Computational and Structural Biotechnology Journal 20 (2022) 5054–5064

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

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

Classification for psychiatric disorders including schizophrenia, bipolar disorder, and major depressive disorder using machine learning

Qingxia Yang [a,*], Qiaowen Xing [a], Qingfang Yang [b], Yaguo Gong [c,*]

[a] Department of Bioinformatics, Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[b] Second Affiliated Hospital, Zhejiang Chinese Medical University, Hangzhou 310005, China
[c] School of Pharmacy, Macau University of Science and Technology, Macau

ARTICLE INFO

Article history:
Received 25 May 2022
Received in revised form 8 September 2022
Accepted 8 September 2022
Available online 12 September 2022

Keywords:
Classification
Psychiatric disorder
Schizophrenia
Bipolar disorder
Major depressive disorder

ABSTRACT

Schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD) are the most common psychiatric disorders. Because there were lots of overlaps among these disorders from genetic epidemiology and molecular genetics, it is hard to realize the diagnoses of these psychiatric disorders. Currently, plenty of studies have been conducted for contributing to the diagnoses of these diseases. However, constructing a classification model with superior performance for differentiating SCZ, BP, and MDD samples is still a great challenge. In this study, the transcriptomic data was applied for discovering key genes and constructing a classification model. In this dataset, there were 268 samples including four groups (67 SCZ patients, 40 BP patients, 57 MDD patients, and 104 healthy controls), which were applied for constructing a classification model. First, 269 probes of differentially expressed genes (DEGs) among four sample groups were identified by the feature selection method. Second, these DEGs were validated by the literature review including disease relevance with the psychiatric disorders of these DEGs, the hub genes in the PPI (protein–protein interaction) network, and GO (gene ontology) terms and pathways. Third, a classification model was constructed using the identified DEGs by machine learning method to classify different groups. The ROC (receiver operator characteristic) curve and AUC (area under the curve) value were used to assess the classification capacity of the model. In summary, this classification model might provide clues for the diagnoses of these psychiatric disorders.

© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

In psychiatric disorders, schizophrenia (SCZ), bipolar disorder (BP), and major depressive disorder (MDD) are multigenic diseases with complex etiology [1]. These four psychiatric disorders are associated with high rates of morbidity, mortality, and suicide. There were evident differences among these psychiatric disorders. SCZ is a severe mental disorder and can cause delusions and hallucinations [2]. SCZ affects approximately 1 % of the world's population and generally appears in subjects aged 15 to 25 years [3]. BP is known as one disabilities worldwide and is characterized by a high suicide rate, sleep problems, and dysfunction of psychological traits [4]. BP is characterized by alternating episodes of mania interspersed with periods of depression [5]. MDD is the leading cause of disability resulting in the overall burden of disease. MDD is characterized by symptoms and causes emotional distress, functional impairment, and suicide [6].

There are many similar symptoms of these psychiatric disorders such as suicidal ideation, sleep disturbances, and cognitive deficits. The diagnostic boundaries among these psychiatric disorders remain difficult to define because of this similarity. Therefore, psychiatry is the last medicine area because the diagnosis only uses the symptoms due to a lack of biomarkers to assist the diagnosis [7]. Using these biomarkers, underlying molecular pathologies using biomarkers is necessary to address the burden of psychiatric diseases. For psychiatric disorders, developing more effective method for objective diagnoses has been a major international public health priority [8,9].

Identification of molecular measures (biomarkers) will provide insight into the biology underlying the shared symptoms and is beneficial to the diagnosis of psychiatric disorders [10]. To seek objective biomarkers, transcriptomic data has become a powerful

* Corresponding authors.
E-mail addresses: yangqx@njupt.edu.cn (Q. Yang), gongyglab@gmail.com (Y. Gong).

https://doi.org/10.1016/j.csbj.2022.09.014
2001-0370/© 2022 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

technology for detecting gene expression [11]. Recently, there are plenty of studies exploring molecular biomarkers based on transcriptomics [12]. For instance, in the research of Lanz *et al.* [13], the STEP level is unchanged in the pre-frontal cortex and associative striatum of post-mortem human brain samples of SCZ, BP, and MDD subjects. As reported by Higgs *et al.* [14], the database including SCZ, BP, and MDD samples can offer an efficient tool for data mining, such as biomarkers elucidation for target discovery. However, a classification model based on machine learning is still highly necessary and beneficial to the diagnoses of psychiatric disorders.

In this work, one combined dataset including SCZ, BP, MDD, and healthy controls was obtained by integrating three transcriptomic studies. First, there were 268 samples in this dataset including 67 SCZ subjects, 40 BP subjects, 57 MDD subjects, and 104 healthy controls. The differentially expressed genes (DEGs) were discovered by the partial least squares-discriminant analysis (PLS-DA), and 269 probes of DEGs were identified for psychiatric disorders (SCZ, BP, MDD, and healthy controls). Second, these DEGs were validated by the literature review including disease relevance with the psychiatric disorders of these DEGs, the hub genes of the PPI (protein–protein interaction) network, GO (gene ontology) terms, and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. Third, a classification model was constructed applying machine learning method for classifying four groups based on the identified DEGs. Based on the independent set, the AUC (area under the curve) value and the ROC (receiver operator characteristic) curve were used for assessing the classification capacity of this model.

2. Materials and methods

2.1. Transcriptomic dataset for the psychiatric disorders

Based on popular databases including GEO (Gene Expression Omnibus) and SMRI (Stanley Medical Research Institute), the datasets of the prefrontal cortex from the Brodmann Area 9, 10, and 46 in the brain were collected by searching the keywords (schizophrenia, bipolar disorder, and major depressive disorder). As a result, three microarray datasets were used in this study, and each dataset included four sample groups (SCZ, BP, MDD, and healthy controls). As shown in Table 1, detailed information on these datasets was provided, such as dataset ID and the number of samples. The data analysis of the raw data for these datasets was performed using the R language. Herein, these three studies were integrated as a comprehensive dataset by matching the probe ID of the gene. After integration, the batch effects were removed for the comprehensive studies [15]. The *combat* function in the *sva* package was used to remove the batch effects for three different datasets [16]. This comprehensive dataset was used to identify the DEGs among different sample groups of SCZ, BP, MDD, and healthy controls.

2.2. Identifying DEGs for SCZ, BP, MDD, and healthy controls

To identify the DEGs among four sample groups, a popular feature selection algorithm, PLS-DA (partial least squares-discriminant analysis) [17] was applied in this study. PLS-DA was one of the most well-known machine learning methods as a useful feature selector [18]. Recently, PLS-DA was widely applied for identify features for omics data [19]. Because of substantial similarity, the identified DEGs were expected to classify SCZ, BP, MDD, and healthy subjects. PLS-DA can select differential features among multiple classes simultaneously. Herein, the DEGs of four sample groups (SCZ, BP, MDD, and healthy groups) were discovered by the PLS-DA model. The VIP (Variable Importance in the Projection > 2) value in the PLS-DA model was applied as the index for the DEGs. And the dysregulated genes among four groups were identified by the VIP value (>2) of the PLS-DA model [20].

2.3. Functional analysis for the DEGs identified in psychiatric disorders

The functional analysis for the DEGs identified in psychiatric disorders was conducted in this work. The analysis was conducted from three different perspectives, including (1) disease relevance of these DEGs, (2) disease relevance of the hub genes of the PPI network, and (3) disease relevance of the gene ontology terms and pathways. For these DEGs, the disease relevance with psychiatric disorders was surveyed by the literature review. A substantial percentage of the disease-related genes was expected for these psychiatric disorders. But a certain number of psychiatric disorder-unrelated genes was unavoidable because of the measurement variations. The disease relevance was represented by the percentage of disease-related genes among all DEGs.

