# Ajustes de valores-P por multiplicidad en el contexto de datos dependientes y mapeo asociativo

Peña Malavera, Andrea Natalia
Gutiérrez, Lucia
Balzarini, Mónica Graciela

Ponencia presentada en el IV Encuentro Iberoamericano de Biometría. XVIII Reunión Científica del Grupo Argentino de Biometría. Mar del Plata, Argentina, 25 al 27 de septiembre de 2013.

# AJUSTES DE VALORES-P POR MULTIPLICIDAD EN EL CONTEXTO DE DATOS DEPENDIENTES Y MAPEO ASOCIATIVO

PEÑA MALAVERA ANDREA[1], GUTIERREZ LUCIA[2], BALZARINI MÓNICA[1]

*[1]Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, CONICET*
*[2]Facultad de Agronomía- Universidad de la República Uruguay*

*andreapema@gmail.com*

## RESUMEN

En mapeo asociativo se utilizan modelos lineales mixtos para evaluar la asociación entre los efectos de múltiples genes y el fenotipo de un individuo. Estos modelos para datos correlacionados han sido exitosamente utilizados ya que permiten contemplar información de la estructura poblacional y parentesco subyacente entre las unidades de análisis. El mapeo asociativo en especies vegetales pretende reconocer QTLs (de su nombre en inglés Quantitative Trait Loci) que codifican para variables de interés. Las pruebas de hipótesis realizadas gen-por-gen, o marcador-por-marcador, son múltiples y tienden a estar altamente correlacionadas cuando existe estructura genética de población, por lo que es necesario identificar una corrección apropiada para los valores p usados para declarar la significancia de la asociación. La corrección por multiplicidad propuesta por Bonferroni, la tasa de descubrimiento de falsos positivos y la estimación del número efectivo de pruebas independientes propuesto por Li y Ji (2005) son herramientas usadas para la corrección de los valores-p en el contexto del análisis de QTL clásico, donde los individuos se suponen igualmente emparentados. El objetivo de este trabajo es evaluar una nueva propuesta de corrección de valores p para el contexto de MA, que toma la idea del número efectivo de pruebas independientes pero éste es deducido luego de ajustar la estructura genética subyacente en las líneas de mapeo bajo diferentes modelos lineales mixtos para datos genéticamente correlaconados.

**Palabras clave: Estructura genética, *pruebas de hipótesis correlacionadas, modelos mixtos, número efectivo de pruebas de hipótesis.***

## Introducción

En estudios de mapeo asociativo (MA), se realizan múltiples pruebas de hipótesis simultáneamente con el fin de reconocer QTLs (de su nombre en inglés Quantitative Trait Loci) asociados a características fenotípicas de interés. Los estudios de MA con alta densidad de datos de marcadores moleculares del tipo SNP son cada vez más populares debido a la disponibilidad de las nuevas tecnologías de genotipado (e.g., Affymetrix e Illumina) con costos relativamente bajos para la producción de datos de marcadores. A su vez, los estudios de MA también son cada vez más frecuente en el mejoramiento genético vegetal porque con ellos es posible mejorar la potencia de las pruebas de asociación sin aumentar el costo de desarrollo de poblaciones específicas de mapeo. Cuando en el análisis de la asociación entre los SNP y el fenotipo no se tiene en cuenta la estructura genética subyacente, se incrementa la cantidad de falsos positivos entre los QTLs identificados. Se han propuesto varios procedimientos para ajustar las pruebas de asociación por la estructura genética subyacente. Un método comúnmente utilizado para identificar dicha estructura es el método bayesiano disponible en el software STRUCTURE [1]. Los modelos lineales mixtos han sido exitosa y ampliamente usados en mapeo asociativo para introducir factores como parentesco y estructura poblacional en el contexto de pruebas de asociación entre marcadores moleculares y el fenotipo [2-4].

