Informatics in Medicine Unlocked 26 (2021) 100740

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

A study of tools for differential co-expression analysis for RNA-Seq data

Tonmoya Sarmah, Dhruba K. Bhattacharyya *

Department of Computer Science and Engineering, Tezpur University, Tezpur, Assam, 784028, India

ARTICLE INFO

Keywords:
RNA-seq data
Co-expression networks
Differential co-expression analysis
GO enrichment analysis
Pathway enrichment analysis

ABSTRACT

A number of methods are being developed and used for analysis of gene expression data such as RNA-Seq data. Most of these tools focus on finding genes that are responsible for the disease conditions. Methods such as co-expression network generation, module detection and differential co-expression analysis are used to look into specific changes in the gene expression data among different conditions. In this paper, a comparative study of four differential co-expression analysis tools are presented, namely, WGCNA, DiffCorr, MODA and CEMiTool, for RNA-Seq data. The different methods used by these tools are studied and tested on schizophrenia and bipolar disorder datasets and their effectiveness in finding the related differentially co-expressed genes and pathways are being discussed. The relevancy of the resultant genes and pathways are decided on the basis of whether the genes and pathways are associated with the given disease conditions.

1. Introduction

Co-expression networks are being increasingly used to find correlations among genes and find modules and intra-modular hub genes. With the use of next-generation sequencing (NGS), such as, microarray and RNA-Seq technologies, there has been rapid development in the study of gene expressions. Gene expression analysis helps in determining the genes affected during the progression of a disease. Computational methods are used to identify those affected genes which serve as biomarkers and helps in determining potential drug targets for the disease. Gene co-expression networks (CEN) are one of such widely used computational methods for gene expression analysis. It can be visualized as an undirected graph, where a node represents a gene and a pair of nodes is connected with an edge if there is a correlation among them. This co-expression network helps to identify correlated genes which can be associated with biological processes or pathways [1]. However, a CEN would provide the correlated gene for only one condition. In situations where the gene expression pattern changes across conditions, such as disease and control, we need to find the difference in expression pattern among the modules of the CEN. Differential expression analysis would consider each gene as an individual entity and provide the individual genes that are differentially expressed among conditions. But this will not work for co-expression module networks where genes are a part of the network and they interact with each other [2]. Thus, this requires the use of differential co-expression analysis. Differential co-expression (DCE) analysis helps to look into condition specific changes of co-expression networks. In other words, if a set of co-expressed genes behave in a certain way or respond in a particular fashion to biological changes; it is termed as differentially co-expressed. Differential co-expression network analysis helps in the study of disease conditions and phenotypic variations in the correlation modules in the network where co-expression patterns vary across different conditions [1].

Different parameters have been studied for the construction of co-expression network and differential co-expression analysis of the co-expression modules. With the increase in use of DCE network analysis, tools are being developed to find differentially coexpressed gene in the DC network. In this study, we look into four tools used for differential co-expression network analysis, namely, WGCNA [3], DiffCorr [4], MODA [5], CEMiTool [6].

This paper presents the outcome of empirical analysis of the four tools and also a comparison of the methods used in the tools. Section 2 includes a literature review of the different popularly available methods used in DCE analysis tools. Section 3 includes an introduction of the tools used in this study and a comparison of their methods. Section 4 presents the analysis of the results which includes GO (Gene Ontology) enrichment analysis, pathway enrichment analysis and hub-gene identification and Section 5 presents a final discussion of the methods used in the DCE analysis tools.

2. Background

In this study, we are focused on differential co-expression analysis of RNA-Seq data. RNA sequencing is a next generation sequencing (NGS) technique that can perform DNA sequencing at a cheaper cost

* Corresponding author.
E-mail address: dkb@tezu.ernet.in (D.K. Bhattacharyya).

https://doi.org/10.1016/j.imu.2021.100740
Received 13 May 2021; Received in revised form 14 September 2021; Accepted 15 September 2021
Available online 22 September 2021

and it can provide better structural variation for co-expression analysis [7]. Dam et al. [8], provides an overview of tools and methods that are used to construct and analyze co-expression networks (CEN) that are constructed from gene expression data. The study also provides details and a brief comparison of tools used for differential co-expression (DCE) analysis. Hussain et al. [1], presents a detailed survey of co-expression, differential co-expression, differential network and connectivity for both RNA-Seq and micro-array data. Liu et al. [2], presents a study of three aspects of differential co-expression network analysis which includes topological comparison of the network, identification of differential co-expression modules and differentially coexpressed genes and identification of gene pairs. The study provides a review of existing tools and discusses their application to cancer research. Measures for proper construction of co-expression network are being studied to find ways to define the connections among nodes or genes in the co-expression network. Instead of assigning binary information, such as 1 for correlated and 0 for not correlated, Zhang et al. describes a framework for soft-thresholding to assign weights for the correlation connection of gene pairs [9]. This concept has been implemented in the DCE analysis tool WGCNA for construction of co-expression network [3]. Much work has been done for co-expression network construction of micro-array data compared to that of RNA-Seq data. In recent times, different concepts have been studied for construction of co-expression network for RNA-Seq data [10–12], such as Guilt-by-Association, Spearman or Pearson correlation coefficient, component-based methods among others. Most of the DCE analysis tools uses Spearman or Pearson correlation coefficient for construction of co-expression networks. These networks are used to study the differential co-expression pattern for genetic changes and disease conditions to find the responsible genes. Different clustering techniques are also used to explore and find patterns among the genes in the co-expression networks. Clustering techniques such as biclustering and triclustering are used on the co-expression networks to find modules which will be further analyzed for differentially coexpressed genes. Different clustering methods are being studied and new methods developed to find their effectiveness using different datasets [13–15].

Tools like DICER [16], DifCoNet [17], DiffCoEx [18], MODA [5], DiffCorr [4], DCGL [19], CEMiTool [6] have used different statistics for DCE analysis such as probabilistic scores, principal component analysis, *z*-score, topological overlap matrix etc. In this study we compare the techniques used by four tools, WGCNA, DiffCorr, MODA, CEMiTool, that has different methods for differential co-expression analysis and look into the results obtained when used on RNA-Seq data.

2.1. Differential co-expression analysis

Fig. 1 shows the tentative workflow followed for differential co-expression network analysis for this study. In this study, we have used RNA-Seq read count data as the input. This input needs to be preprocessed before any downstream analysis. Pre-processing methods such as normalization, missing value estimation, gene selection and batch effect normalization are used. This stage removes any lowly expressed or NaN genes. This preprocessed data is then used in the tools for finding the differentially coexpressed module. The tools first perform a correlation analysis using Pearson/ Spearman's correlation methods. This is used to create the co-expression network from which modules can be identified. The next step is to find the differentially coexpressed network modules. These modules are then used for GO enrichment analysis and pathway analysis for biological interpretation and to find the module hub genes or the differentially coexpressed genes. These hub genes are further analyzed to find its relevance with the disorder being studied. These results are then validated with existing established results.

