Daniela Quijano

NGS Spring 2022

Final Report: Differential Gene Expression Analysis as a response to NRDE2-targeting siRNAs in MDA-MB-231 breast cancer cell lines

Introduction

The goal of this analysis is to investigate the impact that nuclear RNAi defective-2 (NRDE2) transfection has on the overall gene expression of MDA-MB-231 breast cancer cell lines. MDA-MB-231 cells are rated 'a suitable transfection host' and were isolated from an adenocarcinoma patient. Because this study delivered the siRNA via transfection, this was a suitable cell line to use. NRDE2 is known to work closely during the RNA splicing process but human function is not very well understood. The treated samples were transfected with a small interfering RNA to silence the NRDE2 gene. Treated samples were compared to control samples to see what kinds of genes were affected when NRDE2 is silenced. The goal of this analysis is to process the raw reads produced on the NextSeq5000 platform to obtain a set of differentially expressed genes that can relay more information about the effect that silencing NRDE2 has on gene expression.

Methods

Fig.1 Workflow conducted to process files from study SRP161520

In order to conduct this analysis the raw fastq files from the study SRP161520 were obtained from the European Nucleotide Archive (ENA) and were processed both on the NYU Greene HPC Cluster and using R with Rstudio as an IDE. The workflow that was used in order to process the files is summarized in Fig. 1. The single reads were trimmed fastp which automatically detects Illumina adapters, in this case, those used in the Illumina NextSeq500 platform. Given that the average length of the reads was about 75 bp, there was no need to change the parameter to ensure reads reached this length. Reads that were less than 15bp are aromatically removed by fastp. Additionally, because fastp automatically detects and trims the polyG sequence artifacts that are a result of the sequencing process, no specific argument was passed to fastp. As seen in the MULTIQC report, the average N content across the reads for all samples was marginal and there was no need to change the default parameter that removes reads with more than 5 'N' nucleotides. The per-base sequence content section of the MULTIQC showed warnings since the difference between A and T, or G and C is greater than 10% toward the beginning of the reads. After position 12, the reads stabilize and the base content becomes more even as it should be. Because the pattern stabilizes, the warning is not that concerning. In general, some reasons for these kinds of errors include overrepresented sequences or biased library composition.

Before mapping the filtered reads to a reference, Salmon, a pseudoaligner, needs an index. The index was computed using GRCh38 version of the human genome (cDNA). The index was computed with k value 31 which represented the 'minimum acceptable length for a valid match' which worked well in this case since the reads were on average 74bp on average and salmon documentation recommends k=31 for reads of about that length. Salmon was then executed in mapping mode in order to align reads to a computed reference index file corresponding the GRCh38 version of the human genome. Salmon was allowed to infer the library type with the parameter -l A and the single end library was inferred to be stranded. Salmon was given the parameter –validatemappings which, according to documentation, uses a more sensitive algorithm which can improve the sensitivity and specificity of mapping. Salmon outputs the quantification estimates as quant.sf files which were imported into r using tximport which summarizes the counts into a matrix that can be used with DESeq2.

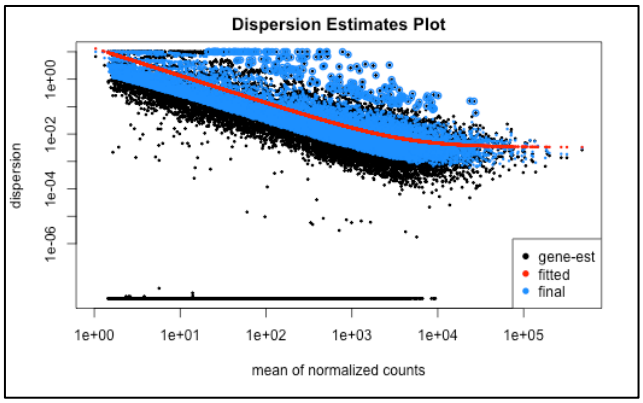

Fig.2 Dispersion Estimates Plot

Deseq2 uses the Benjamin Hochberg (BH) correction method and reports the p values as adjusted p values (Fig.3). P value corrections help minimize the false discovery rate. Although the shrunken values are reported as part of the DESeq results, I am not reporting the shrunken values in the MA plot (Fig.4). The table that contains the DESeq results is in the .tar file provided with this report and is titled 'all_deseq_results_genes.txt'.

