Characterization of Tissue-Specific Gene Expression Differences in the Adult Human Brain

Sihui Huang, BS[1], Florencia Velez-Cortes, BS[1], Nick Giangreco, BS[1], Emily Groopman, BS[1]
[1]Columbia University, New York, New York, USA

Abstract

Central nervous system diseases are difficult to diagnose and to treat, partly due to the lack of common genetic markers and a limited understanding of the pathways that are affected in these disorders. Microarray technology has advanced our understanding of gene expression and found differentially expressed pathways in neurological disorders. However, the interpretation of these studies is complicated by tissue-specific gene expression differences across the heterogeneous structures of the human brain. In this study, we use Gene Expression Omnibus (GEO) microarray data of individual brain structures from Alzheimer's patients to find enriched pathways. We can observe the variance across the 414 structures of the Allen Brain Atlas (ABA) in these enriched pathways. We have resolved differences in gene expression across healthy tissues in the brain by performing principal component analysis on microarray data obtained from the Allen Brain Atlas. Using this method, we could identify structures of the brain that should be subjects of further studies in Alzheimer's.

Introduction

While diseases such as cancer have been well studied and the etiology of different tumor subtypes well documented, the molecular mechanisms driving many brain diseases remain unknown. Neurodegenerative and psychiatric diseases are a major cause of morbidity in the USA[1], but a lack of understanding about their pathogenesis hinders effective diagnosis and treatment[2], resulting in a paucity of disease-specific biomarkers or effective drugs[3]. There is a rich literature examining the effects of various neurological disorders on gene expression on the level of major brain structures[4-6]. However, the brain's high degree of anatomical and functional heterogeneity necessitates examining gene expression at more detailed, sub-structural resolutions. High-throughput technologies for gene expression analysis have improved the study in human brain disease, and allow researchers to view disease manifestation on the molecular level in large populations. Using high-throughput experiments of different brain tissues, we can compare gene expression in different areas of the brain. By understanding the variance in expression across tissues and how expression changes with disease, we can construct gene and pathway interaction networks to delineate and compare the etiology of brain diseases[7-9]. Yet, individual laboratories are underpowered to measure gene expression of more than a handful of tissues for a single study. Here, we use the gene expression data from Gene Expression Omnibus (GEO) [10] and the Allen Brain Atlas[11] to review and synthesize the major gene functions and pathways across different diseased brain tissues, focusing on Alzheimer's disease, which shows the greatest pathogenic differential expression in cortex tissues[4,12,13]. We present a strategy for studying tissue-specific gene expression in a brain-wide fashion.

Methods

Allen Brain Atlas and Gene Expression Omnibus Datasets

To obtain gene expression data for healthy adult brain tissue, we used the R v 3.3.1 package ABAData[7] to download Human Brain dataset from the Allen Brain Atlas1. The Human Brain dataset consists of expression data for protein-coding regions of the genome. Brain tissue samples from 414 regions were collected ≤30h post-mortem from 6 adult donors (5 males, 1 female) aged 18-68 years with no history of neurological or neuropsychiatric disease. Microarray was performed using a custom Agilent 8x60K array, comprising the Agilent 4x44 Whole Human Genome probe set plus an additional 16,000 probes; the resulting array had ≥ 2 probes for 93% of genes with EntrezGeneIDs (29,176 genes). The resulting dataset consists of HGNC-symbol, Entrez-ID, Ensembl-ID, brain region ID as used in the ontology from Allen Brain Atlas (with eight hierarchical tiers), microarray expression value, and the developmental stage of the tissue (here, adult). Expression values were mapped 15,698 unique Ensembl Gene IDs, with the mean value used if multiple probes mapped to one gene, and averaged for the associated brain region within and then between donors, yielding a single value for a given gene in a given region. Alzheimer's expression data was downloaded from the GEO ID GSE48350. The microarray platform was GPL6104 or Illumina humanRef-8 v2.0 expression beadchip. There were 253 patients with 80 Alzheimer's patients and 173 control patients.

Principal Component Analysis

Microarray data from the Allen Brain Atlas Human Brain dataset were conformed to a matrix of 414 columns including both cortex and cerebellum brain structures and 15,698 rows for genes. Principal component analysis was performed using prcomp package of R[14]. Since rigorous normalization was performed on the brain structure data, principal components were calculated while centering the data to zero, and scaled to have unit variance.

