Weighted Gene Co-Expression Network Analysis Identifies Hub Genes Associated with Occurrence and Prognosis of Oral Squamous Cell Carcinoma

You Ge
Wei Li
Qian Ni
Yan He
Jinjin Chu
Pingmin Wei

Background: The aim of this study was to identify biomarkers closely related to the pathogenesis and prognosis of oral squamous cell carcinoma (OSCC) by using weighted gene co-expression network analysis (WGCNA) based on integrative transcriptome datasets.

Material/Methods: Gene expression profiles of OSCC were downloaded from the Gene Expression Omnibus (GEO) database. Differentially expressed genes (DEGs) were obtained and we then performed with Gene ontology (GO) and pathway enrichment analysis as well as protein–protein interactions (PPI) network analysis. WGCNA was used to construct the co-expression network. Multipart results were intersected to acquire the candidate genes, and survival analysis was used to identify the hub genes.

Results: A total of 568 DEGs, including 272 upregulated genes and 296 downregulated genes, were identified. GO and pathway analyses revealed that these DEGs were mainly enriched in extracellular matrix (ECM), ECM organization, structural constituent of muscle, and ECM-receptor interaction. The PPI network of DEGs was established, comprising 428 nodes and 1944 edges. In the co-expression network, pink module was the key module, in which 34 genes with high connectivity were identified. After the intersection of multipart results, 24 common genes were chosen as the candidate genes, among which 7 hub genes (PLAU, SERPINE1, LAMC2, ITGA5, TGFBI, FSCN1, and HLF) were identified using survival analysis.

Conclusions: Seven potential biomarkers were identified as being closely related with the initiation and prognosis of OSCC and might serve as potential targets for early diagnosis and personalized therapy of OSCC.

MeSH Keywords: Carcinoma, Squamous Cell • Gene Expression Profiling • Biological Markers

Full-text PDF: https://www.medscimonit.com/abstract/index/idArt/916025
Background

Oral cavity cancer is a global public health issue and is the sixth most common malignancy of humans, accounting for more than 300,000 new cases annually [1,2]. It has been estimated that approximately 300,400 new cases of oral cancer and 145,400 related fatalities occurred worldwide in 2012 [3]. Despite the relatively low incidence of oral cancer in China, the number of OSCC patients is still large due to the huge population base. It was reported that in 2010, there were over 34,000 new cases and 14,000 people died from oral cancer in China [4].

Oral cancer is growth of cancerous tissue in the oral cavity, and most oral malignancies (approximately 90%) are histologically subtyped as oral squamous cell carcinoma (OSCC) [5]. The main etiological factors are tobacco use, excessive alcohol consumption, chewing betel quid (especially in some Asian areas), and human papilloma virus 16 (HPV16) infection [6–9]. In the last few decades, great efforts have been made to fight OSCC. Despite substantive progress in surgical and medical treatments for OSCC, the overall 5-year survival rate has not significantly improved and remains approximately 50% [5]. Delay in diagnosis of OSCC patients in an early stage leads to progression to an advanced stage [10]. Most OSCC patients had a poor prognosis because of the advanced clinical stage at which they were diagnosed. One of the important reasons for failure in early diagnosis is insufficient research on the mechanisms at molecular levels underlying the carcinogenesis of OSCC. In-depth research on the molecular mechanisms in cancer initiation and progression of OSCC are needed, and this would also benefit OSCC patients at treatable stages. Therefore, it is of great importance to identify novel biomarkers for OSCC and reveal the molecular events contributing to OSCC pathogenesis.

The comparative analysis of differential gene expression between cancer tissues and normal controls will strengthen our exploration of the molecular pathogenesis and thus promote the identification of potential target genes and pathways for OSCC therapy. Previous bioinformatics studies about OSCC either utilized a single dataset with small sample size, or just directly merged multiple datasets, ignoring their batch effects and inherent heterogeneity [11,12]. As an alternative, the ComBat method can address these limitations, and can combine gene expression profiles from different datasets by removing the batch effects [13]. Furthermore, based on the theory that genes with high expression profile similarity may have closely related functional linkages or be involved in interacting pathways [14], the weighted gene co-expression network analysis (WGCNA) algorithm provides a systems biology approach to describe detailed characteristics at the level of genetic networks. WGCNA can establish free-scale gene co-expression networks to screen clusters (modules) of highly correlated genes and construct modules related to sample traits [15]. Zhang et al. used WGCNA to identify 2 modules and 10 hub genes associated with OSCC [16], but they only used a single dataset and the study lacked sufficient representativeness due to its limited sample size. Thus, integrating the data from different independent studies by ComBat method and the construction of a co-expression network based on this data will provide deeper insight into the molecular mechanisms of tumor genes associated with OSCC.

