Selección genómica

Máster en Mejora Genética y Biotecnología de la Reproducción 2010

Profesor: Dra. Noelia Ibáñez Escriche

1. Introducción

En producción animal, la mayoría de los caracteres económicamente importantes son cuantitativos, es decir, muestran distribuciones continuas. Uno de los modelos usados para explicar la variación genética de estos caracteres ha sido el modelo infinitesimal. Este modelo asume que los caracteres están determinados por un número infinito de loci aditivos no ligados con un efecto infinitesimal cada uno (Fisher 1919). Tradicionalmente, la mejora genética animal ha usado este modelo de forma exitosa y es en el que se basa la teoría de estimación del valor de mejora (Henderson 1984). En este caso, son el fenotipo y el parentesco de los individuos la información utilizada para predecir los valores genéticos de mejora.

Otro modelo propuesto para explicar la variación genética observada de estos caracteres cuantitativos es el modelo de loci finitos. Este modelo asume que hay un número finito de loci regulando la variación de los caracteres cuantitativos. De hecho, hay trabajos científicos donde muestran que en las distribuciones de los loci de los caracteres cuantitativos hay unos pocos genes con gran efecto y muchos con pequeño efecto (Shrimpton y Robertson 1998, Hayes y Goddard 2001). Ha sido en el modelo de loci finitos donde se han basado la mayoría de los estudios de búsqueda de loci, particularmente, de esos de moderado a gran efecto. La idea es incrementar la precisión de los valores de mejora usando información molecular, como son las diferencias de secuencia de ADN entre animales.

El uso de la genética molecular en mejora genética animal se ha basado en dos estrategias principalmente. La primera es la llamada gen candidato, donde se asume que un gen implicado en el la fisiología del carácter podría a través de una mutación causar la variación en el carácter. En esta estrategia, el gen o partes del gen, son secuenciadas en un número diferente de animales, y se estudia si hay una asociación entre las secuencias de ADN, ya conocidas, con la variación con el fenotipo del

carácter. Esta estrategia ha tenido algún éxito (ver, Andersson y Georges, 2004), sin embargo tiene dos problemas importantes:

1) Normalmente un gran número de genes afectan al carácter, por lo que hay que secuenciar muchos genes en muchos animales y realizar una gran cantidad de estudios de asociación, lo que conlleva a una elevada muestra de animales.
2) La mutación causal quizás se deba a un gen que no sea el considerado previamente como un candidato para ese particular carácter.

La segunda estrategia utilizada es el mapeo de loci de caracteres cuantitativos (QTL), en la cual se identifican regiones del cromosoma asociadas a variaciones de los fenotipos de los caracteres. Al contrario que en el gen candidato, esta estrategia asume como no conocidos a los genes que afectan al carácter. Se mira la asociación entre la variación alélica de un marcador de ADN (normalmente neutro) y la variación del carácter cuantitativo. Cuando esta asociación se produce, significa que el marcador esta ligado a un QTL, cuyas variantes alélicas causan variación en el carácter cuantitativo. Cuando el número de marcadores por cromosoma es pequeño uno de los problemas que puede haber es que la asociación entre marcadores y QTL persista solamente dentro de familia y, debido a la recombinación, solo por un número limitado de generaciones.

Además de para la identificación y localización de QTL concretos, la información molecular (marcadores de ADN) se ha utilizado para intentar incrementar la precisión en la selección de animales genéticamente superiores. La selección asistida por marcadores (MAS), basada en las dos estrategias previamente mencionadas, ha sido ampliamente estudiada. Sin embargo, tanto su implementación práctica como el incremento de ganancia genética debido a su aplicación han sido muy escasos.

Existen tres tipos de MAS, el basado en la selección de la mutación causada por el efecto del QTL (Gene-MAS o GAS), el basado en el equilibrio de ligamento del marcador con el QTL (LE-MAS), o el basado en el desequilibrio de ligamento (LD) entre QTL y marcador molecular (LD-MAS). Los tres tipos de MAS han sido usados en empresas de mejora genética, debido a la potenciales ventajas de ganancia adicional que presentaban respecto a la mejora genética animal clásica. La aplicación del MAS podría aportar una precisión adicional muy importante en los casos donde la selección tradicional es más difícil o ineficaz, p. ej. caracteres que se expresan solamente en las hembras (tamaño de camada, producción de leche), además de facilitar la reducción del intervalo generacional, al permitir seleccionar a edades más tempranas. Sin embargo, el éxito de su aplicación ha sido más que dudosa. Hay varios factores que determinan el fracaso del MAS:

- La proporción de varianza genética explicada por los marcadores de ADN.

