UNIVERSIDAD NACIONAL DE CÓRDOBA

Facultad de Ciencias Exactas Físicas y Naturales

Tesis Doctoral

Aproximaciones estadísticas para el mapeo asociativo en estudios genéticos

Autora: Lic. Andrea Natalia Peña Malavera

Directora: Ing. Agr. (PhD) Mónica Graciela Balzarini

Mayo de 2015

Aproximaciones estadísticas para el mapeo asociativo en estudios genéticos

por

Lic. Andrea Natalia Peña Malavera

Ing. Agr. (PhD) Mónica Graciela Balzarini
Directora

COMISIÓN ASESORA

Ing. Agr. (PhD) Mónica Graciela Balzarini
FCA – UNC
Biol. (PhD) Cristina Noemí Gardenal
FCEFYN – UNC
Bioing. (PhD) Elmer Andrés Fernández
FCEFYN – UNC

Esta Tesis fue enviada a la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de Córdoba para cumplimentar los requerimientos de obtención del grado académico de Doctor en Ciencias de la Ingeniería.

Córdoba, Argentina
Mayo de 2015

UNIVERSIDAD NACIONAL DE CORDOBA
Facultad de Cs. Exactas, Fisicas y Naturales

ACTA DE EXAMENES

Libro: 00001 Acta: 03106 Hoja 01/01
LLAMADO: 1 07/07/2015
CATEDRA - MESA:

DI002 TESIS DOCTORADO EN CIENCIAS DE LA INGENIERIA

NUMERO	APELLIDO Y NOMBRE	DOCUMENTO	INGRESO	COND.	NOTA	FIRMA
1110446581	PEÑA MALAVERA, Andrea Natalia	Pas: 1110446581	2010	T	Aprobado	

RISK, Marcelo - MANGEAUD, Arnaldo Pedro - SMREKAR, Marcelo R. - ADROVER, Jorge G. -

Observaciones:

Marcelo Risk

Córdoba, ____/____/_____.
Certifico que la/s firma/s que ha/n sido puesta/s en la presente Acta pertenece/n a:_____

Inscriptos	Ausentes	Examinados	Reprobados (0-3)	Aprobados (4-10)
1	0	1	0	1

02/07/2015 10:43:20 Libro/Acta: 0000103106 Hoja: 01/ 01

A mis padres por darme más que la vida

A Duberney por su amor incondicional

A Julian Abel mi impulso de vida

AGRADECIMIENTOS

Agradezco a Dios porque sus bendiciones han sido grandes y por tener estos planes para mi vida haciéndome saber cada día que estaba conmigo.

Agradezco a mi directora de tesis, Dra. Mónica Balzarini por abrirme las puertas desde el primer momento no sólo en el trabajo sino también las de su casa, por su constante apoyo, seguimiento, asistencia y tiempo dedicado de manera generosa durante el desarrollo del presente trabajo, por su generosidad al brindarme la oportunidad de recurrir a su capacidad y experiencia científica permitiéndome crecer profesional y personalmente.

A los profesores Julio Di Rienzo, Margot Tablada, Laura González y Cecilia Bruno de la Cátedra de Estadística y Biometría por estar siempre a disposición para atender mis consultas. Al Dr. Fernando Casanoves por siempre estar. A la Dra. Lucia Gutiérrez de la Universidad de la República-Montevideo por compartir su conocimiento.

A mis compañeros y amigos becarios por transitar conmigo este camino.

A la Facultad de Ciencias Exactas, Físicas y Naturales y a la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba por brindar un espacio de trabajo.

Al Fondo para la Investigación Científica y Tecnológica (FONCyT) y al Consejo Nacional de Investigación Científica y Tecnológica (CONICET) por permitir llevar a cabo este trabajo de investigación a través del otorgamiento de las becas tipo I y tipo II, respectivamente.

Agradezco a los Miembros del Comité Asesor por gentilmente apoyarme y dedicar su valioso tiempo a la revisión de este trabajo. Agradezco a los Miembros del Comité Evaluador por aceptar gentilmente formar parte del tribunal y dedicar su tiempo a la revisión de este trabajo.

A mis padres Martha y Abel por sacrificar mi compañía para que yo cumpliera mis sueños. A mi gran amor Duberney por darme la mano en cada paso y hacerme feliz.

A toda mi familia que desde tierras lejanas creyeron en mí. A mis amigos y hermanos incondicionales por su aliento y porque siempre están.

¡MUCHAS GRACIAS!

RESUMEN

El mapeo asociativo (MA) o GWAS (por sus siglas en inglés, *Genome Wide Association Study*) es usado para encontrar lugares específicos del genoma relacionados con la variación de un carácter fenotípico. Es una práctica difundida en el mejoramiento vegetal, ya que posibilita el uso de poblaciones no diseñadas experimentalmente. Sin embargo, se ha detectado que en poblaciones con estructura genética (EG), la cantidad de falsos positivos en la asociación marcador-carácter puede aumentar significativamente. La modelación estadística que incorpora información sobre la estructura genética poblacional hace más eficiente el MA. Un objetivo de esta tesis es evaluar métodos estadísticos para identificar EG, usar dicha estructura en modelos de mapeo y realizar pruebas de hipótesis sobre la significancia de la asociación marcador-carácter. Se evalúan con este fin métodos multivariados, modelos lineales mixtos (MLM) y métodos de corrección de valor-p por multiplicidad. Como criterios de evaluación se usaron errores de clasificación de métodos orientados a identificar EG, tasas de falsos positivos, potencia estadística y distribución de valores-p para distintas combinaciones de modelos de MA y métodos de corrección por multiplicidad. El uso de mapas auto-organizativos (SOM, *Self-Organizing Maps*) y el algoritmo del software STRUCTURE fueron los más eficientes para identificar EG. La clasificación dada por STRUCTURE usada para contemplar EGP en el modelo de MA, disminuyó la tasa de FDR (*False Discovery Rate*), esta disminución fue mayor cuando estas estrategias se usaron simultáneamente con la matriz de relaciones de parentesco entre individuos como matriz de covarianza del MLM de mapeo. Se propuso un método de corrección de valores-p basado en la estimación del número efectivo de pruebas (pruebas no dependientes), similar al propuesto por Li y Ji (LJ, 1995) y que se denominó MLJ (*Modified Li&Ji*) y resultó más efectivo para disminuir FDR que con los métodos tradicionales Benjamini & Hochberg (1995) y Li & Ji (2005), en escenarios de alta divergencia, principalmente cuando la EGP no forma parte del modelo de MA.

Palabras clave: Estructura genética, Modelos Lineales Mixtos, Correcciones por Multiplicidad

ABSTRACT

The association mapping o GWAS (Genome wide association study), is used to find specific genomic regions related with variation of a phenotypic trait. It is a widespread practice in plant breeding since it allows the use of populations that do not require prior experimental design for its conformation. However, has been detected in genetically structured populations, the number of false positives in the marker-trait association may increase significantly. The analysis of population genetic structure is one of the important analyses previous to perform associative mapping. The statistical modeling that incorporates population genetic structure information makes more efficient the association mapping methodology. The objective in this thesis is to evaluate statistical methods to identify genetic structures, using them on mapping models and to test hypotheses about the significance of the marker-character association. Multivariate methods, linear mixed models and methods for multiplicity of p-values correction are evaluated. As evaluation criterion, classification error rate by methods for identify genetic structure, false discovery rate (FDR), statistical power and cumulative distribution functions of the p-values for each analysis strategy and multiplicity correction, were used. Use self-organizing maps SOM and STRUCTURE algorithm were more efficient to identify genetic structure. Incorporate STRUCTURE classification used to taking account de genetic structure in the association mapping model, decreases the false discovery rates, this was higher when the strategies were used with the relatedness kinship matrix between individuals simultaneously, as covariance matrix in the association mixed linear models. We proposed a method of correcting p-values based on the estimation of the effective number of tests (independent test), similar to that proposed by Li and Ji (LJ, 2005) and was called MLJ (Modified Li&Ji) and resulted more effective to decreases FDR than traditional methods Benjamini & Hochberg (1995) and Li and Ji, in scenarios with high divergence, mainly when genetic structure is not taking account in the association mapping model.

Key words: Genetic structure, Mixed linear models, multiple-testing corrections.

RESUMO

O mapeamento associativo (MA) ou GWAS (pela em inglês, *Genome Wide-Association Study*) é usado para encontrar lugares específicos do genoma relacionados com a variação de um carácter fenotípico. É uma prática difundida no melhoramento vegetal, pois possibilita o uso de populações não desenhadas experimentalmente. No entanto, se detectou que em populações com estrutura genética (EG), a quantidade de falsos positivos na associação marcador-caráter pode aumentar significativamente. A modelação estatística que incorpora informação sobre a estrutura genética faz mais eficiente o MA. Um objetivo desta tese é avaliar métodos estatísticos para identificar EG, usar tal estrutura em modelos de mapeamento e realizar testes de hipóteses sobre a significância da associação marcador-caráter. Avaliam-se com este fim métodos multivariados, modelos lineares mistos (MLM) e métodos de correção de valor-p por multiplicidade. Como critérios de avaliação, se usaram erros de classificação de métodos orientados a identificar EG, taxas de falso positivo, potência estatística e distribuição de valores-p para distintas combinações de modelos de MA e métodos de correção por multiplicidade. O uso de mapas auto organizativos (SOM, Self-Organizing Maps) e o algoritmo do software STRUCTURE foram os mais eficientes para identificar EG. A incorporação de componentes principais como co-variáveis de efeitos aleatórios e a classificação dada por STRUCTURE usadas para contemplar EGP no modelo MA, diminuíram a taxa de falsos positivos, que foi maior quando estas estratégias foram usadas simultaneamente com a matriz de relações de parentesco entre indivíduos com matriz de covariação do MLM de mapeamento. Propôs-se um método de correção de valores-p baseado na estimativa do número efetivo de testes (testes não dependentes) similar ao proposto por Li e Ji (1995) e que se denominou MLJ (Modified Li&Ji) e resultou efetivo para diminuir FDR (False Discovery Rate) os cenários de alta divergência, principalmente quando o EGP não faz parte do modelo de MA.

Palavras chave: Estrutura genética, Modelos Lineares Mistos, Correções por Multiplicidade.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE SÍMBOLOS Y ABREVIATURAS

ACP: Análisis de componentes principales

CER: *Cluster Error Rate*

CP: Componente principal

EGP: Estructura genética poblacional

FN: Falsos negativos

FP: Falsos positivos

GWAS: *Genome-Wide Association Study*

LD: Desequilibrio de ligamiento (*Linkage Disequilibrium*)

FDR: *False Discovery Rate*

MA: Mapeo Asociativo

ML: Máxima verosimilitud

MLM: Modelo lineal mixto

QTL: *Loci* de caracteres cuantitativos (*Quantitative Trait Loci*)

REML: Máxima verosimilitud restringida

RP: Posición relativa

SNP: *Single-Nucleotide Polymorphisms*

TRD: Técnicas de reducción de dimensión

VP: Verdaderos positivos

INTRODUCCIÓN GENERAL

La técnica de mapeo asociativo es un enfoque desarrollado para detectar genes de interés en plantas (Jannink *et al.,* 2001; Kraakman *et al.,* 2004; von Zitzewitz *et al.,* 2011). El mapeo asociativo (MA) o mapeo por desequilibrio de ligamiento (LD) (por sus siglas en inglés, *Linkage Disequilibrium*) es una herramienta analítica para asociar fenotipos (características observadas de los organismos) a genotipos (constitución genética del organismo). Se realiza una examinación de numerosas variantes genéticas comunes en una población de mapeo para identificar si alguna variante está asociada con una característica fenotípica del individuo que interese mapear sobre el genoma. Las asociaciones identificadas pueden contribuir directamente al carácter de interés o pueden estar ligadas (i.e. en desequilibrio de ligamiento) con uno o más *loci* de caracteres cuantitativos (QTL, del inglés *Quantitative Trait Loci*) que contribuyen a explicar la variabilidad del fenotipo de interés.

El LD se define como el grado de asociación no aleatoria entre alelos de distintos *loci* en poblaciones de individuos no relacionados (Yu y Buckler, 2006); se relaciona con la proporción de gametos que no segregan al azar y provee información sobre la historia de la población así como sobre el sistema de selección implementado. A nivel genómico, refleja el impacto de fuerzas evolutivas (selección natural, mutación y migración) que causan cambios en las frecuencias génicas (Lynch y Walsh, 1998; Falconer y Mackay, 1996). En genética, ligamiento indica proximidad física entre genes de un mismo cromosoma, tal cercanía durante el proceso de recombinación deriva en que dichos genes sean heredados siempre juntos (Wu *et al., 2*007). El objetivo del mapeo asociativo es poder identificar el ligamiento físico entre los alelos de dos *loci*, o entre el alelo de un marcador molecular y un QTL y a la vez diferenciarlo de la asociación meramente estadística (Jannink y Walsh, 2002).

El MA ha sido inicialmente utilizado en genética humana (Corder *et al.,* 1994) pero luego se extendió para tratar la identificación de *loci* de interés en la genética de plantas (Thornsberry *et al.,* 2001). La técnica puede aplicarse a cualquier colección de germoplasma, aunque el grado de relacionamiento entre los genotipos sea desconocido, y a diferencia de los métodos tradicionales, no es necesario contar con una población

segregante derivada de cruces entre líneas puras lo cual implica un mayor costo y trabajo previo (Roy *et al., 2*010). En la actualidad el análisis de datos genómicos está a disposición para una gran variedad de especies vegetales, una herramienta importante al momento de estudiar estos datos es el análisis estadístico, el cual explica las variaciones genéticas en rasgos cuantitativos entre diferentes poblaciones e individuos (Wu *et al., 2*007).

El mapeo asociativo constituye una estrategia eficaz para estudiar *loci* que rigen caracteres complejos en un contexto de disponibilidad de abundante información genómica, como es la producida por distintos tipos de marcadores moleculares. Los marcadores del tipo SNP (*Single-Nucleotide Polymorphisms*) (Breen *et al.,* 2000) convocan especial atención (Bernardo, 2007). El MA permite mapear caracteres cuantitativos con alta resolución y de manera potente desde una perspectiva estadística. Cuando un método de mapeo genético requiere demasiado conocimiento sobre los marcadores moleculares y el genoma en estudio, su aplicación puede resultar restrictiva para especies poco estudiadas molecularmente. El análisis de QTL y estudios de asociación genómica (*Genome-Wide Association Studies*, GWAS), un tipo de MA que se realiza usando SNP del genoma entero, son técnicas hoy comunes en el mejoramiento genético vegetal, principalmente para identificar QTL con efectos mayores (Bernardo, 2014).

Desde una perspectiva analítica, el objetivo del MA es evaluar la significancia estadística de las asociaciones entre la información que proveen los marcadores moleculares (variantes genéticas) y la característica fenotípica de interés. Los MA, realizados a nivel poblacional y a partir de frecuencias alélicas, tasas de polimorfismos y LD, han permitido identificar asociaciones reproducibles en numerosas especies vegetales (Flint-García *et al.*, 2005).

Una de las principales ventajas del MA es que no requiere del desarrollo de una población específica de mapeo. Es posible realizarlo a través de un conjunto de genotipos como pueden ser un subconjunto de líneas de un banco de germoplasma (población genotípicamente diversa). Por el contrario, el análisis clásico de QTL se basa en frecuencias de recombinación dentro de poblaciones de mapeo especialmente diseñadas. Mientras el análisis clásico de QTL explota la herencia compartida de polimorfismos funcionales y marcadores moleculares adyacentes, dentro de familias o pedigrís de ancestría conocida o en poblaciones de mapeo derivadas de cruzamientos de dos padres (Yu y Buckler, 2006a), el MA explora asociaciones estadísticas entre variaciones del

genotipo y fenotipo, algunas de ellas pueden ser reales y otras simplemente pueden darse por azar o por la presencia de estructuras en los datos que sugieren correlación. El MA analiza también la herencia compartida de una colección de individuos sin ancestría conocida, donde ha habido recombinación (Yu y Buckler, 2006a), la existencia de múltiples generaciones de recombinación conlleva a una mayor resolución de mapeo con mayor oportunidad de disipación del LD. Los atractivos del mapeo con más alelos que los existentes en las poblaciones biparentales del análisis de QTL clásico consisten en la posibilidad de identificar *loci* ligados a características fenotípicas de interés en cultivos con difíciles patrones de segregación. En síntesis, las ventajas ofrecidas por el MA consisten en mejorar la resolución de estudios de asociación entre marcadores y fenotipo mediante la consideración de una población más amplia, representante del germoplasma elite, que involucra un mayor número de alelos. Ambos análisis, análisis de QTL clásico y MA, se complementan en términos de proveer información relevante, validar cruzamientos y aumentar la potencia estadística en estudios de asociación (Yu y Buckler, 2006a).

Los principales mecanismos que provocan el LD son la mutación y la deriva, mientras que la recombinación reduce el LD (Jannink *et al.,* 2009). Por esto, se explicita que para obtener una exitosa detección de asociaciones la densidad de marcadores debe coincidir con el rango de decaimiento del LD (Jannink *et al.,* 2009). Por ejemplo, si el LD decae de forma rápida, una mayor densidad de marcadores se requiere para capturar marcadores lo suficientemente cerca de los sitios funcionales (Yu y Buckler, 2006a). Dado que el MA depende de la extensión y la distribución del LD entre los marcadores involucrados, resulta necesario calcular empíricamente la magnitud de la correlación entre marcadores, previo a la realización de cualquier análisis. Técnicas gráficas como el *heatmap* (R Core Team, 2013) permiten visualizar patrones de la magnitud de cientos y miles de correlaciones como son las que ocurren en mapeos con alta densidad de marcadores.

Aun pensando en un conjunto de marcadores independientes es importante tener en cuenta que en todo estudio de asociación la significancia estadística de la correlación entre estado de los marcadores y la característica en estudio pueden no ser consecuencia de un ligamiento real entre marcadores y *loci* de interés, es decir generar falsos positivos. Una causa importante de falsos positivos es la correlación entre la información que proveen los individuos generada por el relacionamiento filial entre los genotipos de la colección de estudios (falta de independencia entre los individuos en estudio). Por ejemplo, se supone

que si una mutación incrementa la observación de una característica en una población de individuos, entonces podemos esperar que el alelo asociado a tal característica sea más frecuente entre los individuos emparentados que entre el resto de los individuos. Luego, antes de iniciar estudios de MA será menester estudiar e identificar si existiesen las estructuras que subyacen los datos genéticos en la población de mapeo. Cuando se lleva a cabo un análisis de asociación basado en poblaciones, sin considerar los efectos de la estructura poblacional, se aumenta el riesgo de detectar asociaciones espurias entre marcadores y el fenotipo de interés. En las poblaciones estructuradas puede haber una alta proporción de asociaciones significativas aún cuando muchos marcadores no estén ligados a ningún locus del carácter de interés aumentando la probabilidad de error tipo I (falsos positivos).

Se han desarrollado diferentes estrategias para considerar la relación genética que implica la detección de la estructura de la población y la incorporación de estos agrupamientos en el modelo estadístico (Pritchard *et al.,* 2000; Kraakman *et al.,* 2004). Para el análisis de estructura genética se usan distintos métodos multivariados (Odong *et al.,* 2011), métodos de clasificación difusa de naturaleza bayesiana como los disponibles en el software STRUCTURE (Pritchard *et al.,* 2000) y menos frecuentemente métodos bioinformáticos como aquellos basados en la teoría de redes neuronales (Kohonen, 1997).

Para incrementar la potencia o resolución del MA y disminuir la detección de asociaciones espurias se han utilizado distintos tipos de modelos de asociación, principalmente modelos de regresión carácter vs. marcadores ajustados en el marco teórico de los modelos lineales mixtos (MLM) (Demidenko, 2004; Malosetti *et al.,* 2007) y la estimación REML (Patterson y Thompson, 1971). En el modelo de MA se especifica la relación entre cada marcador y el fenotipo con el propósito de detectar los marcadores significativos, i.e. los potencialmente ligados a un QTL.

Yu *et al.,* (2006) propusieron como modelo de MA, el modelo mixto denominado QK para mapeo de asociación, el cual pretende corregir el LD causado por estructuras poblacionales y relaciones filiales existentes en la población de mapeo, mediante el uso de una matriz Q de estructura poblacional, calculada por el software STRUCTURE (Pritchard, 2000) y que está conformada por las probabilidades de pertenencia a los k grupos. STRUCTURE calcula para cada individuo, la probabilidad de pertenecer a cada una de las subpoblaciones que conforman la población, tal que la sumatoria de las probabilidades de pertenencia por

definición de función de probabilidad da uno. Otra estrategia de modelación, también factible en el contexto del ajuste de un MLM es el uso de la matriz K (Kinship) de parentesco que puede ser provista por el software específico SPAGeDi (Hardy y Vekemans, 2002), por la librería EMMA de R (Kang *et al.*, 2008) o por el software TASSEL (Bradbury *et al.*, 2007); la matriz K es luego incorporada a la estructura de varianzas y covarianzas del MLM. Otro modelo mixto desarrollado para mapeo de asociación es el denominado PK (Zhao *et al.*, 2007) el cual contempla la misma matriz de parentesco K, pero con una matriz distinta a Q como la obtenida desde STRUCTURE para reflejar la estructura poblacional. Otra propuesta frecuentemente usada, es la utilización de componentes principales de los datos genéticos como covariables del modelo de mapeo (Patterson *et al.*, 2006), en este modelo conocido como PK, la estructura se sintetiza en el conjunto de componentes principales de un Análisis de Componentes Principales (ACP; Hotelling 1936) realizado sobre la matriz de frecuencias alélicas. Con ACP es posible obtener un conjunto de nuevas variables, que se genera como combinación lineal de los datos moleculares originales denominadas variables sintéticas o componentes principales (CPs). Tales variables tienen la característica de ser no correlacionadas y óptimas para señalar variabilidad y estructuras entre los genotipos de la población de mapeo. Así, las CPs conformadas a través de la combinación de marcadores moleculares de los genotipos de la población de mapeo permiten señalar diferencias entre genotipos causadas por la existencia de estructuración genética. Las primeras CPs o aquellas estadísticamente significativas, según la prueba de Tracy-Widom (1994), pueden ser usadas como covariables en el modelo de MA. Estas covariables pueden contemplarse en el modelo de mapeo como efectos fijos o aleatorios. Estos desarrollos han generado distintas alternativas de modelación para MA. Si bien existen algunos estudios donde se compara su performance (Gutiérrez *et al.*, 2011; Cappa *et al.*, 2013), aún es necesario desarrollar investigaciones que permitan sugerir o recomendar una u otra aproximación en función de los niveles de LD y de estructura poblacional subyacente en la población de mapeo.

Luego de ajustar el modelo de MA suele aplicarse algún método de corrección de "valores-p" por multiplicidad. Esto es debido a que las pruebas de hipótesis se realizan marcador por marcador y por tanto hay tantos contrastes estadísticos como marcadores haya. La probabilidad de detectar significancias solo por azar aumenta con la acumulación de pruebas de hipótesis realizadas sobre los mismos datos. Distintos métodos de corrección de

valores-p por multiplicidad han sido diseñados para favorecer la adhesión al nivel nominal de las pruebas de hipótesis realizadas y así incrementar la potencia de detección de QTL. La corrección de valores-p por multiplicidad, suele realizarse con los métodos Bonferroni (1935), Benjamini y Hochberg (BH, 1995), Benjamini y Yekutielli (BY, 2001) para pruebas independientes. Tales procedimientos de ajuste de la significancia estadística probaron ser útiles en el análisis de QTL clásico donde los genotipos son independientes. Sin embargo, en MA con poblaciones estructuradas de diferente nivel de parentesco, estas correcciones pueden no desempeñarse apropiadamente. El método propuesto por Li &Ji (LJ, 2005) para pruebas correlacionadas, se ha usado en estos escenarios pero sobre conjuntos de hipótesis que si bien no se suponen independientes, están igualmente correlacionadas.

