BAG | Journal of Basic & Applied Genetics

Vol XXX (1): 17-23; July 2019

Journal of the Argentine Society of Genetics

TEST OF INTERACTION IN THE ANALYSIS OF MOLECULAR VARIANCE

PRUEBA DE INTERACCIÓN EN EL ANÁLISIS MOLECULAR DE VARIANZA

Bruno C. [1,2], Videla M.E. [1,2], Balzarini M. [1,2]

[1] Consejo Nacional de Investigaciones Científicas y Técnicas-CONICET

[2] Estadística y Biometría. Facultad de Ciencias Agropecuarias. Universidad Nacional de Córdoba

Corresponding author:
Cecilia Bruno
cebruno@agro.unc.edu.ar

Cite this article as:
Bruno C., Videla M.E., Balzarini M. 2019. TEST OF INTERACTION IN THE ANALYSIS OF MOLECULAR VARIANCE. BAG. Journal of Basic and Applied Genetics XXX (1): 17-23.

Received: 05/04/2019
Accepted: 06/24/2019

General Editor: Elsa Camadro

DOI: 10.35407/bag.2019.XXX.01.03

ISSN online version: 1852-6233

Available online at
www.sag.org.ar/jbag

ABSTRACT

The genomic diversity, expressed in the differences between molecular haplotypes of a group of individuals, can be divided into components of variability between and within some factor of classification of the individuals. For such variance partitioning, molecular analysis of variance (AMOVA) is used, which is constructed from the multivariate distances between pairs of haplotypes. The classical AMOVA allows the evaluation of the statistical significance of two or more hierarchical factors and consequently there is no interaction test between factors. However, there are situations where the factors that classify individuals are crossed rather than nested, that is, all the levels of a factor are represented in each level of the other one. This paper proposes a statistical test to evaluate the interaction between crossed factors in a Non-Hierarchical AMOVA. The null hypothesis of interaction establishes that the molecular differences between individuals of different levels of a factor are the same for all the levels of the other factor that classifies them. The proposed analysis of interaction in a Non-Hierarchical AMOVA includes: calculation of the distance matrix and partition of it into blocks, subsequent calculation of residuals and analysis of non-parametric variance on the residuals. Its implementation is illustrated in simulated and real scenarios. The results suggest that the proposed interaction test for the Non-Hierarchical AMOVA presents high power.

Key words: genetic variability, non-parametric methods, distances matrix, AMOVA.

RESUMEN

La diversidad genómica, expresada en las diferencias entre haplotipos moleculares de un conjunto de individuos, puede dividirse en componentes de variabilidad entre y dentro de algún factor de clasificación de los individuos. Para tal partición de varianzas, se usa análisis molecular de la varianza (AMOVA), el cual se construye a partir de las distancias multivariadas entre pares de haplotipos. El AMOVA clásico permite evaluar la significancia estadística de dos o más factores jerárquicos y consecuentemente no existe prueba de interacción entre factores. Sin embargo, existen situaciones donde los factores que clasifican a los individuos están cruzados y no anidados, es decir todos los niveles de un factor se encuentran representados en cada nivel del otro factor. Este trabajo propone una prueba estadística para evaluar la interacción entre factores cruzados en un AMOVA No-Jerárquico. La hipótesis nula de interacción establece que las diferencias moleculares entre individuos de distintos niveles de un factor son las mismas para todos los niveles del otro factor que los clasifica. La propuesta de análisis de interacción de factores a partir de distancias en un AMOVA No-Jerárquico comprende: cálculo de la matriz de distancia y partición de la misma en bloques, posterior cálculo de residuos y análisis de varianza no-paramétrico sobre los residuos. Su implementación es ilustrada en escenarios simulados y real. Los resultados sugieren que la prueba de interacción propuesta para el AMOVA No-Jerárquico presenta alta potencia.

Palabras clave: variabilidad genética, métodos no-paramétricos, matrices de distancias. AMOVA.

