Analysis of Microarray Expression Across Brain Diseases Provides Additional Support for the Genetic Relationship Between Tumorigenesis and Aging

Yasutaka Tanaka BA, Chenchao Zang BA, Ian Johnson BA

Columbia University, New York, NY, USA

Abstract:
In this research study, we aim to compare the expression levels of differentially expressed genes across diseases of the human brain. We gather publicly available microarray datasets and use multiple methods of pairwise t-testing including Significance Analysis of Microarrays[1] and Binary Limma from the Bioconductor package[2] to derive our measurements. We also perform pathway enrichment analysis on the differentially expressed genes (DEGs) to explore common underlying molecular mechanisms. Results indicate the presence of several relationships among these conditions, the most significant of which was between samples from glioma tissue and samples from the tissue of aged individuals. These results suggest a common underlying mechanism for these conditions, which is supported in current literature[3,4,5].

1. Introduction

There has been increased interest in microarray experimentation in recent years, with a large number of microarray datasets becoming publicly available for analysis by independent researchers. Several efforts have been undertaken to quantitatively describe the genetic signature of both diseased and healthy brain tissue. These projects include work done at the Allen Institute for Brain Science, and other organizations that have contributed microarray expression datasets to the NCBI Gene Expression Omnibus (GEO).

However, while the data is growing, it remains a difficult task to make comparisons between these datasets, and relatively few studies have been done to integrate this microarray brain data on a large scale. Factors including lab conditions, sampling region, choice of probeset, sample tissue type, cell type, patient history, gender, age, and a large number of additional hidden variables limit further investigation into the relationships between the expression values[6]. It is difficult to attempt to normalize for all of these factors in a simple control vs treatment experiment, and thus it is even more difficult to try to compare across samples from different experiments altogether. Yet the abundance of data and research in this field calls for a way to integrate information from these disparate resources. This study utilizes statistical hypothesis testing to compute quantitative measurements for the degree of relatedness of diseases and conditions of the tissue of the human brain.

2. Methods

2.1 Data Collection

We began this study with relatively little guess as to the expected relationships between brain diseases, and thus we adopted a "cast a wide net" approach for our analysis, and worked towards implementing a pipeline to process as many datasets as possible. This required that we choose datasets that were comparable and had adequate control samples.

We began our search for data by querying the Gene Expression Omnibus geo dataset browser service with the search term "brain", and selecting datasets corresponding to human specimens. We then filtered this list down from 55 to 18 datasets by limiting only to the Affymetrix GPL570 platform. Using one distinct platform helped to ensure that the same number of probes were used and allowed for more robust comparisons. Two datasets were removed due to irregular results from MAPlots and SAM graphs. The final datasets together represented a total of 8 distinct conditions. We used only datasets of the GDS class in R, which is curated by the GEO team to allow for easier analysis. The less robust GSE class contains original author-supplied data and may not be formatted correctly. A more detailed summary of accession numbers, aggregated dataset info, and individual study purpose information can be found in the source code repository:

https://github.com/YasChenSon-Bioinformatics/BrainDiseaseCors/blob/master/GPL_570_metadata.csv

Much of our subsequent work involved exploratory analysis, and processing the data from the different datasets to an integrable form. This work included cleaning and removing missing and invalid columns, as well as ensuring that

datasets have adequate overlap in terms of probes, and not too many N/A values. Presence of control samples in the "disease.state" column of the data frames was a requirement for inclusion in the study. All data handling and analysis was done in R.

