CLASE 12 - ANÁLISIS MULTIVARIANTE Y EXPRESIÓN DE GENES.

Curso Análisis de expresión diferencial de genes e investigación reproducible.

Dr. José Gallardo Matus | https://genomics.pucv.cl/

Pontificia Universidad Católica de Valparaíso

10 November 2022

PLAN DE LA CLASE

1.- Introducción

- ¿Qué son los análisis multivariantes?.
- Estudio de caso: Biomarcadores de genotoxicidad.
- Etapas para realizar análisis multivariante.
- Matriz de distancia euclideana y análisis de cluster jerárquico.
- Análisis de componentes principales.
- PERMANOVA

2). Práctica con R y Rstudio cloud.

- Elaborar análisis multivariante con R.

INTRODUCCIÓN ANÁLISIS MULTIVARIANTE

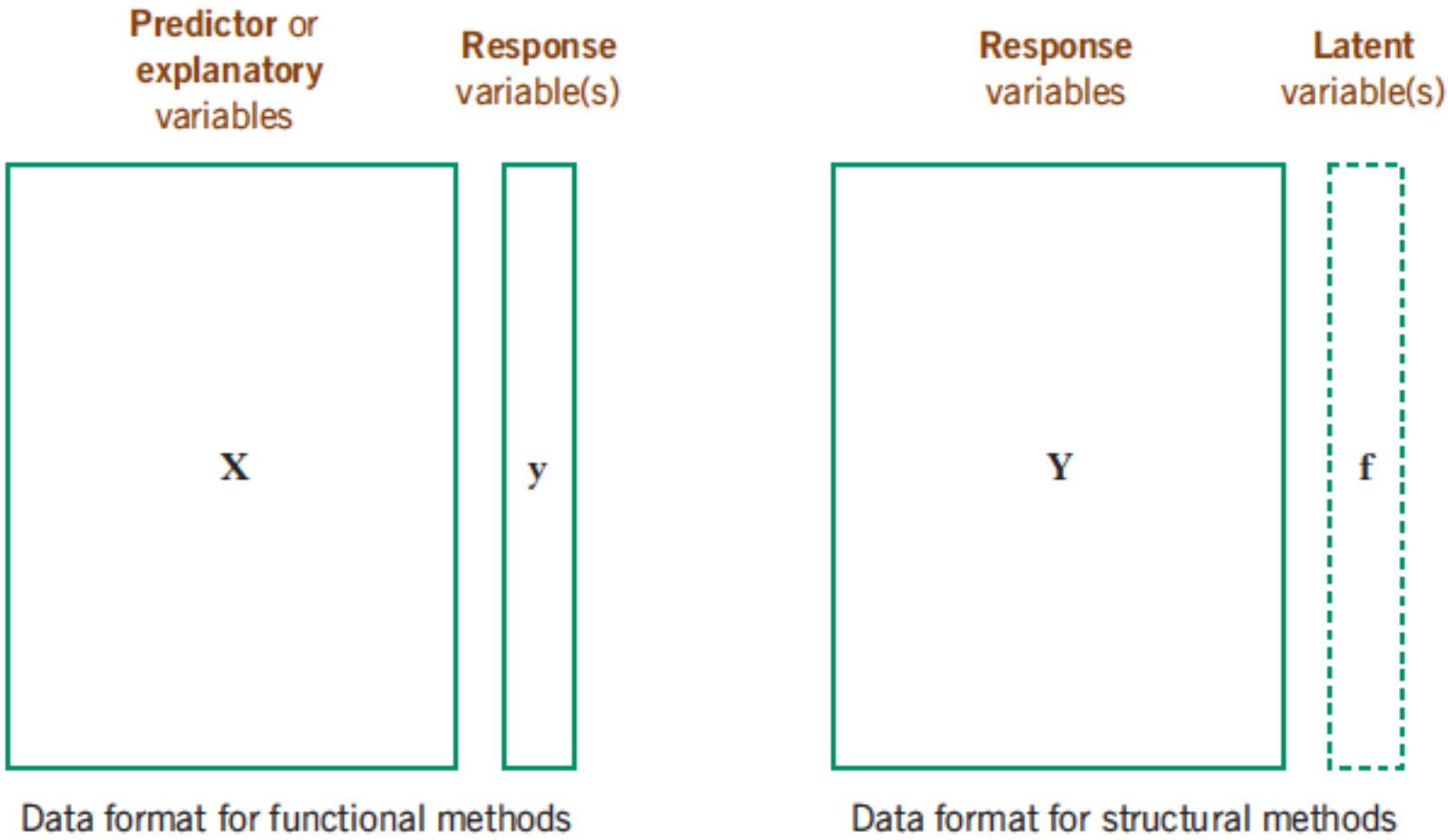
¿Qué son los análisis multivariantes?

Conjunto diverso de métodos estadísticos que observan y estudian el comportamiento simultáneo de múltiples variables.

Table 1: Ejemplo de Ct para multiples genes y muestras.

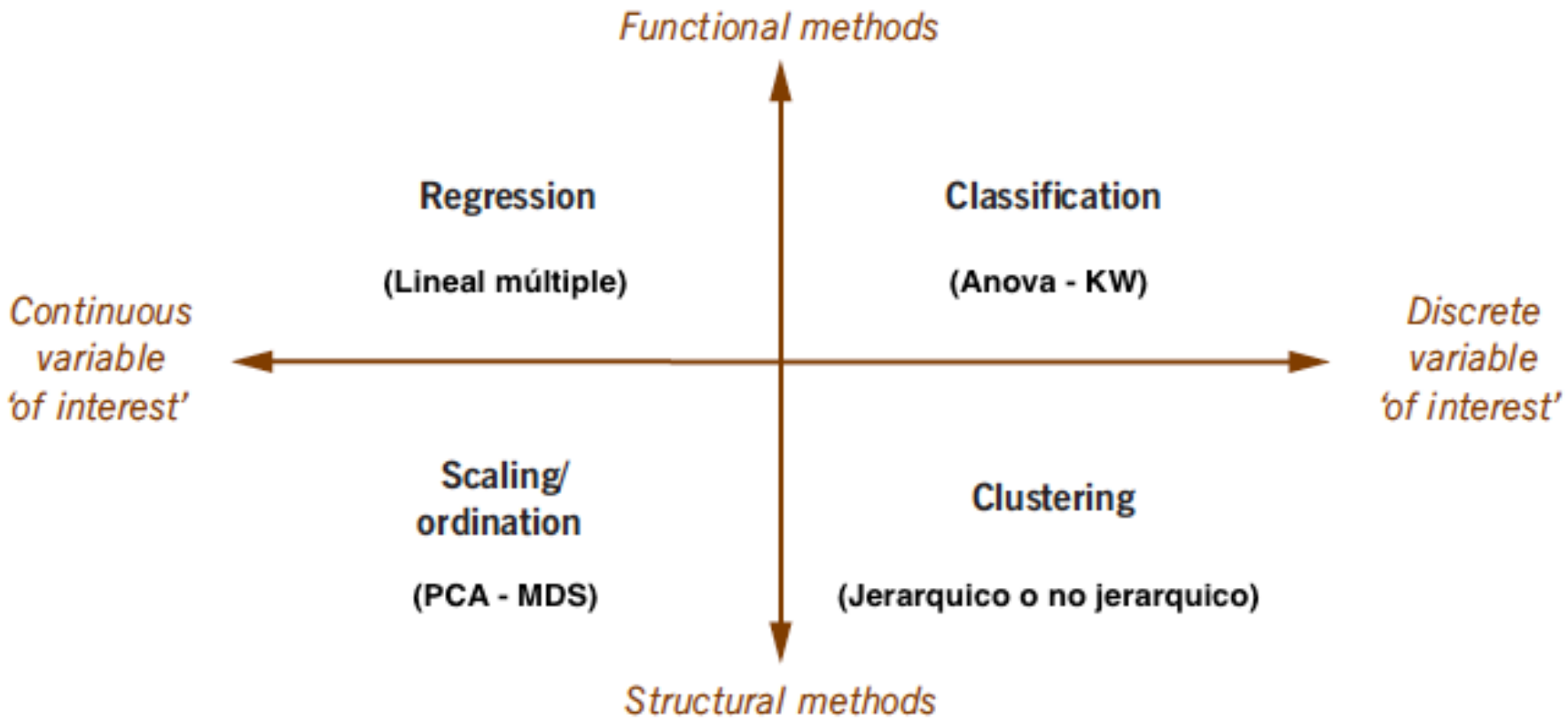
	ACTA2	AEN	B3GNT2	BLOC1S2	BRMS1L
SAMPLE 1	27.9	25.5	22.9	25.2	27.7
SAMPLE 2	27.6	25.4	24.9	25.0	28.2
SAMPLE 3	27.7	24.9	25.2	24.6	28.4
SAMPLE 4	26.4	24.1	24.5	24.0	27.1
SAMPLE 5	26.2	23.3	24.7	23.2	26.7
SAMPLE 6	26.4	23.5	24.4	23.6	27.2

