


Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data

Qingxia Yang, Bo Li, Jing Tang, Xuejiao Cui, Yunxia Wang, Xiaofeng Li, Jie Hu, Yuzong Chen, Weiwei Xue, Yan Lou, Yunqing Qiu and Feng Zhu 

Corresponding authors: Feng Zhu, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China. Tel.: +86-571-88208444. E-mail: zhufeng@zju.edu.cn; Yunqing Qiu, The First Affiliated Hospital, Zhejiang University, Hangzhou, Zhejiang 310000, China. Tel.: +86-571-88236626. E-mail: qiuyq@zju.edu.cn

Abstract

The etiology of schizophrenia (SCZ) is regarded as one of the most fundamental puzzles in current medical research, and its diagnosis is limited by the lack of objective molecular criteria. Although plenty of studies were conducted, SCZ gene signatures identified by these independent studies are found highly inconsistent. As one of the most important factors contributing to this inconsistency, the feature selection methods used currently do not fully consider the reproducibility among the signatures discovered from different datasets. Therefore, it is crucial to develop new bioinformatics tools of novel strategy for ensuring a stable discovery of gene signature for SCZ. In this study, a novel feature selection strategy (1) integrating repeated random sampling with consensus scoring and (2) evaluating the consistency of gene rank among different datasets was constructed. By systematically assessing the identified SCZ signature comprising 135 differentially expressed genes, this newly constructed strategy demonstrated significantly enhanced stability and better differentiating ability compared with the feature selection methods popular in current SCZ research. Based on a first-ever assessment on methods' reproducibility cross-validated by independent datasets from three representative studies, the new strategy stood out among the popular methods by showing superior stability and differentiating ability. Finally, 2 novel and 17 previously reported transcription factors were identified and showed great potential in revealing the etiology of SCZ. In sum, the SCZ signature identified in this study would provide valuable clues for discovering diagnostic molecules and potential targets for SCZ.

Key words: schizophrenia; consistent gene signature; feature selection strategy; transcriptomics; combined analysis

Qingxia Yang, Bo Li, Jing Tang, Xuejiao Cui, Yunxia Wang and Xiaofeng Li are PhD/master candidates of the College of Pharmaceutical Sciences in Zhejiang University, China, and jointly cultivated by the School of Pharmaceutical Sciences in Chongqing University, China. They are interested in the area of bioinformatics.

Yuzong Chen is a professor of National University of Singapore. Jie Hu and Yan Lou are professors of Zhejiang University, China. Weiwei Xue is a professor of the School of Pharmaceutical Sciences in Chongqing University, China. They are interested in the area of system biology and bioinformatics.

Yunqing Qiu is a professor of the First Affiliated Hospital in Zhejiang University, China. He is interested in the area of precision medicine, diagnosis and treatment of liver disease and system biology.

Feng Zhu is a professor of the College of Pharmaceutical Sciences in Zhejiang University, China. His research laboratory (<https://idrblab.org/>) has been working in the fields of bioinformatics, OMIC-based drug discovery, system biology and medicinal chemistry. Welcome to visit his personal website at: <https://idrblab.org/Peoples.php>

Submitted: 20 February 2019; Received (in revised form): 11 March 2019

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Introduction

As one of the most devastating psychiatric disorders, schizophrenia (SCZ) leads to the severe handicap of patients in social engagement and emotional expression [1]. It affects more than 50 million people worldwide [2], whose life expectancy is reduced by about 20 years on average compared with the general population [3]. The etiology of SCZ is regarded as one of the most fundamental puzzles in current biomedical researches [4], and its diagnosis is significantly limited by the lack of objective molecular criteria [5, 6]. To cope with these problems, a variety of studies have been conducted to track down gene signature of this disorder [7–12]. Among these studies, the high-throughput gene expression analysis combining microarray technology and some popular filter feature selection algorithms (such as the Student's *t*-test [13–17], Fisher's exact test [18], analysis of variance [19–21], significant analysis of microarray (SAM) [22–25] and Chi-square test [26]) has emerged as a powerful technique [14–20], and a number of differentially expressed genes (DEGs) between SCZ patients and healthy individuals (such as *SELENBP1* [14] and *CDC42BPB* [19]) are discovered. Based on these DEGs, a series of molecular processes involved in SCZ are discovered, including oxidoreductase activity [14], calcium signaling pathway [19] as well as metabolic and mitochondrial functioning [17].

However, the lists of DEGs for a given disease indication identified by different microarray analyses are highly unstable [27]. Especially for SCZ, the alterations in gene expression are not consistently identified from study to study [14]. As reported in Mistry's pioneer study, there is no overlap among the lists of top-ranked genes identified from seven independent datasets [21]. This inconsistency raises doubts about the reliability of reported signature [28] and significantly hampers its clinical application [29, 30]. Moreover, this may be part of the reason why there is no approved biomarker used for SCZ diagnosis/treatment and why the mechanism underlining SCZ's pathology remains largely unknown [31–33].

This inconsistency among gene signatures by different studies has been attributed to many sources: limited number of samples [34], disease heterogeneity [35], subtle gene expression variation undetected by current feature selection method [14], etc. So far, the analysis combining multiple independent microarray datasets has been used to enlarge the sample size and in turn reduce the disease heterogeneity, which gains a certain level of enhancement in the stability of the identified gene signature (~10% of the DEGs from two separate studies are shared by both) [21, 23]. Recently, the wrapper/embedded feature selection methods (such as SVM-RFE) are proposed as performing better than the filter algorithms [13, 22, 26] in their classification results by involving the classifier using 'Artificial Intelligence' [36–38], and the repeated random sampling is advocated to improve the consistency among signatures identified from various cancer-related datasets [27, 39]. However, the methods currently used for discovering the SCZ signature do not fully consider the consistency among markers discovered by different datasets [21, 27, 40]; it is therefore crucial to develop a new tool of novel strategy for ensuring a stable discovery of the gene signature of SCZ.

In this study, the most comprehensive set of microarray data was constructed by combining information from multiple independent SCZ studies [14–20], and a novel feature selection strategy for stable signature derivation was developed and applied to ensure a reliable discovery. Systematic assessment from multiple perspectives was conducted to guarantee the stability and reliability of this newly proposed strategy, and a cross-validation

among multiple independent studies was further applied to evaluate the reproducibility of this strategy by comparing with traditional methods. Finally, transcription binding motif was analyzed to identify novel transcription factors (TFs) in the development of SCZ. In sum, the findings of this study could facilitate the understanding of SCZ's etiology and the discovery of new diagnostic molecular criteria.