To ensure the hub genes of psychiatric disorders, the STRING database [21] was used to construct protein–protein interaction (PPI) network. Using high confidence (0.7), the DEGs discovered in this study can be mapped into this PPI network. Cytoscape [22] was used for visualizing the interactions of genes in the PPI network. The hub genes were discovered from all genes with high interaction degrees (score $\geq$ 10) for psychiatric disorders. The role of the hub genes in psychiatric disorders was confirmed using the literature review. Moreover, GSEA was used to conduct the enrichment of GO terms and KEGG pathways by the adjusted *p*-value (<0.05) [23]. The GO terms and KEGG pathways overrepresented were identified, and a comprehensive literature review was conducted to reveal the important role of these terms and pathways in psychiatric disorders.

2.4. Constructing classification model using Machine learning

It remains difficult to define the diagnostic boundaries among psychiatric disorders due to the similarity of symptoms. A classification model with superior performance is important for the diagnoses of SCZ, BP, and MDD samples. Therefore, the DEGs identified in this study were used to construct a model for classifying different groups of psychiatric disorders. A popular machine learning method, support vector machine (SVM), was a supervised technique and was applied for classification. Herein, a classification model applying SVM method was constructed based on the identified DEGs for SCZ, BP, and MDD groups. This classification model was validated using fivefold cross-validation. The AUC value and ROC curve of this model were used to assess the classification capacity. Using the comprehensive dataset, the fivefold cross-

Table 1
The transcriptomic datasets were collected from three studies of psychiatric diseases. No. referred to the number of samples. Each dataset contained one cohort of SCZ (schizophrenia), one cohort of BP (bipolar disorder), one cohort of MDD (major depressive disorder), and one cohort of CTRL (control) samples.

ID	No. (SCZ:BP:MDD: CTRL)	Tissue	Reference
GSE92538	128 (31:12:29:56)	Frontal (BA9/46)	*PLoS One*. 13:e0200003,2018.
Stanley AltarC	72 (21:11:11:29)	Frontal (BA10/46)	*BMC Genomics*. 7:70,2006.
GSE53987	68 (15:17:17:19)	Frontal (BA46)	*PLoS One*. 10:e0121744,2015.

validation was applied in the classification model. The AUC value could quantify the classification capacity of the model to distinguish different classes. If the AUC value was 1, the classification capacity of the model to classify different groups was excellent enough. If the AUC value was 0, the classification capacity of the model was poor enough.

To validate the classification capacity for generalizing to other datasets, the independent set was applied in the constructed SVM model. In this model, the combined dataset (Table 1) was as the training set, and the independent sets (GSE127711 [24] and GSE38484 [25]) were as the test set due to a lack of associated datasets. In the independent discovery cohort of the first dataset (GSE127711), there were 124 SCZ patients, 260 BP patients, and 112 MDD patients in the blood samples. In the second dataset (GSE38484), there were 106 SCZ patients and 96 healthy subjects in human whole blood. To obtain all four groups, these two independent datasets were combined as a new independent set. In this dataset, there were 230 SCZ samples, 260 BP samples,112 MDD samples, and 96 healthy samples. The gene expression of the comprehensive dataset for identifying DEGs was detected in the prefrontal cortex of the brain, and the gene expression of the independent set was detected in the blood samples. To generalize the model constructed in this study, the blood samples in the independent set were applied to measure the classification capacity.

3. Results and discussion

3.1. Comprehensive dataset including SCZ, BP, MDD, and healthy groups

As shown in Fig. 1, the flowchart of this study included four parts: (1) the comprehensive transcriptomic dataset; (2) identification of DEGs by PLS-DA; (3) functional analysis; and (4) construction of the classification model. At the beginning of this study, three datasets (Table 1) were collected for the comprehensive transcriptomic dataset. One dataset (Stanley AltarC) was from the SMRI database [14] including 72 (21 SCZ, 11 BP, 11 MDD, and 33 healthy subjects) samples detected by the HG-U133A platform. For dataset GSE92538 [26] from the GEO database, 128 samples (31 SCZ, 12 BP, 29 MDD, and 56 healthy controls) were detected by HG-U133 Plus 2 platform. For dataset GSE53987 [13] from the GEO database, 68 samples (15 SCZ, 17 BP, 17 MDD, and 19 healthy controls) were detected by HG-U133 Plus 2 platform. After each dataset was processed and analyzed using the R language, the comprehensive dataset was combined by removing batch effects. In this comprehensive dataset by combining these three datasets, there were 22,277 probes of genes and 268 samples of prefrontal cortex including 67 SCZ patients, 40 BP patients, 57 MDD patients, and 104 healthy subjects.

3.2. DEGs identified for psychiatric disorders using the comprehensive dataset

DEGs were discovered by the PLS-DA method to classify different groups of psychiatric disorders simultaneously based on the comprehensive transcriptomic data. Using the cutoff of VIP value ($\geq$2) of the PLS-DA model, there were 269 probes of DEGs identified in this study (as shown in Supplementary Figure S1). As shown in Table 2, detailed information on the top 20 DEGs with the highest VIP values was provided. The dysregulated information of all DEGs between two groups (including between SCZ and BP, between SCZ and MDD, as well as between BP and MDD) was shown in Supplementary Table S1. As demonstrated in Fig. 2, the boxplots were applied to visualize and compare the differential expression of the top 9 DEGs among four groups directly. For example, the gene expression of NEK1 with the highest VIP value (VIP = 3.00) has a strong association with a chromosome 4 genetic locus identified as significantly associated with SCZ [27]. It showed an increase in NEK1 after antidepressant treatment in responders [28]. Moreover, it was reported that the expression of CDC42BPA with the second highest VIP value (VIP = 2.95) differed significantly among SCZ, BP, and controls [29].

3.3. Functional analysis for DEGs identified among multiple psychiatric disorders

The functional analysis of the DEGs was performed from three different perspectives including (1) disease relevance for the DEGs,

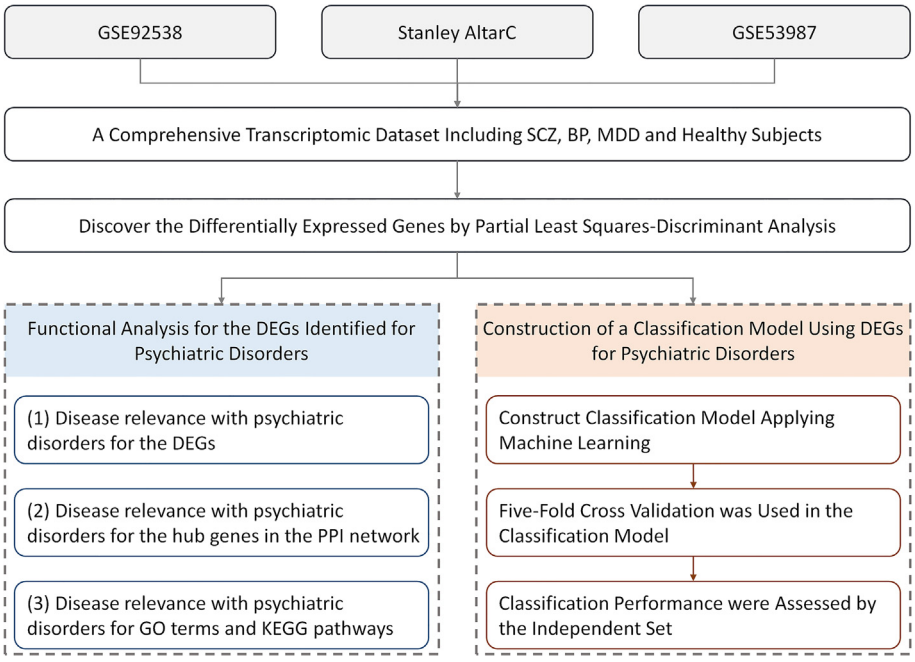

Fig. 1. The detailed information of the flowchart in this study. SCZ: schizophrenia, BP: bipolar disorder, MDD: major depressive disorder, DEGs: differentially expressed genes, ROC: receiver operator characteristic, AUC: area under the curve.