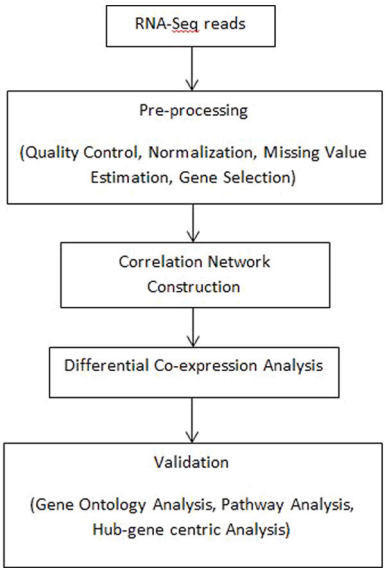

Fig. 1. Workflow for differential co-expression (DCE) network analysis.

3. Methods

3.1. WGCNA

Weighted Gene Co-expression Network Analysis (WGCNA) is one of the most commonly used tool for co-expression module detection and analysis. It detects clusters or modules of highly correlated genes using blockwise module detection. Here, the dataset is divided into blocks and modules are detected one block at a time. The modules among the blocks are summarized using module eigen gene or an intra-modular hub gene, to relate modules to one another. Hierarchical clustering and tree-cutting thresholds are used to identify the modules from the co-expression network. It uses Pearson correlation for computing correlations among genes across different conditions. The WGCNA package includes functions for network construction, module detection, gene selection, topological property calculation, data simulation, visualization and interfacing with external software [3].

3.2. DiffCorr

DiffCorr uses Pearson's correlation coefficient to build correlation network and identify pattern changes among the correlation networks of the two conditions. It calculates correlation matrices for dataset of each condition to build the correlation network, and identifies the eigen-molecule in the network based on its first principal component. It then identifies differential correlations among two conditions using Fisher's z-test [4].

3.3. MODA

MODA or Module Differential Analysis represents the gene co-expression network as a collection of modules and identifies differentially co-expressed sub-networks as conserved or condition specific modules. It is based on the concept of WGCNA wherein edge weights of the co-expression network are taken as the correlation coefficients of the gene pairs and hierarchical clustering method is used to detect the modules in the network. An optimal tree cutting threshold for the clustering is decided based on the average modularity for the weighted networks of the different conditions [5].

Table 1
Comparison of DCE analysis tools.

Method	CEN construction	Module detection	DCE analysis
WGCNA	Pearson correlation, Soft-thresholding power	Block-wise method, Hierarchical clustering, Dynamic tree-cut method	Topological overlap method
DiffCorr	Pearson correlation	Hierarchical clustering, Eigengene	Fisher's z-test
MODA	Pearson correlation	Hierarchical clustering, Density or Modularity	Similarity matrix using Jaccard index
CEMiTool	Pearson correlation, Soft-thresholding power	Hierarchical clustering, Dynamic tree-cut method	*z*-score normalized expression

Table 2
Input, output and availability of DCE analysis tools.

Method	Input	Output	Online/Offline	Reference (URL)
WGCNA	(samples × genes) expression profile	gene list	Offline	https://cran.r-project.org/web/packages/WGCNA/index.html
DiffCorr	(genes × samples) expression profile	text file	Offline	https://cran.r-project.org/web/packages/DiffCorr/index.html
MODA	(genes × samples) expression profile	heatmap, gene list	Offline	https://bioconductor.org/packages/release/bioc/html/MODA.html
CEMiTool	(genes × samples) expression profile, sample annotation file	HTML report	Online, Offline	https://cemitool.sysbio.tools/analysis, https://bioconductor.org/packages/release/bioc/html/CEMiTool.html

Table 3
List of GO terms enriched for the schizophrenia and bipolar disorder using different DCE tools for brain tissue nACC.

Method	GO-term	*p*-value	*q*-value
Schizophrenia			
WGCNA	GO:0030054 cell junction	0.01	1
	GO:0046628 positive regulation of insulin receptor signaling pathway	0.02	1
	GO:0019901 protein kinase binding	0.04	1
DiffCorr	GO:0005829 cytosol	0.004	0.51
	GO:0042059 negative regulation of epidermal growth factor receptor signaling pathway	0.009	1
	GO:0003924 GTPase activity	0.01	1
MODA	GO:0005654 nucleoplasm	1.06E−08	1.10E−06
	GO:0044822 poly(A) RNA binding	1.27E−05	0.001
	GO:0005515 protein binding	5.34E−04	0.01
	GO:0006364 rRNA processing	9.92E−04	0.22
CEMiTool	GO:0043195 terminal bouton	0.001	0.08
	GO:0006890 retrograde vesicle-mediated transport, Golgi to ER	0.003	0.28
	GO:0005829 cytosol	0.003	0.08
	GO:0019904 protein domain specific binding	0.01	0.62
Bipolar Disorder			
WGCNA	GO:0005576 extracellular region	0.004	0.39
	GO:0003682 chromatin binding	0.008	0.77
	GO:0006312 mitotic recombination	0.02	1
DiffCorr	GO:0006997 nucleus organization	1.76E−04	0.08
	GO:0030529 intracellular ribonucleoprotein complex	0.002	0.19
	GO:0044822 poly(A) RNA binding	0.006	0.52
	GO:0005515 protein binding	0.007	0.52
MODA	GO:0004864 protein phosphatase inhibitor activity	0.001	0.12
	GO:0044822 poly(A) RNA binding	0.009	0.29
	GO:0030659 cytoplasmic vesicle membrane	0.02	1
	GO:2001243 negative regulation of intrinsic apoptotic signaling pathway	0.03	1
CEMiTool	GO:0030054 cell junction	6.48E−05	0.001
	GO:0005829 cytosol	4.10E−04	0.009
	GO:0032403 protein complex binding	0.001	0.03
	GO:0006890 retrograde vesicle-mediated transport, Golgi to ER	0.003	0.57

Table 4
List of GO terms enriched for the schizophrenia and bipolar disorder using different DCE tools for brain tissue AnCg.

Method	GO-term	*p*-value	*q*-value
Schizophrenia			
WGCNA	GO:0030054 cell junction	0.002	0.23
	GO:0005524 ATP binding	0.01	1
	GO:0046628 positive regulation of insulin receptor signaling pathway	0.02	1
	GO:0019901 protein kinase binding	0.03	1
DiffCorr	GO:0005515 protein binding	5.87E−04	0.09
	GO:0044822 poly(A) RNA binding	0.01	1
	GO:0071204 histone pre-mRNA 3'end processing complex	0.02	0.94
	GO:0017148 negative regulation of translation	0.02	1
MODA	GO:0043968 histone H2A acetylation	0.001	0.54
	GO:0005654 nucleoplasm	0.003	0.2
	GO:0005829 cytosol	0.005	0.22
	GO:0008565 protein transporter activity	0.03	1
CEMiTool	GO:0043195 terminal bouton	0.001	0.08
	GO:0006890 retrograde vesicle-mediated transport, Golgi to ER	0.003	0.28
	GO:0005829 cytosol	0.003	0.08
	GO:0019904 protein domain specific binding	0.01	0.62
Bipolar Disorder			
WGCNA	GO:0005031 tumor necrosis factor-activated receptor activity	8.36E−04	0.01
	GO:0050684 regulation of mRNA processing	0.01	1
	GO:0005524 ATP binding	0.05	1
DiffCorr	GO:0044822 poly(A) RNA binding	6.65E−04	0.11
	GO:0005524 ATP binding	0.007	0.46
	GO:0005654 nucleoplasm	0.01	0.91
	GO:0045454 cell redox homeostasis	0.03	1
MODA	GO:0045026 plasma membrane fusion	0.01	1
	GO:0000220 vacuolar proton-transporting V-type ATPase, V0 domain	0.01	1
	GO:0015232 heme transporter activity	0.02	1
CEMiTool	GO:0043195 terminal bouton	4.53E−07	4.53E−05
	GO:0005829 cytosol	1.81E−05	6.05E−04
	GO:0005515 protein binding	7.49E−05	0.005

