Construction of a SPP1/PLAU dual genes containing signature as prognosis risk indicator in Oropharyngeal squamous cell carcinoma

Ziwei Gui  
Second Clinical Medical College of ShanXi Medical University

Juan Du  
Second Hospital of ShanXi Medical University

Lifang Shang  
Second Hospital of ShanXi Medical University

Ningning Shen  
Second Hospital of ShanXi Medical University

Zhiqing Yang  
Second Hospital of ShanXi Medical University

Huijun Yang  
Second Clinical Medical College of ShanXi Medical University

Rong Wei  
Second Hospital of ShanXi Medical University

Wenxia Ma  
Second Hospital of ShanXi Medical University

Yanfeng Chen  
Sun Yat-sen University Cancer Center

Chen Wang  ( wangchen@sxmu.edu.cn )  
Second Hospital of ShanXi Medical University

Research Article

Keywords: Oropharyngeal squamous cell carcinoma (OSCC), SPP1 gene, PLAU gene, bioinformatic analysis, risk score indicator, biomarker

Posted Date: April 10th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2744568/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Abstract

Background

Oropharyngeal squamous cell carcinoma (OSCC) has been a common malignancy in head and neck region. Despite the improved understanding of the cancer development attributing to the revealing of significant epidemiological risk factors, the genetic information of the cancer is still lacking and the patients prognosis remains challenging. The study is to explore the transcriptome data of OSCC and to identify promising cancer development responsible genes thus aiding more precise understanding of the disease and screening promising drug targets for clinical medical therapy.

Methods

Multiple bioinformatic serves were used to interpret the genetic events in OSCC development. Firstly, based on GEO OSCC transcriptome data, the genes with changed expression in cancer comparing to normal oral tissues were identified, followed by being grouped according to the changing level. Then, GO/KEGG interpretation, protein-protein interaction (PPI) network construction and modules analysis were in succession performed to interpret the multiple gene groups for selecting promising hub gene clusters, which were next step proceeded by risk score assessment, Kaplan-Meier survival and Cox Regression analysis to scale down the cluster of candidate genes and select credible prognosis relating key genes. Further, detailed information of the key genes including their physicochemical properties, predicted cellular locations, the expression in human cancers, association with immune cells infiltration, relation with OSCC clinical pathological features and the probable signaling pathways involved in the gene’s regulation on cancer development were explored.

Results

A total of 30054 genes were identified to express abnormally in OSCC cancer versus normal oral epithelium. Of the genes, the expression difference of 607/30054 genes were indicated to be over 8-fold, and further module analysis of the 607 genes highlighted a 33-genes containing module which was supported by SurvExpress risk score assessment to be associated with OSCC survival. Moreover, Kaplan-Meier survival and Cox-regression analysis were performed continually to analyze all the 33 genes one by one, and the result revealed SPP1 and PLAU as two independent prognostic indicators in OSCC development. After the validation of changed expression of SPP1 and PLAU in OSCC versus normal tissues using local hospital biobank samples and exploration of the genes’ association with patients clinical pathological features including the relation with HPV infection, detailed information for instance their physicochemical properties, their expression and variation ratio in human cancers, their relation with immune cells infiltration, as well as the probable signaling pathways involved in the genes’ regulation on OSCC development were explored.
Conclusions

Based on online bioinformatic serves as well as local hospital samples validation, we identified SPP1 and PLAU as two independent prognostic indicators in OSCC and preliminary explored their biological features and clinical significance. Although further experiments and rigorous clinical trials are needed to reveal the genes’ potential drug-target role in clinical medical use, the results shall provide inspiring insights into current understanding of the genetic events in OSCC development and provoke next step deeply exploration of the disease.

Background

Human head and neck region comprises of oral cavity, pharynx and larynx, which parts are all mainly covered with squamous cell epithelium explaining the reason that squamous cell carcinoma (HNSCC) is the major cancer type in the region. HNSCC has been the six most common cancer worldwide[1], and nearly 600 ~ 800 thousands of patients were newly diagnosed each year in the past decade. More threatenly, the incidence keeps increasing and was predicted to rise by 30% (around 1.08 million new cases each year) by 2030[2–4]. Meanwhile, the severe clinical occurence of HNSCC has been proved to be associated with several issues including tobacco consumption, alcohol abuse and virus infection most importantly human papillomavirus (HPV) which closely associated with oropharyngeal squamous cell carcinoma (OSCC) and Epstein Barr virus (EBV) which related with nasopharyngeal squamous cell carcinoma[5].

Based on the close association between HPV infection and OSCC occurence, 2017 AJCC/UICC separately state the cancer staging system of HPV positive and P16 negative (free of HPV infection) OSCC patients partly in concern of the significant different prognosis of two patients groups[6]. Actually, the prognosis of HPV positive patients was reported to be much better than the other group[7, 8]. Over the past several decades a considerable improvement has been observed in the 5 year survival rate of OSCC patients from 55% to near 70%, however, the improvement was mainly focused in HPV positive patients, meanwhile, no notably improvement has been received for HPV negative groups[9–11]. It’s clinically important to explore the genetic events during cancer development thus discovering potential prognosis indicators and promising drug targets in OSCC especially HPV negative patients.

Over the past few decades, multiple genetic alterations and signaling pathways dysfunction with or without direct association with HPV infection have been reported to play critical roles in OSCC development, for instance, a reported comparison between HPV positive and HPV negative tumors revealed that the most frequent gene alterations found in HPV positive cases were TRAF gene loss and E2F1 gene amplification, meanwhile, the HPV negative tumors were more likely enriched in the aberrant loss of CDKN2A and P53 genes[12]. Moreover, other reports also discovered a list of inspiring tumor suppressors that were frequently mutated in OSCC including AT1, NOTCH1, KMT2D, NSD1 and TGFBR2[13–15].
Besides the genetic alterations that characterize cancer development, multiple gene targets that benefit OSCC clinical treatment were also discovered, for instance, EGFR which has been one of the most useful gene target for tyrosine kinase inhibitors was reported to be over expressed in 80~90% whole HNSCC cancers[16, 17] and the molecular targeting of EGFR with monoclonal antibodies such as cetuximab has been a FDA-approved treatment strategy. Besides EGFR[18–20], other receptor tyrosine kinases for instance HER2 and MET also occurs in the cancer and contribute to the molecular targeted therapies[21, 22].

However, the current information are still woefully lacking comparing to the highly heterogeneous and complicated nature of malignant tumor, it is vital to keep identifying new potential prognostic indicators as well as promising drug target-able genes thus aiding more precise understanding of the disease. In modern precise medicine era, the emerging molecular pathological technologies specially next generation sequencing (NGS), gene chips and protein microarrays have been bringing in oceans of high throughput disease data, and a considerable portion of these data are openly accessed to the public, making it more convenient to explore the genetic events in cancer development, analyze potential disease causing gene alterations and identify promising drug targets.

In the study, local hospital biobank OSCC tissues with complete clinical information, online public accessed cancer transcriptome datasets and multiple bioinformaic analysis tools were together used to identify promising prognosis indicators. At start, four openly accessed GEO cDNA profiles were used to screen the abnormally expressed genes in OSCC comparing to normal oral epithelium, and the genes were divided into different groups based on the expression abnormal level. Then, the gene cluster with highest expression difference level (> 8 fold) was purposely picked followed by protein-protein interaction (PPI) network construction and function module analysis. Next step, all the genes in the top scored module was analyzed by Kaplan-Meier survival and Cox regression one by on to screen the survival associated key indicators. Further, after the validation of the key genes’ different expression in cancer versus normal oral tissues by local hospital biobank samples, detailed genes information including their physicochemical properties, predicted cellular locations and expression in broad spectrum human cancers, association with immune cells infiltration and the relation with OSCC clinical pathological features were analyzed. The results shall provide inspiring insights into deeper understanding of OSCC development and benefit unearthing of potential drug targets for further clinical treatment.

