

DETECCIÓN DE HUELLAS DE SELECCIÓN UTILIZANDO EL RADIO ESPECTRAL DE LA MATRIZ DE COANCESTRÍA MOLECULAR

Munilla¹, S., González-Rodríguez, A., Mouresan, E. F., Moreno, C., Altarriba, J. y Varona, L.

¹Grupo de Genética Cuantitativa y Mejora Animal. Facultad de Veterinaria, Universidad de Zaragoza. Calle Miguel Servet 177, 50013 Zaragoza, España. Ivarona@unizar.es

INTRODUCCIÓN

La disponibilidad de marcadores moleculares ha revolucionado la mejora genética animal (véase, por ejemplo, la revisión de Dekkers y Hospital, 2002). Entre sus aplicaciones más recientes, el empleo de SNPs con el fin de detectar huellas de selección se ha posicionado como una alternativa valiosa en la búsqueda de genes asociados a caracteres productivos o adaptativos en especies domésticas. Las huellas de selección se definen como patrones distintivos de variación en secuencias de ADN, atribuibles a procesos de selección natural o artificial. En este contexto, los métodos de detección de huellas de selección constituyen una colección de procedimientos estadísticos que ponen a prueba la hipótesis de que la variabilidad observada puede explicarse perfectamente por procesos neutrales de mutación y deriva. En general, estos métodos asumen que las secuencias de marcadores que se utilizan para construir el estadístico provienen de una muestra aleatoria de individuos. En poblaciones animales, sin embargo, los individuos suelen estar emparentados entre sí y, en tal caso, sus genomas no son independientes. Si bien es técnicamente posible evitar el muestreo de individuos emparentados, esta situación puede restringir el tamaño de la muestra y afectar así la potencia de la prueba de hipótesis. Como alternativa, en este trabajo se presenta un método de detección de huellas de selección fundamentado en una matriz de parentescos genómicos. La ventaja de este enfoque radica en que la matriz de parentescos genómicos contempla las relaciones de parentesco entre individuos desde su construcción y, en consecuencia, permitiría relajar el supuesto de muestreo aleatorio de individuos. Más específicamente, la propuesta se basa en calcular una norma matricial, denominada *radio espectral*, a partir de la matriz de parentescos genómicos de una región previsiblemente sujeta a un proceso de selección.

MATERIAL Y MÉTODOS

Considérese una mutación favorable que experimenta una fuerte presión de selección, aumentando su frecuencia hasta reemplazar por completo el alelo ancestral, un proceso denominado barrido alélico (*'selective sweep'*, en inglés). Desde el punto de vista de la genética de poblaciones, un barrido alélico produce un patrón de variación en torno a la posición seleccionada diferente al esperado bajo neutralidad (Walsh y Lynch, 2011). Los métodos de detección de huellas de selección explotan estas diferencias para poner a prueba la hipótesis de neutralidad a través de estadísticos cuyo valor es particularmente sensible a los patrones de variación observados.

En este trabajo nos concentramos en los efectos de un barrido alélico sobre la coancestría molecular en torno a la región seleccionada. En general, las regiones genómicas muestran gran variación en términos de parentesco. Si bien en parte esto es consecuencia de la naturaleza estocástica de los mecanismos hereditarios, las fuerzas evolutivas que afectan a los individuos de una población también contribuyen al fenómeno (Weir et al., 2005). En particular, un barrido alélico aumenta la coancestría molecular en la región de influencia respecto a un proceso neutral porque el haplotipo que contiene la mutación favorable se halla más representado en la población y, en consecuencia, es más probable que dos alelos tomados al azar resulten idénticos por descendencia. Argumentando a la inversa, un 'exceso' de parentesco en una región específica del genoma puede interpretarse como una huella de selección. Para dos individuos en la población, la coancestría molecular se define como la probabilidad de que dos alelos tomados al azar, uno de cada uno de ellos, sean idénticos en estado (Toro et al., 2011). Este concepto se aplica a un locus en particular, pero puede extenderse a una región de interés promediando los valores de los marcadores que hay en ella. En este estudio, los valores de coancestría molecular obtenidos para cada par de individuos en la población se arreglaron en una matriz de coancestría molecular.

