Thomas Murphy

Statistical Analysis of RNA-Seq Experimental Data

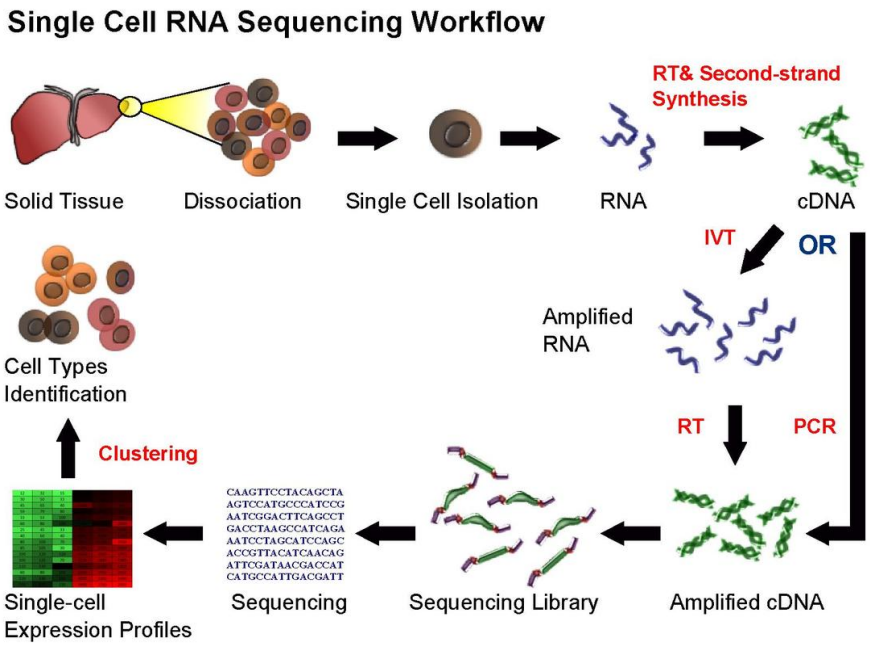
Differential Gene Expression and Aging

Biological Background

Aging is a process that all organisms undergo, which leads to the slow deterioration of many important biological functions. The process of aging is very complex, with many different factors and systems that are involved in the process. A common type of aging in humans is the slow deterioration of mental function and acuity. Parkinson's is a brain disease that leads to mental deterioration with symptoms including: tremors, bradykinesia, rigid muscles, speech changes, and loss of automatic movements[1]. Leucine-Rich Repeat Serine/Threonine-Protein Kinase 2 (LRRK2) is a huge protein with 2 domains, a kinase domain and a ROC-GTPase domain. The enzymatic phosphorylation ability along with the enzymatic GTP-GDP hydrolysis ability suggests that it plays a role in interacting with many other proteins/molecules[2]. One molecule it has shown to interact with is dopamine receptors in the brain. Many Genome-Wide association studies have been shown to link Lrrk2 mutations to the dopaminergic neurodegeneration seen in Parkinson's disease[3[.

RNA-Seq is a method in genomics/transcriptomics that allows scientists the ability to get absolute concentration measurements of transcript abundance, rather than relative concentrations from methods like microarrays. The method involves capturing the transcriptome of a sample and filtering out non-mRNA RNA molecules. The mRNA molecules can be made into a library that represents part of the transcriptome of the sample at the time of RNA extraction. These libraries can be sequenced using next generation sequencing (NGS) technologies. At a basic level, NGS is a method that fragments samples into many small pieces. These pieces are sequenced on special flow cells that can capture which nucleotides are being added during sequence extension. This results in a large collection of small sequences that represent the sequence of the sample provided.[4] So in an RNA-seq experiment, the library that represents the sample's transcriptome is sequenced into a collection of small fragments. These small fragment sequences represent the transcripts, and thus the proteins being expressed, at the time of sample collection. These can be mapped onto a reference genome related to the sample's genome, and the fragments will map to genes annotated in the reference genome (mapping just means that the sequences from the fragments align with sequences from the reference genome). So by sampling an organism, a scientist can get a quantitative measure of which genes are being expressed at any given time(see fig.1).[5]

