MEDICAL SCIENCE MONITOR

LABORATORY TECHNIQUES

e-ISSN 1643-3750
© Med Sci Monit, 2017; 23: 994-1000
DOI: 10.12659/MSM.899690

Received: 2016.05.23
Accepted: 2016.07.11
Published: 2017.02.24

Personalized Analysis by Validation of Monte Carlo for Application of Pathways in Cardioembolic Stroke

Authors' Contribution:
Study Design A
Data Collection B
Statistical Analysis C
Data Interpretation D
Manuscript Preparation E
Literature Search F
Funds Collection G

AE 1 Zhangmin Xing*
BCE 1 Bin Luan*
BCF 2 Ruiying Zhao
DE 1 Zhanbiao Li
A 1 Guojian Sun

1 Department of Rehabilitation Medicine, The People's Hospital of Liaocheng, Liaocheng, Shandong, P.R. China
2 The Blood Center of Liaocheng, Liaocheng, Shandong, P.R. China

* Contributed equally to this work

Corresponding Author: Guojian Sun, e-mail: guojiansunmed@126.com
Source of support: Departmental sources

Background: Cardioembolic stroke (CES), which causes 20% cause of all ischemic strokes, is associated with high mortality. Previous studies suggest that pathways play a critical role in the identification and pathogenesis of diseases. We aimed to develop an integrated approach that is able to construct individual networks of pathway cross-talk to quantify differences between patients with CES and controls.

Material/Methods: One biological data set E-GEOD-58294 was used, including 23 normal controls and 59 CES samples. We used individualized pathway aberrance score (iPAS) to assess pathway statistics of 589 Ingenuity Pathways Analysis (IPA) pathways. Random Forest (RF) classification was implemented to calculate the AUC of every network. These procedures were tested by Monte Carlo Cross-Validation for 50 bootstraps.

Results: A total of 28 networks with AUC >0.9 were found between CES and controls. Among them, 3 networks with AUC=1.0 had the best performance for classification in 50 bootstraps. The 3 pathway networks were able to significantly identify CES versus controls, which showed as biomarkers in the regulation and development of CES.

Conclusions: This novel approach could identify 3 networks able to accurately classify CES and normal samples in individuals. This integrated application needs to be validated in other diseases.

MeSH Keywords: Critical Pathways • Individualized Medicine • Monte Carlo Method • Stroke

Full-text PDF: http://www.medscimonit.com/abstract/index/idArt/899690

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

Background

Cardioembolic stroke (CES), which causes 20% of all ischemic strokes each year, leads to severe neurological deficits [1,2]. CES is associated with high mortality and is a common cause of its atrial fibrillation (AF), which has an increasing incidence with age [3–5]. Panagiota et al. [6] proposed that AF is an important and treatable cause of recurrent stroke and needs to be ruled-out by thorough evaluation before the diagnosis of cryptogenic stroke is assigned. CES is largely preventable through control of major primary cardioembolic risk factors, such as hyperlipidemia and high blood pressure [7]. Giralt et al. [8] offered evidence of significant genetic involvement in ischemic stroke.

In recent years, gene expression profiling of human disease tissues has provided insights into molecular mechanisms and eventually led to the identification of novel therapeutic targets [9]. Currently available high-throughput microarray experiments were developed to analyze genetic expression patterns with differentially expressed genes (DEG) and dysregulated pathways. Canonical reports claimed that gene expression patterns can identify biomarkers of ischemic stroke, which highlighted the relevance of the innate immune system through DEG [10] and signaling pathways [11–13]. However, most methods did not consider regulatory cross-talk among pathways, and treated pathways as independent mechanisms.

Although it is intuitive that interacting pathways could influence each other, the presence of this frame and available technique have not been completely studied yet. Antonio et al. [14] developed an integrated approach to identify functional miRNAs regulating pathway cross-talk in breast cancer with pairs of pathways. Differential protein-protein interaction networks were constructed in CES with Akaike information criterion (AIC) method [7].

