Received April 1, 2020, accepted April 19, 2020, date of publication May 18, 2020, date of current version June 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2994646

Identifying Essential Methylation Patterns and Genes Associated With Stroke

XIANGTIAN YU[1], ZIJUN GAN[2], YAN XU[3], SIBAO WAN[3], MIN LI[3], SHIJIAN DING[3], AND TAO ZENG[4,5]

[1] Clinical Research Center, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai 200233, China
[2] Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China
[3] School of Life Sciences, Shanghai University, Shanghai 200444, China
[4] Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai 201210, China
[5] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China

Corresponding authors: Xiangtian Yu (graceyu1985@163.com) and Tao Zeng (zengtao@bsbii.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61803360 and Grant 11871456, and in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01.

ABSTRACT Stroke is a serious lethal factor for human beings, and thus its unique pathogenic factors and underlying molecular mechanisms must be thoroughly investigated. Current complicated examination and treatment approaches on stroke reflect the complicated pathogenesis of stroke, which involves two major pathogenic factors: phenotypic characteristic and genetic background. Stroke occurrences with different symptoms and pathogenic characteristics may be induced by genetic variations. However, epigenetic contribution and regulation on stroke pathogenesis have been neglected for a long time and thus environmental influence on stroke onset is often underestimated. In this study, relied on our newly presented computational method, we re-screened out the genome methylation data of stroke patients with different subtypes and identified a group of functional methylated or demethylated genes. Recent reports validated the abnormal methylation status of the all identified genes in the pathogenesis of stroke. The genes were associated with biological functions involved in stroke onset. Further functional enrichment analysis confirmed and summarized the novel specific pathogenic roles of ion binding and focal adhesion in the regulation of stroke at the methylation level.

INDEX TERMS Stroke, methylation, pattern, multi-class classification.

I. INTRODUCTION

Stroke is a medical emergency with fast onset. In a typical clinic phenotype of stroke, blood supply to the brain is greatly interrupted or reduced [1], [2]. Stroke cases can be divided into two subgroups according to etiological features: ischemic stroke induced by lack of blood flow [3] and hemorrhagic stroke induced by excessive bleeding [4]. Both subtypes can trigger typical stroke symptoms, such as dysfunction in some brain regions and blockage of blood supply to crucial brain regions [5]. In the USA, stroke is the leading cause of death, affecting more than 800,000 people annually [6]. Therefore, given that stroke is a serious lethal factor for human beings, the unique pathogenic factors and underlying molecular mechanisms of stroke must be deeply explored.

Clinically, high-risk cases of stroke have various typical symptoms [3], [6], and patients with such symptoms are recommended to take medical treatments. For the accurate clinical diagnosis of stroke, various diagnostic approaches are generally applied, including CT (Computed Tomography), MRI (Magnetic Resonance Imaging), carotid ultrasound, and cerebral angiogram [7], [8]. Moreover, therapy approaches for stroke during follow-up clinical treatments after diagnosis vary among different patients due to their respective pathogenesis. For instance, patients suffering from ischemic stroke are generally recommended to be treated with tissue plasminogen activators within 4.5 hours [9]. By contrast, patients suffering from hemorrhagic stroke are usually treated with direct surgery rather than drugs [10].

Such complicated examination and treatment approaches on stroke actually reflect the complicated pathogenesis of stroke. The two major pathogenic factors of stroke are phenotypic characteristic and genetic background [5]. On the

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

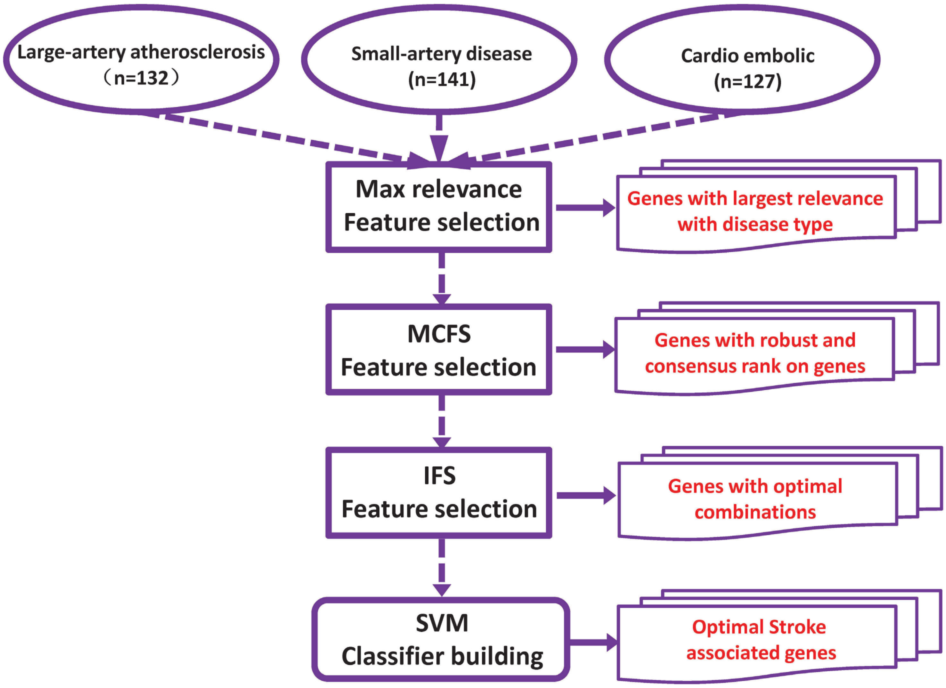

FIGURE 1. **Whole pipeline for classifying samples from different stroke subtypes.**

