Oral Oncology Reports 5 (2023) 100018

Contents lists available at ScienceDirect

Oral Oncology Reports

journal homepage: www.journals.elsevier.com/oral-oncology-reports

Identification of a potential prognostic panel of biomarkers for stratification of head and neck squamous cell carcinoma based on HPV status and TP53 mutational status

Oriana Barros [a,c], Rita Ferreira [b], Vito G. D'Agostino [d], Francisco Amado [b], Lucio Santos [c], Rui Vitorino [a,b,e,*]

[a] Department of Medical Sciences, Institute of Biomedicine iBiMED, University of Aveiro, 3810-193, Aveiro, Portugal
[b] LAQV-REQUIMTE, Department of Chemistry, University of Aveiro, Campus Universitário de Santiago, 3810-193, Aveiro, Portugal
[c] Experimental Pathology and Therapeutics Group, Research Center of IPO Porto (CI-IPOP), RISE@CI-IPOP (Health Research Network) and Surgical Department of Portuguese Oncology Institute of Porto (IPO Porto), Porto Comprehensive Cancer Center (Porto.CCC), Portugal
[d] Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Via Sommarive 9, 38123, Trento, Italy
[e] UnIC, Department of Surgery and Physiology, Faculty of Medicine, University of Porto, 4200-319, Porto, Portugal

ARTICLE INFO

Keywords:
Head and neck squamous cell carcinoma
Prognosis
Biomarkers
Precision medicine

ABSTRACT

Head and Neck Squamous Cell Carcinoma (HNSCC) is a malignant cancer with poor prognosis. Currently, the prognosis of HNSCC is determined by clinical and histopathological criteria. This work focused on identifying a panel of genes that have the potential to be used for prognosis of HNSCC and to improve patient stratification for treatment. To this end, a bibliometric analysis (VosViewer) was applied to identify candidate genes that were further characterized by applying several bioinformatics tools (UALCAN, ToPP). The prognostic potential of the genes of interest was evaluated using the univariate and the multivariate Cox proportional regression models and the transcriptional expression analysis among HNSCC and normal tissues. In HNSCC, the transcriptional levels of candidate genes, were analyzed in HPV-driven HNSCC, HPV-non-driven HNSCC, TP53-mutant HNSCC and TP53-nonmutant HNSCC for selecting the best set of genes for discrimination of HNSCC based on both HPV status and TP53 mutational status. These analyses revealed a signature based on four genes with greater HNSCC prognostic potential: CDKN2A, TGFB1, CD44 and MMP9, being p16 the sole biomarker currently tested. In the future, a molecular signature could facilitate the stratification of patients into high- and low-risk groups as well the wiser adjustment of therapies to each individual response allowing a personalized treatment.

1. Introduction

Head and neck cancer is the sixth most common cancer worldwide and is expected to increase in incidence 30% until 2030. About 90% begins in the squamous cells on the surface of the inner mucosa of that region. Classification can be done according to the place of origin: oral cavity, tongue, salivary glands, pharynx, larynx, nasal cavity, and paranasal sinuses [1]. Because it comprises numerous subtypes, each with a different molecular fingerprint, the characterization of HNSCC is extremely complex. The prognosis in stages I or II is favourable, presenting a cure rate of 80% in stage I and 65% in stage II, with exclusive surgery or radiotherapy treatments. When locally advanced disease, stages III or IV, the 5-year survival rate is less than 50%. The treatment at advanced stages is multimodal and there is a high risk of local relapse and/or disease at a distance [2]. Thus, there is an urgent need of reliable tools for HNSCC prognosis with potential for patient risk stratification.

Molecular signatures have gained increasing importance for their potential in stratifying patients according to their prognosis (prognostic biomarkers) or in predicting the response of a given patient to a treatment (predictive biomarkers). A gene expression signature consists of a set of genes that are correlated with a certain variable of interest, such as diagnosis, treatment response or prognosis. With the advent of multi-omics approaches there has been an explosion of molecular signatures in HNSCC, but none have yet been translated into clinical practice [3] [–] [6]. Tissue $p16^{INK4A}$ has been used as single prognosis biomarker in clinical practice and when combined with TNM staging establish the HNSCC prognosis [7]. $p16^{INK4A}$ immunohistochemistry (IHC) alone showed a sensitivity of 94% (95% CI: 91–97%) and a specificity of 83%

* Corresponding author. Department of Medical Sciences, Institute of Biomedicine iBiMED, University of Aveiro, 3810-193, Aveiro, Portugal.
E-mail address: rvitorino@ua.pt (R. Vitorino).

https://doi.org/10.1016/j.oor.2023.100018
Received 26 December 2022; Accepted 3 March 2023

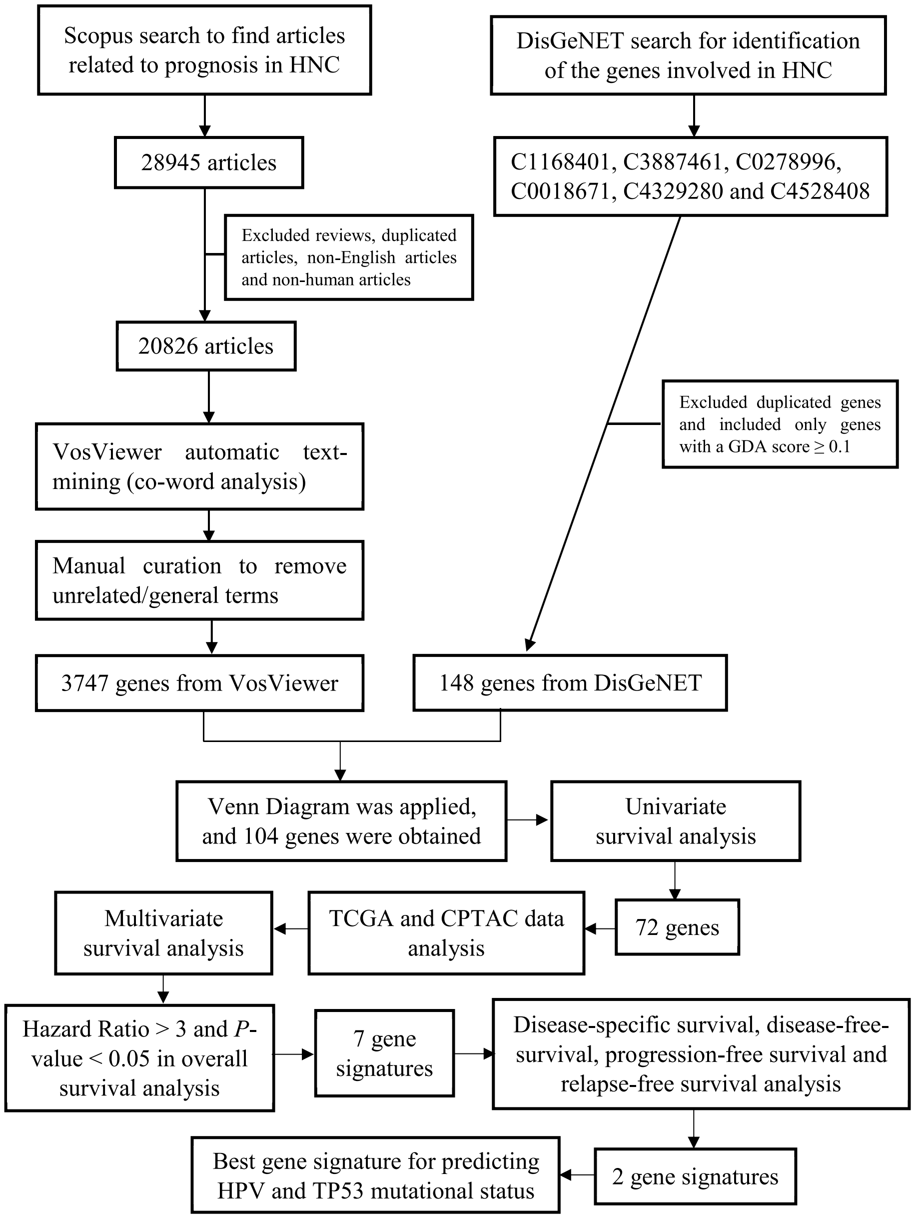


Fig. 1. Flowchart of the methodology of this study.

(95% CI: 78–88%). Although highly sensitive, is moderately specific for HPV-driven HNSCC. Therefore, if p16^{INK4A} IHC is positive, HPV-testing must be done to distinguish HPV-driven HNSCC from HPV-non driven cancers [8,9]. There is already some evidence of the role of gene signatures based on HPV status in predicting the sensitivity or resistance of patients to radiotherapy and certain chemotherapy treatments, as well as in risk stratification of patients with HNSCC [3]. Recent studies have also evaluated the potential of TP53 mutations in predicting the prognosis of HNSCC patients. TP53 gene is the most mutated gene in HNSCC. The mutational profile is site-specific and changes with tumour stage. TP53 mutations are associated with a poorer prognosis and resistance to chemotherapy and radiotherapy treatments. p53 protein is the TP53 gene translation product and has a much higher frequency of mutations in HPV-non-driven HNSCC, since the E6 viral proteins encoded by HPV bind to p53 inactivating it, assuming a crucial role in the pathophysiology of HPV-non-driven HNSCC [10,11].

Our work aims to identify a gene panel that act synergistically in establishing prognosis in HNSCC patients with potential to allow stratification into HPV-driven HNSCC and HPV-non-driven HNSCC, as well as TP53-mutant HNSCC and TP53-nonmutant HNSCC. An algorithm based on a bibliometric analysis and bioinformatic tools was created as a cornerstone for the identification of prognostic biomarkers of HNSCC.

2. Methods

Search strategy and refined data. The publication search was performed using Scopus. The literature type was defined as "all types" and the keywords were set to cover as many results related to the topic under study as possible. The search formula used was the following: "((head AND neck AND cancer) OR (head AND neck AND squamous AND cell AND carcinoma) OR (oral AND squamous AND cell AND carcinoma) OR (nasopharyngeal AND squamous AND cell AND carcinoma) OR (oropharyngeal AND squamous AND cell AND carcinoma) OR (laryngeal AND squamous AND cell AND carcinoma) OR (lip AND squamous AND cell AND carcinoma) OR (tongue AND squamous AND cell AND carcinoma) OR (paranasal AND sinuses AND squamous AND cell AND carcinoma) AND prognosis))". The

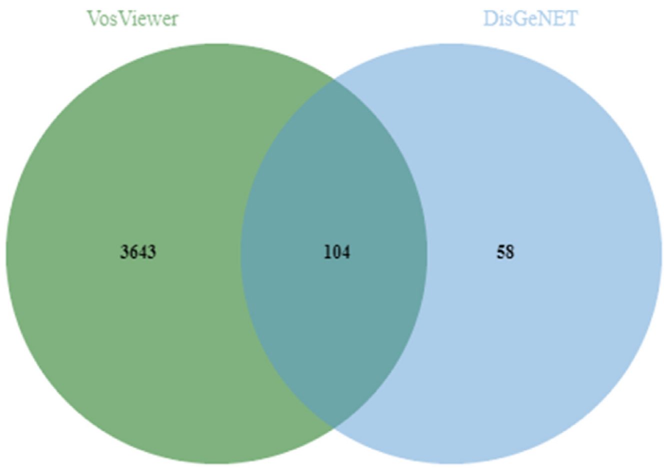


Fig. 2. Venn Diagram of HNSCC common genes identified by VosViewer and DisGeNET.

search referred to scientific articles, excluding review articles to avoid data redundancy. All searches were conducted in July 2022.

