The University of Texas at Austin
College of Natural Sciences

Characterizing Sporadic Parkinson's Disease: Identifying Dysregulated Pathways

Bio 321G

Ericka Salas, Zo-Ann Lee, Kash Bagare, Preston Sundar, Johann Thomas

Mentor: Megan Chan

Introduction

Worldwide, 10 million people are affected by the second most common neurodegenerative disease, Parkinson's Disease(PD) (Parkinson's Foundation). Patients with PD exhibit symptoms such as: tremors, slowed movements, speech changes, rigid muscles, and impaired posture. Currently, there is no cure for PD.

There are two types of PD, familial, meaning it is inherited, and sporadic, meaning it is not inherited. This research is focused on Sporadic PD which is the vast majority of PD cases (Thomas and Beal). In both cases, PD causes the death of neurons that are responsible for producing the neurotransmitter dopamine. This neurotransmitter is vital to the normal functioning of the nervous system, in which the lack of dopamine can lead to the symptoms of PD.

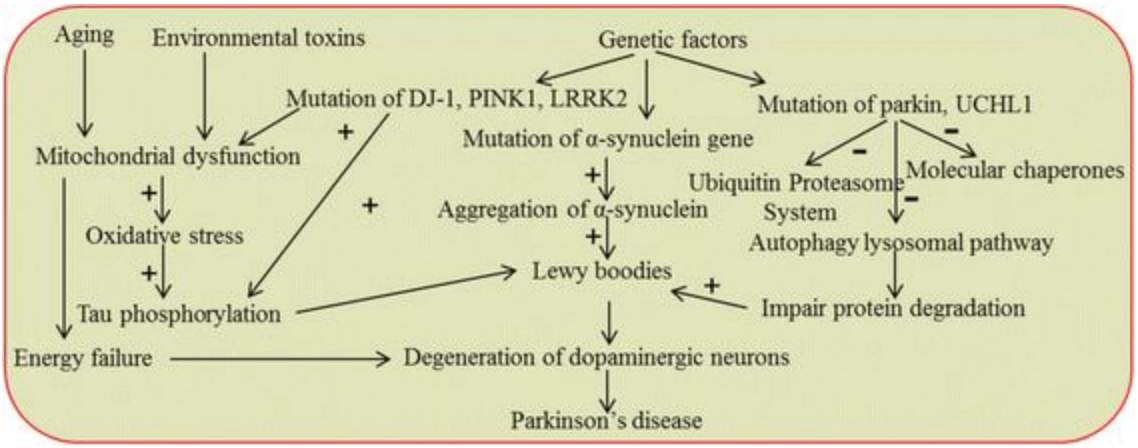


***Figure 1**: A description of the involvement of different factors in the degradation of dopaminergic neurons (Maiti, et al)*

As shown in Figure 1, there are numerous potential causes of PD. While the direct cause of PD is unknown, researchers have identified mutations and/or dysregulations in genes such as LRRK2 , PARK7, PINK1, PRKN, SNCA that can potentially be the trigger of PD. The presence of lewy-bodies and a specific protein called alpha-synuclein, has been identified as potentially having a major role in the cause of PD and is a focus of modern PD research (Gibb, W R, and A J Lees). Lewy Bodies are clumps of misfolded proteins and possibly other materials that take up space within neurons. The presence of these misfolded amyloid proteins compounds the rate of sporadic onset by causing cell death due to oxidative stress, excitotoxicity, and neuroinflammation (Maiti, et al). The build up of lewy bodies in the substantia nigra, where dopaminergic neurons reside, is the main trigger of sporadic PD and has been linked with other neurodegenerative diseases such as dementia and Alzheimer's disease (Wakabayashi, Koichi, et al). Another potential genetically linked cause is Tau protein phosphorylation. Tau proteins are fibers that attach to microtubules in cells, however, if they become misfolded then these fibers begin to aggregate in a similar fashion to the aggregation of alpha-synuclein (Kenney, et al). Aggregation of proteins and other materials within a cell can also be caused by errors in the autophagy lysosomal pathways. These pathways are typically responsible for degrading old and

used proteins. However, errors in these pathways mean that proteins are not being digested by the cell properly and these proteins can aggregate within the cell (Rivero-Ríos, Pilar, et al).

There are multiple ways of protein aggregation in neurons and these could potentially become lewy bodies. The misfolding of proteins have been shown to be caused by mutations in the genes for these proteins (Maiti, et al). Some researchers have shown protein denaturation can affect the central nervous system in a similar fashion to prion diseases (Maiti, et al). Our research is revolved around identifying genes within three different tissues that play a major role with sporadic PD. Once we identified these genes, we analyzed them and determined potential pathways and functions that are related to PD. In addition, our research further included identifying dysregulated genes and biological pathways within each of the three tissues.

