



# A Quantitative Comparison of Microarray Gene Expression Profiles Across Diseases in the Human Brain

Yasutaka Tanaka BS, Chenchao Zang BS, Ian Johnson BA

Columbia University, New York, NY, USA

## Abstract

*Gene expression analyses of human brain have been used to find genes associated with several brain diseases. However, the gene expression commonality among a large amount of brain diseases has not been largely characterized. Most of those studies were restricted either to an individual disease or to a small number of neurodegenerative diseases. Because of different factors involved in the collection and processing of microarray data, current research has been limited by the ability to compare microarray expression values across datasets from different sources and experiments. By drawing from many datasets, and leveraging statistical measures of correlation, we hope to gain some insight into large scale relationships between brain diseases. The methods that we will employ include standard statistical analyses of variance including vector euclidian distance, Pearson correlation, and cosine similarity, which lend themselves to easy interpretation through graphical representation. In addition, we are open to investigating and exploring supplementary methods of investigation including clustering methods such as Ward's minimum variance, the single and complete linkage methods, and centroid clustering.*

## 1. Introduction

The latest developments in the field of neuroscientific research involve several large scale efforts to derive gene expression values, as well as to map connections between all its regions. Both of these initiatives have been taken up by researchers at the Allen Institute for Brain science, as well as by others who have also contributed data to the NCBI Gene Expression Omnibus repository for public use. However, many factors involving in this discrepancy, including lab conditions, sampling region, chosen probesets, sample tissue type, cell type, patient history, gender, age and a large number of additional dimensions limit further research development. While it is difficult enough to attempt to normalize for all of these factors in a simple control vs

treatment experiment, it is even more difficult to try to compare ~~from~~ different samples from different experiments altogether. Yet the abundance of data and research in this field draws us to find a way to integrate information from these disparate resources.

The similarity analysis of several brain diseases at gene expression level has been summarized. There are various methods have been developed identifying expressed genes, correlated genes, and significant functional modules. However, previous research has been restricted to siloed neurodegenerative diseases such as Alzheimer's and Parkinson's disease, without comparison to other diseases. Integrative analysis using combined data from previously uncombined microarray datasets accompanied by normalization to controls makes this study an encouraging attempt to explore more correlations among multiple brain diseases than was possible in previous studies. By attempting to account for the various sources of discrepancy between gene expression datasets, we can increase our chances of drawing meaningful conclusions from the data. It is possible to use normalization methods including subtraction of background expression values, and correlation measures to maximize the information gained about relationships between various disease states in the data.

Much of our work will involve the initial data processing from the different datasets. This work will include cleaning and removing missing and invalid sections, as well as mapping between data sets that overlap in terms of probes, brain regions, and other factors. Most of the datasets include control samples alongside the treated samples, which will allow for normalization. These normalized samples can then be compared to the normalized values from other datasets, with the assumption that the cases and controls were similar enough across the other dimensions to give an accurate representation of just the disease or condition being investigated.



## 2. Methods

### 2.1 Data Sources

Our microarray data is from the NCBI GEO data repository, accessed using the GEOQuery package in R. There are a total of 17 datasets representing a total of 11 distinct conditions or diseases:



- 17 and 19 gestational weeks - PFC, OFC, PTC
- Aging - astrocytes

- Age effect on normal adult brain: frontal cortical regions
- Alzheimer's Disease (x2)
- Glioma
- HIV - associated neurocognitive disorders
- HIV - associated neurocognitive impairment: brain regions
- Multiple Sclerosis - brain lesions
- Parkinsons
- Pediatric glioblastoma
- Primitive neuroectodermal tumors
- Schizophrenia x2
- Schizophrenia - BA10, BA22
- Cardiomyopathy - (for reference)


A summary of accession numbers, aggregated dataset info, and study purpose information can be found at:



- [https://github.com/YasChenSon-Bioinformatics/BrainDiseaseCors/blob/master/GPL\\_data\\_formatted.csv](https://github.com/YasChenSon-Bioinformatics/BrainDiseaseCors/blob/master/GPL_data_formatted.csv)

