Journal of Genetic Engineering and Biotechnology 24 (2026) 100668

Contents lists available at ScienceDirect

Journal of Genetic Engineering and Biotechnology

journal homepage: www.elsevier.com/locate/jgeb

Identification of potential key genes and molecular mechanisms of oral squamous cell carcinoma based on integrated bioinformatics approach

Mohammad Khursheed Alam [a,h,*], Md Rakibul Islam [b,h], Tahsinul Haque [c], Kiran Kumar Ganji [a], Hamzah Ali Babkair [d], Mohammed Enamur Rashid [d], Kawsar Ahmed [e,f,g], Francis M. Bui [f]

[a] Preventive Dentistry Department, College of Dentistry, Jouf University, Sakaka 72345, Saudi Arabia
[b] Department of Software Engineering, Daffodil International University, Daffodil Smart City, Ashulia, Savar, Dhaka, Bangladesh
[c] Department of Preventive Dental Sciences, College of Dentistry, Dar Al Uloom University, Riyadh 13314, Saudi Arabia
[d] Department of Oral and Maxillofacial Diagnostic Sciences, Taibah University, Madinah, Saudi Arabia
[e] Health Informatics Research Lab, Department of Computer Science and Engineering, Daffodil International University, Daffodil Smart City, Birulia, Dhaka 1216, Bangladesh
[f] Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada
[g] Group of Biophotomatiχ, Dept. of ICT, MBSTU, Santosh, Tangail 1902, Bangladesh

ARTICLE INFO

Keywords:
Bioinformatics
Bioinformatics analysis
Key genes
Molecular mechanisms
Oral squamous cell carcinoma

ABSTRACT

Oral Squamous Cell Carcinoma (OSCC) is one of the most occurred cancer types with yearly 377,713 cases and 177,000 deaths. Traditional risk factors of OSCC include smoking, alcohol consumption, excessive sun exposure, family history of cancer, and human papillomavirus (HPV). Last few years, the prevalence of OSCC is growing big in numbers particularly among younger people for their lifestyle. From the Gene Expression Omnibus, 2 gene-expression profiles (GSE23558 and GSE146483) were identified based on some conditions. The GEO2R tool was used to analyze those datasets to extract all the genes. Statistical cut-off criteria were applied to find out DEGs from both datasets, and after that common DEGs were identified by comparing both datasets. Common DEGs were used to perform bioinformatics analysis such as gene ontology and pathway analysis, protein–protein interaction (PPI) network construction, and generating Transcription factor – miRNA network. 265 common DEGs were identified from the datasets including 69 up-regulated and 196 down-regulated DEGs. Using the STRING database and a strong combine score > 0.70, a PPI network is generated including 92 nodes and 226 interactions. Using 3 different hub DEGs seeking algorithm, we identified 9 top hub DEGs. The hub genes are Kinesin Family Member 23 (KIF23), Aurora Kinase A (AURKA), Centromere Protein F (CENPF), Cell Division Cycle 20 (CDC20), Discs Large Associated Protein 5 (DLGAP5), Centrosomal Protein 55 (CEP55), Anillin Actin Binding Protein (ANLN), Non-SMC Condensin I Complex Subunit G (NCAPG), and Kinesin Family Member 14 (KIF14). 3 significant clusters also identified from the PPI network. Previous study shows KIF23 takes part in raising Cell Proliferation in Hepatocellular carcinoma cells and AURKA shows notable overexpression in cancer tissues, which indicates that KIF23 and AURKA showed promising character to become possible biomarkers for OSCC. Further analysis needed to justify the statement.

1. Introduction

Oral carcinoma is mostly designated as oral squamous cell carcinoma (OSCC) and ensues as an ulceroproliferative sore of the oral mucosa affecting any part starting from the lips, tongue, cheeks, floor of the mouth, hard and soft palate, sinuses, and pharynx.[1] According to the GLOBOCAN 2020 report, 377,713 OSCC cases were diagnosed and more than 177,000 deaths occurred. OSCC mostly occurred in men, almost

* Corresponding author.
E-mail addresses: mkalam@ju.edu.sa (M.K. Alam), rakibul35-116@diu.edu.bd (M.R. Islam), mdhaque@dau.edu.sa (T. Haque), dr.kiran.ganji@jodent.org (K.K. Ganji), hbabkair@taibahu.edu.sa (H.A. Babkair), mrashid@taibahu.edu.sa (M.E. Rashid), k.ahmed.bd@ieee.org, kawsar.ict@mbstu.ac.bd, k.ahmed@usask.ca (K. Ahmed), francis.bui@usask.ca (F.M. Bui).
[h] These authors contributed equally to this work and Joint First Author.

https://doi.org/10.1016/j.jgeb.2026.100668
Received 15 October 2024; Received in revised form 27 October 2025; Accepted 24 February 2026

70% of the total cases diagnosed patients were male.[2] Traditional risk factors of OSCC include smoking, alcohol consumption, excessive sun exposure, family history of cancer, and human papillomavirus (HPV). In recent years the prevalence of OSCC is growing big in numbers particularly among younger people for their lifestyle. Despite noticeable improvements in cancer treatment and surgical methods, the 5-year survival rate for patients with OSCC did not show any prosperity in the past two decades.[3] The overall five-year rate of survival for OSCC is somewhere between 50 and 60%, ordinarily; such an impoverished prognosis for OSCC is commonly reported by narration at a late stage of the illness.[4]

Bioinformatics is a field where biology adopts computer science and introduced an emerging technology. Bioinformatics opens plenty of opportunities for researchers, it allows researchers to manage big data in an efficient manner, and now it is easier than ever to access and add new data as they are produced; using bioinformatics techniques lots of tools and resources created to fast the process of data analysis. Various types of biological data can be ingressed using bioinformatics such as expression and sequencing results on a large-scale allowing examination on a whole-genome and whole-exome level in this manner.[5–7]

For the past few years, microarray technology-based high-throughput platforms evolved as reliable and also proficient tools in search of significant genes and epigenetic modifications in carcinogenesis. Currently, microarray technology is largely used to identification of biomarkers for disease diagnosis and prognosis. Gene expression microarrays and gene chips are wide-ranging and proficient observation techniques, which are generally applied to observe genome-wide expression levels of genes in a specific organism or cells, and especially competent for screening differentially expressed genes (DEGs) between two or more samples. Many studies used gene-expression-profile microarrays to explore DEGs of any disease or cancer. In 2020, Zhu et al. studied two different gene-expression-profile (GSE10121 and GSE85195) and got ATG12, BID, NKX2-3 and SPHK1 were differentially expressed among normal tissues and oral cancer tissues. By applying advanced and integrated bioinformatics methods for publicly accessible microarray profiles, more accurate and reliable outcomes may be explored through similar gene-expression profiles, especially in cancer research. Recently, Du et al. identified ANPEP, APOB, GLP1R, and SI as potential biomarkers of oral squamous cell carcinoma (OSCC) using whole transcriptome sequencing and multi-algorithm bioinformatics, providing novel insights into disease mechanisms and therapeutic strategies.[8] In addition, cuproptosis-related lncRNAs (CRLs) have emerged as important regulators of OSCC prognosis and treatment; eight CRLs were recently identified to construct a predictive model for clinical outcomes, immune response, and drug sensitivity, offering new perspectives on immune escape and tumor drug resistance.[9] Furthermore, recent work highlighted the significance of pyroptosis-related genes (CTLA4, CD5, and IL12RB2) in OSCC prognosis, establishing a robust model that links pyroptosis activity with immune infiltration and patient survival outcomes, thereby opening new avenues for targeted therapies.[10]

In the current study, we collect two gene-expression microarray datasets (GSE23558 and GSE146483) from the freely accessible repository gene expression omnibus. These two datasets contained a total of 43 samples (35 OSCC samples, and 8 normal samples). The DEGs were extracted from the samples and screened. Gene Ontology (GO) and pathway enrichment analysis were accomplished using a prominent and reputed online tool. Using the DEGs protein–protein interaction (PPI) network was generated, and the hub candidate DEGs were identified using several methods. A Transcription factor and microRNA combined network were also generated to know more biological function details of identified DEGs. Combining multiple datasets and finding hub genes from encompassing several methods and exploring Transcription factor and microRNA combined network in a single study is very rare in OSCC related bioinformatics analysis. We carefully done all the responsible analysis and extract all the results followed by standard conditions. The outcomes of this study may work as potential biomarkers for early detection or as auspicious targets for the treatment of OSCC.

2. Materials and methods

2.1. The study workflow

The biological characteristics of DEGs related to OSCC are revealed in this study by an extensive bioinformatics approach. All the applied workflow of this study is displayed in Fig. 1.

2.2. Gene expression profile dataset collection

The publicly available and well-founded repository the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/) was accounted for to collect datasets.[11] We searched the GEO repository employing the following keywords: study keyword “Oral Squamous Cell Carcinoma (OSCC)”, organism “Homo Sapiens”, study type “Expression profiling by array”, and a minimum 10 sample size. After a systematic assessment, two gene expression profiles GSE23558 and GSE146483 were elected for bioinformatics study analysis. The GSE23558 dataset contained a total of 32 samples including 5 normal samples and 27 OSCC samples,[12] the dataset was based on the GPL6480 platform (Agilent-014850 Whole Human Genome Microarray 4x44K G4112F). On the other hand, the GSE146483 dataset holds a total of 11 samples including 3 normal samples and 8 OSCC samples,[13] the dataset used the GPL17077 platform (Agilent-039494 SurePrint G3 Human GE v2 8x60K Microarray 039381). In Table 1, the insights of the samples from the identified datasets are shown.

