A comprehensive analysis of candidate gene signatures in oral squamous cell carcinoma

X. LI[1,‡], R. SUN[1], X. GENG[2,‡], S. WANG[1], D. ZEN[1], J. PEI[1], J. YANG[1], Y. FAN[1], H. JIANG[1], P. YANG[3,*], C. LI[1,*]

[1]Department of Head and Neck Surgery, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, 610041, China; [2]Department of internal medicine, Sichuan Provincial People's Hospital, Chengdu, 610072, China; [3]College of Basic Medicine, Chengdu Medical College, Chengdu, 610500, China

*Correspondence: lli_chao3@163.com, wagn_pp@163.com
‡Contributed equally to this work.

This study aimed to unravel the molecular mechanism of oral squamous cell carcinoma (OSCC). With microarray dataset GSE30784, differentially expressed genes (DEGs) were identified between OSCC and control samples. The DEGs overlapped with genes obtained from online database MalaCards were determined as OSCCDEG, followed by Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis. A total of 5177 up-regulated and 6081 down-regulated DEGs were identified between OSCC and control. Out of the DEGs, 451 genes were overlapped with the 704 genes gained from MalaCards and regarded as "OSCCDEG". Up-regulated OSCCDEG were associated with cell cycle pathway, while down-regulated OSCCDEG were linked to ErbB pathway. ANGPT1, ANGPT2 and 3 hub proteins (EGFR, HSP90AA1, RB1) in the PPI network were associated with the survival rates of several tumors. The largest network module with the hub protein EGFR was associated with positive regulation of cell communication. The second largest module with the hub node FN1 was related to angiogenesis. For the third network module in connection with DNA metabolism, the hub protein was PCNA. ErbB and cell cycle pathways were crucial for OSCC. EGFR, FN1, PCNA, ANGPT1 and ANGPT2 might be potential biomarkers for OSCC. These findings help provide guidelines for treating OSCC.

Key words: oral squamous cell carcinoma, network module, differentially expressed genes, protein-protein interaction, gene ontology

As the most frequent tumor arising from oral cavity, oral cancer is of remarkable public health importance with an age-standardized incidence rate of 5.5 per 100, 000 in male [1]. Oral squamous cell carcinoma (OSCC) accounts for more than 90% of oral cancer. OSCC results from the unlimited multiplication of squamous cells lining the oral cavity and oropharynx [2]. The prognosis of patients with OSCC has remained poor for the last 20 years with an overall 5-year survival rate of 50%, largely due to delayed diagnosis and progression of the tumor [3, 4]. Therefore, there is an urgent need to unveil the molecular mechanism of OSCC, and identify possible molecular biomarkers predictive of OSCC for development of novel potent therapies.

Increasing studies are ongoing on the topic of the molecular mechanism of OSCC. For instance, it has been found that miRNA-181a (miR-181a) could suppress the growth of OSCC via inhibiting the expression of K-ras [5]. In a retrospective study enrolling 163 cases with OSCC, prospero homeobox 1 and forkhead box C2 were reported to be significantly correlated with angiogenesis and lymph-angiogenesis, which is critical for the progression of OSCC [6]. Furthermore, a genome-wide study showed that the gene *Talin-1*(*TLN1*) was overexpressed in OSCC, and could serve as a potential target for treating OSCC [7]. In addition to important genes, a number of signaling pathways have been identified to play important roles in OSCC. It has been reported that in OSCC, the tumor-promoting phosphoinositide 3-kinase (PI3K) /v-akt murine thymoma (AKT) pathway in OSCC could be activated by the mutation of phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha (PIK3CA) [8]. Parikh *et al.* have found that in OSCC cell lines, the ATR-CHEK1 pathway is up-regulated and appears to enhance the sensitivity of OSCC patients to the ionizing radiation treatment [9]. A previous study using microarray dataset GSE30784 has screened dif-

ferentially expressed genes (DEGs) between OSCC and control samples with the help of GenePlus software. It then identified potential biomarkers from the screened DEGs using the Forward and stepwise logistic regression for distinguishing OSCC from controls [10]. However, important questions regarding the precise biological function of DEGs involved in OSCC and the genes associated with the prognosis of the disease remain unanswered.

To address these issues, this study performed a secondary study of the microarray dataset GSE30784. Following the DEGs identification, the study gained OSCC-related genes from online database MalaCards. Overlapped DEGs with genes from MalaCards were identified to reduce the false positive rate. Apart from that, we underscored the significant signaling pathways involved in OSCC and uncovered the critical genes associated with the prognosis of patients with OSCC using online platform PPISURV. Protein-protein interaction (PPI) network was constructed as well, followed by network module analysis.

Materials and methods

Microarray data preprocessing. Gene expression profiling dataset GSE30784 [10] based on HG-U133-Plus-2 platform, was retrieved from Gene Expression Omnibus (GEO) database. It included 167 OSCC, 17 dysplasia and 45 normal oral tissues samples. Specifically, OSCC samples were extracted from 167 patients with first primary OSCC between December 1, 2003 and April 17, 2007 at the University of Washington Medical Center, Harbor-view Medical Center and the VA Puget Sound Health Care System in Seattle, Washington. Dysplasia samples were extracted from 17 patients with diagnosed dysplastic lesions at these same medical centers during the same period. Control samples were extracted from patients who had tonsillectomy or oral surgery for treatment of diseases other than cancer, such as obstructive sleep apnea, at the same medical center during the same time period. The robust multiarray average (RMA) algorithm[11] of the affy package in **R** language was utilized for transformation of the raw data in CEL profiles into expression data and normalization of the gene expression data.