Hierarchical Clustering

To determine classification of the 414 brain structures by embryonic origin, we computed hierarchical clustering on the top 500 variable genes using complete agglomeration and a Euclidean distance matrix of the expression of each gene in a given structure compared to the mean of all genes for a given structure. We defined the variability of the pathway by taking the variance in expression of each gene contained in a pathway. Then we subtracted the mean of the variance across structures from the observed pathway variance for each structure to visualize the deviation of variable pathways in a given structure from the average pathway variability across 414 brain structures. The pathways with the top 25 highest and the top 25 lowest variance deviations were then hierarchically clustered.

Differential Expression Analysis

Differences between Alzheimer's patients and control patients was determined using the limma[15] package. The proportion of differential genes used for downstream analysis was defined by the complement proportion of genes determined to be null hypotheses by the pi0est function in the qvalue[16] package in R.

Pathway Enrichment

We computed enrichment in all 1,529 Homo Sapiens Reactome[17] pathways using the differential genes in Alzheimer's patients, where only the subset of genes in common with the ABA dataset, as well as two disease datasets used in an accompanying paper (Paper 2), were used. Pathway enrichment was computed with the fisher's exact test. The top enriched pathways were defined as having an FDR < 0.05.

Predicted Pathway-Structure Co-Localization

We computed the localization of disease enriched pathways that show how variability in the ABA data[18]. For each of the top enriched pathways in Alzheimer's disease, we examined the variability of gene expression for those pathways in all 414 brain structures, only permitted the brain structures with more than the 75th percentile variance in a pathway's gene set expression, and counted the number of structures found highly variable for the top enriched disease pathways.

All analyses and code are uploaded at https://github.com/ngiangre/BINF-4006-Project.

Results

Principal Component Analysis of Allen Brain Atlas Data Shows Clustering by Embryonic Origin of Tissue

To determine whether there was a clustering pattern across different structures of the healthy human brain, we performed principal component analysis with structures from the ABA. Principal components were computed with a matrix of 414 rows for both cortex and cerebellum brain structures and 15,698 rows for genes. The first two principal components accounted for 36% of the variance in gene expression of the tissues sampled. We observed a striking stratification of the brain structures, particularly the metencephalon, an embryonic part of the hindbrain that goes on to develop the cerebellum and the pons, which is quite separated from the remaining cluster of tissues (figure 1). Within the larger cluster, composed mostly of cortex tissues, there was grouping by embryonic origin as well.

To assess whether embryonic origin determined clusters in gene expression in the adult human brain, we performed hierarchical clustering of all 414 ABA structures expressing the 25 pathways with the highest variance as well as cluster tissues based on the lowest 25 variances in pathways. The assumption is that variance in a particular pathway is an indicator of whether the pathway is involved in housekeeping functions of the cell or if the pathway is part of a tissue-specific gene expression signature. Clustering of tissues using high variance pathways yielded more clustering by embryonic tissue than clustering of low variance pathways (figure 2). Interestingly, several mitochondrial maintenance pathways were among the lowest variance pathways in Reactome.