2.3. Dataset preprocessing and differentially expressed gene screening

The DEGs associated with the samples of GSE23558 and GSE146483 were obtained by using GEO2R. GEO2R is a web tool that uses R programming and the limma package to analyze gene expression profiles. To obtain DEGs we used cut-off criteria for adjusted $|\log \text{Fold Change (FC)}| > 2$ and p-value < 0.05. A web tool InteractiVenn (https://www.interactivenn.net/) was employed to show common DEGs between the two datasets.[14]

2.4. Enrichment analysis of DEGs

To uncover the functional insights of identified DEGs, the Gene Ontology (GO) functional outcomes of the biological process (BP), cellular component (CC), and molecular function (MF) were explored by the online annotation tool Database for Annotation, Visualization and Integrated Discovery (DAVID) (https://david.ncifcrf.gov/).[15] DAVID tool also used to complete pathway analysis. Reactome Pathway Database was employed to identify the key pathways.[16] Threshold criteria p-value < 0.05 was set for GO and Pathway terms.

2.5. Protein-protein network construction and hub gene selection

The STRING database (https://string-db.org/) was used to explore internal interactions between all the common DEGs.[17] A strong combine score > 0.70 is used to maintain rigorous analysis. Cytoscape (v3.9.1) software an open-source tool to visualize and analyze biological networks were employed to generate the protein–protein interaction (PPI) networks.[18] To analyze essential topological parameters of the PPI network we use the CytoHubba plugin (a custom package of Cytoscape software).[19] CytoHubba can perform 11 different kinds of topological analysis to identify hub genes from a network. In this study, we have identified hub genes based on three different topological characteristics, according to top 10 Node degree value, top 10 Maximal Clique Centrality (MCC) value and top 10 Maximum Neighborhood Component (MNC) value. To spot clusters from PPI networks, we applied the

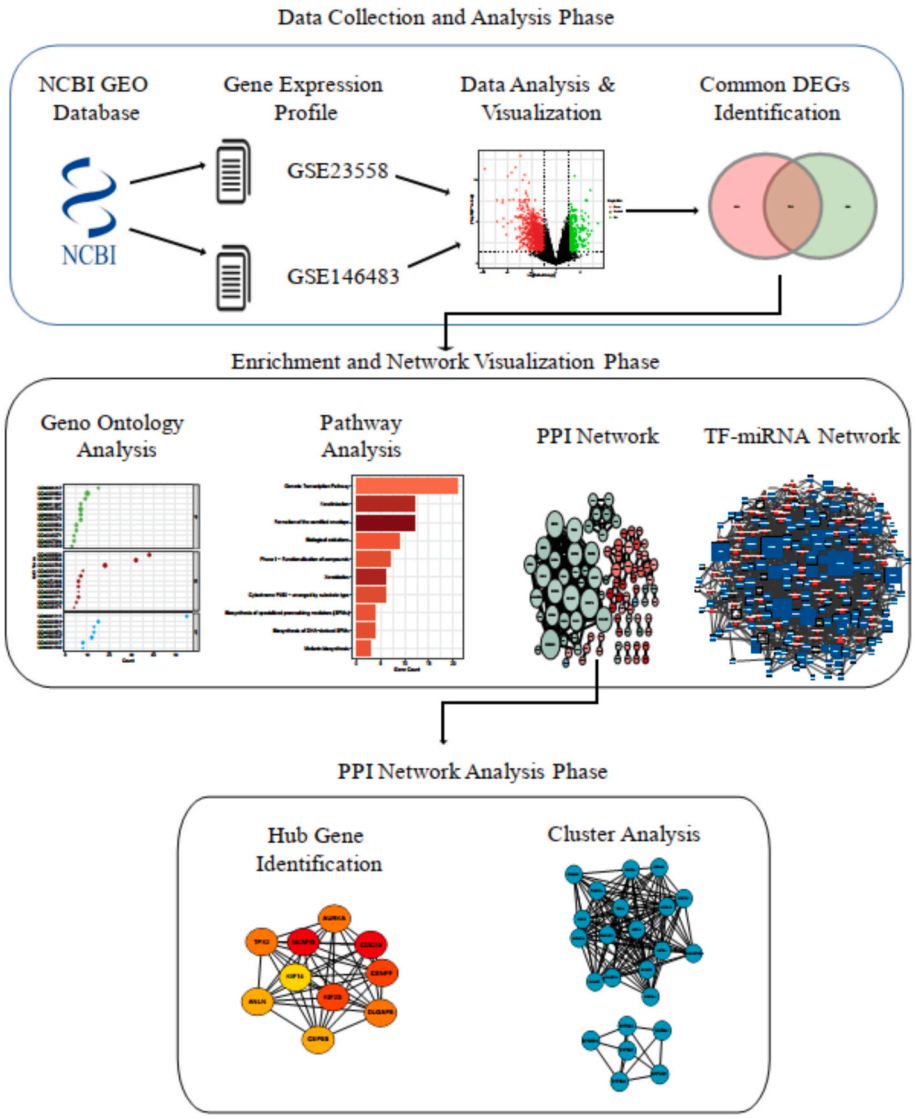

Fig. 1. Pictorial representation of the study. This study combines three different phases, all the phases are covered with step-by-step procedures.

Table 1
Basic information of sample distribution of two gene-expression profiles.

Accession Number	Total Sample	Normal Sample	OSCC Sample	Platform
GSE23558	32	5 (15.6%)	27 (84.4%)	GPL6480
GSE146483	11	3 (27.3%)	8 (72.7%)	GPL17077
Total	43	8 (18.6%)	35 (81.6%)	

Molecular Complex Detection (MCODE) algorithm.[20] We applied the default parameters of the MCODE plugin to identify clusters, the parameters were degree cutoff = 2, node score cutoff = 0:2, k − core = 2, and maximum depth = 100. After exploring a few clusters, a functional pathway analysis was performed by using the REACTOME database.

2.6. TF-miRNA network analysis

We generate a Transcription factor (TF) and microRNA (miRNA) combined network named TF-miRNA network using common DEGs. The RegNetwork database was used to generate the TF-miRNA network.[21] Minimum connectivity >= 2 was followed for the network creation. Cytoscape software was used to display the network.

3. Result and analysis

3.1. Differentially expressed gene screening

The gene expression profiles GSE23558 and GSE146483 were selected from the GEO repository and an amount of 35 OSCC samples were compromised. In Table 1 basic information about both datasets is displayed. The datasets were analyzed by the GEO2R tool. Maintaining the predestine cut-off criteria, 1453 DEGs were marked out from the GSE23558 dataset involving 1120 up-regulated DEGs and 333 down-regulated DEGs. And from the GSE146483 dataset, we got 1870 DEGs marked out involving 1120 up-regulated DEGs and 720 down-regulated

Table 2
DEGs distribution of identified profiles before and after applying cut-off criteria.

Accession Number	Total DEGs	Up-regulated DEGs	Down-regulated DEGs
(A) Before applying the cut-off criteria			
GSE23558	19,563	9514 (48.63%)	10,049 (51.37%)
GSE146483	23,267	13,670 (58.75%)	9597 (41.25%)
(B) After applying (|logFC| > 2 and adjusted P-value < 0.05)			
GSE23558	1453	333 (22.92%)	1120 (77.08%)
GSE146483	1870	720(38.50%)	1150 (61.50%)
Total Common DEGs	265	69(26.04%)	196 (73.96%)

DEGs. In Table 2 and Fig. 2 (A – B), we gave more textual and pictorial representations of the datasets. Considering the web tool InteractiVenn the common up-regulated and down-regulated DEGs were identified. A total of 265 DEGs were identified including 69 up-regulated and 196 down-regulated. In Fig. 2 (C – D) using the Venn diagram the result is displayed.

3.2. Enrichment analysis of DEGs

Identified common DEGs were submitted to the DAVID online tool to perform enrichment analysis. The GO analysis was performed for up-regulated and down-regulated DEGs separately. We choose the top 30 GO terms based on the least p-value. GO analysis reveals the up-regulated DEGs were greatly connected with anterior/posterior pattern specification, embryonic skeletal system morphogenesis, and proximal/distal pattern formation for BP; chromatin, nucleoplasm, and nucleus for CC; sequence-specific double-stranded DNA binding, microtubule binding, and RNA polymerase II transcription factor activity, sequence-specific DNA binding for MF. Table 3 and Fig. 3 demonstrate the full GO analysis result of up-regulated DEGs. Yet GO analysis reveals the down-regulated DEGs were greatly connected with keratinization, regulation of transcription from RNA polymerase II promoter, and xenobiotic metabolic process for BP; endoplasmic reticulum membrane for CC; RNA polymerase II core promoter proximal region sequence-specific DNA binding, aromatase activity, and steroid hydroxylase activity for MF. Table 4 and Fig. 4 demonstrate the full GO analysis result of down-regulated DEGs.

Pathway analysis was also performed separately for up-regulated and down-regulated DEGs and the top 10 Pathway terms were selected from each analysis. Pathway analysis for up-regulated DEGs reveals a strong connection between DEGs with Cell Cycle, Cell Cycle, Mitotic, and M Phase. And the strong connection between down-regulated DEGs with Generic Transcription Pathway, Formation of the cornified envelope, and Keratinization. Table 5(A-B) and Figs. 5 and 6 demonstrate the full Pathway analysis result.

