Aus der

Klinik und Poliklinik für Dermatologie und Allergologie

Klinikum der Ludwig-Maximilians-Universität München

Heterogeneity analysis of low-risk HPV infection and high-risk HPV infection, HPV-positive and HPV-negative cancers

Dissertation

zum Erwerb des Doktorgrades der Medizin

an der Medizinischen Fakultät der Ludwig-Maximilians-Universität zu München

vorgelegt von

Erdong Wei

aus

Guilin, China

Jahr

2023

Mit Genehmigung der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

Erster Gutachter:	Prof. Dr. Markus Reinholz
Zweiter Gutachter:	Priv. Doz. Dr. Jens Wallmichrath
Dritter Gutachter:	Prof. Dr. Ralf Zimmer

Dekan:	Prof. Dr. med. Thomas Gudermann

Tag der mündlichen Prüfung:	15.11.2023

Table of content

Zusammenfassung (Deutsch):

Zahlreiche Studien belegen, dass humane Papillomaviren (HPV) Krebs verursachen können. Allerdings haben nur wenige Studien die Heterogenität von HPV-infizierten oder nicht-infizierten (HPV-pos. und HPV-neg.) Krebserkrankungen untersucht. Zudem werden HPV-Infektionen mit niedrigem Risiko in der Regel mit gutartigen Läsionen in Verbindung gebracht, wobei sich nur wenige Studien mit der Heterogenität von HPV-Infektionen mit niedrigem und hohem Risiko befasst haben. Mit der Entwicklung von Next-Generation-Sequenzierung ist es möglich, Genome auf Einzelzellebene zu amplifizieren und zu sequenzieren, so dass die Heterogenität der Zellen mithilfe der Einzelzellsequenzierungstechnologie beobachtet werden kann. Auf dieser Grundlage werden zukünftig eine genauere Diagnose und Behandlung von HPV-Patienten möglich sein.

In dieser Studie wurde versucht, dieses Problem zu lösen, indem die Einzelzellsequenzierung mit der bulk-RNA-Sequenzierung kombiniert wurde, wobei die Einzelzellsequenzierungsdaten Veränderungen auf zellulärer Ebene lieferten und die bulk-RNA-Sequenzierungsdaten große Stichprobengrößen mit passenden klinischen Informationen enthielten. Die beiden wurden kombiniert, um mit Hilfe mehrerer Methoden wie Seurat für Zellcluster, CIBERSORT für Variationen der Immuninfiltration, WGCNA für charakteristische assoziierte Gencluster usw. zu analysieren.

Zunächst wurde die Hypothese aufgestellt, dass es eine Heterogenität zwischen HPV-pos. und HPV-neg.-Karzinomen gibt, und zwar von der Transkriptionsebene bis hin zur Immuninfiltration. Durch die Untersuchung von Gebärmutterhalskrebs, die mit Hochrisiko-HPV-Infektionen assoziiert ist, wurde bestätigt, dass $CD8^+$ T-Zellen und B-Zellen herunterreguliert wurden, während T-Reg-Zellen, $CD4^+$ T-Zellen und Epithelzellen in der HPV-pos.-Zervixkrebsgruppe hochreguliert waren. Die Analyse gutartiger Läsionen, die mit einer Niedrigrisiko-HPV-Infektion assoziiert sind, ergab Folgendes: Eine Niedrigrisiko-HPV-Infektion weist ähnliche genetische Veränderungen auf wie eine Hochrisiko-HPV-Infektion. Genetische Veränderungen, die durch eine Niedrigrisiko-HPV-Infektion verursacht werden, können auch die Prognose von Krebspatienten beeinflussen. Die Analyse von AIN3 und ASCC, CIN3 und CESC, präkanzerösen Läsionen und Tumoren, die in engem Zusammenhang mit Hochrisiko-HPV-Infektionen stehen, bestätigte auch, dass die Auswirkungen von Niedrigrisiko-HPV-Infektionen und Hochrisiko-HPV-Infektionen Ähnlichkeiten und Unterschiede aufweisen. Die Veränderungen in den Immunzellen waren bei den verschiedenen HPV-Infektionen teilweise gleich, während der Rest der Veränderungen in den Immunzellen durch die Krankheit selbst verursacht werden kann. Die Induktion von oxidativem Stress ist bei den verschiedenen HPV-Infektionen gleich, was zu oxidativem Stress führt, der DNA-Schäden verursacht und das optimale Umfeld für

eine maligne Transformation schafft. Schließlich wurde diese Erkenntnis durch in-vitro-Experimente bestätigt, bei denen sowohl normale als auch Krebszelllinien, die mit HPV transfiziert wurden, eine erhöhte Proliferation und eine hohe Expression von mit oxidativem Stress verbundenen Genen aufwiesen, und die hoch exprimierten Gene waren in der Lage, die Empfindlichkeit von mit oxidativem Stress assoziierten Inhibitoren zu erhöhen.