Table 2
Detailed information on the top 20 DEGs identified by the PLS-DA (partial least squares discriminant analysis) method with the cutoff of Variable Importance in the Projection (VIP > 2). SCZ: schizophrenia, BP: bipolar disorder, and MDD: major depressive disorder.

Order	Probe ID	Entrez ID	Symbol	VIP	Up-or Down -Regulated		
					SCZ *vs* BP	SCZ *vs* MDD	BP *vs* MDD
1	213328_at	4750	NEK1	3.00	Down	Down	Up
2	214464_at	8476	CDC42BPA	2.95	Down	Down	Up
3	205472_s_at	1602	DACH1	2.92	Down	Down	Down
4	208425_s_at	26,115	TANC2	2.84	Down	Down	Down
5	202905_x_at	4683	NBN	2.84	Down	Down	Up
6	208993_s_at	9360	PPIG	2.84	Down	Down	Up
7	219437_s_at	29,123	ANKRD11	2.78	Down	Down	Up
8	212079_s_at	4297	KMT2A	2.78	Down	Down	Up
9	208003_s_at	10,725	NFAT5	2.76	Down	Down	Up
10	213850_s_at	9169	SCAF11	2.75	Down	Down	Up
11	210479_s_at	6095	RORA	2.74	Down	Down	Up
12	213638_at	221,692	PHACTR1	2.74	Down	Down	Down
13	212758_s_at	6935	ZEB1	2.74	Down	Down	Up
14	212650_at	23,301	EHBP1	2.74	Down	Down	Down
15	220462_at	80,034	CSRNP3	2.72	Down	Down	Down
16	202040_s_at	5927	KDM5A	2.72	Down	Down	Up
17	209945_s_at	2932	GSK3B	2.72	Down	Down	Up
18	201996_s_at	23,013	SPEN	2.71	Down	Down	Up
19	220940_at	57,730	ANKRD36B	2.67	Down	Down	Up
20	209376_x_at	9169	SCAF11	2.67	Down	Down	Up

(2) disease relevance for the hub genes of the PPI network, and (3) disease relevance for the enriched GO terms and KEGG pathways.

(1) Disease relevance with psychiatric disorders for the DEGs.

To evaluate the disease relevance of the DEGs discovered among multiple disorders, the top 20 DEGs among different groups were surveyed by the comprehensive literature review. The disease relevance between each DEG and psychiatric disorders (SCZ, BP, MDD, or cognition) was described in Supplementary Table S2. A great disease relevance (90 %) of the top 20 DEGs was verified. For these DEGs, it was reported that DACH1 was a transcription factor acting as a neurogenic cell-fate determining factor [30]. The mutations of TANC2 were associated with both pediatric neurodevelopmental and adult neuropsychiatric disease [31]. ANKRD11 was a nuclear coregulator in the developing brain, which determined precursor proliferation, neurogenesis, and neuronal positioning [32]. It was reported that KMT2A, NFAT5, SCAF11, and GSK3B were upregulated in neurons of BP [33]. Several genetic variants of RORA were associated with BP [34], and the polymorphisms of RORA were associated with risk for various forms of psychopathology including BP and MDD [35]. PHACTR1 showed the association with SCZ in the combined analysis and the locus was located in an SCZ linkage region [36]. ZEB1 was an element of a common pathway involved in SCZ [37]. EHBP1 was downregulated in the medial prefrontal cortex of adult SHANK3-overexpressing mice, and variants of SHANK3 were causally associated with numerous neurodevelopmental and neuropsychiatric disorders including BP and SCZ [38]. CSRNP3 was a mapped gene of 2q24.3 and genome-wide significant loci associated with BP [39]. KDM5A was one of the best candidates for explaining epilepsy, intellectual disability, and SCZ [40]. Seven risk genes (CTCF, HNRNPU, KCNQ3, ZBTB18, TCF12, SPEN, and LEO1) were associated with neurodevelopmental disorders based on the large-scale targeted sequencing [41].

(2) Disease relevance with psychiatric disorders for the hub genes in the PPI network.

STRING database was used for constructing the PPI network [42], and the hub genes were discovered based on the *CytoHubba* [43] of *Cytoscape* [22]. As shown in Fig. 3A, the PPI network for all DEGs was constructed. The degree of nodes in this PPI network was shown in Supplementary Table S3. As shown in Fig. 3B, the top 13 nodes with the highest score ($\geq$10) of the network using the MCC algorithm on *CytoHubba* were marked with red and yellow colors. The intersection between the top 13 genes by the MCC algorithm (as shown in Fig. 3C) and the top 20 genes with high degree ($\geq$10) in this PPI network was regarded as the hub genes. There were 9 hub genes including ESF1, PAK1IP1, SF3B1, RBM25, KRAS, SRRM2, CAMK2G, PIK3R1, and PRPF40A. As shown in Fig. 4, the 9 hub genes were validated to confirm the differential expression among four groups (SCZ group, BP group, MDD group, and healthy group) using the boxplots. From these boxplots, there were significant changes for these DEGs.

As shown in Supplementary Table S4, a great disease relevance (78 %) for the 9 hub genes was discovered between the hub genes and psychiatric disorders by a literature review. It was reported that SF3B1 was associated with SCZ and neurodevelopmental disorders in the largest SCZ genome-wide association study [44]. KRAS mutations were associated with depression severity and higher rates of probable depression in patients with metastatic colorectal cancer [45]. A mechanistic pathway involving CAMK2G was reported in stress and the trauma-related manifestation of anxiety and depression across species [46]. The interaction effects of the polymorphisms in hsa-miR-219, CAKM2G, GRIN2B, and GRIN3A might confer susceptibility to SCZ in the Chinese Han population [47]. PIK3R1 was the shared susceptibility gene for SCZ and BP, which might be a potential diagnostic biomarker for BP [48]. PIK3R1 and PRPF40A were identified as the hub genes in the anterior cingulate cortex regions of the brain for MDD [49]. Therefore, these hub genes discovered using the PPI network had an important role in SCZ, BP, and MDD, which showed the reliability of the DEGs discovered in this work.

(3) Disease relevance with psychiatric disorders for the Enriched GO terms and KEGG pathways.

