Crafting an RNA-Seq Pipeline for Analysis of Gene Expression Profiles of Parkinsonian Samples

Sabyasachi Samantaray[a], Ankit Halder[b], Deeptarub Biswas[b], and Sanjeeva Srivastava[b]

[a]Department of Computer Science and Engineering, IIT Bombay
[b]Department of Biosciences and Bioengineering, IIT Bombay

ABSTRACT

In this study, we craft a robust pipeline for generating the gene and transcript counts from FASTQ reads of the RNA-Seq samples and Visualise the Gene Expression Profiles using DESeq. We analysed RNA-Seq datasets taken from the NCBI GEO, which include transcriptomic data extracted from regions within the substantia nigra pars compacta (SNpc). The ability to use each of this to predict the outcomes of Parkinson's disease, including disease progression and cognitive and motor complications, would be of significant clinical value. Together, these findings identify molecular signatures in Parkinson's disease patients' brain of potential pathophysiologic and prognostic importance.

1. INTRODUCTION

Parkinson's disease is a progressive neurodegenerative disorder characterized by a range of motor and non-motor symptoms. It stands as the second most prevalent neurodegenerative disorder, following Alzheimer's disease, affecting approximately 2–3% of the population over 65 years of age.[1] PD is a complex brain condition impacting movement, mental health, sleep, pain, and other aspects of well-being. The global prevalence of PD is projected to more than double by 2040.[2] As a condition that worsens over time, PD currently lacks a cure, but available therapies and medications aim to alleviate symptoms. Common manifestations include tremors, painful muscle contractions, and difficulty speaking. Additionally, a significant number of PD patients develop dementia.[3] A deeper understanding of the underlying disease mechanisms is essential for the development of novel therapies. All codes and troubleshooting details can be found on github*.

However, it is now widely acknowledged that Parkinson's disease (PD) extends beyond the substantia nigra, impacting both the central and peripheral nervous systems.[5] Braak et al.[6] have proposed a staging system for PD based on the progression of Lewy body (LB) pathology in the brain. Ranging from 1 to 6, this system indicates a gradual increase in LB presence, following a rostral to caudal pattern. This staging framework is believed to reflect disease progression, correlating with the successive emergence of key clinical symptoms as the pathological process extends to different brain regions.[6]

*Github Link: https://github.com/Sachi-27/RNA-Seq-PD

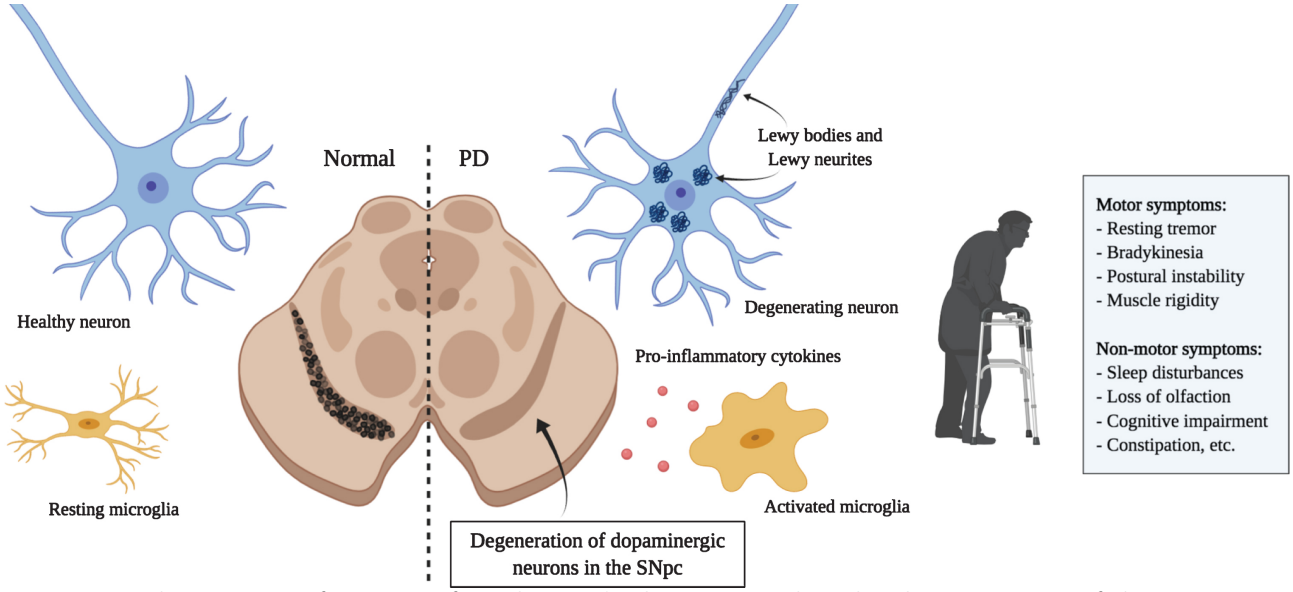


Figure 1: The primary features of Parkinson's disease involve the degeneration of dopaminergic neurons in the substantia nigra pars compacta, resulting in striatal dopamine depletion and the emergence of classical motor symptoms. Afflicted neurons exhibit the presence of insoluble aggregates known as Lewy bodies in their cytoplasm and Lewy neurites in their neurites. These aggregates are primarily constituted of misfolded alpha-synuclein (α-syn).[4]

Various researchers have delved into exploring potential biomarkers and causes of Idiopathic Parkinson's disease. The surge in multi-omics modalities for disease characterization offers complementary perspectives on the disease process. With modern high-throughput omic platforms, biomedical studies increasingly adopt an integrative approach, combining data from genetics, proteomics, and metabolomics. This integrative strategy, facilitated by machine learning-based predictive algorithms, unveils the intricacies of systems biology.[7] Machine learning methods, through data integration and analysis, contribute to biomarker discovery, enabling accurate disease prediction, patient stratification, and precision medicine.[8] Statistical analysis of robust biomarker candidates involves essential steps in the development pipeline.[9] Initially, data visualization aids in identifying outliers and understanding the data nature. Subsequently, data pre-processing addresses outliers, handles missing values, and assesses normality. Once prepared, hypothesis tests reveal differentially expressed genes and proteins.[9]

In this investigation, we examine two RNA sequencing studies conducted on tissue samples from the caudate (focused on stress response) and putamen (emphasizing endothelial pathways) regions of the dorsal striatum. Previous functional imaging studies in Parkinson's disease (PD) patients have revealed that the diminishment of dopaminergic input into the caudate and putamen is associated with cognitive impairment and motor deterioration, respectively.[10]

2. RNA SEQUENCING IN CLINICAL DOMAIN

RNA sequencing (RNA-seq) holds significant promise for revolutionizing clinical diagnostics across various diseases by concurrently profiling global gene transcript levels and diverse RNA

species. Ongoing efforts aim to establish standardized benchmarks for technical and analytical practices in RNA-seq, enhancing precision, reproducibility, and accuracy.[11] This technology offers improved fusion detection, specific exon retention/exclusion insights, and the identification of noncanonical fusion partners that may be overlooked by targeted DNA panels. Potential companion diagnostic biomarkers, including single gene overexpression, transcriptional signatures, and immune activity, are envisioned for future applications.[12] Integrating RNA-seq into diagnostic approaches complements genomic data, addressing challenges such as uncertain significance variants and the interpretation of noncoding variants. Across eight studies, an average diagnostic improvement of 15% has been reported, showcasing the efficacy of RNA-seq in identifying causal variants in Mendelian disorders.[13]

While RNA-seq holds considerable promise, further efforts are required to establish its analytical validity and promote its integration into clinical laboratories.[11] In the realm of clinical trials, RNA-seq demonstrates its capacity to swiftly pinpoint biomarkers of response, guide biomarker-centric strategies, and address the variability in biomarker levels.[12] Clinical applications of RNA-seq with blood, fibroblasts, and muscle biopsies exhibit promise, yielding an average diagnostic improvement of 15%. Despite these advancements, a substantial portion of patients remains undiagnosed, underscoring the importance of incorporating additional omics data, particularly from proteomics. The refinement of analytical tools and the formulation of guidelines for interpreting abnormal RNA phenotypes are imperative and should be seamlessly integrated into existing protocols.[13]

3. PEPTIDE AND PROTEIN QUANTIFICATION

3.1 Where do we get our data from?

We rely on two different approaches to analyse the proteomes of living organisms, cells and tissues: top-down and bottom-up. Both rely on mass spectrometry (MS) for protein identification and characterisation.

The bottom-up approach in proteomics employs proteolytic cleavage to fragment proteins into peptides, which are subsequently analyzed using a mass spectrometer. Techniques like Gel Electrophoresis and Chromatography are employed for validation in bottom-up proteomics, ensuring the effective separation of individual peptides. In contrast, the top-down approach directly applies Mass Spectrometry analysis to intact proteins, eliminating the need for cleavage and thereby preserving the structural characteristics of the protein throughout the analysis process.

3.1.1 Bottom-Up v/s Top-Down

The widely adopted bottom-up proteomics approach facilitates protein quantification with high resolution and simplicity. However, it has limitations, including reduced sequence coverage, leading to the loss of valuable information on post-translational modifications (PTMs) and alternative splice variants (ASVs). In contrast, the top-down approach offers the potential to access the complete protein sequence while preserving structural characteristics, eliminating the time-consuming protein digestion phase of the bottom-up method. Despite its advantages, the top-down approach faces implementation challenges, including high costs associated with

instrumentation, making it less feasible on a large scale due to a lack of integrated intact protein fractionation methods for tandem mass spectrometry.

At our Proteomics Lab, IITBombay, we perform the bottom-up approach for mass spectrometry analysis. This involves various phases: (1) Sample Generation (2) Tissue Lysate (3) Peptide Cleanup (4) QC (5) Mass Spectrometry (5) Data Analysis. A detailed schematic of steps in Bottom-Up Proteomics is given in appendix, Figure 32(generated using draw.io).

3.1.2 Tissue Sample Generation

Figure 2: Sonicator, uses sound energy for homogenizing the tissue contents

In this phase, tissue samples are acquired from medical facilities to ensure subsequent analysis integrity. Initially, a phosphate-buffered saline (PBS) solution is used to eliminate residual blood, ensuring a purified starting material. Following blood removal, a two-fold process of denaturation and digestion prepares tissue proteins for downstream analysis. Denaturation disrupts non-covalent bonds, converting proteins into their primary structure, while digestion breaks down primary amino acid structures. The tissue is weighed, and a lysis buffer with a Protease Inhibitor Complex (PIC) is added to inhibit protease activity, preserving protein integrity. Homogenization follows, using methods like sonication or mechanical homogenization for uniformity. Centrifugation at 4 degrees Celsius and 80,000 rpm separates cellular debris from the liquid supernatant containing desired protein components. Chemical treatments, including Tris(2-carboxyethyl)phosphine (TCEP) as a reducing agent and Iodoacetamide (IAA)

for alkylation, facilitate downstream analysis by disrupting disulfide bonds and preventing their reformation. These modifications collectively prepare proteins for subsequent processing.

3.1.3 Tissue Lysate

Following these preparatory steps, enzymatic digestion is initiated using the enzyme trypsin. Trypsin cleaves peptide bonds specifically at the C-terminal ends of arginine and lysine residues, leading to the generation of peptides. This digestion process is allowed to proceed for a period of 16 hours, enabling thorough breakdown of the proteins into shorter peptide fragments, suitable for subsequent mass spectrometric analysis.

3.1.4 Peptide Cleanup

In this phase, a peptide cleanup procedure is crucial for eliminating salts and other contaminants that could adversely impact our mass spectrometer's precision. The removal of salts is paramount to prevent detrimental effects on both the analytical process and the instrument itself. The peptide cleanup begins by allowing the digested peptides to mix, ensuring thorough homogenization. A SepPak Vac 1 cc (100 mg) column, designed with C18 resin for effective peptide binding, is employed for the desalting process.

To initiate desalting, a conditioning solution is introduced into the column, comprising 90% methanol, 10% water, and 0.1% trifluoroacetic acid (TFA). This solution activates the C18 resin, enhancing optimal peptide binding. The conditioning solution is passed through the column three times for effective activation.

After conditioning, an equilibration solution is introduced to align the column's activation conditions with the peptide solvent environment. The mixing process temporarily halts for approximately 15 minutes during the equilibration step. Once completed, the peptide sample is carefully introduced into the column, facilitating binding interactions between peptides and the C18 resin.

After the sample has been passed through the column, equilibration solution is again passed once more. Subsequently, an elution solution is introduced to liberate the bound peptides from the column matrix. The elution solution consists of a composition of acetonitrile–water–trifluoroacetic acid (TFA) in a ratio of 70:30:0.1 (v/v). To ensure the efficacy of the desalting process, it is imperative to maintain a pH level below 3 throughout the procedure. This acidic pH environment is conducive to efficient desalting and preservation of the peptide samples.