No sólo las estimaciones de variabilidad pueden cambiar en un contexto de datos correlacionados sino que también se pueden ver afectadas las significancias de las asociaciones entre los datos genéticos y el fenotipo. Los procedimientos analíticos que se deben utilizar para mapeo deben considerar la presencia de tales correlaciones en los datos. Es así, como el cuerpo de análisis estadísticos utilizados en estudios de MA es muy variado y muchas veces no es claro para el investigador qué estrategia de modelación es más conveniente usar para analizar estadísticamente sus datos moleculares en combinación con los caracteres fenológicos de interés. Existen trabajos publicados con una gama muy amplia de aproximaciones metodológicas para el fin del mapeo. Este hecho constituye una evidencia de la necesidad que aún existe de investigar el uso de técnicas de análisis estadístico en mapeo genético (Aranzana *et al*., 2005; Breseghello y Sorrells, 2006; D'hoop *et al*., 2008, Cappa *et al*., 2013; Comadran *et al*., 2009, Gutiérrez *et al*., 2011; Locatelli *et al*., 2013; von Zitzewitz *et al*., 2011). Asimismo, existen numerosos estudios de simulación que han sido diseñados para responder preguntas específicas sobre el desempeño, desde criterios estadísticos más que biológicos, de metodologías que alternativamente pueden utilizarse para el mapeo y por tanto para obtener recomendaciones sobre el método más apropiado para el análisis de una situación específica (Wang *et al*., 2012, Feng *et al*., 2013, Bernardo, 2014, Wang *et al*., 2014)

Para elegir el método analítico más apropiado para contestar una pregunta referida a estructura genética es necesario analizar y diferenciar diversos aspectos del problema en cuestión, como son los niveles de estructuración (Excoffier *et al*., 2009, Comadran *et al*.,

2009, Cappa *et al*., 2013), el tamaño de la población de mapeo (Wang *et al*., 2012, Bernardo 2013) y la intensidad de marcadores en uso (Bernardo *et al*., 2008), entre otros. Luego, habrá que introducir esta información en un modelo de asociación que contemple dicha estructura y seleccionar un método de corrección de valores-p por multiplicidad para finalmente, detectar las asociaciones con mayor probabilidad de ligamiento. En esta tesis se hipotetiza que el desempeño de las estrategias de MA alternativas que surgen de la combinación de tales selecciones de métodos, puede depender de factores tales como los niveles de LD, las cantidades de grupos o subpoblaciones que definan la estructura de la población de mapeo y el nivel de divergencia genética entre tales subpoblaciones. En el Capítulo I, se realizará un marco teórico de los métodos y modelos estadísticos usados en un estudio de mapeo asociativo, que incluye análisis multivariado, modelos mixtos, modelos de mapeo y ajustes por multiplicidad. En el Capítulo II, se realiza una comparación de métodos estadísticos, bayesianos y bioinformáticos, para el reconocimiento de estructura genética poblacional en datos de marcadores moleculares simulados y reales de maíz. En el Capítulo III se evaluan modelos mixtos en mapeo asociativo y se realiza una comparación de desempeño en base a tasas de falsos positivos y potencia en datos de marcadores moleculares simulados con diferentes niveles de estructura genética, número de marcadores, tamaño poblacional y en un conjunto de datos reales del tipo SNP con un carácter fenotípico simulado. En el Capítulo IV se escogen los modelos más representativos según el capítulo previo y se prueban y comparan métodos de corrección por multiplicidad adicionando una nueva propuesta, que corrige los valores-p por multiplicidad, teniendo en cuenta la estructura genética poblacional. Por último se presentan conclusiones generales, comentarios finales y futuras investigaciones.

OBJETIVO GENERAL

Contribuir al mapeo asociativo en vegetales mediante el aporte de conocimientos sobre estrategias estadísticas para el estudio de asociaciones entre marcadores moleculares y características fenotípicas de interés agronómico bajo distintos escenario de poblaciones genéticamente diversas.

OBJETIVOS ESPECÍFICOS

1. Comparar e ilustrar, desde su aplicación en escenarios de diferenciación genética contrastante, la información obtenida desde distintos métodos estadísticos para detección de estructura genética.

2. Evaluar estrategias de modelación alternativas para el mapeo asociativo en plantas incluyendo modelos de mapeo con corrección por estructura genética.

3. Proponer y evaluar métodos para la corrección de valores-p por multiplicidad que puedan ser usados en el diseño de un protocolo de análisis estadístico recomendable para realizar mapeo asociativo en plantas.

CAPITULO I

MÉTODOS Y MODELOS ESTADÍSTICOS USADOS EN MAPEO ASOCIATIVO

INTRODUCCIÓN

Este trabajo aborda el análisis de estructura genética poblacional (EGP) en un conjunto de genotipos y la modelización estadística para la detección de asociaciones entre variantes genéticas y fenotipo desde una perspectiva metodológica. Los perfiles marcadores moleculares de una muestra de ADN son de naturaleza multidimensional y para su análisis demandan métodos multivariados y algoritmos desarrollados para encontrar e interpretar patrones de los datos. En este capítulo se describen brevemente técnicas multivariadas (Johnson y Wichern, 1998) usadas en la etapa del reconocimiento de la EGP subyacente en la población de mapeo.

El mapeo asociativo está orientado a localizar lugares del genoma que impactan directamente la variación de un carácter fenotípico, tiene la ventaja respecto al análisis de QTL clásico de no demandar la conformación de poblaciones de mapeo derivadas de cruzamientos biparentales específicos. Sin embargo, se ha detectado que cuando se usan genotipos con alta estructura molecular, la cantidad de falsos positivos (FP) en los estudios de asociación marcador-carácter aumenta significativamente (Malosetti *et al.* 2007). El modelado de asociaciones marcador-fenotipo incluyendo covariables que reflejen la estructura o estableciendo relaciones de parentesco entre individuos, puede realizarse de distintas maneras. El modelo base en la etapa de modelación estadística de asociaciones es un modelo de regresión lineal. En este Capítulo se define este modelo y sus derivados en el contexto teórico de los modelos lineales mixtos (MLM) (Demidenko, 2004); familia de modelos que por su flexibilidad para modelar correlaciones en los datos está siendo usada en el contexto de mapeo asociativo bajo estructura genética poblacional. En esta tesis se

evalúan modelos lineales mixtos (MLM) alternativos para llevar a cabo este proceso. Complementariamente y dado que se analizan numerosas pruebas de hipótesis (una por marcador), se describen aquí algunos de los principales métodos usados en la corrección de valores-p por multiplicidad.

MÉTODOS MULTIVARIADOS Y ESTRUCTURA GENÉTICA

La estadística multivariada es usada para describir y analizar observaciones multidimensionales o multivariadas. Una observación multidimensional se obtiene cuando se releva información sobre varias variables para cada individuo en estudio. La estadística multivariada provee herramientas para comprender la relación (dependencia) entre variables medidas simultáneamente sobre un mismo individuo, para comparar, agrupar y/o clasificar observaciones multivariadas e incluso para comparar, agrupar y clasificar variables. Gran parte de la metodología multivariada se basa en los conceptos de distancia y de dependencia lineal. Las distancias serán usadas como medidas de variabilidad entre pares de puntos que representan los datos multivariados y a partir de ellas es posible analizar similitudes y diferencias entre observaciones y/o variables.

Las aplicaciones del análisis multivariado en datos genéticos son numerosas y diversas lo que ha permitido que sirvan para resolver interesantes problemas (Balzarini *et al.,* 2011). El análisis multivariado provee las herramientas para clasificar y ordenar especies vegetales en función de múltiples características o descriptores; a nivel molecular los estudios genómicos se basan en el reconocimiento de patrones de variabilidad en los espacios mega-dimensionales que genera la información de cientos y miles de marcadores moleculares, esta información debe ser resumida y posteriormente analizada conjuntamente con información proveniente de otros tipos de estudios que se realizan sobre el mismo material.

Para explicar que significa “análisis multivariado” se encuentran en la literatura distintas definiciones, la dada por Johnson y Wichern (1998) dice que el análisis multivariado es una bolsa mixta que contiene métodos apropiados para investigaciones científicas y tecnológicas donde los objetivos son uno o varios de los siguientes: reducción de dimensionalidad, agrupamiento y clasificación, investigación de la dependencia entre

variables, predicción y prueba de múltiples hipótesis. Se prueban hipótesis estadísticas específicas, formuladas en términos de los parámetros de distribuciones multivariadas. Tim y Mieczkowski (1997) discuten y comparan aproximaciones de procedimientos inferenciales univariados y multivariados. Los desarrollos teóricos en análisis multivariado, que sustentan casi la totalidad de numerosas aplicaciones modernas, se produjeron en el siglo XX con una fuerte entrada de las escuelas hindú y norteamericana. Muchas técnicas del análisis multivariado constituyen las principales herramientas del proceso que, en el nuevo milenio, se ha dado a conocer como "knowledge discovery in data bases" (Fayyad *et al*., 1996) y "data mining" (Han y Kamber 2006) y que se sustenta en la capacidad de cálculos intensivos de las nuevas computadoras. El cual involucra además algoritmos de la familia de las redes neuronales (Hecht-Nielsen, 1989).

En el contexto de los datos de marcadores genéticos, el análisis multivariado provee métodos que permitan obtener distancias entre entidades biológicas (Bruno *et al.*, 2003) para realizar análisis de similitudes y diferencias moleculares. Las técnicas multivariadas diseñadas con el objetivo de reducción de dimensión (TRD) se usan para extraer información de datos genéticos provenientes desde múltiples *loci*, desde hace ya más de cuarenta años (Cavalli-Sforza 1966; Smouse *et al.*,1982). Las TRD, permiten resumir la información genética multivariada en pocas variables sintéticas (Balzarini *et al., 2*011) que luego son usadas con fines específicos como puede ser el análisis de estructura genética poblacional. Las TRD más usadas son el Análisis de Componentes Principales (ACP) y el Análisis de Coordenadas Principales (ACoorP) las cuales presentan varias ventajas respecto a otras aproximaciones multivariadas como los algoritmos de agrupamientos. Por ejemplo, comparando las TRD con los métodos bayesianos de agrupamiento disponibles en el software STRUCTURE, debe destacarse que las primeras no requieren supuestos sobre los modelos genéticos subyacentes, como el equilibrio de Hardy -Weinberg o la ausencia de desequilibrio por ligamiento ya que pueden aplicarse igualmente en situaciones donde las variables (marcadores o alelos) están correlacionadas. Las TRD computacionalmente son menos intensivas, por lo que pueden aplicarse a bases de datos masivos, con miles de marcadores (Patterson *et al., 2*006; Teich *et al., 2*011b), de manera relativamente simple.

Cuando se trabaja con datos de marcadores genéticos multilocus-multialélicos interesa estudiar las relaciones entre *n* entidades, que como señalamos anteriormente pueden ser individuos o poblaciones. Cada alelo representa una dimensión de estudio por lo que las

entidades de análisis se pueden representar como puntos en un espacio multidimensional de *p* (número de alelos) dimensiones. En el ACP (Hotelling, 1936) se busca encontrar la dirección, en este espacio multidimensional, en la cual las entidades estén lo más dispersas posible, es decir, la dimensión de máxima variabilidad entre entidades. Luego, en vez de analizar la variabilidad multidimensional entre entidades de análisis, se representa la variabilidad a nivel de las interdistancias de las proyecciones de las unidades de análisis en el eje que sigue esa dirección de máxima varianza. Esta nueva dimensión se conoce como el primer eje o componente principal 1 (CP1) del ACP y está definido por *p* coordenadas que representan los pesos de cada alelo en la definición del valor de la variable sintética CP1. A través del algoritmo de descomposición del valor singular de la matriz de covarianza o correlación de los datos originales, se buscan los ejes (CPs) subsecuentes que maximicen la explicación de la varianza total condicionados a que sean ortogonales a sus ejes previos (es decir cada eje aporta nueva información sobre la variabilidad total). Cada CP está asociada a un valor singular (*eigenvalue*) que proporciona una medida de la cantidad de varianza total explicada por cada eje, que a su vez resulta un buen estimador del porcentaje de varianza total explicado por un subconjunto de CPs que podría ser utilizado en lugar de las *p* variables para reducir la dimensionalidad del problema.

Para sintetizar información molecular de múltiples *loci* y alelos también puede usarse el Análisis de Coordenadas Principales (ACoorP), el cual es una forma de escalamiento multidimensional métrico o clásico (Gower, 1967). A diferencia del ACP, que preserva la distancia Euclídea canónica entre observaciones cuando se aplica a datos continuos, el ACoorP puede utilizarse con otros tipos de datos como los cualitativos ya que opera sobre una matriz de distancia y no sobre una matriz basada en varianzas y covarianzas como el ACP. Así se puede usar para resumir los datos multivariados mediante cualquier distancia entre individuos, incluso métricas basadas en datos cualitativos. Esto representa una ventaja porque se pueden utilizar métricas de diferenciación que estén relacionadas a un modelo poblacional genético en particular o aquellas que sean las más apropiadas según el sistema de marcadores moleculares utilizado. Sin embargo, se pierde la capacidad de representar a las variables en el mismo espacio que las observaciones, como se hace en el ACP, y de analizar los pesos de cada variable (alelo) sobre cada variable sintética. Las CPs, conformadas a través de la combinación de marcadores moleculares de los genotipos

de una población de mapeo pueden ser usadas para señalar diferencias entre genotipos causadas por la existencia de estructuración genética (Patterson, 2006).

En el contexto de alta disponibilidad de datos de marcadores genéticos, métodos que permitan no sólo obtener distancias entre entidades biológicas y realizar estudios de variabilidad genética sino también para clasificarlas. Numerosas investigaciones sobre tipo, abundancia y distribución de los organismos necesitan identificar la estructura subyacente en los datos, es decir el agrupamiento o conglomeración de las entidades en grupos (o clusters) relativamente homogéneos. Por ello, es común el uso de métodos de clasificación, tanto no supervisada (sin conocimiento a priori del análisis de los agrupamientos subyacentes) como supervisada (con conocimiento a priori de la existencia de agrupamientos de los datos). La clasificación no supervisada de genotipos dentro de grupos ya sean éstos de naturaleza discreta o difusa constituye una herramienta comúnmente aplicada en genética de plantas (Balzarini *et al.,* 2011; Peña Malavera *et al.*, 2014a).

Los algoritmos de conglomerado no supervisado son aquellos que demandan información previa sobre los grupos en los que se espera que los individuos se clasifiquen. La mayoría de estos métodos de conglomeración se usan en complementación con cálculos estadísticos desarrollados para estimar la cantidad de conglomerados que subyacen en la estructura de los datos poblacionales y que son de interés identificar (Balzarini *et al.,* 2008). Los conglomerados jerárquicos también son comúnmente aplicados para identificar EG, ya sea directamente sobre los datos moleculares (Odong *et al., 2*011), o luego de realizar un análisis de componentes principales sobre la información molecular para contemplar la correlación entre marcadores y evitar descartar aquellos muy correlacionados (Patterson *et al., 2*006).

Los conglomerados jerárquicos y no-jerárquicos son aplicados en una estrategia de dos pasos. Primero, es necesaria la construcción de una matriz de distancias y luego la conglomeración de individuos desde esa matriz y las matrices de distancia entre pares de individuos y conglomerados que se van recalculando en cada paso de la conglomeración (Balzarini *et al.,* 2008). Estos métodos basados en distancias pueden usar diferentes métricas de similitud multivariada entre pares de genotipos.

La similitud entre un par de genotipos puede depender no solo de la constitución genética, sino también de la del resto de la muestra a través de frecuencias alélicas (Weir y Ott,

1997). El promedio de la distancia genética entre dos genotipos es una interpretación simple del grado de relacionamiento (McVean, 2009). La distancia euclídea al cuadrado entre perfiles de marcadores refleja la cantidad de alelos no idénticos por estado entre dos genotipos y constituye una métrica para el análisis de conglomerados con datos moleculares. Aunque otras distancias basadas en índices de similitud para datos binarios son las más recomendadas (Bruno y Balzarini 2010). La selección de una medida de distancia para usar los métodos de conglomerados basados en distancia será también dependiente de la forma en que se codifique la información molecular (Bruno, 2009).

Otro método de conglomerados usual en la búsqueda de EGP es el de clasificación bayesiana implementado en el software STRUCTURE (Pritchard *et al., 2*000). Dada su naturaleza de clasificación difusa (asigna probabilidades de ocurrencia a uno u otro grupo en lugar de asignar un individuo a sólo un grupo) es apropiado para colecciones de germoplasma con alto nivel de relacionamiento. El software STRUCTURE produce una visualización de los conglomerados resultantes por medio de un gráfico de barras (barplot); cada individuo del conjunto de datos es representado en el gráfico por una barra vertical, la cual es particionada en segmentos de colores (uno por grupo) representando la probabilidad de pertenencia a un conglomerado específico.

No sólo las estimaciones de variabilidad pueden cambiar en un contexto de datos correlacionados como los que podría generar una estructuración genética de la población. También se ve afectada la significancia de las asociaciones de interés. Los procedimientos analíticos que se deben utilizar según los objetivos que se persigan, son diferentes. Es así, como el cuerpo de análisis estadísticos utilizados en estudios de mapeo asociativo es muy variado y muchas veces no es claro para el investigador qué método es más conveniente usar para analizar estadísticamente un problema biológico específico. Esta situación ha conducido a discusiones sobre la selección de metodologías de análisis multivariado y modelación de datos genéticos (Guillot *et al., 2*009, Balzarini *et al.*, 2010). La gran cantidad de trabajos biológicos que discuten las aproximaciones metodológicas sobre análisis de estructura genética, es una evidencia de la necesidad de investigaciones en estadística genética (Vekemans y Hardy 2004; Guillot *et al.*, 2009; Jombart *et al., 2*009b; Francois y Durand 2010; Segelbacher *et al., 2*010; Balzarini *et al., 2*011). Asimismo, existen numerosos estudios de simulación que han sido diseñados para responder preguntas específicas sobre el desempeño, desde criterios estadísticos más que biológicos, de

metodologías que alternativamente pueden utilizarse para un mismo problema y por tanto para obtener recomendaciones sobre el método más apropiado para el análisis de un conjunto específico de datos genéticos (Guillot y Santos, 2009; Bruno, 2009; Guillot y Rousset, 2011; Teich, 2012, Peña Malavera *et al.*, 2014a)

MODELOS LINEALES MIXTOS EN MAPEO GENÉTICO

Los modelos estadísticos conocidos como Modelos Lineales Mixtos (MLM) (Eisenhart, 1947; West *et al.*, 2007) son comúnmente utilizados en el análisis de datos biológicos porque éstos consideran por un lado factores de efectos fijos, como por ejemplo el efecto de distintas variables que explican la respuesta, y por otro factores de efectos aleatorios, como por ejemplo el efecto de unidad experimental o de grupo-cluster de unidad experimental en un estudio. Numerosos problemas de mapeo de QTL han sido ya tratados empleando modelos lineales mixtos de regresión (Malosetti *et al., 2*004; Bradbury *et al., 2*011; Zhang *et al., 2*010). Los MLM son útiles para ajustar los valores fenotípicos según las estructuras experimentales subyacentes (Balzarini *et al.*, 2001). Las medias ajustadas (por la correlación existente) de fenotipo son luego usadas en el análisis de asociación fenotipo-genoma, mejorando así el desempeño de las pruebas estadísticas orientadas a detectar y cuantificar las asociaciones.

La estimación de parámetros en los MLM se hace por métodos basados en la verosimilitud: máxima verosimilitud (ML) o por máxima verosimilitud restringida (REML) (Searle *et al.*, 1992). La mayor ventaja de los MLM es la generalidad en la inferencia luego de modelar la correlación entre los individuos. El MLM ajustado puede contemplar la correlación existente en los datos mediante la incorporación de efectos aleatorios desde los cuales se inducen correlaciones o, alternativamente, con la especificación explícita de la estructura de correlación en la matriz de varianzas-covarianzas de los términos aleatorios del modelo (Balzarini *et al.*, 2004).

La formulación del modelo lineal general de regresión establece que dadas *n* observaciones $y_i, y_2, ..., y_n$ y *p* variables explicativas el modelo para la observación i-ésima puede expresarse como:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad (i = 1, ..., n)$$

En este modelo los coeficientes betas son los coeficientes de regresión que expresan la relación parcial de cada variable predictora con la respuesta. El mismo modelo puede expresarse matricialmente de la siguiente forma:

$$y = X\beta + \varepsilon$$

con *y* vector fila de *n* observaciones, *X* matriz de dimensión $n \times p$, cuyos elementos dependen del valor que asume cada variable regresora en cada individuo, β vector de parámetros de efectos fijos de dimensión *p* y $\varepsilon \sim N(0, \sigma^2 I)$ el vector de errores independientes de dimensión *n* que se supone normalmente distribuido con media cero y varianza constante. El modelo lineal general planteado, debido al supuesto de distribución independiente de las variables aleatorias respuesta, puede resultar restrictivo en ocasiones donde los individuos observados se encuentran correlacionados como es el caso de poblaciones de mapeo donde subyace algún grado de estructuración genética.

Los modelos lineales mixtos (MLM) son más flexibles, al permitir que los elementos del vector de respuestas estén correlacionados; ya sea especificando a la matriz de varianzas y covarianzas como $\varepsilon \sim N(0, R)$ o incluyendo en el análisis efectos aleatorios. El MLM es una extensión del Modelo Lineal General para adicionar efectos aleatorios distintos a los términos de error. Los efectos aleatorios suelen indicarse subrayados.

El MLM se puede escribir de la siguiente manera:

$$y = X\beta + Zu + \varepsilon$$

$$u \sim N(0, G) \qquad e \sim N(0, R) \qquad Cov(u, e) = 0$$

donde:

y: vector de característica fenotípica

X: matriz de diseño de dimensión $n \times p$

Z: matriz de incidencia de dimensión $n \times q$

G: matriz de varianzas y covarianzas del vector de efectos aleatorios *u* de dimensión *q*

R: matriz de varianzas y covarianzas de los errores, vector *e* de dimensión *n*. Si se asume incorrelación y varianza constante, entonces $R = \sigma^2 I$.

Los parámetros del modelo $y = X\beta + Zu + \varepsilon$ son por lo tanto el vector de efectos fijos β y los elementos de las matrices *G* y *R*, llamados en general parámetros de covarianza. Se han desarrollado varios procedimientos con la finalidad de conseguir estimaciones de los efectos fijos y predicciones de los efectos aleatorios de un MLM pero los más utilizados son los que se enfocan en encontrar los mejores estimadores lineales insesgados - MELIs (o BLUEs por su sigla en inglés) y los mejores predictores lineales insesgados - MPLIs (o BLUPs por su sigla en inglés). Son mejores en el sentido que minimizan la varianza, lineales por ser funciones lineales de las observaciones e insesgados porque $E[MELI(\beta)] = \beta$ y $E[MPLI(u)] = u$ (Lynch y Walsh, 1998).

Dado que $E(u) = E(e) = 0$, por definición $E(y) = X\beta$ y como *u* y *e* no están correlacionados entonces Σ, matriz de varianzas y covarianzas de la variable respuesta, se define como:

$$\Sigma = Var(y) = ZGZ' + R$$

El primer término asociado a la variación de los efectos aleatorios y el segundo a la variación de los residuos. Entonces la distribución marginal de *y* es:

$$y \sim N(X\beta, \Sigma)$$

En mapeo asociativo, el vector β o vector de parámetros de efectos fijos, incluye información de los marcadores moleculares y en algunas circunstancias información de la estructura de la población, a la que se denominará *Q*. Por conveniencia, los efectos fijos suelen separarse en dos términos aditivos $X\beta + Qv$, uno correspondiente a los efectos relacionados con los marcadores moleculares y otro a los efectos de la estructura poblacional. En un MLM de mapeo asociativo el vector aleatorio *u*, puede ser entendido como un vector de efectos genéticos aditivos que si está presente induce correlación entre todos los individuos que comparten ese efecto. En algunos trabajos, se usa un modelo que denominaremos "*Naive*" donde se mantiene el supuesto del modelo lineal general que establece que los errores están incorrelacionados y que son homoscedásticos ($R = \sigma^2 I$). La matriz de varianzas y covarianzas de *u* se puede definir como $G = \sigma_A^2 K$ siendo σ_A^2 la varianza genética aditiva y K la matriz de parentesco o coancestría entre los individuos que componen la población de mapeo. Esta matriz modela la estructura de varianza y covarianza entre los individuos en la población (Yu *et al.*, 2006).

MODELOS DE MAPEO ASOCIATIVO

MODELO SIN CORRECCIÓN POR ESTRUCTURA (*NAIVE*)

Se usa para el mapeo, un modelo lineal general de regresión con efectos de marcadores en el vector de coeficientes de regresión y sin explicitar de ninguna manera (ni en la parte fija ni en la parte aleatoria del modelo) la existencia de estructura en los datos.

$$y = X\beta + e$$

donde y es el vector de caracteres fenotípicos de dimensión n×1, *X* es la matriz de marcadores moleculares de dimensión n×p, β es el vector de parámetros (o efectos fijos) de dimensión p×1 y *e* es el vector de errores de dimensión n×1.

MODELO EGP FIJA (P Y Q)

El modelo de mapeo asociativo, propuesto por Pritchard *et al., (*2000), presenta por un lado el componente asociado al efecto de los marcadores moleculares y por otro un componente asociado a la estructura genética poblacional (EGP), ambos tratados como de efectos fijos. En este modelo son usadas alguna de dos importantes matrices 1. La matriz Q proveniente del análisis de estructura realizado con el programa STRUCTURE, que presenta las probabilidades de pertenencia a las p poblaciones en estudio o 2. La matriz P de componentes principales significativas según el estadístico de Tracy-Widom (1994). El modelo se define como:

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

donde y es el vector fenotípico, *X* es la matriz de marcadores moleculares, β es el vector de parámetros (o efectos fijos), *EGP* es la matriz que indica la estructura genética, ya sea Q o P, *v* es el vector de efectos de la población y *e* es el vector de errores.

