Journal of Dental Sciences 19 (2024) S81—S90

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Identify characteristics of Vietnamese oral squamous cell carcinoma patients by machine learning on transcriptome and clinical-histopathological analysis

Huong Thu Duong [a†], Nam Cong-Nhat Huynh [a*†], Chi Thi-Kim Nguyen [a], Linh Gia-Hoang Le [b], Khoa Dang Nguyen [a], Hieu Trong Nguyen [c,d], Lan Ngoc-Ly Tu [c,d], Nam Huynh-Bao Tran [c,d], Hoa Giang [c,d], Hoai-Nghia Nguyen [c,d], Chuong Quoc Ho [b], Hung Trong Hoang [a], Thinh Huy-Quoc Dang [e], Tu Anh Thai [e], Dong Van Cao [f]

[a] *Faculty of Odonto-stomatology, University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Viet Nam*
[b] *Center for Molecular Biomedicine, University of Medicine and Pharmacy at Ho Chi Minh City, Ho Chi Minh City, Viet Nam*
[c] *Gene Solutions, Ho Chi Minh City, Viet Nam*
[d] *Medical Genetics Institute, Ho Chi Minh City, Viet Nam*
[e] *Ho Chi Minh City Oncology Hospital, Ho Chi Minh City, Viet Nam*
[f] *Blood Transfusion Haematology Hospital No. 2, Ho Chi Minh City, Viet Nam*

Received 5 August 2024; Final revision received 19 August 2024
Available online 28 August 2024

KEYWORDS
mRNA;
Oral squamous cell carcinoma;
Machine learning;
Clinical;
Histopathological

Abstract *Background/purpose:* Oral squamous cell carcinoma (OSCC) is notorious for its low survival rates, due to the advanced stage at which it is commonly diagnosed. To enhance early detection and improve prognostic assessments, our study harnesses the power of machine learning (ML) to dissect and interpret complex patterns within mRNA-sequencing (RNA-seq) data and clinical-histopathological features.
Materials and methods: 206 retrospective Vietnamese OSCC formalin-fixed paraffin-embedded (FFPE) tumor samples, of which 101 were subjected to RNA-seq for classification based on gene expression. Then, learning models were built based on clinical-histopathological data to

* Corresponding author. Laboratory of Prosthodontics and Laboratory of Oral-Maxillofacial Biology, Faculty of Odonto-Stomatology, University of Medicine and Pharmacy at Ho Chi Minh City, 652 Nguyen Trai, Ward 11, District 5, Ho Chi Minh City, 749000, Viet Nam.
E-mail address: namhuynh@ump.edu.vn (N.C.-N. Huynh).
† These authors contributed equally to this work.

https://doi.org/10.1016/j.jds.2024.08.013

predict OSCC subtypes and propose potential biomarkers for the remaining 105 samples.
Results: 2 distinct groups of OSCC with different clinical-histopathological characteristics and gene expression. Subgroup 1 was characterized by severe histopathologic features with immune response and apoptosis signatures while subgroup 2 was denoted by more clinical/pathological features, cell division and malignant signatures. XGBoost and SVM (Support Vector Machine) models showed the best performance in predicting subtype OSCC. The study also proposed 12 candidate genes as potential biomarkers for OSCC subtypes (6/group).
Conclusion: The study identified characteristics of Vietnamese OSCC patients through a combination of mRNA sequencing and clinical-histopathological analysis. It contributes to the insight into the tumor microenvironment of OSCC and provides accurate ML models for biomarker prediction using clinical-histopathological features.

Introduction

OSCC is particularly concerning due to its high recurrence rates in Asia, affecting 50—60% of advanced cases and 25—30% of early-stage cases.[1] It highlights the critical need for targeted strategies in prevention and treatment. Based on whole-exome sequencing of 120 Asians, the genetic landscape of OSCC reflects its etiology and therapeutic implications.[2] The advent of ML revolutionizes medicine, enabling precise, data-driven approaches for cancer classification and analysis of complex datasets.[3] They construct predictive models through example-based learning, fine-tuning parameters for optimal performance.[4] In OSCC, a disease marked by pronounced tumor microenvironment heterogeneity, ML synergizes with RNA-seq and clinical-histopathological analysis to forge a powerful alliance.[5,6] This confluence is instrumental in dissecting the unique oncological landscape of each patient, necessitating bespoke treatment modalities for tailored and prognostically favorable therapies.[7]

Given the regional variations in OSCC etiology, it is plausible that the cancer microenvironment in Asia is distinct. In Vietnam, the local OSCC landscape is poorly understood, and insights into genomic, clinical, and pathological data are lacking. Here, we performed RNA-seq on Vietnamese OSCC samples and then clustered and classified by different ML models using clinical-histopathological features suggesting novel biomarkers and clinical evaluation.

Materials and methods

Study design

206 retrospective FFPE tissue samples from primary OSCC patients without prior interventions were obtained at Ho Chi Minh City Oncology Hospital over 2019—2022 (Fig. 1 and Supp. Data S1) with clinical data such as tumor location, clinical/pathologic stage, and tumoral recurrence status. Patients had prior tumor resection and lymph node dissection, providing suitable specimens for histopathology, and RNA-seq (Supp. Doc. S1 and S2).

A subset of 101 patient samples underwent RNA-seq, clinical-histopathological scoring, and ML algorithms to construct predictive models of gene expression based on RNA-seq data, which would subsequently identify potential novel OSCC biomarkers. Then, 105 OSCC non-RNA-seq data with only clinical-histopathological features were analyzed using the predictive models. Study was approved by Ethics Committee (No. 441/2021, 184/2024). Data produced in this study is available at www.oscc.vn.