- La precisión con la que son estimados los efectos de los alelos de los marcadores de los loci de los caracteres quantitativos (QTL).

La mayoría de los caracteres están influenciados por muchos genes, por lo tanto el seguimiento de un número pequeño de ellos a través de los marcadores de ADN sólo explicará una pequeña proporción de la varianza genética. Además, es posible que los genes individuales tengan un pequeño efecto y por lo tanto se necesita un gran número de datos para estimar de forma precisa sus efectos.

1.1 Selección genómica

Una alternativa del MAS donde no se busca un número limitado de QTL a través de los marcadores, sino todos los QTL fue propuesta por Meuwissen et al. (2001). Está variante del MAS fue llamada selección genómica y consiste en dividir el genoma entero en segmentos de cromosoma. La clave de este método es que son usados

todos los marcadores que cubren el genoma y por lo tanto, potencialmente, los marcadores explican toda la varianza genética. Además, dado el alto número de marcadores se asume que alguno de ellos estará están en LD con el QTL. El éxito de la selección genómica se basa en aprovechar LD, la asunción es que los efectos de los segmentos de cromosoma serán igual en todas las poblaciones porque los marcadores están en LD con el QTL que flanquean. En consecuencia, para que la selección genómica sea posible, se necesita una densidad de marcadores suficiente que asegure que todos los QTL estén en LD con uno o varios marcadores. Actualmente, dado el desarrollo tecnológico de genotipado se dispone de una gran cantidad de marcadores (p. ej. SNPs) que posibilitan la aplicación de esta metodología. Por ejemplo, en humanos existen chips de un millón de polimorfismos de nucleótidos individuales (SNPs) y en vacas, ovejas, cerdos y pollos de 50,000 SNps, lo que se traduce en una densidad de 10 a 20 SNPs por cada centimorgan.

La implementación de la selección genómica consiste básicamente en dos pasos: el primero donde se estima los efectos de los segmentos del cromosoma en la población de referencia y el segundo donde se predice el valor genómico (GEBVs) de los animales candidatos a la selección.

El modelo en el que se basa la selección genómica podría derivarse del modelo genético aditivo que se usa habitualmente. Por ejemplo el valor fenotípico de tres individuos podría modelizarse como

$$\textit{Individuo 1} \quad y_1=\mu+a_1+e_1$$
$$\textit{Individuo 2} \quad y_2=\mu+a_2+e_2$$
$$\textit{Individuo 3} \quad y_3=\mu+a_3+e_3$$

De forma generalizada el modelo quedaría

$$y_i=\mu+a_i+e_i \qquad (1)$$

donde y_i es el dato del individuo i, μ es la media general, a_i es el valor genético aditivo del individuo y e_i es el error para el dato *i*. En una situación ideal donde se conocen

todos los genes, el valor aditivo de un individuo se podría descomponer en $a_i=\Sigma_j^n q_{ij}u_j$, donde n es el número de genes, q_{ij} es el genotipo del individuo i para el gen j (por ejemplo, si el genotipo del gen es AA q=1, Aa q=0, o aa q=-1) y u_j es el efecto de sustitución alélicas del gen. Por tanto el modelo anterior quedaría

$$y_i=\mu+\Sigma_{jq}q_{ij}u_j + e_{i,} \quad (2)$$

y una vez estimados $\hat{\mu}$ y $\hat{u}_j$ se podrían estimarse los valores genéticos de los individuos sin necesidad de datos, $GEBV_i= \hat{\mu} + \Sigma_j q_{ij}\hat{u}_j$. Sin embargo, en la realidad no se conocen los genotipos de los genes que regulan un determinado fenotipo, lo que impide que se pueda estimar directamente el efecto del genotipo de un gen. En este caso lo que se conoce es el genotipo de marcadores a lo largo de todo el genoma, por lo que se asume que $\Sigma_j q_{ij}u_j \approx \Sigma_j x_{ij}g_j$, donde x_{ij} es el genotipo del individuo i para el marcador del locus j y g_j es el efecto de sustitución del marcador. Por tanto el primer paso en selección genómica sería estimar los efectos **g** usando el modelo

$$y_i=\mu+\Sigma_j x_{ij}\, g_j + e_{i,} \quad (3)$$

y el segundo paso estimar los valores genéticos utilizando las estimas de $\hat{\boldsymbol{g}}_j$

$$GEBV_i= \hat{\mu} + \Sigma_j x_{ij}\hat{g}_j$$

Nótese, que en el primer paso se necesita tanto el fenotipo como el genotipo de los marcadores de los individuos, mientras que en el segundo solo hace falta el genotipo de los marcadores.