MODELO KINSHIP (K)

Se modela la variación en datos fenotípicos teniendo en cuenta los marcadores como efectos fijos y la matriz de coancestría, matriz “kinship” (K), para indicar relaciones de

parentesco entre los efectos aleatorios. Este modelo fue inicialmente propuesto por Yu y colaboradores (Yu *et al.*, 2006) y luego modificado por Kang *et al.*, (2008) y se define:

$$y = X\beta + \underline{Zu} + e$$

donde y es el vector fenotípico, *X* es la matriz de marcadores moleculares, β es el vector de parámetros (o efectos fijos), *Z* es una matriz de incidencia asociada al vector *u* de efectos poligénicos y *e* es el vector de errores. Se supone que el vector u se distribuye independientemente del vector e y con matriz de varianzas y covarianzas dada por $Var(u) = \sigma_g^2 K$ donde K es la matriz de parentesco (Kinship) obtenido mediante EMMA (Kang *et al.*, 2008) de R-project y $Var(e) = \sigma_e^2 I$ y la matriz de varianzas y covarianzas de los fenotipos será: $V = \sigma_g^2 ZKZ' + \sigma_e^2 I$

MODELO MIXTO UNIFICADO (QK Y PK)

Además de los efectos considerados en el modelo mixto Kinship que incluye la matriz de coancestría (K), este modelo unificado propuesto por Yu *et al.*, (2006), toma en cuenta los efectos fijos relacionados a la estructura genética de la población, ya sea con la matriz Q proveniente del STRUCTURE o con la matriz P proveniente del análisis de componentes principales. Este modelo tiene la forma:

$$y = X\beta + EGPv + \underline{Zu} + e$$

donde y es el vector fenotípico, *X* es la matriz de marcadores moleculares, β es el vector de parámetros (o efectos fijos), *EGP* es la matriz que indica la estructura genética, *v* es el vector de efectos de la población, *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 el vector de términos de error aleatorio. Se supone que el vector u se distribuye independientemente del vector e y con matriz de varianzas y covarianzas dada por $Var(u) = \sigma_g^2 K$ donde K es la matriz de parentesco (Kinship) obtenido mediante EMMA (Kang *et al.*, 2008) de R-project y $Var(e) = \sigma_e^2 I$ y la matriz de varianzas y covarianzas de los fenotipos será: $V = \sigma_g^2 ZKZ' + \sigma_e^2 I$

MODELO ESTRUCTURA ALEATORIA (QA Y PA)

En este caso se considera la información brindada por los marcadores moleculares y la estructura, pero a diferencia del modelo fijo se trabaja con el efecto producido por la estructura como un componente aleatorio (Q o P).

$$y = X\beta + \underline{EGPv} + e$$

Malosetti *et al., (*2007) plantean este modelo basados en la estructura poblacional propuesta por Patterson *et al., (*2006).

MÉTODOS DE CORRECCIÓN POR MULTIPLICIDAD

Uno de los problemas en los estudios de mapeo es la alta tasa de falsos positivos (error tipo I), luego, existe la necesidad de contemplar la correlación durante las pruebas de asociación. El método de control del error tipo I más conocido es la aproximación de Bonferroni (1935), usado en varias áreas del conocimiento, tales como ensayos clínicos de múltiples etapas (Sabatti *et al.*, 2003), experimentos de microarreglos (Tusher *et al.*, 2001), estudios de imágenes cerebrales (Schwartzman *et al.*, 2008) y estudios astronómicos (Miller *et al.*, 2001) resulta excesivamente conservador en estudios de MA por la gran cantidad de pruebas de hipótesis involucradas en escenario de genotipado masivo. Un criterio alternativo es el método de corrección de valores-p propuesto por Benjamini y Hochberg (1995) el cual estima la proporción esperada de hipótesis falsas rechazadas de todas aquellas rechazadas. El método ha sido desarrollado para mantener la tasa de error tipo I para una familia de pruebas de hipótesis realizadas sobre los mismos datos en el valor nominal o pre-especificado por el experimentador. Cheverud (2001) propuso una idea de ajuste de valores-p para pruebas correlacionadas como son las que podría emerger en el caso de LD. La idea es probar las hipótesis como si ellas fueran independientes estimando previamente un número efectivo (Meff) de pruebas independientes. Li y Ji (LJ, 2005) propusieron una estimación del Meff basada en la correlación entre marcadores moleculares y diseñaron un procedimiento basado en el Meff para controlar el error tipo I en la familia de hipótesis que se prueban en análisis de QTL clásico. No se conoce cómo la presencia de EGP podría impactar sobre el desempeño del método.

CAPITULO II

COMPARACIÓN DE MÉTODOS PARA IDENTIFICAR ESTRUCTURA GENÉTICA POBLACIONAL

INTRODUCCIÓN

El análisis de la variabilidad genética orientado a identificar la estructura genética poblacional (EGP) en colecciones vegetales es un paso importante no sólo en la formación de colecciones de recursos genéticos, sino también en estudios de asociación entre las variantes moleculares y los fenotipos (Shriner *et al.,* 2007; Wang *et al.,* 2005). Particularmente, el análisis de conglomerados (Hartigan, 1975; Gordon, 1999) ha demostrado ser una herramienta poderosa para investigar el agrupamiento "natural" de los genotipos de una población. Varios algoritmos han sido desarrollados para clasificar genotipos dentro de sub-poblaciones usando datos genéticos provenientes del genotipado molecular (Odong *et al.,* 2011; Lee *et al., 2*009; Lawson y Falush 2012). Algunos algoritmos de conglomerados tratan la información de marcadores como independiente (marcadores no ligados), pero otros consideran el desequilibrio de ligamiento (LD) (*i.e.* correlación entre marcadores) mientras que en algunos casos con muchos marcadores, suelen aplicarse filtros para descartar algunos de los más correlacionados y reducir la dimensionalidad del problema. En el contexto de alta densidad de marcadores, los métodos que contemplan LD han demostrado ser más potentes (Lawson y Falush 2012).

Entre los conglomerados jerárquicos, el método de conglomerados conocido como encadenamiento promedio (UPGMA de su nombre en inglés, *Unweighted Pair Group Method with Arithmetic Mean*) (Sokal y Michener, 1958) y el método de Ward (Ward, 1963) son los más usados con datos de marcadores moleculares (Balzarini *et al.*, 2010). Dentro de los conglomerados no jerárquicos, uno de los algoritmos de conglomeración más frecuentemente seleccionados es el denominado K-means (MacQueen, 1967), que asigna cada uno de los individuos a clasificar, a uno de los K grupos definidos. Los primeros producen una salida en formato de dendrograma (Mayr *et al.*, 1953) mientras que en los no jerárquicos no se obtiene el dendrograma.

Últimamente, con el advenimiento de mayores capacidades de cómputo se están usando también técnicas multivariadas del tipo bioinformático como son los mapas auto-organizativos (SOM del inglés *Self Organizing Maps*) (Kohonen, 1997), que pertenecen a la familia de redes neuronales como algoritmos de clasificación (Nikolic *et al.*, 2009). En este Capítulo, esta alternativa es evaluada simultáneamente a los otros procedimientos. Este es un algoritmo no supervisado, de la familia de las redes neuronales, con potencialidad para encontrar relaciones entre datos de gran dimensión, agrupándolos y mapeándolos topológicamente. La idea central que soporta el algoritmo SOM es que objetos similares en un espacio de entrada multidimensional serán mapeados cerca uno del otro dentro del mapa resultante de un proceso de organización de información según parecido que se va dando automáticamente. En el arreglo SOM, la estructura poblacional será reconocida como grupos de nodos de la red inicial que terminan juntos o ligados luego del proceso de aprendizaje (conglomerado). La identificación de conglomerados en SOM puede ser realizada a través de métodos de visualización (Ultsch, 1999) o a través de estadísticos que relacionan la variabilidad entre y dentro de grupos. El algoritmo de clasificación SOM-RP-Q (Fernández y Balzarini, 2007), agrega a SOM ese tipo de estadísticos para apoyar la identificación de la estructura subyacente en los datos. Las aplicaciones de SOM, en particular de SOM-RP-Q no son frecuentes en mejoramiento genético vegetal como en otros campos del análisis de datos genéticos (Toronen *et al.*, 1999). Dada la gran diversidad de técnicas de conglomerados disponibles, resulta necesario comparar el desempeño estadístico de los mismos en la aplicación de interés. Otros trabajos han realizado investigaciones metodológicas sobre la identificación de grupos en conjuntos de datos moleculares (Lee *et al.*, 2009; Odong *et al.*, 2011), pero no se han

contemplado simultáneamente, o sobre los mismos conjuntos de datos, métodos de diferente naturaleza como son los conglomerados basados en distancias multivariadas, los bayesianos basados en clasificación difusa y los pertenecientes a la familia de las redes neuronales como SOM o SOM-RP-Q.

Es importante resaltar, que el desempeño relativo de diferentes métodos puede ser función de distintas características específicas de la estructura genética poblacional subyacente como el número de sub-poblaciones (K) y la similitud genética o divergencia genética entre grupos, la cual puede ser medida por el estadístico F_{ST} de Wright (1951), que representa la correlación entre los genes de la subpoblación y los de la población total. Evanno *et al.,* (2005) llevó a cabo un estudio de simulación para evaluar la habilidad de STRUCTURE para reconocer la estructura genética bajo diferentes escenarios biológicos derivados de distintos patrones de migración que consideraban distintos niveles de divergencia genética. Odong *et al., (*2011) evaluaron métodos de conglomerados jerárquicos bajo diferentes niveles de divergencia en colecciones de germoplasma vegetal usando los resultados de STRUCTURE como "*gold standard*". Lee *et al., (*2009) compararon, en un estudio de simulación, Análisis de Componentes Principales (ACP), el algoritmo de STRUCTURE y técnicas de conglomerados no jerárquicos. En un trabajo de Milligan y Cooper (1985), se enumeran varios criterios para la comparación del desempeño de métodos de conglomeración, muchas de éstas usadas en los estudios de simulación mencionados anteriormente. Sin embargo, dado que los métodos que se comparan en este trabajo pertenecen a diferentes familias, difieren en la forma en la cual son agrupados los genotipos y tienen diferentes formas de visualización de resultados, la evaluación se realiza mediante tasas de error de clasificación usando la información disponible sobre la estructura subyacente desde escenarios simulados donde ésta es conocida. El grado de concordancia entre los verdaderos conglomerados de genotipos (simulados) y los obtenidos luego de aplicar una técnica de conglomeración es usado como indicador del desempeño del método. El objetivo de este trabajo fue evaluar el desempeño del algoritmo SOM-RP-Q y de otros métodos ampliamente usados para agrupar genotipos desde datos genéticos no ligados.

MATERIALES Y MÉTODOS

DATOS SIMULADOS

Se simularon datos de marcadores moleculares usando QMSim (Sargolzaei y Schenkel 2009) e involucrado escenarios con *n* cantidad de genotipos y *p* cantidad de marcadores moleculares que imiten datos de mejoramiento genético. Se diseñó un genoma con 300 marcadores multiloci-bialélicos. Se crearon por simulación seis escenarios biológicos, correspondientes a tres niveles de divergencia genética entre poblaciones (bajo, medio y alto F_{ST}), considerando la variación en el número de generaciones desde la población fundadora, y dos números de poblaciones (K=3 y K=5). El número de genotipos por escenario fue de 180. Estos tamaños poblacionales son comunes en estudios de mapeo asociativo en mejoramiento genético vegetal. Los datos simulados fueron creados a partir de una población histórica de 200 individuos y el sistema de cruzamiento fue basado en la unión de gametas muestreadas aleatoriamente por diferente número de generaciones, para lograr un nivel de divergencia genética bajo se simularon 10 generaciones, para el nivel medio se simularon 40 generaciones y para lograr una alta divergencia fueron necesarias 70 simulaciones (Tabla 2.1). Los datos simulados fueron codificados como 0 o 1 para cada marcador. Cada escenario fue repetido 15 veces (Anexo 1). El promedio del estadístico F_{ST} (Wright 1951) derivado desde 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 2.1).

Tabla 2.1. Número de poblaciones y diversidad genética que caracteriza la estructura genética poblacional de seis escenarios simulados de genotipos multilocus-bialélicos.

Escenario	Número de poblaciones (*K*)	Diversidad genética	
		Estadístico F_{ST}	Nivel
I	3	0.07	Bajo
II	5	0.06	Bajo
III	3	0.23	Medio
IV	5	0.17	Medio
V	3	0.38	Alto
VI	5	0.38	Alto

DATOS EXPERIMENTALES

Como conjunto de datos reales se usó un archivo provisto por Natalia de León publicado por Hansey *et al.,* (2011), este archivo cuenta con n=334 líneas de maíz genotipadas mediante p=210 marcadores SNPs. En dicho trabajo sobre estos datos genéticos, Hansey *et al., (*2011) identificó ocho conglomerados o sub-poblaciones que fueron verificadas desde el conocimiento biológico de investigaciones genéticas de maíz. Dicha clasificación fue usada como "*gold estándar*" en este trabajo. Todos los análisis sobre los datos reales se hicieron asumiendo la existencia de ocho conglomerados. Un gráfico LD-*heatmap* fue realizado para evaluar el desequilibrio de ligamiento (DL) (Fig. 1). Se encontró un bajo nivel de correlación entre los 210 SNPs (<5%).

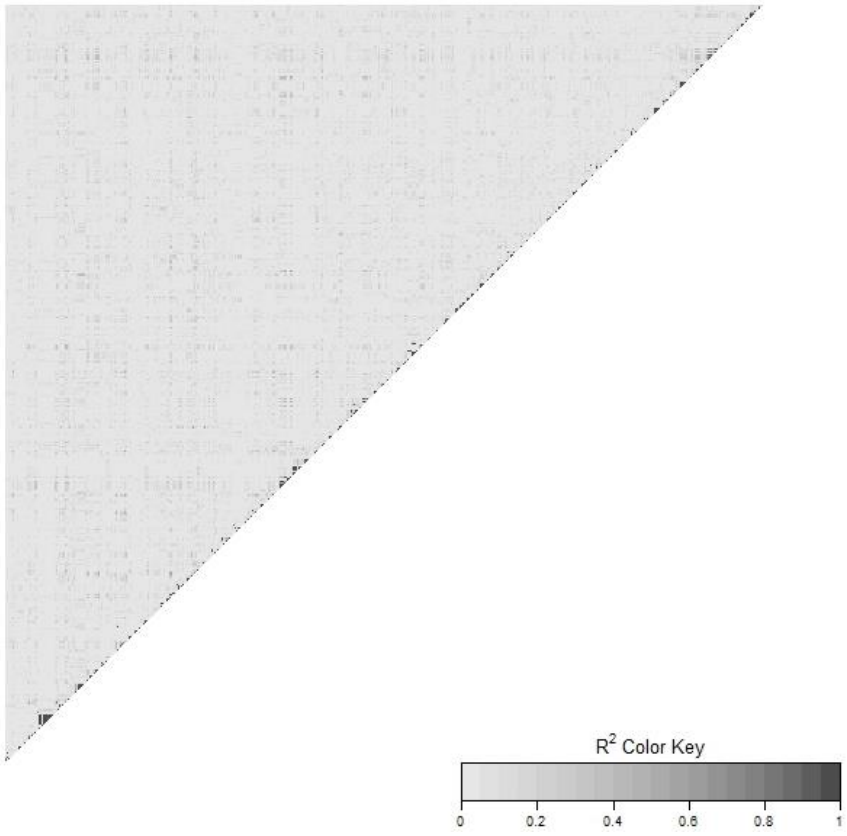

Figura 2.1. LDheatmap para datos de marcadores moleculares en un conjunto de genotipos de maíz. Cada elemento de la matriz triangular superior (R^2) es una medida de desequilibrio de ligamiento DL entre los marcadores. Bajo R^2 indica marcadores no correlacionados.

ALGORITMOS DE CONGLOMERADOS EVALUADOS

CONGLOMERADOS BASADOS EN DISTANCIAS

Se compararon métodos de conglomerados jerárquicos y no jerárquicos. Dos algoritmos jerárquicos: Encadenamiento promedio o UPGMA (Sokal y Michener, 1958) y Ward (Ward, 1963), fueron aplicados directamente sobre los datos de marcadores moleculares. Además, el algoritmo Ward fue aplicado también sobre las componentes principales de los datos moleculares que fueron reconocidas en un paso previo como significativas según la distribución de Tracy-Widom (TW) (Tracy y Widom, 1994). El análisis de componentes principales junto con el estadístico TW provee una forma rápida y efectiva para responder si los datos sugieren la existencia de EGP.

Patterson *et al.*, (2006) modeló los valores propios del ACP usando la distribución de TW, la cual describe el valor propio más grande de la matriz ZZ^T, donde Z es la matriz de datos de marcadores moleculares. Así, el estadístico TW puede ser usado para evaluar significancia de las componentes principales. El método de conglomeración que incluye la obtención de los valores propios, la selección de las componentes principales estadísticamente significativas y la aplicación del algoritmo Ward usando como variables de entrada a las componentes principales seleccionadas fue denominado, en este trabajo, como ACP+Ward.

Con el algoritmo UPGMA el proceso de conglomeración inicia desde una matriz de distancias de todos los pares de genotipos y luego los genotipos son agrupados usando un criterio basado en las distancias promedio entre conglomerados (Balzarini *et al.*, 2008). En el método Ward el proceso es parecido, pero el criterio de agrupamiento es basado en una ponderación (por el tamaño de cada grupo) de las distancias entre los conglomerados (Johnson y Wichern, 1998); este método es particularmente útil para variables de entradas incorrelacionadas como es el caso de las componentes principales.

También se evaluó el método no jerárquico K-means (MacQueen, 1967). Este método de conglomeración inicia el agrupamiento con una partición inicial (aleatoria) de los genotipos en K grupos y continúa reubicando cada genotipo en uno de los conglomerados tal que la distancia entre un genotipo y el centroide del conglomerado al cual éste fue asignado es menor que con los otros centroides de grupo. La diferencia entre grupos es

maximizada y las diferencias dentro de cada conglomerado son minimizadas (función objetivo representada por la suma de cuadrados dentro de grupo).

CONGLOMERADO BAYESIANO

En este trabajo también fue evaluado el método de conglomerado difuso implementado en el software STRUCTURE (Pritchard *et al., 2*000). Este método es usado para asignar genotipos a un grupo K especificado a priori. Los genotipos en una colección son asignados probabilísticamente a estos grupos, o juntamente en dos o más poblaciones si los genotipos indican una mezcla de patrones moleculares. En este método, como en otros bayesianos de conglomeración difusa, las probabilidades a posteriori indican la incertidumbre de la asignación de conglomerados.

CONGLOMERADOS HEURÍSTICOS

El algoritmo mapa auto-organizativo (SOM) (Kohonen, 1997) es una red neuronal artificial capaz de convertir datos de alta dimensión en un mapa de dos dimensiones en el que los puntos de datos que se encuentran muy próximos en el mapa son más similares que los que están más lejos. El SOM consta de dos capas de neuronas artificiales, la capa de entrada y la capa de salida (Paini *et al.,* 2010). En el SOM, la capa de entrada está conformada esencialmente por los datos en bruto y comprende p neuronas (una neurona para cada marcador molecular en nuestro caso), y cada neurona se encuentra conectada a todos los N genotipos. La capa de salida es un mapa de dos dimensiones que comprende un número n artificial de neuronas (nodos), dispuestas en una cuadrícula. Cada uno de los genotipos ocupa un punto particular en el espacio de dimensión p. El SOM proyecta sus nodos en este espacio a través de vectores de ponderación de las neuronas; cada neurona SOM ocupa un punto en el mismo espacio multidimensional de los genotipos, interactuando así con los genotipos (Worner y Gevrey, 2006). Cuando se inicia el análisis, cada dato se evalúa y la neurona que es más cercana a este dato, en el espacio multidimensional, se considera como nodo ganador o nodo de mejor coincidencia. El vector de peso asociado a ese nodo, sufrirá un ajuste y el nodo se moverá cerca del dato. Debido a que todas las neuronas están conectadas entre sí, el proceso de una neurona en movimiento ejerce una fuerza que arrastra a otras neuronas en el SOM. Cuando se completa el análisis, cada dato de entrada

tendrá un nodo ganador (neurona más cercana). Los datos de genotipos que tienen perfiles de marcadores moleculares similares se encontrarán juntos en el espacio multidimensional y tendrán el mismo nodo ganador. Cada nodo o grupo de nodos cercanos en la red entrenada serán entendidos como un conglomerado de genotipos.

En SOM, las agrupaciones se identifican a través de algoritmos de visualización específicos, uno de éstos es el método de RP-Q (Fernández y Balzarini, 2007). En este método, se usa para cada nodo de la red un atributo de adaptación denominado posición relativa (RP). Éste se mueve en un espacio virtual de dos dimensiones imitando el movimiento de las neuronas. La RP final de los nodos de la red se muestra en un gráfico de dispersión. Cada nodo está representado por un círculo cuyo diámetro es proporcional a la "frecuencia de activación" (frecuencia que representa las veces que el nodo ha sido nodo ganador) durante la fase de aprendizaje. Para ayudar en la identificación de grupos de nodos o conglomerados, los nodos más similares están vinculados con un segmento de línea. La longitud del segmento que une dos nodos proporciona información sobre las distancias entre ellos. Los nodos que se vinculan entre sí formarán un conglomerado si la distancia entre ellos es menor que o igual a un valor umbral especificado. El estadístico Q se utiliza para fijar este umbral y la clasificación. El estadístico Q se basa en el cociente entre las sumas de cuadrados entre y dentro de grupos de nodos y además realiza un ajuste para no contar como conglomerados a nodos aislados. Los nodos aislados son frecuentes en procesos de aprendizaje cuando se producen mezclas de los patrones de la estructura de grupos que subyace los datos en el espacio de entrada (nodos de transición).

IMPLEMENTACIÓN DE LOS ALGORITMOS Y CRITERIOS DE COMPARACIÓN

Para implementar los métodos basados en distancia multivariada se utilizó como medida de distancia el cuadrado de la distancia euclídea entre los genotipos. Las agrupaciones jerárquica y no jerárquica se realizaron con Info-Gen (Balzarini y Di Rienzo, 2004) asumiendo un número conocido de grupos (K=3 y K=5 para los datos simulados y K=8 para los datos experimentales). STRUCTURE se implementó bajo el supuesto de un modelo de mezcla con un modelo de la frecuencia alélica independiente utilizando el número de conglomerados (K) sugerido por la EGP simulada o publicada para el caso de

los datos reales y con un número de iteraciones de 50.000 para cada ejecución. Cuando se implementó STRUCTURE, la clasificación de genotipos se basó en la proporción más alta de pertenencia a un grupo registrada para cada individuo o genotipo. El software L-SOM se utilizó para formar la red neuronal en el algoritmo de SOM y la herramienta RelPos desarrollada en Matlab® para implementar el método SOM-RP-Q (Fernández y Balzarini, 2007). La red inicial fue configurada como una matriz de 15×15 nodos. Para SOM-RP-Q, la clasificación de los nodos en grupos (y en consecuencia, la clasificación de los genotipos) se basó en el número de individuos de cada grupo subyacente a un nodo dado. Si la mayor proporción de individuos pertenecía al grupo *i*, a continuación, el nodo se supone que es un componente de la agrupación *i*. La proporción de individuos del grupo *i* en el nodo clasificado como parte de la agrupación *i* (llamado probabilidad posterior) se utilizó como indicador de la certeza en la asignación de clúster. El promedio, a través de nodos, de su complemento se utilizó como una medida de la incertidumbre de la asignación de grupos mediante SOM-RP-Q.

Para comparar simultáneamente el funcionamiento de los métodos evaluados respecto a su desempeño para identificar la agrupación de genotipos subyacente en la población de mapeo, se utilizó una medida denominada tasa de error de la agrupación (CER, del inglés *Cluster Error Rate*). CER se definió de la siguiente manera:

$$CER = \frac{\sum_{i}^{K} Er_i}{N}$$

donde N es el número de genotipos de la población, K es el número de grupos en que se estructura la población en función de los datos genómicos, Er_i es la tasa de error de agrupación en el i-ésimo grupo, con i=1,...,K. Se basa en la diferencia entre el número real de individuos que pertenecen al grupo i-ésimo (N_i) y los individuos clasificados correctamente en ese grupo (C_i). La tasa de error de clasificación para el grupo i-ésimo se estima como $Er_i = (N_i - C_i)/N_i$. CER es el promedio de la tasa de error de clasificación de los K grupos.