Figure 1: RNA-Seq Experimental Workflow



https://en.m.wikipedia.org/wiki/File:RNA-Seq_workflow-5.pdf

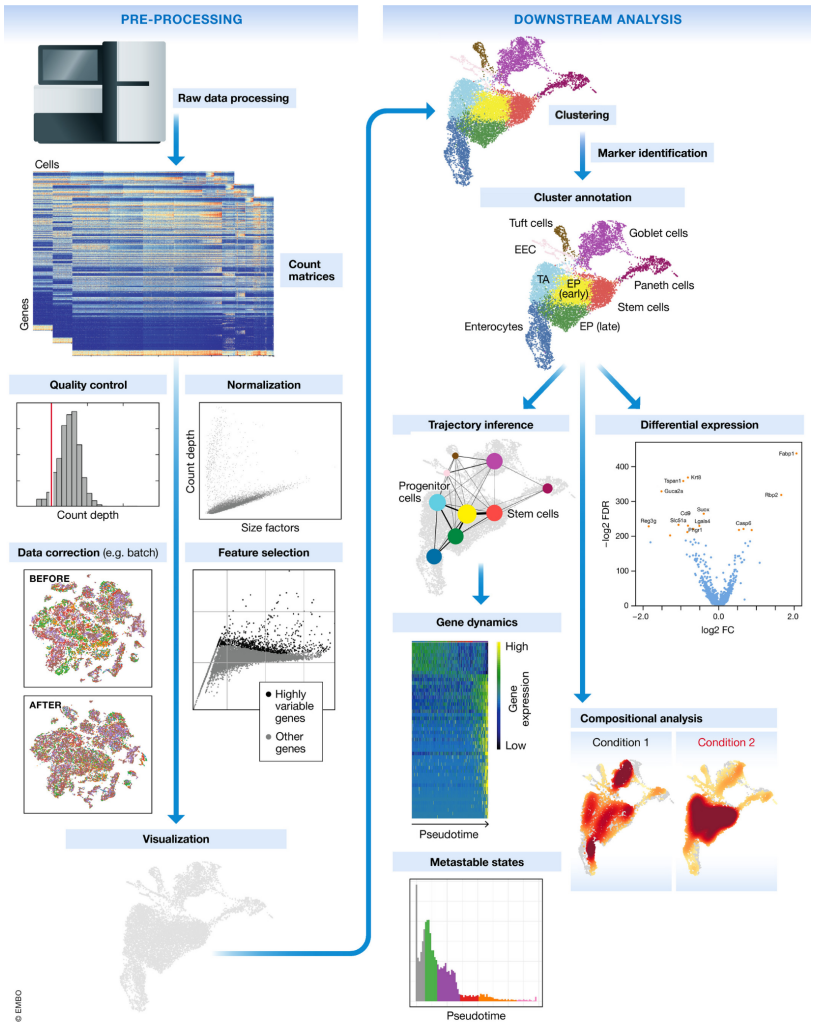
Are there certain genes related to the aging process that can be identified using RNA-Seq gene expression analysis? If an organism is expressing a Lrrk2 knockout mutation that causes dopaminergic neurodegeneration, like what is seen in Parkinson's, then that organism can be thought of as aging faster in a sense. By collecting samples from organisms either expressing or not expressing Lrrk2 over the course of their lifetimes, a scientist could capture at least part of the picture of aging. If the samples are RNA transcripts, then RNA-seq could give an idea of what genes are being expressed differently over the course of the organisms lifetimes.[3,6]

