NGS Final Report: Differential Gene Expression Analysis using DESeq2 of Breast Cancer RNA-seq Samples

Jessie Bologna

5/14/2021

Introduction -

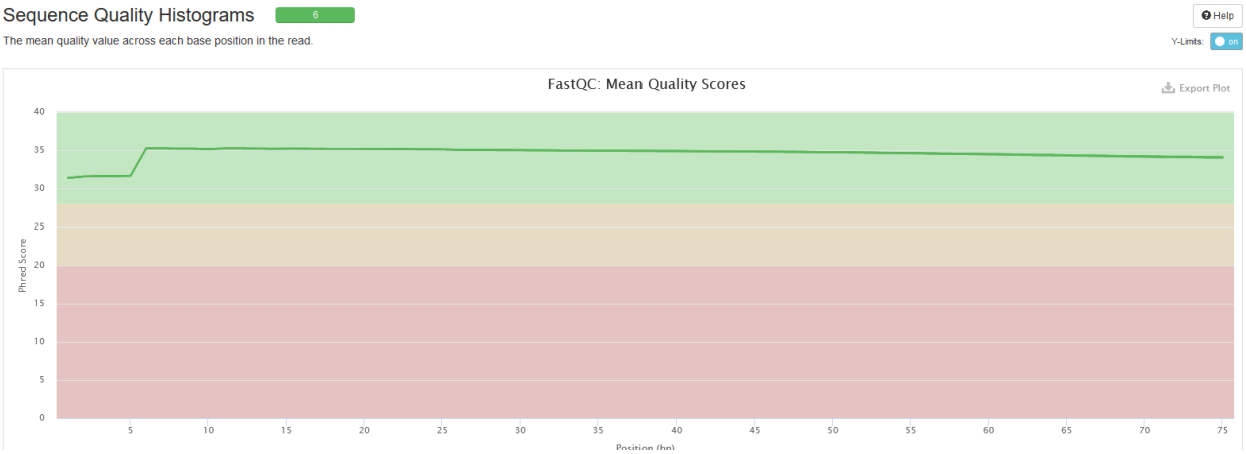
The following report uses an RNA-seq dataset obtained from a study by Jiao, et al., (2019) to investigate the transcriptomic changes between breast cancer cells treated with NRDE2-targeting siRNA's that cause NRDE2 depletion and control breast cancer cells not treated. The dataset contains 6 total fastq files from single end sequencing of 3 treated samples and 3 control samples.

I do this by conducting an analysis to get the count data from RNA-seq data to then detect differentially expressed genes by running an analysis using DESeq2 to compare the transcript expression between the two classes of samples, and to also characterize deferentially expressed genes that couple be potentially be impacted by knocking down NRDE2.

Materials and Methods -

Step 1: QC of fastq sample files - The first step in the analysis is to perform some initial quality control steps on the data obtained from the breast cancer study. To do this I start by running the QC tool fastp to get baseline quality scores and to trim adapter regions, polyG sequences, etc. Next, fastqc and Multiqc was run to produce an html report for review to ensure the quality of the samples and that the downstream analysis and alignment will provide accurate and useful information. See attached supplementary materials file for reports.

After reviewing the reports I note that overall 99.3% of the reads passed filtering. The report noted a high duplication rate as well as a fail for the per-base sequence content. This is acceptable and expected for such samples as they are taken from single end seqeuenced cancer cells. Overall, the baseline quality scores were good and acceptable and all stats were consistent across all samples.



Step 2: Reference file: Download, Normalize, and Index the human reference file - After confirming the quality of our sequenced samples, the next step is to obtain the reference transcript file to be used for our 'psuedo' alignment that we will be conducting using Salmon to quantify transcripts in the following step. Before conducting Salmon we download the reference file from 'http://ftp.ensembl.org/pub/release-104/fasta/homo_sapiens/cdna/Homo_sapiens.GRCh38.cdna.all.fa.gz'.

After downloading the reference file it is necessary to normalize it to get rid of any white-space around the sequence identifiers. I used Picard tools to do this. Now the reference file is ready to be indexed. This step is often required prior to alignment or other analysis steps in order to speed up computational time. Since we are using Salmon to conduct our pseudo alignment we will use Salmon to index our reference file. Note that this step automatically detects the most likely library type of our samples, which was noted as SR, this means the reads come from a stranded single-end protocol where the reads comes from reverse strand. Now our reference file is ready to be used to run Salmon.

