Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Investigation of Differential Gene Expression in Major Depressive Disorder Using Information Theory

Nate Richman [1,*]

[1] Biomedical Engineering, University of Wisconsin-Madison, Madison, 53715, USA

* To whom correspondence should be addressed.

Associate Editor:

Received on 5/11/20

Abstract

Motivation: Major Depressive Disorder (MDD) affects millions of people yearly, yet it is still poorly understood. Some genes that are associated with MDD have been presented in the literature. I hypothesized that gene expression in MDD could be investigated using tools from information theory, such as mutual information and the k-wise interaction information.
Results: I used two Gene Expression Ombnibus (GEO) datasets of post-mortem brain expression profiling Results: I used two Gene Expression Ombnibus (GEO) datasets of post-mortem brain expression profiling in control and mdd-diagnosed indivuduals to investigate the relationship between gene expression and diagnosis, as well as the relationship between gene expression, age and diagnosis. I found 30 genes that are associated with a mdd diagnosis, but no significant genes that are associated with diagnosis and age. I compared these genes with known genes from brainspan using clustering, gene set enrichment analysis and gene ontology. I found that the genes I identified aren't significantly associated with expected cellular functions but they separate the samples better.
Availability: Source available at https://github.com/naterich2/776-project
Contact: nrichman@wisc.edu
Supplementary information: Supplementary data available by contacting author

1 Introduction

Major Depressive Disorder (MDD) is a mental health disorder that cuases a continued feeling of helplessness, sadnes, and loss of interest in doing daily activities (Mayo Clinic., 2020). This disorder affects millions of people nationwide and has its causes as both environmental and genetic (Brainspan., 2020). It is known that many genes affect MDD, and recent studies have undertaken GWAS studies to determine some of these genes (Cai *et al*., 2020; Malhotra., 2020). Additionally, the Brainspan atlas of the developing human brain has identified several genes that are associated with MDD through RNA arrays (Brainspan., 2020). While some of these genes have been identified, GWAS is limited by ungenotyped causal SNPs and RNA arrays limit discovery of new genes or pseudogenes that could be related to MDD. Using a higher throughput and more thorough approach such as RNA seq would be better for identifying genes, but it is not a trivial task to identify causal genes from this data.

Mutual information (MI) is a measure from information theory of the amount of information gained about one random variable from observing the other, and is calculated as follows:

$$MI(X,Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log_2\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

or

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \quad (2)$$

Where H(X) is entropy defined as: $H(X) = \sum_{x\in X} p(x)\log_2 p(x)$ (3)

Mutual information has been used previously to extract information on regulatory modules (Elemento *et al*., 2007). I hypothesized that mutual information could be used to investigate whether certain genes expression levels predict a diagnosis of MDD. Additionally, I hypothesized that gene expression in these genes may be a function of time as well as diagnosis since MDD tends to become apparent in early adulthood (Mayo Clinic., 2020).

While mutual information is only defined to quantify the relationship between two variables, there have been a number of extensions to multiple variables. The first being Total Correlational Information (TCI) (Timme *et al*., 2012; Watkinson *et al*., 2009):

$$TCI(X, Y, Z) = H(X) + H(Y) + H(Z) - H(X, Y, Z) \quad (4)$$

And the second being k-Wise Interaction Information (KWII) (Timme *et al*., 2012):

$$\begin{aligned} KWII(X, Y, Z) = & -H(X) - H(Y) - H(Z) + \\ & H(X, Y) + H(X, Z) + H(Y, Z) \\ & - H(X, Y, Z) \end{aligned} \quad (5)$$

Both of these methods try to gain understanding about the interaction between three random variables, in this case, gene expression, subject age, and subject diagnosis. A positive KWII indicates synergy between the variabes, whereas a negative KWII indicates redundancy between the variables.

I investigated gene expression in relation to age and diagnosis via both methods and found that KWII was a better measure for this relationship, but mutual information between diagnosis and gene expression was much stronger than both multivariate measures. I compared the genes I identified to known gene in brainspan and saw differing results using various metrics.

2 Approach and Methods

2.1 Data

I used the Brainspan database and downloaded RPKM values of control subjects for genes associated with depression (brainspan.org/rnaseq/ (Brainspan., 2020). Additionally, I downloaded two Gene Expression Omnibus datasets (GSE101521 and GSE80655) and normalized expression data using the DeSeq2 library in R (Love *et al*., 2014). Finally, I generated metadata for the datasets.

2.2 Gene significance by MI, TCI, and KWII

I implemented TCI, KWII and MI by using the definition of each from the entropy of marginal or join distributions in python using numpy and pandas. For each gene I calculated either it's MI with diagnosis or age, its TCI with age and diagnosis, or its KWII with age and diagnosis. Then I shuffled the diagnosis labels for 500 permutations and repeated the calculation for each gene. This gave me a background distribution for each gene. I used determined p-values using z-score against a normal distibution and determined significance using an FDR of 0.05 with the Benjamini-Hochberg correction.