In the present study, we first integrated 5 gene expression profile datasets from GEO and removed the batch effects of these datasets by ComBat method. Then, we identified differentially expressed genes (DEGs) between OSCCs and normal samples, which were further assessed with Gene ontology (GO) function and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis, as well as protein–protein interactions (PPI) network analysis. Weighted gene co-expression network analysis (WGCNA) was used to construct a co-expression network of relationships between genes to find network-centric genes. Subsequently, multipart results were intersected to obtain the candidate genes, and then the log-rank test of Kaplan-Meier analysis was performed to identify the hub genes for OSCC. Finally, we used other datasets to demonstrate the value of the hub genes.

Material and Methods

Date search strategy and selection criteria

The gene expression profile datasets were searched from the GEO database with the search terms “Oral squamous cell carcinoma” or “OSCC” and “Homo sapiens” and “Expression profiling by array”. The datasets were eligible if they met the following criteria: (1) mRNA expression profiling by array; (2) datasets compared between OSCC and normal controls; (3) the number of samples more than 20; and (4) accessible gene expression profiles and platform information. The exclusion criteria were: (1) duplicated or non-relevant datasets; (2) non-coding RNA expression profiles; (3) methylation profiles; (4) datasets compared between cell lines; (5) datasets without normal control; (6) the number of samples less than 20; and (7) incomplete gene expression profiles or platform information. Moreover, a manual search of relevant OSCC datasets listed in the Materials and Methods section of the published articles was conducted. The dataset selection procedure is summarized in Figure 1. Six microarray gene expression datasets met the inclusion criteria (GSE30784, GSE13601, GSE47991, GSE31056, GSE9844, and GSE23558) were obtained from the GEO repository and the characteristics of these datasets are listed in Table 1. Five datasets with a larger sample size (GSE30784, GSE13601, GSE47991, GSE31056, and GSE9844) were used to perform an integrative...
transcriptome analysis, and the other dataset (GSE23558) was used to validate the results.

Data preprocessing and differential expression analysis

The raw data were preprocessed for background correction, log2 transformation, quantile normalization, and then converted into expression matrix by using Robust Multi-array Average (RMA) function of the affy R package [17]. Batch effects are inevitable when pooling microarray data across different laboratories, array types, or platforms. ComBat, combining location (mean) and scale (variance) adjustment with empirical Bayes, is a highly effective method of removing batch effects [13]. The batch effects when integrating the 5 datasets were removed by the ComBat function in the SVA R package. Afterwards, DEGs of the integrated datasets were screened through the limma R package with the cut-off criteria adjusted P<0.05 and |log 2-fold change (FC)| >1 based on the Benjamin and Hochberg (BH) procedure [18].

Gene ontology and pathway enrichment analysis

GO and KEGG pathway enrichment analyses were performed using the clusterProfiler package, which offers enrichGO and enrichKEGG methods for enrichment analysis and has the visualization function of profiles for genes and gene clusters [19]. P<0.05 was considered as the threshold value.

Protein–protein interactions (PPI) network construction

The online database STRING (http://string-db.org/) was employed to establish the PPI network of DEGs, with a confidence score ≥0.7. Cytoscape software (National Institute of General Medical Sciences of the National Institutes of Health, Bethesda, MD, USA) was then used to visualize and analyze the network. The Molecular Complex Detection (MCODE) plug-in of Cytoscape was used to find clustered sub-networks (highly interconnected regions) in the PPI network [20]. Degree ≥2, node score ≥0.2, K-core ≥2, and max depth=100 were used as cut-off criteria.
Construction of co-expression network

The integrated gene expression data after removing the batch effects were chosen for constructing the co-expression network using the "WGCNA" R package [15]. First, we constructed the Pearson correlation coefficient matrix between gene pairs based on the gene expression profile. Second, an appropriate soft threshold power (β) was selected in line with scale-free topology criteria. We turned the correlation coefficient matrix into the weighted adjacency matrix through a power function $\beta = |\text{cor}(x, y)| \beta$ (aij-weighted adjacency matrix, cor(x, y)=Pearson correlation coefficient matrix between gene pairs). Third, the adjacency matrix was transformed into a topological overlap matrix (TOM) by the fuction of TOM similarity, and the corresponding TOM-based dissimilarity (1-TOM) was calculated as well. Next, an average linkage hierarchical clustering dendrogram was built on the basis of TOM-based dissimilarity, and clustering module identification was achieved using dynamic tree cut with the minimum module size 50. Moreover, the dissimilarity of module eigengenes (MEs, the first principal component of one module) were calculated, and the cut line of 0.25 for module dendrogram was chosen as the module-merged standard. Finally, we calculated the relationships between module and trait, gene significance (GS), and module membership (MM) to find the key modules. GS, defined as the log$_{10}$ transformation of the P value (lgP) in the linear regression between gene expression and trait information, was used to quantify associations of individual genes with a trait. MM was defined as the correlation between gene expression profile and the ME of a given module. The key modules were identified based on the correlation between MEs and trait.

