Norm_Imputation_strategy_selection

Fay

2023-09-28

Normalize or impute first?

Here is a document on selecting the best order for normalizationa and imputation of the immune gene expression data.

Layout:

1. Correlation of non-normalized and non-imputed gene expression data
2. Correlation of non-normalized and imputed gene expression data
3. Correlation of 1st normalized and sequentially imputed gene expression data
4. Corrrelation of 1st imputed and sequentially normalized gene expression data

Data input

Libraries

```
library(mice)
library(stringr)
library(gridExtra)
library(dplyr)
library(tidyverse)
library(tidyr)
library(janitor)
library(visdat)
library(corrplot)
library(RColorBrewer)
library(ggplot2)
```

Vectors for selecting genes

```
#Lab genes
# The measurements of IL.12 and IRG6 are done with an other assay and will
#ignore for now
Genes_v   <- c("IFNy", "CXCR3", "IL.6", "IL.13", "IL.10",
               "IL1RN","CASP1", "CXCL9", "IDO1", "IRGM1", "MPO",
               "MUC2", "MUC5AC", "MYD88", "NCR1", "PRF1", "RETNLB", "SOCS1",
               "TICAM1", "TNF") #"IL.12", "IRG6")
```

1. Correlation of non-normalized and non-imputed gene expression data

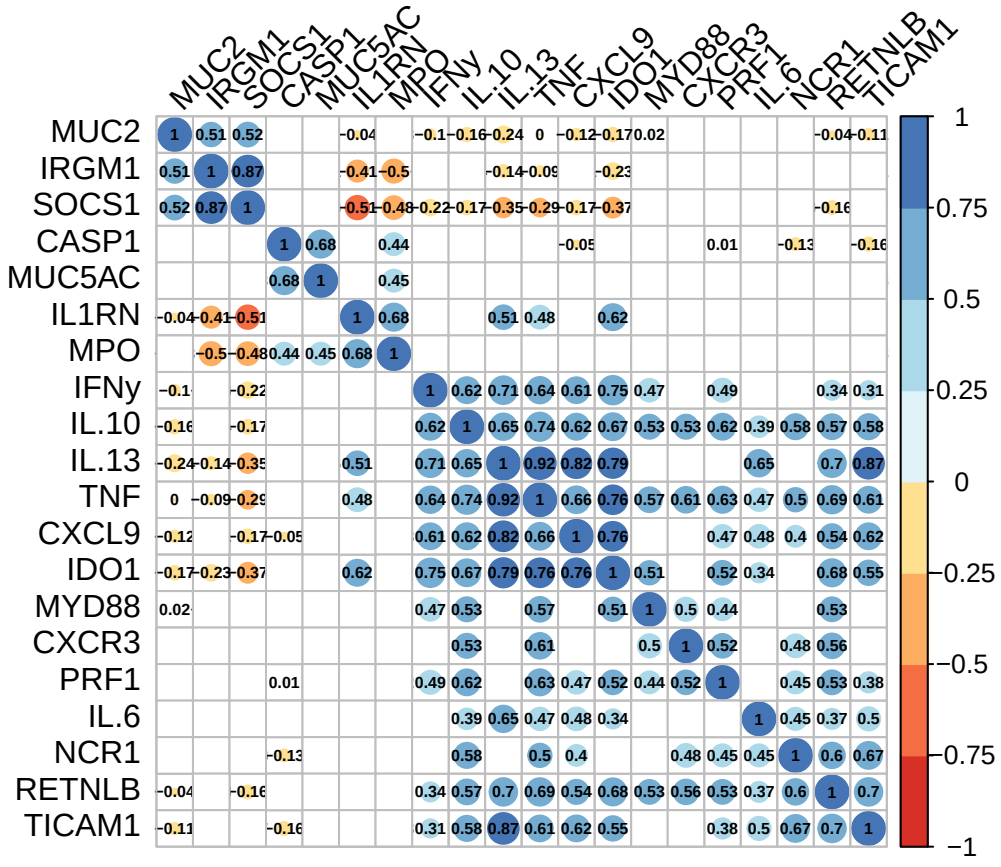
```
gene_correlation <- lab %>%
  filter(infection == "challenge", dpi == max_dpi) %>%
  ungroup() %>%
  dplyr::select(all_of(Genes_v))

# draw correlation between the genes
gene_correlation <- as.matrix(cor(gene_correlation,
                                  use="pairwise.complete.obs"))

# load the function to calculate the p value for correlations
source("R/Functions/p_value_for_correlations.R")

# matrix of the p-value of the correlatio
p.mat <- cor.mtest(gene_correlation)

corrplot(gene_correlation,
         method = "circle",  #method of the plot, "color" would show colour gradient
         tl.col = "black", tl.srt=45, #colour of labels and rotation
         col = brewer.pal(n = 8, name ="RdYlBu"), #colour of matrix
         order="hclust", #hclust reordering
         p.mat = p.mat, sig.level = 0.01, insig = "blank",
         addCoef.col = 'black',
         number.cex=0.5)
```



```
  #Add significance level to the correlogram
#remove the values that are insignificant
```

2. Correlation of non-normalized and imputed gene expression data
3. Correlation of 1st normalized and sequentially imputed gene expression data
4. Corrrelation of 1st imputed and sequentially normalized gene expression data