BAG Journal of Basic and Applied Genetics
Vol XXVII (2). December 2016. 49-58.

STATISTICAL MODELS FOR PHENOTYPE-GENOTYPE ASSOCIATION STUDIES IN GENETICALLY STRUCTURED POPULATIONS

MODELOS ESTADÍSTICOS PARA ESTUDIOS DE ASOCIACIÓN FENOTIPO-GENOTIPO EN POBLACIONES GENÉTICAMENTE ESTRUCTURADAS

Peña Malavera A.[1], Gutierrez L.[2], Balzarini M.[1]*

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Estadística y Biometría, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba. Av. Valparaíso s/n, Ciudad Universitaria, CP: 5000 (509) Córdoba, Argentina.
[2] Department of Agronomy, University of Wisconsin at Madison, 1575 Linden Dr. Madison, WI 53706.

*Corresponding author: mbalzari@agro.unc.edu.ar

ABSTRACT

Association mapping is used to find specific regions in the genome related to changes in a phenotypic trait. However, it has been found that in genetically structured populations, the number of false positives increases. The aim of this study was to compare the performance of several association mapping statistical models that take into account the underlying population genetic structure. Different statistical strategies developed under the mixed model theory were evaluated. The compared association models included the following matrices to model genetic structure: Q-matrix (probability of membership of each individual to each subpopulation), P-matrix (principal components of marker data capturing the structure variance) and K-matrix (containing genetic relationships between the individuals of the mapping population). The columns of Q-matrix and P-matrix were used in the associative mapping model as fixed effect covariates as well as random effect covariates. We also evaluated models including simultaneously Q-matrix and K-matrix, or either as P-matrix and K-matrix. The reference model (naïve model) was a regression model that did not account for genetic structure. Model comparison criteria were the empirical distributions of p-values, the FDR (False Discovery Rate) and the statistical power. The results suggest that the use of the K-matrix, alone or together with the Q-matrix reduced the false positive rate regardless of the level of genetic divergence among underlying subpopulations.

Key words: linear mixed models, population genetic structure, false discovery rate.

RESUMEN

El mapeo asociativo (MA) es usado para encontrar regiones específicas del genoma relacionadas con la variación de un carácter fenotípico. Sin embargo, se ha detectado que en poblaciones con estructura genética poblacional (EGP), la cantidad de falsos positivos en la asociación fenotipo-genotipo aumenta. El objetivo de este trabajo fue evaluar el desempeño de modelos de MA que consideran la EGP mediante distintas estrategias desarrolladas bajo la teoría de los modelos mixtos. Se evaluaron modelos de regresión fenotipo-genotipo incluyendo las siguientes matrices para modelar EGP: matriz Q (probabilidad de pertenencia de cada individuo a cada subpoblación), matriz P (componentes principales de los datos de marcadores), matriz K o de parentesco genético entre las líneas de la población de mapeo. Las columnas de las matrices Q y P fueron usadas en el modelo de MA como covariables de efecto fijo y alternativamente, como efectos aleatorios. También se evaluaron modelos incluyendo simultáneamente las matrices Q y K, así como P y K. El modelo de referencia ("*naive*") fue el modelo de regresión que no contempló EGP. Los criterios de comparación de modelos fueron la función de distribución empírica de valores-p, la tasa FDR (*False Discovery Rate*) y la potencia estadística. Los resultados sugieren que el uso de la matriz K, sola o junto con la matriz Q, fue la estrategia de mayor impacto para disminuir la tasa de detección de falsas asociaciones. Esto se observó independientemente del nivel de divergencia genética, entre las subpoblaciones que constituían la población de mapeo.

Palabras clave: modelos lineales mixtos, estructura genética poblacional, tasa de falsos positivos.

