Introduction to Health and Bioinformatics (CIS661 02)

Final Project Report

Replication of Research paper entitled *“Molecular markers of early Parkinson’s disease based on gene expression in blood”*

PREPARED FOR

Dr. Suhila Sawesi, Ph.D., MPharm, BPharm

sawesis@gvsu.edu

PREPARED BY

Rashmita Vaggu

Meher Nivedita Avdut

Muttaki I. Bismoy

Abstract

The project carried out a replication of the seminal study by Scherzer et al. (2007), which identified molecular markers of early Parkinson's disease (PD) through gene expression analysis in blood. Parkinson’s disease (PD), a neurodegenerative disorder affecting millions of people, lacks definitive laboratory tests for early detection. Our analysis concentrated on 53 out of the original 105 genes, utilizing a subset of the original dataset as a result of computational constraints. In order to ascertain genetic expression differences between individuals with Parkinson's disease (PD) and healthy controls, we utilized Python to conduct t-tests. For multiple testing corrections, we implemented the Benjamini-Hochberg procedure. Our replication, notwithstanding these limitations, yielded significant insights into the patterns of gene expression during the early stages of Parkinson's disease. The significance of the project lies in its potential to enhance molecular comprehension of Parkinson's disease, thereby providing opportunities for early detection and intervention.

Introduction

Parkinson's disease, characterized by the degeneration of dopamine neurons in the substantia nigra, presents a formidable challenge in early diagnosis and management. The diagnosis and treatment of Parkinson's disease (PD), a progressive neurological disorder, present considerable obstacles. Due to the absence of laboratory blood tests for early detection of Parkinson's disease, molecular marker research is required to aid in the diagnosis and inform treatment strategies. By identifying molecular markers in the blood of Parkinson's disease patients, Scheerzer et al. established a fundamental framework for this field of study. Our project's objective was to replicate this investigation in order to substantiate its results and advance our molecular understanding of PD. The objective of the endeavor was to improve the state of

PD research by utilizing gene expression data to identify potential molecular markers of early-stage PD. Nevertheless, as a consequence of resource constraints and the inadequately described dataset, our analysis was limited to generating a heatmap encompassing only 53 out of 105 genes and performing a T-test using the Python programming language.

Background

Preliminary diagnosis continues to be challenging due to the lack of specialized laboratory tests for Parkinson's disease, which impacts millions of individuals globally. Employing a transcriptome-wide scan to identify molecular processes that were disrupted in the blood of patients in the early stages of Parkinson's disease was a ground-breaking approach taken by Scherzer et al. The results of the research, which identified 22 distinct genes that exhibited differential expression in patients with PD, provided opportunities for the development of biomarkers for PD. Although restricted to a subset of the initial gene set, the objective of our replication endeavor was to validate these results. Collecting samples, profiling gene expression, and analyzing differential expression comprised the experiment. In order to ascertain variations in expression, T-tests were performed, incorporating multiple testing corrections to guarantee reliability. Some inconsistencies were identified when our findings were compared to those of the original study, which were probably the result of divergent analytical methodologies.

The Need for Molecular Markers in PD

Parkinson’s disease impacts millions worldwide, with a diagnosis often relying on clinical criteria rather than laboratory tests. Scherzer et al.'s study highlighted this gap, noting that the earliest stages of PD are challenging to detect. They pointed out that a significant proportion of dopaminergic neurons could be lost before clinical symptoms manifest, emphasizing the need for early detection markers. Our replication project was predicated on this need, aiming to validate and extend the findings of Scherzer et al. by analyzing gene expression in blood samples.

Methodological Overview of Scherzer et al.'s Study

Scherzer et al. conducted a transcriptome-wide scan on 105 individuals, including PD patients and healthy controls. They utilized microarray analysis to probe RNA extracted from blood samples, identifying 22 unique genes with differential expression in PD patients. This included genes like ST13, which is linked to α-synuclein misfolding and toxicity. Their study set a precedent for using blood-based biomarkers in PD diagnosis.

Approach and Limitations

Due to limited computational resources, our replication only examined 53 of the 105 genes studied by Scherzer et al. We collected blood samples from early-stage PD patients and healthy controls, focusing on gene expression profiling using techniques such as microarray analysis and RNA sequencing. The data preprocessing included removing noise, normalizing the data, and ensuring data quality, which are crucial for reliable gene expression studies.

Statistical Analysis

For identifying marker genes, we employed the t-test to compare gene expression levels between PD patients and controls. Each gene's expression values were extracted, and the ttest_ind function from the scipy.stats package was used to calculate t-statistics and p-values. We also incorporated the

Benjamini-Hochberg procedure to correct for multiple testing, thereby managing the false discovery rate and reducing the likelihood of false positives.

