Genetic Overlap between Alzheimer's, Parkinson’s, and healthy patients

DSC180B, Section B04: Genetics

Justin Lu, Saroop Samra, Xuanyu Wu

Background

This study aims to find gene expression similarities and differences between patients of Alzheimer's (AD) and Parkinson's (PD). The study uses sequencing data from microRNA (miRNA) found in two of the body's biofluids: cerebrospinal fluid (CSF) and blood serum (SER). Although these disorders are experienced by many people, little is known about what specifically causes the two diseases and how to prevent or cure them. The difficulty in finding these solutions arises from the complexity of the "pathomechanisms" underlying the diseases, as well as their tendencies to have early stages that are asymptomatic, making detection very challenging until symptoms set in [1].

With our analysis, we aim to study the genetic causes underlying the two diseases by inspecting the sequencing data found in patients' bodies in the form of miRNA sequences. Uncovering the genetics behind the diseases can help researchers better their understanding of the development of AD and PD in humans and improve their chances of finding efficient preventative measures for the two disorders. Furthermore, the similarities that we may find between the disorders can aid in the research of neural disorders in general, and contribute to early diagnosis, prevention, and cures.

Psychiatric Disorders

Alzheimer's

Alzheimer's disease is a progressive brain disorder that heavily impacts brain function in that it slowly deteriorates memory and thinking skills, leading to symptoms like forgetting recent events or conversations, to eventually losing the ability to carry out simple tasks or even recognize friends and family members. Difficulty in reasoning and thinking is also a common symptom, especially with abstract concepts like numbers, again making everyday tasks like paying bills challenging. AD also impairs people's abilities to make decisions in everyday situations, like driving. The principal risk factor of Alzheimer's is age; people become more likely to develop AD as they grow older. Family history can also increase a person's chances of developing the disease. A genetic link to increased chance of AD shows in the form of a variation of the gene APOE e4, something we will definitely explore further in our study. As of now, there is no known cure for AD, but medications have been known to improve or slow the effect of the disease, and programs and caregivers help to support AD patients [2].

Parkinson's

Parkinson's disease is a progress nervous system disorder that impacts movement. There are a myriad of symptoms that constitute PD, including a tremor (trembling) in a limb like a hand or fingers, slowed movement, rigid muscles, and impaired posture and balance. PD is also sometimes accompanies by other complications,

like cognitive issues (trouble thinking, dementia), emotional changes (like depression), swallowing, chewing, and eating problems, among others. The principal risk factor of PD is age; PD usually develops in people age 60 or older. Other risk factors include family history and sex (men are more likely to develop it than women). As of now, there are no known cures or even prevention methods for Parkinson's. However, there are medications that can ease the symptoms, as well as surgeries that can regulate parts of the brain to improve symptoms [3].

miRNA

Our sequence data comes from the encodings of microRNA strands. microRNA (miRNA) are a class of non-coding RNAs that regulate gene expression. Specifically, they bind to specific mRNA and prevent those target mRNA from translating the necessary directions to produce certain proteins. Because of the behavior of miRNA, it will be worthwhile to explore which miRNA are binding to which mRNA, and subsequently what proteins are being down-regulated (lowly expressed). These proteins could then be contributing factors to the symptoms and/or development of the two diseases [4].

CSF/SER

The miRNA in our study was sourced from two locations (specifically, fluids found in our body): cerebrospinal fluid (CSF) and blood serum (SER). These two fluids are part of the central nervous system, which are highly impacted by both PD and AD. CSF cushions the brain and is a "shock absorber" for the central nervous system, and also removes waste products from the brain [5]. miRNAs can be found in the CSF and have been found to be instrumental in responding to malignant tumors in the nervous system [6]. Blood serum (or serum) is the fluid that blood cells move through, but without the plasma - it is the clear liquid that remains after blood clots [7]. miRNA is found in serum as "secreted miRNAs", meaning miRNA that has been excreted from cells or tissues [8].

Pipeline

For this project, the basic overview of our pipeline is that we want to access our data, preform necessary quality checks (qc), then merge the inputs into a gene_matrix and feature table, normalize by outputting normalized counts, return the LRT (Likelihood Ratio Test measures how well the model fits [9]) plots in the analysis step, and then finally visualize.

The overall pipeline is shown in Figure 1 below.

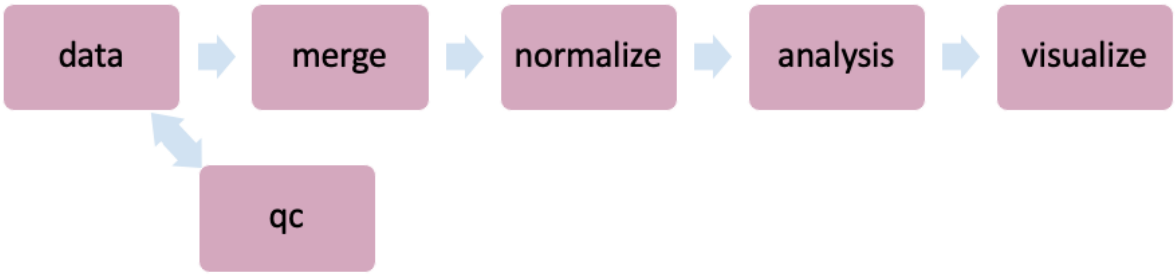


Figure 1: Overview of Targets in Pipeline

Given the restricted access to the raw dataset, we contacted Professor Kendall Jensen, author of the original paper, who showed us the exRNA Atlas, a data repository of the Extracellular RNA Communication Consortium (https://exrna-atlas.org/ (https://exrna-atlas.org/)) which includes small RNA sequencing and qPCR-derived exRNA profiles from human and mouse biofluids. The study's processed dataset was included in this repository including the gene count matrix files for each sample.

The features table is generated from three different sources:

- **SRA_RunTable.csv**: NCBI's SRA Run table that held the age and SRA Run attributes
- **Table_S1.csv**: The attributes table from the research paper that held many other attributes such as expired age, PMI, plaque density, Braak Score.
- **exRNA_Atlas_CORE_Results.csv**: Atlas core table that held information about the condition and biofluid

These tables were all merged in the data step using Pandas merge feature, shown in Figure 2 below.

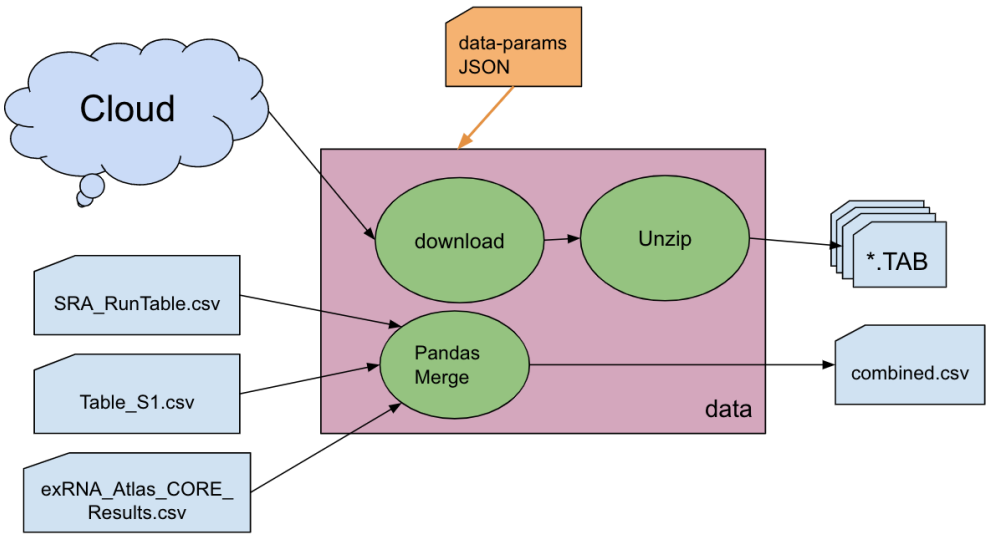


Figure 2: Data Target Implementation

The "qc" target as show in Figure 3 is an optional step that can be used on the source fastq and bam files to generate quality reports using FastQC and Picard. This step was not leveraged as the dataset did not have fastq files but did already have the FastQC quality reports that we were able to analyze.

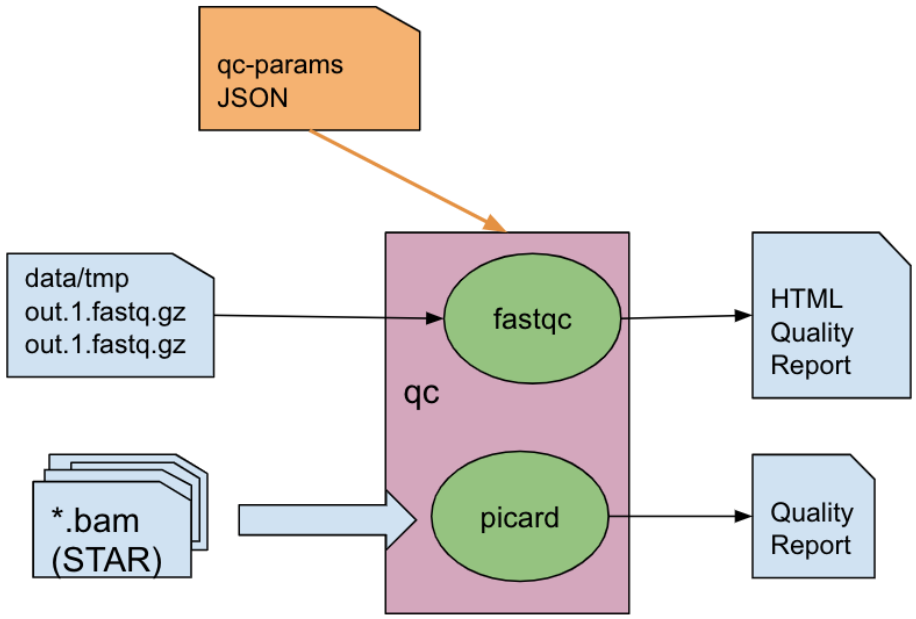


Figure 3: Quality Control Target Implementation

The next step is to execute the "merge" step: which takes all the gene count SRA files and merges them in one gene matrix table using Pandas [10]. The output will also be a gene experiments table which identifies the sample labels and the patient features (age, disorder etc) shown in Figure 4 below.

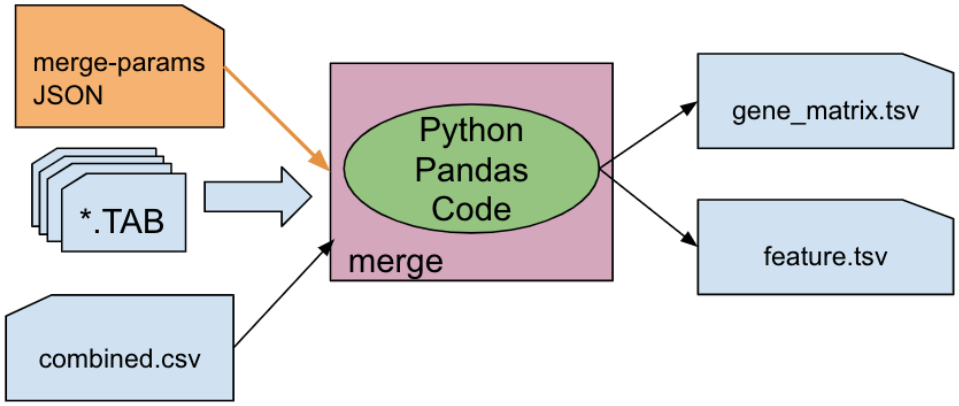


Figure 4: Merge Target Implementation

After merging, we execute the "normalize" step: this imports the merged gene count matrix into a custom R script which uses the DESeq2 [11] module to generate two normalized matrix counts: one uses standard normalization and the other Variable Stabilization Transformation shown in Figure 5 below.

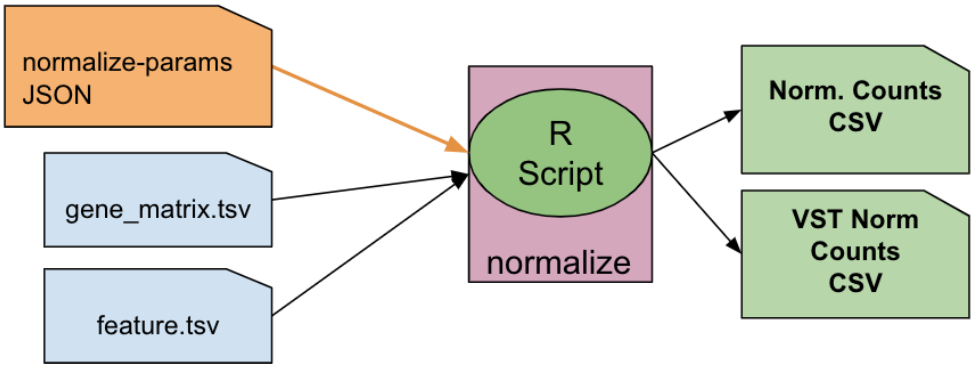


Figure 5: Normalize Target Implementation

After normalizing we execute the "analysis" step: this generates 4 Likelihood Ratio test (LRT), a hypothesis test that compares models in terms of how they fit the available data by comparing the likelihood scores of the models [9]. Each of the 4 LRT's corresponds to one biofluid and one disorder. The LRT will be compared against that disorder versus the control group. Additionally, a MA Plot and Heatmap are generated. This is shown in Figure 6 below.

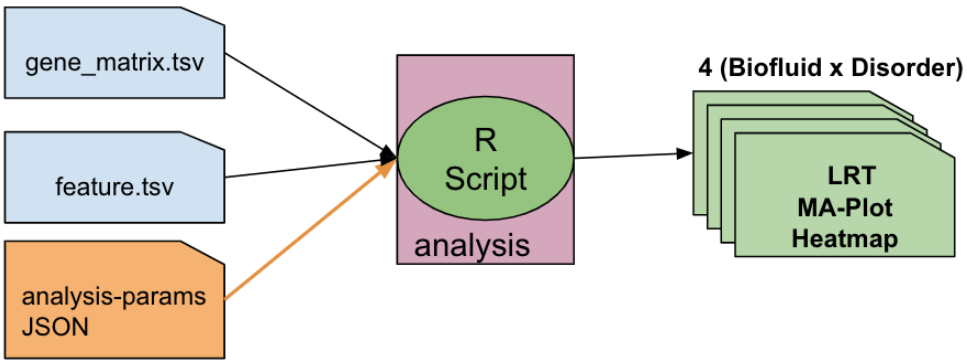


Figure 6: Analysis Target Implementation

The final step in this pipeline process is the "visualize" step. This generates the many charts and visualizations we have implemented to show the results of our findings. This step takes in the VST Norm Counts as well as the LRT MA-plot and Heatmap that are automatically generated and outputs all of our visualizations, shown in Figure 7 below.

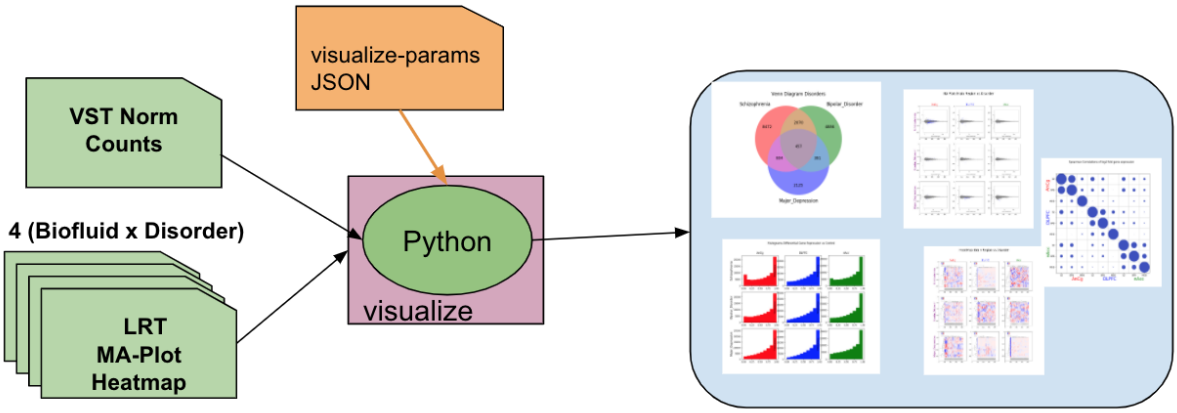


Figure 7: Visualize Target Implementation

Finally, Figure 8 shows the entire pipeline from start to finish. The condition column was either Parkinson's, Alzheimer's or a healthy patient. The biofluid column holds the source of where the miRNA samples were taken from, either serum or the cerebrospinal fluid.

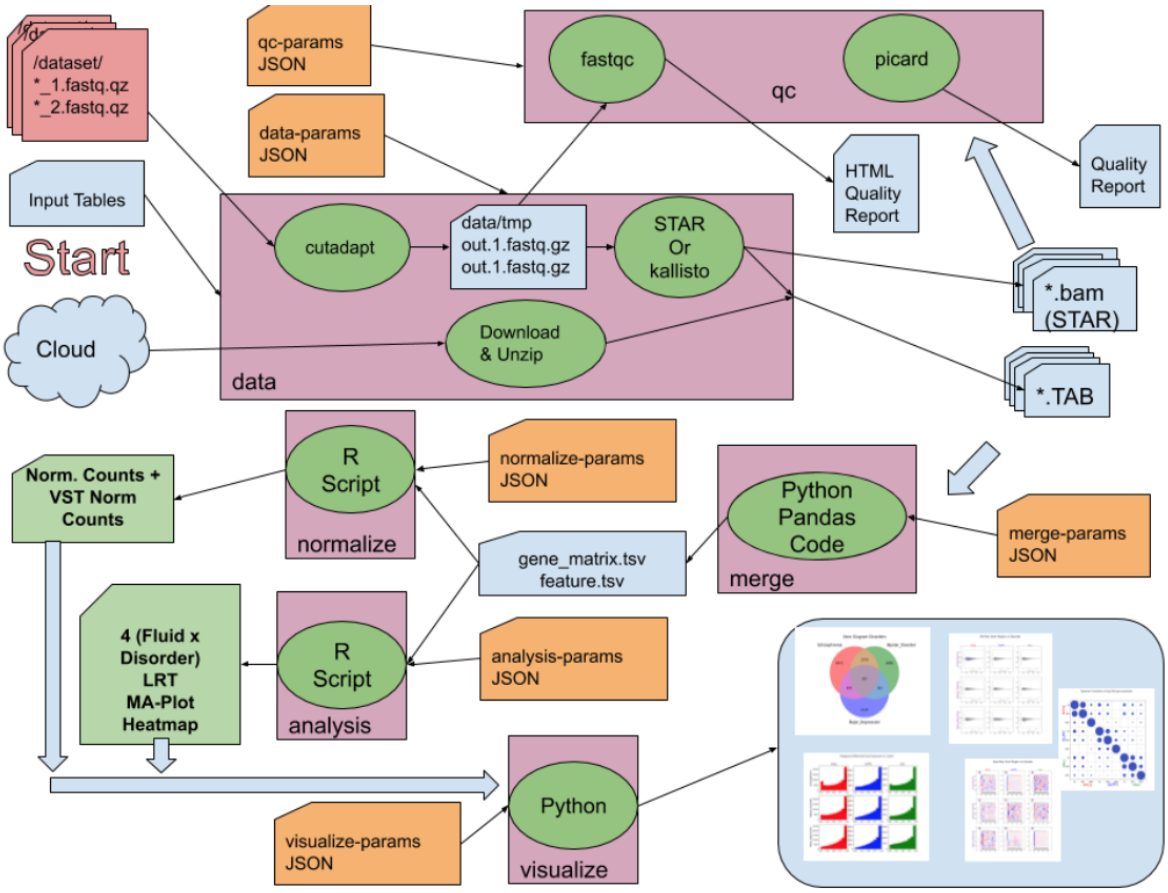


Figure 8: Pipeline Details

Quality Checks

Cutadapt

Cutadapt [12] is a tool that is used by geneticists to perform data cleaning on sequence data. When sequence libraries are prepared, the process adds adapter sequences called "primers" to the actual miRNA sequences. However, those types of sequences are not relevant to our analysis, and can actually negatively affect our data quality and our subsequent results. Therefore, the cutadapt tool removes those sequences, as well as any other low-quality reads so that data is ready for analysis. However, since the data provided to us was not the "raw" data, we could not evaluate the quality of the data itself, nor could we perform any quality control like performing cutadapt. The researchers who provided the data to us most likely already performed quality checks and used a tool like cutadapt in data cleaning.

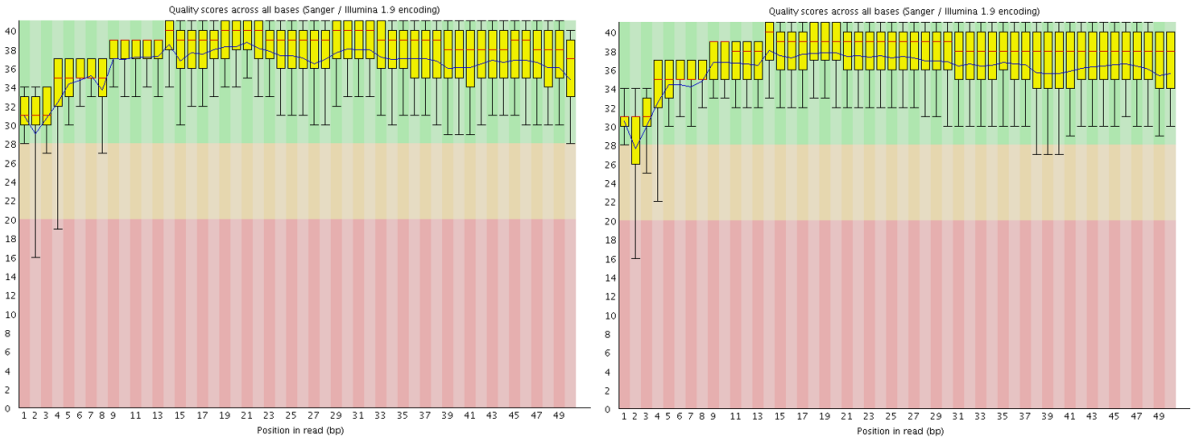
FastQC

FastQC [13] is a tool that is used in checking the quality of raw sequencing data before performing large-scale analysis. The software tool calculates and outputs quality metrics of each of the sequence reads, which allows us to determine whether to keep a sequence read, to “cut away” the extraneous parts of a sequence (using cutadapt), or to leave out the read all together. The metrics (shown as graphs and tables) include “Per sequence quality scores” which indicates the average quality of reads over the sequences of an SRA run, “Sequence length distribution”, the distribution of sequence lengths, and an important factor, “Overrepresented sequences”, which are sequences that are not found to be in the human genome, among other measurements. Specifically for overrepresented sequences, FastQC marks these as “overrepresented” because it cannot find the source of the sequences; however, more often than not, they are adapters that have been “tacked on” during the library preparation of the sequences, and are then caught by FastQC. This is where we would use the cutadapt tool to cut out those unnecessary sections, or just completely leave the reads out of the analysis. In the end, by combining all these factors, our decision to keep or leave out sequence reads follows the ERCC (External RNA Controls Consortium) Quality Control Standards. These standards for our data (which is specifically an exRNA-seq dataset) were drawn up at a Washington, DC conference in November 2015 in order to have a universal set of quality check guidelines [14]:

1. An individual RNA-Seq dataset is required to have a minimum of 100,000 reads that overlap (sense or antisense) with any annotated RNA transcript in the host genome. The annotation includes all small RNAs, such as miRNAs (from miRBase), piRNAs, tRNAs, snoRNAs, and circular RNAs, as well as long transcripts from GENCODE, which includes both protein coding genes and long non-coding RNAs (lncRNAs).
2. The fraction of reads that align to the host genome (after filtering out contaminants, adaptor dimers and ribosomal reads) that also align to any annotated RNA transcript (described in point #1) should be greater than 0.5.

