Análisis Multivariados y Marcadores Moleculares

Luis Valenzuela Villa,
luis.valenz.v@gmail.com

Laboratorio de Genética, Facultad de Ciencias,
Laboratorio de BioMatemática y Ómica Integrativa, Facultad de Medicina,
Universidad de Chile.

30 de Agosto, 2017

Marcador genético: segmento de ADN con una ubicación física identificable (locus) en un cromosoma y cuya herencia genética se puede rastrear.

- Variabilidad

Elección del marcadores en función de la problemática de estudio.

- Expresión

Dominante, Codominante.

- Herencia

ADN nuclear.

ADN mitocondrial.

Análisis Multivariado de Marcadores Moleculares

Marcadores moleculares

Marcadores moleculares:

- Alozimas: Codominante, “neutro”, menor variación.
- Microsatelites: codominante, neutro, mayor variación.
- RAPD (Random Amplified Polimorphism DNA): Dominante (PA), neutro.
- AFLP (Amplified Fragment Length Polymorphism): Dominante.
- SNPs (Single Nucleotide Polimorphism)

Análisis Multivariado de Marcadores Moleculares

"Efficient genetic markers for population biology"

Table 1. Attributes of markers commonly used in molecular population biology[a]

	PCR assay	Single locus	Codominant	Allele genealogy feasible	No. loci readily available	Connectibility of data among studies	Rapid transfer to new taxa	Overall variability[g]
Mitochondrial (and chloroplast)								
Sequence	Yes	Yes	Yes[d]	Yes	Single	Direct	Yes	Low–high
RFLP	No, large	Yes	Yes[d]	Yes	Single	Direct	Yes	Low–moderate
Multilocus nuclear								
Mini- and/or micro-satellite 'fingerprints'	No, large	No	No	No	Many	Limited	Yes	High
RAPD[b]	Yes	No	No	No	Many	Limited	Yes	High
AFLP[b]	Yes	No	No	No	Many	Limited	Yes	High
rDNA[c]	Yes	No	No	No	Few	Limited	Yes	Moderate–high
Single-locus nuclear (single copy nuclear, scn)								
Allozymes	No, protein	Yes	Yes	Rarely	Moderate	Direct	Yes	Low–moderate
Minisatellites	Few	Yes	Yes	Rarely	Moderate	Indirect[e]	Few	High
Microsatellites	Yes	Yes	Yes	Yes	Many	Indirect[e]	Some	High
Anonymous scn	Yes	Yes	Yes	Yes	Many	Indirect[e]	No?[f]	Moderate?[f]
Specific scn	Yes	Yes	Yes	Yes	Moderate	Direct	Yes?[f]	Moderate?[f]
rDNA[c]	Yes	In effect	Yes	Yes	Few	Direct	Yes	Low–moderate

[a]More details in Boxes 2 and 3.
[b]Some RAPD (randomly amplified polymorphic DNA) and AFLP (amplified fragment length polymorphic DNA) bands can be converted to single-locus markers, in which case they behave like 'anonymous scn'or 'specific scn' categories.
[c]rDNA consists of tandem arrays of a few regions. In some taxa the arrays are effectively identical and regions act as single loci, but in some taxa there can be many different sequences within individuals, in which case rDNA acts more like a multilocus system.
[d]mtDNA and chloroplast DNA are haploid and show one of a range of alternative positive states, in contrast to dominant markers that are either present or absent.
[e]Data from these markers are indirectly, but meaningfully, connectible given adequate models of molecular evolution.
[f]Insufficient research effort has been put into these markers.
[g]Variability depends on variation per marker and number of markers obtained readily. The assessment here approximates the outcome of a typical marker system.

Sunnucks P. 2000.

Análisis Multivariado de Marcadores Moleculares

Análisis Multivariado

Análisis Multivariados:

- “Técnicas de reducción de dimensiones”
- “Ordenamientos en espacio reducido”
- “Métodos factoriales”

El propósito es:

- Resumir fuertemente un conjunto de datos en uno pequeño de variables sintéticas no correlacionadas.
- Provee una imagen simplificada pero significativa de información compleja que es imposible percibir.

Análisis Multivariado de Marcadores Moleculares

Métodos más comunes

Vistos:

- Clustering, PCA, KNN, LDA, QDA, SVM, CART, Random Forest…

Difieren en el tipo de datos de entrada y propósito estudio:

- Variables cuantitativas/binarias: Análisis de Componentes Principales (PCA).
- 2 variables categóricas: Análisis de Correspondencia (CA).
- >2 variables categóricas: Análisis de Correspondencia Multiple (MCA).
- Matrices de distancias euclidianas:

Análisis de Coordenadas Principales (PCoA) o

Escalamiento Metrico Multidimensional (MDS)

- y muchos más…

Análisis Multivariado de Marcadores Moleculares

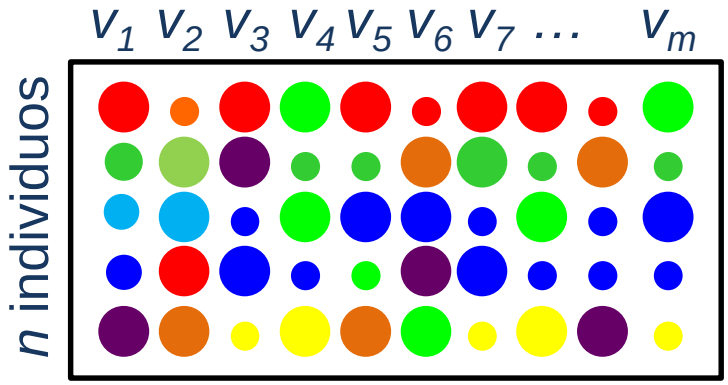
Paquete de R: ade4

Uno de los paquetes más completos de métodos exploratorios (del tipo *duality diagram*). Dray & Dufour, 2007

dudi.pca: *Principal Component Analysis*
dudi.coa: *Correspondence Analysis*
dudi.acm: *Multiple Correspondence Analysis*
dudi.fca: *Fuzzy Correspondence Analysis*
dudi.mix: *mixed analysis (numeric and factors)*
dudi.nsc: *Non Symetric Correspondence Analysis*
etc…

Análisis Multivariado de Marcadores Moleculares
Análisis de Componentes principales

m Variables



Matriz de datos

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{pmatrix}$$

Covarianzas empíricas

$$\sigma_{ij} = \frac{1}{n}\sum_{k=1}^{n}(X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j)$$

Matriz de Covarianzas empíricas

$$\Sigma_X = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{pmatrix}$$

Análisis Multivariado de Marcadores Moleculares

Análisis de Componentes principales

Planteamos nuevo modelo con variables sintéticas $y_i, \ldots, y_i$, que son combinaciones lineales de las variables originales $v_1, \ldots, v_p$

$$y_i = a_{1i}v_1 + a_{2i}v_2 + \cdots + a_{pi}v_p$$

Se eligen los coeficientes aij para maximizar la varianza empírica.