Die Bedeutung dieser Studie liegt in der Erklärung der Heterogenität zwischen HPV-pos. und HPV-neg.-Karzinomen und in der Entdeckung, dass die Gemeinsamkeit zwischen Niedrigrisiko-HPV-Infektionen und Hochrisiko-HPV-Infektionen im oxidativen Stress liegt. Dies könnte eine wichtige Entdeckung für eine künftige Präzisionsmedizin sein, damit Patienten mit HPV-assoziierten Krebserkrankungen präzise gezielte Behandlungen erhalten können.

Abstract (English):

Numerous studies have established the causal relationship between Human Papillomavirus (HPV) and cancer development. Nevertheless, there exists a paucity of research efforts dedicated to investigating the heterogeneity observed in HPV-associated cancers, differentiating between HPV-positive (HPV-pos.) and HPV-negative (HPV-neg.) cases. Furthermore, while low-risk HPV infections have traditionally been linked to benign lesions, investigations exploring the intricacies within both low-risk and high-risk HPV infections have been limited in scope. The advent of next-generation sequencing technologies, particularly single-cell sequencing, has revolutionized the field by facilitating genome amplification and sequencing at the single-cell level. This cutting-edge technology offers the unique capability to elucidate cellular heterogeneity, presenting a promising avenue for enhancing the precision of diagnostic and therapeutic approaches for HPV-infected patients, capitalizing on this newfound understanding of variability within the disease.

This study aimed to address this issue by integrating single-cell sequencing with bulk-RNA sequencing and DNA methylation sequencing. Single-cell sequencing data illuminated cellular-level alterations, while bulk-RNA sequencing and DNA methylation sequencing data encompassed larger sample sizes with accompanying clinical information. The analytical toolbox included methods such as Seurat for cell clustering, CIBERSORT for assessing immune infiltration variations, and WGCNA for identifying clusters of genes with characteristic associations.

Initially, we hypothesized that heterogeneity exists between HPV-pos. and HPV-neg. cancers, spanning from the transcriptional level to immune infiltration. In the context of the most well-established high-risk HPV infection-associated cancer, cervical cancer, $CD8^+$ T cells and B cells were observed to be down-regulated, whereas $T\text{-}_{reg}$ cells, $CD4^+$ T cells, and epithelial cells were up-regulated in the HPV-pos. cervical cancer group. Subsequent analysis of benign lesions associated with low-risk HPV infection revealed shared genetic alterations with high-risk HPV infection, and these genetic alterations impacted the prognosis of cancer patients. Furthermore, the examination of AIN3 and ASCC, CIN3 and CESC, precancerous lesions, and tumors firmly linked to high-risk HPV infection reaffirmed that while low-risk HPV infection and high-risk HPV infection share similarities, differences also exist. Changes in immune cells were partially consistent across different HPV infections, while other immune cell alterations may be attributed to the disease itself. It was observed that oxidative stress, common to various HPV infections, induced DNA damage, creating an environment conducive to malignant transformation. This finding was further corroborated by in vitro experiments, where both

normal and cancer cell lines transfected with HPV exhibited increased proliferation and upregulated expression of oxidative stress-related genes. These highly expressed genes increased sensitivity to oxidative stress-associated inhibitors.

The significance of this study lies in elucidating the heterogeneity between HPV-pos. and HPV-neg. cancers and identifying oxidative stress as a common factor in low-risk and high-risk HPV infections. Additionally, the altered sensitivity to genetic inhibitors resulting from HPV-induced genetic changes offers novel prospects for the treatment of HPV-pos. cancers, potentially paving the way for precision medicine approaches tailored to patients with HPV-associated cancers.

