FACTORES DE BAYES CON GBLUP PARA ESTUDIOS DE ASOCIACIÓN DE GENOMA COMPLETO.

Varona L.*[1] y Legarra A.[2]
[1]Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza
[2] GenPhySE, Institut National de la Recherche Agronomique
*lvarona@unizar.es

INTRODUCCIÓN

El desarrollo de los métodos de genotipado masivo ha permitido la utilización de la información genómica en estudios de asociación a lo largo del genoma (Bush y Moore, 2011). La hipótesis básica de estos estudios parte del supuesto de que se disponen de marcadores SNP que cubren el genoma con una gran densidad y que, si existen genes cuya variación está asociada a caracteres cuantitativos, se dispondrá de alguno de estos marcadores SNP en desequilibrio de ligamiento (DL) con ellos y podrán ser detectados a partir de un test estadístico

El procedimiento más habitual es el denominado regresión de marcador simple (RMS), que realiza un test estadístico para cada marcador SNP, que exige considerar posteriormente procedimientos de corrección para contrastes múltiples. Otra alternativa es utilizar los procedimientos desarrollados para la implementación de la selección genómica (Meuwissen et al., 2001), que implican procedimiento de regularización de los efectos asociados a los SNP (Gianola et al., 2009). Esta aproximación evita la corrección de test múltiples y considera simultáneamente los efectos de todo el genoma. De hecho, en estudios de simulación (López de Maturana et al., 2013) se ha probado que presenta mayor potencia estadística.

Pese a todo, este procedimiento también presenta inconvenientes, ya que también los marcadores se encuentran entre si en desequilibrio de ligamiento, y por lo tanto, generan problemas de co-linealidad. Por este motivo, parece razonable disponer de un test de hipótesis asociado a regiones del genoma donde se localicen varios SNP. En este trabajo se ha generalizado un procedimiento de test de hipótesis basado en Factores de Bayes (García-Cortés et al., 2001; Varona et al., 2001) a una versión multivariante que permita realizar test conjuntos a un conjunto de SNP.

SIMULACIÓN

Se ha simulado una población de tamaño efectivo 100 a lo largo de 1,000 generaciones. Posteriormente, esta población se expandió a 5,000 individuos. Se asumió un genoma de 10 cromosomas de 100 cM cada uno y con 1000 SNP localizados al azar. Además, se simuló un QTL en cada cromosoma (Figura 1). La tasa de mutación fue 2.5 x 10^{-3}, tanto para marcadores SNP como para QTL. Finalmente, se simuló un carácter con heredabilidad 0.25 para los 5,000 individuos.

Tabla 1. Descripción de los QTL simulados

	Cromosoma									
	1	2	3	4	5	6	7	8	9	10
Pos.	50.76	15.61	97.63	80.09	55.34	12.18	64.44	4.72	25.12	50.61
Freq.	0.569	0.123	0.722	0.403	0.689	0.587	0.029	0.468	0.493	0.759
a	20.19	21.56	12.88	10.19	9.65	8.29	22.37	7.08	6.67	7.40
Va	200	100	66.67	50.00	40.00	33.33	28.57	25.00	22.22	20.00
%Va	34.1	17.1	11.4	8.5	6.8	5.7	4.9	4.3	3.8	3.4

Pos. : Posición en cM, Freq. : Frecuencia alélica, a: Efecto aditivo, Va: Varianza aditiva, %Va: Porcentaje de la Varianza Aditiva.

MÉTODOS

El modelo de análisis fue el siguiente modelo lineal:

$$\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{Xa} + \boldsymbol{e}$$

donde se asumió una distribución a priori normal para los efectos asociados a los marcadores (**a**). Es decir, el modelo es de tipo SNP-BLUP o GBLUP (VanRaden, 2008). El análisis se realizó mediante muestreo de Gibbs a través de una única secuencia de 10,000 iteraciones después de descartar las 2,000 primeras, pero se puede realizar usando las ecuaciones del modelo mixto.

