 X=ðn 1Þ ¼ PrKP0 donde las columnas de Pr consisten en vectores propios ortonormales, y K ¼ diag(k1,…,kr) contiene el valor propio correspondiente, ordenado de modo que P tomando las primeras K columnas de Pr. Desde una perspectiva de optimización, cada columna de P, pk (1 £ k £ K), satisface

$$p_k = \arg\max_{p} \mathrm{var}(Xp); \text{ sujeto a } p'p = 1;\ f'p_l = 0 \quad l = 1 \quad (1)$$

donde var(Xpk) denota la varianza muestral del k-ésimo PC (k ¼ 1,…,K). Del criterio y de las dos restricciones, se deduce que los PC tienen varianzas máximas sucesivas y son mutuamente ortogonales. Específicamente, la varianza del k-ésimo PC es igual al k-ésimo valor propio, porque varðXpkÞ ¼ p0 kX0 Xpk=ðn 1Þ ¼ p0 kðPrKP0 rÞpk ¼ kk.

Derivación de componentes latentes en PLS

PLS se basa en las suposiciones de que X ¼ TP¢ +E y y ¼ Tm + f, suponiendo que tanto X como y están centrados en la columna. Aquí, T(n · K) es la matriz de componentes latentes; P(p · K) y m(K · 1) a menudo se denotan como 'cargas X' y 'cargas Y', respectivamente; E(n · p) y f(n · 1) son errores aleatorios.

Esta descomposición indica que la relación entre X e y se transmite a través de T. Los enfoques PLS utilizados con frecuencia incluyen PLS1 (Manne 1987; Helland 1988) y SIMPLS (de Jong 1993). Para una variable de respuesta univariante (como en el estudio actual), las dos son equivalentes (de Jong 1993), con el objetivo de encontrar una matriz de peso W(p · K) tal que su columna k-ésima wk(k ¼ 1, 2,…,K) cumple

$$w_k = \arg\max_{w} \mathrm{corr}^2(y, Xw)\mathrm{var}(Xw); \quad \text{sujeto a } w'w = 1;\ f'Xw_l = 0 \quad l = 1 \quad (2)$$

La solución, Wb , luego se utiliza para construir TðT ¼ XWb Þ, que tiene columnas mutuamente ortogonales como lo indica la segunda restricción.
El vector de carga m se estima resolviendo el

problema de regresión y ¼ Tm+f mediante mínimos cuadrados, dando m^ ¼ ðT0 TÞ T0 y. Esto conduce a m^ ¼ ðX^ bPLS. Este significa que 1 T0 y. El modelo de regresión final ^y ¼ ðWb m^ Þ puede considerarse como una estimación PLS de b, los coeficientes de regresión con respecto a las variables X originales.

Al resolver el problema de optimización (2), es útil obtener una expresión explícita de la función objetivo. Teniendo en cuenta que y y, por lo tanto, var(y) son constantes dados los datos, el problema de maximización original, argmaxwcorr2 (y,Xw)var(Xw), es equivalente

prestado a

argmaxwcorr2 ðy;XwÞvarðXwÞvarðyÞ ¼ cov2 ðXw; yÞ;

y, además, a argmaxww¢X¢yy¢Xw. Esta forma explícita de función objetivo se usará más adelante para describir el PLS disperso. En pocas palabras, en PLS los componentes latentes no son redundantes y tienen una covarianza máxima con la respuesta. Como tales, se espera que tengan un alto poder predictivo.

Predicción.

La principal aplicación de PCR y PLS es utilizar el modelo de regresión final con fines de predicción. El segundo paso de PCR (después del primer paso de PCA) consiste en reemplazar la X original por T y estimar los coeficientes asociados a ella. el modelo es

y ¼ 1nl þ Tg þ e ¼ 1nl þ XðPgÞ þ e; ð3Þ

donde l es la media global, g(K · 1) es un vector de coeficientes a estimar y e es un vector de error. Aquí, (Pg) puede considerarse como bPCR, los coeficientes de regresión con respecto a las variables X originales. Dada una nueva observación x* (p · 1), su respuesta predicha es ^y ¼ l^ þ ðx x Þ umn vector medio promediado sobre las muestras de entrenamiento, donde es el promedio de y debido a las 0 PPg, x y l^ es la col medias de columna cero de X Los coeficientes g en (3) se pueden estimar por mínimos cuadrados ordinarios o métodos de contracción bayesiana. En nuestro análisis, se adoptó el Bayesian Lasso porque produjo modelos con un mejor rendimiento predictivo que los de la estimación por mínimos cuadrados (datos no mostrados).

m^ La predicción para PLS viene dada por ^y ¼ l^ þ ðx x Þ donde 0 Wb , l^ es el promedio de los valores de y; Wb y m^ están fácilmente disponibles después de aplicar un algoritmo PLS a los datos de entrenamiento.

PCR supervisada El paso de selección de SNP en la PCR supervisada, como en Bair et al. (2006) y Chen et al. (2008), se basa al asociarse con el fenotipo de cada SNP.

La asociación podría medirse mediante la estadística t (^ b=s:e:ð^ bÞ) de, por ejemplo, una regresión de marcador único. Luego, se elige un umbral en el valor absoluto de la estadística t para que solo los marcadores SNP cuyo |t| los valores están por encima del umbral se conservan. Si uno está interesado en obtener un solo subconjunto (sin importar su tamaño), el umbral debe elegirse cuidadosamente para maximizar el mejor poder predictivo del subconjunto final. Bair et al. (2006) y Chen et al. (2008) propusieron métodos para la selección del umbral.

