PH240C Final Project

Protein Expression across Disease Groups: Analysis of the Human Brain Proteome

Ramasubramanian Balasubramanian, John Canty, Annan Deng, William Krinsman

December 6th, 2018

Abstract

Contents

1 Introduction

Patients with neurodegenerative disorders such as Alzheimers disease (AD) and Parkinsons disease (PD) share common physiological and symptomatic similarities, however the molecular pathways linking these diseases are incompletely understood. Recently, a comprehensive quantitative proteomic dataset of the human brain in patients with AD, PD, and AD/PD comorbidity has been publicly released. We will perform exploratory data analysis and modelling in order to identify if there are disease-specific differential gene expression patterns that are associated with AD and PD neuropathologies. We will then perform Weighted Correlation Network Analysis (WCNA) of the identified genes to assess if there coexpression relationships between subsets of these genes. Finally, using the PANTHER gene ontology classification system, we will assess whether these genes exhibit functional relationships.

2 Dataset

The dataset[3] that we will be evaluating was obtained by performing Tandem Mass Tag (TMT) isobaric mass spectrometry on brain tissue obtained from individual human donors. Brain tissue samples were obtained from 40 individual patient samples across two separate brain regions, the Frontal Cortex and the Anterior Cingulate Gyrus). The dataset obtained by using a factorial experimental design, where there are 5 experimental batches, with two samples of each phenotype (Control Patient, Alzheimers Patient, Parkinsons Patient, Alzheimers/Parkinsons Patient) within each batch. For each batch, a control sample was generated by pooling fractions of all samples in order to generate a global internal standard (GIS) expression measure. Individual data entries are displayed as base-10 log-transformed ratios of the obtained peptide counts for the ijth sample covariate of interest and the ith covariate of the GIS standard. From the Frontal Cortex, $10,100$ unique protein groups were identified, while from the Anterior Cingulate Gyrus, $10,695$ protein groups were identified.

3 Exploratory analysis and preprocessing

In the initial portion of this project, we will assess data quality, phenotype and inter-batch variability by box-plot visualization. We will then assess the raw data and apply quantile normalization samples. Finally, we assess the whether or not disease-trait groups can be discriminated based on gene expression values by applying dimensionality reduction methods.

1. Data quality and Preprocessing

We performed several quality control steps on the data-set before analysis. A large fraction of the gene rows contained batches that consisted of entirely NA or zero values that needed to be removed. We attribute these values to commonly reported issues of label-free mass spectrometry experiments that involve probe bias and reduced coverage of peptides. In order to preserve genes that contained peptide counts for all batches, we removed any gene that had NAs present in any batch. Next, we repeated the process by removing any gene that contained zeros, since these values would result in undefined log-transformed expression measures. In the Frontal Cortex, this process resulted in a final gene count of $6,004$ from an original dataset of $10,100$ genes. In the Anterior Cingulate Gyrus, this resulted in a final gene count of 6230 genes.

2. Exploratory analysis

In order to assess our processed data, we visualized the distribution of all fold-change (FC) gene expression values for each patient sample across all disease traits (Ctl, AD, PD, and ADPD) **(Fig.1A)**. Upon inspection of the plots, we noticed that all samples had FC gene expression values that were centered around zero, indicating that the experimental effects between batches were significantly reduced by normalizing against a GIS for each batch. However, for a fraction of patient samples, we still observed significant FC gene expression outliers.

Next, we performed Principle Component Analysis (PCA) on the samples in both the Frontal Cortex and the Anterior Cingulate Gyrus **(Fig.2A and B)**. A PCA plot of all disease traits did not indicate any clear separation of classes in both tissues. In order to simplify visualization, we then generated pairwise PCA plots between the Control samples and disease-state samples in both tissues **(Fig.2C and D)**. In both tissue samples, we observed that clear separation of the Control and AD samples. Similarly, we observed separation of the Control and comorbid ADPD samples. However, when when we analyzed the Control and PD samples, we did not observe clear separation between the two traits.

We then quantitatively assessed the separation between the Control and disease-state samples by applying a Support Vector Machine (SVM) classifier to the PCA plots **(Fig.2 E and F)**. Consistent with our observations from the PCA plots, we observed that the SVM was able to accurately classify Control samples versus AD and comorbid PDAD samples. However, SVM classifier performed substantially worse with PD samples. Based on these observations, we concluded that the PD samples do not exhibit noticeable variances in gene expression compared to Control samples.

