Bioinformatics Assignment 2

# Define file paths  
data\_dir <- file.path("../../data", "SRP164913")  
data\_file <- file.path(data\_dir, "SRP164913\_HUGO.tsv")  
metadata\_file <- file.path(data\_dir, "metadata\_SRP164913.tsv")  
results\_dir <- file.path("../../results")  
plots\_dir <- file.path("plots")  
  
# Libraries  
library(DESeq2)

## Loading required package: S4Vectors

## Loading required package: stats4

## Loading required package: BiocGenerics

##   
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':  
##   
## IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':  
##   
## anyDuplicated, aperm, append, as.data.frame, basename, cbind,  
## colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,  
## get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,  
## match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,  
## Position, rank, rbind, Reduce, rownames, sapply, setdiff, table,  
## tapply, union, unique, unsplit, which.max, which.min

##   
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:utils':  
##   
## findMatches

## The following objects are masked from 'package:base':  
##   
## expand.grid, I, unname

## Loading required package: IRanges

## Loading required package: GenomicRanges

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##   
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':  
##   
## colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,  
## colCounts, colCummaxs, colCummins, colCumprods, colCumsums,  
## colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,  
## colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,  
## colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,  
## colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,  
## colWeightedMeans, colWeightedMedians, colWeightedSds,  
## colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,  
## rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,  
## rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,  
## rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,  
## rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,  
## rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,  
## rowWeightedMads, rowWeightedMeans, rowWeightedMedians,  
## rowWeightedSds, rowWeightedVars

## Loading required package: Biobase

## Welcome to Bioconductor  
##   
## Vignettes contain introductory material; view with  
## 'browseVignettes()'. To cite Bioconductor, see  
## 'citation("Biobase")', and for packages 'citation("pkgname")'.

##   
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':  
##   
## rowMedians

## The following objects are masked from 'package:matrixStats':  
##   
## anyMissing, rowMedians

library(ggplot2)  
library(magrittr)

##   
## Attaching package: 'magrittr'

## The following object is masked from 'package:GenomicRanges':  
##   
## subtract

library(M3C)  
library("umap")

##   
## Attaching package: 'umap'

## The following object is masked from 'package:M3C':  
##   
## umap

# Set seed for reproducible results  
set.seed(12345)

**Group:** 5, **Date:** 09/25/2024 Order to run R scripts:  
(task 1) preprocessing.R  
(task 2) PCA\_Plot.R  
(task 3) DifferentialAnalysis.R  
(task 4) heatmap.R  
(task 5-Hannah) gprofiler2.R

## Task 1 - Data analysis

### Sample Size

data\_analysis\_df <- read.delim("../../data/SRP164913/SRP164913\_HUGO.tsv",header = TRUE, row.names = 1, stringsAsFactors = FALSE)  
cat("Number of Genes in the expression matrix: ", dim(data\_analysis\_df)[1], "\n")

## Number of Genes in the expression matrix: 29589

cat("Number of Samples in the expression matrix: ", dim(data\_analysis\_df)[2], "\n")

## Number of Samples in the expression matrix: 88

There are 88 samples, however, there are three disease groups: healthy controls, MS subjects, and ham/tsp subjects. To make this a binary problem, from Task 2 onward, we removed the ham/tsp subjects to compare only MS vs healthy controls (hc). This reduces the number of samples to 62.

### Density Plot of Gene Expressions

# Get the median of expressions by gene  
gene\_median <- apply(data\_analysis\_df, 1, median)  
head(gene\_median)

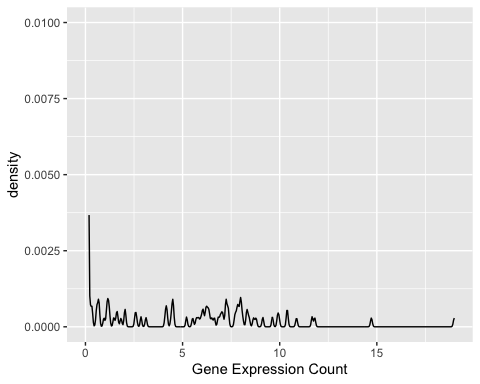
## A1BG A1BG-AS1 A1CF A2M A2M-AS1 A2ML1   
## 0 0 0 0 0 0

gene\_median = log2(gene\_median +1)  
cat("Variance between gene expression medians:", var(gene\_median, na.rm = TRUE))