$$Var(y_i) = \sum_{i=1}^{p} a_{ki}^2 \sigma_{kk} + \sum_{k=1}^{p} \sum_{\substack{j=1 \\ j \neq k}}^{p} a_{ki} a_{ji} \sigma_{kj} = \vec{a}_i^t \Sigma_X \vec{a}_i$$

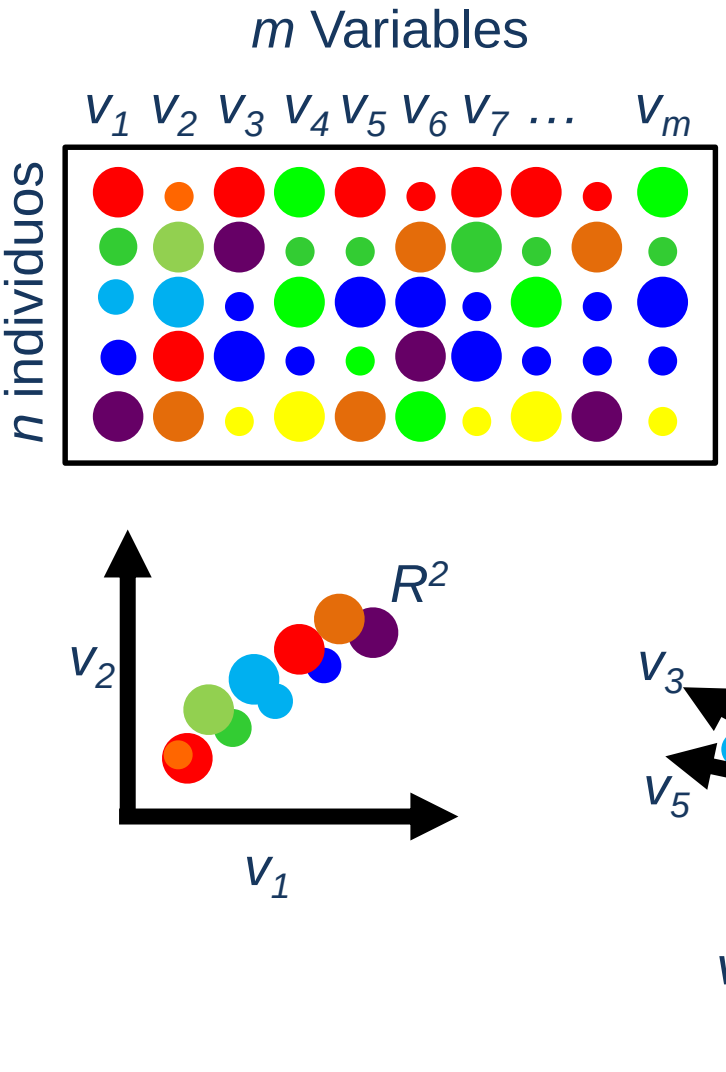
Y que además cumpla que las Covarianzas sean igual a 0.

$Var(y_i) = \lambda_i$ (Valores propios de la matriz de covarianza)

$Cov(y_j, y_k) = 0$

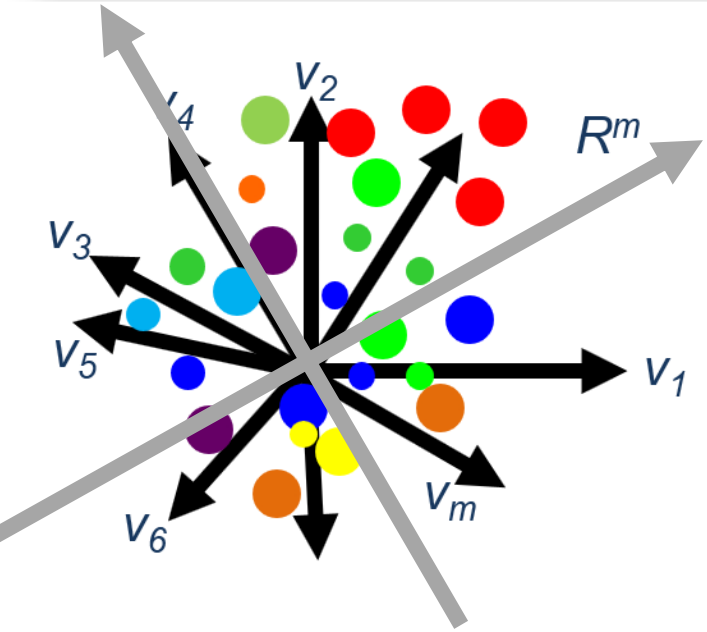
Análisis Multivariado de Marcadores Moleculares

