Twenty-four signature genes predict the prognosis of oral squamous cell carcinoma with high accuracy and repeatability

JIANYONG GAO*, GANG TIAN*, XU HAN and QIANG ZHU

Department of Stomatology, Changhai Hospital, Second Military Medical University, Shanghai 200433, P.R. China

Received February 19, 2017; Accepted August 10, 2017

DOI: 10.3892/mmr.2017.8256

Abstract. Oral squamous cell carcinoma (OSCC) is the sixth most common type cancer worldwide, with poor prognosis. The present study aimed to identify gene signatures that could classify OSCC and predict prognosis in different stages. A training data set (GSE41613) and two validation data sets (GSE42743 and GSE26549) were acquired from the online Gene Expression Omnibus database. In the training data set, patients were classified based on the tumor-node-metastasis staging system, and subsequently grouped into low stage (L) or high stage (H). Signature genes between L and H stages were selected by disparity index analysis, and classification was performed by the expression of these signature genes. The established classification was compared with the L and H classification, and fivefold cross validation was used to evaluate the stability. Enrichment analysis for the signature genes was implemented by the Database for Annotation, Visualization and Integration Discovery. Two validation data sets were used to determine the precise of classification. Survival analysis was conducted followed each classification using the package 'survival' in R software. A set of 24 signature genes was identified based on the classification model with the F_i value of 0.47, which was used to distinguish OSCC samples in two different stages. Overall survival of patients in the H stage was higher than those in the L stage. Signature genes were primarily enriched in 'ether lipid metabolism' pathway and biological processes such as 'positive regulation of adaptive immune response' and 'apoptotic cell clearance'. The results provided a novel 24-gene set that may be used as biomarkers to predict OSCC prognosis with high accuracy, which may be used to determine an appropriate treatment program for patients with OSCC in addition to the traditional evaluation index.

Correspondence to: Dr Qiang Zhu, Department of Stomatology, Changhai Hospital, Second Military Medical University, 168 Changhai Road, Yangpu, Shanghai 200433, P.R. China
E-mail: zhuqiansghhd@aliyun.com

*Contributed equally

Key words: oral squamous cell carcinoma, classification, tumor stage, survival, prognosis

Introduction

Head and neck cancer (HNC) comprises a set of cancers that affect the oral cavity, pharynx and larynx (1), with ~600,000 newly diagnosed cases and ~300,000 mortalities annually (2). Oral squamous cell carcinoma (OSCC) is the most common malignant tumor of the HNCs and the sixth most common cancer worldwide, and accounts for ~90% of all the oral cancers (3,4). OSCC early detection and diagnosis lead to improved survival rates. However, most of OSCC cases are detected in advanced cancer. In this case, delayed detection may result in a high OSCC mortality rate (5). In addition, OSCC has a high recurrence rate in many patients (6). Therefore, the development of novel methods to predict the prognosis of OSCC is urgent.

Several previous studies have revealed that molecular-based classification may improve the prognosis of OSCC. For example, Belbin *et al* (7) were able to distinguish two subgroups of OSCC using a set of molecular signatures that are distinct in the two different groups such as transforming growth factor-β. Another study identified crucial gene expressions in tumors and four subgroups of OSCC using cDNA microarrays (8). Human papilloma virus (HPV) infection has a close relationship with OSCC, and the high risk of infection is associated with the high risk of developing OSCC (9). In addition, the etiologies of HPV-positive and HPV-negative OSCC subtypes are different, and 347 differentially expressed genes have been identified in these two groups, such as thymidylate synthetase, stathmin 1 and cyclin D1 (9). Notably, HPV-positive oral cancers have an improved response to treatment and better prognosis compared with HPV-negative OSCCs (10). One previous study further predicted that HPV types 16 and 18 may be two independent risk factors for oral cancer (11). A recently study used two expression data sets, a training data set and a validation data set to identify genes with distinct expressions between patients with HPV-negative OSCC and normal controls. Subsequently, a set of 131 gene signatures was selected, which was reduced to a total of 13 gene signatures that were identified as the best survival predictors for patients with HPV-negative OSCC (12).

However, none of the aforementioned studies compared the precision of these classifications with the tumor-node-metastasis (TNM) stage. In addition, the patients used in a study by Lohavanichbutr *et al* were in different tumoral stages, and additional data is required (12). Therefore, the present study reanalyzed their data set, GSE41613, and extracted only the

data associated with patients with HPV-negative OSCC. The patients were subsequently classified based on the TNM stage; subsequently, signature genes were identified in the two groups, and patient samples were further classified based on these signature genes. Following this classification, two validation data sets, GSE42743 (12) and GSE26549 (13), were used to detect the precision of the signature-gene-based classification. In addition, survival analysis was performed in each classification. Through these comprehensive analyses, the present study aimed to identify several gene signatures that were able to distinguish patients with HPV-negative OSCC at different TNM stages.

Materials and methods

Data resource and the pretreatments. Data set GSE41613 (12) was obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). This expression profile was based on the Affymetrix Human Genome U133 Plus 2.0 Array platform, and included 97 human samples from patients with HPV-negative OSCC, along with their clinical follow-up information. Among these samples, 30 patients succumbed to OSCC, 21 succumbed to other diseases and the remaining 46 patients survived, and these were classified in stages I to IV.

The data, which were already normalized, were downloaded and 76 of the OSCC-related samples were used in the present study; that is, the 30 patients that succumbed to OSCC and the 46 surviving patients. A set of 54,613 probe values was acquired, following the elimination of probes in empty carriers.

Classification of different samples. Patients were classified based on TNM stage; those in stages I and II were placed in the low (L) stage, whereas those in stages III and IV were placed in the high (H) stage. Survival analysis was an important prognostic analysis and it was implemented according to the survival package in the R software (R 3.4.1; https://www.r-project.org/) (14).