4 Linear modelling to identify differential expression

In order to identify differentially expressed genes between Control samples and disease-state samples, we utilized linear models. We samples as covariates (with parameter coefficientsfor the control, AD, PD, and AD/PD samples) in order to determine the strength of association between each gene and the disease phenotypes. We used two separate linear modelling approaches. First, we modelled each trait as additive categorical random variables:

$$E[Y^i|X^i] = \beta_{Ctl} + \beta_{AD}x^i_{AD} + \beta_{PD}x^i_{PD} + \beta_{ADPD}x^i_{ADPD}$$

Where the $x^i s$ are indicator variables for the presence or absence of each disease-trait in a sample.Using this approach, we performed linear modelling on each gene, and obtained sample estimators $\hat{\beta}_{Ctl}, \hat{\beta}_{AD}, \hat{\beta}_{PD}, and\hat{\beta}_{ADPD}$. We also considered that the changes in gene expression values in the ADPD case may not simply reflect additive changes in from both AD and PD in a linear model. To model this, we considered the comorbid case as an interaction effect between both the AD and PD traits:

$$E[Y^i|X^i] = \beta_{Ctl} + \beta_{AD}x^i_{AD} + \beta_{PD}x^i_{PD} + \beta_{ADPD}x^i_{AD}x^i_{PD}$$

Repeating the same analysis as in the additive case, we obtained sample estimators for the AD, PD, and ADPD cases.

1. Volcano plots to identify differential expression

After obtaining sample estimators $\hat{\beta}_{Ctl}, \hat{\beta}_{AD}, \hat{\beta}_{PD}$, and $\hat{\beta}_{ADPD}$ for each gene in the in Frontal Cortex and Anterior Cingulate Cyrus, we then computed the the log-fold change and p-value of the fold-change. We performed this calculation for each gene with the disease-state sample estimators against the Control sample estimator:

$$\begin{array}{lr}
log_2FC(\hat{\beta}_{Ctl}, \hat{\beta}_{AD}) & Pval(\hat{\beta}_{Ctl}, \hat{\beta}_{AD}) \\
log_2FC(\hat{\beta}_{Ctl}, \hat{\beta}_{PD}) & Pval(\hat{\beta}_{Ctl}, \hat{\beta}_{PD}) \\
log_2FC(\hat{\beta}_{Ctl}, \hat{\beta}_{ADPD}) & Pval(\hat{\beta}_{Ctl}, \hat{\beta}_{ADPD})
\end{array}$$

Next, we plotted our results using volcano plots. We defined a differential expressed gene as any gene expression value where the p-value was less than $\frac{0.05}{nGenes}$ and the log_2 fold-change was larger than 0.25 **(Fig 3.A and B)**.

5 Weighted correlation network analysis

Using the candidate differentially expressed genes for the PD, AD, and PD/AD disease phenotypes, we will apply weighted correlation network analysis in order to determine if the candidate genes exhibit interesting coexpression patterns. First, using the available WGNA R-package[1], we will determine whether there are interesting coexpression clusterings between differentially expressed genes for each phenotype. Next, we will assess the clustering size/structure and compare these properties between phenotypes.

1. Analysis of coexpression clusterings

(a) Size and structure of clusterings

(b) Comparison of clusters between phenotypes

6 Biological functions of differentially expressed genes

Using the PANTHER gene ontology database, we identify the functional classifications of the differential expressed genes. We will perform this analysis between all phenotypes in order to comment on genes that may have unique biological roles related to PD, AD, and PD/AD disease states.

1. PANTHER gene ontology database

The PANTHER project, which we used to identify the biological functions of the differentially expressed genes, has the following purpose[2]:

> The PANTHER (protein annotation through evolutionary relationship) classification system (`http://www.pantherdb.org/`http://www.pantherdb.org/) is a comprehensive system that combines gene function, ontology, pathways and statistical analysis tools that enable biologists to analyze large-scale, genome-wide data from sequencing, proteomics or gene expression experiments. The system is built with 82 complete genomes organized into gene families and subfamilies, and their evolutionary relationships are captured in phylogenetic trees, multiple sequence alignments and statistical models (hidden Markov models or HMMs). Genes are classified according to their function in several different ways: families and subfamilies are annotated with ontology terms (Gene Ontology (GO) and PANTHER protein class), and sequences are assigned to PANTHER pathways.

2. Functional classifications of differentially expressed genes

References

[1] P. Langfelder and S. Horvath, *Wgcna: an r package for weighted correlation network analysis*, BMC Bioinformatics, 9 (2008), p. 559.

[2] H. Mi, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, *Large-scale gene function analysis with the panther classification system*, Nature Protocols, 8 (2013/07/18/online).

[3] L. Ping, D. M. Duong, L. Yin, M. Gearing, J. J. Lah, A. I. Levey, and N. T. Seyfried, *Global quantitative analysis of the human brain proteome in alzheimers and parkinsons disease*, Scientific Data, 5 (2018/03/13/online).