Below, we compare some of the FastQC outputs for reads that were marked as "Pass" versus those marked as "Fail".

FastQC outputs of failed vs passed healthy serum samples



Healthy sample from serum (failed)

Healthy sample from serum (pass)

Figure 9: Comparison of Failed/Passed Base Quality Metrics for Healthy Serum Samples

As we can see in Figure 9, the quality of each base between the two sequence reads does not seem to be very different. It is likely that the "failed" sample on the left did not pass the ERCC quality control check for other reasons, as shown above and outlined by the ERCC. Since we do not have access to the raw sequencing data, there is no way to know for sure how they determined the ERCC pass/fail label.

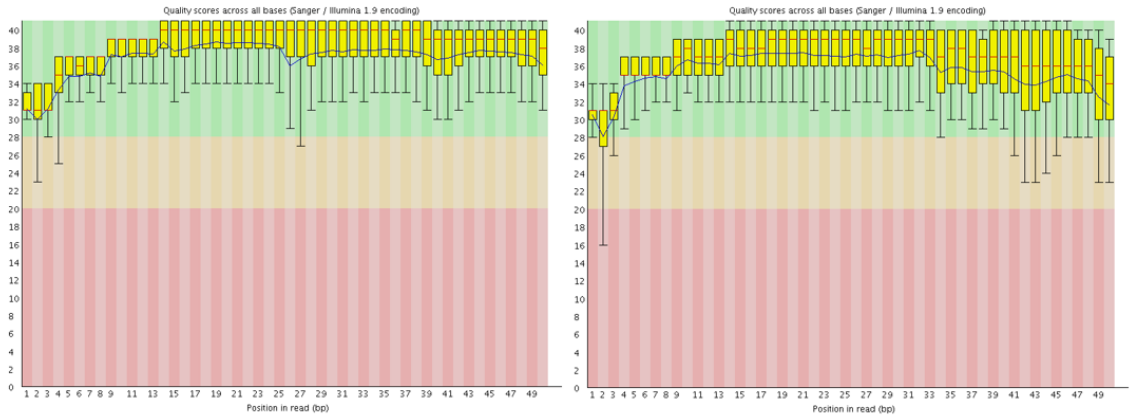
Overrepresented sequences

Sequence	Count	Percentage	Possible Source
CTGACCTATGAATTGACAGCC	13548	13.277926965521297	No Hit
AAGCTGCCAGTTGAAGAACTGT	5852	5.7353431209204775	No Hit
TCAGTGCACTACAGAACTTTGT	3625	3.552737322853167	No Hit
TGGAGTGTGACAATGGTGTTT	3251	3.186192837681557	No Hit
CGGGTTCGACTCCCGGTGTGGGAAC	2816	2.759864358939177	No Hit
TCACCGGGTGTAAATCAGCTTG	2170	2.1267420663700336	No Hit
ACTGGACTTGGAGTCAGAAGGC	1747	1.7121743732481332	No Hit
GGCTGGTCCGATGGTAGTGGGTTATCAGAAC	1660	1.6269086774996568	No Hit
GGCTGGTCCGATGGTAGTGGGTTATCAGAACT	1643	1.610247564537311	No Hit
GGGTTCGACTCCCGGTGTGGGAAC	1552	1.521061606915342	No Hit
TGGAGTGTGACAATGGTGTTTGT	1543	1.5122410176999823	No Hit
CTAATGGATAAGGCACTG	1486	1.456377286002705	No Hit
CTGACCTATGAATTGACAGCCT	1448	1.41913479820452	No Hit
ACTGGACTTGGAGTCAGAAGGCT	1297	1.2711449124801537	No Hit
TGTAAACATCCTCGACTGGA	1119	1.0966932591097085	No Hit
AGCAGAGTGGCGCAGCGGAAGCGTGCTGGGC	1044	1.023188348981712	No Hit
TGGAGTGTGACAATGGTGTTTG	1025	1.0045671050826195	No Hit
CTGACCTATGAATTGACAGC	931	0.9124409510555305	No Hit
CTGACCTCTGAATTGACAGCC	912	0.8938197071564381	No Hit
TCATTGCACTACAGAACTTTGT	770	0.7546504106474313	No Hit
GGGGTTCGATTCCCCGACGGGGAGC	769	0.753670345179058	No Hit
ACTGGACTTGGAGTCAGAAGG	751	0.7360291667483387	No Hit
GTTCGATTCCCCGACGGGGAGC	740	0.7252484465962326	No Hit

Figure 10: Example of 'Overrepresented sequences' list

Shown in Figure 10 is an example of a table/list of overrepresented sequences that FastQC detects in a sequence read. FastQC labels these as either possible primer sequences (by matching with primers that certain tools, like Illumina, use), or just "No Hit", meaning an unknown source. Either way, it is a starting point for cleaning any sequences that failed the FastQC check, by removing these sequences with tools like cutadapt. Otherwise, we could just leave that sequence read out of our analysis.

FastQC outputs of failed vs passed Parkinson's serum samples

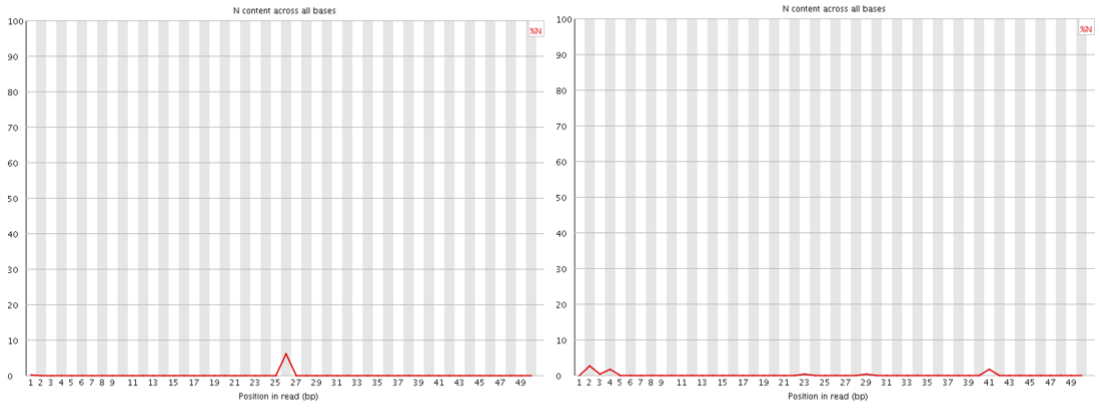


Parkinson's sample from serum (failed)

Parkinson's sample from serum (pass)

Figure 11: Comparison of Failed/Passed Base Quality Metrics for Parkinson's Serum Samples

Figure 11 shows the base quality over the sequences of the two samples is not vastly different. Yet, the sample on the left did not pass the ERCC quality check. So, looking further into the report, we note below that the sample on the left did have an abnormal amount of N bases, which, in DNA sequencing terms, usually means "unspecified", or a base that could be attributed to any of the 4 (A, T, G, C) main bases [15]. This uncertainty is not meaningful, so the sequence itself was removed. There could also have been other factors to this removal, particularly in regard to the ERCC standards.



Parkinson's sample from serum (failed)

Parkinson's sample from serum (pass)

Figure 12: Comparison of Failed/Passed N Content

Finally, we removed 19 samples from the analysis that were marked as failing the ERCC QC variable as shown in Table 1 below.

MEETS ERCC QC STANDARDS?	
PASS	324
FAIL	19

Table 1: Pass and Fail SRA Samples

EDA

Experiments Features Table

The `Run` column in the features table represents a unique "ID" of the sample collection for a patient in our study. Each patient (healthy, Parkinson's, or Alzheimer's) has a sequencing sample run associated with it, and serves to uniquely identify the sample in our study. The `CONDITION` column corresponds to the disease that each patient was afflicted by (Alzheimer's or Parkinson's), or if the patient was part of the control group (healthy). The `BIOFLUID` column designates the source of the sample from the patient's body (CSF or serum). `sex` and `expired_age` are the gender of the subject, and the age of the subject at death, respectively. The `PMI` column stands for "post-mortem interval", which means the amount of time between the subject's death and when the sample was collected from the body of the subject. `PlaqueTotal` , `Plaque density` , and `TangleTotal` all correspond to the amounts of structures called plaques and tangles in the brain. Plaques are dense clumps in the space between nerve cells in the brain, and are known to negatively impact the brain cells around it. When they develop around brain areas like the hippocampus (which is a part of the brain that is fundamental in the process of making memories), it leads to the dementia symptoms of Alzheimer's. Tangles are also structures that develop in the brain that negatively affect the transportation of neurons to and from certain areas, altogether inhibiting brain function [16]. `Braak score` , also referred to as "braak stage" is a score that is used to measure the degree of brain dysfunction for both Alzheimer's and Parkinson's patients [17]. `LB Stage` corresponds to the stage of Lewy Body dementia, which is often related to Parkinson's disease. Lewy bodies are clumps of proteins that develop in areas of the brain responsible for memory and movement - both are impacted by Parkinson's [18]. `NIA-R` is the modified NIA-Reagan diagnosis of Alzheimer's disease is based on consensus recommendations for postmortem diagnosis of Alzheimer's disease. The criteria rely on both neurofibrillary tangles (Braak) and neuritic plaques (CERAD) [19]. Finally, `sn_depigmentation` is short for substantia nigra depigmentation. The substantia nigra is a part of the midbrain; this brain region is usually found to be depigmented in Parkinson's disease patients [20].

By inspecting these features, specifically performing EDA and doing research on what each of the values mean, we can delve further into the analysis by determining what features will be important in the differential gene analysis model in DESeq2, and which of them will significantly differentiate between Alzheimer's, Parkinson's and healthy patients in terms of their genetics.

Run	SRR1568567	SRR1568730	SRR1568666	SRR1568510	SRR1568518
CONDITION	Healthy Control	Parkinson's Disease	Alzheimer's Disease	Parkinson's Disease	Parkinson's Disease
BIOFLUID	Cerebrospinal fluid	Serum	Serum	Cerebrospinal fluid	Cerebrospinal fluid
sex	male	female	female	male	female
expired_age	94	79	81	79	82
PMI	2.5	6	2.5	2.5	4.16
PlaqueTotal	15	2.75	11.5	7.5	0
Plaque density	frequent	sparse	frequent	moderate	zero
TangleTotal	12	3.25	11.1	3	6.5
Braak score	IV	II	V	II	III
LB Stage	No Lewy bodies	Limbic type	No Lewy bodies	Neocortical type	Limbic type

Table 2: Important Feature attributes for first 5 Run Samples

General Patient Population

Uniqueness of the Data

After checking the uniqueness of the data there is only one sample from a single area of interest for each subject. So we need to keep in mind that there may exist cross-subject differences for samples of different biofluids.

Gender & Condition Breakdown of Each Biofluid

We broke down the samples into two groups based on biofluid, and then looked at the distribution of other variables, namely gender, expired age, disease duration, PMI, total number of plaques, and total number of tangles.

This dataset is comprised of 126 healthy control subjects, 110 subjects of Alzheimer's Diseases , and 107 Parkinson's Diseases patients.

Then, we broke the population by gender and biofluids shown in Figure 13 below. Although the gender distribution is mostly balanced, the dataset contains significantly more female samples in the Parkinson's disease group. Therefore, gender bias could potentially affect the result of our analysis.

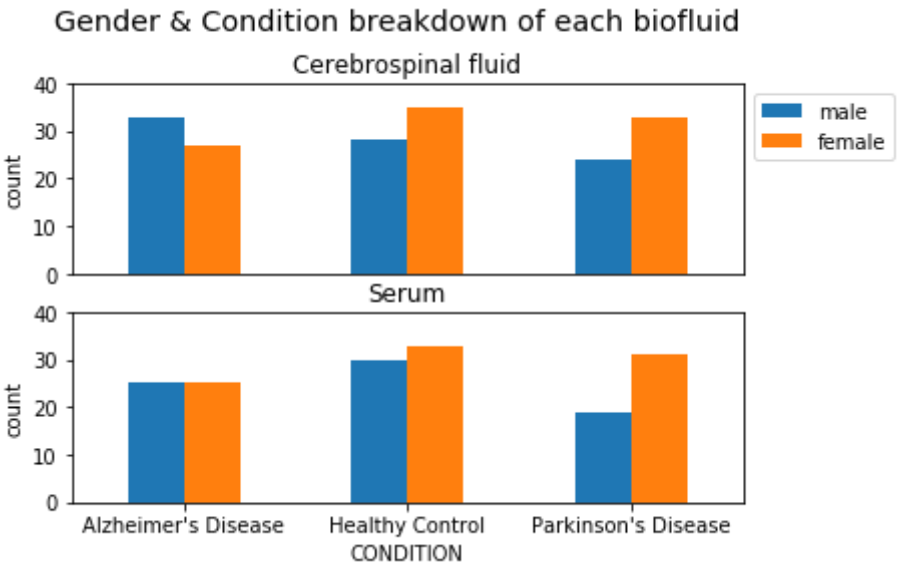


Figure 13: Samples breakdown by gender and biofluid.

Variable distributions Broken Down by Group in CSF and Serum

As shown in Figure 14 below, when comparing across conditions, more samples in the healthy control group have a larger expired age in both CSF and Serum sample populations. The Alzheimer's group have shorter PMI but higher total numbers of plaque and tangle in both CSF and Serum samples. When comparing across biofluids sample populations, the Alzheimer's group has a higher expired age in the CSF population than in the Serum population. The Parkinson's group in the CSF population has slightly larger disease durations.

Although the distributions have slight differences, the distinction between the distribution of the samples of cerebrospinal fluid and that of serum is not too drastic to a degree where we need to handle anything especially in our downstream analysis.

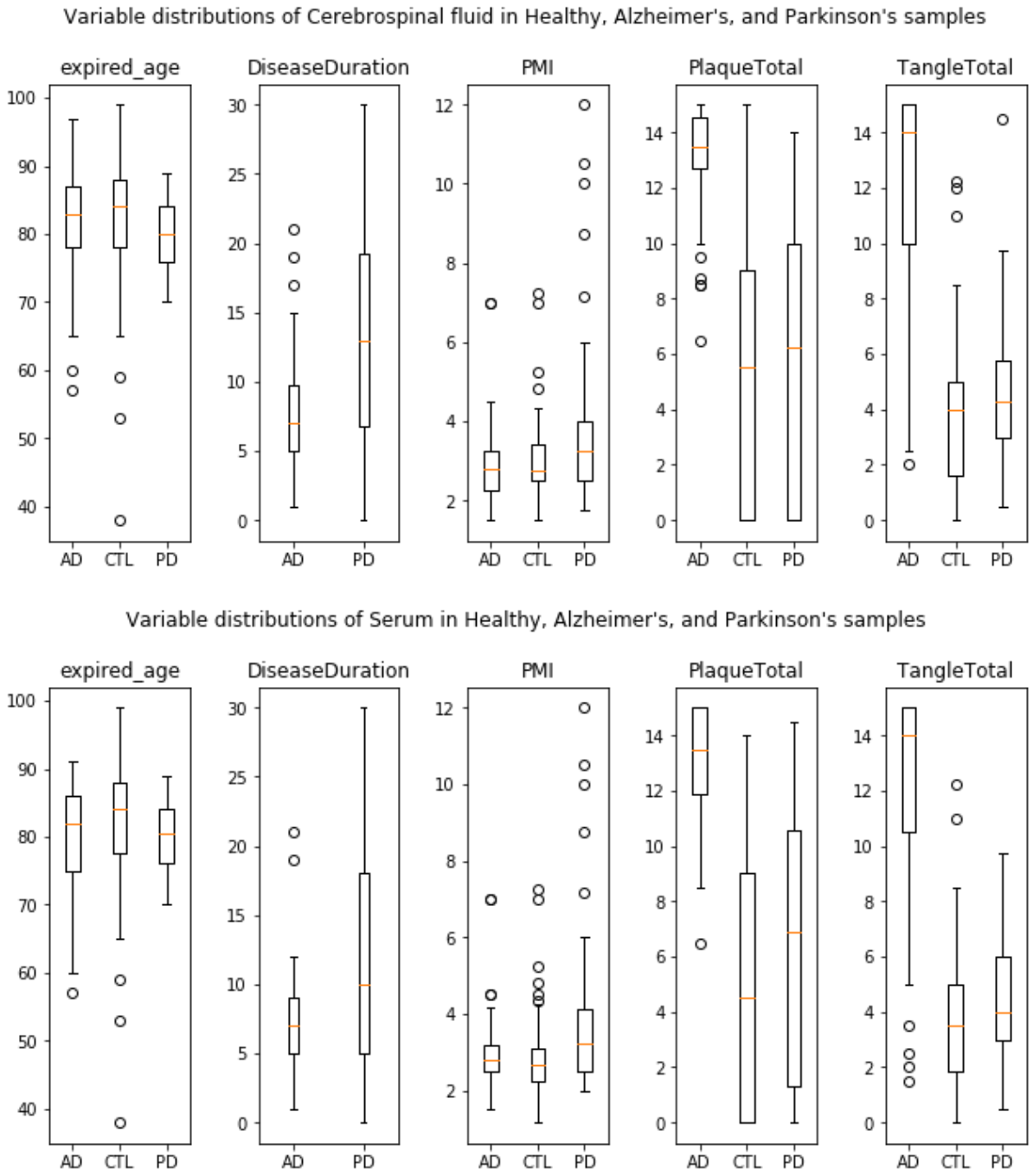


Figure 14: Distributions of the above variables.

"Disorder Markers"

As mentioned previously, NIA-R is a measurement for diagnosing Alzheimer's disease. We wanted to validate if the values of the samples in our dataset reflect this claim.

In Figure 15 below, in panel A, we saw that all the healthy control samples are either under the category "criteria not met" or "no AD". Surprisingly, although it is expected that most of the AD samples are under the category "high" there are some of the PD samples under "intermediate" and "low" categories. It might suggest that there exist some commonalities between AD and PD.

Since Lewy Bodies are closely associated with Parkinson's disease, it is reasonable to see there are only Parkinson's samples in the LB stage categories in panel B below. Among those samples, most of them are under limbic type and neocortical type. Under "No lewy bodies" category, there are no Parkinson's patients, which further proves that lewy bodies are specific to Parkinson's disease in our dataset.

According to Poewe et al., compared to control, Parkinson's disease is defined by sn depigmentation [21]. So it is intuitive to see, in panel C, the Parkinson's group has the most severe cases. However, same as what happened in NIA-R distribution, there is not a clear separation between AD and PD, namely, there are still some AD patients in the "mild", "moderate", or even "severe" categories.

Braak score is used to classify the degree of pathology in both PD and AD. However, in our dataset, although we can see a clearly different distribution for AD patients (the count increases as the stages go higher), there is no clear separation between the PD patients and the healthy control (panel D).

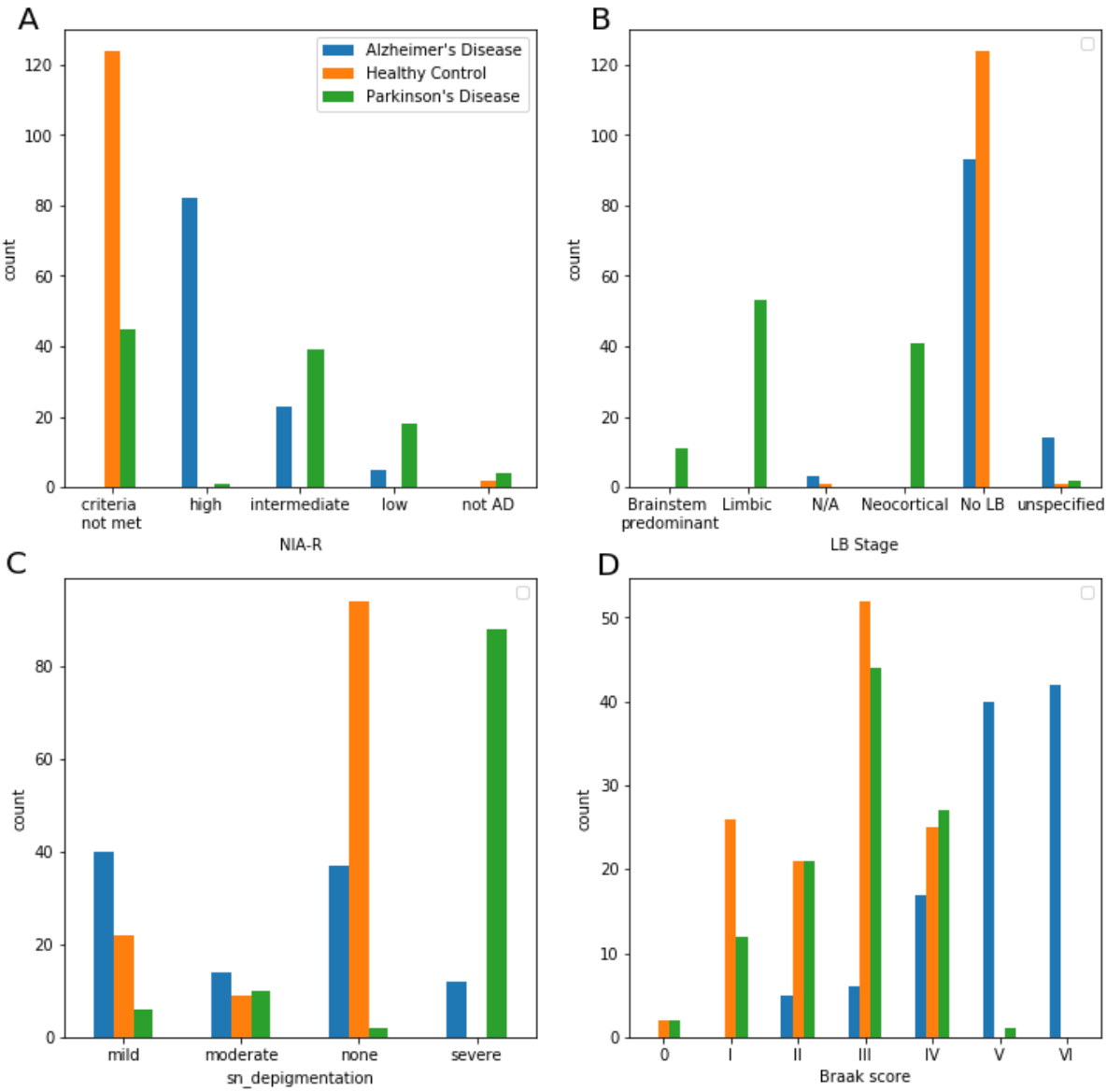


Figure 15: "Disorder Markers" Distribution of Healthy, Alzheimer's, and Parkinson's samples.