Materials and Methods

The RNA-seq dataset published in NCBI GEO (Schulze et al.) contained 60 RPKM normalized samples of gene expression values. The dataset was generated from tissue samples between healthy patients and patients afflicted with sporadic PD. The samples contained three different cells which included midbrain, fibroblasts, and induced pluripotent stem cells (iPSCs). Since PD affects motor functions, neurons were significant to our research due to it being a tissue that is closely related to the disease. Fibroblasts are cells in connective tissue that can be found in bone cells, blood, or anywhere in the body and are producers of collagen. Therefore, fibroblasts are used due to them being easier to access and identify biomarkers. Induced pluripotent stem cells (iPSC) are a type of cell that can be easily generated and derived from adult cells. These are safer to use because we can re-differentiate them into the neurons that we are interested in instead of invasively taking the neurons from patients. Figure 2.1 indicates the total of 60 samples we utilized, which include 28 healthy samples and 32 PD samples. There were the most iPSC samples, 15 healthy and 16 PD, while there were only 8 healthy and 7 PD from the midbrain tissue samples and 5 healthy and 9 PD samples for fibroblasts.

The original dataset was not filtered, but was already normalized. Two filters were applied, one based on mean and the other based on variance. The first filter allowed us to take in the top 75% expressed genes for the t-test in reference to the mean, while the second filter gave us the top 10% varying genes for coexpression analysis. Subsequently, we log transformed all of our data by applying the log10 function to the entire dataset. The data before the log transformation was used in a principal components analysis (PCA), which is a reduction tool that can be used to lessen a large set of variables. The log transformed data that was mean filtered was used in finding differentially expressed genes, and the data that was variance filtered was used in finding clusters of co-expressed genes.

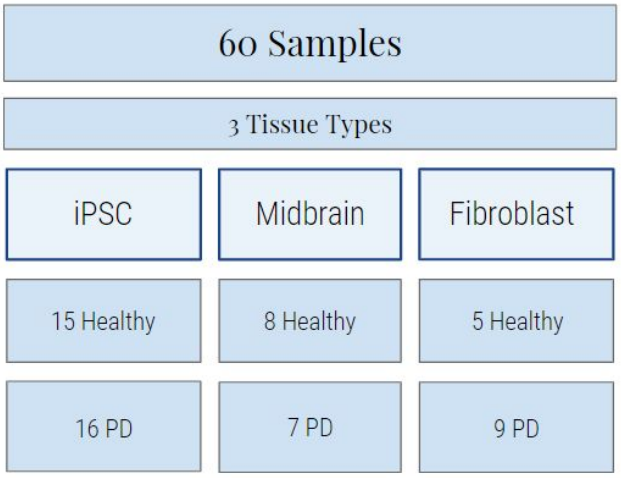


Figure 2.1*: Table of 60 samples broken down in their tissue type and its two different conditions.*

Figure 2.2*: Workflow - This includes our aims and process of performing the research from data filtering to identifying the common biological pathways.*

Differentially Expressed Genes

In order to identify the differentially expressed genes, we performed a t-test for each of the genes expressed in each tissue. There were two conditions amongst the three tissues which were sporadic PD and healthy control. A t-test was conducted comparing PD to control which allowed us to derive the p-values for each of the genes within each tissue type. This allowed us to identify the differentially expressed genes by using a p-value cut-off of .01. Traditionally, the cut-off is .05, but since we did not apply the log2 fold change method, the cut off was made more

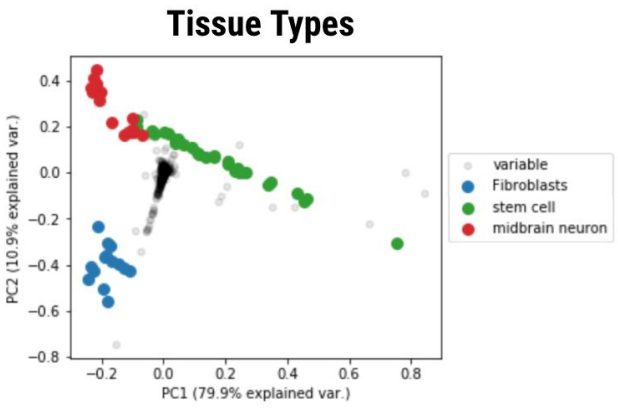
stringent in order to have a more accurate number of differentially expressed genes. After examining the differentially expressed genes for each tissue type, we subsetted the dataframe to only contain the common differentially expressed genes that related to each tissue and conducted pathway analysis using Enrichr.

