Universidad Nacional de Córdoba

Facultad de Ciencias Agropecuarias

Escuela para Graduados

HERRAMIENTAS ESTADÍSTICAS PARA EL ANÁLISIS DE DATOS EN ESTUDIOS DE ASOCIACIÓN GENÓMICA

Lic. María Angélica Rueda Calderón

Tesis

Para optar al Grado Académico de

Doctor en Ciencias Agropecuarias

Universidad Nacional de Córdoba

Córdoba, 2020

HERRAMIENTAS ESTADÍSTICAS PARA EL ANÁLISIS DE DATOS EN ESTUDIOS DE ASOCIACIÓN GENÓMICA

María Angélica Rueda Calderón

Comisión Asesora de Tesis

Director: Ing. Agr. (Dra.) Cecilia Bruno

Asesores: Ing. Agr. (Dra.) Mónica Balzarini

Bioing. (Dr.) Elmer Andrés Fernández

Tribunal Examinador de Tesis

Ing. Agr. (Dra.) Mónica Balzarini

Ing. Agr. (Dra.) María Gabriela Molina

Bioquím. (Dra.) Norma Beatriz Paniego

Presentación formal académica

Junio 2020

Facultad de Ciencias Agropecuarias

Universidad Nacional de Córdoba

AGRADECIMIENTOS

A Dios que me lo ha dado todo,
por su infinita misericordia y amor.

Quisiera agradecer a mi mamá, papá y hermanos por ser mis grandes soportes en el transcurrir de la vida. Por su entrega incondicional de amor, sacrificando momentos que nunca volverán.

A la República de la Argentina por ser un país de grandes bendiciones y brindarme las oportunidades que siempre anhele.

Quisiera agradecer a mi directora de tesis, la Dra. Cecilia Bruno por apoyarme y estar siempre dispuesta a ayudarme cuando más lo he necesitado; por su constante paciencia y porque en varios momentos me ha animado a continuar cuando yo pensaba que no podría lograrlo. También por su tiempo y generosidad al brindarme la oportunidad de recurrir a su capacidad como profesional y en lo personal.

Quisiera agradecer a mi directora de beca la Dra. Mónica Balzarini, por permitirme ser parte de su equipo de trabajo y por enseñarme con amor y dedicación. Por tener la capacidad de ayudar a tantas personas en las que me incluyo y por mostrarme que siempre hay solución para todo. Además, por compartir conmigo muchos momentos y experiencias que nunca se me olvidarán.

A Julio Di Rienzo, por recibirme con cariño en su espacio de trabajo y por estar siempre dispuesto a ayudarme cuando lo he necesitado.

A las profesoras Margot Tablada y Laura González de la Cátedra de Estadística y Biometría por estar siempre a disposición para atender mis consultas.

A la Facultad de Ciencias Agropecuarias de la Universidad Nacional de Córdoba por brindarme un espacio de trabajo.

Al Consejo Nacional de Investigación Científica y Tecnológica (CONICET) por permitir llevar a cabo este trabajo de investigación a través del otorgamiento de la beca interna doctoral.

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 examinador y por dedicar su tiempo a la revisión de este trabajo.

A mi esposo Andrés por su amor y por hacer de mí, una mejor versión de mí misma. Por ser mi amigo y pareja idónea en esta aventura.

A mi hijo Adrián Ricardo por ser esa hermosa bendición que había anhelado desde hace tiempo.

A Soledad Martínez y Ricardo Roldán por su gran cariño, siendo para nosotros como nuestros padres en Argentina.

A los amigos que conocí en este hermoso país y que me han brindado siempre lo mejor de ellos, su amistad.

A mis compañeros, Mati, Fernando, Diego, Franca, Pablo, Miguel, Estefi, Euge, Mariano, Moni Pic y Franco, por compartir lindos momentos.

A todos MUCHAS GRACIAS.

DEDICATORIA

A Dios por sus promesas y
el cumplimento de cada una de ellas.

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 caracter fenotípico. En el mejoramiento vegetal posibilita el uso de poblaciones no diseñadas experimentalmente. Los marcadores significativos pueden ser usados en selección asistida por marcadores. Se supone que la modelación estadística que incorpora información sobre el parentesco y correlaciones ambientales hace más eficiente el MA. Un objetivo de esta tesis es comparar el desempeño, en términos de la estimación de las componentes de varianza, las puntuaciones de los marcadores y los BLUP (*best linear unbiased predictor*) de los efectos de genotipo, de los modelos por ambiente y multiambientales para GWAS que incluyen correlaciones genéticas a través de un pedigrí o una matriz de similitud de marcadores moleculares. Los modelos multiambientales produjeron estimaciones más precisas de la variabilidad genética y ofrecen una estimación de la varianza de la interacción genotipo-ambiente (G×E), razón por la cual son preferidos a la modelación por ambiente. Dada la abundancia de estudios de selección genómica (SG) en vegetales, otros objetivos de la tesis fueron, i) realizar una revisión sistemática de literatura científica publicada sobre SG en vegetales, ii) identificar las principales metodologías de estimación usadas en el contexto de SG, iii) aplicar técnicas propias del meta-análisis para obtener medidas globales de la eficiencia de los modelos SG usados con mayor frecuencia en vegetales y iv) desarrollar un protocolo para meta-análisis orientado a identificar *loci* de efecto mayor en estudios de asociación genómica. Los resultados del meta-análisis sugirieron que la eficiencia de la SG en promedio es del 60% tanto en cereales como en otras especies de importancia agrícola. Respecto al análisis de QTL tradicional donde existen numerosas publicaciones el protocolo propuesto para la identificación de *loci* de efecto mayor representa una herramienta potente para la síntesis de resultados. Implementado en la búsqueda de QTL para resistencia/tolerancia a enfermedades virales en maíz identificó consensos en las publicaciones que sugieren QTL posicionados en el cromosoma 1 con efecto aditivo relativamente importante, mientras que aquellos presentes en los cromosomas 3, 4 y 10 se corresponden con *loci* de efecto relativamente menor.

Palabras clave: datos correlacionados, modelos lineales mixtos, interacción genotipo-ambiente, selección genómica, meta-análisis.

ABSTRACT

Association mapping (AM) or GWAS (Genome Wide Association Study), is used to find specific genomic regions related to phenotypic trait variation. In plant breeding it enables the use of populations not experimentally designed. Significant markers can be used in marker-assisted selection. It is assumed that statistical modeling that incorporates information about kinship and environmental correlations makes the AM more efficient. An objective of this thesis is to compare the performance, in terms of the estimation of the variance components, marker scores, and BLUP (Best Linear Unbiased Predictor) of the genotype effects, in single-environment and multienvironmental models for GWAS which including genetic correlations through a pedigree or a similarity matrix of molecular markers. Multienvironmental models produced more accurate estimates of genetic variability and offered an estimate of the variance of the genotype-environment interaction (G×E), which is why they are preferred to modeling by environment. Given the abundance of Genomic Selection studies (GS) in vegetables, other objectives of the thesis were, i) to conduct a systematic review of published scientific literature on GS in vegetables, ii) to identify the main estimation methodologies used in the context of GS, iii) to apply meta-analysis techniques to obtain global measures of the efficiency of the most frequently used GS models in vegetables, and iv) to develop a protocol for meta-analysis aimed at identifying major-effect loci in genomic association studies. Meta-analysis results suggested that on average of GS efficiency is 60% in both cereals and other species of agricultural importance. Regarding traditional QTL analysis where there are numerous publications, the proposed protocol for the identification of the major-effect loci represents a powerful tool for the synthesis of results. Implemented in the search for resistance/tolerance's QTL to viral diseases in maize, it identified consensus in the publications that suggest QTL positioned on chromosome 1 with a relatively important additive effect, while those present on chromosomes 3, 4, and 10 correspond to loci of relatively minor effect.

Key words: Correlated data, linear mixed models, genotype-environment interaction, genomic selection, meta-analysis.

TABLA DE CONTENIDOS

LISTA DE TABLAS

LISTA DE FIGURAS

LISTA DE ABREVIATURAS

ADN: Ácido DesoxirriboNucleico (DNA, *DeoxyriboNucleic Acid*)

AIC: Criterio de información de Akaike (*Akaike Information Criteria*)

ANOVA: Análisis de la varianza (*ANalysis Of VAriance*)

BIC: Criterio bayesiano de Schwarz (*Bayesian Information Criteria*)

BLUP: *Best Linear Unbiased Predictor*

DArT: *Diversity Array Technology*

E: Ambiente (*Environment*)

G: Genotipo (*Genotype*)

G×E: Interacción Genotipo-Ambiente (*Genotype-Environment interaction*)

G-BLUP*: Genomic-Best Linear Unbiased Predictor*

GEBV: Mérito Genético (*Genome-Estimated Breeding Values*)

GWAS: Estudio de asociación del genoma completo (*Genome-Wide Association Study*)

K: Matriz de relaciones genéticas obtenidas a partir del pedigrí

LASSO: *Least Absolute Shrinkage and Selection Operator*

LD: Desequilibrio de Ligamiento (*Linkage Disequilibrium*)

MA: Mapeo Asociativo (*Association Mapping*)

MAS: Selección Asistida por Marcadores (*Marker-Assited Selection*)

METs: Ensayos Multiambientales (*Multi-Environmental Trials*)

MLM: Modelos Lineales Mixtos (*Linear Mixed Models*)

MM: Marcadores Moleculares (*Molecular Markers*)

NGS: Secuenciación de próxima generación (*Next-Generation Sequencing*)

P: Matriz de componentes principales

Q: Matriz de probabilidad de pertenencia de los genotipos a los subgrupos que estructuran la población.

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

REML: Máxima verosimilitud restringida (*Restricted Maximum Likelihood*)

RKHS: *Reproducing Kernel Hilbert Space*

RR-BLUP: *Ridge Regression-Best Linear Unbiased Prediction*

SG: Selección Genómica (*Genomic Selection*)

SVM: Máquina de Soporte Vectorial (*Support Vector Machine*)

SNP: *Single-Nucleotide Polymorphisms*

TRN: Población de entrenamiento o de referencia (*Training Population*)

TST: Población de validación o prueba (*Testing Population*)

CAPÍTULO 1

INTRODUCCIÓN GENERAL

En las últimas décadas, la genética cuantitativa ha asumido un rol preponderante en ensayos de mejoramiento genéticos de cultivos debido a la incorporación y disponibilidad de información proveniente de marcadores moleculares (MM) basados en ADN, y a una amplia variedad de metodologías estadísticas que evolucionan constantemente para el análisis de la información generada por las nuevas biotecnologías (Gupta *et al*., 2014). Las regiones genómicas que contienen genes que controlan un caracter dado, son conocidos como *loci* de caracteres cuantitativos (QTL, del inglés *Quantitative Trait Loci*) (Yang *et al*., 2013). La relación genotipo-fenotipo ha sido estudiada mediante los enfoques de mapeo de ligamiento tales como, el mapeo de QTL a partir de poblaciones de mapeo basadas en cruzamientos biparentales y el mapeo asociativo (MA) o GWAS (por sus siglas en inglés, *Genome Wide Association Study*) de paneles diversos (Del Carpio *et al*., 2018). El mapeo de QTL es uno de los enfoques más comunes para el estudio genético de caracteres cuantitativos (Yang *et al*., 2013). Este método fue elegido para la identificación de *loci* responsables en los cambios fenotípicos asociados a la evolución del cultivo (Burke *et al*., 2002; Doebley *et al*., 1990). La precisión en el mapeo de QTL depende en gran medida de la variación genética (o antecedentes genéticos) cubierto por: la población de mapeo, el tamaño de la población de mapeo y el número de *loci* usados (Sehgal *et al*., 2016). El principal objetivo del mapeo de QTL es determinar los *loci* que influyen en la variación de los caracteres complejos y cuantitativos. Otro objetivo del análisis de QTL ha sido dar respuesta a la siguiente pregunta, ¿si las diferencias fenotípicas son debidas principalmente a pocos *loci* con efectos mayores o muchos *loci* con efectos menores? (Miles y Wayne, 2008). Una proporción considerable de la variación fenotípica de muchos caracteres cuantitativos puede ser explicada con pocos *loci* de efecto mayor (Remington y Purugganan, 2003; Roff, 2007). Debido a que el mapeo de QTL tradicional requiere genotipar un gran número de individuos provenientes de un cruzamiento biparental y que la generación de esa población de mapeo demanda mucho tiempo y altos costos, hace que esta técnica sea cada vez más relegada (Borevitz y Chory,

2004; Yang *et al*., 2013). Algunas limitantes para usar dicha técnica son: i) la variación alélica en cada cruzamiento está restringida, debido a que es necesario dos padres para crear una población de mapeo de QTL y, ii) dado que se utilizan cruzamientos en la primera generación, el número de eventos de recombinación por cromosoma suele ser pequeño (Sehgal *et al*., 2016).

Los estudios de asociación se pueden dividir en dos categorías: i) mapeo de asociación de genes candidatos, en el que la variación de un gen se prueba a través de la correlación con el caracter fenotípico de interés y ii) mapeo por desequilibrio de ligamiento (LD) (por sus siglas en inglés, *Linkage Disequilibrium*), basado en asociaciones no aleatorias de alelos en diferentes *loci* en el mismo cromosoma (Flint-Garcia *et al*., 2003). El LD es un concepto importante en genética de poblaciones debido a que resume la variación genética que ocurrió dentro de una población a través de su historia evolutiva (Del Carpio *et al*., 2018). El MA permite la identificación de marcadores del tipo SNP (*Single-Nucleotide Polymorphisms*) (Breen *et al*., 2000) que están estrechamente relacionados a un caracter fenotípico, basándose en el principio de LD entre marcadores genéticos y los *loci* que afectan al caracter de interés (Geng *et al*., 2015). Las asociaciones detectadas en el MA son a menudo espurias debido a que dichas asociaciones se basan en el LD, que depende no solamente del ligamiento, sino que también de la estratificación poblacional y la relación entre individuos (Gupta *et al*., 2014). Generalmente, los *loci* que están físicamente juntos presentan un LD más fuerte que los *loci* que están más separados en un cromosoma (Visscher *et al*., 2012). Sehgal *et al*. (2016) presentaron un conjunto de pautas para llevar a cabo un MA: 1) seleccionar un grupo de individuos de una población natural o colección de germoplasma con una amplia cobertura de diversidad genética, 2) medir las características fenotípicas de la población, preferiblemente, en diferentes ambientes y con varias repeticiones; 3) genotipar los individuos de la población de mapeo con MM; 4) cuantificar el LD del genoma de una población seleccionada usando MM; 5) evaluar la estructura poblacional (el nivel de diferenciación genética entre grupos dentro de una población de individuos muestreada) y el parentesco (coeficiente de relación entre pares de individuos dentro de una muestra); y 6) identificar aquellos MM que están posicionados cerca al caracter de interés basado en la información obtenida a través de la cuantificación del LD, la estructura genética poblacional y la correlación entre los datos fenotípicos y genotípicos estimada mediante un modelo estadístico apropiado. El objetivo del MA es lograr identificar el desequilibrio gamético

entre los alelos de dos *loci* (Jannink y Walsh, 2002). Además, el MA permite evaluar la significancia estadística de las asociaciones entre la información genotípica que proveen los MM y el caracter fenotípico de interés. En contraste con el enfoque de mapeo de QTL tradicional, el GWAS usa el desequilibrio de ligamiento entre los marcadores y los caracteres de interés y, por lo tanto, la resolución de mapeo mejora al capturar muchas generaciones de recombinación histórica (Hazzouri *et al*., 2014). No obstante, el MA no debe considerarse como el remplazo del mapeo de QTL tradicional. De hecho, estas dos técnicas poseen ventajas y desventajas complementarias, que pueden conducir a una mejor comprensión del polimorfismo genético causal cuando estos enfoques se aplican de manera conjunta (Chan *et al*., 2010; Mitchell-Olds, 2010).

Aun pensando en un conjunto de marcadores independientes es importante considerar que en todo estudio de asociación la significancia estadística de la correlación entre el estado de los marcadores y el caracter en estudio pueden no ser el resultado de un ligamiento real entre marcadores y *loci* de interés, es decir, pueden generarse falsos positivos. Dicho de otra manera, los falsos positivos pueden producirse por la correlación entre individuos como consecuencia de su relación filial o parentesco. La relación de parentesco puede ser considerada a través de matrices de pedigrí (Piepho *et al*., 2008; Burgueño *et al*., 2012) o matrices de similitud molecular (Ritland, 1996; Lynch y Ritland, 1999; VanRaden, 2007; Burgueño *et al*., 2012). Éste tipo de información debe considerarse en el ajuste de los modelos estadísticos de asociación, ya que introducir información de la estructura genética poblacional disminuye la tasa de falsos positivos (Malosetti *et al*., 2007; Peña Malavera, 2015).

En la actualidad, con el desarrollo de biotecnologías de alto rendimiento a nivel de información genómica. Los GWAS se han convertido en un método eficiente, en el análisis de caracteres cuantitativos controlados por múltiples genes con efectos pequeños (Brachi *et al*., 2011). 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 caracter *vs*. marcadores ajustados en el marco teórico de los modelos lineales mixtos (MLM) (Demidenko, 2004; Malosetti *et al*., 2007; Peña Malavera, 2015) y la estimación por máxima verosimilitud restringida (REML-*Restricted Estimation Maximun Likelihood*) (Patterson y Thompson, 1971). En el modelo de MA se

especifica la relación entre cada marcador y el fenotipo con la finalidad de detectar los marcadores significativos, *i.e*., los potencialmente ligados a un QTL.

Los ensayos multiambientales METs (por sus siglas en inglés, *multi-environment trials*) tienen un rol importante en el contexto del mejoramiento vegetal, ya que permiten evaluar el desempeño de los genotipos a través de diferentes condiciones ambientales, estudiar tanto la interacción genotipo-ambiente (G×E) como la estabilidad de los genotipos y predecir el desempeño de genotipos no evaluados (Burgueño *et al*., 2012; Borgognone *et al*., 2016; Aguate *et al*., 2019). La densa información de MM puede ser usada para estimar la similaridad genética entre individuos (*e.g*., VanRaden, 2007; de los Campos *et al*., 2010). Este tipo de información puede ser incorporada en los modelos multiambientales, de la misma manera que las relaciones genéticas derivadas del pedigrí (Burgueño *et al*., 2012).

La evaluación de fenotipos en múltiples ambientes, dentro de un contexto de gran cantidad de información molecular, puede ser realizada con modelos multiambientales, que incorporan las correlaciones genéticas a partir de la información dada por los MM. Los modelos multiambientales, permiten contemplar otro tipo de correlación en los datos (correlaciones ambientales), que podrían generar estructuras y sugerir asociaciones espurias (incrementar falsos positivos). Cuando los efectos de genotipo (G) se evalúan en METs, los efectos del ambiente (E) y de la interacción entre efectos de genotipo y ambiente (G×E) deben ser estimados separadamente y descontados de la varianza residual para mejorar la precisión. Esto con el fin de evaluar la significancia de los MM y el ordenamiento de los genotipos según su mérito genético. Cuando la predicción del efecto de G es afectada por correlaciones genéticas dentro de la población (Rabier *et al*., 2016), las correlaciones de los perfiles moleculares deben ser contempladas en los modelos. Para el caso de MA a partir de información fenotípica proveniente de un único ambiente, las correlaciones genéticas se han incorporado a través de las matrices, Q (matriz de probabilidad de pertenencia de los genotipos a los subgrupos que estructuran la población) (Pritchard *et al*., 2000; Parisseaux y Bernardo, 2004), P (matriz de componentes principales que identifican subgrupos de genotipos a partir de MM) (Price *et al*., 2006) y/o K (matriz de relaciones genéticas obtenidas a partir del pedigrí) (Kang *et al*., 2008). Si bien existe un desarrollo metodológico en este terreno en lo que respecta al análisis de datos fenotípicos provenientes de un único ambiente (Gutiérrez *et al*., 2011; Cappa *et al*., 2013), la estimación de modelos con interacciones para el caso de METs ha recibido menos atención para el MA. Se han publicado algunos enfoques

sobre cómo podría ser estimada dicha componente de varianza de interacción en el caso de METs donde existe información molecular sobre los genotipos evaluados (Burgueño *et al*., 2012; Malosetti *et al*., 2013). El marco teórico de los MLM es usado también en este contexto, aún con datos fenotípicos desbalanceados (Lado *et al*., 2016). Los MLM permiten estimar la contribución de las principales fuentes de variación y las relaciones entre los distintos tipos de variables que intervienen en el MA. Se supone que un MLM que considere tanto las correlaciones genéticas, extraídas desde los datos moleculares, así como las correlaciones generadas por el diseño de evaluación fenotípica multiambiental, sería la alternativa más eficiente en la identificación de marcadores significativos y de genotipos superiores.

Cuando el interés radica en la predicción de los méritos genéticos de los genotipos para la selección, más que en la identificación de regiones genómicas asociadas al fenotipo de interés, se usan los modelos de selección genómica (SG). La SG permite predecir características complejas tanto en animales (*e.g*., peso, altura, calidad de carne) como en plantas (*e.g*., rendimiento, calidad, estrés abiótico y biótico). Los modelos de SG, basados en información molecular masiva, están orientados a la predicción de méritos genéticos de los genotipos seleccionados (Burgueño *et al*., 2012). La SG supone que los QTL presentes tienden a estar en LD al menos con uno de los tantos marcadores evaluados. Para construir modelos de SG se usan también la metodología MLM-BLUP, métodos bayesianos y métodos de aprendizaje de maquinas (Wang *et al*., 2018). Particularmente los modelos RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*) y G-BLUP (*Genomic-Best Linear Unbiased Predictor*) han mostrado ser competentes para la SG (Heffner *et al*., 2011; Clark y van der Werf, 2013; Thavamanikumar *et al*., 2015; Wang *et al*., 2015; Wang, *et al*., 2018). La eficiencia de la SG es usualmente medida a través de la correlación entre los méritos genéticos predichos por el modelo basado en los marcadores y los verdaderos méritos genéticos (Rabier *et al*., 2016). Esta eficiencia podría depender de aspectos relacionados al diseño y a los modelos de análisis estadístico usados para evaluar las asociaciones entre los MM y el fenotipo (Bhat *et al*., 2016). En esta tesis se realiza una revisión sistemática de trabajos publicados de SG en vegetales y un meta-análisis para obtener una medida global de la eficiencia de la SG en relación a los modelos estadísticos y métodos de estimación usados para evaluar las asociaciones entre marcadores y fenotipo. Dada la abundancia de publicaciones científicas relacionadas a estudios de asociación,

también se ha desarrollado un protocolo analítico para implementar meta-análisis de regiones genómicas que podrían estar asociadas a un fenotipo de interés. La implementación de este tipo de protocolos de meta-análisis puede resultar complementaria en estudios de asociación genotipo-fenotipo sintetizando información publicada para una mejor comprensión de nuevos hallazgos.

En el Capítulo 2, se presenta el marco teórico de los métodos y modelos estadísticos usados en GWAS y en SG, que incluyen enfoques frecuentistas, bayesianos y de aprendizaje de máquinas. En el Capítulo 3 se evalúan modelos por ambiente y multiambientales para GWAS que incluyen correlaciones genéticas a través del pedigrí o de una matriz de similitud molecular y se realiza una comparación del desempeño de las metodologías propuestas en este capítulo. En el Capítulo 4 se realiza una revisión sistemática de literatura referente a la temática de SG en vegetales y se hace un meta-análisis para estimar la eficiencia de los métodos de estimación más usados en el contexto de SG en vegetales. En el Capítulo 5 se presenta un protocolo general para realizar meta-análisis de regiones genómicas que podrían estar asociadas a un fenotipo de interés, el cual se ilustra en la búsqueda de QTL de efecto mayor para enfermedades en maíz. Por último, en el Capítulo 6 se presentan conclusiones generales, comentarios finales y futuras posibles investigaciones.

HIPÓTESIS

La contribución al desarrollo de herramientas estadísticas y bioinformáticas permite identificar a partir de datos genómicos masivos, asociaciones relevantes entre genoma y fenotipo; así como, predecir valores fenotípicos a partir de datos moleculares. Este tipo de información es relevante en las diferentes etapas de los programas de mejoramiento genético vegetal, dado que permite identificar material genético promisorio para un caracter fenotípico de interés, en qué ambientes puede obtenerse un mejor desempeño de los genotipos analizados, qué marcadores moleculares influyen en el rasgo fenotípico a evaluar y qué genotipos pueden ser usados en etapas tempranas de los programas de mejora genética con la finalidad de obtener materiales de mejor performance, en el caracter fenotípico deseado, en una menor cantidad de tiempo.

OBJETIVO GENERAL

Implementar diferentes herramientas estadísticas y bioinformáticas que permitan derivar a partir de datos genómicos masivos, información complementaria que refuerce los resultados obtenidos a partir de la aplicación de estrategias de mapeo de asociación y selección genómica.

OBJETIVOS ESPECÍFICOS

1. Evaluar modelos GWAS que incorporan interacción G×E usando información de parentesco o similaridad genética respecto a su capacidad para identificar MM significativos y para predecir efectos de G para el ordenamiento y selección de los mismos.

2. Desarrollar rutinas de código con software de libre disposición para analizar estudios de asociación genómica.

3. Implementar un meta-análisis para analizar la eficiencia de los métodos de estimación más usados en estudios de SG en vegetales.

4. Desarrollar un protocolo de meta-análisis de estudios genéticos que pueda proveer información complementaria en estudios de asociación.

CAPÍTULO 2

ESTUDIOS DE ASOCIACIÓN Y DE SELECCIÓN GENÓMICA

MARCO TEÓRICO

ESTUDIOS DE ASOCIACIÓN

Desde la década de 1980, con el advenimiento de los marcadores moleculares (MM) y la percepción de sus ventajas, se abrieron nuevas oportunidades para su uso en programas de mejoramiento. El propósito central de usar este tipo de información, es el de contribuir a la selección utilizando información de ADN. El uso de MM fue visto como una alternativa importante para aumentar la comprensión de la arquitectura genética de un caracter cuantitativo, que siempre ha sido difícil de dilucidar (Ferrão *et al*., 2017). Una herramienta muy usada para este fin, es la selección asistida por marcadores (MAS, *marker-assited selection*). La aplicación de MAS fue motivada por la oportunidad de reducir tanto el costo como el tiempo y, en consecuencia, aumentar la ganancia genética esperada (Lande y Thompson, 1990). Los inicios del uso de información molecular en el estudio de caracteres fenotípicos de interés se basaron en la localización o mapeo de QTL (Soller y Plotkin-Hazan, 1977; Soller, 1978). Los caracteres cuantitativos se refieren a fenotipos que son controlados por dos o más genes (esto decir, múltiples genes) y afectados por factores ambientales, lo que da como resultado una variación continua en la población de estudio (Mackay *et al*., 2009). Los QTL son regiones o partes del genoma que albergan genes que afectan a un caracter cuantitativo de interés (Doerge, 2002). El análisis de QTL se caracteriza por dos componentes, i) identificar QTL y ii) estimar sus efectos (Jannink *et al*., 2010).

Los avances de la tecnología de secuenciación aceleran el desarrollo de la teoría de la genética cuantitativa molecular, como el análisis de QTL, el estudio de asociación de todo el genoma (GWAS) y la selección genómica (SG). El mapeo de QTL asume la existencia de pocas regiones del genoma que contienen genes que afectan a un caracter fenotípico. El

objetivo de esta técnica es ubicar dichos QTL en el genoma y estimar la magnitud de sus efectos sobre el caracter. No obstante, el análisis de QTL fue quedando relegado puesto que la proporción de varianza explicada por un QTL puede ser pequeña y difícil de detectar (de los Campos *et al*., 2013). Además, otro factor limitante es que requiere de una gran inversión de tiempo y recursos económicos para generar los datos que permiten emplear esta técnica, esto es, la creación de poblaciones de mapeo de manera experimental a través de cruzamientos biparentales. A partir de los años 2000, las estrategias GWAS y SG se utilizan cada vez más. Estos enfoques se aplican a poblaciones de individuos que no son diseñadas experimentalmente. El propósito de la SG es predecir el fenotipo o mérito genético de un individuo a través de información molecular. El objetivo principal de la SG ya no es la localización y estimación de efectos de los QTL sino el predecir el mérito genético de nuevos individuos en función de la información molecular disponible a veces desde temprana edad. El propósito es incorporar la mayor cantidad de información proveniente de MM a modelos estadísticos calibrados con paneles diversos de individuos para lograr predecir los valores genéticos de nuevos individuos. Tales predicciones posibilitan la selección de los mejores individuos en etapas tempranas de su ciclo de vida ya que, la información molecular puede obtenerse antes de registrar los caracteres fenotípicos. La posibilidad de evitar la observación del fenotipo, se traduce en una significativa reducción de tiempo y costos (de los Campos *et al*., 2009).

MAPEO ASOCIATIVO

El mapeo asociativo (MA) o el estudio de asociación del genoma completo (GWAS, por sus siglas en inglés *Genome-Wide Association Study*) es un método que permite identificar genes subyacentes relacionados con el fenotipo (Zhang *et al*., 2014; Abdullaev *et al*., 2017; Visscher *et al*., 2017; Xu *et al*., 2018) y que no demanda la creación de poblaciones experimentales sino que puede implementarse a través de paneles diversos o colecciones de individuos con suficiente variabilidad genética. Las variantes genéticas identificadas mediante GWAS pueden explicar distinta proporción de la variación fenotípica (Yang *et al*., 2010). Los modelos de MA ofrecen una gran posibilidad para identificar polimorfismos asociados con fenotipos y para comprender la base genética de la variación cuantitativa de los caracteres fenotípicos.

En el MA cada QTL puede ser identificado mediante una prueba de significancia estadística, donde muchos QTL serán ignorados debido a que gran parte de estos tienen un efecto genético menor que la varianza residual y por tanto, no logran alcanzar niveles significativos (Weedon *et al*., 2008; J. Yang *et al*., 2010). Los genes raros y los genes sin gran efecto pueden permanecer sin identificar debido a la falta de potencia estadística (Buckler *et al*., 2009). La potencia estadística está determinada por muchos factores como: el efecto del gen, la frecuencia alélica, el tamaño muestral, la densidad de MM y el error de tipo I (Pe'er *et al*., 2006; Hong y Park, 2012; Shin y Lee, 2015; Wang y Xu, 2019). Un ejemplo de cómo diseñar un MA teniendo en cuenta: la potencia estadística, el tamaño muestral y la estructura de los datos se encuentra en Ball (2013a).

Los métodos estadísticos usados en MA generalmente involucran modelos lineales mixtos (MLM), estos pueden explicar el efecto del marcador en el fenotipo, así como otros efectos, como lo son el ambiente y el grado de relación de un individuo con otros individuos en la población de mapeo. Estos métodos tienen como objetivo cuantificar la evidencia de los efectos genómicos asociados con la variación del caracter, controlando posibles asociaciones espurias o infladas causadas por la estructura poblacional (Ball, 2013b). Varios métodos se han propuesto para controlar la estructura poblacional: control genómico (Devlin y Roeder, 1999), asociación estructurada (Pritchard *et al*., 2000; Falush *et al*., 2003), modelo de regresión a un conjunto de marcadores (Setakis *et al*., 2006), componentes principales (Price *et al*., 2006) y modelo mixto donde se ajusta un conjunto de efectos aleatorios para cada individuo con covarianza basada en una matriz de parentesco estimada, $\widehat{K}$ (Ritland, 1996) o en un pedigrí conocido (Zhang *et al*., 2010). Según Astle y Balding (2009) para que sea efectivo el control genómico, la asociación estructurada y el modelo de regresión requiere aproximadamente 10^2 MM; las componentes principales requieren 10^4 marcadores y el modelo mixto con $\widehat{K}$ requiere de 10^4 a 10^5 marcadores. Se han propuesto modelos estadísticos diferentes para estudiar la asociación marcador-caracter en MA. Los primeros análisis de MA realizados en plantas resaltaron la importancia de contemplar las correlaciones genéticas entre los individuos de la población de mapeo para evitar falsos positivos o detección de asociaciones espurias entre el genotipo y el fenotipo. La estructura de la población se detectó utilizando el software STRUCTURE (Pritchard *et al*., 2000). Las correlaciones genéticas quedan expresadas en una matriz de dimensión (filas) igual al número de genotipos y de un número de columnas igual a la cantidad de grupos de genotipos que conforman la estructura,

denominada matriz Q. Los elementos de esta matriz se corresponden a las probabilidades *a posteriori* de cada genotipo de pertenecer a cada grupo. Más tarde, Parisseaux y Bernardo (2004) integraron la matriz de relaciones de parentesco (K) entre individuos a través de la incorporación de efectos aleatorios de individuo en un MLM donde la parte fija hace referencia al efecto de los marcadores sobre el fenotipo. Posteriormente, surgió el MLM de MA, denominado Q+K, donde Q es la matriz de estructura poblacional dada por STRUCTURE (Yu *et al*., 2006). Este modelo fue tratado y modificado por varios autores para mejorar la eficiencia computacional (Kang *et al*., 2008; Lippert *et al*., 2011; Listgarten *et al*., 2012; Zhou *et al*., 2013). No obstante, este modelo fue cuestionado por Price *et al*. (2006), quienes propusieron un MLM en el que se consideró como efecto fijo la matriz (P), conformada por el conjunto de variables sintéticas o componentes principales obtenidas al realizar el Análisis de Componentes Principales (Hotelling, 1936) sobre la matriz de frecuencias alélicas de los MM, en lugar de la matriz Q. Se observó un mejor desempeño del modelo (P+K) con respecto a los resultados arrojados por el modelo (Q+K) con menores errores de tipo I, mayor potencia para detectar asociaciones verdaderas y menor tiempo de cálculo para obtener la matriz de efectos fijos (Zhao *et al*., 2007). Las componentes principales son variables no correlacionadas y óptimas para señalar variabilidad y estructuras entre los genotipos de la población de mapeo (Peña Malavera, 2015). Las primeras componentes principales se caracterizan por ser las que más contribuyen a la variabilidad total según la prueba de Tracy y Widom (1994) y éstas son usadas como variables de efecto fijo o aleatorio en el modelo de MA. Otro enfoque de modelación fue el propuesto por Zhu y Yu (2009), donde se consideró como efecto fijo la matriz obtenida de un escalamiento multidimensional no métrico en lugar de las matrices Q o P. Se han evaluado comparativamente y desde criterios estadísticos las distintas alternativas de modelación para MA basadas en el marco teórico de los MLM que intentan modelar la variación fenotípica como la suma de los efectos de los MM más los efectos de estructura genética poblacional (Gutiérrez *et al*., 2011; Cappa *et al*., 2013; Peña Malavera *et al*., 2016). Por otra parte, la elección de la matriz a considerar en el MLM como matriz de varianza y covarianza de los efectos aleatorios ha sido ampliamente discutida (Zhao *et al*., 2007; Zhu y Yu, 2009). Una estrategia de modelación posible en el contexto del ajuste de un MLM es usar la matriz de relaciones de parentesco (K) que puede ser obtenida por: el software SPAGeDi (Hardy y

Vekemans, 2002), por el software TASSEL (Bradbury *et al*., 2007) o por la librería EMMA de R (Kang *et al*., 2008).

