scatterplot

Anna Lisa und Wenjia

2024-02-15

Data preparation

```r
# here all three files that we need was read and stored in three data frames: heatshock, df_ATAC and TF
heatshock <- read.delim("complex_yeast_heatshock.tsv")
df_ATAC <- read.delim("ATACcounts_promotor_us500_ds100.tsv")
TFLink <- read.delim('TFLink_Saccharomyces_cerevisiae_interactions_SS_simpleFormat_v1.0.tsv')
TFLink$Name.Target<-mapIds(org.Sc.sgd.db, keys=TFLink$UniprotID.Target, column="ORF", keytype="UNIPROT"
```

```
## ’select()’ returned 1:many mapping between keys and columns
```

```r
# merge the data from heat shock, we only need wildtype, no KD or KO
names <- grep("Wildtype", names(heatshock), value = TRUE)
df_WT <- heatshock[, names]
rownames(df_WT) <- heatshock$gene_id  # add gene_id as rownames


# filter genes that both in ATAC data and heatshock data are
gene_ids <- intersect(rownames(df_WT), rownames(df_ATAC))
ATAC_data <- df_ATAC[, c(1:4, 9:16, 21:28)]
RNA_data <- df_WT[gene_ids,]
order <- order(colnames(RNA_data))
RNA_data <- RNA_data[,order]
# now we have our ATAC_data and RNA_data
# they will be used for following process of DEseq2

# get the subdata from TFLink with only targets of HSF1 and MSN24
TFtargets_HSF1 <- TFLink %>% filter(`Name.TF` == "HSF1")
TF_HSF1<-TFtargets_HSF1$Name.Target
TFtargets_MSN2 <- TFLink %>% filter(`Name.TF` == "MSN2")
TF_MSN2<-TFtargets_MSN2$Name.Target
TFtargets_MSN4 <- TFLink %>% filter(`Name.TF` == "MSN4")
TF_MSN4<-TFtargets_MSN4$Name.Target
TF_MSN24 <- intersect(TF_MSN2, TF_MSN4)
TF_HSF1_MSN24 <- intersect(TF_MSN24, TF_HSF1)
```

