Clase 7:

Análisis de clustering/PCA y categorización de variantes patogénicas, visualización de los datos.

Evelin González F.
evelyn.gonzalez@uoh.cl

Organización de las clases: Parte 2

Clase 5: Introducción de R.

Clase 6: Librería Maftools para interpretación de variantes.

Clase 7: Análisis de clustering/PCA y categorización de variantes patogénicas, visualización de los datos.

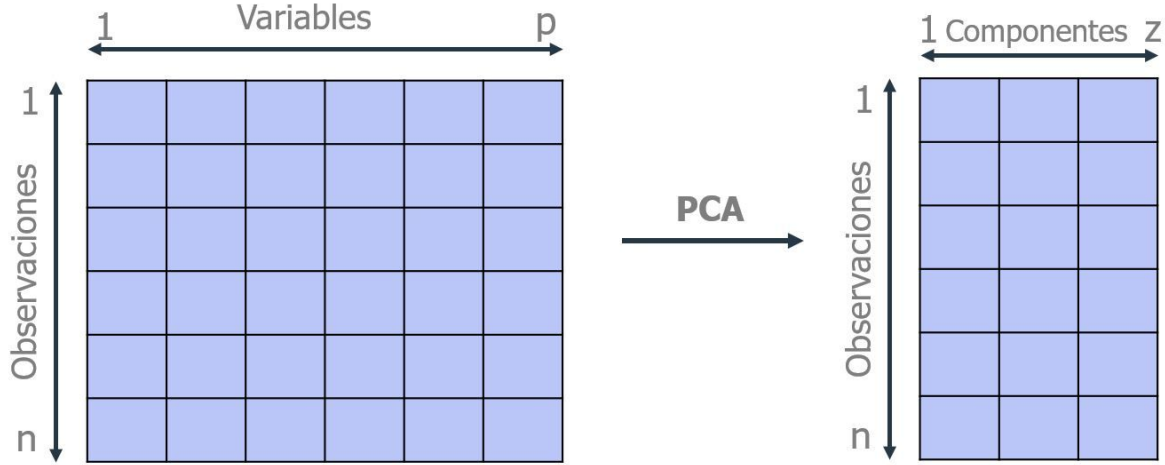
Introducción al Análisis de Datos: PCA, Clustering y Heatmaps

Objetivo: Entender cómo estas técnicas de análisis multivariado pueden ayudar a explorar y visualizar grandes conjuntos de datos.

- **PCA (Análisis de Componentes Principales):**
  Es una técnica de reducción de dimensionalidad que transforma los datos originales en un conjunto de nuevas variables ortogonales (componentes principales), ordenadas por su varianza. Permite identificar patrones y simplificar la visualización de datos complejos.
- **Clustering (Agrupamiento):**
  El clustering es un método de aprendizaje no supervisado que agrupa datos similares en "clusters" o grupos. Esto permite descubrir estructuras o patrones ocultos en los datos sin necesidad de etiquetas predefinidas.
- **Heatmap (Mapa de Calor):**
  Un heatmap es una representación gráfica de datos en una matriz donde los valores son representados mediante colores. Es útil para visualizar patrones, correlaciones y diferencias entre variables o muestras en grandes conjuntos de datos.