Fecha de recepción: 22/03/2015
Fecha de aceptación de versión final: 18/05/2016

INTRODUCCIÓN

En los últimos años se ha incrementado el uso del mapeo asociativo (MA) o mapeo por desequilibrio de ligamiento (LD, del inglés *Linkage disequilibrium*) para la identificación de regiones del genoma responsables de características complejas de interés agronómico. La técnica ha sido ampliamente adoptada en el mejoramiento de especies vegetales para el análisis de los *loci* de caracteres cuantitativos o QTL (Aranzana *et al.*, 2005; Breseghello y Sorrells, 2006; D'hoop *et al.*, 2008; Kraakman *et al.*, 2006; Remington *et al.*, 2001; Stich *et al.*, 2008; Thornsberry *et al.*, 2001; Zhu *et al.*, 2008). Cuando la población de individuos empleada en el análisis de mapeo por LD está estructurada genéticamente, aumenta la cantidad de falsos positivos en la detección de las asociaciones de interés (Malosetti *et al.*, 2007). Esto ocurre porque en una población genética con sub-poblaciones, cualquier carácter presente con mayor frecuencia en una de ellas mostrará asociación positiva con alelos que son más comunes en esta sub-población (Zhang *et al.*, 2010). Consecuentemente, es posible que se detecten marcadores asociados con la composición de la población más que con la característica de interés (Yu *et al.*, 2006). Por ello, se han propuesto distintas estrategias de modelado para los estudios de asociación fenotipo-genotipo, todas tendientes a controlar el aumento en la detección de asociaciones espurias. El modelo de MA básico es un modelo de regresión lineal, donde el fenotipo se asocia al genotipo (marcadores moleculares) mediante coeficientes de regresión. Este modelo básico es luego extendido con el fin de incorporar factores o covariables que representan la estructura genética subyacente en la población de mapeo (Cappa *et al.*, 2013; Muñoz-Amatriaín *et al.*, 2014; Wang *et al.*, 2012). Cuando se usa el programa *Structure* (Pritchard *et al.*, 2000) como herramienta para detectar estructura genetica poblacional (EGP), se obtienen las probabilidades de pertenencia de cada individuo a las sub-poblaciones que componen la población y esta información puede ser incorporada al modelo de MA (Gutiérrez *et al.*, 2011; Yu *et al.*, 2006). Las columnas de la matriz Q resultante del análisis con el programa *Structure* (probabilidades de pertenencia a los distintos grupos o conglomerados) suelen ser usadas como covariables en el modelo de regresión fenotipo-genotipo. Estas covariables proveen información que dimensiona la EGP. Alternativamente se suele usar otra matriz, conocida como matriz P, compuesta por los componentes principales resultantes del análisis de componentes principales (ACP) (Hotelling, 1936) realizado con la matriz de datos de marcadores genéticos (Price *et al.*, 2006). Los componentes principales (CP) significativos, según la prueba de Tracy-Widom (1994), o los primeros CP (los que explican mayor porcentaje de la variabilidad total de los datos moleculares, pueden también ser usados como covariables en el modelo de MA (Peña Malavera *et al.*, 2014). Las covariables que resumen la estructuración genética de los genotipos de la población de mapeo pueden contemplarse en el modelo como efectos fijos o aleatorios. Malosetti *et al.* (2007) recomiendan incluir la estructura genética subyacente en las poblaciones de mapeo como efecto aleatorio, independientemente del procedimiento utilizado para detectar dicha estructura. Otra estrategia ya difundida para contemplar relaciones genéticas durante el MA es la corrección de errores estándares asociados a los coeficientes de regresión mediante la incorporación de la matriz de parentesco genético (matriz K) en la modelación de la estructura de covarianza residual. La matriz K puede ser obtenida con la librería EMMA del programa R (www.Rproject.org) (Kang *et al.*, 2008) como la matriz de covarianza asociada a un vector aleatorio de efectos poligénicos adicionados a cada genotipo en estudio. Usualmente, la matriz K es una matriz de similitudes entre los perfiles moleculares individuales de distintos genotipos. Cualquiera sea la estrategia seleccionada para contemplar la EGP, matriz Q, P o K, la estimación del modelo puede realizarse bajo el marco teórico de los modelos lineales mixtos (MLM) (Peña Malavera, 2015). Sin embargo, poco se ha investigado sobre el impacto de estos modelos alternativos respecto a la detección de QTL en situaciones donde existan bajos o casi nulos niveles de ligamiento como sucede en colecciones de germoplasma. El objetivo de este trabajo fue comparar el desempeño, a nivel de las tasas de falsos positivos y también a nivel de potencia estadística para la detección de QTL, de modelos de MA alternativos bajo escenarios de baja y alta estructuración genética. También, se comparan resultados obtenidos en contextos con distinta cantidad de información con dos niveles de densidad de marcadores (baja y alta). Los modelos comparados fueron: modelos Q y QA (usan matriz Q como covariables de efecto fijo o aleatorio, respectivamente), modelos P y PA (usan matriz P como covariables de efecto fijo o aleatorio, respectivamente), modelo K, QK y PK; estos dos últimos incluyen simultáneamente dos matrices de control de EGP. En cada escenario también se ajustó como modelo de referencia, uno que no contempla corrección por EGP

(Modelo *naive*). La comparación de modelos se realizó usando bases de datos de marcadores moleculares simulados bajo dos escenarios, uno con 300 marcadores moleculares y tamaño poblacional de 150 genotipos los cuales contenían distintos niveles de EGP, y otro con una base de datos experimentales con 511 marcadores moleculares y 504 genotipos de maíz genéticamente estructurados (Hansey *et al.*, 2011).

