Serum microRNA panel excavated by machine learning as a potential biomarker for the detection of gastric cancer

YAO HUANG[1,2*], JIE ZHU[1*], WENSHUAI LI[1], ZIQIANG ZHANG[1],
PANPAN XIONG[1], HONG WANG[2] and JUN ZHANG[1]

[1]Department of Digestive Diseases, Huashan Hospital, Fudan University;
[2]Department of Gastroenterology, Jing'an District Center Hospital of Shanghai (Huashan Hospital,
Fudan University, Jing'an Branch), Shanghai 200040, P.R. China

Received June 21, 2017; Accepted December 12, 2017

DOI: 10.3892/or.2017.6163

Abstract. Early detection of gastric cancer (GC) is crucial to improve the therapeutic effect and prolong the survival of patients. MicroRNAs (miRNAs) are a group of small non-protein-coding RNAs that function as repressors of diverse genes. We aimed to identify a microRNA panel in the serum of patients to predict GC non-invasively with high accuracy and sensitivity. Using six types of classifiers, we selected three markers (miR-21-5p, miR-22-3p and miR-29c-3p) from a published miRNA profiling study (GSE23739) which was treated as a training set. The values of the area under the receiver operating characteristic (ROC) curves (AUCs) were 0.9437, 0.9456 and 0.9563 in the three classifiers [Compound covariate classifier, Diagonal linear discriminant analysis (DLDA) classifier and Support vector machine classifier], respectively. Then the panel was validated further in another two miRNA profiles in GEO (Gene Expression Omnibus) databases (GSE26595, GSE28700) with high AUC values as well. Next, we found that the serum levels of miR-21 were significantly higher in GC patients than levels in healthy controls by quantitative reverse transcription-polymerase chain reaction (qRT-PCR) for confirmation, which was opposite to the serum levels of miR-22 and miR-29c (all P<0.0001). Finally, using bioinformatic tools, their biological mechanisms were elucidated by their predicted targets: Sp1 (miR-21) and PTEN (miR-22 and miR-29c). This miRNA panel is a non-invasive and potential biomarker for GC.

Introduction

Gastric cancer (GC) is one of the most threatening worldwide diseases. An estimated 951,600 new stomach cancer cases and 723,100 deaths occurred in 2012 (1). GC is still the second most frequent malignancy globally despite great improvement in the diagnosis and treatment of GC. In China, which belongs to an area with the highest incidence rate, a total of 679,000 new cases of GC and 498,000 deaths occurred in 2015 (2). The high mortality of GC is attributed to the low rate of early diagnosis. Thus, the majority of patients are diagnosed at an advanced stage with a poor patient prognosis. Therefore, it is critical to improve the sensitivity and specificity of diagnostic tools for the prevention and detection of GC. Although the gold standard diagnostic methods for GC, endoscopy and random biopsy endoscopy or image examination, could facilitate the early diagnosis of GC, the invasive nature, potential sampling errors and high expenditure impact their use only for patients at a high-risk of GC.

MicroRNAs (miRNAs) are small non-coding RNAs comprised of approximately 21 nucleotides in length that crucially participate in regulating the translation and degradation of mRNAs (3). miRNAs preferentially bind to complementary sites at the 3'UTR of their target mRNAs, which contribute to their pivotal regulation in a wide variety of biological processes, including cell growth, development, differentiation and apoptosis (4). In the development of GC, aberrant expression of miRNAs has been found to be correlated with the clinical features of GC, such as occurrence, development and metastasis (5).

Recent studies have aimed to evaluate the microRNAs present in serum/plasma as potential molecular biomarkers on account of their high stability and convenience in biological samples (e.g. miR-21, miR-148a and miR-124) (6,7). MicroRNA expression profiling has been used to achieve this goal in an effective way (8). An innovative screening strategy which is called machine learning was carried out in our study to select miRNAs from profiles and verify their efficacy more efficiently. Machine learning is the programming of computers

Correspondence to: Dr Hong Wang, Department of Gastroenterology, Jing'an District Center Hospital of Shanghai (Huashan Hospital, Fudan University, Jing'an Branch), 259 Xi Kang Road, Shanghai 200040, P.R. China
E-mail: wanghongjzx@aliyun.com

Dr Jun Zhang, Department of Digestive Diseases, Huashan Hospital, Fudan University, 12 Middle Wulumuqi Road, Shanghai 200040, P.R. China
E-mail: archsteed@gmail.com

[*]Contributed equally

Key words: microRNA, machine learning, biomarker, diagnosis, gastric cancer

to optimize a performance criterion using example data or past experience. A mathematical model is built by the theory of statistics, and learning is the execution of a computer program to optimize the parameters of the model using the training data with high speed and efficiency.

Herein, we excavated a profile of three combined serum miRNAs using machine learning and confirmed its excellent accuracy and reliability in the detection of GC.

Materials and methods

Ethics statements. The study was approved by the Human Research Review Committee of Huashan Hospital, Fudan University and written informed consents were obtained from all of the patients. The study conformed to The Code of Ethics of the World Medical Association (Declaration of Helsinki).

Study and microRNA selection. GEO (Gene Expression Omnibus) dataset search engine was used for the GC microRNA expression profiling studies. The following keywords: 'miRNA' OR 'microRNA' OR 'miR', 'gastric' OR 'stomach', 'profiling' OR 'microarray' were used to search for potential studies. We selected a total of three GEO profiles (GSE23739, GSE26595, GSE28700) as their sample sizes were sufficiently large. The GSE23739 profile represented 723 human and 76 human viral miRNAs in 40 normal and 40 cancerous gastric tissues. MicroRNAs identified from 60 primary GC tissues and 8 surrounding non-cancer tissues were used for microarray analysis in GSE26595. A total of 22 paired GC and normal tissues were processed in GSE28700. We eradicated any repetitive miRNAs and normalized data of each profile.

