Nuovi dati con normalizzazione TMM

Mattia Manna

2025-01-19

Indice

1 Importazione dati

1.1 Download dati

I dati sono stati scaricati dal Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/ , in particolare dalla pagina **GSE183635**, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183635.
e salvati localmente.

I dati sono stati trovati leggendo l'articolo: **Immunological Signatures for Early Detection of Human Head and Neck Squamous Cell Carcinoma through RNA Transcriptome Analysis of Blood Platelets**, https://www.mdpi.com/2072-6694/16/13/2399.

1.1.1 Scaricare informazioni sui pazienti

```
library(GEOquery)
library(readr)
# Scaricare informazioni riguardo ai geni
getGEOSuppFiles("GSE183635") # Check for available supplementary files

# Scaricare informazioni riguardanti i samples
patients <- getGEO('GSE183635',GSEMatrix=T)
patients<- pData(phenoData(patients[[1]]))
write_csv(patients,file = "/Users/mattia/Desktop/Università/Magistrale/Tesi/R/Risoluzione Data Leakage/
```

```
patients <- read_csv("/Users/mattia/Desktop/Università/Magistrale/Tesi/R/Risoluzione Data Leakage/Dati/
```

```
## Rows: 1646 Columns: 43
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (40): title, geo_accession, status, submission_date, last_update_date, t...
## dbl  (3): channel_count, taxid_ch1, data_row_count
##
## 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.
```

```
patients_train <- read.csv("~/Desktop/Università/Magistrale/Tesi/R/GSE68086/patients.csv", row.names=1,
dim(patients)
```

```
## [1] 1646   43
```

Controllare che i pazienti siano diversi nei due dataset.

```
IDs <- patients$geo_accession
IDs_train <- patients_train$geo_accession
intersect(IDs,IDs_train)
```

```
## character(0)
```

I pazienti sono totalmente differenti.

```r
table(patients$status)
##
## Public on Aug 05 2022
##                  1646

table(patients_train$status)
##
## Public on Oct 30 2015
##                   285
```

Anche le date confermano che si tratta di sample diversi.

1.1.2 Importare conte dei transcrittomi

```r
load("~/Desktop/Università/Magistrale/Tesi/R/Risoluzione Data Leakage/Dati/GSE183635_TEP_Count_Matrix.R

# Trasformare la count matrix in un dataframe
counts <- as.data.frame(TEP_Count_Matrix)
load("~/Desktop/Università/Magistrale/Tesi/R/Risoluzione Data Leakage/Dati/GSE183635_TEP_Count_Matrix_TS

# Trasformare la count matrix TS00 in un dataframe
countsTS00 <- as.data.frame(TEP_Count_Matrix_TS00)

# Rimuovere le matrici
rm(TEP_Count_Matrix_TS00,TEP_Count_Matrix)
```

1.2 Cleaning dati

1.2.1 Ricodifica nome pazienti del dataset delle conte

```
colnames(counts) <- gsub("^[0-9]+-", "", colnames(counts))
```

1.2.2 Estrazione dei pazienti per ogni patologia

Estrarre le patologie di interesse.
Che si ricordi essere:

- BRCA, breast cancer
- CRC, colorectal cancer
- GBM, glioblastoma multiforme
- HBC, hepatobiliarity
- NSCLC, non small cell lung cancer
- PAAD, Pancreatic adenocarcinoma

Visualizzare i tipi di cancro disponibili in questo database.

```
names(table(patients$`patient group:ch1`))
```

```
##  [1] "Angina Pectoris"            "Asymptomatic Controls"
##  [3] "Bowel Disease"              "Breast Cancer"
##  [5] "Cholangiocarcinoma"         "Colorectal Cancer"
##  [7] "Epilepsy"                   "Esophageal Cancer"
##  [9] "Former Sarcoma"             "Glioma"
## [11] "Head and Neck Cancer"       "Hodgkin Lymphoma"
## [13] "Melanoma"                   "Multiple Myeloma"
## [15] "Multiple Sclerosis"         "Non-Small-Cell Lung Cancer"
## [17] "nSTEMI"                     "Ovarian Cancer"
## [19] "Pancreatic Cancer"          "Pancreatic Disease"
## [21] "Prostate Cancer"            "Pulmonary Hypertension"
## [23] "Renal Cell Carcinoma"       "Sarcoma"
## [25] "Urothelial Carcinoma"
```

Sono tutti disponibili tranne l'HBC.

