COLEGIO DE POSTGRADUADOS

INSTITUCIÓN DE ENSEÑANZA E INVESTIGACIÓN
EN CIENCIAS AGRÍCOLAS

CAMPUS MONTECILLO

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

UNA PRUEBA ESTADÍSTICA PARA GWAS CONSIDERANDO LA NO INDEPENDENCIA DE LOS BLUP

Magin Zúñiga Estrada

T E S I S

PRESENTADA COMO REQUISITO PARCIAL PARA
OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

MONTECILLO,TEXCOCO, ESTADO DE MÉXICO
2019

CARTA DE CONSENTIMIENTO DE USO DE LOS DERECHOS DE AUTOR Y DE LAS REGALIAS COMERCIALES DE PRODUCTOS DE INVESTIGACION

En adición al beneficio ético, moral y académico que he obtenido durante mis estudios en el Colegio de Postgraduados, el que suscribe Magin Zúñiga Estrada, Alumno (a) de esta Institución, estoy de acuerdo en ser participe de las regalías económicas y/o académicas, de procedencia nacional e internacional, que se deriven del trabajo de investigación que realicé en esta institución, bajo la dirección del Profesor Sergio Pérez Elizalde, por lo que otorgo los derechos de autor de mi tesis Una prueba estadística para GWAS considerando la no independencia de los BLUP

y de los productos de dicha investigación al Colegio de Postgraduados. Las patentes y secretos industriales que se puedan derivar serán registrados a nombre del colegio de Postgraduados y las regalías económicas que se deriven serán distribuidas entre la Institución, El Consejero o Director de Tesis y el que suscribe, de acuerdo a las negociaciones entre las tres partes, por ello me comprometo a no realizar ninguna acción que dañe el proceso de explotación comercial de dichos productos a favor de esta Institución.

Montecillo, Mpio. de Texcoco, Edo. de México, a 15 de Julio de 2019

Firma del
Alumno (a)

Sergio Pérez Elizalde
Vo. Bo. del Consejero o Director de Tesis

La presente tesis titulada: **UNA PRUEBA ESTADÍSTICA PARA GWAS CONSIDERANDO LA NO INDEPENDENCIA DE LOS BLUP**, realizada por el alumno: **Magin Zúñiga Estrada**, bajo la dirección del Consejo Particular indicado ha sido aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

DOCTOR EN CIENCIAS

SOCIOECONOMÍA, ESTADÍSTICA E INFORMÁTICA
ESTADÍSTICA

CONSEJO PARTICULAR

CONSEJERO

Dr. Sergio Pérez Elizalde

DIRECTOR DE TESIS

Dr. Juan Andrés Burgueño Ferreira

ASESOR

Dr. José Crossa Hiriart

ASESOR

Dr. José Andrés Christen Gracia

ASESOR

Dr. Juan Manuel González Camacho

Montecillo, Texcoco, México, Julio de 2019

UNA PRUEBA ESTADÍSTICA PARA GWAS CONSIDERANDO LA NO INDEPENDENCIA DE LOS BLUP

Magin Zúñiga Estrada, D. en C.

Colegio de Postgraduados, 2019.

RESUMEN

En el presente trabajo se desarrolló una prueba estadística para análisis de asociación aplicada a plantas basada en el *BLUP*, calculado a partir de un modelo lineal mixto (MLM), y sus propiedades distribucionales. La prueba se ensayó utilizando simulación. Los escenarios simulados se determinaron de acuerdo a: número de *SNPs* a probar, número de *SNPs* con efecto no nulo, número de genotipos, tamaño de bloque, magnitud de varianza genética y magnitud de correlación entre residuales asumiendo una estructura de correlación AR(1). En total se simularon 204 escenarios con 1000 iteraciones cada uno. Los resultados de 180 escenarios fueron comparados con otro método de estudio de asociación implementado en *PLINK*. La comparación se realizó utilizando dos indicadores, la TDV qué es la tasa de detecciones verdaderas y la TDF qué es la tasa de detecciones falsas. Cada indicador fue analizado utilizando *ANOVA* donde se obtuvo que ambos métodos son estadísticamente diferentes para TDV y TDF. El método implementado en *PLINK*resultó tener mejor comportamiento para TDV, sin embargo, cuando el tamaño del efecto fue lo suficientemente grande, la propuesta obtuvo valores de TDV cercanos a uno. Con respecto a TDF, el método propuesto se mantuvo, en todos los casos, por debajo del umbral de 0.05 y el método implementado en *PLINK* tuvo un comportamiento poco deseado ya que obtuvo valores por arriba de 0.2. Los últimos 24 escenarios confirman que la propuesta tiene un comportamiento deseado para TDV cuando el tamaño del efecto es lo suficientemente grande. En conclusión, la prueba estadística, en general, muestra un desempeño deseado en el sentido de que controla de forma óptima la tasa de asociaciones espurias y detecta *SNPs* con tamaños de efectos grandes, lo que arroja alta certeza de estar detectando asociaciones verdaderas.

Palabras clave: Prueba estadística, asociación, BLUP, Correlación, control óptimo.

A STATISTICAL TEST FOR GWAS CONSIDERING THE NON-INDEPENDENCE OF THE BLUP

Magin Zúñiga Estrada, D. en C.

Colegio de Postgraduados, 2019.

ABSTRACT

In the present work a statistical test was developed for analysis of association applied to plants based on the *BLUP*, calculated from a mixed linear model (MLM), and its distributional properties. The test was tested using simulation. The simulated scenarios were determined according to: number of *SNPs* to be tested, number of *SNPs* with non-zero effect, number of genotypes, block size, magnitude of genetic variance and magnitude of correlation between residuals assuming an AR(1) correlation structure. In total, 204 scenarios were simulated with 1000 iterations each. The results of 180 scenarios were compared with another method of association study implemented in *PLINK*. The comparison was made using two indicators, the TDV what is the true detections rate and the TDF what is the false detections rate. Each indicator was analyzed using *ANOVA* where it was obtained that both methods are statistically different for TDV and TDF. The method implemented in *PLINK* turned out to have better behavior for TDV; however, when the effect size was large enough, the proposal obtained values of TDV close to one. With respect to TDF, the proposed method remained, in all cases, below the threshold of 0.05 and the method implemented in *PLINK* had an undesired behavior since it obtained values above 0.2. The last 24 scenarios confirm that the proposal has a desired behavior for TDV when the effect size is large enough. In conclusion, the statistical test shows a desired performance in the sense that it controls in an optimal way the rate of spurious associations and detects *SNPs* with large effect sizes, which gives high certainty of detecting associations true

Keywords: Statistical test, association, BLUP, correlation, optimal control.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico brindado para realizar mis estudios de doctorado.

Al Colegio de Postgraduados, por haberme brindado la oportunidad de seguir mi formación académica y profesional en sus aulas.

A los integrantes de mi Consejo Particular:

- Dr. Sergio Pérez Elizalde.
- Dr. Juan Andrés Burgueño Ferreira.
- Dr. José Crossa Hiriart.
- Dr. José Andrés Christen Gracia.
- Dr. Juan Manuel González Camacho.

por sus sugerencias y tiempo invertido durante el proceso de investigación.

A cada uno de los profesores que contribuyeron en mi formación académica.

A mis compañeros y al personal administrativo que de alguna manera ayudaron a la culminación de este proyecto.

A mi familia.

A mi hijo y mi esposa.

CONTENIDO

LISTA DE CUADROS

LISTA DE FIGURAS

Capítulo 1

INTRODUCCIÓN

Una preocupación que el ser humano ha tenido desde que tiene conciencia sobre los eventos que ocurren a su alrededor, y que estos eventos influyen en su bienestar, es la de buscar explicaciones lógicas a aquellos acontecimientos que son complejos de comprender. Principalmente, del instinto de sobrevivencia nace la curiosidad de querer conocer más allá de lo que los sentidos permiten explorar. Debido a este ímpetu de curiosidad es que en la actualidad se han obtenido grandes logros en todos los campos de investigación en, relativamente, poco tiempo.

El campo de la investigación genética no es la excepción y hoy en día ha alcanzado gran conocimiento sobre las variantes genéticas que se encuentran en estrecha relación con algunos rasgos importantes, tanto en seres humanos como en plantas y animales. Dichas variantes genéticas hace algunos años eran desconocidas debido a la escasa tecnología para escanear el genoma y los altos costos que se generaban al tratar de identificar posibles variantes genéticas a lo largo de todo el genoma. En la actualidad, el avance tecnológico y los bajos costos en el genotipado (Brachi *et al.*, 2011, Ingvarsson y Street, 2011, Kang *et al.*, 2008, Koeleman *et al.*, 2013, Li y Zhu, 2013, Vilhjalmsson y Nordborg, 2012, Xiao *et al.*, 2017), han permitido el desarrollo de nuevas técnicas o métodos más especializados para tratar de identificar relaciones entre rasgos comunes de una población y su constitución genética.

El desarrollo de los métodos que pretenden identificar relaciones entre variantes genéticas comunes y rasgos son motivados por la hipótesis que plantea que los rasgos comunes pueden ser causados, en parte, por variantes genéticas comunes (Witte, 2010,

Xiao *et al.*, 2017); es decir, los rasgos comunes son probablemente influenciados por la variación genética que también es común en la población y se ha visto que para algunas enfermedades, la hipótesis *enfermedad común-variante común* es verdadera (Bush y Moore, 2012). Aunque la hipótesis se restringe al caso de enfermedades es claro que puede extenderse a un amplio rango de características que son de interés. En investigación agrícola; por ejemplo, existen muchos rasgos de interés que pueden ser explicados mediante variantes genéticas a lo largo del genoma. La identificación de regiones del genoma en asociación podría significar una selección más apropiada y mejoramiento de ciertas variedades, donde aún se desconocen sus variantes genéticas asociadas a estas características, y con ello contribuir significativamente a un incremento en el desempeño de muchos cultivos.

Uno de los procedimientos que ha tomado gran relevancia por los buenos resultados que se han obtenido en la identificación de variantes genéticas asociadas a rasgos comunes importantes tanto en seres humanos como en plantas y animales es el estudio de asociación del genoma o *GWAS* por sus siglas en ingles, (*Genome-Wide Association Studies*). En la literatura científica existen varias definiciones de *GWAS* y todas coinciden en que son estudios interesados en identificar asociaciones entre marcadores genéticos y rasgos de los individuos.

Antes de los *GWAS*, los métodos disponibles comprendían el análisis QTL (*Quantitative Trait Loci*), el estudio de ligamiento amplio del genoma *(GWLS)*, y mapeo de desequilibrio de ligamiento *(LDM)* (Levinson, 2009). El concepto QTL o loci de rasgo cuantitativo fue introducido en la década de 1940 y fue utilizado para explicar la herencia de los rasgos no Mendelianos (Long *et al.*, 2008). En términos generales, el análisis QTL se ha utilizado para vincular ciertos fenotipos complejos a regiones específicas de los cromosomas (Miles y Wayne, 2008) mediante la aplicación de un método estadístico que involucra dos tipos de información: datos fenotípicos y datos genotípicos. Durante muchos años se ha considerado este tipo de metodología de vital importancia para identificar QTL, aunque en el presente su aplicación se está volviendo obsoleto (Muhammad *et al.*, 2016). En la actualidad existen los *GWAS* cuyo objetivo es la detección de variantes en locus genómicos que se asocian con rasgos complejos de una población y, en particular, en la detección de asociación entre un polimorfismo de nucleótido único *(SNP)* y un rasgo común (Visscher *et al.*, 2012). El *SNP* es la unidad moderna de variación genética y se define como un cambio en un solo par de bases en la secuencia del ADN que ocurre con alta frecuencia en el genoma (Bush y Moore,

2012). Los *GWAS* se basan en el principio de desequilibrio de ligamiento *(LD)* a nivel de población (Visscher *et al.*, 2012). El desequilibrio de ligamiento es una propiedad de los *SNPs* en un tramo contiguo de secuencia genómica que describe el grado en el que un alelo de un *SNP* se hereda o correlaciona con un alelo de otro *SNP* en la misma población (Bush y Moore, 2012).

La aplicación de los *GWAS* implica tanto técnicas genéticas como teoría estadística. Las técnicas genéticas involucran el genotipado que trata de la identificación de grandes cantidades de *SNPs*. Hoy en día la identificación de variantes genéticas a lo largo del genoma no es un problema gracias al desarrollo tecnológico. Por otro lado, el análisis estadístico para *GWAS* ha representado retos tanto teóricos como computacionales debido al tipo de variables con las que se trabaja y los grandes conjuntos de datos que involucran este tipo de análisis. El enfoque básico de *GWAS* es evaluar asociación entre cada marcador genético y un fenotipo de interés que se ha registrado en un gran número de individuos (Korte y Farlow, 2013). Los fenotipos, que representan la materia prima de los estudios de asociación, pueden ser variables categóricas o cuantitativas. Desde el punto de vista estadístico, los rasgos cuantitativos son preferidos ya que mejoran el poder para detectar un efecto genético. El estudio de asociación del genoma en su conjunto es una serie de pruebas estadísticas de un solo locus, examinando cada *SNP* independientemente para asociación con el fenotipo (Bush y Moore, 2012).

1.1. Planteamiento del problema

Debido a la gran cantidad de factores que intervienen en los estudios de asociación, en el campo de la estadística se han implementado una serie de métodos que ayuden a corregir los resultados obtenidos por posibles factores de confusión. El problema que ha recibido mayor atención es la de tratar de controlar la tasa de falsos-positivos. Un resultado falso-positivo se obtiene cuando estadísticamente hay evidencia de que un *SNP* está en asociación con el fenotipo cuando en la realidad no es así. Si la población bajo estudio es una mezcla de poblaciones que difieren con respecto a las frecuencias alélicas surgirán correlaciones espurias. La estructura de la población es un factor de confusión que provoca la aparición de un gran número de falsos-positivos que debe ser contabilizada mediante la elección de un apropiado diseño de estudio

o controlado en el análisis estadístico (Vilhjalmsson y Nordborg, 2012). El poder de *GWAS* para identificar una verdadera asociación entre un *SNP* y rasgo depende de la varianza fenotípica dentro de la población explicada por el *SNP*. Un método poderoso para explicar el factor de confusión debido a antecedentes genéticos fue desarrollado por primera vez en el campo de la cría de animales y se le conoce como modelo mixto. El modelo mixto contabiliza la estructura de la población por la cantidad de covarianza fenotípica que es debida a la relación genética (incluyendo la relación o parentesco como un término aleatorio dentro del modelo) y, por lo tanto, representa una herramienta importante para el desarrollo de los *GWAS* ya que puede reducir notablemente el número de asociaciones falsas (Korte y Farlow, 2013). Los métodos basados en el modelo lineal mixto han demostrado ser útiles en el control de la estructura de la población y la relación entre individuos dentro de los estudios de asociación de regiones del genoma. Sin embargo, éstos pueden ser computacionalmente intensivos para grandes conjuntos de datos (Zhang *et al*., 2010).

Además, las metodologías estadísticas que se han implementado para *GWAS* se basan en supuestos que en un contexto agrícola son difíciles de sostener y es por ello la necesidad de desarrollar metodologías que puedan ser aplicadas a este tipo de información. Un ejemplo claro y contundente es el supuesto de independencia entre observaciones que son colectadas en campo bajo un diseño experimental que trata de controlar las fuentes de ruido y; por lo tanto, la información colectada tiene cierta estructura de relación entre observaciones que puede enriquecer el análisis estadístico posterior. Ignorar este tipo de propiedades en los datos en el desarrollo de metodologías de análisis estadístico pueden ocasionar pérdida de potencia estadística o un incremento de las tasas de error tipo I y II, y en un caso extremo puede suceder que las metodologías propuestas solo sean aplicadas a escenarios que en la realidad no existen ocasionando una pérdida de tiempo y de capital. Por lo tanto, es fundamental el desarrollo de metodologías que consideren todas las propiedades distribucionales que de forma natural tiene la información.

1.2. Objetivos

En este apartado se presentan los objetivos general y particulares que se pretenden alcanzar en la investigación.

1.2.1. Objetivo General

Desarrollar e implementar computacionalmente una metodología estadística para probar asociación entre rasgo y *SNP* basada en el *BLUP* y sus propiedades distribucionales.

1.2.2. Objetivos particulares

1. Desarrollar una prueba estadística para probar asociación entre *SNP* y rasgo.
2. Escribir un código computacional que permita realizar un análisis de asociación.
3. Ensayar la prueba estadística propuesta mediante simulación.
4. Comparar los resultados de la prueba estadística con otro método de análisis de asociación implementado en el software *PLINK*.

1.3. Hipótesis

Se asume que la inclusión de las propiedades distribucionales de los datos en los estudios de asociación logran una mejora estadística sobre verdaderas detecciones de asociación entre un fenotipo y un marcador genético *SNP*, además de que se logra controlar la tasa de falsos-positivos con la inclusión de factores de confusión que no son intrínsecos de los individuos bajo estudio sino que son resultado de relaciones de convivencia tal como sucede en experimentación agrícola.

1.4. Contenido de la investigación

En los siguientes capítulos se propone y se desarrolla una nueva prueba estadística para estudios de asociación en el contexto agrícola. La investigación se compone de las siguientes partes.

1.4. Contenido de la investigación

En el capítulo dos se presenta una revisión bibliográfica sobre métodos estadísticos implementados para estudios de asociación basados en el modelo lineal mixto.

En el capítulo tres se desarrolla la prueba estadística propuesta para *GWAS*. La metodología consiste de cinco partes fundamentales. La primera parte trata sobre el diseño experimental, el cual es una fuente de ruido sobre las propiedades distribucionales de los datos. La segunda parte consiste en el planteamiento formal de la hipótesis nula. Después se describen los estadísticos de prueba que ayudan, dada una muestra aleatoria, a tomar una decisión sobre la hipótesis nula. Con los datos colectados se ajusta un modelo lineal mixto que es la parte cuatro del capítulo y de donde se obtiene el *BLUP* que es utilizado para calcular el estadístico de prueba. La penúltima parte presenta la distribución del estadístico de prueba. En este apartado se justifica y se muestra la distribución que se usa para obtener un valor crítico y, por lo tanto, una regla de decisión. Por último, se presenta el proceso de simulación partiendo del *BLUP*. En el mismo apartado se identifican los elementos que se cree son determinantes en los resultados de la prueba estadística tanto para el diseño en campo como para valores asumidos de la varianza genética y magnitudes de correlación entre residuales.

En el capítulo cuatro se exponen los resultados de la simulación y se hacen comparaciones entre la metodología propuesta y la metodología implementada en el software *PLINK*. El software *PLINK* se resume en cinco funciones principales, las cuales son: gestión de datos, obtención de estadísticas resumen, estratificación de la población, análisis de asociación y estimación de identidad por descendencia (Purcell *et al.*, 2007). Particularmente, en el análisis de asociación se contempla dos tipos de variables, categórica y continua, lo que determina el tipo de metodología estadística a utilizar. Cuando se asume que el rasgo bajo estudio es una variable categórica y se pretende un prueba de asociación caso-control, entonces se comparan frecuencias alélicas de ambos grupos (casos y controles) mediante, por ejemplo, una prueba exacta de Fisher. Cuando se trata de un rasgo cuantitativo, como se asume en está investigación, se considera al rasgo como la variable dependiente y se ajusta un modelo de regresión lineal incluyendo al *SNP* como variable independiente y en otras covariables se incluyen algunos índices cuantitativos calculados mediante el enfoque de escalamiento multidimensional.

Por último, en el capítulo cinco se presentan las conclusiones. Además, algunas consideraciones generales para utilizar la metodología propuesta.

Capítulo 2

REVISIÓN DE LITERATURA

2.1. Introducción

Hoy en día los estudios de asociación han tomado gran relevancia debido a los buenos resultados que se han obtenido; sin embargo, el camino recorrido no ha sido fácil. Un estudio de asociación involucra gran cantidad de factores tanto de fondo genético como ambientales que dan cuenta del fenotipo de interés, lo que convierte a este tipo de estudios en retos metodológicos para tratar de explicar las relaciones entre fenotipos y variantes genéticas. Los estudios de asociación comprenden una serie de implicaciones a escala individual y global. A escala individual, el estudio de asociación se utiliza para determinar el riesgo que tiene un individuo de desarrollar el rasgo de interés y a escala global, los estudios de asociación ayudan a descubrir asociaciones entre *SNPs* y rasgos de interés que se transmiten de una generación a otra. Dada la enorme complejidad que representa la implementación de los estudios de asociación, su verdadera utilidad depende en gran medida de la comprensión que los investigadores tengan sobre los factores que interactúan detrás de los rasgos complejos.

Un descubrimiento que sirvió de base para el desarrollo de los estudios de asociación fue que los marcadores genéticos polimórficos pueden ser usados para rastrear la herencia genética. A partir de este punto, muchos investigadores comenzaron a construir mapas de ligamiento del genoma. Con cientos de marcadores genéticos distribuidos a lo largo del genoma se logró mapear cientos de caracteres Mendelianos usando análi-

sis de ligamiento; sin embargo, el análisis de ligamiento no funciona tan bien en rasgos complejos. Esto implicó cambiar el enfoque de los estudios de ligamiento basados en la familia a estudios de cohortes y casos-control basados en la población con miles de individuos, lo cual fue el primer paso hacia los estudios de asociación basándose en la idea de que los patrones de desequilibrio de ligamiento a lo largo del genoma pueden usarse para mapear genes de fenotipos (Östensson, 2012).

En rasgos complejos, cada variante de riesgo genético a menudo tiene un pequeño efecto sobre el rasgo y se sabe que los estudios de asociación tienen una mayor potencia estadística para detectar estos pequeños efectos en comparación con el análisis de ligamiento. Los *GWAS* están motivados por la hipótesis variante común - enfermedad común propuesta por Lander (1996), aunque se ha comprobado que dicha afirmación no es cierta para la mayoría de los rasgos complejos.

2.2. Los *GWAS* antes del modelo lineal mixto

Recordando el quehacer de un estudio de asociación, lo que se busca es identificar si un marcador genético está asociado a un rasgo particular, por lo que, si el marcador está asociado al rasgo, las frecuencias de los alelos serán diferentes entre aquellos individuos que tienen el rasgo y los individuos que no lo tienen. Sobre este supuesto se desarrollan los estudios basados en casos y controles, donde se clasifican como casos a aquellos individuos que tienen el rasgo y como controles a aquellos individuos que no tienen el rasgo, pudiendo construir una tabla 2×2 de conteos de alelos en los dos grupos. Una prueba estadística válida para determinar asociación, es la prueba χ^2_1 de Pearson. Otra alternativa, cuando el rasgo que se analiza es binario, es la regresión logística. En regresión logística se pueden incluir muchos factores que podrían estar afectando al rasgo de forma indirecta, por ejemplo factores ambientales o interacciones gen-gen. Otra prueba de asociación que se basa en las frecuencias de alelos es la propuesta por Spealman *et al.* (1993) y se le conoce como prueba de desequilibrio de transmisión o *TDT* por sus siglas en ingles. La prueba *TDT* está basada en el supuesto de que cada uno de los alelos en un locus se transmiten con la misma probabilidad a la descendencia, por lo que para una muestra de heterocigotos se espera que aproximadamente la mitad de uno de los alelos sea transmitido a la descendencia. Cuando sucede que uno de los alelos se transmite con más frecuencia a la descendencia que

cuenta con el rasgo de interés entonces hay indicios de que el alelo se encuentra en asociación con el rasgo.

Aunque las pruebas estadísticas mencionadas funcionan en escenarios de casos y controles; también, en muchas situaciones los resultados presentan asociaciones espurias debido a la falta de contabilización de relaciones entre individuos. Por ejemplo, considere una prueba de asociación caso-control en la cual se comparan las frecuencias de los alelos entre casos y controles, si la muestra proviene de dos poblaciones, los marcadores que no están influyendo en el rasgo tendrán diferencias de manera significativa en la frecuencia de los alelos entre las dos poblaciones dando lugar a una asociación espuria entre marcador genético y rasgo, causada por la existencia de cierta estructura en la muestra.

Dedibo a las altas tasas de falsos-positivos consecuencia de la estructura de la población, se han propuesto varios métodos para corregir este factor de confusión. El primer método para tratar de corregir la estructura de la población fue introducido por Devlin y Roeder (1999). A este método se le conoce como control genómico y consiste en escalar la estadística de prueba para que su mediana se convierta en la mediana esperada. Este enfoque ha sido muy utilizado aunque no resuelve el problema. Un segundo método propuesto para contabilizar la estructura de la población recibe el nombre de asociación estructurada y fue propuesto por Pritchard *et al*. (2000). La asociación estructura utiliza marcadores genéticos aleatorios para estimar la estructura de la población y luego la incorpora a análisis estadísticos posteriores (Yu *et al*., 2006). Otro enfoque, que está basado en el modelo de regresión lineal simple es el análisis de componentes principales o *PCA* por sus siglas en ingles. La idea del método *PCA* para contabilizar la estructura de la población se basa en el modelo infinitesimal de Fisher que dice que los rasgos cuantitativos son el efecto de un número infinito de genes, donde cada gen aporta al rasgo cuantitativo un efecto muy pequeño. De esta forma, se puede tratar de explicar el rasgo cuantitativo mediante una regresión lineal donde la variable respuesta es el rasgo de interés y las variables explicativas son: el marcador genético que se desea probar y otras covariables. En otras covariables se pueden incluir variables de tipo ambiental y la parte para corregir la estructura de la población. En la parte para corregir la estructura de la población, se incluyen los n primeros componentes principales como covariables. Los componentes principales se calculan a partir de la matriz de genotipos. Una de las desventajas de usar componentes principales es que cuando la estructura de la población es demasiada compleja es

necesario incluir muchos componentes principales como covariables lo que ralentiza el método. Este método lo propusieron Price *et al*. (2006) y le dieron el nombre de *EIGENSTRAT*. Otro de los inconvenientes de este método es que se desarrolló sobre el supuesto de que los individuos en la muestra no están relacionados lo que en muchos escenarios no suele ocurrir.

Sobre la premisa del modelo infinitesimal de Fisher, una alternativa para contabilizar la estructura de la población es usar modelos lineales mixtos que incluyan las posibles relaciones entre individuos como un componente aleatorio. Así, surgieron varios enfoques basados en el modelo lineal mixto que dan cuenta de las relaciones entre individuos asumiendo cierta estructura en la matriz de relaciones de parentesco. A continuación se mencionan algunos métodos que han arrojado buenos resultados para controlar la tasa de falsos positivos.

2.3. El modelo lineal mixto en *GWAS*

Después de varios intentos de controlar la estructura de la población y las relaciones entre los individuos en los estudios de asociación, se ha llegado a la posible solución. La idea fundamental de la solución parte del modelo de regresión lineal en el cual se incluye un término poligénico considerado como aleatorio, dando lugar a un modelo lineal mixto (MLM). Entonces, para controlar la estructura de la población, al término poligénico se le asigna una distribución de probabilidad cuya matriz de varianzas y covarianzas estará dada por $\sigma_g^2 K$ y es en la matriz K donde se contabiliza la estructura de la población en el estudio de asociación. σ_g^2 representa la varianza genética y K es una matriz de relaciones de parentesco que generalmente es desconocida cuando hay estructura de la población y una relación críptica entre los individuos de la muestra. En pocas palabras, el enfoque de MLM modela el efecto del genotipo como un término aleatorio mediante la descripción explicita de la estructura de la covarianza entre los individuos.

Un primer acercamiento al MLM para controlar la estructura de la población y en consecuencia controlar la tasa de falsos positivos lo hicieron Yu *et al*. (2006). Ellos desarrollaron un enfoque de modelo mixto unificado para contabilizar los múltiples niveles de relación de forma simultánea de acuerdo a los marcadores genéticos aleatorios.

Según Yu *et al*. (2006), el método muestra un mejor control de las tasas de error tipo I y tipo II; sin embargo, una preocupación importante en el uso del MLM en estudios de asociación es el tiempo consumido por cada análisis y precisamente debido a este inconveniente es que en la actualidad existen muchas propuestas que tratan de optimizar el tiempo de ejecución.

Quizás el primer método, basado en el MLM, que abordó el problema de tiempo de ejecución en un estudio de asociación fue propuesto por Kang *et al*. (2008). El método se le conoce como *EMMA*, por sus siglas en inglés (*Efficient Mixed-Model Association*) y lo que hace es corregir la estructura de la población y la relación genética en el mapeo de asociación. La eficacia del método incrementa aún más al evitar el uso redundante de la matriz computacionalmente costosa en cada iteración mediante el aprovechamiento de la descomposición espectral en el cálculo de la función de verosimilitud, reduciendo el costo computacional de cúbico a lineal. También, es posible converger al óptimo global de la verosimilitud en la estimación de los componentes de varianza con alta confianza combinando la búsqueda vía malla y el algoritmo *Newton-Raphson* (Kang *et al*., 2008). Una parte importante de la propuesta es la estimación de la matriz K. Para estimar K se usa una matriz simple de intercambio de alelos *(IBS)* basada en el supuesto de que cada *SNP* induce el mismo nivel de pequeños cambios aleatorios en el fenotipo y con ello se garantiza que sea semi-definida positiva y converja al óptimo con el adecuado manejo de los alelos faltantes.

Otro enfoque basado en el MLM fue propuesto por Aulchenko *et al*. (2007) y se le conoce como *GRAMMAR* por sus siglas en inglés (*genomewide rapid association using mixed model and regression*). El método *GRAMMAR* se basa en el pedigrí para el estudio de asociación de todo el genoma. La idea básica es realizar un único análisis poligénico usando el pedigrí completo sin tomar en cuenta los datos del marcador genético. Después, los residuales son ajustados mediante una regresión lineal simple como un rasgo cuantitativo que incluyen covarianza poligénica y efectos fijos, para el estudio de asociación con cada uno de los *SNP* usando métodos clásicos (por ejemplo, utilizando la prueba de Wald) bajo el supuesto de observaciones no correlacionadas (Aulchenko *et al*., 2007). Este enfoque es atractivo ya que en la segunda etapa del análisis se pueden usar métodos basados en observaciones no relacionadas. Seguido del enfoque *GRAMMAR* surgió otro método que trata de optimizar el tiempo en las estimaciones de componentes de varianza. Zhang *et al*. (2010) proponen dos enfoques, que al ser implementados conjuntamente se obtiene notablemente una reducción de

tiempo de ejecución y una potencia estadística que sino llega a mejorar se mantiene. El primer enfoque trata con la disminución del tamaño de la muestra mediante el agrupamiento de individuos en conjuntos, a lo cual se le llama *MLM comprimido* (modelo lineal mixto comprimido). El segundo enfoque elimina la necesidad de volver a calcular componentes de varianza y se le conoce como *P3D (population parameters previously determined)*. Debido a que los grandes conjuntos de datos ocupan mucho espacio de memoria y el tiempo de cálculo aumenta al cubo con el número de individuos que se analizan, se ha propuesto abordar este tipo de inconvenientes mediante *MLM comprimido* que consiste en reducir el número de efectos aleatorios mediante la sustitución de n individuos por un número menor de grupos, grupos formados por individuos según parentesco, por lo que el parentesco entre pares de grupos reemplaza el parentesco entre pares de individuos como el efecto aleatorio en un MLM. Por otro lado, siguiendo la línea de ahorro de tiempo, el algoritmo *P3D* es un enfoque que consiste esencialmente de dos pasos. En el primer paso se optimiza un *MLM* con el efecto del marcador genético excluido y en el segundo paso, se ajusta un segundo MLM que incluye el efecto del marcador genético tomando como variable dependiente el rasgo de interés y parámetros de la población previamente estimados en un esquema de MLM (Zhang *et al.*, 2010).

