Thesis progression

Thesis title: Novel Gene Signatures and Therapeutic Target Identification in Human Papillomavirus-Associated Head and Neck Cancer

Thesis progression title: Transcriptomic Profiling and Network Analysis Reveal Key Regulatory Genes Involved in the Progression of Head and Neck Cancer

Student: Siwakorn Boonpok **Student ID:** 675070028-3

Advisor: Dr. Chukkris Heawchaiyaphum

Date: 11th February 2026

1. Introduction

Head and neck cancer (HNC) is a major global health issue, with over 930,000 new cases and approximately 460,000 deaths reported annually worldwide [1]. HNC is the seventh most common cancer and accounts for about 3% of all cancer deaths. HNC primarily originates from the epithelial cells of the oral cavity, pharynx, and larynx, and is strongly associated with environmental and lifestyle risk factors, including tobacco smoking and alcohol consumption [2]. In addition, oncogenic virus infections, particularly Epstein-Barr virus (EBV) and human papillomavirus (HPV), are increasingly recognized as critical contributors to HNC pathogenesis and tumorigenesis, especially in oropharyngeal squamous cell carcinoma (OPSCC) and nasopharyngeal carcinoma (NPC) [3].

Currently, transcriptomics is one of emerging technologies offering powerful tools for revealing the intricate relationships between viruses and host cells. Transcriptomic approaches in HNC research enable comprehensive profiling of gene expression patterns, providing insights into the molecular mechanisms underlying tumorigenesis in both viral and non-viral infections. In HNC cases with viral infection, particularly HPV, transcriptomic analysis helps identify viral-host interactions, elucidate viral gene expression, and elucidate the mechanisms by which HPV infection modulates host cell transcriptional activity, thereby facilitating tumorigenesis [4]. In non-viral HNC, transcriptomics reveals alterations in tumor suppressor genes, signaling pathways, and cell cycle regulation, which may serve as potential therapeutic targets [5].

Recent advancements in omics technologies, such as transcriptomics, have provided new insights into the complex molecular interactions and host factors that drive carcinogenesis. In head and neck cancer (HNC), these technologies have enabled the identification of key alterations in gene expression, signaling pathways, and immune responses that contribute to tumor initiation and progression. However, such approaches have not yet been fully leveraged to investigate the therapeutic potential of targeting host-specific molecular pathways involved in HNC. Furthermore, integrating these multi-omics findings with computational molecular docking and network-based approaches could significantly accelerate the discovery of novel therapeutic compounds and optimize precision treatment strategies.

Alongside these advances, drug repurposing strategies have gained attention as a cost-effective way to accelerate therapy development. Several repurposed drugs, including metformin, statins, and nonsteroidal anti-inflammatory drugs, have shown potential anticancer effects in HNC by modulating cell signaling, metabolism, and immune response [6,7]. However, while drug repurposing has shown anticancer potential in HNC, systematic efforts to identify and validate repurposed drugs that act on specific molecular pathways remain limited, leaving an important gap in the development of precision therapies for HNC.

Therefore, to better understand the molecular mechanisms underlying HNC carcinogenesis, the present study aims to address these gaps by leveraging cutting-edge technologies to elucidate the key molecular drivers of HNC, identify novel gene signatures associated with the disease, and repurpose existing drugs and their targets ultimately contributing to the development of more effective treatment strategies.

2. Objectives

2.2.1 To identify differentially expressed genes and dysregulated biological pathways associated with head and neck cancer using integrative transcriptomic and functional enrichment analyses.

2.2.2 To identify key regulatory hub genes through protein–protein interaction network analysis and to evaluate their prognostic significance using independent expression and survival datasets.

2.2.3 To identify candidate genes involved in head and neck cancer progression as potential therapeutic targets for drug repurposing.

3. Materials and methods

3.1 Data retrieval

To identify differentially expressed genes (DEGs) associated with head and neck cancer (HNC), five publicly available RNA-sequencing datasets (GSE130605, GSE137308, GSE165883, GSE174368, and GSE227919) were retrieved from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. Raw sequencing data were downloaded and subjected to quality control (QC) assessment prior to downstream analysis. QC analysis was performed using FastQC to evaluate key sequencing quality metrics, including per-base sequence quality, GC content, sequence duplication levels, and adapter contamination. High-quality reads were subsequently aligned to the human reference genome (GRCh38) using HISAT2, enabling accurate mapping of sequencing reads to known genes and transcripts. Following alignment, HTSeq was employed to quantify gene-level read counts, generating expression matrices for subsequent differential expression analysis.

