Supplementary materials

BrainEXP: a database featuring with spatiotemporal expression variations and co-expression organizations in human brains

Chuan Jiao[1], Cuihua Xia[1], Kangli Wang[1], Yi Jiang[1], Lingling Huang[1], Rujia Dai[1], Yu Wei[1], Yan Xia[1], Qingtuan Meng[1], Liu Yi[1], Fangyuan Duan[1], Jiacheng Dai[1], Shunan Zhao[1], Chunyu Liu[1, 2, *] and Chao Chen[1, 3, *]

[1]School of Life Science, Central South University, Changsha, Hunan, 410012, China. [2]Department of Psychiatry, SUNY Upstate Medical University, Syracuse, NY 13201, USA.[3]National Clinical Research Center for Geriatric Disorders, Central South University, Changsha, Hunan, 410012, China.

*To whom correspondence should be addressed.

Outline

1. Data Summary

Currently BrainEXP contains 4,567 normal human brain samples of 2,863 individuals with 56 brain regions from our own data and existing public databases.

✧ Analysis workflow

The overall of the analysis workflow shows below.

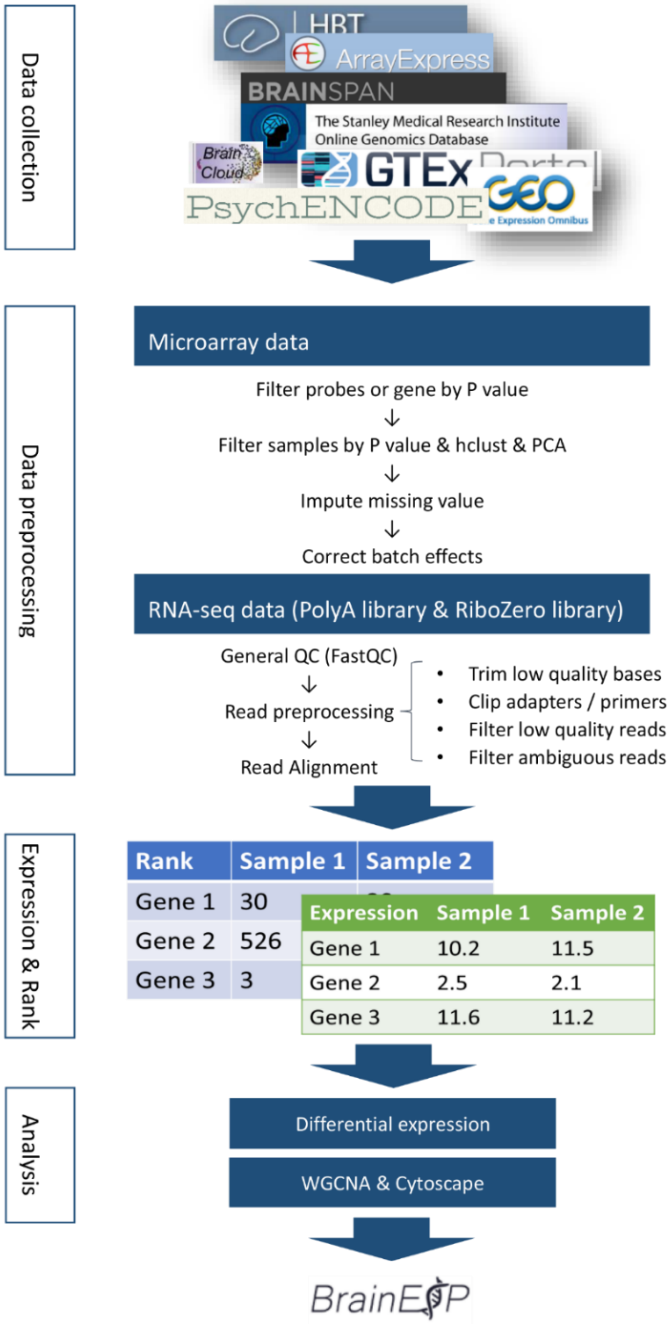

Figure S1. An overview of the analysis workflow for data collection and basic analysis.

- **Data pre-processing:**

Data were collected from vastly different platforms and they are hardly comparable. We normalized the data based on the platform information by the same workflow.

- **HG-U133A (Human Genome U133A 2.0 Array) / HG-U133P (Human Genome U133 Plus 2.0 Array)**

The data pre-processing workflow of platforms HG-U133A and HG-U133P are the same, so we describe these platforms together. After we received the normal human brain gene expression raw data (CEL files), we filtered the probes with a detection P value above 0.06 in at least 80% samples, and then filtered the samples with a detection P value above 0.06 in at least 80% probes. We processed the CEL files using standard tools available within the affy package in R. The CEL files were processed with the expression command to convert the raw probe intensities to probeset expression values. We used the CLL, CLLbatch, simpleaffy packages to obtain the quality control results of samples and filtered the poor-quality samples. After that, we used the impute package in R to impute the NA values in the dataset. Then we used gene-based expression values to replace the probe-based values. In the case of multi-probe-per-gene, we took the max of the probe-based values as the representative values of the gene. After we logit transformed the data and corrected the batch information using *ComBat*, we put the data from different data sets together by *ComBat* method. We corrected batch effects using the *ComBat*, which adjusts for known batches using an empirical Bayesian framework (W.E. Johnson, 2007).