**Materials And Methods**

**Data source: OSCC transcriptome profiles from GEO database**

From GEO online database, we widely screened the OSCC related profiles for exploring the abnormally expressed genes in cancer comparing to normal oral epithelium tissues. The profiles selection criteria were set as: 1. profiles were based on human samples (not animal models); 2. tissues were squamous cell carcinoma from oropharyngeal region, covering both cancer and normal oral epithelium tissues; 3.
sample type was solid tissues (not cell lines nor artificial recombinant nucleic acid); 4. profiles data were mRNA/ cDNA/ transcriptome sequencing data; 5. containing at least 20 cases of samples.

Based on above selection criteria, four GEO profiles were selected for further analysis: GSE30784[23], GSE31056[24], GSE78060[25] and GSE37991[26] containing a total of 256 OSCC cancer and 162 normal oral samples (Table S1). Of the four profiles, GSE30784 and GSE 78060 were both based on GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array, containing 167 cancer/45 normal and 26 cancer/4 normal oral tissues respectively. GSE31056 was based on GPL10526 [HG-U133_Plus_2] Affymetrix GeneChip Human Genome HG-U133 Plus 2 Array [Brainarray Version 12] and contains 23 cancer and 73 normal margin samples. Meanwhile, GSE37991 was based on GPL6883 Illumina HumanRef-8 v3.0 expression beadchip and contains 40 cancer and 40 normal oral epithelium tissues.

Datasets processing: select the abnormally expressing genes in OSCC vs normal oral epithelium tissues

To reveal the abnormally expressed genes in OSCC vs normal oral epithelium tissues, the four GEO profiles were analyzed with GEO2R[27] which has been an effective online data analysis tool provided paring with the profiles. The analysis criteria was set as adjusted P value < 0.05, all the genes that meet the criteria would then be divided into 4 groups based on the |log2FC| value as |log2FC|<1, 1≤|log2FC|<2, 2≤|log2FC|<3 and |log2FC|≥3, namely the expression discrepancy level of the 4 group genes was < 2 fold, 2 ~ 4 fold, 4 ~ 8 fold and > 8 fold in OSCC cancer comparing to normal control respectively.

GO/KEGG interpretation of the abnormally expressed genes

To preliminary understand the basic biological attributes of four groups of genes, Gene ontology analysis (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG)[28] which have been effectively used for annotating lists of genes and interpreting involved signaling pathway networks were performed. Differently groups of genes were separately yet simultaneously analyzed the basic biological functions including their cellular locations, main biological processes, molecular functions and mainly enriched signaling pathways. The difference and potential connections between multiple groups of genes were preliminary explored. Given that immunohistochemistry (IHC) has been one of the most economic effective and well accepted clinical tests, for the sake of further potential clinical medical use, we mainly focused on the gene group with high expression change in cancer vs normal tissues, namely the over 8-fold gene cluster for further analysis.

PPI network construction and function module analysis of the high level differently expressed genes

To next step explore the interaction between different genes in the > 8 fold group and identify the potential candidate hub genes that play critical roles in OSCC development, STRING[29], which is short for Search Tool for the Retrieval of Interacting Genes was used to construct the PPI network of the genes. Based on the PPI network of the genes, we further analyzed the promising function modules (gene clusters sharing similar function) of the genes by Molecular Complex Detection (MCODE) plug-in of
Cytoscape3.6.0 software[30] and identified the top gene module with the highest estimated score, and the genes in the module would be taken into further deeper analysis.

**Risk score assessment of the top module gene cluster**

SurvExpress is a newly developed online database which contains more than 20,000 samples covering over 20 cancer types with complete clinical information and patients outcome[31], and it has been effectively used for survival analysis and risk assessment of various gene lists. In the study, SurvExpress was used to preliminary evaluate the risk indicating value of above top module genes in a whole, which would be followed by deep interpretation of each one of the module genes once the evaluation result supported a certain association between the module as a whole and OSCC survival.

**Kaplan-Meier survival combine with Cox regression analysis of the module genes to identify credible prognostic indicators**

Following the preliminary risk assessment of the top module genes as a whole, each gene in the module was orderly brought for univariate survival analysis by Kaplan-Meier plotter (KM)[32] and UALCAN[33] which have been two effective online services for survival evaluation. The genes that were indicated by both KM and UALCAN to be statistical significantly associated with HNSCC survival would be next step processed for multivariate COX regression analysis based on TCGA data. The genes that were supported by all three analysis to be associating with cancer survival would be identified as credible prognostic gene indicators during HNSCC development and processed for further detailed interpretation.

**Basic physicochemical properties analysis of the key gene indicators**

Before exploring the potential regulation of gene indicators on OSCC development, GeneCards[34], ProtParam[35] and ProtScale[36] were combine used to preliminary explore the basic knowledge of identified genes. GeneCards was performed to interpret the reported knowledge of genes for instance their aliases, predicted cellular location, genetic domains as well as currently reported associated disorders. Meanwhile, ProtParam and ProtScale were applied to analyze the gene indicators’ protein structure including their aminoacid composition, molecular weight, estimated protein half life, theoretical isoelectric point, protein instability index, hydrophobicity and hydrophilicity.

**Gene indicators expression in broad spectrum human cancers**

After understanding the basic physicochemical properties of above identified prognostic gene indicators, Gene Expression Profiling Interactive Analysis (GEPIA)[37] as well as UALCAN which have been two effectively used web-based services constructed based on TCGA and GTEx programs were in succession used to analyze the expression change of gene indicators in human cancers comparing to corresponding normal control samples, especially in OSCC versus normal control oral epithelium.
Immunohistochemistry (IHC) experiment validation of gene indicators’ changed expression in OSCC comparing to normal oral epithelium

Regents and tissue samples

To validate the abnormal expression of gene indicators in OSCC comparing to normal tissues, IHC experiment was conducted on 304 local hospital patients samples which were all collected from routine surgeries. Informed consent of potential scientific application of samples have been obtained from patients, and the study was approved by both donor and application hospitals institutional board (Sun Yat-sen University Cancer Center and Second Hospital of ShanXi Medical University, China).

IHC experiment was performed on VENTANA platform (Roche) using local hospital Pathology Department equipment. The primary antibody of gene indicators: anti-SPP1 and anti-PLAU were purchased from OriGene (SPP1 clone ID: OTI2F2, NO.TA806722; PLAU clone ID: OTI5H4, NO. TA805243 respectively). The secondary antibody (Envision /HRP kit) and DAB detection kit were from ZSBG-Bio, other reagents including H2O2, phosphate-buffered saline (PBS), antigen retrieval citrate solution and hematoxylin stain were all from Pathology Department routinely supplied by hospital Supply Department.

IHC experimental protocol

304 cases of OSCC samples (each with adjacent normal control) were used after firstly HE staining confirmation of the disease diagnosis and evaluation of cancer percentage by local hospital pathologists. Then, the wax blocks of each sample was sectioned for IHC experiment following the operating procedure as: firstly being deparaffinized, and then rehydrated using gradient ethanol, followed by being treated with 0.3% H2O2 for inhibiting endogenous peroxidase activity. Further, the slides were boiled in 10mmol/l citrate buffer for antigen retrieval and incubated with primary antibody overnight at 4℃. Moreover, the slides were incubated with secondary antibody at 37℃ for 40 minutes. Finally, The slides were processed with horseradish peroxidase (HRP) staining and DAB visualizing for results evaluation by two local hospital pathologists.