3.3. Protein-protein network construction and hub gene selection

Ninety-two nodes and a total of 242 interactions were mapped in the constructed PPI network including 46 (50%) up-regulated and 46 (50%) down-regulated DEGs (Fig. 7). We used 3 different methods (1) connectivity value (degree), (2) Maximal Clique Centrality (MCC), and (3) Maximum Neighborhood Component (MNC) to identify hub DEGs. After comparing 10 hub DEGs of 3 different methods we got 9 common DEGs. The top 9 hub DEGs are KIF23, AURKA, CENPF, CDC20, DLGAP5, CEP55, ANLN, NCAPG and KIF14. In Fig. 8 (A-D) we show all the hub DEGs of each method and a Venn diagram that shows the number of common hub DEGs.

A cluster analysis was applied to the PPI network to explore complex sections of the network. We elect the top 3 clusters from 26 clusters. Cluster 1 includes 17 DEGs, Cluster 2 has 7 DEGs and Cluster 3 has 6 DEGs. Moreover, we employed a Pathway analysis to evaluate more functional activities of the 3 clusters. Pathway analysis shows that the clusters are highly connected with Cell Cycle (Cluster 1), Interferon

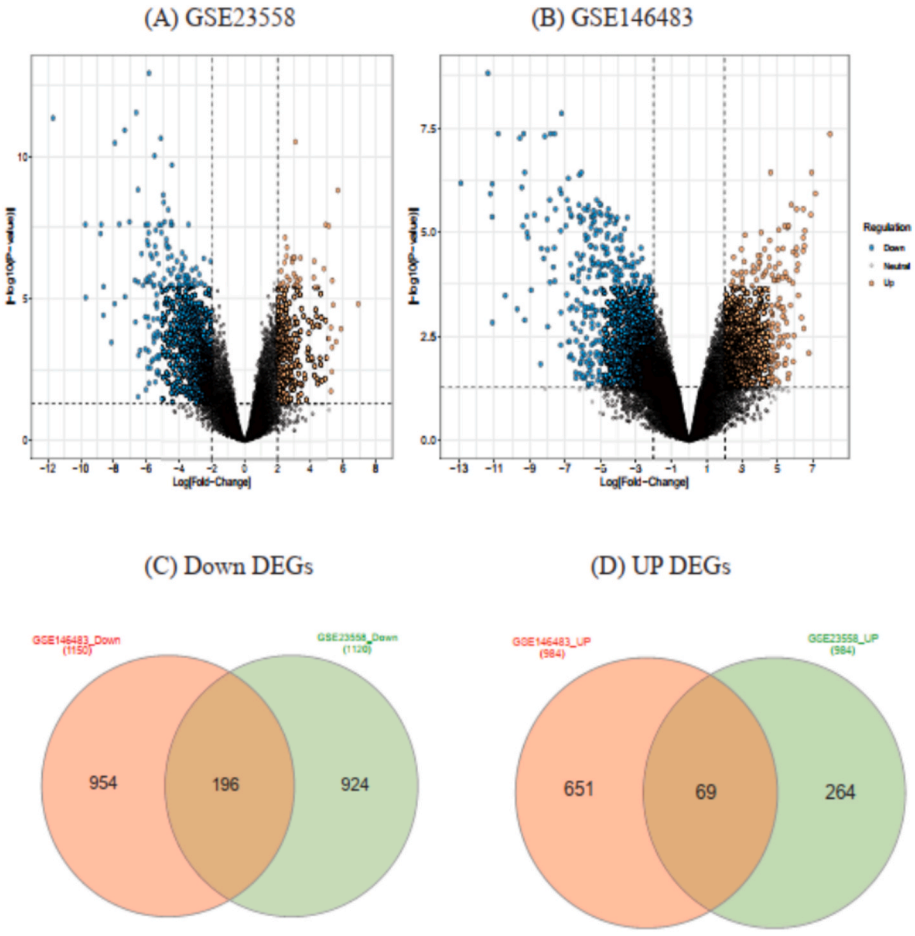

Fig. 2. Identification of DEGs. Volcano plots of the distribution of DEGs in GSE23558 (A) GSE146483 (B). Blue and brown dots represent the down-regulated and up-regulated DEGs, respectively. (C) Represents the Venn diagram of down-regulated common DEGs in both datasets. (D) Represents the Venn diagram of up-regulated common DEGs in both datasets. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
GO analysis result of up-regulated DEGs. (A), (B), and (C) indicate the three categories of GO terms.

Term ID	Term Name	Count	P-Value	Fold Enrichment	FDR
(A) BP					
GO:0009952	anterior/posterior pattern specification	10	2.18E-11	31.757	1.16E-08
GO:0048704	embryonic skeletal system morphogenesis	7	6.06E-09	47.996	1.61E-06
GO:0009954	proximal/distal pattern formation	5	1.46E-06	58.017	2.18E-04
GO:0009615	response to virus	7	1.64E-06	19.025	2.18E-04
GO:0000278	mitotic cell cycle	7	1.09E-05	13.713	0.00116
GO:0051301	cell division	9	3.33E-05	7.0893	0.002952
GO:0007059	chromosome segregation	5	1.11E-04	19.848	0.007367
GO:0051607	defense response to virus	7	1.11E-04	9.0636	0.007367
GO:0007094	mitotic spindle assembly checkpoint	4	1.53E-04	37.711	0.009072
GO:0045071	negative regulation of viral genome replication	4	4.55E-04	26.234	0.024189
GO:0010389	regulation of G2/M transition of mitotic cell cycle	3	5.65E-04	82.278	0.027348
GO:0006357	regulation of transcription from RNA polymerase II promoter	15	0.00110368	2.6264	0.04893
(C) CC					
GO:0000785	chromatin	18	1.54E-08	5.3769	1.96E-06
GO:0005654	nucleoplasm	32	1.29E-07	2.5543	8.22E-06
GO:0005634	nucleus	38	1.54E-06	2.0088	6.54E-05
GO:0015630	microtubule cytoskeleton	7	3.39E-05	11.24	0.001075
GO:0072686	mitotic spindle	6	7.52E-05	13.557	0.00191
GO:0000776	kinetochore	6	1.29E-04	12.089	0.002733
GO:0030496	midbody	6	2.77E-04	10.26	0.005027
GO:0005819	spindle	5	0.001171215	10.687	0.018593
GO:0005813	centrosome	8	0.001535098	4.6181	0.021662
GO:0045171	intercellular bridge	4	0.002964959	13.754	0.037655
GO:0005874	microtubule	6	0.003448848	5.8222	0.039819
(D) MF					
GO:1990837	sequence-specific double-stranded DNA binding	12	1.72E-06	6.4642	2.02E-04
GO:0008017	microtubule binding	8	2.59E-05	9.0078	0.001531
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	15	5.44E-05	3.5099	0.002139
GO:0005515	protein binding	55	2.00E-04	1.3122	0.005899
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	13	4.87E-04	3.2259	0.011482
GO:0003677	DNA binding	13	0.001175739	2.9187	0.023123
GO:0003700	transcription factor activity, sequence-specific DNA binding	8	0.002247447	4.3095	0.037886

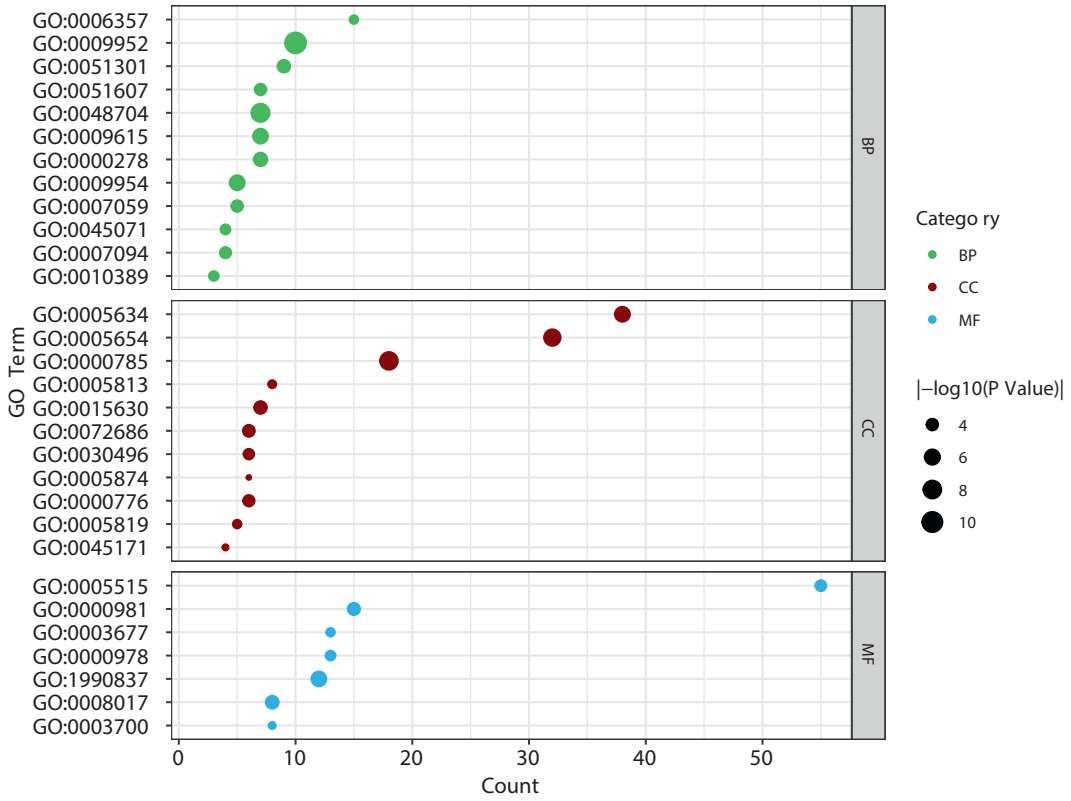

Fig. 3. Gene Ontology function analysis results for up-regulated DEGs. Y-axis represents the GO-term id and X-axis indicates the associated gene number with the term. Bubble size connected with the |−log10(PValue)| value, high |−log10(PValue)| means bigger dot. The color of the dot represents the category of gene ontology.