Luego de ajustar el modelo de MA, es frecuente utilizar algún método de corrección de valores-p por multiplicidad. Esto es necesario debido a las múltiples pruebas de hipótesis que se llevan a cabo marcador por marcador de manera simultánea que generan tantos contrastes estadísticos como marcadores haya. La probabilidad de detectar significancias debidas solo al azar aumenta con la acumulación de pruebas de hipótesis realizadas sobre los mismos datos. Para abordar esta problemática, existen distintos métodos de corrección de valores-p por multiplicidad que 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 puede realizarse con los métodos Bonferroni (1935), Benjamini y Hochberg (1995), Benjamini y Yekutieli (2001) para pruebas independientes. Dichos procedimientos de ajuste de la significancia estadística probaron ser útiles en el análisis de QTL tradicional donde los genotipos son independientes. Sin embargo, en MA con poblaciones estructuradas de diferente nivel de parentesco, estas correcciones pueden no desempeñarse de manera apropiada. El método propuesto por Li y Ji (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. El uso de corrección por multiplicidad se hace necesario en el MA con múltiples marcadores si el objetivo es probar hipótesis sobre los efectos de estos marcadores. Sin embargo, su importancia no es tal en modelos de SG donde la predicción del genotipo es de mayor interés que las pruebas de asociación entre el fenotipo y cada MM.

Numerosos trabajos de MA han sido realizados para un único ambiente por lo que los modelos estadísticos no incluyen el efecto de ambiente ni de interacción genotipo-ambiente (G×E). Hasta la fecha, casi todos los estudios de MA en plantas basados en poblaciones estructuradas examinaron los principales efectos de los factores genéticos sobre el caracter en un solo ambiente, pero no se han estudiado la interacción entre el genotipo y el ambiente en ensayos multiambientales (METs). En este trabajo, se aborda esta problemática desde el enfoque de los MLM con fines de comparar las estimaciones obtenidas en los ajustes de los modelos por ambiente y multiambientales, considerando en la estructura de varianza y covarianza la información del parentesco y de los marcadores moleculares. Esta propuesta permite la estimación de la componente de varianza de interacción G×E en los METs.

SELECCIÓN GENÓMICA

A pesar de la importancia a la hora de explicar las bases genéticas de los *loci* cuantitativos, el análisis de QTL ha mostrado inconvenientes que impiden su frecuente aplicación en los programas de mejoramiento (Bernardo, 2008). Además, este método tiene un mejor desempeño para caracteres controlados por genes de efecto mayor, que es un escenario poco usual en caracteres de importancia agronómica (Goddard y Hayes, 2007). Debido a estos inconvenientes, Meuwissen *et al*. (2001) presentaron la metodología de selección genómica (SG) que permite analizar de manera conjunta información molecular y caracteres fenotípicos de interés; también recibe el nombre de predicción genómica. En la última década, la SG se ha convertido en un enfoque promisorio ya que se ha observado una gran aceptación de dicha técnica, debido a sus bondades, a pesar de que sus inicios fueron en el contexto de mejoramiento animal; desde hace un tiempo, se ha implementado en mejoramiento vegetal (Meuwissen *et al*., 2001; Heffner *et al*., 2009; de los Campos *et al*., 2013). Como se mencionó anteriormente, el mapeo de QTL ha perdido influencia debido a la limitante que impone la creación de poblaciones experimentales en su aplicación práctica (Dekkers, 2004; Bernardo y Yu, 2007; Xu, 2008). Además, el genotipado de alto rendimiento fue impulsado por las técnicas de secuenciación de próxima generación (NGS, *next-generation sequencing*), éstas permiten obtener una alta cantidad de MM de manera más eficiente de tal modo que se reducen los costos en la generación de dicha información (Poland y Rife, 2012). Debida a la alta disponibilidad de MM y su bajo costo a la hora de su obtención, cambio la forma en que la información de ADN podría insertarse en los estudios genéticos (Ferrão *et al*., 2017). Gracias a las NGS, el genotipado se automatizó permitiendo de este modo una aplicación rutinaria y factible; además, se descubrió un gran número de marcadores de polimorfismo de un solo nucleótido (SNP) en todo el genoma de muchas especies (He *et al*., 2014). Esto conllevó a que métodos computacionales y estadísticos lograran manipular información molecular masiva de forma eficiente (Wang *et al*., 2015). Todo esto contribuyó al desarrollo de un nuevo método de selección asistida por marcadores, con mayor éxito. En síntesis, el análisis tradicional de QTL se basa en la detección, el mapeo y el uso de QTL de efecto mayor para la selección de caracteres. Mientras, la SG selecciona simultáneamente cientos o miles de MM que cubren el genoma completo de tal modo que la mayoría de los *loci* de caracteres cuantitativos están en desequilibrio de ligamiento (LD) con algunos de ellos (Meuwissen *et al*., 2001; Goddard y Hayes, 2007). En la SG se evidencia la

ausencia de cualquier prueba estadística para declarar si un marcador tiene un efecto estadísticamente significativo. Incluso los efectos que podrían ser demasiado pequeños se utilizarán para calcular el mérito genético. Además, cuando se utilizan marcadores que cubren todo el genoma, se asume LD entre QTL y marcadores en todas las familias/poblaciones, lo que da como resultado aplicaciones más amplias, incluso para caracteres con baja heredabilidad (Goddard y Hayes, 2007). Para un programa de mejoramiento consolidado, con esquemas de mejoramiento bien definidos, consistentemente respaldados por un buen germoplasma y experimentación, el uso práctico de la predicción genómica puede considerarse de fácil aplicación (Ferrão *et al*., 2017). Este tipo de programas contemplan tres tipos de bases de datos, denominados genéricamente como "poblaciones". El término población, en el contexto de SG, debe interpretarse como un conjunto de genotipos, donde los modelos predictivos serán entrenados, validados y aplicados. Estos conceptos tienen una estrecha relación con los términos comúnmente utilizados en el área de aprendizaje estadístico (*statistical learning*), especialmente cuando se refiere a temas como el remuestreo y la validación cruzada (James *et al*., 2013). El primer conjunto de datos es la población de entrenamiento o de referencia (TRN, *training population*) (Goddard y Hayes, 2007; Nakaya e Isobe, 2012; Desta y Ortiz, 2014). En esta población se define un modelo predictivo y se estiman los efectos alélicos. Los individuos que pertenecen a TRN (líneas, clones, familias, etc.) deben ser genotipados y fenotipados para los caracteres de interés. Un desafío muy frecuente en esta instancia, es definir qué individuos deben componer esta población de referencia. A pesar de esta problemática, se espera que esta población está compuesta por materiales prometedores, sobre los cuales el mejorador tiene particular interés en aplicar métodos de selección y, por lo tanto, obtener nuevos cultivares (Ferrão *et al*., 2017). Esta especificación tendrá importantes consecuencias sobre la capacidad predictiva de la SG. A continuación, se debe definir un segundo conjunto de datos llamado población de validación o prueba (TST, *testing population*) (Goddard y Hayes, 2007; Nakaya e Isobe, 2012; Desta y Ortiz, 2014). Generalmente, esta población es más pequeña que la TRN y también incluye individuos que deben ser genotipados y fenotipados. La contribución al usar la TST, es verificar la ecuación predictiva de eficiencia definida en el paso anterior. El mérito genético (GEBV, *genome-estimated breeding values*) se obtiene al utilizar la estimación del efecto de marcador en el TNR y correlacionarlo con los valores fenotípicos verdaderos (Desta y Ortiz, 2014). Este resultado se llama precisión predictiva

(Ould Estaghvirou *et al*., 2013) y se considera como la métrica estándar para evaluar la eficiencia de la SG. Su magnitud proporcionará una medida importante de la capacidad de la SG para predecir fenotipos basándose únicamente en datos genotípicos (Ferrão *et al*., 2017). El último conjunto de datos se denomina comúnmente población de mejora (Goddard y Hayes, 2007; Nakaya e Isobe, 2012; Desta y Ortiz, 2014). Esta es la población en la que se aplicará SG directamente, siendo muy importante su rol en los programas de mejoramiento. Si la precisión es satisfactoria, los MM se convierten en la unidad de evaluación en el programa de mejoramiento. Los efectos estimados en el TST y validados en el TRN se utilizarán, por lo tanto, para predecir nuevos fenotipos. En este momento, la selección se guiará únicamente por la información del marcador (Lorenz *et al*., 2011). Por esta razón, la selección se puede realizar en las primeras etapas (*e.g*., plántulas dentro de invernaderos), lo que resulta en un ahorro de tiempo y de evaluaciones a campo (suponiendo que los costos de genotipado sean menores) (Ferrão *et al*., 2017). Genotipar y fenotipar son aspectos importantes a tener en cuenta para la implementación práctica. El estado final de un caracter será el resultado acumulativo de una serie de interacciones causales entre la composición genética del genotipo y el ambiente en el que se desarrolló la planta (Malosetti *et al*., 2013). El éxito de la SG depende estrechamente del ambiente en el que se miden los fenotipos. Esta variabilidad del efecto de G a través de los ambientes donde se evalúan, evidencia la presencia de la G×E (Burgueño *et al*., 2012; Cuevas *et al*., 2016, 2017, 2018; Sukumaran *et al*., 2017). La inclusión de la G×E puede aumentar la capacidad de detectar nuevos genes cuyos efectos pueden estar enmascarados en la interacción. Cuando los efectos debidos a la G×E no son contemplados en el modelo de SG, podrían generarse sesgos en la precisión de la predicción (Desta y Ortiz, 2014).

La capacidad predictiva dependerá de los factores genéticos y no genéticos bajo análisis. Un concepto a considerar que está estrechamente relacionado con la definición teórica de SG, es el LD, también conocido como asociación alélica. El LD es la asociación no aleatoria de alelos en diferentes *loci* (Flint-Garcia *et al*., 2003). La correlación entre polimorfismos es causada por su historia compartida de mutación y recombinación. Los términos ligamiento y LD a menudo se confunden. Aunque LD y ligamiento son conceptos relacionados, son intrínsecamente diferentes. El ligamiento se refiere a la herencia correlacionada de *loci* a través de la conexión física en un cromosoma, mientras que el LD se refiere a la correlación entre alelos en una población (Ott *et al*., 2011). En general, todas las fuentes que afectan el

equilibrio de Hardy Weinberg (HW) podrían influir en los patrones de LD (Flint-Garcia *et al*., 2003). En el contexto de SG, el concepto de LD juega un rol importante, ya que de este depende la determinación del número y la densidad de MM, como el diseño experimental apropiado para realizar el análisis de asociación (Flint-Garcia *et al*., 2003; Mackay y Powell, 2007). El éxito de la SG está directamente asociado con la distancia genética entre la población de referencia (TRN), donde se entrena el modelo, y la población de mejora, donde las estimaciones de los efectos de marcador se utilizan como unidad de selección (Ferrão *et al*., 2017). La precisión de las estimaciones de los efectos de marcador aumenta con el tamaño muestral, porque el sesgo y la varianza de dichas estimaciones, disminuyen al aumentar el tamaño de la muestra. Además, se espera que al aumentar el tamaño de la muestra, también aumente la relación genética entre los conjuntos de datos TRN y TST, que se describió previamente como un factor importante (Ferrão *et al*., 2017).

El tamaño de la población ha sido muy variable en los estudios de SG; en una revisión sobre el tema, Nakaya e Isobe (2012) utilizaron, para cereales como el maíz, la cebada y el trigo, un tamaño promedio de 258 individuos en el conjunto de datos de TRN. Por otro lado, la cantidad de observaciones es mayor en estudios forestales, donde, en promedio, 673 individuos constituyen la TRN. Los estudios en plantas han demostrado que se requieren tamaños del conjunto de datos TRN más pequeños, en relación con los usados en animales. Los autores señalan dos factores que suelen ser causales: la estrecha diversidad genética en las poblaciones de plantas, que es causada principalmente por los autocruzamientos, la calidad de las evaluaciones fenotípicas, ya que un buen diseño experimental es más común en plantas que en mejoramiento animal (Ferrão *et al*., 2017).

El mérito genético está altamente correlacionado con la heredabilidad, ésta se define como, la proporción de la varianza genotípica en la varianza total o varianza fenotípica entre individuos de una población. Por lo tanto, en SG se espera que aumente la precisión en los caracteres que son influenciados por factores genéticos y con menos efectos ambientales. La relación directa entre precisión y heredabilidad ha sido evidenciada en estudios por simulación (Daetwyler *et al*., 2013). En el contexto de SG, se propone un modelo estadístico para asociar observaciones fenotípicas con variaciones a nivel de ADN. Se puede definir una gran cantidad de modelos para vincular estas variables (Ferrão *et al*., 2017). Un enfoque simple y común es la regresión lineal de un solo marcador, este modelo se hace marcador por marcador, es decir, se harán tantas regresiones como MM hayan. El problema no puede

abordarse como una regresión múltiple ya que el número de MM (variables) es usualmente mayor al número de individuos de la población (Neves *et al*., 2012). Una solución a esta problemática, es modelar los efectos de los MM como efectos aleatorios y hacer supuestos previamente sobre la varianza explicada. Algunos métodos no lineales como Bayes A, Bayes B (Meuwissen *et al*., 2001; Habier *et al*., 2011; Gianola, 2013) y Bayes C (Habier *et al*., 2011) dan más énfasis a algunas regiones genómicas al permitir que la varianza difiera entre *loci* de MM, mientras que el método G-BLUP (del inglés, *Genomic-Best Linear Unbiased Predictor*) asigna la misma varianza a todos los *loci*, es decir, les da la misma importancia a todos (Clark y van der Werf, 2013). Otro método usado es RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*) que se caracteriza por: i) asumir que todos los MM tienen las mismas varianzas pero con un efecto distinto de cero, ii) estimar la matriz de relaciones a partir de los MM y iii) permitir que algunos QTL estén en LD con algún *loci* mientras que otros no (Meuwissen *et al*., 2001; Heffner *et al*., 2011). Una versión restringida de mínimos cuadrados ordinarios y de selección de variables que se utiliza en el contexto de modelos de SG es LASSO (*Least Absolute Shrinkage and Selection Operator*). Éste es indiferente en la identificación de MM estrechamente correlacionados y tiende a seleccionar algunos MM e ignorar otros (Friedman *et al*., 2010; Li y Sillanpää, 2012). La correlación entre las variables predictoras produce multicolinealidad. En este sentido, el modelo RR-BLUP, a pesar de que no se utiliza para seleccionar variables, es recomendado respecto a la aplicación del método LASSO. El modelo RKHS (*Reproducing Kernel Hilbert Space*) es efectivo para detectar efectos genéticos no aditivos, es decir, dominancia y epistasis. RKHS se basa en distancias genéticas y usa una función kernel que tiene un parámetro de suavizado para regular la distribución de los efectos de los QTL (Gianola y van Kaam, 2008; de los Campos *et al*., 2010). Un método no paramétrico de aprendizaje supervisado son las máquinas de soporte vectorial (SVM, *Support Vector Machine*), que se pueden utilizar con fines de clasificación y regresión. No obstante, en los últimos años, las SVM se han usado para fines predictivos cuando hay bases de datos con alta dimensionalidad y que presentan complejidad en el análisis. Las SVM proponen una posible solución a esta problemática a través del uso de funciones kernel (Wang *et al*., 2018). Si bien el desarrollo de la aplicación de los modelos de SG es relativamente novedoso para distintas especies. En la práctica, la selección de un único método de SG en términos de la eficiencia predictiva es compleja debido a que existen

varios factores que están interrelacionados de manera compleja e integral. En esta tesis se compara la eficiencia predictiva de la SG en programas de mejoramiento en cereales.

CAPÍTULO 3

MODELOS GWAS ESTIMADOS DESDE ENSAYOS MULTIAMBIENTALES

INTRODUCCIÓN

Los estudios de asociación del genoma completo (GWAS, por sus siglas en inglés *Genome-wide association study*) se han convertido en una herramienta importante en la identificación de genotipos superiores basados en la alta disponibilidad de marcadores moleculares (MM) relacionados a caracteres de interés en el mejoramiento genético vegetal (Asoro *et al*., 2011; Bhat *et al*., 2016; Heffner *et al*., 2011; Lorenzana y Bernardo, 2009; Rafalski, 2010; Xavier *et al*., 2018). Los GWAS han sido principalmente usados en el contexto de ensayos de un solo ambiente, evitando la necesidad de estimar correlaciones entre y dentro de ambientes (Burgueño *et al*., 2012). No obstante, al evaluar los efectos de marcador y de genotipo bajo diferentes condiciones ambientales, el impacto del efecto de la interacción genotipo por ambiente (G×E) puede ser medido y las correlaciones dentro ambientes ser usadas para ajustar las pruebas estadísticas para los efectos de MM (Covarrubias-Pazaran, 2016). En los METs, los modelos GWAS pueden ser usados para identificar los mejores genotipos, el mejor subconjunto de marcadores asociados con buenos fenotipos y también para estimar la contribución relativa de los efectos del genotipo (G) y de la interacción (G×E) a la varianza total de los caracteres fenotípicos. Además, los METs permiten analizar no solo el comportamiento de los fenotipos con alto desempeño; sino también la estabilidad fenotípica a través de ambientes (Lopez-Cruz *et al*., 2015). Los modelos de efectos lineales mixtos (West *et al*., 2014), como los modelos de ANAVA mixtos, han sido ampliamente usados para estimar las componentes de varianza de los efectos de G y de G×E en los METs con datos fenotípicos (Balzarini, 2002; Kang *et al*., 2004; Yang, 2007). Sin embargo, el ajuste de los modelos METs que involucran información molecular es más reciente. Sripathi *et al*. (2018) señalaron que modelar los efectos de G×E pueden mejorar la capacidad predictiva de

los efectos de G cuando la varianza del efecto de G×E es considerablemente mayor comparada con la varianza del efecto de G. En ese caso, las componentes de varianza de G y de G×E deben estimarse por separado y los METs son esenciales para la identificación de estas componentes. Sin embargo, la principal función de los modelos GWAS es evaluar los efectos de los marcadores moleculares y la prueba estadística para evaluarlos depende de estas varianzas (Lado *et al*., 2016). Con respecto a la estimación de las correlaciones, varios enfoques han sido aplicados en modelos GWAS (Cappa *et al*., 2013; Gutiérrez *et al*., 2015) en el contexto de un único ambiente, estimando las correlaciones ya sea basado en parentesco "pedigrí" o a través de similitud molecular entre genotipos. Sin embargo, el desempeño de tales modelos GWAS en el contexto de METs han sido menos explorados. Es por ello que el modelo de un solo ambiente, o análisis por ambiente, es el enfoque estadístico más utilizado para evaluar los efectos de los marcadores a través de GWAS. Sin embargo, un único modelo MET de análisis es factible (Covarrubias-Pazaran, 2016). Por lo tanto, es interesante comparar estrategias alternativas para GWAS en el contexto de METs, es decir, modelos por ambiente versus modelos multiambientales.

Cuevas *et al*. (2018) explicaron que la implementación de modelos basados en MM puede ser apropiada tanto para estimar los parámetros del modelo como para predecir G×E. (Lopez-Cruz *et al*., 2015) afirmaron que la presencia de G×E, expresada como un cambio en el desempeño relativo de las líneas genéticas a través de los ambientes, se puede medir por medio de correlaciones entre el ordenamiento de los genotipos entre ambientes. Evaluar G×E en los METs mediante el uso de información de pedigrí para modelar las correlaciones genéticas entre genotipos, ha sido implementada para datos fenotípicos con modelos lineales mixtos (MLM) (Burgueño *et al*., 2007; Crossa *et al*., 2006; Smith *et al*., 2005). Alternativamente, las correlaciones genéticas se han estimado a partir de la información de los MM. El tratamiento estadístico de G×E ha evolucionado a lo largo del tiempo debido al desarrollo de métodos estadísticos que consideran dicha componente (Lopez-Cruz *et al*., 2015). Investigaciones previas para predecir los efectos de genotipo usando información molecular densa ha sido conducida en los METs, concluyendo que evaluar el efecto de G podría beneficiar los modelos multiambientales. Por lo tanto, se puede usar un modelo multiambiental para seleccionar marcadores y genotipos cuyos fenotipos se observan bajo condiciones ambientales variables, incluso cuando la información del MM se evalúa sólo una vez. Sin embargo, los GWAS se llevan a cabo comúnmente vinculando el rendimiento

fenotípico promedio con la información del MM. El objetivo de este capítulo fue comparar el desempeño, en términos de la estimación de las componentes de varianza, las puntuaciones de los marcadores y los BLUP de los efectos de G, de los modelos por ambiente y multiambientales para GWAS que incluyen correlaciones genéticas a través de un pedigrí o una matriz de similitud de marcadores moleculares.

MATERIALES Y MÉTODOS

BASE DE DATOS

Se trabajó con una base de datos pública disponible en el paquete BGLR de R (R Core Team, 2020) conformado por 599 genotipos de trigo evaluados fenotípicamente en 4 ambientes y genotipados con 1279 marcadores DArT (McLaren *et al*., 2000; McLaren *et al*., 2005). El caracter fenotípico estuvo representado por valores de rendimientos estandarizados por ambiente y los datos genotípicos estaban codificados según la presencia/ausencia del marcador molecular como, 1 y 0, respectivamente. Para cada genotipo, se simularon tres repeticiones de observaciones fenotípicas en cada ambiente agregando un término de error aleatorio normalmente distribuido con media cero y varianza igual a 0.5 en todos los ambientes.

MODELACIÓN ESTADÍSTICA

MODELOS ESTADÍSTICOS COMPARADOS

Se ajustaron dos modelos por ambiente y dos modelos multiambientales. En cada tipo de modelo se consideró la estructura genética a través de la matriz de pedigrí o, alternativamente, mediante la similitud molecular. La combinación del tipo de modelo y la forma de incorporar información sobre las correlaciones genéticas entre genotipos dio como resultado cuatro estrategias de modelado: M1–Modelo ajustado por ambiente, incorpora información de pedigrí para tener en cuenta las correlaciones entre líneas; M2–Modelo ajustado por ambiente, con correlaciones entre líneas estimadas a partir de la similitud molecular; M3–Modelo multiambiental con información de pedigrí; M4–Modelo multiambiental con similitud molecular. Para el GWAS de los modelos MET, se llevaron a cabo los siguientes dos pasos. Primero, las estimaciones de la varianza genética (Var (G)),

la varianza de la interacción entre genotipo y ambiente (Var (G×E)) y la componente de la varianza residual ((Var (ε)) fueron estimadas por Máxima verosimilitud restringida (REML) (por sus siglas en inglés, *Restricted Maximum Likelihood*) (Patterson y Thompson, 1971) en el marco del modelo lineal mixto para datos fenotípicos. El modelo lineal mixto incluía los efectos aleatorios de G y G×E, el efecto fijo de ambiente (E) y las correlaciones genéticas entre los efectos de G y de G×E. En segundo lugar, los efectos de marcador se estimaron para cada marcador utilizando las varianzas totales estimadas, es decir, se contemplan las varianzas de G, G×E y las varianzas residuales. Para los modelos por ambiente no se incluyeron los efectos aleatorios G×E en el modelo; pero la correlación entre los efectos genéticos aleatorios se modeló de la misma manera que los modelos multiambientales (utilizando el pedigrí o la similitud de los marcadores moleculares). Todos los modelos fueron ajustados usando el paquete *sommer* de R (Covarrubias-Pazaran, 2016). En este capítulo, se incorporó a modo de anexo la rutina de R que se usó para ajustar los modelos por ambiente y multiambientales (utilizando el pedigrí o la similitud de los marcadores moleculares). Los coeficientes de correlación de Pearson entre los BLUP de los efectos de G obtenidos de los diferentes enfoques de modelado fueron calculados. Además, para la selección de modelos se obtuvieron los criterios AIC y BIC para comparar enfoques alternativos para manejar las correlaciones genéticas dentro de cada tipo de modelo (modelo por ambiente o multiambiental).

ESPECIFICACIÓN DEL MODELO

M1–Modelo por ambiente con información de pedigrí

$$\mathbf{y_{n\times 1}} = \mathbf{X\beta} + \mathbf{Z^1 u_g} + \mathbf{M\tau} + \boldsymbol{\varepsilon} \qquad [3.1]$$

donde $\mathbf{y}_{n\times 1}$ es el vector de la variable de respuesta (rendimiento de la parcela en el ambiente que se está evaluando), n es el número de observaciones (en nuestro caso es igual a 1797, que corresponde a $j = 1,..,599$ genotipos con tres repeticiones cada uno), $\mathbf{X}$ es la matriz de incidencia para el efecto fijo, $\boldsymbol{\beta}$ es el intercepto del efecto fijo, $\mathbf{Z}^1$ es la matriz de incidencia para el efecto aleatorio, $\mathbf{u_g}$ es el efecto aleatorio que contiene la información genética de cada línea, este efecto aleatorio tiene distribución $\mathbf{u_g} \sim N(0, \mathbf{G}^1)$, donde $\mathbf{G}^1 = \mathbf{K}_{j\times j}\sigma^2_g$ y $\mathbf{K}$ es la matriz de parentesco estimada por la información del pedigrí, $\mathbf{M}$ es la matriz de incidencia

asociada a los efectos de marcador, $\boldsymbol{\tau}$ es el vector que modela el efecto aditivo DArT como un efecto fijo y $\boldsymbol{\varepsilon}$ es el vector del término de error con varianza residual $\mathbf{R} = \mathbf{I}_{n\times n}\sigma_{\epsilon}^2$, se supone que este vector está distribuido normalmente $\boldsymbol{\varepsilon}\sim N(0, \mathbf{R})$. Siendo σ_{g}^2 y σ_{ϵ}^2 las varianzas genéticas y residuales, respectivamente.

Para cada marcador, una prueba de significación basada en la varianza fenotípica aplicada a los efectos del marcador evalúa la importancia del marcador en cada ambiente. La matriz de varianza y covarianza de $\mathbf{y}_{n\times 1}$ se calcula como

$$\mathbf{V}(\mathbf{y_{n\times 1}}) = \mathbf{Z^1 G^1 Z^{1'}} + \mathbf{R} \qquad [3.2]$$

y los efectos de marcador se obtuvieron de la siguiente manera

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y_{n\times 1}} \qquad [3.3]$$

Utilizamos un nivel de valor p conservativo para el error tipo I, es decir, el valor p <0,001, correspondiente a una puntuación de asociación (-Log (valor p)) ≥ 3,0.

M2–Modelo por ambiente incluyendo similitud molecular

$$\mathbf{y_{n\times 1}} = \mathbf{X\beta} + \mathbf{Z^1 u_g} + \mathbf{M\tau} + \boldsymbol{\varepsilon} \qquad [3.4]$$

donde $\mathbf{y}_{n\times 1}$ es el vector de la variable de respuesta (rendimiento de la parcela en el ambiente que se está evaluando), n es el número de observaciones (en nuestro caso es igual a 1797, que corresponde a $j = 1,..,599$ genotipos con tres repeticiones cada uno), $\mathbf{X}$ es la matriz de incidencia para el efecto fijo, $\boldsymbol{\beta}$ es el intercepto del efecto fijo, $\mathbf{Z}^1$ es la matriz de incidencia para el efecto aleatorio, $\mathbf{u_g}$ es la información genética de cada línea y es considerado como efecto aleatorio, este efecto aleatorio tiene distribución $\mathbf{u_g}\sim N(0, \mathbf{G}^1)$, donde $\mathbf{G}^1 = \mathbf{A}_{j\times j}\sigma_{\mathrm{g}}^2$ y $\mathbf{A}$ es la matriz de relaciones aditivas estimada por similitud molecular, $\mathbf{M}$ es la matriz de incidencia asociada a los efectos de marcador, $\boldsymbol{\tau}$ es el vector que modela el efecto aditivo DArT como un efecto fijo y $\boldsymbol{\varepsilon}$ es el vector del término de error con varianza residual $\mathbf{R} = \mathbf{I}_{n\times n}\sigma_{\epsilon}^2$, se supone que este vector está distribuido normalmente $\boldsymbol{\varepsilon}\sim N(0, \mathbf{R})$. Siendo σ_{g}^2 y σ_{ϵ}^2 las varianzas genéticas y residuales, respectivamente.

Para cada marcador, una prueba de significación basada en la varianza fenotípica aplicada a los efectos del marcador evalúa la importancia del marcador en cada ambiente. La matriz de varianza y covarianza de $\mathbf{y}_{n\times 1}$ se calcula como

$$\mathbf{V}(\mathbf{y}_{\boldsymbol{n\times 1}}) = \mathbf{Z^1 G^1 Z^{1'}} + \mathbf{R} \qquad [3.5]$$

y los efectos de marcador se obtuvieron de la siguiente manera

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X'V^{-1}X})^{-1}\mathbf{X'V^{-1}y}_{\boldsymbol{n\times 1}} \qquad [3.6]$$

Utilizamos un nivel de valor p conservativo para el error tipo I, es decir, el valor p <0,001, correspondiente a una puntuación de asociación (-Log (valor p)) ≥ 3,0.

M3–Modelo multiambiental con información de pedigrí

$$\mathbf{y}_{\boldsymbol{N\times 1}} = \mathbf{X\beta} + \mathbf{Z^1 u_g} + \mathbf{Z^2 u_{g\times e}} + \mathbf{M\tau} + \boldsymbol{\varepsilon} \qquad [3.7]$$

donde $\mathbf{y}_{N\times 1}$ es el vector de la variable respuesta (rendimiento de la parcela a través de los ambientes) en el *i*-ésimo ambiente (en nuestro caso, i = 1, .., 4) y *j*-ésimo genotipo, siendo N igual a 7,188 (es decir, genotipos × ambientes × repeticiones), $\mathbf{X}$ es la matriz de incidencia para el efecto fijo, $\boldsymbol{\beta}$ es el intercepto y efecto fijo de ambiente, $\mathbf{Z}^1$ es la matriz de incidencia para el efecto aleatorio, $\mathbf{u_g}$ es el efecto aleatorio de la información genética de cada línea, este efecto aleatorio tiene distribución $\mathbf{u_g} \sim N(0, \mathbf{G}^1)$, donde $\mathbf{G}^1 = \mathbf{K}_{j\times j}\sigma_g^2$ y $\mathbf{K}$ es la matriz de parentesco estimada por la información del pedigrí, $\mathbf{Z}^2$ es la matriz de incidencia para el efecto aleatorio de la interacción genotipo por ambiente, $\mathbf{u_{g\times e}}$ es el efecto aleatorio de la interacción genotipo por ambiente, este efecto aleatorio tiene distribución $\mathbf{u_{g\times e}} \sim N(0, \mathbf{G}^2)$, donde $\mathbf{G}^2 = (\mathbf{I}_{i\times i} \otimes \mathbf{K}_{j\times j})\sigma_{ge}^2$ $\mathbf{I}_{i\times i}$ es una matriz identidad, el símbolo $\otimes$ se refiere a el producto de kronecker, $\mathbf{M}$ es la matriz de incidencia asociada a los efectos de marcador, $\boldsymbol{\tau}$ es el vector que modela el efecto aditivo DArT como un efecto fijo y $\boldsymbol{\varepsilon}$ es el vector del término de error con varianza residual $\mathbf{R} = \mathbf{I}_{N\times N}\sigma_\epsilon^2$, se supone que este vector está distribuido normalmente $\boldsymbol{\varepsilon} \sim N(0, \mathbf{R})$. Siendo σ_g^2, σ_{ge}^2 y σ_ϵ^2 las varianzas genéticas, de interacción genotipo por ambiente y residuales, respectivamente.

Para cada marcador, una prueba de significación basada en la varianza fenotípica aplicada a los efectos del marcador evalúa la importancia del marcador en cada ambiente. La matriz de varianza y covarianza de $\mathbf{y}_{N\times 1}$ se calcula como

$$\mathbf{V}(\mathbf{y}_{N\times 1}) = \mathbf{Z}^1\mathbf{G}^1\mathbf{Z}^{1'} + \mathbf{Z}^2\mathbf{G}^2\mathbf{Z}^{2'} + \mathbf{R} \qquad [3.8]$$

y los efectos de marcador se obtuvieron de la siguiente manera

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}_{N\times 1} \qquad [3.9]$$

Utilizamos un nivel de valor p conservativo para el error tipo I, es decir, el valor p <0,001, correspondiente a una puntuación de asociación (-Log (valor p)) ≥ 3,0.

M4–Modelo multiambiental incluyendo similitud molecular

$$\mathbf{y}_{N\times 1} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}^1\mathbf{u}_{\mathbf{g}} + \mathbf{Z}^2\mathbf{u}_{\mathbf{g}\times\mathbf{e}} + \mathbf{M}\boldsymbol{\tau} + \boldsymbol{\varepsilon} \qquad [3.10]$$

donde $\mathbf{y}_{N\times 1}$ es el vector de la variable respuesta (rendimiento de la parcela a través de los ambientes) en el *i*-ésimo ambiente (en nuestro caso, *i* = 1, .., 4) y *j*-ésimo genotipo, siendo *N* igual a 7,188 (es decir, el producto entre los *j* genotipos × *i* ambientes × *n* repeticiones), $\mathbf{X}$ es la matriz de incidencia para el efecto fijo, $\boldsymbol{\beta}$ es el intercepto y efecto fijo de ambiente, $\mathbf{Z}^1$ es la matriz de incidencia para el efecto aleatorio, $\mathbf{u}_{\mathbf{g}}$ es el efecto aleatorio de la información genética de cada línea, este efecto aleatorio tiene distribución $\mathbf{u}_{\mathbf{g}}\sim N(0, \mathbf{G}^1)$, donde $\mathbf{G}^1 = \mathbf{A}_{j\times j}\sigma^2_{\mathrm{g}}$ y $\mathbf{A}$ es la matriz de relaciones aditivas estimada por similitud molecular, $\mathbf{Z}^2$ es la matriz de incidencia para el efecto aleatorio de la interacción genotipo por ambiente, $\mathbf{u}_{\mathbf{g}\times\mathbf{e}}$ es el efecto aleatorio de la interacción genotipo por ambiente, este efecto aleatorio tiene distribución $\mathbf{u}_{\mathbf{g}\times\mathbf{e}}\sim N(0, \mathbf{G}^2)$, donde $\mathbf{G}^2 = (\mathbf{I}_{i\times i} \otimes \mathbf{A}_{j\times j})\sigma^2_{\mathrm{ge}}$ $\mathbf{I}_{i\times i}$ es una matriz identidad, el símbolo $\otimes$ se refiere a el producto de kronecker, $\mathbf{M}$ es la matriz de incidencia asociada a los efectos de marcador, $\boldsymbol{\tau}$ es el vector que modela el efecto aditivo DArT como un efecto fijo y $\boldsymbol{\varepsilon}$ es el vector del término de error con varianza residual $\mathbf{R} = \mathbf{I}_{N\times N}\sigma^2_{\epsilon}$, se supone que este vector está distribuido normalmente $\boldsymbol{\varepsilon}\sim N(0, \mathbf{R})$. Siendo σ^2_{g}, σ^2_{ge} y σ^2_{ϵ} las varianzas genéticas, de interacción genotipo por ambiente y residuales, respectivamente.

Para cada marcador, una prueba de significación basada en la varianza fenotípica aplicada a los efectos del marcador evalúa la importancia del marcador en cada ambiente. La matriz de varianza y covarianza de $\mathbf{y}_{N\times 1}$ se calcula como

$$\mathbf{V}(\mathbf{y}_{N\times 1}) = \mathbf{Z}^1\mathbf{G}^1\mathbf{Z}^{1'} + \mathbf{Z}^2\mathbf{G}^2\mathbf{Z}^{2'} + \mathbf{R} \qquad [3.11]$$

y los efectos de marcador se obtuvieron de la siguiente manera

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}_{N\times 1} \qquad [3.12]$$

Utilizamos un nivel de valor p conservativo para el error tipo I, es decir, el valor p <0,001, correspondiente a una puntuación de asociación (-Log (valor p)) ≥ 3,0.

