RNA-Seq exercise

MRC CSC Bioinformatics Core

10/March/2016

In this exercise we will read in a count table containing counts from RNAseq experiment from erythroblast differentiation in mice This data was downloaded from GEO database (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49843) and aligned to mm9 by Rsubread. More details with regards to this experiment please refer to (http://www.ncbi.nlm.nih.gov/pubmed/24092935). We will perform differential expression analysis and find genes that were changed in knockdown samples versus control.

- Material

(1) Sample description: Exercise_ShortRNAseq_sample.info

(2) Count data: Exercise_ShortRNAseq_counts.csv

- set up the working directory

```
# getwd()    # see your current directory
# setwd()    # set up your working directory
```

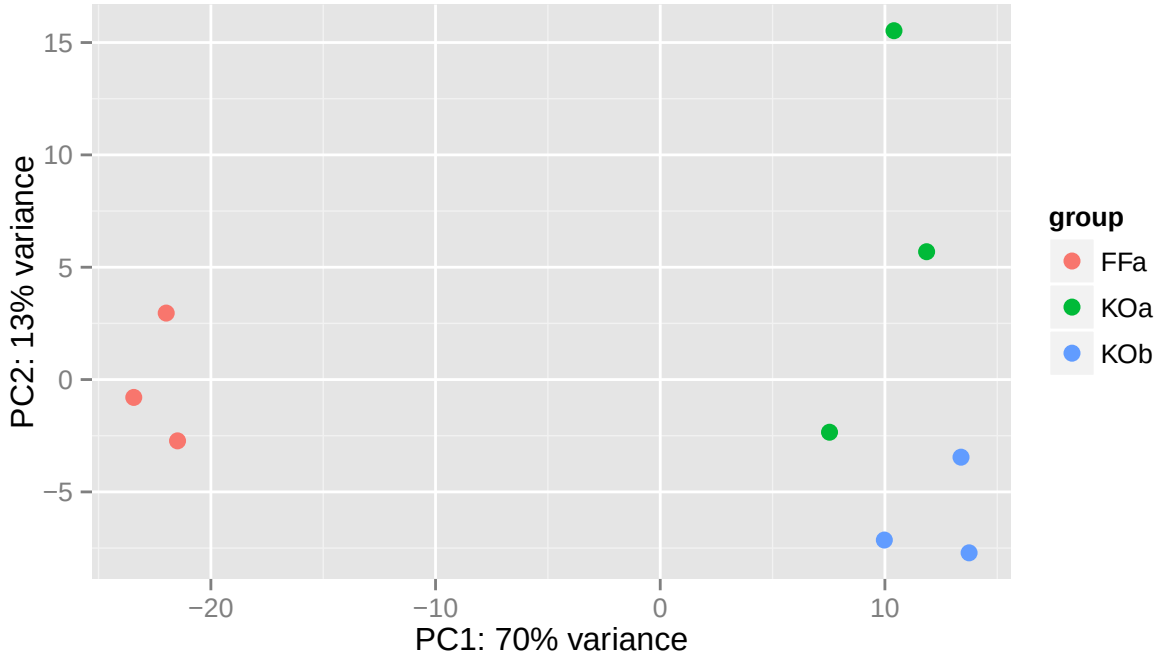
- First read in counts and the sample information.

```
suppressPackageStartupMessages(library(DESeq2))

suppressPackageStartupMessages(library(limma))


targets <- readTargets("Exercise_ShortRNAseq_sample.info")

AllCounts<-read.csv(file="Exercise_ShortRNAseq_counts.csv",header=T,row.names=1)

# see the what is in the counts.csv

head(AllCounts)
##           control_FFa1.bam control_FFa2.bam control_FFa3.bam
## 497097                  16                16                 0
## 100503874               20                 0                 0
## 100038431                0                 0                 2
## 19888                   11                 0                10
## 20671                   14                16                 0
## 27395                  465               193               596
##           mutant_KOa1.bam mutant_KOa2.bam mutant_KOa3.bam mutant_KOb1.bam
## 497097                 21              16              27              20
## 100503874              64               0               4               5
## 100038431               0               0               8               0
## 19888                 113               0              26              14
## 20671                  40               8              33              33
## 27395                 436             686             572            1378
```

```
##           mutant_KOb2.bam mutant_KOb3.bam
## 497097                  0               2
## 100503874               0               2
## 100038431               0               0
## 19888                   6              16
## 20671                  12              24
## 27395                1901            1553

# We provide entrez_id as identifier for this exercise

cData<-data.frame(name=targets$sample,condition=targets$condition,batch=targets$batch)

dds<-DESeqDataSetFromMatrix(countData= AllCounts,colData=cData,design=~batch+condition)
```

- Please perform the PCA analysis.

```
rld<-rlog(dds)

plotPCA(rld, intgroup="condition")
```



Now you have the count table as deseqdataset, you can start to perform the DE analysis.

