BAG | Journal of Basic and Applied Genetics
Vol XXVIV (1): 37-49; January 1-July 31, 2018.



# CONTROL DE FALSOS DESCUBRIMIENTOS EN MAPEO ASOCIATIVO CON POBLACIONES ESTRUCTURADAS

# FALSE DISCOVERY RATE CONTROL IN ASSOCIATION MAPPING WITH GENETICALLY STRUCTURED POPULATIONS

Peña Malavera A.[1,2], Bruno C.[1,2], Balzarini M.[1,2,*]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).
[2] Estadística y Biometría, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Av. Valparaíso s/n, Ciudad Universitaria, CP: 5000 (509) Córdoba, Argentina.

*Autor correspondiente: mbalzari@agro.unc.edu.ar

## ABSTRACT

The association tests between molecular markers and phenotypic traits are crucial for the Quantitative Trait Loci (QTL) identification. Biotechnological advances increased the molecular marker information; consequently, the number of genotype-phenotype association tests required incremented too. The multiple statistical inferences (multiplicity) demand corrections of the p-values obtained for each comparison in order to keep limited the error rates for the family of association tests. However, classic statistical correction methods such as Bonferroni, False Discovery Rate (FDR) and the Effective Number of Independent Test ($M_{eff}$) were developed in the context of independent data. Wherever, when the population genetic structure is present, the data are no longer independent. In this paper, we propose a method of correction for multiplicity based on estimation of the effective number of tests from a model that adjust for the underlying correlation structure. We evaluate the performance of the proposed procedure in the estimation of p-values for a set of simulated QTL. The results suggest that the proposed method provides control of FDR and has more power than other methods for multiplicity correction used in association mapping.

**Key words:** multiplicity, association studies, effective number of hypothesis test, linear models.

## RESUMEN

Las pruebas de asociación entre marcadores moleculares y variables fenotípicas son cruciales para la identificación de QTL (*Quantitative Trait Loci*). Los avances biotecnológicos incrementaron la disponibilidad de marcadores genéticos y consecuentemente el número de pruebas de la asociación fenotipo-genotipo. El incremento de pruebas de significancia estadística a realizar en simultaneo (multiplicidad) demanda correcciones de los valores-p obtenidos para cada prueba de hipótesis de manera de mantener acotada las tasas de error para la familia de pruebas de asociación. Las correcciones estadísticas clásicas para el problema de multiplicidad, como Bonferroni, el método de control de la tasa de falsos descubrimientos (FDR) y el número efectivo de pruebas ($M_{eff}$), son ampliamente usadas, pero fueron desarrolladas para datos independientes. Sin embargo, cuando las poblaciones de mapeo están genéticamente estructuradas los datos dejan de ser independientes. En este trabajo, proponemos un método de corrección por multiplicidad basado en estimación del número efectivo de pruebas desde un modelo que ajusta por la estructura de correlación subyacente. Se evalúa el desempeño del procedimiento propuesto a través del análisis de los valores-p obtenidos para un conjunto de QTL simulados. Los resultados sugieren que el método propuesto provee control de la tasa de falsos positivos y presenta mayor potencia que otros métodos de corrección por multiplicidad usados en mapeo asociativo.

**Palabras clave:** multiplicidad, estudios de asociación, número efectivo de pruebas de hipótesis, modelos lineales

Fecha de recepción: 20/02/2017
Fecha de aceptación de versión final: 12/04/2018

## INTRODUCCIÓN

El análisis conjunto de la información de marcadores moleculares del genoma e información fenotípica permite inferir sobre la existencia de asociaciones entre *loci* de marcadores y expresiones de caracteres cuantitativos de interés agronómico (Spindel *et al.*, 2015;Tadesse *et al.*, 2015; Yan *et al.*, 2011; Zhou *et al.*, 2016). En genética vegetal, la presencia de asociaciones estadísticamente significativas entre el estado del marcador y la variante fenotípica permite identificar los QTL subyacentes en la población de mapeo (Breseghello y Sorrells, 2006; Parisseaux y Bernardo, 2004; Pers *et al.*, 2015). Sin embargo, el análisis de asociaciones bajo estructura genética poblacional (EGP) requiere de conceptos y métodos biológicos y estadísticos específicos orientados a disminuir los descubrimientos de falsos QTL, *i.e.* asociaciones que resultan significativas sólo por azar debido a las correlaciones que genera la estructuración genética de la población de mapeo (Malosetti *et al.*, 2007).

Además, los estudios sobre modelos estadísticos para mapeo asociativo (MA) se realizan con propuestas metodológicas que se encuentran aún en desarrollo, y no es una decisión trivial la elección del modelo de análisis más apropiado para un escenario particular (Bernardo, 2013; Cappa *et al.*, 2013; Gutiérrez *et al.*, 2015; Gutiérrez *et al.*, 2011; Locatelli *et al.*, 2013). La selección de uno u otro modelo debe contemplar aspectos estadísticos como el tamaño de la muestra y número de variables indicadoras o marcadores, y aspectos biológicos, entre ellos el nivel de divergencia genética entre subpoblaciones de la población de mapeo, cuando estas existen (Peña-Malavera, 2015).

Luego de ajustado un modelo de MA, será necesario realizar múltiples pruebas de hipótesis estadísticas sobre la asociación de cada uno de los marcadores con el carácter agronómico de interés. En el caso de los modelos de regresión usados en MA, en $H_0$ (hipótesis nula) se plantea que el coeficiente de regresión asociado al efecto del marcador sobre el fenotipo es nulo, *i.e.* el marcador no se encuentra ligado a un QTL. El segundo paso de la prueba de hipótesis se corresponde con la selección de un estadístico cuya distribución sea conocida cuando $H_0$ es cierta y que se desvíe de modo predecible de dicha distribución cuando $H_0$ no es cierta; el estadístico T de *Student* es apropiado para evaluar la significancia estadística de un coeficiente de regresión (Draper y Smith, 1998). Luego, es necesario calcular el valor del estadístico en la muestra que se tenga. Si el valor de dicho estadístico es diferente de lo que se espera bajo $H_0$, se rechazará $H_0$. El nivel de significación empírico o valor-p asociado al valor observado del estadístico es la probabilidad de obtener en el muestreo (bajo $H_0$) valores tan o más raros que el obtenido. Este valor-p representa una medida del acuerdo (o desacuerdo) de la evidencia muestral con la hipótesis nula. Valores-p pequeños habrán de entenderse como evidencia en contra de la hipótesis nula objeto de contraste. En MA, valores-p pequeños llevan al rechazo de la hipótesis nula que establece que no existe ligamiento entre marcador y QTL, y por tanto sugieren la presencia de una variante genética informativa. Para juzgar si un valor-p es pequeño o no, éste se compara con un nivel de significación pre-especificado, $\alpha$.

