AVANCES EN LA INFERENCIA DE LA ESTRUCTURA GENÉTICA POBLACIONAL EN PRESENCIA DE INDIVIDUOS EMPARENTADOS

S. T. Rodríguez-Ramilo[1,3], M. A. Toro[2] y J. Fernández[3]
[1]CONAFE. Ctra. de Andalucía, Km. 23,6. 28340, Madrid
[2]Dpto. de Producción Animal. ETSI Agrónomos. UPM. 28040, Madrid
[3]Dpto. de Mejora Genética Animal. INIA. Ctra. La Coruña Km. 7,5. 28040, Madrid
E-mail: jmj@inia.es

INTRODUCCIÓN

Los algoritmos de agrupamiento Bayesianos (Pritchard et al., 2000; Corander et al., 2004) son una herramienta computacional muy empleada para inferir estructura genética poblacional. Básicamente, estas metodologías Bayesianas actúan minimizando los desequilibrios de Hardy-Weinberg y de ligamiento dentro de cada subpoblación. Como consecuencia de esta característica, la mayoría de estas metodologías asumen implícitamente que los individuos muestreados en las distintas subpoblaciones no están emparentados. Sin embargo, especialmente en poblaciones con un censo reducido o especies con una elevada fecundidad, es muy probable que existan (y se muestreen) individuos emparentados en una misma subpoblación. Por lo tanto, las asunciones de equilibrio de Hardy-Weinberg y de ligamiento en las metodologías de agrupamiento Bayesianas no se cumplirán, lo que puede reducir la precisión de estos métodos a la hora de inferir la estructura genética poblacional (Anderson y Dunham, 2008; Rodríguez-Ramilo y Wang, 2012).

En el presente estudio se comparó mediante simulación por ordenador el efecto que produce la existencia de individuos emparentados sobre la precisión de dos metodologías para inferir estructura genética poblacional. El primer método evaluado fue el algoritmo Bayesiano implementado en el programa STRUCTURE (Pritchard et al., 2000), y la segunda metodología evaluada está basada en la maximización de la distancia genética entre subpoblaciones (Maximisation of Genetic Distance; MGD) y, por tanto, no asume ni equilibrio de ligamiento ni de Hardy-Weinberg (Rodríguez-Ramilo et al., 2009).

MATERIAL Y MÉTODOS

Datos simulados

Se simularon metapoblaciones compuestas por n = 3 o 5 subpoblaciones de 50 individuos cada una. Se incluyeron 10 (20) marcadores de tipo microsatélite (con 20 alelos) o 100 (200) SNP bialélicos. Las frecuencias alélicas de la metapoblación se generaron asumiendo una distribución Dirichlet. A partir de estas frecuencias y el coeficiente de diferenciación (F_{ST}) deseado (se simularon grados de diferenciación de F_{ST} = 0,1 y 0,2), las frecuencias alélicas de cada subpoblación también se generaron a partir de una distribución Dirichlet. Para crear una estructura de parentescos se realizó una generación previa en la que los genotipos de los individuos con una relación de parentesco determinada (hermanos, medios hermanos, y no emparentados) se generaron para cada marcador siguiendo las reglas de transmisión mendeliana. Se simularon casos con sólo individuos no emparentados, con 1 o 2 familias en la misma subpoblación o 2 familias en subpoblaciones diferentes. El tamaño familiar fue de 4 y 16 individuos. En el caso de familias de medios hermanos solamente se simuló el escenario con 10 microsatélites. Se realizaron 20 réplicas de cada combinación de parámetros.

Algoritmos evaluados

Se evaluó la proporción de réplicas en las que STRUCTURE (Pritchard et al., 2000) y MGD (Rodríguez-Ramilo et al., 2009) identificaron el número correcto de subpoblaciones (3 o 5).

Para la evaluación del método implementado en el STRUCTURE se empleó la proporción de ascendencia (*Q*) de cada individuo *i* (*i* = 1,...,*N*) perteneciente a cada cluster *j* (*j* = 1,...,*K*). Para cada individuo *i*, el cluster con el mayor valor de *Q* es el cluster al que ese individuo pertenece. La media de los valores Q_{ij} para todos los individuos *i* pertenecientes al cluster *j* se denomina $\tilde{Q}^{(j)}$, y el menor valor de $\tilde{Q}^{(j)}$ a través de los clusters evaluados es $\tilde{Q}^{(smc)}$. El número de clústeres inferido fue el mayor valor de *K* en el que $\tilde{Q}^{(smc)}$ > 0,8. Los parámetros empleados para ejecutar el STRUCTURE fueron 5.000 muestras de *burn-in* y 10.000 usadas para calcular la distribución. Se empleó el modelo *admixture* y la opción de frecuencias alélicas correlacionadas. En todos los demás parámetros se dejaron los valores por defecto. El rango de *K*s evaluado fue desde dos hasta el número real de subpoblaciones más uno.

El método MGD implementa un algoritmo de *simulated annealing* para la optimización. Se usaron como parámetros 10.000 soluciones alternativas por temperatura (*T*) y un máximo de 250 diferentes valores de *T*. La tasa de descenso de la temperatura y la temperatura inicial fueron 0,9 y 0,00001, respectivamente. Para cada escenario, el rango de *K*s evaluado fue desde dos hasta seis y el número de *K* inferido se estimó mediante una aproximación similar a la propuesta por Evanno et al. (2005) pero adaptada para las distancias genéticas (ver Rodríguez-Ramilo et al., 2009 para más detalles).