Screening of DEGs. Linear Models for Microarray Analysis (LIMMA) package in R language is a specifically designed package for differential expression analysis of microarray data [12]. It was applied to identify DEGs between OSCC and control samples. Adjusted P-value < 0.05 was set as the threshold for DEGs identification. We screened 5177 up-regulated DEGs and 6087 down-regulated DEGs between OSCC and control samples. MalaCards is an integrated database of human maladies and their annotations, modeled on the architecture and strategy of the GeneCards database of human genes. It combined 44 data sources to generate a computerized card for each of 16919 human diseases [13]. We then retrieved 704 genes possibly related to OSCC from online database MalaCards. In order to reduce the false positive rate of DEGs between OSCC and control samples, we selected the overlapped genes between the DEGs screened

by LIMMA package and 704 genes from MalaCards. These overlapped genes were defined as the genes related to OSCC (OSCCDEG), and used for further analyses in this study.

KEGG pathway enrichment analysis. The Database for Annotation, Visualization and Integrated Discovery (DAVID) is characterized by an systematic biological knowledgebase and analytic tools for functional annotation of genome-scale dataset [14]. In order to map OSCC onto biological metabolic pathways, DAVID was used to perform Kyoto Encyclopedia of Genes and Genomes (KEGG) [15] enrichment analysis for these OSCCDEG obtained above. P-value < 0.05 was set as the strict cutoff after multiple testing correction with Benjamini-Hochberg (BH) approach.

PPI network construction and PPISURV analysis. Online database The Search Tool for the Retrieval of Interacting Genes (STRING) covers 5214234 types of proteins from 1133 organism species and was used to score the interactions between proteins by integrating genomic data, high-throughput experimental data, information of protein coexpressions and literature annotation [16].

In order to evaluate the interactions between OSCCDEG genes, the OSCCDEG list was submitted to the STRING database, which selected protein pairs with interaction score ≥ 0.4. A PPI network was built with these protein pairs selected. In the network, the degree of the protein node represented the number of interactions that a protein possessed. The hub proteins were defined as the protein with the highest degrees in the network.

The top 3 proteins with higher degree than other proteins in the network were selected, and the associations of the 3 proteins with the survival outcome of OSCC patient were then evaluated using online platform PPISURV [17]. PPISURV is a data-mining tool used to evaluate whether the expression of the input gene in 40 datasets of tumor patients samples is significantly associated with survival rates of patients. Subsequently, the relationship of the input gene interactome with the survival rates-related genome of each tumor dataset was detected as well.

Identification and GO enrichment analysis of network modules. Network clustering analysis was performed to identify PPI network modules from the PPI network using clusterMaker plugin [18] from Cytoscape software. Markov cluster (MCL) [19] algorithm was used to assign proteins into different families according to known sequence similarity knowledge. For extracting the biological information from identified network modules, gene ontology (GO) [20] functional enrichment analysis was performed using BinGo [21] plugin from Cytoscape software, followed by multiple testing correction with BH method (adjusted p-value < 0.05).

Results

Identification and KEGG pathway analysis of DEGs. Total 5177 up-regulated DEGs and 6081 down-regulated DEGs were identified between the OSCC samples and the control samples by using LIMMA. Among the 704 genes gained from

MalaCards database, 302 were up-regulated, while 149 were down-regulated. Moreover, 451 DEGs were overlapped with the 704 genes, and were regarded as "OSCCDEG".

The results of the KEGG pathway analysis for OSC-CDEG showed that both up-and down-regulated genes were associated with tumor-related KEGG pathways. Specifically, up-regulated genes were predominantly linked to cell cycle pathway (Table 1a), whereas down-regulated DEGs were primarily enriched in ERBB signaling pathway (Table 1b).