To the best of our knowledge, there are few studies that constructed pathway networks correctly to discriminate controls versus CES. In this work we develop an integrated approach that is able to construct individual networks comprising pathways cross-talk to quantify differences between CES and controls. We used the individualized pathway aberrance score (iPAS) to assess pathway statistics of every Ingenuity Pathways Analysis (IPA) pathway [15]. Random Forest (RF) classification was implemented to calculate the AUC of every network. These procedures were tested by Monte Carlo Cross-Validation (MCCV) for 50 bootstraps. Then we obtained the best network as an individual differential network. Our results may be useful in more integratively and accurately distinguishing CES from normal samples. The novel approach may be the basis of individual medical treatment in CES, serving as therapy targeting markers.

Material and Methods

Step 1: Datasets

One biological dataset, E-GEOD-58294, was derived from the Gene Expression Omnibus (GEO) database (*http://www.ncbi.nlm.nih.gov/geo/*) [16]. There were 23 normal controls and 59 CES samples in total. The platform was A-AFFY-44 – Affymetrix GeneChip Human Genome U133 Plus 2.0, which was used to read the gene chip [17]%h0. The Linear Models for Microarray Data (LIMMA) was then used to preprocess data. After quantile data normalization performed by robust multi-array average (RMA) [18], 20 544 genes were obtained.

Step 2: Pathway enrichment analysis

In order to identify a group of pathways significantly enriched in CES with respect to controls, we collected 589 biological pathways including 5169 genes from the IPA tool (*http://www.ingenuity.com/*). After genes of expression profile were enriched in IPA pathways, we focused on 4929 genes. Fisher Exact test was performed between 4929 genes and genes of every IPA pathway. Then we obtained pathways enriched with $P<0.01$. Raw P-values were adjusted by false-discovery rate (FDR) procedure for multiple testing corrections [19].

Step 3: Pathway-level statistics

A total of 23 accumulated normal samples (ANS) were used to identify IPA pathways as reference. Individual normal sample gene expression was standardized with the mean and standard deviation (SD). For genes of every CES sample,

$$\text{proj}_d q_k = \left(\frac{1}{n}\sum_{j=1}^{n} qkj,\ldots,\frac{1}{n}\sum_{j=1}^{n} qkj\right) \quad (1)$$

as quantile normalization was performed [20].

Average Z equation was recently proved to be a biologically valid modification of pathway analysis methods for iPAS [15]. $Z=(z_1, z_2,\ldots, z_n)$ represents the expression state of a pathway where z_i denotes the standardized expression value of *i*-th gene and the number of genes existing in the pathway is *n*. Gene statistics of each gene from every CES sample:

$$z_i = \frac{g_{Ti} - \text{mean}(g_{nANS})}{\text{stdev}(g_{nANS})} \quad (2)$$

Each IPA pathway statistics:

$$iPAS = \frac{\sum_i^n z_i}{n} \quad (3)$$

z_i represents the standardized gene level statistics of *1-i* gene and the number of genes existing in the pathway is *n*. Z values of every pathway in CES samples were gathered after

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

significance testing. Differentially expressed pathways were selected with Z<0.05.

Step 4: Pathway pairs

The discriminating score (DS) was computed to quantify pathway cross-talk in each sample for the pair of pathways x and y. DS was defined as

$$DS=\frac{|(M_x-M_y)|}{S_x+S_y} \quad (4)$$

where M_x and S_x represent mean and standard deviation of expression levels of genes in a pathway x and M_y and S_y in a pathway y [14]. DS score indicates the relationships between pairs of pathway, with a larger value indicating relatively higher difference of activity between pathways.

DS of normal samples was standardized using the mean and SD as reference. Z values of every pathway pair in CES samples were gathered after significance testing. Differentially expressed pathway pairs were selected with Z<0.25.

Step 5: Construction of network

Z values of differentially expressed pathways and pathway pairs were used to construct individual networks with Cytoscape version 3.2.0. The main network was constructed by selecting the number of edges >5.