Dos criterios de evaluación cobran importancia para evaluar una prueba de hipótesis estadística: la capacidad de mantener su tamaño nominal o nivel de significación $\alpha$, y la potencia de la prueba para detectar una hipótesis nula falsa. El primero está relacionado con el error tipo I, el cual tiene probabilidad de ocurrencia denotada por $\alpha$ y el segundo, con el error tipo II con probabilidad de ocurrencia denotada por $\beta$ (Balzarini *et al.*, 2008). Estos errores pueden ser analizados mediante tasas de error por comparación que representan el valor esperado del cociente entre el número de inferencias erróneas y el número de inferencias realizadas o por experimento; estas últimas estiman la probabilidad de obtener al menos un error dentro de una familia de pruebas de hipótesis.

En estudios de MA, la hipótesis de interés es la hipótesis nula de falta de asociación marcador-fenotipo. Una prueba estadística con baja tasa de error tipo II es aquella con capacidad (o potencia) para detectar asociaciones verdaderas. La mayor potencia de un modelo de MA con respecto a otro no se asocia con un incremento de la tasa de error tipo I, *i.e.*, un incremento en la probabilidad de concluir que existe asociación cuando en realidad no está presente, sino con el tamaño de la población de mapeo y con la cantidad de marcadores o pruebas de hipótesis que se realizan sobre el mismo conjunto de datos.

En MA se ajustan modelos de regresión por cada marcador que se evalúa y por tanto hay múltiples hipótesis a contrastar sobre el mismo conjunto de datos. Este procedimiento debe realizarse siendo consciente de que algunas hipótesis serán objeto de rechazo sólo por azar, con una probabilidad mucho mayor que el nivel de significación nominal empleado para contrastar cada una de ellas. Para una prueba de hipótesis sobre uno de M coeficientes del modelo

de MA y bajo $H_0$, hay probabilidad tan sólo $\alpha$ de que el estadístico T calculado exceda en valor absoluto del cuantil $\alpha/2$ de una distribución T de *Student* con N-M grados de libertad. Pero la probabilidad de que algún estadístico T, desde una miríada de valores T (correspondientes a los M marcadores moleculares), exceda de $t_{\alpha/2,N-M}$, asumiendo independencia, es mayor con Prob (algún $\beta_i \neq 0)=1-(1-\alpha)^m$. Luego, con probabilidad mucho mayor a $\alpha$, algún coeficiente de marcador molecular puede resultar significativo sólo por azar. Esta probabilidad depende de M, es decir, aumenta a medida que se incrementa el número de marcadores moleculares evaluados. Este problema de inferencia simultánea demanda, consecuentemente, de la corrección de los valores-p y debe ser atendido en el contexto de MA para no perder potencia (Xiao *et al.*, 2013). Para el contexto de datos independientes existen métodos de corrección de valores-p por multiplicidad que garantiza que la tasa de falsos positivos sea menor o igual que un valor pre-seleccionado. El método de control del error tipo I más conocido es la aproximación de Bonferroni (Bonferroni, 1935). Sin embargo, este método es excesivamente conservador cuando las pruebas de hipótesis son numerosas. Aplicado a estudios de MA puede reducir drásticamente la cantidad de marcadores positivos, incluso llegar a no detectar ninguna asociación significativa. Una corrección alternativa es la propuesta por Benjamini y Hochberg (Benjamini y Hochberg, 1995) para controlar la proporción esperada de hipótesis mal rechazadas respecto a todas aquellas rechazadas (Miller *et al.*, 2001; Sabatti *et al.*, 2003; Schwartzman *et al.*, 2008; Tusher *et al.*, 2001).

El umbral de significación nominal $\alpha$ es inapropiado para reportar resultados del mapeo asociativo, no sólo por la multiplicidad de pruebas de hipótesis que se realizan sino también por la correlación esperable entre las pruebas debido a la correlación entre marcadores moleculares (pruebas no independientes). En 2001, se propuso un ajuste de valores-p para pruebas correlacionadas que se basa en la determinación del número efectivo ($M_{eff}$) de pruebas independientes (Cheverud, 2001). Li y Ji (2005) propusieron una estimación más exacta del $M_{eff}$ basada en la descomposición por valor singular de una matriz de correlaciones entre marcadores y diseñaron un procedimiento (LJ) basado en este nuevo $M_{eff}$ para controlar el error tipo I. El método LJ ha sido usado con éxito en el contexto del análisis de QTL clásico donde, aunque los marcadores pueden estar correlacionados, los casos son independientes porque provienen de una población de mapeo sin estructura de correlación genética entre los individuos.

En este trabajo se propone una corrección por multiplicidad que contempla la estructura genética de la población de mapeo cuando esta existe. La propuesta está basada en el número de pruebas efectivas o independientes (similar a LJ). La modificación propuesta utiliza los ejes derivados de la descomposición aplicada sobre la matriz de estadísticos de Mantel y Haenszel (MH) (1959) incorporando la información conocida de la estructura genética poblacional. El método propuesto fue comparado con otros métodos de corrección por multiplicidad usando datos simulados. Los modelos de MA ajustados previamente a la corrección por multiplicidad fueron tres, uno que no contempla ninguna corrección por estructura genética poblacional (EGP) (Modelo naive) y otros con distintos tipos de ajuste de la EGP. La comparación se realizó usando bases de datos de marcadores moleculares simulados bajo distintos escenarios biológicos de EGP. Los ajustes de valores-p se realizaron luego de escoger las mejores estrategias de modelación para MA para los escenarios simulados. Para cada combinación modelo de MA-método de ajuste de valor-p, se obtuvieron tasas de falsos positivos y potencia ($\varphi$), bajo dos escenarios con diferente nivel de EGP (bajo y alto $F_{ST}$). El objetivo de este trabajo es evaluar el desempeño de diferentes métodos de corrección de valor-p por multiplicidad cuando ellos son aplicados luego de ajustar modelos de mapeo asociativo, que contemplan o no la EGP subyacente, bajo distintos escenarios biológicos en lo que concierne a tamaño de la población de mapeo, cantidad de marcadores moleculares y nivel de divergencia genética entre subpoblaciones de la población de mapeo.

