Transcriptome

Alejandra

2024-05-05

#load required packages

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

#creating a data frame with coloumn names to make it easier to read.

```
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 5693 transcripts and 4 samples
```

#this code creates the object that will only keep gene variance that are greater than one, and keep the genomic data and exxpression

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

```
## ballgown instance with 5177 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)
```

hoose a transcript to examine more closely (this is a demo, you need to choose another)

```
results_transcripts[results_transcripts$transcriptNames == "gene-PA0044", ]
```

```
##    geneNames transcriptNames    feature id       fc      pval      qval
## 46      exoT     gene-PA0044 transcript 46 1027.041 0.5047568 0.9865547
```

According to the data above, the gene I chose is an exoT, that is considered a transcript with a fold change of 1027, indicating that this gene has a hight expression level. Has a p value > 0.05 of the fold change, with an adjusted p value (q value) of 0.987.

#This code is essentially creating a data frame where it's filtering transcripts samples that have significant results with p values < 0.05, and assorts based on deminsion

```r
sigdiff <- results_transcripts %>% filter(pval<0.05)
dim(sigdiff)
```

```
## [1] 186   7
```

#organize the table. Table is being organized by lowest to greatest fold change and p value.

```r
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
## 955      lpxO2     gene-PA0936  955 1.757415e-01 0.0001111028 0.5751792
## 1895         .     gene-PA1859 1895 2.949667e+01 0.0006401088 0.9865547
## 5334         .     gene-PA5218 5334 3.792541e+01 0.0006429856 0.9865547
## 1584     ccoO2     gene-PA1556 1584 1.197208e+14 0.0012593157 0.9865547
## 2788         .     gene-PA2753 2788 1.649170e+12 0.0014447483 0.9865547
## 4399      sbcB     gene-PA4316 4399 2.184140e+04 0.0018974589 0.9865547
```

#load gene names

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

#pull out gene expression data and visualize

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

```
##               FPKM.plank01 FPKM.plank02 FPKM.biofilm01 FPKM.biofilm02
## .                 1.923647    2.3449283    20.20915947     20.7208696
## gene-PA4673.1     0.000000    0.2259395     0.09972263      0.1650301
## gene-PA5160.1     0.000000    1.3954010     0.00000000      0.7613788
## MSTRG.1         401.516644  396.7171343   229.86084582    179.9326248
## MSTRG.10         30.769351   26.9811414    12.01844430     20.5055004
## MSTRG.100       291.604840  266.1007817   230.95141967    238.6613971
```

#This code is organizing the above table with new column names based on the samples it belongs too. This will allow the table to be easier to read.

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

```
##                  plank01     plank02    biofilm01   biofilm02
## .               1.923647   2.3449283  20.20915947  20.7208696
## gene-PA4673.1   0.000000   0.2259395   0.09972263   0.1650301
## gene-PA5160.1   0.000000   1.3954010   0.00000000   0.7613788
## MSTRG.1       401.516644 396.7171343 229.86084582 179.9326248
## MSTRG.10       30.769351  26.9811414  12.01844430  20.5055004
## MSTRG.100     291.604840 266.1007817 230.95141967 238.6613971
```

```
dim(gene_expression)
```

```
## [1] 4332    4
```

#load the transcript to gene table and determine the number of transcripts and unique genes

```
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))
```