3.1.5 Quality Control (QC)

For quality control (QC), capillary gel electrophoresis (CGE) is harnessed as a robust technique for the separation of proteins. The approach of sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) finds application in mass-based separation strategies. This is chiefly attributed to the uniform imposition of a negative anionic charge on all proteins, thereby rendering separation independent of charge and solely reliant on mass. The role of SDS extends beyond charge modification; it serves as a denaturing agent, effectively dismantling protein structures into primary configurations by disrupting non-covalent bonds.

The procedural tenets of this method encompass several key steps:

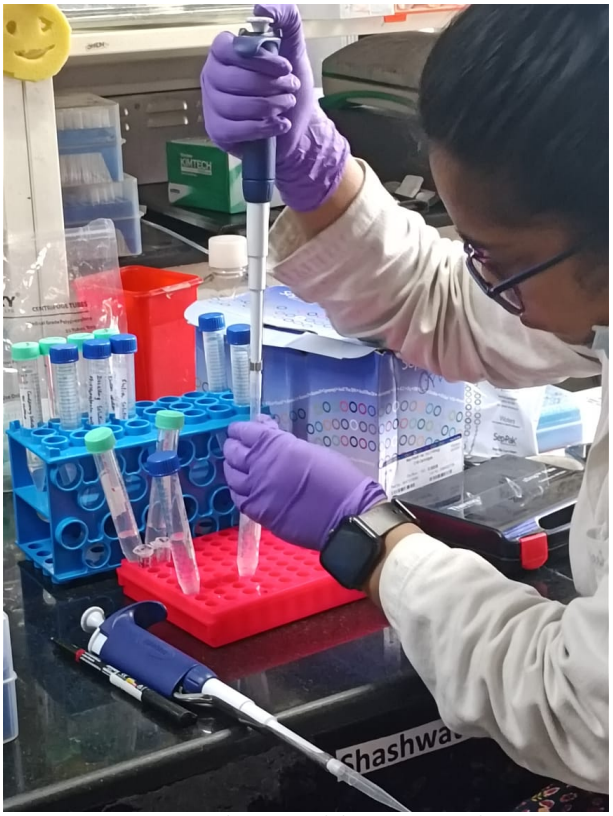


Figure 3: Passing the equilibration solution gently

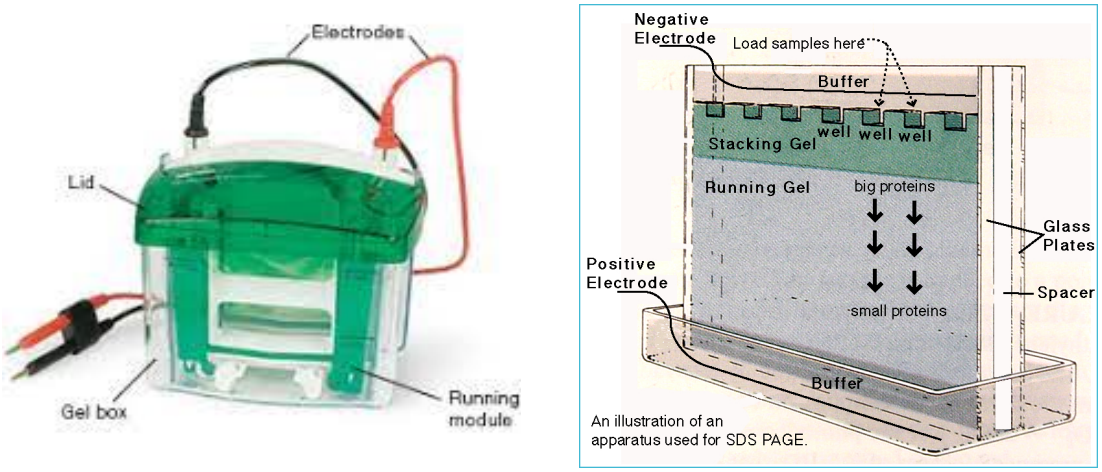


Figure 4: Protein Electrophoresis Equipment

- **Gel Preparation and Apparatus Assembly**:The process begins with the preparation of a gel medium, assembling the gel apparatus. The gel acts as a molecular sieve, with two sections: the stacking gel (upper segment) and the resolving gel (lower section). Stacking gels ensure uniform protein levels before entering the resolving gel. Resolving gel is poured first, followed by alcohol to prevent contact with the atmosphere. The stacking gel is then added. Ammonium persulfate (APS) and tetramethylethylenediamine (TEMED) catalyze

acrylamide and bisacrylamide polymerization for polyacrylamide gel formation.

- **Protein-Sample Mixing and SDS Treatment**: SDS is introduced with the polymerization agent, and the stacking gel has a lower SDS concentration than the resolving gel. This dichotomy aids in efficient protein alignment and resolution. As SDS concentration increases, pore size diminishes, enabling faster migration of smaller proteins. The stacking gel is loaded with combs. The protein sample is mixed with an SDS-containing buffer and subjected to elevated temperatures, uniformly coating proteins with a consistent negative charge.

- **Electrophoresis and Separation**: Electricity is applied to facilitate protein separation based on size, with smaller proteins migrating faster towards the positively charged anode. Migration depends solely on protein mass, as all proteins carry the same negative charge.

- **Quantification**: Bromophenol Blue, Electrophoresis Grade, serves as a color marker in this analytical framework, monitoring polyacrylamide gel electrophoresis progression. With a slight negative charge at moderate pH, it migrates congruently with DNA or proteins within the gel matrix. After separation, proteins are fixed, stained, or de-stained for visualization, and their quantities are quantified. The formed bands undergo further analysis for the identification of novel proteins.

3.1.6 Mass Spectrometry

Mass spectrometry accurately determines the molecular masses and structural characteristics of biomolecules, especially proteins. The process initiates with ionization, employing specific methods to convert biomolecules into ions, enabling their manipulation and analysis within the mass spectrometer. Subsequently, these ions undergo separation based on their mass-to-charge ratio (m/z) through a series of analyzers. The application of electric and magnetic fields causes the ions to follow distinct trajectories. Common components such as quadrupole, time-of-flight (TOF), and ion trap analyzers contribute to the precise separation and characterization of ions within mass spectrometers.

Upon completion of ion separation, the ions are detected and quantified. In the context of proteomic analysis, tandem mass spectrometry (MS/MS) further extends the capabilities. It involves isolating a specific ion of interest, fragmenting it, and analyzing the resulting fragments. This process provides valuable structural information about the molecule, aiding in its identification.

3.2 Protein Quantification

The Bradford assay is a commonly employed method for determining the protein concentration in a solution. This assay relies on the capacity of Coomassie Brilliant Blue dye to bind to proteins, inducing a color alteration that can be quantified spectrophotometrically using a Multi-Detection plate reader.

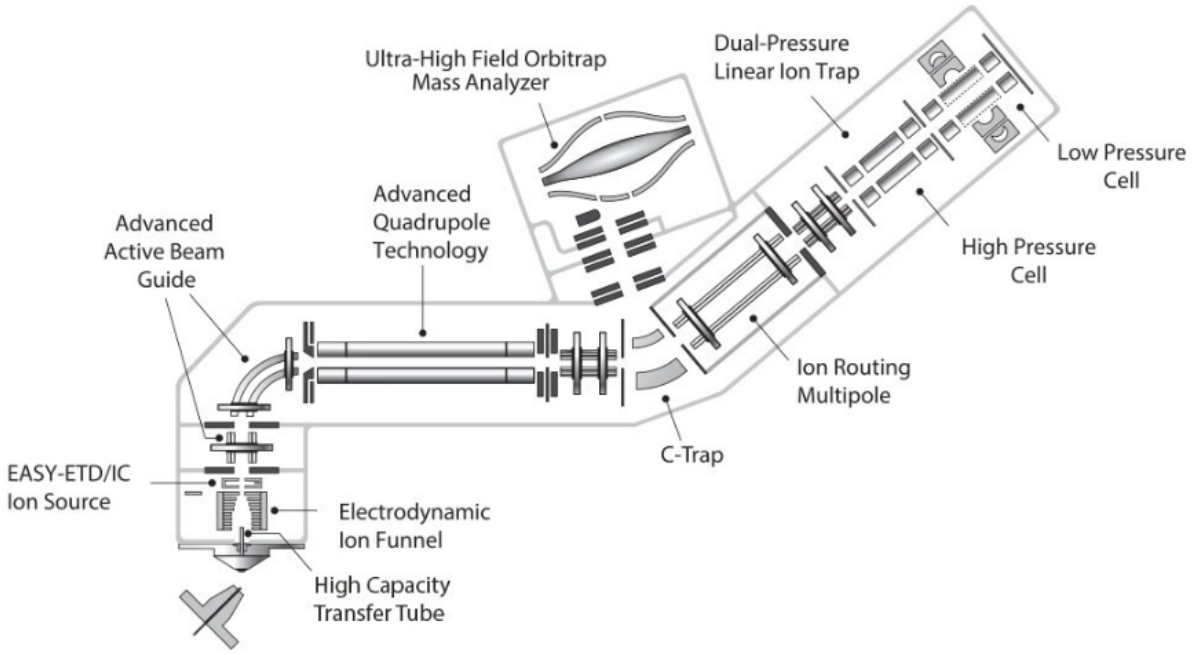


Figure 5: Orbitrap Tribrid Mass Spectrometer, ThermoFisher Scientific

3.2.1 Materials Needed

The materials needed for Conducting the Bradford assay based Protein Quantification test requires the following materials(instruments and chemicals):

- Bradford Reagent
- Bovine Serum Albumin(BSA) standard solutions of known concentrations
- Protein samples taken out from the vacuum after the drying phase
- Multi-Detection plate reader(Here, Multiskan Go spectrophotometer) capable of measuring asborbance (OD values) at wavelength = 595 nm
- Pipettes and disposable round cuvettes
- Microplate wells
- Distilled water (for dilution)

3.2.2 Methodology

Preparing BSA Curve

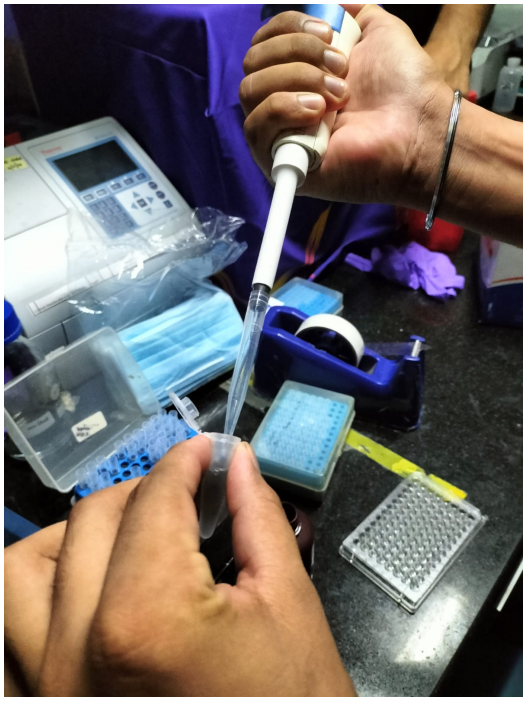
Prepare a series of BSA standard solutions with known concentrations (e.g., 0, 2, 4, 6, 8, and 10 μg/mL). Add 200 μL of each standard solution to separate cuvettes.

Prepare Protein Samples

Dilute your protein samples if necessary to fall within the linear range of the assay. Add 200 μL of each diluted protein sample to separate cuvettes.

Add Bradford Reagent

Add 800 μL of Bradford reagent to each cuvette containing the standards and protein samples. Mix gently but thoroughly to ensure proper mixing of the reagent with the protein solutions. This homogenisation is done using the instrument 6b.



(a) Adding Bradford Reagent to the Cuvette

(b) Mixing the constituents of the cuvette

Incubation

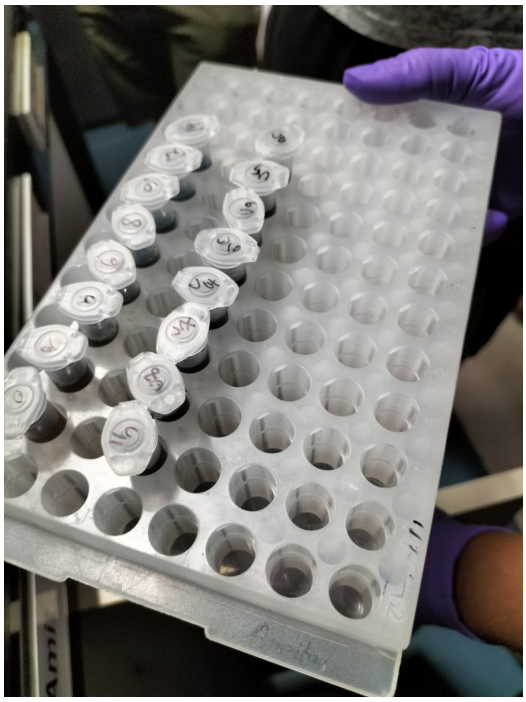
Allow the mixture to incubate at room temperature for about 5 to 10 minutes. During this time, the Coomassie dye binds to the proteins, resulting in a color change from brown to blue.