Es importante remarcar que en la selección genómica todos los efectos de los marcadores se estiman simultáneamente. De esta manera, se evita la sobreestimación de los efectos de los QTL derivada del test múltiple, como ocurre en el MAS. Además la selección genómica no solo se puede usar para predecir el GEBV, si no que también puede usarse para el mapeo de QTLs. También, destacar que la selección genómica se puede realizar usando tanto marcadores individuales (SNP) como haplotipos de marcadores. La única diferencia entre ambos es el número de efectos a estimar por segmento de cromosoma. En el caso de los marcadores individuales, se estimará un sólo efecto por segmento, mientras que para los haplotipos de marcadores puede haber varios efectos por segmento.

Uno de los problemas de la selección genómica es el gran número de efectos de los marcadores a estimar g_j comparado con el número de observaciones fenotípicas de que se disponen, que normalmente son mucho menores. Para poder resolver este problema diferentes métodos, que a continuación se detallan, han sido propuestos.

1.2 Métodos utilizados en selección genómica

En selección genómica se han propuesto diferentes métodos para poder estimar los efectos de los marcadores a lo largo del los segmentos de los cromosomas. En la tabla 1 se observa la precisión obtenida por los diferentes métodos en la estimación de los GEBV a través de los efectos de los marcadores. La diferencia principal entre estos métodos es la asunción que se hace de las varianzas de los efectos marcadores.

Tabla 1. Comparación entre los verdaderos valores de mejora (TBV) de la población de selección y los estimados (EBV) (fuente: Meuwissen, et al. 2001). La población de refrencia para estimar los efectos de los marcadores fue de 2000.

	$r_{TBV,EBV}$+SE	$b_{TBV,EBV}$+SE
LS	0.318 ± 0.018	0.285 ±0.024
BLUP	0.732 ±0.030	0.896 ±0.045
Bayes A	0.798 ±0.018	0.827 ± 0.020
Bayes B	0.848 ± 0.012	0.946 ± 0.018

$r_{TBV,EBV:}$ correlación entre los verdaderos BV y los estimados BV; $b_{TBV,EBV}$, regresion de los verdaderos BV sobre los estimados BV.

1.2. 1 Mínimos cuadrados

Este primer método no hace asunciones de la distribución de los marcadores, porque trata estos efectos como fijos. La selección genómica utilizando mínimos cuadrados conlleva dos pasos (Meuwisen et al., 2001).

1. Se realiza un análisis de regresión simple para cada segmento, *i*, usando el modelo

$$\boldsymbol{y} = \mu \boldsymbol{1}_n + \boldsymbol{x}_i g_i + \boldsymbol{e}$$

donde $\boldsymbol{y}$ es el vector de datos, μ es la media general, $\boldsymbol{1}_n$ es un vector de unos, n es el número de datos, $\boldsymbol{x}_i$ es el genotipo del marcador i, g_i es el efecto genético del marcador y $\boldsymbol{e}$ es el vector de errores.

Por ejemplo para el marcador 1

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ y_n \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 1 \\ . \\ 1 \end{bmatrix} + \begin{bmatrix} x_1 \\ x_1 \\ . \\ x_1 \end{bmatrix} g_1 + \begin{bmatrix} e_1 \\ e_2 \\ . \\ e_3 \end{bmatrix}$$

2. Se selecciona los *m* marcadores más importantes y se estiman sus efectos simultáneamente usando una regresión múltiple

$$\boldsymbol{y} = \mu \boldsymbol{1}_n + \Sigma^{m}_{i=1} \boldsymbol{x}_i g_i + \boldsymbol{e}$$

Se asume que el resto de marcadores no incluidos en el modelo son cero.

La utilización de este método tiene dos problemas importantes. Uno es que la elección del nivel de significación. Este no puede ser muy bajo, porque sino el número de marcadores a estimar es superior al número de datos, en cuyo caso los mínimos cuadrados no pueden usarse. Por otra parte, la estimación de los marcadores por regresión simple produce un problema de sobreestimación de los efectos de los QTL derivado del test múltiple.