MATERIALES Y MÉTODOS

Simulaciones

Los datos de marcadores moleculares fueron simulados usando QMSim (Sargolzaei y Schenkel, 2009) dando origen a dos escenarios con cantidad de genotipos y de marcadores moleculares que imitan datos usuales en mejoramiento genético vegetal. Se simuló un genoma con 300 marcadores multilocus-bialélicos, con diseño de cruzamientos y selección aleatorios para una EGP conformada por cinco poblaciones. Se crearon dos escenarios correspondientes a dos niveles de divergencia genética entre poblaciones (bajo y alto F_{ST}), con un tamaño de población de mapeo (n= 150). Los datos simulados fueron creados a partir de una población de 200 individuos y el sistema de cruzamiento fue basado en la unión aleatoria de gametas por más de 10 generaciones. La coancestría promedio fue baja como sucede en numerosas poblaciones que se usan para MA en vegetales. Al variar el número de generaciones desde la población fundadora, se crearon diferentes niveles de divergencia genética poblacional. Los datos simulados para cada marcador fueron codificados como 0 y 1. El promedio del estadístico F_{ST} (Wright, 1951), provisto por el análisis molecular de la varianza (AMOVA) (Excoffier *et al.*, 2009), fue usado para cuantificar el grado de diferenciación genética entre poblaciones en cada escenario con bajo y alto F_{ST} (F_{ST}= 0,03 y 0,20, respectivamente).

Dada la matriz de marcadores moleculares resultante de QMSim, para cada uno de los dos escenarios, se escogieron aleatoriamente 20 marcadores y con ellos se realizó una combinación lineal con efectos que siguen una distribución $\Gamma(2,5)$ para simular el efecto de los *loci* ligados a un QTL, es decir aquellos con información para determinar el fenotipo. Adicionalmente, se anexó a cada perfil molecular la realización de una variable aleatoria con distribución normal de media 100 (para representar la media del carácter, la cual depende del efecto poligénico de *background*) y varianza 25 (para representar la variabilidad experimental, *i.e.* desvío estándar de 5 es decir no superior al 5 % de la media del carácter fenotípico). A esta variable $\sim N(100,25)$ se le sumaron los efectos de los marcadores ligados extraídos de la distribución gamma.

Datos experimentales

Los modelos de MA también se evaluaron en un conjunto de datos públicos conformados por n= 504 líneas de maíz genotipificadas con p= 511 marcadores del tipo SNPs (Hansey *et al.*, 2011). Hansey *et al.* (2011) identificaron ocho conglomerados o sub-poblaciones en ese conjunto de líneas de maíz, datos que fueron luego verificados por investigaciones genéticas en esa especie. Consecuentemente, los análisis sobre los datos genéticos reales se hicieron asumiendo la existencia de ocho conglomerados. Para implementar el MA se simuló para cada uno de estos genotipos un valor fenotípico adicionando una variable aleatoria $\sim N(100,25)$ y una variable aleatoria gamma $\Gamma(4,2)$, asociada a cada uno de 22 marcadores moleculares elegidos aleatoriamente.

Modelos estadísticos ajustados

Se estimaron ocho modelos de mapeo asociativo para evaluar el efecto del marcador sobre el fenotipo cuya denotación se presentan en la Tabla 1. Los modelos de MA comparados surgen de usar algunos o todos los términos de la siguiente ecuación:

$$y = X\beta + EGPv + Zu + e$$

donde y es el vector de valores fenotípicos (con un dato fenotípico por genotipo), X es el vector de datos para el marcador molecular, β es el efecto del marcador, EGP es la matriz de estructura genética (construida alternativamente como la matriz Q de la salida del programa *Structure* o la matriz P de los componentes principales estadísticamente significativos, ambos realizados previamente sobre los datos moleculares), v es el vector de efectos de la estructura poblacional (en algunas aproximaciones considerado como vector de efectos fijos y en otras como vector de efectos aleatorios), Z es la matriz de incidencia que conecta el vector aleatorio u de efectos poligénicos con los datos fenotípicos (matriz identidad de dimensión igual al número de genotipos que componen la población de mapeo) y e es un vector de términos de error aleatorio, normalmente distribuido con media cero y varianza constante σ_e^2

El vector *u* se distribuye independientemente del vector *e* y con matriz de varianzas y covarianzas dada por $\sigma_e^2 \times K$, siendo *K* la matriz de similitud entre los pares de perfiles moleculares derivadas de la librería EMMA (Kang *et al.*, 2008).

Tabla 1. Ocho modelos comparados en datos reales y simulados.

Matriz de Parentesco	Covariable de Estructura				
	No	Q fijo	ACP fijo	Q aleatorio	ACP aleatorio
No	naive	Q	P	QA	PA
Si	K	QK	PK	--	--

Q es la matriz de probabilidades de pertenencia a los grupos de la EGP, probabilidad calculada con el software *Structure*, P es la matriz de componentes principales retenida mediante el estadístico de Tracy-Widom (1994) y K es la matriz de parentesco propuesta por *Kang et al.* (2008).

Ajuste de modelos y criterios de comparación

Todos los modelos fueron ajustados usando *Info-Gen* (Balzarini y Di Rienzo, 2004) y su interfaz con R (Core Team, 2013). El desempeño de los modelos se evaluó usando como criterio las curvas de distribución acumulada de valores-p (Gutierrez *et al.*, 2011; Cappa *et al.*, 2013; Peña Malavera *et al.*, 2014; Peña Malavera, 2015), las tasas de falsos descubrimientos o FDR (del inglés, *False Discovery Rate*) (Benjamini y Hochberg, 1995) y los cálculos de potencia estadística.