Co-Expression

In order to find clusters of genes that are co-expressed, we used Python functions that allowed us to find the correlation amongst the genes. Clusters were formed using a linkage function that performs hierarchical agglomerative clustering with the metric "correlation". After this, a color key dendrogram using the linkage and the method "average" was created. When creating the dendrogram, the genes and their corresponding colors were saved. Subsequently, we were able to programmatically re-create the clusters by separating the genes based on the color from the dendrogram. Within each cluster, a t-test was performed comparing PD and healthy samples in order to visualize the level of association between the mean gene expression values of the samples and conditions (PD vs. Healthy). This was done to see if the clusters were correlated to these conditions. Afterwards, a pathway analysis was conducted using Enrichr, which is a web-based gene list analysis tool.

Results

Principal Component Analysis



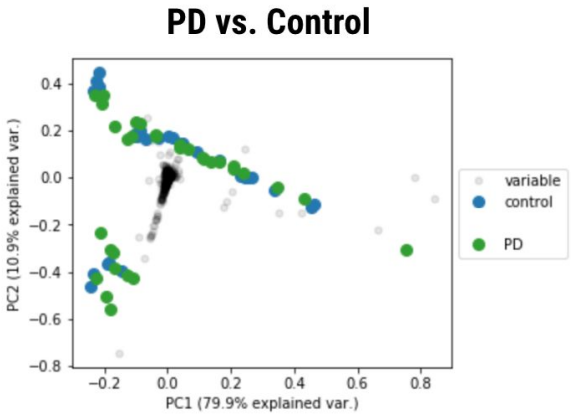


Figure 3.1: *PCA of samples (Labeled by Tissue Type)*

Figure 3.2*: PCA of samples (Labeled by Sample Type)*

Our research began with a Principal Component Analysis (PCA) which was used to determine the largest sources of variation that caused clustering amongst samples. When looking at the *Tissue Types* plot in figure 3.1, each point represents individual samples. The points are colored by tissue type (Fibroblasts, iPSCs, and Midbrain neurons). The "Tissue Types" figure, PC1, accounts for a staggering amount of the total variance, nearly 80%. On the right side of the figure, most of the iPSCs shown are clustering. On the left side, both the Fibroblasts and Midbrain neurons are also clustering. Looking at PC2 on this same figure, we can observe that PC2 can be accredited for roughly 10% of the variance. Fibroblast is clustered at the bottom of the figure, while Midbrain and most of iPSCs are clustered at the top. Given all information derived from the figures, we can conclude that samples from different tissue types cannot be

directly compared with relation to each other. The second PCA (PD vs. Control) is congruent with the *Tissue Types* PCA in the sense that it plots all given samples and is ultimately the reason they maintain the same formations. That being said, this PCA's colors classify whether the given samples were PD or were Control. When analyzing this plot, it is evident that there isn't any clear separation of PD vs Control on either axis. Due to the fact that there is so much clustering as a result of tissue type, further analysis should be done by examining each tissue type individually.

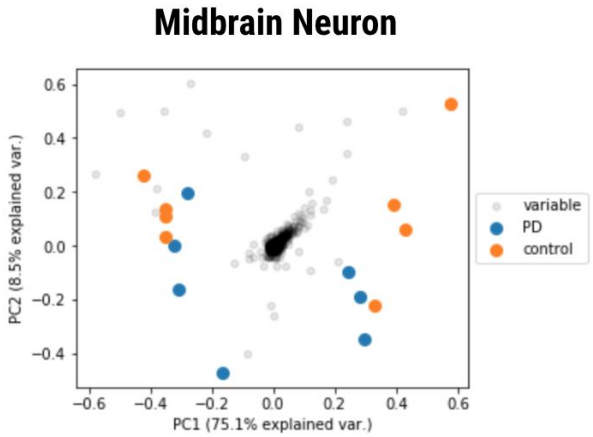


Figure 3.3: *Midbrain Neurons PCA*

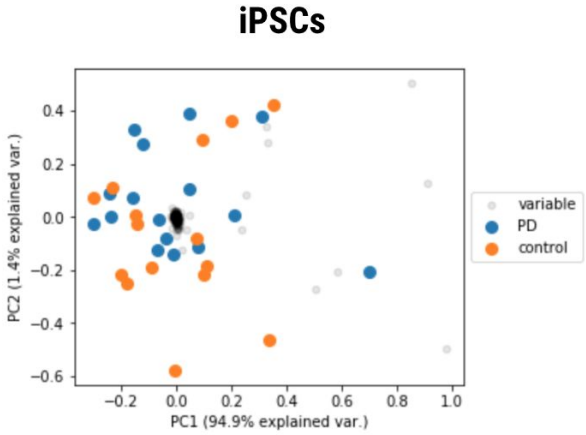


Figure 3.4*: iPSCs PCA*

In the Midbrain PCA, PD samples are typically loading lower on PC2 while Control loads higher. In the iPSC PCA, there is a similar occurrence but Control samples load lower while PD samples load higher on PC2. That being said, there is a possible outlier from both PD (far right on PC1) and control (lower on PC2) samples on the iPSCs plot which potentially alters the results.

