RESEARCH ARTICLE

Analysis of Gene Expression Profiles in the Human Brain Stem, Cerebellum and Cerebral Cortex

Lei Chen[1,2]☯, Chen Chu[3]☯, Yu-Hang Zhang[4], Changming Zhu[2], Xiangyin Kong[4], Tao Huang[4]*, Yu-Dong Cai[1]*

1 School of Life Sciences, Shanghai University, Shanghai, 200444, China, **2** College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China, **3** Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China, **4** Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China

☯ These authors contributed equally to this work.
* cai_yud@126.com (YDC); tohuangtao@126.com (TH)

OPEN ACCESS

Citation: Chen L, Chu C, Zhang Y-H, Zhu C, Kong X, Huang T, et al. (2016) Analysis of Gene Expression Profiles in the Human Brain Stem, Cerebellum and Cerebral Cortex. PLoS ONE 11(7): e0159395. doi:10.1371/journal.pone.0159395

Editor: Ramani Ramchandran, Medical College of Wisconsin, UNITED STATES

Received: March 19, 2016

Accepted: July 1, 2016

Published: July 19, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This study was supported by the National Natural Science Foundation of China (61202021, 31371335), Shanghai Natural Science Foundation (16ZR1414500), Shanghai Sailing Program and The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

The human brain is one of the most mysterious tissues in the body. Our knowledge of the human brain is limited due to the complexity of its structure and the microscopic nature of connections between brain regions and other tissues in the body. In this study, we analyzed the gene expression profiles of three brain regions—the brain stem, cerebellum and cerebral cortex—to identify genes that are differentially expressed among these different brain regions in humans and to obtain a list of robust, region-specific, differentially expressed genes by comparing the expression signatures from different individuals. Feature selection methods, specifically minimum redundancy maximum relevance and incremental feature selection, were employed to analyze the gene expression profiles. Sequential minimal optimization, a machine-learning algorithm, was employed to examine the utility of selected genes. We also performed a literature search, and we discuss the experimental evidence for the important physiological functions of several highly ranked genes, including NR2E1, DAO, and LRRC7, and we give our analyses on a gene (TFAP2B) that have not been investigated or experimentally validated. As a whole, the results of our study will improve our ability to predict and understand genes related to brain regionalization and function.

Introduction

Human brains are distinguished from those of all other species by their incomparable cognitive capacities. Throughout human history, people have questioned how this mysterious and powerful organ functions in such a highly orchestrated manner. Studies of the human brain using cellular and molecular biological techniques have been undertaken for generations, but the mechanisms underlying the development, differentiation and function of the human brain remain quite elusive. It is widely accepted that fine-tuned spatiotemporal gene expression

Competing Interests: The authors have declared that no competing interests exist.

contributes to the proper function of individual tissues [1]. With the advancement of high-throughput technologies, brain transcriptomics studies have gained more attention and have given rise to a large amount of brain gene expression data over the past decade. Comparative studies of gene expression in the brains of different species have shown that the divergence in brain gene expression patterns has contributed to the changes in human brain function during evolution [2], and some human-specific brain gene expression patterns have been gradually revealed [3–5]. Spatially regulated gene expression is another feature of the human brain that is closely related to its development, differentiation and region-specific functions. Previous studies compared brain regional transcriptomes among different species and revealed a correlation between brain region-specific gene expression and function from an evolutionary perspective [6–8]. However, much less is known about intra-specific brain gene expression in humans and how the divergence in gene expression patterns of human brains contributes to differences based on age, race, sex, and other factors. Financial limitations and the limited availability of human samples usually impede the amount of data acquired or analyzed in a given study, and therefore, it is of great significance to perform data mining of previous datasets [9].

In 2013, the Human Brain Project (HBP), a large, 10-year research project, was established to unravel the mystery of human brains. One of its goals is to organize neuroscience data using a variety of computational models and analytical tools and to better serve the integration and analysis of increasingly large-volume, high-dimensional, and fine-grained experimental data. Despite the progress the HBP has made in the past two years, many obstacles in the methodology (data integration, computation, engineering, *etc.*) and overall research paradigm must still be addressed [10]. In addition to the HBP, several other previous or ongoing research projects also provide information regarding human brains. For example, the Allen Human Brain Atlas serves as an integrative and powerful database of brain gene expression data with high anatomical resolution [11] and has provided multimodal data-mining resources for many previous studies [9, 12, 13].

In this study, we further explore the universalities of regional gene expression of human brains, as well as the differential gene expression patterns among individuals of different ages, races and sexes. Compared to previous studies, we sought to better explore region-specific brain gene expression by minimizing the differences among samples. Considering data availability and reliability, we obtained gene expression data from the Allen Human Brain Atlas database. In this study, we employed several feature-selection methods, including minimum redundancy maximum relevance (mRMR) and incremental feature selection (IFS) [14], and a machine learning method, the sequential minimal optimization (SMO) algorithm, [15, 16] to analyze the gene expression profiles of the brain stem (BS), cerebellum (CB) and cerebral cortex (CC) in the brains of six different people. We also performed a literature search to gather evidence to support our analysis.

Materials and Methods

Materials

We downloaded the gene expression profiles of BS, CB and CC samples from six people (H0351.2001, H0351.2002, H0351.1009, H0351.1012, H0351.1015 and H0351.1016) from the Allen Brain Atlas [11] (http://human.brain-map.org/static/download). Detailed information regarding these six individuals is provided in S1 Text, which can be downloaded from http://help.brain-map.org/download/attachments/2818165/CaseQual_and_DonorProfiles.pdf?version=1&modificationDate=1382051848013. The number of samples from each region available for each person can be found in **Table 1**. Depending on the brain region from which the sample was obtained, samples from one person can be divided into three classes, BS, CB and

Table 1. The distribution of samples, obtained from six individuals, among three regions of the human brain.

Individual identification code	Number of samples from each region of the human brain			Total number of samples
	Brain stem (BS)	Cerebellum (CB)	Cerebral cortex (CC)	
H0351.1009	26	42	295	363
H0351.1012	80	48	401	529
H0351.1015	79	62	329	470
H0351.1016	59	80	362	501
H0351.2001	154	53	739	946
H0351.2002	188	83	622	893

doi:10.1371/journal.pone.0159395.t001

CC, and comprise a dataset. For each sample, the expression levels of 20,782 genes were measured using microarray analysis. And the expression level of each gene was the feature used for classifying the samples from different brain regions.

The goal of this study was to identify the differentially expressed genes among different brain regions in each person and then obtain a list of robust, region-specific, differentially expressed genes by comparing the expression signatures from different persons. Each of the six datasets was used as the training dataset for one analysis in which the remaining five datasets were used as test datasets.

For each individual, there were samples from three brain regions including BS, CB and CC. And we try to identify the genes that can classify them.