2.3 Validation: diffusionmap, GSEA, and GO

Diffusionmap was run in R using the package Destiny from Bioconductor. Significant genes were exported from the python script in csv format, diffusionmap was then run and the eigenvectors were exported and plotted using matplotlib. GSEA was run using The Broad Institute GSEA command line software (Subramanian *et al*., 2005). Gene sets were created using the genes determined to be significant using the KWII and MI between diagnosis and expression as well as known genes from Brainspan (Brainspan., 2020). Similarly, GO was run using the PANTHER online database (The Gene Ontology Consortium., 2019)

Table 1. GSEA significance values for Brainspan and MI determined gene sets. The KWII determined gene set (even with nom. $p < 0.01$) was too small as the minimum gene set size is 15 genes, and the KWII gene set was 5 genes

Geneset	Size	ES	NES	Nom. p-val	FDR q-val	FWER p-val
Brainspan	45	-0.23	-0.89	0.549	1.000	0.498
MI	30	-0.26	-0.76	0.819	0.773	0.617

3 Results and Discussion

3.1 KWII vs. TCI vs. MI

After running one iteration of KWII vs. TCI, it was clear that KWII separated the genes much more clearly than TCI, so KWII was used for the rest of the project. After running 500 iterations and determining the background distribution for each gene, KWII did not elucidate any significant genes at an FDR of 0.05, and only suggested 5 genes at a nominal p-value < 0.01. MI between diagnosis and expression level performed much better and found 30 genes to be significant at an FDR of 0.05.

Because KWII did not yield any significant genes, it was not used for further analysis. Potential issues explaining why KWII might not have yielded any significant genes are discussed later.

3.2 Comparison of Brainspan MDD genes vs. those suggested by MI

3.2.1 DiffusionMap Clustering

Figure 1 Figure 2 Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

3.3 Test1

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

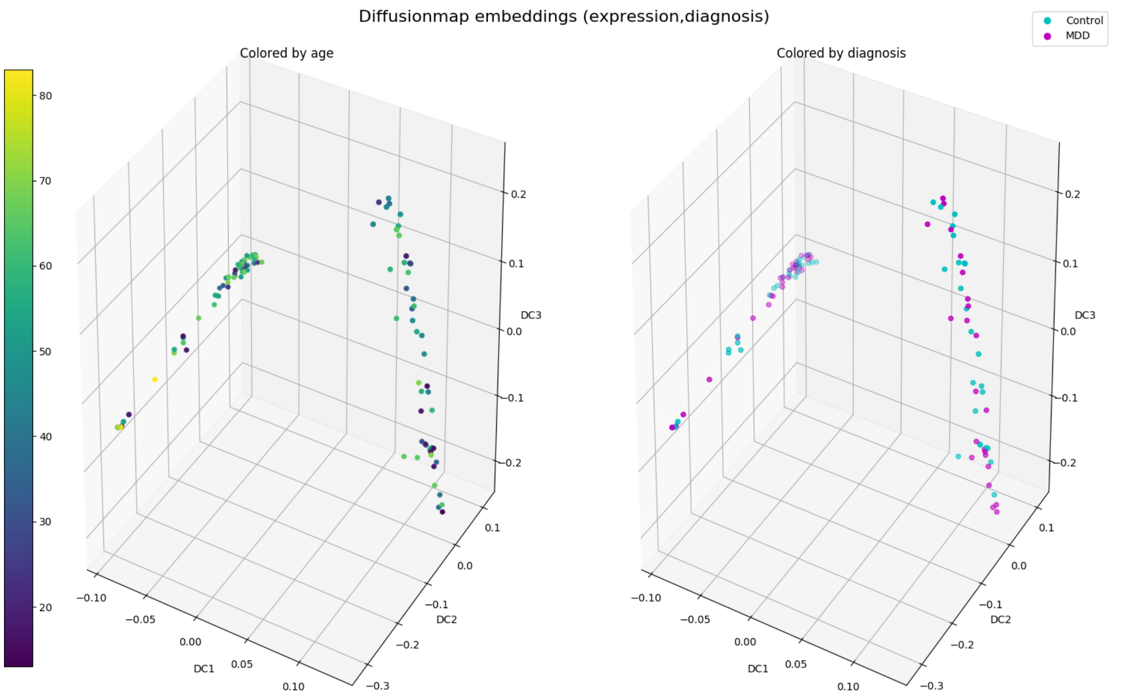


Fig. 1: DiffusionMap clustering on genes identified by MI with diagnosis and expression level. (Left) Points are plotted on top 3 embedding components; points are colored by the age of the diagnosis. (Right) Points are plotted by patient diagnosis

4 Discussion

Figure 1 Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text.