MicroPlate Wells SetUp

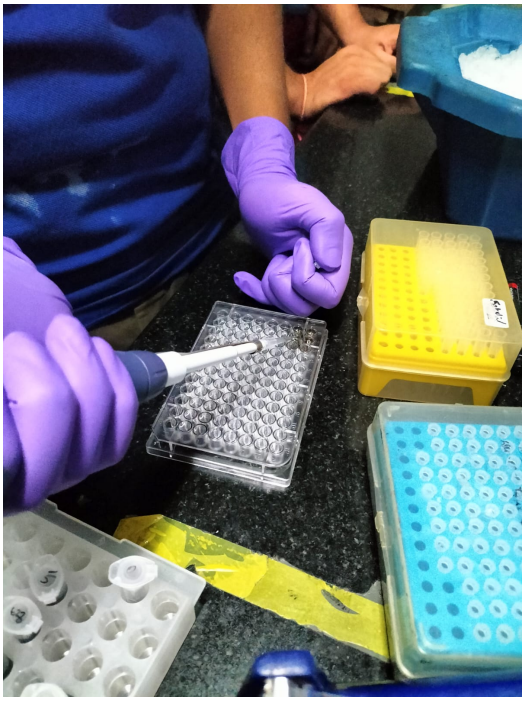
We shall use two wells for each protein sample, so that we can have an average reading obtained from both. Add 180 μL of Bradford (without any protein) to 2 wells to prepare the blank. Now add the remains samples as well, with two wells for each sample.

Measuring Absorbance

Place the microwell plate away from light as the bradford assay may be sensitive to light and



(a) Protein Samples



(b) Adding incubated samples to microplate

this can affect the readings. Now load the microplate into the multi detection plate reader. Measure the absorbance of each standard solution and protein sample at 595 nm against the reference blank. The blank reading will correspond to the color of the Bradford reagent itself.

According to the **Beer-Lambert law**, the absorbance is proportional to the concentration of the absorbing species. In this case, it's the protein-bound Coomassie dye. Subtract the absorbance of the blank well from the absorbance of each corresponding sample well. This corrects for the color of the Bradford reagent itself.

Construct Standard Curve and Calculate Protein Concentration
Plot a graph of the blank-corrected absorbance (y-axis) versus known BSA concentrations (x-axis) for the standards. Determine the linear equation of the standard curve using the plotted data. This equation will help us calculate protein concentrations of unknown samples based on their absorbance values.

3.2.3 Experiment

We conducted the experiment and obtained the following readings.

The plot of the OD vs Conc curve looks as 10. The best fit line comes out as OD = 0.0322(Conc). Concentration is in μg/mL.

Figure 8: Multi detection plate reader

CONC	OD.1	OD.2	AVG
0	0.002900	#########	0.000000
2	0.1278	0.1288	0.128300
4	0.1816	0.1869	0.184250
6	0.2093	0.2338	0.221550
8	0.2891	0.2802	0.284650
10	0.2949	0.3203	0.307600
12	0.3796	0.3899	0.384750
14	0.4122	0.4065	0.409350

Figure 9: Average adjusted OD values for different concentrations

We were then given unknown samples, and we had to measured the OD values, the blank substracted(Adjusted) averaged OD values and the predict concentration using the equation of line curve earlier is shown in 11

Note: The sample 5 seems erroneous due to high difference in OD values from corresponding wells, which possible tells the improper addition of the bradford assay to the microplate, it might have been affected by light or the the presence of bubbles may have the OD values corrupted.

3.3 Peptide Quantification

Peptide quantification provides invaluable insights into the concentration of specific peptides within a given sample. This multi-step process involves absorbance measurements, concentration calculations, and the subsequent interpretation of results across a diverse range of peptide samples.

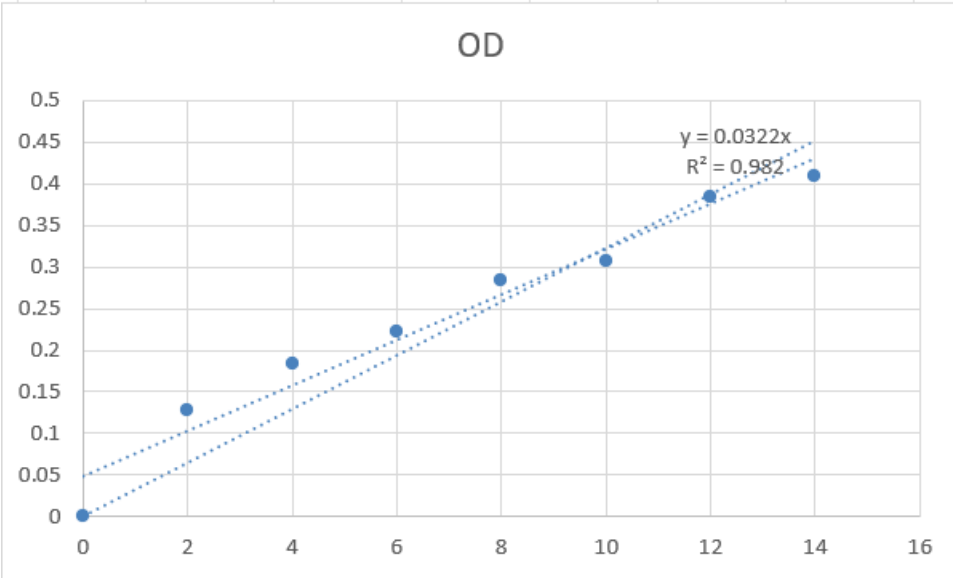


Figure 10: OD vs Conc Standard Curve Plot

Sample	Well 1	Well 2	AVG	Predicted Conc
1	0.3123	0.2869	0.2996	9.304347826
2	0.3820	0.4167	0.39935	12.40217391
3	0.2238	0.2248	0.2243	6.965838509
4	0.4135	0.4357	0.4246	13.1863354
5	0.1998	0.5763	0.38805	12.05124224
6	0.3812	0.2678	0.3245	10.07763975
7	0.3158	0.3175	0.31665	9.833850932

Figure 11: Predicting unknown concentrations

3.3.1 Materials Needed

The materials needed for conducting the Peptide Quantification test requires the following materials(instruments and chemicals):

- Multi-Detection plate reader(Here, Multiskan Go spectrophotometer)
- μDrop Plate
- Peptide Samples - Hela and Mcf7 cell lines
- Pipettes

3.3.2 Methodology

Loading sample
Start by adding the sample into the μDrop plate using a pipette. Place the plate into the spectrophotometer for readings.

Figure 12: μDrop plate

Measuring Absorbance:
Measure the absorbance of each standard solution and protein sample at 595 nm against the reference blank. The blank reading corresponds to the color of the Bradford reagent itself, often in the UV range (e.g., 280 nm). In this procedure, absorbance is measured at both 280 nm and 205 nm.

Calculating Concentration
By rearranging the Beer-Lambert equation, you can solve for concentration $C = \frac{A}{\epsilon l}$. **A** is the solution absorbance, ϵ is the molar absorptivity (also known as molar extinction coefficient) of the substance, which represents how strongly the substance absorbs light at a specific wavelength. **l** is the path length that the light travels through the solution (typically in centimeters)

3.4 Experiment

We obtain the following observations 13. The range of 28 to 35 L/(mol·cm) is considered a common and reasonable range for ϵ values in the context of peptide quantification. If the ϵ value is too low, it might suggest that the peptides being quantified have fewer aromatic amino acids or the assay conditions need adjustment or presence of interfering substances, contaminants, or impurities in the sample. Conversely, if the ϵ value is too high, it could indicate issues with the assay or sample preparation.

4. PARKINSON'S DISEASE AT BRAINPROT

We review the BrainProt3 database,[14] which serves as a comprehensive and robust omics-based knowledgebase, it provides visualisation of data from human's brain and associated neurode-

Sample name	Avg (A205)	Avg (A280)	Avg (A280/A205)	Avg(280/205)*3.85	1-3.85(avg 280/205)	Epsilon=E205(27/[1-(3.85*avg280/205)]	CONC (without multiplying dilution factor) ug/ul	Amount fo 8ul	Plus 0.1% FA
Un0002	0.56475	0.0455	0.08	0.310181	0.689819	39.14	0.288574	6.93063	3.06937
mcf6	0.69825	0.05865	0.08	0.323383	0.676617	39.90	0.349961	5.714921	4.285079
hela1	0.58845	0.0538	0.09	0.351993	0.648007	41.67	0.282459	7.080667	2.919333
hela2	0.565	0.05135	0.09	0.349907	0.650093	41.53	0.272076	7.350889	2.649111
mcf7	0.57455	0.047	0.08	0.314942	0.685058	39.41	0.291556	6.859756	3.140244
hela3	0.54275	0.0457	0.08	0.324173	0.675827	39.95	0.271707	7.360859	2.639141
hela2.1	0.5425	0.0449	0.08	0.318645	0.681355	39.63	0.273804	7.304503	2.695497

Figure 13: Observations for Peptide Quantification Experiment

generative disorders. It also provides clinical trial information of drugs associated with brain deseases and identifies potential theraupetic compounds. It is divided into six sub-domains. The HBDA(Human Brain Disease Atlas) encompasses 56 human brain diseases as of date, including the Parkinson's (MeSH ID* D010300, MedGen UID† 463618) disease, which then links to four other domains - BDMC(Brain Disease Marker Curator), BDTM(Brain Disease Transcriptome Map), BDPM(Brain Disease Proteome Map) and BDDF(Brain Disease Drug Finder).

4.1 BDMC(Brain Disease Marker Curator)

This domain of BrainProt allows users to identify and select markers of PD which will accelerate biomarker discovery and therapeutic target identification. Potential Biomarkers for PD have been identified from three databases, DisGeNET‡, eDGAR, CTD§. Several Scores like the DisGeNET Score¶, CTD Score, Disease 2.0 Score, Pubpopular Score** & Harmonizome Score. Additionally this portal allows viewing scatter plots of top N biomarkers based on combination

*Medical Subject Headings is the National Library of Medicine's controlled vocabulary thesaurus used for indexing articles of MEDLINE/PUBMED

†MedGen organises information related to medical genetics, the corresponding UID is an identifier within the database

‡DisGeNET contains a compilation of genes associated to diseases, that comes from different publicly available databases

§Comparative Toxicogenomics Database is a robust and openly accessible database designed to enhance comprehension of the impact of environmental exposures on human health. It offers meticulously curated data concerning interactions between chemicals and genes/proteins, as well as relationships between chemicals and diseases, and genes and diseases

¶The DisGeNET score for GDAs(gene disease associations) takes into account the number and type of sources and the number of publications supporting the association. Curated sources would include information that has been carefully selected, evaluated, and annotated by experts to establish a reliable link between specific genes and diseases.

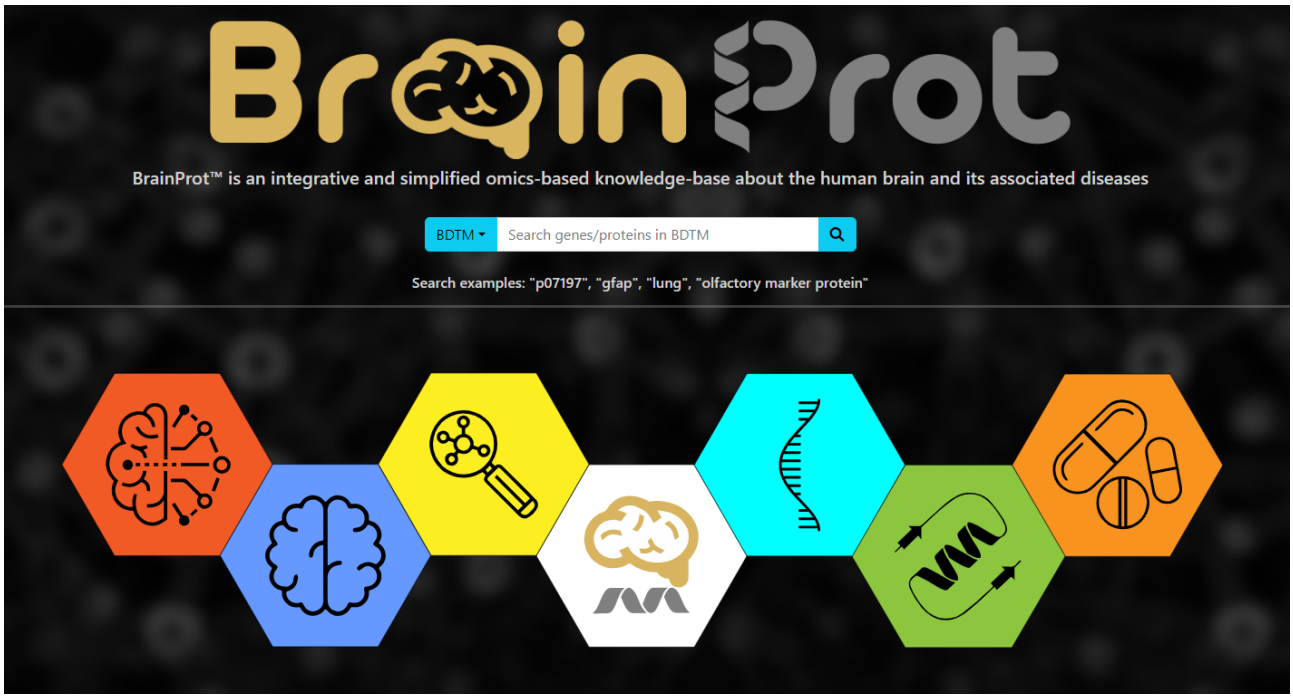


Figure 14: View of BrainProt's Homepage depicting the interfaces for 6 subdomains

Gene Name	UniProt ID	DISGENET Score	Harmonizome Score	Pubpular Score	CTD Score	Disease 2.0 Score	↑ BDMC Score
LRRK2	Q5S007	0.867	0.895	0.951		0.888	9.301
SNCA	P37840	1	0.695	1		1	9.159
TH	P07101	0.947	1	0.756	0.932	0.862	9.018
MAOB	P27338	0.922	0.815	0.854	0.806	0.812	8.955
GDNF	P39905	0.846	0.74	0.76	0.843	0.75	8.497
PINK1	Q9BXM7	0.905	0.609	0.899	0.813	0.838	8.435
PARK7	Q99497	0.915	0.576	0.871	0.794	0.825	8.276
DDC	P20711	0.91	0.661	0.787	0.882	0.688	8.183
SLC18A2	Q05940	0.898	0.602	0.71	0.86	0.7	7.94
DRD2	P14416	0.915	0.605	0.692	0.997	0.688	7.911

Figure 15: Top 10 Biomarkers ranked according to BDMC scores

of any two of the scores. BDMC calculates its own score for GDA, named as BDMC score. Refer 15. Some of the top scoring genes include LRRK2, SNCA, TH, PINK1 and PARK7 genes.