3.4. CEMiTool

CEMiTool or CoExpression Modules identification Tool combines discovery and analysis of co-expression modules into a single function. It provides a fully automated procedure where the tool selects parameters and performs functional analysis of the results. The input data consisting of genes and samples is filtered using a novel unsupervised gene filtering method based on inverse gamma distribution. Parameters such as soft-thresholding power (β) is decided by the tool for carrying out module detection. On providing a sample annotation file, this tool can perform gene set enrichment analysis (GSEA) using the R package *fgsea* (Fast Gene Set Enrichment Analysis) and on providing a gene-pathway list, it performs over representation analysis (ORA) using the R package *clusterProfiler*. All of these results including modules and graphs of the functional analyses are provided in a single HTML report [6] (see Table 1).

3.5. Comparison of tools

The correlation measure used in WGCNA, DiffCorr and CEMiTool, for this study, is the Pearson's correlation coefficient.

For the co-expression network construction, WGCNA uses a scale-free topological criterion to pick soft-thresholding power, β, for network construction. To construct a scale-free network, a lower β value has to be manually chosen by the user by consider the linear regression fit (R^2) and connectivity. The soft-thresholding impacts the network topology wherein a higher β value will lower the mean connectivity of the network. CEMiTool uses the concept of Cauchy sequences to automatically select the β value for a scale-free network construction.

All of these tools uses hierarchical clustering for module detection using the R function *hclust*. Hierarchical clustering is an unsupervised method to identify gene modules where it does not require pre-defined gene sets. It results in a dendrogram where every branch corresponds to a module. WGCNA and CEMiTool uses Dynamic Tree Cut method to select the modules. Although studies are going on for the selection of optimal cutting parameters for this method, the default values used for WGCNA has shown to be working well for several applications [3]. For large datasets, WGCNA provides a block-wise module detection method. The dataset is divided into clusters or blocks using k-means clustering technique and hierarchical clustering is applied to each of these blocks to find modules within the blocks. To summarize the modules across blocks, a weighted average expression profile of the module, called the module eigengene is calculated. Modules with highly correlated eigengenes are merged into a single module. In DiffCorr, this eigengene or eigen molecule is used to determine whether any two correlated modules in the network are significantly different. The eigen molecule is decided based on the first principal component of the data matrix of a module and it represents the correlation pattern within the module [4]. R package pcaMethods is used to perform principal component analysis (PCA), to find the first 10 principal components which are then used to test differential correlations among molecules. In MODA, the cutting height of hierarchical clustering tree is determined using one of the two methods provided in the tool, i.e., density and modularity. The density of a module is based on similarity between any two genes and the number of genes in the module. Modularity is based on the number of edges and degree of the genes in the network. The user has to specify the method to be used for choosing the height of the hierarchical tree.

Table 5
List of GO terms enriched for the schizophrenia and bipolar disorder using different DCE tools for brain tissue DLPFC.

Method	GO-term	*p*-value	*q*-value
Schizophrenia			
WGCNA	GO:0005524 ATP binding	0.01	0.76
	GO:0046628 positive regulation of insulin receptor signaling pathway	0.01	1
	GO:0030054 cell junction	0.04	1
DiffCorr	GO:0045893 positive regulation of transcription, DNA-templated	0.004	1
	GO:0005743 mitochondrial inner membrane	0.005	0.39
	GO:0005515 protein binding	0.02	1
MODA	GO:0006810 transport	0.003	1
	GO:0030117 membrane coat	0.005	0.7
	GO:0016757 transferase activity, transferring glycosyl groups	0.02	0.1
CEMiTool	GO:0019904 protein domain specific binding	4.70E−05	0.002
	GO:0044325 ion channel binding	1.85E−04	0.004
	GO:0005829 cytosol	7.24E−04	0.03
	GO:0007223 Wnt signaling pathway, calcium modulating pathway	7.86E−04	0.06
Bipolar Disorder			
WGCNA	GO:0030054 cell junction	0.01	0.96
	GO:0046628 positive regulation of insulin receptor signaling pathway	0.01	1
	GO:0019901 protein kinase binding	0.02	1
	GO:0005524 ATP binding	0.04	1
DiffCorr	GO:0005829 cytosol	0.01	1
	GO:0006470 protein dephosphorylation	0.02	1
	GO:0005102 receptor binding	0.02	1
	GO:0005515 protein binding	0.04	1
MODA	GO:0008066 glutamate receptor activity	1.24E−04	0.007
	GO:0016021 integral component of membrane	1.55E−04	0.006
	GO:0005886 plasma membrane	5.81E−04	0.01
	GO:0051966 regulation of synaptic transmission, glutamatergic	7.12E−04	0.05
CEMiTool	GO:0043195 terminal bouton	3.50E−05	0.001
	GO:0002576 platelet degranulation	2.02E−04	0.04
	GO:0005829 cytosol	7.24E−04	0.02
	GO:0019904 protein domain specific binding	0.001	0.12

WGCNA uses topological overlap matrix (TOM) to find differential correlations among the modules. DiffCorr uses the top 10 principal component to find differential correlations among modules as well as performs pair-wise differential correlation analysis among molecules using the Fisher's z-test. The correlation coefficients of the two conditions are transformed using Fisher's transformation, and a Z value, representing the difference between the correlations, is calculated. In MODA, pair-wise comparison of modules is done to create a similarity matrix of modules in the co-expression network. The similarity is evaluated using Jaccard index. The similarity matrix is denoted as B, where B_{ij} denotes the similarity between ith module of network N_1, i.e., $N_1(A_i)$ and jth module of network N_2, i.e., $N_2(A_j)$, where N_1 is the background set of genes consisting of samples from all the conditions and N_2 is the set containing all samples except the samples belonging to a condition D. Thus, B is calculated as - $B_{ij} = \frac{N_1(A_i) \cap N_2(A_j)}{N_1(A_i) \cup N_2(A_j)}$. The sum of the rows in B is calculated as $s_i = \sum_j B_{ij}$, where s_i denotes the affect of condition D on the ith module of network N_1. Two thresholds are considered, θ_1 and θ_2. If the frequency of a module is less than $min(s) + \theta_1$, it is considered as a condition specific module, and if it is more than $mas(s) - \theta_2$, it is considered as a condition conserved module [5]. CEMiTool requires a sample annotation file to perform gene set enrichment analysis using the R package $fgsea$ to find the co-expression modules in the dataset. A z-score normalized expression is calculated for samples of each condition to find whether the activity of a modules is altered in different conditions (see Fig. 2 and Table 2).