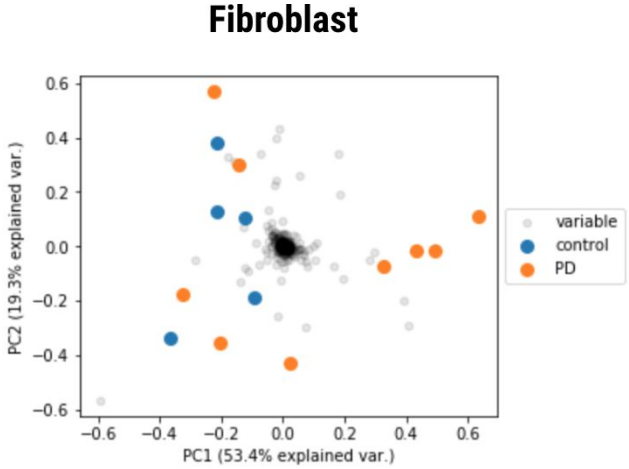


Figure 3.5: *Fibroblasts PCA*

In the Fibroblast PCA, when looking along PC1, the PD samples are mostly loading on the right side while the Control samples are loading on the left side. As a result of the PCAs of these three tissue types, we can tell there is some variation in the data separating control and PD. There isn't as much as we expected, but there is enough to compare the two as separate

conditions.

Differentially Expressed Genes (DEG)

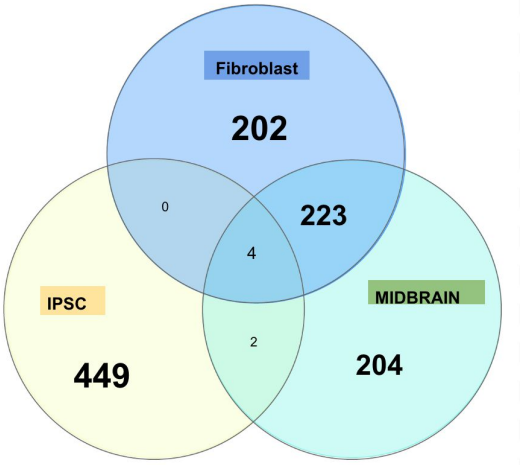


Figure 4.1*: Venn Diagram for overlapping of genes between the three tissues.*

After conducting t-tests for each gene in order to gather the p-values, we identified the differentially expressed genes for each sample using a p-value cutoff of 0.01. The figure above indicates that each tissue has a few hundred differentially expressed genes. Stem cells (iPSC) resulted in a total of 455 DEGS, midbrain 433 DEGs, and fibroblasts 429 DEGs. A great amount of genes overlapped between the fibroblasts and midbrain tissues. Stem cells differed the most amongst the three, however, our PCA foreshadowed its difference. Four genes overlapped between the three tissues which was further analyzed.

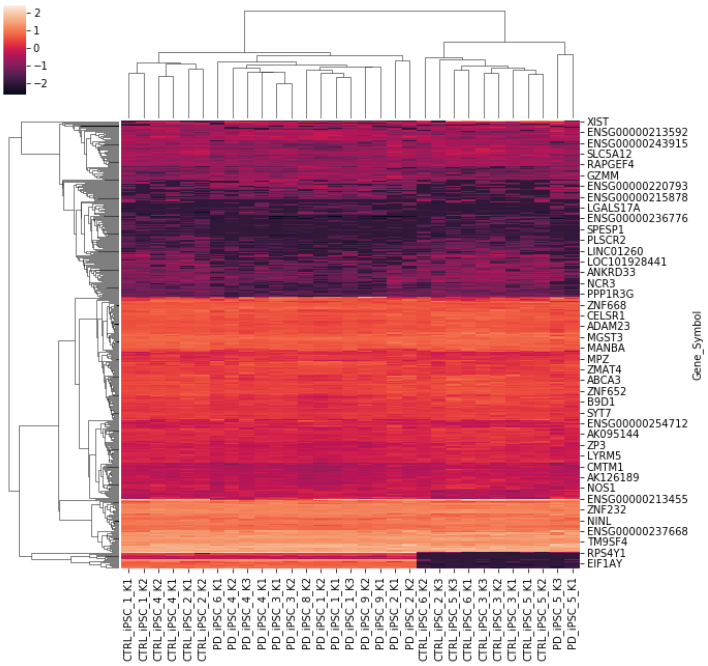


Figure 4.2*: Heatmap for IPSC Differentially Expressed Genes*

The differentially expressed genes in the induced pluripotent stem cells are shown in the figure above through a heat map

After using the cutoff p-value of 0.01, the leftover genes in the iPSC tissue are shown in figure 4.2. It is a bit difficult to notice the different clusters of Sporadic PD and control due to them being immensely clustered together, which was previously noticed in the PCA figures 3.1 and 3.2. Certain genes such as RPS4Y1 and EIF1AY (towards bottom of heatmap) are minute identifiers that make the cluster more noticeable due to the dark purple color. Furthermore, these two genes are highly expressed in some PD and control and lowly expressed in these two conditions as well.