A pesar de los intentos de disminuir el costo computacional en los estudios de asociación, los métodos anteriormente mencionados continúan siendo poco prácticos en el sentido de que consumen mucho tiempo de cálculo y mucha memoria, aún utilizando equipo de computo actual y tamaños de muestra moderados, aunque para lograr la suficiente potencia estadística se requieren tamaños de muestra grandes. Por lo que se ha vuelto una preocupación importante desarrollar metodologías eficientes que ataquen el problema del consumo de tiempo y de memoria sin sacrificar la capacidad de detectar asociaciones verdaderas, sin pérdida de información y disminuyendo la tasa de asociaciones espurias.

Ante este reto computacional, en el presente, existen varios enfoques que pretenden disminuir el tiempo de ejecución (Yang *et al.*, 2014) mediante el adecuado manejo de la estructura de la población y los algoritmos de estimación de componentes de varianza en un MLM. Así es como surge *EMMAX*, como una necesidad de acelerar el tiempo de ejecución de un *GWAS* sin pérdida de información. *EMMAX* fue propuesto por Kang *et al.* (2010) y es una extensión de *EMMA* que; también, fue propuesto por Kang *et al.* (2008), de allí que se le de el nombre de *EMMA eXpedited* para hacer referencia a

que es un algoritmo más rápido que *EMMA*, aún en situaciones de grandes conjuntos de datos. *EMMAX* es un enfoque que se utiliza para corregir la estructura de la muestra basado en un MLM cuya matriz de parentesco es estimada empíricamente para modelar la correlación entre fenotipos de individuos de la muestra (Kang *et al*., 2010). El algoritmo se puede resumir en tres sencillos pasos, en el primer paso se calcula la matriz de relaciones a partir de los *SNPs*, en el segundo paso se ajusta un MLM para estimar la contribución de la estructura de la muestra al fenotipo de interés dando como resultado una matriz de covarianza que modela el efecto de la relación entre los fenotipos. Y en el tercer paso se realiza una *prueba F* o una *prueba score* en cada marcador para detectar asociaciones utilizando la misma matriz de covarianza estimada. Otro algoritmo que también trata de reducir el tiempo de ejecución es *FaST-LMM* propuesto por Lippert *et al*. (2011). El nombre de *FaST-LMM* se derivada de *factored spectrally transformed linear mixed models* y se refiere a la descomposición espectral de la matriz de similitud genética que se hace para acelerar los cálculos en un *GWAS*. El enfoque *FaST-LMM* está basado principalmente en dos formas de interpretar las operaciones en un MLM. Primero, la función de verosimilitud del MLM se escribe en función de una solo parámetro δ, que es la razón de la varianza genética y la varianza residual, por lo que el problema de optimización, ya sea utilizando máxima verosimilitud o máxima verosimilitud restringida, se reduce únicamente a un parámetro. Segundo, solo requiere una única descomposición espectral para probar todos los marcadores genéticos *SNPs*. La idea clave es que la descomposición espectral de la matriz de similitud genética transforma los fenotipos, los *SNPs* y las covariables de tal forma que los datos resultantes no se correlacionan. Estas características logran acelerar la ejecución del *GWAS*, además de que disminuyen el uso de memoria comparado a la aplicación de un modelo lineal mixto estándar. Un método exacto para análisis de asociación fue propuesto por Zhou y Stephens (2012). La principal razón de desarrollar métodos exactos para el estudio de asociación es que en la práctica los métodos aproximados no tienen forma de garantizar que sus resultados sean idénticos a los métodos aproximados. Una consecuencia de la utilización de métodos aproximados es que la potencia estadística podría disminuir. El método exacto propuesto por Zhou y Stephens (2012) recibe el nombre de *GEMMA* y se refiere a *Genome-Efficient Mixed Model Association*. *GEMMA* está basado en el método exacto *EMMA* donde cuya diferencia es que *GEMMA* reemplaza efectivamente el costoso paso de descomposición espectral de *EMMA* con una matriz y multiplicación de vectores, lo que convierte a este método en aproximadamente n veces más rápido (Zhou y Stephens, 2012).

Hasta este punto, los métodos estadísticos basados en el MLM para estudios de asociación están construidos siguiendo el enfoque clásico de la estadística; es decir, se asume una arquitectura genética infinitesimal en donde los tamaños de los efectos tienen una distribución normal. Este supuesto puede reducir el espacio de distribuciones de probabilidad que en algunos entornos puedan resultar más apropiadas que la distribución normal, lo que implicaría una reducción de la potencia de detectar asociaciones verdaderas. Un método que no asume una arquitectura genética infinitesimal y en donde los tamaños de los efectos de los marcadores genéticos asumen una distribución de probabilidad de mezcla Bayesiana a priori fue propuesto por Loh *et al*. (2015). Este método Bayesiano recibe el nombre de *BOLT-LMM* y consiste en adaptar el MLM bajo una perspectiva Bayesiana donde los efectos de *SNPs* son modelados con distribuciones de probabilidad a priori que son adecuadas a la magnitud del efecto según el contexto. Según Loh *et al*. (2015), *BOLT-LMM* es un algoritmo que realiza análisis de modelos mixtos usando un número pequeño de iteraciones y aumenta la potencia estadística mediante el modelado de arquitecturas genéticas no infinitesimales. El algoritmo consiste en asumir un modelo de mezcla Gausiana para los efectos de *SNPs*, y mediante aproximación variacional (Ormerod y Wand, 2010), para calcular residuales de fenotipos aproximados y probar asociación con marcadores candidatos usando un estadístico *score* retrospectivo que proporciona un puente entre el modelado Bayesiano en predicción fenotípica y el estudio de asociación frecuentista. En resumen, *BOLT-LMM* consta de cuatro pasos esenciales. El primero es estimar los parámetros de varianza, en el segundo paso calcula estadísticas de asociación del modelo lineal mixto infinitesimal *(BOLT-LMM-inf)*, en el tercero estima los parámetros de la mezcla Gausiana y en el cuarto calcula las estadísticas de asociación del modelo de mezcla Gausiana *(BOLT-LMM)*. Aunque se ha observado que *BOLT-LMM* muestra una ganancia de potencia estadística aún tiene limitaciones. La ganancia de potencia depende de que los supuestos se cumplan y que el tamaño de la muestra sea lo suficientemente grande.

Los métodos estadísticos para *GWAS* tienen diferentes enfoques que enriquecen el espectro de aplicación y se extienden a diferentes contextos. Sin embargo, aún hace faltan muchas áreas por explorar que son importantes en el desarrollo científico. En el siguiente capítulo se desarrolló una prueba estadística para ser aplicada en el contexto de investigación agrícola y fue ensayada utilizando simulación.

Capítulo 3

MATERIALES Y MÉTODOS

3.1. Introducción

Se ha mencionado la existencia de varios métodos estadísticos dedicados a la identificación de asociaciones entre marcadores genéticos *(SNPs)* y rasgos de interés de los organismos. Sin embargo, gran parte de estos métodos están construidos sobre supuestos que muchas veces no corresponden a la naturaleza de los datos que se están analizando. Por ejemplo, en el caso de investigación agrícola, los datos que se analizan provienen de diseños experimentales que son planeados para cumplir ciertos objetivos. Las propiedades distribucionales que contiene este tipo de información podrían afectar de manera importante los resultados de cualquier análisis estadístico.

Dada la naturaleza de la información es fundamental que en los análisis sean considerados posibles efectos que tienen las observaciones a causa de factores de ruido, ambientales o propias de la existencia misma de la unidad bajo estudio. Aunque en algunas situaciones de análisis estadístico es común que se pasen por alto este tipo de efectos asumiendo que las consecuencias sobre los resultados son mínimas o nulas. Uno de los supuestos que más llama la atención es el de independencia entre observaciones colectadas en escenarios donde con plena seguridad se puede afirmar que los datos o mediciones del rasgo de interés están correlacionadas. En investigación agrícola será siempre conveniente considerar todas las posibles fuentes de ruido que pueden afectar de diferentes formas los resultados. Es por ello la importancia de desa-

rrollar nuevas metodologías que consideren la naturaleza de este tipo de información.

En este capítulo se desarrolla una prueba estadística para *GWAS*, aplicada a plantas, que considera las **propiedades distribucionales** que de forma natural tienen los datos. Lo que se pretende es implementar una metodología que parte de las bases naturales de la información hasta el análisis estadístico según dicta la teoría estadística.

Los diseños experimentales utilizados en agricultura involucran muchas fuentes de variación que convierten la información, con el manejo adecuado y una acertada interpretación estadística, en una invaluable materia prima para la detección de verdaderas asociaciones entre marcadores genéticos y fenotipos. En específico, se pretende explotar el hecho de que las mediciones tomadas en plantas no son independientes por lo que, mediante el desarrollo de una prueba estadística basada en el modelo lineal mixto que considere el hecho de observaciones correlacionadas, se puedan obtener resultados apegados a la realidad y la suficiente potencia estadística.

Para tener el panorama general y específico del desarrollo de la prueba estadística propuesta, el capítulo está compuesto principalmente por dos apartados. El primer apartado consiste del desarrollo de la prueba estadística propuesta para *GWAS* desde el **diseño experimental** pasando por las **hipótesis** que se desean probar, los **estadísticos de prueba**, hasta la obtención de una regla de decisión. En el segundo apartado se ensaya la prueba mediante **simulación**.

3.2. Prueba estadística para *GWAS*

3.2.1. Diseño experimental

Una parte fundamental del quehacer estadístico es el origen de la información debido a que de ello se desprenden los supuestos que ayudan a construir metodologías coherentes con el contexto en el cual se trabaja. En el caso de estudios de asociación es importante tener en cuenta la naturaleza de los datos ya que ayuda en la acertada elección del modelo estadístico. Es fundamental la construcción de un modelo que contemple todos aquellos factores que intervienen en los resultados del experimento

y, también, aquellos factores originados por la mera recolección de la información.

Para los objetivos que se han impuesto en esta investigación se asumirá, sino se menciona lo contrario, que los datos provienen de un diseño experimental y por diseño experimental se entenderá a la forma detallada de ejecutar el experimento en campo para que los datos sean lo más objetivos posibles y así maximizar la calidad de la información.

En estudios de asociación se trabaja con grandes matrices de datos originadas a partir del número de individuos que se someten a estudio y la cantidad de marcadores genéticos que son genotipados, lo que requiere de un diseño experimental eficiente. En experimentación agrícola, el diseño experimental más común es en él que las unidades experimentales se agrupan en bloques, donde cada uno de los bloques contiene todos los tratamientos y se le conoce como diseño de bloques completos. La base teórica que da origen a este tipo de diseños es que el arreglo elimina o controla la heterogeneidad de las unidades experimentales para que puedan ser comparables. Bajo este diseño se asume que la variación entre unidades experimentales dentro de un bloque es menor a la variación entre unidades experimentales de bloques diferentes, de manera que la precisión del experimento incrementa a medida que se controlan las fuentes de error (Steel y Torrie, 1980). Sin embargo, cuando el número de tratamientos es grande es complicado obtener repeticiones completas de los tratamientos dentro de cada bloque, lo que da lugar a la formación de bloques incompletos o si es posible obtener repeticiones completas, el diseño de bloque completo se vuelve ineficiente dado que los bloques son grandes lo que causa que la variación dentro de bloques sea mayor, aumentando el error experimental. En un diseño de bloque incompleto el número de unidades experimentales por bloque es menor que el número total de tratamientos y se considera *resoluble* si las unidades de bloqueo se organizan en repeticiones completas cuando se juntan los bloques de una repetición completa (Barreto *et al*., 1996).

Una clase de diseños de bloques incompletos que son *resolubles* fue introducido por Yates (1936) y se le conoce como *latice cuadrado*. Según Yates (1940) todos los diseños de *latice* no pueden ser menos eficientes que los diseños de bloques completos debido a la propiedad de *resolubilidad*. Aunque, con el diseño de *latice* se logró cubrir una gran cantidad de diseños *resolubles*, aún hacía falta cubrir muchas combinaciones de número de tratamientos y número de repeticiones. Otra clase más general de diseños *resolubles* fue propuesta por Patterson y Williams (1976) y las restricciones

en los tamaños de los bloques disminuyeron en comparación a los *latice*. A este tipo de diseño se le llamó diseño *alfa*. Finalmente, se desarrolló una clase más general de diseños que todos los anteriores. Estos diseños reciben el nombre de diseños *alfa-latice* y no son otra cosa que diseños de bloques incompletos que son *resolubles* cuyo algoritmo de construcción se basa en asignar los tratamientos mediante permutaciones ciclícas del número de tratamientos. Este tipo de diseños tiene más flexibilidad en la elección del número de bloques por cada repetición completa y el tamaño del bloque. Para una revisión más completa de como generar los diseños *alfa-latice* puede consultar Paterson y Patterson (1984).

En nuestro caso, si no se específica a qué tipo de diseño experimental se está haciendo referencia se asumirá que las mediciones del rasgo bajo estudio provienen de un diseño experimental *alfa-latice*.

3.2.2. Planteamiento de la hipótesis nula

Se ha comentado que la idea central de los estudios de asociación es buscar asociaciones en un mapa detallado de variación común a través del genoma; es decir, encontrar una variante común en el genoma cuya existencia esté ligada a la aparición de un rasgo. El proceso para detectar estás asociaciones comprende varias etapas, donde convergen varias disciplinas, tanto de tecnología para la identificación de variaciones en el genoma como de la colaboración de disciplinas tales como genética, medicina, tecnología de laboratorio, bioinformática y estadística.

Antes de empezar con el análisis de la información es necesario plantear de manera formal las hipótesis nulas. El planteamiento de la hipótesis nula se hace asumiendo que no hay asociación entre un marcador genético *SNP* y el rasgo de interés que es común en una población. Explotando la forma en cómo se codifican los *SNPs*, las hipótesis nulas se pueden plantear usando contrastes de medias mediante el agrupamiento de los valores verdaderos de los genotipos de acuerdo al valor que tenga el marcador genético en cada uno de ellos. De esta manera, mediante la comparación de las medias de cada grupo de valores podrá plantearse que no existe diferencia entre las medias involucradas usando un contraste de tal manera que dicho contraste sea igual a cero para decir que no hay asociación entre *SNP* y rasgo.

De manera formal, suponga una matriz M de $g \times m$ de marcadores genéticos *SNPs* tal que m es el número de *SNPs* a probar codificados como -1 cuando el individuo es homocigoto para el primer alelo, 0 cuando el individuo es heterocigoto y 1 cuando el individuo es homocigoto para el segundo alelo, y g el número de genotipos. Entonces, sea x_j el j-ésimo *SNP* tomado de M y sea y el fenotipo de interés medido en escala continua. Ahora, lo que se plantea es la no asociación o el no efecto de x_j sobre el rasgo y mediante el contraste de medias. Además, se pueden probar dos tipos de efectos de los marcadores genéticos sobre el rasgo y , por lo que se plantearan dos hipótesis nulas de la siguiente forma:

- Probar efectos aditivos:

$$\mathsf{H}_{0a} : \mu_{1j} - \mu_{-1j} = 0 \quad \text{vs} \quad \mathsf{H}_{1a} : \mu_{1j} - \mu_{-1j} \neq 0 \tag{3.1}$$

- Probar efectos de heterosis:

$$\mathsf{H}_{0h} : 0.5(\mu_{1j} + \mu_{-1j}) - \mu_{0j} = 0 \quad \text{vs} \quad \mathsf{H}_{1h} : 0.5(\mu_{1j} + \mu_{-1j}) - \mu_{0j} \neq 0 \tag{3.2}$$

donde μ_{ij} representa la media de los valores verdaderos de los g genotipos cuando el j-ésimo *SNP* , $j = 1, 2, ..., m$, toma el valor i para $i = -1, 0, 1$.

Note que el planteamiento de las hipótesis es análogo a lo que se prueba en un análisis de varianza un solo factor. En este caso, el marcador genético *SNP* se interpreta como un factor con diferentes niveles (tres niveles), tal y como sucede cuando se analizan tratamientos en diseños experimentales. Entonces, en términos generales lo que se está probando es que no hay efecto de los niveles del factor sobre la variable respuesta, es decir, el marcador genético x_j no está asociado al fenotipo y. Tomando está analogía del análisis de varianza, en la siguiente sección se exponen las ideas y los estadísticos de prueba correspondientes para probar H_{0a} y H_{0h}, respectivamente.

3.2.3. Estadísticos de prueba

Los estadísticos de prueba propuestos para probar H_{0a} y H_{0h} son análogos a los estadísticos de prueba que se utilizan en comparaciones múltiples de medias y su justifi-

cación se encuentra en la codificación de un marcador genético *SNP* donde implícitamente los valores verdaderos de los genotipos se pueden agrupar de acuerdo al valor que toma el *SNP* en cada uno de ellos y de está forma comparar los grupos utilizando medias observadas.

En un análisis de varianza de un solo factor, el estadístico de prueba que se utiliza para probar la hipótesis nula de que no hay efecto de niveles del factor sobre la variable respuesta se basa en una razón de cuadrados medios. Retomando está idea, para probar H_{0a} y H_{0h}, respectivamente, los estadísticos de prueba se basan en:

$$\frac{\dfrac{T(\mathbf{Y}) - \mathsf{E}(T(\mathbf{Y}))}{\sqrt{\mathsf{Var}\,(T(\mathbf{Y}))}}}{\sqrt{\dfrac{\nu\widehat{\mathsf{Var}}\,(T(\mathbf{Y}))}{\nu\mathsf{Var}\,(T(\mathbf{Y}))}}} = \frac{T(\mathbf{Y}) - \mathsf{E}(T(\mathbf{Y}))}{\sqrt{\widehat{\mathsf{Var}}\,(T(\mathbf{Y}))}} \tag{3.3}$$

donde $T(\mathbf{Y})$ es una función que depende del vector $\mathbf{Y}$ de variables aleatorias, $\mathsf{E}(T(\mathbf{Y}))$ es su valor esperado, $\mathsf{Var}\,(T(\mathbf{Y}))$ es su varianza, $\widehat{\mathsf{Var}}\,(T(\mathbf{Y}))$ es su varianza estimada y ν representa los grados de libertad.

Tomando como base la expresión (3.3), a continuación se proponen los estadísticos de prueba:

- Estadístico de prueba para probar efectos aditivos:

$$T_a = \frac{\overline{U}_{1j} - \overline{U}_{-1j}}{SE(\overline{U}_{1j} - \overline{U}_{-1j})} \tag{3.4}$$

- Estadístico de prueba para probar efectos de heterosis:

$$T_h = \frac{0.5(\overline{U}_{1j} + \overline{U}_{-1j}) - \overline{U}_{0j}}{SE\left(0.5(\overline{U}_{1j} + \overline{U}_{-1j}) - \overline{U}_{0j}\right)} \tag{3.5}$$

donde $\overline{U}_{ij}$ es la media del $EBLUP(\mathbf{U}) = \widehat{\mathbf{U}}$, que son predicciones de los verdaderos valores de los g genotipos, para $i = -1, 0, 1$ y $j = 1, 2, ..., m$. El acrónimo $EBLUP$ se refiere a *Empirical Best Linear Unbiased Predictor* obtenido del ajuste de un MLM.

En el procedimiento de prueba de hipótesis, después de obtener una estimación del estadístico de prueba, dada una muestra, lo que sigue es determinar un valor crítico a

partir de la función de distribución de probabilidad del estadístico de prueba y un nivel de significancia nominal, α, para tomar una decisión sobre la hipótesis nula. Por lo que, lo más lógico en el desarrollo de la prueba es obtener la distribución de probabilidad de los estadísticos de prueba, T_a y T_h. Sin embargo, antes de proseguir con el desarrollo de la prueba estadística, en nuestro caso es necesario describir los componentes del MLM utilizado para obtener $\widehat{\mathbf{U}}$, así como su proceso de estimación ya que para obtener las distribuciones de probabilidad de T_a y T_h es esencial primero exponer los supuestos sobre los cuales se sustenta la obtención del *EBLUP*.

3.2.4. El Modelo Lineal Mixto

La elección del modelo depende principalmente de los objetivos que se persiguen en la investigación y de las propiedades distribucionales que de forma natural tienen los datos. Una herramienta que ha mostrado ser eficiente para controlar la tasa de falsos positivos, que son una consecuencia de trabajar con poblaciones estructuradas, es el **modelo lineal mixto**. La mayoría de los trabajos sobre estudios de asociación coinciden en que los modelos mixtos han demostrado ser una poderosa herramienta para dar cuenta de factores de confusión que intervienen de forma importante en los resultados. Los factores de confusión más importantes que determinan la aparición de falsos-positivos en los estudios de asociación son la estructura de la población y las relaciones de parentesco entre los individuos.

Generalmente se dice que se tiene un falso-positivo cuando se asume como verdadera determinada afirmación sobre una condición que realmente es falsa, después de un proceso de prueba de hipótesis. En el caso de estudios de asociación se tiene un falso-positivo cuando después de una prueba estadística se encuentra significativa la asociación entre marcador genético y rasgo de interés aunque realmente dicha relación no existe. El MLM reduce de forma significativa la tasa de falsos positivos manteniendo la potencia de la prueba estadística.

Como es sabido, el término *mixto* del modelo lineal mixto se refiere a que la ecuación del modelo esta compuesta, esencialmente, por dos tipos de elementos, un elemento fijo y un elemento aleatorio. La determinación de estos dos elementos se basa, principalmente, en la información disponible y en el criterio del investigador. Para saber en que consiste la diferencia entre estos dos términos se dirá que los efectos fijos en

el modelo lineal son aquellos efectos que por si solos se desean conocer y no existe una población detrás de ellos. Por otro lado, los efectos aleatorios son aquellos efectos que solo se incluye un subconjunto de ellos en el modelo y representan una selección aleatoria igualmente representativos de una población. Además, el interés de los efectos aleatorios se centra en examinar la variabilidad de la respuesta debida a toda la población de dichos efectos. Como una sugerencia se dice que los elementos de la estructura del diseño son tratados generalmente como efectos aleatorios, por ejemplo, los bloques que son incluidos en el modelo pertenecen a un conjunto más grande de bloques sobre los cuales se desea hacer inferencia.

Ecuación del modelo lineal mixto

La ecuación del *MLM* es la representación matemática de todos los elementos que se están considerando en el análisis, así como las relaciones existentes entre ellos. La ecuación del modelo es una forma sintética de resumir el gran número de elementos o componentes que se están incluyendo en el análisis. Sin embargo, el hecho de que en un modelo se puedan considerar infinidad de elementos para lograr un buen ajuste es importante recordar el significado de *menos es más* para referirse a que un modelo siempre será más apropiado cuando el número de parámetros a estimar es mínimo sin perder capacidad de predicción. Es obvio que un modelo tendrá más capacidad de predicción a medida que el número de variables explicativas aumente pero en la práctica será más conveniente un modelo con pocos parámetros.

A continuación se expone la ecuación del MLM propuesto para el estudio de asociación, desde su forma más simple.

$$y_{ijk} = \mu + a_i + b_{ji} + (g \times a)_{ki} + e_{ijk} \tag{3.6}$$

donde

- y_{ijk} representa el valor del fenotipo en el ambiente i, bloque j y genotipo k.
- μ representa el valor de la media global.
- a_i representa el efecto del $i-$ésimo ambiente.

- b_{ji} representa el efecto del diseño experimental mediante la inclusión del efecto del bloque $j-$ésimo en el ambiente i.
- $(g \times a)_{ki}$ representa el efecto de la interacción entre el $k-$ésimo genotipo y el $i-$ésimo ambiente.
- e_{ijk} representa el efecto del error residual no observable de la observación ijk.

De los componentes arriba descritos es conveniente decir cuales elementos son considerados como componentes fijos y cuales son considerados como componentes aleatorios. La media global μ, en todo modelo, siempre será considerada como un término fijo. En este caso, el efecto de ambiente a_i no representa un subconjunto de todos los posibles efectos y por lo tanto será considerado como un efecto fijo. El efecto de bloque b_{ji} representa un caso particular extraído de un conjunto de posibles efectos de bloques y por está razón es considerado como un efecto aleatorio. En el caso del efecto de interacción $(g \times a)_{ki}$ genotipo-ambiente, está demostrado que la inclusión de este término ayuda en gran medida con el ajuste del modelo además de que es un término que involucra la contabilización de la estructura de la población. El efecto interacción genotipo-ambiente es considerado un efecto aleatorio. Por último, el término error residual e_{ijk} es considerado un componente aleatorio.

Para fines de estimación es necesario la representación matricial del modelo (3.6) y enumerar los supuestos en los mismos términos.

$$\mathbf{Y} = \mu\mathbf{1} + \mathbf{X}_1\mathbf{a} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2(\mathbf{g} \times \mathbf{a}) + \mathbf{e} \tag{3.7}$$

donde

- $\mathbf{Y}$ representa el vector de dimensión $n \times 1$ de observaciones del fenotipo.
- μ representa la media general y es un escalar. $\mathbf{1}$ es un vector $n \times 1$ donde todos sus elementos son unos.
- $\mathbf{a}$ es el vector de dimensión $p_1 \times 1$ que contiene el efecto de los ambientes. $\mathbf{X}_1$ es una matriz diseño $n \times p_1$ relacionada a los efectos de ambiente.
- $\mathbf{b}$ es el vector $q_1 \times 1$ de efectos de bloque y $\mathbf{Z}_1$ es una matriz diseño $n \times q_1$ relacionada a los efectos de bloque.

- $(\mathbf{g} \times \mathbf{a})$ es un vector $q_2 \times 1$ que representa el efecto de la interacción genotipo-ambiente y $\mathbf{Z}_2$ es una matriz diseño $n \times q_2$ relacionada a los efectos de interacción genotipo-ambiente.
- $\mathbf{e}$ representa el vector de residuales aleatorios no observables de dimensión $n \times 1$.

El modelo en (3.7) está incompleto, hace falta incorporar supuestos sobre las distribuciones de probabilidad de los componentes aleatorios. Es en estos supuestos donde se incluye los componentes de varianza que son parte esencial de este trabajo.

Supuestos y estimación del modelo lineal mixto

Para la estimación del MLM dado en (3.7) se usó la metodología de estimación propuesta por C. R. Henderson. Los supuestos en que se basó Henderson para obtener estimaciones y predicciones del MLM fue asumir que los componentes aleatorios tienen una distribución normal con cierta media y varianza, y además que son independientes entre ellos.

A continuación se enumeran los supuestos de los componentes aleatorios del modelo (3.7):

1. $\mathbf{b} \sim N_{q_1}(\mathbf{0}, \mathbf{D}_1)$.
2. $(\mathbf{g} \times \mathbf{a}) \sim N_{q_2}(\mathbf{0}, \mathbf{D}_2)$.
3. $\mathbf{e} \sim N_n(\mathbf{0}, \mathbf{R})$.
4. Cov$(\mathbf{b}, (\mathbf{g} \times \mathbf{a})^t) = \mathbf{0}$, Cov$(\mathbf{b}, \mathbf{e}^t) = \mathbf{0}$ y Cov$((\mathbf{g} \times \mathbf{a}), \mathbf{e}^t) = \mathbf{0}$.

Derivado de los supuestos se desprende la distribución de probabilidad para el vector respuesta $\mathbf{Y}$ y de allí se parte para la estimación de los componentes del modelo.

- E $(\mathbf{Y}) = E\left(\mu\mathbf{1} + \mathbf{X}_1\mathbf{a} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2(\mathbf{g} \times \mathbf{a}) + \mathbf{e}\right) = \mu\mathbf{1} + \mathbf{X}_1\mathbf{a}$
- Var $(\mathbf{Y}) =$ Var $\left(\mu\mathbf{1} + \mathbf{X}_1\mathbf{a} + \mathbf{Z}_1\mathbf{b} + \mathbf{Z}_2(\mathbf{g} \times \mathbf{a}) + \mathbf{e}\right) = \mathbf{Z}_1\mathbf{D}_1\mathbf{Z}_1^t + \mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2^t + \mathbf{R} = \mathbf{V}$

- Por lo tanto, haciendo a $\mathbf{X} = [\mathbf{1} \quad \mathbf{X}_1]$ y a $\boldsymbol{\beta}^t = (\mu \quad \mathbf{a}^t)$, se tiene que $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$

Por lo tanto, para la estimación de (3.7), se maximiza la función de densidad conjunta entre $\mathbf{Y}$, $\mathbf{b}$ y $(\mathbf{g} \times \mathbf{a})$. Antes de la presentación de la función de densidad conjunta es conveniente hacer un cambio de variable solo para hacer sintético el desarrollo de la estimación, en lugar de usar $(\mathbf{g} \times \mathbf{a})$ de ahora en adelante se usará $\mathbf{U}$ sin perder de vista su significado. La función de densidad conjunta está dada por:

$$\begin{aligned} f(\mathbf{y}, \mathbf{b}, \mathbf{U}) &= f(\mathbf{y} \mid \mathbf{b}, \mathbf{U}) f(\mathbf{b}) f(\mathbf{U}) \\ &= \left[\left(\frac{1}{\sqrt{2\pi}} \right)^n \left(\frac{1}{\mid \mathbf{R} \mid} \right)^{1/2} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{b} - \mathbf{Z}_2\mathbf{U})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{b} \right. \right. \\ &\left. \left. - \mathbf{Z}_2\mathbf{U}) \right\} \right] \times \left[\left(\frac{1}{\sqrt{2\pi}} \right)^{q_1} \left(\frac{1}{|\mathbf{D}_1|} \right)^{1/2} \exp\left\{ -\frac{1}{2} \mathbf{b}^t \mathbf{D}_1^{-1} \mathbf{b} \right\} \right] \\ &\times \left[\left(\frac{1}{\sqrt{2\pi}} \right)^{q_2} \left(\frac{1}{|\mathbf{D}_2|} \right)^{1/2} \exp\left\{ -\frac{1}{2} \mathbf{U}^t \mathbf{D}_2^{-1} \mathbf{U} \right\} \right] \end{aligned} \tag{3.8}$$

Lo que sigue es obtener el logaritmo de la función $f(\mathbf{y}, \mathbf{b}, \mathbf{U})$, el cual está dado por la siguiente función:

$$\begin{aligned} l(\boldsymbol{\beta}, \mathbf{b}, \mathbf{U}) = &- \frac{n}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{R}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{b} - \mathbf{Z}_2\mathbf{U})^t \mathbf{R}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_1\mathbf{b} - \mathbf{Z}_2\mathbf{U}) \\ &- \frac{q_1}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{D}_1| - \frac{1}{2} \mathbf{b}^t \mathbf{D}_1^{-1} \mathbf{b} - \frac{q_2}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{D}_2| - \frac{1}{2} \mathbf{U}^t \mathbf{D}_2^{-1} \mathbf{U} \end{aligned}$$

Enseguida se obtienen las derivadas parciales respecto a cada uno de los elementos desconocidos:

$$\frac{\partial l(\boldsymbol{\beta}, \mathbf{b}, \mathbf{U})}{\partial \boldsymbol{\beta}} = \mathbf{X}^t \mathbf{R}^{-1} \mathbf{y} - \mathbf{X}^t \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta} - \mathbf{X}^t \mathbf{R}^{-1} \mathbf{Z}_1 \mathbf{b} - \mathbf{X}^t \mathbf{R}^{-1} \mathbf{Z}_2 \mathbf{U} \tag{3.9}$$

$$\frac{\partial l(\boldsymbol{\beta}, \mathbf{b}, \mathbf{U})}{\partial \mathbf{b}} = \mathbf{Z}_1^t \mathbf{R}^{-1} \mathbf{y} - \mathbf{Z}_1^t \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta} - \left(\mathbf{Z}_1^t \mathbf{R}^{-1} \mathbf{Z}_1 + \mathbf{D}_1^{-1} \right) \mathbf{b} - \mathbf{Z}_1^t \mathbf{R}^{-1} \mathbf{Z}_2 \mathbf{U} \tag{3.10}$$

$$\frac{\partial l(\boldsymbol{\beta}, \mathbf{b}, \mathbf{U})}{\partial \mathbf{U}} = \mathbf{Z}_2^t \mathbf{R}^{-1} \mathbf{y} - \mathbf{Z}_2^t \mathbf{R}^{-1} \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_2^t \mathbf{R}^{-1} \mathbf{Z}_1 \mathbf{b} - \left(\mathbf{Z}_2^t \mathbf{R}^{-1} \mathbf{Z}_2 + \mathbf{D}_2^{-1} \right) \mathbf{U} \tag{3.11}$$

Igualando a cero cada una de las derivadas parciales se obtienen las famosas ecua-

ciones del modelo lineal mixto de Henderson *(EMLM)*:

$$\begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{D}_1^{-1} & \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{D}_2^{-1} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widetilde{\mathbf{b}} \\ \widetilde{\mathbf{U}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3.12}$$

Note que la solución de (3.12), sobre el supuesto de que la inversa existe, está dada por:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widetilde{\mathbf{b}} \\ \widetilde{\mathbf{U}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{D}_1^{-1} & \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{Z}_2 \\ \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{D}_2^{-1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3.13}$$

Derivado de la expresión (3.13) se obtienen las soluciones individuales para cada componente desconocido:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{P}_2\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{P}_2\mathbf{y} \tag{3.14}$$

$$\widetilde{\mathbf{b}} = \left(\mathbf{Z}_1^t\mathbf{P}_1\mathbf{Z}_1 + \mathbf{D}_1^{-1}\right)^{-1}\mathbf{Z}_1^t\mathbf{P}_1\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{3.15}$$

$$\widetilde{\mathbf{U}} = \left(\mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{D}_2^{-1}\right)^{-1}\mathbf{Z}_2^t\mathbf{R}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{Z}_1\widetilde{\mathbf{b}}\right) \tag{3.16}$$

donde: $\mathbf{P}_1 = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}_2\left(\mathbf{Z}_2^t\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{D}_2^{-1}\right)^{-1}\mathbf{Z}_2^t\mathbf{R}^{-1}$ y $\mathbf{P}_2 = \mathbf{P}_1 - \mathbf{P}_1\mathbf{Z}_1\left(\mathbf{Z}_1^t\mathbf{P}_1\mathbf{Z}_1 + \mathbf{D}_1^{-1}\right)^{-1}\mathbf{Z}_1^t\mathbf{P}_1$

Note la diferenciación que se hace para referirse a los estimadores de los componentes en el modelo. En el caso de la representación para el estimador de $\boldsymbol{\beta}$ se le agrega el símbolo $\wedge$ al estimador de $\boldsymbol{\beta}$ y en el caso de $\mathbf{b}$ y $\mathbf{U}$, en la representación de sus estimadores, se les agrega $\sim$. La cuestión es sencilla, la distinción se hace debido al tipo de término que se está estimando. Por un lado $\boldsymbol{\beta}$ representa el efecto fijo en el modelo, así $\widehat{\boldsymbol{\beta}}$ representa la estimación de los efectos fijos y se le suele llamar **estimador**. Por otro lado, $\mathbf{b}$ y $\mathbf{U}$ representan los efectos aleatorios en el modelo y por esa razón su estimación se representa de forma distinta a $\boldsymbol{\beta}$; es decir, como $\widetilde{\mathbf{b}}$ y $\widetilde{\mathbf{U}}$ y se les llama **predictor**.