Análisis de Componentes Principales (PCA): Reducción de Dimensionalidad

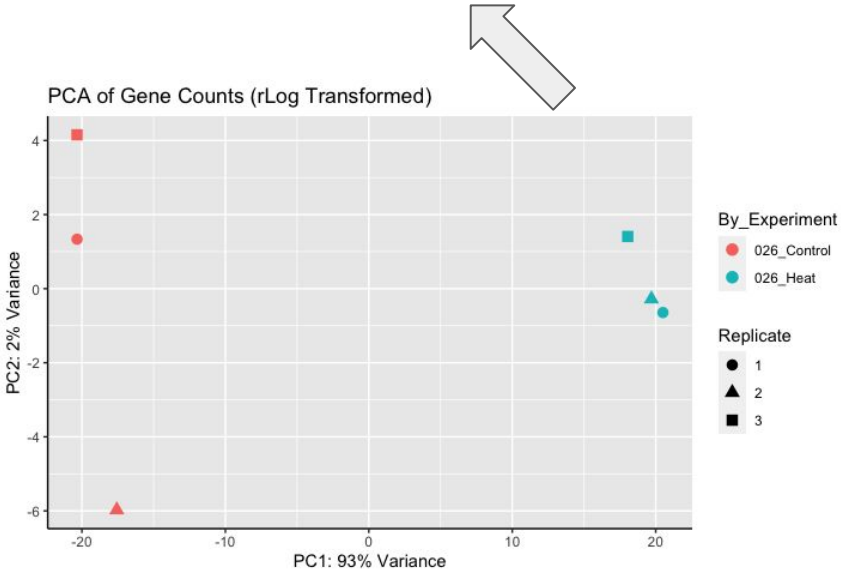
- El análisis de componentes principales (PCA) reduce la cantidad de dimensiones en grandes conjuntos de datos a componentes principales que **conservan** la mayor parte de la **información original**.
- Para ello, transforma las variables potencialmente correlacionadas en un conjunto más pequeño de variables, denominadas **componentes principales**.



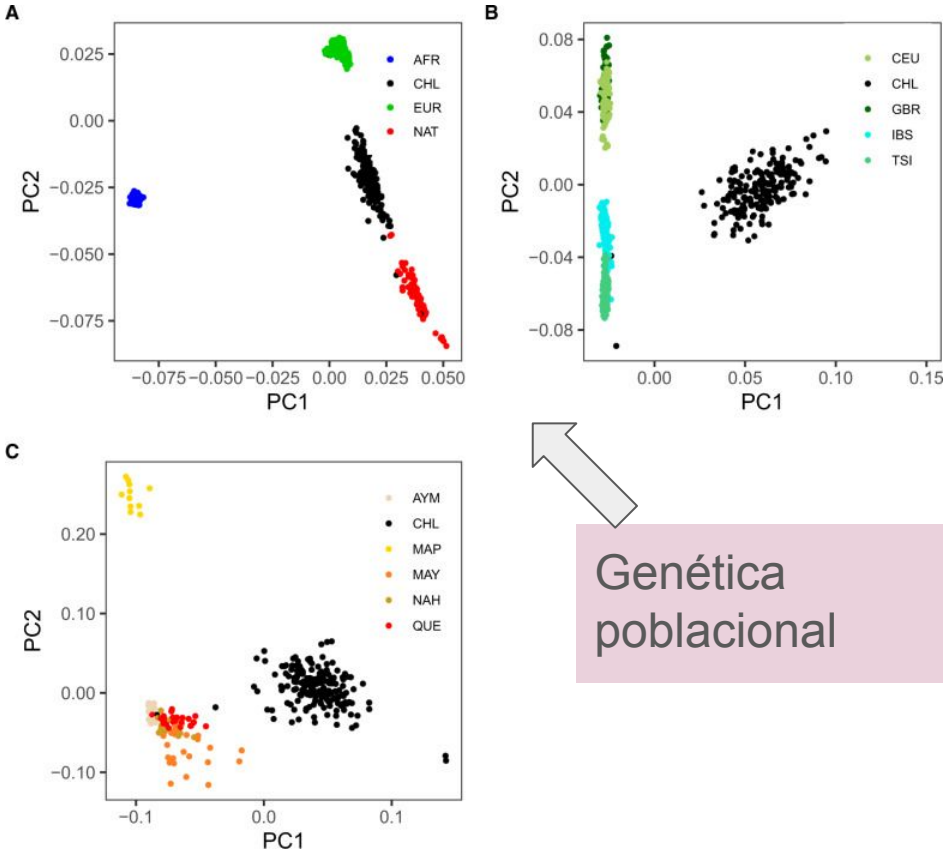
Caso de uso: PCA en genómica

```r
library("FactoMineR")
res.pca <- PCA(decathlon2.active, graph = FALSE)
```