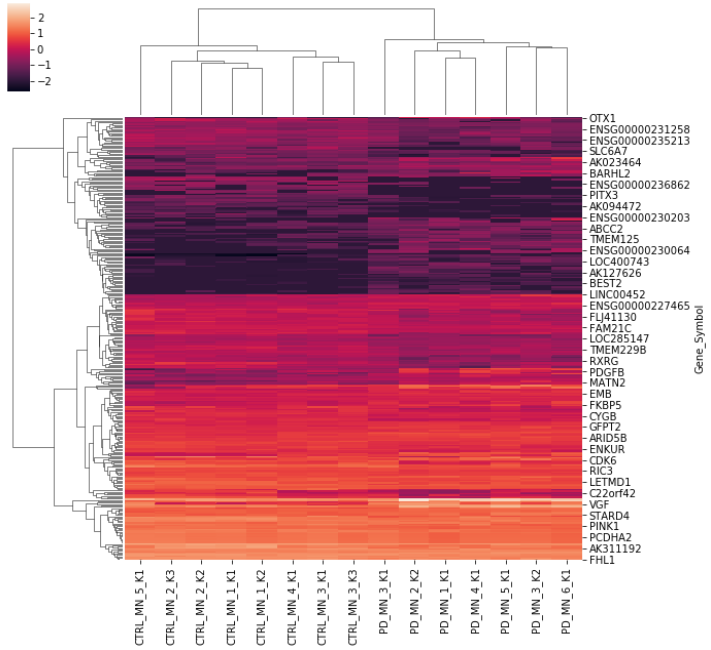


Figure 4.3*: Heatmap for Midbrain Differentially Expressed Genes*

The figure above is similar to the figure 4.2, however these are different genes that establish possible biomarkers for PD and midbrain tissues. A few genes such as OTX1 and PINK1 that have appeared in previous studies are shown in this specific heatmap.

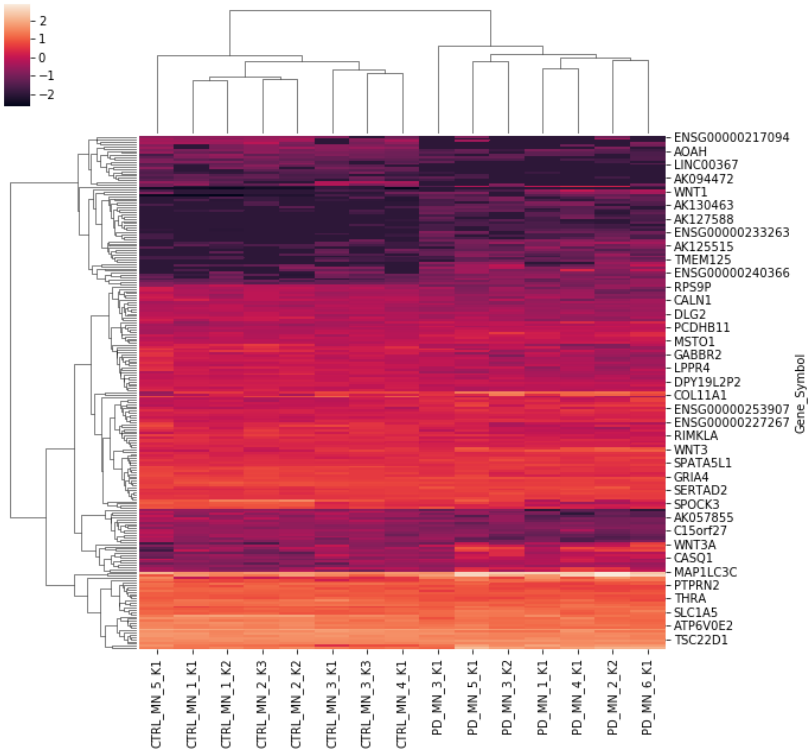


Figure 4.4: *Heat map for Fibroblasts Differentially Expressed Genes*

Finally, figure 4.4 shows the differentially expressed genes between the two main conditions: Sporadic PD and Control. The cluster between these two conditions is more noticeable than the previous.

