Report decisions taken for exam_InTx_addTau.R

Load libraries and files.

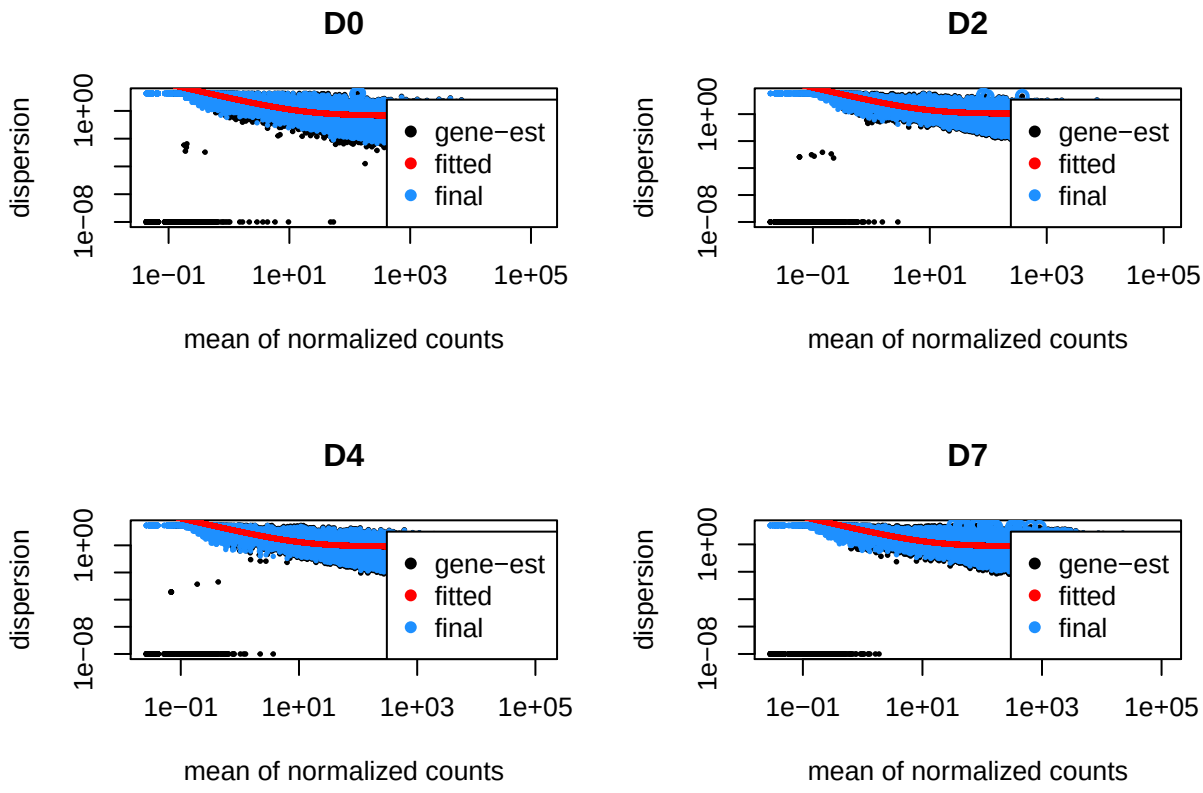
```
setwd("~/BulkAnalysis_plusNetwork/")
library(tidyverse)
library(DESeq2)
library(ggplot2)
library(cowplot)
odir = "exam_INTER_conditions/static/"

daysv = c("D0","D2","D4","D7")
fmat <- readRDS("data/prefiltered_counts.rds")
metadata <- readRDS("data/metadata.rds")
genes_df <- read.table("data/genesinfo.csv", sep="\t", header=T)
consensus_tau <- readRDS(paste0(odir,"conseTau4DEGs/", "Ext_conseTau.rds"))
```

Test diff expr Whole Day Matrices

First check **dispersions** :

```
par(mfrow=c(2,2))
for (d in daysv){
  all <- DESeqDataSetFromMatrix(countData = fmat[, str_detect(colnames(fmat),d)],
                                colData = metadata %>% filter(time==d),
                                design= ~age)
  all$age <- relevel(all$age, ref="Young")
  da <- estimateSizeFactors(all)
  da <- estimateDispersions(da)
  plotDispEsts(da, main=d)
}
```




```
par(mfrow=c(1,1))
```

Then Run analysis: For this report, only testing **D2** for fast checking:

