USO DE DIFERENTES MATRICES DE RELACIONES GENÉTICAS EN ESTUDIOS DE ASOCIACIÓN PARA CARACTERES LECHEROS EN CABRAS DE RAZA FLORIDA

Ramón M.[1], Molina A.[2], Carabaño MJ.[3], Sánchez-Rodríguez M.[2], Serradilla, JM.[2], Díaz C.[3]
[1]CERSYRA-IRIAF; [2]Dpto. Genética (Universidad de Córdoba); [3]Dpto. Mejora Genética (INIA)
mramon@jccm.es

Introducción

Los estudios de asociación a genoma completo (GWAS) presentan varias limitaciones, por un lado asociadas a la existencia de estructuras subyacentes dentro de la población que pueden generar un sesgo en la detección de variantes genéticas y, por otro relacionadas con la cobertura del genoma que dan los chips comerciales así como la magnitud del desequilibrio de ligamiento entre los marcadores y los genes causales. Estas limitantes se puede compensar, en parte, mediante el uso de modelos mixtos con los que se ajusten de forma simultánea un efecto poligénico para cada animal además de los efectos de los marcadores (Calus et al. 2008). En las poblaciones domésticas, la estratificación que resulta de un uso recurrente y desigual de reproductores puede generar una asociación espuria entre marcadores y/o variantes de interés en la forma de efectos confundidos. En este caso, la inclusión del efecto poligénico permite reconocer la existencia de covarianzas entre los individuos, tanto a partir de la información del pedigrí como de los propios marcadores (matriz de parentesco molecular), de manera que además de tomar en cuenta la varianza no explicada por el marcador nos permita corregir la estructura de la población. Así, el procedimiento consiste entonces en construir la matriz de relaciones genéticas (GRM) y realizar el estudio de asociación de los marcadores de forma simultanea (Yang et al. 2014). Un punto crítico es la elección de qué GRM utilizar. Si incluimos en nuestra GRM la información de todos los marcadores moleculares se produce entonces un "doble ajuste" del marcador bajo análisis, ya que una parte importante de la variabilidad del carácter podría ser absorbida por el efecto poligénico causando una pérdida de poder de detección (Listgarten et al. 2012). Este efecto será mayor cuanto mayor sea la magnitud del efecto. El objetivo de este trabajo es evaluar el uso de diferentes matrices de relaciones genéticas en el análisis GWAS de dos caracteres que presentan una arquitectura genética distinta, como son la producción de leche (de carácter marcadamente poligénico) y el contenido en proteína (gobernado por algunos genes de efecto mayor) en cabras de la raza Florida.

Materiales y Métodos

Un total de 625 animales fueron genotipados con el Beadchip 50K para caprino de Illumina Los caracteres considerados fueron la cantidad diaria de leche producida (L; kg/d) y el contenido de proteína de la misma (PP; %/d). Para ambos caracteres, el estudio GWAS se llevó a cabo con los valores genéticos estimados por el modelo de rutina de esta raza una vez de-regresados (Garrick et al. 2009). Como paso previo, se eliminaron los individuos con un *Call Rate* < 0.98 y los marcadores con un "*Call Rate"* < 0.98, una frecuencia para el alelo menos frecuente (MAF) <0.02, y/o una desviación significativa ($p<10^{-6}$) respecto al equilibrio de Hardy-Weinberg (HWE). Finalmente, para los análisis GWAS se utilizaron un total de 623 animales y 48385 SNP's. Los modelos que se estudiaron fueron : (1) Modelo de regresión clásico para cada marcador, sin efecto poligénico (REG); (2) Modelo mixto incluyendo un efecto poligénico con una GRM construida con todos los marcadores moleculares (LMM_{all}); (3) Idem a (2) pero en GRM se excluyen aquellos marcadores con un p-valor nominal $<10^{-5}$ (LMM_{notM}); (4) Similar al anterior y que además excluye aquellos marcadores con un LD>0.3 respecto a aquellos marcadores excluidos en 3 (LMM_{LD}); (5) Modelo mixto incluyendo un efecto poligénico y como GRM la matriz de relaciones aditivas clásica (LMM_{ped}); y (6) Análisis de regresión que usa el residuo de un modelo mixto que incluye las matrices de marcadores moleculares usadas en los puntos 2-5 (GRAMMAR; Aulchenko et al. 2007). Los análisis fueron realizados con el programa GenABEL (Aulchenko et al. 2007). Para cada modelo se calculó además el *Genomic Inflation Factor* (λ_{median}). Valores próximos a 1 de este estadístico se consideran óptimos