Para el estudio actual, los tamaños de los subconjuntos de SNP estaban predeterminados, lo que hacía innecesaria la elección de un umbral óptimo. Por lo tanto, nuestro procedimiento fue utilizar una medida de asociación para clasificar todos los SNP, y se seleccionó un número específico de SNP mejor clasificados. Se consideraron dos medidas.

1. La primera estrategia utilizaba la mencionada |t| de regresión de marcador único de mínimos cuadrados y se denominó PCR I supervisada.
2. La segunda estrategia para clasificar los SNP se basó en un modelo de PCR completo que involucraba a todos los SNP. En concreto, se extrajeron 3000 PC de la matriz 3305 · 32 518 X (genotipo SNP). Los coeficientes de regresión estimados (mediante Bayesian Lasso) de estos PC se transformaron nuevamente en coeficientes para los SNP originales. Las magnitudes de estos coeficientes de SNP se usaron luego para clasificar y seleccionar los SNP. Este procedimiento se denominó PCR II supervisada.

Dado un subconjunto de SNP seleccionados, se ajustó y utilizó un modelo de PCR con Bayesian Lasso como método para estimar los coeficientes de regresión de las PC para predecir los datos de prueba. Se probaron subconjuntos de SNP de diferentes tamaños (300, 500, 1000, 1500, 2000,…, 8000) para los dos métodos de PCR supervisados. Para un subconjunto de p SNP, el número de PC ajustados se incrementó gradualmente desde un valor pequeño a un valor grande [pero no más grande que min (3305, p)], y las predicciones se hicieron en consecuencia.

Sparse PLS

El sparse PLS desarrollado por Chun & Keles¸ (2010) tiene como objetivo producir escasez en las variables originales al imponer una penalización L1 a los vectores de peso de PLS.

Dado el número (K) de componentes latentes a extraer, los vectores de peso K se calculan secuencialmente optimizando algunas funciones objetivo.

A continuación se describe la derivación del primer vector de peso. El procedimiento completo se describe en el Apéndice I.

La función objetivo para el primer vector de peso es

f jw0 Mw þ ð1 jÞðc wÞ min w;c ^Mac WÞ
4Þ
+ k1jjcjj1 + k2jjcjj2 2g; sujeto a w0 w ¼ 1;

donde jjcjj1 ¼ Pp ;M¼X¢ yc¢ Xjcj2 0 < j < 1), k1 (k1 ‡ 0) y k2 (k2 ‡ 0) son tres parámetros de ajuste. Esta formulación impone la penalización L1 no a w, sino a su vector sustituto c, manteniendo c y w cerca uno del otro. La penalización L1 (k1) promueve la escasez en c (estableciendo algunos componentes en cero), mientras que la penalización L2 (k2) supera la singularidad potencial de M cuando se resuelve para c. Como se puede ver en la ecuación (4), cuando j ¼ 1 el PLS disperso se reduce al problema PLS ordinario, es decir, argmaxww¢ X¢ yy¢ Xw (ver "Derivación de componentes latentes en PLS").

Como señalaron Chun & Keles¸ (2010), la motivación de usar (4) en lugar de un criterio más simple como

como

máximo ðw0 MwÞ; sujeto a w0 w ¼ 1; jjwjj1 para;
En

es que la solución de w en la formulación anterior tiende a no ser lo suficientemente escasa. Esta observación se encontró cuando Jolliffe et al. (2003) propusieron una técnica de cargas dispersas para PCA (SCoTLASS).

En (4), la solución ^c se usa como el primer vector de peso PLS disperso estimado. En el caso de y univariado, Chun & Keles¸ (2010) demostraron que ^c depende solo de un nuevo parámetro de ajuste g (0 £ g £ 1) y toma la forma de un estimador de umbral suave:

c^i ¼ juij g max 1 jp jujj yo juij g max signo jujjðuiÞ: ð5Þ
1 JP

Arriba, ui es el i-ésimo elemento del vector u, con u ¼ X¢ y/||X¢ y||; I() es una función indicadora, tomando el valor de 1 si su argumento es verdadero y 0 en caso contrario. Este estimador retiene los componentes de peso que son mayores que una fracción del valor máximo del componente, lo que lleva a un vector de peso disperso con ceros que indican que las variables SNP correspondientes se descartan. Chun & Keles¸ (2010) proporcionaron un algoritmo para implementar el PLS disperso, y su versión univariada se proporciona en el Apéndice I. Después de ajustar el modelo PLS disperso, se puede obtener un vector de coeficientes de regresión con respecto a las variables originales. , que se puede utilizar directamente para la predicción. Como resultado de la selección de variables, algunos de los coeficientes estimados son exactamente cero.

El PLS disperso se implementó utilizando el paquete R spls (http://www.stat.wisc.edu/chungdon/spls/). Todo el procedimiento requiere la especificación de dos ajustes parámetros, el g antes mencionado y el número de componentes latentes deseados, K. Diferentes valores para estos parámetros conducen a la selección de diferentes números de SNP, y es más fácil predefinir un límite superior que un valor exacto para el tamaño del subconjunto.