INTRODUCCIÓN

La estructura genética de poblaciones puede analizarse mediante la comparación de las frecuencias alélicas (Kennington *et al.*, 2003; Hedrick, 2005), mediante métricas de distancias genéticas (Nei, 1973), usando algoritmos de clasificación (Pritchard *et al.*, 2000) y/o con el análisis molecular de la varianza (AMOVA, Excoffier, 1992). La mayoría de los métodos basados en frecuencias alélicas involucran transformaciones no lineales de los datos genéticos y son válidos sólo bajo una serie de supuestos que deben realizarse respecto a los procesos evolutivos subyacentes. Por el contrario, la información sobre divergencia a nivel molecular procesada en el formato de una partición de Análisis de la Varianza (ANAVA) demanda menos supuestos biológicos. Debido a la dimensionalidad de los datos genómicos (naturaleza multivariada), el ANAVA se obtiene a partir de las métricas de distancias entre los pares de haplotipos de ADN. Debido a relaciones entre sumas de cuadrado (SC) y sumas de distancias al cuadrado, la SC asociada con cualquier término de un ANAVA puede ser calculada directamente a partir de las distancias (Gower, 1966; Li, 1976). En el AMOVA se descompone la diversidad genómica (expresada por las diferencias en el total de haplotipos moleculares) como la suma de componentes de variabilidad entre y dentro de grupos de individuos. Estos grupos son conformados por uno o más factores de clasificación usualmente anidados, es decir los niveles de un factor pueden ser distintos para cada nivel del otro factor. Un ejemplo usual de este análisis de variabilidad molecular es el estudio de diferencias entre regiones, entre poblaciones dentro de cada región y dentro de poblaciones. Generalmente, se atribuye una proporción aditiva de la variabilidad total a cada uno de los factores presentes en el diseño del estudio (*i.e.* región, población). La comparación de estos componentes de varianza permite inferir la magnitud de la estructuración genética en el conjunto de todos los haplotipos moleculares bajo estudio.

Los cálculos para estimar componentes de varianza entre y dentro de subgrupos de una estructura jerárquica de factores, se realizan desde la matriz de distancias Euclídeas (al cuadrado) y se usan para contrastar hipótesis sobre variabilidad entre y dentro de grupos (Dyer, 2017). Sin embargo, existen situaciones de estructura de datos, provenientes de estudios experimentales u observacionales, donde los factores que clasifican las muestras se encuentran cruzados en lugar de estar anidados en una estructura jerárquica. Es decir, todos los niveles de un factor se encuentran representados en cada nivel del otro factor. Por ejemplo, en un estudio donde se colecten muestras de haplotipos relacionadas a 3 especies (3 niveles para el factor especie) en cada una de 4 regiones (4 niveles para el factor región), pero los niveles del factor especie son los mismos en cada uno de los niveles del factor región, es necesario evaluar la significancia de la interacción entre especie y región para conocer si las diferencias entre especies dependen de la región desde la que se extrajeron las muestras de haplotipos. Luego, en el ANAVA de perfiles moleculares (AMOVA) No-Jerárquico, es de interés responder si las diferencias moleculares entre los niveles de un factor son las mismos para todos los niveles del otro factor interviniente en la clasificación de los haplotipos. Éste, es el interrogante que se pretende responder con pruebas de interacción entre factores de clasificación no-jerárquicos. Por el contrario, si el muestreo involucra diferentes niveles de un factor en cada uno de los diferentes niveles del otro factor, se usará un AMOVA clásico (AMOVA Jerárquico). En el AMOVA clásico, la prueba de interacción no es factible debido a que no existen muestras moleculares para todas las combinaciones de niveles de los factores intervinientes.

Anderson (2001) propuso un método denominado PERMANOVA, análogo al Análisis de la Varianza Multivariado, particionando las sumas de cuadrado asociadas a sub-matrices de matrices de distancias relacionadas a los factores de clasificación de un conjunto de muestras moleculares. El método puede usarse en diseños con factores de clasificación cruzado, pero sólo en situaciones donde los datos son balanceados, *i.e.*, todos los niveles de los factores tienen la misma cantidad de datos y no falta ningún nivel (Anderson y ter Braack, 2003). La significancia estadística de la interacción entre factores en PERMANOVA es aproximada por permutación.

En este trabajo se propone un método, de inferencia no paramétrica, para evaluar la significancia de la interacción entre factores que es aplicable a estudios moleculares con poblaciones clasificadas a dos vías con estructura cruzada de factores.

