Differential Gene Expression Analysis in cancer cells with silenced NRDE2 gene

Rahi Shah

Introduction

NRDE2 has been shown to be an evolutionarily conserved protein, however, its role in mitotic progression and genomic stability is unknown. In this experiment, researchers investigated the transcriptional profile changes following a knockdown of the *NRDE2* gene using NRDE2-targeting siRNAs. MDA-MB-231 breast cancer cells were transfected with either a 20nM control or the siRNAs, and. I performed differential gene expression analysis (DGE) with the data, which is a technique used to compare gene expression levels between distinct sample groups (Rosati et al, 2024)). This is a very valuable tool since it can allow us to compare and identify the genes that are involved in various biological processes, diseases, and also study response to a particular treatment, which is the case for our study. There are several steps involved in conducting DGE analysis: pre-processing and normalization, selection of model that suits the distribution of the data (in this particular study, I will be using DESeq2), identify genes with significant expression changes, and lastly, visualization of results and using it for further downstream analysis depending on the requirements of the project.

Methods

For this study, six different RNA-seq samples were analyzed: 3 control and 3 experimental (treated with siRNA). The reads that were obtained were processed using the nf-core/rnaseq pipeline (version 3.14.0) on a high performance computing system. The pipeline was executed using Nextflow (version 24.10.4).

The reference transcriptome and annotation files were downloaded using the specification provided in the pipeline documentation, and the latest versions were retrieved from the e!Ensembl.

Trimming and quality control were handled by TrimGalore (version 0.6.7) and the adaptors were trimmed using CutAdapt (version 3.4). The quality control of the reads were performed and assessed by FastQC (version 0.12.1) and a multiQC report was analyzed to look at the quality of the sequencing.

The transcript quantification and indexing was performed using the pseudo-aligner Salmon (version 1.10.1). Ensembl GRCh38 (release 113) was used for indexing, and the output files (quant.sf) were used for downstream analysis of the differentially expressed genes.

The sample sheets for this study were specified in the comma separated values (csv) format, where the sample name, fastq_1, and strandedness was specified. The sample sheet was also formatted to indicate that the reads are sigle-end.

For differential gene expression analysis, the transcript level estimates were summarized using the tximport package. The GTF annotation produced a tx2gene file, and the data set from DESeq2 was created using the function txi(). Differential expression analysis was performed using the DESeq2 (version 1.38.3) package on R studio via the HPC system. The log2 fold changes estimates obtained from the DESeq2 analysis were shrinked using the lfcShrink() function, where the method 'apelgm' was applied. The plots before and after shrinkage were visualized using the function plotMA().

Some other visualizations were conducted using Principal Component Analysis (PCA), where the variance stabilized data (vsd) was used to cluster based on the condition. Dispersion plot was also produced to assess the model fit as well as gene-wise variance estimates. A p-value histogram was created to look at the distribution of raw p-values and get an overall idea of the distribution of significantly different genes. The p-value was also adjusted using the Benjamini-Hochberg False Discovery Rate (FDR) method. Genes with an adjusted p-value (padj) < 0.05 were considered significantly differentially expressed, and then top 10 results for those were examined.

Results

Once the nf-core/rnaseq pipeline was executed, a multiQC report was produced, and it was used to assess the total number of reads per sample as well as the percent of aligned reads. This can be seen in table 1 below.

Samples	Total number of reads	% aligned
control1	61235909	91.10%
control2	63764300	92.28%
control3	55938770	93.24%
treated1	57677275	92.59%
treated2	58907058	92.87%
treated3	46379548	92.85%

Table 1: Total number of reads and % aligned per sample

Once the DESeq2 analysis was conducted, the top 10 most significant genes were visualized. Figure 1 below shows the top 10 genes ordered with respect to decreasing p-adjusted values

	log2FoldChange	pvalue	padj
	<numeric>	<numeric>	<numeric>
ENSG00000175334	1.65818	1.57213e-137	2.53176e-133
ENSG00000163041	1.66799	3.19987e-118	2.57654e-114
ENSG00000196396	1.15095	8.30924e-101	4.46040e-97
ENSG00000105976	1.57137	2.24035e-95	9.01965e-92
ENSG00000128595	1.48572	9.10202e-94	2.93158e-90
ENSG00000101384	1.31508	3.01267e-86	8.08602e-83
ENSG00000124333	1.48742	4.65482e-86	1.07087e-82
ENSG00000117632	1.34686	2.54574e-79	5.12457e-76
ENSG00000145919	1.29112	2.72808e-67	4.68204e-64
ENSG00000213281	1.18569	2.90738e-67	4.68204e-64

