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 normalized data (no imputation)
4. Correlation of 1st normalized and sequentially imputed gene expression data
5. Corrrelation of 1st imputed and sequentially normalized gene expression data

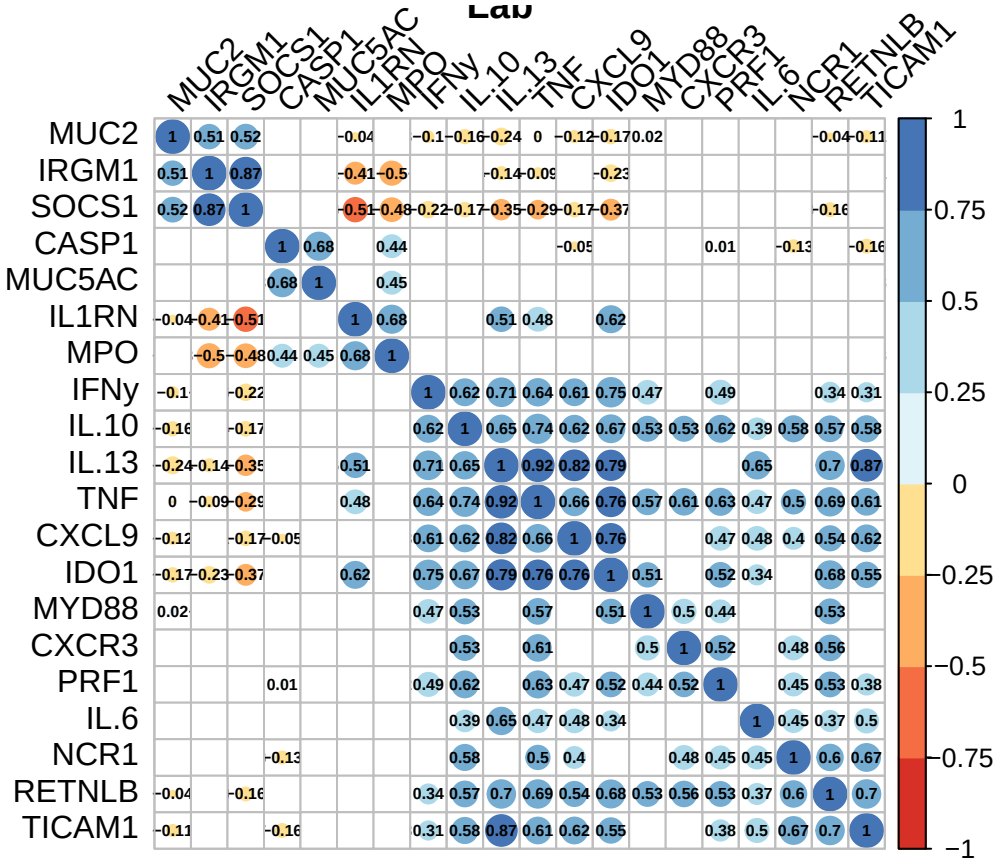
Data input

Libraries

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,
        title = "Lab")
```



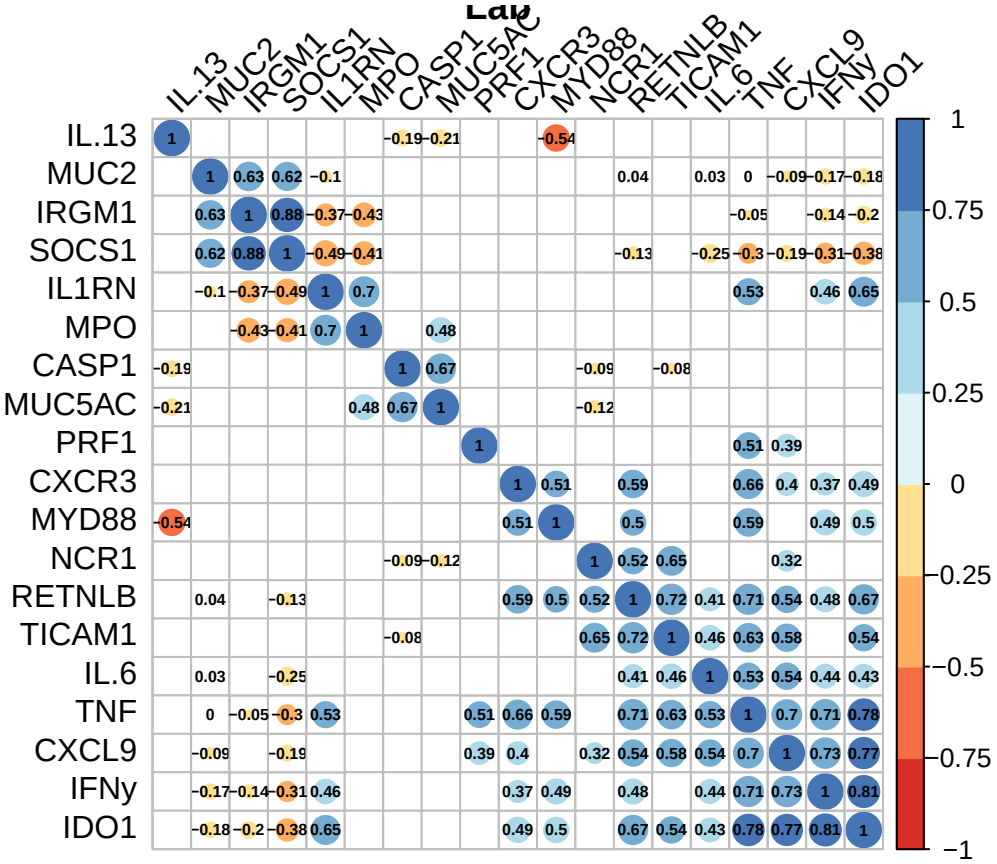

```
  #Add significance level to the correlogram
#remove the values that are insignificant
```

Correlation of non-normalized and non-imputed gene expression data, Results:

a. positive correlations between MUC2, IRGM1 and SOCS1
b. positive correlations between CASP1 and MUC5AC
c. positive correlations between IFNy, IL.10, IL.13, TNF, CXCL9, TICAM1 and IDO1
d. MYD88, PRF1, IL.6, NCR1, RETNLB, TICAM1 positve correlations with group c
e. negative correlations between group a (MUC2, IRGM1, SOCS1) and group c (IFNY, IL.10 etc)
f. negative correlations between IL1RN, MPO and group a
g. negative correlations between NCR1, RETNLB, TICAM1 and group a

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

```
lab <- hm_imp %>%
    filter(origin== "Lab")
gene_correlation <- lab[,Genes_v]