## Variance between gene expression medians: 0.1234497

# Create a Data frame from the numerical array  
gene\_median\_df <- data.frame(Median = gene\_median)  
# Plot the values  
ggplot(gene\_median\_df, aes(x = Median)) + geom\_density() + xlab("Gene Expression Count") + ylim(0,1e-2)

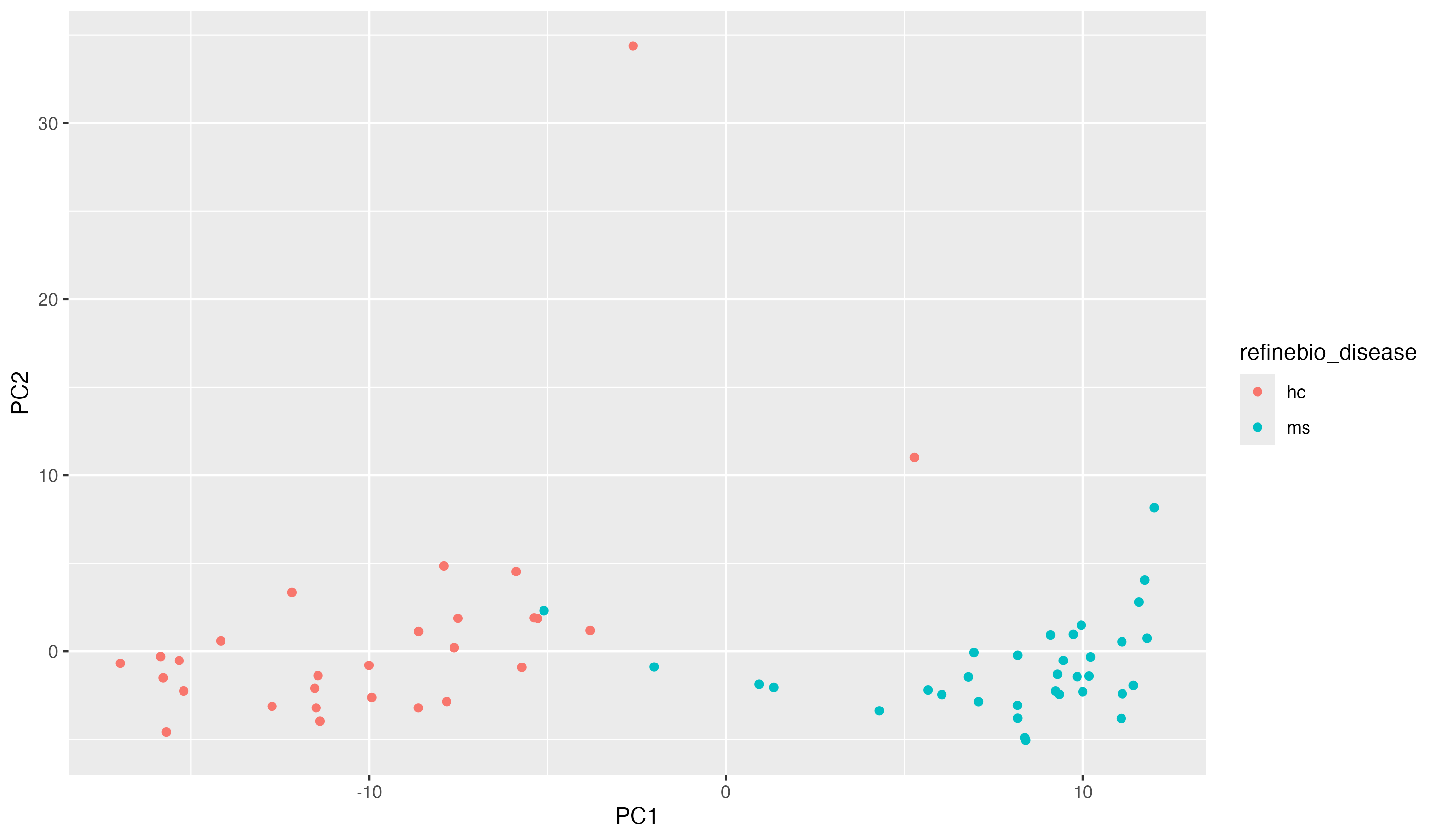
## Warning: Removed 72 rows containing non-finite outside the scale range  
## (`stat\_density()`).