one hand, various phenotypic characteristics including overweight, age, inactive lifestyle, smoking, and drinking have been tightly connected to the stroke onset, according to recent publication reports [11]–[13]. On the other hand, the genetic contribution of stroke has been gradually revealed along with the development of next generation sequencing. Early in 2011, a systematic review on the genetic background of stroke revealed a set of genes with potential functional contribution to the pathogenesis of stroke [14], [15], and these genes were clustered into five groups by onset position, affected tissues and potential biological mechanisms of stroke. Some works suggested that stroke onset is affected by genetic contributions and stroke occurrences with different symptoms and pathogenic characteristics may be induced by genetic variations [16], [17].

Although genetic studies on stroke have increasingly expressed interest in genes and their variants, the epigenetic contribution and regulation on stroke pathogenesis have been neglected for a long time and have not been deeply studied. Therefore, in this study, we re-screened the genome methylation data of stroke patients from a recent publication [18], [19]. Basing on our newly presented computational method, we identified a group of functional methylated or demethylated genes that may participate as potential biomarkers in the pathogenesis of different subtypes of stroke. All predicted the genes and their enriched functions were confirmed to be associated with the stroke cases in recent reports. All in all, all the screened genes identified by our newly presented computational method not only may reflect the methylation pattern of stroke patients but also may contribute to the identification of potential detailed pathogenic mechanisms for stroke initiation and progression.

II. MATERIALS AND METHODS

A. DATASETS

We downloaded the methylation profiles of patients with stroke from Gene Expression Omnibus under accession number of GSE69138 [18], [19]. Previous genome-wide methylation study extracted whole-blood DNA from 404 patients with ischemic stroke, which were distributed across three ischemic stroke subtypes: 132 patients with large-artery atherosclerosis (LAA), 141 patients with small-artery disease (SAD), and 127 patients with cardio embolic (CE). The DNA methylation in CpG sites were measured by Illumina HumanMethylation450 BeadChip array.

B. FEATURE SELECTION

Max relevance feature selection method was first used in the selection of relevant features (i.e., methylation sites), and only relevant features with scores greater than the predefined cutoff were retained. Then, a ranked feature list was obtained by feeding the remaining features into Monte Carlo feature selection (MCFS). Based on the ranked features, increment feature selection (IFS) adopting support vector machine (SVM) as the classifier was further used in the selection of optimum features with the best discrimination performance in classifying samples from the different subtypes of stroke. The whole analysis framework is shown in FIGURE 1.

C. MAX RELEVANCE SCORE

In general, if a feature (i.e. one methylation site) is highly relevant to the outcome variable (i.e. disease phenotype), it is considered important. The maximum relevance score calculates the mutual information (MI) between features and class labels [20]–[22]. The higher the score is, the more important this feature is [23], [24]. The MI of variables x and y is defined as follows:

$$I(x, y) = \iint p(x, y) log \frac{p(x, y)}{p(x)P(y)} dxdy$$

where $p(x)$ and $p(y)$ is the marginal probability density for x and y, respectively, and $p(x, y)$ is the joint probability density. In this study, we used the MI program integrated in the minimum redundancy and maximum relevance (mRMR) package, to calculate the MI between features and class labels.

D. MONTE CARLO FEATURE SELECTION

Monte Carlo feature selection (MCFS) is a newly proposed feature selection approach based on several decision trees on several bootstrap sets [25]–[27]. Each bootstrap set is randomly produced from an original training set [28]. MCFS first generates t feature subsets, each of which contains m features that are randomly selected from original M features. Then, for each feature subset, p decision trees are generated based on the p bootstrap sample sets, in which samples are represented by features in a given feature subset. The abovementioned procedure is executed t times for the t feature subsets. Accordingly, t*p trees are constructed. The relative importance (RI) score for each feature is measured in terms of the number of observation times of this feature in all constructed trees. MCFS package [28] retrieved from http://www.ipipan.eu/staff/m.draminski/mcfs.html was used in this study. For convenience, default parameters of MCFS were used to execute, i.e. $t = 2000$ and $p = 5$. Obviously, such RI score indicates the importance of features relevant to class labels. The higher RI score is, the more important the feature is; thus, features can be further ranked with the decreasing order of their RI scores.

E. INCREMENT FEATURE SELECTION

Not all features are needed to show excellent performance. Thus, IFS [29], [30] was used to select optimum features for a supervised prediction model (i.e. multi-class classifier). It first generates a series of feature subsets with a step interval of 10 according to the ranked features yielded by MCFS. For example, feature subset 1 has only the top 10 features, and feature subset 2 has the top 20 features, and so on. Then, for each feature subset, a supervised classifier (i.e. support vector machine) is trained and evaluated on the samples represented by features in this feature subset. In the end, we selected the feature subset yielding the best performance during 10-fold cross-validation as the optimum feature (e.g., optimal methylated/demethylated genes).

