PH240C Final Project

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

GitHub Repository

Ramasubramanian Balasubramanian, John Canty, Annan Deng, William Krinsman

December 6th, 2018

Abstract

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 protein expression patterns that are associated with AD and PD neuropathologies, reflecting the belief that there may be a relationship between the protein expression patterns between the two diseases. We will then perform Weighted Correlation Network Analysis (WCNA) of the identified protein to assess if there coexpression relationships between subsets of these proteins. Finally, using the PANTHER gene ontology classification system, we will assess whether these proteins exhibit functional relationships. More details of our analysis can be found at our GitHub repository here.

Contents

1 Dataset

The dataset[11] 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.

2 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) protein expression values for each patient sample across all disease traits (Ctl, AD, PD, and ADPD) (see Figure 1). Upon inspection of the plots, we noticed that all samples had FC protein 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 protein expression outliers. Therefore we use quantile normalization (see Figure 2) to normalize the data across all batches and remove the outliers.

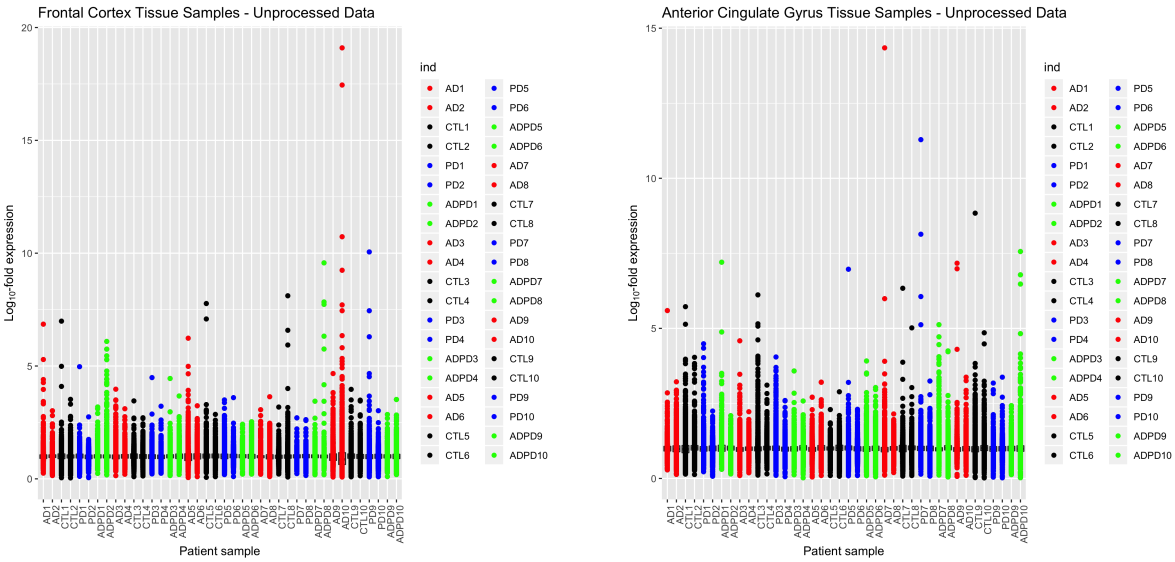


Figure 1: Distribution of all fold-change protein expression values.

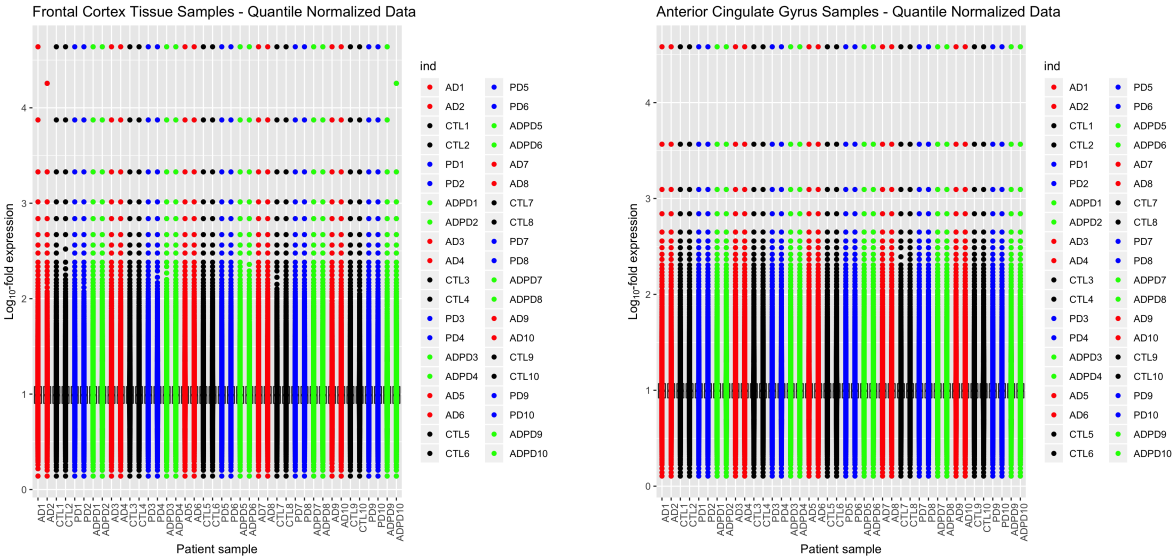


Figure 2: Quantile normalized data.

Next, we performed Principle Component Analysis (PCA) on the samples in both the Frontal Cortex and the Anterior Cingulate Gyrus (see Figure 4). A PCA plot of all disease traits did not indicate any clear separation of classes in both tissues. We also performed PCA plots by batches (see Figure 3). In order to simplify visualization, we then generated pairwise PCA plots between the Control samples and disease-state samples in both tissues (Figures 5 and 6). 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.

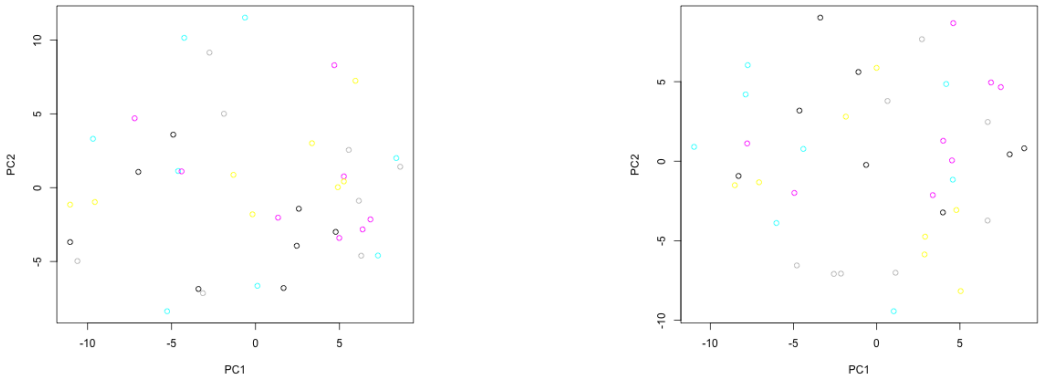


Figure 3: left: Frontal Cortex; right: Anterior Cingulate Gyrus (by batches)

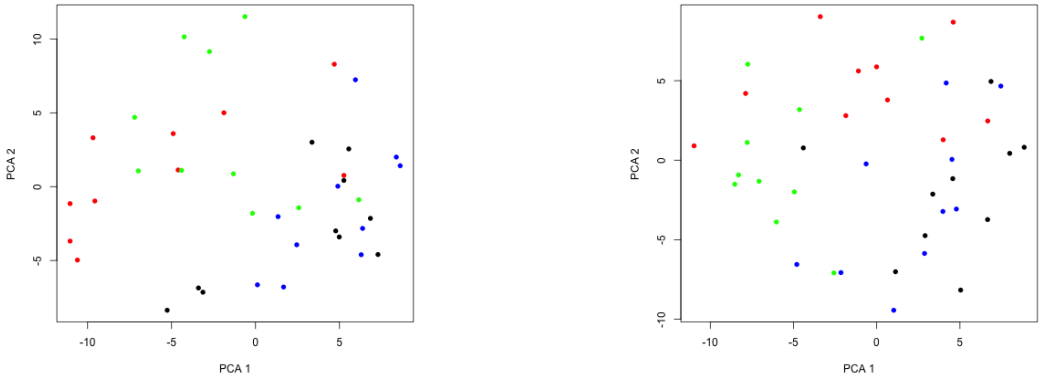


Figure 4: left: Frontal Cortex; right: Anterior Cingulate Gyrus (by classes)