- **HG1.0 (HuGene-1_0-st) /HG1.1 (HuGene-1_1-st)**

The data pre-processing workflow is the same for both platforms HG1.0 and HG1.1, so we describe these platforms together. The raw data is also in the CEL format, so we took the normal human brain CEL files to the expression console to get the gene expression matrixes. We removed the probesets that crosshyb-type=3, which means the probesets perfectly or partially match more than one sequence. We filtered the samples by hierarchical cluster with average linkage –if a sample (or small group of samples) didn't cluster with other samples, we defined the sample as an outlier and removed it. After that, we used the impute package in R to impute the NA values in the dataset. Then we used gene-based expression values to replace the probe-based values. In case of multi-probe-per-gene, we took the max of them as the representative values of the gene. We then corrected batch effects to remove the batch confounder. Finally, we put the data from different data sets together by *ComBat* method.

- **Illumina Human 49K Oligo array (HEEBO-7 set) / Illumina Human HT-12 V3.0 expression beadchip**

The data we received from these two datasets are both non-normalized data with similar data structure, so we preprocessed the data similarly. We filtered samples by *hclust* function in R (similar to how we processed the HG1.0 /HG1.1 data). In the case of multi-probe-per-gene, we took the max of the probe-based values as the representative values of the gene. After that, we use the *ComBat* function to adjust the batch effect within dataset and among datasets.

- **RNA-seq (Ribo-Zero)**

Table S1. The basic tools and workflow of the RNA-seq (Ribo-Zero) dataset analysis.

Workflows	**Purpose**	**Tools**
Quality Control	Generate quality statistics	FastQC v0.11.2
Read Preprocessing	1.Trim low-quality bases	Trimmomatic v0.22
	2.Clip adapters/primers	
	3. Filter low quality reads	
	4. Filter ambiguous reads	
Align Reads to Reference Genome	align trimmed fastq reads to transcript reference	TopHat v2.0.13
Collect Multiple Metrics	Insert Size, Quality Score, Cycle, and RNA seq feature	Picard v1.109
Assemble Transcripts and Quantification	Estimate RNA expression levels by FPKM using TopHat alignment results	Cufflinks v2.2.1
	Read counts using TopHat alignment results	HTSeq

- **RNA-seq (PolyA+)**

RNA-seq was performed using the Illumina TruSeq library construction protocol. This is a non-strand specific polyA+ selected library. The sequencing produced 76-bp paired-end reads. Alignment to the HG19 human genome was performed using TopHat v1.4.1 assisted by the GENCODE v19 transcriptome definition. In post-processing, unaligned reads are reintroduced into the bam. The final bam contains aligned and unaligned reads and marked duplicates. It should be noted that TopHat produces multiple mappings for some reads, but in post-processing one read is flagged as the primary alignment. We filter the probes by 50% samples with FPKM > 0.2 and filter the samples by hclust in R (similar with the HG1.0 /HG1.1 part). In case of multi-probe-per-gene, we take the max of the probe-based values as the representative values of the gene.

- **Differential expression analysis:**

After the basic pre-processing, data were performed the unified adjustment and differential expression analyses pipelines. The linear regression-based adjustment for the chosen covariates and the potential covariates detected by *sva* function. Differential expression among different ages was assessed using the *lm* function in R, with the following inputs: the adjusted gene expression matrix and the age information. We got the differential expression gene list in sex and brain regions by Students' t-test and Analysis of variance (ANOVA) respectively. The resulting P values were then adjusted form multiple hypothesis testing using false discovery rate (FDR) estimation, and the differential expressed genes were determined as those with an estimated FDR<0.05. FDR was calculated by *qvalue* in R (The code can be downloaded from *https://github.com/ChuanJ/DEGlist.R*). Note that the analyses were performed separately by platform.

- **Co-expression analysis:**

We applied weighted gene co-expression network analysis (WGCNA) to the matrix of pairwise gene co-expression values. WGCNA recovers a network that consists of nodes (genes) and edges connecting nodes (i.e., the degree of co-expression for a pair of genes, measured as their correlation after transformation by raising the value to a power, β, that results in an overall scale-free topology). It divides the network into subnetworks called modules, or clusters of genes with more highly correlated expression. By raising the absolute value of the correlation to a power $\beta \geq 1$ (soft thresholding), the weighted gene co-expression network construction emphasizes high correlations at the expense of low correlations. Specifically, we build the signed networks using $a_{ij} = |(1 + cor(x^i, x_j))/ 2|^{\beta}$ (i and j represent gene ID) to calculate the adjacency. The signed networks were constructed using *blockwiseModules* function. Modules were defined using biweight midcorrelation (bicor), with a minimum module size of 50. The power of 8 platforms shows below:

Table S2. The power value we chose in the WGCNA analysis

#	Platform	Power
1	HG-U133P	3
2	HG-U133A	4
3	HuGene-1.0-st	5
4	HuGene-1.1-st	3
5	Illumina Human 49k Oligo Array	8
6	Illumina Human HT-12 V3.0 expression beadchip	5
7	RNA-seq (PloyA+)	5
8	RNA-seq (Ribo-Zero)	2

Note: 49K means Illumina Human 49K Oligo array (HEEBO-7 set), V3.0 means Illumina Human HT-12 V3.0 expression beadchip.

✧ **Data summary**

Here is the detailed sample information.

Table S3. The numbers of gene and samples from each platform.