Por lo tanto, los 17 tamaños de subconjunto elegidos previamente para la PCR supervisada se utilizaron aquí como límites superiores. Dado un límite superior, se compararon diferentes combinaciones de g y K que generaron subconjuntos de tamaños más pequeños que este límite, y la que produjo la correlación de validación cruzada (CV) más alta se usó para construir un modelo PLS disperso para la predicción final en la prueba. datos.

Bayesian Lasso para estimar los coeficientes de regresión Considere un modelo de regresión lineal y ¼ Xb+e, donde X es una matriz de incidencia n · p, b es un vector ap · 1 de coeficientes desconocidos y los errores en e son independientes e idénticamente distribuidos como Nð0; Ir2 eÞ.
Bayesian Lasso asigna la misma distribución previa exponencial doble a cada elemento de b, bj(j ¼ 1,…,p).
Esto es equivalente a los siguientes dos pasos (Park & Casella 2008):

pðbjjsjÞ Nð0; s2Þ j ¼ 1; ... ; p
pðs2 1kÞ Exponencialðk2Þ; j ¼
j

Por lo general, a k se le asigna un Gamma previo con sus hiperparámetros (forma y velocidad) ajustados por el usuario. Los detalles sobre el muestreo de Gibbs para implementar Bayesian Lasso se pueden encontrar en Park & Casella (2008) y de los Campos et al. (2009). Como se señaló, el Bayesian Lasso es una forma de lograr una contracción diferencial de los coeficientes (b). En relación con una distribución anterior normal, la distribución exponencial doble produce una contracción más fuerte de los coeficientes que están cerca de cero y una contracción menor de aquellos con valores absolutos grandes (de los Campos et al. 2009). En nuestro análisis bayesiano, el anterior para la varianza del error, r2 era una distribución chi-cuadrado invertida escalada con df ¼ 0,002 y escala ¼ 1, para k, se usó una distribución Gamma vaga con forma ¼ 1 y tasa ¼ 0.0001 como distribución previa. El muestreo de Monte Carlo de la cadena de Markov se ejecutó durante 50 000 iteraciones con las primeras 30 000 como quemado. El resto de las iteraciones se diluyeron a una tasa de 20. La media posterior (después de quemado y adelgazamiento) de cada parámetro se usó como su estimación.

Resultados y discusión

PCR y PLS utilizando todos los SNP

Vázquez et al. (2010) informaron una correlación predictiva de 0,69 al ajustar un modelo bayesiano de Lasso con

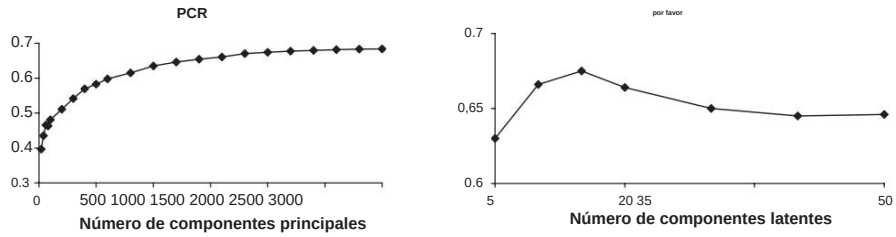

Figura 1 Correlación predictiva (correlación entre la capacidad de transmisión predicha (PTA) de la prueba de progenie y la PTA predicha en toros de prueba) contra el número de componentes principales/componentes latentes para PCR/PLS cuando se usaron los 32 518 SNP.

todos los SNP utilizando los mismos conjuntos de datos de entrenamiento y prueba que en el presente estudio. La Figura 1 muestra las correlaciones predictivas de PCR y PLS usando los 32 518 SNP. Se instalaron diferentes números de PC/componentes latentes. Para PCR, la correlación siguió aumentando con el número de PC hasta que se alcanzó una meseta entre 2000 y 2500 PC, donde la correlación fue de 0,68. Para PLS, la correlación más alta (0,67) se obtuvo con solo 15 componentes latentes.

A medida que se añadían más componentes latentes al modelo, la correlación empezaba a disminuir. En relación con los 32 518 SNP, los 2000 PC y los 15 componentes latentes representaron una gran reducción en el número de parámetros del modelo. Por lo tanto, con PCR/PLS, se puede reducir la dimensión del modelo sin afectar la capacidad predictiva. Esta conclusión también fue apoyada por Macciotta et al. (2010). Por otro lado, para lograr una correlación predictiva similar, la cantidad de PC necesarias para PCR fue mucho mayor que la cantidad de componentes latentes necesarios para PLS, lo que sugiere que los componentes latentes tenían un poder predictivo mucho mayor que las PC. Esto era de esperarse, porque la construcción de PC se basa únicamente en información sobre predictores, mientras que la construcción de componentes latentes también tiene en cuenta la variable de respuesta.

PCR supervisada (I y II)

De manera similar a la PCR que usa todos los SNP, las correlaciones predictivas del procedimiento de PCR supervisado (I y II) alcanzaron una meseta después de una fase creciente, a medida que aumentaba el número de PC. Esto se muestra en la Figura 2, utilizando los resultados del ajuste de un subconjunto de 3000 SNP como ejemplo. La meseta se alcanzó aproximadamente en 1000 para PCR supervisada I y II; el método II fue claramente mejor que el método I de 200 a 2800 PC. Para subconjuntos de SNP de otros tamaños, se encontraron patrones similares. Por lo tanto, las correlaciones predictivas notificadas para la PCR I/II supervisada (que se muestran más adelante en la Figura 4) fueron sus valores de meseta (es decir, los máximos).

