Assignment 2

Using Dataset AntiProfilesData

All questions results must be saved, zipped and uploaded on the blackboard.

Submit R scripts of your code, files and images not the whole project on the blackboard.

=> Question 1:

Explore the data:

Show the type of the data and assign each of feature, phenotype and expression data to different variables

1.a Show the type of each column

1.b Show column names and rows name

1.c Calculate summary of each column

1.d Show frequency of categorical data, taking into the consideration, NA values frequency if any.

1.e Calculate the correlation and covariance between the first 10 columns *only* of our data set and draw full correlation matrix.

1.f For both genes: GSM95478,GSM95473 show the plot with a line of their relation.

=> Question 2:

Using PCA and SVD, Prove by plotting and values that both can return the same result by suitable normalization.

=>Question 3:

256 visual artists were surveyed to find out their zodiac sign. The results were: Aries (29), Taurus (24), Gemini (22), Cancer (19), Leo (21), Virgo (18), Libra (19), Scorpio (20), Sagittarius (23), Capricorn (18), Aquarius (20), Pisces (23).

3.1) Test the hypothesis that zodiac signs are evenly distributed across visual artists.

3.2) Explicitly mention your H1 and Ho assumption.

=> Question 4:

Plot hierarchical clusters on our first 10 columns of edata and apply the kmeans to all the edata columns and show the centroid of the result.

Setups to Download Data:

1- Install bioconductor (refer to the labs in this step)

2- Install antiProfilesData By BiocManager::install("antiProfilesData")

3- To Access our data use antiProfilesData::apColonData

Dataset Documentation:

`apColonData`	*Curated dataset of many colon normal and cancer samples on Affymetrix hgu133plus2 expression arrays.*

Description

Data used in Corrada Bravo, et al. gene expression anti-profiles paper: BMC Bioinformatics 2012, 13:272 doi:10.1186/1471-2105-13-272. Measurements are z-scores obtained from the GeneExpression Barcode in the 'frma' package. Only probes mapped to genes within colon cancer hypomethylation blocks defined in Hansen et al. are included.

format

Data is an `ExpressionSet` object. The `exprs` slot contains gene expression barcode z-scores from `frma` preprocessed data. The `phenoData` slot contains a data frame with the following columns:

`filename`: The CEL filename in the Gene Expression Omnibus (GEO)

`DB_ID`: The GSM sample id in GEO

`ExperimentID`: The GSE experiment id in GEO

`Tissue`: Tissue type, obtained from the gene expression barcode annotation

`SubType`: Sample sub-type, obtained from the gene expression barcode annotation

`ClinicalGroup`: Clinical sample annotation, obtained from the gene expression barcode annotation

`Status`: Normal (0) or Cancer (1) indicator