#	Platform	# of gene	# of samples
1	HG-U133P	21,154	402
2	HG-U133A	12,048	98
3	HuGene-1.0-st	10,015	196
4	HuGene-1.1-st	9,575	626
5	Illumina Human 49k Oligo Array	17,141	269
6	Illumina Human HT-12 V3.0 expression beadchip	25,123	254
7	RNA-seq (Ploy A library)	20,948	1258
8	RNA-seq (RiboZero)	21,866	258

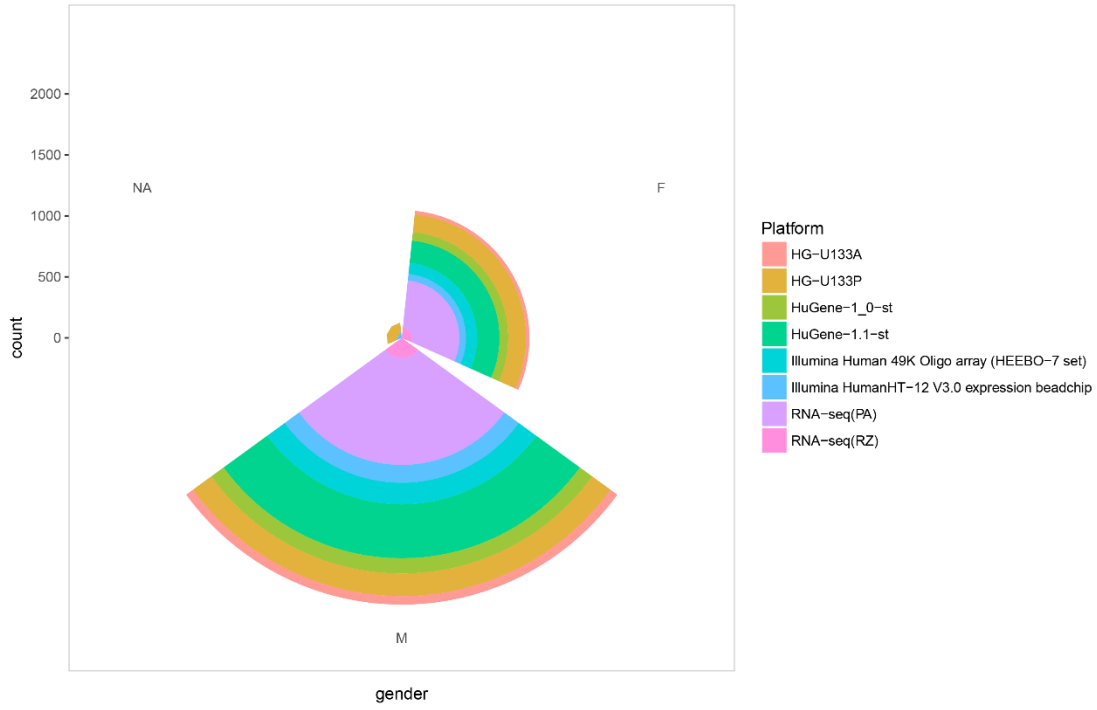

Figure S2. The sample sex distribution in each platform.

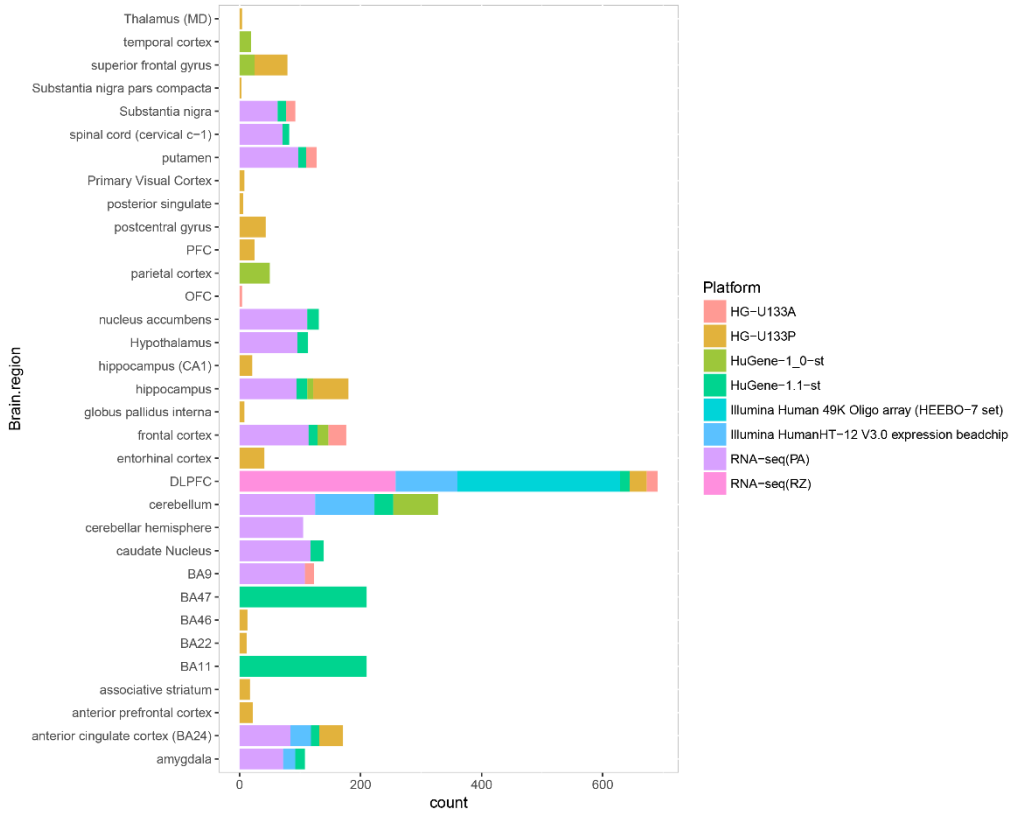

Figure S3. The sample brain region distribution in each platform.