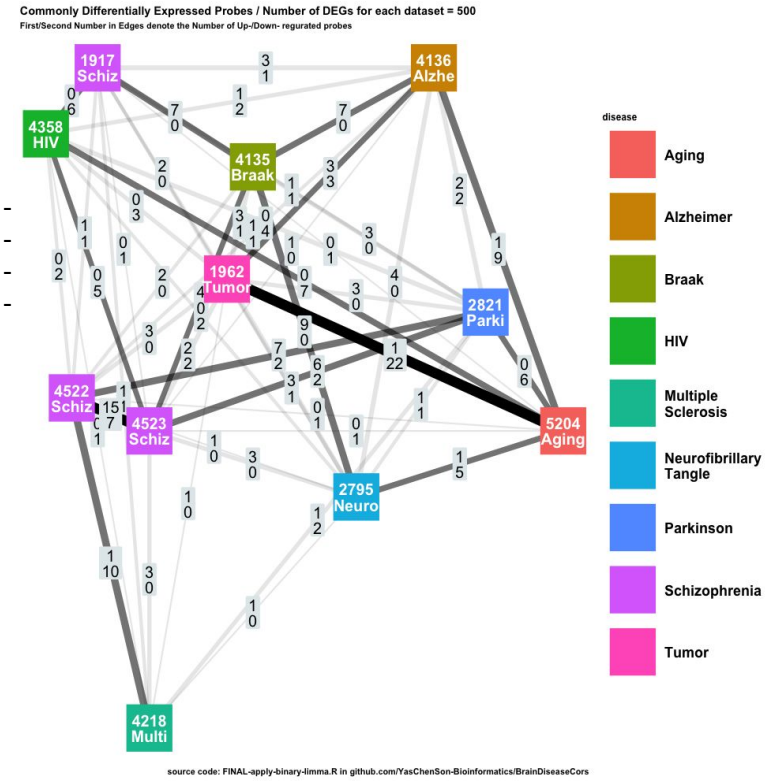
2.2 Statistical analysis

Datasets were passed through our pipeline through additional statistical analysis steps, in which we used linear modeling with empirical bayes from the limma package to produce sets of DEGs for each disease, as well as the SAM package.

Exploratory analysis showed that there was a bias due to the gender variable present in the datasets (Supplementary figure 3). We normalized for this by leaving out the females samples in order to reduce the amount of variation due to underlying variables that may be unbalanced between condition and control cases.

Binary limma was conducted by splitting each of the datasets into condition and control samples, and in certain cases making adjustments for multi-class data, such as in the aging study, in which samples from individuals of age < 70 years were labeled as controls. Tests were performed at various levels of significance by using the topTable function to select out the only the top N most significantly differentially expressed genes (DEGs). Overlap of these DEGs represented the count values for the commonly differentially expressed probes (CDEPs) used in the network graph of disease relationships. A threshold level of the top 500 DEGs was used for further analysis. The reason that we choose the top 500 genes based on the rank of P-value is that the number of DEGs is very small if we limit to P-values less than 0.05. To validate the results of our derived lists of overlapped DEGs, we performed the same analysis with randomly permuted sample labels and compared the original result to the average across 1000 random permutations of the labels. This score indicated the number of times that our randomly permuted samples resulted in a larger number of DEGs than the observed value.

3. Results



Our results for the binary limma analysis (Figures 1 and 2) showed the top number of CDEPs occurring between the following non-control conditions:

- Aging & Glioma - 23 CDEP
- Schizophrenia & Schizophrenia - 22 CDEP
- Schizophrenia & MS - 11 CDEP
- Alzheimer's & Aging - 10 CDEP

Next, steps were taken to investigate two of the data sets that exhibited the most correlated results, GDS1962 (glioma tissues) and GDS5204 (age effect in the frontal cortex). In these subsequent analyses we plotted the expression values of DEGs and performed enrichment analysis using pathways from ReactomeDB[5] to identify common pathways.

Figure 1. CDEP relationships among brain diseases and conditions

	GDS5204 Aging	GDS4522 Schizoph.	GDS4523 Schizoph.	GDS4358 HIV	GDS4218 Sclerosis	GDS4136 Alzheim.	GDS4135 Braak	GDS2821 Parkins.	GDS2795 Neuro	GDS1962 Tumor	GDS1917 Schizoph.
GDS5204 Aging	-			2		1			1	1	
GDS4522 Schizoph.	1	-	15	4	1	1	2	7	1	3	1
GDS4523 Schizoph.	1	7	-	4	3	1	4	6	3	2	
GDS4358 HIV				-		6	4		5	4	2
GDS4218 Sclerosis		10			-			1	1	1	1
GDS4136 Alzheim.	9	1				-	7	2	4	3	3
GDS4135 Braak			1				-	3	9	3	7
GDS2821 Parkins.	6	2	2		2	2		-	1	3	1
GDS2795 Neuro	5							1	-	3	1
GDS1962 Tumor	22		2			3	1		1	-	2
GDS1917 Schizoph.	1	1	1		1	1		1	1		-