## MATERIALES Y MÉTODOS

*Datos*

Los datos de marcadores moleculares usados en este trabajo fueron simulados a través de QMSim (Sargolzaei y Schenkel, 2009) involucrando escenarios con cantidad de genotipos 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 cuatro escenarios biológicos correspondientes a, dos niveles de divergencia genética entre poblaciones (bajo y alto $F_{ST}$), y dos tamaños distintos de

poblaciones de mapeo (n≈150 y n≈300), equivalente a 30 y 60 líneas por población simulada. Los datos simulados fueron creados a partir de una población histórica con un tamaño poblacional de 200 individuos y el sistema de cruzamiento basado en la unión al azar de gametos (cruzamientos aleatorios). La coancestría promedio fue baja como sucede en numerosas poblaciones usadas para MA en vegetales.Variando el número de generaciones desde la población fundadora, se crearon diferentes niveles de divergencia genética poblacional. Los datos simulados fueron codificados como 0 y 1 para cada marcador. 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 (Tabla 1).

Dada la matriz de marcadores moleculares simulados se escogieron aleatoriamente 20 marcadores y con ellos se realizó una combinación lineal con efectos que siguen una distribución gamma con media 2 y varianza 5 [$\Gamma(2,5)$] para simular el efecto de los *loci* ligados a un QTL. Adicionalmente, se anexó a cada perfil molecular la realización de una variable aleatoria con distribución normal de media 100 (representa la media del carácter que depende del efecto poligénico de *background*) y varianza 25 (representa la variabilidad experimental, *i.e.* desvío estándar 5, no superior al 5% de la media del carácter fenotípico). A esta variable simulada se le adicionaron los efectos de los marcadores ligados extraídos de la distribución gamma. Los valores resultantes fueron usados como variable fenotípica para los modelos de MA. La ubicación de cada uno de los 20 QTL simulados sobre los marcadores seleccionados aleatoriamente fue usada para determinar el carácter de verdad de la hipótesis nula.

*Modelo de Mapeo Asociativo*

Se estimaron ocho modelos de mapeo asociativo para evaluar el efecto del marcador sobre el fenotipo (Tabla 2). El modelo básico a partir del cual derivaron los modelos de MA comparados fue:

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

donde $y$ es el vector de valores fenotípicos (conteniendo un dato fenotípico por genotipo), $X$ es la matriz de datos de los marcadores moleculares (tantas columnas como marcadores usados), $\beta$ es un vector desconocido de efectos de los alelos de cada marcador que debe ser estimado para identificar aquellos marcadores asociados con el fenotipo, *EGP* es la matriz de estructura genética (construida alternativamente como la matriz Q de la salida del software *structure* o la matriz P de componentes principales estadísticamente significativas seleccionadas por el estadístico de Tracy-Widom (1994), 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 de poligen 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, que se supone normalmente distribuido con media cero y varianza constante $\sigma^2_e$. Se supone que el vector $u$ se distribuye independientemente del vector $e$ y con matriz de varianzas y covarianzas dada por $\sigma^2_e \square K$, siendo $K$ la matriz de similitud entre todos los pares de perfiles moleculares derivadas del software EMMA (Kang *et al.*, 2008) y que es usada como indicador del parentesco o la filogenética existente entre los genotipos de la población de mapeo.

*Criterios de comparación*

Todos los modelos fueron ajustados usando *Info-Gen* (Balzarini y Di Rienzo, 2004) y su interfaz con R (Team, 2013). El desempeño de los ocho modelos se evaluó usando las curvas de distribución acumulada de valores-p. Para construir las curvas de distribución de valores-p, se usó la opción función de distribución empírica del software *Info-Gen* (Balzarini y Di Rienzo, 2004) usando como variable de análisis el 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 debiera 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.

Luego de aplicarse las correcciones por multiplicidad en los modelos seleccionados, se usó como criterio de evaluación la tasa de falsos descubrimientos o FDR (del inglés, *False Discovery Rate*) (Benjamini y Hochberg, 1995)

**Tabla 1.** Tamaño poblacional y diversidad genética poblacional que caracteriza la estructura genética subyacente en poblaciones de mapeo simuladas con 300 marcadores multilocus-bialélicos como dato genómico.

| Escenario | Diversidad genética | | Tamaño poblacional promedio |
|---|---|---|---|
| | Estadístico $F_{ST}$ | Nivel | |
| I | 0.03 | Bajo | 150 |
| II | 0.03 | Bajo | 300 |
| III | 0.20 | Alto | 150 |
| IV | 0.21 | Alto | 300 |

**Tabla 2.** Ocho modelos de mapeo asociativo para evaluar el efecto del marcador sobre el fenotipo en datos simulados.

| Matriz de Parentesco | Estructura Genética Poblacional | | | | |
|---|---|---|---|---|---|
| | No | Q fijo | ACP fijo | Qaleatorio | ACP aleatorio |
| No | Naive | Q | P | QA | PA |
| Si | K | QK | PK | -- | -- |

Nota: Q es la matriz de probabilidades de pertenencia a los g grupos calculada por el software structure, P es la matriz de componentes principales retenidas mediante el estadístico de Tracy-Widom(1994) y K es la matriz de parentesco propuesta por Kang *et al.*, (2008).

y la potencia estadística. 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,

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

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 fue calculada como,

$$\text{j} = \frac{VP}{VP + FN}$$

Usando los datos de los 4 escenarios simulados consideramos el problema de contrastar simultáneamente *m* hipótesis nulas $H_0^j, j = 1,...,m,$ con *m*= 300. Si *R* es la cantidad de hipótesis rechazadas, los resultados posibles luego del contraste de hipótesis pueden resumirse como en la Tabla 3. Los conjuntos de subíndices que corresponden a hipótesis nulas verdaderas y falsas $\square_0 = \{j : H_0^j \text{ es verdadera}\}$ y $\square_1 = \{j : H_0^j \text{ no es verdadera}\}$ son desconocidos y serán estimados mediante la simulación. El conjunto total de índices es $\square = \{1, 2, ..., m\} = \square_0 \cup \square_1$. Las cantidades de hipótesis nulas verdaderas $m_0 = \#\square_0$ y falsas $m_1 = m - m_0 = \#\square_1$, fueron estimadas por conteo dentro de cada escenario. En cada escenario simulado se estimó la cantidad de hipótesis nulas rechazadas *R* y no rechazadas *m-R* (variables aleatorias observables a través del conjunto de prueba de hipótesis).