4. Results

Dataset used for this study is a RNA-Seq dataset of human post-mortem brain tissues for mental disorders, GSE80655, consisting of 281 samples for 57905 genes. It has gene expressions for bipolar disorder, schizophrenia and control conditions, corresponding to three types of brain tissues — nAcc (nucleus accumbens), AnCg (anterior cingulate cortex), DLPFC (dorsolateral prefrontal cortex) [20].

The resulted modules obtained from the four differential coexpression network analysis tools are analyzed for Gene Ontology (GO) enrichment and pathway enrichment. Functional annotation tool DAVID (Database for Annotation, Visualization and Integrated Discovery) has been used for validation of the co-expression modules. It is a web-based tool that provides GO enrichment analysis and batch annotation to find the most relevant GO terms associated with a given gene list. Also, the DAVID Pathway Viewer uses KEGG and BioCarta pathways to display the gene list on pathway maps [21].

4.1. GO enrichment analysis

The GO enrichment analysis on a set of genes will find the GO terms that are represented using annotations for the gene set. To test its significance a hypergeometric test based method is used to find statistical values like the p and q values to define the GO enrichment of the modules. It tests the null hypothesis to check whether the enrichment of an annotation is purely by chance. The p-value is the probability of obtaining atleast x genes out of the total n number of genes that are annotated to a particular GO term and q-value is a

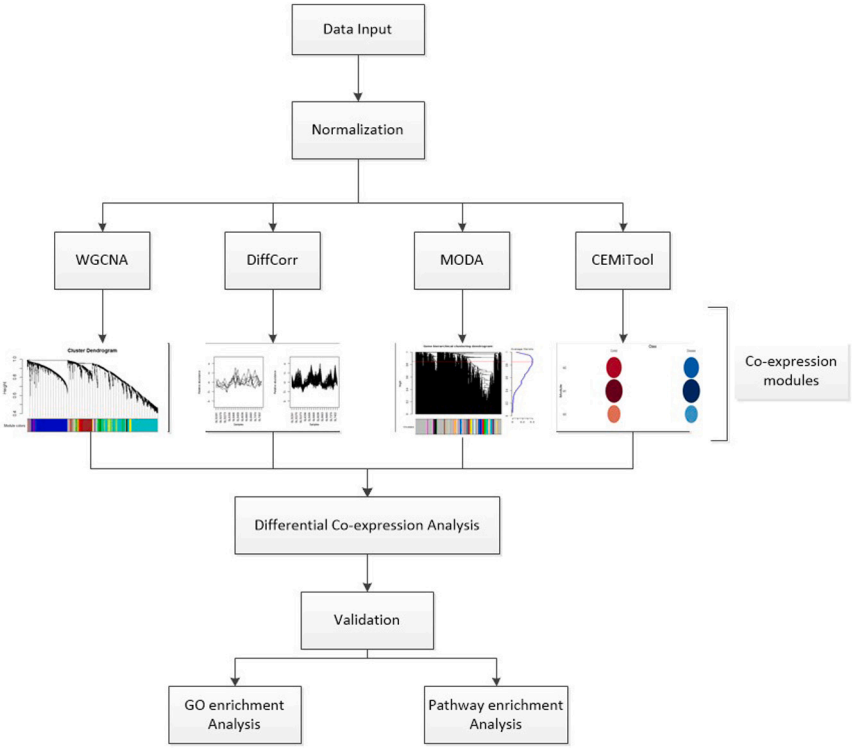

Fig. 2. Workflow for empirical study of tools.

Table 6
p-values and *q*-values for pathways associated with schizophrenia and bipolar disorder for brain tissue nAcc.

Tool	Pathways	*p*-value	*q*-value
Schizophrenia			
WGCNA	hsa04744:Phototransduction	0.05	1
	hsa04727:GABAergic synapse	0.1	1
DiffCorr	hsa04722:Neurotrophin signaling pathway	0.001	0.09
	hsa04668:TNF signaling pathway	0.06	1
MODA	hsa03013:RNA transport	2.66E−04	0.006
	hsa03008:Ribosome biogenesis in eukaryotes	0.03	0.41
CEMiTool	hsa04721:Synaptic vesicle cycle	0.001	0.02
Bipolar disorder			
WGCNA	hsa04722:Neurotrophin signaling pathway	0.02	0.99
	hsa04910:Insulin signaling pathway	8.0.03	0.99
DiffCorr	hsa04144:Endocytosis	0.003	0.18
MODA	hsa04071:Sphingolipid signaling pathway	0.2	1
CEMiTool	hsa04721:Synaptic vesicle cycle	0.006	0.41
	hsa04727:GABAergic synapse	0.01	0.41
	hsa04114:Oocyte meiosis	0.02	0.46

statistical measure that gives the false discovery rate (FDR). Lower the value of *p* and *q*-values, higher is the significance of the module.

The significant GO terms obtained using the different tools and their respective *p* and *q* values are shown in Tables 3–5.

From the results, it has found that the modules obtained by MODA and CEMiTool are more significant according to the respective *p*-values.

4.2. Pathway enrichment analysis

Pathway enrichment analysis identifies biological pathways in a gene set that has the possibility of being more than just a chance occurrence. Pathway analysis is based on the assumption that genes that are involved in the same biological processes or functions are correlated in terms of expression levels.

4.2.1. Brain tissue - nAcc

1. WGCNA — Enriched pathways found for schizophrenia are *Phototransduction* and *GABAergic synapse* and for bipolar disorder are *Neurotrophin signaling pathway, Ras signaling pathway* and *Insulin signaling pathway*.
2. DiffCorr — Enriched pathways found for schizophrenia are *:Neurotrophin signaling pathway* and *TNF signaling pathway* and for bipolar disorder is *Endocytosis*.
3. MODA — Enriched pathways found for schizophrenia are *:RNA transport* and *Ribosome biogenesis in eukaryotes* and for bipolar disorder is *Sphingolipid signaling pathway*.
4. CEMiTool — Enriched pathways found for schizophrenia is *Synaptic vesicle cycle* and for bipolar disorder are *Synaptic vesicle cycle*, *GABAergic synapse* and *Oocyte meiosis*.

Table 7
p-values and *q*-values for pathways associated with schizophrenia and bipolar disorder for brain tissue AnCg.

Tool	Pathways	*p*-value	*q*-value
Schizophrenia			
WGCNA	hsa04744:Phototransduction	0.05	0.97
	hsa04727:GABAergic synapse	0.1	1
DiffCorr	hsa03013:RNA transport	0.03	1
MODA	h_igf1mtorPathway:Skeletal muscle hypertrophy is regulated via AKT/mTOR pathway	0.03	0.54
CEMiTool	hsa04721:Synaptic vesicle cycle	0.001	0.02
Bipolar disorder			
WGCNA	hsa04060:Cytokine–cytokine receptor interaction	0.06	1
DiffCorr	hsa05140:Leishmaniasis	0.03	1
MODA	hsa05110:Vibrio cholerae infection	0.01	0.61
CEMiTool	h_ndkDynaminPathway:Endocytotic role of NDK, Phosphins and Dynamin	0.001	0.03
	hsa04721:Synaptic vesicle cycle	0.003	0.19

Table 8
p-values and *q*-values for pathways associated with schizophrenia and bipolar disorder for brain tissue DLPFC.