CDEPs up-regulated by diseases

CDEPs down-regulated by diseases

Figure 2. CDEP counts among brain diseases and conditions

Pathway Name (Reactome DB)	N of Genes in This Pathway	n_DEG GDS5204	n_DEG GDS1962	p-value GDS5204	p-value GDS1962	FDR GDS5204	FDR GDS1962	Differentially Expressed Genes (GDS5204)	Differentially Expressed Genes (GDS1962)
Trafficking of AMPA receptors	17	5	3	2.20E-05	0.005	0.02	0.24	O60359|P48058|P42261| P24588|Q12959	O60359|Q13554|Q9H4G0
Phase 0 - rapid depolarisation	45	6	4	0.0003	0.01	0.25	0.57	Q08289|P61328| Q9NY72|Q92915| O60359|P62158	P62158|Q13936|O60359| Q13554
MASTL Facilitates Mitotic Progression	10	3	2	0.001	0.01	0.27	0.61	P56211|O43768|P67775	P56211|P67775
Ca2+ pathway	59	6	4	0.001	0.04	0.27	0.74	P63215|P62158|P16298| O14775|Q9UBE8| Q9NQ66	P62158|Q9UBE8|P63098| O14775
Interactions of neurexins and neuroligins at synapses	60	6	8	0.001	4.60E-05	0.27	0.06	Q92796|Q86UE6| Q9HDB5|Q9Y4C0| O14490|Q86YM7	Q9HDB5|Q9Y4C0| Q16623|O43581|Q9HAP6| Q96HC4|Q02410|Q9H4G0
Cam-PDE 1 activation	4	2	2	0.002	0.002	0.27	0.22	P62158 P54750	P62158 P54750
CREB phosphorylation through the activation of	4	2	2	0.002	0.002	0.27	0.22	Q16566|P62158	P62158|Q8N5S9
DARPP-32 events	24	3	3	0.01	0.01	0.66	0.55	P67775|P62158|P16298	P62158|P63098|P67775
Synthesis of IP3 and IP4 in the cytosol	26	3	3	0.01	0.01	0.69	0.61	P62158|P23677| Q9NQ66	O43426|P62158|O15357
CLEC7A (Dectin-1) induces NFAT activation	11	2	2	0.02	0.02	0.70	0.62	P62158|P16298	P62158|P63098
DSCAM interactions	11	2	2	0.02	0.02	0.70	0.62	Q13153|P45983	O95631|Q13153
LGI-ADAM interactions	14	2	3	0.03	0.003	0.96	0.22	O60359|O75077	O75078|Q16623|O60359

After subsetting our analysis to just GDS1962 and GDS5204, our results for pathway enrichment analysis indicated that the ReactomeDB pathway "trafficking of AMPA receptors" was the pathway most significantly enriched for these two conditions, with several additional pathways significantly enriched as well (Figure 3). A notable correlation was found in the CAM-PDE 1 activation pathway, as both datasets resulted in the same two genes, P62158 and P54750 being enriched.

Figure 3. Pathway enrichment analysis results for GDS5204 (Aging), GDS1962 (Glioma)

4. Discussion

The co-occurrence of such a large number of DEGs between several of the top correlated conditions in this study is an interesting finding. The high statistical significance of these results suggests an underlying relationship between several pairs of these conditions (See supplementary tables 1, and 2 for p-values). It should be noted that we would like to perform additional investigation into several inconsistencies including the apparent lack of a relationship

between GDS1917 and the other Schizophrenia datasets, as we would expect these datasets to be more closely related.

It was also expected that we would find the largest CDEP relationship between GDS4522 and GDS4523 (schizophrenia studies conducted by GlaxoSmithKline in the UK), as these two datasets shared a common disease, and were also conducted in the same laboratory. However while these two datasets did exhibit significant DEG overlap, we made the discovery further into our analysis that these two datasets actually used the Affymetrix U133A array, while others used Affymetrix U133plus, and thus the two schizophrenia studies actually had fewer probes overall, so it was likely that there would have been more DEG overlap had these authors also used the U133plus array.

It was surprising that an even more significant result was found between aging and glioma tissue, a result which is supported by current literature[3,4,5]. Because of the results of permutation testing that was performed after computing our original list of DEGs, it is unlikely for these two conditions to have this number of CDEPs by chance alone. The large statistical significance of this result in both the gene expression analysis and pathway analysis ($p < .001$) provides additional support for a relationship between tumors and and an aged state of the human brain, and may suggest a common underlying biological mechanism for the two.