3.2 Differential Gene Expression Profiling in Head and Neck Cancer

To identify differentially expressed genes (DEGs) between normal and head and neck cancer (HNC) samples, RNA-seq data were analyzed using the DESeq2 package in RStudio (version 4.4). DESeq2 models count-based expression data using a negative binomial distribution to estimate fold changes and assess statistical significance. P-values were adjusted for multiple testing using the Benjamini–Hochberg false discovery rate (FDR) method. Genes with an adjusted p-value (FDR) < 0.05 and a log2 fold change ≥ 2 or ≤ -2 were considered significantly differentially expressed.

3.3 Gene ontology (GO) and KEGG enrichment analysis

To elucidate the biological functions, pathways and key regulatory genes associated with differentially expressed genes (DEGs) in head and neck cancer (HNC), functional enrichment analyses were performed. Gene Ontology (GO) enrichment analysis was conducted using the Database for Annotation, Visualization, and Integrated Discovery (DAVID, version 6.8), which categorizes genes into three domains: biological process (BP), cellular component (CC), and molecular function (MF). The Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis will also be conducted to identify significantly affected signaling and metabolic pathways.

3.4 PPI network analysis of differentially expressed genes (DEGs)

To explore the interactions among differentially expressed genes (DEGs), a protein–protein interaction (PPI) network was constructed using the STRING database (version 11). DEGs were mapped to STRING, and interactions with a confidence score above the predefined threshold were retrieved. The resulting PPI network was visualized and analyzed using Cytoscape software (version 3.10.3). To identify hub genes within the protein–protein interaction (PPI) network, the CytoHubba plugin in Cytoscape was employed. CytoHubba applies multiple topological algorithms to evaluate the importance of nodes within a network. In this study, six algorithms were used, including Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Degree, Closeness, and Radiality, each reflecting distinct aspects of network centrality and biological relevance. The top 10% of genes from the entire network, based on their ranking scores, were selected as candidate hub genes. Genes consistently ranked within the top 10% in at least five out of six topological algorithms were defined as final hub genes.

3.5 Validation of gene expression and survival analysis using GEPIA2

To validate gene expression patterns and assess their prognostic significance in head and neck cancer (HNC), the Gene Expression Profiling Interactive Analysis 2 (GEPIA2) web server was utilized. GEPIA2 integrates RNA-sequencing data from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) projects. The hub genes from PPI network was analyzed using the GEPIA2 default pipeline, with a log2 fold change cutoff of 2 and an adjusted p-value < 0.05 as the significance threshold. Survival analysis was performed using the Kaplan–Meier method to evaluate the association between gene expression levels and overall survival in HNC patients. Patients were stratified into high- and low-expression groups based on the median expression level of each gene. Statistical significance was determined using the log-rank test, with $p < 0.05$ considered statistically significant.

3.6 Expression analysis of significant hub genes across disease progression

To further characterize the expression dynamics of prognostically significant hub genes identified from GEPIA2 analysis, normalized RNA-sequencing read counts were obtained from samples representing normal , hyperplasia, dysplasia, and head and neck cancer (HNC). Only genes that demonstrated statistically significant differential expression and survival relevance in the previous analyses were included in this step. Normalized expression values were used to evaluate expression trends across the stages of disease progression. Data were visualized using GraphPad Prism (version 10.6.1). Pairwise comparisons of gene expression levels between groups were performed using unpaired (independent) two-tailed Student’s *t*-tests. Comparisons were conducted between normal vs. hyperplasia, hyperplasia vs. dysplasia, dysplasia vs. cancer, and normal vs. cancer. A *p*-value < 0.05 was considered statistically significant.

Genes displaying a stepwise increase in expression from normal through hyperplasia and dysplasia to cancer, together with statistically significant differences in pairwise comparisons, were considered to exhibit a progressive up-regulation trend associated with disease progression.

4.Results

4.1 Identification of differentially expressed genes (DEGs)

From RNA-seq data, a total of 3,392 gene expression profiles were obtained after normalization, among which 2,796 genes were identified as upregulated and 596 genes as downregulated by differential expression analysis (Fig 1).