MATERIALES Y MÉTODOS

Distancias moleculares

Supongamos que existen n muestras de haplotipos caracterizados con m marcadores moleculares. Si los marcadores se expresan como variables binarias, para cada muestra es posible conformar una observación multivariada (m-dimensional) que lleva valores 1 o 0 para cada uno de los marcadores según el marcador esté presente o ausente en la muestra, respectivamente. El vector booleano m-dimensional denotado como $\mathbf{p'}=[p_1,p_2,..,p_m]$ donde $p_i=1$ con ic1,..., m si el marcador m está presente y $p_i=0$ si está ausente, constituye la observación multivarada a analizar en cada muestra. La diferencia entre dos muestras y_j *vs.* y_k es definida como $\mathbf{p}_j-\mathbf{p}_k$. Se define una métrica de distancia Euclídea entre

las muestras y_j e y_k como $d^2_{jk} = (\mathbf{p}_j - \mathbf{p}_k)' \mathbf{W} (\mathbf{p}_j - \mathbf{p}_k)$ donde **W** es una matriz de pesos que pueden ser diferenciables para los distintos marcadores. Si todos los marcadores se asumen independientes e igualmente informativos entonces **W=I** y la métrica de distancia es igual al número de diferencias entre las dos muestras, *i.e.* complemento a uno del índice de similitud Emparejamiento Simple (Apostol *et al.*, 1993). Una distancia comúnmente usada para análisis de similitud o diferencia molecular es la distancia de Excoffier (Excoffier *et al.*, 1992)

$$D_{Excoffier} = (a + d + c + d)\left(1 - \frac{a+d}{a+d+c+d}\right) \quad [1]$$

donde, *a*, *b*, *c*, y *d* representan las frecuencias de los eventos (1,1), (1,0), (0,1) y (0,0) respectivamente, que surgen al comparar dos individuos para cada alelo. En ésta las disimilitudes o faltas de coincidencia son expresadas como fracción del número total de marcadores que participan en la comparación de cada par de perfiles. Si todos los individuos que se comparan son genotipados con el mismo número de marcadores entonces esta distancia es sólo un múltiplo del complemento a 1 del índice de Emparejamiento Simple (Apostol *et al.*, 1993).

$$D_{ES} = 1 - \left[\frac{a+d}{a+d+c+d}\right] \quad [2]$$

Interacción en Análisis de la Varianza Molecular

A continuación, se propone una prueba estadística desarrollada para evaluar la significancia de la interacción *A*×*B* en un diseño con dos factores A y B cruzados que tiene como *input* a las distancias moleculares entre haplotipos.

La matriz de distancias entre pares de individuos se particiona de manera tal que se identifiquen los siguientes bloques de distancias: 1) Bloque de distancias "Dentro": comprende inter-distancias entre individuos de un mismo grupo, por tanto existen tantos bloques de este tipo como grupos (combinaciones de niveles de los factores) haya; 2) Bloque de distancias "Entre factor *C* dentro de factor *A*", existen tantos bloques como niveles del factor *A*;3) Bloque de distancias "Entre factor *A* dentro de factor *B*" existen tantos bloques como niveles del factor *B* y 4) Bloque de distancias "Entre factor *A* y entre factor *B*", con tantos bloques como (*B*-1) *B* (Fig. 1).

Se obtienen las distancias promedio para cada bloque de distancias (B1 a B4), luego se obtiene el valor absoluto de la diferencia entre una distancia del bloque y la media de las distancias del mismo bloque. Sobre el valor absoluto de la diferencia de cada distancia respecto a su media, se ajusta un análisis de varianza no-paramétrico de Kruskal Wallis (Conover, 1999).