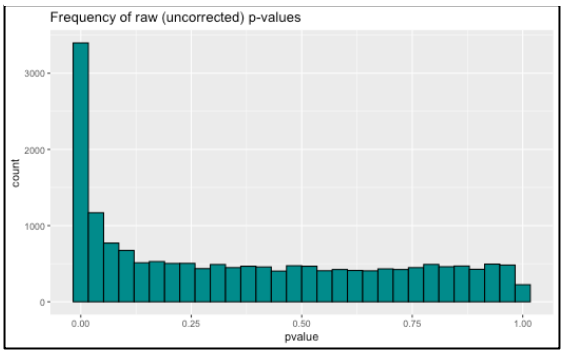

Fig.3 Raw (uncorrected) p value histogram

Results

DESeq was used in order to contrast the differential gene expression of MDA-MB-231 cells that were treated with si-NRDE2 with respect to the gene expression of control samples. When dimensionality analysis was conducted, the principal component with most variance (PC

(45%) was whether a sample belonged to the treated or control group (Fig.5). There was one of the samples, treated3, that although was still belonged to the linearly separable treated group, did not cluster with the other two treated samples. This observation was confirmed using a dendogram with hierarchical clustering that shows treated3 sample in a clade with larger height than the other two samples in the same group. Upon inspection of the mapping rate and read counts from STAR in sample treated3, we can see that this sample has less overall reads than the other samples though it still has comparable mapping rates (Fig.8). Based on the dispersion estimates plot (Fig. 6) the distribution of the confirms the data has a distribution that follows what is expected from a typical RNA Seq experiment. As the mean of the normalized counts increases, the fitted dispersion stabilizes. In general the raw p-value distribution showed that p values close to 0 .05 had the highest p value frequencies-a preliminary indication that there was likely going to be a considerable amount of significantly expressed genes at a 0.05 FDR threshold. The top 10 differentially expressed genes and their functions are summarized in Fig.7. In total there were 2969 significantly expressed genes when using BH-corrected FDR of 0.05 out of the total 13939 genes that were processed by DESeq after removing genes that contained NA values. There were 494 genes with log fold change magnitude greater than 2. Some of the genes with log-fold change magnitude greater than 2 can be seen on the edges of the MA plot (Fig. 5). In the top 10 differentially expressed genes there was a trend for genes involved in cell growth, immune function and epigenetic controls. In general, it seems like si-NRDE2 plays a role in pathways that regulate cell growth, apoptosis and the structural integrity of tissue, and spliceosome machinery. The paper that acted as the data source for this investigation confirms the findings of this analysis namely, the trend seen in the top differentially expressed genes.

In conclusion, it seems that the small RNA that targets the NRDE2 gene leads to the dysregulation of cellular processes involved in cell growth, epigenetic regulation, and control of RNA splicing machinery.