Bulk mRNA sequencing

FFPE samples were cut into 6—8 μm thick slices to prepare for RNA-seq by our reported optimized methods.[8] In brief, total RNA was extracted using PureLink FFPE RNA Isolation Kit (Invitrogen, Waltham, MA, USA). cDNA was synthesized and then sequenced by MiniSeq Sequencing System (Illumina, San Diego, CA, USA). RNA-seq data were then passed quality control and trimming by TrimGalore (v0.6.10) following quantifying abundances of transcripts by Kallisto (v0.46.1) using the human transcriptome index (*Homo sapiens* GRCh38, release 107_k31).[9]

RNA-seq data analysis and unsupervised clustering

Raw count data were transformed to relative gene expression by vst function (DESeq2 v1.42).[10] Since RNA-seq data were generated in multiple batches of sequencing, we performed batch effect correction with Limma (v3.60.2). Gene expression data were fed to a 100-fold consensus K-means clustering (ConsensusClusterPlus v1.66) to initially identify clusters.[11] Only 1,000 most variable genes were included in the clustering step. Differential gene expression analysis was performed in a one-versus-rest strategy between groups using DESeq2.

Pathway analysis and cell type deconvolution

Pathway analysis was performed using the clusterProfiler (v 4.12.0). In short, we employed both the Over-Representation Analysis (ORA) and the Gene Set Enrichment Analysis (GSEA) using the normalized gene expression

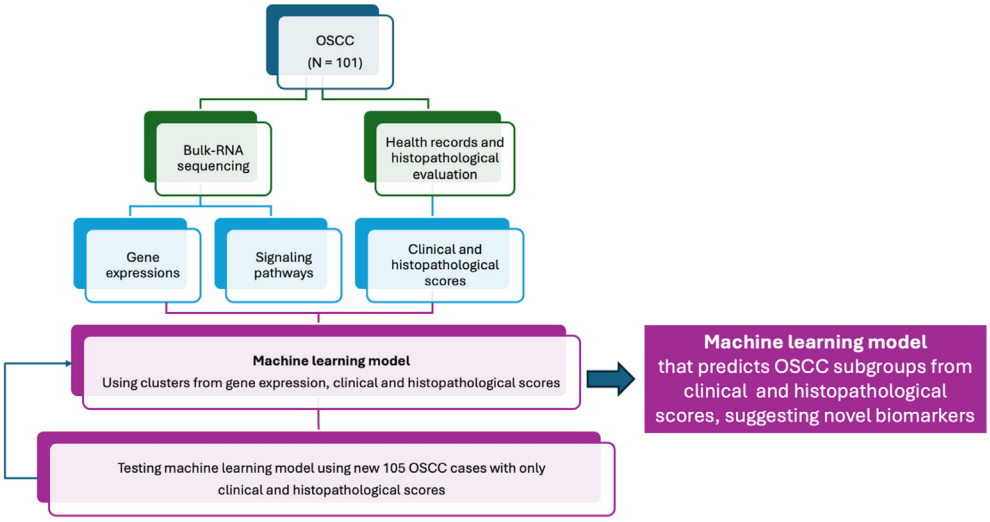

Figure 1 Study design. 101 OSCC samples were RNA-sequenced then classified into subgroups based on gene expression. Clinical, histological features obtained from health records and H&E staining were used to build machine learning models to predict OSCC subgroups from clinical and histopathological scores and suggest novel biomarkers for OSCC with new 105 OSCC non-RNA-seq samples. OSCC, oral squamous cell carcinoma.

matrix of top-1000 most variable genes.[12] For ORA, we ranked the enriched pathways by their GeneRatio. For GSEA, enriched pathways were ranked by their absolute NES values. Our gene list was tested against well-curated gene-set databases: Hallmark, GO, KEGG, C4–C6 MSigDB cancer gene-sets.

Cell type deconvolution was performed by CIBERSORTx.[13] This provided an estimate of cell type abundances presenting in the RNA-seq sample. We used signature genes from head-neck squamous cell carcinoma (HNSCC) single-cell RNA-seq data as reference matrix.[14]

Histopathological analysis

Each FFPE sample was cut into 3–4 μm thick slices for hematoxylin and eosin (H&E) staining and then be scored to assess histopathological features (Supp. Doc. S2) including the growth pattern of the tumor, the depth of invasion (DOI), and specific grading parameters from established systems such as Bryne grading system,[15] World Health Organization (WHO) system, The Brandwein-Gensler risk model. For histopathological analysis and scoring, Extranodal extension (ENE) was excluded due to not being collected in the same FFPE primary tumor samples.[16,17]

Machine learning models

We used 16 clinical and pathological features from health records and 12 histopathological features to construct the ML models. The 55-age cutoff was by average 52–58 diagnosis age of Vietnamese oral carcinoma,[18–20] similar to a report in Taiwan.[21] We scored all features into 1–8 scores for modeling. Minor feature 'levels with low case numbers were integrated or removed to reduce the total levels and sharpen features. From the genomic subtypes obtained from the clustering of transcriptomics data, we merged clusters and then used as labels. We implemented Gaussian Naive Bayes (GaussianNB), Multinomial Naive Bayes (MultinomialNB), Complement Naive Bayes (ComplementNB), Bernoulli Naive Bayes (BernoulliNB), XGBoost, Support Vector Machine (SVM), Logistic Regression (LR) classification models to perform the subtype classification task. To account for the limitation in dataset size, we implemented a leave-one-out cross-validation (LOOCV) strategy. Hyperparameters tuning was performed for each model to select the best combination of parameters. Model predictive performance was measured by the LOOCV accuracy and area-under-the-curve (AUC).

Suggesting novel biomarkers of OSCC groups and survival analysis

From the upregulated gene lists in each group, we selected 12 genes with the highest log2 fold changes, reported roles in various types of cancer but not clearly in OSCC by Human Protein Atlas with immunohistochemistry (IHC) images of the designated genes in HNSCC primary tumors (medium-to-high expression) (www.proteinatlas.org, v.23). These makers also passed gene-protein reliability scores (Enhanced, Supported or Approved). Using HNSCC from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov), survival analysis was carried out based on the gene expression of suggested markers by GEPIA2 with a log-rank test.[22]

Statistical analyses

Data was analyzed by JASP v0.18.3. Multiple unpaired *t*-test and Hom-Sidak correction was used to compared groups. Differential gene expression analysis optimized the number of genes with adjusted *P*-values <0.05. The interrater reliability of the histologic scores obtained from the two oral pathologists was evaluated using Kappa statistics with $\kappa = 0.81–1.00$.[23]

Results

Gene expression identified 2 subgroups of OSCC samples

Consensus clustering of OSCC RNA-seq identified 2—3 reasonable clusters (Fig. 2A). Gene expression of OSCC well-known makers including Catenin β-1 (CTNNB1), Laminin subunit γ-1 (LAMC1), Ki-67 (MKI67) upregulated in cluster 1—3 and P-Cadherin (CDH3), Laminin subunit γ-2 (LAMC2), L-myc-1 proto-oncogene protein (MYCL), Vascular endothelial growth factor A (VEGFA) upregulated in cluster 2 (Supp. Fig. S1). Cluster 1 and 3 shared the similar expression of these markers, then were merged into subgroup 1 (71 cases) and cluster 2 as subgroup 2 (30 cases) for further investigation (Fig. 2B). Differential analysis of RNA-seq data clearly distinguished the expression pattern of these 2 subgroups of OSCC (Fig. 2C and D).