DESeq calculation

```r
# condition: 1 = 25/37_10, 2 = 25/37_30, 3 = 25/42_10, 4 = 25/42_30
# we have four different comparisions and therefor four different DESeq calculation

calculateDESeq <- function(ATAC_data, RNA_data, condition){
  # fold changes from ATAC_data
  # there are four replicates for each condition
  # if condition = 1, it is extracted from the ATAC_data the first four coloumns with values under 25 d
  # and the columns 5-8 as they provide the values under 37 degree 10 min
  # likewise, with condition = 2 we extract column 1-4 and 9-12, containing the values under 37 degree
  compare1 <- ATAC_data[, c(1:4, condition*4+1, condition*4+2, condition*4+3, condition*4+4)]
  coldata <- data.frame(condition = factor(c('control', 'control', 'control', 'control', 'treat', 'trea
  dds <- DESeqDataSetFromMatrix(countData = compare1, colData = coldata, design= ~condition)
  dds1 <- DESeq(dds, fitType = 'mean', minReplicatesForReplace = 7, parallel = FALSE)
  res <- results(dds1, contrast = c('condition', 'treat', 'control'))
  head(res)

  # fold changes from RNA_data (complex_yeast_heatshock)
  # it is basically the same as ATAC_data, only using 3 instead of 4 because there are only 3 replicate
  compare2 <- RNA_data[, c(1:3, condition*3+1, condition*3+2, condition*3+3)]
  coldata2 <- data.frame(condition = factor(c('control','control','control','treat', 'treat','treat'), 
  dds2 <- DESeqDataSetFromMatrix(countData = compare2, colData = coldata2, design= ~condition)
  dds3 <- DESeq(dds2, fitType = 'mean', minReplicatesForReplace = 7, parallel = FALSE)
  res2 <- results(dds3, contrast = c('condition', 'treat', 'control'))
  head(res2)

  # output of the DESeq data in files for further use
  res_df <- data.frame(res, stringsAsFactors = FALSE, check.names = FALSE)
  write.table(res_df, 'ATACData.DESeq2.txt', col.names = NA, sep = '\t', quote = FALSE)
  res_df2 <- data.frame(res2, stringsAsFactors = FALSE, check.names = FALSE)
  write.table(res_df2, 'RNAData.DESeq2.txt', col.names = NA, sep = '\t', quote = FALSE)
  res_ATAC <- res_df
  res_ATAC$gene_id <- rownames(res_ATAC)
  res_RNA <- res_df2
  res_RNA$gene_id <- rownames(res_RNA)

  # combine ATAC data and RNA data in one data frame and this data frame is the output of this function
  join_ATAC_RNA <- inner_join(res_ATAC, res_RNA, by= "gene_id")
  return(join_ATAC_RNA)
}

# getting the results of DESeq for all four conditions
df1 <- calculateDESeq(ATAC_data, RNA_data, 1)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates
## fitting model and testing
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```r
df2 <- calculateDESeq(ATAC_data, RNA_data, 2)
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```r
df3 <- calculateDESeq(ATAC_data, RNA_data, 3)
```

```
## estimating size factors
```

```
## estimating dispersions
```

```
## gene-wise dispersion estimates
## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```
df4 <- calculateDESeq(ATAC_data, RNA_data, 4)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```
# create a new column indicating the condition and combine all four data frames together
df1$condition <- "25-37_10"
df2$condition <- "25-37_30"
df3$condition <- "25-42_10"
df4$condition <- "25-42_30"
DEseqdf <- rbind(df1,df2,df3,df4)
head(DEseqdf)
```

```
##   baseMean.x log2FoldChange.x    lfcSE.x     stat.x    pvalue.x    padj.x
## 1   1124.157      -0.03247036 0.11258145 -0.2884166 0.77302786 0.9981890
## 2   1081.235       0.04122685 0.15341744  0.2687233 0.78814259 0.9981890
## 3   1656.641      -0.08249145 0.07299342 -1.1301218 0.25842491 0.8964312
## 4   1612.265       0.11945128 0.05946342  2.0088195 0.04455628 0.4145398
## 5   1670.778       0.05242060 0.05272500  0.9942266 0.32011255 0.9428576
## 6   1324.018       0.05959501 0.08623224  0.6910989 0.48950338 0.9981890
##     gene_id baseMean.y log2FoldChange.y  lfcSE.y      stat.y  pvalue.y
## 1   YDL248W 3.92895573        0.8399848 0.749346  1.12095721 0.2623061
## 2 YDL247W-A 0.00000000               NA       NA          NA        NA
## 3   YDL247W 0.00000000               NA       NA          NA        NA
## 4   YDL246C 0.00000000               NA       NA          NA        NA
## 5   YDL245C 0.00000000               NA       NA          NA        NA
## 6   YDL244W 0.08239827       -0.1099147 1.731548 -0.06347774 0.9493861
##      padj.y condition
## 1 0.3756183  25-37_10
## 2        NA  25-37_10
## 3        NA  25-37_10
## 4        NA  25-37_10
## 5        NA  25-37_10
## 6        NA  25-37_10
```

```
# Deseqdf will be used for plotting
```