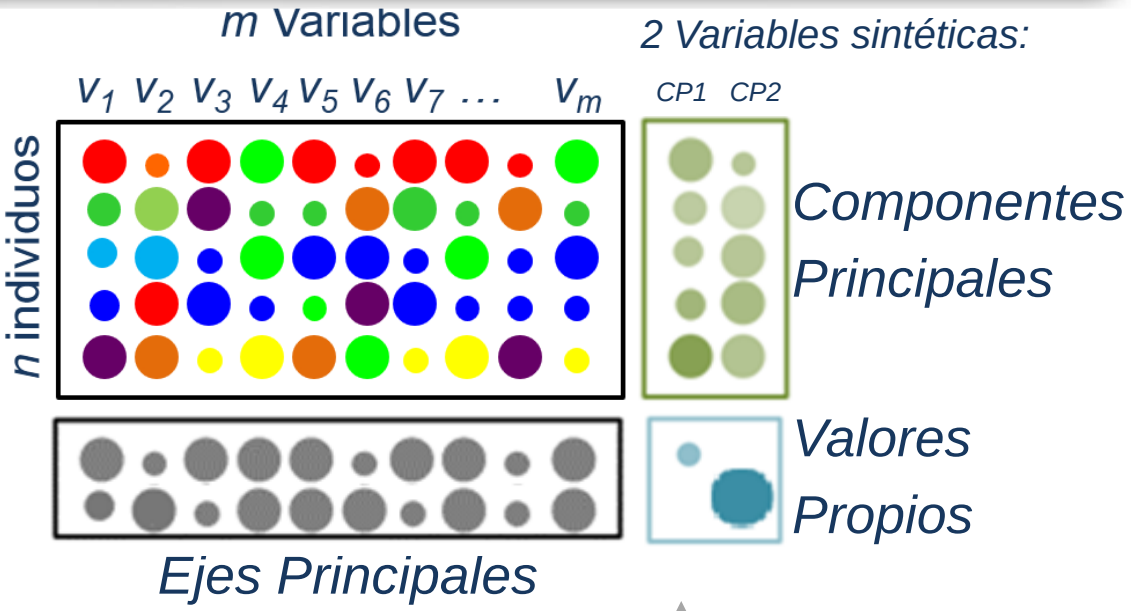
Análisis de Componentes principales



Análisis Multivariado de Marcadores Moleculares

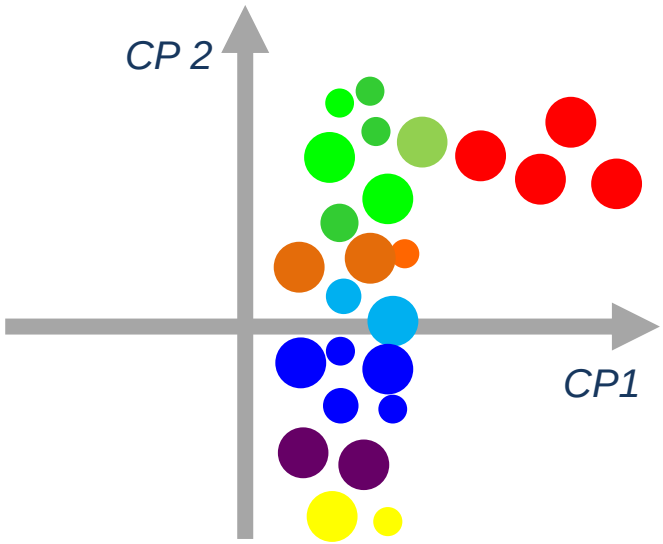
Análisis de Componentes principales





$$CP_i = a_{1i}v_1 + a_{2i}v_2 + \ + a_{mi}v_m$$

$$Cov(CP_j; CP_k) = 0$$



Análisis Multivariado de Marcadores Moleculares

PCA con ade4: dudi.pca

```
pcadatos <- dudi.pca(datos, center=TRUE, scale=FALSE)
pcadatos$
help(dudi.pca)
```

Objeto	Significado
tab	Tabla de datos transformados
cw	Pesos de las columnas
lw	Pesos de las filas
eig	*Eigenvalues* no nulos
rank	Número de *Eigenvalues* no nulos
c1	*Scores* de columnas normadas, i.e., ejes principales.
l1	*Scores* de filas normadas.
co	*Coordenadas de columnas*
li	*Coordenadas de las filas, i.e, componentes principales*

Análisis Multivariado de Marcadores Moleculares

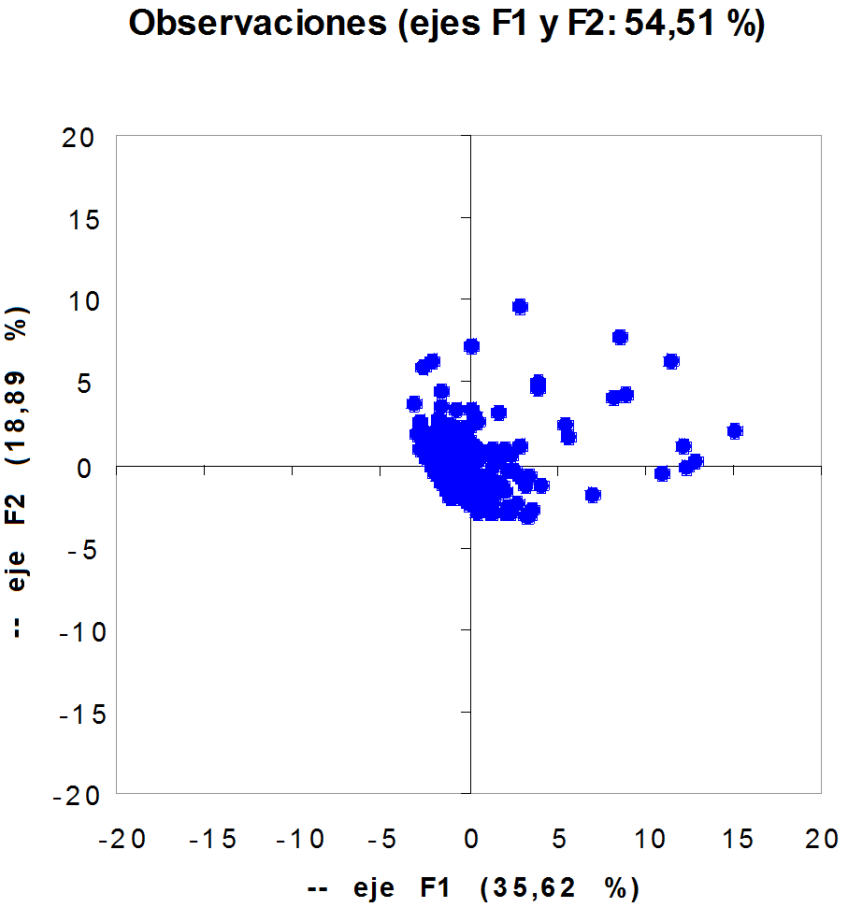
PCA con ade4: dudi.pca

Algunos gráficos:

- Scatterplot: Observaciones.