Escaso PLS

Como se señaló, g controla la escasez en el vector de peso (5), y un valor mayor de g da como resultado menos SNP seleccionados. Por otro lado, el número de SNP seleccionados aumenta con el número de componentes latentes K (Figura 6 en el Apéndice II).

Se llevó a cabo un CV quíntuple sobre los datos de entrenamiento (repetido tres veces para reducir la incertidumbre) para PLS disperso con diferentes valores de (g, K); g se varió entre 0,5, 0,55,…, 0,9 y K se varió entre 1, 3,

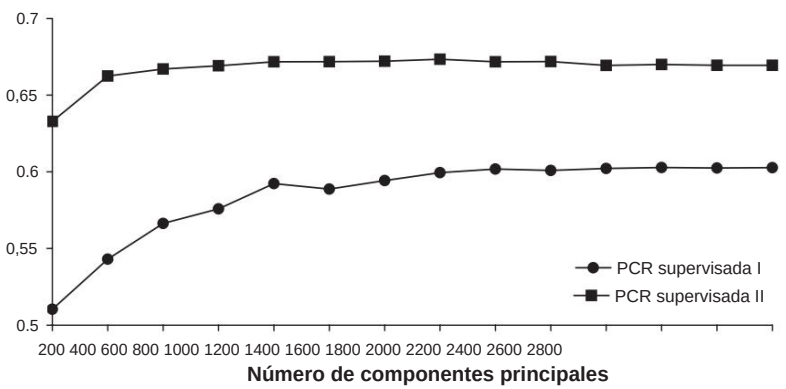

Figura 2 Correlación predictiva (correlación entre la capacidad de transmisión predicha (PTA) de la prueba de progenie y la PTA predicha en toros de prueba) contra el número de componentes principales para PCR supervisada I y II cuando el número de polimorfismos de nucleótido único seleccionados fue 3000 para ambos métodos.

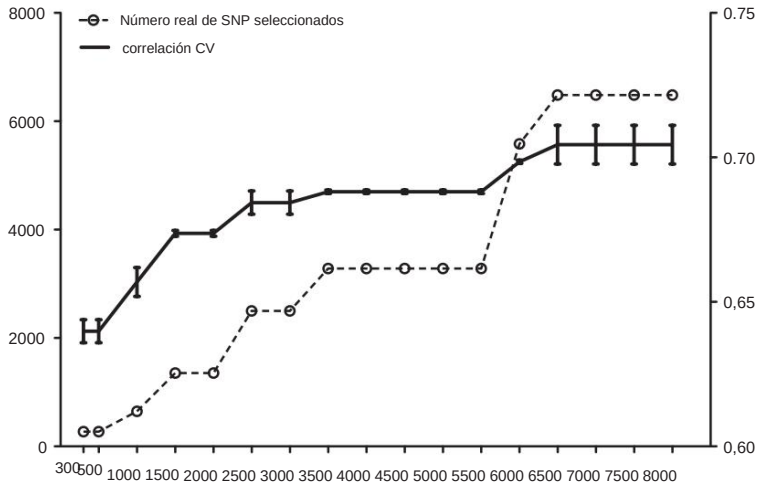

Figura 3 El número real de polimorfismos de un solo nucleótido (SNP) seleccionados y la correspondiente correlación de validación cruzada (CV) para cada límite superior en el número de SNP seleccionados en mínimos cuadrados parciales dispersos. La correlación mostrada (entre las habilidades de transmisión observadas y predichas) fue el promedio de tres réplicas de un CV quíntuple. Cada barra de error representa el error estándar medio. Los errores estándar de algunos puntos son demasiado pequeños para ser visibles.

5, 10, 15, 20, 25, 30, 35, 40, dando un total de 90 combinaciones. Además, el tamaño máximo del subconjunto se restringió a 8000 para ser coherente con el de la PCR supervisada. El criterio de valoración fue la correlación de CV, es decir, la correlación entre los valores de PTA observados y predichos de los datos de retención de 1/5 después de promediar los pliegues y las réplicas. Para cada límite superior del tamaño del subconjunto de SNP, el número real de SNP seleccionados se determinó mediante los mejores valores (g, K), que tenían la correlación de CV más alta. A medida que aumentaba el límite superior, el número real de SNP seleccionados variaba (Figura 3). Por ejemplo, cuando el límite superior estaba entre 3500 y 5500, el tamaño real del subconjunto de SNP mantuvo un valor de 3284, porque los subconjuntos de SNP más grandes no generaron un mejor rendimiento de CV. El subconjunto más grande contenía 6484 SNP, considerablemente más pequeño que el tamaño más grande permitido (8000). Esto también significó que solo siete subconjuntos de SNP (en lugar de 17 en la PCR supervisada) se evaluaron para PLS escasos en la predicción final sobre los datos de prueba.

Comparación de la capacidad predictiva entre métodos

La Figura 4 resume las correlaciones predictivas (en los toros de prueba 1398) utilizando PCR supervisado (I y II) y PLS escaso. La PCR II supervisada fue claramente mejor que la PCR I supervisada, y el PLS disperso se encontraba en el medio. A pesar de las notables diferencias entre los métodos cuando se evaluaron en pequeños subconjuntos de SNP, todos los métodos se acercaron a un nivel similar de correlación predictiva (0,65–0,68) cuando se incluyeron más SNP en el modelo. La PCR II supervisada se estabilizó en 3500 SNP, la PLS escasa en aproximadamente 5600 SNP y la PCR I supervisada en 6500 SNP. Cuando se utilizaron los 32 518 SNP, la correlación fue de 0,68 para PCR y de 0,67 para PLS.