F. CLASSIFICATION MODEL

Support vector machine (SVM) is a widely applied supervised classification model under a statistical framework [31], [32]. It finds an optimal hyperplane between two classes of samples to make the margin maximum. Clearly, the margin is closely related to generalization error. Furthermore, SVM can efficiently handle linear and nonlinear data, as it can map the original data into a linear space with high dimension through kernel trick. The SVM has an effective mathematical theory for solving convex objective function with the global minimum and is appropriate for data with nonlinear structures. Additionally, it requires appropriate kernels and few tuning parameters. In this study, three groups of samples were considered, so a multi-class SVM adopting one-versus-rest strategy for multiclass classification was adopted. The polynomial function was set as the kernel function, and the regularization parameter was set to 1.

G. PERFORMANCE MEASUREMENT

In this work, there are three classes of samples, so that, a multiclass SVM classifier was learned. The individual accuracies of the three classes and the overall accuracy and Matthews correlation coefficient (MCC) [33], [34] were calculated for objective performance evaluation, combined with 10-fold cross-validation. Given that X is the predicted labels of samples and Y is the true labels, the MCC is calculated as

$$\begin{aligned} MCC &= \frac{cov(X, Y)}{\sqrt{cov(X, X)cov(Y, Y)}} \\ &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{C}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}\sum_{j=1}^{C}(y_{ij} - \bar{y}_j)^2}} \end{aligned}$$

where $\bar{x}_j$ and $\bar{y}_j$ are the average values of members in the j-th column of X and j-th column of Y, respectively. Besides, the accuracy (ACC) for each class can also be estimated by

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP and FP indicate the number of true-positive and false-positive samples respectively, while TN and FN indicate the number of true-negative and false-negative samples respectively.

H. BIOLOGICAL ENRICHMENT ANALYSIS

According to the probe annotation table of Illumina HumanMethylation450 BeadChip [18], [19], our selected methylation probes were mapped onto detailed genes, which were then enriched onto GO and KEGG through a hypergeometric test [35]. The GO and KEGG terms with false discovery rate smaller than 0.05 were considered as relevant biological functions with significant enrichment on our selected features.

TABLE 1. Performance of IFS with SVM and RF.

Classifier	SVM	RF
Number of optimum features	4730	790
Overall (MCC)	0.892	0.556
Large-artery atherosclerosis (ACC)	0.953	0.661
Small-artery disease (ACC)	0.936	0.801
Cardio embolic (ACC)	0.894	0.636

III. RESULTS

The input features were ranked as 482,421 methylation sites by their maximum relevance scores, and only 19,491 highly relevant features with score of >0.015 were selected. Then, the discriminative features for classifying stroke samples from different disease subgroups were obtained by further refining these relevant features.

The optimal feature combination was determined through a supervised classifier by running IFS with SVM for sample classification. During this process, a series of feature subsets with a step interval of 10 was generated, and the SVM with 10-fold cross-validation were run on the samples consisting of features from one feature subset and validated. This operation was performed in each feature subset. The best MCC (0.892) was obtained when the top 4730 features were used, and the overall accuracy was 0.928. The features along with their importance scores calculated by MCFS are shown in TABLE S1. Performance corresponding to all feature subsets is provided in TABLE S2. An IFS curve is illustrated (FIGURE 2A) with the MCC value (i.e. y) as y-axis and the number of features (i.e. x∗10) as x-axis. We ran IFS in the same way, with random forest as the supervised classifier, to verify the selection of SVM as a supervised classifier in this work. The best MCC value of 0.556 was obtained when the top 790 features were used, and the overall accuracy was 0.703. Performance corresponding to all the feature subsets is provided in TABLE S2 as well. FIGURE 2B illustrates the MCCs that the RFs yielded when the number of features involved varied. The results in TABLE 1 and FIGURE 2 indicated that SVM outperforms RF and is thus a better choice for classifying samples from dissimilar stroke subtypes.

IV. DISCUSSION

Basing on our newly presented computational method in FIGURE 1, we screened a group of functional genes that have abnormal methylation status contributing to stroke pathogenesis. Of note, we adopted a two-stage feature selection procedure, considering many features required filtering. In the first stage, the Max's relevance score (MI value) of each feature was calculated, thereby discarding lots of irrelevant features. Then in the second stage, the RI score of each remaining feature was computed to obtain final relevant features. Thus, the Max's relevance score and MCFS had different functions and were both necessary in our analysis, providing an efficient feature selection procedure for 482,421 methylation sites. The final optimal genes and enriched functional terms are analyzed and discussed below.

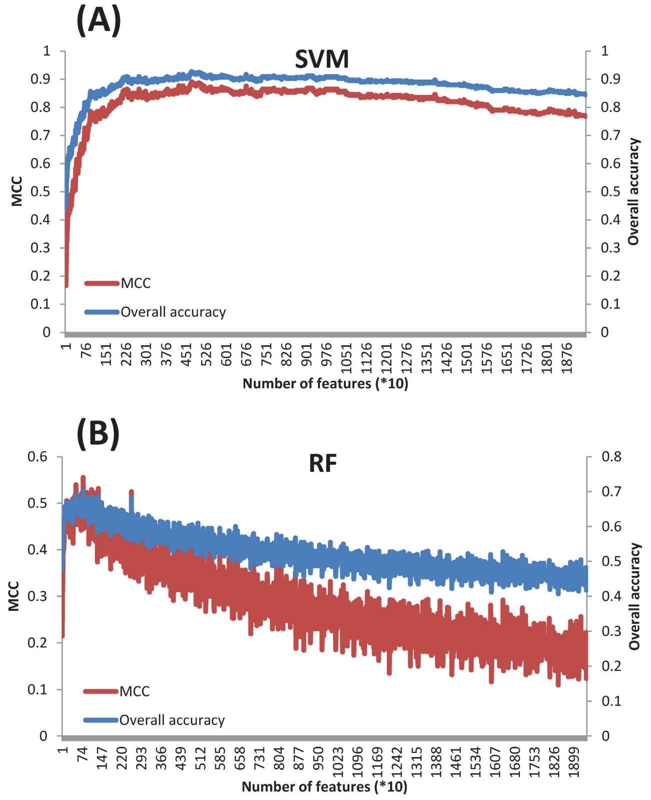

FIGURE 2. IFS curve for SVM and RF with the number of features involved from 19,491 highly relevant features.