TIPOS DE MÉTODOS MULTIVARIANTES



Fuente: Multivariate Statistic, 2014

MÉTODOS MULTIVARIANTES SEGÚN TIPO DE VARIABLE



Fuente: Multivariate Statistic, 2014

ESTUDIO DE CASO: BIOMARCADORES DE GENOTOXICIDAD.

Cho et al. 2018: Validación de un panel de expresión de genes como biomarcadores de genotoxicidad en modelo *in vitro*.

- Objetivo: Reducir el uso de animales en estudios de genotixicidad.
- Células TK6 se exponen a 14 agentes que inducen daños en el ADN (DDI) y a 14 agentes que no indicen daño (non-DDI) por 4 h.

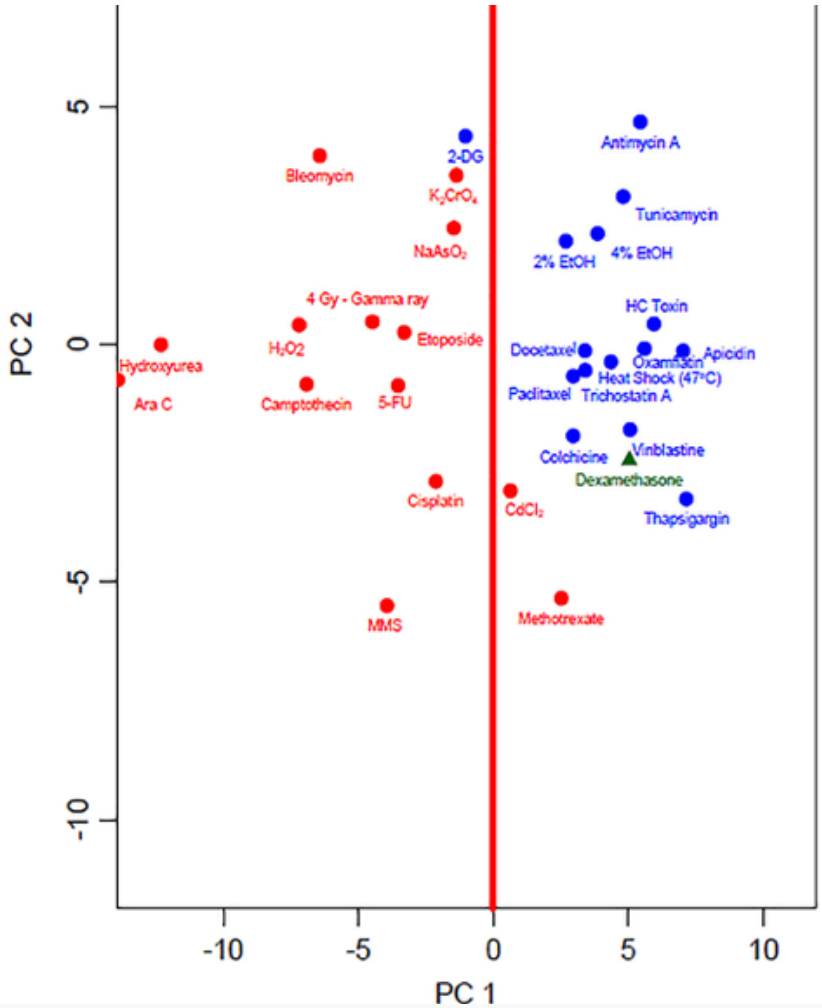
Tipo de agente	Ejemplos
DDI	Gamma irradiation; Cadmium chloride; Potassium chromate; Hydrogen peroxide
non-DDI	Colchicine; Heat shock (47°C); Ethanol; Glycolysis inhibitor

ESTUDIO DE CASO: MAPA DE CALOR LOG10(FOLD CHANGE)

- 61 genes fueron evaluados (DDI:inductor de daño en el ADN; Non-DDI.