Statistical Background

The statistical analysis of RNA-seq expression data is a multi-faceted approach, involving a variety of statistical techniques. First, the amount of data being collected is massive and is highly dimensional. When working with this type of data it is often appropriate to use some type of dimension-reducing approach to make the computational analysis functionally time efficient. Principal Component Analysis(PCA) is a dimension-reducing technique that seeks to maximize variance between data points and align them along different orthogonal vectors. These different orthogonal vectors are the principal components(PCs) and capture a subset of the data within them. By choosing a certain amount of PCs that capture enough information without losing too much of the important data, scientists can reduce large data sets into just very important PCs. PCA is often used when working with RNA-Seq data, and the package planned to be used in this project involves a PCA step. Clustering analysis can also be part of the RNA-seq workflow, which can allow scientists to cluster genes by expression values. An example

of a clustering analysis method often used to help visualize gene expression values is called heat mapping. In this type of plot, genes are clustered by their normalized count values and then hierarchically arranged so that genes with similar expression values are clustered together.[7]

Another important statistical tool used in RNA-Seq analysis is normalization. RNA-seq data comes in the form of read counts for each gene. These read counts need to be normalized in order to account for not only transcript size but also sample variation and library composition. One statistical software package often used in RNA-seq analysis, called DESEQ2, normalizes the read counts by a scaling factor for each data point. A scaling factor is found by log transforming the read counts and finding their new log transformed averages. These averages will be more resistant to outliers compared to more typical library normalization methods such as RPKM (Reads Per Kilobase of transcript, per Million mapped reads), which can help with large variations due to biology. Once these log transformed averages are found, subtracting them from the log transformed values of each of the read counts can be done. What this does mathematically is calculating the ratio of reads from one sample to the average reads across all samples ($log(a[c]) - log(a[t]) = log(\frac{a[c]}{a[t]})$). Calculations of the median values of these ratios in each sample are then performed. Taking the median is another step in accounting for the variability because it gives the same weight to very large changes as it does to very low changes. The algorithm can then take these median values and transform them back to the original data (by raising them to the power of whatever log transformations were done originally). This value can then be used as a scaling factor to normalize the data, i.e. each sample's read counts are divided by the scaling factor (see fig.2[8]).



A final statistical consideration that is important to review when interpreting RNA-seq gene expression data is the hypothesis testing done. Hypothesis testing is the way statistical conclusions are found, as it involves estimating the probability that the sampled data are different from some assumed norm. Said another way, it gives values to the chances that some alternative hypothesis(H_A) is statistically different from some “null” hypothesis(H_0). For gene expression data, this is often done in order to decide if some gene *i* is expressed differently between experimental conditions, with the null hypothesis being that gene

i has the same values across the experimental conditions. The final RNA-Seq data will come out as $\log_2$ transformed datasets, which is done in order to account for the fold differences in gene expression between genes. To determine if a gene is expressed statistically differently than other genes, a null and alternative hypothesis testing is done on the $\log_2$ transformed data. Under the null hypothesis the ratio of gene $i_{treatment}$ and gene $i_{control}$ is 1, meaning that $\log_2(i_{treatment}/i_{control})$ would equal zero. If either gene $i_{treatment}$ or gene $i_{control}$ is larger/smaller, then the $\log_2(i_{treatment}/i_{control})$ would not equal zero. In fact if it was less than zero than would mean there is down-regulation of gene *i* and if it is greater than zero there is up-regulation of gene *i*. So $H_0 = 0$ and $H_A \neq 0$ would be the null and alternative hypothesis respectively.

When working with large and highly complex data, the hypothesis testing done needs to consider that many tests are done simultaneously which can lead to many false positive and false negative results if not accounted for. Controlling for these are done by considering the significance level and power needed for the desired statistical outcomes, as statistical methods cannot account for both perfectly (reducing the likelihood of one type of error leads to the increased possibility of making the other type of error). The DESEQ2 package controls for false positive errors by using a measure based on the ratio of the number of false positives to the total number of significant features, called False Discovery Rate(FDR). Since RNA-seq experiments are looking at the gene expression levels across the whole transcriptome, using FDR allows for controlling the statistical calculations at a level where more genes can be found even at the cost of more false positive errors. The rationale behind this being that a few more false positives are not detrimental to the goals of the experiment, where they might otherwise be if the experiment had high costs for false positives like in a drug testing trial.