Step 3: Running Salmon - Salmon is a tool that can quantify transcripts from RNA-seq data using a quasi-mapping approach without the computationally costly alignment step. Here I will be running Salmon in the mapping-based mode to get our TPM (transcripts per million) count matrix. The first step in the process is to index your reference file, which we already did in the above step.

The next step is to run Salmon quant in order to quantify the reads directly against the indexed reference file. The output of this step produces TPM (transcripts per million), read counts, isofrom lengths, and effective lengths for each sample which will be used in the next steps to run our differential transcript expression using DESeq2 in R.

Step 4: DESeq2: Differential Transcript Expression Analysis - Here we run our differential expression analysis in R using the Bioconductor packages tximport and DESeq2. DESeq2 provides methods for testing differential expression by using quantification and statistical inferences to show the systematic changes between conditions as opposed to within-condition variability (Anders et al., 2021).

"A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, and mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions. This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files." DESeq2 package version: 1.30.1 - (Anders et al., 2021)

The first step is to use the tximport package to convert the Salmon quantification results called TPM's (transcripts per million) into gene-level counts. Note that tximport requires a mapping file to be imported and converted as a dataframe in order to map the transcripts ids to genes (see the .csv file in the supplementary file).

Once the TPM's are converted to gene levels counts, we are ready to create our DESeq2 object that will be used for our analysis. The analysis I will be reporting below uses a 'shrunken' log-fold change estimate. A Wald-test was also conducted on the differential gene expression values using the normalized count data. I also filtered out genes with reads less than or= 10 as they are not useful as they cannot be inferred as differentialy expressed.

```
# which variables and tests were used -
[1] "mean of normalized counts for all samples"
[2] "log2 fold change (MAP): condition Control vs NRDE2_treated"
[3] "standard error: condition Control vs NRDE2_treated"
[4] "Wald statistic: condition Control vs NRDE2_treated"
[5] "Wald test p-value: condition Control vs NRDE2_treated"
[6] "BH adjusted p-values"
```

Summary of running DESeq2 objects on our dataset matrix of quantification scores -

Results -

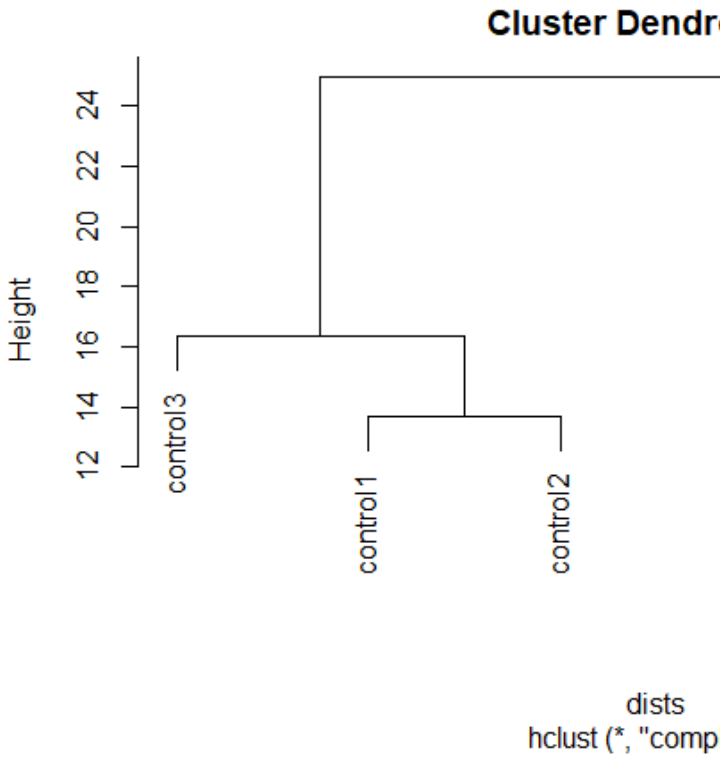
Below is a summary of the differential gene expression analysis done on DESeq2. I have included the results of the top ten deferentially expressed genes between the two groups. Note the results show a table with the log2 fold changes, p-values, and adjusted p-values, which falls below a given FDR cutoff. Also note here, that the top ten differentially expressed genes show very low adjusted p-values. The log2fold change that shows negative represents a down-regulated change between the control compared to the treated samples, whereas the positive log2fold changes show an up-regulated difference in the control vs the treated samples.

```
# Comparing shrunken log fold changes between genes in control vs treated
out of 17961 with nonzero total read count
adjusted p-value < 0.05
LFC > 0 (up)       : 1662, 9.3%
LFC < 0 (down)     : 1381, 7.7%
outliers [1]       : 56, 0.31%
low counts [2]     : 4159, 23%
(mean count < 29)
```

