RNA seq Analysis of Control and NRDE2 Depleted Breast Cancer Cells

Kieran Bissessar

1 Introduction

RNA interference (RNAi), also known as Post-Transcriptional Gene Silencing (PTGS), is a biological process where RNA molecules impede gene expression through the neutralization of specific messenger RNA molecules. A significant gene in this biological process is NRDE2. Although it is known that the gene codes for an evolutionarily conserved protein in many species, not much is known about the specific role of the gene – especially in homo sapiens. To further examine the functionality of the gene, an RNA seq analysis was performed by researchers at the Beth Israel Deaconess Medical Center. RNA sequencing involves extracting messenger RNA from an organism and sequencing it after fragmenting and copying it to yield stable DNA fragments. Once sequenced using high-throughput short-read methods, the data can then be processed using various NGS methods. In the experiment, the transcriptomic changes were examined after depleting the NRDE2 gene. To carry out the RNA sequencing, MDA-MB-231 breast cancer cells were transfected with 20nM control or NRDE2-targeting siRNAs. RNA was then obtained from each sample after 48 hours before being sequenced using the Illumina NextSeq500 platform.

2 Materials and Methods

To begin processing the raw fastq data from the Illumina, pre-processing needed to be done. The pre-processing involved removing adapters from the reads as well as polyG sequences involved on NextSeq platforms. Fastp was used to do this and the minimum length of the reads were set to 75 as recommended by the Salmon manual. After running fastp, a json and html file was also outputted for each sample. Following running fastp, fastqc was then run to generate fastqc reports. These fastqc outputs were then parsed using MultiQC to generate a multi-sample QC report.

A multi-sample QC report is a quality control analysis of the processed fastqs and is usually done to identify outlier libraries. According to the multi-sample QC report, it is confirmed that all of the samples are from the same species since they all have the same guanosine+cytosine percentage. All of the samples also have roughly the same percent of duplicates ranging from 61.0% to 64.9%. Of all of the samples, SRR7819991 had the highest number of reads (62.4 million) and SRR7819995 had the lowest (43.3 million). Since these numbers are not drastically different, it wasn't considered that much of an issue. The mean quality scores for each of them are the same with little dephasing and none of the samples were found to possess any adapter contamination of greater than 0.1%. Something unusual about the multi QC report is that all of the samples seemed to have failed the sequence duplication levels and the per base sequence content.

The Salmon + tximport + DESeq2 workflow was utilized to carryout this analysis. This was actually the recommended approach for conducting DGE with DESeq2. There were many advantages of using Salmon workflow when compared to other workflows. The primary advantage was speed. Also, this workflow lessens the creation of large intermediate files like BAMs. Another reason why this is the recommended approach by DESeq2 developers is that isoform/transcript based quantification tools like

salmon is capable of advanced bias corrections – ensuring that there will be less confusing factors. Salmon also doesn't require the exact mapping location to be identified and has less inflated false positive rates due to transcript usage as opposed to exon union approaches to gene-level analysis.

So, after analyzing the quality control analysis, Salmon was used on the single-end fastq data. Salmon is a reference based tool that requires a reference transcriptome fasta that allows the reads to be pseudo-aligned. Being that Salmon uses pseudo alignment, transcript abundances can be quantified without having to align reads and create SAM/BAM alignment files for each sample. Salmon was set to mapping-based mode to estimate TPM (Transcripts Per Kilobase Million). After running salmon, the mapping rates were determined for each sample and the library type need to be determined.

Salmon was set to automatically infer the library type based on how the first few thousand reads map to the transcriptome. Salmon reports the number of fragments that had at least one mapping compatible with the designated library format, as well as the number that didn't. It also records the strand-bias that provides some information about how strand-specific the computed mappings were. The file also contains a count of the number of mappings that were computed that matched each possible library type. These are counts of mappings, and so a single fragment that maps to the transcriptome in more than one way may contribute to multiple library type counts. According the salmon, the library type was "SR," meaning that it was stranded and came from the reverse strand. It being strand-specific is ideal for performing a DGE analysis. This is because it allows one to determine which strand the RNA is being transcribed from. Because there are numerous cases of overlap (anti-sense) transcription (especially in mammalian genomes), an unstranded library type would make it difficult to determine which of the two transcripts are being upregulated.

After Salmon was run the TPMs were converted to gene counts using tximport. After filtering out read counts less than or equal to 10, the samples were analyzed by clustering the samples hierarchically and Principal component analysis. Then, DESeq2 was used to estimate size factors, dispersions for each gene, and conducts a Wald test of differential gene expression using normalized count data. Next, results were obtained for the adjusted p values and log fold change for each gene. Then, MA plots were plotted to visualize the difference in log fold changes across the entire dataset. DESeq2 log2 fold-change estimates have a tendency to be over-estimated – especially for low expression genes. So, in order to obtain more accurate estimates of the log2 fold-change, the raw estimates were shrunk as recommended by DESeq. To visualize the density distribution of the raw-p values, a histogram was plotted.