Different pathways and cell types highlighted distinct biological processes

Cell type deconvolution revealed the cellular components of OSCC samples (Fig. 3A). Group 1 enhanced significantly CD8 T-cells, dendritic cells, fibroblasts, endothelial cells; group 2 enriched CD4 T-cells, malignant cells (mutated epithelial cell signatures).

Enrichment analysis elucidated the contrast functions in each subgroup. In KEGG pathways, subgroup 1 increased immune responses, apoptosis, p53 signaling. Meanwhile, autophagy, cell metabolism and cycle were upregulated in subgroup 2 (Fig. 3B). In HALLMARK, GO and C4—C6 MSigDB cancer gene-sets pathways, subgroup 1 increased epithelial—mesenchymal transition, p53 signaling while subgroup 2 boosted cell division and metabolism (Fig. 3C and Supp. Fig. S2).

Clinical-histopathological characteristics of 2 OSCC subgroups

Clinical-histopathological features from medical records and scored H&E staining identified distinguished natures of OSCC groups (Fig. 3D and E, Supp. Fig. S3). By medical records, subgroup 1 was characterized by older (≥55-year-old) almost male, higher family cancer history, alcohol drinking, pathological grade and occult lymph node metastasis. They were lower in clinical/pathological tumor size, clinical node metastasis and clinical/pathological stage. Subgroup 2 was characterized by younger (<55-year-old), increased female numbers, higher clinical/pathological tumor size, clinical node metastasis and clinical/pathological stage. This subgroup was lower in family/drinking history, pathological grade and occult lymph node metastasis.

H&E staining of local primary tissues regarding RNA-seq demonstrated the different pictures (Figs. 3D—F and 4A, Supp. Fig. S3). Subgroup 1 enhanced lower depth of invasion (≤5 mm), less keratinization, moderate lymphoplasmacytic infiltration. They were more exophytic growth patterns, high nuclear polymorphism, intermediate Bryne scores, higher WHO system grades, more tumor satellites of worst pattern of invasion, and higher BrandweinGensler risk level. Subgroup 2 improved more endophytic growth pattern, depth of invasion (>5 mm), highly keratinized, marked lymphoplasmacytic infiltration, large separate islands of worst pattern of invasion. This subgroup had less nuclear polymorphism, lower Bryne scores, WHO systems grades, and BrandweinGensler risk levels.

Taken-together, OSCC was categorized into subgroup 1 with older drinking males, more severe histopathologic features, immune response and apoptosis/p53 signatures. Subgroup 2 was denoted by younger less-drinking, more clinical/pathological features, cell division/repair and malignant signatures.

Machine learning model predicted OSCC subgroups

To ascertain whether clinical-histopathological characteristics and their genomic subgroups are synchronized, we predicted the OSCC genomic subtypes for our 101 samples by implementing a LOOCV classification system using 6 ML models. We used the genomic subtypes training labels for the model and a binary representation of the clinical-histopathological scores as training features. The accuracy of each model was measured to assess the predictive power of this system (Fig. 4C). Within 6 models, XGBoost and SVM performed highest average accuracies (71.4% and 70.5%, respectively), then LR and MultinomialNB (69.5% and 68.5%, respectively). XGBoost and SVM models also got similar proportion of predicted subgroups on 105 new OSCC non-RNA-seq samples using clinical-histopathological characteristics while other models perform skew results (Fig. 4D). AUC presented by receiver operating characteristic (ROC) curve plots of XGBoost and SVM were 0.69 and 0.68, respectively, LR and MultinomialNB were 0.59 and 0.69 respectively (Fig. 4E and Supp. Fig. S4A).

Clinical-histopathological characteristics of 2 predicted subgroups using XGBoost and SVM were similar to subgroups 1 and 2 of OSCC RNA-seq samples (Fig. 4B—F and G, Supp. Fig. S4B). However, XGBoost trended to predict better in pathological diagnosis, occult lymph node metastasis, depth of invasion, lymphoplasmacytic infiltration while SVM predicted better in clinical/pathologic tumor size, clinical node metastasis, pathologic stage, drinking, pattern of invasion, worst pattern of invasion (Supp. Figs. S5 and S6). Hence, XGBoost and SVM were the optimal ML models for OSCC genomic subgroup classification using clinical, pathological and histological information as input.

Suggesting novel biomarkers for OSCC subgroup

From the differential expression gene lists that upregulated in each OSCC subgroup, we selected 12 significant genes (6/group) with the highest foldchange, reported roles in various types of cancer but not clearly in OSCC by Human Protein Atlas (Fig. 5A, Supp. Data S2 and S3). These genes were also approved to enhanced in IHC reliability score (Supp. Fig. S7A). Subgroup 1 significantly upregulated ADNP (Activity-dependent neuroprotector homeobox), HNRNPD (Heterogeneous nuclear ribonucleoprotein D), RESF1/KIAA1551 (Retroelement silencing factor 1), SLAIN2 (SLAIN