RESULTADOS

Las estimaciones de las componentes de varianzas para los modelos por ambiente (M1 y M2) y los modelos multiambientales (M3 y M4) se muestran en la Tabla 3.1. La magnitud de las estimaciones de la varianza genética estuvo condicionada a la información suministrada (pedigrí o similitud molecular) para modelar las correlaciones genéticas. La varianza genética en los modelos que usaron la similitud molecular para explicar la correlación genética (M2 y M4) fue mayor que las estimaciones producidas por los modelos M1 y M3, donde la correlación entre individuos se agregó a través de la información del pedigrí. La componente de varianza para la interacción G×E solo puede identificarse en los modelos multiambientales. Para los datos de ejemplo usados en este capítulo, la componente G×E fue mayor que la varianza de G. Los índices para la selección de modelos AIC y BIC indicaron para el modelo por ambiente, que el mejor modelo para un ambiente dado no es el mejor modelo en otro ambiente; por lo tanto, estos índices fueron más bajos para el M1 en los ambientes 3 y 4; pero fueron más altos en los ambientes 3 y 4 de M2. Dado que las matrices K y A usadas en el modelado de los datos fenotípicos en cada ambiente son las mismas y que los modelos son homocedásticos, las diferencias observadas son atribuidas a los efectos G×E, que no pueden ser estimadas para este tipo de modelos; y, por lo tanto, son absorbidas en las estimaciones de las varianzas genéticas. Para los ambientes 3 y 4, las varianzas de G del modelo M2 fueron más altas y la precisión con la cual fueron estimadas fue más baja que la presentada en los ambientes 1 y 2, los resultados sugieren una mayor confusión entre los efectos de interacción y los efectos genéticos en los ambientes 3 y 4, es

decir, se sobreestima la varianza genética. Bajo esta situación, donde los ambientes presentan mayor confusión entre la Var (G) y Var (G×E), como sucede en este trabajo en los ambientes 3 y 4, el modelo M1 presentó mejor ajuste sobre el modelo M2 como lo muestran los criterios AIC y BIC. Estos criterios son útiles para comparar modelos con las mismas observaciones de la variable respuesta, por ello, en este trabajo, M1 siempre se comparó con M2 dentro de cada ambiente. Un valor más bajo de AIC o BIC indica un mejor modelo.

Los modelos GWAS multiambientales permitieron proporcionar una estimación de las componentes de varianza de G y G×E con una buena precisión predictiva. Es importante apreciar que las estimaciones de la Var (G) y Var (G×E) en M4 también fueron más altas que las estimaciones producidas por M3 (Tabla 3.1). Como mostraron los resultados en los modelos ambiente por ambiente, al usar marcadores moleculares para medir las correlaciones genéticas, las varianzas estimadas son más altas que al usar el pedigrí. El mismo patrón se observó para las covarianzas calculadas a partir de los marcadores moleculares, es decir, fueron más altas cuando en el modelo se incorporó la similitud molecular para la estimación de las correlaciones genéticas en lugar de la información de pedigrí.

Tabla 3.1. Componentes de varianza de genotipo (G), genotipo por ambiente (G×E) y residual (ε).

Modelo	Amb	Var(G)	Var(G × E)	Var(ε)	AIC	BIC
			Modelos por ambiente			
1	1	0,785 ± 0,0696	NI	0,308 ± 0,0147	832	838
	2	0,676 ± 0,0617	NI	0,334 ± 0,0159	876	882
	3	0,709 ± 0,0636	NI	0,305 ± 0,0145	775	781
	4	0,716 ± 0,0642	NI	0,308 ± 0,0147	793	798
	Promedio	0,722 ± 0,0324	NI	0,314 ± 0,0075		
2	1	4,377 ± 0,4020	NI	0,307 ± 0,0147	815	820
	2	4,350 ± 0,4013	NI	0,315 ± 0,0151	847	853
	3	5,482 ± 0,4863	NI	0,301 ± 0,0145	887	892
	4	4,546 ± 0,4167	NI	0,314 ± 0,0151	864	869
	Promedio	4,689 ± 0,2133	NI	0,309 ± 0,0074		
			Modelos multiambientales			
3	Todos	0,194 ± 0,0289	0,554 ± 0,0297	0,313 ± 0,0074	3200	3227
4	Todos	1,635 ± 0,2066	3,214 ± 0,1788	0,310 ± 0,0074	3294	3322

Componentes de varianza estimadas por REML ± Error estándar
NI: No Identificable
AIC: Criterio de Información de Akaike (Cuanto más pequeño mejor)
BIC: Criterio de Información Bayesiano (Cuanto más pequeño mejor)

En el modelo por ambiente, la cantidad de marcadores significativos fue mayor cuando la correlación genética se estimó a través de la información de similitud molecular (Figura 3.1). La cantidad de marcadores moleculares significativos cambia ambiente por ambiente. Además de la cantidad también cambia el marcador significativo entre ambientes, es decir, el grupo de marcadores identificados como significativos en un ambiente no es exactamente el mismo que en otro ambiente. Estos cambios, tanto en cantidad como en el grupo de marcadores moleculares puede deberse a la presencia de G×E. Este cambio en el número y grupo de marcadores asociados a la variación del fenotipo se relaciona a las covarianzas más altas que se obtuvieron entre las observaciones cuando se utilizó la información de parentesco o correlación entre las observaciones desde el pedigrí. Por el contrario, en los modelos multiambientales se observa que el impacto de usar K o A fue menor, debido a que la cantidad de marcadores asociados al fenotipo fue prácticamente el mismo, 41 en M3 y 48 en M4 (Figura 3.2). Cuando se ajusta un modelo por ambiente, la cantidad de marcadores significativos puede ser subestimada o sobreestimada con respecto a un modelo multiambiental (Figuras 3.1 y 3.2).

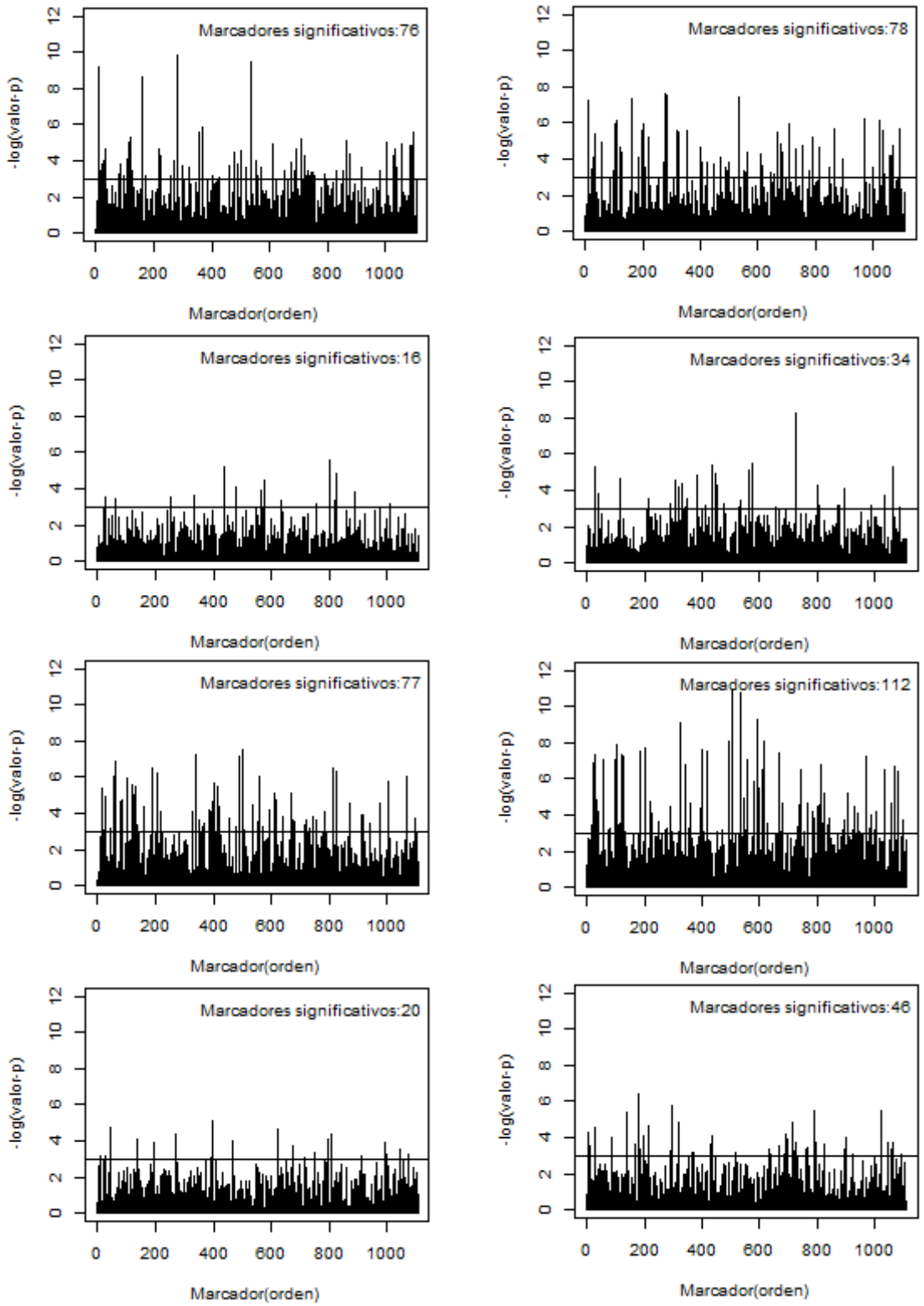

Figura 3.1. Significancia estadística (-log(valor-p)) para cada marcador molecular evaluado en los modelos GWAS, M1 (Izquierda) y M2 (Derecha). En los modelos GWAS M1 y M2, la evaluación fenotípica fue hecha en cuatro ambientes. M1: Modelo por ambiente con el pedigrí para modelar las correlaciones genéticas; M2: Modelo por ambiente con información de marcadores moleculares.

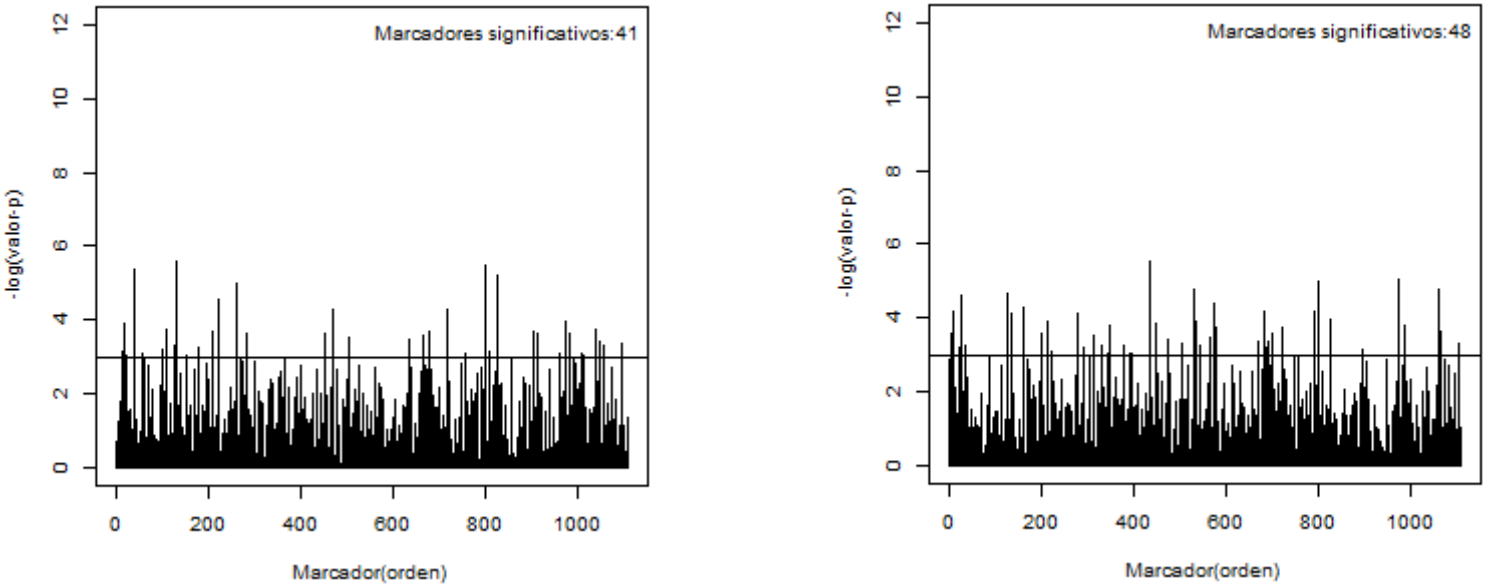

Figura 3.2. Significancia estadística (-log(valor-p)) para cada marcador molecular evaluado en los modelos GWAS, M3 (Izquierda) y M4 (Derecha). En los modelos GWAS M3 y M4, la evaluación fenotípica fue hecha en todos los ambientes. M3: Modelo multiambiental con el pedigrí para modelar las correlaciones genéticas; M4: Modelo multiambiental con información de marcadores moleculares.

Las altas correlaciones obtenidas entre los BLUPs de los modelos comparados en este capítulo, sugieren una alta similaridad entre los predictores de los efectos globales de G en los modelos por ambiente comparados con aquellos derivados de los modelos multiambientales (0,93 *vs*. 0,90, respectivamente) (Figura 3.3). Los BLUP de los efectos de genotipo en los modelos basados en el pedigrí mostraron una mayor habilidad predictiva (alta correlación, r=0,95) comparado con los modelos basados en la similitud molecular (r=0,92) (Figura 3.3).

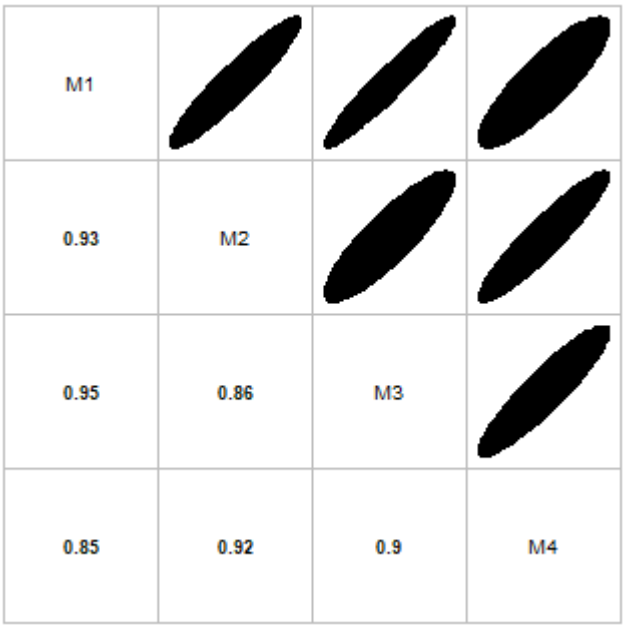

Figura 3.3. Correlación entre los BLUP de genotipo obtenidos para cuatro modelos GWAS, usados en una evaluación fenotípica multiambiental de 599 genotipos. Los modelos son: M1) Modelos por ambiente y correlación por pedigrí M2) Modelos por ambiente y correlación por similitud molecular, M3) Modelo multiambiental y correlación por pedigrí, y M4) Modelo multiambiental y correlación por similitud molecular. Todos los coeficientes son estadísticamente significativos (p<0,05).

DISCUSIÓN

En GWAS, la correlación entre genotipos ha sido contemplada según la información de pedigrí o de la información de marcadores moleculares (similaridad genética) (Burgueño *et al*., 2012; Cuevas *et al*., 2017; Urrestarazu *et al*., 2017). La similaridad genética entre individuos estimada a partir de la información molecular (VanRaden, 2007), se incorpora en los modelos estadísticos analíticos de la misma manera que las relaciones genéticas basadas en el pedigrí (Piepho *et al*., 2008), *i.e*., a través de la matriz de varianza y co-varianza de los efectos aleatorios de genotipo. Los modelos estadísticos que no asumen independencia entre líneas muestran una mayor precisión predictiva, que los modelos que no incluyen ninguna fuente de información sobre las relaciones de parentesco dentro de la población (Cappa *et al*., 2011). En los modelos mixtos infinitesimales aditivos basados en pedigrí (Henderson, 1975), las relaciones genéticas son incorporadas como una matriz de relaciones cuyos valores representan dos veces el coeficiente de relaciones genéticas entre individuos. Sin embargo, los valores esperados y observados de similaridad genética pueden diferir cuando algunas relaciones de ancestros no han sido tenidas en cuenta en la estimación del pedigrí (Burgueño *et al*., 2012). Crossa *et al*. (2010) evaluaron la habilidad predictiva de los modelos en los que se incorporó la información del pedigrí, marcadores moleculares y pedigrí más marcadores moleculares, usando las mismas líneas de trigo involucradas en este capítulo. No obstante, su estimación se realizó en los modelos por ambiente. La información sobre las correlaciones genéticas entre líneas permitió obtener una mayor precisión de predicción de un mérito genético individual. Ellos concluyeron que la adición de información genética también permitía una mejor estimación de los efectos de marcador en el fenotipo. En este capítulo se mostró que la magnitud de la variabilidad genética dependía de la información adicionada a las correlaciones genéticas del modelo. Dado que la varianza residual fue la misma, se espera una mejor repetibilidad del desempeño de G mediante el uso de marcadores moleculares en lugar del pedigrí como una medida de parentesco entre las líneas. Las diferencias entre usar pedigrí o información de los marcadores moleculares para abordar las correlaciones genéticas dependen de qué tan informativos sean los MM o el pedigrí utilizado. El objetivo de este capítulo fue comparar los modelos por ambiente *versus* los modelos multiambientales.

En los METs, los genotipos a menudo se evalúan para determinar el desempeño promedio de G, pero también para explorar la estabilidad de G×E y G (Aguate *et al*., 2019). El modelado simultáneo del desempeño de G en múltiples ambientes utiliza información correlacionada tanto genéticamente como ambientalmente (Wang *et al*., 2018). La inclusión de G×E en el modelo estadístico de factores que contribuyen a rasgos complejos puede aumentar la capacidad de detectar nuevos genes o factores ambientales que influyen en el caracter a través de una interacción, que de lo contrario puede ser indetectable cuando se ignora la interacción. La componente G×E es útil no solo por su capacidad para proporcionar información relevante sobre la variabilidad genética (Kang *et al*., 2004) sino también por su capacidad para predecir el desempeño de un G no evaluado (Burgueño *et al*., 2012). Varios tipos de matrices de varianza y co-varianza se han utilizado para obtener estas predicciones de los METs usando datos fenotípicos (Balzarini, 2002). Al estudiar G×E, la complejidad del modelo está determinada por la escala de medición del factor ambiental debido a que en los METs los ensayos se llevan a cabo en ambientes contrastantes (Ziegler *et al*., 2008). Como consecuencia, en los modelos de componentes de varianza, el efecto G×E suele ser mayor que el efecto de G y/o el efecto de E. En el contexto de GWAS, otros factores pueden explicar la variación genética que no está siendo explicada, como la calidad y precisión de los datos fenotípicos, G×E, efectos epistáticos o variación epigenética (Urrestarazu *et al*., 2017). Actualmente, se han utilizado diferentes estrategias para ajustar los modelos GWAS que contemplan la interacción G×E. Ziegler *et al*. (2008) mostraron algunos ejemplos relacionados a la posible importancia de las interacciones en GWAS. Sin embargo, expusieron que analizar G×E en GWAS es un desafío con respecto a la realización simultánea de múltiples pruebas y a la demanda computacional requerida para este tipo de análisis. A pesar de la importancia de G×E en el contexto de GWAS, hasta el momento se han realizado pocos trabajos donde se desarrollen métodos para detectar este tipo de interacciones (Murcray *et al*., 2008; Gauderman *et al*., 2017; Ferrero-Serrano y Assmann, 2019). Murcray *et al*. (2008) mostraron que en el contexto de GWAS, el uso de información de MM puede llevar a incrementos sustanciales en la capacidad para detectar un gen involucrado en G×E. Los ajustes de los modelos GWAS por ambiente demandan menos tiempo computacional, ya que el número de parámetros a estimar es menor que en los modelos GWAS multiambientales. Los resultados mostraron efectos de G altamente correlacionados con los producidos por el ajuste de los modelos por ambiente. Sin embargo,

los modelos multiambientales proporcionaron componentes de varianza separados para los efectos G y G×E. Además, permiten el préstamo de información entre ambientes correlacionados (Cabrera-Bosquet *et al*., 2012). Por lo tanto, sería factible predecir los efectos de G no evaluados en un E dado en un contexto de datos de desbalanceados.

CONCLUSIÓN

Los modelos GWAS multiambientales proveen información adicional respecto a la modelación por ambiente dado que permiten estimar la varianza de G×E. La prueba ajustada de los efectos de marcador en los modelos GWAS multiambientales produjo un número menor de marcadores significativos, esto podría considerarse como una potencial disminución de la tasa de falsos positivos cuando se ajustan modelos multiambientales y se considera la información del pedigrí para medir la correlación genética. Incorporar la información de pedigrí en los modelos GWAS permite un mejor ajuste respecto a modelos GWAS que consideran la similitud molecular como medida de correlación genética.

CAPÍTULO 4

META-ANÁLISIS DE ESTUDIOS DE SELECCIÓN GENÓMICA

INTRODUCCIÓN

La selección genómica (SG) es usada para predecir el mérito de un genotipo respecto a un caracter cuantitativo a partir de datos moleculares o genómicos. La SG es una técnica con alto potencial para acelerar la tasa de ganancia genética en vegetales (Heffner *et al*., 2009). Valiéndose de modelos estadísticos, permite relacionar vasta cantidad de marcadores moleculares (MM) o información genómica a un caracter fenotípico de interés para predecir luego el mérito genético de cada fenotipo. En SG, se parte de una población de entrenamiento o calibración donde no sólo el genotipo molecular es conocido sino también el fenotipo y se estiman modelos relacionales que, aprendiendo desde dicha población, son luego aplicados a poblaciones de líneas donde no se conoce el fenotipo pero si se desea predecir el mérito genético. Es a partir del modelo estadístico estimado o ajustado que se realizan predicciones para estimar el valor de cría o mérito genético de cada individuo en la población de interés. Así, es posible seleccionar individuos con características promisorias para un determinado caracter usando sólo la información molecular que usualmente proviene del genotipado con marcadores moleculares distribuidos en todo el genoma de cada individuo (Hawkins y Yu, 2018). La información del genotipo y del fenotipo en la población de entrenamiento también es usada para estimar el efecto de cada marcador sobre el caracter de interés. Aún con gran cantidad de MM, los modelos estadísticos usados en SG permiten ajustar el efecto de todos los marcadores simultáneamente. Si los MM se encuentran en desequilibrio de ligamiento con la mutación que afecta el caracter, los marcadores serán capaces de capturar una gran proporción de la varianza genética del caracter de interés (Voss-Fels *et al*., 2019) y el modelo estadístico permitirá asociar los MM con el fenotipo (Hawkins y Yu, 2018). El modelo aplicado sobre la población de mejora se usa para predecir el valor genético del individuo

desde la información molecular distribuida en todo el genoma (*Genome Estimated Breeding Value*-GEBV) (Bhat *et al*., 2016; Hawkins y Yu, 2018). La SG tiene la capacidad de utilizar altas cantidades de MM asociados a cada *loci* incluso de efecto menor (Heffner *et al*., 2009) y así capturar mayor variación genética. La eficiencia de los modelos de SG se evalúa a través de las correlaciones entre los fenotipos observados con los valores de mejora predichos. Hasta la fecha se cuenta con una cantidad considerable de estudios relacionados a SG en vegetales, siendo posible su recopilación a través de la revisión sistemática y posteriormente, ser analizarlos de manera conjunta a través de meta-análisis. La revisión sistemática sintetiza la información científica disponible, incrementa la validez de las conclusiones de estudios primarios e identifica áreas para futuras investigaciones (Ferreira González *et al*., 2011). La revisión sistemática involucra las siguientes acciones: (i) formular la pregunta de investigación, a partir de esta pregunta se realizará el constructo de búsqueda; (ii) realizar la búsqueda de manera exhaustiva y comprensiva de estudios primarios en diferentes bases de datos (búsqueda electrónica); (iii) compactar la información obtenida de las diferentes bases de datos a través de un gestor bibliográfico; (iv) establecer los criterios de exclusión e inclusión para la selección de estudios primarios; (v) determinar la relevancia de los estudios identificados y (vi) extracción de los datos (Pai *et al*., 2004). Después de la revisión sistemática, la información relevante de los estudios individuales es sistematizada en bases de datos con formatos apropiados para ser analizada conjuntamente con modelos estadísticos propios del meta-análisis (Akobeng, 2005; Borenstein *et al*., 2009, 2010; Sánchez-Meca, 2010). El término meta-análisis fue introducido por primera vez por (Glass, 1976), para denotar la síntesis estadística de los resultados de estudios similares. El meta-análisis es una herramienta metodológica que permite: (i) sintetizar los resultados de estudios primarios obtenidos de la revisión sistemática para incrementar la potencia, (ii) estimar el tamaño del efecto de interés, (iii) evaluar heterogeneidad entre estudios y (iv) sea en caso necesario hacer análisis por subgrupos y meta-regresiones (Borenstein *et al*., 2009). El meta-análisis es una herramienta útil de análisis ya que permite analizar de manera conjunta información relevante de diversos tipos de estudios que comparten un mismo tema de investigación. Para realizar el análisis global de los datos es necesario especificar un estadístico que mida el efecto de interés, que sea apropiado para responder a la pregunta de investigación y que sea seleccionado según la naturaleza de los datos (continuos o discretos). Generalmente, el estadístico considera la comparación entre dos grupos, es decir, un grupo

control *vs*. un grupo experimental, donde el tamaño del efecto (*e.g*., diferencia de medias, cociente de medias, cociente de chances, diferencia de riesgos) es la variable a analizar en el meta-análisis. Sin embargo, en este capítulo, el efecto que se midió a través del MA de estudios de SG es el tamaño o magnitud de la correlación entre valores fenotípicos observados y méritos genéticos predichos por los modelos de SG en las poblaciones de mejora. El gráfico *Forest Plot*, permite visualizar los resultados del MA a través de intervalos de confianza para el valor esperado del efecto de interés tanto para cada estudio primario como para el conjunto de éstos. La amplitud de estos intervalos de confianza dependerá de la precisión con que se reportan los resultados de cada estudio primario; ésta es función del tamaño muestral y de la varianza residual. El efecto global es estimado como una media ponderada de los efectos reportados en los estudios primarios y esta ponderación depende de la precisión de cada estudio. En escenarios de alta heterogeneidad entre estudios respecto al tamaño del efecto objeto de estudio, se realizan análisis por subgrupos de estudios primarios relativamente homogéneos. En los estudios primarios relacionados a SG en vegetales es frecuente extraer información referida a la cantidad de genotipos evaluados, la cantidad de marcadores moleculares obtenidos y/o el pedigrí (Wu y Hu, 2012). Esta información puede ser considerada en los modelos ajustados en el meta-análisis a través de un análisis por subgrupos. Metodológicamente, el meta-análisis aplica principios estadísticos que pueden ser considerados sencillos en estadística clásica inferencial, esto es, pruebas de hipótesis e inferencia estadística basada en muestras obtenidas desde estudios primarios publicados y/o literatura gris. Esto conlleva a que la potencia de la estimación del efecto global obtenido desde el meta-análisis se sustente en la suma de los tamaños muestrales (n) de cada uno de los estudios primarios.

Los objetivos de este capítulo fueron: i) realizar una revisión sistemática de literatura científica publicada sobre SG en vegetales, ii) identificar las principales metodologías de estimación usadas en el contexto de SG y iii) aplicar técnicas propias del meta-análisis para obtener medidas globales de la eficiencia de los modelos SG usados con mayor frecuencia en vegetales.

MATERIALES Y MÉTODOS

RECOLECCIÓN DE LA INFORMACIÓN A TRAVÉS DE LA REVISIÓN SISTEMÁTICA

La búsqueda de literatura relacionada a selección genómica en vegetales fue llevada a cabo en múltiples bases de datos electrónicas. Se usaron las siguientes bases de datos: Scopus, Science Direct, ESCOhost, JSTOR, Red de Revistas Científicas de América Latina y el Caribe, España y Portugal y SpringerLink; accesibles desde la biblioteca electrónica de Ciencia y Tecnología de Argentina (http://www.biblioteca.mincyt.gob.ar/recursos/index). De cada base de datos, se consideraron estudios publicados hasta agosto del 2018. Un total de 4.177 estudios primarios no duplicados fueron identificados usando el siguiente constructo de búsqueda: (GS or "*Genomic Selection*") and ("*Plant breeding*") and (*crops*). Luego, se seleccionaron aquellos estudios primarios que, en primera instancia, tuviesen alguna de las palabras claves presentes en el constructo de búsqueda en el título; seguidamente, se leyeron los resúmenes de los trabajos que pasaron el filtrado del título y se eligieron aquellos relacionados a la pregunta de investigación. Los estudios que pasaron la etapa anterior, fueron leídos de manera completa, es decir, todo el texto. Además, se observó en estos estudios si estaba la información necesaria para la conformación de la base de datos necesaria para el meta-análisis. Los datos extraídos de cada estudio primario fueron: (i) especie en la que se realiza la SG, (ii) tamaño de la población involucrada en la calibración del modelo de SG, (iii) caracter fenotípico que se desea predecir, (iv) cantidad de marcadores moleculares involucrados en el modelo, método estadístico usado en la estimación del modelo y la eficiencia de la SG (*prediction accuracy*), es decir, la correlación "r" entre el fenotipo observado y los méritos genéticos (*breeding values*) predichos. Sobre un total de 68 estudios primarios que cumplían con los requisitos de inclusión se conformó la base de datos para el meta-análisis con un total de 232 observaciones (Figura 4.1). La cantidad de observaciones fueron mayores a la cantidad de estudios primarios debido a que en un mismo estudio primario podía encontrarse más de un caracter evaluado, más de una especie y/o más de un método de SG.

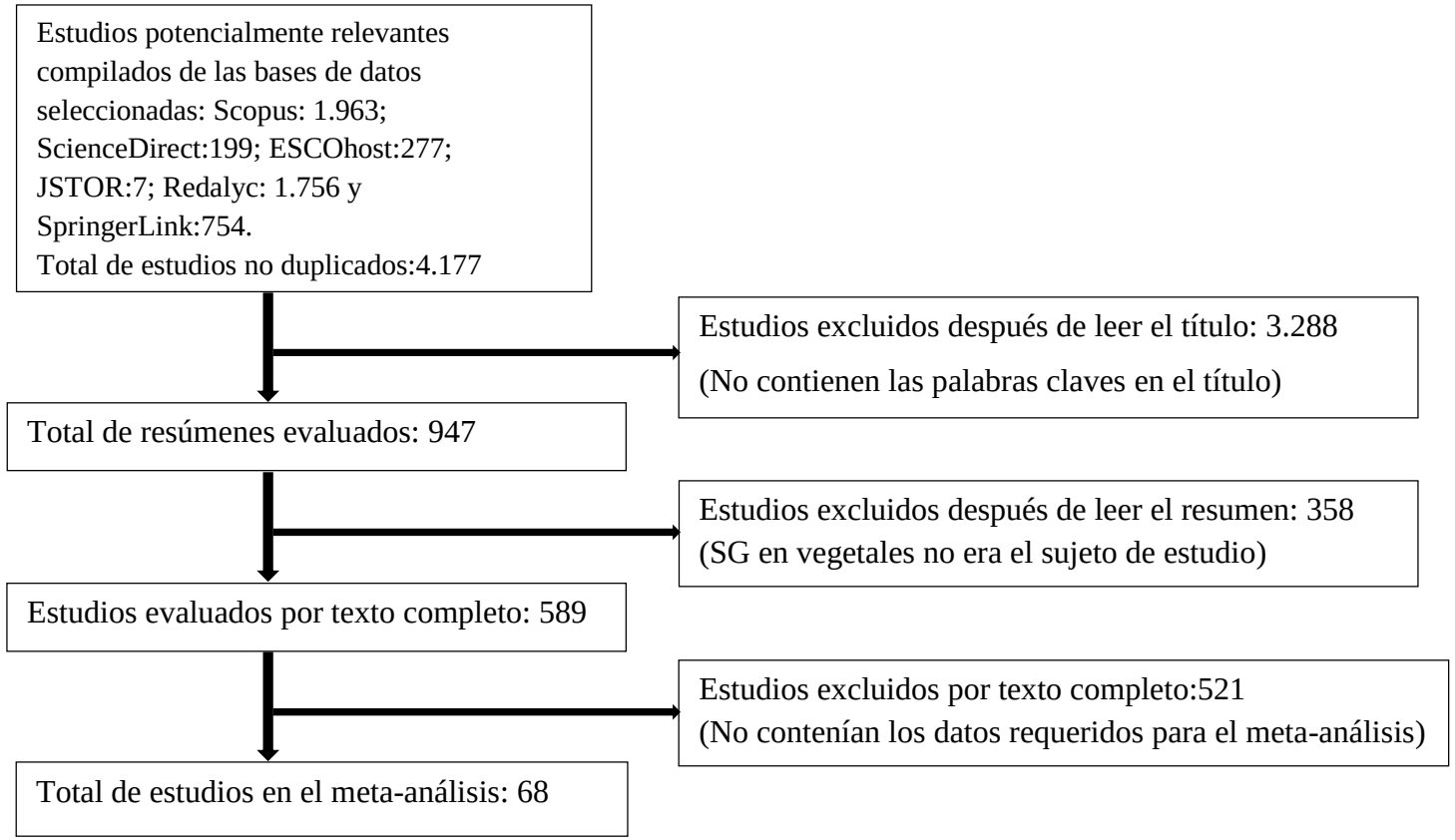

Figura 4.1. Diagrama de flujo del proceso de revisión sistemática.

CONFORMACIÓN DE LA BASE DE DATOS PARA REALIZAR EL META-ANÁLISIS

La base de datos quedó conformada por 68 estudios primarios, aunque en algunos de ellos se evaluaron varios casos, *i.e*., más de un proceso de SG y por tanto se pudieron obtener varios coeficientes de correlación, totalizando 232 casos de SG. La base de datos quedó conformada por 232 observaciones y 11 variables. La primer variable (primera columna) se denominó "Estudios", contiene los autores y el año de la publicación (estudio primario); la segunda columna se nombró "Especies" y contiene la información de las especies evaluadas; la tercera, denominada "Subgrupo_Especies" se formó con la categorización de la variable "Especies" en dos categorías: cereales (trigo, maíz, cebada, arroz y centeno) y otras especies (césped, soja, colza, remolacha, alfalfa, eucalipto, pera, raigrás, pino *Tadea*, mandioca, *Arabidopsis*, caña de azúcar y vid). La cuarta columna que se llamó "caracter", está conformada por las características fenotípicas reportadas en cada especie, es decir, rendimiento de grano, tiempo de floración, altura de planta, contenido de proteína, entre otros. La quinta columna, "n", tiene la información del tamaño de la población, es decir, la cantidad de genotipos evaluados en cada estudio primario. La sexta columna se denominó "Subgrupo_n" y fue generada a partir de la categorización de la cantidad de genotipos (variable "n"). El criterio usado para la categorización fue el percentil (P(33) y P(66)), creando tres categorías: baja≤289, media=(289;515] y alta>515 genotipos. La séptima columna, "Total_Marcadores", posee la información de la cantidad de marcadores moleculares reportados en cada estudio primario. El "Total_Marcadores" se categorizó según el percentil 33 y 66 en tres categorías: baja≤1.700, media=(1.700;17.000] y alta>17.000 de marcadores moleculares y fue considerada la octava columna de la base de datos, denominada "Subgrupo_MM". La novena columna "Método_Estimación", reporta los métodos de estimación usados en SG en vegetales. La décima columna, reagrupa los métodos de estimación a través de dos categorías: basados en BLUP (BLUP, G-BLUP y RR-BLUP) y otros métodos bayesianos o de aprendizaje de máquinas (M-BL, SVR, RKHS, Bayes A y Bayes B). La última columna tiene la correlación "r" entre el fenotipo observado y el mérito genético predicho por SG reportado en cada estudio primario. Un valor de correlación cercano a uno indica alta eficiencia de la SG.