For cross-individual analysis, we used the discriminative genes identified from one individual to classify the samples in other five individuals. The cross-individual analysis could measure the robustness of the discriminative genes and obtain a brain region specific, but not individual specific discriminative gene list. The individual differences could be a confounding factor in brain region specific analysis.

mRMR method

To identify differentially expressed genes, we used a popular feature selection method, mRMR, to analyze the gene expression profiles of three regions in the human brain. The mRMR method, proposed by Peng *et al.* [14], was developed based on two criteria: Max-Relevance and Min-Redundancy. The output of the mRMR program contains two feature lists, the MaxRel feature list and the mRMR feature list, in which all features are sorted. Max-Relevance guarantees that the features that correlate strongly with the target variable receive high ranks, while Min-Redundancy guarantees that a feature with low redundancy to features already in the list is selected in the next round. For the two obtained feature lists, the MaxRel feature list was produced based only on the Max-Relevance criterion, and the mRMR feature list was produced based on both criteria. We define these two feature lists as follows:

$$\begin{cases} \text{MaxRel features list} : F_{\text{MaxRel}} = [f_1^M, f_2^M, \cdots, f_N^M] \\ \text{mRMR features list} : F_{\text{mRMR}} = [f_1^m, f_2^m, \cdots, f_N^m] \end{cases} \tag{1}$$

The MaxRel feature list can provide clues to assess which features are important for distinguishing samples from different classes. However, the combination of a number of top features in this list is not always an optimal combination for classification because redundancies may occur between them. On the other hand, the mRMR feature list considers these factors. Thus, the mRMR feature list is more appropriate for building an optimal classification model and extracting an optimal combination of features. MaxRel and mRMR feature lists have been widely used to address a variety of biological problems [17–26].

Prediction engine

The mRMR method described above only provides lists of features. To extract important features (genes), a prediction engine is necessary, and it, together with the mRMR feature list, was used according to the method described below. In this study, we adopted a powerful machine learning method, SMO, as the prediction engine. SMO is proposed by John Platt [15, 16] and is one of the most popular methods for solving support vector machines (SVMs) in the dual space. It is a type of decomposition method and always uses the smallest possible working set, which contains two dual variables and can be updated very effectively. The optimization problem is divided into a number of smallest possible sub-problems, which are solved analytically [27]. Nowadays, it has been used in many algorithms, especially for C-SVM (SVM for classification) [28–30]. Some published papers have validated the convergence behavior of SMO [31–33]. All of these indicate that SMO is a good method to optimize solving procedures of SVM. Because this study involved three classes, pair-wise coupling [34] was applied to build the multi-class classifier.

In Weka [35], a classifier called SMO implements the SMO method described above. For convenience, this classifier was directly employed as the prediction engine and was used with its default parameters.

Revised IFS method

The original IFS method, which was proposed by Peng *et al.* [14], used the mRMR feature list and a basic prediction engine to extract important features and develop an optimal prediction. All of these procedures are executed only on the training dataset. In this study, we mainly focused on finding differentially expressed genes identified from one individual to classify the samples in other five individuals, but not individual specific differentially expressed genes. We modified the original IFS method by executing it on both the training set and five test datasets. A detailed description is presented below:

1. According to the mRMR feature list $F_{mRMR} = [f_1^m, f_2^m, \ldots, f_N^m]$, N feature sets, denoted as F_1, F_2,. . .,F_N, can be constructed as $F_i = \{f_1^m, f_2^m, \ldots, f_i^m\}(i = 1, 2, \ldots, N)$, *i.e.*, F_i contains the first i features in the mRMR feature list.
2. For each F_i, samples in the training dataset and five test datasets were all represented by features in F_i. Then, the classifier SMO was trained on the training set, and its performance was evaluated using five test datasets. The predicted results for each of the test datasets were counted as the total prediction accuracy and accuracy of each class.
3. For each test dataset, an IFS curve was plotted by setting the total prediction accuracy as the Y-axis and the number of features used as the X-axis.

Theoretically, for a training dataset and a test dataset, we want to find the optimal combination of genes that can accurately evaluate the differences between two people. Thus, a feature set that results in the maximum total prediction accuracy on the test dataset represents the optimal combination of genes for which we are searching. However, this feature set always contains an extremely large number of features (genes), which complicates the analysis. In this study, an inflection point was identified on each IFS curve. The inflection point is defined as the first point on the curve for which the total prediction accuracy at this point is greater than or equal to the total prediction accuracy at the point prior to this point and for which the total prediction accuracy is greater than the total prediction accuracy at the point posterior to this point. Features in the feature set corresponding to this point were deemed to be significant for the problem addressed in this study.

Results and Discussion

Results of the mRMR and revised IFS methods

The gene expression profiles of three brain regions from six individuals were examined in this study, thereby comprising six datasets. The mRMR method was executed on each of these six datasets and yielded the MaxRel feature list and mRMR feature list. Due to the limitations of our computational power, we only required output of the first 500 features in each of the two lists. The obtained lists are provided in S1 Table.

Each of the six datasets was used as a training dataset, with the other five datasets used as test datasets. Thus, the revised IFS method described in Section "Revised IFS method" was executed six times. Each time, the IFS method constructed 500 feature sets according to the mRMR feature list. The features in each set were used to represent samples in the training dataset and five test datasets. The prediction engine SMO was trained on the training dataset, and its performance was evaluated on the five test datasets. For each test dataset, a series of total prediction accuracies and accuracies of three classes were obtained, all of which are provided in S2 Table. According to the third step of the revised IFS method, the predicted results of each test dataset can produce an IFS curve. Thus, for one training dataset and five test datasets, we can obtain five IFS curves. Six groups of IFS curves are illustrated in S1–S6 Figs. These curves show that for a given training dataset and test dataset, the feature set yielding the maximum value for total prediction accuracy on the test dataset can be obtained, and the numbers of features in these sets are listed in S3 Table. A 3-D histogram was plotted in **Fig 1** to show the number of features in the feature set yielding the maximum total prediction accuracy and the corresponding total prediction accuracy.

From S3 Table and **Fig 1**, we can see that the feature set yielding the maximum total prediction accuracy always contains a large number of features (genes), which makes it difficult to further analyze their importance. Thus, as mentioned in Section "Revised IFS method", we selected an inflection point in each IFS curve. To do that, we amplified each IFS curve between 4 and 50 on the X-axis, as illustrated in **Figs 2–7**. The obtained inflection points and their corresponding total prediction accuracies are presented in S4 Table. Additionally, a 3-D histogram was plotted in **Fig 8** to show the number of features of each inflection point and the corresponding total prediction accuracy. It can be observed from S2 Table and **Fig 8** that the number of selected features (genes) at the inflection point was, in most cases, greatly reduced.

Because there is one mRMR feature list for a given training dataset and five inflection points for five test datasets, *i.e.*, we can obtain five sets of the selected features, we took the intersection operation of these sets to extract important features for a given training dataset. Thus, six sets of important features were obtained, involving 20 features (genes), which are listed in column 1 of **Table 2**. To further demonstrate the importance of these 20 genes, the frequency at which each gene occurred in the six important feature sets was counted, as shown in column 3 of **Table 2**. It can be seen that some genes, such as NR2E1, DAO, and LRRC7, occurred several times, indicating that these genes are the most important for evaluating the differences between the brain regions of different people. The following section provides evidence supporting our results, indicating that our method is effective.

Comparison with previous Allen Human Brain Atlas analysis

The Allen Human Brain Atlas dataset has been analyzed by Myers *et al.* [9]. Comparing with their study, there were several major improvements in our work.