La predicción utilizando un pequeño subconjunto de SNP mostró resultados prometedores con PCR II supervisada y PLS escaso. Por ejemplo, la PCR II supervisada logró correlaciones predictivas de 0,54 y 0,59 con 300 y 500 SNP, respectivamente. Estos correspondieron al 80 y 87% de la correlación obtenida con todos los 32 518

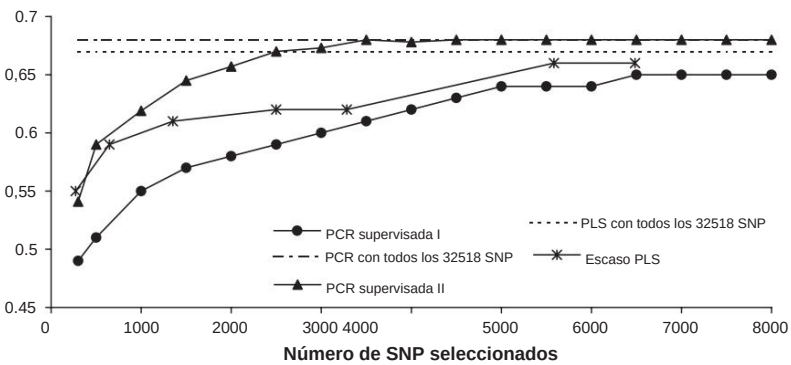

Figura 4 Correlaciones predictivas (correlación entre la capacidad de transmisión predicha (PTA) de las pruebas de progenie y la PTA predicha en los toros de prueba) para los dos métodos de regresión de componentes principales supervisados (PCR) (I y II) y mínimos cuadrados parciales dispersos (PLS), evaluado en diferentes números de SNP.

También se muestran los resultados de PCR y PLS ordinarios con todos los SNP ajustados.

SNP en un modelo de PCR, respectivamente. Del mismo modo, el PLS disperso produjo correlaciones de 0,55 con 272 SNP y de 0,59 con 648 SNP, lo que corresponde al 82 y al 88 % del ajuste de todos los SNP en un modelo PLS.
Sin embargo, la PCR I supervisada se comparó desfavorablemente con los otros dos métodos en la eficiencia de los subconjuntos de SNP seleccionados, especialmente para subconjuntos pequeños.

Discusión

Además de la característica de selección de variables manifestada en la Figura 4, la PCR supervisada y el PLS disperso también redujeron sustancialmente la dimensión del modelo. Como se indicó anteriormente, la cantidad de PC que se pueden instalar en PCR tiene un límite, que es menor entre el tamaño de la muestra y el tamaño del subconjunto SNP. Una estimación conservadora de nuestro análisis indicó que la ganancia en la correlación predictiva fue mínima cuando la cantidad de PC aumentó más allá del 60 % del límite. Para el PLS escaso, la reducción de la dimensión fue aún mayor. En los siete subconjuntos de SNP generados, los primeros cinco (ordenados por tamaño) requerían solo cinco componentes latentes y los dos últimos requerían diez componentes latentes.

Una diferencia importante entre los dos métodos de PCR supervisados fue que el método I evaluó los SNP de uno en uno, mientras que el método II consideró los SNP de forma conjunta al evaluar su importancia. Aquí, la estimación simultánea de todos los efectos de SNP se logró en dos pasos: regresión en las PC de SNP y transformación inversa de los coeficientes de regresión.

Ventajas de hacerlo, en comparación con la regresión en SNP individuales, fueron reducción de dimensión (de 32 518 SNP a 3000 PC) y ahorros computacionales. Se ha observado que los SNP seleccionados por análisis de SNP único pueden producir más falsos positivos que aquellos seleccionados por análisis de SNP múltiples, porque la señal en un SNP cuando se analiza individualmente a menudo se debilita por la inclusión de otros SNP correlacionados (Hoggart et al. 2008).). De acuerdo con esto, se encontró que los SNP mejor clasificados en el método I no presentaban rangos equivalentes (en realidad más bajos) cuando se calificaron en el método II. Esto se muestra en la Figura 5, donde los 300 SNP mejor clasificados seleccionados por el método I cubrieron casi todo el rango de rangos cuando se evaluaron por el método II, lo que indica la inconsistencia en la evaluación de SNP entre los dos métodos. Debido a que el análisis basado en múltiples SNP tiende a brindar estimaciones más confiables de los efectos de SNP, esto puede explicar por qué el método II fue consistentemente mejor que el método I cuando se evaluó en subconjuntos de SNP de varios tamaños.