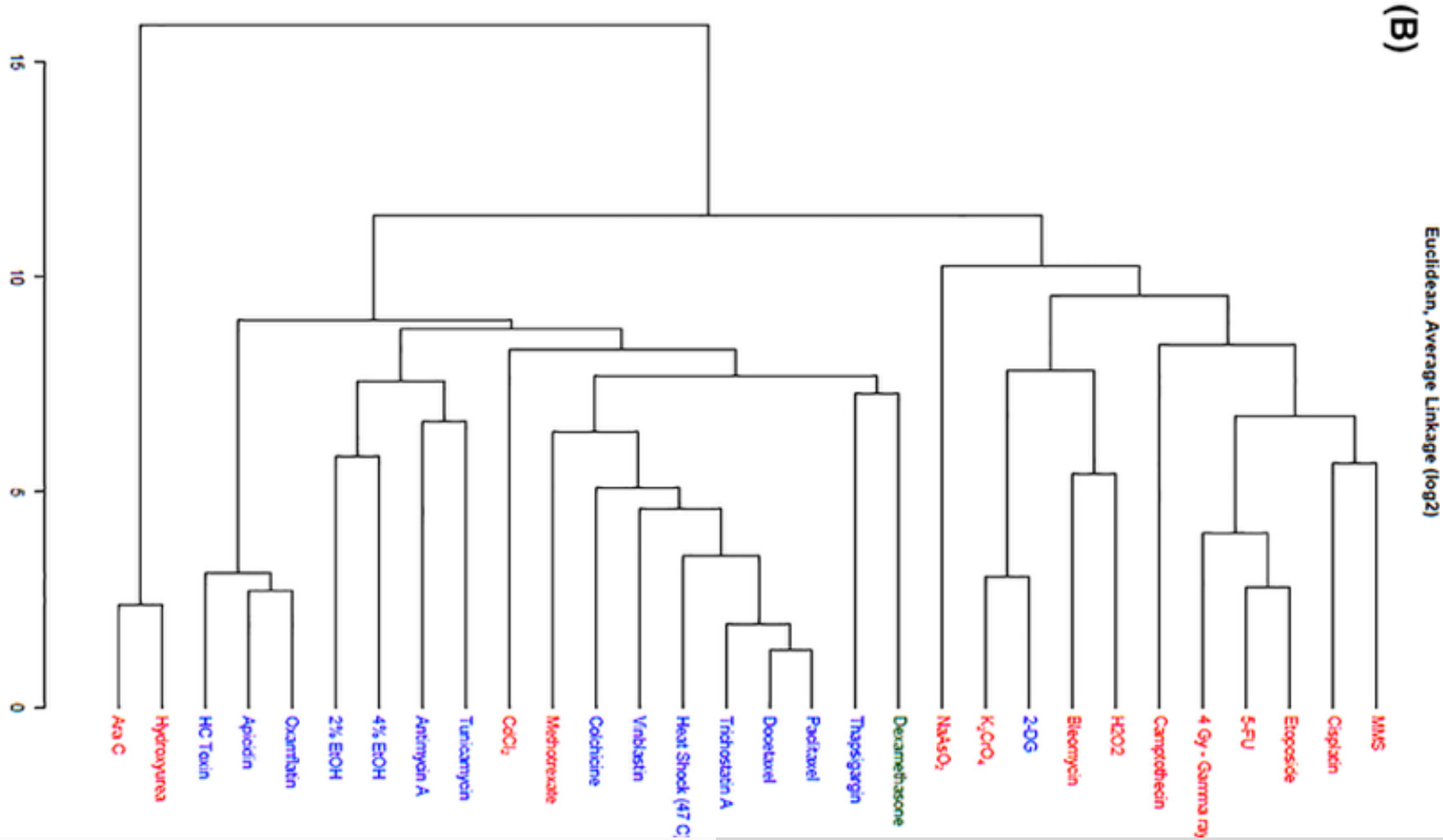
ESTUDIO DE CASO: ACP

PCA permite clasificar agentes (DDI o non DDI) en función de la expresión global de genes (excepciones: 2-DG, Cdcl2 y Methotrexate).



ESTUDIO DE CASO: CLUSTER JERÁRQUICO

Cluster jerárquico permite construir grupos jerarquicos por similitud de expresión de genes (excepciones: 2-DG, Cdcl2 y Methotrexate).



ANÁLISIS MULTIVARIANTE PARA ESTUDIO DE CASO

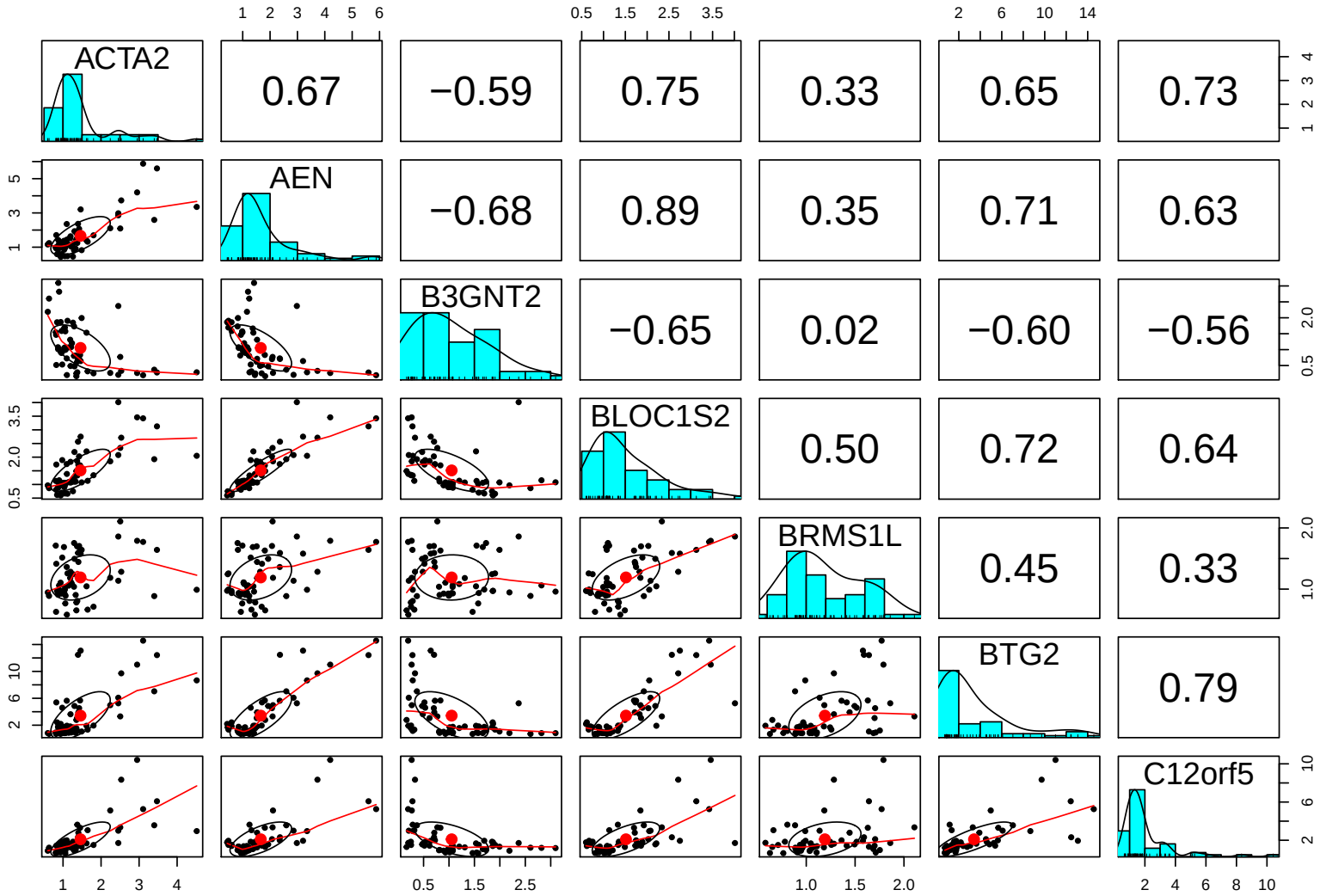
1) Explorar correlación de expresión de genes (Ct o FC). Si no están correlacionadas el PCA no es de utilidad.
2) Explorar matriz de distancia (euclideana) con log(FC).
3) Elaborar e interpretar Cluster jerárquico.
4) Elaborar e interpretar PCA.
5) Elaborar e interpretar PERMANOVA.

FC Y GENOTOXICIDAD

Table 3: FC multiples genes y muestras.

	ACTA2	AEN	B3GNT2	BLOC1S2	BRMS1L
D2	1.38	1.25	0.29	1.33	0.81
D3	1.10	1.44	0.20	1.39	0.58
D4	2.24	2.11	0.27	1.85	1.19
D5	2.95	4.20	0.27	3.46	1.79
D6	2.53	3.73	0.33	2.71	1.28
nD7	1.65	1.32	0.32	1.13	0.66

CORRELACIÓN FC: 7 GENES



MATRIZ DE DISTANCIA O SIMILARIDAD

¿Qué es y para que sirven?

- Las matrices de distancia o similaridad están en la base de todos los análisis multivariados.

Algunas consideraciones

- Las matrices de distancia se pueden elaborar tanto para variables cuantitativas continuas, como discretas.

  - Debido a que las variables pueden tener diferente escala o magnitud es necesario muchas veces transformar o estandarizar las variables antes de calcular las matrices de distancia.

