Differential gene expression analysis in mental disorders

Network based data analysis course report

Rayan Slatni

June 2023

Abstract

We performed differential expression gene analysis using supervised and unsupervised statistical learning methods in order to identify the changes in gene level expression among several mental disease. We used a dataset containing oligonucleotide microarrays from the human prefrontal cortex (BA10) of 50 samples including patients with depression, schizophrenia, bipolar disorder and control. We found three genes that are opposite regulated among the groups: TMP4, DGKA and PDHB.

Introduction

Mental disease is a field whose molecular mechanisms are not yet fully understood, some of them present overlapping symptoms and the phenotype definition is sometimes uncertain. Therefore, investigating the molecular mechanisms deeper is fundamental to improving diagnosis, predisposition, prognosis and therapy. Considering that we are talking about multifactorial and multigenic diseases, with environmental risk and complex mode of inheritance, it is necessary use quantitative strategies that look what happens at a molecular and gene level. For these reasons we performed gene expression analysis using DNA microarray of thousands of genes from the BA-10 located in the prefrontal cortex. The Brodmann Area is involved in the major cognitive functions, emotion processing, behaviour regulation and others, hence it is a crucial area to explore for brain dysfunctions. Among all the mental diseases, bipolar disorder is one of the most complex since it shares signs and symptoms of depressive episodes with major depression, and also shares clinical features with schizophrenia during the mania phase, such as their chronic and relapsing course, and psychotic symptoms[40], so looking at the different gene expression level could be a good starting point for discriminating these pathological conditions. With this report we aim to investigate the uniqueness of bipolar disorder (BD) and its similarity to major depression (MD) and schizophrenia (SD) at the molecular level, using statistical methods of classification, functional enrichment analysis and network-based analysis; this could lead to a thoroughgoing knowledge of the genes and molecular pathways involved in mental disease, plus potentially identifying new genes that could be used as bio-markers.

Methods

The dataset was retrieved from the Gene Expression Omnibus with GEO accession GSE12654[1], containing oligonucleotide microarray according to the platform Affymetrix Human Genome U95 Version 2 Array. All the human samples come from post-mortem prefrontal cortex BA-10 (Brodmann's Area 10) and were donated by the Stanley Foundation Brain Collection. Further information about brain sampling and the microarray procedure can found here [37]. The 50 samples include patients affected by depression (n=11), schizophrenia (n=13), bipolar disorder (n=11) and control group (n=15). Most of the analysis was performed using RStudio [14](full session details can be found in the supplementary materials, SuppFig1). Before each step entailing randomness, *set.seed* function was used in order to obtain reproducible results (seed detail are reported in the code). Extracting data matrix the initial number of genes is 12625 all represented

among the 50 samples (no NA value present). The only pre-processing required was a normalization in log2 scale. The following unsupervised methods were performed: principal component analysis, k-means clustering on log2 transformed data to reduce outlier influence and hierarchical clustering (all from base stats library [13]). Before performing k-means and hierarchical clustering we used the Elbow Method for determining the optimal number of clusters. The method is based on calculating the Within-Cluster-Sum of Squared Errors (WSS) for different number of clusters (k from 1 to 10) and selecting the k for which change in WSS first starts to diminish[34]. From the plot (SuppFig5) the optimal number of cluster is k=4. In the hierarchical clustering we set Euclidean distance and method complete. A first feature selection was performed in order to remove the genes that have a low expression (smaller than 1 on log-transformed data) in at least 40% of the samples, ending up with 4682 genes (genefilter library [15]). A second normalization was applied according to the expression of a housekeeping gene, GAPDH (glyceraldehyde-3-phosphate dehydrogenase)[28] from the control group, this step is important to correct the sample to sample variations. A second feature selection was done by performing a *T-test row by row* pairwise each group filtering for p-value < 0.01 (the p.adjusted was too restricted), in this way we obtained 117 significant genes. At the same time, we calculated the *fold-change* (gtools library [16]) for the significant genes among each group pairs (some samples had being removed since not equally rappresented across the groups). Among the supervised methods we performed random forest (randomForest library[38]), both with and without the second feature selection. At this point we split the dataset into training and validation (*createDataPartition* function with 75% of the samples in the training set using caret library [12]). Then linear discriminant analysis was done using repeated cross-validation (10 fold, repeated 10 times) with metric = "Accuracy" (MASS library [39]), the same parameters were used to fit the random forest again. Lasso regression (glmnet library[11]) was performed using the same set of parameters and setting family = "multinomial" since n=4 groups, a good value for the shrinkage parameter is *lambda* = 0.2 (SuppFig10). The Signature-based Clustering for Diagnostic Purposes (SCUDO on rSCUDO library [25]) was done using data partition with 75% of the samples in the training set, then the model was trained with scudoTrain, scudoTest was used for validation, and the performance with scudoClassify (parameters nTop=25, nBottom=25, N=0.5 where required for returning a connected graph). Using the tool Cytoscape[17] we visualized the resulting network (SuppFig11). The functional enrichment analysis was performed using g:Profiler tool[3] using Bonferroni correction and setting the p-value at 0.05 on the gene list obtained from LDA. Still using g:Profiler we converted the ID probe (Affymetrix Human Genome U95 Version 2 Array) into Ensembl annotation. The online tool DAVID[2] was used to control the results from g:Profiler. In the end, Network based Analysis was performed among the group pairs using the function *pathfinder* (pathfindR library [26], setting iterations = 5) in order to find the genes that are up or down regulated between bipolar respect to the others groups. The input dataframe for pathfider include: gene name, p-value (retrieved from the T-test) and foldchange. We used the online access to the tools String[8] in order to look for other gene networks.