Machine learning. Initially, we treated GSE23739 as a training set and prepared to use GSE26595 and GSE28700 as validation sets. In order to analyze the data from different profiles, total data were standardized using a series of classifiers. These six classifiers (9) were performed to screen specific microRNAs to distinguish GC tissues from normal tissues: Compound covariate classifier (10), Diagonal linear discriminant analysis (DLDA) classifier (11), Bayesian CCP classifier (12,13), 1/3-Nearest Neighbor classifier (14), Nearest centroid classifier (15), and Support vector machines classifier (16). 1-Nearest Neighbor classifier and 3-Nearest Neighbor classifier are different types of this classifier to assess the stability of data. Leave one out cross validation was run to ensure stability and accuracy of the output.

Serum samples. All serum samples were collected from GC patients and healthy individuals treated at Huashan Hospital (Shanghai, China), affiliated to the Fudan University between 2014 and 2016. All of the 24 patients were clinically and pathologically diagnosed with GC. We evaluated the clinical and pathological features of the patients and these data are summarized in Table I.

Serum preparation and microRNA extraction. Venous blood was collected in EDTA anticoagulation vacuum tubes and was centrifuged at 1000 x g for 15 min at 4ºC and then the separated serum was transferred into 1.5 ml RNase-free tubes stored at -80˚C until RNA extraction. Small RNAs were extracted from 200 μl of serum using the miRcute miRNA Isolation kit (Tiangen Biotech Co., Beijing, China) according to the manufacturer's instructions.

Table I. Clinical characteristics of the GC patients vs. normal controls.

Characteristics	GC n=24	Controls n=20	P-value
Age (years), mean ± SD	57.6±14.0	53.2±12.5	P=0.2823
Sex			P=0.2929
Male	13	11	
Female	11	9	
Tumor location			
Cardia	10		
Body	9		
Antrum	3		
Other	2		
Histology			
Adenocarcinoma	24		
Tumor size (cm)			
≥5	13		
<5	11		
TNM stage			
I+II	5		
III+IV	19		
Metastatic status			
Yes	3		
No	21		

GC, gastric cancer; TNM, Tumor, Node, Metastasis.

Quantification of miRNA expression in serum by qRT-PCR. Reverse transcription (RT) reactions were performed using miRcute Plus miRNA First-Strand cDNA Synthesis kit (Tiangen Biotech Co.). Singleplex reactions were conducted in a volume of 20 μl which consisted of 10 μl 2X miRNA RT reaction buffer, 2 μl miRNA RT Enzyme Mix, and 8 μl miRNA template. Then, the RT reaction was carried out in a thermocycler under the following conditions: 42˚C for 60 min, 95˚C for 3 min, followed by a hold at 4˚C. All the operations were performed with caution to exclude RNase contamination, and the total end products were preserved at -20˚C for further analysis.

Quantitative PCR (qPCR) was performed using 7500 Fast Real-Time PCR System (Applied Biosystems, Foster City, CA, USA) with miRcute miRNA qPCR Detection kit (Tiangen Biotech Co.). All forward primers were obtained from Tiangen Biotech Co. Their catalog numbers are CD201-0092 (hsa-miR-21-5p), CD201-0404 (hsa-miR-29c-3p) and CD201-0305 (hsa-miR-22-3p). Then they were diluted to 10 μM before being adding to the PCR reaction mixture. qRT-PCR was performed in 3 duplicate reactions comprising 10 μl 2X miRcute miRNA Premix (with SYBR and ROX), forward and reverse primer,