Para declarar QTLs, es necesario realizar pruebas de hipótesis entre múltiples loci situados a través del genoma y el mismo fenotipo. Cuando muchas pruebas de hipótesis son

realizadas desde los datos, hay dos tipos de tasas de error relacionadas tanto a error tipo I como a error tipo II. La tasa de error tipo I por comparación es la probabilidad de error tipo I para cada prueba de hipótesis. La tasa de error tipo I por experimento es la probabilidad de tener al menos un error de tipo I (falso descubrimiento) al realizar el conjunto de pruebas de hipótesis. Una solución divulgada para controlar la tasa de error tipo I por experimento es la aproximación de Bonferroni [5]. Sin embargo, ésta es demasiado conservadora cuando se aplica a MA dado que pueden requerirse más de 1000 pruebas de hipótesis. Se han propuesto otros métodos para eludir el problema, por ejemplo, métodos que controlan la probabilidad de falso positivo, la que requiere consideración explícita de la probabilidad especificada a priori para cada hipótesis que se contrasta. Para pruebas independientes, Benjamini y Hochberg (1995) [6] propusieron un procedimiento para el control de la tasa de falsos descubrimientos (FDR). El método de control FDR se extendió a los casos de pruebas correlacionadas o dependientes bajo ciertas condiciones, por ejemplo, asumiendo que las pruebas estadísticas son igualmente correlacionadas positiva y normalmente distribuidas. Para pruebas correlacionadas, Benjamini y Yekutieli (2001) [7] propusieron un procedimiento simple pero altamente conservador. En 2001, Cheverud [8] propuso la idea de ajustar pruebas correlacionadas como si fueran independientes de acuerdo con un número efectivo (Meff) de pruebas independientes. Li y Ji (2005) [9] propusieron una estimación más precisa del número Meff, y el diseño basado en procedimientos Meff para controlar la tasa de error tipo I por experimento. La propuesta parte de la matriz de desequilibrio de ligamiento entre los marcadores para identificar redundancias. Sin embargo, esa propuesta se desarrolló en el contexto del análisis de QTL y no en el contexto de MA donde intervienen poblaciones estructuradas y por tanto no considera la correlación entre las pruebas de hipótesis ocasionadas por la presencia de estructura genética entre la líneas de mapeo. En este trabajo, proponemos una estimación del número efectivo de test basada en el uso de modelos lineales mixtos para el ajuste previo de las estructura de correlación subyacente. El nuevo método toma la idea propuesta por Li y Ji (2005), pero modifica la matriz de desequilibrio de ligamiento para los loci por una matriz dada por las pendientes resultante de regresiones lineales realizadas entre pares de marcadores que corrigen por la estructura genética subyacente, luego se estima el número efectivo de pruebas independientes realizando la descomposición de valores propios de esa matriz de coeficientes de regresión. El objetivo del trabajo es evaluar el desempeño del procedimiento propuesto para la selección de QTLs candidatos.

## Materiales y Métodos

Se simularon datos de marcadores moleculares con el programa QMSim [10] para poblaciones alógamas y con el programa Easypop [11] para poblaciones autógamas.

### Datos simulados poblaciones alógamas

Para datos de poblaciones alógamas los datos se simularon asumiendo poblaciones de polinización cruzada y caracterizadas por un genoma de 3000 marcadores bialélicos, otros parámetros de simulación fueron: 3 poblaciones, diseño de selección aleatoria, apareamiento aleatorio, dos niveles de diversidad genética entre poblaciones, uno y dos QTLs por cromosoma, tres niveles de heredabilidad (0.3, 0.5 y 0.8). En total se definieron 12 escenarios biológicos. Cada escenario fue replicado 100 veces y para cada uno de ellos se dispone de información sobre la posición verdadera de cada uno de los QTLs simulados.