Los métodos basados en modelos mixtos y consecuentemente en el mejor predictor lineal insesgado (BLUP) han sido los más usados en SG. El método G-BLUP se caracteriza por asignar la misma varianza a todos los *loci*, es decir otorga la misma importancia a cada alelo del marcador para obtener el predictor del mérito genético, *i.e.*, como suma los efectos alélicos individuales; algunos marcadores pueden asociarse a efectos nulos (Clark y van der Werf, 2013). Por otra parte, el método RR-BLUP asume que todos los *loci* tienen el mismo efecto, distinto de cero, con las mismas varianzas, pero esto no implica que todos los MM tengan el mismo efecto. RR-BLUP estima la matriz de relaciones entre los genotipos a partir de la información provista por los MM, por lo tanto, algunos *loci* pueden aportar al predictor y otros no. Los métodos bayesianos ponderan el efecto de cada *locus* con distinta varianza a diferencia de RR-BLUP. Entre los métodos basados en aprendizaje automáticos, SVM ha sido uno de los más usados en la SG, dado que al usar una función kernel para los cálculos como el producto interno, resuelve el problema de estimación con alta dimensionalidad. Por ello, en estos tipos de modelos, la selección de la función kernel se vuelve un factor clave, dado que la misma debe reflejar la distribución característica de la muestra de entrenamiento (Wang *et al*., 2018).

META-ANÁLISIS

Se estimó la correlación (promedio ponderado de los estudios) entre los valores observados y predichos, teniendo en cuenta todas las especies primero y, en segunda instancia, incluyendo solo los estudios de SG realizados en maíz y trigo. Estas dos especies representaron el 53% del total de observaciones de la base constituida a partir de la revisión sistemática. Se usó un modelo de efectos aleatorios para el meta-análisis, ya que se observó alta heterogeneidad entre estudios respecto a los valores de *r* y la precisión reportada:

$$r_i = \mu + \tau_i + \varepsilon_i \qquad [4.1]$$

donde, r_i es la correlación observada, μ es la correlación esperada entre el fenotipo observado y el mérito genético predicho,τ_i es un efecto aleatorio asociado a cada estudio primario que se supone con distribución $N(0,\tau)$ y ε_i es un término de error aleatorio con distribución $N(0,\sigma^2)$ que mide la precisión dentro de cada caso de SG.

La heterogeneidad entre estudios se evaluó con el estadístico I^2, que permite cuantificar cuánto de la variabilidad total en el estadístico de interés debe ser atribuida a la variación

entre estudios (Higgins *et al*., 2003). Es una medida independiente del número de estudios incluidos en el meta-análisis y de la unidad de medida utilizada para cuantificar el efecto estudiado. El estadístico I^2, se expresa como una proporción, un valor cercano a cero indica que la varianza observada es espuria y, por lo tanto, los estudios primarios pueden considerarse homogéneos. Higgins *et al*. (2003) sugirieron que valores de I^2 hasta el 25% podrían ser indicadores de baja heterogeneidad, entre 25 y 50% de mediana heterogeneidad y más de 75% de alta heterogeneidad. Dado los altos valores encontrados para I^2, se llevaron a cabo análisis por subgrupos considerando en cada análisis diferentes variables de clasificación; cantidad de marcadores moleculares utilizados ("Subgrupo_MM") y el tipo de método de estimación utilizado para la SG ("Método_Estimación") a fin de detectar como estás variables contribuyeron en la estimación global de la eficiencia de la SG. La estrategia de realizar análisis por subgrupo, además de controlar la heterogeneidad entre estudios, permitió detectar cómo estás variables contribuyen en la estimación global de la eficiencia de la SG. Los meta-análisis se realizaron con los datos transformados a través del z de Fisher, pero los resultados de los efectos globales fueron reportados en la métrica de correlaciones. Los datos fueron analizados usando el software R con el paquete meta (R Core Team, 2020).

RESULTADOS

La cantidad de observaciones resultantes de los estudios primarios para cada especie, se encuentran en la Tabla 4.1 junto a otra información que caracteriza a los estudios analizados. Se puede observar que las especies trigo y maíz fueron las que proveyeron una cantidad mayor de observaciones con respecto a otras especies. Además, se observa que para maíz y trigo se incluyen una mayor cantidad de genotipos en las poblaciones usadas para la calibración del modelo de SG. Además, en estas dos especies se ha usado una cantidad relativamente alta de marcadores moleculares (MM). La SG en cereales ha convocado mayor atención que en otras especies agrícolas por el acortamiento aparejado en el ciclo de mejoramiento genético vegetal. Probablemente, este hecho se asocie con la importancia alimentaria de estas especies agrícolas que cuentan con programas de mejoramiento genético vegetal en gran parte del mundo. En este trabajo, la mayoría de observaciones presentes en la base de datos fue de los siguientes cereales: trigo, maíz, cebada, arroz y centeno. Los resultados reportados en la Tabla 4.1 indican predominancia del uso de los métodos de estimación RR-BLUP y G-BLUP en el ajuste del modelo estadístico que permitirá obtener

las predicciones para la SG. Es importante destacar que en las especies más representadas se observó alta variabilidad en los reportes de eficiencia de la SG; *i.e.*, en trigo algunas publicaciones reportaban correlaciones menores al 20% mientras que otras reportaban valores mayores al 80%. No obstante, en la mayoría de los casos analizados la eficiencia de la SG fue cercana al 60%, que en términos estadísticos es mediana, pero en términos prácticos puede ser suficiente. En el 53% de las observaciones la eficiencia de la SG fue aproximadamente del 59% (mediana). En el Anexo II se presentan los *Forest Plot* donde se consideraron todas las especies. Debido a la alta heterogeneidad presente al considerar todas las especies (I^2=99%) y estadísticamente significativa p<0,001 al ajustar un modelo de efectos aleatorios, se llevó a cabo el análisis por subgrupos para las variables de clasificación; cantidad de genotipos evaluados ("Subgrupo_n"), cantidad de marcadores moleculares utilizados ("Subgrupo_MM") y el tipo de método de estimación utilizado para la SG ("Método_Estimación"). No se observaron diferencias significativas en la eficiencia de la SG según los criterios de clasificación como método de estimación de SG, tamaño de la población usada para su calibración y/o densidad de marcadores moleculares en el Anexo II (Figuras 4.4, 4.5 y 4.6, respectivamente).

La representación gráfica de los resultados se realiza con un *Forest Plot* (Figura 4.2) cuyas filas representan cada uno de los estudios primarios y la eficiencia de la SG, en las especies trigo y maíz agrupadas según el método de estimación del modelo de SG (G-BLUP y RR-BLUP). El tamaño del efecto es la magnitud de la asociación entre el valor fenotípico observado y el valor genético predicho. Así, el gráfico permite visualizar la correlación de interés promedio (cuadrado) de cada estudio primario y su intervalo de confianza (IC) con nivel de confianza del 95%. Mientras menor es la amplitud del intervalo de confianza, mayor es la precisión en la estimación del coeficiente de correlación entre el valor fenotípico observado y el valor genético predicho. El cuadrado que representa el tamaño del efecto de cada estudio primario varía entre estudios para reflejar el peso de cada uno en la estimación del efecto global (correlación promedio ponderada). Un estudio con precisión relativamente buena, tendrá asignada mayor ponderación o peso para generar la estimación global. La precisión está gobernada por el tamaño de la muestra y por la varianza residual del estudio. Al final de la lista de estudios, se visualiza el efecto global (rombo). Si la correlación global es estadísticamente distinta de cero, el valor de cría predicho por el modelo se correlaciona con el valor observado y la SG es eficiente. El efecto global de la correlación entre los valores

observados y los valores predichos fue de 0,61, con un intervalo de confianza (IC) de [0,59-0,64] que confirma la eficiencia de la SG. La heterogeneidad entre estudios fue alta I^2=99% y estadísticamente significativa p<0,001; como estrategia analítica para controlar parte de la heterogeneidad, se identificaron subgrupos relacionados al método de estimación (Figura 4.2) y a la cantidad de marcadores moleculares involucrados en la construcción del modelo de SG (Figura 4.3). El intervalo de confianza (IC) para G-BLUP fue [0,57-0,66] y para RR-BLUP [0,59-0,64], en ambos casos no contienen al cero e indican que la eficiencia de la SG es similar entre ellos y que no estuvo condicionada por la selección de uno u otros métodos de estimación basados en BLUP. La superposición de los IC indica que no existen diferencias estadísticamente significativas entre ambos métodos de construcción del modelo para predecir mérito genético desde la información genómica.

Los resultados del meta-análisis, para trigo y maíz, realizado por subgrupos definidos por la densidad de marcadores moleculares (alta con más de 17.000 MM, media con 1.700 hasta 17.000 MM y baja con menos de 1.700 MM), mostraron similitud de la eficiencia alcanzada con distinta cantidad de marcadores moleculares (cerca del 60%) (Figura 4.3). Las ponderaciones o pesos reportados para las categorías alta y media fueron 41,1% y 36,1%, respectivamente, respecto a las ponderaciones de la categoría de baja densidad de marcadores moleculares, que fue de 22,8%. Esto evidenció que la mayor contribución al efecto global de correlación entre los valores observados y predichos, se obtuvieron con densidades de marcadores moleculares altas y medias (Tabla 4.2).

Tabla 4.1. Mediana, mínimo y máximo de genotipos, marcadores moleculares y de correlación para cada especie evaluada en la revisión sistemática.

Especie	Observaciones	Genotipos	Marcadores moleculares	Método de estimación†	r§
Trigo	105	372 90−58.798	4.040 234−90.000	RR-BLUP	0,58 0,16−0,92
Maíz	56	300 97−3.273	16.846 125−158.281	RR-BLUP	0,59 0,24−0,90
Césped	20	515 482−515	16.669 –	RR-BLUP	0,32 -0,09−0,55
Cebada	10	647 140−691	1.536 107−3.072	RR-BLUP	0,69 0,57−0,86
Arroz	8	343 343−413	5604 4011−73.147	BL y RF	0,33 0,30−0,63
Soja	5	301 288−301	52.349 79−52.349	G-BLUP	0,65 0,42−0,69
Colza	4	391 –	253 –	Bayes B y RR-BLUP	0,41 0,34−0,84
Centeno	4	219 219−220	584 –	RR-BLUP	0,77 0,70−0,82
Remolacha	4	310 310−924	384 384−677	BLUP y RR-BLUP	0,80 0,48−0,86
Alfalfa	3	154 124−190	68.972 10.000−77.610	SVR	0,35 0,32−0,51
Eucalipto	3	768 –	24.806 –	G-BLUP	0,55 0,54−0,63
Pera	2	76 –	162 –	Bayes A y Bayes B	0,71 0,71−0,75
Raigrás	2	211 –	10.885 –	RR-BLUP	0,16 0,16−0,53
Pino *Taeda*	2	711 –	4.825 –	G-BLUP	0,74 0,74−0,75
Mandioca	1	–	–	BLUP	–
Arabidopsis	1	–	–	BLUP	–
Caña de azúcar	1	–	–	BL	–
Vid	1	–	–	RR-BLUP	–

† Método de estimación más frecuentemente usado en la construcción del modelo de selección genómica (SG)
§Correlación entre el fenotipo observado y los valores de mejora (*breeding values*) genéticos predichos por el modelo de SG.

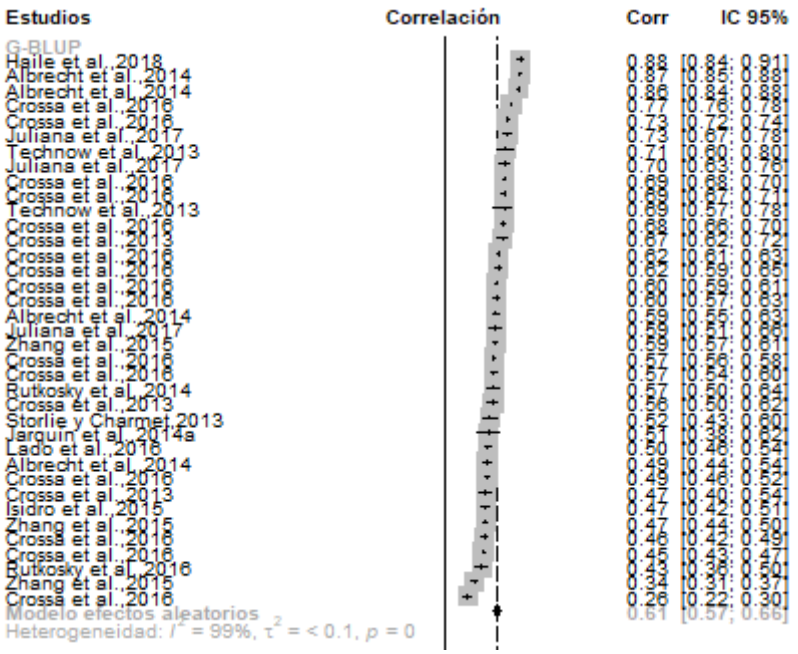

Figura 4.2. *Forest Plot* de la eficiencia de SG para los métodos de estimación G-BLUP y RR-BLUP en trigo y maíz. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos (G-BLUP y RR-BLUP), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada método de estimación.

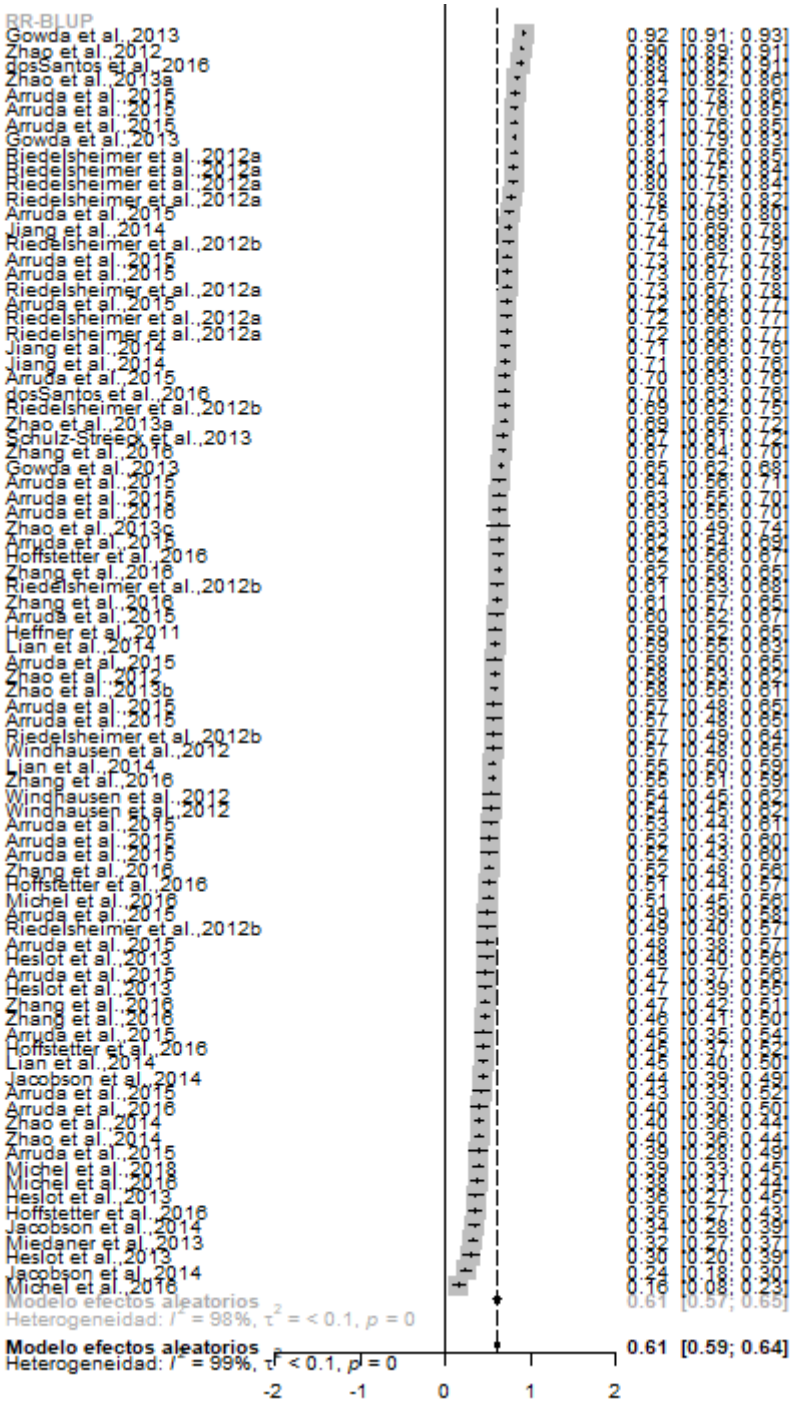

Figura 4.2. *Forest Plot* de la eficiencia de SG para los métodos de estimación G-BLUP y RR-BLUP en trigo y maíz. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos (G-BLUP y RR-BLUP), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada método de estimación. Continuación.

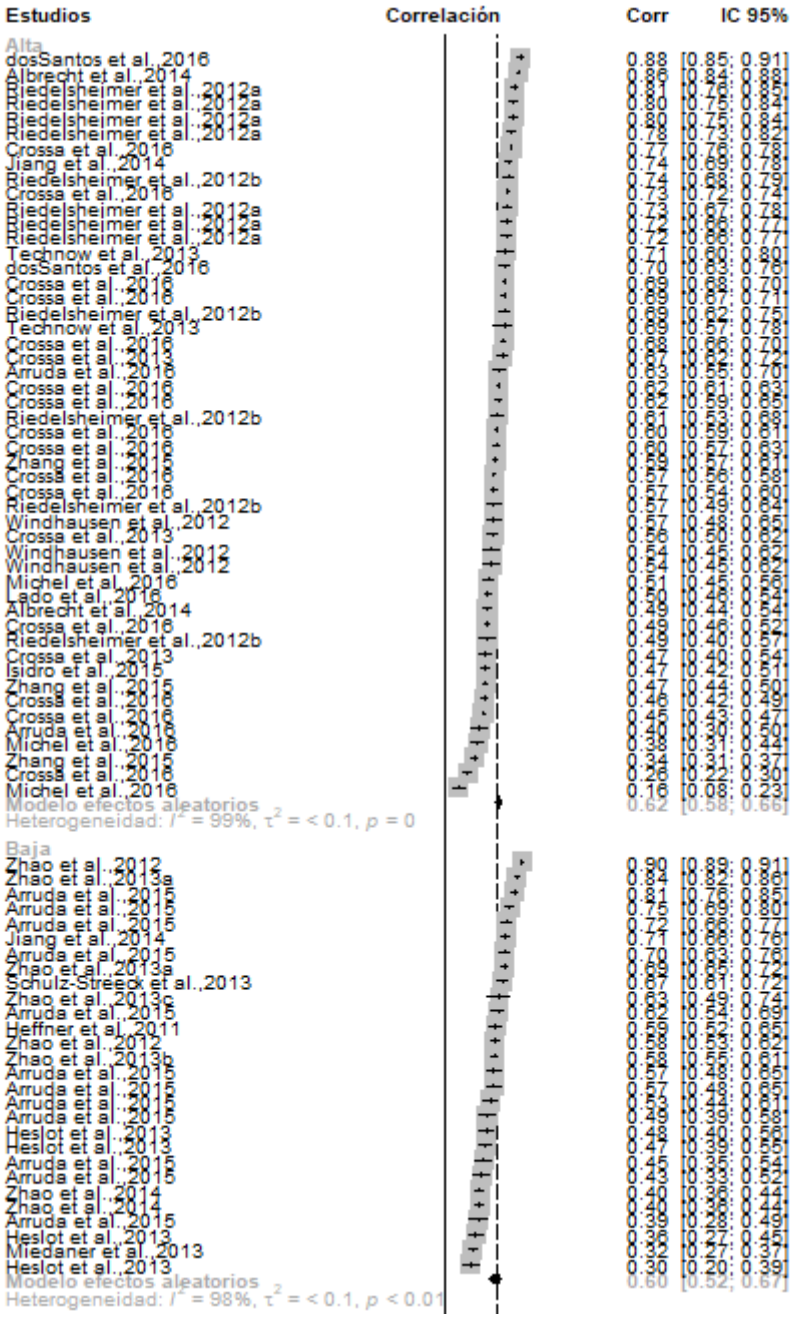

Figura 4.3. *Forest Plot* de la eficiencia de SG para distintas densidades de marcadores moleculares categorizadas en: baja (menos de 1.700), media (entre 1.700 y 17.000) y alta densidad de marcadores moleculares, mayor a 17.000 para estudios primarios de trigo y maíz. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de densidad de marcadores moleculares (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares.

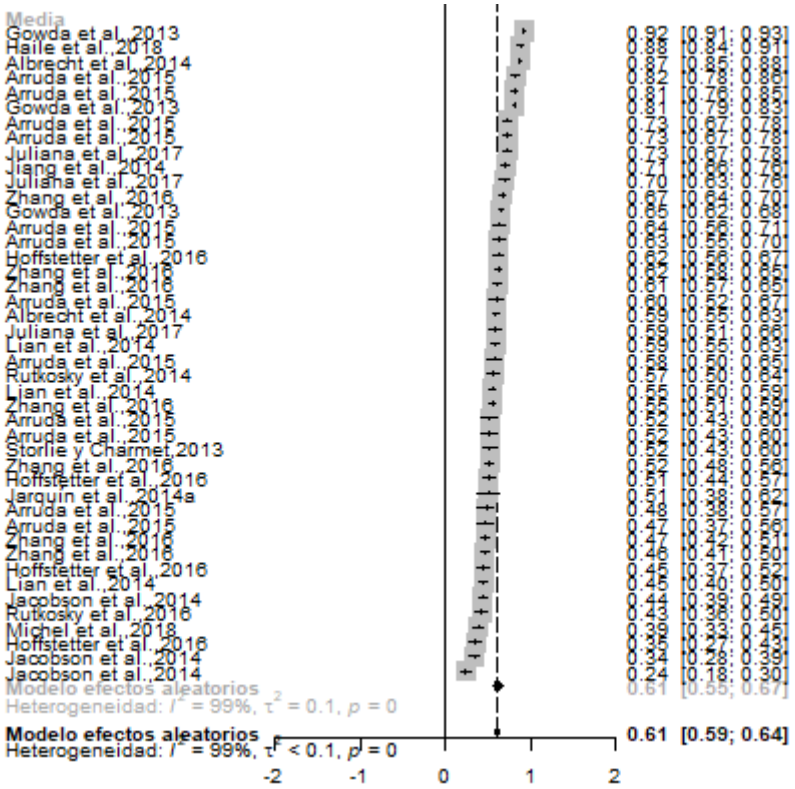

Figura 4.3. *Forest Plot* de la eficiencia de SG para distintas densidades de marcadores moleculares categorizadas en: baja (menos de 1.700), media (entre 1.700 y 17.000) y alta densidad de marcadores moleculares, mayor a 17.000 para estudios primarios de trigo y maíz. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de densidad de marcadores moleculares (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares. Continuación.

Tabla 4.2. Estimación de la correlación entre fenotipo observado y mérito genético predicho desde la información molecular de los métodos de estimación más frecuentes de SG y densidad de marcadores moleculares en maíz y trigo (n=122).

	r[¶]	LI[§]	LS[§]	Ponderación (%)
Método de Estimación de SG				
RR-BLUP	0,61	0,59	0,64	69,4
G-BLUP	0,61	0,57	0,66	30,6
Efecto Global	0,61	0,59	0,64	100
Cantidad de Marcadores Moleculares				
Mayor a 17.000	0,62	0,58	0,66	41,1
Entre 1.700 y 17.000	0,61	0,55	0,67	36,1
Menor o igual a 1.700	0,60	0,52	0,67	22,8
Efecto Global	0,61	0,59	0,64	100

[§] LI y LS son el límite inferior y superior del intervalo de confianza del 95% de la correlación.

[¶] Correlación entre fenotipo observado y mérito genético predicho desde la información molecular.

[ǂ] Pesos resultantes de un modelo de efectos aleatorios para cada uno de los subgrupos respecto a su efecto global.

DISCUSIÓN

Los modelos de SG se comenzaron a aplicar principalmente para la predicción genética en animales (Meuwissen *et al*., 2001); no obstante, rápidamente, su aplicación fue adaptada a modelos de SG para mejoramiento genético en plantas (Thavamanikumar *et al*., 2015). La SG permite predecir de manera efectiva los valores de cría o mérito genético de los individuos a partir de información genómica. Diferentes investigadores a menudo ajustan distintos modelos en sus estudios, sin embargo, han manifestado i) la complejidad a la hora de comparar los resultados de diferentes estudios, y ii) que no siempre se ajusta el modelo más eficiente para un estudio en particular (Wang *et al*., 2015). Una de estas problemáticas puede abordarse desde el contexto del meta-análisis, ya que esta herramienta permite comparar de manera simultánea resultados de diversos estudios relacionados a SG. En el presente trabajo se realizaron estimaciones con un modelo de efectos aleatorios usando las correlaciones puras (sin transformar) y las correlaciones transformadas con el z de Fisher, de los estudios primarios obtenidos mediante la revisión sistemática de literatura científica. Los resultados fueron expresados en la métrica de correlación desde los datos transformados a través del z de Fisher. Se observaron diferencias notables en la correlación global al usar las correlaciones transformadas como en Schmidt *et al*. (1980). A pesar de que algunos autores sugieren que la transformación z de Fisher produce un sesgo hacia la derecha cuando se usa para promediar correlaciones (Silver y Dunlap, 1987; Strube, 1988; Field, 2001, 2005), otros autores sugieren que dicha transformación cumple con su propósito original, el de crear una transformación de la correlación para la cual el error estándar y, por lo tanto, los intervalos de confianza, dependan únicamente del tamaño muestral y no del tamaño de la correlación observada, que se ve afectada por el error de muestreo (Silver y Dunlap, 1987; Field, 2005). Otro autor, señaló que a medida que aumentaba el número de estudios primarios, el sesgo no era significativo al usar las correlaciones transformadas con el z de Fisher o sin transformar (Strube, 1988). En este capítulo, el número de observaciones o total de correlaciones derivadas de los estudios primarios es alto. En el meta-análisis realizado en este trabajo para maíz y trigo, el uso de la transformación z de Fisher produjo intervalos de confianza con menor amplitud (0,59-0,64 *vs*. 0,56-0.62) y mayor efecto global que al considerarse las correlaciones sin transformar (0.61 *vs*. 0.59, respectivamente). No obstante, en cualquiera de las dos escalas, la correlación entre los valores observados y los predichos

por el modelo de SG, dista de ser óptima aunque sea estadísticamente significativa. Futuros desarrollos de estos modelos son necesarios para mejorar la capacidad predictiva del mérito genético a través del uso de infomación genómica. Podría ser importante considerar efectos de interacción entre los efectos de marcadores en los modelos lineales usados para SG con el fin de incrementar la eficiencia de la SG. Conceptualmente estos modelos son entendibles pero computacionalmente difíciles de estimar en la actualidad. El aprendizaje automático podría ser una solución para el tratamiento de interacciones de múltiple orden. Varios métodos de SG abordan los problemas del alta dimensionalidad (mayor cantidad de columnas que de filas) y la complejidad computacional y, por lo tanto, capturan diferentes aspectos de la asociación entre el genotipo y el fenotipo. Sin embargo, el desempeño de diferentes métodos depende de la arquitectura genética subyacente del caracter de interés (Resende *et al*., 2012; Wang *et al*., 2018).

Se ha discutido en la literatura que la precisión de la SG depende de la arquitectura genética de los caracteres, tales como la heredabilidad y la distribución de los genes causales (Desta y Ortiz, 2014), estando la heredabilidad relacionada positivamente a la precisión de la predicción (*r*). No obstante, existe otro factor importante que debe ser considerado, este es, G×E; ya que precisiones sobre el mismo caracter pero evaluados en varios ambientes difieren considerablemente debido a esta interacción (Wang *et al*., 2018). Su *et al*. (2012) usaron predicción genómica (SG) en el contexto de ganado vacuno, con una densidad media (54K) y alta (777K) de SNPs. Ellos observaron que cuando aumentaron la densidad de los marcadores de 54K a 777K, la precisión de la predicción aumentó solo en 0,5-1,0%. Aún cuando se conoce que cuanto más marcadores haya mejor será la predicción, la precisión es difícil de mejorar significativamente cuando la densidad inicial de marcadores ya es alta. El comportamiento observado en este trabajo podría deberse a este suceso, ya que el efecto global estimado para las categorías alta y media de densidad de MM fue muy similar (0,62; 0,61, respectivamente). No obstante, se observó en algunos estudios, para caracteres relacionados a calidad de grano en poblaciones biparentales de trigo, que la precisión de los modelos de SG alcanzó una meseta (se estabilizó, alcanza un cierto nivel donde no varía) a bajas densidades de marcadores moleculares (128-256 MM) (Heffner *et al*., 2011). Para otros tipos de especies, como el *Pinus tadea* L. se presentó en algunos casos, que la máxima capacidad predictiva se alcanzó con pequeños subconjuntos de MM (110-590 marcadores) y disminuyó cuando se adicionaron más MM (Resende *et al*., 2012). El mismo fenomeno se

observó en otro estudio en el que se usó una densidad de 3.490 MM para calibrar los modelos de SG (Oakey *et al*., 2016). Lo mismo se observó en otros estudios (Yang *et al*., 2010; Wang *et al*., 2016) cuando la densidad de MM alcanzó un cierto nivel, la predicción genómica no se vio beneficiada.

Entre los métodos de SG, el más recomendado en la práctica de mejoramiento debido a su robustez y eficiencia computacional es el G-BLUP (Wang *et al*., 2018). No obstante, algunos algoritmos de aprendizaje de máquinas, como, RKHS (*reproducing kernel Hilbert space*) han tomado un papel muy importante en SG (Cuevas *et al*., 2016). El método RR-BLUP, asume que todos los efectos de marcador se distribuyen normalmente y que tales efectos tienen igual varianza (Meuwissen *et al*., 2001), es similar al método Bayes C (Ferrão *et al*., 2017). También en otro trabajo, se aprecia que el método de G-BLUP es similar al BRR (*Bayesian Ridge Regression*) (Gianola, 2013).

En un trabajo donde calibraron modelos de SG basados en G-BLUP y en un método de regresión bayesiana para predecir la resistencia a la roya en trigo (Daetwyler *et al*., 2014), se observó un mejor desempeño del modelo G-BLUP con respecto al modelo de regresión bayesiana. Se ha observado una substancial ganancia genética al aplicar SG en maíz y que dichas ganancias son superiores que las obtenidas al usar selección asistida por marcadores (MAS) (Bernardo y Yu, 2007; Heffner *et al*., 2010). Igual resultado fue observado en papa (*Solanum tuberosum* L.) (Slater *et al*., 2016). En trigo, los modelos SG bayesianos han arrojado resultados prometedores en el contexto de QTL, ya que estos fueron más sensibles al aumento del número de QTL sin disminuir la precisión en contextos de múltiples QTL (Wang *et al*., 2015). La precisión aportada por los métodos G-BLUP y RR-BLUP a menudo se mantiene casi constante, independientemente de la cantidad de QTL (Wang *et al*., 2018). Éstos, son más robustos para trabajar con caracteres fenotípicos controlados por una gran cantidad de genes de efecto menor. Se encontró que el modelo Bayes A era ampliamente adaptable debido a la ponderación modelada que realiza sobre los efectos de los MM (Wang *et al*., 2015). Sin embargo, la estimación de los métodos bayesianos suele llevar mucho tiempo, lo que restringe su aplicación. Friedman *et al*. (2010) muestran una mejor alternativa para este inconveniente, con el algoritmo LASSO, éste logra un equilibrio entre la contracción selectiva (permite graduar/calibrar el parámetro de encogimiento) y la eficiencia computacional. En ganado Nellore mostraron que los modelos de regresión bayesiana eran más precisos que G-BLUP (Neves *et al*., 2014). Sin embargo, Xu *et al*. (2017) señalaron

que el modelo G-BLUP tuvo mejor desempeño que los modelos bayesianos en la predicción de caracteres relacionados con el rendimiento en maíz, y también, en la predicción del rendimiento de grano en trigo. En el presente capítulo de tesis, se observó que la eficiencia de la SG no se ve impactada por la selección de los métodos de estimación RR-BLUP y G-BLUP evidenciando nuevamente que dichos métodos son similares para SG en cereales, resultado consistente con lo publicado por otros autores (Dong *et al*., 2016; Ferrão *et al*., 2017). La revisión sistemática y el meta-análisis realizado en el presente trabajo han permitido reforzar las conclusiones de estudios individuales e identificar incertidumbres en los hallazgos relacionados a SG.

CONCLUSIÓN

Si bien hasta el momento, los modelos estadísticos usados en SG para cultivos de importancia agrícola, han contribuido en mejorar la capacidad predictiva y de esta manera obtener una mayor eficiencia a la hora de seleccionar genotipos promisorios en etapas tempranas de los programas de mejoramiento genético vegetal, aún existe espacio para optimizar dichos modelos y consecuentemente su capacidad predictiva para una mayor eficiencia de selección basada en datos genómicos masivos. Los modelos más usados en maíz y trigo para predecir mérito genético al presente han sido G-BLUP y RR-BLUP. Sin diferencia significativa entre la eficiencia de ambos métodos de calibración de SG. Si bien, la capacidad predictiva aumenta con el número de MM usados, esta respuesta no es lineal y en algunas situaciones el incremento de MM no es redituable en términos de la eficiencia de la SG.