Biofluid Region

During our EDA of the three biofluid regions specified in the replication project, we narrowed our focus to determine how similar the two biofluid regions of the study were when it came to the basic variables explored above (namely `expired_age`, `DiseaseDuration`, `PMI`, `PlaqueTotal`, and `TangleTotal`).

Mean Breakdown of Disorders in Each Biofluid

In Table 3 below, we can determine some important information about patients that suffer from the diseases versus healthy patient samples. Clearly, the plaque and tangle counts are much higher in Alzheimer's patients, which is expected because these structures are found primarily in the brains of people with Alzheimer's. Another important point to note is that the disease duration of Parkinson's patients is significantly higher than the other people, due to the fact that Parkinson's is a slowly progressive disorder and develops gradually (more gradually than Alzheimer's). Therefore, the average disease duration of Parkinson's is much higher than the other patients [22].

CONDITION	BIOFLUID	expired_age	DiseaseDuration	PMI	PlaqueTotal	TangleTotal
Alzheimer's Disease	Cerebrospinal fluid	81.633	7.569	2.942	13.104	12.078
	Serum	80.260	7.146	2.999	13.102	12.034
Healthy Control	Cerebrospinal fluid	81.984	3.286	3.019	5.321	3.910
	Serum	81.714	3.167	2.878	4.943	3.720
Parkinson's Disease	Cerebrospinal fluid	79.895	13.173	3.731	5.768	4.535
	Serum	80.140	11.851	3.876	6.185	4.375

Table 3: Numeric feature attributes Grouped By Condition and Biofluid

Correlation Observation in Serum and Cerebrospinal Biofluids

Here, we inspect the correlation between features in our feature table within both serum and cerebrospinal samples in Figure 16 below.

For serum samples, the most positively correlated variables seem to be `PlaqueTotal` and `TangleTotal` with a high correlation of 0.71; this follows because the presence (or absence) of both the plaque and tangle structures in the brain is related to whether or not someone has Alzheimer's disease, as these structures affect the brain in ways that cause the symptoms of AD. Another set of variables that seem to be positively correlated is `PMI` and `DiseaseDuration` , with a correlation of 0.26. This correlation could be a result of the researchers' process for sample collection, since post-mortem interval and disease duration do not seem to have a genetic or biological relationship. `DiseaseDuration` has a negative correlation with both `PlaqueTotal` and `TangleTotal` . When the duration of the disease is long, then both plaque and tangle amounts are small, and vice versa.

In CSF, the most positively correlated variables are `PlaqueTotal` and `TangleTotal` , both of which are found in similar amounts in Alzheimer's patients. `DiseaseDuration` and `expired_age` were found to be negatively correlated, which follows because if the disease lasts a prolonged period of time, the patient is more likely to die sooner rather than later.

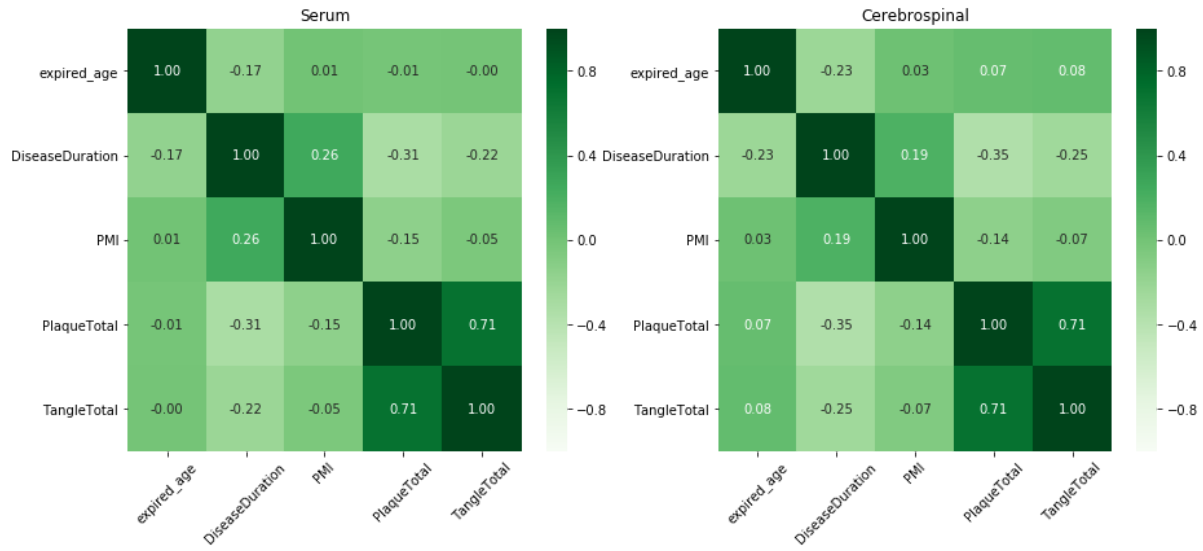


Figure 16: Serum (left) and Cerebrospinal (right) Biofluid Heatmap Correlations.

We have seen that the overall correlation between Plaque Total and Tangle Total in both biofluid samples are high, but after previous results from Table 3, we know that the distribution of those two measurements are different for the AD group with high totals. Therefore, we plot the correlation per disorder in Table 4 and Table 5 that shows that Plaque Total and Tangle Total have a higher correlation of AD compared to PD and healthy; specifically serum has a correlation of 0.68 in AD compared to the Healthy and PD which are 0.42 and 0.51 respectively and similarly cerebrospinal has a correlation of 0.68 in AD compared to the Healthy and PD which are 0.42 and 0.48 respectively.

		PlaqueTotal	TangleTotal
CONDITION			
Alzheimer's Disease	PlaqueTotal	1.00	0.68
	TangleTotal	0.68	1.00
Healthy Control	PlaqueTotal	1.00	0.42
	TangleTotal	0.42	1.00
Parkinson's Disease	PlaqueTotal	1.00	0.51
	TangleTotal	0.51	1.00

Table 4: Serum Biofluid PlaqueTotal and TangleTotal Correlation

CONDITION		PlaqueTotal	TangleTotal
Alzheimer's Disease	PlaqueTotal	1.00	0.68
	TangleTotal	0.68	1.00
Healthy Control	PlaqueTotal	1.00	0.42
	TangleTotal	0.42	1.00
Parkinson's Disease	PlaqueTotal	1.00	0.48
	TangleTotal	0.48	1.00

Table 5: Cerebrospinal Biofluid PlaqueTotal and TangleTotal Correlation

Gene Matrix

Another important step before we step into the formal analysis is to get familar with our gene matrix. Here, we explored some basic properties including counts, missingness, basic distributions and basic correlations. In the full gene matrix that includes all the sequences, we have a lot of NaN values because not every sequence is detected in our sample. We started by looking at how many sequences are missing for one sample and how many samples do not have a certain sequence, that is, the number of missing values of the columns and rows of the matrix.

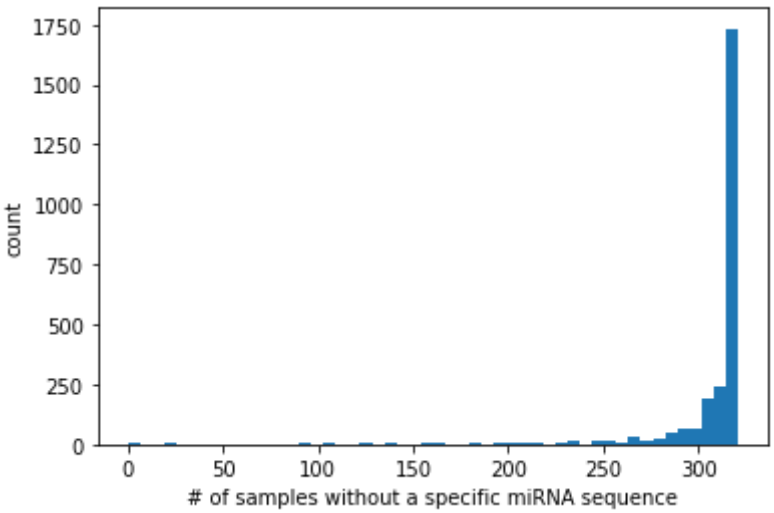


Figure 17: Distribution of miRNA count missingness.

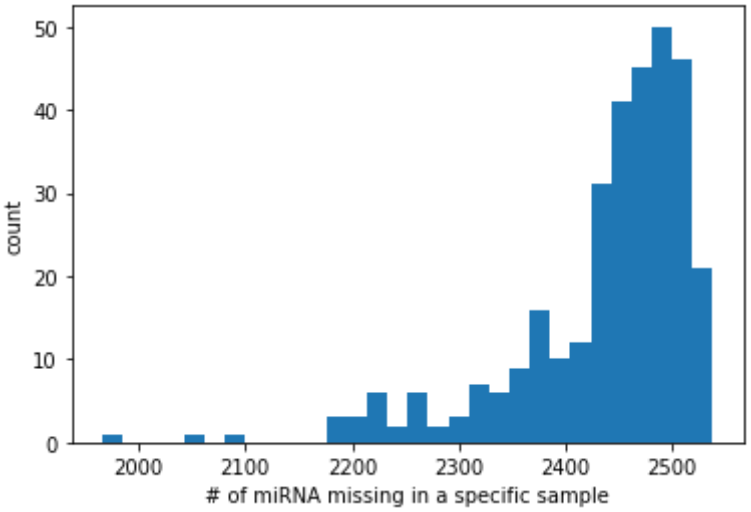


Figure 18: Distribution of missingness of samples.

According to Figure 17 and 18 above, a lot of miRNA sequences only exist in some samples (Figure 17) and most of the samples have a large amount of miRNA sequences missing. It indicates that miRNA sequence set may be specific to individuals, which makes our finding even more interesting if we can find several common miRNAs that are significantly up or down regulated among all samples.

Processing of Data

Merging counts

The data target step automatically downloaded from the exRNA Atlas database a gene count file for each SRA Run sample. This gene count table has a column that indicates the abundance count. The 322 abundance files were merged into one gene count file. Special care was needed to ensure that the columns and rows match up - the columns were the SRA runs, and the rows were the gene counts. The output was a table with over 180K genes, however most of the genes had no overlap amongst the SRA samples. After removing null rows, the final gene matrix table had 400+ miRNA. A portion of the final gene matrix table is shown in Table 6 below.

	SRR1568478	SRR1568692	SRR1568530	SRR1568514
miRNAs				
mir-39-3p	82	104	87	121
mir-54	61	54	61	122
mir-238	87	81	60	80
mir-22-3p	9	18	28	23

Table 6: Subset of the gene matrix of top 4 miRNA's based on count versus first 4 SRA Run's.

Another output of the merge step was to generate the feature experiment table shown in Table 7 which has the features of interest for all the SRA Run's, included in the gene matrix.

Run	Disorder	expired_age	Biofluid	sex	PMI	sn_depigmentation	Braak_Score
SRR1568567	Control	94	Cerebrospinal	male	2.50	none	IV
SRR1568730	Parkinson	79	Serum	female	6.00	severe	II
SRR1568666	Alzheimer	81	Serum	female	2.50	mild	V
SRR1568510	Parkinson	79	Cerebrospinal	male	2.50	severe	II
SRR1568518	Parkinson	82	Cerebrospinal	female	4.16	severe	III

Table 7: Feature table.

Furthermore, the following cleanup was done in the merge step:

- The sample SRR1568391 was removed due to the SRA Run table having two rows which had the same SRA number but with different values.
- The imputation with TangleTotal attribute as it had 2 samples with missing data which was replaced with the mean (6.63).

Normalized Gene Count

The process of normalization used the merged gene matrix and feature experiment table generated from the merge step, and then used DESeq2's transform to generate a normalized count matrix file. As well as outputting the standard normalized count matrix we also performed Variance Stabilization Transformation (VST) to generate an additional normalized matrix which used the parametric fitting type. VST transforms data is by creating new values in terms of y where the variability of the new y-values is unrelated to the x-values [23]. VST finds a function that can be applied to the original x values to generate the new y-values. Methods like VST and normalization allow us to primarily scale our data so that it is in a format that allows us to perform further analysis. The values in our data become more manageable all while still maintaining their original statistical importance and meaning, especially if our variables in our pre-normalized data have different scales [24].

The descriptive statistics for both the normalized and the VST normalized gene matrix is shown in Table 8 and 9 below.

	Average Sample
count	400.000
mean	1.742
std	5.366
min	1.018
25%	1.018
50%	1.018
75%	1.034
max	70.782

Table 8: Normalized gene matrix descriptive statistics summarized for all samples.

	Average Sample
count	400.000
mean	0.263
std	0.679
min	0.087
25%	0.087
50%	0.087
75%	0.101
max	6.045

Table 9: VST gene matrix descriptive statistics summarized for all samples.

The descriptive statistics showed as expected the VST had a smaller range of values. However, we wanted to verify that the two sets of normalized gene count matrices were correlated. For this we took a number of SRA samples from each matrix and compared them against each other. Figure 13 below shows the correlation for SRR1568567 as well as SRR1568584, seen in Figure 19 below. The result of the R^2 Pearson correlation of 0.97 shows a strong indication that the data from both matrices is consistent. For further downstream processing, we used the VST gene matrix.

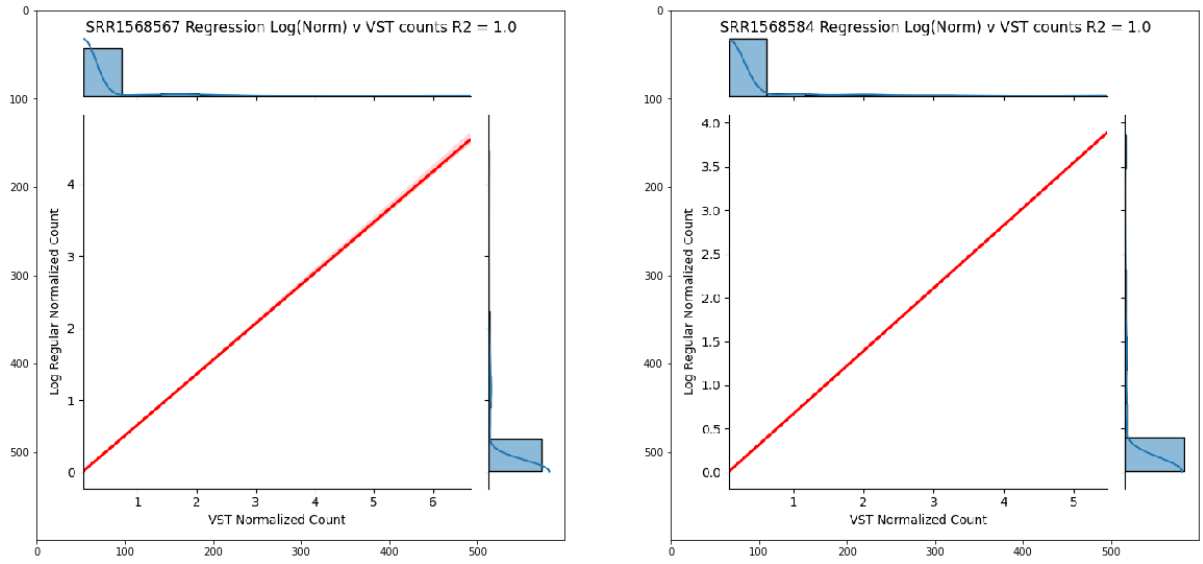


Figure 19: Two SRA Run's regression of log norm vs VST counts.

Remove missing sequences

We merged the counts based on the readCounts_miRNAmature_sense.txt for each sample. This generated a merged gene matrix that has approximately 1,800 rows, but many were repeated miRNAs. So we grouped rows together based on the same miRNA to avoid miRNA counts that were artificially low. Finally, based on the original paper [1] we reduced our focus to the top 400 miRNAs by prioritizing the rows that had the most total gene counts.

PCA

We performed Principal Component Analysis (PCA) on the VST gene matrix for two PCA Plots, one with the grouping set to disorder and another plot with the grouping set to biofluid.

PCA uses linear combinations to explain the variance-covariance structure of a set of variables. Data reduction and data interpretation are the main reasons for the use of PCA with the latter being the method we incorporated in the replication project [25]. For the purposes of this project, we did not remove any samples from our data thus negating the need to do any sort of dimensionality reduction of our data. In doing so, we did not experience any reduction in our data in terms of size and scope which would be common in other PCA implementations. We used PCA purely on an exploratory level where we could observe relationships within our data that may not have been as obvious to us. In Figure 20 below, there are visibly two groups formed in both the left and right charts. However, the spread of every disorder and biofluid, respectively, is relatively equal across the chart.

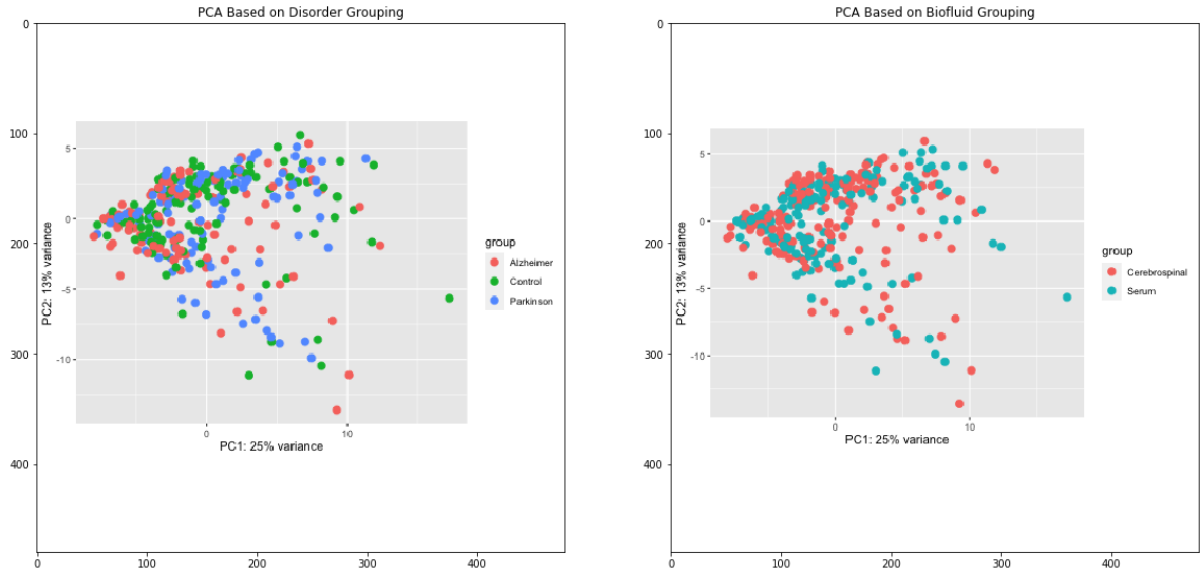


Figure 20: (left) PCA based on grouping by disorder. (right) PCA based on grouping by biofluid.

Missing of Genes

Another analysis we did was with the entire set of counts which not only included miRNA's but also protein coding genes. The combined gene matrix file has approximately 180K genes but most were missing across all samples and are thus not relevant for analysis and compromise quality and performance of the analysis.

Furthermore, we can identify the top and bottom genes based on a statistic for each gene across the samples that measures the spread against the mean normalized count. Genes which had little spread are likely candidates that might not be important as they do not significantly vary across the samples. The spread statistic we developed was the L1 distance against the mean. Table 10 below shows the top 3 and bottom 3 genes and the spread values. The gene with the highest count variance was mir-486a-5p and the bottom ranked gene was mir-2797d.

Top Ranked Genes

	Gene	Spread
0	mir-486a-5p	513.508
1	mir-27b-3p	456.435
2	mir-143	393.827

Bottom Ranked Genes

	Gene	Spread
397	mir-19d	1.53166
398	mir-199a-5p	1.53166
399	mir-2797d	1.53166

Table 10: (left) the top ranked genes and their spread. (right) the bottom ranked genes and their spread.

Data Analysis

The analysis of the data was performed using the same technique in the research paper [1], namely to consider each biofluid separately, and within each biofluid to consider each disorder (versus control) separately. This resulted in 4 combinations of analysis computations that were performed. With the top genes identified, we filtered only the samples for the particular biofluid (Serum, Cerebrospinal fluid) and then further filtered based on one of the 2 disorders (Alzheimer's, Parkinson's) plus the control. The basis of the analysis was the Likelihood Ratio Test (LRT), which is a hypothesis test based on a full and reduced model using the DESeq2 package. The model used the following variables:

- expired_age
- sex
- PMI
- sn_depigmentation
- Braak_Score
- TangleTotal
- Plaque_density
- PlaqueTotal

The outcome variable was Disorder which was not included in the reduced model. The premise of the LRT is to compare models in terms of how they fit the available data by comparing the likelihood scores of the two models [9] using a statistical test of the goodness-of-fit between two models. The full model with Disorder is compared to a reduced model without Disorder [9]. The output from the analysis was a LRT table which included the baseMean, log2 Fold Change, lfcSE, stat, the pvalue, and adjusted pvalue. The descriptive summary of LRT in the cerebrospinal fluid for Parkinson's is shown in Table 11 below.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
count	400.000	400.000	400.000	400.000	400.000	400.0
mean	2.863	0.028	0.430	0.105	0.871	1.0
std	6.173	0.133	0.020	0.305	0.185	0.0
min	2.022	-0.405	0.285	0.000	0.095	1.0
25%	2.031	-0.002	0.429	0.000	0.827	1.0
50%	2.065	0.004	0.436	0.003	0.960	1.0
75%	2.178	0.046	0.437	0.048	0.996	1.0
max	79.698	0.723	0.540	2.792	1.000	1.0

Table 11: The LRT descriptive summary for biofluid Cerebrospinal fluid for Parkinsons.

Visualizations

MA Plot

The LRT data for each of the 4 comparisons was used to generate a 2x2 MA Plot shown in Figure 21 below. This is a scatter plot of the mean of the normalized counts against the log fold change. There does not appear to be any highly significant patterns as the data points appear to have a similar distribution.