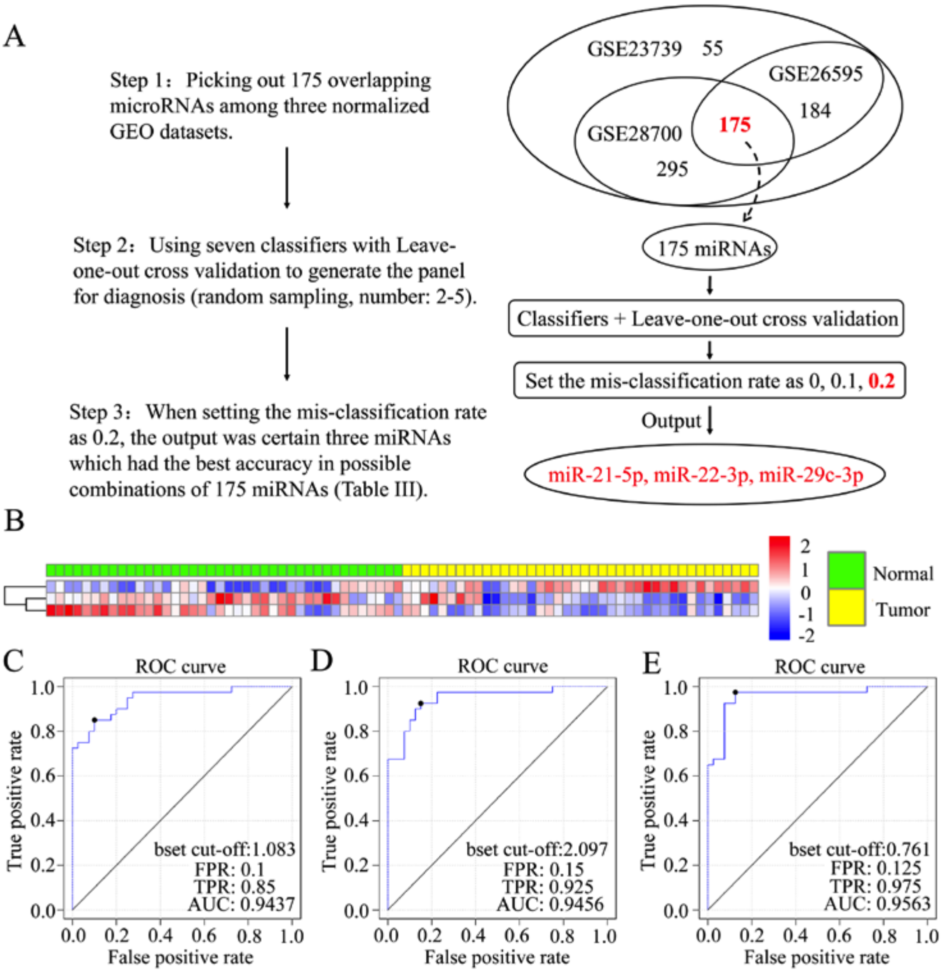

Figure 1. The selection of three miRNAs (miR-21, miR-22 and miR-29c) as markers in the training set. (A) The diagram illustrating the screening of microRNAs in GSE23739 following three steps. Step 1: Selection of 175 overlapping miRNAs (red color). Step 2: Using seven types of classifiers combined with Leave-one-out cross validation to screen a panel of miRNAs (number: 2-5). Step 3: Discovering the most highly accurate panel of three-miRNA group (red color), when setting the mis-classification rate as 0.2. (B) Heat-map diagram consisting of three miRNAs was drawn according to the miRNA profile (GSE23739). The 1-3 lines represent miR-21, miR-22 and miR-29c, respectively. The green color represents normal patients; the yellow color indicates cancer patients. The darkness of the heat map color represents the gene expression level; the color which is closer to red (the number 2) has the higher level of gene expression. (C-E) Receiver operating characteristic (ROC) curves of three linear classifiers are displayed in the following sequence: Compound covariate classifier, Diagonal linear discriminant analysis (DLDA) classifier and Support vector machine classifier.

each with 0.4 μl, 2 μl miRNA First-Strand cDNA template, and 7.2 μl RNase-free ddH_2O. Mixtures were denatured at 94˚C for 2 min and then run for 40 cycles (94˚C for 20 sec, 60˚C for 34 sec). Melting curve analysis was run after all these procedures were completed. The expression levels of miRNAs in serum were normalized to U6 for the next quantification which was calculated using the $2^{-\Delta\Delta CT}$ method.

Target gene analysis. The union of predicted target genes was searched using starBase v2.0 (http://starbase.sysu.edu.cn/index.php). The Gene Ontology and Genome Pathway were processed and produced by OmicsBean (http://www.omicsbean.cn/). Then we created biological networks employing Cytoscape v3.2 open-source software with CyTargetLinker App (17) and we treated miRTarBase (http://mirtarbase.mbc.nctu.edu.tw/) as the tool for selecting targets intersected by the results of three miRNAs in this software. Gene-disease association data were retrieved from the DisGeNET database (http://www.disgenet.org/). The term ‘gastric adenocarcinoma’ (umls: C0278701) was used to identify GC-associated genes. STRING (http://www.string-db.org/) was used to analysis the interaction between different proteins.

Table II. Weight values of three miRNAs.

Genes	Compound covariate	Diagonal linear discriminant analysis	Support vector machines
hsa-miR-21-5p	-6.6501	-0.5599	-0.469
hsa-miR-22-3p	4.8583	0.815	0.2505
hsa-miR-29c-3p	6.9725	0.9744	0.9594

Statistical analysis. The clinical characteristics among groups were compared using the χ^2 test and Fisher's exact test for qualitative data, and t-test for quantitative data. A receiver operating characteristic (ROC) curve was generated for the specificity and sensitivity value calculated by classifiers, which

Table III. The accuracy of miR-21-5p, miR-22-3p and miR-29c-3p in GSE23739 using 7 classifiers (shown as percentages).

Classifier	Compound covariate	Diagonal linear discriminant analysis	1-Nearest neighbor	3-Nearest neighbor	Nearest centroid	Support vector machines	Bayesian CCP
Accuracy	84	88	86	85	80	88	88

Table IV. The accuracy of three miRNAs in validation sets (GSE26595 and GSE28700) using 7 classifiers (shown as percentages).

Classifier	Compound covariate	Diagonal linear discriminant analysis	1-Nearest neighbor	3-Nearest neighbor	Nearest centroid	Support vector machines	Bayesian CCP
GSE26595	88	88	88	88	87	93	93
GSE28700	64	56	56	60	73	69	71

are represented by the area under the curve (AUC) value and 95% confidence intervals (CI). Experimental data are presented as means ± SD. The results were considered to be statistically significant at *P<0.05, **P<0.01, ***P<0.001, ****P<0.001.

Results

Training set marker selection. Firstly, treating GSE23739 as a training set after normalization, six types of classifiers (see Machine learning in Materials and method) were used to select markers. Since Compound covariate classifier, Diagonal linear discriminant analysis (DLDA) classifier and Support vector machine classifier are linear, we achieved the linear discriminant and calculated the gene weight value of diverse classifiers using maximum likelihood estimate (MLE). Using the same method, the threshold values of the Compound covariate classifier, DLDA classifier, and Support vector machine classifier were determined as 1.469, 2.724 and 0.747, respectively. We set each gene weight value as ω_i and the expression of gene as x_i. If a sample's $\sum_i \omega_i x_i$ > threshold value, then it will be classified as cancerous. Following this principle, we calculated the accuracy of single or a small cluster of miRNAs to discriminate GC from normal tissue.

As elaborated in Fig. 1A, we found that the combination (miR-21-5p, miR-22-3p and miR-29c-3p) of these three markers had the greatest accuracy of the total 175 markers following three steps. The weight values of three miRNAs in the linear classifiers are listed in Table II. The accuracy predicted using 7 markers is documented in Table III. The heat map in Fig. 1B represents the expression data clustering analysis of the three markers in all 80 samples of GSE23739. By employing leave-one-out cross validation, we found the results for sensitivity and specificity in diverse classifiers. Then, we drew 3 ROC curves corresponding to Compound covariate classifier (Fig. 1C), DLDA classifier (Fig. 1D) and Support vector machine classifier (Fig. 1E). Thus, we determined the AUC values for the three curves, 0.9437, 0.9456 and 0.9563, which were high and reliable confirming these markers as having potential diagnostic criteria.

Marker validation. We performed 2 validation sets to validate the three markers. As mentioned above, we investigated the accuracy of prediction in GSE26595 and GES28700 by 7 classifiers (Table IV). The heat map of GSE26595 is shown in Fig. 2A. After calculating the sensitivity and specificity of the data in GSE26595, we determined the AUC of three curves by three linear classifiers separately (0.9563, 0.9625, 0.9688), which confirmed the feasibility of the diagnostic criteria (Fig. 2B-D). The three markers were validated again in another GEO microRNA profile (GSE28700). The heat map is presented in Fig. 2E and the AUC of ROC curves are 0.7645, 0.75 and 0.7789 (Fig. 2F-H).