Moreover, 33 KEGG pathways have been enriched using the DEGs discovered in this study (as shown in Fig. 5A and Supplementary Table S5), including regulation of actin cytoskeleton, neurotrophin signaling pathway, focal adhesion, calcium signaling pathway, and insulin signaling pathway. The regulation of the actin cytoskeleton was likely to be shared between SCZ and BP [50]. Rare variants in the neurotrophin signaling pathway were implicated in SCZ risk [51]. The evidence for altered motility and focal adhesion dynamics was consistent with dysregulated gene expression in the FAK signaling pathway. Alterations in cell adhesion dynamics and cell motility can affect the trajectory of brain development in SCZ [52]. A detailed characterization of the risk loci showed that cal-

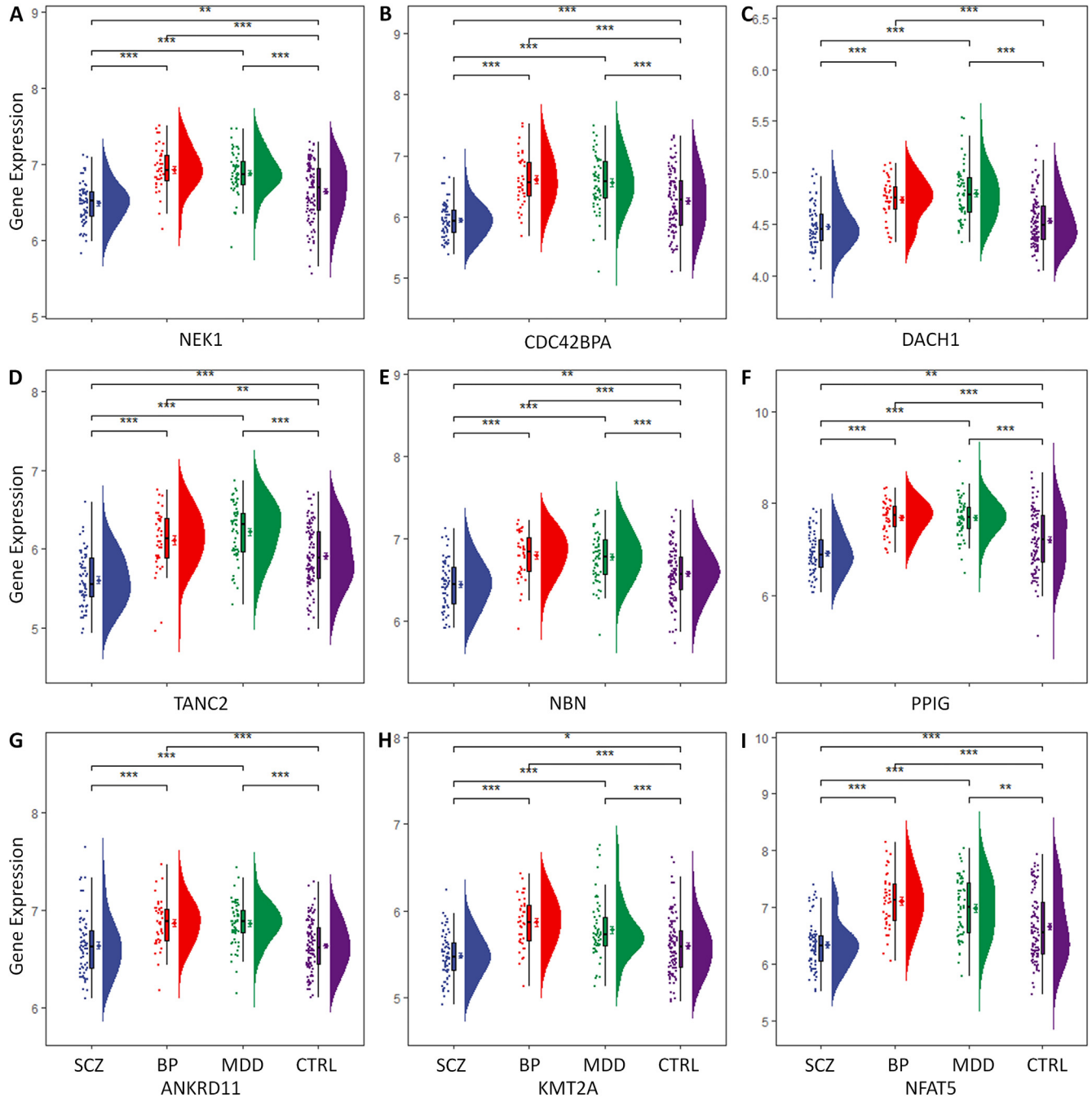

Fig. 2. The boxplots of the top 9 DEGs with the highest VIP (Variable Importance in the Projection) values were applied to visualize the differential expression in different groups. The blue, red, green, and purple indicated the SCZ (schizophrenia), BP (bipolar disorder), MDD (major depressive disorder), and CTRL (healthy controls), respectively. Statistically significant differences in cortical thickness: $^{*}p < 0.05$, $^{**}p < 0.001$, $^{***}p < 0.0001$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

cium signaling pathway genes might play pivotal roles in SCZ [53], and the downregulated signaling pathways in depression mice included the calcium signaling pathway [54]. It was suggested that there existed abnormalities of the insulin signaling pathway in SCZ and that antipsychotic drug effects on this pathway were therapeutic in SCZ [55].

As shown in Fig. 5B and Supplementary Table S6, the enrichment analysis of GO terms was performed to discover the biological processes (BP) terms. For instance, adult neurogenesis concerning regulatory signaling molecules would be helpful to identify how abnormalities might contribute to the pathophysiology of SCZ [56], and altered adult neurogenesis was postulated as an aetiological mechanism for BP [57]. As demonstrated in Fig. 5C and Supplementary Table S7, the molecular functions (MF) terms were enriched using DEGs. Such as, the Alu element in the RNA binding motif protein (RBMX2) was found to be linked to BP [58]. As demonstrated in Fig. 5D and Supplementary Table S8, a lot of key cell components (CC) terms were enriched using the DEGs in this study. And a growing body of evidence connected a dysfunctional microtubule cytoskeleton with neuropsychiatric ill-