IHC results evaluation

IHC experiment result was evaluated by two experienced pathologists registered in local hospital Pathology Department (Second Hospital of ShanXi Medical University, China). Based on GeneCards analysis of cellular location, both gene indicators SPP1 and PLAU were predicted to express in Golgi apparatus or secreted in extracellular region, so cell membrane and cytoplasm staining was regarded as positive, and 5% was set as the cut off criteria. To be more specific, the cases with ≥ 5% membrane or cytoplasm staining would be scored as positive, and the cases with < 5% or none staining were evaluated as negative.

Association between gene indicators expression and OSCC clinical pathological parameters including HPV infection
To next step evaluate the association between expression of gene indicators with OSCC clinical features, two sources of clinical data were used. Firstly, UALCAN which has been an user-friendly web service constructed based on TCGA data for analyzing the clinical influence of multiple genes on human cancers was used. Besides online database analysis, the clinical information of 304 local hospital OSCC cases which were previously used for IHC experiment was also applied for exploring gene indicators’ clinical significance.

Based on above two sources of genes information and clinical data, the relationship between genes indicators expression and OSCC clinical parameters including the chance of HPV infection was preliminary clarified.

**Construction of a nomogram based on gene indicators and clinical parameters**

To combine evaluate the survival prediction value of gene indicators expression and clinical features in OSCC, a nomogram was constructed using R package on the basis of multivariate analysis. In brief, selected gene indicators and clinical features which were supported by multivariate Cox regression analysis to be associated with patients survival were collected, and then the parameters were processed for R analyzing and constructing a nomogram. In the nomogram, a point scale was used to assign points to each variable, and the total score of each sample was recorded as the sum of all variables points, and the probability of patient survival at 1, 3, and 5 years could be predicted by drawing a vertical line from the total point axis straight downward to the outcome axis based on the nomogram. Further, the predictive accuracy of the model was validated by concordance index (C-index) and receiver operating characteristic (ROC) analysis, which have been two most commonly used methods for evaluating the performance of nomogram models.

**Other genetic alterations of gene indicators in OSCC besides mRNA expression change**

cBioPortal[38] has been one of the largest open access cancer genomics data website worldwide which contains over 100 large-scale tumor research projects covering more than 2,8000 cancer samples. In the research, after understanding the clinical significance of changed mRNA expression of gene indicators in OSCC development, other genetic alterations including gene mutations, amplification and deletion ratio, copy number variation, methylation and phosphorylation ratio of the genes were additionally explored based on cBioPortal database following the website operating instruction, assisting more precise understanding of the gene indicators’ roles in OSCC development.

**Association between gene indicators’ expression and immune cells infiltration**

To preliminary understand the regulation of gene indicators on OSCC development, TIMER[39] which has been an effective web resources for evaluating the association between certain gene expression and
immune cells infiltration was used. Based on TIMER database, the correlation between two gene indicators expression and immune cells including CD4/CD8 + T cell, B cell, macrophage, monocyte, neutrophil, NK cell and cancer associated fibroblast infiltration in human cancers was analyzed, and the correlation in HNSCC was especially focused.

Main biological processes and involved signaling pathways analysis

Besides the association with immune cells infiltration, the biological processes and signaling pathways that were centered on selected gene indicators were also explored based on STRING to reveal the regulation of genes on cancer development. Firstly, the protein-protein interaction network centered on gene indicators was constructed to explore the surrounding genes that were mostly related. Then, GO and KEGG analysis were in succession used to annotate the potential biological processes and signaling pathways the related genes mainly enriched in respectively.

Results

GEO data identified 30054 abnormally expressed genes in OSCC comparing to normal oral epithelium

Four GEO transcriptome profiles were combine used to explore the abnormally expressed genes in OSCC comparing to normal oral epithelium, and 18981, 3932, 5695 and 11506 genes were identified in GSE30784, GSE78060, GSE31056 and GSE37991 respectively. Besides the genes that were shared by different profiles, a total of 30054 genes were eventually revealed to be differently expressed in OSCC vs normal control (Table S2).

The genes were then divided into 4 groups based on the expression discrepancy level, considering the heterogeneity of different samples and the possible data deviation caused by different GEO platforms, the 4 profiles genes were analyzed separately. And the result showed that in GSE30784, the expression change of 15326 genes were < 2 fold, 2686 genes were 2 ~ 4 fold, 664 genes were 4 ~ 8 fold and 305 genes were > 8 fold in OSCC comparing to normal oral epithelium tissues. In GSE78060, the gene number was 1024, 2013, 668 and 227 in each group respectively. In GSE31056, the gene number was 3609, 1362, 444 and 280 in each group. Meanwhile, in GSE37991, the gene number was 9173, 1844, 389 and 100 in < 2 fold, 2 ~ 4fold, 4 ~ 8 fold and > 8 fold group respectively (Fig. 1A-1D).

Above all, besides the genes that were shared in multiple profiles, the analysis of 4 GEO profiles indicated that the expression of 23671, 4285, 1491 and 607 genes were credible revealed to be < 2 fold, 2 ~ 4fold, 4 ~ 8 fold and > 8 fold changed in OSCC comparing to normal oral epithelium tissues (Fig. 1E, 1F).

GO/KEGG preliminary interpretation of the OSCC abnormal expressed genes with different changing level
To preliminary interpret the biological functions of the abnormally expressed genes with different expression changing level, the four groups of genes with expressions changed < 2 fold, 2 ~ 4 fold, 4 ~ 8 fold and > 8 fold were processed with GO/KEGG analysis independently. Interestingly, all four groups of genes in four GEO profiles displayed a same trend that more different the genes expression are in cancer comparing to normal control, their cellular location tend to be more outwards from cell nuclear. To be more specific, the < 2 fold group of genes were mainly located in nuclear, 2 ~ 4 fold genes were mostly in the cytoplasm, meanwhile, the 4 ~ 8 fold and > 8 fold genes were tend to locate in cell membrane and outside in extracellular region (Fig. 2A-2D). Actually, this phenomenon has been observed in many other cancers[40–42], and it’s logical considering the biological fact that most function genes are synthesised in nuclear and regulated by various modular factors, the slight change in nuclear protein might result in massive extra nuclear proteins changes.

**PPI network and function module analysis revealed a 33-genes containing gene cluster**

To next step deeply analyze the abnormally expressed genes in OSCC, the connection among each gene was observed by PPI network, and for the convenience of further immunohistochemistry (IHC) experiment validation and potential clinical use, we mainly focused on the high level expression changed > 8 fold group of genes.

A total of 607 > 8 fold genes were screened by four GEO profiles and the PPI network was constructed based on which three top gene clusters were identified (Fig. 3A). And the first module which posses highest module score contains 33 genes that were mainly related with extracellular matrix structural constituent regulation as well as cell growth and maintenance related biological processes (Fig. 3B). Meanwhile, the next module which score was second to the top module contains 39 genes, and the genes were mostly associated with cellular structural molecular activity and chemokine activity (Fig. 3C). Moreover, the third module contains 9 genes which were mainly enriched in catalytic activity and energy metabolism related biological processes (Fig. 3D).

Given the top module possess the highest score and the genes were mostly extracellular matrix constituent regulation related which is consistent with previous GO/KEGG analysis that high level gene clusters tend to locate distal from cell nuclear, we mainly focused on the 33 genes in the first module for further analysis.

