Next generation sequencing analysis final project

Yao Chen

Introduction

NRDE2 (nuclear RNAi defective-2) is an essential protein in human that is responsible for suppressing intron retention in a subset of pre-mRNAs. Essential centriolar satellite protein(CEP131) is regulated by NRDE2 which involves in splicing of CEP131 pre-mRNA. NRDE2 depleted cells have shown significant reduction in CEP131 protein expression. As a result, NRDE2-depletion can result in genomic instability and DNA damage, because NRDE2's regulatory role in centrosome maturation and mitosis.

In this analysis, we investigated the gene expression profile upon NRDE2 gene silencing by transfecting breast cancer cells with NRDE2-targeting siRNAs, the RNA of the si-control and si-treated triplicated cells were harvested and sequenced following 48 hrs incubation. 72 transcripts were identified to have differential expression with greater than 2-fold change and p-value < 0.05 in NRDE2-depleted cells. (The triplicated RNA-seq data for control and NRDE2-depleted cells were downloaded from European Nucleotide Archive under project PRJNA490376).

Materials and methods

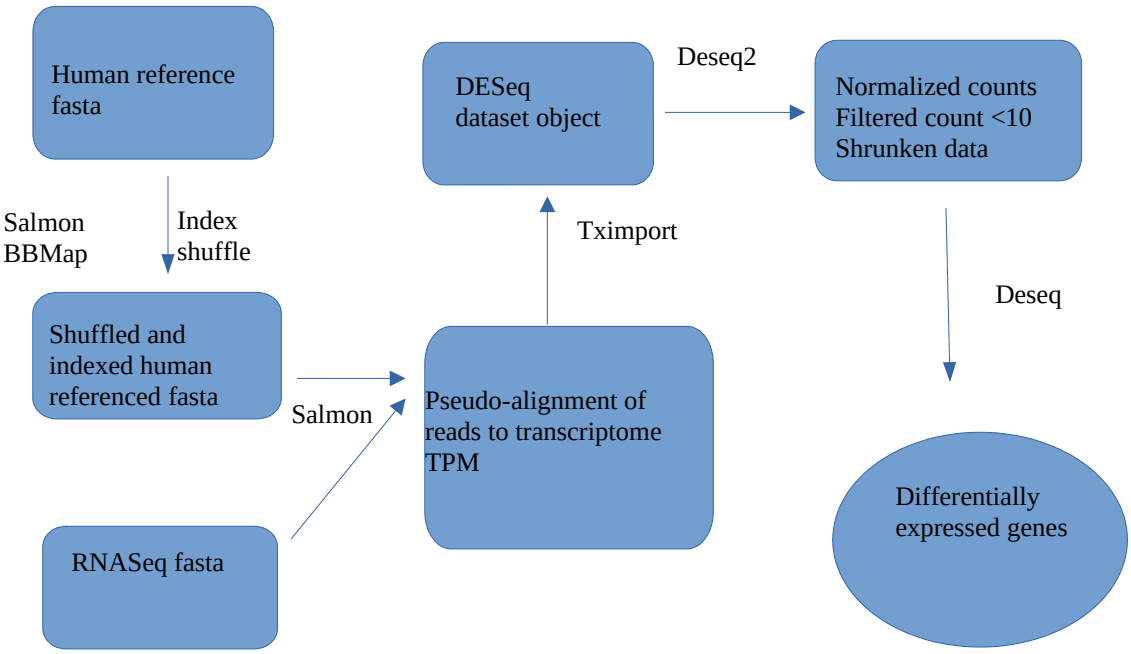

Figure 1. A flowchart description of the overall workflow.

The raw RNA-seq were generated single-end sequenced on Illumina NextSeq. The raw sequencing data were processed through FASTP tool to remove adapter and polyG sequences introduced on NextSeq platforms. The reads with length size <60 were removed. The quality of the processed fastq then assessed by multiqc tool and generated multiqc report. All the samples show sufficient amount of reads with relatively high duplication rates. Phred scores are greater 30 across the reads. Per base N content are low for all samples, which indicates a good sequencing quality. No significant adapter contamination was found in the samples.

The reference cDNA fasta for Homo_sapiens GRCh38 was downloaded from ensemble. The fasta files then normalized by PICARD tool to ensure the sequence names are suitable for downstream process, and 100 bases per line of each read. The normalized reference fasta then shuffled by BBMAP and indexed by salmon tool before pseudo-alignment. The library type was determined by salmon as stranded reverse strand library.