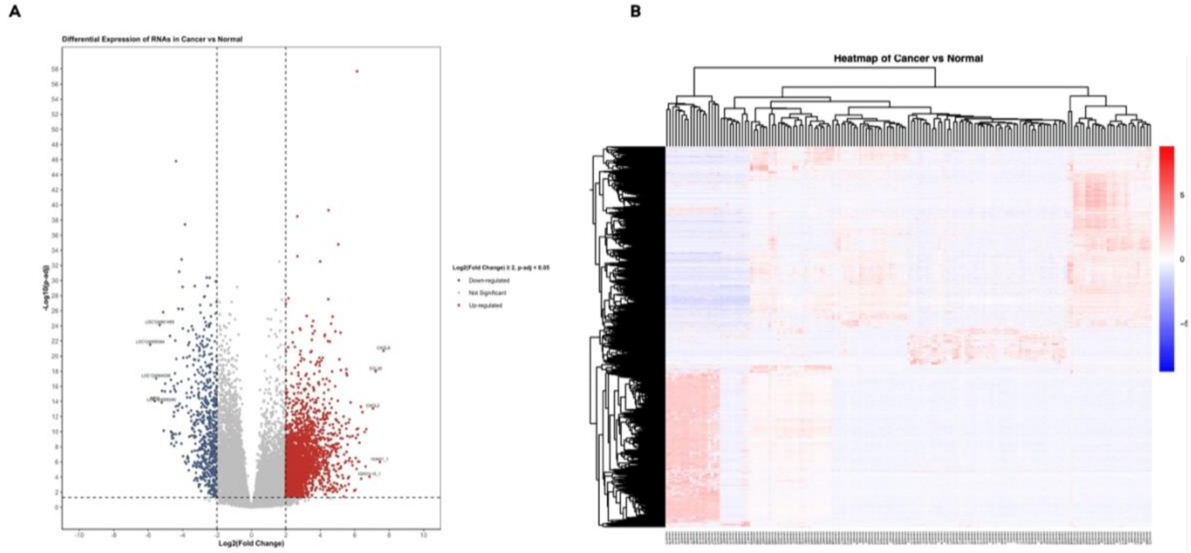

Fig 1. mRNAs expression profiles differences between HNC and normal samples (A) Volcano plot of DEGs between HNC and normal samples (B) Heatmap showing the expression patterns of DEGs between HNC and normal samples. Red indicates relatively high gene expression levels, whereas blue indicates relatively low gene expression levels across samples.

4.2 Gene ontology (GO) and KEGG enrichment analysis

Gene Ontology (GO) biological process enrichment analysis revealed distinct functional patterns between upregulated and downregulated differentially expressed genes. Upregulated genes were significantly enriched in processes related to regulation of apoptosis, cell division, positive regulation of RNA polymerase II–mediated transcription, immune response, angiogenesis, ubiquitin-dependent protein catabolism, mitochondrial translation, RNA splicing, and mRNA processing, suggesting activation of pathways involved in cell survival, proliferation, transcriptional and post-transcriptional regulation, and tumor-associated immune and metabolic activity (Fig 2A). In contrast, downregulated genes were mainly enriched in biological processes associated with intermediate filament and actin cytoskeleton organization, cellular stress responses, fibroblast proliferation, intracellular glutamate homeostasis, positive regulation of lymphocyte proliferation, Notch signaling, calcium ion transport, and regulation of gene expression. These findings indicate reduced activity in cytoskeletal organization, calcium-dependent signaling, stromal and immune regulatory processes, and cellular homeostasis (Fig 2B).

KEGG pathway enrichment analysis demonstrated distinct pathway alterations between upregulated and downregulated differentially expressed genes. Upregulated genes were significantly enriched in NF-**K**B and TNF signaling, cell cycle regulation, p53 signaling, DNA replication, proteasome activity, oxidative phosphorylation, ferroptosis, apoptosis, cellular senescence, and transcriptional misregulation in cancer, indicating enhanced inflammatory signaling, proliferative control, metabolic reprogramming, and stress-response pathways (Fig 2C). In contrast, downregulated genes were mainly enriched in glutamatergic synapse, Ras signaling, calcium signaling, and synaptic vesicle cycle pathways, suggesting reduced activity in calcium-dependent signal transduction and cellular communication processes (Fig 2D).