Formalmente, a las expresiones obtenidas del ajuste del MLM, tanto para $\mathbf{b}$ como para $\tilde{\mathbf{U}}$, se le da el nombre de *BLUP* y se refiere a las propiedades que tienen estos predictores. El acrónimo *BLUP* proviene de *Best Linear Unbiased Predictor* refiriéndose

a que las predicciones $\widetilde{\mathbf{b}}$ y $\widetilde{\mathbf{U}}$ son las mejores de todas las posibles soluciones en el sentido de que tienen varianza mínima, son lineales en los datos $(\mathbf{y})$ y su esperanza es igual al parámetro que se está prediciendo, es decir, $\mathrm{E}(\tilde{\mathbf{b}}) = \mathrm{E}(\mathbf{b})$, y $\mathrm{E}(\tilde{\mathbf{U}}) = \mathrm{E}(\mathbf{U})$, respectivamente (Searle, 1995). Por otro lado, toda combinación lineal de $\boldsymbol{\beta}$, de la forma $\mathbf{X}\boldsymbol{\beta}$, es estimable y la estimación está dada por $\mathbf{X}\widehat{\boldsymbol{\beta}}$. A esta estimación se le conoce como *BLUE* y su significado en inglés es *Best Linear Unbiased Estimator*; también, para describir las propiedades que tiene el estimador. Así, la expresión $\mathbf{X}\widehat{\boldsymbol{\beta}}$ es el mejor estimador en el sentido de que tiene menor cuadrado medio de estimación, es lineal en los datos $\mathbf{y}$ y es insesgado en el sentido de que $\mathrm{E}(\mathbf{X}\widehat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta}$.

Hasta este punto no se ha dicho nada con respecto a los componentes de varianza aunque de manera directa son usados para estimar y predecir los componentes del MLM. Es importante mencionar que para fines de estimación del MLM, los componentes de varianza se asumen como conocidos pero que en la realidad dichos parámetros a menudo son desconocidos y por lo tanto deben ser estimados.

Estimación de componentes de varianza

En la actualidad existen varias alternativas para estimar componentes de varianza de un MLM. Estos procedimientos van desde algoritmos sencillos, como el *ANOVA* hasta algoritmos más complejos como máxima verosimilitud o máxima verosimilitud restringida que se apoyan en aproximaciones numéricas. También, la complejidad de la estimación depende de los supuestos que se hacen sobres las estructuras de las matrices de varianza y covarianza. En nuestro caso, las estructuras de $\mathbf{D}_1$, $\mathbf{D}_2$ y $\mathbf{R}$ se asumen de la siguiente forma:

$$\mathbf{D}_1 = \sigma_b^2 \mathbf{I}_{q_1} \quad \mathbf{D}_2 = \sigma_g^2 \mathbf{G} \quad \text{y} \quad \mathbf{R} = \sigma_e^2 \mathbf{C} \tag{3.17}$$

donde $\mathbf{I}_{q_1}$ representa la matriz identidad, $\mathbf{G}$ es una matriz de relaciones derivada de las frecuencias genotípicas y $\mathbf{C}$ es una matriz con una estructura de correlación espacial conocida.

El método más común para estimar componentes de varianza es el *ANOVA*. El enfoque más recomendado es el propuesto por Henderson (1953) ya que representa el más simple de calcular, incluso cuando se usan calculadoras portátiles (da Silva, 2017). La idea esencial del método *ANOVA* es comparar el error cuadrado medio de cada fuente

de variación con su respectivo valor esperado, de esta forma se obtiene un sistema de ecuaciones lineales, donde las soluciones son las estimaciones de los componentes de varianza. Aunque para derivar las estimaciones de componentes de varianza mediante *ANOVA* no es necesario suponer alguna distribución de probabilidad para los datos subyacentes, el procedimiento no se adapta cuando se manejan gran cantidad de datos y/o los datos provienen de diseños desbalanceados que podrían afectar la insesgadez de los estimadores, además que representa un reto aritmético el hecho de tener tantas fuentes de variación, por lo que se sugiere el uso de métodos más eficientes para este tipo de casos.

Uno de los métodos más conocidos para estimar parámetros asociados a funciones de densidad o de probabilidad, es el llamado método de máxima verosimilitud *(MV)* propuesto por Fisher (1925). Aunque, se cree que el primero en utilizar el método para estimar componentes de varianza fue Crump (1947). A diferencia del *ANOVA*, un requisito básico para la aplicación de *MV* es el supuesto de una función de distribución de probabilidad sobre los datos que se están analizando. Una elección natural de la distribución de probabilidad es la función de distribución normal. Esta distribución es elegida dada su practicidad en el sentido matemático ya que conduce a una metodología manejable incluso en escenarios de datos desbalanceados aunque para algunos tipos de datos no sea necesariamente apropiada (Searle *et al*., 2006). La idea fundamental del método de máxima verosimilitud se puede resumir en los siguientes pasos:

1. Asumir una función de distribución de probabilidad. Así, en el MLM, tanto los términos aleatorios como el término de error residual siguen una distribución normal multivariada con media el vector de ceros y una matriz de varianzas-covarianzas con estructura conocida.

2. Se obtiene el logaritmo natural de la función de densidad conjunta conocida como función de log-verosimilitud.

3. La función de log-verosimilitud se deriva parcialmente una vez con respecto a cada uno de los componentes del modelo que son desconocidos.

4. Cada derivada parcial se iguala a cero para obtener un sistema de ecuaciones lineales.

5. La solución del sistema de ecuaciones lineales se obtiene mediante la elección

de un parámetro desconocido que dependa del resto, para ser sustituido de forma iterativa en cada una de las ecuaciones hasta obtener que su estimación solo dependa de los datos muestrales. Sin embargo, cuando se tienen varios factores y cada factor tiene varios niveles, la solución de este sistema de ecuaciones lineales se vuelve intratable matemáticamente por lo que se debe recurrir a algoritmos de aproximación para obtener las soluciones.

6. Las soluciones del sistema de ecuaciones son las que maximizan la verosimilitud y de allí a que las estimaciones se les conozca como estimadores de máxima verosimilitud.

Otro método que es una modificación del método *MV*, es el método de máxima verosimilitud restringida *(MVR)*. Los primeros en proponer el método de *MVR* para datos desbalanceados fueron Patterson y Thompson (1971) (Searle, 1994) aunque inicialmente fue desarrollado por Anderson y Bancroft (1952) para datos balanceados. El método de máxima verosimilitud restringida utiliza una transformación de los datos, en lugar de los datos originales, para eliminar el efecto de *ruido* originado por la presencia de los efectos fijos en el modelo, aunque toma en cuenta los grados de libertad respecto de los efectos fijos, así lo que maximiza es la función de verosimilitud que solo depende de los componentes de varianza desconocidos. Dos descripciones de remover los efectos fijos son que es un método de estimación que (Searle, 1993):

- Maximiza la parte de la verosimilitud que es invariante al parámetro de localidad (Thompson, 1962), es decir, invariante a los efectos fijos.
- Maximiza la función de verosimilitud del número máximo de contrastes del error que son linealmente independientes (Harville, 1977).

Los métodos MV y MVR utilizados para estimar componentes de varianza han recurrido a algoritmos iterativos para obtener soluciones aproximadas que maximizan la verosimilitud debido a que, en situaciones del MLM donde se involucran muchos factores aleatorios, es casi imposible deducir una solución cerrada mediante el empleo de propiedades algébricas en un sistema de ecuaciones. Por lo tanto, se han desarrollado varios algoritmos de aproximación numérica para resolver las ecuaciones derivadas de la maximización de la verosimilitud que difieren entre sí con respecto a la velocidad de convergencia y las necesidades de cálculo.

Existen varios algoritmos de aproximación numérica que pueden agruparse de acuerdo al uso de derivadas y el número de veces que las emplean. Los algoritmos que son libres de derivadas carecen de eficiencia debido a su lenta convergencia y su complejidad cuando se tiene gran cantidad de parámetros, por lo que su utilización se ha visto reducida en estudios de asociación. Los algoritmos que emplean derivadas son los que resultan atractivos para los estudios de asociación debido a que convergen rápidamente, aunque en algunas situaciones la convergencia no está garantizada.

Con las estimaciones de los componentes de varianza, el MLM (3.7) estará totalmente estimado y dispuesto a ser utilizado. Recuerde que en la presentación de los estadísticos de prueba (3.4) y (3.5) se dijo que dichos estadísticos son funciones del $BLUP(\mathbf{U}) = \widetilde{\mathbf{U}}$ o más aún, del $EBLUP(\mathbf{U}) = \widehat{\mathbf{U}}$ obtenido de la estimación del MLM (3.7), cuando en lugar de usar las matrices de variazas y covarianzas reales $\mathbf{D}_1$, $\mathbf{D}_2$ y $\mathbf{R}$, se utilizan estimaciones $\widehat{\mathbf{D}}_1$, $\widehat{\mathbf{D}}_2$ y $\widehat{\mathbf{R}}$.

3.2.5. Distribución de probabilidad de los estadísticos de prueba

En el proceso de desarrollo de una prueba estadística, un punto importante y fundamental es la apropiada determinación de la función de distribución de probabilidades del estadístico de prueba. Existe dos formas, en nuestro caso, que pueden ayudar a determinar un valor crítico y así tomar una decisión con respecto a la hipótesis nula, dado un nivel de significancia nominal. Una forma es derivar la distribución de probabilidad exacta, la cual representa un reto matemático bastante complicado dado que en lugar de tener la varianza real conocida se tiene una estimación. La otra forma es obtener una distribución de probabilidad aproximada que contemple el hecho de que se está usando una varianza estimada.

En este tipo de casos, lo recomendable es optar por una distribución de probabilidad aproximada debido a que se cuenta con el suficiente tamaño de muestra para obtener distribuciones de probabilidad casi idénticas a las verdaderas, además de evitar el desarrollo matemático que puede resultar bastante tedioso. A continuación, en este apartado, se describe la forma en que se deduce la distribución de probabilidad aproximada para los estadísticos de prueba propuestos. Para una revisión más amplia sobre la aproximación se puede referir a Witkovsky (2012).

Note que los estadísticos de prueba, T_a y T_h, son funciones lineales del $BLUP(\mathbf{U})$ sobre el supuesto de que los componentes de varianza , $\mathbf{D}_1$, $\mathbf{D}_2$ y $\mathbf{R}$, son conocidos. Así, lo que se pretende hacer es inferir sobre funciones lineales de los efectos aleatorios del MLM. La propuesta de Witkovsky (2012) se basa en los elementos de la solución de las ecuaciones del MLM de Henderson (1973). Recuerde que las ecuaciones del modelo mixto fueron derivadas bajo el supuesto de normalidad de $\mathbf{b} \sim N_{q_1}(\mathbf{0}, \mathbf{D}_1)$, $(\mathbf{g} \times \mathbf{a}) \sim N_{q_2}(\mathbf{0}, \mathbf{D}_2)$ y $\mathbf{e} \sim N_n(\mathbf{0}, \mathbf{R})$. Para comprender mejor la aproximación utilizada en la investigación se da una pequeña descripción de la justificación de la aproximación de la distribución de probabilidad desde el punto de vista del autor y para ello se utiliza la misma terminología que se utiliza en el artículo titulado *Estimation, Testing, and Prediction Regions of the Fixed and Random Effects by Solving the Henderson's Mixed Model Equationsn* desarrollado por Witkovsky (2012).

Sea el MLM:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{3.18}$$

El objetivo principal es hacer inferencia sobre:

- $\mathbf{K}^t\mathbf{b}$
- $\mathbf{w} = \mathbf{\Lambda}^t \begin{pmatrix} \mathbf{b}^t & \mathbf{u}^t \end{pmatrix} = \mathbf{K}^t\mathbf{b} + \mathbf{L}^t\mathbf{u}$

La derivación de las ecuaciones del modelo mixto se hace sobre el supuesto de normalidad: $\mathbf{u} \sim (\mathbf{0}, \mathbf{G})$, $\mathbf{e} \sim (\mathbf{0}, \mathbf{R})$ y $\text{Cov}(\mathbf{u}, \mathbf{e}) = \mathbf{0}$ con $\mathbf{G}$ y $\mathbf{R}$ conocidas. Así, las ecuaciones del MLM son:

$$\begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \widetilde{\mathbf{b}} \\ \widetilde{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \tag{3.19}$$

Para obtener una solución para las ecuaciones del MLM es necesario encontrar una inversa generalizada, así sea la matriz $\mathbf{C}$ esa inversa generalizada dada por:

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^t\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{R}^{-1}\mathbf{X} + \mathbf{G}^{-1} \end{bmatrix}^- \tag{3.20}$$

De lo anterior, suponga que $\widetilde{\mathbf{b}}$ y $\widetilde{\mathbf{u}}$ son las soluciones de las ecuaciones del modelo mixto para $\mathbf{b}$ y $\mathbf{u}$, respectivamente. Así, entonces, se tiene que:

- El *BLUE* del vector de funciones lineales estimable de $\mathbf{K}^t\mathbf{b}$ es $BLUE(\mathbf{K}^t\mathbf{b}) = \mathbf{K}^t(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{V}^{-1}\mathbf{y} = \mathbf{K}^t\widetilde{\mathbf{b}}$ donde $\mathbf{K}$ es una matriz de orden $p \times q$ de coeficientes de la función lineal estimable $\mathbf{K}^t\mathbf{b}$, es decir, $\mathbf{K} = \mathbf{X}^t\mathbf{A}$ para alguna matriz $\mathbf{A}$ y $\mathbf{V} = \mathbf{Z}^t\mathbf{G}\mathbf{Z} + \mathbf{R}$.
- El *BLUP* del vector de funciones lineales de los efectos fijos y aleatorios, $\mathbf{K}^t\mathbf{b}+\mathbf{L}^t\mathbf{u}$, es $BLUP(\mathbf{K}^t\mathbf{b}+\mathbf{L}^t\mathbf{u}) = BLUE(\mathbf{K}^t\mathbf{b})+\mathbf{L}^t\mathbf{G}\mathbf{Z}^t\mathbf{V}^{-1}(\mathbf{y}-BLUE(\mathbf{X}\mathbf{b})) = \mathbf{K}^t\widetilde{\mathbf{b}} + \mathbf{L}^t\widetilde{\mathbf{u}}$ donde $\mathbf{L}$ es una matriz de orden $r \times q$ arbitraria de coeficientes y $BLUE(\mathbf{X}\mathbf{b}) = \mathbf{X}\widetilde{\mathbf{b}}$.

Derivado de las soluciones de las ecuaciones del MLM, se enumeran las siguientes propiedades importantes:

1. En la clase de predictores lineales insesgados, el *BLUP* maximiza la correlación entre $\mathbf{u}$ y $\widetilde{\mathbf{u}}$.
2. $\mathbf{K}^t\widetilde{\mathbf{b}}$ es el *BLUE* del conjunto de funciones lineales estimables $\mathbf{K}^t\mathbf{b}$.
3. $\mathrm{E}\left[\mathbf{u} \mid \widetilde{\mathbf{u}}\right] = \widetilde{\mathbf{u}}$.
4. $\widetilde{\mathbf{u}}$ es único.
5. $\mathbf{K}^t\widetilde{\mathbf{b}} + \mathbf{L}^t\widetilde{\mathbf{u}}$ es el *BLUP* de $\mathbf{K}^t\mathbf{b} + \mathbf{L}^t\mathbf{u}$ siempre que $\mathbf{K}^t\mathbf{b}$ sea estimable.
6. $\mathrm{Var}(\mathbf{K}^t\mathbf{b}) = \mathbf{K}^t\mathbf{C}_{11}\mathbf{K}$.
7. $\mathrm{Var}(\mathbf{K}^t\widetilde{\mathbf{b}} + \mathbf{L}^t\widetilde{\mathbf{u}}) = \mathbf{K}^t\mathbf{C}_{11}\mathbf{K} + \mathbf{L}^t(\mathbf{G} - \mathbf{C}_{22})\mathbf{L}$.
8. $\mathrm{Var}\left[(\mathbf{K}^t\widetilde{\mathbf{b}} + \mathbf{L}^t\widetilde{\mathbf{u}}) - (\mathbf{K}^t\mathbf{b} + \mathbf{L}^t\mathbf{u})\right] = (\mathbf{K}^t \quad \mathbf{L}^t)\mathbf{C}(\mathbf{K}^t \quad \mathbf{L}^t)^t$.
9. $\mathrm{Cov}(\mathbf{K}^t\widetilde{\mathbf{b}}, \widetilde{\mathbf{u}}^t) = \mathbf{0}$.
10. $\mathrm{Cov}(\mathbf{K}^t\widetilde{\mathbf{b}}, \mathbf{u}^t) = -\mathbf{K}^t\mathbf{C}_{12}$.
11. $\mathrm{Cov}(\mathbf{K}^t\widetilde{\mathbf{b}}, \mathbf{u}^t - \widetilde{\mathbf{u}}^t) = -\mathbf{K}^t\mathbf{C}_{12}$.
12. $\mathrm{Var}(\widetilde{\mathbf{u}}) = \mathrm{Cov}(\widetilde{\mathbf{u}}, \mathbf{u}^t) = \mathbf{G} - \mathbf{C}_{22}$.
13. $\mathrm{Var}(\widetilde{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{22}$.

Por otro lado, sea $\widetilde{\mathbf{w}} = \mathbf{\Lambda}^t(\widetilde{\mathbf{b}}^t \quad \widetilde{\mathbf{u}}^t) = \mathbf{K}^t\widetilde{\mathbf{b}} + \mathbf{L}^t\widetilde{\mathbf{u}}$, el *BLUP* de $\mathbf{w} = \mathbf{K}^t\mathbf{b} + \mathbf{L}^t\mathbf{u}$. Entonces, de acuerdo a la propiedad (8), el error cuadrado medio *(ECM)* de $\widetilde{\mathbf{w}}$ es:

$$\text{ECM}(\widetilde{\mathbf{w}}) = \text{E}\left[(\widetilde{\mathbf{w}} - \mathbf{w})(\widetilde{\mathbf{w}} - \mathbf{w})^t\right] = \text{Var}(\widetilde{\mathbf{w}} - \mathbf{w}) = \Lambda^t\mathbf{C}\Lambda = \mathbf{M}_{\widetilde{\mathbf{w}}} \tag{3.21}$$

Si los componentes de varianza son conocidos, trivialmente se obtiene la estadística pivote tipo Wald usada para hacer inferencia sobre $\mathbf{w}$ con su distribución (nula) exacta:

$$\mathbf{Q} = (\widetilde{\mathbf{w}} - \mathbf{w})^t(\mathbf{\Lambda}^t\mathbf{C}\mathbf{\Lambda})^{-1}(\widetilde{\mathbf{w}} - \mathbf{w}) \sim \chi^2_q \tag{3.22}$$

con $q = rango(\mathbf{\Lambda}^t)$.

Por otro lado, si los componentes de varianza son desconocidos y los valores estimados están disponibles, $\widehat{\mathbf{C}}$, para hacer inferencia estadística simultanea sobre los efectos fijos y sobre los efectos aleatorios $\mathbf{w} = \mathbf{\Lambda}^t(\mathbf{b}^t \quad \mathbf{u}^t)^t$, basado en el *EBLUP*, $\widehat{\mathbf{w}} = \mathbf{\Lambda}^t(\widehat{\mathbf{b}}^t \quad \widehat{\mathbf{u}}^t)^t$ es natural considerar la siguiente estadística:

$$\mathbf{Q} = (\widetilde{\mathbf{w}} - \mathbf{w})^t(\mathbf{\Lambda}^t\widehat{\mathbf{C}}\mathbf{\Lambda})^{-1}(\widetilde{\mathbf{w}} - \mathbf{w}) \tag{3.23}$$

con $q = rango(\mathbf{\Lambda}^t)$.

Como un caso especial, si w es una función de una sola dimensión dada por $w = \boldsymbol{\lambda}^t(\mathbf{b}^t \quad \mathbf{u}^t)^t = \mathbf{k}^t\mathbf{b} + \mathbf{l}^t\mathbf{u}$, es natural considerar la cantidad pivote:

$$t = \frac{\widehat{w} - w}{\sqrt{\boldsymbol{\lambda}^t\widehat{\mathbf{C}}\boldsymbol{\lambda}}} \tag{3.24}$$

donde $\widehat{w} = \boldsymbol{\lambda}^t(\widehat{\mathbf{b}}^t \quad \widehat{\mathbf{u}}^t)^t$ es el *EBLUP* de w.

La distribución nula de la estadística t es comúnmente aproximada mediante la distribución $t - Student$ con ν grados de libertad estimados mediante la aproximación de Satterthwaite (1946).

Por lo tanto, con ayuda de los resultados obtenidos por Witkovsky (2012), es posible deducir la distribución de probabilidad aproximada de nuestro estadístico de prueba.

Note que, tanto T_a como T_h, se pueden escribir bajo la siguiente forma general:

$$T = \frac{\boldsymbol{\lambda}^t \widehat{\mathbf{U}}}{\sqrt{\text{Var}(\boldsymbol{\lambda}^t \widehat{\mathbf{U}})}} = \frac{\boldsymbol{\lambda}^t \widehat{\mathbf{U}}}{\sqrt{\boldsymbol{\lambda}^t \text{Var}(\widehat{\mathbf{U}}) \boldsymbol{\lambda}}} \tag{3.25}$$

donde $\boldsymbol{\lambda}$ es un vector de constantes conocidas que define un contraste.

Y dado que se desconoce la verdadera varianza de $\widehat{\mathbf{U}}$, en su lugar se utiliza una estimación $\widehat{\text{Var}}(\widehat{\mathbf{U}})$. Por lo tanto:

$$T = \frac{\boldsymbol{\lambda}^t \widehat{\mathbf{U}}}{\sqrt{\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}}) \boldsymbol{\lambda}}} \overset{\text{aprox.}}{\sim} t_{\widehat{\nu}} \tag{3.26}$$

donde $\widehat{\nu}$ es la estimación de los grados de libertad ν mediante el método de Satterthwaite (1946). Y la

$$\widehat{\text{Var}}(\widehat{\mathbf{U}}) = \widehat{\mathbf{D}}_2 \mathbf{Z}_2^t \widehat{\mathbf{P}} \mathbf{Z}_2 \widehat{\mathbf{D}}_2 \tag{3.27}$$

donde $\widehat{\mathbf{P}} = \widehat{\mathbf{V}}^{-1} - \widehat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}^t\widehat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\widehat{\mathbf{V}}^{-1}$ y $\widehat{\mathbf{V}} = \mathbf{Z}_1\widehat{\mathbf{D}}_1\mathbf{Z}_1^t + \mathbf{Z}_2\widehat{\mathbf{D}}_2\mathbf{Z}_2^t + \mathbf{R}$.

La aproximación de Satterthwaite está basado en el supuesto:

$$\frac{\nu(\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}}) \boldsymbol{\lambda})}{\sigma^2} \sim \chi^2_\nu \tag{3.28}$$

para parámetros σ^2 y ν.

Comparando el primer y segundo momentos de la variable aleatoria χ^2_ν por sus respectivos valores teóricos:

$$\text{E}\left[\frac{\nu\left(\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)}{\sigma^2}\right] = \nu \quad \text{y} \quad \text{Var}\left[\frac{\nu\left(\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)}{\sigma^2}\right] = 2\nu$$

se obtiene directamente:

$$\sigma^2 = \text{E}\left[\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right] \quad \text{y} \quad \nu = \frac{2\left[\text{E}\left(\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)\right]^2}{\text{Var}\left(\boldsymbol{\lambda}^t \widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)}$$

Dado que E $\left(\boldsymbol{\lambda}^t\widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)$ y Var $\left(\boldsymbol{\lambda}^t\widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)$ dependen de parámetros desconocidos y pueden ser estimados. Así un estimador natural de ν es:

$$\widehat{\nu} = \frac{2\left(\boldsymbol{\lambda}^t\widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)^2}{\text{Var}\left(\boldsymbol{\lambda}^t\widehat{\text{Var}}(\widehat{\mathbf{U}})\boldsymbol{\lambda}\right)} \tag{3.29}$$

Con la estimación de los grados de libertad de la distribución *t-student* se concluye el desarrollo de la metodología propuesta y con ello, la prueba estadística está totalmente identificada, por lo que dada una muestra aleatoria de un rasgo de interés y una matriz de marcadores genéticos genotipados de una población ya se está en la posibilidad de determinar asociaciones entre *SNP* y rasgo utilizando la propuesta. En el siguiente apartado se obtienen ensayos de la prueba estadística propuesta mediante simulación.

3.3. Simulación

Una parte importante en el proceso de presentación de la prueba es la de verificar que efectivamente está cumpliendo con la función para lo cual fue desarrollada. En nuestro caso se ensayará la prueba mediante simulación y para ello es necesario describir los componentes que la integran así como posibles escenarios sobre los cuales la prueba puede ser aplicada.

Es importante señalar que las configuraciones usadas para ensayar la prueba estadística no son restrictivas a otras situaciones donde la prueba pueda ser aplicable de manera confiable. Los escenarios aquí descritos solo representan ejemplos para mostrar la eficiencia de la prueba y no son escenarios exclusivos para ser aplicada.

3.3.1. Construcción de la simulación

Para la simulación se identifican dos caminos. El primer camino, y el más largo, es simular la variable respuesta y mediante la función de distribución normal asumiendo cierta media y varianza, $\mathbf{y} \sim N(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. Para detectar asociaciones, en el término $\boldsymbol{\mu}_y$ se incluiría el efecto de una media global y los tamaños de efectos de algunos *SNPs*

para que en el ensayo, la prueba estadística propuesta los pueda detectar. Por otro lado, $\boldsymbol{\Sigma}_y$ estaría compuesta por matrices de varianza y covarianza, con una estructura definida, del tipo $\mathbf{D}_1$, $\mathbf{D}_2$ y $\mathbf{R}$ afectadas por un diseño experimental, también, supuesto. Generada $\mathbf{y}$ se estaría en la posibilidad de ajustar un MLM como (3.7) y del ajuste obtener $\widetilde{\mathbf{U}}$. Con $\widetilde{\mathbf{U}}$ ya es posible aplicar la prueba. Sin embargo, esta opción requiere del ajuste de un MLM por cada escenario propuesto, lo que convierte a este camino en algo tedioso de explorar. El segundo camino, y el más apropiado, es partir de $\widetilde{\mathbf{U}}$ considerando que $\widetilde{\mathbf{U}} \sim N_{q_2}(\mathbf{0}, \boldsymbol{\Sigma}_{\widetilde{U}})$ y, después generar $\mathbf{U}_{cem} = \widetilde{\mathbf{U}} + \boldsymbol{\eta}$, donde $\boldsymbol{\eta}$ contiene los tamaños de efectos de algunos *SNPs* que se asumen con efecto no nulo.

Es importante aclarar que la simulación está enfocada, solamente a ensayar la prueba estadística propuesta bajo el supuesto de efectos aditivos, es decir se probará la hipótesis nula H_{0a} usando el estadístico de prueba T_a. La aclaración debe hacerse debido a que en el proceso de simulación, la construcción del vector de tamaños de efectos $\boldsymbol{\eta}$ depende de cuál hipótesis nula se desea contrastar.

Identificación de elementos para la simulación

Note que el mejor predictor lineal insesgado de $\mathbf{U}$ dado por $\widetilde{\mathbf{U}}$ obtenido en (3.16), obedece a la distribución de probabilidad:

$$\widetilde{\mathbf{U}} \sim N_{q_2}\left(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{U}}\right) \tag{3.30}$$

donde $\boldsymbol{\Sigma}_{\widetilde{\mathbf{U}}} = \mathbf{D}_2\mathbf{Z}_2^t\mathbf{P}\mathbf{Z}_2\mathbf{D}_2$ con $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}^t\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{V}^{-1}$ y $\mathbf{V}^{-1} = \left(\mathbf{Z}_1\mathbf{D}_1\mathbf{Z}_1^t + \mathbf{Z}_2\mathbf{D}_2\mathbf{Z}_2^t + \mathbf{R}\right)^{-1}$.