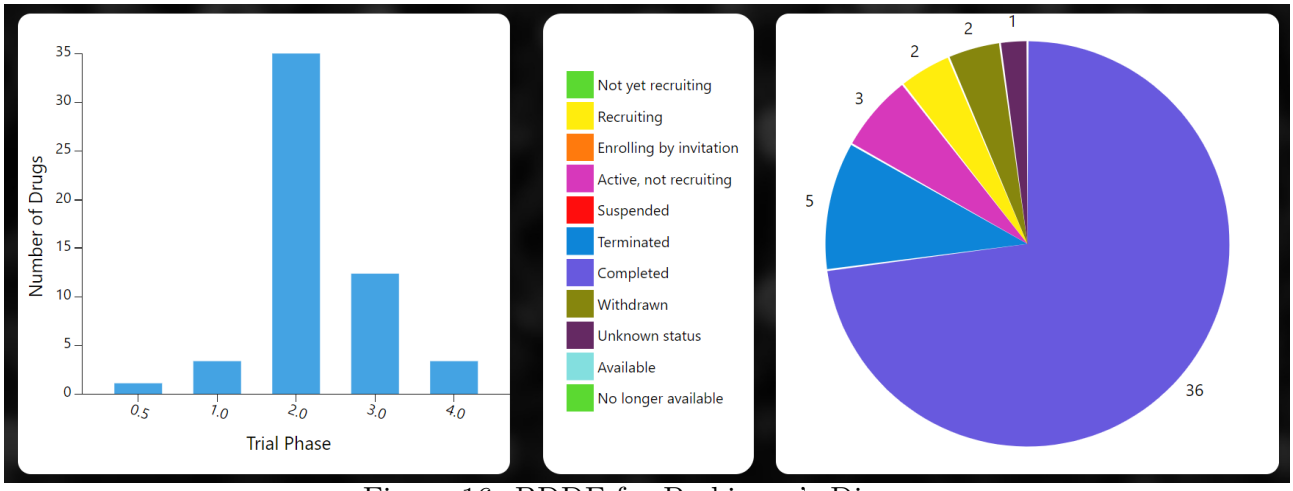


Figure 16: BDDF for Parkinson's Disease

4.2 BDTM(Brain Disease Transcriptome Map)

It integrates transcriptomics profile* of PD. This domain of BrainProt allows users to compare, visualise and understand the expression of genes in PD acquired from public repositories like GEO, OmicsDI and ArrayExpress through simplified, interactive visualisation plots.

4.3 BDPM(Brain Disease Proteome Map)

It allows users to compare, visualise and comprehend the expression of proteins in PD mined from various public repositories such as ProteomeXchange, OmicsDI, and PRIDE. This domain allows users to visualize processed intensity for each dataset. For a better understanding of the disease conditions, users can also access the sample meta-data of each dataset.

4.4 BDDF(Brain Disease Drug Finder)

It allows users to investigate and retrieve clinical trial information related to PD and access therapeutic compounds targeting potential markers associated with these diseases. By integrating multiple databases and resource hubs such as ChEMBL, BindingDB, Therapeutic Target Database(TTD), CLUE, Open Targets, PubChem, and ClinicalTrials.gov, BDDF offers a comprehensive overview of the therapeutic landscape for PD.

4.5 Review

For Parkinson's Disease†, BDPM provides 3 Proteomic datasets taken from ProteomExchange and BDTM provides 14 transcriptomic datasets taken from NCBI GEO. After looking for

*Transcriptomics often involves techniques to measure and quantify the levels of mRNA molecules present in a biological sample. This can help researchers understand which genes are actively being transcribed and translated into proteins under specific conditions. This field provides insights into gene expression patterns and how they relate to various biological processes, cellular functions, and disease states.

†For detailed information about available datasets on BrainProt for PD, refer to https://docs.google.com/spreadsheets/d/1kuHby8L22uxIAK371f1Xoq5-EdeM191jG9INGZKOtXE/edit?usp=sharing

BDMC scores, we investigate the protein and transcriptomic profile of the top biomarker genes, to validate the potential of biomarkers in the distinguishing diseased from control patients. LRRK2(known for its possible role in altered immune function in PD), also the one with highest BDMC score, 8/13 datasets do not contain information about LRRK2 gene, the rest 5/13 do not conclude that a significant difference exists, because the p-value $> 0.05$17. None of the 3 proteomic datasets have LRRK2 gene. The gene TH* stands 3^{rd} according to BDMC score has better statistics 18, with 4/13 datasets showing extremely significant difference(p-value $<$ 0.001), 1/13 with significant difference(p-value$<$ 0.01) and 8/13 with non significant difference(p-value$>$0.05). THe protein expression also shows non significant difference for all 3 proteomic datasets in BrainProt 19.

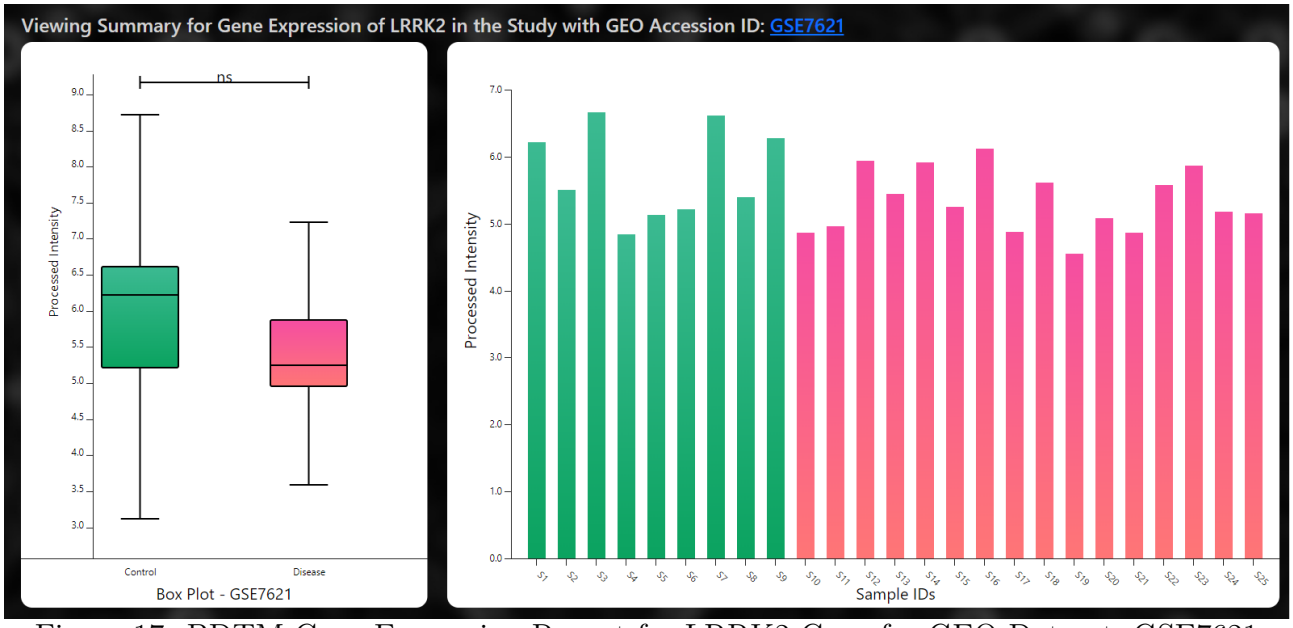


Figure 17: BDTM Gene Expression Report for LRRK2 Gene for GEO Dataset, GSE7621

We observe that although LRRK2(BDMC = 9.301) and TH(BDMC=9.018) are among the top genes according to BDMC score, the available datasets do not show satisfactory significance of them for Parkinson's disease. Potential causes maybe insufficient information in datasets, and hence search for alternate datasets should be performed. This may be due to noise and heterogenity. It also maybe conclusive that BDMC is not a metric of guarantee, and hence such validation tests must be performed to confirm the biomarker panel.

5. DATA MINING

We search on OmicsDI and PUBMED for recently published RNA-Seq and Proteomic datasets for Parkinson's disease under taxonomy HomoSapiens for samples extracted from Tissues, Serum, Plasma and CSF(CerebroSpinal Fluid). We compile the information - metadata and sample

*Tyrosine hydroxylase acts as a catalyst in the formation of L-dihydroxyphenylalanine (L-DOPA), the RDS in the biosynthesis of DA. Hence, PD can be considered as a consequence of TH-deficiency in the striatum