Materials and methods

Collection of microarray data from multiple studies

The prefrontal cortex (PFC) had been widely accepted as the major locus of dysfunction in SCZ by many clinical and neuroimaging studies [41]. In this work, a variety of microarray studies based on tissues from the brain Brodmann areas (BAs) 9, 10 and 46 were therefore collected by searching 'schizophrenia' in such popular SCZ-related data sources as the Gene Expression Omnibus (GEO) [42], Stanley Medical Research Institute (SMRI) [18] and Harvard Brain Bank (HBB) [14]. The collected data should meet the following criteria [21, 33, 43]: (1) the gene expression profiling was conducted using *cDNA* microarray technology for 'Homo sapiens'; (2) the tissues analyzed were based on PFC tissues from BA9, BA10 and BA46; (3) raw data and confounding variables such as gender, age, postmortem interval and brain pH were available for further analysis; and (4) the collected dataset should consist of one group of patients and another group of healthy people. Comprehensive literature search on SCZ microarray studies further yielded additional data [15, 16, 21] of 40 SCZ patients and 35 healthy people. As a result, nine independent microarray studies were collected, and each comprises a cohort of SCZ patients and another cohort of healthy controls. The detailed information of these collected datasets was provided in Table 1, including dataset ID, microarray platform, date of data release, number of samples and brain region as indicated in the original publication. Among these studies, the latest one (GSE62191 [44], 29 patients and 30 healthy controls) was used as independent test dataset and the remaining data (166 patients and 172 controls) were combined to construct classifier and discover the gene signature of SCZ.

Data pre-processing and batch effect removal

Combination of multiple datasets was carried out in R environment (v3.4.3, <http://www.r-project.org>). The raw data (CEL file) was read, log transformed and normalized using R package *affy* [45], and all parameters were set as default. Outliers in each dataset were first checked and removed, and all probe sets were then mapped to their corresponding genes using 'Bioconductor' [46]. The average expression value was retained if a gene was mapped to multiple probes. To remove batch effects among datasets, Z score transformation [47–49] (equation shown below) was used to adjust the gene expression levels in each dataset.

$$Z \text{ score} = \frac{x_i - \bar{x}}{\delta} \quad (1)$$

where x_i refers to the raw intensity of each gene, \bar{x} indicates the average intensity of all genes within a single experiment and δ represents the standard deviation (SD) of all expression intensities in one array. After this procedure, the mean Z score for each array became zero with SD equaling one.

Table 1. A variety of datasets from nine independent microarray studies between SCZ patients and healthy individuals (ordered by the date of dataset release)

Dataset	Microarray platform	Date of data release	No. of samples (patients:controls)	Brain region as indicated in the original publication	Pérez-Santiago et al. (2012)	Mistry et al. (2013)
Haroutunian [16]	HG-U133A/B	Sep 2005	31:29	Frontal (BA10/46)	N.A.	Included
HBB Mclean [14]	HG-U133A	Oct 2005	19:26	Prefrontal cortex (BA9)	Included	Included
Stanley Altarc [18]	HG-U133A	Apr 2006	9:11	Frontal (BA10/46)	Included	Included
Stanley Bahn [18]	HG-U133A	Apr 2006	34:31	Frontal (BA46)	Included	Included
Mirmics [15]	HG-U133A/B	Mar 2008	9:6	Prefrontal cortex (BA46)	Included	Included
GSE17612 [17]	HG-U133 Plus 2	Aug 2009	26:21	Anterior prefrontal cortex (BA10)	Included	Included
GSE21138 [19]	HG-U133 Plus 2	Mar 2010	25:29	Frontal (BA46)	Included	Included
GSE53987 [20]	HG-U133 Plus 2	Jan 2014	13:19	Frontal (BA46)	N.A.	N.A.
GSE62191 [44]	Agilent-014850	Oct 2014	29:30	Frontal cortex (BA46)	N.A.	N.A.

Note: These studies were collected from databases such as GEO [42], SMRI [18] and HBB [14] and a comprehensive literature search on microarray studies of SCZ [15, 16, 21]. All studies were conducted in the prefrontal cortex of postmortem brain tissue, and each dataset comprises a cohort of SCZ subjects and a cohort of healthy people. Patients, no. of schizophrenia patients; Controls, no. of healthy individuals; BA, Brodmann's area; Included, the dataset was included in the corresponding study; N.A., the dataset was not available in the corresponding study; GSE: the accession number in GEO database [42].

Construction of new strategy to ensure the consistent discovery of gene signature

The support vector machine (SVM) showed good performance in classifying microarray data [50], and a wrapper/embedded recursive feature elimination method (RFE-SVM [51]) was widely applied in current studies [40]. In RFE-SVM, a gene ranking function was firstly generated based on a SVM classifier, and the SCZ signature was then identified by eliminating those of no differential expression [51]. In this study, a new strategy based on RFE-SVM was proposed and constructed by (1) integrating the repeated random sampling with consensus scoring and (2) evaluating the ranking consistency among multiple datasets. The workflow of this strategy was illustrated in Figure 1 and demonstrated as follows.

Firstly, the combined dataset was separated into 2000 unique training-test datasets using repeated random sampling [27]. Each training dataset was constructed by a random half of the samples (83 patients and 86 healthy controls) and corresponding test dataset comprised the remaining. Secondly, 2000 datasets were randomly grouped into 20 sampling groups (each with 100 unique training-test datasets). In each sampling group, the gene signature was identified from training dataset using RFE-SVM algorithm. Meanwhile, the classification performance of the signature was evaluated by corresponding test dataset using SVM model with the optimal parameters. Thirdly, to increase the stability among the signatures identified from various datasets, the ranking consistency among 100 training-test datasets in each sampling group were evaluated by a sequential algorithm of consensus scoring. This algorithm included three steps: (1) the genes ranked in the bottom (10%~40% depending on the number of genes selected in different rounds) were selected by making sure that their collective contributions would not exceed the higher-ranked ones; (2) among these selected genes, those ranked in the bottom 50% of previous ranking round were chosen to guarantee that they were consistently low-ranked among several iterations; and (3) the resulting genes of the previous two steps appearing in over 90% of 100 training-test datasets were eliminated. Fourthly, the signature was identified by the highest average classification accuracy among all 100 test datasets. For each sampling set, different parameters were scanned, and various RFE-SVM iterations

were evaluated to find the globally optimized parameters and iterations that gave the highest average class differentiation accuracy for the 500 test sets. Finally, 20 sampling groups were analyzed in the same way, and the stable signature was made up of the DEGs identified simultaneously by all 20 sampling groups.

Systematic assessment on the stability and reliability of the identified SCZ signature

To assess the stability and reliability of the identified signature, a systematic assessment from five different perspectives (i-v) were conducted. The measures were mutually complementary from different perspectives, and all of them were important to assess the stability and reliability of the identified signature.

- (i) The stability among signatures identified from different sampling groups or independent datasets

Firstly, the signature derived from 20 sampling groups was analyzed. A histogram showing the number of DEGs simultaneously discovered by N (1~20) sampling groups was provided. The higher number of DEGs identified by the larger N sampling groups, the more stable the signature identified is. Secondly, to evaluate the stability among signatures identified from independent datasets, the signatures derived from eight studies (Table 1) were analyzed by consistency score (CS) [52] shown below, which quantitatively evaluated the stability among signatures identified from independent datasets [53, 54].

$$CS = \sum_{i=2}^N \sum_{S \in I_i} 2^{i-2} \cdot n_S \quad (2)$$

where N indicates the total number of signatures, I_i refers to a set containing all the intersections of any i signatures and n_S represents the number of DEGs in intersection S . Based on this formula, the value of CS increased exponentially with the accumulation of DEGs in each signature, and the CS values could therefore only be compared (no strict cutoff was available for differentiating 'good' or 'bad' CS values). In sum, the higher

