UNIVERSIDAD DE SALAMANCA

DEPARTAMENTO DE ESTADÍSTICA

TESIS DOCTORAL

DIVERSIDAD GENÉTICA EN BANCOS DE GERMOPLASMA: UN ENFOQUE BIPLOT

JHONNY RAFAEL DEMEY

2008

DIVERSIDAD GENÉTICA EN BANCOS DE GERMOPLASMA: UN ENFOQUE BIPLOT

Memoria que para optar al Grado de Doctor, por el Departamento de Estadística de la Universidad de Salamanca, presenta:

Jhonny Rafael Demey

Salamanca, España

2008

VNiVERSiDAD
D SALAMANCA

Departamento de Estadística

JOSÉ LUIS VICENTE-VILLARDÓN

Profesor Titular del Departamento de Estadística
de la Universidad de Salamanca

MARÍA PURIFICACIÓN GALINDO-VILLARDÓN

Profesora Titular del Departamento de Estadística
de la Universidad de Salamanca

CERTIFICAN: Que **D^{n} Jhonny Rafael Demey**, Magíster en Estadística, ha realizado en el Departamento de Estadística de la Universidad de Salamanca, bajo su dirección, el trabajo que para optar al Grado de Doctor, presenta con el título: "***Diversidad genética en bancos de germoplasma: Un enfoque Biplot***"; y para que conste, firman el presente certificado en Salamanca, en Noviembre de 2008.

A:

JOHA y JOHN mis enanos, las estrellas que iluminan mi camino y fuente de inspiración.

YUSA mi compañera, por sus lecciones permanentes de coraje y valor frente a la adversidad.

AQUILES† por su amistad.

AGRADECIMIENTOS

A mis directores el **Dr. José Luis Vicente-Villardón** y la **Dra. María Purificación Galindo-Villardón** por su guía y apoyo, pero sobre todo por su disposición permanente a compartir sus conocimientos y experiencias tanto en el ámbito académico como profesional.

A la **Dra. Laura E. Pla**, mi maestra y guía, a quien le debo mi amor por la Biometría, gracias por abrirme espacios, por llevarme siempre de la mano y sobre todo por estar siempre para ayudarme.

Al **Dr. Emilio A. Carbonell**, por su apoyo incondicional a mi formación, por compartir su experiencia y sobre todo por brindarme uno de mis mayores tesoros que es su amistad.

A los Profesores **Raúl Macchiavelli**, **Julio Di Rienzo**, **Mónica Balzarini** y **Fernando Casanoves**, por las ideas que han aportado para enriquecer el trabajo y sobre todo por su apoyo solidario y su amistad.

Al **Proyecto de Biotecnología BID-FONACIT II**, por el financiamiento parcial de mis estudios en la Universidad de Salamanca y especialmente a la **Dra. Ariadne Vegas**, quien apoyó mi postulación y libró los obstáculos que me ayudaron a obtener el financiamiento.

A la **Dra. Asia Yusely Zambrano**, por su orientación en los aspectos genéticos de este trabajo, por sus críticas oportunas y sobre todo por ser una fuente de inspiración y ejemplo incansable de amor por el estudio y el trabajo, aun en los momentos más difíciles.

INDICE GENERAL

INDICE DE TABLAS

Página

INDICE DE FIGURAS

Página

Introducción

INTRODUCCION

En la actualidad es común ver reflejado en los medios de comunicación reportajes sobre lo que se considera una de las situaciones más críticas y sin precedentes de los últimos tiempos: "*el hambre y la malnutrición mundial*". En cualquiera de los innumerables estudios que se realizan, la desinversión en agricultura en los países pobres, los subsidios que distorsionan la producción en los ricos, las subvenciones a los biocarburantes y el cambio climático son los factores a los que se les otorga mayor importancia. Sin embargo, el tema de la soberanía alimentaria de los países pobres puede ser abordado desde un punto de vista estratégico a través de la producción de alimentos autóctonos que en condiciones de bajo subsidio energético y en pequeños sistemas de producción garanticen el acceso de las familias del tercer mundo a una alimentación digna y nutricionalmente aceptable. Por esta razón, la preservación de los recursos fitogenéticos entendidos éstos como el material hereditario con valor económico, científico o social contenido en las especies, es de importancia capital en la lucha contra "*el hambre y la malnutrición mundial*". Es aquí donde juegan un papel importante los bancos de genes o germoplasma.

Los bancos de germoplasma resguardan la fuente de variabilidad requerida por los mejoradores de plantas para el desarrollo de cultivares que permitan al agricultor superar las limitaciones naturales a fin de obtener mayores beneficios de su actividad, así como asegurar la fuente contra la erosión genética (Beeching *et al.*, 1994). Los

estudios de la diversidad genética dentro de estos bancos de genes son una de las herramientas que ayudan a tener un control más efectivo sobre la erosión genética. Permiten definir los patrones de variación que determinan la incorporación de individuos a programas de mejoramiento genético, ya sea por sus características promisorias o por susceptibilidad a condiciones bióticas o abióticas, facilitando la incorporación de genes y el establecimiento de la mejor estrategia reproductiva. Es así como, la cuantificación de la variación en general y especialmente dentro de los bancos de germoplasma debe definirse como un complejo que está asociado a un conjunto de caracteres de diferente naturaleza, escalas de medición y resolución.

No obstante, la naturaleza compleja de los estudios de diversidad y específicamente de la caracterización de los genotipos dentro de los bancos de germoplasma, es común observar cómo la mayoría de los trabajos publicados sobre el tema en las principales revistas especializadas solo se limitan a la descripción de las variables o marcadores y a la construcción de grupos. No se contempla el estudio de la naturaleza de las interrelaciones entre genotipos, genotipos-marcadores y marcadores-marcadores. Además, se aprecia una aplicación repetitiva de métodos estadísticos, sin que se explique por qué se han elegido y sin proporcionar detalles de su relación con la naturaleza y propiedades de los datos valorados para cada estudio, empleando rutinariamente las mismas técnicas estadísticas independientemente del tipo de datos.

Bajo estas premisas el presente trabajo tiene como objetivos:

General

Proponer una metodología alternativa para el estudio de la diversidad genética de bancos de germoplasma que permita una mejor comprensión e interpretación de las relaciones entre genotipos, genotipos-marcadores y marcadores-marcadores a través de la aplicación de técnicas de ordenación e integración de subespacios basados en la metodología Biplot.

Específicos

1. *Proponer una metodología para cuantificar la sensibilidad del método de Análisis de Coordenadas Principales (ACoP), a través del estudio de la variabilidad y la calidad de representación de individuos y grupos.*

2. *Proponer una metodología alternativa para la clasificación de genotipos utilizando marcadores moleculares basada en la aplicación de los métodos Biplot que, proyectando las variables sobre los subespacios generados por el método de Análisis de Coordenadas Principales (ACoP), permita el estudio de las relaciones entre genotipos, genotipos-marcadores y marcadores-marcadores.*

3. *Proponer una metodología alternativa basada en la aplicación de los métodos Biplot que permita la proyección de las variables responsables de la definición de grupos homogéneos derivados de métodos de integración de subespacios utilizando marcadores agromorfológicos y moleculares.*

INTRODUCCION

Se ha organizado el trabajo en cuatro capítulos de la forma siguiente:

El primer capítulo comprende una revisión de los estudios de diversidad genética en bancos de germoplasma a través de marcadores agromorfológicos y moleculares. Se describen los diferentes marcadores, su interpretación genética, las medidas de similitud o disimilitud y los métodos de agrupamiento y ordenación comúnmente utilizados.

En el segundo capítulo, se abordan dos métodos: el Análisis de Coordenadas Principales (ACoP) y el Análisis Biplot. Se reivindica el uso del ACoP frente a los estudios de diversidad genética, introduciendo medidas de variabilidad y calidad de la representación de individuos y grupos. En el Análisis Biplot se muestra su aplicabilidad en el estudio de las relaciones entre individuos y variables en aspectos como: facilidad de interpretación, riqueza del análisis y utilidad, ya permite responder las preguntas que no tienen respuesta en los análisis clásicos.

En el tercer capítulo, se demuestra la utilidad del Biplot Logístico en la clasificación de genotipos utilizando marcadores moleculares codificados como variables binarias, detallando su geometría, propiedades, interpretación, medidas de la calidad de representación y se demuestra que la complementariedad entre el Análisis de Coordenadas Principales (ACoP), el Análisis de Conglomerados (AC) y el Biplot Logístico produce una comprensión holística de la estructura de datos y facilita la interpretación de los resultados.

En el cuarto capítulo, se estudia la relación entre caracterización agromorfológica y molecular a través de su integración utilizando el Análisis de Procrustes Generalizado (APG) y se proyectan las variables responsables de la configuración consenso a través del ajuste de la representación Biplot basada en las rotaciones Procrustes.

Los cálculos, simulaciones y representaciones gráficas se ha utilizado InfoStat (InfoStat, 2008), MatLab versión 2008a (The MathWorks Inc, 2008) y R (R Development Core Team, 2008). Además se han desarrollado un conjunto de rutinas bajo MatLab versión 2008a (The MathWorks Inc, 2008) y R (R Development Core Team, 2008) que pueden ser obtenidas a través de http://www.biplot.usal.es.

Nota: Para separar las cifras enteras de los decimales se ha utilizado el punto en lugar de la coma, unificando la notación española con la anglosajona. Regla sobre usos no lingüísticos del punto, Ortografía de la lengua española (Real Academia Española, 1999).

Capítulo I

DIVERSIDAD GENÉTICA EN BANCOS DE GERMOPLASMA

CAPITULO I

1.1 ANALISIS DE LA DIVERSIDAD GENETICA

La fuente de variación genética de las plantas se encuentra en el conjunto de genes que ellas poseen y el espectro de esta variabilidad dentro de especies cultivadas y sus silvestres relacionadas, es comúnmente mantenido en bancos de germoplasma. La importancia del mantenimiento de estos recursos está en la medición y caracterización de dicha diversidad (Cordeiro *et al.*, 2003), y la efectividad en la exploración de ésta varía con el tipo de carácter evaluado, así como por rasgos de naturaleza biométrica, los cuales están codificados por un gran número de genes distribuidos en el genoma, permitiendo explorar mejor la variabilidad que en aquellos de herencia mendeliana. Sin embargo, es complejo, tedioso e impreciso el reconocimiento y enumeración de los diversos genotipos en cada uno de los loci que definen este tipo de caracteres, debido a que la variación se manifiesta como diferencias imperceptibles, y a la presencia de los efectos pleiotrópicos y epistáticos, e inclusive al mismo control poligénico (Bretting y Widrelechner, 1995).

Los bancos de germoplasma tienen como objetivo preservar la diversidad de los recursos fitogenéticos de las especies cultivadas y sus especies relacionadas y corregir la uniformidad derivada de las prácticas de mejoramiento genético que han reducido la base genética de los cultivos y que causan la indefensión de la poblaciones ante el ataque de patógenos para el que no existe resistencia (Martín, 2002).

CAPITULO I

El término germoplasma se refiere al material que se conserva como semillas, cultivo de tejido o plantas establecidas en colecciones de campo que reúne la variabilidad genética intra-específica de los materiales genéticos que pueden perpetuar una especie o una población de un organismo (Graur y Wen-Hsiung, 2000). Además de las funciones de conservación y mantenimiento, los bancos de germoplasma tienen un papel importante en la gerencia de los recursos fitogenéticos ya que su propósito no solo se limita a la conservación de especies sino que además incluye funciones tales como la documentación, caracterización, evaluación de la variabilidad genética, estudios filogenéticos y lo más importante como es el mejoramiento de caracteres deseables y la multiplicación y distribución del germoplasma (Graur y Wen-Hsiung, 2000). La obtención de caracteres deseables y su mejoramiento demanda un conocimiento apropiado de la diversidad genética del germoplasma.

La diversidad genética se puede definir como el grado en el cual el material hereditario diferencia internamente a una colección de plantas. El material hereditario de una planta abarca su DNA genómico y citoplasmático. El material hereditario puede diferenciarse en el nivel de las secuencias del DNA (alelos) y en el nivel de las combinaciones de alelos (genotipos) (Avise, 2004). La diversidad genética tiene una estructura multidimensional compleja y se basa en la semejanza entre pares de individuos valorada a través de caracteres que son compartidos -similitud genética-. Como resultado de la asociación considerable entre los caracteres en colecciones de plantas, es posible describir la estructura de la diversidad genética a un cierto grado describiendo estas colecciones y sus relaciones.

CAPITULO I

El estudio de la diversidad genética entre el germoplasma de una especie tiene importantes aplicaciones tales como: identificación de líneas o poblaciones que deben ser mantenidas para preservar el máximo de la diversidad genética en el banco de genes; relación genética entre líneas o poblaciones -de gran utilidad en la toma de decisiones sobre qué individuos usar para hacer nuevas combinaciones genéticas, contribuyendo por ende a maximizar la respuesta heterótica- (Beeching *et al.,* 1994 y Chavarriaga-Aguirre *et al.,* 1999); el estudio completo del pool de genes de los cultivos (Smartt, 1981, Porter *et al.,* 2005); y la diversidad específica del pool de genes (Prakash *et al.,* 2005). Los grupos que resultan de las descripciones basadas en los diferentes marcadores permiten el estudio de la asociación entre los caracteres o la estructura multi-locus de los grupos (Porter *et al.,* 2005).

Como ha sido mencionado, el estudio de la diversidad genética se basa en el grado de similitud entre individuos lo que permite la formación de grupos homogéneos que comparten un patrón o una estructura de diversidad particular. En este sentido, para medir la diversidad genética, se han empleado dos enfoques básicos el agrupamiento basado en los datos del pedigrí y el basado en marcadores genéticos (Avise, 2004).

Las clasificaciones con datos de pedigrí se basan en el grado de co-ascendencia de dos individuos o la probabilidad de que un alelo de un locus en un individuo sea idéntico por descendencia a otro alelo del mismo locus pero de otro individuo, se cuantifica generalmente a través del coeficiente de parentesco definido por Kempthorne (1969).

Sin embargo, este enfoque tiene la limitación que requiere que el pedigrí de material estudiado sea conocido, lo que raramente sucede.

Para el agrupamiento basado en marcadores genéticos se han empleado diferentes tipos de marcadores. La descripción morfológica de órganos vegetativos y reproductivos y rasgos agronómicos clásicos (descripción fenotípica) ha sido de gran utilidad para la caracterización y evaluación de recursos genéticos. Adicionalmente se consideran datos sobre susceptibilidad a factores de estrés, patógenos y enfermedades. Estos marcadores agromorfológicos pueden ser definidos como atributos de las plantas fácilmente cuantificables e identificables, los cuales pueden o no ser altamente heredables y estar controlados por uno o pocos genes, lo que permite una discriminación rápida de fenotipos (Lowe *et al.*, 1996). No obstante, presentan limitaciones que dificultan su medición y restringen la información genética recuperable, como son los efectos pleiotrópicos, desconocimiento de su base genética, tipo de herencia y su alta susceptibilidad a la influencia del medio ambiente como es el caso de los caracteres de interés agronómico (Pan *et al.*, 2004). Este último aspecto conocido como la variación adaptativa es cuantificada a través de ensayos regionales donde se evalúa la respuesta de distintos genotipos creciendo en las mismas condiciones, con lo que se minimiza la influencia ambiental; a la vez, el ensayo se repite en diferentes localidades para determinar la variación de un mismo genotipo en distintos ambientes (interacción genotipo-ambiente). Estos ensayos se establecen para recoger información sobre parámetros genéticos como la heredabilidad o los coeficientes de variación genética aditiva; sin embargo, su aplicabilidad es solo en cultivos de importancia económica por

lo cual conseguir información para otras especies no comerciales es poco probable (Primack y Kang, 1989).

Otro tipo de marcadores son los asociados a características de la estructura y morfología de los cromosomas (marcadores citogenéticos), estos marcadores son de poco uso dada la complejidad y dificultad de medición (Islam-Faridi *et al.*, 2002).

En los últimos años, el empleo de marcadores producto de la utilización de técnicas bioquímicas y moleculares, ha permitido complementar la información obtenida utilizando caracteres agromorfológicos. Dentro de esas técnicas se encuentran los metabolitos secundarios, proteínas, marcadores de ADN y secuencias de ADN. A través del uso de los marcadores moleculares es posible estimar la diversidad genética neutral. Su evaluación es más compleja que los caracteres morfológicos pero la influencia ambiental es menor y permiten hacer comparaciones entre individuos de una misma especie, entre especies, establecer relaciones de paternidad y parentesco, relaciones filogenéticas y analizar procesos de migración y deriva genética en la poblaciones. Los marcadores moleculares pueden ser clasificados según el tipo de molécula utilizada y la técnica en: los basados en el análisis de proteínas -análisis isoenzimático, polimorfismo posicional de péptidos, y los basados en el análisis del ADN. De estos los más popularizados son los marcadores de ADN ya que permiten la recolección de gran cantidad de información genética porque al tener su origen en variaciones individuales en la secuencia común del ADN, cubren todo el genoma, posibilitan su evaluación en estadios muy tempranos a partir de muestras mínimas que no destruyen el individuo, no

presentan interacciones intergénicas, tienen mayor reproducibilidad y mayor control genético porque son de herencia simple (Avise, 2004).

Los marcadores de ADN se dividen en dos grandes grupos: los revelados mediante hibridación con sondas marcadas y los obtenidos mediante amplificación por PCR (Polymerase Chain Reaction). Los diferentes tipos de marcadores se distinguen por su capacidad de detectar polimorfismos en loci únicos o múltiples y son de tipo dominante o co-dominante **(**Avise, 2004; Garoia *et al.*, 2007) tal como se explicará en el apartado 1.4.2. Entre los más usados para caracterizar y evaluar la variabilidad genética existente en los bancos de germoplasma se encuentran los RFLP "Restriction Fragment Length Polymorphism" (Lu *et al*., 1994a, b; Besse *et al.,* 1997), RAPD "Randomly Amplified Polymorphic DNA" (Harvey y Botha, 1996; Burner *et al.,* 1997; Vijayan *et al.,* 1999), AFLP "Amplified Fragment Length Polymorphism" (Besse *et al.,* 1998; Xu *et al.,* 1999), SSR "Simple Sequence Repeat" (Cordeiro *et al.,* 2000, 2003; Da Silva, 2001), SNP "Single Nucleotide Polymorphism" (Grivet *et al.,* 2001; Cordeiro *et al.*, 2006)**,** TRAP "Target Region Amplification Polymorphism" (Alwala *et al*., 2006).

A diferencia de los marcadores de ADN, la utilización de marcadores proteicos está limitada por su reducido número que no cubre toda la extensión del genoma, por sus interacciones o modificaciones postranscripcionales, y por su diferente expresión en distintos tejidos **(**Avise, 2004). En el caso de las secuencias de ADN aunque su utilización se ha incrementado considerablemente por la resolución de la técnica, su uso

sigue siendo restringido ya que son muy pocas las especies a las que se les ha descrito el genoma (Hunter *et al*., 2004).

La aplicación de los marcadores de ADN a la evaluación de germoplasma ha facilitado la identificación de duplicados en los bancos, la clasificación de los materiales, el cálculo de la distancia genética entre entradas, la identificación de su origen geográfico y la determinación de puntos de máxima variabilidad. Esta información facilita el manejo de las colecciones de germoplasma ya que permite tanto elegir parentales donde buscar nuevos alelos para ampliar la base genética de sus materiales y como una explotación más adecuada de la heterosis (Lee, 1995; Avise, 2004).

El análisis de la diversidad genética dependerá del tipo de datos usados, podrán generarse clasificaciones entre otras; usando información basada en datos de pedigrí (si es conocido), marcadores agromorfológicos utilizando caracteres cualitativos o cuantitativos y marcadores moleculares. En cualquier caso, es posible esperar clasificaciones disimiles porque cada tipo de descriptor reflejará aspectos diferentes de la diversidad genética asociados al tipo de medición y a la resolución del marcador**.** Las divergencias entre los análisis basados en datos agromorfológicos y los basados en datos moleculares se sustentan en que los cambios agronómicos y/o morfológicos no siempre están asociados a variaciones moleculares ya que responden a reglas y presiones evolutivas diferentes, estas incongruencias han originado polémicas respecto a qué tipo de datos pueden proveer de información adecuada para el análisis de la diversidad genética. El principal argumento en favor de la utilización de caracteres

moleculares es que son universales, abundantemente informativos y trabajan directamente con la base genética de la variación. Los estudios que incluyen caracteres morfológicos, en cambio, no permiten establecer diferencias en fases tempranas, son poco informativos ya que la lista de descriptores utilizados en los análisis raramente excede los 100 caracteres y adicionalmente son arbitrarios porque su codificación no sigue ningún criterio, sino que más bien está influenciada por la habilidad o experticia del responsable de la clasificación. Por otro lado, los caracteres agromorfológicos son la base de muchas características varietales que tienen un valor económico directo indudable. Por tanto, en general, estudios que incorporen descriptores morfológicos y marcadores moleculares proveerán una mejor descripción e interpretación de la diversidad genética de los individuos (Wilson *et al.,* 1974, 1977; Hillis y Wiens, 2000; Demey *et al.,* 2003).

Independientemente del marcador utilizado, la estructura de diversidad se puede representar adecuadamente en un modelo jerárquico, y si la información sobre los individuos es suficiente, será posible representarla usando técnicas de agrupamiento y/o clasificación (van Hintum, 1995)**.** La literatura muestra que las técnicas más utilizadas para la caracterización agromorfológica, basada en variables cuantitativas, son el análisis de componentes principales y los dendrogramas, sobre matrices de distancias euclídeas. Para la caracterización usando marcadores moleculares tanto dominantes como codominantes, la técnica de clasificación mayormente utilizada es el árbol generado por el algoritmo UPGMA "Unweighted Pair Group Method with Arithmetic Mean" calculado sobre matrices de similaridad empleando los coeficientes de Jaccard o

Dice. En los trabajos donde se estudian las relaciones entre diferentes marcadores, la amplia mayoría realizan repetidamente correlaciones entre matrices de distancias y/o similaridades para cuantificar la concordancia entre caracterizaciones (ISI Web of Knowledge 2008). Prácticamente ninguno de los métodos revisados incluye información sobre las variables responsables de la clasificación (Hillis y Moritz, 1990; Powell *et al.*, 1996; Graur y Wen-Hsiung, 2000; Infante *et al.*, 2006).

1.2 DISTANCIAS SOBRE LAS MATRICES DE DATOS

La identificación y caracterización de los individuos en los bancos de germoplasma es la metodología que garantizará su gestión y función como fuente de diversidad genética. Ha sido mencionado que la diversidad genética tiene una estructura multidimensional compleja y se basa en la semejanza entre pares de individuos, valorada a través de caracteres que son compartidos. En cada banco de germoplasma es posible generar tablas o matrices rectangulares $\mathbf{X}$ de orden (nxp), donde n son las filas que corresponden a los individuos o genotipos (Unidades Taxonómicas Operativas, UsTO) y p las columnas serán los datos referidos al conjunto de caracteres o atributos que son medidos sobre cada individuo, x_{ij} denotará la medición en el individuo i-ésimo de la variable j-ésima. En algunos casos estas matrices pueden referirse a datos longitudinales, es decir, las filas observadas en t momentos diferentes, o en columnas de la misma variable registrada en t momentos. Adicionalmente, las matrices pueden describir estudios transversales donde los mismos individuos o filas son vistos en diferentes escenarios o conjunto de variables.

CAPITULO I

A partir de la matriz de datos $\mathbf{X}$ de orden (*n*x*p*) es posible generar tres tipos de matrices o formas generales que permiten medir las similaridades o distancias -asociación entre pares de individuos o UsTO- estas son: $\mathbf{S}_{nxn} = (s_{ij})$ matriz de similitud, $\mathbf{\Delta}_{nxn} = (\delta_{ij})$ matriz de disimilitud o distancia y $\mathbf{B}_{nxn} = (b_{ij})$ matriz de productos escalares.

Generalmente, las similaridades están acotadas en el rango cero a uno; un aumento de la similaridad implica un aumento de la semejanza entre unidades, y toda similaridad de una unidad consigo mismo debería ser igual al máximo valor posible, es decir, uno. Las distancias en cambio disminuyen con un aumento del parecido, usualmente no son negativas y la distancia de un elemento consigo mismo es cero. Tanto las similaridades como las distancias son simétricas, es decir, la distancia entre la *i*-ésima y *j*-ésima unidad es la misma, independientemente si se mide a partir de la unidad *i* o desde la unidad *j*. Las matrices $\mathbf{X'X}$ y $\mathbf{XX'}$ representan los productos escalares entre columnas y filas de la matriz $\mathbf{X}$ y son útiles entre otras para medir la variabilidad entre variables (matriz de varianza-covarianza) y la distancia entre individuos, respectivamente.

Dependiendo del método elegido para la ordenación o clasificación y la escala de medición, la asociación entre las UsTO se expresará en términos de similaridad o distancia. No obstante la elección de la forma de asociación, las similaridades pueden transformarse en distancias y viceversa. Para el rango cero-uno, la similaridad s_{ij} puede ser transformada a distancia de la siguiente forma: $\delta_{ij} = 1 - s_{ij}$, $\delta_{ij} = \sqrt{1 - s_{ij}}$ y

$\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$ para $s_{ij's} \neq 0$ es posible transformar la similitud en distancia usando: $\delta_{ij} = -\log\left(s_{ij}\right)$ y $\delta_{ij} = {}^{1}\!\!\diagup\!\!_{S_{ij}} - 1$.

Cuadras (1996), considera dos clases de distancias estadísticas entre individuos y poblaciones: (***i***) los individuos de una población Ω que son representados por una matriz de datos $\mathbf{X}$ de orden (*n*x*p*), donde *p* es el número de variables estadísticas (binarias o categóricas, cualitativas y cuantitativas) y *n* representa una muestra que proviene de una población que puede ser finita o infinita y la distancia $\delta_{ij} = \delta(i, j)$ entre dos individuos o elementos i, j de Ω, es una medida simétrica no negativa que cuantifica la diferencia entre ambos, en relación con la variables; (***ii***) los individuos de cada población están caracterizados por un vector aleatorio, que sigue una distribución de probabilidad $f(x_1,...x_p;\theta)$ y la distancia entre dos individuos i, j, caracterizados por los puntos x_1, x_j de $\mathbb{R}^p$, es una medida simétrica no negativa $\delta(x_i, x_j)$ que dependerá de θ. La distancia entre poblaciones será una medida de divergencia $\delta(\theta_1, \theta_2)$ entre los parámetros que la caracterizan. En cualquiera de los casos lo que interesa es representar el conjunto Ω con las distancias δ, es decir, (θ, δ) mediante un espacio geométrico modelo (V, *d*), donde V representa el espacio euclídeo y *d* es una distancia sobre V. Según sea el método de clasificación y/o ordenación utilizada, la distancia *d* puede ser euclídea, ultramétrica, aditiva, no euclídea, etc. y cumple algunas de las siguientes propiedades:

P.1 -->	$\delta_{ij} > 0$ si $i \neq j$	
P.2 -->	$\delta_{ij} = 0$ si $i = j$	(No negatividad)
P.3 -->	$\delta_{ij} = \delta_{ji}$	(Simetría)
P.4 -->	$\delta_{ij} \leq \delta_{it} + \delta_{jt}$	(Desigualdad triangular)
P.5 -->	δ_{ij} es euclídea	
P.6 -->	$\delta_{ij} \leq \max\left(\delta_{it}, \delta_{jt}\right)$	(Desigualdad ultramétrica)

Según las propiedades que verifiquen las distancias pueden ser calificadas como:

Calificación	**Propiedades**
Disimilaridad	P.1, P.2, P.3
Distancia métrica	P.1, P.2, P.3, P.4
Distancia euclídea	P.1, P.2, P.3, P.5
Distancia ultramétrica	P.1, P.2, P.3, P.6

Observaciones:

1) Toda disimilaridad verifica por lo menos las tres primeras propiedades.
2) $\delta_{ij} = 0 \Leftrightarrow i = j$.
3) Una distancia que es euclídea es también métrica.
4) La condición P.6 implica también P.4 y P.5.

El uso eficaz de los métodos de clasificación y/o de ordenación requiere una comprensión de las propiedades de estos datos x_{ij} –atributos medidos- sobre los individuos y de las medidas de semejanza asociadas a cada tipo de datos. El estudio de la diversidad genética requiere de la colección de datos de diferentes fuentes de

información: caracteres agronómicos, morfológicos, moleculares, etc., que a su vez se corresponden con diferentes formas de variables: binarias (presencia/ausencia), cualitativas (multinomiales y ordinales) y cuantitativas.

A continuación se presentan las diferentes medidas de similitud y distancia calculadas a partir de la matriz $\mathbf{X}$ para datos binarios, cualitativos, cuantitativos y su mezcla.

1.2.1 Datos binarios

Cuando la matriz $\mathbf{X}$ proviene de la observación de p atributos o caracteres cualitativos que se asocian a variables binarias que toman el valor 0 si la característica está ausente y el valor 1 si está presente, la información del grado de asociación entre cualquier par de individuos x_i y x_j puede representarse como una tabla de contingencia 2x2:

		Individuo *j*		
		Presente (1)	Ausente (0)	
Individuo *i*	Presente (1)	*a*	*b*	*a+b*
	Ausente (0)	*c*	*d*	*c+d*
		a+c	*b+d*	*p=a+b+c+d*

donde *a* es el número de caracteres presentes comunes, *b* es el número de caracteres presentes en *i* pero ausentes en *j*, *c* es el número de caracteres ausentes en *i* pero presentes en *j* y *d* en número de caracteres ausentes simultáneamente. Para la matriz $\mathbf{X}$ de orden (*n*x*p*) es posible construir $n(n-1)/2$ tablas de contingencia que definen la similitud entre los individuos en función de las frecuencias *a, b, c* y *d*.

$$S_{ij} = f(a,b,c,d)$$

tal que es creciente en *a,* decreciente y simétrica en *b* y en *c,* S_{ij} tomará igual valor cuando: (***i***) la *i*-ésima unidad está presente y la *j*-ésima ausente y (***ii***) la *i*-ésima unidad está ausente y la *j*-ésima presente. Claramente este es un requisito necesario y suficiente para que el coeficiente de similaridad sea simétrico, es decir, la similaridad entre las unidades x_i y x_j es la misma que la entre x_j y x_i. La mayoría de los coeficientes de similitud S_{ij} están acotados en el rango (0,1), es decir, S_{ij} valdrá 0 cuando todo carácter presente en x_i no está presente en x_j (disimilaridad total), y S_{ij} valdrá 1 cuando todo carácter presente en x_i está presente también en x_j (similaridad total).

Diversos coeficientes de similaridad que verifican estas propiedades han sido propuestos, entre otros Cuadras (1996) menciona a: Jaccard (1908); Rusell y Rao (1940); Sorensen (1948); Sokal y Michener (1958). Sin embargo, existen coeficientes que no verifican las propiedades de simetría y rango tales como el Kulczynski (1970) acotado en el rango (0,∞) y otros que expresan dependencia estocástica entre x_i y x_j como son los de Yule (1912) y el de Pearson (1926), acotados en el rango (-1,1), donde la mayor disimilaridad corresponde a -1, la similaridad total a 1 y el valor 0 se asocia a la independencia estocástica.

Independientemente de las propiedades ya mencionadas, los coeficientes de similaridad pueden ser clasificados en dos grupos: aquellos coeficientes donde tanto la ausencia

como la presencia simultánea del carácter contribuyen a la semejanza entre las unidades; y aquellos en que no se considera como motivo de aumento de la similaridad, la ausencia simultánea.

Cuadras (1996) señala que la utilización de los coeficientes donde tanto la ausencia como la presencia simultánea del carácter contribuyen a la semejanza entre las unidades, es decir, donde aparece *d* en el numerador de S_{ij} puede ocasionar problemas ya que al añadir caracteres arbitrarios no comunes, podrían hacerse falsamente similares individuos que no los son. En estos casos Gower (1971a y b) propone hacer una distinción entre datos binarios, llamando 'dicotómicos' a aquellos en los que la ausencia simultánea del carácter no contribuye a la similitud, reservando el término de datos 'alternativos' en aquellos casos donde la presencia o ausencia de la variable binaria se refieren a dos niveles de una variable cualitativa, situación en la que si tiene importancia tener en cuenta que el carácter no esté presente en dos individuos.

No existe un criterio universal de cuando usar uno u otro coeficiente de similitud, los diferentes autores que han abordado el tema coinciden que la elección de un determinado coeficiente dependerá del peso que se desea dar a las frecuencias de *a*, *b*, *c* y *d*, el tipo de datos que se quieran representar y la situación experimental (Legèndre y Legèndre, 1979; Gower y Legèndre, 1986). En el caso de estudios de la diversidad genética, los coeficientes de similitud para datos binarios son empleados para representar los datos provenientes de marcadores bioquímicos y moleculares. Su uso e interpretación serán discutidos posteriormente.

Una vez definido el coeficiente de similitud, es posible construir la matriz simétrica $\mathbf{S}_{nxn} = (s_{ij})$ que representa la similaridad entre individuos.

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{21} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

También es posible generar $\mathbf{S}_{nxn} = (s_{ij})$ operando la matriz de productos escalares entre filas de la matriz $\mathbf{X}$. Es así como los coeficientes Russel y Rao (1940) y Emparejamiento Simple (Sokal y Michener, 1958) pueden ser expresados como: $\mathbf{S}_{nxn} = \left(\mathbf{X}\mathbf{X}'\right)/p$, $\mathbf{S}_{nxn} = \left[\mathbf{X}\mathbf{X}' + \left(\mathbf{J}-\mathbf{X}\right)\left(\mathbf{J}-\mathbf{X}\right)'\right]\Big/p$, respectivamente, siendo $\mathbf{J}$ matriz de orden *n*x*n* cuyos elementos son todos iguales a 1. Sin embargo, la operación con productos escalares debe ser cuidadosa porque provoca que se realicen análisis no acordes con la naturaleza categórica de los datos de la matriz $\mathbf{X}$.

Si se desea, como es el caso de los estudios de diversidad genética, representar los individuos en un espacio euclídeo o clasificarlos, utilizando alguna técnica de ordenación o clasificación jerárquica, respectivamente, la matriz $\mathbf{S}_{nxn} = (s_{ij})$ debe ser semidefinida o definida positiva y debe verificar (aproximadamente) la propiedad de desigualdad ultramétrica.

Recordemos que para el rango cero-uno, la similaridad s_{ij} puede ser transformada a distancia entre otras formas como: $\delta_{ij} = 1 - s_{ij}$, $\delta_{ij} = \sqrt{1 - s_{ij}}$ y $\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$, sin embargo, para la mayor parte de similaridades utilizadas, Gower (1966) y Cuadras (1996) consideran más aconsejable utilizar $\delta_{ij} = \sqrt{1 - s_{ij}}$ y $\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$, ya que estas expresiones aplicadas sobre matrices de similitud dan lugar a una distancia métrica, incluso euclídea. Que una distancia sea métrica implica que es posible construir, para toda terna de objetos *i*, *j*, *t*, un triángulo con lados igual a δ_{ij}, δ_{it} y δ_{jt} que satisfacen $\delta_{ij} \leq \delta_{it} + \delta_{jt}$, propiedad de la desigualdad triangular. Una matriz de distancias es euclídea si todas las distancias reales pueden representarse como distancias de líneas rectas entre un conjunto de puntos en un espacio real, es decir, $\mathbf{\Delta}_{nxn} = (\delta_{ij})$ la matriz de distancias será euclídea *p*-dimensional, si existen *n* puntos $x_1', x_2', ..., x_n'$ en un espacio $\mathbb{R}^p$ tal que: $\delta_{ij}^2 = \left(x_i - x_j\right)' \left(x_i - x_j\right)$. Operando la matriz de distancia $\mathbf{\Delta}_{nxn} = (\delta_{ij})$ es posible convertirla en una matriz de productos escalares tomando

$$\mathbf{B} = -\frac{1}{2}\mathbf{H}\mathbf{\Delta}^2\mathbf{H}' \qquad [1.1]$$

donde $\mathbf{H}_{nxn}$ es la matriz de centrado:

$$\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}' \qquad [1.2]$$

entonces si $\mathbf{\Delta}_{nxn} = (\delta_{ij})$ es una matriz de distancias y consideramos $\mathbf{B}$ como ha sido definida, $\mathbf{\Delta}_{nxn} = (\delta_{ij})$ será euclídea si y solo si $\mathbf{B}$ es semidefinida positiva.

Si $\mathbf{S}_{nxn} = (s_{ij})$ es una matriz semidefinida positiva, entonces δ_{ij} es euclídea y por lo tanto podremos representar $\left(\Omega, \delta_{ij}\right)$ a través del espacio euclídeo, es decir, si un conjunto de distancias entre *n* unidades es Euclídea, como máximo serán necesarias (*n*-1) dimensiones para representarlas, detalles de la demostración pueden ser consultados en Mardia *et al.* (1979). Esta propiedad es particularmente importante y es la base fundamental en el Análisis de Coordenadas Principales (ACoP) -el método de ordenación más utilizado en los estudios de diversidad genética.

Otra característica deseable de la matriz de disimilitud o distancias es que debe verificar la propiedad de desigualdad ultramétrica. Sin embargo, difícilmente las distancias calculadas a partir de información de datos reales satisfacen esta condición restrictiva *in extremis*, salvo en situaciones o conjuntos de datos particulares donde las distancias entre objetos de una terna particular son tales que, entre sí conforman un triángulo isósceles, siendo la base el lado de longitud menor. Cuando se pretende generar una clasificación basada en métodos jerárquicos es necesario que la matriz de distancias verifique aproximadamente la propiedad de desigualdad ultramétrica. Como ninguna de las matrices generadas con los coeficientes de similitud comúnmente utilizados cumple esta propiedad, los algoritmos de encadenamiento que generan clasificaciones jerárquicas se inician transformando 'razonablemente' la disimilaridad

inicial para convertirla en ultramétrica, y seguidamente luego construir la jerarquía indexada. Por esta razón, la representación de las relaciones entre los objetos generada por la mayoría de estos métodos de clasificación no es exacta.

En la Tabla 1, se presenta la formulación y propiedades de los coeficientes de similaridad más utilizados, en los estudios de diversidad. En orden proporcional decreciente son: el coeficiente de Dice, Jaccard, Emparejamiento Simple y Rogers y Tanimoto. En los dos primeros no se considera como motivo de aumento de la similaridad, la ausencia simultánea y en los dos últimos se consideran a y d simétricas. Existe un conjunto grande de coeficientes de similaridad derivados de los casos clásicos que se muestran; sin embargo, las diferencias entre unos y otros no son relevantes. Una lista extensa de coeficientes puede ser consultada en Sneath y Sokal (1973), Hubálek (1982) y Gower (1985).

Tabla 1. Propiedades de algunos coeficientes de similaridad para variables binarias[1]

Coeficientes de similaridad[2]		**Simetría entre *a* y *d***	**Rango**	**S ≥ 0** [3,4]	**Métrica**[5]	**Euclídea**[6]
Emparejamiento Simple (Sokal y Michener, 1958)	$\frac{a+d}{a+b+c+d}$	Si	0,1	Si	Si	Si
Rogers y Tanimoto (1960)	$\frac{a+d}{a+2b+2c+d}$	Si	0,1	Si	Si	Si
Hamman (1961)	$\frac{(a+d)-(b+c)}{a+b+c+d}$	Si	-1,1	Si	Si	Si
Yule (1912)	$\frac{ad-bc}{ad+bc}$	Si	-1,1	No	No	No
Pearson (1926)	$\frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$	Si	-1,1	Si	Si	Si
Jaccard (1908)	$\frac{a}{a+b+c}$	No	0,1	Si	Si	Si
Kulczynski (1927)	$\frac{a}{b+c}$	No	0,∞	Si	Indefinida	Indefinida
Russel y Rao (1940)	$\frac{a}{a+b+c+d}$	No	0,1	Si	Si	Si
Dice (1945)	$\frac{2a}{2a+b+c}$	No	0,1	Si	Si	Si
Ochiai (1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	No	0,1	Si	Si	Si
Sokal y Sneath (1963)	$\frac{a}{a+2(b+c)}$	No	0,1	Si	Si	Si

1 Modificada de Cuadras (1996).
2 *a*, *b*, *c* y *d* son las frecuencias absolutas de los eventos (1,1), (1,0), (0,1) y (0,0), respectivamente.
3 $\mathbf{S} \geq 0$ la matriz de similaridades es semidefinida positiva.
4 Se puede verificar calculando los valores propios de la matriz de similaridad.
5 La propiedad métrica se refiere a la distancia $\delta_{ij} = \sqrt{1 - s_{ij}}$ y $\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$
6 La distancia δ_{ii} es euclídea.

1.2.2 Datos cuantitativos

Supongamos que sobre la matriz $\mathbf{X}$ se han observado 4 UsTO (**a**, **b**, **c** y **d**) y *2* variables aleatorias cuantitativas x_1 y x_2. La distancia que se observa entre el par de unidades x_i y x_j cuando se representan en el espacio de coordenadas $\mathbb{R}^2$ viene dada entre otras por: $\Delta^2_{a,c} = \left(x_{1,a} - x_{1,c}\right)^2 + \left(x_{2,a} - x_{2,c}\right)^2$ y puede ser representada por la Figura 1.

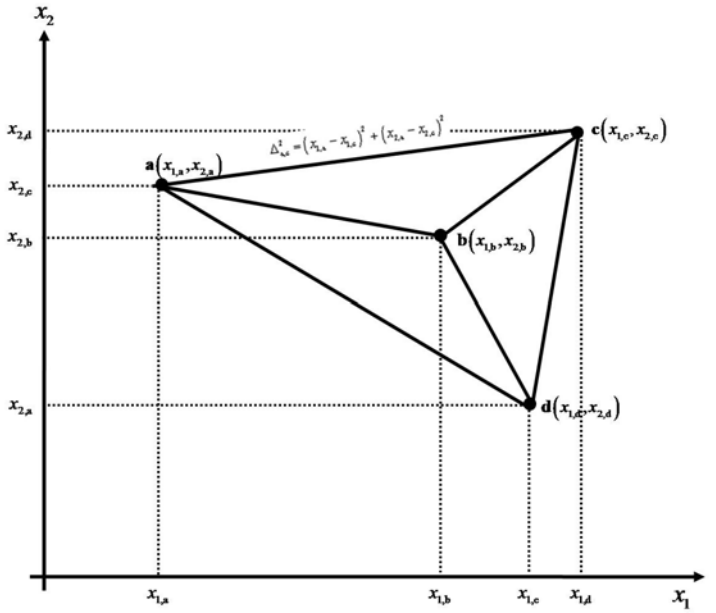


Figura 1. Representación de las 4 UsTO (**a**, **b**, **c** y **d**) como puntos sobre el plano determinado por las variables x_1 y x_2. Arbitrariamente fue asignado el orden **a< b<d<c y d<b<a<c** para la variables x_1 y x_2, respectivamente. Tomado de: Sneath y Sokal (1973).

Generalizando sobre la matriz $\mathbf{X}$ para *n* individuos y *p* variables aleatorias cuantitativas, la distancia usual que se observa entre el par de unidades x_i y x_j cuando

se representan en el espacio de coordenadas $\mathbb{R}^p$ dado por p variables cuantitativas, es conocida como distancia Euclídea, definida por:

$$\delta_{2(i,j)} = \sqrt{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^2} \qquad [1.3]$$

y representa un caso particular de la distancias de Minkowski, dada por:

$$\delta_{q(i,j)} = \sqrt[q]{\sum_{k=1}^{p}\left(x_{ik} - x_{jk}\right)^q} \qquad 1 < q < \infty \qquad [1.4]$$

que aunque verifican P.1, P.2, P.3 y P.4, no son distancias euclídeas, salvo para el caso de $q = 2$.

La distancia euclídea es la más conocida, la de mayor uso y es la herramienta fundamental de cálculo de la mayoría de los métodos multivariantes basados en distancias. Sin embargo, presenta varios inconvenientes: (***i***) no está acotada; (***ii***) es sensible a cambios de escalas y (***iii***) considera las p variables estocásticamente independientes (Cuadras, 1996).

Se han propuesto varias transformaciones que permiten minimizar y/o eliminar estos inconvenientes, entre otras: se recomienda utilizarla en caso de homogeneidad entre la naturaleza física de las variables, cuando esto no es posible se puede estandarizar cada

variable por su rango (r_k) asegurando que la contribución de cualquier variable estará acotada en el rango (0,1). Además puede dividirse por la cantidad de variables obteniendo una distancia media que oscilará en este rango y facilita su inversión a similaridad, la expresión estará definida por:

$$\delta_{2(i,j)} = \sqrt{\frac{1}{p}\sum_{k=1}^{p}\frac{\left(x_{ik} - x_{jk}\right)^2}{r_k^2}} \qquad [1.5]$$

Se han estudiado soluciones adicionales para corregir las imperfecciones de la distancia euclídea como la estandarización de cada variable por su desviación estándar o estandarización por media y desviación –también conocida como distancia euclídea normalizada o distancia de K Pearson. La distancia de Manhattan o métrica *city-block* o ciudad, calculada como la suma de las diferencias absolutas entre unidades para cada variable, es un caso particular de la distancias de Minkowski para $q = 1$. Es menos sensible a valores muy grandes o aberrantes, ya que es función de diferencias absolutas en lugar de diferencias al cuadrado, adicionalmente cada variable puede ser estandarizada por su rango (Cain y Harrison, 1958; Gower, 1971a). Otras distancias derivadas de la Manhattan son las de Bray y Curtis (1957) y la de Canberra (Lance y Williams, 1966).

El problema de la independencia aleatoria se resuelve introduciendo la distancia de Mahalanobis (Mahalanobis, 1936) que tiene en cuenta las correlaciones entre variables y por lo tanto la redundancia que existe entre las mismas; es una distancia general,

perfectamente adecuada para diferenciar individuos, grupos o poblaciones mediante variables aleatorias. Es invariante por transformaciones lineales no singulares de las variables, particularmente es invariante por cambios de escalas, siendo de gran utilidad cuando las variables son muy heterogéneas. (Cuadras, 1996; Digby y Kempton, 1991). Su expresión matricial es:

$$\mathbf{D}_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)' \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right) \qquad [1.6]$$

donde $\boldsymbol{\Sigma}$ es la matriz de varianzas-covarianzas entre las *p* variables. Si las *p* variables son independientes $\boldsymbol{\Sigma}$ sería diagonal y la distancia de Mahalanobis se aproximará a la distancia Euclídea al cuadrado con pesos inversos dados por las varianzas de las *p* variables.

Otra medida utilizada es el coeficiente de correlación de Pearson (1926). Su uso como coeficiente de similaridad en clasificación de genotipos (taxonomía numérica) data de finales de los años 50 y su complemento ha sido usado como una medida de distancia (Sneath y Sokal, 1973). Debido al rango (-1,1) presenta restricciones funcionales puesto que es improbable que para datos reales, en todos los caracteres estudiados existan altas correlaciones negativas entre UsTO, aunque sea posible para algunos. Su notoriedad surge, en parte, porque realiza un ajuste por el valor medio de la unidad ignorando diferencias en sus tamaños en conjunto. Sin embargo, este ajuste es cuestionable para medir distancia entre individuos a menos que todas las variables tengan la misma escala de medida. En general el coeficiente de correlación es usado para calcular la asociación

entre variables (caracteres) cuándo la mayoría, si no todas, presentan más de dos estados, esto hace al coeficiente de correlación muy apropiado para cuantificar las distancias existentes entre variables, es decir, entre columnas de la matriz $\mathbf{X}$.

En la Tabla 2, se presenta la formulación y propiedades de las distancias y disimilaridades no negativas más utilizadas en los estudios de diversidad. Los más comúnmente usados son las distancias: Euclídea, Manhattan y Mahalanobis. Una lista extensa de coeficientes puede ser consultada en Sneath y Sokal (1973) y Gower (1985).

Tabla 2. Propiedades de algunas distancias para variables cuantitativas

Distancias y disimilaridades		**Metrica**	**Euclídea**
Euclídea	$\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}$	Si	Si
Manhattan	$\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert$	Si	No
Bray-Curtis	$\frac{\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert}{\sum_{k=1}^{p}(x_{ik}+x_{jk})}$	Si	No
Canberra	$\sum_{k=1}^{p}\frac{\lvert x_{ik}-x_{jk}\rvert}{(x_{ik}+x_{jk})}$	Si	No
Minkowski	$\sqrt[q]{\sum_{k=1}^{p}\lvert x_{ik}-x_{jk}\rvert^q}$	Si	Si
Mahalanobis	$\sqrt{\sum_{l=1}^{p}\sum_{k=1}^{p}(x_{ij}-x_{jk})\sigma_{kl}^{-1}(x_{il}-x_{jl})}$	Si	Si

→ σ_{kl} elemento de la matriz $\boldsymbol{\Sigma}$ de varianzas-covarianzas entre las p variables.

→ q número entero.

1.2.3 Datos cualitativos

En el apartado 1.2.1 fueron presentadas las propiedades y los coeficientes de similitud para el caso particular de atributos o caracteres cualitativos con respuesta binaria. Supongamos ahora, que la matriz $\mathbf{X}$ proviene de la observación de p atributos o caracteres cualitativos que se asocian a variables del tipo multiestado tal como el color del tallo o tipo de yemas, entre otras. El grado de asociación entre cualquier par de individuos x_i y x_j puede medirse utilizando los coeficientes de similitud propuestos para variables binarias, si las variables cualitativas multiestados son reemplazadas por pseudo variables binarias, que toman el valor 0 si la característica está ausente y el valor 1 si está presente, para el caso del tipo de yemas se generarían cinco variables binarias para las formas: deltoide alargada/no; obconoidal/no; ovalada/no; pentagonal/no y redondeada/no. Sin embargo, esta propuesta metodológica de Sneath y Sokal (1973) tiene el inconveniente de ser artificial ya que tendrán mayor peso en número las variables que posean más multiestados (Digby y Kempton, 1991).

Si por el contrario las categorías para cada variable son codificadas por ejemplo, como *0, 1, 2, 3, …, k.*, el grado de asociación entre cualquier par de individuos x_i y x_j puede medirse a través de la expansión del emparejamiento simple (Gower, 1985) que se expresará como:

$$s_{ij} = \frac{número\,de\,caracateres\,coincidentes}{número\,total\,de\,caracteres}$$

No obstante cuando el cero representa ausencia del carácter es recomendable ignorar el empate de ceros en forma similar como lo hace el coeficiente de Jaccard. Nuevamente aquí también es posible seguir la metodología propuesta por Sneath y Sokal (1973) y reemplazar las variables cualitativas multiestados por pseudo variables binarias que toman el valor 0 si la característica está ausente y el valor 1 si está presente.

En estudios de diversidad genética es común que exista interés en comparar grupos de plantas o poblaciones respecto a caracteres o variables cualitativas, en estos casos es posible representar las variables como frecuencias observadas por grupo o población según sea el caso, es así como las variables pueden ser representas en tablas de contingencia según dos criterios de clasificación:

Grupos o Poblaciones	Caracteres				
	C_1	C_2	...	C_p	
x_1	f_{11}	f_{12}	...	f_{1p}	$f_{1.}$
x_2	f_{21}	f_{22}	...	f_{2p}	$f_{2.}$
⋮	⋮	⋮	...	⋮	⋮
x_n	f_{n1}	f_{n2}	...	f_{np}	$f_{n.}$
	$f_{.1}$	$f_{.2}$	...	$f_{.p}$	$f_{..}$

donde las filas corresponden a los grupos o poblaciones a ser comparadas y las columnas representan los niveles de la variable de interés, el cuerpo de la tabla contiene la frecuencia f_{jk} con que aparece el carácter *k* en el grupo o población *j*, tal que para un grupo o población cualesquiera la distribución de frecuencias de los caracteres viene dada por el vector perfil fila de coordenadas:

$$x_i : \left[\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, ..., \frac{f_{ip}}{f_{i.}} \right] \qquad i = 1, 2, ..., n$$

Por lo que la distancia entre dos grupos o poblaciones *i* y *j* será la euclídea entre los dos vectores perfiles filas correspondientes. Sin embargo, la distancia euclídea tiene el inconveniente que no elimina las distorsiones debidas a las frecuencias dispares entre caracteres, recomendándose para medir distancia entre dos grupos o poblaciones el empleo de la distancia Chi-cuadrado (Benzecri, 1970), dada por:

$$\mathrm{d}^2_{(\mathbf{x}_i, \mathbf{x}_j)} = \sum_{k=1}^{p} \left(\frac{f_{ik}}{\sqrt{f_{.k}}\, f_{i.}} - \frac{f_{jk}}{\sqrt{f_{.k}}\, f_{j.}} \right)^2 \qquad [1.7]$$

Es así como los grupos o poblaciones $\mathbf{x_1, ..., x_n}$ pueden ser representados por una configuración de *n* puntos en un espacio $\mathbb{R}^p$ y cuyas coordenadas estarán dadas por vectores perfil fila corregidos por la distribución de la frecuencias de los caracteres en la población x_i. La construcción de distancias sobre las frecuencias observadas sobre grupos o poblaciones y diferentes caracteres es la base del cálculo de las distancias genéticas.

1.2.4 Datos mixtos

Como ha sido descrito, supongamos que sobre los *n* individuos de la matriz $\mathbf{X}$ se han observado simultáneamente diferentes caracteres que por su naturaleza pueden corresponder a variables binarias, cualitativas, cuantitativas, la distancia entre individuos que presente esta combinación de caracteres puede ser medida a través del coeficiente de similaridad de Gower (1971a). Este coeficiente adaptado de forma general para tipos mixtos de variables además contempla las situaciones de valores faltantes.

Gower (1971a), propone un coeficiente de similaridad S_{ijk} entre el *i*-ésimo y *j*-ésimo individuo para cada *k*-ésima variable, el promedio de las S_{ijk} sobre las *k* variables es la medida de similaridad, cuya expresión es:

$$S_{ij} = \frac{1}{p}\sum_{k=1}^{p} S_{ijk} \qquad [1.8]$$

En el caso de variables binarias o cualitativas, $S_{ijk} = 0$ si $x_{ik} = x_{jk}$ y $S_{ijk} = 1$ si $x_{ik} \neq x_{jk}$. Si la variable es cuantitativa la similaridad entre los individuos estará dada por:

$$\mathrm{S}_{ijk} = 1 - \frac{\left| x_{ik} - x_{jk} \right|}{r_k} \qquad [1.9]$$

donde r_k es el rango (diferencia entre el máximo y el mínimo) del carácter *k*-ésimo sobre toda la población conocida o de la muestra, si $x_{ik} = x_{jk}$ entonces $S_{ijk} = 1$, cuando x_{ik} y x_{jk} se encuentran en los extremos de r_k entonces $S_{ijk} = 0$ y si x_{ik} y x_{jk} son valores intermedios de r_k entonces $0 < S_{ijk} < 1$, una fracción positiva del intervalo.

Usando este razonamiento es posible generar diferentes expresiones que cuantifiquen la similitud derivada del coeficiente de Gower, es así como el coeficiente de emparejamiento simple o de Jaccard para tratar las variables binarias puede expresarse como:

$$S_{ij(Emp.Simple)} = \frac{\sum_{k=1}^{p_1}\left(1 - \frac{\left|x_{ik} - x_{jk}\right|}{r_k}\right) + a + d + \alpha}{p_1 + p_2 + p_3} \qquad [1.10]$$

$$S_{ij(Jaccard)} = \frac{\sum_{k=1}^{p_1}\left(1 - \frac{\left|x_{ik} - x_{jk}\right|}{r_k}\right) + a + \alpha}{p_1 + \left(p_2 - d\right) + p_3} \qquad [1.11]$$

donde p_1 es el número de variables continuas, r_k rango de la *k*-ésima variable continua, p_2 número de variables binarias, *a* es el número de caracteres presentes comunes y *d* número de caracteres ausentes simultáneamente de las variables binarias, p_3 número de variables cualitativas y α número de coincidencias de las variables cualitativas.