Tool	Pathways	*p*-value	*q*-value
Schizophrenia			
WGCNA	hsa04744:Phototransduction	0.04	1
	hsa04727:GABAergic synapse	0.1	1
DiffCorr	hsa05169:Epstein-Barr virus infection	0.02	1
	hsa04146:Peroxisome	0.07	1
MODA	hsa04142:Lysosome	0.1	1
CEMiTool	hsa04713:Circadian entrainment	2.02E−04	0.007
	hsa04114:Oocyte meiosis	3.20E−04	0.007
	hsa04744:Phototransduction	5.25E−04	0.008
Bipolar disorder			
WGCNA	hsa04744:Phototransduction	0.04	1
DiffCorr	map00062:Fatty acid elongation	0.1	6.2E−3
MODA	hsa04080:Neuroactive ligand–receptor interaction	6.78E−04	0.01
	hsa04724:Glutamatergic synapse	0.01	0.14
CEMiTool	h_gpcrPathway:Signaling Pathway from G-Protein Families	0.008	0.25
	hsa04713:Circadian entrainment	0.009	0.23
	hsa04114:Oocyte meiosis	0.01	0.23

Table 6 shows the associated *p*-value and *q*-values for the above pathways of schizophrenia and bipolar disorder for brain tissue nAcc.

4.2.2. Brain tissue — AnCg

1. WGCNA — Enriched pathway for both schizophrenia are *Phototransduction* and *GABAergic synapse* and for bipolar disorder is *Cytokine–cytokine receptor interaction*.
2. DiffCorr — Enriched pathways found for schizophrenia is *RNA transport* and for bipolar disorder is *Leishmaniasis*.
3. MODA — Enriched pathways found for schizophrenia is *Skeletal muscle hypertrophy is regulated via AKT/mTOR pathway* and for bipolar disorder is *Vibrio cholerae infection*.
4. CEMiTool — Enriched pathways found for schizophrenia is *Synaptic vesicle cycle* and for bipolar disorder are *Endocytotic role of NDK, Phosphins and Dynamin* and *:Synaptic vesicle cycle*.

Table 7 shows the associated *p*-value and *q*-values for the above pathways of schizophrenia and bipolar disorder for brain tissue AnCg.

4.2.3. Brain tissue - DLPFC

1. WGCNA — Enriched pathways found for schizophrenia are *Phototransduction* and *GABAergic synapse* and for bipolar disorder is *Phototransduction*.
2. DiffCorr — Enriched pathways found for schizophrenia are *Epstein-Barr virus infection* and *Peroxisom* and for bipolar disorder is *Fatty acid elongation*.
3. MODA — Enriched pathways found for schizophrenia is *Lysosome* and for bipolar disorder are *Neuroactive ligand–receptor interaction* and *Glutamatergic synapse*.
4. CEMiTool — Enriched pathways found for schizophrenia are *Circadian entrainment*, *Oocyte meiosis* and *Phototransduction* and for bipolar disorder are *Signaling Pathway from G-Protein Families*, *Circadian entrainment* and *Oocyte meiosis*.

Table 8 shows the associated *p*-value and *q*-values for the above pathways of schizophrenia and bipolar disorder for brain tissue DLPFC.

Phototransduction and *GABAergic synapse* found to be associated with neurodegenerative diseases such as schizophrenia [22–24] are enriched by WGCNA for all the three tissues nAcc, AnCg and DLPFC. *Cytokine–cytokine receptor interaction*, which is associated with both schizophrenia [22,25,26] and bipolar disorder [27] have been found using WGCNA in the brain tissue AnCg for bipolar disorder. *Neurotrophin signaling pathway* is another pathway associated with both conditions [28–30] and it has been found using DiffCorr for schizophrenia and using WGCNA for bipolar disorder in tissue nAcc. Pathways such as *TNF signaling pathway*, *RNA transport* and *Peroxisome* associated with schizophrenia [22,31,32] has been found to be enriched using

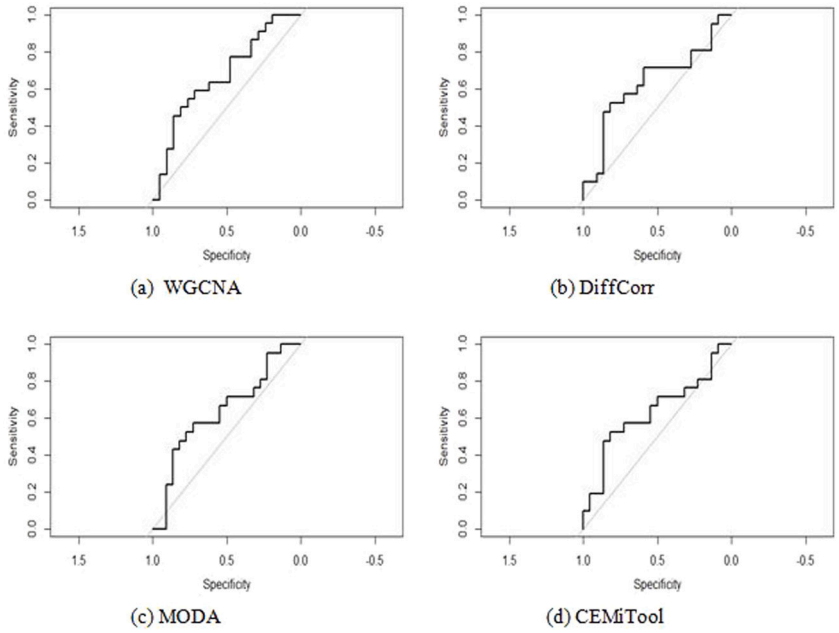

Fig. 3. ROC curve for different tools used in schizophrenia dataset for brain tissue nAcc. AUC - (a) 0.67 (b) 0.64 (c) 0.63 (d)0.63.

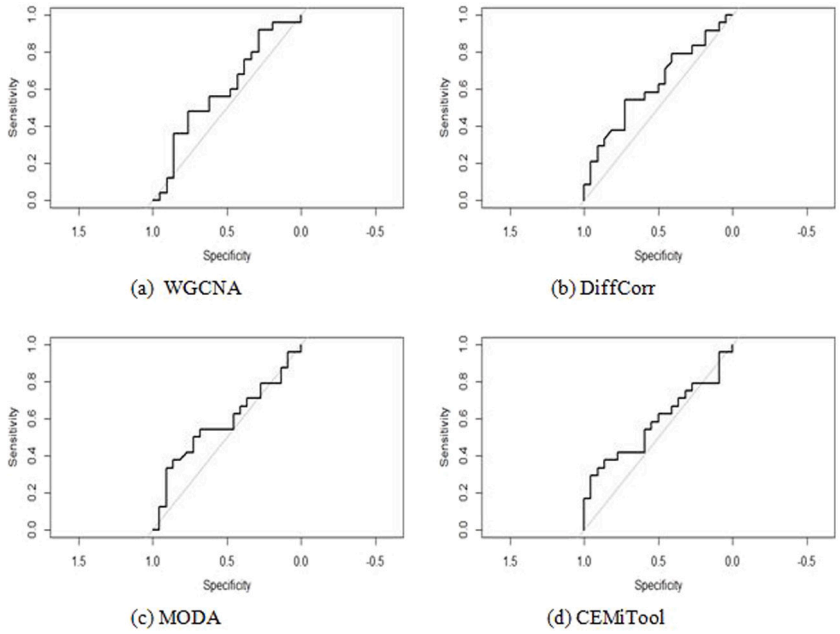

Fig. 4. ROC curve for different tools used in bipolar disorder dataset for brain tissue nAcc. AUC - (a) 0.6 (b) 0.62 (c) 0.6 (d)0.6.