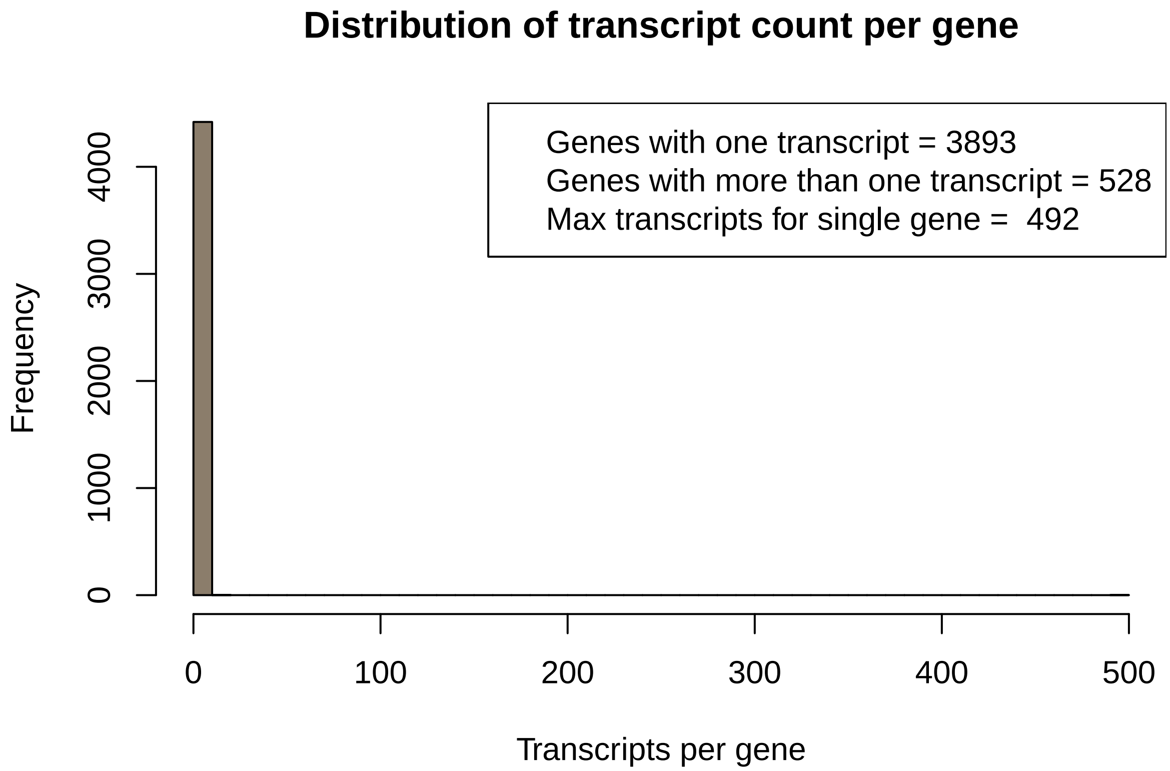
```
## [1] 5693
```

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

```
## [1] 4421
```

plot the number of transcripts per gene

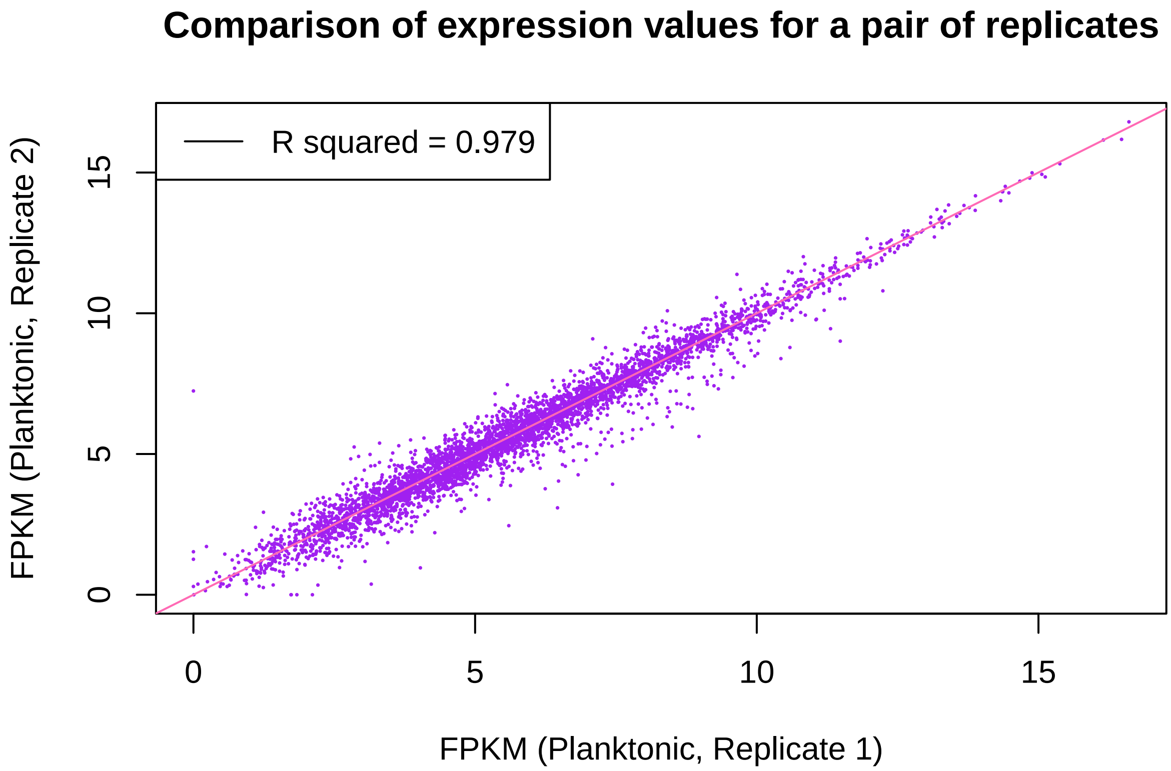
```
counts=table(transcript_gene_table$g_id)
c_one = length(which(counts == 1))
c_more_than_one = length(which(counts > 1))
c_max = max(counts)
hist(counts, breaks=50, col="bisque4", xlab="Transcripts per gene",