Signature gene identification in L and H samples. To identify the optimal signature genes that were able to distinguish samples between the L and H groups, the disparity index s was calculated according to each gene expression, based on the formula: $F_i=(mean_{iH}-mean_{iL})/(SD_{iH} + SD_{iL})$; where i represents a gene and F_i represents its corresponding disparity index in the different samples (15). The significance of each gene to distinguish the different groups of samples was calculated using the permutation test by perm in R (16). The iteration step k with an interval of 0.01 in 0-1 was used to identify the best threshold of Fi. The optimal signature genes were selected based on the criteria $|F|>k$ and $P<0.01$. The selected genes were used to establish the classification model:

$$b_i=(mean_{iH}-mean_{iL})/2.$$

$$V_i=F_i(e_i-b_i)$$

$$PS_j = \sum_{i=0}^{N} V_i \Big/ \sum_{i=0}^{N} |V_i|$$

Where i represents a gene, e_i represents its gene expression value, N represents selected gene numbers and PS_j is the score reflecting the classification of the sample j. Two classifications, positive and negative, were identified based on these scores. The overlap scale of the samples under these two classifications compared with that under the H and L classifications were calculated to identify the optimal signature genes under the classification threshold that had the best consistency.

Fivefold cross validation. To detect the stability of using these signature genes to classify the samples, fivefold cross validation was implemented 10 times, and the overlap scale of the classifications under each cross validation with the classifications of H and L was calculated.

Clinical prognostic analysis for samples classified based on signature genes. To determine the prognostic difference between the samples classified by signature genes, the samples were clustered into two groups based on the PSj scores calculated by the established model, and the survival package in R was used to analyze the prognostic difference of the two clusters (14).

Expression profile analysis of signature genes. Unsupervised hierarchical clustering was conducted for the signature genes, based on their expression values in different samples. Subsequently, the prognostic differences in different clusters were identified using the Kaplan-Meier package in R (17).

Enrichment analyses of signature genes. Function and pathway enrichment analyses of the signature genes were performed based on the Gene Ontology (GO; http://www.geneontology.org) and Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.jp/kegg/pathway.html) databases, respectively, and the Database for Annotation, Visualization and Integration Discovery (DAVID; http://david.abcc.Ncifcrf.gov) online tool (18). A threshold of $P<0.05$ was used to indicate significant function and pathway categories.

Multivariate survival analysis of signature genes. Signature genes were examined by multivariate survival analysis to determine their putative effects on prognosis as a whole. Receiver operating characteristic (ROC) curve was depicted using the SurvivalROC package in R (19).

Validation by individual gene data sets. To validate the reproducibility of the established model in classifying the OSCC samples to different prognosis groups, two independent gene expression profiles, GSE42743 (12) and GSE26549 (13), were downloaded from the GEO database, which were based on the Affymetrix Human Genome U133 Plus 2.0 Array platform and the Affymetrix Human Gene 1.0 ST Array [transcript (gene) version] platform, respectively. A total of 103 samples were in the GSE42743 data set, which also contained follow-up information of 23 patients succumbed to OSCC and 22 patients alive until the final follow-up time point. In this data set, raw data in the CEL format was obtained by ReadAffy in the affy package of R, and was normalized by robust multichip average (20,21). The GSE26549 data set comprised 86 samples: 35 were recurrent patients and 51 were non-recurrent, and the normalized

data in this profile was downloaded. Cox regression was used to analyze these data sets, and to compare prognostic and recurrent differences between different samples using the survival package in R (14).

Results

Survival analysis of H and L samples. Samples were divided into two types, H and L, based on TNM stages. Survival analysis results indicated that there were significant differences between the two classifications, and patients in the L stage had a significantly higher survival probability compared with those in the H stage ($P=2.00x10^{-05}$; Fig. 1).

Threshold of signature genes in different samples. The gene set contained a total of 54,613 probes in the 76 GSE41613 tumor samples used. The overlap scale of the classifications obtained was compared using different F_i values in the classification model and the H and L classifications. As a result, the classification accuracy under different F_i values was not completely consistent: When the F_i was low and more gene sets were contained than others, the accuracy was ~0.86 (Fig. 2A); when F_i was 0.35-0.5, the accuracy was slightly improved and reached a maximum at F_i=0.47. However, when the F_i was >0.5, the accuracy exhibited a linear decline (Fig. 2A). Therefore, F_i=0.47 was used as the cut-off value to classify the samples. Results from fivefold cross-validation analysis indicated that the accuracy was almost always >80%, and the average value was 0.897 (Fig. 2B). This result confirmed the precise classification using gene sets with F_i=0.47).

Signature genes and gene expression profile analysis. Signature gene sets under the threshold of F_i=0.47 were selected, and 24 genes were identified (Table I). These genes were subsequently used to mark each sample, based on the classification model. Samples in H stage had a significantly higher score compared with those in L stage, and '0' was used as the boundary to divide the two samples (Fig. 3A). Survival analysis using 0 as the boundary indicated that there were significant prognostic differences between the two sample clusters ($P=6.30x10^{-07}$; Fig. 3B). Notably, the difference was more significant than those determined in H and L classifications, which suggested that this score-based model was able to adjust the original model with H and L. Unsupervised hierarchical clustering analysis of gene expression profiles revealed that these signature genes could distinctly divide the samples into two classifications: Cluster1 and Cluster2 (Fig. 3C). Kaplan-Meier survival curve of these sample clusters also demonstrated significant differences ($P=2.00x10^{-05}$; Fig. 3D). These data suggested that the samples could be distinguished by using gene expression clustering.

Function enrichment of signature genes. As indicated in Table II, the 24 signature genes were enriched in 1 KEGG pathway, ether lipid metabolism pathway [P=0.0472; genes: 1-acylglycerol-3-phosphate O-acyltransferase 2 (*AGPAT2*) and phosphatidic acid phosphatase type 2B (*PPAP2B*)], and 31 GO functional categories, including 1 cellular component, cell projection part [P=0.0263; genes: protease, serine 12 (*PRSS12*), coiled-coil and C2 domain containing 2A (*CC2D2A*) and adenosine deaminase (*ADA*)] and 30 biological processes (BPs), such as phagocytosis [P=0.0016; genes: thrombospondin 1 (*THBS1*), solute carrier family 11 member 1 (*SLC11A*) and Jumonji domain containing 6 (*JMJD6*)], regulation of dendritic cell antigen processing and presentation (P=0.0025; genes: *THBS1* and *SLC11A1*), apoptotic cell clearance (P=0.0075; genes: *THBS1* and *JMJD6*), T cell activation (P=0.0107; genes: *ADA*, *SLC11A1* and *JMJD6*), lymphocyte activation during immune response (P=0.0224; genes: *ADA*, *SLC11A1*) and leukocyte activation during immune response (P=0.0443; genes: *ADA* and *SLC11A1*).