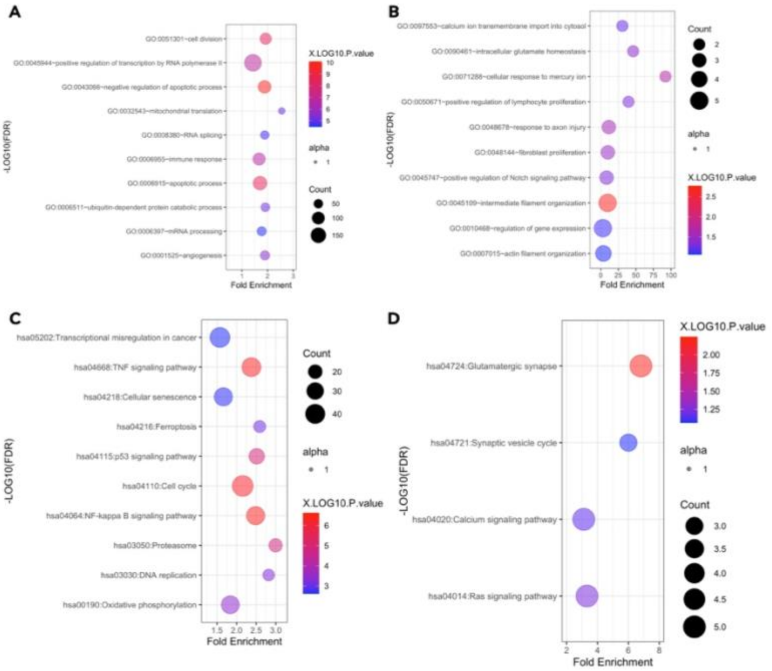

Fig 2. Gene Ontology (GO) biological process and KEGG pathway enrichment analysis of differentially expressed genes. (A) GO biological process enrichment analysis of upregulated genes. (B) GO biological process enrichment analysis of downregulated genes. (C) KEGG pathway enrichment analysis of upregulated genes. (D) KEGG pathway enrichment analysis of downregulated genes.

4.3 PPI network analysis of differentially expressed genes (DEGs)

To investigate the molecular interaction landscape of upregulated differentially expressed genes, a protein–protein interaction (PPI) network was constructed, consisting of 2,227 nodes and 31,638 edges. Network topology was quantitatively assessed using multiple complementary centrality algorithms, including Maximal Clique Centrality (MCC), Maximum Neighborhood Component (MNC), Density of Maximum Neighborhood Component (DMNC), Degree, Closeness, and Radiality (Fig 3A-3F)., to identify highly connected and biologically influential nodes within the network. Genes ranked within the top 10%

according to each topological metric were defined as candidate hub genes. A total of 125 genes met this criterion, and genes consistently ranked by at least five of the six algorithms were subsequently defined as final hub genes (Fig 3G-3H), reflecting robust centrality and potential functional importance within the PPI network.