Tabla 3. Situaciones posibles luego de realizar m pruebas de hipótesis.

| Realidad | Decisión | | Total de hipótesis |
|---|---|---|---|
| | No rechazar hipótesis nula | Rechazar hipótesis nula | |
| Hipótesis nula verdadera | VN | FP(Falsos Positivos) | $m_0$ |
| Hipótesis nula falsa | FN(Falsos negativos) | VP | $m_1$ |
| Total | m-R | R | m |

*Procedimientos de corrección por multiplicidad*

Para cada escenario se implementaron 3 métodos de corrección por multiplicidad y con fines comparativos, también se observaron los resultados luego de contrastar las *m* hipótesis sin corrección por multiplicidad. Los métodos implementados para corregir valores-p del conjunto de pruebas por multiplicidad fueron tres: (1) BH, propuesto por Benjamini y Hochberg (1995); (2) LJ propuesto por Li y Ji (2005); y (3) un nuevo procedimiento propuesto en este trabajo que llamamos Li y Ji Modificado (MLJ). El método de Bonferroni (1935) tradicionalmente usado en problemas de multiplicidad, no fue usado por ser altamente conservador en situaciones como las que se producen en MA donde el número de contrastes de hipótesis asciende a cientos e incluso miles de pruebas.

Para implementar la corrección propuesta por Benjamini y Hochberg (1995) se realizó el siguiente procedimiento:

1. Los valores-p de las m pruebas de hipótesis se ordenaron de menor a mayor.
2. El valor-p mayor no fue ajustado.
3. Cada uno de los restantes valores-p se multiplicó por el número total de marcadores y se dividió por el valor que denota su orden en la lista de valores-p ordenados. Si el valor resultante era menor que 0,05, se rechazaba la hipótesis nula.

Para implementar el método de corrección de Li y Ji (2005), basado en la idea propuesta por Cheverud (2001) para ajustar pruebas de hipótesis correlacionadas, se realizaron los siguientes pasos:

1. Se calculó la matriz de correlación para todos los *loci*.
2. Se calculó el número efectivo ($M_{eff}$) de pruebas independientes a través de la obtención de los valores propios de la matriz de correlación, donde *M* es el número de pruebas y $\lambda_i (i = 1, \ldots, M)$ son los valores propios:

$$M_{eff} = \sum_{i=1}^{M} f(|\lambda_i|)$$

$$f(x) = I(x \geq 1) + (x - [x]), x \geq 0$$

donde $I(x \geq 1)$ es una función indicadora que vale 1 cuando $x \geq 1$ y 0 en otro caso, y $[x]$ es la función parte entera que devuelve el mayor entero posible menor o igual a x.

3. Se ajustó el nivel de significación de la prueba como si hubiera Meff pruebas independientes usando la corrección de Sidak (1967): $\alpha_p = 1 - (1 - \alpha_e)^{1/M_{eff}}$
4. Se realizaron las *m* pruebas de hipótesis locus por locus y cuando el valor-p de alguna prueba era menor que $\alpha_p$, la hipótesis de no asociación fue rechazada.

El método de corrección por multiplicidad propuesto en este trabajo está basado en la aproximación de Li y Ji (2005) con una modificación que contempla la posible EGP que subyace la población de mapeo. En caso de poblaciones estructuradas, la modificación analizará la correlación entre marcadores, controlando por la presencia de los grupos que definen la EGP, para derivar un Meff. Con este fin, la matriz de correlación utilizada en el método LJ es reemplazada por una matriz de estadísticos $\chi^2$ de Mantel y Haenszel (1959). Los estadísticos $\chi^2$ fueron obtenidos a partir de tablas de contingencia construidas entre pares de marcadores, fijando la variable que indica el grupo al cual pertenecen los genotipos como variable de control.

La evaluación del impacto de los métodos de corrección se realizó considerando ambos niveles de FST usados en la simulación, ambos tamaños poblacionales y distintos modelos de MA para generar la lista de valores-p sin corregir. Los modelos ajustados seleccionados para evaluar la corrección por multiplicidad fueron: QK y K (Yu *et al.*, 2006) y el modelo de mapeo de regresión de efectos fijos que incluye los 300 marcadores como variables independientes y no incorpora de ninguna manera explícita el modelado de la EGP (modelo *naive*).

## RESULTADOS

Las funciones de distribución acumulada para los 4 escenarios que involucraron datos genéticos de 300 marcadores moleculares multilocus-bialélicos, consistentemente mostraron que los modelos con mejor ajuste fueron el modelo K y el modelo QK (Figura 1). Con bajo $F_{ST}$ (Figura 1 arriba, escenarios I y II) se observó que el modelo de menor desempeño fue el modelo P. Los modelos se comportaron de manera parecida cuando fueron ajustados en un contexto de alto $F_{ST}$ (Figura 1 abajo, escenarios III y IV), situación en la que se observó menor 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.

Las tasas FDR fueron menores para alto FST en los tres métodos de corrección por multiplicidad, *i.e.*, para los dos modelos de mapeo asociativo seleccionados por tener el mejor desempeño (modelo K y modelo QK) y para el modelo *naive*, seleccionado como modelo de referencia; tanto en poblaciones de tamaño 150 como de 300 se pudo

