Buergeria otai is a species of frog whose tadpoles have been found in hot springs upwards of 40 degrees Celsius. Previous testing has indicated that the genes responsible are located on chromosome 2, but we are unsure where. Perform a differential expression analysis. Detail the steps that you take and provide the necessary intermediate results for each step. Submit a final list of the genes identified as differentially expressed in the reference annotation and provide the expected function of the result with the highest difference in expression value (regardless of significance).

Steps

- 4 RNA Alignment Files and 1 reference annotation
- Process Alignment File
    - Samtools Flagstat - all reads in control and heat samples mapped uniquely to a position in reference
- Get Read Counts of Each Sample using htseq

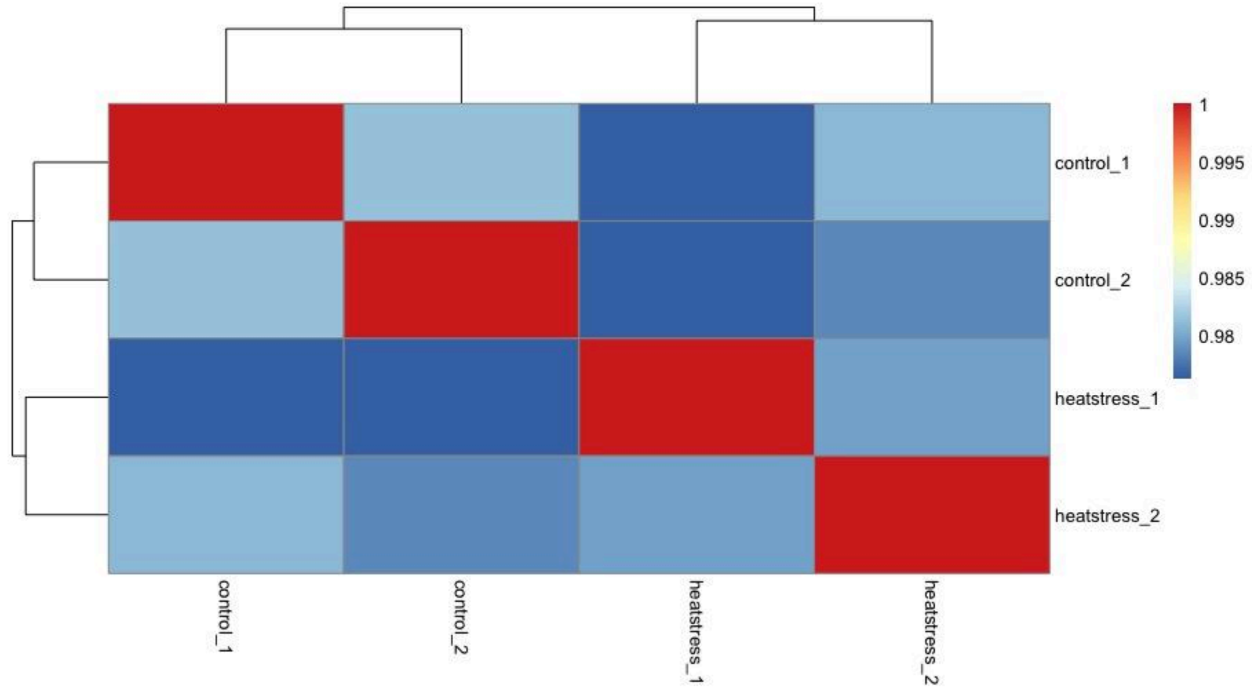
```
#!/bin/bash
#SBATCH --time=0:30:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=8
#SBATCH --mem=8gb
#SBATCH --partition=Centaurus

module purge
module load anaconda3

### Define a shortcut to the annotation file
annot=/users/jaileru/final/Xenopus_tropicalis_annotation.gtf

### Activate the htseq environment
source activate htseq

### Run HTSeq-count to get the read counts
htseq-count -f bam -t gene -i gene_id  control_1.bam  $annot >
./counts-files/control_1.htseq.out
```