TIPOS DE MATRICES DE DISTANCIA

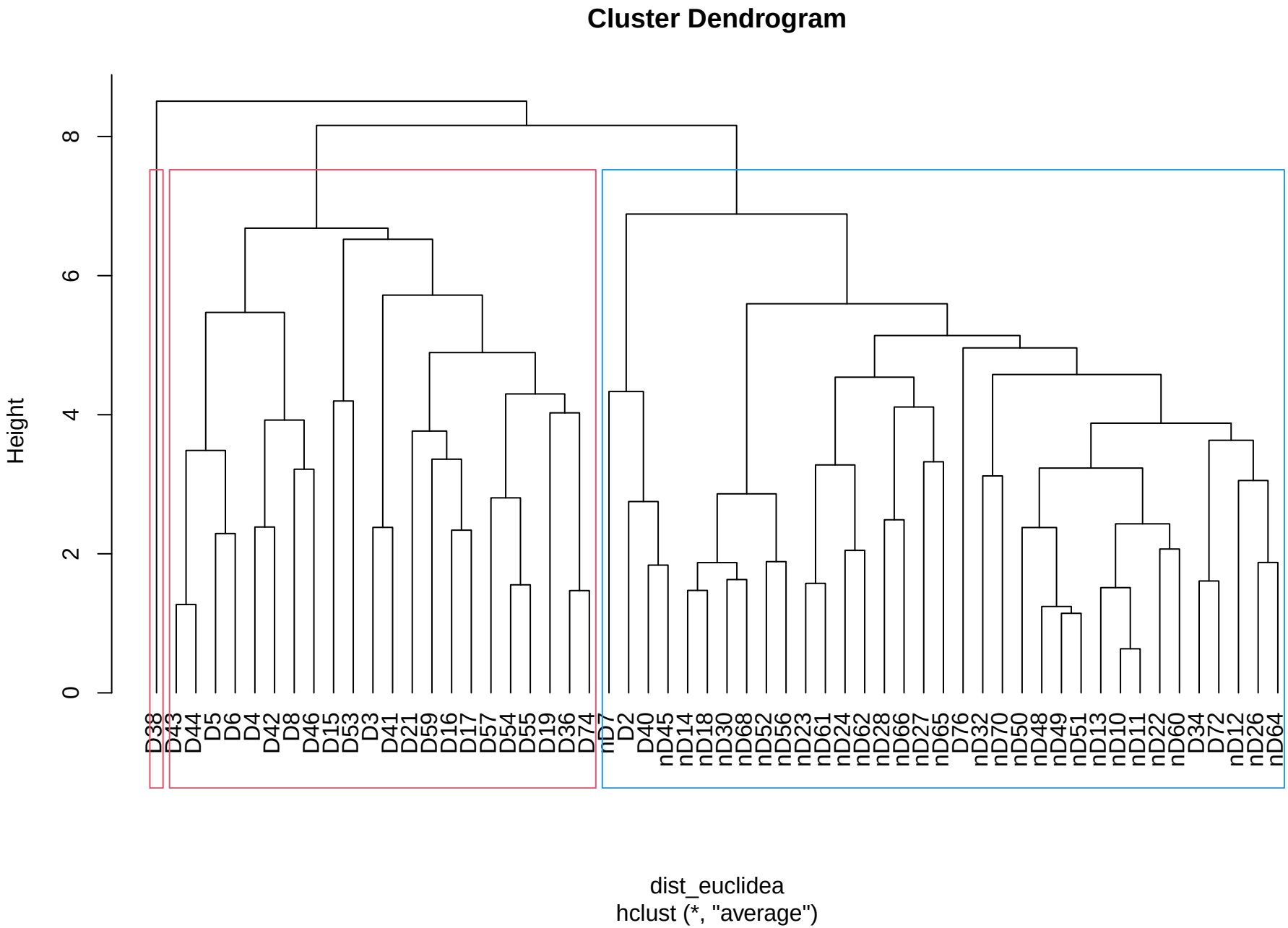
- **Euclideana**: Para variables cuantitativas continuas.
  Con base en el teorema de pitágoras
  $c^2 = a^2 + b^2$
  $a = \sqrt{c^2 - b^2}$
  $b = \sqrt{c^2 - a^2}$
  $c = \sqrt{c^2 + b^2}$
- **No euclideana**: Para variables cuantitativas discretas.

a) Bray-Curtis (datos de conteo).
b) Jacard (binarias)

MATRIZ DE DISTANCIA EUCLIDEANA ENTRE MUESTRAS

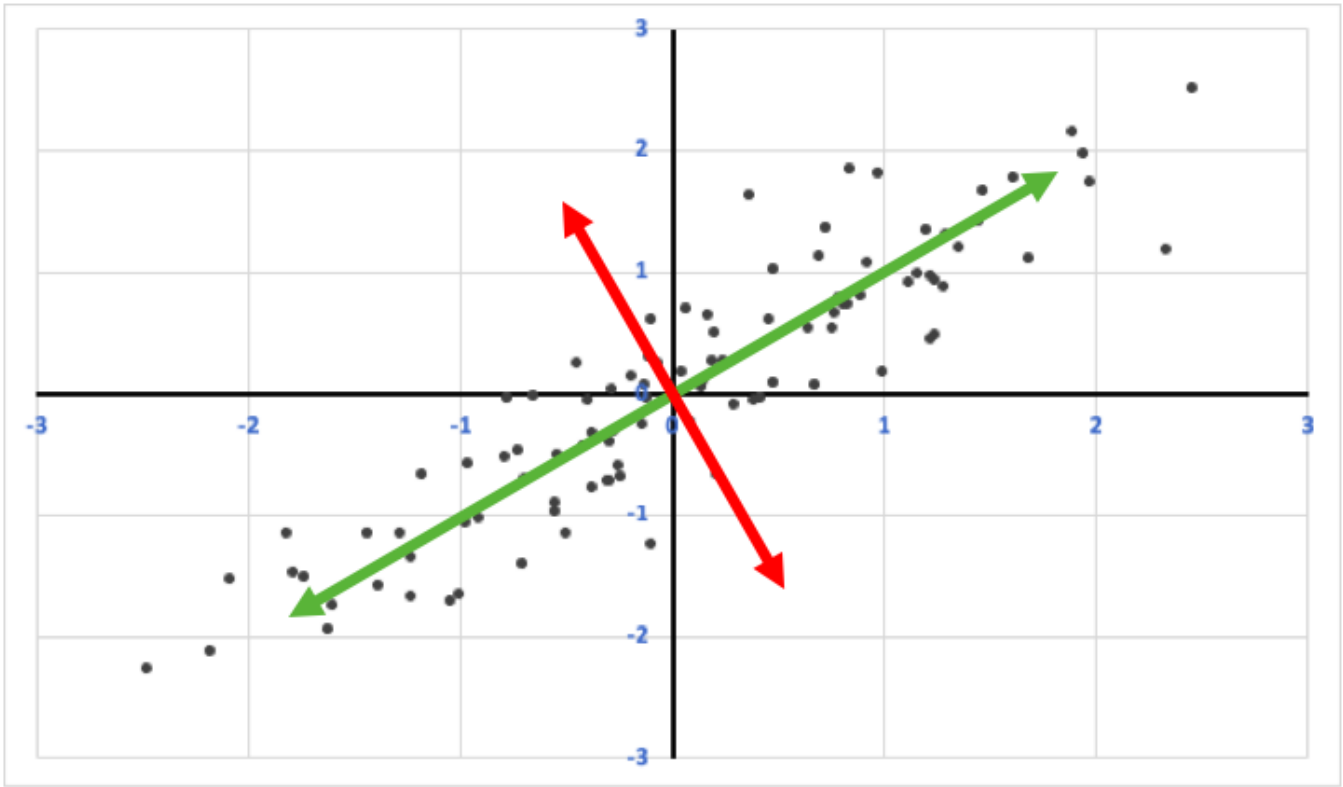
```
##                D2       D3       D4        D5
## D2    0.000000 5.931975 4.888891  8.956738
## D3    5.931975 0.000000 6.814921  9.805088
## D4    4.888891 6.814921 0.000000  6.344714
## D5    8.956738 9.805088 6.344714  0.000000
## D6    7.420962 8.148454 4.815898  2.290639
## nD7   3.285386 6.094782 5.208323 10.113096
## D8    5.735512 5.726718 3.458152  5.553001
## nD10 4.848999 4.810726 7.299157 10.330121
```