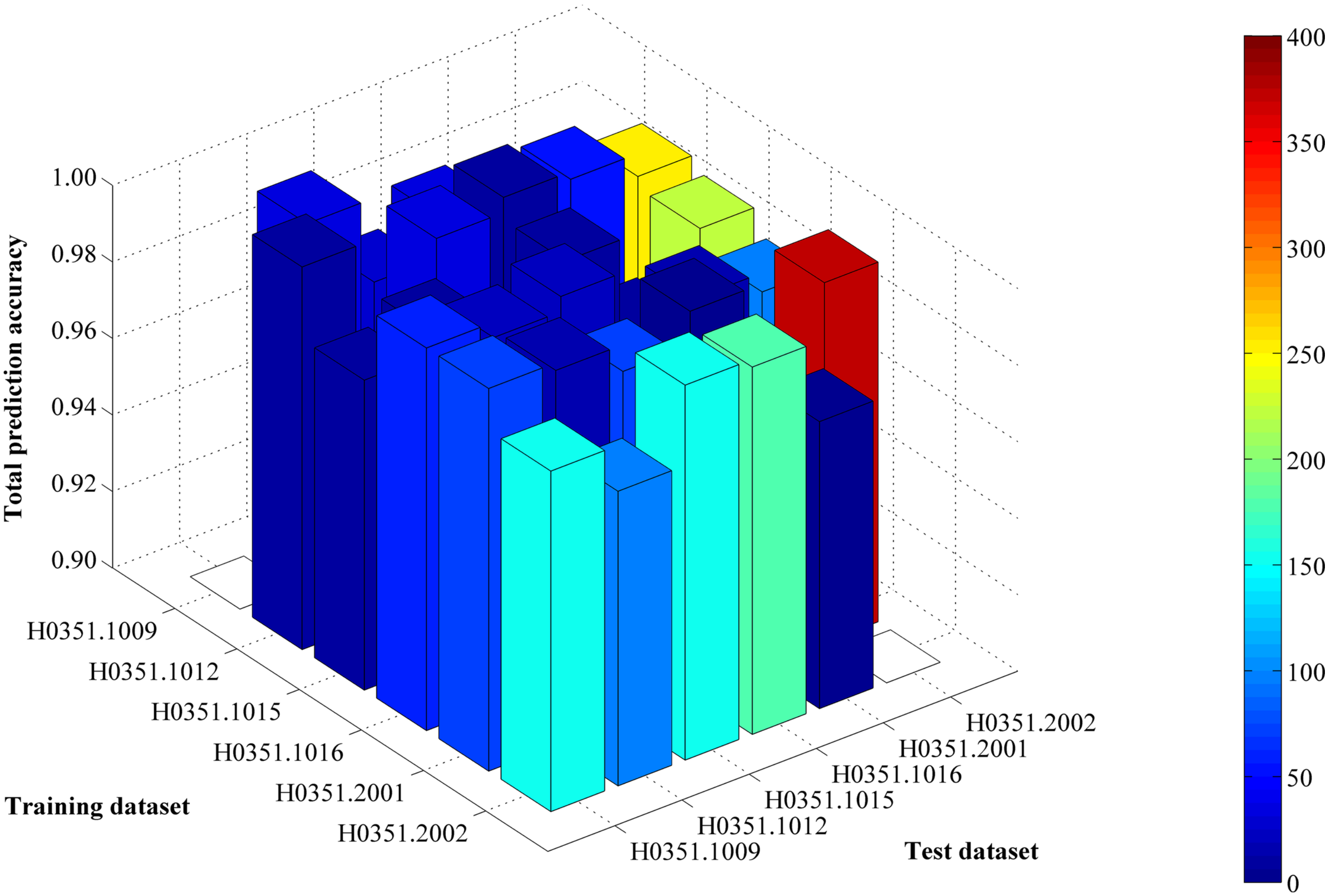

Fig 1. A 3-D histogram illustrating the number of features in the feature set yielding the maximum total prediction accuracy and the corresponding total prediction accuracy. The height of the bar represents the maximum total prediction accuracy, whereas the color of the bar represents the number of features in the feature set yielding the maximum total prediction accuracy.

doi:10.1371/journal.pone.0159395.g001

First, they only used t-test to identify the differentially expressed genes. We adopted the state-of-art machine learning based feature selection method to select the optimal discriminative genes.

Second, they only did statistical test without prediction model. Beside the discriminative genes, the brain region prediction models were constructed in this work.

Third, they did not do cross-individual analysis which is very important for measuring the robustness of selected genes and for analyzing the subtle difference among individuals.

Analysis of the selected genes

In this study, we analyzed the gene expression profiles from three brain regions of six people of different races, ages and sexes. To explore general trends in regionally expressed genes in human brains, we used feature selection methods, specifically the mRMR method and the IFS method, to analyze the gene expression profiles. As a result, 20 genes, along with their frequencies, were identified and are listed in Table 2. Here, we provide evidence to confirm our finding

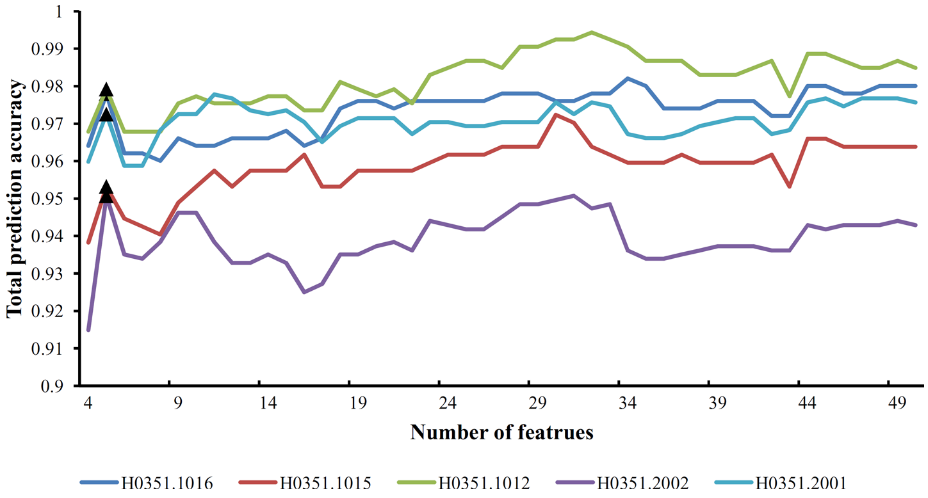

Fig 2. Parts of five IFS-curves using the data of H0351.1009 as the training dataset and the data from other people as the test dataset. The triangle in each curve represents the inflection point.

doi:10.1371/journal.pone.0159395.g002

that these genes are differentially expressed in the brain and important for evaluating the differences between the brain regions of different people.

NR2E1. NR2E1 (nuclear receptor subfamily 2, group E, member 1) was ranked first on our list with a ‘frequency’ of 5. It has been shown to be expressed in a highly regionalized pattern in the brain, and it has important physiological functions in brain development and patterning [36, 37]. Deletion of the *Nr2e1* gene in mouse models results in disorders in brain (hypoplasia of cerebrum and olfactory lobes, abnormal synaptic plasticity and dendritic structure in the mouse dentate gyrus) and eye development and in behavioral abnormalities [38, 39]. Other studies have also suggested that NR2E1 controls the self-renewal of neural stem cells (NSCs) and is involved in the initiation of brain tumors [40].