En el modelado de PCR, la estructura de varianza de los efectos de PC aleatorios es importante para la precisión de la predicción. La PCR de mínimos cuadrados ordinarios refleja el supuesto de igualdad de varianza de los predictores y se demostró que es inferior a la PCR con varianzas de PC heterogéneas (Macciotta et al. 2009). Allí, los autores sugirieron usar kj r2 a=K como la varianza previa para PC j, (kj era el valor propio asociado con PC j, K era el número de PC y r2 la varianza genética aditiva total) ya que se basaba solo en datos y no requería suposiciones previas sobre las distribuciones de los efectos de PC. Con respecto a era

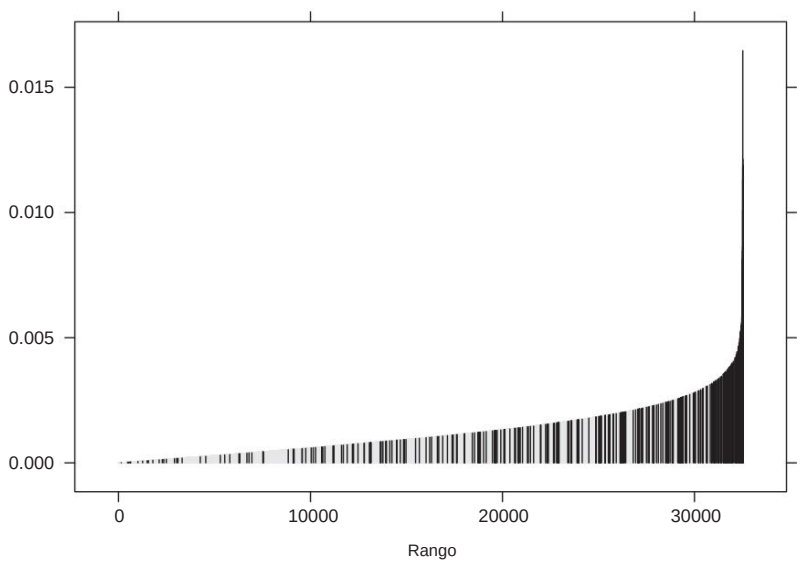

Figura 5 Valores absolutos de los efectos de los 32 518 SNP estimados por PCR II supervisada. Los 300 polimorfismos de un solo nucleótido (SNP) principales seleccionados por regresión de componentes principales supervisada (PCR) I están resaltados con barras sólidas, con sus alturas correspondientes a los tamaños de los efectos y sus posiciones en el eje x correspondientes a los rangos. Los 300 SNP no se agrupan en la región de alto rango, lo que indica discrepancia en la evaluación de SNP entre los dos métodos (I y II).

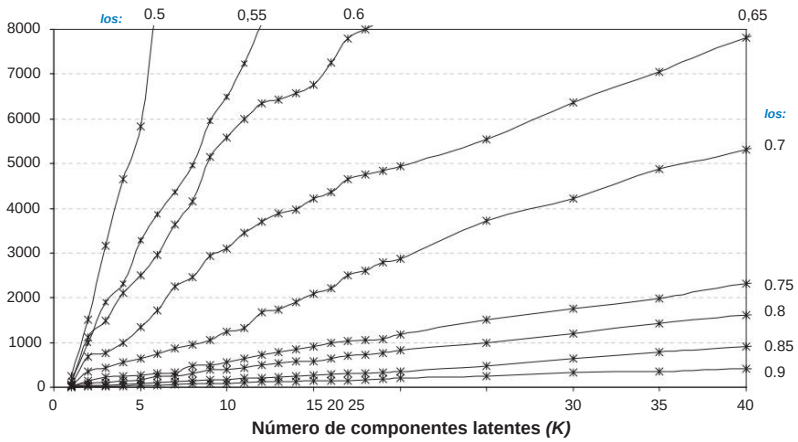

Figura 6 Dos parámetros de sintonización (g, K) en los mínimos cuadrados parciales dispersos controlan la número de polimorfismos de un solo nucleótido seleccionado. g está entre 0 y 1; k es el número de componentes latentes a ser extraído.

esto, una preocupación es que los valores propios contienen información sobre la estructura de correlación SNP solamente, y sus magnitudes no necesariamente reflejan la asociación entre los genotipos SNP y el fenotipo en cuestión. En nuestro estudio, el Las regresiones en las PC se trataron como aleatorias comunes. efectos con la contracción diferencial impuesta por priores exponenciales dobles, que es una forma más general configuración de la estructura de varianza.

El mecanismo de selección de SNP para PLS disperso es diferente al de la PCR supervisada. el anterior está controlado por dos parámetros de sintonización (g y K) cuyos valores determinan el número de seleccionados SNPs, mientras que estos últimos pueden producir cualquier número de SNP ajustando el valor de corte de los efectos SNP. Aunque en PLS escaso el número de SNP seleccionados no puede ser controlado directamente por el usuario, uno puede intentar diferentes combinaciones de (g, k) para encontrar una que produce el subconjunto de SNP cuyo tamaño es cercano al valor deseado o por debajo de algunos predefinidos valor. En este estudio, diferentes límites superiores en lugar que los valores exactos del tamaño del subconjunto se utilizaron para Escaso PLS. Para cada límite superior, el mejor (g, k) valor entre todos los candidatos calificados fue elegido por CV. Por lo tanto, es necesario evaluar una cantidad suficiente número de valores (g, k), aunque eso sería computacionalmente intensiva. Aquí, g osciló entre 0,5 a 0,9 con un tamaño de paso de 0,05; K varió de 1 a 40 en un tamaño de paso de 2 o 5. Es posible que, al intentar más valores para g y K (por ejemplo, reduciendo el paso tamaño), el rendimiento predictivo de PLS disperso puede mejorarse en comparación con el alcanzado en el presente estudio.