Mining of candidate genes

Hub genes in key modules were regarded as genes with high module connectivity measured by the absolute value of the Pearson correlation and the clinical trait relationship (MM) >0.8 and (GS) >0.2. The preprocessed level 3 RNA-seq expression data and clinical information for HNSC were downloaded from TCGA. A total of 499 patients with detailed survival data were included for subsequent analysis. The edgeR package was used to identify the DEGs for TCGA RNA-seq data using the limma R package after preprocessing and removing batch effects. The heatmap was used to evaluate batch effects of the integrated data before and after using ComBat methods (Supplementary Figures 1, 2). Before removing batch effects, the cluster tree of the samples was divided into 5 categories. All samples of the 5 datasets in the cluster tree were evenly mixed together and the batch effects were removed after performing ComBat methods. A total of 568 DEGs, including 272 upregulated genes and 296 downregulated genes, were identified for subsequent analysis.

Gene ontology and pathway enrichment analysis

GO and KEGG pathway enrichment analyses were performed using the clusterProfiler R package. For the biological processes (BP), the top 5 enriched categories among DEGs were extracellular matrix (ECM) organization, extracellular structure organization, collagen metabolic process, muscle filament sliding, and actin-myosin filament sliding. In cellular component (CC) ontology, DEGs were significantly enriched in ECM, proteinaceous ECM, ECM component, contractile fiber component, and contractile fiber component. Molecular function (MF) analysis indicated that structural constituent of muscle, integrin binding, actin binding, heparin binding, and growth factor binding were the top 5 commonly enriched categories (Table 2). KEGG analysis showed that DEGs were mainly enriched in ECM-receptor interaction, focal adhesion, protein digestion and absorption, amoebiasis, and IL-17 signaling pathway (Table 2).

PPI network construction and module selection

The PPI network of DEGs, including 428 nodes and 1944 edges, was constructed using the STRING online database and Cytoscape software (Figure 2). Then, the MCODE plug-in was applied for module selection of the PPI network and module 1, comprising 93 nodes and 818 edges, got the highest score (Figure 3). In module 1, the top 10 genes according to module scores were PLOD2, COL10A1, COL11A1, COL17A1, COL5A2, COL7A1, COL16A1, COL4A5, OAS1, and RSAD2.

Results

DEGs screening in OSCC

The integrated data (GSE30784, GSE13601, GSE37991, GSE31056, and GSE9844) were analyzed using the limma R package after preprocessing and removing batch effects. The heatmap was used to evaluate batch effects of the integrated data before and after using ComBat methods (Supplementary Figures 1, 2). Before removing batch effects, the cluster tree of the samples was divided into 5 categories. All samples of the 5 datasets in the cluster tree were evenly mixed together and the batch effects were removed after performing ComBat methods. A total of 568 DEGs, including 272 upregulated genes and 296 downregulated genes, were identified for subsequent analysis.

Hub genes identification and validation

Log-rank test in Kaplan-Meier analysis was performed to demonstrate the effect of candidate genes expression on prognosis using survival R package. Candidate genes (log-rank P<0.05) were considered as the hub genes. Similarly, GSE23558 was used to verify expression of hub genes. The effect of hub gene expression on patient prognosis was verified by GSE41613.
Weighted co-expression network construction and identification of key modules

The integrated gene expression profiles in which we had been eliminated the batch effects were analyzed by WGCNA R package. The power of \(b=5\) (scale-free \(R^2=0.86\)) was set as the appropriate soft-thresholding value to satisfy the scale-free network criteria. A total of 16 modules, ranging in size from 77 to 1857 genes, were identified and labeled with different colors (Figure 4A). According to the correlation between MEs and trait, the pink module was the most correlated with trait of cancer and was considered as the key module (Figure 4B).