1.2. 2 *Ridge regression* y BLUP

Para poder solucionar el problema de sobreestimación en el contesto del MAS, Whittaker et al. (2000) aplicó el *ridge regression*. Este método asume que los efectos de los marcadores $\boldsymbol{g}$ son aleatorios con una varianza común. Con este método, se puede estimar simultáneamente todos los efectos de los marcadores porque las estimas de g_i son reducidas hacia la media, lo que evita una sobreestimación de sus

efectos. El r*idge regresión* puede aplicarse en selección genómica de la siguiente manera:

$$\hat{\boldsymbol{g}} = (\boldsymbol{X'X}+\lambda \boldsymbol{I})^{-1}\boldsymbol{X'y}$$

donde $\boldsymbol{X}$ es la matriz de los genotipos de los marcadores de cada individuo. La dificultad de este método está en la elección del λ que es arbitraria. Sin embargo, si $\lambda=\sigma^2_e/\sigma^2 g$ en la ecuación del *ridge regresion*, este método es igual al BLUP usado por Meuwisen et al. (2001) y similar al método propuesto por Gianola et al. (2003).

Una cuestión importante de este método es la elección o estimación de la varianza de los efectos de los marcadores σ^2_g. Meuwissen et al. (2001) sustituye está varianza por la la varianza genética esperada de un modelo genético de mutación y deriva y asumiendo la distribución de los efectos de QTL mostrada por Hayes y Goddaed (2001). Sin embargo, cabe remarcar que la varianza calculada por Meuwissen es la varianza genética que es distinta a la varianza de los efectos de los marcadores σ^2_g que es la que se debe utilizar en el BLUP.

Una manera más correcta es estimar la varianza de los efectos de los marcadores. Gianola et al. (2003) muestra como se puede estimar a través de asignar distribuciones a priori a la varianza de los marcadores y a la del error.

En el anejo se muestra un ejemplo práctico de selección genómica usando Ridge regression-BLUP.

1.2.3 BLUP con estimación de las varianzas

En este BLUP el modelo utilizado es similar al utilizado en el apartado anterior y el correspondiente a la expresión (3)

$$y_i=\mu+\Sigma_j X_{ij}\, g_j + e_i\,,$$

donde se asume que los datos condicionados a los parámetros $p(\boldsymbol{y}|\ \mu, \boldsymbol{X}, \boldsymbol{g}, \sigma^2_g, \sigma^2_e)$ se distribuyen como una normal $N(\mu\boldsymbol{1}_n+\Sigma_j\boldsymbol{X}_j\,\boldsymbol{g}, \sigma^2_e)$.

Nótese, que en el texto los parámetros desconocidos se indican en color rojo, mientras que los parámetros conocidos en color negro.

Las distribuciones asumidas a priori para los parámetros no conocidos son:

$\mu \sim$ constante (no es una distribución propia, pero su posterior sí)

$\mathbf{e} \sim N(0, \sigma^2_e)$ y σ^2_e es una chi-cuadrado $v_e S^2_e \chi^{-2}_{ve}$, *donde v son los grados de libertad y* S^2 *el parámetro de escala.*

$\mathbf{g} \sim N(0, \sigma^2_g)$ y σ^2_g es una chi-cuadrado $v_g S^2_g \chi^{-2}_{vg}$, se asume igual para todos los efectos de los marcadores.

Aplicando el teorema de Bayes

La distribución posterior de los parámetros no conocidos es proporcional a

$$p(\mu, \mathbf{g}, \mathbf{e}, \sigma^2_g, \sigma^2_e | \mathbf{y}) \propto p(\mathbf{y} | \mu, \mathbf{g}, \mathbf{e}) p(\mu) p(\mathbf{g} | \sigma^2_g) p(\mathbf{e} | \sigma^2_e) p(\sigma^2_g) p(\sigma^2_e)$$

$$\propto exp\left[-\frac{(\mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}\mathbf{g})'(\mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}\mathbf{g}) + v_e S^2_e}{2\sigma^2_e} \right] exp - \frac{(\mathbf{g}'\mathbf{g} + v_g S^2_g)}{2\sigma^2_g} (\sigma^2_e)^{-(\frac{n+ve}{2}+1)} (\sigma^2_g)^{-(\frac{n+vg}{2}+1)}$$

En este caso las distribuciones condicionales de cada parámetro serían:

$(\mu | \mathbf{g}, \mathbf{e}, \sigma^2_g, \sigma^2_e, \mathbf{y}) \sim N(\mathbf{1}'\mathbf{X}\mathbf{g}/n, \sigma^2_e/n)$