Survival analysis of 24 signature genes. The 24 signature genes exhibited a clustering effect with an area under the ROC curve (AUC) of 0.97 (Fig. 4A), and a significant difference in survival was identified between the H and L classifications ($P=2.48x10^{-19}$; Fig. 4B), which suggested that the 24 genes were able to effectively classify different samples and to predict the prognostic risk.

Validation of the classification by individual data sets. The model containing 24 signature genes were demonstrated to be able to classify samples into two prognosis distinct groups in the GSE42743 data set (AUC=0.994) (Fig. 5A). The survival time between high and low risk samples was different significantly ($P=4.55x10^{-15}$; Fig. 5B).

Similarly, the model classified samples into two recurrence risk groups in the GSE26549 data set (AUC=0.984; Fig. 5C). A Kaplan-Meier curve indicated a significant different recurrence risk between high and low risk samples ($P=1.41x10^{-14}$; Fig. 5D).

Discussion

OSCC has a poor prognosis and molecular-based classification provide improved the prognosis. In the present study, 24 signature genes, including *AGPAT2*, *PPAP2B*, *SLC11A1*, *JMJD6* and *ADA*, were identified that were able to classify the patients with HPV-negative OSCC into two different stages, H and L. They were significantly enriched in the ether lipid metabolism pathway and immune response- or apoptosis-related GO BPs.

The protein encoded by *AGPAT2* (1-AGPAT 2) is specific for lysophosphatidic acid (LPA) (22). It is involved in the lipid metabolism, as it catalyzes LPA conversion into phosphatidic acid (PA), a crucial intermediate step in phospholipid biosynthesis (23). 1-AGPAT 2 is an essential factor for adipogenesis; it serves a role in controlling adipogenesis by mediating the activation of phosphatidylinositol 3-kinase (PI3K)/Akt signaling (24). Mutation of this gene may result in an adipogenic defects (24). One previous study revealed that disruption of *AGPAT2* leads to severe congenital generalized lipodystrophy in humans (25). *PPAP2B* is a phosphatidic acid phosphatase (PAP) family member that converts PA to diacylglycerol, and the PAP2B protein was reported to be involved in the regulation of intracellular lipid metabolism (26).

Lipid metabolism has an important role in cancer progression. It has been proposed that increased lipid metabolic flux may serve as the substrate source for phospholipid synthesis in the rapid growth stage of cancer cells (27). Notably, fatty acid synthase was demonstrated to be necessary

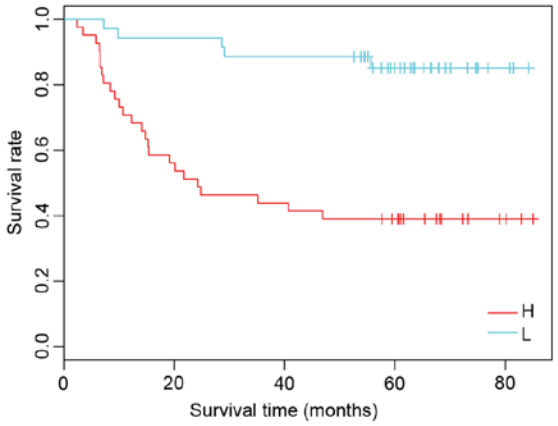

Figure 1. Kaplan-Meier curve analysis indicated a significant difference in survival between H (n=41) and L (n=35) stages samples based on tumor-node-metastasis classification. P=2.00x10^{-05}. H, high; L, low.

during the proliferation of human OSCCs (28). In addition, fatty-acid-binding protein 5 is upregulated in OSCC at the early stage, and this upregulated expression was reported to improve OSCC cell proliferation and invasiveness (29). Results from the present study predicted *AGPAT2* and *PPAP2B* as two signature genes that could differentiate OSCC samples from different stages, and both of these genes were significantly enriched in the ether lipid metabolism pathway. These data suggested that the two genes may serve crucial roles in the progression of OSCC by regulating lipid metabolism, and may be used as therapeutic markers for early stage OSCC prognosis.

The immune system is known to serve crucial roles in the control of cancer development (30,31). Inadequate host immune responses may account for the high incidence rates of cancer and poor prognosis (32). Patients with OSCC in different pathological lymph node (pN) statuses exhibit different overall survival rates; patients with OSCC in pN3 (that is, extracapsular spread) have a lower overall survival compared with patients in other stages, which may be due to a relatively decreased host immune response (32). A previous study using an immunoproteomics method identified several host immune response-related protein candidates in the serum of patients with OSCC, such as clusterin, haptoglobin and complement C3c (33).

The multi-pass membrane protein SLC11A1, also known as natural resistance-associated macrophage protein 1 (NRAMP1), serves important role in host innate immune response against infections (34). ADA functions in the process of A-to-I RNA editing to generate the inosine (I) from adenosine (A); the double-stranded RNA structure may then trigger innate immune responses (35). Elevated *ADA* levels have been detected in a number of cancers, such as colorectal cancer and breast cancer (36,37). In patients with OSCC, the expression levels of *ADA* are significantly increased compared with the healthy control patients (38). Notably, this increased expression level was significantly associated with the histopathological grade (38). The present results indicated that *SLC11A1* and *ADA* were two of the signature genes that were differentially expressed in the two stages of the classification model, and both were enriched in GO BPs, such as positive regulation of adaptive immune response and leukocyte activation during immune response, which suggested that they may function in the development of OSCC through the involvement of immune response-related processes. Based on these results, the present study speculated that *SLC11A1* and *ADA* may also be used as prognostic targets in different tumor stages of OSCC.

