Controlando alpha en la detección de QTLS usando líneas recombinantes puras

J. Pérez-Panadés[1], E. A. Carbonell[1] y M. J. Asíns[2]
[1]Unidad de Biometría. Instituto Valenciano de Investigaciones Agrarias (IVIA)
[2]Laboratorio de Genética. Instituto Valenciano de Investigaciones Agrarias (IVIA)

Antecedentes

Muchos de los caracteres de interés, tales como la producción, la calidad o la resistencia a enfermedades son de naturaleza continua. Definimos un QTL (Quantitative Trait Loci) como una porción del genoma que está implicado en la expresión de un carácter cuantitativo. En la actualidad es posible detectar y localizar QTLs mediante el análisis conjunto de la segregación genotípica para los marcadores y los valores fenotípicos de los individuos o las líneas. Existen varios tipos de diseños experimentales que son adecuados para este tipo de análisis y cuya utilización depende del sistema genético de la especie cultivada de que se trate. La mayor parte de los análisis de QTLs en plantas se realizan en poblaciones derivadas de líneas puras. Las líneas puras recombinantes o RILs (Recombinant Inbred Lines) se obtienen mediante autofecundaciones sucesivas de individuos de una F2 que a su vez derivan del cruce de dos líneas puras. La ventaja principal de este diseño es la replicabilidad de los genotipos, lo que nos permite poder evaluarlos varias veces, en distintas condiciones ambientales o con diferentes razas de patógenos. Un problema común en todos los métodos es la dificultad de determinar un nivel umbral de significación apropiado (valor crítico) contra el que comparar los tests estadísticos, generalmente LOD scores o cocientes de verosimilitud, con el objetivo de detectar un QTL. El origen del problema es doble. Primero, está el problema de determinar (o aproximar) la distribución del test estadístico bajo la hipótesis nula. El segundo problema es la gran cantidad de hipótesis que se utilizan para localizar un QTL a lo largo del genoma. A medida que la tecnología avanza, se dispone de más marcadores para genotipar, incrementando así la densidad de marcadores y acentuando el problema. Un problema similar ocurre cuando se identifica una región de interés y se satura de marcadores. En este caso, el problema todavía es más pronunciado y como los marcadores en esa región están más correlacionados que en otras zonas del genoma, utilizar correcciones clásicas como la de Bonferroni pueden resultar excesivamente conservadoras. El objetivo de este trabajo es evaluar el comportamiento de varios métodos desarrollados para controlar del error de Tipo I en contrastes múltiples sobre poblaciones de líneas puras recombinantes.

Material y métodos

Simulamos RILs partiendo de 200 individuos F2 con un porcentaje de muerte entre generaciones del 20 %. El número final de individuos F7 es de 67, cuyo genoma está formado por 8 grupos de ligamiento (o cromosomas) y en cada uno de ellos disponemos 10 marcadores moleculares equiespaciados (cada 10 cM). Situamos un QTL en cada uno de los 6 primeros grupos y ninguno en los dos últimos. Esto permite estimar tanto la potencia en la detección de QTLs como el error de tipo I. En relación a la distribución de los QTLs, hicimos variar factorialmente tanto la posición de cada QTL como su contribución relativa a la varianza genética total, con tres posiciones diferentes dentro del intervalo (central, intermedia y próxima al marcador) y dos contribuciones (a=efecto aditivo=1.5 ó 0.5). La heredabilidad del carácter se fijó en 0.7. Una vez simulados los genotipos de todos los individuos calculamos los valores genotípicos y simulamos los efectos de la varianza ambiental en función de la heredabilidad para obtener los valores fenotípicos. El experimento se repitió 500 veces. Teóricamente, el genotipo de todos los individuos dentro de una RIL es el mismo tras n generaciones (cuando n es grande). En general, los investigadores determinan una RIL en la séptima u octava generación (F7 o F8). Es habitual que en esa generación todavía hayan algunos marcadores que sigan segregando y las "réplicas" de tales RILs (que se obtienen por autofecundación de la última generación) no son tales réplicas. Por tanto, como todos los individuos todavía no son completamente homocigotos y puesto que el análisis usando mapeo por intervalo con los valores en bruto (valores individuales) no es correcto, debemos definir un genotipo medio. El número de réplicas por RIL que son utilizadas para obtener el genotipo medio es otro de los parámetros de interés. Consideramos los casos 1, 2, 6 y 10. Estudiamos 4 propuestas para definir el genotipo medio en los casos de segregación: A) Eliminar los marcadores que siguen segregando en la última generación, B) Codificar los marcadores segregantes como heterocigoto, Aa, C) Codificar como el genotipo del alelo dominante, AA (ya que hay marcadores que sólo distinguen entre banda, AA-Aa, y no-banda, aa) y D) Si se genotipan todos los individuos, codificando según su frecuencia de aparición de cada genotipo en la RIL. Una vez generados los datos, la detección de QTLs se efectuó mediante mapeo por intervalo usando máxima verosimilitud (Carbonell et al., 1992). Este método calcula un cociente de verosimilitud (LR) cada 2 cM (en nuestro caso) a lo largo del genoma y utiliza las información de los dos marcadores que flanquean cada posición de un posible QTL. Se detecta un QTL si sobre cualquier punto de un intervalo el LR excede de un umbral fijado. La aproximación tradicional para trabajar con comparaciones múltiples, el familywise error rate (FWER), se controla colocando un umbral suficientemente estricto para que la probabilidad de rechazar erróneamente al menos una hipótesis nula sea menor que cierto valor, normalmente 0.05. Para ello utilizamos la Corrección de Bonferroni a dos niveles: uno más suave, considerando el número de intervalos en cada grupo de ligamiento, Bc, y otro, más estricto, que

considera el número total de intervalos, Bg.