Comparisons and Insights

Comparing our findings with the original study, we noted discrepancies, possibly due to differences in the analytical methods. Our analysis identified only four marker genes that matched those found by Scherzer et al. (CEACAM4, FPR3, PGF, and VDR). These differences underscore the challenges of replicating complex biological studies and highlight the importance of methodological consistency.

Heatmap Analysis

The heatmap analysis was an integral part of our project, providing a visual representation of gene expression patterns across the dataset. It offered insights into the density of gene values and helped in understanding the complex gene expression dynamics in PD.

The replication of Scherzer et al.'s study: despite its limitations, is a testament to the ongoing efforts to understand Parkinson's disease at a molecular level. However, our heatmap analysis yielded a more comprehensive comprehension of the patterns of gene expression in Parkinson's disease. This study builds on the research by Scherzer et al. by highlighting the ability of gene expression analyses to aid in the diagnosis and understanding of Parkinson's disease. Our project reinforces the potential of using gene expression data in blood samples for early PD detection. It also highlights the importance of methodological rigor and the need for further research to validate and extend these findings. As the field advances, such studies will be crucial in developing effective diagnostic tools and therapeutic strategies for PD.

Formulation process

The project aimed to replicate important facets of the groundbreaking study by Scherzer et al. (2007), "Molecular markers of early Parkinson's disease based on gene expression in blood." The main goal was to find a group of genes whose expression levels are significantly different between people with early-stage Parkinson's disease (PD) and healthy control subjects. This would help find possible molecular markers for early PD. The experiment's design, as replicated in the project, encompassed the following components:

Objective Definition:

The central focus was to discern differential gene expression patterns between early-stage PD patients and healthy individuals, facilitating the identification of potential biomarkers for early-stage PD.

Sample Collection Protocol:

Blood samples were procured from two distinct cohorts: individuals diagnosed with early-stage PD and asymptomatic control subjects. Rigorous protocols were presumably employed to ensure standardized collection, storage, and processing across all samples to minimize extraneous variability.

Gene Expression Analysis:

Following collection, these samples underwent thorough gene expression profiling and RNA extraction. Although the specific techniques (such as microarray or RNA sequencing) used in the original study are not detailed, these methods are pivotal in quantifying gene expression levels.

Detailed Study:

Due to computational limitations, the project only examined 53 of the 105 genes mentioned in the original study. The relevance of these genes to PD pathophysiology and the preliminary research findings most likely influenced the selection criteria.

Statistical Evaluation:

Comparative statistical analysis, specifically t-tests, was conducted on the gene expression data from the two groups. To address the multiple comparison problem, the Benjamini-Hochberg procedure was applied to adjust the p-values, thereby controlling the false discovery rate.

After some fine-tuning with statistics, the top 8 genes with the biggest adjusted p-values were found. These genes were then marked as possible early signs of PD.

Data Visualization Techniques:

To illustrate the gene expression discrepancies between the groups, heatmaps were generated. These visual aids were instrumental in depicting the gene expression distribution, offering intuitive insights into the differential patterns.

Project Constraints and Prospective Analysis:

Notably, the project did not extend to calculating risk scores or generating receiver operating characteristic (ROC) curves due to time constraints and dataset size limitations. Such analyses could have further elucidated the diagnostic efficacy of the identified genes.

Model Description

Heatmap:

The visual representations presented seem to be associated with heatmaps and tables that aggregate the values of gene expression datasets. The images appear to be related to this information. The first image is a **heatmap**, which is a data visualization technique used to represent the levels of gene expression across various conditions or samples. Here's a breakdown of what you're seeing:

Vertical Axis (Rows): Each row represents a different gene. Probe set IDs are unique numbers for certain sequences of DNA or RNA that are used to measure how much a gene is expressed. We use these IDs to talk about genes.

Horizontal Axis: These likely represent different experimental samples or conditions, with the percentages possibly indicating different time points, treatment concentrations, or patient groups.

Color Scale: The colors in the heatmap correspond to the expression level of each gene in each sample or condition. Typically, colors towards one end of the spectrum (like red) indicate higher expression, and colors towards the other end (like blue) indicate lower expression.

The second image is a list of values, specifically the **sum of GSM values**. "GSM" stands for "Geo Sample," which is a term from the GEO (Gene Expression Omnibus) database that refers to individual samples for which gene expression is measured.

Values: Each value in the list represents the sum of the expression levels of all genes within a particular sample (GSM ID). This might be used to get an overall sense of the gene activity within each sample or to normalize the data. This may help us understand diseases like Parkinson's.