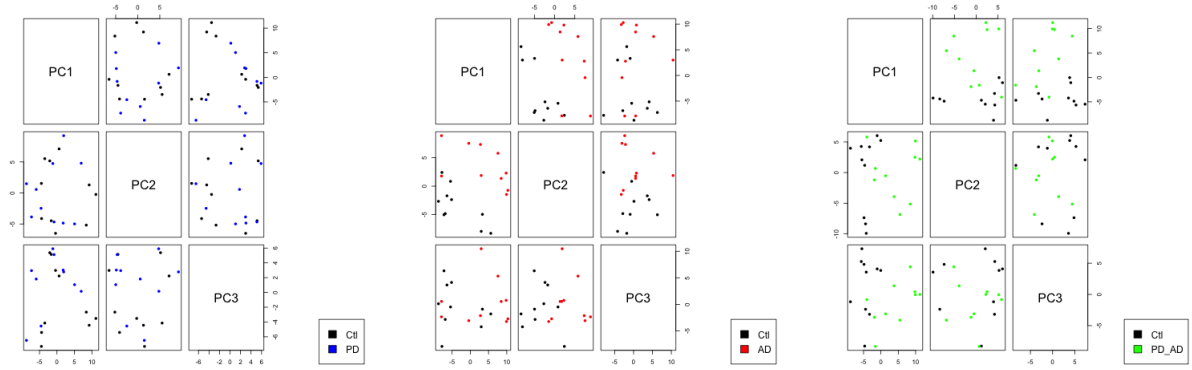


Figure 5: Pairwise PCA for Frontal Cortex

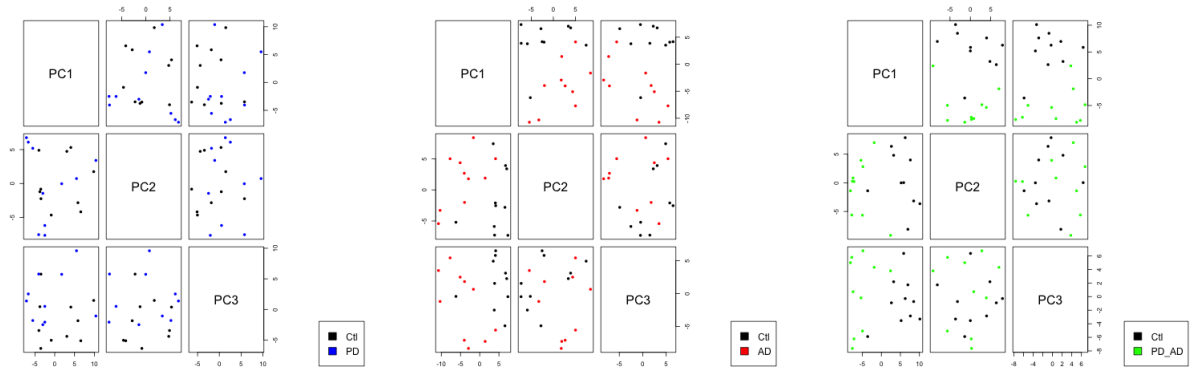
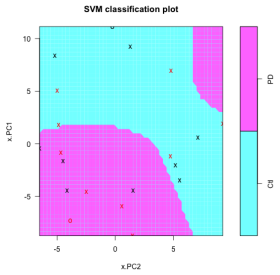
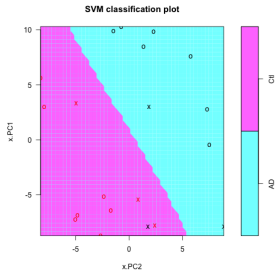


Figure 6: Pairwise PCA for Anterior Cingulate Gyrus

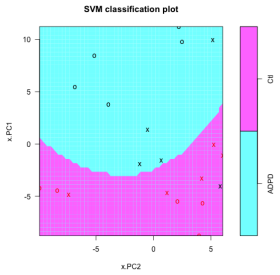
We then quantitatively assessed the separation between the Control and disease-state samples by applying a Support Vector Machine (SVM) classifier to the PCA plots (Figures 7 and 8). Consistent with our observations from the PCA plots, we observed that the SVM was able to accurately classify Control 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.



	Ctl	PD
Ctl	6	4
PD	4	6

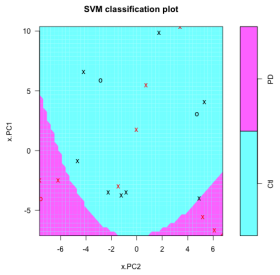


	AD	Ctl
AD	9	1
Ctl	0	10

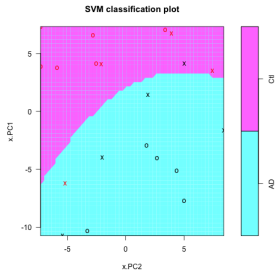


	ADPD	Ctl
ADPD	9	1
Ctl	0	10

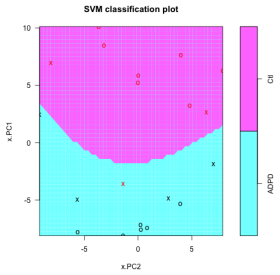
Figure 7: pairwise classification for Frontal Cortex



	Ctl	PD
Ctl	6	4
PD	4	6



	AD	Ctl
AD	9	1
Ctl	0	10



	ADPD	Ctl
ADPD	9	1
Ctl	0	10

Figure 8: pairwise classification for Anterior-Cingulate-Gyrus

3 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:

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

Where the x^is 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:

$$\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{AD}}, \hat{\beta}_{\text{PD}}, \text{ and } \hat{\beta}_{\text{ADPD}}\,.$$

(a) Ctl vs. PD (b) Ctl vs. AD (c) Ctl vs. ADPD

Figure 9: Frontal Cortex Volcano Additive Model

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:

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

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:

$$\begin{aligned} \log_2 FC(\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{AD}}) \quad & \text{Pval}(\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{AD}}) \\ \log_2 FC(\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{PD}}) \quad & \text{Pval}(\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{PD}}) \\ \log_2 FC(\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{ADPD}}) \quad & \text{Pval}(\hat{\beta}_{\text{Ctl}}, \hat{\beta}_{\text{ADPD}}) \end{aligned}$$

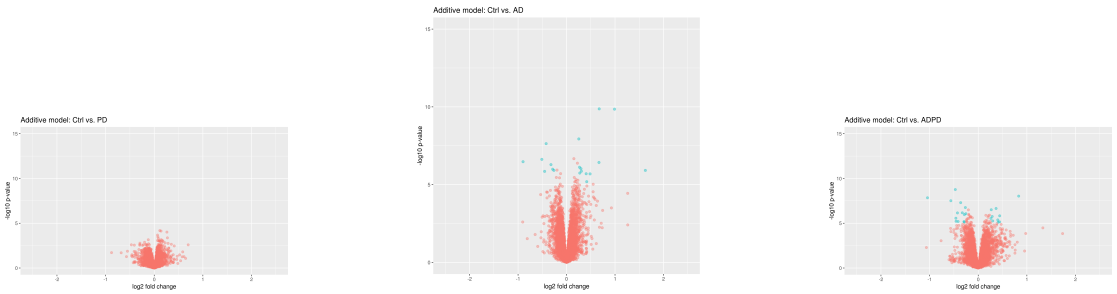
Next, we plotted our results using volcano plots. We defined a differential expressed gene as any gene expression value where the p-value was less than $\frac{0.05}{\text{nGenes}}$ and the $\log_2$ fold-change was larger than 0.25 (see Figures 9 and 10).

4 Weighted correlation network analysis

Our goal here was to employ weighted correlation network analysis in order to determine if the candidate differentially expressed genes for the PD, AD, and PD/AD disease phenotypes exhibit interesting coexpression patterns.

Correlation networks are constructed on the basis of correlations between quantitative measurements that can be described by an n X m matrix X, where the row indices correspond to network nodes/genes ($i = 1, \ldots, n$) and the column indices ($l = 1, \ldots, m$) correspond to sample measurements. The rationale behind correlation network methodology is to use network language to describe the pairwise relationships (correlations) between the rows of X. We now describe the steps.

