Molecular Therapy
Nucleic Acids
Original Article

Implications of Newly Identified Brain eQTL Genes and Their Interactors in Schizophrenia

Lei Cai,[1,2,3,9] Tao Huang,[1,4,9] Jingjing Su,[5,9] Xinxin Zhang,[1] Wenzhong Chen,[1] Fuquan Zhang,[6] Lin He,[1,3] and Kuo-Chen Chou[1,2,7,8]

[1]Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Genetics and Development, Shanghai Mental Health Center, Shanghai Jiaotong University, Shanghai 200240, China; [2]Gordon Life Science Institute, Boston, MA 02478, USA; [3]Shanghai Center for Women and Children's Health, Shanghai 200062, China; [4]Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; [5]Department of Neurology, Shanghai Ninth People's Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200011, China; [6]Department of Psychiatry, Wuxi Mental Health Center, Nanjing Medical University, Wuxi 214015, China; [7]Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, 610054, China; [8]Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Schizophrenia (SCZ) is a devastating genetic mental disorder. Identification of the SCZ risk genes in brains is helpful to understand this disease. Thus, we first used the minimum Redundancy-Maximum Relevance (mRMR) approach to integrate the genome-wide sequence analysis results on SCZ and the expression quantitative trait locus (eQTL) data from ten brain tissues to identify the genes related to SCZ. Second, we adopted the variance inflation factor regression algorithm to identify their interacting genes in brains. Third, using multiple analysis methods, we explored and validated their roles. By means of the aforementioned procedures, we have found that (1) the cerebellum may play a crucial role in the pathogenesis of SCZ and (2) *ITIH4* may be utilized as a clinical biomarker for the diagnosis of SCZ. These interesting findings may stimulate novel strategy for developing new drugs against SCZ. It has not escaped our notice that the approach reported here is of use for studying many other genome diseases as well.

INTRODUCTION

Schizophrenia (SCZ) is a devastating chronic psychiatric disorder, characterized by a group of symptoms including hallucinations and delusions, severely inappropriate emotional and behavioral responses, substantial cognitive changes, the division of thought, and impaired coordination of social or occupational function.[1] Despite its low prevalence (about 1% of the population), SCZ imposes a substantial burden on the family and society.[2] Now, it is widely considered to be of a complex genetic disease, which is affected by environmental factors together with multiple micro- or intermediate-effect genes.[3,4] Although the studies by the genome-wide association study (GWAS) analysis have identified a number of significantly associated variants with SCZ, most of them are located in noncoding regions and their effects remain elusive.

In 2001, the mRNA expression in the whole genome was proposed as a quantitative trait. Meanwhile, the first expression quantitative trait locus (eQTL) mapping analysis, which relates SNP allelic variation to target transcript abundance, was performed.[5] Because the gene expression is tissue specific and influenced by environmental factors, integration of eQTL data and the variants associated with a specific disease in specific tissue may reveal some problematic genes causing diseases. Furthermore, many studies have indicated that significant changes in gene expression rather than alterations in protein structure and/or function play a crucial role in SCZ susceptibility.[6–8] Accordingly, SCZ-susceptible variants could be eQTLs that would influence the expression of some genes.

In the present study, we used the minimum Redundancy-Maximum Relevance (mRMR) algorithm to identify the potential eQTL genes for SCZ by integrating eQTL data from 10 human brain tissues from the Genotype-Tissue Expression (GTEx) project with the results from a meta-analysis of GWASs.[9,10] Compared with common classifiers of the Naive Bayes, a library for support vector machines (LIBSVM) version (v.)3.22, linear discriminant analysis (LDA), and logistic regression, mRMR algorithm has the advantages of reducing mutual redundancy within the selected genes and effectively selecting the genes to be more representative of the target phenotypes.[11–14] In addition to eQTL genes, their target genes may also have some effects on SCZ; we subsequently used the identified genes to explore their

Received 11 March 2018; accepted 30 May 2018;
https://doi.org/10.1016/j.omtn.2018.05.026.

[9]These authors contributed equally to this work.

Correspondence: Lei Cai, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Genetics and Development, Shanghai Mental Health Center, Shanghai Jiaotong University, 55 Guangyuan Xi Road, Shanghai 200240, China.
E-mail: lcai@sjtu.edu.cn
Correspondence: Lin He, Bio-X Institutes, Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders (Ministry of Education), Collaborative Innovation Center for Genetics and Development, Shanghai Mental Health CenterShanghai Jiaotong University, 55 Guangyuan Xi Road, Shanghai 200240, China.
E-mail: helinhelin123@yeah.net

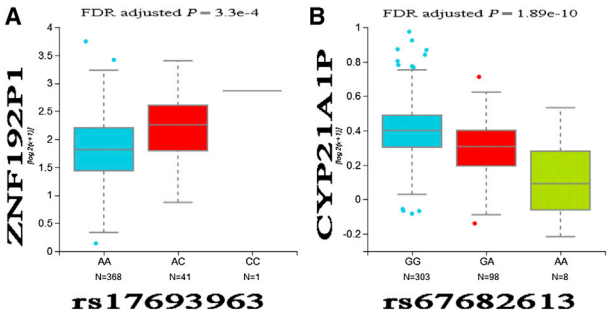

Figure 1. Association of eQTL with Corresponding Genes Based on the BrainCloud eQTL Database
(A) rs17693963 with ZNF 192P1. (B) rs67682613 with CYP21A1P.

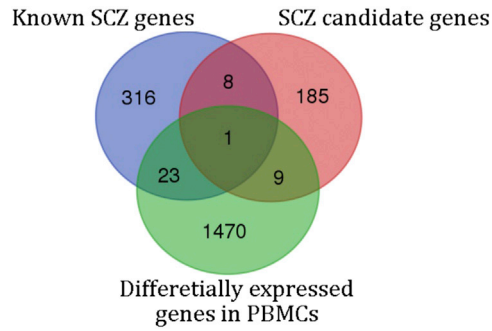

Figure 2. Venn Diagram Comparison among Three Groups of Genes
Known SCZ genes reported by GWASs, identified SCZ candidate genes in the present study, and differentially expressed genes in PBMCs. Error bars mean SD.

target genes in corresponding tissues, and we determined their putative roles in the brain.

RESULTS

SCZ Risk Genes Based on the Integration Analysis of eQTL and GWASs

A total of 10,301 SNPs met the GWAS significant threshold of p < 10e−8. From the 10 brain tissues, 492,401 eQTL SNPs, which affected the expression of 22,832 genes, were collected. Of these, only 134 SNPs exhibited positive expression SNPs (eSNPs). Thus, for each of 10,000 SNP benchmark datasets, there were 134 positive eSNPs and 134 negative randomly selected eSNPs. Subsequently, based on the MaxRel scores of the eQTL gene feature in the mRMR analysis, we identified the most discriminative eQTL gene features from different brain tissues for the positive eSNPs of SCZ. Using the average MaxRel score of greater than 0.01 and the frequency of gene feature reappearance in the top 500 among all tested eSNP-gene pair matrix more than 70%, we identified 22 eQTL gene features, which included 12 candidate genes in eight different brain tissues, excluding the anterior cingulate cortex BA24 and the caudate basal ganglia (see Table S1).

Furthermore, these 12 genes were supported by at least one item of evidence from the GWASs, gene differential expression ones, and/or alternative eQTL data for replication. These genes may play crucial roles in the pathogenesis of SCZ, and they can serve as potential putative genes that increase the risk of developing SCZ. Of these, the gene with the highest average MaxRel score was *PRSS16* from cerebellum (average MaxRel = 0.0311), which exhibited the most significant association with SCZ and was only supported to be risk for SCZ by the results of GWASs. Furthermore, this gene was also found to increase the risk for SCZ in the cerebellar hemisphere and hippocampus. The second most significant SCZ eQTL gene was complement factor 4A (*C4A*) in the cerebellum and the frontal cortex BA9 (average MaxRel = 0.0165 and 0.0129, respectively), which was only supported to be risk for SCZ by the results of the gene differential expression study GEO: GSE53987. Interestingly, the *AS3MT* gene was found to be a potential risk gene for SCZ in the most number of tissues, i.e., the cerebellum, cerebellar hemisphere, and cortex. Furthermore, in the cerebellum noncoding RNA *lnc-CNNM2-1* targeting the *CNNM2* gene (average MaxRel = 0.0127) and *CYP21A1P* (average MaxRel = 0.0147) in the cerebellum and *ZNF192P1* in the cerebellum and cortex (both average MaxRel = 0.011) were identified to be SCZ risk genes in the present study (Figure 1).

Potential Genes Interacted with SCZ Risk Genes Identified above

To determine the target interacting genes of the SCZ risk genes identified, we first identified their coexpressed genes in each of the corresponding brain tissues using the variance inflation factor (VIF) regression algorithm, and then we used adjusted R^2 to select the potential interactors. In total, 186 genes were identified to interact with the nine SCZ candidate genes (i.e., *ARL3*, *AS3MT*, *C10orf32*, *C4A*, *CYP21A1P*, *HLA-DMA*, *PRSS16*, *ARL6IP4*, and *SNX19*) in the three brain tissues of cerebellum, frontal Cortex BA9, and nucleus accumbens basal ganglia (see Table S2). Of these, *ARL6IP4* in the nucleus accumbens basal ganglia exhibited the largest number (174) of functionally relevant target genes. Moreover, the nucleus accumbens basal ganglia interactor gene *SNX19* had 96 target genes that probably participate in a wide variety of physiological processes relevant for SNX19. Another interactor gene, *C4A*, was identified with nine target genes in the cerebellum and with four target genes in the frontal cortex BA9. In the present study, only nine genes of all these identified genes overlapped with known SCZ genes (Figure 2).

Enrichment Analysis

Gene enrichment analysis of the genes expressed in the brain indicated that the candidate risk genes are significantly enriched within the known SCZ genes[15] (p = 0.015). Furthermore, gene ontology (GO) enrichment analysis demonstrated that the genes were involved in a variety of physiological and pathophysiological processes. Within the molecular function GO category, all the above genes were significantly enriched in protein binding (false discovery rate [FDR]-adjusted p = 5.75E−06) and poly(A) RNA binding (FDR-adjusted p = 2.02E−03). Within the cellular component GO category, the significantly enriched terms were cytosol (FDR-adjusted p = 1.06E−05), mitochondrion (FDR-adjusted p = 1.07E−04),

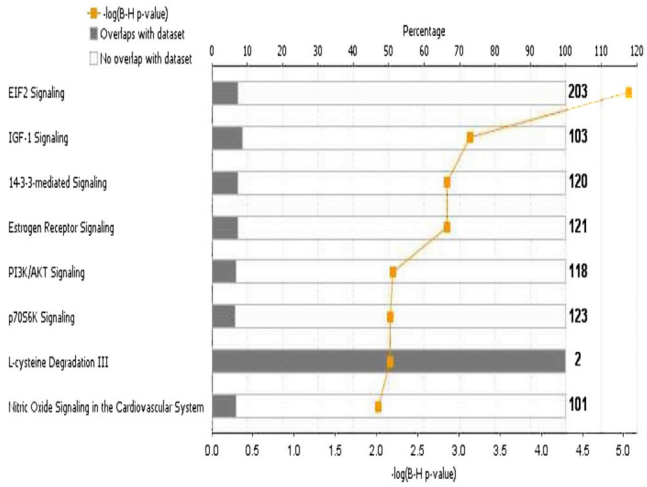

Figure 3. The Top Eight Signaling Pathways in which All Identified Genes in the Present Study Are Enriched

extracellular exosome (FDR-adjusted p = 2.85E−04), and myelin sheath (FDR-adjusted p = 3.61E−03). Within the biological process GO category, three enriched terms, specifically SRP-dependent co-translational protein targeting to the membrane (FDR-adjusted p = 8.18E−03), viral transcription (FDR-adjusted p = 2.99E−02), and nuclear-transcribed mRNA catabolic process, nonsense-mediated decay (FDR-adjusted p = 4.64E−02), were revealed (see Table S2).

Results from pathway enrichment analysis, performed using the hypergeometric test, are illustrated in Figure 3. Eight of these pathways fulfilled the criterion that $-\log p > 2$. The top three pathways associated with SCZ were EIF2 signaling, IGF-1 signaling, and 14-3-3-mediated signaling. Moreover, interestingly, all proteins in L-cysteine Degradation III pathways (i.e., MPST and GOT1) were among the candidate SCZ proteins (Table S3).

Systematic Review of 14-3-3 Isoforms

Because 14-3-3 protein includes seven isoforms (β, ε, γ, σ, η, θ, and ζ) and the 14-3-3-mediated pathway is involved in SCZ, we attempted to identify the isoforms that might play a role in SCZ. To that end, we performed an updated systemic review of 14-3-3 isoforms with SCZ. The previous results are listed in Table S4. In total, 11 studies meeting the analysis criteria were included; they concerned six isoforms, namely, β, ε, γ, η, θ, and ζ. Among these studies, p values were calculated on the basis of either Student's t test or a multivariate analysis of covariance. All studies of the θ isoform had p values less than 0.05. Since a multivariate analysis of covariance is more strict, after excluding the studies using Student's t test, all six isoforms were significantly associated with SCZ; and, furthermore, the average fold changes (FCs) for the six isoforms β, ε, γ, η, θ, and ζ were 0.89, 1.42, 0.741, 1.135, 0.787, and 0.879, respectively. According to one study,[16] a variation of a minimum 40% is viewed as significant regulation; thus, the results suggest that the ε and θ isoforms tend to play important roles in SCZ.

Potential Candidate Genes for Clinical Diagnosis

Among the genes identified in brain tissues, inter-α-trypsin inhibitor H4 (*ITIH4*), *MOSPD3*, *SNAP25*, *RNPEPL1*, *UBE4A*, *SLC25A39*, *ZNF688*, *ANK2*, *BAD*, and *THAP7* were found to be significantly dysregulated in peripheral blood mononuclear cells (PBMCs) of patients with SCZ with the FC > 1.5 (see Table S5). Furthermore, *ITIH4* ($p_{adj.}$ = 0.010, logFC = −1.102), *SNAP25* ($p_{adj.}$ = 0.026, logFC = −1.373), *RNPEPL1* ($p_{adj.}$ = 0.028, logFC = −0.725), *UBE4A* ($p_{adj.}$ = 0.044, logFC = −0.780), *BAD* ($p_{adj.}$ = 0.030, logFC = −0.671), and *THAP7* ($p_{adj.}$ = 0.048, logFC = −0.878) were found to be downregulated significantly in patients with SCZ, whereas the significantly upregulated genes were *SLC25A39* ($p_{adj.}$ = 0.0002, logFC = 0.979), *ZNF688* ($p_{adj.}$ = 0.012, logFC = 0.606), *ANK2* ($p_{adj.}$ = 0.026, logFC = 0.637), and *MOSPD3* ($p_{adj.}$ = 0.003, logFC = 0.611). The common genes among those known for SCZ, candidate genes in the brain and dysregulated expressed genes in PBMCs are also depicted in Figure 1. However, only *ITIH4* was found to display an overlap among these three groups, and, therefore, it may serve as a potential putative gene for diagnosing SCZ by a blood test.

DISCUSSION

Schizophrenia is a multifactorial and polygenic psychiatric disorder. Due to limited sample size, several GWASs on SCZ reported various independent genomic loci exceeding genome-wide significance, i.e., $p < 10^{-8}$.[17–20] Furthermore, most of the identified risk variants are located in noncoding regions. How these risk variants contribute to SCZ susceptibility remains unidentified. Therefore, the Schizophrenia Working Group of the Psychiatric Genomics Consortium (PGC) was created to combine all available SCZ samples with published or unpublished GWAS analysis genotypes into a single, systematic meta-analysis.[10] Since many studies implicated that changes in gene expression rather than alterations in protein structure and/or function play critical roles in SCZ susceptibility,[7,8] it was suggested that those risk variants in GWAS may alter the expression of SCZ-related genes rather than protein function. Furthermore, it is well known that brain tissues appear to be most relevant to SCZ; however, so far which part of the brain plays a significant role in SCZ remains elusive. In the present study, we integrated eQTL data from 10 brain tissues and genetic association findings from the largest meta-GWAS on SCZ with a total of 150,064 subjects using mRMR method, and we identified the potential putative interactors in corresponding brain tissues.

The simple mRMR approach is one of the most potent methods proposed by Peng et al. to use mutual information (MI) for gene feature selection based on microarray gene expression data.[11,21–24] Probably, mRMR is much faster and in practice more robust, since this algorithm is theoretically more efficient to perform an optimal max-dependency selection and produce a feature set with little pairwise redundancy, and usually mRMR yields more excellent classification accuracy than other classifiers (e.g., LIBSVM, LDA, Naive Bayes, logistic regression, etc.). Using this algorithm, we found that cerebellum is the most closely linked to SCZ since the most amount of genes was identified within it than other brain tissues. Cerebellum has been established to be associated with the auditory, cognitive,

and social behavior of SCZ in addition to motor function.[25] Furthermore, our results found that *C4A*, known for its role in immunity, is an eQTL gene in cerebellum and frontal cortex, which supports that *C4A* is an authentic risk gene for SCZ. The *C4A* gene, located in major histocompatibility complex (MHC) class III region on chromosome 6, encodes the acidic form of complement factor 4, which is the primary effector of the innate and the adaptive immune system, and is involved in the classical pathway of complement activation system.[26] Recently, studies reported that C4 might play essential roles in the pathogenesis of SCZ.[26,27] It was also suggested that some *C4* variants in the brain caused significant differential expression of *C4A* and *C4B* and the SCZ-related common *C4* allele is more likely to cause higher expression of *C4A*.[26]

In the current study, at least three items of evidence support that *ITIH4* is a risk gene for SCZ. Also, interestingly, *ITIH4*, which was also identified as an eQTL gene in putamen basal ganglia, was found to be significantly decreased in the serum of patients suffering acute-phase processes.[28] ITIH4 is one of the heavy chains of inter-α-trypsin inhibitor (ITI), which encodes the ITI family molecules with four other homologous heavy chains and one light chain. It has been demonstrated that the ITIH3-ITIH4 region is one of the most significantly associated with SCZ and bipolar disorder.[20] Also, we have identified that the SNPs rs2239547, rs4687552, and rs2535627 in *ITIH4* exceed the GWAS threshold and regulate expression of *ITIH4*. Over the last decade, many research groups have been interested in finding a reliable clinical biomarker for the early detection of SCZ.[2,29,30] Although significant differences between patients with SCZ and healthy controls have been found in brain structure, functional brain imaging, gene expression, and genetic polymorphisms, etc., the overlap of reported abnormalities between patients and healthy controls indicates that there is no valid diagnostic test for establishing a concrete early diagnosis of SCZ.[29] Here, supported by the above multiple findings, *ITIH4* is suggested to be a potential clinical biomarker for the diagnosis of SCZ through a blood test, which can provide easy operation and objective diagnosis criteria. Furthermore, we identified two risk genes risk for SCZ, including *CYP21A1P* and *ZNF192P1*. These are pseudogenes, whose products function as regulatory elements. Although we identified three target genes for CYP21A1P, further work is warranted to investigate the mechanism underlying these genes.

Since SCZ is a complex disease, multiple genes/pathways are involved in disease progression. Thus, we further explored target genes within the eQTL-corresponding brain tissues using the VIF regression algorithm, which provided a list of prioritized genes. With all identified genes in the brain, we performed pathway analysis. The most relevant pathway of eIF2 signaling was suggested. eIF2 is a multimeric protein consisting of α, β, and γ subunits, and it is generally considered to affect the maintenance of a rate-limiting step in mRNA translation.[31] eIF2 signaling has important roles in the pathogenesis of SCZ as the corresponding stressors (starvation, virus, cytokines, and oxidative and endoplasmic reticulum stress) activate eIF2α kinases, which ultimately suppress protein synthesis through a series of reactions of phosphorylated eIF2-alpha.[32] The second significant pathway identified was IGF1 signaling. IGF1, insulin-like growth factors 1, is a multifunctional protein whose amino terminus is highly homologous to the insulin B chain, which makes it possible to promote the consumption of glucose in adipose tissue via the insulin/IGF1 axis.[33] Previous studies on IGF1 signaling in human neuroblastoma cells demonstrated that IGF1 signaling is involved in SCZ, as the pharmacological stimulation of muscarinic and insulin/IGF1 receptors reverses the expression levels of the specific subunits of disordered genes in SCZ.[34] Another critical pathway including 14-3-3 proteins was also identified, which is a family of highly conserved, multifunctional proteins highly expressed in the brain during development. Moreover, many studies have examined the 14-3-3 family gene and protein expression in the brain of patients with SCZ, and 14-3-3 proteins include seven isoforms, β, ε, γ, η, σ, θ, and ζ;[35,36] however, conflicting results have been obtained, and which isoform plays a role in SCZ remains to be elucidated.[16,35–38] Our results suggested that the isoforms of ε and θ have essential roles in SCZ, although other isoforms required more data to validate.

There are some limitations to the present analysis that need to be acknowledged and addressed. First, in addition to SNPs, other variants, such as copy number variation (CNV) and chromosomal aberration, may contribute to gene expression alteration and SCZ. In the present study, GTEx project V6p eQTL only provides the full data about the SNPs. If more data about other variants were available, the data could be used in the mRMR analysis. Second, as pointed out in Chou and Shen[39] and demonstrated in a series of recent publications[40–72], user-friendly and publicly accessible web servers represent the future direction for developing practically more useful prediction methods and computational tools. Actually, many practically useful web servers have increasing impacts on medical science,[73] driving medicinal chemistry into an unprecedented revolution.[74] We shall make efforts in our future work to provide a web server for the prediction method presented in this paper. (Once the web server has been established, an announcement will be made in the official website of Bio-X Institutes and via the MTNA journal.) Last, it is widely considered that environmental factors and genetic factors work together to induce many diseases, including SCZ. Gene expression is the direct result from environmental and genetic factors. Although here we focus on the integration of data from eQTL and GWAS to identify SCZ risk genes, we do not identify those genes associated with a certain environmental condition, and further studies are required to explore the specific genes for an environmental factor related to SCZ.

Conclusions

Our analysis by integrating the data from brain eQTL and GWAS of SCZ using the mRMR algorithm has indicated that cerebellum may play a crucial role in the pathogenesis of SCZ. Also, *ITIH4* may be utilized as a clinical biomarker for the diagnosis of SCZ, since its

quantity has been observed significantly decreased in the serum. Furthermore, three major pathways, i.e., EIF2 signaling, IGF-1 signaling, and 14-3-3-mediated signaling, have been identified to confer risk of SCZ. Further in-depth studies, both experimental and theoretical, are needed to reveal the molecular mechanism of such important findings.

MATERIALS AND METHODS

Benchmark Dataset

According to the 5-step rule[75] widely used in performing various genome or proteome analyses[40–44,76–90], the first important thing is to construct or select an effective benchmark dataset.

In this study, the raw eQTL data were taken from 10 human brain tissues, i.e., anterior cingulate cortex BA24, caudate basal ganglia, cerebellar hemisphere, cerebellum, cortex, frontal cortex BA9, hippocampus, hypothalamus, nucleus accumbens basal ganglia, and putamen basal ganglia, from the GTEx and association information on SNPs was taken from the genome-wide meta-analysis about SCZ.[9,10] In this meta-analysis,[10] a total of 36,989 cases with SCZ and 113,075 healthy controls was considered, and p values of a total of 9,444,231 SNPs were calculated for their genetic association with SCZ. The GTEx project (V6p eQTL) (https://gtexportal.org/home/datasets), which is currently the most massive eQTL project including the gene expression and genotype data of 53 normal human tissues from 544 donors, provides the association p values for SNPs regulating the gene expression.[9] The p value for each SNP-gene pair in GTEx databases was transformed into $-\log 10(p \text{ value})$. Then, when a variant had no significant effects on gene expression, the $-\log 10(p \text{ value})$ was set to be 0, i.e., p value of a corresponding SNP = 1. When a variant had significant effects on gene expression, i.e., eQTL, the $-\log 10(p \text{ value})$ for this eQTL was more than 0. Moreover, those eQTLs that were significantly associated with SCZ were classified as positive eSNPs, and those that were not were referred to as negative eSNPs. Since the number of negative eSNPs was much higher than that of the positive eSNP set, we randomly selected 10,000 negative eSNP sets, each of which matched the number of the total positive eSNP. Then, a benchmark dataset was constructed by the total positive eSNPs and each randomly selected negative eSNP set with the same number. Thus, overall, there were 10,000 eSNP benchmark datasets.

Based on each benchmark dataset, an eSNP-gene matrix was constructed for the next analysis. In this matrix, the rows were eSNPs, whereas the columns were class of eSNP, i.e., positive or negative ones, and genes regulated by the eSNPs from the 10 brain tissues. Totally, 22,832 eQTL genes were included in this matrix for each eSNP.

The mRMR Method Integrating Brain eQTL and GWASs

The mRMR algorithm has been widely used in computational biology for genome and proteome analyses.[41,91–95] Here we also used the mRMR approach to identify the potential eQTL genes for SCZ by calculating the MI between two features and ranking these features.[11]

Given two variables x and y, their MI value can be calculated according to the following equation:

$$I(x,y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy, \quad (1)$$

where $p(x)$ and $p(y)$ are the marginal probabilities of x and y; and $p\ (x,\ y)$ is their joint probabilistic distribution. Using the value of MI, the distance between two variables can also be quantitatively measured. Based on the definition of MI, the MaxRel distance can be formulated as the distance between a given feature and the target classes, which reflects the relevance between the eQTL gene features from 10 brain tissues and positive eSNPs. A larger MaxRel score, which is highly interpretative and can reveal the difference between target classes, is indicative of a stronger relevance. Since there were 10,000 benchmark datasets, there were 10,000 MaxRel scores for each eQTL gene. We ranked the eQTL genes based on both the MaxRel scores for each benchmark dataset and the average of the MaxRel scores for all tested benchmark datasets.

Furthermore, the identified candidate genes were evaluated by searching more evidence for them as potential SCZ risk genes in the SZDB database (http://www.szdb.org/). In this database,[96] SCZ risk genes reaching the genome-wide significance level were extracted from multiple GWASs and 5 microarray datasets, including GEO: GSE53987 (114 samples of prefrontal cortex, striatum, and hippocampus),[97] GEO: GSE12649 (69 post mortem samples of prefrontal cortex),[98] GEO: GSE21138 (59 postmortem samples of prefrontal cortex s),[99] GEO: GSE35978 (195 samples of cerebellum and parietal cortex brain),[100] and GEO: GSE62191 (59 samples of frontal cortex)[101] (https://www.ncbi.nlm.nih.gov/geo/). Moreover, BrainCloud eQTL database,[102] which contains the eQTL data from the human post mortem dorsolateral prefrontal cortex (DLPFC) of 261 normal human subjects in Caucasians and African Americans, was used for replication analysis of the eQTL association.

Identify the Potential Interactors of SCZ Risk Genes in the Brain

To identify the target interacting genes of each eQTL gene, the VIF regression algorithm, an efficient and accurate method, was adopted.[103] The objective of this algorithm used here was to select the optimal genes as interactors that can fit the expression pattern of the interesting genes. We tried to identify the optimal that could minimize the penalized sum of squared errors, l_0, using the algorithm represented by the following equation:

$$\underset{\beta}{\operatorname{argmin}} \left\{ \| y - X\beta \|_2^2 + \lambda_0 \| \beta \|_{l_0} \right\}, \quad (2)$$

where $\mathrm{y} = (y_1, \ldots, y_n)^{'}$ are n observations of the target gene, $X = (X_1, \ldots, X_p)$ are p interactors, $\| \beta \|_{l_0} = \sum_{i=1}^{p} I_{\{\beta \neq 0\}}$. This algorithm calculates the correlations of each candidate interactor with the interesting genes using a small presampled dataset, and it searches the optimal interactor subset by applying t-statistic with a correction procedure when adding or removing one interactor at a time. The

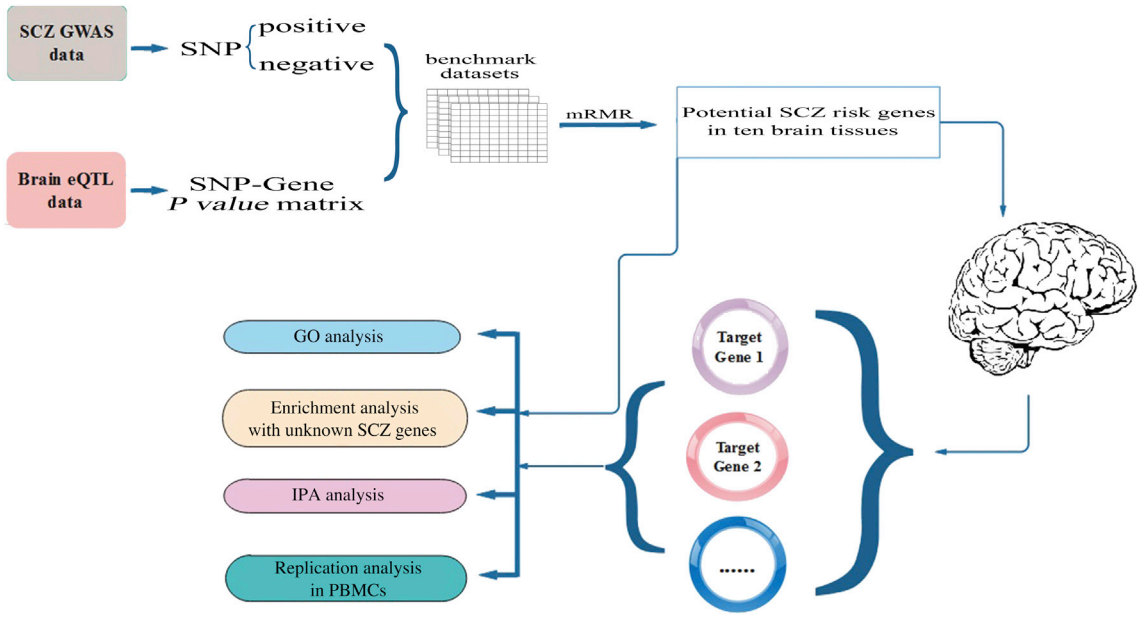

Figure 4. Flow Chart Detailing the Inclusion Process to the Present Study

R package http://cran.r-project.org/web/packages/VIF/ was used to implement the VIF method.

Furthermore, to assess the goodness of fit for VIF regression models, we calculated the adjusted coefficient of determination, also known as adjusted R^2,[104] which measures how well the regression model fits the real data points and considers the number of interactors that have been used. In the present study, the regression models with adjusted R^2 values greater than 0.6 were considered. The scheme for the exploration of candidate SCZ genes is shown in Figure 4.

Enrichment Analysis

To gain a better understanding of the biological effects of all the identified genes, we performed GO enrichment analysis.[105] Using a hypergeometric test, we analyzed whether all the above genes, including eQTL genes and their interactors, significantly overlapped certain GO terms.[106] For each specific GO gene set, the hypergeometric test p value was caudated as

$$P = \sum_{k=m}^{n} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \quad (3)$$

where N is the number of all human genes, M is the number of GO genes, n is the number of interesting genes, and m is the number of interesting genes that are GO disease genes. To control the FDR, the p values of the hypergeometric test were adjusted with the Benjamini-Hochberg method.[107]

Furthermore, not only the overlap with GO but also the overlap with the reported SCZ genes was evaluated. The known SCZ genes reported by GWASs and genes expressed in brain tissues are listed in Table S1. In addition, we identified the canonical pathways associated with these SCZ candidate genes using the Ingenuity Pathway Analysis (IPA) suite (https://www.qiagenbioinformatics.com/). In canonical pathway-based analysis, the criteria for involved significant pathways was set as $-\log p > 2$.

Systematic Review of 14-3-3 Isoforms Associated with SCZ

Further, to determine the association of 14-3-3 isoforms with SCZ, we performed an updated systematic review with a literature search of studies published between January 1990 and December 2017 in six English-language databases (PubMed, Embase, Web of Science, ScienceDirect, SpringerLink, and EBSCO) and two Chinese databases (Wanfang and Chinese National Knowledge Infrastructure databases). The following keywords were used: 14-3-3 or YWHA and SCZ. The scheme for this systematic review is described in Figure S1.

Data extraction was independently performed by two investigators; any discrepancies between the two reviewers were resolved through discussion, and a consensus was reached by a third party who was from a different organization. Inclusion criteria for the analysis were as follows: (1) detailed diagnosis definition of SCZ; (2) sample size, FC, and p value; and (3) at least three qualifying studies per isoform. The strength of the associations between gene expression levels and SCZ was measured by calculating the FC and p value.

Potential Candidate Genes for Diagnosis

To identify the potential candidate genes for the blood test, the gene expression profile of PBMCs was examined in our previous study.[108]

Briefly, blood samples from 18 first-onset SCZ patients (8 males and 10 females, aged 14.78 ± 1.70 years) and 12 healthy controls (6 males and 6 females, aged 14.75 ± 2.14 years) were collected. The patients were untreated and drug naive and were independently diagnosed by at least two experienced psychiatrists according to the Diagnosis and Statistical Manual of Mental Disorders Fourth Edition (DSM-IV) criteria for SCZ. Agilent Human LncRNA Microarray v.2.0 and 17,200 valid probes were used to identify the putative clinical gene biomarkers. All participants have provided informed consent in accordance with the approval of the Bioethics Committee of Bio-X Institutes of Shanghai Jiaotong University and the principles set forth by the Declaration of Helsinki.

SUPPLEMENTAL INFORMATION

Supplemental Information includes five tables and one figure and can be found with this article online at https://doi.org/10.1016/j.omtn.2018.05.026.

AUTHOR CONTRIBUTIONS

L.C., L.H., and K.-C.C. conceived the study. L.C., T.H., and J.S. designed and performed the analyses. L.C., T.H., and X.Z. drafted the manuscript. W.C. and F.Z. provided the data and performed gene expression tests. L.C., and K.-C.C. finalized the paper.

CONFLICTS OF INTEREST

The authors have no conflict of interest.

ACKNOWLEDGMENTS

The authors wish to thank the two anonymous reviewers for their constructive comments and Ms. Wen Wen for her preparing the first figure. The authors also wish to thank other participants in this study as well as the Shanghai Key Laboratory of Psychotic Disorders (13dz2260500). This work was supported by Ministry of Science and Technology Project (2017YFC1001302, 2016YFC0906400, and 2017YFC09092), the Grant of Shanghai Brain-Intelligence Project from STCSM (16JC1420500), Shanghai Jiaotong University Medical Engineering Cross Research Foundation (YG2014MS07), National Natural Science Foundation of China (31701151), and Shanghai Sailing Program from Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245).

REFERENCES

1. Cai, L., Chen, T., Yang, J., Zhou, K., Yan, X., Chen, W., Sun, L., Li, L., Qin, S., Wang, P., et al. (2015). Serum trace element differences between Schizophrenia patients and controls in the Han Chinese population. Sci. Rep. *5*, 15013.

2. Cai, L., Yang, Y.H., He, L., and Chou, K.C. (2016). Modulation of Cytokine Network in the Comorbidity of Schizophrenia and Tuberculosis. Curr. Top. Med. Chem. *16*, 655–665.

3. Flint, J., and Munafò, M. (2014). Schizophrenia: genesis of a complex disease. Nature *511*, 412–413.

4. Huang, T., Liu, C.L., Li, L.L., Cai, M.H., Chen, W.Z., Xu, Y.F., O'Reilly, P.F., Cai, L., and He, L. (2016). A new method for identifying causal genes of schizophrenia and anti-tuberculosis drug-induced hepatotoxicity. Sci. Rep. *6*, 32571.

5. Jansen, R.C., and Nap, J.P. (2001). Genetical genomics: the added value from segregation. Trends Genet. *17*, 388–391.

6. Xu, Y., Yue, W., Yao Shugart, Y., Li, S., Cai, L., Li, Q., Cheng, Z., Wang, G., Zhou, Z., Jin, C., et al. (2016). Exploring Transcription Factors-microRNAs Co-regulation Networks in Schizophrenia. Schizophr. Bull. *42*, 1037–1045.

7. Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. Nat. Rev. Genet. *10*, 184–194.

8. Kim, Y., Xia, K., Tao, R., Giusti-Rodriguez, P., Vladimirov, V., van den Oord, E., and Sullivan, P.F. (2014). A meta-analysis of gene expression quantitative trait loci in brain. Transl. Psychiatry *4*, e459.

9. Ardlie, K.G., DeLuca, D.S., Segre, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., Lek, M., et al.; GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

10. Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427.

11. Ding, C., and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. *3*, 185–205.

12. Zhang, N., Zhou, Y., Huang, T., Zhang, Y.C., Li, B.Q., Chen, L., and Cai, Y.D. (2014). Discriminating between lysine sumoylation and lysine acetylation using mRMR feature selection and analysis. PLoS ONE *9*, e107464.

13. Zhou, Y., Zhang, N., Li, B.Q., Huang, T., Cai, Y.D., and Kong, X.Y. (2015). A method to distinguish between lysine acetylation and lysine ubiquitination with feature selection and analysis. J. Biomol. Struct. Dyn. *33*, 2479–2490.

14. Zhang, Y., Xu, J., Zheng, W., Zhang, C., Qiu, X., Chen, K., and Ruan, J. (2014). newDNA-Prot: Prediction of DNA-binding proteins by employing support vector machine and a comprehensive sequence representation. Comput. Biol. Chem. *52*, 51–59.

15. Lin, S., Lin, Y., Nery, J.R., Urich, M.A., Breschi, A., Davis, C.A., Dobin, A., Zaleski, C., Beer, M.A., Chapman, W.C., et al. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. Proc. Natl. Acad. Sci. USA *111*, 17224–17229.

16. Martins-de-Souza, D., Gattaz, W.F., Schmitt, A., Rewerts, C., Marangoni, S., Novello, J.C., Maccarrone, G., Turck, C.W., and Dias-Neto, E. (2009). Alterations in oligodendrocyte proteins, calcium homeostasis and new potential markers in schizophrenia anterior temporal lobe are revealed by shotgun proteome analysis. J. Neural Transm. (Vienna) *116*, 275–289.

17. Yu, H., Yan, H., Li, J., Li, Z., Zhang, X., Ma, Y., Mei, L., Liu, C., Cai, L., Wang, Q., et al.; Chinese Schizophrenia Collaboration Group (2017). Common variants on 2p16.1, 6p22.1 and 10q24.32 are associated with schizophrenia in Han Chinese population. Mol. Psychiatry *22*, 954–960.

18. Shi, Y., Li, Z., Xu, Q., Wang, T., Li, T., Shen, J., Zhang, F., Chen, J., Zhou, G., Ji, W., et al. (2011). Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. Nat. Genet. *43*, 1224–1227.

19. Yue, W.H., Wang, H.F., Sun, L.D., Tang, F.L., Liu, Z.H., Zhang, H.X., Li, W.Q., Zhang, Y.L., Zhang, Y., Ma, C.C., et al. (2011). Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2. Nat. Genet. *43*, 1228–1231.

20. Ripke, S., Sanders, A.R., Kendler, K.S., Levinson, D.F., Sklar, P., Holmans, P.A., Lin, D.-Y., Duan, J., Ophoff, R.A., Andreassen, O.A., et al.; Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011). Genome-wide association study identifies five new schizophrenia loci. Nat. Genet. *43*, 969–976.

21. Song, J., Wang, H., Wang, J., Leier, A., Marquez-Lago, T., Yang, B., Zhang, Z., Akutsu, T., Webb, G.I., and Daly, R.J. (2017). PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. Sci. Rep. *7*, 6862.

22. Wang, H., Feng, L., Zhang, Z., Webb, G.I., Lin, D., and Song, J. (2016). Crysalis: an integrated server for computational analysis and design of protein crystallization. Sci. Rep. *6*, 21383.

23. Li, F., Li, C., Wang, M., Webb, G.I., Zhang, Y., Whisstock, J.C., and Song, J. (2015). GlycoMine: a machine learning-based approach for predicting N-, C- and O-linked glycosylation in the human proteome. Bioinformatics *31*, 1411–1419.

24. Qin, W., Li, Y., Li, J., Yu, L., Wu, D., Jing, R., Pu, X., Guo, Y., and Li, M. (2012). Predicting deleterious non-synonymous single nucleotide polymorphisms in signal peptides based on hybrid sequence attributes. Comput. Biol. Chem. *36*, 31–35.

25. Yeganeh-Doost, P., Gruber, O., Falkai, P., and Schmitt, A. (2011). The role of the cerebellum in schizophrenia: from cognition to molecular pathways. Clinics (São Paulo) *66* (*Suppl 1*), 71–77.

26. Sekar, A., Bialas, A.R., de Rivera, H., Davis, A., Hammond, T.R., Kamitaki, N., Tooley, K., Presumey, J., Baum, M., Van Doren, V., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2016). Schizophrenia risk from complex variation of complement component 4. Nature *530*, 177–183.

27. Hakobyan, S., Boyajyan, A., and Sim, R.B. (2005). Classical pathway complement activity in schizophrenia. Neurosci. Lett. *374*, 35–37.

28. Piñeiro, M., Alava, M.A., González-Ramón, N., Osada, J., Lasierra, P., Larrad, L., Piñeiro, A., and Lampreave, F. (1999). ITIH4 serum concentration increases during acute-phase processes in human patients and is up-regulated by interleukin-6 in hepatocarcinoma HepG2 cells. Biochem. Biophys. Res. Commun. *263*, 224–229.

29. Mohammadi, A., Rashidi, E., and Amooeian, V.G. (2018). Brain, blood, cerebrospinal fluid, and serum biomarkers in schizophrenia. Psychiatry Res. *265*, 25–38.

30. Cai, L., Cai, M.-H., Wang, M.-Y., Xu, Y.-F., Chen, W.-Z., Qin, S.-Y., Wan, C.-L., and He, L. (2015). Meta-Analysis-Based Preliminary Exploration of the Connection between ATDILI and Schizophrenia by GSTM1/T1 Gene Polymorphisms. PLoS One *10*, e0128643.

31. Kimball, S.R. (1999). Eukaryotic initiation factor eIF2. Int. J. Biochem. Cell Biol. *31*, 25–29.

32. Carter, C.J. (2007). eIF2B and oligodendrocyte survival: where nature and nurture meet in bipolar disorder and schizophrenia? Schizophr. Bull. *33*, 1343–1353.

33. L-López, F., Sarmento-Cabral, A., Herrero-Aguayo, V., Gahete, M.D., Castaño, J.P., and Luque, R.M. (2017). Obesity and metabolic dysfunction severely influence prostate cell function: role of insulin and IGF1. J. Cell. Mol. Med. *21*, 1893–1904.

34. Altar, C.A., Hunt, R.A., Jurata, L.W., Webster, M.J., Derby, E., Gallagher, P., Lemire, A., Brockman, J., and Laeng, P. (2008). Insulin, IGF-1, and muscarinic agonists modulate schizophrenia-associated genes in human neuroblastoma cells. Biol. Psychiatry *64*, 1077–1087.

35. Qing, Y., Sun, L., Yang, C., Jiang, J., Yang, X., Hu, X., Cui, D., Xu, Y., He, L., Han, D., and Wan, C. (2016). Dysregulated 14-3-3 Family in Peripheral Blood Leukocytes of Patients with Schizophrenia. Sci. Rep. *6*, 23791.

36. Wong, A.H., Likhodi, O., Trakalo, J., Yusuf, M., Sinha, A., Pato, C.N., Pato, M.T., Van Tol, H.H., and Kennedy, J.L. (2005). Genetic and post-mortem mRNA analysis of the 14-3-3 genes that encode phosphoserine/threonine-binding regulatory proteins in schizophrenia and bipolar disorder. Schizophr. Res. *78*, 137–146.

37. English, J.A., Dicker, P., Föcking, M., Dunn, M.J., and Cotter, D.R. (2009). 2-D DIGE analysis implicates cytoskeletal abnormalities in psychiatric disease. Proteomics *9*, 3368–3382.

38. Saia-Cereda, V.M., Santana, A.G., Schmitt, A., Falkai, P., and Martins-de-Souza, D. (2017). The Nuclear Proteome of White and Gray Matter from Schizophrenia Postmortem Brains. Mol. Neuropsychiatry *3*, 37–52.

39. Chou, K.C., and Shen, H.B. (2009). Recent advances in developing web-servers for predicting protein attributes. Nat. Sci. *1*, 63–92.

40. Chen, W., Feng, P.M., Lin, H., and Chou, K.C. (2013). iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. *41*, e68.

41. Lin, H., Deng, E.Z., Ding, H., Chen, W., and Chou, K.C. (2014). iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. *42*, 12961–12972.

42. Chen, W., Tang, H., Ye, J., Lin, H., and Chou, K.C. (2016). iRNA-PseU: Identifying RNA pseudouridine sites. Mol. Ther. Nucleic Acids *5*, e332.

43. Feng, P., Ding, H., Yang, H., Chen, W., Lin, H., and Chou, K.C. (2017). iRNA-PseColl: Identifying the occurrence sites of different RNA modifications by incorporating collective effects of nucleotides into PseKNC. Mol. Ther. Nucleic Acids *7*, 155–163.

44. Liu, B., Yang, F., and Chou, K.C. (2017). 2L-piRNA: A two-layer ensemble classifier for identifying piwi-interacting RNAs and their function. Mol. Ther. Nucleic Acids *7*, 267–277.

45. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2015). iPPI-Esml: An ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theor. Biol. *377*, 47–56.

46. Liu, B., Wang, S., Long, R., and Chou, K.C. (2017). iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics *33*, 35–41.

47. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition. J. Biomol. Struct. Dyn. *34*, 1946–1961.

48. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal. Biochem. *497*, 48–56.

49. Liu, L.M., Xu, Y., and Chou, K.C. (2017). iPGK-PseAAC: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general PseAAC. Med. Chem. *13*, 552–559.

50. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, Z.C., Jia, J.H., and Chou, K.C. (2017). iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier. Genomics, S0888-7543(17)30138-6.

51. Xu, Y., Wang, Z., Li, C., and Chou, K.C. (2017). iPreny-PseAAC: identify C-terminal cysteine prenylation sites in proteins by incorporating two tiers of sequence couplings into PseAAC. Med. Chem. *13*, 544–551.

52. Feng, P., Yang, H., Ding, H., Lin, H., Chen, W., and Chou, K.C. (2018). iDNA6mA-PseKNC: Identifying DNA N^6-methyladenosine sites by incorporating nucleotide physicochemical properties into PseKNC. Genomics, S0888-7543(18)30009-0.

53. Liu, B., Yang, F., Huang, D.S., and Chou, K.C. (2018). iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. Bioinformatics *34*, 33–40.

54. Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., and Chou, K.C. (2015). Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE *10*, e0121501.

55. Liu, B., Fang, L., Long, R., Lan, X., and Chou, K.C. (2016). iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics *32*, 362–369.

56. Cheng, X., Xiao, X., and Chou, K.C. (2017). pLoc-mPlant: predict subcellular localization of multi-location plant proteins by incorporating the optimal GO information into general PseAAC. Mol. Biosyst. *13*, 1722–1727.

57. Cheng, X., Xiao, X., and Chou, K.C. (2017). pLoc-mVirus: predict subcellular localization of multi-location virus proteins via incorporating the optimal GO information into general PseAAC. Gene *628*, 315–321.

58. Cheng, X., Xiao, X., and Chou, K.C. (2018). pLoc-mEuk: Predict subcellular localization of multi-label eukaryotic proteins by extracting the key GO information into general PseAAC. Genomics *110*, 50–58.

59. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J. Theor. Biol. *394*, 223–230.

60. Cheng, X., Xiao, X., and Chou, K.C. (2017). pLoc-mGneg: Predict subcellular localization of Gram-negative bacterial proteins by deep gene ontology learning via general PseAAC. Genomics, S0888-7543(17)30102-7.

61. Xiao, X., Ye, H.X., Liu, Z., Jia, J.H., and Chou, K.C. (2016). iROS-gPseKNC: Predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget *7*, 34180–34189.

62. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iCar-PseCp: identify carbonylation sites in proteins by Monte Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget *7*, 34558–34570.

63. Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., Deng, N.Y., and Chou, K.C. (2014). iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE *9*, e105018.

64. Cheng, X., Zhao, S.G., Lin, W.Z., Xiao, X., and Chou, K.C. (2017). pLoc-mAnimal: predict subcellular localization of animal proteins with both single and multiple sites. Bioinformatics *33*, 3524–3531.

65. Jia, J., Zhang, L., Liu, Z., Xiao, X., and Chou, K.C. (2016). pSumo-CD: predicting sumoylation sites in proteins with covariance discriminant algorithm by incorporating sequence-coupled effects into general PseAAC. Bioinformatics *32*, 3133–3141.

66. Liu, Z., Xiao, X., Yu, D.J., Jia, J., Qiu, W.R., and Chou, K.C. (2016). pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. Anal. Biochem. *497*, 60–67.

67. Xiao, X., Cheng, X., Su, S., Mao, Q., and Chou, K.C. (2017). pLoc-mGpos: incorporate key gene ontology information into general PseAAC for predicting subcellular localization of Gram-positive bacterial proteins. Nat. Sci. *9*, 330–349.

68. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2017). iRNA-AI: identifying the adenosine to inosine editing sites in RNA sequences. Oncotarget *8*, 4208–4217.

69. Cheng, X., Zhao, S.G., Xiao, X., and Chou, K.C. (2017). iATC-mISF: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinformatics *33*, 341–346.

70. Jia, J., Liu, Z., Xiao, X., Liu, B., and Chou, K.C. (2016). iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules *21*, E95.

71. Qiu, W.R., Jiang, S.Y., Xu, Z.C., Xiao, X., and Chou, K.C. (2017). iRNAm5C-PseDNC: identifying RNA 5-methylcytosine sites by incorporating physical-chemical properties into pseudo dinucleotide composition. Oncotarget *8*, 41178–41188.

72. Qiu, W.R., Sun, B.Q., Xiao, X., Xu, D., and Chou, K.C. (2017). iPhos-PseEvo: identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory. Mol. Inform. *36*, 1600010.

73. Chou, K.C. (2015). Impacts of bioinformatics to medicinal chemistry. Med. Chem. *11*, 218–234.

74. Chou, K.C. (2017). An unprecedented revolution in medicinal chemistry driven by the progress of biological science. Curr. Top. Med. Chem. *17*, 2337–2358.

75. Chou, K.C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. *273*, 236–247.

76. Ju, Z., Cao, J.Z., and Gu, H. (2016). Predicting lysine phosphoglycerylation with fuzzy SVM by incorporating k-spaced amino acid pairs into Chou's general PseAAC. J. Theor. Biol. *397*, 145–150.

77. Meher, P.K., Sahu, T.K., Saini, V., and Rao, A.R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. Sci. Rep. *7*, 42362.

78. Huo, H., Li, T., Wang, S., Lv, Y., Zuo, Y., and Yang, L. (2017). Prediction of presynaptic and postsynaptic neurotoxins by combining various Chou's pseudo components. Sci. Rep. *7*, 5827.

79. Tripathi, P., and Pandey, P.N. (2017). A novel alignment-free method to classify protein folding types by combining spectral graph clustering with Chou's pseudo amino acid composition. J. Theor. Biol. *424*, 49–54.

80. Zhang, S., and Duan, X. (2018). Prediction of protein subcellular localization with oversampling approach and Chou's general PseAAC. J. Theor. Biol. *437*, 239–250.

81. Arif, M., Hayat, M., and Jan, Z. (2018). iMem-2LSAAC: A two-level model for discrimination of membrane proteins and their types by extending the notion of SAAC into chou's pseudo amino acid composition. J. Theor. Biol. *442*, 11–21.

82. Mei, J., and Zhao, J. (2018). Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. Sci. Rep. *8*, 2359.

83. Zhang, L., and Kong, L. (2018). iRSpot-ADPM: Identify recombination spots by incorporating the associated dinucleotide product model into Chou's pseudo components. J. Theor. Biol. *441*, 1–8.

84. Chen, W., Feng, P., Yang, H., Ding, H., Lin, H., and Chou, K.C. (2018). iRNA-3typeA: identifying three types of modification at RNA's adenosine sites. Mol. Ther. Nucleic Acids *11*, 468–474.

85. Cheng, X., Xiao, X., and Chou, K.C. (2018). pLoc-mHum: predict subcellular localization of multi-location human proteins via general PseAAC to winnow out the crucial GO information. Bioinformatics *34*, 1448–1456.

86. Khan, Y.D., Rasool, N., Hussain, W., Khan, S.A., and Chou, K.C. (2018). iPhosT-PseAAC: Identify phosphothreonine sites by incorporating sequence statistical moments into PseAAC. Anal. Biochem. *550*, 109–116.

87. Liu, B., Weng, F., Huang, D.S., and Chou, K.C. (2018). iRO-3wPseKNC: Identify DNA replication origins by three-window-based PseKNC. Bioinformatics. Published online April 19, 2018. https://doi.org/10.1093/bioinformatics/bty312.

88. Song, J., Li, F., Takemoto, K., Haffari, G., Akutsu, T., Chou, K.C., and Webb, G.I. (2018). PREvaIL, an integrative approach for inferring catalytic residues using sequence, structural, and network features in a machine-learning framework. J. Theor. Biol. *443*, 125–137.

89. Song, J., Wang, Y., Li, F., Akutsu, T., Rawlings, N.D., Webb, G.I., and Chou, K.C. (2018). iProt-Sub: a comprehensive package for accurately mapping and predicting protease-specific substrates and cleavage sites. Brief. Bioinform. , Published online April 19, 2018. https://doi.org/10.1093/bib/bby028.

90. Yang, H., Qiu, W.R., Liu, G., Guo, F.B., Chen, W., Chou, K.C., and Lin, H. (2018). iRSpot-Pse6NC: Identifying recombination spots in Saccharomyces cerevisiae by incorporating hexamer composition into general PseKNC. Int. J. Biol. Sci. *14*, 883–891.

91. Hu, L., Huang, T., Shi, X., Lu, W.C., Cai, Y.D., and Chou, K.C. (2011). Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties. PLoS ONE *6*, e14556.

92. Huang, T., Niu, S., Xu, Z., Huang, Y., Kong, X., Cai, Y.D., and Chou, K.C. (2011). Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. PLoS ONE *6*, e22940.

93. Wang, P., Hu, L., Liu, G., Jiang, N., Chen, X., Xu, J., Zheng, W., Li, L., Tan, M., Chen, Z., et al. (2011). Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS ONE *6*, e18476.

94. Li, B.Q., Huang, T., Liu, L., Cai, Y.D., and Chou, K.C. (2012). Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. PLoS ONE *7*, e33393.

95. Zheng, L.L., Li, Y.X., Ding, J., Guo, X.K., Feng, K.Y., Wang, Y.J., Hu, L.L., Cai, Y.D., Hao, P., and Chou, K.C. (2012). A comparison of computational methods for identifying virulence factors. PLoS ONE *7*, e42517.

96. Wu, Y., Yao, Y.G., and Luo, X.J. (2017). SZDB: A Database for Schizophrenia Genetic Research. Schizophr. Bull. *43*, 459–471.

97. Lanz, T.A., Joshi, J.J., Reinhart, V., Johnson, K., Grantham, L.E., 2nd, and Volfson, D. (2015). 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 *10*, e0121744.

98. Iwamoto, K., Bundo, M., and Kato, T. (2005). 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. *14*, 241–253.

99. Narayan, S., Tang, B., Head, S.R., Gilmartin, T.J., Sutcliffe, J.G., Dean, B., and Thomas, E.A. (2008). Molecular profiles of schizophrenia in the CNS at different stages of illness. Brain Res. *1239*, 235–248.

100. Chen, C., Cheng, L., Grennan, K., Pibiri, F., Zhang, C., Badner, J.A., Gershon, E.S., and Liu, C.; Members of the Bipolar Disorder Genome Study (BiGS) Consortium (2013). Two gene co-expression modules differentiate psychotics and controls. Mol. Psychiatry *18*, 1308–1314.

101. de Baumont, A., Maschietto, M., Lima, L., Carraro, D.M., Olivieri, E.H., Fiorini, A., Barreta, L.A., Palha, J.A., Belmonte-de-Abreu, P., Moreira Filho, C.A., and Brentani, H. (2015). Innate immune response is differentially dysregulated between bipolar disease and schizophrenia. Schizophr. Res. *161*, 215–221.

102. BrainSeq: A Human Brain Genomics Consortium. Electronic address: drweinberger@libd.org; BrainSeq: A Human Brain Genomics Consortium (2015). BrainSeq: Neurogenomics to Drive Novel Target Discovery for Neuropsychiatric Disorders. Neuron *88*, 1078–1083.

103. Lin, D., Foster, D.P., and Ungar, L.H. (2011). VIF Regression: A Fast Regression Algorithm for Large Data. J. Am. Stat. Assoc. *106*, 232–247.

104. Chen, L., Zhang, Y.H., Zheng, M., Huang, T., and Cai, Y.D. (2016). Identification of compound-protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds. Mol. Genet. Genomics *291*, 2065–2079.

105. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

106. Cai, L., Pan, H., Trzciński, K., Thompson, C.M., Wu, Q., and Kramnik, I. (2010). MYBBP1A: a new Ipr1's binding protein in mice. Mol. Biol. Rep. *37*, 3863–3868.

107. Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Series B *57*, 289–300.

108. Sun, L., Cheng, Z., Zhang, F., and Xu, Y. (2015). Gene expression profiling in peripheral blood mononuclear cells of early-onset schizophrenia. Genom. Data *5*, 169–170.