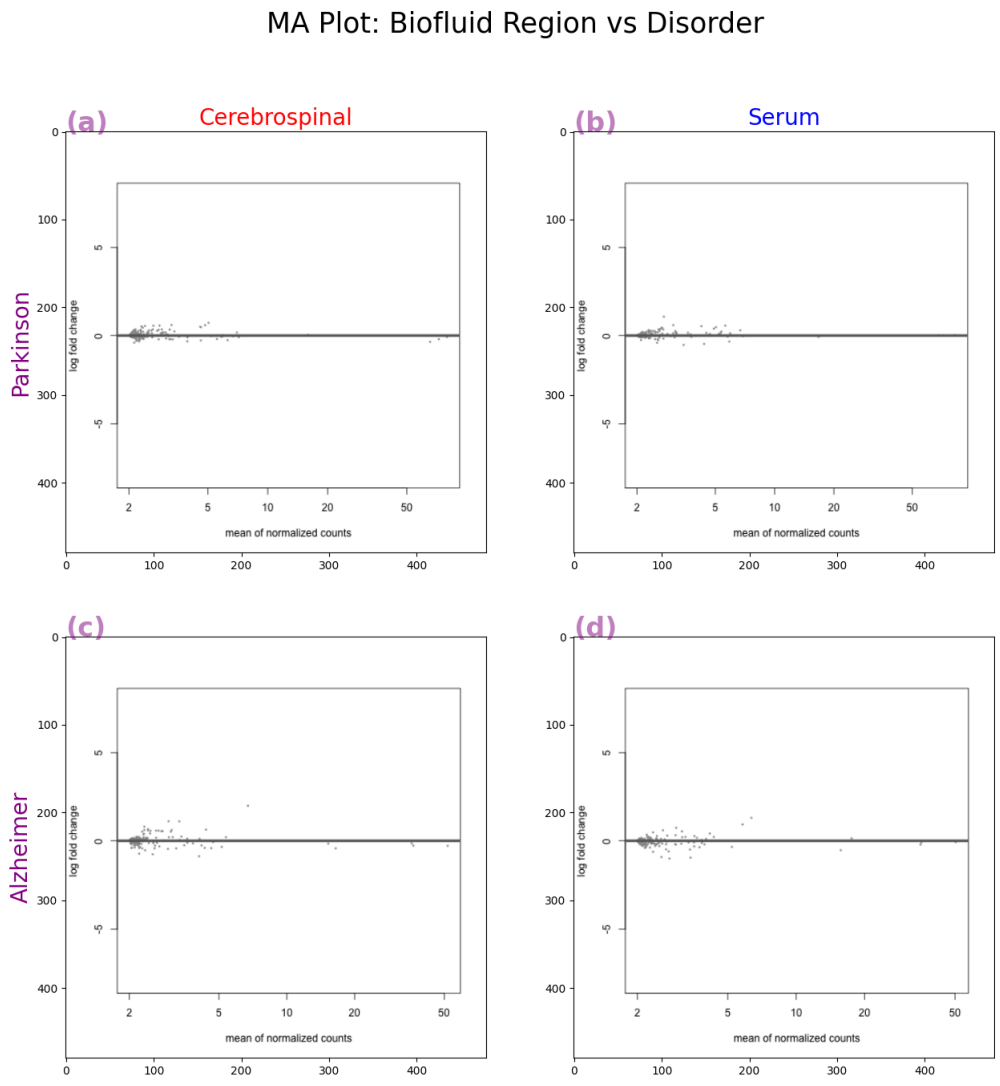


Figure 21: MA Plot for each biofluid versus each disorder.

Heat Maps

Heatmaps were generated from the LRT data by selecting the top 20 miRNA's based on the average mean normalized count and then plotting against the SRA samples. Each of the 4 analysis regions is plotted in a 2x2 heatmaps in Figure 22 below. Heatmaps allow us to observe specific values of interest across two axis variables in the form of a grid with colored cells [26]. The variables we have used for our axes are the patients and the miRNAs identified across our patient samples, with the main value of interest being the miRNA expression between patients of different disorders. By observing our value of interest, miRNA expression, we can determine if any patterns or associations exist within our 4 analysis regions. It appears that Parkinson's across the biofluids are more similar and Alzheimers across the biofluids are more similar, this is based on the general trends seeing that Parkinson's has less intensity with raw Z-scores closer to 0 and Alzheimers having higher intensity with extreme raw Z-scores -4 and 4. However, one noticeable difference is that the placement of the coloring across the heat map is more similar between the biofluids. Meaning that cerebrospinal fluid across Parkinsons and Alzheimers has a more similar color mapping across starting with higher Z-scores (red) then to lower Z-scores (blue). Similarly, serum across Parkinsons and Alzheimers has a more similar color mapping across starting with lower Z-scores (blue) then to higher Z-scores (red). This means that Alzheimers has higher raw Z-scores than Parkinsons, but more importantly the notable similarities in the color scheme across biofluids means that there is some commonality between diseases.

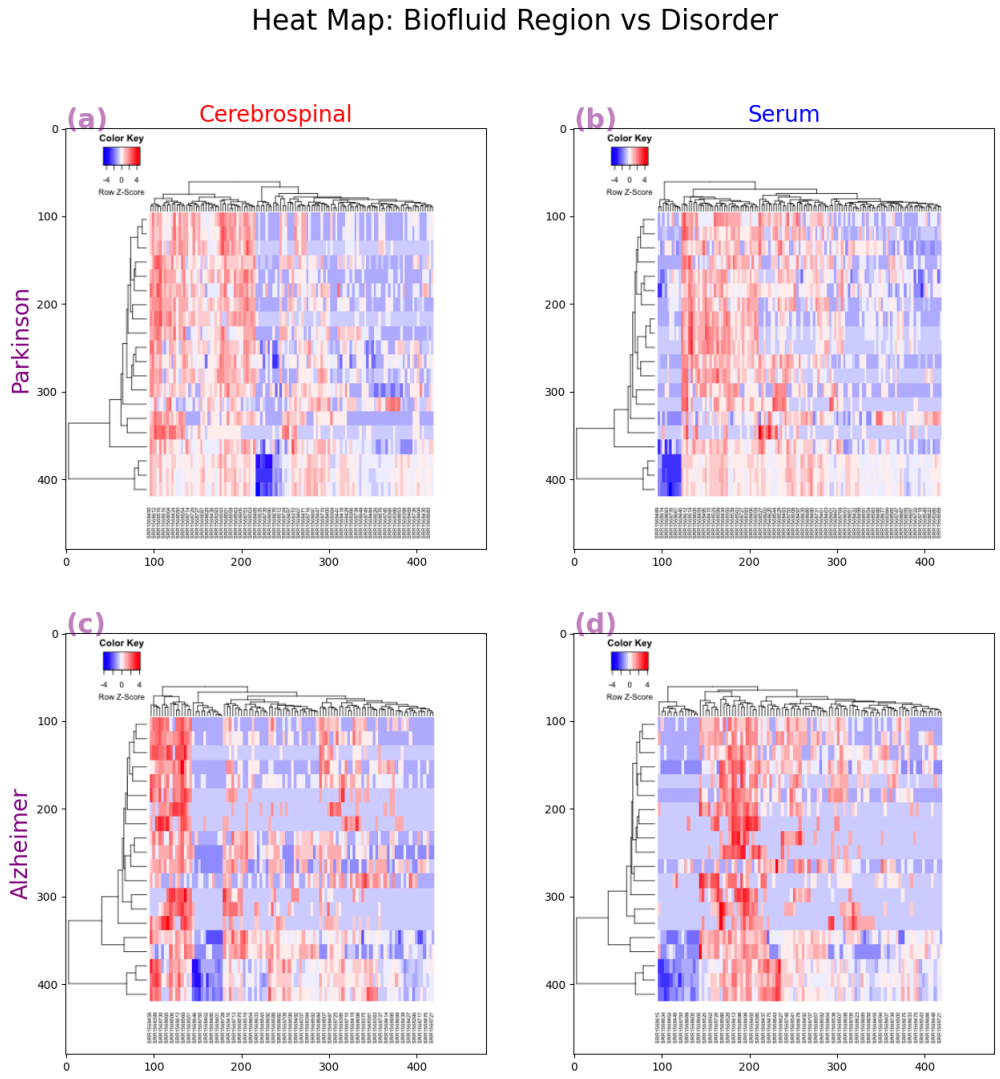


Figure 22: Top 20 Expressed miRNA Heatmap for each biofluid versus each disorder.

Histograms of case versus control differential expression

We plot the histogram distribution of the pvalues for each of the 4 regions shown in Figure 23 below. Each plot below has its peak at 1.0 with a general increase across. Interestingly, the charts for cerebrospinal vs Parkinsons and Alzheimers look similar and the charts for serum vs Parkinsons and Alzheimers are more similar. This is significant to note because rather than the disease being more similar across biofluids, it happens to be that the biofluids are more similar across the two disorders.

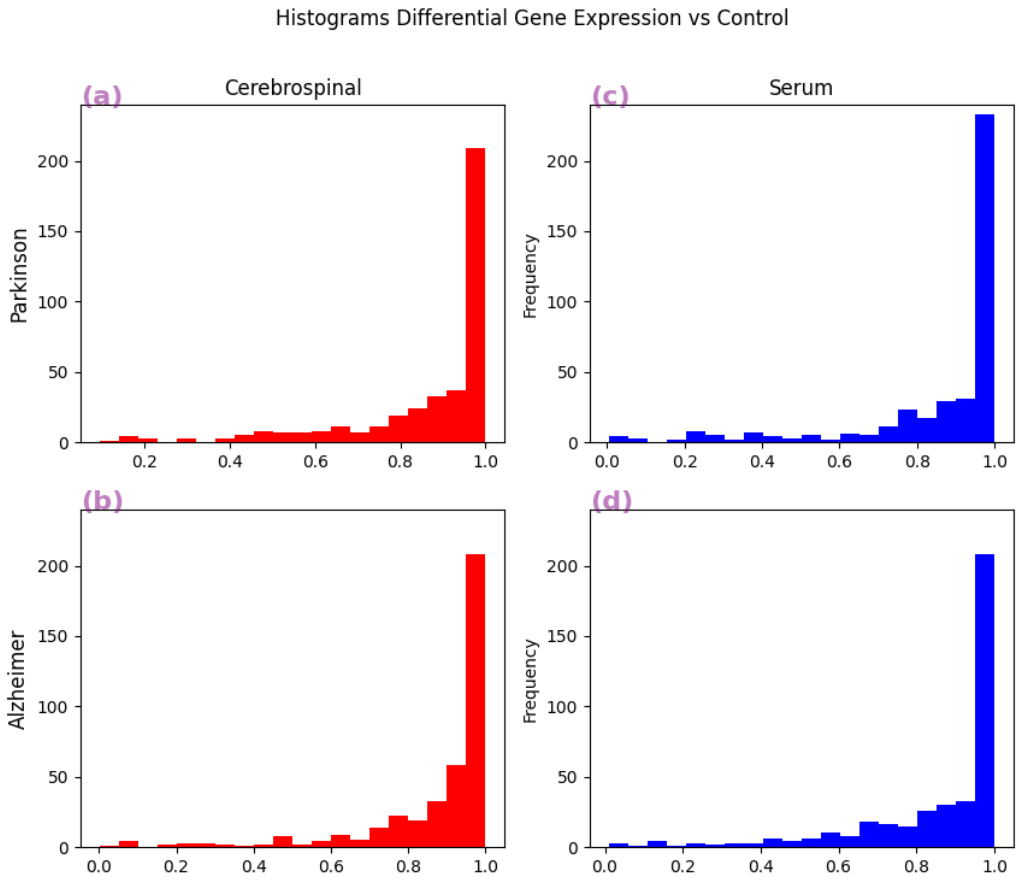


Figure 23: Histogram of pvalue for each biofluid versus each disorder.

Venn Diagram of Disorders

Another visualization performed is a venn diagram seen in Figure 24 below that shows overlap of miRNAs differentially expressed between Parkinson (red) and Alzheimer (green). Majority of miRNAs are in Parkinson's with 17 listed. There are 14 miRNAs in Alzheimer's. But between the two disorders there are 13 miRNAs shared. Therefore, there is still a good amount of miRNA overlap between the diseases.

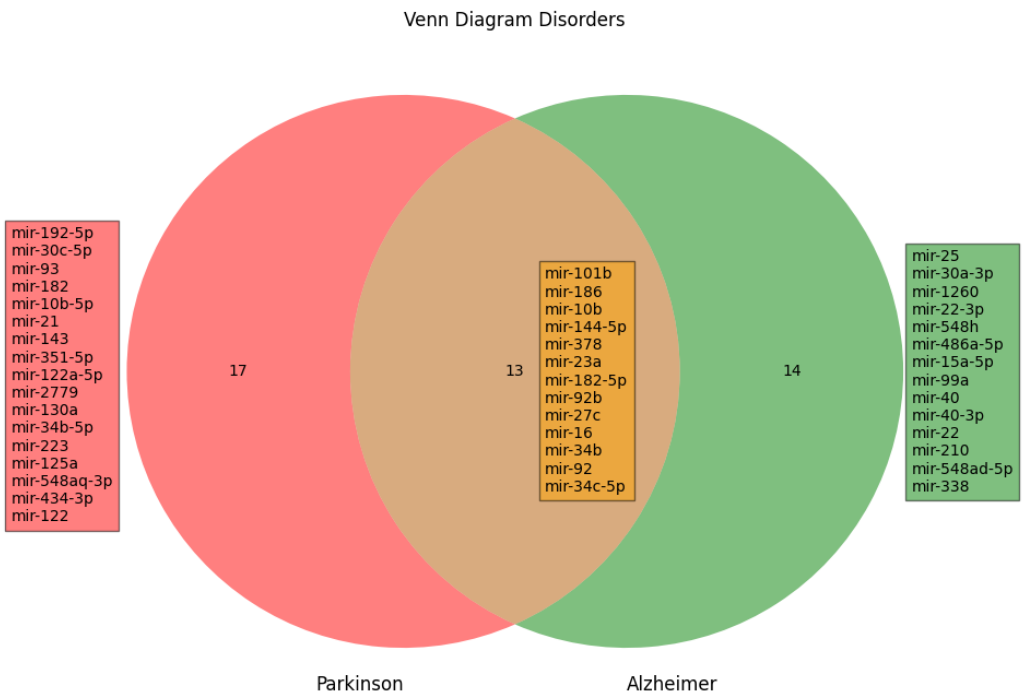


Figure 24: Venn Diagram showing miRNA overlap between Parkinson's and Alzheimer's in both biofluids.

Spearman correlations of log2 fold gene expression

A spearman correlation matrix is shown in Figure 25 which is a pairwise Spearman correlation of log2 fold gene expression changes between each disorder and CTL in each biofluid. The circle sizes are scaled to reflect absolute Spearman correlations. To produce this plot we took the log2 fold gene expressions column from each of the 4 LRT analyses we performed and then used Pandas correlation function to generate a R^2 Pearson correlation number.

The striking amount of correlation is between Parkinson and Alzheimers in the cerebrospinal fluid region. This is very important as it implies that there is a significant amount of correlation between the two disorders for this biofluid with the highest correlation of 0.30. There is also a significant amount of correlation between Parkinson's and Alzheimer's in the serum region with a correlation of around 0.27. There is high correlation between the biofluids and the two disorders, especially in the cerebrospinal fluid region.

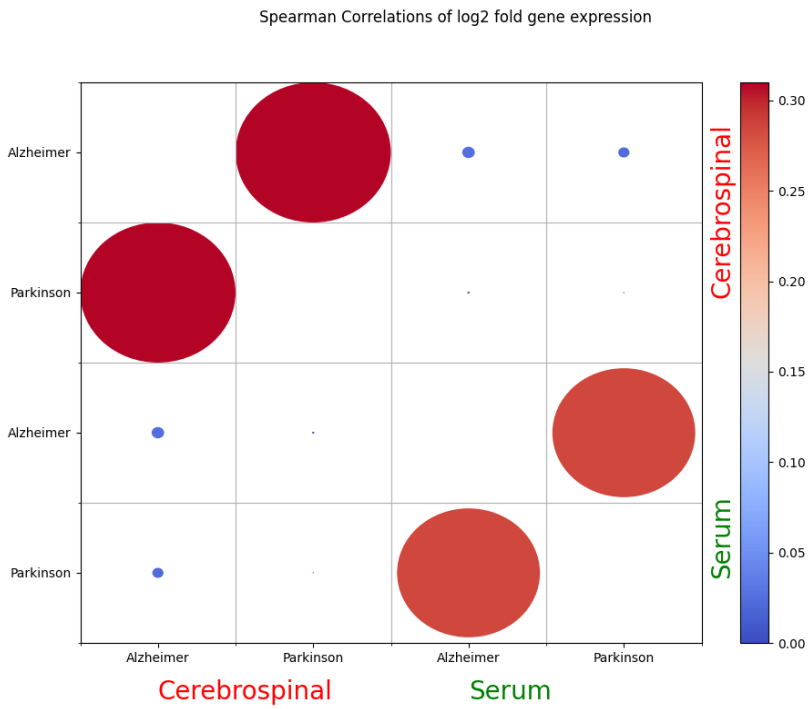


Figure 25: Spearman correlation of log2 fold gene expression for each biofluid against each disorder.

Volcano Plot

Figure 26 shows a volcano plot with the two disorders against the two biofluid regions. Volcano plots help to differentiate the down and up regulated miRNA sequences with respect to the control group (healthy patients). It is worthwhile to note that when miRNAs are downregulated, this means that there is less miRNA expression - because miRNA regulates mRNA expression, this results in more mRNA expression. mRNA, of course, dictates what proteins are synthesized in the body. In that same vein, when miRNA are upregulated, this means that they regulate mRNA expression at a higher rate, causing less mRNA expression. Chart (a) shows only up regulated miRNAs. Chart (b) shows both down and up regulated miRNAs, with mostly up regulated. Chart (c) shows mostly up regulated miRNAs with few down regulated ones as well. Finally, chart (d) shows a higher amount of down regulated miRNAs with some up regulated ones as well.

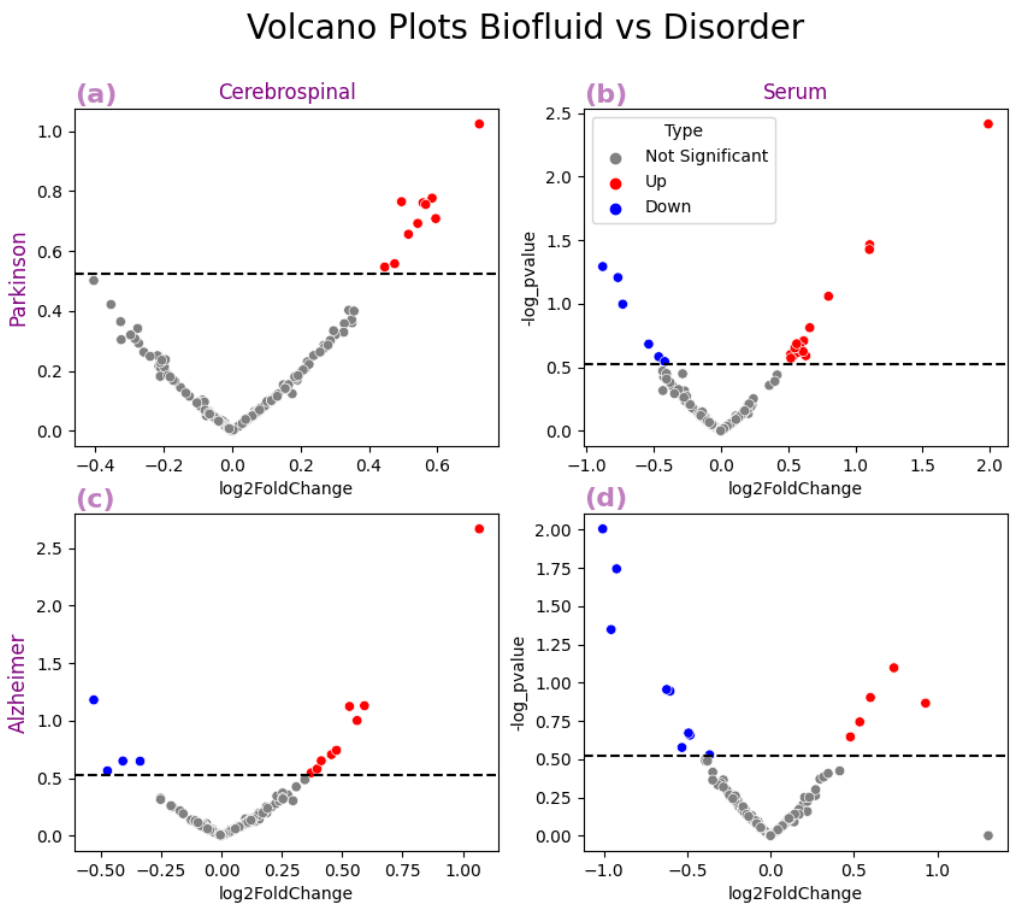


Figure 26: Volcano plot Biofluid versus Disorder.

Additional detailed volcano plots were made to that included labels for the potentially important miRNA's. Figure 27, shows the Alzheimer vs Serum detailed Volcano plot which shows the miRNA that are upregulated and downregulated. These miRNA's are further analysed by mapping them to the specific mRNAs that they regulate the expression of in the following sections.

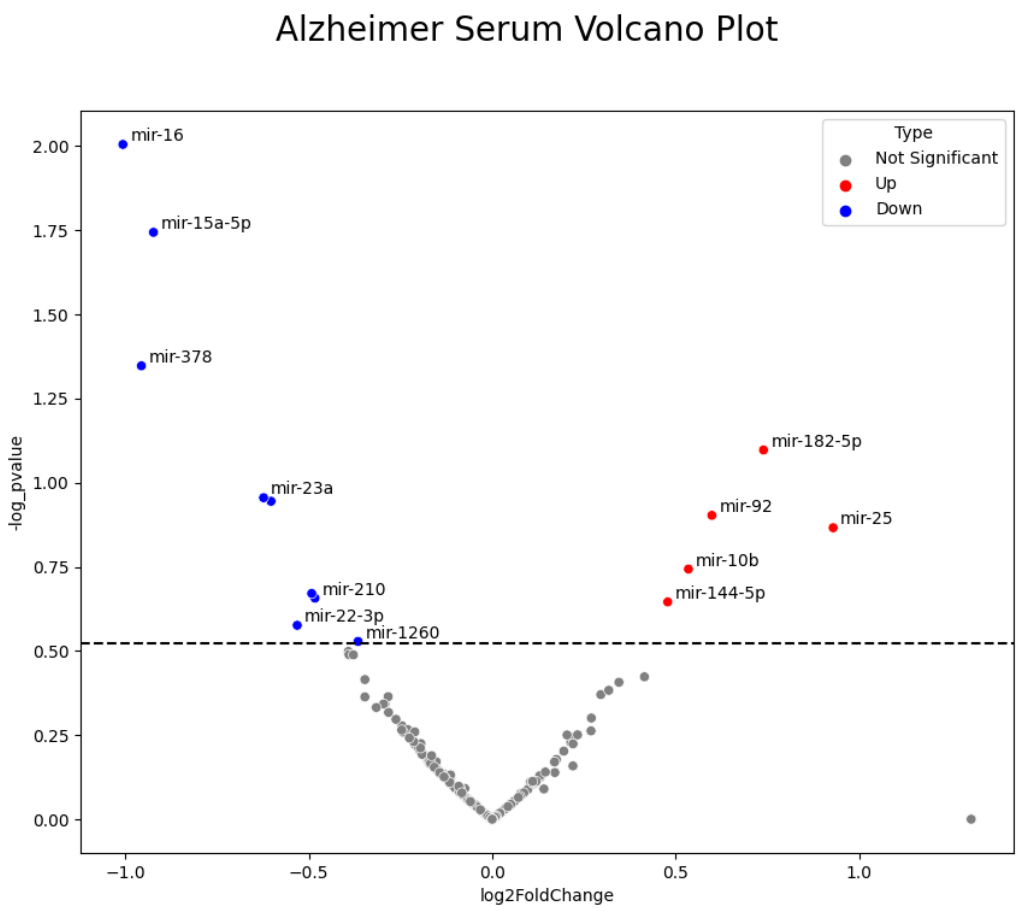


Figure 27: Alzheimer vs Serum Detailed Volcano plot with miRNAs.

Mappings

In oue final analysis, we mapped the overlapping miRNAs of interest to their target mRNAs, and subsequently, to the genes they encode.