Data analysis. Data results were obtained on Scopus using the above search formula and exported in a CSV format that was uploaded in VosViewer (v1.6.15) to construct a bibliometric network. In this case, we created a co-occurrence network of all keywords using a full counting method by importing the bibliographic information from Scopus. The established minimum number of occurrences of a keyword was 5 and the number of selected terms was 9499. From the bibliometric network, the relationship between the selected keywords by text-mining was explored. Each node is associated with a different keyword and the node size indicate the occurrence frequency of the term and its relative weight in the network. The lines between nodes represent the strength of interaction between terms, while the different colors allow the identification of different clusters of related keywords. For each keyword, all the genes and their encoded proteins were extracted using UNIPROT (https://www.uniprot.org/). To ensure that only HNSCC – associated genes were integrated into this study, a parallel search was done in DisGeNET (https://www.disgenet.org/). A filter was applied for gene-disease association (GDA) score greater than 0.1, as very low values of GDA score are associated to genes with very limited expression in HNSCC. An interactive Venn diagram (http://jvenn.toulouse.inra.fr/app/index.html), was used to perform an intersection analysis of the genes obtained from VOSViewer and DisGeNET, which allowed us to identify the genes from the bibliometric network with a relationship well established to HNSCC.

Identification of the prognostic genes. Not every gene is transcriptionally active. To get a deeper insight about the patterns of gene expression associated to our genes of interest the corresponding expressed RNAs (TCGA) and proteins (CPTAC) were evaluated in UALCAN (http://ualcan.path.uab.edu/), in GEPIA2 (http://gepia2.cancer-pku.cn/#index) and in tumor prognostic analysis platform (ToPP) (http://www.biostatistics.online/topp/index.php) [12–14]. UALCAN is a bioinformatic tool based on level 3 RNA-seq and clinical evidence from TCGA and CPTAC database. It is a very versatile web tool that allows cancer multi-omics data to be made public in a more intuitive way. In this study, UALCAN was used to understand how the transcriptional levels of candidate genes are modulated by HPV status and TP53 mutational status and to evaluate the protein expression of the proteins encoded by the genes of interest in HNSCC [15]. GEPIA2 was used to complement the results obtained at UALCAN in terms of analysing the differential expression of genes of interest in tumors and normal tissues using "Expression analysis – Differential genes" module with the following conditions: HNSCC dataset,

Table 1
Genes selected from Venn Diagram.

Gene symbol	Protein encoded by the gene of interest	Expression in HNSCC tissue in comparison to normal tissue		Overall survival analysis[1]
		Gene[1]	Protein[2]	
ABCB1	ATP-dependent translocase ABCB1	↓	ns	HR = 0.549 (0.420–0.717)*
ABCG2	Broad substrate specificity ATP-binding cassette transporter ABCG2	↓	↓	HR = 1.33 (1.02–1.75)*
AKT1	RAC-alpha serine/threonine-protein kinase	↑	↑	HR = 1.56 (1.19–2.03)*
ALDH1A1	Aldehyde Dehydrogenase 1 Family Member A1	↓	↓	Ns
ALDH2	Aldehyde Dehydrogenase 2 Family Member	↓	↓	HR = 0.549 (0.384–0.785)*
ANO1	Anoctamin-1	↑	↑	HR = 1.76 (1.35–2.32)*
AREG	Amphiregulin	ns	N/A	HR = 1.72 (1.32–2.25)*
ATM	ATM Serine/Threonine Kinase	ns	↑	HR = 0.621 (0.424–0.908)*
ATP7B	Copper-transporting ATPase 2	↓	N/A	HR = 1.51 (0.999–2.28)*
B2M	Beta-2-Microglobulin	↑	↑	HR = 1.41 (1.05–1.88)*
BAP1	Ubiquitin carboxyl-terminal hydrolase BAP1	ns	ns	HR = 0.667 (0.496–0.898)*
BCL2	BCL2 Apoptosis Regulator	↓	ns	HR = 0.6 (0.444–0.811)*
BCL2L1	Bcl-2-like protein 1	↓	N/A	Ns
BMI1	Polycomb complex protein BMI-1	ns	N/A	ns
BRAF	B-Raf Proto-Oncogene Serine/Threonine Kinase	↓	ns	ns
CASP8	Caspase 8	↑	↑	HR = 1.49 (1.01–2.2)*
CCNA1	Cyclin A1	↑	N/A	HR = 1.77 (1.35–2.31)*
CCNA2	Cyclin A2	↑	↑	HR = 1.69 (1.06–2.67)*
CCNB1	Cyclin B1	↑	↑	HR = 1.95 (1.15–3.29)*
CCND1	Cyclin D1	ns	↑	HR = 1.8 (1.34–2.43)*
CD44	Cluster of Differentiation 44	↑	↑	HR = 1.63 (1.15–2.31)*
CD274	Cluster of Differentiation 274	↑	↑	HR = 1.36 (1.04–1.78)*
CDK4	Cyclin-dependent kinase 4	↑	ns	No
CDKN2A	Cyclin-Dependent Kinase Inhibitor 2A	↑	↓	HR = 0.562 (0.401–0.787)*
CKAP4	Cytoskeleton-associated protein 4	↑	↑	No
CSF3	Granulocyte colony-stimulating factor	ns	N/A	HR = 1.38 (1.05–1.81)*
CTLA4	Cytotoxic T-lymphocyte protein 4	↑	N/A	HR = 0.573 (0.439–0.747)*
CTNNB1	Catenin Beta 1	↓	↓	HR = 1.54 (1.1–2.16)*
CTTN	Src substrate cortactin	ns	↑	HR = 1.79 (1.37–2.34)*
CYLD	Ubiquitin carboxyl-terminal hydrolase CYLD	ns	ns	HR = 0.796 (0.607–1.04)*
CYP1A1	Cytochrome P450 Family 1 Subfamily A Member 1	↓	N/A	HR = 1.41 (1.04–1.9)*
DPYD	Dihydropyrimidine dehydrogenase [NADP(+)]	↓	ns	HR = 1.55 (1.07–2.25)*
EGFR	Epidermal Growth Factor Receptor	↑	↑	HR = 1.6 (1.11–2.31)*
EP300	Histone acetyltransferase p300	ns	↑	HR = 0.713 (0.466–1.09)*
ERBB2	Receptor tyrosine-protein kinase erbB-2	↓	↓	HR = 0.68 (0.519–0.892)*

(continued on next page)

Table 1 (*continued*)

Gene symbol	Protein encoded by the gene of interest	Expression in HNSCC tissue in comparison to normal tissue		Overall survival analysis[1]
		Gene[1]	Protein[2]	
ERBB3	Receptor tyrosine-protein kinase erbB-3	↓	ns	HR = 0.658 (0.504–0.86)*
ERCC1	DNA excision repair protein ERCC-1	ns	ns	HR = 1.29 (0.955–1.75)*
ERCC2	General transcription and DNA repair factor IIH helicase subunit XPD	↑	↑	HR = 1.66 (1.23–2.24)*
FANCA	Fanconi anemia group A protein	↑	ns	ns
FANCB	Fanconi anemia group B protein	↑	N/A	HR = 0.575 (0.355–0.932)*
FANCC	Fanconi anemia group C protein	↑	N/A	HR = 0.485 (0.303–0.777)*
FANCD2	Fanconi anemia group D2 protein	↑	ns	HR = 0.547 (0.345–0.868)*
FANCE	Fanconi anemia group E protein	↑	N/A	HR = 0.597 (0.401–0.888)*
FANCF	Fanconi anemia group F protein	ns	N/A	ns
FANCG	Fanconi anemia group G protein	↑	N/A	HR = 0.678 (0.477–0.962)*
FANCI	Fanconi anemia group I protein	↑	ns	HR = 0.693 (0.486–0.988)*
FANCL	E3 ubiquitin-protein ligase FANCL	↑	N/A	HR = 0.555 (0.36–0.856)*
FANCM	Fanconi anemia group M protein	↑	N/A	HR = 0.669 (0.465–0.961)*
FAT1	Very long-chain fatty acid transport protein	↑	↑	HR = 1.62 (1.15–2.29)*
FBXW7	F-box/WD repeat-containing protein 7	↓	N/A	HR = 0.663 (0.488–0.9)*
GNAS	Guanine nucleotide-binding protein G(s) subunit alpha isoforms short	↓	ns	HR = 1.51 (1.15–1.99)*
GPX1	Glutathione peroxidase 1	↓	↓	ns
GSTM1	Glutathione S-Transferase Mu 1	ns	↓	HR = 0.712 (0.507–0.999)*
GSTP1	Glutathione S-Transferase Pi 1	↑	↑	HR = 1.57 (1.19–2.07)*
GSTT1	Glutathione S-Transferase Theta 1	↓	↓	ns
HGF	Hepatocyte growth factor	↓	↓	HR = 0.677 (0.516–0.89)*
HIF1A	Hypoxia Inducible Factor 1 Subunit Alpha	↑	↑	ns
HPGDS	Hematopoietic prostaglandin D synthase	↓	↓	HR = 0.57 (0.429–0.758)*
HRAS	GTPase HRas	ns	ns	ns
IDH2	Isocitrate dehydrogenase [NADP], mitochondrial	↓	↓	ns
IGF1	Insulin-Like Growth Factor 1	↓	↓	ns
IL1A	Interleukin 1 Alpha	↑	ns	HR = 1.67 (1.21–2.31)*
IL6	Interleukin 6	↓	N/A	HR = 1.66 (1.19–2.32)*
KRAS	KRAS Proto-Oncogene, GTPase	ns	ns	ns
MAP2K1	Mitogen-Activated Protein Kinase Kinase 1	↑	↑	HR = 1.6 (1.22–2.08)*
MAP2K2	Mitogen-Activated Protein Kinase Kinase 2	↓	ns	HR = 0.639 (0.45–0.905)*
MAPK1	Mitogen-Activated Protein Kinase 1	ns	ns	Ns
MAPK3	Mitogen-Activated Protein Kinase 3	↓	↓	ns
MAPK9	Mitogen-Activated Protein Kinase 9	ns	↓	HR = 1.63 (1.22–2.18)*
MERTK	MER Proto-Oncogene, Tyrosine Kinase	ns	N/A	ns
MET	Indolethylamine N-methyltransferase	↑	↑	HR = 1.64 (1.24–2.18)*
MGMT	O-6-Methylguanine-DNA Methyltransferase	↓	↓	ns
MLH1	MutL Homolog 1	↓	↑	HR = 0.681 (0.468–0.991)*
MMP2	Matrix Metallopeptidase 2	↑	↑	ns
MMP9	Matrix Metallopeptidase 9	↑	↑	HR = 1.42 (1.01–1.98)*
MTOR	Serine/threonine-protein kinase mTOR	ns	↑	HR = 0.75 (0.575–0.98)*
NFE2L2	Nuclear Factor, Erythroid 2 Like 2	↓	N/A	ns
NOTCH1	Notch Receptor 1	ns	ns	ns
PDCD1	Programmed cell death protein 1	ns	N/A	HR = 0.656 (0.498–0.862)*
PIK3CA	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Alpha	↑	↑	HR = 1.51 (1.05–2.17)*
PIK3CB	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Beta	↑	↓	HR = 0.762 (0.585–0.993)*
PIK3CD	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Delta	↑	↑	HR = 0.75 (0.567–0.992)*
PIK3CG	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Gamma	ns	ns	HR = 0.594 (0.425–0.829)*
PRAME	Melanoma antigen preferentially expressed in tumors	↑	ns	HR = 1.72 (1.31–2.26)*
PTEN	Phosphatase and Tensin Homolog	ns	↑	ns
PTGS2	Prostaglandin-Endoperoxide Synthase 2	↑	↑	HR = 0.55 (0.379–0.799)*
RAC1	Ras-related C3 botulinum toxin substrate 1	ns	↑	ns
RAD51	DNA repair protein RAD51 homolog 1	↑	ns	ns
RARB	Retinoic acid receptor beta	↓	N/A	ns
SMAD4	SMAD Family Member 4	↓	↓	ns
SOX2	SRY-Box Transcription Factor 2	ns	↓	ns
STAT3	Signal Transducer and Activator of Transcription 3	↓	↑	HR = 0.65 (0.468–0.903)*
STAT6	Signal transducer and activator of transcription 6	ns	ns	ns
TGFA	Transforming Growth Factor Alpha	↑	N/A	HR = 1.42 (1.05–1.92)*
TGFB1	Transforming Growth Factor Beta 1	↑	↑	HR = 1.95 (1.35–2.81)*
TNF	Tumour Necrosis Factor	ns	N/A	HR = 0.696 (0.522–0.928)*
TNFRSF10B	Tumor necrosis factor receptor superfamily member 10B	↑	N/A	ns
TP53	Tumour Protein P53	ns	↑	HR = 0.601 (0.387–0.934)*
TP63	Tumour Protein P63	↑	↑	ns
TYMS	Thymidylate synthase	↑	↑	HR = 1.85 (1.13–3.04)*
VEGFA	Vascular Endothelial Growth Factor A	↑	ns	ns
VIM	Vimentin	ns	↓	HR = 1.43 (1.01–2.02)*