Figure 1: top 10 most significant differentially expressed genes (in reference to control)

Statistically significant genes at an FDR of <0.05 were also identified, which amounted to a total of 3690 that were significantly different in control vs. treated samples. To identify the genes that were significant as well as upregulated in the samples treated with siRNA, an adjusted p-value threshold of 0.05 and a log2 fold change value of >0 was chosen. A total of 1900 genes were identified to be significantly upregulated in the treated samples. On the other hand, using the same criteria, 1709 genes were identified to be significantly downregulated.

In figure 2 below, a PCA plot shows the clustering patterns between the samples based on the condition (control vs. treated). The data used for clustering for variance stabilized. On the x-axis is principal component 1, which explains 47% of the variance in the data, and on the y-axis is principal component 2, which explains 21% of the variance in the data. A very distinct clustering pattern is visible between the control and treated samples along PC1, which is shown in figure 2 by red and blue dots respectively.

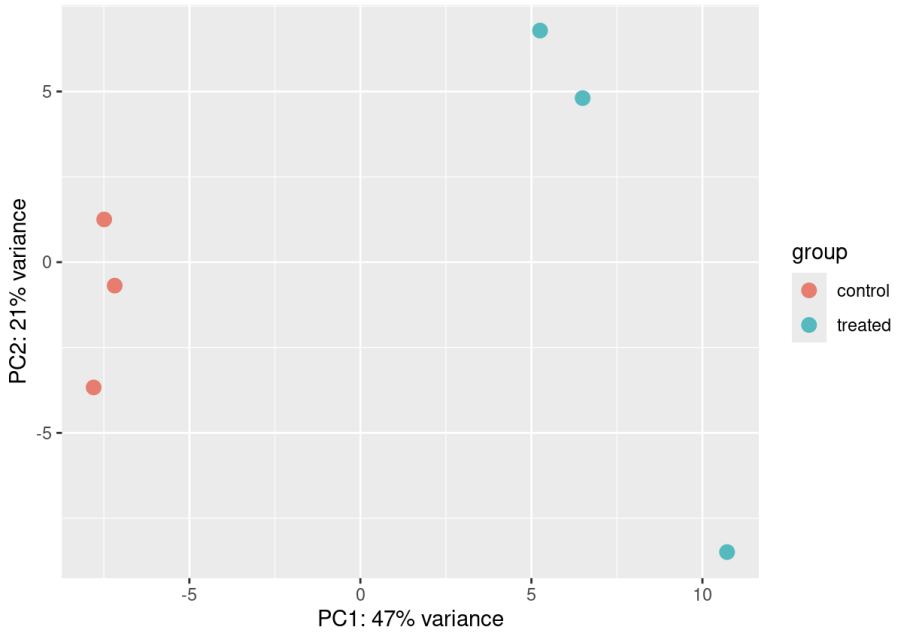

Figure 2: a PCA plot depicting the clustering pattern between control (red dots) and treated (blue dots) samples.

To reduce the variance in the log2 fold changes of the genes, lfc shrinkage was performed. Shrinkage helps with stabilizing the LFC estimates, especially in the genes that are lowly expressed. This can be seen in Figure 3, where MA plot A on the left shows high variance in lowly expressed genes, whereas MA plot B on the right shows most stabilized graph with reduced noise, which also helps in solidifying the differentially expressed genes that were identified.