Matriz de datos

$$X = \begin{matrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{matrix}$$

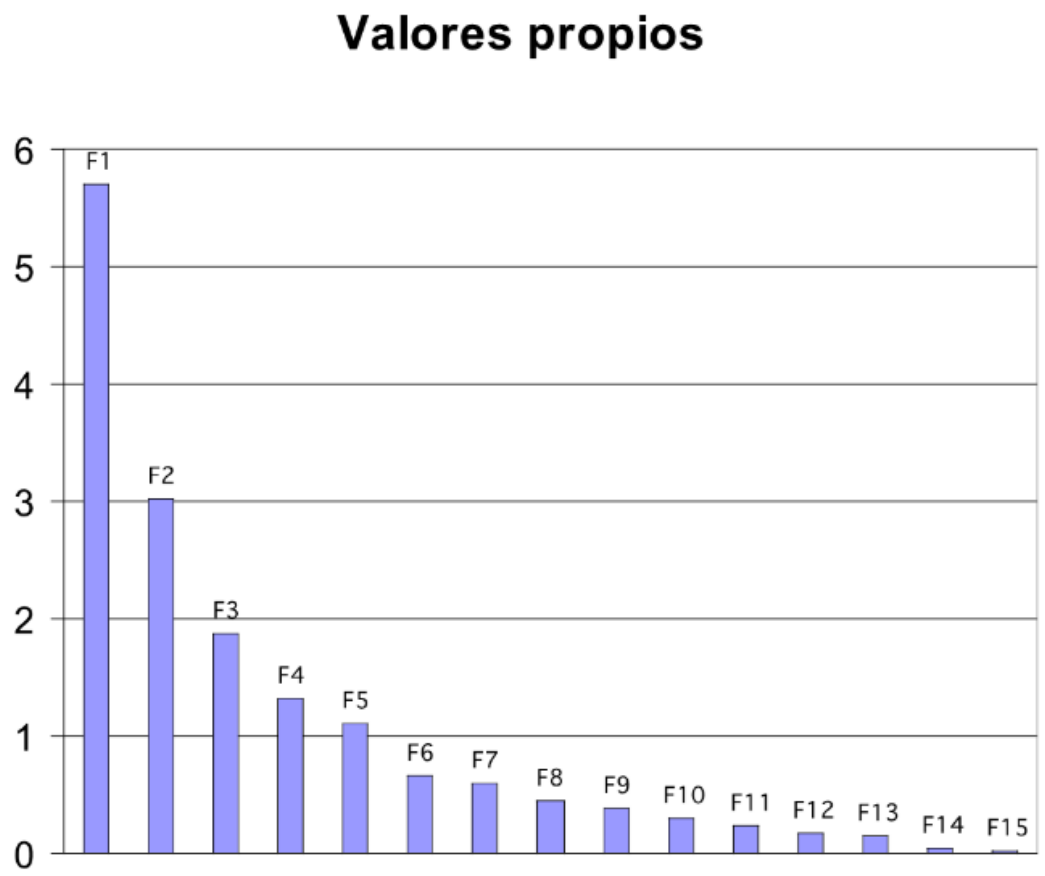


Análisis Multivariado de Marcadores Moleculares

PCA con ade4: dudi.pca

- Screeplot: barra de valores propios (*eigenvalues*).

¿Con cuántos quedarse?



Valores propios de la matriz de covarianza

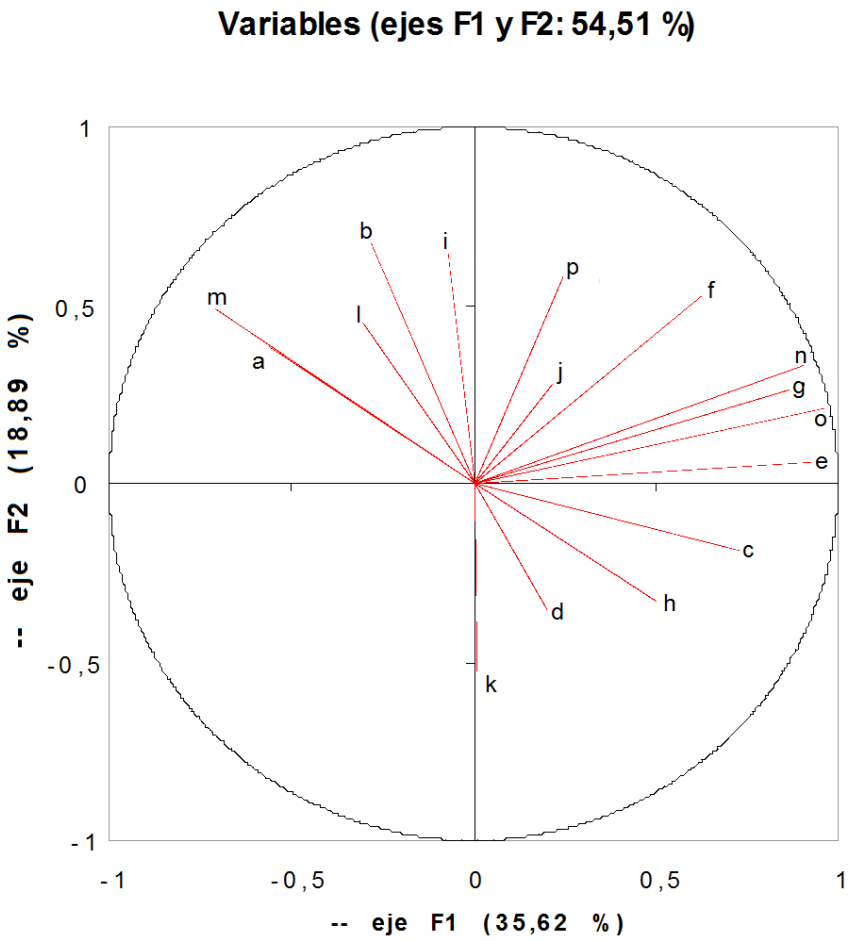
$$Var(y_i) = \lambda_i$$

F_1: λ_1 de CP_1
F_2: λ_2 de CP_2
F_3: λ_3 de CP_3
…
F_{15}: λ_{15} de CP_{15}

Análisis Multivariado de Marcadores Moleculares

PCA con ade4: dudi.pca

- Círculo de correlaciones (coordenadas de variables).



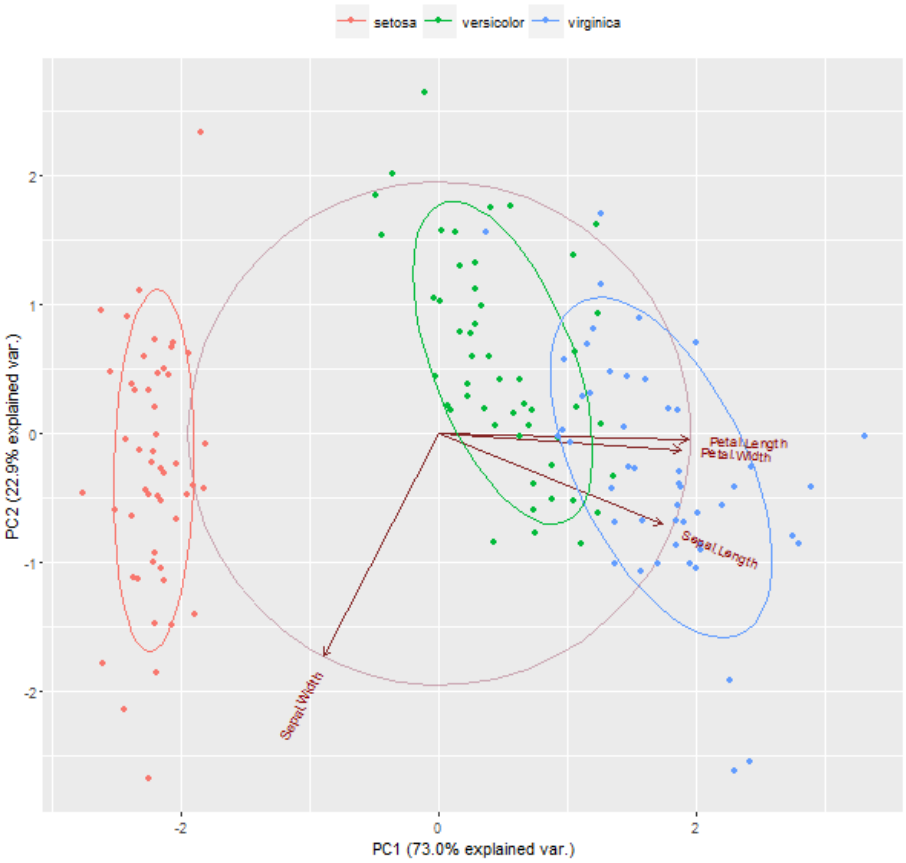
Se modifica X restando a cada valor en la j-ésima columna el promedio de dicha columna y dividiendo por $\sqrt{\sigma_{jj}}$