Table 1a. Top 10 KEGG pathways enriched with up-regulated OSCCDEG

KEGG PATHWAY	Count of genes	Genes	FDR
hsa05200:Pathways in cancer	68	E2F1, E2F3, PTGS2, PDGFA, MMP9, TGFB3, NFKB1, MMP2, PTEN, MMP1, TGFB1, TGFB2, CCNE2, AKT1, FOS, CCNE1, CDKN2A, SLC2A1, PIK3CA, TGFA, RARA, MYC, CHUK, AKT3, PRKCA, EGFR, HSP90AA1, SKP2, FADD, CDK6, RB1, CDK4, CDK2, VEGFB, VEGFC, HIF1A, JUN, VEGFA, LAMC2, XIAP, NFKBIA, IGF1R, PTK2, ITGAV, FN1, CSF1R, COL4A2, IL6, BMP2, COL4A1, EPAS1, IL8, TGFBR1, MET, ITGA2, BIRC5, ITGA3, BIRC2, STAT3, LAMA1, HSP90B1, CDKN1A, ITGA6, LAMA5, ETS1, BAX, ABL1, CRK	7.71E-29
hsa05222:Small cell lung cancer	31	E2F1, E2F3, XIAP, PTGS2, NFKBIA, NFKB1, PTEN, CCNE2, AKT1, CCNE1, PTK2, ITGAV, PIK3CA, MYC, CHUK, AKT3, FN1, COL4A2, COL4A1, SKP2, ITGA2, CDK6, ITGA3, RB1, CDK4, BIRC2, CDK2, LAMA1, ITGA6, LAMA5, LAMC2	1.74E-19
hsa04510:Focal adhesion	44	IBSP, CAV1, XIAP, PDGFA, TNC, ITGB4, PTEN, PXN, VCL, AKT1, IGF1R, PTK2, ITGAV, ILK, ITGB6, PIK3CA, ZYX, THBS1, AKT3, SPP1, FN1, PRKCA, EGFR, COL4A2, COL4A1, FLT1, MET, ITGA2, ITGA3, ITGA4, VAV2, BIRC2, VEGFB, VEGFC, LAMA1, CCND3, ITGA6, LAMA5, ITGA5, JUN, VEGFA, LAMC2, COL1A1, CRK	9.40E-19
hsa04110:Cell cycle	29	E2F1, E2F3, TGFB3, PRKDC, SFN, TGFB1, TGFB2, CCNE2, CCNE1, CDKN2A, RAD21, MCM7, CCNA2, MYC, CDC7, CDC6, CDK1, SKP2, CDK6, RB1, CDK4, CDC25C, CDK2, CDC25A, CDKN1A, CCND3, PCNA, BUB1B, ABL1	1.81E-12
hsa05219:Bladder cancer	17	E2F1, EGFR, E2F3, IL8, MMP9, RB1, CDK4, MMP2, MMP1, VEGFB, VEGFC, TYMP, CDKN1A, CDKN2A, VEGFA, THBS1, MYC	6.49E-11
hsa05212:Pancreatic cancer	21	E2F1, EGFR, E2F3, TGFBR1, TGFB3, CDK6, NFKB1, RB1, CDK4, STAT3, TGFB1, TGFB2, AKT1, VEGFB, VEGFC, CDKN2A, VEGFA, TGFA, PIK3CA, CHUK, AKT3	8.22E-11
hsa05220:Chronic myeloid leukemia	20	E2F1, E2F3, TGFBR1, TGFB3, NFKBIA, CDK6, NFKB1, RB1, CDK4, TGFB1, TGFB2, AKT1, CDKN1A, CDKN2A, PIK3CA, ABL1, CRK, MYC, CHUK, AKT3	1.48E-09
hsa04512:ECM-receptor interaction	21	IBSP, COL4A2, COL4A1, TNC, ITGB4, ITGA2, ITGA3, ITGA4, HMMR, LAMA1, CD44, ITGA6, ITGA5, LAMA5, ITGAV, ITGB6, LAMC2, COL1A1, THBS1, FN1, SPP1	1.37E-09
hsa04060:Cytokine-cy-tokine receptor interaction	36	IL1R1, CCL2, PDGFA, CCR1, CXCL2, TGFB3, CXCR1, CCL7, TGFB1, IL10, TGFB2, TNFRSF11B, CXCR4, IFNG, IL1B, CSF1R, EGFR, IL18R1, IL6, BMP2, FLT1, IL8, TGFBR1, MET, IL6R, CCL11, VEGFB, INHBA, VEGFC, CCR7, TNFSF10, CCR5, CXCL13, CCR3, VEGFA, NGFR	3.94E-09
hsa05215:Prostate cancer	20	E2F1, EGFR, E2F3, HSP90AA1, PDGFA, NFKBIA, NFKB1, RB1, PTEN, CDK2, AKT1, CCNE2, CCNE1, IGF1R, CDKN1A, HSP90B1, TGFA, PIK3CA, CHUK, AKT3	2.44E-08

KEGG: Kyoto Encyclopedia of Genes and Genomes; FDR: false discovery rate

Table 1b. Top10 KEGG pathways enriched with down-regulated DEGs

KEGG PATHWAY	Count of genes	Genes	FDR
hsa05200:Pathways in cancer	21	AR, FGF7, FGFR3, BRAF, ERBB2, CYCS, PPARG, MLH1, RAF1, KIT, CTNNA1, WNT1, CDKN1B, KRAS, CDKN2B, MAPK3, NKX3-1, RARB, EGF, AKT2, APC	7.69E-06
hsa05213:Endometrial cancer	10	KRAS, BRAF, ERBB2, MAPK3, MLH1, RAF1, EGF, CTNNA1, AKT2, APC	3.91E-06
hsa04012:ErbB signaling pathway	11	CDKN1B, KRAS, ERBB4, BRAF, ERBB3, ERBB2, MAPK3, RAF1, RPS6KB1, EGF, AKT2	2.56E-05
hsa05223:Non-small cell lung cancer	9	FHIT, KRAS, BRAF, ERBB2, MAPK3, RAF1, RARB, EGF, AKT2	3.81E-05
hsa05215:Prostate cancer	10	AR, CDKN1B, KRAS, BRAF, ERBB2, MAPK3, NKX3-1, RAF1, EGF, AKT2	1.70E-04
hsa05221:Acute myeloid leukemia	8	KRAS, BRAF, MAPK3, PIM1, RAF1, RPS6KB1, KIT, AKT2	4.87E-04
hsa05219:Bladder cancer	7	FGFR3, KRAS, BRAF, ERBB2, MAPK3, RAF1, EGF	6.13E-04
hsa04910:Insulin signaling pathway	10	KRAS, SLC2A4, TSC1, BRAF, MAPK3, FASN, RAF1, RPS6KB1, RPS6, AKT2	2.93E-03
hsa05210:Colorectal cancer	8	KRAS, BRAF, MAPK3, CYCS, MLH1, RAF1, AKT2, APC	3.53E-03
hsa05218:Melanoma	7	FGF7, KRAS, BRAF, MAPK3, RAF1, EGF, AKT2	8.16E-03

KEGG: Kyoto Encyclopedia of Genes and Genomes; FDR: false discovery rate

Table 2a. Relationship of 3 hub genes with clinical outcome of patients with tumors