Sampling Design

Although sample design is always an important part of any experiment, when considering RNA-seq experimental design it requires even more consideration. Sampling pool depth and replicates are important for any experiment but need special considerations for RNA-seq data. This is because variation can come not only from experimental techniques and sensitivities but also from biological variation. Gene expression values can vary widely for some transcripts between similar biological samples, i.e. the same transcript from different samples can have wide variability due to biological variation of that gene (while also having variations in expression from gene to gene). So accounting for this potential source of variation is important to consider when deciding sample size. Even more confounding, RNA-seq runs are relatively expensive and have finite sequencing depth capacities. So determining the weights given to sample size versus cost is an important consideration, especially when considering the sequencing depth needed. To handle variations due to experimental techniques and instrument sensitivities technical replicates are used, which are replicates from the same sample (organism). To handle biological variation biological replicates from different samples (organisms) are used. The trick from RNA-seq experimental design is how much to account for each one based on the amount of available funds and needed sequencing depth for questions of interest, i.e. isoform specific reads offen need more sequencing depth to capture.

Methods

The data set used in this analysis is from a series of Illumina RNA-seq runs between wildtype and treatment samples. Two populations of *Mus musculus*(breed:C57BL/6J) were grown under similar conditions over the course of 30 months. One population were mice expressing wild-type Lrrk2 and the other population of mice were transgenic mice expressing a knockout version of Lrrk2. They handle biological variation by pooling RNA extracts from randomly picked mice, then make a series of technical replicates by creating copies of the pools described above[6]. More specifically, this analysis investigates the wildtype and mutant Lrrk2 populations at the 30 month mark with just bladder cells being considered (the bladder SRA runs were the smallest datasets, the others were as large as >1000Mb).

RNA-seq data comes from Illumina sequencing outputs in the form of many small sequences, referred to as reads. These reads need to be mapped onto a reference genome and the number of reads mapped to the genome are called counts. The DESEQ2 algorithm is used in the analysis; where read counts from sample j assigned to gene i can be modeled and then sample j and j' can be compared between condition k, in order to estimate differential gene expression. DESEQ2 uses a negative binomial distribution for its model because a normal Poisson distribution does not sufficiently account for possible type I errors (false positives). This is because negative binomial distributions have parameters accounting for both mean and variance, rathan than simply the mean (as in the case with Poisson distributions). The relationship between mean(μ) and variance($\boldsymbol{\sigma}^2$) from gene *i* in sample *j*, is estimated using the data and it provides a better fit for the model to the data. Estimates for μ and $\boldsymbol{\sigma}^2$ were found using the data as if three assumptions were made: First the μ_{ij} parameters is estimated by $\mu_{ij} = q_{i,k(j)}s_j,^{-1}$ where s_j = size factor and $q_{i,k(j)}$ = condition specific gene value. The size factor accounts for sequencing depth and library composition and normalizes $q_{i,k(j)}$, which is proportional to the expected number of reads of gene *i* from condition *k*. Second, the variance parameter is estimated by $\boldsymbol{\sigma}^2_{ij} = \mu_{ij} + s^2_j v_{i,k(j)}$, where $v_{i,k(j)}$ = raw variance and μ_{ij} = shot noise. Finally, the raw variance parameter is a smooth function of the condition specific gene value parameter. This last assumption helps account for low technical replicate numbers due to high cost by grouping genes with similar expression levels into larger pools during the variance parameter estimation. So overall, the model estimates read counts by estimating variance and mean parameters under a negative binomial distribution ($RC_{ij} \sim NB(q_{i,k(j)}s_j,^{-1}, \mu_{ij} + s^2_j v_{i,k(j)})$.

