library(GEOquery)

library(DESeq2)

library(dplyr)

library(ggplot)

library(tidyverse)

data <- getGEO(GEO = "GSE152075")

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

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 = " ")

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

With the above command please check the colnames and rownames of counts data do the same for clinicaldata

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

#the outcome should be TRUE

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

#the outcome should be TRUE

If the outcome of either of them is false do the following steps

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

rownames(clindata) <- clindata$[title]

This is the very important step where you are matching the colnames and rownames of your two file one is clinicaldata (Metadata) and the other one is counts data.

#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 []

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)

All these steps are to rename the conames/rownames for the sake of easiness you can skip upto you

dds <- DESeqDataSetFromMatrix(countData = raw\_counts,

+

+ colData = clindata,

+

+ design = formula(~positivity)) Here you will add the name of the column which specifes both samples e.g in this particular case in the variable positivity both pos and neg of covid are written so I use positive here.

dds <- estimateSizeFactors(dds)

sizeFactors(dds)

res <- results(dds)

summary(res)

plotMA(res) This should plot the volcano plot displaying the differential expressed genes