A

Matriz de distancia		Factor A	A_1	A_1	A_1	A_1	A_1	A_1	A_2	A_2	A_2	A_2	A_2	A_2
		Factor B	B_1	B_1	B_1	B_2	B_2	B_2	B_1	B_1	B_1	B_2	B_2	B_2
Factor A	Factor B	Ind	1	2	3	4	5	6	7	8	9	10	11	12
A_1	B_1	1	0											
A_1	B_1	2	1	0										
A_1	B_1	3	1	2	0									
A_1	B_2	4	9	8	8	0								
A_1	B_2	5	9	8	8	2	0							
A_1	B_2	6	10	9	9	1	1	0						
A_2	B_1	7	6	5	5	3	5	4	0					
A_2	B_1	8	6	5	5	3	3	4	2	0				
A_2	B_1	9	7	6	6	2	4	3	1	3	0			
A_2	B_2	10	3	2	4	6	6	7	3	3	4	0		
A_2	B_2	11	3	2	4	6	6	7	3	3	4	2	0	
A_2	B_2	12	4	3	5	5	5	6	2	2	3	1	1	0

B

Bloques		Factor A	A_1	A_1	A_1	A_1	A_1	A_1	A_2	A_2	A_2	A_2	A_2	A_2
		Factor B	B_1	B_1	B_1	B_2	B_2	B_2	B_1	B_1	B_1	B_2	B_2	B_2
Factor A	Factor B	Ind	1	2	3	4	5	6	7	8	9	10	11	12
A_1	B_1	1	B_1			B_2			B_3			B_4		
A_1	B_1	2												
A_1	B_1	3												
A_1	B_2	4	B_2			B_1			B_4			B_3		
A_1	B_2	5												
A_1	B_2	6												
A_2	B_1	7	B_3			B_4			B_1			B_2		
A_2	B_1	8												
A_2	B_1	9												
A_2	B_2	10	B_4			B_3			B_2			B_1		
A_2	B_2	11												
A_2	B_2	12												

C

		Factor A	A_1	A_1	A_1	A_1	A_1	A_1	A_2	A_2	A_2	A_2	A_2	A_2
		Factor B	B_1	B_1	B_1	B_2	B_2	B_2	B_1	B_1	B_1	B_2	B_2	B_2
Factor A	Factor B	Ind	1	2	3	4	5	6	7	8	9	10	11	12
A_1	B_1	1	d_{w1}			$d_{B[A]}$								
A_1	B_1	2												
A_1	B_1	3												
A_1	B_2	4	$d_{B[A]}$			d_{w2}								
A_1	B_2	5												
A_1	B_2	6												
A_2	B_1	7	$d_{A[B]}$			d_{AB}			d_{w1}					
A_2	B_1	8												
A_2	B_1	9												
A_2	B_2	10	d_{AB}			$d_{A[B]}$						d_{w2}		
A_2	B_2	11												
A_2	B_2	12												

Figura 1. Esquema ilustrativo de las distancias entre individuos y los bloques conformados por la partición de la matriz de distancia usados para calcular el Análisis Molecular de la Varianza (AMOVA) No-jerárquico. Panel A distancias entre individuos. Panel B partición de la matriz de distancia en bloques. Panel C distancias promedio para cada bloque conformado por la partición de la matriz de distancia.

Nota: B_1: Bloques de distancia "Dentro", B_2: Bloque de distancias entre factor *B*[*A*], B_3: Bloque de distancia entre *A*[*B*], B_4: Bloque de distancia entre factor *A* y entre factor *B*. Los valores d_{w1} y d_{w2} representan los valores absolutos de las diferencias entre una distancia entre individuos del mismo grupo, conformado por la misma combinación de niveles de ambos factores, respecto a la distancia promedio del mismo bloque. $d_{B[A]}$ es el valor absoluto de la diferencia de cada distancia entre los individuos del factor B dentro del factor A respecto a la distancia promedio del bloque, $d_{A[B]}$ es el valor absoluto de las diferencias entre individuos del factor A dentro del factor B respecto a la distancia promedio del bloque, d_{AB} valor absoluto de la diferencia entre factor A y entre factor B respecto a la distancia promedio del bloque.

Evaluación de la prueba de interacción

El método propuesto se evaluó sobre datos simulados y se ilustró su aplicación sobre un conjunto de datos experimentales. Para la evaluación por simulación de la prueba propuesta para la interacción, se simularon dos situaciones hipotéticas para una matriz $n \times m$ donde n=24 observaciones y m=20 marcadores moleculares codificados como binarios (presencia/ausencia). El diseño de experimentos fue bifactorial con dos niveles cada factor. Las situaciones hipotéticas fueron construidas bajo dos escenarios: (A) Interacción estadísticamente significativa. Esto representa que la hipótesis nula es falsa, es decir que al menos una de las diferencias entre los bloques definidos sobre la matriz de distancias es distinta de cero y (B) No hay interacción entre los factores. En términos estadísticos, este escenario implica que la hipótesis nula es verdadera.