To fit the model to the data the size factor was estimated. This was done by calculating the medians of the ratios of RC_{ij} vs the geometric mean across samples. The estimates for s_j and $s_{j'}$ for some gene *i* should be roughly equal if gene *i* is not differentially expressed (or if *j* and *j'* were replicates). By taking the medium of the read counts normalized by geometric means not only helps account for sequencing depth issues but also helps nullify the potential impact of influential outliers (very low or high expression values). Estimates of q_{ik} were calculated using the average read counts from samples $j_{1,2,...,m}$ for each condition *k* scaled by estimated size factor. Finally the estimates of sample variance (w_{ik}) were calculated and scaled using the estimated size factor. Once q_{ik} and w_{ik} parameters were estimated, a local regression of (q_{ik}, w_{ik}) was used to create a smooth function of $w_k(q)$ which can be used to estimate the raw variance ($v_{i,k(j)}$). This gives us our overall model $RC_{ij} \sim NB(\widehat{q_{ik}}, \widehat{v_k}[\widehat{q_{ik}}])$.

Once the model is calculated significance testing of the different gene expression values between the samples in the different conditions *k* needed to be performed. Under the null hypothesis, it was expected that the differential expression values of a gene *i* between treatment conditions and control conditions were roughly equal i.e. $q_{iA} = q_{iB}$ (meaning they were not differentially expressed). To test this DESEQ2

uses a test statistic that uses the sum total read counts for gene *i* across samples *j* in condition A (RC_{iA}) and the sum total read counts for gene *i* across samples *j* in condition B (RC_{iB}). The P-value(adjusted) of (RC_{iA}, RC_{iB}) would be the sum of all probabilities less than or equal to p(RC_{iA}, RC_{iB}). This method helps to account for a small number of technical replicates in a method that is similar to the Fisher's exact test, which tries to help account for small sample sizes.

When gene expression values were calculated and statistically differentiated gene expression values were found. They were plotted using the Volcano Plots and Heatmap R software packages. Volcano Plots are "A type of scatter plot represents differential expression of features (genes for example)" and Heatmaps are "a graphical representation of data where the individual values contained in a matrix are represented as colors"(see packages section). Finally, go enrichment analysis was performed using the R software package Enrichment Analysis Tool which is "A comprehensive system that combines gene function, ontology, pathways and statistical analysis tools to enable biologists to analyze large-scale genome-wide experimental data."(see packages section)

Results

The raw read count data from the RNA-seq Ilumina runs were analyzed using the fastQC package in Galaxy. Regions with poor quality scores were dropped from further analysis and adapter region sequences were trimmed using the Trimmomatic package in Galaxy. The resulting reads were aligned against the *Mus musculus* reference genome GRCm39 using the HISAT2 package in Galaxy(see table 1).

Table 1: Matrix of SRA runs between wildtype and mutant Lrrk2 *Mus musculus* populations' HISAT2 summary stats