A. OPTIMAL MEHTYLATED/DEMETHYLATED GENES

Recent publication confirmed the identified optimal methylated/demethylated genes with high ranks (i.e., the top 5 genes).

TSPY4, encoding a specific functional gene participating in sperm differentiation and proliferation, is a potential stroke-associated gene with respective methylation abnormality [36]. The methylation regulation of TSPY4 and its homologues have been widely identified during tumorigenesis and found to interact with Y chromosome-located oncogenes [37]–[39]. Moreover, the specific gene expression patterns of genes in the Y chromosome may directly participate in the pathogenesis of ischemic stroke [40]. Therefore, indirectly, our predicted TSPY4 with differential methylation may influence the expression of Y chromosome-located genes during the pathogenesis of stroke.

LHB, as a glycoprotein-hormone-encoding gene, has been predicted to be abnormally regulated at the methylation level [41], [42] during the initiation and progression of stroke. As reported, an abnormal sex hormone (e.g., testosterone in

TABLE 2. Representative functional terms of enrichment analysis.

Gene.set ID	Gene.set Name	FDR	P-value	Category
GO:0009653	anatomical structure morphogenesis	2.77E-15	1.75E-19	GO(BP)
GO:0007399	nervous system development	1.95E-14	2.46E-18	GO(BP)
GO:0048856	anatomical structure development	1.05E-12	1.99E-16	GO(BP)
GO:0032502	developmental process	1.28E-11	3.24E-15	GO(BP)
GO:0007275	multicellular organism development	1.03E-09	3.27E-13	GO(BP)
GO:0048731	system development	1.65E-09	6.27E-13	GO(BP)
GO:0005622	intracellular	7.33E-06	9.11E-09	GO(CC)
GO:0043226	organelle	7.33E-06	1.16E-08	GO(CC)
GO:0043227	membrane-bounded organelle	7.33E-06	1.01E-08	GO(CC)
GO:0044424	intracellular part	1.96E-05	4.12E-08	GO(CC)
GO:0097458	neuron part	6.66E-05	1.75E-07	GO(CC)
GO:0005488	binding	9.59E-10	2.08E-13	GO(MF)
GO:0043167	ion binding	6.02E-09	3.92E-12	GO(MF)
GO:0043565	sequence-specific DNA binding	6.02E-09	2.72E-12	GO(MF)
GO:0046872	metal ion binding	5.30E-07	4.61E-10	GO(MF)
GO:0043169	cation binding	4.77E-06	5.18E-09	GO(MF)
hsa04360	Axon guidance	0.038882	0.000263	KEGG
hsa04510	Focal adhesion	0.038882	0.000611	KEGG
hsa04512	ECM-receptor interaction	0.038882	0.000421	KEGG
hsa04940	Type I diabetes mellitus	0.038882	0.000336	KEGG
hsa05222	Small cell lung cancer	0.038882	0.000489	KEGG
hsa05220	Chronic myeloid leukemia	0.048533	0.000916	KEGG

the blood) may participate in the pathogenesis of stroke [43]. Given that LHB is the encoding gene of a quiet effective sex hormone (i.e., luteinizing hormone), its methylation status may affect gene/protein expression level and is potentially pathogenic for our objective disease/symptom, stroke.

FAM153C is another identified stroke-associated gene. Binding to its related regulatory factors, FAM153C has only been reported to participate in human height regulation. According to a recent publication on voltage-gated proton channels in sperm, which is functionally connected to FAM153C [44], the proton channels are potential targets of novel drugs with specific side effects in stroke. Therefore, despite the lack of evidence, the methylation status of our predicted gene FAM153C may be functionally related to stroke symptoms.

GRIN2A, encoding a member of the glutamate-gated ion channel, has been predicted to contribute to stroke-related abnormally methylation regulation. A systematic analysis on a typical disease named Landau-Kleffner [45] confirmed that the expression and mutational pattern of GRIN2A may contribute to the pathogenesis of the disease as one of the core driver at the genetic and epigenetic levels. Further studies identified that stroke is one of its major complications of Landau-Kleffner which would be induced by the defected mitochondrial respiratory chain [46]. Therefore, these facts indicate the methylation of GRIN2A has potential to involve in the pathogenesis of stroke.

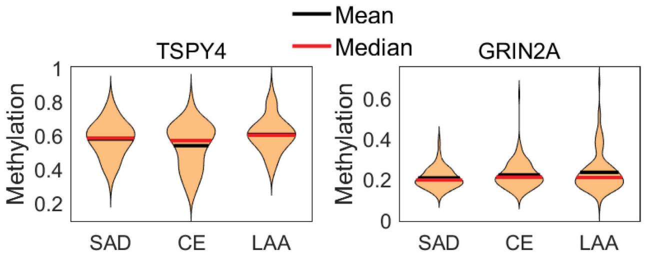

FIGURE 3. Cases of differential methylation distributions in Small-artery disease (SAD), Cardio embolic (CE), and Large-artery atherosclerosis (LAA).