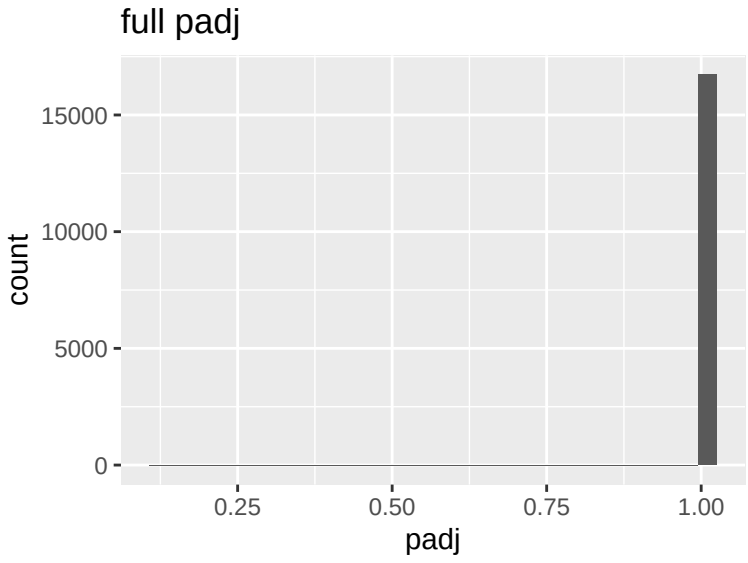
```
deresfull <- list()
d = "D2"
dmx <- fmat[, str_detect(colnames(fmat),d)]
dmt <- metadata %>% filter(time==d)
keep.d <- apply(dmx, 1, function(w) ifelse(sum(w >= 5)>=3,T,F))
dmx <- dmx[keep.d, ]
ddsm <- DESeqDataSetFromMatrix(countData=dmx, colData=dmt, design = ~age)
ddsm$age <- relevel(ddsm$age, ref="Young")
res <- DESeq(ddsm, full = ~age)
restab <- as_tibble(results(res))
restab$symbol = genes_df[match(rownames(res), genes_df$Geneid),]$symbol
deresfull[[d]] <- restab %>% filter(! is.na(padj))
```

```
print(summary(deresfull[["D2"]]$padj))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1125  0.9998  0.9998  0.9997  0.9998  0.9999
```

```
ggplot(data=deresfull[["D2"]],aes(padj)) + geom_histogram() + labs(title="full padj")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

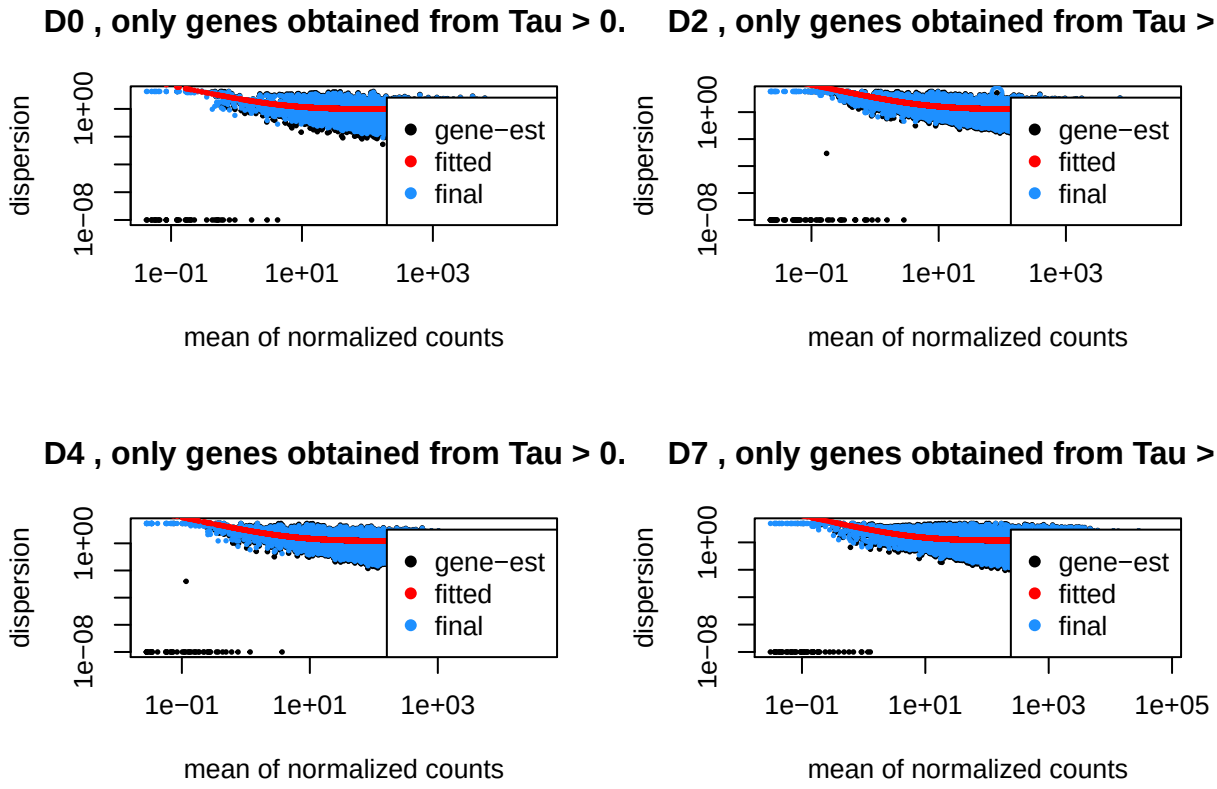


Padj values were very bad, all close to 1.

Test diff exp on Tau filtered matrix :