main="Distribution of transcript count per gene")
legend_text = c(paste("Genes with one transcript =", 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 majority of samples contained about 1 transcript per gene and a frequency of over 4000.

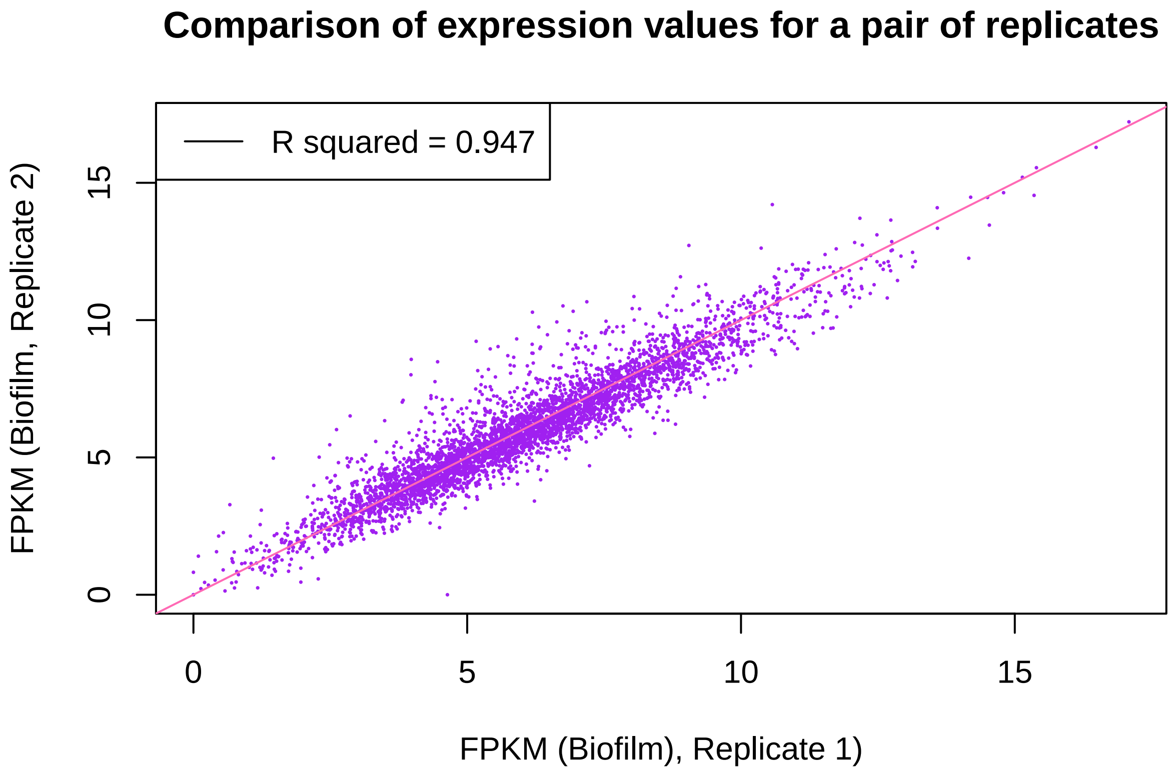
#create a plot of how similar the two replicates are for one another. We have two data sets…how can you modify this code in another chunk to create a plot of the other set?