 The density plot shows a left skewed distribution of gene expression counts with most of the expressions being between 0 and 10 but with a few outliers between 10 and 15. We downloaded the un-normalized data, however when we opened the raw file, most count values were 0 or very close to 0. We suspect the data came pre-normalized, and this is causing some issues. This carries over through the entire assignment.

## Task 2 - Principal Component Analysis

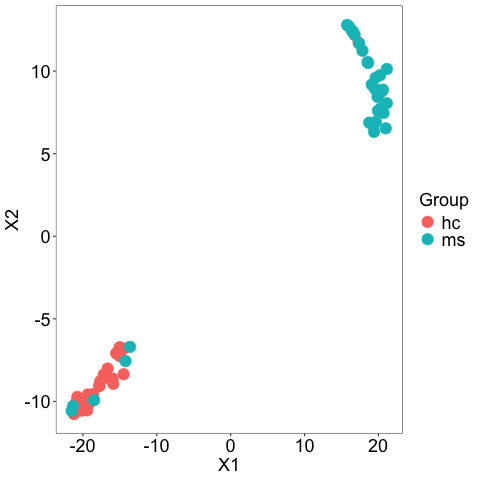
We were able to get a PCA, tsne, and umap plot following the tutorials. All three show a clear difference between the healthy control (hs) and MS (ms) groups, with a few samples in the midst of the other cluster. The number of crossovers is different.