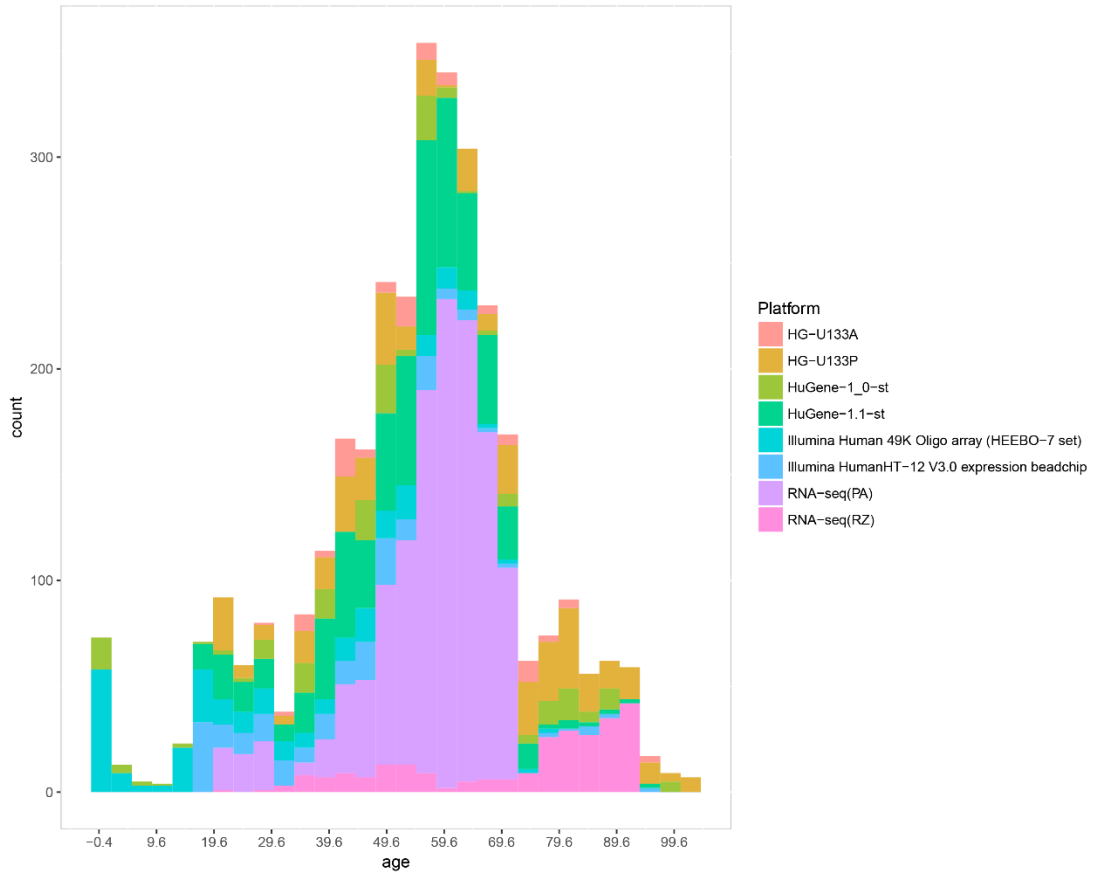

Figure S4. The sample age distribution in each platform.

2. Search for data

✧ Quick search

If you want to scan or browse a single gene's expression value and its correlation with other genes in brain rapidly, you can use a quick search. You can search the gene by gene symbol, gene ID, or ensemble ID. In the quick search, we have default values on the retrieval restriction. For sex, age, and brain region modules, the default values are "select all". For platform module, the default value is "RNA-seq (PolyA+)." You can also download the search results.

✧ Search strategy

Gene symbol: Vague search and automatic matching search are supported. For example, the search is not case-sensitive, if you enter "SCN" or "scn," all of the genes whose gene symbol containing uppercase, lowercase or a mixed case of "SCN" will be listed in the results. When you enter "DRD" or "drd," all of the genes whose name begin with "DRD" or "drd," such as "DRD1", "DRD2", "DRD4", "DRD5" will be listed for the remainder.