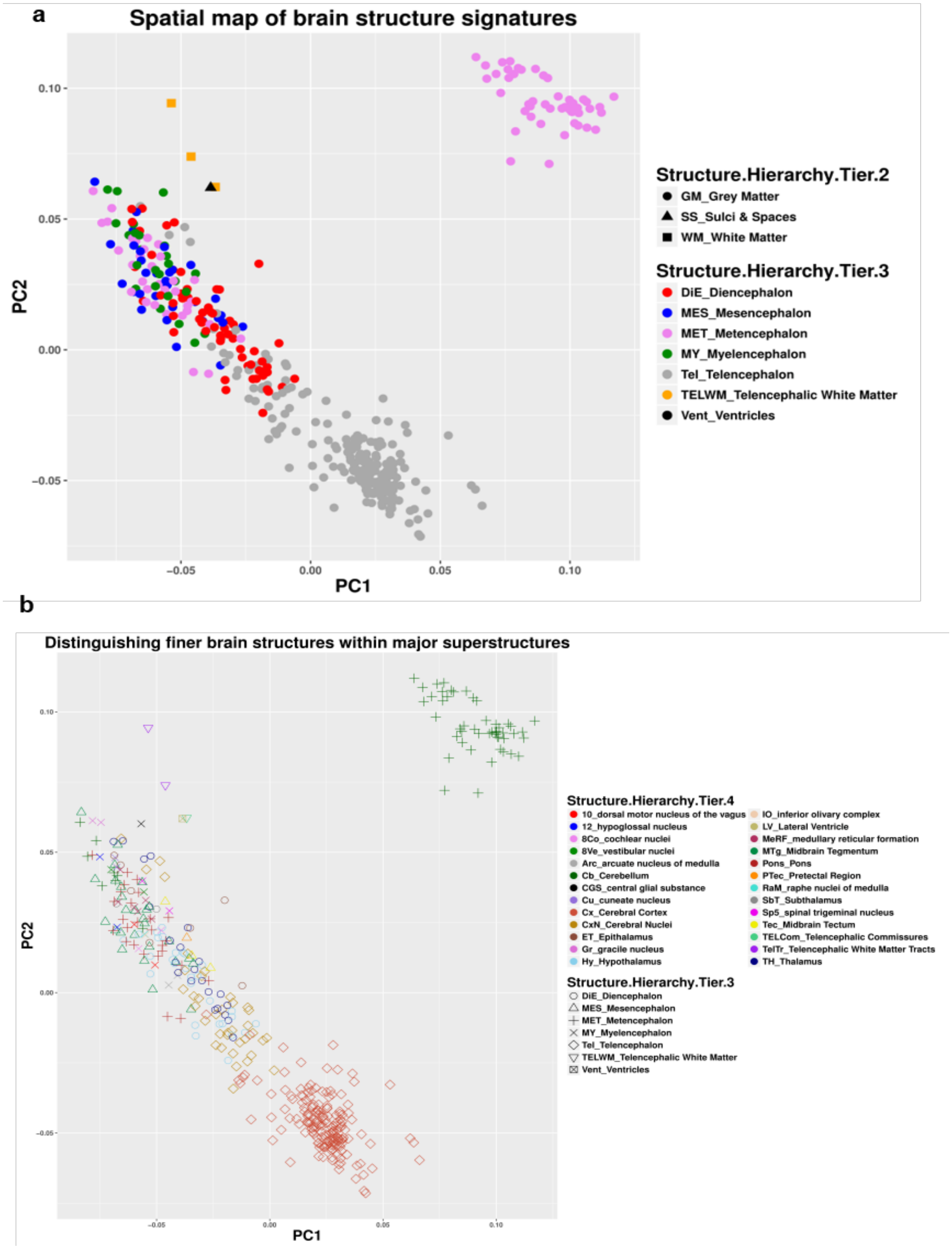

Figure 1. Principal component analysis of 414 structures of healthy adult brains from ABA shows two main clusters. a) The tissues are mainly grouped on the basis of embryonic origin and are b) separated into a cluster of mostly cortex tissues and another cluster of solely cerebellum tissue.