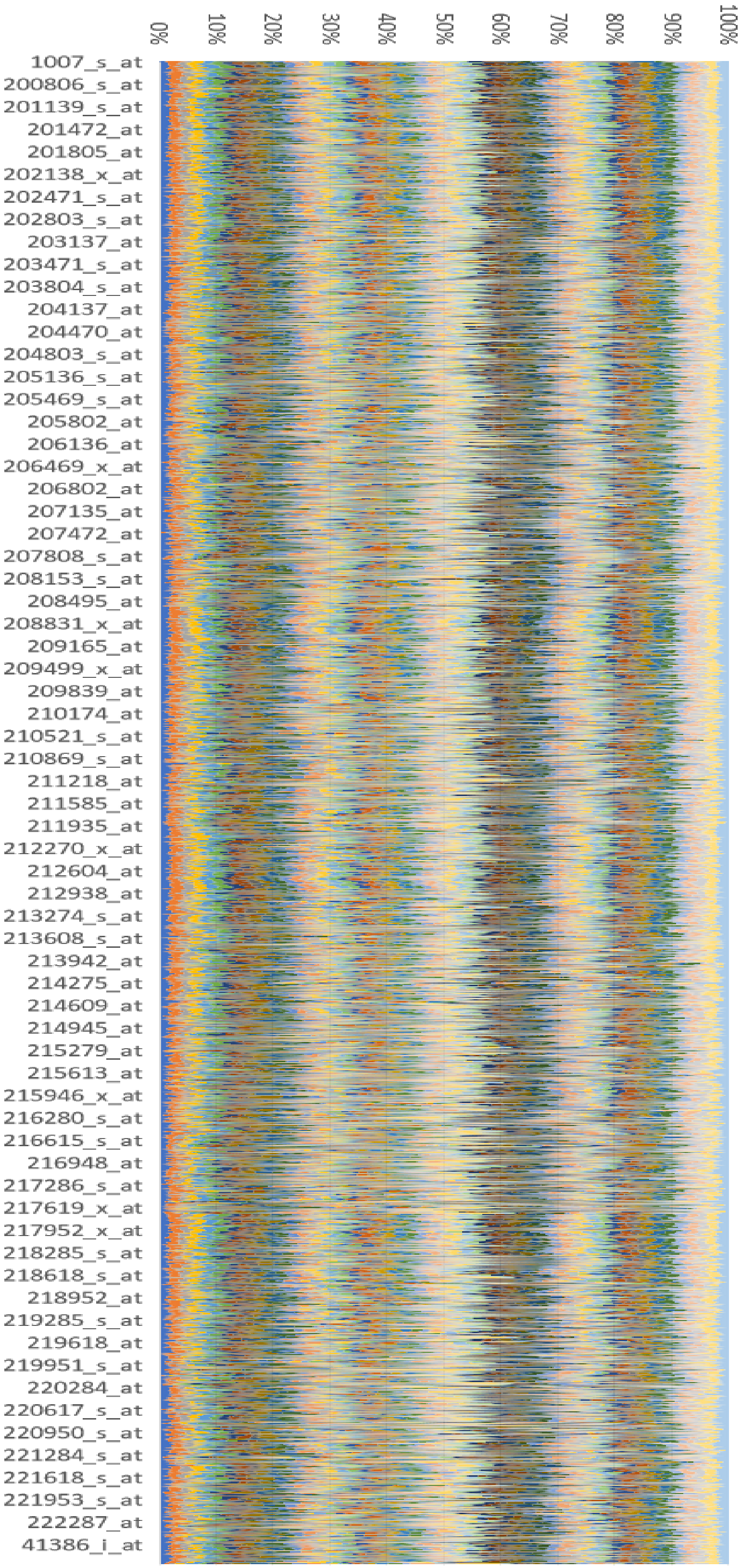
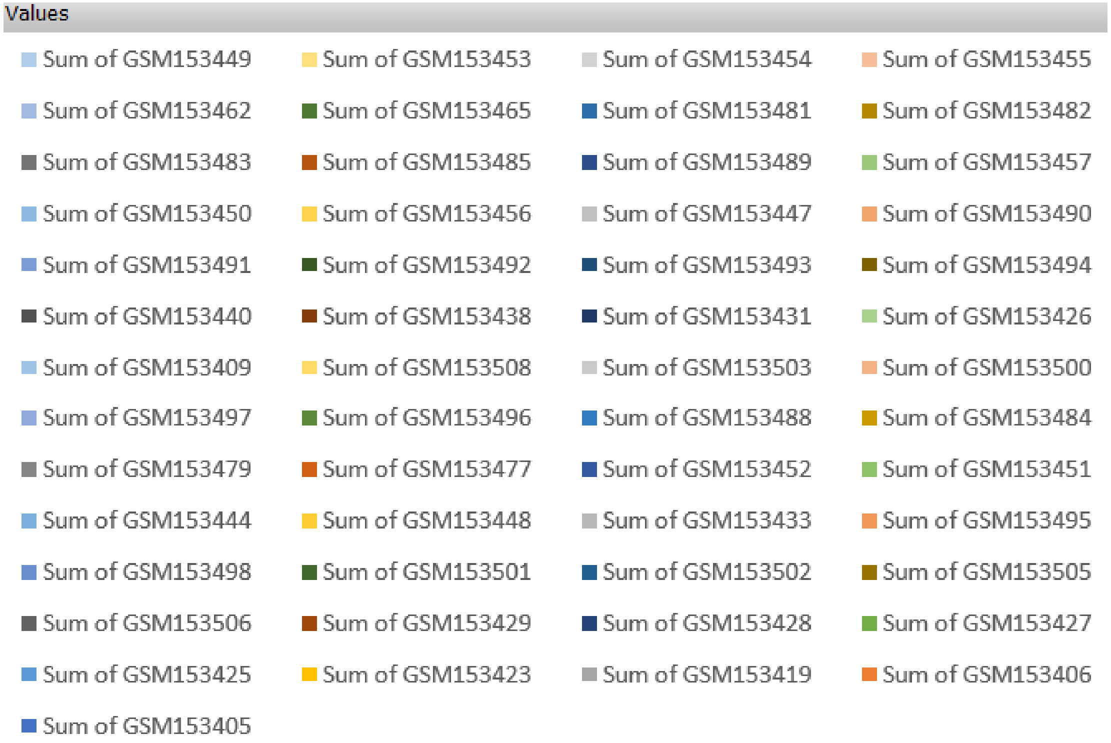


