Khan Inan

Methods

How I started this project was by first trimming the fastq files using fastp. By specifying the required setting it is possible to automatically remove adapters from single-end reads as well as polyG sequences. I used a similar method to what was used for my week 2 assignment. Originally I was having some trouble with some of the fastq files but once the professor told us about the new corrected fastqs2 file, I no longer had any issues. After the trimming, I was able to run fastqc on the processed reads in order to eventually generate the MultiQC report. The main unusual thing I noticed in my MultiQC report was there are an extra 2 samples (along with the original 6) and I believe those were there because they were also statistically significant genes to consider. After generating the report, I downloaded the required files from ensembl and saved it onto my local machine. After unzipping the files in question I ran picard tool and normalized the fasta data. What this did was it stripped everything that came after the transcript id in the data set. From there I had to create a salmon index for my newly normalized data and samples. It is important to note that I had to specify in commands to run salmon is mapping mode since the data was single-end, and also salmon inferred that my library type should be unstranded. The last few steps were that I had to convert the salmon TPMs txt file that I obtained using tximport and then conduct DGE on the file along with DESeq2. Since tximport needs a mapping file input in order to map ids to genes, I had to import the specified tx2gene file as a data.frame, since it was stored as a csv. After running DESeq2 I had to specify that the changes should be reported as shrunken. The main statistical correction method that I decided to use in addition to the multiple-test correction method was the transcript abundance estimation method.

Results

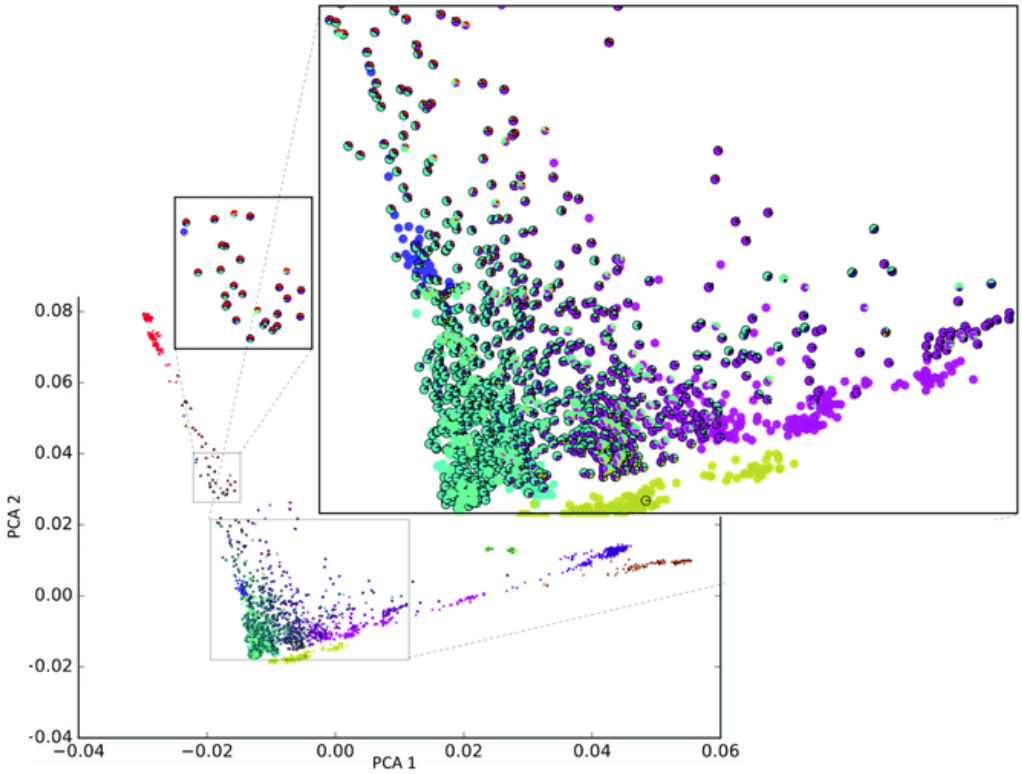
```
SRR7819990= 88.72301%
SRR7819991=91.98127%
SRR7819992=78.6781%
SRR7819993=85.9172%
SRR7819994=93.24825%
SRR7819995=95.67219%

# The mapping rate for all 6 samples seem to be at fairly good percentages and at the higher end of the spectrum, The kmer l
engths seem to be lax enough. In order To increased the mapping rates even more, I could re-run salmon and choose a smaller
k-mer length, but I believe it is fine as is
```