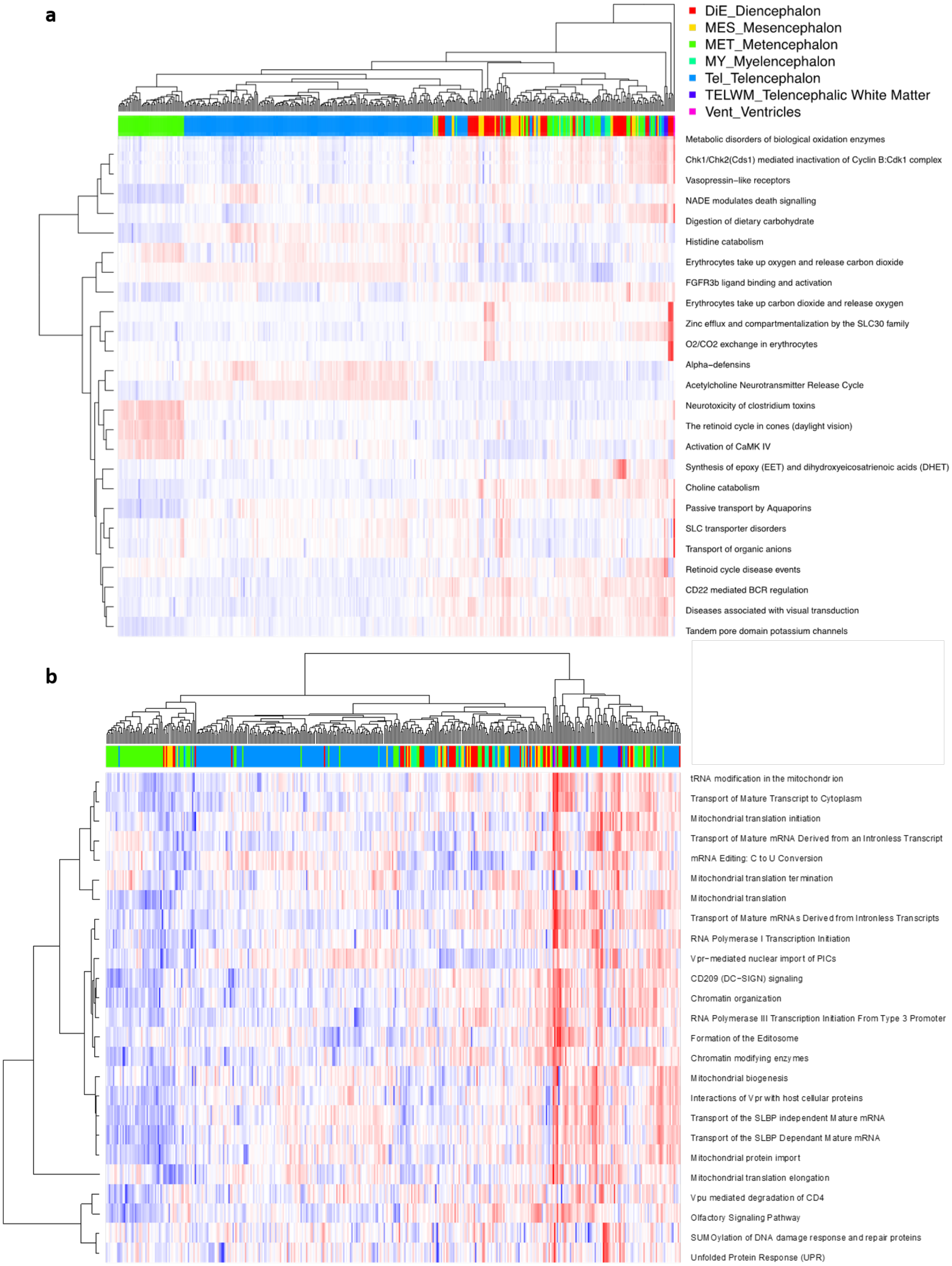

Figure 2. Hierarchical clustering of ABA tissue gene expression of Reactome pathways. a) Using the 25 highest variance pathways, tissues were clustered according to embryonic origin, in particular metencephalon and telencephalon tissues. b) Tissues did not group as clearly by embryonic origin when clustered on the basis of the 25 lowest variance pathways.

Variance of Pathways Enriched in Alzheimer's Across Brain Structures

We next calculated which genes were differentially expressed in Alzheimer's from a GEO dataset, and determined the enriched pathways in the structures studied in that particular dataset, namely, samples of the gyri, hippocampus, and cortex. From the Reactome database, we found 97 pathways that were enriched with an FDR of less than 5% (table 1). Among the eight enriched pathways with the highest significance we found immune system and signaling pathways, both of which have been previously implicated in Alzheimer's disease[19,20]. The immune system pathway variance was plotted across the 414 structures of the ABA (figure 3). We can thus obtain the proportion of high variance pathways predicted to be enriched in each structure, and rank them according to highest proportion of high variance pathways (table 2). Among the structures containing the highest proportion of high variance pathways enriched in Alzheimer's is the cortex, which is one of the tissues that is greatly deteriorated in Alzheimer's[21].

Discussion

Tissue-specific study of the brain is essential for understanding differences in gene expression in neurological disorders. However, laboratories are unable to perform these time- and cost-intensive experiments on a brain-wide scale. We suggest using ABA along with analysis of pathway enrichment to bridge the understanding we currently have of diseased tissue gene expression in the brain.

Using principal component analysis and hierarchical clustering, which suggested gene expression of tissues was associated to embryonic origin. We also performed independent component analysis of the ABA gene expression data, where we observed only one cluster with two independent components (figure 4). However, we were unable to obtain information about how much of the variation in gene expression was accounted for using each principal component, which made it difficult to compare to our principal component analysis. Still, using hierarchical clustering we found that high-variance pathways were more likely to cluster by embryonic origin than low-variance pathways. This may be attributed to the involvement of high-variance pathways in tissue-specific functions, whereas low-variance pathways may be more likely to be part of cellular housekeeping. The variance of a pathway may also suggest how tightly regulated the pathway is. Of note, mitochondrial maintenance pathways were highly represented in the 25 lowest-variance pathways. Further study should be performed on these pathways, which have been implicated in morphological changes to the mitochondria of Alzheimer's patients[22].