Salmon generated TPM from the pseudo-alignment to transcriptome, the TPMs then converted to gene counts and created a DESeq Data set in R by tximport. The raw count data then normalized using the median-of-ratios per feature by DESeq2 shrunken estimates were used to obtain better estimates of log2 fold-change for low expression genes, hierarchical cluster was plotted to see if the treatment and control form a cluster, PCA plot was also used to confirm if batch effect exists. The shrunken MA plot is more consistent in a broad range of counts, shrunken remove the noise associated with log2 fold changes from low count genes without requiring arbitrary filtering thresholds, which is easier for interpretation of log fold change. The candidate genes are selected by log fold change > 1 and p-adjusted value < 0.05 for biological significant. Gene-wise dispersion plot was used to estimate the correlation between normalized count and dispersion, the dispersion should be decrease with higher counts.

Results

Samples	Total reads	Mapping rate
SRR7819990	54937907	91.166%
SRR7819991	57962873	91.7893%
SRR7819992	51364911	92.7966%
SRR7819993	52,597,484	91.9953%
SRR7819994	53,912,823	92.2498%
SRR7819995	40,331,365	92.3223%

Table 1. QC metrics for total reads and mapping rate of each samples, all samples has good mapping rate.

72 statistically significant genes were chosen based on log fold change value greater than 1 and adjusted-p value less than 0.05.

Samples	Base mean	log2FoldChange	padj
ENSG00000196396.10	6552.33	1.15	2.44309e-159
ENSG00000175334.8	6391.07	1.65	3.86263e-150
ENSG00000206286.11	2773.56	-1.39350	8.18905e-124
ENSG00000163041.11	7909.36	1.66	2.33357e-108
ENSG00000105976.15	9418.09	1.53	3.72572e-100
ENSG00000101384.12	11619.72	1.29	4.91814e-99
ENSG00000124333.16	2719.53	1.48	9.21800e-94
ENSG00000128595.17	22621.17	1.46	9.21800e-94
ENSG00000117632.23	16619.17	1.34	1.12753e-87
ENSG00000213281.5	6883	1.18	8.26813e-78

Table 2. The 10 most significantly differentially expressed genes with log fold change value greater than 1 and adjusted P value less than 0.05.

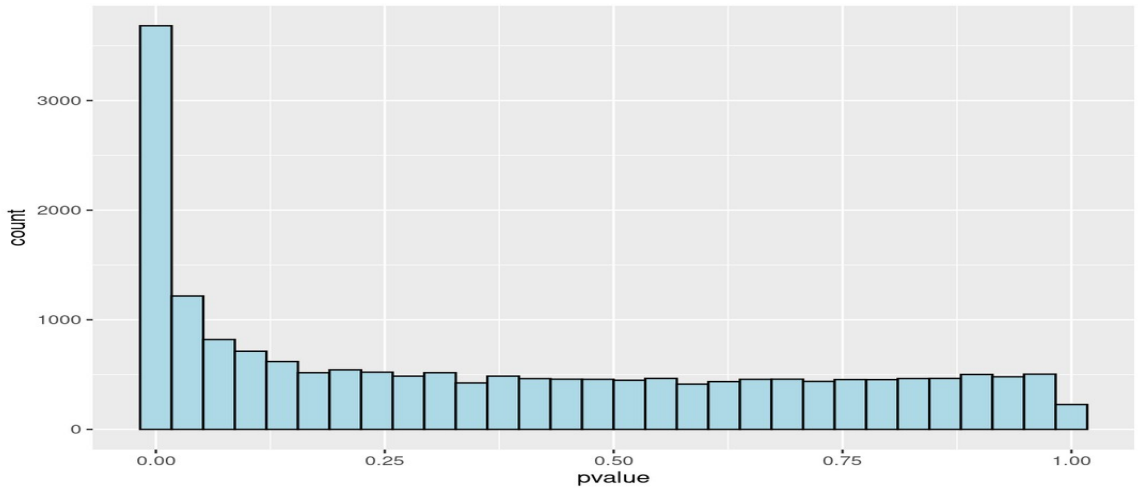

Figure 2. Raw p-value histogram shows enrichment of low p-values, it shows a large class of differentially expressed genes between treatment and control.