Results

The PCA wasn't performing well in separating the groups (SuppFig4), in fact the first 3 dimensions explain over 18.37% of the variance where only the 9.39% was due PC1. Also, the k-means clustering didn't return good results, in fact most of the groups were misclassified (SuppFig6) (0.28 accuracy). Additionally, the hierarchical clustering didn't separate the groups well, leading to the construction of several single-samples branches (SuppFig7) (0.34 accuracy). From the random forest we retrieved that only the first 100 genes seem to be relevant (Supp9Fig), but the model is not performing well, in fact the OOB error rate is 66%(SuppFig8). Performing LDA we obtain a good separation between the groups (Figure 1) (0.76 accuracy). Running again RF (using the data after the second feature selection) the accuracy is 0.68. Using lasso regression, we get a classifier with decent performance (0.67 accuracy). We finally performed SCUDO analysis, which yields to very bad results (0.5 accuracy). LDA is the model that performs the best group separation and has the highest accuracy Table 1, but considering that after the second feature selection

we ended up with only 117 genes, we decided to keep all of them for the downstream analysis.

Methods	Accuracy
K-means	0.28
H. Clustering	0.34
RF	0.68
LDA	0.76
Lasso	0.67
Scudo	0.50

Tabella 1

The functional enrichments analysis returns 8 main terms, the most significant indicate general function about cytoplasm (padj 3.014×10-6) and protein binding (padj 1.724×10-4), the only specific function for nervous system is for axon (padj 7.678×10-3) Figure 2. Among the genes dysregulated in the axon we have found: FKBP4, know to have significant change in depression[19] and is downregulated in our data (SuppTab1), LMTK2, important for nerve growth[8], downregulated in SD[21] (SuppTab1), NFIB already associated with BD and SD[36] and intellectual disability [29], MAG is downregulated in SD [9] (SuppTab1) and from STRING network is linked to SOX10 (SuppFig14), a transcription factor that activates expression of myelin genes[8] and tended to be highly methylated in brains of patients with SD[5] (SuppTab1). RUFY3 (FSCN1) is decreased in MD[7] (SuppTab1), MAP3K12 a gene whose knockdown increases neuronal survival[18] interestingly is downregulated only in bipolar and not in MD or SD (SuppTab1), and PLXND1 involved in neurodevelopmental processes[30] decrease in BD (SuppTab1). Performing network enrichment analysis we found several genes upregulated in bipolar respect to the control group: NCAM1, important for the development of the central nervous system and synaptic plasticity[22], related to prion disease pathway (Figure 3), but the level of NCAM1 in BD are low (SuppTab1 *), STIP1, related also to MD[20], and linked to FKBP4[8] (SuppFig13), but in our results the level of STIP1 in BD are low (SuppTab1 *), ARHGEF1, participate to proliferation, differentiation[4], but the level of ARHGEF1 in BD are low (SuppTab1 *). The genes downregulated are: TFE3 where low levels of these genes are related to autophagy dysfunction and neurodegeneration of dopaminergic neurons in Parkinson's Disease[10], RPS6KA1, a Serine/threonine-protein kinase that repress pro-apoptotic function of DAPK1[8], this gene is upregulated only in the MD group and is interesting because it has being shown that knockdown of DAPK1 reduces depression like behaviour[6] and SLC29A1[27]. Comparing the bipolar group against the schizophrenia we found two genes that are oppositely regulated in the two diseases and are involved in the cardiac contraction and uniquitin mediated proteolysis(Figure 4): NEDD4L, an ubiquitin ligase that promotes the degradation of NRG1 a gene which decrease level has been related to depression-like behaviours[23], but in our results NEDD4L level are low (SuppTab1 *). TPM4, a cytoskeleton constituents that have shown broad differential expressions in schizophrenia[33], is down regulated in bipolar and up regulated in schizophrenia (SuppTab1). Comparing the bipolar group against depression we found two genes that are opposite regulated in the two disease and are involved in pyruvate metabolism and citrate cycle (Figure 5): PDHB is increased in BD[24] and decreased both in SD[32] and MD (SuppTab1), DGKA seems to play a role in regulating depression[35] and in our data is dowregulated in BD and upregulated both in MD and SD. Among the 117 significant genes also CACAN1C is associated to bipolar disorder[31], but only two of them are reported in the article that used the same dataset (GSE12654) [37]: SLC29A1 and ARHGEF2.

