Deseq2 with gene lists walkthrough

Author: Mary Allen

Step 1) Copy our scripts

Log in to the super computer
Make a directory /scratch/User/yourusername/day7/
Cd into the new directory
Copy the two scripts we have provided in
/scratch/Shares/public/sread2021/scripts/day10/Deseq2_to_go/

```
[maallen3@ip-172-31-38-192 day10]$ mkdir -p /scratch/Users/maallen3/day10/
[maallen3@ip-172-31-38-192 day10]$ scp /scratch/Shares/public/sread2021/scripts/
day10/Deseq2_to_go /scratch/Users/maallen3/day10/
cp: omitting directory '/scratch/Shares/public/sread2021/scripts/day10/Deseq2_to
_go'
[maallen3@ip-172-31-38-192 day10]$ scp -r /scratch/Shares/public/sread2021/scrip
ts/day10/Deseq2_to_go /scratch/Users/maallen3/day10/
[maallen3@ip-172-31-38-192 day10]$
```

Step 2) Edit Our scripts

In the R script Change your working directory
In the sbatch script change your email and error and output files
####Make sure the error and output directory exist before you run!!!!!!

```
[maallen3@ip-172-31-38-192 Deseq2_to_go]$ vi DESeq2_example_withgenelists.R
[maallen3@ip-172-31-38-192 Deseq2_to_go]$ vi sr_deseq2.sbatch
```

```
#!/bin/bash
#SBATCH --job-name=deseq                                    # Job name
#SBATCH --mail-type=ALL                          # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=username@colorado.edu                   # Where to send mail
#SBATCH --nodes=1                                           # Number of cores job will run on
#SBATCH --ntasks=4                                          # Number of CPU (processers, tasks)
#SBATCH --time=2:00:00                                      # Time limit hrs:min:sec
#SBATCH --partition compute                                 # Job queue
#SBATCH --mem=4gb                                           # Memory limit
#SBATCH --output=/scratch/Users/username/eofiles/%x_%j.out
#SBATCH --error=/scratch/Users/username/eofiles/%x_%j.err
```

becomes

```
#!/bin/bash
#SBATCH --job-name=deseq                                    # Job name
#SBATCH --mail-type=ALL                           # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=allenma@colorado.edu                    # Where to send mail
#SBATCH --nodes=1                                           # Number of cores job will run on
#SBATCH --ntasks=4                                          # Number of CPU (processers, tasks)
#SBATCH --time=2:00:00                                      # Time limit hrs:min:sec
#SBATCH --partition compute                                 # Job queue
#SBATCH --mem=4gb                                           # Memory limit
#SBATCH --output=/scratch/Users/maallen3/eofiles/%x_%j.out
#SBATCH --error=/scratch/Users/maallen3/eofiles/%x_%j.err
```

```
# For more information see: http://www.bioconductor.org/help/workflows/rnaseqGene/
library("tidyverse")
library("DESeq2")
#set working dir
workdir <- '/scratch/Users/username/day10/Deseq2_to_go/'
dir.create(workdir, showWarnings = FALSE)
setwd(workdir)
getwd()
outdir <- paste(workdir, 'deseqresults', '/', sep='') ##naming our outdir
dir.create(outdir, showWarnings = FALSE) ###creating the directory
```

becomes

```
# For more information see: http://www.bioconductor.org/help/workflows/rnaseqGene/
library("tidyverse")
library("DESeq2")
#set working dir
workdir <- '/scratch/Users/maallen3/day10/Deseq2_to_go/'
dir.create(workdir, showWarnings = FALSE)
setwd(workdir)
getwd()
outdir <- paste(workdir, 'deseqresults', '/', sep='') ##naming our outdir
dir.create(outdir, showWarnings = FALSE) ###creating the directory
```