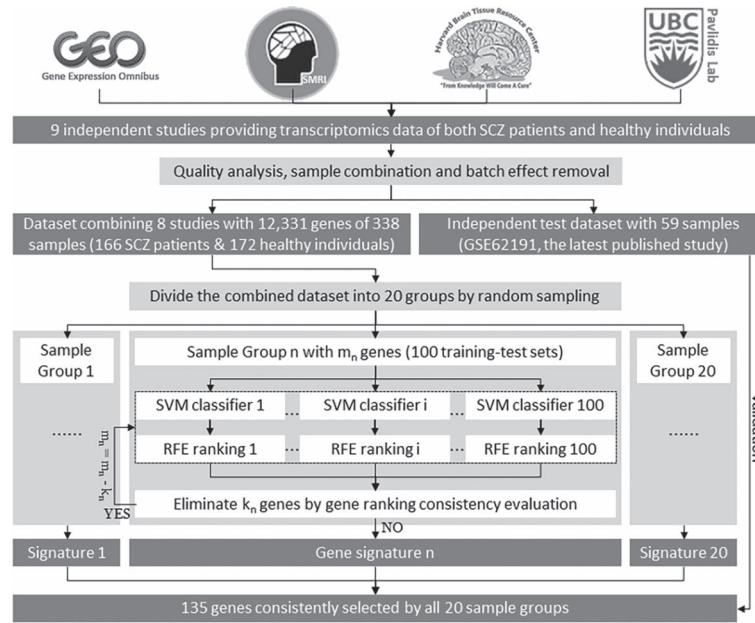


Figure 1. Flowchart of this study and the newly constructed feature selection strategy.

the CS, the more DEGs are identified in common among independent datasets. The CS value of the newly proposed strategy was compared with that of Student's t-test (corrected by Benjamini-Hochberg algorithm) [14, 55, 56] and of the SAM [22–25].

- (ii) The level of disease relevance (DR) of the identified signature

In a complex disorder such as SCZ, the identified signature is expected to contain a substantial percentage of the SCZ-related genes [57–59]. But a certain number of irrelevant genes may be inevitably selected due to measurement variability. Herein, comprehensive literature reviews were performed to investigate the DR of the identified signature, which was represented by the percentage of SCZ-related genes among all DEGs in the identified signature.

- (iii) The role in SCZ played by the hub genes identified from Protein-protein interaction (PPI) network

STRING database [60] was applied to construct PPI network with high confidence level (>0.7). The signature identified in this study was then mapped into this network, and Cytoscape [61] was utilized to visualize the interactions among DEGs. DEGs of high interaction degree (≥ 5) were selected as the hub gene in SCZ.

- (iv) The role in SCZ played by the identified signature based on enrichment analysis

Enrichment analysis on the identified signature was conducted to identify the significantly overrepresented GO terms and KEGG pathways using hypergeometric test ($P < 0.05$) provided by Gene Set Enrichment Analysis (GSEA) tool [62]. Based on the comprehensive literature review on the pathways playing important role in SCZ, the pathways enriched in this study were compared with that of previous reports to validate the signature identified in this work.

- (v) The classification capacity of the identified signature assessed by the independent test dataset

Classification performance of signature identified from the combined dataset was evaluated by predicting the SCZ outcomes of independent test dataset (GSE62191 [44], the most recently published data among all datasets in Table 1) based on the SVM classifier. The performance was assessed by two popular metrics [accuracy (ACC) and Matthews correlation coefficient (MCC)]. ACC indicates the number of true samples successfully predicted divided by the number of samples in the independent test dataset:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

where TP, TN, FP and FN represent the number of true cases, true controls, false cases and false controls, respectively. MCC reflects the stability of classifier based on the identified signature, which is considered as one of the most comprehensive metrics due to its full consideration of TP, TN, FP and FN.

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP) * (TN + FP) * (TP + FN) * (TN + FN)}} \quad (4)$$

ACC and MCC are within the range of [0,1] and [-1,1], respectively. The higher value of each indicates better predictive performance. MCC of -1 represents total disagreement between the prediction results and independent test dataset, 0 denotes no better than random prediction and 1 refers to perfect prediction. The performance of the signature identified in this study was compared with that of two pioneer studies [21, 23].

Reproducibility of the new strategy cross-validated by multiple studies

The stability among signatures identified based on independent studies [52] and the predictive performance of one study on another and vice versa [63, 64] were two critical criteria for assessing the reproducibility of the applied feature selection methods. Herein, three representative studies of large sample size (>50) were firstly chosen from Table 1. Each of these three studies was used as a training dataset, and the remaining studies were used as two independent test datasets, which resulted in six sets of unique training-test data for cross-validation. Secondly, CS (2) was used to assess the stability among signatures identified from those three studies. Thirdly, ACC (3) and MCC (4) of classifier were used to assess the predictive performance among three studies (performance of one study on another and vice versa). Finally, a systematic comparison on the reproducibility among popular feature selection methods (SAM [25] and Student's t-test [14] with FDR-BH [55]) and the new strategy constructed in this study was conducted.

Transcription binding motif analysis

Enrichment analysis on the identified signature was conducted to identify the significantly overrepresented TF binding site using hypergeometric test ($P < 0.05$) in GSEA [62]. The catalog is based on the work reporting 57 commonly conserved regulatory motifs in the promoter regions of human genes [65] and makes it possible to link changes in a microarray experiment to a conserved, putative cis-regulatory element. There were 615 gene sets that share upstream cis-regulatory motifs, which function as the potential TF binding site. Based on the comprehensive literature review on the TFs playing key role in SCZ, those TFs enriched in this study were compared with that of previous reports to validate the signature identified here. Moreover, the newly identified TFs were further proposed as novel factors regulating the development of SCZ.

Results and discussion

Consistent signature derived using the newly constructed strategy

The most comprehensive set of microarray data from eight published studies (Table 1) were combined by pre-processing and batch effect removal. The resulting dataset contained 166 SCZ patients and 172 healthy controls with 12 331 genes after quality control, and no marked distinction in gender, age and postmortem interval between case and control was discovered (Supplementary Table S1). Meanwhile, the level of brain pH was statistically lower in SCZ patients than control, but it should not be a confounding factor here since no strong correlation with each DEG ($|r| \leq 0.38$, Supplementary Table S1) was observed, and the relatively low level of brain pH in SCZ patients was reported to come from the increased anaerobic respiration and hypoxic conditions [21, 23, 66]. As a result, a signature comprising 135 DEGs (Supplementary Table S2) was consistently identified by all 20 sampling groups using the new strategy constructed in this study, and the total number of down-regulated genes (76 genes) was larger than the up-regulated ones (59 genes), which

agreed well with previous studies [67]. Additionally, due to the lack of medication information, it was impossible to incorporate it into this study. However, a comparative analysis between the 135 DEGs identified in this work and the DEGs representing antipsychotic drug function [68] further revealed that there was no overlap between these two signatures, which might indirectly indicate that the signature identified was unaffected by this extraneous factor. Despite the antipsychotics, illicit drugs and smoking are also possible factors that confounded the study of SCZ-related gene expression, but the completely absence of such information made it unlikely to be integrated into this study. As a result, since the influence of these factors could not be entirely excluded, it should be admitted that the lack of such information might be a limitation of the analysis conducted in this study.