Confirmation of the selected miRNAs in serum of GC patients. There were 20 samples from healthy volunteers regarded as control subjects and 24 serum samples from GC patients in this study. No significant differences in sex or age (Table I) were noted between the GC patients and the healthy volunteers (P=0.2823, P=0.2929, Student's t-test, respectively).

The expression of three candidate miRNAs (miR-21-5p, miR-22-3p and miR-29c-3p) was assessed by qRT-PCR in individual serum samples. The level of miR-21 was significantly upregulated in GC (P<0.0001) (Fig. 3A). Reversely, in Fig. 3B and C, levels of miR-22 and miR-29c were downregulated in the tumor group (both P<0.0001). These changes were consistent with the results in the training and validation set. Furthermore, we explored the relationship between the expression of these miRNAs with the clinical and pathological features of GC (Table V). There was a higher expression of miR-21 in GC patients with larger tumor sizes (≥5 cm) as previous reported (18).

Target prediction and analysis. Initially, StarBase v2.0 was used to search for the targets of miR-21, miR-22, and miR-29c. All predicted targets of the deregulated miRNAs are illustrated in Fig. 3D. An overview of the Gene Ontology (GO) analysis indicated that the binding attribute of molecular function was high by OmicsBean website (data not shown). Fig. 3E further shows that the highest percentage of genes are involved in the nucleic acid binding activity of enriched processes of level 4. The top 20 highly enriched KEGG pathways are listed in Table VI. Noteworthy, the tumor-suppressor gene, PTEN, was discovered in the class of enriched KEGG pathways, such as: focal adhesion and PI3K-Akt signaling pathway.

Table V. Association between the three selected microRNAs and clinicopathological features of the GC patients.

Clinicopathological features	miR-21		miR-22		miR-29c	
	ΔCt	P-value	ΔCt	P-value	ΔCt	P-value
Tumor location		0.07		0.5		0.35
Cardia	3.07±0.93		4.87±0.83		0.19±0.86	
Not in cardia	3.78±0.86		6.02±1.07		0.83±0.96	
Tumor size (cm)		0.05		0.25		0.99
≥5	3.14±1.05		5.19±1.20		0.23±0.91	
<5	3.89±0.62		5.95±0.89		0.96±0.89	
TNM stage		0.38		0.29		0.27
Early (I+II)	3.14±0.49		4.77±0.69		-0.26±0.52	
Later (III+IV)	3.58±1.03		5.74±1.14		0.78±0.95	
Metastatic status		0.38		0.85		0.74
Yes	3.04±1.25		4.80±1.02		-0.14±0.81	
No	3.55±0.89		5.65±1.11		0.66±0.95	

[a]P<0.05 compared to the control group. GC, gastric cancer; TNM, Tumor, Node, Metastasis.

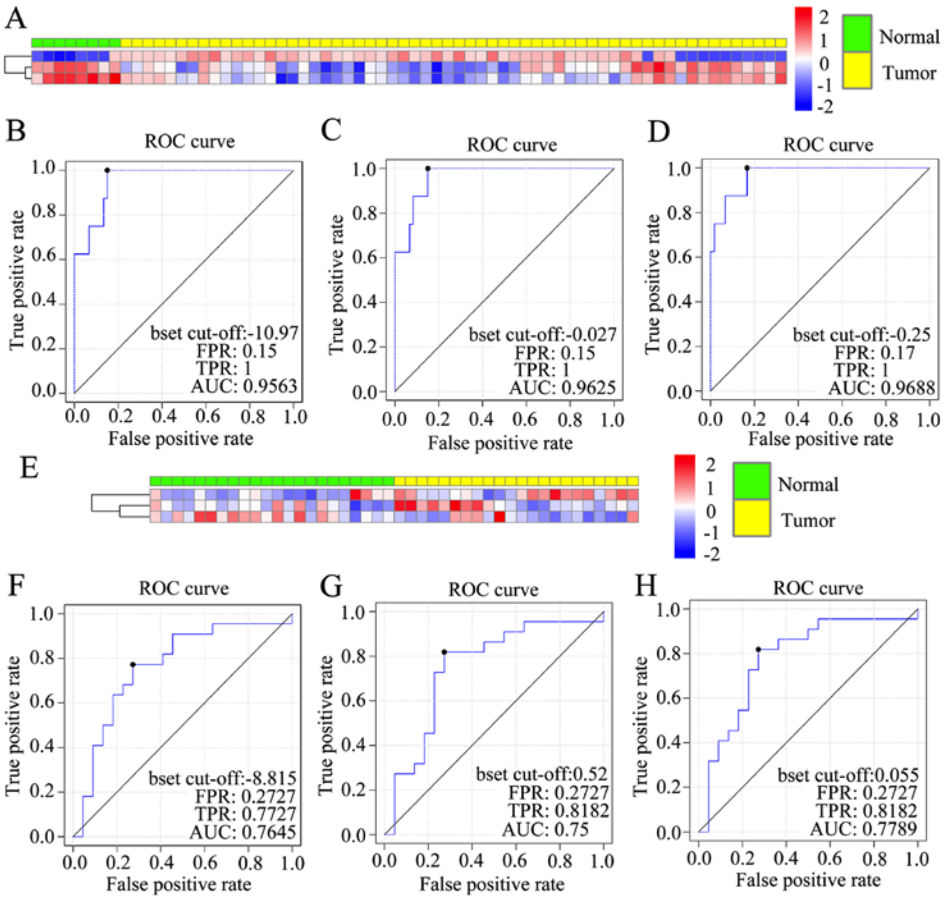

Figure 2. The validation of two prediction sets. Heat map of GSE26595 (A) and GSE28700 (E) was created according to the relative expression of three miRNAs. The lines represent miR-21, miR-22 and miR-29c from the first to the last row. The discriminative ability of the three-miRNA panel is shown in GSE26595 (B-D) and GSE28700 (F-H) using ROC analysis following the sequence mentioned above. FPR, false-positive rate; TPR, true-positive rate; AUC, area under the receiver operating characteristic (ROC) curve.