A partir de la forma de $\boldsymbol{\Sigma}_{\widetilde{U}}$ es posible identificar los elementos necesarios para generar $\widetilde{\mathbf{U}}$, los cuales se enumeran enseguida:

1. Matrices generadas por el diseño experimental: $\mathbf{X}$, $\mathbf{Z}_1$ y $\mathbf{Z}_2$. Estás matrices son generadas a partir de un diseño *alfa-latice* determinado por la elección de número de genotipos, número de repeticiones y tamaño de bloque. Con ayuda de la librería *agricolae* instalada en el entorno *R-project* se generó este diseño.

2. Matrices generadas por supuestos en estructuras de varianzas y covarianzas: $\mathbf{D}_1$, $\mathbf{D}_2$ y $\mathbf{R}$. Para presentar las estructuras de las matrices $\mathbf{D}_1$, $\mathbf{D}_2$ y $\mathbf{R}$ es impor-

tante señalar que para la simulación solo se consideró un ambiente, por lo que $\mathbf{D}_2$ representa la matriz de varianzas y covarianzas de los genotipos ya que no hay efecto de interacción genotipo-ambiente. Estas matrices tienen la siguiente forma:

- Var$(\mathbf{b}) = \mathbf{D}_1 = \sigma_b^2 \mathbf{I}_{q_1}$, donde σ_b^2 es la varianza entre bloques y $\mathbf{I}_{q_1}$ es la matriz identidad de dimensión $q_1 \times q_1$. Para todos los escenarios, la varianza entre bloques se consideró fija, $\sigma_b^2 = 2$.
- Var$(\mathbf{g}) = \mathbf{D}_2 = \sigma_g^2 \mathbf{G}$, donde σ_g^2 es la varianza genética y $\mathbf{G}$ es una matriz de relaciones genéticas de dimensión $q_2 \times q_2$. La matriz $\mathbf{G}$ se construye a partir de las frecuencias alélicas de los marcadores genéticos *SNPs* tal como lo propone VanRaden (2008).
- Var$(\mathbf{e}) = \mathbf{R} = \sigma_e^2 \mathbf{C}$, donde σ_e^2 es la varianza residual y $\mathbf{C}$ es una matriz de correlaciones entre residuales. Para la simulación se consideraron dos estructuras de correlación para $\mathbf{C}$, la primera es tomar $\mathbf{C} = \mathbf{I}_n$ y la segunda es $\mathbf{C} = AR(1)$, es decir, considerar una estructura de correlación espacial ρ como Autorregresivo de primer orden, donde ρ toma diferentes magnitudes de correlación.

Los elementos para obtener $\widetilde{\mathbf{U}}$ están completamente identificados y determinados, lo que sigue es describir la forma en cómo se construye el vector $\boldsymbol{\eta}$ que contiene los efectos de los *SNPs* seleccionados de antemano para que la prueba propuesta los detecte y así evaluar su capacidad de detectar verdaderas asociaciones y su capacidad de controlar la tasa de falsos positivos. A continuación se enumeran los pasos para obtener $\boldsymbol{\eta}$:

1. Se seleccionan sistemáticamente x_{j_1}, x_{j_2},...,$x_{j_{ns}}$ marcadores genéticos *SNPs* de la matriz $\mathbf{M}$ que contiene m marcadores genéticos.
2. Dado los ns marcadores se construye el vector $\boldsymbol{\gamma}$ que contiene el tamaño del efecto de cada x_{j_i} con $i = 1, 2, ..., ns$.
   - La asignación de los tamaños de los efectos se basa en:

$$\sigma_g^2 = \sum_{i=1}^{ns} \gamma_i^2$$

   donde γ_i representa el tamaño del efecto del *SNP* i con $i = 1, 2, ..., ns$.

- Note que la expresión anterior se puede escribir como:

$$\sigma_g^2 = \sum_{i=1}^{ns} a_i^2 \gamma^2$$

donde a_i es el peso del *SNP* i y γ es el tamaño total del efecto de los ns *SNPs*.

- La asignación de tamaños de efectos parte del hecho de que:

$$\gamma = \sqrt{\frac{\sigma_g^2}{\sum_{i=1}^{ns} a_i^2}}$$

- Entonces, el tamaño del efecto del *SNP* i, para $i = 1, 2, ..., ns$, es:

$$\gamma_i = a_i \gamma$$

- En la posición $j_i \in \{1, 2, ..., m\}$ de $\boldsymbol{\gamma}$ se coloca el valor γ_i, para $i = 1, 2, ..., ns$, y en el resto de posiciones se coloca cero. Note que $\boldsymbol{\gamma}$ es de dimensión m.
- Por lo tanto $\boldsymbol{\eta}$, para probar efectos aditivos, está dada por:

$$\boldsymbol{\eta} = \mathbf{M}\boldsymbol{\gamma}$$

donde $\mathbf{M}$ es la matriz de marcadores *SNPs* de dimensión $q_2 \times m$.

- Si en lugar de probar efectos aditivos se deseara probar efectos de heterosis, la construcción de $\boldsymbol{\eta}$ tendría una ligera modificación. En cada columna $j_i \in \{1, 2, ..., m\}$ de $\mathbf{M}$, para $i = 1, 2, ..., ns$, los ceros serían sustituidos por un valor constante β. Con está modificación se obtendría una matriz $\mathbf{M}^*$ y entonces $\boldsymbol{\eta} = \mathbf{M}^*\boldsymbol{\gamma}$. En la práctica se sugiere usar β =1.5.

La derivación de $\boldsymbol{\eta}$ pareciera ser complicada, sin embargo en la práctica es muy sencilla. Dado el valor σ_g^2 y los pesos a_i que tendrán cada uno de los ns *SNPs* seleccionados sistemáticamente de $\mathbf{M}$ es sencillo y rápido de calcular.

3.3.2. Escenarios a simular

Los escenarios simulados fueron seleccionados tratando de obtener la mayor representatividad. Un ejemplo para obtener representatividad en los escenarios a simular para contabilizar el efecto del diseño experimental en los resultados del *GWAS* es variar el tamaño del bloque y por lo tanto el número de bloques. Por otro lado, para obtener representatividad de los factores genéticos que afectan directamente al rasgo se asumirán diferentes valores de la varianza genética σ_g^2 y con ello se estará considerando diferentes heredabilidades h^2.

Elección de valores para determinar el diseño experimental

Para construir el diseño experimental *alfa-latice* es necesario conocer tres elementos esenciales, el número de tratamientos, en nuestro caso el número de genotipos, el número de repeticiones y el tamaño de bloque. Con estos tres componentes es posible obtener el arreglo en campo y, por lo tanto, conocer la estructura de $\mathbf{X}$, $\mathbf{Z}_1$ y $\mathbf{Z}_2$. En el cuadro (3.1) se exponen los valores asumidos a cada elemento del diseño experimental y sus combinaciones.

Genotipos	Tamaño de bloque	Repeticiones	Bloques totales
1000	5	2	400
1000	10	2	200
1000	20	2	100
3000	5	2	1200
3000	10	2	600
3000	20	2	300

Cuadro 3.1: Elección de valores para el diseño experimental *alfa-latice.*

Con las combinaciones de valores asumidos se obtienen varios escenarios (seis escenarios), donde cada uno determina la configuración de las matrices $\mathbf{X}$, $\mathbf{Z}_1$, $\mathbf{Z}_2$.

Elección de valores para componentes de varianza y correlación

Recuerde que $\sigma_b^2 = 2$ y $\sigma_e^2 = 1$ fueron fijados para todos los escenarios a simular. Para σ_g^2 se tomaron en cuenta tres valores, 0.3, 1 y 3, que corresponden a tres valores de h^2, 0.38, 0.67 y 0.86. La heredabilidad h^2 es la proporción de la varianza fenotípica explicada por factores genéticos. $\mathbf{D}_1$ se asume como $\sigma_b^2\mathbf{I}_{q_1}$ y $\mathbf{D}_2 = \sigma_g^2\mathbf{G}$ con $\mathbf{G}$ calculada a partir de las frecuencias alélicas de $\mathbf{M}$. Estas estructuras de matrices se mantienen a lo largo de todos los escenarios. Las entradas de $\mathbf{G}$ cambian conforme cambia $\mathbf{M}$.

Por otra parte, para especificar la estructura de $\mathbf{R}$ es necesario partir del supuesto que $\mathbf{R} = \sigma_e^2\mathbf{C}$ donde $\mathbf{C}$ es una matriz de correlaciones entre residuales, de la cuál se asumen dos tipos de estructuras.

1. $\mathbf{C} = \mathbf{I}_n$ representa el caso de observaciones independientes.
2. $\mathbf{C} = AR(1)$ donde $AR(1)$ representa una estructura de correlación espacial autorregresiva de primer orden. Esta estructura solo depende de ρ. La entrada (i, j) de $\mathbf{C}$ será $\rho^{|i-j|}$ para $i, j = 1, 2, ..., n$. Los valores para ρ son 0.25 y 0.6.

En el cuadro (3.2) se muestran las combinaciones de valores entre σ_g^2 y $\mathbf{C}$ para determinar escenarios para los componentes de varianza.

σ_g^2	$\mathbf{C}$
0.3	$\mathbf{I}_n$
0.3	$AR(1)$, ρ =0.25, 0.60
1.0	$\mathbf{I}_n$
1.0	$AR(1)$, ρ =0.25, 0.60
3.0	$\mathbf{I}_n$
3.0	$AR(1)$, ρ =0.25, 0.60

Cuadro 3.2: Combinaciones entre la varianza genética σ_g^2 y la correlación $\mathbf{C}$.

De las combinaciones entre σ_g^2 y $\mathbf{C}$ se obtienen nueve escenarios a simular. Retomando los seis escenarios para el diseño experimental y considerando que cada diseño experimental será implementado con cada una de las combinaciones entre σ_g^2 y $\mathbf{C}$, el total de escenarios a simular, hasta este momento, es de 54. Sin embargo, aún hace falta incorporar las configuraciones para determinar $\boldsymbol{\eta}$.

Configuraciones para determinar $\boldsymbol{\eta}$

La construcción de $\boldsymbol{\eta}$ involucra varios elementos. En primer lugar la matriz $\mathbf{M}$ fue simulada asumiendo un diseño de mejoramiento que consiste primero de cruzar dos padres contrastantes generando un híbrido. De este híbrido, por autofecundación, se genera una población segregante. El proceso que genera la variabilidad genética imita lo que pasa en los organismos vivos, en que el número de recombinaciones proviene de una distribución de Poisson y las posiciones de estas recombinaciones es uniformemente distribuida en los cromosomas. Se fijo en 10 el número de cromosomas a simular y en cada cromosoma habían $m/10$ marcadores genéticos simulados. Se consideraron tres tamaños para m: 1000, 5000 y 15000. Todos los marcadores genéticos fueron simulados considerando control de calidad.

De $\mathbf{M}$ se seleccionó un subconjunto de *SNPs* para asignarles efectos diferentes de cero. El tamaño del subconjunto es representado por ns que tomó tres valores: 5, 20 y 200. Otro elemento es el peso a_i asignado a cada uno de los ns *SNPs*. La determinación de estos pesos tienen dos objetivos, uno es comparar escenarios donde cada *SNP* tienen el mismo peso y dos, comparar escenarios donde los *SNPs* tienen pesos distintos. En el primer caso $a_i = 1/ns$ para $i = 1, 2, ..., ns$. En el Cuadro (3.3) se muestran las combinaciones de los elementos que generan $\boldsymbol{\eta}$.

ns	a_i	σ_g^2	Tamaño del efecto total γ	Tamaño del efecto individual γ_i
5	0.2	0.3	1.2247	0.2449
	0.2	1.0	2.2361	0.4472
	0.2	3.0	3.8730	0.7746
20	0.05	0.3	2.4495	0.1225
	0.05	1.0	4.4721	0.2236
	0.05	3.0	7.7460	0.3873
200	0.005	0.3	7.7460	0.0387
	0.005	1.0	14.1421	0.0707
	0.005	3.0	24.4949	0.1225

Cuadro 3.3: Combinaciones entre ns, a_i y γ_i.

Hasta este momento, los escenarios descritos se pueden resumir como:

- m =1000 *SNPs* a probar:
  - q_2 =1000 genotipos, $ns = 5$, 20, 200, tamaño de bloque 5, 10 y 20, varianza genética $\sigma_g^2 =$ 0.3, 1.0 y 3.0, y $\mathbf{C} = \mathbf{I}_n$, $AR(1)$ con $\rho =$ 0.25 y 0.6. Total de escenarios = 81.
  - q_2 =3000 genotipos, tamaño de bloque 20, $ns = 5$, 20 ,200, varianza genética $\sigma_g^2 =$ 0.3, 1.0 y 3.0, y $\mathbf{C} = \mathbf{I}_n$, $AR(1)$ con $\rho =$ 0.25 y 0.6. Total de escenarios = 27.
- m =5000 *SNPs* a probar:
  - q_2 =1000 genotipos, tamaño de bloque 20, $ns = 5$, 20, 200, varianza genética $\sigma_g^2 =$ 0.3, 1.0 y 3.0, y $\mathbf{C} = \mathbf{I}_n$, $AR(1)$ con $\rho =$ 0.25. Total de escenarios = 18.
  - q_2 =3000 genotipos, tamaño de bloque 20, $ns = 5$, 20, 200, varianza genética $\sigma_g^2 =$ 0.3, 1.0 y 3.0, y $\mathbf{C} = \mathbf{I}_n$, $AR(1)$ con $\rho =$ 0.25. Total de escenarios = 18.
- m =15000 *SNPs* a probar:
  - q_2 =1000 genotipos, tamaño de bloque 20, $ns = 5$, 20, 200, varianza genética $\sigma_g^2 =$ 0.3, 1.0 y 3.0, y $\mathbf{C} = \mathbf{I}_n$, $AR(1)$ con $\rho =$ 0.25. Total de escenarios = 18.
  - q_2 =3000 genotipos, tamaño de bloque 20, $ns = 5$, 20, 200, varianza genética $\sigma_g^2 =$ 0.3, 1.0 y 3.0, y $\mathbf{C} = \mathbf{I}_n$, $AR(1)$ con $\rho =$ 0.25. Total de escenarios = 18.
- Por lo tanto, son $81 + 27 + 18 + 18 + 18 + 18 = 180$ escenarios simulados.

La determinación de los siguientes escenarios tiene por objetivo mostrar que el desempeño de la prueba propuesta para detectar verdaderas asociaciones tiene relación directa con el tamaño del efecto del *SNP* probado. Así; entonces, se fijó ns =200, σ_g^2 =1 y se asignaron pesos a_i de la siguiente forma:

1. Un solo *SNP* con un peso de 0.2 y 199 *SNPs* con un peso de 0.8/199.
2. Dos *SNPs*, cada uno con un peso de 0.2 y 198 *SNPs* con un peso de 0.6/198.
3. Tres *SNPs*, cada uno con un peso de 0.2 y 197 *SNPs* con un peso de 0.4/197.

4. Cuatro *SNPs*, cada uno con un peso de 0.2 y 196 *SNPs* con un peso de 0.2/196.

En el Cuadro (3.4) se muestran los efectos correspondientes a cada *SNP*. El resto de elementos involucrados en la simulación son presentados en el Cuadro (3.5). Considerando las combinaciones de los elementos descritos se obtuvieron 24 escenarios a simular.

SNPs con a_i =0.2	Tamaño del efecto para cada *SNP* con a_i =0.2	Tamaño del efecto total para todos los *SNPs* con a_i =0.2	Tamaño del efecto para cada *SNP* restante
1	2.8284	2.8284	0.0569
2	2.8284	5.6569	0.0429
3	2.8284	8.4853	0.0287
4	2.8284	11.3137	0.0144

Cuadro 3.4: Tamaños del efecto para *SNPs* cuyo peso es 0.2 y para el resto de los *SNPs*, con $ns = 200$.

SNPs	Genotipos	Tamaño de bloque	σ_g^2	$\mathbf{C}$
1000	1000	20	1.0	$AR(1)$, ρ =0.25
	3000	20	1.0	$AR(1)$, ρ =0.25
5000	1000	20	1.0	$AR(1)$, ρ =0.25
	3000	20	1.0	$AR(1)$, ρ =0.25
15000	1000	20	1.0	$AR(1)$, ρ =0.25
	3000	20	1.0	$AR(1)$, ρ =0.25

Cuadro 3.5: Combinaciones a simular considerando cada escenario del cuadro (3.4).

En total se simularon 180+24=204 escenarios. Los primeros 180 escenarios fueron analizados usando dos métodos de estudios de asociación. El primer método es el desarrollado en está investigación y el segundo es el desarrollado en el software *PLINK* asumiendo rasgo cuantitativo. Para los escenarios restantes solamente se aplicó el método propuesto. Los resultados son presentados en el siguiente capítulo.

Capítulo 4

RESULTADOS Y DISCUSIÓN

La simulación tiene el objetivo de mostrar la habilidad de la prueba estadística propuesta para discriminar entre *SNPs* aquellos con efecto significativo y aquellos con efecto nulo. Además, comparar resultados con el método implementado en *PLINK* es esencial, ya que se deben tener referencias para poder evaluar la metodología.

Antes de comenzar con la presentación y análisis de resultados es conveniente explicar de forma precisa sobre qué cálculos se hace el análisis. Como primer punto, considerar que cada escenario fue simulado 1000 veces y en cada escenario se calcularon dos cocientes, uno relacionado a la habilidad de la prueba de detectar verdaderas asociaciones y el otro relacionado con la capacidad de la prueba de no detectar asociaciones espurias.

Sea m el total de *SNPs* probados, de los cuales ns tienen efecto no nulo y $m - ns$ con efecto cero. Dado un método de estudio de asociación se tienen m *p-values*, cada uno correspondiente a cada marcador genético probado. Derivado de los *p-values*, un nivel de significancia nominal $\alpha = 0.05$ y la aplicación de una regla de decisión se tiene, m_1, el número de *SNPs* significativos y s_1 el número de *SNPs* con efecto no nulo, además $m_1 - s_1$ es el número de marcadores genéticos con efecto nulo. En el Cuadro (4.1) se presenta el panorama general resultado de aplicar una regla de decisión.

En teoría estadística, precisamente en pruebas de hipótesis se conocen dos tipos de errores. El error tipo I se comete cuando se rechaza la hipótesis nula H_0 siendo que es verdadera *(falso-positivo)* y el error tipo II se comete cuando no se rechaza

SNPs	Con efecto	Sin efecto	Total
Significativos	s_1	$m_1 - s_1$	m_1
No significativos	$ns - s_1$	$m - ns - m_1 + s_1$	$m - m_1$
Total	ns	$m - ns$	m

Cuadro 4.1: Resultado general en un estudio de asociación.

H_0 siendo que es falsa *(falso-negativo)*; sin embargo, estos errores se dan para una sola hipótesis nula. En nuestro caso, se tiene tantas hipótesis nulas como marcadores genéticos a probar y para poder tomar una decisión es necesario ajustar el valor crítico. Para el ajuste se utiliza la corrección de Bonferroni para pruebas múltiples; así, dado un nivel de significancia α y m *SNPs* a probar, la corrección de Bonferroni consiste en comparar el *p-value* del j-ésimo *SNP*, $j = 1, 2, ..., m$, con $\alpha^* = \alpha/m$. Por lo tanto, la regla de decisión consiste en rechazar H_{0j} si $p_j \leq \alpha^*$, de otro modo no se rechaza, con $p_j = \text{mín}\{p - value_j, 1\}$.

Regresando al cálculo de los cocientes. El primer cociente dado por s_1/ns es la razón del número de *SNPs* detectados por la prueba estadística como significativos y que de antemano se les asignó un tamaño de efecto distinto de cero (s_1) entre el total de *SNPs* que de antemano se les asignó un tamaño de efecto distinto de cero (ns). Note que este cociente se refiere a la proporción de verdaderas detecciones. El segundo cociente dado por $(m_1 - s_1)/(m - ns)$ es la división del número de *SNPs* detectados por la prueba estadística como significativos y que de antemano se sabe que tienen un tamaño de efecto cero $(m_1 - s_1)$ entre el total de *SNPs* que de antemano se les asignó un tamaño de efecto cero $(m - ns)$. El último cociente se puede interpretar como la tasa de falsos positivos en una prueba de estudio de asociación ya que representa la proporción de aquellos *SNPs* que estadísticamente tienen efecto sobre el rasgo de interés cuando en la realidad su efecto es nulo.

Con los dos cocientes calculados se derivaron dos indicadores que, simplemente, son los promedios de las 1000 iteraciones para cada cociente y se definen como sigue: el indicador TDV es el promedio de verdaderas detecciones y el indicador TDF es el promedio de falsas detecciones. Cada uno se calcula de la siguiente manera, respectivamente.

- $TDV = \frac{1}{1000} \sum_{i=1}^{1000} \frac{s_{1i}}{(ns)_i}$
- $TDF = \frac{1}{1000} \sum_{i=1}^{1000} \frac{(m_1 - s_1)_i}{(m - ns)_i}$

Ambos indicadores fueron calculados tanto para la propuesta como para el método implementado en *PLINK*. TDV y TDF ayudaron a comparar ambos métodos y sobre todo, evaluar el método propuesto. A continuación se analizan ambos indicadores por separado.

4.1. Indicador TDV

El indicador TDV tiene por objeto evaluar el desempeño de la prueba estadística propuesta con respecto a su capacidad de detectar *SNPs* que tienen efecto no nulo frente a otro método de asociación.

Un panorama general de la TDV entre ambos métodos se muestra en la Figura (4.1).

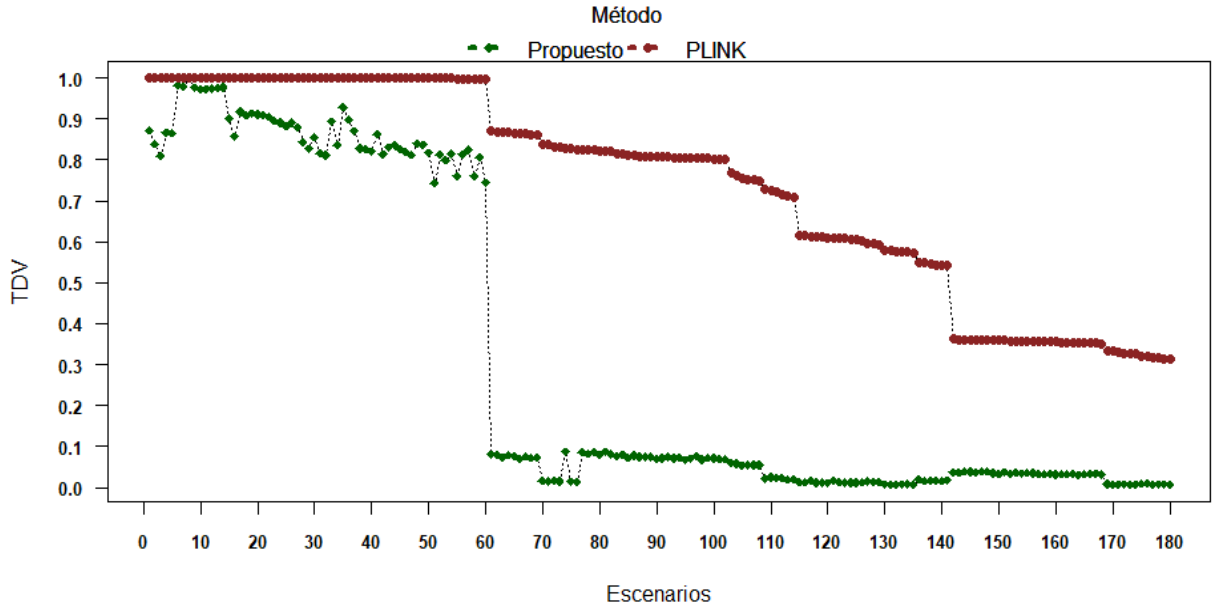

Figura 4.1: Comportamiento de la TDV entre métodos de estudios de asociación.

La Figura (4.1) exhibe diferencia entre ambos métodos, aunque se desconocen las

causantes específicas de está diferencia. Conocer las causantes que hacen diferentes, en términos estadísticos, ambos métodos permitirá determinar las fortalezas y debilidades de la propuesta.

Un análisis utilizado para identificar los factores que afectan la tasa de verdaderas detecciones es el *ANOVA*. En el Cuadro (4.2) se muestran los resultados del *ANOVA* solo para aquellos términos que tienen significancia estadística sobre la respuesta TDV.

Fuente	Grados de libertad	Sumas de cuadrados	Estadística F	*P-value*
MET	1	16.159	9233.6830	<2.2e-16
SNP	2	0.114	32.4450	1.308e-13
MCE	2	31.987	9139.3330	<2.2e-16
GEN	1	0.369	210.9926	<2.2e-16
MET:MCE	2	5.319	1519.6542	<2.2e-16
MET:GEN	1	0.099	56.7966	4.528e-13
SNP:MCE	4	0.078	11.0727	1.900e-08
SNP:GEN	2	0.019	5.4411	0.004725
MCE:GEN	2	0.071	20.1912	5.239e-09
Error	336	0.588		

Cuadro 4.2: ANOVA para TDV. MET = Método, SNP = *SNPs* a probar, MCE = *SNPs* con efecto, GEN = Genotipos y el resto son las interacciones.

Los resultados del *ANOVA* muestran que únicamente cuatro elementos influyen, por si solos, directamente sobre la tasa de detecciones verdaderas, los cuales son el método (el método propuesto y el método implementado en *PLINK*), el número de *SNPs* a probar, el número de *SNPs* con efecto y el número de genotipos. Las interacciones que resultaron significativas fueron cinco: método y *SNPs* con efecto (MET:MCE), método y genotipos (MET:GEN), número de *SNPs* a probar y número de *SNPs* con efecto (SNP:MCE), número de *SNPs* a probar y número de genotipos (SNP:GEN), y número de *SNPs* con efecto y número de genotipos (MCE:GEN). Solamente se consideraron interacciones entre dos factores.

El *ANOVA* encontró que el factor método es significativo; es decir, el efecto método está influenciando los resultados de TDV y, por lo tanto, existen diferencias entre los

dos niveles del factor. También, el número de *SNPs* a probar tiene efecto sobre la tasa de detecciones verdaderas; este efecto se debe a los niveles del factor y no hay interacción con el factor método. En el caso del número de *SNPs* con efecto se encontró que hay efecto de los niveles del factor sobre la TDV, además de que la interacción con el factor método es significativo. Esto último significa que existe efecto, también, entre niveles del factor MCE cuando se combina con el efecto del método. El caso del factor genotipos es similar al número de *SNPs* con efecto; sin embargo, su aporte a la suma de cuadrados es 86.7 veces menor que el factor MCE. Por último, las interacciones SNP:MCE, SNP:GEN y MCE:GEN reafirman la importancia que tienen en primer lugar el número de *SNPs* que son sometidos a prueba, luego el número de *SNPs* con efecto no nulo. Ambos factores influyen significativamente en la tasa de detecciones verdaderas, el primero influye en la regla de decisión mediante la corrección por comparaciones múltiples y el segundo está relacionado con el tamaño del efecto, a menor número de *SNPs* con efecto mayor es el tamaño del efecto y viceversa. Después está el número de genotipos que en términos prácticos es el tamaño de la muestra y tal como se esperaba, incide significativamente sobre la capacidad de la prueba de detectar asociaciones verdaderas, tanto como un efecto principal como interactuando con el factor método, número de *SNPs* a probar y número de *SNPs* con tamaño de efecto no nulo.

Un análisis posterior al *ANOVA* que ayuda a identificar diferencias entre niveles de los factores es la comparación múltiple de medias. El método utilizado para hacer las comparaciones es el enfoque HSD *(Honestly Significant Difference)* de Tukey. Los resultados obtenidos a través de este método se muestran a continuación ordenados con respecto al orden en como fueron incluidos en el *ANOVA*. Para diferenciar el efecto de los métodos se asignó la letra A para el método propuesto y la letra B para el método implementado en *PLINK*.

Contraste	Estimación	SE	g.l.	Razón T	P-value
A - B	-0.434	0.00462	336	-93.927	$<.0001$

Cuadro 4.3: Prueba HSD de Tukey para contrastar el método A y B.

Derivado de la prueba estadística (Cuadro 4.3) se encontraron diferencias significativas entre ambos métodos. La Figura (4.2) confirma está situación. Note que el método implementado en *PLINK* tiene una media mayor y la dispersión de los valores de TDV es menor comparado con el método A. El método A obtiene valores de TDV tan dis-

persos que se pueden agrupar en dos conjuntos, uno a partir de valores mayores a 0.6 y el otro a partir de valores menores a 0.2.

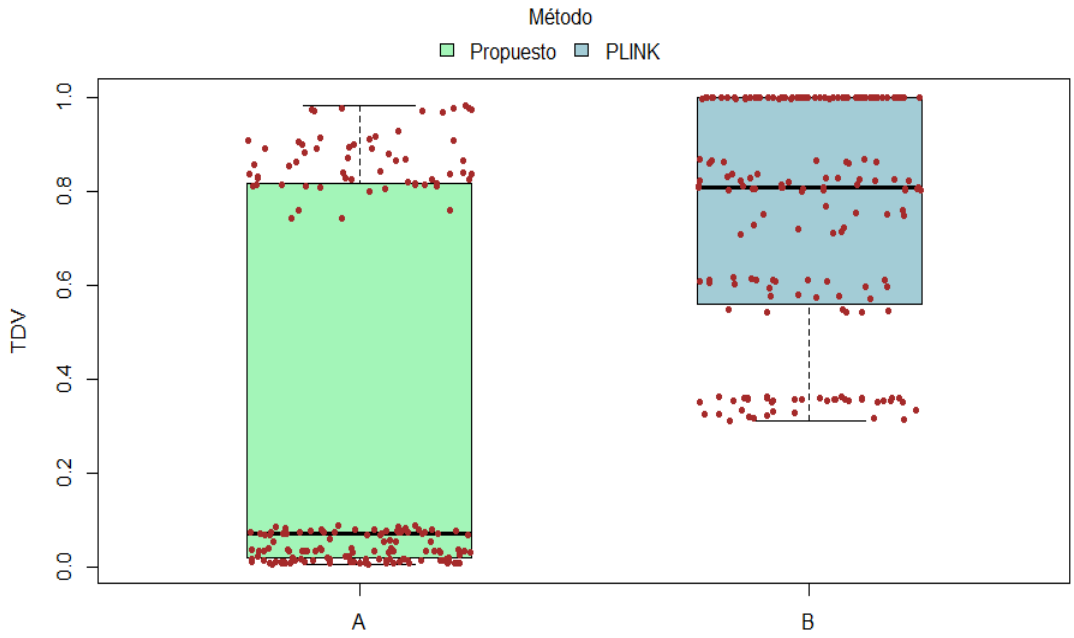

Figura 4.2: Comportamiento de la TDV entre métodos A y B.

Las comparaciones múltiples del factor *número de SNPs a probar* arrojan diferencias significativas entre todos los niveles del factor; sin embargo, esto debe tomarse con reserva ya que en el diseño se tienen más observaciones para el nivel 1000. La Figura (4.3) exhibe el comportamiento de la TDV agrupada por número de *SNPs* a probar. Note que la media de la TDV en el grupo 1000 es mayor a las medias agrupadas por 5000 y 15000, está situación se explica a que en el grupo 1000 existe mayor concentración de puntos por arriba del umbral 0.4, lo que probablemente esté haciendo que la media incremente. Sin embargo, la dispersión de los datos en los tres grupos apuntan a que los valores de TDV se pueden agrupar en cinco conjuntos bien definidos, es decir, existe influencia de otros factores en el *ANOVA* que explican este comportamiento. Entre niveles 5000 y 15000 no hay diferencia de acuerdo a la Figura (4.3).