Para construir las curvas de distribución de valores-p, se usó la función de distribución empírica de la variable "valor-p asociado a cada una de las pruebas de hipótesis realizadas en un escenario". En cada escenario hay tantas pruebas de hipótesis de asociación como marcadores. Es importante resaltar que en una distribución acumulada de valores-p se espera que si la modelación ha sido buena, la distribución se aproxime a una línea recta de 45 grados, ya que la distribución de los valores-p debe ser simétrica. Una distribución asimétrica hacia valores-p pequeños indica mayor significancia de la esperada, lo que sugiere un posible incremento de falsos positivos, es decir presencia de asociaciones espurias.

La tasa FDR se calculó en base a las proporciones de falsos positivos (FP) y verdaderos positivos (VP). Los FP son todos aquellos valores-p significativos vinculados a marcadores que no están asociados al fenotipo (no ligados a un QTL) y los VP son todos aquellos marcadores positivos que efectivamente están asociados al fenotipo (ligados a un QTL), de esta forma tenemos que:

$$FDR = \frac{FP}{VP + FP}$$

La potencia estadística en la detección de marcadores asociados con el fenotipo está referida a una medida de eficacia de los modelos y es la probabilidad de que la hipótesis nula H_0 sea rechazada cuando esta es falsa, o dicho de otra manera cuando la hipótesis alternativa H_a es verdadera. La potencia estadística (φ) puede interpretarse como la probabilidad de no cometer error del tipo II (error que producen los eventos conocidos como falsos negativos, FN). La potencia en este trabajo fue calculada de la siguiente manera:

$$\varphi = \frac{VP}{VP + FN}$$

RESULTADOS

Datos genéticos simulados

En la Figura 1 se muestran los gráficos de dispersión de los dos primeros ejes resultantes del escalamiento multidimensional métrico (Gower, 1967) obtenido desde los datos moleculares observados bajo cada uno de los dos escenarios con diferentes niveles de F_{ST}, con 300 marcadores moleculares. La figura proporciona información con respecto a la distancia genética entre los genotipos y el grupo al que estos genotipos fueron asignados. En el escenario I se observa baja divergencia genética mientras que en el escenario II, los grupos o subpoblaciones que estructuran la población se presentan más distanciados por la mayor divergencia genética con la que fueron simulados.

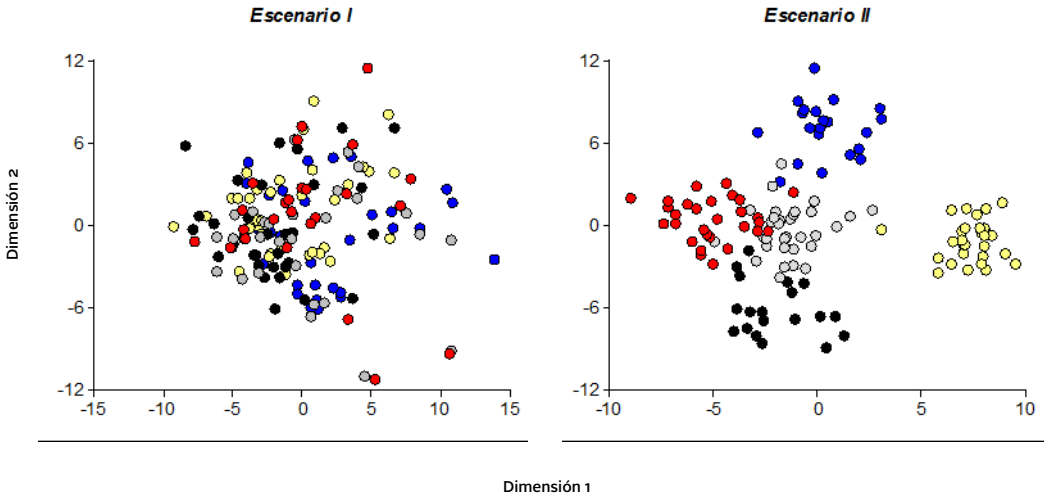

Figura 1. Gráficos de dispersión de los dos primeros ejes resultantes de un análisis de coordenadas principales (escalamiento multidimensional) de los datos moleculares (300 MM). En la columna de la izquierda bajo FST. y en la columna derecha alto FST. Los colores identifican los cinco grupos que definen la EGP.

Evaluación de modelos de mapeo asociativo

En la Figura 2 se muestran las funciones de distribución acumulada para los dos escenarios. Se puede ver que consistentemente en los dos escenarios con diferentes niveles de F_{ST}, los modelos con mejor ajuste son el modelo K y el modelo QK.