## 2.2 Preliminary Methods

We spent a large portion of our initial analysis time going through the datasets to find those that would be most amenable to comparison to one another. One of the main criteria for this filtering was the choice of microarray platform. We chose the GPL570 Affymetrix Human Genome U133 Plus 2.0 Array, as it was both the most common, and most comprehensive array used across the datasets. As part of our preliminary analysis, we performed the following steps on our chosen datasets:

1. Subset the data to only include probes that exist in all datasets
2. Remove the control cases, as we are interested in samples from the disease state 
3. Remove any extra rows including one with metadata regarding the experimental conditions
4. Compute correlations within the samples from each dataset to give an indication of the correlation between its members



5. Average the gene expression values across all cases from each dataset, to obtain a single gene expression vector
6. Merge these vectors into a final data frame, and compute a correlation matrix for this final merged set

After implementing a pipeline to perform these analyses on our datasets, there were several relationships between the gene expression values that could be confirmed by known disease relationships.





## 2.3 Proposed Methods


We would like to improve upon the methods of correlation described previously, and implement some degree of normalization across the datasets that would prevent our results from showing up as spuriously correlated. This may be done by using the controls from the original datasets as a normalization background, or possibly by bringing in additional data from the Allen Atlas if necessary, although this would require some additional normalization. In addition to the correlation methods described previously

However, we are not restricting ourselves to the three simple methods of correlation already mentioned, and methods of clustering including BICLIC biclustering, using several R packages to get the unified ID structures,  and/or ICA, and also potentially a machine learning model  to predict disease based on gene expression values are all possible contenders for additional analysis.

## 2.4 Evaluation Strategy

Our evaluation strategy is based around research of the  current literature to confirm our findings on diseases and conditions with highly correlated expression values. In addition, in the event that we build a machine learning model to classify expression profiles as diseases, we will use standard k-fold  cross validation as provided by the sklearn package.