Resultados y Discusión

Los resultados obtenidos de los estudios de asociación fueron diferentes en función del modelo usado. El análisis de regresión clásico (REG) fue el que mayor número de marcadores por debajo de un umbral dado (para un *p-valor* dado) y de mayor magnitud identificó (Figura 1). Si embargo, es claro que existió una sobreestimación de los efectos de los marcadores y por lo tanto un número alto de falsos positivos como refleja los altos valores de λ_{median} , de 3 o superiores, obtenidos para este modelo (Tabla 1). El modelo de regresión clásico no permite descomponer la varianza atribuida al marcador del resto del genoma y además no tiene en cuenta la posible existencia de una estratificación en la población por lo que puede existir un sesgo en la estima del efecto del marcador. Los *p-valores* correspondientes a valores de FDR en el rango del 1 al 10 % fueron más altos, que para el resto de modelos, así como el número de marcadores que detectó como significativos. En el lado opuesto tenemos los modelos en 2 pasos (GRAMMAR), los cuales usan como variable para el GWAS el residuo de un modelo mixto que contiene un efecto poligénico con una estructura de (co)varianzas reconocidas por la GRM. Este tipo de modelos que persiguen corregir la estratificación de la población, son muy conservadores en el sentido que reducen el número de falsos positivos pero a costa de incrementar los falsos negativos provocando una subestimación de los efectos de los marcadores y una pérdida de poder de detección. Los GRAMMAR, con la excepción del $GRAMMAR_{ped}$, presentaron valores de λ_{median} bajos, en torno a 0.6 y los *p-valores* umbrales para FDR en el rango de 1 al 10% fueron altos y similares a los obtenidos en el modelo REG, aunque en el caso de los modelos GRAMMAR el número de marcadores que se identificaron por debajo de esos *p-valores* umbrales fue muy inferior y el efecto estimado mucho más bajo (Tabla 1 y Figura 1). Como ya se avanzó en la introducción, la menor variabilidad del residuo resultante del modelo mixto va a depender de la matriz de relaciones genéticas (GRM) que usemos en dicho modelo mixto. Excepto para el análisis en el que se usó la información del pedigrí para construir la matriz GRM, para el resto de análisis prácticamente no se encontraron marcadores con efecto significativo. Un aspecto importante a tener en cuenta es que los estudios de asociación llevados a cabo en este trabajo usaron como variable dependiente el valor genético de-regresado, por lo que es esperable que una parte muy importante de la varianza de esa variable sea absorbida por el efecto poligénico y que esa "absorción" sea mayor cuando usamos la matriz de de marcadores moleculares completa.

Como solución a las limitaciones de los modelos anteriores se ha propuesto el uso de modelos mixtos, los cuales a partir de la matriz GRM consideran de forma simultánea la estructura de población y el efecto poligénico. En este estudio, los valores de λ_{median} se han situado muy próximos al valor de 1 (Tabla 1), indicativo de que se ha tenido en cuenta la estructura poblacional de forma apropiada sin que exista un exceso (modelo REG) o reducción (modelos GRAMMAR) de la varianza asociada a los marcadores. Al ajustar un efecto poligénico teniendo en cuenta la estructura de (co)varianzas entre los mismos mediante el uso de una matriz de GRM tiene como objetivo remover la variabilidad genética del carácter no asociada al marcador y corregir el ruido que una posible estratificación de la población puede generar. En este proceso los LMM reducen el numero de falsos positivos, sin embargo en función de la magnitud de los efectos el poder de detección de estos modelos va a estar influenciado por la matriz GRM que utilicemos (Listgarten et al. 2012, Yang el al., 2013). Para el carácter leche no se encontraron marcadores que permitieran mantener una tasa de falsos positivos por debajo de un umbral prefijado, excepto para REG y $LMM_{PED.}$ Para el carácter proteína los modelos LMM identificaron un mayor número de marcadores asociados al carácter para un nivel de FDR dado, siendo LMM_{notM} y LMM_{LD} los que identificaron un mayor número de marcadores. Entre estos modelos no se observaron diferencias importantes, pudiendo deberse en parte al bajo LD observado en esta raza para este carácter y muy circunscrito a regiones próximas del genoma. La tercera alternativa utilizó la información del pedigrí en la matriz GRM. Si bien este modelo identificó un mayor número de marcadores asociados al carácter, los valores por encima de 1 de λ_{median} obtenidos parecen indicarnos que la estructura poblacional no habría sido considerada correctamente. Esto podría ser consecuencia de que hay un 33% de animales de los que se desconoce su pedigrí y que según la información molecular podrían estar emparentados.

Bibliografía

Aulchenko YS, Ripke S, Isaacs A, et al. 2007. Bioinformatics. 23(10):1294-6 · Calus MPL, Meuwissen THE, de Roos APW et al. 2008. Genetics. 178:553–561 · Listgarten J, Lippert C, Kadie CM, et al. 2012. Nature Methods 9, 525-526 · Yang J, Zaitlen NA, Goddard ME, et al. 2014. Nat. Genet. 46:100-106.

Tabla 1. P-valores umbrales asociados a diferentes niveles de tasa de falsos positivos (FDR), número de marcadores declarados significativos para ese umbral (entre paréntesis), y *Genomic Inflation Factor* (λ_{median}) para los modelos de GWAS incluidos en este trabajo. Ver Materiales y Métodos para una descripción de cada modelo.