Contraste	Estimación	SE	g.l.	Razón T	P-value
1000 - 5000	0.0356	0.00668	336	5.329	$<.0001$
1000 - 15000	0.0521	0.00668	336	7.803	$<.0001$
5000 - 15000	0.0165	0.00697	336	2.369	0.0482

Cuadro 4.4: Prueba HSD de Tukey para contrastar el factor SNP.

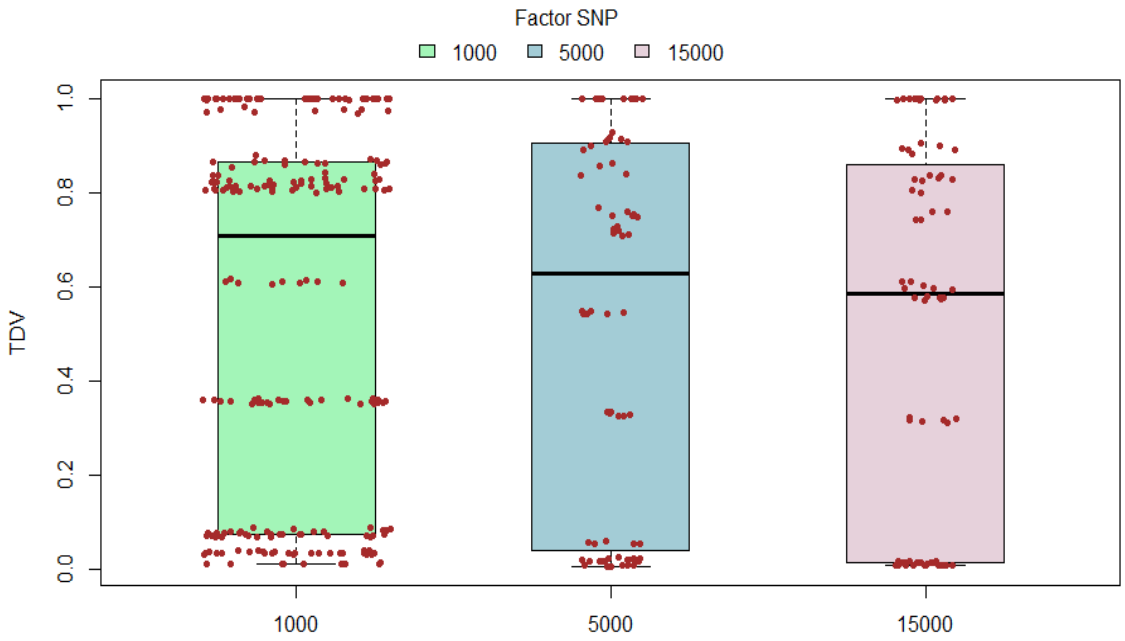

Figura 4.3: Comportamiento de la TDV entre niveles del factor SNP.

El factor SNP junto al factor MCE tienen significancia estadística, según *ANOVA*, por lo que el interés ahora es identificar en cuáles combinaciones de la interacción se encuentran esas diferencias. Se podrían hacer muchas comparaciones derivado de las interacciones; sin embargo, nuestro interés se restringe a identificar diferencias entre número de *SNPs* a probar agrupado por número de marcadores genéticos con efecto no nulo. Con estas restricciones se tienen nueve contrastes cuyos resultados se muestran en el Cuadro (4.5).

Contraste (SNP:MCE)	Estimación	SE	g.l.	Razón T	P-value
1000:5 - 5000:5	0.00352	0.0106	336	0.332	1.0000
1000:5 - 15000:5	0.03311	0.0106	336	3.126	0.0172
5000:5 - 15000:5	0.02959	0.0121	336	2.450	0.1254
1000:20 - 5000:20	0.07422	0.0106	336	7.006	$<.0001$
1000:20 - 15000:20	0.09649	0.0106	336	9.108	$<.0001$
5000:20 - 15000:20	0.02227	0.0121	336	1.844	0.4593
1000:200 - 5000:200	0.02897	0.0106	336	2.734	0.0577
1000:200 - 15000:200	0.02666	0.0106	336	2.516	0.1057
5000:200 - 15000:200	-0.00231	0.0121	336	-0.191	1.0000

Cuadro 4.5: Prueba HSD de Tukey para contrastar la interacción SNP:MCE.

Las comparaciones múltiples arrojaron que en los grupos 5 y 200 de *SNPs* con efecto distinto de cero no se encontró diferencia significativa entre niveles del número de *SNPs* a probar. En el caso del grupo 20 se encontró diferencia significativa en dos contrastes. Se encontró diferencia significativa entre el nivel 1000 y 5000 y entre el nivel 1000 y 15000. Está situación indica que después de cierto número de marcadores genéticos sometidos a prueba, este factor no tendría efecto sobre la detección de asociaciones. Una forma de visualizar los contrastes es mediante la Figura (4.4). Se confirman los resultados obtenidos en el Cuadro (4.4). En el grupo 20 es donde la media del nivel 1000 del factor SNP es ligeramente mayor comparada a los niveles 5000 y 15000.

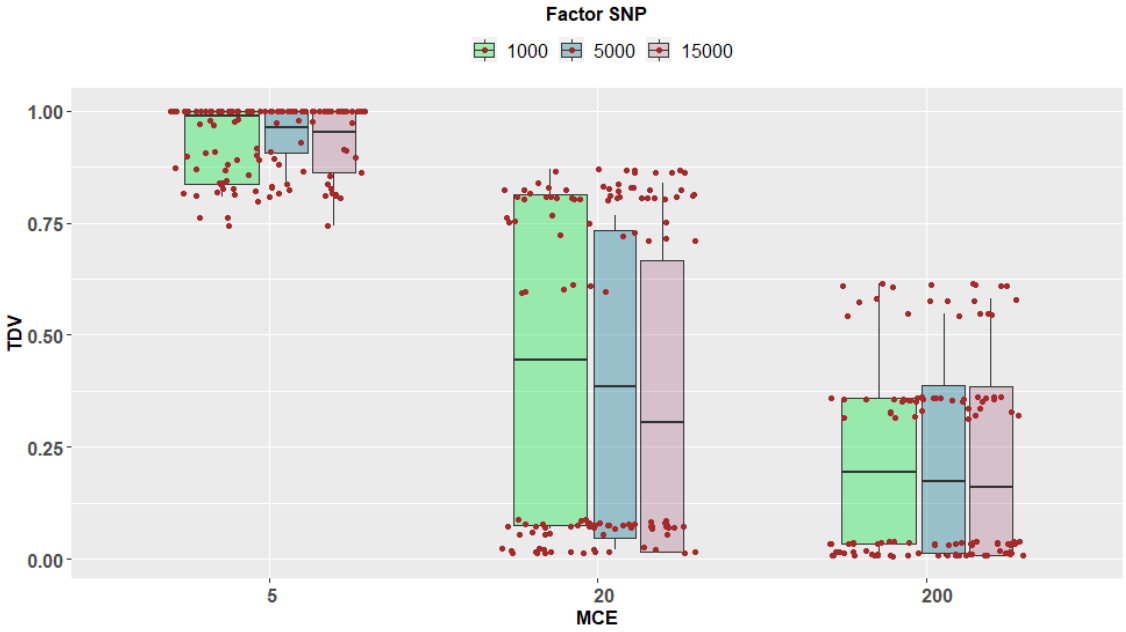

Figura 4.4: Comportamiento de la TDV entre niveles de la interacción SNP:MCE.

Los resultados en el Cuadro (4.6) corresponden a comparaciones múltiples de la interacción SNP:GEN utilizando contrastes agrupados por el número de genotipos. De todas las comparaciones solamente en una se encontró que no hay diferencia significativa. Esta situación se presenta en el nivel 3000 de GEN, entre el nivel 5000 y 15000 del factor SNP. El hecho de encontrar diferencias significativas entre todos lo niveles del factor SNP dentro del grupo de 1000 genotipos se debe a la importancia del tamaño de la muestra; es decir, la TDV es sensible a tamaños de muestra relativamente pequeños.

Contraste (SNP:GEN)	Estimación	SE	g.l.	Razón T	P-value
1000:1000 - 5000:1000	0.02859	0.00922	336	3.100	0.0125
1000:1000 - 15000:1000	0.06745	0.00922	336	7.313	$<$.0001
5000:1000 - 15000:1000	0.03885	0.00986	336	3.940	0.0006
1000:3000 - 5000:3000	0.04255	0.00922	336	4.613	$<$.0001
1000:3000 - 15000:3000	0.03673	0.00922	336	3.982	0.0005
5000:3000 - 15000:3000	-0.00582	0.00986	336	-0.590	0.9923

Cuadro 4.6: Prueba HSD de Tukey para contrastar la interacción SNP:GEN.

Las comparaciones de la interacción SNP:GEN se muestran en la Figura (4.5). La significancia estadística de los contrastes deben de tomarse con reserva ya que observando la Figura (4.5), el comportamiento de la TDV es bastante disperso en todos los grupos y no se puede asegurar si los valores de la TDV son superiores o tienen cierta homogeneidad en algunos grupos con respecto a otros.

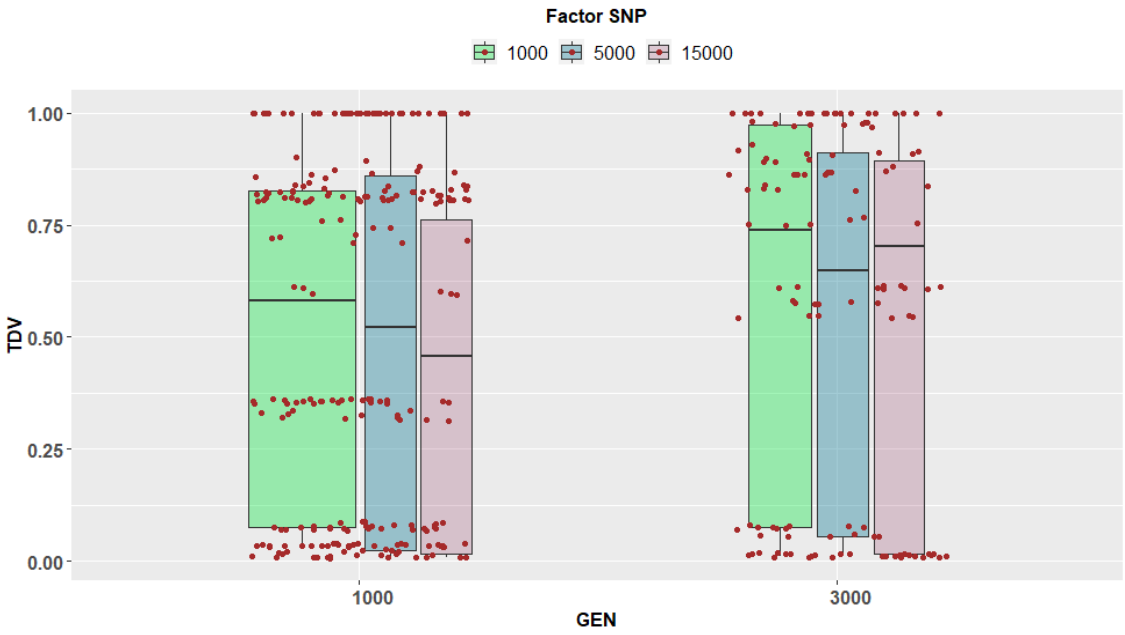

Figura 4.5: Comportamiento de la TDV entre niveles de la interacción SNP:GEN.

Otro componente que fue variante en el proceso de simulación es el número de *SNPs* con efecto no nulo (MCE) y que además es significativo de acuerdo al *ANOVA*. Por lo tanto, el siguiente paso es determinar entre cuáles niveles de este factor se encuentran las diferencias. Los resultados se muestran en el Cuadro (4.7). Las comparaciones múltiples muestran que existe diferencia en términos estadísticos entre todos los niveles del factor número de *SNPs* con tamaño de efecto diferente de cero. Note como está

situación se ve reflejada en la Figura (4.6). La TDV disminuye conforme aumenta el número de *SNPs* con efecto no nulo, lo que tiene que ver directamente con el tamaño del efecto. En la simulación se explicó la forma en que se asignaron los tamaños de efecto a cada uno de los *SNPs* que tienen efecto no nulo y, de manera general, se dedujo que conforme aumentaba el número de marcadores genéticos con efecto, el tamaño del efecto disminuía. Por lo tanto, la TDV es mejor para tamaños de efectos relativamente grandes y no deseada para tamaños de efectos relativamente pequeños.

Contraste	Estimación	SE	g.l.	Razón T	P-value
5 - 20	0.531	0.00623	336	85.230	$<.0001$
5 - 200	0.701	0.00623	336	112.534	$<.0001$
20 - 200	0.170	0.00623	336	27.304	$<.0001$

Cuadro 4.7: Prueba HSD de Tukey para contrastar el factor MCE.

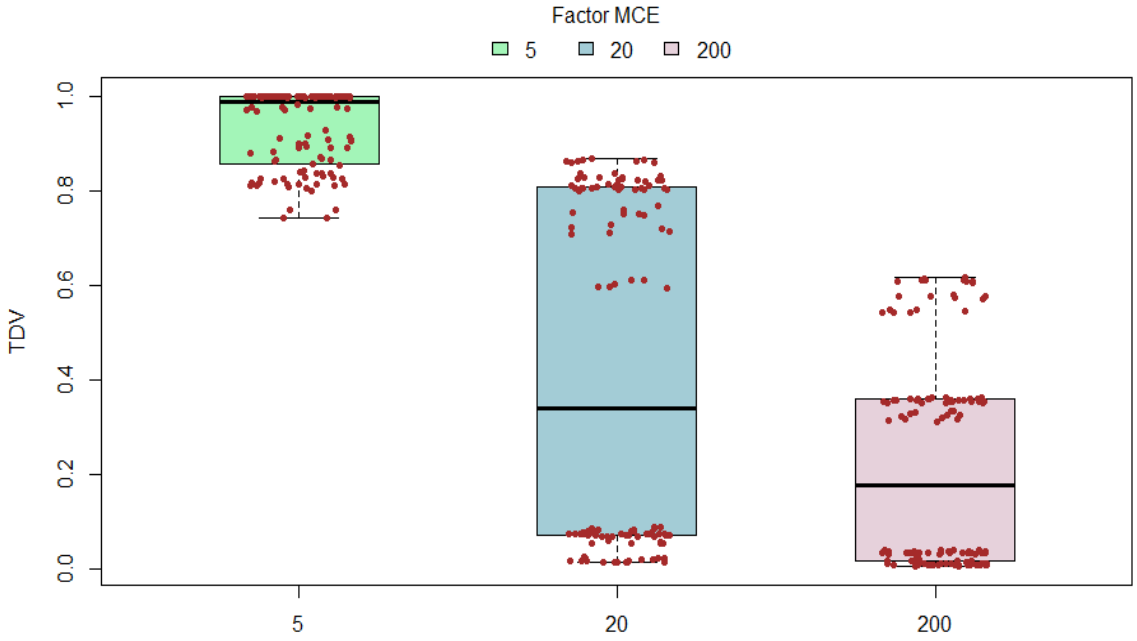

Figura 4.6: Comportamiento de la TDV entre niveles del factor MCE.

Continuando con el análisis del factor MCE, el siguiente paso es comparar ambos métodos de estudios de asociación dentro de cada nivel del factor MCE. En el Cuadro (4.8) se presentan los resultados de las pruebas estadísticas al comparar los valores de TDV de ambos métodos fijando los niveles de MCE. Los resultados exhiben diferencia significativa entre ambos métodos en cada nivel de MCE. Es decir, el comportamiento de la TDV cambia conforme se usa uno u otro método.

Contraste (MET:MCE)	Estimación	SE	g.l.	Razón T	P-value
A:5 - B:5	-0.145	0.00776	336	-18.659	$<$.0001
A:20 - B:20	-0.740	0.00776	336	-95.285	$<$.0001
A:200 - B:200	-0.418	0.00776	336	-53.854	$<$.0001

Cuadro 4.8: Prueba HSD de Tukey para contrastar la interacción MET:MCE.

De acuerdo a la Figura (4.7), la metodología propuesta tiene un desempeño deseado cuando el número de *SNPs* con efecto no nulo es menor y disminuye conforme el número de *SNPs* con efecto aumenta. Este comportamiento está estrechamente relacionado con el tamaño del efecto asignado a cada uno de los *SNPs*. El hecho de que la TDV disminuya para valores grandes del factor MCE no es un efecto directo del número de *SNPs* con efecto, sino es el efecto directo del tamaño del efecto. La prueba estadística propuesta arroja buenos resultados cuando el tamaño del efecto es lo suficientemente grande, lo que puede interpretarse como una característica óptima de la prueba ya que un *SNP* con tamaño de efecto grande es muy probable que sea responsable de la manifestación del rasgo bajo estudio, por lo que se estaría en una situación de verdaderas asociaciones utilizando la metodología propuesta.

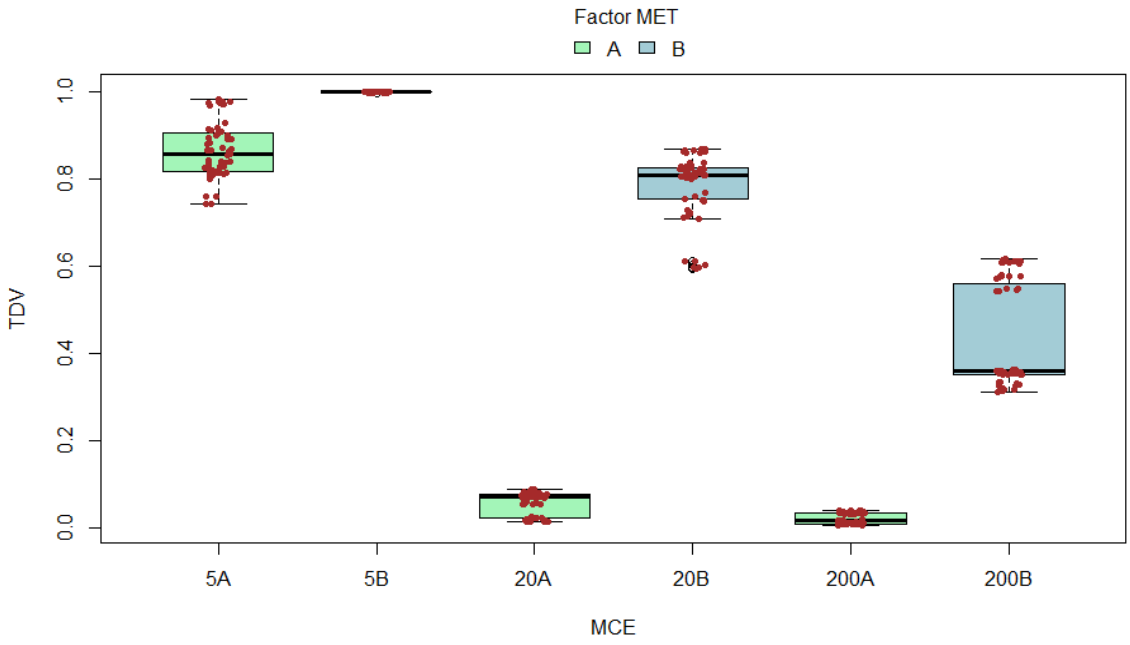

Figura 4.7: Comportamiento de la TDV entre niveles de la interacción MET:MCE.

Otro factor importante involucrado en cambios en los valores de la TDV es la interacción marcadores genéticos con efecto y número de genotipos (MCE:GEN). El resultado de las comparaciones múltiples se muestran en el Cuadro (4.9).

Contraste (MCE:GEN)	Estimación	SE	g.l.	Razón T	P-value
5:1000 - 5:3000	-0.0605	0.00863	336	-7.006	$<.0001$
20:1000 - 20:3000	-0.0515	0.00863	336	-5.969	$<.0001$
200:1000 - 200:3000	-0.1200	0.00863	336	-13.900	$<.0001$

Cuadro 4.9: Prueba HSD de Tukey para contrastar la interacción MCE:GEN.

Las pruebas de los contrastes en el Cuadro (4.9) exhiben diferencia significativa entre tamaños de muestra para cada uno de los niveles de MCE. Y la Figura (4.8) da cuenta de que el tamaño de muestra 3000 arroja mejores resultados de la TDV en todos los niveles del factor MCE con respecto al tamaño de muestra 1000. Los resultados son consistentes con lo esperado ya que para tamaños de muestra grandes se espera que la prueba tenga mejor desempeño.

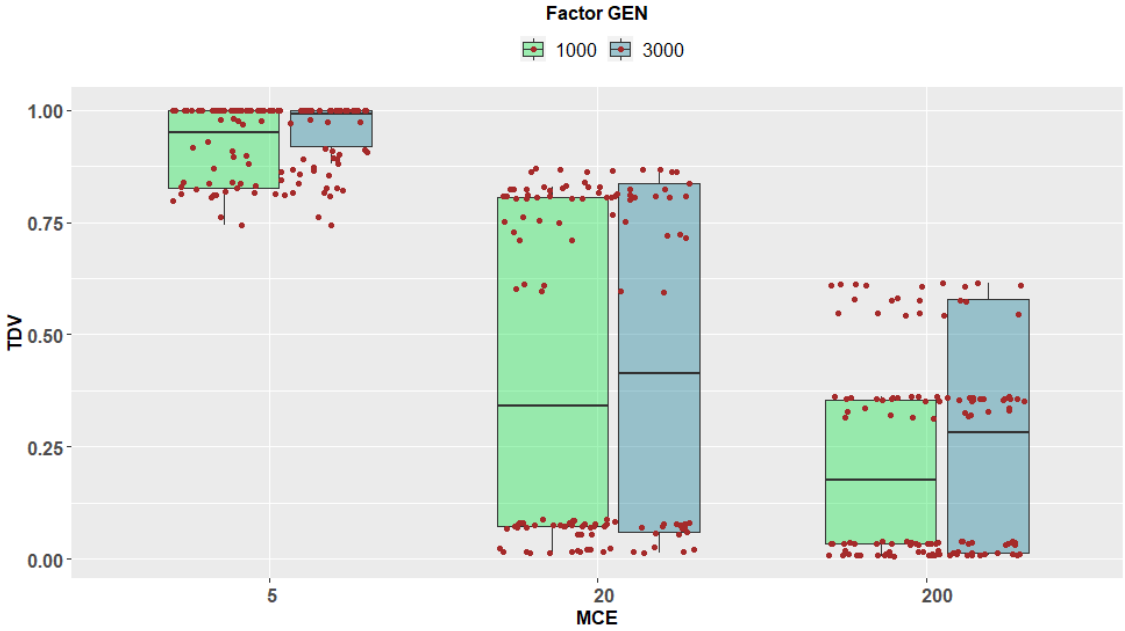

Figura 4.8: Comportamiento de la TDV entre niveles de la interacción MCE:GEN.

El siguiente factor a analizar es el tamaño de la muestra (GEN). En primer lugar se muestran (Cuadro 4.10) los resultados derivados de las comparaciones múltiples entre niveles del factor.

Contraste	Estimación	SE	g.l.	Razón T	P-value
1000 - 3000	-0.0774	0.00537	336	-14.413	$<.0001$

Cuadro 4.10: Prueba HSD de Tukey para contrastar el factor GEN.

Note que existe diferencia significativa entre los dos niveles del factor GEN considerando a la tasa de detecciones verdaderas TDV como respuesta. Independientemente del método de estudio de asociación utilizado, la TDV tiene un mejor desempeñado en escenarios donde el tamaño de la muestra es mayor (Figura 4.9).

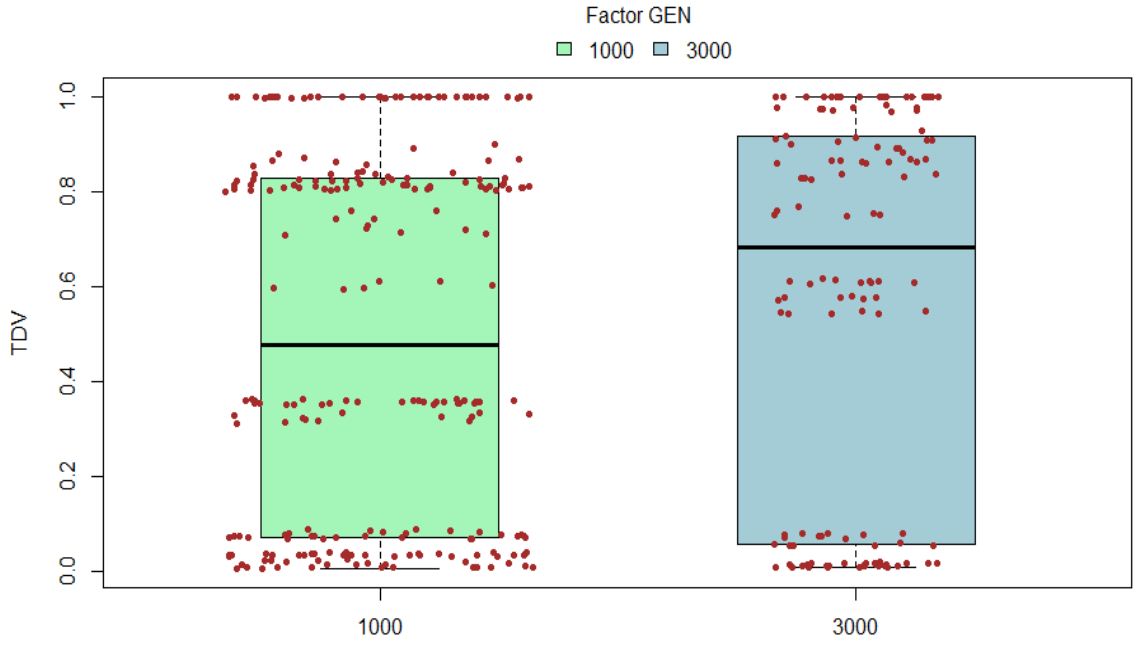

Figura 4.9: Comportamiento de la TDV entre niveles del factor GEN.

Por último, el análisis del factor interacción MET:GEN servirá para identificar el desempeño de ambos métodos tomando como referencia el tamaño de la muestra. Los resultados obtenidos (Cuadro 4.11) muestran que efectivamente existe diferencia significativa entre ambos métodos en cada uno de los niveles del tamaño de la muestra. En la Figura (4.10) se muestran las diferencias detectadas por la prueba estadística.

Contraste (MET:GEN)	Estimación	SE	g.l.	Razón T	P-value
A:1000 - B:1000	-0.399	0.00547	336	-73.013	$<$.0001
A:3000 - B:3000	-0.469	0.00745	336	-62.925	$<$.0001

Cuadro 4.11: Prueba HSD de Tukey para contrastar la interacción MET:GEN.

El mejor desempeño de la TDV se da para el método implementado en *PLINK* y con el tamaño de muestra mayor. Sin embargo, estos resultados deben tomarse con su debida reserva, ya que el hecho de que un marcador genético tenga asignado un tamaño de efecto diferente de cero no garantiza que estadísticamente esté asociado con el rasgo de interés.

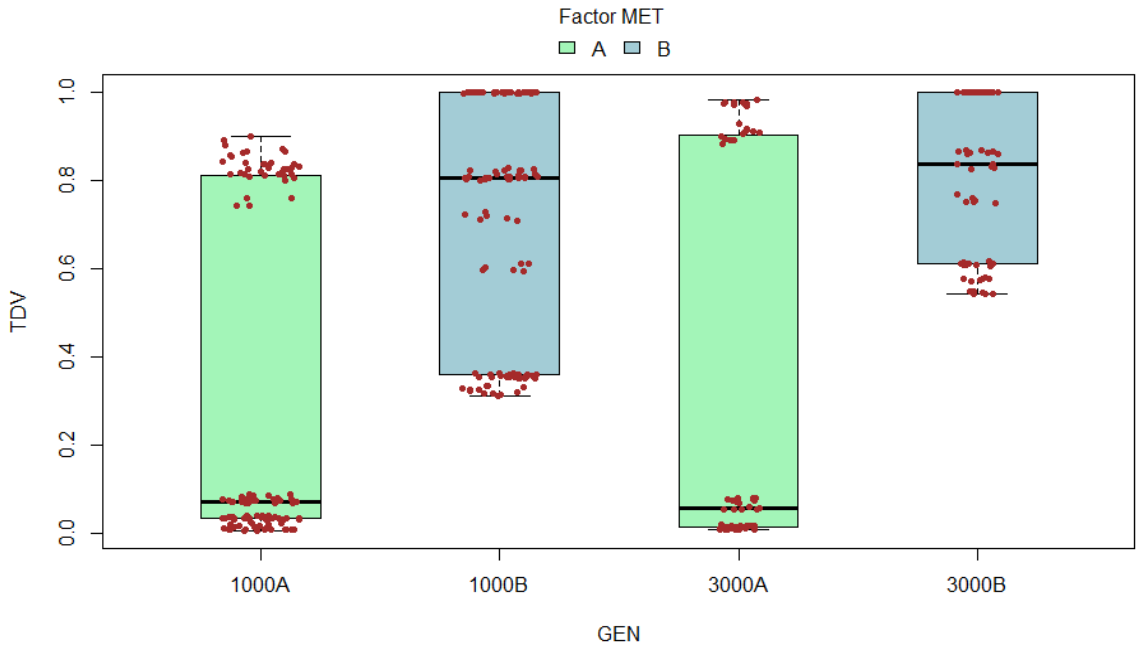

Figura 4.10: Comportamiento de la TDV entre niveles de la interacción MET:GEN.

Las comparaciones entre métodos tomando como indicador de análisis a la tasa de detecciones verdaderas arrojaron resultados interesantes. En términos generales, el desempeño de la TDV es mejor para el método implementado en *PLINK* con respecto a la metodología propuesta en está investigación. Sin embargo, los valores de TDV están sujetos al tamaño del efecto y un tamaño de efecto distinto de cero no significa que esté asociado al rasgo, por lo que la prueba estadística propuesta obtiene resultados deseados de la TDV cuando el tamaño del efecto es grande y en consecuencia, se obtiene un desempeño deseado de la propuesta. Con respecto al tamaño de la muestra, la metodología propuesta es consistente con lo que dice la teoría estadística ya que cuando el tamaño de la muestra es mayor, los resultados de TDV mejoran considerablemente aunado a un tamaño de efecto los suficientemente grande. En el caso de *SNPs* a probar, ambos métodos tienen comportamientos similares; es decir, cuando el número de marcadores es grande, el factor método conjuntamente con el factor SNP no afectan la TDV en términos estadísticos.

4.2. Indicador TDF

El otro indicador propuesto para evaluar la prueba estadística es la TDF. La TDF como ya se dijo es la tasa de falsas detecciones calculada como el promedio de las

mil simulaciones de la proporción de marcadores genéticos que son detectados por la prueba como estadísticamente significativos cuando en la realidad no lo son, ya que tienen tamaño de efecto cero. El comportamiento general de la TDF, a lo largo de los 180 escenarios, entre el método propuesto en está investigación y el método implementado en *PLINK* se muestra en la Figura (4.11).