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

# 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,
         title = "Lab")
```




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

2. Correlation of non-normalized and imputed gene expression data, Results:

a. positive correlations between MUC2, IRGM1, SOCS1
b. positive correlations between IL1RN, MPO
c. positve correlations between CASP1, MUC5AC
d. positve correlations between CXCR3, MYD88
e. positive correlations between RETNLB, TICAM1, IL.6, TNF, CXCL9, IFNy, IDO1
f. positive correlations between IL1RN and TNF, IFN< and IDO1
g. negative correlations between IRGM1, SOCS2 and IL1RN, MPO
h. negative correlations between group MUC2, IRGM1, SOC1 vs IL.6, TNF, CXCL9, IFNy and IDO1
i. negative correlations between IL.13 and CASP1, MUC5AC and MYD88

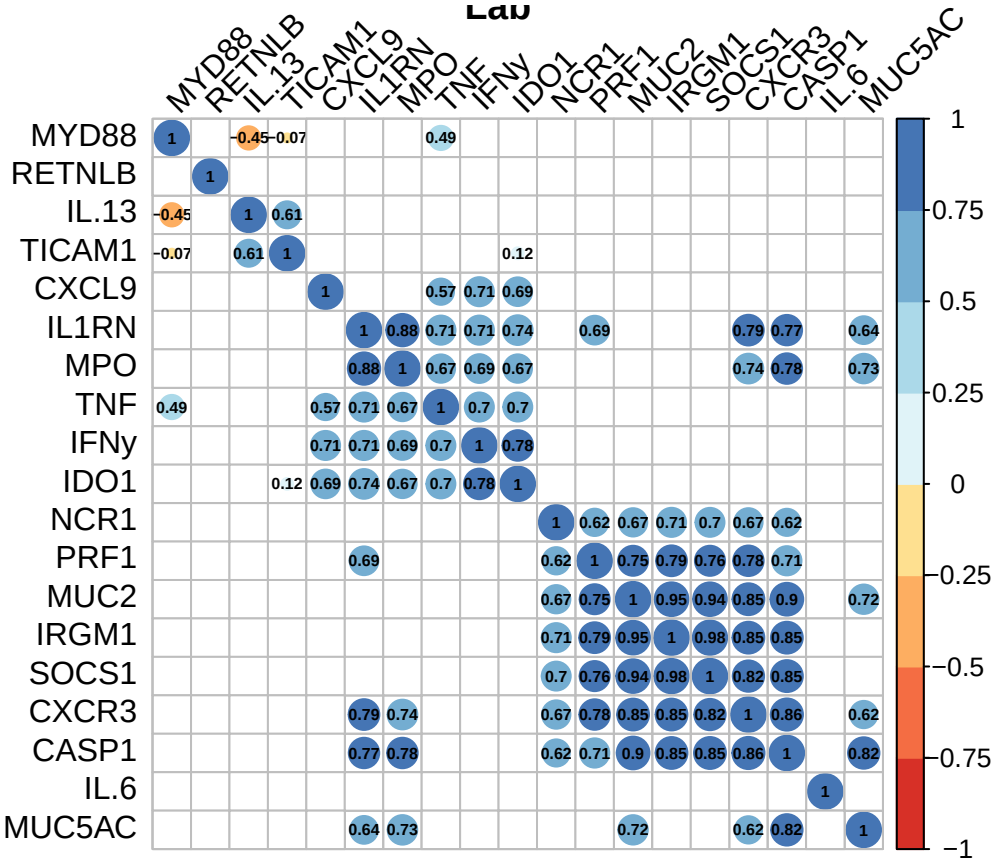
3. Correlation of normalized data (no imputation)

```
#DCt (sample) = Ct(reference gene) - Ct(gene of interest)

calculate_delta_ct <- function(df, HKG) {
  # Extract the column of the housekeeping gene
  reference_gene <- df[[HKG]]
  Mouse_ID <- df$Mouse_ID
  g <- df[, colnames(df) %in% Genes_v]

  delta_ct <- sapply(g, function(gene) reference_gene - gene)
  delta_ct <- as.data.frame(cbind(Mouse_ID, delta_ct))
  return(delta_ct)
}

# more positive = higher expression

# Use the function
norm_field <- calculate_delta_ct(field, "GAPDH")

norm_lab <- calculate_delta_ct(lab, "PPIB")

norm_g <- rbind(norm_field, norm_lab)

hm_norm <- hm %>%
    dplyr::select(-all_of(Genes_v)) %>%
    left_join(norm_g, by = "Mouse_ID")

rm(result, norm_g)
```

Correlations normalised genes (no imputation)

```
lab <- hm_norm%>%
    filter(origin== "Lab")

gene_correlation <- sapply(lab[,Genes_v], as.numeric)

# draw correlation between the genes
```

```r
gene_correlation <- as.matrix(cor(gene_correlation,
                                  use="pairwise.complete.obs"))

# 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,
         title = "Lab")
```




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

3. Correlation of normalized (not imputed) gene expression data, Results:

a. • correlation between IL.13 and TICAM1

b. • correlation between CXCL9 and TNF, IFNy and IDO1

c. • correlation between IL1RN, MPO, TNF, IFNy, IDO1

d. • correlation between NCR1, PRF1, MUC2, IRGM1, SOCS1, CXCR3, CASP1

e. • correlation between MUC5AC and CXCL9, IL1RN, MUC2, CXCR3, IL.16

f. • correlation between MYD88 and IL.13, TICAM1

4. Correlation of 1st normalized and sequentially imputed gene expression data

```
hm_genes <- hm_norm[,c("Mouse_ID", Genes_v)]


genes <- hm_genes[, -1]