LHX9, encoding a member of the LIM homeobox gene family, acts as a development-associated transcription factor [47], [48]. The methylation status of LHX9 contributes to the formation of the testicular cord [49]. Therefore, speculating that the abnormal expression/methylation of LHX9 may lead to the endogenous defect of the testicular cord and results in hemorrhagic shock in some cases is reasonable.

As shown in FIGURE 3 and FIGURE S1, the above discussed genes have different methylation distributions in different patient groups. In addition, we also validated the differential expression of these genes between control samples and cardioembolic stroke samples, on an independent dataset of GEO GSE58294 (TABLE S4). Actually, we found FAM153C, GRIN2A, and LHX9 have significantly differential expressions (P value < 0.05), which might be caused by their differential methylation reported here.

B. FUNCTIONAL ENRICHMENT ANALYSIS OF OPTIMAL GENES

All the optimal screened genes directly and indirectly participate in stroke-associated biological processes. For the accurate identification and analysis of these potential pathological processes, gene ontology and KEGG enrichment analyses were performed on the selected optimal genes by the R function phyper, whose results are supplied in TABLE 2 and TABLE S3.

Two specific GO terms (GO: 0009653 anatomical structure morphogenesis, and GO: 0048856 anatomical structure development biological process) are quite significant for stroke symptoms, indicating the specific role of unique anatomical structure in such processes. For instance, various specific anatomical structures, such as the testicular cord, participate in the pathogenesis of stroke.

Ion binding (described by GO: 0043167) has also been screened out as one of the enriched items. In fact, various publications have systematically confirmed that ion binding processes, such as gating in TRPV1 channels [50] and potassium ion for Na(+)/K(+)-ATPase regulation [51], may directly participate in various complicated pathological processes related to stroke. Another GO-term-like membrane-bounded organelle has been found to have core biological processes, reflecting the complicated biological mechanisms for the onset of stroke.

We identified a few functional KEGG pathways that are significantly enriched. Among them, hsa04510 (focal adhesion) is the only one that is quite significant for the pathogenesis of stroke. A specific regulator of a focal adhesion, named as carcinoembryonic antigen-related cell adhesion molecule 1, has been identified as a pathogenic factor for stroke [52]. This finding indicates the detailed connection of focal adhesion and our objective disease symptom.

V. CONCLUSION

Based on our newly presented computational method, all the identified genes have been validated to have abnormal methylation status during the pathogenesis of stroke and to be associated with biological functions involved in stroke onset. Further functional enrichment analysis confirmed and summarized the novel specific pathogenic roles of ion binding and focal adhesion for stroke regulated at the methylation level.

ACKNOWLEDGMENT

Supporting information: TABLE S1: Top 4730 features with their importance scores calculated by the MCFS; TABLE S2: Ten-fold cross-validation performance of IFS with SVM and RF respectively; TABLE S3: Functional enrichment analysis results; FIGURE S1: Cases of differential methylation distributions in different stroke subtypes; TABLE S4 Expression profiles of several genes in an independent data.

REFERENCES

[1] G. C. Fonarow, X. Zhao, E. E. Smith, J. L. Saver, M. J. Reeves, D. L. Bhatt, Y. Xian, A. F. Hernandez, E. D. Peterson, and L. H. Schwamm, "Door-to-needle times for tissue plasminogen activator administration and clinical outcomes in acute ischemic stroke before and after a quality improvement initiative," *JAMA*, vol. 311, no. 16, pp. 1632–1640, Apr. 2014.

[2] R. G. Nogueira *et al.*, "Predictors and clinical relevance of hemorrhagic transformation after endovascular therapy for anterior circulation large vessel occlusion strokes: A multicenter retrospective analysis of 1122 patients," *J. NeuroInterventional Surg.*, vol. 7, no. 1, pp. 16–21, Jan. 2015.

[3] N. A. M. M. Maaijwee, L. C. A. Rutten-Jacobs, P. Schaapsmeerders, E. J. van Dijk, and F.-E. de Leeuw, "Ischaemic stroke in young adults: Risk factors and long-term consequences," *Nature Rev. Neurol.*, vol. 10, no. 6, pp. 315–325, Jun. 2014.

[4] S. Chen, L. Zeng, and Z. Hu, "Progressing haemorrhagic stroke: Categories, causes, mechanisms and managements," *J. Neurol.*, vol. 261, no. 11, pp. 2061–2078, Nov. 2014.

[5] A. Siniscalchi, A. Bonci, N. Mercuri, A. Siena, G. Sarro, G. Malferrari, M. Diana, and L. Gallelli, "Cocaine dependence and stroke: Pathogenesis and management," *Current Neurovascular Res.*, vol. 12, no. 2, pp. 163–172, Mar. 2015.

[6] V. L. Feigin, M. H. Forouzanfar, R. Krishnamurthi, G. A. Mensah, M. Connor, A. B. Derrick, A. E. Moran, R. L. Sacco, L. Anderson, T. Truelsen, M. O'Donnell, N. Venketasubramanian, S. Barker-Collo, C. M. M. Lawes, W. Wang, Y. Shinohara, E. Witt, M. Ezzati, M. Naghavi, and C. Murray, "Global and regional burden of stroke during 1990–2010: Findings from the global burden of disease study 2010," *LANCET*, vol. 383, no. 9913, pp. 245–254, Jan. 2014.

