# ESTIMACIÓN DE PARÁMETROS DE UN QTL MEDIANTE GENOTIPADO DE LOTES SELECTOS DE ADN

Carleos C., Corral N., López T., Baro J.A., Canon J.

**Resumen**

El genotipado por lotes (*pool genotyping*) es una técnica de laboratorio que permite determinar, con cierto grado de error, frecuencias alélicas en un grupo de individuos a partir de un único genotipado. Dada la reducción de costos que supone, las posibilidades de su aplicación en estudios de análisis de ligamiento y asociación son considerables. Se pretende estudiar la estimación de parámetros relativos a un gen con influencia en un carácter cuantitativo (cuantigén; quantitative trait locus, QTL) mediante la información obtenida del genotipado por lotes para un marcador polimórfico. En concreto, se aborda la estimación de las medias y las varianzas de los diferentes grupos genotípicos, y la estimación de las frecuencias de los alelos del cuantigén. Por simplicidad, el modelo asume dominancia completa y determinación exacta de las frecuencias respecto a la herencia paterna. Se concluye que: *a*) las estimaciones de la media relativa al genotipo dominante y de la varianza son centradas y bantante precisas; *b*) la estimación de la media relativa al genotipo recesivo es centrada y menos precisa; *c*) la estimación de la frecuencia de los alelos del cuantigén no es factible.

## Introducción

Se han sugerido diversos métodos para reducir costos en la experimentación sobre la relación entre el genotipo y el fenotipo. Cabe destacar el genotipado selectivo (Lebowitz & al. 1987), consistente en genotipar únicamente los animales con fenotipo extremo. El problema de la estimación de parámetros genéticos en un experimento de genotipado selectivo ha sido estudiado, entre otros, por Darvasi & Soller (1992) y Muranty & Goffinet (1997), quienes presentaron aproximaciones a las estimaciones máximo-verosímiles. Un paso más drástico es el genotipado selectivo por lotes; se trata de obtener estimaciones de frecuencias alélicas en una muestra combinada de múltiples individuos. Darvasi & Soller (1994) presentan estimadores para el genotipado por lotes en el caso homocedástico. En el presente trabajo se enfoca el problema desde las fórmulas de Hill (1998) sobre verosimilitud de recuentos alélicos tras selección.

# Métodos

Se estudiará un carácter cuantitativo controlado por un cuantigén ligado a un marcador polimórfico. Se supondrá que el *fenotipo* $Y$ de un individuo es una variable aleatoria continua de distribución dependiente del genotipo del individuo, $X$ (variable aleatoria discreta cuyo recorrido es G). La función de distribución de $Y$ se denotará $\Phi_Y$, y la función de densidad $\phi_Y$; la función de distribución de $Y$ condicionado al genotipo $X = x$ se denotará $\Phi_x$, y la función de densidad $\phi_x$:

$$\phi_Y = \sum_x \Pr[X = x]\phi_x$$

.

Sea una muestra de $N$ individuos; si de cada individuo se conoce el fenotipo $y_i$ y el genotipo $x_i$ $(i = 1 \dots N)$, la verosimilitud de la muestra es:

$$\Pr[(x_i, y_i)_{i=1\dots N}] = \prod_{i=1}^{N} \Pr[X = x_i]\phi_{x_i}(y_i)$$

En la técnica de *genotipado selectivo individual*, se ordenan los $N$ individuos por su fenotipo y se agrupan en tres conjuntos: los $L$ inferiores, los $U$ superiores y los $N - L - U$ restantes. La información sobre $X$ se pierde para los $N - L - U$ individuos restantes. Una estrategia alternativa de selección consiste en fijar dos umbrales $l$ y $u$ y obtener los genotipos $X$ de los individuos más bajos que $l$ o más altos que $u$; $L$ sería ahora el número de individuos menores que $l$, y $U$ el número de individuos mayores que $u$. La verosimilitud de esta muestra es:

$$\Pr[(y_i)_{\substack{i=1,\dots,L,\\ N-U+1,\dots,N}}, (x_i, y_i)_{i=L+1,\dots,N-U}] =$$

$$\prod_{\substack{i=1,\dots,L,\\ N-U+1,\dots,N}} \left(\sum_x \Pr[X = x]\phi_x(y_i)\right) \cdot \prod_{i=L+1}^{N-U} (\Pr[X = x_i]\phi_{x_i}(y_i)) \quad (1)$$

En la técnica de *genotipado de lotes selectos* se dispone de todos los fenotipos $y_i (i = 1, \dots, N)$, pero la información genotípica se limita a las frecuencias de los genotipos en los grupos formados. Es decir, no hay información individual sobre genotipos. En principio, imitando el genotipado selectivo individual, supondremos los individuos de la muestra clasificados en tres grupos. Sean $l_x$, $c_x$, $u_x$ el número de individuos con genotipo $x$ entre los de la parte inferior, central y superior, respectivamente, de la muestra ordenada de fenotipos.