Para comparar diferentes matrices de coancestría molecular es conveniente recurrir a un estadístico que resuma en un único valor la información contenida. Las normas matriciales constituyen una alternativa ideal para este propósito, dado que extienden el concepto de

distancia vectorial a espacios multidimensionales (Golub y Van Loan, 1996). El estadístico elegido para este estudio fue el denominado radio espectral, cuyo valor se corresponde con la norma-2 en el caso de matrices simétricas. El radio espectral es equivalente al máximo valor propio, en términos absolutos, de una matriz. Además de su facilidad de cómputo, una propiedad importante de este estadístico es que su valor es menor o igual a cualquier otra norma de la misma matriz.

Para evaluar la potencia estadística de un test de neutralidad basado en el radio espectral de la matriz de coancestría molecular en una región de interés se desarrolló un experimento de simulación estocástica. Primero, se simuló una genealogía de secuencias de marcadores moleculares bialélicos en una población. Una vez alcanzado el equilibrio mutación-deriva, se seleccionó a favor de una mutación causal durante diez generaciones. Para construir la hipótesis nula, también se replicó la genealogía sin selección. La potencia se calculó superponiendo las distribuciones empíricas del estadístico en ambas situaciones (selección vs neutralidad) bajo diferentes escenarios poblacionales, que se obtuvieron modificando los siguientes parámetros: intensidad de selección, densidad de SNPs, frecuencia del alelo derivado al comienzo de la selección y tamaño de muestra (Tabla 1). Por último, los resultados fueron comparados contra dos métodos clásicos de detección de huellas de selección: la prueba de Tajima (1989) y el estadístico rEHH (del inglés 'relative extended haplotype homozygosity', Sabeti et al., 2002). Para llevar adelante los cálculos se escribieron programas específicos en Fortran 90 y R (R Core Team, 2014).

RESULTADOS Y DISCUSIÓN

Los resultados de la prueba de potencia se presentan en la Figura 1. Cada barra representa la potencia estadística del método de detección de huellas de selección, calculada a partir de la distribución del estadístico en 5.000 réplicas de la genealogía. En el caso del estadístico rEHH, fue necesario determinar previamente el marcador *base* ('core', Sabeti et al., 2002). Para ello, se seleccionó aquel SNP de la región que maximizaba la potencia estadística en un análisis de ventanas deslizantes. En términos generales, la potencia estadística de la prueba de Tajima (1989) fue baja para todos los escenarios simulados (en torno a 0,2 – 0,4), al menos para el tamaño de muestra de referencia (50 individuos). Resultados similares fueron reportados por Simianer et al. (2014) en un estudio comparativo de diferentes métodos de detección de huellas de selección aplicados a ganado doméstico. Bajo un enfoque de coalescencia, Simonsen et al. (1995) observaron que la prueba de Tajima sólo tiene potencia estadística en una ventana acotada de tiempo, que seguramente no captura el escenario de referencia. Por su parte, la prueba de hipótesis basada en el estadístico rEHH mostró mayor potencia estadística en general, con valores por encima de 0,6 para la mayoría de los escenarios simulados. Este método evalúa la caída del desequilibrio de ligamiento en la región circundante al marcador (o haplotipo) bajo selección (Sabeti et al., 2002). Los métodos basados en este enfoque también manifestaron una mayor potencia estadística en el estudio comparativo de Simianer (2014).

El método alternativo propuesto en este trabajo, basado en el radio espectral de la matriz de coancestría molecular construida a partir de un conjunto de marcadores moleculares en torno a la región de interés, superó en potencia a la prueba de Tajima (1989) en todos los escenarios simulados. En cambio, la potencia estadística sólo fue mayor a la de la prueba rEHH en situaciones puntuales, especialmente con una baja frecuencia de la mutación favorecida al comienzo de la selección y una fuerte presión de selección subsiguiente. Se concluye que el método propuesto es comparativamente útil para detectar un barrido alélico, particularmente cuando éste es intenso.

Tabla 1. Parámetros y rango de los valores utilizados para representar diferentes escenarios poblacionales. Valores de referencia en negrilla.