DAO. DAO (D-amino-acid oxidase) had a ‘frequency’ of 4 and ranked second on our list. The regional expression and distribution of the DAO protein in the rat brain has been investigated and reported in several previous studies [41, 42]. Hindbrain neurons, especially Golgi

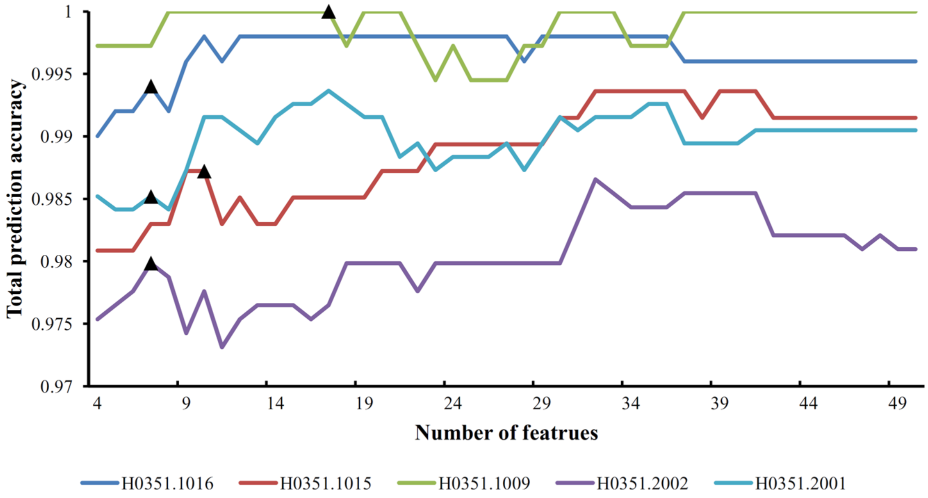

Fig 3. Parts of five IFS-curves using the data of H0351.1012 as the training dataset and the data from other people as the test dataset. The triangle in each curve represents the inflection point.

doi:10.1371/journal.pone.0159395.g003

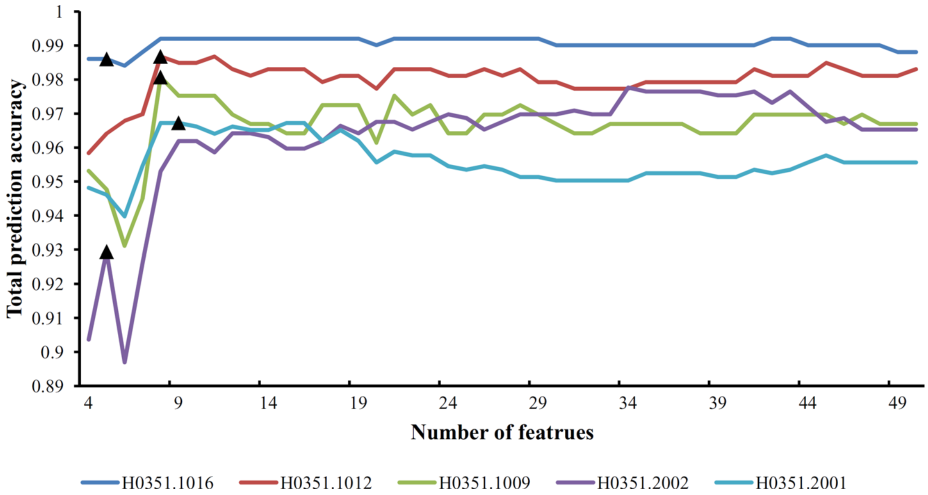

Fig 4. Parts of five IFS-curves using the data of H0351.1015 as the training dataset and the data from other people as the test dataset. The triangle in each curve represents the inflection point.

doi:10.1371/journal.pone.0159395.g004

and Purkinje cells, have been shown to express higher levels of DAO than forebrain neurons. In our study, DAO was also shown to be expressed in a region-specific manner in all 6 human brains. The DAO protein in the rat brain was postulated to function in the elimination of its substrate, D-serine [43], which has been shown to be a major endogenous coagonist of the N-methyl D-aspartate (NMDA) type of glutamate receptors [44]. Because a higher level of DAO is found in the hindbrain, D-serine is more abundant in the forebrain [45]. Previous studies have shown that the hypofunction of the N-Methyl-D-Aspartate (NMDA) type of glutamate receptors in the brain might lead to symptoms similar to those seen in schizophrenia [46].

LRRC7. LRRC7 (leucine rich repeat containing 7, also known as DENSIN) had a 'frequency' of 3 and ranked third on our list. The LRRC7 protein was first purified from rat brains and was identified as a brain-specific synaptic protein of the O-sialoglycoprotein family [47]. Rat densin was more abundant in the forebrain than in the cerebellum [47], which was

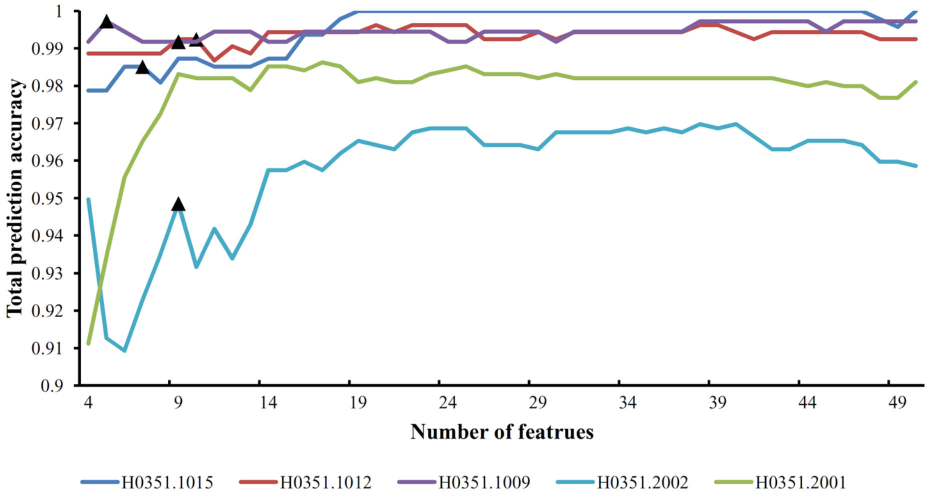

Fig 5. Parts of five IFS-curves using the data of H0351.1016 as the training dataset and the data from other people as the test dataset. The triangle in each curve represents the inflection point.