El tratamiento de los datos faltantes, la consideración de la ausencia simultánea del carácter y la ponderación relativa de caracteres son debilidades metodológicas que afectan las medidas de similitud y distancias referidas hasta ahora y se destacan independientemente del tipo de datos. En este sentido, Gower introdujo en la ecuación [1.8] ponderaciones denominadas w_{ijk} como una función que depende de cada par de datos x_{ik} y x_{jk}, obteniendo la siguiente expresión general:

$$S_{ij} = \frac{\sum_{k=1}^{p} w_{ijk} S_{ijk}}{\sum_{k=1}^{p} w_{ijk}} \qquad [1.12]$$

Corrigiendo así las debilidades metodológicas mencionadas tal como sigue: (***i***) si x_{ik} o x_{jk} son faltantes solo bastará con tomar $w_{ijk} = 0$ en la expresión general para que los valores faltantes sean tratados individualmente sin necesidad de omitir toda la variable; (***ii***) si $x_{ik} = x_{jk} = 0$ tal como sucede en los coeficientes de Jaccard, Dice, Sokal y Sneath entre otros, donde no se considera la simetría entre *a* y *d*, se podrá ignorar el empate de doble cero haciendo $w_{ijk} = 0$; y (***iii***) con la forma general es posible considerar que ciertos caracteres son más importantes que otros asignándoles diferentes valores a w_{ijk}. La asignación de las ponderaciones dependerá de la situación experimental y del tipo de relaciones que se quieran establecer entre los individuos, tal como ocurre en las comparaciones de medias hechas *a posteriori*.

Como se deduce de la expresión general, algunos de los coeficientes de similaridad o de distancias entre pares de individuos que han sido descritos pueden ser considerados como un caso particular de la expresión general del coeficiente de similaridad de Gower.

Una alternativa adicional -poco conocida en el contexto del análisis de la diversidad genética- para el tratamiento conjunto de variables cuantitativas y cualitativas es la codificación de variables continuas propuesta por Escofier (1979). Usando un criterio similar al descrito previamente para variables cualitativas multiestado, Escofier (1979) propone transformar cada variable cuantitativa en dos nuevas variables $\frac{1-x_i}{2}$ y $\frac{1+x_i}{2}$, donde x_i representa el valor estandarizado de la variable para el individuo *i*. Nótese que aquí al igual que para una variable dicotómica, ésta queda definida por dos columnas, que para un mismo carácter siempre suma 1, para todos los individuos. Las variables cuantitativas codificadas pueden ser analizadas conjuntamente con las variables cualitativas binarias y cualitativas multiestado, utilizando cualquiera de los coeficientes de similitud discutidos previamente para este tipo de datos. Este método de discretización ha sido desarrollado con el fin de poder utilizar Análisis de Correspondencias Múltiples (ACM) bajo un conjunto mixto de variables. Tiene la ventaja de garantizar que las variables cuantitativas no sufran alteración o pérdida de información en el proceso de reestructuración de la matriz $\mathbf{X}$.

1.3 DISTANCIAS GENETICAS SOBRE LAS MATRICES DE DATOS

El concepto de distancia genética está referido exclusivamente a la medida de la diferencia genética entre las poblaciones y es útil para determinar qué tan disímil son éstas, respecto a su composición genética. Adicionalmente, es necesario considerar en la estimación de las distancia entre poblaciones la variación genética intra-poblacional que pueda existir, ya que variaciones altas dentro de la población podrán afectar la diferencia o la medida de distancia genética entre poblaciones (Weir, 1996; Hartl y Clark, 2006).

Basado en el principio de equilibrio de Hardy-Weinberg, la composición genética de una población no cambiará mientras no se produzcan alteraciones debidas a la selección natural, factores de stress o mutaciones. En otras palabras, bajo ciertas condiciones, tras una generación de apareamiento al azar, las frecuencias de los genotipos de un locus individual se fijarán en un valor de equilibro particular y éstas en función de las frecuencias alélicas en ese locus (Falconer y Mackay, 1996; Zintzaras, 2008).

Entonces, la distancia genética $\left(\delta_{ij(gen)}\right)$ entre dos poblaciones *i* y *j*, son distancias estadísticas que se calculan sobre datos basados en frecuencias alélicas de diferentes loci y formalmente, puede ser expresada como: "Dados *n* sucesos mutuamente excluyentes $A_1,...,A_n$, una distancia genética es una medida de divergencia entre dos distribuciones de probabilidad $p=\left(p_1,...,p_n\right)'$ y $q=\left(q_1,...,q_n\right)'$" (Cuadras, 1996), que deberá satisfacer las propiedades: (***i***) no negativa $\left(\delta_{ij(gen)} \geq 0 \ \forall \ i \neq j\right)$ y

$\left(\delta_{ij(gen)} = 0 \Leftrightarrow i = j\right)$; (***ii***) simétrica $\delta_{ij(gen)} = \delta_{ji(gen)}$ y (***iii***) métrica, es decir para tres poblaciones cualquiera *i*, *j*, *t* es posible construir un triangulo con lados igual a $\delta_{ij(gen)}$, $\delta_{it(gen)}$, $\delta_{jt(gen)}$ tal que se cumple $\delta_{ij(gen)} \leq \delta_{it(gen)} + \delta_{jt(gen)}$, desigualdad triangular. La primera condición implica inmediatamente que la definición de $\delta_{ij(gen)}$ debe considerar la variación genética intra-poblacional en cada población.

Aunque los métodos para detectar la diversidad genética hayan cambiado considerablemente desde el advenimiento de las técnicas moleculares, los estudios de distancias genéticas entre poblaciones preceden al descubrimiento de los marcadores genéticos. En este sentido, se ha propuesto una variedad de medidas para transformar frecuencias alélicas y genotípicas en datos de distancias genéticas y se han utilizado para una variedad de propósitos (Smith, 1977; Nei, 1978; Wright, 1984; Jorde, 1985; Lalouel, 1980; Chakraborty y Rao, 1991).

De acuerdo a su aplicación, las medidas de distancia genética pueden ser clasificadas en dos grandes grupos: (***i***) las usadas para la clasificación de las poblaciones y (***ii***) las utilizadas con el objeto de estudiar patrones evolutivos. En el primer grupo se incluyen las medidas de distancias geométricas debidas a: Bhattacharyya (1946), Cavalli-Sforza y Edwards (1967), Rogers (1972), Prevosti (Wright, 1978) y también algunas de las distancias mencionadas para datos cuantitativos como: la distancia de Manhattan, la de Mahanalobis o el coeficiente de Pearson, entre otras. En el segundo grupo se encuentran las medidas de distancias debidas a: Nei (1972 y 1978), Hillis (1984), Swofford y Olsen

(1990), los coeficientes de diferenciación genética, los índices de parentesco o los índices coancestrales, entre otros (Nei, 1987; Felsenstein, 1991; Weir, 1996; Nei y Kumar 2000).

Aunque esta clasificación tenga poca incidencia en la cuantificación de la distancia entre poblaciones, es importante resaltar que el uso de una medida estará asociado al conocimiento de la dinámica evolutiva. Las medidas de distancias ubicadas en el primer grupo no muestran un patrón o tendencia en el tiempo, en términos de mecanismos evolutivos definidos, tales como son las mutaciones, la deriva genética, etc. Por el contrario, los coeficientes e índices referidos en el segundo grupo se han estudiado en el contexto de modelos evolutivos específicos, de modo que es conocida su tendencia prevista con el tiempo, puesto que la divergencia entre las poblaciones está bien descrita (Chakraborty y Rao, 1991; Nei, 1987; Weir, 1996).

En el contexto del estudio de la diversidad genética en bancos de germoplasma se desconocen los modelos evolutivos específicos y en la mayoría de los casos se trabaja con poblaciones únicas, por lo que las medidas de distancias genéticas entre poblaciones son de uso más restringido. No obstante, nos referiremos brevemente a las más utilizadas y presentaremos un resumen con algunas propiedades importantes mostrando la relación entre las diferentes medidas, tal como se ha hecho para los casos de las distancias sobre matrices para los diferentes tipos de datos.

CAPITULO I

Antes de empezar a describir las distancias genéticas más utilizadas, es importante señalar que en muchas de ellas no pueden ser calificadas como distancias, sino como disimilaridades, ya que es muy raro que se cumpla la propiedad de la desigualdad triangular. Además, presentan la desventaja que las distancias pueden tender a infinito en los casos de discrepancia total entre poblaciones y son dependientes de la posición en el espacio de los vectores, lo que les da un carácter de medidas angulares.

La distancia de Bhattacharyya (1946), una distancia geométrica que define lo que se considera la medida precursora de la distancia entre dos poblaciones, está basada en el número de ocurrencias de *k* caracteres en comparación. Considera que un individuo en la población posee exactamente uno de los *g* caracteres, es decir, pertenece exactamente a una de las *k* clases que pueden ocurrir en una población, entonces ésta puede ser definida con respecto a los *k* caracteres por un vector $n = (n_1,...,n_k)$ tal que $\sum_{i=1}^{k} n_i = N$, dónde *N* es el tamaño de la población y *n* seguirá una distribución multinomial como consecuencia de una generalización de la distribución binomial para *k* clases (*k*>2). Estos vectores de ocurrencias pueden convertirse en distribuciones de frecuencias, como se refirió en la definición formal para el cálculo de la distancia entre dos poblaciones.

Para dos poblaciones con distribución multinomial y distribución de probabilidad dada por $(\pi_1,...,\pi_k)$ y $(\pi'_1,...,\pi'_k)$, respectivamente, donde $\sum \pi_i = \sum \pi'_i = 1$, estas distribuciones pueden ser representadas geométricamente como puntos en el espacio *k*-dimensional tomando $\left(\sqrt{\pi_1},...,\sqrt{\pi_k}\right)$ y $\left(\sqrt{\pi'_1},...,\sqrt{\pi'_k}\right)$, conformando dos líneas rectas

que atraviesan el origen, formando ángulos entre un punto y otro. Si Δ es el ángulo entre las dos líneas, entonces: $\cos\Delta = \sum_{j=1}^{k}\sqrt{\pi_i\pi_i'}$. El cuadrado del ángulo Δ que forman dos distribuciones de frecuencias, se puede considerar como una medida de divergencia entre las dos poblaciones, tal que:

$$\Delta^2 = \left(\cos^{-1}\left(\sum_{j=1}^{k}\sqrt{\pi_i\pi_i'}\right)\right)^2 \qquad [1.13]$$

Bhattacharyya (1946) también definió una distancia Δ' basada en proporciones para dos muestras, sin embargo demuestra que Δ'^2 no es un estimador insesgado de Δ^2, la distancia definida previamente entre pares de distribuciones poblacionales. Geométricamente interpreto a Δ' como distancia de la cuerda, acoplando los dos puntos de la muestra situados en la intersección de las dos líneas muestrales, con una hiperesfera de radio 1 en el espacio *k*-dimensional. Es claro que, aunque la medida de distancia no haya sido formulada en el contexto de la genética poblacional, si una población es definida por sus frecuencias alélicas para un locus dado o para un conjunto de loci, la analogía a la distribución multinomial es directa, por lo que esta medida es perfectamente aplicable.

Tal como Bhattacharyya, Cavalli-Sforza y Edwards (1967) formulan una medida de distancia geométrica pensando en las raíces cuadradas de la distribución de frecuencias como puntos en el espacio *k*-dimensional. Esta medida generalizada a *k* alelos considera

la distancia angular entre dos poblaciones de la misma manera que Bhattacharyya (1946). Sin embargo, para que la distancia pueda representar una sustitución del gen, transforman las unidades en radianes acotándola. Así mismo, para facilitar el estudio en cierto tipo de situaciones, y para evitar los inconvenientes de incluir la función arcocoseno en la expresión, Cavalli-Sforza y Edwards (1967) usan como medida de distancia, la longitud de la cuerda asociada al ángulo Δ. La transformación que incorpora tanto la distancia del arco como su asociada de la cuerda, estandariza las distancias con respecto a tendencias aleatorias, haciendo independiente la tasa de incremento en distancia genética de los valores iníciales de las frecuencias alélicas. Felsenstein (1985) demuestra que aunque tiene derivaciones diferentes al final las formulaciones de Bhattacharyya y Cavalli-Sforza y Edwards son muy similares.

Hemos querido referir las medidas de Cavalli-Sforza y Edwards en el contexto, o como derivación de las medidas de Bhattacharyya, que si bien se puede considerar como una de las precursoras de las medidas de distancia genéticas que entiende el problema de la correlación existente entre los vectores, no ha sido tomada en cuenta o ha permanecido oculta en la mayoría de las bibliografías genéticas.

Las distancias de Nei de 1972 y 1978, suponiendo que la tasa de sustitución de genes por locus es uniforme a lo largo de loci y linajes, cuantifica el número de sustituciones por locus que han tenido lugar después de la divergencia entre un par de poblaciones. Las dos versiones son muy similares, solo que en la segunda se corrigen los problemas de sesgos cuando se trabaja con tamaños de muestra pequeños y son las de mayor uso y

difusión en los programas informáticos especializados en estudios de población (Swofford y Olsen, 1990; Raymond y Rousset, 1995; Yeh y Boyle, 1997).

Hillis (1984) define una nueva distancia modificando las propuestas por Nei donde se considera la falta de uniformidad de la tasa de sustitución; consiste en calcular la distancia de Nei para cada locus y luego promediarla sobre todos los loci.

Swofford y Olsen (1990) corrigen el sesgo de la distancia de Hillis (1984) para aprovechar su uso cuando se trabaja con muestras pequeñas como sugiriera Nei (1978).

Rogers (1972) propone una medida basada en las distancias euclídeas entre los vectores de frecuencias alélicas por locus de las poblaciones, al igual que las de Nei tiene la desventaja de ser muy sensible a los niveles de homocigocidad dentro de las poblaciones, es decir la distancia entre dos poblaciones con alelos alternativos en homocigosis será siempre mayor que cuando cualesquiera de las dos o ambas poblaciones presentan elevada heterocigosis pero no tengan alelos en común. Similarmente Prevosti (Wright, 1984) desarrolla una medida análoga que consiste en hallar la media aritmética sobre la distancia de Manhattan calculada para cada locus.

Otras medidas de similitud y/o distancia genética entre poblaciones han sido implementadas sobre los vectores de frecuencias alélicas por locus/loci, entre las cuales podemos mencionar: Latter (1973a y b), Reynolds *et al.* (1983), Bhattacharyya y Nei (1987), la proporción de alelos compartidos (1994), Goldstein *et al.* (1995) y Shriver *et*

al. (1995) (Bowcock *et al.*, 1994; Goldstein *et al.*, 1995; Shriver *et al.*, 1995). Se ha incrementado su uso en la medida que los desarrollos informáticos las han puesto a disposición de los usuarios. Si bien las definiciones de estas distancias se basan en diversas premisas, la mayoría están analíticamente relacionadas, incluso las que son matemáticamente bastante disímiles con respecto a relaciones entre las poblaciones, por lo menos para el caso de las poblaciones genéticamente cercanas. Por lo tanto, las características estadísticas de tales medidas de la distancia deben todavía ser exploradas con mayor rigurosidad.

Adicionalmente, a partir de la popularización de técnicas basadas en la secuenciación de la molécula de ADN se han implementado medidas de distancia que estiman divergencia entre poblaciones siguiendo modelos evolutivos específicos Nei y Li (1979).

En la Tabla 3, se presenta la formulación y propiedades de las distancias genéticas entre poblaciones más utilizadas.

Tabla 3. Propiedades de algunas distancias genéticas

Distancias[1]		Rango	Tipo angular	Métrica	Euclídea
Cavalli-Sforza y Edwards (1967) (distancia del arco)	$\sqrt{\frac{1}{p}\sum_{k=1}^{p}\left(\frac{2}{\pi}\cos^{-1}\sum_{l=1}^{q}\sqrt{x_{ikl}.x_{jkl}}\right)^2}$	0,1	No	Si	Si
Cavalli-Sforza y Edwards (1967) (distancia de la cuerda)	$\sqrt{2.\left(1-\frac{1}{p}\sum_{k=1}^{p}\left(\sum_{l=1}^{q}\sqrt{x_{ikl}.x_{jkl}}\right)\right)}$	0,2	No	Si	Si
Nei (1972)	$-\ln\frac{\sum_{k=1}^{p}\sum_{l=1}^{q}x_{ikl}.x_{jkl}}{\sqrt{\sum_{k=1}^{p}\sum_{l=1}^{q}x_{ikl}^{2}.\sum_{k=1}^{p}\sum_{l=1}^{q}x_{jkl}^{2}}}$	0,∞	Si	No	No
Nei (1978)	$-\ln\frac{\sum_{k=1}^{p}\sum_{l=1}^{q}x_{ikl}.x_{jkl}}{\sqrt{\frac{\left(2n_i\sum_{k=1}^{p}\sum_{l=1}^{q}x_{ikl}^{2}\right)-1}{2n_i-1}.\frac{\left(2n_j\sum_{k=1}^{p}\sum_{l=1}^{q}x_{jkl}^{2}\right)-1}{2n_j-1}}}$	0,1	Si	No	No
Hillis (1984)	$-\ln\left(\frac{1}{p}\sum_{k=1}^{p}\frac{\sum_{l=1}^{q}x_{ikl}.x_{jkl}}{\sqrt{\sum_{l=1}^{q}x_{ikl}^{2}\sum_{l=1}^{q}x_{jkl}^{2}}}\right)$	0,1	Si	No	No
Swofford – Olsen (1990)	$-\ln\left(\frac{1}{p}\sum_{k=1}^{p}\frac{\sum_{l=1}^{q}x_{ikl}.x_{jkl}}{\sqrt{\frac{\left(2n_i\sum_{l=1}^{q}x_{ikl}^{2}\right)-1}{2n_i-1}\frac{\left(2n_j\sum_{l=1}^{q}x_{jlk}^{2}\right)-1}{2n_j-1}}}\right)$	0,1	Si	No	No
Rogers (1972)	$\frac{1}{p}\sum_{k=1}^{p}\sqrt{\sum_{l=1}^{q}(x_{ikl}-x_{jkl})^2}$	0,1	Si	Si	No
Prevosti (Wright, 1978)	$\frac{1}{p}\sum_{k=1}^{p}\sum_{l=1}^{q}\lvert x_{ikl}-x_{jkl}\rvert$	0,1	Si	Si	No

1. Ver referencias en el texto

1.4 PROPIEDADES DE LOS DATOS

Se han descrito medidas de similitud y distancia asociadas a los diferentes tipos de datos; así mismo, en apartados anteriores se han descrito los diferentes tipos de marcadores que ayudan a comprender la diversidad genética. Existe una relación entre el tipo de dato, el tipo de marcador y la medida de similitud o distancia que puede ser calculada para representar a los individuos y poder estudiar las relaciones entre ellos.

A continuación se considerarán brevemente algunos aspectos importantes acerca de la naturaleza y las propiedades de los datos y para esto se englobarán los marcadores en dos grupos. Los derivados de la descripción morfológica de órganos vegetativos y reproductivos, cuantificación de variables agronómicas clásicas y datos sobre susceptibilidad a factores de estrés, patógenos y enfermedades (caracteres agromorfológicos) y los derivados de las técnicas bioquímicas y moleculares cuantificados a través de un patrón de bandas o alelos observados en un mismo locus o diferentes loci (caracteres bioquímicos y moleculares).

1.4.1 Caracteres agromorfológicos

Están conformados por características fenotípicas de fácil identificación visual o medición tales como: color, forma, tipo, número y tamaño de inflorescencias o frutos; tipo, forma y altura de crecimiento de las plantas; biomasa; rendimiento y susceptibilidad o resistencia a stress hídrico, plagas o enfermedades. Estos caracteres suelen estar definidos por descriptores para cada cultivo aprobados por los organismos internacionales encargados de ello.

Los datos registrados pueden ser del tipo binario, cuando denoten la presencia o ausencia de un carácter; del tipo multinomial cuando un carácter tenga más de dos clases; del tipo ordinal cuando el carácter además de tener más de dos clases éstas estén referidas a una escala de valoración; y cuantitativos cuando la medición del carácter provenga de un registro continuo, por ejemplo de un aparato de precisión (Sneath y Sokal, 1973).

A pesar de la variabilidad de la información agromorfológica, -con las herramientas provistas en los apartados anteriores- es posible cuantificar la similitud o distancia entre pares de individuos o poblaciones, según sea el caso. No obstante, se recomienda: no utilizar un excesivo número de clases ya que artificialmente podemos estar dando mayor peso a unas variables respecto a otras; en el caso de que se evalúe simultáneamente la susceptibilidad o resistencia de los individuos a varias plagas o enfermedades deberá utilizarse una escala única de valoración y el cero (0) debe ser considerado como el nivel más bajo; se recomienda expresar los caracteres cuantitativos en unidades métricas similares y si no es posible la mejor alternativa es la estandarización. En cualquier caso, y dependiendo del método de codificación del carácter, es posible construir un vector $\mathbf{x}_{nx1}$ o $\mathbf{x}_{nxk}$, donde *k* es el número de clases.

1.4.2 Caracteres bioquímicos y moleculares

Si es conocido el patrón genético de la progenie de los individuos bajo estudio, los caracteres bioquímicos y/o moleculares estarán conformados por la expresión genotípica cuantificada, a través del patrón multibanda observado para cada individuo

después de su visualización mediante la electroforesis de proteínas o de productos de amplificación de la molécula de ADN. Sin embargo, en estudios de diversidad genética de bancos de germoplasma y especialmente los asociados a especies tropicales y subtropicales no existen estudios previos de segregación por lo que la interpretación que se realice sobre el patrón multibanda dependerá del tipo de marcador y la especie (Telles *et al.*, 2006).

Como fue mencionado en el apartado 1.1 de acuerdo a su capacidad de detectar polimorfismos o capacidad informativa pueden diferenciarse dos tipos de marcadores: Los dominantes, donde las bandas detectadas se comportan como si se tratase de la expresión fenotípica de un par de alelos donde uno es nulo y por tanto recesivo frente al otro alelo, de tal forma que el homocigoto dominante y el heterocigoto presentarán productos de amplificación o banda, mientras que el homocigoto recesivo no. Es decir, en estos casos no se pueden discriminar los homocigotos dominantes de los heterocigotos para un segmento particular (Zintzaras, 2008; Zhao *et al.*, 2008). La estimación de las frecuencias alélicas se debe hacer de manera indirecta, asumiendo equilibrio de Hardy-Weinberg (Sato *et al.*, 2006; Zintzaras, 2008). Los codominantes, donde los genotipos homocigotos y heterocigotos pueden ser distinguidos con mucha precisión (Sato *et al.*, 2006).

En la Figura 2ab, se representan los esquemas probables de amplificación para un marcador molecular dominante y uno codominante, para un organismo diploide con loci bialélicos. En el caso del marcador dominante las bandas son del mismo grueso y

representan la alternativa alélica A_1A_1, A_1A_2 y no se puede diferenciar si el alelo es homocigoto dominante o heterocigoto (Figura 2a) y la ausencia de banda se interpreta como genotipo A_2A_2. Para el marcador codominante (Figura 2b) los individuos homocigotos se representan como una banda gruesa a distinta altura, mientras que los heterocigotos con dos bandas más finas de forma simultánea.

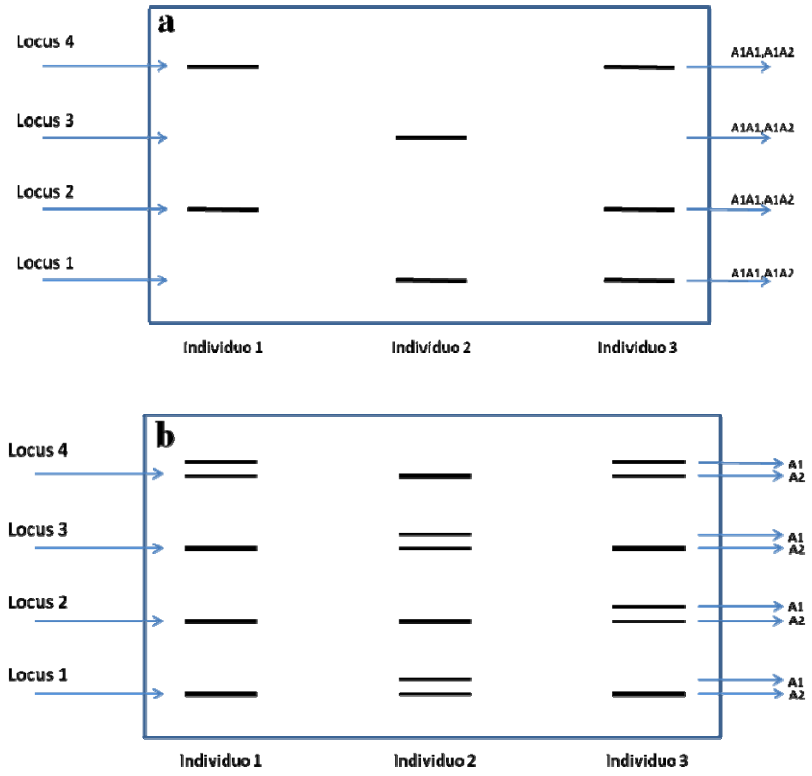


Figura 2. Esquema de amplificación para un marcador molecular en un organismo diploide con loci bialélicos: (**a**) dominante y (**b**) codominante.

Así, el individuo 1 de la Figura 2a para el locus 4 será indistinguible A_1A_1 o A_1A_2 mientras que el individuo 2 para ese mismo locus será A_2A_2. En la Figura 2b, el

individuo 1 tiene genotipo A_1A_2 para el locus 4 y el individuo 2 tiene genotipo A_2A_2 para ese mismo locus. Estas diferencias respecto a los productos de amplificación observables suponen también diferencias en el análisis de cada tipo de marcador. Por esta razón, al analizar los datos obtenidos de marcadores dominantes se deben tener en cuenta los siguientes supuestos: (***i***) cada uno de los marcadores representa un locus mendeliano en el cual el marcador visible, el alelo dominante, está en equilibrio de Hardy-Weinberg con un alelo recesivo, es decir, cada patrón de banda observado corresponderá a loci específicos o a la región del genoma que un loci particular explore. (***ii***) los alelos marcados para diferentes loci no migran a la misma posición en el gel, es decir, la presencia/ausencia observada para cada banda se interpreta como variación en loci individuales y (***iii***) los alelos observados sobre un locus particular estarán más correlacionados entre sí que los observados en diferentes locus (Lynch y Milligan, 1994; Avise, 2004).

En el esquema de amplificación presentado (Figura 2ab) si se sigue la codificación binomial típica que utiliza 1 para indicar la presencia y 0 para la ausencia. Para el marcador dominante cada locus generara una columna o variable, mientras que para el marcador codominante se generará una columna por alelo o su equivalente dos columnas por locus ya que es posible la diferenciación de las alternativas alélicas A_1A_1 y A_1A_2 (Tabla 4).

Tabla 4. Expresión del genotipo y codificación de los fragmentos de amplificación para un organismo diploide con loci bialélicos, utilizando un marcador dominante y uno codominante.

		EXPRESION DEL GENOTIPO				CODIFICACION			
	Individuo	**Locus 1**	**Locus 2**	**Locus 3**	**Locus 4**	**Locus 1**	**Locus 2**	**Locus 3**	**Locus 4**
M.D	**1**	A_2A_2	A_1A_1,A_1A_2	A_2A_2	A_1A_1,A_1A_2	0	1	0	1
	2	A_1A_1,A_1A_2	A_2A_2	A_1A_1,A_1A_2	A_2A_2	1	0	1	0
	3	A_1A_1,A_1A_2	A_1A_1,A_1A_2	A_2A_2	A_1A_1,A_1A_2	1	1	0	1
M.COD	**1**	A_2A_2	A_2A_2	A_2A_2	A_1A_2	0,1	0,1	0,1	1,1
	2	A_1A_2	A_2A_2	A_1A_2	A_2A_2	1,1	0,1	1,1	0,1
	3	A_2A_2	A_1A_2	A_2A_2	A_1A_2	0,1	1,1	0,1	1,1

M.D →Marcador dominante
M.COD →Marcador codominante

Por lo tanto, la capacidad informativa dependerá del tipo de marcador y está asociada con la capacidad de diferenciación de las alternativas alélicas, es así como marcadores dominantes ofrecerán menos información genética por locus que marcadores codominantes.

Las consideraciones presentadas suponen que la capacidad informativa de un marcador afectará la clasificación de los individuos. A continuación ilustraremos de forma empírica y con propósitos académicos cómo afecta el empleo de uno u otro marcador a la formación de grupos. Con este propósito se utilizarán dos escenarios simulados como se describe a continuación.

1.4.2.1 Estudio de simulación

- **Escenario 1:** se simularon genotipos de individuos diploides, suponiendo una población apareada al azar con respecto a cada locus bialélico, sin consanguinidad y que los genotipos para cada locus seguían el equilibrio de Hardy-Weinberg, es decir, $p^2 + 2pq + q^2 = 1$, donde p es la frecuencia poblacional de A_1 y q la de A_2.

- **Escenario 2:** se simularon genotipos de individuos diploides, suponiendo una población no apareada al azar con respecto a cada locus bialélico, es decir, con consanguinidad hasta la presencia de líneas puras **y** sin equilibrio**,** este procedimiento se realiza para no poner en ventaja la presencia del alelo en el caso de los codominantes.

En cada escenario se generaron en forma aleatoria 3 grupos de 15 individuos y 10 loci con frecuencias alélicas poblacionales iguales (Tabla 5). Cada escenario fue repetido 1000 veces.

Tabla 5. Frecuencias genotípicas por escenario y grupos simulados.

	Escenario 1			**Escenario 2**		
	$p^2 + 2pq + q^2 = 1$			$p^2 + 2pq + q^2 \neq 1$		
Grupos	A_1A_1	A_1A_2	A_2A_2	A_1A_1	A_1A_2	A_2A_2
1	0.01	0.18	0.81	0.90	0.05	0.05
2	0.25	0.50	0.25	0.05	0.90	0.05
3	0.81	0.18	0.01	0.05	0.05	0.90

Las alternativas alélicas generadas por la simulación de los individuos en cada escenario considerado, fueron codificadas siguiendo el mismo procedimiento que se mostró en la Tabla 4, es decir, para el caso de un marcador dominante $A_1A_1 = A_1A_2 = 1$ y $A_2A_2 = 0$,

generando una sola columna por locus. Para el marcador codominante se consideró todas las alternativas genotípicas A_1A_1, A_1A_2 y A_2A_2 generando dos columnas por locus, una para el alelo A_1 y otra para el alelo A_2. Las alternativas alélicas fueron perturbadas utilizando un error aleatorio del 5%.

Para valorar cómo el tipo de marcador afecta la clasificación de los individuos -capacidad informativa-, con los patrones generados para el conjunto de loci por cada tipo-de-marcador/escenario/repetición se estudiaron sus relaciones a través del análisis de conglomerados. Se calcularon las disimilitudes de las matrices utilizando los coeficientes de Jaccard, de Emparejamiento Simple, Dice y Rogers y Tanimoto, y fueron representadas gráficamente por un árbol jerárquico utilizando el encadenamiento promedio o UPGMA **(**Sneath y Sokal, 1973). Se seleccionó este método porque es el de mayor uso en estudios de clasificación de genotipos utilizando marcadores moleculares. Las tasas de error de clasificación se usaron para estimar el error aparente al clasificar los individuos bajo las combinaciones tipo-de-marcador/escenario/repetición respecto a la clasificación de referencia o grupos definidos a *priori* en las simulaciones.

En los resultados de las simulaciones (Figura 3), se destaca que independientemente del coeficiente de similitud y del equilibrio poblacional, el porcentaje de error al ubicar los individuos en los grupos establecidos *a priori* es mayor cuando el marcador es dominante, siendo la tasa de error en algunos casos hasta 3 veces superior respecto al marcador codominante. Este error está asociado al bajo poder discriminante de este tipo de marcador ya que incluye prácticamente a todos los individuos en un solo grupo. En

el caso del marcador codominante los errores de clasificación están asociados, en su mayoría, al grupo 2 del escenario en equilibrio debido a que la frecuencia poblacional de heterocigotos es 2 veces la frecuencias de las alternativas alélicas A_1A_1 y A_2A_2, respectivamente. Adicionalmente, aquellos coeficientes -Emparejamiento Simple y Rogers y Tanimoto- donde tanto la ausencia como la presencia simultánea del carácter contribuyen a la semejanza entre los individuos son los que presentan menor tasa de error, independientemente del escenario y el tipo de marcador. Aunque Cuadras (1996) previene que la utilización de estos coeficientes puede ocasionar problemas ya que al añadir caracteres arbitrarios no comunes, podrían hacerse falsamente similares individuos que no lo son. En el caso de la utilización de marcadores moleculares para detectar similitudes entre individuos la ausencia simultánea de un carácter (locus, alelos) nunca será posible porque cuando sucede este fenómeno se considera que el marcador no fue capaz de extraer información del individuo a través de expresión genotípica. Además, por la codificación empleada, es necesario considerar similitud el hecho de la ausencia simultánea tal como hacen esos coeficientes puesto que, de lo contrario, serían más similares dos individuos heterocigotos A_1A_2 (dos coincidencias con codificación 1,1) que dos homocigotos A_1A_1 (una coincidencia con codificación 1,0 si el doble 0 no contara como coincidencia).

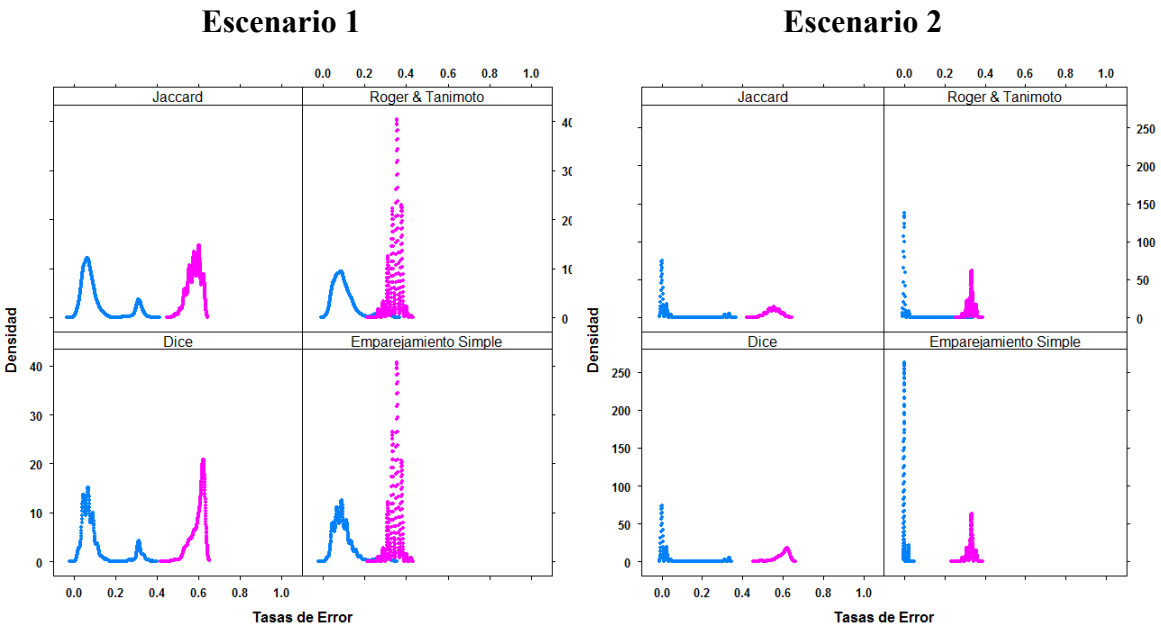


Figura 3. Distribución de las tasas de error de clasificación por tipo de marcador y coeficientes de similitud: (■) Marcador Dominante y (■) Marcador Codominante.

No obstante los resultados de este ejercicio empírico, es posible suponer que la ventaja informativa mostrada por el marcador codominante se debe a la forma de codificación ya que se duplica el número de variables por considerar dos columnas por locus bialélico. Por esta razón, y a los fines de complementar la información, se consideró la alternativa de codificación multinomial asignándole los valores de 1, 2 y 3 para las alternativas alélicas A_1A_1, A_1A_2 y A_2A_2, respectivamente, obteniéndose una columna para cada locus. Es así como, utilizando los dos escenarios considerados, los coeficientes de emparejamiento simple, el de Gower y el mismo método de clasificación se obtuvo una distribución del error de clasificación similar a la obtenida para marcadores codominantes cuando se utiliza el coeficiente de Gower (Figura 4). En el

caso del coeficiente de emparejamiento simple la tasa de error tiene la mayor amplitud registrada debiéndose exclusivamente a la tasa de error de clasificación del grupo 2 que fue más del 60% y esto como ya se detalló debe a la frecuencia poblacional de heterocigotos considerados.

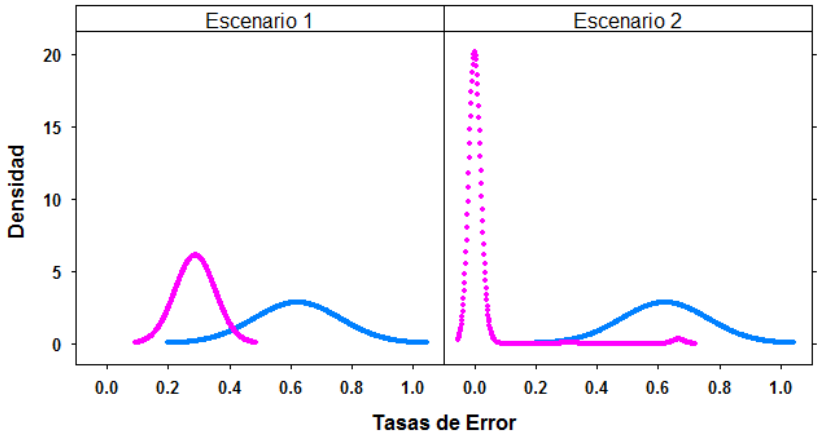


Figura 4. Distribución de las tasas de error de clasificación para la alternativa de codificación multinomial y los Coeficientes (■) Emparejamiento Simple y (■) Gower.

Esta alternativa multinomial de codificación, de solo registrar el número de una de las alternativas alélicas para cada locus e individuo, bajo el supuesto de loci bialélicos es sencilla, provee información suficiente para reescribir el genoma completo y facilita la interpretación y cálculo de las frecuencias cuando se comparan grupos o poblaciones y se quiere hacer uso de alguna de las distancias genéticas descritas en apartados previos.

De los resultados presentados destacan entre otros, que el tipo de marcador, el coeficiente de similitud, la forma de codificación y el tipo de organismo constituyen un continuo relacionado que debe ser considerado a la hora de construir las matrices de

similitud entre pares de individuos. Para marcadores dominantes tales como los RAPD, AFLP y ISSR los fragmentos de amplificación deben ser codificados por locus utilizando la codificación binomial típica que utiliza 1 para indicar la presencia y 0 para la ausencia. Para el caso de marcadores codominantes, con organismos diploides, en ausencia de análisis de segregación y si no se realiza ninguna suposición sobre la naturaleza de los alelos puede utilizarse codificación binomial típica pero sobre los alelos, generando dos columnas por locus (Cordeiro *et al.*, 2003). Cuando se desconocen tanto el tipo de herencia como la naturaleza de los alelos y el organismo es poliploide, lo recomendable es utilizar una alternativa multinomial de codificación por locus, asignándole diferentes valores a las alternativas alélicas observadas. Si se utiliza la alternativa de codificación binaria debe tenerse en cuenta el nivel de ploidía puesto que en marcadores dominantes puede subestimarse la diversidad genética al ocultar la presencia de alternativas alélicas; y en el caso de los marcadores codominantes, deberán usarse muestras de gran tamaño para que puedan detectarse todos los genotipos posibles.

Así mismo, las medidas de similitud o distancia que pueden ser definidas entre dos individuos o grupos también dependerán del tipo de marcador molecular y los individuos en estudio. Kosman y Leonard (2005), analizan los diferentes coeficientes de similaridad utilizados para estudiar las relaciones genéticas entre individuos a través de marcadores moleculares y demuestran que no existe un criterio universal. Puesto que diferentes medidas de similaridad podrán ser usadas en organismos con diferente grado de poliploidía, tipo de marcador y objetivos del estudio. Rohlf y Sokal (1981) usando un

argumento más estadístico consideran que se puede usar como criterio de selección del mejor coeficiente, la relación entre la métrica de las técnicas de ordenación o clasificación y la distancia observada entre los individuos.

En resumen, utilizando los criterios de codificación mencionados, para evaluar el grado de similitud entre individuos o genotipos para marcadores dominantes, pueden utilizarse los coeficientes de similaridad de: Jaccard, Emparejamiento Simple, Dice o Roger y Tanimoto, entre otros. Para el caso de marcadores codominantes, se pueden utilizar los mismos coeficientes de similaridad que para marcadores dominantes cuando se codifique sobre alelos y los coeficientes de similitud propios para variables multiestado cuando se codifique sobre locus. En cualquier caso, las medidas de distancia genéticas solo podrán utilizarse si en el estudio son reconocibles dos o más poblaciones o grupos de individuos, ya que solo sobre éstas es posible calcular las frecuencias alélicas necesarias para determinar las distancias.

1.5 TECNICAS DE AGRUPAMIENTO

En los apartados previos se han presentado y discutido aspectos relevantes sobre la diversidad genética referidos a su descripción y al cálculo de las relaciones dentro y entre poblaciones. A continuación presentaremos un resumen del estado actual del arte sobre las metodologías más utilizadas para la representación de relaciones entre individuos y poblaciones, y se realizan algunas consideraciones prácticas.

CAPITULO I

La clasificación o el agrupamiento de individuos o poblaciones con patrones similares y la ordenación o la representación espacial de los individuos o poblaciones en un sistema de coordenadas son las dos grandes técnicas multivariantes de uso general en la visualización de la estructura de diversidad genética. Y aunque la selección del método de agrupación u ordenación debe estar asociado entre otros, al tipo de caracteres evaluados, a la diversidad, al sistema de reproducción, al nivel de ploidía y a los niveles de heterocigosidad de la especie. Además de la concordancia de las clasificaciones generadas respecto al conocimiento del patrón natural o filogenético de los individuos o poblaciones en estudio (Boontong *et al.,* 2008), en la bibliografía especializada se observa un atavismo a las mismas estrategias de análisis, bien sea por el desconocimiento de nuevos métodos que han sido desarrollados para ciertas particularidades de los estudios de diversidad genética y/o a la proliferación de paquetes informáticos que han facilitado la aplicación de una determinada metodología.

Es así como, revisando los artículos sobre estudios de diversidad genética que han sido publicados en revistas internacionales en los últimos diez años y que se encuentran registrados en la base de datos "Web of Science database (ISI Web of Knowledge 2008), encontramos que las técnicas más utilizadas para la representación de relaciones entre individuos y poblaciones, con marcadores agromorfológicos basadas en variables cuantitativas, son el Análisis de Componentes Principales (ACP) y los árboles jerárquicos generados del análisis de conglomerados sobre matrices de distancias euclídeas utilizando el algoritmo UPGMA (Xiaoyan *et al.*, 2008; Sarwat y Srivastava, 2008; Nghia *et al.*, 2008; Gökirmak *et al.,* 2008; Brito *et al*., 2008 ; Kar *et al*., 2008 ; Li

et al., 2008 y Badea *et al*., 2008). Para la caracterización usando marcadores moleculares tanto dominantes como codominantes, la técnica de clasificación mayormente utilizada también es la de los árboles jerárquicos derivados del análisis de conglomerados sobre matrices de similitud utilizando el algoritmo UPGMA (Kaundun y Park. 2002; Rouf Mian *et al*., 2005; Franco *et al*., 2006; Wang *et al.*, 2006; Perumal *et al*., 2007; Terzopoulos y Bebeli, 2008). En menor proporción se emplea la clasificación utilizando árboles no-jerárquicos generados sobre matrices de similitud pero aplicando el método de "Neighbor-joining" (Tar'an *et al*., 2005; Gillaspie *et al*., 2005; Barkley *et al*., 2006; Barry *et al*., 2007; Payn *et al*., 2008) y la representación generada a través del Análisis de Coordenadas Principales (ACoP) (Reif *et al*., 2004; Dreisigacker *et al*., 2005; Alwala *et al*., 2006; Teklewold y Becker. 2006; Tamiru *et al*., 2007; Ofori *et al*., 2008; Terzopoulos y Bebeli. 2008).

Cuando se estudian las relaciones entre diferentes marcadores, la amplia mayoría realiza repetidamente correlaciones entre matrices de distancias y/o similaridades para cuantificar la concordancia entre caracterizaciones (Tar'an *et al*., 2005; Garcia *et al.*, 2007; Syamkumar y Sasikumar. 2007; Kalita *et al*., 2007), y son pocas las referencias donde se realiza un tratamiento de consenso entre caracterizaciones (Bramardi *et al.,* 2005; Esposito *et al*., 2007).

Con excepción de cuando se utiliza el Análisis de Componentes Principales (ACP) basado en variables cuantitativas, ninguno de los métodos revisados incluye

información sobre las variables responsables de las clasificaciones (Hillis y Moritz, 1990; Powell *et al.*, 1996; Graur y Wen-Hsiung, 2000; Infante *et al.*, 2006).

El uso prácticamente universal del Análisis de Conglomerados (AC) se debe a que el método construye grupos de individuos sin que exista un conocimiento previo de esas estructuras, solamente es requerido el conocimiento de la similaridad entre los pares de objetos de acuerdo a cualquiera de los métodos estudiados previamente. Sin embargo, el problema con éste análisis es que puede identificar o generar grupos que no existen en forma natural (Everitt, 1979 y 2001). Como consecuencia de este defecto, se han desarrollado múltiples algoritmos de agrupación para satisfacer diversos criterios y por consiguiente, cada uno produce diferentes patrones de agrupación. Adicionalmente, los árboles jerárquicos derivados del Análisis de Conglomerados (AC) requieren que los datos satisfagan ciertas propiedades que no son de fácil cumplimiento. Por ejemplo, los algoritmos de generación de árboles ultramétricos como es el caso del UPGMA, suponen que las matrices de similitud o disimilitud generadas por los diferentes coeficientes sean ultramétricas. Si esto es cierto, entonces la representación mediante árboles jerárquicos es exacta, pero si no es así, como ocurre en la generalidad de los casos, se estará introduciendo un error por la adecuación de una distancia no ultramétrica a un árbol ultramétrico. El error de representación está presente, independientemente que el algoritmo de construcción del árbol trate de transformar "razonablemente" la disimilaridad original en ultramétrica. Ya fue mencionado en apartados previos que la propiedad de la desigualdad ultramétrica es una condición muy difícil de satisfacer y ninguna de las medidas de similitud utilizadas para clasificar

genotipos, utilizando por ejemplo, marcadores moleculares la cumplen por definición, a no ser para conjuntos de datos particulares (Swofford y Olsen, 1990; Hall, 2001).

Otro aspecto contradictorio del uso de algoritmos de generación de árboles ultramétricos, como es el caso del UPGMA, es que éstos suponen un modelo único de ultrametricidad, ya que está comprobado que no todas las regiones del genoma evolucionan concertadamente, debido a que unas están sometidas a fuertes presiones selectivas y otras a una variación prácticamente neutral. En otras palabras, la topología o forma del árbol la mayor parte de las veces es incorrecta ya que la tasa de evolución no es constante a lo largo de las distintas ramas (Avise, 2004; Graur y Wen-Hsiung, 2000). En estos casos si se insiste en usar el análisis de conglomerados, lo más recomendable es probar modelos aditivos que se ajusten perfectamente a árboles no-jerárquicos como el "Neighbor-joining". Debe resaltarse que la aditividad de las distancias entre individuos no es un supuesto que se cumple a *priori*, aunque esta condición es menos restrictiva que la de ultrametricidad. (Saitou y Nei, 1987).

Estas ambigüedades determinan que el agrupamiento logrado para identificar clases existentes en relación a los individuos, dependerá no sólo de la medida de similitud o disimilitud seleccionada, sino del número de grupos que deben ser formados (cuando esta información exista) así como también del método de agrupación y del algoritmo de agregación elegido.

CAPITULO I

Los inconvenientes metodológicos que acarrea el uso del análisis de conglomerados han impulsado el desarrollado de procedimientos que permiten validar la exactitud, confiabilidad y estabilidad de las clasificaciones obtenidas. Entre los más utilizados se encuentran: la validación externa, que consiste en comparar la matriz de distancia con otra información que no se haya usado en los cálculos de agrupación, como puede ser la información sobre el pedigrí (Avise, 2004); el coeficiente de correlación cofenética, que cuantifica la distorsión debida al método de agrupación empleado, comparando la matriz de distancias original y la derivada del árbol llamada matriz cofenética (Sokal y Rohlf, 1962); y los árboles de consenso, árboles que resumen las configuraciones de dos o más árboles y que reúne toda o una parte de las coincidencias reflejadas en ellos, pueden construirse a través de métodos de remuestreo (Jain y Moreau, 1987) o métodos bayesianos (Brouat *et al.*, 2004).

Entre los criterios mencionados, los árboles de consenso construidos a través de métodos de remuestreo son los más populares debido al desarrollo de paquetes informáticos que permiten su aplicación. Se basan en realizar pseudo-réplicas del conjunto original de datos, para cada una de las cuales se reconstruye un árbol por el método que se esté utilizando y posteriormente se estudia el grado de estabilidad de las distintas ramas del árbol. Esto produce conjuntos de matrices con el mismo número de variables, pero en los que unas columnas aparecen más de una vez y otras no aparecen. Como cada pseudo-réplica es distinta, el resultado es ligeramente distinto. Cuando se superponen las réplicas unas partes del árbol aparecen en todas las réplicas o en un número elevado de ellas, lo que significa que están soportadas por un número de

residuos estadísticamente significativo, y otras no. Felsenstein (2004) sugiere que como el supuesto de independencia de los alelos (columnas) no puede ser satisfecho, su uso no es siempre adecuado.

Consideramos que las ventajas que el análisis de conglomerados ofrece a la taxonomía tradicional y a la biología evolutiva con respecto a la calidad de clasificaciones y de su interpretación, varían en el contexto de la clasificación de genotipos y en especial cuando se basa en marcadores moleculares. Por otra parte, estas técnicas de agrupación no ofrecen información de las relaciones entre individuo-marcador y marcador-marcador.

Aunque su uso sigue siendo restringido, ya que son muy pocas las especies vegetales a las que se les ha descrito el genoma, los estudios filogenéticos basados en secuencias de ADN han recibido en los últimos años la mayor atención. Este tipo de información, que está mucho más íntimamente ligada a la evolución de los organismos que cualquier otro tipo de los marcadores mencionados, permite la construcción de filogenias introduciendo modelos evolutivos que mejoran las fallas metodológicas consideradas en este apartado. Según nuestro criterio, esto ha desviado la atención de los problemas de la clasificación de genotipos usando marcadores moleculares clásicos, sin que aun se hayan respondido muchas preguntas metodológicas. En el caso de especies altamente comerciales esto no es un problema, pero en el caso de la mayoría de las especies tropicales y subtropicales los análisis y el enfoque metodológico deben seguir siendo objeto de estudio para mejorarlos.

Capítulo II

CLASIFICACIÓN DE GENOTIPOS Y TÉCNICAS DE ORDENACIÓN

CAPITULO II

El capítulo anterior finalizó con algunas consideraciones metodológicas sobre el uso casi exclusivo, que la mayoría de las publicaciones relacionadas con estudios de diversidad genética hacen, del Análisis de Conglomerados (AC) en sus distintas versiones. En este capítulo y el siguiente construiremos nuestra estrategia metodológica, que propone el uso combinado de técnicas de ordenación, de clasificación y de modelaje desde una perspectiva biológica que permiten responder las preguntas que un genético, mejorador o simplemente un responsable del mantenimiento de un banco de germoplasma hacen sobre: qué individuos son iguales; cuántos grupos pueden formarse; cuáles son las variables, caracteres o alelos que definen los grupos; cómo interactúan estas variables. Alguno de estos interrogantes no tienen respuesta con la aplicación de los análisis clásicos que se han utilizado hasta ahora en las diferentes publicaciones.

Adicionalmente, es conveniente aclarar que los análisis comúnmente realizados sobre caracteres cuantitativos, que en su mayoría usan el Análisis de Componentes Principales (ACP), no requieren mayor discusión a los fines de la innovación metodológica que proponemos. En cualquier caso esta técnica ofrece ventajas sobre el análisis de conglomerados en el sentido de que es posible estudiar, a través de las proyecciones de las variables, la relación de éstas con los individuos y su importancia en la formación de grupos.

En este orden de ideas, se abordaran dos técnicas de ordenación ampliamente conocidas, desde la perspectiva del carácter interrelacionado de las variables estudiadas, aspecto esencial del Análisis Multivariante. La primera es el Análisis de Coordenadas Principales (ACoP), donde reivindicaremos su aplicación a los estudios de diversidad genética, introduciendo medidas de variabilidad y calidad de la representación de individuos y grupos. La segunda son los métodos Biplot donde mostraremos su aplicabilidad en el estudio de las relaciones entre individuos y variables en aspectos como: facilidad de interpretación, riqueza del análisis y utilidad, ya que permiten responder las preguntas que no tienen respuesta en los análisis clásicos.

La estrategia metodológica se orientará a la operación con matrices de similitud derivadas de la observación de variables del tipo binario, variables en su mayoría asociadas a marcadores moleculares, aspecto al que debe prestarse la mayor atención desde el punto de vista metodológico. En este sentido, para los marcadores dominantes, se asume la codificación binomial por locus y para el caso de marcadores codominantes se codifica sobre los alelos, generando dos columnas por locus.

2.1 ANALISIS DE COORDENADAS PRINCIPALES (ACoP)

El Análisis de Coordenadas Principales (ACoP) es un procedimiento geométrico que permite encontrar una configuración $\mathbf{Y}$ en un espacio euclídeo $\mathbb{R}^k$ de baja dimensión tal que la distancia inter-puntos de la matriz estimada $\mathbf{D}$, sea lo más cercana posible a la matriz observada Δ. La aproximación *k*-dimensional se encontrará en las primeras *k* columnas de $\mathbf{Y}$ que son llamadas coordenadas principales (Gower, 1966).

Formalmente, consideramos para el cálculo de $\mathbf{Y}$ a: $\mathbf{S}_{nxn} = (s_{ij})$ la matriz simétrica que contiene las similitudes entre los *n* individuos o genotipos (Unidades Taxonómicas Operativas, UsTO), obtenida de la matriz de datos $\mathbf{X}$ de orden (*n*x*p*), siendo *p*>*k* el conjunto de caracteres o atributos que se miden sobre cada individuo, donde x_{ij} denotará la medición en el individuo *i*-ésimo de la variable *j*-ésima. *Sea* $\Delta = (\delta_{ij})$ una matriz de distancias euclídeas que ha sido derivada por alguna de las transformaciones siguientes: $\delta_{ij} = \sqrt{1 - s_{ij}}$ o $\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$. Sean $\mathbf{B}$ y $\mathbf{H}$, la matriz de productos escalares de Δ y la matriz de centrado como han sido definidas en [1.1] y [1.2], respectivamente; siendo que $\mathbf{B} \geq 0$ (semidefinida positiva), entonces para un conjunto de puntos de una configuración $\mathbf{Y} = \left(y_1, ..., y_n\right)'$, la distancia euclídea entre estos puede ser representada por:

$$b_{ij} = \left(y_i - \bar{y}\right)' \left(y_j - \bar{y}\right) \qquad i, j = 1, ..., n \qquad [2.1]$$

y en forma matricial como:

$$\mathbf{B} = \left(\mathbf{HY}\right)\left(\mathbf{HY}\right)' \qquad [2.2]$$

La matriz $\mathbf{B} \geq 0$ será la matriz de productos escalares de la configuración $\mathbf{Y}$. Si $\mathbf{B}$ es una matriz semidefinida positiva de rango *p*, entonces una configuración correspondiente a $\mathbf{B}$ puede construirse a partir de los valores y vectores propios de $\mathbf{B}$ tal que:

$$\mathbf{Y} = \mathbf{U}\Lambda^{1/2} \qquad [2.3]$$

donde $\mathbf{B} = \mathbf{U}\Lambda\mathbf{U}'$ es la descomposición espectral de $\mathbf{B}$ y $\mathbf{U}'\mathbf{U} = \mathbf{I}$, siendo $\mathbf{U}$ una matriz de rango *p* que contiene los vectores propios en las columnas y $\mathbf{\Lambda} = diag\left(\lambda_1,...,\lambda_p\right)$ es una matriz diagonal de rango *p* que contiene los valores propios ordenados en forma creciente $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$.