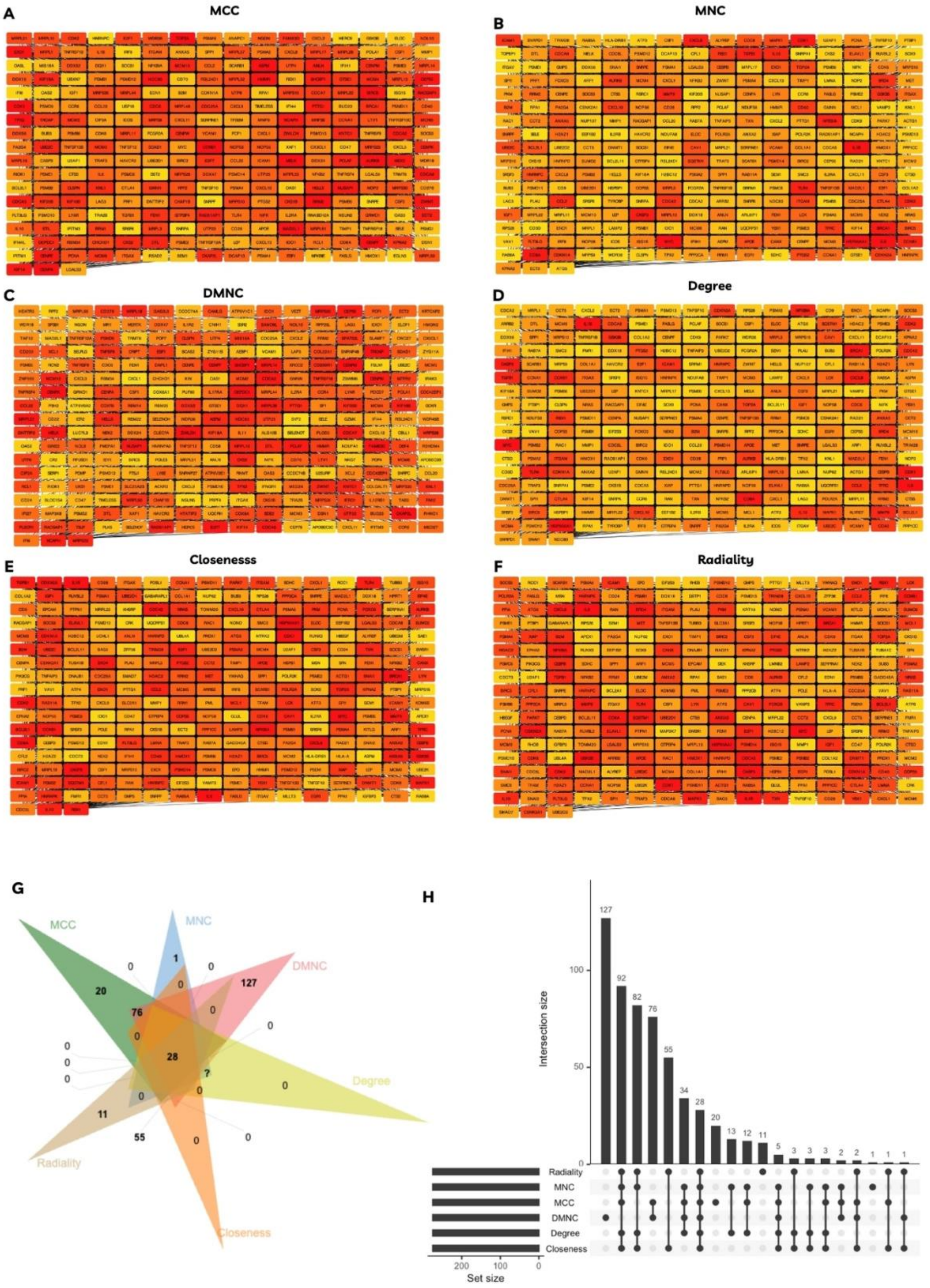

Fig 3. Identification of hub genes using protein–protein interaction (PPI) network analysis. (A–F) Hub gene ranking based on six topological algorithms: (A) Maximal Clique Centrality (MCC), (B) Maximum Neighborhood Component (MNC), (C) Density of Maximum Neighborhood Component (DMNC), (D) Degree, (E) Closeness, and(F) Radiality. (G) Venn diagram

illustrating the overlap of hub genes identified by the six topological algorithms. (H) UpSet plot showing the intersection patterns and consistency of hub gene selection across the six topological measures.

4.4 Validation of gene expression and survival analysis using GEPIA2

To validate the expression profiles and prognostic significance of hub genes identified from the PPI network analysis, a total of 125 hub genes were subjected to differential expression and survival analyses using the GEPIA2 database. Differential expression analysis between head and neck cancer (HNC) and normal tissues was conducted using a threshold of $|\log_2$ fold change$| \geq 2$ and a p-value < 0.05. Based on these criteria, 60 hub genes were identified as significantly differentially expressed. Subsequent Kaplan–Meier survival analysis of the significantly expressed hub genes demonstrated that nine genes (CTLA4, SCARB1, SOCS1, SPP1, CCNA1, MMP1, IDO1, ANXA5, and SERPINE1) were significantly associated with overall survival in patients with HNC (Fig 5 & Fig 6).