### Datos simulados poblaciones autógamas

Para datos de poblaciones autógamas los datos se simularon asumiendo 2100 marcadores, 3 poblaciones sin cruzamientos aleatorios, proporción clonal igual a cero, autofecundación caracterizada por 1, un modelo de migración de islas jerárquico, dos valores de proporción de migración dentro de grupos: 0.2 y 0.1, proporción de migración entre grupos de 0.1, 0.2 y 0.3, tasa de recombinación entre loci de 0.03, y tasa de mutación igual a cero. Sobre los datos simulados en Easypop se seleccionaron aleatoriamente 10 marcadores de la matriz genética, en base a estas posiciones se generaron fenotipos distribuidos normalmente, este

proceso nos permite saber la ubicación de los marcadores que afectan a la realización del fenotipo e identificar posteriormente los falsos positivos

### Mapeo Asociativo

Se corrieron cinco modelos lineales mixtos propuestos en el contexto de mapeo asociativo para encontrar el efecto del marcador sobre el fenotipo de interés corrigiendo por relacionamiento genético en las siguientes formas: 1) utilizando covariables de efectos fijo o aleatorio (Q= los scores de los ejes significativos de un análisis de componentes principales y factores indicando pertenecía a grupos dados por la simulación); 2) incluyendo una matriz de parentesco con coeficientes de coancestría entre genotipos utilizando un modelo mixto con efecto aleatorio de genotipo (Z); 3) utilizando tanto una covariable de estructura genética como una matriz de parentesco También se hizo un análisis sin considerar correcciones por estructura o parentesco (modelo naive).

### Comparación de Métodos de detección de QTls

Se aplicaron a los resultados de los cinco modelos las correcciones de Bonferroni, Benjamini y Hochberg, Li y Ji y nuestra propuesta. Se calcularon las tasas de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos para cada corrección implementada.

## Resultados y Conclusiones

Se espera que el método de Li and Ji modificado propuesto en este trabajo permita disminuir la tasa de falsos positivos (FDR) y que sea menos conservador que los métodos de Bonferroni o Benjamini y Yekutielli. También se espera que el procedimiento sea más apropiado para poblaciones de MA, ya que tiene en cuenta la posible estructura poblacional que se encuentra en los genotipos objeto de análisis.

## Bibliografía


1. Pritchard, J., M. Stephens, y P. Donnelly, "*Inference of population structure using multilocus genotype data".* Genetics, 2000. **155**: p. 945-959.
2. Malosetti, M., et al., "*A Mixed-Model Approach to Association Mapping Using Pedigree Information With an Illustration of Resistance to Phytophthora infestans in Potato."* Genetics, 2007. **175**(2): p. 879-889.
3. Stich, B., Möhring J., Piepho H., Heckenberger M., Buckler E.S. y Melchinger A.E. , "*Comparison of mixed-model approaches for association mapping".* Genetics, 2008. **178**: p. 9.
4. Yu, J., et al., "*A unified mixed-model method for association mapping that accounts for multiple levels of relatedness".* Nat Genet, 2006. **38**(2): p. 203-208.
5. Bonferroni, C.E., "*Il calcolo delle assicurazioni su gruppi di teste".* Studi in Onore del Professore Salvatore Ortu Carboni. , 1935: p. 13-60.
6. Benjamini, Y. y Y. Hochberg., "*Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing."* Journal of the Royal Statistical Society., 1995. **57**(1): p. 11.
7. Benjamini, Y. y D. Yekutieli., "*The Control of the False Discovery Rate in Multiple Testing under Dependency."* The Annals of Statistics 2001. **29**(4): p. 23.
8. Cheverud, J.M., "*A simple correction for multiple comparisons in interval mapping genome scans."* Heredity, 2001. **87**(Pt 1): p. 52-58.
9. Li, J. y L. Ji, "*Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix."* Heredity, 2005. **95**(3): p. 221-227.
10. Sargolzaei, M. y F. Schenkel, "*QMSim: a large-scale genome simulator for livestock."* . Bioinformatics 2009. **25**: p. 680-681.
11. Balloux, F., "*Easypop (version 1.7): a computer program for population genetics simulations."* Journal of Heredity, 2001. **92**: p. 301-302.