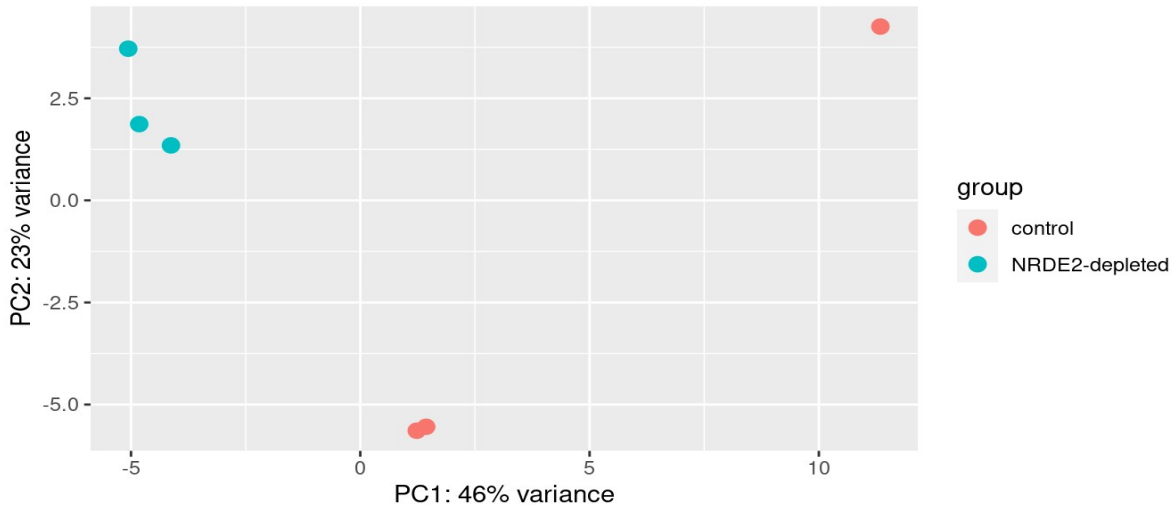

Figure 3. Principal Components Analysis (PCA) plot. This plot indicates the distances between sample gene expression profiles and help detect batch effects. One of the control samples has greater PC1 and PC2 than the rest of the control samples. Batch effect unlikely to happen in this case.

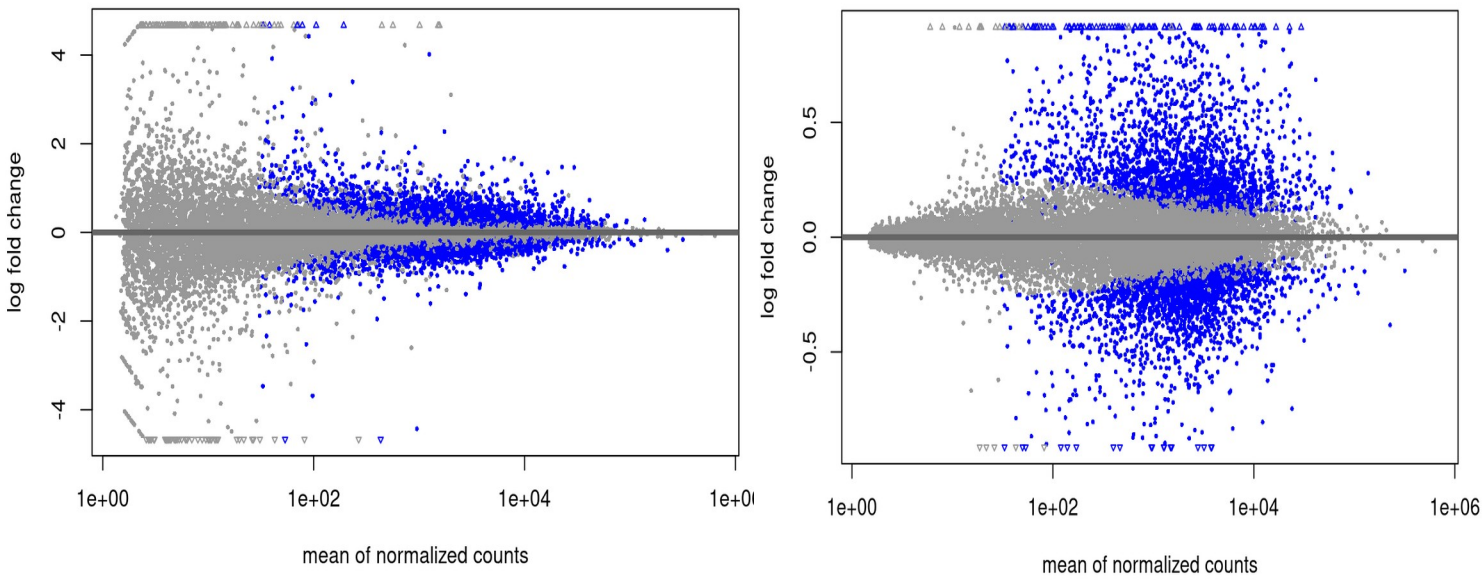

Figure 4. MA plot, on the left is a unshrunked MA plot, on the right is a shrunken MA plot, the blue dots show both up-regulated and down-regulated genes between NRDE2-depleted breast cancer cells and control breast cancer cells. The shrunken plot is more consistent in broad range of counts, shrunken remove the noise associated with log2 fold changes from low count genes without requiring arbitrary filtering thresholds. Easier for interpretation of log fold change.