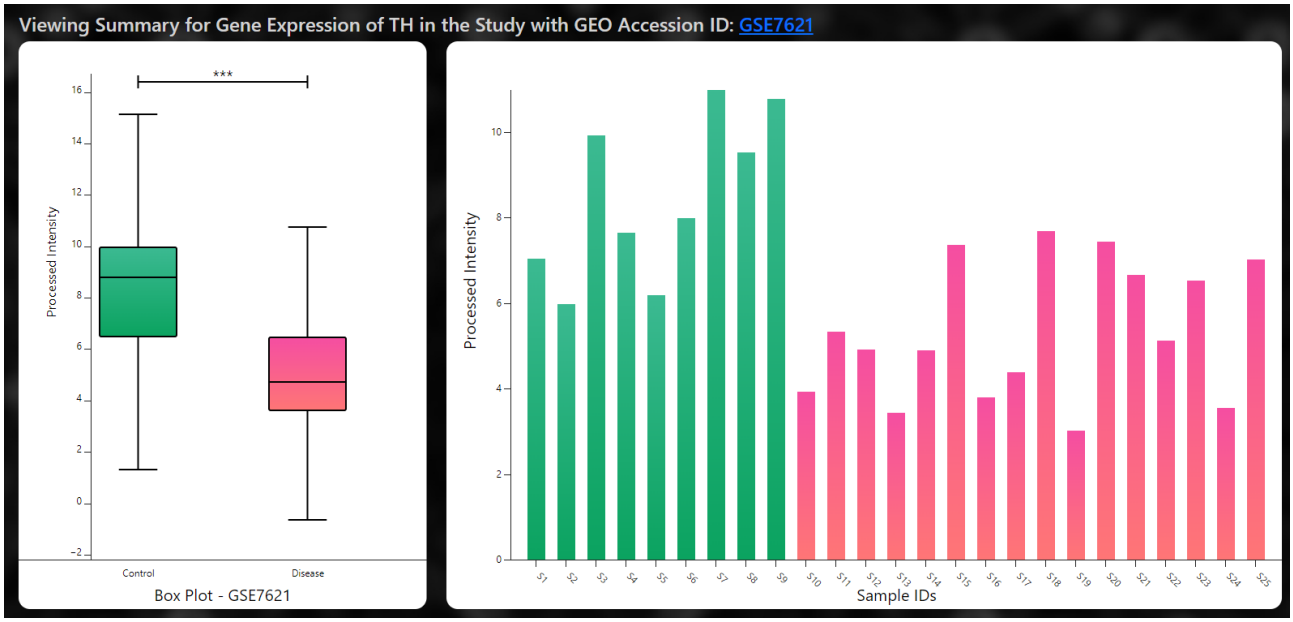


Figure 18: BDTM Gene Expression Report for TH Gene for GEO Dataset, GSE7621

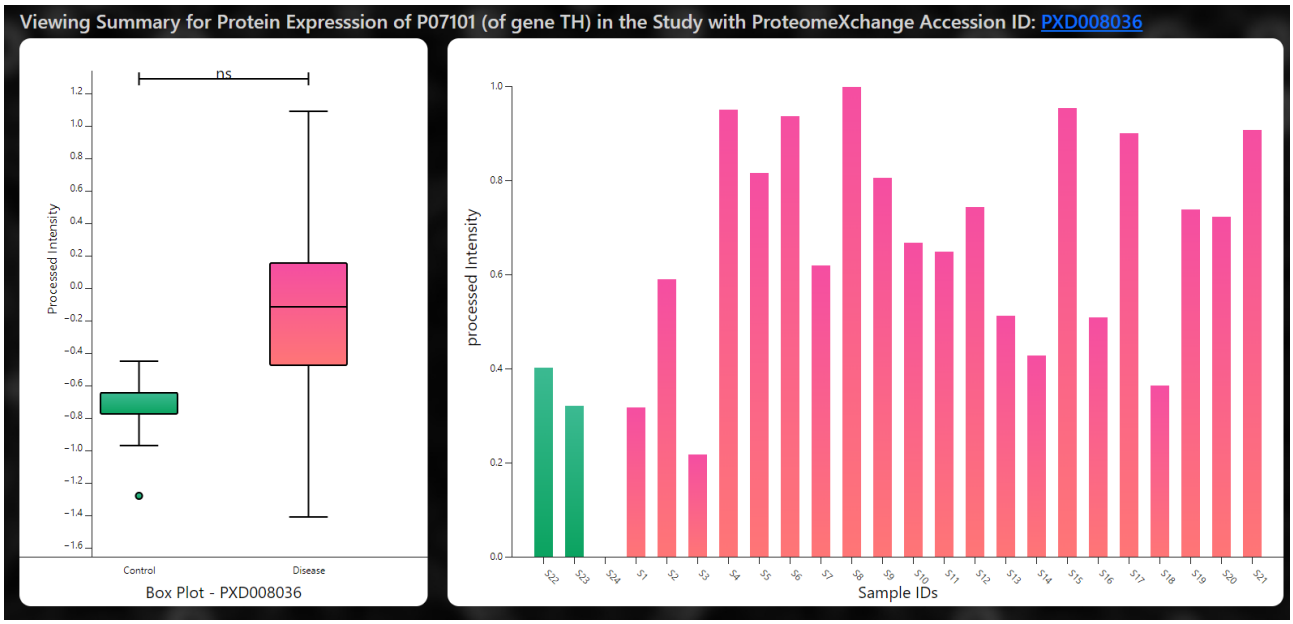


Figure 19: BDPM Protein Expression Report for TH on PRIDE Dataset, PXD021630

Study Information		Study 1	Study 2
Name of Person		Sabyasachi	Sabyasachi
Dataset Information	Dataset ID (ProteomeXchange)	PXD034120	PXD038555
	Secondary Link (If any)	NA	NA
	PMID/Link	36224378	36797805
	Paper Publication Date	2022 Oct 25	2023 Feb 20
	LFQ/Labelled/Anything else	-	LFQ
	Instrument	Orbitrap Fusion Lumos, Q Exactive	Q Exactive
	File Type	.raw	.raw
	Any Comments	Lip-MS for the detection of global protein structural changes, which is a structural analysis technique	30 samples in total. 10 for each PD, AD and MS
Sample and Study Information	Sample Type	Brain CSF	Blood Serum
	Patient/Sample Info Availability		
	Enzymes	-	Lys-C/Trypsin
	Fracationation	-	-
	Number of Sample	103	10
	Control/Healthy	51	-
	Diseased	52	10

Figure 20: Glimpse of Mining Sheet for BDPM

type, distribution onto a mining sheet. A glimpse of the sheet can be seen in Figure 20. Entire Mining Sheet for both BDPM and BDTM can be found here. We mined 5(3 CSF, 2 blood) new proteomic datasets post 2022, and 3 transcriptomic datasets derived from tissues. We explain the mined transcriptomic datasets below.

5.1 GSE205450: Transcriptome changes in human post-mortem PD striatum

GSE 205450 - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE205450
PUBMED - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10322907/
CodeBase - https://github.com/carynhale/PD_RNAseq
The dataset GSE205450 investigates transcriptome changes in the post-mortem striatum of individuals with Parkinson's disease (PD) in comparison to controls. The study encompasses 40 control samples and 35 PD samples, all confirmed through autopsy. High-throughput sequencing using the Illumina NovaSeq 6000 platform for RNA sequencing is employed in this dataset. The associated publication[10] delves into the molecular basis of clinical heterogeneity in PD by

conducting bulk RNA sequencing on postmortem caudate and putamen regions from 35 PD and 40 control brains. The study identifies common gene expression changes in both regions related to miRNA activity, immune response, synaptic signaling, mitochondrial dynamics, and lipid metabolism. Region-specific alterations in the caudate and putamen are associated with cognitive decline and levodopa-induced dyskinesia, respectively. Distinct molecular patterns are observed in late-onset and early-onset PD, and disease duration correlates with changes in caudate (oligodendrocyte development) and putamen (cellular senescence). Notably, transcriptome patterns in postmortem PD brains are reflected in antemortem peripheral blood, correlating with clinical features and highlighting potential blood-based biomarkers. The findings offer valuable insights into the molecular processes underlying the clinical diversity of PD, emphasizing the relevance of blood transcriptomics as a diagnostic and prognostic tool.

5.2 GSE169755: Polyomic analyses of dopaminergic neurons isolated from human substantia nigra in Parkinson's disease

GSE 169755 - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169755
PUBMED - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9561004/
The above study investigates the molecular profile of dopaminergic neurons in the substantia nigra pars compacta (SNpc) in Parkinson's disease (PD) using laser microdissection. Dopaminergic neurons were extracted from 10 human SNpc samples obtained at autopsy in PD patients and control subjects. RNA and proteins extracted were identified through RNA sequencing (Illumina NovaSeq 6000 platform) and nano-LC-MS/MS analyses, revealing differential expression in 52 genes and 33 proteins, including known PD-associated molecules.[15] Despite utilizing the same samples, the correlation between RNA and protein expression is reported to be low. This study employs an exploratory multiomic approach, integrating transcriptomics and proteomics, to gain insights into the specific molecular changes in PD-affected dopaminergic neurons. It represents the first effort to simultaneously analyze gene and protein expression in laser-dissected neuronal parts from the SNpc in PD, contributing valuable data to the field.

5.3 GSE68719: mRNA-Seq expression and MS3 proteomics profiling of human post-mortem BA9 brain tissue for PD and neurologically normal individuals

GSE 68719 - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68719 PUBMED - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722694/ The dataset incorporates mRNA-Seq expression and three-stage Mass Spectrometry Tandem Mass Tag Proteomics profiling of post-mortem BA9 brain tissue. The study encompasses 29 Parkinson's disease (PD) and 44 control samples for mRNA-Seq, and 12 PD and 12 control samples for proteomics (a subset of the RNA-Seq set). The investigation identified 3,558 unique proteins, with 7.9% exhibiting significant differences between PD and controls. Similarly, RNA-sequencing identified 17,580 protein-coding genes, with 6.2% being significantly different. Notably, only 0.94% of the significant genes overlapped between the two approaches. The integrative analysis suggests implications in mitochondrial processes, protein folding pathways, and GWAS loci in Parkinson's disease, underscoring the value of combining multiple genome-wide platforms for comprehensive insights into the pathological processes associated with PD.[16]

6. RNA-SEQ ANALYSIS PIPELINE

We choose the pipeline as described in the paper.[17] In this paper, they study potential pathways and molecular mechanisms involved in glioma related seizures revealed by RNA Sequencing of Intraoperative Perimtumoral Tissues. RNA Samples are extracted from IPBT resected from 12 LGG patients with or without seizures, 6 in each class. The authors choose only those RNA samples with good RIN value (> 7) for sequencing. Differential Expression are performed using R packages to identify DEGs(Differentially Expressed Genes) in pGRS(with seizures) and PGNS(without seizures). Sequences were quality-checked using fastQC and low-quality bases and reads were excluded from further analysis. RNA sequences were aligned to Human genome GRCh38 assembly using hisat2 aligner. The transcripts were assembled using stringtie and the raw count matrix was used for diferential gene expression analysis using deseq2 and edgeR packages. Common DEGs as identified in both deseq2 and edgeR(fold-change ≥ 2, padj/FDR < 0.001) were used for downstream pathway analysis. A schematic of RNA-seq data analysis pipeline is depicted in Fig. 21.

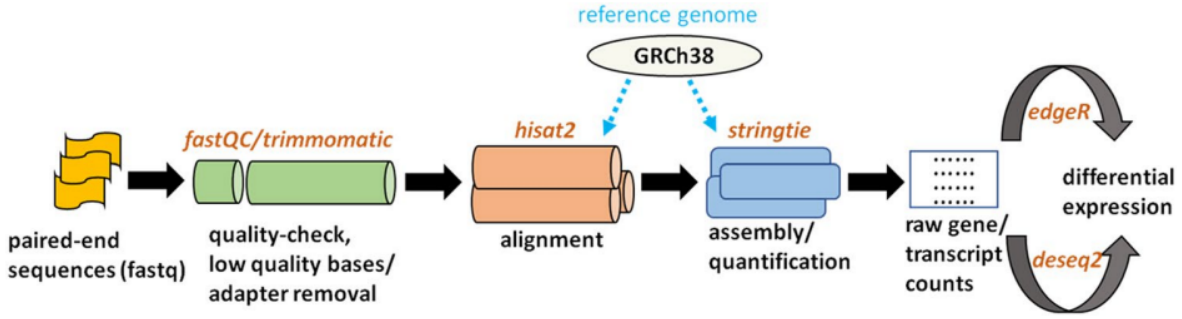


Figure 21: Pipeline from[17] to be employed for our purpose.

We follow the same pipeline and craft it to be efficient using multiple threading. The versions of the tools used, installation troubleshooting and code for both single-ended fastq reads and paired-ended fastq reads can be found on github. To summarise our pipeline follows the below order:

1. Download fastq reads using the SRA Toolkit using prefetech and fasterq-dump for a particular SRR
2. (Optional) Run Quality Check on the fastq reads for assessment
3. Use Trimmomatic to trim adapters and low-quality bases from the reads and (optionally) run the Quality Check on the trimmed reads for comparison.
4. Align trimmed paired-end reads to the reference genome using Hisat2 into a SAM alignment file
5. Sort the SAM file, convert it to BAM format, and indexes it using Samtools.
6. Run StringTie to assemble and quantify transcripts based on the aligned reads.

7. Remove all files except the stringtie gtf alignment output and repeat for the next SRR

After collecting gtf alignments for all samples, we generate raw gene and transcript counts text files using prepDE.py, provided by stringtie. The generated output is now subjected to Differential Expression Analysis. Exact details and parameters used can be found on github.

7. RESULTS

7.1 GSE205450: Transcriptome changes in human post-mortem PD striatum

This is in regard with Dataset GSE205450. The dataset contains transcriptomic paired-ended RNA-Seq data from caudate and putamen regions in postmortem brain of 40 control donors and 35 PD donors. Figure 22 shows the two regions. The publication linked with this dataset is PMID 37407548.[10] The authors analyse transcriptomic patterns and differences in PD caudate and putamen. We perform bulk RNA Sequencing on all 150 data samples. We generate the metadata.csv file needed for the DESeq Analysis from the GSE Matrix File provided in NCBI repository. We then pass the gene counts and the metadata through the DESeq analysis, while generating a bunch of useful plots. We first study the dataset in two cohorts separately, and then do a comparative analysis. Codes for all are made available on github.