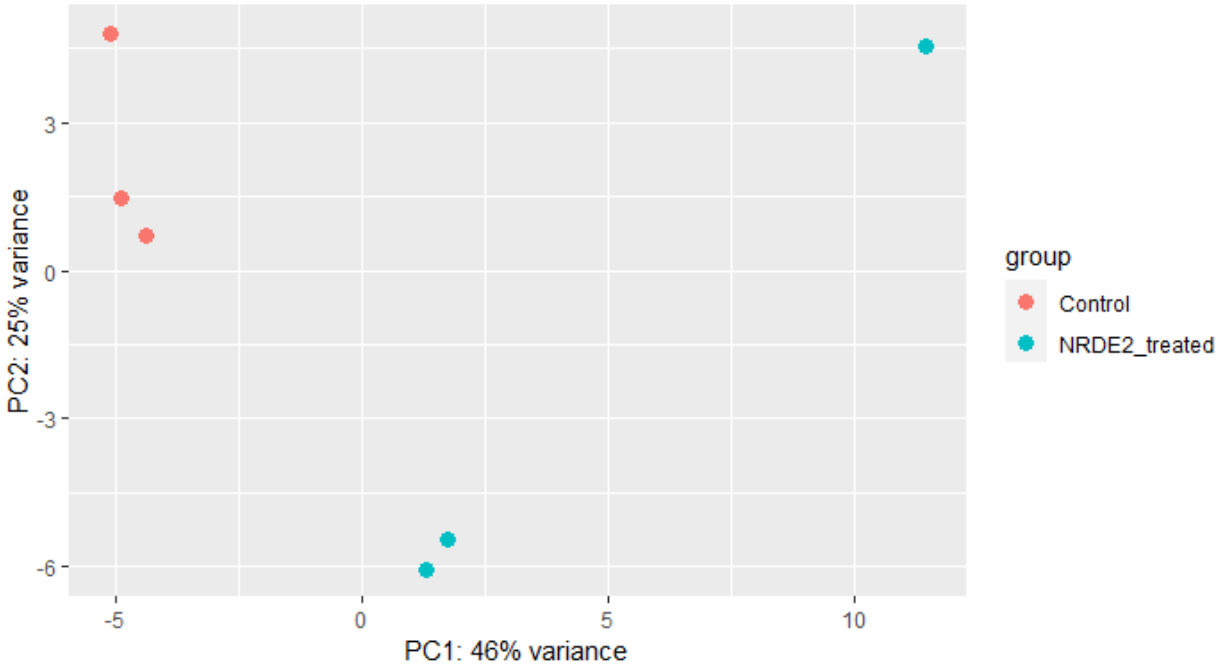
Summary of differential expressed gene:

Top 10 differentially expressed genes:

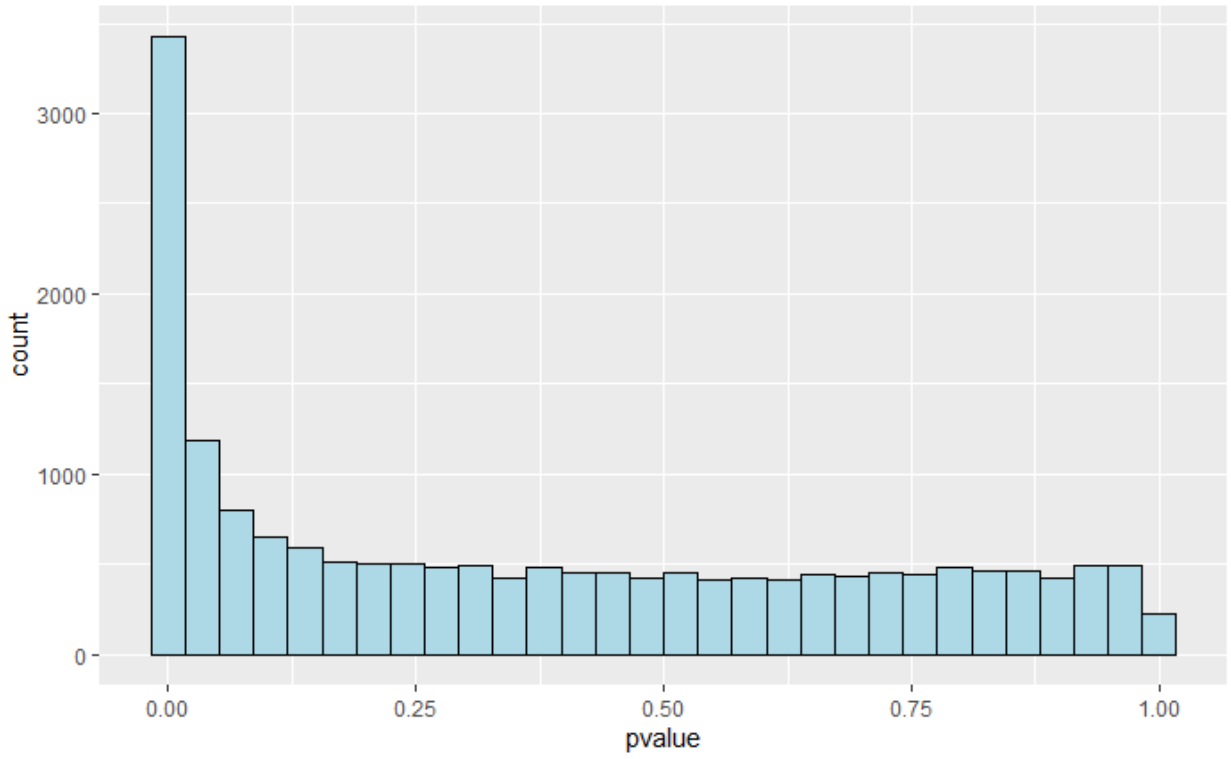
```
log2 fold change (MAP): condition Control vs NRDE2_treated
Wald test p-value: condition Control vs NRDE2_treated
DataFrame with 10 rows and 6 columns
                    baseMean log2FoldChange     lfcSE      stat       pvalue         padj
                   <numeric>      <numeric> <numeric> <numeric>    <numeric>    <numeric>
ENSG00000196396.10  6634.677       -1.14597 0.0407896  -28.0921 1.22490e-173 1.68375e-169
ENSG00000175334.8   6467.739       -1.62684 0.0618447  -26.2997 1.93534e-152 1.33016e-148
ENSG00000206286.11  2809.415        1.36937 0.0562119   24.3488 5.96982e-131 2.73537e-127
ENSG00000119720.18   997.138        3.79060 0.1576105   23.7555 9.63241e-125 3.31018e-121
ENSG00000101384.12 11773.506       -1.28050 0.0585187  -21.8803 4.00061e-106 1.09985e-102
ENSG00000128595.17 22876.387       -1.43526 0.0666856  -21.5219 9.71282e-103  2.22521e-99
ENSG00000124333.16  2742.898       -1.45559 0.0683968  -21.2721 2.06049e-100  4.04621e-97
ENSG00000143384.13 21358.666       -1.04060 0.0493345  -21.0924  9.34863e-99  1.60633e-95
ENSG00000117632.23 16816.515       -1.31972 0.0655333  -20.1373  3.47550e-90  5.30824e-87
ENSG00000164066.13  1269.180        1.47100 0.0741811   19.8110  2.39456e-87  3.29157e-84
```



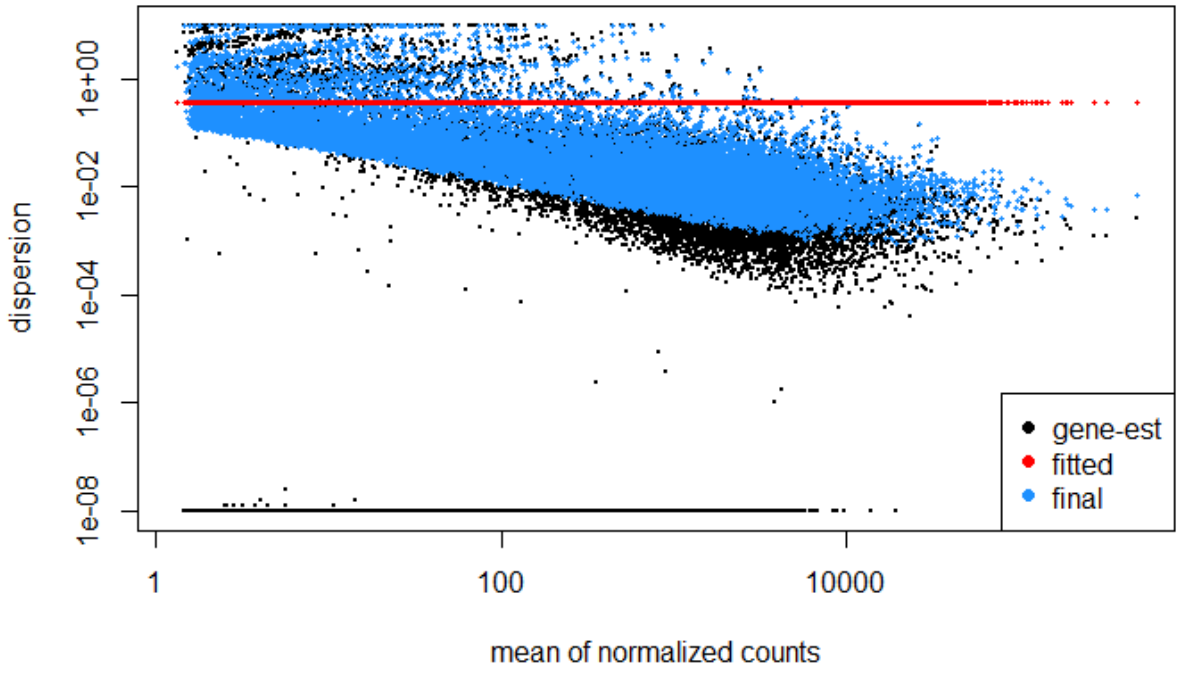
clutering of relationship the 6 samples in our dataset:



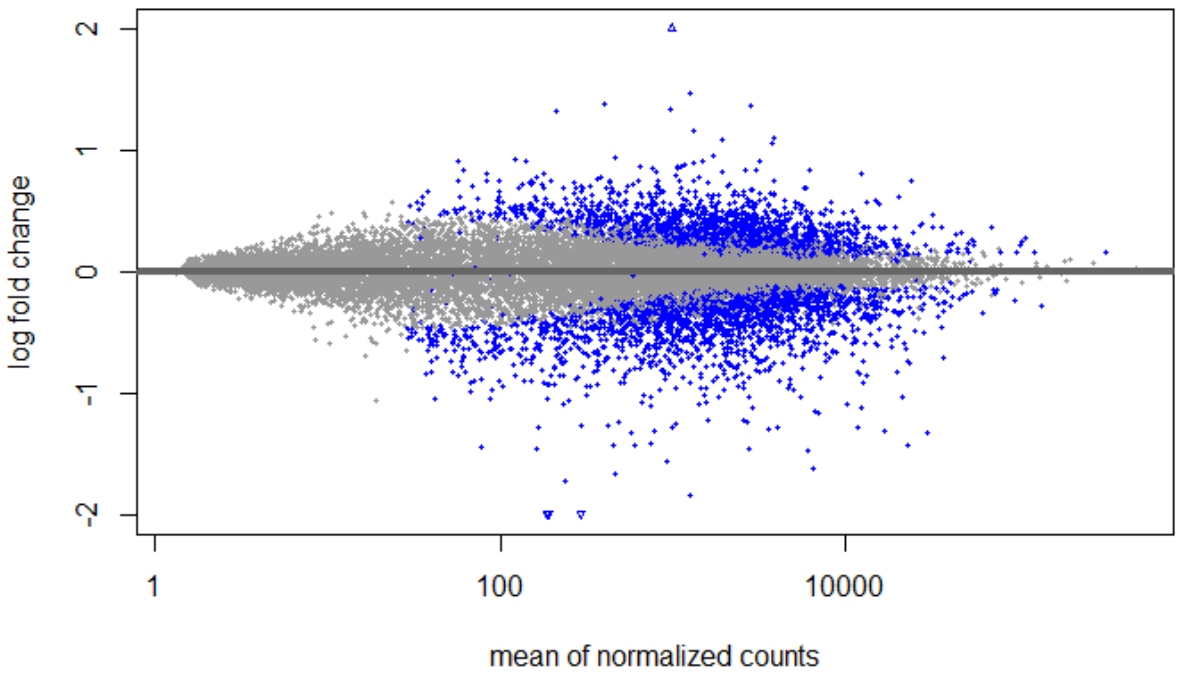
Histogram of the raw p-value's for the data: This shows that there is a large amount of statistically significant differentially expressed genes between the two groups in our analysis