$$\tilde{X}_{ij} = (X_{ij} - \bar{X}_j) / \sqrt{\sigma_{jj}}$$

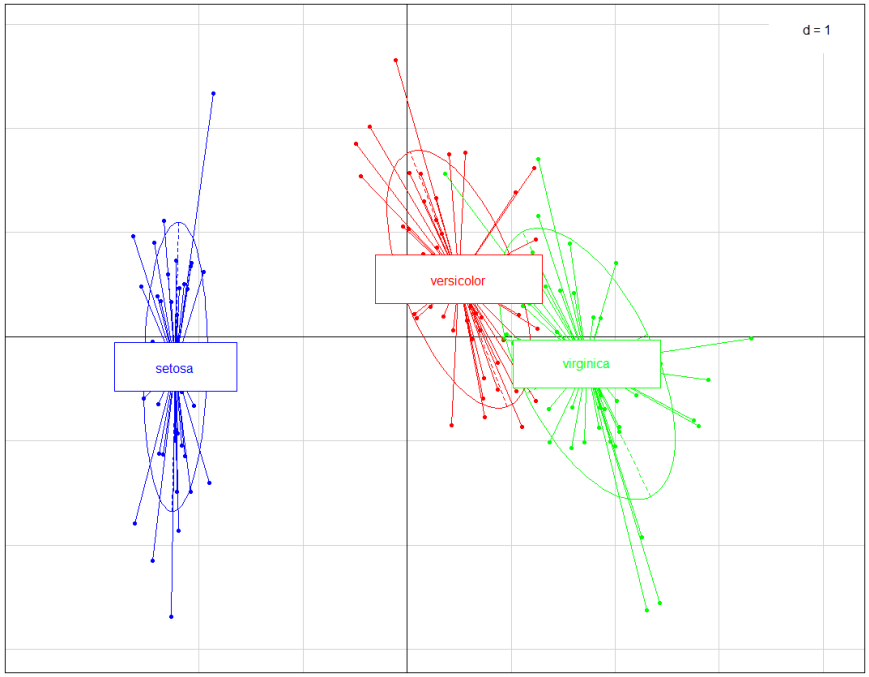
Análisis Multivariado de Marcadores Moleculares

PCA con ade4: dudi.pca

ir.pca <-prcomp(ir,center = TRUE,scale. = TRUE)



ir.pca <-dudi.pca(ir,center = TRUE,scale = TRUE)

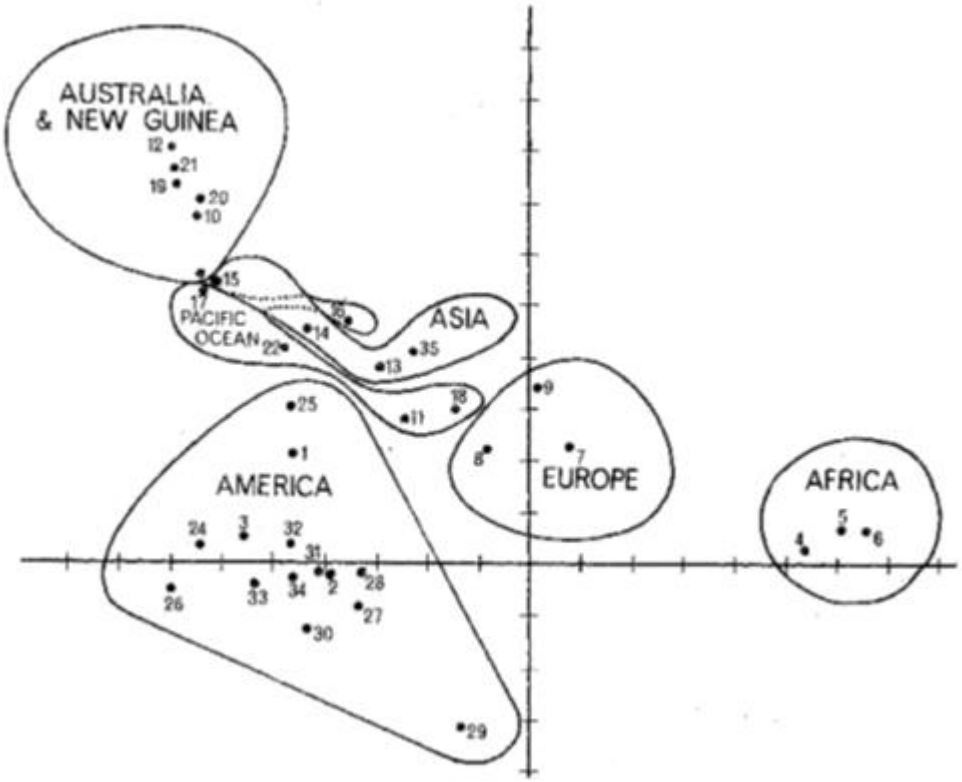


Análisis Multivariado de Marcadores Moleculares

Primera aplicación de PCA en datos genéticos (alozimas)

Alozimas de poblaciones nativas humanas (Cavalli-Sforza, 1966).

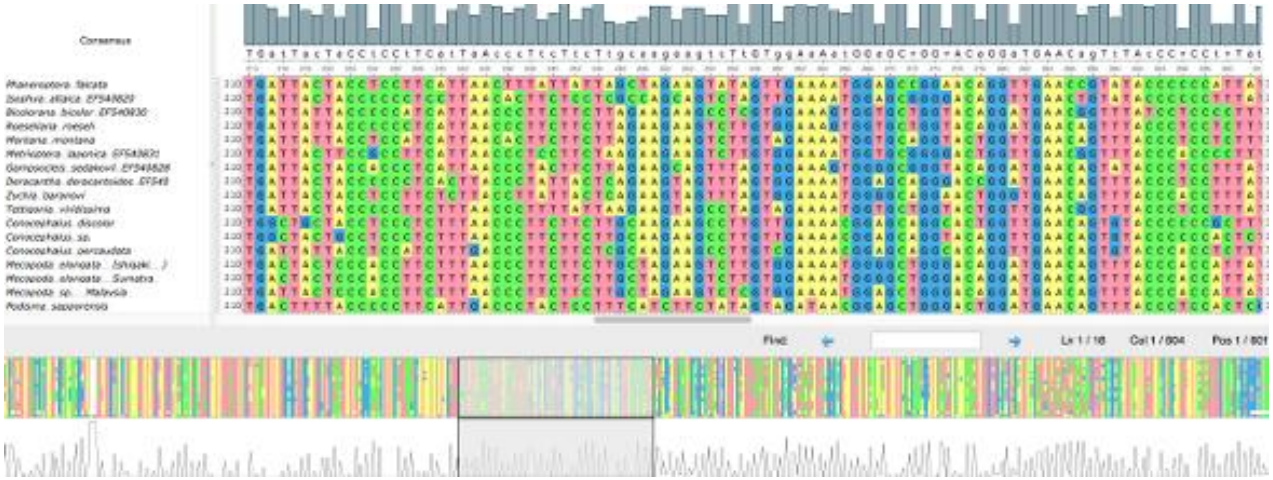
Componentes principales separan poblaciones por continentes.