functions for plotting

```
# inputs of the function are:
# DEseqdf: the dataframe that contains all the results from DESeq2 in all four conditions
# highlights: a list of gene ids, will be marked in the scatterplot with a different color
# targets: a name of this gene list, could be a pathway name
# highlights2 and targets2: likewise, they are the second list of genes that will be highlighted with a

makeplot2 <- function(DEseqdf, highlights, targets, highlights2, targets2){
  highlights <- DEseqdf %>% filter(gene_id %in% highlights)  # extract the subdata from DEseqdf with on
  highlights2 <- DEseqdf %>% filter(gene_id %in% highlights2) # extract the subdata from with DESeq onl
  title <- "foldchanges of the same gene from ATAC and heat shock data"
  ggplot(DEseqdf, aes(x = log2FoldChange.x, y = log2FoldChange.y), size = 0.5) +
      geom_point() +  # draw points of all the genes
      geom_point(data = highlights, aes(color = targets), size = 2) +  # mark the first hightlights
      geom_point(data = highlights2, aes(color = targets2), size = 2) +  # mark the second hightlights
      geom_hline(yintercept=0, linetype="dashed", color = "blue") +  # the dashed blue line of x = 0 an
      geom_vline(xintercept=0, linetype="dashed", color = "blue") +
      labs(x = "ATAC", y = "heat shock", title = title) +
      facet_wrap(vars(condition))  # wrap all four scatterplots of four conditions in one plot
}

# the same as makeplots2, but with three different kinds of highlights instead of two kinds of highligh
makeplot3 <- function(DEseqdf, highlights, targets, highlights2, targets2, highlights3, targets3){
  highlights <- DEseqdf %>% filter(gene_id %in% highlights)
  highlights2 <- DEseqdf %>% filter(gene_id %in% highlights2)
  highlights3 <- DEseqdf %>% filter(gene_id %in% highlights3)
  title <- "foldchanges of the same gene from ATAC and heat shock data"
  ggplot(DEseqdf, aes(x = log2FoldChange.x, y = log2FoldChange.y), size = 0.5) +
      geom_point() +
      geom_point(data = highlights, aes(color = targets), size = 2) +
      geom_point(data = highlights2, aes(color = targets2), size = 2) +
      geom_point(data = highlights3, aes(color = targets3), size = 2) +
      geom_hline(yintercept=0, linetype="dashed", color = "blue") +
      geom_vline(xintercept=0, linetype="dashed", color = "blue") +
      labs(x = "ATAC", y = "heat shock", title = title) +
      facet_wrap(vars(condition))
}
```

plotting

```
# first, we need to get our highlights from the results of Gene Enrichment
enrich_HSF1 <- read_csv("Enrichment_Results_for_Targets_HSF1.csv")  # read the file
```

```
## Rows: 10 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): PathwayID, Description, GeneRatio, GeneIDs
## dbl (3): pvalue, p.adjust, qvalue
##
## 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.
```

```
chaperone <- enrich_HSF1$GeneIDs[1]  # get the gene ids from first row of chaperone activities
list_of_chaperone <- unlist(strsplit(chaperone, split = "/"),use.names=FALSE) # split the ids
list_of_chaperone <- mapIds(org.Sc.sgd.db, keys=list_of_chaperone, column="ORF", keytype="ENTREZID", mul
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

```
# transfer ENTREZID to ORF

#likewise do it for common targets of HSF1 and MSN24
enrich_HSF1MSN24 <- read_csv("Enrichment_Results_for_Targets_HSF1MSN24.csv")
```

```
## Rows: 8 Columns: 7
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): PathwayID, Description, GeneRatio, GeneIDs
## dbl (3): pvalue, p.adjust, qvalue
##
## 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.
```

```
metabolism <- enrich_HSF1MSN24$GeneIDs[3]
list_of_metabolism <- unlist(strsplit(metabolism, split = "/"),use.names=FALSE)
list_of_metabolism <- mapIds(org.Sc.sgd.db, keys=list_of_metabolism, column="ORF", keytype="ENTREZID", m
## ’select()’ returned 1:1 mapping between keys and columns
```