CLUSTER JERARQUIDO



COMPONENTES PRINCIPALES

- ¿Qué son los componentes principales (PC)?
  Combinación lineal de las variables originales no correlacionadas entre si (perpendiculares / ortogonales).
- Permite reducir la dimensionalidad de un conjunto de datos con muchas variables respuesta, sin perder mucha información



COMPONENTES PRINCIPALES: COMO SE CALCULAN

- Cada componente principal se obtiene por combinación lineal de las variables originales (FC).

Tabla de Datos

$$\begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \longrightarrow \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n1} & C_{n2} & \cdots & C_{np} \end{bmatrix}$$

Componentes

100% de la información

80% 16% 0.02%

Fuente: Rodriguez, 2009

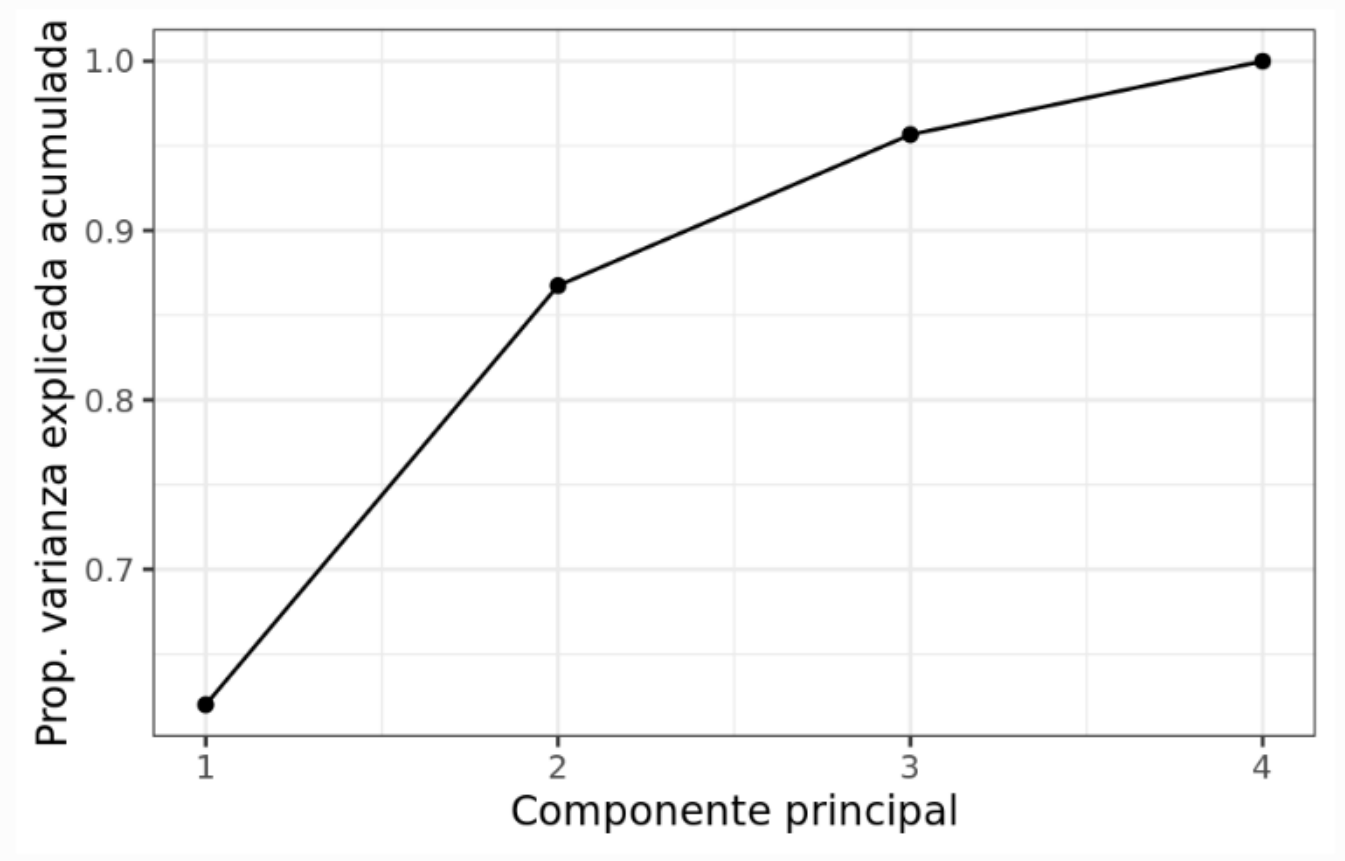
PCA: COMPONENTES PRINCIPALES

- Extracto matrix CP (eigenvector).