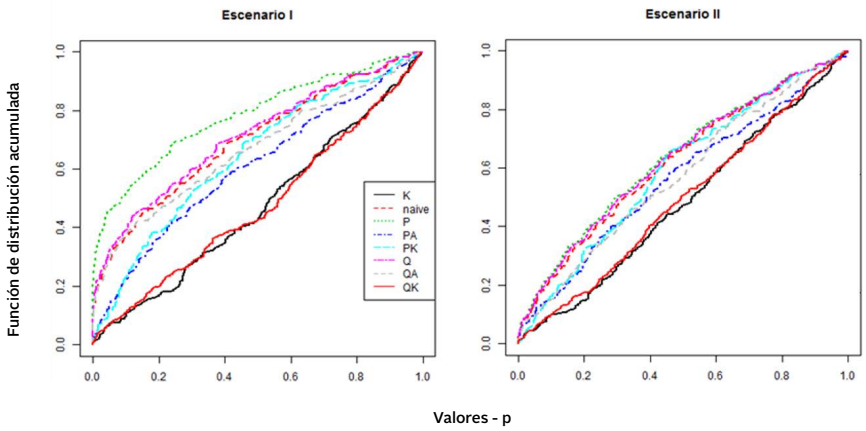

Figura 2. Gráfico de distribución acumulada de los valores-p de los ocho modelos evaluados en los escenarios simulados. En la columna de la izquierda escenarios con FST. bajo y en la columna derecha FST. alto.

Con bajo F_{ST} (izquierda, escenario I) se observa que el modelo de menor desempeño fue el modelo P (Figura 2). Los modelos se comportan de manera parecida cuando se trabaja en el contexto de alto F_{ST} (escenario II), situación en la que se observan menores diferencias entre los ajustes, principalmente para el caso de mayor estructuración relativa en la población, correspondiente a un valor de F_{ST} de 0,20 y a una población de 150 individuos.

Existen para el modelo P una cantidad mayor a la esperada de valores-p pequeños, es decir mayor probabilidad de detecciones falsas. Los modelos *naive* y Q también mostraron alta asimetría hacia pequeños valores-p. Con bajo nivel de EGP no se desempeñaron bien los modelos que incorporaron más parámetros para modelar su estructura de media (efectos fijos). Los modelos que incluyen la matriz K y los que incluyen covariables de efectos aleatorios para incorporar la estructura, fueron los de mejor desempeño. Con alto nivel de divergencia genética, QA se desempeñó mejor que PA mientras que con bajo nivel de divergencia la relación entre ambos fue inversa. Entre los modelos que usan la matriz K y que también incorporan covariables de efectos fijos, el modelo QK mostró menor FDR y una función de distribución acumulada de valores-p más cercana a la esperada bajo la hipótesis nula. Es importante notar que en ninguno de los escenarios simulados la estrategia de modelado conjunto mejoró el desempeño del modelo que sólo usa la matriz K en términos de valores-p.

En la Figura 3 se muestran las funciones de distribución acumulada de los valores-p correspondiente a las pruebas de hipótesis implementadas para el conjunto de datos experimentales de maíz (Hansey *et al.*, 2011). Se observa que los modelos, excepto por PK, tienen un comportamiento relativo similar. Todos muestran menor cantidad de la esperada de valores-p, comportándose similar al modelo que no implementa ninguna corrección por estructura (modelo *naive*). Aún cuando se ha reconocido previamente la existencia de ocho grupos en los datos moleculares, los distintos modelos de MA ajustados sobre estos datos no provocaron significativos cambios en los valores-p (a excepción de PK que sobredimensionó la cantidad de pequeños valores-p). La estrategia de modelación más diferente del modelo *naive*, en este dominio de valores-p, fue el modelo QK. Estos resultados podrían estar asociados al bajo nivel de divergencia genética existente entre las ocho subpoblaciones (F_{ST}= 0,02), o a una variabilidad relativamente alta de los datos fenotípicos que enmascara el efecto de los QTL.

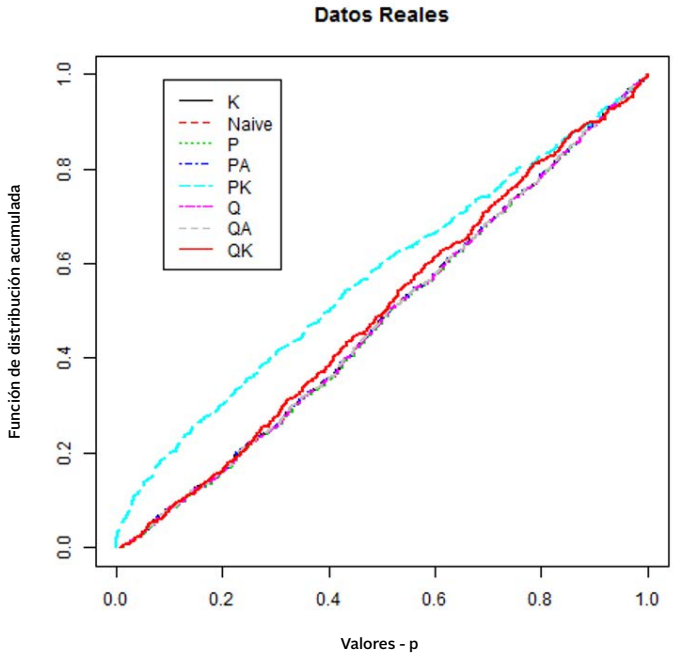

Figura 2. Gráfico de distribución acumulada de los valores-p de los ocho modelos evaluados en datos de 511 marcadores SNP sobre 504 genotipos de maíz estructurados genéticamente con un nivel de FST. relativamente bajo (FST.= 0,02).