DiffCorr for nAcc, AnCg and DLPFC respectively. *Epstein-Barr virus infection* that has been found to infect the central nervous system resulting in increased risk of schizophrenia [33], is enriched using DiffCorr for DLPFC. Pathways enriched using MODA such as *Ribosome biogenesis in eukaryotes* for nAcc, *AKT/mTOR pathway* for AnCg and *Lysosome* for DLPFC are found to be associated with schizophrenia [22,26,34] and pathways *Sphingolipid signaling pathway* and *Glutamatergic synapse* have been found to be associated with bipolar disorder [35–37]. *Synaptic vesicle cycle* found to be associated with schizophrenia and bipolar disorder [38] is enriched using CEMiTool for nAcc and AnCg. *Phototransduction* is also found using CEMiTool in DLPFC for schizophrenia and *Circadian entrainment* associated with bipolar disorder is found using CEMiTool for DLPFC [39].

The Figs. 3 to 8 shows the ROC curves for results obtained for differential coexpression analysis using the four tools. The AUC of the ROC curves are above 0.6 which indicates that the diagnosis is not a chance occurrence.

4.3. Hub gene identification

The gene with the highest degree of connectivity in a co-expression network is called the hub gene for the particular co-expression module. The set of hub genes act as the informative genes. The co-expression analysis has been extended to find hub genes for each of the disease conditions.

In nAcc, hub genes PPHLN1 [40] and EIF2 A [41] are associated with schizophrenia. Hub genes COX15, involved in mitochondrial respiratory chain [42], KDM1 A, involved with Wnt signaling pathway [43], EIF4ENIF1 [44] and MCCC1 [45] are associated with bipolar disorder.

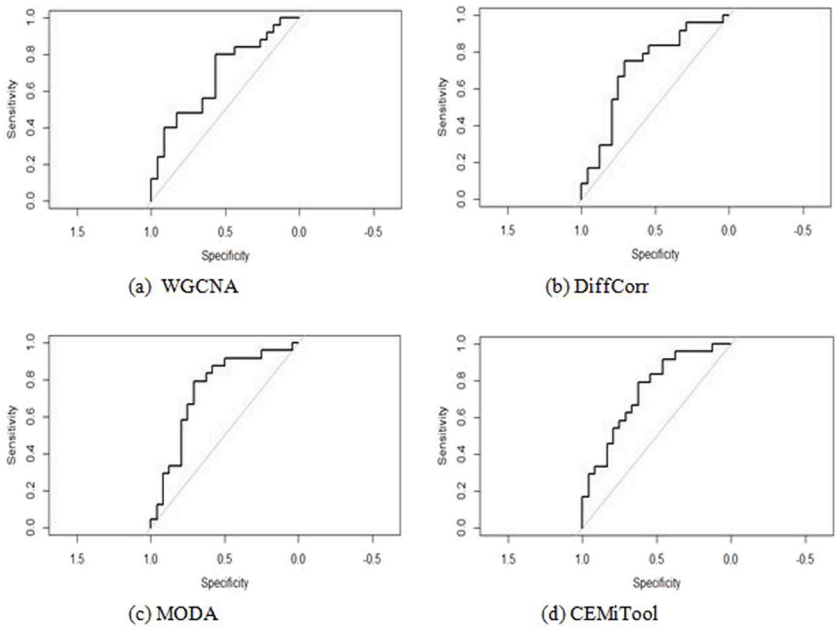

Fig. 5. ROC curve for different tools used in schizophrenia dataset for brain tissue AnCg. AUC - (a) 0.68 (b) 0.71 (c) 0.74 (d)0.74.

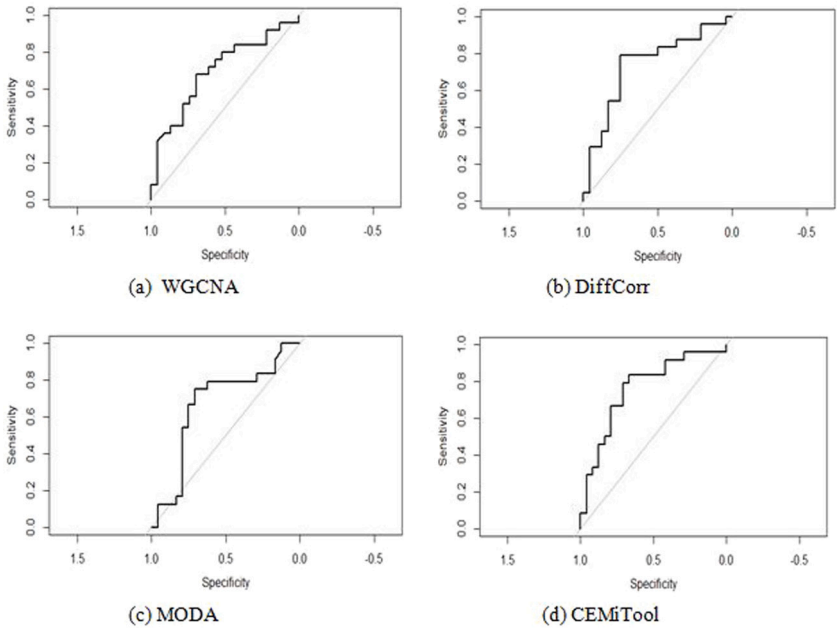

Fig. 6. ROC curve for different tools used in bipolar disorder dataset for brain tissue AnCg. AUC - (a) 0.7 (b) 0.74 (c) 0.67 (d) 0.76.

In AnCg, hub genes TLK2, associated with intellectual disability [46], CIAO1, involved in cytosolic iron-sulfur cluster protein assembly pathway [47], RAB18, involved in neurodevelopment [48], EIF4ENIF1 [44] and NRDC [49] are associated with schizophrenia. Hub genes EIF2AK1, involved in oxidative stress pathway [50] and ACK [51] are associated with bipolar disorder.

In DLPFC, hub genes RNF34, involved in protein-DNA interactions [52], EIF2B1, involved in stress related pathways [50], MTREX, involved in RNA metabolism [53], TFCP2 [54], WDR33 [55] and CHCHD3 [56] are associated with schizophrenia. Hub genes RAB7 A [57] and MICU1 [58] are associated with bipolar disorder.

5. Discussion

Based on the analysis, the four tools resulted in statistically significant differentially coexpressed genes that are also enriched with functional annotations. According to the GO enrichment analysis modules found using MODA and CEMiTool were found to be more significant in most of the cases as compared to others in terms of *p*-values. In pathway enrichment analysis, pathways associated with the given conditions were found by the tools used. It has been found that DiffCorr and MODA focuses on providing less but more precise result in most of the cases.