[7] K. S. Yew and E. M. Cheng, "Diagnosis of acute stroke," *Amer. Family Physician*, vol. 91, no. 8, pp. 528–536, Apr. 2015.

[8] T. F. Hasan, A. A. Rabinstein, E. H. Middlebrooks, N. Haranhalli, S. L. Silliman, J. F. Meschia, and R. G. Tawk, "Diagnosis and management of acute ischemic stroke," *Mayo Clinic Proc.*, vol. 93, no. 4, pp. 523–538, Apr. 2018.

[9] A. L. Berkowitz, M. K. Mittal, H. C. McLane, G. C. Shen, R. Muralidharan, J. L. Lyons, R. T. Shinohara, A. Shuaib, and F. J. Mateen, "Worldwide reported use of IV tissue plasminogen activator for acute ischemic stroke," *Int. J. Stroke*, vol. 9, no. 3, pp. 349–355, Apr. 2014.

[10] Y. Xia, Y. Ju, J. Chen, and C. You, "Hemorrhagic stroke and cerebral paragonimiasis," *Stroke*, vol. 45, no. 11, pp. 3420–3422, Nov. 2014.

[11] Z. Zhang, G. Xu, Y. Wei, W. Zhu, and X. Liu, "Nut consumption and risk of stroke," *Eur. J. Epidemiol.*, vol. 30, no. 3, pp. 189–196, Mar. 2015.

[12] R. Renna, F. Pilato, P. Profice, G. Della Marca, A. Broccolini, R. Morosetti, G. Frisullo, E. Rossi, V. de Stefano, and V. Di Lazzaro, "Risk factor and etiology analysis of ischemic stroke in young adult patients," *J. Stroke Cerebrovascular Diseases*, vol. 23, no. 3, pp. e221–e227, Mar. 2014.

[13] J. M. Abraham and S. J. Connolly, "Atrial fibrillation in heart failure: Stroke risk stratification and anticoagulation," *Heart Failure Rev.*, vol. 19, no. 3, pp. 305–313, May 2014.

[14] H. S. Markus, "Stroke genetics," *Hum. Mol. Genet.*, vol. 20, no. 2, pp. R124–R131, Oct. 2011.

[15] H. S. Markus, "Unravelling the genetics of ischaemic stroke," *PLoS Med.*, vol. 7, no. 3, 2010, Art. no. e1000225.

[16] Z. Szolnoki, F. Somogyvári, A. Kondacs, M. Szabó, and L. Fodor, "Evaluation of the interactions of common genetic mutations in stroke subtypes," *J. Neurol.*, vol. 249, no. 10, pp. 1391–1397, Oct. 2002.

[17] D. B. Gould, F. C. Phalan, S. E. van Mil, J. P. Sundberg, K. Vahedi, P. Massin, M. G. Bousser, P. Heutink, J. H. Miner, E. Tournier-Lasserve, and S. W. M. John, "Role of COL4A1 in small-vessel disease and hemorrhagic stroke," *New England J. Med.*, vol. 354, no. 14, pp. 1489–1496, Apr. 2006.

[18] C. Soriano-Tárraga, J. Jiménez-Conde, E. Giralt-Steinhauer, M. Mola-Caminal, R. M. Vivanco-Hidalgo, A. Ois, A. Rodríguez-Campello, E. Cuadrado-Godia, S. Sayols-Baixeras, R. Elosua, and J. Roquer, "Epigenome-wide association study identifiesTXNIPgene associated with type 2 diabetes mellitus and sustained hyperglycemia," *Hum. Mol. Genet.*, vol. 25, no. 3, pp. 609–619, Feb. 2016.

[19] C. Soriano-Tárraga, E. Giralt-Steinhauer, M. Mola-Caminal, A. Ois, A. Rodríguez-Campello, E. Cuadrado-Godia, I. Fernández-Cadenas, N. Cullell, J. Roquer, and J. Jiménez-Conde, ''Biological age is a predictor of mortality in ischemic stroke,'' *Sci. Rep.*, vol. 8, no. 1, p. 4148, Dec. 2018.

[20] L. Cai, T. Huang, J. Su, X. Zhang, W. Chen, F. Zhang, L. He, and K.-C. Chou, ''Implications of newly identified brain eQTL genes and their interactors in schizophrenia,'' *Mol. Therapy-Nucleic Acids*, vol. 12, pp. 433–442, Sep. 2018.

[21] S. Zhang, X. Pan, T. Zeng, W. Guo, Z. Gan, Y.-H. Zhang, L. Chen, Y. Zhang, T. Huang, and Y.-D. Cai, ''Copy number variation pattern for discriminating MACROD2 states of colorectal cancer subtypes,'' *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 407, Dec. 2019.

[22] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, ''Predicting diabetes mellitus with machine learning techniques,'' *Frontiers Genet.*, vol. 9, p. 515, Nov. 2018.

[23] L. Chen, X. Pan, Y. H. Zhang, M. Liu, T. Huang, and Y. D. Cai, ''Classification of widely and rarely expressed genes with recurrent neural network,'' *Comput. Struct. Biotechnol. J.*, vol. 17, pp. 49–60, Dec. 2018, doi: 10.1016/j.csbj.2018.12.002.