Análisis Multivariado de Marcadores Moleculares

PCA: aplicación en datos genéticos

Secuencias de ADN son una fuente rica de información de la dinámica espacio-temporal de poblaciones biológicas.



Ejemplo:

Muchos (cientos, miles de) individuos-secuencias (observaciones).

Muchos (cientos, miles, millones de) SNPs (variables)

Análisis Multivariado se usa para resumir la diversidad genética.

Análisis Multivariado de Marcadores Moleculares

Asegurar reproducibilidad

1. Citar referencias.
2. Abreviaciones correctas.
3. Explicitar las transformaciones iniciales de los datos.
4. Gráficos.

Temas específicos por métodos:

1. Análisis de Discriminante
2. PCA
    - Interpretación de estructuras genéticas
    - Escalamiento de los datos

Análisis Multivariado de Marcadores Moleculares

Ejemplos:

Table 1 Multivariate analyses applied to genetic markers

Method	*Criterion*	*Application*	*Data*
Principal component analysis (PCA)	Variance (same as squared Euclidean distances)	Cavalli-Sforza (1966)	Allozymes
Principal coordinates analysis (PCoA)	Any Euclidean distance	Sanchez-Mazas and Langaney (1988)	Allozymes
Non-metric dimensional scaling (NMDS)	Ordering of objects	Lessa (1990)	Roger's and Nei's distances
Correspondence analysis (CA)	χ^2 distance	She *et al.* (1987)	Allozymes
Discriminant analysis (DA)	Variance between groups/total variance	Smouse *et al.* (1982)	Allozymes
Constant-row total multiple correspondence analysis (CRT-MCA)	Correlation ratio	Guinand (1996)	Allozymes
Factor analysis (FA)	'Common effect' in allele frequencies	Taylor and Mitton (1974)	Allozymes
Canonical correspondence analysis (CCA)	χ^2 distances in predicted data	Angers *et al.* (1999)	Microsatellites
Redundancy analysis (RDA)	Variance of predicted data	Kölliker *et al.* (2008)	AFLP and SSR
Canonical correlation analysis (CCorA)	Squared correlation between pairs of scores	Johnson and Schaffer (1973)	Allozymes
Co-inertia analysis (COA)	Squared covariance between pairs of scores	Jarraud *et al.* (2002)	AFLP
Multiple co-inertia analysis (MCOA)	Squared covariance between a set of scores	Laloë *et al.* (2007)	Microsatellites
Spatial principal component analysis (sPCA)	Product of variance and spatial autocorrelation	Jombart *et al.* (2008)	Microsatellites

Abbreviations: AFLP, amplified fragment length polymorphism; SSR, single sequence repeats.
Each method is indicated by its most frequent name and abbreviation. The 'criterion' is the quantity optimized by the principal components of the method. The 'application' column gives the reference of an early and representative publication using the method to analyse genetic markers.

Análisis Multivariado de Marcadores Moleculares

PCA: Diferentes funciones para optimizar

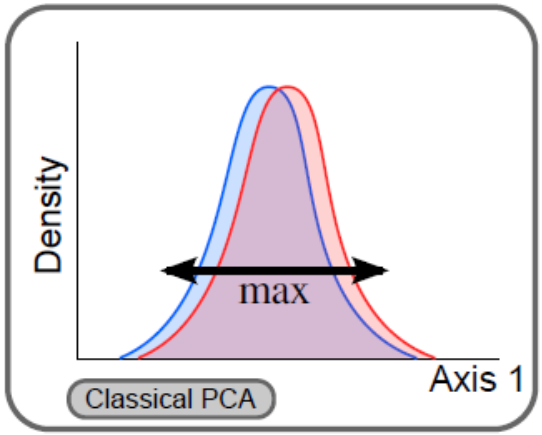
Varianza total = (Varianza entre grupos) + (Varianza dentro de grupos)

PCA o PCoA/MDS: Varianza Total

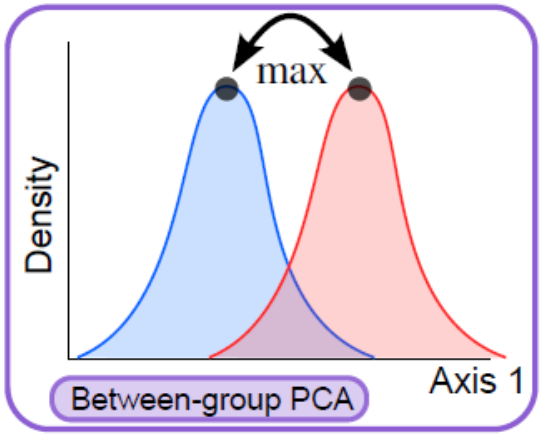
Between-group PCA: Varianza entre grupos

Within-group PCA: Varianza dentro de grupos

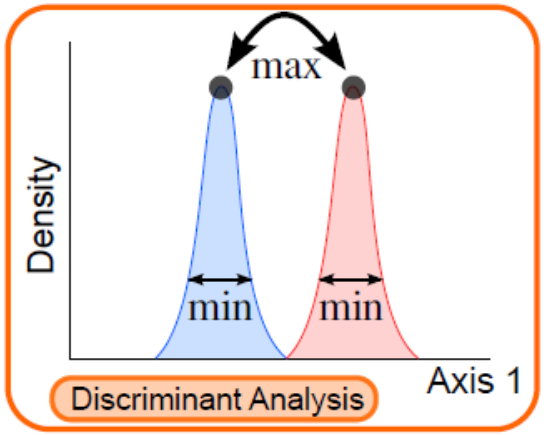
Análisis de discriminante: Varianza entre grupos/dentro de grupos



Max. diversidad total



Max. diversidad entre grupos



Max. separación de grupos

Análisis Multivariado de Marcadores Moleculares

Análisis Discriminante (DA)

DA:

Requiere menos variables (alelos) que observaciones (individuos).

Requiere variables no correlacionadas.

Solución (DAPC):

Reducción previa de las variables con PCA, y luego DA.

DAPC (Jombart et al. 2010))

PCA sobre los datos.

Requiere de grupos a priori.