RESULTADOS

En la Figura 2.2 se presenta el resultado de un escalamiento multidimensional métrico (Análisis de Coordenadas Principales) realizado sobre los datos genómicos con el objetivo de mostrar el nivel de divergencia genética alcanzado en cada uno de los escenarios simulados. Se muestran los gráficos de dispersión de los dos primeros ejes resultantes del escalamiento multidimensional de los datos de seis escenarios con diferentes niveles de F_{ST} (bajo, medio y alto, de arriba hacia abajo, respectivamente) y el número de grupos (K=3 y K=5).

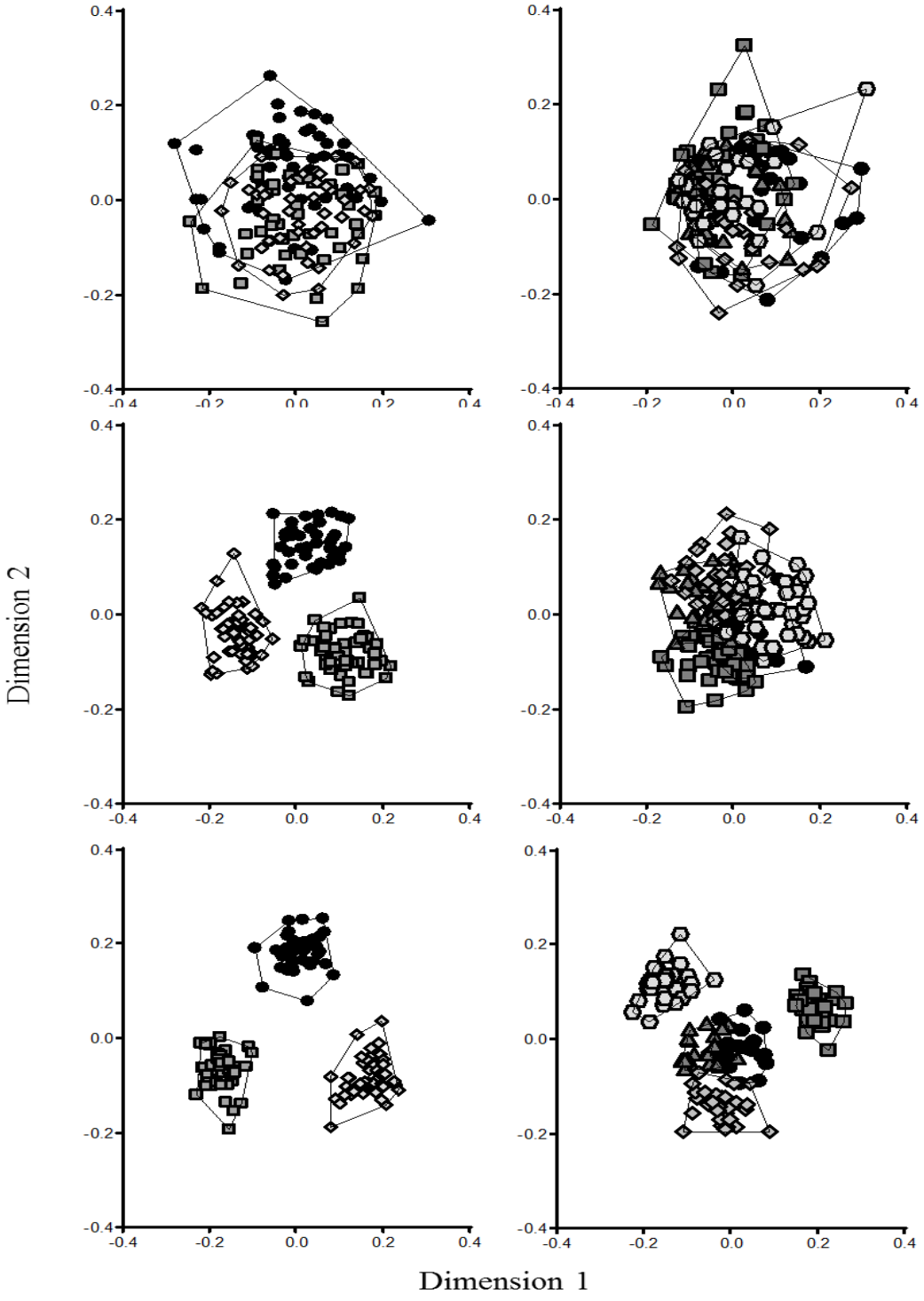

Figura 2.2. Gráficos de dispersión de los dos primeros ejes resultantes de un análisis de coordenadas principales (escalamiento multidimensional) de los datos moleculares. En la columna de la izquierda para tres poblaciones, mientras en la columna de la derecha para cinco poblaciones. De arriba hacia abajo, bajo (F_{ST}=0.06–0.07), medio (F_{ST}=0.23–0.17), y alto (F_{ST}=0.38) F_{ST}.

En la Figura 2.3 se puede visualizar la clasificación de los datos experimentales de perfiles genómicos de maíz. Esta figura proporciona información con respecto a la distancia genética entre los genotipos y el grupo al que estos genotipos fueron clasificados por Hansey *et al.*, (2011). El estadístico F_{ST} calculado a partir de los datos genéticos de esta población de mapeo fue de 0.02, es decir, cercano al de los datos simulados para escenarios que indican baja divergencia genética.

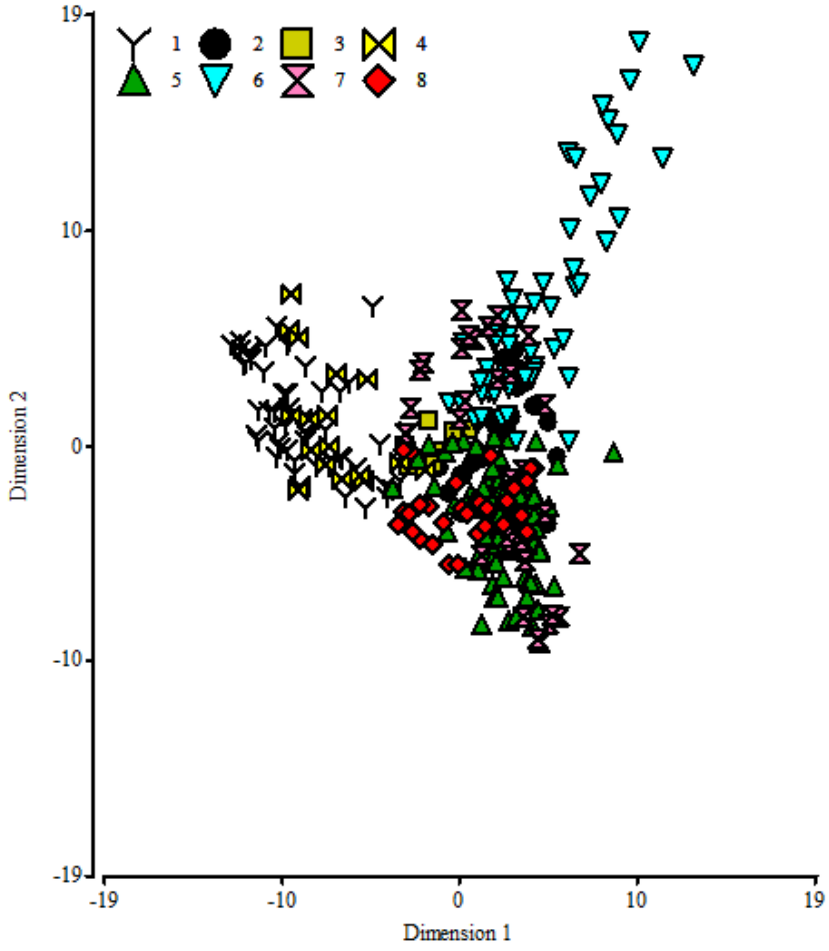

Figura 2.3. Gráficos de dispersión de los dos primeros ejes resultantes de un análisis de coordenadas principales (escalamiento multidimensional) de los datos moleculares (SNPs) de ocho grupos de los datos experimentales de maíz.

ANÁLISIS DE RESULTADOS EN LOS ESCENARIOS SIMULADOS

En escenarios donde la divergencia genética fue baja (F_{ST}<0.10), el método SOM-RP-Q produjo mejores resultados con un valor de CER por debajo de 0.25, mientras que el error de clasificación de genotipos en los grupos que conforman la EGP según la clasificación realizada por los métodos jerárquicos alcanzó 0.73 (Tabla 2.2). La estimación de la probabilidad posterior en el algoritmo SOM-RP-Q indica que, incluso en condiciones de poca diferenciación genética, la incertidumbre de la asignación de conglomerados fue inferior a 0.10.

Para el método ACP+Ward, el número medio de componentes significativos disminuyó cuando el estadístico F_{ST} aumentó, es decir, cuando existían mayores divergencias entre las poblaciones o mayor nivel de EGP. En los escenarios I y II, con baja diferencia genética, el estadístico TW identificó más de 20 componentes significativas. Los errores de agrupación fueron similares a los observados para los otros métodos jerárquicos sin previo análisis de componentes principales.

Para estructuras más complejas (K=5), el valor de CER aumentó en todos los métodos (Fig. 2.4). En escenarios con una fuerte estructura de la población (escenarios V y VI) todos los métodos tuvieron un buen desempeño, con bajo error en la identificación de la EGP subyacente (CER <0.02) (Tabla 2.2).

Los métodos de conglomerados jerárquicos fueron competitivos sólo en escenarios con altos niveles de divergencia genética (escenarios V y VI), tanto cuando se aplicaron directamente a los datos genómicos como cuando se aplicaron sobre variables sintéticas o componentes principales estadísticamente significativas según el estadístico TW.

Tabla 2.2. Proporción de error de clasificación de los algoritmos evaluados para identificar EGP en poblaciones de perfiles moleculares simuladas bajo distintos escenarios biológicos.

Divergencia genética		Poblaciones	Método de Conglomerados					
Nivel	Estadístico F_{ST}	*K*	UPGMA*	Ward†	ACP+Ward‡	K-Means§	STRUCTURE¶	SOM-RP-Q ††
Bajo	0.07	3	0.61	0.56	0.57(22)	0.52	0.48	0.17
Bajo	0.06	5	0.73	0.68	0.71(20)	0.69	0.68	0.22
Medio	0.23	3	0.09	0.03	0.09(8)	0.01	0.02	0.01
Medio	0.17	5	0.47	0.22	0.15(9)	0.08	0.09	0.03
Alto	0.38	3	0	0	0(7)	0	0	0
Alto	0.38	5	0.04	0.01	0(6)	0	0	0

* Unweighted Pair Group Method using Arithmetic Average. † Método Ward. ‡ Método Ward aplicado a las componentes principales (CP) que fueron significativas según el estadístico Tracy-Widom (TW), el número promedio de CPs usadas se muestra en paréntesis. § Método no-jerárquico K-Means. ¶ Método Bayesiano implementado en STRUCTURE. ††Método de mapa auto-organizativo con posición relativa.

La tendencia en la CER fue dependiente del nivel de divergencia genética en población (Fig. 2.4). El método SOM-RP-Q (línea negra continua) mostró el error de clasificación más bajo (Fig. 2.4). Para los valores de F_{ST} cercanos a 0.3, el error llegó a valores despreciables. Los valores de CER para K-means y STRUCTURE fueron similares para ambos valores de K. Sin embargo, la probabilidad posterior dada por STRUCTURE indica que la incertidumbre de conglomeración fue sólo despreciable con alto F_{ST}, pero los valores medios alcanzados fueron de 0.40 en escenarios con bajo F_{ST}. En estos conjuntos de datos simulados, STRUCTURE y SOM-RP-Q tuvieron desempeños estadísticos similares.

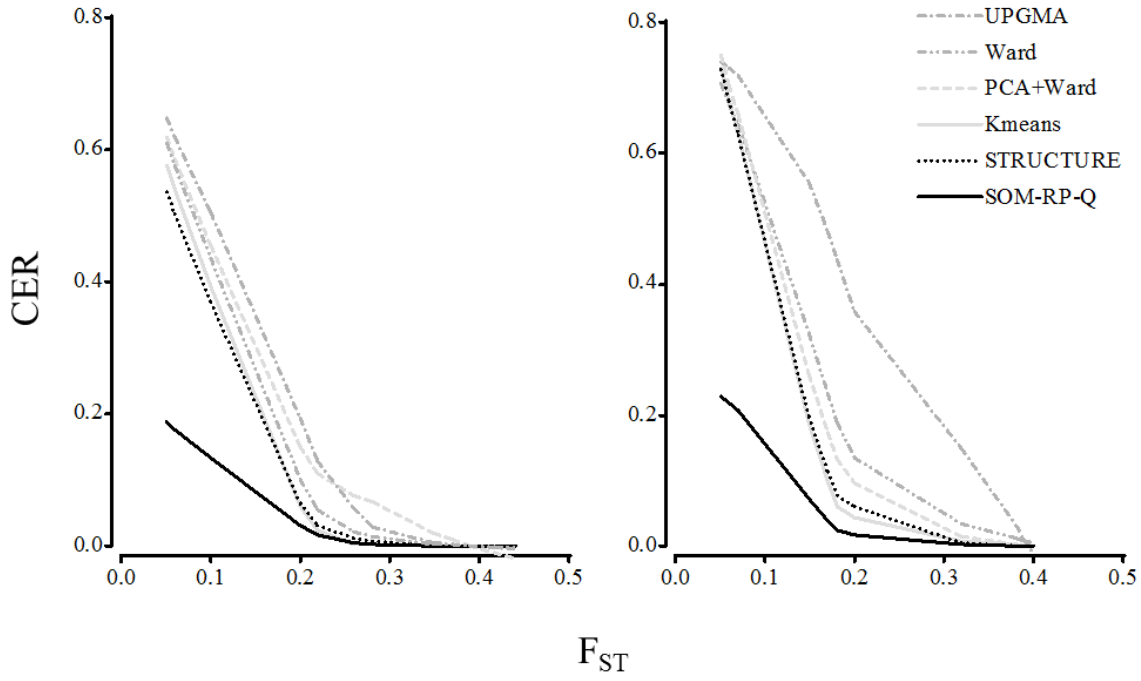

Figura 2.4. Tasa de error de clasificación (CER) con respecto al nivel de divergencia genética (F_{ST}) entre poblaciones para EGP con tres poblaciones (izquierda) y cinco poblaciones (derecha). Seis procedimientos de agrupamiento fueron comparados.

ANÁLISIS DE RESULTADOS EN LOS ESCENARIOS DE DATOS REALES

Los métodos SOM-RP-Q, Ward y STRUCTURE reconocieron la EGP subyacente en el conjunto de datos experimentales de maíz, con errores de agrupamiento de 0.09, 0.21 y 0.23, respectivamente. Por el contrario, los métodos UPGMA, K-means y ACP+Ward, no fueron capaces de identificar la EGP en estos perfiles moleculares. El dendrograma, que se muestra en la figura 2.5, es el diagrama resultante del análisis de agrupamiento jerárquico. El eje x en el dendrograma indica la distancia entre los genotipos para nuestro caso distancia euclídea al cuadrado y grupos de genotipos.

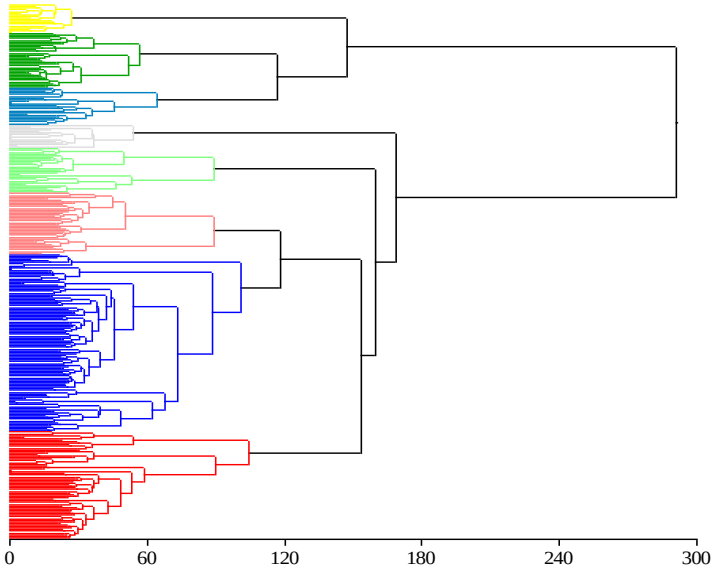

Figura 2.5 Dendrograma del método Ward aplicado a los datos experimentales de maíz.

En la Figura 2.6 se muestra un gráfico de barras obtenido como salida del software STRUCTURE, cada línea vertical representa un individuo y cada color representa una población. Las barras que representan cada individuo van de 0 a 1 y cuando un individuo está representado por un solo color significa que la proporción de pertenencia a dicho grupo es completa o igual a 1, mientras que cuando una barra es de 2 o más colores indica que el individuo representado por dicha barra tiene proporción de pertenencia a varios grupos lo que indica que se compone de una mezcla de poblaciones.

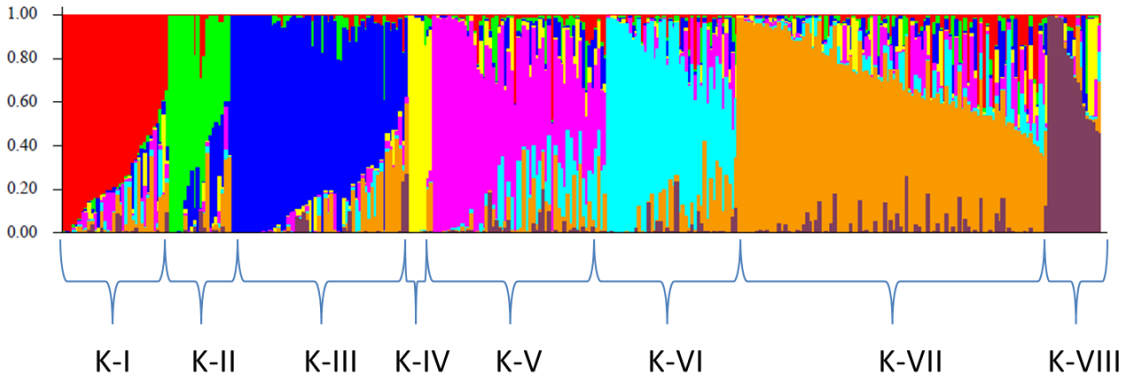

Figura 2.6 Gráfico Dendrograma del método Ward aplicado a los datos experimentales de maíz.

La Figura 2.7 también muestra la salida gráfica de SOM-RP-Q. En el gráfico SOM-RP-Q (panel superior-izquierda), pequeños círculos representan los nodos y se agrupan en racimos (marcadas con elipses) según su posición relativa. Los tamaños de nodo son proporcionales a la frecuencia que indica las veces que el nodo fue identificado como nodo ganador durante la fase de entrenamiento, y los números en cada nodo representan su posición en la estructura SOM. Los eje x y el eje y son las coordenadas del espacio de las posiciones relativas, sus valores son arbitrarios pero permiten ver los nodos que se conglomeran por similitud. Sobre el mismo espacio se han dibujado elipses indicando los grupos de nodo que forman un conglomerado. Se visualizan ocho conglomerados. Los nodos restantes, por fuera de estos grupos, son considerados como nodos de transición.

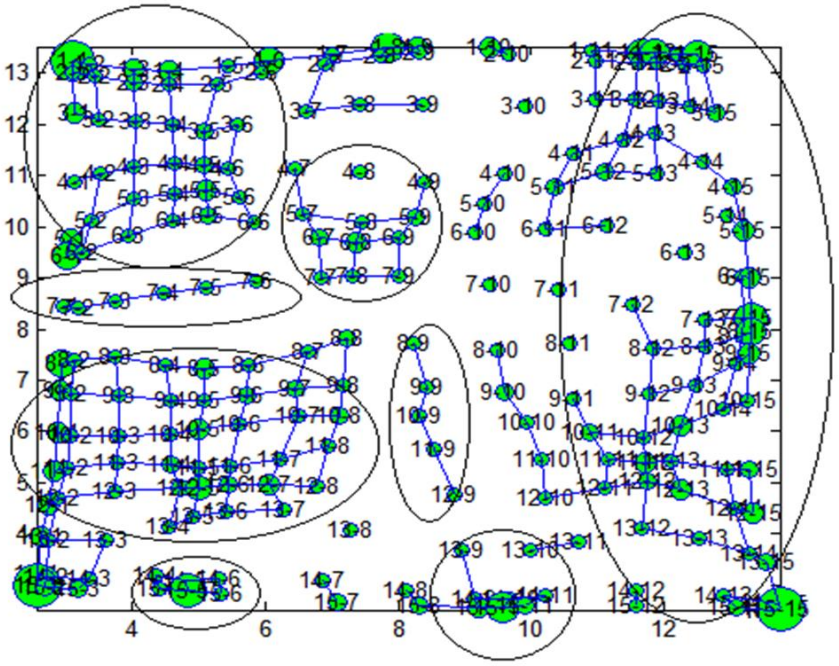

Figura 2.7 Salida gráfica de SOM-RP-Q aplicado a los datos experimentales de maíz.

DISCUSIÓN

La identificación de grupos de perfiles de genotipos dados por los marcadores moleculares es un objetivo común en genética y biología molecular. El análisis de conglomerados ha demostrado ser una herramienta poderosa para investigar los grupos de genotipos

moleculares que subyacen una población. En la práctica del análisis de conglomerados, los grupos no se conocen a priori, y el interés se centra en la búsqueda de ellos sin la ayuda de una variable de respuesta (aprendizaje no supervisado). En numerosos trabajos con datos genéticos, se han utilizado algoritmos de agrupamiento jerárquicos, como UPGMA y Ward, y no jerárquicos como K-means ya que éstos proporcionan una herramienta sencilla y potente para la determinación de la EGP de las colecciones de germoplasma utilizando datos de marcadores moleculares (Odong *et al., 2*011). El análisis se puede realizar directamente desde el conjunto de datos moleculares y usualmente es seguido por otros procedimientos para estimar el número de conglomerados óptimo para describir la EGP (Gordon 1999, Still *et al.*, 2004, Tibshirani *et al.*, 2001).

Estudios recientes sobre identificación de EGP, muestran que la información molecular puede ser manejada de manera efectiva por métodos de conglomerados basados en modelos. Éstos asignan (probabilísticamente) individuos de la población a una de las subpoblaciones, o conjuntamente a dos o más subpoblaciones si sus genotipos indican que son mezcla. El software STRUCTURE permite implementar un método de clasificación basado en modelo aplicables directamente a datos de marcadores genéticos, cuando éstos no están estrechamente relacionados (Pritchard *et al.*, 2000); el supuesto biológico de equilibrio de ligamiento es un requisito difícil de cumplir a medida que la intensidad del genotipado se hace mayor. En este trabajo, el algoritmo de STRUCTURE mostró ser efectivo para determinar la EGP para los datos experimentales de maíz y también para la mayoría de los escenarios simulados.

Otra aproximación para el estudio de EGP es el uso de técnicas de reducción de dimensión, como el análisis de componentes principales (ACP) previo al análisis de conglomerado. Acoplado al estadístico TW (Tracy y Widon, 1994) se obtiene una prueba formal para EGP (Patterson *et al.*, 2006). Los resultados de este trabajo sugieren que el desempeño de los métodos de conglomeración basados en distancia, ya sea aplicado directamente a los datos de los marcadores o a las componentes principales resulta dependiente de la diferenciación genética de los subgrupos. Nosotros comparamos ambos tipos de algoritmos, los basados en modelos y los basados en distancias multivariada entre ellos, y con el método de agrupación de la familia de redes neuronales artificiales (SOM-RP-Q) (Fernández y Balzarini, 2007). El desempeño global de los métodos de identificación de EGP fue evaluado en varios escenarios de datos de marcadores simulados y utilizando datos

experimentales de líneas puras de maíz (Hansey *et al., 2*011) con EGP conocida a partir de estudios previos. SOM-RP-Q produjo una buena estimación del número de grupos en los conjuntos de datos simulados. La identificación de los nodos que pertenecen a un mismo conglomerado hizo que sea posible calcular coeficientes de incertidumbre para la asignación de un individuo a la estructura de conglomerados.

Las tasas de error de clasificación mostraron que el agrupamiento jerárquico UPGMA aplicado a los datos moleculares no es eficiente para detectar estructura genética de población en los casos de baja divergencia genética. En niveles medios de divergencia genética ($0.1<F_{ST}<0.25$), UPGMA mejoró su rendimiento relativo en las situaciones de estructura poblacional simple (por ejemplo: 3 subpoblaciones), pero mantuvo altas tasas de error de la agrupación con estructuras más complejas (por ejemplo: 5 subpoblaciones). Estos métodos no pueden ser recomendados para la determinación de EGP en poblaciones de mapeo en vegetales donde la divergencia genética entre grupos es usualmente baja o a lo sumo intermedia. Los métodos basados en distancia sólo se desempeñaron bien en escenarios en los cuales existía alta divergencia entre los grupos de genotipos. Odong *et al., (*2011) sugirieron que el pobre desempeño del método UPGMA en la recuperación de la estructura original con baja divergencia se debe a que produce conglomerados con alto desbalance. Según los resultados de este trabajo, con niveles altos de divergencia genética ($F_{ST}>0.2$), el rendimiento de todos los procedimientos es similar y aceptable.