Co-Expression

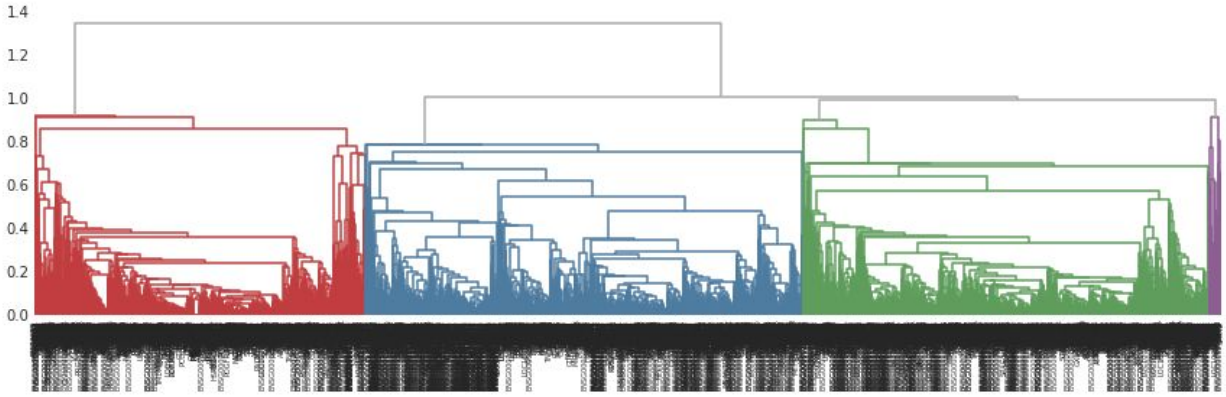


Figure 5.1: *Clustering using a color key dendrogram*

Cluster #	# of genes	Correlation Coeff	P-Value	Pathways
Cluster 1	1157	0.1328	0.3117	Focal Adhesion WP306, miRNA targets in ECM and membrane receptors WP2911, Oncostatin M Signaling Pathway WP2374
Cluster 2	1420	-0.049	0.7100	mRNA Processing WP411, Retinoblastoma Gene in Cancer WP2446, Cell Cycle WP179
Cluster 3	1536	-0.1286	0.3274	Cytoskeletal regulation by Rho GTPase_Homo sapiens_P00016, Cadherin signaling pathway_Homo sapiens_P00012, Endogenous cannabinoid signaling_Homo sapiens_P05730
Cluster 4	41	-0.0519	0.6933	Statin Pathway WP430, PPAR signaling pathway WP3942, and Alzheimers Disease WP2059

Figure 5.2: *Table with each cluster and their size, correlation coefficient and p-value from the t-test, and the top 3 pathways from pathway analysis*

After clustering by examining co-expressed genes (fig 5.1), there were three large clusters (1536, 1420, 1157 genes) and one smaller cluster (41 genes). Unfortunately, the p-values from the t-test performed on each cluster using the conditions PD vs. control were all significantly higher than 0.05 (fig 5.2), which did not give evidence against the null hypothesis (that there is no relationship between the conditions and the expression value for each cluster). These high p-values (0.3117, 0.71, 0.3724, and 0.6933), supported by the corresponding correlation coefficients, may have been caused by using the mean of the gene expression values for each sample when performing the t-test comparing the PD and healthy samples. Nevertheless, pathway analysis was still performed using Enrichr where multiple pathways were found to be involved with each cluster of genes.

Discussion

After gathering data between our two aims, we analyzed the dysregulated pathways identified from our results. First of all, the differentially expressed genes indicated a certain pathway that correlates to Parkinson's disease (Figure *4.1*). According to this venn diagram (Figure 4.1), it is significant to remark the high amount of genes expressed between midbrain and fibroblast is immense compared to any other pair of tissues, which had previously been hinted at by the different PCAs. The iPSC clustering in the *Tissue Type* PCA is very spread out while the fibroblasts and midbrain are closer to each other. This PCA foreshadowed that there is minimal overlap of differentially expressed genes between the iPSCs and other tissues, which indicates that their genes are distinct. This is accommodated by the fact that approximately 90% of the variation in the data can be explain by difference in iPSCs and the two other tissue types. Therefore, we could hypothesize this result by simply looking at the PCA for the tissue types.

The Venn diagram in Figure 4.1 also allows us to visualize the number of differentially expressed genes that overlap between the three different tissues. We narrowed down to four genes which include NLRC3, CASQ1, TPM3P9, GLL3P in which are all involved in the calcium regulatory pathway in the cardiac cells, which was discovered by inputting the genes in Enrichr. The calcium regulatory pathway we found is based in the heart which is completely

different from the tissues we are dealing with. However, previous research from the University of Cambridge has shown that an excess amount of calcium can lead to toxic clusters that can cause sporadic Parkinson's disease. Even though this pathway originates within the cardiac region, the circulatory system is responsible for transporting these clusters to the brain. According to the Michael J Fox research foundation, dysfunction of the blood brain barrier has lead to cause parkinson's disease, which allows toxic clusters such as calcium to be easily transported in the brain. Figure 5.3 below shows the calcium regulatory pathway with the identified genes. The main genes we must point out amongst these four is the NLRC3 gene and CASQ1 gene. The CASQ1 genes, also known as Calsequestrin gene, is a calcium-binding protein that is responsible for trigger muscle contraction. This explains why people with PD experience these kinds of symptoms of stiff muscles and tremor due to the high amounts of calcium that stimulate muscle contraction. In addition, the NLRC3 gene is known as the NOD-like intracellular protein that plays a role in the immune system. This indicates the reason why many patients with PD have low immune response and are more susceptible to experience exacerbations with other diseases.