Hill (1998) da la verosimilitud para una muestra en que se seleccionan los individuos superiores. Aquí la extendemos teniendo en cuenta la selección de individuos inferiores y superiores. La verosimilitud de una muestra de genotipos

conjuntos, prescindiendo de los valores de los fenotipos, vendrá dada por:

$$\Pr[\{l_x, c_x, u_x\}_x] = N! \prod_x \frac{q_x^{l_x+c_x+u_x}}{l_x! c_x! u_x!} \cdot$$
$$\cdot \int_{l=-\infty}^{\infty} \int_{u=l}^{\infty} \prod_x \{\Phi_x(l)^{l_x} [1-\Phi_x(u)]^{u_x} [\Phi_x(u)-\Phi_x(l)]^{c_x}\} \cdot$$
$$\cdot \sum_x \sum_z \frac{l_x u_z \phi_x(l) \phi_z(u)}{\Phi_x(l)[1-\Phi_z(u)]} du dl \quad (2)$$

donde $q_x = \Pr[X = x]$.

Se describen a continuación ciertas variantes de (2). Si los $c_i$ son desconocidos, se tiene

$$\Pr[\{l_x, u_x\}_x] = \frac{N!}{(N-L-U)!} \prod_x \frac{q_x^{l_x+u_x}}{l_x! u_x!} \cdot$$
$$\cdot \int_{l=-\infty}^{\infty} \int_{u=l}^{\infty} \prod_x \{\Phi_x(l)^{l_x} [1-\Phi_x(u)]^{u_x}\} \left\{ \sum_x q_x [\Phi_x(u) - \Phi_x(l)] \right\}^{N-L-U} \cdot$$
$$\cdot \left\{ \sum_x \sum_z \frac{l_x u_z \phi_x(l) \phi_z(u)}{\Phi_x(l)[1-\Phi_z(u)]} \right\} du dl \quad (3)$$

sumando en (2) sobre todos los $c_x$ posibles.

Si, en vez de un solo grupo central, se dispone de las frecuencias $c_x^k$ ($k = 1, \ldots, K$) en $K$ grupos centrales, la verosimilitud es:

$$\Pr[\{l_x, c_x^1, \ldots, c_x^K, u_x\}_x] = N! \prod_x \frac{q_x^{l_x+c_x^1+\cdots+c_x^K+u_x}}{l_x! c_x^1! \cdots c_x^K! u_x!} \cdot$$
$$\cdot \int_{c_0=-\infty}^{\infty} \int_{c_1=c_0}^{\infty} \cdots \int_{c_K=c_{K-1}}^{\infty} \prod_x \{\Phi_x(c_0)^{l_x} [1-\Phi_x(c_K)]^{u_x} \prod_{k=1}^{K} [\Phi_x(c_k) - \Phi_x(c_{k-1})]^{c_x}\} \cdot$$
$$\cdot \sum_{x_0} \cdots \sum_{x_K} \frac{\prod_{k=0}^{K} c_{x_k}^k \phi_{x_k}(c_k)}{\Phi_{x_0}(c_0) \prod_{k=1}^{K} [\Phi_{x_k}(c_k) - \Phi_{x_k}(c_{k-1})]} \prod_{k=0}^{K} dx_k \quad (4)$$

donde $c_k$ recorre los posibles fenotipos del individuo más alto del grupo central $k$-ésimo ($k = 1 \ldots K$) y $c_0$ recorre el posible fenotipo del individuo más alto de la cola inferior.