The protein encoded by *JMJD6* was previously considered to serve a role in the phagocytosis of apoptotic cells (39). However, subsequent studies failed to confirm this function and, conversely, indicated that it translocates to the nucleus and serves as a histone arginine demethylase (40), or it may be responsible for angiogenic sprouting (41). Another study reported that *JMJD6* may improve cancer stem cell (CSC)

Table I. List of 24 signature genes.

Gene symbol	Gene name
ADA	Adenosine deaminase
CC2D2A	Coiled-coil and C2-domain containing 2A
C9ORF102	Chromosome 9 open reading frame 102
PRSS12	Protease, serine 12 (also known as neurotrypsin and motopsin)
TNXB	Tenascin XB
SLC11A1	Solute carrier family 11 member 1 (also known as natural resistance-associated macrophage protein 1)
GAPVD1	GTPase activating protein and VPS9 domains 1
THBS1	Thrombospondin 1
C19ORF53	Chromosome 19 open reading frame 53
IGSF10	Immunoglobulin superfamily, member 10
PLGLB2	Plasminogen-like B2
ROD1	ROD1 regulator of differentiation 1 (also known as polypyrimidine tract-binding protein 3)
AGPAT2	1-acylglycerol-3-phosphate O-acyltransferase 2 (also known as lysophosphatidic acid acyltransferase β)
SESN3	Sestrin 3
CSNK1G1	Casein kinase 1 γ1
HMGN3	High-mobility group nucleosomal-binding domain 3
SLC2A3	Solute carrier family 2 member 3
FAM161A	Family with sequence similarity 161 member A
DDX31	DEAD-box helicase 31
JMJD6	Jumonji-domain containing 6 (also known as arginine demethylase and lysine hydrolase)
PPAP2B	Phosphatidic acid phosphatase type 2B (also known as phospholipid phosphatase 3)
YEATS2	YEATS-domain containing 2
SERTAD4	SERTA-domain containing 4
NAPEPLD	N-acyl phosphatidylethanolamine phospholipase D

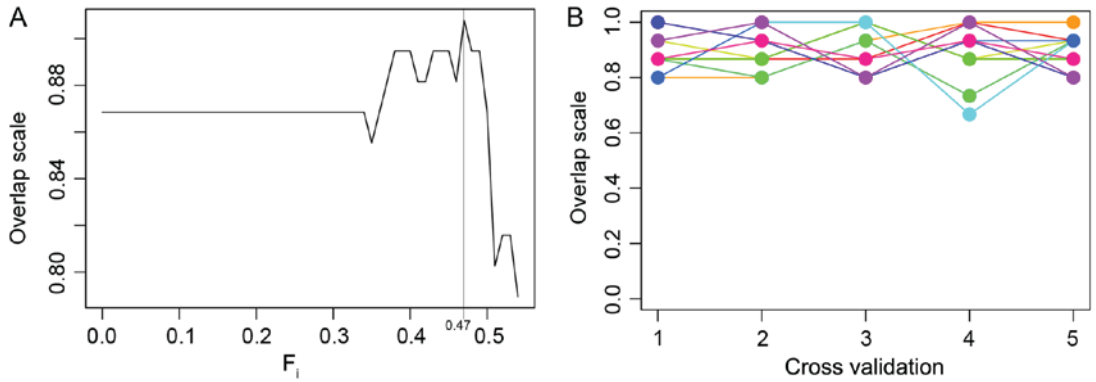

Figure 2. Optimal threshold selection in the classification model. (A) Overlap scale of classifications using the model with the High and Low classifications. Gene sets with $|F_i|>k$ were selected, and k from 0-1 was set with a step size of 0.01; F_i=0.47 (vertical line) was used as the cut-off value to classify the samples. (B) 5-fold cross validation results for 10 iterations, which are indicated by the different colored lines. F_i, disparity index; k, iteration step.