Gene	GEO accession	GEO dataset	Cancer Type	GENE (Probe ID)	P-value	Effect Sign
RB1	GSE26712	a gene signature predicting for survival in suboptimally debulked patients with ovarian cancer	ovarian cancer	203132_AT	9.28E-04	Negative
RB1	GSE2034	breast cancer relapse free survival	breast cancer	211540_S_AT	1.90E-03	Positive
RB1	GSE13876	survival related profile, pathways and transcription factors in ovarian cancer	ovarian cancer	5050	1.17E-02	Negative
RB1	GSE25065	validation cohort for genomic predictor of response and survival following neoadjuvant taxane-anthracycline chemotherapy in breast cancer	breast cancer	203132_AT	4.48E-02	Positive
RB1	GSE16181	human meningioma fixed tumour tissue	meningioma	73	0.0451	Negative
HSP90AA1	GSE3494	an expression signature for p53 in breast cancer predicts mutation status, transcriptional effects, and patient survival	breast cancer	214328_S_AT	0.00238	Negative
HSP90AA1	GSE11121	the humoral immune system has a key prognostic impact in node-negative breast cancer	breast cancer	214328_S_AT	0.00265	Negative
HSP90AA1	GSE17536	metastasis gene expression profile predicts recurrence and death in colon cancer patients (moffitt samples)	colon cancer	211968_S_AT	0.00432	Negative
HSP90AA1	GSE1456	gene expression of breast cancer tissue in a large population-based cohort of swedish patients	breast cancer	210211_S_AT	0.00535	Negative
HSP90AA1	GSE2034	breast cancer relapse free survival	breast cancer	214328_S_AT	0.00575	Negative
HSP90AA1	GSE17538	experimentally derived metastasis gene expression profile predicts recurrence and death in colon cancer patients	colon cancer	211968_S_AT	0.0112	Negative
HSP90AA1	GSE24450	183 breast tumors from the helsinki univeriuty central hospital with survival information	breast cancer	ILMN_2373515	0.0128	Negative
HSP90AA1	GSE10846	prediction of survival in diffuse large b cell lymphoma treated with chemotherapy plus rituximab	diffuse large b cell lymphoma	211969_AT	0.0133	Positive
HSP90AA1	GSE30929	whole-transcript expression data for liposarcoma	liposarcoma	211969_AT	0.015	Negative
HSP90AA1	GSE39671	expression data from untreated cll patients	chronic lymphocytic leukemia	211968_S_AT	0.0163	Positive
HSP90AA1	GSE17705	endocrine sensitivity index validation dataset	breast cancer	211969_AT	0.0332	Negative
HSP90AA1	GSE30760	cervical cancer	cervical cancer	CG10833014	0.0361	Negative
EGFR	GSE16181	human meningioma fixed tumour tissue	meningioma	352	2.71E-10	Negative
EGFR	GSE25055	discovery cohort for genomic predictor of response and survival following neoadjuvant taxane-anthracycline chemotherapy in breast cancer	breast cancer	201984_S_AT	6.28E-05	Negative
EGFR	GSE30929	whole-transcript expression data for liposarcoma	liposarcoma	211607_X_AT	0.000485	Positive
EGFR	GSE25065	validation cohort for genomic predictor of response and survival following neoadjuvant taxane-anthracycline chemotherapy in breast cancer	breast cancer	201983_S_AT	0.0017	Negative
EGFR	GSE10846	prediction of survival in diffuse large b cell lymphoma treated with chemotherapy plus rituximab	diffuse large b cell lymphoma	201983_S_AT	0.00239	Positive
EGFR	GSE30682	search for a gene-expression signature of breast cancer local recurrence in young women	breast cancer	ILMN_1798975	0.00455	Negative
EGFR	GSE14764	a prognostic gene expression index in ovarian cancer	ovarian cancer	211551_AT	0.00523	Negative
EGFR	GSE17538	experimentally derived metastasis gene expression profile predicts recurrence and death in colon cancer patients	colon cancer	210984_X_AT	0.00532	Positive
EGFR	GSE22762	an eight-gene expression signature for the prediction of survival and time to treatment in chronic lymphocytic leukemia	chronic lymphocytic leukemia	201983_S_AT	0.00832	Negative
EGFR	GSE17536	metastasis gene expression profile predicts recurrence and death in colon cancer patients (moffitt samples)	colon cancer	210984_X_AT	0.0113	Positive
EGFR	GSE19783	molecular characterization of breast cancer subtypes derived from joint analysis of high throughput mirna and mrna data	breast cancer	A_23_P215790	0.0126	Negative
EGFR	GSE13213	relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis	lung cancer	A_23_P215790	0.0213	Negative
EGFR	GSE21653	a gene expression signature identifies two prognostic subgroups of basal breast cancer	breast cancer	210984_X_AT	0.0383	Positive
EGFR	GSE18166	genome-wide profiling of astrocytic gliomas	astrocytic gliomas	34391	0.0405	Positive
EGFR	GSE24080	maqc-ii project: multiple myeloma (mm) data set	multiple myeloma	210984_X_AT	0.0446	Positive
EGFR	GSE17705	endocrine sensitivity index validation dataset	breast cancer	201984_S_AT	0.0486	Positive

GENE (Probe ID): the probe identity of datasets in which expression of input gene is significantly associated with the clinical outcome of patients with tumor; Effect sign, indicate whether the input gene is positively or negatively associated with the clinical outcome of patients .

Table 2b. Correlation of interacted partner genes of 3 hub genes with clinical outcome of patients with tumors

Gene	Interaction source	Number of partners	Significant (p-value<0.05) hits	Best p-value
RB1	Reactome, HPRD, NCI_ NATURE, IntAct, PhosphoSitePlus	220	17	2.41E-06
EGFR	Reactome, HPRD, NCI_ NATURE ,IntAct, PhosphoSitePlus, HumanCyc	1385	14	0.00015
HSP90AA1	Reactome, HPRD, NCI_ NATURE, IntAct	382	11	4.12E-08

Interaction source, interaction data source; Number of partners, the number of interaction partner genes of the input gene; significant hits: the number of tumor gene expression datasets with survival rates-related genome overlapped with genes that encoded interaction partner proteins; Best P-value, the minimum p-value of all p-values arising from datasets with overlapped genome;

By means of literature-mining, we found that up-regulated *angiopoietin (ANGPT)-1* and *ANGPT2* have been reported to be associated with prognosis of OSCC patients. *ANGPT2* expression displayed marked difference (logFC = 1.124, adjusted p-value= $1.03e^{-20}$) between OSCC and normal samples, while the expression difference of *ANGPT1* was relatively small (logFC = 0.204, adjusted p-value = 0.025), indicating that *ANGPT2* might be a critical gene in OSCC.