Las diferencias en el error de clasificación entre Ward y UPGMA en datos reales confirman las conclusiones de los resultados de simulación que muestran un mejor desempeño de Ward que de UPGMA bajo divergencia genética inferior. Jobson (1992) analiza la capacidad de Ward para evitar adhesiones periféricas dentro de las agrupaciones. Por lo tanto, el algoritmo de Ward fue la opción de mejor desempeño entre los métodos de agrupamiento jerárquico. Sin embargo, el método SOM-RP-Q realizó siempre la mejor clasificación de perfiles de marcadores moleculares, identificando las poblaciones subyacentes aun cuando existía relativamente baja divergencia entre subpoblaciones ($F_{ST}<0.1$). SOM-RP-Q proporcionó una herramienta para mapear los perfiles de marcadores moleculares en un espacio de dos dimensiones. La estructura interna de la red mapeada permite inferir sobre el número de grupos emergentes en la red entrenada.

El desempeño de SOM-RP-Q fue próximo al de STRUCTURE, mejorando el primero los tiempos de implementación. El algoritmo del software STRUCTURE estima parámetros de manera computacionalmente más intensiva que el algoritmo SOM (Lee *et al., 2*009; Roux *et al., 2*007). Wang *et al., (*2002) concluyeron que mediante el uso de redes neuronales, como un paso intermedio para analizar los datos de expresión de genes del genoma completo, los patrones de expresión génica pueden ser revelados más fácilmente que mediante el uso de los métodos jerárquicos y no jerárquicos. Evanno *et al., (*2005) afirmó que STRUCTURE fue capaz de detectar las subpoblaciones en conjuntos de datos simulados según un modelo de migración de islas. Nuestros resultados están de acuerdo con los hallazgos de Evanno *et al., (*2005), pero sólo cuando la divergencia entre subpoblaciones no es baja. Zhao *et al., (*2007) informaron que el método de agrupamiento STRUCTURE no podía definir adecuadamente la complejidad estructural de las poblaciones.

Lawson y Falush (2012) mostraron que los algoritmos basados en modelos podrían superar a los conglomerados con enfoques genéricos bajo LD entre los marcadores genéticos y la alta relación entre los individuos. Ellos discutieron también el problema de la elevada correlación estadística entre sitios cercanos físicamente y la subsecuente pérdida de información para calcular las similitudes entre los individuos. El algoritmo basado en modelos llamado FineSTRUCTURE (Lawson *et al., 2*012), es recomendable en casos que hay LD entre los marcadores. Este método no se utilizó en este trabajo porque, dada la baja cantidad de marcadores (200-300 SNPs), los conjuntos de datos pueden ser en gran medida no ligados. Lawson y Falush (2012) trabajaron en el contexto de datos provenientes de secuenciamientos completos, y con datos simulados pero que datos que contienen 500,000 SNPs para un pequeño número de individuos (n=140). Los métodos incluidos en nuestro documento asumen que cada marcador es independiente de la mayoría de los otros marcadores en el panel (>95%) o la correlación es baja. Mediante el uso de QMSim, se simularon los datos en los que la correlación estadística entre los marcadores era baja, y en los datos de maíz experimentales utilizadas como ilustración, los niveles de ancestría y de LD también fueron bajos. Cuando se espera fuerte coancestría, las medidas de similitud que incluyen el concepto de ligamiento pueden proporcionar un beneficio adicional durante el proceso de conglomeración (Lawson y Falush, 2012).

El presente estudio también mostró que el método de conglomerado no jerárquico K-means se desempeña relativamente bien con respecto a SOM-RP-Q bajo diferentes escenarios. Resultados similares fueron reportados por Lee *et al., (*2009), sin embargo el uso del conglomerado no-jerárquicas K-means no tuvo éxito al reconocer la EGP en el conjunto de datos experimentales.

En base a los resultados de este trabajo y para inferir EGP desde datos de marcadores moleculares, se recomienda utilizar el algoritmo SOM-RP-Q y el implementado en el software STRUCTURE, especialmente en poblaciones de mapeo caracterizadas por baja divergencia genética. SOM es un método eficiente para analizar sistemas gobernados por relaciones no lineales complejas y ofrece una alternativa a los métodos estadísticos tradicionales de clasificación de datos complejos (Park *et al., 2*003a, Worner y Gevrey, 2006). SOM ha sido utilizado para el reconocimiento de patrones y la agrupación y visualización de grandes conjuntos de datos multidimensionales. Worner y Gevrey (2006) encontraron que SOM fue capaz de reducir datos de alta dimensionalidad rescatando patrones de utilidad para la interpretación del problema. El algoritmo SOM-RP-Q no sólo pudo reducir la dimensionalidad del conjunto de datos de marcadores, sino que también proporcionó información sobre los perfiles de marcadores típicos de cada grupo.

CAPITULO III

MODELOS DE MAPEO ASOCIATIVO

INTRODUCCIÓN

En los últimos años se ha incrementado el uso del mapeo asociativo (MA) para la identificación de genes responsables de características complejas de interés agronómico y adoptada en el mejoramiento de especies vegetales para el análisis de los *loci* de caracteres cuantitativos o QTL (Remington *et al.*, 2001; Kraakman *et al.*, 2006; Aranzana *et al.*, 2005; Breseghello y Sorrells, 2006; D'hoop *et al.*, 2008; 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 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 el mapeo asociativo, todas tendientes a controlar el aumento en la detección de asociaciones espurias. El modelo de MA básico es un modelo de regresión lineal múltiple, donde el fenotipo se asocia a los múltiples marcadores a través de 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 (Wang *et al.*, 2012; Cappa *et al.*, 2013; Muñoz-Amatriaín *et al.*, 2014).

Por ejemplo, cuando se usa STRUCTURE (Pritchard *et al.*, 2000) para estimar la probabilidad de pertenencia de cada individuo a las sub-poblaciones que componen la meta-población, la información resultante de la clasificación (i.e. las probabilidades de

pertenencia de cada individuo a cada conglomerado) puede ser incorporada al modelo como covariables (Yu *et al.*, 2006; Gutiérrez *et al.*, 2011). Estas covariables que proveen información que dimensiona la EGP, se dispone en una matriz de diseño usualmente denominada matriz Q. Alternativamente, se suele usar con el mismo propósito otra matriz, conocida como matriz P, compuesta por las componentes principales resultantes del análisis de componentes principales (ACP) (Hotelling, 1936) realizado sobre la matriz de datos de marcadores genéticos (Price *et al.*, 2006). Las CPs significativas, según la prueba de Tracy-Widom (1994), o las primeras CP (las que explican mayor porcentaje de la variabilidad total de los datos moleculares) son usadas como covariables en el modelo de mapeo asociativo (Peña Malavera *et al.*, 2014b). Tales covariables pueden contemplarse en el modelo de mapeo como efectos fijos o aleatorios, para este último caso la estimación del modelo se realiza en el contexto de un modelo lineal mixto (MLM) (Demidenko, 2004). 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 y del nivel de LD subyacente. 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 de especies genotipadas con cantidades relativamente bajas de marcadores. El objetivo de estudio en este Capítulo fue comparar el desempeño, a nivel de las tasas de falsos positivos y también a nivel de potencia para la detección de QTL de diferentes modelos biométricos de mapeo asociativo bajo escenarios de baja y alta estructuración genética y con dos niveles de densidad de marcadores (baja y alta).

Los modelos de MA comparados incluyen uno que no contempla ninguna corrección por EGP (Modelo *Naive*), otro con corrección por estructura genética mediante la matriz Q derivada de la clasificación difusa por modelo bayesiano resultante de la aplicación del software STRUCTURE sobre los datos de marcadores moleculares y otro basado en el uso de una matriz P derivada de un ACP aplicado también sobre los datos moleculares y una selección posterior de CPs significativas según el estadístico TW (Tracy y Widom, 1994). Ambos tipos de covariables (matriz Q y matriz P) fueron introducidas al modelo como efectos fijos y alternativamente como efectos aleatorios. El estudio también incluyó un modelo con corrección por parentesco mediante la matriz de parentesco genético conocida como matriz K obtenida con la librería EMMA de R propuesta por Kang *et al.*, (2008)

asociada a un vector aleatorio de efectos poligénicos de *background*, la matriz K es una matriz de similitudes entre los perfiles moleculares individuales que contienen información clave para construir la matriz de varianzas y covarianzas del vector de efectos poligénicos. Así mismo, se evaluaron estrategias combinadas como son los modelos denominados QK y PK que incluyen corrección por estructura en un modelo de efectos fijos más la matriz de parentesco K en la parte aleatoria del modelo lineal mixto usado para MA. La comparación se realizó usando bases de datos de marcadores moleculares simulados bajo distintos escenarios biológicos de EGP y bases de datos de marcadores moleculares reales conformada por 8 subpoblaciones (Hansey *et al.,* 2011).

MATERIALES Y MÉTODOS

DATOS SIMULADOS

Los datos de marcadores moleculares fueron simulados usando QMSim (Sargolzaei y Schenkel, 2009) e involucrado escenarios con *n* cantidad de genotipos y *p* cantidad de marcadores moleculares que imitan datos usuales en mejoramiento genético vegetal. Se simuló un genoma con 300 y otro con 3000 marcadores multiloci-bialélicos, con diseño de cruzamientos y selección aleatorios para una EGP conformada por cinco poblaciones. Se crearon para ambos casos de números de marcadores, cuatro escenarios biológicos, correspondientes a dos niveles de divergencia genética entre poblaciones (bajo y alto F_{ST}) y dos distintos tamaños de poblaciones de mapeo ($n \approx 150$ y 300). Los datos simulados fueron creados a partir de una población histórica de 200 individuos y el sistema de cruzamiento fue basado en la unión de gametas muestreadas aleatoriamente para muchas generaciones. La coancestría promedio fue baja como sucede en numerosas poblaciones que se usarán para MA en vegetales. Variando el número de generaciones desde la población fundadora, se crearon diferentes niveles de divergencia genética poblacional, para lograr un nivel de divergencia genética bajo se simularon 10 generaciones, para el nivel medio se simularon 40 generaciones y para lograr una alta divergencia fueron necesarias 70 simulaciones. 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 3.1).

Tabla 3.1. Número de poblaciones y diversidad que caracteriza la estructura genética subyacente en poblaciones de mapeo simuladas con 300 y 3000 marcadores multilocus-multialélicos como dato genómico.

Escenario	Número de marcadores	Diversidad genética		Tamaño poblacional promedio
		Estadístico F_{ST}	Nivel	
I	300	0.03	Bajo	150
II	300	0.03	Bajo	300
III	300	0.20	Alto	150
IV	300	0.21	Alto	300
V	3000	0.04	Bajo	150
VI	3000	0.03	Bajo	300
VII	3000	0.20	Alto	150
VIII	3000	0.20	Alto	300

Dada la matriz de marcadores moleculares resultante de QMSim, para los escenario del I al IV, correspondientes a los escenario con 300 marcadores, 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, 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.

Para los escenarios del V al VIII, correspondientes a situaciones con 3000 marcadores moleculares se siguió el mismo procedimiento pero se escogieron aleatoriamente 200 marcadores a los cuales se les asignó un efecto por muestreo aleatorio desde una distribución $\Gamma(2,100)$ y además a cada genotipo molecular se le asignó la realización de una variable aleatoria independiente distribuida como una $\sim N(1000, 2500)$.

DATOS MOLECULARES REALES

Los modelos de MA también se evaluaron usando un conjunto de datos de marcadores moleculares provisto por Natalia de León publicado por Hansey *et al.,* (2011), este archivo cuenta con n=334 líneas de maíz genotipadas mediante p=210 marcadores SNPs. En dicho trabajo sobre estos datos genéticos, Hansey *et al., (*2011) identificó ocho conglomerados o sub-poblaciones que fueron verificadas desde el conocimiento biológico de investigaciones genéticas de maíz. Todos los análisis sobre los datos reales se hicieron asumiendo la existencia de 8 conglomerados en los datos genéticos. Para implementar el MA se simuló para cada uno de estos genotipos un valor fenotípico adicionando una variable aleatoria $\sim N(40,144)$ 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 carácter fenotípico cuya denotación se presentan en la Tabla 3.2. El modelo básico a partir del cual derivan los modelos de MA comparados, es:

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

donde y es el vector de valores fenotípicos (conteniendo un dato fenotípico por genotipo), X es la matriz de datos de los marcadores moleculares (la cantidad de columnas es igual a la cantidad de marcadores usados), β 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, 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 σ_e^2. Se supone que el vector u se distribuye independientemente del vector e y con matriz de varianzas y covarianzas dada por $\sigma_e^2 * K$, siendo K la matriz de similitud entre 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.

Tabla 3.2. Ocho modelos comparados en datos reales y simulados.

Matriz de Parentesco	Estructura Genética Poblacional				
	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 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).

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 (R Core Team, 2013). En el Anexo 2 se presenta el código para el ajuste de estos modelos de mapeo asociativo, el cual invoca desde la interfaz mencionada los *scripts* desarrollados por Gutiérrez (2011). El desempeño de los modelos se evaluó usando como criterio la tasa de falsos descubrimientos o FDR (del inglés, *False Discovery Rate*) (Benjamini y Hochberg, 1995), la potencia estadística y las curvas de distribución acumulada de valores-p.

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 esta tesis fue calculada de la siguiente manera:

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

Finalmente, 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 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.

RESULTADOS

En las Figuras 3.1 y 3.2, se presentan los histogramas de la variable aleatoria que expresa la característica fenotípica simulada bajo distintos escenarios de datos moleculares. Se observa que los datos fenotípicos se distribuyen con una función de densidad simétrica esperable en función de las distribuciones usadas en la simulación. La variabilidad del

fenotipo es similar en los cuatro escenarios creados con 300 marcadores moleculares (Fig. 3.1).

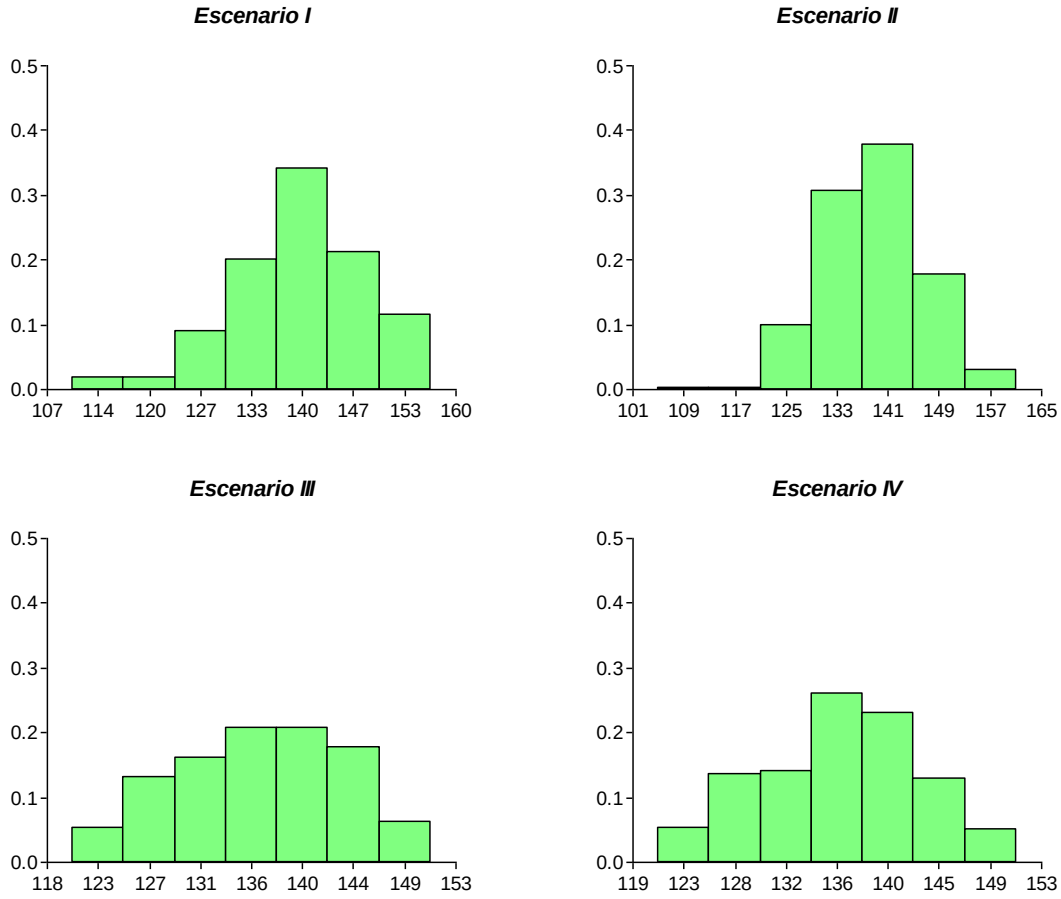

Figura 3.1. Histogramas para la variable fenotipo en cuatro escenarios creados vía simulación con 300 marcadores moleculares. Los cuatro escenarios corresponden a los presentados en la Tabla 3.1.

También se observa similar simetría en la distribución de datos fenotípicos de los cuatro escenarios generados con 3000 marcadores moleculares (Fig. 3.2), solo que los valores fenotípicos creados a partir de la suma de los efectos de loci de marcadores ligados a QTL son mayores a los de escenarios con 300 marcadores porque la suma de sus efectos tiene mayor cantidad de términos y la simulación se realizó adicionando una normal de mayor media. No obstante la varianza se incrementó de manera de mantener variabilidades relativas similares en los distintos escenarios.

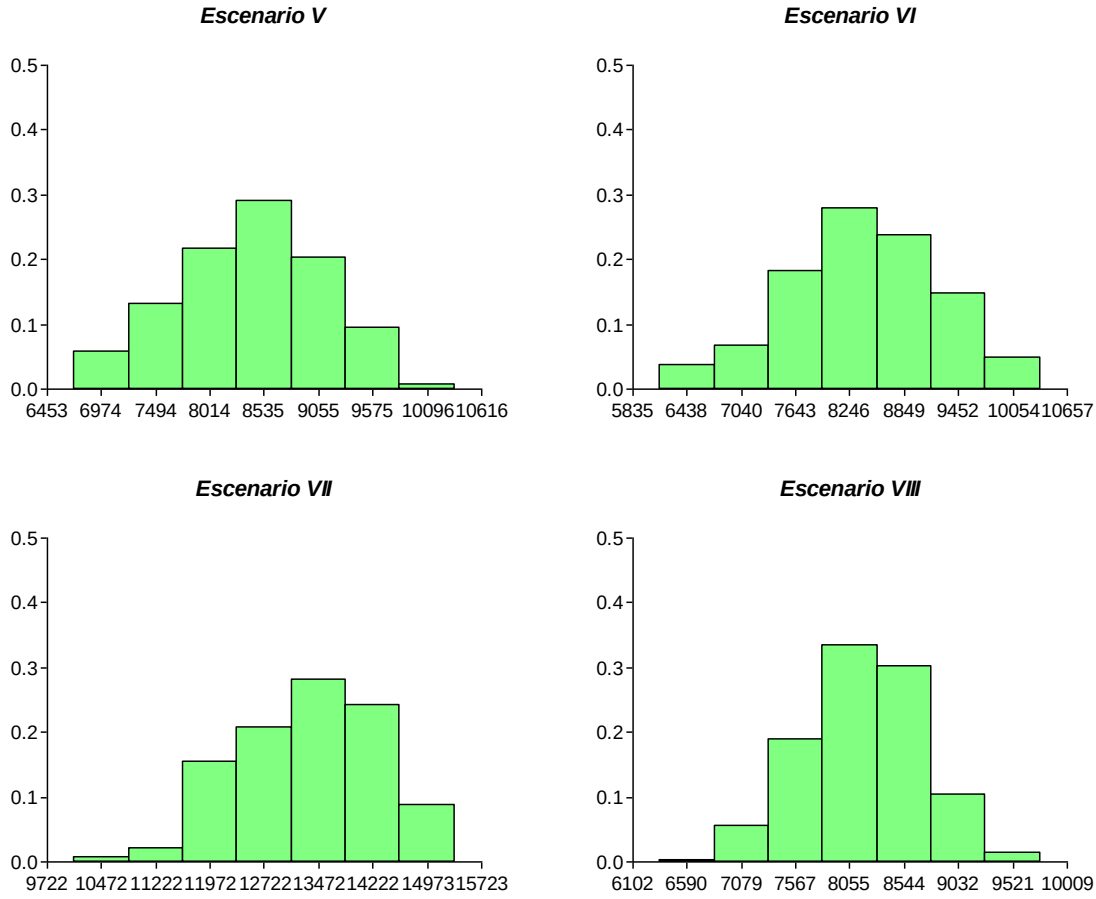

Figura 3.2. Histogramas para la variable simulada para representar el fenotipo en cuatro escenarios creados vía simulación con 3000 marcadores moleculares. Los escenarios corresponden a los descritos en la Tabla 3.2**.**

ANÁLISIS DE DATOS GENÉTICOS SIMULADOS

En la Figura 3.3 se muestran los gráficos de dispersión de los dos primeros ejes resultantes del escalamiento multidimensional métrico (Gower, 1967) obtenido desde los datos moleculares observado bajo cada uno de 4 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 los escenarios I y II se observa baja divergencia genética mientras que en los escenarios III y IV, los grupos o subpoblaciones que estructuran la población se presentan más distanciados por la mayor divergencia genética con la que fueron simulados.

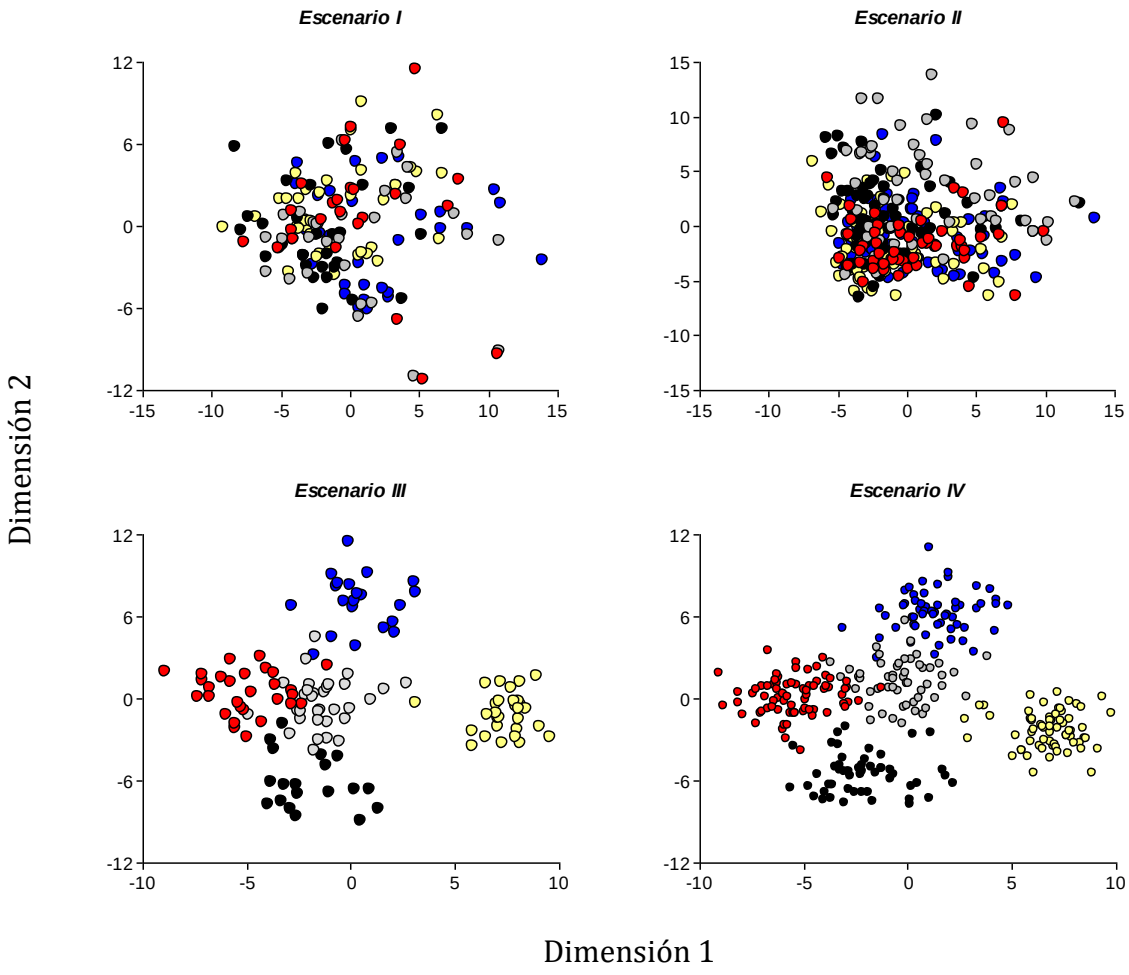

Figura 3.3. Gráficos de dispersión de los dos primeros ejes resultantes de un análisis de coordenadas principales (escalamiento multidimensional) de los datos moleculares (300 MM). En la columna de la izquierda tamaño poblacional de 150, mientras en la columna de la derecha para tamaño poblacional de 300. Arriba bajo F_{ST} y abajo alto F_{ST}. Los colores identifican los cinco grupos que definen la EG

En la Figura 3.4 se muestran los gráficos de dispersión de los dos primeros ejes resultantes del escalamiento multidimensional de los datos de cuatro escenarios con diferentes niveles de F_{ST} con 3000 marcadores moleculares. Al igual que con el caso de 300 marcadores mostrados arriba, en las figuras correspondientes a escenarios de mayor divergencia genética las subpoblaciones aparecen más distanciadas que en los escenarios de menor divergencia genética para un mismo tamaño poblacional. Con 3000 marcadores se observa mejor que con 300, la variabilidad entre subpoblaciones que implica un mismo valor de F_{ST}, la cual no es la misma para los dos tamaños poblacionales. Mayor distanciamiento entre subpoblaciones existe en poblaciones de menor tamaño con el mismo valor de F_{ST}.