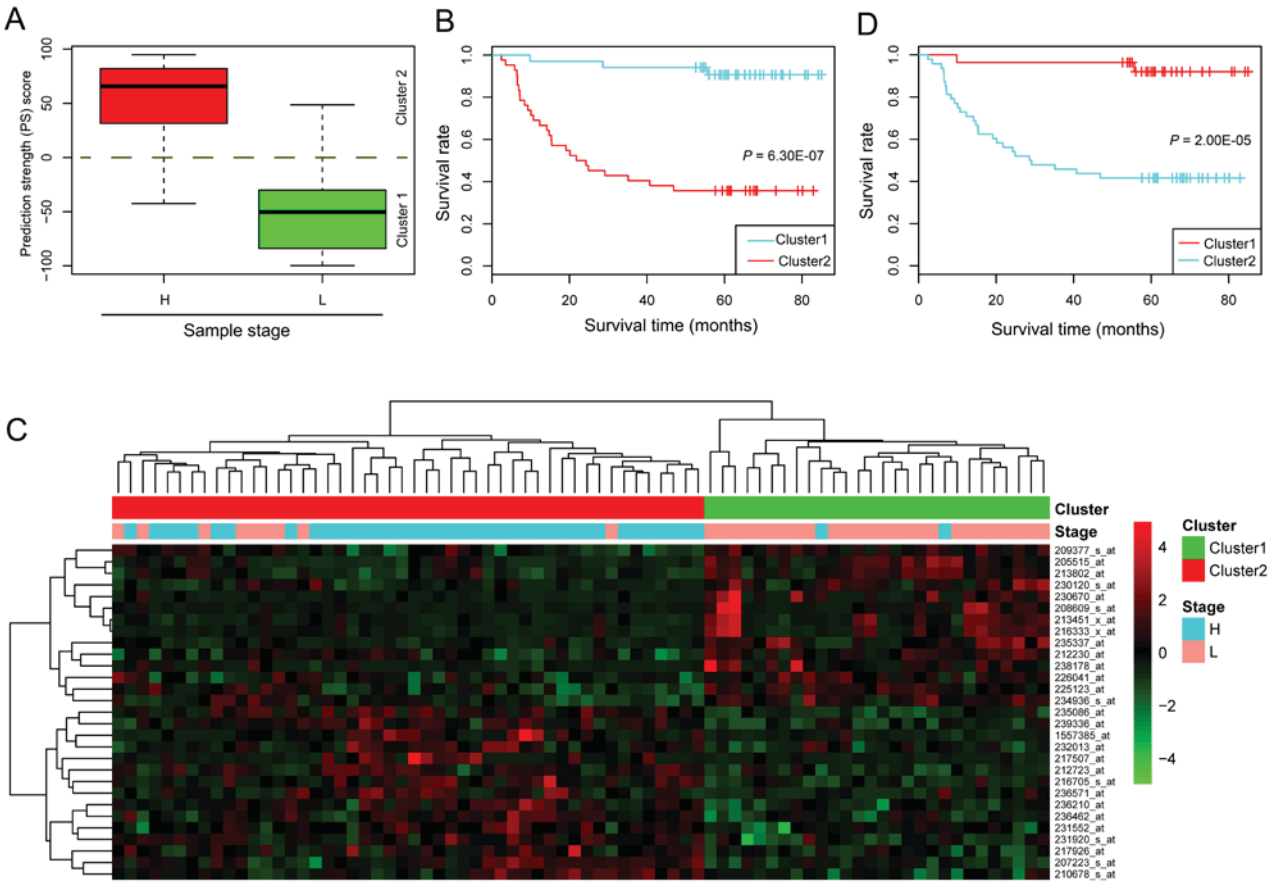

Figure 3. Clustering analysis of signature gene expressions and corresponding survival analyses. (A) Signature genes were used to mark samples in H and L classifications under the score-classification model. (B) Kaplan-Meier curve indicating a significant difference in survival between Cluster 1 and Cluster2 classified with the boundary of '0'. (C) Heat map of signature gene expressions in two cluster samples. (D) Kaplan-Meier curve indicating a significant difference in survival between Cluster1 and Cluster2 that were identified by expression profiling of the signature genes. H, high; L, low.

phenotypes in OSCC cells (42). In addition, increased *JMJD6* expression was previously demonstrated in CSC-enriched populations of OSCC cell lines, which may contribute to oral carcinogenesis and, therefore, it has been suggested as a potential biomarker of oral cancer (43). In the present study, *JMJD6* was amongst the 24 signature genes able to classify patients with OSCC into different tumor stages. Notably, *JMJD6* was significantly enriched in the GO BP category apoptotic cell clearance, which indicated a potential role for JMJD6 in clearing apoptotic cells, at least in OSCC cells. Therefore, *JMJD6* may be considered another novel biomarkers for OSCC prognosis, relating to different tumor stages.

Although the accuracy of using these 24 signature genes to classify different tumor stages of OSCC was validated by other data sets and provided satisfactory results, there were several limitations to the present study: i) None of the identified gene

Table II. GO function term and KEGG pathway enrichment analysis of 24 signature genes.

A, GOTERM_BP	P-value	Genes
GO:0006909~phagocytosis	0.0016	*THBS1*, *SLC11A1* and *JMJD6*
GO:0006897~endocytosis	0.0024	*THBS1*, *SLC11A1*, and *JMJD6*
GO:0010324~membrane invagination	0.0024	*THBS1*, *SLC11A1*, and *JMJD6*
GO:0002604~regulation of dendritic cell antigen processing and presentation	0.0025	*THBS1* and *SLC11A1*
GO:0002577~regulation of antigen processing and presentation	0.0025	*THBS1* and *SLC11A1*
GO:0051240~positive regulation of multicellular organismal process	0.0033	*THBS1*, *AGPAT2*, *ADA* and *SLC11A1*
GO:0001819~positive regulation of cytokine production	0.0056	*THBS1*, *AGPAT2* and *SLC11A1*
GO:0043277~apoptotic cell clearance	0.0075	*THBS1* and *JMJD6*
GO:0042110~T cell activation	0.0107	*ADA*, *SLC11A1* and *JMJD6*
GO:0016044~membrane organization	0.0112	*THBS1*, *SLC11A1*, and *JMJD6*
GO:0051241~negative regulation of multicellular organismal process	0.0176	*THBS1*, *ADA* and *SLC11A1*
GO:0042116~macrophage activation	0.0187	*SLC11A1* and *JMJD6*
GO:0001817~regulation of cytokine production	0.0212	*THBS1*, *AGPAT2* and *SLC11A1*
GO:0002285~lymphocyte activation during immune response	0.0224	*ADA* and *SLC11A1*
GO:0006644~phospholipid metabolic process	0.0232	*AGPAT2*, *PPAP2B* and *NAPEPLD*
GO:0002685~regulation of leukocyte migration	0.0249	*THBS1* and *ADA*
GO:0046649~lymphocyte activation	0.0253	*ADA*, *SLC11A1* and *JMJD6*
GO:0019637~organophosphate metabolic process	0.0256	*AGPAT2*, *PPAP2B* and *NAPEPLD*
GO:0016192~vesicle-mediated transport	0.0335	*THBS1*, *SLC11A1*, and *JMJD6*
GO:0048584~positive regulation of response to stimulus	0.0347	*THBS1*, *ADA* and *SLC11A1*
GO:0002684~positive regulation of immune system process	0.0352	*THBS1*, *ADA* and *SLC11A1*
GO:0045321~leukocyte activation	0.0363	*ADA*, *SLC11A1* and *JMJD6*
GO:0002824~positive regulation of adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	0.0371	*ADA* and *SLC11A1*
GO:0001568~blood vessel development	0.0372	*THBS1*, *PPAP2B* and *JMJD6*
GO:0002821~positive regulation of adaptive immune response	0.0383	*ADA* and *SLC11A1*
GO:0001944~vasculature development	0.0388	*THBS1*, *PPAP2B* and *JMJD6*
GO:0002366~leukocyte activation during immune response	0.0443	*ADA* and *SLC11A1*
GO:0002263~cell activation during immune response	0.0443	*ADA* and *SLC11A1*
GO:0001818~negative regulation of cytokine production	0.0467	*THBS1* and *SLC11A1*
GO:0001775~cell activation	0.0495	*ADA*, *SLC11A1* and *JMJD6*
B, GOTERM_CC	**P-value**	**Genes**
GO:0044463~cell projection part	0.0263	*PRSS12*, *CC2D2A* and *ADA*
C, KEGG_PATHWAY	**P-value**	**Genes**
hsa00565: Ether lipid metabolism	0.0472	*AGPAT2* and *PPAP2B*