5. Conclusion

With this research, we hope to have made progress towards the classification of the relationships between various disease and condition states in the brain, and the quantification of their genetic correlations. Additionally we hope that these results may provide insight to help researchers pursue further investigation of our proposed relationship between glioma development and the aging processes, and that this study will help other researchers to utilize similar methods of dataset integration in order to further characterize relationships amongst brain diseases, or to diseases of other etiologies. In addition, future work for researchers might involve investigating the specific genes that are found to be related to both of these two conditions, as well as their association with the upregulated pathways.

6. Supplemental Data and Figures:

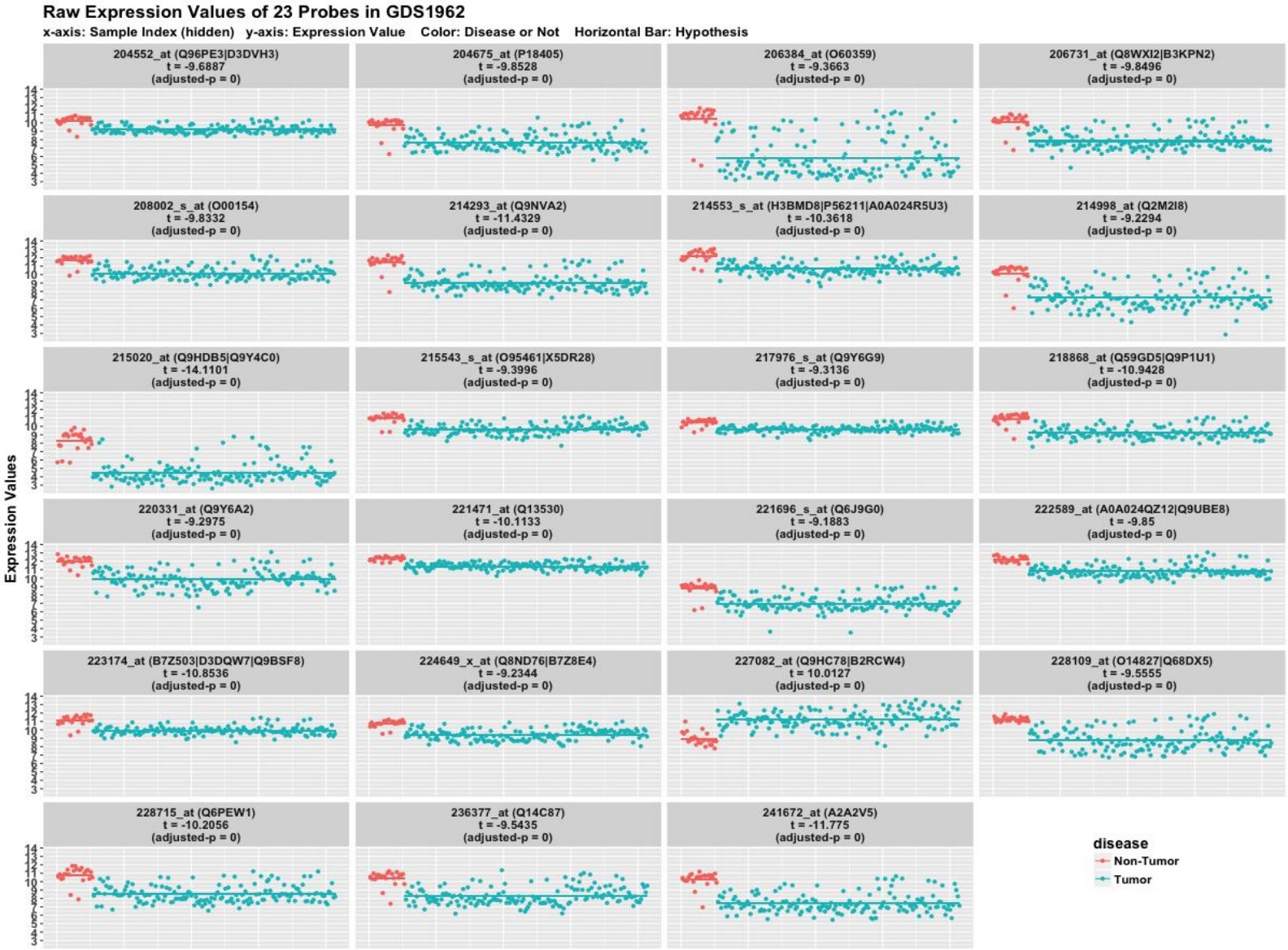
Github: https://github.com/YasChenSon-Bioinformatics/BrainDiseaseCors/

