Transcriptome Demo

Tricia

2025-04-28

Load required packages (you might have to figure out how to install some of these first. . .)

```
library(ballgown)
library(RColorBrewer)
library(genefilter)
library(dplyr)
library(devtools)
```

This code is creatse a data frame with two columns: ids and stage.

```
pheno_data<-data.frame(ids = c("plank01", "plank02", "biofilm01", "biofilm02"),
                       stage = c("planktonic", "planktonic", "biofilm", "biofilm"))
```

create Ballgown object and check transcript number

```
samples.c <- paste('ballgown', pheno_data$ids, sep = '/')
bg <- ballgown(samples = samples.c, meas='all', pData = pheno_data)
bg
```

```
## ballgown instance with 5727 transcripts and 4 samples
```

This code is filters the Ballgown object to keep only genes or transcripts with expression variance > 1 across samples.

```
bg_filt = subset(bg,"rowVars(texpr(bg)) >1",genomesubset=TRUE)
bg_filt
```

```
## ballgown instance with 4743 transcripts and 4 samples
```

create a table of transcripts

```
results_transcripts<- stattest(bg_filt, feature = "transcript", covariate = "stage",
getFC = TRUE, meas = "FPKM")
results_transcripts<-data.frame(geneNames=geneNames(bg_filt),
transcriptNames=transcriptNames(bg_filt), results_transcripts)
```

The transcript I chose was gene-PA0001.

```
results_transcripts[results_transcripts$transcriptNames == "gene-PA0001", ]
```

```
##   geneNames transcriptNames    feature id        fc      pval      qval
## 1      dnaA     gene-PA0001 transcript  1 0.9862819 0.6906561 0.6906561
```

The transcript gene-PA0001 is for dnaA. It's id number is 1 and it's fc is 0.9862819. This repesents the how much the expression level of a transcript/gene changes between conditions.The pval and qval are equal at 0.6906561 which means there is no signifance.

This code filters for significantly differentially expressed transcripts and returns how many of them were found in a table format.

```
sigdiff <- results_transcripts %>% filter(pval<0.7)
dim(sigdiff)
```

```
## [1] 4743    7
```

organize the table

```
o = order(sigdiff[,"pval"], -abs(sigdiff[,"fc"]), decreasing=FALSE)
output = sigdiff[o,c("geneNames","transcriptNames", "id","fc","pval","qval")]
write.table(output, file="SigDiff.txt", sep="\t", row.names=FALSE, quote=FALSE)
head(output)
```

```
##      geneNames transcriptNames   id       fc      pval      qval
## 1693         .     gene-PA1660 1693 1.000000 0.6906561 0.6906561
## 950       gacS     gene-PA0928  950 1.000017 0.6906561 0.6906561
## 269          .     gene-PA0266  269 1.000016 0.6906561 0.6906561
## 3176      purF     gene-PA3108 3176 1.000105 0.6906561 0.6906561
## 393          .     gene-PA0389  393 1.000138 0.6906561 0.6906561
## 3825      dsbC     gene-PA3737 3825 1.000189 0.6906561 0.6906561
```

load gene names

```
bg_table = texpr(bg_filt, 'all')
bg_gene_names = unique(bg_table[, 9:10])
```

pull out gene expression data and visualize

```
gene_expression = as.data.frame(gexpr(bg_filt))
head(gene_expression)
```

```
##                FPKM.plank01 FPKM.plank02 FPKM.biofilm01 FPKM.biofilm02
## .                 1.2848601    0.7715227   3.155843e+00      1.2848601
## gene-PA1781.1     0.2154666    0.4325117   4.788977e-02      0.2154666
## MSTRG.1        2347.2883407 2316.1708024   1.817852e+03   2347.2883407
```

```
## MSTRG.10        14.6400744   12.8207084   9.278507e+00     14.6400744
## MSTRG.100       82.4897736   66.6706324   4.423373e+02     82.4897736
## MSTRG.1000       5.7265984    4.4001802   1.777665e+01      5.7265984
```