$(g_j | \mu, \mathbf{g}_{-j}, \mathbf{e}, \sigma^2_g, \sigma^2_e, \mathbf{y}) \sim N(\hat{g}_j, \sigma^2_e/c_j)$ *donde*

$$\hat{g}_j = \frac{\mathbf{x}'_j \left(\mathbf{y} - \mu \mathbf{1}_n - \sum_{-j,j} \mathbf{x}_{-j} g_{-j} \right)}{c_j} \quad \text{y} \quad c_j = \mathbf{x}'_j \mathbf{x}_j + \frac{\sigma^2_e}{\sigma^2_g}$$

$(\sigma^2_e | \mu, \mathbf{g}, \mathbf{e}, \sigma^2_g, \mathbf{y}) \sim \hat{v}_e \hat{S}^2_e \chi^{-2}_{ve}$ donde

$$\mathbf{e} = (\mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}\mathbf{g}) \quad \text{y} \quad \hat{S}^2_e = \left[(\mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}\mathbf{g})'(\mathbf{y} - \mu \mathbf{1}_n - \mathbf{X}\mathbf{g}) + v_e S^2_e \right] / \hat{v}_e, \quad \hat{v}_e = n + v_e$$

v_e son los grados de libertad y *n* es el número de datos

$$(\sigma^2_g | \mu, \mathbf{g}, \mathbf{e}, \sigma^2_g, \mathbf{y}) \sim \hat{v}_g \hat{S}^2_g \chi^{-2}_g \quad \text{donde} \quad \hat{S}^2_g = \left[\mathbf{g}'\mathbf{g} + v_g S^2_g \right] / \hat{v}_g, \quad \hat{v}_g = q + v_g$$

v_g son los grados de libertad y q *es* el número de marcadores.

La posterior condicional de g_j corresponde a la posterior condicional de cada marcador *(j=1,.. q*), sin embargo también se podría utilizar la posterior condicional conjunta de todos los marcadores, que en este caso sería

$$(\mathbf{g} \mid \mu, \mathbf{e}, \sigma^2_g, \sigma^2_e, \mathbf{y}) \sim N(\hat{\mathbf{g}}, \sigma^2_e / c)$$ donde

$$\hat{\mathbf{g}} = \frac{\mathbf{X}'\mathbf{y}}{c} \text{ y } c = \mathbf{X}'\mathbf{X} + \frac{\sigma^2_e}{\sigma^2_g}\mathbf{I}$$

Nótese, que en este caso la expresión de $\hat{g}$ es idéntica a la utilizada en el *rigde regession-BLUP* de Meuwissen (2001).

Generalmente, el muestreo de los parámetros se hace con el algoritmo de Gibbs, que se basa en el muestreo de la distribución condicional posterior de cada parámetro.

Ejemplo de algoritmo de muestreo de Gibbs

1. Se les adjudica un valor inicial (0) a todos los parámetros desconocidos.
2. Se muestrea la $\mu^{(1)}$ de una distribución Normal, distribución que corresponde a la posterior de μ condicionada a los parámetros iníciales

$$(\mu^{(1)} \mid \mathbf{g}^{(0)}, \mathbf{e}^{(0)}, \sigma^2_g{}^{(0)}, \sigma^2_e{}^{(0)}, \mathbf{y}) \sim N(\mathbf{1'y - 1'Xg}^{(0)}/n, \sigma^2_e{}^{(0)}/n).$$

3. Se muestrea la $g_j^{(1)}$ de una distribución Normal, distribución que corresponde a la posterior de g_j condicionada a los demás parámetros

$$(g_j^{(1)} \mid \mu^{(1)}, \mathbf{g}_{-j}^{(0)}, \mathbf{e}^{(0)}, \sigma^2_g{}^{(0)}, \sigma^2_e{}^{(0)}, \mathbf{y}) \sim N(\hat{g}_j^{(1)}, \sigma^2_e{}^{(0)} / c_j)$$

donde

$$\hat{g}_j^{(1)} = \frac{\mathbf{x}'_j(\mathbf{y} - \mu^{(1)}\mathbf{1}_n - \sum_{j' \neq j} \mathbf{x}_{j'} g^{(0)}_{j'})}{c_j} \text{ y } c_j = \mathbf{x}'_j\mathbf{x}_j + \frac{\sigma^{2(0)}_e}{\sigma^{2(0)}_g}.$$