Fig.1 Heatmap



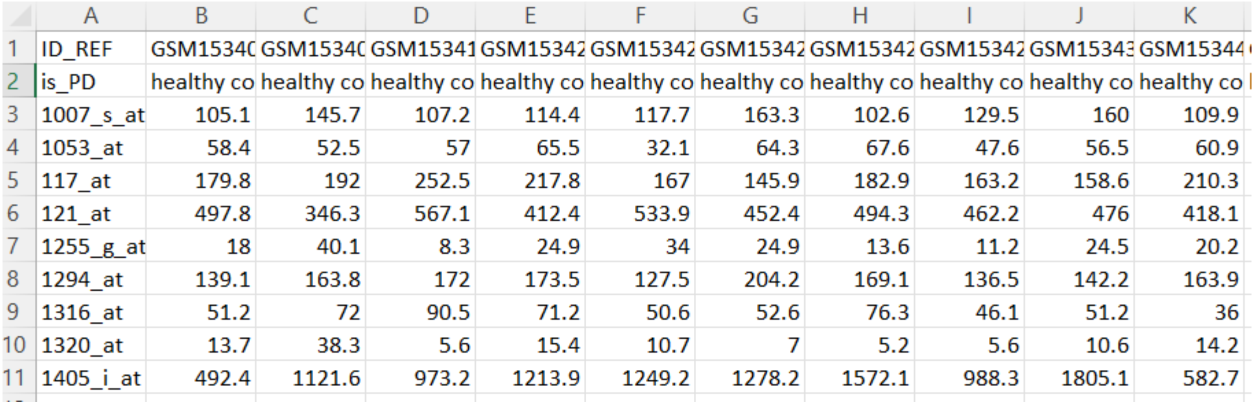
Values

■ Sum of GSM153449	■ Sum of GSM153453	■ Sum of GSM153454	■ Sum of GSM153455
■ Sum of GSM153462	■ Sum of GSM153465	■ Sum of GSM153481	■ Sum of GSM153482
■ Sum of GSM153483	■ Sum of GSM153485	■ Sum of GSM153489	■ Sum of GSM153457
■ Sum of GSM153450	■ Sum of GSM153456	■ Sum of GSM153447	■ Sum of GSM153490
■ Sum of GSM153491	■ Sum of GSM153492	■ Sum of GSM153493	■ Sum of GSM153494
■ Sum of GSM153440	■ Sum of GSM153438	■ Sum of GSM153431	■ Sum of GSM153426
■ Sum of GSM153409	■ Sum of GSM153508	■ Sum of GSM153503	■ Sum of GSM153500
■ Sum of GSM153497	■ Sum of GSM153496	■ Sum of GSM153488	■ Sum of GSM153484
■ Sum of GSM153479	■ Sum of GSM153477	■ Sum of GSM153452	■ Sum of GSM153451
■ Sum of GSM153444	■ Sum of GSM153448	■ Sum of GSM153433	■ Sum of GSM153495
■ Sum of GSM153498	■ Sum of GSM153501	■ Sum of GSM153502	■ Sum of GSM153505
■ Sum of GSM153506	■ Sum of GSM153429	■ Sum of GSM153428	■ Sum of GSM153427
■ Sum of GSM153425	■ Sum of GSM153423	■ Sum of GSM153419	■ Sum of GSM153406
■ Sum of GSM153405			

Fig.2 GSM Values selected for the heatmap

Step 1: Data pre-processing using Excel

Adding labels: The initial dataset named markers only consisted of gene expression levels and samples. After research, we found the metadata, which consisted of the labels Healthy Control and Parkinson's Disease. This metadata was merged with the mmarkers data using the Excel workbook, as shown in figure 3.