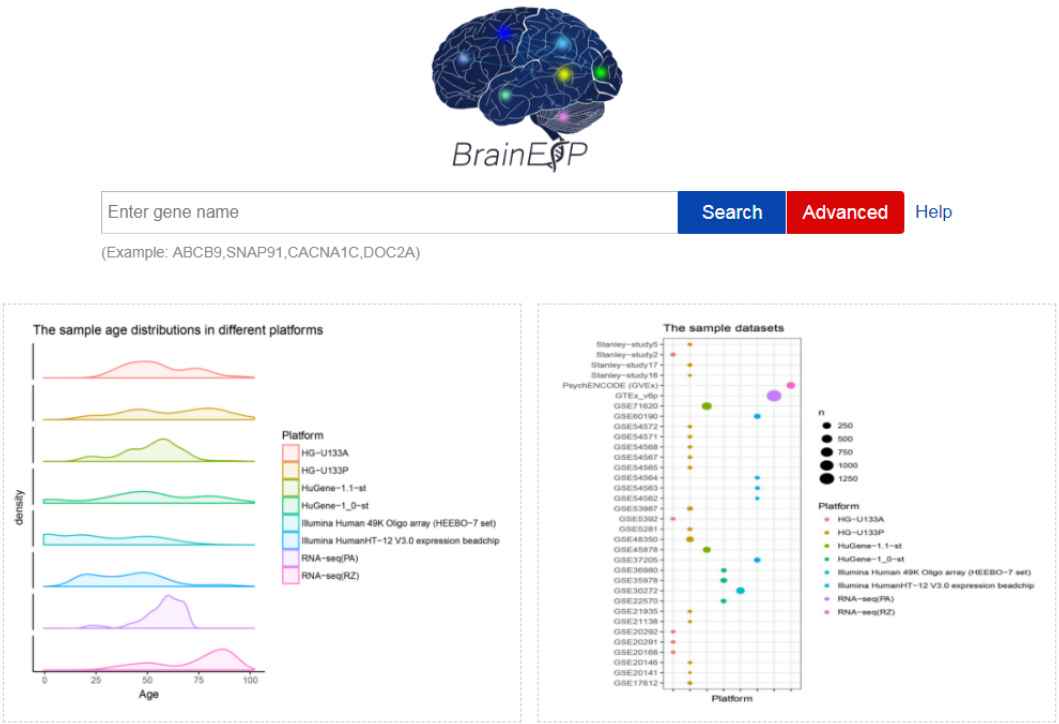

Figure S5. The quick search page with the help and advanced button.

✧ Advanced search

If you are interested in a list of genes, or the gene expression for only a particular age, brain region, or more restrictions, you can select the advanced search. You can search any gene in the human brain for age, sex, brain region, or platform. When you input multiple genes, each item should be separated by a comma. In the age module, you can select any numbers from -0.5 to 110 (ys), such as “20~30” or “-0.5~0”. The negative number indicates the samples are from the embryonic and fetal individuals. In the brain region module, you can select one or more brain regions at a time. In the platform module, if you select "HG-U133P", related data in BrainEXP will be displayed. If you don’t choose a platform, it will have an “RNA-seq (PolyA+)” default. You can also download all the search results.

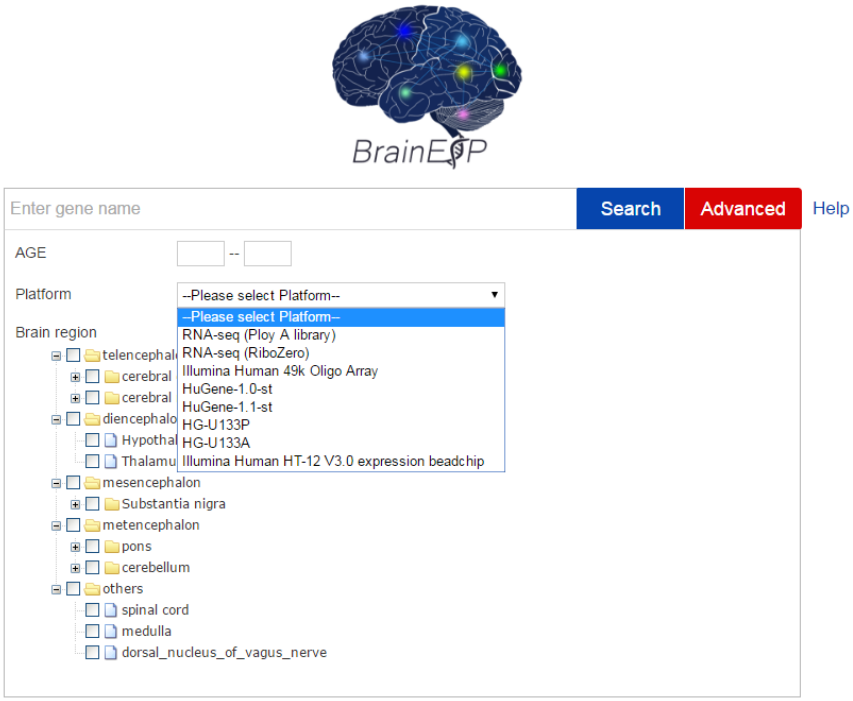

Figure S6. The advanced search page with the age range, platform, and brain region chosen.

3. BrainEXP search results

You can obtain the basic gene information and gene ontology at first glance when you open the search page. The basic information of gene is obtained from NCBI.

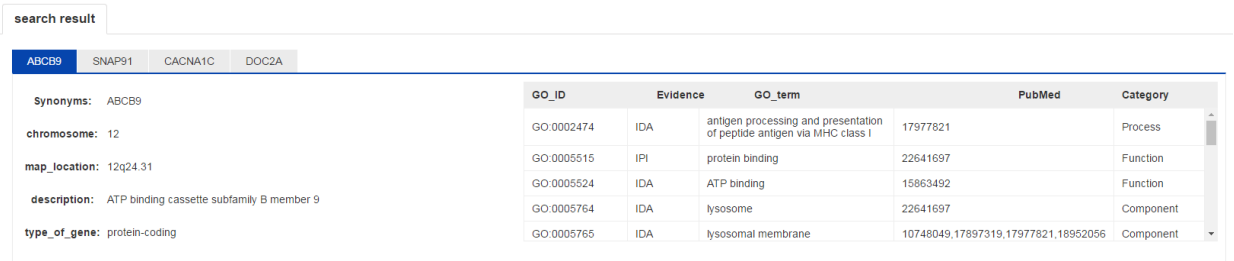

Figure S7. The gene information interface.