Our study found from GEO microarray gene expression data of post-mortem Alzheimer's tissue that the top eight enriched pathways ranked by significance included immune system and signaling pathways, in accordance with the literature. We analyzed the variance in these enriched pathways across the 414 healthy tissues in ABA, which produces a brain-wide view of variance, which we used as an indication of tissue-specific activity in that pathway. Among the tissues with the highest proportion of high-variance enriched pathways, we identified tissues that are known to be affected in Alzheimer's disease, in particular, the cortex. The cortex is subject to extensive degeneration in Alzheimer's, which results in the loss of cognitive functions that is typical of the disease. We would like to evaluate the effects of disease on these high-variance pathways, and to determine whether variance can be used as proxy for tissue-specific gene expression functionality.

Top 8 Enriched Reactome Pathways in the Alzheimer's dataset	Corrected p-values
Immune System	4.37e-07
Gene Expression	4.83e-06
Signaling by Rho GTPases	8.04e-05
Signaling by NGF	1.61e-04
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins	1.68e-04
Cellular responses to stress	1.68e-04
Innate Immune System	5.78e-04

Table 1. Top 8 enriched Reactome pathways in Alzheimer's GEO dataset, ranked by order of significance.

Brain structure category	Structures Harboring Enriched High-Variance Pathways	P-value
Cerebral Cortex	39.7%	2.0E-03
Cerebellum	13.2%	3.2E-2
Cerebral Nuclei	9.9%	1.4E-2
Pons	8.8%	0.95
Hypothalamus	7.2%	6.6E-4

Table 2. Proportion of structures in each brain category harboring high-variance enriched pathways.

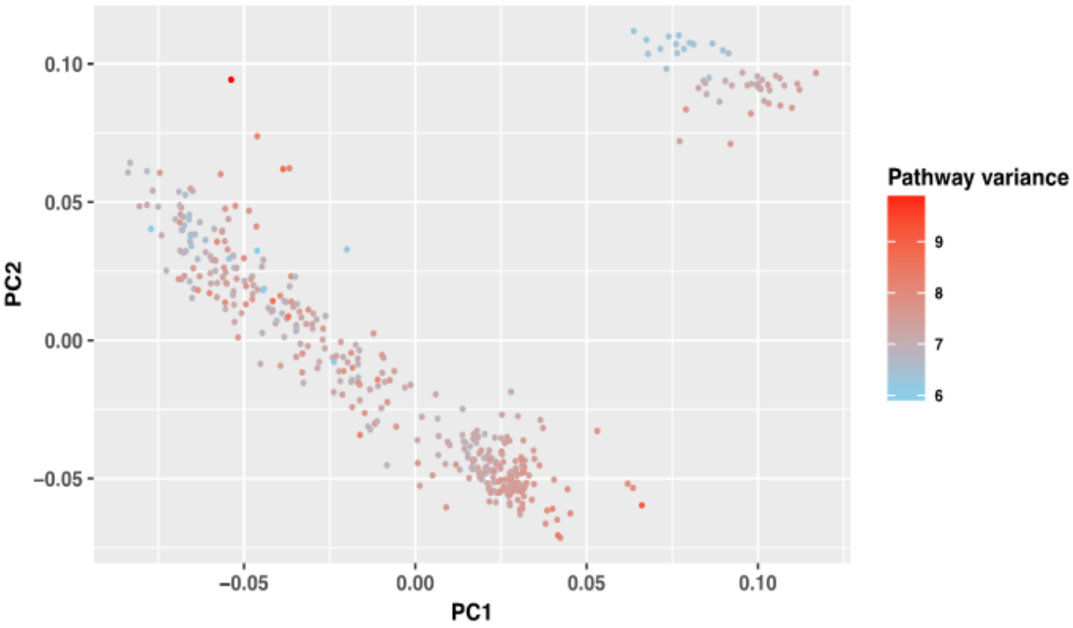

Figure 3. Variance of the immune system Reactome pathways plotted as a function of color on the principal component analysis scatterplot of ABA structures. Red indicates high variance and blue indicates low variance.