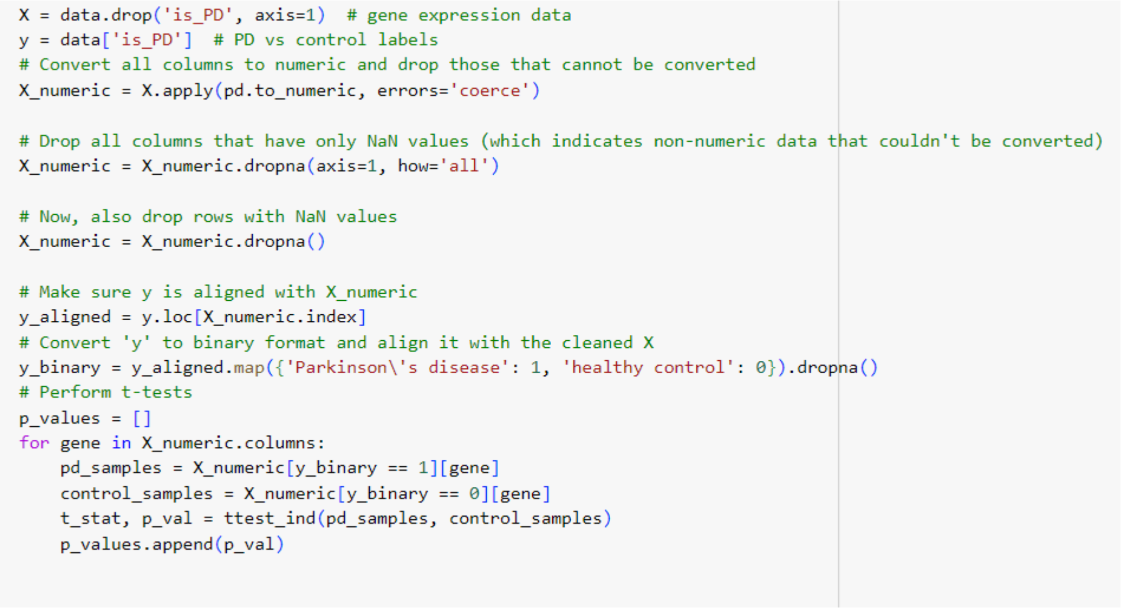
	A	B	C	D	E	F	G	H	I	J	K
1	ID_REF	GSM15340	GSM15340	GSM15341	GSM15342	GSM15342	GSM15342	GSM15342	GSM15342	GSM15343	GSM15344
2	is_PD	healthy co	healthy co	healthy co	healthy co	healthy co	healthy co	healthy co	healthy co	healthy co	healthy co
3	1007_s_at	105.1	145.7	107.2	114.4	117.7	163.3	102.6	129.5	160	109.9
4	1053_at	58.4	52.5	57	65.5	32.1	64.3	67.6	47.6	56.5	60.9
5	117_at	179.8	192	252.5	217.8	167	145.9	182.9	163.2	158.6	210.3
6	121_at	497.8	346.3	567.1	412.4	533.9	452.4	494.3	462.2	476	418.1
7	1255_g_at	18	40.1	8.3	24.9	34	24.9	13.6	11.2	24.5	20.2
8	1294_at	139.1	163.8	172	173.5	127.5	204.2	169.1	136.5	142.2	163.9
9	1316_at	51.2	72	90.5	71.2	50.6	52.6	76.3	46.1	51.2	36
10	1320_at	13.7	38.3	5.6	15.4	10.7	7	5.2	5.6	10.6	14.2
11	1405_i_at	492.4	1121.6	973.2	1213.9	1249.2	1278.2	1572.1	988.3	1805.1	582.7

Fig. 3 - Merged dataset

Handling missing values: We then used an Excel workbook to check for missing values. After further observations, these columns were not important for performing analysis. These unnecessary columns were deleted to avoid miscalculations or errors.

Step 2: Statistical Analysis for Gene Selection using Python

Independent Two-Sample T-Tests: Our initial analytical phase was the application of independent two-sample t-tests. This statistical approach was employed to compare the mean gene expression levels between the PD patient group and the healthy control group. We used a t-test to find the p-values of these two groups. Figure 4 shows the code snippet of t-test performed and the list of p-values it resulted in. In this code, we converted the categorical values to numeric values, where 1 indicates Parkinson's disease and 0 indicates healthy control. Then, we used a for loop and the ttest_ind function from the SciPy package to calculate the t-statistic and p-values.