Churchill y Doerge (1994) propusieron estimar empíricamente el umbral de rechazo FWER generando diferentes muestras a partir de los datos mediante permutaciones de los valores del carácter respecto a los genotipos de los marcadores. Como el valor del carácter es ahora aleatorio para cada individuo respecto a los genotipos de los marcadores, se cumple la hipótesis nula de no ligamiento entre los marcadores y el QTL. El nivel apropiado de rechazo para un error de Tipo I deseado se calcula a partir de la distribución empírica del test estadístico. La ventaja de este método es que no requiere cumplir ninguna hipótesis sobre la distribución del carácter. Denotaremos por Pc y Pg a estos métodos. Benjamini y Hochberg (1995) propusieron controlar False Discovery Rate (FDR) como una alternativa a controlar el FWER en problemas de contrastes múltiples. Definieron el FDR como: "la proporción esperada de hipótesis erróneamente rechazadas sobre todas las hipótesis rechazadas". Cuando no todas las hipótesis nulas son ciertas, se mostró que este método controla el FDR a un nivel menor que alpha cuando los tests estadísticos son independientes (Benjamini y Hochberg, 1995) y cuando tienen una regresión positiva de dependencia (Benjamini y Yekutieli, 2001). Denotaremos este método por BHc y BHg. Benjamini y Yekutieli (2001) adicionalmente propusieron un ajuste del método anterior para tests dependientes. Se mostró que este procedimiento controla el FDR a un nivel menor que alpha bajo cualquier estructura de dependencia. Denotaremos a este método por BYc y BYg. En mapeo por intervalo, detectamos un QTL en un intervalo si el máximo LR del intervalo excede del umbral. Teniendo en cuenta este hecho, aplicamos los dos métodos anteriores también únicamente sobre los máximos de cada intervalo. Los denotaremos por MBHc, MBHg, MBYc y MBYg respectivamente. El análisis que presentamos en este trabajo está centrado sobre los cromosomas 7 y 8 de la simulación en los que no hay simulado ningún QTL y por tanto, todas las hipótesis nulas de los contrastes múltiples son ciertas. En este caso, controlar el FDR es equivalente a controlar el FWER.

Resultados

Para comparar todos los métodos anteriores analizaremos el número de detecciones que hemos declarado en cada cromosoma en las 500 simulaciones realizadas. Comprobamos que en las estrategias A y C, el número de detecciones es mucho mayor que en las estrategias B y D, excepto para los métodos Pg y Pc (permutaciones). Este resultado era previsible puesto que los genotipos medios de las dos primeras estrategias son completamente homocigotos cosa que no ocurre en las otras dos, por definición. Por tanto, como los tests estadísticos se distribuyen aproximadamente según una chi cuadrado con 1 y 2 grados de libertad respectivamente, los p-valores serán menores en las dos primeras estrategias. El número de detecciones aumenta en todos los métodos al trabajar a nivel de grupo de ligamiento excepto en BH. Los métodos Bg, Pg, BYg,

BYc son los más restrictivos en todas las estrategias no llegando en ningún caso al 2 % de falsos positivos (10 detecciones sobre 500 simulaciones). Respecto al efecto del número de réplicas por RIL en las estrategias A y C parece que el número de detecciones aumenta con el número de RILs mientras que el efecto no es tan notable en las estrategias B y D. Trabajar con los máximos de cada intervalo aumenta el número de detecciones en todos los casos de control del FDR. Comprobamos que el método Pc es el que se mantiene más cerca del 5 % de las detecciones en todas las estrategias.

Conclusiones

En nuestro contexto, para controlar el error de tipo I se recomienda el método basado en tests de permutaciones propuesto por Churchill y Doerge (1994) aplicándolo a nivel cromosómico ya que a nivel del todo el genoma es demasiado restrictivo. Éste método mantiene el error de tipo I en torno al 5 % en todos las estrategias analizadas. El principal problema de aplicar este método es que en algunos paquetes habituales para análisis de QTLs es necesario analizar de forma independiente cromosoma por cromosoma y además requiere de mucho de tiempo de computación. El método de Benjamini y Hochberg (1995), aplicado sobre los valores máximos de cada intervalo y a nivel cromosómico, MBHc, también mantiene el error de tipo I en torno al 5 % en las estrategias B y D. Respecto a la elección de la estrategia, a la vista de los resultados, aconsejamos utilizar la estrategia B o la D. Parece que en estas estrategias todos los métodos para controlar el error de tipo I funcionan mejor. Entre ellas nos decantamos por la B ya que es también mejor desde el punto de vista económico. Para definir el genotipo medio de una RIL, utilizando esta estrategia, no es necesario tener genotipados a todos los individuos que la forman como en la estrategia D. Sólo es necesario genotipar un "pool" de todos ellos. Este es el comienzo de un trabajo mucho más extenso. Debemos seguir estudiando el comportamiento de estos métodos y estrategias en los cromosomas con QTLs para comprobar la potencia de los métodos y el control del FDR. Por último estudiaremos los efectos de algunos parámetros tanto estadísticos como genéticos que intervienen en la detección de QTLs en RILs y que son de interés para los investigadores.

Referencias

Benjamini, Y. y Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 51(1):289–300.

Benjamini, Y. y Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29(4):1165–1188.

Carbonell, E. A., Gerig, T. M., Balansard, E. y Asíns, M. J. (1992). Interval mapping in the analysis of nonadditive quantitative trait loci. *Biometrics* 48:305–315.

Churchill, G. A. y Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* 138(3):963–971.