doi:10.1371/journal.pone.0159395.g005

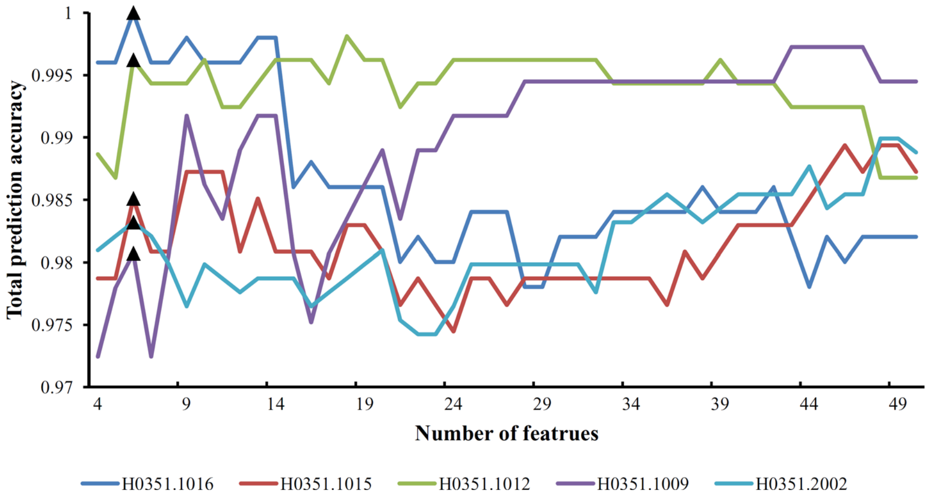

Fig 6. Parts of five IFS-curves using the data of H0351.2001 as the training dataset and the data from other people as the test dataset. The triangle in each curve represents the inflection point.

doi:10.1371/journal.pone.0159395.g006

consistent with our finding that LRRC7 was differentially expressed in the human brain. In the postsynaptic density (PSD), densin can form a high-affinity functional complex with αCaMKII and α-actinin [48]. Loss of densin results in reduced levels of alpha-actinin in the brain and reduced localization of mGluR5 and DISC1 in the PSD fraction, which may play a role in the behavioral endophenotypes of mental illness, as revealed in *Lrrc7* null-mutation mouse models [48]. Interestingly, LRRC7 was observed to be regionally expressed in the brains of the two African American men and the Hispanic woman but not in the three Caucasian men, which suggested that the differential brain expression of LRRC7 might also somewhat be related to race.

Hox family genes. Three Hox (homeobox) family genes were on our list, including HOXA2, with a 'frequency' of 2, and HOXA3 and HOXB2, each with a 'frequency' of 1. Differential expression of these Hox family genes was observed in some of the human brains (*e.g.*,

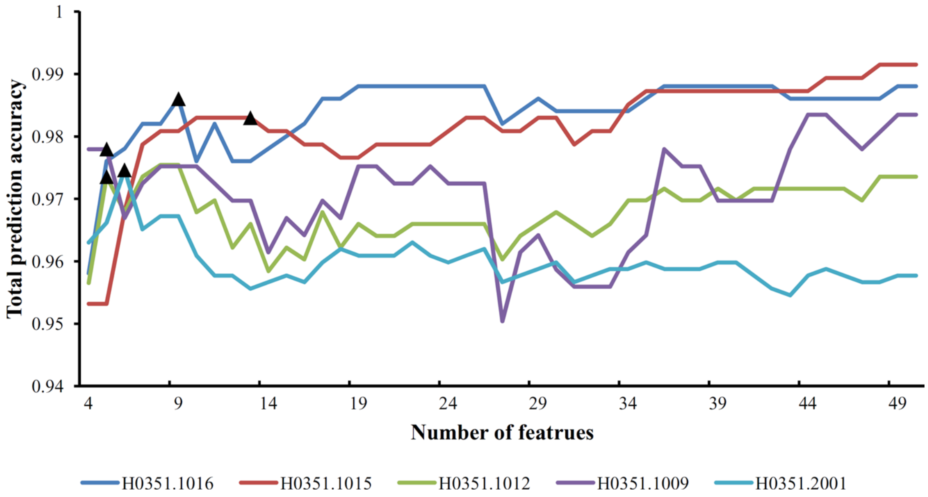

Fig 7. Parts of five IFS-curves using the data of H0351.2002 as the training dataset and the data from other people as the test dataset. The triangle in each curve represents the inflection point.

doi:10.1371/journal.pone.0159395.g007

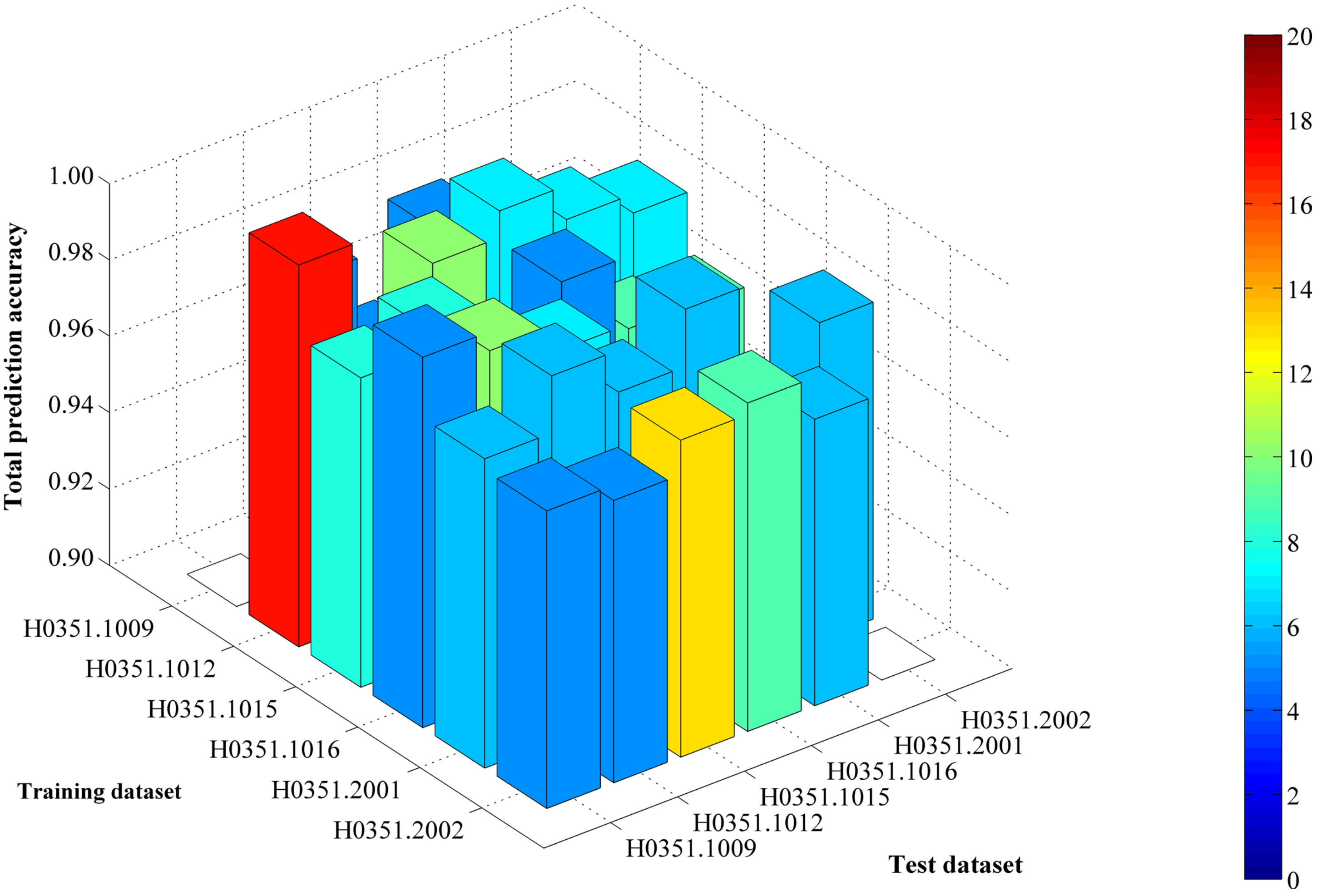

Fig 8. A 3-D histogram illustrating the number of features of each inflection point and the corresponding total prediction accuracy. The height of the bar represents the total prediction accuracy, whereas the color of the bar represents the number of features of each inflection point.

doi:10.1371/journal.pone.0159395.g008

HOXA2 and HOXB2 in the 31-year-old Caucasian and 24-year-old African American men, HOXA3 in the 39-year-old African American man, and HOXA4 in the 57-year-old Caucasian man). Hox genes play pivotal roles in postembryonic brain development and regionalization through their spatiotemporal expression in the central nervous system [49, 50]. Similarly, in our study, HOXA2 and HOXB2 were among the top differentially expressed genes in younger people (age 24 and 31), irrespective of race. In contrast, other Hox family genes were differentially expressed in older people (age 39 and 57).