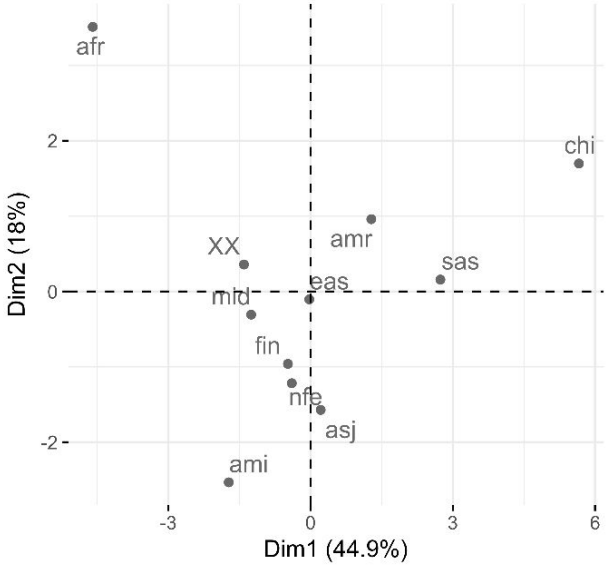
RNA-seq

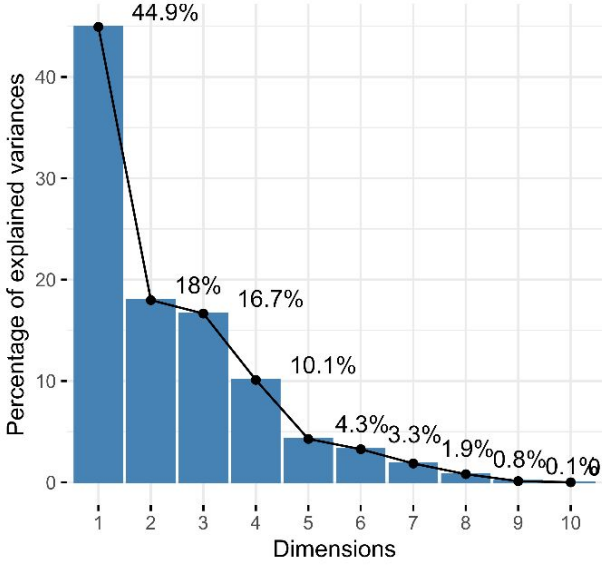


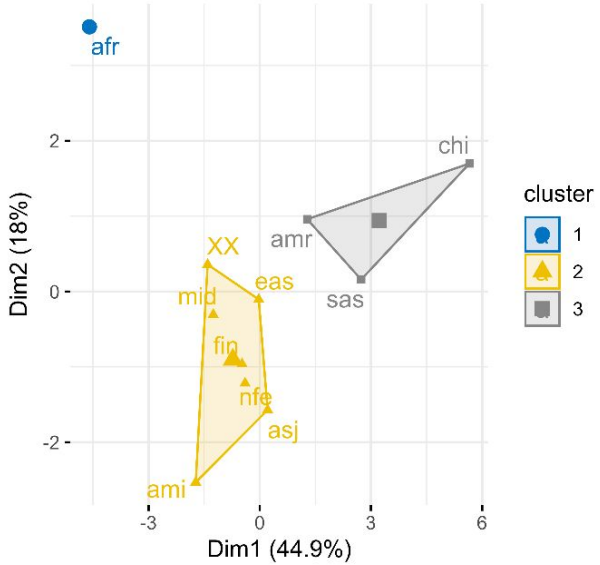
Genética poblacional