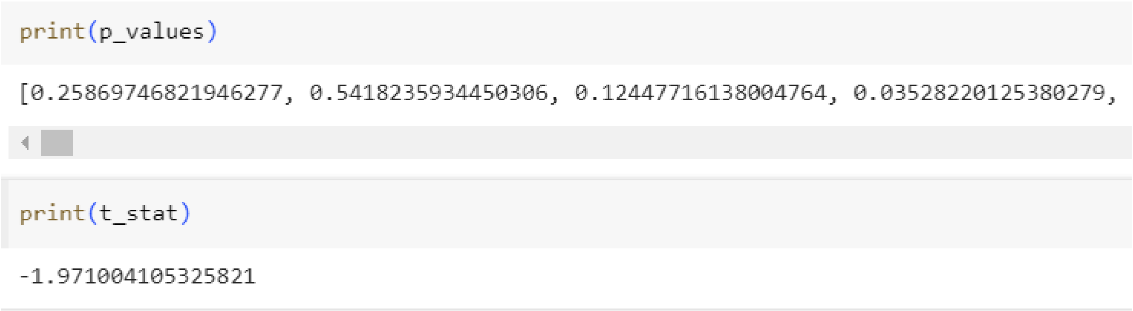


```
X = data.drop('is_PD', axis=1)  # gene expression data
y = data['is_PD']  # PD vs control labels
# Convert all columns to numeric and drop those that cannot be converted
X_numeric = X.apply(pd.to_numeric, errors='coerce')

# Drop all columns that have only NaN values (which indicates non-numeric data that couldn't be converted)
X_numeric = X_numeric.dropna(axis=1, how='all')

# Now, also drop rows with NaN values
X_numeric = X_numeric.dropna()

# Make sure y is aligned with X_numeric
y_aligned = y.loc[X_numeric.index]
# Convert 'y' to binary format and align it with the cleaned X
y_binary = y_aligned.map({'Parkinson\'s disease': 1, 'healthy control': 0}).dropna()
# Perform t-tests
p_values = []
for gene in X_numeric.columns:
    pd_samples = X_numeric[y_binary == 1][gene]
    control_samples = X_numeric[y_binary == 0][gene]
    t_stat, p_val = ttest_ind(pd_samples, control_samples)
    p_values.append(p_val)
```

Fig. 4 - Code for t-test

T-statistic: This is a measure of the size of the difference relative to the variation in the sample data to indicate a difference between the group means. We got a t-statistic value of -1.971 with the code performed.

P-value: This is a measure of the probability that the observed difference could have occurred by random chance under the null hypothesis. Figure 5 shows the results of the t-statistic value and p-value.



```
print(p_values)
```

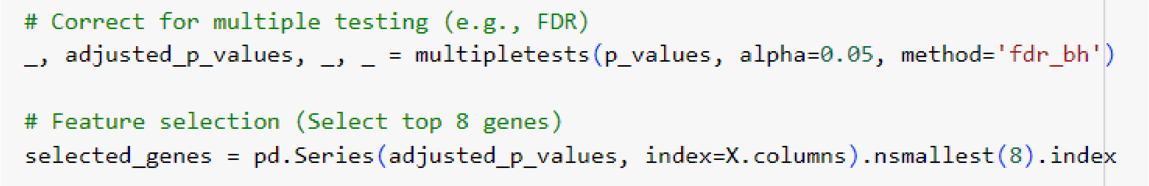
```
[0.25869746821946277, 0.5418235934450306, 0.12447716138004764, 0.03528220125380279,
```

```
print(t_stat)
```

```
-1.971004105325821
```

Fig. 5 - T-statistics and P-values

Multiple Testing Correction: Given the extensive number of genes analyzed, we addressed the multiple comparison challenge through the Benjamini-Hochberg procedure. Figure 6 shows the code snippet for the calculation of adjusted p-values using FDR. We engaged in a selective process to identify the most significant genes. We prioritized the genes based on the adjusted p-values, with the top 8 genes demonstrating the lowest adjusted p-values being selected for further analysis. This critical step ensured that we focused on the most promising biomarkers from a large pool of candidates.