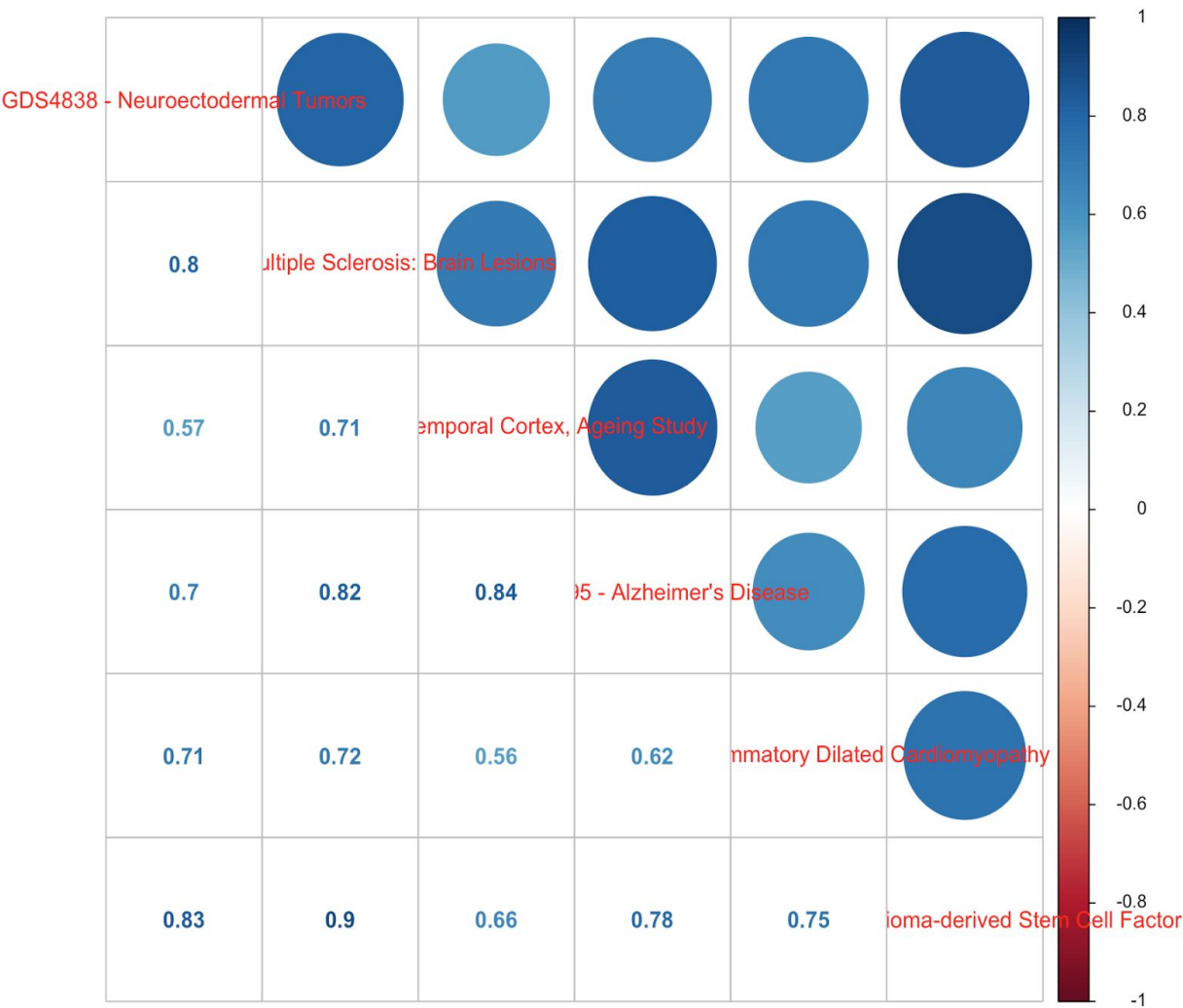
## 3. Preliminary Results

As part of our preliminary analysis, we found that several of the  average values from the cases in the experiments were correlated with one another. The highest correlation (0.90) was between

the Multiple Sclerosis lesioned tissue and Glioma damaged tissue. The next top four correlations were:

- 1. Alzheimer's Disease & Aging tissues (0.84)
- 2. Neuroectodermal Tumors & Glioma tissue (0.83)
- 3. Alzheimer’s Disease & Multiple Sclerosis Lesions (0.82)
- 4. Neuroectodermal Tumors & Brain Lesions (0.80)

These results can be viewed in the following correlation figure:




## 4. Discussion

### 4.1 Summary

The problem that we are trying to address is the lack of information regarding relationships among gene expression values between brain disease states. Our solution to this problem is to give quantitative measurements based on existing datasets that can best describe these relationships, while making adjustments for discrepancies between the experimental conditions of each of the results. It is a difficult task to try to account for each and every source of discrepancy, but we believe that factoring out the background by subtracting disease state from control will provide a certain level of improvement on the isolation of just the disease state. From our preliminary analysis we can see that there are some reasonable findings in our data. The highest correlation (0.9) between MS-lesioned tissue and Glioma-damaged tissue may make sense in light of the fact that both of these tissues were undergoing cell damage, and thus may have expressed similar factors to handle this stress. In addition, the association between aging and Alzheimer's is well-documented, and thus it makes sense that tissues from the disease states of these two cases would have some additional correlation. The subsequent three relationships are valid for similar reasons to the top relationship. The endomyocardial biopsy expression data was included for reference, and showed less correlation with any of the brain disease states than these states showed with each other, as expected.

While these results do show some initial findings, we will perform additional statistical tests to quantify their precise values, as well as put the data through additional normalization to account for possible underlying variables or correlates.