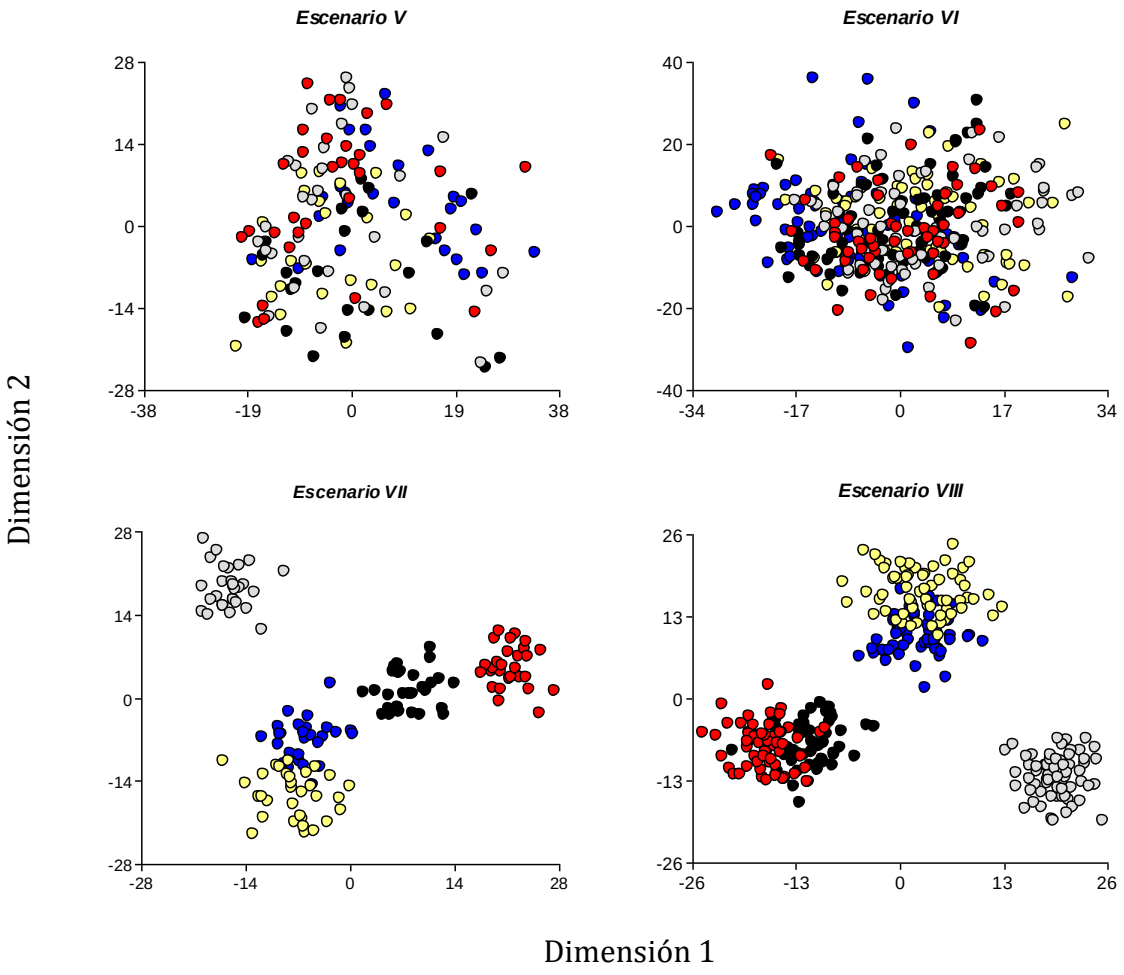

Figura 3.4. Gráficos de dispersión de los dos primeros ejes resultantes de un análisis de coordenadas principales (escalamiento multidimensional) de los datos moleculares (3000 MM). En la columna de la izquierda tamaño poblacional de 150, mientras en la columna de la derecha para tamaño poblacional de 300. Arriba bajo F_{ST} y abajo alto F_{ST}.

EVALUACIÓN DE MODELOS DE MAPEO ASOCIATIVO

ESTRUCTURA GENÉTICA SIMULADA

En la Figura 3.5 se muestran las funciones de distribución acumulada para los 4 escenarios que involucran datos genéticos de 300 marcadores moleculares. Se puede ver que consistentemente en los cuatro escenarios los modelos con mejor ajuste son el modelo K y el modelo QK.

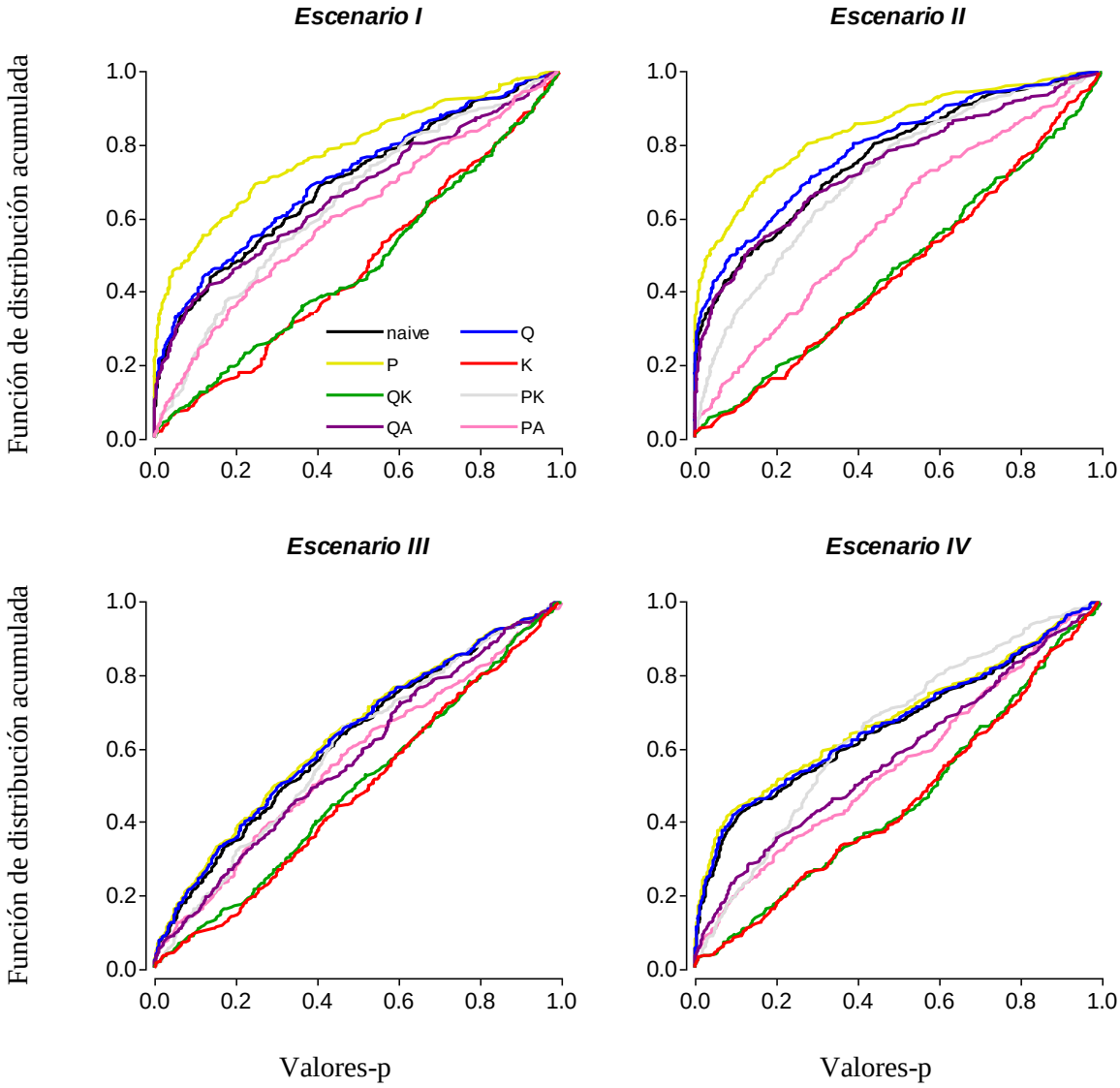

Figura 3.5. Gráfico de distribución acumulada de los valores-p de los ocho modelos evaluados en los escenarios que contienen 300 marcadores moleculares. En la columna de la izquierda escenarios con tamaño poblacional de 150 y en la columna derecha tamaño poblacional de 300. Arriba F_{ST} bajo y abajo F_{ST} alto.

Con bajo F_{ST} (arriba, escenario I y II) se observa que el modelo de menor desempeño fue el modelo P (Fig. 3.5). Los modelos se comportan de manera parecida cuando se trabaja en el contexto de alto F_{ST} (escenario III y IV), situación en la que se observa 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.

En la Figura 3.6 se muestran las funciones de distribución acumulada para los 4 escenarios correspondientes a los casos que involucran 3000 marcadores moleculares. Se puede ver

que consistentemente en los cuatro escenarios los modelos con mejor ajuste son nuevamente el modelo K y el modelo QK. Con bajo F_{ST} (arriba, escenario V y VI) se observa que el modelo de menor desempeño fue nuevamente el modelos que intenta modelar la estructura incluyendo como covariables de efectos fijos a variables sintéticas derivadas de un ACP (modelo P). Existen para este modelo una cantidad mayor a la esperada de valores-p pequeños, es decir mayor probabilidad de detecciones falsas. El peor de los escenarios fue el de menor nivel de EGP (escenario VI). Los modelos *Naive* y Q también mostraron alta asimetría hacia valores-p pequeños. Con bajo nivel de EGP, los modelos que incorporaron más parámetros en su estructura de media (efectos fijos) para modelar estructura no se desempeñaron bien. Por el contrario, cuando la estructura de incorpora a las varianzas-covarianzas del modelo, como es el caso de los modelos que incluyen la matriz K e incluso los que incluyen covariables de efectos aleatorios, principalmente PA 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. Para el modelado de EGP de manera combinada, es decir tanto a nivel de la estructura de medias (efectos fijos) como de la estructura de varianzas-covarianzas del modelo (efectos aleatorios), el uso de covariables derivadas de la clasificación bayesiana difusa que produce STRUCTURE fue siempre mejor que el uso de covariables derivadas de un ACP como estrategia que acompaña la incorporación de la matriz K para los efectos aleatorios de poligen. Aunque, es importante notar que la estrategia de modelado conjunto no mejoró el desempeño del modelo que solo usa la matriz K en términos de valores-p en ninguno de los ocho escenarios simulados.

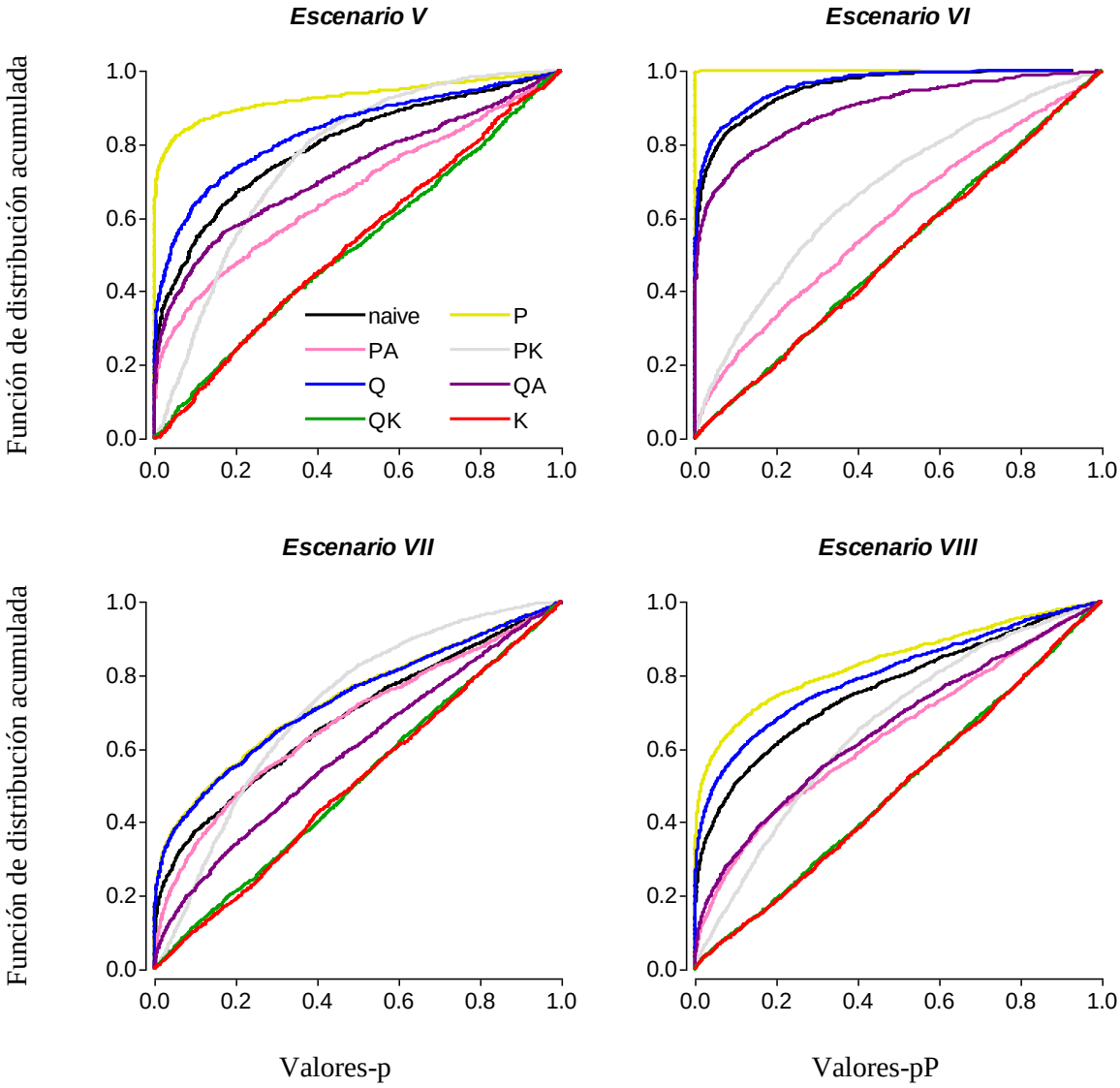

Figura 3.6. Gráfico de distribución acumulada de los valores-p de los ocho modelos evaluados en los escenarios que contienen 3000 marcadores moleculares. En la columna de la izquierda escenarios con tamaño poblacional de 150 y en la columna derecha tamaño poblacional de 300. Arriba F_{ST} bajo y abajo F_{ST} alto.

ESTRUCTURA GENÉTICA REAL

En la Figura 3.7 se muestran las funciones de distribución acumulada de los valores-p correspondiente a las pruebas de hipótesis implementadas para un conjunto de datos experimentales de maíz estructurado en 8 subpoblaciones (Hansey *et al.*, 2011). Por simplicidad y para facilitar la visualización de los principales resultados se muestran sólo los valores-p menores a 0.01.

Se observa que los modelos, excepto por PK, tienen un comportamiento relativo similar. Todos muestran menor cantidad de la esperada de estos valores-p, comportándose similar al modelo que no implementa ninguna corrección por estructura (modelo *Naive*). Aún cuando se ha reconocido previamente la existencia de 8 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 valores-p pequeños). 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 8 subpoblaciones (F_{ST}=0.02) o a una variabilidad relativamente alta de los datos fenotípicos que enmascara el efecto de los QTL.

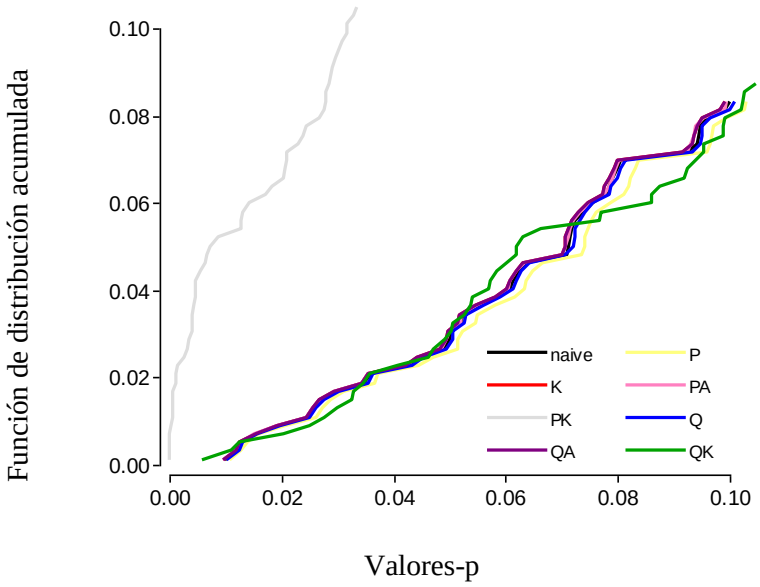

Figura 3.7. Gráfico de distribución acumulada de los valores-p de los ocho modelos evaluados en datos de 511 marcadores SNP sobre 504 genotipos de maíz estructurados genéticamente con un nivel de F_{ST} relativamente bajo (F_{ST} =0.02).

TASAS FDR Y POTENCIA

DATOS SIMULADOS

En la Tabla 3.3 se observa que, para los escenarios con 300 marcadores y n=150, 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 falsas detecciones.

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, de estos últimos el modelo QK fue el de menor pérdida relativa de potencia si se considera la disminución de FDR que este provocó.

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

En la Tabla 3.4 se observan los mismos resultados para situaciones donde el tamaño de la población aumenta (n=300). Comparando los escenarios con n=150 con los que involucran un mayor tamaño de la población de mapeo (n=300), se observa a partir del modelo *Naive* un resultado también esperado desde la teoría de errores tipo I y tipo II en pruebas de hipótesis. Este es que el aumento del tamaño de muestra (en este caso de 150 a 300 individuos en la población de mapeo) no se asocia con la probabilidad de error tipo I y por tanto no modifica la tasa FDR, pero sí implica una mayor potencia para la detección de QTL, *i.e.* menor probabilidad de error tipo II.

Tabla 3.3. Tasas de falsos positivos y potencia de ocho modelos de mapeo asociativo para dos niveles de estructura genética poblacional (Bajo y Alto F_{ST}), 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 a través de CPs 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 y PA: con la matriz P como covariables de efectos aleatorios.

Con 300 marcadores y n=300 se verifica nuevamente 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 un modelo de covariables de efectos fijos, tanto con la matriz Q (0.34) como con la matriz P (0.37) incrementa los falsos positivos con respecto al modelo *Naive* (0.31). Los modelos K y QK son los modelos con menor tasa de FDR (0.05 y 0.07, respectivamente) en escenarios con bajo F_{ST} y 0.04 y 0.02 con altos valores de F_{ST}. El modelo PA provoca una buena disminución de la FDR y su potencia no es tan baja.

Tabla 3.4. Tasas de falsos positivos y potencia de ocho modelos de mapeo asociativo para dos niveles de estructura genética poblacional (Bajo y Alto F_{ST}) y 300 marcadores moleculares y n=300.

	FDR		Potencia	
Modelo	Bajo F_{ST}	Alto F_{ST}	Bajo F_{ST}	Alto F_{ST}
Naive	0.31	0.25	0.75	0.75
Q	0.34	0.25	0.80	0.80
P	0.37	0.29	0.85	0.80
K	0.05	0.04	0.40	0.45
QK	0.07	0.02	0.45	0.45
PK	0.27	0.16	0.55	0.45
QA	0.32	0.15	0.75	0.65
PA	0.12	0.10	0.55	0.65

*_Naive_: sin corrección por estructura, Q: con corrección mediante la matriz de probabilidades a posteriori obtenida con el software STRUCTURE, P: con corrección a través de CPs 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 y PA: con la matriz P como covariables de efectos aleatorios.

En la Tabla 3.5 se observa que los modelos que involucran la matriz K en el análisis fueron los únicos que disminuyeron los valores de FDR para los dos niveles de F_{ST} y de éstos el modelo PK tuvo la menor perdida de potencia.

Tabla 3.5. Tasas de falsos positivos y potencia de ocho modelos de mapeo asociativo para dos niveles de estructura genética poblacional (Bajo y Alto F_{ST}), 3000 marcadores moleculares y n=150.

	FDR		Potencia	
Modelo	Bajo F_{ST}	Alto F_{ST}	Bajo F_{ST}	Alto F_{ST}
Naive	0.45	0.42	0.51	0.40
Q	0.47	0.44	0.60	0.47
P	0.49	0.44	0.86	0.47
K	0.23	0.17	0.16	0.20
QK	0.25	0.18	0.17	0.23
PK	0.33	0.22	0.26	0.32
QA	0.44	0.32	0.48	0.29
PA	0.48	0.38	0.32	0.36

**Naive*: sin corrección por estructura, Q: con corrección mediante la matriz de probabilidades a posteriori obtenida con el software STRUCTURE, P: con corrección a través de CPs 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 y PA: con la matriz P como covariables de efectos aleatorios.

Dado el aumento en pruebas de hipótesis debida al incremento en la densidad del genotipado (300 vs 3000 para las situaciones anteriores), se registró un incremento en la tasa FDR para ambos niveles de divergencia genética. Nuevamente, los modelos que involucran la matriz K en el análisis fueron los que mejor controlaron la tasa FDR pero con pérdidas de potencia significativas. Ninguno de los modelos de MA tuvo un desempeño aceptable si se consideran simultáneamente las tasas de error tipo I y tipo II.

Tabla 3.6. Tasas de falsos positivos y potencia de ocho modelos de mapeo asociativo para dos niveles de estructura genética poblacional (Bajo y Alto F_{ST}), 3000 marcadores moleculares y n=300.

	FDR		Potencia	
Modelo	Bajo F_{ST}	Alto F_{ST}	Bajo F_{ST}	Alto F_{ST}
Naive	0.48	0.42	0.85	0.65
Q	0.48	0.44	0.88	0.64
P	0.50	0.45	1.00	0.71
K	0.15	0.12	0.27	0.29
QK	0.13	0.14	0.29	0.28
PK	0.24	0.16	0.44	0.47
QA	0.45	0.31	0.80	0.47
PA	0.31	0.30	0.28	0.42

**Naive*: sin corrección por estructura, Q: con corrección mediante la matriz de probabilidades a posteriori obtenida con el software STRUCTURE, P: con corrección a través de CPs 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 y PA: con la matriz P como covariables de efectos aleatorios.

En la Tabla 3.6 se observa que el incremento en el número de individuos de la población de mapeo (n=300) no impactó positivamente la tasa de FDR para escenarios de alta densidad de marcadores (3000). Si bien los modelos que involucran la matriz K disminuyeron la tasa FDR respecto al modelo *Naive*, la pérdida de potencia fue significativa (a excepción de PK). Los modelos de efectos aleatorios (QA y PA) disminuyeron los valores de FDR, no más que los modelos que involucran la matriz K, pero con menor pérdida de potencia.