L2 Candidate genes mining for OSCC

Based on the hub genes screening criteria in the key module (\(|GS|>0.2\) and \(|MM|>0.8\)), 34 genes in the pink module were regarded as hub genes in the co-expression network (Figure 4C, Supplementary Table 1). TCGA RNA-seq expression data were analyzed using the edgeR package. There were 6712 DEGs in total under the threshold value of FDR <0.05 and \(|\log_{2}FC|>1.5\), in which 4060 upregulated and 2652 downregulated genes were identified. Using the intersect function in Venny 2.1.0, we identified 24 genes in all 3 parts of the results (DEGs of 5 gene microarray data, DEGs of TCGA data, and hub genes in the co-expression network) as the candidate genes (Figure 4D, Supplementary Table 1).

Hub genes identification and validation

Log-rank analysis was used to evaluate the difference in overall survival between high expression and low expression of these candidate genes. This procedure showed that 7 genes (PLAU, SERPINE1, LAMC2, ITGA5, TGFBI, FSCN1, and HLF) were significantly associated with the prognosis of OSCC patients, and these were defined as hub genes (Figure 5, Supplementary Figures 3, 4). Survival analysis demonstrated that OSCC patients with high expression of 6 hub genes (PLAU, SERPINE1, LAMC2, ITGA5, TGFBI, and FSCN1) had lower overall survival than those with low expression. In contrast, patients with high expression of gene HLF displayed remarkably longer overall survival compared to those with low expression.

| Category | Term/pathway ID | Description | Count | P-value |
|----------|----------------|-------------|-------|---------|
| BP       | GO: 0030198    | Extracellular matrix organization | 64    | 1.46E-27 |
| BP       | GO: 0043062    | Extracellular structure organization | 59    | 5.25E-27 |
| BP       | GO: 0032963    | Collagen metabolic process | 29    | 1.73E-19 |
| BP       | GO: 0030049    | Muscle filament sliding | 19    | 9.50E-19 |
| BP       | GO: 0033275    | Actin-myosin filament sliding | 19    | 9.50E-19 |
| CC       | GO: 0031012    | Extracellular matrix | 84    | 2.11E-40 |
| CC       | GO: 0005578    | Proteinaceous extracellular matrix | 69    | 1.50E-34 |
| CC       | GO: 0044420    | Extracellular matrix component | 31    | 5.15E-23 |
| CC       | GO: 0043292    | Contractile fiber | 44    | 1.82E-22 |
| CC       | GO: 0044449    | Contractile fiber part | 42    | 4.39E-21 |
| MF       | GO: 0008307    | Structural constituent of muscle | 18    | 7.59E-16 |
| MF       | GO: 0005178    | Integrin binding | 20    | 1.73E-10 |
| MF       | GO: 0003779    | Actin binding | 39    | 6.54E-10 |
| MF       | GO: 0008201    | Heparin binding | 23    | 7.87E-10 |
| MF       | GO: 0019838    | Growth factor binding | 21    | 9.45E-10 |
| KEGG     | hsa04512      | ECM-receptor interaction | 21    | 4.76E-11 |
| KEGG     | hsa04510      | Focal adhesion | 28    | 1.34E-08 |
| KEGG     | hsa04974      | Protein digestion and absorption | 19    | 6.94E-08 |
| KEGG     | hsa05146      | Amoebiasis | 18    | 2.22E-07 |
| KEGG     | hsa04657      | IL-17 signaling pathway | 16    | 3.48E-06 |

GO – Gene ontology; KEGG – Kyoto Encyclopedia of Genes and Genomes; BP – biological processes; CC – cellular component; MF – molecular function.
Figure 2. PPI network construction. PPI network of DEGs containing 428 nodes and 1944 edges was constructed. Upregulated and downregulated genes are shaded in red color and green, respectively. PPI – protein–protein interaction; DEGs – differentially expressed genes.

Figure 3. Module analysis of PPI network. Module 1 with the highest score was identified by MCODE plug-in of Cytoscape software, including 93 nodes and 818 edges. Upregulated and downregulated genes are shaded red and green, respectively. PPI – protein–protein interaction; MCODE – Molecular Complex Detection.
GSE23558 was used to confirm the differences in expression of 7 hub genes between normal and OSCC tissues. The results suggested that the expression of hub genes was significantly higher in OSCC tissues than in normal tissues, except for gene HL, which had lower expression in OSCC tissues compared with normal tissues (Table 3). The GSE41613 dataset was successfully used to validate the effect of hub genes on OSCC patient prognosis (Figure 6).