The stability and reliability of the identified signature confirmed by five lines of evidence

Evidence a: a high-stability among signatures from different sampling groups or datasets

Firstly, a highly stable signature among all 20 sampling groups was identified using the new strategy. As shown in Figure 2A and Supplementary Table S3, the number of DEGs identified by each sampling group varied from 191 to 203, and 135 DEGs were consistently discovered by all sampling groups (taking up to 66.5%–70.7% of all DEGs identified). Twenty-one genes were only selected by a single group (taking up to <2.1% of all DEGs identified), which suggested a highly stable signature among all sampling groups.

Secondly, the stability among signatures of independent datasets identified in this study was substantially enhanced compared with the traditional feature selection method. The number of DEGs identified from eight independent datasets using the new strategy varied from 191 to 245, and the resulting CS among signatures equaled to 1412. By contrast, the Student's t-test and SAM were applied, and their CSs among eight datasets increased with the enlargement of the selected top-ranked DEGs (Figure 2B; from 11 to 308 for Student's t-test, from 17 to 402 for SAM). However, these CSs were substantially lower than that of the new strategy, which indicated a great enhancement in the stability of signatures identified from independent datasets.

Evidence b: a great DR of the identified SCZ signature

The SCZ signature comprising 135 DEGs was consistently identified by 20 sampling groups, and a great DR (53.3%, Table 2 and Supplementary Table S4) of the identified SCZ signature was discovered. Compared with the DRs of the signatures identified by two pioneer studies [21, 23], the DR of this study slightly outperformed that of Mistry's study [21] (DR=37.8%, Table 2 and Supplementary Table S5) and Perez-Santiago's study [23] (DR=43.1%, Table 2 and Supplementary Table S6). Among these identified DEGs, eight genes (AGT, CYP26B1, LPL, MYL5, SCN1B, SELENBP1, SNN and TIAL1) were identified by recent prominent studies [69–71]. CYP26B1 was involved in the transport of retinoic acid, which was implicated in SCZ's pathogenesis [72]. LPL was an attractive candidate gene for SCZ and SNPs in LPL may confer risk for SCZ [73]. SCN1B was identified as a dysregulated gene in SCZ patients

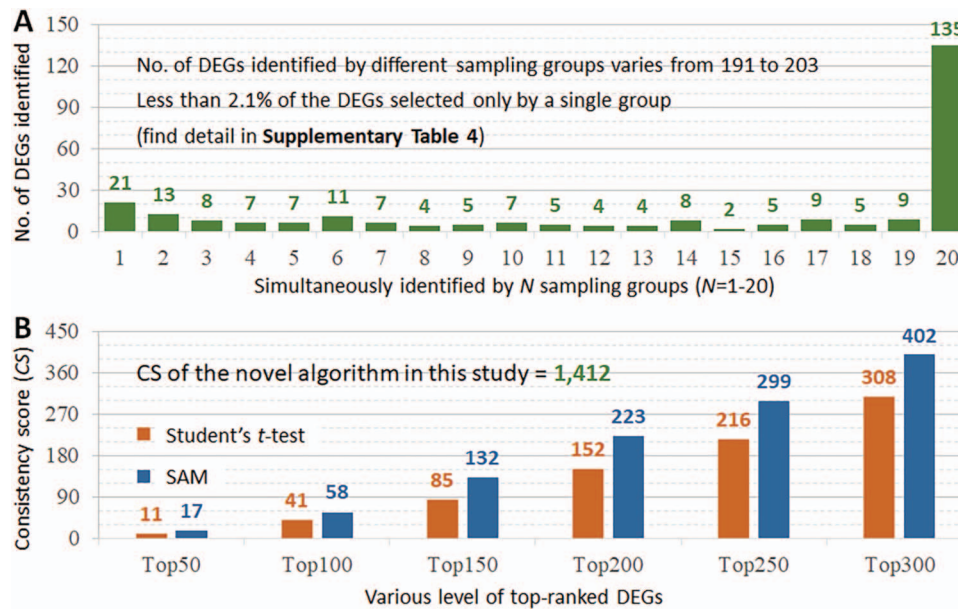


Figure 2. The high-stability among signatures identified from (A) sampling groups and (B) independent datasets. (A) A highly stable signature among all 20 sampling groups was identified using the new strategy. A total of 135 DEGs (>66.5%) were consistently discovered, while 21 (<2.1%) were only selected by a single group. (B) The stability among signatures identified from eight independent datasets by new strategy was substantially enhanced compared with the Student's t-test (orange bars) and SAM (blue bars).

Table 2. Performance comparison of three different lists of DEGs identified by two pioneer combined microarray studies [21, 23] and this study from three different perspectives: (1) the level of relevance between the identified DEGs and SCZ, (2) the level of relevance between the enriched pathways and SCZ and (3) the predictive performance of the identified DEGs on independent test dataset (GSE62191 [44] of 29 patients and 30 controls)

Study	DR of genes associated with SCZ	DR of pathways associated with SCZ	Predictive performance on independent test dataset					ACC (%)	MCC
			TP	FP	TN	FN			
Pérez-Santiago et al. 2012	43.06%	62.50%	19	11	18	11	62.71	0.25	
Mistry et al. 2013	37.84%	66.67%	12	2	27	18	62.71	0.29	
This study	53.33%	100.00%	21	7	22	9	72.88	0.46	

Note: ACC, classification accuracy; MCC, Matthews correlation coefficient.

by microarray study [23]. The level of mRNA expression of *SELENBP1* was significantly up-regulated in the dorsolateral PFC of SCZ patients [14]. *SNN* was found down-regulated with high confidence in SCZ patients compared with the normal controls [17].

Evidence c: the hub genes discovered in this study played a key role in SCZ's development

The topological characteristics of a PPI network could give great insights into the development of SCZ at the molecular level [66]. Herein, a PPI network was constructed using those 135 DEGs, and a network with 29 PPIs were constructed (Figure 3). As a result, two hub genes (*CDC42* and *LPAR1*) were identified, both of which were reported to play a key role in SCZ's development. Particularly, the reduced expression of *CDC42* contributed to the decreased density of dendritic spines in the PFC of SCZ patients [74], and the altered *CDC42* signaling promoted the spine deficits observed in the layer three pyramidal neurons in SCZ patients [75]. Moreover, the prenatal exposure to lysophosphatidic acid (LPA) alone phenocopied many SCZ-like alterations in serum model, whereas the treatment with antagonist against *LPAR1*

could prevent many of those behavioral and neurochemical alterations [76].