	Leche (kg/d)				Proteína (%)			
	λ_{med}	FDR			λ_{med}	FDR		
		0.01	**0.05**	**0.10**		**0.01**	**0.05**	**0.10**
REG	3.24	(2132) $4.4\cdot10^{-4}$	(5810) 0.006	(8754) 0.018	2.97	(1269) $2.6\cdot10^{-4}$	(4442) 0.004	(7205) 0.015
LMM_{all}	0.99	.	.	.	1.00	(3) $6.2\cdot10^{-7}$	(4) $4.1\cdot10^{-6}$	(4) $8.1\cdot10^{-6}$
LMM_{notM}	1.03	.	.	.	1.03	(3) $6.2\cdot10^{-7}$	(5) $5.2\cdot10^{-6}$	(23) $4.8\cdot10^{-4}$
LMM_{LD}	1.03	.	.	.	1.01	(3) $6.2\cdot10^{-7}$	(5) $5.2\cdot10^{-6}$	(23) $4.8\cdot10^{-5}$
LMM_{ped}	1.75	(3) $6.2\cdot10^{-7}$	(212) $2.2\cdot10^{-4}$	(718) 0.001	1.77	(48) $9.9\cdot10^{-6}$	(224) $2.3\cdot10^{-4}$	(628) 0.001
$GRAMMAR_{all}$	0.61	.	.	.	0.56	.	.	.
$GRAMMAR_{notM}$	0.63	.	.	.	0.58	.	.	.
$GRAMMAR_{LD}$	0.63	.	.	.	0.57	.	.	.
$GRAMMAR_{ped}$	1.38	.	(2) $3.4\cdot10^{-6}$	(12) $2.5\cdot10^{-5}$	1.39	(8) $1.7\cdot10^{-6}$	(30) $3.1\cdot10^{-5}$	(57) $1.2\cdot10^{-4}$

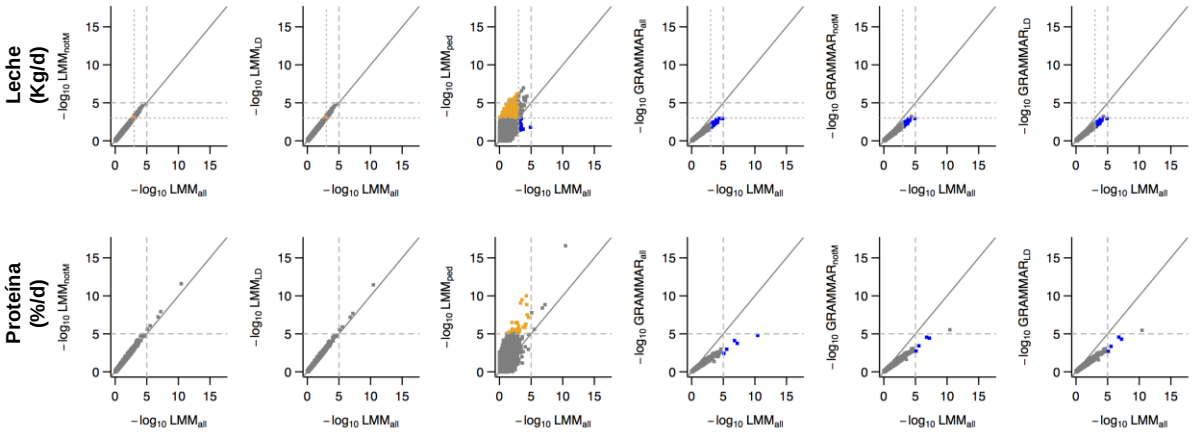

Figura 1. Comparación de los p-valores obtenidos del modelo mixto que incluyó la matriz de marcadores moleculares completa (LMM_{all}; eje de las *x*) con respecto al resto de modelos (excepto Reg; eje de las *y*), para los caracteres cantidad de leche (fila superior) y porcentaje de proteína (fila inferior). Los puntos en naranja representan aquellos marcadores con un *–log₁₀(p-valor)* menor al umbral fijado (10^{-3} para leche y 10^{-5} para proteína) para el modelo LMM_{all} y mayor para los otros modelos. Para los puntos en azul la interpretación es la opuesta *(–log₁₀(p-valor)* mayor al umbral fijado para el modelo LMM_{all} y menor para los otros modelos). Ver Materiales y Métodos para una descripción de cada modelo

THE USE OF DIFFERENT GENOMIC RELATIONSHIP MATRICES ON GWAS STUDIES FOR MILKS TRAITS IN FLORIDA GOATS

The choice of a method to be used for conducting genetic association studies (GWAS) is a key step. In this study, as reported in other studies, the use of classical regression models led to an inflation of the estimated effects of the markers, since they do not take into account the population structure. The two-step approach proposed to account for that structure led to the opposite effect, that is, a subestimation of the effects of markers. Thus, Linear Mixed Models (LMM) have resulted to be the best choice for conducting association studies given that these models consider the populations structure and the information of molecular markers at the same time. For these models, the source of information used to build the genomic relationship matrix has resulted in differences in terms of power of detection and the magnitude of the association. The models that excluded the candidate markers and those in LD with them performed better than the model that included the information from all the markers.