Si se incluye en la fórmula (2) la información de los fenotipos del individuo más alto de la cola inferior y del individuo más bajo de la cola superior,

desaparecen las integrales:

$$\Pr[\{l_x, c_x, u_x\}_x, l, u] = N! \prod_x \frac{q_x^{l_x+c_x+u_x}}{l_x! c_x! u_x!} \cdot \cdot \prod_x \{\Phi_x(l)^{l_x}[1-\Phi_x(u)]^{u_x}[\Phi_x(u)-\Phi_x(l)]^{c_x}\} \cdot \cdot \sum_x \sum_z \frac{l_x u_z \phi_x(l)\phi_z(u)}{\Phi_x(l)[1-\Phi_z(u)]} \quad (5)$$

lo cual evita el problema de la integración numérica. Sin embargo, se muestra más adelante cómo la integración puede aportar ventajas dependiendo del modelo paramétrico. Si $l$ y $u$ representan umbrales de selección fijos en lugar de fenotipos de individuos, la probabilidad (5) corresponde a una multinomial de parámetros $N$ y $(q_x\Phi_x(l), q_x[\Phi_x(u)-\Phi_x(l)], q_x[1-\Phi_x(u)])_x$:

$$\Pr[\{l_x, c_x, u_x\}_x, l, u] = N! \prod_x \frac{q_x^{l_x+c_x+u_x}}{l_x! c_x! u_x!} \cdot \cdot \prod_x \{\Phi_x(l)^{l_x}[1-\Phi_x(u)]^{u_x}[\Phi_x(u)-\Phi_x(l)]^{c_x}\}$$

Si se incluyen en la verosimilitud los fenotipos de todos los individuos, indexando los genotipos como enteros, la expresión toma la forma:

$$\Pr[(y_i)_{i=1,\ldots,N}, (l_x, c_x, u_x)_x] = \sum_{\tau\in\mathcal{S}_L} \prod_x \prod_{i=\sum_{j=1}^{x-1} l_j+1}^{\sum_{j=1}^{x} l_j} \left(\Pr[x]\phi_x(y_{\tau(i)})\right) \cdot \sum_{\tau\in\mathcal{S}_C} \prod_x \prod_{i=\sum_{j=1}^{x-1} c_j+1}^{\sum_{j=1}^{x} c_j} \left(\Pr[x]\phi_x(y_{\tau(i)})\right) \cdot \sum_{\tau\in\mathcal{S}_U} \prod_x \prod_{i=\sum_{j=1}^{x-1} u_j+1}^{\sum_{j=1}^{x} u_j} \left(\Pr[x]\phi_x(y_{\tau(i)})\right) \quad (6)$$

donde $S_a$ indica el grupo de permutaciones de $a$ elementos.

Es habitual parametrizar la distribución del fenotipo condicionado a un genotipo a través de la distribución normal, $\Phi_x \to \mathrm{N}(\mu_x, \sigma_x)$. Se pueden obtener estimaciones máximo-verosímiles de los parámetros, dos (media y desviación típica) por cada genotipo del cuantigén. Suele suponerse homocedasticidad, en cuyo caso el número de parámetros es uno más el número de genotipos posibles. Suponiendo que existen sólo dos genotipos (hipótesis razonable para ciertos diseños experimentales, como veremos más adelante) tendríamos tres

parámetros que estimar en (5): $\mu_1$, $\mu_2$, $\sigma$. Si se considera la fórmula con integrales, (2), se puede hacer el cambio de variables $l' = \frac{l-\frac{\mu_1+\mu_2}{2}}{\sigma}$, $u' = \frac{u-\frac{\mu_1+\mu_2}{2}}{\sigma}$, con lo que la verosimilitud quedaría en función de un único parámetro $\delta = \frac{|\mu_1-\mu_2|}{2\sigma}$, o sea, la semidiferencia de medias tipificada. Esto simplifica el cálculo del máximo, al hacerse la búsqueda en una sola dimensión.

Presentamos un enfoque que permite estimar las diferentes varianzas asociadas a cada genotipo. Denótese $\pi_x^u = \Pr[Y > u|X = x] = 1 - \Phi_x(u)$, $\pi_x^l = \Pr[Y < l|X = x] = \Phi_x(l)$.Entonces se verifica

$$\left.\begin{array}{r} \pi_x^l = \Phi\left(\frac{l-\mu_x}{\sigma_x}\right) \\ \pi_x^u = 1 - \Phi\left(\frac{u-\mu_x}{\sigma_x}\right) \end{array}\right\} \Longrightarrow \left\{\begin{array}{l} \sigma_x = \frac{u-l}{\Phi^{-1}(1-\pi_x^u)-\Phi^{-1}(\pi_x^l)} =: h_\sigma(\pi_x^l, \pi_x^u) \\ \mu_x = \frac{l\Phi^{-1}(1-\pi_x^u)-u\Phi^{-1}(\pi_x^l)}{\Phi^{-1}(1-\pi_x^u)-\Phi^{-1}(\pi_x^l)} =: h_\mu(\pi_x^l, \pi_x^u) \end{array}\right.$$