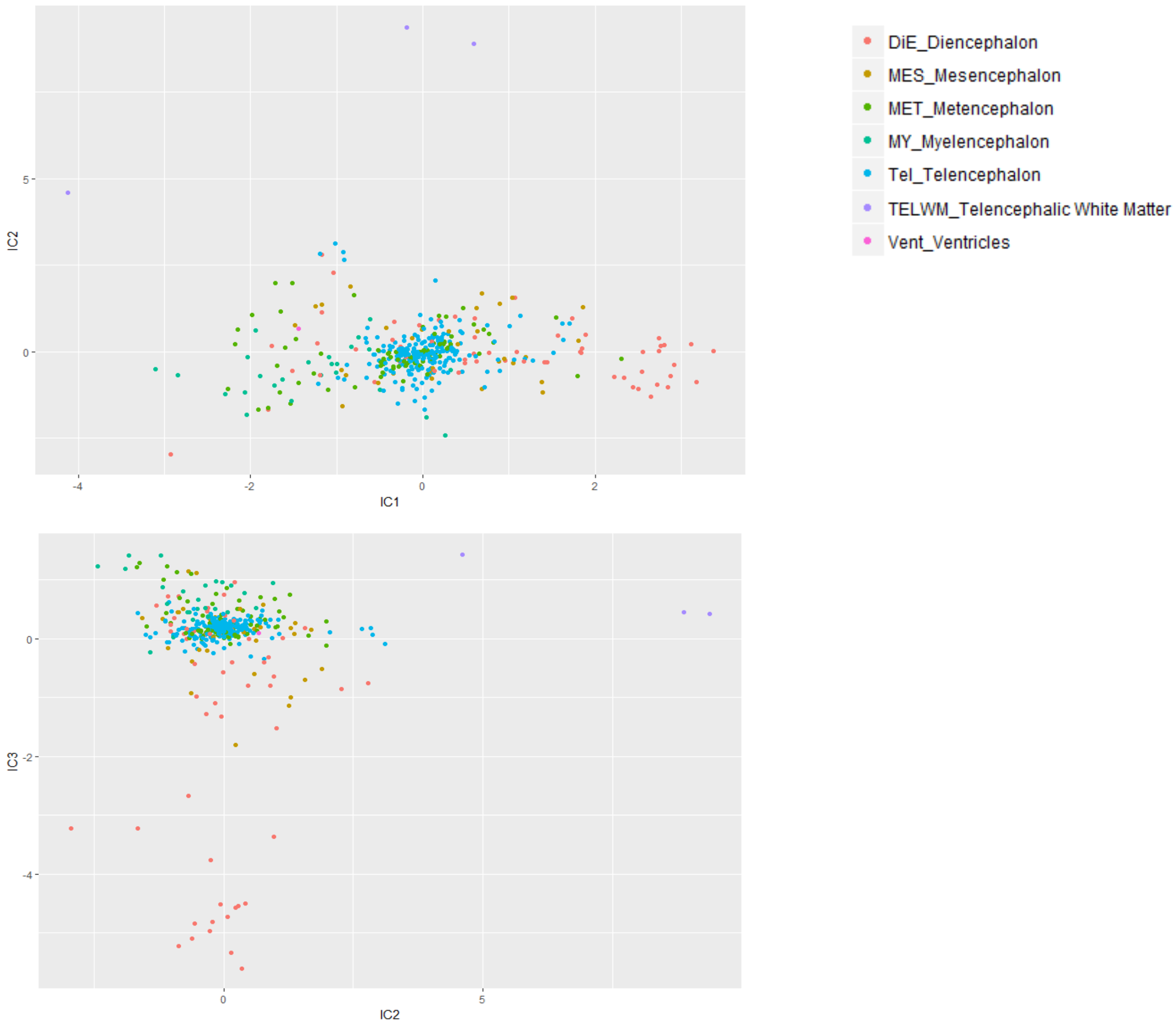

Figure 4. Independent component analysis of the ABA gene expression data, where we observed only one cluster with two independent components.

Conclusion

Despite the prominence of central nervous system disorders in the population and their cost in terms of health care and quality of life of affected individuals, there is a lack of understanding of the pathways affected by these diseases across brain structures. The human brain has a surprising degree of complexity both in terms of communication between brain structures and the variety of functions of the structures themselves. This is reflected in the heterogeneity of gene expression we observe across brain tissues. Using principal component analysis and pathway enrichment analysis we were able to glean trends in the variance of these pathways across all 414 structures of the brain surveyed in ABA. Future work will focus on determining the role of low-variance pathways, in particular mitochondrial maintenance pathways, in Alzheimer's disease, and to study the variance of gene expression in enriched pathways in Alzheimer's in individuals affected by the disease.