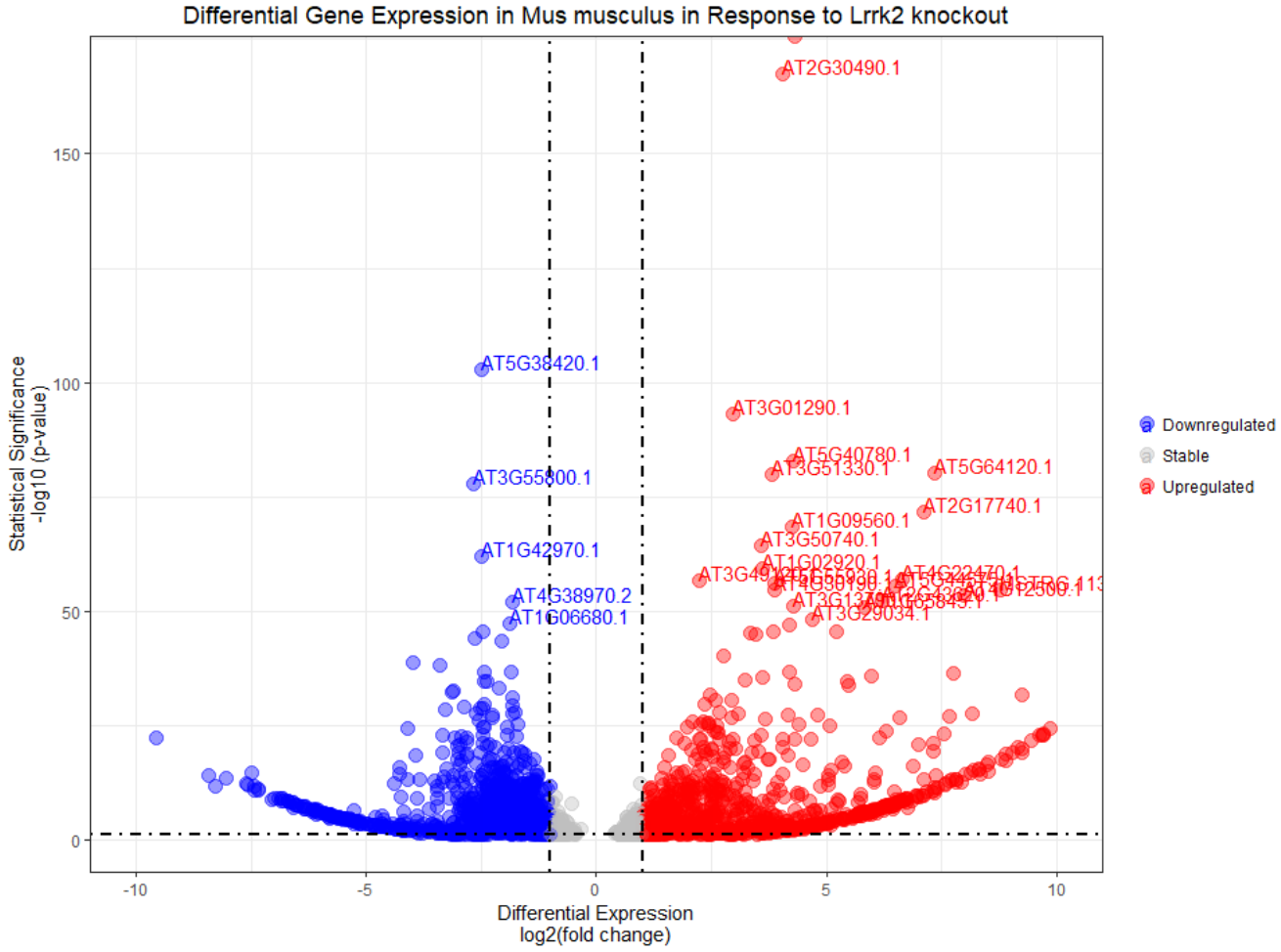
	Mut Lrrk2 SRR6519510	*Wt* Lrrk2 SRR6519511	*Mut* Lrrk2 SRR6519512	*Wt* Lrrk2 SRR6519513	*Mut* Lrrk2 SRR6519514	*Wt* Lrrk2 SRR6519515
Total reads	2756019	1736401	1668286	1626585	1021886	2485542
Aligned 0 time	172118 (6.25%)	108267 (6.24%)	117971 (7.07%)	128165 (7.88%)	26162 (2.56%)	79342 (3.19%)
Aligned 1 time	2330860 (84.57%)	1480663 (85.27%)	1139737 (68.32%)	1354407 (83.27%	767784 (75.13%)	1677104 (67.47%)
Aligned >1 times	253041 (9.18%)	147471 (8.49%)	410578 (24.61%)	144013 (8.85%)	227940 (22.31%)	729096 (29.33%)
Overall alignment rate	93.75%	93.76%	92.93%	92.12%	97.44%	96.81%

