**BIOL792-1017 Independent Project Report (Ryan Kyger)**

**Introduction:**

ER stress has been implicated numerous diseases such as autoimmune diseases, neurodegeneratve diseases, cardiovascular diseases, diabetes, and cancer (Zhang and Kaufman 2008; Hotamisligil 2010; Cnop et al. 2012; Wang and Kaufman 2014; Hetz and Saxena 2017). ER stress is disrupted ER homeostasis; common causes of ER stress include infection, hypoxia, glucose starvation, and protein folding diseases (Cnop et al. 2012; Wang and Kaufman 2014; Almanza et al. 2019). Moreover, the ER also communicates with the mitochondria via physical contact, where this communication is involved in regulating processes such as calcium signaling, lipid metabolism, autophagy, and apoptosis (de Brito and Scorrano 2008; Verfaillie et al. 2012; Malty et al. 2017, Lebeau et al. 2018; Almanza et al. 2019).

Furthermore, under ER stress there is increased and ER-mitochondria contact, as well as changes in mitochondrial morphology and metabolism (Bravo et al. 2011; Verfaillie et al. 2012; Lebeau et al. 2018). However, no one has looked specifically for changes in mitochondrial gene expression.

The main goal of my project was to answer this question: *How does ER Stress Affect Mitochondrial Gene Expression?* For my project, I re-analyzed RNA-seq data from Kimmig et al. 2012, where the researchers had looked at changes in RNA expression in ER stressed *S. pombe* (fission yeast) as compared to an unstressed control. The focus of their paper was on the genes that are regulated by the ER stress sensing protein Ire1. Furthermore, I was able to get through the entire RNA-seq analysis workflow, as I was able to make FastQC quality graphs, a PCA plot, a volcano plot, and heatmaps, and a table of differentially expressed genes. However, there were quality issues with the RNA-seq data, which leave my main question largely unanswered, and casts serious doubt on the results and conclusions of Kimmig et al. 2012. Nonetheless, this project was a valuable learning experience.

**Results:**

Detailed information about the RNA-seq data used in this project can be found in Table 1 in the Main Figures section.

FastQC was used to assess the quality of the RNA-seq data. The results of this quality assessment showed that all four samples: Untreated Replicate 1, Untreated Replicate 2, DTT Treated Replicate 1,

DTT Treated Replicate 2, have high per-base sequence quality as determined by phred score. Thus, all samples appeared to be of good quality. Additional details about the FastQC reports can be found in the Methods section.

HISAT2 was used to align the RNA-seq reads for each of the four samples to the genome. The results of this alignment (Table 2 Main Figures) showed high sequence alignment for Untreated Replicate 1 (90.13%) and DTT Treated Replicate 1 (94.56%). But, Untreated Replicate 2 (1.44%) and DTT Treated Replicate 2 (4.01%) had very low sequence alignments.

Principal component analysis (PCA) was conducted on all four samples using the “DeSeq2” R package.

The results of this analysis (Figure 1 Main Figures) showed that the samples do not cluster by treatment, as neither Untreated Replicate 1 and Untreated Replicate 2 nor DTT Treated Replicate 1 and DTT Treated Replicate 2 cluster with each other.

A correlation heatmap for all four samples was constructed using the “DeSeq2”, “pheatmap”, and “RcolorBrewer” R packages. The results of this analysis (Figure 2 Main Figures) showed that the samples do not correlate by treatment, as neither Untreated Replicate 1 and Untreated Replicate 2 nor DTT Treated Replicate 1 and DTT Treated Replicate 2 correlate with each other as determined by Pearson’s coefficients.

An MA plot of the differential expression results was made using the “DeSeq2” R package.

This plot (Figure 3 Main Figures) shows that there are 17 genes that are down-regulated during ER stress as compared to the untreated control condition.

A volcano plot of the differential expression results was made using the “DeSeq2” R package.

diff heat map. This plot (Figure 4 Main Figures) shows that there are 17 genes that are down-regulated during ER stress as compared to the untreated control condition.

A differential expression heat map of the differential expression results was made using the “DeSeq2” R package. This heatmap (Figure 5 Main Figures), like the other two plots, that there are 17 genes that are down-regulated during ER stress as compared to the untreated control condition.

A summary table was created that lists the 17 differentially expressed genes identified using the “DeSeq2” R package. Only one mitochondrial gene, rip1, was identified, and it is down-regulated during ER stress. The rip1 gene is involved in the electron transport chain.