```
heatstress2 <- enrich_HSF1MSN24$GeneIDs[5]
list_of_heatstress2 <- unlist(strsplit(heatstress2, split = "/"),use.names=FALSE)
list_of_heatstress2 <- mapIds(org.Sc.sgd.db, keys=list_of_heatstress2, column="ORF", keytype="ENTREZID"
```

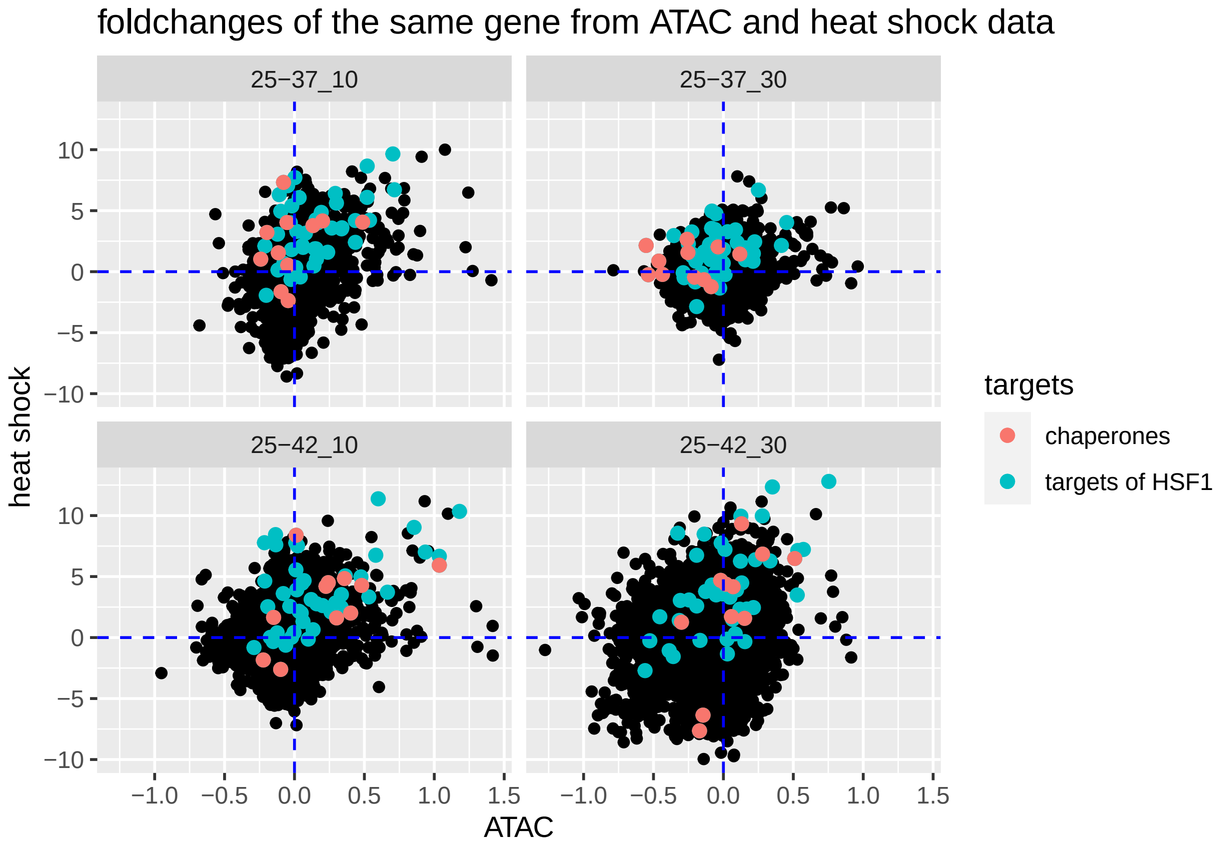
```
## ’select()’ returned 1:1 mapping between keys and columns
```

```
# why using makeplot2 for HSF1 and makeplot3 for HSF1_MSN24
# because on the scatterplot of HSF1 there are only two different highlights: targets of HSF1 and genes
# on the scatterplot of HSF1_MSN24 there are three highlights: common targets of HSF1 and MSN24, genes
# in one word, use makeplots2 when there are two kinds of highlights and makeplots3 when there are thre