Table 4
GO analysis result of down-regulated DEGs. (A), (B), and (C) indicate the three categories of GO terms.

Term ID	Term Name	Count	P Value	Fold Enrichment	FDR
(A) BP					
GO:0031424	keratinization	8	4.44E-06	12.01213	0.003945
GO:0006357	regulation of transcription from RNA polymerase II promoter	33	2.43E-05	2.214366	0.008436
GO:0006805	xenobiotic metabolic process	8	2.85E-05	9.067982	0.008436
GO:0006355	regulation of transcription, DNA-templated	22	9.95E-05	2.606116	0.022104
GO:0008210	estrogen metabolic process	5	1.38E-04	18.64787	0.024594
GO:0070989	oxidative demethylation	4	2.63E-04	30.83114	0.038996
GO:0006082	organic acid metabolic process	4	7.41E-04	22.02224	0.085414
GO:0042178	xenobiotic catabolic process	4	9.74E-04	20.10726	0.085414
GO:0042737	drug catabolic process	3	0.0010774	57.80838	0.085414
GO:0042738	exogenous drug catabolic process	4	0.0011061	19.26946	0.085414
GO:0008202	steroid metabolic process	5	0.0011115	10.90724	0.085414
GO:0042572	retinol metabolic process	5	0.0011925	10.70526	0.085414
GO:0042573	retinoic acid metabolic process	4	0.001249	18.49868	0.085414
GO:0002933	lipid hydroxylation	3	0.0014998	49.55004	0.09524
(B) CC					
GO:0005789	endoplasmic reticulum membrane	22	2.55E-04	2.438567	0.047007
(C) MF					
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	31	2.75E-07	2.867587	7.93E-05
GO:0070330	aromatase activity	6	5.27E-06	23.10018	7.58E-04
GO:0008395	steroid hydroxylase activity	6	1.02E-05	20.30016	9.82E-04
GO:0004252	serine-type endopeptidase activity	10	5.94E-05	5.78502	0.003269
GO:0004497	monooxygenase activity	7	5.95E-05	10.28363	0.003269
GO:0016712	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, reduced flavin or flavoprotein as one donor, and incorporation of one atom of oxygen	5	7.78E-05	21.47132	0.003269
GO:0008401	retinoic acid 4-hydroxylase activity	4	7.95E-05	44.66036	0.003269
GO:0008236	serine-type peptidase activity	6	1.79E-04	11.35433	0.006442
GO:0019825	oxygen binding	5	2.86E-04	15.50707	0.009151
GO:0000981	RNA polymerase II transcription factor activity, sequence-specific DNA binding	25	4.23E-04	2.180681	0.012173
GO:0050649	testosterone 6-beta-hydroxylase activity	3	4.67E-04	83.73817	0.012235
GO:0005198	structural molecule activity	8	0.0014516	4.776509	0.034839
GO:0005506	iron ion binding	7	0.0017825	5.465428	0.038781
GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen	5	0.0018852	9.46194	0.038781
GO:0101020	estrogen 16-alpha-hydroxylase activity	3	0.0021301	41.86908	0.040899

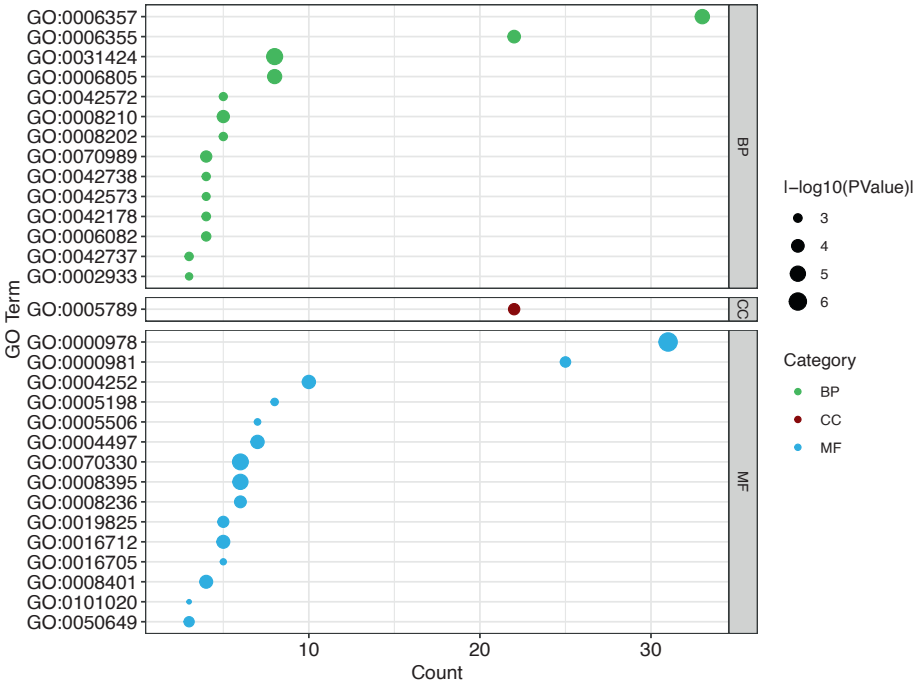

Fig. 4. Gene Ontology function analysis results for down-regulated DEGs. Y-axis represents the GO-term id and X-axis indicates the associated gene number with the term. Bubble size connected with the |−log10(PValue)| value, high |−log10(PValue)| means bigger dot. The color of the dot represents the category of gene ontology.

alpha/beta signaling (Cluster 2), and Xenobiotics (Cluster 3). In Table 6 (A-C) and Fig. 9 (A-D) demonstrates all the 3 clusters and their associated Pathways.

3.4. TF-miRNA network analysis

A TF-miRNA network was generated including 226 nodes and 994

Table 5
Pathway analysis result of (A) up-regulated and (B) down-regulated DEGs.

Term	Term ID	Count	P-Value	Fold Enrichment	FDR
(A) Up-regulated DEGs					
Cell Cycle	R-HSA-1640170	19	7.66E-11	6.505435	2.53E-08
Cell Cycle, Mitotic	R-HSA-69278	16	3.21E-09	6.757498	5.30E-07
M Phase	R-HSA-68886	10	4.27E-05	5.668296	0.001162
Mitotic Prometaphase	R-HSA-68877	9	1.56E-06	10.45301	1.03E-04
Resolution of Sister Chromatid Cohesion	R-HSA-2500257	8	7.36E-07	15.04348	8.09E-05
RHO GTPases Activate Formins	R-HSA-5663220	8	1.50E-06	13.53913	1.03E-04
Separation of Sister Chromatids	R-HSA-2467813	8	1.17E-05	9.92397	4.85E-04
EML4 and NUDC in mitotic spindle formation	R-HSA-9648025	7	7.79E-06	14.17559	4.28E-04
Interferon alpha/beta signaling	R-HSA-909733	6	1.09E-05	19.74457	4.85E-04
Amplification of signal from the kinetochores	R-HSA-141424	6	4.41E-05	14.80842	0.001162
(B) Down-regulated DEGs					
Generic Transcription Pathway	R-HSA-212436	21	0.010465	1.786554	0.286743
Formation of the cornified envelope	R-HSA-6809371	12	2.90E-08	9.863348	7.96E-06
Keratinization	R-HSA-6805567	12	4.37E-06	5.991754	3.99E-04
Biological oxidations	R-HSA-211859	9	0.001039	4.331876	0.031639
Phase I – Functionalization of compounds	R-HSA-211945	7	4.41E-04	7.056326	0.021826
Xenobiotics	R-HSA-211981	6	2.84E-06	25.64471	3.89E-04
Cytochrome P450 – arranged by substrate type	R-HSA-211897	6	3.53E-04	9.713904	0.021826
Biosynthesis of DHA-derived SPMs	R-HSA-9018677	4	4.78E-04	25.14187	0.021826
Biosynthesis of specialized proresolving mediators (SPMs)	R-HSA-9018678	4	6.72E-04	22.49536	0.026302
Melanin biosynthesis	R-HSA-5662702	3	8.35E-04	64.11176	0.028598