Table 1 shows that Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text. Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text.

5 Conclusion

(Table 1) Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

1. this is item, use enumerate
2. this is item, use enumerate
3. this is item, use enumerate

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text Text Text Text Text Text Text Text Text. **?** might want to know about text text text text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text.

Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text Text. Figure 2 shows that the above method Text Text Text Text

Acknowledgements

Text Text Text Text Text Text Text Text. **?** might want to know about text text text text

Funding

This work has been supported by the... Text Text Text Text.

References

Ashburner et al., “Gene ontology: tool for the unifiation of biology,” *Nat Genet.*, vol. 25, no. 1, 2000.

Brainspan, “Developmental Transcriptome,” Allen Institute for Brain Science, 2020.

N. Cai et al., “Minimal phenotyping yields GWAS hits of reduced specificity for major depression,” *bioRxiv*, [preprint], 2020.

O. Elemento, N. Slonim and S. Tavazoie., "A universal framework for regulatory element discovery across all genomes and data types," *Molecular Cell*, vol. 28. no. 2, p. 337-350, 2007.

The Gene Ontology Consortium, “The Gene Ontology Resource: 20 years and still Going strong,” *Nucleic Acids Res.*, vol. 47, no. D1, p. D330-D338, 2019.

Krebs et al., “Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect,” *Psychological Medicine*, pg. 1-12, 2019.

M.I. Love, W. Huber, S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” *Genome Biology*, vol. 15, pg. 550, 2014.

A. K. Malhotra, “The pharmacogenetics of depression: Enter the GWAS,” *Am. J. Psychiatry*, vol. 167, no. 5, pp. 493–495, 2010.

Mayo Clinic, “Depression (major depressive disorder),” Mayo Foundation for Medical Education and Research (MFMER), 2020.

T. E. Oliphant, “Python for Scientific Computing”, Comp. Sci. Eng., vol. 9, pg. 10-20, 2007.

Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *JMLR*, p. 2825-2830, 2011.

T. H. Pham, T. B. Ho, Q. D. Nguyen, D. H. Tran, and V. H. Nguyen, “Multivariate mutual information measures for discovering biological networks,” *IEEE Conf. Res. Innov. Vis. Futur.*, 2012.

Ramaker et al. “Post-mortem molecular profiling of three psychiatric disorders,” *Genome Medicine*, vol. 9, no. 72, 2017.

A. Subramanian, et al., “Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Nat. Acad. Sci.*, vol. 102, pg. 15545-15550, 2005.

N. Timme, W. Alford, B. Flecker, J. M. Beggs, "Multivariate information measures: an experimentalist’s perspective," *arXiv*, 2012.

K. C, Vadodaria et al., “Seratonin-induced hyperactivity in SSRI-resistant major depressive disorder patient-derived neurons,” *Molecular Psychiatry*, vol. 24, no. 6, pg. 795-807, 2019.

J. Watkinson, K. Liang, X. Wang, T. Zheng, D. Anastassiou, "Inference of Regulatory Gene Interactions from Expression Data Using Three-Way Mutual Information," *Ann. N. Y. Acad. Sci.*, vol 1158, p. 302-313, 2009

Bag,M., Name2, Name3 (2001) Article title, *Journal Name*, **99**, 33-54.

Yoo,M.S. *et al*. (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson’s disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Lehmann,E.L. (1986) Chapter title. *Book Title*. Vol. 1, 2nd edn. Springer-Verlag, New York.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer management: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al*. (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d’obesite infantile avec polydactylie et retinite pigmentaire (contribution a l’etude des formes cliniques de l’obesite hypophysaire). PhD Thesis, name of institution, Paris, France.