Furthermore, Cytoscape v3.2 software was applied to focus on the shared genes by both miR-22 and miR-29c.

Their decreased level may contribute to GC by upregulating various oncogenes. By applying miRTarBase database in

Table VI. KEGG pathway analysis of shared target genes of the three miRNAs.

Pathway name	Pathway ID	P-value	Genes
Protein digestion and absorption	hsa04974	1.30E-10	COL7A1; COL3A1
ECM-receptor interaction	hsa04512	1.38E-06	COL4A5; COL4A4
Focal adhesion	hsa04510	6.21E-06	PTEN; COL4A5; COL4A4
PI3K-Akt signaling pathway	hsa04151	3.19E-05	PTEN; COL4A5; COL4A4
Small cell lung cancer	hsa05222	8.28E-05	PTEN; COL4A5; COL4A4
Amoebiasis	hsa05146	2.94E-04	COL3A1; COL4A5; COL4A4
Insulin resistance	hsa04931	5.89E-04	PTEN; PPARA; RPS6KA3
Proteoglycans in cancer	hsa05205	1.04E-03	ESR1; CBL; FLNA; HGF
Phosphatidylinositol signaling system	hsa04070	4.31E-03	PTEN; PIKFYVE; CALM1
MAPK signaling pathway	hsa04010	7.61E-03	RASGRP1; RPS6KA3; FLNA
Pathways in cancer	hsa05200	8.69E-03	PTEN; RASGRP1; COL4A5
N-Glycan biosynthesis	hsa00510	9.19E-03	ALG9; ALG1; MAN1A2
Inositol phosphate metabolism	hsa00562	1.10E-02	PTEN; PIKFYVE; MTMR2
Melanoma	hsa05218	1.10E-02	PTEN; HGF; PTEN; CDK6
Neurotrophin signaling pathway	hsa04722	1.41E-02	CAMK4; CALM1; RPS6KA3
AGE-RAGE signaling pathway in diabetic complications	hsa04933	1.74E-02	COL3A1; COL4A5; COL4A4
Glioma	hsa05214	2.84E-02	PTEN; CALM1; PTEN; CDK6
Long-term potentiation	hsa04720	3.01E-02	CAMK4; GRM5; CALM1
Prostate cancer	hsa05215	3.03E-02	PTEN; CREB1; PTEN; CREB5

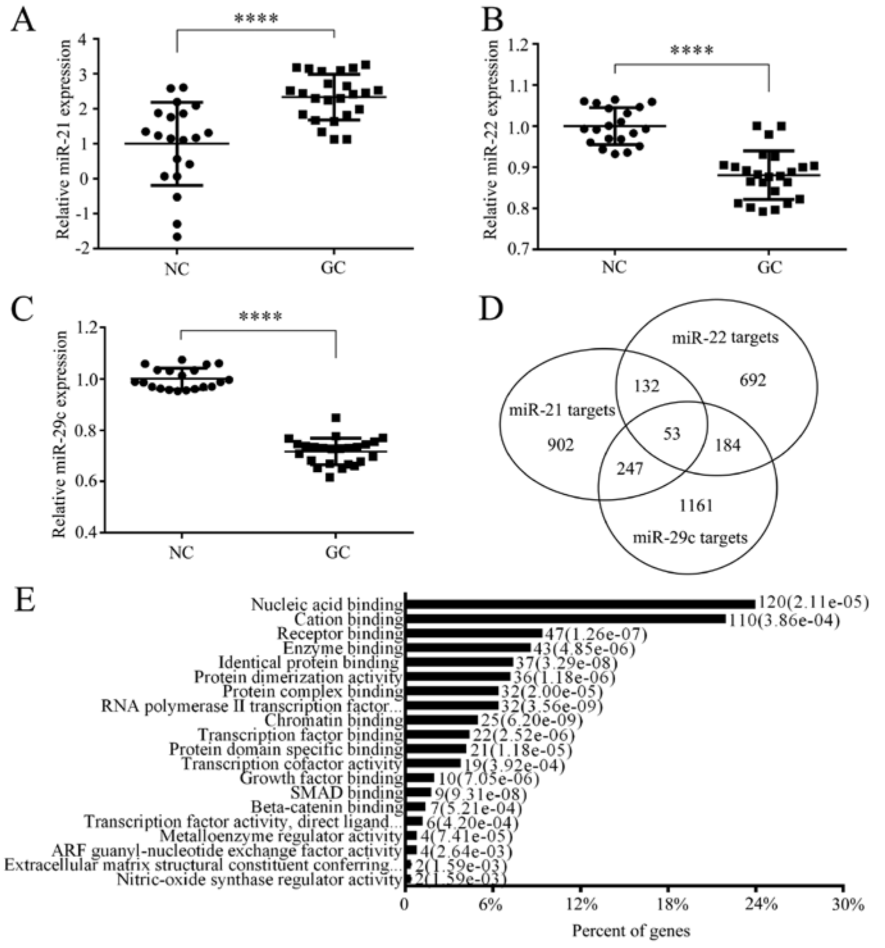

Figure 3. The expression of candidate miRNAs in serum samples and analysis of target genes of the miRNAs. (A-C) The expression of miR-21 was significantly upregulated, while the expression of miR-22 and miR-29c were downregulated in 24 GC tissues compared to 20 control subjects. ****P<0.0001 compared to the healthy control group. Each value is the mean ± SD of three experiments. (D) Venn diagram illustrating the target genes targeted by three miRNAs, which were predicted by Starbase website. (E) Gene ontology (GO) analysis was conducted and the highest percentage of genes are involved in the nucleic acid binding activity of enriched processes of level 4. GC, gastric cancer; NC, normal controls.