CAPÍTULO 5

IDENTIFICACIÓN DE *LOCI* DE EFECTO MAYOR: UN PROTOCOLO BASADO EN META-ÁNALISIS

INTRODUCCIÓN

El meta-análisis permite combinar la evidencia de estudios individuales entorno a una pregunta de investigación de manera tal que sea posible analizarlos conjuntamente a través de técnicas estadísticas (Borenstein *et al*., 2009). La cantidad de información involucrada en un meta-análisis está compuesta por la suma de las observaciones de cada uno de los estudios primarios que serán analizadas conjuntamente (Borenstein *et al*., 2009, 2010). En general, las bases de datos se conforman a partir de información recopilada desde estudios primarios obtenidos a través de una revisión sistemática electrónica. La revisión sistemática implica una búsqueda electrónica y exhaustiva de todos los estudios primarios (investigaciones científicas) potencialmente relevantes para un tema de investigación; considerando de manera global, evidencia de estudios primarios entorno a una pregunta de investigación (Sargeant *et al*., 2006). La pregunta de investigación, proveerá las pautas para establecer los criterios de inclusión y/o exclusión de los estudios primarios disponibles en los repositorios electrónicos que pueden contener tanto literatura científica publicada como literatura gris. Es importante destacar que la pregunta de investigación es fundamental para la elaboración del constructo de búsqueda que debe especificar el tipo de estudio sobre los que se colectará la información, el tipo población, de intervención realizada y los resultados sobres los cuales se realizará el meta-análisis. Diferentes constructos de búsqueda, pueden conducir a resultados disímiles para una misma pregunta de investigación. Una comparación, en el contexto de medicina, entre los resultados arrojados por los constructos de búsqueda PICO (*Population-Intervention-Comparison-Outcome*) y PICOS (*Population-Intervention-Comparison-Outcome-Study*) fue realizada por Methley *et al*. (2014). En dicha comparación, encontraron mayor cantidad de estudios primarios recolectados usando PICO como

constructo de búsqueda, respecto a PICOS. Sin embargo, al utilizar el constructo de búsqueda que incluía los estudios, *i.e.*, PICOS en lugar de PICO, la cantidad de estudios irrelevantes disminuyó. Otros estudios, en el área de las Ciencias Agrícolas, con fines de detectar QTL de efecto mayor para la resistencia/ tolerancia a enfermedades virósicas y fúngicas en maíz, Rossi *et al*. (2018) utilizaron para la recolección de información el constructo de búsqueda PIO, *i.e.*, no tuvieron en cuenta para el constructo de búsqueda el tipo de diseño de experimento, que podría representar el tipo de estudio (*Study*) en el acrónimo del constructo, ni los casos (*Cases* o *Comparison*), como por ejemplo tipo de población. Pai *et al*. (2004) presentaron una ilustración paso a paso de cómo realizar revisiones sistemáticas y meta-análisis en el contexto de la medicina. Ellos mostraron algunos aspectos claves a tener en cuenta al realizar una revisión sistemática. Estos aspectos incluyen: i) formulación de la pregunta de investigación, ii) búsqueda comprensiva y exhaustiva, e inclusión de estudios primarios relevantes, iii) evaluación de la calidad de los estudios incluidos y extracción de los datos para realizar el meta-análisis. Finalizada la etapa de la construcción de la base de datos a partir de los estudios primarios obtenidos del proceso de selección resultante de la revisión sistemática, el meta-análisis consiste en una serie de metodologías estadísticas aplicables a dichas tablas de datos (Borenstein *et al*., 2009, 2010). Si bien, el meta-análisis surgió desde las áreas de las ciencias médicas, sociales y del comportamiento (Hedges y Olkin, 1985; Olkin, 1995), también se ha implementado en áreas relacionadas a las Ciencias Agropecuarias (Miguez y Bollero, 2005; Rotundo y Westgate, 2009; Pittelkow *et al*., 2015). Goffinet y Gerber (2000) fueron pioneros al utilizar el meta-análisis como una herramienta para seleccionar el mejor modelo ajustado sobre diferentes estudios primarios que buscaban identificar QTL relacionados a un mismo caracter y mapeados sobre el mismo grupo de ligamiento en el cultivo de maíz. Wisser *et al*. (2006) extrajeron la información de 50 estudios primarios sobre resistencia a enfermedades en maíz provenientes de distintos germoplasmas de este cultivo a partir de los cuales construyeron un mapa de *loci* de resistencia a enfermedades para estudiar su distribución. Ellos usaron este mapa con el fin de identificar regiones que contengan *loci* para la resistencia a enfermedades en maíz y conglomerar características similares presentes en los diez cromosomas (genoma completo del maíz). Otros autores presentaron meta-análisis sobre la búsqueda de QTL asociados con tolerancia a estrés abiótico en cebada (Li *et al*., 2013), meta-análisis de QTL asociados con la resistencia al tizón de la espiga en trigo causado por

Fusarium (Liu *et al*., 2009) y meta-análisis de QTL relacionados a la arquitectura genética de la raíz del arroz en condiciones de sequía (Courtois *et al*., 2009). Wu *et al*. (2011) realizaron meta-análisis para estudios de asociación y estudios de mapeo de QTL usando modelos paramétricos y no paramétricos en ganado lechero. Wu y Hu (2012) resaltaron como un desafío el combinar resultados de mapeo de QTL a través de varios estudios, ya que existen una serie de aspectos a considerar que difieren de un estudio a otro; tales como, densidad de marcadores moleculares, grupos de ligamientos, tamaños muestrales, tipo de poblaciones desde las cuales se realizaron los cruzamientos parentales, diseños de experimentos y métodos estadísticos.

Una técnica muy usada para el estudio de *loci* de efecto mayor es la denominada meta-QTL. Un paquete llamado MetaQTL fue presentado por Veyrieras *et al*. (2007), este consiste en la aplicación de métodos computacionales para la integración de múltiples experimentos de mapeo de QTL independientes. Para realizar un meta-QTL es necesario contar con la información de poblaciones biparentales y mapas de referencias. El mapa de consenso debe contar con alta densidad de marcadores moleculares (Hong *et al*., 2010; Shirasawa *et al*., 2013) y los QTL deben ser independientes para el mismo caracter identificado desde diferente información genética y ambientes (Goffinet y Gerber, 2000). El consenso de QTL obtenidos al realizar un meta-análisis basado en la mayoría de QTL relacionados a un caracter de interés en un intervalo de confianza del 95% son llamados meta-QTL (Lu *et al*., 2018). La técnica meta-QTL ha sido aplicada para caracteres relacionados a tiempo de floración en trigo de invierno (Griffiths *et al*., 2009), rendimiento en arroz (Wu *et al*., 2016; Carrijo *et al*., 2017), resistencia a enfermedades en maíz (Xiang *et al*., 2012; Wang *et al*., 2016) y calidad de semilla en soja (Qi *et al*., 2011). Guo *et al*. (2018) usaron la técnica meta-QTL para el análisis y la identificación de genes candidatos relacionados a caracteres de la raíz en maíz. En Chen *et al*. (2017) recolectaron información de QTL relacionados a características de rendimiento en maíz, estos fueron analizados con la técnica de meta-análisis para obtener meta-QTL a través de todo el genoma del maíz. Li *et al*. (2019) realizaron tanto la identificación de QTL como el análisis de efectos epistáticos de caracteres relacionados a el tamaño y peso de la semilla de maíz usando la técnica meta-QTL. A pesar de que el meta-QTL es una técnica muy utilizada, el genoma de referencia y la posición física de varios genes/QTL podrían no estar disponibles para la especie que se pretende

estudiar (Ali y Yan, 2012). Estos requisitos resultan una limitante para el uso de meta-QTL en todas las especies.

En la última década ha habido una gran contribución de estudios relacionados a experimentos de mapeo de QTL en animales y plantas, existe una gran variabilidad en éstos estudios de QTL en cuanto a las poblaciones de referencia sobre las cuales fueron reportados, los diseños de experimentos sobre los cuales fueron conducidos y los tamaños muestrales usados. Estas diferencias entre los estudios primarios generan cierta incertidumbre al momento de comparar los QTL informados en cada uno de ellos. El meta-análisis, surge como una herramienta que permite integrar la información relevante de múltiples estudios de genes candidatos y de experimentos de mapeo de QTL para ser analizados de manera consensuada. Etzel y Guerra (2002) desarrollaron un enfoque en el que aplicaron herramientas de meta-análisis con el fin de superar la heterogeneidad entre estudios y así mejorar aspectos como la ubicación de QTL y la magnitud de los efectos genéticos. Los estudios primarios pueden proporcionar imágenes variadas del gen candidato o QTL, en algunos estudios de QTL, los tamaños muestrales han sido tan escasos que no fue posible detectar QTL de efecto menor incluso cuando éstos existían (Ball, 2005). La combinación de resultados de varios estudios puede llevar a una conclusión más consistente y más fuerte en relación con los estudios individuales (Goffinet y Gerber, 2000; Wang *et al*., 2016), logrando una mayor potencia estadística para la detección de QTL y estimaciones más precisas de sus efectos genéticos. Así, el meta-análisis permite una comparación más objetiva de la evidencia, resolviendo la discrepancia existente entre los estudios primarios. Además, en el meta-análisis se puede estimar la variabilidad entre estudios, identificar características de los estudios primarios asociados con QTL particulares y lograr una mayor comprensión de la arquitectura genética de caracteres complejos (Salih y Adelson, 2009). Diferentes métodos han sido usados para identificar QTL y genes relacionados a resistencia a enfermedades en maíz tales como mapeo de ligamiento en poblaciones biparentales y mapeo de asociación en poblaciones diversas (Warburton *et al*., 2015). Wang *et al*. (2016) realizaron un meta-análisis para QTL de resistencia a enfermedades virales en maíz y para otros caracteres fenotípicos. La presencia de heterogeneidad entre los estudios primarios podría deberse a la diversa información presente en estos hallazgos, es decir, diferentes poblaciones, cantidades de genotipos, ambientes, fenotipos y marcadores moleculares, entre otros.

El objetivo de este capítulo es especificar paso a paso las acciones a seguir para implementar un meta-análisis orientado a identificar *loci* de efectos mayores en estudios de QTL. Se presenta una propuesta metodológica (protocolo) de análisis para estimar el efecto global de *loci* de efecto mayor en estudios de asociación orientados a identificar QTL. El protocolo delinea las acciones a seguir en dos partes, revisión sistemática y meta-análisis. La estrategia se ilustra con un ejemplo orientado a identificar QTL de efectos principales asociados a la tolerancia del maíz (*Zea mays* L.) frente a enfermedades transmitidas por virus.

MATERIALES Y MÉTODOS

PROTOCOLO PROPUESTO

PARTE 1: REVISIÓN SISTEMÁTICA

Paso I. Construcción de la pregunta de investigación (PICOS) y del constructo de búsqueda

La pregunta de investigación debe ser elaborada de tal manera que su estructura sea concisa, clara y bien definida. La pregunta debe contemplar especificaciones relacionadas a los siguientes términos: Población (P), Intervención (I), Casos (C), Resultado (O, *Outcome*) y Estudios (S, *Studies*).

Paso II. Buscar publicaciones/estudios primarios en diferentes bases de datos y usar el gestor bibliográfico para descargar información masivamente

El medio más eficiente de búsqueda de información para identificar estudios potencialmente relevantes es mediante el uso de bases de datos bibliográficas electrónicas. Existe una cantidad considerable de bases de datos, estás se encuentran en repositorios digitales y/o en bibliotecas virtuales. La selección de la base de datos, depende de las disciplinas temáticas. Para organizar los estudios primarios identificados en la revisión sistemática, se debe usar un gestor bibliográfico en el que sea posible: recopilar, seleccionar, compactar duplicados y por último exportar lista de títulos, autores y años de publicación.

Paso III. Seleccionar/excluir estudios primarios

Los criterios para seleccionar o descartar los estudios primarios que serán sometidos a meta-análisis deben ser explícitos. La primera exclusión se realiza por el título del estudio primario. Luego, en aquellos estudios primarios no excluidos por título se revisa el resumen, si no contiene información relevante para el análisis también son excluidos. Sobre los estudios restantes, *i.e.*, después de excluir por título y resumen, se lee el texto completo, si los mismos no contienen resultados que luego serán sometidos al meta-análisis, también son descartados.

PARTE 2: META-ANÁLISIS

Paso IV. Recolección de datos (construcción de la base de datos)

De cada estudio primario, las siguientes cantidades deberán ser extraídas por cromosoma: 1) la cantidad de QTL reportados en el cromosoma que está siendo estudiado (Ne) y en los otros cromosomas (Nc), 2) la cantidad de QTL de efecto mayor (R^2 del efecto aditivo de QTL mayor al percentil 75 de los R^2 de efectos aditivos del conjunto de QTL reportado en el estudio). Esta cantidad debe ser reportada tanto para el cromosoma que está siendo estudiado (Ee) como para el conjunto conformado por el resto de los cromosomas (Ec).

Paso V. Calcular el promedio ponderado de la diferencia de riesgos en un *Forest Plot*

La diferencia de riesgo es el estadístico seleccionado para responder a la pregunta ¿Existe en el cromosoma *loci* de efectos mayores para el caracter estudiado? La diferencia de riesgos estima la diferencia entre la probabilidad de tener un QTL de efecto mayor en el cromosoma estudiado (Grupo Experimental) y la probabilidad de tener un QTL de efecto mayor en los otros cromosomas (Grupo Control). Si el intervalo de confianza que contiene el efecto global estimado por el modelo seleccionado, para dicho cromosoma, no contiene al cero, entonces se concluye que existen diferencias estadísticamente significativas entre el grupo experimental y el grupo control. La diferencia de riesgos se estimó bajo un modelo de efectos aleatorios para cada uno de los cromosomas ya que se esperaba alta heterogeneidad entre los diferentes estudios de QTL. Esta heterogeneidad podría deberse al tipo de población de mapeo usada, por ejemplo, poblaciones biparentales o poblaciones conformadas por colecciones de líneas diversas. Para cuantificar la heterogeneidad entre estudios, se utilizó el estadístico I^2 propuesto por Higgins y Thompson (2002). Este estadístico no es sensible a la métrica del tamaño del efecto (en este caso es la diferencia de riesgos) y tampoco es sensible al número de estudios que están siendo evaluados en el meta-análisis (Borenstein *et al*., 2009). Para determinar la significancia estadística de la heterogeneidad entre estudios se considera que cuando el I^2 es menor al 25% la heterogeneidad entre estudios es baja; si el I^2 tiene valores cercanos al 50% entonces la heterogeneidad es considerada como media y si I^2 es mayor al 75% la heterogeneidad es alta. Para presentar los resultados obtenidos de promediar la diferencia de riesgos de estudios individuales bajo un modelo de efectos

aleatorios, se propuso construir un *Forest Plot* para cada cromosoma. Los datos fueron analizados usando el software R con el paquete meta (R Core Team, 2020).

RESULTADOS

Paso I. Construcción de la pregunta de investigación (PICOS) y del constructo de búsqueda

Para el caso en estudio, el constructo fue escrito de la siguiente manera: ("*Zea mays*" OR *maize* OR *corn*) AND ((*tolerance* OR *resistance*) AND "*virus disease*") AND (QTL OR *loci* OR "*Quantitative trait loci*"). Un total de 624 estudios sobre QTL relacionados a enfermedades virales en maíz fueron identificados. De éstos, 349 (56%) correspondían a mapeo de QTL de poblaciones experimentales obtenidas de cruces biparentales y el resto de poblaciones conformadas por líneas diversas.

Paso II. Buscar publicaciones/estudios primarios en diferentes bases de datos y usar el gestor bibliográfico para descargar información masivamente

La búsqueda electrónica se realizó en las siguientes bases de datos: cabo usando Scopus, Science Direct, ESCOhost, SciELO, Agrícola, JSTOR, y Red de Revistas Científicas de América Latina y el Caribe, España y Portugal "Redalyc". En este paso fue usado el gestor bibliográfico Zotero (https://www.zotero.org/) debido a su libre acceso y habilidad de detectar estudios duplicados. Con el constructo de búsqueda usado en la revisión sistemática se obtuvieron para cada base de datos distinto número de estudios primarios que se describen a continuación. Para Scopus se hallaron 290 estudios primarios, para Science Direct 201, Redalyc reportó 126 estudios primarios y Ebscohost 9. Las bases de datos Agrícola y JSTOR reportaron 6 estudios primarios cada una, Pubmed 5 y Scielo 2. La búsqueda de información y descarga masiva fue llevada a cabo por dos personas de manera independiente.

Paso III. Seleccionar/excluir estudios primarios

Un total de 509 estudios primarios fueron excluidos por su título, dado que en el mismo no contenía ninguna palabra usada en el constructo de búsqueda. Se revisó el resumen de los 115 estudios primarios restantes, si en los mismos no se informaba algún QTL para enfermedades virales en maíz, eran excluidos. De esta selección, quedaron 58 estudios

primarios que fueron evaluados en su texto completo, observando en cada uno que reportara información sobre el QTL, la posición y el efecto aditivo del mismo. Finalmente, 20 estudios primarios con toda la información necesaria fueron seleccionados para construir la base de datos sobre la cual se realizará el meta-análisis.

Paso IV. Recolección de datos originales (construcción de la base de datos)

Se construyeron 10 bases de datos, una para cada cromosoma, todas con el mismo formato. La primer variable (columna) es una variable de clasificación que identifica el estudio primario, en general con el apellido del primer autor y el año de publicación. Esta variable se denominó en este ejemplo de ilustración "*Papers*". A continuación, se colocaron las columnas que identifican los datos con los cuales se realiza el meta-análisis, denominadas Ne, Ee, Nc y Ec, en ese orden. En forma opcional, se pueden agregar otras variables con información recabada desde cada estudio primario, como por ejemplo el caracter medido que contiene información sobre la forma en que fue medida la enfermedad (síntoma, severidad, incidencia o índice de severidad de la enfermedad). También la población de estudio (población biparental experimental o población sobre líneas diversas) u toda otra información que se considere pueda ser utilizada en el meta-análisis. Cada base de datos tiene tantas filas como estudios primarios se hayan encontrado para dicho cromosoma, excepto cuando en un mismo estudio primario se reporte información sobre distintos QTL.

Paso V. Calcular el promedio ponderado de la diferencia de riesgos en un *Forest Plot*

Los resultados muestran que en los cromosomas 1, 3, 4 y 10 existe un efecto global estadísticamente significativo entre el grupo experimental y el grupo control. Solamente en el cromosoma uno, la proporción de QTL de efecto mayor presentes en el grupo experimental fue mayor con respecto a la presentada en el grupo control, debido a que el signo del efecto global es positivo (0,40). Esto significa que en este cromosoma hay mayor probabilidad de encontrar QTL de efecto mayores ($p<0,05$) que en el resto de los cromosomas. Los cromosomas 3, 4 y 10 presentaron efectos globales negativos (-0,20; -0,39 y -0,14, respectivamente), *i.e.*, la mayor proporción de QTL de efecto mayor para la resistencia/tolerancia a enfermedades virales en maíz, estuvo presente en el grupo control.

Se encontró heterogeneidad entre estudios solo en los cromosomas uno y tres. El análisis por subgrupos fue realizado para estos dos cromosomas según los niveles de la variable

Generación que contiene tres categorías: F2: la población era un conjunto de familias F2 y F3 provenientes de un cruzamiento biparental; Líneas: líneas diversas de una población de mapeo por asociación y RIL: la población de mapeo era un conjunto de líneas endocriadas recombinantes provenientes de un cruzamiento biparental. El cromosoma uno presentó alta heterogeneidad promedio (I^2=89%), siendo la categoría F2 la más influyente respecto a las categorías restantes, debido a la presencia de un efecto global (0,63) estadísticamente significativo, ya que el intervalo de confianza para este grupo no contiene el cero, [0,02;1,25] y una alta contribución (peso o ponderación) (54,2%). Además, esta categoría presentó la mayor varianza entre estudios ($\tau^2 = 0{,}6261$) y el estadístico I^2 estimado fue superior (93%) al obtenido cuando se analizan todos los estudios en simultáneo (89%). Esto evidencia que la variabilidad en esta categoría contribuye a la alta heterogeneidad presente a nivel global en el cromosoma uno. A diferencia de lo hallado en el cromosoma uno, en el cromosoma tres la categoría de mayor influencia fue RIL con un intervalo de confianza entre [-0,51; -0,01] y un efecto global negativo de -0,26. Esta categoría fue la de menor varianza entre estudios ($\tau^2 = 0{,}0045$), asimismo presentó heterogeneidad baja (I^2=8%) y la misma no fue estadísticamente significativa. En este cromosoma la categoría de mayor contribución fue F2 (65,1%), sin embargo, esta categoría no presentó un efecto global estadísticamente significativo para este cromosoma al igual que la categoría Líneas. Los *Forest Plot* resultantes de dichos ajustes para cada cromosoma se muestran en la Figura 5.1.

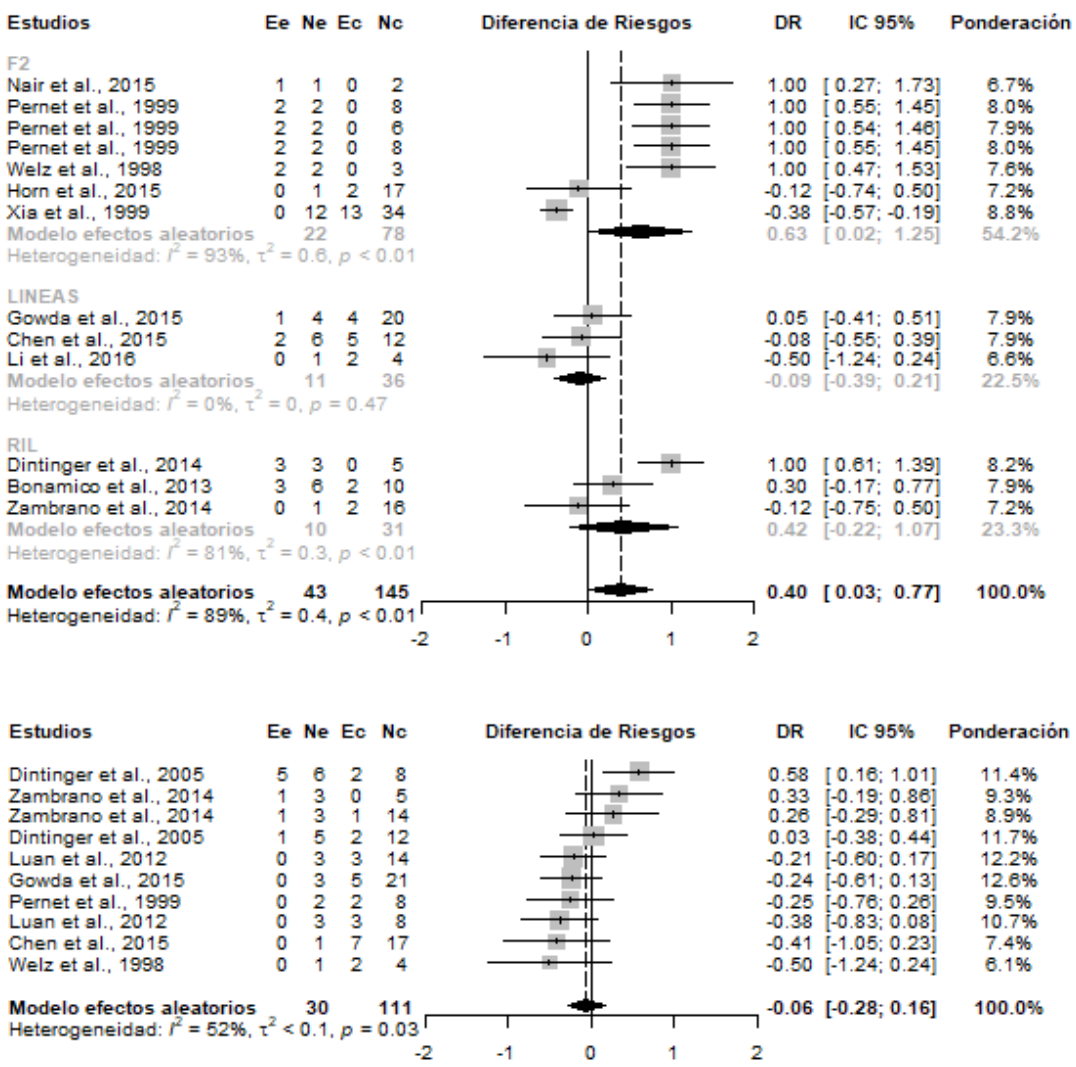

Figura 5.1. *Forest Plots* del modelo de efectos aleatorios para la diferencia de riesgos. Cromosomas 1 (arriba) y 2 (abajo). El cromosoma 1 presenta el análisis por subgrupos según las categorías F2 (población de familias F2 y F3 provenientes de un cruzamiento biparental), Líneas (líneas diversas de una población de mapeo por asociación) y RIL (población de mapeo de líneas endocriadas recombinantes provenientes de un cruzamiento biparental). Las diferencias de riesgos se presentan ordenadas de mayor a menor.

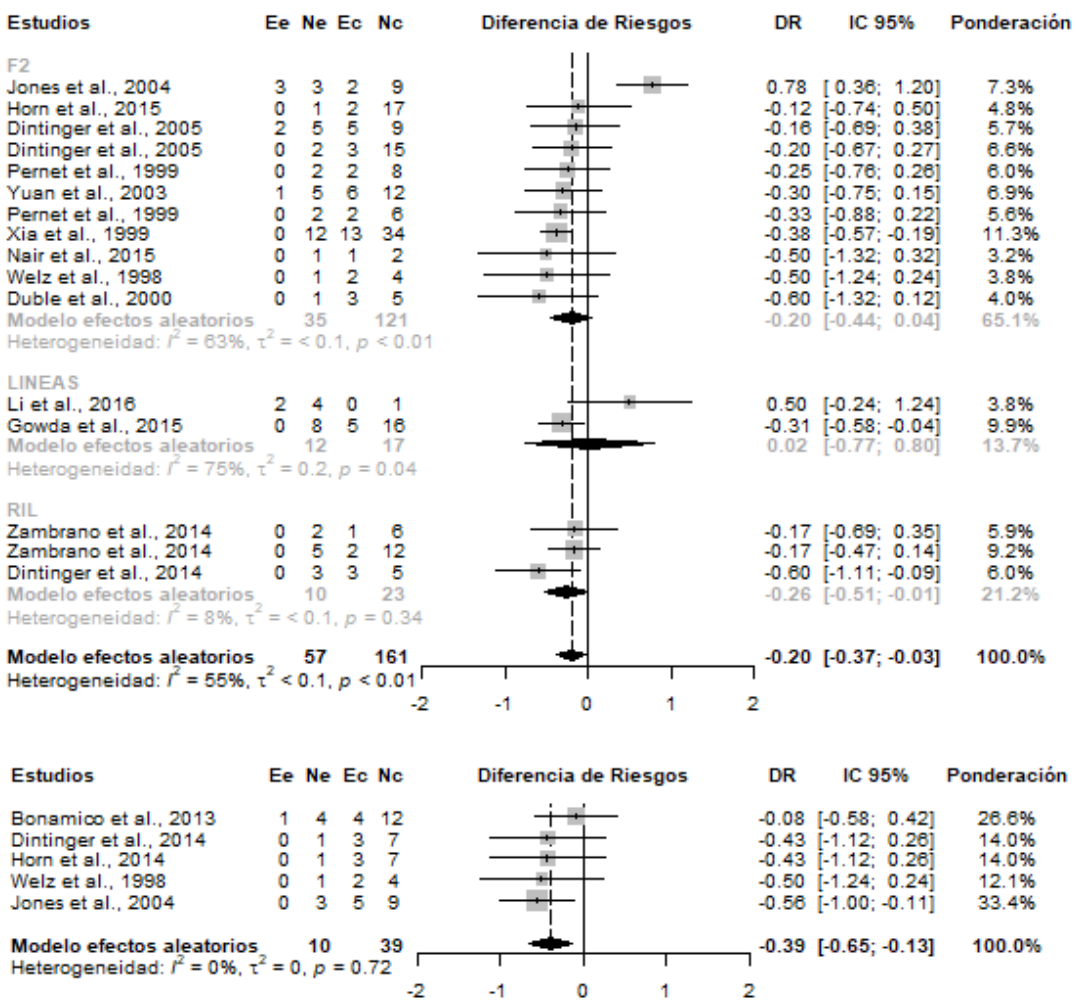

Figura 5.1. *Forest Plots* del modelo de efectos aleatorios para la diferencia de riesgos. Cromosomas 3 (arriba) y 4 (abajo). El cromosoma 3 presenta el análisis por subgrupos según las categorías F2 (población de familias F2 y F3 provenientes de un cruzamiento biparental), Líneas (líneas diversas de una población de mapeo por asociación) y RIL (población de mapeo de líneas endocriadas recombinantes provenientes de un cruzamiento biparental). Las diferencias de riesgos se presentan ordenadas de mayor a menor. Continuación.

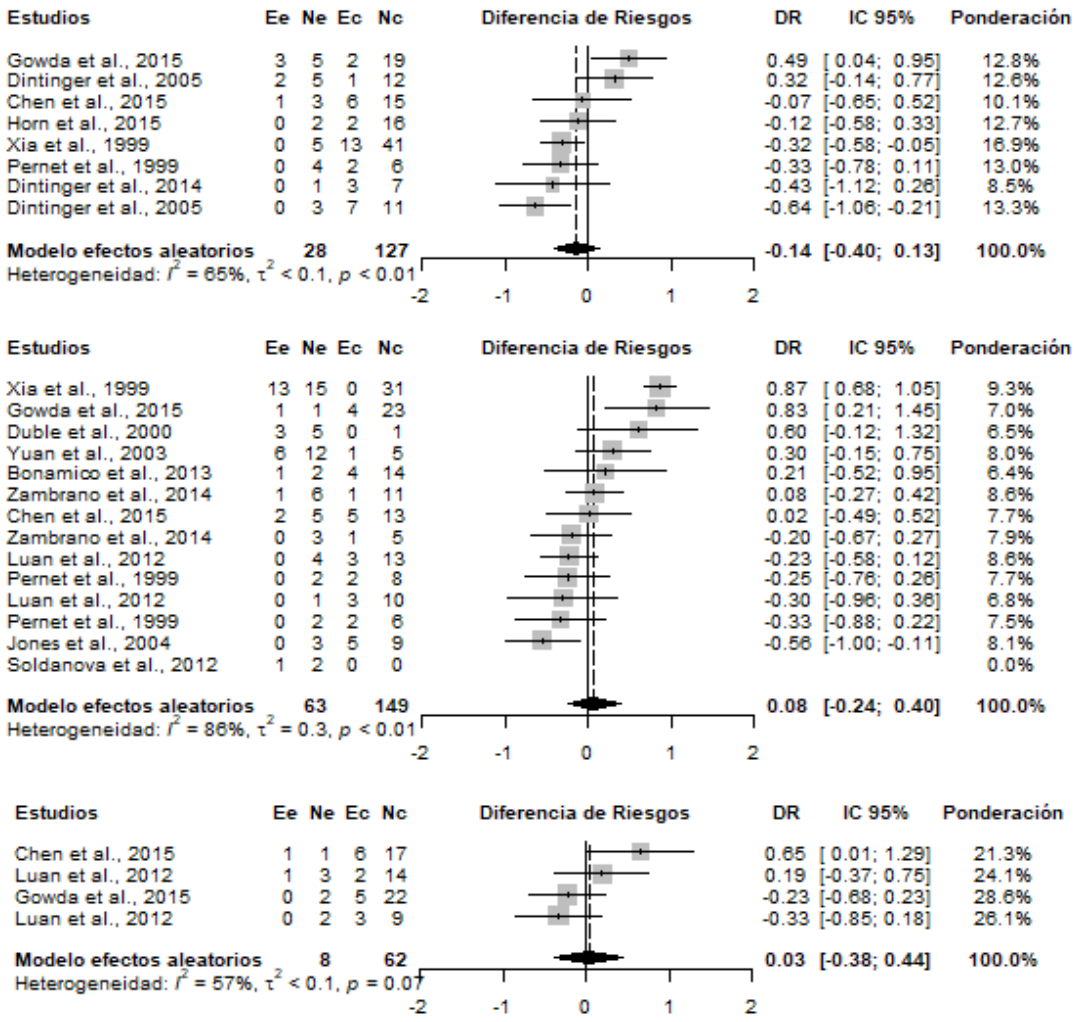

Figura 5.1. *Forest Plots* del modelo de efectos aleatorios para la diferencia de riesgos. Cromosomas 5 al 7 (de arriba hacia abajo). Las diferencias de riesgos se presentan ordenadas de mayor a menor. Continuación.

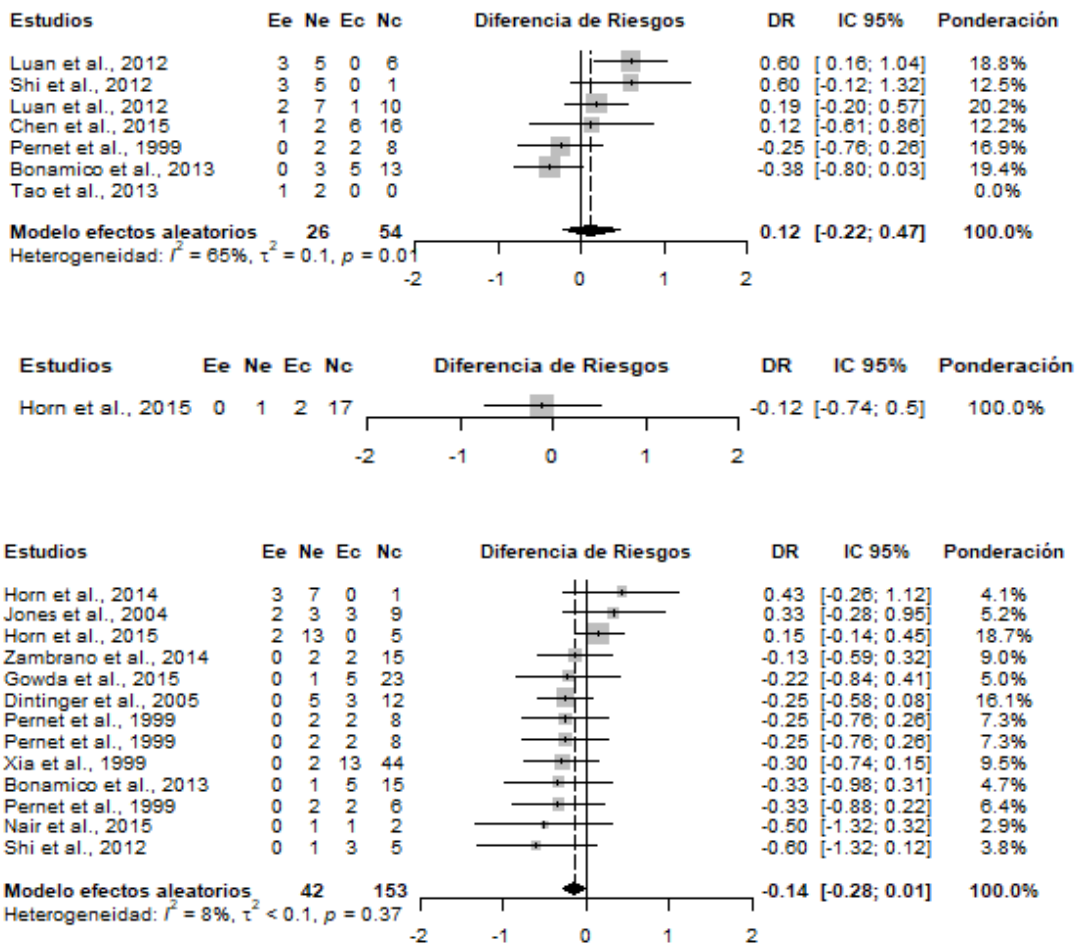

Figura 5.1. *Forest Plots* del modelo de efectos aleatorios para la diferencia de riesgos. Cromosomas 8 al 10 (de arriba hacia abajo). Las diferencias de riesgos se presentan ordenadas de mayor a menor. Continuación.

DISCUSIÓN

En el protocolo analítico presentado en este trabajo permite identificar *loci* de efecto mayor que han sido consensuados por publicaciones de estudios científicos previos. Los *loci* publicados fueron clasificados según el tamaño del efecto sea relativamente mayor (ubicados en el cuarto superior de la lista de efectos) o no mayor (efectos no destacables por su tamaño en comparación a los publicados). Finalmente, para cada cromosoma o para cada grupo de ligamiento se analiza las posiciones donde han sido reportado los genes de efecto mayor a través de una diferencia de riesgos (diferencias en la probabilidad de encontrar un QTL de efecto mayor en el cromosoma estudiado respecto de encontrarlo en el resto del genoma). No obstante, podría haberse usado el cociente de chances (Rossi *et al*., 2018) u otra medida de riesgo relativo (Pai *et al*., 2004). En el caso de usar el cociente de chances como el tamaño del efecto en el meta-análisis, se debe tener en cuenta la cantidad de estudios primarios u observaciones; ya que, para su estimación, se realiza el cociente de un cociente de probabilidades y si en algunos estudios primarios la cantidad de información es cero para el caracter estudiado, la estimación no será posible. La cantidad de estudios primarios obtenidos en este trabajo, a través de la revisión sistemática relacionada con QTL para resistencia/tolerancia a enfermedades virales en maíz fue menor a la registrada por Rossi *et al*. (2018) donde enfermedades fúngicas y bacterianas fueron también incluidas.