Para la selección genómica, el objetivo principal de seleccionar un subconjunto de marcadores del genoma completo es reducir costo de genotipado manteniendo la capacidad predictiva obtenidos con los marcadores seleccionados a un precio razonable nivel. Por lo tanto, es natural basar la elección en asociaciones entre los marcadores y el rasgo de interés, y el conjunto de marcadores seleccionado será específico del rasgo. Nuestros estudios presentados aquí siguieron este lógica. En una línea alternativa de investigación (Weigel et al. 2010a,b), uno comienza eligiendo un pequeño número de marcadores igualmente espaciados a lo largo del genoma y luego imputar genotipos para el resto de los marcadores utilizando una población de referencia que contiene genotipos en todos los marcadores. uno puede esperar que la capacidad predictiva de estos marcadores se compararía desfavorablemente con la de los marcadores seleccionados basado en los efectos sobre el rasgo, especialmente cuando el el número de marcadores es pequeño, por ejemplo, <750 en Weigel et al. (2010b).

En cuanto a la elección de la variable de respuesta en un modelo de predicción genómica, varios estudios de simulación (eg Pimentel et al. 2009;Macciotta et al. 2010) have demostró que la precisión (correlación entre BV verdadero y BV predicho) de BV genómicos predichos del uso de fenotipos crudos como variable de respuesta es incluso más alto que el del uso de BLUP-estimado BV (EBV). Esto puede ser atribuible a la baja confiabilidad de EBV o al suavizado excesivo por el infinitesimal modelo. A diferencia de un entorno de simulación, la verdadera precisión (a diferencia de la precisión basada en el modelo) del Los PTA genómicos pronosticados en un estudio de datos reales no pueden evaluarse, porque se desconocen los BV verdaderos. Sin embargo, no esperamos una superioridad obvia de usar fenotipos crudos en lugar de usar PTA, porque la fiabilidad de los PTA de toros en nuestro los datos fueron bastante altos (84% de los toros tenían una confiabilidad de PTA superior a 0.8). Por otro lado, una mejor la opción es usar versiones ponderadas de PTA incorporando información de confiabilidad (Garrick et al. 2009), debido a que los datos comprendían toros con diferentes confiabilidades

Conclusiones

En comparación con la regresión en una gran cantidad de SNP, la dimensión del modelo en la PCR supervisada y el PLS disperso se redujo considerablemente al usar una cantidad menor de PC o componentes latentes como predictores. La PCR II supervisada fue claramente mejor que la PCR I supervisada en capacidad predictiva cuando se evaluó en subconjuntos de SNP de varios tamaños, y el PLS escaso estaba en el medio. El aumento del número de SNP mejoró la capacidad de predicción para todos los métodos, y la PCR II supervisada mostró el aumento más rápido a medida que se incluían más SNP. Cuando el tamaño del subconjunto de SNP estaba por debajo de 1000, el PLS disperso presentó un rendimiento predictivo similar al de la PCR II supervisada. Sin embargo, con más SNP utilizados, la capacidad de PLS disperso para aumentar la correlación de predicción fue inferior a la de la PCR II supervisada. En conjunto, este estudio demostró el potencial de combinar la reducción de dimensiones y la selección de variables para una predicción precisa y rentable de la VB genómica.

Agradecimientos

Se agradece el apoyo de la Estación Experimental Agrícola de Wisconsin y de las subvenciones NRICGP/USDA 2003-35205-12833, NSF DEB-0089742 y NSF DMS-044371. Se agradece a AI Va´zquez por la limpieza y edición de datos.

Referencias

Bair E., Hastie T., Paul D., Tibshirani R. (2006) Predicción por componentes principales supervisados. Mermelada. Estadística Asociación, 101, 119–137.

Boulesteix A.-L., Strimmer K. (2007) Mínimos cuadrados parciales: una herramienta versátil para el análisis de datos genómicos de alta dimensión. Informes Bioinformática, 8, 32–44.

Cadima J., Jolliffe IT (1995) Cargas y correlaciones en la interpretación de componentes principales. Aplicación J. Estado, 22, 203–214.

de los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel KA, Cotes J. (2009) Predicción de rasgos cuantitativos con modelos de regresión para marcadores moleculares densos y pedigríes. Genética, 182, 375–385.

Chen X., Wang L., Smith JD, Zhang B. (2008). súper visó el análisis de componentes principales para el enriquecimiento de conjuntos de genes de datos de micromatrices con resultados continuos o de supervivencia. Bioinformática, 24, 2474–2481.

Chun H., Keles¸ S. (2009) Mapeo de loci de rasgos cuantitativos de expresión con regresión de mínimos cuadrados parciales. Genética, 182, 79–90.

Chun H. y Keles¸ S. (2010) Regresión de mínimos cuadrados parciales dispersos para la reducción simultánea de dimensiones y la selección de variables. Estado JR. Soc. B, 72, 3–25.

Frank IE, Friedman JH (1993) Una visión estadística de algunas herramientas de regresión quimiométrica. Tecnometría, 35, 109–135.

Garrick DJ, Taylor JF, Fernando RL (2009) Regresión ing valores de cría estimados e información ponderada de análisis de regresión genómica. Gineta. sel. Evol., 41, 1.

Habier D., Fernando RL, Dekkers JCM (2009) Geno selección de micrófono usando paneles de marcadores de baja densidad. Genética, 182, 343–353.

Helland, ES (1988). Sobre la estructura de la regresión de mínimos cuadrados parciales. común Estado, 17, 581–607.

Hoggart CJ, Whittaker JC, Iorio MD, Calvo DJ (2008) Análisis simultáneo de todos los SNP en estudios de asociación de genoma completo y resecuenciación. PLoS Genet., 4, e1000130.