We started off by constructing an adjacency matrix, which is a symmetric $n \times n$ matrix with entries in $[0, 1]$ whose component a_{ij} encodes the network connection strength between nodes i and j. To calculate the



(a) Ctl vs. PD (b) Ctl vs. AD (c) Ctl vs. ADPD

Figure 10: Anterior Cingulate Cyrus Additive Model

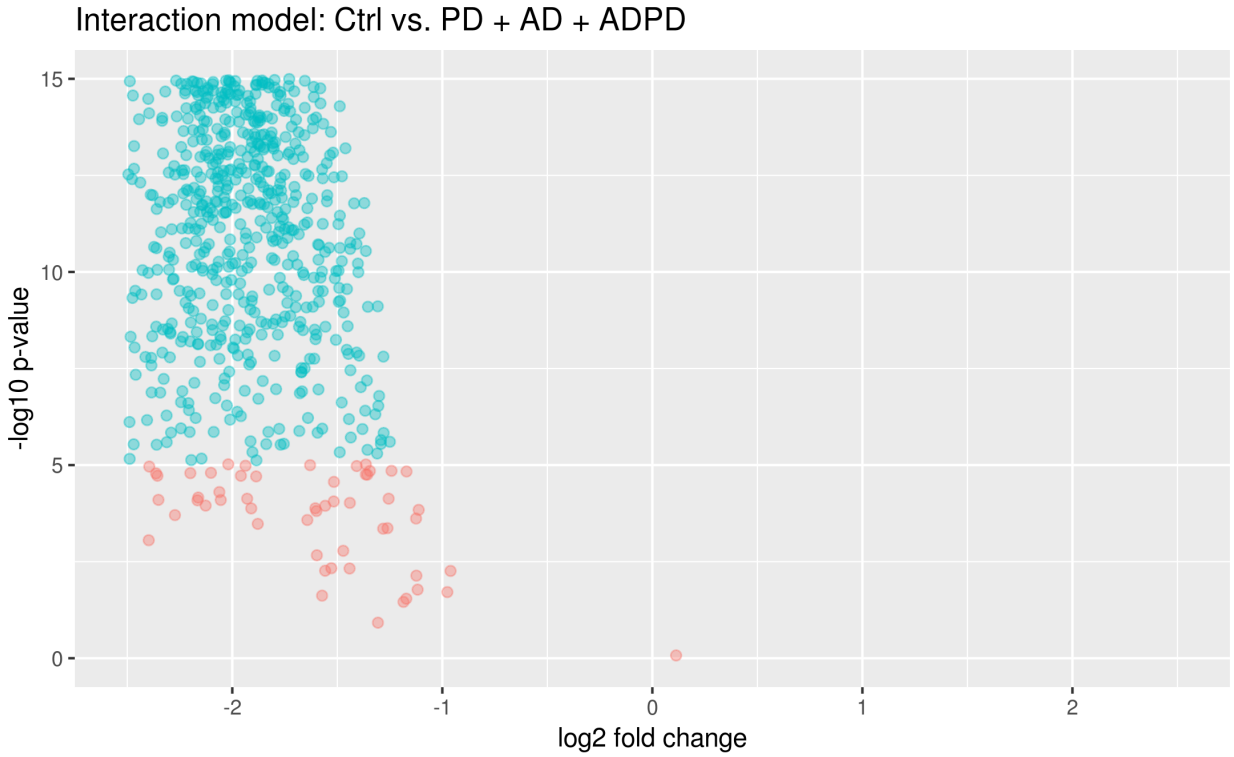


Figure 11: Interaction model for Anterior Cingulate Gyrus

adjacency matrix, an intermediate quantity called the co-expression similarity is first defined. The default method defines the coexpression similarity s_{ij} as the absolute value of the correlation coefficient between the profiles of nodes i and j:

$$s_{ij} = |\operatorname{cor}(x_i, x_j)|.$$

A weighed network adjacency is defined by raising the co-expression similarity to a power: $a_{ij} = s_{ij}^{\beta}$. Hence, the weighted adjacency between two genes is proportional to their similarity on a logarithmic scale. One way to visualize a weighted network is to plot its heatmap (Figure 12). Each row and column of the heatmap corresponds to a single gene. The heatmap can depict adjacencies or topological overlaps, with light colors denoting low adjacency (overlap) and darker colors higher adjacency (overlap). In addition, the gene dendrograms and module colors are plotted along the top and left side of the heatmap.

To find the optimal value of β, we examined the plot of power vs R^2 for a scale-free topology fit (Figure 13). A scale-free network is a network whose degree distribution follows a power law, at least asymptotically. That is, the fraction $P(k)$ of nodes in the network having k connections to other nodes goes for large values of k as $k^{-\gamma}$.

Based on the plots, we decided to go ahead with 18 as the power for FC (trade-off between R^2 and power) and 10 as the power for AFC (the lowest power which achieves an R^2 value above 0.7, which we set as a threshold). To minimize effects of noise and spurious associations, we then transformed the adjacency into a Topological Overlap Matrix. The m-th order generalized topological overlap measure (GTOM) is defined by counting the number of m-step neighbors that a pair of nodes share and normalizing it to take a value between 0 and 1. It allows one to trade-off sensitivity versus specificity when it comes to defining pairwise interconnectedness and network modules. Several studies have shown that two proteins having a higher topological overlap are more likely to belong to the same functional class than proteins having a lower topological overlap. Defining these mathematically, $N_m(i) = \{j \neq i | dist(i,j) \leq m\}$ is the set of nodes (excluding i itself) that are reachable from i within a path of length m. The mth order generalized TOM matrix has elements:

$$t_{ij}^m = \frac{|N_m(i) \cap N_m(j)| + a_{ij}}{\min(|N_m(i)|, |N_m(j)|) + 1 - a_{ij}}$$

We then calculated the corresponding dissimilarity matrix (1-TOM), and used hierarchical clustering to produce a dendrogram of the genes. Each leaf, that is a short vertical line, corresponds to a gene. Branches of the dendrogram group together densely interconnected, highly co-expressed genes. Module identification amounts to the identification of individual branches (cutting the branches off the dendrogram). We used the Dynamic Tree Cut for branch cutting, which implements an adaptive, iterative process of cluster decomposition and combination and stops when the number of clusters becomes stable. It starts by obtaining a few large clusters by the static tree cut. The joining heights of each cluster are analyzed for a characteristic pattern of fluctuations indicating a sub-cluster structure; clusters exhibiting this pattern are recursively split. To avoid over-splitting, very small clusters are joined to their neighboring major clusters.

The function returned 10 clusters for the FC case (refer the R notebook), and we observed that all the clusters had an even distribution of the differentially expressed genes in the AD and the AD/PD case. The case was similar for AGC as well. We decided to merge some of these clusters, and to do that we found out the eigen genes of each cluster, which are representative of the cluster. They are nothing but the first PCA direction of the corresponding expression matrix in each cluster. We wanted to merge modules whose expression profiles are very similar since their genes are highly co-expressed. We clustered the eigengenes on correlation. We got three clusters after that, again having pretty even distributions of the differentially expressed genes from the different volcano plots. Figure 14 has color panels that show the original clusters in the topmost panel, the clusters after merging in the second panel, and the volcano-plot membership (or