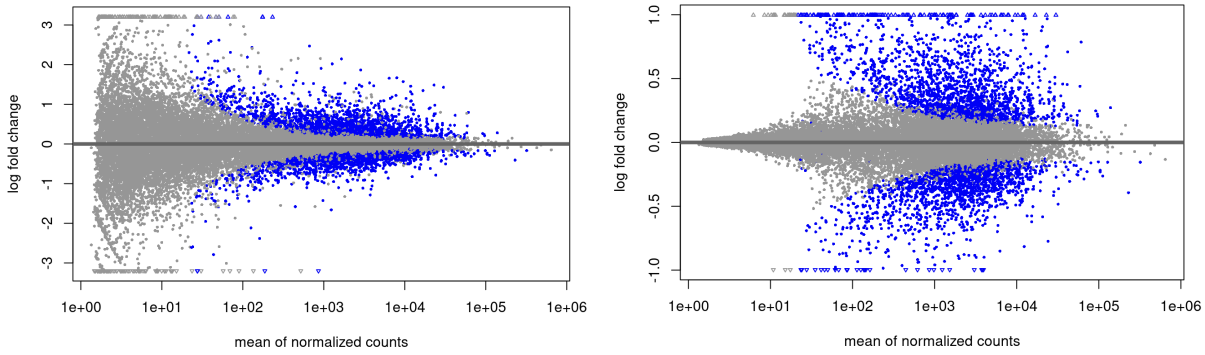

Figure 3A: MA plot using the log2FC values before shrinkage, and 3B is after shrinkage

Figure 4 shows the dispersion plot estimates, with the gene-wise dispersion estimates on the y-axis and the mean of normalized counts on the x-axis. The black dots in the graph reflect the individual gene estimates based on their counts, the red dots form the fitted line which captures the dispersion trends and smooths the noise, and the blue dots are the shrunken estimates that are used in the DESeq2 model for each gene. The blue dots, as expected, have a much lower dispersion, which again reduces noise in our data and gives us more confidence in the differentially expressed genes that are identified.

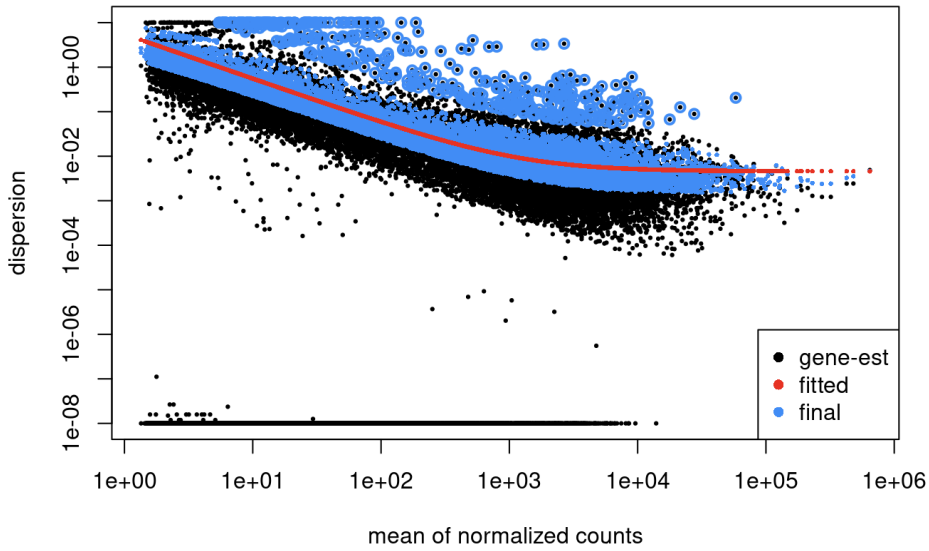

Figure 4: gene-wise dispersion estimates

Lastly, figure 5 below shows the graph of raw p-values obtained after the DESeq2 analysis. The y-axis shows the frequency or the number of genes (within each bin), and the x-axis shows the distribution of p-values. The left skewness in the graph is informative and tells us that there are indeed differentially expressed genes in our dataset, since the peak closer to 0 suggests a lower p-value. A uniform distribution, on the other hand, would've meant a lack of DEGs between our conditions.