La solución consiste en seleccionar los primeros vectores propios correspondientes a los valores propios más grandes. Se obtendrá una mejor representación de $\mathbf{\Delta}$ en el espacio *k*-dimensional reducido, en la medida que los primeros valores propios sean considerablemente mayores que el resto, es decir, se obtendrá la mejor representación con la menor pérdida de información.

Las coordenadas generadas pueden representarse arbitrariamente en el espacio sin alterar las distancias ajustadas entre puntos. Por convención, los puntos se centran al

origen y los ejes se rotan de modo que las primeras *k dimensiones* den el mejor ajuste. En el caso de que se encuentren valores propios negativos, es decir, que no exista la representación euclídea de la distancia $\Delta = (\delta_{ij})$, se recomienda revisar la transformación hecha sobre la matriz $\mathbf{S}_{nxn} = (s_{ij})$ o usar algunos de los procedimientos clásicos, como es sumar una constante hasta conseguir que la matriz sea semidefinida positiva, o buscar la matriz semidefinida positiva que más se aproxime a $\mathbf{B}$ igualando a cero los valores propios negativos y reconstruyendo la matriz de partida (Gower, 1966). En cualquier caso, si las transformaciones no funcionan es posible recurrir al método de escalas multidimensionales.

Al Análisis de Coordenadas Principales (ACoP) como técnica de ordenación, en el contexto de estudios de diversidad genética, le concierne el problema de la construcción de una configuración de *n* puntos (Unidades Taxonómicas Operativas, UsTO) en el espacio euclídeo, a partir de una matriz de distancia Δ, de manera tal que la distancia entre dos puntos cualesquiera de la configuración aproxime tanto como sea posible la disimilitud entre las UsTO representada por estos puntos, respetando la estructura de similaridades definida por la matriz de similaridades $\mathbf{S}$. La representación de los individuos en dimensión reducida simplifica el análisis de la dispersión permitiendo poner en evidencia posibles agrupamientos. Adicionalmente, el ACoP representa la mejor alternativa entre las técnicas de ordenación cuando *n* es más pequeño que *p* (Krzanowski, 2000), situación que involucra a la mayoría de los estudios de clasificación de genotipos usando marcadores moleculares.

La proporción de la variación explicada o bondad de ajuste total de la representación está dada por:

$$\left(\frac{\sum_{i=1}^{k} \lambda_i^2}{\sum_{i=1}^{n-1} \lambda_i^2} \right) x100\% \qquad [2.4]$$

donde k representa la dimensión retenida para la representación final y λ_i los autovalores de la matriz $\mathbf{B}$. Las consideraciones teóricas y las demostraciones del método pueden ser encontradas en Mardia *et al*. (1979).

Una vez obtenidas las coordenadas principales es necesario decidir cuántas de las *k*-dimensiones serán retenidas, es decir, cuánto de la información estamos dispuestos a perder en función de la reducción de la dimensión. Los diferentes criterios para seleccionar el número de dimensiones en el Análisis de Componentes Principales (ACP) (Bartlett, 1950; Anderson, 1963; Frontier, 1976), no son aplicables porque suponen la formulación de hipótesis sobre las variables originales cuantitativas, que no es el caso del ACoP. Por esta razón, graficar la distribución de los valores propios según su orden de magnitud y buscar en el gráfico el 'codo' que permita descartar la varianza explicada por el resto (Cattell, 1966), es el criterio más usado.

Hemos considerado como criterio adecuado para la selección del número *k* de dimensiones a ser retenidas un procedimiento análogo al de la correlación cofenética, el cual consiste en calcular la correlación lineal de Pearson entre los *n(n-1)/2* elementos

distintos fuera de la diagonal de las matrices de distancias observada Δ y estimada $\mathbf{D}$ para distintas combinaciones de *k*-dimensiones retenidas. Es así como, se sugiere descartar valores de $r \leq 0.8$ que indican una distorsión notable entre las disimilaridades iniciales y las estimadas.

Este criterio relaciona la métrica de la técnica de ordenación con la distancia observada entre los *n* individuos por lo que puede ser empleado para seleccionar el coeficiente de similitud que en menor dimensión, refleje la mayor coherencia entre las matrices de distancias observadas y estimadas. Las correlaciones graficadas deberán seguir un modelo exponencial con asíntota aproximadamente igual a uno. En el caso de variables del tipo binario, como la mayoría de las variables asociadas a marcadores moleculares, este criterio permitirá adicionalmente seleccionar el mejor coeficiente de similitud que represente la estructura natural de las relaciones entre los individuos.

2.1.1 Construcción de grupos

A diferencia del Análisis de Componentes Principales (ACP) en el Análisis de Coordenadas Principales (ACoP) las coordenadas no contienen información sobre las variables; sin embargo, cada objeto o individuo es identificado únicamente por sus coordenadas principales (Gower y Harding, 1988). Esta ventaja metodológica favorece la generación de clasificaciones o la definición de grupos homogéneos de individuos a través del Análisis de Conglomerados (AC).

Es así como, utilizando cualquier algoritmo de encadenamiento sobre las coordenadas principales $\mathbf{Y}$ o sobre la matriz de distancia estimada $\mathbf{D}$, es posible generar una clasificación de los individuos y representar las particiones obtenidas en la ordenación del Análisis de Coordenadas Principales (ACoP) usando envolventes convexas (*convex hulls*) de los puntos que pertenecen a cada grupo.

Puede argumentarse que usar las coordenadas principales $\mathbf{Y}$ de la matriz de distancia estimada $\mathbf{D}$ para el análisis adicional puede dar lugar a una pérdida de información, pero este procedimiento se puede también interpretar como una forma de separar la señal del ruido. Es decir, la pérdida de información que se derive por el uso de coordenadas principales para generar la clasificación es compensada porque el nivel de ruido se reduce.

Chae y Warde (2006) demuestran -usando simulaciones con grupos definidos *a priori* y tres tipos de estandarización-, que la capacidad de recuperación de información de algoritmos de agrupamiento se mejora considerablemente usando las coordenadas principales de los individuos, y sugieren una revisión cuidadosa de los grupos generados utilizando los algoritmos de agrupamiento sobre los datos originales ya que éstos son afectados por el ruido. Entonces se esperaría que para un mismo método de aglomeración, la clasificación generada utilizando las coordenadas principales de las *k*-dimensiones retenidas o la matriz de distancia estimada $\mathbf{D}$ sea similar o superior a la generada utilizando la matriz de distancia observada Δ.

A continuación ilustraremos de forma empírica la capacidad de recuperación de información del algoritmo de agrupamiento UPGMA utilizando las primeras dos coordenadas principales retenidas para generar las clasificaciones de genotipos usando datos típicos de marcadores moleculares.

2.1.1.1 Estudio de simulación

Se simularon en forma aleatoria 3 grupos de 15 individuos diploides y 10 loci, suponiendo en cada caso una población no apareada al azar con respecto a cada locus bialélico. Las frecuencias alélicas poblacionales asignadas a cada grupo fueron: $A_1A_1 = 0.90$, $A_1A_2 = 0.05$ y $A_2A_2 = 0.05$; $A_1A_1 = 0.05$, $A_1A_2 = 0.90$ y $A_2A_2 = 0.05$ y $A_1A_1 = 0.05$, $A_1A_2 = 0.05$ y $A_2A_2 = 0.90$, para los grupos 1, 2 y 3, respectivamente. Para la codificación se consideraron todas las alternativas alélicas generando dos columnas por loci. Se obtuvo una matriz $\mathbf{X}$ de orden (*45*x20) con estructura de grupo conocida que fue denominada matriz de señal, la cual fue perturbada utilizando un error aleatorio interno del 5%.

Adicionalmente, fueron generadas tres matrices de ruido externo denominadas $\mathbf{R}^{100\%}_{(45x20)}$, $\mathbf{R}^{200\%}_{(45x40)}$ y $\mathbf{R}^{300\%}_{(45x60)}$, conformadas por conjuntos de loci bialélicos suplementarios en proporciones de 100, 200 y 300% respecto a los 10 loci considerados como señal; es decir, se añadieron 10, 20 o 30 loci suplementarios al conjunto de datos original, equivalentes a 20, 40 y 60 columnas adicionales. En este caso, los loci bialélicos fueron generados utilizando una distribución uniforme (0,1). Los alelos se consideraron

presentes si el valor simulado $x_i \geq 0.5$ y ausente en el caso contrario. Este esquema de simulación produjo cuatro matrices binarias diferentes: (***i***) $\mathbf{X}^{i}_{(45x20)}$ (matriz de señal); (***ii***) $\mathbf{X}^{ii}_{(45x40)}$ (matriz de señal + 100% ruido externo), (***iii***) $\mathbf{X}^{iii}_{(45x60)}$ (matriz de señal + 200% ruido externo) y (***iv***) $\mathbf{X}^{iv}_{(45x80)}$ (matriz de señal + 300% ruido externo. La simulación fue repetida 1000 veces.

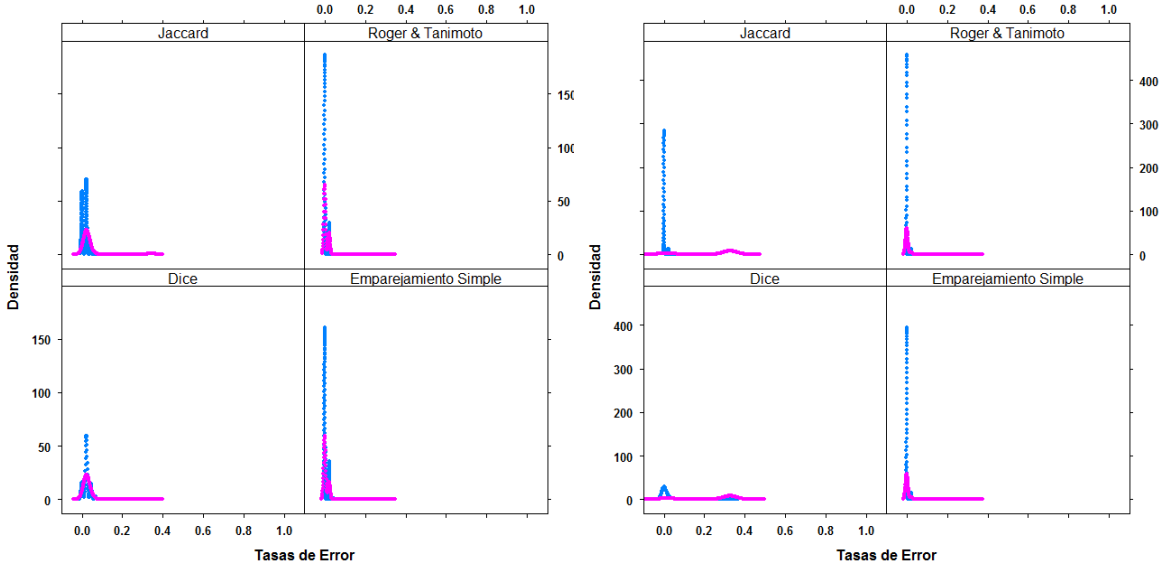
Para valorar capacidad de recuperación de información del algoritmo de agrupamiento seleccionado, con los patrones generados para las cuatro matrices binarias, se calcularon las disimilitudes utilizando los coeficientes de Jaccard, de Emparejamiento Simple, Dice y Rogers y Tanimoto **(**Sneath y Sokal, 1973). Los individuos fueron clasificados utilizando las matrices de similitud directamente y las coordenadas principales de las dos primeras dimensiones. Las tasas de error de clasificación se utilizaron para medir el error aparente al clasificar los individuos bajo las diferentes combinaciones de loci respecto a la clasificación de referencia o grupos definidos *a priori* en las simulaciones.

Los resultados de las simulaciones destacan que para las clasificaciones generadas usando las coordenadas principales de las dos primeras dimensiones e independientemente del nivel de ruido y el coeficiente de similitud, las tasas de error son menores o iguales a las obtenidas utilizando las matrices originales directamente, Figura 5. Adicionalmente, es notable cómo la utilización de las coordenadas principales para la generación de grupos, incrementa los porcentajes de clasificación correcta en la medida que el ruido incrementa.

En todos los casos estudiados se detecta un patrón de error asociado a los coeficientes que no consideran como motivo de aumento de la similaridad, la ausencia simultánea, tales como Dice y Jaccard. Siendo el Dice el que genera la mayor tasa de error de clasificación, entre otras cosas debido a la doble ponderación que hace sobre la presencia simultánea. Estos resultados corroboran los obtenidos por Chae y Warde (2006) y garantizan que la combinación de las dos técnicas, ordenación y clasificación favorece la eliminación del ruido, sin perder información relevante para la clasificación de los individuos.

$\mathbf{X}^{i}_{(45x20)}$
(matriz de señal)

$\mathbf{X}^{ii}_{(45x40)}$
(matriz de señal + 100% ruido externo)



$\mathbf{X}^{iii}_{(45x60)}$
(matriz de señal + 200% ruido externo)

$\mathbf{X}^{iv}_{(45x80)}$
(matriz de señal + 300% ruido externo)

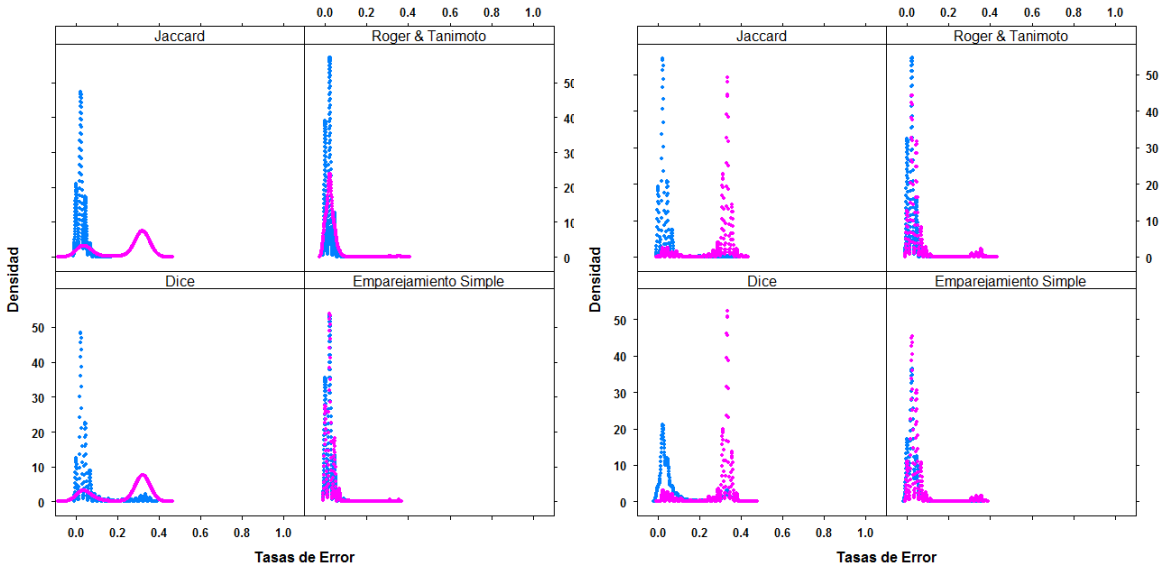


Figura 5. Distribución de las tasas de error de clasificación para las diferentes matrices binarias y coeficientes de similitud: (■) Datos originales y (■) Primeras dos coordenadas retenidas.

2.1.2 Medidas de la calidad de representación de individuos y grupos

La bondad de ajuste total puede ser considerada como una calidad de ajuste medio, -o calidad de la representación-, de los *n* individuos en la representación gráfica. Sin embargo, no todos los individuos tienen la misma calidad de representación, ya que no todos retienen la misma cantidad de información en la dimensión reducida, debido a las divergencias de los *n(n-1)/2* elementos entre las matrices de distancias observada $\boldsymbol{\Delta}$ y estimada $\mathbf{D}$.

Se considerará que un individuo está bien representado, cuando la mayoría de su información -medida a través de la variabilidad- se concentre en las *k*-dimensiones retenidas. Dado que la representación se centra en el origen, la variabilidad de un individuo estará dada por la distancia al cuadrado desde el punto que ocupe en la representación hasta el centro de la misma. Es así como, la calidad de representación será el cociente entre la distancia ajustada en la dimensión reducida y la distancia ajustada en el espacio completo, cuya expresión es:

$$CR_i^k = \left(\frac{\sum_{l=1}^{k} y_{il}^2}{\sum_{j=1}^{n-1} y_{ij}^2} \right) x100\% \qquad [2.5]$$

donde y_{ij} es la coordenada principal del *i*-ésimo individuo en la *j*-ésima dimensión. Geométricamente, debe ser interpretado como el coseno cuadrado del ángulo entre el vector en el espacio completo y su proyección en el espacio de representación (Figura 6). Así, la calidad de la representación puede ser expresada como:

$$\cos^2\theta = \left(\frac{d^2(\hat{y}_i - 0)}{d^2(y - 0)} \right)$$

donde $\cos\theta_{12} = \cos\theta_1 + \cos\theta_2$ y $\cos\theta_1 + \cos\theta_2 + \cos\theta_3 = 1$, pudiéndose derivar la importancia relativa de cada individuo de la expresión anterior.

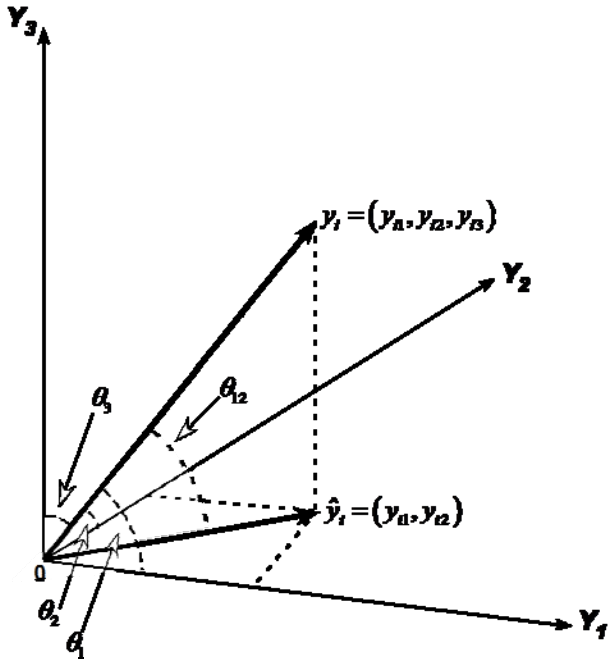


Figura 6. Interpretación geométrica de la calidad de representación del *i*-ésimo individuo.

Para los grupos obtenidos por cualquier técnica de clasificación, la calidad de la representación es calculada en forma similar pero y_{ij} es reemplazada por $\overline{y}_{gj}$ la media de las coordenadas en la *j*-ésima dimensión para el grupo *g*.

2.1.3 Variabilidad muestral

Este apartado se enfoca sobre la premisa de que los resultados de cualquier análisis de datos no está completo si no ofrece información sobre la estabilidad de la solución. En este sentido, Gifi (1990) menciona que existen varios tipos estabilidad o sensibilidad y diferentes enfoques para su análisis. No obstante, la mayor parte de las técnicas que permiten medir la estabilidad o la sensibilidad de los análisis tienen en común que estudian el efecto que tiene una pequeña perturbación de los datos sobre la solución o partes de ésta. En este sentido, se considerará que un método es estable para una aplicación en particular, si pequeños cambios en los datos producen solamente pequeños cambios en la solución. Por ejemplo, en el análisis estadístico clásico el cálculo de errores estándar o intervalos de confianza, entre otros, pertenece a una clase particular de análisis de sensibilidad, que estudia las perturbaciones inducidas por el muestreo al azar.

En el caso del Análisis de Coordenadas Principales (ACoP), los estudios sobre la estabilidad de la soluciones han recibido poca o ninguna atención. Aún a pesar de que se han difundido en los últimos años dentro de las técnicas multivariantes, tales como: el Análisis de Conglomerados (AC), el Análisis de Componentes Principales (ACP), el Análisis de Correspondencias Simple o Múltiple (ACS, ACM), y el Escalamiento Multidimensional (EM), entre otros. Es así como, en la actualidad, es casi imposible ver una publicación que incluya un dendrograma derivado de algún algoritmo de encadenamiento donde no aparezcan valores que nos indiquen la estabilidad de los nodos de la representación.

CAPITULO II

En el contexto de las técnicas exploratorias que implican la Descomposición en Valores Singulares (DVS), Greenacre (1993) considera que son detectables dos tipos de estabilidad. La externa que mide si los datos representan a la población bajo estudio y solo es posible cuantificarla si se dispone de muestreos sucesivos de una población, y la interna que mide la calidad y estabilidad de los resultados del análisis, la cual es afectada por la selección, tipo, unidad de medida y peso de las variables, así como por los errores derivados de los métodos de medición. En el caso de estudios de diversidad genética la estabilidad externa no requiere atención, por lo que se centrará la atención en el estudio de la estabilidad interna o como hemos titulado este aparte, en la variabilidad muestral.

La variabilidad muestral, se estudia a través de la medición de las alteraciones o perturbaciones que producen modificaciones sobre la matriz $\mathbf{X}$ de orden (nxp) o sus transformaciones; es decir, el procedimiento consiste en eliminar un elemento, cambiarlo, o simular errores en los datos para comprobar o verificar la estabilidad respecto a una configuración inicial. Abascal-Fernández y Landaluce-Calvo (2002), señalan que las alteraciones que se pueden producir sobre la matriz $\mathbf{X}$ pueden ser: (***i***) el agrupamiento de varias columnas; (***ii***) la elección de las variables que definen el problema; (***iii***) la codificación o definición de las variables; (***iv***) la eliminación de filas y (***v***) la alteración de datos mediante la suma de errores aleatorios o la introducción de errores de medidas en las variables.

CAPITULO II

Los métodos de remuestreo tales como el Jackknife (Tukey, 1958) y más específicamente el Bootstrap (Efron y Tibshirani, 1993) han demostrado su potencialidad para determinar la estabilidad de las configuraciones generadas por las técnicas que involucren la Descomposición en Valores Singulares (DVS) (Greenacre, 1984; Meulman, 1984; Leeuw y Meulman, 1986; Ringrose, 1992; Reiczigel, 1996; Lebart *et al.*, 2000; Tan *et al.*, 2004; Lebart, 2004 y 2007). Estos métodos permiten la construcción de regiones de confianza para los *n* elementos o variables representadas en los ejes principales sin el conocimiento previo de su distribución y donde el modelaje paramétrico o el análisis tradicional o bien no son aplicables o no son fiables (Milan y Whittaker, 1995; Krzanowski, 2006; Lebart, 2007). Es decir, en ausencia de cualquier información respecto a la distribución poblacional, la distribución de los valores encontrados en una muestra aleatoria constituye la mejor orientación en cuanto a la distribución de esa población, al no utilizar más que los valores observados en la muestra; por esta razón, estos métodos son considerados como autosuficientes.

En el Análisis de Coordenadas Principales (ACoP) donde el objetivo principal es la representación de *n* puntos en un espacio de dimensión reducida, y donde además, a diferencia de otras técnicas de reducción de la dimensionalidad, los planos principales no contienen información sobre las variables, no es posible o es ilógico generar medidas de estabilidad para variables y ejes, ya que estas medidas no revisten importancia por carecer de interpretación. En este orden, solo nos ocuparemos de cómo cuantificar la estabilidad de las representaciones a través de métodos que permitan detectar la variabilidad muestral de los *n* individuos o grupos de individuos, esta última utilizando

la misma idea geométrica empleada para el cálculo de la calidad de la representación de grupos referidos en el apartado anterior.

En este sentido, las alteraciones o perturbaciones que deben introducirse para producir modificaciones sobre la matriz $\mathbf{X}$, o sus transformaciones, que permitirán medir la estabilidad del Análisis de Coordenadas Principales (ACoP) se harán sobre los individuos. Este enfoque de alteraciones o perturbaciones sobre individuos tiene la ventaja que no necesita suponer independencia entre columnas. Recordemos, que en el caso de los árboles de consenso popularmente utilizados en estudios de diversidad genética utilizando marcadores moleculares, las técnicas de remuestreo utilizadas para medir estabilidad suponen independencia de los alelos (columnas), que son los remuestreados, supuesto que con excepción de datos particulares no puede ser satisfecho.

Según el propósito y grado de complejidad del análisis se pueden utilizar varios tipos de técnicas de remuestreo para determinar la calidad de las visualizaciones obtenidas, entre otras podemos mencionar: Jackknife, Bootstrap parcial, Bootstrap total y Bootstrap específico (Leeuw y Meulman, 1986; Greenacre, 1984; Lebart, 2004 y 2007). De estas técnicas el Bootstrap total, representa la mejor estrategia metodológica cuando se trata es de construir regiones de confianza para los n elementos. Consiste en generar B muestras aleatorias de matrices de disimilitud a partir de la matriz original, haciendo el muestreo sobre los individuos. Esta estrategia puede aplicarse porque puede inferirse la geometría de los individuos sobre los ejes principales, como es el caso del Análisis de

Coordenadas Principales (ACoP), donde el objetivo es la representación de *n* puntos. No obstante, esta técnica de remuestreo tiene la desventaja que cada nueva matriz puede generar planos principales con direcciones diferentes, con el inconveniente adicional de que estos son impredecibles de una muestra a otra y subyacen de la técnica de muestreo en sí misma.

Para corregir estos cambios y poder comparar las nuevas configuraciones generadas con respecto a la configuración original, es necesario realizar un conjunto de transformaciones sobre los B conjuntos de coordenadas. Lebart (2007) indica que se pueden realizar tres tipos de transformaciones que dan origen a tres pruebas diferentes para evaluar la estabilidad de la estructura observada y las denomina Bootstrap total tipo 1, 2 y 3, respectivamente.

La primera transformación, conservadora, permite corregir las reflexiones de los ejes haciendo cambio de signos -cuando sea necesario- en las nuevas coordenadas encontradas para cada muestra. El procedimiento es sencillo y consiste en hacer el producto escalar entre los ejes originales y sus replicas homologas. La segunda transformación, menos conservadora o corrección para los posibles intercambios de ejes, permite solo la validación de los considerados como variables latentes, hace transformaciones asignando secuencialmente los ejes originales a los ejes derivados del remuestreo donde tengan correlación máxima. Y por último, la tercera transformación, o transformación Procrustes, consiste en superponer tanto como sea posible la configuración original y las generadas del muestreo. El procedimiento traslada, rota y el

re-escala los ejes principales de las configuraciones de las muestras hasta generar un consenso entre éstas y la original, puede realizarse de a pares o de manera general (Gower y Dijksterhuis, 2004). Detalle de la transformación Procrustes será presentado en el Capítulo IV. Greenacre (1984) indica que en los casos donde las distancias entre individuos tienen significado tanto absoluto como relativo como en el escalamiento multidimensional es preferible omitir el re-escalamiento.

Las otras técnicas tales como el Jackknife, Bootstrap parcial y Bootstrap específico, no se han detallado porque su aplicabilidad al Análisis de Coordenadas Principales (ACoP) y más específicamente en el contexto de estudios de diversidad genética tiene algunas desventajas. El Jackknife, presenta limitaciones en el caso de la construcción de intervalos de confianza ya que suponen distribución idéntica e independencia de los pseudovalores y en la mayoría de los casos de matrices de datos de marcadores las columnas o alelos no pueden ser consideradas como *n* variables independientes e idénticamente distribuidas. Adicionalmente, en la mayoría de los casos la aplicación del Jackknife da lugar a estimadores con menor sesgo y que en algunos casos, tienen menor varianza o al menos menor error, por lo que siempre se encuentran en ventaja sobre los otros métodos. Greenacre (1984) indica que si hay grandes conjuntos de datos, es necesario primero construir grupos de éstos y después investigar la estabilidad con respecto a la omisión de alguno. En cuanto al Bootstrap parcial, aunque presenta menos complejidad y al igual que el Bootstrap total tipo 2, permite validar la estabilidad de las configuraciones a través de los planos principales, tiene el problema que causa un efecto de expansión en las coordenadas obtenidas debido al aumento de la inercia total

en las configuraciones muestrales. En el caso del Bootstrap específico, aunque las aplicaciones revisadas están referidas a datos textuales, se destaca de la publicación de Lebart (2007), que su aplicación es adecuada en el caso de tener respuestas multinivel.

En resumen, consideraremos dos técnicas de remuestreo para construir de regiones de confianza para los *n* elementos y medir la calidad de las visualizaciones obtenidas en el Análisis de Coordenadas Principales (ACoP). El Bootstrap total tipo 1, que permite la validación de estructuras estables y robustas y supone que cada muestra generará configuraciones de igual rango a la original y el Bootstrap total tipo 3, que permite la validación del subespacio completo y además cuantifica la proximidad de las configuraciones.

Otra estrategia que permite generar las alteraciones o perturbaciones, consiste en realizar permutaciones, que producen variaciones en el orden o la disposición del número de elementos la matriz $\mathbf{X}$ o de sus transformaciones. Este procedimiento, en combinación con cualquiera de las alternativas de transformación propuestas, permite generar configuraciones de consenso capaces de reproducir la estructura original y producir elipses de confianza para cada uno de los *n* individuos.

2.1.3.1 Formulación

Sea $\Delta = (\delta_{ij})$ la matriz de distancia euclídea que ha sido derivada de alguna de las transformaciones siguientes: $\delta_{ij} = \sqrt{1 - s_{ij}}$ o $\delta_{ij} = \sqrt{s_{ii} - 2s_{ij} + s_{jj}}$, donde $\mathbf{S}_{nxn} = (s_{ij})$ es la matriz simétrica que contiene las similitudes entre los *n* individuos obtenida de la

matriz de datos $\mathbf{X}$ utilizando cualesquiera de los coeficientes descritos según sea el caso. Sea $\mathbf{Y}$ la matriz que contiene el conjunto de puntos o coordenadas principales que representan la mejor aproximación *k*-dimensional de los individuos en el espacio euclidiano. Para estudiar la variabilidad muestral de los los *n* individuos de la representación obtenida vía Análisis de Coordenadas Principales (ACoP), deberemos generar tantas veces como sea necesario matrices $\Delta = (\delta_{ij})$ que permitan la obtención de B configuraciones de la matriz $\mathbf{Y}$. Dos tipos de enfoques pueden ser utilizados para generar las B configuraciones de la matriz $\mathbf{Y}$, para producir alteraciones sobre los individuos o sobre los residuales.

En este punto es necesario aclarar, con el objeto de no confundir al lector con la terminología empleada, que denominaremos "*remuestreo sobre*" o "*permutación aleatoria sobre*" a la estrategia de alteración sobre los *n* individuos o residuales y "*método de transformación"* a la estrategia de corrección de los cambios que permitan comparar las configuraciones generadas por las alteraciones con la configuración original o de referencia.

Algoritmo sobre los individuos

Se realizan B remuestreos con reemplazamiento sobre los individuos de la matriz $\mathbf{X}$ generando B nuevas matrices de distancia que denominaremos Δ_i^*, $i=1,\ldots, B$, estas matrices de distancia permiten representaciones *k*-dimensionales de los individuos en el espacio euclidiano y las denominaremos $\mathbf{Y}^*_{i(-m)}$, el subíndice (-*m*), nos indica que existirán individuos que debido a la naturaleza del remuestreo no serán seleccionados,

por lo que no todas las parejas de las distancias originales estarán contenidas en la matriz Δ_i^*. Es así como, en el espacio *k*-dimensional generado por $\mathbf{Y}_{i(-m)}^*$, no estarán representados todos los *n* individuos, por lo que ésta no es comparable con la configuración original $\mathbf{Y}$ y que para fines del algoritmo denominaremos como $\mathbf{Y}^0$. Los individuos que por azar no han sido muestreados son incluidos en la configuración utilizando la *"add-a-point formula"*. Este método fue desarrollado en el contexto del escalamiento multidimensional (Gower, 1966), y permite actualizar la configuración $\mathbf{Y}_{i(-m)}^*$, obteniendo una nueva configuración $\mathbf{Y}_i^*$, que es comparable con $\mathbf{Y}^0$. Este procedimiento ha sido usado en ajustes de Biplots no lineales y generalizados (Gower y Harding, 1988; Gower, 1992), en análisis de regresión basado en distancias (Cuadras y Arenas, 1990; Cuadras y Fortiana, 1995) y en análisis canónico (Krzanowski, 1994). Sobre las configuraciones $\mathbf{Y}_i^*$ son aplicadas cualesquiera de las transformaciones propuestas para la validación del espacio, tales como el Bootstrap total tipo 1 y el Bootstrap total tipo 3, que de ahora en adelante, denominaremos reflexión y transformación Procrustes, respectivamente, Figura 7a.

Algoritmo sobre los residuales

En este procedimiento se supone que $\mathbf{\Delta}$ puede ser descompuesta en $\mathbf{\Delta} = \hat{\mathbf{\Delta}} + \mathbf{E}$, donde $\mathbf{E}$, es una matriz de residuales con las mismas propiedades de $\mathbf{\Delta}$. Remuestreando B veces los $n(n-1)/2$ elementos fuera de la diagonal de la matriz, son generadas las replicas $\Delta_i^* = \hat{\Delta} + \mathbf{E}_i^*$, *i=1,…, B*. Con estas matrices y siguiendo el procedimiento descrito previamente, se generan las nuevas configuraciones $\mathbf{Y}_i^*$ que son

posteriormente comparadas con $\mathbf{Y}^0$. A diferencia del algoritmo sobre los individuos este procedimiento no necesita actualizar las configuraciones utilizando la *"add-a-point formula"*, Figura 7b. Los procedimientos de remuestreo sobre los residuales han sido empleados para hacer estimaciones del coeficiente de determinación (Efron y Tibshirani, 1993) y en la estimación de parámetros de modelos bilineales que incorporan la Descomposición en Valores Singulares (DVS) (Milan y Whittaker, 1995).

Otro procedimiento que permite medir la variabilidad muestral consiste en sustituir el remuestreo sobre la matriz $\mathbf{E}$ por perturbaciones aleatorias de los $n(n-1)/2$ elementos, similar al anteriormente descrito para los residuales, que permite la generación de $\mathbf{Y}_i^*$ nuevas configuraciones que son posteriormente comparadas con $\mathbf{Y}^0$, Figura 7b.

Aunque la distribución de los residuales no es independiente y las condiciones de ortogonalidad de cualquier técnica que implique la Descomposición en Valores Singulares (DVS) causan las distorsiones en las regiones de confianza, consideramos que las técnicas revisadas y las propuestas en este trabajo reducen sustancialmente las distorsiones derivadas del muestreo y permiten calcular la estabilidad de las configuraciones con relativa fiabilidad.

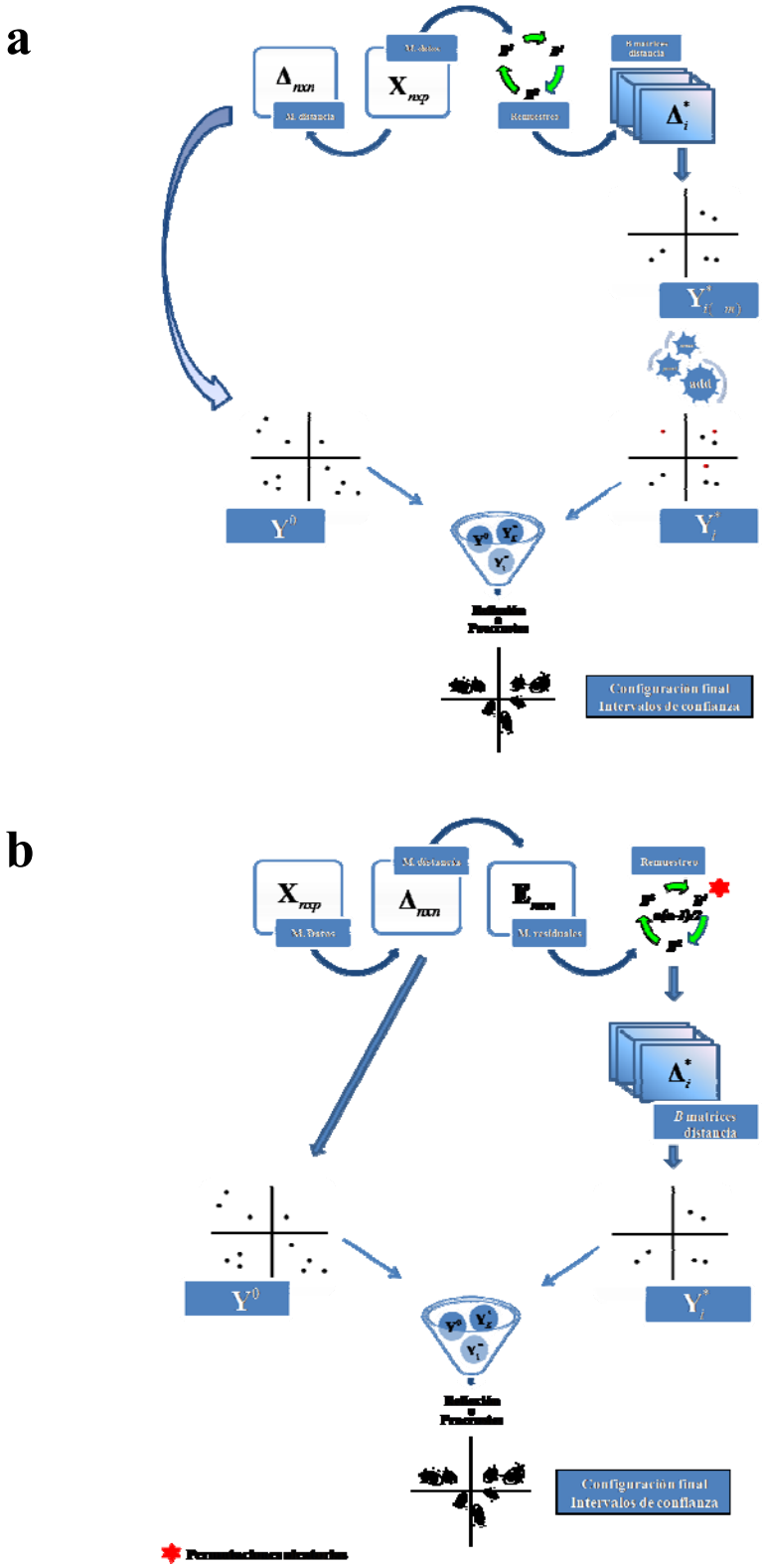


Figura 7. Algoritmo para el cálculo de la estabilidad y la construcción de regiones de confianza en un Análisis de Coordenadas Principales (ACoP). (**a**) remuestreo sobre los individuos y (**b**) remuestreo sobre los residuales

De los algoritmos considerados, son reconocibles seis rutinas metodológicas que pueden ser utilizadas para estudiar la variabilidad muestral, a saber:

(i) remuestreo sobre los *n* individuos de la matriz **X** y transformar las configuraciones obtenidas utilizando el método de reflexión;

(ii) remuestreo sobre los *n* individuos de la matriz **X** y transformar las configuraciones obtenidas utilizando el método de Procrustes;

(iii) remuestreo sobre los $n(n-1)/2$ elementos fuera de la diagonal de la matriz **E** (remuestreo sobre los residuales) y transformar las configuraciones obtenidas utilizando el método de reflexión;

(iv) remuestreo sobre los $n(n-1)/2$ elementos fuera de la diagonal de la matriz **E** (remuestreo sobre los residuales) y transformar las configuraciones obtenidas utilizando el método de Procrustes;

(v) permutación aleatoria sobre los $n(n-1)/2$ elementos fuera de la diagonal de la matriz **E** (permutación aleatoria sobre los residuales) y transformar las configuraciones obtenidas utilizando el método de reflexión;

(vi) permutación aleatoria sobre los $n(n-1)/2$ elementos fuera de la diagonal de la matriz $\mathbf{E}$ (permutación aleatoria sobre los residuales) y transformar las configuraciones obtenidas utilizando el método de Procrustes.

En cada caso es posible calcular la estabilidad de la forma:

$$\mathbf{EST} = 1 - \frac{\sum_{b=1}^{B}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\mathbf{Y}_{ij}^{*} - \mathbf{Y}_{ij}^{0}\right)^{2}}{\sum_{b=1}^{B}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\mathbf{Y}_{ij}^{*}\right)^{2}} \qquad [2.6]$$

donde $\mathbf{Y}_{ij}^{*}$ es la representación para cada punto en las configuraciones de los diferentes remuestreos y $\mathbf{Y}_{ij}^{0}$ es la configuración inicial. **EST** puede ser interpretada como la proporción de variabilidad asociada a la capacidad del método de análisis para reproducir la configuración original. El numerador de la fracción de la ecuación 2.6 tiende a cero cuando la representación resultante de las perturbaciones ejercidas sobre los datos o sobre los residuales reproduce la configuración original y por lo tanto la EST tiende a uno (su valor máximo). De la misma forma, cuando el numerador de la ecuación 2.6 aumenta, es decir que la perturbación produce una variabilidad del mismo orden que la de los datos mismos, generando una representación muy diferente de la original, la fracción tiende a uno y la EST tiende a cero (su valor teórico mínimo).

Habiendo esbozado la importancia de conocer la estabilidad de las soluciones y las herramientas metodológicas disponibles para su estudio, a continuación ilustraremos el comportamiento de la estabilidad de las seis rutinas metodológicas consideradas para el cálculo de la variabilidad muestral.

2.1.3.2 Estudio de simulación

Con este propósito, fue utilizada la matriz binaria descrita en el apartado 2.1.1, $\mathbf{X}^{iii}_{(45x60)}$ -matriz de señal + 200% ruido externo- considerando diferentes combinaciones de los coeficientes de similitud de Jaccard, de Emparejamiento Simple, Dice y Rogers y Tanimoto **(**Sneath y Sokal, 1973); dos y tres dimensiones en la generación de la configuración $\mathbf{Y}_i^*$ (2D y 3D) y dos alternativas de la matriz $\mathbf{Y}^0$; $\mathbf{Y}^{0i}$ o configuración original -cuando utilizamos la configuración que se produce sin ninguna perturbación- y $\mathbf{Y}^{0c}$ o configuración consenso -cuando utilizamos la configuración que se produce del consenso de todas las $\mathbf{Y}_i^*$ configuraciones, resultando un total de 96 escenarios, Figura 8. El procedimiento fue repetido 1000 veces y en cada uno se realizaron 100 remuestreos o perturbaciones $B = 100$.

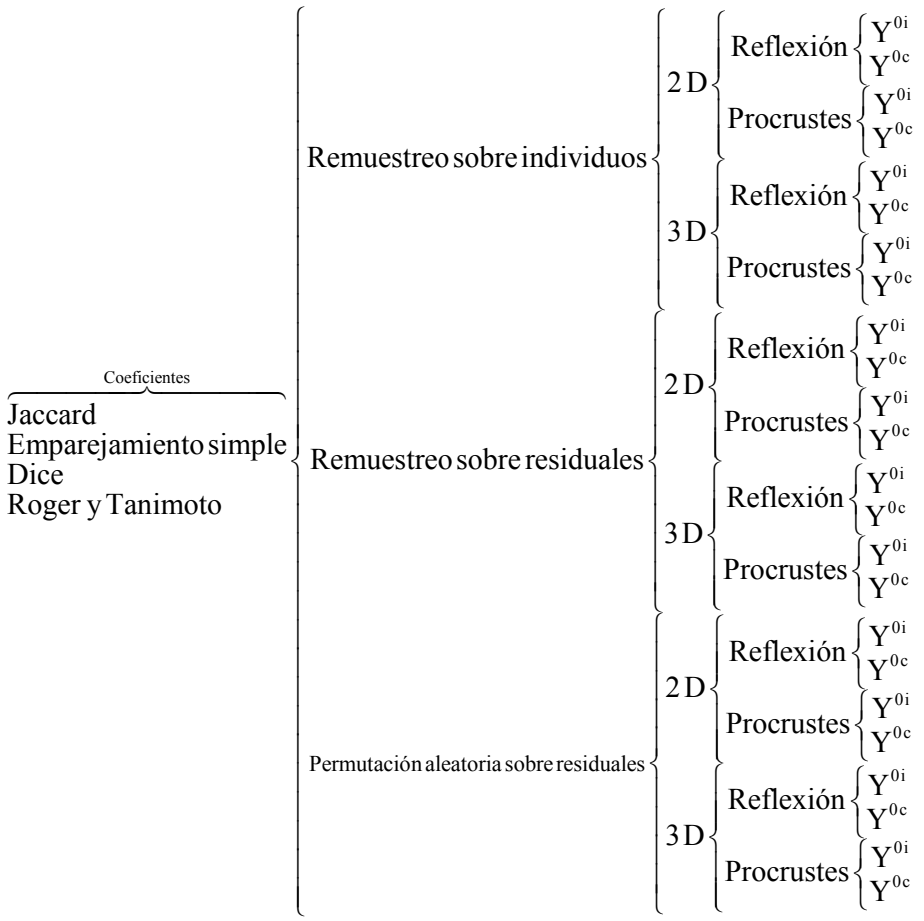


Figura 8. Escenarios utilizados para estudiar el comportamiento de la estabilidad del Análisis de Coordenadas Principales (ACoP).

La Figura 9, muestra la distribución de la estabilidad (EST) para los 96 escenarios estudiados, los coeficientes de similitud probados no presentan diferencias que puedan ser consideradas importantes o que indiquen que se producen cambios en la solución o parte de esta. Sin embargo, el promedio general para el efecto coeficiente fue superior para el de Roger y Tanimoto (92.19%) que para los coeficientes de Dice (90.94%), Emparejamientos simple (90.55%) o Jaccard (90.27%) que son casi iguales. En cualquier caso los valores de estabilidad fueron superiores al 75%, 85% y 90% en por lo

menos 95%, 70% y 50% de las 24 combinaciones de coeficientes valoradas, respectivamente.

Respecto al efecto estrategia de alteración se aprecian dos grupos, el primero y de más baja estabilidad (86.95%) formado por el remuestreo sobre los residuales, y el segundo con una estabilidad superior al 92% formado por remuestreo sobre los *n* individuos y la permutación aleatoria de residuales.

En los dos tipos de dimensiones utilizadas (2D y 3D) para la generación de la configuración $\mathbf{Y}_i^*$, no se aprecian diferencias. Este resultado contribuye a corroborar la hipótesis de que dos dimensiones suelen ser suficientes para lograr buenas representaciones de los individuos, que hemos desarrollado en apartados anteriores, aunque, debemos recordar que si se utilizaran más grupos probablemente necesitaríamos más dimensiones.

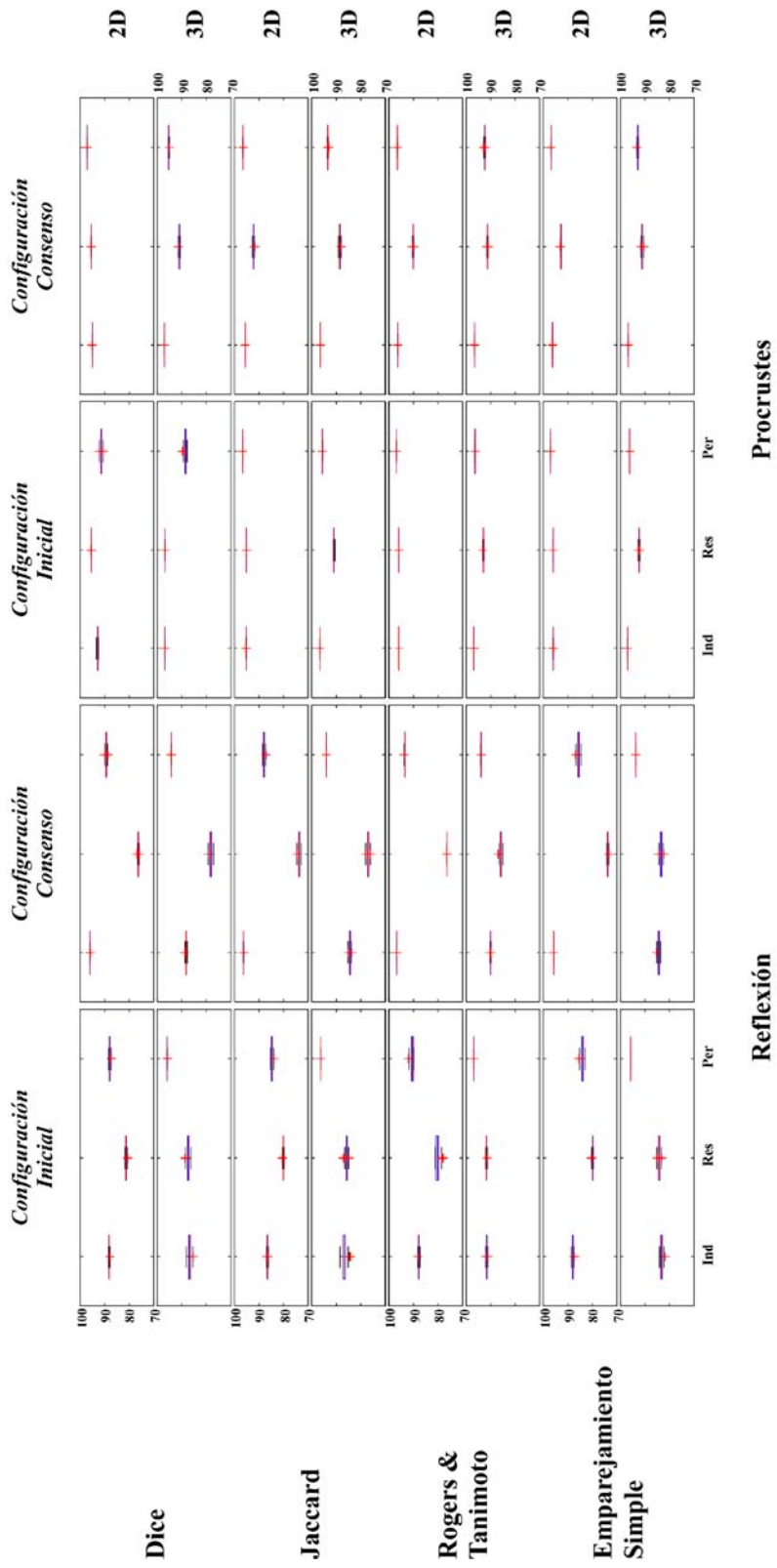


Figura 9. Coplot de estabilidad para los diferentes escenarios estudiados

En relación a los métodos de transformación, el método de Procrustes muestra los mejores resultados de estabilidad siendo en la mayoría de los casos superior al 90% y tanto este método como el de reflexión no son sensibles al efecto configuración de referencia.

A pesar de que hemos discutido los efectos en forma separada -solo para proveer una visión general sobre su comportamiento-, existe interacción y aunque el objetivo de este ejercicio no es recomendar ningún escenario, es evidente que, de las seis rutinas metodológicas, la que realiza el remuestreo sobre los residuales y la transformación utilizando el método de reflexión produce las configuraciones menos estables independientemente del coeficiente de similitud, de la dimensión que se retenga o de la configuración de referencia que se use para hacer la comparación. Esto sugiere que las configuraciones generadas del remuestreo sobre los residuales no son estables y que la sola reflexión de los ejes no corrige los cambios para hacer las configuraciones comparables. En cualquier caso, los mejores resultados se obtienen utilizando la transformación debida al método de Procrustes, en el siguiente orden decreciente: remuestreo sobre los residuales, perturbaciones aleatorias sobre los residuales y remuestreo sobre los individuos.

De manera general podemos resumir que lo que más afecta la estabilidad es el método que se utiliza para transformar las configuraciones generadas del remuestreo o permutación antes de ser comparadas con la configuración de referencia. Realizar

solamente la reflexión de las configuraciones no solo no permite validar el subespacio completo y sino que aumenta la distancia entre configuraciones.

Las Figuras 10-15 ilustran, a modo de ejemplo, el análisis detallado de la variabilidad muestral para las seis rutinas metodológicas, el coeficiente de similitud de Dice, y dos dimensiones (2D) para la generación de la configuraciones $\mathbf{Y}_i^*$ y $\mathbf{Y}^{0i}$ como configuración de referencia.

Aunque en algunos casos la descripción de la representación gráfica no coincida con los resultados obtenidos para el análisis de estabilidad, el análisis de datos gráficos permite determinar la calidad de las visualizaciones y ofrece una mejor aproximación de cómo interactúan la estrategias de alteración de los datos y los métodos de transformación. Este enfoque intuitivo permite además comparar las rutinas valoradas a través de la observación de la distribución de autovalores, comparar la magnitud de los sesgos entre la configuración inicial y la $\mathbf{Y}_i^*$ y las regiones de confianza proyectadas para cada individuo en los ejes principales.

Es así como, si comparamos los gráficos de sedimentación (Figuras 10a-15a) podemos observar que a partir del quinto autovalor la amplitud de los intervalos es mayor cuando la alteración se hace sobre los residuales, independientemente de si por remuestreo o permutación aleatoria y método de transformación. En el caso de las dos primeros autovalores, los resultados son contrapuestos, ya que la amplitud del intervalo de confianza generado por los autovalores de las $\mathbf{Y}_i^*$ configuraciones es mayor cuando el

remuestreo es sobre los *n* individuos de la matriz $\mathbf{X}$ (Figuras 10b-15b), sin embargo, esta estrategia de alteración produce estimaciones menos sesgadas respecto a la configuración original (Figuras 10c-15c). Nótese que cuando se hace el remuestreo sobre los residuales o se permutan aleatoriamente, el valor original del autovalor (línea roja) se aleja considerablemente del histograma generado por los autovalores de las $\mathbf{Y}_i^*$ configuraciones (la línea verde representa la media), siendo esta diferencia de mayor magnitud para el segundo auto valor (Figuras 10d-15d). En ningún caso se detecta efecto método de transformación. En otras palabras, las perturbaciones sobre los residuales independientemente del método de transformación producen un efecto de aumento sobre las inercias.

El grado de efecto de las rutinas sobre la calidad de la estimación para cada coordenada por individuo se muestra en las Figuras 10e-15e y 10f-15f, para el primero y segundo eje, respectivamente. Si se utiliza la proporción de veces que el valor original -línea roja en el histograma- coincide aproximadamente con el valor medio estimado de las $\mathbf{Y}_i^*$ configuraciones -línea verde en el histograma-, en el caso de la primera coordenada se obtienen resultados similares a los presentados para el primer autovalor, es decir menor sesgo cuando se realiza el remuestreo sobre los individuos. Adicionalmente, se observa que existe una clara interacción con los métodos de transformación, siendo que, en el remuestreo sobre los individuos el método de reflexión de los ejes tiene mejor comportamiento y cuando la perturbación se hace sobre los residuales la mejor transformación es la debida al método de Procrustes. Para la segunda coordenada la

situación es inversa aunque se mantiene el tipo de interacción entre métodos de perturbación y transformación.

A pesar del comportamiento respecto a la configuración inicial, cuando observamos las elipses de confianza que delinean la variabilidad muestral en el plano, Figuras 10f-15f, el remuestreo sobre los individuos es el que produce las peores visualizaciones, mostrando elipses de mayor tamaño asociadas por lo tanto a mayor variabilidad. Las rutinas que utilizan el remuestreo o permutación aleatoria sobre los residuales ofrecen resultados considerablemente superiores, independientemente del método de transformación; no obstante, la transformación Procrustes es la que delinea elipses de menor tamaño. Nótese que, en todos los casos el eje de mayor trazo de las elipses está en el mismo sentido de la segunda coordenada principal, geometría asociada a que el primer eje recoge la mayor parte de la variabilidad. Otro aspecto importante a destacar es que la magnitud de las elipses en el caso del remuestreo sobre los individuos coincide con los resultados presentados por Greenacre (1994) y Lebart (2007) en el contexto de análisis que involucran la Descomposición en Valores Singulares (DVS).

Estos resultados, aunque parezcan contradictorios se sustentan en el hecho de que aunque las configuraciones generadas por perturbaciones de los residuales son mas sesgadas, la distribución de las coordenadas de las $\mathbf{Y}_i^*$ configuraciones son menos dispersas.

Estas interpretaciones son corroboradas cuando se proyecta la variabilidad de los individuos agregando la tercera dimensión, Figura 16. La adición del tercer eje permite observar el volumen que forma cada punto o individuo, haciéndose más evidentes las diferencias entre las rutinas estudiadas. Así mismo, se facilita la separación en términos de magnitud entre solo hacer la reflexión de los ejes o hacer la transformación Procrustes.