(*continued on next page*)

Table 1 (*continued*)

Gene symbol	Protein encoded by the gene of interest	Expression in HNSCC tissue in comparison to normal tissue		Overall survival analysis[1]
		Gene[1]	Protein[2]	
XRCC1	DNA repair protein XRCC1	↑	↑	HR = 0.464 (0.27–0.798)*
YAP1	Transcriptional coactivator YAP1	↓	↓	ns

Abbreviations: [1], data extracted from ToPP; [2], data extracted from UALCAN; HR, hazard ratio; N/A, no results available; ns, not statistically significant; *, statistically significant (*P*-value <0.05).

log2FC (fold change) cutoff 1, and *P*-value cutoff 0.01. The transcriptional analysis of the genes in tumour and healthy individuals was also evaluated on ToPP (http://www.biostatistics.online/topp/index.php).

Gene Survival-Associated Analysis. Initially, the impact of genes identified in the Venn Diagram on the overall survival (OS) of HNSCC patients was assessed using the online ToPP and the TCGA-HNSC dataset (521 samples). ToPP is a user-friendly bioinformatic tool that provides prognostic analysis using multi-omics data and clinical data of 55 tumor types. For survival analysis in ToPP, the "Univariate analysis" function was used, and "Best cutoff" option was selected to slit patients. A risk score or prognostic index (PI) was built based on a linear component of the Cox model, PI = $\beta 1x1 + \beta 2x2 + \ldots + \beta pxp$, where β is a Cox coefficient (risk coefficient) and x is the gene expression value. The database

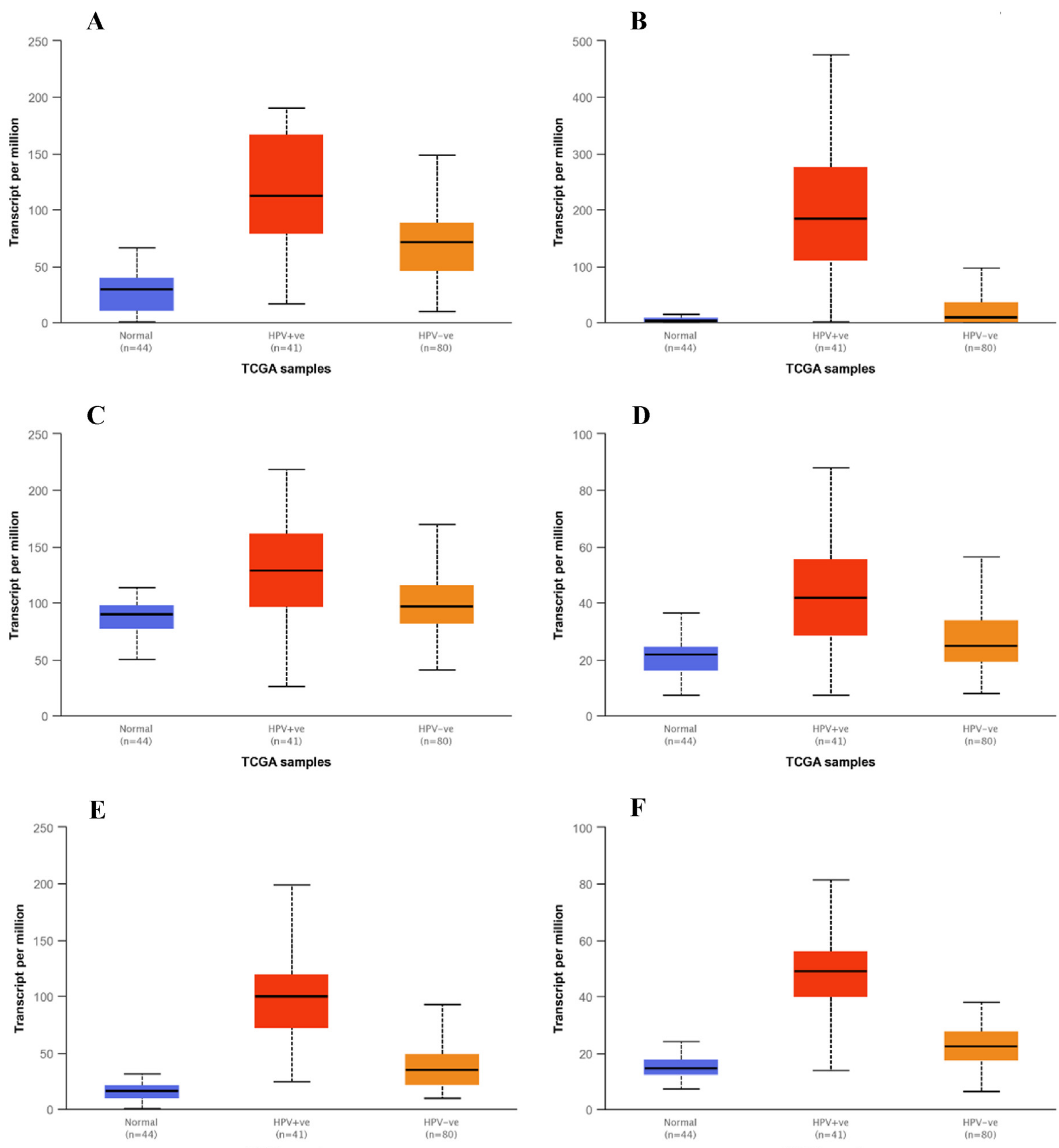


Fig. 3. Overexpressed genes in HPV-driven HNSCC. Box plot of expression of CCNB1 (A), CDKN2A (B), MAP2K2 (C), PIK2CB (D), TYMS (E) and XRCC1 (F) in HPV-driven HNSCC (red), HPV-non-driven HNSCC (orange) and healthy individuals (blue). The significance difference between groups was estimated by Student's *t*-test with *P*-value as shown in Table 2. *P*-value <0.05 was considered to be statistically significant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 2
Significant changes in the transcriptional expression of the selected genes in HPV-non-driven HNSCC, HPV-driven HNSCC and normal tissues (UALCAN).

Gene symbol	Transcript per million (median)			*P*-value from transcriptional analysis		
	Normal tissue	HPV^+ HNSCC tissue	HPV^- HNSCC tissue	Normal vs HPV^+ HNSCC	Normal vs HPV^- HNSCC	HPV^+ vs HPV^- HNSCC
ABCB1	0.667	0.793	0.393	ns	ns	ns
ABCG2	0.756	0.699	0.91	ns	ns	ns
AKT1	68.189	98.363	108.478	1.61×10^{-5}	6.57×10^{-12}	ns
ALDH2	124.012	108.139	67.616	ns	1.89×10^{-7}	1.38×10^{-3}
ANO1	9.242	15.125	34.987	ns	1.90×10^{-5}	9.37×10^{-4}
AREG	17.08	8.142	53.445	4.22×10^{-2}	4.83×10^{-4}	4.70×10^{-8}
ATM	3.318	5.927	5.018	8.03×10^{-5}	2.30×10^{-7}	ns
ATP7B	0.579	0.587	0.784	3.62×10^{-2}	1.69×10^{-2}	ns
B2M	2073.941	5424.817	4470.892	2.75×10^{-2}	8.70×10^{-11}	ns
BAP1	40.266	57.413	49.196	1.30×10^{-7}	6.42×10^{-7}	ns
BCL2	2.287	5.196	1.908	1.97×10^{-4}	ns	7.52×10^{-5}
CASP8	9.37	19.667	16.415	9.81×10^{-11}	1.24×10^{-11}	ns
CCNA1	0.556	0.155	5.301	ns	1.55×10^{-5}	3.75×10^{-5}
CCNA2	9.213	35.959	21.63	5.35×10^{-12}	1.01×10^{-14}	2.38×10^{-5}
CCNB1	29.534	112.231	71.282	1.63×10^{-12}	1.33×10^{-15}	5.66×10^{-7}
CCND1	65.228	26.012	61.132	ns	4.34×10^{-3}	1.93×10^{-2}
CD44	147.878	152.734	304.931	ns	1.89×10^{-12}	1.44×10^{-9}
CD274	1.691	4.299	3.702	9.97×10^{-4}	3.34×10^{-6}	ns
CDKN2A	3.42	184.245	10.04	1.26×10^{-11}	2.01×10^{-7}	2.29×10^{-10}
CSF3	2.8	1.849	3.028	ns	ns	ns
CTLA4	0.479	6.063	2.503	4.09×10^{-9}	4.17×10^{-11}	5.82×10^{-4}
CTNNB1	124.528	147.268	162.971	ns	1.93×10^{-2}	ns
CTTN	105.405	80.577	133.819	ns	5.99×10^{-5}	2.84×10^{-3}
CYLD	13.105	14.16	16.599	4.78×10^{-2}	4.28×10^{-5}	ns
CYP1A1	0.034	0.026	0.051	ns	ns	ns
DPYD	9.65	11.384	9.757	ns	ns	ns
EGFR	20.284	21.466	39.072	ns	8.40×10^{-3}	1.29×10^{-2}
EP300	14.764	21.745	20.82	5.52×10^{-3}	2.03×10^{-6}	ns
ERBB2	84.347	50.597	49.08	ns	2.05×10^{-4}	ns
ERBB3	42.006	40.141	29.222	ns	3.03×10^{-4}	3.12×10^{-3}
ERCC1	46.083	68.678	59.048	6.40×10^{-4}	1.41×10^{-2}	ns
ERCC2	6.888	11.451	11.381	8.79×10^{-8}	2.93×10^{-9}	ns
FANCB	0.321	1.572	0.972	1.80×10^{-10}	$<1 \times 10^{-12}$	9.71×10^{-5}
FANCC	2.178	10.062	4.375	2.55×10^{-15}	1.62×10^{-12}	9.01×10^{-10}
FANCD2	2.915	12.848	5.457	2.15×10^{-12}	6.70×10^{-10}	1.50×10^{-8}
FANCE	6.665	19.89	13.82	3.16×10^{-11}	1.63×10^{-12}	6.47×10^{-5}
FANCF	3.547	5.822	5.203	5.68×10^{-5}	4.81×10^{-4}	ns
FANCG	8.145	34.718	15.116	1.31×10^{-11}	4.95×10^{-11}	1.77×10^{-6}
FANCI	6.396	31.538	16.483	4.44×10^{-16}	2.36×10^{-14}	5.36×10^{-7}
FANCL	5.398	28.678	8.593	3.64×10^{-10}	1.09×10^{-8}	3.07×10^{-8}
FANCM	1.095	2.614	1.783	1.49×10^{-9}	2.03×10^{-11}	8.62×10^{-4}
FAT1	27.653	56.791	98.567	1.03×10^{-4}	1.33×10^{-15}	2.46×10^{-3}
FBXW7	10.721	9.541	7.96	ns	3.38×10^{-3}	3.44×10^{-2}
GNAS	618.619	886.499	783.637	ns	ns	3.43×10^{-2}
GSTM1	0.389	0.541	0.936	ns	ns	ns
GSTP1	2096.081	2395.898	2797.054	8.04×10^{-4}	7.64×10^{-9}	ns
HGF	0.385	0.298	0.479	ns	ns	ns
HPGDS	0.855	0.902	0.77	ns	ns	ns
IL1A	2.429	5.588	9.822	7.62×10^{-3}	5.96×10^{-8}	1.62×10^{-5}
IL6	5.879	3.309	8.032	3.02×10^{-3}	ns	1.99×10^{-2}
MAP2K1	33.213	56.734	48.766	1.86×10^{-7}	1.77×10^{-10}	ns
MAP2K2	90.344	128.898	97.169	8.05×10^{-7}	6.84×10^{-4}	3.90×10^{-3}
MAPK9	9.706	14.03	10.741	2.16×10^{-7}	4.98×10^{-3}	2.80×10^{-3}
MET	10.67	13.65	36.453	4.50×10^{-2}	2.67×10^{-12}	3.80×10^{-5}
MLH1	17.74	30.43	18.063	3.95×10^{-8}	ns	4.48×10^{-8}
MMP9	2.637	55.726	60.887	7.82×10^{-5}	1.63×10^{-5}	ns
MTOR	14.979	24.682	19.07	3.13×10^{-8}	4.59×10^{-7}	ns
PDCD1	0.853	5.111	1.043	2.39×10^{-7}	1.36×10^{-4}	8.25×10^{-5}
PIK3CA	6.355	13.074	12.534	1.09×10^{-6}	4.14×10^{-11}	ns
PIK3CB	21.792	41.88	24.726	4.51×10^{-8}	6.21×10^{-4}	9.65×10^{-5}
PIK3CD	3.307	8.788	9.989	3.11×10^{-8}	$<1 \times 10^{-12}$	ns
PIK3CG	0.35	0.889	0.575	3.24×10^{-3}	1.37×10^{-4}	ns
PRAME	0.025	0.575	1.488	6.48×10^{-3}	3.23×10^{-6}	ns
PTGS2	2.984	3.526	8.104	3.01×10^{-2}	5.90×10^{-3}	ns
STAT3	80.694	103.858	97.685	4.16×10^{-3}	3.71×10^{-2}	ns
TGFA	14.312	14.398	31.581	ns	2.01×10^{-11}	4.23×10^{-5}
TGFB1	22.536	52.722	91.1	4.07×10^{-9}	1.62×10^{-12}	5.57×10^{-5}
TNF	1.359	2.779	2.633	5.88×10^{-3}	3.96×10^{-3}	ns
TP53	39.067	97.015	34.066	6.43×10^{-11}	ns	1.75×10^{-10}
TYMS	16.249	99.942	34.833	1.68×10^{-12}	3.93×10^{-7}	4.88×10^{8}
VIM	205.48	373.398	490.789	ns	4.21×10^{-4}	ns
XRCC1	14.684	48.945	22.322	4.54×10^{-12}	3.92×10^{-10}	3.39×10^{-9}