	5204	4522	4523	4358	4218	4136	4135	2821	2795	1962	1917
5204	NA	1	1	0.267	1	0.746	1	1	0.622	0.638	1
4522	0.558	NA	0.002	0.104	0.687	0.785	0.391	0.01	0.528	0.149	0.581
4523	0.565	0.004	NA	0.09	0.205	0.779	0.12	0.013	0.193	0.367	1
4358	1	1	1	NA	1	0.039	0.122	1	0.045	0.105	0.246
4218	1	0.009	1	1	NA	1	1	0.735	0.694	0.692	0.688
4136	0	0.387	1	1	1	NA	0.029	0.554	0.145	0.299	0.239
4135	1	1	0.539	1	1	1	NA	0.284	0.006	0.261	0.017
2821	0.01	0.139	0.278	1	0.159	0.169	1	NA	0.699	0.228	0.739
2795	0.037	1	1	1	1	1	1	0.581	NA	0.23	0.674
1962	0	1	0.278	1	1	0.048	0.463	1	0.521	NA	0.343
1917	0.606	0.528	0.663	1	0.511	0.52	1	0.612	0.655	1	NA

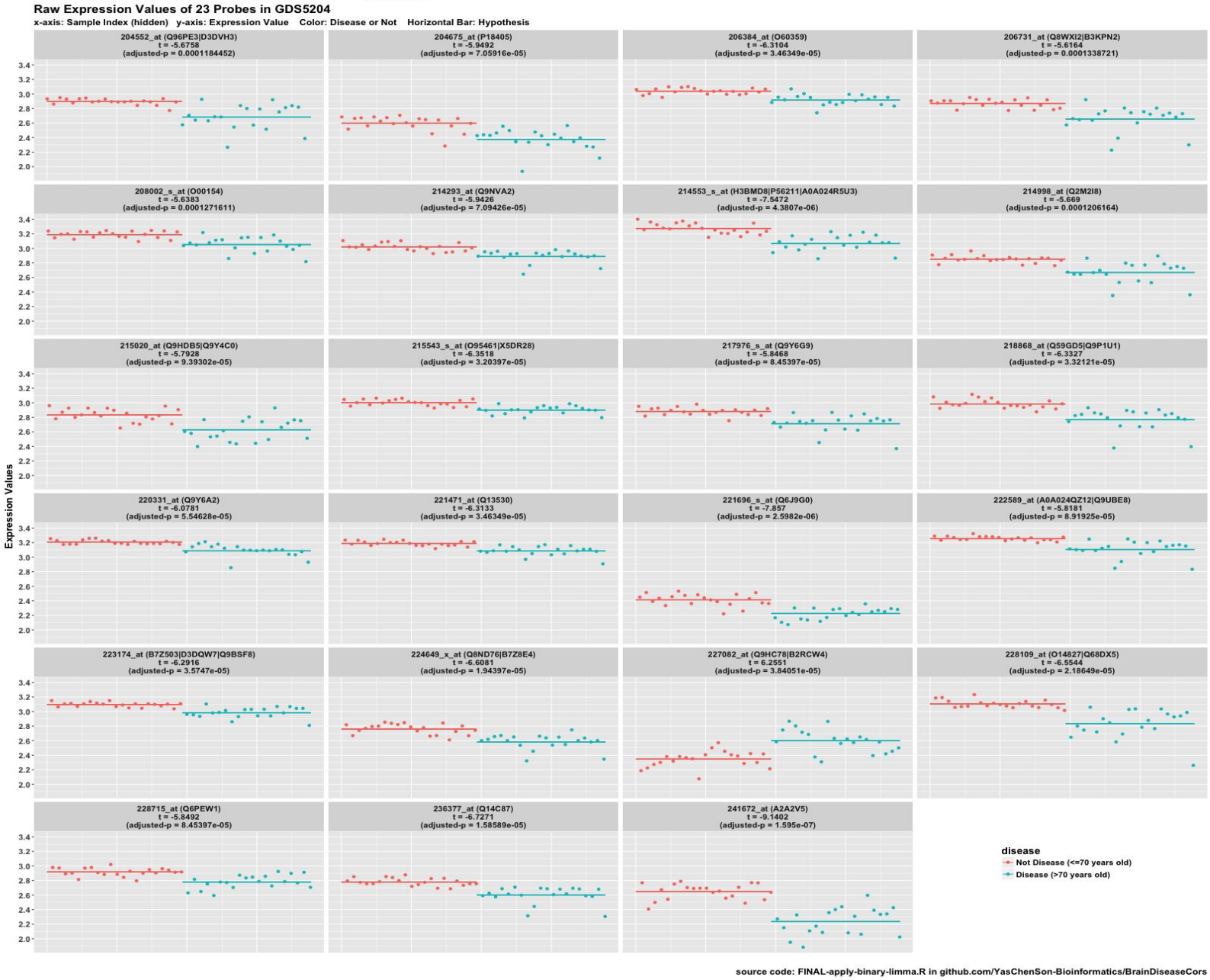
Supplemental Table 1: P-values for binary limma. Scores indicate the percent chance that a random permutation of the labels of the datasets would result in a larger number of common DEGs between the datasets than was observed when running with the correct labels.