#init <- mice(genes, maxit = 0)
```

Error in edit.setup(data, setup, ...) : `mice` detected constant and/or collinear variables. No predictors were left after their removal.

The treshold for colinearity in the package "MICE" is set to a max correlation of 0.99 by default.

The normalised gene expression values present a maximum correlation of

```
max(gene_correlation[gene_correlation < 1], na.rm = TRUE)
```

```
## [1] 0.9843948
```

between genes.

I can't run mice as genes become "too" correlated.

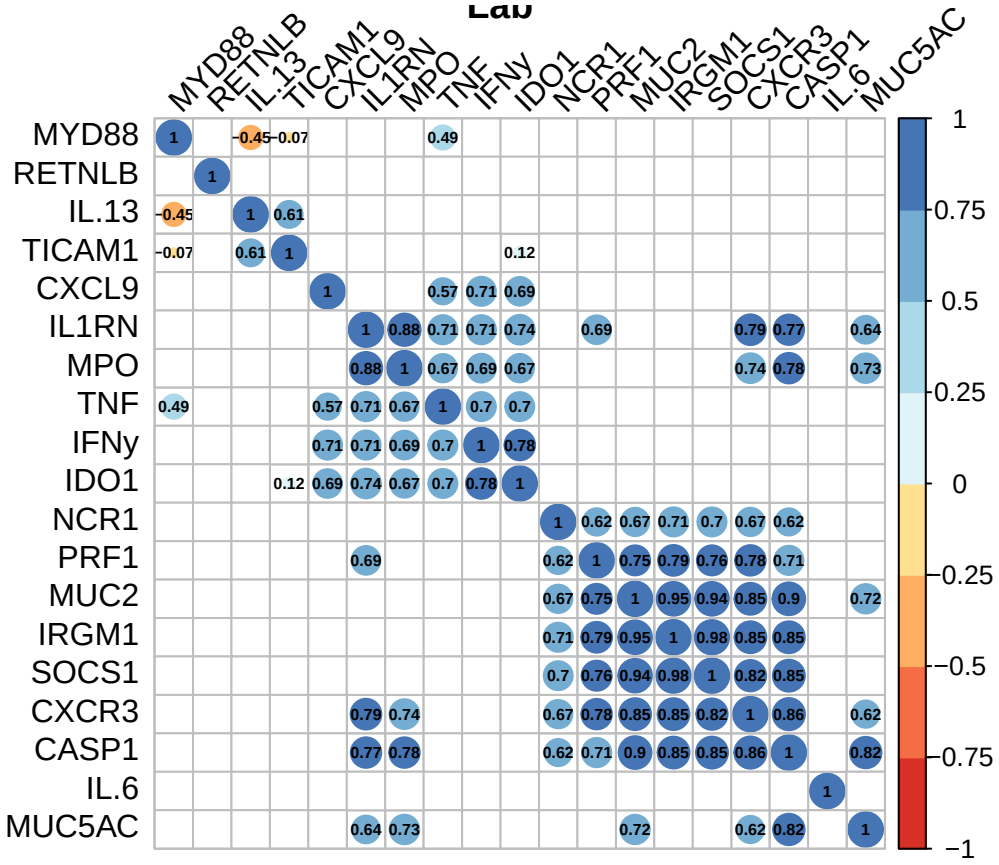
I can't get this to work: https://github.com/amices/mice/issues/278

5. Corrrelation of 1st imputed and sequentially normalized gene expression data

```
field <- hm_imp %>%
    dplyr::filter(origin == "Field")

lab <- hm_imp %>%
    dplyr::filter(origin == "Lab")

#DeltaCt (sample) = Ct(reference gene) - Ct(gene of interest)

# more positive = higher expression

# Use the function
norm_field <- calculate_delta_ct(field, "GAPDH")

norm_lab <- calculate_delta_ct(lab, "PPIB")

norm_g <- rbind(norm_field, norm_lab)
```

```
hm_norm <- hm %>%
    dplyr::select(-all_of(Genes_v)) %>%
    left_join(norm_g, by = "Mouse_ID")
```

```
lab <- hm_norm%>%
    filter(origin== "Lab")

gene_correlation <- sapply(lab[,Genes_v], as.numeric)

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

# 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,
         title = "Lab")
```



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