Besides the above genes, we also found experimental results supporting the regional expression pattern of several other genes in our list, including PAX3 [51, 52], PCP2 [53], ARHGAP4 [54] and NECAB1 [55]. These lines of evidence indicated the effectiveness of our method.

Using our method, we have also identified genes with previously unknown regional expression patterns in the brain, such as TFAP2B, ABLIM1, KLK8, SEC13 and STON1. Our study shed light on the potential important biological functions of these genes in the human brain. Here, we took TFAP2B as an example and gave our analyses.

TFAP2B. The transcription factor AP-2 beta (TFAP2B, also known as AP-2B or AP2-B) was another potential race-specific gene, and it was only observed to be differentially expressed

Table 2. Frequencies of the selected genes.

Gene symbol	Description	Frequency
NR2E1	Nuclear Receptor Subfamily 2, Group E, Member 1	5
DAO	D-Amino-Acid Oxidase	4
LRRC7	Leucine Rich Repeat Containing 7	3
HOXA2	Homeobox A2	2
PAX3	Paired Box 3	2
PCP2	Purkinje Cell Protein 2	2
TFAP2B	Transcription Factor AP-2 Beta (Activating Enhancer Binding Protein 2 Beta)	2
ABLIM1	Actin Binding LIM Protein 1	1
ARHGAP4	Rho GTPase Activating Protein 4	1
FLJ43663	Estrogen Receptor 1	1
HOXA3	Homeobox A3	1
HOXB2	Homeobox B2	1
KHDRBS1	KH Domain Containing, RNA Binding, Signal Transduction Associated 1	1
KLK8	Kallikrein-Related Peptidase 8	1
LOC100287521	—	1
NECAB1	N-Terminal EF-Hand Calcium Binding Protein 1	1
SEC13	Homolog of SEC13 (*S. cerevisiae*)	1
STON1	Stonin 1	1
TRAF3IP1	TNF Receptor-Associated Factor 3 Interacting Protein 1	1
WDR48	WD Repeat Domain 48	1

doi:10.1371/journal.pone.0159395.t002

in the brains of the two Caucasians (age 31 and 57) and not the two African American men or the Hispanic woman. TFAP2B encodes a transcription factor expressed in neural crest cells, stimulating cell proliferation and suppressing terminal differentiation [56]. Interestingly, loss of TFAP2B has been reported to be linked to several race-specific syndromes and diseases. For example, mutation of TFAP2B has been shown to be related to Char syndrome, a familial form of facial dysmorphism, as well as to patent ductus arteriosus (PDA) and hand anomalies (aplasia or hypoplasia of the middle phalanges of the fifth fingers) [57]. Other studies observed mutations of TFAP2B in similar congenital heart disease patients in southern China [58] and south India [59]. These lines of evidence suggest that genetic polymorphism of TFAP2B plays a role in various diseases.

Conclusions

This study investigated the gene expression profiles of three human brain regions from six people of different races, ages, and sexes. Utilizing two major feature selection methods, a number of key genes were identified. These findings may contribute to further studies aimed at elucidating the mechanisms of development of the human brain from a genomics perspective.

Supporting Information

S1 Fig. Five IFS curves using the data of H0351.1009 as the training dataset and the data from other people as the test dataset.
(TIF)

S2 Fig. Five IFS curves using the data of H0351.1012 as the training dataset and the data from other people as the test dataset.
(TIF)

S3 Fig. Five IFS curves using the data of H0351.1015 as the training dataset and the data from other people as the test dataset.
(TIF)

S4 Fig. Five IFS curves using the data of H0351.1016 as the training dataset and the data from other people as the test dataset.
(TIF)

S5 Fig. Five IFS curves using the data of H0351.2001 as the training dataset and the data from other people as the test dataset.
(TIF)

S6 Fig. Five IFS curves using the data of H0351.2002 as the training dataset and the data from other people as the test dataset.
(TIF)

S1 Table. mRMR results for the gene expression profiles of human brains from six individuals.
(DOCX)

S2 Table. The accuracies obtained using the revised IFS method.
(DOCX)

S3 Table. The numbers of features (genes) yielding the maximum total prediction accuracy for pairs of training and test datasets.
(DOCX)

S4 Table. The inflection point for each IFS curve and its corresponding total prediction accuracy.
(DOCX)

S1 Text. Detailed information about the six individuals whose samples were studied.
(PDF)

Acknowledgments

This study was supported by the National Natural Science Foundation of China (61202021, 31371335), Shanghai Natural Science Foundation (16ZR1414500), Shanghai Sailing Program and The Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245).

Author Contributions

Conceived and designed the experiments: LC TH YDC. Performed the experiments: LC XK YDC. Analyzed the data: CC YHZ TH. Contributed reagents/materials/analysis tools: LC CC. Wrote the paper: LC CC CZ.

References

1. Gofflot F, Chartoire N, Vasseur L, Heikkinen S, Dembele D, Le Merrer J, et al. Systematic gene expression mapping clusters nuclear receptors according to their function in the brain. Cell. 2007; 131(2):405–18. Epub 2007/10/25. doi: 10.1016/j.cell.2007.09.012 PMID: 17956739.

2. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science. 1975; 188 (4184):107–16. Epub 1975/04/11. PMID: 1090005.

3. Konopka G, Friedrich T, Davis-Turak J, Winden K, Oldham MC, Gao F, et al. Human-specific transcriptional networks in the brain. Neuron. 2012; 75(4):601–17. Epub 2012/08/28. doi: 10.1016/j.neuron.2012.05.034 PMID: 22920253; PubMed Central PMCID: PMC3645834.

4. Preuss TM, Caceres M, Oldham MC, Geschwind DH. Human brain evolution: insights from microarrays. Nat Rev Genet. 2004; 5(11):850–60. Epub 2004/11/03. doi: 10.1038/nrg1469 PMID: 15520794.

5. Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, et al. Intra- and interspecific variation in primate gene expression patterns. Science. 2002; 296(5566):340–3. Epub 2002/04/16. doi: 10.1126/science.1068996 PMID: 11951044.