El estudio empírico presentado no pretende ofrecer una recomendación sobre cuál debe ser el mejor método para evaluar la variabilidad en el Análisis de Coordenadas Principales (ACoP), ya que consideramos que debe probarse una mayor cantidad de escenarios que incluyan la adición de grupos, número de individuos y alternativas alélicas. Hemos presentado un estudio comparativo que cuantifica cómo pequeños cambios o alteraciones en los datos afectan las soluciones y hemos desarrollado una metodología que permite profundizar en el estudio de la variabilidad muestral del Análisis de Coordenadas Principales (ACoP) del cual existen pocas o ninguna referencia, considerando que éste es uno de los aportes fundamentales de este trabajo.

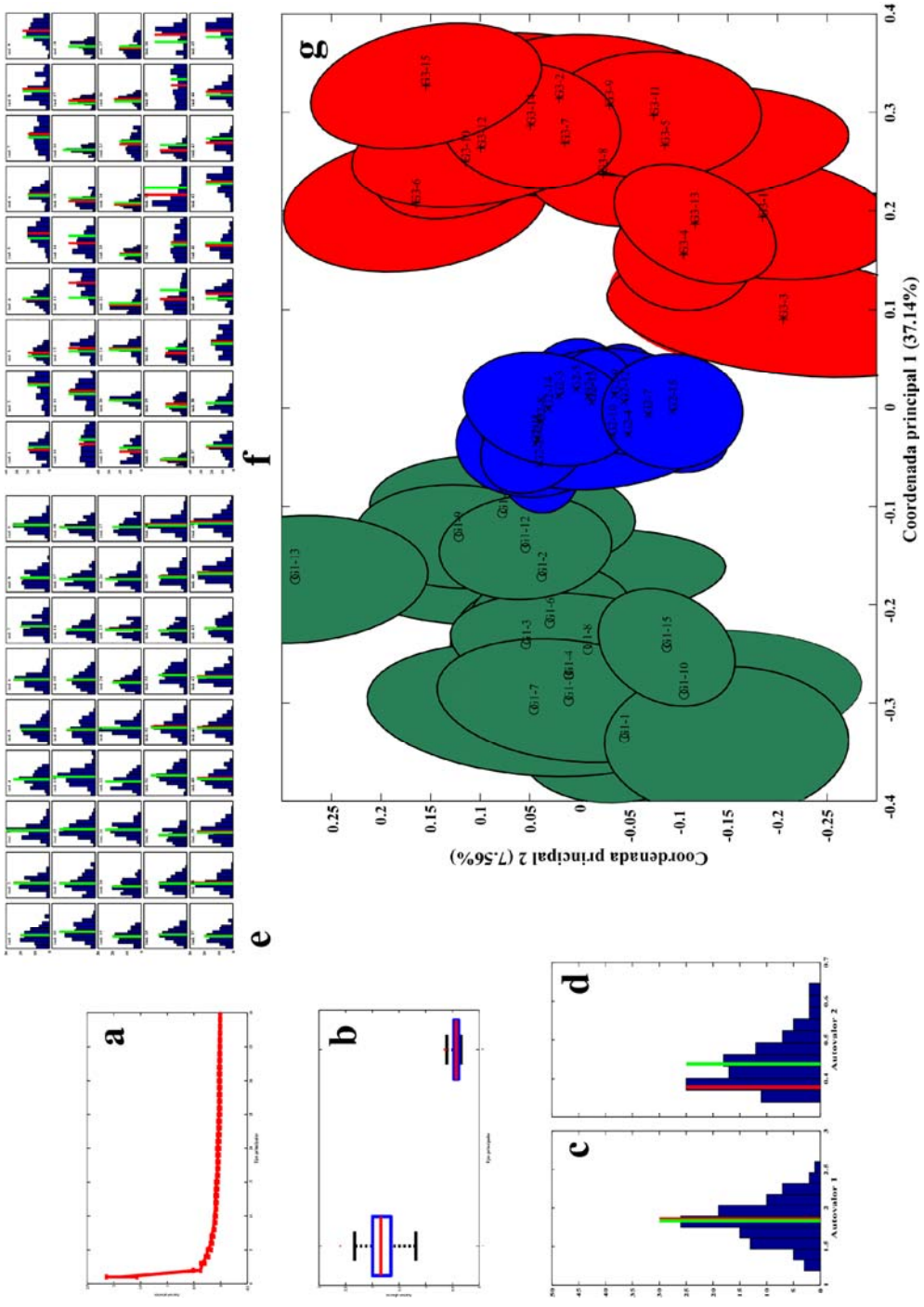


Figura 10. Variabilidad muestral de autovalores e individuos basada en el coeficiente de Dice, dos dimensiones, remuestreo sobre los individuos, transformación a través del método de reflexión y $\mathbf{Y}^{0i}$ como configuración de referencia.

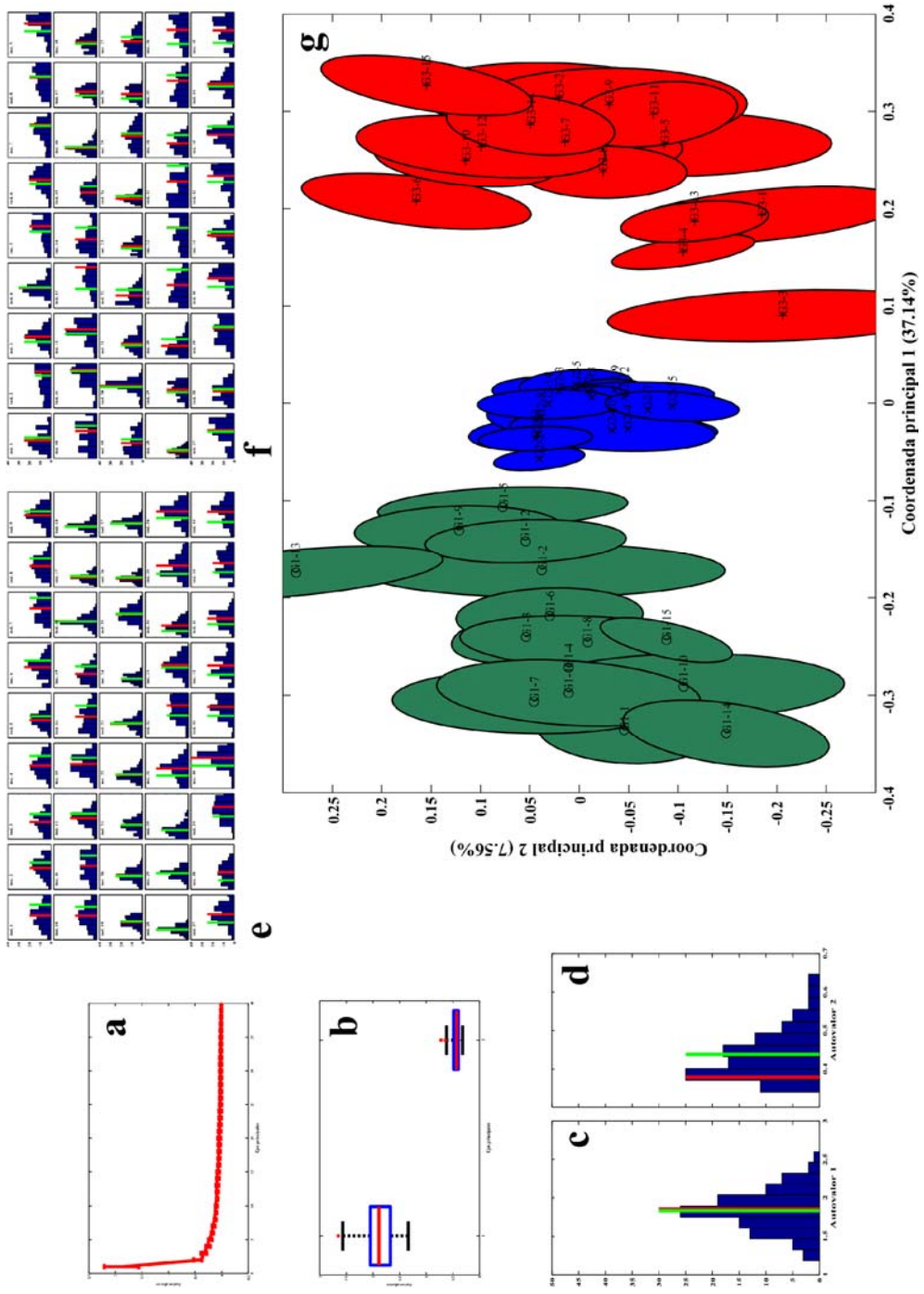


Figura 11. Variabilidad muestral de autovalores e individuos basada en el coeficiente de Dice, dos dimensiones, remuestreo sobre los individuos, transformación a través del método de Procrustes y $\mathbf{Y}^{oi}$ como configuración de referencia.

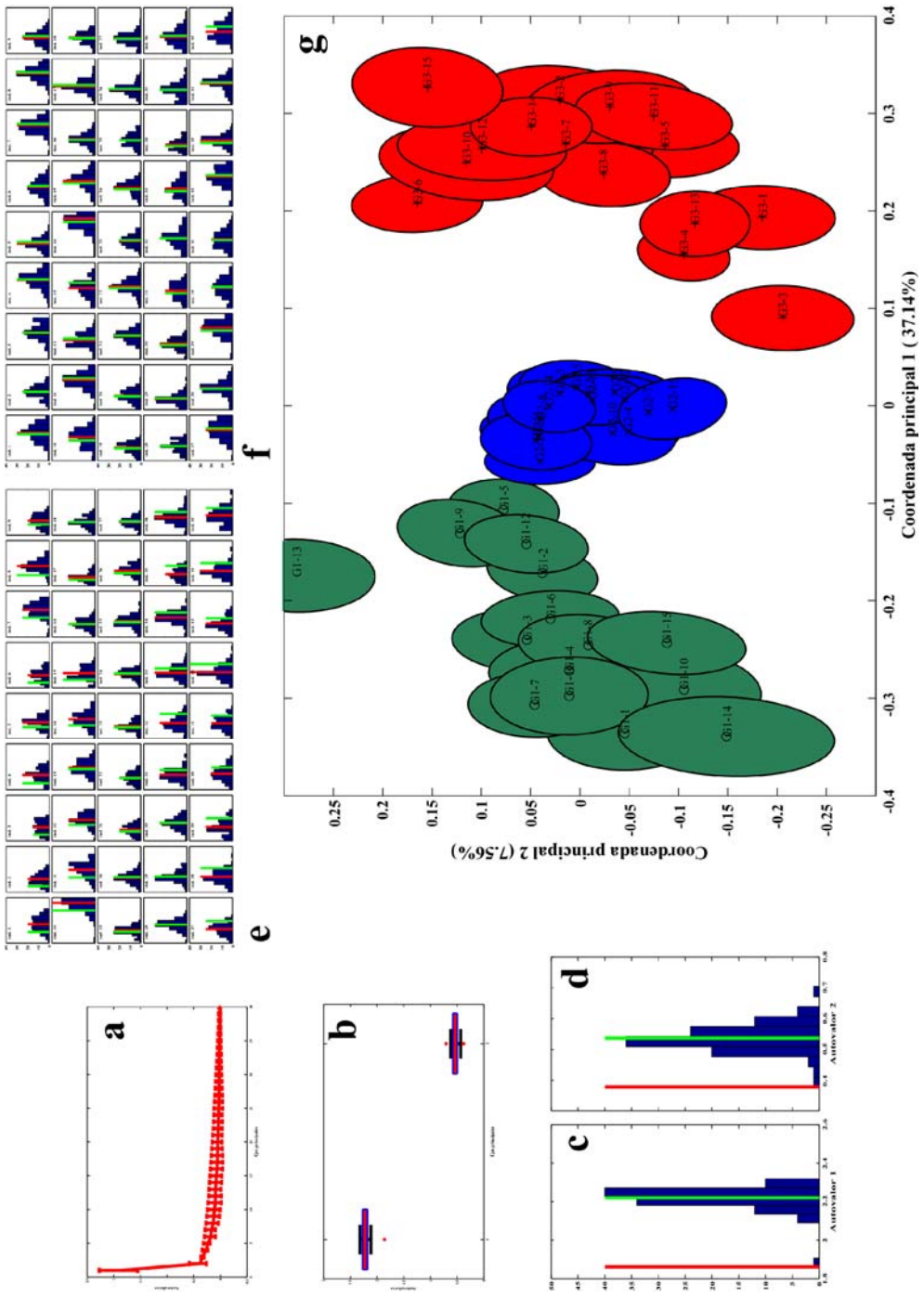


Figura 12. Variabilidad muestral de autovalores e individuos basada en el coeficiente de Dice, dos dimensiones, remuestreo sobre los residuales, transformación a través del método de reflexión y $\mathbf{Y}^{oi}$ como configuración de referencia.

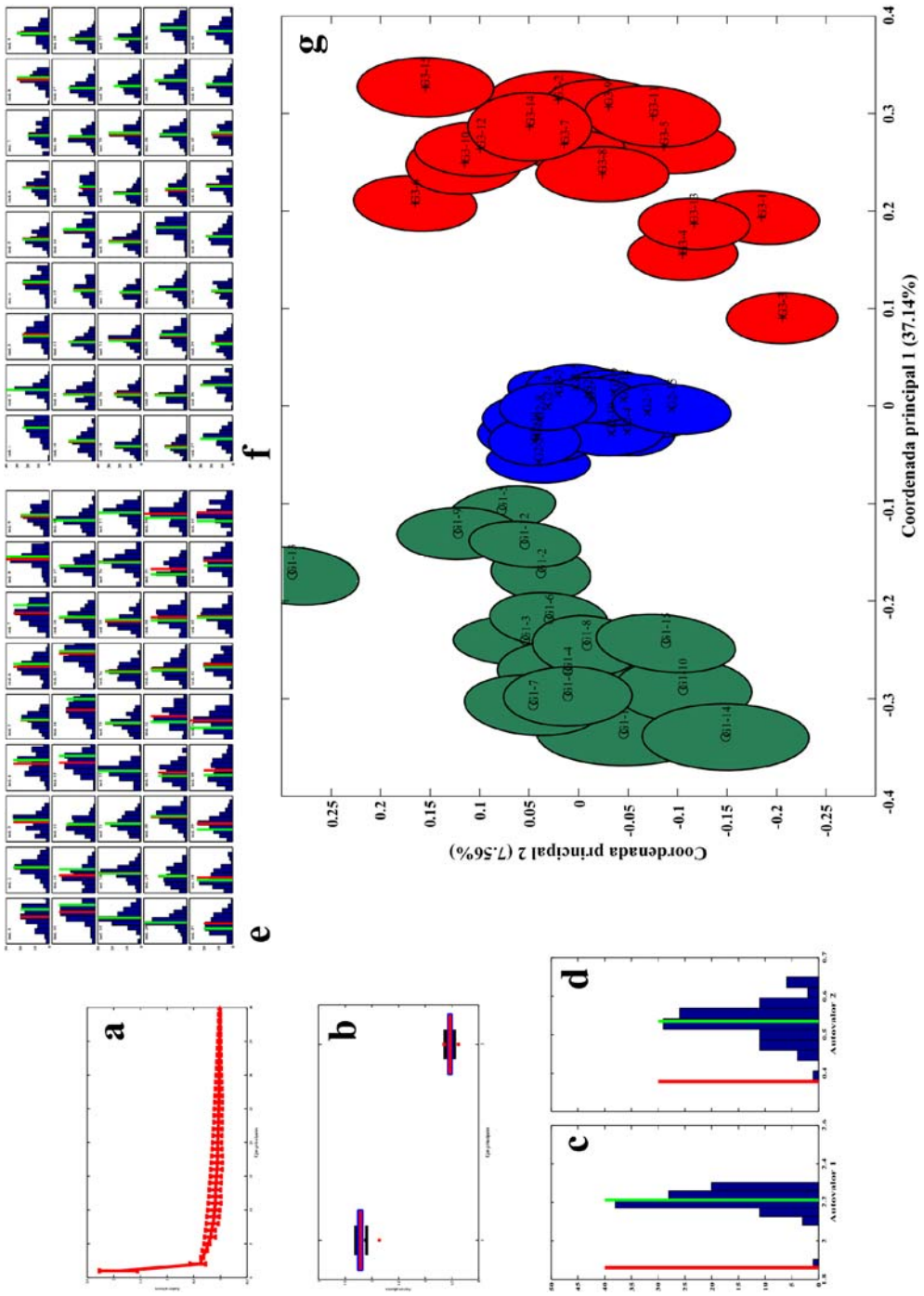


Figura 13. Variabilidad muestral de autovalores e individuos basada en el coeficiente de Dice, dos dimensiones, remuestreo sobre los residuales, transformación a través del método de Procrustes y $\mathbf{Y}^{oi}$ como configuración de referencia.

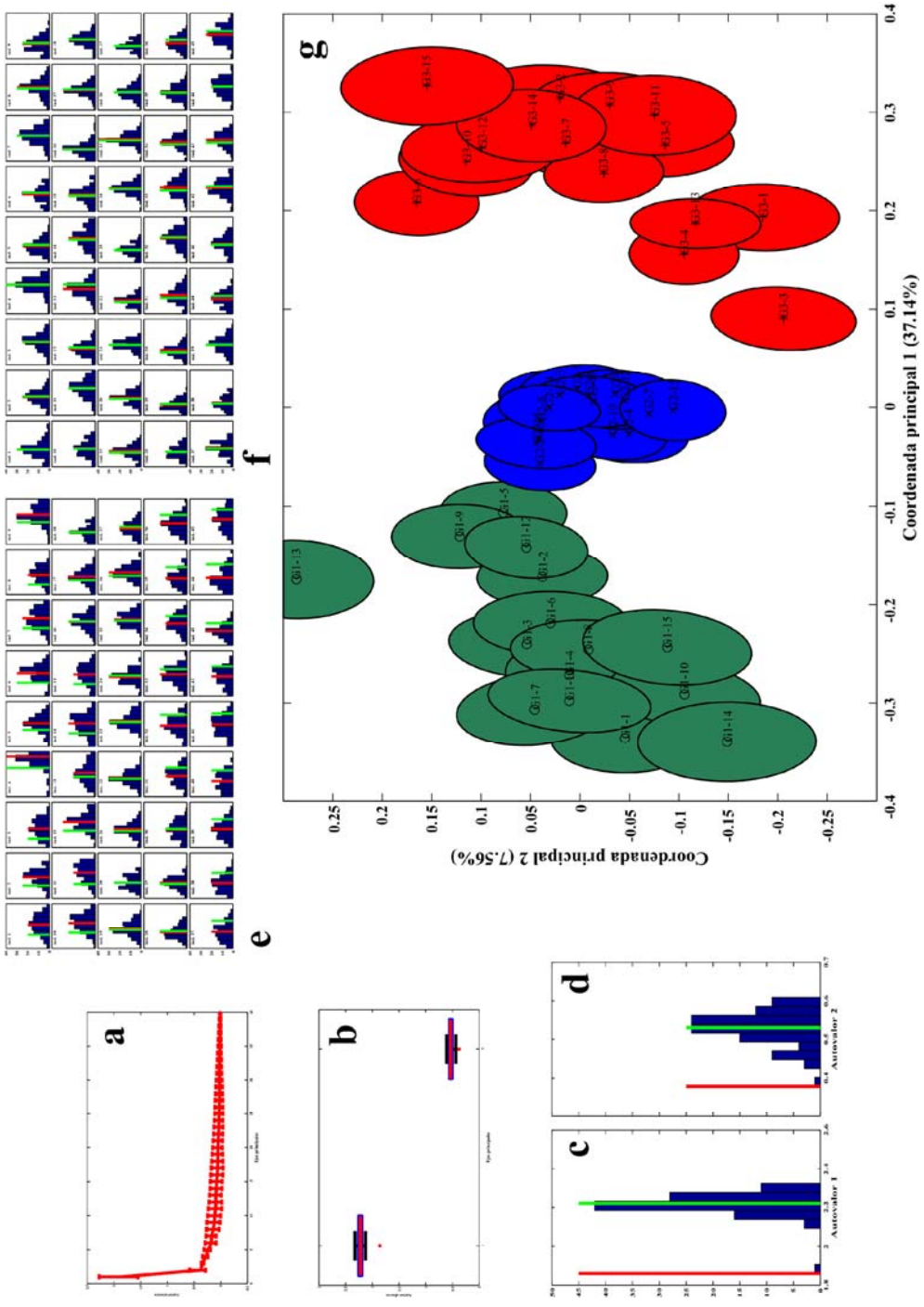


Figura 14. Variabilidad muestral de autovalores e individuos basada en el coeficiente de Dice, dos dimensiones, permutación aleatoria sobre los residuales, transformación a través del método de reflexión y $\mathbf{Y}^{oi}$ como configuración de referencia.

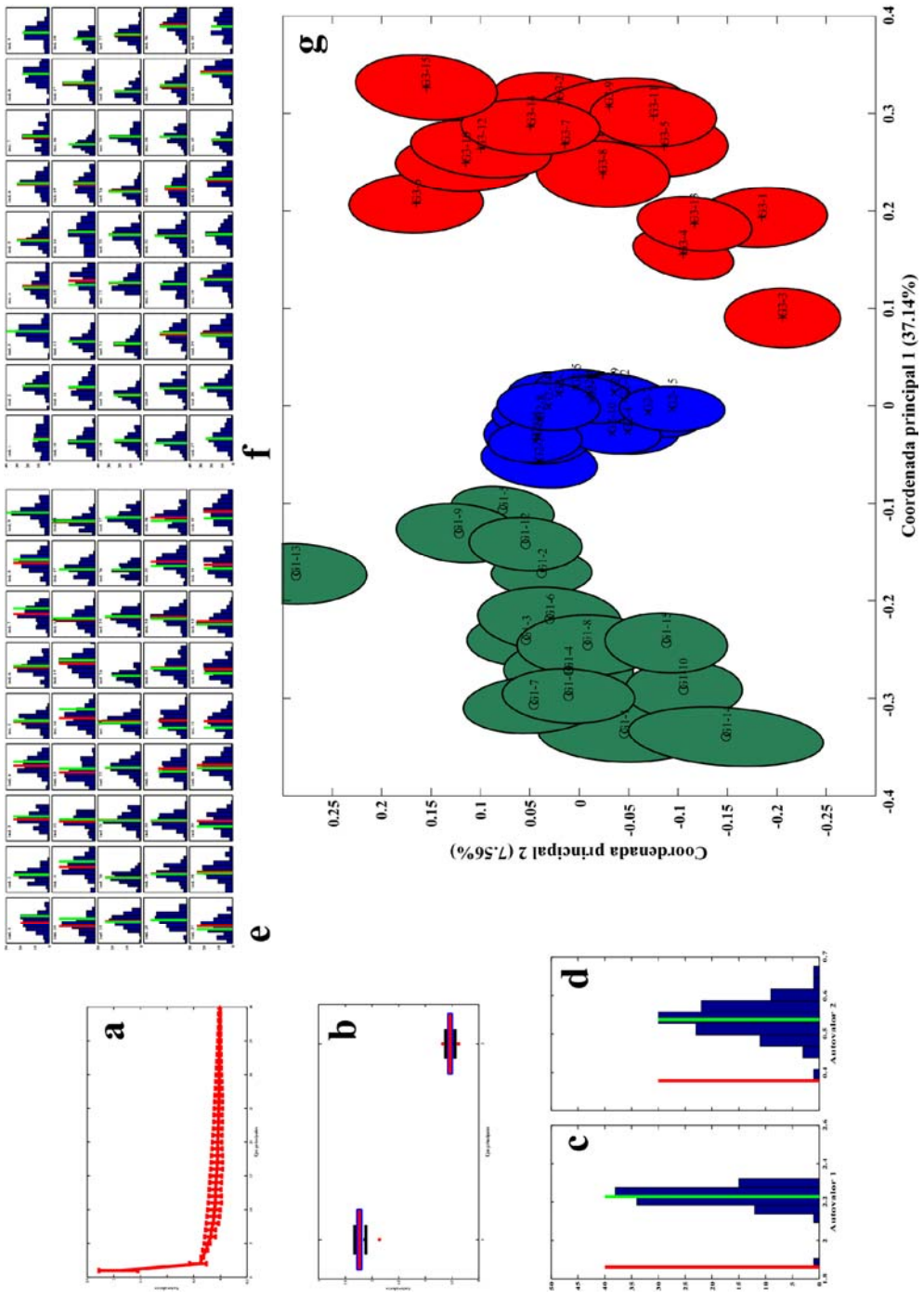


Figura 15. Variabilidad muestral de autovalores e individuos basada en el coeficiente de Dice, dos dimensiones, permutación aleatoria sobre los residuales, transformación a través del método de Procrustes y $\mathbf{Y}^{oi}$ como configuración de referencia.

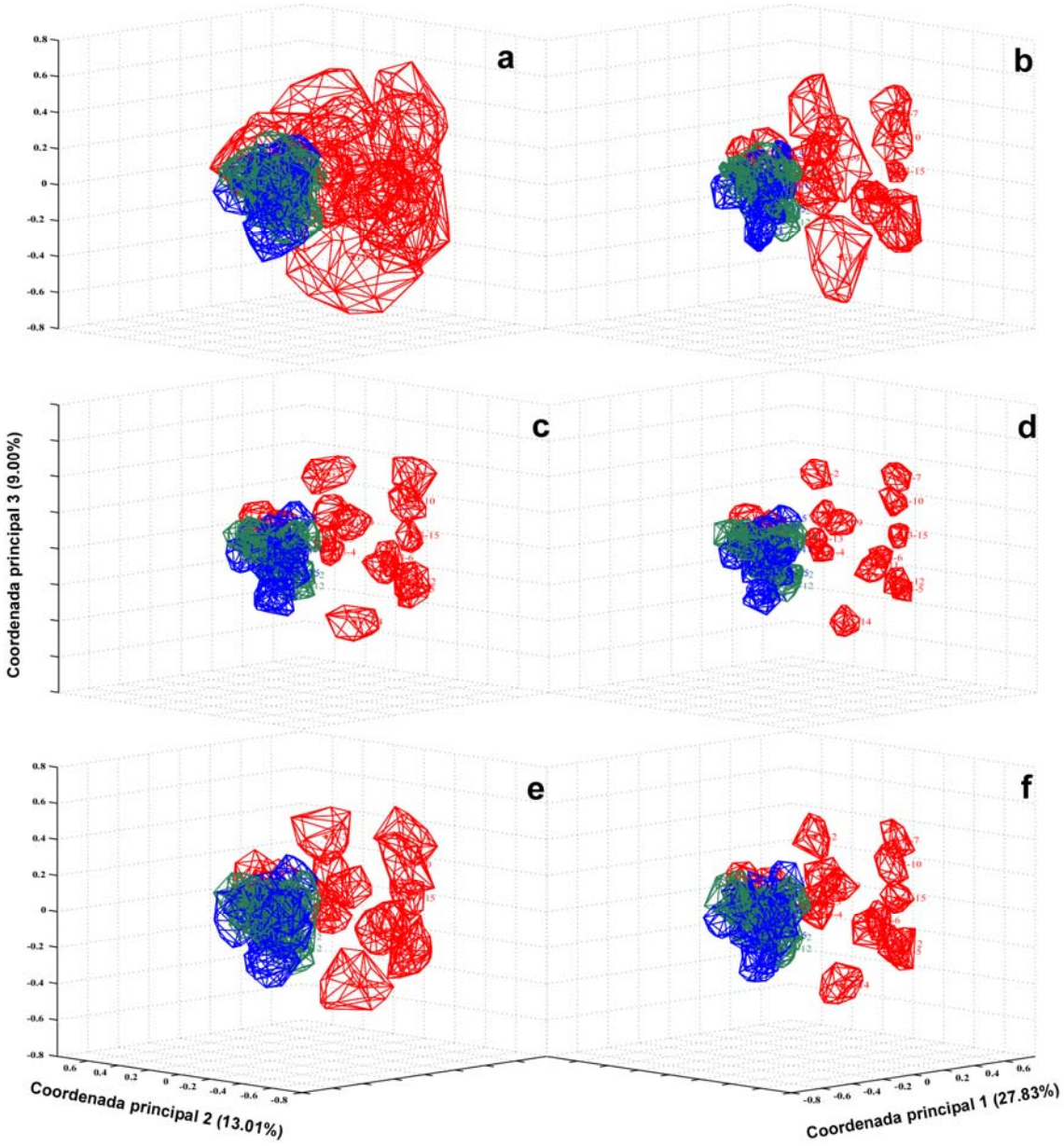


Figura 16. Proyección tridimensional de la variabilidad muestral de individuos basada en la disimilaridad de Dice y $\mathbf{Y}^{oi}$ como configuración de referencia. (**a,b**) Remuestreo sobre los individuos y transformación a través de los métodos de reflexión y Procrustes. (**c,d**) Remuestreo sobre los residuales y transformación a través de los métodos de reflexión y Procrustes. (**e,f**) Permutación aleatoria sobre los residuales y transformación a través de los métodos de reflexión y Procrustes.

2.2 METODOS BIPLOT

En el apartado 2.1 se ha descrito el ACoP y adicionalmente se han presentado procedimientos que le otorgan valor agregado en términos de la calidad de representación y cuantificación de la variabilidad de individuos y grupos; sin embargo, aunque le hemos aplicado métodos que mejoran el estudio de las relaciones entre individuos, tanto la representación gráfica como el análisis carecen de información sobre las variables.

2.2.1 Formulación

Consideramos en su definición clásica y más general a los métodos Biplot como una aproximación gráfica de una matriz de datos multivariantes -matriz de datos $\mathbf{X}$ de orden (*n*x*p*)-, que permite estudiar las relaciones entre individuos y variables. Es posible generar esta aproximación Biplot a través de un modelo bilineal general del tipo multiplicativo:

$$\mathbf{X} = \mathbf{Y}\boldsymbol{\beta}' + \mathbf{E} \qquad [2.7]$$

Que puede entenderse como una regresión multivariante de $\mathbf{X}$ sobre las coordenadas de los individuos $\mathbf{Y}$, cuando éstas están fijadas, o como una regresión multivariante de $\mathbf{X}'$ sobre las coordenadas de las variables $\boldsymbol{\beta}$, cuando están fijadas. La aplicación alternada de ambas regresiones multivariantes converge a la misma solución que la sostiene a partir de la Descomposición de Valores Singulares (DVS) (Vicente-Villardón *et. al*., 2006). Es posible obtener la misma solución cuando $\mathbf{Y}$ son las coordenadas principales

de $\mathbf{X}$ calculadas a partir de la distancia euclídea ordinaria y realizando solamente la primera de las regresiones multivariantes (Gabriel, 1971; Gower y Hand, 1996).

La representación gráfica de los *n* individuos y grupos generada vía Análisis de Coordenadas Principales (ACoP), permite que se proyecten las variables de la matriz $\mathbf{X}$, logrando así una representación conjunta, donde es posible visualizar las relaciones individuos-individuos, individuos-variables y variables-variables, en un espacio de dimensión reducida y con la menor pérdida de información. Este tipo de aproximación Biplot, a través de modelos bilineales, han sido denominados Biplots de Regresión ó Biplot Predicción (Gower y Hand, 1996; Cárdenas *et al.*, 2006; Vicente-Villardón *et al*., 2006).

Esta aproximación Biplot dependerá de las restricciones que se impongan para su ajuste, no obstante, en el caso del modelo [2.7] el vector de parámetros o coeficientes de la regresión de $\mathbf{X}$ sobre las coordenadas principales obtenidas usando una métrica euclídea puede ser estimado, como en un modelo de regresión, mediante:

$$\boldsymbol{\beta}' = \left(\mathbf{Y}'\mathbf{Y}\right)^{-1}\mathbf{Y}'\mathbf{X} \qquad [2.8]$$

Por otro lado, en los Biplots la matriz de datos $\mathbf{X}$ de orden (*n*x*p*) siempre puede descomponerse según la Descomposición en Valores Singulares (DVS) (Eckart y Young, 1936), en la forma:

$$\mathbf{X} = \mathbf{U}\,\boldsymbol{\Lambda}^{1/2}\,\mathbf{V}' \qquad [2.9]$$

donde:

$\mathbf{U}$: ya definida en [2.3], contiene los vectores propios de $\mathbf{XX}'$ y cumple $\mathbf{U}'\mathbf{U} = \mathbf{I}$.

$\mathbf{\Lambda}^{1/2}$: ya definida en [2.3], contiene los valores singulares de $\mathbf{X}$ o raíces cuadradas no negativas de los valores propios $\lambda_1, \lambda_2, ..., \lambda_p$ de $\mathbf{X}'\mathbf{X}$.

$\mathbf{V}$: contiene los vectores propios de $\mathbf{X}'\mathbf{X}$ y cumple $\mathbf{V}'\mathbf{V} = \mathbf{I}$.

Operando sobre $\mathbf{Y}$ y $\mathbf{X}$, usando [2.3] y [2.8], respectivamente, $\boldsymbol{\beta}'$ podrá tomar la forma:

$$\boldsymbol{\beta}' = \left(\left(\mathbf{U}\mathbf{\Lambda}^{1/2} \right)' \left(\mathbf{U}\mathbf{\Lambda}^{1/2} \right) \right)^{-1} \left(\mathbf{U}\mathbf{\Lambda}^{1/2} \right)' \left(\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}' \right) = \mathbf{V}' \qquad [2.10]$$

Ahora podemos expresar $E(\mathbf{X})$ como:

$$E(\mathbf{X}) = \mathbf{Y}\mathbf{V}' \qquad [2.11]$$

Nótese que las matrices obtenidas en [2.10] concuerdan con los marcadores (empleando la nomenclatura Biplot, no confundir con la nomenclatura genética) filas y columnas que se derivan de la factorización JK-Biplot de Gabriel (1971):

$$\mathbf{X} = \mathbf{J}\mathbf{K} \qquad [2.12]$$

donde $\mathbf{J} = \mathbf{U}\mathbf{\Lambda}^{1/2}$ y $\mathbf{K} = \mathbf{V}'$. Este procedimiento puede ser generalizable aun cuando $\mathbf{Y}$ se calcule utilizando métricas diferentes a la euclídea, con el inconveniente que se desconocen sus propiedades.

La aproximación del Biplot que permitirá proyectar las variables sobre la representación gráfica de los *n* individuos y grupos generada vía ACoP, estará dada por:

$$x_{i(q)} = \text{Proy}\left(\frac{\mathbf{x}_i}{\mathbf{V}_{(q)}}\right) = \sum_{k=1}^{q}\left(\mathbf{x}_i'\,\mathbf{V}_k\right)\mathbf{V}_k \qquad [2.13]$$

donde *i* y *k* representan la *i*-ésima fila y el rango de la matriz $\mathbf{X}$, respectivamente. El ajuste bilineal para cada elemento fila de la matriz $\mathbf{X}$ vienen dado por:

$$x_{ij(q)} \cong \left(\mathbf{y}_{i\,(q)}\right)'\boldsymbol{\beta}_{j(q)} = \sum_{k=1}^{q}\left(\mathbf{x}_i'\,\mathbf{V}_k\right)v_{jk} \qquad [2.14]$$

El error de esta aproximación del JK-Biplot como en cualquier regresión lineal, estará asociado a los desvíos entre los valores de las variables originales y sus proyecciones en la representación Biplot, es así que:

$$\sum_{i=1}^{n}\sum_{j=1}^{p}\left(residuos\right)^2 = traza\left(\mathbf{Y}'\mathbf{Y}\right) - \sum_{k=1}^{q}\lambda_k \qquad [2.15]$$

Los marcadores verifican las siguientes propiedades:

- Son equivalentes los productos escalares de las filas de la matriz $\mathbf{X}$ y de los marcadores $\mathbf{j}$, $\mathbf{XX}' = \mathbf{JJ}'$

- Son equivalentes la distancia euclídea entre dos filas de la matriz $\mathbf{X}$ y la distancias euclídeas entre los marcadores $\mathbf{j}$,

$$d_{ij}^2 = \left(\mathbf{X}_i - \mathbf{X}_j\right)' \left(\mathbf{X}_i - \mathbf{X}_j\right) = \left(\mathbf{j}_i - \mathbf{j}_j\right)' \left(\mathbf{j}_i - \mathbf{j}_j\right)$$

- Los marcadores para las filas coinciden con las coordenadas de los individuos en el espacio de las componentes principales de las variables.
- La calidad de representación para las filas es óptima y no preserva la métrica Euclídea entre columnas.

2.2.2 Geometría

Si consideramos a cada vector fila o columna de la matriz $\mathbf{X}$ como puntos *n* o *p* dimensionales en el espacio euclídeo, la aproximación Biplot a través del modelo de regresión lineal definido en [2.6] consistirá en realizar regresiones lineales simples para cada columna de la matriz $\mathbf{X}$ a partir de las coordenadas de los individuos generadas a través del Análisis de Coordenadas Principales (ACoP) que están contenidas en la matriz $\mathbf{Y}$. Los coeficientes de regresión de cada variable coinciden con sus coordenadas en la representación Biplot y se calculan como:

$$\mathbf{b}_j = \left(\mathbf{Y}'\mathbf{Y}\right)^{-1} \mathbf{Y}' x_j \qquad [2.16]$$

Vicente-Villardón *et al.* (2006), describen la geometría del Biplot ajustado a través de modelos de regresión lineal, llamando $\mathcal{L}$ al espacio generado por las columnas de $\mathbf{Y}$, y muestran que, sin pérdida de generalidad, el ajuste de los puntos al plano tridimensional de la regresión forma una superficie de respuesta lineal a la que denominan $\mathcal{H}$.

Así mismo, muestran que geométricamente el conjunto de puntos de $\mathcal{H}$ que predice un valor fijo de la variable x_j, está dado por la intersección entre el plano normal al tercer eje para el valor particular de x_j y el plano de regresión, y que para diferentes valores a predecir se obtienen rectas paralelas en el plano $\mathcal{H}$. Al eje de referencia que permite predecir los valores de x_j y que representa la dirección de $\mathcal{H}$ normal a todas esas rectas paralelas en el plano de regresión se le denomina ξ_j. Los puntos en $\mathcal{L}$ que predicen diversos valores de la variable - coeficientes de regresión de x_j sobre y_i - están también en líneas rectas paralelas; la proyección de ξ_j sobre $\mathcal{L}$ es normal a todas las líneas y se denomina eje Biplot β_j, Figura 17.

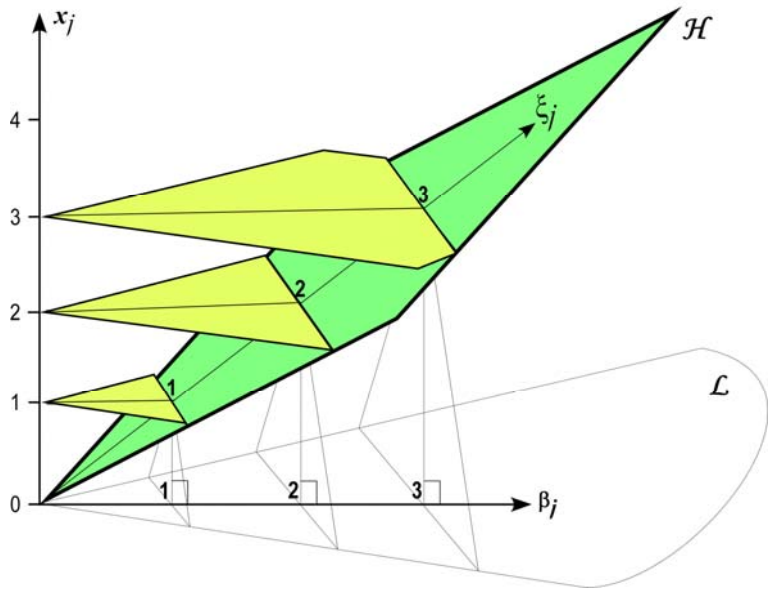


Figura 17. Geometría del Biplot ajustado a través de modelos de regresión lineal. Tomado de: Vicente-Villardón *et al.* (2006).

La proyección de los marcadores filas sobre el eje Biplot $\beta_j = \left(b_{j1}, b_{j2}\right)$ permite derivar el eje de predicción $\mathcal{L}$ para diferentes puntos a través de interpolación. Es así, que para encontrar un marcador β_j que permita predecir un valor fijo μ de la variable observada debemos encontrar un punto (*x*, *y*) que verifique:

$$y = \frac{b_{j2}}{b_{j1}} x \qquad \text{y} \qquad \mu = b_{j0} + b_{j1}x + b_{j2}y \tag{2.17}$$

resolviendo el sistema para *x* e *y*, obtenemos:

$$x = \mu \frac{b_{j1}}{b_{j1}^2 + b_{j2}^2} \qquad\qquad y = \mu \frac{b_{j2}}{b_{j1}^2 + b_{j2}^2} \tag{2.18}$$

y en forma general tenemos:

$$(x, y) = \mu \frac{\mathbf{b}_j}{\mathbf{b}_j' \mathbf{b}} \tag{2.19}$$

Por lo que el marcador que permite predecir un valor fijo de la variable *j*-ésima viene dado por la razón entre las coordenadas del β_j y su longitud ajustada. La calidad de representación de cada variable se mide a través de los coeficientes de determinación R_j^2 derivados de cada regresión que se interpretan de la misma manera que en el análisis de correspondencias o el análisis Biplot clásico.

Las proyecciones de los marcadores fila sobre los marcadores columna permiten una ordenación de los individuos respecto a cada una de las variables consideradas en el ajuste y si como fue referido a apartes previos, los individuos son clasificados en grupos producto de la aplicación de algún algoritmo de agrupamiento sobre las primeras coordenadas principales, se obtendrá una representación gráfica resultado de la combinación de las tres técnicas: Análisis de Coordenadas Principales, Análisis de Conglomerados (AC) y los métodos Biplot que favorece el estudio simultáneo de las relaciones entre individuos, individuos-variables y variables-variables, incrementando la cantidad y calidad de la información sobre los métodos unánimemente utilizados para los estudios de diversidad genética. A continuación presentamos un cuadro comparativo que detalla las diferencias entre las dos los enfoques.

Tabla 6. Comparación entre los enfoques de clasificación

Criterios de comparación	**Análisis de conglomerados (enfoque clásico)**	**Análisis combinado de: ACoP, Análisis de conglomerados y Métodos Biplot (enfoque propuesto)**
1. Estudio de las relaciones entre individuos.	✓	✓
2. Estudio de la calidad de representación de los individuos.		✓
3. Estudios de la calidad de representación de los grupos.	✓	✓
4. Selección de las variables responsables de la formación de grupos.		✓
5. Estudio de la relación entre las variables responsables de la formación de grupos.		✓

Como se ha señalado, las derivaciones, geometría, y bondad de ajuste, entre otros aspectos, del desarrollo Biplot presentado se encuentran extensamente desarrolladas en la bibliografía. No obstante, nuestra intención ha sido mostrar que es posible, al igual

que en el Análisis de Componentes Principales (ACP), proyectar las variables en el Análisis de Coordenadas Principales (ACoP), con la ventaja adicional que se puede utilizar cualquiera de la medidas de similitud/disimilitud presentadas en apartados anteriores para establecer las relaciones entre individuos para distintos tipos de variables.

Capítulo III

IDENTIFICACIÓN DE LOS MARCADORES MOLECULARES ASOCIADOS CON LA CLASIFICACIÓN DE GENOTIPOS

CAPITULO III

En el capítulo anterior hemos demostrado que es posible mejorar la interpretación del Análisis de Coordenadas Principales (ACoP) con la proyección de variables cuantitativas a través del ajuste de Biplot de regresión o más explícitamente a través del ajuste de regresiones lineales simples para cada columna de la matriz $\mathbf{X}$, a partir de las coordenadas de los individuos generadas a través del Análisis de Coordenadas Principales (ACoP) y que están contenidas en la matriz $\mathbf{Y}$.

Sin embargo, cuando la matriz $\mathbf{X}$ proviene de la observación de p atributos o caracteres cualitativos que se asocian a variables binarias que toman el valor 0 si la característica está ausente y el valor 1 si está presente, como es el caso de la codificación de las diferentes alternativas alélicas en el análisis de la información molecular, la aplicación del Biplot lineal clásico así como el Análisis de Componentes Principales (ACP) no es conveniente, porque en ambos se supone que la respuesta a lo largo de las dimensiones es lineal. Esto debe ser entendido por la misma razón que no es apropiado ajustar una regresión lineal cuando la variable respuesta es binaria o categórica.

Se han utilizado diversas estrategias para ajustar Biplots utilizando matrices de datos binarios. El Análisis de Correspondencias Múltiples (ACM) puede ser considerarse como una forma particular de ajuste Biplot para una matriz binaria, donde las regiones de predicción se basan en las distancias entre individuos y categorías (Gower y Hand, 1996). No obstante, en el contexto de estudios de diversidad genética utilizando

marcadores moleculares esta técnica no debe ser aplicada puesto que no refleja la estructura de los datos ya que está basada en la distancia χ^2.

Otras estrategias se fundamentan en: (***i***) modelaje de la respuesta binaria a través de regresiones generalizadas alternadas, estimando cada columna de $\mathbf{X}$ bajo el supuesto de independencia entre individuos, así como de los parámetros de cada una de las variables; (***ii***) estimación conjunta de todas las columnas de la matriz $\mathbf{X}$ a través de regresión bilineal generalizada; y (***iii***) estimación conjunta y en forma simultánea de todas las filas y columnas de $\mathbf{X}$ (van Eeuwijk, 1995ab; Blázquez, 1998; Gabriel, 1998; Vicente-Villardón *et al.*, 2006).

En este sentido Vicente-Villardón *et al.* (2006) -bajo el enfoque de regresiones o interpolaciones alternadas- proponen el ajuste de un Biplot Logístico (BL) lineal para datos binarios, en el cual la respuesta a lo largo de las dimensiones es logística. En este Biplot Logístico (BL) los individuos se representan como puntos y las variables a través de vectores centrados en el origen. Desde el punto de vista geométrico se considera que la proyección de un individuo sobre una dirección de un vector predice la probabilidad de la presencia de ese carácter o variable. El método tiene la ventaja que se relaciona con la regresión logística de la misma manera que el método de Biplot está relacionado con la regresión lineal y a diferencia de las propuestas de van Eeuwijk (1995ab), Gabriel (1998) o Falguerolles (1998) otorga un enfoque exploratorio donde el objetivo principal es analizar la matriz de datos (individuos por variables) y no modelar una tabla de dos vías. Este enfoque se puede considerar más cercano al Análisis de

Correspondencias Múltiples (ACM) y algunos procedimientos sobre variables latentes muy utilizados en psicometría como la Teoría de la Respuesta al Item (TRI).

Orientados en la idea de Vicente-Villardón *et al.* (2006) y siguiendo la estrategia metodológica que hemos venido desarrollando a lo largo del trabajo proponemos el uso combinado del Análisis de Coordenadas Principales (ACoP), Análisis de Conglomerados (AC) y el ajuste de un Biplot Logístico Externo (BLE) sobre las coordenadas principales como mejor vía para identificar las alternativas alélicas que son responsables de la clasificación de genotipos en estudios de diversidad genética que involucren información de marcadores moleculares.

Esta propuesta se basa en el hecho de que la regresión alternada para datos binarios de las columnas de la matriz $\mathbf{X}$ introducida por Vicente-Villardón *et al.* (2006) es análoga a ajustar regresiones logísticas simples para cada columna de la matriz $\mathbf{X}$ sobre la configuración *k*-dimensional obtenida del Análisis de Coordenadas Principales (ACoP).

Aunque podría haberse utilizado el procedimiento de regresión alternada para datos binarios, el Análisis de Coordenadas Principales (ACoP) es más simple, más accesible a los usuarios, proporciona una alternativa más flexible porque permite utilizar diversas medidas de la similitud/disimilitud y, en nuestra experiencia, los resultados son similares. Adicionalmente, los procedimientos de regresiones alternadas aunque comparten la misma geometría necesitan algunas adaptaciones para operar con matrices de datos binarios. Por lo que, en el contexto de estudios de diversidad genética

utilizando marcadores moleculares, su uso puede ser lo suficientemente complejo como para dificultar su aplicabilidad.

3.1. BIPLOT LOGISTICO EXTERNO

3.1.1 Formulación

Sea $\mathbf{X}$ la matriz de datos de orden (*nxp*) que proviene de la observación de *n* individuos a los que se les cuantifican p atributos o caracteres cualitativos que se asocian a variables binarias -fragmentos de amplificación- que toman el valor 0 si la característica (alelo o banda) está ausente y el valor 1 si está presente. Sea $\pi_{ij} = E(x_{ij})$ la probabilidad de que el j-ésimo alelo esté presente en un genotipo cualquiera, con coordenadas y_{is} $(i = 1,...,n;\ s = 1,...,k)$ y que está representado en el plano *k*-dimensional generado por el Análisis de Coordenadas Principales (ACoP), π_{ij} puede escribirse en función de las coordenadas principales como:

$$\pi_{ij} = \frac{e^{b_{j0} + \sum_{s=1}^{k} b_{js} y_{is}}}{1 + e^{b_{j0} + \sum_{s=1}^{k} b_{js} y_{is}}} \qquad [3.1]$$

donde b_{js} $(j = 1,..., p)$ son los coeficientes de la regresión logística que corresponden a la *j*-ésima variable (alelos o bandas) en la en la *k*-ésima dimensión. El modelo [3.1] es equivalente al modelo lineal generalizado que utiliza la función logit, como función de enlace para evitar problemas de escala.

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = b_{j0} + \sum_{s=1}^{k} b_{js} y_{is} = b_{j0} + \mathbf{y}_i' \mathbf{b}_j \qquad [3.2]$$

donde $\mathbf{y}_i = (y_{i1},...,y_{ik})'$ y $\mathbf{b}_j = (b_{j1},...,b_{jk})'$ definen a un Biplot en escala logit. El procedimiento se denomina Biplot Logístico Externo (BLE) porque las coordenadas de los *n* individuos (genotipos) se calculan en un procedimiento externo como el Análisis de Coordenadas Principales (ACoP). Es así como, si las y_s' son variables conocidas cuyo número sólo depende de las *k*-dimensiones que se deseen retener, los parámetros b_s' se obtienen ajustando regresiones logísticas simples utilizando la *j*-ésima columna de la matriz $\mathbf{X}$ como variable dependiente y las y_s' como regresoras.

Este procedimiento permite generar una grafico bi o tri dimensional, donde las y_s' son representadas como puntos (genotipos) y los b_s' estimados para cada alelo son representados como vectores los cuales determinan las direcciones de los ejes Biplot. La proyección de cada uno de los genotipos sobre el segmento que representa a cada alelo, permite obtener la probabilidad estimada de presencia de un alelo en particular para cada genotipo.

Como en cualquier problema de modelaje no todas las variables (alelos) estarán asociados significativamente a la configuración. En el contexto de la clasificación de genotipos usando marcadores moleculares, como se estudia un número elevado de

alelos, solo se deben proyectar aquellos que se relacionan directamente con la configuración, es decir, aquellas cuyos parámetros presenten la mejor calidad de representación después de ajustar la regresión logística. En este sentido, el pseudo R^2 de Nagelkerke/Cragg & Uhler's (Long, 1997) para regresiones de variables categóricas se utiliza como medida de la "calidad de la representación" y se interpreta de la misma forma que en Análisis de Correspondencias (Tenenhaus y Young, 1985). Adicionalmente, la corrección de Bonferroni puede ser utilizada como criterio de selección de alelos con alta capacidad discriminatoria. Con este método, solo aquellos alelos que tienen un nivel de significación dado; por ejemplo, para un α=0.05 serán proyectados en el Biplot aquellos alelos con $p \leq$(0.05/número total de alelos). No obstante para grandes conjuntos de datos, los p valores son afectados considerablemente por el tamaño de muestra y el número de alelos. En estos casos, Demey *et al.* (2008) recomiendan utilizar el pseudo R^2 con un valor altamente restrictivo porque es menos sensible al tamaño de muestra.

Desde el punto de vista de la ordenación, la calidad de representación o bondad de ajuste representa el "Porcentaje de Clasificación Correcta (PCC)", el porcentaje de coincidencias entre la matriz de los datos binarios original y la estimada de los modelos de regresión logística. Este puede ser calculado en forma global para la representación Biplot o para fila o columna separadamente.

De esta manera, la identificación de los alelos asociados con la clasificación de genotipos o grupo de éstos, es equivalente a la selección de las variables asociadas a la ordenación generada vía Análisis de Coordenadas Principales (ACoP).

3.1.2 Geometría del Biplot Logístico Externo

Al igual que los Biplot ajustados a través de modelos de regresión lineal, en el Biplot Logístico Externo (BLE) o Biplot de probabilidad, el ajuste al hiperplano genera una superficie de respuesta sigmoidea. Las proyecciones de las curvas de respuesta sobre el subespacio de mejor ajuste generan ejes Biplot de predicción lineal, aunque la respuesta ajustada sea no lineal. Vicente-Villardón *et al.* (2006) demuestran que la proyección de la curva de respuesta no lineal sobre un subespacio de baja dimensión es siempre lineal, aunque la escala de predicción en el eje Biplot no se encuentre igualmente espaciada. Consiguientemente, la predicción de las probabilidades se hace de la misma forma que en un Biplot lineal. En nuestro contexto, se interpreta como que la proyección de un genotipo en la dirección de un vector (alelo) cualquiera predice la probabilidad de la presencia de ese alelo en el genotipo.

Para facilitar la interpretación gráfica, en los extremos de cada vector se fijan puntos de predicción con probabilidad conocida, es así como el 0.50 se fija como punto corte para la predicción de la presencia y 0.75 para la dirección de mayor probabilidad creciente. La longitud del vector debe ser interpretada como una medida inversa de la capacidad discriminatoria de los alelos, en decir, vectores más cortos corresponden con alelos que discriminan mejor a los genotipos. La relación entre los diferentes alelos proyectados

sobre el plano Biplot, se interpreta según el ángulo que formen. Cuando dos alelos tengan el mismo sentido de predicción se dice que están positivamente correlacionados, cuando tengan direcciones opuestas se correlacionan negativamente, y cuando formen un ángulo cerca de 90° se dice que son independientes, Figura 18.

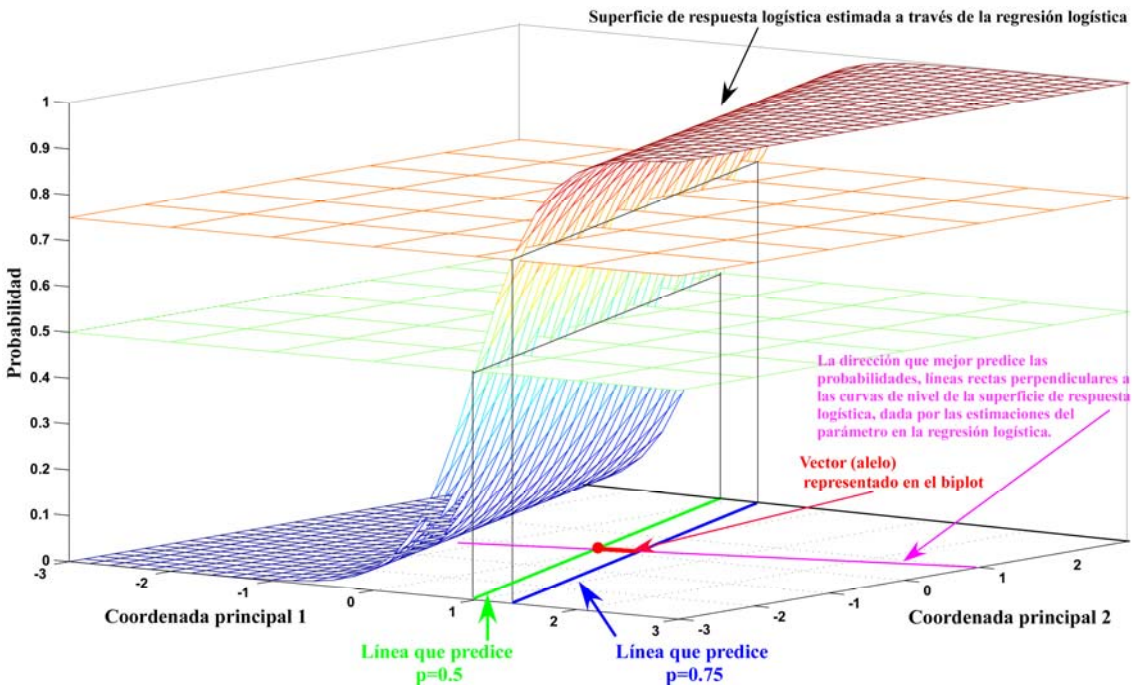


Figura 18. Geometría de la curva de respuesta logística ajustada

Como ilustración, la Figura 19 muestra paso a paso cómo interpretar las proyecciones usando el Biplot Logístico Externo.