En primer lugar, se calculó el factor de Bayes para cada SNP por separado a partir de la siguiente expresión:

$$BF = \frac{N(0|0, \sigma^2_{a0})}{N\left(0|\hat{a}_i, Var(\hat{a}_i)\right)}$$

Donde σ^2_{a0} es la varianza a priori de los efectos aditivos asociados a los marcadores SNP, $\hat{a}_i$ es la media posterior de efecto asociado al *iesimo* SNP y $Var(\hat{a}_i)$ es su varianza posterior obtenida mediante el procedimiento de McMC.

Posteriormente, se definió el Factor de Bayes para segmentos de *k* marcadores como:

$$BF = \frac{MVN(\mathbf{0}_k|\mathbf{0}_k, \boldsymbol{I}_k\sigma^2_{a0})}{MVN\left(\mathbf{0}_k|\widehat{\boldsymbol{a}}_{[i,i+k]}, \boldsymbol{T}_k\right)}$$

Donde $\widehat{\boldsymbol{a}}_{[i,i+k]}$ es el vector de medias posteriores para *k* efectos a partir del *iesimo* marcador**,** $\mathbf{0}_k$ es un vector de zeros e $\mathbf{I}_k$ es una matriz identidad de tamaño *k* y $\mathbf{T}_k$ es la matriz de varianzas y covarianzas posteriores para los efectos de los *k* marcadores obtenidas del McMC. Dicha matriz considera la cantidad de datos y la colinealidad de los marcadores. En el análisis, se asumió como σ^2_{a0} el estimador obtenido a partir del propio análisis mediante McMC.

RESULTADOS

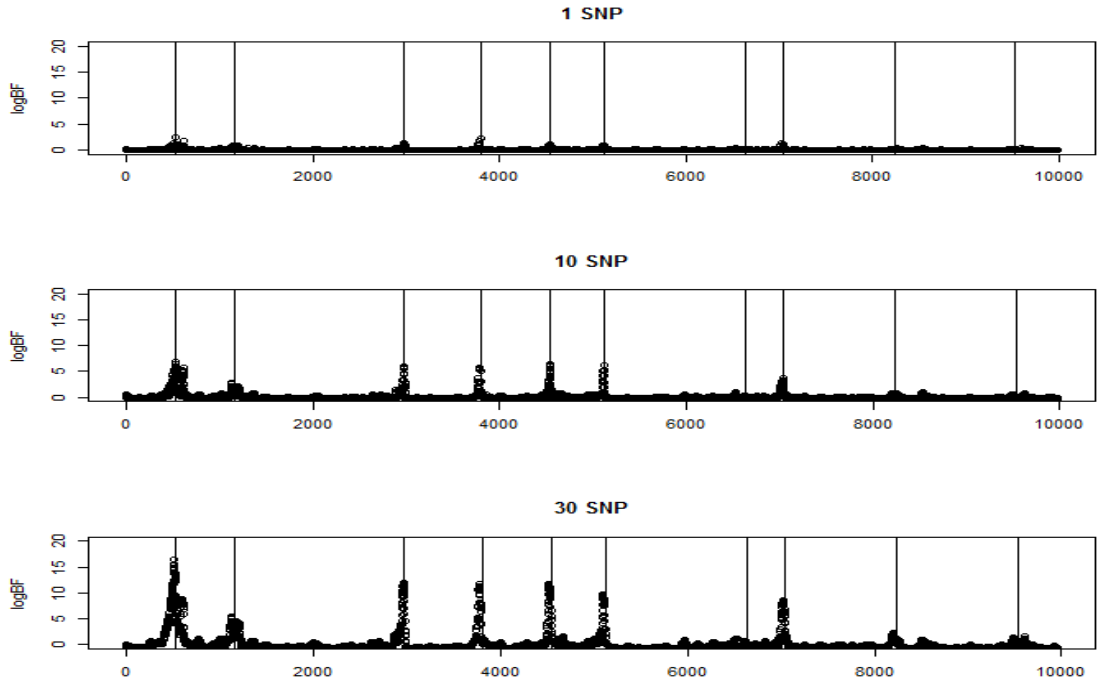
En la Figura 1 se presentan los resultados del logaritmo del Factor de Bayes para los análisis realizados para regiones del genoma definidas por 1, 10 y 30 marcadores. Se observa como la evidencia en favor del modelo que postula la presencia de un gen asociado al carácter se incrementa de manera muy considerable a media que aumenta el número de marcadores que se localizan en la región genómica. Si se retienen los BF superiores a 150 como “muy relevantes” (Legarra et al. 2014) , con 30 marcadores se encuentran 7 de los 10 QTLs, mientras que con 10 o 1 marcadores se encuentran 5 y ningumo respectivamente.