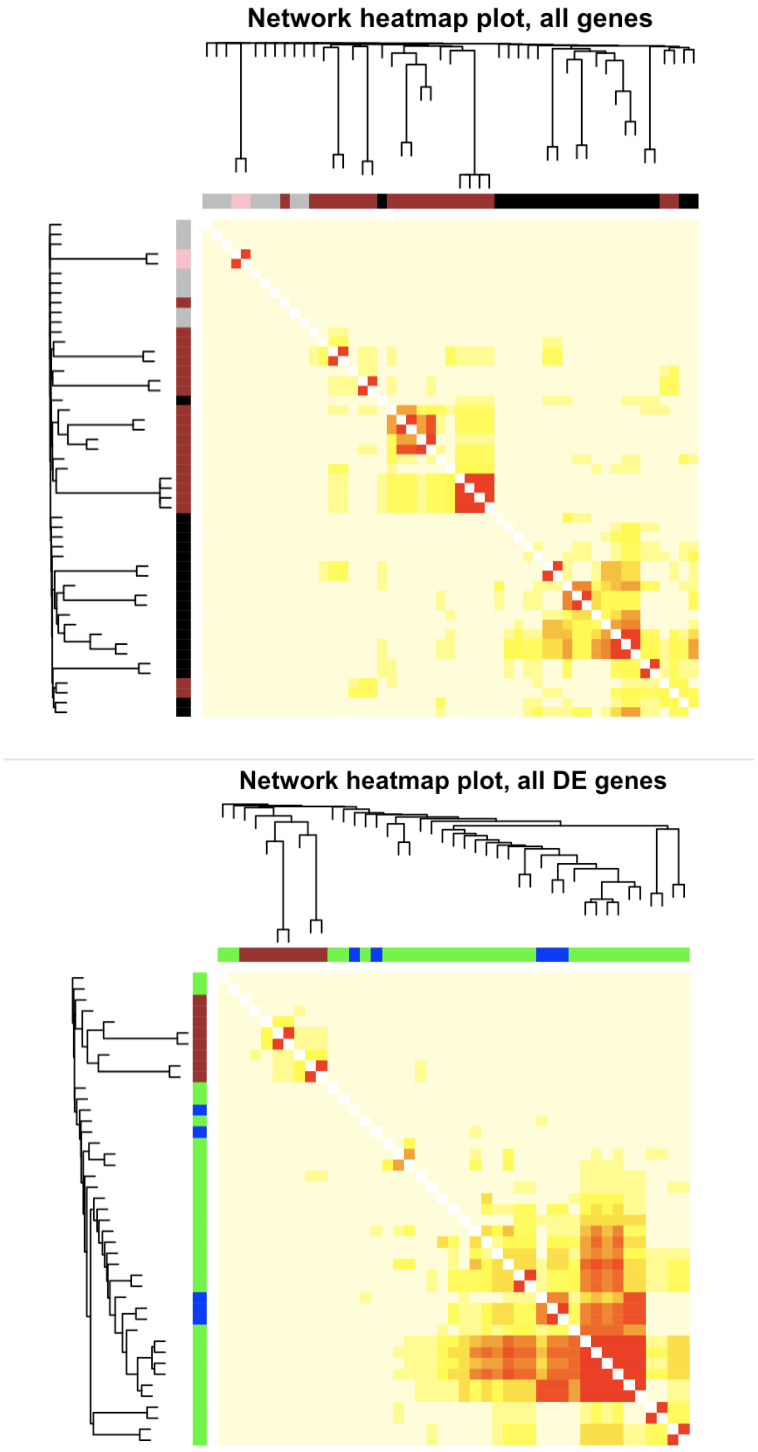


Figure 12: Top: Frontal Cortex, Bottom: Anterior Cingulate Gyrus

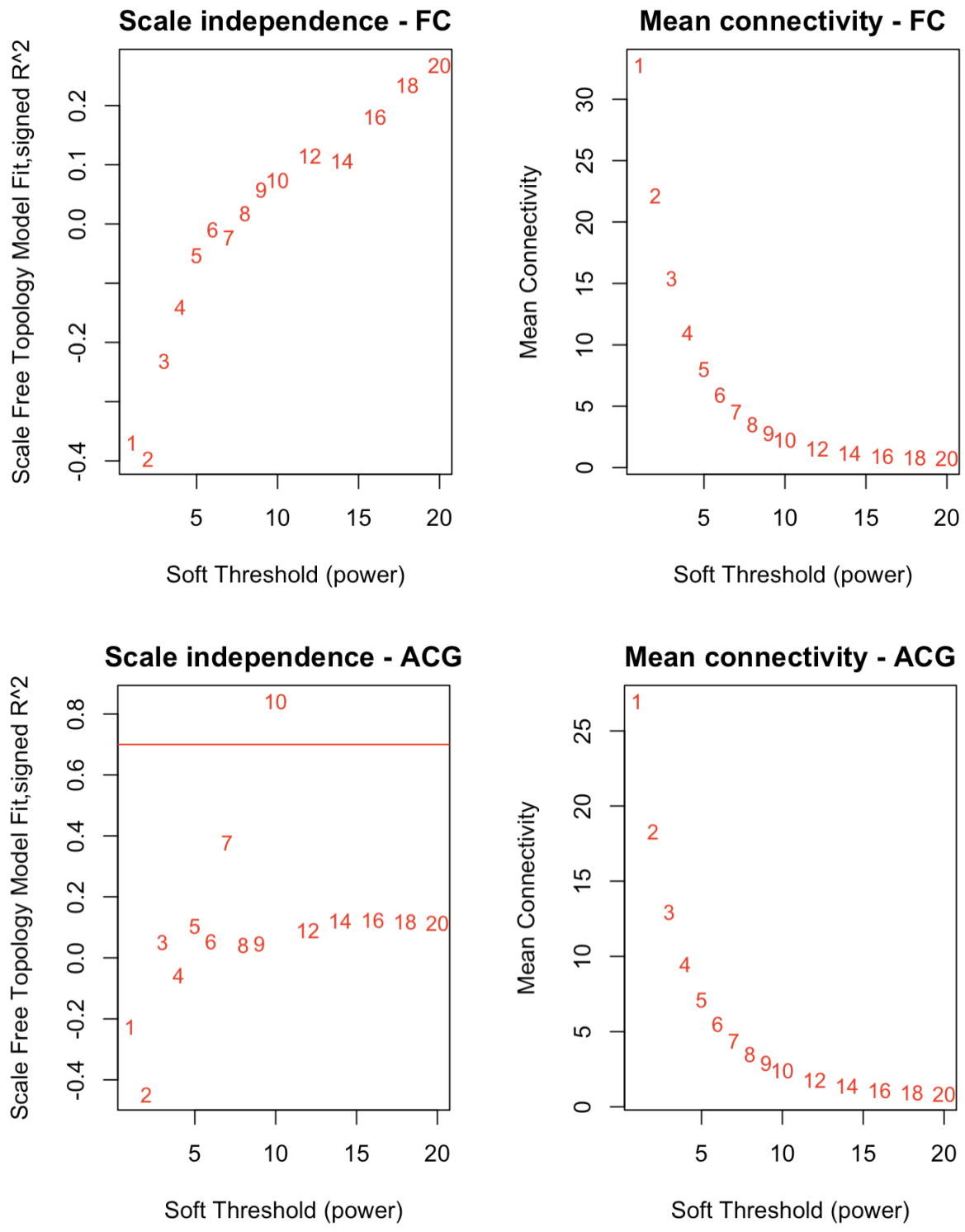


Figure 13: Scale-free topology determination for both the parts of the brain

rather, for which phenotype was this a differentially expressed gene) in the third panel. Given that there is no homogeneity in the clusters based on which phenotype's differentially expressed genes those are (i.e. each cluster has an even distribution from all the relevant volcano plots), we concluded that the comorbidity is driven by the AD genes, which are coexpressed with the AD/PD genes, which corroborates the findings of the authors.

5 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[8]:

> 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 proteins in frontal cortex

(a) Alzheimer's disease versus controls (Additive and interaction model)

Proteins identified as differentially expressed by both additive and interaction model

- P10636: Microtubule-associated protein tau, associated with microtubule binding and neuron projection.
- Q9Y2J0: Rabphilin-3A, function undescribed for humans in PANTHER.
- Q9HCH3: Copine-5, function undescribed for humans in PANTHER.
- Q9H4F8: SPARC-related modular calcium-binding protein 1, associated with the binding of calcium ions.
- P09211: Glutathione S-transferase P, function undescribed for humans in PANTHER.
- Q9UIW2: Also known as PLXNA1 (Plexin-A1), associated with axon guidance mediated by semaphorins, nervous system development, signal transducer activity, receptor activity, GTPase activity, pyrophosphatase activity, and many others.
- Q96Q06: Perilipin-4, function undescribed for humans in PANTHER.
- P61764: Syntaxin-binding protein 1, also known as STXBP1, associated with synaptic vesicle exocytosis and trafficking (as part of synaptic transmission).
- P48729: Also known as CSNK1A1 (Casein kinase I isoform alpha)), associated with protein kinase activity. Well known to be associated with Parkinson's disease.