4. Se muestrea la $\sigma^2_e{}^{(1)}$ de una distribución chi-cuadrado, distribución que corresponde a la posterior de σ^2_e condicionada a los demás parámetros

$$(\sigma^2_e{}^{(1)} \mid \mu^{(1)}, \mathbf{g}^{(1)}, \mathbf{e}^{(1)}, \sigma^2_g, \mathbf{y}) \sim \hat{v}_e \hat{S}^2_e \chi^{-2}_{ve}$$

donde $\mathbf{e}^{(1)} = (\mathbf{y} - \mu^{(1)}\mathbf{1}_n - \mathbf{X}\mathbf{g}^{(1)})$ y

$$\hat{S}^2_e = \left[(\mathbf{y} - \mu^{(1)}\mathbf{1}_n - \mathbf{X}\mathbf{g}^{(1)})'(\mathbf{y} - \mu^{(1)}\mathbf{1}_n - \mathbf{X}\mathbf{g}^{(1)}) + v_e S^2_e\right] / \hat{v}_e, \quad \hat{v}_e = n + v_e \quad .$$

5. Se muestrea la $\sigma_g^{2(1)}$ de una distribución chi-cuadrado, distribución que corresponde a la posterior de σ_g^2 condicionada a los demás parámetros muestreados previamente

$$(\sigma^2{}_g^{(1)} \mid \mu^{(1)}, \mathbf{g}^{(1)}, \mathbf{e}^{(1)}, \sigma^2{}_e^{(1)}, \mathbf{y}) \sim \hat{v}_g \hat{S}_g^2 \chi_g^{-2} \text{ donde}$$

$$\hat{S}_g^2 = \left[g^{(1)\prime} g^{(1)} + v_g S_g^2 \right] / \hat{v}_g, \quad \hat{v}_g = q + v_g \quad .$$

6. Se repite, un establecido número de veces, los pasos del punto 1 al 4.

1.2. 4 Bayes A

El denominado Bayes A se diferencia del BLUP en que para cada g_j se asume una varianza diferente. En este caso las distribuciones a priori de g_j y $\sigma^2{}_g$ serían $g_j \sim N(0, \sigma^2{}_{gj})$, $\sigma^2{}_{gj} \sim v_{gj} S^2{}_{ggj} X^{-2}{}_{vgj}$ y por tanto las distribuciones posteriores serían

$$(g_j \mid \mu, \mathbf{g}_{-j}, \mathbf{e}, \sigma^2{}_g, \sigma^2{}_e, \mathbf{y}) \sim N(\hat{g}_j, \sigma^2{}_e / c_j) \text{ donde}$$

$$\hat{g}_j = \frac{\mathbf{x}'_j (\mathbf{y} - \mu \mathbf{1}_n - \sum_{j' \neq j} \mathbf{x}_{j'} g_{j'})}{c_j} \text{ y } c_j = \mathbf{x}'_j \mathbf{x}_j + \frac{\sigma_e^2}{\sigma_{gj}^2},$$

$$(\sigma^2{}_{gj} \mid \mu, \mathbf{g}, \mathbf{e}, \sigma^2{}_{gj}, \mathbf{y}) \sim \nu_{gj} \hat{S}^2_{gj} \chi^{-2}_{gj} \text{ donde}$$

$$\hat{S}^2_{gj} = \left[g'_j g_j + v_{gj} S^2_{gj} \right] / \hat{v}_{gj}, \quad \hat{v}_g = 1 + v_{gj}$$

ν_{gj} son los grados de libertad asignados a priori.

1.2. 5 Bayes B

El Bayes B a diferencia del Bayes A asume con probabilidad π, dependiente de la tasa de mutación, que hay un porcentaje de marcadores que no tienen efecto ni varianza, mientras que el resto de marcadores sí que tienen efecto y varianza.