PCA en muestras de BRCA1/2

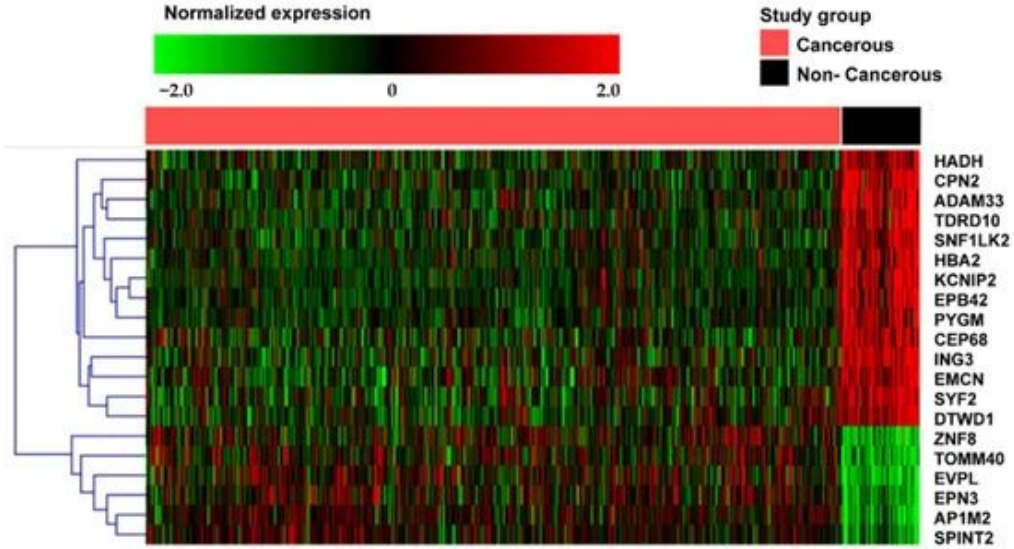






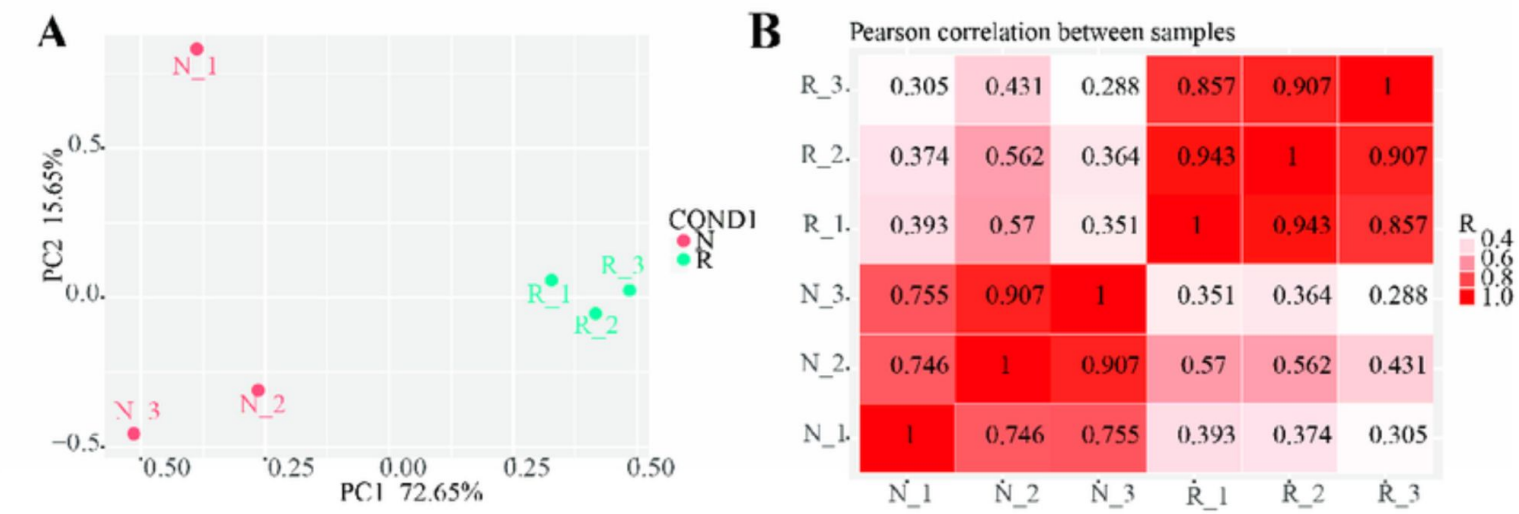
Heatmap (Mapa de Calor)