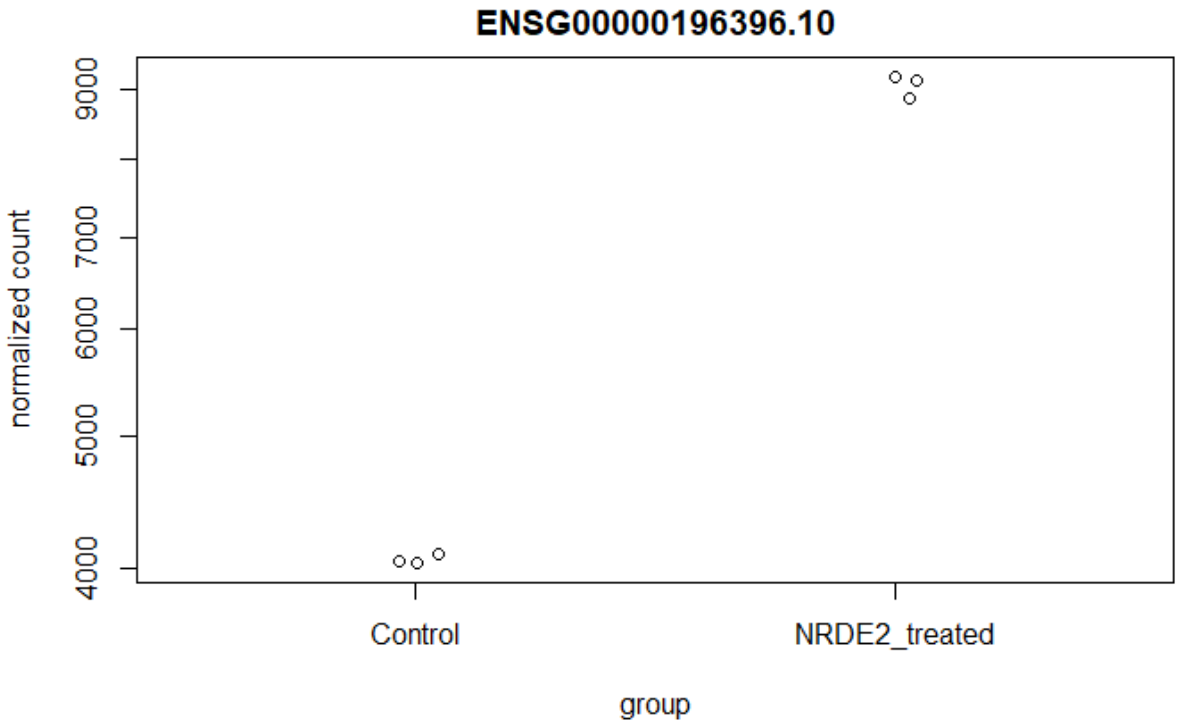
Dispersion Plot: This plot shows the differential expression/ bias (blue) compared to the expected gene expression estimates (black) if there were to be no differentail expression between the groups.



MA-Plot: The below plot is an MA-plot which shows the shrunken log2 fold changes attributable to a given variable over the mean of normalized counts for all the samples. Note that points colored blue = padj values lower than 0.1, points that fall out of the window are plotted as open triangles . I used the shrunken log2 fold changes which removes the noise associated with log2 fold change from low count genes without requiring arbitrary filtering thresholds.



Top differentially expressed gene: See an example of the expression level differences difference between groups for the gene with the highest log fold change between the two groups



Discussion -

One of the downsides of using a count based approach to quantify the transcript expression levels is ambiguous alignment and assumption of transcript length, this can be problematic when the relative abundances of isoforms vary between groups and can cause false positives. However, Salmon solves this issue by estimating TPM's across isoforms for each gene, and then by using tximport and DESeq2 we were able to transfer the TPM matrix to be able to use standard count-based gene levels for our analysis without the standard count-based gene expression analysis issues.

The results of the differential gene expression analysis show that there is a statistically significant difference between the two groups, control and treated, for over 3,043 of the 17,961 genes. The results of the analysis can further be parsed to determine if there is significant expression level changes for a specific gene in question. For example, if a gene that is important in a function of the disease isn't included in our top ten list, we can select the gene in question and note the log2 fold changes and p-values to determine if there is a statistically and or biologically significant difference in expression between the control group compared to the treatment group. The information gained from these results can be helpful in better understanding how breast cancer and the treatment with NRDE2 can affect gene expression levels.

References -

Anders, Simon, and Wolfgang Huber. (2010.) "Differential Expression Analysis for Sequence Count Data." Genome Biology 11: R106. http://genomebiology.com/2010/11/10/R106.

Anders, S., Huber W., Love., M. (Feb 19, 2021). Analyzing RNA-seq data with DESeq2. Vinette. Retrieved from http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#references