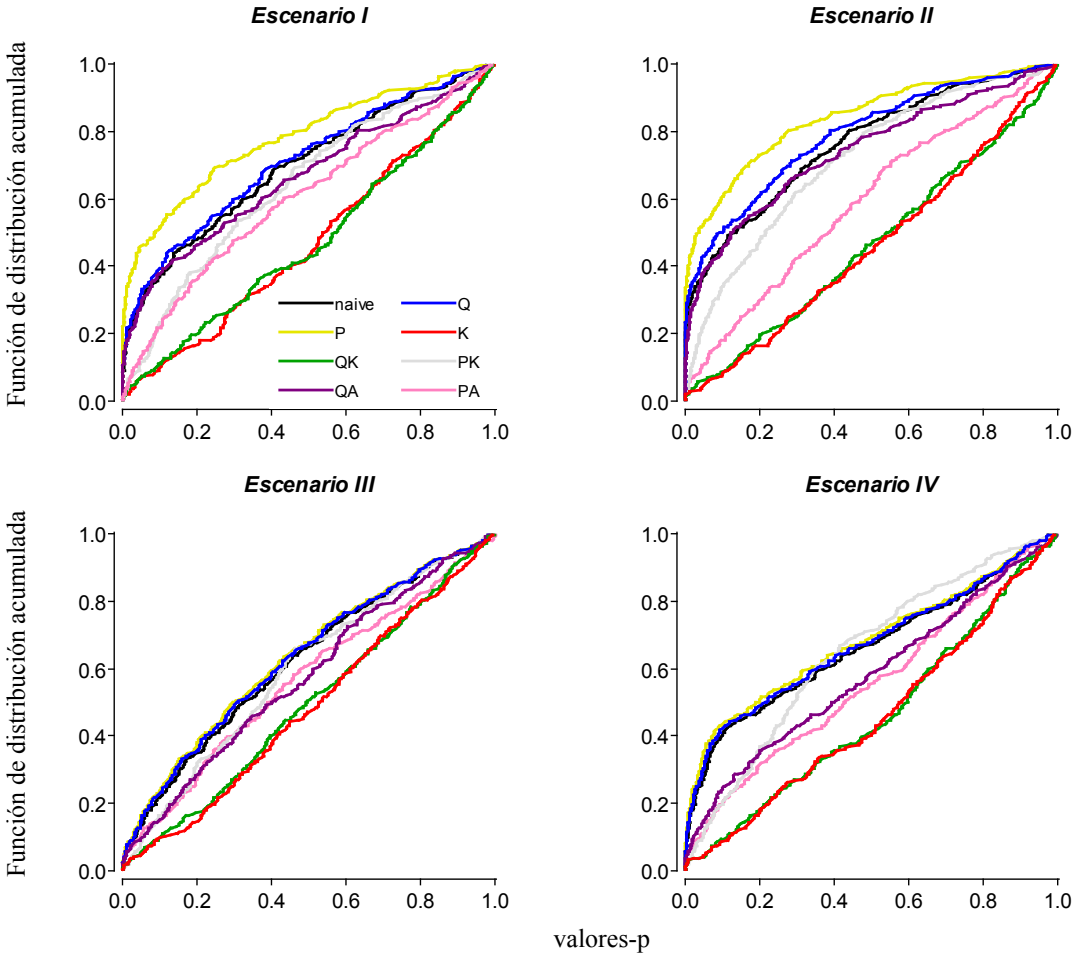


**Figura 1.** Función de distribución acumulada de los valores-p para cada uno de los ocho modelos evaluados en cuatro escenarios simulados que contienen 300 marcadores moleculares multilocus-bialélicos. En la columna de la izquierda escenarios con tamaño poblacional de 150 y en la columna derecha tamaño poblacional de 300. Arriba $F_{ST}$ bajo y abajo $F_{ST}$ alto.

observar que las tasas FDR fueron menores para alto FST. Cuando nos ubicamos en la situación de no corrección por estructura ni parentesco en el modelado, es decir cuando ajustamos un modelo *naive*, podemos observar que la tasa FDR disminuye con todas las correcciones respecto a sin corrección (SC), pero con alto FST baja en mayor medida corrigiendo con MLJ que con BH y LJ, esto se debe a que la estructura es grande y no fue corregida previamente en el modelado. Cuando la estructura es baja, es decir el nivel de convergencia entre poblaciones es alto, es más importante incluir la corrección por estructura en el modelado que en la corrección por multiplicidad, si bien las tasas FDR igualmente bajan con las correcciones por multiplicidad, dicha disminución no se produce de forma tan drástica como en la situación de alta estructura genética poblacional (Tabla 4 y Tabla 5).

Los resultados indicaron que aún para el caso de no corrección por EGP, es decir con el modelo *naive*, la aplicación de métodos de ajustes de valor-p por multiplicidad reduce la potencia significativamente. Potencias excesivamente bajas se observaron en escenarios correspondientes al menor tamaño poblacional (150 individuos) cuando las subpoblaciones tenían poca divergencia. Cuando se ajustaron los modelos de mapeo K o QK, la corrección de valores-p también produjo reducciones importantes de potencia. Estas reducciones fueron de mayor magnitud que las producidas por el ajuste de un modelo de MA que controla EGP y sin corrección por multiplicidad. Es importante, mencionar que los métodos presentados para corrección por multiplicidad han sido diseñados para controlar el error de tipo I en una familia de pruebas y no para aumentar la probabilidad de detectar verdaderos positivos. Por la relación teórica existente entre los errores de tipo I y de tipo II en las pruebas de hipótesis es de esperar que la reducción significativa que estos métodos producen a nivel de FDR se encuentre asociada a pérdida de potencia. No obstante, la potencia con alto FST para el método MLJ fue igual o superior a la de los otros dos métodos de corrección de valores-p por multiplicidad (Tabla 6 y Tabla 7).

Con el mayor de los tamaños poblacionales (300), la aplicación del método MLJ directamente sobre los valores-p derivados del modelo más simple (de efectos fijos y sin corrección por EG, *i.e.*, *naive*) produjo potencias similares a las obtenidas con el modelo de mapeo QK y sin ninguna corrección de valores-p. Las mayores pérdidas de potencia se obtuvieron con las dos estrategias usadas para controlar por EGP simultáneamente, es decir en el momento del modelado y al utilizar los valores-p para determinar significancia. MLJ produjo menor FDR que LJ en escenarios de alta estructura genética y no mostró mayores reducciones de potencia que LJ.

**Tabla 4.** Tasa de falsos descubrimientos (FDR) para tres modelos de mapeo asociativo, tres opciones de corrección de valores-p por inferencia simultánea bajo dos niveles de estructura genética poblacional, baja ($F_{ST}$= 0,03) y alta ($F_{ST}$= 0,2) divergencia genética, con un tamaño poblacional de 150.

| | **Bajo $F_{ST}$** | | | | **Alto $F_{ST}$** | | | |
|---|---|---|---|---|---|---|---|---|
| | Correcciones* | | | | | | | |
| Modelo** | SC | BH | LJ | MLJ | SC | BH | LJ | MLJ |
| *naive* | 0.30 | 0.21 | 0.34 | 0.30 | 0.24 | 0.18 | 0.18 | 0.15 |
| K | 0.12 | 0.07 | 0.07 | 0.07 | 0.11 | 0 | 0.07 | 0 |
| QK | 0.13 | 0.07 | 0.07 | 0.07 | 0.11 | 0 | 0.03 | 0 |

*SC: Sin corrección por multiplicidad, BH: Benjamini y Hochberg, LJ: Li y Ji, MLJ: Li y Ji Modificado. ***naive:* sin corrección por estructura, K: con corrección por matriz de parentesco y QK: modelo mixto con Q, corrección mediante la matriz de probabilidades a posteriori obtenida con el software *Structure*, como factor de efectos fijos y K factor de efectos aleatorios.