**Risk score assessment of the 33-genes containing module**

Before deeply interpretation of each of the 33 genes in first module, SurvExpress was in advance used to evaluate the prognostic indicating value of the module as a whole, and the result supported that based on the module genes expression, OSCC patients could be significantly divided into two groups with high and low prognostic risks respectively. The prediction concordance index equals 66.86, risk group hazard ratio equals 2.71 between high and low survival patients (Fig. 4A-4D). Moreover, the relative expression of 33 genes expression in low and high risk patient groups were also explored (Fig. 4E).
**KM survival combine with Cox regression analysis revealed SPP1 and PLAU as two credible OSCC prognostic indicators**

To further scale down the candidate genes and identify credible key genes during cancer development, KM survival which is based on GEO, EGA and TCGA data, UALCAN database which is based on TCGA, MET500 and CPTAC data and multivariate cox regression analysis were combine used to select credible survival related key genes from the 33 module genes. And the results showed that 7 of 33 genes were supported by both KM analysis and Ualcan data to be associated with OSCC survival, namely SPP1, SERPINH1, PLAU, THBS1, CXCL13, CSF2 and CXCL1. Further, the 7 genes were processed for multivariate Cox regression analysis which result indicated that SPP1 and PLAU genes were the only two genes could be regarded as independent prognostic indicators in OSCC development (Table 1), and both genes relate with not only patients overall survival but also progress free survival (Fig. 5A-5D).

<table>
<thead>
<tr>
<th>Module Genes</th>
<th>Survival analysis (p value)</th>
<th>KM survival</th>
<th>UALCAN survival</th>
<th>COX Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPP1</td>
<td>0.0039</td>
<td>0.028</td>
<td>0.048*</td>
<td></td>
</tr>
<tr>
<td>SERPINH1</td>
<td>0.00014</td>
<td>0.142</td>
<td>0.559</td>
<td></td>
</tr>
<tr>
<td>PLAU</td>
<td>0.00005</td>
<td>0.001</td>
<td>0.021*</td>
<td></td>
</tr>
<tr>
<td>THBS1</td>
<td>0.0014</td>
<td>0.009</td>
<td>0.122</td>
<td></td>
</tr>
<tr>
<td>CXCL13</td>
<td>0.00045</td>
<td>0.397</td>
<td>0.275</td>
<td></td>
</tr>
<tr>
<td>CSF2</td>
<td>0.0058</td>
<td>0.136</td>
<td>0.769</td>
<td></td>
</tr>
<tr>
<td>CXCL1</td>
<td>0.0035</td>
<td>0.238</td>
<td>0.475</td>
<td></td>
</tr>
</tbody>
</table>

**Basic physicochemical properties of SPP1 and PLAU genes**

ProtParam, ProtScale and GeneCards analysis were successively used to interpret the physiochemical properties of SPP1 and PLAU genes. The results indicated that SPP1 which is short for Secreted Phosphoprotein 1, locates in 4q22.1 and encodes a protein composed of 314 amino acids including 75 negatively charged amino acid residues (ASP + Glu) and 29 positively charged amino acid residues (Arg + Lys), the estimated molecular weight of the protein is 35.4KD with theoretical isoelectric point computed as 4.37. Moreover, the estimated instability index of the protein is 53.37 and grand average of hydrophobic value is -1.10 indicating SPP1 works as a cellular unstable and hydrophilic protein which is consistent with the ProtScale analysis result of SPP1 structure showing that the protein harbors more hydrophilic regions than hydrophobicity regions (Fig. 5F). Meanwhile, the prediction result of GeneCards
indicated that SPP1 locates in cellular Golgi apparatus or to be secreted in the extracellular region (Fig. 5E).

As for PLAU gene whose full name is Urokinase-type plasminogen activator, the gene locates in 10q22.2 and encodes a protein composed of 431 amino acids including 39 negatively charged amino acid residues (ASP + Glu) and 52 positively charged amino acid residues (Arg + Lys), the estimated molecular weight of the protein is 48.5KD with theoretical isoelectric point computed as 8.78. The estimated instability index of the protein is 42.71 and grand average of hydrophobic value is -0.470 indicating SPP1 also works as a cellular unstable and hydrophilic protein (Fig. 5H). Similar to SPP1 gene, the prediction of GeneCards also indicated that PLAU locates in cellular Golgi apparatus or to be secreted in extracellular region (Fig. 5G).

Aberrant changed expression of SPP1 and PLAU in human cancers including OSCC

To validate the expression change of SPP1 and PLAU genes in OSCC comparing to normal oral epithelium, both online database and local hospital samples were used. As for SPP1 gene, online GEPIA analysis revealed that although the gene expression various in broad spectrum human cancers, for instance, the expression was lower in sarcoma comparing to control samples, its expression in most other types of tumors including HNSCC was significantly higher than matched normal tissues, especially in glioblastoma multiforme and cholangio carcinoma, the expression was more than 50 times higher in cancer comparing to corresponding normal control samples (Fig. 6A), and in HNSCC, SPP1 expression was nearly 20 times higher in cancer versus normal oral epithelium (Fig. 6B). Meanwhile, as for PLAU gene, a similar trend was observed that although the expression various (Fig. 6C), it expresses almost 10 times significantly higher in HNSCC comparing to normal oral epithelium samples (Fig. 6D).

Besides online GEPIA analysis which was based on TCGA data, the result of IHC experiment which was conducted using 304 local hospital patients samples (Table S3) also validated the aberrant gain of expression of SPP1 and PLAU in OSCC. IHC experiment result showed that the positive ratio of both SPP1 and PLAU were much higher in cancer (69.7% and 54.8% respectively) than matched normal oral epithelium samples (both less than 5%), supporting the aberrant gain of expression of SPP1 and PLAU in OSCC (Fig. 6E-6H).

Association between SPP1 and PLAU expression with OSCC clinical pathological parameters

To access the association between SPP1 and PLAU expression with OSCC clinical parameters, both TCGA online data and local hospital patients information were used. Firstly, UALCAN information which is based on TCGA data revealed that not only SPP1 and PLAU genes express markedly higher in cancer comparing to normal control, but also both genes expression keep increasing as the cancer stage and grade advancing. More interestingly, both genes expression were associated with HPV infection, as a matter of fact, the genes tend to express higher in patients with no HPV infection indicating their potential
roles as gene targets in HPV(-) OSCC patients. Additionally, PLAU expression tends be higher in patients with TP53 mutation. Meanwhile, no significance relationship was found between neither SPP1 nor PLAU expression with patients age, gender, race and nodal metastasis (Fig. 7A-7P).

Meanwhile, the information of 304 local hospital patients which were previously used for IHC experiment revealed that besides the keeping increasing genes expression as the cancer stages advancing which is consistent with above UALCAN result, the genes expression were also associated with cancer differentiation as well as chemo and radiotherapy history. Moreover, no association was found between neither genes expression and patients smoking, drinking, tobacco chewing habit nor with tumor location, intravascular tumor thrombus existence and tumor nerve invasion (Table 2, 3).
Table 2
The association between SPP1 and OSCC clinical pathological features