- Read counts in counts-files
    - control_1.htseq.out

    - control_2.htseq.out
    - Heatstress_1.htseq.out
    - Heatstress_2.htseq.out

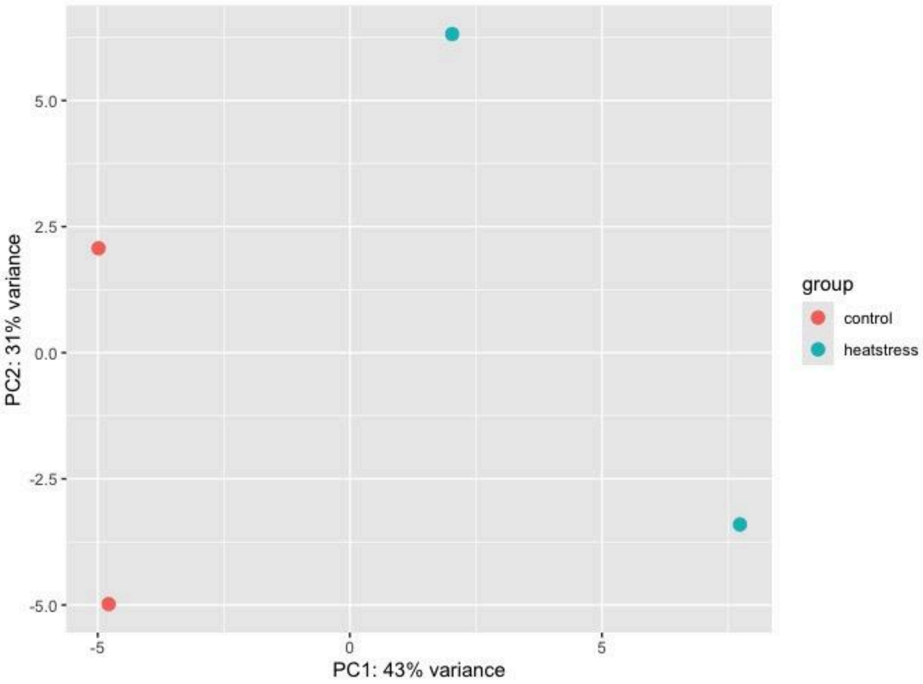
- Use DESeq2 for Differential Gene Expression Analysis

```
#Load Libraries
library(DESeq2)
library(pheatmap)
library(RColorBrewer)
### Load the table of counts from the HT-seq files
sampleFiles = list.files("~/final_MSA/counts-files/", pattern="*.htseq.out")
sampleNames = gsub("\\.htseq\\.out", "", sampleFiles)
sampleStates = unlist(lapply(strsplit(sampleNames, split="_"), `[[`, 1))
sampleReps = unlist(lapply(strsplit(sampleNames, split="_"), `[[`, 2))
sampleTable = data.frame(sampleName = sampleNames,
                         fileName = sampleFiles,
                         condition = sampleStates,
                         rep = sampleReps)
### Create the DESeq2 object
ddsHTSeq = DESeqDataSetFromHTSeqCount(sampleTable = sampleTable, directory = "~/final_MSA/counts-files", design = ~condition)

### Determine the size factors needed for normalization
dds = estimateSizeFactors(ddsHTSeq)

### Do Variance Stabilizing transformation
vsd = vst(dds, blind = T,nsub=600)

### Extract the vst matrix from the vsd object
vsd_mat = assay(vsd)

### Compute the pairwise correlation values and plot the heatmap with dendrogram
vsd_cor = cor(vsd_mat)
pheatmap(vsd_cor)
```