ESTRUCTURA GENÉTICA REAL

En la Figura 3.8 se presenta un gráfico de barras en el cual se muestra la diferencia que existe entre la cantidad de *loci* de QTL existentes en el conjunto de datos de marcadores

SNP de maíz usado para esta comparación de modelos y la cantidad de marcadores identificados como significativos por MA. En la Figura 3.8 no se ha incluido el modelo PK por presentar fuerte sobre-estimación de la cantidad de *loci* significativos (más del doble de los que realmente definieron el fenotipo). La cantidad de componentes retenidas 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 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.

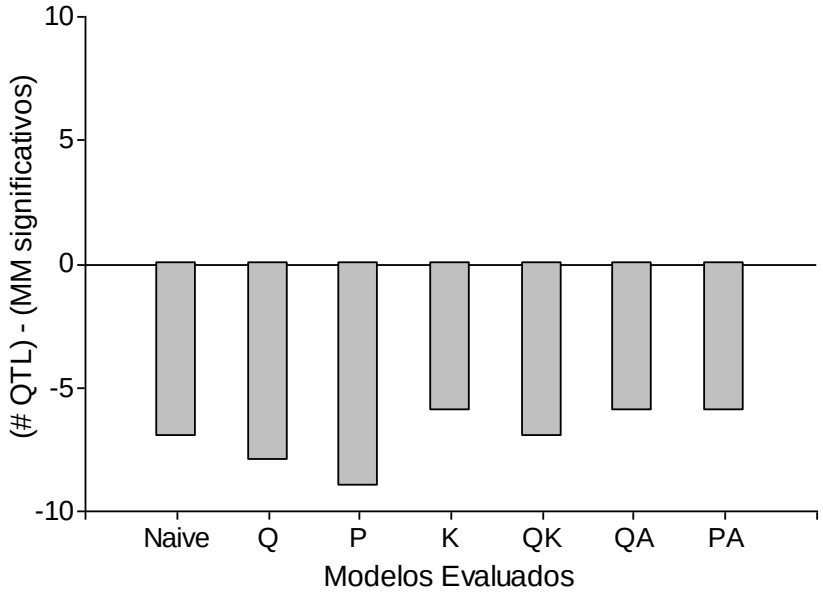

Figura 3.8. Diferencia entre la cantidades de QTL simulados y la cantidad de marcadores encontrados como significativos para distintos modelos de MA evaluados (ver códigos de modelos en Tabla 3.2).

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 del 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 perdida de potencia en los escenarios de estructura genética simulados. No obstante, mostró tanto en situaciones de EGP simulada como real, menor capacidad de controlar FDR que K y QK, sobre todo en situaciones donde la EGP es baja y se demanda un número alto de componentes principales para alcanzar una explicación significativa de la variabilidad molecular en la población de mapeo. Aún cuando en los escenarios de bajo F_{ST} los modelos P y Q mostraron más potencia que el modelo sin corrección (*Naive*) para detectar QTL, 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 marcador-carácter si es que 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 las CPs 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*, 2014b). 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 CPs 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 estudiados. Es importante notar que la EGP era alta y fue bien identificada en el trabajo con eucaliptus; tanto con STRUCTURE como con ACP donde las dos primeras componentes principales explicaron cerca del 50% de la variabilidad de los perfiles moleculares. Sólo con dos CPs se pudo ordenar claramente los genotipos en tres grupos esperados según el conocimiento previo de sus procedencias. Con alto nivel de EGP y pocas CP, el modelo P puede resultar una buena estrategia, sobre todo para la detección de verdaderos QTL.

La incorporación de las componentes principales como covariables de efectos aleatorios (modelo PA) disminuyó también en nuestro trabajo 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 variables de salida del software STRUCTURE (Pritchard *et al.,* 2000). En este trabajo, el modelo PA siempre controló mejor que el modelo Q la tasa FDR. 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 software STRUCTURE. Wang *et al.,* (2012) también demostraron la efectividad del modelo con componentes principales aleatorias (PA) para contemplar la estructura poblacional previa al MA y disminuir los FP.

CAPITULO IV

AJUSTES DE VALORES-P POR MULTIPLICIDAD DE PRUEBAS DE HIPÓTESIS

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. 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. 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.

Los estudios sobre modelos estadísticos para MA, son variados y las propuestas se encuentran aún en activo desarrollo, lo cual dificulta la elección del modelo de análisis más apropiado para un escenario particular. Los resultados del capítulo previo sugieren que la aplicación de uno u otro modelo debe contemplar aspectos estadísticos como el tamaño de la muestra y número de variables indicadoras o marcadores, además de aspectos biológicos como son el nivel de divergencia genética que estructura la población de mapeo.

En todos los modelos de MA será necesario realizar múltiples pruebas de hipótesis estadísticas. La teoría sobre contrastes de hipótesis provee un protocolo metodológico que involucra, como primer paso, el planteo de una hipótesis H_0 sobre los parámetros del modelo. En el caso de los modelos de regresión usados en MA, en H_0 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 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 un estadístico apropiado para evaluar la significancia estadística de un coeficiente de regresión (Drapper y Smith, 1998). Luego, es necesario calcular el valor del estadístico en la muestra que se tenga. Si el valor de dicho estadístico es anómalo respecto de lo que se esperará 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 muy pequeños habrán así de entenderse como evidencia en contra de la hipótesis nula objeto de contraste. En MA, valores-p pequeños llevan al rechazo de la hipótesis 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 α.

Dos criterios de evaluación cobran importancia para evaluar una prueba de hipótesis estadística: la capacidad de mantener su tamaño nominal o nivel de significación α, 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 α y, el segundo, con el error tipo II con probabilidad de ocurrencia denotada por β (Balzarini *et al.*, 2012). 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 prueba 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 mapeo asociativo (MA) se ajustan modelos de regresión con un número grande de parámetros relativos a los tantos marcadores que se evalúan de forma simultánea, 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 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 regresión múltiple y bajo H_0, hay probabilidad tan sólo α de que el estadístico T calculado exceda en valor absoluto del cuantil α/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: Prob(Algún βi distinto de 0) = $1 - (1 - \alpha)^m$. Luego, con probabilidad mucho mayor a α, algún coeficiente de regresión puede resultar significativo sólo por azar. Esta probabilidad depende de M aumentando a medida que aumenta el número de marcadores moleculares evaluados. Este problema, conocido como problema de inferencia simultánea que demanda correcciones de los valores-p por multiplicidad, debe ser atendido en el contexto de MA para no perder potencia (Xiao *et al.*, 2013). Para el contexto de pruebas independientes existes 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, usado en varias áreas del conocimiento. 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 (1995) para controlar la proporción esperada de hipótesis mal rechazadas respecto a todas aquellas rechazadas (Sabatti *et al*., 2003, Tusher *et al*., 2001, Miller *et al.*, 2001, Schwartzman *et al*., 2008).

El umbral de significación nominal α es inapropiado para reportar resultados del MA 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, Cheverud propuso una idea de ajuste para pruebas correlacionadas ajustando los valores-p luego de determinar un número efectivo (Meff) de pruebas independientes (Cheverud, 2001). Li y Ji (LJ) (2005) propusieron una estimación más exacta del Meff basada en la descomposición por valor singular de una matriz de correlaciones entre marcadores, y diseñaron un procedimiento basado en este nuevo Meff para controlar el error tipo I, que ha sido usado en el contexto

del análisis de QTL clásico. El objetivo de trabajo en este Capítulo, es evaluar el desempeño de diferentes métodos de ajustes de valor-p por multiplicidad cuando ellos son aplicados luego de un modelo de MA y escenario biológicos específico.

Se propone una corrección por multiplicidad 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 Hanzel (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 ajustes de valores-p se realizaron luego de aplicar diferentes estrategias de modelación para MA descriptas en el Capítulo 3 de esta tesis. Para cada combinación modelo de MA-método de ajuste de valor-p, se obtuvieron tasas de falsos positivos y potencia (φ), bajo dos escenarios con diferente nivel de EGP (bajo y alto F_{ST}).

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 multiloci-bialelicos, 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 distintos tamaños de poblaciones de mapeo (n≈150 y 300). Los datos simulados fueron creados a partir de una población histórica de 200 individuos y el sistema de cruzamiento fue basado en la unión de gametas muestreadas aleatoriamente para muchas generaciones. La coancestría promedio fue baja como sucede en numerosas poblaciones que se usarán 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 3.1).

Tabla 3.1. Tamaño poblacional y diversidad que caracteriza la estructura genética subyacente en poblaciones de mapeo simuladas con 300 marcadores multi-locus multialélicos 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

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 (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. Los valores resultantes fueron usados como variable fenotípica para el MA y 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.

PROCEDIMIENTOS

Usando los datos de los 4 escenarios simulados consideramos el problema de contrastar simultáneamente m hipótesis nulas $H_{0j}, j = 1,...,m,$ 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 4.1. Los conjuntos de subíndices que corresponden a hipótesis nulas verdaderas y falsas $\Delta_0 = \{ j : H_{0j} \text{ es verdadera} \}$ y $\Delta_1 = \{ j : H_{0j} \text{ no es verdadera} \}$ son desconocidos y serán estimados mediante la simulación. El conjunto total de índices es $\Delta = \{1, 2, ..., m\} = \Delta_0 \cup \Delta_1$. Las cantidades de hipótesis nulas verdaderas $m_0 = \#\Delta_0$ y falsas $m_1 = m - m_0 = \#\Delta_1$, fueron estimadas por conteo dentro de cada escenario.

Tabla 4.1. Situaciones posibles luego de realizar *m* pruebas de hipótesis

Realidad	Decisión		Total
	hipótesis nula no rechazada	hipótesis nula rechazada	
Hipótesis nula verdadera	U	V (Falsos Positivos)	m_0
Hipótesis nula falsa	T (Falsos negativos)	S	m_1
Total	m-R	R	m

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). 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 pruebas de hipótesis relativas a marcadores que no fueron asociados a un QTL en la simulación de los datos. Los VP son todas aquellas pruebas de hipótesis significativas a marcadores positivos que efectivamente estaban asociados a uno de los QTL simulados:

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

La potencia estadística se interpretó como la probabilidad de no cometer error de tipo II (error que produce los eventos conocidos como falsos negativos, FN):

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

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 el propuesto Benjamini y Hochberg (1995),

el método propuesto por Li y Ji (2005) y un nuevo procedimiento propuesto en este trabajo que llamamos Li&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 dan en MA donde el número de pruebas de hipótesis es de varios cientos e incluso miles.

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

1. Los valores-p de las m pruebas de hipótesis se ordenaron desde 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 dividido por el valor que denota su orden en la lista de valores-p ordenados. Si el valor resultante era menor que 0.05, se rechazó la hipótesis nula.

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

1. Se calculó la matriz de correlación para todos los *loci*.
2. Se calculó el número efectivo (Meff) de pruebas independientes a través de la obtención de los valores propios de la matriz de correlación mencionada arriba, según ecuación (1), donde M es el número de pruebas y $\lambda_i (i = 1,...,M)$ son los valores propios

$$M_{eff} = \sum_{i=1}^{M} f(|\lambda_i|)$$
$$f(x) = I(x \geq 1) + (x - \lfloor x \rfloor), x \geq 0 \quad (1)$$

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

3. Se ajustó el nivel de significación de la prueba como si hubiera M_{eff} 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 de marcador y cuando el valor-p de alguna prueba era menor que α_p, la hipótesis de no asociación fue rechazada.

La propuesta que realizamos en este capítulo está basada en la aproximación de Li&Ji (2005) con una modificación pensada para contemplar la posible EGP que subyace la población de mapeo. En caso de poblaciones estructuradas, la modificación analizará la correlación entre marcadores para derivar un Meff pero controlando por la presencia de los grupos que definen la EG. Con este fin, la matriz de correlación utilizada en el método LJ es reemplazada por una matriz de estadísticos χ^2 de Mantel y Haenzel (1959). Los estadísticos χ^2 fueron obtenidos a partir de tablas de contingencia construidas entre cada par 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 realizaron considerando ambos niveles de F_{ST} usados en la simulación, ambos tamaños poblacionales y distintos modelos de MA para generar la lista de valores-p sin corregir. Los modelos ajustados fueron: QK y K (Yu *et a*l., 2006) y el modelo de mapeo de regresión múltiple 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

En la Tabla 4.2 se observan las tasas FDR para cuatro modelos de mapeo asociativo evaluados en el Capítulo 3 y para tres métodos de corrección por multiplicidad en condiciones de bajo y alto F_{ST}, se puede ver que las tasas son menores para alto F_{ST} en todas las correcciones. Cuando nos ubicamos en la situación de no corrección por estructura ni parentesco en el modelado, es decir cuando realizamos un modelo *Naive*, podemos observar que la tasa FDR baja con todas las correcciones respecto a sin corrección (SC), pero con alto F_{ST} baja en mayor medida 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 correcciones por multiplicidad, si bien las tasas igualmente bajan, no se ve de una forma tan drástica como en la situación de alta estructura genética.

Tabla 4.2. 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 Alto (F_{ST}=0.2) divergencia genética, con un tamaño poblacional de 150.

Modelo**	Bajo F_{ST}				Alto F_{ST}			
	Correcciones*							
	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&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 4.3. 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 Alto (F_{ST}=0.2) divergencia genética, con un tamaño poblacional de 300.

Modelo**	Bajo F_{ST}				Alto F_{ST}			
	Correcciones*							
	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&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.

En las Tablas 4.4 y 4.5 se presentan las potencias alcanzadas con cada combinación de modelo estadístico usado para derivar los valores-p y el método de ajuste de valores-p para determinar significancia estadística en el contexto de inferencia simultánea, tanto en escenarios de baja y alta divergencia genética entre las subpoblaciones que definen la EGP en la población de mapeo. Se observa que aún para el caso de no corrección por EG, 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 cuando las subpoblaciones tenían poca divergencia.

Cuando se ha usado el modelo K o QK como modelo de mapeo, la corrección de valores-p también produce reducciones importantes de potencia. Estas reducciones son 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 tests 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 la tasa FDR se encuentre asociada a pérdida de potencia. No obstante, la potencia con alto F_{ST} para el método MLJ fue igual o superior a la de los otros dos métodos de corrección de valores-p por multiplicidad.

Tabla 4.4. 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 150.

	Bajo F_{ST}				Alto F_{ST}			
Modelo	Correcciones							
	SC	BH	LJ	MLJ	SC	BH	LJ	MLJ
N	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&Ji Modificado. **N: *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 4.5. 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}			
Modelo	Correcciones							
	SC	BH	LJ	MLJ	SC	BH	LJ	MLJ
N	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&Ji Modificado. **N: *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.

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

DISCUSIÓN

Han sido propuestos y utilizados en mapeo de QTL clásico, un número importante de métodos para abordar el problema de inferencia simultánea presente en el estudio de asociaciones marcadores-fenotipo. No obstante, ninguno de ellos fue desarrollado pensando en pruebas no independientes a causa de lo que podría ser una EGP en la población de MA. Cheverud (2001), propuso estimaciones de dependencia en el contexto de querer controlar el error de tipo I, en poblaciones de mapeo de QTL clásico con LD

entre marcadores; a partir de este procedimiento se derivaron otras propuestas como la de Li y Ji (2005) que también fueron pensadas en el contexto de mapeo de QTL clásico. Piepho (2001) propuso una forma de calcular un umbral para valores-p para ser usado tanto en mapeo de QTL por intervalos simple como en mapeo por intervalo compuesto. La tasa FDR involucra un balance entre FP y VP, dependencia que puede ser problemática para su interpretación en el contexto de mapeo por ligamiento, debido a que los FP y VP se producen en grandes cluster o conjuntos de datos dependientes, por eso su uso para declarar asociaciones significativas para un carácter es dudoso (Chen y Sorey, 2006).

En este trabajo extendemos la propuesta de Li y Ji (2005) para contemplar la EGP subyacente también al momento de hacer el ajuste de valores-p.

La corrección de Bonferroni (1935), no fue incluida en el trabajo de comparación simultánea de métodos por ser una prueba demasiado conservadora en el contexto de MA. En su lugar, se incluyó BH (Benjamini y Hochberg, 1995), uno de los desarrollos metodológicos más importantes de pruebas de hipótesis múltiples que ha jugado un papel exitoso en muchos estudios de mapeo asociativo (Gutiérrez *et al.*, 2011; Wang *et al.*, 2012; Muñoz-Amatriaín *et al.*, 2014; Olukolu *et al.*, 2014). Sun y Cai (2009) propusieron un índice local de significación (LIS) como procedimiento de ajuste de valores-p por multiplicidad. Este método utiliza un modelo Markoviano para representar la estructura de dependencia en los datos y ha mostrado ser óptimo bajo ciertas condiciones así como un sólido desempeño empírico. No obstante, su aplicación aún no es de rutina en análisis de mapeo en vegetales.

El índice LIS se generalizó como un índice local de significación agrupada (PLIS) para el análisis de asociaciones en múltiples cromosomas o grupos de ligamiento (Wei *et al.*, (2009). Para su implementación el modelo Markoviano debe implementarse para cromosomas específicos. Su utilidad está siendo demostrada en análisis de datos de marcadores como SNPs derivados de GWAS a gran escala (Xiao *et al.*, (2013). Varios métodos han sido diseñados y probados en función de un número efectivo de pruebas estadísticas (Meff), aunque las evaluaciones se han realizado usando un número limitado de marcadores genéticos (Cheverud, 2001; Li y Ji, 2005; Nyholt, 2004). Recientemente, también se han propuesto otros métodos basados en reducir el número de dimensiones que explican la EGP y que tienen como objetivo el GWAS como el simpleM (Gao *et al.*, 2008) y Keff (Moskvina y Schmidt, 2008). El método simpleM no requiere el conocimiento de

ninguna distribución estadística (en este contexto suele presentarse como un método no paramétrico), sin embargo ha sido reportado como mejor que otros métodos basados en Meff (Gao *et al.*, 2010; Gao *et al.*, 2008). La propuesta realizada en este trabajo (MLJ) representa una extensión del modelo LJ para considerar la EGP que está presente en los datos genómicos que, como se observa a partir de los resultados de simulaciones de datos genéticos bajo distintos escenarios biológicos, ha representado una mejora en la tasa de FDR, principalmente cuando la EGP no ha sido incorporada en el modelo de mapeo a partir del cual se obtuvieron los valores-p para las pruebas de significancia de asociaciones.

COMENTARIOS FINALES

El mapeo asociativo se presentó como una técnica cuyo objetivo radica en identificar marcadores de herencia simple próximos a *loci* que afectan características cuantitativas. En los últimos años se ha incrementado el uso del mapeo asociativo para identificar QTL que codifican para los caracteres de interés y esta identificación se realiza mediante pruebas de hipótesis estadística de las asociaciones entre los marcadores y dichos caracteres basados en el LD.

Muchos autores afirman que los atractivos del MA consisten en la falta de cruzamientos diseñados entre parentales o líneas mejoradas, y en la posibilidad de identificar QTL en cultivos con difíciles patrones de segregación. A su vez, mencionan que para el MA se puede presentar más de dos alelos por locus y los alelos de QTL pueden ser mapeados en germoplasma que sea relevante para programas de mejoramiento (Malosetti *et al.* 2007). Por lo tanto, las ventajas ofrecidas por el MA consisten en mejorar la resolución del mapeo, disminuir el tiempo de investigación, considerar una población más amplia y mejorar el número alélico. Por otro lado, se destaca que ambos análisis se complementan en términos de proveer información de gran relevancia, validación de cruzamientos y poder estadístico (Yu y Buckler, 2006a).

Se ha mostrado en esta tesis, como en antecedentes de la literatura sobre MA, que la inclusión de la estructura genética es importante y altamente recomendada, es por esto que reconocer la estructura genética poblacional es un paso previo de vital importancia para el mapeo asociativo. En esta tesis se observó que SOM-RP-Q, un algoritmo de aprendizaje no supervisado, puede ser usado para descubrir EG. Su naturaleza basada en el algoritmo de red neuronal que lo constituye, permite realizar visualizaciones de la EGP de los datos moleculares en planos que preservan la topología de las agrupaciones de un espacio de mayor dimensión. El algoritmo SOM-RP-Q utilizado resultó más efectivo que otros métodos de conglomeración usuales evaluados simultáneamente sobre conjuntos de datos genéticos con bajo nivel de ligamiento tanto sean reales o simulados bajo distintos escenarios biológicos. Su desempeño fue próximo al del algoritmo bayesiano embebido en el software STRUCTURE el cual está bien difundido para estudios de EGP en colecciones

de germoplasma vegetal. Sin embargo, este último implica mayor demanda computacional para datos de alta dimensionalidad.

Los modelos de mapeo asociativo fueron evaluados teniendo en cuenta escenarios caracterizados bajo dos distintos niveles de estructuración genética factibles de ocurrir en poblaciones de mapeo usadas en el mejoramiento vegetal de cultivos y bajo dos tamaños poblacionales usuales en la práctica en nuestro medio. También se consideraron dos cantidades diferentes de marcadores moleculares para representar la densidad del mapeo. Las combinaciones de los factores usados en las simulaciones permitieron tener una visión abarcadora del desempeño de ocho modelos de mapeo asociativo. Se observó que los modelos que mejores resultados en cuanto a tasa de falsos positivos fueron los modelos K y QK, en escenarios con bajo y alto F_{ST}. Para los datos reales estos resultados fueron consistentes.

Luego de realizar el análisis de mapeo asociativo se vio que es de importancia aplicar ajustes de valores-p por multiplicidad de pruebas de hipótesis ya que esto permite mejorar el desempeño del análisis en vista de no seleccionar QTL para futuras investigaciones que pudieran ser falsos descubrimientos, lo que para un programa de mejoramiento puede ser de gran pérdida, tanto económica como de tiempo y recursos vegetales. Cuando se desconoce la estructura poblacional y el relacionamiento genético y no se pueden involucran a ciencia cierta en el modelo de MA, el ajuste por multiplicidad MLJ aquí propuesto surge como una herramienta de control de los falsos positivos. Cuando se conocen algunas de las posibles estructuras igualmente el MLJ puede ser usado con buenos resultados, mejorando los presentados por Benjamini & Hochberg (1995) y por Li & Ji (2005).

La propuesta MLJ se está evaluando con el objetivo de optimizarla computacionalmente e incluirla en el software *Info-Gen* (Balzarini y Di Rienzo, 2004) haciendo uso de la interfaz con el lenguaje R (Di Rienzo, 2010). En futuras líneas de investigación será necesario incluir como covariables de los modelos de mapeo asociativo a las frecuencias de asignación de individuos a los grupos de nodos de la red SOM-RP-Q como modelo alternativo a los modelos K y QK. Se espera que la investigación metodológica sobre el uso de estas aproximaciones en el contexto de mapeo asociativo y el desarrollo de herramientas de software que faciliten la implementación del protocolo de análisis completo en un único ambiente contribuyan a mejorar la calidad de la información

derivada de mapeos asociativos implementados para el mejoramiento genético vegetal de cultivos de importancia para el desarrollo de nuestras sociedades.

BIBLIOGRAFÍA

Agresti, A. (1990). *Categorical Data Analysis.* New York.: John Wiley & Sons, Inc., .

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., Bruno, C., Peña, A., Teich, I., Di Rienzo, J. (Eds) (2010). *Estadística en Biotecnología. Aplicaciones en Info-Gen.* Córdoba: Encuentro.

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

Balzarini, M., Macchiavelli, R., Casanoves, F. (2004).Documentación sobre Aplicaciones de Modelos Mixtos en Agricultura y Forestería. In *Curso de Capacitacion Centro Agronomico Tropical de Investigación y Enseñanza*, 210 Turrialba- Costa Rica: CATIE.

Balzarini, M., Milligan, S., MS., K. (2001).Best linear unbiased prediction: A mixed model approach in multi-environment trials (Chapter 12). In *Crop Improvement: Challenges in the 21st Century*, 102-113 (Ed M. Kang). Binghamton, NY. : Food Products Press Inc.

Balzarini, M., Teich, I., Bruno, C., Peña Malavera, A. (2011). Making genetic biodiversity measurable: A review of statistical multivariate methods to study variability at gene level. *Revista de la Facultad de Ciencias Agrarias de la Universidad Nacional de Cuyo*, 43: 261-275.

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

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.

Benjamini, Y., Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Stat* 29: 1165 - 1188.

Bernardo, R. (2008). Molecular Markers and Selection for Complex Traits in Plants: Learning from the Last 20 Years. *Crop Sci.* 48: 16.

Bernardo, R. (2013). Genomewide markers as cofactors for precision mapping of quantitative trait loci. *Theoretical and Applied Genetics* 126(4): 999-1009.

Bernardo, R. (2014). Genomewide Selection when Major Genes Are Known. *Crop Science* 54(1): 68-75.

Bernardo, R., Yu, J. (2007). Prospects for genome-wide selection for quantitative traits in maize. Crop Sci 47: 8

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

Bradbury, PJ., Zhang, Z., Kroon, DE., Casstevens, TM., Ramdoss, Y., Buckler, ES. (2007) TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23: 2633-2635.