<This code creates a table that is a cleaned up version of table 1. It contains simplified column names. >

```
colnames(gene_expression) <- c("plank01", "plank02", "biofilm01", "biofilm02")
head(gene_expression)
```

```
##                   plank01      plank02    biofilm01    biofilm02
## .                1.2848601    0.7715227 3.155843e+00    1.2848601
## gene-PA1781.1    0.2154666    0.4325117 4.788977e-02    0.2154666
## MSTRG.1       2347.2883407 2316.1708024 1.817852e+03 2347.2883407
## MSTRG.10        14.6400744   12.8207084 9.278507e+00   14.6400744
## MSTRG.100       82.4897736   66.6706324 4.423373e+02   82.4897736
## MSTRG.1000       5.7265984    4.4001802 1.777665e+01    5.7265984
```

```
dim(gene_expression)
```

```
## [1] 4311    4
```

There are five unique genes and 6 transcripts.

```
transcript_gene_table = indexes(bg)$t2g
head(transcript_gene_table)
```

```
##   t_id    g_id
## 1    1 MSTRG.1
## 2    2 MSTRG.2
## 3    3 MSTRG.3
## 4    4 MSTRG.3
## 5    5 MSTRG.4
## 6    6 MSTRG.5
```

```
length(row.names(transcript_gene_table))
```

```
## [1] 5727
```

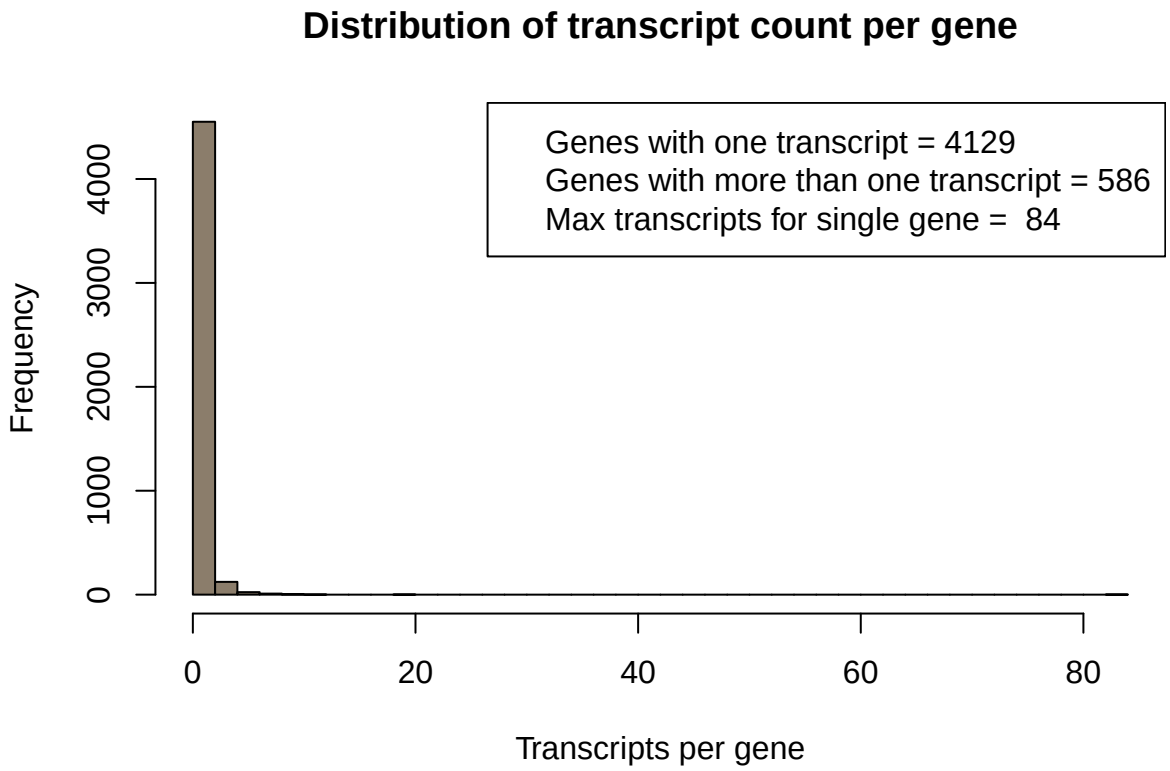
```
length(unique(transcript_gene_table[,"g_id"]))
```

```
## [1] 4715
```