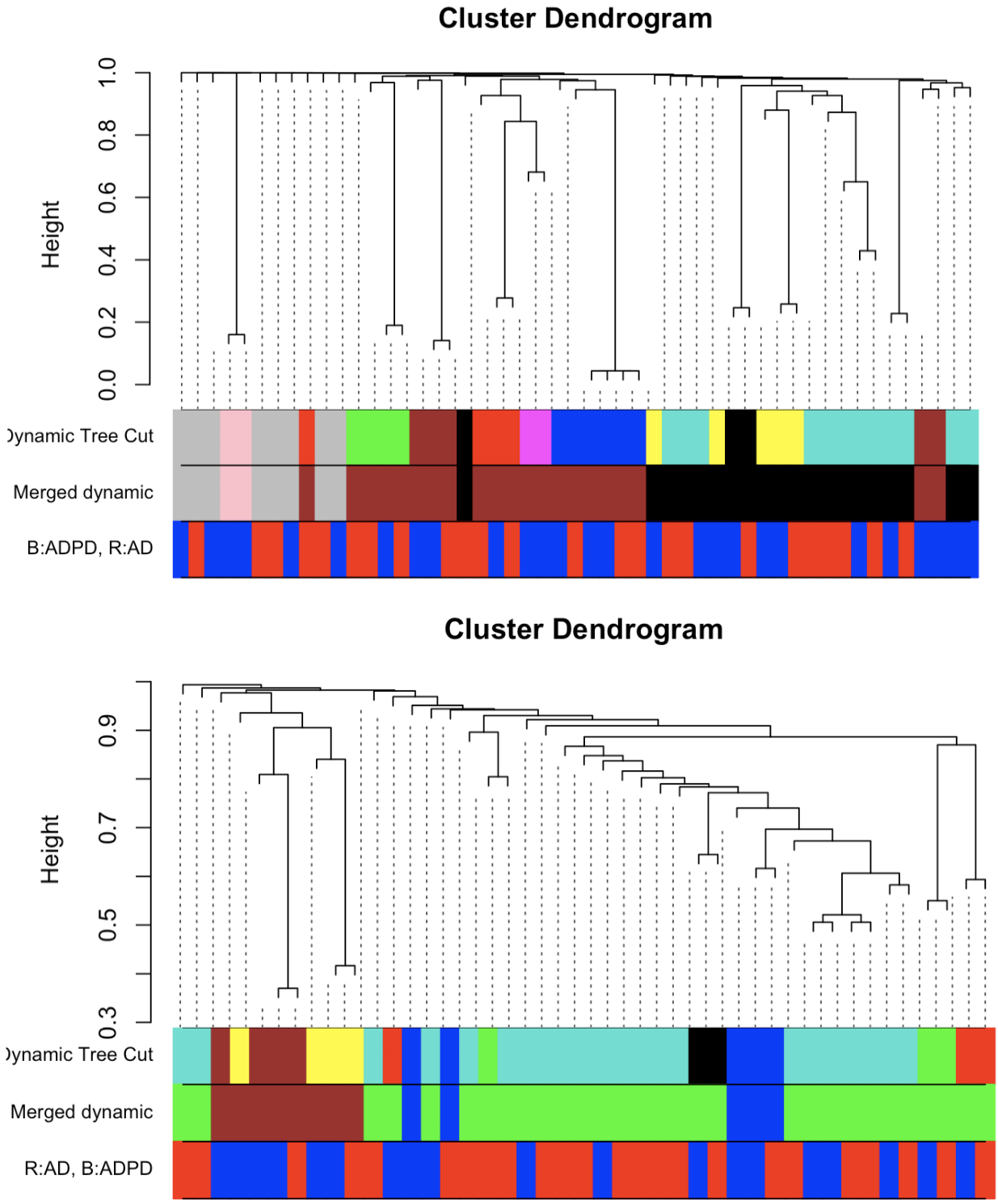


Figure 14: Top: Frontal Cortex, Bottom: Anterior Cingulate Gyrus

- P30041: Peroxiredoxin-6, an enzyme which catalyzes the oxidation via hydrogen peroxide.
- P78369: Claudin-10, a tight junction protein known to be found in the plasma membrane.
- O43175: D-3-phosphoglycerate dehydrogenase, associated with the synthesis of serine and glycine.
- Q9Y617: Phosphoserine aminotransferase, also associated with the synthesis of serine and glycine, as well as pyridoxal-5 phosphate synthesis and vitamin B6 metabolism.
- P04080: Cystatin-B, known to be a cysteine protease inhibitor.
- O95452: Gap junction beta-6 protein, associated with gap junction channel activity in the plasma membrane.
- P31431: Syndecan-4, a cytoskeletal protein involved in cell adhesion and the extracellular matrix, as well as membrane-bound signalling.
- Q13501: Sequestosome-1, known to be associated with protein binding in vacuoles in the cytoplasm.
- P04792: Heat shock protein beta-1, involved in multiple signalling pathways as well as sensory and in particular visual perception.
- Q14019: Coactosin-like protein, involved in the intracellular actin cytoskeleton, in particular with the binding of actin as well as a cytoskeletal component itself.
- Q9UBI6: Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-12, involved in GTPase activity, protein binding, and numerous signalling pathways.
- Q8N987: N-terminal EF-hand calcium-binding protein 1, known to be found in the cytoplasm and involved in the regulation of cellular metabolism, given the naming presumably via the binding of calcium. Otherwise little else appears to be described in PANTHER.

Identified by interaction model but not additive model

- A0A087WWT2: Neuritin, also known as NRN1, function undescribed for humans in PANTHER.
- B5MCG9: R3H domain-containing protein 2, also known as R3HDM2, function unknown in humans.
- Q9UPV7: Unnamed, also known as PHF24 (PHD Finger Protein 24)), function undescribed for humans in PANTHER.

Identified by interaction model but not additive model None.

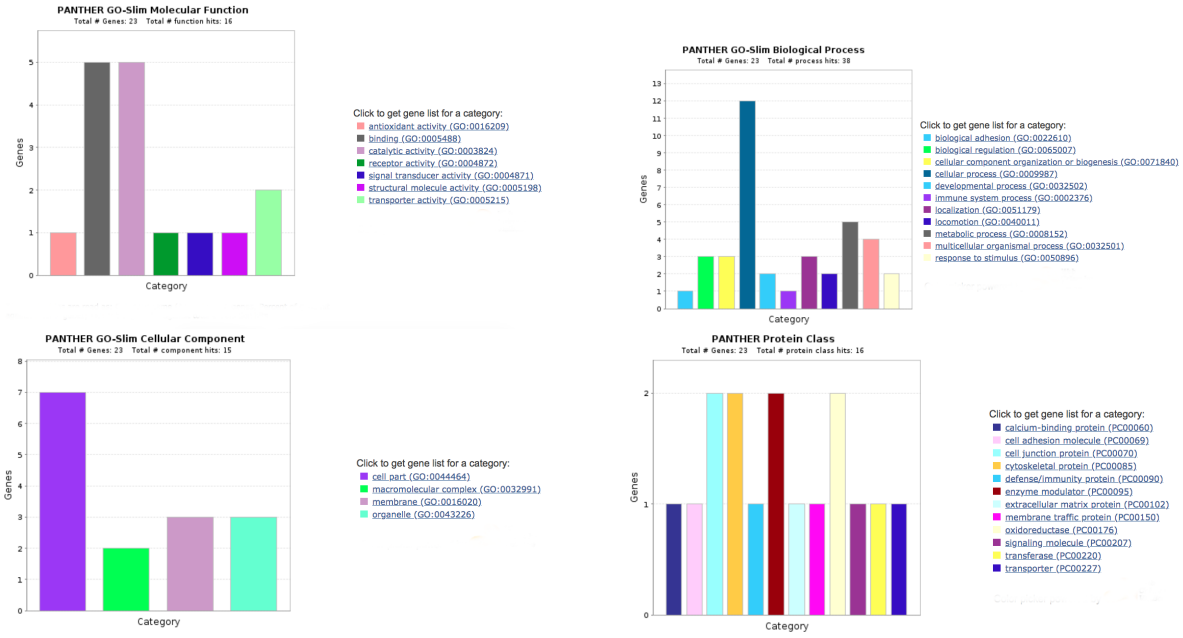


Figure 15: Gene ontologies for proteins found in (a).

Comments: First, it is a good validation of our methodology that it detected the differential expression of the tau proteins, given the extensive prior literature relating these proteins to Alzheimer's disease. It is also noteworthy that protein already well-known to be associated with Parkinson's disease, namely Casein kinase I isoform alpha, was found to be differentially expressed in the frontal cortex of Alzheimer's disease patients who did not also have Parkinson's. This is for two reasons: first, Parkinson's does not seem to affect the frontal cortex, so it is interesting that it might be involved with Alzheimer's in that region of the brain, and second, it provides support for the authors' conjecture that there may be structural similarity/relationship between the development of Alzheimer's and Parkinson's diseases.

As for some possible common themes that seem worth mentioning with regards to the biological functions of the differentially expressed proteins, these include calcium-binding, GTPase activity, synthesis of serine and glycine, signalling pathways, protein binding, cellular membrane functions, and the internal and external cytoskeletons.