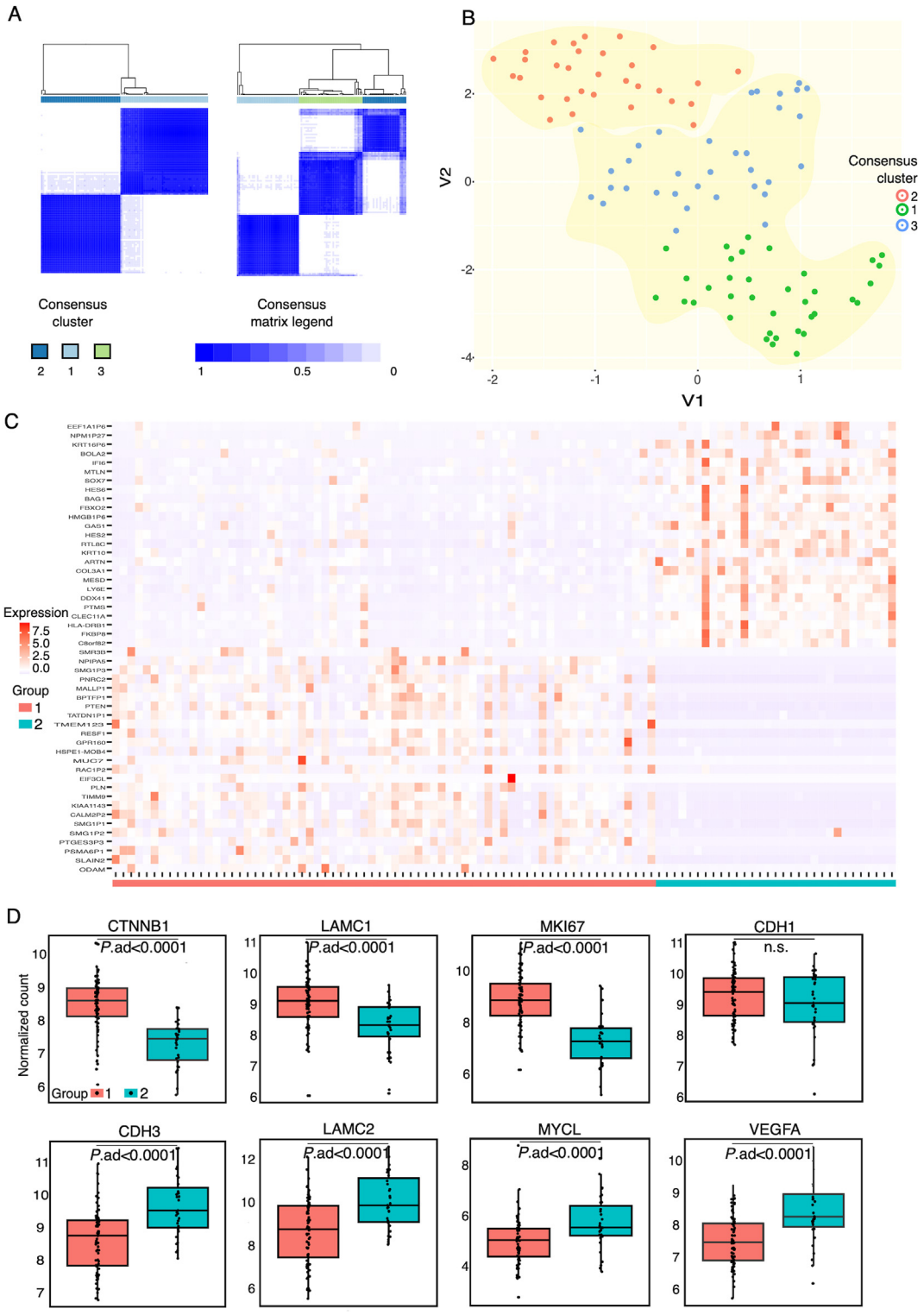

Figure 2 OSCC RNA-seq data was divided into 2 groups. (A) Consensus matrix heatmaps of 101 OSCC RNA-seq samples that were clustered by consensus for k = 2 (left panel) and k = 3 (right panel). The consensus degree was represented by color gradients ranging from 0 to 1. (B) UMAP plot of 101 OSCC RNA-seq samples based on gene expression. Cluster 1 and 3 were grouped into group 1, cluster 2 as group 2. (C) Heatmap of top 50 significant differential genes in 2 subgroups of OSCC. The scaled expression degree was represented by colour gradients. (D) Expression comparison of selected known OSCC markers in 2 subgroups of OSCC RNA-seq samples. OSCC, oral squamous cell carcinoma; *P*.ad, adjusted *P*-values; n.s., not significant.