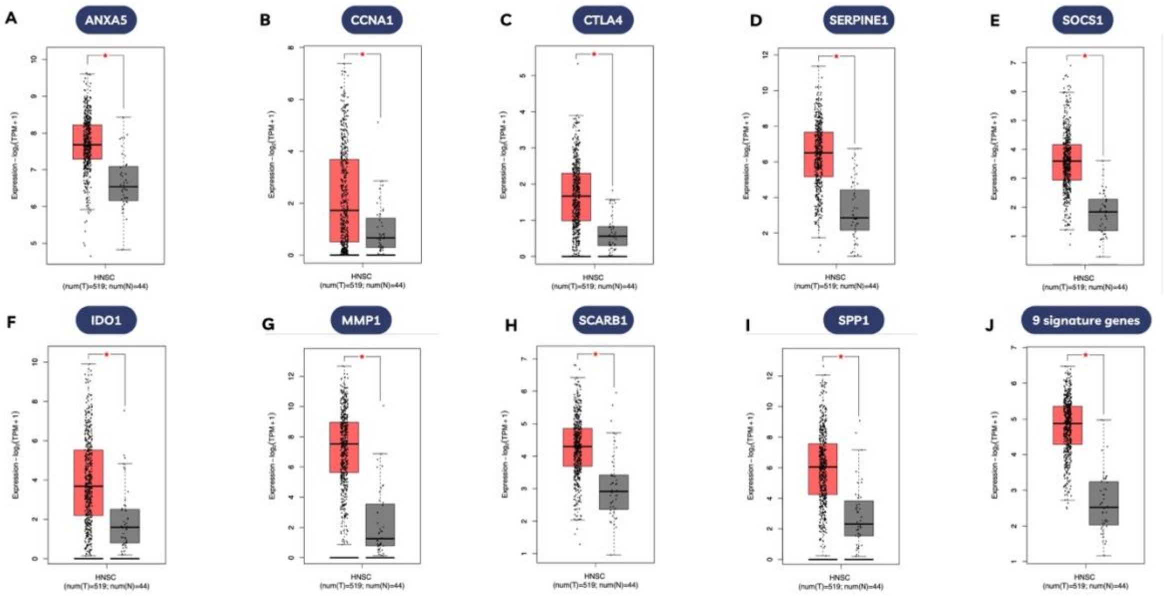

Fig 5. Validation of hub gene expression between head and neck cancer (HNC) and normal tissues using GEPIA2. (A) ANXA5, (B) CCNA1, (C) CTLA4, (D) SERPINE1, (E) SOCS1, (F) IDO1, (G) MMP1, (H) SCARB1, (I) SPP1, and (J) combined expression profile of the nine-gene signature.