En cada una de las situaciones (A) y (B), se simularon 100 bases de datos a partir de una perturbación introducida a través de una distribución binomial con parámetros n=1 y p=0.20. Sobre cada una de las 100 nuevas matrices de datos binarios creados para cada una de las situaciones, se calcularon tres medidas de distancia entre todos los pares de individuos: la distancia de Excoffier, el complemento a uno del índice de emparejamiento simple y la distancia de Bray-Curtis (Bray y Curtis, 1957).

Se calcularon las distancias promedio para cada bloque de distancias a partir del cual se reagruparon los datos. Sobre el valor absoluto de la diferencia de la distancia entre todos los pares de individuos y la media del bloque al que pertenece según el re-agrupamiento, se ajustó un análisis de la varianza no-paramétrico de Kruskal Wallis. Para estimar el tamaño de muestra se simularon 100 corridas o ajustes del análisis propuesto y bajo hipótesis nula verdadera, se contó la cantidad de veces que la prueba no aceptó la hipótesis es decir donde se rechazó una hipótesis verdadera que suponía la no existencia de interacción. Este valor estimado empíricamente representa el "tamaño de la prueba".

Para estimar la potencia de la prueba, al menos empíricamente, se simularon para la situación (A) diferentes niveles de interacción: (A.a) interacción alta, donde las diferencias entre los individuos de los niveles del factor B para un mismo nivel del factor A eran altas respecto a las diferencias entre los niveles de los factores B con el otro nivel de A donde esas diferencias eran prácticamente nulas, (A.b) interacción media, donde la diferencia entre individuos de los niveles de B que se encuentran bajo un mismo nivel del factor A presentan una diferencia máxima en el 60% de sus loci y (A.c) interacción baja; la diferencia máxima se da solo en un 30% de los loci. Para conocer la potencia de la prueba, se contó en cada situación la cantidad de veces que el valor p del análisis de la varianza no-paramétrico arrojó valores por debajo del nivel de significación (α =0.05).

Los resultados obtenidos fueron comparados con los resultados arrojados por el software PERMANOVA con la distancia de Bray-Curtis y distancia Euclídea (Anderson, 2001).

El conjunto de datos experimentales usado para ilustrar la prueba de interacción involucra muestras de ADN provenientes de un patógeno que habita en distintos cultivos agrícolas. Las muestras fueron recolectadas a partir de 12 plantas infectadas, que se clasificaron según el tipo de hospedero (cultivo de invierno y cultivo de verano), y también según la región. Los dos tipos de hospederos se encontraban en cada región. Se obtuvieron 9 marcadores moleculares polimórficos que fueron codificados como binarios (presencia/ausencia) para cada muestra.

RESULTADOS

En la Tabla 1 se presenta para cada situación simulada (A.a: interacción alta, A.b: interacción media, A.c: interacción baja y B: interacción nula) las tasas de error estimadas para la prueba propuesta en este trabajo (AMOVA No-Jerárquico) y para la prueba de interacción implementada en el software PERMANOVA v.1.6 (Anderson y ter Braak, 2003). Ambas fueron aplicadas teniendo como *input* una matriz de distancias entre individuos de dimensión 24×24 y calculada en base a la métrica de Excoffier.

Los resultados en Tabla 1 sugieren que la prueba de interacción propuesta para el AMOVA No-Jerárquico presenta una alta potencia, complemento a uno de la Tasa de Error de Tipo II, para detectar el efecto de interacción entre los factores, aun cuando este efecto es bajo. PERMANOVA v.1.6 a través de 4999 corridas de permutación de filas, produjo también buena potencia aunque siempre menor que el método propuesto. En escenarios donde la diferenciación entre grupos es mayor, como los casos de alta y media interacción, las Tasas de Error II alcanzada por el método propuesto fueron bajas (0.05 en condiciones de alta interacción y del 0.2 en situaciones de media interacción). La ventaja del método propuesto para evaluar interacción en un AMOVA No-Jerárquico es que no necesita que el diseño del estudio tenga el mismo número de repeticiones para cada nivel de factores como requiere PERMANOVA.

Para implementar la prueba estadística de interacción a través del AMOVA No-Jeráquico propuesto, se construyeron las matrices de distancia entre los perfiles moleculares de los genomas del patógeno genotipado en hospedantes de verano e invierno en dos zonas agrícolas (Fig. 2). El ANOVA No-Jerárquico detectó una interacción estadísticamente significativa entre los niveles de los factores, es decir que el perfil molecular del patógeno es diferente según el cultivo donde se hospede y la región

en la que se encuentre. El valor del estadístico Kruskal-Wallis fue de 4.4262 con un valor-p=0.03539. Para el mismo conjunto de datos, PERMANOVA no encontró una interacción estadísticamente significativa. Los análisis se realizaron en el software R (R Core Team, 2013) con el código que ponemos a disposición del lector en el repositorio GitHub @estadistica-aplicada (https://github.com/estadistica-aplicada). El algoritmo para evaluar la interacción en AMOVA No Jerárquico también fue incorporado al software Info-Gen (Balzarini y Di Rienzo, 2018).

Tabla 1. Tasas de Error Tipo II obtenidas por simulación para la prueba de hipótesis de la interacción, bajo situaciones de alta, media y baja interacción y Tasas de Error Tipo I bajo interacción nula, de los procedimientos AMOVA No-Jerárquico y PERMANOVA con el método de permutación de filas.

Procedimiento	Interacción			
	Alta	Media	Baja	Nula
FIDA	0.05	0.2	0.28	0.097
PERMANOVA	0	0	0.15	0.05

Nota: por definición el Error de Tipo I se calcula bajo hipótesis nula cierta, es decir, interacción nula y el Error de Tipo II se calcula bajo hipótesis nula falsa, *i.e.*, hay interacción.

A

Matriz de distancia		Factor A	A_1	A_1	A_1	A_1	A_1	A_1	A_2	A_2	A_2	A_2	A_2	A_2
		Factor B	B_1	B_1	B_1	B_2	B_2	B_2	B_1	B_1	B_1	B_2	B_2	B_2
Factor A	Factor B	Ind	1	2	3	4	5	6	7	8	9	10	11	12
A_1	B_1	1	**0**											
A_1	B_1	2	**3**	**0**										
A_1	B_1	3	**2**	**3**	**0**									
A_1	B_2	4	6	5	8	**0**								
A_1	B_2	5	4	5	6	**4**	**0**							
A_1	B_2	6	5	6	3	**7**	**7**	**0**						
A_2	B_1	7	2	1	2	6	6	5	**0**					
A_2	B_1	8	2	1	2	6	6	5	**0**	**0**				
A_2	B_1	9	5	6	7	3	1	6	**7**	**7**	**0**			
A_2	B_2	10	6	5	6	2	6	5	6	6	5	**0**		
A_2	B_2	11	6	5	6	4	2	5	6	6	1	**6**	**0**	
A_2	B_2	12	3	6	5	5	3	4	5	5	2	**7**	**3**	**0**

B

Matriz de residuos		Factor A	A_1	A_1	A_1	A_1	A_1	A_1	A_2	A_2	A_2	A_2	A_2	A_2
		Factor B	B_1	B_1	B_1	B_2	B_2	B_2	B_1	B_1	B_1	B_2	B_2	B_2
Factor A	Factor B	Ind	1	2	3	4	5	6	7	8	9	10	11	12
A_1	B_1	1	0											
A_1	B_1	2	0.33	0										
A_1	B_1	3	0.67	0.33	0									
A_1	B_2	4				0								
A_1	B_2	5				2	0							
A_1	B_2	6				1	1	0						
A_2	B_1	7							0					
A_2	B_1	8							4.67	0				
A_2	B_1	9							2.33	2.33	0			
A_2	B_2	10										0		
A_2	B_2	11										0.67	0	
A_2	B_2	12										1.67	2.33	0

Figura 2. Esquema ilustrativo de las distancias entre individuos estimadas desde perfiles multivariados codificados como binarios y los bloques conformados por la partición de la matriz de distancia usados para evaluar interacción en (AMOVA) No-jerárquico. Referencias: El Factor A representa la región del cultivo, con dos niveles (A_1 y A_2) y el Factor B representa si el cultivo hospedero del patógeno es de invierno (B_1) o verano (B_2). Panel A distancias entre individuos. Panel B residuos estimados como la diferencia de la distancia entre un par de individuos dentro del mismo bloque y la distancia promedio del bloque.