When mapping certain miRNAs to the specific mRNAs that they regulate, one issue we ran into was that a miRNA can regulate hundreds of mRNAs. This is because miRNAs only require a small amount of nucleotide matches to be able to latch onto an mRNA and suppress its expression. Therefore, miRNAs can have many target mRNAs. Of course, we are more interested in a small amount of relevant mRNAs, specifically those that encode proteins involved in neurological processes. So, we chose to focus our analysis on the top 3 target mRNAs for each miRNA, ranked in terms of the mRNA's (gene's) "Target Rank" and "Target Score", dictated by miRDB.org. The site allowed us to search miRNA sequences for their target genes. As mentioned above, since miRNA can have hundreds of matches, these are just a few matches, some selected because of their high "Target Ranks" and "Target Scores", and others because of past research done by scientists on the proteins' involvement with neurological processes.

Note: The miRNA and their mappings included here are not the *full* list of miRNAs and their proteins. Those included in each section are the miRNA that repressed mRNA more relevant to the disorders we are studying, primarily encoding proteins responsible for certain brain functions.

Cerebrospinal & Alzheimer's Mappings

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*, *phospholipase D1 (PLD1)*, *sortilin related receptor 1 (SORL1)*

mir-34b: insulin induced gene 1 (INSIG1), protein phosphatase 6 regulatory subunit 3 (PPP6R3), furin, paired basic amino acid cleaving enzyme (FURIN), *neuroplastin (NPTN)*

mir-338: Cbl proto-oncogene (CBL), galectin like (LGALSL), RAB14, member RAS oncogene family (RAB14), neuropilin 1 (NRP1), *phosphatidylinositol binding clathrin assembly protein (PICALM)*

mir-548h: CREB binding protein (CREBBP), ubiquitin conjugating enzyme E2 D1 (UBE2D1), zinc finger CCHC-type containing 14 (ZCCHC14), *neuron navigator 2 (NAV2)*

mir-34c-5p: family with sequence similarity 76 member A (FAM76A), delta like canonical Notch ligand 1 (DLL1), MDM4, p53 regulator (MDM4), *neuron navigator 1 (NAV1)*, *neuron navigator 3 (NAV3)*, microtubule associated protein tau (MAPT)

Based on prior research, Alzheimer's disease has primarily been linked with a protein called apolipoprotein E (APOE) [27]. Researchers have found that the presence of this particular gene has been associated with the formation of amyloid plaques (the presence of which we studied and incorporated into our DESeq2 model). These protein clumps "clog up" the brain and lead to the death of nerve cells. Unfortunately, the mapping of the differentially expressed miRNA in the CSF of Alzheimer's patients did not uncover this protein. Another gene commonly linked with AD is the tau protein, which contributes to the "tangles" in Alzheimer's brains (we also studied this in our analysis). In this vein, some of the genes that were found to be affected by one of the up-regulated miRNAs were tau tubulin kinase 2 (TTBK2) and microtubule associated protein tau (MAPT). These genes contribute to making the protein that forms the tau tangles in the brain. However, there are other proteins that have been found to be related to the onset of Alzheimer's [28]. One such gene was complement C3b/C4b receptor 1 (CR1). The upregulated miRNA represses the expression of this gene, which decreases the production of a protein in the brain that is partly responsible for controlling brain inflammation. The absence of this protein can result in inflammation, a possible cause of Alzheimer's. Another such gene is the one that encodes phosphatidylinositol binding clathrin assembly protein (PICALM), which contributes to the process of neurons communicating signals to each other and ensuring that the right communication happens in order for the body to function properly, as well as the process of memory formation [28]. The absence of this protein can negatively affect these processes. There are also neurological proteins that are affected by the up-regulation of miRNA. For example, the amount of neuroplastin available in the brain is affected by some miRNA. Neuroplastin is a protein that is important in neuron and synaptic functions; in other words, they are significant in the process of cells communicating with each other [29].

Cerebrospinal & Parkinson's Mappings

mir-34b: insulin induced gene 1 (INSIG1), protein phosphatase 6 regulatory subunit 3 (PPP6R3), furin, paired basic amino acid cleaving enzyme (FURIN), *neuroplastin (NPTN)*

mir-34c-5p: family with sequence similarity 76 member A (FAM76A), delta like canonical Notch ligand 1 (DLL1), MDM4, p53 regulator (MDM4), *neuron navigator 1 (NAV1)*

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*

mir-130a: gap junction protein alpha 1 (GJA1), cytoplasmic polyadenylation element binding protein 1 (CPEB1), SKI/DACH domain containing 1 (SKIDA1), *leucine rich repeat kinase 2 (LRRK2)*

mir-34b-5p: teneurin transmembrane protein 1 (TENM1), ELMO domain containing 1 (ELMOD1), regulatory factor X3 (RFX3), *parkin RBR E3 ubiquitin protein ligase (PRKN)*

mir-23a: zinc finger protein 99 (ZNF9), semaphorin 6D (SEMA6D), family with sequence similarity 234 member B (FAM234B)

Based on prior research, researchers have pinpointed some possible genes that, coupled with family history, can mutate and cause Parkinson's disease [30]. We were able to find miRNA that affected the expression of some of these genes, and consequently the proteins that they help to create. One such gene is leucine rich repeat kinase 2 (LRRK2). This gene encodes a protein called dardarin, which plays a big role in biological processes that require inter-protein interaction, like the transmitting of signals between neurons or assembling a cell's cytoskeleton (its physical framework) [31]. Another protein found to be related to Parkinson's is parkin RBR E3 ubiquitin protein ligase (PRKN). This gene encodes the protein parkin, which helps in the cell by tagging unneeded proteins with markers called ubiquitin. This lets other parts of the cell know that those proteins are unneeded, so they are properly disposed in structures called proteasomes. With the absence of parkin, this system is compromised, and the build-up of unnecessary proteins may lead to issues that cause physical movement and balance problems associated with PD. The failure of the ubiquitin-proteasome system can also affect normal cell activities and the cells themselves, specifically those that produce dopamine. Decrease of dopamine production is a tell-tale sign of Parkinson's [32]. One neurologically-related protein that we found to be affected by miRNA expression was neurofilament medium (NEFM). The neurofilament medium protein encodes the protein neurofilament, which is used by cells to mark neurons that are damaged. If this system is affected, there would be no way to distinguish between working and damaged neurons, gravely affecting neuronal activity [33].

Serum & Alzheimer's Mappings

mir-16: pappalysin 1 (PAPPA), fatty acid synthase (FASN), unc-80 homolog, NALCN channel complex subunit (UNC80), *clusterin (CLU)*, *triggering receptor expressed on myeloid cells 1 (TREM1)*, *neurofibromin 1 (NF1)*

mir-186: RUN and FYVE domain containing 3 (RUFY3), zinc finger CCCH-type containing 11A (ZC3H11A), zinc finger protein 644 (ZNF644), *neuronal growth regulator 1 (NEGR1)*

mir-92: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *sortilin related receptor 1 (SORL1)*

mir-10b: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6), *brain derived neurotrophic factor (BNDF)*

mir-22: glutamate metabotropic receptor 5 (GRM5), fucosyltransferase 9 (FUT9), *neuroepithelial cell transforming 1 (NET1)*

mir-210: insulin like growth factor 2 (IGF2), iron-sulfur cluster assembly enzyme (ISCU), galanin receptor 2 (GALR2), brain derived neurotrophic factor (BDNF), *neuronal pentraxin 1 (NPTX1)*

mir-144-5p: zinc finger protein 292 (ZNF292), ATPase H+ transporting V1 subunit C1 (ATP6V1C1), HIC ZBTB transcriptional repressor 1 (HIC1), neurotrophic receptor tyrosine kinase 2 (NTRK2), *neuregulin 3 (NRG3)*

As mentioned above when inspecting Alzheimer's samples from CSF, researchers have already pinpointed some genes that have a likely connection with Alzheimer's. One of these genes is the triggering receptor expressed on myeloid cells 1 (TREM1). Receptors on myeloid cells are responsible for controlling inflammation and neurological development [34]. Inflammation, especially in the brain, as mentioned earlier, is a tell-tale sign of Alzheimer's. Clusterin (CLU) is a gene that helps to regulate amyloid-beta amounts in the brain - these, as we know, make up the plaque structures that are found in Alzheimer's brains. An imbalance in the production and movement of amyloid-beta is then crucial to the development of Alzheimer's [28]. Another such gene is sortilin related receptor 1 (SORL1). SORL1 is a gene that is involved in the production of amyloid-beta peptides, which are the same plaque structures that are found in the brain of Alzheimer's patients [35]. There were also neurologically-related genes that were found to be affected by the differentially expressed miRNA. For example, the count of brain derived neurotrophic factor (BDNF) was found to be decreased by an up-regulated miRNA. The BDNF protein is largely responsible for promoting the growth of and dealing with the maintenance of nerve cells [36]. One target gene of many miRNA in this group was neuregulin 3 (NRG3), which is a group of signaling proteins that helps to oversee cellular functions of neuronal systems, like survival, proliferation, and differentiation of nerve cells [37]. Neuronal pentraxin 1 (NP1) is another gene that was affected by miRNA. The miRNA that targeted NP1 was actually found to be down-regulated, which increases the production of NP1; NP1 is involved in the process of inducing neuronal cell death, and a surplus of NP1 could result in more neurons being destroyed prematurely [38].

Serum & Parkinson's Mappings

mir-192-5p: NIPA like domain containing 1 (NIPAL1), basic helix-loop-helix family member e22 (BHLHE22), protein kinase D3 (PRKD3), *neurofilament light (NEFL)*

mir-182-5p: protein kinase cAMP-activated catalytic subunit beta (PRKACB), regulator of G protein signaling 17 (RGS17), basonuclin 2 (BNC2), *neurocalcin delta (NCALD)*

mir-10b-5p: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6), *brain derived neurotrophic factor (BDNF)*

mir-144-5p: zinc finger protein 292 (ZNF292), ATPase H+ transporting V1 subunit C1 (ATP6V1C1), HIC ZBTB transcriptional repressor 1 (HIC1), *neuregulin 3 (NRG3)*

mir-92: folliculin interacting protein 1 (FNIP1), CD69 molecule (CD69), G3BP stress granule assembly factor 2 (G3BP2), *neurofilament medium (NEFM)*

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*

mir-30c-5p: twinfilin actin binding protein 1 (TWF1), UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5 (B3GNT5), embryonic ectoderm development (EED), *neural cell adhesion molecule 1 (NCAM1)*, *leucine rich repeat kinase 2 (LRRK2)*

mir-548aq-3p: polyhomeotic homolog 3 (PHC3), CREB3 regulatory factor (CREBRF), protein tyrosine phosphatase, receptor type K (PTPRK), *synuclein alpha (SNCA)*

mir-186: RUN and FYVE domain containing 3 (RUFY3), zinc finger CCCH-type containing 11A (ZC3H11A), zinc finger protein 644 (ZNF644), *neuronal growth regulator 1 (NEGR1)*

mir-16: pappalysin 1 (PAPPA), fatty acid synthase (FASN), unc-80 homolog, NALCN channel complex subunit (UNC80), clusterin (CLU), *triggering receptor expressed on myeloid cells 1 (TREM1)*, neurofibromin 1 (NF1)

mir-223: F-box and WD repeat domain containing 7 (FBXW7), SP3 transcription factor (SP3), *synuclein alpha (SNCA)*, *neuron derived neurotrophic factor (NDNF)*

In this group of data/results, something that stood out was the overlap of many proteins with the Alzheimer's in serum group, as well as the two CSF groups. Proteins like neurofilament, neuregulin, BDNF, and neuronal growth regulators have been affected by up- and down-regulated miRNAs in both the CSF and serum groups for the two disease conditions. One new protein that seems to be related to Parkinson's and serum specifically, though, is neurocalcin delta (NCALD). A decrease in NCALD has been shown to protect against spinal muscular atrophy, a symptom tangentially related to the symptoms of Parkinson's [39]. Another target protein that has also been pinpointed by researchers in the past to be connected to Parkinson's is synuclein alpha (SNCA). This is one of the most common proteins linked to Parkinson's - mutations of this protein can disrupt cell homeostasis and neuron death [40].

Overlapping miRNA

These mappings are from the Venn Diagram in Figure 24.

mir-16: mir-16: pappalysin 1 (PAPPA), fatty acid synthase (FASN), unc-80 homolog, NALCN channel complex subunit (UNC80), *clusterin (CLU)*, *triggering receptor expressed on myeloid cells 1 (TREM1)*, *leucine rich repeat kinase 1 (LRRK1)

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*

mir-34b: insulin induced gene 1 (INSIG1), protein phosphatase 6 regulatory subunit 3 (PPP6R3), furin, paired basic amino acid cleaving enzyme (FURIN), *neuroplastin (NPTN)*

mir-182-5p: protein kinase cAMP-activated catalytic subunit beta (PRKACB), regulator of G protein signaling 17 (RGS17), basonuclin 2 (BNC2), *neurocalcin delta (NCALD)*

mir-34c-5p: family with sequence similarity 76 member A (FAM76A), delta like canonical Notch ligand 1 (DLL1), MDM4, p53 regulator (MDM4), *neuron navigator 1 (NAV1)*, *neuron navigator 3 (NAV3)*, *microtubule associated protein tau (MAPT)*

mir-10b: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6), *brain derived neurotrophic factor (BNDF)*

mir-186: RUN and FYVE domain containing 3 (RUFY3), zinc finger CCCH-type containing 11A (ZC3H11A), zinc finger protein 644 (ZNF644), *neuronal growth regulator 1 (NEGR1)*

mir-144-5p: zinc finger protein 292 (ZNF292), ATPase H+ transporting V1 subunit C1 (ATP6V1C1), HIC ZBTB transcriptional repressor 1 (HIC1), *neuregulin 3 (NRG3)*

We found that the overlapping miRNA that are differentially expressed are all up-regulated, meaning that they restrict the amounts of the target proteins listed above. These overlapping miRNA are the central focus of our research project. We have discussed the importance of many of these genes in the previous sections for the individual conditions and data sources, and the neural functions that they contribute are all affected in the development of both Alzheimer's and Parkinson's diseases. Specifically for Alzheimer's, for example, clusterin helps to regulate the processing of amyloid-beta structures, which make up plaques in Alzheimer's brains. Another Alzheimer's-focused gene is TREM1, which encodes receptors on myeloid cells. These receptors are responsible for controlling brain inflammation, which is a tell-tale sign of Alzheimer's. Specifically for Parkinson's, LRRK1 plays a big role in biological processes that require inter-protein interaction, like the transmitting of signals between neurons - the repression of this gene can play a big factor in the movement problems of Parkinson's patients. What is most interesting, however, are the overlapping miRNA that seem to affect both Alzheimer's and Parkinson's disease patients, particularly their brains. The decrease in the amounts of the following proteins can affect the brain function of those affected. For example, neurofilament medium encodes neurofilament, the protein that is used by cells to mark neurons that are damaged. Neuroplastin is a protein that is important in neuron and synaptic functions. Neurocalcin delta has been shown to protect against spinal muscular atrophy, a symptom tangentially related to the symptoms of Parkinson's. It has also been linked with the creation of neurons in the hippocampus of adults, which is the part of the brain that is responsible for memory. This, of course, is negatively affected in those with Alzheimer's. The BDNF protein is largely

responsible for promoting the growth of and dealing with the maintenance of nerve cells. Finally, neuregulin 3 (NRG3), which is a group of signaling proteins that helps to oversee cellular functions of neuronal systems, like survival, proliferation, and differentiation of nerve cells.

Conclusion

Our goal for this study was to find genetic overlapping in Alzheimer's and Parkinson's in order to guide future research with key miRNA that are present in both diseases. We identified 13 up and down-regulated miRNAs in the CSF of Alzheimer's patients, 10 up and down-regulated miRNAs in the CSF of Parkinson's patients, 14 up and down-regulated miRNAs in the serum of Alzheimer's patients, and 22 up and down-regulated miRNAs in the serum of Parkinson's patients. From those, we identified 13 miRNAs that were shared between the diseases and between the biofluids. We mapped all the up-regulated, down-regulated and overlapping miRNAs to the top 3 target mRNAs that they are binding to (ranked in terms of the mRNA's (gene's) "Target Rank" and "Target Score"). As stated, however, these top 3 mRNAs are not necessarily relevant to our studies of brain disorders, so we also identified mRNAs that were tangentially related to neural functions. With careful analysis we discovered that the overlapping miRNA's were all up-regulated, meaning the miRNAs are restricting the amounts of target proteins that we found to be produced at lower amounts.

The important target proteins we have found in Alzheimer's are:

- clusterin (CLU) [26]
- triggering receptor expressed on myeloid cells 1 (TREM1) [34]
- microtubule associated protein tau (MAPT) [28]

These have all been linked as key to Alzheimer's Disease by previous studies.

The important target protein we have found in Parkinson's is:

- leucine rich repeat kinase 1 (LRRK1) [31]

This has been linked as key to Parkinson's Disease by previous studies.

However, the most significant proteins we have found are listed below, these have not to our knowledge been identified yet in other studies. These are common to both Alzheimer's and Parkinson's diseases:

- neurofilament medium (NEFM)
- neuroplastin (NPTN)
- neurocalcin delta (NCALD)
- brain derived neurotrophic factor (BNDF)
- neuregulin (NRG3)

We hope that these significant proteins that we found as genetically overlapped in the cerebrospinal and serum biofluid regions will help future researchers and scientists to gain a better understanding at how these two diseases are linked and that future progress can be made in order to target these proteins to inhibit or lessen the effects of both Alzheimer's and Parkinson's Diseases.

Appendix

Project Targets

Running the project

• To install the dependencies, run the following command from the root directory of the project:

```
pip install -r requirements.txt
```

target: data

• To process the data, from the root project directory run the command:

```
python3 run.py data
```

• The data pipeline step takes the .fastq compressed files as input and then applies two transformations: process and align

• This pipeline step also uses an additional CSV file that is the SRA run database, a sample looks like as follows:

```
Run expired_age    CONDITION    BIOFLUID
SRR1568567  40  Parkinson's Disease Cerebrospinal
```

• The configuration files for the data step are stored in config/data-params.json. These include the parameters for the tools as well as the directories used for storing the raw, temporary and output files.

```
"raw_data_directory": "./data/raw",
"tmp_data_directory": "./data/tmp",
"out_data_directory": "./data/out",
```

• The configuration also includes an attribute to the SRA run input database (described above), and an attribute of where to store that in the data folder. Additional filter attributes are included for ease of use to avoid processing all patients, if this filter_enable is set it will only process a subset of SRA rows (filter_start_row to filter_start_row + filter_num_rows).

```
"sra_runs" : {
    "input_database" : "/datasets/SRP046292/exRNA_Atlas_CORE_Results.csv",
    "input_database2" : "/datasets/SRP046292/SraRunTable.csv",
    "input_database3" : "/datasets/SRP046292/Table_S1.csv",
    "output_database" : "data/raw/exRNA_Atlas_CORE_Results.csv",
    "filter_enable" : 0,
    "filter_start_row" : 120,
    "filter_num_rows" : 10
},
```

• An optional transformation of the data is "process" that uses the following data configuration below that will invoke cutadapt which finds and remove adapter sequences. The attributes include the adapters (r1 and r2) to identify the start and end of pairs are a JSON array. The attribute enable allows to disable this cleaning step, instead it will simply copy the paired files from the source dataset. The arguments attribute allows flexible setting of any additional attribute to the cutadapt process. Finally, we have two wildcard paths that indicate the location of the SRA fastq pair files (fastq1 and fastq2).

```
"process" : {
    "enable" : 1,
    "tool" : "/opt/conda/bin/cutadapt",
    "r1_adapters" : ["AAAAA", "GGGG"],
    "r2_adapters" : ["CCCCC", "TTTT"],
    "arguments" : "--pair-adapters --cores=4",
    "fastq1_path" : "/datasets/srp073813/%run_1.fastq.gz",
    "fastq2_path" : "/datasets/srp073813/%run_2.fastq.gz"
},
```

• The second transformation of the data is "aligncount" that can be set to either use download, STAR or Kallisto. The choice is controlled by the aligncount attribute:

```
"aligncount" : "download",
```

• download step will use the ftp location of the gzip file in the Sra table and download using the curl command and unzips and extracts the readCounts_gencode_sense.txt which represents the gene counts for the sample.

```
"download" : {
    "enable" : 1,
    "tool" : "curl",
    "arguments" : "-L -R",
    "read_counts_file" : "readCounts_gencode_sense.txt"
},
```

• kallisto uses the index_file attribute, the location of the directory of the reference genome, which for this replication project was GRCh37_E75. The arguments attribute allows flexible setting of any additional attribute to the kallisto process. Including the bootstrap samples.The attribute enable allows to disable this alignment

step, this is useful for debugging the process prior step, for example, you can run quality checks on the processed fastq files before proceeding to alignment.

```
"kallisto" : {
    "enable" : 1,
    "tool" : "/opt/kallisto_linux-v0.42.4/kallisto",
    "index_file" : "/datasets/srp073813/reference/kallisto_transcripts.idx",
    "arguments" : "quant -b 8 -t 8"
},
```

• STAR uses the gene_path attribute is the location of the directory of the reference genome, which for this replication project was GRCh37_E75 as described in the reference_gene attribute. The arguments attribute allows flexible setting of any additional attribute to the STAR process. Including TranscriptomeSAM in the quantMode arguments will also output bam files. Additionally, the log file gets outputted which has PRUA (percentage of reads uniquely aligned). The attribute enable allows to disable this alignment step, this is useful for debugging the process prior step, for example, you can run quality checks on the processed fastq files before proceeding to alignment.

```
"STAR" : {
    "enable" : 1,
    "tool" : "/opt/STAR-2.5.2b/bin/Linux_x86_64_static/STAR",
    "reference_gene" : "GRCh37_E75",
    "gene_path" : "/path/to/genomeDir",
    "arguments" : "--runMode alignReads --quantMode GeneCounts --genomeLoad
 LoadAndKeep --readFilesCommand zcat --runThreadN 8"
},
```

• The process and align transformation work on each of the samples. After each sample iteration, the temporary fastq files will be deleted to reduce storage requirements.

• Example processing:

```
python3 run.py data

# -----------------------------------------------------
# Process
# -----------------------------------------------------
# -----------------------------------------------------
# Starting sample # 1 out of 1
# -----------------------------------------------------
# Starting sample # 1 out of 343
curl-proxy -L -R -o ./data/tmp/SRR1568613.tgz ftp://ftp.genboree.org/exRNA-a
tlas/grp/Extracellular%20RNA%20Atlas/db/exRNA%20Repository%20-%20hg19/file/e
xRNA-atlas/exceRptPipeline_v4.6.2/KJENS1-Alzheimers_Parkinsons-2016-10-17/sa
mple_SAMPLE_1022_CONTROL_SER_fastq/CORE_RESULTS/sample_SAMPLE_1022_CONTROL_S
ER_fastq_KJENS1-Alzheimers_Parkinsons-2016-10-17_CORE_RESULTS_v4.6.2.tgz
sh: curl-proxy: command not found
mkdir ./data/tmp/SRR1568613
tar -C ./data/tmp/SRR1568613 -xzf ./data/tmp/SRR1568613.tgz
cp ./data/tmp/SRR1568613/data/readCounts_gencode_sense.txt ./data/tmp/SRR156
8613_ReadsPerGene.out.tab
# -----------------------------------------------------
# Starting sample # 2 out of 343
curl-proxy -L -R -o ./data/tmp/SRR1568457.tgz ftp://ftp.genboree.org/exRNA-a
tlas/grp/Extracellular%20RNA%20Atlas/db/exRNA%20Repository%20-%20hg19/file/e
xRNA-atlas/exceRptPipeline_v4.6.2/KJENS1-Alzheimers_Parkinsons-2016-10-17/sa
mple_SAMPLE_0427_PD_CSF_fastq/CORE_RESULTS/sample_SAMPLE_0427_PD_CSF_fastq_K
JENS1-Alzheimers_Parkinsons-2016-10-17_CORE_RESULTS_v4.6.2.tgz
sh: curl-proxy: command not found
mkdir ./data/tmp/SRR1568457
tar -C ./data/tmp/SRR1568457 -xzf ./data/tmp/SRR1568457.tgz
cp ./data/tmp/SRR1568457/data/readCounts_gencode_sense.txt ./data/tmp/SRR156
8457_ReadsPerGene.out.tab
# -----------------------------------------------------
```

target: merge

• To merge gene count and/or BAM files generated from the data target, from the root project directory run the command:

```
python3 run.py merge
```

• The configuration files for the data step are stored in config/count-params.json. These include the parameters for the count merge and bam merge and it's associated arguments.