RNA-seq

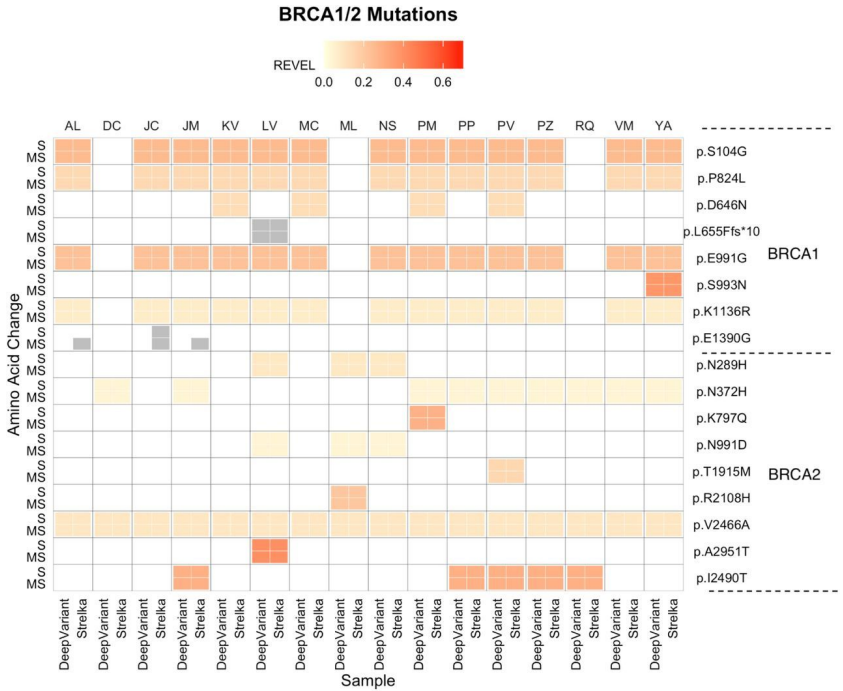


Un heatmap es una representación gráfica de datos en una **matriz** donde los valores son representados mediante colores. Es útil para visualizar **patrones, correlaciones** y **diferencias** entre variables o muestras en grandes conjuntos de datos.

PCA y heatmap para análisis exploratorio inicial



Caso uso: heatmap en variantes



Example data

Load data and subset for demonstration purposes.

```
example_file <- "https://davetang.org/file/TagSeqExample.tab"
data <- read.delim(example_file, header = TRUE, row.names = "gene")
data_subset <- as.matrix(data[rowSums(data)>50000,])
dim(data_subset)
```

```
[1] 49  6
```

Default heatmap

Default heatmap using `pheatmap`.

```
pheatmap(data_subset)
```

La librería `pheatmap` de R es una herramienta ampliamente utilizada para crear mapas de calor (heatmaps) con un alto grado de personalización en cuanto a dimensiones y apariencia.

Clustering: Análisis de Agrupamiento

Definición:
El **clustering** es una técnica de análisis de datos no supervisado que agrupa observaciones similares en "clusters" o grupos. Su objetivo es identificar patrones ocultos o relaciones dentro de los datos sin la necesidad de etiquetas predefinidas.

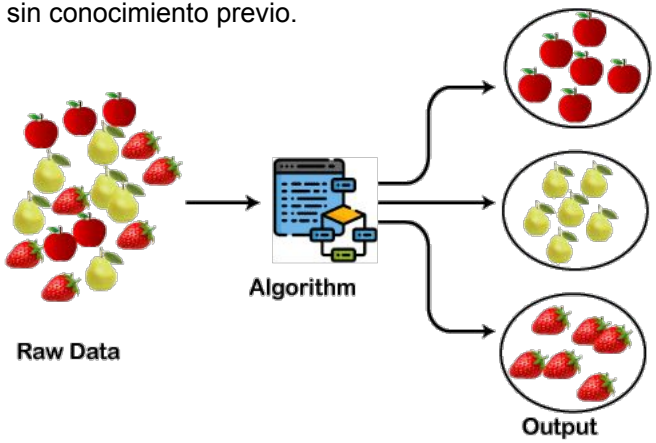
"El análisis de agrupamiento de datos, o clustering, puede revelar relaciones entre las observaciones y proporcionar una visión más profunda de los datos."

Beneficios:

- **Descubrimiento de patrones**: Permite identificar subgrupos naturales dentro de los datos.
- **Segmentación**: Útil para dividir grandes volúmenes de datos en categorías manejables.
- **Exploración de datos**: Ayuda a comprender mejor la estructura de los datos sin conocimiento previo.