PPI network and PPISURV analysis. The constructed PPI network contained 332 proteins encoding by OSCCDEG and 1059 pairs of protein interactions. The top 3 proteins sorted by degree were epidermal growth factor receptor (EGFR, degree = 41), heat shock protein 90AA1 (HSP90AA1, degree = 41), retinoblastoma protein1 (RB1, degree = 35). Besides, there were direct links between ANGPT1 and integrin alpha-5 (ITGA5), and between ANGPT2 and vascular endothelial growth factor A (VEGFA) (Supplementary Figure 1).

As shown in Table 2a, the result of PPISURV analysis delineated that the 3 proteins were associated with the clinical outcome of patients with diverse tumors. Of note is that the expression of EGFR was significantly associated with the survival rates of patients in 16 gene expression datasets covering various tumors such as colorectal cancer, breast cancer and lung cancer. In 8 of the 16 gene expression datasets, EGFR was negatively correlated with prognosis of patients. In addition to that, interacted partner proteins of EGFR were remarkably overlapped with the survival rates-related genome of 14 tumor gene expression datasets (Table 2b).

Moreover, ANGPT2 was associated with prognosis of patients in 8 gene expression datasets. In 7 of the 8 gene expression datasets negative associations were observed. ANGPT1 was associated with prognosis of patients in 10 datasets with negative associations observed in 4 datasets.

PPI network module identification and functional analysis. From the PPI network, three network modules were identified. Among the three modules, the largest module included 24 proteins and 42 pairs of protein interactions, in which EGFR was the hub protein (Figure 1) and was significantly associated with positive regulation of cell communication according to the result of GO function analysis

Figure 1. The largest network module with hub protein EGFR. White node represents protein encoded by up-regulated gene; Gray node represents protein encoded by down-regulated gene; a link represents each pairwise protein interaction with interaction score≥ 0.4.

Figure 2. The second largest network module with hub protein FN1. White round node represents protein encoded by up-regulated gene; Gray node represents protein encoded by down-regulated gene; a link represents each pairwise protein interaction with interaction score≥ 0.4.

Figure 3. The third network module with hub protein PCNA. White round node represents protein encoded by up-regulated gene; Gray node represents protein encoded by down-regulated gene; a link represents each pairwise protein interaction with interaction score≥ 0.4.

704 genes related to OSCC from MalaCards database, out of which 302 were up-regulated, while 149 were down-regulated. Furthermore, 451 DEGs overlapped with the 704 genes were defined as "OSCCDEG". Up-regulated OS-CCDEG were significantly enriched in cell cycle pathway, while down-regulated OSCCDEGs were primarily associated with ERBB pathway. In the PPI network, 3 hub proteins were identified: EGFR1, RB1 and HSP90AA1. The result of PPISURV analysis showed that the 3 proteins were associated with the survival rates of several tumors. From the PPI network, three network modules were identified, followed by GO enrichment analysis.

EGFR, also termed ErbB-1, belongs to ErbB family and is the cell-surface receptor that binds to EGF. It has long been established that EGFR could serve as predictor for the prognosis of patients with OSCC [22]. Furthermore, a 2012 report has found that overexpression of EGFR located on membrane and in cytoplasm might be closely associated with the prognosis of OSCC at early stage [23]. Similarly, in the study, the result of PPISURV analysis showed that the expression of EGFR was significantly associated with the survival rates of various types of tumors such as colorectal cancer, breast cancer and lung cancer. Besides, the involvement of EGFR in OSCC has been confirmed by other analyses, in which EGFR was hub node in the PPI network and the first largest module. Genes of the first largest module were linked to positive regulation of cell communication. Consistently, ErbB signaling pathway was associated with down-regulated OSCCDEG. These findings suggest that EGFR might participate in modulating cell communication that affects the OSCC development.

Fibronectin1 encoded by gene *FN1* is an important constituent of ECM that binds to integrin. It takes part in a diversity of activities such as cell adhesion, proliferation and differentiation [24]. Kaspar *et al.* have reported that deregulated expression of fibronectin was partly responsible for morphological alterations of tumors and involved in the development of carcinoma [25]. In OSCC, fibronectin1 participates in regulating the migration and invasion of cancer cells [26], and could serve as a potential biomarker for this carcinoma [27]. Consistently, this study showed that *FN1* was identified as a hub gene in the second largest network module extracted from the PPI network, which was closely associated with angiogenesis. These findings

(Table 3, adjusted P-value = $8.3e^{-5}$). As shown in Figure 2, the second largest module had 13 proteins and 16 pairs of protein interactions, with fibronectin1 (FN1) as the hub node. Thrombospondin (THBS1) was also included in the module. Genes involved in the module was significantly enriched in the GO function term, angiogenesis (adjusted P-value = $8.3e^{-5}$). The third one included 13 protein interaction pairs and 11 proteins among which the hub protein was proliferating cell nuclear antigen (PNCA) (Figure 3). The third network module was significantly related with DNA metabolic process (adjusted P-value = $6.8e^{-6}$).