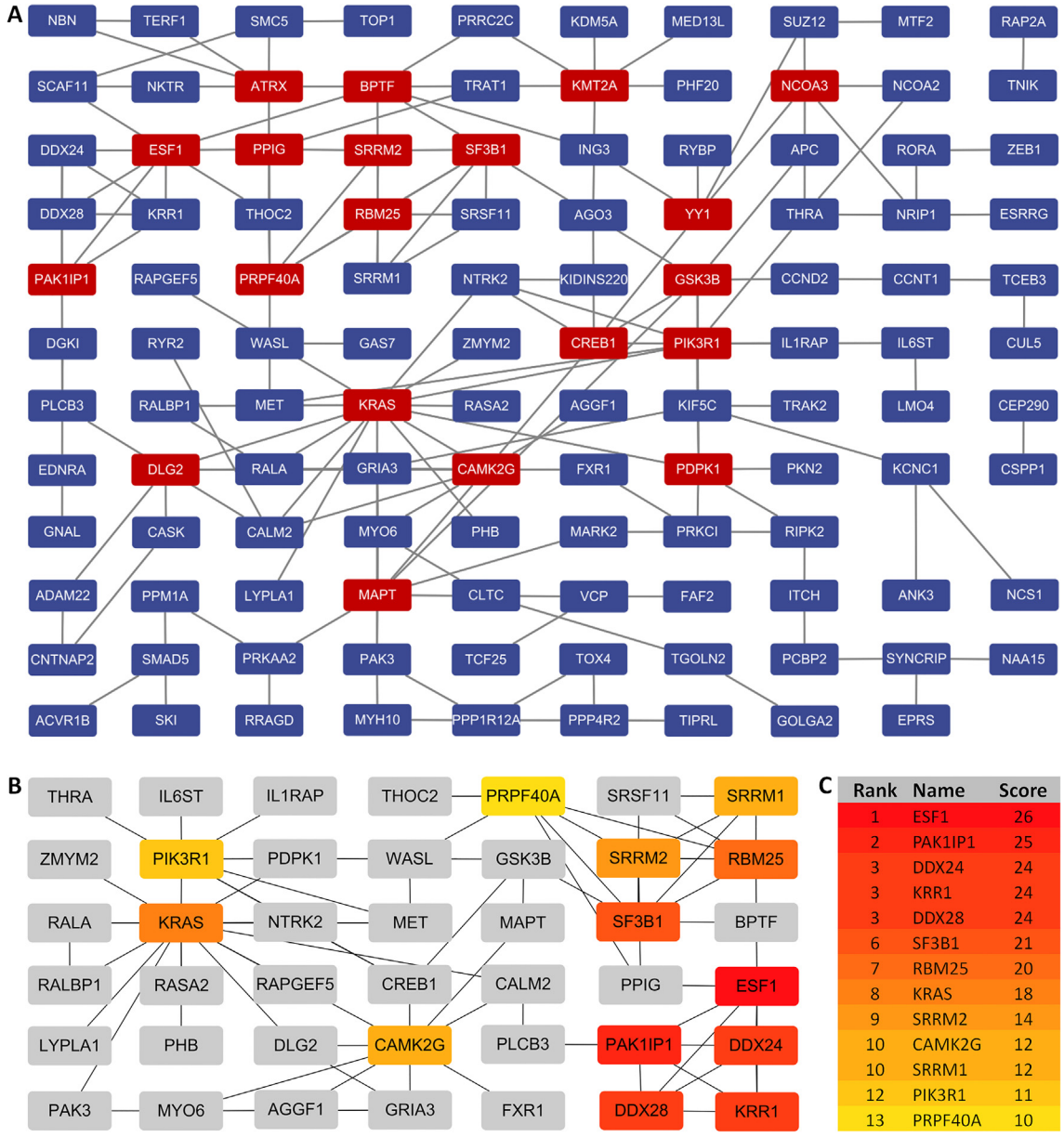

Fig. 3. (A) The PPI network was constructed using DEGs (differentially expressed genes) among schizophrenia, bipolar disorder, major depressive disorder, and healthy controls. (B) The top 13 hub nodes with the highest MCC (score $\geq$ 10) in the network were marked with red and yellow colors using the MCC algorithm on *CytoHubba*. (C) The scores for the top 13 hub nodes were ranked by the MCC. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

nesses [59]. Using the literature review, the GO terms and KEGG pathways enriched were validated that they played an important role in the development of psychiatric disorders.

3.4. Constructing the classification model for multiple psychiatric disorders

As one of the supervised machine learning algorithms, SVM can be used to construct a classification model. The classification of SVM can be applied for two or more classes using the *e1071* package. A single SVM does binary classification and can classify samples between two classes. SVM can be applied for classifying multiple groups using the One-to-Rest approach. To classify multiple classes, each binary classifier is set to per each class. In this approach, the classifier can use *m* SVM models and each model will predict membership in one of the *m* classes. In this study, the SVM method was used to construct a model for classifying multiple groups (SCZ, BP, MDD, and healthy controls). The classification model was constructed for classifying samples of SCZ, BP, MDD, and healthy groups based on the DEGs identified by the PLS-DA method using the comprehensive dataset (Table 1). Because it was hard to obtain good performance when using all genes due to the interference of the irrelevant genes, these DEGs differential among four groups were applied for constructing well-performed classification model. In the multi-class classification models, there were four SVM models for SCZ, BP, MDD, and healthy groups. And the total model was obtained using the micro value of all SVM models. For the combined dataset (Table 1), the performance of the classification model was assessed by 5-fold cross-validation. The AUC value and ROC curve were used to assess the classification

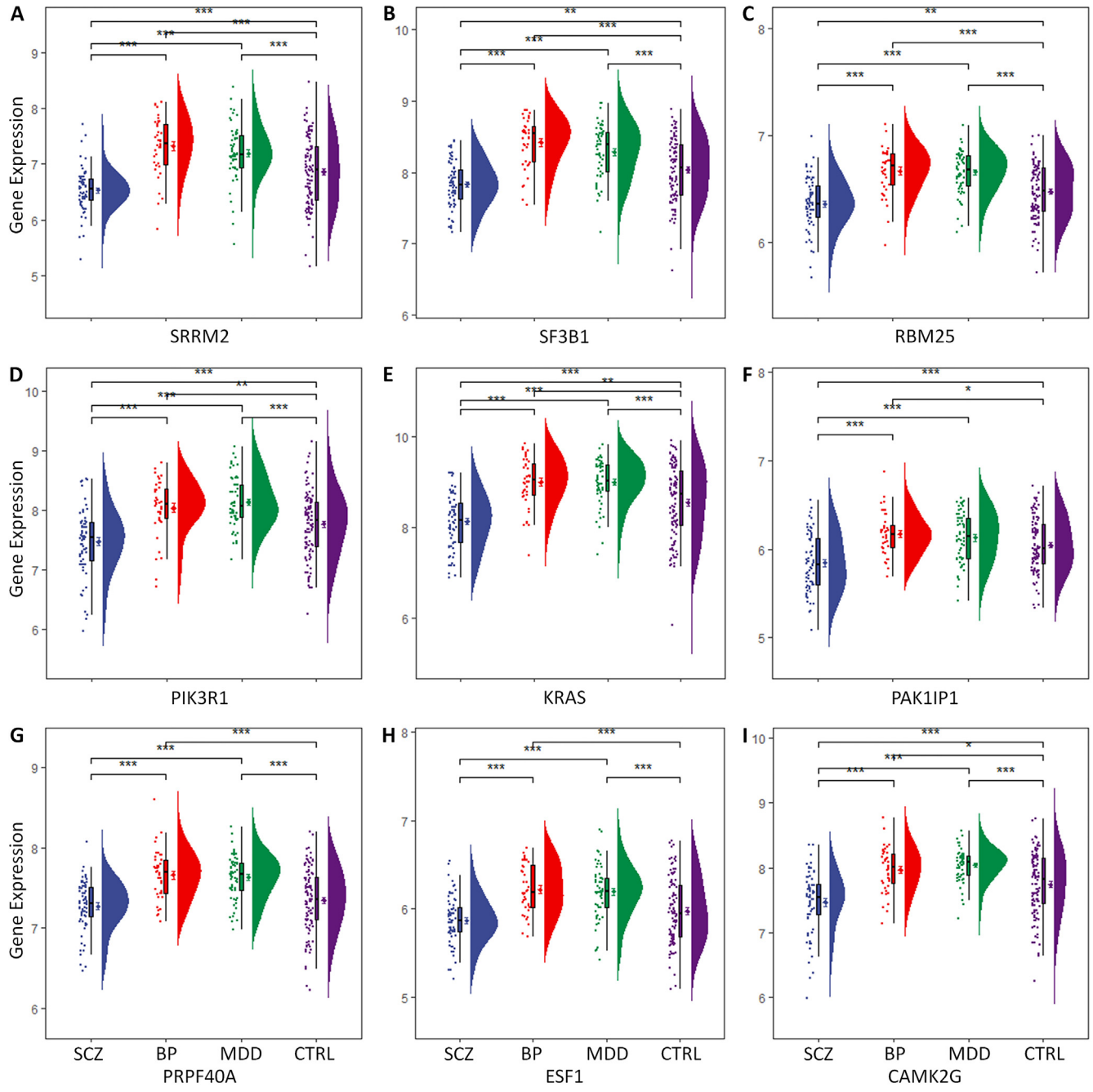

Fig. 4. The boxplots of the 9 hub genes of the PPI network using the intersection between the genes with the highest degree (score ≥ 10) of the PPI network and the top 13 hub nodes ranked by the MCC (score ≥ 10) in the network using the MCC algorithm on CytoHubba software. The blue, red, green, and purple indicated the SCZ (schizophrenia), BP (bipolar disorder), MDD (major depressive disorder), and CTRL (healthy controls), respectively. Statistically significant differences in cortical thickness: $^{*}p < 0.05$, $^{**}p < 0.001$, $^{***}p < 0.0001$. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

performance. Based on the comprehensive dataset, the AUC values and ROC curves are interpreted for SCZ groups (Fig. 6A), BP groups (Fig. 6B), MDD groups (Fig. 6C), healthy groups (Fig. 6D), and (Fig. 6E) total micro value for all groups using the 5-fold cross-validation. Overall, the AUC value of 5-fold cross-validation for four groups was 0.94 in the SVM model using the comprehensive dataset.