• The format attribute informs if to process download, kallisto (or STAR) files. The gene counts are merged into a TSV file and as well as a feature table based on the SRA run table. Additional STAR attributes in the JSON allow you to specify skiprows used when processing the gene count files as well as identifying the column from the

gene matrix file to use as the column used to. There is an additional imputes attribute that allows you to impute any column with missing data. The attributes also include an optional "filter_names" gene table used to remove genes as well as removing false-positive genes. Finally, we can rename the feature columns before we save out the feature table.

```
"count" : {
    "enable" : 1,
    "format" : "download",
    "skiprows" : 4,
    "column_count" : 1,
    "skip_samples" : ["SRR1568391"],
    "enable_filter" : 0,
    "filter_keep_genes" : "NM_",
    "filter_remove_genes" : ["chrX", "chrY"],
    "filter_names" : "/datasets/srp073813/reference/Gene_Naming.csv",
    "run_database" : "data/raw/exRNA_Atlas_CORE_Results.csv",
    "imputes" : ["TangleTotal"],
    "features" : ["Run", "CONDITION", "expired_age", "BIOFLUID", "sex", "PM
I", "sn_depigmentation", "Braak score", "TangleTotal", "Plaque density", "Pl
aqueTotal"],
    "rename" : {"CONDITION" : "Disorder", "BIOFLUID" : "Biofluid", "Braak sc
ore" : "Braak_Score", "Plaque density" : "Plaque_density"},
    "replace" : {"from":["Parkinson's Disease", "Alzheimer's Disease", "Cere
brospinal fluid", "Healthy Control"], "to":["Parkinson", "Alzheimer", "Cereb
rospinal", "Control"]},
    "output_matrix" : "data/out/gene_matrix.tsv",
    "output_features" : "data/out/features.tsv"
},
```

• For bam merging, which should not be enabled by default, we use the "samtools" merge feature that takes all the BAM files and combines them into one merged BAM file.

```
"bam" : {
    "enable" : 0,
    "output" : "data/tmp/merged.bam",
    "tool" : "/usr/local/bin/samtools",
    "arguments" : "merge --threads 8"
},
```

• Example processing:

```
python3 run.py merge

# ------------------------------------------------------
# Merge
Input: SRR3438605_ReadsPerGene.out.tab
Input: SRR3438604_ReadsPerGene.out.tab
Output: data/out/gene_matrix.tsv data/out/features.tsv
# Finished
# ------------------------------------------------------
```

target: normalize

• To normalize the aligned merge counts, from the root project directory run the command:

```
python3 run.py normalize
```

• The configuration files for the data step are stored in config/normalize-params.json.

• We use a custom R script which uses the DESeq2 module to take the input merged gene counts and the experiment features and outputs two normalized counts files. The analysis is done for all samples in the SRA run table. The output_dir sets the output location for the normalized count matrix files. One file is the standard normalized counts using the DESeq2 module, and the second normalized count file is after a Variable Stabilization Transform (LRT). We also have a "max_genes" attribute that will filter the genes and removes ones that have little to no variance across disorder versus control.

• The data JSON configuration file also holds an array of samples, a sample looks like as follows:

```
{
    "output_dir" : "data/out",
    "DESeq2" : {
        "Rscript" : "/opt/conda/envs/r-bio/bin/Rscript",
        "source" : "src/data/normalize.r",
        "input_counts" : "data/out/gene_matrix.tsv",
        "input_features" : "data/out/features.tsv",
        "max_genes" : 8000
    },
    "cleanup" : 0,
    "verbose": 1
}
```

• Example processing:

```
python3 run.py normalize

# -----------------------------------------------------
# Normalize
Rscript  src/data/normalize.r data/out/gene_matrix.tsv data/out/features.tsv
data/out/
[1] "Output data/out/normalized_counts.tsv data/out/vst_transformed_counts.t
sv"
# Finished
# -----------------------------------------------------
```

target: analysis

• To perform the analysis for the gene counts, from the root project directory run the command:

```
python3 run.py analysis
```

• The configuration files for the data step are stored in config/analysis-params.json.

• We use a custom R script which uses the DESeq2 module to take the input merged gene counts and the experiment features and outputs 2 sets of files for each biofluid region. Each biofluid region will compare a disorder versus Control. This will result in a total of 4 sets of files (2 biofluid regions x 2 disorder pair comparisons). Each output set includes a Likelihood Ratio Test (LRT) using the full and reduced model as specified in the attributes below as well as a MA-Plot and Heatmap. The additional attributes include the property of doing parallel processing for DESeq2.

```
{
    "output_prefix" : "data/out/%biofluid_region%",
    "DESeq2" : {
        "Rscript" : "/opt/conda/envs/r-bio/bin/Rscript",
        "biofluid_regions" : ["Cerebrospinal", "Serum"],
        "disorders" : ["Parkinson", "Alzheimer"],
        "control" : "Control",
        "input_counts" : "data/out/pca_normalized_counts.tsv",
        "input_features" : "data/out/features.tsv",
        "source" : "src/analysis/analysis.r",
        "full" : "expired_age+sex+PMI+sn_depigmentation+Braak_Score+TangleTo
tal+Plaque_density+PlaqueTotal+Disorder",
        "reduced" : "expired_age+sex+PMI+sn_depigmentation+Braak_Score+Tangl
eTotal+Plaque_density+PlaqueTotal",
        "parallel" : 0
    },
    "cleanup" : 0,
    "verbose": 1
}
```

• Example processing:

```
python3 run.py analysis

# ------------------------------------------------
# Analysis
Cerebrospinal x Parkinson vs Control
Rscript src/analysis/analysis.r data/out/Cerebrospinal/Parkinson/gene_matri
x.tsv data/out/Cerebrospinal/Parkinson/features.tsv data/out/Cerebrospinal/P
arkinson/ full=expired_age+sex+PMI+sn_depigmentation+Braak_Score+TangleTotal
+Plaque_density+PlaqueTotal+Disorder reduced=expired_age+sex+PMI+sn_depigmen
tation+Braak_Score+TangleTotal+Plaque_density+PlaqueTotal charts=1 parallel=
0
```

target: visualize

• The visualize pipeline step can be invoked as follows:

```
python3 run.py visualize
```

• The configuration files for the data step are stored in config/visualize-params.json. The output will include multiple sets of charts: Gene Spread Variance Histogram, SRA Linear Correlation between SRA chart, MA-Plot 2x2 chart, Heat Map 2x2 chart, 2x2 Histogram, 4x4 Correlation Matrix and a Disorder Venn Diagram. Each chart type has flexible settings to control the input and layout for the charts as shown below:

```
"gene_hist" : {
    "enable" : 1,
    "max_genes" : 8000,
    "nbins" : 100,
    "title" : "Distribution of Genes Based on Spread Metric: All vs Top Gene
s"
},
"missing_plot" : {
    "enable" : 1,
    "title" : "Percentage of Missing Genes over"
},
"sra_lm" : {
    "enable" : 1,
    "sra" : ["SRR1568567", "SRR1568584"],
    "normalized_counts" : "data/out/normalized_counts.tsv",
    "vst_counts" : "data/out/vst_transformed_counts.tsv",
    "title" : "%sra% Regression Log(Norm) v VST counts"
},
"ma_plot" : {
    "enable" : 1,
    "biofluid_regions" : ["Cerebrospinal", "Serum"],
    "disorders" : ["Parkinson", "Alzheimer"],
    "src_image" : "MAplot.png",
    "title" : "MA Plot: Biofluid Region vs Disorder"
},
"heat_map" : {
    "enable" : 1,
    "biofluid_regions" : ["Cerebrospinal", "Serum"],
    "disorders" : ["Parkinson", "Alzheimer"],
    "src_image" : "heatmap.png",
    "title" : "Heat Map: Biofluid Region vs Disorder"
},
"histogram" : {
    "enable" : 1,
    "biofluid_regions" : ["Cerebrospinal", "Serum"],
    "disorders" : ["Parkinson", "Alzheimer"],
    "title" : "Histograms Differential Gene Expression vs Control",
    "ylim" : 55
},
"corrmatrix" : {
    "enable" : 1,
    "title" : "Spearman Correlations of log2 fold gene expression"
},
"venn" : {
    "enable" : 1,
    "biofluid_regions" : ["Cerebrospinal", "Serum"],
    "disorders" : ["Parkinson", "Alzheimer"],
    "pvalue_cutoff" : 0.05,
    "title" : "Venn Diagram Disorders"
```

```
},
```

• Example processing:

```
python3 run.py visualize

# ------------------------------------------------
# Visualize
# Finished
# ------------------------------------------------
```

target: qc

• The quality pipeline step can be invoked as follows:

```
python3 run.py qc
```

• The configuration files for the data step are stored in config/qc-params.json. These include the parameters for the output directory where the quality HTML reports will be outputted.

```
"outdir" : "data/out",
"inputs" : "data/tmp",
```

• For fastq files, the quality tool attribute is set to fastqc and that includes attributes to extract reports or keep them in a zip file. To enable this quality check make sure you set the cleanup to 0 in the data configuration pipeline as well as to disable the STAR processing, this will retain the fastq.qz files after the data pipeline step is executed.

```
"fastq" : {
    "enable" : 1,
    "tool" : "/opt/FastQC/fastqc",
    "extract" : 1
},
```

• For bam files, the quality tool attribute is set to picard and that includes attributes such as collecting alignment summary metrics. To enable this quality check make sure you set the cleanup to 0 in the data configuration pipeline and add 'TranscriptomeSAM' to the arguments for STAR which will then output BAM files that will be retained after the data pipeline step is executed.

```
"bam" : {
    "enable" : 1,
    "tool" : "java",
    "jar" : "/opt/picard-tools-1.88/CollectAlignmentSummaryMetrics.jar"
},
```

• Example processing:

```
python3 run.py qc

# -----------------------------------------------------
# Quality Check
fastqc data/tmp/out.1.fastq.gz --outdir=data/out --extract
fastqc data/tmp/out.2.fastq.gz --outdir=data/out --extract
java -jar /opt/picard-tools-1.88/CollectAlignmentSummaryMetrics.jar INPUT=da
ta/tmp/SRR3438604_Aligned.bam OUTPUT=data/out/SRR3438604_Aligned.bam.txt
java -jar /opt/picard-tools-1.88/CollectAlignmentSummaryMetrics.jar INPUT=da
ta/tmp/SRR3438605_Aligned.bam OUTPUT=data/out/SRR3438605_Aligned.bam.txt
# Finished
# -----------------------------------------------------
```

target: report

• To generate the report from the notebook, run this command:

```
python3 run.py report
```

• The configuration files for the data step are stored in config/report-params.json.

```
{
    "tool": "jupyter",
    "args": "nbconvert --no-input --to html --output report.html notebooks/r
eport.ipynb",
    "verbose" : 1
}
```

target: clean

• To clean the data (remove it from the working project), from the root project directory run the command:

python3 run.py clean

target: all

• The all target will execute the following steps in sequence: data, merge, normalize, analysis and visualize. It can be executed as follows:

python3 run.py all

Appendix

Additional EDA Analysis

Basic numerical features broken down by gender and biofluids

	expired_age	DiseaseDuration	PMI	PlaqueTotal	TangleTotal
count	343.000	218.000	343.000	343.000	341.000
mean	81.006	9.477	3.220	7.947	6.628
std	8.190	6.876	1.559	5.531	4.795
min	38.000	0.000	1.160	0.000	0.000
25%	76.000	5.000	2.415	1.500	3.000
50%	82.000	8.000	2.830	9.000	5.000
75%	87.000	13.000	3.500	13.000	10.500
max	99.000	30.000	12.000	15.000	15.000

Supplementary Table 1: Descriptive statistics

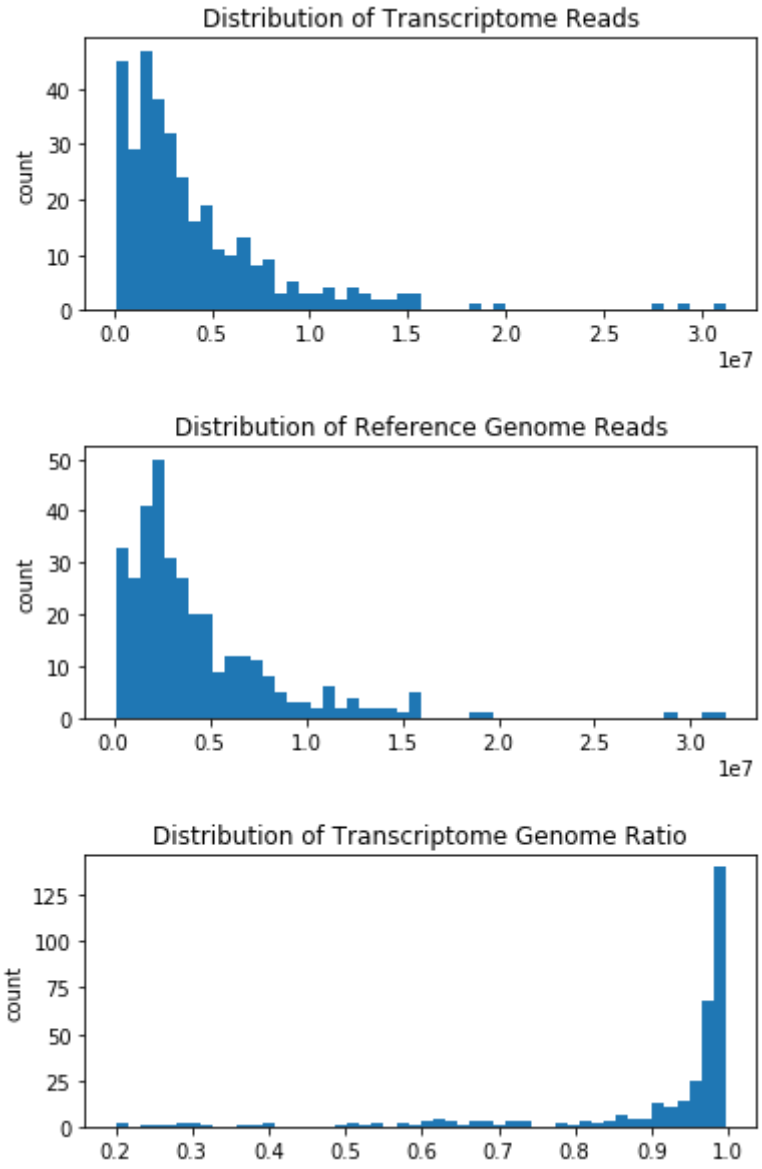
Data Description broken down by Gender and Clinical Diagnosis

	expired_age		DiseaseDuration		PMI		PlaqueTotal		TangleTotal	
sex	female	male	female	male	female	male	female	male	female	male
CONDITION										
Alzheimer's Disease	80.310	81.788	7.946	6.740	2.894	3.051	13.197	12.998	12.469	11.581
Healthy Control	83.690	80.279	1.333	4.857	3.029	2.879	4.846	5.376	3.893	3.748
Parkinson's Disease	81.535	78.984	14.947	11.049	4.221	3.516	6.285	5.746	5.506	3.758

Supplementary Table 2: Feature attributes Grouped By Gender and Condition

Distribution of reads information in SRA_run table

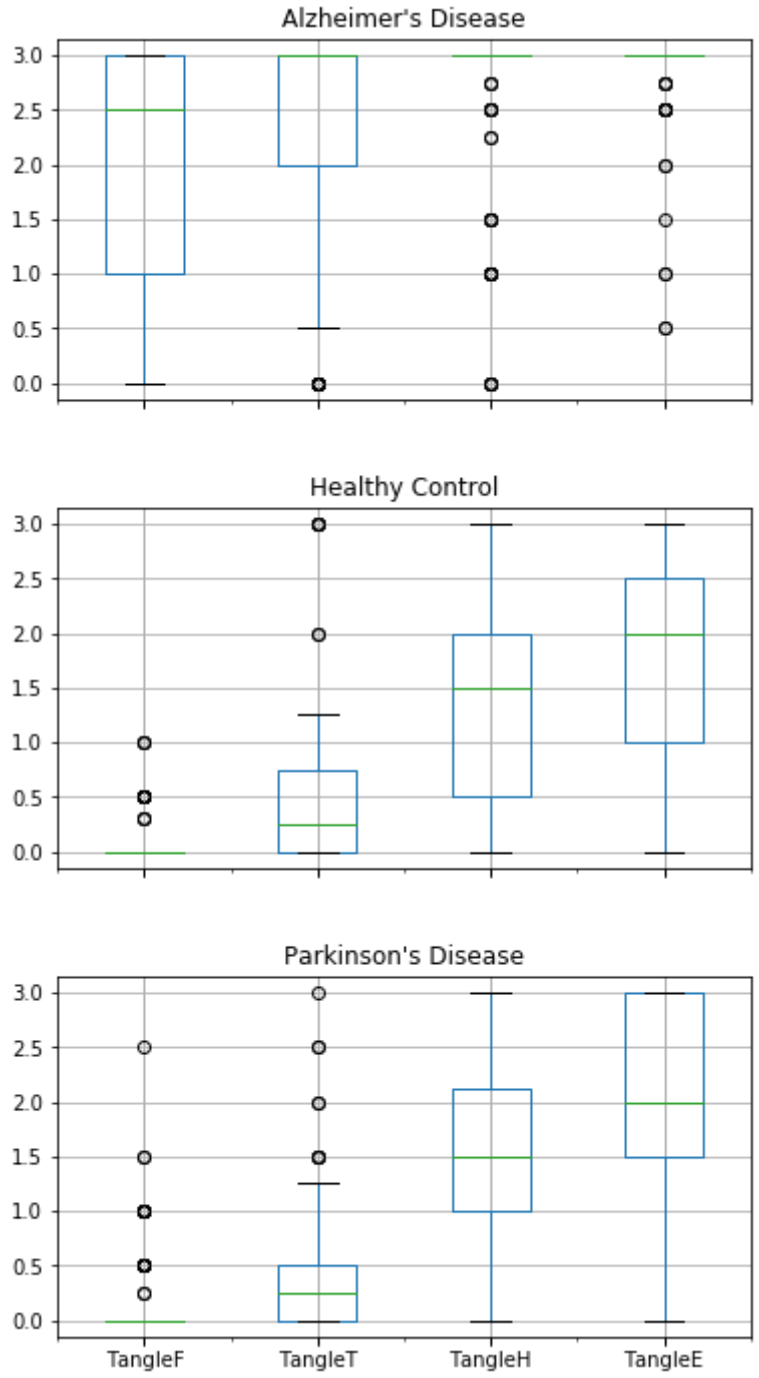
The SRA run table includes the transcriptome reads, reference genome reads, and transcriptome genome ratio for each sample. Here are the distributions of those.



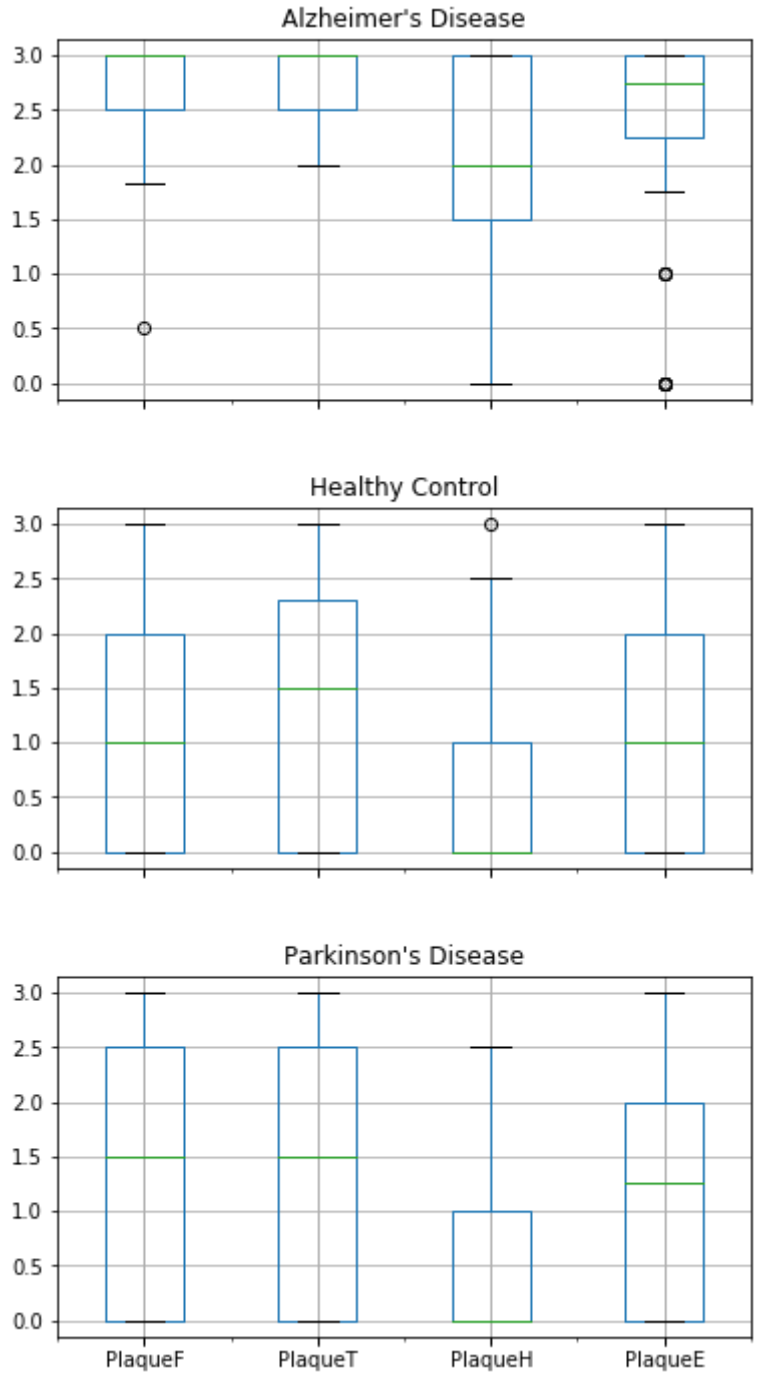
Supplementary Figure 1: Distributions of reads and transcriptome genome ratio.

Tangle & Plaque Counts Distribution in Each Brain Region Broken Down By Conditions

We chose four brain regions(Frontal, Temporal, Hippocampal, and Entorhinal) and plotted the distributions of plaques and tangles of all three groups respectively.



Supplementary Figure 2: Tangle counts in each brain region.