Discussion

OSCC is linked to remarkable mortality and morbidity, persecuting the health of people. In this study, 5177 up-regulated DEGs and 6081 down-regulated DEGs were identified between OSCC and control samples. We gained

Table 3. GO biological processes significantly associated with 3 network modules

Module	p-value	Corrected p-value	Number of genes	GO Biological Process	Genes in test set
module 1	2.91E-13	2.86E-10	12	positive regulation of cell communication	EGFR|IGF1R|ERBB4|ERBB3|RIPK1|ERBB2|LGALS1|TLR2|TGFA|ADAM17|KIT|EGF
module 2	1.52E-07	8.38E-05	5	angiogenesis	COL4A1|PDGFA|ITGA5|THBS1|FN1
module 3	2.25E-08	6.89E-06	7	DNA metabolic process	CDC7|TYMS|CDC6|XPC|MCM7|MGMT|PCNA

GO, gene ontology

lead to a conclusion that FN1 might affect OSCC-related angiogenesis.

ANGPT1 and ANGPT2 were angiopoietins encoded by *ANGPT1* and *ANGPT2*, respectively, both of which interacts with TEK receptor tyrosine kinase and are involved in angiogenesis and vascular remodeling [28, 29]. Emerging studies have suggested that expression of ANGPT2 was related to the OSCC-related angiogenesis, vessel maturation, lymphangiogenesis and prognosis of patients with OSCC [30, 31]. Consistently, in the current study, *ANGPT2* and *ANGPT1* was identified DEGs for OSCC, and both were associated with prognosis of tumor patients according to the result of PPISURV analysis. Besides, interactions between ANGPT1 and ITGA5, and between ANGPT2 and VEGFA were observed in the PPI network. Moreover, in the second largest network module, THBS1 was also observed, directly interacting with FN1. THBS1 was an adhesive glycoprotein, participating in angiogenesis as well. It has been found that in the remodeling of heart, disruption of THBS1 results in inhibition on the up-regulation of ANGPT2, whereas THBS1 stimulation leads to increase of ANGPT2 in macrophage [32]. There findings reveal that THBS might regulate the expression of ANGPT2 and VEGFA, and then instruct the angiogenesis in OSCC.

Proliferating cell nuclear antigen (PCNA) is a homotrimer that acts as a processivity factor involved in DNA replication and DNA repair. PCNA protein resides in nucleus, and functions as a cofactor of DNA polymerase delta [33]. It has long been reported that PCNA could serve as a proliferation marker of OSCC [34]. Furthermore, a recent report suggested that PCNA expression located at the invasive front of OSCC was closely associated with the prognosis of patients [35]. In consistence, in the present study PCNA was determined as the hub protein in the third network module in which the genes were significantly linked to DNA metabolic process. We could infer that PCNA might affect the proliferation of cancer cells via modulating DNA metabolism.

Conclusion

In summary, this study suggested that ErbB pathway and cell cycle pathway might play major roles in OSCC, and *EGFR*, *FN1*, *PCNA*, *ANGPT1* and *ANGPT2* could be potential biomarkers for prognosis prediction of OSCC. These findings would contribute to the development of novel potent therapies that target the molecular biomarkers identified in this study. Inevitably, there are limitations in this study due to the small sample size. *In vitro* and *in vivo* studies are required to further validate the results from this study.

References

[1] FERLAY J, SOERJOMATARAM I, ERVIK M, DIKSHIT R, ESER S et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Lyon, France: International Agency for Research on Cancer; 2013. Last accessed April 2014.

[2] SALAHSHOURIFAR I, VINCENT-CHONG VK, KALLARAKKAL TG, ZAIN RB. Genomic DNA copy number alterations from precursor oral lesions to oral squamous cell carcinoma. Oral Oncol 2014; 50: 404–412. https://doi.org/10.1016/j.oraloncology.2014.02.005

[3] ESMAELBEIGI F, HADJI M, HARIRCHI I, OMRANIPOUR R, VAND RM et al. Factors affecting professional delay in diagnosis and treatment of oral cancer in iran. Arch Iran Med 2014; 17: 253–257

[4] SCOTT S, MCGURK M, GRUNFELD E. Patient delay for potentially malignant oral symptoms. Eur J Oral Sci 2008; 116: 141–147. https://doi.org/10.1111/j.1600-0722.2007.00520.x

[5] SHIN K-H, BAE SD, HONG HS, KIM RH, KANG MK et al. miR-181a shows tumor suppressive effect against oral squamous cell carcinoma cells by downregulating K-ras. Biochem Biophys Res Commun 2011; 404: 896–902. https://doi.org/10.1016/j.bbrc.2010.12.055

[6] SASAHIRA T, UEDA N, YAMAMOTO K, KURIHARA M, MATSUSHIMA S et al. Prox1 and FOXC2 Act as Regulators of Lymphangiogenesis and Angiogenesis in Oral Squamous Cell Carcinoma. PloS One 2014; 9: e92534. https://doi.org/10.1371/journal.pone.0092534

[7] LAI MT, HUA CH, TSAI MH, WAN L, LIN YJ et al. Talin-1 overexpression defines high risk for aggressive oral squamous cell carcinoma and promotes cancer metastasis. J Pathol 2011; 224: 367–376. https://doi.org/10.1002/path.2867

[8] COHEN Y, GOLDENBERG-COHEN N, SHALMON B, SHANI T, OREN S et al. Mutational analysis of PTEN/PIK3CA/AKT pathway in oral squamous cell carcinoma. Oral Oncol 2011; 47: 946–950. https://doi.org/10.1016/j.oraloncology.2011.07.013

[9] PARIKH RA, APPLEMAN LJ, BAUMAN JE, SANKUNNY M, LEWIS DW et al. Upregulation of the ATR-CHEK1 pathway in oral squamous cell carcinomas. Genes Chromosomes Cancer 2014; 53: 25–37. https://doi.org/10.1002/gcc.22115