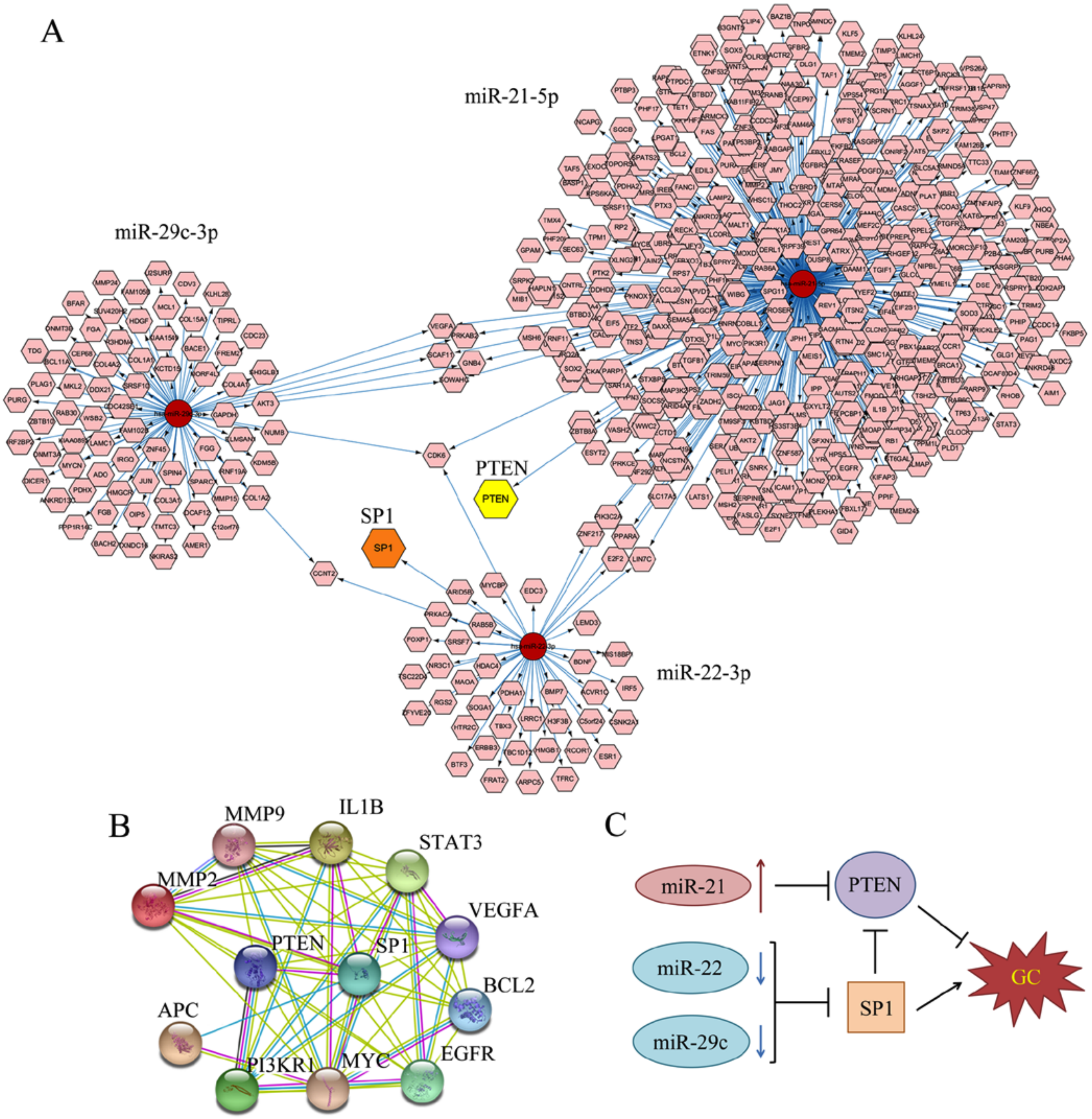

Figure 4. The network of candidate miRNAs and the possible mechanisms. (A) The network includes the individual miRNAs (red circles) and their predicted mRNA target genes (hexagons), obtained from miRTarBase databases. The orange and yellow colors highlight the predicted genes (SP1 and PTEN). (B) The protein-protein-interaction was drawn using STRING website. (C) The diagram illustrates the possible mechanism involving the three miRNAs. GC, gastric cancer.

CyTargetLinker App to search for targets, we created a biological network with numerous nodes which stand for target genes and gave attention to the shared targets between two miRNAs (Fig. 4A). However, there may have been several missing targets as this database has not been included. Sp1, which is a transcription factor and may be associated with poor prognosis of GC patients, was selected from the targets of miR-22 and its node is indicated in orange (19). Recent research also reveals that miR-29c could target Sp1 in lung cancer (20). Therefore, SP1 has been confirmed to be a shared target between two downregulated miRNAs (miR-22 and miR-29c).

STRING was used to screen genes between Sp1 and 476 targets of miR-21 predicted by miRTarBase. A total of 38 relevant genes were selected out. Next, we search for the gene list related to gastric adenocarcinoma in DisGeNet database (umls: C0149826). Then we took the intersection between 38 genes related to Sp1 and 284 GC-related genes. Twelve genes were selected out and STRING was used to predict the interaction between them (Fig. 4B). Finally, we gave attention to PTEN, which has been reported to be targeted by miR-21 in GC and was transcriptionally inhibited by Sp1 (21-24). Sp1 and PTEN were consistent with the results we found in GO analysis and KEGG pathway. In conclusion, Fig. 4C illustrates that the higher level of SP1 and lower level of PTEN may contribute to the progression of GC.

Discussion

It is crucial to identify practical biomarkers for the detection of gastric cancer (GC) in order to improve patient outcomes. In comparison to tissue biopsy, using serum miRNAs as biomarkers is simple, has a lower cost and is non-invasive, which has benefit for the screening and monitoring of tumors (25). A combination of miR-21, miR-22 and miR-29c was identified using machine learning. Their ROC analyses in the training set revealed marked AUC (0.9437, 0.9456 and 0.9563) with more than 80% positive predictive value (PPV) and negative predictive value (NPV) in three linear classifiers. Then we validated them in two training sets and the serum of patients for further confirmation. It should be noted that the number of samples was few and we required further proof to validate these three biomarkers in GC. In the final step, we used tools to identify their targets and elucidate the possible mechanisms.