<table>
<thead>
<tr>
<th>Parameters</th>
<th>OSCC(%)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- (%)</td>
<td>+ (%)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>34 (33.0%)</td>
<td>69 (67.0%)</td>
</tr>
<tr>
<td>male</td>
<td>54 (27.0%)</td>
<td>146 (73.0%)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 60</td>
<td>62 (29.1%)</td>
<td>151 (70.9%)</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>26 (28.9%)</td>
<td>64 (71.1%)</td>
</tr>
<tr>
<td>Smoking habit</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>63 (31.3%)</td>
<td>138 (68.7%)</td>
</tr>
<tr>
<td>Yes</td>
<td>25 (24.5%)</td>
<td>77 (75.5%)</td>
</tr>
<tr>
<td>Drinking alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>76 (31.1%)</td>
<td>168 (68.9%)</td>
</tr>
<tr>
<td>Yes</td>
<td>12 (20.3%)</td>
<td>47 (79.7%)</td>
</tr>
<tr>
<td>Chewing tobacco</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>83 (29.1%)</td>
<td>202 (70.9%)</td>
</tr>
<tr>
<td>Yes</td>
<td>5 (27.8%)</td>
<td>13 (72.2%)</td>
</tr>
<tr>
<td>Tumor site</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left</td>
<td>42 (30.2%)</td>
<td>97 (69.8%)</td>
</tr>
<tr>
<td>Right</td>
<td>43 (29.5%)</td>
<td>103 (70.5%)</td>
</tr>
<tr>
<td>Radiation history</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>85 (32.3%)</td>
<td>180 (67.9%)</td>
</tr>
<tr>
<td>Yes</td>
<td>3 (7.9%)</td>
<td>15 (83.3%)</td>
</tr>
<tr>
<td>Intravascular tumor cells</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>84 (29.8%)</td>
<td>198 (70.2%)</td>
</tr>
<tr>
<td>Yes</td>
<td>4 (19.0%)</td>
<td>17 (81.0%)</td>
</tr>
<tr>
<td>Nerve affection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameters</td>
<td>OSCC(%)</td>
<td>P Value</td>
</tr>
<tr>
<td>-----------------------------</td>
<td>---------------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>- (%)</td>
<td>+ (%)</td>
</tr>
<tr>
<td>No</td>
<td>64 (31.5%)</td>
<td>139 (68.5%)</td>
</tr>
<tr>
<td>Yes</td>
<td>24 (24.0%)</td>
<td>76 (76.0%)</td>
</tr>
<tr>
<td>Lymph node infiltration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>86 (29.7%)</td>
<td>204 (70.3%)</td>
</tr>
<tr>
<td>Yes</td>
<td>2 (15.4%)</td>
<td>11 (84.6%)</td>
</tr>
<tr>
<td>Differentiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well</td>
<td>32 (21.3%)</td>
<td>118 (78.7%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>39 (32.5%)</td>
<td>81 (67.5%)</td>
</tr>
<tr>
<td>Poor</td>
<td>14 (50.0%)</td>
<td>14 (50.0%)</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cT1</td>
<td>62 (41.6%)</td>
<td>87 (58.4%)</td>
</tr>
<tr>
<td>cT2</td>
<td>20 (18.7%)</td>
<td>87 (81.3%)</td>
</tr>
<tr>
<td>cT3</td>
<td>3 (8.8%)</td>
<td>31 (91.2%)</td>
</tr>
<tr>
<td>cT4</td>
<td>1 (12.5%)</td>
<td>7 (87.5%)</td>
</tr>
</tbody>
</table>
Table 3  
The association between PLAU and OSCC clinical pathological features

<table>
<thead>
<tr>
<th>Parameters</th>
<th>OSCC(%)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>- (%)</td>
<td>+ (%)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>30(29.1%)</td>
<td>73(70.9%)</td>
</tr>
<tr>
<td>male</td>
<td>58(29.0%)</td>
<td>142(71%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤ 60</td>
<td>57(26.8%)</td>
<td>156(73.2%)</td>
</tr>
<tr>
<td>&gt; 60</td>
<td>31(34.4%)</td>
<td>59(65.6%)</td>
</tr>
<tr>
<td><strong>Smoking habit</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>64(31.8%)</td>
<td>137(68.2%)</td>
</tr>
<tr>
<td>Yes</td>
<td>24(23.5%)</td>
<td>78(76.5%)</td>
</tr>
<tr>
<td><strong>Drinking alcohol</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>74(30.3%)</td>
<td>170(69.7%)</td>
</tr>
<tr>
<td>Yes</td>
<td>14(23.7%)</td>
<td>45(76.3%)</td>
</tr>
<tr>
<td><strong>Chewing tobacco</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>85(29.8%)</td>
<td>200(70.2%)</td>
</tr>
<tr>
<td>Yes</td>
<td>3(16.7%)</td>
<td>15(83.3%)</td>
</tr>
<tr>
<td><strong>Tumor site</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left</td>
<td>46(33.1%)</td>
<td>93(66.9%)</td>
</tr>
<tr>
<td>Right</td>
<td>42(27.4%)</td>
<td>122(72.6%)</td>
</tr>
<tr>
<td><strong>Radiation history</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>80(30.2%)</td>
<td>185(69.8%)</td>
</tr>
<tr>
<td>Yes</td>
<td>8(21.1%)</td>
<td>30(78.9%)</td>
</tr>
<tr>
<td><strong>Intravascular tumor cells</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>79(28.0%)</td>
<td>203(72.0%)</td>
</tr>
<tr>
<td>Yes</td>
<td>9(42.9%)</td>
<td>12(57.1%)</td>
</tr>
<tr>
<td><strong>Nerve affection</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameters</td>
<td>OSCC(%)</td>
<td>P Value</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td></td>
<td>- (%)</td>
<td>+ (%)</td>
</tr>
<tr>
<td>No</td>
<td>57(28.1%)</td>
<td>146(71.9%)</td>
</tr>
<tr>
<td>Yes</td>
<td>31(31.0%)</td>
<td>69(69.0%)</td>
</tr>
<tr>
<td>Lymph node infiltration</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>84(29.0%)</td>
<td>206(71.0%)</td>
</tr>
<tr>
<td>Yes</td>
<td>4(30.8%)</td>
<td>9(69.2%)</td>
</tr>
<tr>
<td>Differentiation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Well</td>
<td>33(22.0%)</td>
<td>117(78.0%)</td>
</tr>
<tr>
<td>Moderate</td>
<td>37(30.8%)</td>
<td>83(69.2%)</td>
</tr>
<tr>
<td>Poor</td>
<td>15(53.6%)</td>
<td>13(46.4%)</td>
</tr>
<tr>
<td>Stage</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cT1</td>
<td>57(38.3%)</td>
<td>92(61.7%)</td>
</tr>
<tr>
<td>cT2</td>
<td>25(23.4%)</td>
<td>82(76.6%)</td>
</tr>
<tr>
<td>cT3</td>
<td>3(8.8%)</td>
<td>31(91.2%)</td>
</tr>
<tr>
<td>cT4</td>
<td>1(12.5%)</td>
<td>7(87.5%)</td>
</tr>
</tbody>
</table>

Construction of a nomogram based on SPP1 and PLAU expression as well as OSCC clinical parameters

For the clinical convenience of patients survival prediction, a nomogram combining gene indicators and OSCC clinical features was constructed. Firstly, based on multivariate cox regression analysis, three clinical parameters including patients age, extracapsular invasion and radiation therapy history were identified to be independent risk indicators in OSCC development (Table 4).
Table 4
Association between OSCC key parameters with patients overall survival

<table>
<thead>
<tr>
<th>OSCC parameters</th>
<th>P Value</th>
<th>B value</th>
<th>HR</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Univariate analysis</td>
<td>Multivariate analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SPP1 expression</td>
<td>0.001</td>
<td>0.021*</td>
<td>0.435</td>
<td>1.545</td>
</tr>
<tr>
<td>PLAU expression</td>
<td>0.028</td>
<td>0.048*</td>
<td>0.319</td>
<td>1.376</td>
</tr>
<tr>
<td>Age</td>
<td>0.014</td>
<td>0.007*</td>
<td>0.678</td>
<td>1.970</td>
</tr>
<tr>
<td>Extracapsular</td>
<td>0.001</td>
<td>0.001*</td>
<td>0.965</td>
<td>2.624</td>
</tr>
<tr>
<td>Perineural</td>
<td>0.002</td>
<td>0.101</td>
<td>0.407</td>
<td>1.502</td>
</tr>
<tr>
<td>Radiations</td>
<td>0.002</td>
<td>&lt; 0.001*</td>
<td>−0.963</td>
<td>0.382</td>
</tr>
<tr>
<td>Grade</td>
<td>0.043</td>
<td>0.800</td>
<td>0.056</td>
<td>1.057</td>
</tr>
<tr>
<td>Stage</td>
<td>0.012</td>
<td>0.043*</td>
<td>0.354</td>
<td>1.424</td>
</tr>
</tbody>
</table>

* Represents the parameters that were supported by both univariate and multivariate survival analysis to be independent survival indicators.