6. Khaitovich P, Muetzel B, She X, Lachmann M, Hellmann I, Dietzsch J, et al. Regional patterns of gene expression in human and chimpanzee brains. Genome Res. 2004; 14(8):1462–73. Epub 2004/08/04. doi: 10.1101/gr.2538704 PMID: 15289471; PubMed Central PMCID: PMC509255.

7. Giger T, Khaitovich P, Somel M, Lorenc A, Lizano E, Harris LW, et al. Evolution of neuronal and endothelial transcriptomes in primates. Genome Biol Evol. 2010; 2:284–92. Epub 2010/07/14. doi: 10.1093/gbe/evq018 PMID: 20624733; PubMed Central PMCID: PMC2998193.

8. Bauernfeind AL, Soderblom EJ, Turner ME, Moseley MA, Ely JJ, Hof PR, et al. Evolutionary Divergence of Gene and Protein Expression in the Brains of Humans and Chimpanzees. Genome Biol Evol. 2015; 7(8):2276–88. Epub 2015/07/15. doi: 10.1093/gbe/evv132 PMID: 26163674; PubMed Central PMCID: PMC4558850.

9. Myers EM, Bartlett CW, Machiraju R, Bohland JW. An integrative analysis of regional gene expression profiles in the human brain. Methods. 2015; 73:54–70. Epub 2014/12/20. doi: 10.1016/j.ymeth.2014.12.010 PMID: 25524419.

10. The Lancet N. The Human Brain Project: mutiny on the flagship. Lancet Neurol. 2014; 13(9):855. Epub 2014/08/22. doi: 10.1016/S1474-4422(14)70181-4 PMID: 25142450.

11. Shen EH, Overly CC, Jones AR. The Allen Human Brain Atlas: comprehensive gene expression mapping of the human brain. Trends Neurosci. 2012; 35(12):711–4. Epub 2012/10/09. doi: 10.1016/j.tins.2012.09.005 PMID: 23041053.

12. Fernandez-Irigoyen J, Zelaya MV, Perez-Valderrama E, Santamaria E. New insights into the human brain proteome: Protein expression profiling of deep brain stimulation target areas. Journal of proteomics. 2015. Epub 2015/04/08. doi: 10.1016/j.jprot.2015.03.032 PMID: 25845585.

13. Mahfouz A, van de Giessen M, van der Maaten L, Huisman S, Reinders M, Hawrylycz MJ, et al. Visualizing the spatial gene expression organization in the brain through non-linear similarity embeddings. Methods. 2015; 73:79–89. Epub 2014/12/03. doi: 10.1016/j.ymeth.2014.10.004 PMID: 25449901.

14. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(8):1226–38. PMID: 16119262

15. Platt J, editor. Fast training of support vector machines using sequential minimal optimization. Cambridge, MA: MIT Press; 1998.

16. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation. 2001; 13(3):637–49.

17. Chen L, Zeng W-M, Cai Y-D, Huang T. Prediction of Metabolic Pathway Using Graph Property, Chemical Functional Group and Chemical Structural Set. Current Bioinformatics. 2013; 8(2):200–7.

18. Mohabatkar H, Mohammad Beigi M, Abdolahi K, Mohsenzadeh S. Prediction of Allergenic Proteins by Means of the Concept of Chous Pseudo Amino Acid Composition and a Machine Learning Approach. Medicinal Chemistry. 2013; 9(1):133–7. PMID: 22931491

19. Chen L, Chu C, Huang T, Kong X, Cai Y-D. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. Amino acids. 2015; 47(7):1485–93. doi: 10.1007/s00726-015-1974-5 PMID: 25894890

20. Mohabatkar H, Mohammad Beigi M, Esmaeili A. Prediction of GABAA receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. Journal of Theoretical Biology. 2011; 281(1):18–23. doi: 10.1016/j.jtbi.2011.04.017 PMID: 21536049

21. Li B-Q, Feng K-Y, Chen L, Huang T, Cai Y-D. Prediction of Protein-Protein Interaction Sites by Random Forest Algorithm with mRMR and IFS. PLoS ONE. 2012; 7(8):e43927. doi: 10.1371/journal.pone.0043927 PMID: 22937126

22. Liu T, Hu L, Ma C, Wang Z-Y, Chen H-L. A fast approach for detection of erythemato-squamous diseases based on extreme learning machine with maximum relevance minimum redundancy feature selection. International Journal of Systems Science. 2015; 46(5):919–31. doi: 10.1080/00207721.2013.801096

23. Chen L, Chu C, Lu J, Kong X, Huang T, Cai YD. Gene Ontology and KEGG Pathway Enrichment Analysis of a Drug Target-Based Classification System. PLoS ONE. 2015; 10(5):e0126492. Epub 2015/05/08. doi: 10.1371/journal.pone.0126492 PMID: 25951454; PubMed Central PMCID: PMC4423955.

24. Gui T, Dong X, Li R, Li Y, Wang Z. Identification of Hepatocellular Carcinoma–Related Genes with a Machine Learning and Network Analysis. Journal of Computational Biology. 2015; 22(1):63–71. doi: 10.1089/cmb.2014.0122 PMID: 25247452

25. Chen L, Chu C, Feng K. Predicting the types of metabolic pathway of compounds using molecular fragments and sequential minimal optimizatio. Combinatorial Chemistry & High Throughput Screening. 2016; 19(2):136–43.

26. Zhang PW, Chen L, Huang T, Zhang N, Kong XY, Cai YD. Classifying ten types of major cancers based on reverse phase protein array profiles. PLoS ONE. 2015; 10(3):e0123147. Epub 2015/03/31. doi: 10.1371/journal.pone.0123147 PMID: 25822500.

27. Xu Z, Dai M, Meng D. Fast and efficient strategies for model selection of Gaussian support vector machine. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on. 2009; 39 (5):1292–307.

28. Fan R-E, Chen P-H, Lin C-J. Working set selection using second order information for training support vector machines. The Journal of Machine Learning Research. 2005; 6:1889–918.

29. Torii Y, Abe S. Decomposition techniques for training linear programming support vector machines. Neurocomputing. 2009; 72(4):973–84.

30. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2(3):27.

31. Chang C-C, Hsu C-W, Lin C-J. The analysis of decomposition methods for support vector machines. IEEE Transactions on Neural Networks. 2000; 11(4):1003–8. doi: 10.1109/72.857780 PMID: 18249827

32. Lin C-J. On the convergence of the decomposition method for support vector machines. IEEE Transactions on Neural Networks. 2001; 12(6):1288–98. doi: 10.1109/72.963765 PMID: 18249958

33. Keerthi SS, Gilbert EG. Convergence of a generalized SMO algorithm for SVM classifier design. Machine Learning. 2002; 46(1–3):351–60.

34. Hastie T, Tibshirani R. Classification by pairwise coupling. Proceedings of the 1997 conference on Advances in neural information processing systems 10; Denver, Colorado, USA. 302744: MIT Press; 1998. p. 507–13.

35. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann Pub; 2005.

36. Stenman J, Yu RT, Evans RM, Campbell K. Tlx and Pax6 co-operate genetically to establish the pallio-subpallial boundary in the embryonic mouse telencephalon. Development. 2003; 130(6):1113–22. Epub 2003/02/07. PMID: 12571103.