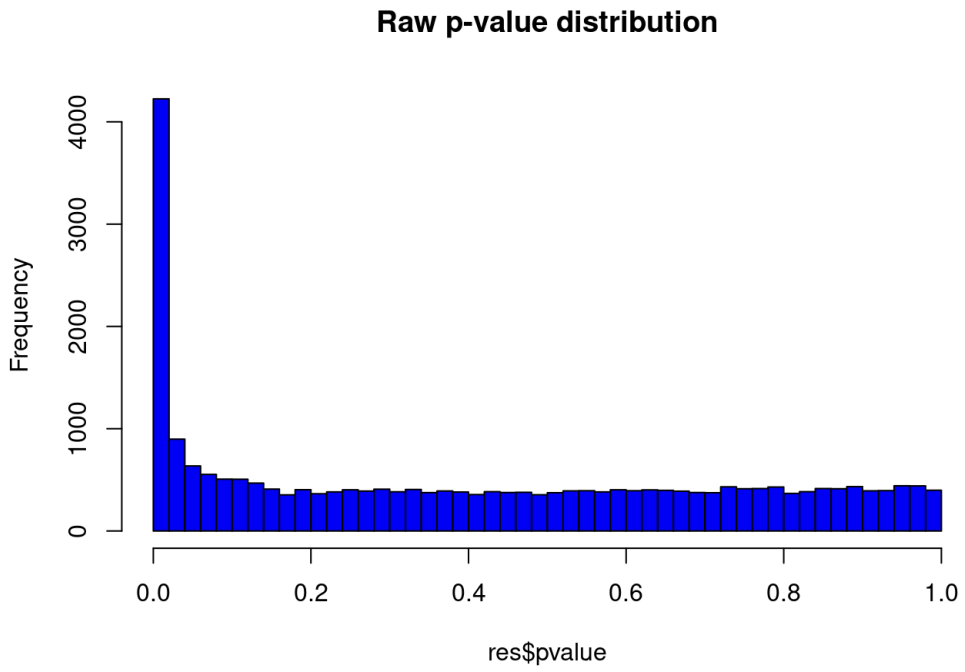

Figure 5: distribution of raw p-values graphed against the frequency

Discussion

Initial quality assessment using FastQC, which were then summarized through MultiQC, revealed several notable issues across samples.

Firstly, all six samples exhibited high levels of duplication and failed the module check. This may potentially indicate PCR amplification bias or over-sequencing of low-complexity libraries, which can reduce the diversity of reads. This can also skew expression estimates of some of the genes and interfere with downstream analysis, such as while identifying DEGs.

The "Per base sequence content" module in FastQC reported red alerts in all samples, meaning that they failed the module check. A failure in this module typically indicates a disproportionate distribution in the nucleotides (A and T, C and G), the difference sometimes exceeding 20% (Babraham Bioinformatics). Inspecting the graphs in detail from the multiQC report revealed a drop in nucleotide A towards the tail end of the sequence (the 3’ end), from nucleotides 70-75, as well as a high variability in the first 10 nucleotides in a read. Although the variability in nucleotides composition at the 5’ end of the reads is normal, it still warrants considerations in regards to downstream bioinformatics analysis.

Lastly, some samples triggered warnings in sequence length distribution, suggesting the presence of reads of variable lengths. Despite these issues, the overall read quality and

post-trimming statistics were acceptable for downstream differential expression analysis, especially some of the other important statistics such as: per sequence quality scores, GC content, and some other general statistics regarding the percentage of sequence aligned. The duplication rate statistic is worrisome, however, since there's anywhere from 63-67% duplications in all the samples, and the library prep needs to be optimized for better future sequencing runs.

A total of 3690 genes were identified to be differentially expressed in control vs. siRNA treated samples. It would be interesting to look into these genes in detail, and further downstream analysis can be conducted, such as looking at gene networks, pathways, and performing gene ontology enrichment, to understand NRDE2's impact on the gene expression network. If key regulatory genes or biomarkers are identified, further validation can be performed experimentally and the significance of certain genes and pathways can be searched for using previously published literature.