RESULTADOS Y DISCUSIÓN

La Figura 1 muestra la proporción de réplicas en las que el STRUCTURE (columna izquierda) y MGD (columna derecha) infieren *K* = 3 cuando *n* = 3 en presencia de familias de hermanos, los dos coeficientes de diferenciación considerados y usando 10 o 20 microsatélites. La proporción de *K* correctos es elevada en ambas metodologías en situaciones en las que no existen individuos emparentados y en casos con familias de 4 hermanos. Sin embargo, en las situaciones con 16 hermanos por familia, la proporción de réplicas en las que se estima *K* correctamente es mayor en MGD que con el STRUCTURE. En las dos metodologías la precisión con 20 microsatélites es mayor que con 10 microsatélites. Además, con un coeficiente de diferenciación elevado la precisión de ambos métodos mejora.

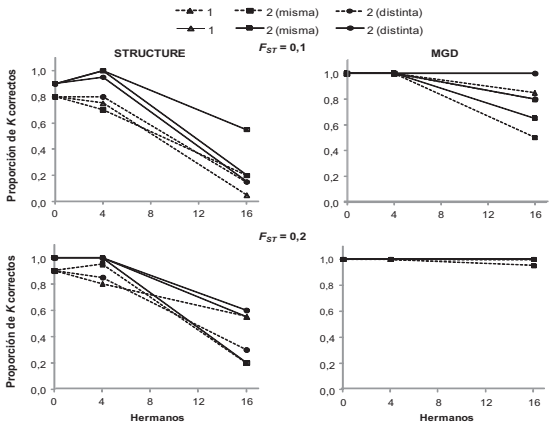

Figura 1. *Proporción de réplicas en las que el STRUCTURE (columna izquierda) y MGD (columna derecha) infieren K = 3 cuando n = 3 en el caso de hermanos y distintos niveles de diferenciación. Las líneas discontinuas indican 10 microsatélites. Las líneas continuas indican 20 microsatélites. Los triángulos representan una familia, los cuadrados indican dos familias en la misma subpoblación, y los círculos representan dos familias en subpoblaciones diferentes.*

La Figura 2 muestra la proporción de réplicas en las que el STRUCTURE (columna izquierda) y MGD (columna derecha) infieren *K* = 3 cuando *n* = 3 en los casos con familias

de medios hermanos con distintos coeficientes de diferenciación y 10 microsatélites. En general, el comportamiento de ambas metodologías es muy similar al caso de las familias de hermanos, indicando que la metodología MGD no reduce apreciablemente su precisión en presencia de individuos emparentados. Además, los resultados obtenidos con 100 y 200 SNPs y cinco subpoblaciones fueron muy similares a los anteriores (datos no mostrados).

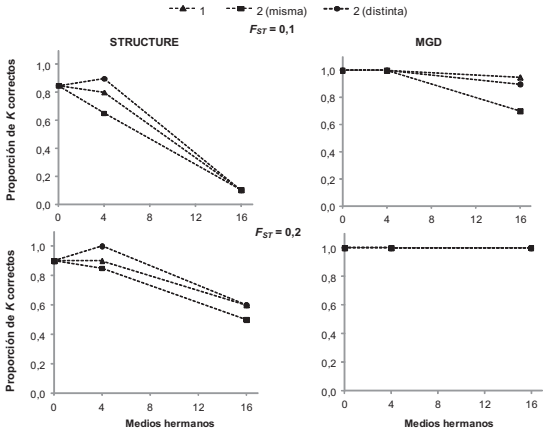

Figura 2. *Proporción de réplicas en las que el STRUCTURE (columna izquierda) y MGD (columna derecha) infieren K = 3 cuando n = 3 en el caso de medios hermanos y distintos niveles de diferenciación. Las líneas discontinuas indican 10 microsatélites. Los triángulos representan una familia, los cuadrados indican dos familias en la misma subpoblación, y los círculos representan dos familias en subpoblaciones diferentes.*

En resumen, los resultados indican que una metodología que no asume ni equilibrio de Hardy-Weinberg ni de ligamiento es más precisa a la hora de inferir estructura genética poblacional en presencia de individuos emparentados que una aproximación Bayesiana que si los asume. Por tanto, si se sospecha de la existencia de individuos emparentados en una muestra, sería conveniente emplear una metodología del primer tipo a la hora de inferir estructura genética poblacional.

REFERENCIAS BIBLIOGRÁFICAS

• Anderson, E.C. & Dunham K.K. 2008. Mol. Ecol. Res. 8: 1219-1229.• Corander, J., Waldmann, P., Marttinen, P. & Sillanpaa, M.J. 2004. Bioinformatics 20: 2363-2369.• Evanno, G., Regnaut, S. & Goudet, J. 2005. Mol. Ecol. 14: 2611-2620.• Pritchard, J.K., Stephens, M. & Donnelly, P. 2000. Genetics 155: 945-959.• Rodríguez-Ramilo, S.T., Toro, M.A. & Fernández, J. 2009. Genet. Sel. Evol. 41: 49.• Rodríguez-Ramilo, S.T. & Wang, J. 2012. Mol. Ecol. Res. 12: 873-884.

ADVANCES ON THE INFERENCE OF POPULATION GENETIC STRUCTURE IN PRESENCE OF RELATED INDIVIDUALS

ABSTRACT: This study aims to compare two methodologies for the inference of population genetic structure in presence of related individuals. The first method implements a Bayesian approach to minimise Hardy-Weinberg and linkage equilibrium within subpopulations. The second methodology maximises genetic distance between subpopulations and does not make Hardy-Weinberg and linkage equilibrium assumptions. Using simulated data, the results indicate that the second approach is less influenced by the presence of close relatives, and is more appropriated when close relatives are supposed to be present in a sample.

Keywords: population structure, related individuals, molecular markers.