Si estraggano.

```
# BRCA, breast cancer
BRCA <- patients[patients$`patient group:ch1`=="Breast Cancer",]

# CRC, colorectal cancer
CRC <- patients[patients$`patient group:ch1`=="Colorectal Cancer",]

Glioma <- patients[patients$`patient group:ch1`=="Glioma",]
```

```r
# GBM, glioblastoma multiforme
GBM <- Glioma[grepl("\\bGBM\\b", Glioma$source_name_ch1), ]

# NSCLC, non small cell lung cancer
NSCLC <- patients[patients$`patient group:ch1`=="Non-Small-Cell Lung Cancer",]
NSCLC <- NSCLC[grepl("\\bNSCLC\\b", NSCLC$title), ]

# PAAD, Pancreatic adenocarcinoma
PAAD <- patients[patients$`patient group:ch1`=="Pancreatic Cancer",]

# Sani
HC <- patients[patients$`patient group:ch1`=="Asymptomatic Controls",]

# Metterli tutti insieme PANCANCER
ALL <-  rbind(BRCA,CRC,GBM,NSCLC,PAAD,HC)
```

2 Normalizzazione dati

3 Preparare dati per esportazione

```
conditions <- ALL[,c("title","patient group:ch1")]
conditions[1:5,]
```

```
## # A tibble: 5 x 2
##   title                 `patient group:ch1`
##   <chr>                 <chr>
## 1 MGH-BrCa-H76-TR469    Breast Cancer
## 2 MGH-BrCa-P28-TR499    Breast Cancer
## 3 MGH-BrCa-P35-TR620    Breast Cancer
## 4 Vumc-BRMETA-13-TR1451 Breast Cancer
## 5 Vumc-BRMETA-12-TR1450 Breast Cancer
```

```
dataframe.formato.classificazione <- function(dataframe){

  # Trasporre il dataframe delle conte di test, in questo modo si hanno le conte
  # disposte nel modo giusto per la classificazione: pazienti sulle righe e geni sulle colonne
  dataframe <- as.data.frame(t(dataframe))

  # Aggiungere le label y (cancer sano) alle conte
  dataframe <- merge(dataframe,conditions,by.x="row.names",
                                    by.y="title",all = FALSE)
  #print(dataframe)
  # Sistemare il dataframe dopo l'operazione di merge, rimettere i rownames al posto giusto
  rownames(dataframe) <- dataframe$Row.names



  # Eliminare la colonna rownames
  dataframe <- dataframe %>% dplyr::select(-"Row.names")
  dataframe$cancer.type.ch1 <- dataframe$`patient group:ch1`
  dataframe <- dataframe %>% dplyr::select(-"patient group:ch1")

  return(dataframe)
}
```

3.1 Creazione samples

3.1.1 GBM

```
# Prendere il nome pazienti GBM
pazienti.GBM <- GBM$title

# Prendere il nome dei pazienti SANI
pazienti.sani <- HC$title#[1:57]

# Estrarre dalle conte i pazienti GBM e sani
countsGBM <- counts[,colnames(counts) %in% c(pazienti.GBM,pazienti.sani)]
dim(countsGBM)
```

```
## [1] 5440  313
```

```
# Mettere nel formato adatto alla classificazione le conte
countsGBM <- dataframe.formato.classificazione(countsGBM)

# Estrarsi il vettore dei vecchi nomi
nuovi.nomi.patologia <- countsGBM$cancer.type.ch1

# Sostituire Glioma con GBM
nuovi.nomi.patologia <- gsub("Glioma", "GBM", nuovi.nomi.patologia)

# Sostituire Asymptomatic Controls con  HC
nuovi.nomi.patologia <- gsub("Asymptomatic Controls", "HC", nuovi.nomi.patologia)

# Assegnare la nuova nomenclatura
countsGBM$cancer.type.ch1 <- nuovi.nomi.patologia

table(countsGBM$cancer.type.ch1)
```

```
##
## GBM  HC
##  57 256
```

```
#write.csv(countsGBM,file="/Users/mattia/Desktop/Università/Magistrale/Tesi/Python TMM/Data/GBMtest.csv
```