By restriction to retrieve, we get the gene expression-related results. This detailed information contains "Spatiotemporal expression variations" and "Gene co-expression."

✧ **Spatiotemporal expression variations**

Basic spatiotemporal expression variation information includes differential expression of sex, age stage, and brain region. See details below:

➢ **Differential expression analysis**

Differential expression among age, sex, and brain region

	Age	Sex	Region
ABCB9	-	-	***
CACNA1C	**	-	*
DOC2A	-	-	*
SNAP91	***	-	**

*** means FDR<0.001, ** means 0.001≤FDR <0.01, *means 0.01≤FDR <0.05, - means FDR>0.05.
Note: The results are based on all the data in the platform you selected and are not specific to the certain age range or brain region.

Figure S8. The gene expression in different ages, sexes and brain regions

➢ **Gene expression in different brain regions and sexes**

This part displays a box plot by calling all the expression data in BrainEXP. It can exhibit gene expression difference among different brain regions and sexes.

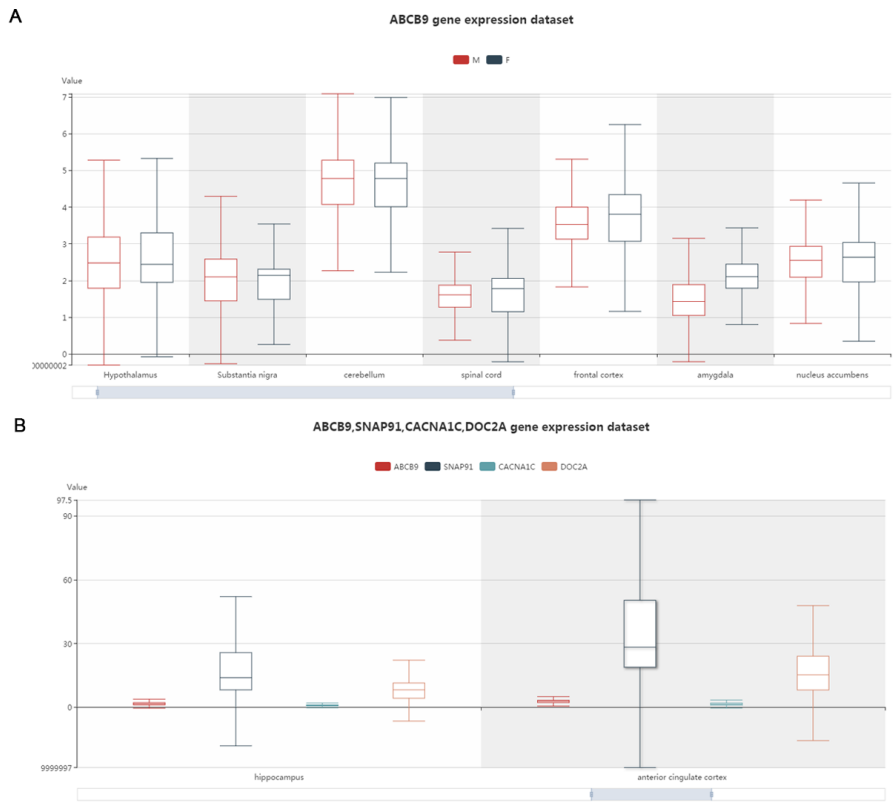

Figure S9. The gene expression in different brain regions and sexes interface. You can drag the image to get the complete information. A. One gene search results. The different colors represent different sexes. B. Multiple genes search results. The different colors represent different genes.

➢ Gene expression in different brain regions and ages

This part shows a scatter plot by calling all the expression data in BrainEXP. It can exhibit brain gene expression level among different ages and brain regions. See the graph below.

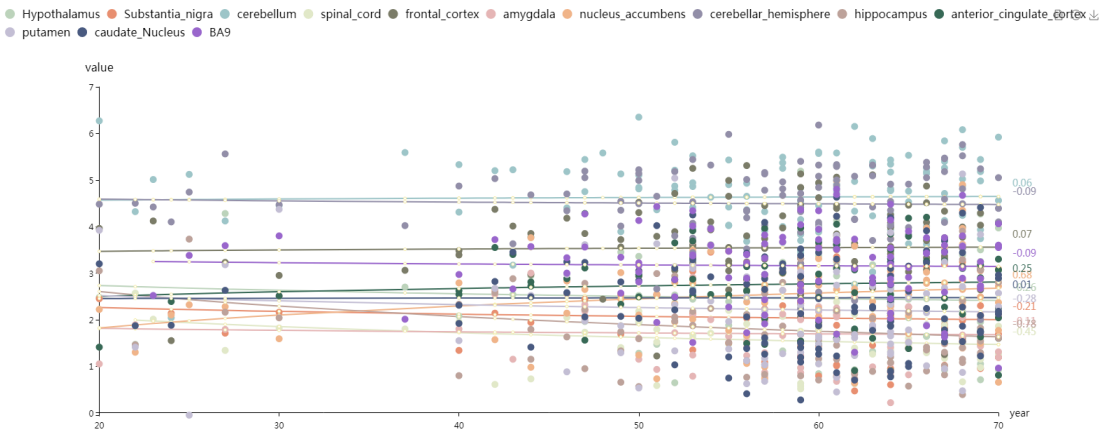

Figure S10. The Gene expression in different brain regions and age interface.
You can click the legend to delete the related points. The number sideward means the

gradient of the trend line. The different colors represent different brain regions.