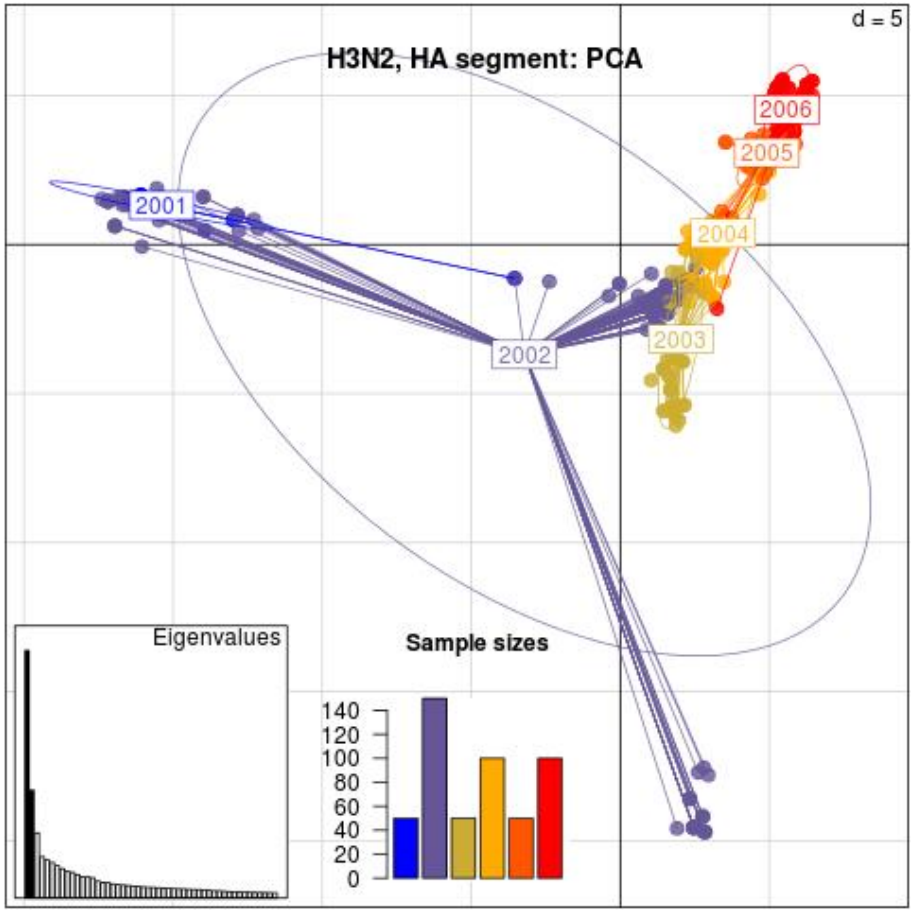
K-means que maximice la varianza entre grupos.

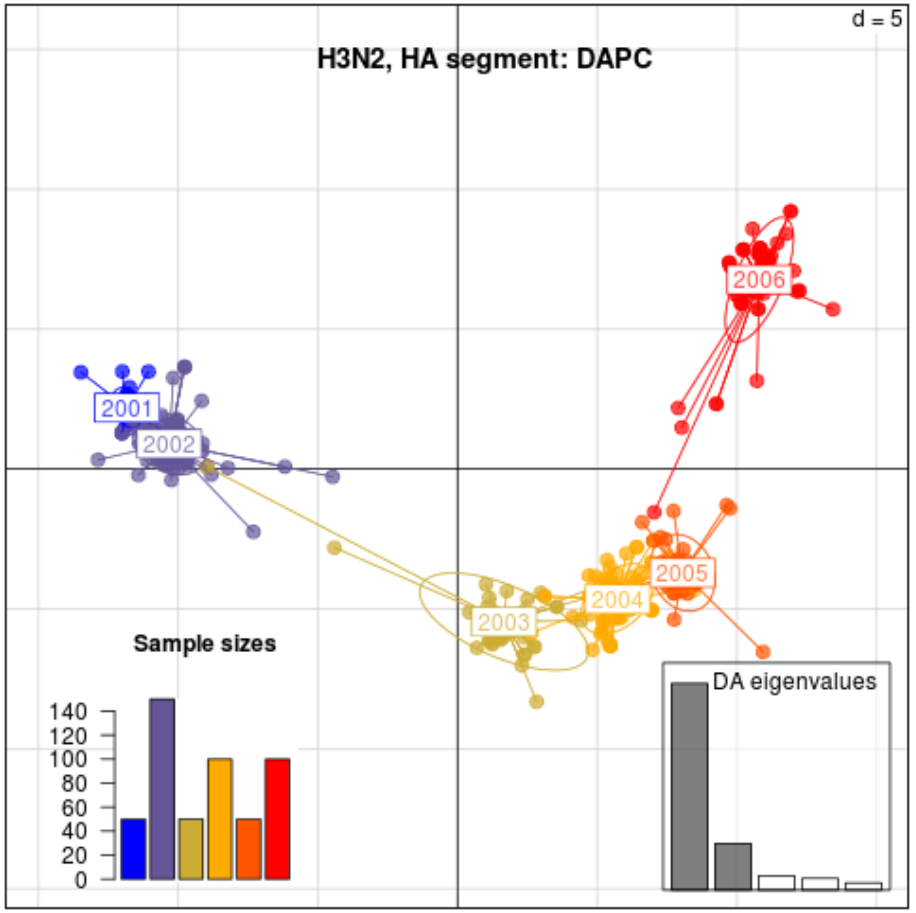
Criterio para identificar número óptimo: BIC.

Análisis de Discriminantes.

Análisis Multivariado de Marcadores Moleculares

DAPC





Análisis Multivariado de Marcadores Moleculares

Análisis Discriminante (DA)

I de Moran

$$I(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{L} \mathbf{x}}{n} \frac{1}{\mathrm{var}(\mathbf{x})}$$

- Autocorrelación espacial positiva corresponde a una mayor similaridad genética entre individuos cercanos geográficamente.

- Autocorrelación espacial negativa corresponde a una mayor diferencia genética entre individuos cercanos geográficamente.