Also, there were several proteins identified as differentially expressed whose function is still relatively unknown, including (but not necessarily limited to) Rabphilin-3A, Copine-5, Glutathione S-transferase P, and Perilipin-4. Our analysis strongly suggests that future research clarifying the biological function of these specific proteins could be relevant to understanding the pathology behind Alzheimer's disease.

We see an extremely large overlap between proteins already identified via the additive model as differentially expressed in either Alzheimer's patients or patients with comorbid Alzheimer's and Parkinson's, with the only exception being R3H domain-containing protein 2. Thus this seems to lend further support to the importance of these genes, if only to the extent that their being identified as differentially expressed does not appear contingent upon the exact assumptions of the linear model being used.

(b) Parkinson's Disease versus Controls (Additive and Interaction Model)

No genes were differentially expressed. This is also a validation of our methodology, since it corresponds to previous work which suggests that there is no manifestation of Parkinson's in the frontal cortex.

(c) Comorbid Alzheimer's and Parkinson's Diseases versus Controls (Additive and Interaction Model)

- O15240: Neurosecretory protein VGF, function undescribed for humans in PANTHER.
- Q9H4F8: SPARC-related modular calcium-binding protein 1, associated with the binding of calcium ions.
- O14829: Serine/threonine-protein phosphatase with EF-hands 1, associated with apoptosis and intracellular signalling.
- A0A087WWT2: Neuritin, also known as NRN1, function undescribed for humans in PANTHER.
- Q92743: Serine protease HTRA1, associated with serine-type proteolysis (peptide and protein breakdown).
- P04792: Heat shock protein beta-1, involved in multiple signalling pathways as well as sensory and in particular visual perception.
- Q13501: Sequestosome-1, known to be associated with protein binding in vacuoles in the cytoplasm.
- Q96Q06: Perilipin-4, function undescribed for humans in PANTHER.
- Q9Y2J0: Rabphilin-3A, function undescribed for humans in PANTHER.
- Q92563: Testican-2, associated with calcium ion binding, and the inhibition of proteolysis.
- P78369: Claudin-10, a tight junction protein known to be found in the plasma membrane.
- O75936: Gamma-butyrobetaine dioxygenase, known to be associated with the synthesis of certain vitamins.
- Q99784: Noelin, a phosphate ion receptor and transporter.

- Q14693: Phosphatidate phosphatase LPIN1, known to be associated with the regulation of transcription.
- O75891: Cytosolic 10-formyltetrahydrofolate dehydrogenase, a catalyst for the transfer of electrons from one molecule to another (oxidation-reduction reactions).
- K7EJH8: Alpha-actinin-4, also known as ACTN4, associated with the morphogenesis of the extracellular matrix/cytoskeleton.
- Q96RR4: Calcium/calmodulin-dependent protein kinase 2, associated with calcium-mediated synaptic transmission.
- Q9UPV7: Unnamed, also known as PHF24 (PHD Finger Protein 24), function undescribed for humans in PANTHER.
- P30043: Flavin reductase (NADPH), another oxidation-reduction catalyst.
- Q15102: Platelet-activating factor acetylhydrolase IB subunit gamma, specific function undescribed for humans in PANTHER.
- P40123: Adenylyl cyclase-associated protein 2, known to be involved in the organization and regulation of the intracellular actin cytoskeleton in the cytoplasm.
- Q9NQ86: E3 ubiquitin-protein ligase TRIM36, specific function undescribed for humans in PANTHER.
- Q9UPR0: Inactive phospholipase C-like protein 2, associated with GTPase regulation and several intracellular signalling pathways.

Even though some of the results are given via gene ID instead of protein ID or vice versa, ultimately the results here are literally identically the same as those for the additive model. This suggests that the two models also should have given identical output for the Alzheimer's only group as well (they gave identical output for the Parkinson's only group too), which strongly suggests that we may have made an error in our analysis (which unfortunately due to time constraints cannot be addressed).

Some common themes include regulation and development of the actin cytoskeleton, oxidation-reduction reaction catalysts, and several of the categories mentioned above (e.g. calcium ion binding, GTPase regulators, signalling pathways, vitamin synthesis, protein binding, etc.).

The occurrence of several proteins in both this group as well as those differentially expressed with Alzheimer's (e.g. heat-shock protein beta-1, Sequestosome-1, SPARC-related modular calcium-binding protein 1) only could be seen to validate our methodology, since we would expect these two groups to have similar protein expression profiles. Similarly, the occurrence of the unknown function proteins Perilipin-4 and Rabphilin-3A as differentially expressed in both of these groups strongly argues in favor of follow-up studies investigating the function and structure of these proteins.

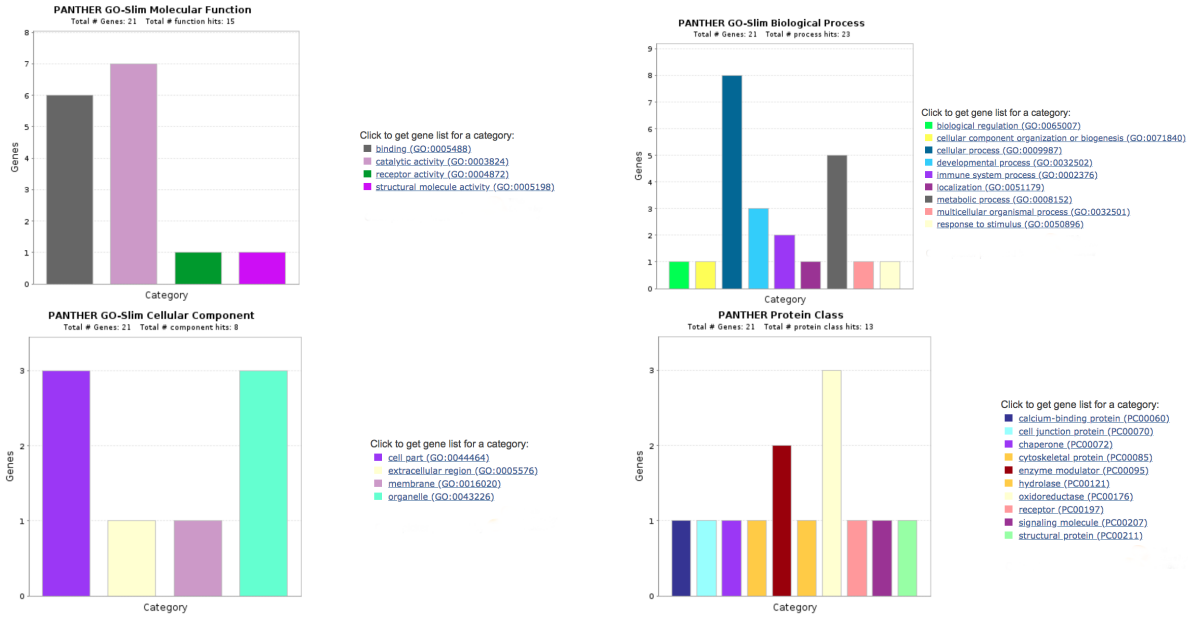


Figure 16: Gene ontologies for proteins found in (c).

3. Differentially expressed proteins in comorbid versus non-comorbid Alzheimer's

(a) Common proteins

- Q9Y2J0: Rabphilin-3A, function undescribed for humans in PANTHER.
- Q9H4F8: SPARC-related modular calcium-binding protein 1, associated with the binding of calcium ions.
- Q96Q06: Perilipin-4, function undescribed for humans in PANTHER.
- P78369: Claudin-10, a tight junction protein known to be found in the plasma membrane.
- Q13501: Sequestosome-1, known to be associated with protein binding in vacuoles in the cytoplasm.
- P04792: Heat shock protein beta-1, involved in multiple signalling pathways as well as sensory and in particular visual perception.
- A0A087WWT2: Neuritin, also known as NRN1, function undescribed for humans in PANTHER.
- Q9UPV7: Unnamed, also known as PHF24 (PHD Finger Protein 24)), function undescribed for humans in PANTHER.