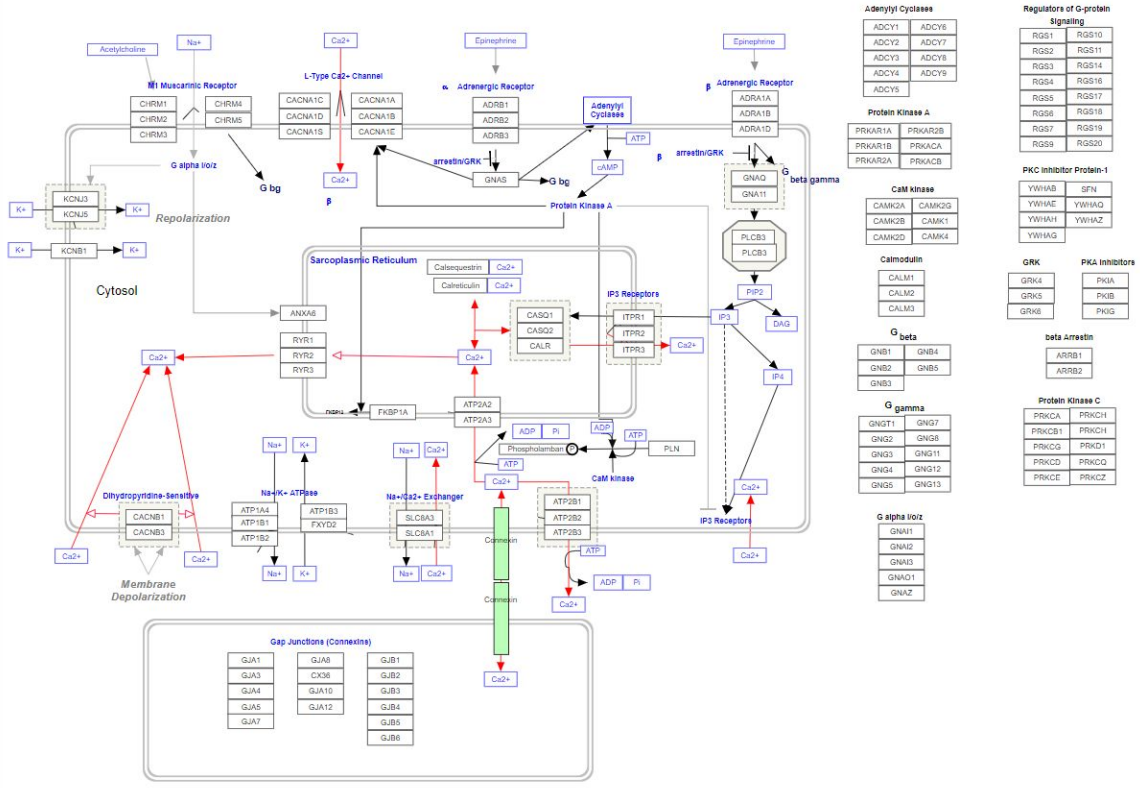


Figure 5.3: *Calcium Regulatory Pathway*

From the co-expressed gene analysis, multiple pathways for each cluster were found (fig 5.1). The top three pathways that were found from each co-expressed gene cluster were included in the table, and some of them were mentioned in the study we got our data from (highlighted in yellow). Pathways such as (miRNA targets in ECM and membrane receptors WP2911, mRNA Processing WP411, and Alzheimer's Disease WP2059) were common between both our results. Finding these pathways not only helps better our understanding of which body functions are

involved with PD, it also is important as we were able to replicate some of the results from the original study, which provides more support for the idea that these pathways are involved with PD. In a recent study that examined the effect of miRNAs on PD, they found that while miRNAs normally play a large role in the development of certain populations of neurons in the CNS, distinct miRNA are involved with the progression of PD. MiRNAs also *repress target gene functions both by regulating target mRNA levels as well as repressing the target mRNA translation* (Schulze et al.). This correlates to the mRNA pathway found because any dysregulation in the miRNA pathway may have lead to repressed translation involving mRNA. It is interesting that we were able to find that the genes that were co-expressed similarly are a part of a network that is involved with PD. Another pathway we found that had been identified before was Alzheimer's Disease. The study that we got our data from also mentioned how their research on the pathways involved in PD supports the hypothesis that pathways that are altered in PD also have been reported for Alzheimer's disease. There may be overlapping pathways between these diseases as both may involve similar buildup in the brain as mentioned previously. Additionally, a common molecule that seems to be deregulated between the two diseases is piRNA, which forms RNA-protein complexes by interacting with piwi proteins. In another study that used the same dataset we used, researchers found piRNAs to be dysregulated in PD and mentioned that dysregulation of piRNA also occurred in Alzheimer's disease samples (Schulze et al.). Unfortunately, we were unable to find differentially expressed genes that were involved with piRNA across all tissue types when examining co-expressed and differentially expressed genes.