Conclusion

About the statistical methods, we can say that overall the supervised methods perform better than the unsupervised, in particular the LDA is the one with the highest accuracy. However, not all methods are well performing, this was expected considering the molecular complexity of mental disorder. Through this analysis, we identify 21 genes up and down regulated in bipolar disorder, schizophrenia and depression. Some of them are already known to play a role in mental disorder, but for others the function is still undefined or not confirmed. In particular, we have found some genes that are oppositely regulated among the three disease (TMP4, DGKA and PDHB); these are particularly interesting because they could be used to distinguish one disease from another at a genotype level. Further exploration about the molecular mechanism behind these genes should be done before considering their application as possible biomarkers.

References

[1] *GEO*. URL: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE12654. (accessed: 20.03.2023).

[2] *GEO*. URL: https://david.ncifcfr.gov. (accessed: 07.06.2023).

[3] *gprofiler*. URL: https://biit.cs.ut.ee/gprofiler/gost. (accessed: 04.06.2023).

[4] et. al Hervé M. "Translational Identification of Transcriptional Signatures of Major Depression and Antidepressant Response". In: (). DOI: 10.3389/fnmol.2017.00248.

[5] et. al Iwamoto K. "DNA methylation status of SOX10 correlates with its downregulation and oligodendrocyte dysfunction in schizophrenia". In: (). DOI: 10.1523/JNEUROSCI.0766-05.2005.

[6] Xu LZ Li SX Han Y. "Uncoupling DAPK1 from NMDA receptor GluN2B subunit exerts rapid antidepressant-like effects". In: (). DOI: 10.1038/mp.2017.85.

[7] et. al Notaras M. "The proteomic architecture of schizophrenia iPSC-derived cerebral organoids reveals alterations in GWAS and neuronal development factors". In: (). DOI: 10.1038/s41398-021-01664-5.

[8] *String*. URL: https://string-db.org/. (accessed: 04.06.2023).

[9] et. al Tkachev Bsc a c. "Oligodendrocyte dysfunction in schizophrenia and bipolar disorder". In: (). DOI: 10.1016/S0140-6736(03)14289-4.

[10] et. al Zheng Q. "TFE3-Mediated Autophagy is Involved in Dopaminergic Neurodegeneration in Parkinson's Disease". In: (). DOI: doi:10.3389/fcell.2021.761773.

[11] Trevor Hastie. "An introduction to glmnet". In: (2023).

[12] Max Kuhn. "Classification and Regression Training". In: (2023).

[13] "R Core Team. R: A Language and Environment for Statistical Computing". In: (2023).

[14] "RStudio Team. RStudio: Integrated Development Environment for R". In: (2023).

[15] Robert Gentleman et al. "genefilter: methods for filtering genes from high-throughput experiments". In: (2022).

[16] Ben Bolker. "Various R Programming Tools". In: (2022).