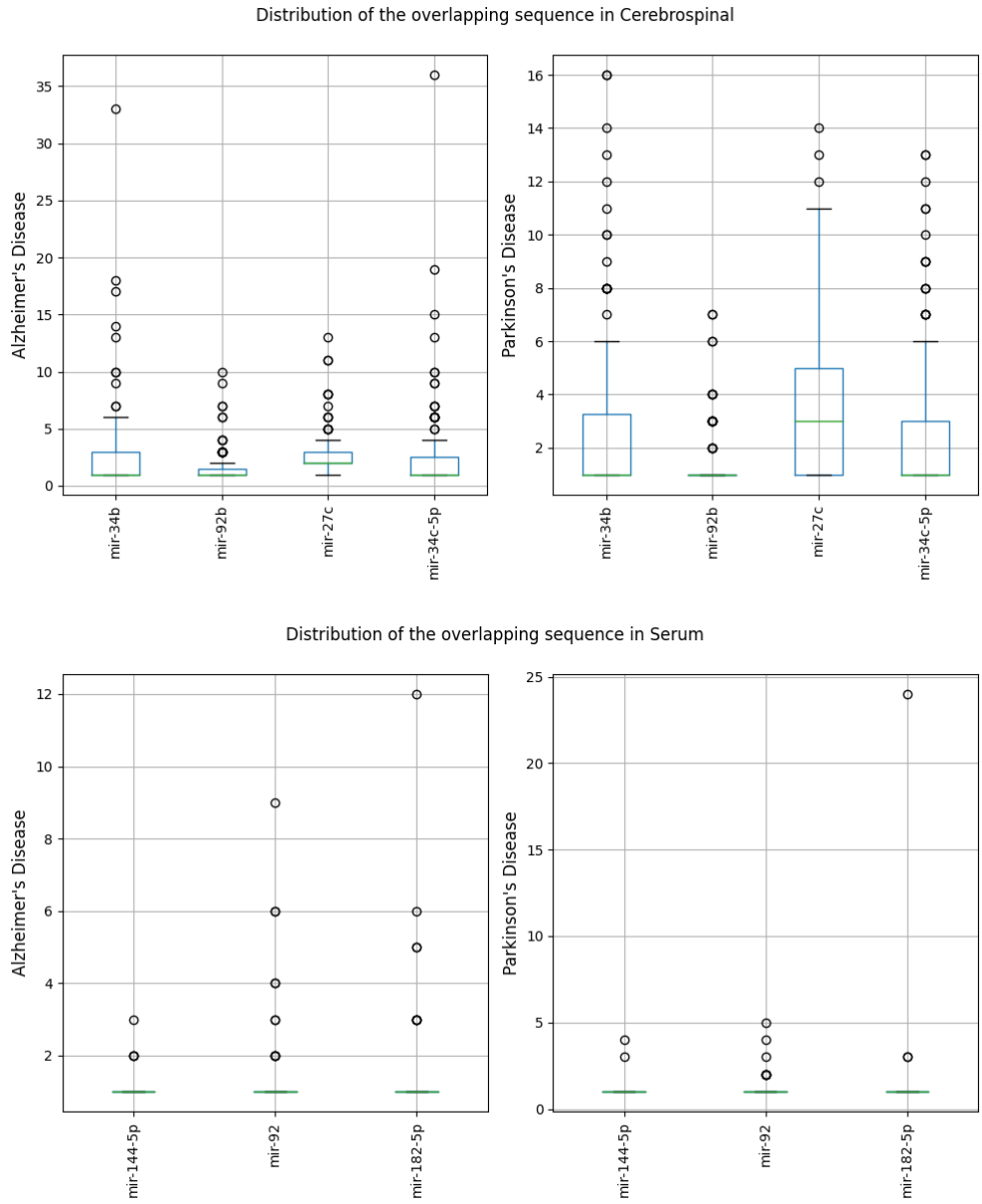


Supplementary Figure 3: Plaque counts in each brain region.

The plots above show that the distribution of tangles and plaques of AD group is different to other groups as expected. However, the difference between PD and healthy control is not significant.

Distributions of the overlapping miRNA Sequences

Below shows the distributions of the overlapping miRNA sequences. It is clear that the distributions of these sequences are significantly different between AD and PD groups in CSF samples.



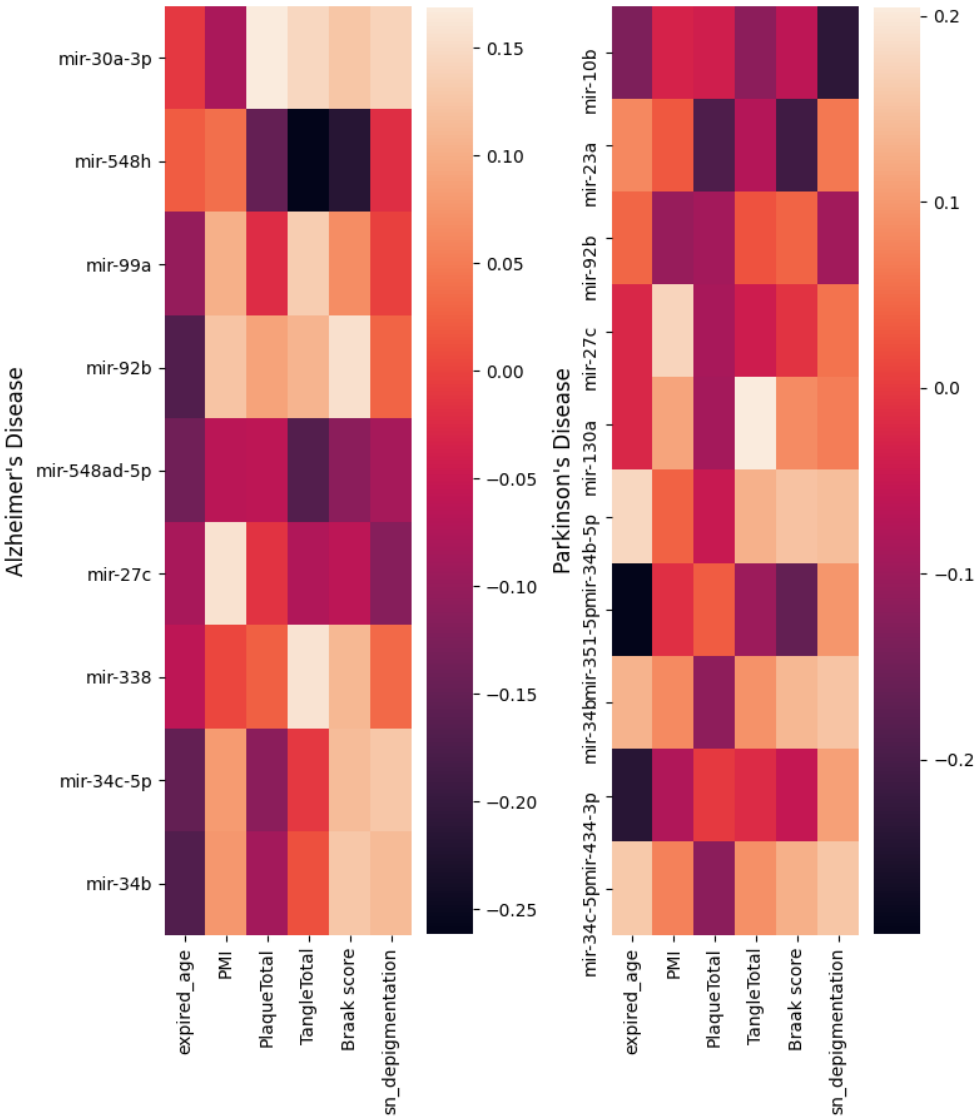
Supplementary Figure 4: Distribution of overlapping miRNA in CSF and Serum.

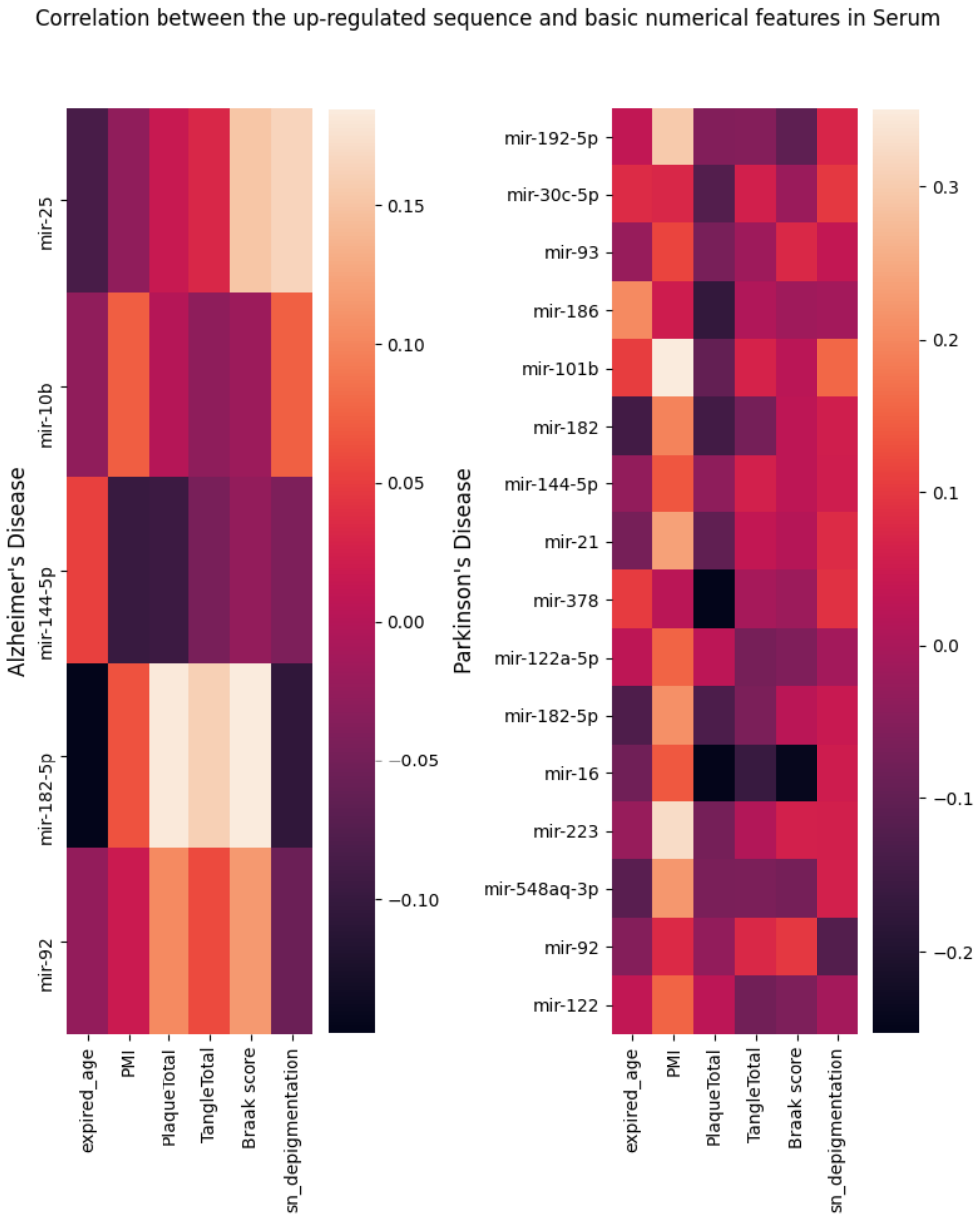
Correlation between significantly regulated miRNA and numerical features

The plots below show the correlation between the significantly regulated miRNA found in the volcano plots above and the basic numerical features we used in the DESeq model.

It appears that most of the up-regulated sequences in Cerebrospinal fluid of the Alzheimer's Disease group are weakly positively correlated with these numerical features, especially with `PlaqueTotal`, `TangleTotal`, `Braak score`, and `sn_depigmentation`. And the sequences in Serum of the Parkinson's Disease group are mostly negatively correlated. However, the correlations with numerical features are not particularly strong.

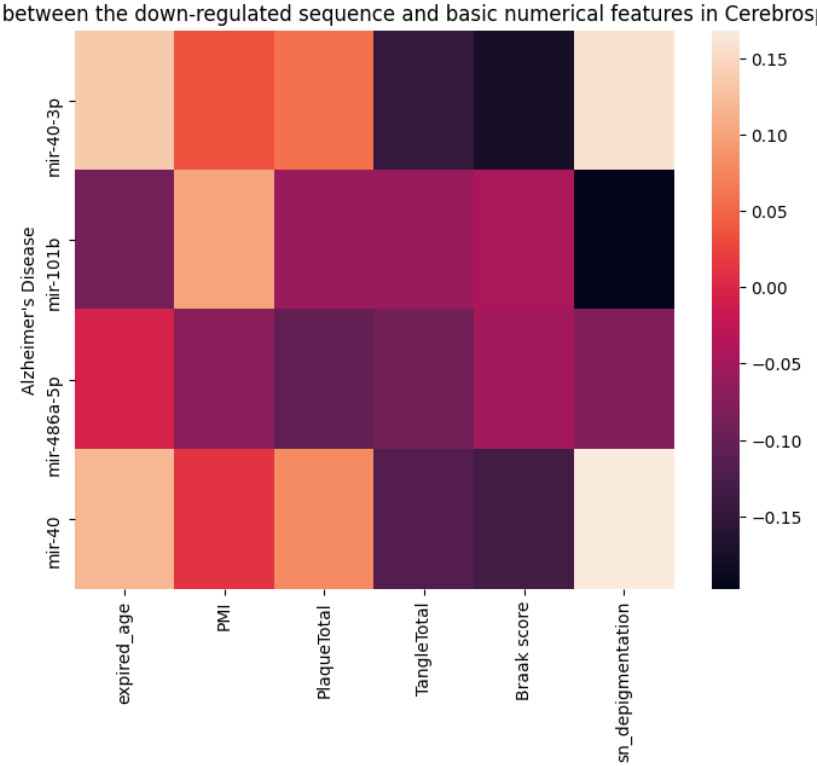
Correlation between the up-regulated sequence and basic numerical features in Cerebrospinal

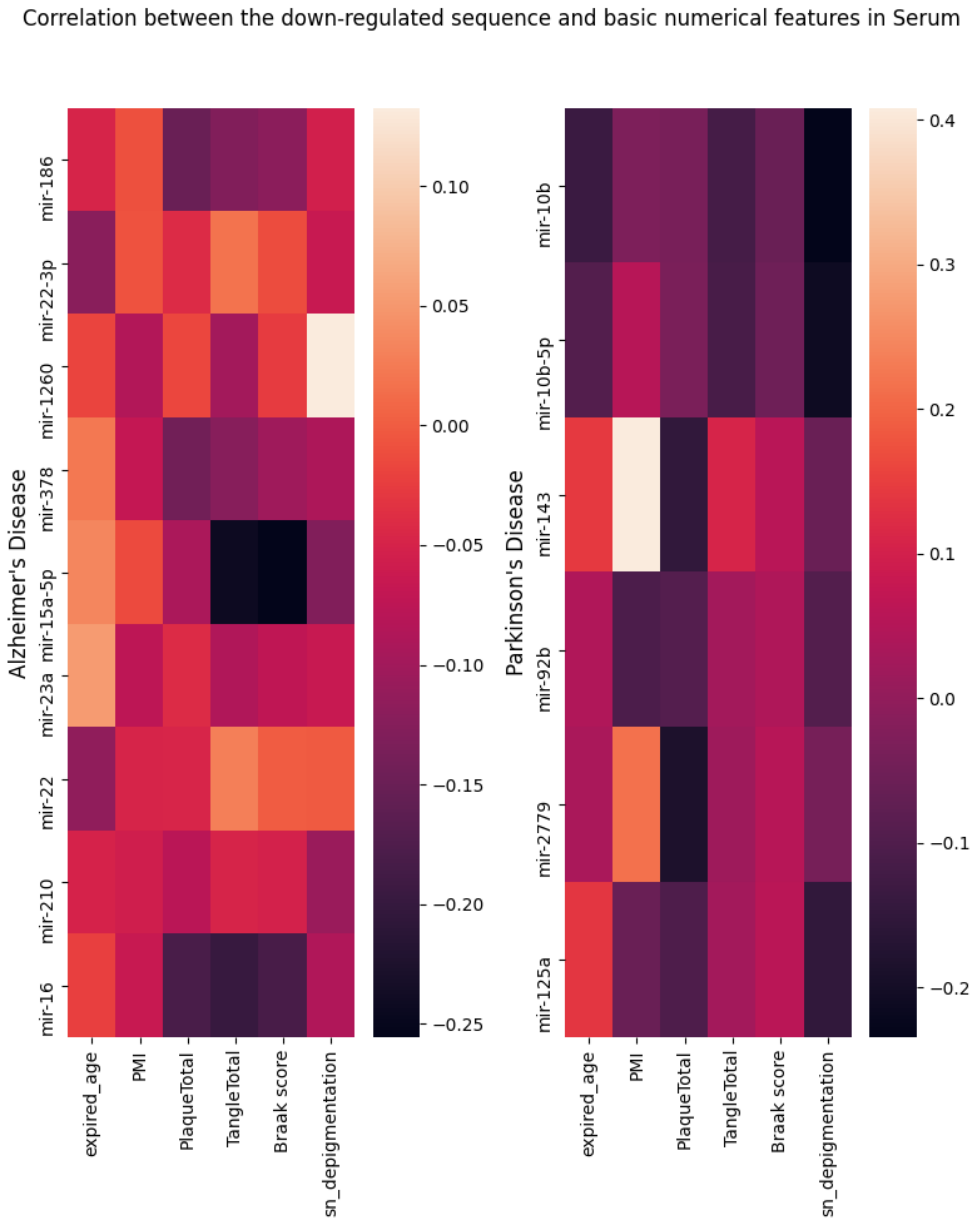




Supplementary Figure 5: Correlation between up-regulated miRNA and numerical features in CSF and Serum.

Correlation between the down-regulated sequence and basic numerical features in Cerebrospinal



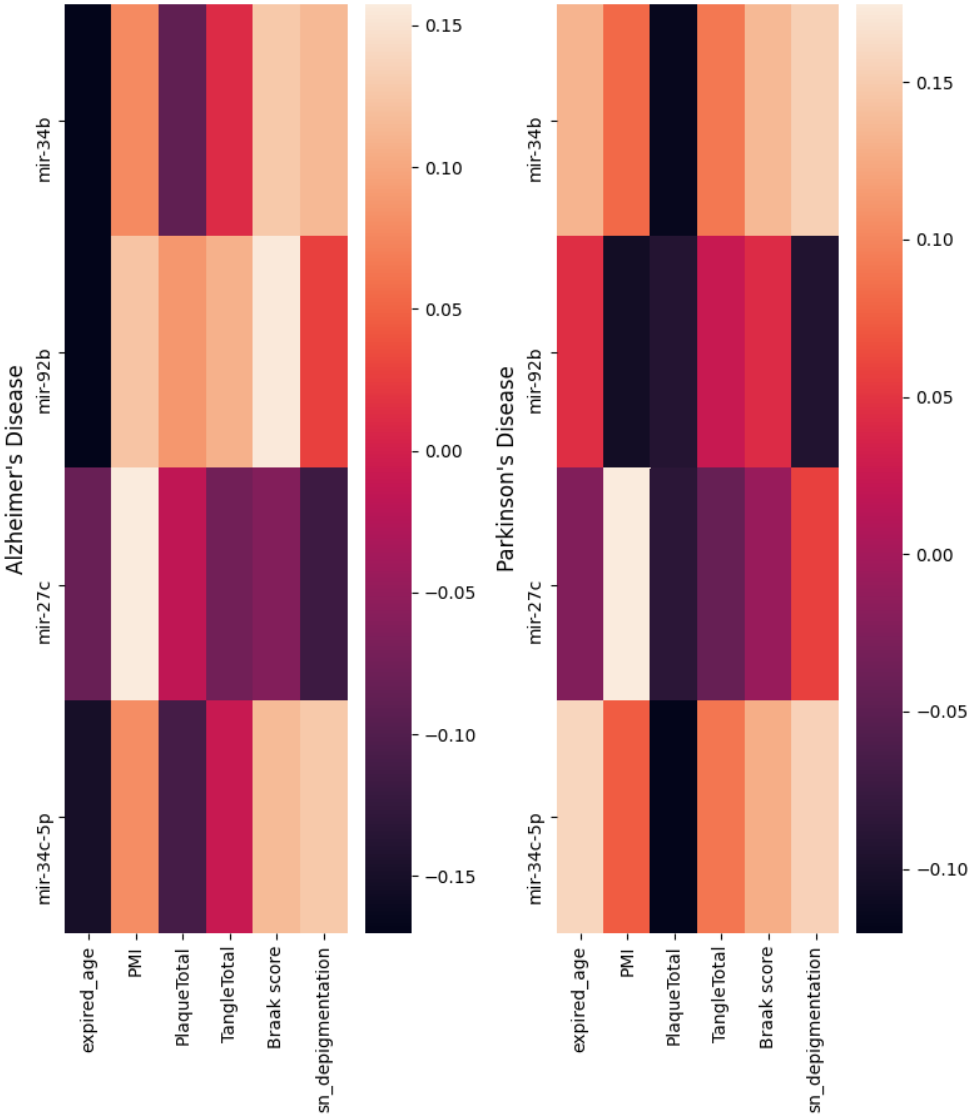


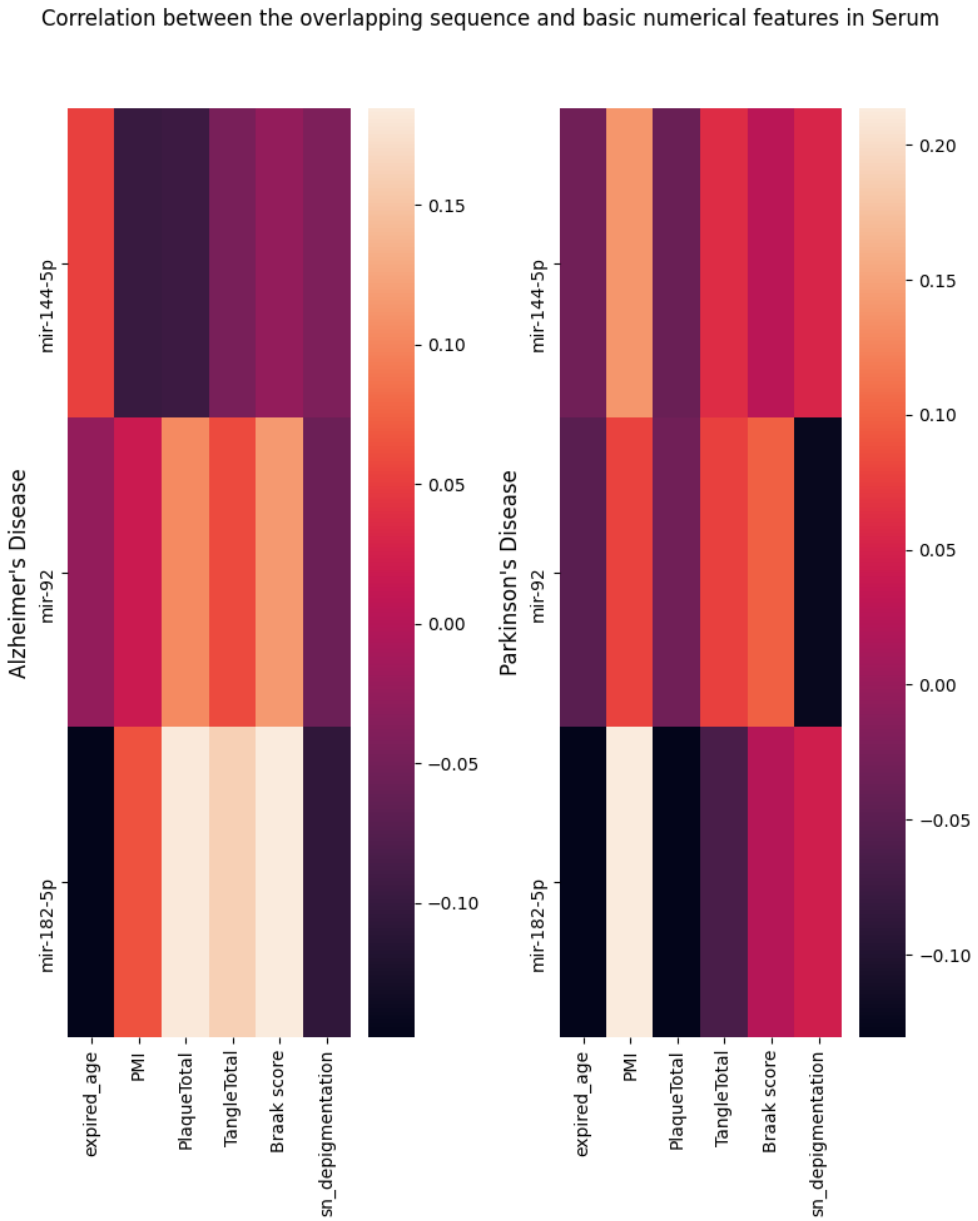
Supplementary Figure 6: Correlation between down-regulated miRNA and numerical features in CSF and Serum.