3.1.2 BRCA

```
# Prendere il nome pazienti GBM
pazienti.BRCA <- BRCA$title

# Prendere il nome dei pazienti SANI
pazienti.sani <- HC$title#[1:77]

# Estrarre dalle conte i pazienti GBM e sani
countsBRCA <- counts[,colnames(counts) %in% c(pazienti.BRCA,pazienti.sani)]
dim(countsBRCA)
```

```
## [1] 5440  333
```

```
# Mettere nel formato adatto alla classificazione le conte
countsBRCA <- dataframe.formato.classificazione(countsBRCA)
dim(countsBRCA)
```

```
## [1]  333 5441
```

```
# Estrarsi il vettore dei vecchi nomi
nuovi.nomi.patologia <- countsBRCA$cancer.type.ch1

# Sostituire Glioma con GBM
nuovi.nomi.patologia <- gsub("Breast Cancer", "Breast", nuovi.nomi.patologia)

# Sostituire Asymptomatic Controls con  HC
nuovi.nomi.patologia <- gsub("Asymptomatic Controls", "HC", nuovi.nomi.patologia)

# Assegnare la nuova nomenclatura
countsBRCA$cancer.type.ch1 <- nuovi.nomi.patologia

table(countsBRCA$cancer.type.ch1)
```

```
##
## Breast     HC
##     77    256
```

```
#write.csv(countsBRCA,file="/Users/mattia/Desktop/Università/Magistrale/Tesi/Python TMM/Data/BRCAtest.c
```

3.1.3 CRC

```
# Prendere il nome pazienti GBM
pazienti.CRC <- CRC$title

# Prendere il nome dei pazienti SANI
pazienti.sani <- HC$title#[1:44]

# Estrarre dalle conte i pazienti GBM e sani
countsCRC <- counts[,colnames(counts) %in% c(pazienti.CRC,pazienti.sani)]
dim(countsCRC)
```

```
## [1] 5440  300
```

```
# Mettere nel formato adatto alla classificazione le conte
countsCRC <- dataframe.formato.classificazione(countsCRC)
dim(countsCRC)
```

```
## [1]  300 5441
```

```
# Estrarsi il vettore dei vecchi nomi
nuovi.nomi.patologia <- countsCRC$cancer.type.ch1

# Sostituire Glioma con GBM
nuovi.nomi.patologia <- gsub("Colorectal Cancer", "CRC", nuovi.nomi.patologia)

# Sostituire Asymptomatic Controls con  HC
nuovi.nomi.patologia <- gsub("Asymptomatic Controls", "HC", nuovi.nomi.patologia)

# Assegnare la nuova nomenclatura
countsCRC$cancer.type.ch1 <- nuovi.nomi.patologia

table(countsCRC$cancer.type.ch1)
```

```
##
## CRC  HC
##  44 256
```

```
#write.csv(countsCRC,file="/Users/mattia/Desktop/Università/Magistrale/Tesi/Python TMM/Data/CRCtest.csv
```

3.1.4 NSCLC

```
# Prendere il nome pazienti GBM
pazienti.NSCLC <- NSCLC$title#[1:256]

# Prendere il nome dei pazienti SANI
pazienti.sani <- HC$title#[1:44]

# Estrarre dalle conte i pazienti GBM e sani
countsNSCLC <- counts[,colnames(counts) %in% c(pazienti.NSCLC,pazienti.sani)]
dim(countsNSCLC)
```

```
## [1] 5440  688
```

```
# Mettere nel formato adatto alla classificazione le conte
countsNSCLC <- dataframe.formato.classificazione(countsNSCLC)
dim(countsNSCLC)
```

```
## [1]  688 5441
```

```
# Estrarsi il vettore dei vecchi nomi
nuovi.nomi.patologia <- countsNSCLC$cancer.type.ch1

# Sostituire Glioma con GBM
nuovi.nomi.patologia <- gsub("Non-Small-Cell Lung Cancer", "Lung", nuovi.nomi.patologia)

# Sostituire Asymptomatic Controls con  HC
nuovi.nomi.patologia <- gsub("Asymptomatic Controls", "HC", nuovi.nomi.patologia)

# Assegnare la nuova nomenclatura
countsNSCLC$cancer.type.ch1 <- nuovi.nomi.patologia

table(countsNSCLC$cancer.type.ch1)
```

```
##
##   HC Lung
##  256  432
```

```
#write.csv(countsNSCLC,file="/Users/mattia/Desktop/Università/Magistrale/Tesi/Python TMM/Data/NSCLCtest
```