Sea $n_x^l$ = "n obs.$< l$ y de tipo $i$"; $n_x^l \to B(N, \pi_x^l q_x)$. Sea $n_x^u$ = "n obs. $> u$ y de tipo $i$"; $n_x^u \to B(N, \pi_x^u q_x)$. Entonces

$$\frac{n_x^l}{N} \xrightarrow[N\to\infty]{c.s.} \pi_x^l q_x \Longrightarrow \hat{\pi}_x^l := \frac{n_x^l}{q_x N} \xrightarrow[N\to\infty]{c.s.} \pi_x^l$$

A partir de los estimadores $\hat{\pi}_x^l$ de $\pi_x^l$ y $\hat{\pi}_x^u$ de $\pi_x^u$, se obtienen los estimadores $\hat{\sigma}_x := h_\sigma(\hat{\pi}_x^l, \hat{\pi}_x^u)$ y $\hat{\mu}_x := h_\mu(\hat{\pi}_x^l, \hat{\pi}_x^u)$. Dado que $\hat{\pi}_x^l \xrightarrow[N\to\infty]{c.s.} \pi_x^l$ y $\hat{\pi}_x^u \xrightarrow[N\to\infty]{c.s.} \pi_x^u$, y como $h_\sigma$ y $h_\mu$ son continuas en $(\pi_x^l, \pi_x^u)$ (siempre que $l < u$), se tiene $\hat{\mu}_x \xrightarrow[N\to\infty]{c.s.}$ $\mu_x$ y $\hat{\sigma}_x \xrightarrow[N\to\infty]{c.s.} \sigma_x$.

Aplicando el método $\delta$ se pueden hallar las varianzas asintóticas. Como $(n_x^l, n_x^u)$ sigue una distribución multinomial $\mathrm{M}\big(N, (\pi_x^l, \pi_x^u)\big)$, su matriz de varianzas y covarianzas es

$$\begin{pmatrix} Nq_x\pi_x^l(1-q_x\pi_x^l) & -Nq_x^2\pi_x^l\pi_x^u \\ -Nq_x^2\pi_x^l\pi_x^u & Nq_x\pi_x^u(1-q_x\pi_x^u) \end{pmatrix}$$

por lo que la varianza de $(\hat{\pi}_x^l, \hat{\pi}_x^u)$ es

$$\begin{pmatrix} \frac{\pi_x^l(1-q_x\pi_x^l)}{Nq_x} & \frac{-\pi_x^l\pi_x^u}{N} \\ \frac{-\pi_x^l\pi_x^u}{N} & \frac{\pi_x^u(1-q_x\pi_x^u)}{Nq_x} \end{pmatrix}$$

y la varianza de $(\hat{\mu}_x, \hat{\sigma}_x)$ tiene una expresión compleja; así, la varianza de $\hat{\mu}_g$ es

$$\begin{aligned}\operatorname{var}(\hat{\mu}_g) \approx \pi(u-l)^2 \Bigg( & \frac{1}{2} \mathrm{e}^{\Phi^{-1}[U_g/N_g]^2} \frac{U_g}{N_g} \\ & \left(1 - \frac{U_g}{N_g} \Pr[g]\right) \Phi^{-1}[L_g/N_g]^2 \\ & - \mathrm{e}^{\frac{1}{2}\left(\Phi^{-1}[L_g/N_g]^2 + \Phi^{-1}[U_g/N_g]^2\right)} \\ & \frac{L_g}{N_g} \frac{U_g}{N_g} \Pr[g] \Phi^{-1}[L_g/N_g] \Phi^{-1}[U_g/N_g] \\ & + \frac{1}{2} \mathrm{e}^{\Phi^{-1}[L_g/N_g]^2} \frac{L_g}{N_g} \left(1 - \frac{L_g}{N_g} \Pr[g]\right) \\ & \Phi^{-1}[U_g/N_g]^2 \Bigg) / \\ & \left( N \Pr[g] \left[ \frac{\Phi^{-1}[U_g/N_g]}{\sqrt{2}} - \frac{\Phi^{-1}[L_g/N_g]}{\sqrt{2}} \right]^4 \right)\end{aligned}$$

En el caso de dos únicos grupos genotípicos, una estimación del parámetro $\delta$ (diferencia de medias tipificada) que no requiere conocimento de $l$ y $u$ es

$$\hat{\delta} = \lambda \left| \Phi^{-1}(\pi_1^l) - \Phi^{-1}(\pi_2^l) \right| + (1-\lambda) \left| \Phi^{-1}(1-\pi_1^u) - \Phi^{-1}(1-\pi_2^u) \right|$$

con $\lambda \in [0,1]$.