Aplicaciones comunes:

- Análisis de grupos de genes en biología
- Reducción de dimensionalidad en imágenes



Anotación de variantes con gnomAD

Filter

	aaChange	afr	amr	ami	asj	eas	sas	fin	mid	nfe	XX	chi
1	p.A2951T	1.200e-03	0.0508	0.0011	0.0014	0.0000	0.0141	0.0008	0.0034	0.0051	0.0072	0.04166667
2	p.E991G	1.799e-01	0.3192	0.2851	0.3489	0.3687	0.4921	0.4029	0.3401	0.3313	0.2938	0.39583333
3	p.H743H	2.140e-02	0.0978	0.0263	0.0441	0.0998	0.1139	0.0146	0.0340	0.0344	0.0384	0.10416667
4	p.I2490T	3.200e-03	0.0767	0.0000	0.0000	0.0021	0.0012	0.0000	0.0000	0.0002	0.0080	0.12500000
5	p.I2944F	4.090e-02	0.0037	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0123	0.02083333
6	p.K1132K	2.303e-01	0.2241	0.1762	0.3073	0.3979	0.2931	0.3182	0.2959	0.3165	0.2828	0.16666667
7	p.K1136R	2.325e-01	0.3281	0.2841	0.3589	0.3693	0.4932	0.4039	0.3503	0.3301	0.3091	0.39583333
8	p.L1521L	9.260e-01	0.9925	1.0000	1.0000	0.9998	0.9994	1.0000	0.9762	0.9997	0.9779	1.00000000

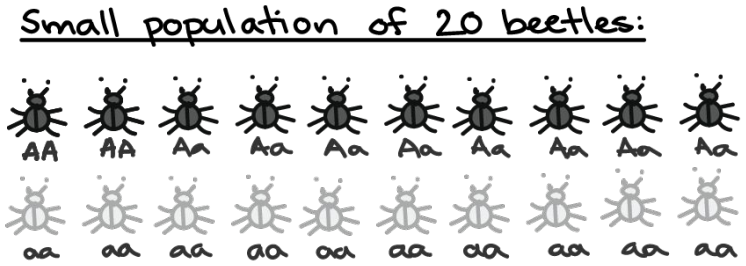
Estas subpoblaciones ayudan a:

- **Calcular frecuencias alélicas:** Identificar qué variantes son específicas o más comunes en ciertos grupos.
- **Detectar variantes patogénicas:** Asociar mutaciones con enfermedades prevalentes en subpoblaciones específicas.
- **Evitar sesgos:** Garantizar que el análisis genómico sea inclusivo y representativo de la diversidad genética global.

Subpoblaciones principales:

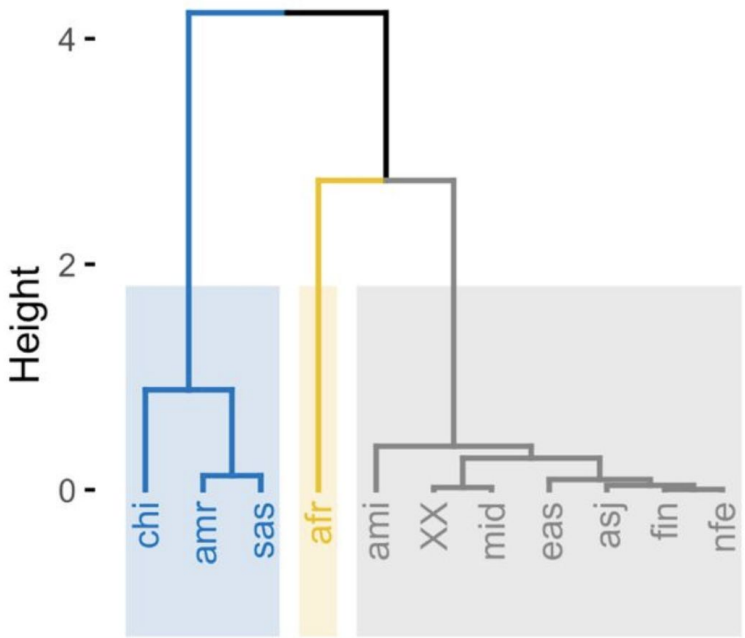
AFR (African)
AMR (Admixed American)
ASJ (Ashkenazi Jewish):
EAS (East Asian):
SAS (South Asian):

Cálculo de la Frecuencia Alélica Menor (MAF)



El cálculo de la **Frecuencia Alélica Menor (MAF)** se realiza dividiendo la cantidad de veces que aparece el alelo menos frecuente de una variante genética por el total de alelos observados en la población (dos por individuo en un genoma diploide).