In this study, the independent dataset was applied to generalize the constructed SVM model. The combined dataset (Table 1) was regarded as the training set, and the independent set by combining GSE127711 and GSE38484 was regarded as the test set. In this classification, the micro value was calculated for four groups of psychiatric disorders (SCZ, BP, MDD, and healthy controls). The AUC value of the independent set was 0.71 in the classification model. As shown in Fig. 6F, the AUC value and ROC curve were used to assess the performance of model using the independent set (Table 1). From the results, the classification performance is only good (AUC > 0.7) for classifying four groups simultaneously. The genes of the training set were detected in the prefrontal cortex, and the genes of the independent set were detected in the blood samples.

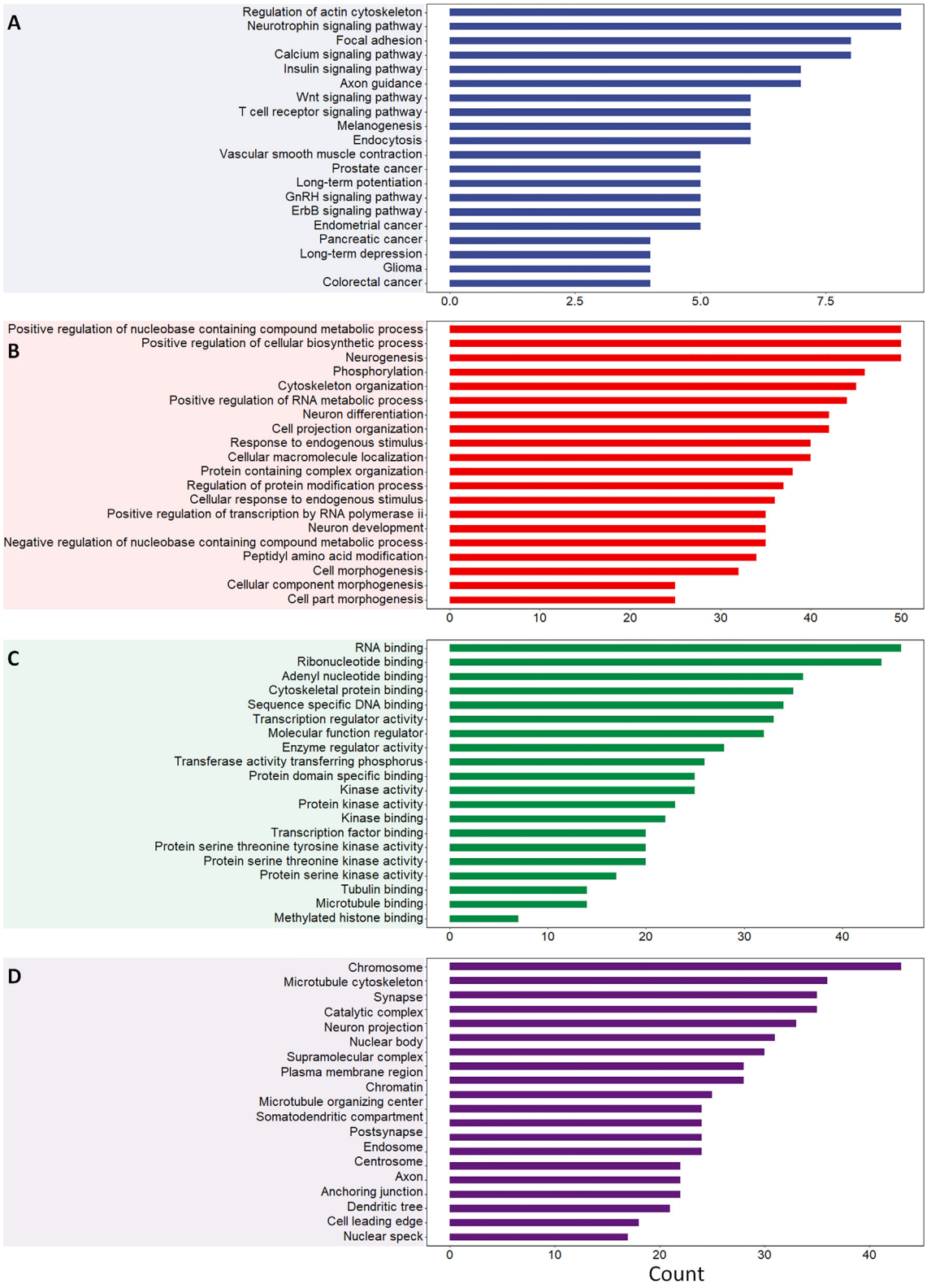

Fig. 5. The enrichment analysis was performed using differentially expressed genes. The top 20 terms of (A) biological processes, (B) molecular functions, and (C) cell components of GO (gene ontology) enrichment. (D) The top 20 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways were enriched in this study.

Because of the differences in the data type between the training set and test set, it is very difficult to obtain superior performance for the classification capacity by the independent test. In the future, the classification model with superior performance can be developed using other machine learning methods, which will be helpful for the diagnoses of psychiatric disorders.