## Aplicación a familias de medio-hermanos

Para la resolución de este problema, se tendrá en cuenta solamente la herencia conjunta de dos genes: un cuantigén (*quantitative trait locus*, QTL) y un marcador. Se supondrá que existen dos alelos del cuantigén: $Q$ y $q$; y tres (se justificará más adelante) alelos del marcador: $M$, $m$ y $m'$. Se denominará *haplotipo* a la combinación de alelo del cuantigén con alelo del marcador. Todo *individuo* es portador de una pareja de haplotipos (*genotipo*). Se considera *una* familia de medio-hermanos de padre, esto es, un grupo de individuos hijos del mismo padre y de distintas madres escogidas *al azar* de una determinada población.

Se supone *ligamiento completo* entre el locus marcador y el cuantigén, o sea, la *tasa de recombinación* entre ambos es $\theta = 0$. El padre se supone *doble heterozigoto*. La población de madres se halla en *equilibrio de ligamiento*, esto es, la herencia de alelos marcadores es independiente de la herencia de alelos del cuantigén. El padre tiene genotipo $MQ/mq$; por tanto, trasmite haplotipos $MQ$ y $mq$, con probabilidad 0'5 cada uno. Nótese que solo es informativo para el ligamiento si es doble heterozigoto. Las madres trasmiten los alelos $M$, $m$ y $m'$ (éste representa cualquier otro alelo *real* distinto de los que porta el padre) del marcador con probabilidades respectivas $f$, $g$ y $1-f-g$, y los

alelos $Q$ y $q$ del cuantigén con probabilidades respectivas $t$ y $1-t$; equivalentemente: trasmiten los haplotipos $MQ, Mq, mQ, mq, m'Q, m'q$ con probabilidades respectivas $ft, f(1-t), gt, g(1-t), (1-f-g)t, (1-f-g)(1-t)$.

Por simplicidad en la notación, si $X_Q$ es el genotipo relativo al cuantigén, $X_M$ es el genotipo relativo al marcador, y $X_M X_Q$ es el genotipo conjunto, esto es, $X_M \in \{MM,\ Mm,\ Mm',\ mm,\ mm'\}$, $X_Q \in \{QQ, Qq, qq\}$, y $X_M X_Q \in \mathcal{G} := \{MMQQ, MMQq, mmqq, mmQq, MmQQ,$
$MmQq, Mmqq, mm'qq, mm'Qq, Mm'QQ, Mm'Qq\}$, se indicará $\Phi_{X_M X_Q} = \Phi_{X_Q}$ y $\phi_{X_M X_Q} = \phi_{X_Q}$.

La fórmula (2) da la verosimilitud para una muestra de genotipos conjuntos. La verosimilitud de una muestra de genotipos marcadores vendrá dada a partir de la anterior, como

$$\begin{aligned}
&\Pr[l_{MM}, l_{Mm}, l_{mm}, l_{Mm'}, l_{mm'}, c_{MM}, c_{Mm}, c_{mm}, c_{Mm'}, c_{mm'},\\
&u_{MM}, u_{Mm}, u_{mm}, u_{Mm'}, u_{mm'}] = \sum \Big\{ \Pr[\{l_i, c_i, u_i\}_{i\in\mathcal{G}}] \Big| l_{MMQQ} + l_{MMQq} =\\
&= l_{MM}, l_{mmqq} + l_{mmQq} = l_{mm}, l_{MmQQ} + l_{MmQq} + l_{Mmqq} = l_{Mm}, l_{Mm'QQ} +\\
&+ l_{Mm'Qq} = l_{Mm'}, l_{mm'qq} + l_{mm'Qq} = l_{mm'}, c_{MMQQ} + c_{MMQq} = c_{MM}, c_{mmqq} +\\
&+ c_{mmQq} = c_{mm}, c_{MmQQ} + c_{MmQq} + c_{Mmqq} = c_{Mm}, c_{Mm'QQ} + c_{Mm'Qq} = c_{Mm'},\\
&c_{mm'qq} + c_{mm'Qq} = c_{mm'}, u_{MMQQ} + u_{MMQq} = u_{MM}, u_{mmqq} + u_{mmQq} = u_{mm},\\
&u_{MmQQ} + u_{MmQq} + u_{Mmqq} = u_{Mm}, u_{Mm'QQ} + u_{Mm'Qq} = u_{Mm'},\\
&u_{mm'qq} + u_{mm'Qq} = u_{mm'} \Big\}
\end{aligned}$$

donde $l$, $c$ y $u$ representan los recuentos para la parte inferior, central y superior de la distribución fenotípica, respectivamente, y los subíndices denotan los genotipos sobre los que se hace el recuento.