And as for the gene indicators, besides the previous observation that both SPP1 and PLAU genes associated with OSCC survival, great utility of the two genes were found in potential further OSCC clinical medical use. Firstly, both online data and IHC experiment showed no obvious correlation between the two genes expression indicating they regulate cancer development via independent and most likely different signaling pathways (Fig. 8B). And what's inspiring is that based on the two genes expression, OSCC patients survival could be divided into four groups: the patients with both SPP1 and PLAU high expression harbor the worst survival, next the patients with either SPP1 or PLAU gene high expression and the other gene low, meanwhile, the survival of patients with both genes low expression was the best in four groups of patients, supporting the potential combine use of two genes as survival indicators in clinical medical treatment (Fig. 8C).

Moreover, a genes signature was at start constructed based on the two genes expression: genes score = 0.0405*SPP1 expression + 0.179*PLAU expression, and the signature was calculated by LASSO analysis to be the best genes equation of the two genes which posses highest prognosis prediction value among other equations (Fig. 8D-8G). However, considering the inconvenience that a complicate genes calculation might caused to clinical medical use, we compared the gene signature with an easier signature that: gene score = median SPP1 expression + median PLAU expression, surprisingly, only little difference was found between the two gene signatures indicating the potential utility of median SPP1 + median PLAU which will be of much more convenience in further clinical medical use (Fig. 8H, 8I).

Further, to combine evaluate the survival prediction value of gene indicators and clinical features in OSCC, a nomogram was constructed. And in the nomogram, a point scale was assigned for each of
above variables, and the sum of the all the variable points equal the final score of each patient, and the 1, 3 and 5 years of survival could be predicted by drawing a vertical line from the total point axis straight downward to the outcome axis. Since for different patients the points of multiple variables vary, the nomogram could be used to predict the survival risk of different OSCC individuals (Fig. 8A). For example, a OSCC patient who is younger than 65 years old, without extracapsular invasion and no radiation therapy history, meanwhile both of the SPP1 and PLAU gene were high expressed, the total points of the patient would be scored as 58 + 40 + 0 + 0 = 98, then the potential of 1-year survival was less than 20%, 3 years survival rate was much less that 10% and 5 years survival probability < 5%.

Other genetic alterations of SPP1 and PLAU in OSCC

Besides mRNA expression, other types of gene alterations including mutation ratio, amplification, deletion and protein structure variant were also preliminary explores based on cBioPortal database. The results revealed that as for both SPP1 and PLAU genes, multiple types of gene alterations exist in human tumors, a certain percent of gene mutation, deletion and amplification occurs in various tumors. In HNSCC, besides gene amplification, a certain percent of cases were diagnosed with SPP1 or PLAU gene mutations, mostly missense mutations and nonsense mutations including SPP1(E282*) nonsense mutation as well as SPP1(A213V), PLAU(R130Q) and PLAU(S376L) missense mutations (Fig. 9A, 9B).

Association between SPP1 and PLAU expression with different types of immune cells infiltration

Immune cells infiltration has been a characteristic feature of tumor microenvironment, which not only relates with the ability of cancer initiation, progression and metastasis, but also associates with the effect of immune targeting therapy. Given that SPP1 and PLAU were supported by IMMPORT data which has been a worldwide commonly used immune analyzing database to be associated with cancer immune regulation, TIMER and TISIDB database were combine used to evaluate the potential association between SPP1 and PLAU expression with OSCC immune cell infiltration. And the result revealed that mild positive association was found between SPP1 expression and cancer associated fibroblasts as well as macrophages infiltration, meanwhile, slight negative association was discovered between the gene expression and B cells infiltration (Fig. 9C, 9D). As for PLAU gene, moderate association was found between gene expression and myeloid dendritic cells, CD4 + T cells as well as cancer associated fibroblasts infiltration (Fig. 9E, 9F).

What's worth of emphasizing is that besides the association between genes expression with immune cells infiltration, SPP1 and PLAU expression was also preliminary revealed to be different in multiple OSCC immune subtypes. Based on TISIDB analysis result, SPP1 expresses the lowest in inflammatory subtype and highest in lymphocyte depleted subtype. Meanwhile, the expression of PLAU was the lowest in lymphocyte depleted subtype and highest in IFN-gamma dominant type (data not shown) indicating their potential roles in OSCC immune microenvironment regulation, although the mechanism remains to be deeper researched.
SPP1 and PLAU potentially involved biological processes and signaling pathways

To preliminary explore the potential biological functions of SPP1 and PLAU genes in OSCC and the probable signaling pathways involved, STRING was used to construct the SPP1 and PLAU centering PPI network respectively for revealing the surrounding potentially connecting genes followed by GO and KEGG analyzing the probable signaling pathways these genes enriched in. And GO results showed that the biological processes SPP1 gene participated in were mainly focused on extracellular matrix organization and cell matrix adhesion, and KEGG analysis revealed the signaling pathways SPP1 gene involved were mostly ECM-receptor interaction, PI3K-AKT signaling pathway and human papillomavirus infection related signaling pathways (Fig. 10A, 10B).

Meanwhile, the biological processes that PLAU gene participated in were mainly focused on cell migration and protein metabolic process regulation, and the signaling pathways PLAU gene involved were mostly proteoglycans metabolism, ErbB signaling pathway and human papillomavirus infection related signaling pathways (Fig. 10C, 10D).

Although deeper analysis are needed to validate the association between SPP1 and PLAU with above signaling pathways, current results shall provide promising directions for further exploring the mechanism behind the genes regulation on OSCC development. It's of encouraging insight to better understand the genetic events during OSCC development and discovering promising disease indicators and probable drug targets.

Discussion

HNSCC has been the six most common malignancy worldwide, and a big part of the cancer is OSCC which is different from the other head and neck regions cancers including nasopharyngeal squamous cell carcinoma considering its close association with HPV infection. Although the reason for cancer occurrence has been gradually clear since the discovery of its close relation with unhealthy living habit for instance obacco consumption, alcohol abuse and HPV infection partly caused by penilingus, no notable survival improvement has been received especially for HPV negative OSCC patients[5]. It's of clinical realistic significance to keep exploring the genetic information of OSCC thus screening potential prognosis indicators and promising drug targets in the cancer especially HPV negative patients. In the study, based on multiple public online datasets as well as local hospital biobank patients samples, we identified two genes: SPP1 and PLAU which were credible related with OSCC patients survival, especially the genes expression were much higher in HPV negative patients than the others indicating their potential clinical medical value.

During the discovery of the two genes, firstly, four OSCC transcriptome GEO profiles namely GSE30784, GSE31056, GSE78060 and GSE37991 containing a total of 256 OSCC cancer and 162 normal oral samples were used to analyze the aberrant differently expressed genes in OSCC cancer comparing to
normal control samples. And the result identified a total of 30054 genes including 23671 genes with expression change < 2 fold, 4285 genes 2 ~ 4 fold, 1491 genes 4 ~ 8 fold and 607 genes whose expression discrepancy was > 8 fold in OSCC versus normal oral epithelium.