✧ **Gene co-expression**

By spatiotemporal co-expression analysis, a network of co-expression will be displayed. You can set some parameters to get the network you want. The color of nodes is listed from dark to light; a darker color denotes a closer correlation with the central node. You can set maximum nodes of the network. In this case, these genes which locate closest to the central gene will be displayed on the website. See graph below.

➢ Co-expression gene network

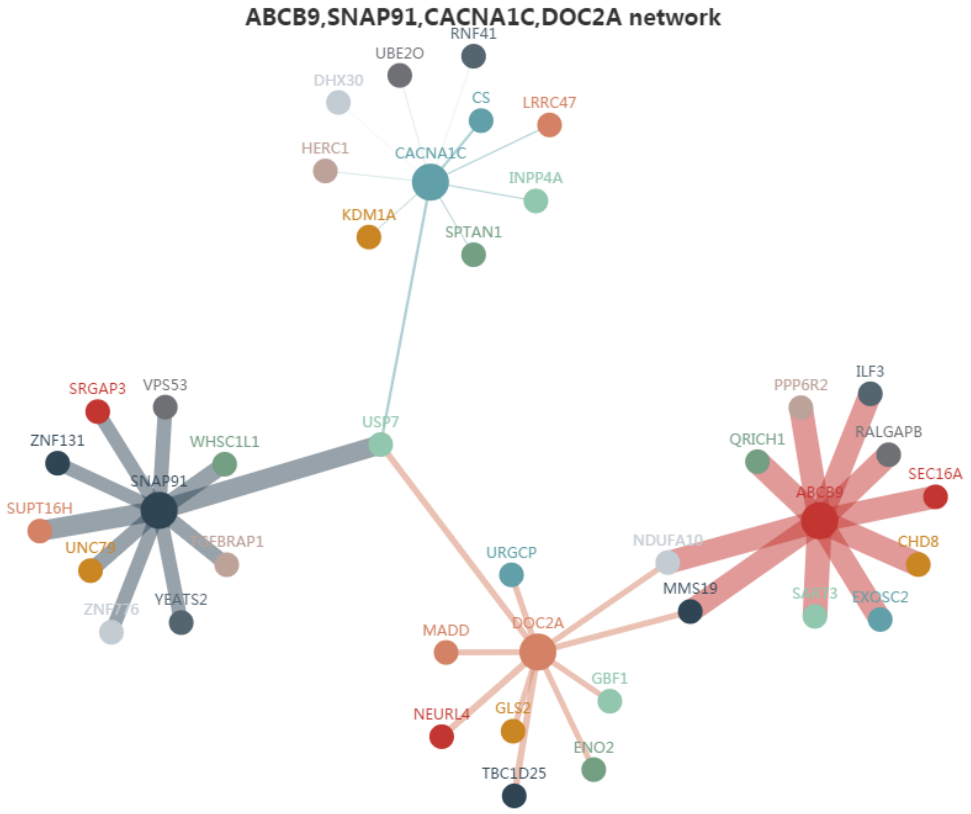

Figure S11. The network constructed by the top 10 genes searched related. The different colors represent different genes. The weight-value calculated by WGCNA orders the line thickness. The more related to each other, the line will be much thicker.