Appendix

Code for running the nf-core/rnaseq pipeline:

```
#!/bin/bash
#
#SBATCH --nodes=1
#SBATCH --tasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --time=8:00:00
#SBATCH --mem=4GB
#SBATCH --job-name=rna_seq
#SBATCH --account=pr_284_general
#SBATCH --mail-type=FAIL
#SBATCH --mail-user=rs8579@nyu.edu
module purge
module load nextflow/24.10.4
sample_sheet="/scratch/rs8579/final_project/project.2025/sample_sheet.csv"
genome_dir="/scratch/rs8579/final_project/project.2025"
json_params="/scratch/work/courses/BI7653/hw8.2025/rnaseq.json"
config_file="/scratch/work/courses/BI7653/hw8.2025/rnaseq.config"
nextflow run nf-core/rnaseq -r 3.14.0 \
  --input $sample_sheet \
  --outdir res \
  --fasta $genome_dir/Homo_sapiens.GRCh38.dna_sm.primary_assembly.fa.gz \
  --gtf $genome_dir/Homo_sapiens.GRCh38.113.gtf.gz \
  --extra_salmon_quant_args="--gcBias" \
  --save_reference \
```

```
  --skip_alignment true \ #since we are only using salmon for pseudo alignment
  --pseudo_aligner salmon \
  -profile nyu_hpc \
  --account pr_284_general \
  -c $config_file \
  --paramsFile $json_params
module purge
echo "_ESTATUS_ [NEXTFLOW]: $?"
```

Sample sheet:

sample	fastq_1	strandedness
control1	/scratch/rs8579/final_project/project.2025/SRR7819990.fastq.gz	auto
control2	/scratch/rs8579/final_project/project.2025/SRR7819991.fastq.gz	auto
control3	/scratch/rs8579/final_project/project.2025/SRR7819992.fastq.gz	auto
treated1	/scratch/rs8579/final_project/project.2025/SRR7819993.fastq.gz	auto
treated2	/scratch/rs8579/final_project/project.2025/SRR7819994.fastq.gz	auto
treated3	/scratch/rs8579/final_project/project.2025/SRR7819995.fastq.gz	auto

Software versions used in the nf-core/rnaseq workflow:

Process Name	Software	Version
CUSTOM_DUMPSOFTWAREVERSIONS	python	3.11.7
	yaml	5.4.1
CUSTOM_GETCHROMSIZES	getchromsizes	1.16.1
DESEQ2_QC_PSEUDO	bioconductor-deseq2	1.28.0
	r-base	4.0.3
FASTQC	fastqc	0.12.1
FQ_SUBSAMPLE	fq	0.9.1 (2022-02-22)
GTF2BED	perl	5.26.2
GTF_FILTER	python	3.9.5
GUNZIP_FASTA	gunzip	1.10
GUNZIP_GTF	gunzip	1.10
MAKE_TRANSCRIPTS_FASTA	rsem	1.3.1
	star	2.7.10a
SALMON_INDEX	salmon	1.10.1
SALMON_QUANT	salmon	1.10.1
SE_GENE	bioconductor-summarizedexperiment	1.24.0
	r-base	4.1.1
TRIMGALORE	cutadapt	3.4
	trimgalore	0.6.7
TX2GENE	python	3.9.5
TXIMPORT	bioconductor-tximeta	1.12.0
	r-base	4.1.1
Workflow	Nextflow	24.10.4
	nf-core/rnaseq	3.14.0

References

Ensembl genome browser 113. (n.d.-b). http://www.ensembl.org

Jiao1, A. L., Perales3, R., Umbreit4, N. T., Haswell1, J. R., Piper6, M. E., Adams1, B. D., Pellman4, D., & and, S. K. (1970, January 1). *Human nuclear RNAi-defective 2 (NRDE2) is an essential RNA splicing factor*. RNA. https://rnajournal.cshlp.org/content/25/3/352.full

Michael Love [aut, cre]. (2021, February 22). *Variancestabilizingtransformation: Apply a variance stabilizing transformation (VST) to the... in DESEQ2: Differential gene*

expression analysis based on the negative binomial distribution. varianceStabilizingTransformation: Apply a variance stabilizing transformation (VST) to the... in DESeq2: Differential gene expression analysis based on the negative binomial distribution. https://rdrr.io/bioc/DESeq2/man/varianceStabilizingTransformation.html

Per Base Sequence Content. Per base sequence content. (n.d.). https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html

Rosati, D., Palmieri, M., Brunelli, G., Morrione, A., Iannelli, F., Frullanti, E., & Giordano, A. (2024). Differential gene expression analysis pipelines and bioinformatic tools for the identification of specific biomarkers: A review. *Computational and structural biotechnology journal, 23,* 1154–1168. https://doi.org/10.1016/j.csbj.2024.02.018