In determining which genes are differentially expressed statistically, the results with the lowest adjusted p-values were chosen. It is important to make the distinction between the raw p value and the adjusted p value. Known as the multiple testing problem, since there are large numbers of genes, there will be large numbers of false positives (type I errors). To combat this the false discovery rate (FDR) was used to create adjusted p values. Choosing alpha to be 0.05, genes with an adjusted p value greater than 0.05 were rejected from being differentially expressed. To determine if these genes are biologically relevant, gene expression of two-fold or greater was chosen (log2 fold greater than 1 or less than -1)

3 Results

SAMPLE	CONDITION	NO. OF READS	MAPPING RATE
SRR7819990	*Control*	*54239068*	*90.9621%*
SRR7819991	*Control*	*57176805*	*91.5983%*
SRR7819992	*Control*	*50707558*	*92.6122%*
SRR7819993	*Treated*	*51914655*	*91.7819%*
SRR7819994	*Treated*	*53253369*	*92.0313%*
SRR7819995	*Treated*	*39835621*	*92.0995%*

Table 1: This table shows the total number of reads and mapping rate of each sample in the analysis. The condition for each sample was also include. Each sample has around the same number of reads except for the last one. Also, the mapping rate for each of the samples are relatively high – suggesting that the kmer length was chosen adequately.

The number of statistically significant genes after filtering out genes with an adjusted p-value more than alpha=0.05 was ***3327***.

The number of biologically relevant differentially expressed genes was determined by having a change in gene expression two-fold or greater. So, in addition to filtering out genes with an adjusted p value of 0.05, genes with a log fold change between 1 and -1 were also filtered out. After this filtering process, only ***63*** genes were left.

```
log2 fold change (MAP): condition control vs treated
Wald test p-value: condition control vs treated
DataFrame with 10 rows and 6 columns
                    baseMean     log2FoldChange             lfcSE             stat               pvalue
                   <numeric>          <numeric>         <numeric>        <numeric>            <numeric>
ENSG00000163041  7925.2807750077  -1.69005213322698 0.0637441561911091  -26.507629708261 7.91556389715159e-155
ENSG00000196396 6477.19649567436  -1.14175512465188 0.0433031995593011 -26.3638760349442 3.55843109013999e-153
ENSG00000175334 6336.43984126653  -1.61916754973746 0.0658570383450477 -24.5807289953233 2.03052834599172e-133
ENSG00000206286 2779.40845846847   1.38253745339608  0.058433180816662  23.6473340116926 1.25726838395057e-123
ENSG00000128595  21771.695910053  -1.42824475037082 0.0611993737925754 -23.3365272635837 1.88856964307094e-120
ENSG00000119720  911.86575789695   3.54960197933373  0.151676906695676  23.0879704051421 6.11583875206732e-118
ENSG00000117868 12161.4564709006  -1.18098332209562 0.0548427694633696 -21.5327003860213 7.69185734885256e-103
ENSG00000101384 11456.2571696796  -1.27095896613276 0.0599151705415841 -21.2109866962127 7.56104853945298e-100
ENSG00000075785 8511.50435706466 -0.899287511509002 0.0438396195111614 -20.5121828126567  1.67601454678462e-93
ENSG00000105976 9333.59469732736  -1.49972332743962 0.0739240988611648 -20.2847314859941  1.75418121514145e-91
                          padj
                       <numeric>
ENSG00000163041 1.14712351997521e-150
ENSG00000196396 2.57843916791544e-149
ENSG00000175334 9.80880559670402e-130
ENSG00000206286 4.55508335505293e-120
ENSG00000128595 5.47383025347683e-117
ENSG00000119720 1.47717891991599e-114
ENSG00000117868   1.5924342385653e-99
ENSG00000101384  1.36968394292191e-96
ENSG00000075785  2.69875586800031e-90
ENSG00000105976  2.54215941698299e-88
```

Figure 1: This table is a screen shot from the R console. It displays the 10 most highly significant genes. This table was produces by sorting the results from lowest to highest adjusted p value, and then returning the first 10 rows.