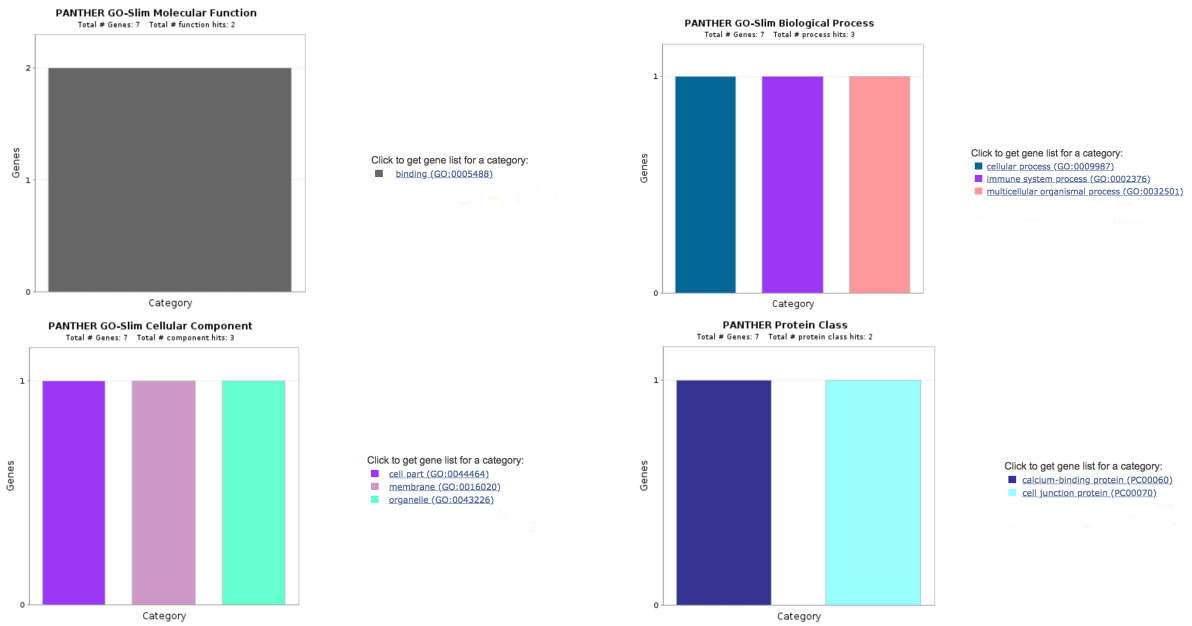


Figure 17: Gene ontologies for proteins found in (a).

(b) Identified in comorbid but not in non-comorbid

- O15240: Neurosecretory protein VGF, function undescribed for humans in PANTHER.
- O14829: Serine/threonine-protein phosphatase with EF-hands 1, associated with apoptosis and intracellular signalling.
- Q92743: Serine protease HTRA1, associated with serine-type proteolysis (peptide and protein breakdown).
- Q92563: Testican-2, associated with calcium ion binding, and the inhibition of proteolysis.
- O75936: Gamma-butyrobetaine dioxygenase, known to be associated with the synthesis of certain vitamins.
- Q99784: Noelin, a phosphate ion receptor and transporter.
- Q14693: Phosphatidate phosphatase LPIN1, known to be associated with the regulation of transcription.
- O75891: Cytosolic 10-formyltetrahydrofolate dehydrogenase, a catalyst for the transfer of electrons from one molecule to another (oxidation-reduction reactions).
- K7EJH8: Alpha-actinin-4, also known as ACTN4, associated with the morphogenesis of the extracellular matrix/cytoskeleton.
- Q96RR4: Calcium/calmodulin-dependent protein kinase 2, associated with calcium-mediated synaptic transmission.
- P30043: Flavin reductase (NADPH), another oxidation-reduction catalyst.
- Q15102: Platelet-activating factor acetylhydrolase IB subunit gamma, specific function undescribed for humans in PANTHER.
- P40123: Adenylyl cyclase-associated protein 2, known to be involved in the organization and regulation of the intracellular actin cytoskeleton in the cytoplasm.
- Q9NQ86: E3 ubiquitin-protein ligase TRIM36, specific function undescribed for humans in PANTHER.
- Q9UPR0: Inactive phospholipase C-like protein 2, associated with GTPase regulation and several intracellular signalling pathways.

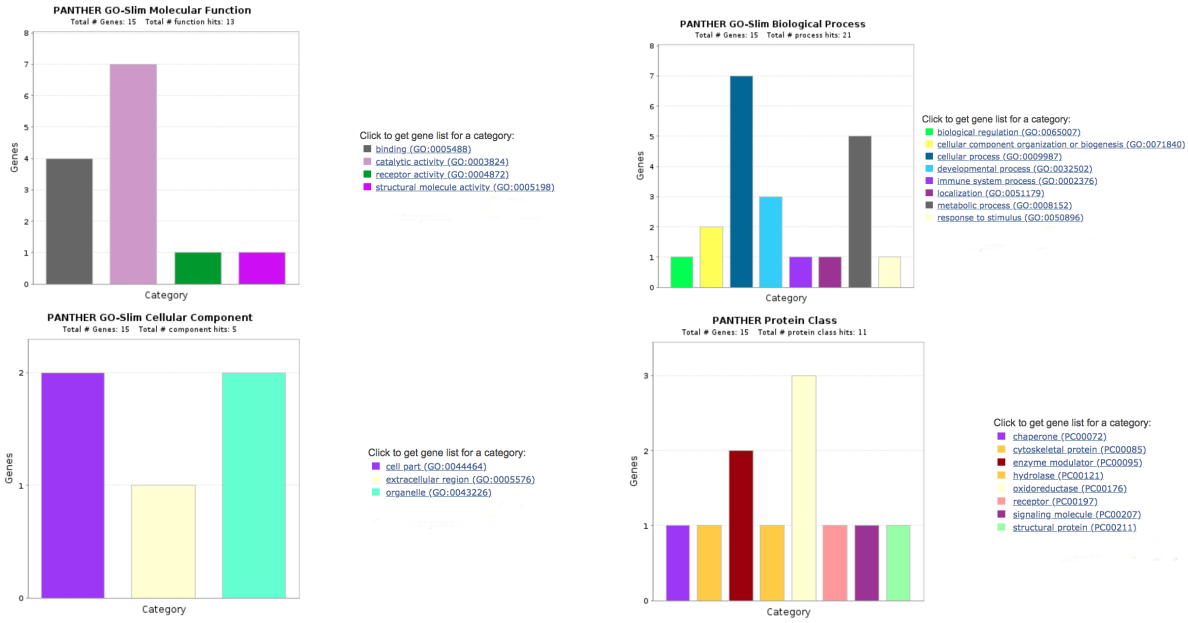


Figure 18: Gene ontologies for proteins found in (b).

(c) Identified in in non-comorbid but not in comorbid

- P10636: Microtubule-associated protein tau, associated with microtubule binding and neuron projection.
- Q9HCH3: Copine-5, function undescribed for humans in PANTHER.
- P09211: Glutathione S-transferase P, function undescribed for humans in PANTHER.
- Q9UIW2: Also known as PLXNA1 (Plexin-A1), associated with axon guidance mediated by semaphorins, nervous system development, signal transducer activity, receptor activity, GTPase activity, pyrophosphatase activity, and many others.
- P61764: Syntaxin-binding protein 1, also known as STXBP1, associated with synaptic vesicle exocytosis and trafficking (as part of synaptic transmission).
- P48729: Also known as CSNK1A1 (Casein kinase I isoform alpha)), associated with protein kinase activity. Well known to be associated with Parkinson's disease.
- P30041: Peroxiredoxin-6, an enzyme which catalyzes the oxidation via hydrogen peroxide.
- O43175: D-3-phosphoglycerate dehydrogenase, associated with the synthesis of serine and glycine.
- Q9Y617: Phosphoserine aminotransferase, also associated with the synthesis of serine and glycine, as well as pyridoxal-5 phosphate synthesis and vitamin B6 metabolism.
- P04080: Cystatin-B, known to be a cysteine protease inhibitor.
- O95452: Gap junction beta-6 protein, associated with gap junction channel activity in the plasma membrane.
- P31431: Syndecan-4, a cytoskeletal protein involved in cell adhesion and the extracellular matrix, as well as membrane-bound signalling.
- Q14019: Coactosin-like protein, involved in the intracellular actin cytoskeleton, in particular with the binding of actin as well as a cytoskeletal component itself.
- Q9UBI6: Guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-12, involved in GTPase activity, protein binding, and numerous signalling pathways.
- Q8N987: N-terminal EF-hand calcium-binding protein 1, known to be found in the cytoplasm and involved in the regulation of cellular metabolism, given the naming presumably via the binding of calcium. Otherwise little else appears to be described in PANTHER.

- B5MCG9: R3H domain-containing protein 2, also known as R3HDM2, function unknown in humans.