Application of machine learning to large databases is also called data excavation which means that a large volume of raw data are processed into a small amount of precious material using classifier models. During recent years, machine learning has been widely utilized as a method to predict the progression, susceptibility and recurrence of cancerous conditions (26). For example, machine learning models, including Support vector machine (SVM) classifier, was used to predict childhood acute lymphoblastic leukemia (ALL) relapse based on medical data (27). In the field of diagnostics, Chen *et al* (28) applied four classical machine learning-based classifications to estimate the stage of hepatic fibrosis. Radiomic machine-learning classifiers were applied for prognostic biomarkers of advanced nasopharyngeal carcinoma (29). In this study, we fully utilized the predominance of machine learning and used it for the screening of biological biomarkers in GC. The panel of miR-21, miR-22, and miR-29c was found to have the highest accuracy in predicting GC tissues (Fig. 1A). Machine learning is a novel method that is worthy to be popularized in identifying biomarkers in different types of disease.

Among the three miRNAs identified in this study, miR-21-5p was upregulated in the serum of GC patients, which was consistent with previous research (6,30). As early as 2008, Zhang *et al* (31) found that miR-21 could regulate GC cell invasion and migration. Concurrently, Chan *et al* (32) verified that miR-21 was overexpressed in 92% (34/37) of GC samples and PTEN may be a target gene of miR-21 (21). *H. pylori* infection was found to induce miR-21 and the level of miR-21 was upregulated in gastric juice of GC patients (33). Accumulating evidence indicates that miR-21 can serve as a diagnostic candidate for GC. In contrast, the levels of miR-22 and miR-29c were decreased in our serum samples, which was in accordance with results in other research (34). miR-29 family plays a vital role in tumor-related changes including cell proliferation, cell cycle, cell differentiation, apoptosis and metastasis (35). Han *et al* (36) showed that miR-29c suppressed the initiation of gastric carcinogenesis in transgenic mouse models. Sufficient evidence revealed that the level of miR-22 was downregulated in GC, which was related to lymph node metastasis, poor prognosis in patients (37), and acted as a metastasis suppressor by directly targeting Sp1 (38). Therefore, the combination of these three miRNAs would achieve more specificity than separate miRNAs in the prediction of GC, achieving high accuracy.

Next, we aimed to explain how these three miRNAs work together. Sp1, which functions as a transcription factor, is a ubiquitously expressed, zinc finger-containing DNA binding protein that can activate or repress transcription in a variety of diseases (39). It is overexpressed in GC and is closely correlated with poor outcome (40). miR-22 targets Sp1 and represses GC (38), while miR-29c may function in the same way (41). PTEN is one of the well-known tumor suppressor gene that plays a crucial role in various types of tumors including GC (42) and was validated to be targeted by miR-21 (21-24). Sp1 can inhibit PTEN promoter activity through a specific Sp1-binding site at the PTEN core promote (43). The mechanism is summarized in Fig. 4C. miR-22 and miR-29c both suppress the level of Sp1 and miR-21 suppresses the expression of PTEN inhibited by Sp1, which contributes to the development of GC.

In summary, our study revealed that miRNAs or other biomarkers could be excavated effectively by machine learning. Three miRNAs were screened: miR-21, miR-22 and miR-29c. Their diagnostic potential was evaluated by various classifiers and AUC curves. We then verified their differential expression in the serum of patients and explained this phenomenon by predicting their targets. Further studies will aid in confirming this serum miRNA panel for the diagnosis of GC.

Acknowledgements

This research was supported by a grant from the Shanghai Municipal Commission of Health and Family Planning (no. 20134132).

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J and Jemal A: Global cancer statistics, 2012. CA Cancer J Clin 65: 87-108, 2015.
2. Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, Jemal A, Yu XQ and He J: Cancer statistics in China, 2015. CA Cancer J Clin 66: 115-132, 2016.
3. Bartel DP: MicroRNAs: Target recognition and regulatory functions. Cell 136: 215-233, 2009.
4. Ambros V: The functions of animal microRNAs. Nature 431: 350-355, 2004.
5. Song JH and Meltzer SJ: MicroRNAs in pathogenesis, diagnosis, and treatment of gastroesophageal cancers. Gastroenterology 143: 35-47.e2, 2012.
6. Li BS, Zhao YL, Guo G, Li W, Zhu ED, Luo X, Mao XH, Zou QM, Yu PW, Zuo QF, *et al*: Plasma microRNAs, miR-223, miR-21 and miR-218, as novel potential biomarkers for gastric cancer detection. PLoS One 7: e41629, 2012.
7. Ventura A and Jacks T: MicroRNAs and cancer: Short RNAs go a long way. Cell 136: 586-591, 2009.
8. Shrestha S, Hsu SD, Huang WY, Huang HY, Chen W, Weng SL and Huang HD: A systematic review of microRNA expression profiling studies in human gastric cancer. Cancer Med 3: 878-888, 2014.
9. Berrar DP, Dubitzky W and Granzow M (eds): A practical approach to microarray data analysis. Springer, New York, p368, 2003.