[10] CHEN C, MENDEZ E, HOUCK J, FAN W, LOHAVANICHBUTR P et al. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev 2008; 17: 2152–2162. https://doi.org/10.1158/1055-9965.EPI-07-2893

[11] IRIZARRY RA, HOBBS B, COLLIN F, BEAZERBARCLAY YD, ANTONELLIS KJ et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4: 249–264. https://doi.org/10.1093/biostatistics/4.2.249

[12] RITCHIE ME, PHIPSON B, WU D, HU Y, LAW CW et al. limma powers differential expression analyses for RNA-

sequencing and microarray studies. Nucleic Acids Res 2015; 43: e47. https://doi.org/10.1093/nar/gkv007

[13] RAPPAPORT N, NATIV N, STELZER G, TWIK M, GUAN-GOLAN Y et al. MalaCards: an integrated compendium for diseases and their annotation. Database (Oxford) 2013; 2013:bat018, https://doi.org/10.1093/database/bat018

[14] HUANG DA W, SHERMAN BT, LEMPICKI RA. Systematic and integrative analysis of large gene lists using DAVID bio-informatics resources. Nat Protoc 2009; 4: 44–57. https://doi.org/10.1038/nprot.2008.211

[15] ARAKAWA K, KONO N, YAMADA Y, MORI H, TOMITA M. KEGG-based pathway visualization tool for complex omics data. In Silico Biol 2005; 5: 419–423.

[16] FRANCESCHINI A, SZKLARCZYK D, FRANKILD S, KUHN M, SIMONOVIC M et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 2013; 41: D808–815. https://doi.org/10.1093/nar/gks1094

[17] ANTONOV AV, KRESTYANINOVA M, KNIGHT RA, RODCHENKOV I, MELINO G et al. PPISURV: a novel bioin-formatics tool for uncovering the hidden role of specific genes in cancer survival outcome. Oncogene 2014; 33: 1621–1628. https://doi.org/10.1038/onc.2013.119

[18] MORRIS JH, APELTSIN L, NEWMAN AM, BAUMBACH J, WITTKOP T et al. clusterMaker: a multi-algorithm cluster-ing plugin for Cytoscape. BMC Bioinformatics 2011; 12: 436. https://doi.org/10.1186/1471-2105-12-436

[19] ENRIGHT AJ, VAN DONGEN S, OUZOUNIS CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res 2002; 30: 1575–1584. https://doi.org/10.1093/nar/30.7.1575

[20] ASHBURNER M, BALL CA, BLAKE JA, BOTSTEIN D, BUTLER H et al. Gene Ontology: tool for the unifica-tion of biology. Nat Genet 2000; 25: 25–29. https://doi.org/10.1038/75556

[21] MAERE S, HEYMANS K, KUIPER M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 2005; 21: 3448–3449. https://doi.org/10.1093/bioinformatics/bti551

[22] XIA W, LAU Y-K, ZHANG H-Z, XIAO F-Y, JOHNSTON DA et al. Combination of EGFR, HER-2/neu, and HER-3 is a stronger predictor for the outcome of oral squamous cell carcinoma than any individual family members. Clin Cancer Res 1999; 5: 4164–4174.

[23] MONTEIRO LS, DINIZ-FREITAS M, GARCIA-CABALLE-RO T, WARNAKULASURIYA S, FORTEZA J et al. Combined cytoplasmic and membranous EGFR and p53 overexpression is a poor prognostic marker in early stage oral squamous cell carcinoma. J Oral Pathol Med 2012; 41: 559–567. https://doi.org/10.1111/j.1600-0714.2012.01142.x

[24] PANKOV R, YAMADA KM. Fibronectin at a glance. J Cell Sci 2002; 115: 3861–3863. https://doi.org/10.1242/jcs.00059

[25] KASPAR M, ZARDI L, NERI D. Fibronectin as target for tumor therapy. Int J Cancer 2006; 118: 1331–1339. https://doi.org/10.1002/ijc.21677

[26] KAMARAJAN P, GARCIA-PARDO A, D'SILVA NJ, KAPILA YL. The CS1 segment of fibronectin is involved in human OSCC pathogenesis by mediating OSCC cell spreading, migration, and invasion. BMC Cancer 2010; 10: 330. https://doi.org/10.1186/1471-2407-10-330

[27] YEN CY, HUANG CY, HOU MF, YANG YH, CHANG CH et al. Evaluating the performance of fibronectin 1 (FN1), integrin α4β1 (ITGA4), syndecan-2 (SDC2), and glycoprotein CD44 as the potential biomarkers of oral squamous cell carcinoma (OSCC). Biomarkers 2013; 18: 63–72. https://doi.org/10.3109/1354750X.2012.737025

[28] THURSTON G, RUDGE JS, IOFFE E, ZHOU H, ROSS L et al. Angiopoietin-1 protects the adult vasculature against plasma leakage. Nat Med 2000; 6: 460–463. https://doi.org/10.1038/74725

[29] GALE NW, THURSTON G, HACKETT SF, RENARD R, WANG Q et al. Angiopoietin-2 is required for postnatal angiogenesis and lymphatic patterning, and only the latter role is rescued by Angiopoietin-1. Dev Cell 2002; 3: 411–423. https://doi.org/10.1016/S1534-5807(02)00217-4

[30] LI C, SUN C-J, FAN J-C, GENG N, LI C-H et al. Angiopoi-etin-2 expression is correlated with angiogenesis and overall survival in oral squamous cell carcinoma. Med Oncol 2013; 30: 571. https://doi.org/10.1007/s12032-013-0571-2

[31] LI C, FAN J, SONG X, ZHANG B, CHEN Y et al. Expression of angiopoietin-2 and vascular endothelial growth factor recep-tor-3 correlates with lymphangiogenesis and angiogenesis and affects survival of oral squamous cell carcinoma. PloS One 2013; 8: e75388. https://doi.org/10.1371/journal.pone.0075388

