APLICACIÓN DEL REMUESTREO EN EL ANÁLISIS DE DATOS DE *DNA POOLING*

C. Carleos,* J.A. Baro,† N. Corral,* T. López,* J. Cañón ‡

31 de enero de 2001

Resumen

Se aplican métodos de remuestreo para la obtención de intervalos de confianza y el contraste de hipótesis en la estimación de parámetros de un gen cuantitativo (QTL) cuando los datos provienen de un genotipado de lotes selectos de ADN (*DNA pooling*).

Introducción

Este trabajo es continuación de los estudios sobre la potencia [1] y la precisión de los estimadores [3] en el análisis de ligamiento entre un QTL y un marcador cuando la información sobre genotipos se obtiene mediante genotipado por lotes [5].

En el genotipado por lotes, los datos disponibles son: los fenotipos de todos los animales y las frecuencias alélicas del marcador en grupos de esos animales. Los grupos están formados por animales correlativos en orden fenotípico. Supondremos que se pueden estimar con precisión las frecuencias alélicas [4].

Modelos y estimadores

Se supone que el fenotipo tiene una distribución normal con media determinada por el genotipo respecto a un gen cuantitativo con dos alelos (Q y q); éste se supone completamente ligado ($\theta = 0$) con un marcador de tres o más alelos (M, m, m'). Sea un padre doble heterozigoto (MQ/mq) que se cruza con madres de una población en que el cuantigén y el marcador están en equilibrios de Hardy & Weinberg y de ligamiento; sea t la frecuencia de Q en la población de madres, f la de M, g la de m.

El diseño experimental consiste en una familia de medio-hermanos de padre. Tras ordenar los fenotipos, se forman grupos (p. ej. tres: los de fenotipo mayor, intermedio y menor) y se emplea el genotipado por lotes para estimar las frecuencias de los alelos del marcador en cada grupo.

Se tendrán en cuenta el modelo con el cuantigén aditivo ($\mu_{Qq} = \frac{\mu_{QQ}+\mu_{qq}}{2}$) y el modelo con el cuantigén completamente dominante ($\mu_{Qq} = \mu_{QQ}$).

Máxima verosimilitud

La verosimilitud para el diseño descrito con tres grupos es proporcional a:

$$\prod_x \frac{\Pr[x]^{l_x+c_x+u_x}}{l_x!c_x!u_x!} \cdot \prod_x \{\Phi_x(l)^{l_x}[1-\Phi_x(u)]^{u_x}[\Phi_x(u)-\Phi_x(l)]^{c_x}\} \cdot \sum_x \sum_z \frac{l_x u_z \phi_x(l)\phi_z(u)}{\Phi_x(l)[1-\Phi_z(u)]}$$

*Depto. Estadística, Univ. de Oviedo

†Depto. Ciencias Agroforestales, Univ. de Valladolid

‡Depto. Producción Animal, Univ. Complutense

donde x, z recorren los posibles genotipos; l, c y u son los recuentos de los genotipos dados en los grupos inferior, central y superior; l es el fenotipo mayor del grupo inferior; u es el fenotipo menor del grupo superior; ϕ y Φ son la función de densidad y la de distribución de una normal con la media dada por el genotipo. Para discusiones sobre variantes de la fórmula, ver [3].

Método analógico

Parte de la aproximación

$$\frac{l_M}{l_M + c_M + u_M} \approx t\Phi\left(\frac{l-\mu_{QQ}}{\sigma}\right) + (1-t)\Phi\left(\frac{l-\mu_{Qq}}{\sigma}\right)$$

y similares. Con el modelo de dominancia completa, es posible obtener expresiones explícitas de los estimadores de σ y μ_{QQ} (se demuestra que son estimadores fuertemente consistentes) [3].

Regresión

Pérez-Enciso sugirió [6] la utilización de lotes más pequeños para soslayar las perturbaciones de la estimación de las frecuencias alélicas. Su enfoque es adecuado para lotes muy pequeños (dos, tres, ... individuos). Un procedimiento alternativo (descrito en detalle en [2]) es el siguiente: Para cada fenotipo y,

$$\frac{\Pr[M|y]}{\Pr[m|y]} = \frac{\phi_M(y)}{\phi_m(y)} = e^{\frac{-1}{\sigma^2}\left[(y-\mu_M)^2-(y-\mu_m)^2\right]} = e^{\frac{-1}{\sigma^2}(\mu_M-\mu_m)[\mu_M+\mu_m-2y]} = e^{\frac{2\delta(y-\mu)}{\sigma^2}}$$