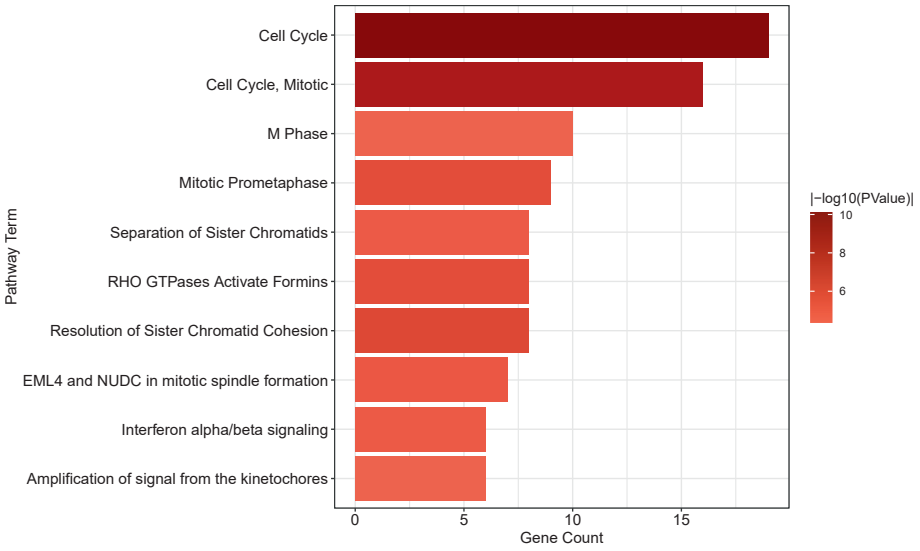

Fig. 5. Pathway enrichment analysis results for up-regulated DEGs. Y-axis represents the pathway term name and X-axis indicates the associated gene number with the term. The color of the bar relies on |-log(PValue)|, a high value of |-log(PValue)| means high opacity and low opacity means low |-log(PValue)| value.

edges. 74 miRNAs and 152 TFs were mapped in the network. The top 5 miRNA and TF were selected based on the highest degree value. The top 5 miRNAs are hsa-miR-19b, hsa-miR-218, hsa-miR-19a, hsa-miR-340, and hsa-miR-137; and the top 5 TFs are FOXP2, TRPS1, PDGFRA, MAF, and SATB1. In Fig. 10 (A-B) TF-miRNA network and the top 5 miRNA and TF are visualized.

4. Discussion

OSCC is a type of lethal tumor with remarkable characteristics of impetuous development, lofty metastatic retention, and soaring mortality. OSCC progression requires a multistep method for the procurement of diverse genetic violations, affected by a patient's genetic proneness as well as by environmental effects like smoking, consuming alcohol, and viral infection.[21,22] Despite notable developments in modern therapeutics, biomarkers or targets for OSCC detection and therapy are currently inadequate compared to other cancers. Recently lots of changes happened in chemoradiation and surgery for OSCC treatments but the etiopathogenesis of OSCC is not perfectly convinced till now, and five-year survival rates for OSCC remain low.[22,23] Exploring the causative and molecular mechanisms of OSCC is crucial seriousness for prevention and therapy. In the last few years, bioinformatics analysis has played very promising role in cancer research, it's become a very popular and common feature to identify and screen crucial DEGs in any tumor or cancer.

The NCBI-GEO is one of the most trusted and freely accessible repositories for microarray profiling and next-generation sequencing. In the current study, two different gene-expression profile datasets (GSE23558 and GSE146483) were collected by maintaining a few general conditions. The GEO2R is a tool for the analysis of microarray data that gives a visual representation to understand the dataset more clearly and compare various groups of samples using efficient statistical analysis. GEO2R also provides a DEGs list but without proper filtration of cut-off parameters. After filtering (applying the cut-off criteria) and comparing between two profiles a total of 265 DEGs were collected including 69 up-regulated and 196 down-regulated DEGs.

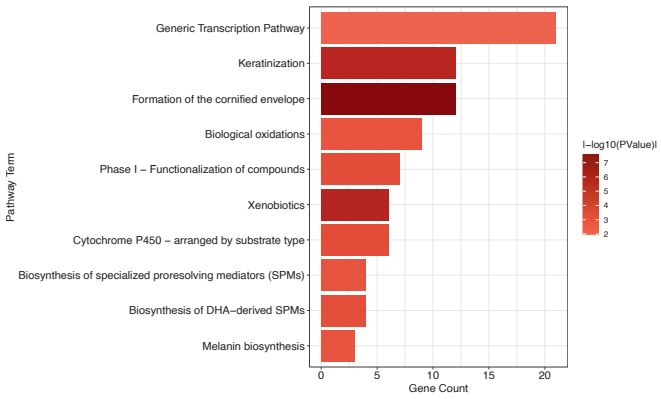

Fig. 6. Pathway enrichment analysis results for down-regulated DEGs. Y-axis represents the pathway term name and X-axis indicates the associated gene number with the term. The color of the bar relies on |-log(PValue)|, a high value of |-log(PValue)| means high opacity and low opacity means low |-log (PValue)| value.

The GO analysis showed the top 30 ontology terms for identified up and down-regulated DEGs. Pathway analysis was used to reveal the top 10 most attached terms to the up and down-regulated DEGs. Pathway terms show the up-regulated DEGs were mostly connected with some very crucial activities like Cell Cycle, Cell Cycle, Mitotic, and M Phase; and the down-regulated DEGs were connected with Generic Transcription Pathway, Formation of the cornified envelope, and Keratinization. Keratinization enacts the cytoplasmic instances that take place in the cytoplasm of epidermal keratinocytes throughout their terminal disintegration. In 2017 a study, by Wolfer et al. revealed that Keratinization Is an exclusive Prognostic component in OSCC.[23,24] In 2015, Cooper et al showed that nonkeratinized tumors have elevated survival than keratinized tumors in throat cancer.[24,25]

The PPI network was constructed using a strong combined score. In the network, 92 nodes and 242 interactions were present. Three distinct methods were applied to identify the hub DEGs, and after that, we figure out the common DEGs in the three methods' results. The hub DEGs were KIF23, AURKA, CENPF, CDC20, DLGAP5, CEP55, ANLN, NCAPG and KIF14. KIF23 is a protein-coding gene and the 23rd member of the kinesin-like protein family. Study shows the KIF23 gene is greatly expressed in different types of tumors, like breast cancer, gastric cancer, and lung cancer. The overexpression of KIF23 is mainly connected with tumor grade, invasion, and prognosis in breast cancer.[25–27] In a study, Cheng and et al. disclose that KIF14 and KIF23 takes part to raise Cell Proliferation Chemoresistance in Hepatocellular carcinoma cells.[27,28] Unfortunately, no studies are found available that discussed the regulatory role of KIF23 in OSCC development, but we think KIF23 may have the potential to be a therapeutic biomarker of OSCC. More study is needed to prove our statement.

AURKA is a member of serine/threonine kinases family, its activation

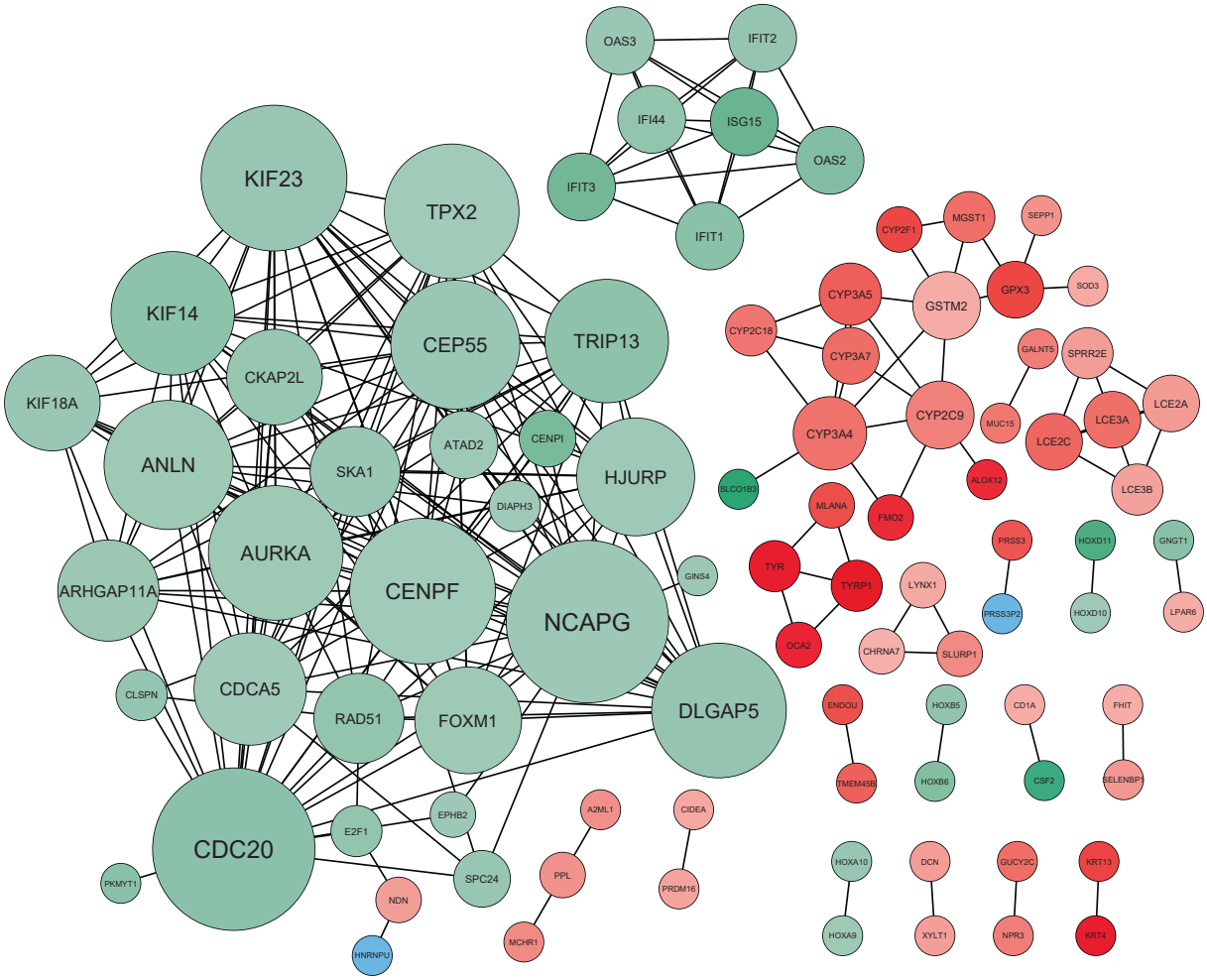

Fig. 7. Protein-protein interaction network using the common DEGs. The network is constructed with 92 nodes and 242 edges. The green and red colors represent up-regulated and down-regulated DEGs. The circle size of nodes depends on the connection value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