Esta asunción se realiza a través de la distribución a priori de g_j y de σ^2_{gj}

$$(g_j \mid \pi, \sigma^2_{gj}) \begin{cases} \sim N(0, \sigma^2_{gj}) & \text{con probabilidad } (1 - \pi), \\ = 0 & \text{con probabilidad } \pi \end{cases}$$

y $\sigma^2{}_{gj} \mid v_{gj} S^2{}_{gj} \sim v_{gj} S^2{}_{gj} \chi^{-2}{}_{vgj}$

entonces

$$(\sigma^2_{gj} \mid \pi) \begin{cases} \sim univariante\text{-}t(0, S^2_{gj}, v_{gj}) & con\ probabilidad\ (1-\pi), \\ = 0 & con\ probabilidad\ \pi \end{cases}$$

En este caso la posterior de g_j y σ^2_{gj} son distribuciones mixtas con forma no conocida. La solución propuesta por Meuwissen et al. (2001) fue muestrear g_j y σ^2_{gj} conjuntamente usando el algoritmo de Metropolis-Hastings. La estrategia de muestreo sería la siguiente:

1. Muestreo de μ y σ^2_e de su distribución normal con el algoritmo de Gibbs.
2. Muestrear conjuntamente con el algoritmo de Metropolis-Hastings g_j y σ^2_{gj} de su distribución condicional

$$(g_j, \sigma^2_{gj} \mid \mu, \mathbf{e}, \mathbf{g}_{-j}, \boldsymbol{\sigma}^2_{\mathbf{g}-j}, \sigma^2_e, \mathbf{y}) \sim N(\mu \mathbf{1}_n + \sum_j \mathbf{x}_j g_{-j}, (\mathbf{X}'\sigma^2_{gj}\mathbf{X} + \mathbf{I}\sigma^2_e)^{-1}) \ .$$

1.2. 6 Bayes C

Uno de los problemas del Bayes B es que la probabilidad π con la que se asigna que el marcador y la varianza del marcador toman valor cero se escoge arbitrariamente. Además, otro problema que presenta tanto del Bayes A como el Bayes B es la gran dependencia de las varianzas del prior. Esto se debe a que se estima una varianza por marcador y la información para ello es limitada. El Bayes C se diferencia del Bayes B porque se estima el valor de π conjuntamente con los otros parámetros desconocidos, además se considera una misma varianza para todos los efectos g_j distintos de cero, lo que hace que disminuya la dependía del prior.

Así en este caso

$$(g_j \mid \pi, \sigma^2_g) \begin{cases} \sim N(0, \sigma^2_g) & con\ probabilidad\ (1-\pi), \\ = 0 & con\ probabilidad\ \pi \end{cases}$$

y $\sigma^2_g \mid v_g S^2_g \sim v_g S^2_g \chi^{-2}_{vg}$

Además

$\pi \sim Uniforme\ (0, 1)$

En este caso de la probabilidad (1-π) todas las distribuciones posteriores son conocidas, por lo que el muestreo se haría con el algoritmo de Gibbs de la siguiente manera:

- El Muestreo de los parámetros de μ, σ^2_e y $\boldsymbol{g}$ serian de distribuciones normales al igual que en el BLUP.
- El muestreo de σ^2_g se haría de una chi-cuadrado $(\sigma^2_g|\boldsymbol{y}, \mu, \boldsymbol{g}, \pi, \sigma^2_e) \sim \hat{v}_g \hat{S}^2_g \chi^{-2}_g$

  donde $\hat{v}_g = q + v_g$ y $\hat{S}^2_g = \dfrac{\boldsymbol{g'g} + v_g S^2_g}{\hat{v}_g}$
- El muestreo de π se haría de la siguiente distribución conocida: $f(\pi|\boldsymbol{y}, \mu, \boldsymbol{g}, \sigma^2_g, \sigma^2_e) = \pi^{(q-m)}(1-\pi)^m$, donde $m=\boldsymbol{g'g}$ y q es el número de marcadores. La distribución de muestreo de π corresponde a una distribución beta con $a=q-m+1$ y $b=m+1$.

2. Bibliografía.

Andersson L, Georges M. 2004. Domestic-animal genomics: deciphering the genetics of complex traits. *Nat Rev Genet.* 5(3):202-212.

Fischer, R. A. 1918. The correlation between relatives: the supposition of mendelain inheritance. *Transactions of the royal society of Edinburgh*. 52:399.

Gianola, D., Perez-Enciso, M. Toro, M. A. 2003. Genomic assisted prediction of genetic value: Beyond the ridge. *Genetics* 163:347-365.

Hayes, B. J. and Goddard, M.E. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* 33: 209-229.

Henderson, C. R. 1984. Applications of linear models in animal breeding. *Can. Catal. Publ. Data, Univ Guelph, Canada*.

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

Shrimpton, A. E., Robertson, A. 1988. The Isolation of Polygenic Factors Controlling Bristle Score in Drosophila melanogaster. II. Distribution of Third Chromosome Bristle Effects Within Chromosome Sections. *Genetics* 118: 445-459.