Abbreviations: N/A, no results available; ns, not statistically significant.

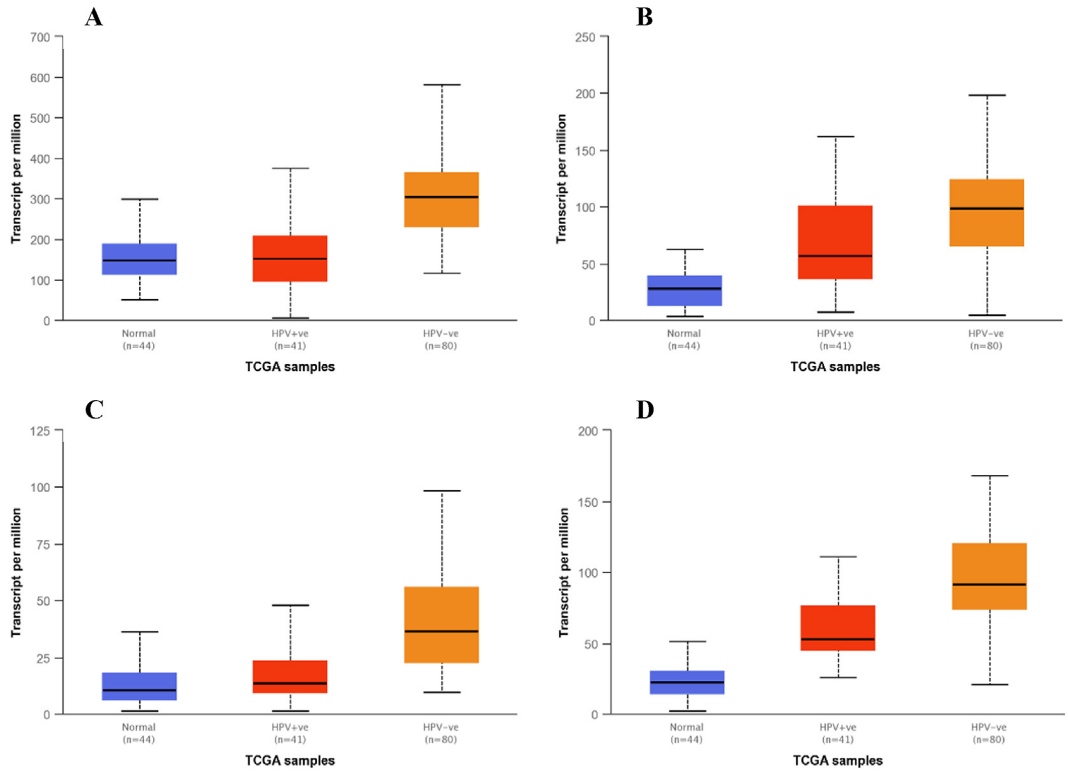


Fig. 4. Overexpressed genes in HPV-non-driven HNSCC. Box plot of expression of CD44 (A), FAT1 (B), MET (C) and TGFB1 (D) in HPV-driven HNSCC (red), HPV-non-driven HNSCC (orange) and healthy individuals (blue). The significance difference between groups was estimated by Student's *t*-test with *P*-value as shown in Table 2. *P*-value <0.05 was considered to be statistically significant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

selected was "TCGA - HNSC" (521 samples). Genes whose Kaplan-Meyer showed a statistically significant impact on OS were studied in terms of variation of their expression according to HPV status and TP53 mutational status using UALCAN. The genes most strongly modulated by HPV and TP53 mutational status were used to construct several combinations of genes for prognosis of HNSCC. To select the best prognostic gene signature with ability to mirror HPV status and TP53 mutational status, "Multivariate analysis" function was used to analyze OS and "Best cutoff" option was selected to slit patients. Gene signatures with a hazard ratio (HR) > 3 and *P*-value <0.05 were further evaluated in terms of disease-free survival (DFS), disease-specific survival (DSS), progression-free survival (PFS) and relapse-free survival (RFS). Among the gene signatures with statistically significant impact in all survival types, the choice fell on the gene signatures with higher capacity to reflect the HPV and TP53 mutational status in HNSCC patients.

Functional analysis of the chosen gene signature. Functional enrichment analysis was performed with STRING (string-db.org) and g:Profiler (https://biit.cs.ut.ee/gprofiler) for identification of the molecular mechanisms underlying the chosen gene signature [16,17]. Enrich items with *P*-value <0.05 were considered significant.

3. Results

3.1. Bibliometric analysis and molecular interaction network

In this study, we performed a multistep bioinformatic analysis to screen key genes of HNSCC. The flowchart is displayed in Fig. 1. Firstly, we performed an analysis in VosViewer to extract all the genes and their encoded proteins. Then, DisGeNET was used to verify which genes selected from VosViewer are associated with HNSCC. The detailed analysis of the genes extracted from VosViewer network is shown in Supplementary Table S1. The listed genes in C1168401, C3887461, C0278996, C0018671, C4329280 and C4528408 datasets downloaded from DisGeNET can be found in Supplementary Table S2.

The DisGeNET platform contained 2328 gene-disease associations. After applying the GDA score filter, 162 disease associations genes were obtained. The results from VosViewer and DisGeNET were analyzed, and the information was intercepted using a Venn Diagram (Fig. 2), identifying 104 common genes.

3.2. Construction and validation of a prognostic risk model

The 104 genes were studied using UALCAN and ToPP, the results of which are shown in Table 1. From the initial 104 genes, 72 genes were shown to have a statistically significant impact on OS of HNSCC patients. Subsequently, we studied how gene expression patterns are influenced by HPV status and TP53 mutational status. The most overexpressed genes in HPV-driven HNSCC were CCNB1, CDKN2A, MAP2K2, PIK3CB, TYMS and XRCC1, as shown in Fig. 3 and Table 2. In the case of HPV-non-driven HNSCC, the genes with a significant increase in expression were CD44, FAT1, MET and TGFB1 (Fig. 4 and Table 2). TP53 mutational status was also shown to influence the expression of certain genes. In TP53-mutant HNSCC, AKT1, ANO1, CD44, CTTN, MET, MMP9, TGFA and TGFB1 were the most upregulated genes as described in Fig. 5 and Table 3. In TP53-nonmutant HNSCC, CDKN2A and TYMS were the genes with significantly increased expression compared to TP53-mutant HNSCC and healthy patients (Fig. 6 and Table 3).

For the identification of the best gene combination, CDKN2A and TGFB1 were selected as fixed elements for their high discriminative power for HPV^{+}/HPV^{-} HNSCC and TP53-mutant/TP53-nonmutant HSNCC. The remaining genes that were shown to be significantly modulated by HPV status and TP53 mutational status were used to