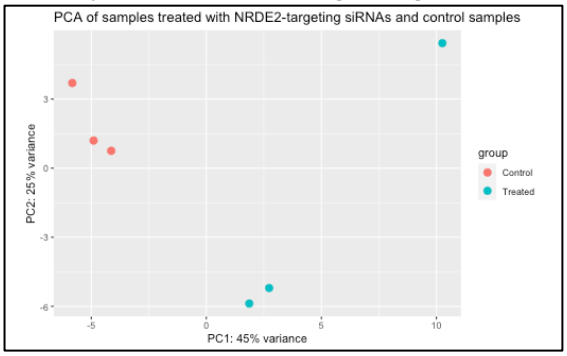

Fig.4 Principal Component Analysis of samples

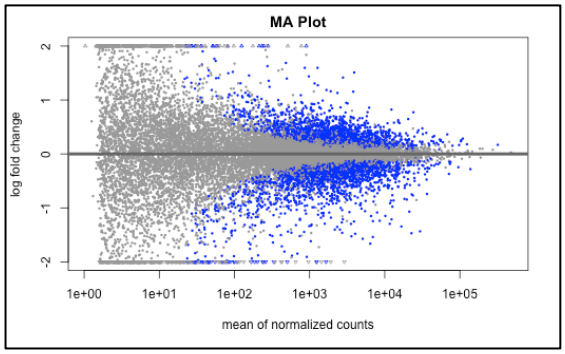

Fig.5 MA Plot of uncorrected p values

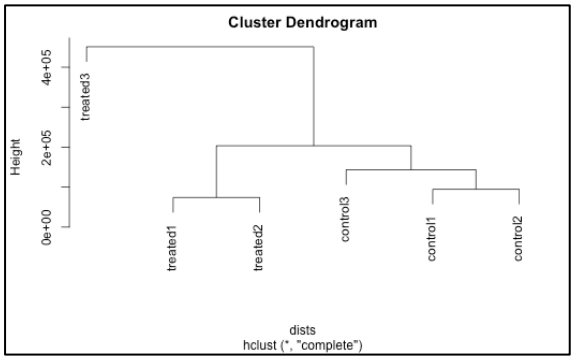

Fig.6 Dendogram showing hierarchically clustered samples

STAR Alignment Statistics

Sample	Mapping Rate	Read Count
Control 1	91.29%	55,279,804
Control 2	91.95%	58,328,433
Control 3	92.93%	51,663,295
Treated 1	92.17%	52,896,723
Treated 2	92.42%	54,192,653
Treated3	92.49%	40,548,619

Fig.8 STAR mapping statistics

Top 10 Significantly Expressed Genes (By p-value)

Gene	Gene Function	baseMean	log2FoldChange	lfcSE	stat	pvalue
ENSG00000196396.10	Tyrosine-protein phosphatase which . Mediates dephosphorylation of EIF2AK3/PERK	6573.21474	-1.151273092	0.04092245	-28.133047	3.86E-174
ENSG00000175334.8	BAF nuclear assembly factor 1,roles in nuclear assembly, chromatin organization	6417.20625	-1.661613007	0.06309592	-26.334714	7.68E-153
ENSG00000163041.11	Variant histone H3 epigenetic imprint of transcriptionally active chromatin.	6943.15662	-1.540007765	0.06346564	-24.26522	4.57E-130
ENSG00000206286.11	Vacuolar protein sorting-associated protein 52 homolog, involved in vesicle trafficking	2780.66195	1.415834147	0.05899082	24.0009225	2.72E-127
ENSG00000101384.12	involved in the mediation of Notch signaling, involved in cell-fate decisions during hematopoiesis	11654.9029	-1.295561676	0.05909364	-21.923876	1.54E-106
ENSG00000143384.13	Involved in the regulation of apoptosis versus cell survival	21215.4432	-1.06936988	0.04901283	-21.818161	1.56E-105
ENSG00000128595.17	Vitamin K-dependent carboxylation of N-term. glu residues.	22673.9544	-1.469162183	0.06771202	-21.697215	2.18E-104
ENSG00000117632.23	Prevents assembly and promotes disassembly of microtubules.	16649.9119	-1.344897819	0.06625475	-20.298889	1.32E-91
ENSG00000124333.16	Targeting and/or fusion of transport vesicles to their target membrane. Needed for eosinophil and neutrophil degranulation.	2710.21936	-1.461995652	0.07607126	-19.218765	2.58E-82
ENSG00000144028.15	Small nuclear ribonucleoprotein U5 subunit 200, involved in pre-mRNA splicing and spliceosome complexes.	12273.7426	0.8226123	0.04332191	18.9883652	2.13E-80

Fig.7 Top 10 Differentially expressed genes and corresponding function and statistical values

Works Cited

ATCC. (n.d.). *MDA-MB-231*. Retrieved May 12, 2022, from https://www.atcc.org

Embl-Ebi. (n.d.). *Project: PRJNA490376*. Ena Browser. Retrieved May 12, 2022, from https://www.ebi.ac.uk/ena/browser/view

Love, M. S. I. A. (2022, April 26). *Analyzing RNA-seq data with DESeq2*. Deseq2. Retrieved May 12, 2022, from http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html

Salmon 1.8.0 documentation. (n.d.). Salmon Documentation. Retrieved May 12, 2022, from https://salmon.readthedocs.io/en/latest/