Supongamos un alelo cualquiera que ha sido nombrado como A_1A_1, que está representado por un pequeño segmento y además que existen tres grupos de genotipos marcados **A**, **B** y **C** (Figura 19a).

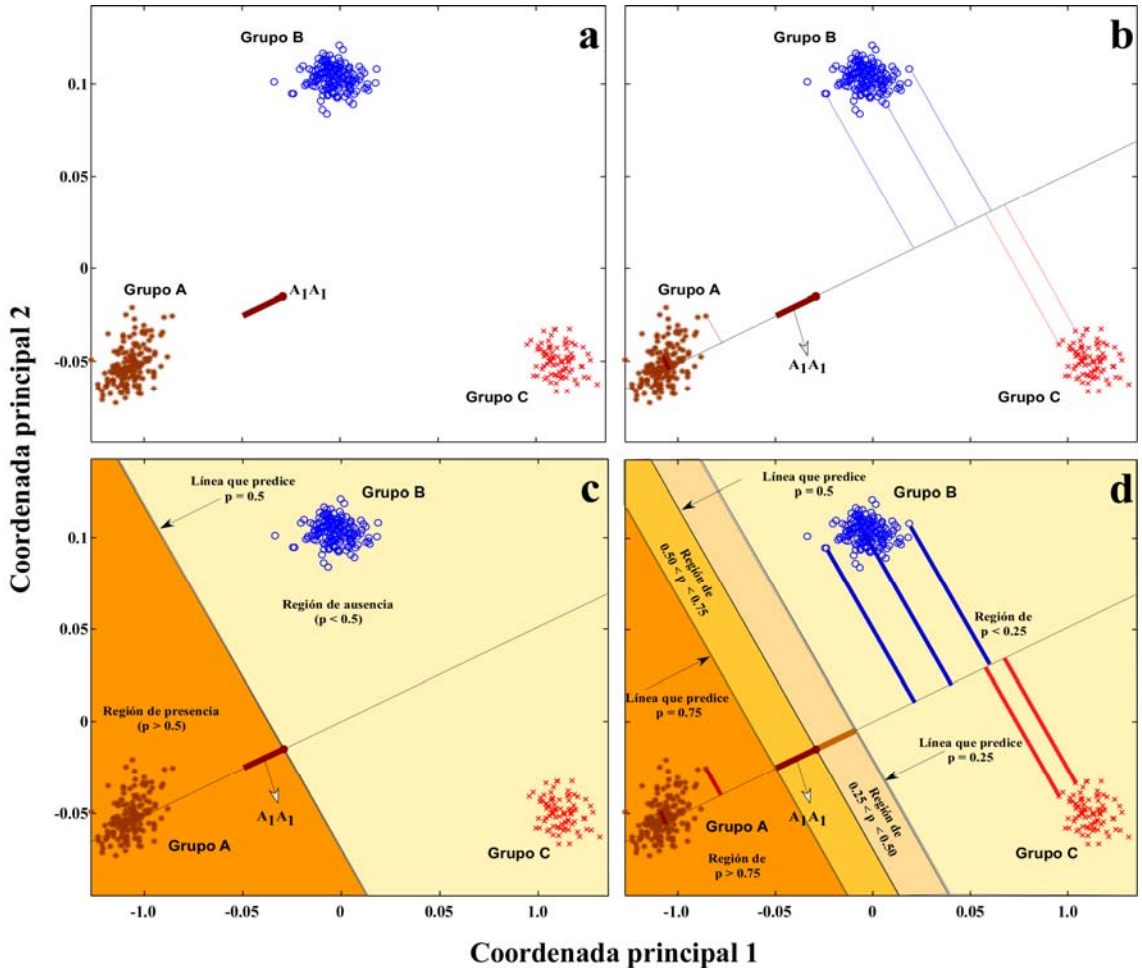


Figura 19. Proyecciones usando el Biplot Logístico Externo

El primer paso consiste en extender la proyección del alelo de forma que atraviese completamente el gráfico y corte los extremos. La dirección del alelo se debe entender como una serie continua que cubre la escala de probabilidades (0,1). No obstante, para

simplificar la representación gráfica solamente se representan los puntos que predicen 0.5 (el principio del segmento marcado con un pequeño círculo), y 0.75 el final del segmento. Es así como, si se proyecta el alelo hacia la derecha del gráfico, la línea cubrirá probabilidades inferiores a 0.5 y al contrario, hacia el extremo izquierdo, probabilidades superiores a 0.75. Luego se toman genotipos ilustrativos de cada uno de los grupos y se hacen proyecciones perpendiculares de cada uno sobre el alelo. La proyección perpendicular de un genotipo sobre esa dirección aproxima la probabilidad estimada; se entiende que la posición de un genotipo respecto al gráfico estará representando al grupo de genotipos que presentan un mismo patrón de ADN, Figura 19b.

El segundo paso consiste en dividir el gráfico dibujando una línea perpendicular al alelo que atraviese el punto de predicción de 0.5. Esta línea divide el espacio en dos regiones: la región dónde es más probable que el alelo esté presente ($p>0.5$) y la región dónde es más probable que el alelo esté ausente ($p<0.5$), Figura 19c. Por ejemplo, todos los genotipos del grupo **A** tienen mayor probabilidad de tener el alelo presente, mientras que los genotipos del resto de los grupos tienen mayor probabilidad de que el alelo esté ausente. Si modificáramos la dirección de mayor probabilidad del alelo, entonces la interpretación se invertiría.

A continuación, se proyecta una imagen del alelo en el sentido de menor probabilidad y se dibujan líneas perpendiculares al extremo del alelo original y de su imagen, respectivamente. La primera línea representa $p=0.75$ y la segunda $p=0.25$. Es importante

señalar que es posible dibujar tantas líneas perpendiculares como regiones de probabilidad se quieran representar en el gráfico, en nuestro ejemplo el gráfico ha sido dividido en cuatro regiones a saber: $p<0.25$, $0.25<p<0.50$, $0.50<p<0.75$ y $p>0.75$, Figura 19d. De esta forma se aumenta la precisión de la interpretación, y es así como puede afirmarse que la probabilidad de que el alelo esté presente en el grupo **A** es mayor que 0.75 ($p>0.75$).

Es recomendable, cuando se quieren hacer interpretaciones sobre los grupos, tomar el centroide para hacer las proyecciones sobre los alelos. Cuanto más compacto sea el grupo mejor será el centroide para representar las coordenadas de los genotipos de ese grupo.

Para una mejor comprensión de la geometría de la solución, la curva de respuesta para un eje de coordenadas principales se muestra en la Figura 20. El segmento marcado en azul en el eje principal, es el equivalente al segmento representado en el plano, Figura 19. Puede observarse que espacios de probabilidades iguales, no corresponden a marcadores equidistantes en la dirección del alelo, porque la respuesta logística es no lineal, es así como la longitud del alelo que predice las regiones de probabilidad de $0.20<p<0.30$, es diferente a la longitud del segmento que predice la región de $0.80<p<0.90$. El equivalente para una representación bidimensional, con sólo dos coordenadas principales retenidas se mostró en la Figura 18 y aunque la interpretación es un poco más compleja, el razonamiento geométrico es similar.

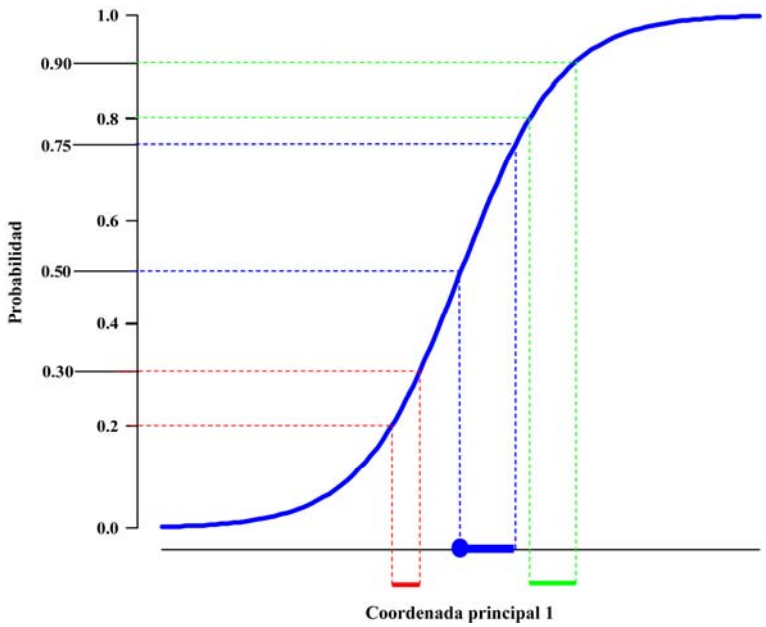


Figura 20. Geometría de la curva logística

Adicionalmente se ilustra cómo los alelos más cortos tienen mayor capacidad discriminatoria, para esto se comparan las dos curvas logísticas representadas en la Figura 21. La curva azul tiene una capacidad más alta de discriminar entre presencias y ausencias porque la pendiente de la curva logística en el punto que predice 0.5 es más alta; es decir, las probabilidades de la presencia contra la ausencia incrementan más rápidamente en la curva azul. Formalmente esta capacidad discriminatoria puede ser definida como la pendiente de la tangente a la respuesta logística, en el punto que predice p=0.5. Se reconoce fácilmente que la curva azul tiene un segmento más corto y entonces tiene una mayor capacidad discriminatoria.

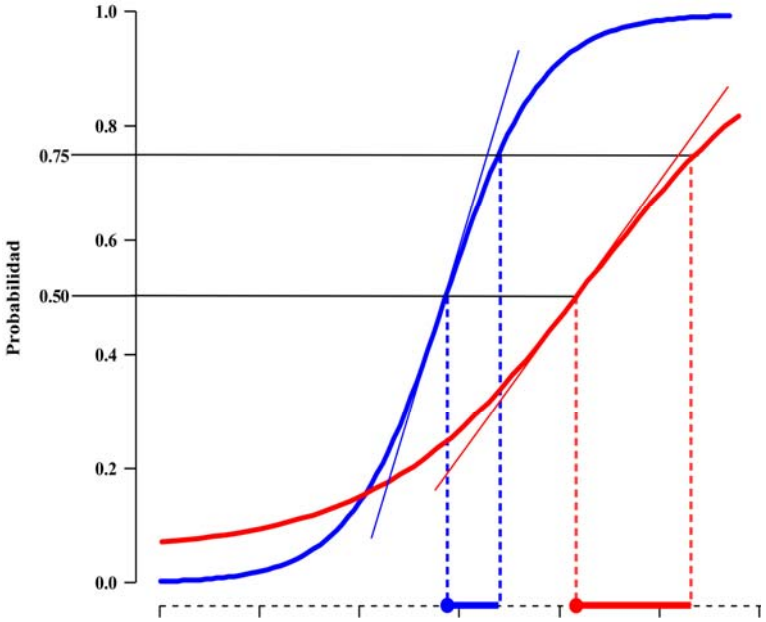


Figura 21. Interpretación de las longitudes de los alelos

Sin embargo, la interpretación de la longitud del vector debe hacerse con precaución, porque en algunos casos, longitudes más cortas también podrían estar asociadas al hecho que los segmentos señalan en la dirección de otro eje; es por esta razón, que no debe ser tomada en cuenta solo la longitud del alelo para su selección, sino que además deben investigarse todas las dimensiones retenidas y considerar los alelos que en el ajuste de la regresión logística obtengan un alto R^2.

Otro criterio que permite evitar una falsa interpretación, es calcular el coseno de la dirección del alelo con todas las coordenadas principales retenidas. Los cosenos son el equivalente a las correlaciones entre las variables originales y los componentes

principales en el Análisis de Componentes Principales (ACP) para datos continuos, donde solo aquellos componentes principales con cosenos más altos se deben interpretar en cada proyección parcial.

La Figura 22 demuestra la proyección para el alelo en la Figura 18 y permite ilustrar la importancia de calcular el coseno de la dirección del alelo con todas las coordenadas principales retenidas. Es así como se observa, que el alelo se asocia principalmente a la primera dimensión de las coordenadas principales; sin embargo, la proyección es más corta para la segunda, siendo este un ejemplo donde un segmento es corto en la proyección de un plano porque señala otra dimensión de la solución. Por esta razón, las interpretaciones de las longitudes cortas sobre proyecciones en el segundo eje podrían llevar a conclusiones incorrectas.

Este enfoque, aplicado a la clasificación de genotipos usando marcadores moleculares, permite considerar las dimensiones de la solución de las coordenadas principales como gradientes genéticos latentes y estos se podrán interpretar usando los alelos con los cosenos más altos (correlaciones) con cada dimensión. En el ejemplo de la Figura 19, observamos que la primera dimensión representa un gradiente genético que discrimina entre los grupos **A** y **C**, con el grupo **B** ocupando posiciones intermedias. Podríamos interpretar entonces, que alelos altamente correlacionados con la primera dimensión estarán presentes en el grupo **A** y ausentes en el grupo **C** o presentes en el grupo **C** y ausentes en el grupo **A**, en el grupo **B** existe una mezcla de patrones para esos alelos. Respecto a la segunda dimensión, separa el grupo **B** del resto, es decir, los alelos más

correlacionados con la segunda dimensión mostrarán las diferencias genéticas entre el grupo **B** y el resto.

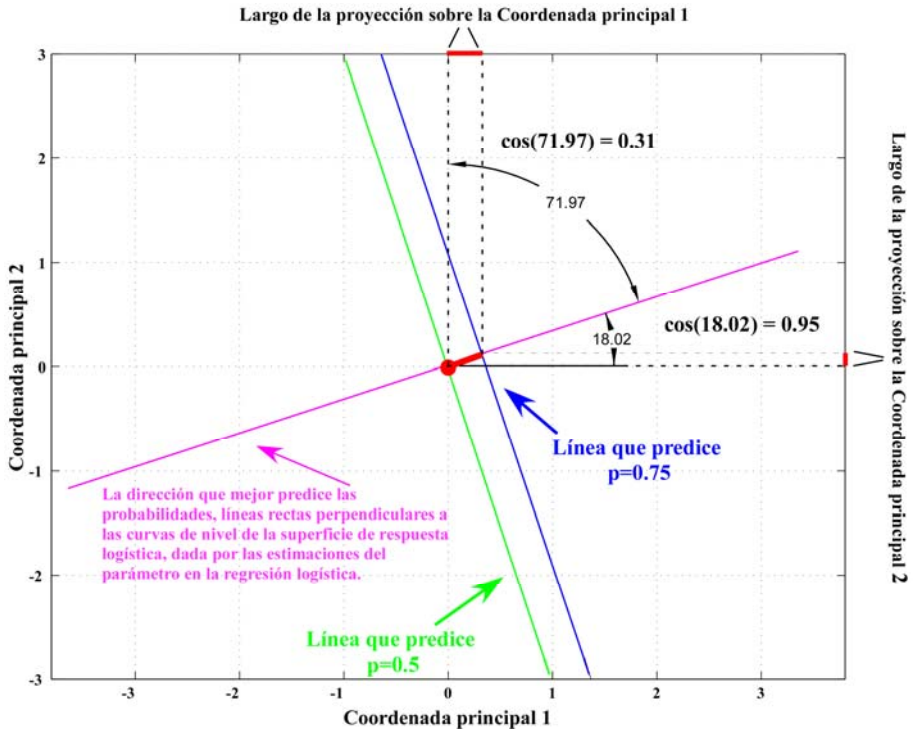


Figura 22. Interpretación de las relaciones entre alelos y ejes principales.

El software desarrollado para esta aplicación tiene presente todas estas consideraciones analíticas y geométricas analizadas, adicionalmente como ayuda a la interpretación y a la visualización, calcula los cosenos de los ángulos formados por las direcciones y las dimensiones, la dimensión con el coseno más alto y el cuadrante al que pertenece cada alelo. También permite ocultar los alelos no relacionados con la ordenación siguiendo los criterios de selección mencionados.

3.2 ESTUDIO DE SIMULACION

A continuación ilustraremos de forma empírica el comportamiento del Biplot Logístico Externo (BLE) en la identificación de los marcadores moleculares asociados a la clasificación de genotipos, al igual que en los casos ya desarrollados, utilizamos la estrategia de la simulación modelando escenarios donde se conocen *a priori* tanto la estructura de grupos como las variables responsables de su formación.

3.2.1 Método

Se generaron matrices binarias con estructura de grupo, simulando en forma aleatoria individuos diploides con número de loci variable, suponiendo en cada caso una población no apareada al azar con respecto a cada locus bialélico. Las alternativas alélicas generadas por la simulación de los individuos fueron codificadas suponiendo un marcador dominante, es decir, $A_1A_1 = A_1A_2 = 1$ y $A_2A_2 = 0$. Se generó una sola columna por loci garantizando que los alelos presentes en un grupo estuviesen ausente en el resto para permitir la configuración de tres escenarios básicos de simulación. ***S1***, ***S2*** y ***S3*** que dependen del número de grupos generados, Tabla 7.

A los escenarios básicos descritos le fueron agregados dos tipos de ruido. Un ruido externo que consistió en agregar un conjunto de loci bialélicos suplementarios que representaban el 250% del total de los utilizados para formar las diferentes estructuras de grupos o matrices de señal. Al igual que en otro de los casos presentados, los loci bialélicos fueron generados utilizando una distribución uniforme (0,1), los alelos se consideraron presentes si el valor simulado $x_i \geq 0.5$ y ausente en el caso contrario.

El segundo tipo de ruido -ruido interno- consintió en modificar en forma aleatoria en niveles del 5%, 10% y 20% los valores asignados a cada individuo por alelo, es así como, los valores de presencia fueron sustituidos por ausencia o viceversa en los porcentajes indicados. Obsérvese, que los porcentajes de ruido ensayados pueden considerarse altos a la vista de las nuevas tecnologías de marcadores moleculares disponibles; sin embargo, se han ensayado escenarios extremos para garantizar que las conclusiones sobre el comportamiento del Biplot Logístico Externo (BLE) que se generen sean lo más confiables posibles.

Para cada uno de tres los escenarios básicos se generaron tres combinaciones de ruido externo y tres de ruido interno, produciendo en total 27 alternativas. Para facilitar la visualización de las diferentes alternativas en el Biplot se utilizó un número moderado de individuos (n=50) y se considera a cada columna de la matriz binaria como un alelo, Tabla 7. Cada Biplot fue ajustado utilizando el coeficiente de similitud de Dice para generar las coordenadas principales y el algoritmo de agrupamiento UPGMA para confirmar las estructuras de grupos definidas. El procedimiento se repitió 1000 veces para cada escenario.

La estrategia metodológica o combinación del Análisis de Coordenadas Principales (ACoP), Análisis de Conglomerados (AC) y el ajuste de un Biplot Logístico Externo (BLE) sobre las coordenadas principales fue valorada usando los siguientes criterios: la absorción de variancia retenida en las dos primeras dimensiones y su comparación con el estándar para este tipo de experimentos; la proyección de un grupo en la dirección de

un vector del alelo que esté presente en el grupo –conocido *a priori*- y por lo tanto que predice teóricamente la probabilidad de la presencia de este alelo en el grupo; la bondad de ajuste medida a través de la calidad de la representación de los alelos (CRA); el porcentaje de clasificación correcta de los alelos (PCCA); la calidad de la representación de los individuos (CRI) y la tasa de error de clasificación (TEC).

Tabla 7. Escenarios simulados

Escenarios	Número de grupos	Número de individuos/ grupos	Número de alelos*	Alelos suplementarios / Ruido externo	Número total de alelos	Ruido interno (%)
***S1*(a_1,b_1,c_1)**	2	(20,30)	12	30	42	(a_1:5,b_1:10,c_1:20)
***S1*(a_2,b_2,c_2)**	2	(20,30)	12	48	60	(a_2:5,b_2:10,c_2:20)
***S1*(a_3,b_3,c_3)**	2	(20,30)	12	66	78	(a_3:5,b_3:10,c_3:20)
***S2*(a_1,b_1,c_1)**	3	(10,15,25)	20	50	70	(a_1:5,b_1:10,c_1:20)
***S2*(a_2,b_2,c_2)**	3	(10,15,25)	20	80	100	(a_2:5,b_2:10,c_2:20)
***S2*(a_3,b_3,c_3)**	3	(10,15,25)	20	110	130	(a_3:5,b_3:10,c_3:20)
***S3*(a_1,b_1,c_1)**	4	(6,10,14,20)	34	85	119	(a_1:5,b_1:10,c_1:20)
***S3*(a_2,b_2,c_2)**	4	(6,10,14,20)	34	136	170	(a_2:5,b_2:10,c_2:20)
***S3*(a_3,b_3,c_3)**	4	(6,10,14,20)	34	187	221	(a_3:5,b_3:10,c_3:20)

*Alelos que definen la estructura de grupo

3.2.2 Resultados

La Figura 23 muestra la distribución de la varianza para los diferentes escenarios. Los valores propios y su patrón se encuentran dentro de los límites de los reportados para estudios de diversidad genética usando marcadores moleculares. Una amplificación del plano principal usado en la representación de Biplot se muestra en la Figura 24. Los escenarios ***S1*** logran acumular la mayor variación, se observa además que los tipos de ruido (externo e interno) influyen en forma diferencial. La tasa de acumulación de la varianza se afecta más debido al ruido interno que respecto al número de alelos

suplementarios que sean agregados. Esto puede estar indicando que la estructura de variación es más afectada por los errores debido a la técnica molecular que al número de loci que se prueben de manera simultánea; así mismo, se comprueba una vez más la capacidad que posee el Análisis de Coordenadas Principales de separar el ruido de la señal, aspecto que fue demostrado en apartes anteriores.

Las Figuras 25, 26, 27 y 28 muestran la calidad de la representación y el porcentaje de clasificación correcta de los alelos (CRA y PCCA) con y sin estructura de grupos, de los individuos (CRI), y la tasa de error de clasificación (TEC). Al igual que con la varianza acumulada se observa cómo, independientemente del grado de ruido externo, los escenarios con mayor ruido interno son los que presentan los valores más bajos, siendo en los escenarios con errores internos del 5% dónde la calidad de la representación y el porcentaje de clasificación correcta de los alelos con estructura de grupos resulta superior al 80% y 90%, respectivamente. Nótese que como es de esperar, estos criterios poseen valores bajos cuando los alelos no presentan estructura de grupo. Respecto a la calidad de la representación media de los individuos, aunque mantiene el mismo patrón respecto a la influencia del error interno, los valores que se obtienen son bajos debido a la influencia que tiene el ruido externo sobre la precisión de las coordenadas.

Otro aspecto a considerar y que muestra la bondad del método es que las estructuras de grupo diseñadas previamente se mantienen. Esto indica que el número de alelos suplementarios no afecta a la estructura de los grupos definidos previamente; es decir, la reducción de la dimensionalidad y la representación de los individuos en el plano

bidimensional no están afectadas incluso cuando se obtengan valores bajos de absorción de varianza. Heoa y Gabriel (2001) demostraron que las representaciones gráficas de datos multivariantes exhiben a menudo las estructuras esperadas, revelando características como formación de grupos o patrones de correlación, aunque por la naturaleza misma de los datos muestren ajustes pobres.

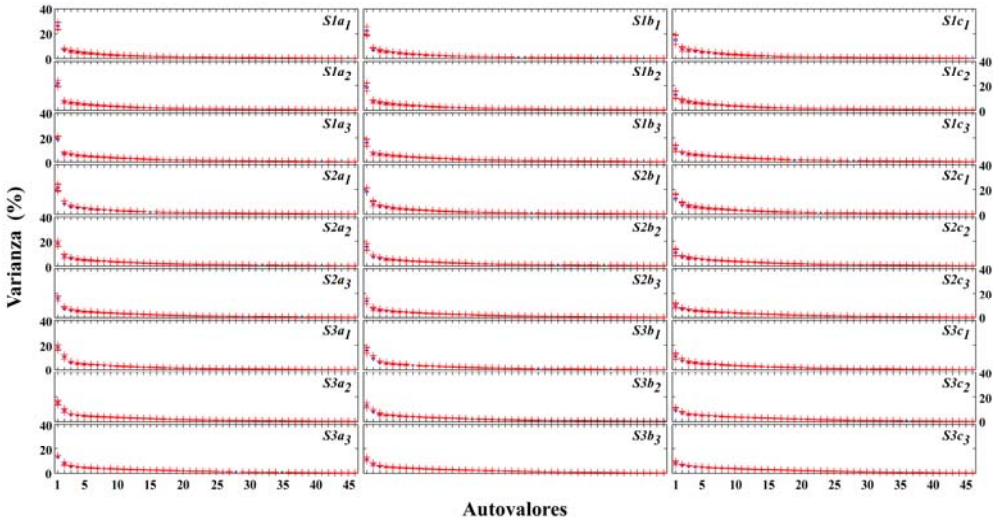


Figura 23. Distribución de la varianza para todas las dimensiones en los diferentes escenarios.

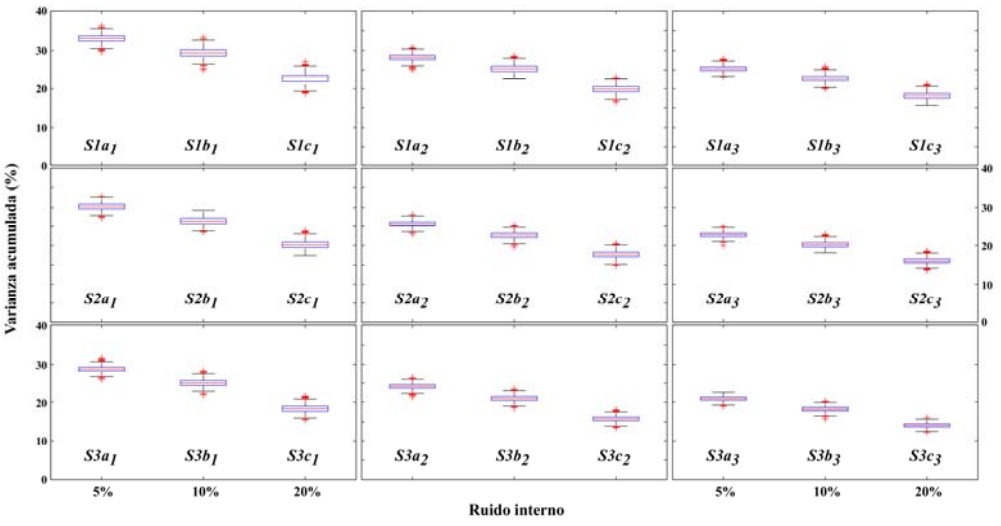


Figura 24. Distribución de la varianza acumulada para las primeras dos dimensiones en los diferentes escenarios

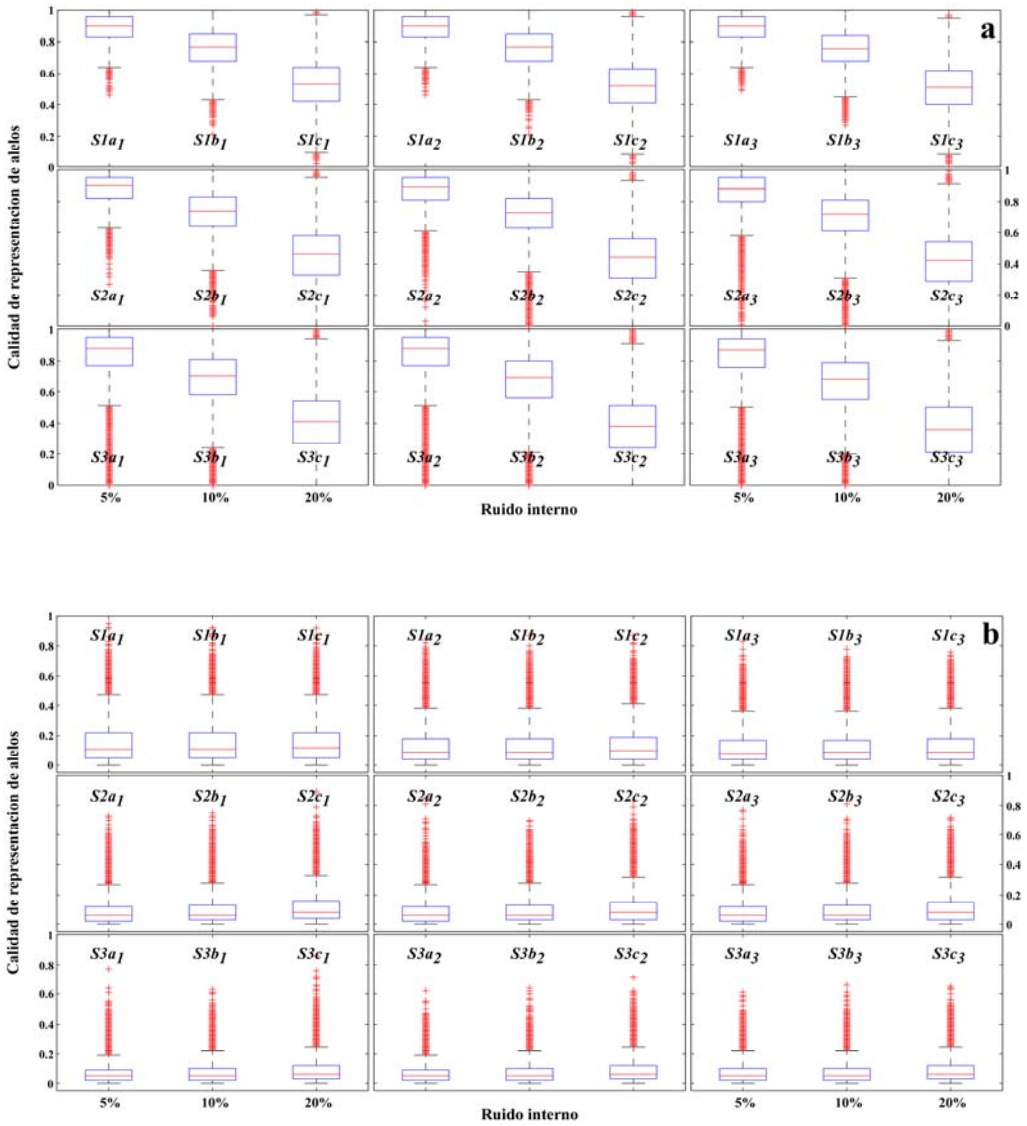


Figura 25. Calidad de la representación de alelos (variables) en los diferentes escenarios: (**a**) alelos con estructura de grupo y (**b**) alelos suplementarios.

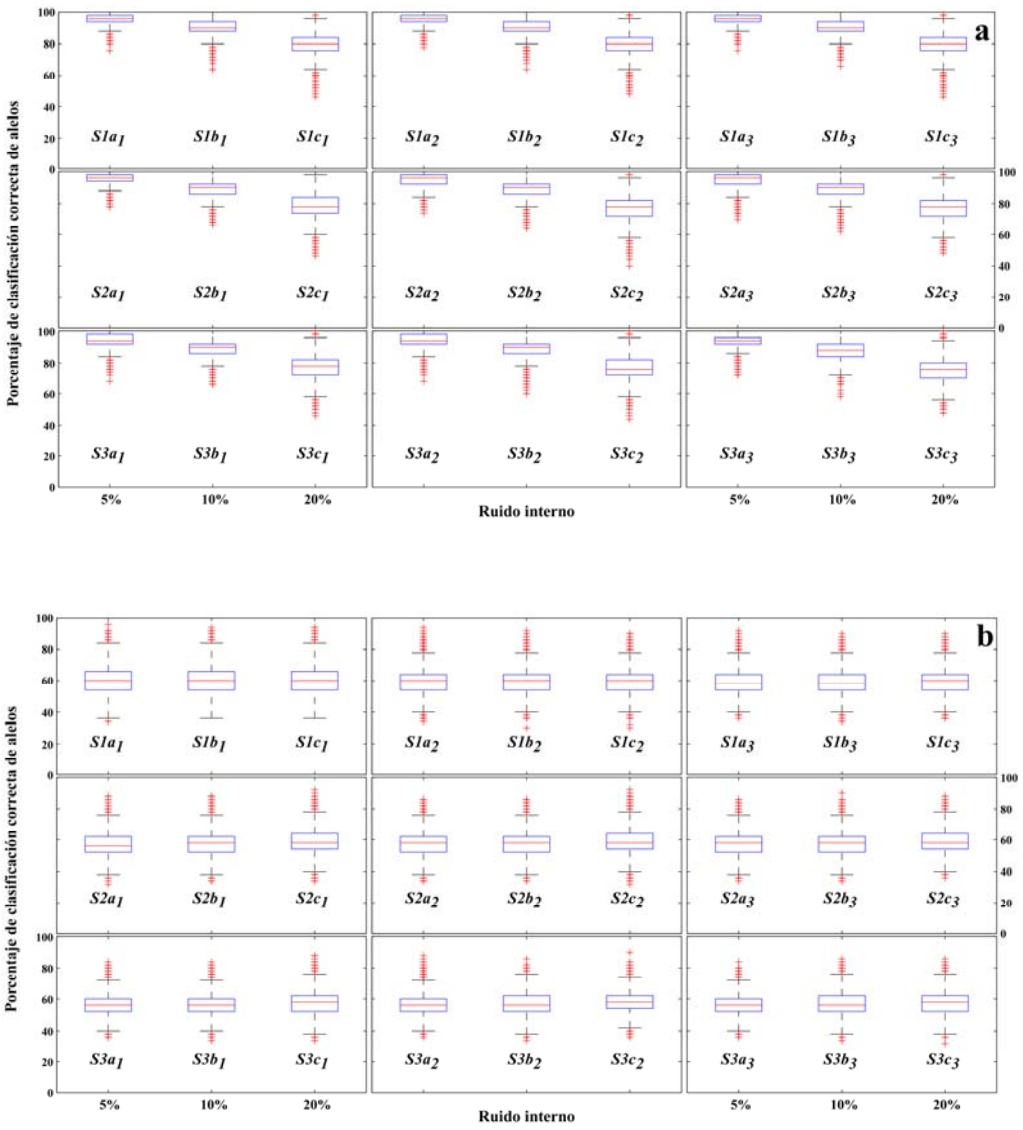


Figura 26. Porcentaje de clasificación correcta de alelos (variables) en los diferentes escenarios: (**a**) alelos con estructura de grupo y (**b**) alelos suplementarios.

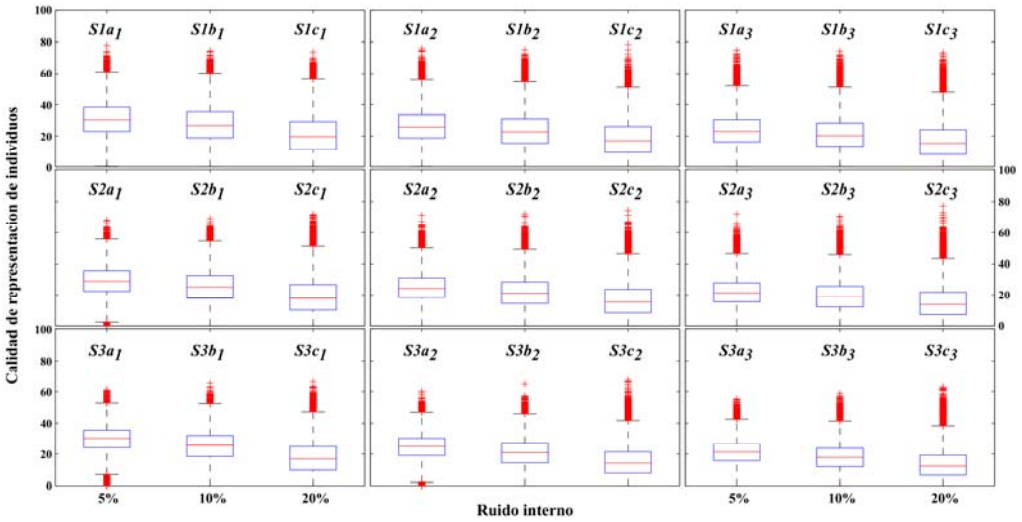


Figura 27. Calidad de la representación de los individuos en los diferentes escenarios.

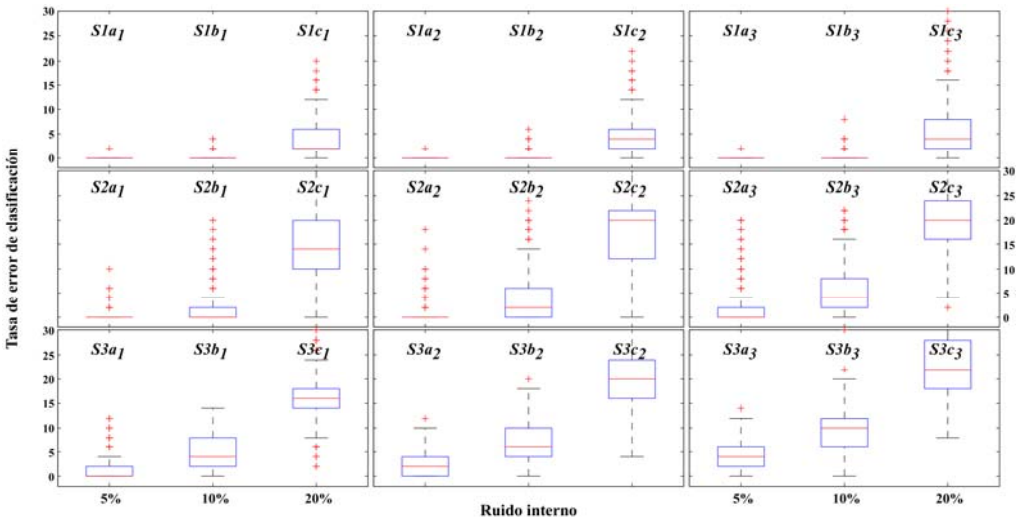


Figura 28. Tasa de error de clasificación en los diferentes escenarios**.**

Como ejemplo, la Figura 29 muestra la representación de Biplot de las relaciones entre los individuos y los alelos que determinan la estructura del grupo para algunos de los escenarios estudiados. Se observa que los alelos que definen los grupos están proyectados en las direcciones de la mayor probabilidad para estos –presentan el mismo color que el grupo que definen. Así mismo, se observa que una vez aplicada la corrección de Bonferroni la mayoría de los alelos suplementarios son eliminados y los que aun son proyectados en el Biplot, aparecen en forma irregular y con longitudes mayores lo que indica una baja capacidad discriminatoria –vectores de color negro.

Únicamente en el caso de la Figura 29c, el grupo 2 no muestra los alelos asociados a su formación, y esto es debido a que generalmente son necesarios al menos ***g-1*** ejes para conservar la estructura de grupos. En este caso es probable que se necesitase incluir un tercer eje para poder observar más claramente las variables que se asocian a la estructura de este grupo. Todos los alelos que definen la estructura de los grupos –como fue mostrado en las Figuras 25 y 26, presentan una alta calidad de representación y un alto porcentaje de clasificación correcta, indicando que existe un buen nivel de coincidencias entre la matriz de los datos binarios original y la matriz estimada con los modelos de regresión logística.

Si bien los grupos fueron definidos en la configuración de cada uno de los escenarios, se utilizó el algoritmo de agrupamiento UPGMA para confirmar estas estructuras a través del cálculo de la tasa de error de clasificación (TEC). Esta tasa permite comprobar la sensibilidad del método cuando la estructura no se sabe *a priori*. Es así que, de manera

semejante a los otros criterios de calidad, la tasa de error de clasificación tiene un comportamiento similar respecto a ruido interno. Quedando demostrado claramente la sensibilidad del método por el hecho que en todos los escenarios las tasas de error no superaron el 25% en el peor de casos y fueron menores del 9% cuando el error interno o error debido a la técnica molecular fue menor del 10%.

Finalmente, podemos concluir que independientemente del número de alelos o fragmentos de la amplificación que se evalúen, el ajuste de un Biplot Logístico Externo (BLE) sobre las coordenadas principales permite identificar los alelos de mayor importancia en la definición de la estructura natural de los individuos en las primeras coordenadas principales y que el algoritmo de agrupamiento asigna correctamente, en los grupos conocidos *a priori*, a prácticamente el 100% de los genotipos.

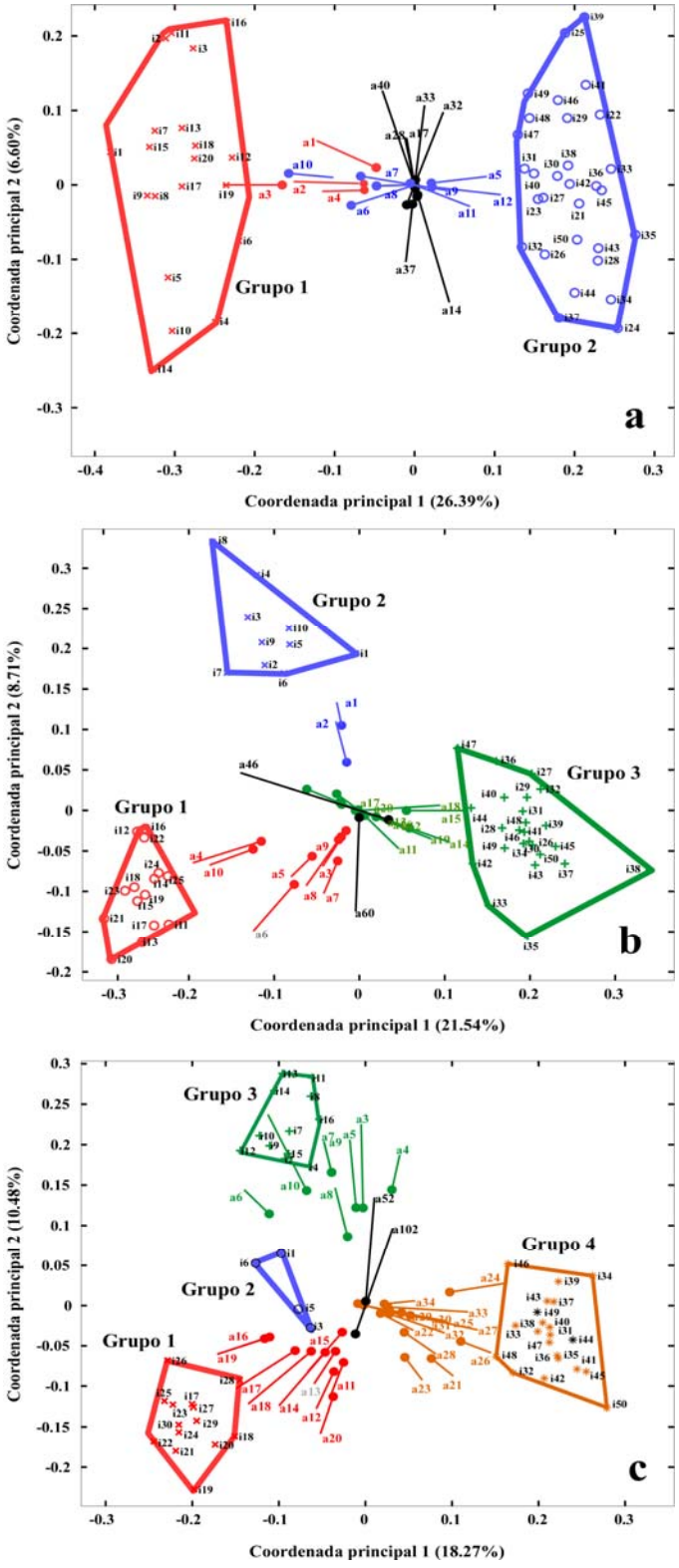


Figura 29. Representación Biplot mostrando la relación entre individuos y alelos, basada en el coeficiente de disimilitud de Dice para los escenarios: (**a**) ***S1a*$_1$**, (**b**) ***S2a*$_1$** y (**c**) ***S3a*$_1$**.

3.3 APLICACIÓN A DATOS REALES

A continuación y con el objeto de ilustrar el desarrollo metodológico propuesto se evaluó la diversidad genética existente en el banco nacional de germoplasma de caña de azúcar del Instituto Nacional de Investigaciones Agrícolas (INIA) de Venezuela.

La caña de azúcar representa para Venezuela la única fuente en la producción de sacarosa y además es utilizada en generación de bioetanol, constituyendo así uno de los cultivos industriales más importantes del país. Por esta razón, cualquier estudio que contribuya a la organización funcional de la información de los genotipos susceptibles de conservación, investigación, mejoramiento y transformación es de transcendental importancia en los programas agrícolas.

Las variedades cultivadas de caña de azúcar (*Saccharum* spp) son altamente poliploides (2n=100-130), derivadas de hibridaciones inter específicas entre las especies silvestres *Saccharum spontaneum* (2n=40-128) y las especies productoras *Saccharum officinarum* (2n=60 o 80) (Butterfield *et al.*, 2001; D'Hont y Glaszmann, 2001). La complejidad del genoma de estas variedades está determinada por la ocurrencia de eventos de poliploidia, aneuploidia y origen multiespecífico (Besse *et al.*, 1997).

En la caracterización y evaluación de la diversidad genética de caña de azúcar han sido usadas varias técnicas moleculares, incluyendo RFLP (Lu *et al.*, 1994ab; Besse *et al.*, 1997), RAPD (Harvey y Botha, 1996; Burner *et al.*, 1997; Vijayan *et al.*, 1999), AFLP (Besse *et al.*, 1998; Xu *et al.*; 1999; Selvi *et al.*, 2005) y SSR (Da Silva, 2001; Schenck

et al., 2004; Cordeiro *et al.*, 2000, 2003, y 2006). De todas estas técnicas, RAPD ha sido la más empleada ya que es simple y fácil a pesar de las incógnitas referentes a su reproducibilidad (Williams *et al.,* 1990 y 1993 y Tingey y del Tufo, 1993).

En este sentido se evaluó la diversidad genética existente entre cincuenta variedades de Caña de Azúcar a través de marcadores moleculares RAPD.

3.3.1 Materiales y métodos

Material Vegetal, Extracción y Amplificación ADN: Se realizó la extracción de ADN genómico total según metodología de Zambrano *et al.* (2002) en tres muestras de tejido foliar joven de los cultivares B37-161, B47-47, B44-341, B41-227, B43-62, B49-119, B64-129, B67-49, B64-136, B75-542, B75-49, B75-403, B76-226, B82-157, SP71-1406, C323-68, C371-67, Co421, Co740, CP56-59, CP72-1210, CP74-2005, CP72-2086, CL41223, MEX641487, POJ2878, POJ29-61, Ragnar, PR1013, PR61-632, PR62-258, PR69-2176, PR980, V58-4, V64-10, V68-74, V71-39, V74-7, V75-6, V77-12, V77-24, V78-1, 4-51-48, 4-51-33, 31-53-1, 4-51-32, 3-54-3, 1-54-2, 118-53-19 y 48-55-4, los cuales constituyen una muestra representativa de las entradas del banco nacional de germoplasma de caña de azúcar ubicado en CIAE-Yaritagua, Yaracuy, Venezuela. La amplificación fue realizada según metodología descrita por Zambrano *et al.* (2003) utilizando ocho iniciadores RAPD de Operon Technologies Inc: OPA-07, OPM-04, OPM-16, OPM-18, OPY-04, OPY-07, OPY-09 y OPY-17.

Análisis de Datos: Debido a la naturaleza poliploide de la caña de azúcar y en ausencia de análisis de segregación no se hizo ningún supuesto sobre la naturaleza genética de los alelos (Cordeiro *et al*., 2003). Los fragmentos de amplificación fueron codificados de acuerdo a un marcador dominante, es decir, $A_1A_1 = A_1A_2 = 1$ y $A_2A_2 = 0$, generando una columna por locus para cada iniciador. La relación genética entre los 50 cultivares fueron estudiadas usando el Análisis de Coordenadas Principales (ACoP), Análisis de Conglomerados (AC) y el ajuste de un Biplot Logístico Externo (BLE) sobre datos de disimilitud utilizando los coeficientes de Jaccard, Emparejamiento simple, Dice y Rogers y Tanimoto (Sneath y Sokal, 1973). El número *k* de dimensiones a ser retenidas, el coeficiente de similitud que mejor define la estructura de los datos y las medidas de la calidad fueron calculados utilizando los procedimientos descritos en apartados previos.

3.3.2 Resultados

Los ocho iniciadores aleatorios (RAPD) utilizados para la amplificación produjeron un total de 103 fragmentos de amplificación polimórficos, Tabla 8, con un tamaño entre 200-4270 bp. El menor número de fragmentos fue generado con el iniciador OPM-18 con un total de siete y el mayor número fue amplificado con el iniciador OPA-07 con un total de veinte. La Figura 30, muestra la huella digital de los 103 fragmentos polimórficos; la alternativa alélica $A_1A_1 = A_1A_2 = 1$ y $A_2A_2 = 0$ se muestran en negro y blanco respectivamente. Visualmente la huella ofrece una idea de la variabilidad esperada entre los 50 cultivares.

Tabla 8. Fragmentos Amplificados por cada iniciador en los cultivares caña de azúcar

Iniciador RAPD	Rango de Amplificación (bp)	Total de Fragmentos amplificados	Fragmentos amplificados
OPA-07	**400-4200**	**20**	OPA-07(4200pb), OPA-07(1904pb), OPA-07(1800pb), OPA-07(1700pb), OPA-07(1650pb), OPA-07(1600bp), OPA-07(1375bp), OPA-07(1300bp), OPA-07(1250bp), OPA-07(1200bp), OPA-07(1150bp), OPA-07(1000bp), OPA-07(950bp), OPA-07(900bp), OPA-07(850bp), OPA-07(800bp), OPA-07(700bp), OPA-07(564bp), OPA-07(500bp), OPA-07(400bp),
OPM-04	**700-3530**	**12**	OPM-04(3530bp), OPM-04(2000bp), OPM-04(1904bp), OPM-04(1800bp), OPM-04(1790bp), OPM-04(1584bp), OPM-04(1450bp), OPM-04(1375bp), OPM-04(1200bp), OPM-04(1190bp), OPM-04(947bp), OPM-04(700bp)
OPM-16	**831-2027**	**10**	OPM-16(2027bp), OPM-16(1950bp), OPM-16(1900bp), OPM-16(1800bp), OPM-16(1750bp), OPM-16(1700bp), OPM-16(1375bp), OPM-16(1100bp), OPM-16(947bp), OPM-16(831bp)
OPM-18	**560-1450**	**7**	OPM-18(1450bp), OPM-18(1000bp), OPM-18(950bp), OPM-18(900bp), OPM-18(831bp), OPM-18(800bp), OPM-18(560bp)
OPY-04	**560-1950**	**11**	OPY-04(1950bp), OPY-04(1900bp), OPY-04(1650bp), OPY-04(1375bp), OPY-04(1200bp), OPY-04(900bp), OPY-04(850bp), OPY-04(800bp), OPY-04(750bp), OPY-04(600bp), OPY-04(560bp)
OPY-07	**200-2100**	**14**	OPY-07(2100bp), OPY-07(2000bp), OPY-07(1800bp), OPY-07(1400bp), OPY-07(1300bp), OPY-07(1200bp), OPY-07(1000bp), OPY-07(940bp), OPY-07(800bp), OPY-07(600bp), OPY-07(560bp), OPY-07(400bp), OPY-07(300bp), OPY-07(200bp)
OPY-09	**560-4270**	**16**	OPY-09(4270bp), OPY-09(4000bp), OPY-09(3500bp), OPY-09(2027bp), OPY-09(2000bp), OPY-09(1900bp), OPY-09(1800bp), OPY-09(1700bp), OPY-09(1600bp), OPY-09(1590bp), OPY-09(1400bp), OPY-09(1375bp), OPY-09(1300bp), OPY-09(1200bp), OPY-09(940bp), OPY-09(560bp)
OPY-17	**400-3530**	**13**	OPY-17(3530bp), OPY-17(2100bp), OPY-17(1700bp), OPY-17(1600bp), OPY-17(1400bp), OPY-17(1200bp), OPY-17(1000bp), OPY-17(947bp), OPY-17(831bp), OPY-17(800bp), OPY-17(600bp), OPY-17(564bp), OPY-17(400bp)
Total		**103**	

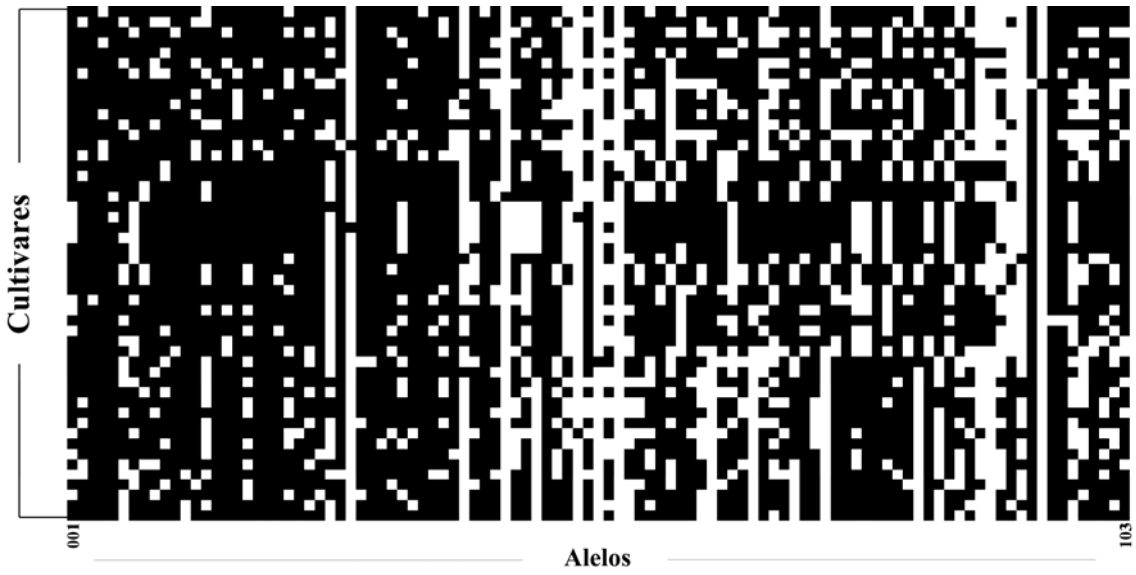


Figura 30. Huella digital de los 103 fragmentos polimórficos para los 50 cultivares.

La Figura 31, muestra la distribución de los coeficientes de correlación lineal de Pearson entre los *n(n-1)/2* elementos distintos fuera de la diagonal de las matrices de distancias observada $\boldsymbol{\Delta}$ y estimada $\mathbf{D}$, para distintas combinaciones de *k*-dimensiones retenidas luego de aplicar el Análisis de Coordenadas Principales (ACoP). Para las dos primeras dimensiones el valor más alto ($\boldsymbol{r}$=0.8187) se obtuvo cuando la matriz $\boldsymbol{\Delta}$ fue calculada utilizando el coeficiente de similitud de Dice, indicando que esta opción es la que refleja la mayor coherencia entre la matriz de distancias observadas y estimadas. Esta combinación permite reconocer mejor la relación entre individuos, entendida en términos de similitud, es decir, utilizando el coeficiente de similitud de Dice y las dos primeras coordenadas. Así, dos individuos con posiciones más cercanas en la representación bidimensional, tendrán patrones más similares de ADN respecto a las secuencias aleatorias utilizadas. Obsérvese además que para alcanzar valores de $\boldsymbol{r}\geq 0.8$

cuando se utiliza el coeficiente de Emparejamiento simple o los coeficientes de Jaccard y Rogers y Tanimoto deben retenerse al menos tres o seis coordenadas, respectivamente.