**Tabla 5.** Tasa de falsos descubrimientos (FDR) para tres modelos de mapeo asociativo, tres opciones de corrección de valores-p por inferencia simultánea bajo dos niveles de estructura genética poblacional, baja ($F_{ST}$= 0,03) y alta ($F_{ST}$= 0,2) divergencia genética, con un tamaño poblacional de 300.

| | **Bajo $F_{ST}$** | | | | **Alto $F_{ST}$** | | | |
|---|---|---|---|---|---|---|---|---|
| | Correcciones* | | | | | | | |
| Modelo** | SC | BH | LJ | MLJ | SC | BH | LJ | MLJ |
| *naive* | 0.31 | 0.28 | 0.24 | 0.24 | 0.24 | 0.18 | 0.09 | 0.05 |
| K | 0.05 | 0 | 0.03 | 0 | 0.11 | 0 | 0 | 0 |
| QK | 0.07 | 0 | 0 | 0 | 0.11 | 0 | 0 | 0 |

*SC: Sin corrección por multiplicidad, BH: Benjamini y Hochberg, LJ: Li y Ji, MLJ: Li y Ji Modificado. ***naive:* sin corrección por estructura, K: con corrección por matriz de parentesco y QK: modelo mixto con Q, corrección mediante la matriz de probabilidades a posteriori obtenida con el software *Structure*, como factor de efectos fijos y K factor de efectos aleatorios.

**Tabla 6.** Tasa de falsos descubrimientos (FDR) para tres modelos de mapeo asociativo, tres opciones de corrección de valores-p por inferencia simultánea bajo dos niveles de estructura genética poblacional, baja ($F_{ST}$= 0,03) y alta ($F_{ST}$= 0,2) divergencia genética, con un tamaño poblacional de 300.

| | **Bajo $F_{ST}$** | | | | **Alto $F_{ST}$** | | | |
|---|---|---|---|---|---|---|---|---|
| | Correcciones* | | | | | | | |
| Modelo** | SC | BH | LJ | MLJ | SC | BH | LJ | MLJ |
| *naive* | 0.60 | 0.15 | 0.15 | 0.15 | 0.55 | 0.05 | 0.10 | 0.10 |
| K | 0.35 | 0 | 0.05 | 0.05 | 0.25 | 0.05 | 0.05 | 0.05 |
| QK | 0.35 | 0 | 0.05 | 0.05 | 0.30 | 0.05 | 0.10 | 0.05 |

*SC: Sin corrección por multiplicidad, BH: Benjamini y Hochberg, LJ: Li y Ji, MLJ: Li y Ji Modificado. ***naive:* sin corrección por estructura, K: con corrección por matriz de parentesco y QK: modelo mixto con Q, corrección mediante la matriz de probabilidades a posteriori obtenida con el software *Structure*, como factor de efectos fijos y K factor de efectos aleatorios.

**Tabla 7.** Potencia estadística para tres modelos de mapeo asociativo, tres opciones de corrección de valores-p por inferencia simultánea bajo dos niveles de estructura genética poblacional, baja ($F_{ST}$= 0,03) y Alto ($F_{ST}$= 0,2) divergencia genética, con un tamaño poblacional de 300.

| | **Bajo $F_{ST}$** | | | | **Alto $F_{ST}$** | | | |
|---|---|---|---|---|---|---|---|---|
| | Correcciones* | | | | | | | |
| Modelo** | SC | BH | LJ | MLJ | SC | BH | LJ | MLJ |
| *naive* | 0.75 | 0.35 | 0.55 | 0.45 | 0.75 | 0.25 | 0.35 | 0.45 |
| K | 0.40 | 0.10 | 0.20 | 0.10 | 0.45 | 0.05 | 0.20 | 0.20 |
| QK | 0.45 | 0.05 | 0.15 | 0.15 | 0.45 | 0.05 | 0.20 | 0.20 |

*SC: Sin corrección por multiplicidad, BH: Benjamini y Hochberg, LJ: Li y Ji, MLJ: Li y Ji Modificado. ***naive:* sin corrección por estructura, K: con corrección por matriz de parentesco y QK: modelo mixto con Q, corrección mediante la matriz de probabilidades a posteriori obtenida con el software *Structure*, como factor de efectos fijos y K factor de efectos aleatorios.

## DISCUSIÓN

La corrección por multiplicidad con el método MLJ, propuesto en este trabajo, fue más potente que los métodos de corrección LJ y BH con los que fue comparado bajos distintos escenarios en relación al tamaño poblacional y al nivel de diferenciación genética entre las subpoblaciones de la población de mapeo. MLJ disminuyó la FDR en mayor proporción que BH y LJ en escenarios con alta divergencia genética poblacional aun cuando la EGP no había sido incluida en el ajuste del modelo, *i.e.* modelo *naive*. Para estudios de asociación genómica donde los marcadores pueden estar asociados, Li y Ji (2005) propusieron determinar el número efectivo de pruebas independientes ($M_{eff}$) para usarlo posteriormente en la corrección de los valores-p de las pruebas de hipótesis realizadas para descubrir asociaciones entre el estado de cada marcador y el fenotipo. Li y Ji *op. cit.*, usaron la descomposición espectral de la matriz de correlación entre marcadores para determinar el número de pruebas independientes. Esta aproximación es la base del método de corrección MLJ, aunque a diferencia de LJ, MLJ incorpora la estructura genética poblacional (EGP) en el cálculo de $M_{eff}$. La estrategia metodológica para contemplar la falta de independencia entre las pruebas de hipótesis que es ocasionada por la presencia de EGP, es ajustar la asociación entre cualquier par de marcadores por la estructura de grupo que determina la EGP. Así, la matriz de correlación usada por LJ es reemplazada por la matriz de medidas de asociación calculadas mediante el estadístico $\chi^2$ de Mantel y Haenszel (1959) fijando como variable de control el grupo al que pertenece cada individuo de la población de mapeo.