Shown above are the mapping rates for the 6 samples. They were at good enough percentages and if they were less than 30% I would consider re-running the salmon, but as it is I believe these are acceptable mapping rates with the k-mer lengths I chose. I should add that the total number

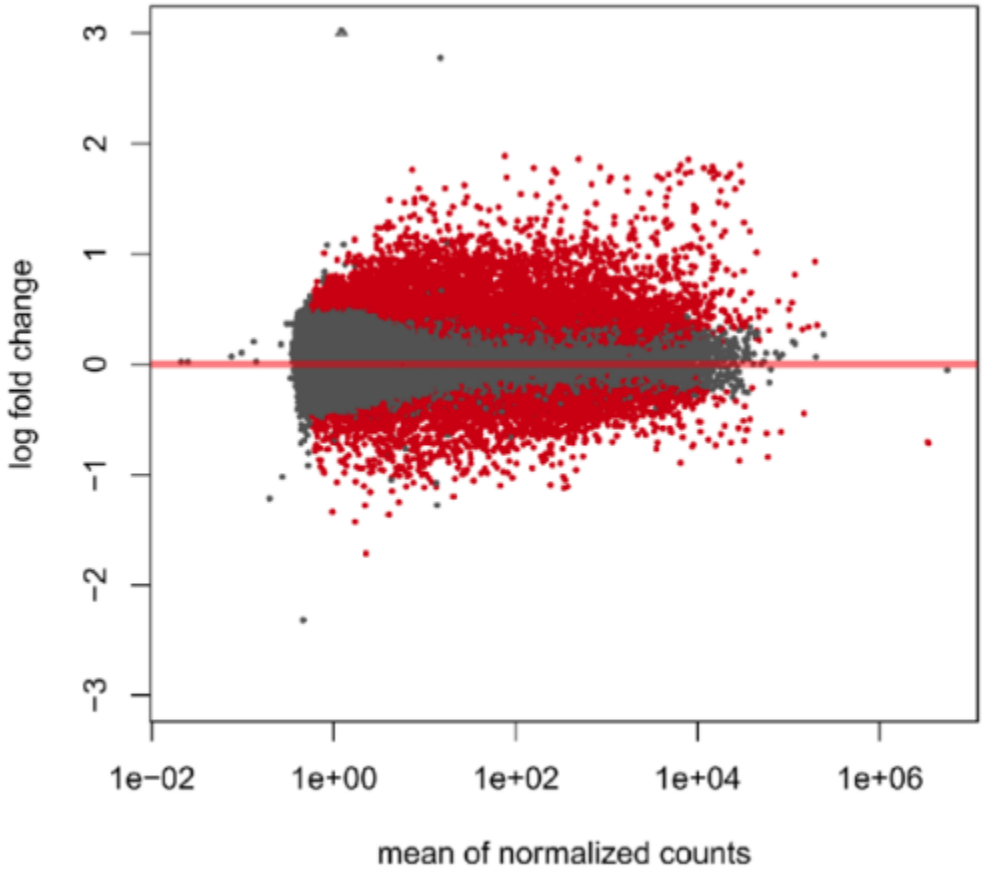
of reads for the Homo_sapiens.GRCh38.cdna.all.fa.gz were about 58359 reads. According to my MultiQC report the number of statistically significant genes at my FDR was 8

```
log2 fold change (MMSE): condition LowCount vs HighCount
Wald test p-value: condition LowCount vs HighCount
DataFrame with 2 rows and 5 columns
                     baseMean log2FoldChange     lfcSE      pvalue        padj
                    <numeric>      <numeric> <numeric>   <numeric>   <numeric>
SRR3192657  6705.780       8.291839  0.683121 4.21831e-55 7.12937e-45
SRR3192658  6989.783       6.985437  0.451921 2.21932e-22 5.12125e-21
```