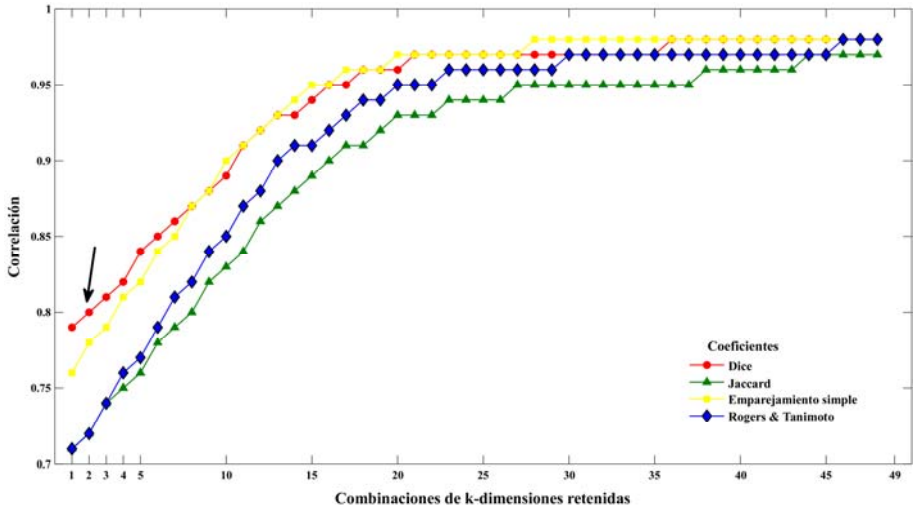


Figura 31. Distribución de las correlaciones entre las matrices de distancias observadas y estimadas para diferentes coeficientes de similitud y combinaciones de *k*-dimensiones retenidas.

La Figura 32, muestra en forma de pasos la estrategia metodológica que se propone: se inicia con el estudio de las relaciones entre individuos a través de su proyección bidimensional utilizando el Análisis de Coordenadas Principales (ACoP), se continua con la clasificación en grupos generados por el algoritmo de agrupamiento UPGMA sobre las dos primeras coordenadas, luego se estima la variabilidad muestral de los individuos relacionada con el grupo al que pertenece y por último se realizan las distintas fases del ajuste del Biplot Logístico Externo (BLE) hasta llegar a la selección

de las variables (alelos) que definen las estructuras de grupos y permiten observar sus relaciones.

Es así como en la Figura 32ab, se muestra el espacio bidimensional obtenido del Análisis de Coordenadas Principales (ACoP): las dos primeras dimensiones explican el 37.50% de la variabilidad total y permiten la formación de cinco grupos de cultivares utilizando los ocho iniciadores RAPD. El primer grupo formado por los cultivares: B43-62, C323-68, C371-67, CP56-59, CP72-1210, CP74-2005, CP72-2086, CL41223, MEX641487, Ragnar, PR1013, PR69-2176 y PR980, el segundo grupo formado por los cultivares B44-341, B41-227, B67-49, B76-226, PR61-632, PR62-258 y V58-4; el tercer grupo formado por los cultivares B37-161, B49-119, B64-129, B64-136, B75-542, B75-49, B75-403, B82-157, SP71-1406, Co421 y Co740, el cuarto grupo formado por los cultivares B47-47, POJ2878, POJ29-61, V64-10, V68-74, V74-7, V77-24, 31-53-1, 4-51-32 y 3-54-3 y el quito grupo formado por los cultivares V71-39, V75-6, V77-12, V78-1, 4-51-48, 4-51-33, 1-54-2, 118-53-19, 48-55-4, con una similaridad genética media de 0.7005±0.0125, 0.5867±0.0332, 0.5958±0.0134, 0.6642±0.0128, 0.6742±0.0114**,** y una calidad de representación calculada con las dos primeras dimensiones de 96.70%, 38.78%, 89.69%, 60.66%, 90.83%, para el primero, segundo, tercero, cuarto y quinto grupo, respectivamente. Las bajas calidades de representación observadas para los grupos 2 y 4, sugieren que estos puedan tener una mejor representación en otra dimensión

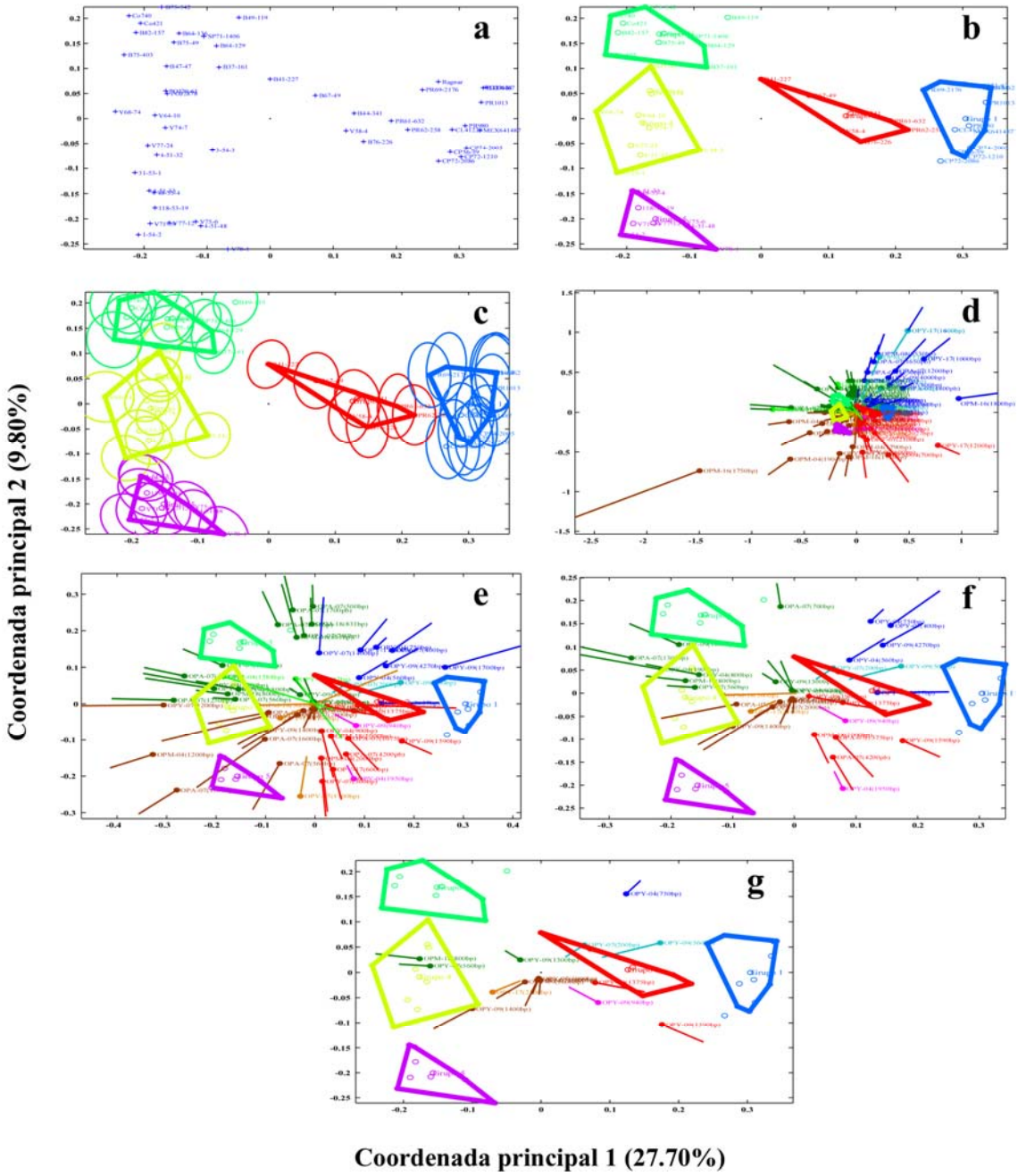


Figura 32. Relaciones genéticas entre los 50 cultivares de caña de azúcar basada en la disimilaridad debida al coeficiente de Dice y los ocho iniciadores RAPD: (**a**) representación en el plano las coordenadas principales; (**b**) grupos obtenidos bajo el algoritmo UPGMA utilizando las dos primeras coordenadas principales retenidas; (**c**) variabilidad muestral de los individuos; (**d**) representación después del ajuste del Biplot Logístico Externo (BLE); (**e**) representación después del ajuste Biplot corregida por el *p*-valor; (**f**) representación después del ajuste Biplot corregida por el *p*-valor y Bonferroni y (**g**) representación después del ajuste Biplot corregida por el *p*-valor, Bonferroni y el pseudo R^2 de Nagelkerke/Cragg & Uhler's.

La Figura 32c, muestra la variabilidad muestral de la configuración generada a través del remuestreo sobre los residuales usando como método de transformación el Procrustes: se observa que para una confianza de 75% las elipses muestran, en la mayoría de los casos, una amplitud moderada siendo mayor en el primer eje que en el segundo. Así mismo, el análisis revela que la configuración presenta una estabilidad de 97.51%; es decir, los individuos serán proyectados en esas coordenadas, en promedio, con un error menor del 3%. Indicando que los ejes retenidos son suficientes para recoger información sobre la visualización de los individuos en la representación bidimensional. Estos resultados confirman los obtenidos por Heoa y Gabriel (2001) así como los del estudio de simulación presentando en el aparte anterior, ratificando que la reducción de la dimensionalidad y la representación de los individuos en el plano bidimensional no están afectadas incluso cuando se obtengan valores bajos de absorción de varianza.

Las Figura 32d, muestra el primer paso del ajuste del Biplot Logístico Externo (BLE) sin haber iniciado el proceso de selección de variables, obsérvese que la escala de los ejes incrementan considerablemente para poder proyectar variables que se alejan o no tienen ninguna relación con la clasificación. Nótese que una vez aplicados acumulativamente los criterios de corrección del ajuste por el p-valor ($p \leq 0.01$), Bonferroni y el pseudo R^2 de Nagelkerke/Cragg & Uhler's ($R^2 \geq 0.75$), disminuye considerablemente el número de variables y solo se muestran las que están relacionadas con la clasificación de los individuos, Figuras 32efg.

Después del ajuste en el Biplot, el porcentaje de coincidencias entre la matriz de los datos binarios original y la estimada de los modelos logístico o Porcentaje de Clasificación Correcta (PCC) es de 85.75%. La media de calidad de representación y el porcentaje de clasificación correcta de los individuos fue de 36.86%± 2.34 y 84.13% ±0.9388, respectivamente, de los 50 cultivares proyectados en la representación el 75% mostró una bondad de ajuste superior al 80%.

La Tabla 9, muestra la lista de alelos seleccionados después del ajuste Biplot y como ayuda a la interpretación presenta los cosenos de los ángulos formados por las direcciones y las dimensiones, la dimensión donde el coseno es más alto, el cuadrante donde está ubicado y la dirección de cada alelo.

Tabla 9. Alelos seleccionados después del ajuste Biplot corregido por el *p*-valor, Bonferroni y el pseudo R^2 de Nagelkerke/Cragg & Uhler's

Variables	p-valor	Pseudo R^2	Coordenadas		Coseno		Eje de máximo coseno	Situación plano(1-2)	
			Eje 1	Eje 2	Eje 1	Eje 2		Cuadrante	Dirección
OPM-18(900bp)	5.8291E-12	0.8179	-0.0009	-0.0157	-0.0599	-0.9982	2	3	1
OPM-18(800bp)	3.6087E-12	0.7643	-0.1765	0.0271	-0.9884	0.1516	1	2	1
OPY-04(750bp)	1.1102E-16	0.9538	0.1238	0.1552	0.6236	0.7817	2	1	1
OPY-07(1800bp)	4.7588E-11	0.7825	-0.0053	-0.0151	-0.3302	-0.9439	2	3	1
OPY-07(1000bp)	8.1080E-12	0.8258	-0.0034	-0.0116	-0.2830	-0.9591	2	3	1
OPY-07(560bp)	0.0000	0.9182	-0.1612	0.0127	-0.9969	0.0788	1	2	1
OPY-07(200bp)	0.0000	0.9459	0.0630	0.0544	0.7567	0.6537	1	1	-1
OPY-09(1600bp)	4.7042E-12	0.8091	-0.0233	-0.0186	-0.7826	-0.6225	1	3	1
OPY-09(1590bp)	7.1831E-13	0.8308	0.1758	-0.1033	0.8622	-0.5066	1	4	1
OPY-09(1400bp)	1.8088E-11	0.7624	-0.1001	-0.0716	-0.8134	-0.5817	1	3	1
OPY-09(1375bp)	2.9058E-12	0.8294	0.0787	-0.0200	0.9692	-0.2465	1	4	1
OPY-09(1300bp)	0.0000	0.9590	-0.0302	0.0249	-0.7715	0.6362	1	2	1
OPY-09(940bp)	4.4409E-16	0.9025	0.0829	-0.0594	0.8131	-0.5821	1	4	-1
OPY-09(560bp)	9.6940E-11	0.7607	0.1732	0.0585	0.9475	0.3198	1	1	-1
OPY-17(2100bp)	0.0000	0.9045	-0.0705	-0.0390	-0.8752	-0.4838	1	3	-1

Si bien los resultados presentados muestran que dos ejes son suficientes para el estudio de la diversidad genética de los 50 cultivares de caña de azúcar, como ilustración, la

Figura 33ab muestra la proyección tridimensional de la variabilidad muestral de los individuos al 75% de confianza y las regiones generadas para los cinco grupos al 95% de confianza y el conjunto de alelos seleccionados. Obsérvese que en esta representación el grupo 4 presenta el eje 3 con mayor dimensión respecto al resto de grupos, esta variación geométrica permite intuir la razón de la baja calidad de representación de este grupo al utilizar solo dos ejes en la representación. Este comportamiento muy probablemente se deba a la influencia del cultivar B47-47, que si bien en el plano bidimensional se observa más cerca del grupo 3 en realidad su calidad de representación incrementa al aumentar las dimensiones retenidas. Respecto al grupo 2 no se detecta la influencia del tercer eje.

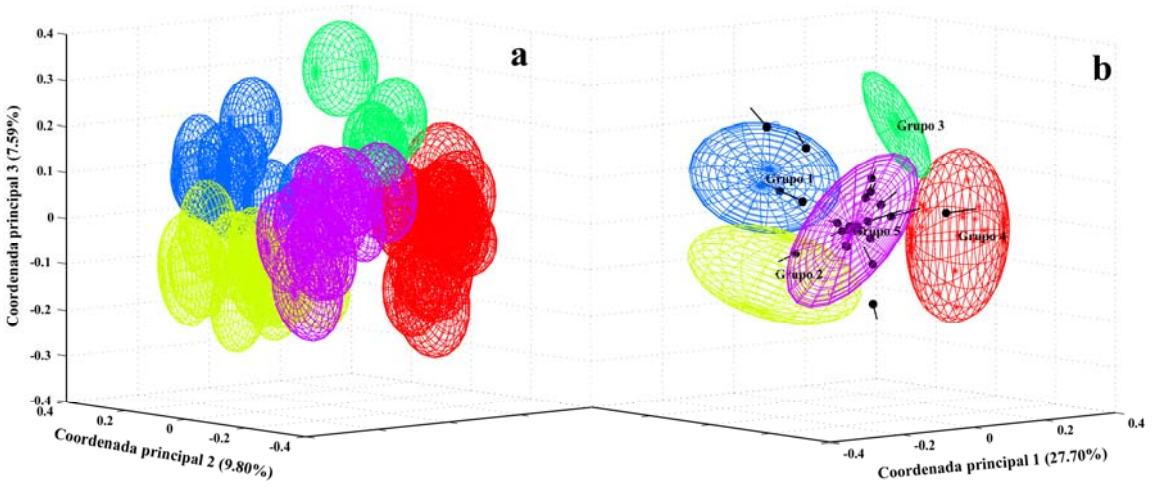


Figura 33. Relaciones genéticas entre los 50 cultivares de caña de azúcar basada en la disimilaridad debida al coeficiente de Dice y los ocho iniciadores RAPD: (**a**) variabilidad muestral de los individuos y (**b**) representación después del ajuste Biplot corregida por el *p*-valor, Bonferroni y el pseudo R^2 de Nagelkerke/Cragg & Uhler's.

El tipo de representación Biplot generada, demuestra que es necesario conocer las variables responsables de la clasificación, ya que un conjunto de estas solo aportan ruido. En el caso de estudios de diversidad genética con el objetivo de reconocer alelos o grupos de estos asociados a características específicas del cultivo, este tipo de representación permitirá ir afinado el proceso de selección hasta obtener el grupo de alelos susceptibles de ser estudiados en profundidad.

Sin pretender que los resultados presentados abordan en su totalidad la discusión requerida en cualquier estudio de diversidad genética, el ejemplo presentado ha permitido demostrar que comparado con el enfoque clásico, el uso combinado del Análisis de Coordenadas Principales (ACoP), Análisis de Conglomerados (AC) y el ajuste de un Biplot Logístico Externo (BLE) ofrece una comprensión holística de la estructura de datos, facilita las interpretaciones de los resultados y puede ser altamente recomendado para una descripción cuidadosa de datos en estudios de la diversidad genética usando marcadores de ADN.

Adicionalmente, el procedimiento descrito ha sido aplicado satisfactoriamente a estudios de diversidad genética de poblaciones humanas y en grandes conjuntos de datos. En este sentido y como material anexo a esta memoria se presenta el trabajo publicado en la revista Bioinformatics, (Demey *et al*., 2008), donde se ilustra la aplicación de la propuesta metodológica con datos reales de cuatro poblaciones de África, Asia y Europa, usando marcadores SNPs generados por el International HapMap Consortium (2003).

Capítulo IV

RELACIÓN ENTRE MARCADORES

CAPITULO IV

Se ha abordado, en apartados anteriores, la estructura multidimensional que tiene el complejo genético de un individuo o grupo de individuos. Por esta razón, estudios que incorporen diferentes marcadores, tales como descriptores agromorfológicos y marcadores moleculares o simplemente diferentes tipos de marcadores moleculares, proveen una mejor descripción e interpretación de la diversidad genética de los individuos, bien sea porque se incorporan características asociadas a caracteres altamente heredables que no interaccionan con el ambiente y tienen importancia económica o porque la región del genoma que se explore sea mayor (Wilson *et al.*, 1974, 1977; Russell *et al*., 1997; Hillis y Wiens, 2000; Demey *et al.*, 2003). En otras palabras, si se estudian en forma conjunta los subespacios que generan descriptores agronómicos, morfológicos, marcadores dominantes o codominantes se obtendrá una mejor valoración de la diversidad genética de los individuos y adicionalmente será posible cuantificar el grado de asociación entre los diferentes descriptores y/o marcadores. Sin embargo, Wilson *et al.*, han demostrado que las caracterizaciones generadas a través de descriptores morfológicos y marcadores moleculares suelen ser independientes respondiendo en cada caso a reglas y presiones evolutivas diferentes.

No obstante esta premisa, se ha considerado que la independencia no es un problema de índole genético, sino de que no son utilizadas las herramientas metodológicas que permitan cuantificar y tipificar el tipo de relación y determinar cuánto de la variabilidad es explicada por uno u otro marcador. En este sentido, cuando se estudian las relaciones entre diferentes marcadores, en la amplia mayoría de la literatura, se percibe que la

concordancia entre caracterizaciones generadas por diferentes tipos de marcadores, se valora a través de la correlaciones entre matrices de distancias y/o similaridades (Lanza *et al.*, 1997; Franco *et al*., 2001; Roldán-Ruiz *et al*., 2001; Tar'an *et al*., 2005; García *et al*., 2007; Syamkumar y Sasikumar, 2007; Kalita *et al*., 2007).

La razón de este enfoque es que el análisis conjunto y las relaciones que generan los diferentes subespacios no es simple. El problema se centra en encontrar una medida de distancia o de similitud adecuada que permita relacionar a los individuos con las diversas características que han sido observadas de manera simultánea y medidas en diferentes escalas, dimensiones o espacios. Así por ejemplo, si se observa la proyección bidimensional generada por el Análisis de Coordenadas Principales (ACoP) de un conjunto de individuos a los cuáles se les ha observado simultáneamente su expresión frente a un marcador dominante y a otro codominante, es posible que no solo se generen sistemas de agrupaciones diferentes sino que además los ejes de las coordenadas principales tengan escalas disímiles que no permitan comparación o representación en el mismo sistema de coordenadas, sin antes realizar alguna transformación. Esto se debe tanto en el Análisis de Coordenadas Principales (ACoP) como en cualquiera de las otras técnicas de ordenación, a que las direcciones de las dimensiones que se generan son arbitrarias y dependen estrechamente del tipo de atributo y de las escalas de medición, entre otros. En estos casos, una alternativa que puede ser explorada, es estudiar simultáneamente los subespacios generados por los diferentes descriptores y/o marcadores a través de lo que denominamos en apartado 1.2.4, coeficiente de similaridad de Gower (1971a) o coeficiente para datos mixtos. Sin embargo, aunque

este coeficiente permite estudiar de manera simultánea todas las variables, en el caso de su forma general no es posible asignar a priori valores a w_{ijk}, ya que esto implicaría conocer el tipo de relaciones que se quieran establecer entre los individuos, y adicionalmente significa suponer que existe para todos los descriptores y/o marcadores un único sistema de referencia, suposición que obviamente no es aplicable al contexto de estudios de diversidad genética.

Una estrategia metodológica, que permite corregir los problemas mencionados es la generación de configuraciones de consenso a partir de los subespacios considerados, aplicando métodos que permiten comparar ordenaciones correspondientes al mismo conjunto de objetos, donde se aplicaron diferentes métodos o se utilizaron distintas medidas de distancia. Estos métodos, que en la terminología para datos de tres modos, se conocen como "Multiway Set of Data", tratan de forma simultánea las matrices que generan los diferentes tipos de descriptores y/o marcadores y evitan el problema de recurrir a la correlación entre matrices de distancia donde la información no es independiente.

Entre otros, el Análisis de Procrustes Generalizado (Gower, 1975; Gower y Dijksterhuis, 2004), STATIS (Structuration des Tableaux A Trois Indices de la Statistique; ACT stands for Analyse Conjointe de Tableaux) (Lavit *et al*., 1994), Meta-Componentes Principales (Krzanowsky, 1979 y 1984) y Análisis de Componentes Principales Comunes (Flury, 1984 y 1988), son métodos estadísticos que permiten la integración de los subespacios generados por la medición de diferentes atributos sobre

el mismo grupo de individuos, los cuales se basan en la búsqueda de una configuración consenso a través de la aproximación de las diferentes configuraciones asociadas a cada tipo de marcador. No obstante, la popularidad que han adquirido estos métodos para el tratamiento de conjuntos múltiples de datos, son pocas las referencias donde se realiza un tratamiento de consenso bien sea entre marcadores morfológicos y moleculares o entre diferentes marcadores moleculares (Faccioli *et al.*, 1995, Milbourne *et al.*, 1997; Bramardi *et al.*, 2005; Lopes *et al.*, 2006; Esposito *et al.*, 2007), y ninguna referencia incluye en el análisis la proyección de las variables responsables de la configuración consenso.

En este orden de ideas se estudiará el análisis simultáneo y la relación entre marcadores a través de su integración utilizando el Análisis de Procrustes Generalizado (PGA) y se desarrollará una metodología alternativa basada en la aplicación de los métodos Biplot que permite la proyección de las variables responsables de la definición de grupos homogéneos en el espacio de consenso que facilita una mejor comprensión de la diversidad genética de bancos de germoplasma.

CAPITULO IV

4.1. ANALISIS DE PROCRUSTES GENERALIZADO

Sean $\mathbf{Y}_1$ y $\mathbf{Y}_2$ las matrices formadas por las *k* componentes o coordenadas principales para las filas de las matrices $\mathbf{X}_1$ y $\mathbf{X}_2$, resultantes del Análisis de Componentes Principales (ACP) o el Análisis de Coordenadas Principales, según la naturaleza de los descriptores y/o marcadores.

El Análisis Procrustes (AP) consiste en la búsqueda de un consenso entre $\mathbf{Y}_1$ y $\mathbf{Y}_2$, que represente la verdadera estructura subyacente a los datos, aplicando transformaciones o movimientos sobre la matriz $\mathbf{Y}_2$, manteniendo $\mathbf{Y}_1$ como matriz de referencia, de modo que se preserven las interdistancias entre los puntos hasta obtener el mejor ajuste entre ambas matrices. Desde el punto de vista algebraico el problema se reduce a buscar la transformación Procrustes que minimiza la función de pérdida L que cuantifica las discrepancias entre las matrices. Los movimientos que preservan las distancias entre puntos o que minimizan la función de pérdida son: la traslación y rotación de los ejes, y las reflexiones y dilataciones de la configuración completa (Tucker, 1958; Gower, 1975; Gower y Dijksterhuis, 2004).

Adicionalmente, el Análisis Procrustes (AP) permite identificar la estructura de grupos presentes, y determinar si la variabilidad resultante de la falta de integración se debe a las diferencias propias que aporta cada configuración o a que se están comparando espacios no equivalentes.

4.1.1 Transformación Procrustes

Consiste en transformar las coordenadas de $\mathbf{Y}_2$, manteniendo $\mathbf{Y}_1$ como configuración de referencia, hasta encontrar una nueva matriz $\mathbf{Z} = \rho \mathbf{Y}_2 \mathbf{T}$, donde ρ y $\mathbf{T}$ son un factor de escala y una matriz de transformación ortogonal, respectivamente, tal que las discrepancias entre $\mathbf{Y}_1$ y $\mathbf{Z}$ sean mínimas.

Es así como, deberán elegirse valores de ρ y $\mathbf{T}$ de forma que L se minimice, siendo L la suma de cuadrados de la interdistancias entre las nubes de puntos de las $\mathbf{Y}_1$ y $\mathbf{Y}_2$ configuraciones, cuya expresión es:

$$L = tr\left[\left(\mathbf{Y}_1 - r\mathbf{Y}_2\mathbf{T} \right)' \left(\mathbf{Y}_1 - r\mathbf{Y}_2\mathbf{T} \right) \right] \qquad [4.1]$$

Dado que la transformación de $\mathbf{Y}_2$ en $\mathbf{Y}_2\mathbf{T}$ afecta la configuración de la nube de puntos, se le impone a $\mathbf{T}$ la restricción de que sea ortogonal de tal forma que la transformación sobre $\mathbf{Y}_2$ pueda ser solo o una reflexión o una rotación. Es así que la expresión de la matriz de transformación que minimiza L estará dada por:

$$\mathbf{T} = \mathbf{U}_2 \mathbf{U}_1' \qquad [4.2]$$

donde $\mathbf{U}_1$ y $\mathbf{U}_2$ son las matrices de vectores propios de $\mathbf{Y}_1'\mathbf{Y}_2\mathbf{Y}_2'\mathbf{Y}_1$ y $\mathbf{Y}_2'\mathbf{Y}_1\mathbf{Y}_1'\mathbf{Y}_2$, respectivamente, adicionalmente solo se consideran en el cálculo de $\mathbf{T}$ los autovalores

no negativos, garantizando que la $tr\left(\mathbf{Y}_2\mathbf{T}\mathbf{Y}_1'\right)$ sea tan grande como sea posible. Este tipo de transformación se denomina rotación Procrustes.

Obsérvese, que como $\mathbf{Y}_1$ y $\mathbf{Y}_2$ son matrices formadas por *k* componentes o coordenadas principales referidos al mismo conjunto de individuos, sin pérdida de generalidad, se supondrá que todas las configuraciones están centradas para evitar el problema de la traslación de las configuraciones.

Respecto al reescalamiento su importancia está en el hecho de que aunque la correlación entre $\mathbf{Y}_1$ y $\mathbf{Y}_2\mathbf{T}$ fuese igual a 1, no necesariamente sus escalas tienen que coincidir, por esta razón es necesario dilatar o comprimir la matriz $\mathbf{Z}$ utilizando el escalar ρ, de forma tal que las escalas de $\mathbf{Y}_1$ y $\mathbf{Y}_2$ sean comparables, siendo ρ de la forma:

$$\rho = \frac{tr\left(\mathbf{T}'\mathbf{Y}_2'\mathbf{Z}\mathbf{Y}_1\right)}{tr\left(\mathbf{Y}_2'\mathbf{Z}\mathbf{Y}_1\right)} \qquad [4.3]$$

Si $\mathbf{Y}_1$ y $\mathbf{Y}_2$ no tienen la misma dimensión, porque son necesarias *k*-dimensiones diferentes para explicar la variabilidad en cada configuración, la matriz de más baja dimensión deberá completarse con ceros hasta alcanzar la dimensión de la mayor. Así mismo, para garantizar la simetría de las transformaciones se recomienda su

estandarización al inicio del proceso de transformación. Nótese que tanto la rotación como el reescalado pueden ser aplicados de forma independientemente.

La idea presentada para dos configuraciones ha sido generalizada por Gower (1975) para *M* matrices, donde todas las configuraciones son comparadas a una configuración de referencia, obteniéndose una única configuración de consenso. La configuración de referencia es la media de las configuraciones individuales, esta transformación para *M* configuraciones es lo que se denomina Análisis de Procrustes Generalizado (APG).

Igual que para el caso de dos configuraciones, sin pérdida de generalidad, se supone que todas las configuraciones están centradas para evitar el problema de la traslación de las configuraciones y se busca una configuración consenso $\mathbf{Z}$ a través de la rotación y reescalamiento de las configuraciones individuales para que sean lo más similares posible a la configuración media. El modelo puede escribirse como:

$$\mathbf{Z}_m = \rho_m \mathbf{Y}_m \mathbf{T}_m = \mathbf{Z} + \mathbf{E}_m \qquad [4.4]$$

donde $\mathbf{Z}_m, \rho_m, \mathbf{Y}_m$ y $\mathbf{T}_m$ son los definidos previamente y $\mathbf{E}_m$ representa el error; todos para m=1,…, M configuraciones.

La Figura 34, muestra en forma gráfica los pasos descritos en el proceso de transformación Procrustes, para cuatro configuraciones (M=4) y tres individuos en cada una (n=3). La Figura 34a representa las configuraciones originales de las matrices

$\mathbf{X}_1,...\mathbf{X}_M$, como fue referido en el contexto de los análisis de diversidad genética se trabajará directamente con $\mathbf{Y}_1,...\mathbf{Y}_M$, las matrices formadas por *k* componentes o coordenadas principales referidos al mismo conjunto de individuos, es así como, el primer paso del análisis se inicia en la Figura 34b, donde se supondrá que todas las configuraciones están centradas y así se evita tener que trasladarlas. Con todas las configuraciones estandarizadas y trasladadas al origen comienza el proceso de rotado y escalado iterativo, el cual consiste en hacer sucesivas rotaciones y escalamientos generando cada vez una nueva configuración de consenso media, así sucesivamente hasta que el cambio en la suma de cuadrados residual entre iteraciones sucesivas es menor que una valor de tolerancia establecido. La Figura 34c muestra cómo la rotación y el escalamiento mueven las configuraciones hasta que los individuos se sitúen lo más cerca posible, formando un nube de puntos de varianza mínima, Figura 34d.

En el contexto de estudios de diversidad genética, el conjunto de transformaciones o movimientos que se han realizado sobre las matrices $\mathbf{Y}_1,...\mathbf{Y}_M$, permite para cada individuo, observar cómo es su variabilidad respecto a los diferentes descriptores o marcadores que se le han observado simultáneamente, conocer si es posible organizar su estructura multidimensional a través de una proyección consenso de los diferentes descriptores o marcadores y determinar con cuanto contribuye cada descriptor o marcador a esa proyección. Estas interrogantes son valoradas a través del estudio de la variabilidad total o de la cuantificación de la importancia relativa del efecto de configuraciones y de individuos.

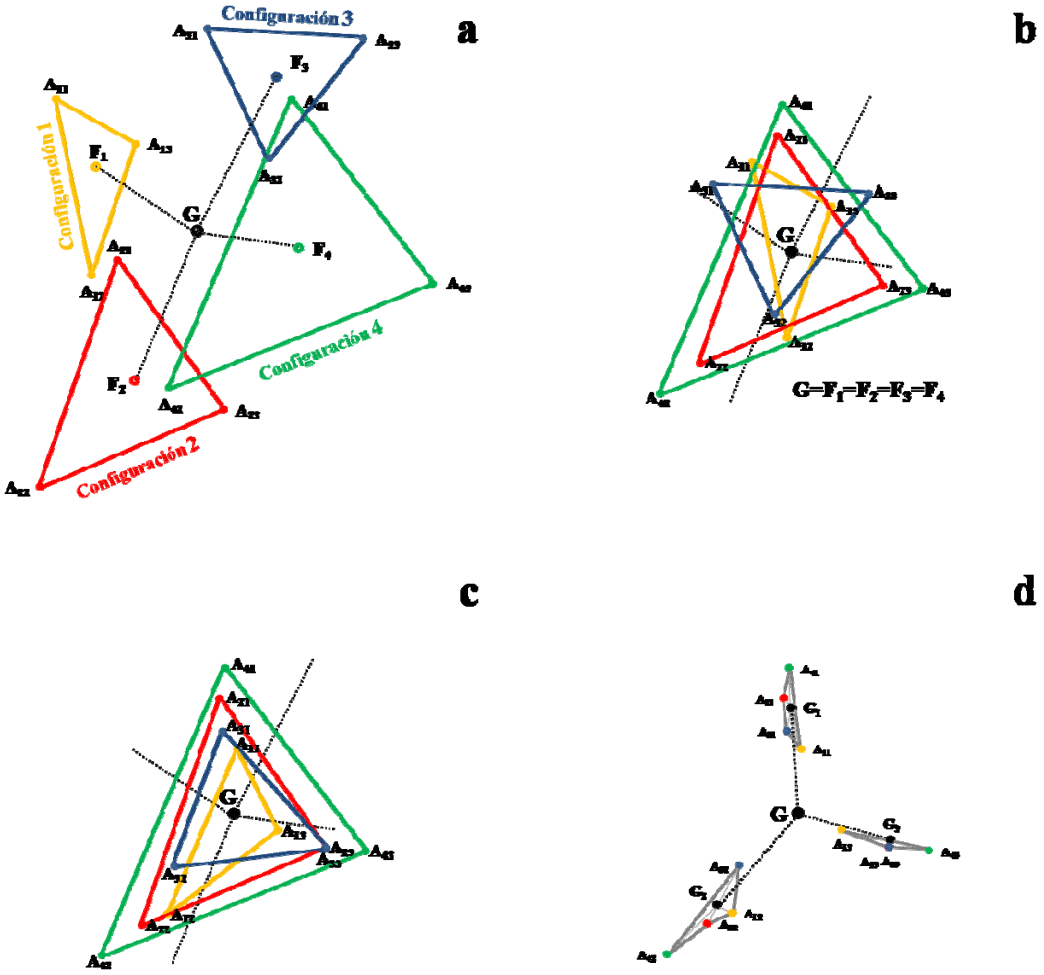


Figura 34. Geometría del Análisis de Procrustes Generalizado (APG).

Se debe entender a la variabilidad total como la suma de cuadrados de las distancias de cada observación respecto al centroide de todas las configuraciones marcado como $\mathbf{G}$ en la Figura 34, este es el equivalente a la suma de cuadros total del Análisis de la Varianza clásico. Gower (1975), propone completar el Análisis de Procrustes Generalizados (APG) realizando un análisis de la varianza, descomponiendo la variabilidad total en una componente debida a la traslación y otra a la orientación que se divide en un término de consenso y uno residual. Obsérvese que bajo el enfoque que se le ha dado a la transformación donde se ha supuesto que las configuraciones están centradas, el efecto de la traslación incluido en el análisis de la varianza propuesto por Gower (1975) no tiene sentido, es así como la variabilidad total puede ser descompuesta como:

$$\mathrm{VTotal} = \mathrm{VConsenso} + \mathrm{VResidual} \qquad [4.5]$$

donde $\mathrm{VTotal} = \sum_{i,m} d^2\left(\mathrm{A}_{im}, \mathbf{G}\right)$, $\mathrm{VConsenso} = \sum_{i} d^2\left(\mathrm{G}_i, \mathbf{G}\right)$ -variabilidad entre individuos-, $\mathrm{VResidual} = \sum_{i,m} d^2\left(\mathrm{A}_{im}, \mathrm{G}_i\right)$ variabilidad dentro de individuos-, $\mathbf{G}$ es el centroide de todas las configuraciones, G_i es el centroide de i-ésimo individuo y A_{im} representa al *i*-ésimo individuo de la *m*-ésima configuración. Geométricamente, las distancias que intervienen en el cálculo de VTotal, $\mathrm{VConsenso}$ y $\mathrm{VResidual}$, son mostradas en las Figuras 35abc, respectivamente.

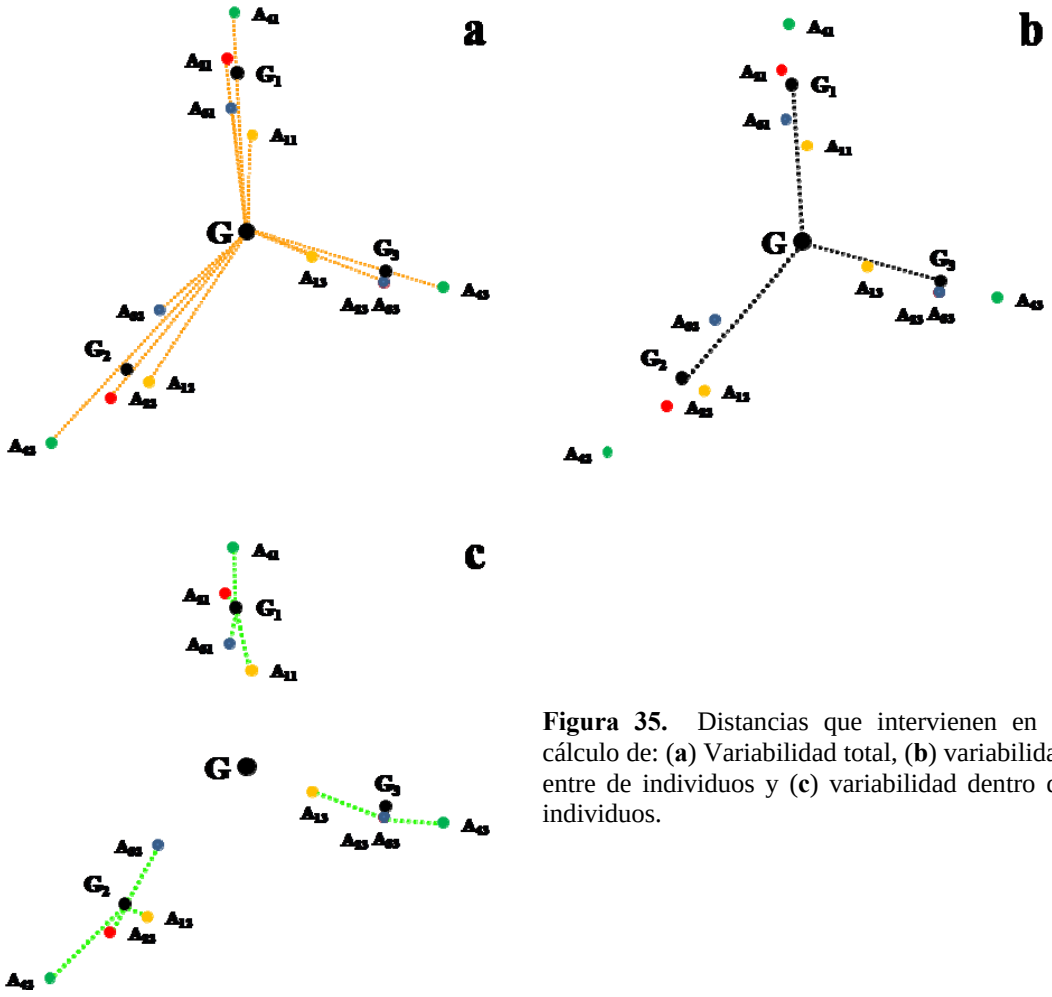


Figura 35. Distancias que intervienen en el cálculo de: (**a**) Variabilidad total, (**b**) variabilidad entre de individuos y (**c**) variabilidad dentro de individuos.

Las coordenadas de los puntos A_{im} de cada una de las configuraciones (m=1,…, *M*) están contenidas en las matrices $\mathbf{Z}_{im}$, es decir, las coordenadas del punto A_{im} están en el vector $\mathbf{z}_{im} = \left(z_{i1(m)},...,z_{ik(m)}\right)'$, $\mathbf{z}_{ij(m)}$ es la coordenada del *i*-ésimo individuo en la *j*-ésima dimensión para la *m*-ésima configuración. Adicionalmente la suma de cuadrados debida a la orientación puede descomponerse por dimensiones o por individuos.

La Tabla 10, muestra las sumas de cuadrados para los efectos considerados en términos de las coordenadas de la matriz de consenso, siendo: $\overline{\mathbf{z}}_i = \frac{1}{M}\sum_m \mathbf{z}_{i(m)} = \left(\overline{z}_{i1},...,\overline{z}_{ik}\right)'$ y $\overline{\mathbf{z}} = \frac{1}{n}\sum_i \overline{\mathbf{z}}_i = \left(\overline{z}_1,...,\overline{z}_k\right)'$.

Podrán calcularse tantas sumas de cuadrados asociadas a la dimensión (segunda columna de Tabla 10) como dimensiones tenga la matriz de consenso; y tantas sumas de cuadrados asociadas a los individuos (tercera columna de Tabla 10) como individuos conformen la matriz de consenso.

Tabla 10. Descomposición de la suma de cuadrados en el Análisis de Procrustes Generalizado (APG)

Fuentes de variación	**Sumas de Cuadrados**		
	Orientación	**Dimensión**	**Individuos**
Consenso	$M\sum_{i=1}^{n}\sum_{j=1}^{k}\left(\overline{z}_{ij}-\overline{z}_{j}\right)^2$	$M\sum_{i=1}^{n}\left(\overline{z}_{ij}-\overline{z}_{j}\right)^2$	$M\sum_{j=1}^{k}\left(\overline{z}_{ij}-\overline{z}_{j}\right)^2$
Residual	$\sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(z_{ij(m)}-\overline{z}_{ij}\right)^2$	$\sum_{m=1}^{M}\sum_{i=1}^{n}\left(z_{ij(m)}-\overline{z}_{ij}\right)^2$	$\sum_{m=1}^{M}\sum_{j=1}^{k}\left(z_{ij(m)}-\overline{z}_{ij}\right)^2$
Total	$\sum_{m=1}^{M}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(z_{ij(m)}-\overline{z}_{j}\right)^2$	$\sum_{m=1}^{M}\sum_{i=1}^{n}\left(z_{ij(m)}-\overline{z}_{j}\right)^2$	$\sum_{m=1}^{M}\sum_{j=1}^{k}\left(z_{ij(m)}-\overline{z}_{j}\right)^2$

La expresión que permite estimar la contribución de las configuraciones a la orientación se denotará como:

$$\text{VConsenso}_{configuración} = \left[\sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(1)} - \overline{z}_j \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(2)} - \overline{z}_j \right)^2 + \ldots + \sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(M)} - \overline{z}_j \right)^2 \right] - \left[\sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(1)} - \overline{z}_{ij} \right)^2 + \sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(2)} - \overline{z}_{ij} \right)^2 + \ldots + \sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(M)} - \overline{z}_{ij} \right)^2 \right] . \quad [4.6]$$

Es posible reordenar la ecuación 4.6 agrupando los términos de cada configuración y encontrar la contribución relativa de cada configuración al consenso dividiendo por la suma de cuadrados de las orientaciones. Por ejemplo, para la configuración m=1 sería:

$$\text{Contribucion}_{(m=1)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(1)} - \overline{z}_j \right)^2 - \sum_{i=1}^{n} \sum_{j=1}^{k} \left(z_{ij(1)} - \overline{z}_{ij} \right)^2}{\text{VConsenso}_{configuración}} .$$

Con algo más de álgebra es posible reagrupar los términos de cada individuo y compararlos con la suma de cuadrados de las orientaciones. Por ejemplo para el individuo n=3 sería:

$$\text{Contribución}_{(n=3)} = \frac{\sum_{m=1}^{M} \sum_{j=1}^{k} \left(z_{3j(m)} - \overline{z}_j \right)^2 - \sum_{m=1}^{M} \sum_{j=1}^{k} \left(z_{3j(m)} - \overline{z}_{3j} \right)^2}{\text{VConsenso}_{configuración}} .$$

En el contexto que nos concierne, la ecuación 4.6 es una medida global de la contribución de los diferentes conjuntos de descriptores a la estructura de grupos de los individuos; mientras que la contribución relativa de un dado conjunto de descriptores permite cuantificar cuánto éstos contribuyen a la estructura de grupos. Los marcadores de un conjunto de baja contribución relativa no podrán interpretarse en el plano de consenso, ya sea porque solo reflejan variabilidad aleatoria o porque se asocian de manera diferente al resto de los marcadores o al conjunto de variables morfológicas. De la misma forma, los individuos cuya contribución relativa sea muy baja estarán mal representados en el consenso porque los conjuntos de marcadores o de variables morfológicas no se expresan en ese individuo de la misma forma que en el resto. Esta información puede también ser aprovechada para investigar la causa de la marginalidad de un individuo.

El consenso entre configuraciones puede entenderse como la contribución relativa de los diferentes descriptores o marcadores. Por esta razón, identificar si esta importancia relativa es significativa, suele ser un indicador importante de la estabilidad de los resultados. Sin embargo, King y Arents (1991) han señalado que el Análisis de Procrustes Generalizado (APG) produce siempre un espacio consenso incluso cuando los datos se generan aleatoriamente, por esta razón diferencias observadas en el consenso pueden deberse al azar y no a que realmente existan. En este sentido se han diseñado pruebas de significación para determinar cuánto de la variación total es debida al consenso (Rc), en otras palabras la prueba tiene como objeto demostrar cuanto del consenso puede ser asociado a un patrón subyacente común. Se han utilizado dos

enfoques en la prueba de significación, el de King y Arents (1991) y el Wakeling *et al.* (1992). El primero genera por simulación muestras aleatorias de distribución uniforme a las cuales se les aplica Análisis de Procrustes Generalizado (APG); si el Rc de los datos originales es mayor que el percentil 95 de la distribución de la muestra, se infiere que el consenso generado es verdadero. Este procedimiento se considera muy conservador porque las muestras aleatorias producen Rc muy altos. El segundo enfoque consiste en hacer permutaciones sobre las filas de las matrices de partida utilizadas para hacer el Análisis de Procrustes Generalizado (APG). Este enfoque tiene la ventaja de preservar la estructura de correlación entre las variables. Para este caso un Rc de los datos menor que el percentil 50 de la distribución empírica resultado de la permutación indica que el consenso no es verdadero sino producto del azar; es decir, la configuración consenso representa menos del 50% de la variabilidad total. Este enfoque tiene la desventaja de que no existe una hipótesis nula implícita para cada prueba pero tiene la ventaja de reproducir la estructura de los datos y por lo tanto ofrecer un estadístico más realista. En el caso que nos ocupa, cuando las configuraciones provienen de un Análisis de Coordenadas Principales y no existe asociación entre las 'variables', se espera que ambos enfoques produzcan resultados coincidentes.

4.2. REPRESENTACIÓN BIPLOT BASADA EN LA ROTACION PROCRUSTES

Como se ha mencionado no es práctica común en las publicaciones que incluyen el Análisis de Procrustes Generalizados (APG) incluir en el análisis la proyección de las variables responsables de la configuración consenso. En este sentido, se considerará para la aproximación Biplot la misma idea formulada en el apartado 2.1.1. La aproximación gráfica de la matriz de datos multivariantes -matriz de datos $\mathbf{X}$ de orden (*n*x*p*)-, que permite estudiar las relaciones entre individuos y variables en una configuración consenso, se hará a través de un modelo bilineal generalizado multiplicativo:

$$g\left(E\left[\mathbf{X}\right]\right) = \mathbf{Z}\boldsymbol{\beta}' + \mathbf{E} \qquad [4.7]$$

que como fue explicado en los apartados anteriores, se entenderá como una regresión generalizada multivariante de $\mathbf{X}$ sobre las coordenadas consenso de los individuos $\mathbf{Z}$, cuando éstas están fijadas. Permitiendo que, se proyecten las variables de la matriz $\mathbf{X}$ sobre la representación gráfica de los n individuos y grupos generada por la matriz de consenso $\mathbf{Z}$ calculada a través del Análisis de Procrustes Generalizados (APG). Se logran así una representación conjunta, donde es posible visualizar las relaciones entre individuos-individuos, individuos-variables y variables-variables. Obsérvese, que las regresiones individuales dependerán de la escala en que fueron medidas las variables originales, pudiéndose entonces representar simultáneamente variables cuantitativas y cualitativas, combinando la función y la distribución de probabilidad asociada.

La proyección de descriptores agromorfológicos y marcadores moleculares conjuntamente permite ver la asociación de aquellos respecto al genoma en estudio. Es en este principio de asociación en el que se basa el mapeo genético de caracteres cuantitativos, una de las herramientas fundamentales de mejoramiento genético (Asíns 2002; Martínez-Gómez *et al.*, 2005; Xu *et al.*, 2005; Céron-Rojas y Sahagún-Castellanos, 2007; Mora *et al.*, 2008). En este sentido la proyección de variables sobre el consenso generado por descriptores agromorfológicos y marcadores moleculares puede considerarse como una herramienta inicial que permite la identificación de asociaciones subyacentes a los datos y la ubicación de caracteres cuantitativos (QTLs) con la ventaja adicional que pueden realizarse con más de un carácter simultáneamente. Además, al ser un enfoque eminentemente descriptivo no es necesario asociar ningún tipo de distribución a los caracteres cuantitativos.

4.3 EJEMPLO ILUSTRATIVO

A continuación se presenta la aplicación del Análisis de Procrustes Generalizados (APG) en la evaluación de la diversidad genética existente en el banco nacional de germoplasma de yuca (*Manihot esculenta* Crantz) del Instituto Nacional de Investigaciones Agrícolas (INIA) de Venezuela.

La yuca (*Manihot esculenta* Crantz) representa una de las fuentes principales de calorías en la dieta de los venezolanos, después del maíz y el arroz. Es un cultivo tropical, producido principalmente por pequeños productores y se usa en forma industrial a

pequeña escala. A nivel mundial ocupa el cuarto lugar como fuente de calorías en la dieta humana, alimentando a más de 500 millones de personas en África, Asia y América Latina (Roa *et al.*, 1997). Las condiciones que presenta la yuca como especie monoica altamente prolífera, con flores de tamaño adecuado para su fácil manipulación tanto en autofecundaciones como en polinización cruzada, con presencia de esterilidad masculina y la posibilidad de propagarse vegetativamente, hacen posible la aplicación de prácticamente todos los métodos de mejoramiento existentes.

La descripción morfológica de órganos vegetativos, reproductivos y rasgos agronómicos clásicos ha sido tradicionalmente utilizada en la caracterización y evaluación de los bancos de germoplasma de yuca. Estas descripciones son limitadas del tiempo necesario para la evaluación completa del cultivo, la cual es necesariamente lenta, dado el gran número de caracteres que hay que observar. Sin embargo, son útiles por ser fácilmente cuantificables e identificables y permitir una discriminación rápida de fenotipos (Lowe *et al.*, 1996).

En los últimos años la utilización de técnicas moleculares ha permitido complementar la información obtenida a través de la caracterización agromorfológica. Dentro de las técnicas de marcadores moleculares más usadas para caracterizar y evaluar la variabilidad genética existente en los bancos de germoplasma de yuca se encuentran la amplificación aleatoria de ADN polimórfico (RAPD) y los microsatélites. Los primeros tiene la gran ventaja que pueden ser utilizados sin conocimiento previo del genoma y los segundos son altamente polimórficos, se encuentran distribuidos al azar a lo largo

del genoma y son abundantes, multialélicos, codominantes, de herencia mendeliana y somáticamente estables (Beeching *et al.*, 1994; Fregene *et al.*, 1994; Lowe *et al.*, 1996; Fregene *et al.*, 1997; Chavarriaga-Aguirre *et al.*, 1999; Zambrano *et al.*, 2003, 2007, Fortes *et al.*, 2008).

Sin embargo, la mayoría de los estudios se hacen en forma independiente y los que incluyen tanto descriptores agromorfológicos como marcadores moleculares, no los estudian en forma conjunta o se limitan, como fue referido, al estudio de la correlación entre matrices de atributos usando las pruebas univariadas clásicas. Demey *et al.* (2003), aunque utilizan un enfoque multivariante, no integran los subespacios generados por los descriptores agromorfológicos y los marcadores moleculares.

En este sentido se evaluó la diversidad genética existente entre treinta clones de yuca a través descriptores agromorfológicos y marcadores moleculares, y se estudio su relación a través del espacio y de las variables responsables del consenso calculadas usando Análisis de Procrustes Generalizados (APG) y los métodos Biplot, respectivamente.

4.3.1 Materiales y métodos

Material Vegetal y Extracción ADN: Se realizó la extracción de ADN genómico total según metodología de Zambrano *et al.* (2002) en tres muestras de tejido foliar joven de las entradas Amacuro 130 (AMA-130), Amacuro 144 (AMA-144), Amacuro 166 (AMA-166), Amacuro 168 (AMA-168), Amazonas 188 (AMAZ-188), Bolívar 043 (BOL-43), Bolívar 050 (BOL-50), Bolívar 059 (BOL-59), Bolívar 089 (BOL-89),

Brasileña 012 (BRA-12), Cogollo Verde (COGO), Juliana Catira (JULCAT), Lengua'e Pájaro (L-PAJA), M-285 (M-285), M-291 (M-291), M-306 (M-306), M-365 (M-365), M-366 (M-366), M-388 (M-388), M-393 (M-393), M-395 (M-395), M-422 (M-422), M-433 (M-433), M-434 (M-434), M-440 (M-440), M-478 (M-478), Meven 150 (MEV-150), Meven 180 (MEV-180), Querepa Blanca (QBLANCO), y Remigio (REMIGIO), las cuales constituyen una muestra representativa del banco nacional de germoplasma de yuca (*Manihot esculenta* Crantz) del Instituto Nacional de Investigaciones Agrícolas (INIA) de Venezuela.

Caracterización molecular: Se utilizaron los iniciadores RAPD: OPA-04, OPA-07, OPB-07, OPK-03, OPK-05, OPM-04, OPM-18 y OPM-20 de Operon Technologies Inc y los SSR: SSRY4, SSRY9, SSRY12, SSRY19, SSRY63, SSRY102, SSRY103 y SSRY110 de IDT Integrated DNA Technologies. Su amplificación fue realizada según metodologías descritas por Zambrano *et al.* (2003) y Chavarriaga-Aguirre *et al.* (1998), para los RAPD y SSR, respectivamente.

Caracterización agromorfológica: Se evaluaron 10 descriptores agromorfológicos recomendados por Fukuda y Guevara (1988). Las mediciones fueron realizadas en el campo en las 30 entradas seleccionadas del banco de germoplasma (*ex situ*), utilizando el promedio de 11 plantas por entrada en un periodo de tres años. Los descriptores utilizados para la caracterización fueron: número de estacas comerciales (NE), largo promedio de raíces en centímetros (LMR), diámetro promedio de raíces en centímetros (DMR), peso promedio de raíz por planta en gramos (PMRP), rendimiento de raíces

comerciales en gramos (RRC) y rendimiento de raíces no comerciales en gramos (RRNC).

Análisis de Datos: En ausencia de análisis de segregación no se hizo ningún supuesto sobre la naturaleza genética de los alelos; en este sentido para cada marcador molecular se generaron sendas matrices binarias producto de la codificación de los productos de amplificación como $A_1A_1 = A_1A_2 = 1$ y $A_2A_2 = 0$ para el caso del marcador dominante, generando una sola columna por locus, para el marcador codominante se consideraron todas las alternativas alélicas A_1A_1 y A_1A_2 generando una columna por alelo. La relación genética entre las 30 entradas fue estudiada usando el Análisis de Coordenadas Principales (ACoP), sobre datos de disimilitud utilizando los coeficientes de Dice y Emparejamiento simple para el caso del marcador dominante y codominante, respectivamente. Para el caso de los descriptores agromorfológicos se utilizó el Análisis de Componentes Principales sobre las 10 variables estandarizadas. Como criterio para la selección del número de dimensiones a ser retenidas, se utilizó el mayor valor *k* de las tres ordenaciones, donde el porcentaje de variación explicada fuese superior al 40%, valor considerado lo suficientemente grande en este tipo de experimentos. Este procedimiento se utiliza a fin de homologar el número de dimensiones a ser comparadas, ya que debido a la naturaleza de cada marcador se necesitarán retener más o menos dimensiones según sea el caso; aunque metodológicamente es correcto completar con ceros aquellas configuraciones que tenga menor dimensión, cuando en los análisis de ordenación no existe una clara separación entre las dos primeras

dimensiones y el resto, esta opción permite explotar mejor la información sobre la variabilidad.