Evidence d: GO terms and pathways enriched by the identified signature played a key role in SCZ

As demonstrated in Supplementary Table S7, many enriched GO terms were related to such SCZ-related biological process as ion channel activity [77] and metabolic process [78]. Meanwhile, 21 KEGG pathways were enriched (Supplementary Table S8), 10 out of which agreed with Gardiner's study on miRNA expression profile [79]. As shown, the calcium signaling pathway was discovered to be involved in SCZ's etiology based on a genome-wide association studies [80], and the tyrosine metabolism pathway was found to be peripheral marker of dopamine synthesis associated with SCZ [78]. Moreover, a comprehensive literature review reveals that all 21 pathways were reported at least once by previous studies as SCZ-related (Table 2 and Supplementary Table S9). Compared to the pathways enriched based on two pioneer studies [21, 23], the percentage of SCZ-related pathway of this study was

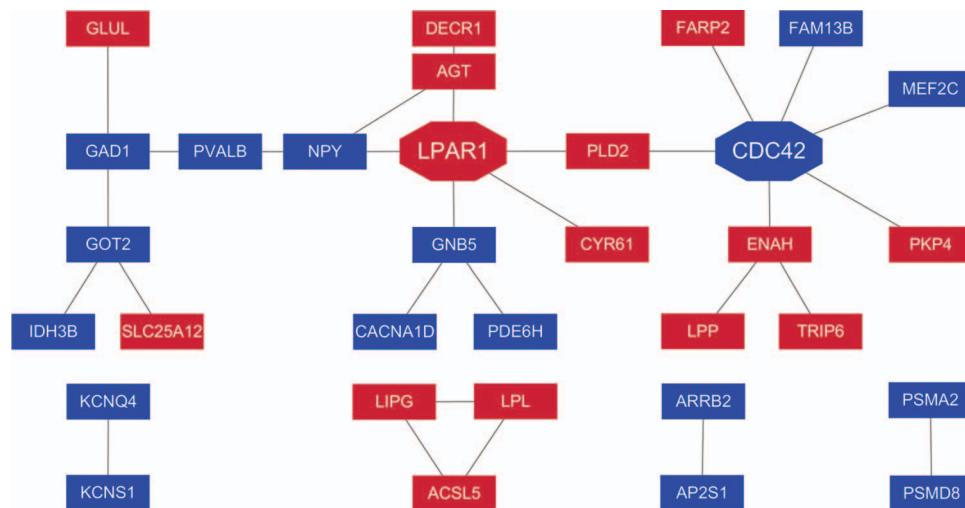


Figure 3. PPI network constructed using 135 gene signatures identified in this study. Up-regulated and down-regulated genes in SCZ patients were illustrated in red and blue, respectively. Genes (LPAR1 and CDC42) of the highest degree were identified as the hubs of this network.

Table 3. The reproducibility of two popular feature selection methods (Student's t-test [14] with FDR-BH [55] and SAM [25]) and the newly proposed strategy assessed by (1) the CS among signatures identified from three representative studies and (2) the differentiating ability (ACC and MCC) of one study on another and vice versa

Train and test sets	Measure	This study	Student's t-test [14] (FDR-BH [55])			SAM [25]		
			Top 100	Top 200	Top 300	Top 100	Top 200	Top 300
The CS among three signatures identified by different method		98	13	38	68	17	51	91
Training: Stanley Bahn [18]	ACC (%)	75.9	63.0	68.5	68.5	61.1	66.7	64.8
Test: GSE21138 [19]	MCC	0.52	0.28	0.37	0.36	0.25	0.36	0.34
Training: Stanley Bahn [18]	ACC (%)	77.4	56.7	61.7	58.3	60.0	58.3	56.7
Test: Haroutunian [16]	MCC	0.53	0.15	0.24	0.17	0.21	0.25	0.22
Training: GSE21138 [19]	ACC (%)	75.4	60.0	69.2	67.7	61.5	69.2	70.8
Test: Stanley Bahn [18]	MCC	0.51	0.20	0.38	0.35	0.26	0.38	0.47
Training: GSE21138 [19]	ACC (%)	71.7	63.3	56.7	68.3	61.7	61.7	65.0
Test: Haroutunian [16]	MCC	0.51	0.27	0.19	0.39	0.25	0.24	0.31
Training: Haroutunian [16]	ACC (%)	75.9	53.8	58.5	53.8	50.8	58.5	58.5
Test: Stanley Bahn [18]	MCC	0.52	0.20	0.17	0.13	0.17	0.16	0.20
Training: Haroutunian [16]	ACC (%)	77.8	57.4	57.4	57.4	59.3	57.4	57.4
Test: GSE21138 [19]	MCC	0.50	0.21	0.21	0.21	0.19	0.21	0.21

Note: FDR-BH, false discovery rate corrected using Benjamini-Hochberg algorithm.

higher than that of Mistry's study [21] (66.7%, Table 2 and Supplementary Table S10) and Perez-Santiago's study [23] (62.5%, Table 2 and Supplementary Table S11).

Moreover, the pathways of Reactome database were enriched by hypergeometric test (FDR $P < 0.05$) based on all 135 DEGs (Supplementary Table S12). Some enriched pathways were validated by reported publications: an enrichment of miRNA targets in 'axon guidance' (identified by this study) was discovered as reflecting key cellular effects in SCZ [81], dysfunction of the 'gamma-aminobutyric acid (GABA)ergic neuronal system' was reported to contribute to the pathogenesis of SCZ [82], a dysregulation of both 'innate and adaptive immune systems' was found contributing to the SCZ symptoms [83, 84], 'transmission across chemical synapses' was discovered as related to SCZ-related gene [85], decreased expression of 'regulation of ornithine decarboxylase' in the ornithine-polyamine metabolism was found to be responsible for higher

concentration of ornithine in SCZ [86] and the etiology of SCZ had been linked to an altered 'metabolism of lipids' in neuronal membranes resulting from an increase in the activity of phospholipase A2 [87, 88].

Evidence e: the SCZ signature demonstrated strong ability to differentiate patients from controls

Differentiating ability of the identified signature was evaluated by independent test dataset (GSE62191 [44]) using the SVM classifier. As shown in Table 2, ACCs of this study, Mistry's study [21] and Perez-Santiago's study [23] were 72.88%, 62.71% and 62.71%, respectively, with their corresponding MCCs equaling to 0.46, 0.25 and 0.29. These results suggest that the SCZ signature identified by the new strategy demonstrated a much better ability to differentiate patients from controls compared with previous reports [21, 23].

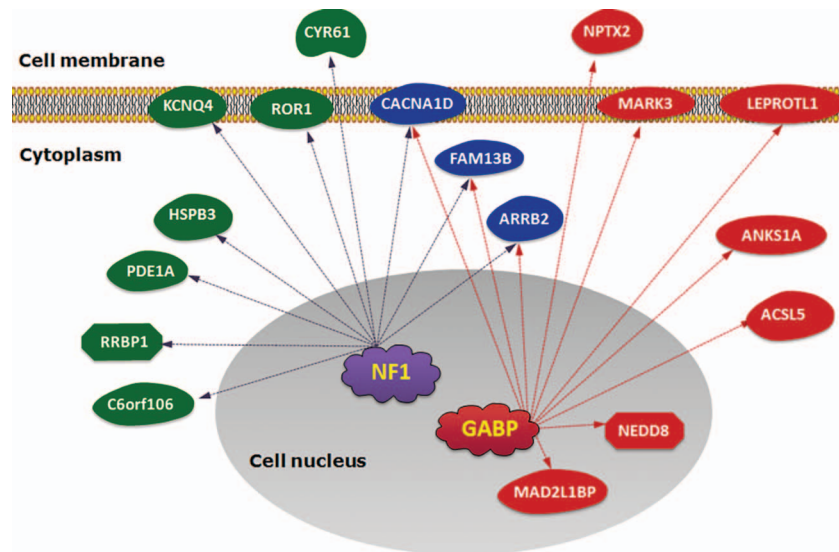


Figure 4. Two novel TFs (NF1 and GABP) identified in this study together with 17 DEGs regulated by them. Co-regulated DEGs were colored in blue, and the DEGs solely regulated by NF1 and GABP were colored in green and red, respectively.