- Find the number of genes that are changed in knockdown samples versus control at FDR 0.05 irrespective of fold change.
- Find the number of genes that are changed because of batch at FDR 0.05 irrespective of fold change.

```
dds<-DESeq(dds)
## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing

res1<-results(dds, contrast=c("condition","KOa","FFa"))

res2<-results(dds, contrast=c("condition","KOb","FFa"))

res3<-results(dds, contrast=c("batch","b","a"))

summary(res1,alpha=0.05)
##
## out of 24695 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 845, 3.4%
## LFC < 0 (down)   : 1015, 4.1%
## outliers [1]     : 0, 0%
## low counts [2]   : 7157, 29%
## (mean count < 9)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

- This time, use likelihood ratio test instead of the wald test.
- Find the number of genes that are changed because of condition at FDR 0.05 irrespective of fold change.
- Find the number of genes that are changed because of batch at FDR 0.05 irrespective of fold change.

```
ddsLRT<-DESeqDataSetFromMatrix(countData=AllCounts,colData=cData,design=~batch+condition)

# LRT analysis for the condition effect
ddsLRT_con <- DESeq(ddsLRT, test="LRT", full=~batch+condition, reduced=~batch)
## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing

resddsLRT_con<-results(ddsLRT_con)

summary(resddsLRT_con,alpha=0.05)
##
## out of 24695 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 1628, 6.6%
## LFC < 0 (down)   : 1567, 6.3%
## outliers [1]     : 0, 0%
## low counts [2]   : 7157, 29%
## (mean count < 9)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
resLRTorder<-resddsLRT_con[order(resddsLRT_con$padj),]

# LRT analysis for the batch effect

ddsLRT_batch <- DESeq(ddsLRT, test="LRT", full=~batch+condition, reduced=~condition)
## estimating size factors
## estimating dispersions
## gene-wise dispersion estimates
## mean-dispersion relationship
## final dispersion estimates
## fitting model and testing

resddsLRT_batch<-results(ddsLRT_batch)

summary(resddsLRT_batch,alpha=0.05)
##
## out of 24695 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)     : 31, 0.13%
## LFC < 0 (down)   : 38, 0.15%
## outliers [1]     : 0, 0%
## low counts [2]   : 4295, 17%
## (mean count < 4)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Using the genes that are changed because of condition at FDR 0.05 irrespective of fold change as our differentially expressed genes, perform the Gene Ontology and Pathway Enrichment Analysis.

```
suppressPackageStartupMessages(library(KEGG.db))
suppressPackageStartupMessages(library(goseq))

# remove the NAs

resdat<- resLRTorder[complete.cases(resLRTorder$padj),]

degenes<-as.integer(resdat$padj<0.05)
names(degenes)<-rownames(resdat)

# remove duplicate gene names
degenes<-degenes[match(unique(names(degenes)),names(degenes))]
table(degenes)
```

```
## degenes
##     0     1
## 14343  3195
```

```
# Fitting the probability weighting function (PWF)
# note, we use Entrez Gene ID as identifiler for this exercise
# we need to choose the correct "id" for the nullp function
# more details see
?nullp
```

```
pwf=nullp(degenes,'mm9','knownGene', plot.fit=FALSE)
```

```
## Loading mm9 length data...
```

```
# Calculate the over and under expressed GO categories among DE genes
go<-goseq(pwf,'mm9','knownGene', test.cats=c("GO:BP","GO:MF","KEGG"))
```

```
## Fetching GO annotations...
## For 919 genes, we could not find any categories. These genes will be excluded.
## To force their use, please run with use_genes_without_cat=TRUE (see documentation).
## This was the default behavior for version 1.15.1 and earlier.
## Calculating the p-values...
## 'select()' returned 1:1 mapping between keys and columns
```

Change the Keggpath id to name in the goseq output

```
# function that converts KEGG id to KEGG description
xx <- as.list(KEGGPATHID2NAME)
temp <- cbind(names(xx),unlist(xx))

addKeggTogoseq <- function(JX,temp){
  for(l in 1:nrow(JX)){
      if(JX[l,1] %in% temp[,1]){
          JX[l,"term"] <- temp[temp[,1] %in% JX[l,1],2]
          JX[l,"ontology"] <- "KEGG"
      }
  }
  return(JX)
}

restemp<-addKeggTogoseq(go,temp)

head(restemp)
```

```
##        category over_represented_pvalue under_represented_pvalue
## 839  GO:0002376           1.624638e-21                         1
## 9315 GO:0050896           3.088017e-13                         1
## 963  GO:0002682           9.109990e-13                         1
## 8039 GO:0045321           4.658370e-12                         1
## 8563 GO:0046649           8.493103e-12                         1
## 8931 GO:0048534           1.574758e-11                         1
##      numDEInCat numInCat                                      term
## 839         245      741                    immune system process
## 9315        497     2039                     response to stimulus
## 963         113      319      regulation of immune system process
## 8039        109      315                      leukocyte activation
## 8563         97      272                    lymphocyte activation
## 8931        110      326 hematopoietic or lymphoid organ development
##      ontology
## 839        BP
## 9315       BP
```

```
## 963        BP
## 8039       BP
## 8563       BP
## 8931       BP
```

```
# save the goseq result
write.csv(restemp,file="Exercise_ShortRNAseq_GO_Kegg_Wallenius.csv", row.names=F)
```