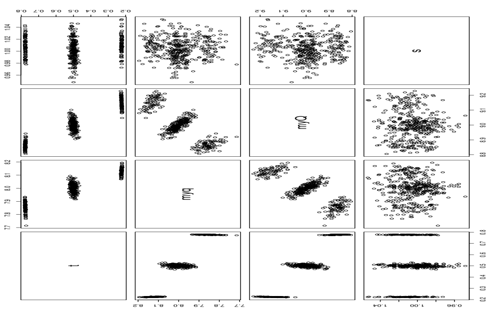
y, tomando logaritmos, se obtiene una relación lineal. Si como valores de la variable independiente se consideran las marcas de clases fenotípicas de los lotes formados, se puede estimar $\mu_Q - \mu_q$ mediante regresión.

Intervalos de confianza y constraste de hipótesis

Se han utilizado métodos de remuestreo debido a la complejidad analítica de las expresiones manejadas. Los tres usados fueron: pruebas de permuta (permutación de fenotipos y genotipos), autosuficiente (*bootstrap*, remuestreo con repetición) y autosuficiente paramétrico (considerando que la distribución de los fenotipos es gaussiana).

Resultados

Comentaremos los resultados para los siguientes valores de los parámetros: familia de $N = 500$ ó 5000 medio-hermanos; efectos gen dominante: $\mu_{qq} = 8$, $\mu_{Qq} = \mu_{QQ} = 9$; efectos gen aditivo: $\mu_{qq} = 8$, $\mu_{Qq} = 8'5$, $\mu_{QQ} = 9$; $t = 0'5$ (luego $\delta = 0'25$); $f = g = 0$.



La estimación de t es muy inestable y se localiza en los extremos (0 ó 1) con mucha frecuencia (50% con el modelo aditivo); sólo tiene forma regular y acampanada cuando se simula un cuantigén con dominancia completa y la expresión de verosimilitud refleja la dominancia completa.

Cuando se utiliza la verosimiltud o función de ajuste (estimadores analógicos) del modelo que no es el empleado para generar los datos simulados (es decir, modedo aditivo de análisis sobre datos simulados con modelo de dominancia completa, o viceversa), la función presenta varios máximos (mínimos) locales, por lo que la estimación según estos métodos requiere precauciones adicionales.

La aplicación de las distintas técnicas de remuestreo condujo a conclusiones similiares, que confirmaron los resultados de [1]. Cabe señalar que bajo la hipótesis nula de igualdad de medias, t desaparece y las funciones muestrales que se optimizan resultan casi planas respecto a t. Esto conduce a estimaciones espurias de t que desestabilizan a los estimadores de las medias $\hat{\mu}_{QQ}$ y $\hat{\mu}_{qq}$.

Como habíamos visto en [3], el método analógico y el de máxima verosimilitud coinciden en todos los casos. En cuanto al método de regresión, es necesario ajustarlo para la situación en que supuestamente debería ser más útil (muchos lotes), pues es muy sensible a la aparición de frecuencias extremas. Así, el error típico de $\hat{\delta}$ sube de 0'0354 (con 10 clases) a 0'0404 (con 100 clases). El aumento del número de clases tampoco había supuesto mayor precisión en el caso de la verosimilitud [3].

Discusión

El estudio confirma los resultados previos [1, 3] con la aplicación de técnicas alternativas. La estimación de t es precisa sólo con N grande (5000) y cuando existe dominancia completa.

Agradecimientos

Este estudio ha sido financiado por el proyecto FEDER 1FD97-0042.

Referencias

[1] Baro JA, Carleos C, Corral N, López T, Cañón J, 2001, Power analysis of QTL detection in half-sib families using selective DNA pooling, GSE (aceptado).

[2] Carleos C, 1998, Análisis de datos de genotipado por lotes en familias de medio-hermanos, Trabajo de Investigación de Doctorado, Universidad de Oviedo.

[3] Carleos, Baro, Corral, López, Cañón, Estimación de parámetros de un QTL mediante genotipado de lotes selectos de ADN, 2000, X Reunión Nacional de Mejora Genética Animal, Caldes de Montbui.

[4] Checa ML, Carleos C, Baro JA, Dunner S, Cañón J, 2000, Estimación de frecuencias alélicas en pools de ADN mediante el procesamiento de señales de electroforesis capilar, 2000, X Reunión Nacional de Mejora Genética Animal, Caldes de Montbui.

[5] Darvasi A, Soller M, Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus, 1994, Genetics 138: 1365-1373.

[6] Pérez-Enciso M, 1998, Sequential bulked typing: a rapid approach for detecting QTLs, TAG 96: 551-557.