Further, GO/KEGG was performed to preliminary interpret the basic biological features of the four groups of genes which result showed a specific phenomenon that the more the genes expression were changed in cancer versus normal control samples, the more they tend to locate distal from cell nuclear. To be specific, the < 2 fold and 2 ~ 4 fold groups of genes were shown to focus mainly in nuclear and cytoplasm, 4 ~ 8 fold genes were displayed to be locating mainly in cytoplasm and cell membrane, meanwhile, > 8 fold fold genes tend to be mostly locating on the cellular membrane or out in extracellular region. Actually, similar phenomenon has been discovered in many others types of cancers we recently studied, and we do think the phenomenon logical considering the biological fact that most human functional proteins were synthesized in nuclear abide by the classic “central dogma” DNA-RNA-protein direction, slight change in nuclear proteins who potentially work as transcription regulatory factors shall result in massive extracellular locating proteins’ expression change.

To next step explore the relationship among the abnormally expressed genes, the PPI network was constructed. Considering the feasibility of further clinical validation using IHC experiment, the highest level expression changed genes were mainly focused. As a result, the PPI network of 607 genes whose expression change were > 8 fold revealed four promising gene modules involving multiple different genes and various signaling pathways. Of the four modules, the top module with highest computed module score and contains 33 genes was focused for further analysis after SurvExpress validating the prognosis prediction value of the module as a whole.

To further scale down candidate genes and identify credible prognosis related key genes during OSCC development, KM survival, UALCAN survival and multivariate Cox Regression analysis were combine used to evaluate the association between each of the 33 genes with OSCC survival, and the result highlighted two genes: SPP1 and PLAU, which were supported by all three analysis to be associated with OSCC patients survival and worked potentially as independent prognostic indicators in cancer development. Although SPP1 and PLAU have been separately reported to be participated in HNSCC development previously, detailed information about the two genes are still lacking and their potential collaboration are worth of deeply exploring.

SPP1, which is short for Secreted Phosphoprotein 1, locates in 4q22.1 and encodes a cellular unstable and hydrophilic protein composed of 314 amino acids weighting 35.4KD, and based on the computed physicochemical parameters of the protein, the theoretical isoelectric point is 4.37 and estimated half-time is 30h in mammal cells. Actually, SPP1 has been characterized by UniProtKA as a cytokine that involved in enhancing production of IFN-γ and IL12 as well as reducing production of IL10, and the gene potentially play a role in the regulation of type I immunity. Moreover, SPP1 has been reported to be associated with several diseases including pediatric systemic lupus erythematosus, nephrolithiasis and calcium oxalate[43]. In the study, we mainly focused on its potential regulation on OSCC development.
Meanwhile, as for PLAU, which is short for plasminogen activator urokinase, locates in 10q22.2 and encodes a golgi apparatus and extracellular region locating protein. PLAU protein is composed of 431 amino acids weighting 48.5KD with estimated theoretical isoelectric point as 8.78. According to GeneCards interpretation of the gene, PLAU is a secreted serine protease that specifically cleaves the zymogen plasminogen to form active enzyme plasmin, and the gene has been reported to be associated with multiple diseases for instance quebec platelet disorder, alzheimer disease as well as HNSCC participating in several signaling pathways including FGF signaling pathway and DNA damage response related signaling pathway[44, 45].

Both SPP1 and PLAU have been validated by online datasets as well as local hospital patients samples to be aberrant up regulated in OSCC comparing to normal oral epithelium. An inspiring fact is that although no obvious relation was found between SPP1 and PLAU expression, a combine effect was observed that based on the two genes expression, OSCC patients can be divided into 4 groups, the patients with neither SPP1 nor PLAU expression shows statistically significantly better prognosis than the patients with either one gene expression, and the survival of patients with both genes expression was the worst in 4 groups, indicating the high potential of combine use of the two genes in further clinical medical.

Moreover, based on TCGA data and 304 local hospital patients information, the association between genes expression and OSCC clinical parameters were evaluated which result revealed that both SPP1 and PLAU were related with cancer stage and grade, the genes expression increases as the cancer stage and grade advancing. An inspiring discovery is that both genes were associated with HPV infection, as a matter of fact, the genes tend to express higher in patients with no HPV infection indicating their potential roles as gene targets in HPV(-) OSCC patients. Further, a nomogram was constructed integrating two genes expression and clinical parameters which were supported by multivariate cox regression to be associated with patients survival to individualize the 1 year, 3 years and 5 years survival predication of each patient.

Further, other types of gene variations of SPP1 and PLAU were analyzed, and the result revealed that besides mRNA expression, occasional gene mutation and deletion were detected in OSCC samples. Meanwhile, moderate association was found between genes expression and certain immune cells infiltration including B cells, macrophages, dendritic cells and cancer associated fibroblasts indicating their potential value in cancer immune microenvironment regulation.

Additionally, to preliminary explore the potential mechanism of SPP1 and PLAU regulation on OSCC development, both genes centered PPI networks were constructed respectively to explore the potential signaling pathways SPP1 and PLAU as well as surrounding genes involved. And the result revealed that the biological processes SPP1 involved were mostly extracellular matrix organization and cell matrix adhesion related, and the signaling pathways were mostly ECM-receptor interaction, PI3K-AKT signaling and human papillomavirus infection related signaling pathways. Meanwhile, the biological processes that PLAU gene participated in were mainly focused on cell migration and protein metabolic process.
regulation, and the signaling pathways were mostly proteoglycans metabolism, ErbB signaling pathway and human papillomavirus infection related signaling pathways.

Although the current result is not yet enough to classify SPP1 and PLAU as new useful clinical drug targets for OSCC HPV(-) patients, comprehensive studies and clinical trials are needed to confirm the findings before promoting the clinical utility of the genes in clinical treatment, the results shall provide meaningful insight into better understanding of the disease.

**Conclusion**

In conclusion, based on four GEO transcriptome profiles, we identified 30054 genes that were abnormally expressed in OSCC comparing to normal oral samples, and further series of bioinformatic analysis of the genes highlighted 2 genes: SPP1 and PLAU as credible prognostic indicators in cancer development. Both openly accessed online data analysis and local hospital IHC experiment validated the aberrant up regulation of the two genes in OSCC. All three of KM survival, UALCAN data and cox regression analysis supported the survival predication value of the two genes. Moreover, basic physiochemical properties, other types variations besides mRNA expression, association with immune cells infiltration and the genes centered biological processes as well as involved signaling pathways were preliminary explored. Above results shall provide inspiring insights into better understanding the molecular mechanism behind OSCC development.

**Abbreviations**

HNSCC  Head and Neck squamous cell carcinoma  
OSCC  Oropharyngeal squamous cell carcinoma  
SPP1  Secreted Phosphoprotein 1  
PLAU  Plasminogen activator urokinase  
GEO  Gene Expression Omnibus  
GO  Gene ontology  
KEGG  Kyoto Encyclopedia of Gene and Genome  
PPI  Protein-protein interaction network  
HPV  Human papillomavirus  
IHC  Immunochemistry experiment  
KM  Kaplan-Meier
OS    overall survival rate
RFS    progress free survival rate
GEPIA  Gene Expression Profiling Interactive Analysis

**Declarations**

**Ethnic approval and consent to participate**

All of the local hospital OSCC patients samples used in IHC experiment were originated from Sun Yat-sen University Cancer Center and shipped to Second Hospital of ShanXi Medical University for further collaborative analysis. Informed consent of the potential scientific application of the surgery samples have been obtained from patients at the same time they donated the samples to the hospital, and the paper agreement signed with the donors’ signatures were kept by the Sun Yat-sen University Cancer Center. And the OSCC samples being used in this research was approval by both Sun Yat-sen University Cancer Center and Second Hospital of ShanXi Medical University Institutional Board. All methods were carried out in accordance with relevant guidelines and regulations or declaration of Helsinki.