El procedimiento descrito generó dos matrices de coordenadas y una de componentes principales de igual dimensión que fueron utilizadas para explorar las relaciones y representar gráficamente la configuración consenso generada a través de los marcadores moleculares y los descriptores agromorfológicos utilizando Análisis de Procrustes Generalizados (APG). Para probar si la configuración consenso representa más del 50% de la variabilidad total se utilizó el procedimiento debido a Wakeling *et al.* (1992), usando 500 permutaciones. Las variables fueron proyectadas ajustando Biplots a través de regresiones lineales simples o logísticas utilizando como criterio de selección el R^2 o pseudo R^2 dependiendo de la naturaleza de las variables.

En base al objetivo de este apartado, no se detalla la diversidad genética derivada por cada tipo de marcador o descriptor, sino que se muestra cómo debe ser la estrategia para el estudio de sus relaciones; no obstante, debe recordarse que según la metodología se debe realizar primero un estudio detallado por marcador, siguiendo las propuestas que se han desarrollado a lo largo del manuscrito.

4.3.2 Resultados

Los iniciadores aleatorios RAPD produjeron un total de 65 fragmentos de amplificación polimórficos con un tamaño entre 220-1700pb y los SSR a su vez produjeron 16 fragmentos polimórficos con tamaños entre 168-285pb. En los RAPD el menor número de fragmentos polimórficos fue generado con el iniciador OPA-04 con un total de cuatro y el mayor número fue amplificado con el iniciador OPM-18 con un total de diez. Todos los SSR amplificaron dos fragmentos por iniciador. Respecto a los descriptores agromorfológicos se obtuvieron valores medios de 10.93±0.75, 28.70±1.23, 5.53±0.17, 1183.00±118.44, 899.20±103.45 y 363.00±97.50, para el número de estacas comerciales (NE), el largo promedio de raíces (LMR), el diámetro promedio de raíces (DMR), el peso promedio de raíz por planta (PMRP), el rendimiento de raíces comerciales (RRC) y el rendimiento de raíces no comerciales (RRNC), respectivamente.

La Figura 36abc, muestra el espacio bidimensional obtenido del Análisis de Coordenadas Principales (ACoP) para los marcadores RAPD y SSR y el Análisis de Componentes Principales (ACP) para los descriptores agromorfológicos. Las dos primeras dimensiones explican el 25.70%, 35.84% y 74.86% de la variabilidad total y permiten la formación de cuatro y tres grupos de entradas utilizando los iniciadores RAPD y SSR y descriptores agromorfológicos, respectivamente. El mayor valor *k* de las tres ordenaciones, donde el porcentaje de variación explicada es superior al 40% es igual a cuatro (*k*=4) y se extrae de la ordenación debida a los iniciadores RAPD.

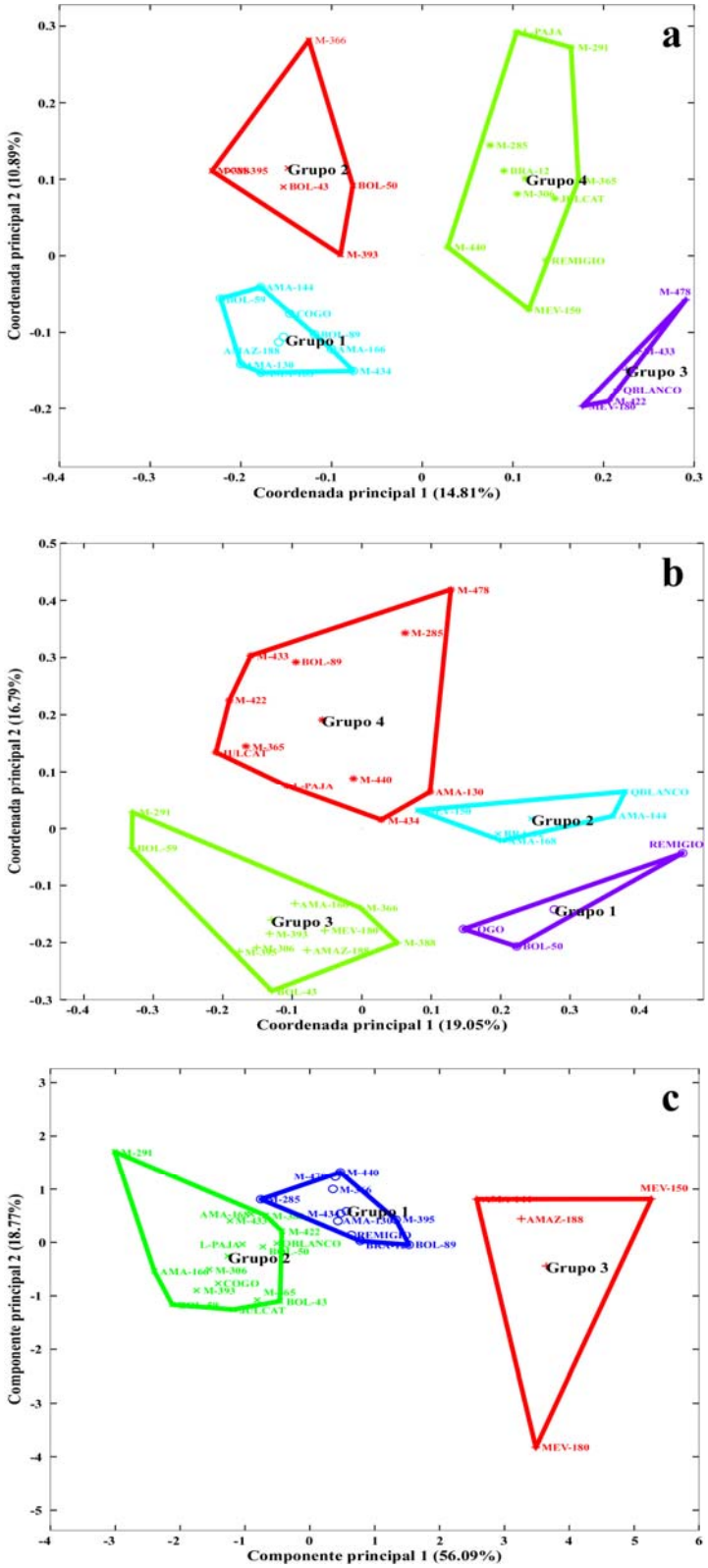


Figura 36. Relaciones genéticas entre las diferentes entradas de yuca: (**a**) Iniciadores RAPD; (**b**) Iniciadores SSR y (**c**) Descriptores agromorfológicos.

La inercia absorbida por el plano principal, las escalas y direcciones de los ejes, así como las relaciones entre las diferentes entradas muestran que las configuraciones por los marcadores moleculares y los descriptores agromorfológicos son diferentes inhabilitando su representación en un mismo sistema de coordenadas. La Tabla 11 y la Figura 37, revelan la independencia entre las clasificaciones generadas, esto se debe a que los marcadores y descriptores utilizados extraen información diferente que altera la proyección de las entradas en el plano principal. Nótese además, que no es posible determinar cuál de las configuraciones ofrece los mejores resultados, aunque algunas detecten mayor variabilidad que otras. Esto sucede ya que se desconoce *a priori* la estructura de grupos existente, como es el caso en la mayoría de los estudios de diversidad genética.

Tabla 11. Distribución de las entradas para las diferentes configuraciones

Marcador o descriptor	Grupo	Entradas	Similitud genética[1]	Calidad de representación[2]
RAPD	1	AMA-130, AMA-144, AMA-166, AMA-168, AMAZ-188, BOL-59, BOL-89, COGO, M-434	0.7067±0.0094	92.18
	2	BOL-43, BOL-50, M-366, M-388, M-393, M-395	0.7087±0.0185	72.19
	3	M-422, M-433, M-478, MEV-180, QBLANCO	0.7101±0.0153	91.74
	4	BRA-12, JULCAT, L-PAJA, M-285, M-291, M-306, M-365, M-440, MEV-150, REMIGIO	0.6979±0.0084	86.09
SSR	1	BOL-50, COGO, REMIGIO	0.7500±0.0625	66.44
	2	AMA-144, AMA-168, BRA-12, MEV-150, QBLANCO	0.7750±0.0212	58.21
	3	AMA-166, AMAZ-188, BOL-43, BOL-59, M-291, M-306, M-366, M-388, M-393, M-395, MEV-180	0.7318±0.0124	99.02
	4	AMA-130, BOL-89, JULCAT, L-PAJA, M-285, M-365, M-422, M-433, M-434, M-440, M-478	0.6023±0.0171	93.09
AGRO MORFOLOGICOS	1	AMA-130, BOL-89, BRA-12, M-285, M-366, M-395, M-434, M-440, M-478, REMIGIO	0.5200±0.0428	96.11
	2	AMA-166, AMA-168, BOL-43, BOL-50, BOL-59, COGO, JULCAT, L-PAJA, M-291, M-306, M-365, M-388, M-393, M-422, M-433, QBLANCO	0.4338±0.0235	99.14
	3	AMA-144, AMAZ-188, MEV-150, MEV-180	0.4107±0.1106	99.99

1. Similitud genética entre entradas
2. Calculada con las *k*-dimensiones retenidas

Adicionalmente, la Figura 37, muestra cómo el uso del coeficiente de correlación entre las matrices de distancias puede llevar a conclusiones inexactas, ya que no solo es afectado por el tamaño de las muestras a comparar, sino que además supone que las configuraciones pertenecen al mismo sistema de referencia. Es así, que con un coeficiente de correlación (***r***=0.1461) se detecta asociación entre la configuración generada por los iniciadores RAPD y los descriptores agromorfológicos, a pesar de que el diagrama de dispersión muestre que el comportamiento es diferente.

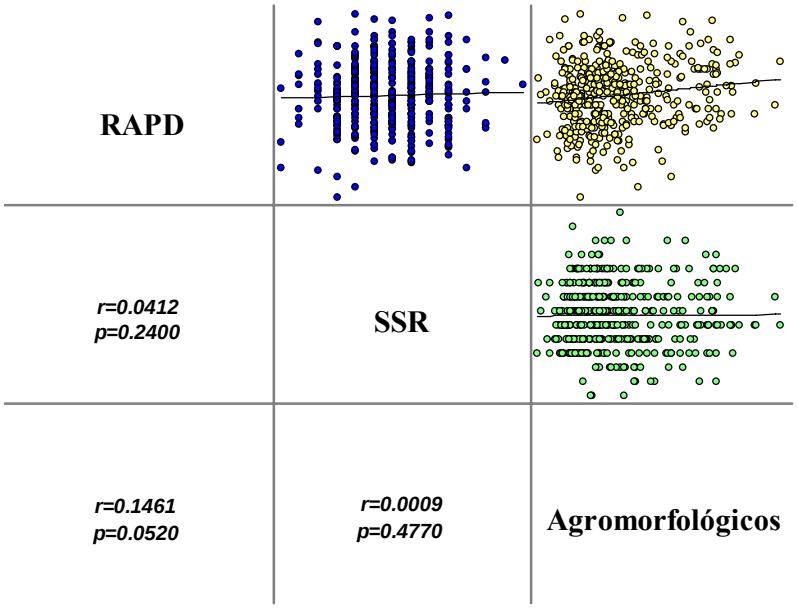


Figura 37. Matriz de dispersión y correlaciones entre las configuraciones.

Los resultados de las tres caracterizaciones y del estudio de sus relaciones indican que los marcadores moleculares y los descriptores morfológicos, ofrecen información que puede ser considerada complementaria ya que no se origina un patrón único de asociación entre las entradas, corroborando la importancia que tiene su estudio conjunto

para obtener una mejor descripción e interpretación de la diversidad genética de los individuos. En este sentido la Figura 38a, presenta la representación bidimensional de la configuración consenso generada de la aplicación del Análisis de Procrustes Generalizado (APG) sobre las configuraciones $\mathbf{Y}_{RAPD}, \mathbf{Y}_{SSR}, \mathbf{Y}_{AGROM}$ para las cuatro dimensiones retenidas. El plano principal representa el 57.10% de la variabilidad y permite la formación de cuatro grupos. El primer grupo formado por las entradas M-478, QBLANCO y REMIGIO, el segundo grupo formado por las entradas JULCAT, L-PAJA, M-285, M-291, M-365, M-422, M-433 y M-440; el tercer grupo formado por los cultivares AMA-166, BOL-43, BOL-50, BOL-59, COGO, M-306, M-393 y M-395 y el cuarto grupo formado por las entradas AMA-130, AMA-144, AMA-168, AMAZ-188, BOL-89, BRA-12, M-366, M-388, M-434, MEV-150, MEV-180, con una similaridad genética media de 0.7922±0.1002, 0.5151±0.0664, 0.5935±0.0479, 0.4825±0.0406**,** y una calidad de representación calculada con las *k* dimensiones retenidas de 92.77%, 98.06%, 99.68%, 99.89%, para el primero, segundo, tercero y cuarto grupo, respectivamente y una varianza explicada por el consenso de 54.95%.

Así mismo, el Análisis de Procrustes Generalizado (APG) permite concluir que la contribución relativa de las configuraciones $\mathbf{Y}_{RAPD}, \mathbf{Y}_{SSR}, \mathbf{Y}_{AGROM}$ al consenso es similar y todas las configuraciones son explicadas en más del 50% por el consenso, en valores de 52.10%, 53.90% y 56.98% para $\mathbf{Y}_{RAPD}, \mathbf{Y}_{SSR}, \mathbf{Y}_{AGROM}$, respectivamente. Respecto a las cuatro dimensiones valoradas contribuyen al consenso en el orden de 36.62%, 24.27%, 19.46% y 19.65%.

Nótese que aunque las trayectorias que definen la variabilidad de cada individuo respecto al consenso, son irregulares en magnitud y en muchos de los casos denotan dispersión considerable, Figura 38b, el porcentaje de variabilidad media de cada individuo explicado por el consenso fue superior al 51%.

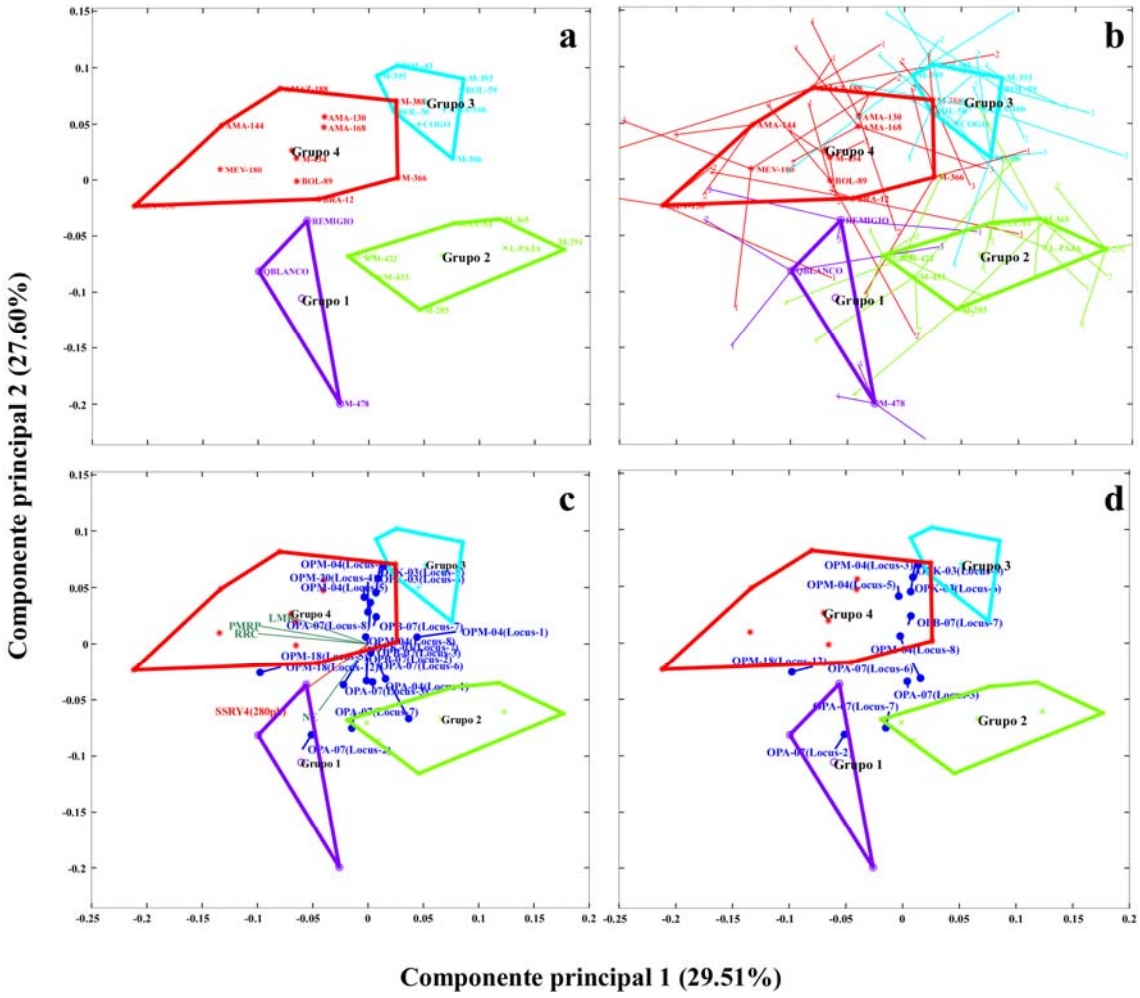


Figura 38. Relaciones genéticas entre las diferentes entradas de yuca: (**a**) En la configuración consenso; (**b**) En la configuración consenso mostrando la variabilidad de los individuos; (**c, d**) En la configuración consenso mostrando las variables que fueron proyectadas ajustando Biplots a través de regresiones lineales simples o logísticas utilizando como criterio de selección un R^2 o pseudo R^2 mayor o igual a 0.60 y 0.75, respectivamente.

Las Figuras 38cd muestran las variables que fueron proyectadas ajustando Biplots a través de regresiones lineales simples o logísticas utilizando como criterio de selección un R^2 o pseudo R^2 mayor o igual a 0.60 y 0.75, respectivamente. Nótese que en el primer caso, menos restrictivo, las variables asociadas al rendimiento como: el largo promedio de raíces (LMR), el peso promedio de raíz por planta (PMRP) y el rendimiento de raíces comerciales (RRC), aparecen representadas y están asociadas a la formación de dos grandes grupos: los grupos 2 y 3 donde se esperaría encontrar entradas con rendimientos bajos y los grupos 1 y 4 con más alto rendimiento. El número de estacas comerciales (NE) contribuye a la separación del grupo 1 de los demás. Un solo alelo asociado a los iniciadores SSR permite separar el grupo 1 del resto, respecto al comportamiento de los indicadores RAPD se observa que pueden intervenir en la formación de diferentes combinaciones de individuos. Cuando se restringe la proyección de variables aumentando el valor del R^2 o el pseudo R^2, se observa que solo las variables asociadas a los iniciadores RAPD intervienen en la definición de grupos de entradas.

Estos resultados aparentemente podrían contradecir el grado de contribución que tienen las configuraciones individuales al consenso, puesto que se esperaría que existiese un número proporcional de variables por cada configuración que definieran grupos homogéneos. Esto puede ser atribuido a dos causas, a la magnitud del consenso encontrado y al tipo de marcadores utilizados. Respecto a la magnitud del consenso, este representa más del 50% de la variabilidad total y puede considerarse que es verdadero y no producto del azar, como se demuestra en la distribución de los

consensos generados a través de la permutación de las configuraciones utilizadas para hacer el Análisis de Procrustes Generalizado, Figura 39. Destaca que el consenso observado (línea verde), supera al consenso teórico (línea roja) y se encuentra dentro de los límites del intervalo de confianza estimado para las 500 muestras valoradas.

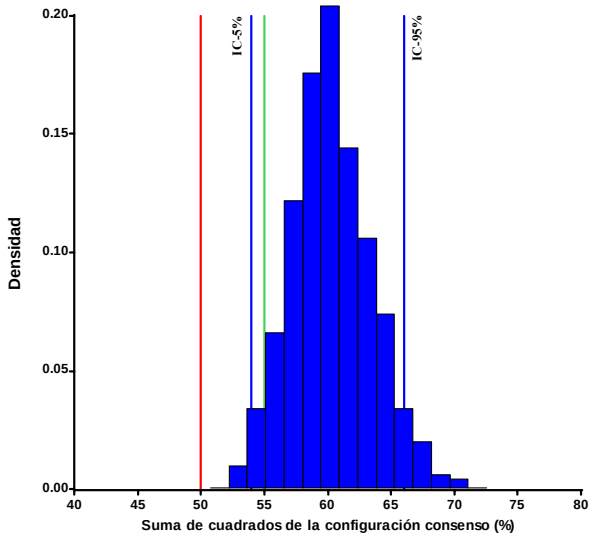


Figura 39. Distribución de la variabilidad debida al consenso generada a través del procedimiento del Wakeling *et al.* (1992), usando 500 permutaciones.

En cuanto a la influencia del tipo de marcadores, la proyección de una mayor cantidad de variables asociadas a los iniciadores RAPD, en la configuración consenso, es atribuible a que como marcadores aleatorios cubren una base del genoma más amplia, al contrario de los marcadores SSR que al ser específicos evalúan pequeños segmentos y también a que el análisis se realiza desconociendo los mecanismos de segregación, disminuyendo la ventaja de codominancia de estos marcadores. Sin embargo, en

estudios amplios de diversidad, donde se desconoce la estructura genética de la población estos resultados están ajustados a la realidad.

La yuca es un cultivo altamente complejo porque no tiene un solo mecanismo de reproducción, lo que hace que muchas de las entradas distinguidas en el banco de germoplasma, sean producto de variaciones somaclonales seleccionadas por pequeñas alteraciones que afectan rasgos agronómicos favorables, que se traducen en pequeñas diferencias a nivel molecular. La metodología propuesta permite que, sobre cada individuo, se observe cómo es su variabilidad respecto a los diferentes descriptores o marcadores que se le han observado simultáneamente y organizar su estructura multidimensional en el sentido de conocer con cuánto contribuye cada descriptor o marcador a su relación con el resto.

Conclusiones

CONCLUSIONES

El trabajo representa una contribución a la compresión de los estudios de diversidad genética en bancos de germoplasma. Bajo un enfoque heurístico construye una estrategia metodológica tratada desde la revisión del problema, la descripción de las propiedades de los datos, la simulación de escenarios probables y la presentación de ejemplos reales, hasta ofrecer un algoritmo de solución que permite responder a las preguntas clave de los mejoradores genéticos y que hasta ahora no tienen respuesta mediante los análisis clásicos.

Bajo esta premisa se concluye:

1. La mejora de las técnicas genéticas ha aumentado la cantidad y la calidad de los atributos que se utilizan para evaluar la diversidad genética de los bancos de germoplasma. Sin embargo, en la bibliografía especializada se observa un atavismo a las mismas estrategias de análisis, bien sea por el desconocimiento de nuevos métodos que han sido desarrollados para las particularidades de los marcadores y/o a la proliferación de paquetes informáticos que han facilitado la aplicación de una determinada metodología sin ningún espíritu crítico.

2. La clasificación de genotipos tiene dos aspectos importantes: la técnica en sí misma y la ayuda que ofrece a los investigadores en la interpretación de los resultados. El énfasis en la primera aproximación requiere la elección de métodos apropiados para

determinar relaciones entre los individuos; mientras que usando la segunda aproximación se debe seleccionar la técnica que permita que se estudien apropiadamente las relaciones individuo-individuo, individuo-marcador y marcador-marcador. Ambos enfoques no son excluyentes.

3. Se demuestra que las ventajas que el Análisis de Conglomerados (CA) –método unánimemente usado en los estudios de diversidad genética- ofrece a la taxonomía tradicional y a la biología evolutiva con respecto a la calidad de clasificaciones y de su interpretación, varían en el contexto de la clasificación de genotipos; especialmente cuando se basa en marcadores moleculares, ya que esta técnica de clasificación no ofrece información de las relaciones entre individuo-marcador y marcador-marcador.

4. Se reivindica la aplicación del Análisis de Coordenadas Principales (ACoP) en estudios de diversidad genética y se propone como procedimiento inicial de un algoritmo de análisis que, empleado conjuntamente con los Métodos Biplot, permite un estudio apropiado de la diversidad genética estableciendo las relaciones entre individuo- individuo, individuo-marcador y marcador-marcador. Ventaja importante ya que las dos últimas relaciones no han recibido la suficiente atención en la literatura especializada en estudios de diversidad genética.

En este sentido se demuestra que:

5. Para datos provenientes de la cuantificación de productos de ampliación de ADN, la capacidad de recuperación de información de los algoritmos de agrupamiento cuando se utilizan las coordenadas principales de las dos primeras dimensiones, es igual o superior a la obtenida utilizando las matrices binarias originales directamente, independientemente del nivel de ruido y del coeficiente de similitud. La utilización de las coordenadas principales para la generación de grupos incrementa los porcentajes de clasificación correcta a medida que el ruido incrementa; así mismo, la reducción de la dimensionalidad y la representación de los individuos en el plano bidimensional no están afectadas, incluso aún cuando se obtengan valores bajos de absorción de varianza.

6. Con los algoritmos propuestos es posible estudiar la variabilidad muestral de los individuos en el plano bidimensional. Se concluye que la estabilidad de las configuraciones está más afectada por la estrategia de alteración de los individuos y el método de corrección que por el coeficiente de similitud y las dimensiones que se retengan.

7. A través del ajuste de Biplots de Regresión o Biplot de Predicción es posible proyectar las variables de la matriz **X** sobre la representación gráfica de los individuos o grupos de individuos, generada vía Análisis de Coordenadas Principales (ACoP), logrando así una representación conjunta. En esta

representación conjunta es posible visualizar individuos y variables de forma simultánea en un espacio de dimensión reducida, con la menor pérdida de información y con la ventaja adicional de poder utilizar cualquiera de la medidas de similitud/disimilitud, dependiendo de la naturaleza de las variables.

8. La representación gráfica producto de la combinación de las tres técnicas, Análisis de Coordenadas Principales (ACoP), Análisis de Conglomerados (AC) y los métodos Biplot, favorece el estudio simultáneo de las relaciones entre individuos, individuos-variables y variables-variables, incrementando la cantidad y la calidad de la información en relación a la suministrada por los métodos unánimemente utilizados para los estudios de diversidad genética.

9. Para matrices binarias generadas de la cuantificación de productos de ampliación de ADN, el ajuste de un Biplot Logístico Externo (BLE) sobre las coordenadas principales permite identificar los alelos de mayor importancia en la definición de la estructura natural de los individuos, en las primeras coordenadas principales.

10. Comparado con el enfoque clásico, el uso combinado del Análisis de Coordenadas Principales (ACoP), Análisis de Conglomerados (AC) y el ajuste de un Biplot Logístico Externo (BLE) ofrece una comprensión holística de la estructura de datos, facilita las interpretaciones de los resultados y puede ser altamente recomendado para una descripción cuidadosa de datos en estudios de la diversidad genética usando marcadores de ADN.

CONCLUSIONES

Respecto a la organización de la estructura multidimensional de los individuos a los cuales se le han observado simultáneamente diferentes tipos de marcadores se demuestra que:

11. Como alternativa al enfoque clásico de aplicar matrices de correlación para medir las relaciones entre los diferentes tipos de marcadores, es posible el estudio conjunto sobre un espacio de consenso a través del Análisis de Procrustes Generalizado (APG). Sobre este espacio común es posible determinar la contribución relativa de cada descriptor o marcador a la formación del consenso y comprobar si la contribución es significativa, a través del estudio de la estabilidad de los resultados.

12. En la configuración consenso proveniente de las matrices de las diferentes estrategias de ordenaciones para cada descriptor o marcador generada través del Análisis de Procrustes Generalizado (APG), es posible proyectar las variables responsables del consenso a través del ajuste de Biplots de Regresión. Se pueden entonces representar simultáneamente variables cuantitativas y cualitativas, combinando la función y la distribución de probabilidad asociada de cada una. Este enfoque puede ser muy útil para su aplicación, entre otros, a la localización de QTLs independientemente de la distribución del carácter.

Bibliografía

BIBLIOGRAFIA

Abascal-Fernández, E.; Landaluce-Calvo, M.I. (2002). Análisis factorial múltiple como técnica de estudio de la estabilidad de los resultados de un análisis de componentes principales. *Questiió,* 26:109-122.

Alwala, S.; Suman, A.; Arro, J.A.; Veremis, J.C.; Kimbeng, C.A. (2006). Target Region Amplification Polymorphism (TRAP) for assessing genetic diversity in sugarcane germplasm collections. *Crop Science,* 46:448-455.

Anderson, T.W. (1963). Asymptotic theory for principal components analysis. *Annals of Mathematical Statistics,* 34:122-148.

Asíns, M.J. (2002). Present and future of quantitative trait locus analysis in plant breeding. *Plant Breeding,* 121**(4):**281-291.

Avise, J.C. (2004). *Molecular Markers, Natural History and Evolution.* Sinauer Associates. 2nd edition. Sunderland. USA. 684 p.

Badea, A.; Eudes, F.; Graf, R.J.; Laroche, A.; Gaudet, D.A.; Sadasivaiah, R.S. (2008). Phenotypic and marker-assisted evaluation of spring and winter wheat germplasm for resistance to fusarium head blight. *Euphytica,* 164:803-819.

Barkley, N.A.; Roose, M.L.; Krueger, R.R.; Federici, C.T. (2006). Assessing genetic diversity and population structure in a citrus germplasm collection utilizing simple sequence repeat markers (SSRs). *Theoretical and Applied Genetics,* 112**(8):**1519-1531.

BIBLIOGRAFIA

Barry, M.B.; Pham, J.L.; Noyer, J.L.; Billot, C.; Courtois, B.; Ahmadi, N. (2007). Genetic diversity of the two cultivated rice species (*O. sativa* and *O. glaberrima*) in Maritime Guinea. Evidence for interspecific recombination. *Euphytica,* 154:127-137.

Bartlett, M.S. (1950). Test of significance in factor analysis**.** *British Journal Psychology (Statistic section),* 3:77-85.

Beeching, J.R.; Marmey, P.; Hughes, M.A.; Charrier, A. (1994). *Evaluation of molecular approaches for determining genetic diversity in Cassava germplasm.* In: Proceedings of the second international Scientific Meeting. The Cassava Biotechnology Network. Bogor. Indonesia. pp 22-26.

Benzecri, J.P. (1970). Distance distributionnelle et metrique chi-deux en analyse factorielle des correspondances. Paris. Laboratoire de Statistique Mathématique.

Besse, P.; McIntrey, C.L.; Burner, D.; Dealmeida, C.G. (1997). Using genomic slot hybridization to assess intergeneric Saccharum x Erianthus hybrids (Andropogoneae-Saccharinae). *Genome,* 40:428-432.

Besse, P.; Taylor, G.; Carrol, B.; Berbing, N.; Burner, D.; McIntrye, C.L. (1998). Assessing genetic diversity in a sugarcane germplasm collection using an automated AFLP analysis. *Genetica,* 104:143-153.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhya,* 7:401-406.

Blazquez, A. (1998). *Análisis Biplot basado en modelos lineales generalizados.* Tesis Doctoral. Universidad de Salamanca. España. 240 p.

BIBLIOGRAFIA

Boontong, C.; Pandey, M.; Changtragoon, S. (2008). Isolation and characterization of microsatellite markers in Indian neem (*Azadirachta indica* var. indica A. Juss) and cross-amplification in Thai neem (*A. indica* var. siamensis Valenton). *Conservation Genetics,* DOI 10.1007/s10592-008-9610-5.

Bowcock, A.M.; Ruíz-Linares, A.; Tomfohrde, J.; Minch, E.; Kidd, J.R.; Cavalli-Sforza, L.L. (1994). High resolution human evolutionary trees with polymorphic microsatellites. *Nature,* 368:455-457.

Bramardi, S.J.; Bernet, G.P.; Asíns, M.J.; Carbonell, E.A. (2005). Simultaneous Agronomic and Molecular Characterization of Genotypes via the Generalised Procrustes Analysis. An Application to Cucumber. *Crop Science,* 45:1603-1609.

Bray, J.R.; Curtis, J.T. (1957). An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs,* 27:325-349.

Bretting, P.K.; Widrlechner, M.P. (1995). Genetic markers and horticultural germplasm management. *HortScience,*. 30**(7):**1349-1356.

Brito, G.; Loureiro, J.; Lopes, T.; Rodriguez, E.; Santos, C. (2008). Genetic characterisation of olive trees from Madeira Archipelago using flow cytometry and microsatellite markers. *Genetic Resources and Crop Evolution,* 55**(5):**657-664.

Brouat, C.; MacKey, D.; Douzery, E.J.P. (2004). Differentiation in a geographical mosaic of plants coevolving with ants: phylogeny of the *Leonardoxa africana* complex (Fabaceae: Caesalpinioideae) using amplified fragment length polymorphism markers. *Molecular Ecology,* 13**(5):**1157-1171.

BIBLIOGRAFIA

Burner, D.M.; Pan, Y.B.; Webster, R.D. (1997). Genetic diversity of North American and Old World *Saccharum* assessed by RAPD analysis. *Genetic Resources and Crop Evolution*, 44:235-240.

Butterfield, M.K.; D'Hont, A.; Berding, N. (2001). The sugarcane genome: a synthesis of current understanding, and lessons for breeding and biotechnology. *Proceeding International Society African Sugarcane Technologists Association*, 75:1-5.

Cain, A.J.; Harrison, G.A. (1958). An analysis of the taxonomists' judgment of affinity. *Proceedings of the Zoological Society of London*, 131:85-98.

Cárdenas, O.; Noguera, C.; Galindo, P.; Vicente-Villardón, J.L. (2006). Alternativa a la regresión con componentes principales basada en biplots de regresión. *Interciencia*, 31**(3):**160-167.

Cattell, R.B. (1966). *The meaning and strategic use of factor analysis.* In: Handbook of multivariate Experimental Psychology. Cattell RB (Ed). Rand McNally. Chicago. 241 p.

Cavalli-Sforza, L.L.; Edwards, A.W.F. (1967). Phylogenetic Analysis: Models and estimation procedures. *Evolution*, 32:550-570.

Céron-Rojas, J.J.; Sahagún-Castellanos, J. (2007). Estimating QTL biometrics parameters in F2 populations: a new approach. *Agrociencia*, 41:57-73.

Chae, S.S.; Warde, W.D. (2006). Effect of using principal coordinates and principal components on retrieval of clusters. *Computational Statistics and Data Analysis*, 50:1407-1417.

Chakraborty, R.; Rao, C.R. (1991). *Measurement of genetic variation for evolutionary studies.* Handbook of Statistics. Vol. 8. Elsevier. Amsterdam. pp. 271-316.

BIBLIOGRAFIA

Chavarriaga-Aguirre, P.; Maya, M.; Bonierbale, M.; Kresovich, S.; Fregene, M.; Tohme, J.; Kochert, G. (1998). Microsatellites in Cassava (*Manihot esculenta* Crantz): discovery, inheritance and variability. *Theoretical and Applied Genetics*, 97: 493-501.

Chavarriaga-Aguirre, P.; Maya, M.; Tohme, J.; Duque, M.; Iglesias, C.; Bonierbale, M.; Kresovich, S.; Kochert, G. (1999). Using microsatellites, isozymes and AFLPs to evaluate genetic diversity and redundancy in cassava core collection and to assess the usefulness of DNA-based markers to maintain germplasm collections. *Molecular Breeding*, 5:263-273.

Cordeiro, G.M.; Taylor, G.O.; Henry, R.J. (2000). Characterization of microsatellite markers from sugarcane (*Saccharum* sp.), a highly polyploid species. *Plant Science*, 155:161-168.

Cordeiro, G.M.; Pan, Y.B.; Henry, R.J. (2003). Sugarcane microsatellites for the assessment of genetic diversity in sugarcane germplasm. *Plant Science*, 165**(1):**181-189.

Cordeiro, G.M.; Eliott, F.; McIntyre, C.; Casu, R.; Henry, R.J. (2006). Characterisation of single nucleotide polymorphisms in sugarcane ESTs. *Theoretical and Applied of Genetics*, 113**(2):**331-343.

Cuadras, C.M. (1996). *Métodos de análisis multivariante*. EUB (Ed), SL. Barcelona. 642 p.

Cuadras, C.M.; Arenas, C.A. (1990). A distance based model for prediction with mixed data. Communications in Statistics. *Theory Methods*, 19:2261-2279.

Cuadras, C.M.; Fortiana, J. (1995). A continuous metric scaling solution for a random variable. *Journal of Multivariate Analysis*, 52:1-14.

BIBLIOGRAFIA

Da Silva, J.A.G. (2001). Preliminary analysis of microsatellites markers derived from sugarcane expressed sequence tags (ESTs). *Genetics and Molecular Biology*, 24 **(1-4):**155-159.

Demey, J.R.; Zambrano, A.Y.; Fuenmayor, F.; Segovia, V. (2003). Relación entre caracterizaciones molecular y morfológica en una colección de Yuca. *Interciencia*, 28**(12):**1-7.

Demey, J.R.; Vicente-Villardon, J.L.; Galindo, M.P.; Zambrano, A.Y. (2008). Identifying molecular markers associated with classification of genotypes by External Logistic Biplots. *Bioinformatics*, DOI: 10.1093/bioinformatics/btn552.

D'Hont, A.; Glaszmann, J.C. (2001). Sugarcane genome analysis with molecular markers, a first decade of research. *Proceedings International Society of Sugarcane Technologist*, 24:556-559.

Dice, L.R. (1945). Measures of the amount of ecologic assocaition between species. *Ecology*, 26: 297-302.

Digby, P.G.N.; Kempton, R.A. (1991). *Multivariate analysis of ecological communities.* Chapman and Hall. London. England. 206 p.

Dreisigacker, S.; Zhang, P.; Warburton, M.L.; Skovmand, B.; Hoisington, D.; Melchinger, A.E. 2005. Genetic Diversity among and within CIMMYT Wheat Landrace Accessions Investigated with SSRs and Implications for Plant Genetic Resources Management. *Crop Science*, 45:653-661.

Eckart, V.; Young, G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika*, 1**(3):**211-218.

Efron, B.; Tibshirani, RJ. (1993). *An introduction to the bootstrap.* New York: Chapman and Hall. 436p.

BIBLIOGRAFIA

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. *Les cahiers de l'analyse des donées,* 4:137-146.

Esposito, M.A.; Martin, E.A.; Cravero, V.P.; Cointry, E. (2007). Characterization of pea accessions by SRAP's markers. *Scientia Horticulturae,* 113**(4):**329-335.

Everitt, B.S. (1979). Unresolved problems in cluster analysis. *Biometrics,* 35:169-181.

Everitt, B.S.; Landau, S.; Leese, M. (2001). *Cluster analysis.* Hodder Arnold Publication. 4th edition. Oxford University Press. Oxford. UK. 248 p.

Faccioli, P.; Terzi, V.; Monetti, A.; Nicola, J.; Pecchioni, N. (1995). B-hordein STS markers for barley genotype identification : comparison with RFLPs, hordein A-PAGE and morpho-physiological traits. *Seed Science Technology,* 23:415-427.

Falconer, D.S.; Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics.* Ed 4. Longmans Green. Harlow. Essex. UK. pp 122-324.

Falguerolles, A de. (1998). *Log-bilinear biplots in action.* In: Visualization of Categorical Data. Blasius, J.; Greenacre, M. (Eds). Academic Press. San Diego. USA. 594 p.

Felsenstein, J. (1985). Phylogenies from gene frequencies: A statistical problem. *Systematic Zoology,* 34:300-311.

Felsenstein, J. (1991). *PHYLIP (Phylogeny inference package)* v.3×4. University of Washington. Seattle. USA.

Felsenstein, J. (2004). *Inferring Phylogenies.* Sinauer associates. Sunderland. 664 p.

Flury, B.D. (1984). Common Principle Components in K groups. *Journal of the American Statistical Associations,* 79. 892-898.

Flury, B.D. (1988). *Common principal components and related multivariate models.* John Wiley and Sons, New York. 258 p.

BIBLIOGRAFIA

Fortes, C.; Alves, E.; Nogueira, K.; Jungahans, D.; Kenji, A.; Santos, V.; Pereira, R.; Henrique, P.; Soares, E.; Fukuda, W. (2008). Molecular characterization of Cassava (*Manihot esculenta* Crantz) with yellow-orange roots for beta-carotene. *Crop Breeding and Applied Biotechnology*, 8:23-29.

Franco, J.; Crossa, J.; Ribaut, J.M.; Betran, J.; Warburton, M.L.; Khairallah, M. (2001). A method for combining molecular markers and phenotypic attributes for classifying plant genotypes. *Theoretical Applied of Genetics*, 103:944-952.

Franco, J.; Crossa, J.; Warburton, M.L.; Taba, S. (2006). Sampling Strategies for Conserving Maize Diversity When Forming Core Subsets Using Genetic Markers. *Crop Science*, 46:854-864.

Fregene, M.; Vargas, J.; Ikea, J.; Angel, F.; Tohme, J.; Asiedu, R.A.; Akoroda, M.O.; Roca, W.M. (1994). Variability of chloroplast DNA and nuclear ribosomal DNA in cassava (Manihot esculenta Crantz). *Theoretical Applied Genetics*, 89:719-727.

Fregene, M.; Angel, F.; Gomez, R.; Rodríguez, F.; Chavarriaga, P.; Roca, W.; Tohme, J.; Bonierbale, M.W. (1997). A molecular genetic map for cassava (*Manihot esculenta* Crantz). *Theoretical Applied Genetics*, 95:431-441.

Frontier, S. (1976). Étude de la décroissance des valeurs propes dans une analyse en composantes principales: comparaison avec le modèle du baton brisé. *Journal Experimental Marine Biology and Ecology*, 26:67-75.

Fukuda, W.M.G.; Guevara, C.L. (1988). *Descritores morfológicos e agronômicos para a caracterizcao de mandioca (Manihot esculenta Crantz).* EMBRAPA-CNPMF. Brasil. 38 p.

Gabriel, K.R. (1971). The biplot-graphic display of matrices with application to principal component analysis. *Biometrika*, 58:453-467.

BIBLIOGRAFIA

Gabriel, K.R. (1998). Generalised bilinear regression. *Biometrika*, 85:689-700.

Garcia, M.V.; Balatti, P.A.; Arturi, M.J. (2007). Genetic variability in natural populations of *Paspalum dilatatum* Poir. Analyzed by means of morphological traits and molecular markers. *Genetic Resources and Crop Evolution*, 54**(5):**935-946.

Garoia, F.; Guarniero, I.; Grifoni, D.; Marzola, S.; Tinti, F. (2007). Comparative analysis of AFLPs and SSRs efficiency in resolving population genetic structure of Mediterranean *Solea vulgaris*. *Molecular Ecology*, 16**(7):**1377-1387.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester. England. Wiley. 602 p.

Gillaspie, A.G.; Hopkins, M.S.; Dean, R.E. (2005). Determining genetic diversity between lines of *Vigna unguiculata* subspecies by AFLP and SSR markers. *Genetic Resources and Crop Evolution,* 52:245-247.

Gökirmak, T.; Mehlenbacher, S.A.; Bassil, N.V. (2008). Characterization of European hazelnut (*Corylus avellana*) cultivars using SSR markers. *Genetic Resources and Crop Evolution*, DOI 10.1007/s10722-008-9352-8.

Goldstein, D.B.; Ruíz-Linares, A.; Feldman, M.; Cavalli-Sforza, L.L. (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proceeding National Academy of Science USA,* 92**(15):**6723–6727.

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

Gower, J.C. (1971a). A general coefficient of similaririty and some of its properties. *Biometrics,* 27:857-874.

BIBLIOGRAFIA

Gower, J.C. (1971b). S*tatistical methods of comparing different multivariate analysis of the same data.* In: Mathematics in the archaeological and historical sciences. Hodson F.R.; Kendal, D.G.; Tautu, P. (Eds). Edinburgh university press. Edinburgh. pp 138-149.

Gower, J.C. (1975). Generalized Procrustes analysis . *Psychometrika,* 40:33-51.

Gower, J.C. (1985). *Measures of similarity, dissimilarity and distance. In: Encyclopedia of Statistics*. Vol. 5. Johnson, N.L.; Kotz, S.; Read, C.B. (Eds). Wiley. New York. pp 397-405.

Gower, J.C. (1992). Generalized biplots. *Biometrika*, 79**(3):**475-493.

Gower, J.C.; Legendre, P. (1986). Metric and Euclidean Properties of Dissimilarity coefficients. *Journal of Classification,* 3:5-48.

Gower, J.C.; Harding, S.A. (1988). Nonlinear biplots. *Biometrika*, 75**(3):**445-455.

Gower, J.; Hand, D. (1996). *Biplots.* Monographs on statistics and applied probability 54 Chapman and Hall. London. UK. 277 p.

Gower, J.C.; Dijksterhuis, G.B. (2004). *Procrustes Problems.* Oxford University Press. Oxford. UK. 248 p.

Graur, D.; Wen-Hsiung, R.I. (2000). *Fundamentals of Molecular Evolution.* Sinauer Associates Inc. USA. 481p.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis.* Academic Press London. 363p.

Greenacre, M.J. (1993). *Correspondence analysis in practique.* Academic Press London. 195p.

BIBLIOGRAFIA

Greenacre, M.J. (1994). *Correspondence analysis and its interpretation.* In: Greenacre, MJ.; Blasius, J. (Eds). Correspondence analysis in the social sciences. Academic Press. London. 3-22.

Grivet, L.; Glaszmann, J.C.; Arruda, P. (2001). Sequence polymorphism from EST data in sugarcane: a fine analysis of 6-phosphogluconate dehydrogenase genes. *Genetics and Molecular Biology*, 24:1-4.

Hall, B.G. (2001). *Phylogenetics trees made easy: A how-to manual for molecular biologists.* Sinauer Assoc. USA. 179 p.

Hamann, U. (1961) Merkmalsbestand und verwandtschaftsbeziehungen der farinosae. ein beitrag zum system der monokotyledonen. *Willdenowia,* 2:639-768.

Hartl, D.L.; Clark, A.G. (2006). *Principles of Population Genetics.* Sinauer Associates Inc. 4th edition. 545 p.

Harvey, M.; Botha, F.C. (1996). Use of PCR-based methodologies for determination of DNA diversity between *Saccharum* varieties. *Euphytica*, 89:257-265.

Heoa, M.; Gabriel, K.B. (2001). The fit of graphical displays to patterns of expectations. *Computational Statistics and Data Analysis*, 36:47-67.

Hillis, D.M. (1984). Misure and modification of Nei's genetic distance. *Systematic Zoology,* 33:238-240.

Hillis, D.M.; Moritz, C. (1990). *Molecular systematic: context and controversies. In: Molecular Systematics.* Hillis, D.M.; Morizt, C. (Eds). Sinauer Associates Inc. Massachusetts. USA. pp 1-11.

Hillis, D.M.; Wiens, J.J. (2000). *Molecules versus morphology in systematics. In: Phylogenetic analysis of morphological data.* Wiens, J.J. (Ed). Smithsonian Institution Press. Washington. USA. pp 1-19.

BIBLIOGRAFIA

Hubálek, Z. (1982). Coefficients of association and similarity, based on binary (presence-absence data): an evaluation. *Biological Reviews*, 57:669-689.

Hunter, B.F.; Hirsh, A.E.; Wall, D.P.; Eisen, M.B. (2004). Coevolution of gene expression among interacting proteins. *Proceeding National Academy of Science of USA*, 101**(24):**9033-9038.

Infante, D.; Molina, S.; Demey, J.R.; Gámez, E. (2006). Asexual genetic variability in Agavaceae determined with Inverse Sequence-Tagged Repeats and Amplification Fragment Length Polymorphism analysis. *Plant Molecular Biology Reporter*, 24**(2):**205-217.

Islam-Faridi, M.N.; Childs, K.L.; Klein, P.E.; Hodnett, G.; Menz, M.A.; Klein, R.R.; Rooney, W.L.; Mullet, J.E.; Stelly, D.M.; Price, H.J. (2002). A Molecular Cytogenetic Map of Sorghum Chromosome 1: Fluorescence in Situ Hybridization Analysis with Mapped Bacterial Artificial Chromosomes. *Genetics,*161:345-353.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin Society Vaudoise Science Natural*, 44: 223-270.

Jain, A.K.; Moreau, J.V. (1987). Bootstrap technique in cluster analysis. *Pattern Recognition*, 20**(5):**547-568.

Jorde, L. (1985). Human genetic distance studies: present status and future prospects. *Annual Review of Anthropology*, 14:343-373.

BIBLIOGRAFIA

Kalita, M.C.; Mohapatra, T.; Dhandapani, X.; Yadava, D.K.; Srinivasan, K.; Mukherjee, A.K.; Sharma, R.P. (2007). Comparative evaluation of RAPD, ISSR and anchored-SSR markers in the assessment of genetic diversity and fingerprinting of oilseed Brassica genotypes. *Journal of Plant Biochemistry and Biotechnology,* 16**(1):**41-48.

Kar, P.K.; Srivastava, P.P.; Awasthi, A.K.; Urs, S.R. (2008). Genetic variability and association of ISSR markers with some biochemical traits in mulberry (*Morus* spp.) genetic resources available in India. *Tree Genetics and Genomes,* 4:75-83.

Kaundun, S.S.; Park, Y.G. (2002). Genetic Structure of Six Korean Tea Populations as Revealed by RAPD-PCR Markers. *Crop Science,* 42:594-601.

Kempthorne, O. (1969). *An introduction to genetic statistics.* The Iowa University Press. Ames Iowa. USA. 228 p.

King, B.M.; Arents, P.A. (1991). Statistical test of consensus obtained from generalized procrustes analysis of sensory data. *Journal of Sensory Studies,* 6:37-48.

Kosman, E.; Leonard, K.J. (2005). Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid and polyploid species. *Molecular Ecology,* 14:415-424.

Krzanowski, WJ. (1979). Between-groups comparison of principal components. *Journal American Statistic Association,* 74**(367):**703-707.

Krzanowski, W.J. (1984). Principal Component Analysis in the presence of groups structure. *Applied Statistics,* 33:164-168.

Krzanowski, W.J. (1994). Ordination in the presence of group-structure, for general multivariate data. *Journal Classification,* 11:195-207.

BIBLIOGRAFIA

Krzanowski, W.J. (2000). *Principles of Multivariate Analysis: A User's Perspective,* revised edition. Oxford University Press. Oxford. UK. 608p.

Krzanowski, W.J. (2006). Sensitivity in Metric Scaling and Analysis of Distance. *Biometrics,* 62:239-244.

Kurczynski, T.W. (1970). Generalized distance and discrete variables. *Biometrics,* 26:525-534.

Lalouel, J.M. (1980). *Distance analysis and multidimensional scaling.* In: Current Developments in Anthropological Genetics: Theory and Methods. Vol. 1. Mielke J.H.; Crawford, M.H. (Eds). Plenum. New York. pp. 209-250.

Lance, G.N.; Williams, W.T. (1966). Computer programs for hierarchical polythetic classification. *Computer Journal,* 9:64-64.

Lanza, L.L.B.; de Souza, C.L. Jr; Ottoboni, L.M.M,; Vieira, M.L.C.; de Souza, A.P. (1997). Genetic distance of inbred lines and prediction of maize single-cross performance using RAPD markers. *Theoretical and Applied Genetics,* 94**(8):**1023-1030.

Latter, B.D.H. (1973a). The estimation of genetic divergence between populations based on gene frequency data. *American Journal Human Genetics,* 25:247-261.

Latter, B.D.H. (1973b). *Measures of genetic distance.* In: Genetic Structure of Populations. NE Morton (Ed). University Press of Hawaii. Honolulu. pp. 27-37.

Lavit, C. ; Escoufier, Y. ; Sabatier, R.; Traissac, P. (1994). The ACT (STATIS) method, Computational. *Statistics Data Analysis,* 18:97-119.

Lebart, L. (2004). *Validité des visualisations de données textuelles.* Actes des JADT. 7es Journées internationales d´Analyse statistique des Données Textuelles. Louvain, Belgium. 708-715.

BIBLIOGRAFIA

Lebart, L. (2007). *Which Bootstrap for Principal Axes Methods.* In: Part VII, Selected Contributions in Data Analysis and Classification. Brito, P.; Cucumel, G.; Bertrand, P.; de Carvalho, F. (Eds). Springer Berlin Heidelberg. 581-588.

Lebart, L. ; Morineau, A. ; Piron, M. (2000). *Statistique exploratoire multidimensionnelle.* Ed. Dunod. París. Francia. 108p.

Lee, M. (1995). DNA markers and plant breeding programs. *Advances in Agronomy*, 55:265-341.

Leeuw, J.; Meulman, J. (1986). A Special Jackknife for Multidimensional Scaling. *Journal of Classification,* 3:97-112.

Legèndre, L.; Legèndre, P. (1979). *Ecologie Numérique.* Vol. 1 and 2. Masson Editeur. In: Masson et Les Presses de l'université du Québec. Paris. pp 197-284.

Li, Y.; Shan, X.; Liu, X.; Hu, L.; Guo, W.; Liu, B. (2008). Utility of the methylation-sensitive amplified polymorphism (MSAP) marker for detection of DNA methylation polymorphism and epigenetic population structure in a wild barley species (*Hordeum brevisubulatum*). *Ecological Research,* 23:927-930.

Long, J.S. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Sage Publications. London. UK. 328pp.

Lopes, C.A.; Rodríguez, M.E.; Querol, A.; Bramardi, S.; Caballero, A.C. (2006). Relationship between molecular and enological features of Patagonian wine yeasts: relevance in selection protocols World. *Journal of Microbiology and Biotechnology*, 22:827-833.

BIBLIOGRAFIA

Lowe, A.J.; Hanotte, O.; Garino, L. (1996). Standardization of molecular genetic techniques for the characterization of germplasm collection: the case of random amplified polymorphic DNA (RAPD). *Plant Genetic Resources Newsletter*, 107:50-54.

Lu, Y.H.; D'Hont, A.; Walker, D.I.T.; Rao, P.S.; Feldmann, P.; Glassmann, J.C. (1994a). Relationships among ancestral species of sugarcane revealed with RFLP using single copy maize nuclear probes. *Euphytica*, 78:7-18.

Lu, Y.H.; D'Hont, A.; Paulet, F.; Grivet, L.; Arnaud, M.; Glassmann, J.C. (1994b). Molecular diversity and genome structure in modern sugarcane varieties. *Euphytica*, 78:217-226.

Lynch, M.; Milligan, B.G. (1994). Analysis of population genetic structure with RAPD markers. *Molecular Ecology*, 3:91-99.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceeding National Institute Science India.* 2:49-55.

Mardia, K.V.; Kent, J.T.; Bibby, M.J. (1979). *Multivariate analysis.* Academic Press. Londres. 521 p.

Martín, A. (2002). *Los marcadores genéticos en la Mejora Vegetal.* En: Genómica y Mejora Vegetal. Nuez, F.; Carrillo, J.M.; Lozano, R. (Eds). Mundi-Prensa. Sevilla. pp. 37-64.

Martínez-Gómez, P.; Sánchez-Pérez, R.; Rubio, M.; Dicenta, F.; Gradziel, T.M.; Sozzi, G.O. (2005). Application of recent biotechnologies to Pronus tree crop genetic improvement. *Ciencia Investigación Agraria*, 32:73-96.

Meulman, J.J. (1984). *Correspondence Analysis and Stability.* Research Report RR 84-01. Leiden University Press: Dept of Data Theory. Netherland.

BIBLIOGRAFIA

Milan, L.; Whittaker, J. (1995). Application of the Parametric Bootstrap to Models that Incorporate a Singular Value Decomposition. *Applied Statistics,* 44**(1):**31-49.

Milbourne, D.; Meyer, R.; Bradshaw, J.E.; Baird, E.; Bonar, N.; Provan, J.; Powell, W.; Waugh, R. (1997). Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. *Molecular Breeding,* 3:127-136.

Mora, F.; Santos, A.I.; Scapin, C.A. (2008). Mapeo de *loci* de caracteres cuantitativos (QTL) usando un enfoque multivariado. *Ciencia Investigación Agraria,* 35**(2):**137-145.

Nei, M. (1972). Genetic distance between populations. *American Naturalist,* 106:283-292.

Nei, M. (1978). The theory of genetic distance and evolution of human races. *Japanese Journal of Human Genetics,* 23:341-369.