[17] G.-M.; Chen Ge B.-K.; Hu. "MSClustering: A Cytoscape Tool for Multi-Level Clustering of Biological Networks". In: (2022). DOI: 10.3390/ijms232214240.

[18] et. al Kurishev AO. "CRISPR/Cas-Based Approaches to Study Schizophrenia and Other Neurodevelopmental Disorders". In: (2022). DOI: 10.3390/ijms24010241.

[19] et. al Dawid Szczepankiewicz. "Genes involved in glucocorticoid receptor signalling affect susceptibility to mood disorders". In: (2021). DOI: 10.1080/15622975.2020.1766109.

[20] et. al Dawid Szczepankiewicz. "Genes involved in glucocorticoid receptor signalling affect susceptibility to mood disorders". In: (2021). DOI: 0.1080/15622975.2020.1766109.

[21] et. al Ditsiou A. "The multifaceted role of lemur tyrosine kinase 3 in health and disease". In: (2021). DOI: 10.1098/rsob.210218.

[22] et. al Blessed Raj Jesudas. "Relationship of elevated neural cell adhesion molecule 1 with interleukin-10 and disease severity in bipolar disorder". In: (2020). DOI: 10.1016/j.ajp.2019.101849.

[23] et. al Xu Jia Guo Cuiping. "Nedd4l downregulation of NRG1 in the mPFC induces depression-like behaviour in CSDS mice". In: (2020). DOI: DOI:10.1038/s41398-020-00935-x.

[24] et. al Harry Campbell. "A pyruvate dehydrogenase complex disorder hypothesis for bipolar disorder". In: (2019). DOI: 10.1016/j.mehy.2019.109263.

[25] Thomas Cantore Matteo Ciciani and Mario Lauria. "rScudo: an R package for classification of molecular profiles using rank-based signatures". In: *Bioinformatics* (2019).

[26] Sezerman OU Ulgen E Ozisik O. "An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks". In: (2019). DOI: 10.3389/fgene.2019.00858.

[27] et. al Catharine E. Krebs. "Whole blood transcriptome analysis in bipolar disorder reveals strong lithium effect". In: (2018). DOI: 10.1017/S0033291719002745.

[28] Keertan Dheda. "Validation of housekeeping genes for normalizing RNA expression in real-time PCA". In: (2018). DOI: 10.2144/04371RR03.

[29] et. al Schanze I. "NFIB Haploinsufficiency Is Associated with Intellectual Disability and Macrocephaly". In: (2018). DOI: 10.1016/j.ajhg.2018.10.006.

[30] et. al Suhas Ganesh MD. "Exome sequencing in families with severe mental illness identifies novel and rare variants in genes implicated in Mendelian neuropsychiatric syndromes". In: (2018). DOI: 10.1111/pcn.12788.

[31] et. al Xie Z. Yang X. "A Genome-Wide Association Study and Complex Network Identify Four Core Hub Genes in Bipolar Disorder". In: (2017). DOI: 10.3390/ijms18122763.

[32] et. al Dean B. Thomas N. "Evidence for impaired glucose metabolism in the striatum, obtained postmortem, from some subjects with schizophrenia". In: (2016). DOI: 10.1038/tp.2016.226.

[33] et. al Nascimento JM. "The proteome of schizophrenia". In: (2015). DOI: 10.1038/npjschz.2014.3.

[34] Purnima Bholowalia and Arvind Kumar. "EBK-means: A clustering technique based on elbow method and k-means in WSN". In: *International Journal of Computer Applications* (2014).

[35] et. al Redei E. Andrus B. "Blood transcriptomic biomarkers in adult primary care patients with major depressive disorder undergoing cognitive behavioral therapy". In: (2014). DOI: 10.1038/tp.2014.66.

[36] et. al H. Le-Niculescu. "Convergent functional genomics of genome-wide association data for bipolar disorder: Comprehensive identification of candidate genes, pathways and mechanisms". In: (2008). DOI: 10.1002/ajmg.b.30887.

[37] et. al Iwamoto K Kakiuchi C. "Molecular characterization of bipolar disorder by comparing gene expression profiles of postmortem brains of major mental disorders". In: *Mol Psychiatry* (2004). DOI: doi:10.1038/sj.mp.4001437.

[38] Andy Liaw and Matthew Wiener. "Classification and Regression by randomForest". In: *R news* (2003).