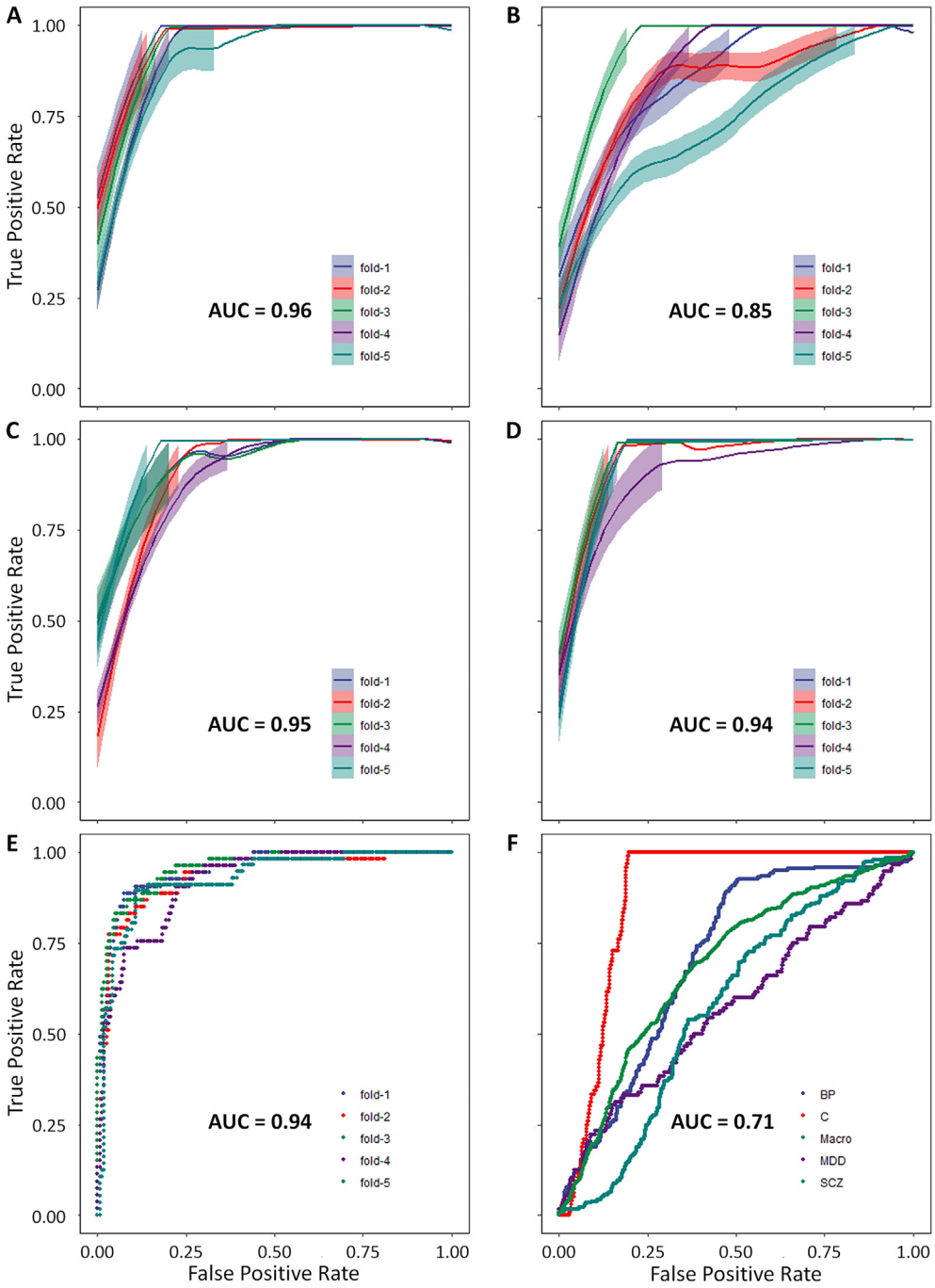

Fig. 6. The classification model was constructed for psychiatric disorders including schizophrenia, bipolar disorder, major depressive disorder, and healthy controls using machine learning. Based on the combined dataset, the ROC curves and AUC values for (A) SCZ groups, (B) BP groups, (C) MDD groups, (D) healthy groups, and (E) total micro value for all groups was obtained using the fivefold cross-validation. (F) the ROC curve and AUC value for the independent set for the classification model.

4. Conclusions

In this work, a combined dataset comprising 67 SCZ patients, 40 BP patients, 57 MDD patients, and 104 healthy controls was collected. First, 269 probes of DEGs were discovered based on the PLS-DA method to classify the samples into four groups. Second, these DEGs were validated by the literature review including disease relevance with the psychiatric disorders of these DEGs, the hub genes of the PPI network, and enriched GO terms and KEGG pathways. Third, a classification model was constructed by machine learning method using the DEGs identified in four groups. By ROC curve and AUC value, a strong capacity to classify samples among multiple groups was demonstrated. Moreover, the constructed SVM model was generalized using the independent set. In sum, the classification model constructed might provide clues for the diagnoses of these psychiatric disorders.

CRediT authorship contribution statement

Qingxia Yang: Conceptualization, Methodology, Software, Writing – original draft. **Qiaowen Xing:** Visualization, Investigation. **Qingfang Yang:** Software, Validation. **Yaguo Gong:** Data curation, Writing – review & editing.

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.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (62201289), the National Natural Science Foundation of Jiangsu (BK20210597), and the NUPTSF (Grant No. NY220169).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.09.014.

References:

[1] McGuinness AJ, Davis JA, Dawson SL, Loughman A, Collier F, O'Hely M, et al. A systematic review of gut microbiota composition in observational studies of major depressive disorder, bipolar disorder and schizophrenia. Mol Psychiatry 2022;27(4):1920–35.
[2] Jauhar S, Johnstone M, McKenna PJ. Schizophrenia. Lancet 2022;399 (10323):473–86.
[3] Bighelli I, Rodolico A, Garcia-Mieres H, Pitschel-Walz G, Hansen WP, Schneider-Thoma J, et al. Psychosocial and psychological interventions for relapse prevention in schizophrenia: a systematic review and network meta-analysis. Lancet Psychiatry 2021;8(11):969–80.
[4] Rantala MJ, Luoto S, Borraz-Leon JI, Krams I. Bipolar disorder: An evolutionary psychoneuroimmunological approach. Neurosci Biobehav Rev 2021;122:28–37.
[5] Zhang C, Xiao X, Li T, Li M. Translational genomics and beyond in bipolar disorder. Mol Psychiatry 2021;26(1):186–202.
[6] McCarron RM, Shapiro B, Rawles J, Luo J. Depression. Ann Intern Med 2021;174 (5):65–80.
[7] Wolfers T, Doan NT, Kaufmann T, Alnaes D, Moberget T, Agartz I, et al. Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. JAMA Psychiatry 2018;75(11):1146–55.
[8] Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science 2018;362(6420):8127.
[9] Hebert M, Merette C, Gagne AM, Paccalet T, Moreau I, Lavoie J, et al. The electroretinogram may differentiate schizophrenia from bipolar disorder. Biol Psychiatry 2020;87(3):263–70.
[10] Ruderfer DM, Fanous AH, Ripke S, McQuillin A, Amdur RL, et al. Polygenic dissection of diagnosis and clinical dimensions of bipolar disorder and schizophrenia. Mol Psychiatry 2014;19(9):1017–24.
[11] Hoseth EZ, Krull F, Dieset I, Morch RH, Hope S, Gardsjord ES, et al. Exploring the Wnt signaling pathway in schizophrenia and bipolar disorder. Transl Psychiatry 2018;8(1):55.
[12] Birnbaum R, Weinberger DR. Genetic insights into the neurodevelopmental origins of schizophrenia. Nat Rev Neurosci 2017;18(12):727–40.
[13] Lanz TA, Joshi JJ, Reinhart V, Johnson K, Grantham 2nd LE, Volfson D. STEP levels are unchanged in pre-frontal cortex and associative striatum in post-mortem human brain samples from subjects with schizophrenia, bipolar disorder and major depressive disorder. PLoS ONE 2015;10(3):e0121744.
[14] Higgs BW, Elashoff M, Richman S, Barci B. An online database for brain disease research. BMC Genomics 2006;7:70.
[15] Yang Q, Gong Y. Construction of the classification model using key genes identified between benign and malignant thyroid nodules from comprehensive transcriptomic data. Front Genet 2022;12:791349.
[16] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 2007;8 (1):118–27.
[17] Lee LC, Liong CY, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst 2018;143(15):3526–39.
[18] Le Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinf 2011;12:253.
[19] Yang Q, Li Y, Li B, Gong Y. A novel multi-class classification model for schizophrenia, bipolar disorder and healthy controls using comprehensive transcriptomic data. Comput Biol Med 2022;148:105956.
[20] Belmonte-Sanchez JR, Romero-Gonzalez R, Arrebola FJ, Vidal JLM, Garrido FA. An innovative metabolomic approach for golden rum classification combining ultrahigh-performance liquid chromatography-orbitrap mass spectrometry and chemometric strategies. J Agric Food Chem 2019;67(4):1302–11.
[21] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43(D1):447–52.
[22] Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504.
[23] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 2005;102(43):15545–50.
[24] Niculescu AB, Le-Niculescu H, Roseberry K, Wang S, Hart J, Kaur A, et al. Blood biomarkers for memory: toward early detection of risk for Alzheimer disease, pharmacogenomics, and repurposed drugs. Mol Psychiatry 2020;25 (8):1651–72.
[25] van Eijk KR, de Jong S, Strengman E, Buizer-Voskamp JE, Kahn RS, Boks MP, et al. Identification of schizophrenia-associated loci by combining DNA methylation and gene expression data from whole blood. Eur J Hum Genet 2015;23(8):1106–10.
[26] Hagenauer MH, Schulmann A, Li JZ, Vawter MP, Walsh DM, Thompson RC, et al. Inference of cell type content from human brain transcriptomic datasets illuminates the effects of age, manner of death, dissection, and psychiatric diagnosis. PLoS ONE 2018;13(7):e0200003.
[27] Lehner T, Miller BL, State MW. Genomics, circuits, and pathways in clinical neuropsychiatry, 2016;doi:10.1016/C2013-0-13583-0.
[28] Turck CW, Guest PC, Maccarrone G, Ising M, Kloiber S, Lucae S, et al. Proteomic differences in blood plasma associated with antidepressant treatment response. Front Mol Neurosci 2017;10:272.
[29] Konopaske GT, Balu DT, Presti KT, Chan G, Benes FM, Coyle JT. Dysbindin-1 contributes to prefrontal cortical dendritic arbor pathology in schizophrenia. Schizophr Res 2018;201:270–7.
[30] Schormair B, Zhao C, Bell S, Tilch E, Salminen AV, Putz B, et al. Identification of novel risk loci for restless legs syndrome in genome-wide association studies in individuals of European ancestry: a meta-analysis. Lancet Neurol 2017;16 (11):898–907.
[31] Guo H, Bettella E, Marcogliese PC, Zhao R, Andrews JC, Nowakowski TJ, et al. Disruptive mutations in TANC2 define a neurodevelopmental syndrome associated with psychiatric disorders. Nat Commun 2019;10(1):4679.
[32] Gallagher D, Voronova A, Zander MA, Cancino GI, Bramall A, Krause MP, et al. Ankrd11 is a chromatin regulator involved in autism that is essential for neural development. Dev Cell 2015;32(1):31–42.
[33] Kim KH, Liu J, Sells Galvin RJ, Dage JL, Egeland JA, Smith RC, et al. Transcriptomic analysis of induced pluripotent stem cells derived from patients with bipolar disorder from an old order amish pedigree. PLoS ONE 2015;10(11):e0142693.
[34] Lai YC, Kao CF, Lu ML, Chen HC, Chen PY, Chen CH, et al. Investigation of associations between NR1D1, RORA and RORB genes and bipolar disorder. PLoS ONE 2015;10(3):e0121245.
[35] Amstadter AB, Sumner JA, Acierno R, Ruggiero KJ, Koenen KC, Kilpatrick DG, et al. Support for association of RORA variant and post traumatic stress symptoms in a population-based study of hurricane exposed adults. Mol Psychiatry 2013;18(11):1148–9.
[36] Athanasiu L, Mattingsdal M, Kahler AK, Brown A, Gustafsson O, Agartz I, et al. Gene variants associated with schizophrenia in a Norwegian genome-wide study are replicated in a large European cohort. J Psychiatr Res 2010;44 (12):748–53.
[37] Borglum AD, Demontis D, Grove J, Pallesen J, Hollegaard MV, Pedersen CB, et al. Genome-wide study of association and interaction with maternal cytomegalovirus infection suggests new schizophrenia loci. Mol Psychiatry 2014;19(3):325–33.
[38] Jin C, Kang H, Ryu JR, Kim S, Zhang Y, Lee Y, et al. Integrative brain transcriptome analysis reveals region-specific and broad molecular changes in shank3-overexpressing mice. Front Mol Neurosci 2018;11:250.
[39] Gordovez FJA, McMahon FJ. The genetics of bipolar disorder. Mol Psychiatry 2020;25(3):544–59.
[40] Han JY, Park J. Variable phenotypes of epilepsy, intellectual disability, and schizophrenia caused by 12p13.33-p13.32 terminal microdeletion in a korean family: a case report and literature review. Genes (Basel) 2021;12(7):1001.
[41] Wang T, Hoekzema K, Vecchio D, Wu H, Sulovari A, Coe BP, et al. Large-scale targeted sequencing identifies risk genes for neurodevelopmental disorders. Nat Commun 2020;11(1):4932.
[42] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021;49(D1):605–12.
[43] Chin CH, Chen SH, Wu HH, Ho CW, Ko MT, Lin CY. cytoHubba: identifying hub objects and sub-networks from complex interactome. BMC Syst Biol 2014;8 (S4):S11.
[44] Ingason A, Giegling I, Hartmann AM, Genius J, Konte B, Friedl M, et al. Expression analysis in a rat psychosis model identifies novel candidate genes validated in a large case-control sample of schizophrenia. Transl Psychiatry 2015;5:e656.
[45] Zhou Y, Gu X, Wen F, Chen J, Wei W, Zhang ZH, et al. Association of KRAS gene mutations with depression in older metastatic colorectal cancer patients. Int Psychogeriatr 2016;28(12):2019–28.