**Main Figures:**

|  |
| --- |
|  |
| **Figure 1: PCA plot of untreated and DTT treated samples**  The x-axis is principal component one and the y-axis is principal component two. Principal component one accounts for 64% of the variance among samples. Principal component two accounts for 27% of the variance among samples. Untreated replicates are shown as light blue dots and DTT treated replicates are shown as light red dots. |

|  |
| --- |
|  |
| **Figure 2: Correlation heat map of untreated and DTT treated samples**  The multi-color legend to the right of the heat map shows which Pearson’s coefficient of correlation corresponds to which color on the heat map. The condition legend to the right of the multi-color legend indicates that light blue corresponds to a sample that was untreated and that light red corresponds to a sample that was DTT treated. The dendrogram at the top and on the left side of the heat map both show how the different samples cluster by Pearson’s coefficient of correlation. |

|  |
| --- |
|  |
| **Figure 3: MA plot of differential expression results between untreated and DTT treated samples**  The x-axis is the mean of normalized counts and the y-axis is the log base 2 of the fold change in gene expression. Red dots indicate statistically detectable genes that are differentially expressed under ER stress as compared to the untreated control. |

|  |
| --- |
|  |
| **Figure 4: Volcano plot of differential expression results between untreated and DTT treated samples** The x-axis is the log2 fold change and the y-axis is the negative log base 10 of the adjusted p-value. Light blue dots indicate statistically detectable genes that are differentially expressed under ER stress as compared to the untreated control. Light red dots indicate genes that are not differentially expressed. |

|  |
| --- |
|  |
| **Figure 5: Differential expression heat map of untreated and DTT treated samples**  The multi-color legend to the right of the heat map shows which Z-score corresponds to which color on the heat map. The condition legend to the right of the multi-color legend indicates that light blue corresponds to a sample that was untreated and that light red corresponds to a sample that was DTT treated. The dendrogram at the top of the heat map shows how the samples cluster by condition. The dendrogram on the left side of the heat map shows how different samples cluster by expression pattern as indicated by Z-score. |

**Main Tables:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **GSM** | **SRA** | **Description** |
| Untreated Replicate 1 | GSM991008 | SRR546421 | polyA+ enriched mRNA, WT -DTT Rep 1 |
| Untreated Replicate 2 | GSM991011 | SRR546424 | polyA+ enriched mRNA, WT -DTT Rep 2 |
| DTT Treated Replicate 1 | GSM991009 | SRR546422 | polyA+ enriched mRNA, WT +DTT Rep 1 |
| DTT Treated Replicate 2 | GSM991012 | SRR546425 | polyA+ enriched mRNA, WT +DTT Rep 2 |

**Table 1: RNA-seq Data Description Summary Table**

This tables shows the names, ID numbers, and treatment description for the four RNA-seq data samples used in this project. All RNA-seq data is single-end and was sequenced using an Illumina HiSeq 2000.

|  |  |
| --- | --- |
| **Untreated Replicate 1**  18880403 reads; of these:  18880403 (100.00%) were unpaired; of these:  1862859 (9.87%) aligned 0 times  16212069 (85.87%) aligned exactly 1 time  805475 (4.27%) aligned >1 times  90.13% overall alignment rate | **DTT Treated Replicate 1**  20552669 reads; of these:  20552669 (100.00%) were unpaired; of these:  1118902 (5.44%) aligned 0 times  18631040 (90.65%) aligned exactly 1 time  802727 (3.91%) aligned >1 times  94.56% overall alignment rate |
| **Untreated Replicate 2**  46193880 reads; of these:  46193880 (100.00%) were unpaired; of these:  45530441 (98.56%) aligned 0 times  608530 (1.32%) aligned exactly 1 time  54909 (0.12%) aligned >1 times  1.44% overall alignment rate | **DTT Treated Replicate 2**  45691740 reads; of these:  45691740 (100.00%) were unpaired; of these:  43860887 (95.99%) aligned 0 times  1713840 (3.75%) aligned exactly 1 time  117013 (0.26%) aligned >1 times  4.01% overall alignment rate |

**Table 2: HISAT2 Results Summary Table**

This table shows the results for each of the HISAT2 alignments performed as part of this project. This includes the total read count per sample and the percentage of aligned reads.

|  |  |
| --- | --- |
|  | **Table 3: DeSeq2 Results Summary Table**  This table lists the 17 differentially expressed genes by ID, common name, and function that were identified using DeSeq2. The rip1 gene is highlighted in red. |

**Discussion:**

The main goal of my project was to answer this question: *How does ER Stress Affect Mitochondrial Gene Expression?* However, there were quality issues with the RNA-seq data, which leave my main question largely unanswered, and casts serious doubt on the results and conclusions of Kimmig et al. 2012.