**Discussion**

The transformative process of normal stratified squamous oral mucosa into squamous cell carcinoma contains several steps and factors in which accumulated genetic alterations intervene with the normal functions of oncogenes and tumor suppressor genes [22]. However, the underlying molecular mechanisms involved in the process are unclear. In the present study, by integrating 5 individual cohorts of gene expression...
profile datasets and using multiple bioinformatics methods, we identified key pathways and 7 hub genes in OSCC tissues not found in the normal controls.

Based on the GO analysis, the most significant enrichment in biological process was extracellular matrix (ECM) organization, a process occurring at the cellular level, resulting in the assembly, arrangement of constituents, or disassembly of ECM. Through a literature research, we found that the process of ECM organization appears to be associated with cancer-associated fibroblasts (CAFs). Comparison of genes or proteins between CAFs and normal fibroblasts revealed that most of the enriched genes or proteins are related to ECM organization [23,24]. As a major element of tumor stroma, CAFs can promote tumor growth and invasiveness by affecting ECM remodeling through producing MMPs [25,26]. Several studies have also indicated the vital role of CAFs in OSCC development and metastasis [27,28]. Other biological processes such as the collagen metabolic process were also enriched, which might be due to the production of MMPs [29]. Similarly, the top 3 cellular components were associated with ECM. ECM is a highly dynamic structure, continuously performing the remodeling process, including ECM components deposition, degradation, or other modification [26]. These results of biological process and cellular component ontology highlight the pivotal effect of abnormal ECM dynamics on OSCC occurrence and progression. For the molecular function ontology, the enriched categories of structural constituent of muscle, actin binding,
Collagen is one of the most important components of ECM, and increasing evidence suggests collagen affects tumor growth and metastasis through multiple mechanisms [38]. Some studies have explored the mechanism of collagen XVI and XVII facilitation of proliferation and invasion of OSCC cells [39,40]. Additionally, the overexpression of some collagen-related genes is also possibly due to the net result of overproduction and degradation.

In this study, we integrated 3 parts of results and combined them with survival analysis to identify 7 genes (PLAU, SERPINE1, LAMC2, TGFBI, ITGA5, FSCN1, and HLF) as hub genes. PLAU (plasminogen activator, urokinase) and SERPINE1 (plasminogen activator inhibitor-1) are both part of the plasminogen activation system. LAMC2, TGFBI, and ITGA5 are ECM protein-related gene. The remaining 2 genes (FSCN1 and HLF) are novel molecules that have received little research interest to date.

PLAU encodes the urokinase-type plasmin activator (uPA), a serine protease which converts inactive plasminogen to plasmin by binding to its receptor (uPAR). uPA has been shown to be involved in tissue remodeling and migration under physiological conditions and tumorigenesis [41,42]. uPA/uPAR plays critical roles in promoting extracellular proteolysis, regulation of cell/ECM interactions, and cell migration, all of which are related to the malignant progression of various tumors [43]. Elevated expression of uPA was observed in OSCC tissues, which was correlated with increased invasiveness of OSCC [44].

SERPINE1, also known as plasminogen activator inhibitor-1 (PAI-1), is the principal inhibitor of tissue plasminogen activator (tPA) and uPA. However, higher expression of SERPINE1 has been described as a poor prognostic marker in several...
cancers. Especially, SERPINE1 had been validated as a biological marker for treatment regimen selection in patients with node-negative breast cancer [45]. Bajou et al. observed that absence of SERPINE1 expression in mice had the effect of inhibiting malignant cell invasion and angiogenesis after transplantation of malignant keratinocytes, whereas invasion and vascularization recurred when mice were injected with adenovirus carrying human SERPINE1 [46]. These observations indicate that SERPINE1 has a multifunctional role in promoting tumor development, invasion, and metastasis, independent of the ability to function as a protease inhibitor. Recent findings have suggested that SERPINE1 can regulate apoptosis and make the tumor cells detach from vitronectin and the ECM by displacing vitronectin from the uPA receptor-vitronectin interaction, thus enhancing tumor cell migration and metastasis [47,48].

TGFBI has dual functions in tumor progression, acting as a tumor suppressor in breast cancer and mesothelioma [60], and, compared with normal mouth mucosa, its expression is increased in precancerous and, more significantly, in OSCC [62]. Our results indicate that TGFBI might have a positive regulatory effect, and higher expression of TGFBI in OSCC tissues suggests a poor prognosis.