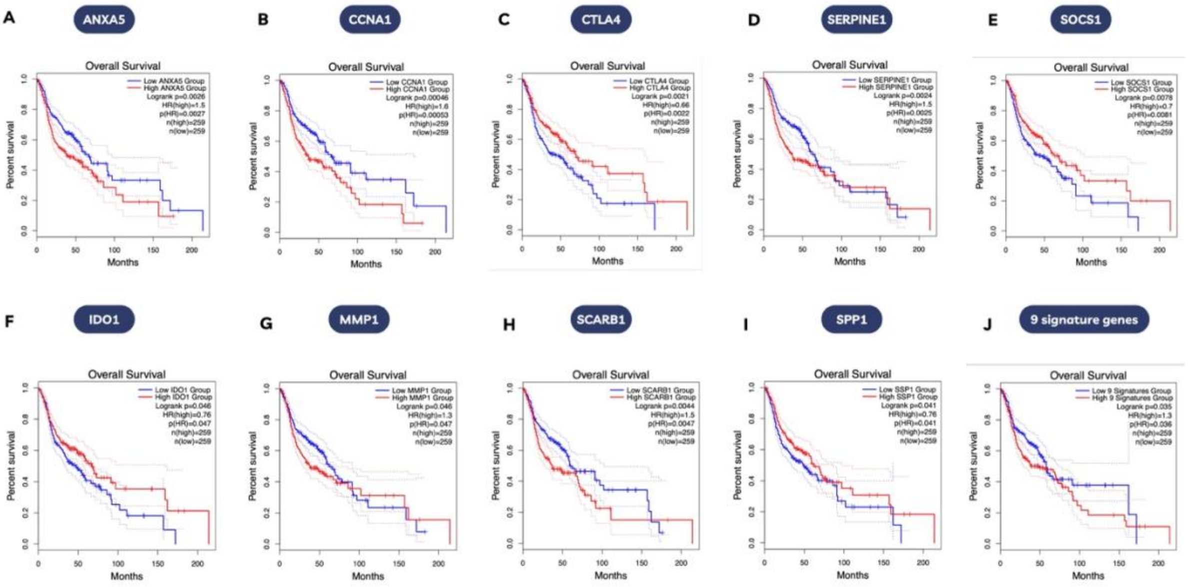

Fig 6. Survival analysis of hub genes in head and neck cancer (HNC) using GEPIA2. Kaplan–Meier overall survival curves for (A) ANXA5, (B) CCNA1, (C) CTLA4, (D) SERPINE1, (E) SOCS1, (F) IDO1, (G) MMP1, (H) SCARB1, (I) SPP1, and (J) the combined nine-gene signature, generated using the GEPIA2 database.

4.5 Expression analysis of significant hub genes across disease progression

To further evaluate expression patterns across disease progression, the nine hub genes identified as significantly associated with overall survival were subjected to expression analysis using normalized RNA-sequencing read counts from samples representing normal epithelium, hyperplasia, dysplasia, and head and neck cancer. Among these, five genes (CCNA1, MMP1, IDO1, ANXA5, and SERPINE1) exhibited statistically significant differential expression across disease stages in the study cohort (Fig 7). These genes were subsequently selected as candidate targets for drug repurposing screening.