Cross-validating the reproducibility of the new feature selection strategy

Two critical criteria were considered when assessing the reproducibility of feature selection methods: (1) the stability among signatures identified from different independent studies [52] and (2) the differentiating ability of one study on another and vice versa [63]. Herein, three representative studies of large sample size (>50) were selected to discover SCZ signatures (Table 1). Each of these three studies was selected as training dataset, and the remaining ones were chosen as two independent test datasets, which resulted in six sets of unique training-test data for cross-validation. Firstly, the stability (CS) of the new strategy equaled to 98 (Table 3), which was higher than that of Student's t-test (13, 38 and 68 for top 100, top 200 and top 300, respectively) and that of SAM (17, 51 and 91 for top 100, top 200 and top 300, respectively). As the top 100 DEGs were the most frequently applied and widely accepted criterion in DEG selection [89], the new strategy showed a substantially higher stability than these traditional feature selection methods. Secondly, the differentiating abilities of Student's t-test, SAM and the new strategy among three representative datasets were assessed by the ACC and MCC on the corresponding cross-validating datasets (Table 3). The ACCs for Student's t-test, SAM and the new strategy were in the range of 53.8%–69.2%, 50.8%–70.8% and 71.7%–77.8%, respectively, and the ranges of MCCs were 0.13–0.39, 0.16–0.47 and 0.50–0.53, respectively. It was evident that there were great improvements on the differentiating ability of the new strategy compared with those traditional methods. As the top 100 DEGs were the most frequently used and widely accepted criterion in the identification of DEG [89], the ACCs for Student's t-test and SAM would further decrease to 53.8%–63.3% and 50.8%–61.7% with MCCs reduced to 0.15–0.28 and 0.17–0.26. This finding showed significantly higher differentiating ability of the new strategy than the popular methods used in current SCZ studies. The possible reasons of the enhanced performance of this newly proposed strategy could be as follows: (1) integrating repeated random sampling with consensus scoring and evaluating the ranking consistency among multiple datasets have the ability to avoid erroneous elimination of pre-

dictor genes due to noises in microarray data, which ensured the high stability for the selected signatures; and (2) a huge number of parameters were scanned and a variety of feature elimination iterations were assessed for each sampling, and the globally optimized parameters and suitable iteration runs were discovered, which resulted in the highest differentiating ability of the signature. Thus, these steps made sure the better performance on stability and differentiating ability of the new strategy than those traditional methods in current SCZ studies.

Discovering the key TFs in the development of SCZ

TF-binding sites with hypergeometric test $P < 0.005$ using GSEA [62] were enriched based on 135 DEGs, and 19 TFs were overrepresented (Supplementary Table S13). A total of 17 (89.47%) out of these 19 were reported to be closely related to SCZ by previous study, and 2 novel TFs (NF1 and GABP) without any relation to SCZ reported so far were identified. These two TFs may have strong association with the mechanism of SCZ, since NF1 was found to be susceptible to autism [90] and GABP was regulated by the ERK signaling [91] (which was found involved in the pathology of SCZ [92]). Among those 135 DEGs identified, 17 were regulated by NF1 and GABP (Figure 4) with 3 (ARRB2, CACNA1D and FAM13B) co-regulated by both. As shown in Supplementary Table S4, ARRB2 was associated with tardive dyskinesia in Chinese SCZ patients [93], and CACNA1D was over-expressed in SCZ animal models [94]. These findings could be additional strong support to the novel and key roles played by NF1 and GABP in SCZ.

Conclusion

The SCZ signature identified in this study using the newly constructed strategy demonstrated significantly enhanced stability and better differentiating ability compared with those popular feature selection methods used in current SCZ research. Therefore, those identified 135 DEGs underneath this SCZ signature

would provide important clues for discovering the etiology, diagnostic molecule and potential drug target of SCZ. Moreover, 2 novel TFs and 17 previously reported TFs identified here showed great potential in revealing the developmental mechanism and etiology of SCZ, which required additional in-depth study in the future.

Author Contributions

FZ conceived the idea and supervised the work. QY performed the researches. QY and BL developed and wrote the C++ scripts. QY, BL, JT, XC, YW, XL, JH, YZC, WX, YL and YQ prepared and analyzed the data. FZ wrote the manuscript.

Key Points

- The etiology of SCZ is regarded as one of the most fundamental puzzles in current medical research. However, its diagnosis is limited by the lack of objective molecular criteria, and the SCZ gene signatures identified by the independent studies are found highly inconsistent.
- It is crucial to develop new tools of novel strategy for ensuring a stable discovery of gene signature for SCZ, and a novel feature selection strategy was therefore constructed.
- This new strategy demonstrated significantly enhanced stability and better differentiating ability compared with the feature selection methods popular in current SCZ research, and 2 novel and 17 previously reported TFs were identified for revealing the etiology of SCZ.

Supplementary Data

Supplementary data are available online at <https://academic.oup.com/bib>.

Funding

National Key Research and Development Program of China (2018YFC0910500); National Natural Science Foundation of China (81872798); Fundamental Research Funds for Central Universities (2018QNA7023, 10611CDJXZ238826 and 2018CDQYSG0007); Innovation Project on Industrial Generic Key Technologies of Chongqing (cstc2015zdcy-ztx120003).