### PCA Plot



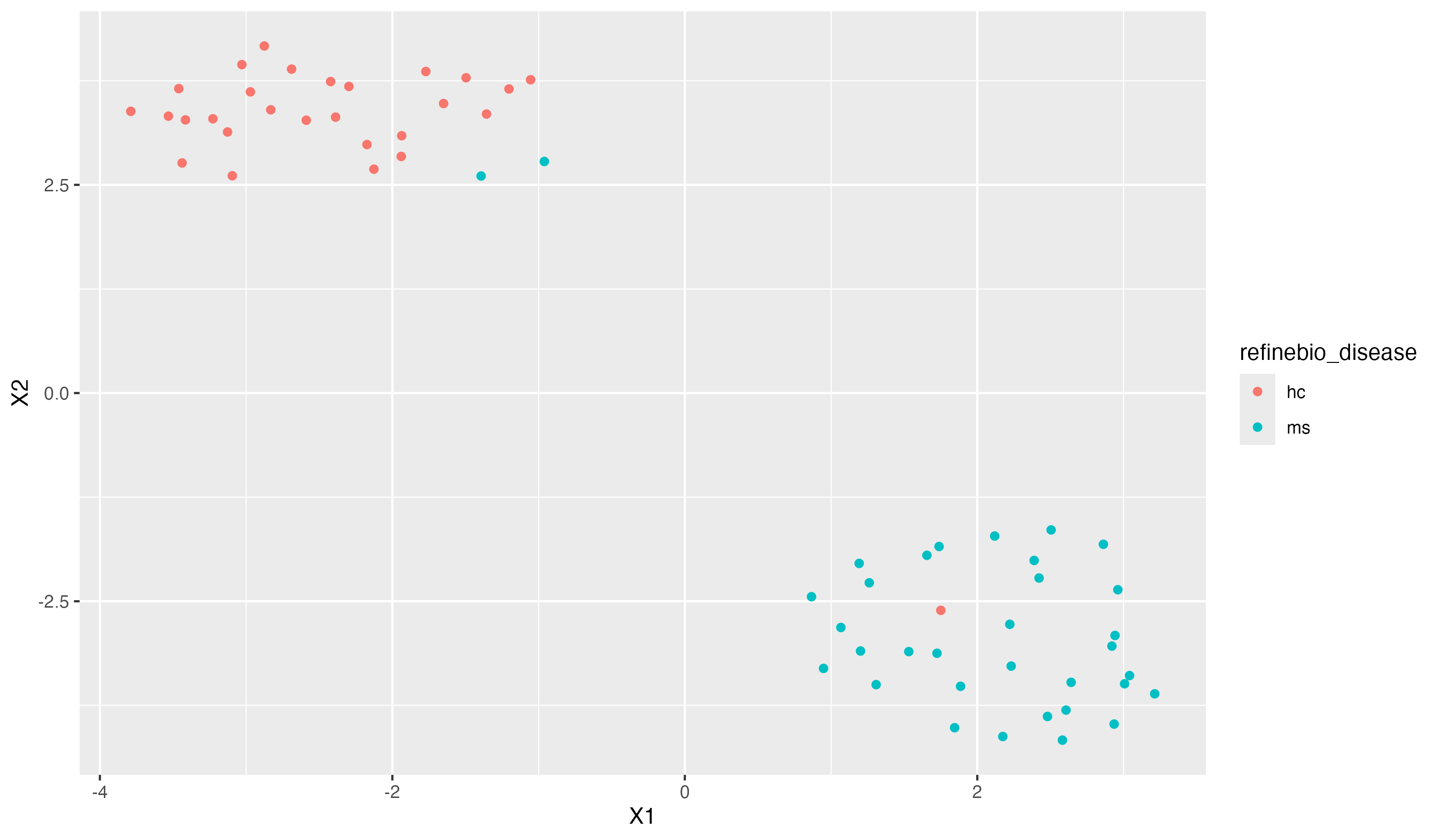
PCA Plot

### TSNE Plot



tsne Plot

### UMAP Plot



umap Plot

## Task 3 - Differential Analysis

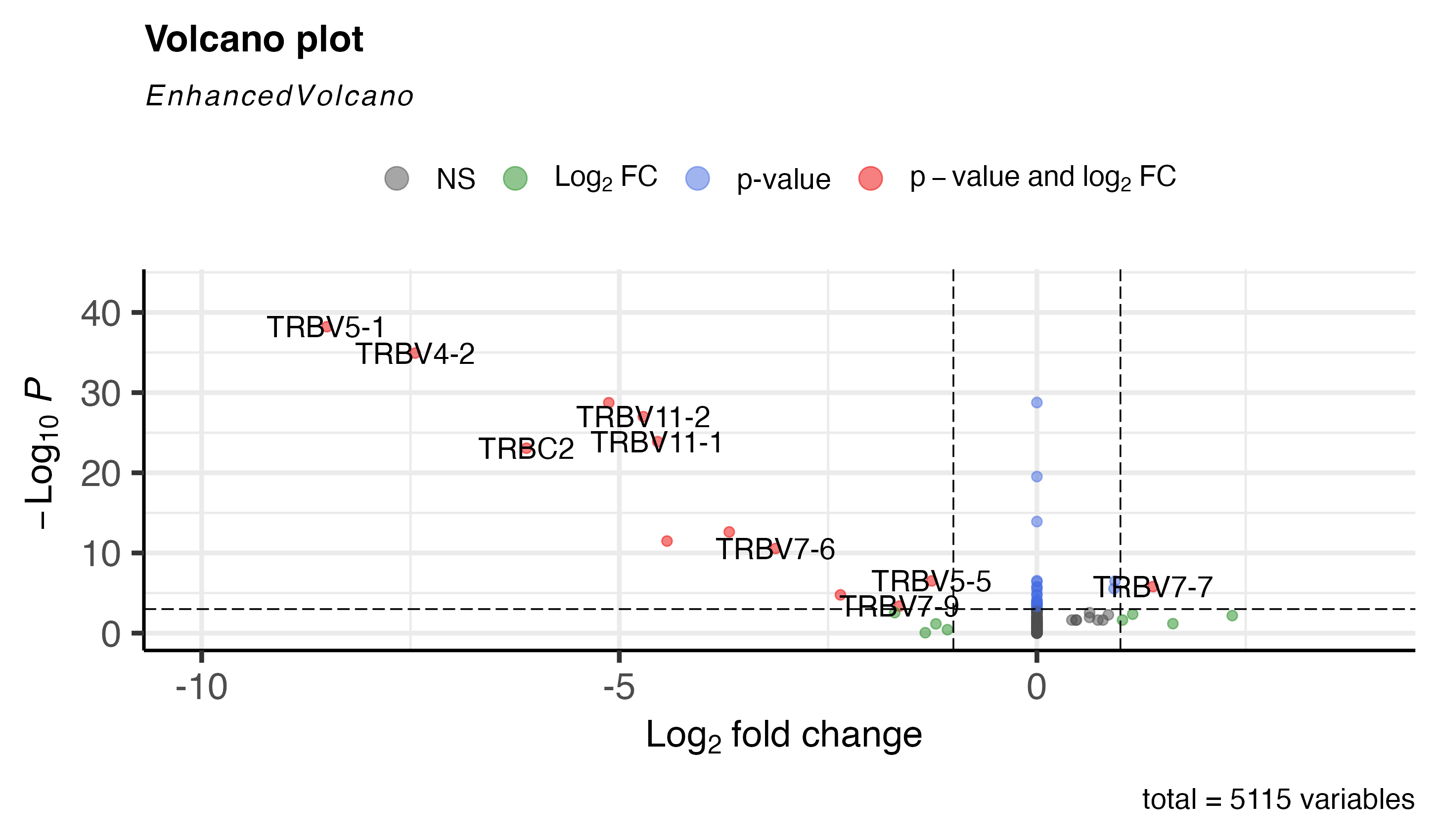
When we did our differential analysis, because the data was so small and normalized, we had to adjust the filter value to 0.01 or else only ~1000 rows were making it to the DESeqDataSetFromMatrix() step. This was causing our volcano plot to look very sparse. We were able to get a list of the top 50 differentially expressed genes. The results, both the entire table and only the top 50, are stored in the results folder.

top\_50\_file <- file.path(results\_dir, "SRP164913\_diff\_expr\_top\_50\_results.tsv")  
top\_50 <- readr::read\_tsv(top\_50\_file)

## Rows: 50 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (1): Gene  
## dbl (5): baseMean, log2FoldChange, lfcSE, pvalue, padj  
## lgl (1): threshold  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

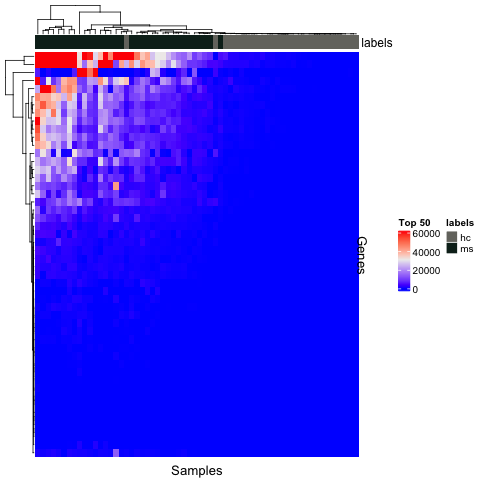
head(top\_50)