ADA, adenosine deaminase; *AGPAT2*, 1-acylglycerol-3-phosphate O-acyltransferase 2; BP, biological process; CC, cellular component; *CC2D2A*, coiled-coil and C2-domain containing 2A; GO, gene ontology; *JMJD6*, Jumonji-domain containing 6; KEGG, Kyoto Encyclopedia of Genes and Genomes; *NAPEPLD*, N-acyl phosphatidylethanolamine phospholipase D; *PPAP2B*, phosphatidic acid phosphatase type 2B; *PRSS12*, protease, serine 12; *SLC11A1*, solute carrier family 11 member 1; *THBX*, thrombospondin 1.

expressions in OSCC, particularly in HPV-negative OSCC, were validated experimentally; and ii) the obtained OSCC sample data were classified into L and H stages, which may have caused deviations from the true TNM stages. Further studies using additional data sets are required to confirm the precision of our classification and the signature gene functions.

In conclusion, a novel 24-gene set was identified that was able to predict OSCC prognosis with high accuracy,

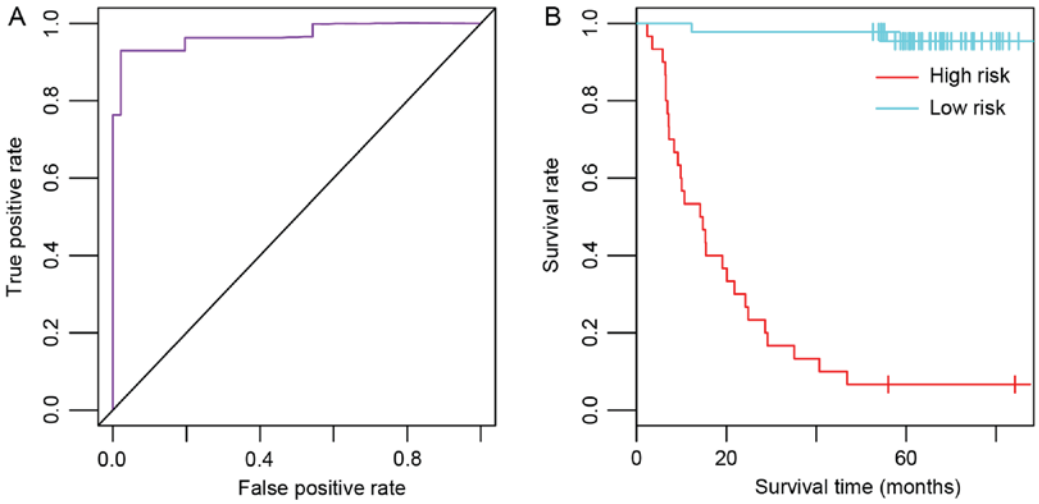

Figure 4. Multivariate survival analysis of 24 signature genes. (A) ROC curve; AUC=0.97. (B) Kaplan-Meier curve indicating a significant difference in survival between high-risk and low-risk samples identified by the multivariate prognosis of 24 genes; $P=2.48x10^{-19}$. AUC, area under the ROC curve; ROC, receiver operating characteristic.