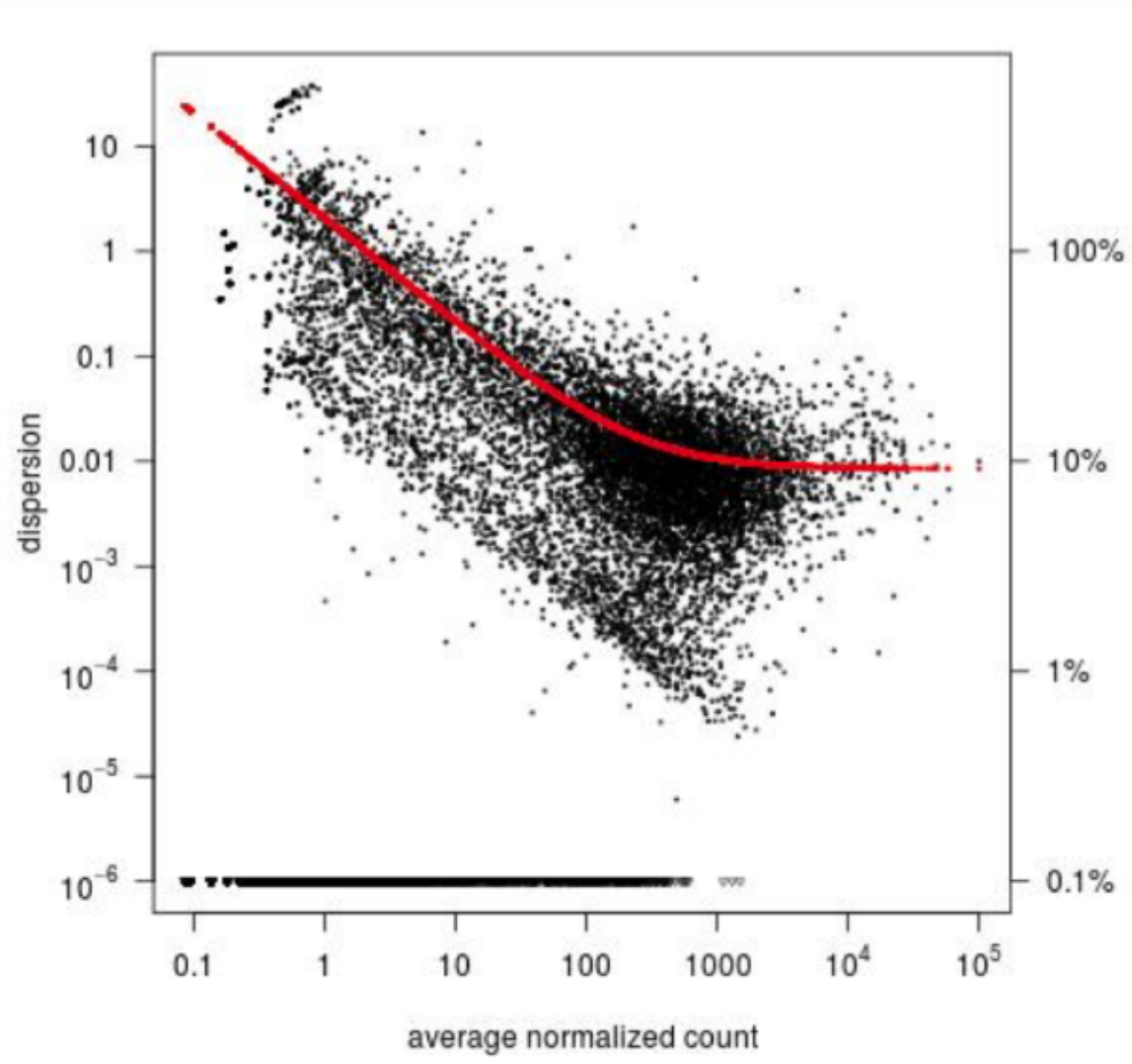
The top two most biologically relevant differentially expressed genes that I found are shown above, and they were included in my MultiQC file as well. The reason why I consider these genes to be differentially expressed is because they have a fairly high log2 FC value, that is also greater than 1. What this means is that the genes that are differentially expressed show a change in expression that is at least 2 fold. Both of these genes that I selected also have a comparatively lower adjusted p-value (or padj on the table.) The FDR threshold that I chose (which is 0.01) means that 1 percent or less of the genes should be false positives. Both of these genes have an adjusted p-value of greater than 0.01 , and when combined with the high log2 FC value, I think it is safe to consider these two genes to be differentially expressed and biologically significant.