37. Stenman JM, Wang B, Campbell K. Tlx controls proliferation and patterning of lateral telencephalic progenitor domains. J Neurosci. 2003; 23(33):10568–76. Epub 2003/11/25. PMID: 14627641.

38. Young KA, Berry ML, Mahaffey CL, Saionz JR, Hawes NL, Chang B, et al. Fierce: a new mouse deletion of Nr2e1; violent behaviour and ocular abnormalities are background-dependent. Behav Brain Res. 2002; 132(2):145–58. Epub 2002/05/09. PMID: 11997145; PubMed Central PMCID: PMC2862907.

39. Christie BR, Li AM, Redila VA, Booth H, Wong BK, Eadie BD, et al. Deletion of the nuclear receptor Nr2e1 impairs synaptic plasticity and dendritic structure in the mouse dentate gyrus. Neuroscience. 2006; 137(3):1031–7. Epub 2005/11/18. doi: 10.1016/j.neuroscience.2005.08.091 PMID: 16289828.

40. O'Loghlen A, Martin N, Krusche B, Pemberton H, Alonso MM, Chandler H, et al. The nuclear receptor NR2E1/TLX controls senescence. Oncogene. 2014; 34(31):4069–77. Epub 2014/10/21. doi: 10.1038/onc.2014.335 PMID: 25328137; PubMed Central PMCID: PMC4305339.

41. Horiike K, Tojo H, Arai R, Nozaki M, Maeda T. D-amino-acid oxidase is confined to the lower brain stem and cerebellum in rat brain: regional differentiation of astrocytes. Brain research. 1994; 652(2):297–303. Epub 1994/08/01. PMID: 7953743.

42. Moreno S, Nardacci R, Cimini A, Ceru MP. Immunocytochemical localization of D-amino acid oxidase in rat brain. J Neurocytol. 1999; 28(3):169–85. Epub 2000/01/05. PMID: 10617900.

43. Nagata Y. Involvement of D-amino acid oxidase in elimination of D-serine in mouse brain. Experientia. 1992; 48(8):753–5. Epub 1992/08/15. PMID: 1355447.

44. Madeira C, Freitas ME, Vargas-Lopes C, Wolosker H, Panizzutti R. Increased brain D-amino acid oxidase (DAAO) activity in schizophrenia. Schizophr Res. 2008; 101(1–3):76–83. Epub 2008/04/02. doi: 10.1016/j.schres.2008.02.002 PMID: 18378121.

45. Yamanaka M, Miyoshi Y, Ohide H, Hamase K, Konno R. D-Amino acids in the brain and mutant rodents lacking D-amino-acid oxidase activity. Amino Acids. 2012; 43(5):1811–21. Epub 2012/08/16. doi: 10.1007/s00726-012-1384-x PMID: 22892863.

46. Strick CA, Li C, Scott L, Harvey B, Hajos M, Steyn SJ, et al. Modulation of NMDA receptor function by inhibition of D-amino acid oxidase in rodent brain. Neuropharmacology. 2011; 61(5–6):1001–15. Epub 2011/07/19. doi: 10.1016/j.neuropharm.2011.06.029 PMID: 21763704.

47. Apperson ML, Moon IS, Kennedy MB. Characterization of densin-180, a new brain-specific synaptic protein of the O-sialoglycoprotein family. J Neurosci. 1996; 16(21):6839–52. Epub 1996/11/01. PMID: 8824323.

48. Carlisle HJ, Luong TN, Medina-Marino A, Schenker L, Khorosheva E, Indersmitten T, et al. Deletion of densin-180 results in abnormal behaviors associated with mental illness and reduces mGluR5 and DISC1 in the postsynaptic density fraction. J Neurosci. 2011; 31(45):16194–207. Epub 2011/11/11. doi: 10.1523/JNEUROSCI.5877-10.2011 PMID: 22072671; PubMed Central PMCID: PMC3235477.

49. Reichert H, Bello B. Hox genes and brain development in Drosophila. Advances in experimental medicine and biology. 2010; 689:145–53. Epub 2010/08/28. PMID: 20795329.

50. Krumlauf R, Holland PW, McVey JH, Hogan BL. Developmental and spatial patterns of expression of the mouse homeobox gene, Hox 2.1. Development. 1987; 99(4):603–17. Epub 1987/04/01. PMID: 2889591.

51. Joven A, Morona R, Gonzalez A, Moreno N. Spatiotemporal patterns of Pax3, Pax6, and Pax7 expression in the developing brain of a urodele amphibian, Pleurodeles waltl. J Comp Neurol. 2013; 521 (17):3913–53. Epub 2013/06/21. doi: 10.1002/cne.23385 PMID: 23784810.

52. Matsunaga E, Araki I, Nakamura H. Role of Pax3/7 in the tectum regionalization. Development. 2001; 128(20):4069–77. Epub 2001/10/20. PMID: 11641229.

53. Mohn AR, Feddersen RM, Nguyen MS, Koller BH. Phenotypic analysis of mice lacking the highly abundant Purkinje cell- and bipolar neuron-specific PCP2 protein. Mol Cell Neurosci. 1997; 9(1):63–76. Epub 1997/01/01. doi: 10.1006/mcne.1997.0606 PMID: 9204480.

54. Foletta VC, Brown FD, Young WS 3rd. Cloning of rat ARHGAP4/C1, a RhoGAP family member expressed in the nervous system that colocalizes with the Golgi complex and microtubules. Brain research Molecular brain research. 2002; 107(1):65–79. Epub 2002/11/05. PMID: 12414125.

55. Wu H, Li D, Shan Y, Wan B, Hexige S, Guo J, et al. EFCBP1/NECAB1, a brain-specifically expressed gene with highest abundance in temporal lobe, encodes a protein containing EF-hand and antibiotic biosynthesis monooxygenase domains. DNA Seq. 2007; 18(1):73–9. Epub 2007/03/17. doi: 10.1080/10425170500511271 PMID: 17364817.

56. Eckert D, Buhl S, Weber S, Jager R, Schorle H. The AP-2 family of transcription factors. Genome Biol. 2005; 6(13):246. Epub 2006/01/20. doi: 10.1186/gb-2005-6-13-246 PMID: 16420676; PubMed Central PMCID: PMC1414101.

57. Satoda M, Zhao F, Diaz GA, Burn J, Goodship J, Davidson HR, et al. Mutations in TFAP2B cause Char syndrome, a familial form of patent ductus arteriosus. Nat Genet. 2000; 25(1):42–6. Epub 2000/05/10. doi: 10.1038/75578 PMID: 10802654.

58. Xiong F, Li Q, Zhang C, Chen Y, Li P, Wei X, et al. Analyses of GATA4, NKX2.5, and TFAP2B genes in subjects from southern China with sporadic congenital heart disease. Cardiovasc Pathol. 2013; 22 (2):141–5. Epub 2012/09/11. doi: 10.1016/j.carpath.2012.07.001 PMID: 22959235.

59. Lingaiah K, Sosalagere DM, Mysore SR, Krishnamurthy B, Narayanappa D, Nallur RB. Mutations of TFAP2B in congenital heart disease patients in Mysore, South India. Indian J Med Res. 2011; 134 (5):621–6. Epub 2011/12/27. doi: 10.4103/0971-5916.90986 PMID: 22199100; PubMed Central PMCID: PMC3249959.