El protocolo analítico presentado tiene en cuenta aspectos particulares para ser aplicados en el área de la agronomía, a diferencia de las pautas propuestas por Pai *et al*. (2004) para el área de medicina. Las bases de datos electrónicas que se encuentran en los repositorios digitales suelen estar clasificadas por áreas temáticas, por ello, aquellas que suelen ser usadas en el contexto de agronomía difieren de aquellas usadas en un contexto de medicina. Por ejemplo, bases electrónicas como Scopus, Science Direct, ESCOhost y JSTOR contienen prácticamente todos sus trabajos científicos referidos al área de investigaciones agrícolas, mientras que Pubmed, Embase, NLM Gateway presentan trabajos científicos predominantemente basados en investigaciones sobre enfermedades en humanos. Cualquiera sea la plataforma utilizada para la búsqueda del material digital, el uso de un gestor bibliográfico es crucial para el manejo organizado de los datos recolectados. La selección de un gestor bibliográfico debe realizarse considerando la disponibilidad de herramientas que permitan la importación y exportación de las referencias bibliográficas

para obtener mayores facilidades a la hora de combinar, compactar y unificar estudios primarios duplicados por estar presentes en distintos repositorios electrónicos. Además, el gestor debe vincularse de manera fácil a navegadores de amplia distribución mundial. Para el meta-análisis, se utilizó un software estadístico de libre disposición como lo es R (R Core Team, 2020) pero cualquier otro software estadístico con posibilidad de estimar estadísticos de asociación para datos continuos discretos puede ser utilizado.

El meta-análisis implementado permitió identificar 43 QTL de efecto mayor en el cromosoma 1 con respecto a los QTL de efecto mayor presentes en los otros cromosomas, donde se encontraron 145 QTL para resistencia/tolerancia a enfermedades virales en maíz. Estos resultados coinciden por los encontrados por Rossi *et al*. (2018). En el presente trabajo el efecto global para el cromosoma 3 también fue estadísticamente significativo, pero negativo, indicando mayor probabilidad de obtener un QTL de efecto mayor en los cromosomas restantes que en el evaluado. Otros estudios también reportaron la presencia de QTL en el cromosoma 1, como Di Renzo *et al*. (2004) que detectaron QTL para resistencia al virus del Mal de Río Cuarto (MRCV) en maíz en dicho cromosoma (bin 1,03) y en el cromosoma 8 (bins 8,03 y 8,04). Para este mismo virus, Bonamico *et al*. (2013) identificaron QTL para el índice de severidad de enfermedad (ISE) que es un indicador de la severidad e incidencia de la enfermedad del maíz causada por el MRCV. Los QTL detectados por Bonamico *op. cit*. pertenecen a los cromosomas 1, 4, 6, 8 y 10, mostrando que uno de los QTL con mayor efecto estaba presente en el cromosoma 1 (bin 1,03), siendo este QTL significativo de manera consistente en la mayoría de los ambientes evaluados. Más tarde, Rossi *et al*. (2015) identificaron para el mismo caracter, resistencia al virus del (MRCV), QTL ubicados en los cromosomas 1, 6, 8 y 10. Por otra parte, Yang *et al*. (2017) reportaron genes ubicados en los cromosomas 1, 3, 4, 6, 8 y 10 relacionados a resistencia en enfermedades en maíz y Redinbaugh y Pratt (2009) identificaron *loci* en los cromosomas 1, 3 y 10 para resistencia a virus en maíz. A diferencia de Yang *et al*. (2017), en el presente trabajo, los efectos globales para los cromosomas 3, 4 y 10 fueron negativos, es decir, la probabilidad de encontrar QTL de efecto mayor en dichos cromosomas es menor que la probabilidad de encontrarlos en los cromosomas restantes. Es decir, que los genes reportados en el cromosoma 3, 4 y 10 podrían no estar asociados a *loci* de efecto mayor.

No se encontraron QTL de efecto mayor en el cromosoma 8 en este trabajo. Sin embargo, otros trabajos reportaron la presencia de QTL en dicho cromosoma. Por ejemplo, Shi *et al*.

(2012) identificaron un QTL para la resistencia al virus RBSDV (*rice black streaked dwarf virus*) en el cromosoma 8 (bin 8,03) en maíz mientras que Tao *et al*. (2013) reportaron para el mismo cromosoma y bin (bin 8,03) un QTL para la resistencia al virus MRDD (*maize rough dwarf disease*) en diferentes poblaciones de mapeo. Chen *et al*. (2015) identificaron para el virus MRDD en maíz estudiado por Tao *et al*. (2013), 17 *loci* significativos ubicados en los cromosomas 1, 2, 5, 6, 7 y 8; donde la mayor cantidad de *loci* significativos estuvieron en los cromosomas 1 y 6 (cinco *loci* en cada uno). En el trabajo reportado por McMullen y Simcox (1995) indicaron la presencia de conglomerados de *loci* para resistencia a enfermedades en maíz en todo el genoma; excepto en los cromosomas 7 y 9; los mismos, fueron analizados a través de software BioMercator (Zhao *et al*., 2015).

Es importante destacar que el cromosoma 1 es reportado como la ubicación del genoma de maíz que se encuentra ligada a la resistencia/tolerancia de enfermedades de este cultivo, causadas por virus. Estos resultados se desprenden del efecto global significativo y positivo encontrado por el meta-análisis realizado en el presente trabajo. Sin embargo, a pesar de que hay varios estudios primarios que reportan la presencia de QTL ligados a la resistencia/tolerancia de virus en maíz, en nuestros resultados del consenso de estudios primarios, estos cromosomas no tuvieron QTL ligados a este caracter debido a su efecto global estadísticamente significativo, pero negativo.

CONCLUSIÓN

El protocolo propuesto integra herramientas metodológicas de análisis en cada una de sus etapas con el propósito de identificar la posición de QTL con efecto mayor desde una base de datos conformada mediante revisión sistemática de estudios primarios de QTL. El protocolo representa un nuevo enfoque para identificar QTL con efectos significativos para la resistencia/tolerancia en enfermedades del genoma completo sin la necesidad de usar un mapa genético de referencia para su implementación. La conformación de este protocolo brinda la posibilidad de su aplicación en diferentes especies de importancia agrícola.

CAPÍTULO 6

CONCLUSIONES GENERALES

En este trabajo de tesis se destaca la importancia de la modelación estadística en estudios de asociación. Estos estudios han experimentado cambios en su implementación a través del tiempo debido al incremento en la generación de información molecular causada por el auge de las NGS como así también del aumento de las posibilidades de cómputo para la lectura y procesamiento de este tipo de información. La generación continua de información, tanto molecular como fenotípica, no solo demanda alta capacidad computacional sino también el desarrollo de nuevas metodologías estadísticas para su análisis. Los métodos estadísticos aplicados en los estudios de asociación han ido adaptándose tanto al incremento de información como a la presencia de factores que deben ser considerados en el análisis. Los avances en las biotecnologías de secuenciación aceleraron el desarrollo en los estudios estadísticos-metodológicos para el análisis de asociación, tales como, el mapeo de QTL tradicional, el GWAS y la SG. El mapeo de QTL tradicional permite la detección de QTL, pero requiere la construcción de mapas genéticos a partir de poblaciones biparentales previamente diseñadas experimentalmente y el uso de los QTL de efecto mayor para seleccionar caracteres de interés en el contexto de mejoramiento vegetal. El objetivo de este tipo de estudio es determinar los *loci* responsables en la variación de los caracteres complejos y cuantitativos. En algunas situaciones, la determinación del número de *loci*, la localización de los mismos en el genoma y la interacción de estos *loci* es el objetivo final de estudio; no obstante, también son de interés la identificación de los genes reales y sus funciones. A pesar de su amplio uso, este tipo de estudio mostraba limitantes respecto a la disponibilidad de MM debido a que el genotipado era tedioso y costoso. El advenimiento de otro tipo de estudio denominado GWAS brindó la oportunidad de descifrar arquitecturas genéticas de caracteres complejos en los cultivos sin necesidad de contar con poblaciones diseñadas experimentalmente. A través de modelos GWAS fue posible identificar variantes genéticas ligadas al fenotipo que no habían sido evaluadas en el mapeo de QTL tradicional. Esto fue posible a partir de la implementación de modelos estadísticos con la finalidad de analizar

asociaciones marcador-caracter. Los MLM se han convertido en un marco de análisis muy usado para este tipo de estudios, debido a su capacidad al considerar la estructura genética poblacional subyacente y la relación existente entre los individuos de la población de mapeo a través de matrices de parentesco o la matriz Q de pertenencia. No obstante, el uso de estos modelos aplicados a datos provenientes de METs ha sido poco estudiado en el contexto de GWAS debido a la demanda computacional que estos requieren a la hora de estimar la componente de G×E. Dicha estimación se hace más compleja, si se considera en la estructura de varianza y covarianza distintas correlaciones genéticas dadas por la matriz de parentesco (K) o por la similitud genética estimada a partir de los MM. En esta tesis se comparó el desempeño de los modelos GWAS desde un enfoque por ambiente respecto al abordaje multiambiental considerando ambos tipos de correlaciones genéticas (pedigrí o similitud molecular). La modelación multiambiental permitió estimar las componentes de varianza del efecto de G, de la G×E y de la varianza residual, siendo posible la disección entre las varianzas de los efectos de G y de G×E; no siendo esto posible en la modelación por ambiente donde la varianza de G queda confundida con la varianza de la interacción. No obstante, la demanda computacional al ajustar los modelos GWAS multiambientales fue mayor a la requerida por la modelación por ambiente debido a la mayor cantidad de parámetros a estimar. Sin embargo, el uso de la modelación multiambiental es recomendable cuando se desea conocer cuánto de la variabilidad total está compuesta por cada uno de estos efectos (G y G×E). Además, este tipo de modelos GWAS multiambientales ajustan la varianza fenotípica que se utilizará en la prueba de los efectos de marcador ya que al considerar la varianza G×E, se obtendrá una significativa reducción en la tasa de falsos positivos que no es posible detectar en dicha prueba para la modelación por ambiente. Al ajustar la prueba donde se obtuvieron los efectos de marcador, el modelo GWAS multiambiental que consideró como correlación genética el pedigrí permitió identificar una menor cantidad de marcadores estadísticamente significativos (potencial diminución de la tasa de falsos positivos) en comparación a la modelación que uso la similitud molecular. Los modelos que incluyeron la información de pedigrí mostraron un mejor ajuste que aquellos modelos que consideraron la similitud molecular como medida de correlación genética. Los modelos GWAS representaron, desde su implementación estadística-metodológica, la posibilidad de realizar estudios de asociación en una amplia variedad de cultivos sin necesidad de contar con una población de mapeo experimental, resultando una ventaja

respecto a los estudios de QTL. Si bien, los modelos GWAS permiten detectar los marcadores asociados al efecto fenotípico, no son capaces de inferir el comportamiento de un genotipo no evaluado. Los modelos de SG surgen como una estrategia para predecir el comportamiento fenotípico a partir de datos genómicos, incrementando la eficiencia de los programas de mejoramiento genético vegetal en una amplia variedad de especies vegetales. El objetivo de la SG es predecir genotipos que tengan un alto mérito genético y que puedan ser usados por mejoradores genéticos sin necesidad de evaluaciones a campo. La creciente motivación de usar SG es que se reduce drásticamente el tiempo y los gastos involucrados en el fenotipado de las líneas de mejora, la fase más lenta y costosa del ciclo de mejoramiento genético vegetal. Los estudios de SG son aplicados en poblaciones de individuos que pueden no estar relacionados, en cuyo caso la coancestría entre individuos de la población es pequeña o desconocida. Diferentes estrategias metodológicas han sido propuestas para aumentar la capacidad predictiva del mérito genético a través de los modelos de SG. Hasta el momento, los modelos más usados en maíz y trigo para predecir el mérito genético son G-BLUP y RR-BLUP. Sin embargo, no hay un método que sea considerado el de mejor predicción, dado que el desempeño de cada uno depende de la estructura de correlación subyacente entre los marcadores y entre los genotipos; como así también, de la cantidad de genotipos y/o MM. En este trabajo de tesis, se implementó un meta-análisis para comparar la eficiencia de los métodos de estimación usados en SG en vegetales y cómo la capacidad predictiva a la hora de calibrarlos, cambia al considerar una cantidad determinada de genotipos o de marcadores moleculares u otro tipo de información que puede afectar o incidir en la eficiencia de dichos modelos. Su aplicación permitió analizar simultáneamente resultados de estudios distintos, entorno a una pregunta de investigación; obteniendo así, un tamaño muestral mayor y resultados más potentes que al considerar los estudios individualmente. Es sabido que la capacidad predictiva de un modelo de SG aumenta cuando crece el número de MM. Sin embargo, a partir de este trabajo de tesis, se evidenció que esta respuesta no es lineal y en algunas situaciones el incremento de MM no es redituable en términos de la ganancia relativa de la eficiencia en la SG. Es por ello, que desde un punto de vista estadístico-metodológico aún existe espacio para mejorar los modelos analíticos usados en SG y consecuentemente, su capacidad predictiva para una mayor eficiencia de selección basada en datos genómicos.

A pesar de que la aplicación del análisis de QTL, GWAS y SG puede ser vista en una secuencia temporal, cada uno proporciona información diferente y son utilizados con objetivos de investigación distintos. Para estudios de resistencia/tolerancia a enfermedades virales en maíz existen una considerable cantidad de literatura científica que utiliza el análisis de QTL tradicional para detectar *loci* asociados a esta característica. Los resultados provistos por cada estudio científico son diversos y en algunos casos pueden resultar contradictorios, provocando confusión en la generación de conocimiento en cuanto a la ubicación de los genes asociados a las características de resistencia/tolerancia de interés. En este trabajo de tesis, se propuso un protocolo que integra herramientas metodológicas de análisis en cada una de sus instancias con el propósito de identificar la posición de QTL con efecto mayor a partir de una base de datos conformada mediante revisión sistemática de estudios de QTL. Éste, representa un enfoque distinto a la hora de identificar QTL con efectos significativos para la resistencia/tolerancia en enfermedades del genoma completo del maíz sin la necesidad de usar un mapa genético de referencia para su implementación. Siendo posible su aplicación en diferentes especies de importancia agrícola.

REFERENCIAS BIBLIOGRÁFICAS

Abdullaev, A. A., Salakhutdinov, I. B., Egamberdiev, S. S., Khurshut, E. E., Rizaeva, S. M., Ulloa, M., y Abdurakhmonov, I. Y. (2017). Genetic diversity, linkage disequilibrium, and association mapping analyses of *Gossypium barbadense* L. germplasm. *PLoS One, 12*(11), e0188125.

Aguate, F., Crossa, J., y Balzarini, M. (2019). Effect of missing values on variance component estimates in multienvironment trials. *Crop Science, 59*(2), 508-517.

Akobeng, A. K. (2005). Understanding systematic reviews and meta-analysis. *Archives of Disease in Childhood, 90*(8), 845-848.

Ali, F., y Yan, J. (2012). Disease resistance in maize and the role of molecular breeding in defending against global threat. *Journal of Integrative Plant Biology, 54*(3), 134-151.

Asoro, F. G., Newell, M. A., Beavis, W. D., Scott, M. P., y Jannink, J.-L. (2011). Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. *The Plant Genome, 4*, 132-144.

Astle, W., y Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science, 24*(4), 451-471.

Ball, R. D. (2005). Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. *Genetics, 170*(2), 859-873.

Ball, R. D. (2013a). Designing a GWAS: power, sample size, and data structure. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 37-98). Totowa, NJ: Humana Press.

Ball, R. D. (2013b). Statistical analysis of genomic data. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 171-192). Totowa, NJ: Humana Press.

Balzarini, M. (2002). Applications of mixed models in plant breeding. En M. S. Kang (Ed.), *Quantitative Genetics, Genomics and Plant Breeding* (CAB Int., pp. 353–363).

Benjamini, Y., y Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*: *series B (Methodological), 57*(1), 289-300.

Benjamini, Y., y Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165-1188.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science, 48*(5), 1649-1664.

Bernardo, R., y Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, *47*(3), 1082-1090.

Bhat, J. A., Ali, S., Salgotra, R. K., Mir, Z. A., Dutta, S., Jadon, V., Tyagi, A., Mushtaq, M., Jain, N., Singh, P. K., Singh, G. P., y Prabhu, K. V. (2016). Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Frontiers in Genetics*, *7*, 221.

Bonamico, N. C., Di Renzo, M. A., Borghi, M. L., Ibañez, M. A., Díaz, D. G., Salerno, J. C., y Balzarini, M. G. (2013). Mapeo de QTL para una medida multivariada de la reacción al virus del mal de Río cuarto. *BAG. Journal of Basic and Applied Genetics*, *24*(2), 11-21.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. *Studi in onore del professore salvatore ortu carboni*, 13-60.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., y Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons. Ltd.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., y Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97-111.

Borevitz, J. O., y Chory, J. (2004). Genomics tools for QTL analysis and gene discovery. *Current Opinion in Plant Biology*, *7*(2), 132-136.

Borgognone, M. G., Butler, D. G., Ogbonnaya, F. C., y Dreccer, M. F. (2016). Molecular marker information in the analysis of multi-environment trials helps differentiate superior genotypes from promising parents. *Crop Science*, *56*, 2612-2628.

Brachi, B., Morris, G. P., y Borevitz, J. O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biology*, *12*(10), 232.

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., y Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633-2635.

Breen, G., Harold, D., Ralston, S., Shaw, D., y Clair, D. S. (2000). Determining SNP allele frequencies in DNA pools. *Biotechniques*, *28*(3), 464-470.

Buckler, E. S., Holland, J. B., Bradbury, P. J., Acharya, C. B., Brown, P. J., Browne, C., Ersoz, E., Flint-Garcia, S., Garcia, A., Glaubitz, J. C., Goodman, M. M., Harjes, C., Guill, K., Kroon, D. E., Larsson, S., Lepak, N. K., Li, H., Mitchell, S. E., Pressoir, G., Peiffer, J. A., Rosas, M. O., Rocheford, T. R., Romay, M. C., Romero, S., Salvo, S., Villeda, H. S., Sofia da Silva, H., Sun, Q., Tian, F., Upadyayula, N., Ware, D., Yates, H., Yu, J., Zhang, Z., Kresovich, S., y McMullen, M. D. (2009). The genetic architecture of maize flowering time. *Science*, *325*(5941), 714-718.

Burgueño, J., Crossa, J., Cornelius, P. L., Trethowan, R., McLaren, G., y Krishnamachari, A. (2007). Modeling additive×environment and additive× additive×environment using genetic covariances of relatives of wheat genotypes. *Crop Science*, *47*, 311-320.

Burgueño, J., de los Campos, G., Weigel, K., y Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype×environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*, 707-719.

Burke, J. M., Tang, S., Knapp, S. J., y Rieseberg, L. H. (2002). Genetic analysis of sunflower domestication. *Genetics*, *161*(3), 1257-1267.

Cabrera-Bosquet, L., Crossa, J., von Zitzewitz, J., Serret, M. D., y Luis Araus, J. (2012). High-throughput phenotyping and genomic selection: the frontiers of crop breeding converge. *Journal of Integrative Plant Biology*, *54*(5), 312-320.

Cao, K., Zhou, Z., Wang, Q., Guo, J., Zhao, P., Zhu, G., Fang, W., Chen, C., Wang, X., Wang, X., Tian, Z., y Wang, L. (2016). Genome-wide association study of 12 agronomic traits in peach. *Nature Communications*, *7*, 13246.

Cappa, E. P., El-Kassaby, Y. A., Garcia, M. N., Acuña, C., Borralho, N. M. G., Grattapaglia, D., y 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 *Eucalyptus globulus*. *PLoS One*, *8*(11), e81267.

Cappa, E. P., Martínez, M. C., Garcia, M. N., Villalba, P. V., y Poltri, S. N. M. (2011). Effect of population structure and kinship relationships on the results of association mapping tests of growth and wood quality traits in four Eucalyptus populations. *BMC proceedings*, *5*(7), P23. BioMed Central.

Carrijo, D. R., Lundy, M. E., y Linquist, B. A. (2017). Rice yields and water use under alternate wetting and drying irrigation: a meta-analysis. *Field Crops Research*, *203*, 173-180.

Chan, E. K. F., Rowe, H. C., y Kliebenstein, D. J. (2010). Understanding the evolution of defense metabolites in *Arabidopsis thaliana* using genome-wide association mapping. *Genetics*, *185*(3), 991-1007.

Chen, G., Wang, X., Hao, J., Yan, J., y Ding, J. (2015). Genome-wide association implicates candidate genes conferring resistance to maize rough dwarf disease in maize. *PLoS One*, *10*(11), e0142001.

Chen, L., An, Y., Li, Y.-X., Li, C., Shi, Y., Song, Y., Zhang, D., Wang, T., y Li, Y. (2017). Candidate loci for yield-related traits in maize revealed by a combination of MetaQTL analysis and regional association mapping. *Frontiers in Plant Science*, *8*, 2190.

Clark, S. A., y van der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. En *Genome-Wide Association Studies and Genomic Prediction* (pp. 321-330). Springer.

Courtois, B., Ahmadi, N., Khowaja, F., Price, A.H., Rami, J.F., Frouin, J., Hamelin, C., y Ruiz, M. (2009). Rice root genetic architecture: meta-analysis from a drought QTL database. *Rice*, *2*(2), 115-128.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One*, *11*(6), e0156744.

Crossa, J., Burgueño, J., Cornelius, P. L., McLaren, G., Trethowan, R., y Krishnamachari, A. (2006). Modeling genotype×environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Science*, *46*, 1722-1733.

Crossa, J., Vargas, M., y Joshi, A. K. (2010). Linear, bilinear, and linear-bilinear fixed and mixed models for analyzing genotype×environment interaction in plant breeding and agronomy. *Canadian Journal of Plant Science*, *90*(5), 561-574.

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., y de los Campos, G. (2017). Bayesian genomic prediction with genotype×environment interaction kernel models. *G3: Genes, Genomes, Genetics*, *7*(1), 41-53.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., Montesinos-López, O.A., y Burgueño, J. (2016). Genomic prediction of genotype × environment interaction kernel regression models. *The Plant Genome*, *9*(3).

Cuevas, J., Granato, I., Fritsche-Neto, R., Montesinos-Lopez, O. A., Burgueño, J., e Sousa, M. B., y Crossa, J. (2018). Genomic-enabled prediction kernel models with random intercepts for multi-environment trials. *G3: Genes, Genomes, Genetics*, *8*(4), 1347-1365.

Daetwyler, H. D., Bansal, U. K., Bariana, H. S., Hayden, M. J., y Hayes, B. J. (2014). Genomic prediction for rust resistance in diverse wheat landraces. *Theoretical and Applied Genetics*, *127*(8), 1795-1803.

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., y Hickey, J. M. (2013). Genomic prediction in animals and plants: simulation of data, validation, reporting, and benchmarking. *Genetics*, *193*(2), 347-365.

de los Campos, G., Gianola, D., Rosa, G. J. M., Weigel, K. A., y Crossa, J. (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research*, *92*(4), 295-308.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., y Calus, M. P. L. (2013). Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, *193*(2), 327-345.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., y Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, *182*(1), 375-385.

Dekkers, J. C. M. (2004). Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. *Journal of Animal Science*, *82*(suppl_13), E313-E328.

Del Carpio, D. P., Lozano, R., Wolfe, M. D., y Jannink, J.-L. (2018). Genome-wide association studies and heritability estimation in the functional genomics era. En *Population Genomics* (pp. 361-425). Springer.

Demidenko, E. (2004). *Mixed models: theory and application*. New Jersey: John Wiley & Sons.

Desta, Z. A., y Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, *19*(9), 592-601.

Devlin, B., y Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997-1004.

Di Renzo, M. A., Bonamico, N. C., Díaz, D. G., Ibañez, M. A., Faricelli, M. E., Balzarini, M. G., y Salerno, J. C. (2004). Microsatellite markers linked to QTL for resistance to Mal de Rio Cuarto disease in *Zea mays* L. *The Journal of Agricultural Science*, *142*(3), 289-295.

Doebley, J., Stec, A., Wendel, J., y Edwards, M. (1990). Genetic and morphological analysis of a maize-teosinte F2 population: implications for the origin of maize. *Proceedings of the National Academy of Sciences*, *87*(24), 9888-9892.

Doerge, R. W. (2002). Multifactorial genetics: mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics*, *3*(1), 43-52.

Dong, L., Xiao, S., Wang, Q., y Wang, Z. (2016). Comparative analysis of the GBLUP, emBayesB, and GWAS algorithms to predict genetic values in large yellow croaker (*Larimichthys crocea*). *BMC Genomics*, *17*(1), 460.

Etzel, C. J., y Guerra, R. (2002). Meta-analysis of genetic-linkage analysis of quantitative-trait loci. *The American Journal of Human Genetics*, *71*(1), 56-65.

Falush, D., Stephens, M., y Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567-1587.

Fernando, R. L., y Garrick, D. (2013). Bayesian methods applied to GWAS. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 237-274). Totowa, NJ: Humana Press.

Ferrão, L. F. V., Ortiz, R., y Garcia, A. A. F. (2017). Genomic selection: state of the art. En *Genetic Improvement of Tropical Crops* (pp. 19-54). Springer.

Ferreira González, I., Urrútia, G., y Alonso-Coello, P. (2011). Revisiones sistemáticas y metaanálisis: bases conceptuales e interpretación. *Revista Española de Cardiología*, *64*(8), 688-696.

Ferrero-Serrano, Á., y Assmann, S. (2019). Phenotypic and genome-wide association with the local environment of Arabidopsis. *Nature Ecology & Evolution*, *3*, 274-285.

Field, A. P. (2001). Meta-analysis of correlation coefficients: a Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*(2), 161-180.

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*(4), 444-467.

Flint-Garcia, S. A., Thornsberry, J. M., y Buckler IV, E. S. (2003). Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology*, *54*(1), 357-374.

Friedman, J., Hastie, T., y Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1-22.

Gauderman, W. J., Mukherjee, B., Aschard, H., Hsu, L., Lewinger, J. P., Patel, C. J., Witte, J. S., Amos, C., Tai, C. G., Conti, D., Torgerson, D. G., Lee, S., y Chatterjee, N. (2017). Update on the state of the science for analytical methods for gene-environment interactions. *American Journal of Epidemiology*, *186*(7), 762-770.

Geng, X., Sha, J., Liu, S., Bao, L., Zhang, J., Wang, R., Yao, J., Li, C., Feng, J., y Sun, F. (2015). A genome-wide association study in catfish reveals the presence of functional hubs of related genes within QTLs for columnaris disease resistance. *BMC Genomics*, *16*(1), 196.

Gianola, D. (2013). Priors in whole-genome regression: the Bayesian alphabet returns. *Genetics*, *194*(3), 573-596.

Gianola, D., y van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, *178*(4), 2289-2303.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*(10), 3-8.

Goddard, M. E., y Hayes, B. J. (2007). Genomic selection. *Journal of Animal Breeding and Genetics*, *124*(6), 323-330.

Goffinet, B., y Gerber, S. (2000). Quantitative trait loci: a meta-analysis. *Genetics*, *155*(1), 463-473.

Gondro, C., Lee, S. H., Lee, H. K., y Porto-Neto, L. R. (2013). Quality control for genome-wide association studies. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 129-147). Totowa, NJ: Humana Press.

Gondro, C., Porto-Neto, L. R., y Lee, S. H. (2013). R for genome-wide association studies. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 1-17). Totowa, NJ: Humana Press.

Griffiths, S., Simmonds, J., Leverington, M., Wang, Y., Fish, L., Sayers, L., Alibert, L., Orford, S., Wingen, L., y Herry, L. (2009). Meta-QTL analysis of the genetic control of ear emergence in elite European winter wheat germplasm. *Theoretical and Applied Genetics*, *119*(3), 383-395.

Guo, J., Chen, L., Li, Y., Shi, Y., Song, Y., Zhang, D., Li, Y., Wang, T., Yang, D., y Li, C. (2018). Meta-QTL analysis and identification of candidate genes related to root traits in maize. *Euphytica*, *214*(12), 223.

Gupta, P. K., Kulwal, P. L., y Jaiswal, V. (2014). Association mapping in crop plants: opportunities and challenges. En *Advances in Genetics* (Vol. 85, pp. 109-147). Academic Press.

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

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

Habier, D., Fernando, R. L., Kizilkaya, K., y Garrick, D. J. (2011). Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics, 12*(1), 186.

Hardy, O. J., y Vekemans, X. (2002). SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. *Molecular Ecology Notes, 2*(4), 618-620.

Hawkins, C., y Yu, L.-X. (2018). Recent progress in alfalfa (Medicago sativa L.) genomics and genomic selection. *The Crop Journal*, *6*(6), 565-575.

Hayes, B. (2013). Overview of statistical methods for genome-wide association studies (GWAS). En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 149-169). Totowa, NJ: Humana Press.

Hazzouri, K. M., Purugganan, M. D., y Flowers, J. M. (2014). Population genomics of plant species. En *Advances in Botanical Research* (Vol. 69, pp. 311-334). Elsevier.

He, J., Zhao, X., Laroche, A., Lu, Z.-X., Liu, H., y Li, Z. (2014). Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Frontiers in Plant Science, 5*, 484.

Hedges, L. V., y Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Heffner, E. L., Jannink, J.-L., Iwata, H., Souza, E., y Sorrells, M. E. (2011). Genomic selection accuracy for grain quality traits in biparental wheat populations. *Crop Science*, *51*, 2597-2606.

Heffner, E. L., Jannink, J.-L., y Sorrells, M. E. (2011). Genomic selection accuracy using multifamily prediction models in a wheat breeding program. *The Plant Genome*, *4*, 65-75.

Heffner, E. L., Lorenz, A. J., Jannink, J.-L., y Sorrells, M. E. (2010). Plant breeding with genomic selection: gain per unit time and cost. *Crop Science, 50*, 1681-1690.

Heffner, E. L., Sorrells, M. E., y Jannink, J.-L. (2009). Genomic selection for crop improvement. *Crop Science, 49*(1), 1-12.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, *31*(2), 423-447.

Henshall, J. M. (2013). Validation of genome-wide association studies (GWAS) results. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 411-421). Totowa, NJ: Humana Press.

Higgins, J. P. T., y Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539-1558.

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., y Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560.

Hong, E. P., y Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & Informatics*, *10*(2), 117-122.

Hong, Y., Chen, X., Liang, X., Liu, H., Zhou, G., Li, S., Wen, S., Holbrook, C.C., y Guo, B. (2010). A SSR-based composite genetic linkage map for the cultivated peanut (*Arachis hypogaea* L.) genome. *BMC Plant Biology*, *10*(1), 17.

Hotelling, H. (1936). Simplified calculation of principal components. *Psychometrika*, *1*(1), 27-35.

Huang, C., Nie, X., Shen, C., You, C., Li, W., Zhao, W., Zhang, X., y Lin, Z. (2017). Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. *Plant Biotechnology Journal*, *15*(11), 1374-1386.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). Statistical learning. En *An Introduction to Statistical Learning* (pp. 15-57). Springer.

Jannink, J.-L., Lorenz, A. J., y Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. *Briefings in Functional Genomics*, *9*(2), 166-177.

Jannink, J.-L., y Walsh, B. (2002). Association mapping in plant populations. En *Quantitative Genetics, Genomics and Plant Breeding* (pp. 59-68). CAB International: New York, NY, USA.

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

Kang, M. S., Balzarini, M. G., y Guerra, J. L. L. (2004). Genotype-by-environment interaction. En *Genetic Analysis of Complex Traits Using SAS* (Cary: SAS, pp. 78–118).

Kelley, G. A., y Kelley, K. S. (2012). Statistical models for meta-analysis: a brief tutorial. *World Journal of Methodology*, *2*(4), 27-32.

Lado, B., Barrios, P. G., Quincke, M., Silva, P., y Gutiérrez, L. (2016). Modeling genotype×environment interaction for genomic selection with unbalanced data from a wheat breeding program. *Crop Science*, *56*(5), 2165-2179.

Lande, R., y Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, *124*(3), 743-756.

Li, D., Zhao, X., Han, Y., Li, W., y Xie, F. (2019). Genome-wide association mapping for seed protein and oil contents using a large panel of soybean accessions. *Genomics*, *111*(1), 90-95.

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

Li, L., Li, X., Li, L., Schnable, J., Gu, R., y Wang, J. (2019). QTL identification and epistatic effect analysis of seed size-and weight-related traits in *Zea mays* L. *Molecular Breeding*, *39*(5), 67.

Li, W. T., Liu, C., Liu, Y. -X., Pu, Z. E., Dai, S. -F., Wang, J. -R., Lan, X. -J., Zheng, Y. -L., y Wei, Y. -M. (2013). Meta-analysis of QTL associated with tolerance to abiotic stresses in barley. *Euphytica*, *189*(1), 31-49.

Li, Z., y Sillanpää, M. J. (2012). Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*, *125*(3), 419-435.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C. M., Davidson, R. I., y Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. *Nature Methods*, *8*(10), 833-835.

Listgarten, J., Lippert, C., Kadie, C. M., Davidson, R. I., Eskin, E., y Heckerman, D. (2012). Improved linear mixed models for genome-wide association studies. *Nature Methods*, *9*(6), 525-526.

Liu, S., Hall, M. D., Griffey, C. A., y McKendry, A. L. (2009). Meta-analysis of QTL associated with Fusarium head blight resistance in wheat. *Crop Science*, *49*(6), 1955-1968.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J. -L., Singh, R. P., Autrique, E., y de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker×environment interaction genomic selection model. *G3: Genes, Genomes, Genetics*, *5*(4), 569-582.

Lorenz, A. J., Chao, S., Asoro, F. G., Heffner, E. L., Hayashi, T., Iwata, H., Smith, K. P., Sorrells, M. E., y Jannink, J. -L. (2011). Genomic selection in plant breeding: knowledge and prospects. En *Advances in Agronomy* (Vol. 110, pp. 77-123). Elsevier.

Lorenzana, R. E., y Bernardo, R. (2009). Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and Applied Genetics*, *120*(1), 151-161.

Lu, Q., Liu, H., Hong, Y., Li, H., Liu, H., Li, X., Wen, S., Zhou, G., Li, S., y Chen, X. (2018). Consensus map integration and QTL meta-analysis narrowed a locus for yield traits to 0.7 cM and refined a region for late leaf spot resistance traits to 0.38 cM on linkage group A05 in peanut (*Arachis hypogaea* L.). *BMC Genomics*, *19*(1), 887.

Lynch, M., y Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, *152*(4), 1753-1766.

Mackay, I., y Powell, W. (2007). Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science*, *12*(2), 57-63.

Mackay, T. F. C., Stone, E. A., y Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, *10*(8), 565-577.

Malosetti, M., Ribaut, J. M., y van Eeuwijk, F. A. (2013). The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in Physiology*, *4*, 44.

Malosetti, M., van der Linden, C. G., Vosman, B., y van Eeuwijk, F. A. (2007). A mixed-model approach to association mapping using pedigree information with an illustration of resistance to Phytophthora infestans in potato. *Genetics*, *175*(2), 879-889.

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., y Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, *27*(2), e1608.

McLaren, C. G., Bruskiewich, R. M., Portugal, A. M., y Cosico, A. B. (2005). The international rice information system. A platform for meta-analysis of rice crop data. *Plant Physiology*, *139*(2), 637-642.

McLaren, C. G., Ramos, L., López, C., y Eusebio, W. (2000). *Applications of the genealogy management system. ICIS International Crop Information System: Technical Development Manual-version 6*. Mexico DF (Mexico): Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT), Mexico DF.

McMullen, M. D., y Simcox, K. D. (1995). Genomic organization of disease and insect resistance genes in maize. *Molecular Plant Microbe interactions*, *8*(6), 811-815.

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., y Cheraghi-Sohi, S. (2014). PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC Health Services Research*, *14*(1), 579.

Meuwissen, T. H., Hayes, B. J., y Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819-1829.

Miguez, F. E., y Bollero, G. A. (2005). Review of corn yield response under winter cover cropping systems using meta-analytic methods. *Crop Science*, *45*, 2318-2329.

Miles, C. M., y Wayne, M. (2008). Quantitative trait locus (QTL) analysis. *Nature Education*, *1*(1), 208.

Mitchell-Olds, T. (2010). Complex-trait analysis in plants. *Genome Biology*, *11*, 1-3.

Murcray, C. E., Lewinger, J. P., y Gauderman, W. J. (2008). Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology*, *169*(2), 219-226.