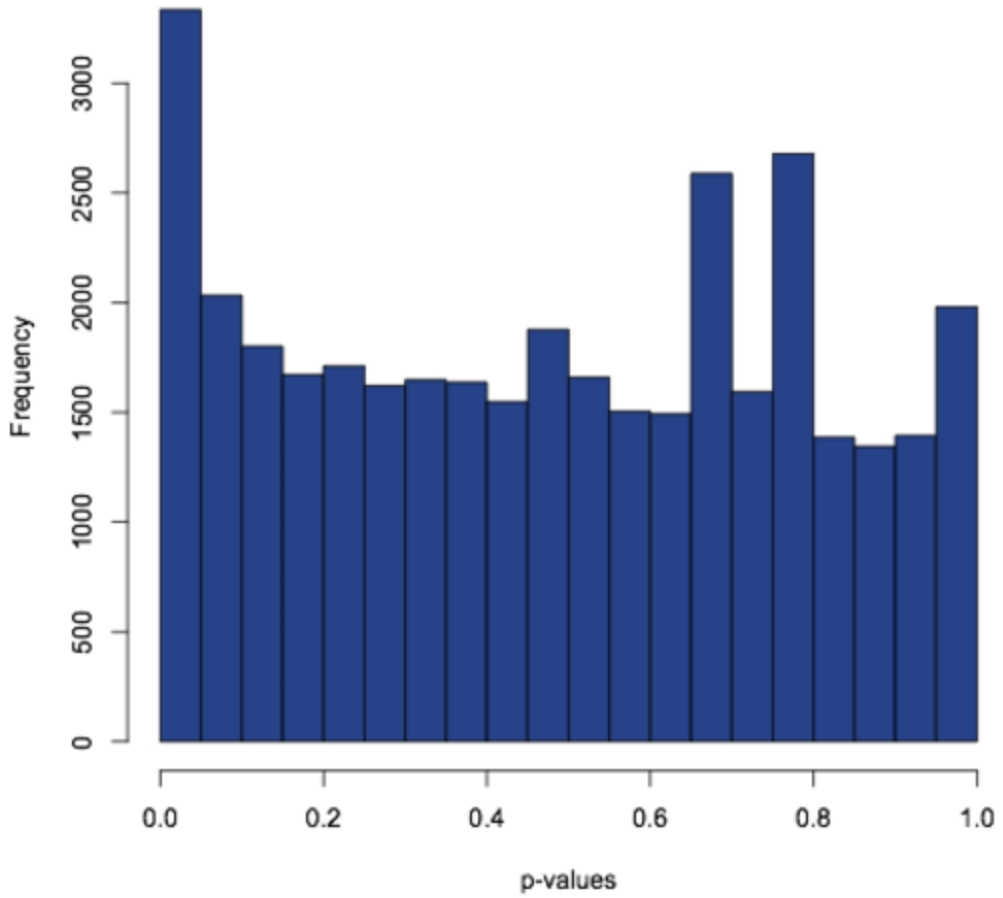
Shown in the previous page is the PCA plot that I was able to generate. What this plot shows is that essentially there's a tightly packed cluster in the lower left side of the graph and as the gene data points get farther away from this cluster, they spread out more and more. What this suggests is that most of the genes (including our 6 samples) have very similar expression levels, and there is a much smaller amount of genes (outliers) that vary in expression levels from our main cluster.



Shown above is the MA plot that I was able to generate. What the color coding represents is the plotting of gene LFC values pre-normalization and post-normalization. What the MA plot tell us is similar to the PCA plot, most of the genes tend to be clustered around the center (when the LFC is at 0) and from there are a much smaller amount of outliers that stray further and further from the initial cluster, and essentially these outlier genes are exceptions to what our average mean and LFC value is.



My Dispersion-by-mean plot tells a similar story to the other plots I was able to generate. It seems that most of the gene data points are somewhere near the curved mean line (in red) Although it seems that there are a decent amount of genes that stray from this line, My calculations and previous plots imply that a majority of the gene dispersion values do not stray too far from the mean.



Lastly this is the raw p-value histogram that I obtained from DESeq2. Interestingly enough, what this histogram is telling me is that frequency for the p-values across the range stays at around 1500. I was expecting the frequency to deviate heavily across the board and give me a very cluttered and chaotic p-value histogram, But judging from my results, it seems to me that highest frequency is at the lower end of the p-value spectrum and it gradually gets lower as the p-value gets higher (with some spikes between 0.6 and 0.8 p-value)