The Deseq2 script should like about like this…

```
# For more information see: http://www.bioconductor.org/help/workflows/rnaseqGene/
library("tidyverse")
library("DESeq2")
#set working dir
workdir <- '/scratch/Users/username/day10/Deseq2_to_go/'
dir.create(workdir, showWarnings = FALSE)
setwd(workdir)
getwd()
outdir <- paste(workdir, 'deseqresults', '/', sep='') ##naming our outdir
dir.create(outdir, showWarnings = FALSE) ###creating the directory
```

```
counts <- read.csv("/scratch/Shares/public/sread2021/cookingShow/day8/RNAseqextras/counts/featureCounts.txt", row.names=1, sep="\t")
head(counts) #your rowname should be the gene ids. Your colnames should match some column of your metadata table (in this case filetable)
filetable <- read.csv('/scratch/Shares/public/sread2021/cookingShow/day8/RNAseqextras/meta.txt', sep="\t")
head(counts)
filetable$chr21 <- factor(filetable$chr21)
filetable$bamfiles <- paste0(filetable$Run, ".sorted.bam") #making a column the files
filelist<- filetable$bamfiles #creating a vector that is the file list

#Your metatdata columns and your counts rows must be in the same order!!!!!!
counts <- counts %>% select(as.vector(filetable$bamfiles))

# Generate DESeqDataSet from count matrix generated by featureCounts
ddsMat <- DESeqDataSetFromMatrix(countData = counts, colData = filetable, design=~chr21)
dds <- ddsMat


### Run DESeq on the DESeqDataSet object
DEdds <- DESeq(dds)

### output the results for a specified alpha value
alpha_val <- 0.05
comparison <- c("chr21", "Disomic", "Trisomic")
res <- results(DEdds, alpha = alpha_val, contrast = comparison)

res_shrink <- lfcShrink(DEdds, contrast = comparison, res = res)

### MA plot
name <- "MA_tri_vs_ctrl_DEA"
limits <- c(-10, 10)
pdf(paste0(outdir, name, ".pdf"))
maplot <- plotMA(res_shrink, main="Disomic vs Trisomic", alpha=alpha_val, ylim=limits)
dev.off()
```

```
### disp plot
name <- "disp_tri_vs_ctrl_DEA"
limits <- c(-10, 10)
pdf(paste0(outdir, name, ".pdf"))
maplot <- plotDispEsts(DEdds, main="Disomic vs Trisomic")
dev.off()

#### sort by sig
res_shrink<- res_shrink[ order( res_shrink$padj ), ]

### Take subset of results that are significant
res_shrink_Sig <- subset(res_shrink, padj < alpha_val)


write.csv(res_shrink, file = paste0(outdir,"all_results.csv"))
write.csv(res_shrink_Sig, file = paste0(outdir,"sig_results.csv"))


#for go and enricher and gsea
res_shrink_expressed <- as.data.frame(res_shrink)
res_shrink_expressed <- res_shrink_expressed[!is.na(res_shrink_expressed$padj),]
write.csv(rownames(res_shrink_expressed), file = paste0(outdir,"backgroundgenes.csv"),row.names = FALSE, col.names = FALSE, quote = FALSE)
write.csv(rownames(res_shrink_Sig), file = paste0(outdir,"siggenes.csv"),row.names = FALSE, col.names = FALSE, quote = FALSE)

rnkdf <- tibble(gene = rownames(res_shrink),
                                rnk = -log(res$pvalue) * sign(res$log2FoldChange)) %>%
        arrange(desc(rnk)) %>% drop_na()

## Write out the table without any additional information
write.table(rnkdf, file = paste0(outdir,"deseq_res_for_gsea.rnk"),
                        append = FALSE, col.names = FALSE, row.names = FALSE,
                        quote = FALSE, sep = "\t")

```

Step 3) Submit the sbatch script to the queue

The sbatch script runs the R script… how?

```
[maallen3@ip-172-31-38-192 day7]$ sbatch sr_deseq2.sbatch
```

Look at the number of genes in each csv

Step 4) If it works you will end up with a directory named deseqresults in your working directory. In the deseqresults directory you will end up with many files.

```
[maallen3@ip-172-31-38-192 deseqresults]$ ls -lahtr
total 4.5M
drwxrwxr-x 2 maallen3 maallen3 6.0K Jul 29 20:39 .
-rw-rw-r-- 1 maallen3 maallen3 152K Jul 29 21:00 MA_tri_vs_ctrl_DEA.pdf
-rw-rw-r-- 1 maallen3 maallen3 459K Jul 29 21:00 disp_tri_vs_ctrl_DEA.pdf
-rw-rw-r-- 1 maallen3 maallen3 3.0M Jul 29 21:00 all_results.csv
-rw-rw-r-- 1 maallen3 maallen3  423 Jul 29 21:00 sig_results.csv
-rw-rw-r-- 1 maallen3 maallen3 186K Jul 29 21:00 backgroundgenes.csv
-rw-rw-r-- 1 maallen3 maallen3   25 Jul 29 21:00 siggenes.csv
-rw-rw-r-- 1 maallen3 maallen3 633K Jul 29 21:00 deseq_res_for_gsea.rnk
drwxrwxr-x 3 maallen3 maallen3 6.0K Jul 29 21:26 ..
[maallen3@ip-172-31-38-192 deseqresults]$ wc -l all_results.csv
33122 all_results.csv
[maallen3@ip-172-31-38-192 deseqresults]$ wc -l backgroundgenes.csv
24544 backgroundgenes.csv
[maallen3@ip-172-31-38-192 deseqresults]$ wc -l siggenes.csv
4 siggenes.csv
```

Why are there less background genes than all_results genes?