[24] L. Chen, Y.-H. Zhang, X. Pan, M. Liu, S. Wang, T. Huang, and Y.-D. Cai, ''Tissue expression difference between mRNAs and lncRNAs,'' *Int. J. Mol. Sci.*, vol. 19, no. 11, p. 3416, 2018.

[25] Y.-D. Cai, S. Zhang, Y.-H. Zhang, X. Pan, K. Feng, L. Chen, T. Huang, and X. Kong, ''Identification of the gene expression rules that define the subtypes in glioma,'' *J. Clin. Med.*, vol. 7, no. 10, p. 350, 2018.

[26] X. Pan, T. Zeng, F. Yuan, Y.-H. Zhang, L. Chen, L. Zhu, S. Wan, T. Huang, and Y.-D. Cai, ''Screening of methylation signature and gene functions associated with the subtypes of isocitrate dehydrogenase-mutation gliomas,'' *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 339, Nov. 2019.

[27] L. Chen, T. Zeng, X. Pan, Y.-H. Zhang, T. Huang, and Y.-D. Cai, ''Identifying methylation pattern and genes associated with breast cancer subtypes,'' *Int. J. Mol. Sci.*, vol. 20, no. 17, p. 4269, 2019.

[28] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, ''Monte Carlo feature selection for supervised classification,'' *Bioinformatics*, vol. 24, no. 1, pp. 110–117, Jan. 2008.

[29] L. Chen, Y.-H. Zhang, G. Lu, T. Huang, and Y.-D. Cai, ''Analysis of cancer-related LncRNAs using gene ontology and KEGG pathways,'' *Artif. Intell. Med.*, vol. 76, pp. 27–36, Feb. 2017.

[30] L. Chen, X. Pan, T. Zeng, Y.-H. Zhang, Y. Zhang, T. Huang, and Y.-D. Cai, ''Immunosignature screening for multiple cancer subtypes based on expression rule,'' *Frontiers Bioeng. Biotechnol.*, vol. 7, p. 370, Nov. 2019.

[31] C. Cortes and V. Vapnik, ''Support-vector networks,'' *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[32] T. Zeng and J. Liu, ''Mixture classification model based on clinical markers for breast cancer prognosis,'' *Artif. Intell. Med.*, vol. 48, nos. 2–3, pp. 129–137, Feb. 2010.

[33] B. W. Matthews, ''Comparison of the predicted and observed secondary structure of t4 phage lysozyme,'' *Biochimica Et Biophysica Acta (BBA)-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[34] J. Gorodkin, ''Comparing two K-category assignments by a K-category correlation coefficient,'' *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004.

[35] X. Yu, T. Zeng, and G. Li, ''Integrative enrichment analysis: A new computational method to detect dysregulated pathways in heterogeneous samples,'' *BMC Genomics*, vol. 16, no. 1, p. 918, Dec. 2015.

[36] T. A. Jacot, I. Zalenskaya, C. Mauck, D. F. Archer, and G. F. Doncel, ''TSPY4 is a novel sperm-specific biomarker of semen exposure in human cervicovaginal fluids; potential use in HIV prevention and contraception studies,'' *Contraception*, vol. 88, no. 3, pp. 387–395, Sep. 2013.

[37] Y. Wang, Q. Yu, A. H. Cho, G. Rondeau, J. Welsh, E. Adamson, D. Mercola, and M. McClelland, ''Survey of differentially methylated promoters in prostate cancer cell lines,'' *Neoplasia*, vol. 7, no. 8, pp. 748–768, Aug. 2005.

[38] T. Kido and Y.-F.-C. Lau, ''Identification of a TSPY co-expression network associated with DNA hypomethylation and tumor gene expression in somatic cancers,'' *J. Genet. Genomics*, vol. 43, no. 10, pp. 577–585, Oct. 2016.

[39] V. K. Dasari, D. Deng, G. Perinchery, C.-C. Yeh, and R. Dahiya, ''DNA methylation regulates the expression of Y chromosome specific genes in prostate canCER,'' *J. Urol.*, vol. 167, pp. 335–338, Jan. 2002.

[40] Y. Tian, B. Stamova, G. C. Jickling, H. Xu, D. Liu, B. P. Ander, C. Bushnell, X. Zhan, R. J. Turner, R. R. Davis, P. Verro, W. C. Pevec, N. Hedayati, D. L. Dawson, J. Khoury, E. C. Jauch, A. Pancioli, J. P. Broderick, F. R. Sharp, ''Y chromosome gene expression in the blood of male patients with ischemic stroke compared with male controls,'' *Gender Med.*, vol. 9, no. 2, pp. 68–75.e3, Apr. 2012.

[41] P. S. Suresh and R. Medhamurthy, ''Luteinizing hormone regulates inhibin-α subunit expression through multiple signaling pathways involving steroidogenic factor-1 and beta-catenin in the macaque corpus luteum,'' *Growth Factors*, vol. 30, no. 3, pp. 192–206, Jun. 2012.

[42] F. L. Jayes, K. A. Burns, K. F. Rodriguez, G. E. Kissling, and K. S. Korach, ''The naturally occurring luteinizing hormone surge is diminished in mice lacking estrogen receptor Beta in the ovary,'' *Biol. Reproduction*, vol. 90, no. 2, p. 24, Feb. 2014.

[43] L. L. Jeppesen, H. S. Jørgensen, H. Nakayama, H. O. Raaschou, T. S. Olsen, and K. Winther, ''Decreased serum testosterone in men with acute ischemic stroke,'' *Arteriosclerosis, Thrombosis, Vascular Biol.*, vol. 16, no. 6, pp. 749–754, Jun. 1996.