**Consent for publication**

Not applicable

**Availability of data and materials**

Publicly available datasets were analyzed in this study. The data can be found here:

And 304 cases of local hospital patients samples were also included in the manuscript, detailed information of these samples were listed in Table S3.

All data generated or analyzed based on above datasets and other experiments during this study are included in this published article.

**Competing interests**

All of the authors agreed the publication of the paper and declare no conflicts of interests.

**Funding**
The work was supported by the China central government funds for guiding local scientific and technological development (YDZJSX2021A042), the fund of Shanxi Medical key scientific research project (2021XM34), and two funds from ShanXi province scientific research projects in China (201901D211498 and 2023103).

Authors' contributions

ZG, JD and LS designed the study and drafted the manuscript, contributed equally to the whole study. NS and ZY performed the data collecting and analysis. HY and RW participated in data interpretation, performing experiments and study design, WM, YC and CW were involved in the drafting and critical revision of manuscript. As the corresponding author, all WM, YC and CW have full access to all data of the manuscript, CW made the eventual decision to submit the article for publication. All authors read and approved the final manuscript.

Acknowledgements

We sincerely thank the researchers for providing their GEO data online, and honestly appreciate the hospital patients for donating their samples to hospital Biobank and agreeing to the scientific use of their donations, it is our honest honor to acknowledge their contributions.

References


**Figures**
Figure 1

The differently expressed genes in OSCC vs normal oral epithelium identified based on GEO datasets

From GEO datasets (A) GSE30784, (B) GSE31056, (C) GSE78060 and (D) GSE37991, the up-regulated (right-sided) and down-regulated (left sided) aberrant differently expressed genes in OSCC comparing to normal control samples were identified and classified into four groups based on expression difference
level as: <2 fold genes (black-colored spots), 2~4 fold genes (yellow-colored spots), 4~8 fold genes (green-colored spots) and >8 fold genes (red-colored spots), and the blue-colored spots represent the leftover genes with p value>0.05. (E) The intersection of all the differently expressed genes in four GEO profiles. (F) The intersection of >8 fold group of genes in four GEO profiles, the genes that were > 8 fold differently expressed were mainly focused for deeper analysis.

Figure 2

GO/KEGG interpretation of the abnormally expressed genes in OSCC

The cellular components the abnormally expressed genes mainly enriched in (A) GSE30784, (B) GSE31056, (C) GSE78060 and (D) GSE37991 respectively.
Figure 3

Construction of the PPI network of >8 fold genes and function gene modules analysis

(A) The PPI network of 607 >8 fold differently expressed genes was constructed based on which three genes modules were identified (three red circles and each represents one gene module). (B, C, D) The diagrammatic sketch (left diagram) and the detailed information (right table) including the module score,
module description and detailed involving genes of three main modules in the PPI network. (*The top module with the highest module score was mainly focused for further analysis).

Figure 4
Risk score assessment of top module genes as a whole
The Risk score assessment of 33-genes containing top module which result separates the OSCC patients into two groups with high and low survival risk respectively. (A, B) Risk score assessment displayed according to the prognostic index of each patient. (C) Relative expression of 33 modules genes in high and low risk patients groups respectively displayed in heatmap. (D) Survival analysis of the OSCC patients grouped by module risk score (high risk group of patients versus low risk group). (E) Relative expression of 33 module genes in high and low risk groups respectively displayed in histogram.
Figure 5

Survival and basic physicochemical properties analysis of SPP1 and PLAU genes

The overall survival and (B) progress free survival of SPP1 gene in OSCC. (C) The overall survival and (D) progress free survival of PLAU gene in OSCC. (E) The predicted cellular location of SPP1 protein in human cells. (F) The hydrophilcity / hydrophobicity analysis of SPP1 protein. (G) The predicted cellular location of PLAU protein in human cells. (H) The hydrophilcity / hydrophobicity analysis of PLAU protein.
**Figure 6**

UALCAN prediction of SPP1 gene expression in broad spectrum human cancers. (B) GEPIA prediction of SPP1 expression in HNSCC comparing to normal oral epithelium. (C) UALCAN prediction of PLAU gene expression in broad spectrum human cancers. (D) GEPIA prediction of PLAU expression in HNSCC comparing to normal oral epithelium. (E) HE staining and (F) IHC experiment using local hospital patients samples to reveal the aberrant gain of expression of SPP1 in OSCC comparing to normal oral squamous
Abnormal SPP1 and PLAU expressions in OSCC comparing to normal oral epithelium samples
**Figure 7**

The association between SPP1 and PLAU genes expression with OSCC clinical parameters

(A) Relative SPP1 expression in OSCC versus normal oral squamous cell epithelium. And the association between SPP1 expression and OSCC (B) patients' age, (C) gender, (D) cancer grade, (E) cancer stage, (F) HPV infection, (G) TP53 gene mutation, (H) lymph node metastasis. (I) Relative PLAU expression in OSCC versus normal oral squamous cell epithelium. And the association between PLAU expression and OSCC (J) patients' age, (K) gender, (L) cancer grade, (M) cancer stage, (N) HPV infection, (O) TP53 gene mutation, (P) lymph node metastasis. (* p<0.05, **p<0.01, ***p<0.001. The first layer * which is right above the error bar representing comparison to normal group, and the above layers * which were above a secondary line represent the comparison between corresponding groups that were covered by the line).
Figure 8

Construction of a prognosis prediction nomogram based on genes expression and OSCC clinical parameters

OSCC patients prognosis prediction nomogram constructed based on key genes signature (median SPP1 + median PLAU) and clinical parameters which were supported by Cox Regression to be independently
related with patients survival. (B) Correlation analysis between SPP1 and PLAU genes expression in HNSCC. (C) Overall survival analysis of four group of OSCC patients with SPP1-/PLAU-, SPP1A-/PLAU+, SPP1+/PLAU- and SPP1+/PLAU+ expression respectively. (D, E) LASSO analysis to calculate the best genes equation of SPP1 and PLAU which posses highest prognosis prediction value. (F) TCGA OSCC patients ordered by gene signature points. (G) Survival analysis of OSCC patients with high (above median) and low (below median) genes signature respectively. (H) ROC curve to compare the prognosis prediction value of two gene signatures (“0.0405*SPP1+0.179*PLAU” versus “0.5*SPP1+0.5*PLAU”). (I) ROC curving compare the prediction value of gene signature and other clinical parameters.
Figure 9

SPP1 and PLAU genes variations and the association between genes expression with immune cells infiltration in OSCC

Different types of (A) SPP1 and (B) PLAU variations in human cancers analyzed based on cBioPortal dataset. (C) The association between SPP1 expression and broad spectrum immune cells infiltration in
human cancers. (D) Association between SPP1 expression and CD8+ T cell, CD4+ T cell, B cell, dentic cell, macrophage cell and cancer associated fibroblast infiltration in OSCC. (E) Association between PLAU expression and broad spectrum immune cells infiltration in human cancers. (F) Association between PLAU expression and CD8+ T cell, CD4+ T cell, B cell, dentic cell, macrophage cell and cancer associated fibroblast infiltration in OSCC.

**Figure 10**

SPP1 and PLAU centered biological functions and signaling pathways

The PPI network which is centered on SPP1 gene for analyzing (B) the main biological functions and potential signaling pathways SPP1 and its connected genes participated in. (C) The PPI network which is
centered on PLAU gene for analyzing (D) the main biological functions and potential signaling pathways SPP1 and its connected genes participated in.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTables12.doc
- TableS3Originaldataof304localhospitalpatientsinformationforlHCexperiment.pdf