

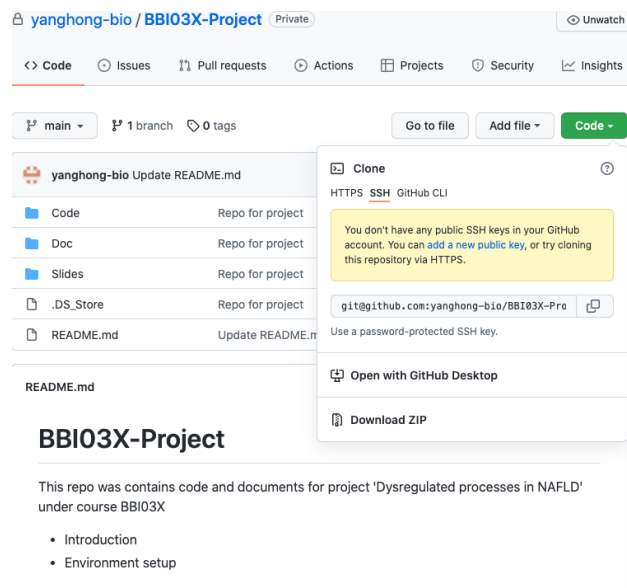
Dysregulated processes in NAFLD (BB103X)

Hong Yang

03/02/2022

Project materials

- download from Github (<https://github.com/yanghong-bio/BB103X-Project>)



Download metadata (1)

- Rstudio
- install package 'BiocManager' (only need once)

```
if (!require("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install()
```

- install other packages (only need once)

```
BiocManager::install(c("DESeq2", "pheatmap", "tidyverse", "xlsx", "readxl", "gplots", "ggbiplot",
  "piano", "venn", "clusterProfiler", "GEOquery"), ask = F, update = F)
```

```
## Warning: package 'ggbiplot' is not available for this version of R
##
## A version of this package for your version of R might be available elsewhere,
## see the ideas at
## https://cran.r-project.org/doc/manuals/r-patched/R-admin.html#Installing-packages
##
```

```
## The downloaded binary packages are in
## /var/folders/y8/hp0xlycj3bx97_g8n9zdfjpss22r3p/T//RtmpHxHqTB/downloaded_packages
```

Download metadata (2)

- Rstudio
- load packages (need each time you run your code when you come back from)

```
library(DESeq2)
library(pheatmap)
library(tidyverse)
library(xlsx)
library(readxl)
library(gplots)
library(ggbiplot)
library(piano)
library(venn)
library(clusterProfiler)
library(GEOquery)
library(openxlsx)
library(GEOquery)
```

Download metadata (2)

- Rstudio
- download datasets from GEO database

```
GSE135251 <- getGEO('GSE135251',GSEMatrix=TRUE)
GSE135251 <- GSE135251[[1]]
sampleInfo <- pData(GSE135251)
write.xlsx(sampleInfo, file = "./data/DGSE135251_metadata.xlsx")
sampleInfo[1:4,1:3]
```

```
##               title geo_accession      status
## GSM3998167  Liver patient 97      GSM3998167 Public on Dec 03 2020
## GSM3998168  Liver patient 98      GSM3998168 Public on Dec 03 2020
## GSM3998169  Liver patient 101     GSM3998169 Public on Dec 03 2020
## GSM3998170  Liver patient 102     GSM3998170 Public on Dec 03 2020
```

Load count data (pre-processed by Authors)

- Rstudio

```
# enter data folder
cd data

# download files
wget https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE135251&format=file -O GSE135251_RAW.tar

# create a new folder for count files
mkdir countfiles

# unzip raw data file and put all the files that it has into new folder
tar xzf GSE135251_RAW.tar -C countfiles

# enter the folder and unzip all files
```

```
cd countfiles
gunzip *.gz
```

```
##
```

```
## 2022-02-07 17:40:13 (256 KB/s) - 'GSE135251_RAW.tar' saved [45854720/45854720]
```

Save count data

```
result_dir = './data/countfiles/'
```

```
count_tr = ''
```

```
for(i in list.files(result_dir,recursive=F)){
  temp = read.csv(paste0(result_dir, i),sep='\t',stringsAsFactors = F)
  id = gsub("_.*", "",i)
```

```
  colnames(temp) = c("ID","Count")
  temp$Sample = id
  count_tr = rbind(count_tr,temp)
}
```

```
count_tr = count_tr[-1,]
```

```
count_tr_reshape = count_tr %>% spread(Sample, Count)
```

```
write.xlsx(count_tr_reshape, file = './data/DGSE135251_sample2count_str.xlsx')
```