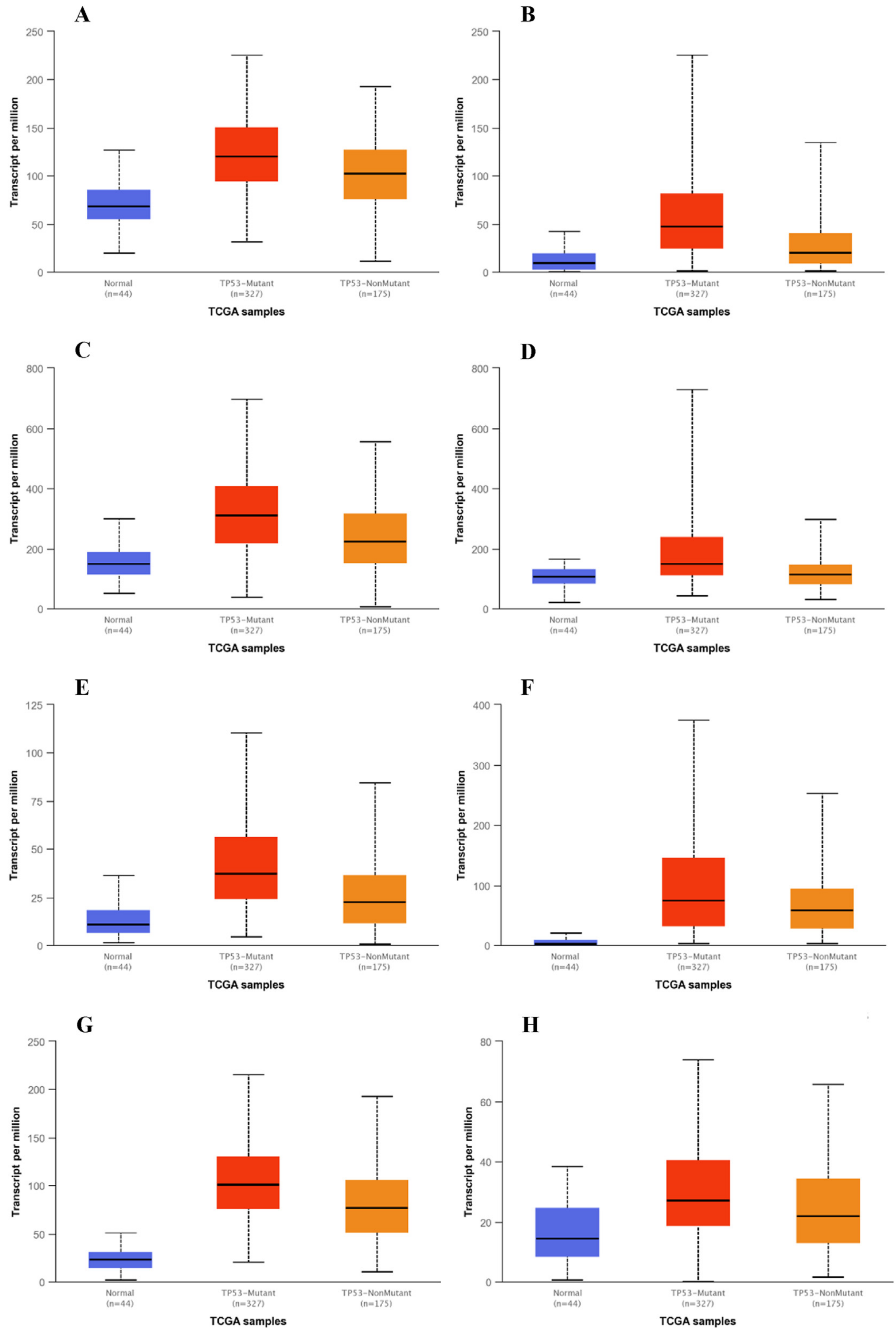


Fig. 5. Overexpressed genes in TP53-mutant HNSCC. Box plot of expression of AKT1 (A), ANO1 (B), CD44 (C), CTTN (D), MET (E), MMP9 (F), TGFB1(G) and TGFA (H) in TP53-mutant HNSCC (red), TP53-nonmutant HNSCC (orange) and healthy individuals (blue). The significance difference between groups was estimated by Student's *t*-test with *P*-value as shown in Table 3. *P*-value <0.05 was considered to be statistically significant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 3
Significant changes in the transcriptional expression of the selected genes in TP53-mutant HNSCC, TP53-nonmutant HNSCC and normal tissues (UALCAN).

Gene symbol	Transcript per million (median)			*P*-value from transcriptional analysis		
	Normal tissue	TP53 M HNSCC tissue	TP53 NonM HNSCC tissue	Normal vs TP53 M HNSCC	Normal vs TP53 NonM HNSCC	TP53 M vs TP53 NonM HNSCC
ABCB1	0.669	0.333	0.52	1.44 × 10-2	ns	8.28 × 10-3
ABCG2	0.746	0.741	0.619	ns	ns	ns
AKT1	68.216	120.014	102.236	1.62 × 10-12	6.04 × 10-13	1.02 × 10-6
ALDH2	124.012	66.607	78.721	9.42 × 10-9	5.02 × 10-3	2.48 × 10-3
ANO1	9.638	47.516	20.429	1.62 × 10-12	1.90 × 10-8	3.16 × 10-4
AREG	16.685	48.109	25.226	1.43 × 10-4	ns	4.12 × 10-3
ATM	3.355	4.86	4.639	1.62 × 10-6	1.70 × 10-5	ns
ATP7B	0.628	0.703	0.707	1.49 × 10-2	2.04 × 10-2	ns
B2M	2105.987	5129.913	5362.382	1.62 × 10-12	1.62 × 10-12	ns
BAP1	40.266	49.134	52.512	1.47 × 10-9	7.90 × 10-11	ns
BCL2	2.262	1.266	1.775	4.69 × 10-2	2.55 × 10-4	1.35 × 10-6
CASP8	9.374	16.423	17.078	<1 × 10-12	1.62 × 10-12	ns
CCNA1	0.531	2.908	0.959	1.76 × 10-12	4.02 × 10-4	2.99 × 10-4
CCNA2	9.288	23.41	23.916	1.62 × 10-12	<1 × 10-12	1.87 × 10-2
CCNB1	31.362	71.269	81.182	<1 × 10-12	1.62 × 10-12	ns
CCND1	68.631	71.674	46.055	2.46 × 10-10	ns	2.60 × 10-6
CD44	148.668	309.755	222.861	1.62 × 10-12	4.77 × 10-9	1.96 × 10-4
CD274	1.706	3.26	4.378	1.02 × 10-10	3.69 × 10-6	ns
CDKN2A	3.445	15.527	47.501	1.62 × 10-12	1.62 × 10-12	5.55 × 10-8
CSF3	2.647	2.012	2.477	ns	ns	ns
CTLA4	0.41	2.197	3.268	1.62 × 10-12	<1 × 10-12	2.41 × 10-4
CTNNB1	124.894	147.592	148.606	3.63 × 10-2	1.90 × 10-2	ns
CTTN	106.825	149.915	114.507	1.62 × 10-12	1.80 × 10-7	1.82 × 10-3
CYLD	13.202	16.234	16.078	4.23 × 10-6	7.73 × 10-5	ns
CYP1A1	0.034	0.031	0.021	ns	ns	ns
DPYD	9.727	8.675	10.416	ns	ns	4.05 × 10-2
EGFR	20.398	37.16	30.416	1.72 × 10-10	1.20 × 10-4	3.66 × 10-3
EP300	14.892	19.405	19.805	3.92 × 10-7	4.43 × 10-5	ns
ERBB2	85.599	47.385	54.844	3.56 × 10-3	ns	ns
ERBB3	42.897	26.313	30.592	2.34 × 10-5	2.76 × 10-3	4.22 × 10-4
ERCC1	46.256	64.429	66.296	2.21 × 10-4	1.68 × 10-5	ns
ERCC2	6.888	11.798	10.838	<1 × 10-12	4.28 × 10-12	9.56 × 10-3
FANCB	0.325	1.112	0.986	1.62 × 10-12	1.62 × 10-12	ns
FANCC	2.196	4.601	5.369	1.62 × 10-12	1.62 × 10-12	3.71 × 10-6
FANCD2	2.927	5.63	6.765	1.62 × 10-12	1.62 × 10-12	9.90 × 10-7
FANCE	6.675	14.422	15.28	<1 × 10-12	1.11 × 10-16	ns
FANCF	3.569	5.8	5.406	9.01 × 10-8	1.76 × 10-5	ns
FANCG	8.266	17.093	22.275	1.62 × 10-12	<1 × 10-12	1.88 × 10-3
FANCI	6.472	18.958	20.839	1.62 × 10-12	1.62 × 10-12	4.94 × 10-3
FANCL	5.483	9.548	9.877	1.62 × 10-12	1.62 × 10-12	4.48 × 10-6
FANCM	1.096	1.968	1.744	1.62 × 10-12	<1 × 10-12	ns
FAT1	27.791	89.847	87.613	<1 × 10-12	1.62 × 10-12	ns
FBXW7	10.497	8.18	9.374	1.80 × 10-2	ns	ns
GNAS	600.245	789.105	800.503	ns	ns	ns
GSTM1	0.389	0.356	0.56	3.53 × 10-5	2.64 × 10-2	ns
GSTP1	2102.774	2905.781	2729.842	<1 × 10-12	6.99 × 10-11	4.32 × 10-3
HGF	0.38	0.375	0.373	ns	ns	ns
HPGDS	0.84	0.675	0.753	ns	ns	ns
IL1A	2.412	9.922	7.101	<1 × 10-12	1.98 × 10-10	ns
IL6	5.64	5.838	3.595	ns	3.50 × 10-2	ns
MAP2K1	33.925	51.812	49.476	1.62 × 10-12	1.81 × 10-12	ns
MAP2K2	90.395	103.591	106.242	3.83 × 10-8	1.01 × 10-9	ns
MAPK9	9.68	11.589	12.259	5.76 × 10-5	3.74 × 10-6	ns
MET	10.903	37.163	22.392	1.62 × 10-12	6.39 × 10-8	4.13 × 10-5
MLH1	17.681	17.819	21.406	ns	6.61 × 10-8	4.98 × 10-10
MMP9	2.874	74.272	58.706	1.62 × 10-12	1.66 × 10-12	2.15 × 10-2
MTOR	15.041	20.281	20.579	7.29 × 10-13	6.71 × 10-11	ns
PDCD1	0.853	0.985	2.235	3.59 × 10-6	1.66 × 10-12	2.97 × 10-8
PIK3CA	6.414	12.292	10.736	<1 × 10-12	1.64 × 10-12	1.90 × 10-2
PIK3CB	22.197	26.777	28.124	1.63 × 10-7	4.69 × 10-8	ns
PIK3CD	3.363	9.209	9.282	<1 × 10-12	1.67 × 10-12	ns
PIK3CG	0.368	0.489	0.77	3.50 × 10-3	3.50 × 10-7	1.02 × 10-3
PRAME	0.023	1.718	0.196	<1 × 10-12	2.03 × 10-8	ns
PTGS2	2.581	6.305	4.677	1.70 × 10-10	1.79 × 10-6	ns
STAT3	80.942	91.132	94.38	ns	4.58 × 10-3	1.86 × 10-2
TGFA	14.318	26.971	21.807	6.04 × 10-14	7.47 × 10-7	1.49 × 10-2
TGFB1	23.029	100.812	76.704	<1 × 10-12	1.62 × 10-12	5.49 × 10-7
TNF	1.337	2.363	2.199	2.11 × 10-7	1.20 × 10-4	ns
TP53	41.119	38.493	54.242	1.21 × 10-2	4.04 × 10-9	1.53 × 10-5
TYMS	16.618	37.063	49.839	1.62 × 10-12	<1 × 10-12	2.16 × 10-5
VIM	205.48	454.124	457.081	3.90 × 10-5	4.32 × 10-3	ns
XRCC1	14.803	25.481	28.007	<1 × 10-12	<1 × 10-12	1.59 × 10-7

Abbreviations: N/A, no results available; ns, not statistically significant; TP53 NonM HNSCC tissue, TP53-nonmutant HNSCC tissue; TP53 M HNSCC tissue, TP53-mutant HNSCC tissue.

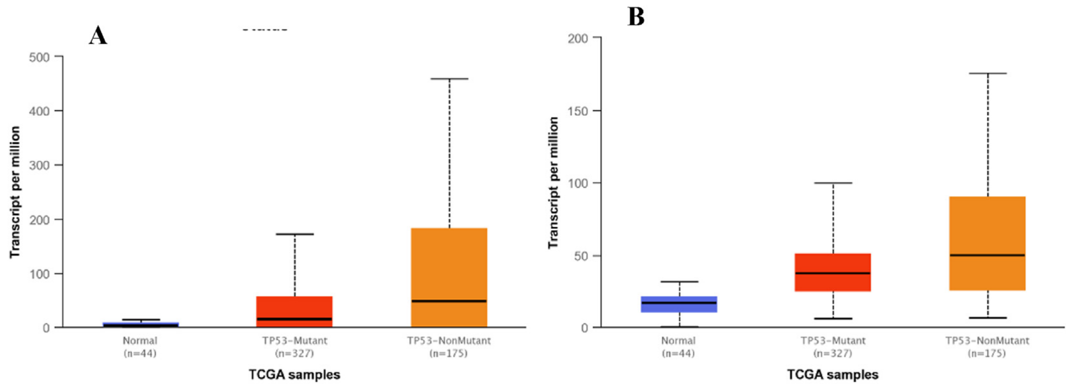


Fig. 6. Overexpressed genes in TP53-nonmutant HNSCC. Box plot of expression of CDKN2A (A) and TYMS (B) in TP53-mutant HNSCC (red), TP53-nonmutant HNSCC (orange) and healthy individuals (blue). The significance difference between groups was estimated by Student's *t*-test with *P*-value as shown in Table 3. *P*-value <0.05 was considered to be statistically significant. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

construct several gene combinations. The performance of these prognostic signatures was assessed using multivariate analysis in ToPP platform and the results are shown in Supplementary Table S3. Of all possible combinations, seven gene signatures showed a HR > 3 and a *P*-value <0.05 in the overall survival analysis, as shown in Fig. 7.