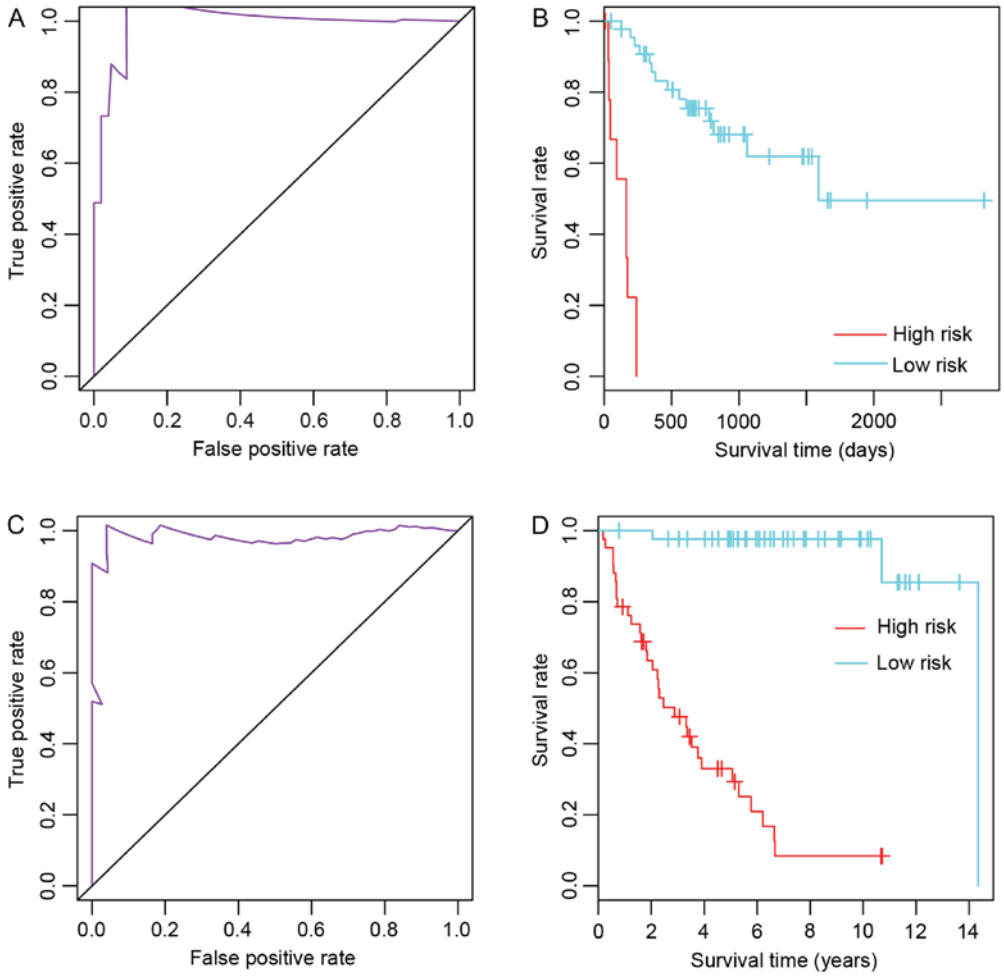

Figure 5. Validation of survival analysis using two additional data sets. (A) ROC curve of the classification in the GSE42743 data set; AUC=0.994. (B) Kaplan-Meier curve indicating a significant difference in survival between high and low risk samples in GSE42743; $P=4.55x10^{-15}$. (C) ROC curve of the recurrent classification in the GSE26549 data set; AUC=0.984. (D) Kaplan-Meier curve indicating a significant difference between high and low risk samples in GSE26549; $P=1.41x10^{-14}$. AUC, area under the ROC curve; ROC, receiver operating characteristic.

which may have the benefit of aiding in the determination of an appropriate treatment program for patients with OSCC, in addition to the traditional evaluation index. *AGPAT2*, *PPAP2B*, *SLC11A1*, *ADA* and *JMJD6* may be biomarkers for

OSCC prognosis; however, further experimental validation is required to confirm these predictions.

Acknowledgements

The study was supported by a grant from The General Program of Health Bureau of Shanghai (grant no. 20133124), and The 1255 project of Changhai Hospital of Second Military Medical University (grant no. CH125541800).