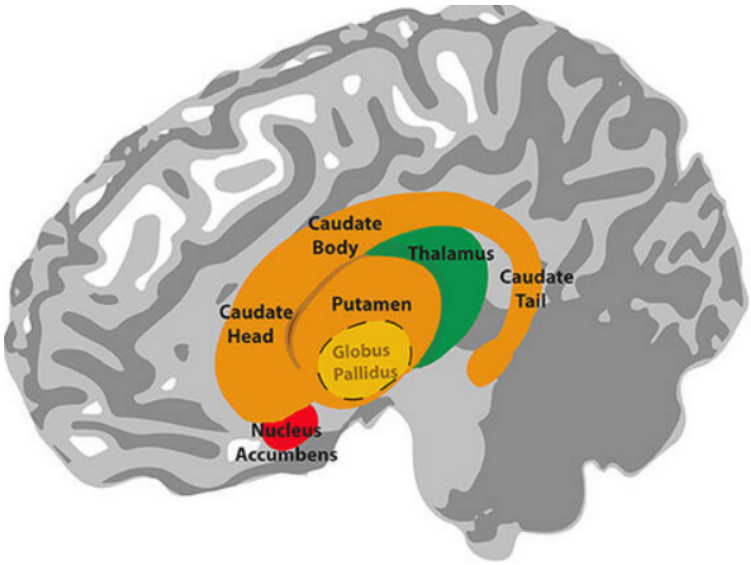


Figure 22: Caudate and Putamen Regions of dorsal striatum, Source: BrainInjury

DESeq2 employs the Wald test to assign p-values to individual genes, testing the null hypothesis that there is no differential expression between sample groups (PD vs control). If the p-value is small (e.g., $p < 0.01$), the null hypothesis is rejected, indicating a low probability (1%) that it is true. However, when testing numerous genes, there's a 1% chance that some genes may show significant p-values purely by chance, even if there's no true differential expression. To address this issue of multiple testing, DESeq2 adjusts p-values using the Benjamini and Hochberg method. The resulting BH-adjusted p-values, also known as False Discovery Rate (FDR), are found in the 'padj' column of the results object. These adjusted values help control

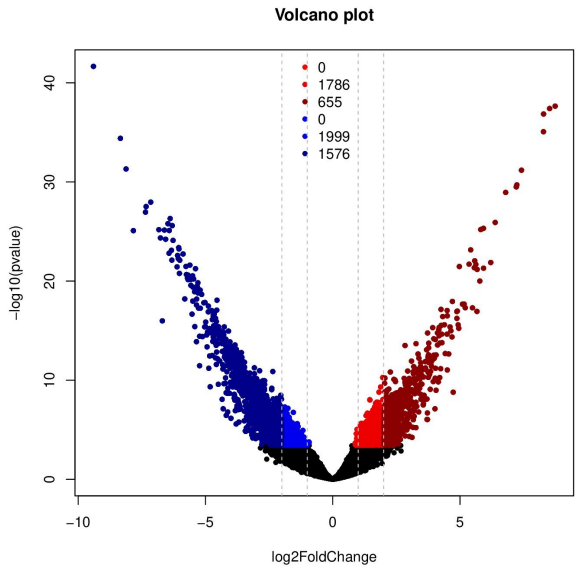


Figure 23: Volcano Plots (-log10(padj) vs log2FoldChange) of Caudate putamen transcriptomic samples. Helps identify genes with large fold changes(PD/CTRL) that are also statistically significant.

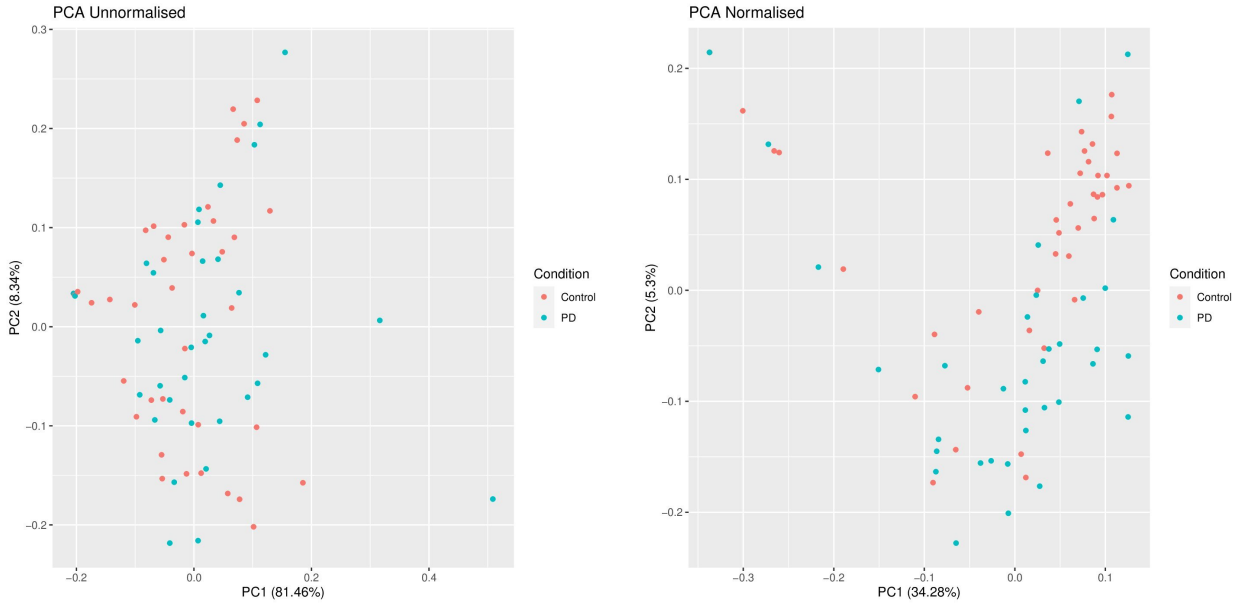
the risk of false positives associated with observing significant p-values by random chance when analyzing a large set of genes.

Volcano plots shows significantly changed RNAs in PD caudate and putamen compared to their respective controls. RNAs with significant changes in expression with FDR < 0.01 and an absolute value of log2 fold change > 1 and 2 are colored according to direction. Differential gene expression analysis revealed 2441 and 3575 up and downregulated RNAs, respectively, in the PD caudate compared to controls.

In Figure 24, the observed clustering does not align with expectations for control and treatment subjects. A Density Plot, which visualizes data distribution over a continuous interval, can reveal batch effects in RNA-Seq data. Batch effects, stemming from various experimental factors, can introduce spurious variability unrelated to the studied condition. Strategies to address batch effects are comprehensively discussed by Jaffe et al.[18] However, Figure 25 illustrates that the density plot is not optimal. To correct batch effects, normalization is employed, which scales numeric values in a dataset to a common scale without distorting differences, reducing data sparsity. The Variance Stabilizing Transformation (VST) method is chosen for normalization due to its efficacy with large samples, resulting in improved density plots and PCA clustering.

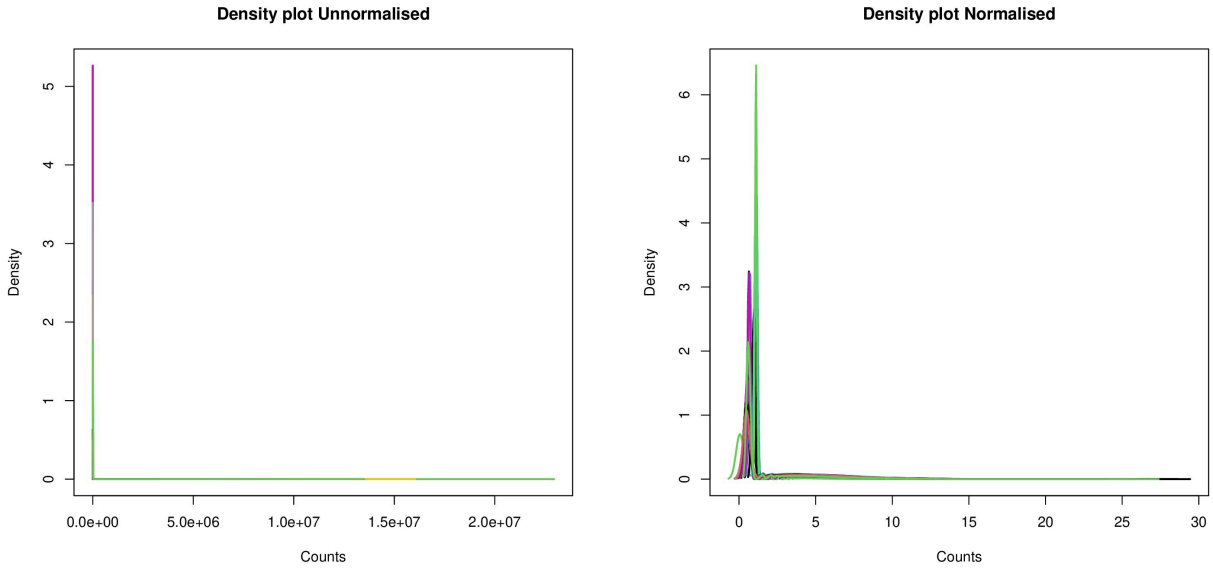
Heatmaps (Figure 33) provide a broad view of similarities and dissimilarities among samples. Calculating Euclidean distance between samples reveals considerable similarity between PD and control samples, suggesting that Euclidean similarity might not be an accurate metric.

MA plots (Figure 26), displaying log ratios (M) versus averages (A), visualize differences between two groups. Generally, gene expression is expected to remain consistent between con-



(a) (b)

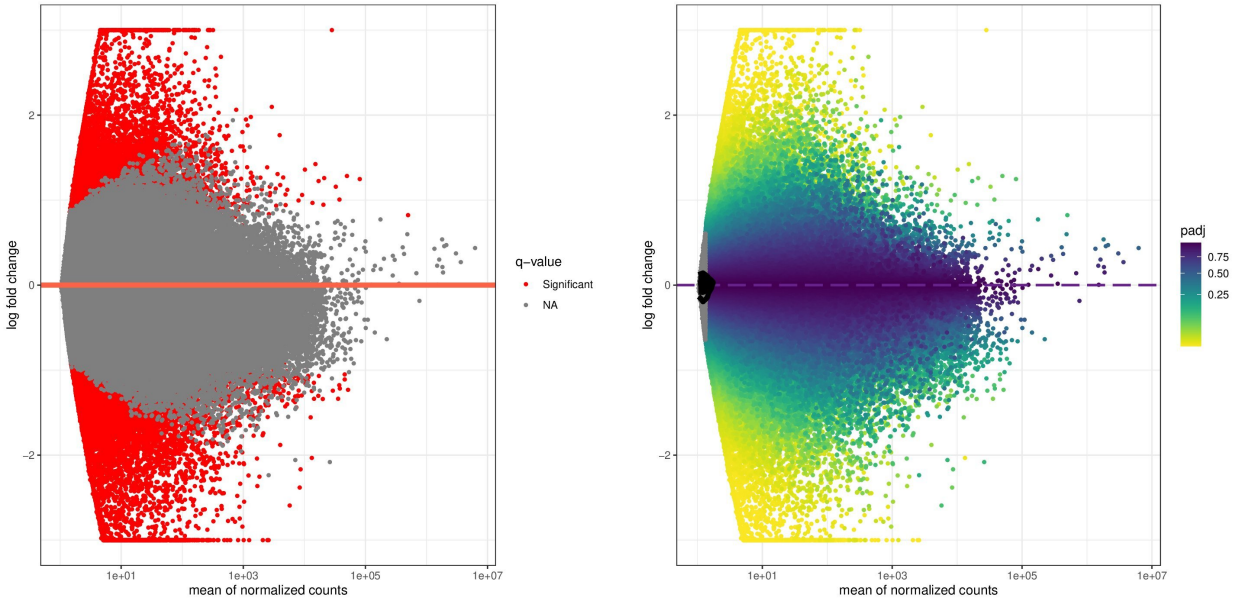
Figure 24: Principal Component Analysis showing separation of disease and control groups in Caudate region. (a) Unnormalised data and (b) Normalised data.



(a) (b)

Figure 25: Density plots for disease and control groups in Caudate region. (a) Unnormalised data and (b) Normalised data

ditions, resulting in an MA plot resembling a trumpet shape. DESeq2 offers a built-in method for constructing MA plots, but in this case, ggplot2 is used for better visualization. The MA plots for caudate samples are depicted in Figure 26.visualisation. Figure 26 are the MAplots for the caudate samples.



(a) (b)

Figure 26: MA Plots for results from caudate samples created using ggplot2 customised as to our purpuse (a) Differentiating significant from other genes (b) Showing contours of significance

Dispersion refers to within-group variability for a gene, which can influence its significance. For strongly expressed genes, dispersion can be seen as a squared coefficient of variation, where a dispersion value of 0.01 implies that the gene's expression typically differs by approximately $\sqrt{0.01} = 10\%$ between samples of the same treatment group. Weak genes, in addition to dispersion, also experience Poisson noise as an additional source of variability. The function **plotDispEsts** visually represents DESeq2's dispersion estimates in Figure 27. Black points on the plot represent dispersion estimates for individual genes, exhibiting notable fluctuations, especially with limited samples. To address this, a red trend line, illustrating the dependence of dispersion on the mean, is fitted. Gene estimates are then shrunk toward this line, resulting in final blue points used for hypothesis testing. Blue circles above the main cluster identify genes with high dispersion (outliers), exempt from the trend line shrinkage.

7.2 GSE169755: Polyomic analyses of dopaminergic neurons isolated from human substantia nigra in PD

This is in regard with dataset GSE169755, which contains single ended RNA seq data extracted from dopaminergic neturons using laser dissection. The publication linked with this dataset is PMID 34528139. Overall distribution contains 12 analyses, 6 for each groups of samples: PD and Control and 3 replicated samples per group.

