Tuberculosis Bulk RNA-seq Analysis

Table of contents

1 Download and Setup

(download data from GSE, establish needed local directories, format data objects)

```
library(tidyverse) # dplyr, tidyr, tibble, readr
library(readxl)
library(Biobase)
library(GEOquery)

dir.create(file.path("data_raw"), showWarnings = FALSE) # establish directory to accomodate GEO downl

supplemental_dir <- file.path("data_raw", "GSE107994")

if (!dir.exists(supplemental_dir) || length(list.files(supplemental_dir)) == 0) { # checks for supple
  getGEOSuppFiles("GSE107994", baseDir = "data_raw") # accesses GEO for supplemental files (gene cour
}

gse <- getGEO("GSE107994", GSEMatrix = TRUE) # grabs the series matrix

counts_file <- list.files( # grabs the specific file name for the raw gene counts data
  path = supplemental_dir,
  pattern = "Raw_counts.*Leicester.*\\.xlsx$",
  full.names = TRUE
)

phenotypes <- Biobase::pData(gse[[1]]) |> as_tibble() # formats GSE download as tabular sample metada
counts <- readxl::read_xlsx(counts_file) # function to import .xlsx file as data table

if (!dir.exists(file.path("data_processed"))) { # checks for extant data_processed directory
  dir.create(file.path("data_processed"))
}

readr::write_tsv(phenotypes, file.path("data_processed", "GSE107994_phenotypes.tsv"))
readr::write_tsv(counts, file.path("data_processed", "GSE107994_counts_raw.tsv"))
```

2 Data Inspection

(snapshot of counts and phenotypes tables,)