The data that Kimmig et al. 2012 used is of questionable quality. While the data had high per-base sequence quality, as determined by FastQC, the HISAT2 sequence alignment showed very poor alignment for Untreated Replicate 2 (1.44%) and DTT Treated Replicate 2 (4.01%) (Table 2). Additionally, the PCA results (Figure 1) and the correlation heatmap results (Figure 2) show that there is no clustering or correlation of the samples by treatment. Taken together, there seems to be serious quality issues with the RNA-seq data. Furthermore, Kimmig et al. 2012 used SOAP2 to align their RNA-seq reads. SOAP2 was released in 2008 and is supposedly good at aligning short reads to genome (Li et al. 2009). However, TopHat1 was released in 2009, so Kimmig et al. 2012 could have used it instead (Trapnell et al. 2009). So, why didn’t they do this? I don’t know, but it seems that they may not have fully understood what they were doing.

My differential expression results (Figure 3, 4, 5, Table 3) are inconsistent with the results of Kimmig et al. 2012. I found 17 differentially expressed genes that were down-regulated during ER stress as compared to the untreated control condition. However, this is far fewer than the dozens of down-regulated genes reported in Kimmig et al. 2012. But, this inconsistency can be explained by low quality data and differences in the analysis methods. Due to all aforementioned issues, I no longer believe the results and conclusions of Kimmig et al. 2012.

My main question, *How does ER Stress Affect Mitochondrial Gene Expression?,* remains largely unanswered. I found one gene, rip1, that is involved in electron transport to be down-regulated during ER stress. But, it is unclear what this really means, as this result could simply be a fluke do to low quality data. So, if I were to attempt to answer my main question again, I would try to find higher quality RNA-seq data on the SRA database. This RNA-seq data would need to have reads from ER stress and control conditions with at least three replicates per condition, as well as pass quality control with FastQC, and pass alignment analysis with HISAT2. If I could find data that satisfies these conditions, I would then analyze the rest of this new data in the same way that I did for this project. In general, I believe that my main question is answerable, and answering it might provide insight into many different diseases.

**Methods:**

S. pombe genome data was downloaded from Ensembl: <ftp://ftp.ensemblgenomes.org/pub/fungi/release-42/fasta/schizosaccharomyces_pombe/>

The RNA-seq data from Kimmig et al. 2012 was downloaded as SRA files and converted to fastq files using the NCBI SRA Toolkit. Detailed information about the data can be found on GEO:

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40298>

FastQC (v0.11.8) was used to assess the RNA-seq data in the fastq files (Andrews 2010). A short “summary” of the FastQC reports can be found in Supplementary\_Material/Supplementary\_Figures as “per-base-quality-of-all-4-samples.png”. The full FastQC reports can be found in the same place.

HISAT2 (v2.1.0) was used to map the RNA-seq reads to the genome (Sirén et al. 2014). The results of this alignment can be found in Table 2 (above in the Results section). The script that was used to accomplish this task is “pombe-reads-hisat2.sh”, which can be found in Supplementary\_Material/Supplementary\_Code.

Samtools (v1.9) was used to convert the SAM files (Li et al. 2009), produced as the output from the HISAT alignment, to BAM files, to be used with Stringtie. The script that was used to accomplish this task is “pombe-sam-to-bam.sh”, which can be found in Supplementary\_Material/Supplementary\_Code.

Stringtie (v1.3.5) was used to assemble and quantify the transcripts (Pertea et al. 2015). The scripts that were used to accomplish these tasks are “pombe-stringtie.sh” and “pombe-stringtie-abundances-2.sh”, which can be found in Supplementary\_Material/Supplementary\_Code.

The script prepDE was used to convert the data produced by Stringtie in to count tables that could be imported into R. The count matrix files produced from this step, “1-transcript\_count\_matrix-2” and “2-gene\_count\_matrix-2”, can be found in Supplementary\_Material/Supplementary\_Tables.

All further statistical data analysis was conducted R (v3.5.1) with Rstudio (v1.1.463).