Whittaker, J. C., Thompson, R., Denham, M. C. 2000. Marker-assisted selection using ridge regresión. *Genet. Res*. 75:249-252.

Anejo I.

Ejemplo de selección genómica usando Ridge regression-BLUP

En este ejemplo se considera datos correspondientes a 6 animales con tres marcadores SNP. La varianza del error de los datos es σ^2_e=2 y la de los marcadores σ^2_g *=1.*

Tabla 2. Fichero de Datos

Animal	SNP1	SNP2	SNP3	Fenotipo
1	*AA*	*Bb*	*CC*	11
2	*Aa*	*BB*	*cc*	10.5
3	*aa*	*BB*	*cc*	9
4	*AA*	*bb*	*cc*	9.5
5	*Aa*	*bb*	*Cc*	8.5
6	*aa*	*bb*	*Cc*	10

El modelo del cual se ha simulado los datos es el siguiente,

$$\boldsymbol{y}=\mu\boldsymbol{1}_n+\boldsymbol{X}\,\boldsymbol{g}\ +\boldsymbol{e},$$

donde el vector de $\boldsymbol{1}_n$**'** corresponde a [1 1 1 1 1 1]

Existen diversas formas de codificar los SNPs, en este caso utilizaremos la siguiente codificación:

- *homocigoto dominante =1,*
- *heterocigoto =0,*
- *homocigoto recesivo =-1.*

Por tanto el diseño de la matriz ***X*** (en rojo) para estos datos es

Animal	**SNP1**	**SNP2**	**SNP3**
1	**1**	**0**	**1**
2	**0**	**1**	**-1**
3	**-1**	**1**	**-1**
4	**1**	**-1**	**-1**
5	**0**	**-1**	**0**
6	**-1**	**-1**	**0**

Para estimar los parámetros desconocidos μ y **g** se construyen las ecuaciones del modelo mixto

$$\begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\mathbf{X} \\ \mathbf{X}'\mathbf{1}_n & \mathbf{X}'\mathbf{X} + \lambda\mathbf{I} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_n'\mathbf{y} \\ \mathbf{X}'\mathbf{y} \end{bmatrix},$$

donde $\lambda=\sigma^2_e/\sigma^2_g$ =2 y **I** es una matriz de identidad de dimensiones 3x3 (Nº SNPsx Nº SNPs).

Las ecuaciones del modelo mixto en nuestro caso son

$$\begin{bmatrix} 6 & 0 & 0 & -2 \\ 2 & 6 & 0 & 3 \\ 2 & 0 & 6 & 1 \\ 0 & 3 & 1 & 6 \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} 58.5 \\ 1.5 \\ 1.5 \\ -18 \end{bmatrix}$$

Y por tanto las estimas de para μ y ***g*** son

$$\begin{bmatrix} \hat{\mu} \\ \hat{g}_1 \\ \hat{g}_2 \\ \hat{g}_3 \end{bmatrix} = \begin{bmatrix} 9.24 \\ -2.06 \\ 2.57 \\ 1.54 \end{bmatrix},$$

Ahora, una vez estimados los efectos ***g*** podríamos estimar el valor genético de un individuo conociendo sus marcadores.

$$\text{GEBV} = \mathbf{X}\hat{\mathbf{g}}$$

Por ejemplo, considerando los animales de la tabla 3.

Tabla 3. Animales genotipados

Animal	SNP1	SNP2	SNP3
7	*AA*	*Bb*	*Cc*
8	*AA*	*Bb*	*cc*
9	*aa*	*Bb*	*CC*
10	*Aa*	*BB*	*cc*
11	*Aa*	*bb*	*CC*
12	*Aa*	*bb*	*Cc*

El diseño de la matriz ***X*** (en rojo) para los animales de la tabla 3 es

Animal	SNP1	SNP2	SNP3
7	**1**	**0**	**0**
8	**1**	**0**	**-1**
9	**-1**	**1**	**1**
10	**0**	**1**	**-1**
11	**0**	**-1**	**1**
12	**0**	**-1**	**0**

Usando los valores $\hat{\boldsymbol{g}}$ estimados anteriormente obtendríamos los valores genotípicos de cada

Animal	GEBV
7	**-2.06**
8	**-0.51**
9	**-2.06**
10	**-1.03**
11	**1.03**
12	**2.57**

Nótese, que también podríamos predecir el fenotipo sumando la $\hat{\mu}$ al GEBV.