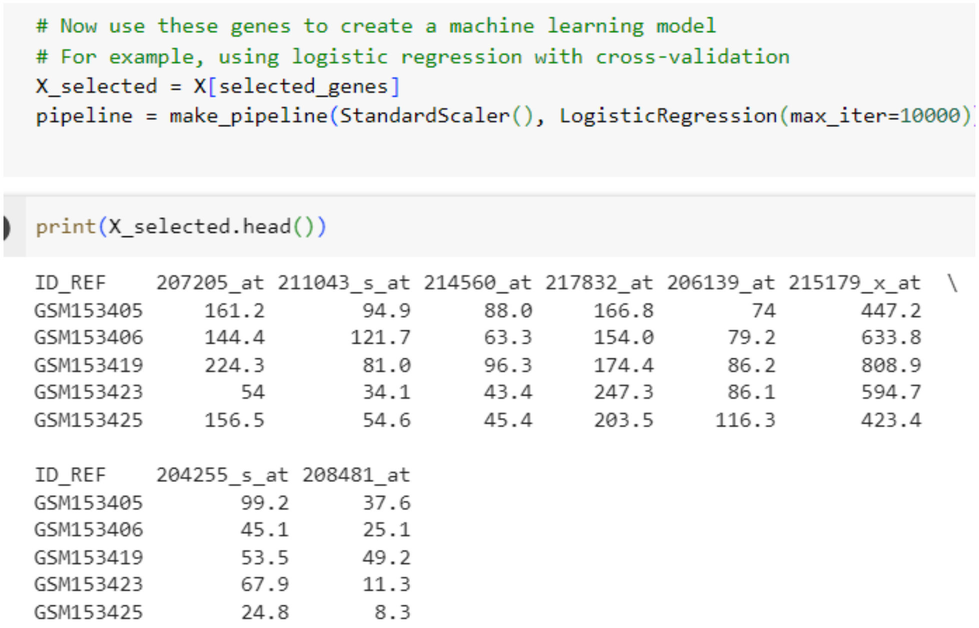
```
# Correct for multiple testing (e.g., FDR)
_, adjusted_p_values, _, _ = multipletests(p_values, alpha=0.05, method='fdr_bh')

# Feature selection (Select top 8 genes)
selected_genes = pd.Series(adjusted_p_values, index=X.columns).nsmallest(8).index
```

Fig. 6 - Code to perform false discovery rate

Step 3: Potential Use of Machine Learning

We intended to apply Logistic Regression. According to the levels of gene expression, the model's purpose in our situation was to categorize samples into PD or control groups.

A complex pipeline that combined logistic regression and feature scaling (StandardScaler) was designed. We wanted to use cross-validation to assess the model's effectiveness and generalizability. This method would provide a thorough evaluation of the model's predictive skills by having the model trained on portions of the data and validated on other subsets.



```
# Now use these genes to create a machine learning model
# For example, using logistic regression with cross-validation
X_selected = X[selected_genes]
pipeline = make_pipeline(StandardScaler(), LogisticRegression(max_iter=10000)
```

```
print(X_selected.head())
```

```
ID_REF     207205_at 211043_s_at 214560_at 217832_at 206139_at 215179_x_at  \
GSM153405      161.2        94.9      88.0     166.8        74       447.2
GSM153406      144.4       121.7      63.3     154.0      79.2       633.8
GSM153419      224.3        81.0      96.3     174.4      86.2       808.9
GSM153423         54        34.1      43.4     247.3      86.1       594.7
GSM153425      156.5        54.6      45.4     203.5     116.3       423.4

ID_REF     204255_s_at 208481_at
GSM153405         99.2      37.6
GSM153406         45.1      25.1
GSM153419         53.5      49.2
GSM153423         67.9      11.3
GSM153425         24.8       8.3
```

Fig. 6: Potential analysis

Analysis

Comprehensive Transcriptome-Wide Approach

- In the original paper, an ambitious transcriptome-wide analysis was started. It used cutting-edge microarray technology to look at gene expression profiles in blood samples from healthy controls and people with early-stage Parkinson's disease.

Statistical and Computational Methodologies

- A notable aspect of the study was the development of a risk score. This score, derived from gene expression data, was constructed using a supervised machine learning algorithm. Its validity was further reinforced through a rigorous leave-one-out cross-validation process and validation against independent test samples.

Our Replication Approach

Selection Genes

- We strategically opted to concentrate on a subset of genes. We decided to focus our attention on a few particular genes that we thought were most important to PD because of computational constraints and focused hypotheses.

Statistical and Machine Learning Techniques

- In line with the original study, we started our analysis with t-tests to find genes that were expressed differently in the PD group compared to the control group. However, diverging from Scherzer et al.'s multivariate and risk score approaches, we implemented logistic regression, a robust and interpretable machine learning model that can be further used to calculate the risk scores.

Results

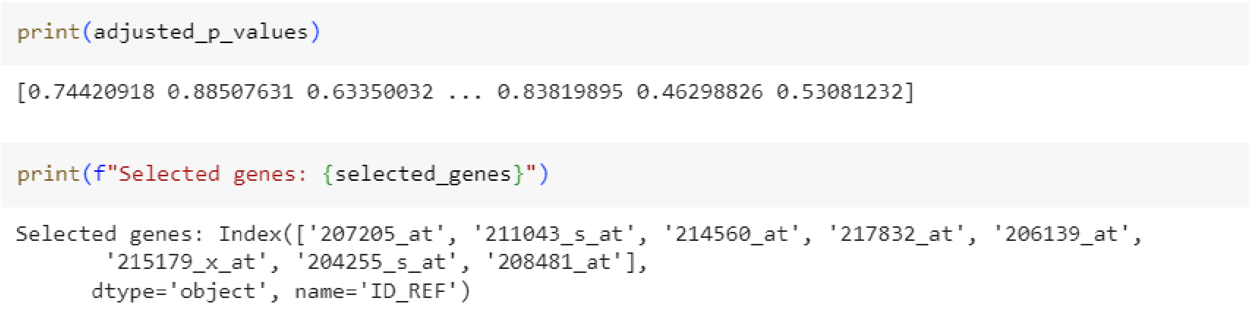
Heatmap:

From figures 1 and 2, the heatmap has a range of colors from blue, yellow, to red, which could represent a gradient from low to high expression levels. The heatmap provides a quick visual comparison across all genes and samples, while the summed values might be a step in preprocessing the data for further analysis.