La razón de estas diferencias entre procedimientos es que los marcadores SNP no se encuentran en DL completo con las mutaciones causales. Además, los SNP también se encuentran en DL entre ellos y, por este motivo, la varianza explicada por el gen se divide entre todos aquellos SNP que presenten algún grado de DL con él. En el método de regresión simple, se observaría que varios marcadores contiguos muestran evidencia del gen (*e.g.* Rupp et al., 2015). Esta colinealidad no impide una buena capacidad predictiva de los valores mejorantes (precisión de 0.88).

El procedimiento se ha puesto a punto con la aproximación de GBLUP, que además permitiría un cálculo sencillo y exacto del test mediante las ecuaciones del modelo mixto (sin utilizar métodos de MonteCarlo), pero la estrategia puede ser extendida con sencillez para otro tipo de regularizaciones como Bayes A (Meuwissen et al., 2001),

Bayes B (Meuwissen et al., 2001) o Bayesian Lasso (De los Campos et al., 2009), entre otras. Otra ventaja del procedimiento es que no necesita variables indicadoras de "pertenencia" al modelo, que necesitan una implementación complicada y distribuciones a priori difíciles de concebir. El *a priori* en el caso del GBLUP es extremadamente sencillo y ampliamente utilizado en evaluación genómica (VanRaden, 2008).

Figura 1. Logaritmo de los Factores de Bayes a lo largo del genoma con segmentos de 1, 10 y 30 SNP.



CONCLUSIÓN

El Factor de Bayes en la forma presentada permite realizar estudios de asociación multimarcadores, con estadísticos y umbrales exactos, robustos al desequilibrio de ligamiento entre marcadores y de manera eficiente computacionalmente.

REFERENCIAS

Meuwissen T. H. E., Hayes B. J., Goddard M. E., 2001 157: 1819–1829. ● Garcia-Cortes L.A., Cabrillo C. Moreno C., Varona L. 2001 Genet. Sel. Evol. 33: 3–16 ● Varona L, Garcia-Cortes L.A., Perez-Enciso M. 2001 Genet. Sel. Evol. 33: 133–152 ● VanRaden P. 2008 J. Dairy Sci. 91:4414–4423 ● de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra et al., 2009b Genetics 182: 375–385. ● Gianola, D., G. de los Campos, W. G. Hill, E. Manfredi, and R. L. Fernando, 2009 Genetics 183: 347–363. ● Bush W. S., Moore, J. H. 2012. PLoS Comput Biol. 8(12):e1002822. ● López de Maturana E., Ibáñez-Escriche N., González-Recio O., Marenne G., Mehrban H., Chanock S. J., Goddard M.E., Malats N. 2013 Hum Genet ● Legarra A., Croiseau P., Sanchez M.P., Teyssèdre S., Sallé G., Allais S., Fritz S., Moreno C.R., Ricard A., Elsen J.M..2015. Genetics Selection Evolution, 47:6 ● Rupp R, Senin P, Sarry J, Allain C, Tasca C, Ligat L, et al. 2015 PLoS Genet 11(12): e1005629.