Nakaya, A., e Isobe, S. N. (2012). Will genomic selection be a practical method for plant breeding? *Annals of Botany*, *110*(6), 1303-1316.

Neves, H. H. R., Carvalheiro, R., O'Brien, A. M. P., Utsunomiya, Y. T., do Carmo, A. S., Schenkel, F. S., Sölkner, J., McEwan, J. C., Van Tassell, C. P., Cole, J. B., da Silva, M. V. G. B., Queiroz, S. A., Sonstegard, T. S., y Garcia, J. F. (2014). Accuracy of genomic predictions in *Bos indicus* (Nellore) cattle. *Genetics Selection Evolution*, *46*(1), 17.

Neves, H. H. R., Carvalheiro, R., y Queiroz, S. A. (2012). A comparison of statistical methods for genomic selection in a mice population. *BMC Genetics*, *13*(1), 100.

Oakey, H., Cullis, B., Thompson, R., Comadran, J., Halpin, C., y Waugh, R. (2016). Genomic selection in multi-environment crop trials. *G3: Genes, Genomes, Genetics*, *6*(5), 1313-1326.

Olkin, I. (1995). Statistical and theoretical considerations in meta-analysis. *Journal of Clinical Epidemiology*, *48*(1), 133-146.

Ott, J., Kamatani, Y., y Lathrop, M. (2011). Family-based designs for genome-wide association studies. *Nature Reviews Genetics*, *12*(7), 465-474.

Ould Estaghvirou, S. B., Ogutu, J. O., Schulz-Streeck, T., Knaak, C., Ouzunova, M., Gordillo, A., y Piepho, H.-P. (2013). Evaluation of approaches for estimating the accuracy of genomic prediction in plant breeding. *BMC Genomics*, *14*, 1-21.

Pai, M., McCulloch, M., Gorman, J. D., Pai, N., Enanoria, W., Kennedy, G., Tharyan, P., y Colford, J. M. (2004). Systematic reviews and meta-analyses: an illustrated, step-by-step guide. *The National Medical Journal of India*, *17*(2), 86-95.

Parisseaux, B., y Bernardo, R. (2004). In silico mapping of quantitative trait loci in maize. *Theoretical and Applied Genetics*, *109*(3), 508-514.

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

Pe'er, I., de Bakker, P. I. W., Maller, J., Yelensky, R., Altshuler, D., y Daly, M. J. (2006). Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genetics*, *38*(6), 663-667.

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

Peña Malavera, A., Gutierrez, L., y Balzarini, M. (2016). Modelos estadísticos para estudios de asociación fenotipo-genotipo en poblaciones genéticamente estructuradas. *BAG. Journal of Basic and Applied Genetics*, *27*(2), 49-58.

Petersen, A., Spratt, J., y Tintle, N. L. (2013). Incorporating prior knowledge to increase the power of genome-wide association studies. En C. Gondro, J. van der Werf, y B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols)* (pp. 519-541). Totowa, NJ: Humana Press.

Piepho, H. P., Möhring, J., Melchinger, A. E., y Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, *161*(1), 209-228.

Pittelkow, C. M., Linquist, B. A., Lundy, M. E., Liang, X., van Groenigen, K. J., Lee, J., van Gestel, N., Six, J., Venterea, R. T., y van Kessel, C. (2015). When does no-till yield more? A global meta-analysis. *Field Crops Research*, *183*, 156-168.

Poland, J. A., y Rife, T. W. (2012). Genotyping-by-sequencing for plant breeding and genetics. *The Plant Genome*, *5*(3), 92-102.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., y Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904-909.

Pritchard, J. K., Stephens, M., y Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

Qi, Z. -M., Sun, Y. -N., Wu, Q., Liu, C. -Y., Hu, G. -H., y Chen, Q. -S. (2011). A meta-analysis of seed protein concentration QTL in soybean. *Canadian Journal of Plant Science*, *91*(1), 221-230.

R Core Team. (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. Vienna.

Rabier, C.-E., Barre, P., Asp, T., Charmet, G., y Mangin, B. (2016). On the accuracy of genomic selection. *PLoS One*, *11*(6), e0156086.

Rafalski, J. A. (2010). Association genetics in crop improvement. *Current Opinion in Plant Biology*, *13*(2), 174-180.

Redinbaugh, M. G., y Pratt, R. C. (2009). Virus resistance. En *Handbook of maize: Its biology* (pp. 251-270). Springer.

Remington, D. L., y Purugganan, M. D. (2003). Candidate genes, quantitative trait loci, and functional trait evolution in plants. *International Journal of Plant Sciences*, *164*(S3), S7-S20.

Resende, M. F. R., Muñoz, P., Resende, M. D. V., Garrick, D. J., Fernando, R. L., Davis, J. M., Jokela, E. J., Martin, T. A., Peter, G. F., y Kirst, M. (2012). Accuracy of genomic selection methods in a standard data set of Loblolly Pine (*Pinus taeda* L.). *Genetics*, *190*(4), 1503-1510.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetics Research*, *67*(2), 175-185.

Roff, D. A. (2007). A centennial celebration for quantitative genetics. *Evolution*, *61*(5), 1017-1032.

Rossi, E. A., Borghi, M. L., Di Renzo, M. A., y Bonamico, N. C. (2015). Quantitative Trait loci (QTL) Identification for resistance to Mal de Rio Cuarto Virus (MRCV) in maize based on segregate population. *The Open Agriculture Journal*, *9*(1), 48-55.

Rossi, E. A., Ruiz, M., Rueda Calderón, M. A., Bruno, C. I., Bonamico, N. C., y Balzarini, M. G. (2018). Meta-analysis of QTL studies for resistance to fungi and viruses in maize. *Crop Science*, *59*(1), 125-139.

Rotundo, J. L., y Westgate, M. E. (2009). Meta-analysis of environmental effects on soybean seed composition. *Field Crops Research*, *110*(2), 147-156.

Saïdou, A.-A., Thuillet, A.-C., Couderc, M., Mariac, C., y Vigouroux, Y. (2014). Association studies including genotype by environment interactions: prospects and limits. *BMC Genetics*, *15*(1), 3.

Salih, H., y Adelson, D. L. (2009). QTL global meta-analysis: are trait determining genes clustered? *BMC Genomics*, *10*(1), 184.

Sánchez-Meca, J. (2010). Cómo realizar una revisión sistemática y un meta-análisis. *Aula Abierta*, *38*(2), 53–64.

Sargeant, J. M., Rajic, A., Read, S., y Ohlsson, A. (2006). The process of systematic review and its application in agri-food public-health. *Preventive Veterinary Medicine*, *75*(3), 141-151.

Schmidt, F. L., Gast-Rosenberg, I., y Hunter, J. E. (1980). Validity generalization results for computer programmers. *Journal of Applied Psychology*, *65*(6), 643-661.

Sehgal, D., Singh, R., y Rajpal, V. R. (2016). Quantitative trait loci mapping in plants: concepts and approaches. En *Molecular Breeding for Sustainable Crop Improvement* (pp. 31-59). Springer.

Setakis, E., Stirnadel, H., y Balding, D. J. (2006). Logistic regression protects against population structure in genetic association studies. *Genome Research*, *16*(2), 290-296.

Shi, L. -Y., Hao, Z. -F., Weng, J. -F., Xie, C. -X., Liu, C. -L., Zhang, D. -G., Li, M. -S., Bai, L., Li, X. -H., y Zhang, S. -H. (2012). Identification of a major quantitative trait locus for resistance to maize rough dwarf virus in a Chinese maize inbred line X178 using a linkage map based on 514 gene-derived single nucleotide polymorphisms. *Molecular Breeding*, *30*(2), 615-625.

Shin, J., y Lee, C. (2015). Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. *Genomics*, *105*(1), 1-4.

Shirasawa, K., Bertioli, D. J., Varshney, R. K., Moretzsohn, M. C., Leal-Bertioli, S. C. M., Thudi, M., Pandey, M. K., Rami, J.-F., Foncéka, D., y Gowda, M. V. C. (2013). Integrated consensus map of cultivated peanut and wild relatives reveals structures of the A and B genomes of Arachis and divergence of the legume genomes. *DNA Research*, *20*(2), 173-184.

Silver, N. C., y Dunlap, W. P. (1987). Averaging correlation coefficients: should Fisher's z transformation be used? *Journal of Applied Psychology*, *72*(1), 146-148.

Singh, A., Sharma, V., Dikshit, H. K., Aski, M., Kumar, H., Thirunavukkarasu, N., Patil, B.S., Kumar, S., y Sarker, A. (2017). Association mapping unveils favorable alleles for grain iron and zinc concentrations in lentil (*Lens culinaris* subsp. *culinaris*). *PLoS One*, *12*(11), e0188296-e0188296.

Slater, A. T., Cogan, N. O. I., Forster, J. W., Hayes, B. J., y Daetwyler, H. D. (2016). Improving genetic gain with genomic selection in autotetraploid potato. *The Plant Genome*, *9*(3).

Smith, A. B., Cullis, B. R., y Thompson, R. (2005). The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *The Journal of Agricultural Science*, *143*(6), 449-462.

Soller, M. (1978). The use of loci associated with quantitative effects in dairy cattle improvement. *Animal Science*, *27*(2), 133-139.

Soller, M., y Plotkin-Hazan, J. (1977). The use marker alleles for the introgression of linked quantitative alleles. *Theoretical and Applied Genetics*, *51*(3), 133-137.

Sripathi, R., Conaghan, P., Grogan, D., y Casler, M. D. (2018). Modeling genotype×environment correlation structures in long-term multilocation forage yield trials. *Crop Science*, *58*(4), 1447-1457.

Strube, M. J. (1988). Averaging correlation coefficients: influence of heterogeneity and set size. *Journal of Applied Psychology*, *73*(3), 559-568.

Su, G., Brøndum, R. F., Ma, P., Guldbrandtsen, B., Aamand, G. P., y Lund, M. S. (2012). Comparison of genomic predictions using medium-density (∼54,000) and high-density (∼777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. *Journal of Dairy Science*, *95*(8), 4657-4665.

Sukumaran, S., Crossa, J., Jarquin, D., Lopes, M., y Reynolds, M. P. (2017). Genomic prediction with pedigree and genotype× environment interaction in spring wheat grown in South and West Asia, North Africa, and Mexico. *G3: Genes*, *Genomes*, *Genetics*, *7*(2), 481-495.

Tao, Y., Liu, Q., Wang, H., Zhang, Y., Huang, X., Wang, B., Lai, J., Ye, J., Liu, B., y Xu, M. (2013). Identification and fine-mapping of a QTL, qMrdd1, that confers recessive resistance to maize rough dwarf disease. *BMC Plant Biology*, *13*(1), 145.

Thavamanikumar, S., Dolferus, R., y Thumma, B. R. (2015). Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. *G3: Genes, Genomes, Genetics*, *5*(10), 1991-1998.

Tracy, C. A., y Widom, H. (1994). Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, *159*(1), 151-174.

Urrestarazu, J., Muranty, H., Denancé, C., Leforestier, D., Ravon, E., Guyader, A., Guisnel, R., Feugey, L., Aubourg, S., Celton, J. -M., Daccord, N., Dondini, L., Gregori, R., Lateur, M., Houben, P., Ordidge, M., Paprstein, F., Sedlak, J., Nybom, H., Garkava-Gustavsson, L., Troggio, M., Bianco, L., Velasco, R., Poncet, C., Théron, A., Moriya, S., Bink, M. C. A. M., Laurens, F., Tartarini, S., y Durel, C. -E. (2017). Genome-wide association mapping of flowering and ripening periods in apple. *Frontiers in Plant Science*, *8*, 1923.

VanRaden, P. M. (2007). Genomic measures of relationship and inbreeding. *Proceedings Interbull Meeting*, (37), 33-36. Dublin, Ireland.

Veyrieras, J. -B., Goffinet, B., y Charcosset, A. (2007). MetaQTL: a package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinformatics*, *8*(1), 49.

Visscher, P. M., Brown, M. A., McCarthy, M. I., y Yang, J. (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, *90*(1), 7-24.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., y Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *American Journal of Human Genetics*, *101*(1), 5-22.

Voss-Fels, K. P., Cooper, M., y Hayes, B. J. (2019). Accelerating crop genetic gains with genomic selection. *Theoretical and Applied Genetics*, *132*(3), 669-686.

Wang, M., y Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity*, *123*(3), 287-306.

Wang, Q., Tian, F., Pan, Y., Buckler, E. S., y Zhang, Z. (2014). A SUPER powerful method for genome wide association study. *PLoS One*, *9*(9), e107684.

Wang, X., Li, L., Yang, Z., Zheng, X., Yu, S., Xu, C., y Hu, Z. (2016). Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. *Heredity*, *118*, 302–310.

Wang, X., Xu, Y., Hu, Z., y Xu, C. (2018). Genomic selection methods for crop improvement: Current status and prospects. *The Crop Journal*, *6*(4), 330-340.

Wang, X., Yang, Z., y Xu, C. (2015). A comparison of genomic selection methods for breeding value prediction. *Science Bulletin*, *60*(10), 925-935.

Wang, Y., Xu, J., Deng, D., Ding, H., Bian, Y., Yin, Z., Wu, Y., Zhou, B., y Zhao, Y. (2016). A comprehensive meta-analysis of plant morphology, yield, stay-green, and virus disease resistance QTL in maize (*Zea mays* L.). *Planta*, *243*(2), 459-471.

Warburton, M. L., Tang, J. D., Windham, G. L., Hawkins, L. K., Murray, S. C., Xu, W., Boykin, D., Perkins, A., y Williams, W. P. (2015). Genome-wide association mapping of Aspergillus flavus and aflatoxin accumulation resistance in maize. *Crop Science*, *55*(5), 1857-1867.

West, B. T., Welch, K. B., y Galecki, A. T. (2014). *Linear mixed models: a practical guide using statistical software*. New York: Chapman and Hall/CRC.

Wisser, R. J., Balint-Kurti, P. J., y Nelson, R. J. (2006). The genetic architecture of disease resistance in maize: a synthesis of published studies. *Phytopathology*, *96*(2), 120-129.

Wu, X.-L., Gianola, D., Hu, Z.-L., y Reecy, J. M. (2011). Meta-analysis of quantitative trait association and mapping studies using parametric and non-parametric models. *Journal of Biometrics & Biostatistics*, *S1*, 1-9.

Wu, X.-L., y Hu, Z. L. (2012). Meta-analysis of QTL mapping experiments. En S. A. Rifkin (Ed.), *Quantitative Trait Loci (QTL): Methods and Protocols* (pp. 145-171). Totowa, NJ: Humana Press.

Wu, Y., Huang, M., Tao, X., Guo, T., Chen, Z., y Xiao, W. (2016). Quantitative trait loci identification and meta-analysis for rice panicle-related traits. *Molecular Genetics and Genomics*, *291*(5), 1927-1940.

Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J. E., Graef, G. L., Beavis, W. D., Diers, B. W., Song, Q., y Cregan, P. B. (2018). Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. *G3: Genes, Genomes, Genetics*, *8*(2), 519-529.

Xiang, K., Reid, L. M., Zhang, Z.-M., Zhu, X.-Y., y Pan, G.-T. (2012). Characterization of correlation between grain moisture and ear rot resistance in maize by QTL meta-analysis. *Euphytica*, *183*(2), 185-195.

Xu, S. (2008). Quantitative trait locus mapping can benefit from segregation distortion. *Genetics*, *180*(4), 2201-2208.

Xu, Y., Xu, C., y Xu, S. (2017). Prediction and association mapping of agronomic traits in maize using multiple omic data. *Heredity*, *119*(3), 174-184.

Xu, Y., Yang, T., Zhou, Y., Yin, S., Li, P., Liu, J., Xu, S., Yang, Z., y Xu, C. (2018). Genome-wide association mapping of starch pasting properties in maize using single-locus and multi-locus models. *Frontiers in Pant Science*, *9*, 1311.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., y Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*, 565.

Yang, Q., Balint-Kurti, P., y Xu, M. (2017). Quantitative disease resistance: dissection and adoption in maize. *Molecular Plant*, *10*(3), 402-413.

Yang, R.-C. (2007). Mixed-model analysis of crossover genotype–environment interactions. *Crop Science*, *47*(3), 1051-1062.

Yang, Z., Huang, D., Tang, W., Zheng, Y., Liang, K., Cutler, A. J., y Wu, W. (2013). Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled extremes. *PLoS One*, *8*(7).

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., y Holland, J. B. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, *38*(2), 203.

Zhang, P., Liu, X., Tong, H., Lu, Y., y Li, J. (2014). Association mapping for important agronomic traits in core collection of rice (Oryza sativa L.) with SSR markers. *PLoS One*, *9*(10), e111508-e111508.

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

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

Zhao, L., Liu, H. J., Zhang, C. X., Wang, Q. Y., y Li, X. H. (2015). Meta-analysis of constitutive QTLs for disease resistance in maize and its synteny conservation in the rice genome. *Genetics and Molecular Research*, *14*(1), 961-970.

Zhou, X., Carbonetto, P., y Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genetics*, *9*(2), e1003264.

Zhu, C., y Yu, J. (2009). Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics*, *182*(3), 875-888.

Ziegler, A., König, I. R., y Thompson, J. R. (2008). Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, *50*(1), 8-28.

ANEXOS

ANEXO I

RUTINAS DE R UTILIZADAS EN LOS ANÁLISIS ESTADÍSTICOS

RUTINAS DE R PARA AJUSTAR LOS MODELOS GWAS DEL CAPÍTULO 3

Las siguientes rutinas o códigos fueron escritas para ser implementadas en el software R con el paquete *Sommer* (Covarrubias-Pazaran, 2016). A continuación, se presentan a modo de ejemplo, los comandos para ajustar modelos GWAS por ambiente y multiambientales, en ambos casos, con correlaciones genéticas por pedigrí y/o similitud molecular.

MODELO GWAS POR AMBIENTE CON CORRELACIONES GENÉTICAS OBTENIDAS A PARTIR DE INFORMACIÓN DE PEDIGRÍ

 # Con la función library se cargan las funciones contenidas en la biblioteca "sommer"

```
library(sommer)
```

 # La función read.table permite leer una base de datos con extensión .txt y convertirlo en un objeto data.frame.

```
DT<-read.table("DT.txt", sep=" ", header = T, dec=".")
```

#El data.frame es un arreglo de filas y columnas, tendrá una columna que identifique el #genotipo, otra que identifique el número de repetición del genotipo según el diseño #experimental, otra con el valor observado del fenotipo y tantas columnas como marcadores #moleculares haya. Este data.frame tendrá tantas filas como genotipos×repeticiones se #hayan evaluado para un ambiente.

 # Acondicionamiento de la matriz de pedigrí (Kinship) para poder ser introducida en la # función GWAS de la biblioteca Sommer

```
K<- read.table("pedigree.txt", sep="\t", header = T, dec=",")
rownames(K)<-unique(as.factor(DT$Name))
colnames(K)<-unique(as.factor(DT$Name))
K <- as.matrix(K)
```

#Sommer no acepta nombre repetidos, como cada genotipo tiene su repetición, es necesario #unificar el nombre del genotipo con la función unique. Es necesario nombrar las filas y #columnas de la matriz de parentesco (K) con los nombres de los genotipos provistos en el #data.frame DT. Debido a que K es una matriz cuadrada, es necesario aplicar la función #unique tanto en las filas como en las columnas. Finalmente, debe aplicarse la función #as.matrix para convertirla en un arreglo de número de clase matriz.

 # Función GWAS para ajustar un modelo GWAS por ambiente usando el pedigrí como
 # correlación genética en Sommer

```
M1_NR <- GWAS(Yield~1,random= ~ vs(Name, Gu=K),
              method= "NR",rcov= ~ units, M=M,
              gTerm = "u:Name", data=DT)
```

 # Valores obtenidos a partir del modelo GWAS ajustado

 # Estima las componentes de varianza del modelo ajustado

```
M1_NR$sigma
```

 # Estima la varianza genotípica

```
Vg_M1_NR<- round(M1_NR$sigma[[1]],3)
```

 # Estima la varianza residual

```
Ve_M1_NR <- round(M1_NR$sigma[[2]],3)
```

 # Se obtiene el criterio de información de Akaike (AIC)

```
M1_NR$AIC
```

 # Se obtiene el criterio de información bayesiano (BIC)

```
M1_NR$BIC
```

 # Estima los errores estándar de las componentes de varianza

```
SE_M1<- round(sqrt(diag(M1_NR$sigmaSE)),4)
```

#Obtiene los BLUP de genotipo

```
BLUP_M1_NR<- randef(M1_NR)
```

#Obtiene los efectos de marcador

```
beta_M1_NR<- as.data.frame(M1_NR$scores[1,1:1109])
```

MODELO GWAS POR AMBIENTE CON CORRELACIONES GENÉTICAS OBTENIDAS A PARTIR DE INFORMACIÓN DE MARCADORES MOLECULARES

 # Con la función library se cargan las funciones contenidas en la biblioteca "sommer"

```
library(sommer)
```

#Base de datos

```
DT<-read.table("DT.txt", sep=" ", header = T, dec=".")
```

#El data.frame es un arreglo de filas y columnas, tendrá una columna que identifique el #genotipo, otra que identifique el número de repetición del genotipo según el diseño #experimental, otra con el valor observado del fenotipo y tantas columnas como marcadores #moleculares haya. Este data.frame tendrá tantas filas como genotipos×repeticiones se #hayan evaluado para un ambiente.

 # Selección y colocación del nombre de los genotipos a las filas de la matriz de
 # marcadores moleculares

```
M<- DT[1:599,3:1281]
rownames(M) <- unique(as.factor(DT$Name))
M
```

#Desde el data.frame DT, se toma el perfil molecular de cada genotipo y todas las columnas #que contengan la información de los marcadores moleculares. En las filas del DT, cada #genotipo está repetido tantas veces como repeticiones se hayan realizado. Por ello #es necesario aplicar la función unique.

 # La función A.mat estima las relaciones aditivas a partir del cálculo de similitud
 # molecular entre genotipos a través de la matriz de marcadores moleculares

```
A <- A.mat(M)
```

 # Función GWAS para ajustar un modelo GWAS por ambiente usando la similitud molecular como correlación genética

```
M2_NR <- GWAS(Yield~1,random= ~ vs(Name, Gu=A),
              method = "NR",rcov= ~ units, M=M,
              gTerm = "u:Name",data=DT)
```

 # Valores obtenidos a partir del modelo GWAS ajustado

 # Estima las componentes de varianza estimadas por el modelo

```
M2_NR$sigma
```

 # Estima la varianza genotípica

```
Vg_M2_NR<- round(M2_NR$sigma[[1]],3)
```

 # Estima la varianza residual

```
Ve_M2_NR<- round(M2_NR$sigma[[2]],3)
```

 # Se obtiene el criterio de información de Akaike (AIC)

```
M2_NR$AIC
```

 # Se obtiene el criterio de información bayesiano (BIC)

```
M2_NR$BIC
```

 # Estima los errores estándar de las componentes de varianza

```
SE_M2<- round(sqrt(diag(M2_NR$sigmaSE)),4)
```

#Obtiene los BLUP de genotipo

```
BLUP_M2_NR<- randef(M2_NR)
```

#Obtiene los efectos de marcador

```
beta_M2_NR<- as.data.frame(M2_NR$scores[1,1:1109])
```

MODELO GWAS MULTIAMBIENTAL CON CORRELACIONES GENÉTICAS OBTENIDAS A PARTIR DE INFORMACIÓN DE PEDIGRÍ

 # Con la función library se cargan las funciones contenidas en la biblioteca "sommer"

```
library(sommer)
```

#Base de datos

```
DT<-read.table("DT.txt", sep=" ", header = T, dec=".")
```

El data.frame es un arreglo de filas y columnas, tendrá una columna que identifique el
ambiente, otra el genotipo, otra columna que identifique el número de repetición del
genotipo según el diseño experimental, otra con el valor observado del fenotipo y tantas
columnas como marcadores moleculares haya. Este data.frame tendrá tantas filas como
genotipos×repeticiones×ambientes se hayan evaluado.

 # Acondicionamiento de la matriz de pedigrí (Kinship) para poder ser introducida en la # función GWAS de la biblioteca Sommer

```
K<- read.table("pedigree.txt", sep="\t", header = T, dec=",")
rownames(K)<-unique(as.factor(DT$Name))
colnames(K)<-unique(as.factor(DT$Name))
K <- as.matrix(K)
```

 # Indica la cantidad de genotipos

```
nind <- length(unique(DT$Name))
```

 # Indica la cantidad de ambientes

```
nenv <- length(unique(DT$Env))
```

 # Realiza la matriz de correlaciones entre ambientes

```
Rho<- diag(nenv)
colnames(Rho) <- rownames(Rho) <- unique(DT$Env)
Rho
```

 # Realiza el producto Kronecker entre las matrices de correlaciones y la matriz de # pedigrí

```
Rho_K<- kronecker(Rho,K,make.dimnames=TRUE)
```

 # Función GWAS para ajustar un modelo GWAS multiambiental usando el pedigrí como correlación genética en Sommer

```
M3_NR_Rho_K <- GWAS(Yield~Env,random= ~ vs(Name, Gu=K) +
vs(Env:Name, Gu=Rho_K),rcov= ~ units,M=M, gTerm = "u:Name",
method = "NR",data=DT)
```

```
# Valores obtenidos a partir del modelo GWAS ajustado
    # Estima las componentes de varianza estimadas por el modelo
    M3_NR_Rho_K$sigma
    # Estima la varianza genotípica
    Vg_M3_NR_Rho_K<- round(M3_NR_Rho_K$sigma[[1]],3)
    # Estima la varianza de interacción genotipo-ambiente
    Vge_M3_NR_Rho_K<- round(M3_NR_Rho_K$sigma[[2]],3)
    # Estima la varianza residual
    Ve_M3_NR_Rho_K<- round(M3_NR_Rho_K$sigma[[3]],3)
    # Se obtiene el criterio de información de Akaike (AIC)
    M3_NR_Rho_K$AIC
    # Se obtiene el criterio de información bayesiano (BIC)
    M3_NR_Rho_K$BIC
    # Estima los errores estándar de las componentes de varianza
    SE_M3_NR_Rho_K<- round(sqrt(diag(M3_NR_Rho_K$sigmaSE)),4)
    #Obtiene los BLUP de genotipo
    BLUP_G_M3_NR_Rho_K<- randef(M3_NR_Rho_K)[[1]]
    #Obtiene los BLUP de interacción genotipo-ambiente
    BLUP_GE_M3_NR_Rho_K<- randef(M3_NR_Rho_K)[[2]]
    #Obtiene los efectos de marcador
    beta_M3_NR_Rho_K<-as.data.frame(M3_NR_Rho_K$scores[1,1:1109])
```

MODELO GWAS MULTIAMBIENTAL CON CORRELACIONES GENÉTICAS OBTENIDAS A PARTIR DE INFORMACIÓN DE MARCADORES MOLECULARES

```
# Con la función library se cargan las funciones contenidas en la biblioteca "sommer"
library(sommer)
#Base de datos
DT<-read.table("DT.txt", sep=" ", header = T, dec=".")
# El data.frame es un arreglo de filas y columnas, tendrá una columna que identifique el
# ambiente, otra el genotipo, otra columna que identifique el número de repetición del
# genotipo según el diseño experimental, otra con el valor observado del fenotipo y tantas
# columnas como marcadores moleculares haya. Este data.frame tendrá tantas filas como
# genotipos×repeticiones×ambientes se hayan evaluado.
```

 # Selección y colocación del nombre de los genotipos a las filas de la matriz de
 # marcadores moleculares

```
M<- DT[1:599,3:1281]
rownames(M) <- unique(as.factor(DT$Name))
M
```

#Desde el data.frame DT, se toma el perfil molecular de cada genotipo y todas las columnas #que contengan la información de los marcadores moleculares. En las filas del DT, cada #genotipo está repetido tantas veces como repeticiones se hayan realizado. Por ello #es necesario aplicar la función unique.

 # La función A.mat estima las relaciones aditivas a partir del cálculo de similitud
 # molecular entre genotipos a través de la matriz de marcadores moleculares

```
A <- A.mat(M)
```

 # Indica la cantidad de genotipos

```
nind <- length(unique(DT$Name))
```

 # Indica la cantidad de ambientes

```
nenv <- length(unique(DT$Env))
```

 # Realiza la matriz de correlaciones entre ambientes

```
Rho<- diag(nenv)
colnames(Rho) <- rownames(Rho) <- unique(DT$Env)
Rho
```

 # Realiza el producto Kronecker entre las matrices de correlaciones y la matriz de
 # similitud molecular

```
Rho_A<- kronecker(Rho,A,make.dimnames=TRUE)
```

 # Función GWAS para ajustar un modelo GWAS multiambiental usando la similitud
molecular como correlación genética

```
M4_NR_Rho_A<-GWAS(Yield~Env,random=~vs(Name,Gu=A)+vs(Env:Name,Gu=Rho_A),
rcov= ~ units, M=M, gTerm = "u:Name",method = "NR",data=DT)
```

 # Valores obtenidos a partir del modelo GWAS ajustado

 # Estima las componentes de varianza estimadas por el modelo

```
M4_NR_Rho_A$sigma
```

 # Estima la varianza genotípica

```
Vg_M4_NR_Rho_A<- round(M4_NR_Rho_A$sigma[[1]],3)
```

 # Estima la varianza de interacción genotipo-ambiente

```
Vge_M4_NR_Rho_A<- round(M4_NR_Rho_A$sigma[[2]],3)
```

```
# Estima la varianza residual
Ve_M4_NR_Rho_A<- round(M4_NR_Rho_A$sigma[[3]],3)
# Se obtiene el criterio de información de Akaike (AIC)
M4_NR_Rho_A$AIC
# Se obtiene el criterio de información bayesiano (BIC)
M4_NR_Rho_A$BIC
# Estima los errores estándar de las componentes de varianza
SE_M4_NR_Rho_A<- round(sqrt(diag(M4_NR_Rho_A$sigmaSE)),4)
#Obtiene los BLUP de genotipo
BLUP_G_M4_NR_Rho_A<- randef(M4_NR_Rho_A)[[1]]
#Obtiene los BLUP de interacción genotipo-ambiente
BLUP_GE_M4_NR_Rho_A<- randef(M4_NR_Rho_A)[[2]]
#Obtiene los efectos de marcador
beta_M4_NR_Rho_A<-as.data.frame(M4_NR_Rho_A$scores[1,1:1109])
```

RUTINAS DE R PARA REALIZAR META-ANÁLISIS Y SU VISUALIZACIÓN MEDIANTE *FOREST PLOT* PARA CORRELACIONES DEL MÉRITO GENÉTICO ABORDADAS EN EL CONTEXTO DE SELECCIÓN GENÓMICA EN EL CAPÍTULO 4

Las siguientes rutinas o códigos fueron escritas para ser implementadas en el software R (R Core Team, 2020) con el paquete meta. A continuación, se presentan a modo de ejemplo, los comandos para ajustar el modelo de efectos aleatorios y el análisis por subgrupos para la cantidad de genotipos, la densidad de marcadores moleculares y los métodos de estimación para las especies de maíz y trigo. Además, la visualización de los resultados a través del *Forest Plot.*

```
# Con la función library se cargan las funciones contenidas en las bibliotecas "meta"
library(meta)

#Base de datos
datos <- read.table("Datos_Cap3.txt", sep=" ", header = T, dec=".")
```

 # Función metacor para ajustar un modelo de efectos aleatorios para la correlación
 # transformada con el z de Fisher

```
Corr_MEA_Z <- metacor(datos$r,datos$n, studlab= Estudios,
                      sm="ZCOR",comb.fixed=FALSE,
                      comb.random=TRUE,method.tau="DL",
                      backtransf=TRUE,data = datos)
```

 # Realiza el Forest Plot para el objeto Corr_MEA_Z en formato .pdf

```
pdf("Corr_MEA_Z.pdf", width = 10, height = 20, paper = "special")
forest(Corr_MEA_Z,fs.study.labels=8,fs.study=8,
       ff.study=1, col.study="black",
       col.inside="black",fs.hetstat=8,fs.test.overall=8,
       fontsize=12,spacing=0.7,fs.heading=8,
       fs.random.labels=8,fs.random=8,leftcols="studlab",
       col.diamond.random="blue",xlim= c(-2,2),
       lty.random=5, col.random="blue",hetstat=TRUE)
dev.off()
```

 # Realiza el análisis por subgrupos para especie

```
ASub_Especie_Z <-  metacor(datos$r,datos$n,sm="ZCOR",
                   method.tau="DL", studlab = paste(Estudios),
                   byvar= datos$Especie, print.byvar=F,
                   tau.common=FALSE,backtransf=TRUE,
                   comb.fixed=FALSE,comb.random=TRUE,
                   data=datos)
```

 # Realiza el Forest Plot para el objeto ASub_Especie_Z en formato .pdf

```
pdf("ASub_Especie_Z.pdf", width = 10, height = 25, paper = "special")
forest(ASub_Especie_Z,fs.study.labels=8,fs.study=8,ff.study=1,
       col.study="black",col.inside="black",fs.hetstat=8,
       fs.test.overall=8,fontsize=12,spacing=0.7,fs.heading=8,
       xlim= c(-2,2),leftcols="studlab",fs.random.labels=8,
       fs.random=8,col.diamond.random="blue",lty.random=5,
       col.random="blue",hetstat= FALSE,print.byvar = FALSE)
dev.off()
```

 # Realiza el análisis por subgrupos para genotipo

```
ASub_genotipo_Z <-  metacor(datos$r,datos$n,sm="ZCOR",
                    method.tau="DL", studlab = paste(Estudios),
                    byvar= datos$Subgrupo_n, print.byvar=F,
                    tau.common=FALSE,backtransf=TRUE,
                    comb.fixed=FALSE, comb.random=TRUE,data=datos)
```

 # Realiza el Forest Plot para el objeto ASub_genotipo_Z en formato .pdf

```
pdf("ASub_genotipo_Z.pdf", width = 10, height = 25, paper = "special")
forest(ASub_genotipo_Z,fs.study.labels=8,fs.study=8,ff.study=1,
       col.study="black",col.inside="black",fs.hetstat=8,
       fs.test.overall=8,fontsize=12,spacing=0.7,fs.heading=8,
       xlim= c(-2,2),leftcols="studlab",fs.random.labels=8,
       fs.random=8,col.diamond.random="blue",lty.random=5,
       col.random="blue",hetstat= FALSE,print.byvar = FALSE)
dev.off()
```

 # Realiza el análisis por subgrupos para la densidad de marcadores moleculares

```
ASub_Marcadores_Z <-  metacor(datos$r,datos$n,sm="ZCOR",
                      method.tau="DL",studlab = paste(Estudios),
                      byvar= datos$Subgrupo_MM,tau.common=FALSE,
                      backtransf=TRUE,comb.fixed=FALSE,
                      comb.random=TRUE,data=datos)
```

 # Realiza el Forest Plot para el objeto ASub_Marcadores_Z en formato .pdf

```
pdf("ASub_Marcadores_Z.pdf", width = 10, height = 25,paper = "special")
forest(ASub_Marcadores_Z,fs.study.labels=8,fs.study=8,
       ff.study=1,col.study="black",col.inside="black",
       fs.hetstat=8,fs.test.overall=8,fontsize=12,
       spacing=0.7,fs.heading=8, xlim= c(-2,2),
       leftcols="studlab",fs.random.labels=8,
       fs.random=8,col.diamond.random="blue",lty.random=5,
       col.random="blue",hetstat= FALSE,print.byvar = FALSE)
dev.off()
```

 # Realiza el análisis por subgrupos para métodos de estimación usados en SG

```
ASub_ME_Z<- metacor(datos$r,datos$n,sm="ZCOR",
                    method.tau="DL", studlab = paste(Estudios),
                    byvar= datos$Método_Estimación,tau.common=FALSE,
                    backtransf=TRUE,comb.fixed=FALSE,comb.random=TRUE,
                    data=datos)
```

 # Realiza el Forest Plot para el objeto ASub_ME_Z en formato .pdf

```
pdf("ASub_ME_Z.pdf", width = 10, height = 25, paper = "special")
forest(ASub_ME_Z,fs.study.labels=8,fs.study=8,ff.study=1,
       col.study="black",col.inside="black",fs.hetstat=8,
       fs.test.overall=8,fontsize=12,spacing=0.7,
       fs.heading=8,xlim= c(-2,2),leftcols="studlab",
       fs.random.labels=8,fs.random=8,col.diamond.random="blue",
       lty.random=5, col.random="blue",
       hetstat= FALSE,print.byvar = FALSE)
dev.off()
```

RUTINAS DE R PARA REALIZAR META-ANÁLISIS Y SU VISUALIZACIÓN MEDIANTE *FOREST PLOT* PARA DIFERENCIAS DE RIESGOS EN EL CONTEXTO DE IDENTIFICACIÓN DE QTL PARA RESISTENCIA/TOLERANCIA DE VIRUS EN MAÍZ EN EL CAPÍTULO 5

Las siguientes rutinas o códigos fueron escritas para ser implementadas en el software R (R Core Team, 2020) con el paquete meta. A continuación, se presentan a modo de ejemplo, los comandos para ajustar el modelo de efectos aleatorios y el análisis por subgrupos para la variable generación. Esto se puede hacer para cada uno de los cromosomas del maíz, la visualización de los resultados es a través del *Forest Plot*.