plot the number of transcripts per gene

```
counts=table(transcript_gene_table[,"g_id"])
c_one = length(which(counts == 1))
c_more_than_one = length(which(counts > 1))
c_max = max(counts)
hist(counts, breaks=50, col="bisque4", xlab="Transcripts per gene",
main="Distribution of transcript count per gene")
legend_text = c(paste("Genes with one transcript =", c_one),
paste("Genes with more than one transcript =", c_more_than_one),
```

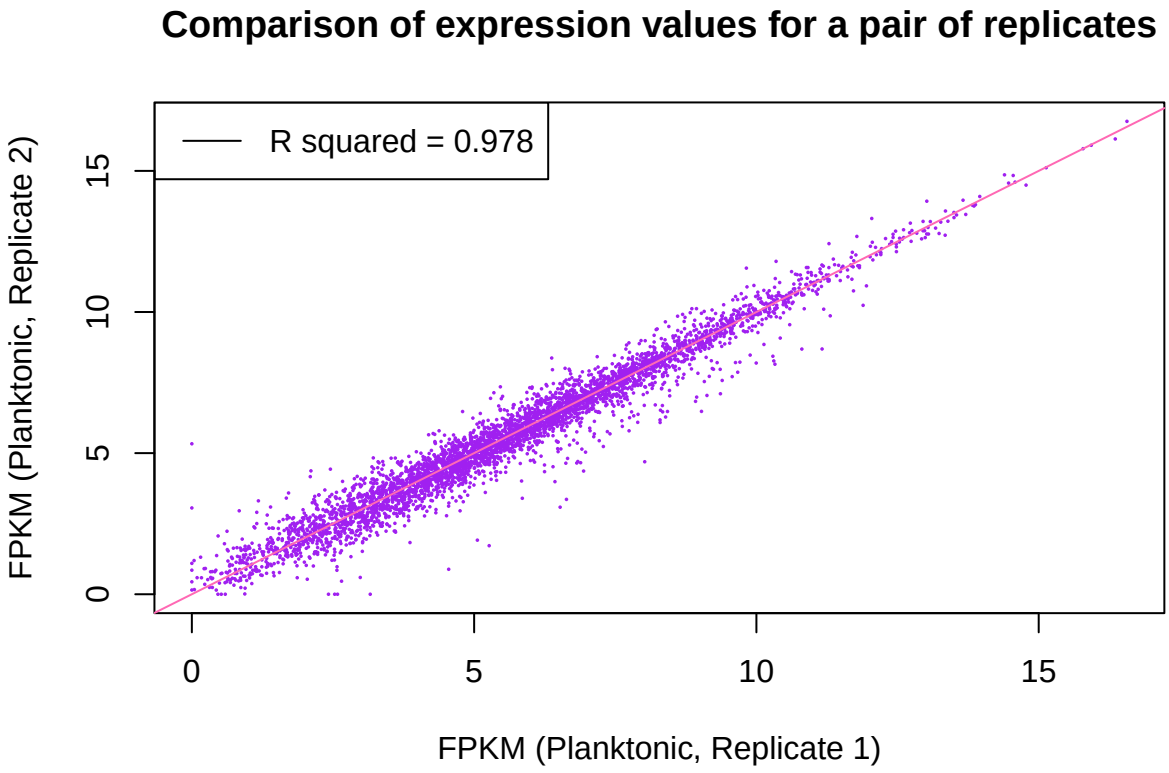
```
paste("Max transcripts for single gene = ", c_max))
legend("topright", legend_text, lty=NULL)
```



<The x axis shows the transcripts per gene and the y axis is the frequency of the gene. Based on the graph there are approximately 0-10 transcripts per gene at a high frenqcy of over 400.>

create a plot of how similar the two replicates are for one another. To modify the chart you can substitute the variables x and y with the new dataset columns and adjust the plot labels.