These signatures were further characterized in terms of disease-free survival (DFS), disease-specific survival (DSS), progression-free survival (PFS) and relapse-free survival (RFS) in Supplementary Figures S1-S4, whose results are summarized in Table 4. It was possible to observe that only two gene signatures showed a statistically significant correlation with all types of survival analysis. These two combinations were CDKN2A + TGFB1 + CD44 + MMP9 and CDKN2A + TGFB1 + MAP2K2 + TYMS. Of these two gene combinations, CDKN2A + TGFB1 + CD44 + MMP9 was chosen because it had three genes overexpressed in TP53-mutant HNSCC (TGFB1, CD44 and MMP9), one gene overexpressed in TP53-nonmutant HNSCC (CDKN2A), one gene overexpressed in HPV-driven HNSCC (CDKN2A), and two genes overexpressed in HPV-non-driven HNSCC (TGFB1 and CD44). The risk score of the selected gene combination was (0.0706 x CD44) + (−0.0534 x CDKN2A) + (−0.0503 x MMP9) + (0.273 x TGFB1). After selecting the best gene combination, the prognostic accuracy of the four-gene signature risk score compared with other clinical factors was assessed, the results of which are shown in Fig. 8. It was possible to verify that the combination of our gene signature with other variables such as gender, race and histological grade showed a statistically significant impact on the OS of HNSCC patients. Regarding gender, female HNSCC patients presenting the chosen prognostic genes showed a higher risk (HR: 4.42 vs 2.69) compared to male HNSCC patients presenting the same gene signature. Regarding race, Asian patients expressing the genes of interest showed a much higher impact (HR: 4.92×10^9 vs 1.9) on survival compared to Caucasian patients. When assessing the impact on OS of histological grade in HNSCC patients presenting the chosen gene signature, a higher risk score was observed for Gx and G1 histological grades (HR(G1): 5.12, HR(G2): 1.66, HR(G3): 4.33 and HR(Gx): 5.36). From the gene signature, three of the four genes of interest correspond to differentially expressed genes in HNSCC, as shown in Table 5.

3.3. Functional enrichment

The PPI network obtained in STRING (Fig. 9A) allows verifying the close relationship among all genes that are part of the chosen signature. When the extended version of the PPI network (Fig. 9B) was obtained it was possible to observe that this set of genes is strongly associated to the TP53 pathway, which is in line with the results described above. g:Profiler analysis shown that the enriched items were mainly related to regulation of endopeptidase activity involved in apoptosis, regulation of protein modifications (phosphorylation, proteolysis), regulation of DNA damage response and collagen binding, as shown in Fig. 10.

4. Discussion

In the present study, a four-gene signature for HNSCC prognosis was identified. This signature contains genes whose expression changes significantly according to HPV status and TP53 mutational status, potentially discriminating HNSCC into HPV-driven HNSCC versus HPV-non-driven HNSCC and TP53-mutant HNSCC versus TP53-nonmutant HNSCC. In addition, this gene signature contains CDKN2A gene encoding p16, which is currently the only biomarker used in the clinic to establish prognosis. No study to date has identified a gene signature that allows this type of prognostic stratification considering HPV and TP53 mutational status. Therefore, in this work, an innovative methodology based on an automatic text mining feature of VosViewer was used. All proteins and genes of each keyword of the bibliometric network generated with this software were extracted using UNIPROT. In combination with DisGeNET, 104 genes with a well-established relationship with HNSCC (GDA score >0.1) were selected. Analyzing the expression profiles of the genes of interest (TCGA and CPTAC) and the impact on survival, a four-gene signature was identified that among all those studied is the one with the most potential in predicting prognosis in patients with HNSCC, as well as HPV and TP53 mutational status. The TP53 gene is the most frequently mutated gene in HPV-non-driven HNSCC. The associated TP53 mutations play a major role in the early stages of carcinogenesis and tumor progression. TP53 mutations are associated with a worse prognosis, poorer response to chemotherapy treatments, and higher tumor recurrence rates [18] [–] [21].

The genes that constitute the chosen prognostic signature are: CDKN2A, TGFB1, CD44 and MMP9. This signature has a risk group hazard ratio of 3.04 (IC 95%: 1.73–5.32), demonstrating an increased risk of death in patients who present this gene signature. The OS of the high-risk group is worse than the low-risk group ($P < 0.0001$) allowing a risk stratification of the HNSCC patients for wiser adjustment of the treatment schemes and follow-up orientations. The relationship between the expression of each of the genes and HPV status and TP53 mutational status was studied, and it was possible to observe that TGFB1, CD44 and MMP9 were overexpressed in TP53-mutant HNSCC, while CDKN2A was overexpressed in TP53-nonmutant HNSCC. In Fig. 9, we can observe that our signature genes are strongly linked to TP53, which consolidates the potential of this signature to reflect TP53 mutational status. In Fig. 10, signal transduction in DNA damage response by p53 is one of the main pathways associated with the chosen gene signature, which reinforces

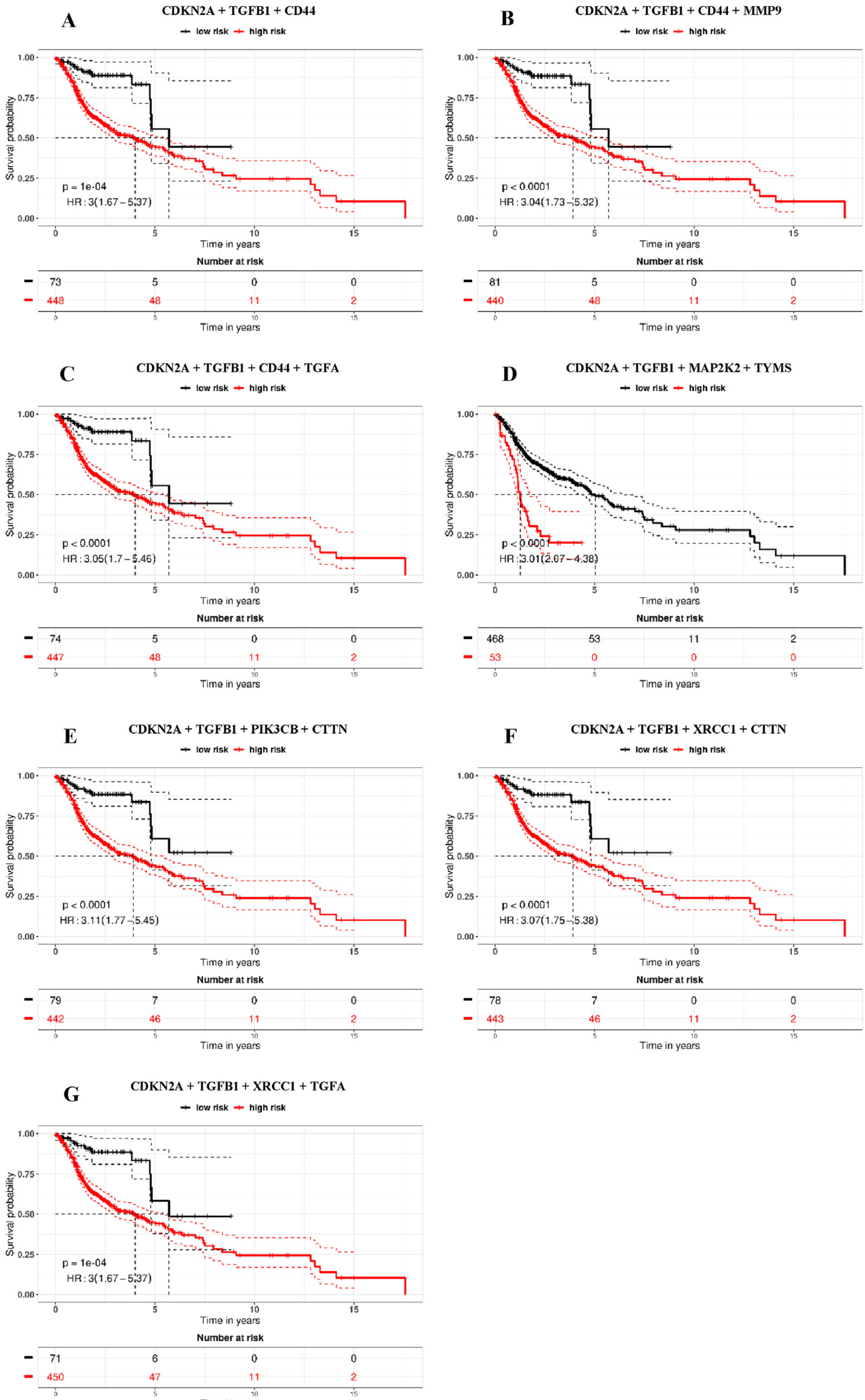


Fig. 7. Kaplan-Meier survival analysis in ToPP Database. OS analysis of the following gene combinations: CDKN2A + TGFB1 + CD44 (A), CDKN2A + TGFB1 + CD44 + MMP9 (B), CDKN2A + TGFB1 + CD44 + TGFA (C), CDKN2A + TGFB1 + MAP2K2 + TYMS (D), CDKN2A + TGFB1 + PIK3CB + CTTN (E), CDKN2A + TGFB1 + XRCC1 + CTTN (F) and CDKN2A + TGFB1 + XRCC1 + TGFA (G) using the HNSCC dataset. A red line indicates the survival curve of the patient group at higher risk of death. A black line indicates the survival curve of the patient group with lower risk of death. Tick marks indicate censored data points; *P*-values are determined by log-rank tests. The size of each patient group, the hazard ratio of the two groups of patients and the log-rank *P*-value are reported and summarized in Table 4. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Survival analysis of the top seven gene signatures in ToPP platform.

	OS	PFI	DSS	DFI	RFS
CDKN2A TGFB1 CD44	HR = 3 (1.67–5.37) P = 0.0001	HR = 1.8 (1.21–2.69) P = 0.0034	HR = 2.28 (1.5–3.47) P < 0.0001	HR = 1.67 (0.733–3.78) P = 0.22	HR = 1.51 (0.979–2.32) P = 0.061
CDKN2A TGFB1 CD44 MMP9	HR = 3.04 (1.73–5.32) P < 0.0001	HR = 1.96 (1.4–2.72) P < 0.0001	HR = 2.14 (1.5–3.03) P < 0.0001	HR = 2.17 (0.994–4.72) P = 0.046	HR = 1.96 (1.05–3.66) P = 0.032
CDKN2A TGFB1 CD44 TGFA	HR = 3.05 (1.7–5.46) P < 0.0001	HR = 1.94 (1.26–2.97) P = 0.002	HR = 1.86 (1.31–2.64) P = 0.00039	HR = 1.71 (0.752–3.89) P = 0.2	HR = 1.59 (0.957–2.64) P = 0.071
CDKN2A TGFB1 MAP2K2 TYMS	HR = 3.01 (2.07–4.38) P < 0.0001	HR = 1.5 (1.12–2.01) P = 0.0062	HR = 2.2 (1.55–3.11) P < 0.0001	HR = 2.46 (1.15–5.23) P = 0.016	HR = 1.91 (1.12–3.25) P = 0.016
CDKN2A TGFB1 PIK3CB CTTN	HR = 3.11 (1.77–5.45) P < 0.0001	HR = 1.84 (1.23–2.75) P = 0.0024	HR = 1.89 (1.33–2.68) P = 0.00034	HR = 1.95 (0.906–4.19) P = 0.082	HR = 2.21 (1.21–4.04) P = 0.0078
CDKN2A TGFB1 XRCC1 CTTN	HR = 3.07 (1.75–5.38) P < 0.0001	HR = 2.32 (1.32–4.06) P = 0.0026	HR = 1.96 (1.36–2.81) P = 0.00022	HR = 5.19 (0.0704–38.2) P = 0.07	HR = 2.02 (1.06–3.88) P = 0.03
CDKN2A TGFB1 XRCC1 TGFA	HR = 3 (1.67–5.37) P = 0.0001	HR = 1.87 (1.17–3.01) P = 0.0084	HR = 1.97 (1.34–2.88) P = 0.00039	HR = 2.05(0.872–4.84) P = 0.093	HR = 1.67 (1.01–2.77) P = 0.045

Abbreviations: DFI, disease-free survival; DSS, disease-specific survival; HR, hazard ratio; P, P-value; PFI, progression-free survival; OS, overall survival; RFS, relapse-free survival.

the results obtained in this work. Regarding HPV status, we observed that the expression of CDKN2A was increased in HPV-driven HNSCC, while the expression of TGFB1 and CD44 was increased in HPV-non-driven HNSCC. These genes were also characterised in terms of biological and molecular processes to understand the role of this signature in HNSCC (Fig. 10).