Jolliffe IT, Trendafilov NT, Uddin M. (2003) A modi Técnica de componentes principales fied basada en el lazo. J. Cómputo. Grafico. Estadística 12, 531–547.

de Jong S. (1993). SIMPLS: un enfoque alternativo a la regresión de mínimos cuadrados parciales. Quimio. Intel. Laboratorio. Syst., 18, 251–263.

Long N., Gianola D., Rosa GJM, Weigel KA, Avendaño S. (2007) Procedimiento de clasificación de aprendizaje automático para seleccionar SNP en la selección genómica: aplicación a la mortalidad temprana en pollos de engorde. J. Anim. Criar. Genet., 124, 377–389.

Macciotta N., Gaspa G., Steri R., Pieramati C., Carnier P., Dimauro C. (2009) Preselección de los SNP más significativos para la estimación de valores genéticos genéticos. BMC Proc., 3 (suplemento 1), S14.

Macciotta N., Gaspa G., Steri R., Nicolazzi E., Dimauro C., Pieramati C., Cappio-Borlino A. (2010) Uso de valores propios como priores de varianza en la predicción de valores genéticos genómicos mediante análisis de componentes principales. j Dairy Sci., 93, 2765–2774.

Manne R. (1987) Análisis de dos mínimos cuadrados parciales algoritmos para la calibración multivariada. Quimio. Laboratorio inteligente. Syst., 2, 187–197.

Massy WF (1965) Regresión de componentes principales en la investigación estadística exploratoria. Mermelada. Estadística Asociado, 60, 234–256.

McIntosh AR, Bookstein FL, Haxby JV, Grady CL (1996) Análisis de patrones espaciales de imágenes cerebrales funcionales utilizando mínimos cuadrados parciales. Neuroimagen, 3, 143–157.

Meuwissen THE, Hayes BJ, Goddard ME (2001) Pre dicción del valor genético total utilizando mapas de marcadores densos de todo el genoma. Genética, 157, 1819–1829.

Parque T., Casella G. (2008). El lazo bayesiano. Mermelada. Estadística Asociación, 103, 681–686.

Pimentel ECG, Konig S., Schenkel FS, Simianer H. (2009) Comparación de procedimientos estadísticos para estimación

efectos poligénicos usando un marcador de genoma denso datos. BMC Proc., 3 (suplemento 1), S12.

Solberg T., Sonesson A., Woolliams J., Meuwissen T. (2009) Reducción de la dimensionalidad para la predicción de los valores genéticos de todo el genoma. Gineta. sel. Evol., 41, 29.

Usai MG, Goddard ME, Hayes BJ (2009) LASSO con validación cruzada para la selección genómica. Gineta. Res., 91, 427–436.

Va´zquez A.I., Rosa G.J.M., Weigel K.A., de los Campos G., Gianola D., Allison DB (2010) Capacidad predictiva de subconjuntos de polimorfismos de un solo nucleótido con y sin promedio de padres en Holsteins de EE.UU. J. Ciencias de la leche, 93, 5942–5949.

Weigel K.A., de los Campos G., Gonza´lez-Recio O., Naya H., Wu XL, Long N., Rosa GJM, Gianola, D. (2009) Capacidad predictiva de valores genómicos directos para toda la vida mérito neto de toros Holstein utilizando subconjuntos seleccionados de marcadores de polimorfismo de un solo nucleótido. J. Dairy Sci., 92, 5248–5257.

Weigel K., de los Campos G., Va´zquez A., Rosa G., Gian ola D., Tassell CV (2010a) Precisión de la genómica directa valores derivados de genotipos de polimorfismo de nucleótido único imputados en ganado Jersey. J. Dairy Sci., 93, 5423–5435.

Weigel K., Tassell CV, O'Connell J., VanRaden P., Wig gans G. (2010b) Predicción de genotipos de polimorfismo de un solo nucleótido no observados del ganado Jersey usando paneles de referencia e imputación basada en la población algoritmos J. Dairy Sci., 93, 2229–2238.

Wold H. (1985) Mínimos cuadrados parciales. En: S. Kotz, Países Bajos Johnson (eds), Enciclopedia de Ciencias Estadísticas, Volumen 6. Wiley, Nueva York, NY, págs. 581–591.

Apéndices

I. Implementación de PLS disperso con univariado respuesta.

Dados los valores de g y K (número de componentes latentes), el PLS disperso se puede utilizar para la selección de variables como sigue (Chun & Keles¸ 2010).

Defina A para que sea un conjunto de índices para variables seleccionadas y XA sea una submatriz de X(n · p) que contenga variables indexadas en A.

1. conjunto $\hat{b}PLS = 0$, A = fg, k = 1, y1 = y.
2. Mientras (k £ K)
   - 2.1. Encuentre ^c [dado por (5)], usando y1 como respuesta variable.
   - 2.2. Actualizar A como fi : ^ci 6¼ 0g[fi : b^PLS 6¼ 0g; yo ¼ 1; ...; pags.
   - 2.3. Ajuste PLS en XA usando k componentes latentes T ¼ {t1,…,tk}; actualice ^bPLS a través de la regresión de mínimos cuadrados de y1 en T ('Resumen de PCR y PLS').
   - 2.4. Actualizar b1 ¼ y X^bPLS.
   - 2.5. Actualice k a k + 1.

II. Número de SNP seleccionados controlados por los dos parámetros de ajuste en PLS disperso. Consulte la Figura 6.