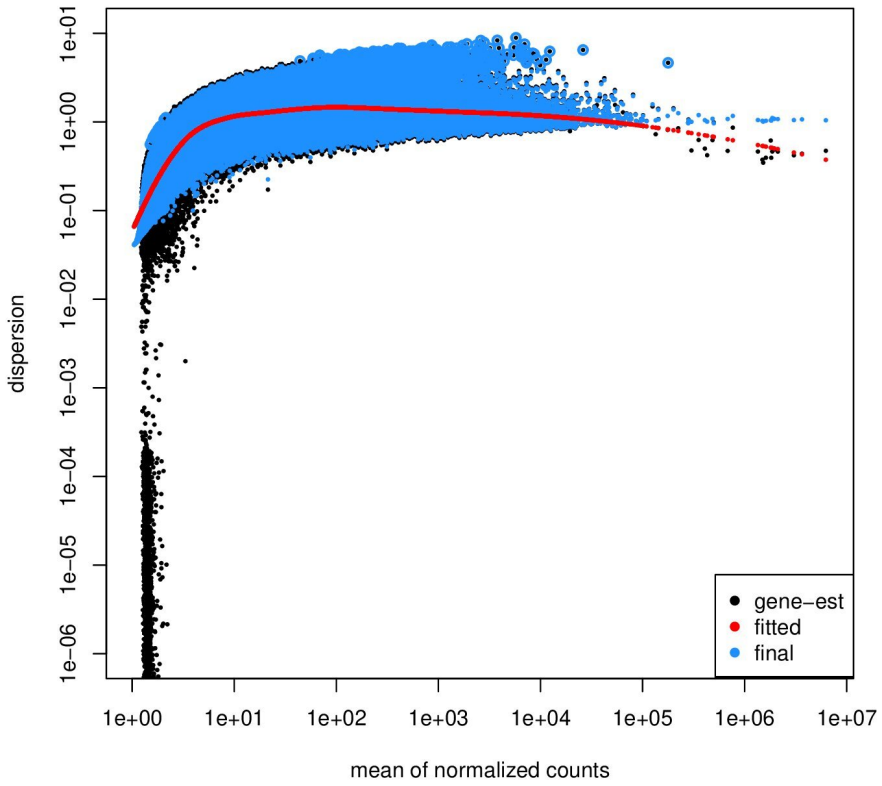


Figure 27: Dispersion plot to visualise dispersion estimates in caudate data

Figure 29, we see the number of upregulated and downregulated genes are 930, 716 respectively. We have a lot of genes with missing values, and hence many have NA adjusted pvalues. The regions in Volcano plot are broken into two shades of colors based on thresholding log2FoldChange at 2.

Dispersion plot in Figure 31 shows a lot of outliers above the main cloud.

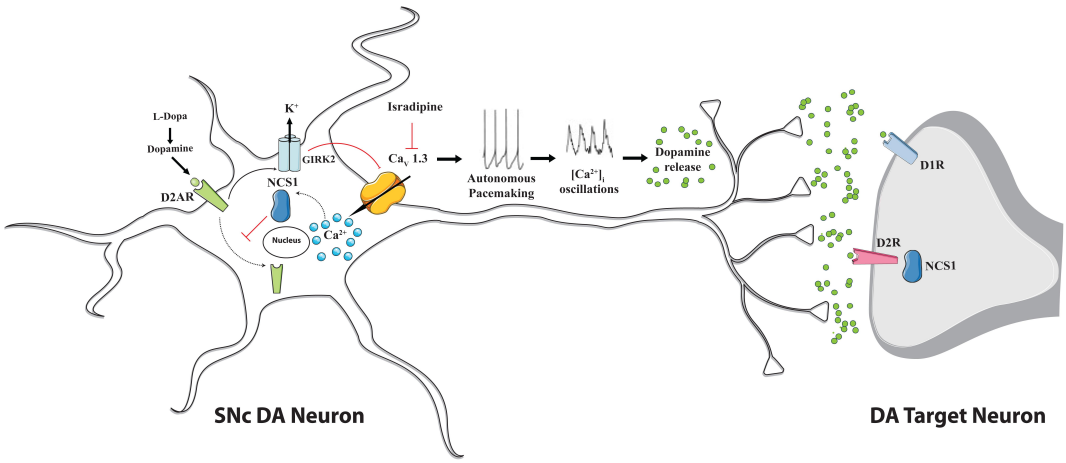
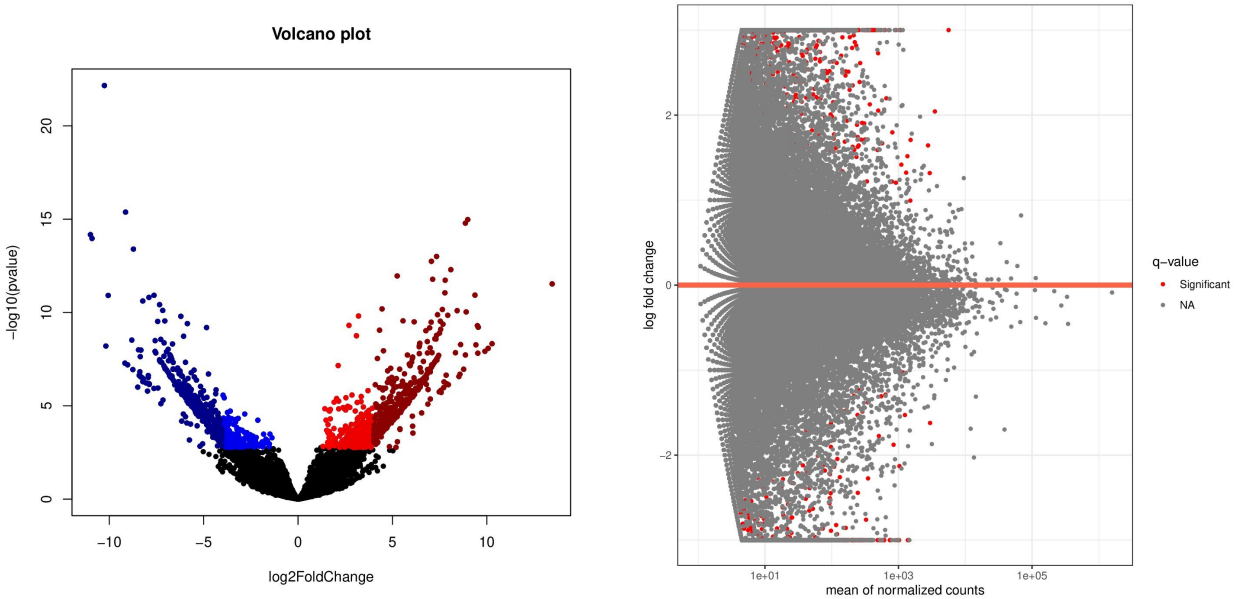


Figure 28: The schematic shows a proposed mechanism for Cav1.3 L-type Ca2+ channel action and NCS-1 contribute during autonomous firing of substantia nigra pars compactadopaminergic neurons.[19]

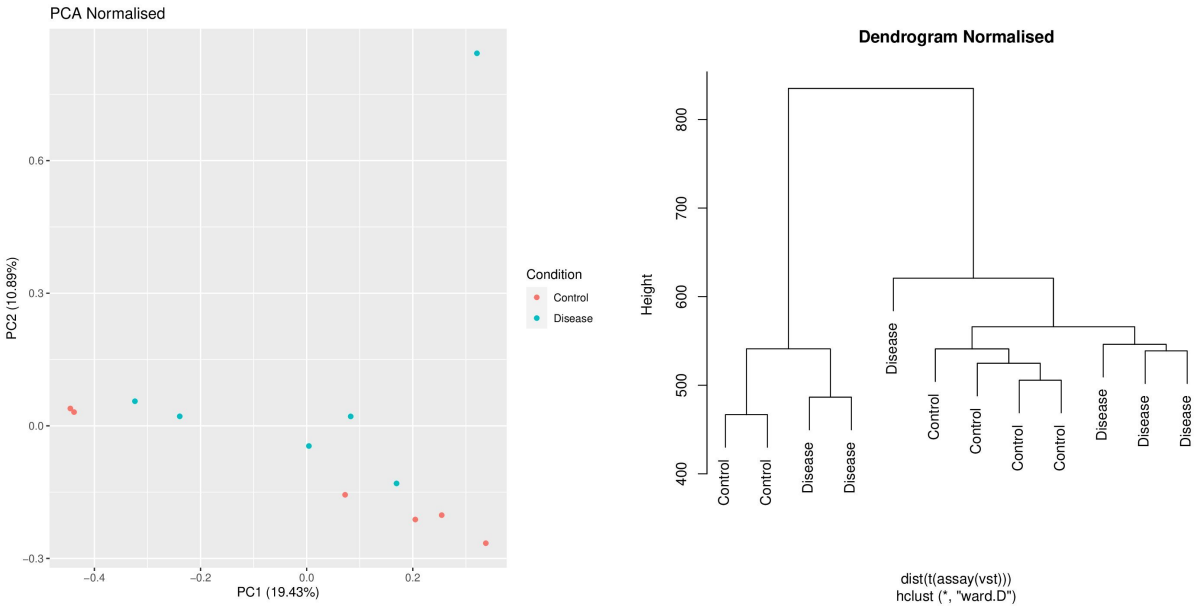


(a) (b)

Figure 29: Volcano Plot and MA Plot indicate that number of differentially expressed genes are relatively lower than the previous study

8. IN CONCLUSION

In conclusion, this project focused on the development of a robust bulk RNA sequencing pipeline, which was successfully tested on two datasets with promising results. A significant portion of the effort was dedicated to addressing challenges encountered during RNA sequencing, particularly when gene counts were zero. The implementation of pseudo-counts proved effective in mitigating this issue, contributing to the overall success of the analysis. Throughout the project, we encountered various troubleshooting challenges, and solutions were meticulously compiled.



(a) (b)

Figure 30: Normalised PCA Plot and Dendogram for the dataset

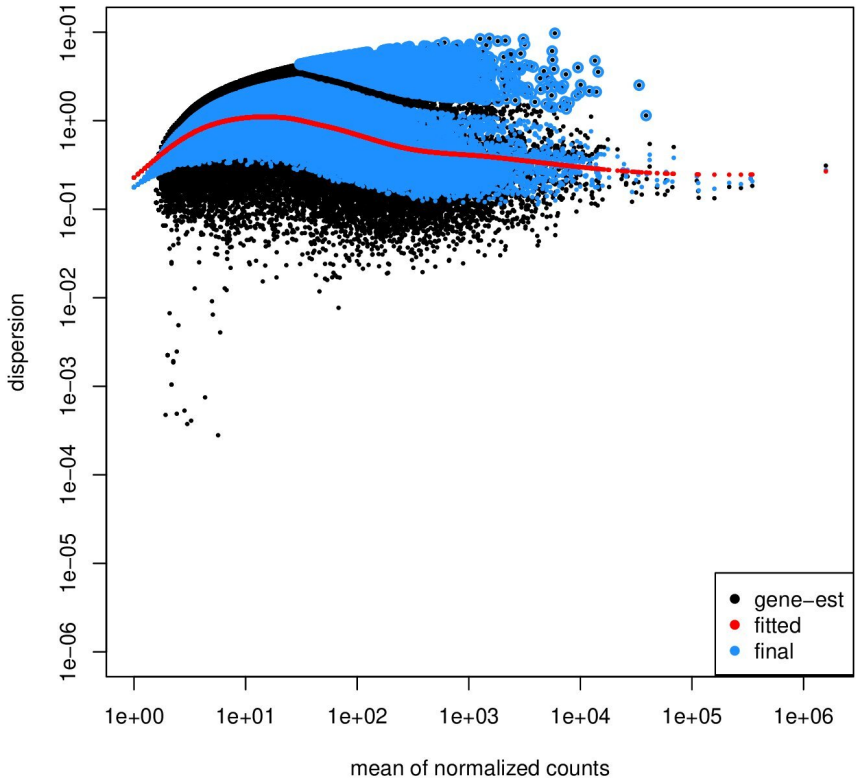


Figure 31: Dispersion plot

GSE205450, gave promising results from the transcriptomic samples extracted from caudate region. We can also extend this to analysis of samples from putamen region(analysis still underway), and do a multi-cohort comparative analysis.

Future work on this project could involve Gene Enrichment Analysis using Gene Ontology (GO) terms to unveil molecular pathways associated with disease progression. Additionally, implementing a quality check by considering only samples with high RNA Integrity Number (RIN) values before DESeq analysis could help eliminate unnecessary noise in the data. Exploring integrative analyses combining RNA sequencing and proteomics data could provide a more comprehensive understanding of the molecular mechanisms underlying the observed changes. Additionally, incorporating machine learning techniques for predictive modeling and classification based on gene expression patterns may offer valuable insights for diagnostic and therapeutic purposes. Beyond this, investigating alternative preprocessing methods, such as normalization techniques tailored for specific experimental conditions, could be explored to enhance the accuracy of downstream analyses. Moreover, exploring the potential impact of different statistical approaches for differential expression analysis could offer valuable insights into the robustness of the findings.

In the realm of data visualization, the development of interactive and user-friendly tools tailored to biologists and clinicians could facilitate the interpretation of complex results. Creating visually informative representations, such as pathway enrichment maps, and interactive plots, can enhance the accessibility of findings and aid in the communication to diverse audiences.