Bradbury, P., Parker, T., Hamblin, M. T., Jannink, J.-L. (2011). Assessment of Power and False Discovery Rate in Genome-Wide Association Studies using the BarleyCAP Germplasm. *Crop Sci.* 51(1): 52-59.

Breen, G., Harold, D., Ralston, S., Shaw, D. and St. Clair, D. (2000). Determining SNP allele frequencies in DNA pools. *Biotechniques.* 28: 2.

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.

Bruno, C. (2009).Evaluación de métodos de análisis de datos de "marcadores" moleculares. Su aplicación en mejoramiento genético. In *FCA-UNC. Escuela para Graduados.* , Vol. Dr. en Ciencias Agropecuarias., 177 Córdoba: Argentina.

Bruno, C., Arroyo, A., Di Rienzo, J., Balzarini, M. (2003).Ordenamiento de Perfiles Moleculares RAPD: Una Aplicación del análisis Procrustes Generalizado. In *XXXII Congreso Argentino de Genética, IV Jornadas Argentino-Chilenas de Genética y XXXIV Reunión Anual Sociedad de Genética de Chile.*, S2-116 Huerta Grande, Córdoba, Argentina.

Bruno, C., Balzarini, M. (2010). Distancias genéticas entre perfiles moleculares obtenidos desde marcadores multilocus multialélicos. *Revista de la Facultad de Ciencias Agrarias UNCuyo* 41(3): 11.

Bruno, C., Balzarini, M., Di Rienzo, J. (2003). Comparación de Medidas de Distancia entre Perfiles RAPD individuales. *Journal of Basic & Applied Genetics* 15(2): 69-78.

Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D., Marcucci Poltri, S. N. (2013). Impacts of Population Structure and Analytical Models in Genome-Wide Association Studies of Complex Traits in Forest Trees: A Case Study in <italic>Eucalyptus globulus</italic>. *PLoS ONE* 8(11): e81267.

Cavalli-Sforza, L. L. (1966). Population structure and human evolution. *Proceedings of the Royal Society of London Series B-Biological Sciences* 164: 362-379.

Comadran, J., Thomas, W. T. B., Eeuwijk, F. Á., Ceccarelli, S., Grando, S., Stanca, A. M., Pecchioni, N., Akar, T., Al-Yassin, A., Benbelkacem, A., Ouabbou, H., Bort, J., Romagosa, I., Hackett, C. A., Russell, J. R. (2009). Patterns of genetic diversity and linkage disequilibrium in a highly structured Hordeum vulgare association-mapping population for the Mediterranean basin. *Theoretical and Applied Genetics* 119(1): 175-187.

Corder, E., Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Rimmler JB, Locke PA, Conneally PM, Schmader KE, Small GW, Roses AD, Haines JL and Pericak-vance MA. (1994). Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer's disease. *Nature Genetics* 7: 4.

Chen, L., Storey, J. D. (2006). Relaxed Significance Criteria for Linkage Analysis. *Genetics* 173(4): 2371-2381.

Chen, Z., Huang, H., Ng, H. K. (2013). Testing for association in case-control genome-wide association studies with shared controls. *Stat Methods Med Res* 1: 1.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 87(Pt 1): 52-58.

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.

Demidenko, E. (2004). *Mixed Models: Theory and Application.* New Jersey: John Wiley & Sons.

Di Rienzo, J. A. (2010).RUNNER. A simple interface to R. Argentina.

Draper, N. R., Smith, H. (1998). *Applied Regression Analysis, 3rd Edition.* New York: Wiley.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* 3(1): 1-21.

Evanno, G., Regnaut, S., Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular Ecology* 14(8): 2611-2620.

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

Falconer, D. S., Mackay, T. F. C. (1996). *Introduction to quantitative genetics.* Harlow, UK.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (1996). *Advances in knowledge and data mining.* Cambridge (Massachussets).

Feng, J.-Y., Zhang, J., Zhang, W.-J., Wang, S.-B., Han, S.-F., Zhang, Y.-M. (2013). An Efficient Hierarchical Generalized Linear Mixed Model for Mapping QTL of Ordinal Traits in Crop Cultivars. *PLoS ONE* 8(4): e59541.

Fernández, E. A., Balzarini, M. (2007). Improving cluster visualization in self-organizing maps: Application in gene expression data analysis. *Comput. Biol. Med.* 37(12): 1677-1689.

Flint-Garcia, S., Thuillet A.C., Yu J., Pressoir G., Romero S.M., Mitchell S.E., Doebley J., Kresovich S., Goodman M.M., Buckler E.S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal* 44: 10.

Francois, O., Durand, E. (2010). Spatially explicit Bayesian clustering models in population genetics. *Mol Ecol Resour* 10(5): 773-784.

Gordon, A. (1999). *Clustering.* London: Chapman & Hall/HRC Press.

Gower, J. C. (1967). Multivariate Analysis and Multidimensional Geometry. *Journal of the Royal Statistical Society. Series D (The Statistician)* 17(1): 13-28.

Guillot, G., Leblois, R., Coulon, A., A.C., F. (2009). Statistical methods in spatial genetics. *Molecular Ecology* 18: 4734-4756.

Guillot, G., Rousset, F. (2011). On the use of the simple and partial Mantel tests in presence of spatial auto-correlation. *Systematic Biology*.

Gutiérrez, L., Cuesta-Marcos, A., Castro, A. J., von Zitzewitz, 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(3): 256-272.

Han, J., Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Illinois: Morgan Kaufmann Publishers.

Han, J., Kamber, M., Pei, J. (2011). *Data Mining: Concepts and Techniques: Concepts and Techniques.*: Elsevier Science.

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.

Hardy, O., Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population level. *Mol Ecol Notes* 2: 618 - 620.

Hartigan, J. A. (1975). *Cluster Algorithms.* New York: Wiley.

Hecht-Nielsen, R. (1989).Theory of the backpropagation neural network. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, 593-605 vol.591.

Hotelling, H. (1936). Relations Between Two Sets of Variables. *Biometrika* 28: 321-377.

Jannink, J.-L., Bink, M. C., Jansen, R. C. (2001). Using complex plant pedigrees to map valuable genes. *Trends in plant science* 6(8): 337-342.

Jannink, J.-L., Iwata, H., Bhat, P. R., Chao, S., Wenzl, P., Muehlbauer, G. J. (2009). Marker imputation in barley association studies. *The Plant Genome* 2(1): 11-22.

Jobson, J. D. (1992). *Applied Multivariate Data Analysis: Categorical and multivariate methods.*: Springer-Verlag.

Johnson, R., Wichern, D. (Eds) (1998). *Applied Multivariate Statistical Analysis.* New Jersey: Prentice Hall.

Jombart, T., Eggo, R. M., Dodd, P., Balloux, F. (2009a). Spatiotemporal dynamics in the early stages of the 2009 A/H1N1 influenza pandemic. *PLoS Curr* 1: RRN1026.

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.

Kohonen, T. (1997). Self-organizing maps. 117.

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.

Kraakman, A. T. W., Niks, R. E., Van den Berg, P. M. M. M., Stam, P., Van Eeuwijk, F. A. (2004). Linkage Disequilibrium Mapping of Yield and Yield Stability in Modern Spring Barley Cultivars. *Genetics* 168(1): 435-446.

Lawson, D. J., Falush, D. (2012). Population Identification Using Genetic Data. *Annual Review of Genomics and Human Genetics* 13(1): 337-361.

Lee, C., Abdool, A., Huang, C.-H. (2009). PCA-based population structure inference with generic clustering algorithms. *BMC Bioinformatics* 10(Suppl 1): S73.

Li, H., Wei, Z., Maris, J. (2010). A hidden Markov random field model for genome-wide association studies. *Biostatistics* 11(1): 139-150.

Li, J., Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 95(3): 221-227.

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. *Molecular Breeding* 31(3): 631-654.

Lynch, M., Walsh, B. (Eds) (1998). *Genetics and Analysis of Quantitative Traits.* Massachusetts.

MacQueen, J. (1967). Some Methods for classification and Analysis of Multivariate Observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability* 1: 17.

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

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, S., van Eeuwijk, F. (2004). Mixed models including environmental covariables for studying QTL by environment interaction. *Euphytica* 137(1): 139-145.

Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22(4): 719-748.

Mayr, E., Linsley, E. G., Usinger, R. L. (1953). *Methods and principles of systematic zoology.* New York: McGraw-Hill.

McVean, G. (2009). A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet* 5(10): e1000686.

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.

Milligan, G., Cooper, M. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrik* 50(2): 20.

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., Muehlbauer, G. J. (2014). The USDA Barley Core Collection: Genetic Diversity, Population Structure, and Potential for Genome-Wide Association Studies. *PLoS ONE* 9(4): e94688.

Nikolic, N., Park, Y. S., Sancristobal, M., Lek, S., Chevalet, C. (2009). What do artificial neural networks tell us about the genetic structure of populations? The example of European pig populations. *Genet Res (Camb)* 91(02): 121-132.

Nyholt, D. R. (2004). A Simple Correction for Multiple Testing for Single-Nucleotide Polymorphisms in Linkage Disequilibrium with Each Other. *American journal of human genetics* 74(4): 765-769.

Odong, T., van Heerwaarden, J., Jansen, J., van Hintum, T., van Eeuwijk, F. (2011). Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? *TAG Theoretical and Applied Genetics* 123(2): 195-205.

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. J., Johal, G. (2014). A Genome-Wide Association Study of the Maize Hypersensitive Defense Response Identifies Genes That Cluster in Related Pathways. *PLoS Genet* 10(8): e1004562.

Paini, D. R., Worner, S. P., Cook, D. C., De Barro, P. J., Thomas, M. B. (2010). Using a self-organizing map to predict invasive species: sensitivity to data errors and a comparison with expert opinion. *Journal of Applied Ecology* 47(2): 290-298.

Park, Y.-S., Céréghino, R., Compin, A., Lek, S. (2003). Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters. *Ecological Modelling* 160(3): 265-280.

Patterson, H. D., Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* 58(3): 545-554.

Patterson, N., Price, A. L., Reich, D. (2006). Population Structure and Eigenanalysis. *PLoS Genet* 2(12): e190.

Peña-Malavera, A., Bruno, C., Fernandez, E., Balzarini, M. (2014a).Comparison of algorithms to infer genetic population structure from unlinked molecular markers. In *Statistical Applications in Genetics and Molecular Biology*, Vol. 13, 391.

Peña Malavera, A., Gutierrez, L., Balzarini, M. (2014b). Componentes principales en mapeo asociativo. *BAG. Journal of basic and applied genetics* 25: 32-40.

Piepho, H. P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* 157(1): 425-432.

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

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

Roux, O., Gevrey, M., Arvanitakis, L., Gers, C., Bordat, D., Legal, L. (2007). ISSR-PCR: Tool for discrimination and genetic structure analysis of Plutella xylostella populations native to different geographical areas. *Molecular Phylogenetics and Evolution* 43(1): 240-250.

Roy, J., Smith, K., Muehlbauer, G., Chao, S., Close, T., Steffenson, B. (2010). Association mapping of spot blotch resistance in wild barley. *Molecular Breeding* 26(2): 243-256.

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., Taylor, J. (2008). False discovery rate analysis of brain diffusion direction maps. *Ann Stat* 2: 153 - 175.

Searle, S. R., Casella, G., McCulloch, C. E. (2008).Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML). In *Variance Components*, 232-257: John Wiley & Sons, Inc.

Searle, S. R., Casella, G., McCullogh, C. E. (1992). *Variance components.* New York: Wiley.

Segelbacher, G., Cushman, S. A., Epperson, B. K., Fortin, M. J., Francois, O., Hardy, O. J., Holderegger, R., Taberlet, P., Waits, L. P., Manel, S. (2010). Applications of landscape genetics in conservation biology: concepts and challenges. *Conservation Genetics* 11(2): 375-385.

Shriner, D., Vaughan, L., Padilla, M., Tiwari, H. (2007). Problems with Genome-Wide Association Studies. *Science* 316: 1840-1842.

Sidak, Z. (1967). Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. *Journal of the American Statistical Association* 62(318): 626-633.

Smouse, P., Spielman, R., Park, M. (1982). Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. *American Naturalist* 119: 445-463.

Sokal, R. R., Michener, C. D. (1958). A Statistical Methods for Evaluating Systematic Relationships. *University of Kansas Science Bulletin* 38: 1409-1438.

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., Möhring, J., Piepho, H.-P., Heckenberger, M., Buckler, E. S., Melchinger, A. E. (2008). Comparison of Mixed-Model Approaches for Association Mapping. *Genetics* 178(3): 1745-1754.

Sun, W., Cai, T. (2009). Large-scale multiple testing under dependence. *J R Stat Soc Ser B* 71: 393 - 424.

Team, R. D. C. (2013). R: A language and environment for statistical computing. (Ed R. F. f. S. Computing). Vienna, Austria.

Teich, I., Planchuelo, A., Balzarini, M. (2011b).Análisis de asociaciones en escenarios de datos masivos. . In *Anales del Congreso de AgroInformática*, 166-177.

Teich, I. (2012).Análisis de la estructura genética espacial de especies arbóreas y su asociación con la variabilidad fenotípica y ambiental. In *Facultad de Ciencias Exactas y Naturales.*, Vol. Doctorado Buenos Aires: Universidad de Buenos Aires.

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

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2): 411-423.

Toronen, P., Kolehmainen, M., Wong, G., Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters* 451: 142 - 146.

Tracy, C. A., Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Comm. Math. Phys.* 159(1): 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(9): 5116-5121.

Ultsch, A. (Ed) (1999). *Data mining and knowledge discovery with emergent selforganizing feature maps for multivariate time series.* . E. Oja, S. Kaski.

Vekemans, X., Hardy, O. J. (2004). New insights from fine-scale spatial genetic structure analyses in plant populations. *Molecular Ecology* 13(4): 921-935.

von Zitzewitz, J., Cuesta-Marcos, A., Condon, F., Castro, A. J., Chao, S., Corey, A., Filichkin, T., Fisk, S. P., Gutierrez, L., Haggard, K., Karsai, I., Muehlbauer, G. J., Smith, K. P., Veisz, O., Hayes*, P. M. (2011). The Genetics of Winterhardiness in Barley: Perspectives from Genome-Wide Association Mapping. *Plant Gen.* 4(1): 76-91.

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. *Theoretical and Applied Genetics* 124: 111-124.

Wang, J., Delabie, J., Aasheim, H., Smeland, E., Myklebost, O. (2002). Clustering of the SOM easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics* 3(1): 36.

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., Zhang, Z. (2014). A SUPER Powerful Method for Genome Wide Association Study. *PLoS ONE* 9(9): e107684.

Wang, W. Y., Barratt, B. J., Clayton, D. G., Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* 6(2): 109-118.

Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. . *Journal of the American Statistical Association.* 58: 8.

Wei, Z., Sun, W., Wang, K., Hakonarson, H. (2009). Multiple testing in genome-wide association studies via hidden Markov models. *Bioinformatics* 25(21): 2802 - 2808.

Weir, B. S., Ott, J. (1997). Genetic data analysis II. *Trends in Genetics* 13(9): 379.

West, B., Welch, K., Galecki, A. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software.* Boca Raton, FL.

Worner, S. P., Gevrey, M. (2006). Modelling global insect pest species assemblages to determine risk of invasion *Journal of Applied Ecology* 43(5): 858-867.

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

Wu, R., Casella, G., Ma, C.-X. (2007). *Statistical Genetics of Quantitative Traits. Linkage, Maps, and QTL.* New York: Springer

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(1): 282.

Yu, J., Buckler, E. (2006). Genetic association mapping and genome organization of maize. *Curr Opin Biotech* 17: 155 - 160.

Yu, J., Pressoir, G., Briggs, W., Bi, I., Yamasaki, M., Doebley, J., McMullen, M., Gaut, B., Nielsen, D., Holland, J., Kresovich, S., Buckler, E. (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. J., Yu, J., Arnett, D. K., Ordovas, J. M., Buckler, E. S. (2010). Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42(4): 355-360.

Zhao, K., Aranzana, M., Kim, S., Lister, C., Shindo, C., Tang, C., Toomajin, C., Zheng, H., Dean, C., Marjoram, P., Nordborg, M. (2007). An Arabidopsis example of association mapping in structured samples. *PLoS Genet* 3(1): e4.

Zhu, C., Gore, M., Buckler, E., Yu, J. (2008). Status and Prospects of Association Mapping in Plants. *The Plant Genome* 1(1): 16.

ANEXO I

Códigos de QMSim para simulación de datos genéticos.

Se presenta el codigo usado para simular los datos genéticos dadas 10 generaciones, 10 cromosomas y 30 marcadores por cromosoma para un genoma de 300 marcadores moleculares.

```
/*******************************
 **     Global parameters     **
 *******************************/
title = "LD";
nrep  = 1;                  //Number of replicates
h2    = 0.2;                //Heritability
qtlh2 = 0.1;                //QTL heritability
phvar = 1.0;                //Phenotypic variance
/*******************************
 **   Historical population   **
 *******************************/
begin_hp;
  hg_size = 500 [0]         //Size of the historical generations
            200 [1];
nmlhg   = 50;               //Number of males in the last generation
end_hp;
/*******************************
 **       Populations       **
 *******************************/
begin_pop = "Line 1";
  begin_founder;
    male   [n =  50, pop = "hp"];
    female [n = 150, pop = "hp"];
  end_founder;
  ls  = 1 2 [0.5] 3 [0.1];   //Litter size
  pmp = 0.5 /fix;            //Proportion of male progeny
  ng  = 10;                  //Number of generations
  md  = rnd;                 //Mating design
  sr  = 1;                  //Replacement ratio for sires
```

```
  dr  = 1;              //Replacement ratio for dams
  sd  = rnd;            //Selection design
  begin_popoutput;
     ld /bin 10 /maft 0.05 /gen 10;
     data /gen 10;
     genotype /snp_code /gen 10;
     allele_freq /gen 10;
  end_popoutput;
end_pop;
begin_pop = "Line 2";
  begin_founder;
    male   [n =  50, pop = "hp"];
    female [n = 150, pop = "hp"];
  end_founder;
  ls  = 1 2 [0.5] 3 [0.1];   //Litter size
  pmp = 0.5 /fix;           //Proportion of male progeny
  ng  = 10;                 //Number of generations
  md  = rnd;                //Mating design
  sr  = 1;                  //Replacement ratio for sires
  dr  = 1;                  //Replacement ratio for dams
  sd  = rnd;            //Selection design
  begin_popoutput;
     ld /bin 10 /maft 0.05 /gen 10;
     data /gen 10;
     genotype /snp_code /gen 10;
     allele_freq /gen 10;
  end_popoutput;
end_pop;
begin_pop = "Line 3";
  begin_founder;
    male   [n =  50, pop = "hp"];
    female [n = 150, pop = "hp"];
```

```
  end_founder;
  ls  = 1 2 [0.5] 3 [0.1];   //Litter size
  pmp = 0.5 /fix;           //Proportion of male progeny
  ng  = 10;                 //Number of generations
  md  = rnd;                //Mating design
  sr  = 1;                  //Replacement ratio for sires
  dr  = 1;                  //Replacement ratio for dams
  sd  = rnd;                //Selection design
  begin_popoutput;
     ld /bin 10 /maft 0.05 /gen 10;
     data /gen 10;
     genotype /snp_code /gen 10;
     allele_freq /gen 10;
  end_popoutput;
end_pop;
/*******************************
 **         Genome          **
 *******************************/
begin_genome;
  begin_chr = 10;
    chrlen = 100;          //Chromosome length
    nmloci = 30;           //Number of markers
    mpos   = rnd;          //Marker positions
    nma    = all 2;        //Number of marker alleles
    maf    = eql;          //Marker allele frequencies
    nqloci = 3;            //Number of QTL
    qpos   = rnd;          //QTL positions
    nqa    = all 2;     //Number of QTL alleles
    qaf    = eql;          //QTL allele frequencies
    qae    = rndg 2;      //QTL allele effects
    cld   =mq;             //LD
  end_chr;
```

```
  mmutr     = 2.5e-5 /recurrent; //Marker mutation rate
  qmutr     = 2.5e-5;           //QTL mutation rate
  interference = 25;
  r_mpos_g;                 //Randomize marker positions across genome
  r_qpos_g;                 //Randomize QTL positions across genome
end_genome;
/*******************************
 **       Output options      **
 *******************************/
begin_output;
     linkage_map;
     allele_effect;
end_output;
```

Anexo II

Códigos R. Análisis de mapeo asociativo

Se presentan los códigos de R para estudiar la asociación entre los datos de marcadores moleculares y los caracteres fenotípicos mediante modelos lineales mixtos (MLM). Se plantea el ajuste de ocho modelos y la presentación gráfica de los resultados para el modelo *Naive*.

#Carga el directorio de trabajo

```
setwd("nombre del directorio")
getwd()
```

#Carga en el espacio de trabajo los archivos de genotipos, fenotipos y mapa

```
qtl.data<-
load.data(P.file="QAssociation_pheno.txt",G.file="QAssociation_geno.txt",
map.file="QAssociation_map.txt", cross="am", heterozygotes="FALSE")
```

#Devuelve la estadística descriptiva de los archivos cargados previamente

```
summary(qtl.data)
```

#Realiza en análisis de componentes principales y selecciona la cantidad de ejes significativas según Tracy-Widom

```
pca<-pca.analysis(file=qtl.data, p.val=0.05)
```

#Realiza el gráfico de desequilibrio de ligamiento

```
LD.plots(file=qtl.data, structure="FALSE", heterocigotes="TRUE")
```

#Declaración de covariables para los modelos

```
covar5<-
(read.table(file="QAssociation_phenoK_5.txt",header=TRUE))[,(2:6)]
cova5<-as.matrix(covar5)
```

#Modelo con matriz de STRUCTURE para K=5 (Modelo Q)

```
(str5.am<-am(file=qtl.data, method="fixed", provide.K=FALSE,
covariates=cova5, trait="yield",threshold=0.05,p=0.05,
out.file="str5"))$selected
```

#Modelo con matriz de STRUCTURE para K=5 aleatorio (Modelo QA)

```
(str5R.am<-am(file=qtl.data,
method="mixed.random",provide.K=FALSE,covariates= cova5,
trait="yield",threshold=0.05,p=0.05,out.file="str5R"))$selected
```

#Modelo con covariables aleatorias provenientes del análisis de ACP (Modelo PA)

```
(pcaR.am<-am(file=qtl.data, method="mixed.random",
provide.K=FALSE,covariates= pca$scores, trait="yield", threshold=0.05,
p=0.05, out.file="AM fixed PCAmodel"))$selected
```

#Modelo con covariables fijas provenientes del análisis de ACP (Modelo P)

```
(pca.am<-am(file=qtl.data, method="fixed",
provide.K=FALSE,covariates=covariate,trait="yield",threshold=0.05,p=0.05,
out.file="AM fixed PCAmodel"))$selected
```

#Modelo con matriz de parentesco (Modelo K)

```
(k.am<-
am(file=qtl.data,method="mixed.nostructure",provide.K=FALSE,covariates
=FALSE, trait="yield", threshold=0.05, p=0.05, out.file="AM mixed model
nocovariate"))$selected
```

#Modelo con matriz de parentesco y corrección por estructura mediante CPs (Modelo PK)

```
(k_pca.am<-am(file=qtl.data, method="mixed",
provide.K=FALSE,covariates=pca$scores, trait="pheno", threshold=0.05,
p=0.05, out.file="AM mixed model nocovariate"))$selected
```

#Modelo con matriz de parentesco y corrección por estructura mediante Q (Modelo QK)

```
(k_str5.am<-am(file=qtl.data, method="mixed",
provide.K=FALSE,covariates=cova5, trait="pheno", threshold=0.05, p=0.05,
out.file="AM mixed model nocovariate"))$selected
```

#Modelo sin corrección por estructura ni parentesco (Modelo *Naive*)

```
(naive.am<-am(file=qtl.data, method="naive", provide.K=FALSE, covariates=
FALSE,trait="yield", threshold=0.05, p=0.05, out.file="AM naïve
model"))$selected
```

#Gráfico de resultados se presenta sólo para el modelo naive

```
p.file<-naive.am$p.val
pdf(file = "5_AM naive.pdf" , onefile = TRUE)
print(xyplot(-log10(p.file[,3])~ p.file[,2] | factor(p.file[,1]),
type="h",layout=c(length (unique(p.file[,1])),1), col="red",
xlab="Posición del cromosoma", ylab="-log10(P)", main="Mapeo sin
corrección", scales =list(x = "free"), ylim=c(0,(max(-
log10(p.file[,3]))+0.5))))
dev.off()
write.table(p.file, file=" Mapeo sin
corrección.txt",append=TRUE,row.names= FALSE,col.names=FALSE,quote =
FALSE)
dim(naive.am$selected)
```