DISCUSIÓN

En este trabajo se propone una prueba estadística para analizar la interacción entre factores en un AMOVA No-Jerárquico. La hipótesis nula de interacción establece que las diferencias entre perfiles moleculares agrupados por un factor de clasificación son las mismas para cada nivel del factor de clasificación con el que se ha cruzado. A diferencia del AMOVA de Excoffier (Excoffier *et al.*, 1992) y de PERMANOVA (Anderson, 2001) métodos que permiten evaluar el efecto de factores de clasificación a partir de distancias moleculares, la prueba estadística presentada en nuestro trabajo es específica de interacción y puede implementarse incluso en situaciones de desbalance de datos. Cuando la interacción es estadísticamente significativa, no se recomienda realizar pruebas de efectos principales.

La prueba de interacción que proponemos no demanda la elección de un método de permutación para hallar un pseudo-F que determine la significancia estadística de la interacción de factores, como lo hace PERMANOVA y goza de las propiedades del método no paramétrico de Kruskal-Wallis (Conover, 1999). El algoritmo comprende: cálculo de la matriz de distancia y partición de la misma en bloques de distinto tipos de distancia, posterior cálculo de residuos, análisis de varianza no-paramétrico sobre esos residuos y cálculo del valor-p para determinar la significancia estadística. A través del análisis de varianza no-paramétrico (AMOVA No-Jerárquico) se evalúa la hipótesis de homogeneidad de varianzas dentro de los distintos bloques. El valor promedio de la variable de análisis (rangos de los valores

absolutos de los residuos) será el mismo para todos los bloques sólo cuando la variabilidad dentro de cada uno de los cuatro bloques sea la misma. Si los valores esperados en los bloques que contiene distancias "entre" (B2, B3, B4) no difieren significativamente de los valores esperados en el bloque B1 (que contiene distancias "dentro" de un grupo de individuos o entre individuos de un mismo grupo) entonces no existe evidencia para rechazar la hipótesis nula y suponer que hay interacción entre los factores.

La interacción se produce cuando las diferencias entre los niveles de un factor dentro de un mismo nivel de un segundo factor, son distintas a las diferencias obtenidas para otros niveles del segundo factor. Si esto ocurre (*i.e.* si hay interacción entre factores) los valores absolutos de los residuos provenientes de distancias que involucran perfiles de distintos niveles de uno o ambos factores (Bloques B2, B3 y B4), serán mayores que aquellos provenientes de distancias del bloque B1, donde sólo se estima la variabilidad residual y no la debida a la interacción. Para evitar supuestos distribucionales en la variable de análisis (valor absoluto de los residuos entre distancias observadas y distancias esperada sin interacción) se ajusta una prueba no paramétrica, basada en rango, por lo que el valor de significancia es obtenido como en Kruskal Wallis.

Warton *et al.* (2012) discuten que una propiedad crítica de los datos discretos es que la media y la varianza suelen estar relacionadas. Postulan que si la dispersión o variabilidad entre individuos, se define en función de los cambios que sufre la relación entre la media y la varianza, el efecto de la variabilidad genética puede confundirse con el efecto de la variabilidad subyacente. Para evitar ese efecto confundido, es importante utilizar una métrica de distancia entre individuos de diferente grupo o taxón que contemple apropiadamente la relación entre media y varianza. Dicho de otra manera, una consecuencia de la selección incorrecta de la métrica, es que en grupos de individuos con alta varianza, las diferencias entre grupos serán detectada con baja potencia. Este último resultado es indeseable si el objetivo es la búsqueda de estructura genética (Warton *et al.*, 2012). Por otro lado, Jost (2008) discutió que usar una medida de diferenciación genética como un estimador de divergencia corregido por el sesgo del muestreo, evita cualquier impacto que pueda tener la diversidad dentro de grupo para poder estimar la diferenciación entre grupos de poblaciones (Bird *et al.*, 2014). Si bien la métrica de Excoffier es ampliamente utilizada para estudios de variabilidad genética entre y dentro de grupos en el AMOVA Jerárquico como otros índices de similitud para datos binarios (Bruno *et al.*, 2003) pueden ser transformados a distancias para este tipo de análisis de la varianza molecular .