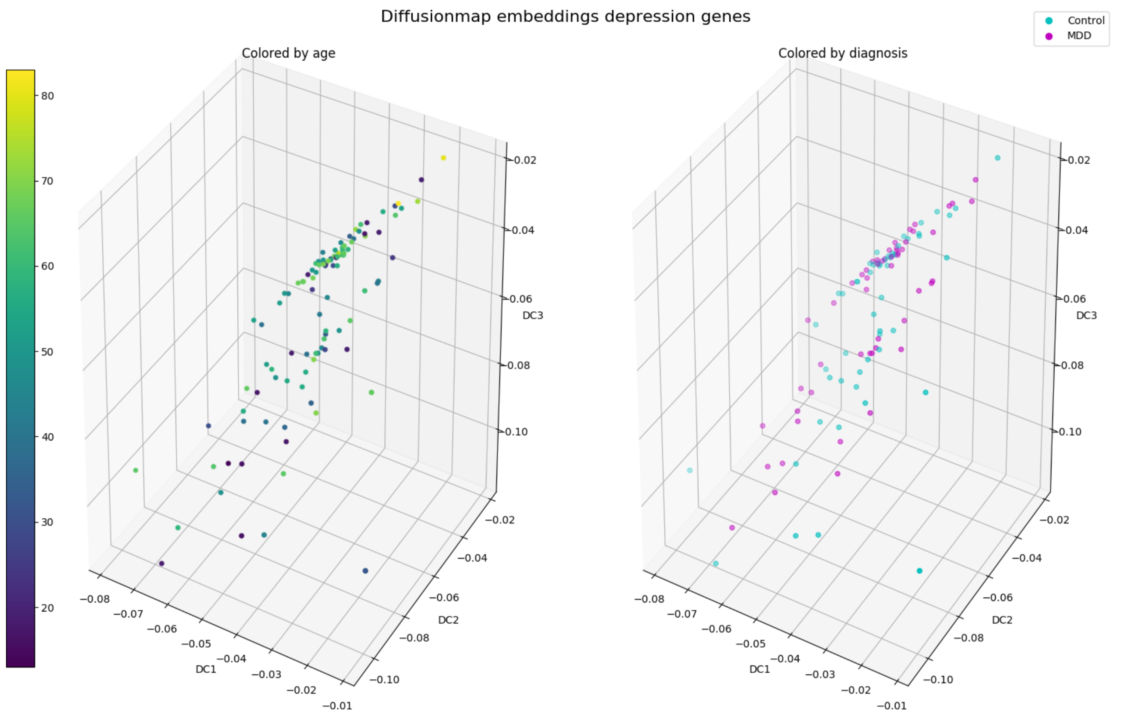
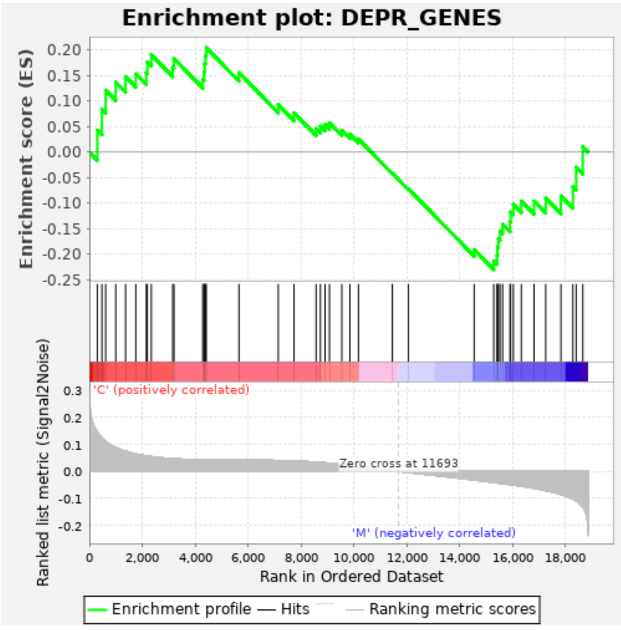
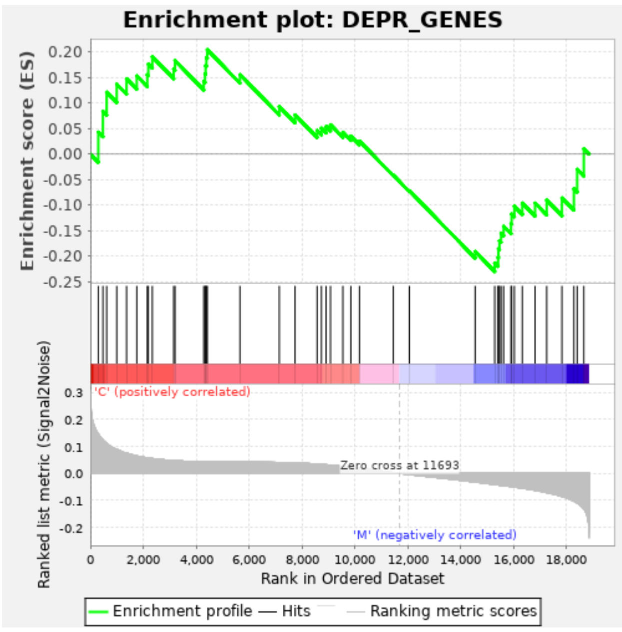


Fig. 2: DiffusionMap clustering on genes associated with depression according to Brainspan. (Left) Points are plotted on top 3 embedding components; points are colored by the age of the diagnosis. (Right) Points are plotted by patient diagnosis

tph



(a) Title A



(b) Title B

Fig. 3: Title for both