El resultado de un experimento con genotipado por lotes (no individual) ofrece recuentos de alelos, no de genotipos marcadores. Por ello, la verosimilitud de una observación experimental será función de las frecuencias alélicas en cada uno de los tres grupos en que se dividió la muestra fenotípica:

$$\begin{aligned}
&\Pr[l_M, l_m, l_{m'}, c_M, c_m, c_{m'}, u_M, u_m, u_{m'}] =\\
&\sum \Big\{ \Pr[l_{MM}, l_{Mm}, l_{mm}, l_{Mm'}, l_{mm'}, \quad c_{MM}, c_{Mm}, c_{mm}, c_{Mm'}, c_{mm'},\\
&u_{MM}, u_{Mm}, u_{mm}, u_{Mm'}, u_{mm'}] \Big| 2l_{MM} + l_{Mm} + l_{Mm'} = l_M,\\
&2l_{mm} + l_{Mm} + l_{mm'} = l_m, \quad l_{Mm'} + l_{mm'} = l_{m'},\\
&2c_{MM} + c_{Mm} + c_{Mm'} = c_M, \quad 2c_{mm} + c_{Mm} + c_{mm'} = c_m,\\
&c_{Mm'} + c_{mm'} = c_{m'}, \quad 2u_{MM} + u_{Mm} + u_{Mm'} = u_M,\\
&2u_{mm} + u_{Mm} + u_{mm'} = u_m, \quad u_{Mm'} + u_{mm'} = u_{m'} \Big\} \qquad (7)
\end{aligned}$$

Dada la complejidad analítica, en la práctica se ha venido empleando un modelo simplificado, en el que interviene únicamente la herencia paterna. Como

la población de madres se supone en equilibrio de ligamiento, es decir, la herencia del cuantigén es independiente de la herencia del marcador, y el laboratorio sólo puede ofrecer información sobre alelos marcadores, tiene sentido excluir del modelo la herencia materna. De este modo, solo se considerarían dos grupos de genotipos: $\mathcal{G} = \{MQ, mq\}$. Con este $\mathcal{G}$, las fórmula (2) es igualmente válida, pero (7) se convierte en :

$$\begin{aligned}
&\Pr[l_M, l_m, l_{m'}, c_M, c_m, c_{m'}, u_M, u_m, u_{m'}] = \\
&= \sum_{l_{MQ}=\max\{0,N-l_m\}}^{\min\{N,l_M\}} \sum_{c_{MQ}=\max\{0,N-c_m\}}^{\min\{N,c_M\}} \sum_{u_{MQ}=\max\{0,N-u_m\}}^{\min\{N,u_M\}} \\
&\Pr[l_{MQ}, l_{mq} = N - l_{MQ}, c_{MQ}, c_{mq} = N - c_{MQ}, u_{MQ}, u_{mq} = N - u_{MQ}] \cdot \\
&\cdot \Pr[l_M, l_m, l_{m'}, c_M, c_m, c_{m'}, u_M, u_m, u_{m'}] = \\
&= \sum \Big\{ \Pr[l_{MQ}, l_{mq}, c_{MQ}, c_{mq}, u_{MQ}, u_{mq}] \cdot \\
&\cdot f^{l_M + c_M + u_M - l_{MQ} - c_{MQ} - u_{MQ}} g^{l_m + c_m + c_m - l_{mq} - c_{mq} - u_{mq}} \cdot \\
&\cdot (1-f-g)^{l_{m'} + c_{m'} + u_{m'}} | \\
&| l_{MQ} = \max\{0, N - l_m\}, \ldots, \min\{N, l_M\}, \\
&c_{MQ} = \max\{0, N - c_m\}, \ldots, \min\{N, c_M\}, \\
&u_{MQ} = \max\{0, N - u_m\}, \ldots, \min\{N, u_M\} \Big\}
\end{aligned}$$