References

1. Chin D, Boyle GM, Porceddu S, Theile DR, Parsons PG and Coman WB: Head and neck cancer: Past, present and future. Expert Rev Anticancer Ther 6: 1111-1118, 2006.
2. Wyss AB, Hashibe M, Lee YA, Chuang SC, Muscat J, Chen C, Schwartz SM, Smith E, Zhang ZF, Morgenstern H, *et al*: Smokeless Tobacco Use and the Risk of Head and Neck Cancer: Pooled Analysis of US Studies in the INHANCE Consortium. Am J Epidemiol: Oct 15, 2016 (Epub ahead of print).
3. Cannonier SA, Gonzales CB, Ely K, Guelcher SA and Sterling JA: Hedgehog and TGFβ signaling converge on Gli2 to control bony invasion and bone destruction in oral squamous cell carcinoma. Oncotarget 7: 76062-76075, 2016.
4. Yakob M, Fuentes L, Wang MB, Abemayor E and Wong DTW: Salivary biomarkers for detection of oral squamous cell carcinoma-current state and recent advances. Curr Oral Health Rep 1: 133-141, 2014.
5. Zini A, Czerninski R and Sgan-Cohen HD: Oral cancer over four decades: Epidemiology, trends, histology, and survival by anatomical sites. J Oral Pathol Med 39: 299-305, 2010.
6. Wu JY, Yi C, Chung HR, Wang DJ, Chang WC, Lee SY, Lin CT, Yang YC and Yang WC: Potential biomarkers in saliva for oral squamous cell carcinoma. Oral Oncol 46: 226-231, 2010.
7. Belbin TJ, Singh BI, Socci N, Wenig B, Smith R, Prystowsky MB and Childs G: Molecular classification of head and neck squamous cell carcinoma using cDNA microarrays. Cancer Res 62: 1184-1190, 2002.
8. Chung CH, Parker JS, Karaca G, Wu J, Funkhouser WK, Moore D, Butterfoss D, Xiang D, Zanation A, Yin X, *et al*: Molecular classification of head and neck squamous cell carcinomas using patterns of gene expression. Cancer Cell 5: 489-500, 2004.
9. Lohavanichbutr P, Houck J, Fan W, Yueh B, Mendez E, Futran N, Doody DR, Upton MP, Farwell DG, Schwartz SM, *et al*: Genome-wide gene expression profiles of HPV-positive and HPV-negative oropharyngeal cancer: Potential implications for treatment choices. Arch Otolaryngol Head Neck Surg 135: 180-188, 2009.
10. Fakhry C, Westra WH, Li S, Cmelak A, Ridge JA, Pinto H, Forastiere A and Gillison ML: Improved survival of patients with human papillomavirus-positive head and neck squamous cell carcinoma in a prospective clinical trial. J Natl Cancer Inst 100: 261-269, 2008.
11. Al-Malkey MK, Abbas AAH and Yaseen NY: Human papilloma virus types 16 and 18 in a sample of iraqis patients presented with oral cancer. Iraqi J Med Sci 14: 174-181, 2016.
12. Lohavanichbutr P, Méndez E, Holsinger FC, Rue TC, Zhang Y, Houck J, Upton MP, Futran N, Schwartz SM, Wang P and Chen C: A 13-gene signature prognostic of HPV-negative OSCC: Discovery and external validation. Clin Cancer Res 19: 1197-1203, 2013.
13. Saintigny P, Zhang L, Fan YH, El-naggar AK, Papadimitrakopoulou VA, Feng L, Lee JJ, Kim ES, Ki Hong W and Mao L: Gene expression profiling predicts the development of oral cancer. Cancer Prev Res (Phila) 4: 218-229, 2011.
14. O'Quigley J and Moreau T: Cox's regression model: Computing a goodness of fit statistic. Comput Methods Programs Biomed 22: 253-256, 1986.
15. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al*: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286: 531-537, 1999.
16. Fay MP and Shaw PA: Exact and asymptotic weighted logrank tests for interval censored data: The interval R package. J Stat Softw 36: pii: i02, 2010.
17. Lacny S, Wilson T, Clement F, Roberts DJ, Faris PD, Ghali WA and Marshall DA: Kaplan-Meier survival analysis overestimates the risk of revision arthroplasty: A meta-analysis. Clin Orthop Relat Res 473: 3431-3442, 2015.
18. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA: DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol 4: P3, 2003.
19. Heagerty PJ, Thomas L and Pepe MS: Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics 56: 337-344, 2000.
20. Bolstad BM, Irizarry RA, Astrand M and Speed TP: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193, 2003.
21. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264, 2003.
22. Kitson AP, Stark KD and Duncan RE: Enzymes in brain phospholipid docosahexaenoic acid accretion: A PL-ethora of potential PL-ayers. Prostaglandins Leukot Essent Fatty Acids 87: 1-10, 2012.
23. Boutet E, El Mourabit H, Prot M, Nemani M, Khallouf E, Colard O, Maurice M, Durand-Schneider AM, Chrétien Y, Grès S, *et al*: Seipin deficiency alters fatty acid Delta9 desaturation and lipid droplet formation in Berardinelli-Seip congenital lipodystrophy. Biochimie 91: 796-803, 2009.
24. Subauste AR, Das AK, Li X, Elliott BG, Evans C, El Azzouny M, Treutelaar M, Oral E, Leff T and Burant CF: Alterations in lipid signaling underlie lipodystrophy secondary to AGPAT2 mutations. Diabetes 61: 2922-2931, 2012.
25. Talukder MM, Sim MF, O'Rahilly S, Edwardson JM and Rochford JJ: Seipin oligomers can interact directly with AGPAT2 and lipin 1, physically scaffolding critical regulators of adipogenesis. Mol Metab 4: 199-209, 2015.
26. Smyth SS, Mueller P, Yang F, Brandon JA and Morris AJ: Arguing the case for the autotaxin-lysophosphatidic acid-lipid phosphate phosphatase 3-signaling nexus in the development and complications of atherosclerosis. Arterioscler Thromb Vasc Biol 34: 479-486, 2014.
27. Chen Y and Li P: Fatty acid metabolism and cancer development. Sci Bull 61: 1473-1479, 2016.
28. Agostini M, Silva SD, Zecchin KG, Coletta RD, Jorge J, Loda M and Graner E: Fatty acid synthase is required for the proliferation of human oral squamous cancer cells. Oral Oncol 40: 728-735, 2004.
29. Fang LY, Wong TY, Chiang WF and Chen YL: Fatty-acid-binding protein 5 promotes cell proliferation and invasion in oral squamous cell carcinoma. J Oral Pathol Med 39: 342-348, 2010.
30. Huang X, Qin J and Lu S: Kanglaite stimulates anticancer immune responses and inhibits HepG2 cell transplantation-induced tumor growth. Mol Med Rep 10: 2153-2159, 2014.
31. Wang J, Wang L, Lin Z, Tao L and Chen M: More efficient induction of antitumor T cell immunity by exosomes from CD40L gene-modified lung tumor cells. Mol Med Rep 9: 125-131, 2014.
32. Shaw RJ, Lowe D, Woolgar JA, Brown JS, Vaughan ED, Evans C, Lewis-Jones H, Hanlon R, Hall GL and Rogers SN: Extracapsular spread in oral squamous cell carcinoma. Head Neck 32: 714-722, 2009.
33. Chen Y, Azman SN, Kerishnan JP, Zain RB, Chen YN, Wong YL and Gopinath SCB: Identification of host-immune response protein candidates in the sera of human oral squamous cell carcinoma patients. PLoS One 9: e109012, 2014.
34. Li X, Yang Y, Zhou F, Zhang Y, Lu H, Jin Q and Gao L: SLC11A1 (NRAMP1) polymorphisms and tuberculosis susceptibility: Updated systematic review and meta-analysis. PLoS One 6: e15831, 2011.
35. George CX, Ramaswami G, Li JB and Samuel CE: Editing of cellular self RNAs by adenosine deaminase ADAR1 suppresses innate immune stress responses. J Biol Chem 291: 6158-6168, 2016.
36. Suchitra MM, Reddy P, Sudhakar GM, Ramesh B, Sambasivaiah K, Bitla AR and Srinivasa RAO Pvln: Evaluation of serum adenosine deaminase as a tumor marker in gastric cancer. Res J Med Med Sci, 2009.
37. Mahajan M, Tiwari N, Sharma R, Kaur S and Singh N: Oxidative stress and its relationship with adenosine deaminase activity in various stages of breast cancer. Indian J Clin Biochem 28: 51-54, 2013.

38. Kelgandre DC, Pathak J, Patel S, Ingale P and Swain N: Adenosine deaminase-a novel diagnostic and prognostic biomarker for oral squamous cell carcinoma. Asian Pac J Cancer Prev 17: 1865-1868, 2016.
39. Fadok VA, Bratton DL, Rose DM, Pearson A, Ezekewitz RA and Henson PM: A receptor for phosphatidylserine-specific clearance of apoptotic cells. Nature 405: 85-90, 2000.
40. Chang B, Chen Y, Zhao Y and Bruick RK: JMJD6 is a histone arginine demethylase. Science 318: 444-447, 2007.
41. Boeckel JN, Guarani V, Koyanagi M, Roexe T, Lengeling A, Schermuly RT, Gellert P, Braun T, Zeiher A and Dimmeler S: Jumonji domain-containing protein 6 (Jmjd6) is required for angiogenic sprouting and regulates splicing of VEGF-receptor 1. Proc Natl Acad Sci USA 108: 3276-3281, 2011.
42. Lee CR: Histone demethylase JMJD6 enhances cancer stem cell phenotype in oral squamous cell carcinoma cells. Dissertations Theses-Gradworks: 57, 2014.
43. Lee CR, Lee SH, Rigas NK, Kim RH, Kang MK, Park NH and Shin KH: Elevated expression of JMJD6 is associated with oral carcinogenesis and maintains cancer stemness properties. Carcinogenesis 37: 119-128, 2016.