Step 6: Random Forest (RF) classification

Random Forest (RF) classification was implemented using the R-package. Parameters were adopted with mtry=$\sqrt{2}$ and ntree=500. Classification was applied on DS of pathway pairs in the main network. The AUC of the main network was calculated by 10-fold cross-validation method.

Step 7: Selection of the best network

We developed MCCV to circulate step 3–6 of the proposed methodology. It randomly selected expression data in proportion 6:4 to form the training and testing set [14]. Then the process was repeated in 50 bootstraps, randomly generating new training and test partitions each time. Each bootstrap achieved an individual network, main network, and their AUC values. The number of main networks appearing in the 50 bootstraps was counted by ranking all networks with their AUC values.

Results

In the present study we developed an integrated approach that was sufficient to construct individual networks comprising pathways cross-talk to quantify differences between CES and controls. We used iPAS to evaluate pathway statistics of each IPA pathway [15]. RF classification was implemented to calculate AUC of every network, which was tested by MCCV for 50 bootstraps. Then we obtained the best network as an individual differential network.

Figure 1 shows the results for each bootstrap of MCCV. We obtained a heatmap in which pink squares indicate pathway pairs for classification in the training dataset for that bootstrap (the frequency >6). There were 4 pairs of pathways in 46 bootstraps: Cholesterol Biosynthesis I and Cholesterol Biosynthesis II, Cholesterol Biosynthesis I and Cholesterol Biosynthesis III, Cholesterol Biosynthesis II and Cholesterol Biosynthesis III, Uracil Degradation II and Thymine Degradation.

Individual networks were ordered with respect to their AUC and 28 networks with AUC >0.9 were found between CES and controls. Among them, 3 networks with AUC=1.0 had the best performance for classification of CES and normal samples for all 50 bootstraps. As shown in Figure 2, the best individual networks were in 4, 10, and 23 bootstraps. Therefore, the 3 pathway networks were able to significantly identify CES versus controls, which showed as biomarkers in the regulation and development of CES. Then we found there were 22 pairs of pathways that commonly appeared in 3 networks (Table 1), which revealed that the pathway pairs were important in regulating CES.

Discussions

Given the substantial difference in the activities of main networks between CES and controls, we examined its effectiveness in classifying CES and normal samples based on their profiles.

In the best 3 networks, we focused on pathways that had multi-cross-talk with others. The MSP-RON Signaling Pathway had the most cross-talk, which played an important interaction role in the best networks. A previous study has reported that MSP-RON Signaling is important for the invasive growth of many types of cancers and appeared to have potential as a therapeutic target [21].

Pathway analysis has become the first choice for extracting and explaining the underlying pathology for high-throughput molecular measurements [22]. Personalized identification of altered pathway pairs is important for understanding disease mechanisms and for the future application of custom therapeutic decisions. Existing pathway analysis methods are not suitable for identifying the pathway aberrance that may occur in an individual sample [15]. Therefore, we employed the iPAS to analyze the personalized identification of networks, taking advantage of a vast number of normal samples.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

Figure 1. Heatmap of pathway pairs in each bootstrap. Bootstraps were clustered with the abscissa and pairs of pathways were clustered with the ordinate.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