3.1.5 PAAD

```
# Prendere il nome pazienti GBM
pazienti.PAAD <- PAAD$title

#set.seed(123) # Per rendere il risultato riproducibile
numeri_casuali <- sample(1:283, 86, replace = TRUE)

# Prendere il nome dei pazienti SANI
#pazienti.sani <- HC$title[1:86]
pazienti.sani <- HC$title#[numeri_casuali]

# Estrarre dalle conte i pazienti GBM e sani
countsPAAD <- counts[,colnames(counts) %in% c(pazienti.PAAD,pazienti.sani)]
dim(countsPAAD)
```

```
## [1] 5440  342
```

```
# Mettere nel formato adatto alla classificazione le conte
countsPAAD <- dataframe.formato.classificazione(countsPAAD)
dim(countsPAAD)
```

```
## [1]  342 5441
```

```
# Estrarsi il vettore dei vecchi nomi
nuovi.nomi.patologia <- countsPAAD$cancer.type.ch1

# Sostituire Glioma con GBM
nuovi.nomi.patologia <- gsub("Pancreatic Cancer", "Pancreas", nuovi.nomi.patologia)

# Sostituire Asymptomatic Controls con  HC
nuovi.nomi.patologia <- gsub("Asymptomatic Controls", "HC", nuovi.nomi.patologia)

# Assegnare la nuova nomenclatura
countsPAAD$cancer.type.ch1 <- nuovi.nomi.patologia

table(countsPAAD$cancer.type.ch1)
```

```
##
##       HC Pancreas
##      256       86
```

```
#write.csv(countsPAAD,file="/Users/mattia/Desktop/Università/Magistrale/Tesi/Python TMM/Data/PAADtest.c
```

3.1.6 PANCANCER

```
# Prendere il nome pazienti GBM
#pazienti.PANCANCER <- c(BRCA$title,CRC$title,GBM$title,NSCLC$title,PAAD$title)

pazienti.PANCANCER <- c(BRCA$title#[1:53]
                        ,CRC$title#[1:53]
                        ,GBM$title#[1:53]
                        ,NSCLC$title#[1:53]
                        ,PAAD$title#[1:53]
                        )

# Prendere il nome dei pazienti SANI
pazienti.sani <- HC$title#[1:256]

# Estrarre dalle conte i pazienti GBM e sani
countsPANCANCER <- counts[,colnames(counts) %in% c(pazienti.PANCANCER,pazienti.sani)]
dim(countsPANCANCER)
```

```
## [1] 5440  952
```

```
# Mettere nel formato adatto alla classificazione le conte
countsPANCANCER <- dataframe.formato.classificazione(countsPANCANCER)
dim(countsPANCANCER)
```

```
## [1]  952 5441
```

```
# Estrarsi il vettore dei vecchi nomi
nuovi.nomi.patologia <- countsPANCANCER$cancer.type.ch1

# Sostituire Glioma con GBM
nuovi.nomi.patologia <- gsub("Breast Cancer", "Cancer", nuovi.nomi.patologia)
nuovi.nomi.patologia <- gsub("Colorectal Cancer", "Cancer", nuovi.nomi.patologia)
nuovi.nomi.patologia <- gsub("Glioma", "Cancer", nuovi.nomi.patologia)
nuovi.nomi.patologia <- gsub("Non-Small-Cell Lung Cancer", "Cancer", nuovi.nomi.patologia)
nuovi.nomi.patologia <- gsub("Pancreatic Cancer", "Cancer", nuovi.nomi.patologia)

# Sostituire Asymptomatic Controls con  HC
nuovi.nomi.patologia <- gsub("Asymptomatic Controls", "HC", nuovi.nomi.patologia)

# Assegnare la nuova nomenclatura
countsPANCANCER$cancer.type.ch1 <- nuovi.nomi.patologia

table(countsPANCANCER$cancer.type.ch1)
```

```
##
## Cancer     HC
##    696    256
```

```
#write.csv(countsPANCANCER,file="/Users/mattia/Desktop/Università/Magistrale/Tesi/Python TMM/Data/PANCA
```