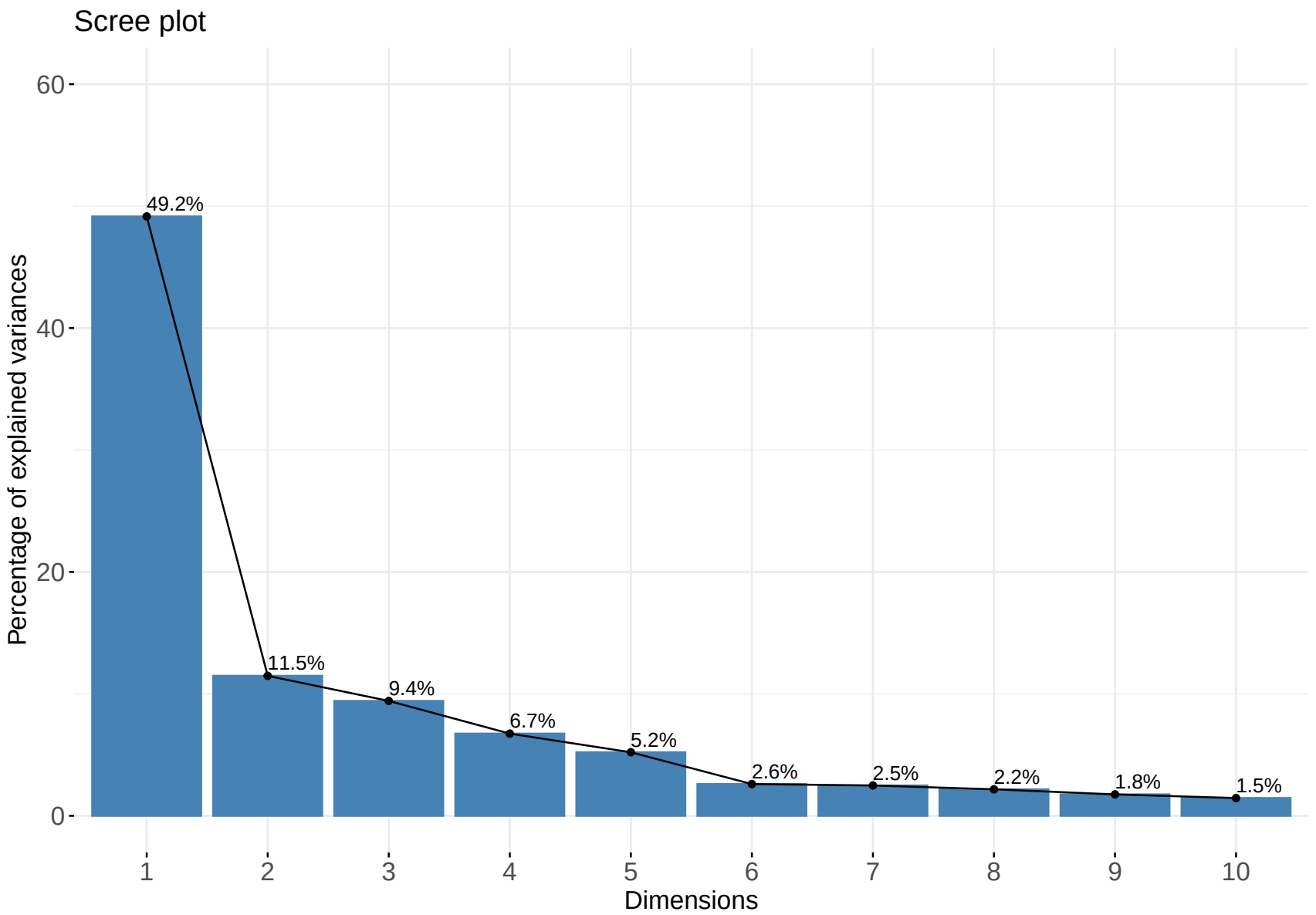
```
##                PC1          PC2         PC3
## ACTA2    -0.14961865  0.044953973 -0.02315913
## AEN      -0.16592638 -0.008884413 -0.02949912
## B3GNT2    0.13388846  0.085437650  0.23794488
## BLOC1S2  -0.16666856 -0.005713705  0.05834965
## BRMS1L   -0.08756488  0.042302329  0.29375674
## BTG2     -0.16050506 -0.033088233  0.09921226
## C12orf5  -0.15237221  0.110854448 -0.01130368
## CBLB      0.09134749  0.108602231 -0.06711369
```

VARIANZA EXPLICADA

Cada eigenvalue estima la varianza explicada por cada CP. Note que en este ejemplo los primeros dos componentes principales pueden capturar mucha de la varianza explicada por todas las variables analizadas.