The aligned reads found using the HISAT2 package represent reads that map to gene model regions within the reference genome. However, the reads only represent a fraction of the gene mRNA transcript sequences. This is because the RNA-seq method used to collect the data was from a Tag-Seq library preparation method, which only targets the 3' regions of the mRNA transcripts. As a result, the reads

only cover a small portion of the full mRNA transcript sequence. To produce the full mRNA transcript sequences associated with the aligned reads, a Galaxy package called Stringtie was used. This package produced a list of known transcripts predicted by the aligned reads, along with potentially novel transcripts as well. These transcripts were then reintegrated into the reference genome using the Stringtie Merge package in Galaxy. The resulting genome (with new gene models) was compared to the original reference genome (GRCm39) using the GFFCompare tool in Galaxy. This analysis found 134 potential novel exons, 71 novel introns, and 93 novel loci. A GO enrichment analysis of these 134 novel exons predicted some had high subcellular localizations in mitochondria, an important organelle often associated with neurodegenerative diseases.

The novel transcripts were removed from the Stringtie merge file in order to perform the DESEQ2 analysis. This was done so that only reference transcripts were present in the file, so the DESEQ2 algorithm doesn't have to worry about some novel transcript *i* not being present in all the samples. DESEQ2 analysis revealed 2889 genes were being differentially expressed between the wildtype Lrrk2 *Mus musculus* population and the knockout mutant Lrrk2 *Mus musculus* population, at a significance level $\alpha = 0.05$. Differential gene expression was visualized using Heatmap2 and Volcano Plot packages(see Fig. 5, heatmap omitted for space). [Note: Volcano plot software in Galaxy was not working, so I plotted the volcano plot using R code, which can be found in package information] Finally, GO enrichment analysis was performed on the top 25 most differentially expressed genes between the wildtype and mutant Lrrk2 *Mus musculus* populations.

Figure 5: Volcano Plot of RNA-Seq Differential Gene Expression Results



Discussion

The RNA-Seq Illumina runs provided data that was able to be assembled into a list of 30127 mRNA transcripts that were expressed in the experimental *Mus musculus* populations. 134 of these transcripts were flagged as potentially novel forms, and were related to mitochondrial function based on the subsequent GO enrichment analysis. This is a good indication that the Lrrk2 was driving the phenotypic expression of aging (well a phenotypic proxy, i.e. neurodegeneration) based on mitochondrial dysfunction having been associated with many types of neurodegenerative diseases.[9] There was a significant amount of differential gene expression between the wild type population of *Mus musculus* which had the normal version of Lrrk2, and the mutant population of *Mus musculus* which had a recombinant knockout out version of Lrrk2. A portion of these differentially expressed genes should be related to the aging process based on the novel transcript GO enrichment analysis.

DESEQ2 differential gene expression analysis was successfully able to determine 2889 transcripts in the 30127 provided mRNA transcripts that were differentially expressed between the two populations of *Mus musculus*, given a significance level of $\alpha = 0.05$. This was done through a process of normalization, model selection, and multiple hypothesis testing described in the background and methods sections. GO term enrichment analysis of the transcripts showed that the most differential gene expression values, either upregulated or downregulated, had associations with a plethora of aging related functions. Some of these functions included pathways relating to reactive oxidative stress responses, genomic instability, and immune response suppression. These results show the power of the statistical techniques we learn about in this class.

Just from RNA-Seq count data, which is nothing but a list of short sequences, large complex biological pathways were able to be identified that responded differentially between the two populations. These responses showed a strong association with many different biological pathways that have been seen to change during the aging process. Of course, since error was accounted for using a False Discovery Rate (FDR) method, there were some genes that could have surreptitiously been associated with the aging process. However, this was done as a global view to demonstrate the power of dimension reduction, model generation/selection, multiple hypothesis testing, and other statistical techniques we learned in this class. Elucidation of the data for a given problem can be driven by understanding how each of the different statistical factors affects outcomes. The data generated from the analysis can also be further analyzed using other techniques, if more understanding of the respective functions are required.