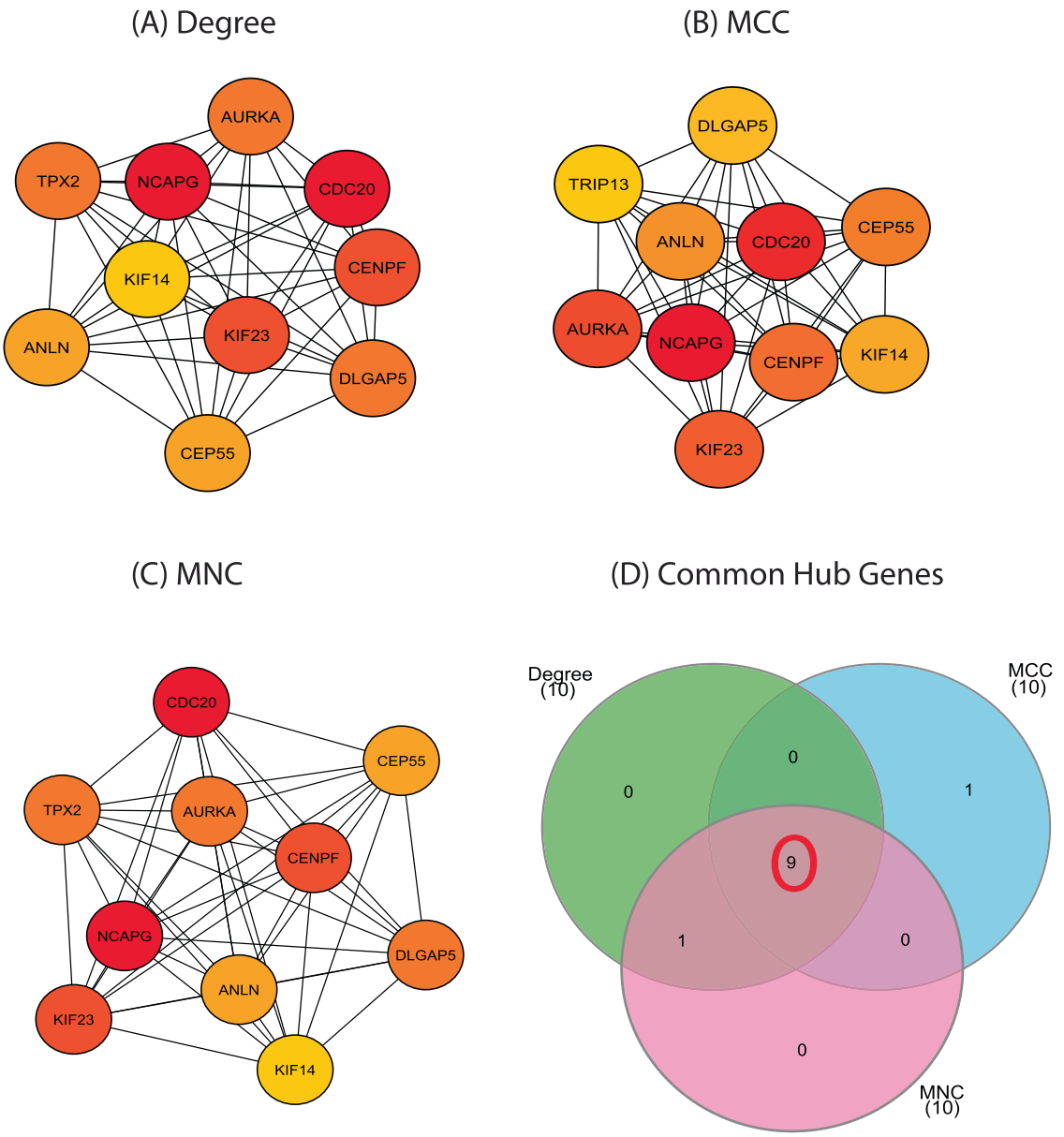

Fig. 8. Top 10 hub genes according to (A) degree value, (B) MCC, and (C) MNC method. (D) Venn diagram that identifies common hub genes between the results of 3 methods.

is crucial for cell division procedure. According to The Cancer Genome Atlas (TCGA) database data, AURKA shows remarkable overexpression in cancer tissues compared to normal tissues.[28,29] Overexpression of AURKA can encourage cellular transformation and may potentiate the action of various oncogenes, further enhancing the chance of tumorigenesis.[29,30] Dawei et al. identified that AURKA associates with the development of OSCC by modulating epithelial-to-mesenchymal transition and apoptosis via the regulation of ROS.[10,31] In 2013 in a study, Tanaka et al. finds a meaningful connection between the expression of AURKA and histological differentiation and lymph node metastasis from OSCC samples.[11,32] In a study in 2017, genetic polymorphisms of AURKA were identified as a significant factor in OSCC progression in Taiwan.[12,33] According to the discussion, AURKA may play a crucial role in OSCC development. The CENPF gene has connections with several tumor and cancer types. Study shows high expression of CENPF relates to slow prediction of cancer and tumor bone metastasis in breast cancer.[13,34,39] High expression of CENPF is also connected with slow prediction and development of lung adenocarcinoma.[14,35] Cell division cycle 20′s (CDC20) divergent expression is connected with aggressive development and slow detection of tumors in several types of cancer, such as pancreatic ductal adenocarcinoma, gastric cancer, urothelial bladder cancer, astrocytoma, hepatocellular carcinoma, lung adenocarcinoma, and OSCC.[15,36] CDC20 may play a significant role in OSCC progression and medication. Centrosomal Protein 55 (CEP55) showed high expression in head and neck squamous cell carcinoma, and oral cavity squamous cell carcinoma.[16,17,37,38,40] Furthermore, we generate a TF-miRNA network with 226 nodes and 994 edges to understand deeper knowledge about the DEGs but no significant result was found to describe. Bioinformatics tools offer significant insights into biological processes, yet their predictive nature has inherent limitations. These include biases in algorithms, dependence on incomplete databases, and an inability to fully capture the complexity of dynamic cellular environments. Without experimental validation, computational predictions may generate false positives or fail to detect functionally important variations. Moreover, structural and functional annotations based solely on sequence homology can be misleading due to evolutionary divergence. To address these challenges, future research should emphasize the integration of multi-omics data, enhancement of machine learning

Table 6
Pathway analysis result of identifying clusters of PPI network. (A), (B), and (C) indicate the cluster number.

Term ID	Term	Count	P Value	Fold Enrichment	FDR
(A) Cluster 1					
R-HSA-1640170	Cell Cycle	10	8.72E-09	11.29851	6.28E-07
R-HSA-69278	Cell Cycle, Mitotic	9	4.64E-08	12.54316	1.67E-06
R-HSA-68877	Mitotic Prometaphase	5	7.34E-05	19.16317	0.001425
R-HSA-68886	M Phase	6	7.91E-05	11.22283	0.001425
R-HSA-2500257	Resolution of Sister Chromatid Cohesion	4	3.92E-04	24.82086	0.005638
(B) Cluster 2					
R-HSA-909733	Interferon alpha/beta signaling	6	1.07E-11	152.0278	2.46E-10
R-HSA-913531	Interferon Signaling	6	1.94E-09	54.73	2.23E-08
R-HSA-1280215	Cytokine Signaling in Immune system	6	1.30E-06	14.99452	9.99E-06
R-HSA-1169410	Antiviral mechanism by IFN-stimulated genes	4	3.72E-06	91.21667	2.14E-05
R-HSA-168256	Immune System	6	2.15E-04	5.408103	9.90E-04
(C) Cluster 3					
R-HSA-211981	Xenobiotics	5	1.06E-10	364.8667	1.58E-09
R-HSA-211859	Biological oxidations	6	3.28E-09	49.30631	2.46E-08
R-HSA-211897	Cytochrome P450 − arranged by substrate type	5	6.00E-09	138.2071	3.00E-08
R-HSA-211945	Phase I − Functionalization of compounds	5	4.12E-08	86.05346	1.55E-07
R-HSA-1430728	Metabolism	6	2.70E-04	5.168083	8.11E-04