Clustering: Agrupación Basada en MAF en Diferentes Poblaciones



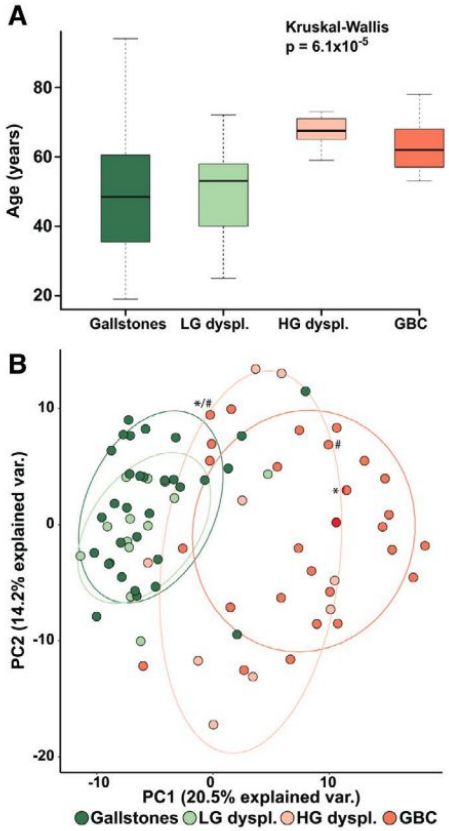
¿Qué es el Clustering Jerárquico?
Es una técnica de agrupamiento que organiza los datos en una jerarquía de grupos basándose en sus similitudes.

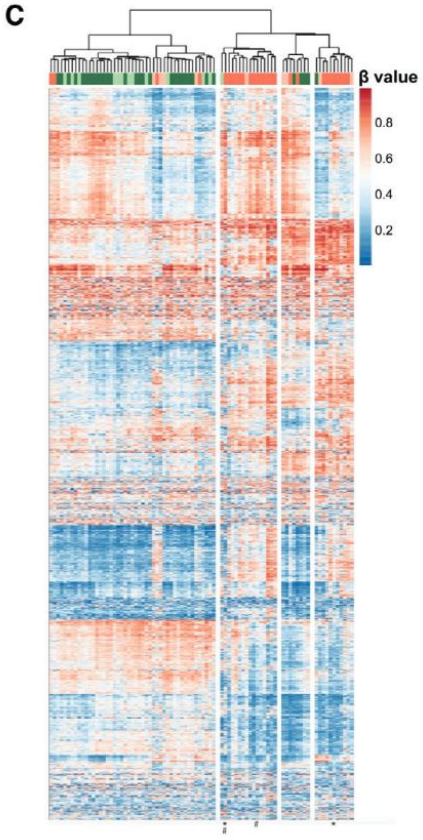
- **Algoritmo:** Une iterativamente las observaciones o grupos más similares.
- **Representación:** El resultado se visualiza en un dendograma, una estructura tipo árbol.

Características principales:

- **Flexibilidad:** Permite identificar patrones en múltiples niveles.
- **Distancias:** Basado en métricas como Euclidiana o Manhattan.
- **Métodos de enlace:** Completo, promedio o simple

Caso de uso: PCA, clustering y heatmap





HEPATOLOGY
AASLD
HEPATOLOGY, VOL. 73, NO. 6, 2021

Epigenome-Wide Analysis of Methylation Changes in the Sequence of Gallstone Disease, Dysplasia, and Gallbladder Cancer

Johannes Brägelmann,[1-3] Carol Barahona Ponce,[1,4] Katherine Marcelain,[4] Stephanie Roessler,[5] Benjamin Goeppert,[5] Ivan Gallegos,[6] Alicia Colombo,[4,6] Verónica Sanhueza,[7] Erik Morales,[8] María Teresa Rivera,[9] Gonzalo de Toro,[10] Alejandro Ortega,[11] Bettina Müller,[12] Fernando Gabler,[13] Dominique Scherer,[1] Melanie Waldenberger,[14] Eva Reischl,[14] Felix Boekstegers,[1] Valentina Garate-Calderon,[1,4] Sinan U. Umu,[15] Trine B. Rounge,[15,16] Odilia Popanda,[17] and Justo Lorenzo Bermejo[1]