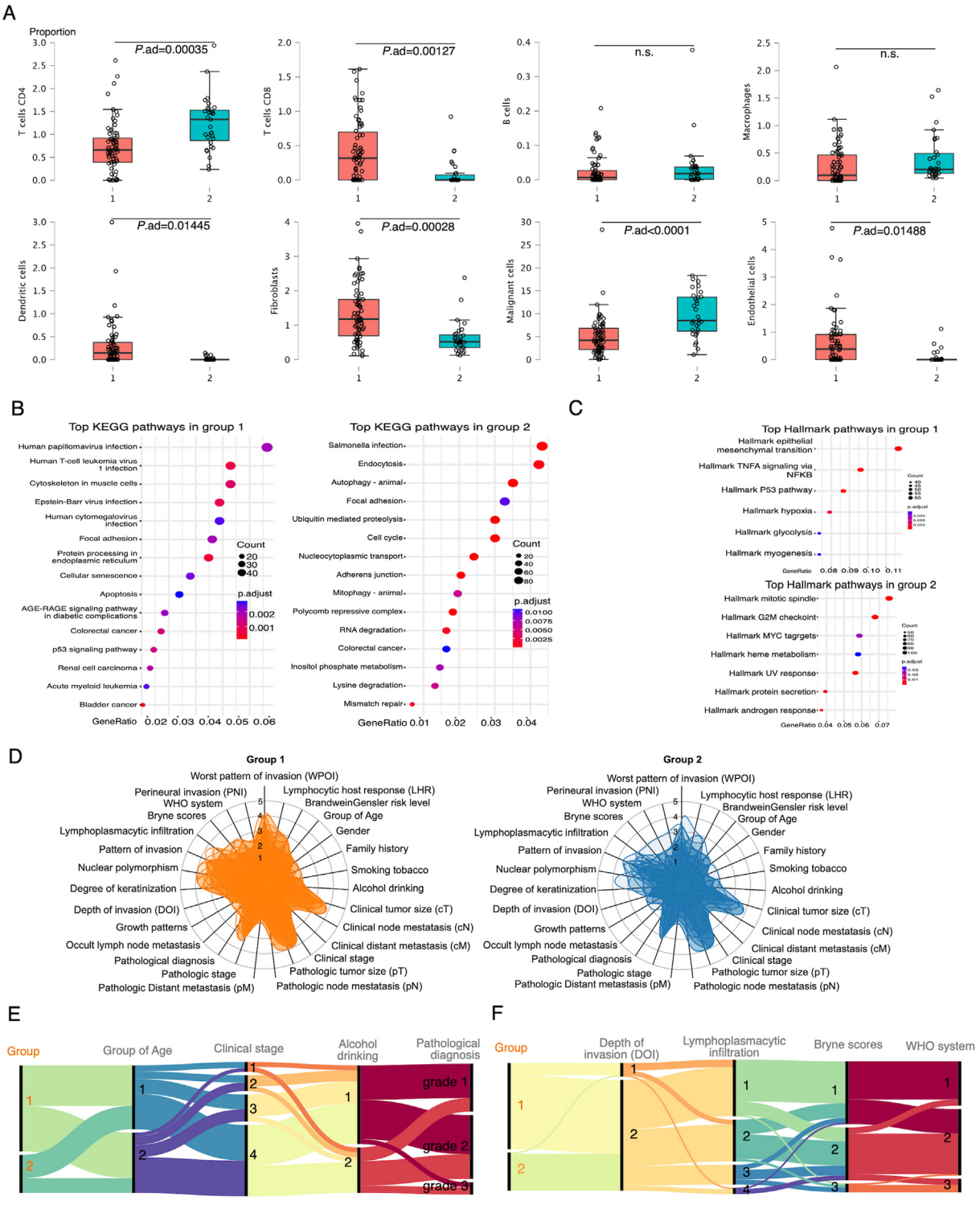

Figure 3 Characteristics of 2 OSCC sub groups. (A) CIBERSORTX deconvolution of proportions of different cell population in 2 subgroups of OSCC RNA-seq samples. Myocytes and mast cells were not shown (not significant). (B) Enrichment analysis of top KEGG pathways upregulated in group 1 (left panel) and group 2 (right panel) of OSCC RNA-seq samples. (C) Enrichment analysis of top HALLMARK pathways upregulated in group 1 (upper panel) and group 2 (lower panel) of OSCC RNA-seq samples. (D) Radar plots of distribution of clinical and histopathological features of group 1 (left panel) and group 2 (right panel) of OSCC RNA-seq samples. Primary tumor site features were not included for the clear and concise data visualization. (E, F) Alluvial diagrams of selected features from health records (E) and H&E staining (F) of 2 subgroups of OSCC RNA-seq samples (See supplementary Fig.S3 for full information). OSCC, oral squamous cell carcinoma; *P*.ad, adjusted *P*-values; n.s., not significant.