[44] T. Seredenina, N. Demaurex, and K.-H. Krause, ''Voltage-gated proton channels as novel drug targets: From NADPH oxidase regulation to sperm biology,'' *Antioxidants Redox Signaling*, vol. 23, no. 5, pp. 490–513, Aug. 2015.

[45] J. Conroy, P. A. McGettigan, D. McCreary, N. Shah, K. Collins, B. Parry-Fielder, M. Moran, D. Hanrahan, T. W. Deonna, C. M. Korff, D. Webb, S. Ennis, S. A. Lynch, and M. D. King, ''Towards the identification of a genetic basis for Landau-Kleffner syndrome,'' *Epilepsia*, vol. 55, no. 6, pp. 858–865, Jun. 2014.

[46] Y. M. Lee, H. C. Kang, J. S. Lee, S. H. Kim, E. Y. Kim, S. K. Lee, A. Slama, and H. D. Kim, ''Mitochondrial respiratory chain defects: Underlying etiology in various epileptic conditions,'' *Epilepsia*, vol. 49, no. 4, pp. 685–690, Apr. 2008.

[47] N. Radó-Trilla, K. Arató, C. Pegueroles, A. Raya, S. de la Luna, and M. M. Albà, ''Key role of amino acid repeat expansions in the functional diversification of duplicated transcription factors,'' *Mol. Biol. Evol.*, vol. 32, no. 9, pp. 2263–2272, Sep. 2015.

[48] V. Failli, M. Rogard, M.-G. Mattei, P. Vernier, and S. Rétaux, ''Lhx9 and Lhx9α LIM-homeodomain factors: Genomic structure, expression patterns, chromosomal localization, and phylogenetic analysis,'' *Genomics*, vol. 64, no. 3, pp. 307–317, Mar. 2000.

[49] T. Mizukami, Y. Kanai, M. Fujisawa, M. Kanai-Azuma, M. Kurohmaru, and Y. Hayashi, ''Five azacytidine, a DNA methyltransferase inhibitor, specifically inhibits testicular cord formation and sertoli cell differentiation *in vitro*,'' *Mol. Reproduction Develop.*, vol. 75, no. 6, pp. 1002–1010, Jun. 2008.

[50] A. Jara-Oseguera, C. Bae, and K. J. Swartz, ''An external sodium ion binding site controls allosteric gating in TRPV1 channels,'' *Elife*, vol. 5, Feb. 2016, Art. no. e13356.

[51] J. P. Castillo, H. Rui, D. Basilio, A. Das, B. Roux, R. Latorre, F. Bezanilla, and M. Holmgren, ''Mechanism of potassium ion uptake by the Na+/K+−ATPase,'' *Nature Commun.*, vol. 6, no. 1, p. 7622, Nov. 2015.

[52] P. Ludewig, J. Sedlacik, M. Gelderblom, C. Bernreuther, Y. Korkusuz, C. Wagener, C. Gerloff, J. Fiehler, T. Magnus, and A. K. Horst, ''Carcinoembryonic antigen–related cell adhesion molecule 1 inhibits MMP-9–mediated blood–brain–barrier breakdown in a mouse model for ischemic stroke,'' *Circulat. Res.*, vol. 113, no. 8, pp. 1013–1022, Sep. 2013.

XIANGTIAN YU received the B.S., M.S., and Ph.D. degrees from Shandong University, Jinan, China, in 2008, 2011, and 2015, respectively. She has been a Postdoctoral Researcher with the Shanghai Institutes of Biochemistry and Cell Biology, Chinese Academy of Sciences. She is currently an Assistant Professor with Shanghai Jiao Tong University Affiliated Sixth People's Hospital. Her research fields mainly include bioinformatics and computational biology.

ZIJUN GAN was born in Zhejiang, China, in 1993. She received the B.S. degree in biopharmaceutical from Zhejiang A & F University, in 2016, and the master's degree in genetics from the Shanghai Institute for Biological Sciences, in 2019. Her research interests include bioinformatics and genetics.

YAN XU was born in Jiangsu, China, in 1995. She received the bachelor's degree in agriculture from the College of Plant Protection, Hunan Agricultural University, in 2017. She is currently pursuing the master's degree with the School of Life Sciences, Shanghai University. Her research interest is bioinformatics.

SIBAO WAN received the B.S. degree from Shandong Agricultural University, in 2002, and the Ph.D. degree in major food science from China Agricultural University, in 2007. He was a Visiting Scholar with Cornell University, from 2013 to 2014. He is currently an Associate Professor with Shanghai University. His research interests include molecular biology and bioinformatics.

MIN LI was born in Yulin, China, in 1997. She received the B.S. degree in life science and technology from Yan'an University, in 2018. After that, she joined the School of Life Science, Shanghai University, for her master education. Her research interests include bioinformatics and cancers.

SHIJIAN DING was born in Anhui, China, in 1996. He received the B.S. degree in biotechnology from Anqing Normal University, in 2019. He is currently pursuing the master's degree in biology with Shanghai University. His research interests include bioinformatics and machine learning.

TAO ZENG received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 2003, 2006, and 2010, respectively. Since 2013, he has been an Associate Professor with the Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He is currently with the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China. His research interests include precision medicine, bioinformatics, network biology, computational biology, machine learning, and graph theory.

• • •