# Get GEO DATSET

data <- getGEO(GEO = ["GSE152075"])

#replace the text between [] with the GSE of your choice and remove the [].

#print de first five rows of the matrix to see matrix information

head(data)

#output (do not run this piece of script)

> head(data)

$GSE152075\_series\_matrix.txt.gz

ExpressionSet (storageMode: lockedEnvironment)

assayData: 0 features, 484 samples

  element names: exprs

protocolData: none

phenoData

  sampleNames: GSM4602241 GSM4602242 ... GSM4602725 (484 total)

  varLabels: title geo\_accession ... sequencing\_batch:ch1 (43 total)

  varMetadata: labelDescription

featureData: none

experimentData: use 'experimentData(object)'

  pubMedIds: 32898168

Annotation: GPL18573

# You may now extract the phenotypic/clinical data matrix from the series matrix:

clindata <- data[["GSE152075\_series\_matrix.txt.gz"]]@phenoData@data

#replace the GSE with the one of your choice. Do not remove the [] in this line.

#print de first five rows of the matrix to see the information included in columns of interest

head(clindata[,c(1,2,8,40,39,42)])

#output (do no run this piece of script)

> head(clindata[,c(1,2,8,40,39,42)])

title geo\_accession source\_name\_ch1 gender:ch1 age:ch1 sars-cov-2 positivity:ch1

GSM4602241 POS\_001 GSM4602241 Nasopharyngeal Swab M 64 pos

GSM4602242 POS\_002 GSM4602242 Nasopharyngeal Swab F 30 pos

GSM4602243 POS\_003 GSM4602243 Nasopharyngeal Swab M 47 pos

GSM4602244 POS\_004 GSM4602244 Nasopharyngeal Swab F 67 pos

GSM4602245 POS\_005 GSM4602245 Nasopharyngeal Swab M 62 pos

GSM4602246 POS\_006 GSM4602246 Nasopharyngeal Swab F 52 pos

# Download and save on your computer the raw-counts matrix from GEO website. This matrix is a tab-delimited txt. file containing the counts for every gene aligned from a RNA-seq experiment. After downloading it, load the matrix into RStudio:

raw\_counts <- read.delim("[C:/Users/File/Location/GSE152075\_raw\_counts\_GEO.txt.gz]", stringsAsFactors=FALSE, sep = " ")

#replace the text between [] with the directory path to the GSE\_raw\_counts\_GEO.txt.gz file you downloaded and remove the [].

#another way to download the raw count matrix directly from RStudio is running the following command:

url="https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE152075&format=file&file=GSE152075%5Fraw%5Fcounts%5FGEO%2Etxt%2Egz"

download.file(url, "raw\_reads.gz")

raw\_counts <- read.delim("raw\_reads.gz", stringsAsFactors=FALSE, sep = " ")

#print de first five rows of the raw counts matrix to see how information is organized

head(raw\_counts[,c(1:10)])

#output (do no run this piece of script)

> head(raw\_counts[,c(1:10)])

POS\_001 POS\_002 POS\_003 POS\_004 POS\_005 POS\_006 POS\_007 POS\_008 POS\_009 POS\_010

A1BG 0 1 0 0 18 8 0 1 0 1

A1CF 0 0 2 0 0 0 0 0 0 0

A2M 69 36 84 42 83 46 26 0 93 6

A2ML1 2 0 0 0 3 30 0 32 6 0

A2MP1 0 0 0 0 21 0 0 0 0 0

A3GALT2 0 0 0 0 0 0 0 0 0 0

Gene expression normalization:

1. Before sample normalization, data should be converted and organized to the format required for further analysis (data format and organization might vary for different packages)

raw\_counts <- as.matrix(raw\_counts)

rownames(clindata) <- clindata$[title]

#replace the rownames of clindata (sampleID) with the same sample name [title] of [raw\_counts]. This will help to match sample names in both matrixes. Remove the []

all(rownames(clindata) %in% colnames(raw\_counts))

#the outcome should be TRUE

all(colnames(raw\_counts) %in% rownames(clindata))

#the outcome should be TRUE

1. Make sure that the grouping variables are factors. We also changed the original names of the columns containing the relevant variables to make them shorter and easier to work with.

colnames(clindata)[colnames(clindata) == "sequencing\_batch:ch1"] <- "batch"

clindata$batch <- as.factor(clindata$batch)

colnames(clindata)[colnames(clindata) == "n1\_ct:ch1"] <- "ct"

colnames(clindata)[colnames(clindata) == "sars-cov-2 positivity:ch1"] <- "positivity"

clindata$positivity[clindata$positivity == "pos"] <- "COVID19"

clindata$positivity[clindata$positivity == "neg"] <- "HEALTHY"

clindata$positivity <- as.factor(clindata$positivity)

1. Merge the read counts and clinical data matrixes into a DESeqDataSet object using the DESeq2 package: