BrainEXP-NPD: a database of transcriptomic profiles of human brains of six neuropsychiatric disorders

Cuihua Xia[1], Teng Ma[2], Chuan Jiao[1], Chao Chen[1,4], Chunyu Liu[1,3,5]

[1] Center for Medical Genetics, School of Life Science, Central South University, Changsha, China
[2] Department of Anatomy and Neurobiology, School of Basic Medical Sciences, Central South University, Changsha, China.
[3] Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA.
[4] National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University, Changsha, China.
[5] School of Psychology, Shaanxi Normal University, Xi'an, Shaanxi, China.
Corresponding author: Chunyu Liu, Email: liuch@upstate.edu ; Chao Chen, Email: chenchao@sklmg.edu.cn

Abstract

Background: Spatio-temporal gene expression has been widely used to study gene functions and biological mechanisms in diseases. Numerous microarray and RNA sequencing data focusing on brain transcriptomes in neuropsychiatric disorders have accumulated. However, their consistency, reproducibility has not been properly evaluated. Except for a few psychiatric disorders, like schizophrenia, bipolar disorder and autism, most have not been compared to each other for cross-disorder comparisons.

Methods: We organized 48 human brain transcriptome datasets from six sources. The original brain donors include patients with schizophrenia (SCZ, N=427), bipolar disorder (BD, N=312), major depressive disorder (MDD, N=219), autism spectrum disorder (ASD, N=53), Alzheimer's disease (AD, N=765), Parkinson's disease (PD, N=163) as well as controls as unaffected by such disorders (CTRL, N=6,378), making it a total of 8,317 samples. Raw data included multiple brain regions of both sexes, with ages ranging from embryonic to seniors. After standardization, quality control, filtering and removal of known and unknown covariates, we performed comprehensive meta- and mega- analyses, including gene differential expression and gene co-expression network.

Results: A total of 6922, 3011, 2703, 4389, 3507, 4279 significantly differentially expressed genes (FDR $q < 0.05$) were detected in the comparisons of 6 brain regions of SCZ-CTRL, 5 brain regions of BD-CTRL, 6 brain regions of MDD-CTRL, 4 brain regions of ASD-CTRL, 7 brain regions of AD-CTRL, and 6 brain regions of PD-CTRL, respectively. Most differentially expressed genes were brain region-specific and disease-specific. SCZ and BD have a maximal transcriptome similarity in striatum ($\rho=0.42$) among the four brain regions, as measured by Spearman's correlation of differential expression log2 FC values. SCZ and MDD have a maximal transcriptome similarity in hippocampus ($\rho=0.30$) among the five brain regions. BD and MDD have a maximal transcriptome similarity in frontal cortex ($\rho=0.45$) among the five brain regions. Other disease pairs have a less transcriptome similarity ($\rho<0.1$) in all brain regions. PD is negatively correlated with SCZ, BD, and MDD in

cerebellum and striatum. We also performed coexpression network analyses for different disorders and controls separately. We developed a database named BrainEXP-NPD (http://brainexpnpd.org:8088/BrainEXPNPD/), to provide a user-friendly web interface for accessing the data, and analytical results of meta- and mega- analyses, including gene differential expression and gene co-expression networks between cases and controls on different brain regions, sexes and age groups. Discussion: BrainEXP-NPD compiled the largest collection of brain transcriptomic data of major neuropsychiatric disorders and presented lists of differentially expressed genes and coexpression modules in multiple brain regions of six major disorders.

Keywords

Transcriptomic profiles; Neuropsychiatric disorder; Human brain; BrainEXP-NPD

Introduction

It is widely accepted that gene expression and gene-gene interaction are the basic tools to study and understand molecular normal and abnormal functions in organisms. Though genetic studies have revealed lots of disease risk variants, it still remains unclear how genetic factors impact on the disease. Transcriptome can help identify convergent molecular pathology[1].

A number of public databases have been developed for studying expression variations and co-expression patterns in neuropsychiatric disorders, such as SZDB, AlzData, SMRIDB. SZDB is a specific database which focuses on SNP and gene annotation, spatio-temporal expression pattern analysis, network (PPI and co-expression) and pathway analysis, brain eQTL analysis and gene prioritization of schizophrenia[2]. AlzData covers high-throughput omic data, including genomics, transcriptomics, proteomics and functional genomics of Alzheimer's disease[3]. However, both SZDB and AlzData only focus on a specific neuropsychiatric disease. SMRIDB (The Stanley Medical Research Institute Online Genomics Database) contains 988 samples from 12 studies and 6 platforms aimed at gene differential analysis and pathway/GO analysis of schizophrenia, bipolar, depression and healthy controls[4]. But cross-disorder analysis and co-expression analysis are not provided. Since these databases are constructed primarily on specific neuropsychiatric disorders or only gene expression, a more integrative data source and comprehensive research for neuropsychiatric disorders are urgently needed.

Table 1. Number of individuals across disease state on transcriptome per brain project

Name	# of Individuals (# of samples)	Age	# of Brain regions	Disease state (# of individuals)							
				CTRL	SCZ	BD	MDD	ASD	AD	PD	Others
BrainCloud	269	Lifespan	1	269							
BrainSpan	42	Lifespan	6	42							
UKBEC	134	Postnatal	6	134							
GTEx	714 (2418)	Adult	14	714							
BrainEXP	2863 (4567)	Lifespan	56	4,567							
CMC	621	Postnatal	1	291	275	47					8
BrainSeq	746	Lifespan	3	341	184	69	152				
ROSMAP	748	Adult	1	432					316		
SZDB2.0	537	Postnatal	1	279	258						
AlzData	1246		4	562					684		
SMRIDB	979		5								
PsychENCODE	1695	Postnatal	2	899	531	217		48			

Here we developed a database which focuses on six complex neuropsychiatric disorders' transcriptome profiles' analyses. We first collected totally 8,317 brain samples from GEO (Gene Expression Omnibus), ArrayExpress and our own lab. It consists of six complex neuropsychiatric disorders (427 schizophrenia, 312 bipolar disorder, 163 Parkinson's disease, 765 Alzheimer's disease, 219 major depressive disorder, 53 autism spectrum disorder) and 6,378 controls (no neuropsychiatric-related disorders) across 55 brain regions from 48 individual datasets and different platforms (microarrays and RNA-seq). Then a consistent process workflow was applied to each individual dataset, and a combined-data analysis was conducted to uncover gene variations and gene co-expression patterns between cases and controls across different brain regions, sexes and age stages. Besides, a weighted gene co-expression network analysis (WGCNA)[5] was carried out to figure out the disease-related modules and pathway enrichment. Finally, an integrative database of Transcriptomic Profiling in human Brains for six NeuroPsychiatric disorders (BrainEXP-NPD) was developed as a useful open free resource and reference for researchers and clinical workers.

Material and Methods

Study Design

The study was designed into three stages: Data Collection, Data Analysis and Visualization, as shown in Figure 1. We first collected the brain transcriptomic data from GEO[6], ArrayExpress[7], SMRIDB[4], GTEx[8], ROSMAP[9] and PsychENCODE[10]. And then analyses were conducted for the collected data. Here we divided the analysis into two parts: differential gene expression (DEG) analysis,

including individual dataset DEG analysis, DEG meta-analysis, DEG mega analysis; and gene co-expression analysis. We first processed each individual dataset according to the consistent workflow as the Figure 1 shows. And then, the datasets were sorted according to different brain regions. Next, each of the three types of DEG analyses was conducted across three aspects: brain region, sex and age stage. Besides, gene network analysis was conducted using weighted gene co-expression network analysis (WGCNA)[5]. In the third stage, we developed a database named Brain EXPression in NeuroPsychiatric Disorders (BrainEXP-NPD), which visualized the analysis results through various tables and figures.

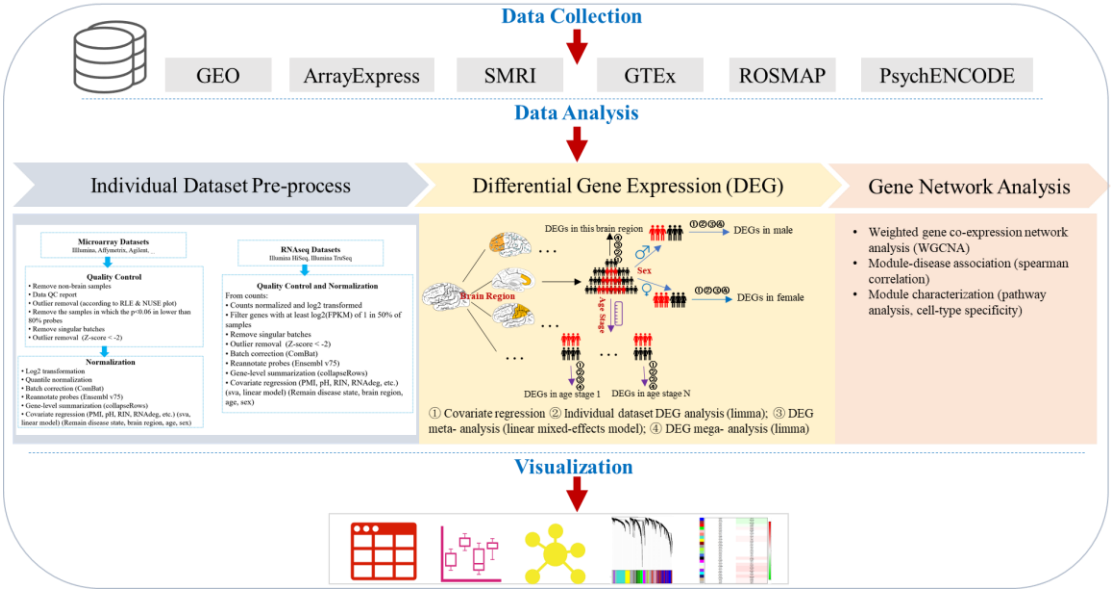

Figure 1. The study design of the BrainEXP-NPD database. DEGs: differential expressed genes

Datasets

All the transcriptomic datasets included in the BrainEXP-NPD database are shown in Table 1. The data distributions across disease state, brain region, sex, age stage are shown in Figure 2. Totally 8,317 brain samples were collected from GEO, ArrayExpress and our own lab. It consists of six complex neuropsychiatric disorders (427 schizophrenia, 312 bipolar disorder, 163 Parkinson's disease, 765 Alzheimer's disease, 219 major depressive disorder, 53 autism spectrum disorder) and 6,378 controls (no neuropsychiatric-related disorders) across 55 brain regions from 48 individual datasets and different platforms (microarrays and RNA-seq) (Figure 2).

Table 1. All the transcriptomic datasets included in the BrainEXP-NPD database

Accession	Platform	Disease State	# of Disease Samples	# of CTL Samples	# of Genes
GSE4036	HG-U133P	SCZ	SCZ 11	14	22601
GSE7621	HG-U133P	PD	PD 11	6	22601
GSE17612	HG-U133P	SCZ	SCZ 25	21	22601
GSE21138	HG-U133P	SCZ	SCZ 27	28	22601
GSE21935	HG-U133P	SCZ	SCZ 21	17	22601

GSE48350	HG-U133P	AD	AD 75	166	22601
GSE53987	HG-U133P	BD, MDD, SCZ	BD 51, MDD 49, SCZ 43	53	22601
GSE54565	HG-U133P	MDD	MDD 15	16	22601
GSE54567	HG-U133P	MDD	MDD 14	13	22601
GSE54568	HG-U133P	MDD	MDD 13	13	22601
GSE54571	HG-U133P	MDD	MDD 10	11	22601
GSE54572	HG-U133P	MDD	MDD 12	9	22601
study5	HG-U133P	BD, SCZ	BD 21, SCZ 18	18	22601
GSE5392	HG-U133A	BD	BD 32	35	13252
GSE8397-GPL96	HG-U133A	PD	PD 28	17	13252
GSE8397-GPL97	HG-U133B	PD	PD 28	17	10413
GSE20291	HG-U133A	PD	PD 15	17	13252
GSE20164	HG-U133A	PD	PD 6	4	13252
GSE20314	HG-U133A	PD	PD 4	2	13252
GSE54570	HG-U133A	MDD	MDD 7	12	13252
GSE54575	HG-U133A	MDD	MDD 11	10	13252
study2	HG-U133A	BD, SCZ, Depression	BD 11, Depression 10, SCZ 18	27	13252
study3	HG-U133A	BD, SCZ	BD 27, SCZ 27	28	13252
study7	HG-U133A	BD, SCZ	BD 29, SCZ 35	32	13252
GSE12685	HG-U133A	AD	AD 5	5	13252
GSE22570	HG-1_0st	CTL	0	48	23823
GSE35978	HG-1_0st	BD, SCZ, Depression	BD 73, Depression 25, SCZ 94	94	23823
GSE36980	HG-1_0st	AD	AD 25	44	23823
GSE45878	HG-1_1st	CTL	0	196	23847
GSE71620	HG-1_1st	CTL	0	385	23847
GSE30453	HuEx-1_0-st	CTL	0	86	15927
GSE30422	HuEx-1_0-st	CTL	0	86	15927
GSE37263	HuEx-1_0-st	AD	AD 7	8	15927
GSE60862(UKBEC)	HuEx-1_0-st	CTL	0	1163	15927
GSE25219-GPL5175	HuEx-1_0-st	CTL	0	1268	15927
GSE28521	Illumina HumanRef-8 v3.0 expression beadchip	ASD	ASD 36	38	17686

GSE26927	Illumina HumanRef-8 v2.0 expression beadchip	AD, PD, SCZ	AD 18, PD 19, SCZ 18		0	18190
GSE28894	Illumina HumanRef-8 v2.0 expression beadchip	PD	PD 52		57	15761
GSE29378	Illumina HumanHT-12 V3.0 expression beadchip	AD	AD 30		31	37804
GSE37205(Illumina Probe Annotation)	Illumina HumanHT-12 V3.0 expression beadchip	CTL		0	96	25531
GSE54562	Illumina HumanHT-12 V3.0 expression beadchip	MDD	MDD 9		9	25531
GSE54563	Illumina HumanHT-12 V3.0 expression beadchip	MDD	MDD 25		24	25531
GSE54564	Illumina HumanHT-12 V3.0 expression beadchip	MDD	MDD 19		20	25531
GSE38322(SubSeries of GSE38609, methyl)	Illumina HumanHT-12 V4.0 expression beadchip	ASD	ASD 17		18	31426
GSE30272(BrainCloud)	Illumina Human 49K Oligo array (HEEBO-7 set)	CTL		0	269	17161
ROSMAP	non-stranded, Illumina HiSeq	AD	AD 605		7	16263
GTEx	non-stranded, Illumina TruSeq library construction protocol	CTL		0	1629	19449

BrainGVEX	non-stranded, Illumina HiSeq2000	SCZ, BD	SCZ 90, BD 68	233	16660

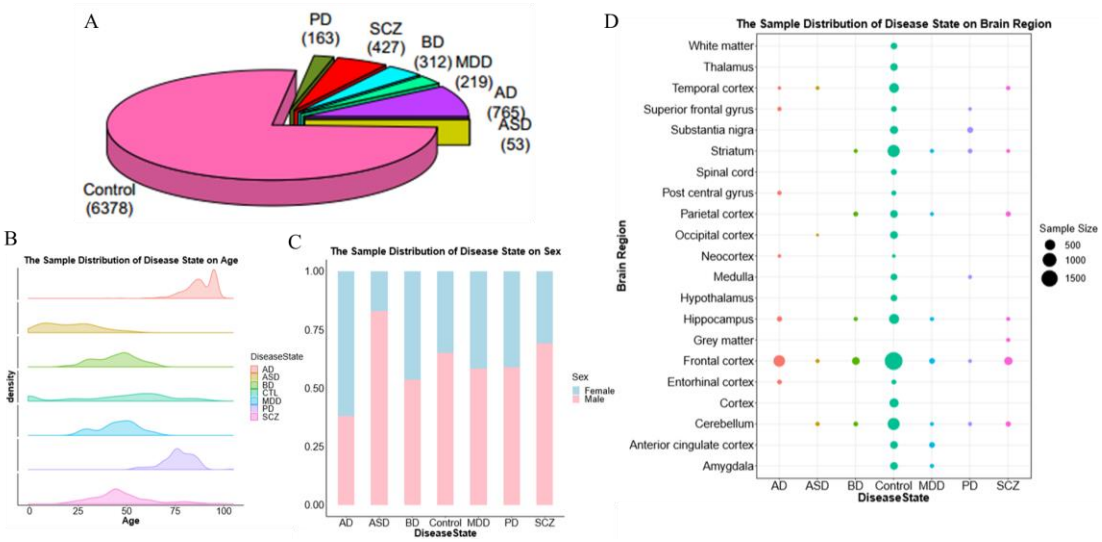

Figure 2. The data distribution in the BrainEXP-NPD database. A. The percentage of samples in different disease states. In BrainEXP-NPD, totally six neuropsychiatric disorders are included. B. The sample distribution in different ages of different disease states. C. The sample distribution in different sexes of different disease states. D. The sample distribution in different brain regions of different disease states.

Identification of Differentially Expressed Genes

Limma package[11] was used for the individual dataset DEG analysis and DEG mega analysis. For the meta-analysis, consistent differential expressed genes between cases and controls were figured out using linear mixed-effects model according to Michael J. Gandal's workflow[12]. The significant threshold for DEGs was FDR $q<0.05$.

Identification of Co-expressed Modules

Weighted gene co-expression network analysis (WGCNA)[5] was used for identifying the co-expressed modules between cases and controls. Spearman correlation analysis was conducted between module eigengene and disease state to find disease related modules. The significant threshold for correlation was FDR $q<0.05$.

Visualization of Analysis Results

The main languages for building the website are MySQL, HTML and Java. Apache was used for the server environment. Traditional MVC structure was used for the web design, the front-end page is implemented by the mainstream HTML+CSS+JS language combination, the Bootstrap component is used to customize the style of the page. Java language is used for the back-end to implement the logical functions of the website, and ECharts[13] was used to display the data in a more intuitive way. The underlying layer uses MySQL to build and store our data.

Results

Differentially Expressed Genes

Totally 6922, 3011, 2703, 4389, 3507, 4279 significantly differentially expressed genes (FDR $q < 0.05$) were detected in the comparisons of 6 brain regions of SCZ-CTRL, 5 brain regions of BD-CTRL, 6 brain regions of MDD-CTRL, 4 brain regions of ASD-CTRL, 7 brain regions of AD-CTRL, and 6 brain regions of PD-CTRL, respectively. The numbers of significantly expressed genes of different brain regions between the six disorders and controls were shown in table 2.

Table 2. Numbers of significantly expressed genes of different brain regions between the six disorders and controls

Disorder	Brain Region	# of DEGs	# of Up-regulated	# of Down-regulated
	Cerebellum	179	79	100
	Frontal cortex	2056	1093	963
	Hippocampus	671	342	329
	Parietal cortex	138	57	81
	Striatum	954	402	552
SCZ	Temporal cortex	4234	2047	2187
	Cerebellum	135	62	73
	Frontal cortex	2858	1702	1156
	Hippocampus	18	10	8
	Parietal cortex	17	10	7
BD	Striatum	5	3	2
	Anterior cingulate cortex	989	617	372
	Cerebellum	130	78	52
	Frontal cortex	1648	659	989
	Hippocampus	12	5	7
	Parietal cortex	4	2	2
MDD	Striatum	9	5	4
	Cerebellum	3476	1515	1961
	Frontal cortex	701	424	277
	Occipital cortex	71	27	44
ASD	Temporal cortex	692	398	294
	Entorhinal cortex	2	1	1
	Frontal cortex	326	186	140
	Hippocampus	1550	723	827
	Neocortex	1670	1026	644
	Post central gyrus	0	0	0
	Superior frontal gyrus	149	45	104
AD	Temporal cortex	127	94	33
	Cerebellum	500	260	240
	Frontal cortex	301	147	154
PD	Medulla	115	46	69

Striatum	1239	598	641
Substantia nigra	2812	1369	1443
Superior frontal gyrus	2	0	2

Table 3. Numbers of significantly expressed genes of different sexes in different brain regions between the six disorders and controls

Disorder	Brain Region	# of DEGs in Male	# of DEGs in Female
SCZ	Cerebellum	30	0
	Frontal cortex	1024	48
	Temporal cortex	1628	522
	Straitum	9	0
	Hippocampus	14	14
	Parietal cortex	3	0
BD	Cerebellum	9	0
	Parietal cortex	0	0
	Frontal cortex	983	547
	Straitum	0	0
	Hippocampus	0	0
MDD	Frontal cortex	537	56
	Straitum	0	0
	Hippocampus	0	0
	Parietal cortex	0	0
	Cerebellum	1	5
ASD	Cerebellum	3451	0
	Frontal cortex	361	1
	Temporal cortex	456	1
	Occipital cortex	71	
	Entorhinal cortex	0	1
	Hippocampus	324	65
AD	Post central gyrus	0	0
	Superior frontal gyrus	0	121
	Frontal cortex	0	0
	Temporal cortex	2	0
	Neocortex	64	2
PD	Substantia nigra	1328	34
	Cerebellum	5	0
	Superior frontal gyrus	3	
	Frontal cortex	0	0
	Medulla	0	0
	Straitum	210	0

Table 4. Numbers of significantly expressed genes of different age stages in different brain regions between the six disorders and controls

Disorder	Brain Region	# of DEGs					
		1-6Y	6-12Y	12-20Y	20-40Y	40-60Y	> 60Y
SCZ	Cerebellum				1	16	0
	Frontal cortex				134	854	1
	Temporal cortex					1	2181
	Straitum					402	
	Hippocampus				0	125	
	Parietal cortex				0	13	0
BD	Cerebellum				0	0	0
	Parietal cortex				0	0	0
	Frontal cortex				493	1077	
	Straitum				0	8	
	Hippocampus				0	1	
MDD	Frontal cortex				0	384	
	Straitum				0	6	
	Hippocampus				0	12	
	Parietal cortex				0	0	
	Cerebellum				0	11	
ASD	Cerebellum	24			583		
	Frontal cortex			0	1	0	
	Temporal cortex			37	26		
AD	Entorhinal cortex						0
	Hippocampus						1315
	Post central gyrus						0
	Superior frontal gyrus						50
	Frontal cortex						279
	Temporal cortex						83
PD	Neocortex						572
	Substantia nigra						2162
	Cerebellum						390
	Frontal cortex						604
	Medulla						639
	Straitum						1004

Co-expressed Modules

Totally 4, 0, 0, 8, 13, 7 disease-related modules (FDR q < 0.05) were detected in SCZ-CTRL, BD-CTRL, MDD-CTRL, ASD-CTRL, AD-CTRL, and PD-CTRL dataset, respectively. The numbers of co-expressed modules detected between the six disorders and controls are shown in table 5.

Table 5. Numbers of co-expressed modules detected between the six disorders and controls

Disorder	# of Modules	# of Significant Modules	Enriched GO/KEGG Pathways (Top 3) of the Most Significant Module
SCZ	20	4	Cytokine-mediated signaling pathway, Interferon alpha/beta signaling, cytokine production.
BD	18	0	
MDD	20	0	
ASD	18	8	Vasculature development, cytokine-mediated signaling pathway, response to wounding.
AD	19	13	Response to growth factor, response to peptide, regulation of cell adhesion.
PD	22	7	Organelle localization, mitochondrion organization, nucleoside phosphate metabolic process.

Visualization

The BrainEXP-NPD database (http://brainexpnpd.org:8088/BrainEXPNPD/index.html) was developed in an easy-to-use mode. We provide the DEG results and co-expression results on three aspects and six neuropsychiatric disorders. The result page for single gene search was shown in Figure 3. It includes totally four parts: Gene Basic Information, Gene Normalized Expression, Differential Gene Expression Summary Statistics, and Gene Co-expression. Here, we took human DRD2 as an example to demonstrate the usage of BrainEXP-NPD database, as shown in Figure 4.