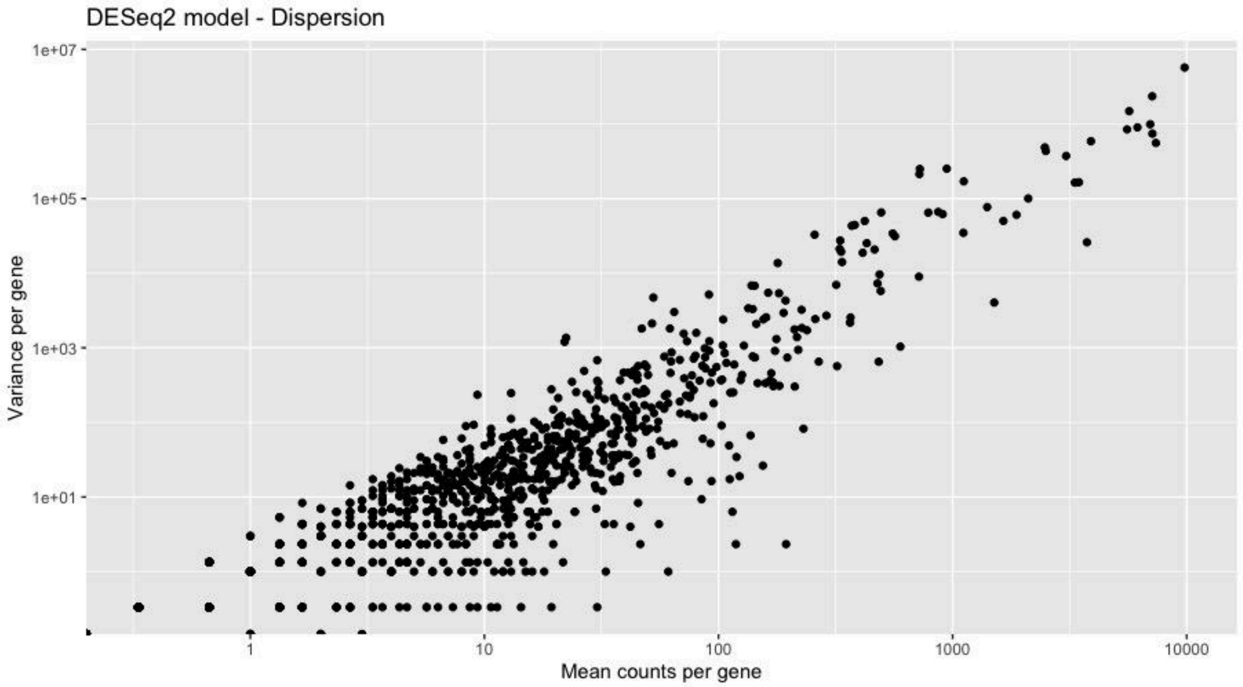
In the pairwise correlation plot there is a high level of correlation between the control and experimental samples but hierarchical clustering heatmap still clusters the control samples together and the experimental samples together which means that they are more similar within the groups compared to between the groups.

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



Again we can see good clustering of the control and heat stress samples meaning they are similar with replicates of the same group rather than replicates between different groups.

```
### Calculate the mean for each gene
readCounts = counts(ddsHTSeq)
mean_readCounts = apply(readCounts[,1:3], 1, mean)


### Calculate the variance for each gene
var_readCounts = apply(readCounts[,1:3], 1, var)

### Plot the mean versus variance in read count data
df = data.frame(mean_readCounts, var_readCounts)

library(ggplot2)

ggplot(df) +
  geom_point(aes(x=mean_readCounts, y= var_readCounts)) +
  scale_y_log10() +
  scale_x_log10() +
  xlab("Mean counts per gene") +
  ylab("Variance per gene") +
  labs(title = "DESeq2 model - Dispersion")
dds = DESeq(dds)
res = results(dds)
write.table(res, file="DESeq_Results", quote=FALSE, sep="\t", row.names=TRUE,
col.names=TRUE
```



Using RNAseq data you would expect to see the variance differ from the mean in some way and not be equal as in the Poisson Distribution. Data follows the negative binomial distribution

Highest Difference in Gene Expression Value (log2FC): -4.128281

Gene ID: ENSXETG00000015457 (slain 1)

Description : Microtubule plus-end tracking protein that might be involved in the regulation of cytoplasmic microtubule dynamics, microtubule organization and microtubule elongation.

Attach gene_names.txt (list of differentially expressed genes)