With the increase in use of RNA-Seq data, there is a demand for tools that can specifically work on RNA-Seq data to find precise and functionally enriched differentially coexpressed genes. Newer tools are also focused on providing a user-friendly output, as seen in CEMiTool where the modules and its analysis are presented in a single report.

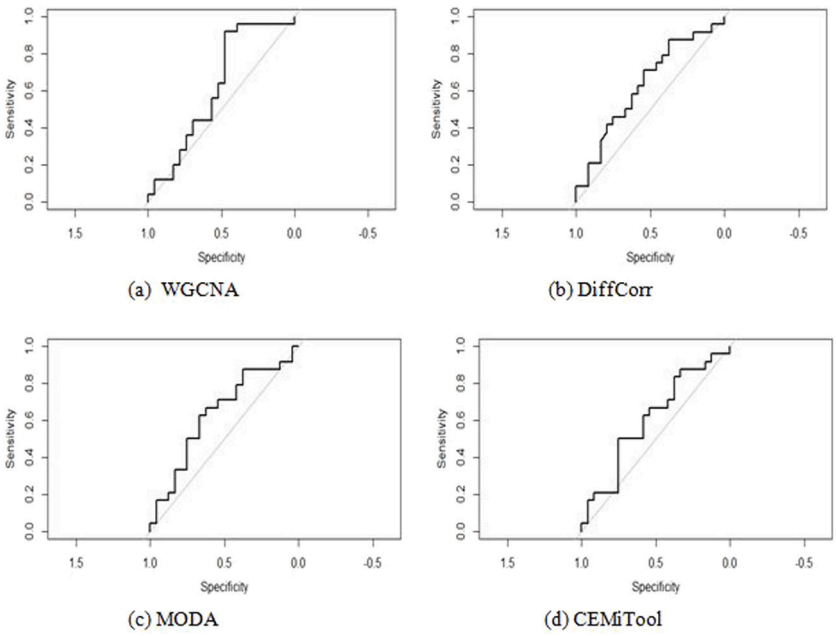

Fig. 7. ROC curve for different tools used in schizophrenia dataset for brain tissue DLPFC. AUC - (a) 0.62 (b) 0.63 (c) 0.63 (d)0.6.

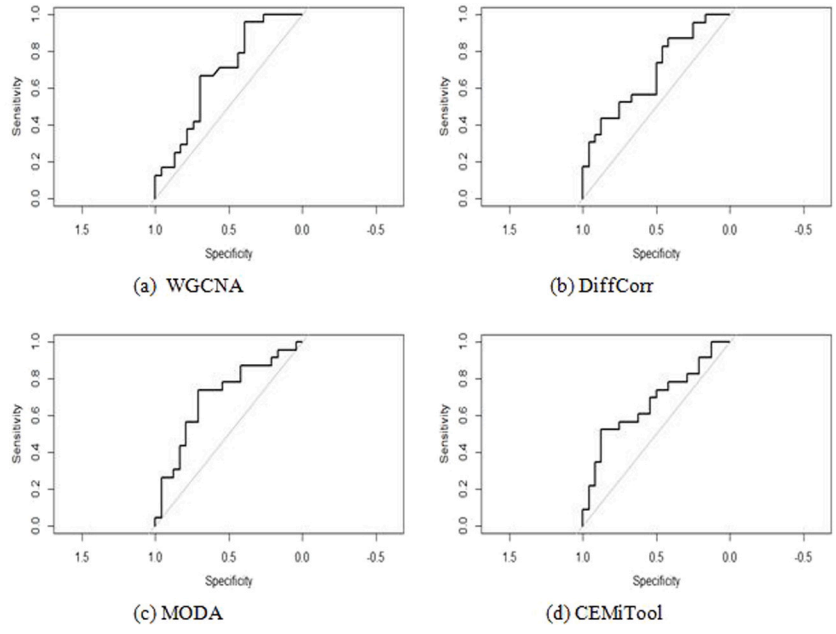

Fig. 8. ROC curve for different tools used in bipolar disorder dataset for brain tissue DLPFC. AUC - (a) 0.68 (b) 0.68 (c) 0.7 (d) 0.67.

Most of the tools uses Pearson correlation for the correlation network construction and hierarchical clustering for the module detection. Additional methods, such as dynamic tree cut method and density, are used to improve the results. All the tools used in this study uses different methods for the differential coexpression analysis. Emphasis is given on finding precise and more specific differentially coexpressed genes.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Ahmed Chowdhury Hussain, Bhattacharyya Dhruba Kumar, Kalita Jugal Kumar. (Differential) co-expression analysis of gene expression: a survey of best practices. IEEE/ACM Trans Comput Biol Bioinform 2019;17(4):1154–73.

[2] Bao-Hong Liu. Differential coexpression network analysis for gene expression data. In: Computational systems biology. New York, NY: Humana Press; 2018, p. 155–65.

[3] Peter Langfelder, Horvath Steve. WGCNA: an R Package for weighted correlation network analysis. BMC Bioinformatics 2008;9(1):1–13.

[4] Atsushi Fukushima. DiffCorr: an R Package to analyze and visualize differential correlations in biological networks. Gene 2013;518(1):209–14.

[5] Dong Li, et al. MODA: Module differential analysis for weighted gene co-expression network. 2016, arXiv preprint arXiv:1605.04739.

[6] Russo Pedro ST, et al. CEMiTool: a Bioconductor Package for performing comprehensive modular co-expression analyses. BMC Bioinformatics 2018;19(1):1–13.