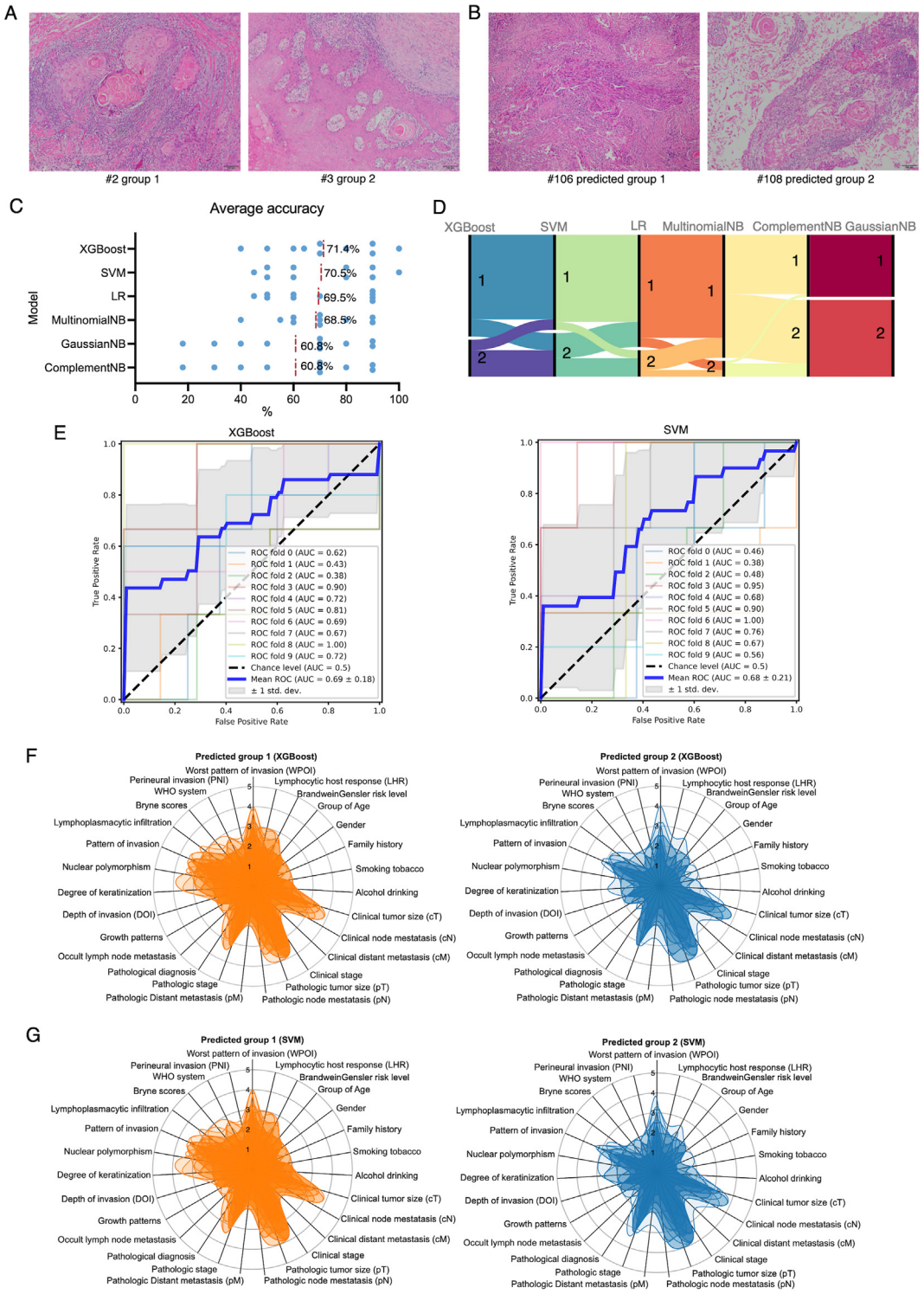

Figure 4 Building machine learning models for OSCC subgroup prediction. (A) H&E staining of cases #2 targeted group 1 (left panel) and #3 targeted group 2 (right panel) of OSCC RNA-seq sample (Scale bar 100 μm). (B) H&E staining of cases #106 predicted group 1 (left panel) and #108 predicted group 2 (right panel) of OSCC non-RNA-seq samples. These samples were predicted identically by using XGBoost and SVM models. (C) Average accuracy of 6 different machine learning models. 0–9 folds were performed to calculate the accuracy. (D) Alluvial diagrams of the distribution of predicted groups (group 1 and group 2) of OSCC non-RNA-seq samples using 6 different models. (E) Receiver operating characteristic (ROC) curve plots of XGBoost (left panel) and SVM (right panel) models and area-under-the-curve (AUC) values. 0–9 folds were performed to calculate the AUC. (F–G) Radar plots of distribution of clinical and histopathological features of predicted group 1 (left panel) and predicted group 2 (right panel) of OSCC RNA-seq samples using XGBoost (F) and SVM (G) models. Primary tumor site features were not included for the clear and concise data visualization. OSCC, oral squamous cell carcinoma; GaussianNB, Gaussian Naive Bayes; MultinomialNB, Multinomial Naive Bayes; ComplementNB, Complement Naive Bayes, BernoulliNB, Bernoulli Naive Bayes; SVM, Support Vector Machine; LR, Logistic Regression.

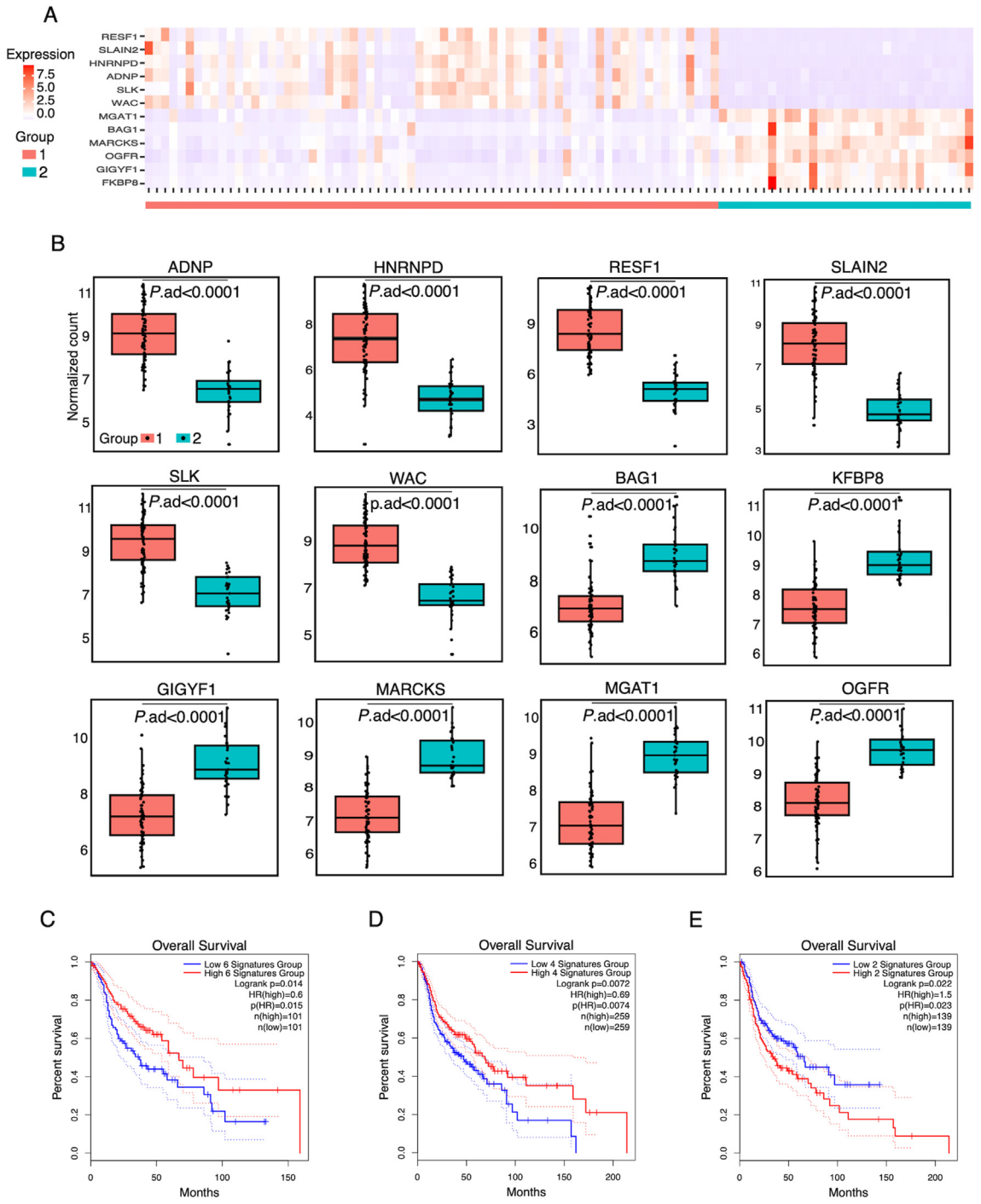

Figure 5 Novel markers of OSCC sub groups were investigated. (A) Heatmap of 12 significant differential suggested novel markers in 2 subgroups of OSCC. The scaled expression degree was represented by colour gradients. (B) Expression comparison of 12 suggested novel markers in 2 subgroups of OSCC RNA-seq samples (*P*.ad, adjusted *P* values). (C) Survival analysis of signature genes (ADNP, HNRNPD, RESF1/KIAA1551, SLAIN2, SLK and WAC) on TCGA-HNSCC tumor samples. (D) Survival analysis of signature genes (BAG1, FKBP8, GIGYF1 and OGFR) on TCGA-HNSCC tumor samples. E Survival analysis of signature genes (MARCKS and MGAT1) on TCGA-HNSCC tumor samples. HNSCC, head and neck squamous cell carcinoma; TCGA, The Cancer Genome Atlas; *P*.ad, adjusted *P*-values.