	5204	4522	4523	4358	4218	4136	4135	2821	2795	1962	1917
5204	0.09	1	1	0.09	0.06	0.31	0.57	0.15	0.42	0	0.18
4522	1	1	0.27	0.44	0.66	0.09	1	1	1	1	0.54
4523	1	0.27	0.62	1	1	0.58	0.51	0.14	0.06	1	0.13
4358	0.09	0.44	1	0.56	0.31	1	0.18	0.28	0.17	0.45	0.06
4218	0.06	0.66	1	0.31	0.61	0.62	1	0.61	1	0	0.57
4136	0.31	0.09	0.58	1	0.62	0.29	1	0.51	0.61	0.18	1
4135	0.57	1	0.51	0.18	1	1	0.78	0.66	0.21	0.3	0.55
2821	0.15	1	0.14	0.28	0.61	0.51	0.66	0.26	0.2	0.08	0.71
2795	0.42	1	0.06	0.17	1	0.61	0.21	0.2	0.02	0	0.09
1962	0	1	1	0.45	0	0.18	0.3	0.08	0	0.02	0.58
1917	0.18	0.54	0.13	0.06	0.57	1	0.55	0.71	0.09	0.58	0.05

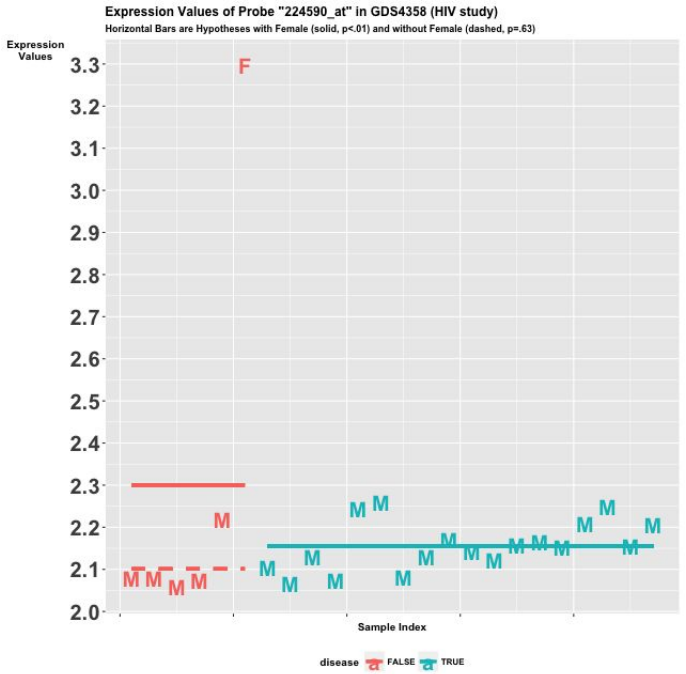
Supplemental Table 2: P-values for pathway enrichment. Scores indicate the percent chance that a random permutation of the labels of the datasets would result in a larger number of common pathways than was observed with the correct labels.