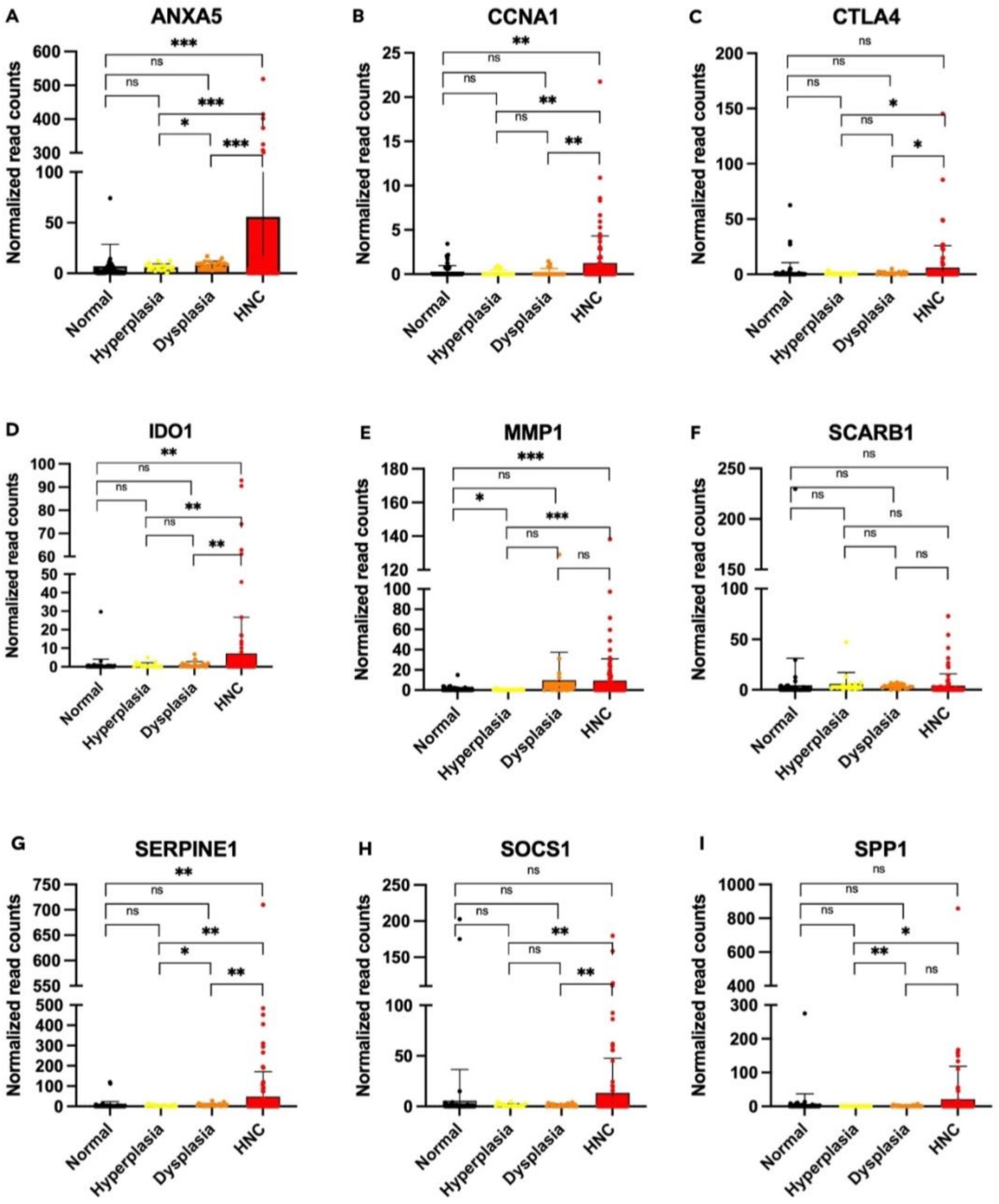

Fig 7. Expression of survival-associated hub genes across disease progression. Normalized RNA-sequencing read counts of nine hub genes in normal epithelium, hyperplasia, dysplasia, and head and neck cancer samples: (A) ANXA5, (B) CCNA1, (C) CTLA4, (D) IDO1, (E) MMP1, (F) SCARB1, (G) SERPINE1, (H) SOCS1, and (I) SPP1.

5. Conclusions

This study provides a comprehensive transcriptomic overview of head and neck cancer (HNC), revealing extensive transcriptional reprogramming characterized by 2,796 upregulated and 596 downregulated genes. Functional enrichment analyses indicated that upregulated genes are mainly involved in cell cycle regulation, immune and inflammatory responses, apoptosis, and metabolic reprogramming, whereas downregulated genes are associated with cytoskeletal organization, calcium signaling, and cellular communication. Network analysis identified 125 candidate hub genes, of which 60 were validated as significantly dysregulated in HNC. Survival analysis highlighted nine hub genes with prognostic significance, and five genes (CCNA1, MMP1, IDO1, ANXA5, and SERPINE1) showed consistent expression changes across disease progression. These genes were therefore prioritized for drug repurposing screening. Overall, this integrative bioinformatics analysis identifies clinically relevant molecular drivers of HNC and provides a strong foundation for subsequent drug repurposing and functional validation studies.

6. References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209–49. https://doi.org/10.3322/caac.21660.

[2] Hashibe M, Brennan P, Chuang S, Boccia S, Castellsague X, Chen C, et al. Interaction between Tobacco and Alcohol Use and the Risk of Head and Neck Cancer: Pooled Analysis in the International Head and Neck Cancer Epidemiology Consortium. Cancer Epidemiology, Biomarkers & Prevention 2009;18:541–50. https://doi.org/10.1158/1055-9965.EPI-08-0347.

[3] Blanco R, Carrillo-Beltrán D, Corvalán AH, Aguayo F. High-risk human papillomavirus and epstein–barr virus coinfection: A potential role in head and neck carcinogenesis. Biology (Basel) 2021;10. https://doi.org/10.3390/biology10121232.

[4] Serafini MS, Lopez-Perez L, Fico G, Licitra L, De Cecco L, Resteghini C. Transcriptomics and Epigenomics in head and neck cancer: available repositories and molecular signatures. Cancers Head Neck 2020;5. https://doi.org/10.1186/s41199-020-0047-y.

[5] Farah CS. Molecular landscape of head and neck cancer and implications for therapy. Ann Transl Med 2021;9:915–915. https://doi.org/10.21037/atm-20-6264.

[6] Nakhaei A, Marzoughi S, Ghoflchi S, Hosseini H, Afshari AR, Jalili-Nik M, et al. An exploration of molecular signaling in drug reprocessing for Oral Squamous Cell Carcinoma. Eur J Med Chem 2025;295:117816. https://doi.org/10.1016/j.ejmech.2025.117816.

[7] Saka Herrán C, Jané-Salas E, Estrugo Devesa A, López-López J. Protective effects of metformin, statins and anti-inflammatory drugs on head and neck cancer: A systematic review. Oral Oncol 2018;85:68–81. https://doi.org/10.1016/j.oraloncology.2018.08.015.