Nei, M. (1987). *Molecular Evolutionary Genetics.* Columbia University Press, New York. 512 p.

Nei, M.; Li, W.H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceeding National Academy Sciences of USA.* 76:5269-5273.

Nei, M.; Kumar, S. (2000). *Molecular evolution and phylogenetics.* Oxford University Press, Oxford. UK. 266 p.

Nghia, NA.; Kadir, J.; Sunderasan, E.; Abdullah, M.P., Malik, A.; Napis, S. (2008). Morphological and Inter Simple Sequence Repeat (ISSR) Markers Analyses of *Corynespora cassiicola* Isolates from Rubber Plantations in Malaysia. *Mycopathologia,* 166:189-201.

BIBLIOGRAFIA

Ofori, A.; Becker, H.C.; Kopisch-Obuch, F.J. (2008). Effect of crop improvement on genetic diversity in oilseed *Brassica rapa* (turnip-rape) cultivars, detected by SSR markers. *Journal of Applied Genetics,* 49**(3):**207-212.

Pan, Y.B.; Burner, D.M.; Legendre, B.L.; Grisham, M.P.; White, W.H. (2004). An assessment of the genetic diversity within a collection of *Saccharum spontaneum* L. with RAPD-PCR. *Genetic Resources and Crop Evolution,* 51**(1):**895-903.

Payn, K.G.; Dvorak, W.S.; Janse, B.J.H.; Myburg, A.A. (2008). Microsatellite diversity and genetic structure of the commercially important tropical tree species *Eucalyptus urophylla*, endemic to seven islands in eastern Indonesia. *Tree Genetics and Genomes,* 4:519-530.

Pearson, K. (1926). On the coefficient of racial likeness. *Biometrica,* 18:337-343.

Perumal, R.; Krishnaramanujam, R.; Menz, M.A.; Katilé, S.; Dahlberg, J.; Magill, C.W.; Rooney, W.L. (2007). Genetic Diversity among Sorghum Races and Working Groups Based on AFLPs and SSRs. *Crop Science,* 47:1375-1383.

Porter, M.L.; Pérez-Losada, M.; Crandall, K.A. (2005). Model-based multi-locus estimation of decapods phylogeny and divergence times. *Molecular Phylogenetics and Evolution,* 37**(2):**355-369.

Powell, W.; Morgante, M.; Andre, C.; Hanafey, M.; Vogel, J.; Tingey, S.; Rafalsky, A. (1996). The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. *Molecular Breeding,* 2:225-238.

BIBLIOGRAFIA

Prakash, N.S.; Combes, M.C.; Dussert, S.; Naveen, S.; Lashermes, P. (2005). Analysis of genetic diversity in Indian robusta coffee genepool (*Coffea canephora*) in comparison with a representative core collection using SSRs and AFLPs. *Genetic Resources and Crop Evolution,* 52**(3):**333-343.

Primack, R.B., Kang, H. (1989). Measuring Fitness and Natural Selection in Wild Plant Populations. *Annual Review of Ecology and Systematics,* 20:367-396.

Raymond, M.; Rousset, F. (1995). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicist. *Journal of Heredity,* 86:248-249.

R Development Core Team. (2008). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Real Academia Española. (1999). *Ortografía de la lengua española.* 55 p.

Reiczigel, J. (1996). Bootstrap tests in Correspondence Analysis. *Applied Stochastic Models and Data Analysis,* 12:107-117.

Reif, J.C.; Xia, X.C.; Melchinger, A.E.; Warburton, M.L.; Hoisington, D.A.; Beck, D.; Bohn, M.; Frisch, M. (2004). Genetic Diversity Determined within and among CIMMYT Maize Populations of Tropical, Subtropical, and Temperate Germplasm by SSR Markers. *Crop Science,* 44:326-334.

Reynolds, J.B.; Weir, B.S.; Cockerham, C.C. (1983). Estimation of the co-ancestor co-efficient; basis for a short term genetic distance. *Genetics,* 105:767-779.

Ringrose, T.J. (1992). Bootstrapping and Correspondence Analysis in Archaeology. Journal of Archaeological. *Science,* 19:615-629.

BIBLIOGRAFIA

Roa, A.C.; Maya, M.M.; Duque, M.C.; Tohme, J.; Allen, A.C.; Bonierbale, M.V. (1997). AFLP analysis of relationships among cassava and other Manihot species. *Theoretical Applied Genetics,* 95:741-750.

Rogers, J.S. (1972). Measures of genetic similarity and genetic distance. *Studies in Genetics*, 7213:145-153.

Rogers, D.J.; Tanimoto, T.T. (1960). A computer program for classifying plants. *Science,* 132:1115-1118.

Rohlf, F.J.; Sokal, R.R. (1981). Comparing numerical taxonomic studies. *Systematic Zoology,* 30**(4):**459-490.

Rouf Mian, M.A.; Zwonitzer, J.C.; Chen, Y.; Saha, M.C.; Hopkins, A.A. (2005). AFLP Diversity within and among Hardinggrass Populations. *Crop Science*, 45:2591-2597.

Russel, T.S.; Rao, T.R. (1940). On habitat and association of species of Anotheline larvae in Southeastern Madras. *Indian Journal of Malariology*, 3:153-178.

Russell, J.R.; Fuller, J.D.; Macaulay, M.; Hatz, B.G.; Jahoor, A.; Powell, W.; Waugh, R. (1997). Direct comparison of levels of genetic variation among barley accessions detected by RFLPs, AFLPs, SSRs and RAPDs. *Theoretical Applied Genetics,* 95:714-722.

Saitou, S.N.; Nei, M. (1987). The neighbord-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology Evolution,* 4:406-425.

Sarwat, M.; Das, S.; Srivastava, P.S. (2008). Analysis of genetic diversity through AFLP, SAMPL, ISSR and RAPD markers in *Tribulus terrestris*, a medicinal herb. *Plant Cell Reports*, 27:519-528.

BIBLIOGRAFIA

Sato, Y.; Suganami, H.; Hamada, C.; Yoshimura, I.; Sakamoto, H.; Yoshida, T.; Yoshimura, K. (2006). The confidence interval of allelic odds ratios under the Hardy-Weinberg disequilibrium. *Journal of Human Genetics,* 51:772-780.

Schenck, S.; Crepeau, M.W.; Wu, K.K.; Moore, P.H.; Yu, Q.; Ming, R. (2004). Genetic diversity and relationships in native Hawaiian *Saccharum officinarum* sugarcane. *Journal of Heredity*, 95**(4):**327-331.

Selvi, A.; Nair, N.V.; Noyer, J.L.; Singh, N.K.; Balasundaram, N.; Bansal, K.C.; Koundal, K.R.; Mohapatra, T. (2005). Genomic constitution and genetic relationship among the tropical and subtropical Indian sugarcane cultivars revealed by AFLP. *Crop Science*, 45:1750-1757.

Shriver, M.D.; Jin, L.; Boerwinkle, E.; Deka, R.; Ferrell, R.E.; Chakraborty, R. (1995). A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Molecular Biology Evolution,* 12**(5):**914-920.

Smartt, J. (1981). Evolving gene pools in crop plants. *Euphytica,* 30**(2):**415-418.

Smith, C.A.B. (1977). A note on genetic distance. *Annals of Human Genetics,* 40:463-479.

Sneath, P.H.A.; Sokal, R.R. (1973). *Numerical taxonomy: The principles and practice of numerical classification.* Freeman W.H. and Co. San Francisco. USA. 573p.

Sokal, R.R.; Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University Kansas Science Bulletin,* 38:1409-1438.

Sokal, R.R.; Rohlf, F.J. (1962). The comparison of dendograms by objective methods. *Taxon,* 11:33-40.

Sokal, R.R.; Sneath, P.H.A. (1963). *Numerical taxonomy.* Freeman W.H. and Co. San Francisco. pp 359.

Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its applications to analysis of the vegetation on Danish commons. *Biological Skr,* 15:201-293.

Swofford, D.L.; Olsen, G.J. (1990). *Phylogenetic reconstruction.* In: Molecular systematics. Hillis, D.M.; Moritz, C. (Eds). Sinauer Associates Inc. Sunderland, Massachusetts. USA. pp 411-501.

Syamkumar, S.; Sasikumar, B. (2007). Molecular marker based genetic diversity analysis of Curcuma species from India. *Scientia Horticulturae,* 112**(2):**235-241.

Tamiru, M.; Becker, H.C.; Maass, B.L. (2007). Genetic Diversity in Yam Germplasm from Ethiopia and Their Relatedness to the Main Cultivated Dioscorea Species Assessed by AFLP Markers. *Crop Science,* 47:1744-1753.

Tan, Q.; Brusgaard, K.; Kruse, B.L.; Oakeley, E.; Hemmings, B.; Beck-Nielsen, H.; Hansen, L.; Gaster, M. (2004). Correspondence analysis of microarray time-course data in case-control design. *Journal of Biomedical Informatics,* 37**(5):**358-365.

Tar'an, B.; Zhang, C.; Warkentin, T.; Tullu, A.; Vandenberg, A. (2005). Genetic diversity among varieties and wild species accessions of pea (*Pisum sativum* L.) based on molecular markers, and morphological and physiological characters. *Genome,* 48**(2):**257-272.

Teklewold, A.; Becker, H.C. (2006). Geographic Pattern of Genetic Diversity Among 43 Ethiopian Mustard (*Brassica carinata* A. Braun) Accessions as Revealed by RAPD Analysis. *Genetic Resources and Crop Evolution,* 53:1173-1185.

Telles, M.P.C.; Bastos, R.P.; Soares, T.N.; Resende, L.V.; Diniz-Filho, J.A.F. (2006). RAPD variation and population genetic structure of *Physalaemus cuvieri* (Anura: Leptodactylidae) in Central Brazil. *Genetica*, 128:323-332.

Tenenhaus, M.; Young, F.W. (1985). An analysis and synthesis of multiple correspondence analyses, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika,* 50: 91-119.

Terzopoulos, P.J.; Bebeli, P.J. (2008). Genetic diversity analysis of Mediterranean faba bean (*Vicia faba* L.) with ISSR markers. *Field Crops Research,* 108**(1):**39-44.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature*, 426, 789-796.

The MathWorks Inc. (2008). *MATLAB Programming.* Natick, USA.

Tingey, S.V.; del Tufo, J.P. (1993). Genetic analysis with random amplified polymorphic DNA markers. *Plant Physiology*, 101:349-352.

Tucker, L.R. (1958). An inter-battery method of factor analysis. *Psychometrika,* 23**(2):**111-136.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics,* 29:614.

van Eeuwijk, F.A. (1995a). Multiplicative interaction in generalized linear models. *Biometrics*, 51:1017-1032.

van Eeuwijk, F.A. (1995b). Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica,* 84:1-7.

van Hintum, Th.J.L. (1995). *Hierarchical approaches to the analysis of genetic diversity in crop plants.* In: Core Collections of Plant Genetic Resources. Hodgkin, T.; Brown, A.D.H.; van Hintum Th.J.L. (Eds). IPGRI p 23-34.

Vicente-Villardón, J.L.; Galindo M.P., Blázquez-Zaballos, A. (2006). *Logistic Biplots.* In: Multiple correspondence analysis and related methods. Greenacre M.; Blasius, J. (Eds) Cham-man and Hall / CRC. Boca de Raton. USA. 608 p.

Vijayan, N.; Nair, S.; Screenivasan, T.V.; Mohan, M. (1999). Analysis of genetic diversity and phylogeny in *Saccharum* and related genera using RAPD markers. *Genetic Resource and Crop Evolution*, 46:73-79.

Wakeling, I.N.; Raats, M.M., MacFie, H.J.H. (1992). A new significance test for consensus in generalized procrustes analysis. *Journal of Sensory Studies*, 7:91-96.

Wang, L.; Guan, R.; Zhangxiong, L.; Chang, R.; Qiu, L. (2006). Genetic Diversity of Chinese Cultivated Soybean Revealed by SSR Markers. *Crop Science*, 46:1032-1038.

Weir, B.S. (1996). *Genetic data analysis 2: Methods for discrete population genetic data*. Sinauer Associates, 2nd edition. 445p.

Wilson, A.C.; Sarich, V.M.; Maxson, L.R. (1974). The importance of gene rearrangement in evolution: evidence from studies of rates of chromosomal, protein and anatomical evolution. *Proceeding National Academy Science*, 71:3028-3030.

Wilson, A.C.; Carlson, S.S.; White, T.J. (1977). Biochemical evolution. *Annual Review Biochemistry*, 46:473-639.

BIBLIOGRAFIA

Williams, J.G.K.; Kubelik, A.R.; Livak, K.J.; Rafalsky, J.A.; Tingey, S.V. (1990). Polymorphisms amplified by arbitrary initiators are useful as genetic markers. *Nucleic Acids Research,* 18:6531-6535.

Williams, J.G.K.; Hanafey, M.K.; Rafalsky, J.A.; Tingey, S.V. (1993). Genetic analysis using random amplified polymorphic DNA markers. *Methods in Enzymology,* 218:704-740.

Wright, S. (1978). *Evolution and the genetics populations.* Vol. 4, Variability in and among natural populations. University Chicago Press. Chicago. 580p.

Wright, S. (1984). *Evolution and the genetics populations: Genetic and Biometric Foundations*. Vol. 1. University Chicago Press. Chicago. 480 p.

Xiaoyan, Z.; Blair, M.W.; Wang, S. (2008). Genetic diversity of Chinese common bean (*Phaseolus vulgaris* L.) landraces assessed with simple sequence repeat markers. *Theoretical and Applied Genetics,* 117:629-640.

Xu, M.L.; Melchinger, A.E.; Lübberstedt, T. (1999). High-resolution mapping of loci conferring resistance to sugarcane mosaic virus in maize using RFLP, SSR, and AFLP markers. *Molecular General Genetics,* 261:574-581.

Xu, Z.; Zou, F.; Vision, T.J. (2005). Improving quantitative trait loci zapping resolution in experimental crosses by the use of genotypically selected samples. *Genetics,* 170:401-408.

Yeh, F.C.; Boyle, T.J. (1997). Population genetic analysis of co-dominant and dominant markers and quantitative traits Belgian. *Journal of Botany,* 129:157-166.

Yule, G.U. (1912). On the methods of measuring association between two attributes (with discussion). *Journal of the Royal Statistical Society,* 75:579-642.

BIBLIOGRAFIA

Zambrano, A.Y.; Demey, J.R.; Martínez, G.; Fuenmayor, F.; Gutiérrez, Z.; Saldaña, G.; Torrealba, M. (2002). Método rápido, económico y confiable de minipreparación de ADN para amplificaciones por RAPD en bancos de germoplasma. *Agronomía Tropical,* 52**(2):**235-243.

Zambrano, A.Y.; Demey, J.R.; Fuchs, M.; González, V.; Rea, R.; Desousa, O.; Gutiérrez, Z. (2003). Selection of sugarcane plants resistant to SCMV. *Plant Science,* 165**(1):**221-225.

Zambrano, A.Y.; Demey, J.R.; Fuenmayor, F.; Vicente-Villardón, J.L.; Ruiz, L.; Moreno, R.; Gutiérrez, Z.; Márquez, A.; Rodríguez, A. (2007). Genetic diversity of Venezuelan cassava collection. *Acta Horticulturae,* 738:729-733.

Zhao, L.; Shao, C.; Liao, X.; Chen, S. (2008). Isolation and characterization of polymorphic microsatellite loci from a dinucleotide-enriched genomic library of seven-band grouper (*Epinephelus septemfasciatus*) and cross-species amplification. *Conservation Genetics*, DOI 10.1007/s10592-008-9593-2.

Zintzaras, E. (2008). Variante estimation of allele-based odds ratio in the absence of Ardí-Weinberg equilibrium. *European Journal of Epidemiology,* 23:323-326.

Anexo

BIOINFORMATICS ORIGINAL PAPER

2008, pages 1–7
doi:10.1093/bioinformatics/btn552

Genome analysis

Identifying molecular markers associated with classification of genotypes by External Logistic Biplots

J. R. Demey[1,*], J. L. Vicente-Villardón[2], M. P. Galindo-Villardón[2] and A. Y. Zambrano[3]

[1]Centro de Biotecnología, Instituto de Estudios Avanzados (IDEA), Caracas, Venezuela, [2]Departamento de Estadística, Universidad de Salamanca, Salamanca, España and [3]Instituto Nacional de Investigaciones Agrícolas (INIA-CENIAP), Maracay, Venezuela

Received on April 26, 2008; revised on September 30, 2008; accepted on Ocotber 22, 2008

Associate Editor: Dmitrij Frishman

ABSTRACT

For characterization of genetic diversity in genotypes several molecular techniques, usually resulting in a binary data matrix, have been used. Despite the fact that in Cluster Analysis (CA) and Principal Coordinates Analysis (PCoA) the interpretation of the variables responsible for grouping is not straightforward, these methods are commonly used to classify genotypes using DNA molecular markers. In this article, we present a novel algorithm that uses a combination of PCoA, CA and Logistic Regression (LR), as a better way to interpret the variables (alleles or bands) associated to the classification of genotypes. The combination of three standard techniques with some new ideas about the geometry of the procedures, allows constructing an External Logistic Biplot (ELB) that helps in the interpretation of the variables responsible for the classification or ordination. An application of the method to study the genetic diversity of four populations from Africa, Asia and Europe, using the HapMap data is included.

Availability: The Matlab code for implementing the methods may be obtained from the web site: http://biplot.usal.es.

Contact: jhonny.demey@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

To characterize and evaluate the genetic diversity, various molecular techniques have been employed, including Restriction Fragment Length Polymorphisms (RFLP), Random Amplified Polymorphic DNAs (RAPD), Amplified Fragment Length Polymorphisms (AFLP), Sequence Tagged Sites (STS), Simple Sequence Repeats (SSR) or microsatellites, Single Nucleotide Polymorphisms (SNPs) (Avise, 2004).

Molecular information is analyzed according to the type of marker and organism. Generally, a binary data matrix is obtained from amplified fragments coded as 'presence' (1's) or 'absence' (0's). Genetic relationships among genotypes are investigated using different techniques of classification/ordination such as UPGMA or Neighbor-joining clustering algorithm (Saitou and Nei, 1987; Sneath and Sokal, 1973) and Principal Coordinates Analysis (PCoA) (Gower, 1966). Despite its extensive use, these techniques of classification/ordination do not permit to study genotype–allele and allele–allele relations appropriately, i.e. it is not possible to determine which alleles or bands are responsible for (or associated to) the classification of genotypes.

An approach that facilitates the genetic interpretation, compared to the classic techniques of classification/ordination, is provided by the Biplot methods (Chapman *et al.*, 2002; Gabriel, 1971; Sharov *et al.*, 2005), that is, a simultaneous graphical representation of the rows (individuals) and the columns (variables) of a given data matrix. The main uses are exploratory, although it has also been used as a graphical representation for more formal models (Gabriel, 1998). The biplot can be fitted by performing alternating regressions and interpolations (Gabriel and Zamir, 1979; Gower and Hand, 1996; Jongman *et al.*, 1995). However, when data are binary, like those obtained in the analysis of molecular information, Classical Linear Biplots and Principal Components Analysis (PCA) are not suitable because the response along the dimensions is linear. This is the same reason why linear regression is not appropriate for binary or categorical data.

Several strategies can be used in order to fit biplots from binary data matrices: Multiple Correspondence Analysis (MCA) can be considered as a particular form of biplot for a binary matrix, where the prediction regions are based on distances from the individual points to the category points (Gower and Hand, 1996); when modeling a two-way table using a bilinear model, the parameter estimates are obtained by an iterative process of alternating generalized row and column regressions (Falguerolles, 1998; Gabriel, 1998; van Eeuwijk, 1995a, b; Vicente-Villardón *et al*., 2006). Nevertheless MCA depends on the chi-squared distance, which does not reflect the structure of our data. In this case, PCoA provides a more flexible alternative, because we can use different similarity/dissimilarity measures to extract the genetic relationships among genotypes.

Vicente-Villardón *et al*. (2006) described the geometry of a linear biplot for binary data in which the response along the dimensions is logistic (Logistic Biplots, LB). In the LB, each individual is represented as a point and each variable as a direction through the origin. The projection of an individual point onto a character direction predicts the probability of presence of that character. The method is related to LR in the same way that biplot analysis is related to linear regression.

*To whom correspondence should be addressed.

In this article, we use a combination of PCoA, Cluster Analysis (CA) and External Logistic Biplots (ELB) on the principal coordinates, as a better way to identify the alleles or bands that are responsible for the classification of genotypes. The proposal is based on the fact that the column regression in the alternating procedure for binary data is simply a LR that can be fitted to the configuration obtained from PCoA. Although the whole alternating procedure could have been used, PCoA is simpler, more accessible to applied researchers and, in the authors' experience, the results are similar. On the other hand, the alternating procedures, as described in the literature, although share the same geometry; need some adaptations for binary data matrices.

We have taken an exploratory point of view as opposed to the modeling approach in papers by van Eeuwijk (1995a, b), Gabriel (1998) or Falguerolles (1998). The main aim is to analyze a data matrix (individuals by variables) rather than to model a two-way (contingency) table using a bilinear model. Our proposal is closely related to MCA and some psychometric latent variable procedures such as item response theory or latent traits.

Measures of the quality of the representation of individuals, groups of individuals and variables (alleles or bands) are also defined. It is shown that this approach facilitates the genetic interpretation as compared to traditional clustering methods.

Some theory is developed for the proposal and a simulation study shows the performance of the methodology. An application of the method to study the genetic diversity of four populations from Africa, Asia and Europe, using the HapMap data is included.

2 METHODS

Let $\mathbf{X}$ be the matrix of binary data obtained from amplified fragments that were scored as present or absent (1 or 0), in which the rows correspond to n individuals or entries (genotypes) and the columns to p binary characters (alleles or bands). Let $\mathbf{S}=(s_{ij})$ be a matrix containing the similarities among genotypes, obtained from the binary data matrix $\mathbf{X}$, and let $\Delta=(\delta_{ij})$ be the corresponding dissimilarity/distance matrix, taking for example $\delta_{ij}=1-s_{ij}$.

The algorithm starts with a PCoA, as a technique of ordination of the individuals (genotypes). PCoA is concerned with the problem of constructing a configuration of n points in an Euclidean space in such a way that the distance between any two points of the configuration approximates, as closely as possible, the dissimilarity (δ_{ij}) between genotypes represented by these points. The objective is then to find a configuration $\mathbf{Y}$ in a lower dimensional Euclidean space $\mathbb{R}^k$ whose inter-point distance matrix $\mathbf{D}$ is as close as possible to Δ. When the observed dissimilarity/distance measured is 'Euclidean', it is possible to find an exact configuration in $n-1$ dimensions. A lower dimensional approximation can be obtained projecting onto the first k principal coordinates (usually $k=2$). The theoretical considerations and demonstrations of the method can be found in Mardia *et al*. (1979).

In PCoA, it is known that the proportion of the total variance explained by k dimensions (overall goodness of fit or overall quality of representation) can be considered as an average of the n points in the graphical representation. However, a good overall fit does not imply that all the individuals have the same quality of representation and then that the interpretation of the positions of all the points in the diagram is equally reliable—this has not received sufficient attention in published research articles and major statistical packages. We consider that an individual is well represented when most of its information (measured through the variability) is accounted for in the reduced dimension. As the representation is centered at the origin, the variability of each individual is measured by its squared distance to the center, so that the quality of representation can be measured by the ratio between the squared distance in the reduced dimension and the squared distance in the complete space, that is:

$$CR_i^k=\frac{\sum_{l=1}^{k} y_{il}^2}{\sum_{j=1}^{n-1} y_{ij}^2}\times 100\% \qquad (1)$$

where y_{ij} *d*enotes the principal coordinates of individual i in the j-th dimension. Geometrically, it is the squared cosine of the angle between the vector in the complete space and its projection onto the representation space. For groups of individuals, the quality of representation (on the PCoA ordination diagram) is calculated as in (1) using its centroid- $\bar{y}_{gj}$ the average of the coordinates in j-th dimension for the group g.

Unlike PCA in its Biplot version, where the new axes can be interpreted in terms of the original variables, in PCoA, the axes have no direct meaning. Therefore it is not possible to interpret the relationship between genotype–allele/band and allele/band–allele/band. It can be shown that PCA configurations are also obtained applying PCoA to the matrix of Euclidean distances. A classical biplot is obtained by fitting linear regressions to that configuration as described in Vicente-Villardón *et al*. (2006). Hence, an immediate heuristic generalization in this context is to use LRs and its graphical representation on the PCoA, called ELB, rather than linear regressions.

To search for the variables associated to the ordination obtained in PCoA, we can look for the directions in the ordination diagram that better predict the probability of presence of each allele.

More formally, define $\pi_{ij}=E(x_{ij})$ as the expected probability that the allele j be present at genotype for a genotype with coordinates y_{is} ($i=1,\ldots,n$; $s=1,\ldots,k$) on the ordination diagram, then

$$\pi_{ij}=\frac{e^{b_{j0}+\sum_{s=1}^{k} b_{js}y_{is}}}{1+e^{b_{j0}+\sum_{s=1} b_{js}y_{is}}}$$

where b_{js} ($j=1,\ldots,p$) are the LR coefficients that correspond to the j-th variable (alleles or bands) in the s-th dimension. The model is a generalized linear model having the logit as a link function.

$$logit(\pi_{ij})=log(\frac{\pi_{ij}}{1-\pi_{ij}})=b_{j0}+\sum_{s=1}^{k} b_{js}y_{is}=b_{j0}+\mathbf{y}_i^{\intercal}\mathbf{b}_j$$

where $\mathbf{y}_i=(y_{i1},\ldots,y_{ik})^T$ and $\mathbf{b}_j=(b_{j1}\ldots b_{jk})^T$, $\mathbf{y}$'s and $\mathbf{b}$'s define a biplot in logit scale. This is called External Logistic Biplot because the coordinates of the genotypes are calculated in an external procedure (PCoA). Given that the $\mathbf{y}$'s are known from PCoA, obtaining the b's is equivalent to performing a LR using the j-th column of $\mathbf{X}$ as a response variable and the columns of $\mathbf{y}$ as regressors.

The regression equation predicts the probability that an allele will be present in that genotype. Geometrically, the y's can be represented as points in the reduced dimension space and the $b's$ are the vectors showing the directions that best predict the probability of presence of each allele π_j. For a complete explanation of the geometrical properties of the ELB (see Vicente-Villardón *et al*., 2006).

The prediction of the probabilities is made in the same way as in a linear Biplot, i.e. the projection of a genotype point on the direction of an allele vector predicts the probability of presence of that allele in the genotype. To facilitate the interpretation of the graph, fixed prediction probabilities points are situated on each allele vector. To simplify the graph, in our application, a vector joining the points for 0.5 and 0.75 are placed; this shows the cut point for prediction of presence and the direction of increasing probabilities. The length of the vector can be interpreted as an inverse measure of the discriminatory power of the alleles or bands, in the sense that shorter vectors correspond to alleles that better differentiate individuals. Two alleles pointing in the same direction are highly correlated, two alleles

pointing in opposite directions are negatively correlated, and two alleles forming an angle close to 90° are almost uncorrelated.

For each allele, the ordination diagram can be divided into two separate regions predicting presence or absence, the two regions are separated by the line that is perpendicular to the allele vector in the Biplot and cuts the vector at the point predicting 0.5. The alleles associated to the configuration are those that predict the presences adequately.

In a practical situation not all the alleles are associated to the ordination. Due to the high number of alleles usually studied, it is convenient to situate on the graph only those that are related to the configuration, i.e. those that have an adequate goodness of fit after adjusting the LR.

A goodnees-of-fit (or quality of the representation) criterion, to select the alleles, is the 'percentage of correct classifications' calculated as the percentage of coincidences between the binary data matrix and the expected binary matrix obtained from the LR models. When the percentage of correct classification is added for all the alleles, the overall goodnees-of-fit of the logistic Biplot is obtained. Additionally, the pseudo R^2-Squared performed according to Nagelkerke/Cragg & Uhler's (Long, 1997) for the regressions of categorical outcome variables are used as measures of the 'quality of the representation' and this is interpreted in the manner commonly used in correspondence analysis (Tenenhaus and Young, 1985).

Additionally, a Bonferroni correction can be used as criterion of selection of alleles with higher discriminatory power. With this method, only those alleles that have a given significance level ($P \leqslant 0.05$/ total number of alleles) will be included in the biplot.

For large data sets P-values are highly affected by the sample size and the number of alleles. In these cases, it is better to use the pseudo R^2 with a highly restrictive value, for example $R^2 \geqslant 0.9$, because pseudo R^2 is less sensitive to the sample size.

Frequently the analysis also obtains groupings; Cluster Analysis (CA) can be applied using the initial distance matrix $\mathbf{\Delta}$ or the fitted Euclidean distance matrix $\mathbf{D}$ obtained from the PCoA. The partition obtained is represented on the PCoA ordination using the convex hulls of the points belonging to each cluster. It could be argued that using the principal coordinates $\mathbf{Y}$ or the fitted distance $\mathbf{D}$ for additional analysis can result in a loss of information. This can also be thought as a way to separate the signal from the noise; the loss of information that entails the use of principal coordinates it is compensated by the noise level that is reduced, additionally we guarantee the orthogonality of the regressors. Chae and Warde (2006) show that the retrieval abilities of the known agglomerative clustering algorithms are improved by using principal coordinates as prior data.

The ELB is then used to evaluate the separation of the groups and to search for the alleles associated to the groups.

In summary, the general algorithm for ELB works as follows: (i) make a PCoA of the binary data matrix, using the most adequate similarity coefficient for the data; (ii) calculate standard LR using the principal coordinates as independents variables and each allele or band as dependent; (iii) plot the Biplot filtering the variables using the Bonferroni correction; and (iv) draw the groups using the principal coordinates and the most adequate clustering algorithm.

3 SIMULATION STUDY

3.1 Method

In order to show the behavior of ELB in identifying molecular markers associated with the classification of genotypes simulated data will be used. Simulations are powerful tools for evaluating the performance of a method because we know the a priori structure of groups and the variables responsible for the classification. Three basic scenarios typical of real data sets (worst case scenarios) were investigated. In order to facilitate the visualization of the Biplot representation, a moderate number of genotypes ($n = 50$) was included.

Matrices of binary data with known group structure were generated and scored as present or absent (1 or 0); the rows correspond to 50 individuals or entries (genotypes) and the columns to p binary characters (alleles or bands). Each group was characterized by a set of alleles in such a way that the characteristic alleles were present in that group and absent in the rest. External and internal noises were added to the matrices. The external noise consisted of adding a set of supplementary alleles. These supplementary alleles were generated using a uniform distribution (0,1), the alleles were considered present if the simulated value of the samples $x_i \geqslant 0.5$; and absent otherwise. An internal noise was added to the total set of alleles. It consisted in modifying, at random, 5%, 10% and 20% of the values assigned to each individual by allele, i.e. the values of presence in the matrices were replaced by absence or vice versa in the indicated percentage. The internal noise is just the percentage of genotyping errors. Although the 20% error can be considered too large for advanced genotyping technology, that percentage would be useful for another application like the detection call in Affymetrix microarrays. Using a high percentage is also useful to assess the reliability of the proposal even in extreme situations.

Dice similarity coefficient and the UPGMA clustering algorithm were used.

Twenty-seven scenarios were simulated in total. Each scenario was repeated 1000 times.

The performance of the proposed method was evaluated using the following criteria: the variance accounted by the first two dimensions and its comparison with the standard for this type of experiments, the projection of a group on the direction of an allele vector that theoretically is present in the group and therefore predicts the probability of presence of allele in the group, the sensitivity of the method measured through the quality of the representation of the individuals (QRIndividuals), the quality of the representation of the alleles (QRAlleles), the percentage of correct classifications of the alleles (%CCAlleles) and the error rate of classification (ERC). Table 1 shows the different combinations generated by the three basic scenarios.

All the analyses were computed with programs especially written for this purpose using MatLab versión 2008a (The MathWorks Inc, 2008). (Available at http://biplot.usal.es).

3.2 Results

Figure 1a shows the distribution of the accounted variance for the different scenarios. The eigenvalues and their distribution pattern are within the limits for the studies of genetic diversity using molecular markers. An amplification of the first principal plane used in the Biplot representation is shown in Figure 1b, c and d. The ***S1*** scenarios accounted for more variance, additionally, both types of noise (internal and external) considered in the simulations influence the different forms of accumulated variance. The accounted variance values decrease more as internal noise increases than when the total number of supplementary alleles or added external noise increase.

Table 2 shows the quality-of-the-representation-of-the-individuals, which has similar behavior as the accounted variance: it is most affected by internal noise. However, the group structures previously designed is maintained. This fact indicates that the number of supplementary alleles does not affect the structure of groups defined previously, i.e. the reduction of the dimensionality and the representation of the individuals in the bidimensional plane

Table 1. Scenarios simulated

Scenarios	Groups	Number of individuals/groups	Number of alleles[a]	Supplementary alleles/ external noise	Total number of alleles	Internal noise (%)
$S1(a_1,b_1,c_1)$	2	(20,30)	12	30	42	$(a_1{:}5,b_1{:}10,c_1{:}20)$
$S1(a_2,b_2,c_2)$	2	(20,30)	12	48	60	$(a_2{:}5,b_2{:}10,c_2{:}20)$
$S1(a_3,b_3,c_3)$	2	(20,30)	12	66	78	$(a_3{:}5,b_3{:}10,c_3{:}20)$
$S2(a_1,b_1,c_1)$	3	(10,15,25)	20	50	70	$(a_1{:}5,b_1{:}10,c_1{:}20)$
$S2(a_2,b_2,c_2)$	3	(10,15,25)	20	80	100	$(a_2{:}5,b_2{:}10,c_2{:}20)$
$S2(a_3,b_3,c_3)$	3	(10,15,25)	20	110	130	$(a_3{:}5,b_3{:}10,c_3{:}20)$
$S3(a_1,b_1,c_1)$	4	(6,10,14,20)	34	85	119	$(a_1{:}5,b_1{:}10,c_1{:}20)$
$S3(a_2,b_2,c_2)$	4	(6,10,14,20)	34	136	170	$(a_2{:}5,b_2{:}10,c_2{:}20)$
$S3(a_3,b_3,c_3)$	4	(6,10,14,20)	34	187	221	$(a_3{:}5,b_3{:}10,c_3{:}20)$

[a]Alleles that define group's structure.

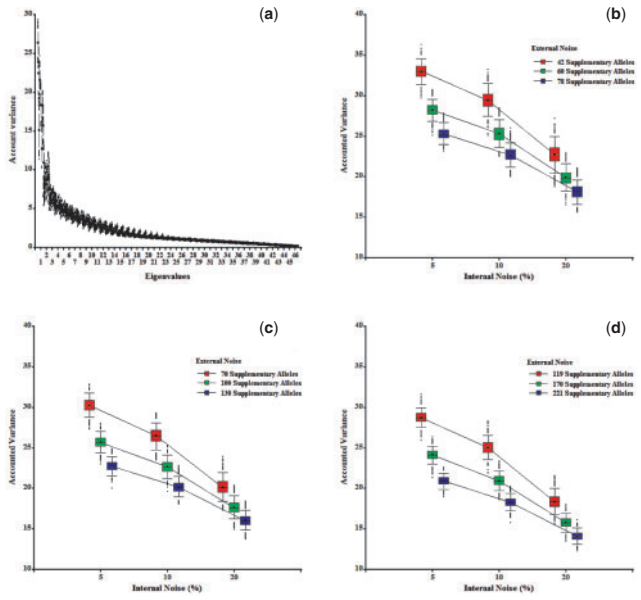


Fig. 1. Distribution of the accounted variance values for different scenarios: (**a**) all eigenvalues and (**b, c, d**) first two dimensions.

Table 2. Simulation results

Scenarios	Average QR individuals	QRAlleles/ groups	QRAlleles/ supplementary	% CCAlleles/ groups	% CCAlleles/ supplementary	ERC
$S1a_1$	30.87	0.89	0.15	95.16	60.66	0.00
$S1a_2$	26.55	0.89	0.13	95.10	59.65	0.01
$S1a_3$	23.90	0.88	0.12	95.11	59.09	0.02
$S1b_1$	27.34	0.76	0.15	90.11	60.69	0.13
$S1b_2$	23.70	0.76	0.13	90.12	59.65	0.20
$S1b_3$	21.41	0.76	0.12	90.03	59.20	0.27
$S1c_1$	20.99	0.53	0.15	79.99	60.74	3.81
$S1c_2$	18.57	0.52	0.13	79.63	59.87	4.71
$S1c_3$	17.06	0.51	0.12	79.31	59.42	5.90
$S2a_1$	29.16	0.88	0.08	95.03	57.20	0.25
$S2a_2$	24.84	0.87	0.08	94.82	57.27	0.72
$S2a_3$	22.06	0.86	0.08	94.59	57.34	1.68
$S2b_1$	25.42	0.73	0.09	89.59	57.61	1.69
$S2b_2$	21.85	0.72	0.09	89.21	57.76	3.60
$S2b_3$	19.52	0.70	0.09	88.76	57.74	5.86
$S2c_1$	19.22	0.45	0.11	78.08	59.00	14.26
$S2c_2$	16.92	0.43	0.11	77.22	58.71	17.65
$S2c_3$	15.44	0.41	0.10	76.58	58.47	20.04
$S3a_1$	28.50	0.82	0.07	94.20	56.54	1.52
$S3a_2$	23.90	0.81	0.07	94.10	56.65	2.55
$S3a_3$	20.75	0.81	0.07	93.89	56.65	4.00
$S3b_1$	24.73	0.67	0.07	88.88	56.78	4.86
$S3b_2$	20.71	0.66	0.07	88.59	56.86	6.87
$S3b_3$	18.07	0.65	0.07	88.18	56.89	8.78
$S3c_1$	17.97	0.40	0.09	77.53	57.71	16.27
$S3c_2$	15.44	0.38	0.09	76.56	57.71	19.94
$S3c_3$	13.82	0.36	0.09	75.73	57.63	23.15

are not affected even with small variance absorption. Heoa and Gabriel (2001) demonstrated that graphical displays of multivariate data often clearly exhibit features of the expectations even though the data themselves are poorly fitted by the displays. Thus, it often occurs that ordinations and biplots that poorly fit the sample data still reveal salient characteristics such as clusters of similar individuals and patterns of correlation.

As an example, Figure 2 shows the Biplot representation of the relations among the individuals and the alleles that determine the group structure in some of the simulations. It can be observed that the alleles that have been marked as those of importance in the formation of the groups are projected on the directions of greater probability for the groups that they define. The supplementary alleles are projected on the Biplot in irregular form and with higher lengths indicating a low discriminatory power (see practical interpretation rules in the Supplementary Material). After Bonferroni corrections (5%) almost all supplementary alleles were eliminated.

Only in the case of the Figure 2c, Group 2 does not show the alleles associated with its formation, and this is due to the fact that, generally, $g-1$ or less axes are needed to correctly retain the structure of groups. In this case, it is probable that we would need to include a third axis to be able to observe more clearly the variables that are associated with the structure of this group. All the alleles that defined groups had high quality of representation and percentage of correct classifications, indicating that there exists a high degree of coincidences between the binary data matrix and the expected binary matrix obtained from the LR models.

Although the group structure was known, the UPGMA method was used here to calculate the error rate of classification in order to check the sensitivity of the method when the structure is not known 'a priori'. In the same manner as the other sensitivity criteria, the error rate of classification had a similar behavior with respect to internal noise. The good sensitivity of the method is clearly demonstrated by the fact that in all scenarios the error rates are not >25% in the worst case and <9% when the genotyping error is <10%.

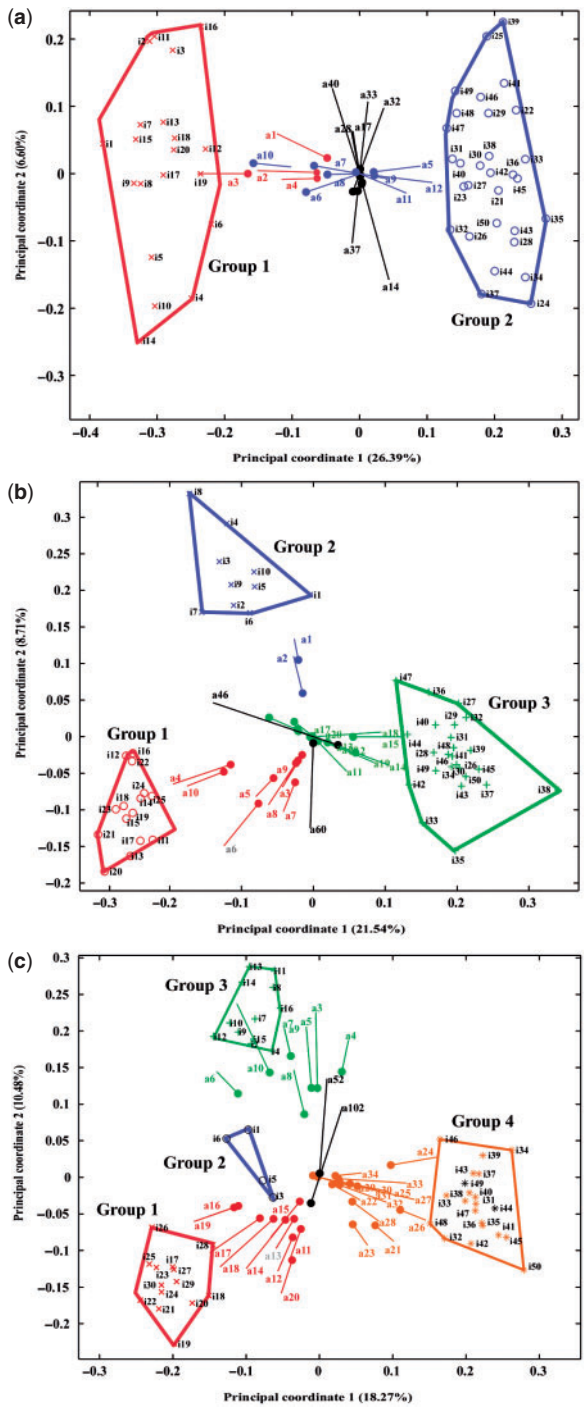


Fig. 2. Biplot representation showing the relationships among individuals and alleles based on Dice dissimilarity matrix: (**a**) scenario $S1a_1$, (**b**) scenario $S2a_1$ and (**c**) scenario $S3a_1$.

Finally, we can conclude that independently of the number of alleles or amplification fragments that are evaluated, the method recovers those of importance in the definition of the structure or natural grouping of the individuals in the first principal coordinates and that the CA almost always achieves nearly 100% accuracy for assigning individuals to their true groups when the genotyping error is small.

4 APPLICATION TO HAPMAP DATA

The practical merit of our methodology is additionally illustrated in a real data study of the genetic diversity of four populations from Africa, Asia and Europe, using the genotype data generated by the International HapMap Consortium (2003).

4.1 Genotype data

All our analyses are based on the HapMap phase 3 genotypes for chromosome 22 from samples of four populations: 171 Maasai (MKK); 82 Han Chinese in Beijing, China (CHB); 82 Japanese in Tokyo, Japan (JPT) and 162 with Western European ancestry (CEU), available for bulk downloads at http://www.hapmap.org/ (Supplementary Table 1). Previously to the analysis, the matrix was depurated eliminating the monomorphic SNPs and those with missing values. Data were scored like a binary matrix using only SNPs common to all populations obtaining a matrix of 497 individuals by 14 666 alleles corresponding to 7333 genotypes. The alleles were labeled using the SNP name and the allele: rs1314-T and rs1314-G, respectively. (The complete list of variables is shown in Supplementary Table 2).

4.2 Method

Genetic relationships among the 497 individuals were investigated using PCoA, CA and ELB as described above. To compare their behavior in the proposed context, the following coefficients were used: Dice, Jaccard, Rogers and Tanimoto and Simple Matching (Sneath and Sokal, 1973). The correlation between the observed and fitted distances, based on genetic dissimilarity, was used to establish a criterion for the selection of the number of axes, measure the grouping analysis performance and to evaluate the stability of the constructed relationship. Although CA is not necessary because the group structure is known, the UPGMA cluster algorithm was used to check the group structure obtained from the principal axes retained. Measures of quality of representation of groups and variables (alleles) were obtained as described above.

4.3 Results

The distribution of the correlation values among observed and expected distance matrices for different dissimilarity coefficients and several combinations of principal coordinates retained showed that for the three principal coordinates, the simple matching coefficient is the one that better defines the DNA sequence variation patterns using high-density SNPs genotyping arrays. The correlation values among observed and expected distance matrices are interpreted in a similar form to Cophenetic correlation values.

Figure 3 shows 3D representation of the first three principal coordinates. The four populations are clearly separated using the first three principal coordinates. The first five dimensions account for 15.96%, 10.00%, 1.27%, 0.97% and 0.91% of the total variance. Plane 1–2 differentiates between MKK, CEU and the Asian (CHB and JPT, that are mixed in the graphical representation). In our analysis, when we introduce the third axis it is possible to separate the CHB from JPT. Further axes do not contribute with

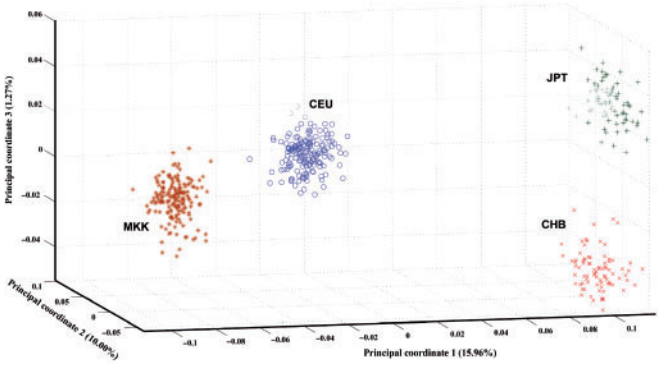


Fig. 3. 3D representation of the first three principal coordinates showing the genetic relationships among individuals based on simple matching dissimilarity matrix, whole dataset.

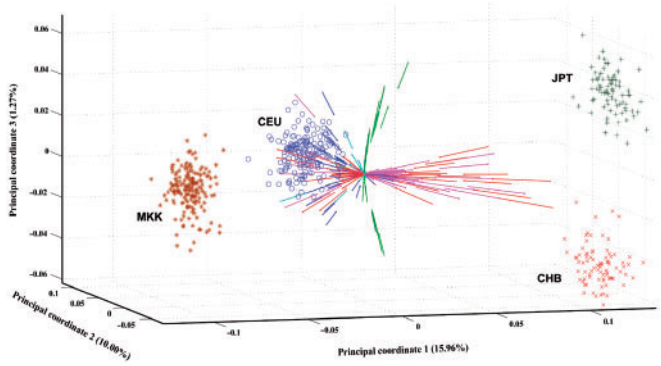


Fig. 4. The 3D Biplot projection of the alleles with higher discriminatory power onto the first three principal coordinates, using whole dataset.

discriminatory information, i.e. adding new axes does not introduce any information about the differences among populations.

The centroid of a group could be taken as the representative of the group, the more compact is the group the better is the summary. In the graphical representation, all the individuals belonging to the same group are close in the graph, showing that all share a common DNA pattern. The fact that the four populations are clearly separated using just the first three principal coordinates means that the main source of variation in the SNPs patterns is the differentiation among groups, i.e. the differences among populations are higher than the differences within populations. Supplementary Figure 1 shows more specifically the projection obtained from PCoA of the individuals on the planes 1–2 (1a), 1–3 (1b) and 4–5 (1c).

The projection of the alleles on the principal coordinates solution using ELB, on the planes 1–2 (2a) and 1–3 (2b) is showed in Supplementary Figures 2. The alleles are represented by lines pointing in the direction of increasing predictions of the probabilities. As mentioned previously, the beginning of the line is the point predicting 0.5 and the end, the point predicting 0.75. The length of each is related to capacity to predict presence or absence of the alleles in each individual (see practical interpretation rules in the Supplementary Material). After adjusting the ELB, the global goodness-of-fit as a percentage of correct classifications was 81.04%, i.e. if the biplot is used to predict each genotype for each individual, 81.04% of them will be correct; the percentage is high even when many of the alleles are not associated with any pattern in data. If we consider each allele separately, the percent of correct classifications of 75% of the alleles was higher than 69.42%.

Using the Bonferroni correction, 9319 alleles (63.55%) were selected (Supplementary Figure 3). As a consequence of the large sample size, most of those *P*-values are associated with low explanatory power and are probably not very useful for the discrimination among populations. The quality of the representation measured through the pseudo R^2, was <0.5 in 8219 (92.83%) of the selected alleles and >0.9 in 1.78% of the cases, i.e. from the 14 666 alleles, 260 have the highest discriminatory power. This more restrictive selection criterion is less sensitive to the sample size and permitted selecting the most important alleles to determine the patterns of DNA sequence variation among the four populations. Figure 4 shows the 3D representation of the alleles with a $R^2 > 0.9$. The bidimensional projections into planes 1–2 (4a) and 1–3 (4b) are depicted in Supplementary Figure 4 and the list of selected alleles is shown in Table 3 of the same.

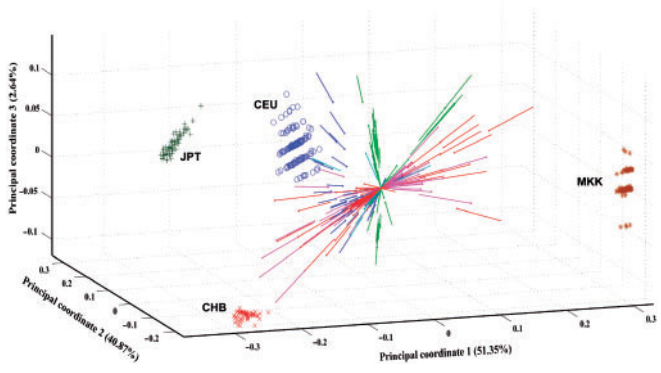


Fig. 5. The 3D Biplot projection of the alleles onto the first three principal coordinates, using reduced dataset.

The quality of representation of groups calculated with the three first retained dimensions was >99% for all the populations. The patterns of DNA sequence variation among the four populations generated with all the SNPs (14 666) and with the small subset of SNPs (260) are similar (Fig. 5). Supplementary Figure 5 shows the bidimensional projections into planes 1–2 (5a) and 1–3 (5b). Global goodness-of-fit as a percentage of correct classifications for the reduced matrix was 98.69%.

It has been demonstrated that the proposed methodology is useful when evaluating large datasets such as data from the HapMap project. It allows recovering the structure of the studied populations with small dataset. This helps to reduce the problem of multiple comparisons that arises from testing tens to hundreds of thousands of SNPs and haplotypes for disease associations. In such settings, it is often desirable to reduce the number of markers needed for structure identification and the identification of the functionally important SNPs. Additionally, the method helps to visualize the data structure in a reduced dimension.

The described procedure has been successfully applied to different types of binary data and contexts—with or without knowing a previous group structure—for example: detection call in Affymetrix microarrays (Vicente-Villardón *et al*., 2006) and genetic diversity studies in plants collections (Demey *et al.*, 2006).

Selection procedures using univariate statistics to compare the populations with each allele are commonly used by most researchers. For example, Weir *et al.* (2005) using F_{ST} as a measurement of genetic population structure also found high similarity between the HapMap Han Chinese from Beijing (HCB) and Japanese from Tokyo (JPT) for the majority of the chromosomes studied including the 22. However, it is not clear how this approach can be applied to the selection of informative markers when the parental information is unknown (Rosenberg *et al*., 2003).

Our approach uses the multivariate nature of the data offering some advantages over the classical methods: (i) by using the principal coordinates the main patterns of genetic variation among populations are summarized in just three combined variables; (ii) the graphical representations permit not only global exploration of the main patterns and the variables associated to the discrimination, but also the direction of the association and the selection of small subsets of SNPs that have a similar behavior in relation to the discrimination; (iii) it is possible to study the correlation structure among alleles; and (iv) it is possible to know the population structure without any prior knowledge about the parental information.

Paschou *et al.* (2007, 2008) used a multivariate approach (PCA) to infer population structure using data from the HapMap project; however, although they indicate that the algorithm can be used to identify a small set of structure informative markers, they do not use the Biplot properties of the Singular Value Decomposition to interpret the SNPs responsible for the discrimination. Additionally they use a linear technique for continuous data to a categorical data matrix in which the elements coded as +1, 0 or −1 is questionable. When the data are already genotyped, and therefore it is categorical, PCA is not suitable. Our approach is a generalization of the PCA method and the Singular Value Decomposition (Biplot) that can handle the genotyped data in an appropriate way.

5 FINAL REMARKS

In summary, the proposed methodology using a combination of PCoA, CA and ELB represents an improvement over traditional methods for classifying genotypes using DNA molecular markers, since it produces groups, calculates a measure of the quality of the groups, identifies the alleles or bands that are responsible for the classification of genotypes (allowing the study of individual–individual, individual–variable and variable–variable relations more appropriately), and facilitates the genetic interpretation of the results.

The complementarity nature between PCoA, CA and LB yields a holistic comprehension of the data structure and facilitates the interpretations of the results. Consequently, a combined use of this set of techniques is highly recommended for a thorough description of data in studies of genetic diversity using DNA markers.

ACKNOWLEDGEMENTS

We are thankful to Prof John Gower by the critical reading and the suggestions made about this article.

Funding: Proyecto de Biotechnología BID-FONACIT II, Venezuela.

Conflict of Interest: none declared.

REFERENCES

Avise,J.C. (2004) *Molecular Markers, Natural History and Evolution*, 2nd edn. Sinauer Associates, 684 pp.

Chae,S.S. and Warde,W.D. (2006) Effect of using principal coordinates and principal components on retrieval of clusters. *Comput. Stat. Data Analysis*, **50**, 1407–1417.

Chapman,S. *et al.* (2002) Using biplots to interpret gene expression patterns in plants. *Bioinformatics*, **18**, 202–204.

Demey, *et al.* (2006) *Classifying genotypes using Molecular Markers: A Biplot Methodogy Apoach*. In XXIIIrd International Biometric Conference, Montreal, Québec, Canada. 16–21 June 2006, THP3.326.

Falguerolles,A.de. (1998) Log-bilinear biplots in action. In J. Blasius and M. Greenacre (eds) *Visualization of Categorical Data*. Academic Press, 594 pp.

Gabriel,K.R. (1971) The biplot - graphic display of matrices with applications to principal component analysis. *Biometrika*, **58**, 453–467.

Gabriel,K.R. and Zamir,S. (1979) Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics*, **21**, 489–498.

Gabriel,K.R. (1998) Generalised bilinear regression. *Biometrika*, **85**, 689–700.

Gower,J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

Gower,J.C. and Hand,D. (1996) *Biplots*. Chapman & Hall, 280 pp.

Heoa,M. and Gabriel,K.B. (2001) The fit of graphical displays to patterns of expectations. *Comput. Stat. Data Analysis*, **36**, 47–67.

Jongman,R.H.G. *et al.* (1995) *Data analysis in Community and Landscape Ecology*. Cambridge University Press, 321 pp.

Long,J.S. (1997) *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, 328 pp.

Mardia,K.V. *et al*. (1979) *Multivariate analysis*. Academic Press, 521 pp.

Paschou, *et al.* (2007) PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.*, **3**, e160.

Paschou, *et al*. (2008) Tracing sub-structure in the European American population with PCA-informative markers. *PLoS Genet.*, **4**, e1000114.

Rosenberg, *et al*. (2003) Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.*, **73**, 1402–1422.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.

Sharov,A.A. *et al.* (2005) A web-based tool for principal component and significance analysis of microarray data. *Bioinformatics*, **21**, 2548–2549.

Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical Taxonomy*. Freeman, 573 pp.

Tenenhaus,M. and Young,F.W. (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, **50**, 91–119.

The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.

The MathWorks Inc. (2008) *MATLAB Programming*. Natick, USA.

van Eeuwijk,F.A. (1995a) Multiplicative interaction in generalized linear models. *Biometrics*, **51**, 1017–1032.

van Eeuwijk,F.A. (1995b) Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. *Euphytica*, **84**, 1–7.

Vicente-Villardón,J.L. *et al*. (2006) Logistic Biplots. In M. Greenacre and J. Blasius (eds) *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC, 608 pp.

Weir, *et al.* (2005) Measures of human population structure show heterogeneity among genomic regions. *Genome Res.*, **15**, 1468–1476.