Li *et al.* (2012) también trabajaron con la matriz de correlación entre marcadores, pero propusieron particionar dicha matriz en bloques de grupos de ligamiento para acelerar los tiempos computacionales relativos a la descomposición espectral. La estratificación en grupos de ligamiento permitió obtener una tasa de error de tipo I con valores cercanos al correcto de 0,05. Mientras mayor era la cantidad de estratos, más se acercaba al nivel de significación deseado (Li *et al.*, 2012). La propuesta fue comparada con otros métodos de corrección por multiplicidad, incluyendo el método de Li y Ji (2005) y el método de Moskvina y Schmidt (2008) que también estima el número de pruebas independientes a partir de la matriz de correlación entre marcadores, pero agrega el grado de independencia estadística entre las pruebas de hipótesis de un marcador respecto a los marcadores que lo preceden ($K_{eff}$). LJ resultó, en esta comparación, más liberal que el método de Moskvina y Schmidt (2008) el cual fue menos conservador a medida que aumentaba la cantidad de bloques.

BH (Benjamini y Hochberg, 1995), es otro de los desarrollos metodológicos difundidos para la corrección de valores-p en casos de múltiples pruebas de hipótesis y de extenso uso en mapeo asociativo (Gutiérrez *et al.*, 2011; Muñoz-Amatriaín *et al.*, 2014; Olukolu *et al.*, 2014; Wang *et al.*, 2012). BH fue concebido como un método para corrección de la tasa de falsos descubrimientos. Para todos los métodos de control de multiplicidad, la potencia o capacidad de detectar asociaciones verdaderas disminuye a medida que aumenta el número de pruebas (Benjamini y Hochberg, 1995). En nuestro estudio, la pérdida de potencia fue mayor en BH y LJ que en MLJ y fue más abrupta para BH en el escenario con menor tamaño poblacional y bajo nivel de divergencia genética.

Wang *et al.* (2012) estudiaron el efecto del tamaño poblacional en la habilidad para detectar QTL usando un modelo que corregía por estructura sobre líneas endocriadas de cebada con EGP. A partir de las líneas disponibles generaron poblaciones de tamaños diferentes: 96, 192, 288, 384, 480, 576 y 672 individuos, encontrando que reducir el tamaño poblacional por debajo de 384 individuos produce una tasa alta de falsos descubrimientos. Cuando el tamaño de la población de mapeo disminuyó de 480 a 288 individuos, la FDR aumentó un 18%. Nuestros hallazgos también mostraron incremento de la FDR (7%), cuando el tamaño poblacional disminuyó de 300 a 150 individuos.

Beavis (1998) postuló que la construcción de un resultado estadístico en un análisis de asociación donde se pretende identificar QTL, puede caracterizarse según los valores de error de tipo I asociado a la tasa de falsos descubrimientos y según los valores de potencia para las pruebas de asociación. Beavis *op. cit.* estudió el efecto de la potencia simulando tres tamaños poblacionales (100, 500 y 1000 individuos para una progenie F2) para detectar 10 y 40 QTL, bajo tres niveles de heredabilidad expresado como variabilidad fenotípica explicada por los QTL (30, 63, 95%). Los valores de potencia estimados fueron menores a 6% con tamaños poblacionales de 100 individuos aun con QTL de alta heredabilidad. A medida que aumentaba el tamaño poblacional, los valores de potencia estimados aumentaron. Con 500 individuos las potencias fueron mayores al 50% sólo con 10 QTL, pero con 40 QTL se necesitaron 1000 individuos para alcanzar potencias

mayores al 50% en la detección de QTL de mediana a alta heredabilidad. Sin embargo, los QTL de menor efecto no fueron bien detectados aun con 1000 individuos en la población. Bradbury *et al.* (2011) usando datos genotípicos del programa de cebada (BarleyCAP), simularon datos con efectos fenotípicos para diferente cantidad de QTL con tres niveles de heredabilidad. En cada escenario ellos calcularon la potencia y el FDR para tamaños muestrales de 100 y 300 individuos. Bajo el modelo K para mapeo asociativo, las simulaciones con 100 líneas se desempeñaron pobremente para la detección de QTL, pero simulaciones con 300 líneas se desempeñaron adecuadamente. Las simulaciones con 300 líneas resultaron suficientemente potentes para detectar QTL cuando el carácter fenotípico fue controlado por pocos QTL con tamaño de efecto grande, aunque la potencia para detectar QTL de tamaño de efecto pequeño fue deficiente. En nuestro trabajo, las potencias fueron bajas debido a los tamaños poblaciones. La corrección por multiplicidad disminuye la FDR y provoca una pérdida de potencia en cualquiera de los métodos a los que se ha hecho referencia. Sin embargo, la corrección por multiplicidad con MLJ, usada en modelos donde no se ha descontado el efecto de la EGP previamente, fue la opción que condujo a la menor pérdida de potencia en poblaciones con alta divergencia genética y mayor tamaño poblacional. En poblaciones de mapeo de interés agronómico es frecuente la presencia de ancestros en común y por lo tanto la existencia de EGP, por lo que la práctica de usar un modelo que descuente su efecto sobre la dependencia en la pruebas de hipótesis es recomendable para disminuir la tasa de falsos descubrimientos y trabajar con poblaciones de mapeos de más de 300 individuos para evitar que la tasa de no detección de QTL sea alta, sobre todo en contextos donde se esperan varios QTL de moderado o bajo efecto.

## BIBLIOGRAFÍA

Balzarini M., Di Rienzo J. (2004) Info-Gen. Universidad Nacional de Cordoba, Córdoba.

Balzarini M.G., Gonzalez L., Tablada M., Casanoves F., Di Rienzo J.A., Robledo C.W. (2008) Infostat. Manual del Usuario, Córdoba, Argentina.

Beavis W.B. (1998) QTL analyses: power, precision, and accuracy. In: Patterson A.H. (Ed.) Molecular dissection of complex traits. CRC Press, Boca Raton.

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

Bernardo R. (2013) Genome wide markers as cofactors for precision mapping of quantitative trait loci. Theor. Appl. Genet. 126: 999-1009. doi:10.1007/s00122-012-2032-2.

Bonferroni C.E. (1935) Il calcolo delle assicurazioni su gruppi di teste. Studi in Onore del Professore Salvatore Ortu Carboni, pp. 13-60.