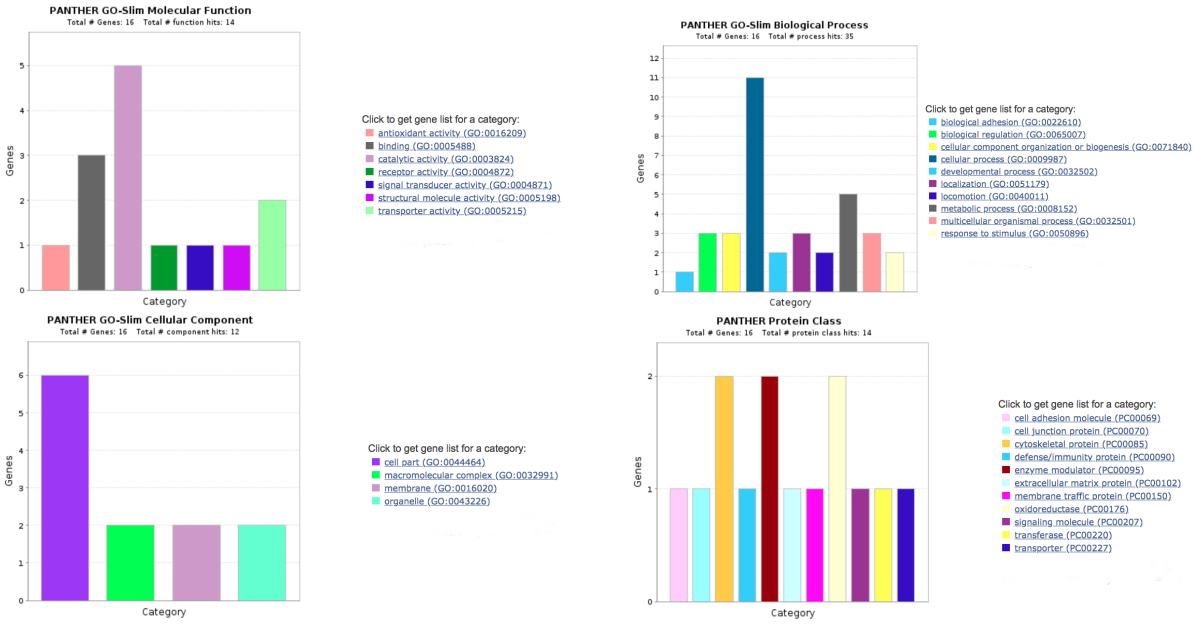


Figure 19: Gene ontologies for proteins found in (c).

Comments: First, it is very surprising that the tau protein associated so closely with Alzheimer's is not Identified as differentially expressed among the comorbid group. It is also strange (but less so, since we don't expect manifestation of Parkinson's in the frontal cortex) that the casein kinase associated with Parkinson's is not identified as differentially expressed among the comorbid group.

What is perhaps strangest here is the relatively small overlap between genes identified as differentially expressed in both the comorbid and non-comorbid group, and that there are so many genes identified as differentially expressed in the non-comorbid but not in the comorbid group. This is because we expect the comorbid group to be manifesting the same disease as the non-comorbid group, plus another disease, such that the differentially expressed genes in the comorbid group would be expected to be approximately a strict superset of those in the non-comorbid group – yet this is not what happened at all.

This counter-intuitive finding invites deeper scrutiny of our methodology as well as the assumptions implicit in using that methodology. However, if the finding was valid, it seemingly would possibly be of genuine interest to domain experts.

4. Summary

Some of our findings, e.g. that tau protein is differentially expressed, and that Parkinson's does not seem to manifest in the frontal cortex, correspond well to existing research. Others are more surprising, some even counterintuitive. There are other aspects of our results which argue in our favor. For example, we found several proteins involved in serine synthesis and metabolism to be differentially expressed, which corresponds to previous research linking Alzheimer's and serine[4, 6, 7, 9]. The same is also true of the identification of GTPase proteins as possibly relevant[10, 2, 1].

Based on our findings, we would also recommend future studies of Rabphilin-3A and Perilipin-4 and their relationships to Alzheimer's and Parkinson's to follow those which have already been made [15, 13, 14] [3, 12, 5] (as well as updates of the PANTHER database to more fully reflect current knowledge of these proteins). In any case, the existence of prior work studying these connections suggests at the very least that our analysis is not completely unreasonable in finding these proteins to possibly be important.

References

[1] B. J. Aguilar, Y. Zhu, and Q. Lu, *Rho gtpases as therapeutic targets in alzheimer's disease*, Alzheimer's Research & Therapy, 9 (2017), p. 97.

[2] S. Bolognin, E. Lorenzetto, G. Diana, and M. Buffelli, *The potential role of rho gtpases in alzheimer's disease pathogenesis*, Molecular Neurobiology, 50 (2014), pp. 406–422.

[3] X. Han, J. Zhu, X. Zhang, Q. Song, J. Ding, M. Lu, S. Sun, and G. Hu, *Plin4-dependent lipid droplets hamper neuronal mitophagy in the mptp/p-induced mouse model of parkinsons disease*, Frontiers in Neuroscience, 12 (2018), p. 397.

[4] K. Hashimoto, T. Fukushima, E. Shimizu, S. ichi Okada, N. Komatsu, N. Okamura, K. Koike, H. Koizumi, C. Kumakiri, K. Imai, and M. Iyo, *Possible role of d-serine in the pathophysiology of alzheimer's disease*, Progress in Neuro-Psychopharmacology and Biological Psychiatry, 28 (2004), pp. 385 – 388.

[5] M. V. Heck, M. Azizov, T. Stehning, M. Walter, N. Kedersha, and G. Auburger, *Dysregulated expression of lipid storage and membrane dynamics factors in tia1 knockout mouse nervous tissue*, neurogenetics, 15 (2014), pp. 135–144.

[6] L. Katsouri, A. Ashraf, J. de Belleroche, and M. Sastre, *D-serine synthesis and metabolism in the alzheimer's brain*, Alzheimer's & Dementia, 7 (2011), p. S702.

[7] C. Madeira, M. V. Lourenco, C. Vargas-Lopes, C. K. Suemoto, C. O. Brandão, T. Reis, R. E. P. Leite, J. Laks, W. Jacob-Filho, C. A. Pasqualucci, L. T. Grinberg, S. T. Ferreira, and R. Panizzutti, *d-serine levels in alzheimer's disease: implications for novel biomarker development*, Translational Psychiatry, 5 (2015), pp. e561 EP –. Original Article.

[8] 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).

[9] NIH, *Phase iia l-serine trial for ead - full text view*.

[10] I. Nishimoto, T. Okamoto, Y. Matsuura, S. Takahashi, T. Okamoto, Y. Murayama, and E. Ogata, *Alzheimer amyloid protein precursor complexes with brain gtp-binding protein go*, Nature, 362 (1993), pp. 75 EP –.

[11] 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).

[12] M. K. Shimabukuro, L. G. P. Langhi, I. Cordeiro, J. M. Brito, C. M. d. C. Batista, M. P. Mattson, and V. de Mello Coelho, *Lipid-laden cells differentially distributed in the aging brain are functionally active and correspond to distinct phenotypes*, Scientific Reports, 6 (2016), pp. 23795 EP –. Article.

[13] J. Stanic, M. Carta, I. Eberini, S. Pelucchi, E. Marcello, A. A. Genazzani, C. Racca, C. Mulle, M. Di Luca, and F. Gardoni, *Rabphilin 3a retains nmda receptors at synaptic sites through interaction with glun2a/psd-95 complex*, Nature Communications, 6 (2015), pp. 10181 EP –. Article.

[14] J. Stanic, M. Mellone, F. Napolitano, C. Racca, E. Zianni, D. Minocci, V. Ghiglieri, M.-L. Thiolat, Q. Li, A. Longhi, A. D. Rosa, B. Picconi, E. Bezard, P. Calabresi, M. D. Luca, A. Usiello, and F. Gardoni, *Rabphilin 3a: A novel target for the treatment of levodopa-induced dyskinesias*, Neurobiology of Disease, 108 (2017), pp. 54 – 64.

[15] M. G. Tan, C. Lee, J. H. Lee, P. T. Francis, R. J. Williams, M. J. Ramrez, C. P. Chen, P. T.-H. Wong, and M. K. Lai, *Decreased rabphilin 3a immunoreactivity in alzheimers disease is associated with a burden*, Neurochemistry International, 64 (2014), pp. 29 – 36.