Limitations

One limitation is that when we performed a PCA for the iPSCs, we noticed some outliers that could have altered the results. Given more time, those outliers would have been removed and the PCAs would have been rerun. Another limitation is that when we performed the t-test on the clusters of co-expressed genes using the conditions PD vs. Control, instead of using all the gene expression values for each gene for each sample, we just found the mean gene expression level for each sample. This means that some variation could have been compressed and differences could have been lost. Additionally, it could have been beneficial to perform a t-test using different conditions (age, sex, etc.) to see if there might be another cause for the co-expression, but we were not able to find enough metadata for the samples to do so. Having this additional metadata would have also been beneficial because in the PCAs for each of the tissues there was some variable that was causing the most variation for PC1. Having this metadata could potentially help us find the variable that was causing the variation as statistical analyses could be performed using that condition. Another limitation was that when we found the differentially expressed genes, we did not subset using Log2FC and p-value cutoffs, only p-value. Fold change tells us how one is being more expressed than another. This is a limitation because this means that the genes we found to be differentially expressed could be significant but not that different.

Conclusion

In this research, we were able to identify multiple pathways involved with PD. Some of the results from the original study or other previous studies overlapped with our own results, which provided more support for the idea that these pathways are involved with PD. The calcium regulatory pathway provided reinforcement to ideas that could investigate the causes of PD. The

pathway investigated from the co-expressed gene analysis allowed us to examine dysregulation and associate it with other diseases such as Alzheimer's. Finding these pathways helps to better our understanding of which body functions are involved with PD. One of the most key findings was that differential gene expression was highly overlapped between fibroblasts and midbrain neurons. This is because it is highly invasive to extract cells from a person's midbrain, however it should be much easier to extract samples from fibroblasts. This could allow scientists and doctors to check for markers of PD without extracting cells from a person's brain. We might not know the cure for Parkinson's, but every step of research is a a step closer to making the lives of the affected 10 million people in desire for improvement.

Works Cited

"Calcium May Play a Role in the Development of Parkinson's Disease." *ScienceDaily*, ScienceDaily, 19 Feb. 2018, www.sciencedaily.com/releases/2018/02/180219071758.htm.

"Dysfunction of the Blood-Brain Barrier: The Cause of Parkinson's Disease." *The Michael J. Fox Foundation for Parkinson's Research | Parkinson's Disease*, www.michaeljfox.org/foundation/grant-detail.php?grant_id=188.

Gibb, W R, and A J Lees. "The Relevance of the Lewy Body to the Pathogenesis of Idiopathic Parkinson's Disease." *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 51, no. 6, 1988, pp. 745–752., doi:10.1136/jnnp.51.6.745.

Kenney, Daniel L., and Eduardo E. Benarroch. "The Autophagy-Lysosomal Pathway." *Neurology*, vol. 85, no. 7, 2015, pp. 634–645., doi:10.1212/wnl.0000000000001860.

Maiti, Panchanan, et al. "Current Understanding of the Molecular Mechanisms in Parkinson's Disease: Targets for Potential Treatments." *Translational Neurodegeneration*, vol. 6, no. 1, 2017, doi:10.1186/s40035-017-0099-z.

Rivero-Ríos, Pilar, et al. "Targeting the Autophagy/Lysosomal Degradation Pathway in Parkinson's Disease." *Current Neuropharmacology*, Bentham Science Publishers, Apr. 2016, www.ncbi.nlm.nih.gov/pubmed/26517050.

Schulze, Markus, et al. "Sporadic Parkinson's Disease Derived Neuronal Cells Show Disease-Specific MRNA and Small RNA Signatures with Abundant Deregulation of PiRNAs." *Acta Neuropathologica Communications*, BioMed Central, 10 July 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6038190/#CR12.

"Statistics." *Parkinson's Foundation*, 28 Mar. 2019, parkinson.org/Understanding-Parkinsons/Statistics.

Thomas, Bobby, and M Flint Beal. "Parkinson's Disease." *Human Molecular Genetics*, U.S. National Library of Medicine, 15 Oct. 2007, www.ncbi.nlm.nih.gov/pubmed/17911161.

Wakabayashi, Koichi, et al. “The Lewy Body in Parkinson's Disease and Related Neurodegenerative Disorders.” *Molecular Neurobiology*, U.S. National Library of Medicine, Apr. 2013, www.ncbi.nlm.nih.gov/pubmed/22622968.