FDR Y POTENCIA

Datos simulados

En la Tabla 2 se observa que los modelos que involucran la matriz K en el análisis, tienen los menores valores de FDR para los dos niveles de F_{ST}, mientras que la corrección por estructura con la matriz P incrementó el porcentaje de falsos descubrimientos de asociaciones con respecto al modelo *naive* (34 % *vs.* 30 %).

Como es de esperar por las relaciones teóricas existente entre los errores tipo I y tipo II de las pruebas de hipótesis estadísticas (Balzarini *et al.*, 2012), los métodos que mejor controlan FDR son los que más potencia pierden o los que tienen menor probabilidad de detectar mayor cantidad de QTL verdaderos. La introducción de la matriz K en el modelo de MA, redujo en aproximadamente un tercio para el caso de bajo F_{ST} y en un medio para el caso de mayor F_{ST}, la probabilidad de detecciones falsas.

En escenarios de alta divergencia genética, todas las aproximaciones metodológicas representaron una mejora a nivel de FDR respecto al modelo *naive*, algunos sin pérdida de potencia y otros (los que introducen la matriz K) con pérdida de potencia. No obstante, el modelo QK fue el de menor pérdida relativa de potencia si se considera la disminución de FDR que provocó.

Para los escenarios de bajo F_{ST}, el modelo K fue el que mejor se desempeñó, reduciendo significativamente los valores de FDR respecto al modelo *naive* con la consecuente pérdida de potencia, aunque ésta no fue mayor que para otros modelos con menor impacto sobre la FDR.

Tabla 2. Tasas de falsos positivos y potencia de ocho modelos de mapeo asociativo para dos niveles de estructura genética poblacional (Bajo y Alto FST.), 300 marcadores moleculares y n= 150.

Modelo	FDR		Potencia	
	Bajo F_{ST}	Alto F_{ST}	Bajo F_{ST}	Alto F_{ST}
Naive	0,30	0,24	0,60	0,55
Q	0,31	0,20	0,65	0,50
P	0,34	0,22	0,85	0,50
K	0,12	0,11	0,35	0,25
QK	0,13	0,11	0,35	0,30
PK	0,22	0,18	0,35	0,30
QA	0,32	0,18	0,55	0,35
PA	0,25	0,20	0,35	0,35

Naive: sin corrección por estructura; Q: con corrección mediante la matriz de probabilidades a *posteriori* obtenida con el software *Structure*; P: con corrección por CP como covariables de efectos fijos; K: con corrección por matriz de parentesco; QK: modelo mixto con Q como factor de efectos fijos y K factor de efectos aleatorios; PK: modelo mixto con P como factor de efectos fijos y K factor de efectos aleatorios; QA: modelo con la matriz Q como efectos aleatorios; PA: con la matriz P como covariables de efectos aleatorios.

Estructura genética real

La cantidad de componentes retenidos por Tracy-Widom para estos datos fue de 36; este alto número es necesario para explicar una EGP con bajo nivel de divergencia, pero a su vez produce una sobre-parametrización del modelo PK y quita grados de libertad a la estimación de la varianza residual con las consecuencias negativas observadas a nivel de la detección de QTL. Para los otros modelos se observa una sub-detección de QTL, siendo los modelos K, QK y los que incluyen la EGP en la porción aleatoria del modelo de MA, los de mejor desempeño. La baja divergencia genética de los datos de marcadores (F_{ST}= 0,02) podrían explicar estos resultados.

DISCUSIÓN

Se observó que todos los métodos, excepto Q y P, contribuyen a bajar la FDR. El uso de la matriz K fue la estrategia estadística de mayor impacto para bajar la FDR, tanto usada sola como con las correcciones por estructura genética Q y P. No hubo diferencia entre las tasas de FDR para alto y bajo F_{ST}, FDR es una característica más asociada al modelo de MA (parametrización de las estructuras de medias y varianzas que realiza cada modelo) que al nivel de estructura subyacente. Sin embargo se puede ver que el nivel de estructura impacta en la potencia del modelo. Con un nivel alto de estructura todos los modelos excepto P y Q, si bien controlaron la tasa de falsos positivos (FP) respecto al modelo *naive*, mostraron pérdida de potencia. El método PK que combina la matriz P proveniente del ACP y la matriz K de parentesco, fue el de menor pérdida de potencia en los escenarios simulados. No obstante, PK mostró en todos los escenarios menor capacidad de controlar FDR que K y QK. Aún, cuando en los escenarios de bajo F_{ST} los modelos P y Q mostraron más potencia para detectar QTL que el modelo sin corrección (*naive*), su incapacidad para reducir la tasa de FP no los hace recomendables.