```
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")
```



##plot similarity of biofilms

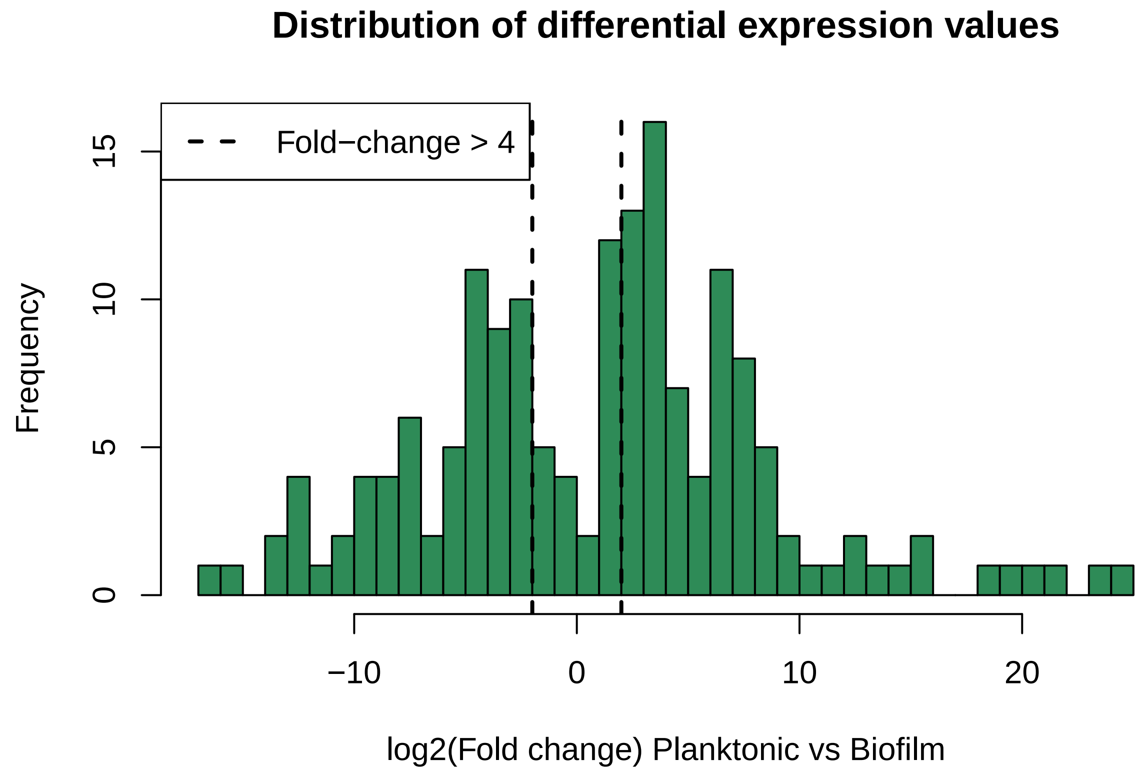
```
x = gene_expression[,"biofilm01"]
y = gene_expression[,"biofilm02"]
min_nonzero=1
plot(x=log2(x+min_nonzero), y=log2(y+min_nonzero), pch=16, col="purple", cex=0.25,
xlab="FPKM (Biofilm), Replicate 1)", ylab="FPKM (Biofilm, Replicate 2)",
main="Comparison of expression values for a pair of replicates")
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")
```



## What does it mean if the two data sets are similar? #Similarity between two data sets indicate shared expression patterns and biological features between the samples.

create plot of differential gene expression between the conditions

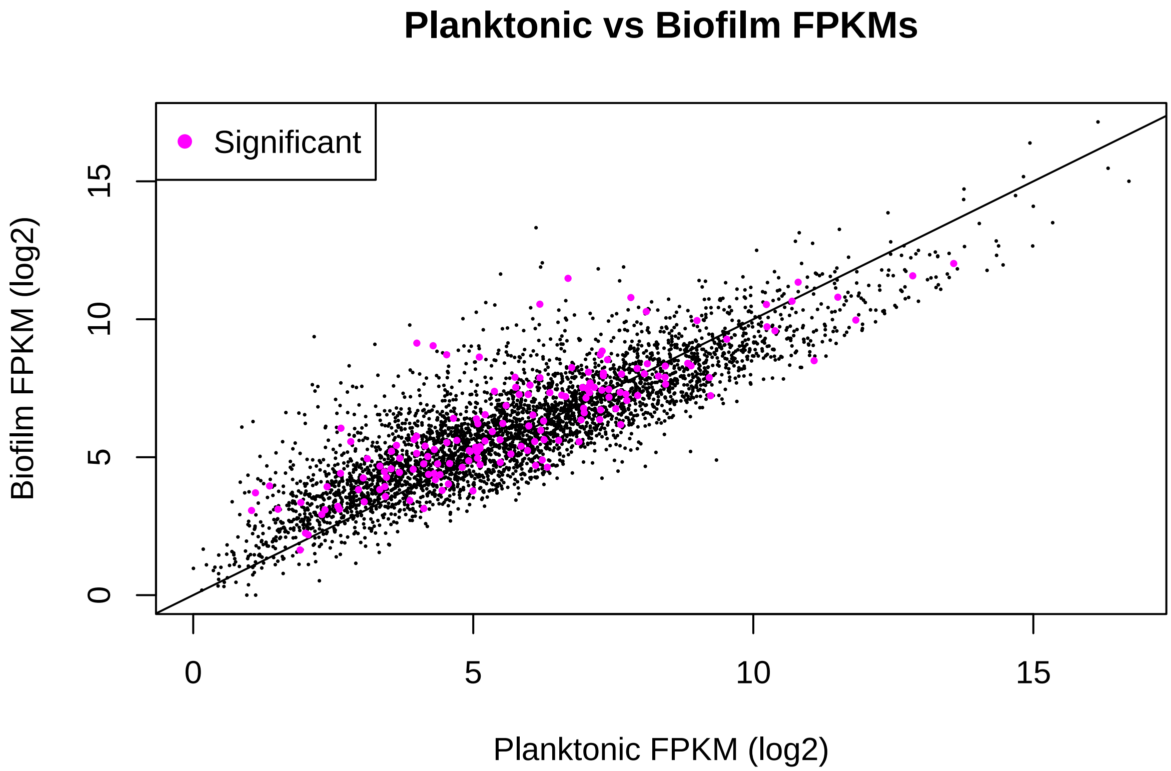
```
results_genes = stattest(bg_filt, feature="gene", covariate="stage", getFC=TRUE, meas="FPKM")
results_genes = merge(results_genes,bg_gene_names,by.x=c("id"),by.y=c("gene_id"))
sig=which(results_genes$pval<0.05)
results_genes[,"de"] = log2(results_genes[,"fc"])
hist(results_genes[sig,"de"], breaks=50, 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)
```



interpret the above figure: The bar graph above displays gene results of gene-PA0044 expression from Planktonic and Biofilm samples and their frequency. Fold change parameter was set to -4 to 4. Gene frequencies outside these threshold are significantly different with highest frequency seen greater than 4.

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 (pick a different number than I did)