➢ Co-expression pattern

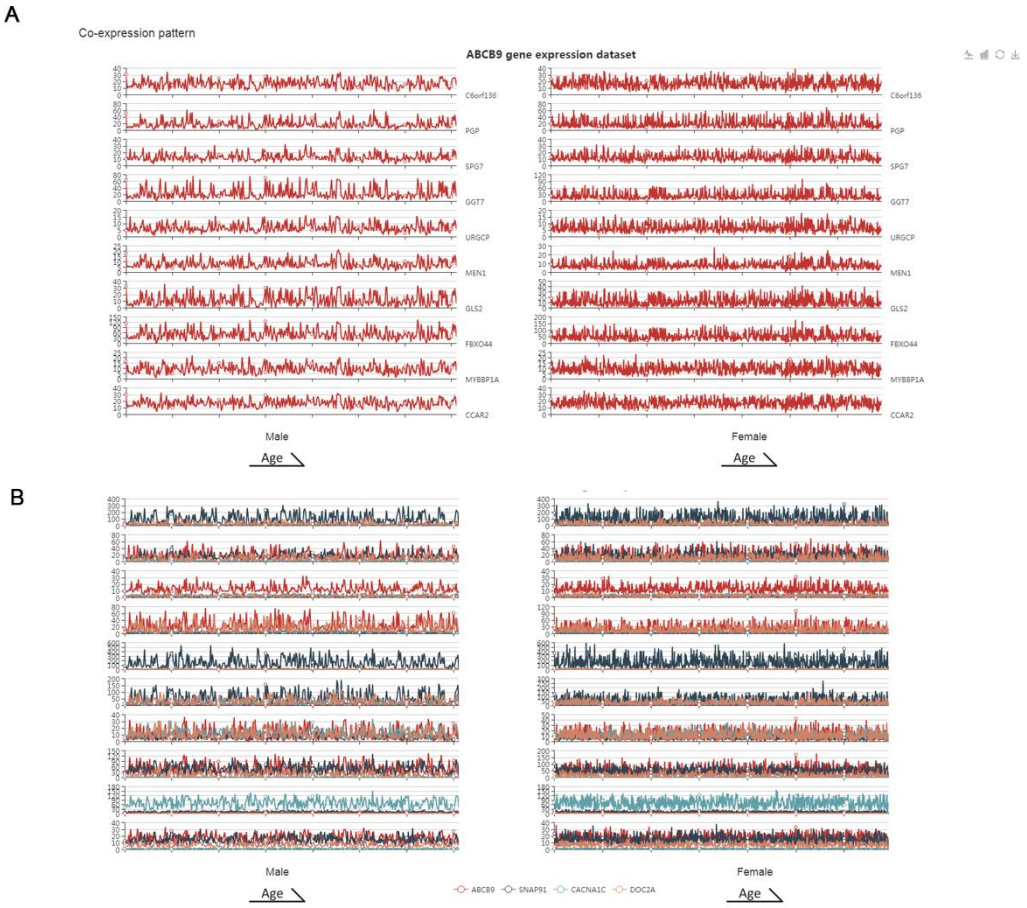

Figure S12. The co-expression pattern interface. Users can get the expression degree of the related search gene. The x-axis represents the samples in the searched platform. The left is the male expression pattern, and the right is female. Expression pattern plot in each part is ordered by age so that the users can see the similarity of co-expression genes pattern. The y-axis represents the expression value of each related gene. A. One gene search results. The different color represents different related gene. B. Multiple genes search results.

➢ Correlation of co-expression gene

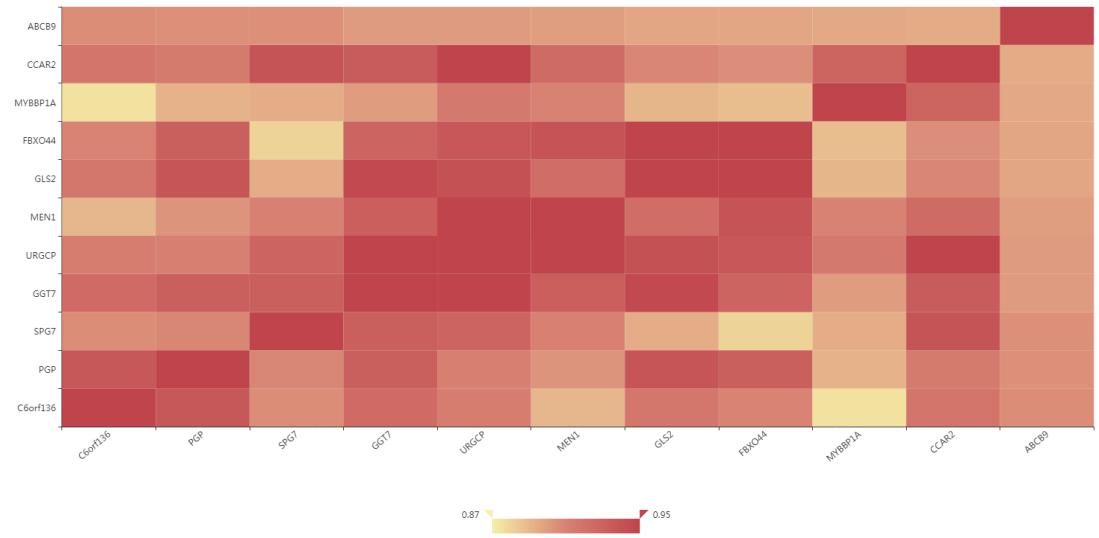

Figure S13. The correlation of co-expression gene interface. The different colors represent different degrees of correlation value. Users can click the cells to get detailed correlation values.

➢ WGCNA cluster dendrogram

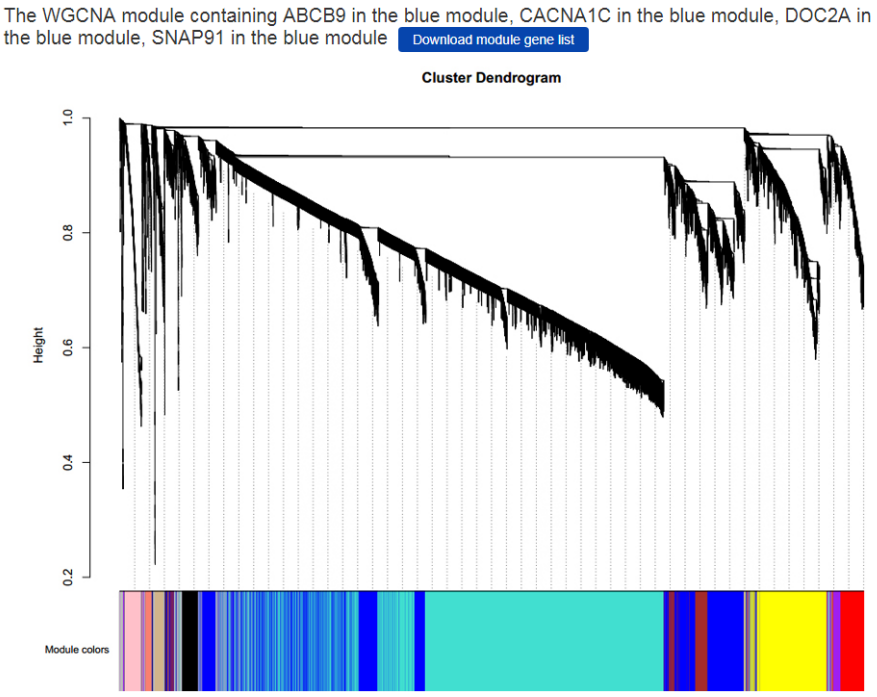

Figure S14. The WGCNA cluster dendrogram interface. The different colors represent different modules. The chart shows in which module the gene was detected. Users can see how many genes are in the same module with the gene of interest.