Bradbury P., Parker T., Hamblin M.T., Jannink J.L. (2011) Assessment of power and false discovery in genome-wide association studiesusing the BarleyCAP germplasm. Crop Sci. 51: 52-59.

Breseghello F., Sorrells M.E. (2006). Association Mapping of Kernel Size and Milling Quality in Wheat (*Triticum aestivum* L.) Cultivars. Genetics, *172*(2), 1165–1177. http://doi.org/10.1534/genetics.105.044586.

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: e81267. doi:10.1371/journal.pone.0081267.

Cheverud J.M. (2001) A simple correction for multiple comparisons in interval mapping genome scans. Heredity 87: 52-58.

Draper N.R., Smith H. (1998) Applied regression analysis, 3rd Edition Wiley, New York.

Excoffier L., Hofer T., Foll M. (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103: 285-298. doi:http://www.nature.com/hdy/journal/v103/n4/suppinfo/hdy200974s1.html.

Gutiérrez L., Cuesta-Marcos A., Castro A.J., von Zitze-

witz J., Schmitt M., Hayes P.M. (2011) Association mapping of malting quality Quantitative Trait Loci in winter barley: positive signals from small germplasm arrays. Plant Gen. 4: 256-272. doi:10.3835/plantgenome2011.07.0020.

Gutiérrez L., Germán S., Pereyra S., Hayes P., Pérez C., Capettini F., Locatelli A., Berberian N.M., Falconi E.E., Estrada R., Fros D., Gonza V., Altamirano H., Huerta-Espino J., Neyra E., Orjeda G., Sandoval-Islas S., Singh R., Turkington K., Castro A.J. (2015) Multi-environment multi-QTL association mapping identifies disease resistance QTL in barley germplasm from Latin America. Theor. Appl. Genet. 128 (3): 501-516. doi:10.1007/s00122-014-2448-y.

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: 1709-1723. doi:10.1534/genetics.107.080101.

Li J., Ji L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. Heredity 95: 221-227.

Li M.X., Yeung J.M.Y., Cherny S.S., Sham P. (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum. Genet. 131(5): 747. https://doi.org/10.1007/s00439-011-1118-2.

Locatelli A., Cuesta-Marcos A., Gutiérrez L., Hayes P., Smith K., Castro A. (2013) Genome-wide association mapping of agronomic traits in relevant barley germplasm in Uruguay. Mol. Breeding 31: 631-654. doi:10.1007/s11032-012-9820-x.

Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA (2007) A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. Genetics 175:879–889.

Mantel N., Haenszel W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute 22: 719-748.

Miller C.J., Genovese C., Nichol R.C., Wasserman L., Connolly A., Reichart D., Hopkins A., Schneider J., Moore A. (2001) Controlling the false-discovery rate in astrophysical data analysis. The Astronomical Journal 122 (6): 3492-3505.

Moskvina V., Schmidt K.M. (2008) On multiple-testing correction in genome-wide association studies. Genet. Epidemiol. 32: 567-573.

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.M., Muehlbaueret G.J. (2014) The USDA Barley Core Collection: Genetic diversity, population structure, and potential for genome-wide association studies. PLoS ONE 9 (4): e94688. doi:10.1371/journal.pone.0094688.

Olukolu B.A., Wang G.F., Vontimitta V., Venkata B.P., Marla S., Ji J., Gachomo E., Chu K., Negeri A., Benson J., Nelson R., Bradbury P., Nielsen D., Holland J.B., Balint-Kurti P., Gurmukh J. (2014) A Genome-Wide Association Study of the maize hypersensitive defense response identifies genes that cluster in related pathways. PLoS Genet 10: e1004562. doi:10.1371/journal.pgen.1004562.

Parisseaux B., Bernardo R. (2004) *In silico* mapping of quantitative trait loci in maize. Theor. Appl. Genet. 109:508–514.

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

Pers T. H., Karjalainen J. M., Chan Y., Westra H. J., Wood A. R., Yang J., LuiJ. C., Vedantam S., Gustafsson S., Esko T., Frayling T., Speliotes E.K. Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Boehnke M., Raychaudhuri S., Fehrmann R., Hirschhorn J., Franke L. (2015) Biological interpretation of genome-wide association studies using predicted gene functions. Nature communications 6: 5890.

Sabatti C., Service S., Freimer N. (2003) False discovery

rate in linkage and association genome screens for complex disorders. Genetics 164: 829-833.

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

Schwartzman A., Dougherty R.F., Taylor J.E. (2008) False discovery rate analysis of brain diffusion direction maps. The Annals of Applied Statistics 153-175. doi:10.1214/07-aoas133.

Sidak Z. (1967) Rectangular confidence regions for the means of multivariate normal distributions. Journal of the American Statistical Association 62: 626-633. doi:10.2307/2283989.

Spindel J., Begum H., Akdemir D., Virk P., Collard B., Redoña E., Atlin G., Jannink JL., McCouch S. (2015) Genomic selection and association mapping in rice (*Oryza sativa*): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. PLoS Genet.11: e1004982.

Tadesse W., Ogbonnaya F.C., Jighly A., Sanchez-Garcia M., Sohail Q., Rajaram S., Baum M. (2015) Genome-Wide Association Mapping of Yield and Grain Quality Traits in Winter Wheat Genotypes. Plos One 10(10): e0141339. https://doi.org/10.1371/journal.pone.0141339.

Team R.D.C. (2013) R: A language and environment for statistical computing, Vienna, Austria.

Tracy C.A., Widom H. (1994) Level-spacing distributions and the Airy kernel. Comm. Math. Phys. 159: 23.

Tusher V.G., Tibshirani R., Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences 98: 5116-5121.

Wang H., Smith K., Combs E., Blake T., Horsley R., Muehlbauer G. (2012) Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. Theor. Appl. Genet. 124: 111-124. doi:10.1007/s00122-011-1691-8.

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

Xiao J., Zhu W., Guo J. (2013) Large-scale multiple testing in genome-wide association studies via region-specific hidden Markov models. BMC Bioinformatics 14: 282.

Yan J., Warburton M., Crouch J. (2011) Association mapping for enhancing maize (*Zea mays* L.) genetic improvement. Crop Sci 51: 433–449.

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.

Zhou G. F., Broughton S., Zhang X. Q., Zhou M. X., Li C. D. (2016) Genome-wide association mapping of acid soil resistance in barley (*Hordeum vulgare* L.).Front. Plant Sci.7:406 10.3389/fpls.2016.00406.