[7] Tulika Kakati, et al. Comparison of methods for differential co-expression analysis for disease biomarker prediction. Comput Biol Med 2019;113:103380.
[8] Van Dam Sipko, et al. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform 2018;19(4):575–92.
[9] Bin Zhang, Horvath Steve. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 2005;4(1).
[10] Sara Ballouz, Verleyen Wim, Gillis Jesse. Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. Bioinformatics 2015;31(13):2123–30.
[11] Iancu Ovidiu D, et al. Utilizing RNA-seq data for de novo coexpression network inference. Bioinformatics 2012;28(12):1592–7.
[12] Shengjun Hong, et al. Canonical correlation analysis for RNA-seq co-expression networks. Nucleic Acids Res 2013;41(8):e95.
[13] Chandrasekhar T, Thangavel K, Elayaraja E. Effective clustering algorithms for gene expression data. 2012, arXiv preprint arXiv:1201.4914.
[14] Mahanta P, et al. Triclustering in gene expression data analysis: a selected survey. In: 2011 2nd national conference on emerging trends and applications in computer science. IEEE; 2011.
[15] Sauravjoyti Sarmah, Bhattacharyya Dhruba K. An effective technique for clustering incremental gene expression data. IJCSI Int J Comput Sci Issues 2010;7(3):31–41.
[16] David Amar, Safer Hershel, Shamir Ron. Dissection of regulatory networks that are altered in disease via differential co-expression. PLoS Comput Biol 2013;9(3):e1002955.
[17] Elpidio-Emmanuel Gonzalez-Valbuena, Treviño Víctor. Metrics to estimate differential co-expression networks. BioData Min 2017;10(1):1–15.
[18] Tesson Bruno M, Breitling Rainer, Jansen Ritsert C. DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules. BMC Bioinformatics 2010;11(1):1–9.
[19] Bao-Hong Liu, et al. DCGL: an r package for identifying differentially coexpressed genes and links from gene expression microarray data. Bioinformatics 2010;26(20):2637–8.
[20] Ramaker Ryne C, et al. Post-mortem molecular profiling of three psychiatric disorders. Genome Med 2017;9(1):1–12.
[21] Glynn Dennis, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4(9):1–11.
[22] Schizophrenia pathways, http://www.polygenicpathways.co.uk/keggszgenes.htm.
[23] Cameron Lenahan, et al. Rhodopsin: A potential biomarker for neurodegenerative diseases. Front Neurosci 2020;14:326.
[24] de Jonge Jeroen C, et al. GABAergic Mechanisms in schizophrenia: linking postmortem and in vivo studies. Front Psychiatry 2017;8:118.
[25] Turrin Nicolas P, Plata-Salamán Carlos R. Cytokine–cytokine interactions and the brain. Brain Res Bull 2000;51(1):3–9.
[26] Carter CJ. Schizophrenia: a pathogenetic autoimmune disease caused by viruses and pathogens and dependent on genes. J Pathogens 2011;2011.
[27] Guimarães Barbosa Izabela, et al. Cytokines in bipolar disorder: paving the way for neuroprogression. Neural Plast 2014;2014.
[28] Neurotrophin Signaling Pathway - Creative Diagnostic, https://www.creative-diagnostics.com/neurotrophin-signaling-pathway.htm.
[29] Mariela Mitre, Mariga Abigail, Chao Moses V. Neurotrophin signalling: novel insights into mechanisms and pathophysiology. Clin Sci 2017;131(1):13–23.
[30] Galit Shaltiel, Chen Guang, Manji Husseini K. Neurotrophic signaling cascades in the pathophysiology and treatment of bipolar disorder. Curr Opin Pharmacol 2007;7(1):22–6.
[31] Zsuzsanna HosethEva, et al. A study of TNF pathway activation in schizophrenia and bipolar disorder in plasma and brain tissue. Schizophr Bull 2017;43(4):881–90.
[32] Johannes Berger, et al. Peroxisomes in brain development and function. Biochim Biophys Acta (BBA) Mol. Cell Res. 1863;5(2016):934–55.
[33] Faith Dickerson, et al. Schizophrenia is associated with an aberrant immune response to Epstein–Barr virus. Schizophr Bull 2019;45(5):1112–9.
[34] Radhika Chadha, S. James Meador-Woodruff. S 192. AKT-mTOR signaling pathway is downregulated in schizophrenia. Schizophr Bull 2018;44(1):S400.
[35] Sujatha Narayan, Thomas Elizabeth A. Sphingolipid abnormalities in psychiatric disorders: a missing link in pathology. Front Biosci 2011;16:1797–810.
[36] Guang Chen, Henter Ioline D, Manji Husseini K. Presynaptic glutamatergic dysfunction in bipolar disorder. Biol Psychiat 2010;67(11):1007.
[37] Carlos Zarate, et al. Glutamatergic modulators: the future of treating mood disorders? Harvard Rev Psychiatry 2010;18(5):293–303.
[38] Willcyn Tang, et al. Stimulation of synaptic vesicle exocytosis by the mental disease gene DISC1 is mediated by N-type voltage-gated calcium channels. Front Synaptic Neurosci 2016;8:15.
[39] Wehr Thomas A. Bipolar mood cycles associated with lunar entrainment of a circadian rhythm. Transl Psychiatry 2018;8(1):1–6.
[40] Legge Sophie E, et al. Genome-wide common and rare variant analysis provides novel insights into clozapine-associated neutropenia. Mol Psychiatry 2017;22(10):1502–8.
[41] Trinh Mimi A, et al. Brain-specific disruption of the eif2α kinase PERK decreases ATF4 expression and impairs behavioral flexibility. Cell Reports 2012;1(6):676–88.
[42] Suzanne Gonzalez. The role of mitonuclear incompatibility in bipolar disorder susceptibility and resilience against environmental stressors. Front Genet 2021;12.
[43] Emily Ricq. Chemical neurobiology of the histone lysine demethylase KDM1A. Diss. 2016.
[44] Fuquan Zhang, et al. Systematic association analysis of microrna machinery genes with schizophrenia informs further study. Neurosci Lett 2012;520(1):47–50.
[45] Melanie Föcking, et al. Proteomic analysis of the postsynaptic density implicates synaptic function and energy pathways in bipolar disorder. Transl Psychiatry 2016;6(11):e959.
[46] Cicek M, Cicek E, Erson-Bensan AE. TLK2 (tousled like kinase 2), atlas genet cytogenet oncol haematol. 2020, in press.
[47] Ciao1 (cytosolic iron-sulfur assembly component 1) Ictidomys tridecemlineatus, https://rgd.mcw.edu/rgdweb/report/gene/main.html?id=12721123.
[48] Chih-Ya Cheng, et al. The association of RAB18 gene polymorphism (rs3765133) with cerebellar volume in healthy adults. Cerebellum 2014;13(5):616–22.
[49] Bernstein H-G, et al. Nardilysin in human brain diseases: both friend and foe. Amino Acids 2013;45(2):269–78.
[50] Carter Christopher J. eIF2B and oligodendrocyte survival: where nature and nurture meet in bipolar disorder and schizophrenia? Schizophr Bull 2007;33(6):1343–53.
[51] Thompson AGB, et al. Genome-wide association study of behavioural and psychiatric features in human prion disease. Transl Psychiatry 2015;5(4):e552.
[52] Gil Stelzer, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinform 2016;54(1):1–30.
[53] Marta Barrera-Conde, et al. Cannabis use induces distinctive proteomic alterations in olfactory neuroepithelial cells of Schizophrenia patients. J Pers Med 2021;11(3):160.
[54] Upasana Bhattacharyya, et al. Revisiting schizophrenia from an evolutionary perspective: An association study of recent evolutionary markers and Schizophrenia. Schizophr Bull 2021;47(3):827–36.
[55] Fengbiao Mao, et al. Post-transcriptionally impaired de novo mutations contribute to the genetic etiology of four neuropsychiatric disorders. Biorxiv 2019;175844.
[56] Estefanía Piñero-Martos, et al. Disrupted in schizophrenia 1 (DISC1) is a constituent of the mammalian mitochondrial contact site and cristae organizing system (MICOS) complex, and is essential for oxidative phosphorylation. Hum Mol Gen 2016;25(19):4157–69.
[57] Jin RheeSang, et al. Comparison of serum protein profiles between major depressive disorder and bipolar disorder. BMC Psychiatry 2020;20(1):1–11.
[58] Roghaiyeh Safari, et al. Mutation/SNP analysis in EF-hand calcium binding domain of mitochondrial Ca 2+ uptake 1 gene in bipolar disorder patients. J. Integr. Neurosci. 2016;15(02):163–73.