Con la hipótesis de distribución normal dentro de grupo genotípico de cuantigén, se tiene que $\Phi_{MQ} \to t\mathrm{N}(\mu_{QQ}, \sigma_{QQ}) + (1-t)\mathrm{N}(\mu_{Qq}, \sigma_{Qq})$ y $\Phi_{mq} \to t\mathrm{N}(\mu_{Qq}, \sigma_{Qq}) + (1-t)\mathrm{N}(\mu_{qq}, \sigma_{qq})$. Lo habitual es aproximar las mixturas mediante distribuciones normales:

$$\Phi_{MQ} \tilde{\to} \mathrm{N}\left(t\mu_{QQ} + (1-t)\mu_{Qq}, \sqrt{t\sigma^2_{QQ} + (1-t)\sigma^2_{Qq} + t(1-t)(\mu_{QQ} - \mu_{Qq})^2}\right)$$

$$\Phi_{mq} \tilde{\to} \mathrm{N}\left(t\mu_{Qq} + (1-t)\mu_{qq}, \sqrt{t\sigma^2_{QQ} + (1-t)\sigma^2_{Qq} + t(1-t)(\mu_{QQ} - \mu_{Qq})^2}\right)$$

Darvasi & Soller (1994) utilizan directamente (2) con estimaciones heurísticas de $l_{MQ}$,$l_{mq}$, $c_{MQ}$, $c_{mq}$, $u_{MQ}$, $u_{mq}$.

Siguiendo el método heurístico utilizado ya en el caso unidimensional, se tiene ($f = g = 0$):

$$\begin{array}{lcll}
\Pr[Y < l|M] & = & \Phi\left(\frac{l-\mu_Q}{\sigma}\right) & \approx \frac{l_M}{n_M} \\
\Pr[Y < l|m] & = & t \cdot \Phi\left(\frac{l-\mu_Q}{\sigma}\right) + (1-t)\Phi\left(\frac{l-\mu_q}{\sigma}\right) & \approx \frac{l_m}{n_m} \\
\Pr[Y > u|M] & = & 1 - \Phi\left(\frac{u-\mu_Q}{\sigma}\right) & \approx \frac{u_M}{n_M} \\
\Pr[Y > u|m] & = & 1 - t \cdot \Phi\left(\frac{l-\mu_Q}{\sigma}\right) - (1-t)\Phi\left(\frac{l-\mu_q}{\sigma}\right) & \approx \frac{l_m}{n_m}
\end{array}$$

de donde

$$\hat{\mu}_Q = \frac{l\Phi^{-1}\left(1-\frac{u_M}{n_M}\right) - u\Phi(-1)\left(\frac{l_M}{n_M}\right)}{\Phi^{-1}\left(1-\frac{u_M}{n_M}\right) - \Phi(-1)\left(\frac{l_M}{n_M}\right)}$$

$$\hat{\sigma} = \frac{u-l}{\Phi^{-1}\left(1-\frac{u_M}{n_M}\right) - \Phi(-1)\left(\frac{l_M}{n_M}\right)}$$

Para estimar $\mu_q$, sin embargo, no se puede obtener una versión explícita de los estimadores. Se trata de minimizar la diferencias cuadráticas entre ciertas expresiones poblacionales y las análogas muestrales. En principio se consideraron las expresiones similares a () pero relativas a la herencia paterna del alelo recesivo:

$$\Pr[X < l|m] = \\ = t\Phi\left(\frac{l-\mu_Q}{\sigma}\right) + (1-t)\Phi\left(\frac{l-\mu_q}{\sigma}\right) \approx \\ \approx \frac{L_m}{N_m}$$

$$\Pr[X > u|m] = \\ = t\left(1-\Phi\left(\frac{u-\mu_Q}{\sigma}\right)\right) + (1-t)\left(\Phi\left(\frac{u-\mu_q}{\sigma}\right)\right) \approx \\ \approx \frac{U_m}{N_m}$$

Dada la poca precisión de la estimación, se incluyeron en el sistema las ecuaciones relativas a la media y a la varianza.

# Resultados

La función de verosimilitud respecto al parámetro $\delta$ es una curva unimodal suave, por lo que su maximización no presenta problema. La distribución del estimador máximo verosímil obtenida a través de la extracción iterada de 10 000 muestras seudoaleatorias por ordenador se ajusta notablemente a una distribución normal, centrada en el valor del parámetro y con dispersión igual a 0'10 para $N = 500$, $L/N = U/N = 0'25$, $\delta = 0, 0'125$ (*condiciones normales*) y 0'16 para $N = 200$. En el rango de casos estudiados mediante simulación, el estimador $\hat{\delta}$ sugerido proporciona valores muy similares al estimador máximo-verosímil: en las condiciones de antes, la mayor diferencia con $N = 500$ fue 0'011; con $N = 200$: 0'030.