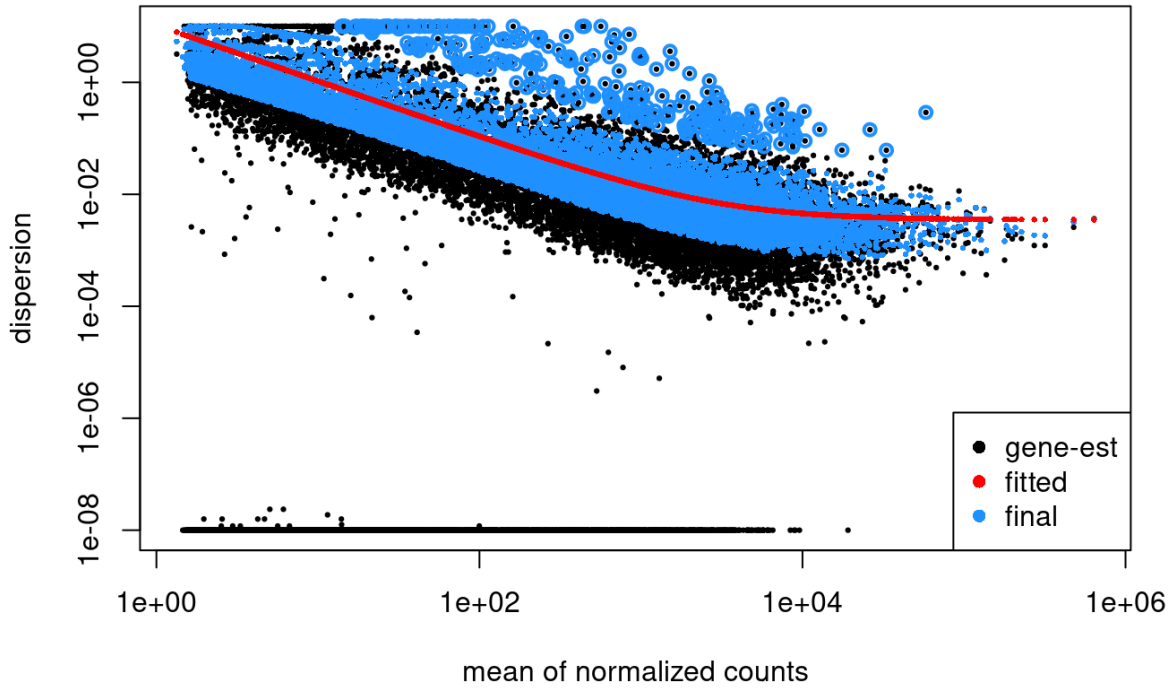

Figure 5. Dispersion-by-mean plot, the blue dots are shrunken estimates of dispersion per gene, the red line is curve fit, the black dots are original Maximum Likelihood estimate. Dispersion parameter in the negative binomial model is a key parameter that must be estimated by statistical packages of DGE based on count data. DESeq2 will typically shrink the initial dispersion values by fitting a line and then shrinking the dispersion values towards the fitted line. The curve fit shows increasing count concordant with decreasing dispersion, all the outliers shrunk towards the fitted value. The shrinkage method works in this case.

Discussion:

The analysis of the RNA-seq data is successful as the raw RNA-seq data shows high mapping rate > 90% and abundant of high quality reads without adapter contamination. DESeq2 analysis of the TPM data from salmon tool is robust and reliable. The QC metrics for DESeq2 data are good, most of the raw differentially express genes shows low p-Value. PCA plot shows the experiment not experiencing batch effect. MA plot and dispersion-by-mean plot indicates the shrinkage method works to increase the signal to noise ratio which assist the interpretation of the data. At the end, 72 differentially expressed genes are identified based on the biological and statistically significant metrics with high level of confidence.

Reference

Jiao AL, Perales R, Umbreit NT, et al. Human nuclear RNAi-defective 2 (NRDE2) is an essential RNA splicing factor. RNA (New York, N.Y.). 2019 Mar;25(3):352-363. DOI: 10.1261/rna.069773.118.