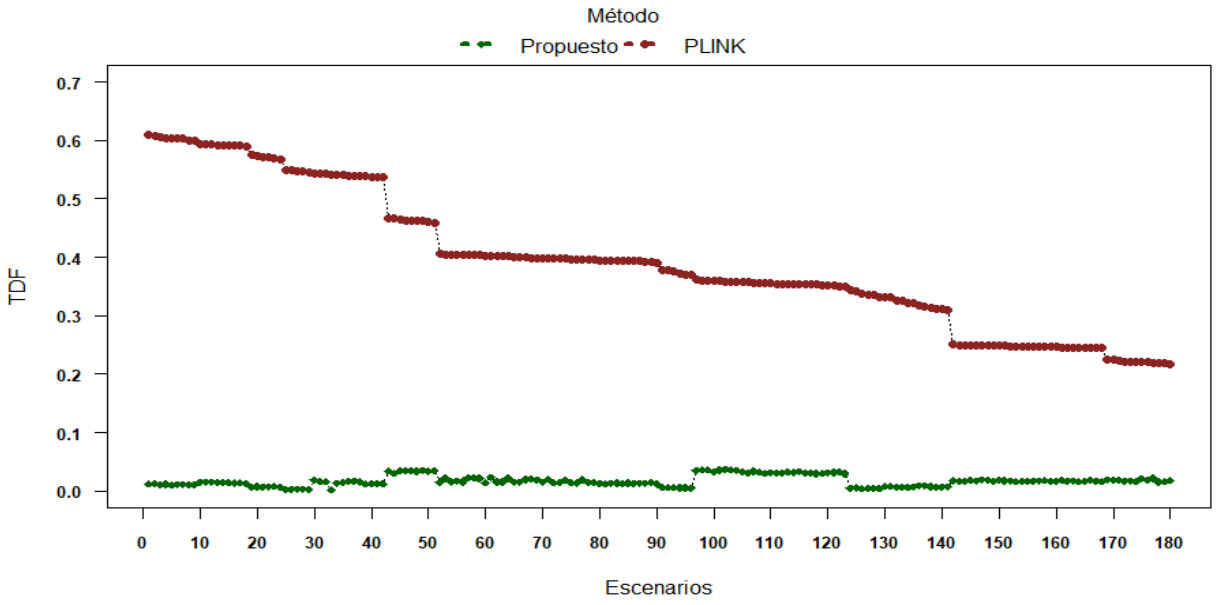

Figura 4.11: Comportamiento de la TDF entre métodos de estudios de asociación.

Note que, Figura (4.11), se tiene diferencia entre ambos métodos comparándolos por tasa de falsas detecciones. Para describir los factores que directamente afectan a la TDF y son diferentes entre métodos se realiza un análisis de varianza para determinar significancia estadística. Los resultados del *ANOVA* se exponen en el cuadro (4.12).

Los resultados del *ANOVA* para TDF son similares a los obtenidos para TDV, aunque en este último análisis se obtiene significancia estadística para el factor varianza genética (VG) y para la interacción método y número de *SNPs* a probar (MET:SNP). Y de igual forma como en la sección anterior, en este caso se analizaron cada uno de los factores para determinar en cuales niveles existen diferencias que afectan a la respuesta TDF.

Fuente	Grados de libertad	Sumas de cuadrados	Estadística F	*P-value*
MET	1	12.6757	1.2809e+05	<2.2e-16
SNP	2	0.0380	1.9182e+02	<2.2e-16
MCE	2	0.3212	1.6229e+03	<2.2e-16
GEN	1	0.6954	7.0275e+03	<2.2e-16
VG	2	0.0007	3.6077	0.028176
MET:SNP	2	0.0213	1.0763e+02	< 2.2e-16
MET:MCE	2	0.3864	1.9521e+03	< 2.2e-16
MET:GEN	1	0.8424	8.5122e+03	< 2.2e-16
SNP:MCE	4	0.0015	3.8955e	0.004164
SNP:GEN	2	0.0024	1.1902e+01	1.016e-05
MCE:GEN	2	0.0055	2.7884e+01	6.322e-12
Error	334	0.0331		

Cuadro 4.12: ANOVA para TDF. MET = Método, SNP = *SNPs* a probar, MCE = *SNPs* con efecto, GEN = Genotipos, VG = Varianza genética y el resto son las interacciones.

El primer factor a analizar es el método. En el Cuadro (4.13) se encuentra el *p-value* obtenido de la prueba estadística de comparar el método A y el método B, y dado que es menor al valor de significancia nominal $\alpha = 0.05$ se tiene que ambos métodos son estadísticamente diferentes con respecto a la variable TDF.

Contraste	Estimación	SE	g.l.	Razón T	P-value
A - B	-0.398	0.00121	334	-329.336	<.0001

Cuadro 4.13: Prueba HSD de Tukey aplicada a TDF para contrastar el método A y B.

En la Figura (4.12) se confirma el resultado de la prueba estadística, la metodología propuesta controla de forma óptima la tasa de detecciones falsas ya que en todos los escenarios la TDF se mantiene por debajo de un umbral mínimo deseado. Por otro lado, el método implementado en *PLINK* obtiene valores de la TDF en el intervalo $(0.2, 0.6)$, lo que se considera como altas tasas de falsos positivos que representan sesgos en los resultados.

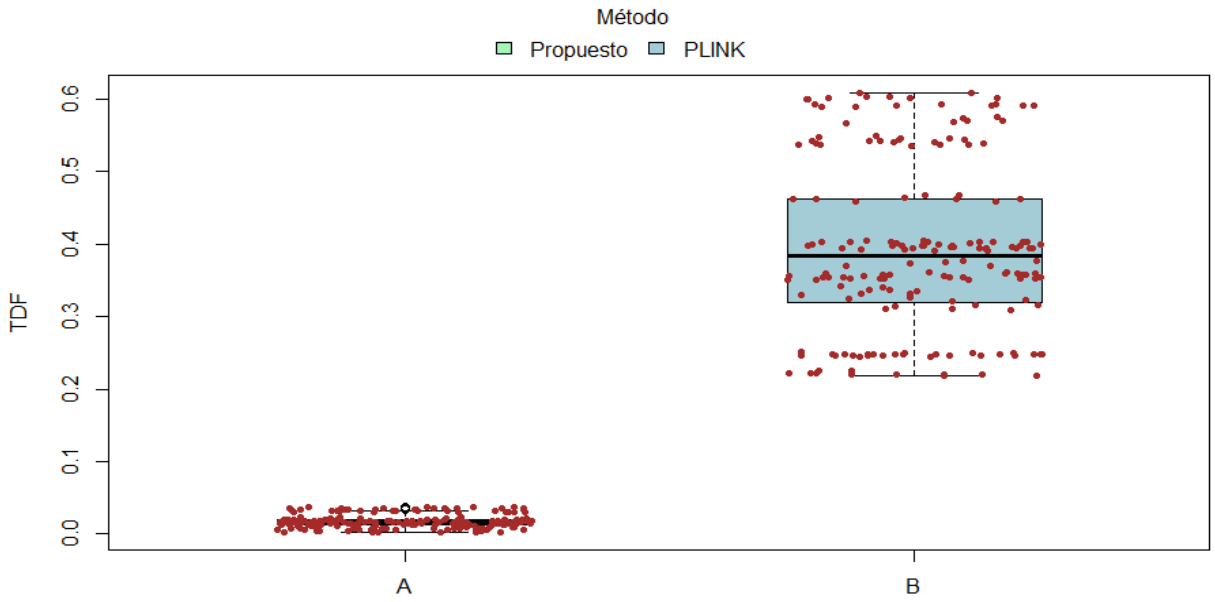

Figura 4.12: Comportamiento de la TDF entre métodos A y B.

El siguiente factor a analizar es el número de *SNPs* a probar. Del Cuadro (4.14) se encontraron diferencias significativas entre los niveles 1000 y 5000, y 1000 y 15000. Estos resultados coinciden con los obtenidos para la TDV por lo que se concluye que cuando se tienen varios miles de *SNPs* a probar, el factor ya no tiene efecto sobre el desempeño de las metodologías aquí analizadas.

Contraste	Estimación	SE	g.l.	Razón T	P-value
1000 - 5000	0.02536	0.00159	334	15.975	$<$.0001
1000 - 15000	0.02755	0.00159	334	17.356	$<$.0001
5000 - 15000	0.00219	0.00166	334	1.322	0.3840

Cuadro 4.14: Prueba HSD de Tukey aplicada a TDF para contrastar el factor SNP.

La Figura (4.13) muestra mayor dispersión para el nivel 1000 con respecto a los dos niveles más. También se observa que la media empírica del nivel 1000 es ligeramente mayor al resto de niveles, por lo que la prueba estadística los establece como diferentes. Por otro lado, el factor 5000 y 15000 tienen la misma dispersión y medias de los datos por lo que se asumen iguales en términos estadísticos.

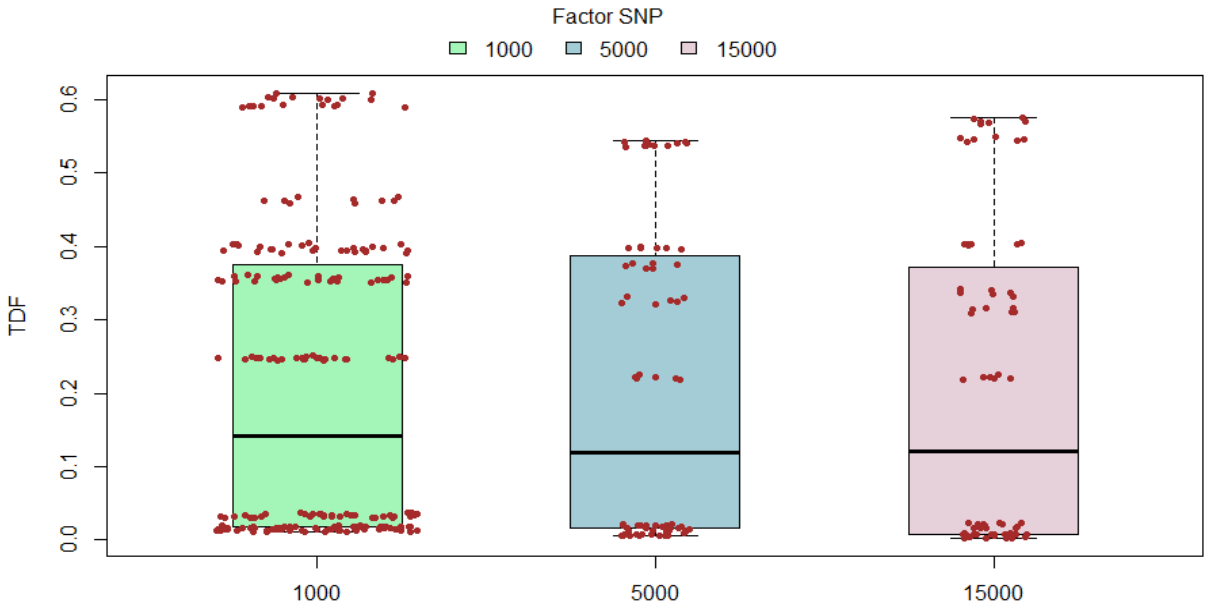

Figura 4.13: Comportamiento de la TDF entre niveles del factor SNP.

El siguiente análisis es para determinar el efecto del factor número de *SNPs* a probar entre métodos, por lo que se compararán las medias de la TDF entre métodos por niveles del factor SNP. Los resultados se presentan en el Cuadro (4.15).

Contraste (MET:SNP)	Estimación	SE	g.l.	Razón T	P-value
A:1000 - B:1000	-0.420	0.00147	334	-285.870	$<$.0001
A:5000 - B:5000	-0.386	0.00234	334	-164.686	$<$.0001
A:15000 - B:15000	-0.388	0.00234	334	-165.667	$<$.0001

Cuadro 4.15: Prueba HSD de Tukey aplicada a TDF para contrastar la interacción MET:SNP.

En la comparación el método propuesto en está investigación es diferente al método implementado en *PLINK* en todos los niveles del factor SNP. La Figura (4.14) muestra las diferencias encontradas, exhibe de manera clara y contundente que la metodología propuesta en está investigación controla de manera óptima las asociaciones espurias al mantener los valores de TDF siempre por debajo a 0.1 con respecto al método implementado en *PLINK* que ofrece un panorama bastante no deseado.

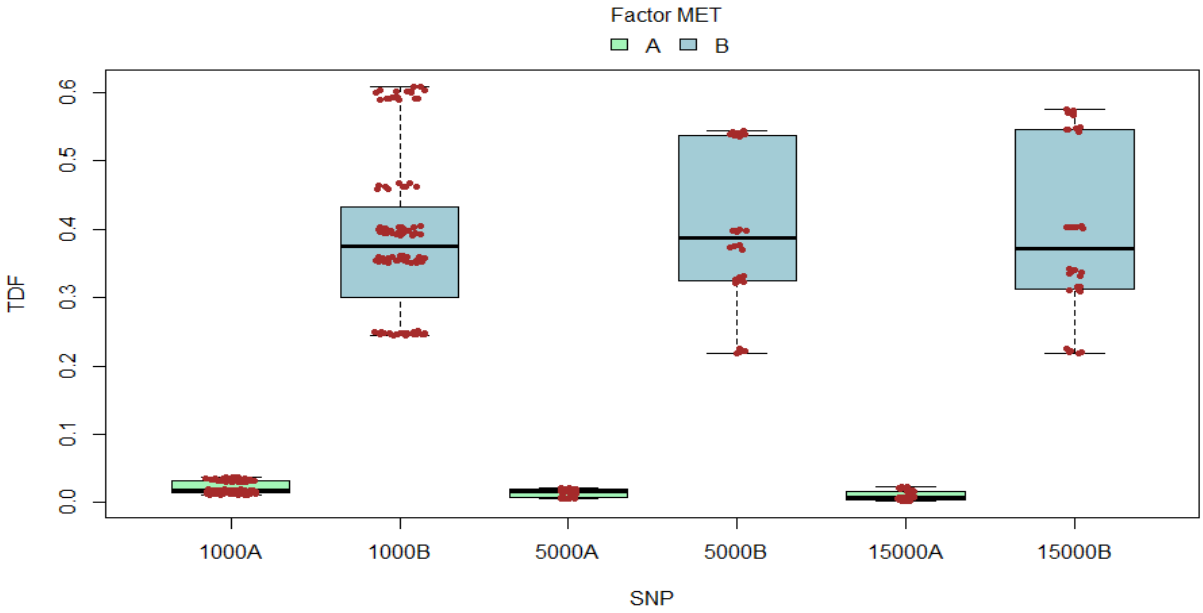

Figura 4.14: Comportamiento de la TDF entre niveles de la interacción MET:SNP.

La otra interacción que incluye el factor SNP es la SNP:MCE que se relaciona; también, con el número de marcadores con efecto no nulo. De igual forma que en casos anteriores se ejecutaron comparaciones múltiples para detectar aquellos niveles de la interacción relevantes. Los resultados obtenidos de dichas comparaciones se muestran en el Cuadro (4.16).

Contraste (SNP:MCE)	Estimación	SE	g.l.	Razón T	P-value
1000:5 - 5000:5	0.02393	0.00252	334	9.497	$<$.0001
1000:5 - 15000:5	0.02251	0.00252	334	8.937	$<$.0001
5000:5 - 15000:5	-0.00141	0.00287	334	-0.492	0.9998
1000:20 - 5000:20	0.02226	0.00252	334	8.836	$<$.0001
1000:20 - 15000:20	0.03220	0.00252	334	12.780	$<$.0001
5000:20 - 15000:20	0.00994	0.00287	334	3.460	0.0055
1000:200 - 5000:200	0.02989	0.00252	334	11.864	$<$.0001
1000:200 - 15000:200	0.02794	0.00252	334	11.090	$<$.0001
5000:200 - 15000:200	-0.00195	0.00287	334	-0.679	0.9980

Cuadro 4.16: Prueba HSD de Tukey aplicada a TDF para contrastar la interacción SNP:MCE.

Note que los resultados de las comparaciones múltiples confirman lo previamente dicho, la tasa de falsas detecciones no es afectada por el número de marcadores a probar cuando este valor es grande, ya que no se encontró diferencia significativa entre los niveles 5000 y 15000 de *SNPs* a probar, aún, cuando fueron probados agrupados

por número de marcadores con efecto no nulo. Un comportamiento interesante y generalizado de la TDF, derivado de la Figura (4.15), es que en cada uno de los niveles de MCE su media tiende a disminuir conforme aumenta el número *SNPs* a probar.

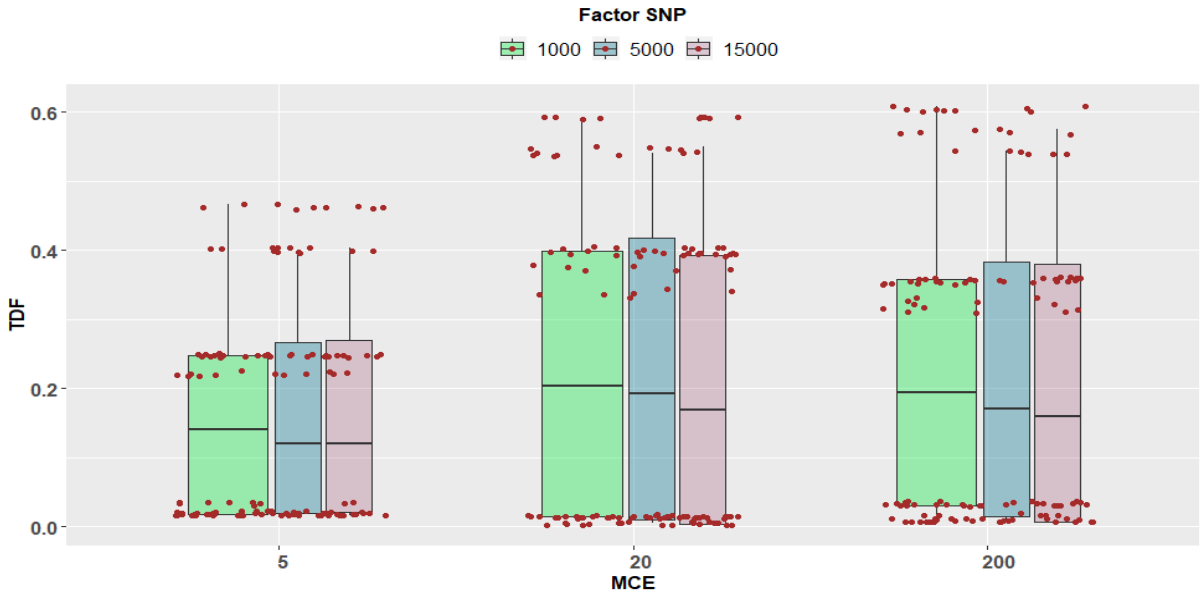

Figura 4.15: Comportamiento de la TDF entre niveles de la interacción SNP:MCE.

Los resultados de comparar los niveles de la interacción SNP:GEN se muestran en el Cuadro (4.17).

Contraste (SNP:GEN)	Estimación	SE	g.l.	Razón T	P-value
1000:1000 - 5000:1000	0.01852	0.00219	334	8.442	$<$.0001
1000:1000 - 15000:1000	0.02736	0.00219	334	12.473	$<$.0001
5000:1000 - 15000:1000	0.00884	0.00234	334	3.771	0.0012
1000:3000 - 5000:3000	0.03220	0.00219	334	14.682	$<$.0001
1000:3000 - 15000:3000	0.02774	0.00219	334	12.649	$<$.0001
5000:3000 - 15000:3000	-0.00446	0.00234	334	-1.901	0.3018

Cuadro 4.17: Prueba HSD de Tukey aplicada a TDF para contrastar la interacción SNP:GEN.

Cuando se compararon los niveles del factor SNP dentro de cada nivel del factor GEN se encontraron diferencias significativas entre los tres niveles del factor *número de SNPs a probar* en el nivel 1000 del número de genotipos. En el caso del nivel 3000 se encontraron diferencias significativas entre el nivel 1000 y 5000, y 1000 y 15000; sin embargo, cuando se compararon el nivel 5000 y 15000 fijando el tamaño de muestra en 3000 no se encontraron diferencias significativas de la TDF. Está situación es de

relevancia fundamental para caracterizar la prueba estadística propuesta ya que su desempeño con respecto al indicador TDF es deseado debido a que se mantiene al incrementar el número de marcadores genéticos a probar y mantener fijo un tamaño de muestra lo suficientemente grande.

En la Figura (4.16) no se perciben diferencias marcadas entre los tres niveles del factor agrupado de acuerdo al número de genotipos; sin embargo, se nota una diferencia muy marcada entre valores de la TDF cuando se comparan ambos tamaños de muestra.

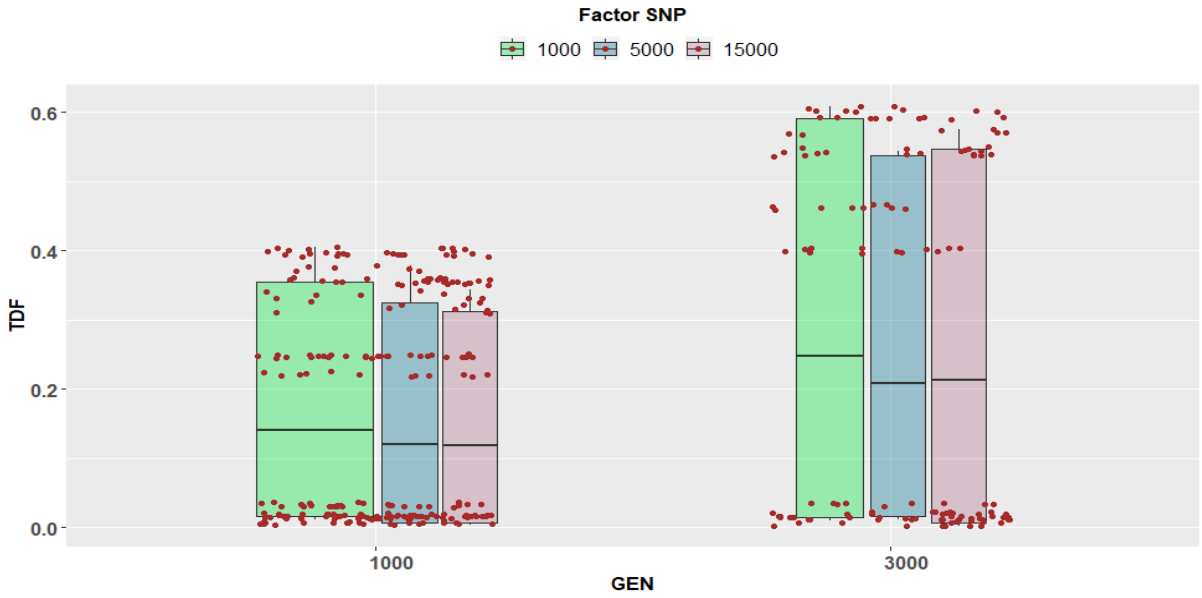

Figura 4.16: Comportamiento de la TDF entre niveles de la interacción SNP:GEN.

Otro factor que resultó ser estadísticamente significativo es el número de *SNPs* con tamaño de efecto diferente de cero (MCE).

Contraste	Estimación	SE	g.l.	Razón T	P-value
5 - 20	-0.06390	0.00148	334	-43.163	$<$.0001
5 - 200	-0.06043	0.00148	334	-40.814	$<$.0001
20 - 200	0.00348	0.00148	334	2.349	0.0507

Cuadro 4.18: Prueba HSD de Tukey aplicada a TDF para contrastar el factor MCE.

En el Cuadro (4.18) se muestran los resultados derivados de las comparaciones múltiples entre niveles del factor MCE. Existe diferencia significativa entre el nivel 5 y 20, y entre el nivel 5 y 200. En la Figura (4.17) se notan dichas diferencias, además de que se observa mayor dispersión de los datos en los niveles 20 y 200, lo que se puede

explicar por la presencia de muchos *SNPs* de efectos pequeños.

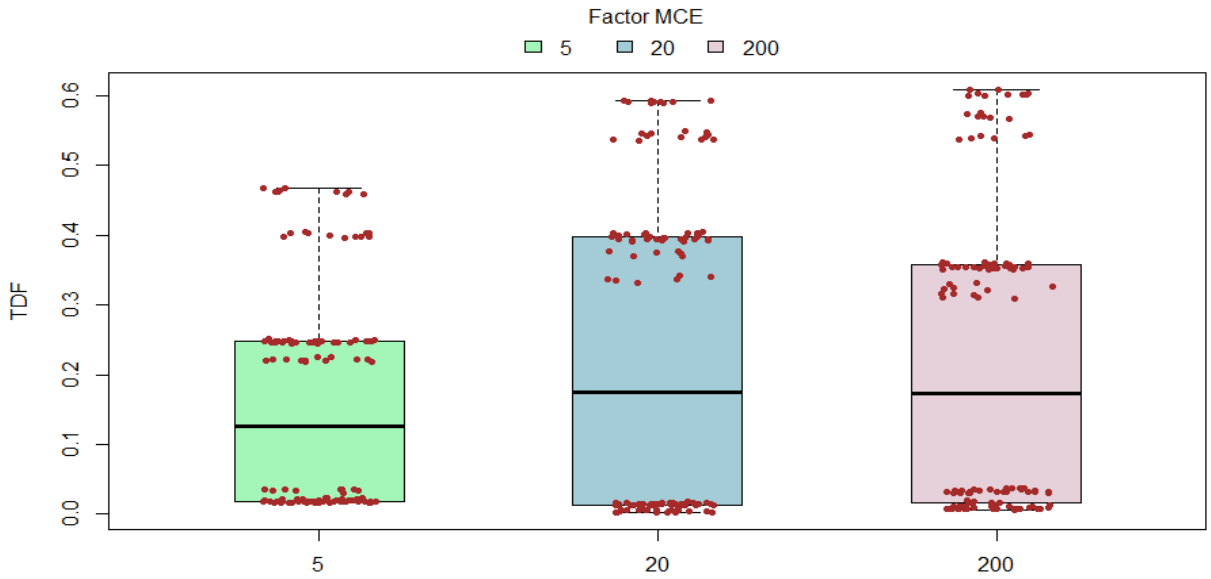

Figura 4.17: Comportamiento de la TDF entre niveles del factor MCE.

Otra interacción fundamental que ayuda a comparar el desempeño de los métodos de asociación involucrados en el análisis es MET:MCE con respecto a la tasa de detecciones falsas y que significa el efecto conjunto del método y el número de marcadores genéticos con efecto no nulo. Para comenzar con el análisis el primer paso es ejecutar comparaciones múltiples para determinar entre cuáles niveles de la interacción se encuentran las diferencias, sobre todo, aquellos contrastes que comparan ambos métodos con respecto a un nivel dado del número de marcadores genéticos con efecto no nulo.

Contraste (MET:MCE)	Estimación	SE	g.l.	Razón T	P-value
A:5 - B:5	-0.307	0.00191	334	-160.542	$<.0001$
A:20 - B:20	-0.459	0.00191	334	-239.848	$<.0001$
A:200 - B:200	-0.428	0.00191	334	-223.867	$<.0001$

Cuadro 4.19: Prueba HSD de Tukey aplicada a TDF para contrastar la interacción MET:MCE.

En el Cuadro (4.19) se nota que el método A (método propuesto) y el método B (método implementado en el software *PLINK*) son diferentes en términos estadísticos en cada uno de los niveles del factor MCE para TDF; es decir, el efecto del método cambia de manera significativa el comportamiento de la tasa de detecciones falsas. La

Figura (4.18) ayuda a comprender dichas diferencias y permite identificar qué método tiene mejor control de la tasa y, no cabe duda, que el mejor método para dicho fin es el propuesto es está investigación.

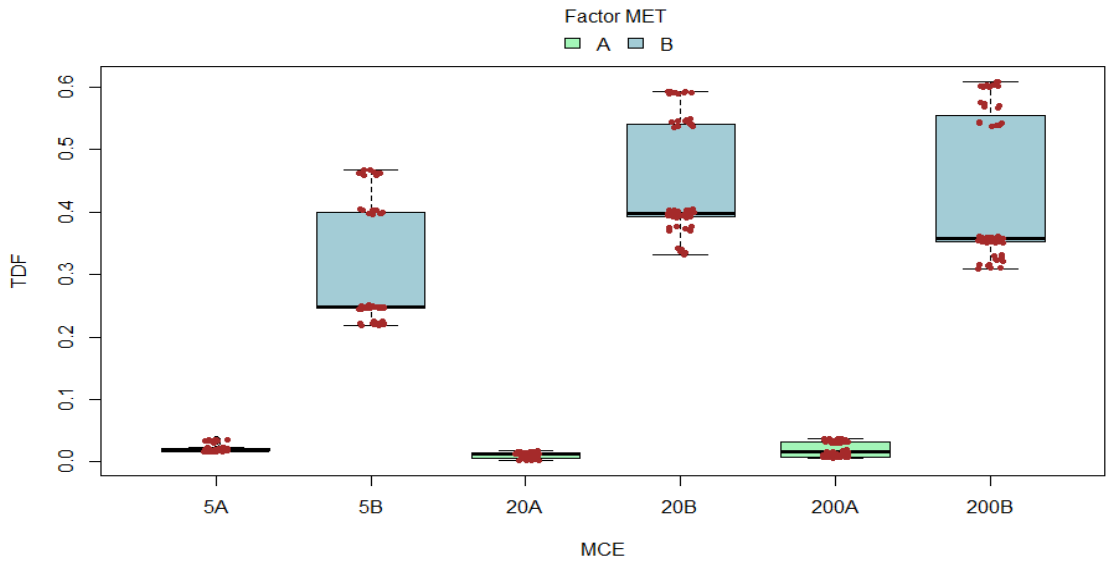

Figura 4.18: Comportamiento de la TDF entre niveles de la interacción MET:MCE.

Otra interacción que resultó ser significativa es la MCE:GEN y se refiere al efecto combinado del número de marcadores genéticos con efecto y el número de genotipos. En este caso se compararon los dos tamaños de muestra para un nivel dado de MCE y los resultados se muestran en el Cuadro (4.20).

Contraste (MCE:GEN)	Estimación	SE	g.l.	Razón T	P-value
5:1000 - 5:3000	-0.1030	0.00205	334	-50.159	$<.0001$
20:1000 - 20:3000	-0.0948	0.00205	334	-46.152	$<.0001$
200:1000 - 200:3000	-0.1154	0.00205	334	-56.215	$<.0001$

Cuadro 4.20: Prueba HSD de Tukey aplicada a TDF para contrastar la interacción MCE:GEN.

Note que en todos los niveles de MCE se encontraron diferencias significativas entre los dos tamaños de muestra analizados. Los resultados confirman lo previamente dicho, cuando el tamaño de la muestra es grande la dispersión de los datos para la TDF es grande comparado a un tamaño de muestra mucho menor y aumenta en mayor medida cuando se tiene una cantidad considerable de marcadores genéticos con efectos pequeños (ver Figura 4.19). También, debe tenerse presente que los valores de TDF mayores a 0.2 son originados por el método implementado en *PLINK*.

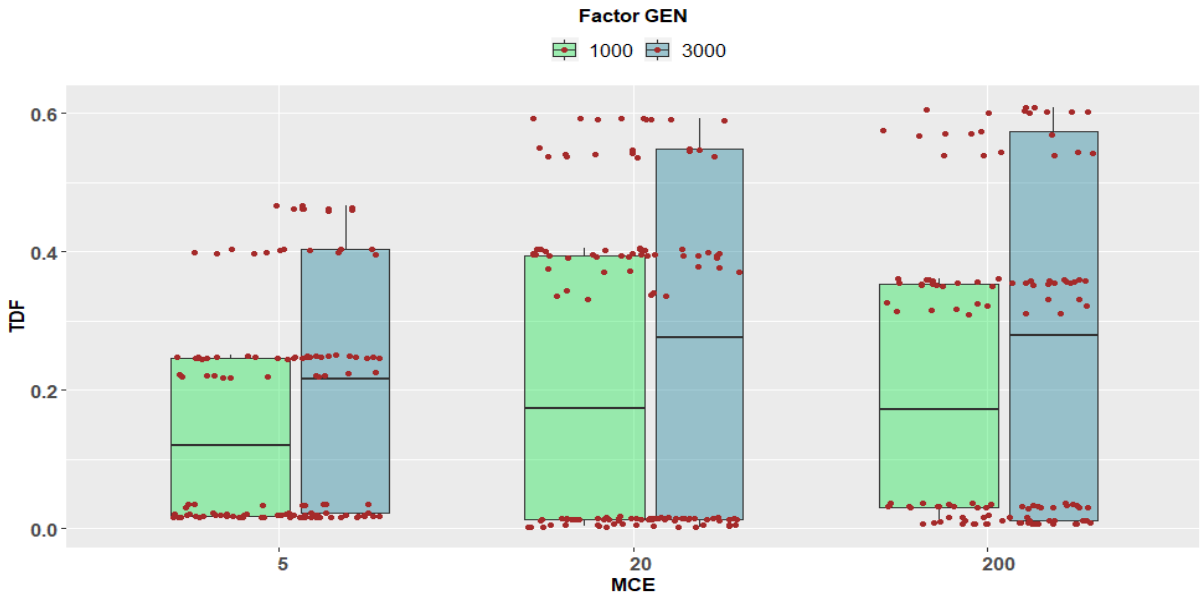

Figura 4.19: Comportamiento de la TDF entre niveles de la interacción MCE:GEN.