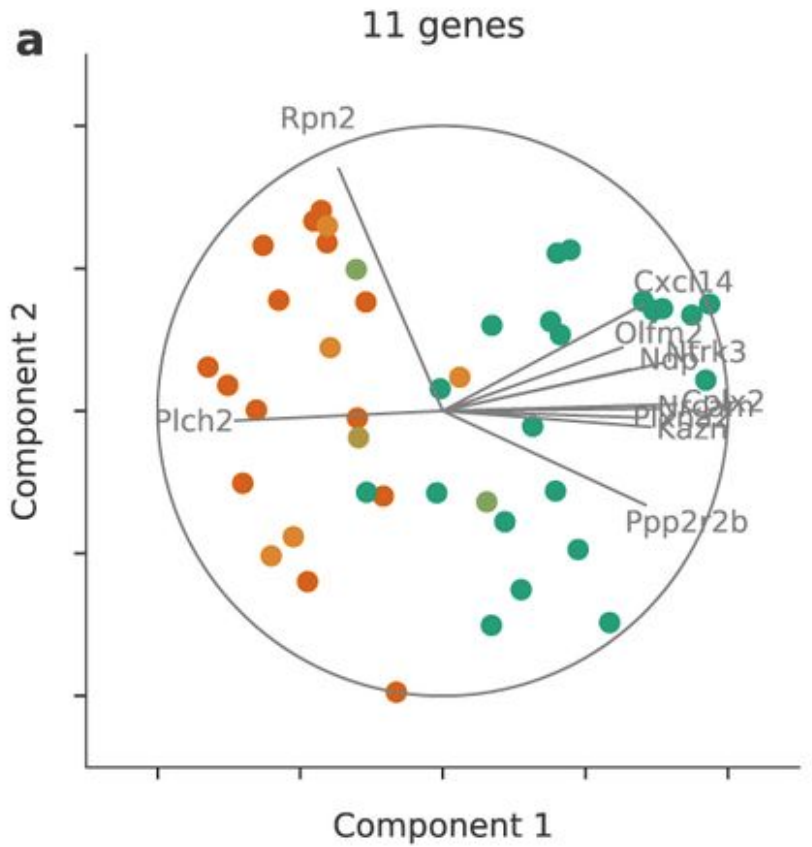


PCA: VARIANZA EXPLICADA



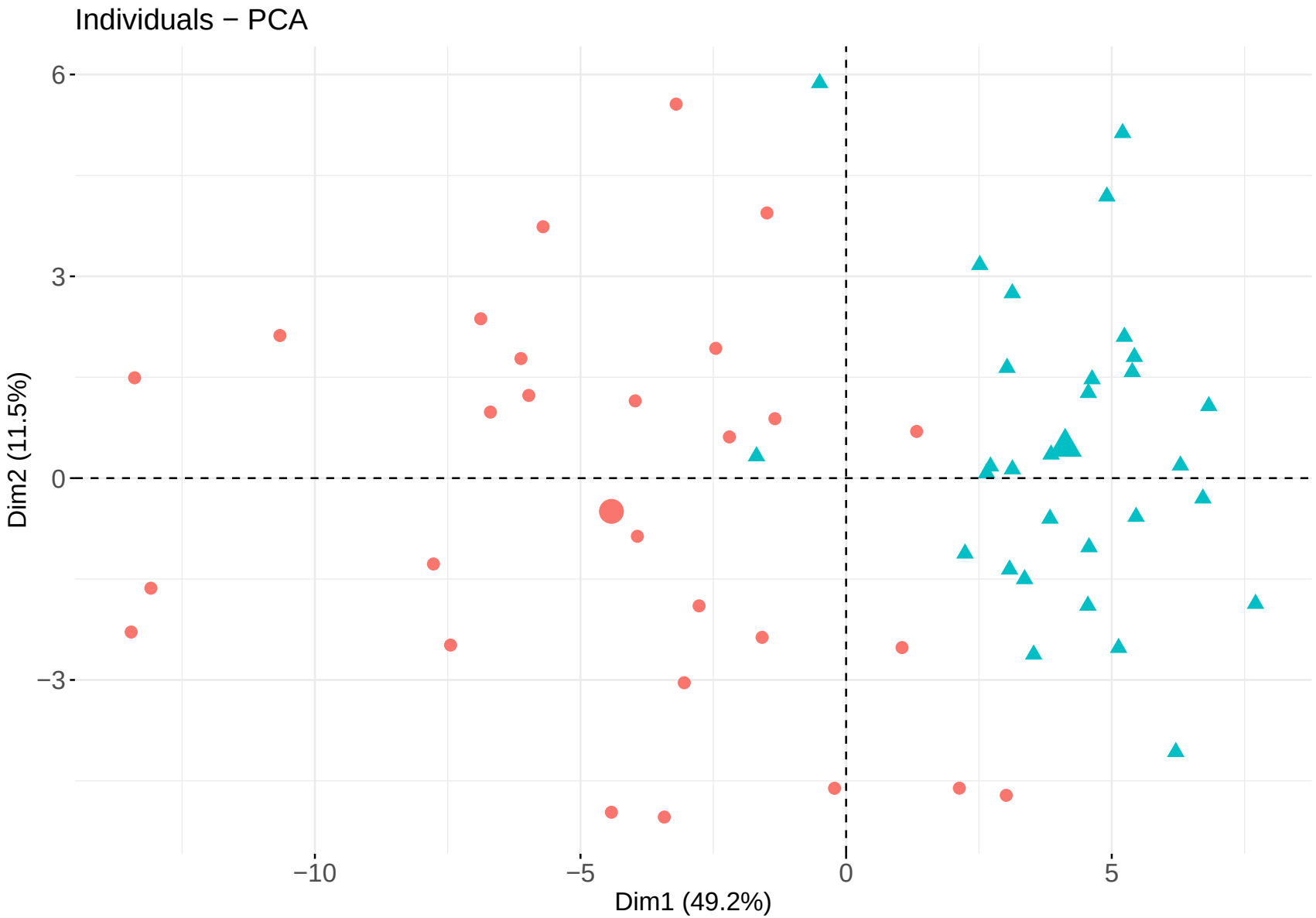
GRÁFICAS BI-PLOT

- 2 eigenvector o componentes principales para cada variable.
- Correlación de expresión de genes + observaciones (Color naranjo y verde = tipos de celulas).