References

1. Services, H. Global Health and Aging. *NIH Publ. no 117737* 1, 273–277 (2011).
2. Altar, C. A., Vawter, M. P. & Ginsberg, S. D. Target identification for CNS diseases by transcriptional profiling. *Neuropsychopharmacology* **34**, 18–54 (2009).
3. Herrmann, N., Chau, S. A., Kircanski, I. & Lanctôt, K. L. Current and emerging drug treatment options for alzheimers disease: A systematic review. *Drugs* 71, 2031–2065 (2011).
4. Colangelo, V. *et al.* Gene expression profiling of 12633 genes in Alzheimer hippocampal CA1: Transcription and neurotrophic factor down-regulation and up-regulation of apoptotic and pro-inflammatory signaling. *J. Neurosci. Res.* **70**, 462–473 (2002).
5. Shinn, A. K., Baker, J. T., Lewandowski, K. E., Öngür, D. & Cohen, B. M. Aberrant cerebellar connectivity in motor and association networks in schizophrenia. *Front. Hum. Neurosci.* **9**, 134 (2015).
6. Wu, T. & Hallett, M. The cerebellum in Parkinson's disease. *Brain* **136**, 696–709 (2013).
7. Aubry, S. *et al.* Assembly and interrogation of Alzheimer's disease genetic networks reveal novel regulators of progression. *PLoS One* **10**, 1–25 (2015).
8. Hartley, D. *et al.* Down syndrome and Alzheimer's disease: Common pathways, common goals. *Alzheimer's and Dementia* **11**, 700–709 (2015).
9. Chang, J., Gilman, S. R., Chiang, A. H., Sanders, S. J. & Vitkup, D. Genotype to phenotype relationships in autism spectrum disorders. *Nat. Neurosci.* **18**, 191–198 (2015).
10. Edgar, R., Domrachev, M. & Lash, A. E. *Gene Expression Omnibus: NCBI gene expression and hybridization array data repository*. *Nucleic Acids Res* **30**, 207–210 (2002).
11. Sunkin, S. M. *et al.* Allen Brain Atlas: An integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Res.* **41**, (2013).
12. Liang, W. S. *et al.* Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. *Proc. Natl. Acad. Sci.* **105**, 4441–4446 (2008).
13. Loring, J. F., Wen, X., Lee, J. M., Seilhamer, J. & Somogyi, R. A gene expression profile of Alzheimer's disease. *DNA Cell Biol.* **20**, 683–95 (2001).
14. R Core Team. R: A Language and Environment for Statistical Computing. *R Found. Stat. Comput.* **version 3**, 3503 (2016).
15. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **47**, 43-7 (2015).
16. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A.**100(16)**, 9440-5 (2003).
17. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway Knowledgebase. Nucleic Acids Res. **44(D1)**, D481-7(2016).
18. Grote S. ABAEnrichment: Gene expression enrichment in human brain regions. R package version 1.2.2.; 2016.
19. Marsh, Samuel E., Alborz Karimzadeh, Edsel M. Abud, Anita Lakatos, et al. The Adaptive Immune System Critically Regulates Alzheimer's□□ Disease Pathogenesis By Modulating Microglial Function.*Alzheimer's & Dementia*. **113**, 9 (2016).
20. Godoy, J. A., Rios, J. A., Zolezzi, J. M., Braidy, N., & Inestrosa, N. C. Signaling pathway cross talk in Alzheimer's disease. *Cell Communication and Signaling : CCS*. ***12***, 23 (2014).
21. Thompson, P. M., M. S. Mega, R. P. Woods, C. I. Zoumalan, C. J. Lindshield, R. E. Blanton, J. Moussai, C. J. Holmes, J. L. Cummings, and A. W. Toga. "Cortical Change in Alzheimer's

Disease Detected with a Disease-specific Population-based Brain Atlas." *Cerebral Cortex* **11.1**, 1-16 (2001).

22. Paula I. Moreira, Cristina Carvalho, Xiongwei Zhu, Mark A. Smith, George Perry. Mitochon- drial dysfunction is a trigger of Alzheimer's disease pathophysiology. BBA - Molecular Basis of Disease, Elsevier, 2009, 1802 (1), pp.2.