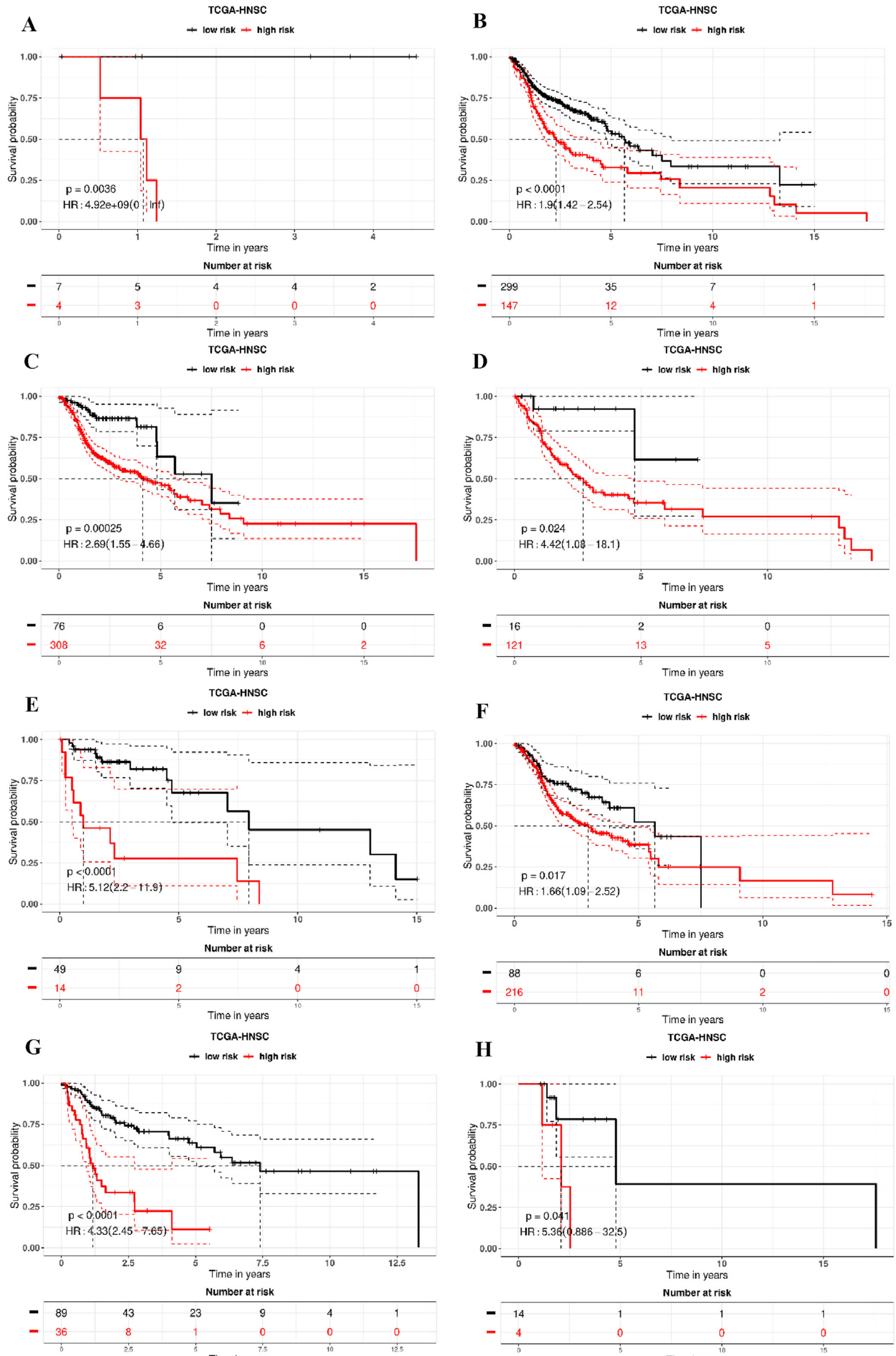
CDKN2A is a CDK inhibitor that interacts with both CDK4 and CDK6, preventing their binding to cyclins D and consequently inhibiting RB1 phosphorylation. It works as a tumor suppressor as it induces cell cycle arrest at G1 and G2/M checkpoints [22]. There is evidence in the literature supporting the role of CDKN2A in the prognosis of HNSCC. Some studies have shown that hypermethylation and copy number loss of CDKN2A gene are associated with worse OS in patients with HNSCC [23, 24]. The p16INK4A encoded by CDKN2A gene besides being a biomarker with high sensitivity for HPV status may reflect the genetic alterations of CDKN2A in patients with HNSCC. On the other hand, patients who are positive for p16INK4A and negative for p53 have a better prognosis than patients who are positive only for p16INK4A [25]. Studies demonstrating the detection and quantification of this protein in biological fluids as prognostic biomarkers of HNSCC are still scarce. However, some studies have detected hypermethylated p16INK4A in saliva and blood samples and demonstrated its correlation with HNSCC prognosis [26].

TGFB1 is a multifunctional peptide belonging to the transforming growth factor beta superfamily of cytokines. This polypeptide binds to TGFB receptors leading to the activation of the SMAD signalling pathway which regulates the transcription of hundreds of genes. The protein encoded by TGFB1 is involved in various cellular processes such as cell growth, cell differentiation, cell migration and apoptosis [27]. Elahi and Rakhshan have shown that high levels of TGFB1 are associated with a better prognosis in patients with oral squamous cell carcinoma [28]. The TGFB1 rs1800470 and TGFB1 rs1982073 polymorphisms were studied using peripheral blood samples from patients with HNSCC and shown to be associated with better DFS and OS [29–31]. The impact of TGFB1 on HNSCC is not yet established because of the dual role of this biomarker in suppressing abnormal cell proliferation in normal cells and promoting the ability to invade and metastasize in cancer [22].

CD44 is a surface glycoprotein that is overexpressed in several types of cancer. This cell-surface receptor helps cells to elaborate their response to changes in the tumour microenvironment, as it is associated with the regulation of cell-cell interactions, cell adhesion and migration [32]. A meta-analysis performed by Chen et al. showed that CD44 is associated to a worse prognosis for cancer of the larynx and pharynx. Regarding oral cancer the results were not conclusive [33]. A few studies have evaluated the role of CD44 obtained from saliva and blood samples of patients with HNSCC, with most linking elevated solCD44 levels to worse PFS and OS [34,35].

MMP9 is a Zn^{2+} dependent endopeptidase is secreted as a zymogen and activated by the plasminogen/plasmin system. This metalloproteinase is involved in the degradation of extracellular matrix proteins and leukocyte migration. MMP9 cleaves type IV and V collagens in shorter fragments and degrades fibronectin. CD44 binds to MMP2 and MMP9, which promote CD44 gene tail cleavage with CD44 intracytoplasmic domain (CD44ICD) release which is associated with cell migration and invasion [36]. A meta-analysis by Thangaraj et al. supports that high levels of MMP9 protein are associated with a worse prognosis of oral tongue squamous cell carcinoma [37]. This biomarker has been shown to have much potential in predicting OSCC recurrence in cases where surgical resections of the tumor are performed with histologically negative surgical margins [38]. Ruokolainen et al. demonstrated that serum MMP9 correlates with tissue MMP9 in patients with HNSCC. Patients with high MMP9 levels had the shortest cause-specific survival, RFS and OS [39]. One study evaluated salivary MMP9 levels before and after surgical treatment to patients with HNSCC. A statistically significant decrease in MMP9 levels was observed in patients after surgery, indicating that this biomarker has potential to be used in prognosis [40].

Albeit our study shows promising results, these need to be validated envisioning the translation of this set of genes to the clinical practice. Moreover, the use of several bioinformatic tools, each one based on a different methodology, may be a source of bias, supporting the need of data validation. Gene signatures could be articulated with clinical and histopathological data through the construction of a prognostic nomogram to obtain more reliable prognostic models. Furthermore, there is a great potential to enhance bioinformatics analysis with artificial intelligence envisioning the integration of clinical and histopathological data, multi-omics data and pharmacometrics. The ultimate goal is the creation of decision algorithms to tailor treatment choices to each patient's "omics" profile [41]. Omics encompasses multiple levels of molecular analysis, and the future will see machine learning approach to multi-omics disease data and decision-supporting tool.



(caption on next page)

Fig. 8. Kaplan–Meier survival curves for validation *in silico* of the four-gene model based on different clinical characteristics in ToPP. Overall survival for Asian HNSCC patients presenting the selected gene signature (A). Overall survival for Caucasian HNSCC patients presenting the selected gene signature (B). Overall survival for male HNSCC patients presenting the selected gene signature (C). Overall survival for female HNSCC patients presenting the selected gene signature (D). Overall survival for G1 HNSCC patients presenting the selected gene signature (E). Overall survival for G2 HNSCC patients presenting the selected gene signature (F). Overall survival for G3 HNSCC patients presenting the selected gene signature (G). Overall survival for GX HNSCC patients presenting the selected gene signature (H). A red line indicates the survival curve of the patient group at higher risk of death. A black line indicates the survival curve of the patient group with lower risk of death. Tick marks indicate censored data points; *P*-values are determined by log-rank tests. The size of each patient group, the hazard ratio of the two groups of patients and the log-rank *P*-value are reported. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Differentially Expressed Genes from the selected gene signature in HNSCC in GEPIA2.

	Median TPM (Tumor)	Median TPM (Normal)	Log2(FC)	adjP
CDKN2A	25.100	4.623	2.377	3.83×10^{-5}
TGFB1	67.982	17.279	1.916	2.38×10^{-51}
MMP9	56.261	2.529	4.020	7.10×10^{-36}

Abbreviations: adjP, adjusted *P*-value, Log2(FC), fold change; Median TPM, median transcript per million.

5. Conclusion

In this study, a gene-based signature composed by CDKN2A, CD44, MMP9 and TGFB1 genes was identified for prognosis and risk stratification of HNSCC using data from online free databases. In addition, it has the potential to mirror both HPV status and TP53 mutational status, proposing a novel strategy/gene panel to be used during the patient risk stratification process and allowing the development of integrative tools to advance precision medicine. If validated in large independent studies and studied their predictive power, these biomarkers may be useful as prognostic and predictive biomarkers in HNSCC.