Packages and Code Used

- HISAT2 : “A fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes as well as to a single reference genome.” - http://daehwankimlab.github.io/hisat2/
- fastQC : “Aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.” - https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Trimmomatic : “Performs a variety of useful trimming tasks for illumina paired-end and single ended data.” - http://www.usadellab.org/cms/?page=trimmomatic
- Stringtie : “A fast and highly efficient assembler of RNA-Seq alignments into potential transcripts.” - https://ccb.jhu.edu/software/stringtie/
- DESeq2 : “Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.” - https://bioconductor.org/packages/release/bioc/html/DESeq2.html
- Volcano Plots : “A type of scatter plot represents differential expression of features (genes for example)” - https://biocorecrg.github.io/CRG_RIntroduction/volcano-plots.html
- Enrichment Analysis Tool : “A comprehensive system that combines gene function, ontology, pathways and statistical analysis tools to enable biologists to analyze large-scale genome-wide experimental data.” - http://www.pantherdb.org/
- Heatmap : “a graphical representation of data where the individual values contained in a matrix are represented as colors” - https://www.r-graph-gallery.com/heatmap

(Volcano Plot Code)

```
library(dplyr)
library(ggplot2)
library(ggrepel)

data = `DifferentiallyExpressedGenes(a=0.05).StatProj_TM...Sheet1`
head(data)

data$expression = ifelse(data$V7 < 0.05 & abs(data$V3) >= 1,
                         ifelse(data$V3> 1 ,'Upregulated','Downregulated'),
                         'Stable')

plot = ggplot(data = data,
              aes(x = V3,
                  y = -log10(data$V7),
                  colour=expression,
```

```
                    label = data$V1)) +
  geom_point(alpha=0.4, size=3.5) +
  geom_text(aes(label=ifelse(data$V7<5.003100e-48,as.character(data$V1),'')),hjust=0,vjust=0) +
  scale_color_manual(values=c("blue", "grey","red"))+
  xlim(c(-10, 10)) +
  geom_vline(xintercept=c(-1,1),lty=4,col="black",lwd=0.8) +
  geom_hline(yintercept = 1.301,lty=4,col="black",lwd=0.8) +
  labs(x="Differential Expression
log2(fold change)",
       y="Statistical Significance
-log10 (p-value)",
       title="Differential Gene Expression in Mus musculus in Response to Lrrk2 knockout")  +
  theme_bw()+
  theme(plot.title = element_text(hjust = 0.5),
       legend.position="right",
       legend.title = element_blank())
Plot
```

Sources

1. Parkinson's Disease. Mayo Clinic, Patient Care & Health Information, Diseases & Conditions. https://www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055 (accessed Feb 17, 2020)
2. Li, JQ., Tan, L. & Yu, JT. The role of the LRRK2 gene in Parkinsonism. *Mol Neurodegeneration* **9**: 47 (2014). https://doi.org/10.1186/1750-1326-9-47
3. An Phu Tran Nguyen, Elpida Tsika, Kaela Kelly, et. al. Dopaminergic neurodegeneration induced by Parkinson's disease-linked G2019S LRRK2 is dependent on kinase and GTPase activity. *Proceedings of the National Academy of Sciences.* **117**:29, 17296-17307, (2020); DOI: 10.1073/pnas.1922184117
4. Behjati S, Tarpey PS. What is next generation sequencing?. *Arch Dis Child Educ Pract Ed*. **98**:6, 236-238. (2013) doi:10.1136/archdischild-2013-304340
5. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat Rev Genet* **20,** 631–656 (2019). https://doi.org/10.1038/s41576-019-0150-2
6. The Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020). https://doi.org/10.1038/s41586-020-2496-1
7. Sarah Bonnin. Heatmap.2 function from gplots package. 2020-03-09. https://biocorecrg.github.io/CRG_RIntroduction/heatmap-2-function-from-gplots-package.html (accessed Mar. 9 2022)
8. Malte D Luecken; Fabian J Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Sys. Bio.* **15**:e8746 (2019). https://doi.org/10.15252/msb.20188746
9. A. Trifunovic; N.-G. Larsson. Mitochondrial dysfunction as a cause of ageing. *Int. Med.* **263**:(2), 167-178 (2008). https://doi.org/10.1111/j.1365-2796.2007.01905.x