[46] Wingo AP, Velasco ER, Florido A, Lori A, Choi DC, Jovanovic T, et al. Expression of the PPM1F gene is regulated by stress and associated eith anxiety and depression. Biol Psychiatry 2018;83(3):284–95.
[47] Zhang Y, Fan M, Wang Q, He G, Fu Y, Li H, et al. Polymorphisms in microRNA genes and genes involving in NMDAR signaling and schizophrenia: a case-control study in Chinese Han population. Sci Rep 2015;5:12984.
[48] Huang J, Chen Z, Zhu L, Wu X, Guo X, Yang J, et al. Phosphoinositide-3-kinase regulatory subunit 1 gene polymorphisms are associated with schizophrenia and bipolar disorder in the Han Chinese population. Metab Brain Dis 2020;35(5):785–92.
[49] Wei Y, Qi K, Yu Y, Lu W, Xu W, Yang C, et al. Analysis of differentially expressed genes in the dentate gyrus and anterior cingulate cortex in a mouse model of depression. Biomed Res Int 2021;2021:5013565.
[50] Zhao Z, Xu J, Chen J, Kim S, Reimers M, Bacanu SA, et al. Transcriptome sequencing and genome-wide association analyses reveal lysosomal function and actin cytoskeleton remodeling in schizophrenia and bipolar disorder. Mol Psychiatry 2015;20(5):563–72.
[51] Kranz TM, Goetz RR, Walsh-Messinger J, Goetz D, Antonius D, Dolgalev I, et al. Rare variants in the neurotrophin signaling pathway implicated in schizophrenia risk. Schizophr Res 2015;168(1–2):421–8.
[52] Fan Y, Abrahamsen G, Mills R, Calderon CC, Tee JY, Leyton L, et al. Focal adhesion dynamics are altered in schizophrenia. Biol Psychiatry 2013;74(6):418–26.
[53] Xie Y, Huang D, Wei L, Luo XJ. Further evidence for the genetic association between CACNA1I and schizophrenia. Hereditas 2018;155:16.
[54] Si Y, Song Z, Sun X, Wang JH. microRNA and mRNA profiles in nucleus accumbens underlying depression versus resilience in response to chronic stress. Am J Med Genet B Neuropsychiatr Genet 2018;177(6):563–79.
[55] Girgis RR, Javitch JA, Lieberman JA. Antipsychotic drug mechanisms: links between therapeutic effects, metabolic side effects and the insulin signaling pathway. Mol Psychiatry 2008;13(10):918–29.
[56] Weissleder C, North HF, Shannon WC. Important unanswered questions about adult neurogenesis in schizophrenia. Curr Opin Psychiatry 2019;32(3):170–8.
[57] Cinar RK. Neuroserpin in bipolar disorder. Curr Top Med Chem 2020;20(7):518–23.
[58] Laine P, Rowell WJ, Paulin L, Kujawa S, Raterman D, Mayhew G, et al. Alu element in the RNA binding motif protein, X-linked 2 (RBMX2) gene found to be linked to bipolar disorder. PLoS ONE 2021;16(12):e0261170.
[59] Marchisella F, Coffey ET, Hollos P. Microtubule and microtubule associated protein anomalies in psychiatric disease. Cytoskeleton (Hoboken) 2016;73(10):596–611.