Nosotros usamos PERMANOVA con la distancia de Bray-Curtis (Anderson, 2001), una técnica ampliamente utilizada para evaluar simultáneamente la respuesta de abundancia en datos multivariados para uno o más factores (Andreson, 2001) para comparar el desempeño del método propuesto. Al igual que Warton *et al.* (2012) los valores-p calculados por PERMANOVA tendieron a ser más pequeños que con la prueba de interacción propuesta (mayor significancia) en conjuntos de datos donde la varianza entre individuos de distinto grupo es alta. Por el contrario, los valores-p que calcula PERMANOVA tienden a ser más grande, es decir, menos significativos, cuando el efecto entre grupos tiene menos variabilidad. Esto indica que la potencia de PERMANOVA para detectar diferencias entre grupos cuando la variabilidad entre ellos es menos variable, es más pequeña que cuando el efecto entre grupos presenta mayor variabilidad (Warton *et al.*, 2012). Con el método no paramétrico propuesto para la prueba de interacción en AMOVA No-Jerárquico, la potencia también disminuye conforme disminuye el efecto de la interacción, pero esta disminución es menor que la que se evidencia con PERMANOVA.

BIBLIOGRAFÍA

Anderson M.J. (2001) A new method for non-parametric multivariate analysis of variance. Austral Ecology 26:32-46.

Anderson M.J., ter Braak T. (2003) Permutation test for multi-factorial analysis of variance. Journal of Statistical Computation and Simulation. 73 (2): 85-113.

Apostol B.L., Black W.C., Miller B.R., Reiter P., Beaty B.J. (1993) Estimation of the number of full sibling families at an oviposition site using RAPD-PCR markers: applications to the mosquito Aedes aegypti. Theoretical and Applied Genetics 86: 991-1000.

Balzarini M.G., Di Rienzo J.A. Info-Gen versión 2018. FCA, Universidad Nacional de Córdoba, Argentina. URL http://www.info-gen.com.ar

Bird C.E, Karl S.A., Smouse P.E., Toonen R.J. (2014). Detectin and measuring genetic differentiation. Cust Issues 19:31-39.

Bray J.R., Curtis J.L. (1957) An ordination of the upland forest communities of southern Wiscosin. Ecol. Monogr. 27: 325-349.

Bruno C, Balzarini M., Di Rienzo J. (2003). Comparación de Medidas de Distancia entre Perfiles RAPD individuales. Journal of Basic & Applied Genetics, 15 (2):69-78. ISSN BAG-1666-0390

Conover W.J. (1999) Practical nonparametric statistics, 3rd edition. New York: John Wiley & Sons.

Dyer R. (2017) Applied Population Genetics. Disponible en https://dyerlab.github.io/applied_population_genetics/index.html#interactive-content

Excoffier L., Smouse P.E., Quattro J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479-491.

Gower J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53 (3 and 4): 325-338.

Hedrick P. (2005) Genetic of Populations. Arizona State University. Jones and Bartlett publishers. Third Edition.

Kennington W.J., Gockel J., Partridge L. (2003) Testing for Asymmetrical Gene Flow in a Drosophila melanogaster Body-Size Cline. Genetics 165: 667-673.

Li W.H. (1976) A mixed model of mutation for electrophoretic identity of proteins within and between populations. Genetics 83:423-432

Jost L. (2008). G_{ST} and its relatives do not measure differentiation. Mol Ecol. 17: 4015-4026.

Nei M. (1973) The theory and estimation of genetic distance. In: MORTON, N. (ed.), Genetic structure of Populations University of Hawaii, Honolulu, pp. 45-54.

Pritchard J.K., Stephens M., Donnelly P. (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155: 945-959.

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Warton D.I., Wright S.T., Wang Y. (2012) Distance-based multivariate analyses confound location and dispersion effect. Methods in Ecology and Evolution 3 :89-101.