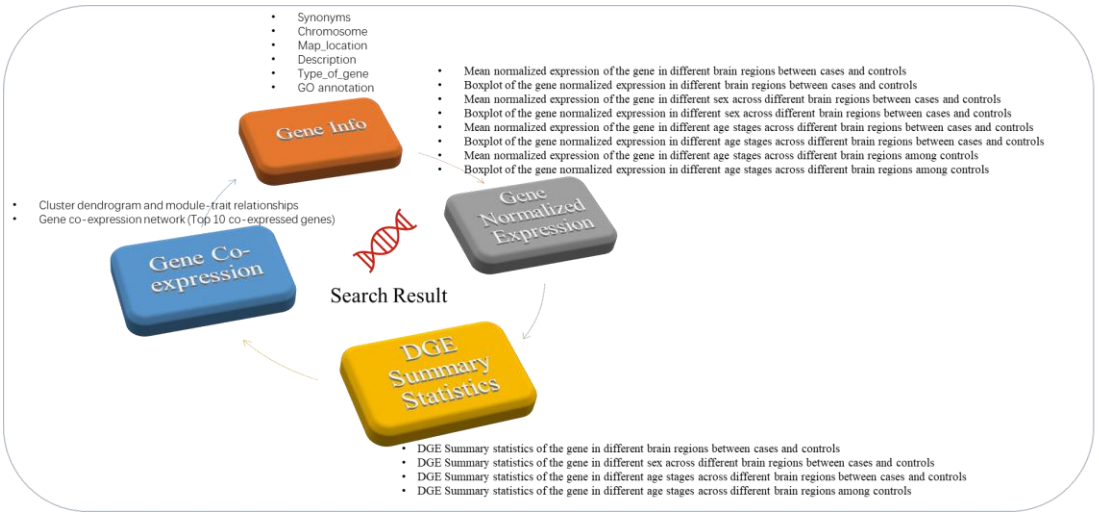

Figure 3. Data visualization content of the BrainEXP-NPD database.

Figure 4. Search results and display. A. Gene differential expression retrieval results. B. Gene co-expression retrieval results.

Summary and Perspectives

Currently, we provide a searchable & downloadable web entrance for results of normalized brain gene expression profiling, gene differential expression between

cases and controls and co-expression. More microarray and RNA-seq data for brain expression of the major neuropsychiatric disorders are under continuously updating.

Data Availability

The software used for data processing in this study is R 3.5.0, and the code for data analysis is updated on https://github.com/CuihuaXia/BrainEXP-NPD. All the data information can be freely accessed on http://brainexpnpd.org:8088/BrainEXPNPD/index.html.

Authors' contributions

Chunyu Liu and Chao Chen conceived, designed, and supervised the study. Cuihua Xia performed the overall analysis. Teng Ma developed the database webserver. Chuan Jiao contributed to the data collection and data analysis. Guihu Zhao contributed to develop the database webserver. Cuihua Xia wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared no competing interests.

Acknowledgement

This work was supported by National Natural Science Foundation of China grants 81401114, 31571312, the National Key Plan for Scientific Research and Development of China (2016YFC1306000), Innovation-Driven Project of Central South University (No. 2015CXS034，2018CX033) (to C. Chen), and NIH grants 1 U01 MH103340-01, 1R01ES024988 (to C. Liu). All the data contributors are sincerely appreciated for data submitted in the GEO and other databases. We are grateful to Professor Jinchen Li and National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University for hosting our webserver.

References

[1] VOINEAGU I, WANG X, JOHNSTON P, et al. Transcriptomic analysis of autistic brain reveals convergent molecular pathology [J]. Nature, 2011, 474(7351): 380-4.

[2] WU Y, YAO Y G, LUO X J. SZDB: A Database for Schizophrenia Genetic Research [J]. Schizophr Bull, 2017, 43(2): 459-71.

[3] XU M, ZHANG D F, LUO R, et al. A systematic integrated analysis of brain expression profiles reveals YAP1 and other prioritized hub genes as important upstream regulators in Alzheimer's disease [J]. Alzheimers Dement, 2018, 14(2): 215-29.

[4] HIGGS B W, ELASHOFF M, RICHMAN S, et al. An online database for brain disease research [J]. BMC Genomics, 2006, 7(70.

[5] LANGFELDER P, HORVATH S. WGCNA: an R package for weighted correlation network analysis [J]. BMC Bioinformatics, 2008, 9(559.

[6] BARRETT T, WILHITE S E, LEDOUX P, et al. NCBI GEO: archive for functional genomics data sets--update [J]. Nucleic Acids Res, 2013, 41(Database issue): D991-5.

[7] ATHAR A, FULLGRABE A, GEORGE N, et al. ArrayExpress update - from bulk to single-cell expression data [J]. Nucleic Acids Res, 2019, 47(D1): D711-D5.

[8] CONSORTIUM G T. The Genotype-Tissue Expression (GTEx) project [J]. Nat Genet, 2013, 45(6): 580-5.

[9] DE JAGER P L, MA Y, MCCABE C, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research [J]. Sci Data, 2018, 5(180142.

[10] PSYCH E C, AKBARIAN S, LIU C, et al. The PsychENCODE project [J]. Nat Neurosci, 2015, 18(12): 1707-12.

[11] RITCHIE M E, PHIPSON B, WU D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies [J]. Nucleic Acids Res, 2015, 43(7): e47.

[12] GANDAL M J. Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap [J]. 2018, 359(6376): 693-697.

[13] DEQING LI E A. ECharts: A Declarative Framework for Rapid Construction of Web-based Visualization [J]. Visual Informatics, 2018, 2(2): 136-146.