motif family member 2), SLK (STE20 like kinas), WAC (WW domain containing adaptor with coiled-coil). Subgroup 2 significantly upregulated BAG1 (BAG cochaperone 1, anti-apoptotic activity), FKBP8 (FKBP prolyl isomerase 8), GIGYF1 (GRB10 interacting GYF protein 1), MARCKS (substrate for protein kinase C), MGAT1 (Myristoylated alanine-rich protein kinase C substrate), OGFR (Opioid growth factor receptor) (Fig. 5B).

Finally, we analyzed the overall survival of TCGA-HNSCC patients with the combination of expression or single gene expression using Kaplan–Meier plots (Fig. 5C–E, Supp. Fig. S7B). Interestingly, survival analysis of subgroup 1 markers (6 genes) demonstrated high cumulative survival probability with 40% up to 150 months (>6 years) in high expression group (Fig. 5C). In subgroup 2, combination of BAG1, FKBP8, GIGYF1 and OGFR high expression increased the HNSCC cumulative survival probability up to 40% at 150 months and 20% at 200 months (>8 years) (Fig. 5D). In contrast, MARCKS and MGAT1 combination reduces the cumulative survival probability to only 20-10% after 150–200 months (Fig. 5E). All the above findings indicated the different roles of these novel suggested markers in OSCC genomic subgroups.

Discussion

For the first time, we presented the use of ML combined with genomic data decoding and clinical-histopathological analysis to characterize Vietnamese OSCC patients allowing for a less expensive clinical histology-based method of cancer investigation. SVM, effective with small datasets, identifies optimal hyperplanes for binary classification and handles high-dimensional spaces using Kernel functions, and has been used successfully in the subgroup of rheumatoid arthritis with 45 samples.[24] XGBoost, with advantages like higher accuracy, robustness with missing data, and parallelization.[25]

The study evaluates gene expression within the tumor microenvironment, bridging molecular factors to clinical and anatomical aspects in OSCC. We identified two distinct subgroups based on RNA-seq data, consistent with classical clustering seen in oral premalignant lesions.[26] The concept of addressing interpatient tumor heterogeneity and subclassifying disease based on parameters affecting prognosis, predicting susceptibility to immunotherapy, and achieving optimal therapy for each case has reinforced the need for personalized cancer management.[27,28] Our research indicates that subgroup 2, characterized by a higher proportion of female patients, lower alcohol consumption history, larger tumor size, and increased clinical node metastasis and clinical/pathological stage, presents a phenotype that may be considered as a potential candidate for immunotherapeutic interventions. Therefore, we selected 12 significant genes with the highest fold change.

In subgroup 1, novel markers are mainly in apoptosis and transcription regulation functions with reported roles in liver, lung, renal and pancreatic cancers. ADNP, HNRNP and SLAIN2 are oncogenes impacting the development and resistance of bladder, ovarian and colorectal cancers.[29–31] RESF1, SLK and WAC play critical roles in tumor promoter and metastasis.[32–34] In subgroup 2, novel markers are characterized by autophagy, cell metabolism and cycle, and RNA repair. BAG-1, FKBP8, GIGYF1 and OGFR are anti-apoptotic proteins correlating with the prognosis of ovarian, kidney renal and gastric cancers. It holds promise as a prognostic marker and represents an intriguing therapeutic target.[35–38] MARCKS and MGAT1 (a novel transcriptional target of the Wnt/β-catenin pathway) involved in cell processes like adhesion and motility, contribute to cancer development, metastasis, and treatment resistance by promoting cancer stem cell renewal and immunosuppression.[39,40] These markers, although novel in the context of OSCC, exert critical roles across various other cancers, significantly impacting prognosis, metastasis, and drug response. Their distinct gene expression patterns differentiated 2 subgroups with different combined roles in the survival of HNSCC patients, suggesting their potential as a promising treatment strategy for managing OSCC. However, these markers require further experimental validation including IHC imaging data and functional studies to establish their diagnostic and prognostic value.

Our study contributed insights into the tumor microenvironment of OSCC and provided an accurate ML model for predicting biomarkers using only clinical-histopathological features. These findings highlight the potential of integrating advanced technologies like machine learning with traditional diagnostic methods to improve the understanding and management of OSCC in Vietnamese patients. Novel biomarkers were suggested for each OSCC subgroup, providing potential targets for future research and clinical applications.

Declaration of competing interest

The authors declare that there are no conflicts of interest.

Acknowledgments

This study was granted by the Department of Science and Technology (DOST), Ho Chi Minh City, Vietnam (No. 07/2022/HĐ-QKHCN and 1038/QĐ-SKHCN).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jds.2024.08.013.

References

1. Shetty KSR, Kurle V, Greeshma P, et al. Salvage surgery in recurrent oral squamous cell carcinoma. *Front Oral Health* 2021;2:815606.
2. Su SC, Lin CW, Liu YF, et al. Exome sequencing of oral squamous cell carcinoma reveals molecular subgroups and novel therapeutic opportunities. *Theranostics* 2017;7:1088–99.
3. Tseng YJ, Wang YC, Hsueh PC, Wu CC. Development and validation of machine learning-based risk prediction models of oral squamous cell carcinoma using salivary autoantibody biomarkers. *BMC Oral Health* 2022;22:534.
4. Alharbi F, Vakanski A. Machine learning methods for cancer classification using gene expression data: a review. *Bioengineering* 2023;10:1–26.
5. Huynh NCN. Analysis of single-cell RNA sequencing data revealed the upregulation of Wnt signaling pathway and potential biomarkers in oral squamous cell carcinoma cancer-associated fibroblasts. *MedPharmRes* 2023;7:15–22.
6. Huynh NCN, Huang TT, Nguyen CTK, Fk L. Comprehensive integrated single-cell whole transcriptome analysis revealed the p-EMT tumor cells-CAFs communication in oral squamous cell carcinoma. *Int J Mol Sci* 2022;23:1–16.