Several studies have shown that TGFBI is upregulated in OSCC tissues [61,62], and, compared with normal mouth mucosa, its expression is increased in precancerous and, more significantly, in OSCC [62]. Several studies have shown that TGFBI is upregulated in OSCC tissues [61,62], and, compared with normal mouth mucosa, its expression is increased in precancerous and, more significantly, in OSCC [62]. Our results indicate that TGFBI might have a positive regulatory effect, and higher expression of TGFBI in OSCC tissues suggests a poor prognosis.

FSCN1, known as fascin actin-bundling protein 1, exerts a critical function in regulating cell migration, cell motility, and cell-to-cell interactions [63]. FSCN1 is usually absent in normal epithelial tissues but is overexpressed in many human carcinomas, suggesting aggressive, metastatic carcinomas and poor prognosis [64–66]. FSCN1 was found to participate in EMT and promote invasive filopodia formation, which confers increased motility and metastatic properties to cancerous cells [67]. Given the lack of studies showing involvement of FSCN1 in OSCC, further studies are required.

Integrin-α5 (ITGA5) is a member of the integrins family, which has been demonstrated to regulate various complex biological events such as differentiation, development, cell adhesion, and control of cancer growth and progression. Integrin α5, which is associates with integrin β1 to form a fibronectin receptor, has been shown to exert a pivotal role in certain cancers such as non-small cell cancer, esophageal squamous cell carcinoma, and breast cancer [52–54]. Integrin α5β1 can promote cancer cell migration and invasion through activating the focal adhesion kinase (FAK) and Src [55]. In addition, Claudia et al. found that integrin α5β1 enhances cancer cell invasiveness by facilitating the generation of contractile forces [56]. However, its expression pattern and function in OSCC are still elusive.

HLF (hepatic leukemia factor) is a member of the proline and acid-rich (PAR) bZIP transcription factor family, which can form homodimers or heterodimers among each other and regulates transcriptional activity [68]. Chromosomal translocations fuse portions of HLF with the E2A gene to form E2A-HLF, a chimeric transcription factor created by the t(17;19) gene, which contributes to leukemogenesis through its potential to inhibit apoptosis [69,70]. Moreover, HLF can increase miR-132 expression through the HLF binding site BS1 of mir-132 promoter, suppressing the proliferation of glioma cells, metastasis, and radioresistance via inhibiting a downstream factor TTK protein kinase [71]. Intriguingly, in our study, the expression of HLF in OSCC tissues was downregulated compared to normal tissues, and patients with low expression of the HLF gene had shorter overall survival compared to patients with high expression. This result indicates HLF might have potential value in OSCC treatment. Unfortunately, there has been no study discussing the regulating role of HLF in OSCC, and further studies are needed.

Zhang et al. recently used WGCNA to identify 10 hub genes (MMP1, TNFRSF12A, PLA2, FSCN1, PDN1, KRT78, EVPL, GGTL6, SMIM5, and CYSRT1) that are associated with OSCC carcinogenesis and undertook survival analysis to validate the prognostic value of these genes [16]. Among these 10 hub genes, 4 genes (PLA2, FSCN1, MMP1, and PDN1) were involved in our analysis. PLA2 and FSCN1 were also identified as hub genes in our study, which might have diagnostic and prognostic perspectives for OSCC patients. However, in our study, MMP1 and PDN1 were considered to be the candidate genes and had no effect on the prognosis of OSCC. These different conclusions regarding MMP1 and PDN1 might be due to the difference in data used for survival analysis. The other 6 hub genes (TNFRSF12A, KRT78, EVPL, GGTL6, SMIM5, and CYSRT1) identified by Zhang et al. were not included in our 2 parts of this study.
results (DEGs of 5 gene microarray data and hub genes in the co-expression network). Several factors might be responsible for this difference between the 2 studies. First, Zhang et al. used a single dataset (GSE30784) to construct the gene co-expression network to identify these 6 genes. In our study, we integrated 5 datasets (GSE30784, GSE13601, GSE37991, GSE31056, and GSE9844) for subsequent analysis. Therefore, these 6 genes might not be statistically significant in our integrative data, and cannot meet our cut-off criteria. Second, Zhang et al. used WGCNA to study the genes related to the transformation of normal mucosa to oral dysplasia and oral dysplasia to carcinoma. However, among the 5 datasets obtained from GEO, only GSE30784 contains normal mucosa, oral dysplasia, and OSCC samples. Thus, we could merely identify hub genes related to the process of normal to OSCC, which might also lead to the different results.