## # A tibble: 6 × 7  
## Gene baseMean log2FoldChange lfcSE pvalue padj threshold  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <lgl>   
## 1 TRBV23-1 9.45 2.34 0.614 0.000168 0.00652 TRUE   
## 2 TRBV16 18.9 1.63 0.424 0.0633 0.0663 FALSE   
## 3 TRBV7-7 557. 1.39 0.269 0.0000000137 0.00000158 TRUE   
## 4 TRBV7-4 31.7 1.15 0.362 0.000108 0.00435 TRUE   
## 5 TRBV12-4 389. 1.02 0.509 0.000837 0.0234 TRUE   
## 6 TRBV15 729. 0.936 0.172 0.00000000247 0.000000326 TRUE

 The volcano plot, while still somewhat sparse, is far more populated than it was with the highly filtered data. Most significant results have a negative log fold change.

## Task 4 - Heatmap

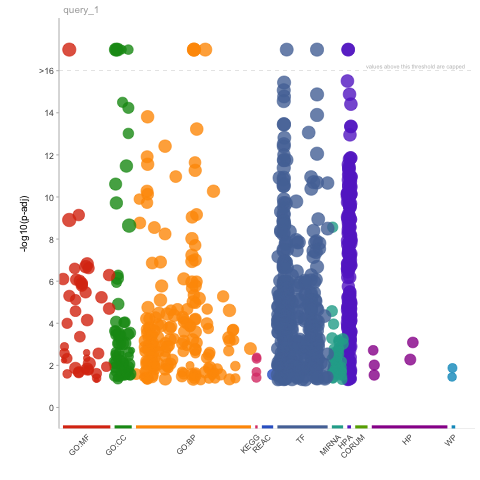
A heatmap was generated for the top 50 differentially expressed genes. The counts for each gene in the top 50 were extracted and put into the heatmap plot. It was annotated with the two disease groups: ms subjects and healthy control subjects. Since many of our counts are 0 and therefore are not expressed differently, the map is largely blue.



Heatmap

## Task 5 - Enrichment Analysis

Hannah- gprofiler2 The function gost was used from gprofiler2 to perform enrichment analysis on the differentially expressed data. Disease ontology was used, as the data was differentially expressed between the MS subjects and healthy controls. Only genes with adjusted p-values below 0.05 were put into the function. Two gost results were generated: one that contains all results, significant or not, and one that contained only significant results. The plot (created with gostplot) for the significant results is shown below. It seems there are a few points above the significance threshold.

 However, when publish\_gosttable(gost\_res3) was run, the function would not complete. It would continue to run for over 15 minutes and never produce results. I was unable to generate a table from the results, and when I tried to print the results, it would return NULL for both significant and non-significant gost results. The NULL result is shown below along with the gost function that was run. It is possible that no truly significant results were generated and so the results were empty.

results\_dir <- file.path("../../results")  
diff\_expr\_file <- file.path(results\_dir, "SRP164913\_diff\_expr\_results.tsv")  
library(gprofiler2)  
diff\_expr\_df <- readr::read\_tsv(diff\_expr\_file)

## Rows: 5115 Columns: 7  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (1): Gene  
## dbl (5): baseMean, log2FoldChange, lfcSE, pvalue, padj  
## lgl (1): threshold  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

gost\_res3 <- gost(  
 query = unlist(diff\_expr\_df[diff\_expr\_df$padj<0.05, 'Gene']),   
 significant=TRUE,   
 organism = "hsapiens"  
 )   
head(gost\_res3$results)

## NULL

When I had initially tried to run gost with every gene name, no p-value factor, it would produce a result that could be printed, but that is not as relevant to our dataset. This does prove that something is off with specifically the differential data, and the function does work.

Unfortunatly, due to the circumstances described above, no table was generated using the gprofiler2 method.

## Task 6/7 - Joint Enrichment Analysis