Otro factor que tiene efecto sobre los valores de TDF es el tamaño de la muestra (GEN). Se encontró diferencia significativa entre tamaños de muestra y según la Figura (4.20) la media mayor de la TDF corresponde al nivel 3000 de número de genotipos.

Contraste	Estimación	SE	g.l.	Razón T	P-value
1000 - 3000	-0.104	0.00128	334	-81.794	$<.0001$

Cuadro 4.21: Prueba HSD de Tukey aplicada a TDF para contrastar el factor GEN.

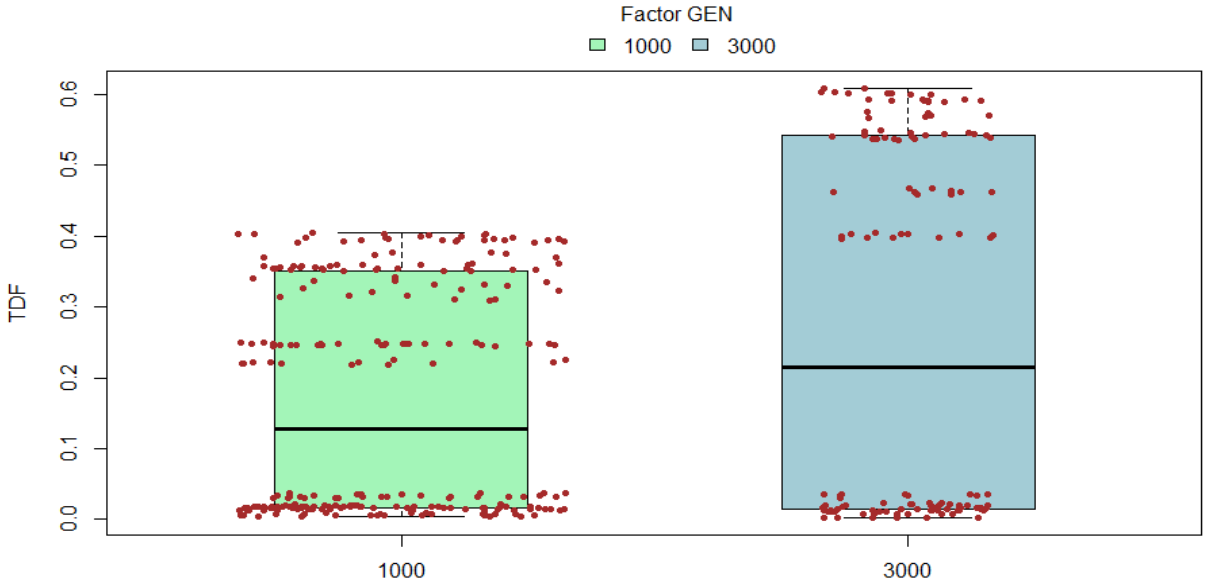

Figura 4.20: Comportamiento de la TDF entre niveles del factor GEN.

El análisis que arroja más información, dados los objetivos de la investigación, es comparar ambos métodos de estudio de asociación fijando un tamaño de muestra. Éste análisis corresponde a la interacción método-número de genotipos (MET:GEN). Los resultados obtenidos se muestran en el Cuadro (4.22) y Figura (4.21).

Contraste (MET:GEN)	Estimación	SE	g.l.	Razón T	P-value
A:1000 - B:1000	-0.293	0.00152	334	-192.329	$<.0001$
A:3000 - B:3000	-0.503	0.00178	334	-281.826	$<.0001$

Cuadro 4.22: Prueba HSD de Tukey aplicada a TDF para contrastar la interacción MET:GEN.

Las comparaciones múltiples entre métodos arrojaron resultados alentadores para la prueba estadística desarrollada en está investigación en términos de control óptimo de la tasa de asociaciones espurias. La Figura (4.21) muestra el comportamiento de la TDF entre métodos y entre tamaños de muestra. Los resultados obtenidos, considerando el tamaño de muestra como un factor de efecto sobre el desempeño del método de asociación, se inclinan a favor de la propuesta.

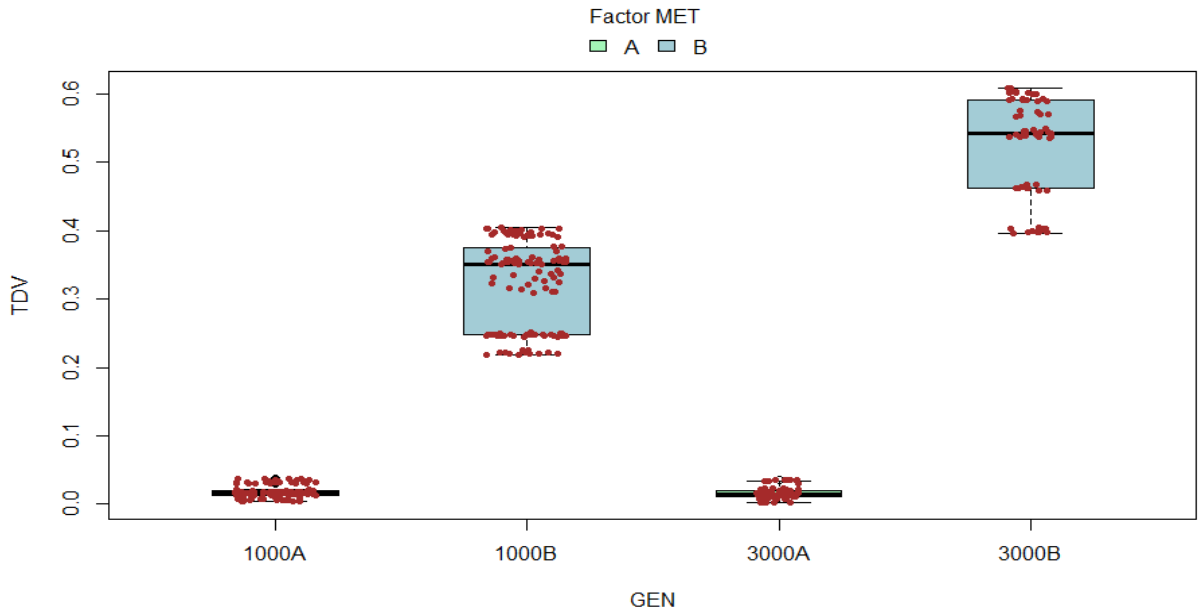

Figura 4.21: Comportamiento de la TDF entre niveles de la interación MET:GEN.

Para finalizar la sección de análisis para TDF se analiza el último factor significativo, VG, que se refiere a la magnitud de varianza genética supuesta en los escenarios simulados.

Los resultados estrictamente apuntan a que existe diferencia significativa entre los ni-

veles 0.3 y 3.0 de la magnitud de varianza genética; sin embargo, este resultado debe tomarse con precaución ya que el análisis gráfico (Figura 4.22) exhibe un comportamiento de la tasa de detecciones falsas igual entre los tres niveles del factor VG. En términos prácticos, no hay diferencia entre valores de la TDF originados por cambios en el factor VG.

Contraste	Estimación	SE	g.l.	Razón T	P-value
0.3 - 1.0	0.00214	0.00128	334	1.669	0.2187
0.3 - 3.0	0.00341	0.00128	334	2.657	0.0224
1.0 - 3.0	0.00127	0.00128	334	0.988	0.5848

Cuadro 4.23: Prueba HSD de Tukey aplicada a TDF para contrastar el factor VG.

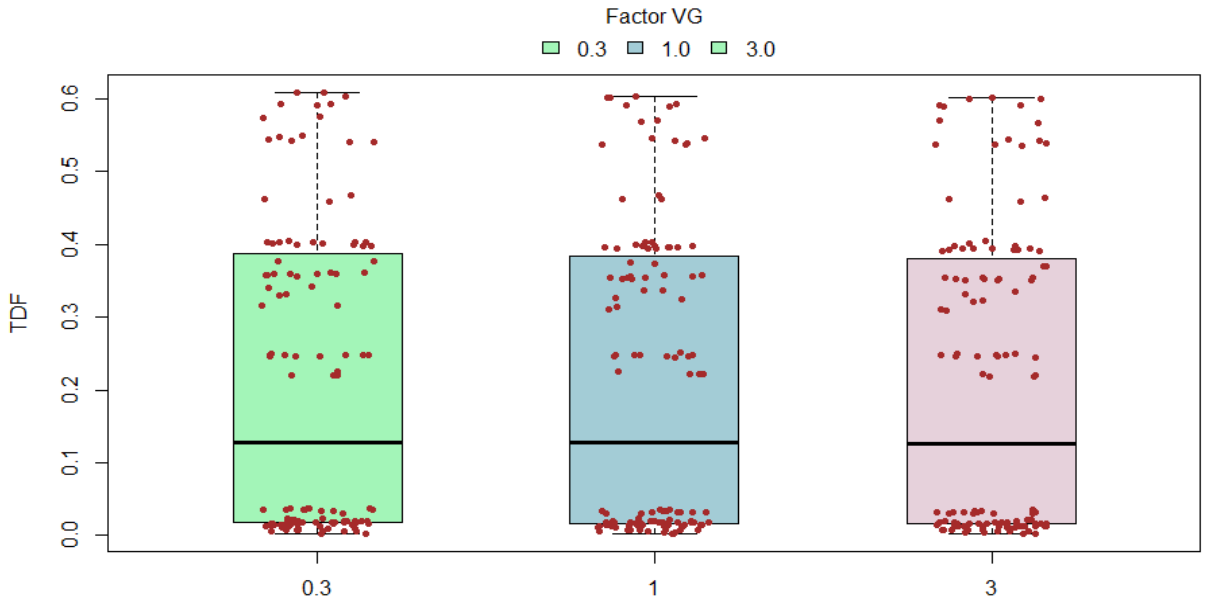

Figura 4.22: Comportamiento de la TDF entre niveles del factor VG.

En términos generales la TDF representa una estimación de la tasa de error de una familia de hipótesis y que en conjunto dicha estimación se encuentra siempre por debajo de un nivel de significancia nominal comúnmente fijado en 0.05. Mientras, para el método implementado en *PLINK* ocurre que no hay control de este error ya que en ninguno de los escenarios simulados se obtuvo un valor de TDF abajo o cercano de 0.05 y más aún, algunos valores sobrepasaron el nivel de 0.6.

Las comparaciones entre ambos métodos arrojaron resultados interesantes. En primer lugar, los resultados obtenidos de TDV fueron a favor del método implementado en

PLINK, ya que para todos los escenarios están por arriba de los resultados obtenidos a partir de la metodología propuesta. Sin embargo, es de fundamental importancia dejar claro que los tamaños de efecto asignados a los *SNPs* seleccionados en la simulación no garantizan que sean estadísticamente significativos. En segundo lugar, el análisis correspondiente a TDF muestra una clara ventaja de la metodología propuesta sobre el método implementado en *PLINK*.

Para reforzar el hecho de que la metodología propuesta es sensible al tamaño del efecto de los *SNPs* a probar se simularon 24 escenarios más fijando en 200 el número de *SNPs* con efecto y asignando pesos a_i distintos. Los resultados obtenidos se analizaron en la siguiente sección.

NOTA: Los resultados completos para TDV y TDF se encuentran en el ANEXO A.

4.3. TDV y TDF usando diferentes pesos

La idea central de las 24 simulaciones más es comparar escenarios donde el tamaño del efecto es el mismo para todos los *SNPs* con efecto no nulo y escenarios donde los *SNPs* con efecto no nulo tienen tamaños diferentes.

Para los 24 escenarios simulados solamente se aplicó la prueba estadística propuesta. Los resultados obtenidos para TDV se muestran en el Cuadro (4.24).

Escenario	*SNPs*	Genotipos	TDV				
			1 snp	2 snp	3 snp	4 snp	5 snp
1	1000	1000	1.0000	1.0000	0.9493	0.7508	0.8206
2	1000	3000	1.0000	0.9995	0.9820	0.8438	0.9714
3	5000	1000	1.0000	0.9955	0.9363	0.3835	0.8574
4	5000	3000	1.0000	0.9985	0.8807	0.8365	0.9138
5	15000	1000	1.0000	0.9785	0.6210	0.3585	0.7604
6	15000	3000	1.0000	0.9825	0.9080	0.6333	0.8910

Cuadro 4.24: TDV para los 24 escenarios descritos.

Note que los resultados obtenidos tienen la misma tendencia entre los seis escenarios agrupados por número y pesos distintos de los *SNPs*. Los valores menores, 0.3585 y 0.3835, se obtienen en combinaciones donde el número de *SNPs* a probar y el número de genotipos es marcadamente diferente. El escenario que peor comportamiento tiene de la TDV es el escenario 5 donde se tiene un menor número de genotipos y mayor número de marcadores genéticos a probar, lo que podría explicar dicho comportamiento. En la Figura (4.23) puede verse claramente lo descrito. Los seis escenarios comparados entre sí coinciden en que los valores menores de la TDV se obtiene cuando se seleccionan cuatro *SNPs* con un peso de 0.2 cada uno, dejando al resto, 196 *SNPs*, con un peso de 0.2/196. La baja de TDV en estos escenarios podría explicarse por la presencia de muchos *SNPs* con efectos pequeños lo que está provocando mayor incertidumbre y una disminución de la potencia.

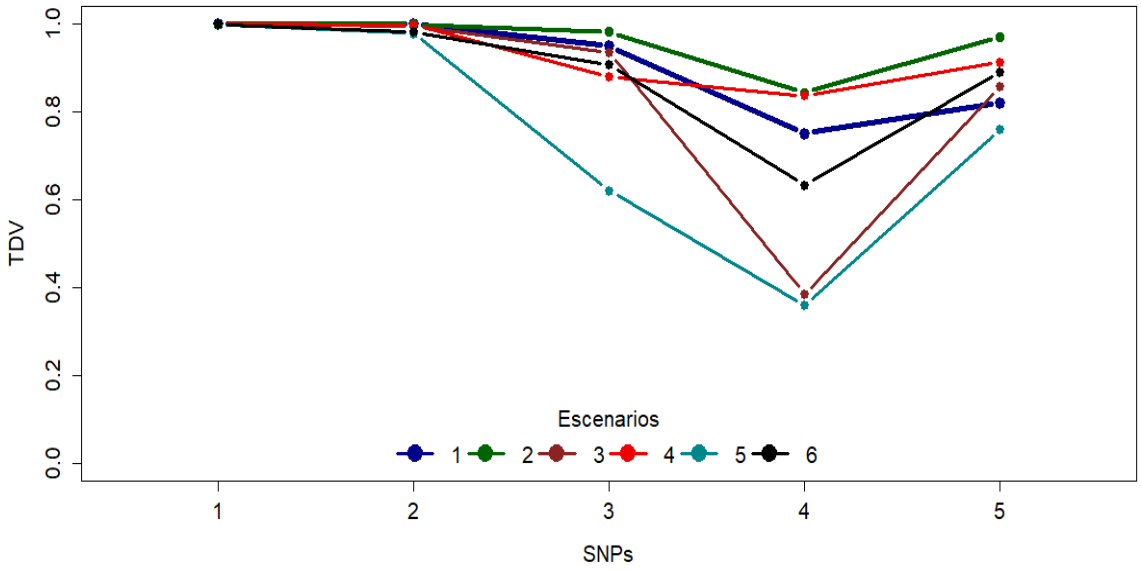

Figura 4.23: Comparación de la TDV para los 24 escenarios.

Por otro lado, el indicador TDF también arrojó algunos resultados singulares. En el Cuadro (4.25) se muestran los resultados obtenidos de la TDF para los 24 escenarios. Cuando se asume solamente un marcador genético con peso 0.2, los resultados son alentadores y cuando son dos los *SNPs* con peso 0.2, el valor de TDF incrementa aunque no en valores que podría preocupar. El mejor escenario presente es el escenario seis (ver Figura 4.24) ya que es donde se observan los valores más bajos de la TDF. Los elementos que determinan el escenario seis son 15000 *SNPs* a probar y 3000 genotipos, lo que representa un panorama alentador para la prueba estadística propuesta ya que estás situaciones podrían presentarse con mayor frecuencia en la vida real.

Escenario	*SNPs*	Genotipos	TDF				
			1 snp	2 snp	3 snp	4 snp	5 snp
1	1000	1000	0.0206	0.0434	0.0500	0.0402	0.0169
2	1000	3000	0.0210	0.0538	0.0481	0.0490	0.0346
3	5000	1000	0.0156	0.0387	0.0396	0.0193	0.0191
4	5000	3000	0.0204	0.0577	0.0280	0.0505	0.0196
5	15000	1000	0.0155	0.0467	0.0153	0.0131	0.0164
6	15000	3000	0.0166	0.0363	0.0301	0.0259	0.0218

Cuadro 4.25: TDF para los 24 escenarios descritos.

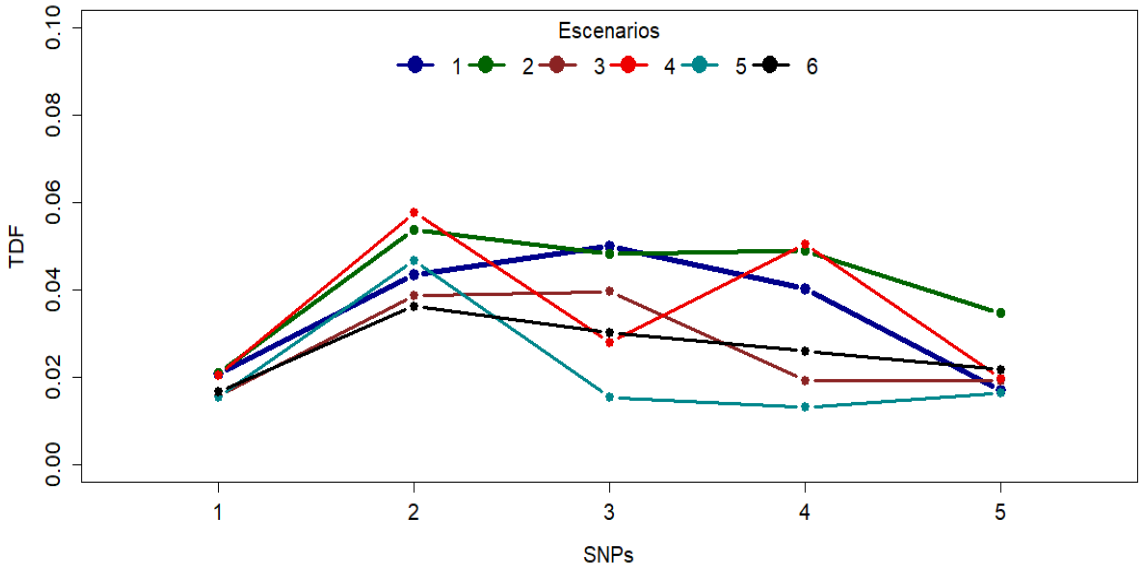

Figura 4.24: Comparación de la TDF para los 24 escenarios.

Los resultados obtenidos de los indicadores TDV y TDF son prometedores para la metodología propuesta en está investigación. Por un lado, su poder de detección está estrechamente relacionado con el tamaño del efecto lo que da tranquilidad ya que estaría discriminando de forma contundente entre *SNPs*, aquellos con un efecto verdadero sobre el rasgo y aquellos con efecto estadísticamente nulo. Por otro lado, aunque haya *SNPs* con diferentes tamaños de efecto no nulo, la prueba estadística mantiene de forma apropiada la proporción de asociaciones espurias.

Capítulo 5

CONCLUSIONES

En una prueba estadística lo que se busca son esencialmente dos cosas. En primer lugar que la prueba estadística tenga la potencia necesaria para identificar hipótesis verdaderas; es decir, que la probabilidad de rechazar la hipótesis nula cuando es falsa sea alta. En segundo lugar, que la prueba estadística tenga la capacidad necesaria para discriminar una hipótesis falsa; es decir, que la probabilidad de rechazar la hipótesis nula cuando es verdadera sea mínima.

En nuestro caso se desarrolló una prueba estadística para estudios de asociación en el contexto de investigación agrícola que contabiliza las relaciones intrínsecas y ambientales de los organismos en el análisis, por lo que en la construcción del modelo estadístico se incluyeron las relaciones genéticas de los individuos mediante una matriz construida a partir de las frecuencias alélicas de los marcadores genéticos y una matriz de relaciones originada por el ambiente, precisamente una estructura de autocorrelación espacial de grado 1. El algoritmo para análisis de asociación quedo totalmente desarrollado al identificar y estimar un MLM, proponer y calcular un estadístico de prueba como función del $BLUP$ y su matriz de covarianza. Estos dos elementos contienen el efecto de todos los factores de ruido incluidos en el modelo. Después, la distribución de probabilidad fue aproximada utilizando un distribución t-Student con los grados de libertad estimados vía Satterthwaite y para la regla decisión se utilizó la corrección de Bonferroni. Este algoritmo fue implementado en un código computacional escrito en el software de análisis estadístico. La descripción así como el código computacional se incluyen en el ANEXO B de este documento.

Por otro lado, la prueba estadística desarrollada en está investigación mostró, vía simulación, tres elementos que influyen significativamente en su capacidad de detección de asociaciones verdaderas, o en su defecto, en su capacidad de identificar marcadores genéticos con un tamaño de efecto distinto de cero. Los elementos son el número de *SNPs* a probar, el número de *SNPs* con efecto distinto de cero y el número de genotipos.

La metodología propuesta funciona de forma apropiada en escenarios donde el número de marcadores genéticos a probar es grande y conforme el número de genotipos incrementa. La combinación de ambos elementos en estudios de asociación en la vida real podría presentarse con bastante frecuencia, y más aún, puede haber muchas situaciones donde, para los dos elementos, su número es mucho mayor al estudiado aquí, generando así mayor potencia de la prueba estadística propuesta.

La conclusión con respecto al número de *SNPs* con efecto distinto de cero debe comprender los elementos que hay detrás de este elemento. De forma directa el número de marcadores genéticos con efecto relaciona a la varianza genética y el tamaño del efecto asignado a cada uno de los *SNPs*, así, en términos generales el tamaño del efecto depende tanto de la cantidad de *SNPs* con efecto como de la magnitud de la varianza genética. Por lo tanto, si se mantiene fija la varianza genética y varía el número de marcadores genéticos también varía el tamaño del efecto por la forma en cómo se calculan. Cuando el número de *SNPs* con tamaño de efecto distinto de cero incrementa, disminuye el tamaño del efecto del *SNP* y viceversa. Entonces, los escenarios donde se tiene una menor proporción de *SNPs* significativos, aunque tengan efecto diferente de cero son donde el número de marcadores genéticos es mayor, por lo tanto, el tamaño del efecto es menor y con ello, la prueba estadística propuesta no los determina como significativos. Por otro lado, los escenarios donde el número de *SNPs* con efecto es menor, la prueba estadística propuesta tiene altas tasas de *SNPs* con efecto identificados como significativos. En este punto es importante remarcar que el hecho de que un marcador genético se le asignará un tamaño de efecto diferente de cero durante el proceso de simulación no es garantía de que sea estadísticamente significativo. Por lo tanto, la prueba estadística propuesta funciona de manera apropiada para situaciones donde el rasgo de interés es explicado por marcadores genéticos con tamaños de efecto lo suficientemente grande, es decir, es ideal en contextos donde se requiere un especial cuidado con las asociaciones espurias y las asociaciones significativas sean reales.

Otra característica fundamental que debe cumplir la prueba estadística es minimizar, tanto como sea posible, la proporción de falsos positivos y en ese sentido la metodología propuesta cumple de manera tal, que no solo mantiene la proporción de asociaciones espurias por debajo de un umbral deseado sino que cuando incrementa el tamaño de la muestra la proporción disminuye. Por lo tanto, está característica convierte a la metodología propuesta en un método de estudio de asociación altamente atractivo para escenarios donde se busca minimizar la tasa de falsos positivos.

En términos generales, la metodología presentada aquí cumple con las características teóricas para ser una alternativa confiable en estudios de asociación. En su aplicación debe considerarse la naturaleza sobre la que está construida la prueba y los aspectos implícitos en el MLM. Una conclusión general sobre la metodología propuesta es que, aunque falta por ensayar otros escenarios, muestra un comportamiento bastante prometedor en estudios de asociación en plantas y que puede significar una alternativa con bastante aceptación dentro de los *GWAS* en general y, además, su aplicación es sencilla ya que todos los cálculos son realizados en el software de análisis estadístico R.

LITERATURA CITADA

Anderson, R. D. y Bancroft, T. A. (1952). *Statistical Theory in Research.*. McGraw-Hill, New York.

Aulchenko, Y. S., de Koning, D.-J. y Haley, C. (2007). Genomewide Rapid Association Using Mixed Model and Regression: A Fast and Simple Method For Genomewide Pedigree-Based Quantitative Trait Loci Association Analysis. *Genetics*, 177, 577–585.

Barreto, H., Edmeades, G., Chapman, S. y Crossa, J. (1996). The alpha lattice design in plant breeding and agronomy: Generation and analysis. *Developing Drought- and Low N-Tolerant Maize*, 544–551.

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*, 12:232.

Bush, W. S. y Moore, J. H. (2012). Chapter 11: Genome-Wide Association Studies. *PLOS Computational Biology*, 8, 11.

Crump, S. L. (1947). *The estimation of components of variance in multiple classifications*. Tesis Doctoral, Iowa State College.

da Silva, A. J. M. (2017). *Variance Components Estimation In Mixed Linear Models*. Proyecto Fin de Carrera, Faculdade de Ciencias e Tecnologia, Universidade Nova de Lisboa.

Devlin, B. y Roeder, K. (1999). Genomic Control for Association Studies. *BIOMETRICS*, 55, 997–1004.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. of the Cambridge Philosophical Soc*, 22, 700–725.

Harville, D. A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *J. A mer. Statist. Assoc.*, 72, 320–340.

Henderson, C. R. (1953). Estimation of Variance and Covariance Components. *Biometrics*, 9, 226–252.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Animal Science (Symposium)*, 10–41.

Ingvarsson, P. K. y Street, N. R. (2011). Association genetics of complex traits in plants. *New Phytologist*, 189, 909–922.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., yee Kong, S., Freimer, N. B., Sabatti, C. y Eskin, E. (2010). Variance component model to accoun for sample structure in genome-wide association studies. *Nature Genetics*, 42, 348–354.

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, 1709–1723.

Koeleman, B. P. C., Al-Ali, A., van der Laan, S. W. y Asselbergs, F. W. (2013). A Concise History of Genome-Wide Association Studies. *Saudi Journal of Medicine and Medical Sciencies*, 1, 4–10.

Korte, A. y Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review.

Lander, E. S. (1996). The New Genomics: Global Views of Biology. *Sciencie*, 274, 536–539.

Levinson, D. F. (2009). Genomewide association studies: History, rationale and prospects for psychiatric disorders. *American Journal of Psychiary*, 5, 540–556.

Li, G. y Zhu, H. (2013). Genetic Studies: The Linear Mixed Models in Genome-wide Association Studies. *The Open Bioinformatics*, 7, 27–33.

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, 833–835.

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., VilhjÃ¡lmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N. y Price, A. L. (2015). Efficient Bayesian mixed analysis increases association power in large cohorts. *Nat Genet.*, 47, 284–290.

Long, Y., Zhang, C. y Meng, J. (2008). Challenges for QTL Analysis in Crops. *Journal of Crop Science and Biotechnology*, 11, 7–12.

Miles, C. M. y Wayne, M. (2008). Quantitative Trait Locus (QTL) Analysis. *Nature Education*, 1, 208.

Muhammad, J., Aamir, A., Khalid, A., Abdul, A., Alvina, G. y Abdul, M.-K. (2016). QTL Analysis in Plants: Ancient and Modern Perspectives. 59–82.

Ormerod, J. T. y Wand, M. P. (2010). Explaining Variational Approximations. *The American Statistician*, 64, 140–153.

Östensson, M. (2012). *Statistical Methods for Genome Wide Association Studies*. Tesis Doctoral, Chalmers University of Technology and University of Gothenburg Göteborg.

Paterson, L. J. y Patterson, H. D. (1984). An algorithm for generating alpha-lattice designs. *ARS Combinatoria*, 16A, 87–98.

Patterson, H. D. y Thompson, R. (1971). Recovery of Inter-Block Information when Block Sizes are Unequal. *Biometrika*, 58, 545–554.

Patterson, H. D. y Williams, E. R. (1976). A New Class of Resolvable Incomplete Block Designs. *Biometrika*, 63, 83–92.

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, 904–909.

Pritchard, J. K., Stephens, M., Rosenberg, N. A. y Donnelly, P. (2000). Association Mapping in Structured Populations. *Am. J. Hum. Genet.*, 67, 170–181.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Julian Maller, P. S., de Bakker, P. I. W., Dal, M. J. y Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81, 559–575.

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *International Biometric Society*, 2, 110–114.

Searle, S. R. (1993). Applying the EM algorithm to calculating ML and REML estimates of variance components. *Technical Report BU-1213-M, Biometrics Unit*.

Searle, S. R. (1994). An overview of variance component estimation. *Technical Report BU-1231-M, Biometrics Unit*.

Searle, S. R. (1995). The matrix handling of BLUE and BLUP in the mixed linear model. *Technical Report BU- 1275-MA, Biometrics Unit*.

Searle, S. R., Casella, G. y McMculloch, C. E. (2006). *Variance Components*. Jonh Wiley and Sons, Inc.

Spealman, R. S., McGinnis, R. E. y Ewens, W. J. (1993). Transmission Test for Linkage Disequilibrium: The Insulin Gene Region and Insulin-dependent Diabetes Mellitus (IDDM). *Am. J. Hum. Genet.*, 52, 506–516.

Steel, R. G. y Torrie, J. H. (1980). *Principles and Procedures of Statistics. A Biometrical Approach*. McGraw-Hill, New York, segunda edición.

Thompson, W. A. J. (1962). The problem of negative estimates of variance components. *Ann. Math Statist*, , 33, 273–289.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, 91, 4414–4423.

Vilhjalmsson, B. J. y Nordborg, M. (2012). The nature of confounding in Genome-Wide Association Studies. *Nature Reviews Genetics*, 1–2.

Visscher, P. M., Brown, M. A., McCarthy, M. I. y Yang., J. (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 7–24.

Witkovsky, V. (2012). Estimation, Testing, and Prediction Regions of the Fixed and Random Effects by Solving the Henderson's Mixed Model Equations. *Measurement Science Review*, 12, 234–248.

Witte, J. S. (2010). Genome-Wide Association Studies and Beyond. *Annual Reviews Public Health*, 31, 9–20.

Xiao, Y., Liu, H., Wu, L., Warburton, M. y Yan, J. (2017). Genome-wide Association Studies in Maize: Praise and Stargaze. *Molecular Plant*, 10, 359–374.

Yang, J., Zaitlen, N. A., Goddard, M. E., Visschar, P. M. y Price, A. L. (2014). Advantages and pitfalls in the application of mixed model association methods. *Nat Genet.*, 46, 100–106.

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *Agricultural Science*, 26, 424–455.

Yates, F. (1940). THE RECOVERY OF INTER-BLOCK INFORMATION IN BALANCED INCOMPLETE BLOCK DESIGNS. *Annals of Eugenics*, 10, 317–325.

Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. y Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *nature genetics*, 38, 203–208.

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

Zhou, X. y Stephens, M. (2012). Genome-wide Efficient Mixed Model Analysis for Association Studies. *Nat Genet.*, 44, 821–824.

ANEXOS

ANEXO A: Indicadores TDV y TDF

SNPs[1]	MCE[2]	GEN[3]	TB[4]	VG[5]	EC[6]	TDV[7]		TDF[8]	
						MP[9]	MIP[10]	MP	MIP
1000	5	1000	5	0.3	$\mathbf{I}$	0.8794	0.9998	0.0191	0.2482
1000	5	1000	5	0.3	AR(1), $\rho = 0.25$	0.8694	0.9996	0.0191	0.2485
1000	5	1000	5	0.3	AR(1), $\rho = 0.60$	0.8436	0.9998	0.0177	0.2482
1000	5	1000	5	1.0	$\mathbf{I}$	0.8372	0.9992	0.0176	0.2511
1000	5	1000	5	1.0	AR(1), $\rho = 0.25$	0.8274	0.9998	0.0170	0.2460
1000	5	1000	5	1.0	AR(1), $\rho = 0.60$	0.8246	0.9982	0.0173	0.2478
1000	5	1000	5	3.0	$\mathbf{I}$	0.8140	0.9994	0.0168	0.2496
1000	5	1000	5	3.0	AR(1), $\rho = 0.25$	0.8280	0.9996	0.0169	0.2483
1000	5	1000	5	3.0	AR(1), $\rho = 0.60$	0.8148	0.9988	0.0165	0.2478
1000	5	1000	10	0.3	$\mathbf{I}$	0.8722	1.0000	0.0186	0.2461
1000	5	1000	10	0.3	AR(1), $\rho = 0.25$	0.8554	0.9998	0.0184	0.2485
1000	5	1000	10	0.3	AR(1), $\rho = 0.60$	0.8392	1.0000	0.0171	0.2487
1000	5	1000	10	1.0	$\mathbf{I}$	0.8262	0.9996	0.0170	0.2468
1000	5	1000	10	1.0	AR(1), $\rho = 0.25$	0.8308	0.9994	0.0169	0.2477
1000	5	1000	10	1.0	AR(1), $\rho = 0.60$	0.8090	1.0000	0.0170	0.2475
1000	5	1000	10	3.0	$\mathbf{I}$	0.8118	0.9990	0.0169	0.2497
1000	5	1000	10	3.0	AR(1), $\rho = 0.25$	0.8140	0.9984	0.0167	0.2475
1000	5	1000	10	3.0	AR(1), $\rho = 0.60$	0.8158	0.9998	0.0170	0.2459
1000	5	1000	20	0.3	$\mathbf{I}$	0.8662	1.0000	0.0191	0.2458
1000	5	1000	20	0.3	AR(1), $\rho = 0.25$	0.8652	1.0000	0.0184	0.2470
1000	5	1000	20	0.3	AR(1), $\rho = 0.60$	0.8362	0.9994	0.0179	0.2496
1000	5	1000	20	1.0	$\mathbf{I}$	0.8262	0.9994	0.0170	0.2451
1000	5	1000	20	1.0	AR(1), $\rho = 0.25$	0.8206	0.9996	0.0169	0.2473
1000	5	1000	20	1.0	AR(1), $\rho = 0.60$	0.8114	0.9998	0.0164	0.2462
1000	5	1000	20	3.0	$\mathbf{I}$	0.8186	0.9994	0.0162	0.2447
1000	5	1000	20	3.0	AR(1), $\rho = 0.25$	0.8166	0.9992	0.0165	0.2459
1000	5	1000	20	3.0	AR(1), $\rho = 0.60$	0.8118	0.9994	0.0161	0.2459

SNPs	MCE	GEN	TB	VG	EC	TDV		TDF	
						MP	MIP	MP	MIP
1000	20	1000	5	0.3	$\mathbf{I}$	0.0873	0.8287	0.0171	0.4038
1000	20	1000	5	0.3	AR(1), $\rho = 0.25$	0.0856	0.8247	0.0158	0.4022
1000	20	1000	5	0.3	AR(1), $\rho = 0.60$	0.0815	0.8209	0.0155	0.4017
1000	20	1000	5	1.0	$\mathbf{I}$	0.0784	0.8104	0.0147	0.3981
1000	20	1000	5	1.0	AR(1), $\rho = 0.25$	0.0766	0.8049	0.0145	0.3952
1000	20	1000	5	1.0	AR(1), $\rho = 0.60$	0.0715	0.8073	0.0135	0.3938
1000	20	1000	5	3.0	$\mathbf{I}$	0.0706	0.8050	0.0122	0.3943
1000	20	1000	5	3.0	AR(1), $\rho = 0.25$	0.0729	0.8038	0.0133	0.3933
1000	20	1000	5	3.0	AR(1), $\rho = 0.60$	0.0692	0.8000	0.0124	0.3905
1000	20	1000	10	0.3	$\mathbf{I}$	0.0871	0.8232	0.0156	0.4037
1000	20	1000	10	0.3	AR(1), $\rho = 0.25$	0.0866	0.8225	0.0156	0.4040
1000	20	1000	10	0.3	AR(1), $\rho = 0.60$	0.0806	0.8141	0.0151	0.3994
1000	20	1000	10	1.0	$\mathbf{I}$	0.0732	0.8114	0.0136	0.3958
1000	20	1000	10	1.0	AR(1), $\rho = 0.25$	0.0747	0.8068	0.0142	0.3973
1000	20	1000	10	1.0	AR(1), $\rho = 0.60$	0.0746	0.8085	0.0141	0.3973
1000	20	1000	10	3.0	$\mathbf{I}$	0.0707	0.8077	0.0133	0.3939
1000	20	1000	10	3.0	AR(1), $\rho = 0.25$	0.0718	0.8023	0.0138	0.3917
1000	20	1000	10	3.0	AR(1), $\rho = 0.60$	0.0697	0.8023	0.0130	0.3920
1000	20	1000	20	0.3	$\mathbf{I}$	0.0820	0.8239	0.0153	0.4058
1000	20	1000	20	0.3	AR(1), $\rho = 0.25$	0.0806	0.8229	0.0146	0.4005
1000	20	1000	20	0.3	AR(1), $\rho = 0.60$	0.0776	0.8153	0.0143	0.4029
1000	20	1000	20	1.0	$\mathbf{I}$	0.0750	0.8085	0.0139	0.3956
1000	20	1000	20	1.0	AR(1), $\rho = 0.25$	0.0752	0.8083	0.0142	0.3959
1000	20	1000	20	1.0	AR(1), $\rho = 0.60$	0.0714	0.8063	0.0129	0.3950
1000	20	1000	20	3.0	$\mathbf{I}$	0.0728	0.8061	0.0134	0.3940
1000	20	1000	20	3.0	AR(1), $\rho = 0.25$	0.0675	0.8051	0.0126	0.3942
1000	20	1000	20	3.0	AR(1), $\rho = 0.60$	0.0683	0.8040	0.0129	0.3935
1000	200	1000	5	0.3	$\mathbf{I}$	0.0393	0.3602	0.0367	0.3585
1000	200	1000	5	0.3	AR(1), $\rho = 0.25$	0.0385	0.3612	0.0361	0.3606
1000	200	1000	5	0.3	AR(1), $\rho = 0.60$	0.0355	0.3614	0.0331	0.3600
1000	200	1000	5	1.0	$\mathbf{I}$	0.0353	0.3591	0.0330	0.3578
1000	200	1000	5	1.0	AR(1), $\rho = 0.25$	0.0347	0.3571	0.0324	0.3560

SNPs	MCE	GEN	TB	VG	EC	TDV		TDF	
						MP	MIP	MP	MIP
1000	200	1000	5	1.0	AR(1), $\rho = 0.60$	0.0326	0.3579	0.0304	0.3557
1000	200	1000	5	3.0	$\mathbf{I}$	0.0334	0.3557	0.0312	0.3547
1000	200	1000	5	3.0	AR(1), $\rho = 0.25$	0.0331	0.3526	0.0309	0.3521
1000	200	1000	5	3.0	AR(1), $\rho = 0.60$	0.0320	0.3508	0.0297	0.3501
1000	200	1000	10	0.3	$\mathbf{I}$	0.0373	0.3624	0.0351	0.3614
1000	200	1000	10	0.3	AR(1), $\rho = 0.25$	0.0380	0.3611	0.0356	0.3597
1000	200	1000	10	0.3	AR(1), $\rho = 0.60$	0.0360	0.3572	0.0337	0.3566
1000	200	1000	10	1.0	$\mathbf{I}$	0.0332	0.3591	0.0310	0.3576
1000	200	1000	10	1.0	AR(1), $\rho = 0.25$	0.0329	0.3557	0.0309	0.3550
1000	200	1000	10	1.0	AR(1), $\rho = 0.60$	0.0334	0.3548	0.0313	0.3534
1000	200	1000	10	3.0	$\mathbf{I}$	0.0317	0.3554	0.0299	0.3547
1000	200	1000	10	3.0	AR(1), $\rho = 0.25$	0.0314	0.3535	0.0295	0.3529
1000	200	1000	10	3.0	AR(1), $\rho = 0.60$	0.0340	0.3522	0.0319	0.3516
1000	200	1000	20	0.3	$\mathbf{I}$	0.0395	0.3599	0.0371	0.3588
1000	200	1000	20	0.3	AR(1), $\rho = 0.25$	0.0389	0.3614	0.0366	0.3602
1000	200	1000	20	0.3	AR(1), $\rho = 0.60$	0.0376	0.3590	0.0352	0.3580
1000	200	1000	20	1.0	$\mathbf{I}$	0.0351	0.3558	0.0325	0.3543
1000	200	1000	20	1.0	AR(1), $\rho = 0.25$	0.0357	0.3560	0.0333	0.3542
1000	200	1000	20	1.0	AR(1), $\rho = 0.60$	0.0330	0.3539	0.0307	0.3530
1000	200	1000	20	3.0	$\mathbf{I}$	0.0341	0.3555	0.0319	0.3543
1000	200	1000	20	3.0	AR(1), $\rho = 0.25$	0.0329	0.3542	0.0306	0.3522
1000	200	1000	20	3.0	AR(1), $\rho = 0.60$	0.0340	0.3523	0.0320	0.3506
1000	5	3000	20	0.3	$\mathbf{I}$	0.9818	1.0000	0.0355	0.4617
1000	5	3000	20	0.3	AR(1), $\rho = 0.25$	0.9780	1.0000	0.0354	0.4587
1000	5	3000	20	0.3	AR(1), $\rho = 0.60$	0.9688	1.0000	0.0301	0.4670
1000	5	3000	20	1.0	$\mathbf{I}$	0.9768	1.0000	0.0344	0.4671
1000	5	3000	20	1.0	AR(1), $\rho = 0.25$	0.9714	1.0000	0.0346	0.4621
1000	5	3000	20	1.0	AR(1), $\rho = 0.60$	0.9724	1.0000	0.0338	0.4618
1000	5	3000	20	3.0	$\mathbf{I}$	0.9736	1.0000	0.0340	0.4642
1000	5	3000	20	3.0	AR(1), $\rho = 0.25$	0.9746	1.0000	0.0344	0.4625
1000	5	3000	20	3.0	AR(1), $\rho = 0.60$	0.9772	1.0000	0.0336	0.4598

SNPs	MCE	GEN	TB	VG	EC	TDV		TDF	
						MP	MIP	MP	MIP
1000	20	3000	20	0.3	$\mathbf{I}$	0.0800	0.8696	0.0152	0.5918
1000	20	3000	20	0.3	AR(1), $\rho = 0.25$	0.0786	0.8677	0.0152	0.5933
1000	20	3000	20	0.3	AR(1), $\rho = 0.60$	0.0786	0.8663	0.0146	0.5930
1000	20	3000	20	1.0	$\mathbf{I}$	0.0742	0.8669	0.0141	0.5916
1000	20	3000	20	1.0	AR(1), $\rho = 0.25$	0.0763	0.8636	0.0146	0.5927
1000	20	3000	20	1.0	AR(1), $\rho = 0.60$	0.0729	0.8615	0.0141	0.5905
1000	20	3000	20	3.0	$\mathbf{I}$	0.0727	0.8612	0.0134	0.5915
1000	20	3000	20	3.0	AR(1), $\rho = 0.25$	0.0690	0.8636	0.0132	0.5893
1000	20	3000	20	3.0	AR(1), $\rho = 0.60$	0.0765	0.8630	0.0149	0.5919
1000	200	3000	20	0.3	$\mathbf{I}$	0.0131	0.6153	0.0122	0.6082
1000	200	3000	20	0.3	AR(1), $\rho = 0.25$	0.0127	0.6158	0.0118	0.6089
1000	200	3000	20	0.3	AR(1), $\rho = 0.60$	0.0120	0.6117	0.0114	0.6042
1000	200	3000	20	1.0	$\mathbf{I}$	0.0116	0.6095	0.0108	0.6026
1000	200	3000	20	1.0	AR(1), $\rho = 0.25$	0.0120	0.6120	0.0113	0.6045
1000	200	3000	20	1.0	AR(1), $\rho = 0.60$	0.0112	0.6105	0.0104	0.6028
1000	200	3000	20	3.0	$\mathbf{I}$	0.0118	0.6070	0.0110	0.5997
1000	200	3000	20	3.0	AR(1), $\rho = 0.25$	0.0116	0.6073	0.0107	0.5997
1000	200	3000	20	3.0	AR(1), $\rho = 0.60$	0.0122	0.6100	0.0112	0.6026
5000	5	1000	20	0.3	$\mathbf{I}$	0.9008	1.0000	0.0215	0.2197
5000	5	1000	20	0.3	AR(1), $\rho = 0.25$	0.8928	0.9998	0.0208	0.2208
5000	5	1000	20	1.0	$\mathbf{I}$	0.8624	0.9996	0.0191	0.2245
5000	5	1000	20	1.0	AR(1), $\rho = 0.25$	0.8574	1.0000	0.0191	0.2223
5000	5	1000	20	3.0	$\mathbf{I}$	0.8362	0.9998	0.0174	0.2216
5000	5	1000	20	3.0	AR(1), $\rho = 0.25$	0.8400	0.9994	0.0186	0.2179
5000	20	1000	20	0.3	$\mathbf{I}$	0.0252	0.7242	0.0063	0.3775
5000	20	1000	20	0.3	AR(1), $\rho = 0.25$	0.0220	0.7277	0.0060	0.3778
5000	20	1000	20	1.0	$\mathbf{I}$	0.0227	0.7152	0.0057	0.3729
5000	20	1000	20	1.0	AR(1), $\rho = 0.25$	0.0226	0.7205	0.0060	0.3755
5000	20	1000	20	3.0	$\mathbf{I}$	0.0211	0.7094	0.0053	0.3700
5000	20	1000	20	3.0	AR(1), $\rho = 0.25$	0.0200	0.7111	0.0056	0.3704
5000	200	1000	20	0.3	$\mathbf{I}$	0.0082	0.3347	0.0079	0.3309
5000	200	1000	20	0.3	AR(1), $\rho = 0.25$	0.0087	0.3350	0.0083	0.3308

SNPs	MCE	GEN	TB	VG	EC	TDV		TDF	
						MP	MIP	MP	MIP
5000	200	1000	20	1.0	$\mathbf{I}$	0.0075	0.3286	0.0071	0.3250
500	200	1000	20	1.0	AR(1), $\rho = 0.25$	0.0076	0.3304	0.0073	0.3261
5000	200	1000	20	3.0	$\mathbf{I}$	0.0063	0.3268	0.0060	0.3223
5000	200	1000	20	3.0	AR(1), $\rho = 0.25$	0.0069	0.3259	0.0067	0.3218
15000	5	1000	20	0.3	$\mathbf{I}$	0.8054	0.9980	0.0192	0.2255
15000	5	1000	20	0.3	AR(1), $\rho = 0.25$	0.7990	0.9990	0.0184	0.2200
15000	5	1000	20	1.0	$\mathbf{I}$	0.7604	0.9982	0.0167	0.2215
15000	5	1000	20	1.0	AR(1), $\rho = 0.25$	0.7604	0.9986	0.0164	0.2214
15000	5	1000	20	3.0	$\mathbf{I}$	0.7430	0.9992	0.0159	0.2185
15000	5	1000	20	3.0	AR(1), $\rho = 0.25$	0.7440	0.9980	0.0158	0.2193
15000	20	1000	20	0.3	$\mathbf{I}$	0.0156	0.6125	0.0052	0.3430
15000	20	1000	20	0.3	AR(1), $\rho = 0.25$	0.0176	0.6103	0.0055	0.3412
15000	20	1000	20	1.0	$\mathbf{I}$	0.0152	0.5973	0.0045	0.3363
15000	20	1000	20	1.0	AR(1), $\rho = 0.25$	0.0135	0.6021	0.0043	0.3376
15000	20	1000	20	3.0	$\mathbf{I}$	0.0132	0.5968	0.0044	0.3316
15000	20	1000	20	3.0	AR(1), $\rho = 0.25$	0.0129	0.5938	0.0041	0.3354
15000	200	1000	20	0.3	$\mathbf{I}$	0.0101	0.3207	0.0091	0.3157
15000	200	1000	20	0.3	AR(1), $\rho = 0.25$	0.0096	0.3214	0.0088	0.3167
15000	200	1000	20	1.0	$\mathbf{I}$	0.0081	0.3183	0.0074	0.3143
15000	200	1000	20	1.0	AR(1), $\rho = 0.25$	0.0073	0.3160	0.0066	0.3110
15000	200	1000	20	3.0	$\mathbf{I}$	0.0073	0.3150	0.0067	0.3109
15000	200	1000	20	3.0	AR(1), $\rho = 0.25$	0.0073	0.3127	0.0066	0.3087
5000	5	3000	20	0.3	$\mathbf{I}$	0.9172	1.0000	0.0203	0.3984
5000	5	3000	20	0.3	AR(1), $\rho = 0.25$	0.9282	0.9998	0.0203	0.3974
5000	5	3000	20	1.0	$\mathbf{I}$	0.9084	1.000	0.0195	0.3993
5000	5	3000	20	1.0	AR(1), $\rho = 0.25$	0.9138	1.0000	0.0196	0.3958
5000	5	3000	20	3.0	$\mathbf{I}$	0.9106	1.0000	0.0191	0.3983
5000	5	3000	20	3.0	AR(1), $\rho = 0.25$	0.9096	1.0000	0.0193	0.3972
5000	20	3000	20	0.3	$\mathbf{I}$	0.0601	0.7674	0.0145	0.5401
5000	20	3000	20	0.3	AR(1), $\rho = 0.25$	0.0581	0.7615	0.0131	0.5407
5000	20	3000	20	1.0	$\mathbf{I}$	0.0542	0.7551	0.0121	0.5370
5000	20	3000	20	1.0	AR(1), $\rho = 0.25$	0.0552	0.7522	0.0122	0.5382

SNPs	MCE	GEN	TB	VG	EC	TDV		TDF	
						MP	MIP	MP	MIP
5000	20	3000	20	3.0	$\mathbf{I}$	0.0551	0.7491	0.0119	0.5365
5000	20	3000	20	3.0	AR(1), $\rho = 0.25$	0.0552	0.7515	0.0126	0.5377
5000	200	3000	20	0.3	$\mathbf{I}$	0.0193	0.5478	0.0190	0.5436
5000	200	3000	20	0.3	AR(1), $\rho = 0.25$	0.0163	0.5477	0.0160	0.5431
5000	200	3000	20	1.0	$\mathbf{I}$	0.0164	0.5471	0.0159	0.5426
5000	200	3000	20	1.0	AR(1), $\rho = 0.25$	0.0172	0.5433	0.0169	0.5389
5000	200	3000	20	3.0	$\mathbf{I}$	0.0160	0.5435	0.0158	0.5383
5000	200	3000	20	3.0	AR(1), $\rho = 0.25$	0.0166	0.5442	0.0163	0.5394
15000	5	3000	20	0.3	$\mathbf{I}$	0.8990	0.9998	0.0232	0.4025
15000	5	3000	20	0.3	AR(1), $\rho = 0.25$	0.9050	1.0000	0.0233	0.4036
15000	5	3000	20	1.0	$\mathbf{I}$	0.8948	1.0000	0.0223	0.4036
15000	5	3000	20	1.0	AR(1), $\rho = 0.25$	0.8910	1.0000	0.0218	0.4034
15000	5	3000	20	3.0	$\mathbf{I}$	0.8816	1.0000	0.0217	0.4042
15000	5	3000	20	3.0	AR(1), $\rho = 0.25$	0.8912	1.0000	0.0218	0.4017
15000	20	3000	20	0.3	$\mathbf{I}$	0.0162	0.8385	0.0026	0.5499
15000	20	3000	20	0.3	AR(1), $\rho = 0.25$	0.0154	0.8369	0.0026	0.5488
15000	20	3000	20	1.0	$\mathbf{I}$	0.0164	0.8306	0.0027	0.5467
15000	20	3000	20	1.0	AR(1), $\rho = 0.25$	0.0151	0.8297	0.0026	0.5465
15000	20	3000	20	3.0	$\mathbf{I}$	0.0142	0.8260	0.0025	0.5420
15000	20	3000	20	3.0	AR(1), $\rho = 0.25$	0.0146	0.8275	0.0025	0.5446
15000	200	3000	20	0.3	$\mathbf{I}$	0.0079	0.5802	0.0073	0.5750
15000	200	3000	20	0.3	AR(1), $\rho = 0.25$	0.0081	0.5779	0.0076	0.5737
15000	200	3000	20	1.0	$\mathbf{I}$	0.0086	0.5746	0.0079	0.5696
15000	200	3000	20	1.0	AR(1), $\rho = 0.25$	0.0077	0.5759	0.0072	0.5707
15000	200	3000	20	3.0	$\mathbf{I}$	0.0075	0.5723	0.0069	0.5674
15000	200	3000	20	3.0	AR(1), $\rho = 0.25$	0.0072	0.5758	0.0067	0.5703

Notas

[1]Número de SNPs a probar.

[2]Número de SNPs con tamaño de efecto distinto de cero.

[3]Número de genotipos.

[4]Tamaño de bloque.

[5]Varianza genética.

[6]Estructura de la matriz $\mathbf{C}$.

[7]Tasa de detecciones verdaderas.

[8]Tasa de detecciones falsas.

[9]Prueba estadística propuesta.

[10]Método implementado en PLINK.

ANEXO B: Código computacional R

El código R usado para probar asociación entre un fenotipo y y un conjunto de marcadores genéticos cuyas entradas son -1, 0 y 1 utilizando el $BLUP$ y su matriz de varianzas y covarianzas se divide en dos funciones principales de acuerdo a la hipótesis nula que se desea probar. Una función para probar aditividad y otra función para probar heterosis.

En el siguiente código R se muestra un ejemplo de como utilizar las funciones para probar asociación. En él se encuentran comentarios que ayudan a entender los elementos de entrada y los elementos de salida de las funciones.

```
rm(list = objects()); ls() ## Limpiar espacio de trabajo
setwd('C:/Users/~')        ## Ubicación carpeta de trabajo.

Map<-read.table(file="./Data/MAPA_SNP5000.txt",
     header=TRUE,sep="\t",as.is=TRUE) ## Marcadores y cromosomas
MMK<-read.table(file="./Data/MMN_SNP5000_GEN3000.txt",
     header=TRUE,sep="\t",as.is=TRUE) ## Matriz de SNPs que incluye
                                      ## nombre de líneas.
FEN<-read.table(file="./Data/SNP5000_MS5_GEN3000_TB20_VG1_VE.25.txt",
     header=TRUE,sep="\t",as.is=TRUE) ## Matriz que contiene 1000 BLUPs
                                      ## simulados.
COV<-read.table(file="./Data/COV_SNP5000_MS5_GEN3000_TB20_VG1_VE.25.txt",
    header=FALSE,sep="\t",as.is=TRUE) ## Covarianza simulada del BLUP.

S<-as.matrix(COV)       ## Matriz gxg de covarianza del BLUP.
U<-FEN$y1000            ## Vector BLUP de tamaño g.
M<-MMK[,-1]             ## Matriz de m SNPs y g genotipos.

###PRUEBA ADITIVIDAD###
source("./R/Funciones_A.R")   ## Archivo R que contiene todas las
                              ## funciones utilizadas en la función
                              ## principal TGWAS_A.
GWAS_ANALISIS<-TGWAS_A(M,U,S) ## Función R para probar la hipótesis de
                              ## aditividad. Los valores de entrada son
                              ## la matriz M de SNPs, el BLUP U y la
                              ## matriz de covarianza del BLUP.
```

```
## SALIDA DE LA FUNCIÓN ##
str(GWAS_ANALISIS)
List of 4
$ Marker: chr [1:5000] "snp1" "snp2" "snp3" ...## Nombre del SNP.
$ Stat_T: num [1:5000] 0.989 -1.028 -0.858 ... ## Estadística T.
$ df    : num [1:5000] 2241 2236 2220 ...   ## Grados de libertad.
$ pvalue: num [1:5000] 0.323 0.304 0.391 ... ## Valor-p no ajustado.

###PRUEBA HETEROSIS###
source("./R/Funciones_H.R")    ## Archivo R que contiene todas las
                               ## funciones utilizadas en la función
                               ## principal TGWAS_H.
TGWAS_H(M,U,S) ## Función R para probar la hipótesis de heterosis.
               ## Los valores de entrada son la matriz M de SNPs,
               ## el BLUP U y la matriz de covarianza del BLUP.

## SALIDA DE LA FUNCIÓN ##
List of 4
$ Marker: chr [1:5000] "snp1" "snp2" "snp3" ... ## Nombre del SNP.
$ Stat_T: num [1:5000] 1.47 1.58 1.47 1.76 ... ## Estadística T.
$ df    : num [1:5000] 1366 1398 1444 1439 ... ## Grados de libertad.
$ pvalue: num [1:5000] 0.142 0.115 0.141 0.079 ... ## Valor-p no ajustado.
```

A continuación se describen el conjunto de funciones que componen cada una de las funciones por separado, empezando por la función para probar efectos aditivos.

```
## EL ARCHIVO Funciones_A.R CONTIENE LAS SIGUIENTES FUNCIONES ##
genA_vlambda<-function(x) ## Esta función genera el vector lambda que
{ind1<-which(x==1)        ## define un contraste para el SNP x.
ind2<-which(x==-1)
n1<-length(ind1)
n2<-length(ind2)
n<-length(x)
factor<-1/(n1*n2)
lambda<-rep(0,n)
lambda[ind1]<-n2
lambda[ind2]<-(-1*n1 )
return(factor*lambda)}
```

```
genA_nu<-function(x,U) ## Esta función tiene como entrada el vector x
{ind1<-which(x==1)     ## de SNP y el BLUP U, y la salida son los
ind2<-which(x==-1)     ## grados de libertad estimados vía
y1<-U[ind1]            ## Satterthwaite.
y2<-U[ind2]
n1<-length(y1)
n2<-length(y2)
v1<-round(var(y1),5)
v2<-round(var(y2),5)
nu<-((v1/n1+v2/n2)^2)/(v1^2/(n1^2*(n1-1))+v2^2/(n2^2*(n2-1)))
return(nu)}

gen_T<-function(lambda,U,S)  ## Esta función genera el estadístico T.
{valor_T<-round((t(lambda)%*%U)/sqrt(t(lambda)%*%S%*%lambda),4)
return(valor_T)}             ## Tiene como entradas el vector de
                             ## contraste lambda, el BLUP U y la
                             ## matriz S de covarianza del BLUP.

test_A<-function(x,U,S)  ## Esta función calcula el p-value a partir de
{lambda<-genA_vlambda(x) ## la distribución T. Las entradas son el
est_T<-gen_T(lambda,U,S) ## vector SNP x, el BLUP U y la covarianza S.
nu<-genA_nu(x,U)
pvalue<-2*pt(q=abs(est_T), df=nu, lower.tail = FALSE, log.p = FALSE)
return(c(est_T,nu,pvalue))} ## La función devuelve tres valores, la
                            ## estadística T, los grados de libertad nu
                            ## y el valor p de la prueba.

TGWAS_A<-function(M,U,S) ## Función principal para probar efectos
{m<-length(M[1,])        ## aditivos.
RES<-matrix(NA,nrow=m,ncol=3)
Marker<-colnames(M)
i<-1
while(i<=m)
{x<-M[,i]
RES[i,]<-test_A(x,U,S)
i<-i+1 }
return(list(Marker=Marker,Stat_T=RES[,1],df=RES[,2],pvalue=RES[,3]))}
```

Por último, el conjunto de funciones que componen la función principal para probar efectos de heterosis.

```
genH_vlambda<-function(x) ## Genera el vector lambda de contraste a
{ind1<-which(x==-1)       ## partir del SNP x.
ind2<-which(x==0)
ind3<-which(x==1)
n1<-length(ind1)
n2<-length(ind2)
n3<-length(ind3)
n<-n1+n2+n3
lambda<-rep(0,n)
lambda[ind1]<-0.5*n3*n2
lambda[ind2]<-(-1)*n3*n1
lambda[ind3]<-0.5*n2*n1
factor<-1/(n1*n2*n3)
return(factor*lambda)}

genH_nu<-function(x,U)  ## Genera los grados de libertad utilizando
{ind1<-which(x==-1)     ## el SNP x y el BLUP U.
ind2<-which(x==0)
ind3<-which(x==1)
y1<-U[ind1]
y2<-U[ind2]
y3<-U[ind3]
n1<-length(y1)
n2<-length(y2)
n3<-length(y3)
v1<-round(var(y1),5)
v2<-round(var(y2),5)
v3<-round(var(y3),5)
a<-((0.5^2)*v1/n1+v2/n2+(0.5^2)*v3/n3)^2
b1<-(0.5^4)*(v1^2)/(n1^2*(n1-1))
b2<-(v2^2)/((n2^2)*(n2-1))
b3<-(0.5^4)*(v3^2)/(n3^2*(n3-1))
b<-b1+b2+b3
return(a/b)}

gen_T<-function(lambda,U,S) ## Genera la estadística T.
```

```
{valor_T<-round((t(lambda)%*%U)/sqrt(t(lambda)%*%S%*%lambda),4)
return(valor_T)}              ## Los elementos de entrada son el vector
                              ## lambda, el BLUP U y la covarianza S.

test_H<-function(x,U,S)  ## Genera el valor p a partir de la
{lambda<-genH_vlambda(x) ## distribución T.
est_T<-gen_T(lambda,U,S)
nu<-genH_nu(x,U)
pvalue<-2*pt(q=abs(est_T), df=nu, lower.tail = FALSE, log.p = FALSE)
return(c(est_T,nu,pvalue))} ## Los valores de salida son la estadística
                              ## T, los grados de libertad nu y el valor-p
                              ## de la prueba.

TGWAS_H<-function(M,U,S)  ## Función principal para probar efectos de
{m<-length(M[1,])         ## heterosis.
RES<-matrix(NA,nrow=m,ncol=3)
Marker<-colnames(M)
i<-1
while(i<=m)
{x<-M[,i]
RES[i,]<-test_H(x,U,S)
print(i)
i<-i+1 }
return(list(Marker=Marker,Stat_T=RES[,1],df=RES[,2],pvalue=RES[,3]))}
```