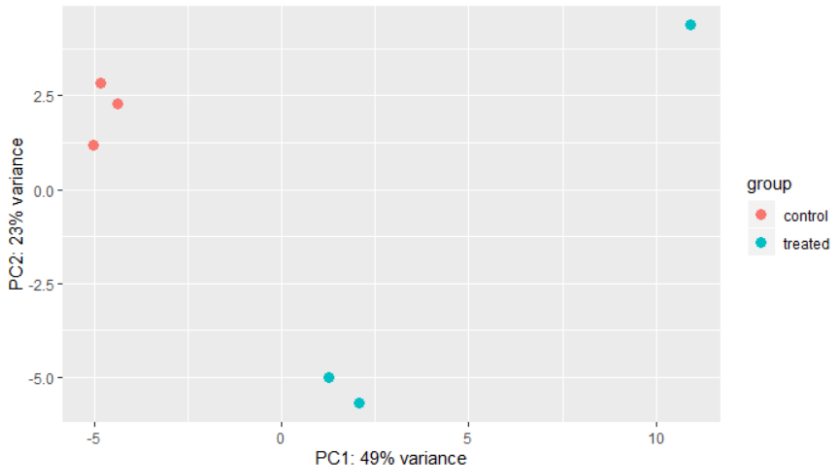

Figure 2: This figure is a plot that shows the Principle Component Analysis (PCA) of the samples in the analysis. PCA helps determine an early indication of the distances between sample gene expression profiles and detect possible batch effects. For the most part all of the samples are close together. However one of the treated samples is away from the other two. This may be a batch effect.

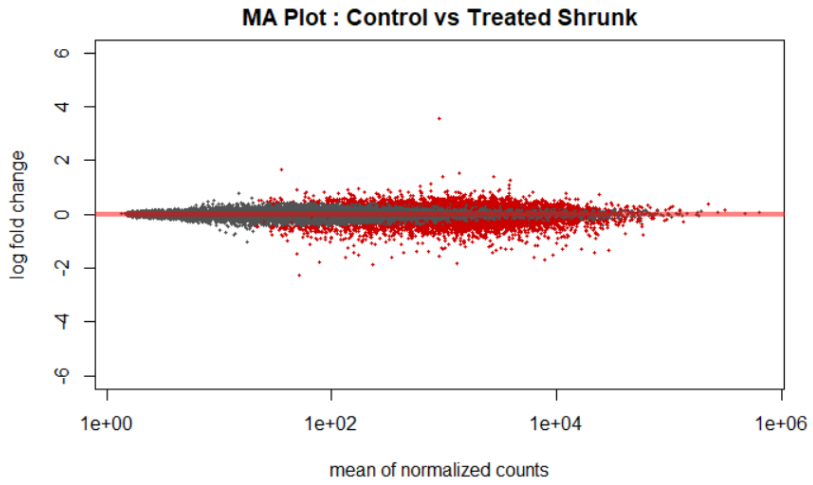

Figure 3: This plot shows the MA plot for the comparison between the control and treated samples. MA plots help visualize the difference in log fold changes across the entire dataset. The log fold change estimates were also shrunk because of DESeq's tendency to overestimate log fold change estimates. All of the points are close to the horizontal line with only a few points far away.

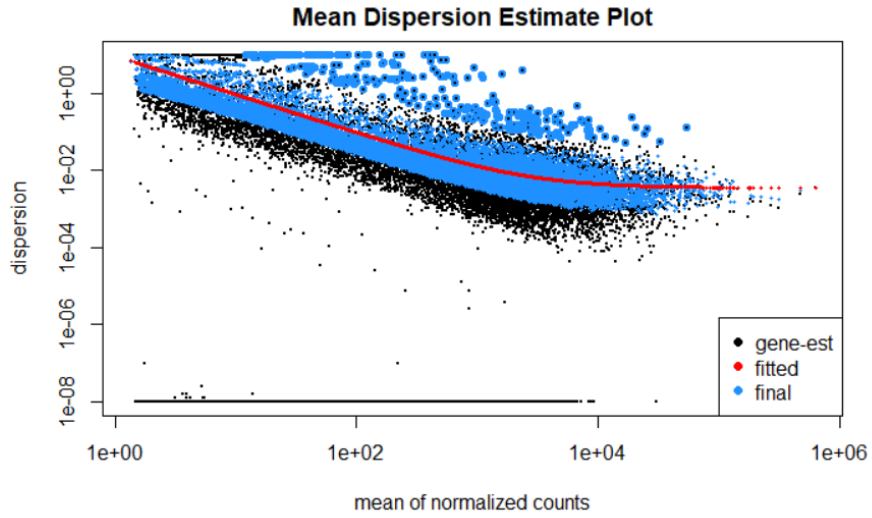

Figure 4: This figure shows the dispersion estimate plot calculated by mean. A dispersion plot allows one to visualize how much the variance deviates from the mean. The curve enables the researcher to more accurately identify differentially expressed genes when the sample sizes are small. Ideally, the data should scatter around the curve with the dispersion decreasing as the mean expression levels rise. Because this is the case, the dispersion plot looks ideal.