There are several limitations worth noting in our study. First, in a co-expression network, a connection between 2 genes cannot be assumed to correspond with a connection in regulatory or PPI networks. When compared with other biological networks where the edges represent well-defined biological interactions, the edges in a co-expression network might be a limited representation of the correlation of the data. Therefore, experimental or clinical studies are needed to further validate these findings. Second, due to the lack of specific clinical information in these 5 datasets from GEO, we failed to construct a co-expression network to explore the relationship between genes and clinical features. Furthermore, we simply studied the effect of hub genes expression on OSCC patient prognosis, while more clinical parameters such as HPV infection and cancer staging should be included in further analysis. Third, among the 5 datasets obtained from GEO, only GSE30784 consists of normal mucosa, oral dysplasia, and OSCC samples. Thus, we could merely analyze the transformation from normal to carcinoma. To include the dynamic analysis from normal to precancerous lesions and further to OSCC would be much better. Fourth, this study only used TCGA and GEO data, and more data from other databases should be assessed to produce a more comprehensive analysis.

Conclusions

We have identified 7 hub genes (PLAU, SERPINE1, LAMC2, TGFBI, ITGA5, FSCN1, and HLF) by using WGCNA and found they are closely correlated with the initiation and prognosis of OSCC. The GO and KEGG pathway enrichment analysis combined with the hub genes has emphasized the essential role of ECM in OSCC occurrence and progression. Among the 7 genes, ITGA5, FSCN1, and HLF are relatively new biomarkers for OSCC, and few studies about their roles in OSCC are currently available, so this topic needs further experimental verification. These novel biomarkers will greatly contribute to the early diagnosis and prognosis prediction in OSCC.

Acknowledgements

The authors gratefully acknowledge the researchers who provided the datasets used in this analysis.

Conflicts of interests

None.
Supplementary Figure 1. The heatmap of the integrated data before removing batch effects. Each column and row represent one sample and one gene, respectively. Red shows high relative expression and green suggests low relative expression. The cluster tree shows that the samples was divided into 5 categories.
Supplementary Figure 2. The heatmap of the integrated data after removing batch effects. Each column and row represent one sample and one gene, respectively. Red shows high relative expression and green suggests low relative expression. The cluster tree shows that all the samples of 5 datasets were evenly mixed together.
**Supplementary Table 1.** Thirty-four hub genes in the co-expression network.

| Gene names | GS. Cancer | MM. turquoise |
|------------|------------|---------------|
| LAMC2*     | 0.764      | 0.911         |
| PLAU*      | 0.754      | 0.892         |
| MMP1*      | 0.788      | 0.891         |
| SERPINE1*  | 0.714      | 0.882         |
| PDPN*      | 0.717      | 0.879         |
| PTHLH*     | 0.721      | 0.874         |
| FSCN1*     | 0.707      | 0.873         |
| HOMER3*    | 0.729      | 0.873         |
| ITGA3*     | 0.674      | 0.867         |
| MMPI0*     | 0.700      | 0.864         |
| CDH3       | 0.750      | 0.860         |
| MYO1B*     | 0.720      | 0.858         |
| TMSB10     | 0.709      | 0.828         |
| MYH9       | 0.613      | 0.817         |
| SLC3A2     | 0.649      | 0.817         |
| PTK7*      | 0.667      | 0.814         |
| CYP27B1*   | 0.626      | 0.807         |
| SNAIL2*    | 0.670      | 0.807         |
| ITGB4      | 0.674      | 0.805         |
| IL24*      | 0.653      | 0.802         |
| ITGA6*     | 0.653      | 0.801         |
| NCOA1      | –0.671     | –0.801        |
| EPHX2*     | –0.705     | –0.816        |
| RRAGD*     | –0.702     | –0.834        |
| UBL3*      | –0.741     | –0.845        |
| ACADSB*    | –0.686     | –0.846        |
| HLF*       | –0.747     | –0.888        |
| GPD1L*     | –0.733     | –0.896        |

* 24 candidate genes identified through intersection of three parts of results (DEGs of three gene microarray data, DEGs of TCGA data, hub genes in the co-expression network). GS – gene significance; MM – module membership.

**Supplementary Figure 3.** Survival analysis of the other 17 genes.
Supplementary Figure 4. Survival analysis of the other 17 genes.