### 4.2 Anticipated Results

We anticipate that we will find additional relationships with more disease states included in our results. We also anticipate to find some statistically significant results after performing pairwise t-testing to determine precise p-values for the difference between the means of the expression values of the samples.  Our findings can confirm known relationships between the expression values of these additional diseases as confirmed by the literature, then we can be more confident

in the relationships that show up between diseases that had previously not been known to have similar expression profiles.

In the case that we observe statistically significant results that go against our expectations, we will have to revisit the initial steps of the data pipeline for those datasets, and ensure that our findings are not erroneous. This is entirely possible, given that we are drawing from a large number of datasets, and there may be underlying variables that have not been accounted for which are providing false relationships between the data. It will take careful inspection of the methods of the previous investigators to ensure that the datasets are indeed comparable.

### **4.3 Conclusion**

With this research, we hope to make some tangible progress towards the classification of the relationships between various disease states in the brain, and the quantification of their genetic correlations. In the process we will gain insight into the power of our statistical methods and the power of gene expression microarray data to suggest possible connections between diseases and their underlying mechanisms.

## **References**

Github Repo: <https://github.com/YasChenSon-Bioinformatics/BrainDiseaseCors/>

Relevant Papers:

1. Kim KY, Kim BJ, Yi GS. Reuse of imputed data in microarray analysis increases imputation efficiency. BMC Bioinformatics. 2004;5:160. doi: 10.1186/1471-2105-5-160. [PMC free article][PubMed] [Cross Ref]
2. Hwang T, Sun CH, Yun T, Yi GS. FiGS: a filter-based gene selection workbench for microarray data. BMC Bioinformatics. 2010;11:50. doi: 10.1186/1471-2105-11-50. [PMC free article][PubMed] [Cross Ref]
3. Yun T, Hwang T, Cha K, Yi GS. CLIC: clustering analysis of large microarray datasets with individual dimension-based clustering. Nucleic acids research. 2010;38(Web Server):W246–253. doi: 10.1093/nar/gkq516. [PMC free article] [PubMed] [Cross Ref]

4. Yun T, Yi GS. Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion. BMC Genomics. 2013;14:144. doi: 10.1186/1471-2164-14-144.[PMC free article] [PubMed] [Cross Ref]
5. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America. 2005;102(43):15545–15550. doi: 10.1073/pnas.0506580102. [PMC free article] [PubMed] [Cross Ref]
6. Sun CH, Hwang T, Oh K, Yi GS. DynaMod: dynamic functional modularity analysis. Nucleic acids research. 2010;38(Web Server):W103–108. doi: 10.1093/nar/gkq362. [PMC free article] [PubMed][Cross Ref]
7. <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-490>
8. [https://lvdmaaten.github.io/publications/papers/Methods\\_2014.pdf](https://lvdmaaten.github.io/publications/papers/Methods_2014.pdf)
9. <http://www.pnas.org/content/101/7/2173.full>
10. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3029380/>
11. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3387376/>
12. <https://www.ncbi.nlm.nih.gov/pubmed/11476888>
13. <https://www.ncbi.nlm.nih.gov/pubmed/21321390>
14. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3029380/>
15. <http://onlinelibrary.wiley.com/doi/10.1002/humu.23040/full>
16. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4460778/>
17. [http://genomics.princeton.edu/storeylab/papers/ETST\\_JASA\\_2001.pdf](http://genomics.princeton.edu/storeylab/papers/ETST_JASA_2001.pdf)
18. <https://www.ncbi.nlm.nih.gov/pubmed/26068849>

#### Resources:

1. <https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/JMiller/>
2. [http://human.brain-map.org/microarray/search/show?search\\_term=91&search\\_type=gene\\_classification](http://human.brain-map.org/microarray/search/show?search_term=91&search_type=gene_classification)



3. <http://www.scientificamerican.com/article/massive-u-k-brain-mapping-project-releases-first-results/>
4. <https://www.ncbi.nlm.nih.gov/pubmed/12454644> (normalizing expression datasets)
5. <http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf> (distance calculations p235)
6. <https://www.biostars.org/p/1216/>