Funding

This work was supported by the Portuguese Foundation for Science

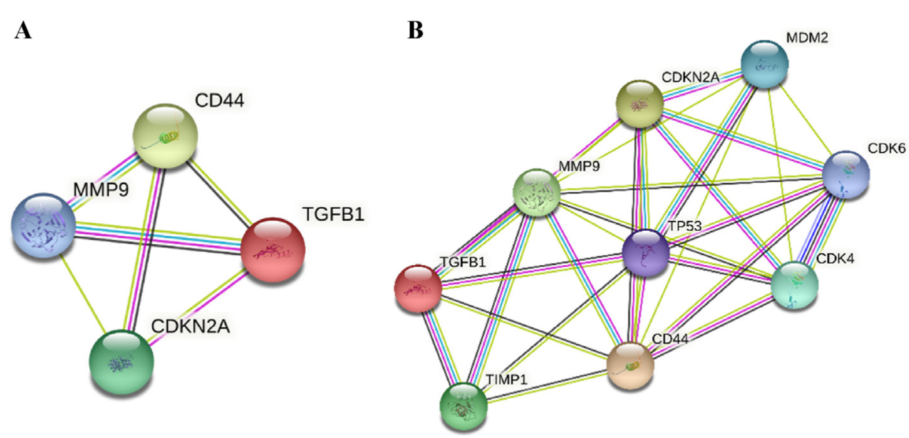


Fig. 9. PPI network in STRING. Protein-Protein interactions of the query proteins (A). Protein-Protein interaction of the query proteins with other major signal pathways (B).

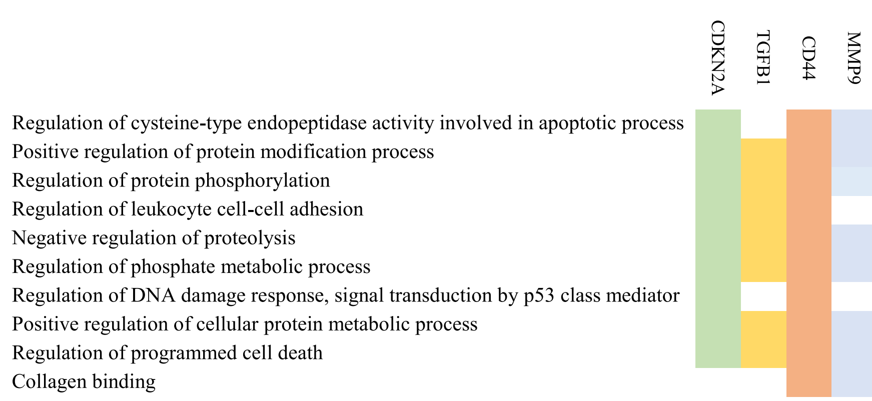


Fig. 10. Gene signature functional enrichment. Top up regulated pathways of the four-gene signature using g:Profiler web server (*P*-value <0.05).

and Technology (FCT), European Union, QREN, FEDER, and COMPETE for funding Institute of Biomedicine (iBiMED) (UIDB/04501/2020, POCI-01-0145-FEDER-007628), LAQV-REQUIMTE (UIDB/50006/2020), Instituto Português de Oncologia do Porto Francisco Gentil, EPE (CI-IPOP-134-2020), COST Action INTERCEPTOR and by an individual scholarship FCT-UA-ECIU from Oriana Barros.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. All authors have read the journal's authorship agreement and policy on disclosure of potential conflicts of interest.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.oor.2023.100018.

References

[1] Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Rev Dis Prim 2020;6(1).
[2] Bel'skaya LV, Sarf EA, Solomatin DV, Kosenok VK. Diagnostic and prognostic value of salivary biochemical markers in oral squamous cell carcinoma. Diagnostics 2020;10(10).
[3] Dhawan A, Scott J, Sundaresan P, Veness M, Porceddu S, Hau E, et al. Role of gene signatures combined with pathology in classification of oropharynx head and neck cancer. Sci Rep 2020 Dec 1;10(1).
[4] Qian Y, Daza J, Itzel T, Betge J, Zhan T, Marmé F, et al. Prognostic cancer gene expression signatures: current status and challenges. MD 2021;10(Cells):1–17.
[5] Michiels S, Ternès N, Rotolo F. Statistical controversies in clinical research: prognostic gene signatures are not (yet) useful in clinical practice. Ann Oncol 2016 Dec 1;27(12):2160–7.
[6] Ma W, Cao Q, She W. Identification and clinical validation of gene signatures with grade and survival in head and neck carcinomas. Braz J Med Biol Res 2021;54(11).
[7] Albers AE, Qian X, Kaufmann AM, Coordes A. Meta analysis: HPV and p16 pattern determines survival in patients with HNSCC and identifies potential new biologic subtype. Sci Rep 2017;7(1):1–14.
[8] Albers AE, Qian X, Kaufmann AM, Coordes A. Meta analysis: HPV and p16 pattern determines survival in patients with HNSCC and identifies potential new biologic subtype. Sci Rep 2017;7(1):1–14.
[9] Prigge ES, Arbyn M, von Knebel Doeberitz M, Reuschenbach M. Diagnostic accuracy of p16INK4a immunohistochemistry in oropharyngeal squamous cell carcinomas: a systematic review and meta-analysis. Int J Cancer 2017;140(5):1186–98.
[10] Zhou Ge, Liu Zhiyi, JNM. TP53 mutations in head and neck squamous cell carcinoma and their impact on disease progression and treatment response. Physiol Behav 2019;176(3):139–48.
[11] Shi C, Liu S, Tian X, Wang X, Gao P. A TP53 mutation model for the prediction of prognosis and therapeutic responses in head and neck squamous cell carcinoma. 2021. p. 1–13.
[12] Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 2019 Jul 1;47(W1):W556–60.
[13] Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia (United States) [Internet] 2017;19(8):649–58. https://doi.org/10.1016/j.neo.2017.05.002. Available from:.
[14] Ouyang J, Qin G, Liu Z, Jian X, Shi T, Xie L. ToPP: tumor online prognostic analysis platform for prognostic feature selection and clinical patient subgroup selection. iScience [Internet] 2022;25(5):104190. https://doi.org/10.1016/j.isci.2022.104190. Available from:.
[15] Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, et al. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 2017;19(8):649–58.
[16] von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B. STRING: a database of predicted functional associations between proteins. Nucleic Acids Res 2003;31(1):258–61.
[17] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019;47(W1):W191–8.
[18] Nathan CA, Khandelwal AR, Wolf GT, Rodrigo JP, Mäkitie AA, Saba NF, et al. TP53 mutations in head and neck cancer. Mol Carcinog 2022 Apr 1;61(4):385–91.
[19] Klinakis A, Rampias T. TP53 mutational landscape of metastatic head and neck cancer reveals patterns of mutation selection. EBioMedicine 2020 Aug 1:58.
[20] Masica DL, Li S, Douville C, Manola J, Ferris RL, Burtness B, et al. Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. Hum Genet 2015 May 1;134(5):497–507.
[21] Zhou G, Liu Z, Myers JN. TP53 mutations in head and neck squamous cell carcinoma and their impact on disease progression and treatment response. J Cell Biochem 2016 Dec 1:2682–92.
[22] Romagosa C, Simonetti S, López-Vicente L, Mazo A, Lleonart ME, Castellvi J, et al. P16Ink4a overexpression in cancer: a tumor suppressor gene associated with senescence and high-grade tumors. Oncogene 2011;30(18):2087–97.
[23] Chen WS, Bindra RS, Mo A, Hayman T, Husain Z, Contessa JN, et al. CDKN2A copy number loss is an independent prognostic factor in HPV-negative head and neck squamous cell carcinoma. Front Oncol 2018 Apr 4;8(APR).
[24] Zhou C, Shen Z, Ye D, Li Q, Deng H, Liu H, et al. The association and clinical significance of CDKN2A promoter methylation in head and neck squamous cell carcinoma: a meta-analysis. Cell Physiol Biochem 2018 Oct 1;50(3):868–82.
[25] Shinohara S, Kikuchi M, Tona R, Kanazawa Y, Kishimoto I, Harada H, et al. Prognostic impact of p16 and p53 expression in oropharyngeal squamous cell carcinomas. Jpn J Clin Oncol 2014 Mar 1;44(3):232–40.
[26] Chai RC, Lim Y, Frazer IH, Wan Y, Perry C, Jones L, et al. A pilot study to compare the detection of HPV-16 biomarkers in salivary oral rinses with tumour p16INK4a expression in head and neck squamous cell carcinoma patients. BMC Cancer 2016;16(1).
[27] [Internet] The Human Protein Atlas [cited 2022 Sep 4]. Available from: https://www.proteinatlas.org/ENSG00000105329-TGFB1; 2022.
[28] Lundberg M, Leivo I, Saarilahti K, Mäkitie AA, Mattila PS. Transforming growth factor beta 1 genotype and p16 as prognostic factors in head and neck squamous cell carcinoma. Acta Otolaryngol 2012 Sep;132(9):1006–12.
[29] Lundberg M, Leivo I, Saarilahti K, Mäkitie AA, Mattila PS. Transforming growth factor beta 1 genotype and p16 as prognostic factors in head and neck squamous cell carcinoma. 2012;(February. p. 1006–12.
[30] Lundberg M, Pajusto M, Koskinen WJ, Mä Kitie AA, Aaltonen LM, Mattila PS. ASSOCIATION BETWEEN TRANSFORMING GROWTH FACTOR b1 GENETIC POLYMORPHISM AND RESPONSE TO CHEMORADIOTHERAPY IN HEAD AND NECK SQUAMOUS CELL CANCER. 2009. Available from: www.interscience.wiley.com.
[31] Tao Y, Sturgis EM, Huang Z, Wang Y, Wei P, Wang JR, et al. TGFb1 genetic variants predict clinical outcomes of hpv-positive oropharyngeal cancer patients after definitive radiotherapy. Clin Cancer Res 2018 May 1;24(9):2225–33.
[32] [Internet] The Human Protein Atlas [cited 2022 Sep 4]. Available from: https://www.proteinatlas.org/ENSG00000026508-CD44; 2022.
[33] Chen J, Zhou J, Lu J, Xiong H, Shi X, Gong L. Significance of CD44 expression in head and neck cancer: a systemic review and meta-analysis [Internet]. 2014. Available from: http://getdata-graph-digitizer.com/.
[34] Sawant S, Ahire C, Dongre H, Joshi S, Jamghare S, Rane P, et al. Prognostic significance of elevated serum CD44 levels in patients with oral squamous cell carcinoma. J Oral Pathol Med 2018 Aug 1;47(7):665–73.
[35] Cohen ER, Reis IM, Gomez-Fernandez C, Smith D, Pereira L, Freiser ME, et al. CD44 and associated markers in oral rinses and tissues from oral and oropharyngeal cancer patients. Oral Oncol 2020 Jul 1:106.
[36] Chen J, Zhou J, Lu J, Xiong H, Shi X, Gong L. Significance of CD44 expression in head and neck cancer: a systemic review and meta-analysis. BMC Cancer 2014;14(1):1–9.
[37] Thangaraj SV, Shyamsundar V, Krishnamurthy A, Ramani P, Ganesan K, Muthuswami M, et al. Molecular portrait of oral tongue squamous cell carcinoma shown by integrative meta-analysis of expression profiles with validations. PLoS One 2016 Jun 1;11(6).
[38] Ogbureke KUE, Weinberger PM, Looney SW, Li L, Fisher LW. Expressions of matrix metalloproteinase-9 (MMP-9), dentin sialophosphoprotein (DSPP), and osteopontin (OPN) at histologically negative surgical margins may predict recurrence of oral squamous cell carcinoma [Internet]. 3. Oncotarget; 2012. Available from: www.impactjournals.com/oncotarget.
[39] Ruokolainen H, Pääkkö P, Turpeenniemi-Hujanen T. Serum matrix metalloproteinase-9 in head and neck squamous cell carcinoma is a prognostic marker. Int J Cancer 2005 Sep 1;116(3):422–7.
[40] Shin YJ, Vu H, Lee JH, Kim HD. Diagnostic and prognostic ability of salivary MMP-9 for oral squamous cell carcinoma: a pre-/post-surgery case and matched control study. PLoS One 2021 Mar 1;(3 March):16.
[41] Zeng H, Chen L, Huang Y, Luo Y, Ma X. Integrative models of histopathological image features and omics data predict survival in head and neck squamous cell carcinoma. Front Cell Dev Biol 2020;8(October:1–12.