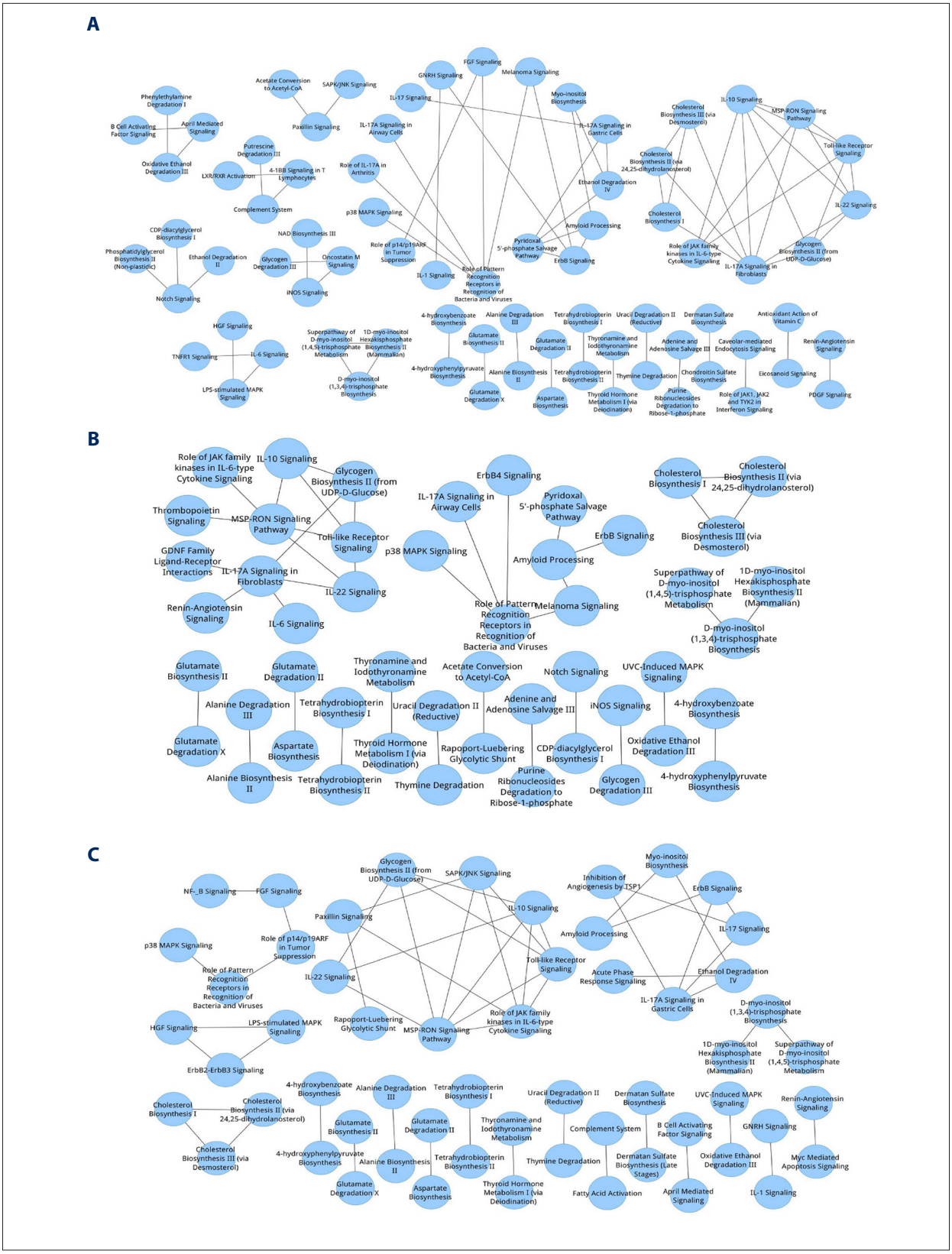

Figure 2. The best individual differential networks repeated 50 bootstraps. (**A**) The individual network in 10 bootstraps. (**B**) The individual network in 10 bootstraps. (C) The individual network in 23 bootstraps.

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

Table 1. Common pairs of pathways in best three networks.