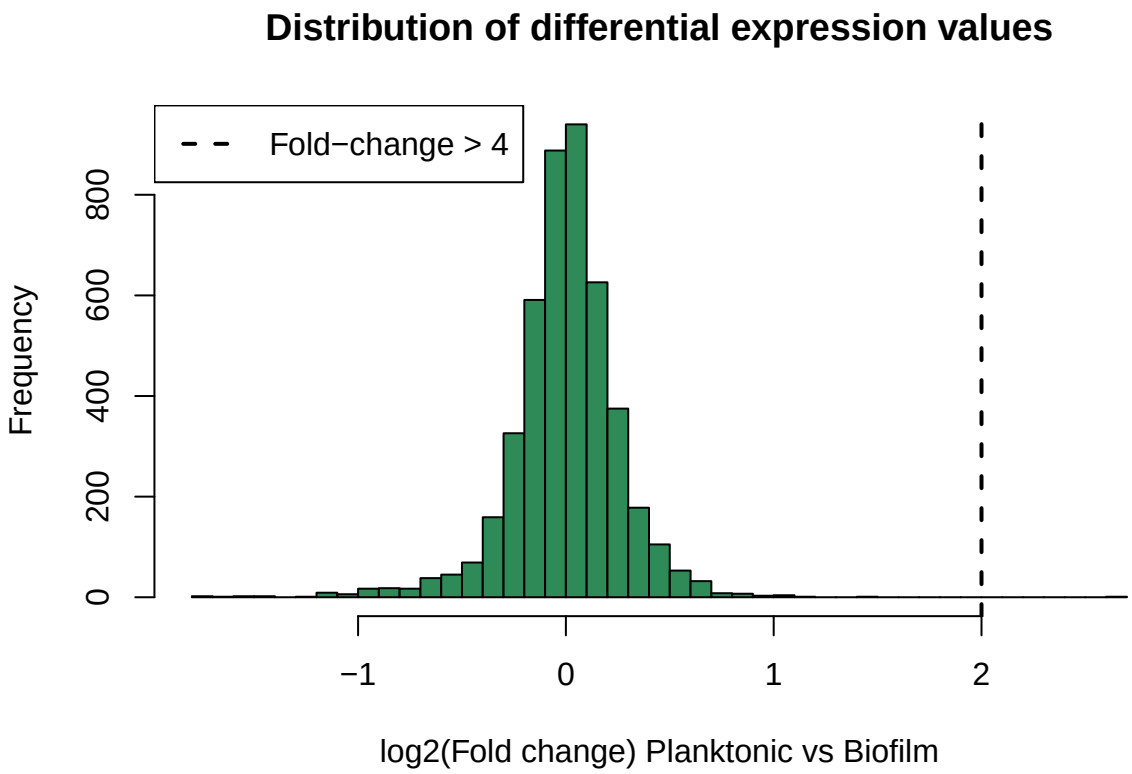
```
x = gene_expression[,"plank01"]
y = gene_expression[,"plank02"]
min_nonzero=1
plot(x=log2(x+min_nonzero), y=log2(y+min_nonzero), pch=16, col="purple", cex=0.25,
xlab="FPKM (Planktonic, Replicate 1)", ylab="FPKM (Planktonic, Replicate 2)",
main="Comparison of expression values for a pair of replicates")
abline(a=0,b=1, col = "hotpink")
rs=cor(x,y)^2
legend("topleft", paste("R squared = ", round(rs, digits=3), sep=""), lwd=1, col="black")
```



If the two data sets are similiar it means there is a high reproducibility and a strong correlation.