En el caso concreto de familias de medio-hermanos, suponiendo dominancia completa y que el conocimiento de la herencia paterna puede determinarse unívocamente (o sea, cuando $f = g = 0$ o $f = 1$ o $g = 1$), los fenotipos herederos

de $MQ$ se distribuirían normalmente, y los herederos de $mq$ seguirían una mixtura. Escribiendo la verosimilitud en función de $\delta$ y $t$, la superficie generada alcanza el máximo a lo largo de una curva bien definida; el valor de la verosimilitud, sin embargo, es prácticamente constante, lo que hace que estimaciones máximo-verosímil sean muy inestables.

La siguiente tabla muestra los resultados de las estimaciones en caso de dominancia:

| | simulado | estimado | error típico |
|---|---|---|---|
| $\mu_q$ | 10'0 | 10'00 | 0'178 |
| $\mu_Q$ | 10'5 | 10'50 | 0'072 |
| $\sigma$ | 1'0 | 1'00 | 0'075 |
| $t$ | 0'5 | 0'39 | *0'343 |

Las estimaciones de la varianza y de la media asociada al alelo dominante son bastante precisas. Se logra estimar bastante bien la media del alelo recesivo. La estimación de las frecuencias de los alelos del cuantigén es muy sesgada.

La inclusión de las frecuencias alélicas en la parte central de la distribución no añade precisión a las estimaciones. La inclusión de información genotípica sobre la parte central de la distribución de fenotipos (esto es, la utilización de la fórmula (2) en lugar de la (3)) no se traduce en mayor precisión de la estimación (observada una diferencia máxima de 0'006 entre las estimaciones máximo-verosímiles para ambas funciones de verosimilitud, es decir, del orden de la precisión de la estimación; las estimaciones empíricas de los errores típicos coinciden hasta la cuarta cifra significativa).

La utilidad de agrupar los fenotipos en más de tres grupos (o sea, incluyendo varios grupos centrales) se comprobó mediante un procedimiento de regresión (no mediante estimaciones máximo-verosímiles, por la complejidad de cómputo). En las condiciones normales, la potencia aumentaba de 0'71 a 0'75 cuando se dividía la muestra en cinco o más categorías.

La utilización del modelo simplificado no implica pérdida en la potencia. Dieron el mismo resultado en lo que concierne al contraste $\mathrm{H}_0 : \delta = 0$, $\mathrm{H}_1 : \delta \neq 0$ (con $\delta = 0'125$) el análisis de datos simulados bajo el modelo complejo (incluyendo genotipos maternos) tanto mediante el modelo complejo como mediante el modelo simple.

# Discusión

Más allá de la simple detección de asociación entre un locus marcador y un carácter, el genotipado por lotes permite estimar los efectos sobre el carácter. Las estimaciones de las medias y de la desviación típica, para el caso de dominancia completa y homocedasticidad, son centradas, con más precisión para la media asociada al alelo dominante. Los métodos presentados no permiten la estimación de las frecuencias alélicas del cuantigén.

Se considerará la estimación con modelos genéticos más complejos (dominancia parcial, recombinación entre marcador y cuantigén) y con fuentes adicionales

de error (error técnico). Se buscarán estimadores de las frecuencias alélicas del cuantigén que soslayen la falta de precisión de los estimadores presentados aquí. Debe explorarse la posibilidad de modelar ciertos caracteres con distribuciones no gaussianas, lo que analíticamente no plantea ningún problema con el método de máxima verosimilitud.

# Referencias


- Darvasi A, M Soller, 1992, Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus, TAG 85:353-359.
- Darvasi A, M Soller, 1994, Selective dna pooling for determination of linkage between a molecular marker and a quantitative trait locus, Genetics 138:1365-1373.
- Hill WG, 1998, A note on the theory of artificial selection in finite populations and application to qtl detecton by bulk segregant analysis, Genet Res 72:55-58.
- Lebowitz RJ, M Soller, JS Beckmann, 1987, Trait-based analyses for the detection of linkage between marker loci and quantiative trait loci in crosses between inbred lines, TAG 73:556-562.
- Muranty H, B Goffinet, 1997, Selective genotyping for location and estimation of the effect of a quantitative trait locus, Biometrics 53:629-643.