Las funciones de distribución de los valores-p estimadas en este trabajo sugieren que los modelos K y QK son los de mejor desempeño en la identificación de asociaciones fenotipo-genotipo cuando se consideran tanto las tasas de error tipo I como las tasa de error tipo II. Este resultado coincide con los presentados por Yu *et al.* (2006) quienes trabajaron con tres variables fenotípicas medidas en 277 líneas endocriadas de maíz genotipadas con 553 SNP en un escenario de estructura genética media a alta. El modelo QK fue también el elegido para el MA realizado con los datos moleculares reales considerados en este trabajo por Hansey *et al.* (2011). El modelo QK propuesto por Yu *et al.* (2006) fue inicialmente usado en poblaciones humanas y en especies alógamas. No obstante, Stich y Mechinger (2009) probaron este modelo en especies con distintos sistemas de reproducción usando colecciones de germoplasma no sólo de maíz sino también de papa, remolacha, nabo y arabidopsis, concluyendo que el modelo QK había resultado apropiado para todas las especies evaluadas.

Otros estudios han mostrado, al igual que los resultados observados en el presente trabajo, que en contextos de bajo LD y escasa estructuración genética, la incorporación de los CP como efecto fijo (P) puede producir una sobre-parametrización del modelo que conduce a un incremento en la tasa de FP (Peña Malavera *et al.*, 2014). En nuestro estudio, al igual que lo informado por Malosetti *et al.* (2007) y Gutiérrez *et al.* (2011), el modelo P (modelo de efectos fijos de CP que representan estructura) produjeron un gran número de falsos positivos o asociaciones espurias. Este resultado debería ser frecuente en situaciones donde la estructura poblacional es de poca magnitud y el número de componentes necesario para resumirla, de manera estadísticamente significativa, resulta alto. En tal contexto, el modelo lineal a estimar usa una cantidad alta de parámetros para describir poca variabilidad entre los genotipos moleculares. Sin embargo, Cappa *et al.* (2013) al estudiar el comportamiento de diversos modelos de MA en poblaciones de eucalipto de Argentina y Uruguay concluyeron que, para cuatro de los seis caracteres estudiados, el modelo P resultó una estrategia buena para controlar la marcada estructura familiar que existía entre los genotipos. Es importante notar que la EGP era alta y fue bien identificada en el trabajo con eucaliptus; tanto con *Structure* como con ACP donde los dos primeros componentes principales explicaron cerca del 50% de la variabilidad de los perfiles moleculares. Sólo con dos CP se pudo ordenar claramente los genotipos en tres grupos esperados según el conocimiento previo de sus procedencias. Con alto nivel de EGP y pocos CP, el modelo P puede resultar una buena estrategia para la detección de QTL verdaderos.

La incorporación de los componentes principales como covariables de efectos aleatorios (modelo PA) disminuyó también la tasa de falsos positivos respecto a los modelos

naive y P, pero con menor pérdida de potencia que K y QK. Gutiérrez *et al.* (2011) probaron métodos de corrección de la estructura poblacional, incluyendo entre ellos al modelo PA encontrando comportamientos similares entre este modelo y el modelo Q que usaba como covariables las columnas de salida del programa *Structure* (Pritchard *et al.*, 2000). En este trabajo, el modelo PA siempre controló mejor la tasa FDR que el modelo Q. Es de destacar que la implementación del control por estructura con análisis de componentes principales es significativamente más eficiente en tiempo computacional que el uso de información resultante del programa *Structure*. Wang *et al.* (2012) también mostraron la efectividad del modelo con componentes principales aleatorias (PA) para contemplar la estructura poblacional previa al MA y disminuir los FP.

BIBLIOGRAFÍA

Aranzana M.J., Kim S., Zhao K., Bakker E., Horton M., Jakob K., Lister C., Molitor J., Shindo C., Tang C., Toomajian C., Traw B., Zheng H., Bergelson J., Dean C., Marjoram P., Nordborg M. (2005) Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 1 (5): e60.

Balzarini M., Di Rienzo J. (2004) Info-Gen: Software estadístico para análisis de datos genéticos. Universidad Nacional de Cordoba, Córdoba.

Balzarini M., Di Rienzo J., Tablada M., Gonzalez L., Bruno C., Córdoba M., Robledo W., Casanoves F. (2012) Estadística y Biometría. Córdoba: Encuentro Grupo de Editores. pag.380. ISBN 978-987-591- 301-1

Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57: 289-300.

Breseghello F., Sorrells M. (2006) Association mapping of kernel size and milling quality in wheat (*Triticum aestivum* L.) cultivars. Genetics 172: 1165-1177.

Cappa E.P., El-Kassaby Y.A., Garcia M.N., Acuña C., Borralho N.M.G., Grattapaglia D., Marcucci Poltri S. (2013) Impacts of population structure and analytical models in genome-wide association studies of complex traits in forest trees: A Case Study in *Eucalyptus globulus*. PLoS ONE 8 (11): e81267.