create plot of differential gene expression between the conditions

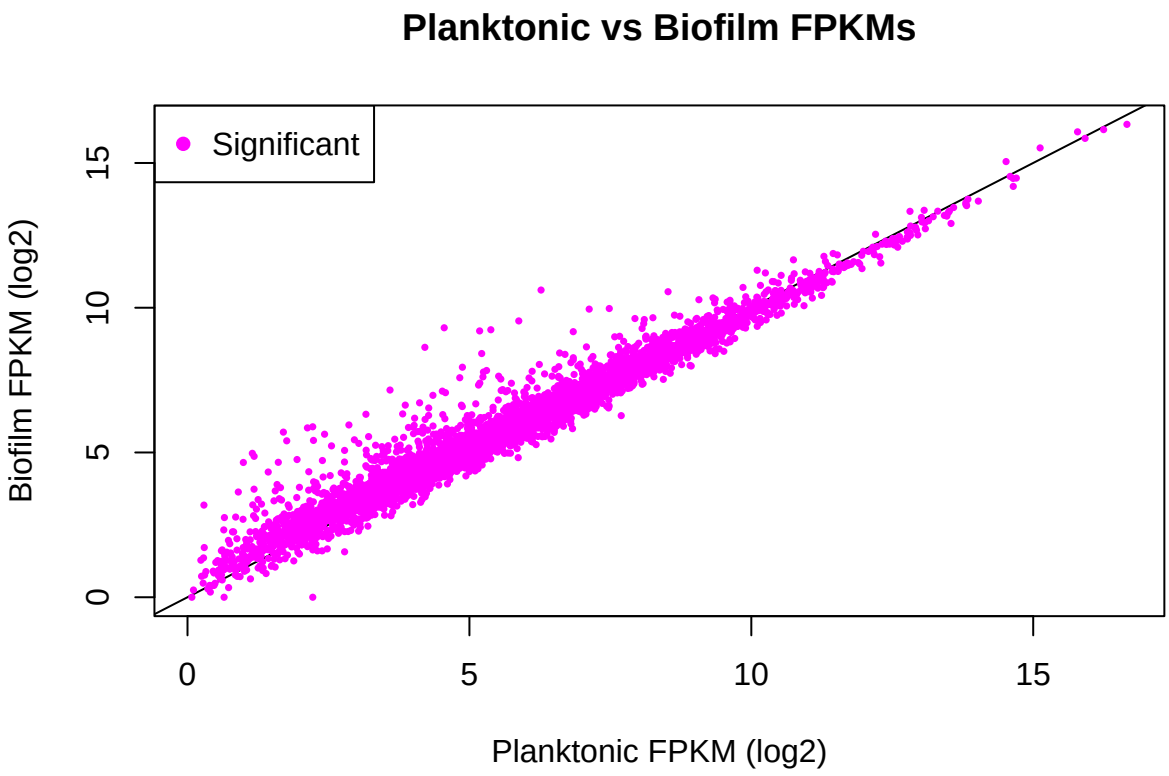
```
results_genes = stattest(bg_filt, feature="gene", covariate="stage", getFC=TRUE, meas="FPKM")
results_genes = merge(results_genes,bg_gene_names,by.x=c("id"),by.y=c("gene_id"))
sig=which(results_genes$pval<0.7)
results_genes[,"de"] = log2(results_genes[,"fc"])
hist(results_genes[sig,"de"], breaks=50, col="seagreen",
xlab="log2(Fold change) Planktonic vs Biofilm",
main="Distribution of differential expression values")
abline(v=-2, col="black", lwd=2, lty=2)
abline(v=2, col="black", lwd=2, lty=2)
legend("topleft", "Fold-change > 4", lwd=2, lty=2)
```



The histogram shows that most the genes have minimal differential expression centered around zero, with few genes having a strong upregulation.

Plot total gene expression highlighting differentially expressed genes

```
gene_expression[,"plank"]=apply(gene_expression[,c(1:2)], 1, mean)
gene_expression[,"biofilm"]=apply(gene_expression[,c(3:4)], 1, mean)
x=log2(gene_expression[,"plank"]+min_nonzero)
y=log2(gene_expression[,"biofilm"]+min_nonzero)
plot(x=x, y=y, pch=16, cex=0.25, xlab="Planktonic FPKM (log2)", ylab="Biofilm FPKM (log2)",
main="Planktonic vs Biofilm FPKMs")
abline(a=0, b=1)
xsig=x[sig]
ysig=y[sig]
points(x=xsig, y=ysig, col="magenta", pch=16, cex=0.5)
legend("topleft", "Significant", col="magenta", pch=16)
```