Prueba de Mantel:

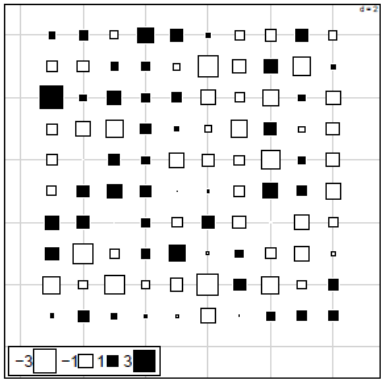
Ho: No hay correlación

$$r_{\mathrm{M}} = \frac{1}{d-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left(\frac{x_{ij} - \bar{x}}{s_x}\right)\left(\frac{y_{ij} - \bar{y}}{s_y}\right)$$

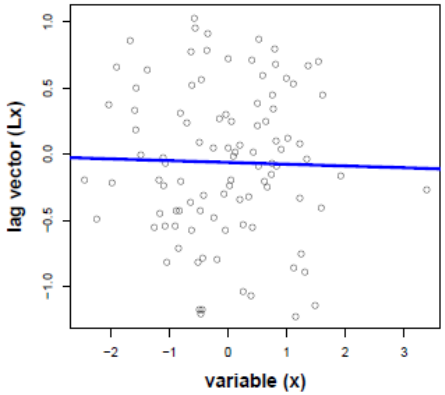
Análisis Multivariado de Marcadores Moleculares

Análisis Discriminante (DA)

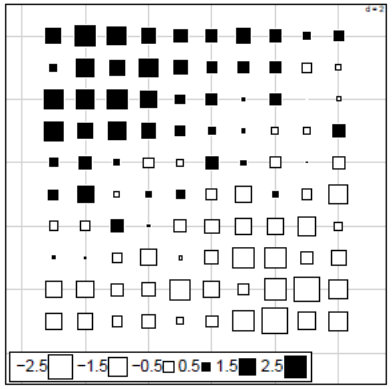
Distribucíón al azar

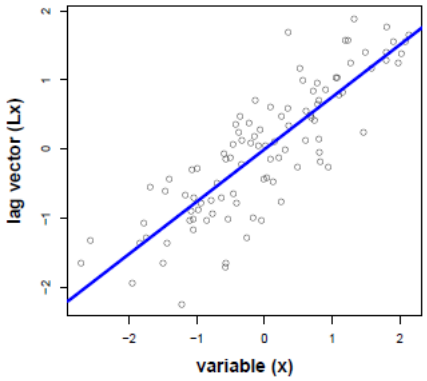


Lag vector (valor prom de vecinos de la variable)

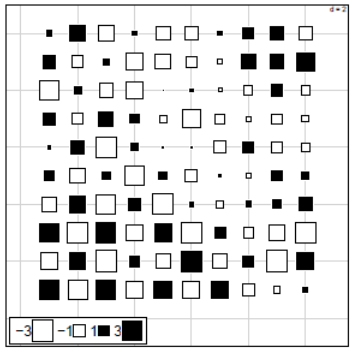


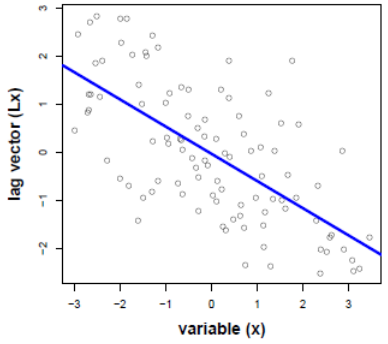
Autocorrelacion +





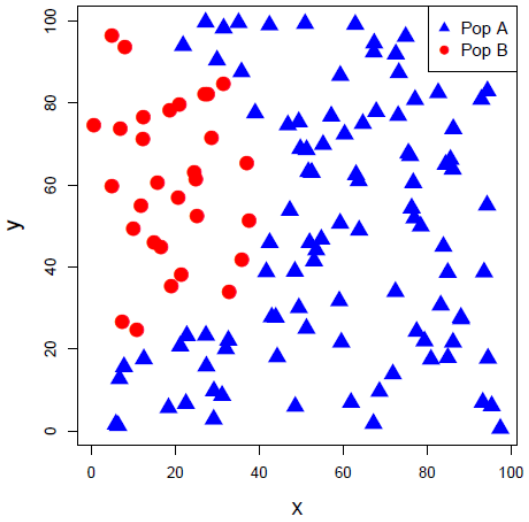
Autocorrelacion -



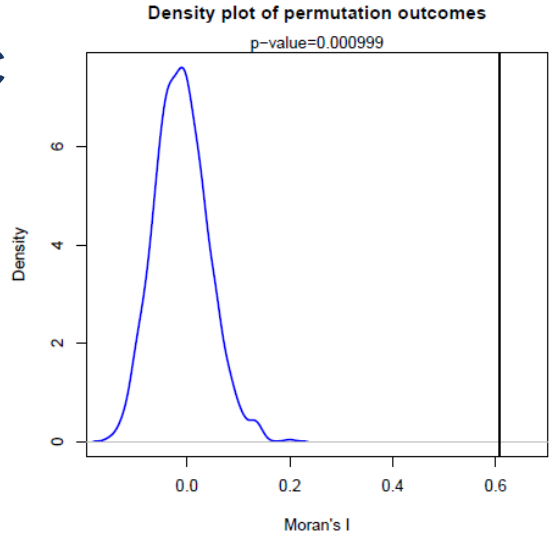


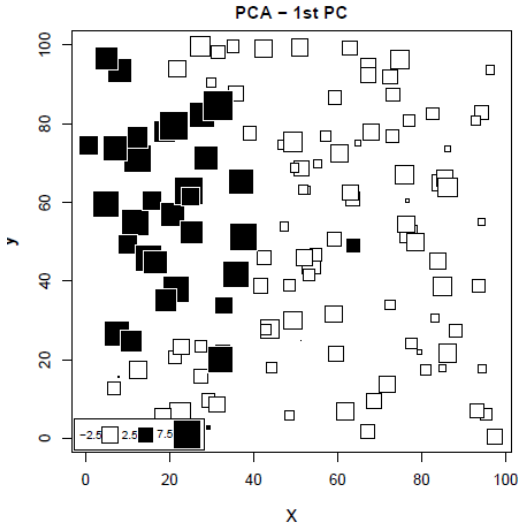
Análisis Multivariado de Marcadores Moleculares

Análisis Discriminante (DA)

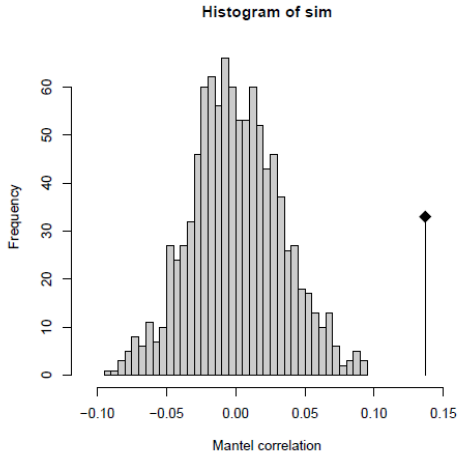


I de Moran del PC





Mantel test del PC1



Análisis Multivariado de Marcadores Moleculares

Análisis Discriminante (DA)

sPCA: Se desarrolló para investigar patrones genéticos espaciales ocultos o no obvios.

sPCA descompone: (varianza total) x (Moran's I)

Diferencias con PCA.

Eigenvalues + y – debido a que se optimiza la autocorrelación.

- Autocorrelación espacial positiva corresponde a una mayor similaridad genética entre individuos cercanos geográficamente.

- Autocorrelación espacial negativa corresponde a una mayor diferencia genética entre individuos cercanos geográficamente.

Análisis Multivariado de Marcadores Moleculares
Análisis Discriminante (DA)