Core Team (2013) R: A language and environment for statistical computing. Computing (Editor), Vienna, Austria.

D'hoop B., Paulo M., Mank R., Eck H., Eeuwijk F. (2008) Association mapping of quality traits in potato (*Solanum tuberosum* L.). Euphytica 161(1-2): 47-60.

Excoffier L., Hofer T., Foll M. (2009) Detecting *loci* under selection in a hierarchically structured population. Heredity 103 (4): 285-298.

Gower J.C. (1967) Multivariate analysis and multidimensional geometry. Journal of the Royal Statistical Society, Series D (The Statistician) 17 (1): 13-28.

Gutiérrez L., Cuesta-Marcos A., Castro A., von Zitzewitz J., Schmitt M., Hayes P. (2011) Association mapping of malting quality quantitative trait *loci* in winter barley: positive signals from small germplasm arrays. Plant Gen. 4 (3): 256-272.

Hansey C.N., Johnson J.M., Sekhon R.S., Kaeppler S.M., Leon N.D. (2011) Genetic diversity of a maize association population with restricted phenology. Crop Sci. 51 (2): 704-715.

Hotelling H. (1936) Relations between two sets of variables. Biometrika 28: 321-377.

Kang H.M., Zaitlen N.A., Wade C.M., Kirby A., Heckerman D., Daly M.J., Eskin E. (2008) Efficient control of population structure in model organism association mapping. Genetics 178 (3): 1709-1723.

Kraakman A.T.W., Martínez F., Mussiraliev B., Eeuwijk F.A., Niks R.E. (2006) Linkage disequilibrium mapping of morphological, resistance, and other agronomically relevant traits in modern spring barley cultivars. Molecular Breeding 17 (1): 41-58.

Malosetti M., Linden C., Vosman B., van Eeuwijk F. (2007)

BAG | Journal of Basic and Applied Genetics
Vol XXVII (2). December 2016. 49-58.

A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175: 879-889.

Muñoz-Amatriaín M., Cuesta-Marcos A., Endelman J.B., Comadran J., Bonman J.M., Bockelman H.E., Chao S., Russell J., Waugh R., Hayes P., Muehlbauer G. (2014) The USDA Barley Core Collection: Genetic diversity, population structure, and potential for genome-wide association studies. PLoS ONE 9 (4): e94688.

Peña Malavera A. (2015) Aproximaciones estadísticas para el mapeo asociativo en estudios genéticos. Universidad Nacional de Córdoba, Córdoba, 117 pp.

Peña Malavera A., Gutierrez L., Balzarini M. (2014) Componentes principales en mapeo asociativo. JBAG 25: 32-40.

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., Reich D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38 (8): 5.

Pritchard J., Stephens M., Donnelly P. (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Remington D., Thornsberry J., Matsuoka Y., Wilson L., Whitt S., Doebley J., Kresovich S., Goodman M.M., Buckler E.S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl. Acad. Sci. USA 98 (20): 11479-11484.

Sargolzaei M., Schenkel F. (2009) QMSim: a large-scale genome simulator for livestock. Bioinformatics 25: 680-681.

Stich B., Melchinger A. (2009) Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and *Arabidopsis*. BMC Genomics 10 (1): 1-14.

Stich B., Melchinger A., Heckenberger M., Möhring J., Schechert A., Piepho H.P. (2008) Association mapping in multiple segregating populations of sugar beet (*Beta vulgaris* L.). Theor. Appl. Genet. 117 (7): 1167-1179.

Thornsberry J., Goodman M., Doebley J., Kresovich S., Nielsen D., Buckler E. (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28 (3): 286-289.

Tracy C.A., Widom H. (1994) Level-spacing distributions and the Airy kernel. Comm. Math. Phys. 159 (1): 151-174.

Wang M., Jiang N., Jia T., Leach L., Cockram J., Waugh R., Ramsay L., Thomas B., Luo Z. (2012) Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor. Appl. Genet. 124 (2): 233-246.

Wright S. (1951) The genetical structure of populations. Ann. Eugen. 15: 31.

Yu J., Pressoir G., Briggs W., Bi I., Yamasaki M., Doebley J., McMullen M.D., Gaut B.S., Nielsen D.N., Holland J.B., Kresovich S., Buckler E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 2: 203-208.

Zhang Z., Ersoz E., Lai C.Q., Todhunter R.J., Tiwari H.K., Gore M.A., Bradbury P., Yu J., Arnett D., Ordovas J., Buckler E. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat. Genet. 42 (4): 355-360.

Zhu C., Gore M., Buckler E., Yu J. (2008) Status and prospects of association mapping in plants. Plant Genome 1 (1): 16.

AGRADECIMIENTOS

El presente trabajo es parte de la tesis de Andrea Peña Malavera para el cumplimiento de los requisitos del Doctorado en Ciencias de la Ingeniería de la Universidad Nacional de Córdoba y del programa de becas de posgrado del Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).