References

1. Kim Y, Giusti-Rodriguez P, Crowley JJ, et al. Comparative genomic evidence for the involvement of schizophrenia risk genes in antipsychotic effects. *Mol Psychiatry* 2018;**23**:708–12.
2. Sullivan CR, Koene RH, Hasselfeld K, et al. Neuron-specific deficits of bioenergetic processes in the dorsolateral prefrontal cortex in schizophrenia. *Mol Psychiatry* 2018. doi: [10.1038/s41380-018-0035-3](https://doi.org/10.1038/s41380-018-0035-3).
3. Laursen TM, Nordentoft M, Mortensen PB. Excess early mortality in schizophrenia. *Annu Rev Clin Psychol* 2014;**10**:425–48.
4. Kennedy D, Norman C. What don't we know? *Science* 2005;**309**:75.
5. Modai S, Shomron N. Molecular risk factors for schizophrenia. *Trends Mol Med* 2016;**22**:242–53.
6. Li YH, Yu CY, Li XX, et al. Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics. *Nucleic Acids Res* 2018;**46**:D1121–7.
7. Guillozet-Bongaarts AL, Hyde TM, Dalley RA, et al. Altered gene expression in the dorsolateral prefrontal cortex of individuals with schizophrenia. *Mol Psychiatry* 2014;**19**:478–85.
8. Wang S, Che T, Levit A, et al. Structure of the D2 dopamine receptor bound to the atypical antipsychotic drug risperidone. *Nature* 2018;**555**:269–73.
9. Zandersen M, Parnas J. Identity disturbance, feelings of emptiness, and the boundaries of the schizophrenia spectrum. *Schizophr Bull* 2019;**45**:106–13.
10. Allardyce J, Leonenko G, Hamshere M, et al. Association between schizophrenia-related polygenic liability and the occurrence and level of mood-incongruent psychotic symptoms in bipolar disorder. *JAMA Psychiat* 2018;**75**:28–35.
11. Schnack HG. Improving individual predictions: machine learning approaches for detecting and attacking heterogeneity in schizophrenia (and other psychiatric diseases). *Schizophr Res* 2017. doi: [10.1016/j.schres.2017.10.023](https://doi.org/10.1016/j.schres.2017.10.023).
12. Smucny J, Lesh TA, Newton K, et al. Levels of cognitive control: a functional magnetic resonance imaging-based test of an RDoC domain across bipolar disorder and schizophrenia. *Neuropsychopharmacology* 2018;**43**:598–606.
13. Choi KH, Elashoff M, Higgs BW, et al. Putative psychosis genes in the prefrontal cortex: combined analysis of gene expression microarrays. *BMC Psychiatry* 2008;**8**:87.
14. Glatt SJ, Everall IP, Kremen WS, et al. Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 up-regulation in schizophrenia. *Proc Natl Acad Sci U S A* 2005;**102**:15533–8.
15. Garbett K, Gal-Chis R, Gaszner G, et al. Transcriptome alterations in the prefrontal cortex of subjects with schizophrenia who committed suicide. *Neuropsychopharmacol Hung* 2008;**10**:9–14.
16. Katsel P, Davis K, Gorman J, et al. Variations in differential gene expression patterns across multiple brain regions in schizophrenia. *Schizophr Res* 2005;**77**:241–52.
17. Maycox PR, Kelly F, Taylor A, et al. Analysis of gene expression in two large schizophrenia cohorts identifies multiple changes associated with nerve terminal function. *Mol Psychiatry* 2009;**14**:1083–94.
18. Higgs BW, Elashoff M, Richman S, et al. An online database for brain disease research. *BMC Genomics* 2006;**7**:70.
19. Narayan S, Tang B, Head SR, et al. Molecular profiles of schizophrenia in the CNS at different stages of illness. *Brain Res* 2008;**1239**:235–48.
20. Lanz TA, Joshi JJ, Reinhart V, et al. 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**:e0121744.
21. Mistry M, Gillis J, Pavlidis P. Genome-wide expression profiling of schizophrenia using a large combined cohort. *Mol Psychiatry* 2013;**18**:215–25.
22. Hill MJ, Killick R, Navarrete K, et al. Knockdown of the schizophrenia susceptibility gene TCF4 alters gene expression and proliferation of progenitor cells from the developing human neocortex. *J Psychiatry Neurosci* 2017;**42**:181–8.

23. Perez-Santiago J, Diez-Alarcia R, Callado LF, et al. A combined analysis of microarray gene expression studies of the human prefrontal cortex identifies genes implicated in schizophrenia. *J Psychiatr Res* 2012;**46**:1464–74.
24. Strazisar M, Cammaerts S, van der K, et al. MIR137 variants identified in psychiatric patients affect synaptogenesis and neuronal transmission gene sets. *Mol Psychiatry* 2015;**20**:472–81.
25. Iwamoto K, Bundo M, Kato T. Altered expression of mitochondria-related genes in postmortem brains of patients with bipolar disorder or schizophrenia, as revealed by large-scale DNA microarray analysis. *Hum Mol Genet* 2005;**14**:241–53.
26. Chen X, Long F, Cai B, et al. A novel relationship for schizophrenia, bipolar and major depressive disorder part 3: evidence from chromosome 3 high density association screen. *J Comp Neurol* 2018;**526**:59–79.
27. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005;**365**:488–92.
28. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* 2006;**103**:5923–8.
29. Pickard BS. Schizophrenia biomarkers: translating the descriptive into the diagnostic. *J Psychopharmacol* 2015;**29**:138–43.
30. Zhu F, Li XX, Yang SY, et al. Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol Sci* 2018;**39**:229–31.
31. Geaghan M, Cairns MJ. MicroRNA and posttranscriptional dysregulation in psychiatry. *Biol Psychiatry* 2015;**78**:231–9.
32. de FM, Bouwkamp CG, Gunhanlar N, et al. Candidate CSPG4 mutations and induced pluripotent stem cell modeling implicate oligodendrocyte progenitor cell dysfunction in familial schizophrenia. *Mol Psychiatry* 2018. doi: [10.1038/s41380-017-0004-2](https://doi.org/10.1038/s41380-017-0004-2).
33. Li YH, Li XX, Hong JJ, et al. Clinical trials, progression-speed differentiating features and swiftness rule of the innovative targets of first-in-class drugs. *Brief Bioinform* 2019. doi: [10.1093/bib/bby130](https://doi.org/10.1093/bib/bby130).
34. Osborn D, Burton A, Hunter R, et al. Clinical and cost-effectiveness of an intervention for reducing cholesterol and cardiovascular risk for people with severe mental illness in English primary care: a cluster randomised controlled trial. *Lancet Psychiatry* 2018;**5**:145–54.
35. Schwalbe EC, Lindsey JC, Nakjang S, et al. Novel molecular subgroups for clinical classification and outcome prediction in childhood medulloblastoma: a cohort study. *Lancet Oncol* 2017;**18**:958–71.
36. Juneja A, Rana B, Agrawal RK. fMRI based computer aided diagnosis of schizophrenia using fuzzy kernel feature extraction and hybrid feature selection. *Multimed Tools Appl* 2018;**77**:3963–89.
37. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007;**23**:2507–17.
38. He W, Jia C, Zou Q. 4mCPred: machine learning methods for DNA N4-methylcytosine sites prediction. *Bioinformatics* 2019;**35**:593–601.
39. Zeller G, Tap J, Voigt AY, et al. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* 2014;**10**:766.
40. Tang ZQ, Han LY, Lin HH, et al. Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation. *Cancer Res* 2007;**67**:9996–10003.
41. Mirnics K, Middleton FA, Marquez A, et al. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* 2000;**28**:53–67.
42. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 2013;**41**:D991–5.
43. Li S, Roupheal N, Duraisingham S, et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat Immunol* 2014;**15**:195–204.
44. de A, Maschietto M, Lima L, et al. Innate immune response is differentially dysregulated between bipolar disease and schizophrenia. *Schizophr Res* 2015;**161**:215–21.
45. Gautier L, Cope L, Bolstad BM, et al. Affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004;**20**:307–15.
46. Tippmann S. Programming tools: adventures with R. *Nature* 2015;**517**:109–10.
47. Lazar C, Meganck S, Taminou J, et al. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform* 2013;**14**:469–90.
48. Tang J, Fu J, Wang Y, et al. ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* 2019. doi: [10.1093/bib/bby127](https://doi.org/10.1093/bib/bby127).
49. Li B, Tang J, Yang Q, et al. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. *Sci Rep* 2016;**6**.
50. Pochet N, De F, Suykens JA, et al. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* 2004;**20**:3185–95.
51. Inza I, Larranaga P, Blanco R, et al. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif Intell Med* 2004;**31**:91–103.
52. Wang X, Gardiner EJ, Cairns MJ. Optimal consistency in microRNA expression analysis using reference-gene-based normalization. *Mol Biosyst* 2015;**11**:1235–40.
53. Li B, Tang J, Yang Q, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res* 2017;**45**:W162–70.
54. Schwarzenbach H, da AM, Calin G, et al. Data normalization strategies for MicroRNA quantification. *Clin Chem* 2015;**61**:1333–42.
55. Shi Z, Zhang Q, Chen H, et al. STAT4 polymorphisms are associated with neuromyelitis optica spectrum disorders. *Neuromolecular Med* 2017;**19**:493–500.
56. Han Z, Xue W, Tao L, et al. Identification of key Long non-coding RNAs in the pathology of Alzheimer's disease and their functions based on genome-wide associations study, microarray, and RNA-seq data. *J Alzheimers Dis* 2019;**68**:339–55.
57. Grames MS, Dayton RD, Jackson KL, et al. Cre-dependent AAV vectors for highly targeted expression of disease-related proteins and neurodegeneration in the substantia nigra. *FASEB J* 2018;**32**:4420–7.
58. Wu Y, Yao YG, Luo XJ. SZDB: a database for schizophrenia genetic research. *Schizophr Bull* 2017;**43**:459–71.
59. Yang H, Qin C, Li YH, et al. Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information. *Nucleic Acids Res* 2016;**44**:D1069–74.

60. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;**43**:D447–52.
61. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498–504.
62. Subramanian A, Tamayo P, Mootha VK, 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**:15545–50.
63. Kuo CS, Liu CY, Pavlidis S, et al. Unique immune gene expression patterns in Bronchoalveolar lavage and tumor adjacent non-neoplastic lung tissue in non-small cell lung cancer. *Front Immunol* 2018;**9**:232.
64. Zou Q, Xing P, Wei L, et al. Gene2vec: gene subsequence embedding for prediction of mammalian N (6)-methyladenosine sites from mRNA. *RNA* 2019;**25**:205–18.
65. Xie X, Lu J, Kulbokas EJ, et al. Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals. *Nature* 2005;**434**:338–45.
66. Prabakaran S, Swatton J, Ryan M, et al. Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol Psychiatry* 2004;**9**:684–97.
67. Arion D, Unger T, Lewis DA, et al. Molecular evidence for increased expression of genes related to immune and chaperone function in the prefrontal cortex in schizophrenia. *Biol Psychiatry* 2007;**62**:711–21.
68. Thomas EA. Molecular profiling of antipsychotic drug function: convergent mechanisms in the pathology and treatment of psychiatric disorders. *Mol Neurobiol* 2006;**34**:109–28.
69. Fromer M, Rousos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* 2016;**19**:1442–53.
70. Jaffe AE, Tao R, Norris AL, et al. qSVA framework for RNA quality correction in differential expression analysis. *Proc Natl Acad Sci U S A* 2017;**114**:7130–5.
71. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* 2014;**511**:421–7.
72. Wan C, Shi Y, Zhao X, et al. Positive association between ALDH1A2 and schizophrenia in the Chinese population. *Prog Neuropsychopharmacol Biol Psychiatry* 2009;**33**:1491–5.
73. Xie C, Wang ZC, Liu XF, et al. Association between schizophrenia and single nucleotide polymorphisms in lipoprotein lipase gene in a Han Chinese population. *Psychiatr Genet* 2011;**21**:307–14.
74. Hill J, Hashimoto T, Lewis D. Molecular mechanisms contributing to dendritic spine alterations in the prefrontal cortex of subjects with schizophrenia. *Mol Psychiatry* 2006;**11**:557–66.
75. Ide M, Lewis DA. Altered cortical CDC42 signaling pathways in schizophrenia: implications for dendritic spine deficits. *Biol Psychiatry* 2010;**68**:25–32.
76. Mirendil H, Thomas E, De C, et al. LPA signaling initiates schizophrenia-like brain and behavioral changes in a mouse model of prenatal brain hemorrhage. *Transl Psychiatry* 2015;**5**:e541.
77. Purcell SM, Moran JL, Fromer M, et al. A polygenic burden of rare disruptive mutations in schizophrenia. *Nature* 2014;**506**:185–90.
78. Felger JC, Li L, Marvar PJ, et al. Tyrosine metabolism during interferon-alpha administration: association with fatigue and CSF dopamine concentrations. *Brain Behav Immun* 2013;**31**:153–60.
79. Gardiner E, Beveridge N, Wu J, et al. Imprinted DLK1-DIO3 region of 14q32 defines a schizophrenia-associated miRNA signature in peripheral blood mononuclear cells. *Mol Psychiatry* 2012;**17**:827–40.
80. Ripke S, O'Dushlaine C, Chambert K, et al. Genome-wide association analysis identifies 13 new risk loci for schizophrenia. *Nat Genet* 2013;**45**:1150–9.
81. Santarelli DM, Beveridge NJ, Tooney PA, et al. Upregulation of dicer and microRNA expression in the dorsolateral prefrontal cortex Brodmann area 46 in schizophrenia. *Biol Psychiatry* 2011;**69**:180–7.
82. Umeda K, Iritani S, Fujishiro H, et al. Immunohistochemical evaluation of the GABAergic neuronal system in the prefrontal cortex of a DISC1 knockout mouse model of schizophrenia. *Synapse* 2016;**70**:508–18.
83. Michel M, Schmidt MJ, Mirmics K. Immune system gene dysregulation in autism and schizophrenia. *Dev Neurobiol* 2012;**72**:1277–87.
84. Horvath S, Mirmics K. Immune system disturbances in schizophrenia. *Biol Psychiatry* 2014;**75**:316–23.
85. Sundararajan T, Manzardo AM, Butler MG. Functional analysis of schizophrenia genes using GeneAnalytics program and integrated databases. *Gene* 2018;**641**:25–34.
86. He Y, Yu Z, Giegling I, et al. Schizophrenia shows a unique metabolomics signature in plasma. *Transl Psychiatry* 2012;**2**:e149.
87. Gattaz WF, Hubner CV, Nevalainen TJ, et al. Increased serum phospholipase A2 activity in schizophrenia: a replication study. *Biol Psychiatry* 1990;**28**:495–501.
88. van SJ, Smuts CM, Hon D, et al. Changes in erythrocyte membrane fatty acids during a clinical trial of eicosapentaenoic acid (EPA) supplementation in schizophrenia. *Metab Brain Dis* 2009;**24**:659–72.
89. Zhou H, Skolnick J. A knowledge-based approach for predicting gene-disease associations. *Bioinformatics* 2016;**32**:2831–8.
90. Marui T, Hashimoto O, Nanba E, et al. Association between the neurofibromatosis-1 (NF1) locus and autism in the Japanese population. *Am J Med Genet B Neuropsychiatr Genet* 2004;**131B**:43–7.
91. Hoffmeyer A, Avots A, Flory E, et al. The GABP-responsive element of the interleukin-2 enhancer is regulated by JNK/SAPK-activating pathways in T lymphocytes. *J Biol Chem* 1998;**273**:10112–9.
92. Ono T, Hashimoto E, Ukai W, et al. The role of neural stem cells for in vitro models of schizophrenia: neuroprotection via Akt/ERK signal regulation. *Schizophr Res* 2010;**122**:239–47.
93. Liou YJ, Wang YC, Chen JY, et al. The coding-synonymous polymorphism rs1045280 (Ser280Ser) in beta-arrestin 2 (ARRB2) gene is associated with tardive dyskinesia in Chinese patients with schizophrenia. *Eur J Neurol* 2008;**15**:1406–8.
94. Genis-Mendoza A, Gallegos-Silva I, Tovilla-Zarate CA, et al. Comparative analysis of gene expression profiles involved in calcium signaling pathways using the NLVH animal model of schizophrenia. *J Mol Neurosci* 2018;**64**:111–6.