No.	Pairs of pathways	AUC of 10 bootstrap
1	Toll-like receptor signaling Glycogen biosynthesis II (from UDP-D-glucose)	0.286
2	IL-10 signaling Glycogen biosynthesis II (from UDP-D-glucose)	0.281
3	D-myo-inositol hexakisphosphate biosynthesis II (Mammalian) D-myo-inositol (134)-trisphosphate biosynthesisi	0.262
4	IL-10 signaling Toll-like receptor signaling	0.261
5	IL-10 signaling MSP-RON signaling pathway	0.260
6	MSP-RON signaling pathway IL-22 signaling	0.259
7	MSP-RON signaling pathway Role of JAK family kinases in IL-6-type cytokine signaling	0.256
8	p38 MAPK signaling Role of pattern recognition receptors in recognition of bacteria and viruses	0.255
9	Superpathway of D-myo-inositol (145)-trisphosphate metabolism D-myo-inositol (134)-trisphosphate biosynthesisi	0.253
10	MSP-RON signaling pathway Toll-like receptor signaling	0.253
11	Adenine and adenosine salvage III Purine ribonucleosides degradation to ribose-1-phosphate	0.252
12	ErbB signaling Amyloid processing	0.251
13	Cholesterol biosynthesis I Cholesterol biosynthesis II (via 2425 dihydrolanosterol)	0.243
14	Cholesterol biosynthesis I Cholesterol biosynthesis III (via desmosterol)	0.243
15	Cholesterol biosynthesis II (via 2425-dihydrolanosterol) Cholesterol biosynthesis III (via desmosterol)	0.243
16	Uracil degradation II (reductive) Thymine degradation	0.243
17	Thyronamine and iodothyronamine metabolism Thyroid hormone metabolism I (via deiodination)	0.243
18	Tetrahydrobiopterin biosynthesis I Tetrahydrobiopterin biosynthesis II	0.243
19	Glutamate degradation II Aspartate biosynthesis	0.243
20	Alanine degradation III Alanine biosynthesis II	0.243
21	Glutamate biosynthesis II Glutamate degradation X	0.243
22	4-hydroxybenzoate biosynthesis 4-hydroxyphenylpyruvate biosynthesis	0.243

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]

A key innovation of the method is iPAS using ANS in CES. Ahn et al. [15] proved that the Average Z equation can efficiently reveal noticeable aberrance in expression profiles and clinical significance, which sufficed to confirm the best averaged validation rate and distinguish a known survival-relevant pathway statistically. Furthermore, ANS data is expected to be available in more fields of medicine along with rapid advances in high-throughput databases. DS obtained lightly more improvement than the Euclidean distance as a metric to quantify pathway cross-talk [14].

In recent years, different validation technologies have been generally used to evaluate performance of pathways and networks in medical regression analysis [14,23]. The MCCV pays attention to a notable part of the sample at a time during network building and validation with multi-repeats. Compared with conventional validation tests for capturing the best predictor variables, MCCV showed superior performance, resulting from a form of cross-validation based on vast combinations of data sets [24]. Interestingly, MCCV has not been utilized in individual networks comprising pathways cross-talk in CES patients. In this study we developed an integrated approach to quantify differences between CES and controls with the MCCV test, which suggests that MCCV worked better, based on strong predictive ability. Screened networks were efficient in distinguishing differences among individual CES samples, and can provide broader carcinogenic insight in personalized medicine [25]. The final purpose of our approach was to detect the best network able to discriminate CES versus controls. We found that the 3 best networks were similar and had 22 common pairs of pathways. We tended to select network 10 to differentiate CES disease from normal samples, with the fewest pairs of pathways (Figure 2B).

Conclusions

Our novel approach identified 3 networks able to accurately classify CES and normal samples in individuals. We propose the integrated method should be further validated in more diseases.

References:

1. Bogousslavsky J, Cachin C, Regli F et al: Cardiac sources of embolism and cerebral infarction – clinical consequences and vascular concomitants: The Lausanne Stroke Registry. Neurology, 1991; 41: 855–59
2. Freeman WD, Aguilar MI: Prevention of cardioembolic stroke. Neurotherapeutics, 2011; 8: 488–502
3. Hajat C, Heuschmann PU, Coshall C et al: Incidence of aetiological subtypes of stroke in a multi-ethnic population based study: The South London Stroke Register. J Neurol Neurosurg Psychiatry, 2011; 82: 527–33
4. Palm F, Urbanek C, Wolf J et al: Etiology, risk factors and sex differences in ischemic stroke in the Ludwigshafen Stroke Study, a population-based stroke registry. Cerebrovasc Dis, 2012; 33: 69–75
5. Palm F, Kraus M, Safer A et al: Management of oral anticoagulation after cardioembolic stroke and stroke survival data from a population based stroke registry (LuSSt). BMC Neurol, 2014; 14: 199
6. Christia P, Katsa I, Ocava L, Faillace R: Atrial fibrillation identified during echocardiography in a patient with recurrent cardioembolic events: A case report. Am J Case Rep, 2016; 17: 129–32
7. Wong YH, Wu CC, Lai HY et al: Identification of network-based biomarkers of cardioembolic stroke using a systems biology approach with time series data. BMC Syst Biol, 2015; 9(Suppl. 6): S4
8. Giralt D, Domingues-Montanari S, Mendioroz M et al: The gender gap in stroke: A meta-analysis. Acta Neurol Scand, 2012; 125: 83–90
9. Chen ZH, Kim HP, Ryter SW, Choi AM: Identifying targets for COPD treatment through gene expression analyses. Int J Chron Obstruct Pulmon Dis, 2008; 3: 359–70
10. Bi BL, Wang HJ, Bian H, Tian ZT: Identification of therapeutic targets of ischemic stroke with DNA microarray. Eur Rev Med Pharmacol Sci, 2015; 19: 4012–19
11. Tang Y, Xu H, Du X et al: Gene expression in blood changes rapidly in neutrophils and monocytes after ischemic stroke in humans: a microarray study. J Cereb Blood Flow Metab, 2006; 26: 1089–102
12. Grond-Ginsbach C, Hummel M, Wiest T et al: Gene expression in human peripheral blood mononuclear cells upon acute ischemic stroke. J Neurol, 2008; 255: 723–31
13. Barr TL, Conley Y, Ding J et al: Genomic biomarkers and cellular pathways of ischemic stroke by RNA gene expression profiling. Neurology, 2010; 75: 1009–14
14. Colaprico A, Cava C, Bertoli G et al: Integrative analysis with Monte Carlo cross-validation reveals miRNAs regulating pathways cross-talk in aggressive breast cancer. Biomed Res Int, 2015; 2015: 831314
15. Ahn T, Lee E, Huh N, Park T: Personalized identification of altered pathways in cancer using accumulated normal tissue data. Bioinformatics, 2014; 30: i422–29
16. Stamova B, Jickling GC, Ander BP et al: Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. PLoS One, 2014; 9: e102550
17. Gautier L, Cope L, Bolstad BM, Irizarry RA: affy – analysis of Affymetrix GeneChip data at the probe level. Bioinformatics, 2004; 20: 307–15
18. Irizarry RA, Hobbs B, Collin F et al: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 2003; 4: 249–64
19. Benjamini Y, Drai D, Elmer G et al: Controlling the false discovery rate in behavior genetics research. Behav Brain Res, 2001; 125: 279–84
20. Bolstad BM, Irizarry RA, Astrand M, Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 2003; 19: 185–93
21. Yao HP, Zhou YQ, Zhang R, Wang MH: MSP-RON signalling in cancer: Pathogenesis and therapeutic potential. Nat Rev Cancer, 2013; 13: 466–81
22. Khatri P, Sirota M, Butte AJ: Ten years of pathway analysis: Current approaches and outstanding challenges. PLoS Comput Biol, 2012; 8: e1002375
23. Muto S, Tanabe T: Reverse Monte Carlo analysis of extended energy-loss fine structure for disordered structures of tetrahedrally coordinated materials: Its applicability. J Electron Microsc (Tokyo), 2003; 52: 125–32
24. Theodorou D, Meligotsidou L, Karavoltsos S et al: Comparison of ISO-GUM and Monte Carlo methods for the evaluation of measurement uncertainty: Application to direct cadmium measurement in water by GFAAS. Talanta, 2011; 83: 1568–74
25. Slattery ML, Herrick JS, Mullany LE et al: Improved survival among colon cancer patients with increased differentially expressed pathways. BMC Med, 2015; 13: 75

Indexed in: [Current Contents/Clinical Medicine] [SCI Expanded] [ISI Alerting System] [ISI Journals Master List] [Index Medicus/MEDLINE] [EMBASE/Excerpta Medica] [Chemical Abstracts/CAS] [Index Copernicus]