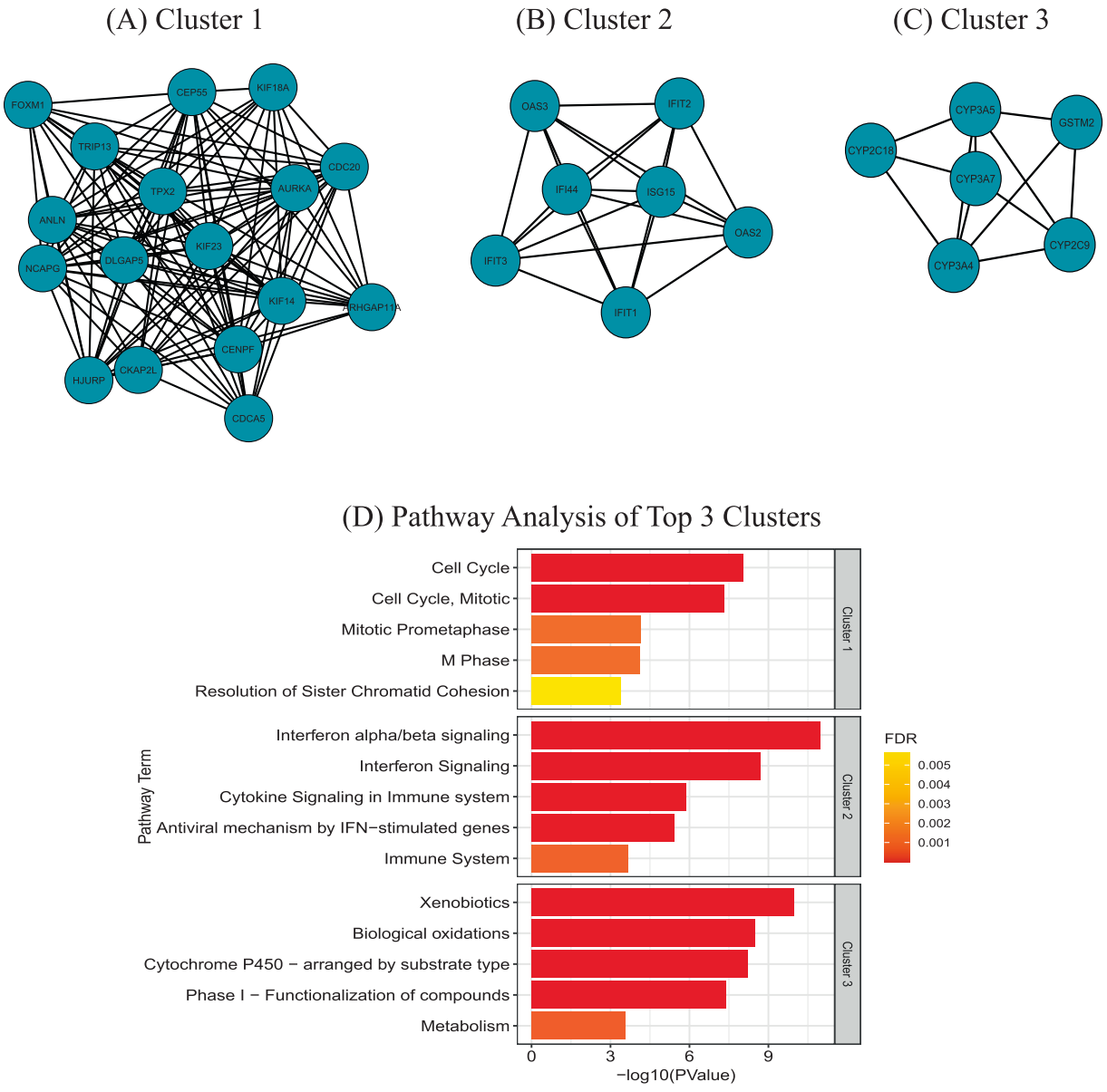

Fig. 9. Top 3 clusters from the PPI network (A–C). Pathway analysis of clusters (D). The red color bar means a low FDR value and the yellow color bar means a high FDR value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

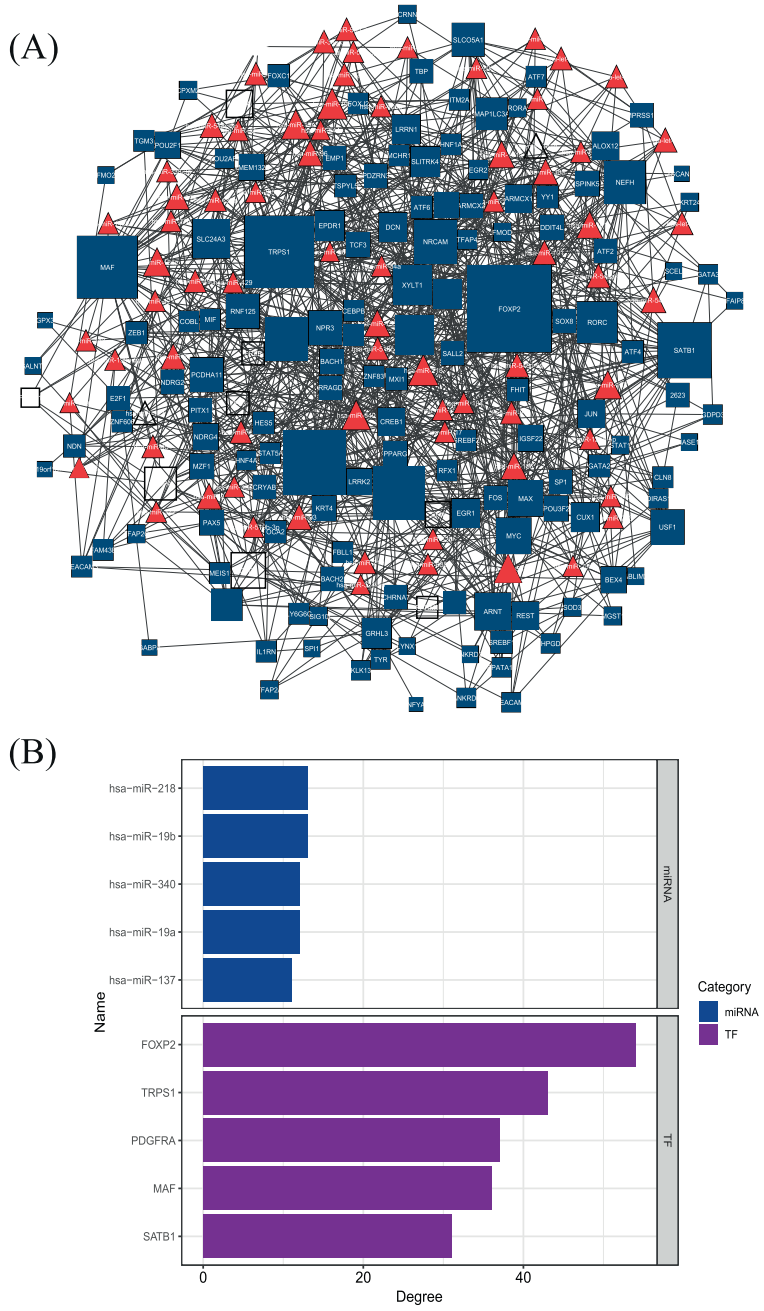

Fig. 10. TF-miRNA network includes 226 nodes and 994 edges (A). 74 miRNAs and 152 TFs were mapped in the network. The triangle shape represents the miRNA and the square shape represents the TF. (B) Top 5 miRNA and TF visualized with indicating their degree value. Different color represents whether it is miRNA or TF.

models with experimentally validated datasets, and the development of hybrid approaches that merge computational predictions with high-throughput experimental screening. Advances in single-cell sequencing, cryo-electron microscopy, and synthetic biology can facilitate direct experimental validation of bioinformatics-driven hypotheses. Strengthening collaborations between computational biologists and experimental researchers will be essential to bridging the gap between in silico predictions and in vivo or in vitro validation. Limitations of this study, firstly we have used two different datasets but those are imbalanced in class between OSCC infected and normal. Secondly, we could not offer any lab related experiment for that we can't validate our results. Third, our datasets are not appropriate to do survival related studies that's why we didn't any ROC and AUC analysis for in-depth results. But in future we will cover these limitation.

5. Conclusions

In conclusion, we applied a systematic bioinformatics analysis to identify significant biomarkers for OSCC. We identify KIF23, AURKA, and CDC20 may play significant roles in OSCC progression and therapeutic solution. The limitation of the study is the unavailability of gene expression profiles according to our conditions. We could not present any medical validation of our result which is also a limitation of our study. We hope the results of this study will be helpful to the researchers to identify biomarkers of OSCC.

Funding

"This work was funded by the Deanship of Graduate Studies and Scientific Research at Jouf University under grant No. (DGSSR-2023-01-02144)".