make a table of FPKM values

```
fpkm = texpr(bg_filt,meas="FPKM")
```

choose a gene to determine individual expression

```
ballgown::transcriptNames(bg_filt)[4]
```

```
##             4
## "gene-PA0004"
```

```
ballgown::geneNames(bg_filt)[4]
```

```
##      4
## "gyrB"
```

transform to log2

```
transformed_fpkm <- log2(fpkm[2, ] + 1)
```

make sure values are properly coded as numbers

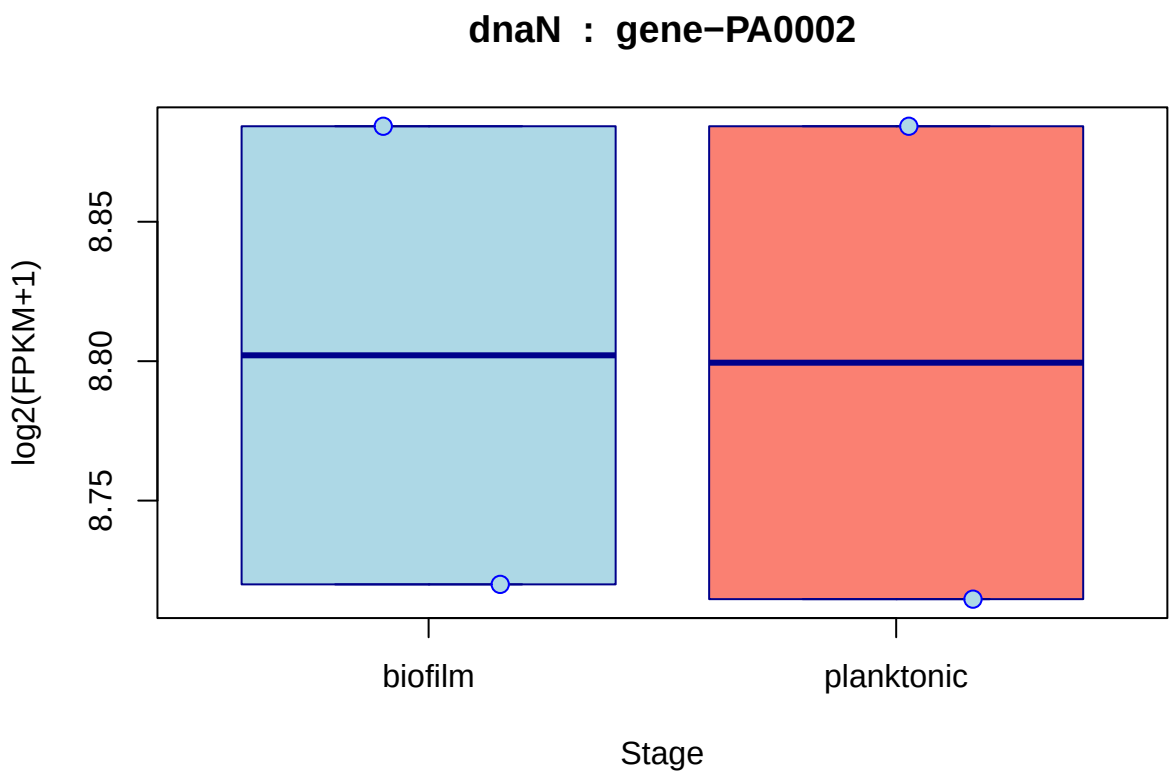
```
numeric_stages <- as.numeric(factor(pheno_data$stage))

jittered_stages <- jitter(numeric_stages)
```

plot expression of individual gene

```
boxplot(transformed_fpkm ~ pheno_data$stage,
        main=paste(ballgown::geneNames(bg_filt)[2], ' : ', ballgown::transcriptNames(bg_filt)[2]),
        xlab="Stage",
        ylab="log2(FPKM+1)",
        col=c("lightblue", "salmon"),
        border="darkblue")

points(transformed_fpkm ~ jittered_stages,
       pch=21, col="blue", bg="lightblue", cex=1.2)
```



The gene dnaN gene-PA0002 show similar expression levels between biofilm and planktonic conditions. There is no major differences observed between the groups.