```
ballgown::transcriptNames(bg_filt)[666]
```

```
##           753
## "gene-PA0741"
```

```
ballgown::geneNames(bg_filt)[666]
```

```
## 753
## "."
```

transform to log2

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

make sure values are properly coded as numbers

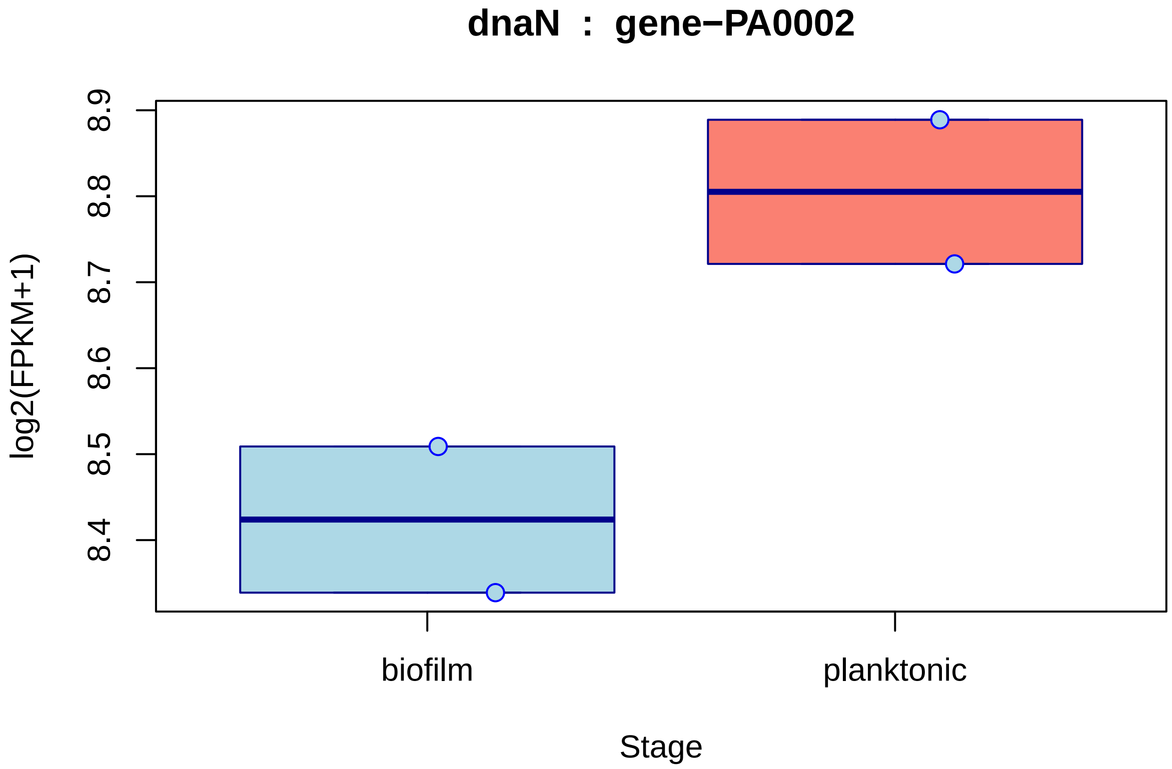
```
numeric_stages <- as.numeric(factor(pheno_data$stage))

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)
```



interpret the above figure

#Box plot of gene-PA0002 expression Aafter being log transformed PA0002. Based on graph, Planktonic samples showed a greater transcript abundance when comapred to biofilm samples.