```
CUTOFFTAU <- 0.3
# extract only Tau > CUTOFFTAU and deduplicate with max tau
dedup <- list()
for (d in daysv){
  tf <- consensus_tau[[d]] %>% filter(Tau >= CUTOFFTAU) %>%
      group_by(symbol) %>% slice_max(Tau)
  dedup[[d]] <- tf
}
```

```
par(mfrow=c(2,2))
for (d in daysv){
  zo <- left_join(dedup[[d]], genes_df, by="symbol")
  tmpo <- DESeqDataSetFromMatrix(countData = fmat[rownames(fmat) %in% zo$Geneid,
                                                  str_detect(colnames(fmat),d)],
                                 colData = metadata %>% filter(time==d),
                                 design= ~age)
  tmpo$age <- relevel(tmpo$age, ref="Young")
  da <- estimateSizeFactors(tmpo)
  da <- estimateDispersions(da)
  plotDispEsts(da, main=paste(d, ", only genes obtained from Tau >", CUTOFFTAU))
  rm(tmpo)
}
```




```
par(mfrow=c(1,1))
```

Same as done for full gene list, test on **D2** for this filtered matrix:

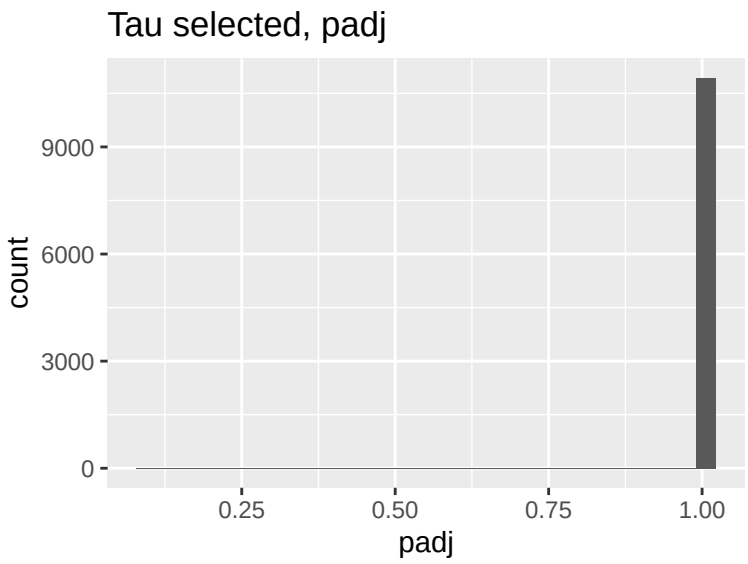
```
d = "D2"
resUniqMx <- list()
zo <- left_join(dedup[[d]], genes_df, by="symbol")
smx <- fmat[rownames(fmat) %in% zo$Geneid, str_detect(colnames(fmat),d)]
smm <- metadata %>% filter(time==d)
#  rownames for this day must be unique, set as symbols then:
if (length(unique(rownames(smx))) == length(rownames(smx))){
  print("ok, rownames are unique (ensemblid), setting rownames as symbols")
  chrw = zo[match(rownames(smx), zo$Geneid),]$symbol
  names(chrw) =  zo[match(rownames(smx), zo$Geneid),]$whichMAX
  rownames(smx) = unname(chrw)
}else{
  print("stop, rownames must be set unique!")
  stop()
}
keep.s <- apply(smx, 1, function(w) ifelse(sum(w >= 5)>= 3, T, F ))
smx <- smx[keep.s,]
ds.s <- DESeqDataSetFromMatrix(countData=smx, colData=smm, design = ~age)
colData(ds.s)$age <- factor(colData(ds.s)$age,
                            levels=c("Young","Old"))
res <- DESeq(ds.s, full = ~age)
restab <- as_tibble(results(res))
restab$symbol = rownames(results(res))
resUniqMx[[d]] <- restab %>% filter(! is.na(padj))
```

```
print(summary(resUniqMx[["D2"]]$padj))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08729 0.99996 0.99996 0.99974 0.99996 0.99996
```

```
ggplot(data=resUniqMx[["D2"]],aes(padj)) + geom_histogram() + labs(title="Tau selected, padj")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



IN CONCLUSION

Padj (BH method result for p values) are invariably extremely bad when entire day genes are tested for difference in expression at Young vs Old conditions.

In any case, non binary method GSEA will be tested to find Young vs Old pathway enrichment differences, which relies on ranked genes by their log2FoldChange, and NOT on each respective padj value.

http://rmarkdown.rstudio.com.