makeplot2(DEseqdf, TF_HSF1, "targets of HSF1", list_of_chaperone, "chaperones")
```

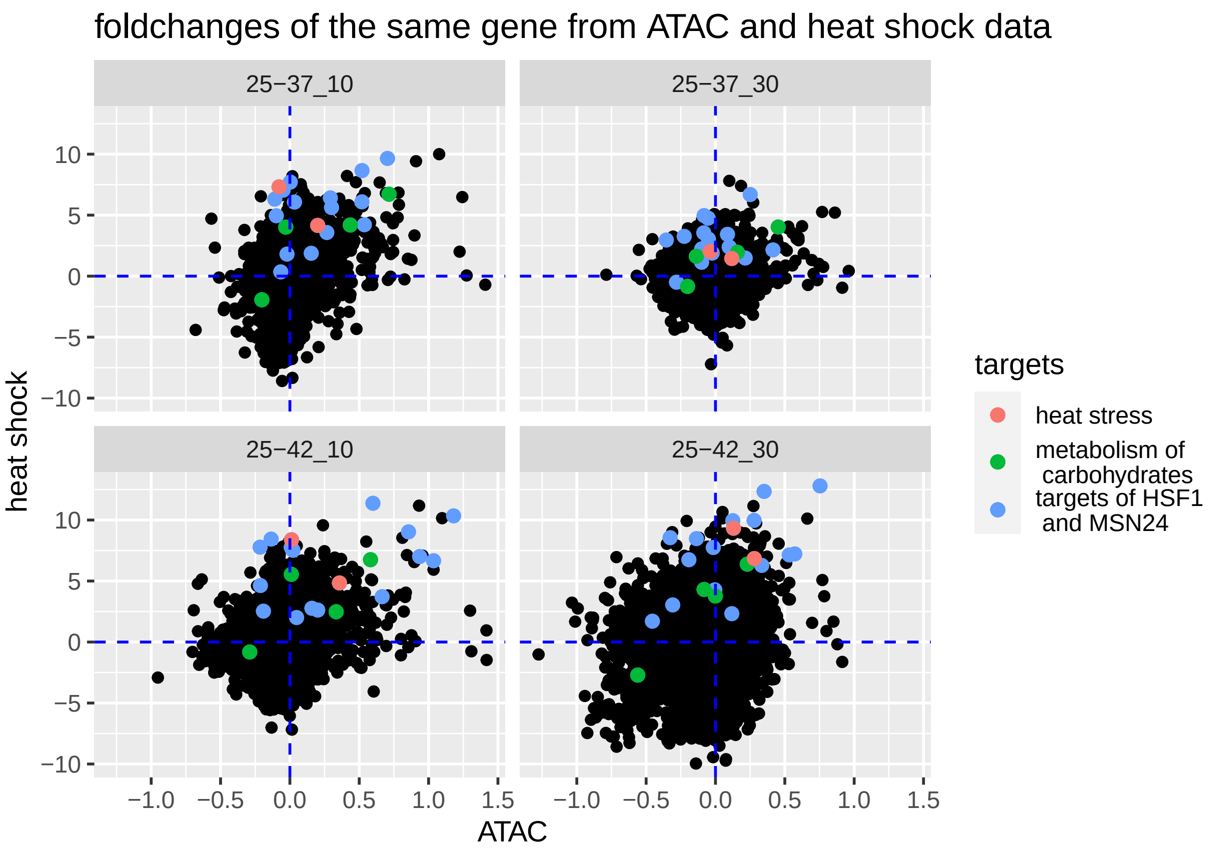
```
## Warning: Removed 2872 rows containing missing values (‘geom_point()‘).
```

```
## Warning: Removed 4 rows containing missing values (‘geom_point()‘).
```



```
makeplot3(DEseqdf, TF_HSF1_MSN24, "targets of HSF1 \n and MSN24", list_of_metabolism, "metabolism of \n
```

```
## Warning: Removed 2872 rows containing missing values (‘geom_point()‘).
```


foldchanges of the same gene from ATAC and heat shock data
25−37_10
25−37_30
25−42_10
25−42_30
heat shock
10
5
0
−5
−10
ATAC
−1.0
−0.5
0.0
0.5
1.0
1.5
targets
heat stress
metabolism of carbohydrates
targets of HSF1 and MSN24