Correlation between overlapping miRNA and numerical features

The plot below shows the correlation between overlapping miRNA and some basic numerical features. As stated in the report, none of the correlations is particularly strong enough to lead to meaningful conclusions.

Correlation between the overlapping sequence and basic numerical features in Cerebrospinal





Supplementary Figure 7: Correlation between overlapping miRNA and numerical features in CSF and Serum.

Differentially Expressed miRNA in CSF of Alzheimer's Patients + Full mRNA/Protein Mappings

	miRNA	Type	log2FoldChange	-log_pvalue
0	mir-99a	Up	1.068476	2.667435
1	mir-40-3p	Down	-0.529650	1.180250
2	mir-92b	Up	0.592280	1.129815
3	mir-27c	Up	0.530398	1.123541
4	mir-548ad-5p	Up	0.561702	1.001056
5	mir-34b	Up	0.476389	0.742151
6	mir-338	Up	0.455333	0.704723
7	mir-548h	Up	0.413335	0.651008
8	mir-101b	Down	-0.409032	0.649200
9	mir-40	Down	-0.337495	0.647403
10	mir-34c-5p	Up	0.396187	0.579773
11	mir-486a-5p	Down	-0.472859	0.563278
12	mir-30a-3p	Up	0.371358	0.545930

Supplementary Table 3: Cerebrospinal & Alzheimer's Upregulated and Downregulated miRNA's

mir-99a: tribbles pseudokinase 2 (TRIB2), kelch repeat and BTB domain containing 8 (KBTBD8), SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5 (SMARCA5)

mir-40-3p: not found

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*, *phospholipase D1 (PLD1)*, *sortilin related receptor 1 (SORL1)*

mir-27c: not found

mir-548ad-5p: family with sequence similarity 135 member A (FAM135A), nuclear factor of activated T cells 5 (NFAT5), neuronal growth regulator 1 (NEGR1), tau tubulin kinase 2 (TTBK2), *complement C3b/C4b receptor 1 (Knops blood group) (CR1)*

mir-34b: insulin induced gene 1 (INSIG1), protein phosphatase 6 regulatory subunit 3 (PPP6R3), furin, paired basic amino acid cleaving enzyme (FURIN), *neuroplastin (NPTN)*

mir-338: Cbl proto-oncogene (CBL), galectin like (LGALSL), RAB14, member RAS oncogene family (RAB14), neuropilin 1 (NRP1), *phosphatidylinositol binding clathrin assembly protein (PICALM)*

mir-548h: CREB binding protein (CREBBP), ubiquitin conjugating enzyme E2 D1 (UBE2D1), zinc finger CCHC-type containing 14 (ZCCHC14), *neuron navigator 2 (NAV2)*

mir-101b: not found

mir-40: not found

mir-34c-5p: family with sequence similarity 76 member A (FAM76A), delta like canonical Notch ligand 1 (DLL1), MDM4, p53 regulator (MDM4), *neuron navigator 1 (NAV1)*, neuron navigator 3 (NAV3), microtubule associated protein tau (MAPT)

mir-486a-5p: not found

mir-30a-3p: cell division cycle 73 (CDC73), zinc finger E-box binding homeobox 2 (ZEB2), nuclear FMR1 interacting protein 2 (NUFIP2)

Differentially Expressed miRNA in CSF of Parkinson's Patients + Full mRNA/Protein Mappings

	miRNA	Type	log2FoldChange	-log_pvalue
0	mir-34b	Up	0.722977	1.023605
1	mir-34c-5p	Up	0.584796	0.776372
2	mir-27c	Up	0.495344	0.764454
3	mir-434-3p	Up	0.558104	0.760882
4	mir-92b	Up	0.566194	0.755235
5	mir-130a	Up	0.595353	0.708097
6	mir-351-5p	Up	0.542760	0.692046
7	mir-10b	Up	0.515659	0.655791
8	mir-34b-5p	Up	0.475180	0.557489
9	mir-23a	Up	0.446270	0.546405

Supplementary Table 4: Cerebrospinal & Parkinson's Upregulated and Downregulated miRNA's

mir-34b: insulin induced gene 1 (INSIG1), protein phosphatase 6 regulatory subunit 3 (PPP6R3), furin, paired basic amino acid cleaving enzyme (FURIN), *neuroplastin (NPTN)*

mir-34c-5p: family with sequence similarity 76 member A (FAM76A), delta like canonical Notch ligand 1 (DLL1), MDM4, p53 regulator (MDM4), *neuron navigator 1 (NAV1)*

mir-27c: not found

mir-434-3p: not found

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*

mir-130a: gap junction protein alpha 1 (GJA1), cytoplasmic polyadenylation element binding protein 1 (CPEB1), SKI/DACH domain containing 1 (SKIDA1), *leucine rich repeat kinase 2 (LRRK2)*

mir-351-5p: not found

mir-10b: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6)

mir-34b-5p: teneurin transmembrane protein 1 (TENM1), ELMO domain containing 1 (ELMOD1), regulatory factor X3 (RFX3), *parkin RBR E3 ubiquitin protein ligase (PRKN)*

mir-23a: zinc finger protein 99 (ZNF9), semaphorin 6D (SEMA6D), family with sequence similarity 234 member B (FAM234B)

Differentially Expressed miRNA in Serum of Alzheimer's Patients + Full mRNA/Protein Mappings

	miRNA	Type	log2FoldChange	-log_pvalue
0	mir-16	Down	-1.006115	2.005127
1	mir-15a-5p	Down	-0.922835	1.743917
2	mir-378	Down	-0.955715	1.347237
3	mir-182-5p	Up	0.739354	1.097146
4	mir-23a	Down	-0.622807	0.955343
5	mir-186	Down	-0.602617	0.944766
6	mir-92	Up	0.598591	0.902923
7	mir-25	Up	0.929064	0.866180
8	mir-10b	Up	0.534783	0.743404
9	mir-22	Down	-0.491724	0.670908
10	mir-210	Down	-0.482772	0.657003
11	mir-144-5p	Up	0.478381	0.645686
12	mir-22-3p	Down	-0.531189	0.576222
13	mir-1260	Down	-0.365382	0.528260

Supplementary Table 5: Serum & Alzheimer's Upregulated and Downregulated miRNA's

mir-16: pappalysin 1 (PAPPA), fatty acid synthase (FASN), unc-80 homolog, NALCN channel complex subunit (UNC80), *clusterin (CLU)*, *triggering receptor expressed on myeloid cells 1 (TREM1)*, *neurofibromin 1 (NF1)*

mir-15a-5p: pappalysin 1 (PAPPA), fatty acid synthase (FASN), unc-80 homolog, NALCN channel complex subunit (UNC80), *neuritin 1 (NRN1)*, *neuropilin 2 (NRP2)*

mir-378: ubinuclein 2 (UBN2), vestigial like family member 3 (VGLL3), M-phase specific PLK1 interacting protein (MPLKIP)

mir-182-5p: protein kinase cAMP-activated catalytic subunit beta (PRKACB), regulator of G protein signaling 17 (RGS17), basonuclin 2 (BNC2)

mir-23a: zinc finger protein 99 (ZNF99), semaphorin 6D (SEMA6D), family with sequence similarity 234 member B (FAM234B), *neuroligin 4 X-linked (NLGN4X)*

mir-186: RUN and FYVE domain containing 3 (RUFY3), zinc finger CCCH-type containing 11A (ZC3H11A), zinc finger protein 644 (ZNF644), neuronal growth regulator 1 (NEGR1)

mir-92: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *sortilin related receptor 1 (SORL1)*

mir-25: CD69 molecule (CD69), solute carrier family 12 member 5 (SLC12A5), mannosidase alpha class 2A member 1 (MAN2A1)

mir-10b: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6), *brain derived neurotrophic factor (BNDF)*

mir-22: glutamate metabotropic receptor 5 (GRM5), fucosyltransferase 9 (FUT9), *neuroepithelial cell transforming 1 (NET1)*

mir-210: insulin like growth factor 2 (IGF2), iron-sulfur cluster assembly enzyme (ISCU), galanin receptor 2 (GALR2), brain derived neurotrophic factor (BDNF), *neuronal pentraxin 1 (NPTX1)*

mir-144-5p: zinc finger protein 292 (ZNF292), ATPase H+ transporting V1 subunit C1 (ATP6V1C1), HIC ZBTB transcriptional repressor 1 (HIC1), neurotrophic receptor tyrosine kinase 2 (NTRK2), *neuregulin 3 (NRG3)*

mir-22-3p: glutamate metabotropic receptor 5 (GRM5), fucosyltransferase 9 (FUT9), *neuroepithelial cell transforming 1 (NET1)*

mir-1260: zinc finger protein 268 (ZNF268), zinc finger protein 763 (ZNF763), cutaneous T cell lymphoma-associated antigen 1 (CTAGE1), *complement C3b/C4b receptor 1 (Knops blood group) (CR1)*

Differentially Expressed miRNA in Serum of Parkinson's Patients + Full mRNA/Protein Mappings

	miRNA	Type	log2FoldChange	-log_pvalue
0	mir-192-5p	Up	1.987334	2.412864
1	mir-182-5p	Up	1.106859	1.463374
2	mir-93	Up	1.104689	1.425775
3	mir-143	Down	-0.875965	1.291720
4	mir-10b-5p	Down	-0.763093	1.205186
5	mir-144-5p	Up	0.800832	1.057860
6	mir-125a	Down	-0.727192	0.995720
7	mir-182	Up	0.661768	0.810613
8	mir-21	Up	0.616150	0.707095
9	mir-92	Up	0.564453	0.684278
10	mir-92b	Down	-0.534518	0.681643
11	mir-30c-5p	Up	0.590171	0.658381
12	mir-548aq-3p	Up	0.550847	0.651495
13	mir-186	Up	0.574868	0.651354
14	mir-378	Up	0.611943	0.623439
15	mir-16	Up	0.575121	0.617367
16	mir-122a-5p	Up	0.519159	0.600663
17	mir-101b	Up	0.545156	0.594463
18	mir-122	Up	0.632374	0.589976
19	mir-10b	Down	-0.460216	0.582377
20	mir-223	Up	0.521058	0.570865
21	mir-2779	Down	-0.415505	0.544440

Supplementary Table 6: Serum & Parkinson's Upregulated and Downregulated miRNA's

mir-192-5p: NIPA like domain containing 1 (NIPAL1), basic helix-loop-helix family member e22 (BHLHE22), protein kinase D3 (PRKD3), *neurofilament light (NEFL)*

mir-182-5p: protein kinase cAMP-activated catalytic subunit beta (PRKACB), regulator of G protein signaling 17 (RGS17), basonuclin 2 (BNC2), *neurocalcin delta (NCALD)*

mir-93: ectonucleotide pyrophosphatase/phosphodiesterase 5 (putative) (ENPP5), FYVE and coiled-coil domain containing 1 (FYCO1), dynein cytoplasmic 1 light intermediate chain 2 (DYNC1LI2)

mir-143: ABL proto-oncogene 2, non-receptor tyrosine kinase (ABL2), vasohibin 1 (VASH1), DENN domain containing 1B (DENND1B)

mir-10b-5p: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6), brain derived neurotrophic factor (BDNF)

mir-144-5p: zinc finger protein 292 (ZNF292), ATPase H+ transporting V1 subunit C1 (ATP6V1C1), HIC ZBTB transcriptional repressor 1 (HIC1), neuregulin 3 (NRG3)

mir-125a: DTW domain containing 1 (DTWD1), BCL2 family apoptosis regulator BOK (BOK), BRCA1, DNA repair associated (BRCA), *neuronal vesicle trafficking associated 2*

mir-182: protein kinase cAMP-activated catalytic subunit beta (PRKACB), regulator of G protein signaling 17 (RGS17), basonuclin 2 (BNC2)

mir-21: YOD1 deubiquitinase (YOD1), Fas ligand (FASLG), PR/SET domain 11 (PRDM11), *neurotrophin 3* (NTF3)

mir-92: folliculin interacting protein 1 (FNIP1), CD69 molecule (CD69), G3BP stress granule assembly factor 2 (G3BP2), neurofilament medium (NEFM)

mir-92b: beta-1,3-galactosyltransferase 2 (B3GALT2), mannosidase alpha class 2A member 1 (MAN2A1), F-box and WD repeat domain containing 7 (FBXW7), *neurofilament medium (NEFM)*

mir-30c-5p: twinfilin actin binding protein 1 (TWF1), UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 5 (B3GNT5), embryonic ectoderm development (EED), *neural cell adhesion molecule 1 (NCAM1)*, *leucine rich repeat kinase 2 (LRRK2)*

mir-548aq-3p: polyhomeotic homolog 3 (PHC3), CREB3 regulatory factor (CREBRF), protein tyrosine phosphatase, receptor type K (PTPRK), *synuclein alpha (SNCA)*

mir-186: RUN and FYVE domain containing 3 (RUFY3), zinc finger CCCH-type containing 11A (ZC3H11A), zinc finger protein 644 (ZNF644), *neuronal growth regulator 1 (NEGR1)*

mir-378: ubinuclein 2 (UBN2), vestigial like family member 3 (VGLL3), M-phase specific PLK1 interacting protein (MPLKIP)

mir-16: pappalysin 1 (PAPPA), fatty acid synthase (FASN), unc-80 homolog, NALCN channel complex subunit (UNC80), clusterin (CLU), triggering receptor expressed on myeloid cells 1 (TREM1), neurofibromin 1 (NF1)

mir-122a-5p: not found

mir-101b: not found

mir-122: heterogeneous nuclear ribonucleoprotein U (HNRNPU), cytoplasmic polyadenylation element binding protein 1 (CPEB1), CD40 ligand (CD40LG)

mir-10b: cell adhesion molecule 2 (CADM2), transcription factor AP-2 gamma (TFAP2C), CCR4-NOT transcription complex subunit 6 (CNOT6)

mir-223: F-box and WD repeat domain containing 7 (FBXW7), SP3 transcription factor (SP3), *synuclein alpha (SNCA)*, *neuron derived neurotrophic factor (NDNF)*

mir-2779: not found

References

[1] Burgos, Kasandra, et al. “Profiles of Extracellular MiRNA in Cerebrospinal Fluid and Serum from Patients with Alzheimer's and Parkinson's Diseases Correlate with Disease Status and Features of Pathology.” PLOS One, vol. 9, no. 5, 5 May 2014, doi:10.1371/journal.pone.0094839.

[2] Mayo Clinic Staff. “Alzheimer's Disease.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 29 Dec. 2020, www.mayoclinic.org/diseases-conditions/alzheimers-disease/symptoms-causes/syc-20350447.

[3] Mayo Clinic Staff. “Parkinson's Disease.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 8 Dec. 2020, www.mayoclinic.org/diseases-conditions/parkinsons-disease/symptoms-causes/syc-20376055.

[4] Charlotte Hewel, et al. “Common MiRNA Patterns of Alzheimer’s Disease and Parkinson’s Disease and Their Putative Impact on Commensal Gut Microbiota.” Frontiers in Neuroscience, 5 Mar. 2019, doi:10.3389/fnins.2019.00113.

[5] “Cerebrospinal Fluid (CSF).” National Multiple Sclerosis Society, www.nationalmssociety.org/Symptoms-Diagnosis/Diagnosing-Tools/Cerebrospinal-Fluid-(CSF).

[6] Kopkova, Alena. “MicroRNA Isolation and Quantification in Cerebrospinal Fluid: A Comparative Methodical Study.” PLOS One, vol. 13, no. 12, 7 Dec. 2018, doi:10.1371/journal.pone.0208580.

[7] Flournoy, Blake. “What Is Serum?” Sciencing, 21 Jan. 2020, sciencing.com/what-is-serum-4673561.html.

[8] “Analysis of Circulating MiRNA in Plasma or Serum.” 3D-Gene, www.3d-gene.com/en/products/dna/dna_004.html

[9] “Likelihood Ratio Test.” Evolution and Genomics, 3 July 2016, evomics.org/resources/likelihood-ratio-test/.

[10] "pandas.DataFrame.merge", Pandas, https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.merge.html.

[11] I. Lönnstedt, T. Speed, et al. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” Genome Biology, BioMed Central, 1 Jan. 1970, genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8.

[12] Martin, Marcel. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” EMBnet.journal, journal.embnet.org/index.php/embnetjournal/article/view/200/479.

[13] Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data, www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[14] “Data Quality Control Standards.” ExRNA Research Portal, exrna.org/resources/data/data-quality-control-standards/.

[15] Moss, G. P. “Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences.” Nomenclature Committee of the International Union of Biochemistry (NC-IUB), 1984, www.qmul.ac.uk/sbcs/iubmb/misc/naseq.html.

[16] Smith, Yolanda. “Alzheimer’s Tangles and Plaques: What’s the Difference?” Medical News, 26 Feb. 2019, www.news-medical.net/health/Alzheimers-tangles-and-plaques-whats-the-difference.aspx.

[17] Burke, Robert E, et al. “A Critical Evaluation of The Braak Staging Scheme for Parkinson’s Disease.” Annals of Neurology, vol. 64, no. 5, 1 Nov. 2009, doi:10.1002/ana.21541.

[18] “What Is Lewy Body Dementia?” National Institute on Aging, U.S. Department of Health and Human Services, www.nia.nih.gov/health/what-lewy-body-dementia.

[19] Bennett, D A, et al. “Neuropathology of older persons without cognitive impairment from two community-based studies”. Neurology, vol. 66, no. 12, 27 Jun 2006, doi: 10.1212/01.wnl.0000219668.47116.e6.

[20] Hack, Nawaz, et al. “Substantia nigra depigmentation and exposure to encephalitis lethargica.” Annals of Neurology, vol. 76, no. 6, 1 Dec. 2013, doi:10.1002/ana.23697.

[21] Poewe, W., Seppi,K., Tanner, C. et al. "Parkinson disease". Nature Reviews Disease Primers, vol. 3, no. 17013, 23 March 2017, doi: 10.1038/nrdp.2017.13.

[22] “How Does Parkinson's Progress?” Parkinson's Victoria, www.parkinsonsvic.org.au/parkinsons-and-you/how-does-parkinsons-progress/.

[23] “Variance-Stabilizing Transformation.” Wikipedia, Wikimedia Foundation, 12 Dec. 2020, en.wikipedia.org/wiki/Variance-stabilizing_transformation.

[24] “How, When, and Why Should You Normalize / Standardize / Rescale Your Data?” Towards AI - The Best of Tech, Science, and Engineering, 29 May 2020, towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

[25] CB. Caldwell, II. Gottesman, et al. “Post-Mortem Molecular Profiling of Three Psychiatric Disorders.” Genome Medicine, BioMed Central, 1 Jan. 1990, genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0458-5.

[26] Yi, Mike. “A Complete Guide to Heatmaps.” Chartio, 11 Nov. 2019, chartio.com/learn/charts/heatmap-complete-guide/.

[27] "APOE gene." MedlinePlus, https://medlineplus.gov/genetics/gene/apoe/#conditions

[28] “The Role of Genes in Your Alzheimer's Risk.” Mayo Clinic, Mayo Foundation for Medical Education and Research, 19 Apr. 2019, www.mayoclinic.org/diseases-conditions/alzheimers-disease/in-depth/alzheimers-genes/art-20046552.

[29] Beesley, Philip W, et al. “The Neuroplastin adhesion molecules: key regulators of neuronal plasticity and synaptic function”. Journal of Neurochemistry, vol. 131, no. 3, 14 Aug. 2014, doi: 10.1111/jnc.12816.

[30] “Parkinson Disease: MedlinePlus Genetics.” MedlinePlus, U.S. National Library of Medicine, 18 Aug. 2020, medlineplus.gov/genetics/condition/parkinson-disease/#causes.

[31] “LRRK2 Gene: MedlinePlus Genetics.” MedlinePlus, U.S. National Library of Medicine, 18 Aug. 2020, medlineplus.gov/genetics/gene/lrrk2/#conditions.

[32] “PRKN Gene: MedlinePlus Genetics.” MedlinePlus, U.S. National Library of Medicine, 18 Aug. 2020, medlineplus.gov/genetics/gene/prkn/.

[33] “NEFM neurofilament medium.” National Center for Biotechnology Information (NCBI), https://www.ncbi.nlm.nih.gov/gene/4741.

[34] Genua, Marco, et al. “The Triggering Receptor Expressed on Myeloid Cells (TREM) in Inflammatory Bowel Disease Pathogenesis.” Journal of Translational Medicine, vol. 12, no. 1, 28 Oct. 2014, doi:10.1186/s12967-014-0293-z.

[35] “Sortilin-Related Receptor.” UniProt Consortium, 10 Feb. 2021, www.uniprot.org/uniprot/Q92673.

[36] “BDNF gene.” MedlinePlus, https://medlineplus.gov/genetics/gene/bdnf/.

[37] Seroogy, Kim B, et al. “Neuregulins”. Handbook of Biologically Active Peptides (Second Edition), 2013.

[38] Joana, Figueiro-Silva, et al. “Neuronal Pentraxin 1 Negatively Regulates Excitatory Synapse Density and Synaptic Plasticity.” The Journal of Neuroscience, vol. 35, no. 14, 8 Apr. 2015, doi:10.1523/JNEUROSCI.2548-14.2015.

[39] Upadhyay, Aaradhita, et al. “Neurocalcin Delta Knockout Impairs Adult Neurogenesis Whereas Half Reduction Is Not Pathological”. Frontiers in Molecular Neuroscience, vol. 12, no. 19, 12 Feb. 2019, doi: 10.3389/fnmol.2019.00019.

[40] Stefanis, Leonidas. “α-Synuclein in Parkinson's Disease.” Cold Spring Harbor Perspectives in Medicine, vol.