The “tidyverse” (v1.2.1) package was used to organize the data. The “DESeq2” (v1.22.2) package (Love et al. 2015) in combination with “pheatmap” (v1.0.12) and “RColorBrewer” (v1.1-2) were used to make the PCA plot, MA plot, counts-vs-dispersion graph, volcano plot, and heat maps. The Rmarkdown file and Rnotebook HTML file produced from this analysis can be found in Supplementary\_Material/Supplementary\_Code as “pombe-project-deseq2-2.Rmd” and “pombe-project-deseq2-2.nb”. The metadata file, “3-pombe-metadata2”, used as an input for this analysis, and the result tables, “4-pombe\_DE\_results\_frame\_det” and “5-p\_DE\_frame\_log”, generated from this analysis can be found in Supplementary\_Material/Supplementary\_Tables.

The script “pombe-tdf-script.sh” was used to create TDF files for visualization, which can be found in Supplementary\_Material/Supplementary\_Code, and the TDF files can be found in Supplementary\_Material/Supplementary\_Data. IGV (v2.4.19) was used with the TDF files to make the visualization of coverage using for gas2 and rip1 and these visualizations can be found in Supplementary\_Material/Supplementary\_Figures as “coverage-data-for-gas2” and “coverage-data-for-rip1”.

**References:**

Almanza, A., Carlesso, A., Chintha, C., Creedican, S., Doultsinos, D., Leuzzi, B., Luís, A., McCarthy, N., Montibeller, L., More, S., et al. (2019). Endoplasmic reticulum stress signalling - from basic mechanisms to clinical applications. FEBS J. *286*, 241–278.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data.

Bravo, R., Vicencio, J.M., Parra, V., Troncoso, R., Munoz, J.P., Bui, M., Quiroga, C., Rodriguez, A.E., Verdejo, H.E., Ferreira, J., et al. (2011). Increased ER–mitochondrial coupling promotes mitochondrial respiration and bioenergetics during early phases of ER stress. J. Cell Sci. *124*, 2511–2511.

de Brito, O.M., and Scorrano, L. (2008). Mitofusin 2 tethers endoplasmic reticulum to mitochondria. Nature *456*, 605–610.

Cnop, M., Foufelle, F., and Velloso, L.A. (2012). Endoplasmic reticulum stress, obesity and diabetes. Trends Mol. Med. *18*, 59–68.

Hetz, C., and Saxena, S. (2017). ER stress and the unfolded protein response in neurodegeneration. Nat. Rev. Neurol. *13*, 477–491.  
  
Hotamisligil, G.S. (2010). Endoplasmic reticulum stress and the inflammatory basis of metabolic disease. Cell *140*, 900–917.

Kimmig, P., Diaz, M., Zheng, J., Williams, C.C., Lang, A., Aragón, T., Li, H., and Walter, P. (2012). The unfolded protein response in fission yeast modulates stability of select mRNAs to maintain protein homeostasis. Elife *1*, e00048.

Lebeau, J., Saunders, J.M., Moraes, V.W.R., Madhavan, A., Madrazo, N., Anthony, M.C., and Wiseman, R.L. (2018). The PERK Arm of the Unfolded Protein Response Regulates Mitochondrial Morphology during Acute Endoplasmic Reticulum Stress. Cell Rep. *22*, 2827–2836.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics *25*, 1966–1967.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Malty, R.H., Aoki, H., Kumar, A., Phanse, S., Amin, S., Zhang, Q., Minic, Z., Goebels, F., Musso, G., Wu, Z., et al. (2017). A Map of Human Mitochondrial Protein Interactions Linked to Neurodegeneration Reveals New Mechanisms of Redox Homeostasis and NF-κB Signaling. Cell Syst *5*, 564–577.e12.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. *33*, 290–295.

Sirén, J., Välimäki, N., and Mäkinen, V. (2014). HISAT2-fast and sensitive alignment against general human population. IEEE/ACM Trans. Comput. Biol. Bioinform. *11*, 375–388.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Verfaillie, T., Rubio, N., Garg, A.D., Bultynck, G., Rizzuto, R., Decuypere, J.-P., Piette, J., Linehan, C., Gupta, S., Samali, A., et al. (2012). PERK is required at the ER-mitochondrial contact sites to convey apoptosis after ROS-based ER stress. Cell Death Differ. *19*, 1880–1891.

Wang, M., and Kaufman, R.J. (2014). The impact of the endoplasmic reticulum protein-folding environment on cancer development. Nat. Rev. Cancer *14*, 581–597.

Zhang, K., and Kaufman, R.J. (2008). From endoplasmic-reticulum stress to the inflammatory response. Nature *454*, 455–462.