[39] W. N. Venables and B. D. Ripley. "Modern Applied Statistics with S. Springer". In: (2002).

[40] et. al Berrettini WH. "Are schizophrenic and bipolar disorders related? A review of family and molecular studies". In: *Biol Psychiatry* (2000). DOI: 10.1016/s0006-3223(00)00883-0.

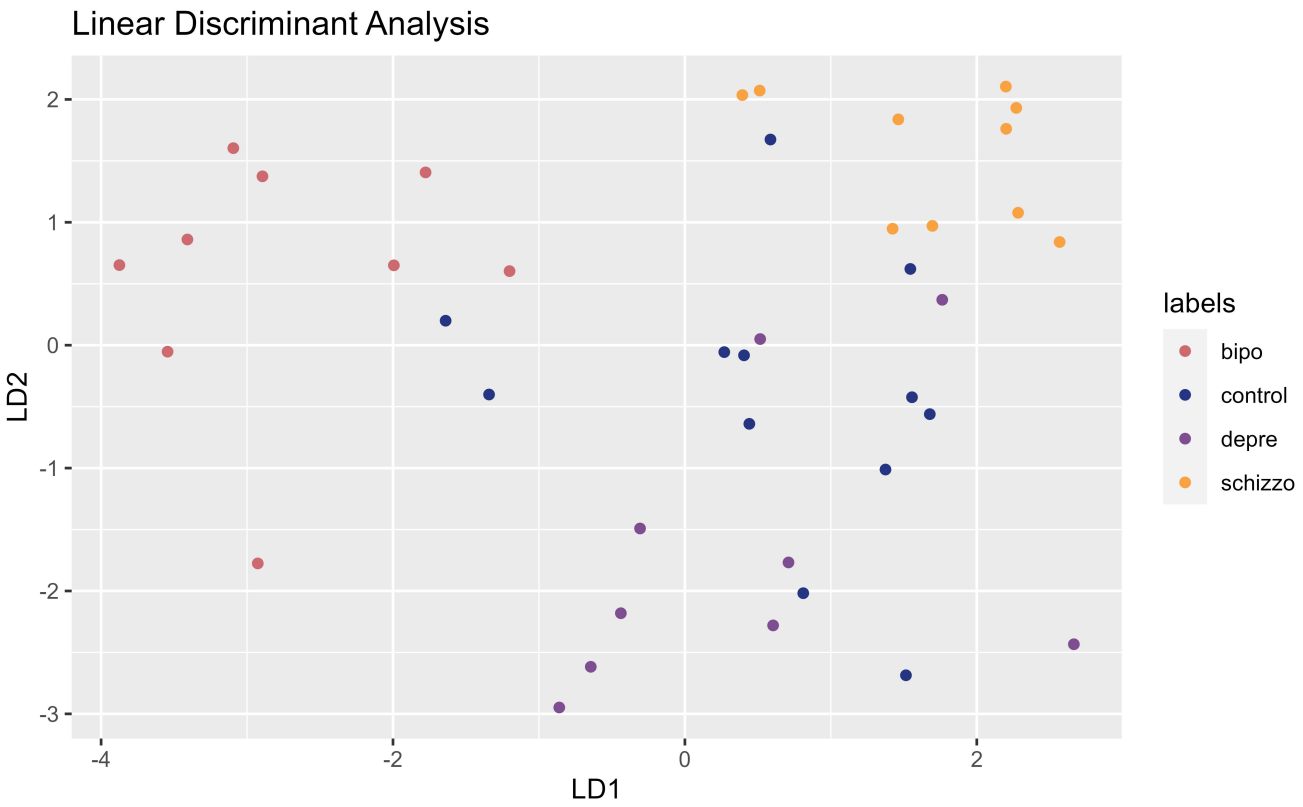
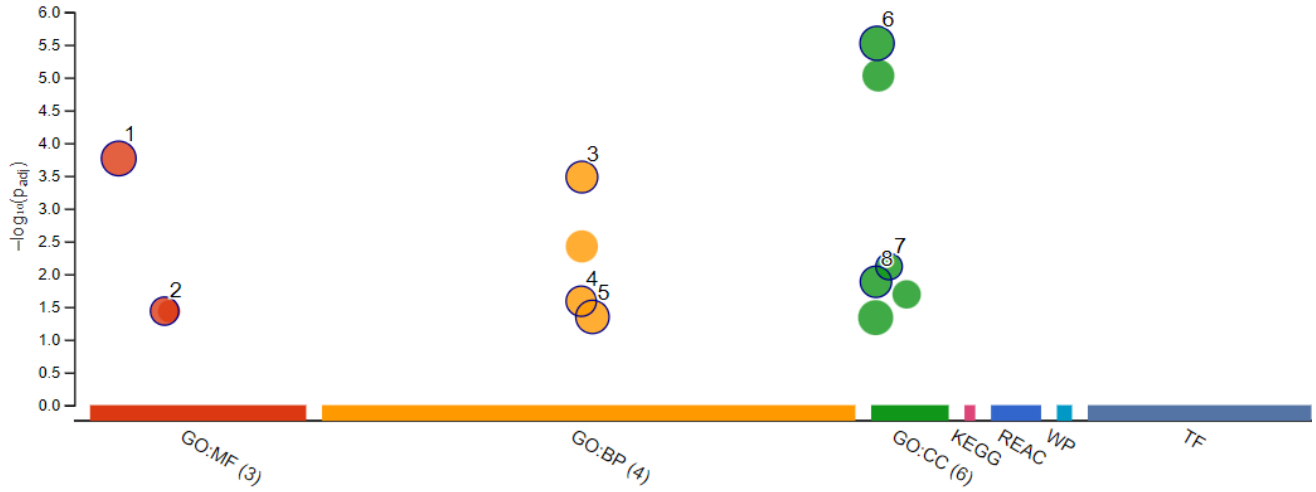