REFERENCES

[1] Poewe, W., Seppi, K., Tanner, C. M., Halliday, G. M., Brundin, P., Volkmann, J., Schrag, A.-E., and Lang, A. E., "Parkinson disease," *Nature reviews Disease primers* **3**(1), 1–21 (2017).

[2] Dorsey, E., Sherer, T., Okun, M. S., and Bloem, B. R., "The emerging evidence of the parkinson pandemic," *Journal of Parkinson's disease* **8**(s1), S3–S8 (2018).

[3] Organization, W. H., "Parkinson's disease," (2019). Accessed: ¡date¿.

[4] Castonguay, A.-M., Gravel, C., and Levesque, M., "Treating parkinson's disease with antibodies: Previous studies and future directions," *Journal of Parkinson's Disease* **11**, 1–22 (10 2020).

[5] Cappelletti, C., Henriksen, S., Geut, H., Rozemuller, A., Vande Berg, W., Pihlstrøm, L., and Toft, M., "Transcriptomic profiling of parkinson's disease brains reveals disease stage specific gene expression changes," *Acta Neuropathologica* **146**, 1–18 (06 2023).

[6] Braak, H., Del Tredici, K., Rüb, U., De Vos, R. A., Steur, E. N. J., and Braak, E., "Staging of brain pathology related to sporadic parkinson's disease," *Neurobiology of aging* **24**(2), 197–211 (2003).

[7] Glaab, E., "Computational systems biology approaches for parkinson's disease," *Cell and Tissue Research* **373**, 1–19 (07 2018).

[8] Reel, P. S., Reel, S., Pearson, E., Trucco, E., and Jefferson, E., "Using machine learning approaches for multi-omics data analysis: A review," *Biotechnology Advances* **49**, 107739 (2021).

[9] Spratt, H. and Ju, H., [*Statistical Approaches to Candidate Biomarker Panel Selection*], vol. 919, 463–492 (12 2016).

[10] Irmady, K., Hale, C., Qadri, R., Fak, J., Simelane, S., Carroll, T., Przedborski, S., and Darnell, R., "Blood transcriptomic signatures associated with molecular changes in the brain and clinical outcomes in parkinson's disease," *Nature Communications* **14** (07 2023).

[11] Byron, S., Keuren-Jensen, K., Engelthaler, D., Carpten, J., and Craig, D., "Translating rna sequencing into clinical diagnostics: Opportunities and challenges," *Nature reviews. Genetics* **17** (03 2016).

[12] Joshua SK Bell, "The value of rna sequencing in drug discovery," (2023). Accessed: Nov 27, 2023.

[13] Peymani, F., Farzeen, A., and Prokisch, H., "Rna sequencing role and application in clinical diagnostic," *Pediatric Investigation* **6** (03 2022).

[14] Biswas, D., Shenoy, S. V., Chauhan, A., Halder, A., Ghosh, B., Padhye, A., Auromahima, S., Yadav, D., Sasmal, S., Dutta, S., Kumari, N., Bhavaskar, H., Mukherjee, A. P., Kumar, T. R., and Srivastava, S., "Brainprot(™) 3.0: Understanding human brain diseases using comprehensively curated & integrated omics datasets," *bioRxiv* (2023).

[15] Zaccaria, A., Antinori Malaspina, P., Licker, V., Kovari, E., Lobrinus, J., and Burkhard, P., "Multiomic analyses of dopaminergic neurons isolated from human substantia nigra in parkinson's disease: A descriptive and exploratory study," *Cellular and Molecular Neurobiology* **42** (09 2021).

[16] Dumitriu, A., Golji, J., Labadorf, A., Gao, B., Beach, T., Myers, R., Longo, K., and Latourelle, J., "Integrative analyses of proteomics and rna transcriptomics implicate mitochondrial processes, protein folding pathways and gwas loci in parkinson disease," *BMC Medical Genomics* **9** (01 2016).

[17] Kumar, K., Dubey, V., Zaidi, S. S., Tripathi, M., Siraj, F., Sharma, M. C., Chandra, P. S., Doddamani, R., Dixit, A. B., and Banerjee, J., "RNA sequencing of intraoperative peritumoral tissues reveals potential pathways involved in glioma-related seizures," *J. Mol. Neurosci.* **73**, 437–447 (June 2023).

[18] Jaffe, A., Hyde, T., Kleinman, J., Weinbergern, D., Chenoweth, J., McKay, R., Leek, J., and Colantuoni, C., "Practical impacts of genomic data "cleaning" on biological discovery using surrogate variable analysis," *BMC bioinformatics* **16**, 372 (11 2015).

[19] Catoni, C., Calì, T., and Brini, M., "Calcium, dopamine and neuronal calcium sensor 1: Their contribution to parkinson's disease," *Frontiers in Molecular Neuroscience* **12** (03 2019).

9. PUBLIC REFERENCES

1. Lashlock's DESeq2 R Tutorial
2. RNA-seq analyses pipeline for cancer research, by Albert Doughan
3. Differential expression with DEseq2 by Griffith Labs
4. RNA-Seq differential expression work flow using DESeq2, by STHDA
5. Bottom Up and Top Down Proteomics

APPENDIX

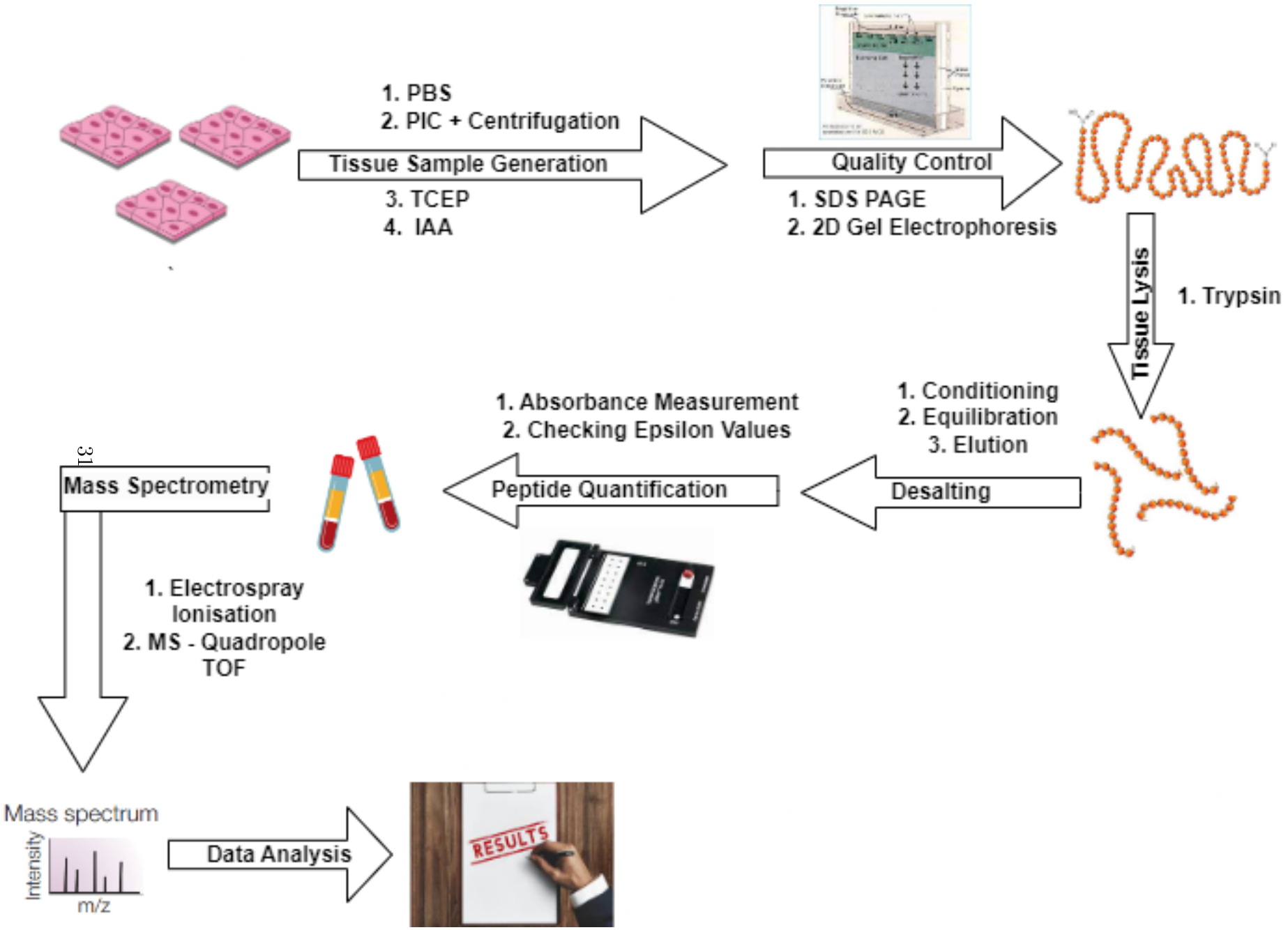


Figure 32: Schematic of steps in Bottom-Up Approach of Proteomics

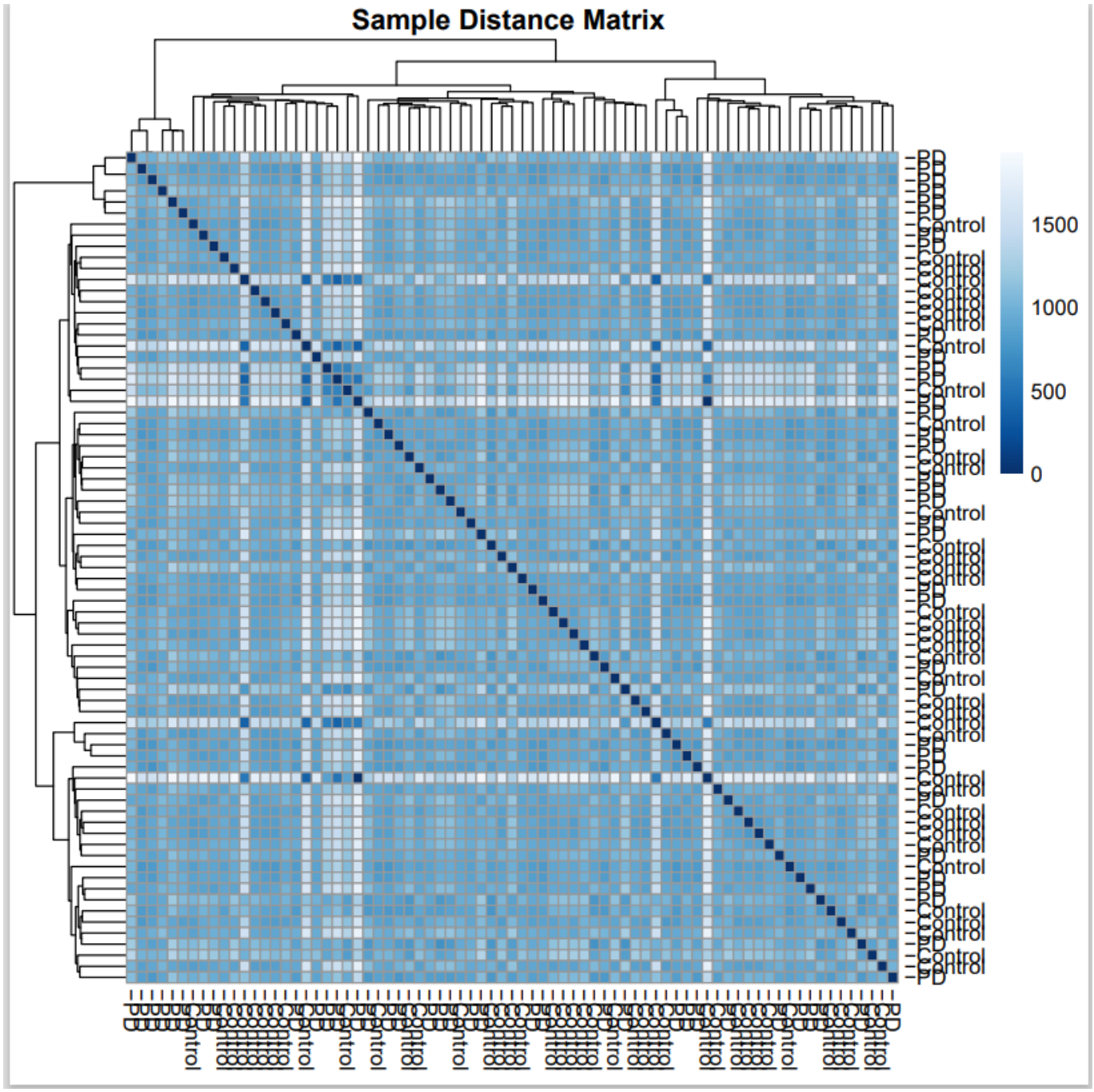


Figure 33: Heatmap to explore the sample-to-sample distance within the caudate data