[32] GONZALEZ-QUESADA C, CAVALERA M, BIERNACKA A, KONG P, LEE D-W et al. Thrombospondin-1 induction in the diabetic myocardium stabilizes the cardiac matrix in addition to promoting vascular rarefaction through angiopoietin-2 upregulation. Circ Res 2013; 113: 1331–1344. https://doi.org/10.1161/CIRCRESAHA.113.302593

[33] MAGA G, H BSCHER U. Proliferating cell nuclear antigen (PCNA): a dancer with many partners. J Cell Sci 2003; 116: 3051–3060. https://doi.org/10.1242/jcs.00653

[34] MYOUNG H, KIM MJ, LEE JH, OK YJ, PAENG JY et al. Correlation of proliferative markers (Ki-67 and PCNA) with survival and lymph node metastasis in oral squamous cell carcinoma: a clinical and histopathological analysis of 113 patients. Int J Oral Maxillofac Surg 2006; 35: 1005–1010. https://doi.org/10.1016/j.ijom.2006.07.016

[35] KATO K, KAWASHIRI S, YOSHIZAWA K, KITAHARA H, OKAMUNE A et al. Expression form of p53 and PCNA at the invasive front in oral squamous cell carcinoma: correlation with clinicopathological features and prognosis. J Oral Pathol Med 2011; 40: 693–698. https://doi.org/10.1111/j.1600-0714.2011.01032.x
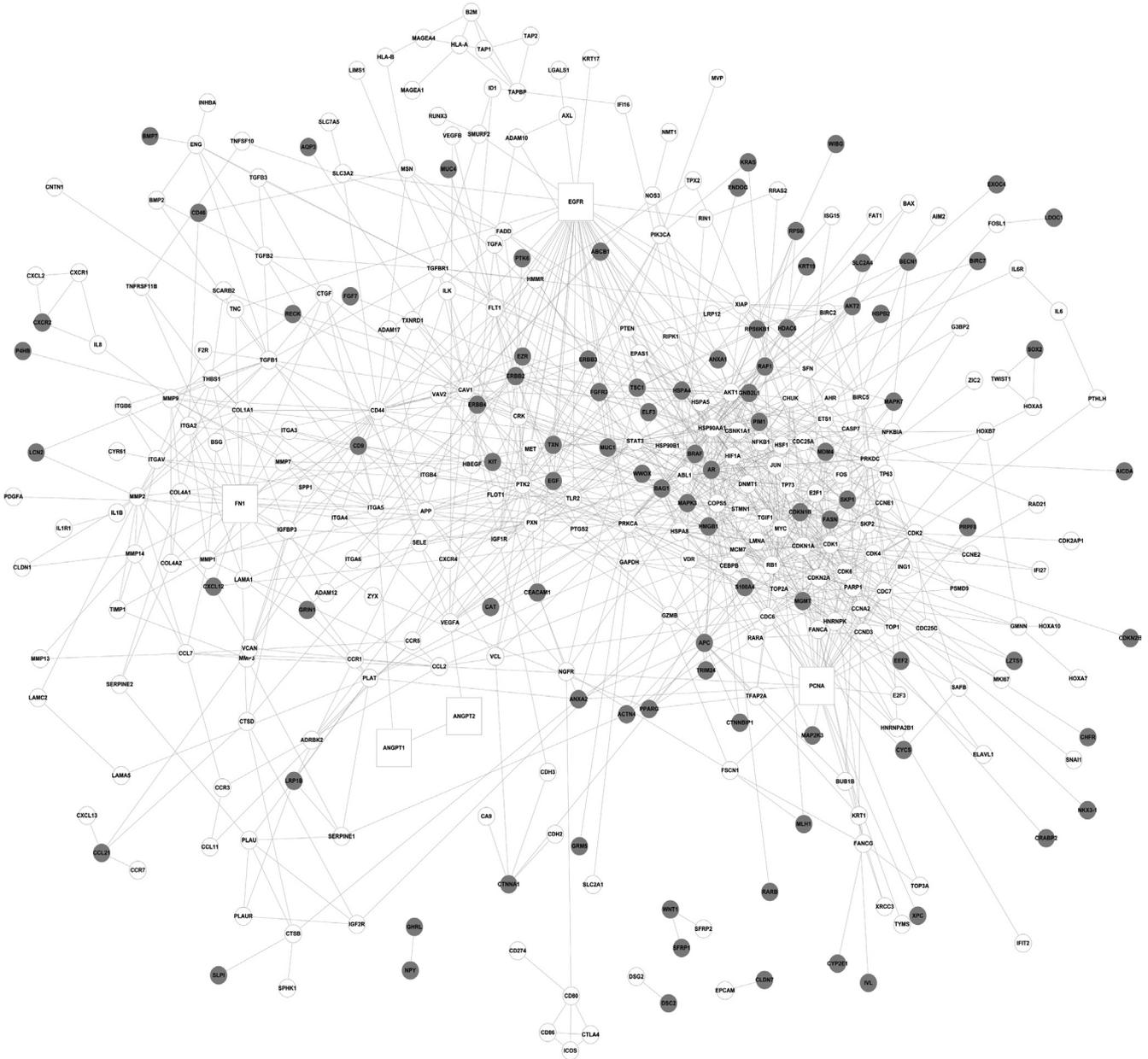
Supplementary Figure 1. Protein-protein interaction network. White nodes denote proteins encoded by up-regulated DEGs; Gray nodes denote pro¬teins encoded by down-regulated DEGs; Square nodes denote important proteins; a link denotes an interaction between proteins with interaction score ≥ 0.4.