References:

1. Ferlay J, Soerjomataram I, Dikshit R et al: Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer, 2015; 136(5): E359–86
2. Thomson PJ: Perspectives on oral squamous cell carcinoma prevention-proliferation, position, progression and prediction. J Oral Pathol Med, 2018; 47(9): 803–7
3. Torre LA, Bray F, Siegel RL et al: Global cancer statistics, 2012. Cancer J Clin, 2015; 65(2): 87–108
4. Zheng CM, Ge MH, Zhang SS et al: Oral cavity cancer incidence and mortality in China, 2010. J Cancer Res Ther, 2015; 11(Suppl. 2): C149–54
5. Warnakulasuriya S: Global epidemiology of oral and oropharyngeal cancer. Oral Oncol, 2009; 45(4–5): 309–16
6. Smokeless tobacco and some tobacco-specific N-nitrosamines. IARC Monogr Eval Carcinog Risks Hum, 2007; 89: 1–592
7. Reidy J, McHugh E, Stassen LF: A review of the relationship between alcohol and oral cancer. Surgeon, 2011; 9(5): 278–83
8. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans: Betel-quid and areca-nut chewing and some areca-nut derived nitrosamines. IARC Monogr Eval Carcinog Risks Hum, 2004; 85: 1–334
9. Dalianis T: Human papillomavirus and oropharyngeal cancer, the epidemiics, and significance of additional clinical biomarkers for prediction of response to therapy (Review). Int J Oncol, 2014; 44(6): 1799–805
10. Gomez I, Seoane J, Varela-Centelles P et al: Is diagnostic delay related to advanced-stage oral cancer? A meta-analysis. Eur J Oral Sci, 2009; 117(5): 541–46
11. Li G, Li X, Yang M et al: Prediction of biomarkers of oral squamous cell carcinoma using microarray technology. Sci Rep, 2017; 7: 42105
12. Zhao X, Sun S, Zeng X, Cui L: Expression profiles analysis identifies a novel three-mRNA signature to predict overall survival in oral squamous cell carcinoma. Am J Cancer Res, 2018; 8(3): 450–61
34. Bohanes P, Yang D, Loupakis F et al: Integrin genetic variants and stage-
33. Martinez-Outschoorn UE, Lisanti MP, Sotgia F: Catabolic cancer-associat
32. Madar S, Goldstein I, Rotter V: ‘Cancer associated fibroblasts’ – more than
30. Kuzet SE, Gaggioli C: Fibroblast activation in cancer: When seed fertilizes
29. Van Doren SR: Matrix metalloproteinase interactions with collagen and
27. Kawashiri S, Tanaka A, Noguchi N et al: Significance of stromal desmopla
26. Lu P, Takai K, Weaver VM, Werb Z: Extracellular matrix degradation and
22. Thomson PJ: Perspectives on oral squamous cell carcinoma prevention-pro
21. Robinson MD, McCarthy DJ, Smyth GK: edgeR: A Bioconductor package for dif
20. Bader GD, Hogue CW: An automated method for finding molecular com
16. Zhang X, Feng H, Li Z et al: Application of weighted gene co-expression net
13. Johnson WE, Li C, Rabinovic A: Adjusting batch effects in microarray expres

META-ANALYSIS
This work is licensed under Creative Common Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0)
63. Hashimoto Y, Kim DJ, Adams JC: The roles of fascins in health and disease. J Pathol, 2011; 224(3): 289–300
64. Liu C, Gao H, Cao L et al: The role of FSCN1 in migration and invasion of pituitary adenomas. Mol Cell Endocrinol, 2016; 419: 217–24
65. Hanker LC, Karn T, Holtrich U et al: Prognostic impact of fascin-1 (FSCN1) in epithelial ovarian cancer. Anticancer Res, 2013; 33(2): 371–77
66. Luo A, Yin Y, Li X et al: The clinical significance of FSCN1 in non-small cell lung cancer. Biomed Pharmacother, 2015; 73: 75–79
67. Machesky LM, Li A: Fascin: Invasive filopodia promoting metastasis. Commun Integr Biol, 2010; 3(3): 263–70
68. Xiang DM, Sun W, Ning BF et al: The HLF/IL-6/STAT3 feedforward circuit drives hepatic stellate cell activation to promote liver fibrosis. Gut, 2018; 67(9): 1704–15
69. Inukai T, Inaba T, Dang J et al: TEF, an antiapoptotic bZIP transcription factor related to the oncogenic E2A-HLF chimera, inhibits cell growth by down-regulating expression of the common beta chain of cytokine receptors. Blood, 2005; 105(11): 4437–44
70. Hunger SP, Li S, Fall MZ et al: The proto-oncogene HLF and the related basic leucine zipper protein TEF display highly similar DNA-binding and transcriptional regulatory properties. Blood, 1996; 87(11): 4607–17
71. Chen S, Wang Y, Ji C et al: HLF/miR-132/TTK axis regulates cell proliferation, metastasis and radiosensitivity of glioma cells. Biomed Pharmacother, 2016; 83: 898–904