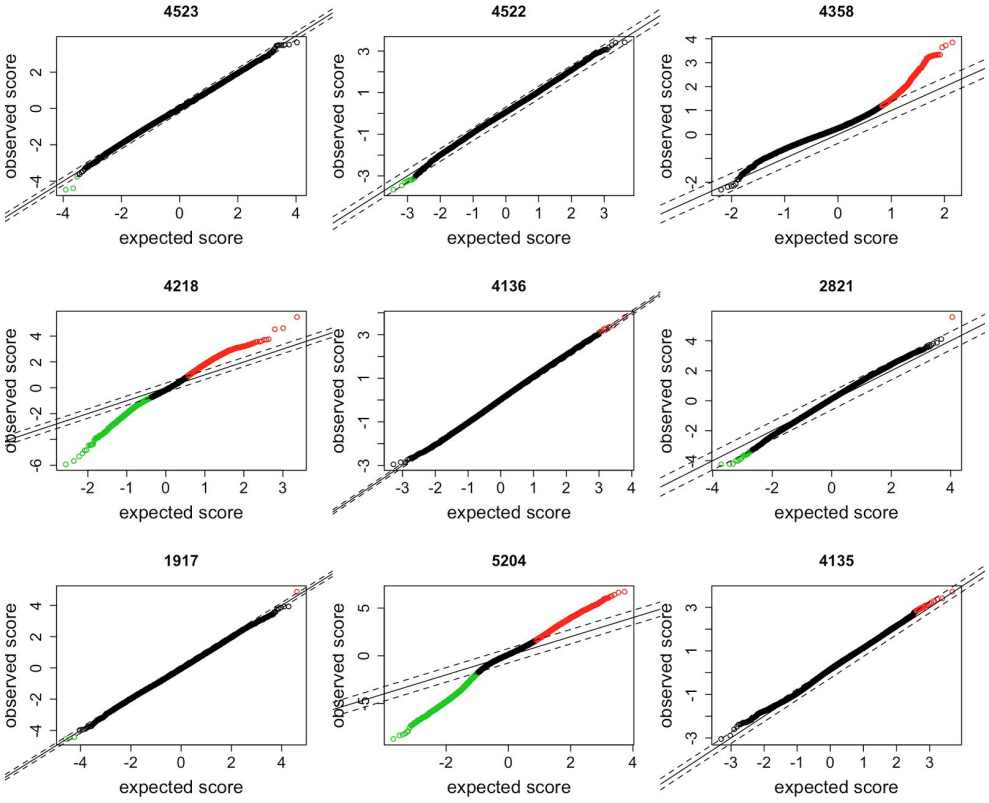
Supplemental Figure 1: Expression values of CDEGs for Glioma vs control cases in 1962 (glioma tumorigenesis).



Supplemental Figure 2: Expression values of CDEGs for aged versus control cases in 5204 (aging study).



Supplemental Figure 3: Confounding gender effect exhibited by DEP 224590_at in GDS4358



Supplemental Figure 4: Significance Analysis of Microarray result for datasets with binary disease.state predictors

References

1. http://statweb.stanford.edu/~tibs/SAM/
2. https://www.bioconductor.org/
3. Toren Finkel, Manuel Serrano, and Maria A. Blasco, The common biology of cancer and ageing, Nature 448 (2007), no. 7155, 767–774.
4. William B. Ershler and Dan L. Longo, Aging and cancer: Issues of basic and clinical science, Journal of the National Cancer Institute 89 (1997), no. 20, 1489–1497.
5. Anisimov VN. Biology of aging and cancer. Cancer Control. 2007;14:23–31.
6. Maycox P. R. et al. . Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. Mol Psychiatry 14, 1083–1094, (2009).10.1038/mp.2009.18
7. Konradi C. Gene expression microarray studies in polygenic psychiatric disorders: applications and data analysis. *Brain Res Rev* 2005; **50**: 142–155.
8. http://www.reactome.org/PathwayBrowser/
9. https://www.ncbi.nlm.nih.gov/sites/GDSbrowser
10. Madan Babu. "11." *An Introduction to Microarray Data Analysis*. N.p.: n.p., n.d. 225-36. Print.
11. Mirnics K, Middleton FA, Lewis DA, Levitt P. Analysis of complex brain disorders with gene expression microarrays: schizophrenia as a disease of the synapse. Trends Neurosci. 2001;24:479–486.
12. https://www.bioconductor.org/packages/devel/bioc/vignettes/affyPLM/inst/doc/MAplots.pdf