7. Cabassi A, Kirk PDW. Multiple kernel learning for integrative consensus clustering of omic datasets. *Bioinformatics* 2020;36:4789—96.
8. Duong HT, Pham PM, Tran NHB, et al. Optimizing RNA extraction and library preparation from oral squamous cell carcinoma FFPE samples for next-generation RNA sequencing. *Biomed Res and Therapy* 2023;10:5987—94.
9. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 2016;34:525—7.
10. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;15:550.
11. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* 2010;26:1572—3.
12. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545—50.
13. Newman AM, Steen CB, Liu CL, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* 2019;37:773—82.
14. Puram SV, Tirosh I, Parikh AS, et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* 2017;171:1611—24.
15. Bryne M, Koppang HS, Lilleng R, Kjaerheim A. Malignancy grading of the deep invasive margins of oral squamous cell carcinomas has high prognostic value. *J Pathol* 1992;166:375—81.
16. Brandwein-Gensler M, Teixeira MS, Lewis CM, et al. Oral squamous cell carcinoma: histologic risk assessment, but not margin status, is strongly predictive of local disease-free and overall survival. *Am J Surg Pathol* 2005;29:167—78.
17. Julian RS, Woo BM, Rabey EC. Oral cavity and oropharyngeal cancer: etiology, diagnosis and staging. *J Calif Dent Assoc* 2021;49:163—70.
18. Le VQ, Ngo QD, Le TD, Xq N. Evaluation of cervical lymph nodes metastasis and its relationship with features of oral cavity cancer. *Vietnam Med J* 2021;500:249—52.
19. Nguyen HN, Vs L. Clinical symptoms of oral cavity cancer. *Vietnam Med J* 2022;510:33—6.
20. Nguyen VM, Nguyen HL, Tka D. Evaluation of clinical and paraclinical features in patients with carcinoma of oral cavity. *Hue J Med Pharm* 2022;6:56—60.
21. Chen TC, Chang HL, Yang TL, et al. Impact of dysplastic surgical margins for patients with oral squamous cell carcinoma. *Oral Oncol* 2019;97:1—6.
22. Kirtane K, Rodriguez CP. Postoperative combined modality treatment in high risk resected locally advanced squamous cell carcinomas of the head and neck (HNSCC). *Front Oncol* 2018;8:588.
23. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22:276—82.
24. Orange DE, Agius P, DiCarlo EF, et al. Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data. *Arthritis Rheumatol* 2018;70:690—701.
25. Guan X, Du Y, Ma R, et al. Construction of the XGBoost model for early lung cancer prediction based on metabolic indices. *BMC Med Inf Decis Making* 2023;23:107.
26. Foy JP, Bertolus C, Ortiz-Cuaran S, et al. Immunological and classical subtypes of oral premalignant lesions. *OncoImmunology* 2018;7:e1496880.
27. Hsieh Y, Wu Y, Cheng S, et al. Single-cell RNA sequencing analysis for oncogenic mechanisms underlying oral squamous cell carcinoma carcinogenesis with candida albicans infection. *Int J Mol Sci* 2022;23:4833.
28. Huynh NCN, Pham AL, Pham NVT, Phn L. Differential gene expression analysis of TCGA mRNA sequencing data from male patients with and without lymph node metastasis in tongue cancer. *Arch Orofac Sci* 2024;9:1—14.
29. Wang X, Peng H, Zhang G, et al. ADNP is associated with immune infiltration and radiosensitivity in hepatocellular carcinoma for predicting the prognosis. *BMC Med Genom* 2023;16:178.
30. Kumar V, Kumar A, Kumar M, Lone MR, Mishra D, Chauhan SS. NFkappaB (RelA) mediates transactivation of hnRNPD in oral cancer cells. *Sci Rep* 2022;12:5944.
31. Zhuang M, Zhao S, Jiang Z, et al. MALAT1 sponges miR-106b-5p to promote the invasion and metastasis of colorectal cancer via SLAIN2 enhanced microtubules mobility. *EBioMedicine* 2019;41:286—98.
32. Majocha MR, Jackson DE, Ha NH, et al. Resf1 is a compound G4 quadruplex-associated tumor suppressor for triple negative breast cancer. *PLoS Genet* 2024;20:e1011236.
33. Yang Z, Liu Z, Lu W, Guo H, Chen J, Zhang Y. LncRNA WAC-AS1 promotes osteosarcoma Metastasis and stemness by sponging miR-5047 to upregulate SOX2. *Biol Direct* 2023;18:74.
34. Al-Zahrani KN, Abou-Hamad J, Cook DP, et al. Loss of the Ste20-like kinase induces a basal/stem-like phenotype in HER2-positive breast cancers. *Oncogene* 2020;39:4592—602.
35. Wu H, Liu M, He Y, Meng G, Guo W, Guo Q. Expression of BAG1 is associated with prognosis in kidney renal clear cell carcinoma based on bioinformatics. *BMC Cancer* 2021;21:160.
36. Zhang J, Yin Y, Wang J, et al. Prohibitin regulates mTOR pathway via interaction with FKBP8. *Front Med* 2021;15:448—59.
37. Zhu L, Yao Z, Luo Q, et al. Low expression of GIGYF1 inhibits metastasis, proliferation, and promotes apoptosis and autophagy of gastric cancer cells. *Int J Med Sci* 2023;20:1038—45.
38. Hankins GR, Harris RT. The opioid growth factor in growth regulation and immune responses in cancer. *Adv Neurobiol* 2024;35:45—85.
39. Chiu CL, Zhao H, Chen CH, Wu R, Brooks JD. The role of MARCKS in metastasis and treatment resistance of solid tumors. *Cancers* 2022;14:4925.
40. Akiva I, Birgul Iyison N. MGAT1 is a novel transcriptional target of Wnt/beta-catenin signaling pathway. *BMC Cancer* 2018;18:60.