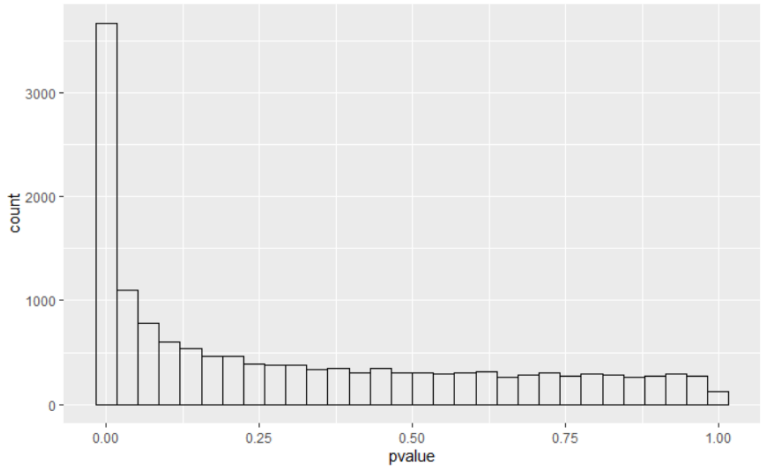

Figure 5: This figure shows the raw p-value histogram of the results. This histogram suggests an enrichment of low p-values. This is the expected result if there is an ideal amount of differentially expressed genes between the control and the treated sample.

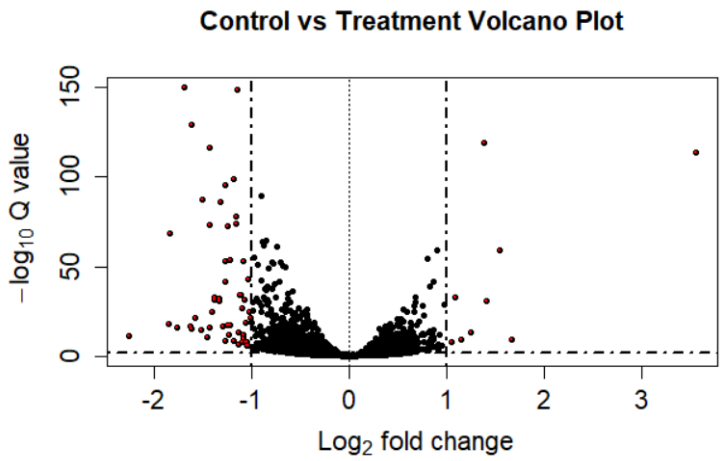

Figure 6: This figure shows a volcano plot for the results of determining differentially expressed genes. Volcano Plots are useful in visualizing differentially expressed genes. Every point represents a gene. If the gene had a log 2 fold change greater than 1 or less than -1, then it would be outside the two vertical lines and colored red. If the point is red then it was determined to be considered differentially expressed.

4 Discussion

In this experiment, the functionality of the NRDE2 gene was investigated by conducting an RNA seq analysis. The analysis focused on the comparison between MDA-MB-231 breast cancer cells that were transfected with 20nM control and transfected with NRDE2-targeting siRNAs. After conducting the analysis, 63 differentially expressed genes were identified statistically as well as considering for biological relevancy. Statistically speaking, the genes that were considered to be differentially expressed had to have an adjusted p value less than 0.5 – which left 3327 genes. But after considering each genes biological relevancy by only keeping genes with a gene expression two fold or greater, 63 genes were left. The analysis included determining the total number of reads and mapping rate of each sample when using Salmon (Figure 1). It also included using PCA to analyze the samples (Figure 2), making an MA Plot (Figure 3) to visualize the difference in log fold changes across the entire dataset, plotting the dispersion estimates to visualize how much the variance deviates from the mean (Figure 4), and constructing a raw p-value histogram (Figure 5) to determine if there is an enrichment of low p-values. These methods allow us to assess our dataset. In determining differentially expressed genes, a volcano plot (Figure 6) was made to visualize the results which are the differentially expressed genes.

Next steps in this analysis may be to begin interpreting the genes that are differentially expressed. A way to do this might be to cluster the differentially expressed genes to construct heatmaps to visualize which genes are being over/under expressed in each sample. Also, a GO term enrichment analysis may help determine the functions of the differentially expressed genes. Finding a pattern may help provide insight into the role of the NRDE2 gene itself in RNA interference.

5 References

https://www.ncbi.nlm.nih.gov/probe/docs/techrnai/

https://www.ebi.ac.uk/ena/data/view/PRJNA490376