 # Con la función library se cargan las funciones contenidas en la biblioteca "sommer"

```
library(meta)
```

 # La función read.table permite leer una base de datos con extensión .txt y convertirlo en un objeto data.frame.

```
datos <- read.table("Cromosoma.txt", sep=" ", header = T, dec=".")
```

 # Función metabin para ajustar un modelo de efectos aleatorios para la diferencia de riesgos

```
FIT_RD_MEA <- metabin(Ee, Ne, Ec, Nc, studlab = Papers ,
                      sm="RD",comb.fixed=FALSE,comb.random= TRUE,
                      method="I",data=datos)
```

 # Realiza el Forest Plot para el objeto FIT_RD_MEA y lo guarda en formato .pdf

```
pdf("ForestPlot_MEA.pdf",width = 10 ,height = 10)
forest(FIT_RD_MEA ,xlim= c(-2,2),
       hetstat= TRUE,fs.study.labels=9,
       fs.study=9,ff.study=9,fs.hetstat=9,
       fs.test.overall=9,fontsize=12)
dev.off()
```

 # Realiza el análisis por subgrupos para la variable generación

```
ASub_Ge <- metabin(Ee, Ne, Ec, Nc, sm="RD", method="I",
                   studlab = paste(Papers),byvar= Generación,
                   print.byvar=F,comb.fixed=FALSE,
                   comb.random= TRUE,data=datos)
```

 # Realiza el Forest Plot para el objeto ASub_Ge y se visualiza en formato .pdf

```
pdf("ForestPlot_Análisis_Sub_Gen.pdf",width = 10,height = 10)
forest(ASub_Ge,fs.study.labels=9,fs.study=9,ff.study=9,
       xlab= "Análisis por subgrupos para Generación",
       xlim= c(-2,2),hetstat= TRUE,fs.hetstat=9,
       fs.test.overall=9,fontsize=12)
dev.off()
```

ANEXO II

VISUALIZACIÓN DEL META-ANÁLISIS EN ESTUDIOS DE SELECCIÓN GENÓMICA

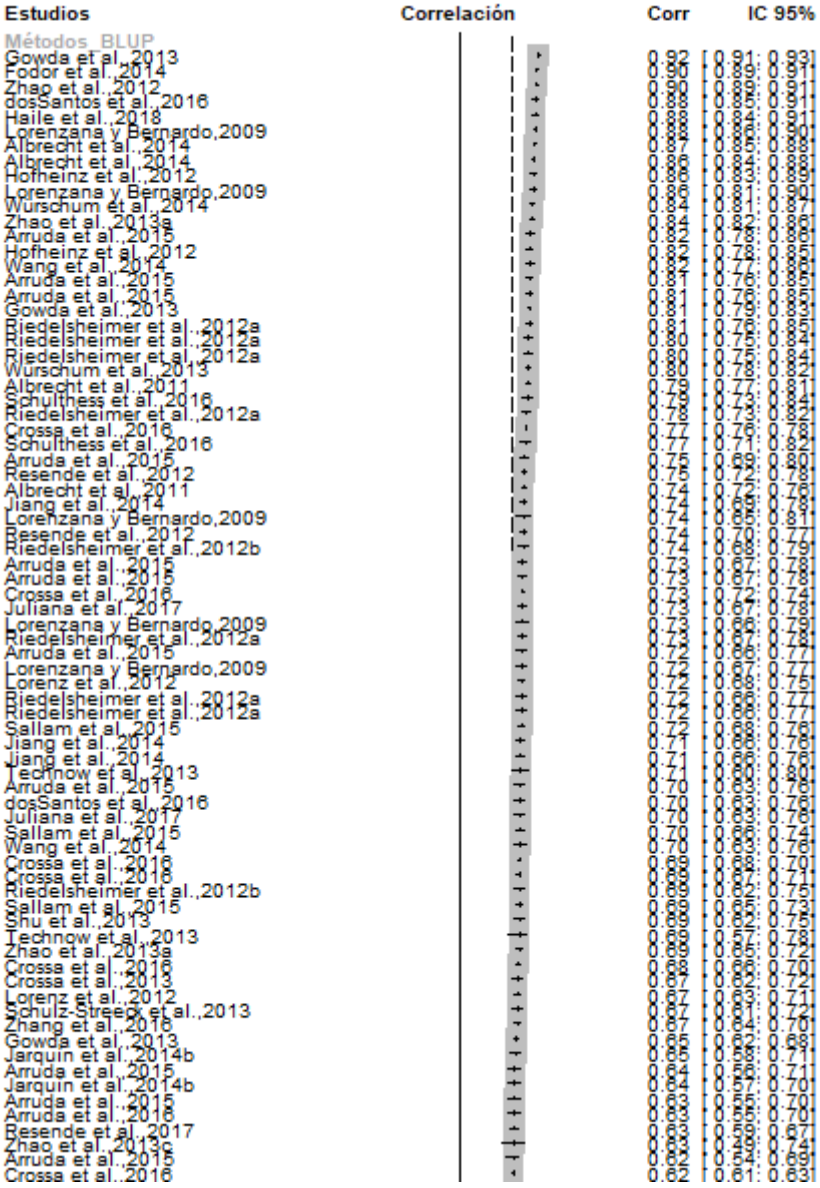

Figura 4.4. *Forest Plot* de la eficiencia de SG para los métodos de estimación "Métodos BLUP" y "Otros" en todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos (Métodos BLUP y Otros), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada método de estimación.

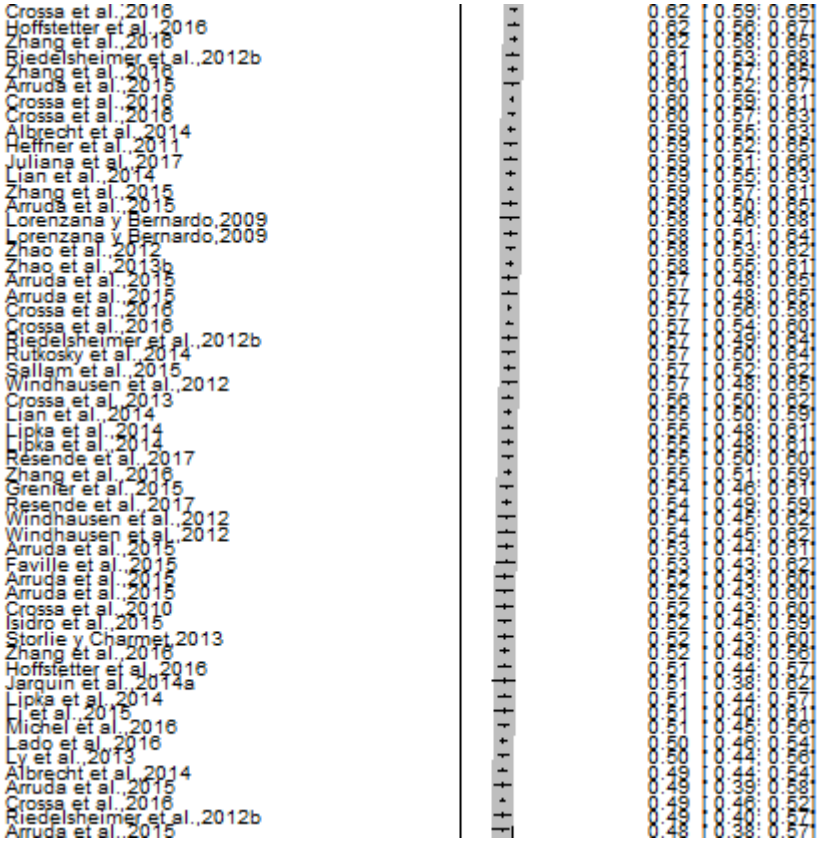

Figura 4.4. *Forest Plot* de la eficiencia de SG para los métodos de estimación “Métodos BLUP” y “Otros” en todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos (Métodos BLUP y Otros), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada método de estimación. Continuación.

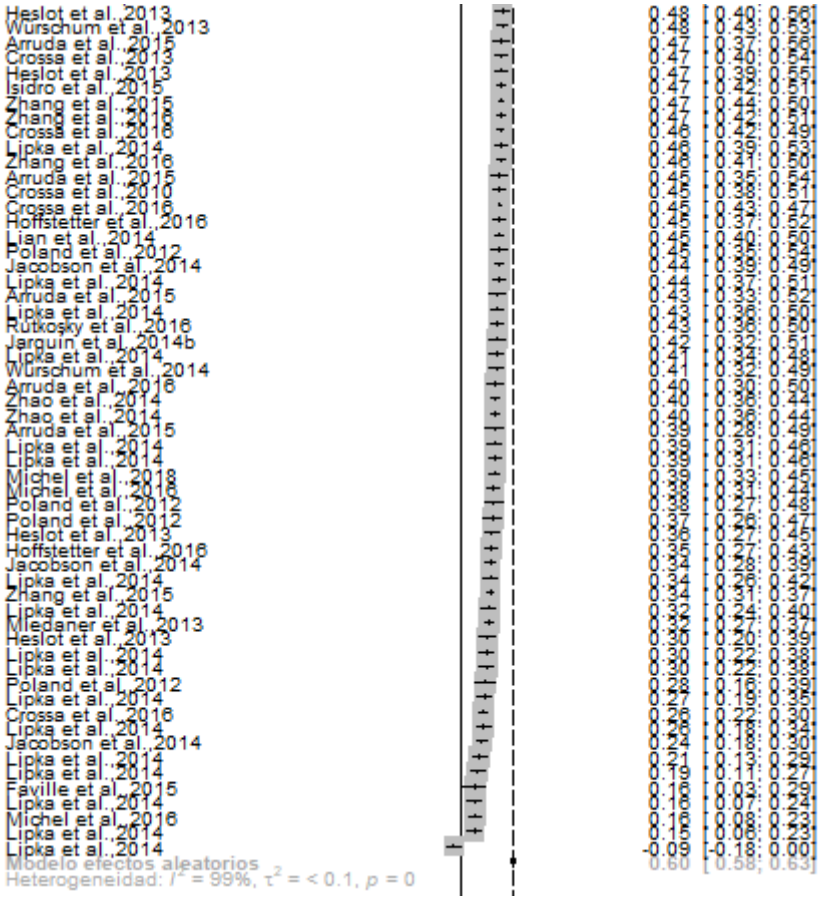

Figura 4.4. *Forest Plot* de la eficiencia de SG para los métodos de estimación “Métodos BLUP” y “Otros” en todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos (Métodos BLUP y Otros), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada método de estimación. Continuación.

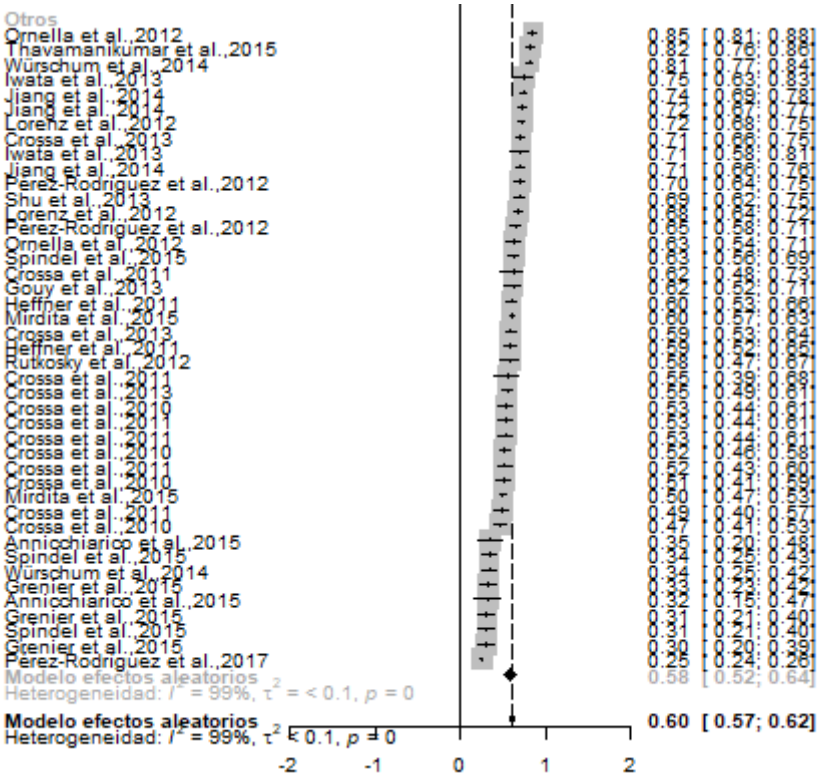

Figura 4.4. *Forest Plot* de la eficiencia de SG para los métodos de estimación "Métodos BLUP" y "Otros" en todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos (Métodos BLUP y Otros), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada método de estimación. Continuación.

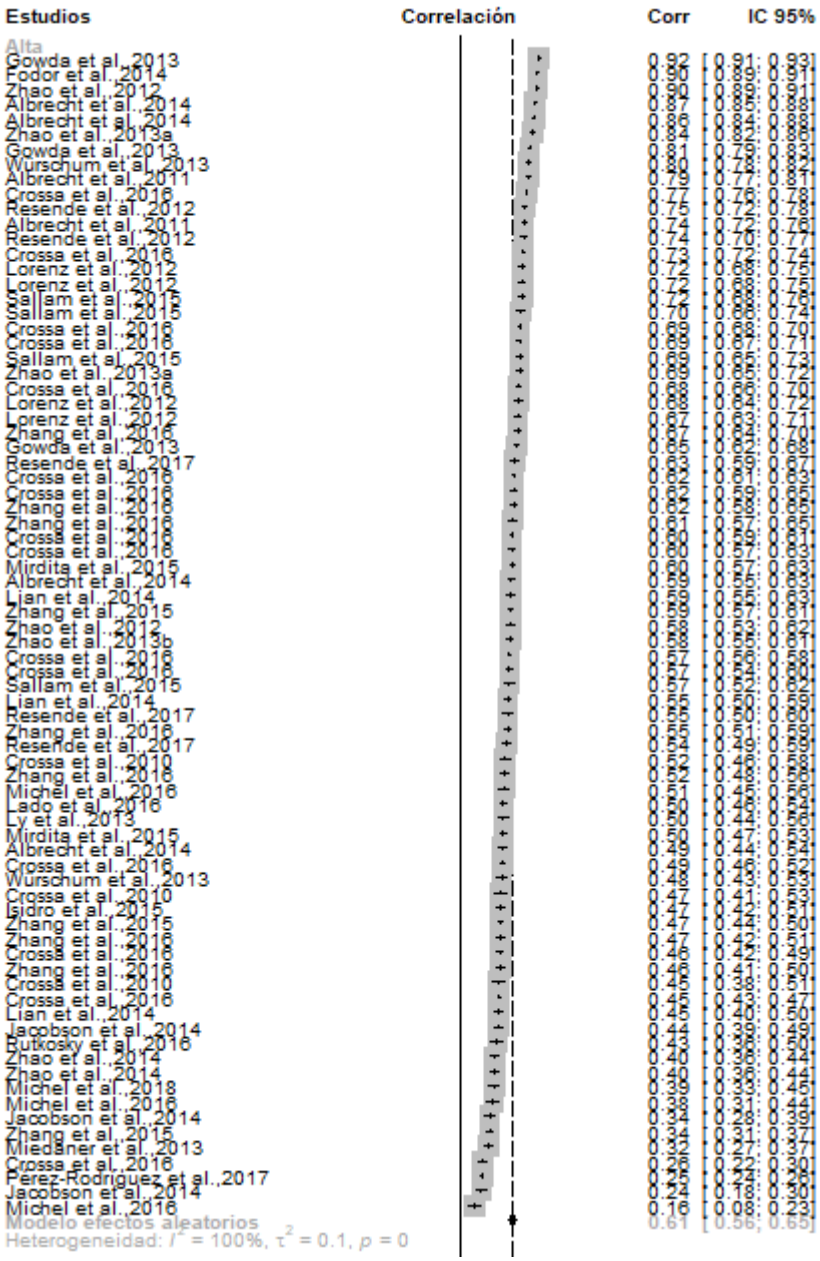

Figura 4.5. *Forest Plot* para distintas cantidades de genotipos categorizadas en: baja (menos de 289), media (entre 289 y 515) y alta (mayor a 515) para estudios primarios de todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de cantidad de genotipos (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares.

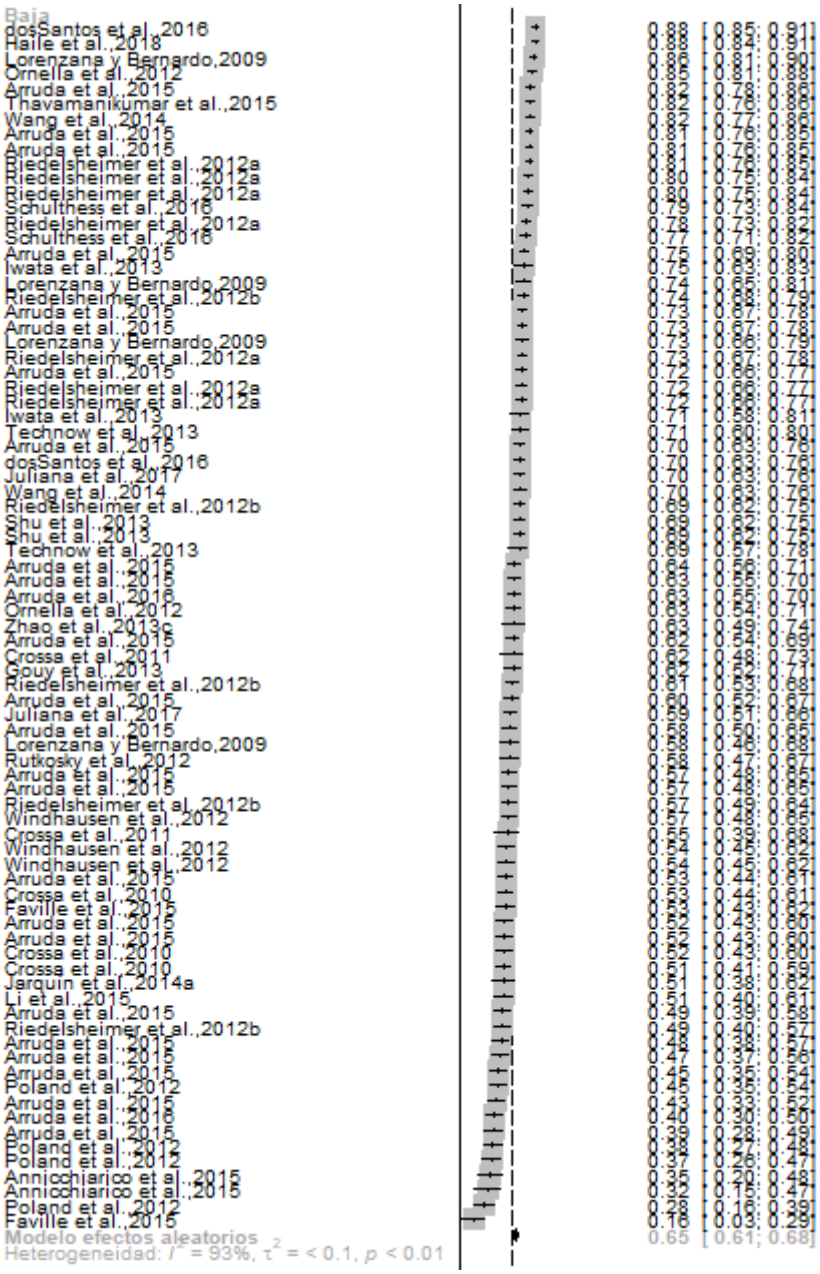

Figura 4.5. *Forest Plot* para distintas cantidades de genotipos categorizadas en: baja (menos de 289), media (entre 289 y 515) y alta (mayor a 515) para estudios primarios de todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de cantidad de genotipos (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares. Continuación.

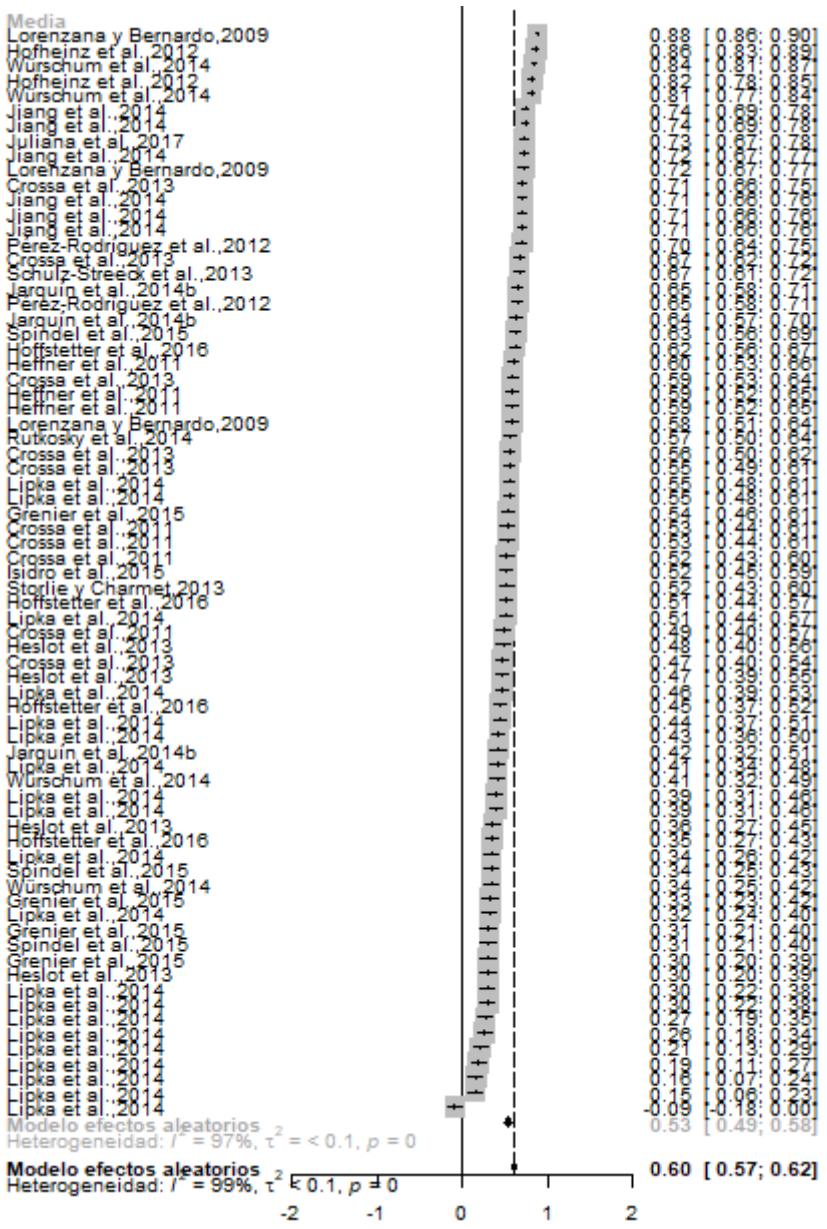

Figura 4.5. *Forest Plot* para distintas cantidades de genotipos categorizadas en: baja (menos de 289), media (entre 289 y 515) y alta (mayor a 515) para estudios primarios de todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de cantidad de genotipos (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares. Continuación.

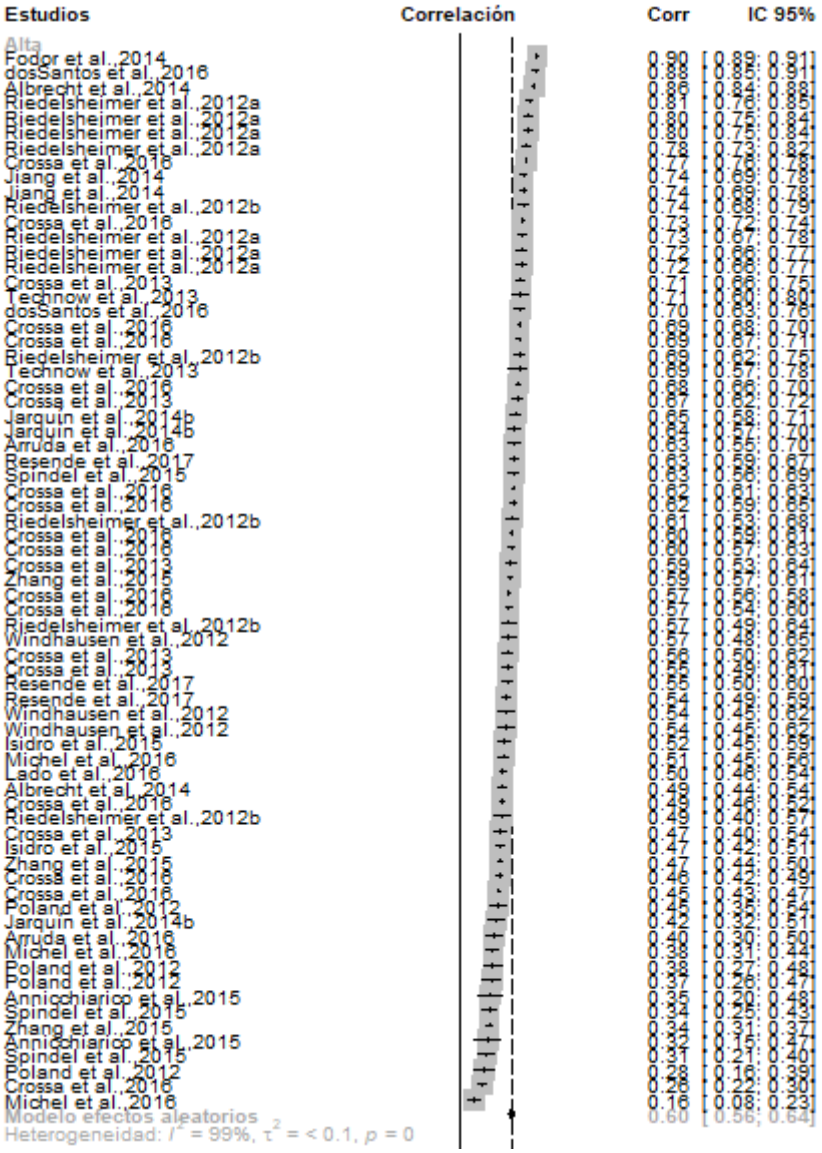

Figura 4.6. *Forest Plot* de la eficiencia de SG para distintas densidades de marcadores moleculares categorizadas en: baja (menos de 1.700), media (entre 1.700 y 17.000) y alta densidad de marcadores moleculares, mayor a 17.000 para estudios primarios de todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de densidad de marcadores moleculares (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares.

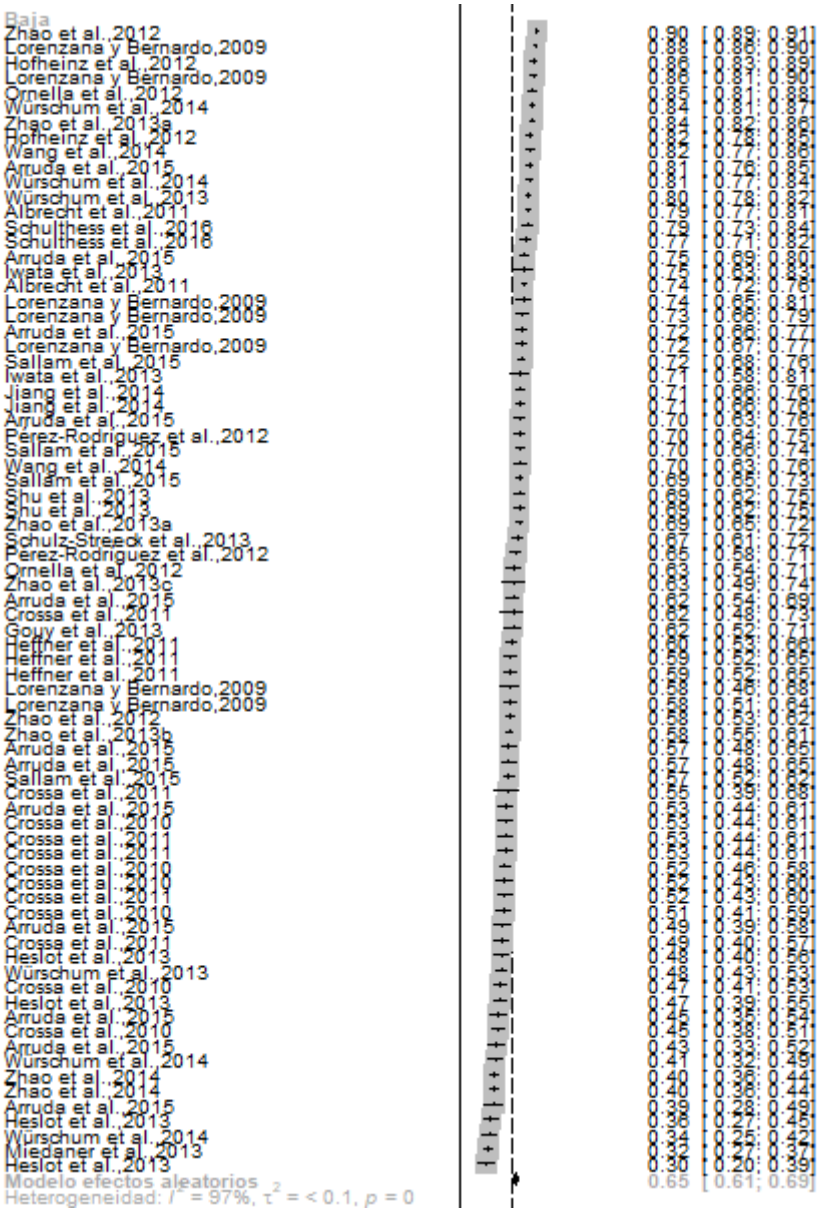

Figura 4.6. *Forest Plot* de la eficiencia de SG para distintas densidades de marcadores moleculares categorizadas en: baja (menos de 1.700), media (entre 1.700 y 17.000) y alta densidad de marcadores moleculares, mayor a 17.000 para estudios primarios de todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de densidad de marcadores moleculares (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares. Continuación.

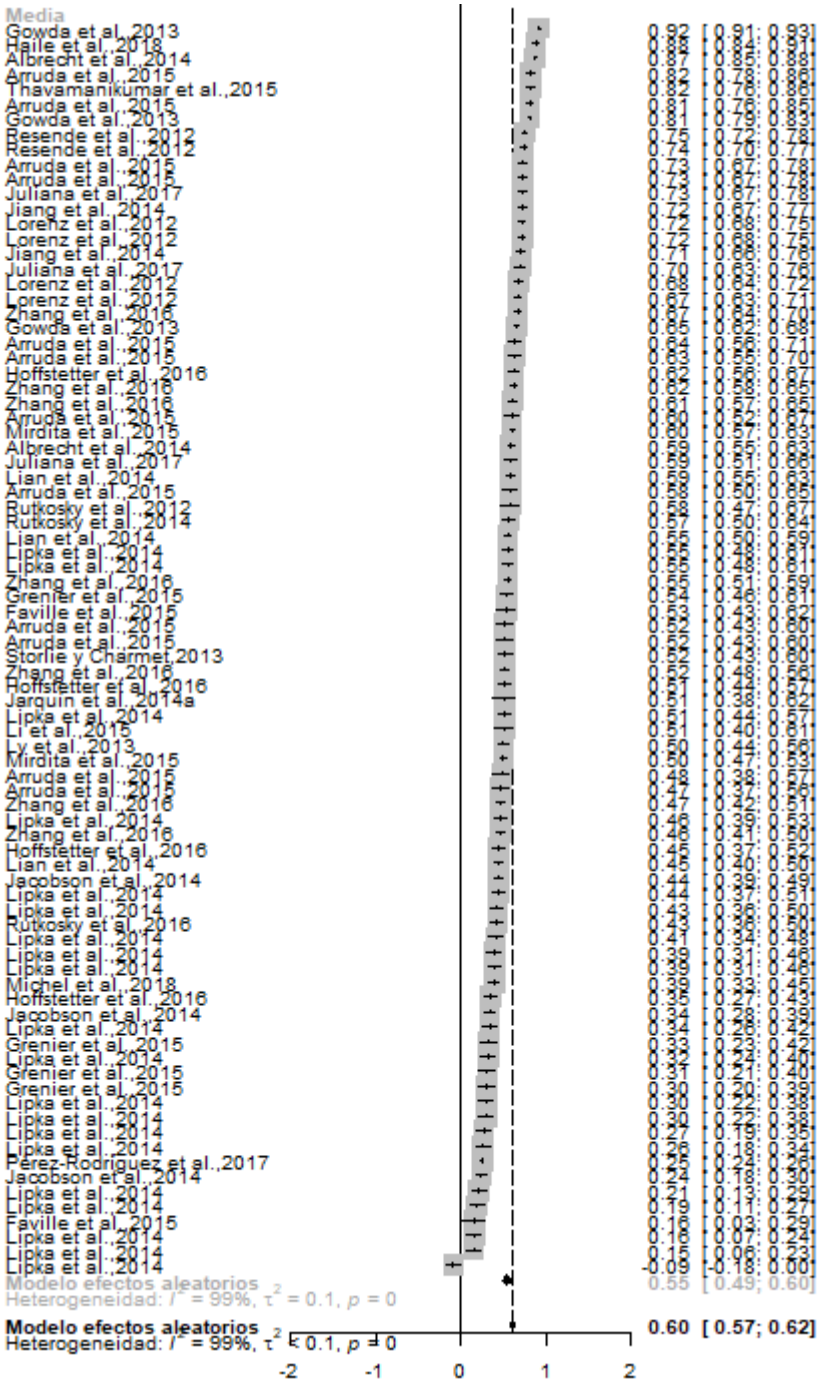

Figura 4.6. *Forest Plot* de la eficiencia de SG para distintas densidades de marcadores moleculares categorizadas en: baja (menos de 1.700), media (entre 1.700 y 17.000) y alta densidad de marcadores moleculares, mayor a 17.000 para estudios primarios de todas las especies. El modelo de meta-análisis ajustado fue un modelo de efectos aleatorios por subgrupos de densidad de marcadores moleculares (Alta, Baja y Media), contemplando de esta forma la heterogeneidad entre estudios primarios y entre grupos. Las correlaciones se presentan ordenadas de mayor a menor dentro de cada categoría de densidad de marcadores moleculares. Continuación.