Parámetro	Rango de valores
Intensidad de selección (definida a través de coeficientes en contra del genotipo 11/12/22)	Débil- (1,0/0,8/0,8); Débil+ (1,0/0,5/0,5); Fuerte- (1,0/0,5/0,2); Fuerte+ (1,0/0,5/0,0)
Densidad de marcadores (SNP × cM ⁻¹)	200; 400 ; 1000
Rango de frecuencia del alelo seleccionado	0,1–0,2 ; 0,2–0,3; 0,4–0,5; 0,5–0,6; 0,6–0,7
Tamaño de muestra (2 × tamaño efectivo)	100 ; 200; 400

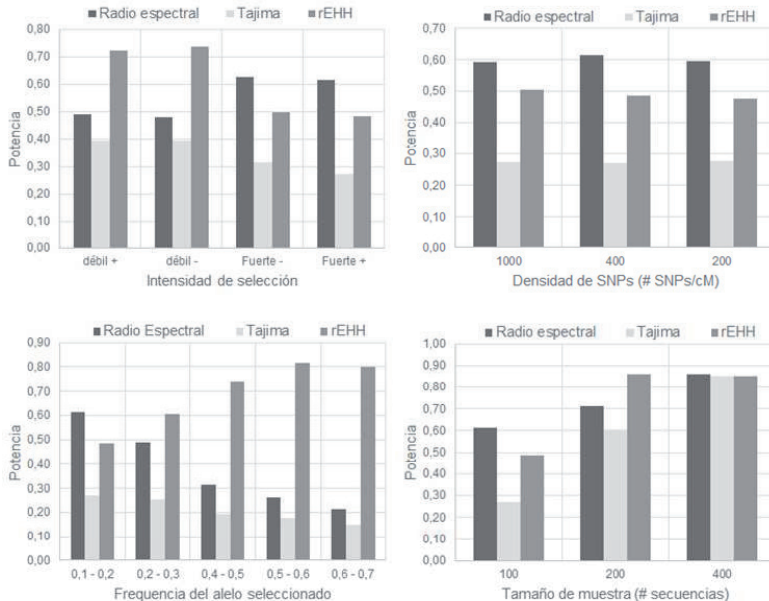


Figura 1. Potencia estadística de los métodos de detección de huellas de selección bajo los diferentes escenarios poblacionales ensayados.

REFERENCIAS BIBLIOGRÁFICAS

• Dekkers & Hospital. 2002. *Nature Rev. Genet.* 3: 22–32 • Golub & Van Loan. 1996. *Matrix Computations*. 3ra edición • R Core Team. 2014. URL: <http://www.R-project.org/> • Sabeti et al. 2002. *Nature* 419: 832–837 • Simianer et al. 2014. Proc. 10th WCGALP • Simonsen et al. 1995. *Genetics* 141: 413–429 • Tajima, F. 1989. *Genetics*. 123: 585–95 • Toro et al. 2011. *Genet. Sel. Evol.* 43:27 • Walsh & Lynch. 2011. *Evolution and Selection of Quantitative Traits*. Draft. • Weir et al. 2005. *Genome Res.* 15:1468–76.

Agradecimientos: S. Munilla agradece el financiamiento recibido con cargo al proyecto FITE 2012/2013 para el grupo DGA A51.

DETECTING SIGNATURES OF SELECTION BY MEANS OF THE SPECTRAL RADIUS OF THE MOLECULAR COANCESTRY MATRIX

ABSTRACT: In this study we introduce the spectral radius of the molecular coancestry matrix, computed for an enclosed set of contiguous markers, as a selection signature test. We emphasize that this test is useful when the sampled individuals are related (a typical situation in animal breeding) as compared to other tests that rest on the assumption that sampled sequences are randomly drawn. This is because the genomic relationship matrix intrinsically accounts for relatedness by construction. We explore the statistical power of the spectral radius test versus other standard neutrality tests through a stochastic simulation study under several population scenarios. We found that the test based in the spectral radius outperformed the Tajima’s test in all the scenarios assayed. In turn, the power was only greater than the one obtained with the rEHH statistic for certain specific situations, noticeably when the frequency of the causal mutation at the start of the selection is low and subsequent selection intensity is strong. We conclude that the method is comparatively useful to detect intense selective sweeps.

Keywords: hard sweep, molecular coancestry, matrix norms, statistical power.