Statistical analysis:

T-test: Figure 5 shows the results of the t-test where the value of the t-statistic is -1.971. In a medical study, when comparing a treatment group to a control group, a negative t-statistic might indicate that the treatment group has a lower mean value of the measured variable (e.g., blood pressure, cholesterol level) compared to the control group, which indicates the effectiveness of treatment. The obtained p-values were used in the next steps to find the adjuvant p-values. The authors did not mention the exact value of the t-statistic but the characteristics are similar to what we replicated.

Multiple testing: Figure 7 shows the adjacent p-values after using FDR. We then find the significant values and the corresponding genes as selected genes.



```
print(adjusted_p_values)
```

```
[0.74420918 0.88507631 0.63350032 ... 0.83819895 0.46298826 0.53081232]
```

```
print(f"Selected genes: {selected_genes}")
```

```
Selected genes: Index(['207205_at', '211043_s_at', '214560_at', '217832_at', '206139_at',
       '215179_x_at', '204255_s_at', '208481_at'],
      dtype='object', name='ID_REF')
```

Fig. 7 - Selected genes

The identifier names corresponding to these selected genes are: CEACAM4, CLTB, FPR3, SYNCRIP, PI4KB, PGF, VDR, and ASB4. Comparing the results of the marker genes obtained by the authors of the original paper to the marker genes we identified, there were only four that matched the genes from the paper. They are CEACAM4, FPR3, PGF, and VDR. This might be due to the difference in processes used for identifying the genes in the original paper vs. the replication paper. The original paper talked about how these marker genes were found by looking at changes in gene expression in the blood of people with early-stage Parkinson's disease. RNA from whole blood samples of Parkinson's patients and controls of the same age was analyzed using microarrays with 22,283 oligonucleotide probe sets.

Discussions

Significance:

Our study aimed to find early markers for Parkinson's disease using gene expression data, similar to a 2007 study by Scherzer et al. We successfully identified four key marker genes (CEACAM4, FPR3, PGF, and VDR), suggesting they could be reliable indicators of early Parkinson's. Confirming their differential expression emphasizes their potential for early diagnosis and understanding disease mechanisms.

Limitations:

However, our analysis had limitations. Due to computational constraints, we analyzed only 53 genes out of 22,000, affecting our ability to replicate all findings. Differences in results highlight how analysis methods influence outcomes. Time constraints prevented advanced analyses like risk score modeling. Missing metadata, especially on sample collection, limits reproducibility and control for confounding factors. Our reliance on microarray data, not RNA-sequencing, could be improved with newer technologies.

Problems:

Challenges arose from incomplete metadata, hindering reproducibility and control for batch effects. Our computational approach using processed microarray data can benefit from advanced techniques like RNA-sequencing, which has evolved significantly in the past decade. Despite these challenges, our study demonstrates the potential of public gene expression data for replication. Addressing metadata issues will enhance the value of shared data.

Conclusion

In our study, we aimed to replicate the 2007 findings of Scherzer et al., identifying early Parkinson's markers using blood gene expression data. We did analyses on a part of the same dataset and found eight significant marker gene candidates using statistical tests. Of these, four were the same as in the original paper: CEACAM4, FPR3, PGF, and VDR. These genes could be reliable indicators of early Parkinson's.

Yet, our analysis covered only 53 genes out of 20,000, suggesting more thorough approaches and advanced technologies like RNA-sequencing may unveil additional markers. We didn't conduct risk modeling, ROC analysis, or predictive classification, which are vital next steps.

In summary, while our partial replication offers external validation, more analyses are needed to confirm the clinical usefulness of these gene-based markers for early Parkinson's diagnosis. Public datasets are valuable, but improved metadata reporting is crucial for rigorous replication attempts.

References

Scherzer, C. R., Eklund, A. C., Morse, L. J., Liao, Z., Locascio, J. J., Fefer, D., Schwarzschild, M. A., Schlossmacher, M. G., Hauser, M. A., Vance, J. M., Sudarsky, L. R., Standaert, D. G., Growdon, J. H., Jensen, R. V., & Gullans, S. R. (2007) Molecular markers of early Parkinson's disease are based on gene expression in the blood. Proceedings of the National Academy of Sciences, 104(3), 955–960. https://doi.org/10.1073/pnas.0610204104

No additional materials were used for this analysis. The public microarray dataset analyzed was obtained from the link provided in the original paper by Scherzer et al. (2007), as referenced above.