10. Radmacher MD, McShane LM and Simon R: A paradigm for class prediction using gene expression profiles. J Comput Biol 9: 505-511, 2002.
11. Dudoit S, Fridlyand J and Speed TP: Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 97: 77-87, 2002.
12. Efron B, Tibshirani R, Storey JD and Tusher V: Empirical Bayes analysis of a microarray experiment. J Am Stat Assoc 96: 1151-1160, 2001.
13. Wright G, Tan B, Rosenwald A, Hurt EH, Wiestner A and Staudt LM: A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma. Proc Natl Acad Sci USA 100: 9991-9996, 2003.
14. Li L, Darden TA, Weinberg CR, Levine AJ and Pedersen LG: Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method. Comb Chem High Throughput Screen 4: 727-739, 2001.
15. Pal M: Modified nearest neighbour classifier for hyperspectral data classification. Int J Remote Sens 32: 9207-9217, 2011.
16. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M and Haussler D: Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16: 906-914, 2000.
17. Kutmon M, Kelder T, Mandaviya P, Evelo CT and Coort SL: CyTargetLinker: A cytoscape app to integrate regulatory interactions in network analysis. PLoS One 8: e82160, 2013.
18. Wang JL, Hu Y, Kong X, Wang ZH, Chen HY, Xu J and Fang JY: Candidate microRNA biomarkers in human gastric cancer: A systematic review and validation study. PLoS One 8: e73683, 2013.
19. Jiang W, Jin Z, Zhou F, Cui J, Wang L and Wang L: High co-expression of Sp1 and HER-2 is correlated with poor prognosis of gastric cancer patients. Surg Oncol 24: 220-225, 2015.
20. Zhang HW, Wang EW, Li LX, Yi SH, Li LC, Xu FL, Wang DL, Wu YZ and Nian WQ: A regulatory loop involving miR-29c and Sp1 elevates the TGF-β1 mediated epithelial-to-mesenchymal transition in lung cancer. Oncotarget 7: 85905-85916, 2016.
21. Zhang BG, Li JF, Yu BQ, Zhu ZG, Liu BY and Yan M: microRNA-21 promotes tumor proliferation and invasion in gastric cancer by targeting PTEN. Oncol Rep 27: 1019-1026, 2012.
22. Zheng P, Chen L, Yuan X, Luo Q, Liu Y, Xie G, Ma Y and Shen L: Exosomal transfer of tumor-associated macrophage-derived miR-21 confers cisplatin resistance in gastric cancer cells. J Exp Clin Cancer Res 36: 53, 2017.
23. Eto K, Iwatsuki M, Watanabe M, Ida S, Ishimoto T, Iwagami S, Baba Y, Sakamoto Y, Miyamoto Y, Yoshida N, *et al*: The microRNA-21/PTEN pathway regulates the sensitivity of HER2-positive gastric cancer cells to trastuzumab. Ann Surg Oncol 21: 343-350, 2014.
24. Yang SM, Huang C, Li XF, Yu MZ, He Y and Li J: miR-21 confers cisplatin resistance in gastric cancer cells by regulating PTEN. Toxicology 306: 162-168, 2013.
25. Cai H, Yuan Y, Hao YF, Guo TK, Wei X and Zhang YM: Plasma microRNAs serve as novel potential biomarkers for early detection of gastric cancer. Med Oncol 30: 452, 2013.
26. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV and Fotiadis DI: Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13: 8-17, 2014.
27. Pan L, Liu G, Lin F, Zhong S, Xia H, Sun X and Liang H: Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Sci Rep 7: 7402, 2017.
28. Chen Y, Luo Y, Huang W, Hu D, Zheng RQ, Cong SZ, Meng FK, Yang H, Lin HJ, Sun Y, *et al*: Machine-learning-based classification of real-time tissue elastography for hepatic fibrosis in patients with chronic hepatitis B. Comput Biol Med 89: 18-23, 2017.
29. Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, Mo X, Huang W, Tian J and Zhang S: Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. Cancer Lett 403: 21-27, 2017.
30. Sekar D, Krishnan R, Thirugnanasambantham K, Rajasekaran B, Islam VIH and Sekar P: Significance of microRNA 21 in gastric cancer. Clin Res Hepatol Gastroenterol 40: 538-545, 2016.
31. Zhang Z, Li Z, Gao C, Chen P, Chen J, Liu W, Xiao S and Lu H: miR-21 plays a pivotal role in gastric cancer pathogenesis and progression. Lab Invest 88: 1358-1366, 2008.
32. Chan SH, Wu CW, Li AF, Chi CW and Lin WC: miR-21 microRNA expression in human gastric carcinomas and its clinical association. Anticancer Res 28: 907-911, 2008.
33. Karimi Kurdistani Z, Saberi S, Tsai KW and Mohammadi M: MicroRNA-21: Mechanisms of Oncogenesis and its Application in Diagnosis and Prognosis of Gastric Cancer. Arch Iran Med 18: 524-536, 2015.
34. Wang D, Fan Z, Liu F and Zuo J: Hsa-miR-21 and Hsa-miR-29 in tissue as potential diagnostic and prognostic biomarkers for gastric cancer. Cell Physiol Biochem 37: 1454-1462, 2015.
35. Wang Y, Zhang X, Li H, Yu J and Ren X: The role of miRNA-29 family in cancer. Eur J Cell Biol 92: 123-128, 2013.
36. Han TS, Hur K, Xu G, Choi B, Okugawa Y, Toiyama Y, Oshima H, Oshima M, Lee HJ, Kim VN, *et al*: MicroRNA-29c mediates initiation of gastric carcinogenesis by directly targeting ITGB1. Gut 64: 203-214, 2015.
37. Wang W, Li F, Zhang Y, Tu Y, Yang Q and Gao X: Reduced expression of miR-22 in gastric cancer is related to clinicopathologic characteristics or patient prognosis. Diagn Pathol 8: 102, 2013.
38. Guo MM, Hu LH, Wang YQ, Chen P, Huang JG, Lu N, He JH and Liao CG: miR-22 is down-regulated in gastric cancer, and its overexpression inhibits cell migration and invasion via targeting transcription factor Sp1. Med Oncol 30: 542, 2013.
39. Tan NY and Khachigian LM: Sp1 phosphorylation and its regulation of gene transcription. Mol Cell Biol 29: 2483-2488, 2009.
40. Wang L, Wei D, Huang S, Peng Z, Le X, Wu TT, Yao J, Ajani J and Xie K: Transcription factor Sp1 expression is a significant predictor of survival in human gastric cancer. Clin Cancer Res 9: 6371-6380, 2003.
41. Xiao S, Yang Z, Qiu X, Lv R, Liu J, Wu M, Liao Y and Liu Q: miR-29c contribute to glioma cells temozolomide sensitivity by targeting O6-methylguanine-DNA methyltransferases indirectely. Oncotarget 7: 50229-50238, 2016.
42. Li C, Song L, Zhang Z, Bai XX, Cui MF and Ma LJ: MicroRNA-21 promotes TGF-β1-induced epithelial-mesenchymal transition in gastric cancer through up-regulating PTEN expression. Oncotarget 7: 66989-67003, 2016.
43. Kou XX, Hao T, Meng Z, Zhou YH and Gan YH: Acetylated Sp1 inhibits PTEN expression through binding to PTEN core promoter and recruitment of HDAC1 and promotes cancer cell migration and invasion. Carcinogenesis 34: 58-67, 2013.