Figure 1: *Linear Discriminant Analysis of the two main components*



ID	Source	Term ID	Term Name	p_{adj} (query_1)
6	GO:CC	GO:0005737	cytoplasm	3.014×10^{-6}
1	GO:MF	GO:0005515	protein binding	1.724×10^{-4}
3	GO:BP	GO:0048523	negative regulation of cellular process	3.308×10^{-4}
7	GO:CC	GO:0030424	axon	7.678×10^{-3}
8	GO:CC	GO:0005654	nucleoplasm	1.313×10^{-2}
4	GO:BP	GO:0048468	cell development	2.599×10^{-2}
2	GO:MF	GO:0030234	enzyme regulator activity	3.677×10^{-2}
5	GO:BP	GO:0050896	response to stimulus	4.507×10^{-2}

Figure 2: *Gprofiler on the 117 significant genes*

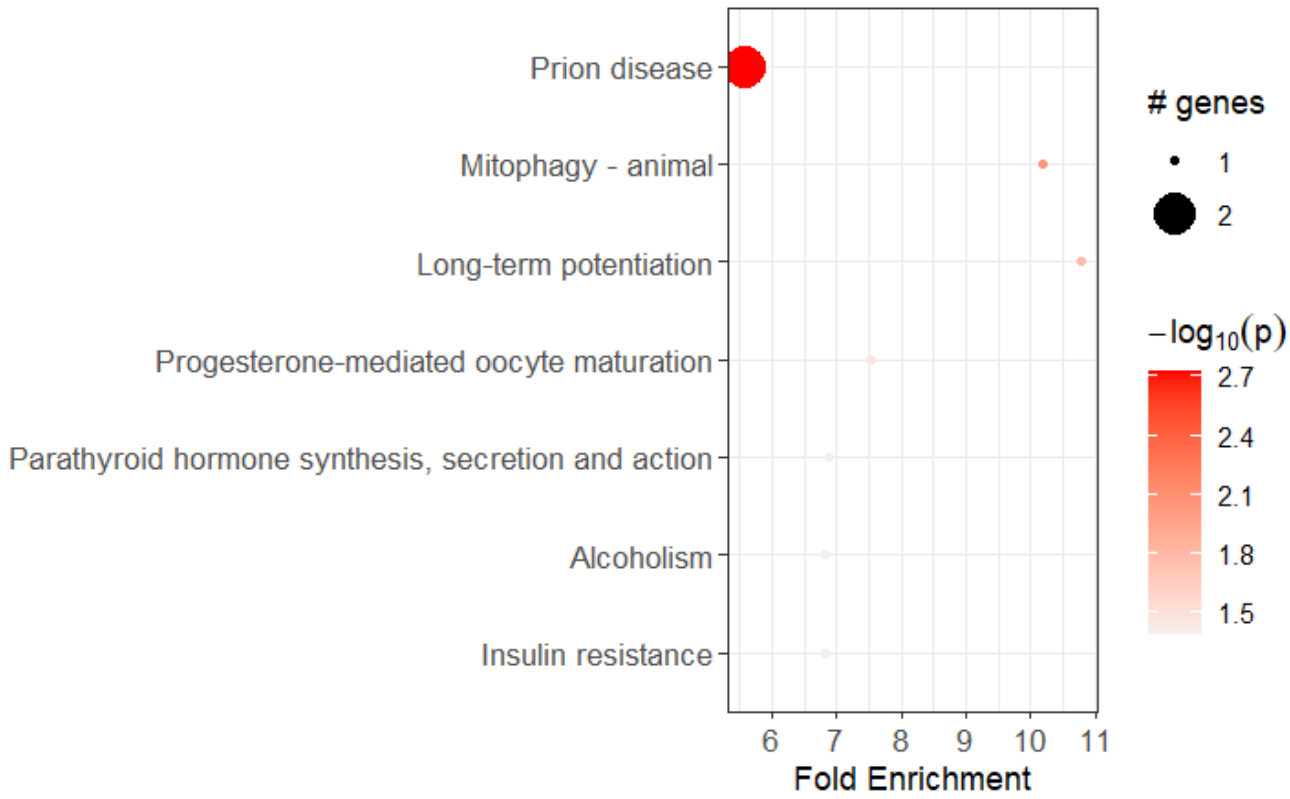


Figure 3: *Pathfinder plot: bipolar vs control*

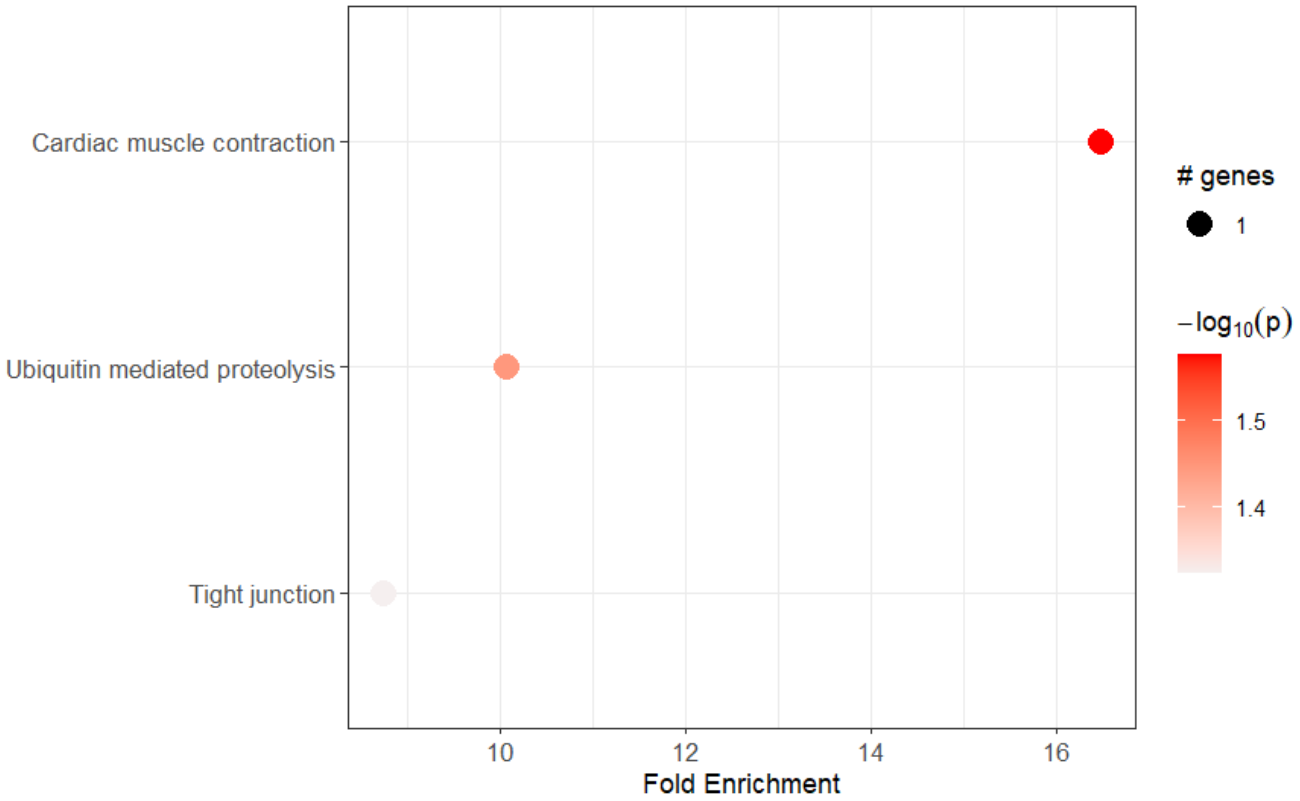


Figure 4: *Pathfinder plot: bipolar vs schizophrenia*

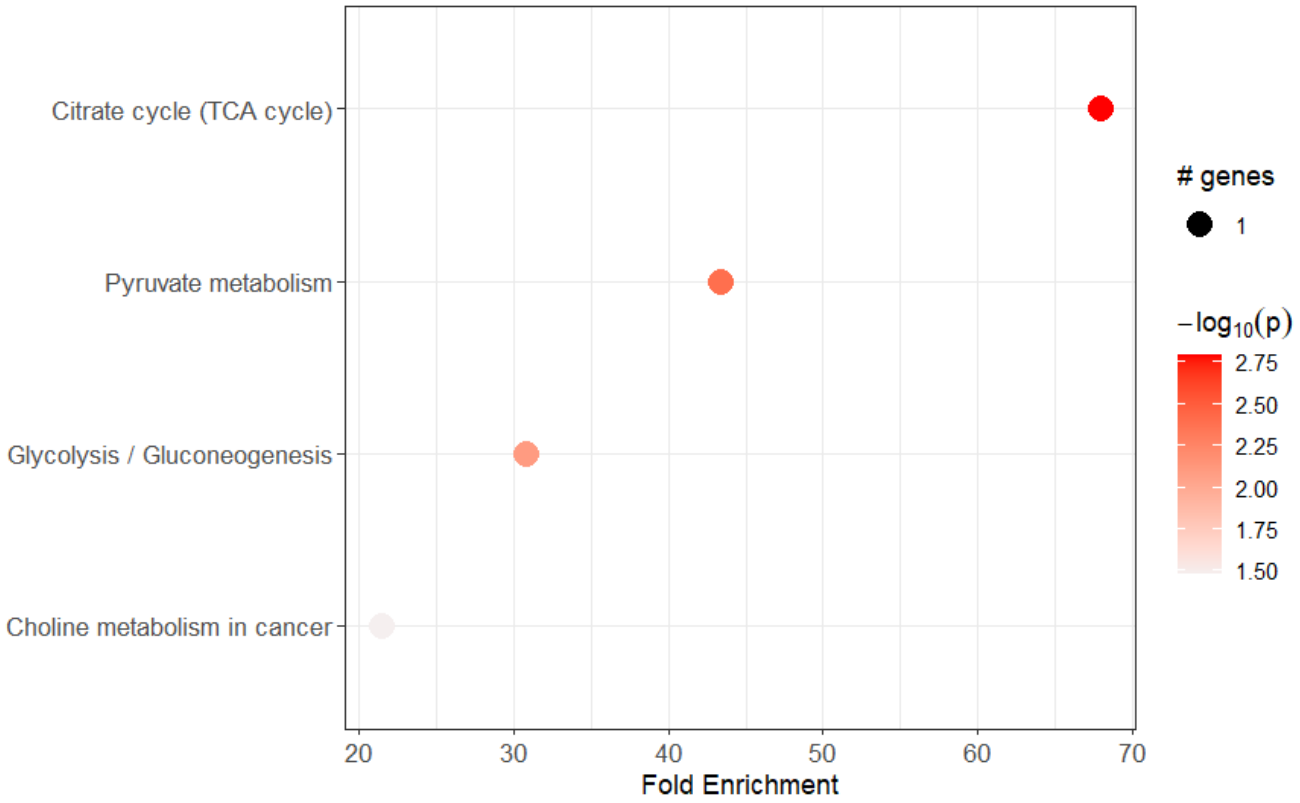


Figure 5: *Pathfinder plot: bipolar vs depression*