CRediT authorship contribution statement

Mohammad Khursheed Alam: Writing – review & editing, Writing – original draft, Conceptualization. **Md Rakibul Islam:** Writing – review & editing, Writing – original draft, Conceptualization. **Tahsinul Haque:** Writing – review & editing, Writing – original draft, Conceptualization. **Kiran Kumar Ganji:** Writing – review & editing, Writing – original draft, Conceptualization. **Hamzah Ali Babkair:** Writing – review & editing, Writing – original draft, Conceptualization. **Mohammed Enamur Rashid:** Writing – review & editing, Writing – original draft, Conceptualization. **Kawsar Ahmed:** Writing – review & editing, Writing – original draft, Conceptualization. **Francis M. Bui:** Writing – review & editing, Writing – original draft, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Mohammad Khursheed Alam reports financial support was provided by Jouf University. Mohammad Khursheed Alam reports a relationship with Jouf University that includes: funding grants. Mohammad Khursheed Alam has patent None pending to None. None If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Deshmukh V, Shekar K. Oral squamous cell carcinoma: diagnosis and treatment planning. *Oral and Maxillofacial Surgery for the Clinician*. 2021:1853–1867. https://doi.org/10.1007/978-981-15-1346-6_81.
2. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin 2021;71(3):209-249. Doi: 10.3322/caac.21660.
3. Malik UU, Zarina S, Pennington SR. Oral squamous cell carcinoma: key clinical questions, biomarker discovery, and the role of proteomics. *Arch Oral Biol*. 2016;63:53–65. https://doi.org/10.1016/j.archoralbio.2015.11.017.
4. Rogers SN, Brown JS, Woolgar JA, et al. Survival following primary surgery for oral cancer. *Oral Oncol*. 2009;45(3):201–211. https://doi.org/10.1016/j.oraloncology.2008.05.008.
5. Blohm DH, Guiseppi-Elie A. New developments in microarray technology. *Curr Opin Biotechnol*. 2001;12(1):41–47. https://doi.org/10.1016/S0958-1669(00)00175-0.
6. Duggan DJ, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. *Nat Genet*. 1999;21(1):10–14. https://doi.org/10.1038/4434.
7. Lipshutz RJ, Fodor S, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet*. 1999;21(1):20–24. https://doi.org/10.1038/4447.
8. Du H, Wang Z, Qi M, et al. Identification of novel biomarkers involved in oral squamous cell carcinoma by whole transcriptome sequencing and bioinformatics analysis. *Cancer Cell Int*. 2025;25(1):277.
9. Liang S, Ji L, Zhenyuan Yu, et al. Bioinformatic analysis and experimental validation of cuproptosis-related LncRNA as a novel biomarker for prognosis and immunotherapy of oral squamous cell carcinoma. *Hereditas*. 2024;161(1):10.
10. Qi Lu, Tang Z. Prognostic model revealing pyroptosis-related signatures in oral squamous cell carcinoma based on bioinformatics analysis. *Sci Rep*. 2024;14(1):6149.
11. Clough E, Barrett T. The gene expression omnibus database. In: *Statistical Genomics*. New York, NY: Humana Press; 2016:93–110, 10.1007/978-1-4939-3578-9_5.
12. Ambatipudi S, Gerstung M, Pandey M, et al. Genome-wide expression and copy number analysis identifies driver genes in gingivobuccal cancers. *Genes Chromosom Cancer*. 2012;51(2):161–173. https://doi.org/10.1002/gcc.20940.
13. Kase Y, Uzawa K, Wagai S, et al. Engineered exosomes delivering specific tumor-suppressive RNAi attenuate oral cancer progression. Sci Rep 2021;11(1):1-12. doi: 10.1038/s41598-021-85242-1.
14. Heberle H, Meirelles GV, da Silva FR, Telles GP, Minghim R. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC Bioinf*. 2015;16(1):1–7. https://doi.org/10.1186/s12859-015-0611-3.
15. Dennis G, Sherman BT, Hosack DA, et al. DAVID: database for annotation, visualization, and integrated discovery. Genome Biol 2003;4(9):1-11. Doi: 10.1186/gb-2003-4-9-r60.
16. Joshi-Tope G, Gillespie M, Vastrik I, et al. Reactome: a knowledgebase of biological pathways. Nucl Acids Res 2005;33(1): 428-432. Doi: 10.1093/nar/gki072.
17. Szklarczyk D, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*. 2021;49(1):605–612. https://doi.org/10.1093/nar/gkaa1074.
18. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*. 2011;27(3):431–432. https://doi.org/10.1093/bioinformatics/btq675.
19. Chin C-H, Chen S-H, Hsin-Hung Wu, Ho C-W, Ko M-T, Lin C-Y. cytoHubba: identifying hub objects and sub-networks from complex interactome. *BMC Syst Biol*. 2014;8(4):1–7. https://doi.org/10.1186/1752-0509-8-S4-S11.
20. Bader GD, Hogue CWV. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinf*. 2003;4(1):1–27. https://doi.org/10.1186/1471-2105-4-2.
21. Liu Z-P, Canglin Wu, Miao H, Hulin Wu. RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database*. 2015;2015. https://doi.org/10.1093/database/bav095.
22. Choi S, Myers JN. Molecular pathogenesis of oral squamous cell carcinoma: implications for therapy. *J Dent Res*. 2008;87(1):14–32. https://doi.org/10.1177/154405910808700104.
23. Omura K. Current status of oral cancer treatment strategies: surgical treatments for oral squamous cell carcinoma. *Int J Clin Oncol*. 2014;19:423–430. https://doi.org/10.1007/s10147-014-0689-z.
24. Wolfer S, Elstner S, Schultze-Mosgau S. Degree of keratinization is an independent prognostic factor in oral squamous cell carcinoma. *J Oral Maxillofac Surg*. 2018;76(2):444–454. https://doi.org/10.1016/j.joms.2017.06.034.
25. Cooper T, Biron VL, Adam B, et al. Association of keratinization with 5-year disease-specific survival in oropharyngeal squamous cell carcinoma. *JAMA Otolaryngology–Head & Neck Surgery*. 2015;141(3):250–256. https://doi.org/10.1001/jamaoto.2014.3335.
26. Hu Y, Zheng M, Wang C, et al. Identification of KIF23 as a prognostic signature for ovarian cancer based on large-scale sampling and clinical validation. *Am J Transl Res*. 2020;12(9):4955.
27. Zou JX, Duan Z, Wang J, et al. Kinesin family deregulation coordinated by bromodomain protein ANCCA and histone methyltransferase MLL for breast cancer cell growth, survival, and tamoxifen ResistanceANCCA mediates coordinated deregulation of kinesins. Mol Cancer Res 2014;12(4):539-549. Doi: 10.1158/1541-7786.MCR-13-0459.
28. Cheng C, Xingxing Wu, Shen Yu, Li Q. KIF14 and KIF23 promote cell proliferation and chemoresistance in HCC cells, and predict worse prognosis of patients with HCC. *Cancer Manag Res*. 2020;12:13241. https://doi.org/10.2147/CMAR.S285367.
29. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in cancer: molecular mechanisms and opportunities for cancer therapy. *Mol Cancer*. 2021;20:1–27. https://doi.org/10.1186/s12943-020-01305-3.
30. Tatsuka M, Sato S, Kitajima S, et al. Overexpression of Aurora-a potentiates HRAS-mediated oncogenic transformation and is implicated in oral carcinogenesis. *Oncogene*. 2005;24(6):1122–1127. https://doi.org/10.1038/sj.onc.1208293.
31. Dawei H, Honggang D, Qian W. AURKA contributes to the progression of oral squamous cell carcinoma (OSCC) through modulating epithelial-to-mesenchymal transition (EMT) and apoptosis via the regulation of ROS. *Biochem Biophys Res Commun*. 2018;507(1–4):83–90. https://doi.org/10.1016/j.bbrc.2018.10.170.
32. Tanaka H, Nakashiro K-I, Iwamoto K, et al. Targeting Aurora kinase a suppresses the growth of human oral squamous cell carcinoma cells in vitro and in vivo. *Oral Oncol*. 2013;49(6):551–559. https://doi.org/10.1016/j.oraloncology.2013.02.002.
33. Chou C-H, Chou Y-E, Chuang C-Y, Yang S-F, Lin C-W. Combined effect of genetic polymorphisms of AURKA and environmental factors on oral cancer development in Taiwan. *PLoS One*. 2017;12(2), e0171583. https://doi.org/10.1371/journal.pone.0171583.
34. Sun J, Huang J, Lan J, et al. Overexpression of CENPF correlates with poor prognosis and tumor bone metastasis in breast cancer. *Cancer Cell Int*. 2019;19(1):1–11. https://doi.org/10.1186/s12935-019-0986-8.
35. Li M-X, Zhang M-Y, Dong H-H, et al. Overexpression of CENPF is associated with progression and poor prognosis of lung adenocarcinoma. *Int J Med Sci*. 2021;18(2):494. https://doi.org/10.7150/ijms.49041.
36. Cheng S, Castillo V, Sliva D. CDC20 associated with cancer metastasis and novel mushroom-derived CDC20 inhibitors with antimetastatic activity. *Int J Oncol*. 2019;54(6):2250–2256. https://doi.org/10.3892/ijo.2019.4791.
37. Waseem A, Ali M, Odell EW, Fortune F, Teh M-T. Downstream targets of FOXM1: CEP55 and HELLS are cancer progression markers of head and neck squamous cell carcinoma. *Oral Oncol*. 2010;46(7):536–542. https://doi.org/10.1016/j.oraloncology.2010.03.022.
38. Chen CH, Chien CY, Huang CC, et al. Expression of FLJ10540 is correlated with aggressiveness of oral cavity squamous cell carcinoma by stimulating cell migration and invasion through increased FOXM1 and MMP-2 activity. *Oncogene*. 2009;28(30):2723–2737. https://doi.org/10.1038/onc.2009.128.
39. Hamad HA, Enezei HH, Alrawas A, et al. Identification of potential chemical substrates as fuel for hypoxic tumors that may be linked to invadopodium formation in hypoxia-induced MDA-MB-231 breast-cancer cell line. *Molecules*. 2020;25(17):3876. https://doi.org/10.3390/molecules25173876.
40. Al-Kubaisi AA, Ghazi MA, Majeed NS, Aldelaimi ER, Enezei HH. Soluble urokinase plasminogen activator receptor (suPAR) is a potential biomarker of stage III-IV, grade C periodontitis through the impact of post-radiotherapy on head and neck cancer patients. *BMC Oral Health*. 2024;24(1):1144. https://doi.org/10.1186/s12903-024-04801-w.