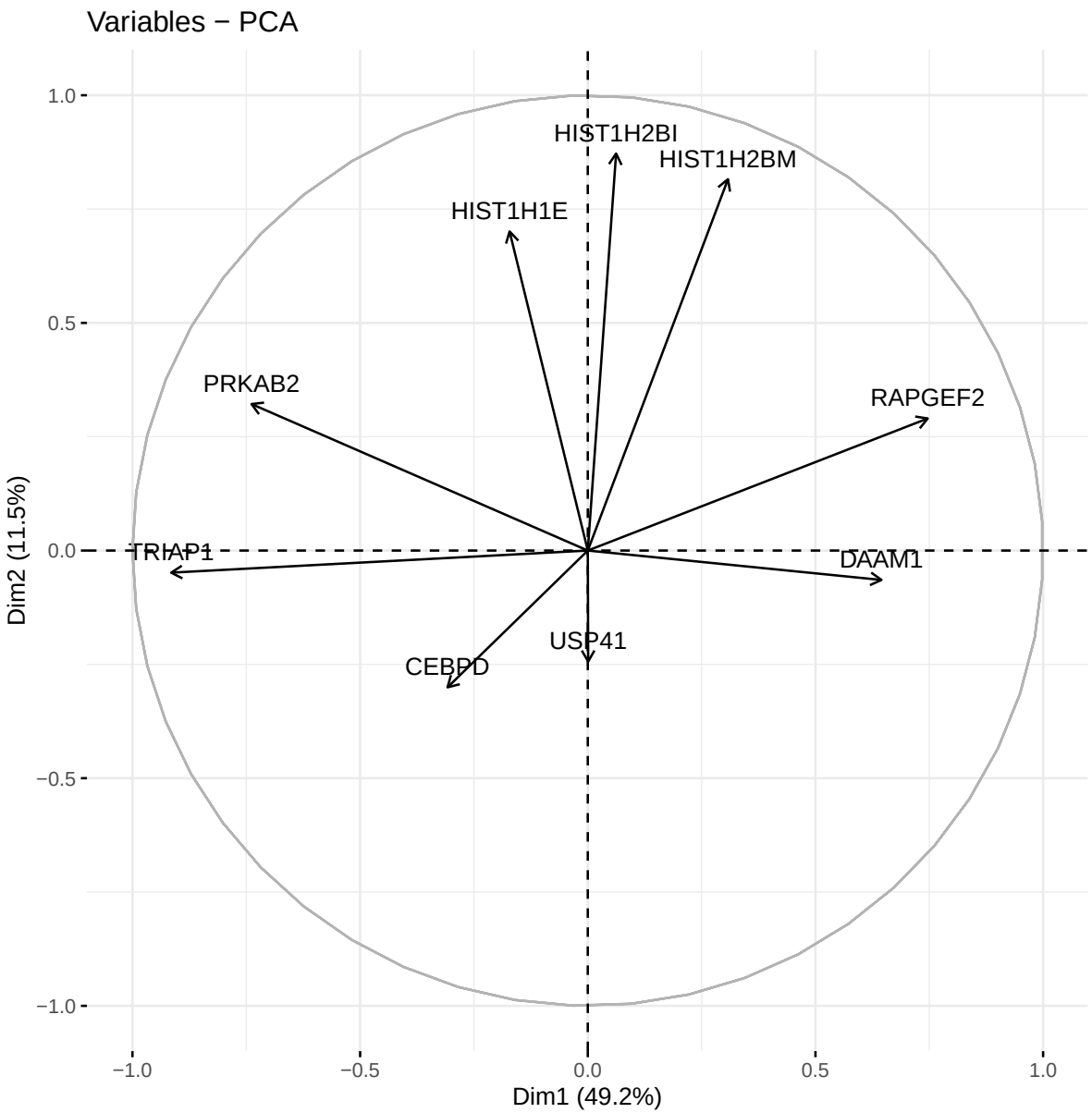


Fuente: Kovak, et al. 2018

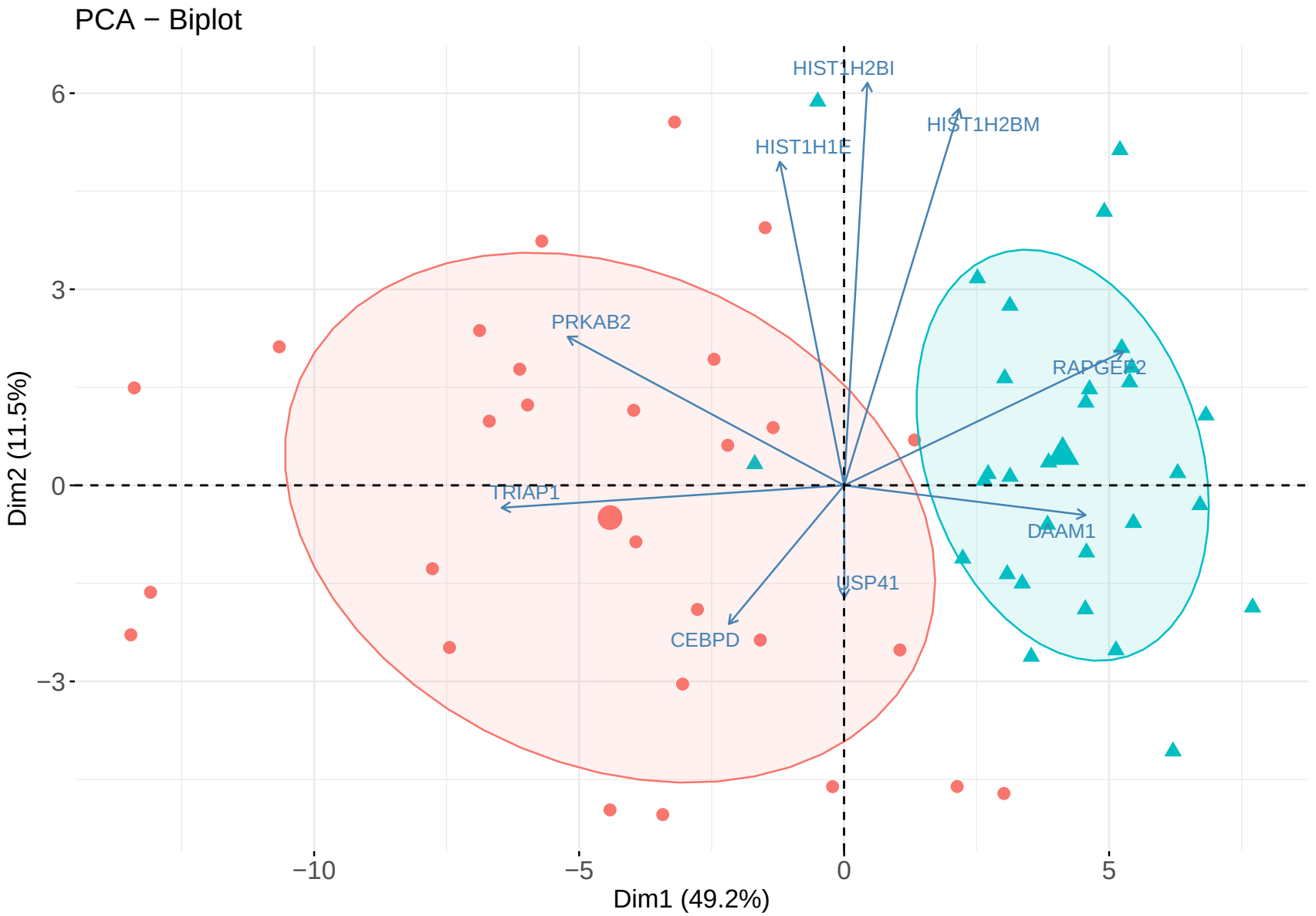
PCA: GRÁFICA POR MUESTRA



PCA: GRÁFICA POR VARIABLE



PCA: GRÁFICAS BI-PLOT

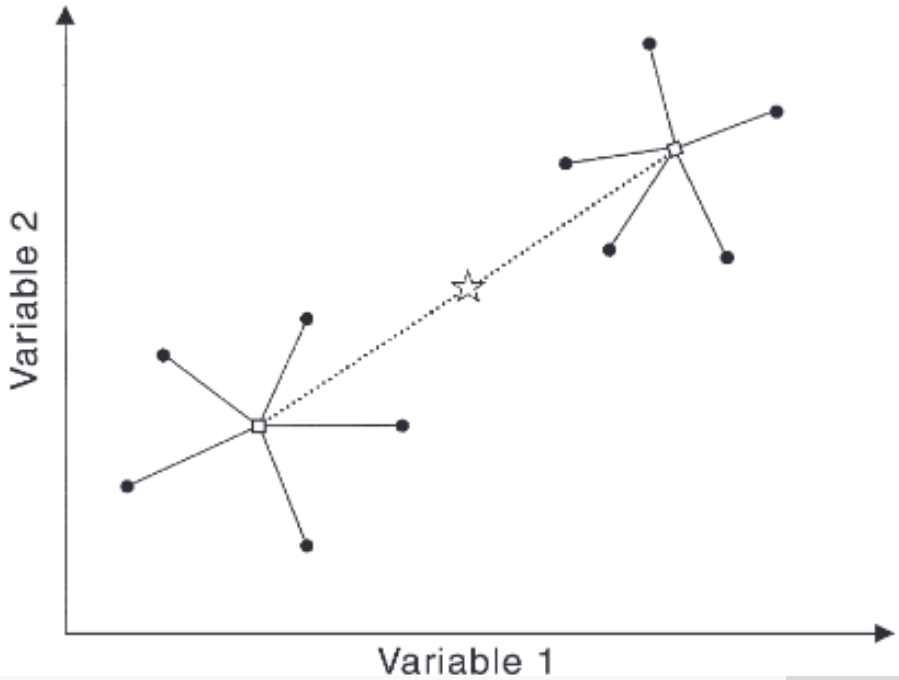


ANÁLISIS DE VARIANZA MULTIVARIANTE PERMUTACIONAL

- **¿Qué es un PERMANOVA?**

a) Es una prueba estadística multivariante No paramétrica.

b) Determina, en términos simples, si el centroides de un conjunto de observacione difiere del centróide de otro grupo.

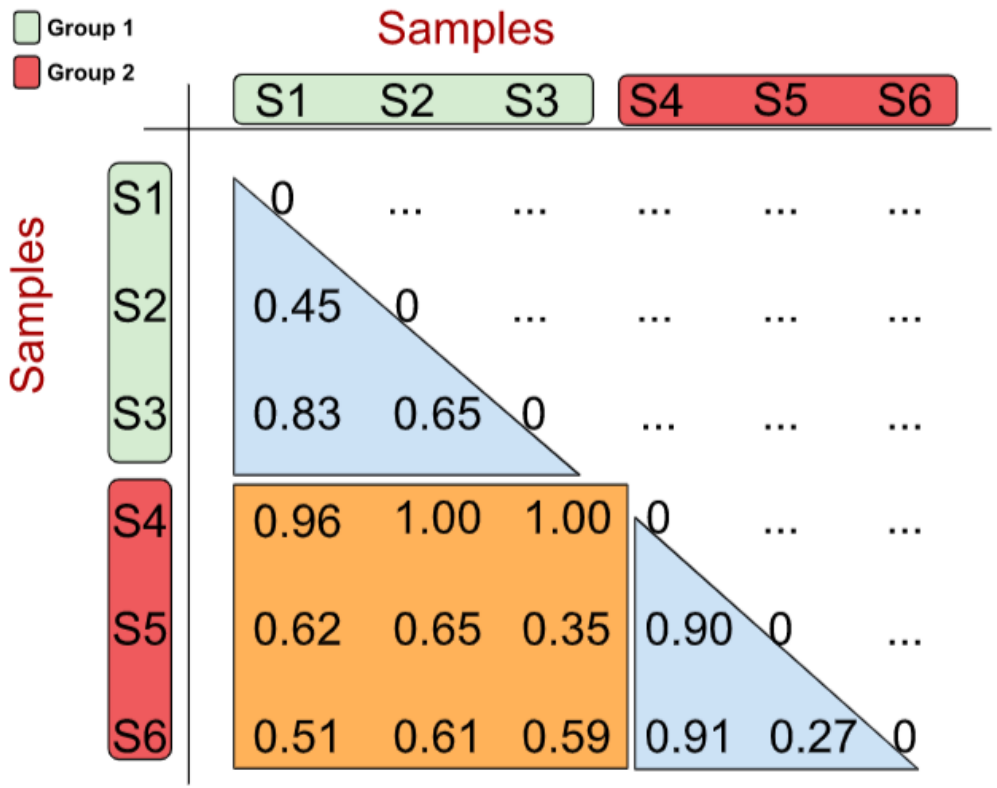


HIPÓTESIS PERMANOVA

- Hipótesis.

a) H_0= No existe diferencia entre los grupos.
b) H_1= Al menos dos grupos son diferentes.

- Datos: matriz de distancia.



PERMANOVA DATOS GENOTOXICIDAD

```
## Permutation: free
## Number of permutations: 999
##
## Terms added sequentially (first to last)
##
##           Df SumsOfSqs MeanSqs F.Model      R2 Pr(>F)
## Treatment  1    1.4351 1.43506  30.304 0.35113  0.001 **
## Residuals 56    2.6519 0.04736         0.64887
## Total     57    4.0870                 1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

RESUMEN DE LA CLASE

- ¿Qué son los análisis multivariados?.
- ¿Qué es un análisis de componentes principales?.
- ¿Qué son los componentes principales?.
- Etapas para realizar un ACP.
- Varianza explicada.
- Graficas biplot.