List of figures

List of abbreviations

HPV	Human Papillomavirus
µg	Microgram
µm	Micrometer
°C	Centigrade
Bp	Base pair
ORFs	Open-reading frames
LCR	Long control region
WHO	World Health Organization
CESC	Cervical cancer
HNSCC	Head and neck squamous cell carcinoma
ASCC	Anal squamous cell carcinoma
SCC	Squamous cell carcinoma
ACS	American Cancer Society
HER	Human epidermal growth factor receptor
NGS	Next-generation sequencing
scRNA-seq	Single-cell RNA sequencing
Bulk RNA-seq	Bulk RNA-sequencing
HPV-pos.	HPV-positive
HPV-neg.	HPV-negative
GEO	Gene Expression Omnibus
AIN3	Grade 3 anal intraepithelial neoplasia
CIN3	Grade 3 cervical intraepithelial neoplasia
SKCM	Skin cutaneous melanoma
TCGA	Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
vst	Variance stabilizing transformation
PCA	Principal component analysis
UMAP	Uniform Manifold Approximation and Projection

t-SNE	t-distributed stochastic neighbor embedding
DEGs	Differentially expressed genes
BEAM	Branch Expression Analysis Modeling
TPM	Transcripts per million
NA	Not available
ESTIMATE	Estimation of Stromal and Immune cells in Malignant Tumors using Expression data
ssGSEA	Single-sample gene set-enrichment analysis
CIBERSORT	Cell-type Identification by Estimating Relative Subsets of RNA Transcripts
FC	Fold change
WGCNA	Weighted gene co-expression network analysis
TOM	Topological overlap measure
ME	Module eigengene
MM	Module membership
Lasso	Least Absolute Shrinkage and Selection Operator
GO	Gene Ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
PPI	Protein-protein interaction
MCODE	Molecular complex detection
CTRP	Cancer Therapeutics Response Portal
GEPIA2021	Gene Expression Profiling Interactive Analysis 2021
WST-1	Water-Soluble Tetrazolium 1
OS	Overall survival
KM	Kaplan–Meier estimator
HR	Hazard ratio
BLCA	bladder urothelial carcinoma
COAD	colon adenocarcinoma
PRAD	prostate adenocarcinoma
EGFR	Epidermal growth factor receptor

OPSCC	Squamous Cell Carcinomas of the Oropharynx
pDCs	Plasmacytoid dendritic cells
ROS	Reactive oxygen species
DMP	Differential Methylation Probes
DMR	Differential Methylation Regions
DMB	Differential Methylation Blocks

1. Introduction

1.1. Human Papillomavirus Infection

1.1.1. Structure and classification of human papillomavirus

Human papillomavirus (HPV) is an epitheliotropic, non-enveloped, double-stranded DNA virus with a diameter of approximately 52-55 nm [1]. Viral particles consist of a single double-stranded DNA molecule comprising approximately 8000 base pairs (bp), which is bound to cellular histones and enclosed within a protein capsid composed of 72 pentameric capsomers. The circular double-stranded DNA genome encodes about 8 open-reading frames (ORFs), categorized into six early ORFs (E1, E2, E4, E5, E6, and E7), two late ORFs (L1 and L2), and a non-coding long control region (LCR) [2].

Notably, 80% of the papillomavirus capsid consists of the L1 protein, which harbors a conserved ORF. A new papillomavirus variant is defined if the complete genome, when cloned, deviates by less than 2% in DNA sequence from the closest known type. Subtypes are distinguished by homology differences ranging from 2% to 10%, while a novel isolate is recognized when differences exceed 10% [3]. This classification system has led to the identification of over 200 genotypes, with numerous variants coexisting within the same genotype [4]. High-risk HPV and low-risk HPV oncogenic types are categorized based on their oncogenic potential. High-risk HPV types, such as HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82, are associated with anogenital cancers, whereas low-risk types, including 6, 11, 40, 41, 42, 43, 44, 54, 61, 70, 72, 81, and 89, typically cause benign warts [5, 6].

E6 and E7 are the two primary oncoproteins of high-risk HPV and target several negative regulators of the cell cycle, primarily p105Rb and p53, respectively. Throughout the viral life cycle, E6 and E7 play essential roles in maintaining the stability of viral episomes and inducing differentiated cells to re-enter the S phase. This re-entry is crucial for sustaining the differentiated host keratinocyte in a state conducive to the amplification of viral genome replication and subsequent expression of late genes [3, 7]. E6 binds to host E6-associated proteins, recruiting a protein ligase with ubiquitin ligase activity that acts on ubiquitinated p53, resulting in the inhibition of p53's transcriptional activity and the suppression of p53-induced apoptosis. Additionally, E6 can impact transcriptional pathways, disrupt cell adhesion and structure, inhibit apoptosis, counteract DNA damage responses, induce genome instability, and immortalize cells by binding to numerous other cellular proteins [3, 8]. On the other hand, E7 competes with the tumor-suppressor protein p105Rb, releasing p105Rb's control over E2F and propelling the cell cycle forward.

Moreover, E7's ability to destabilize centrosomes allows it to interfere with transcriptional and signal transduction pathways, as well as DNA repair processes [3].

1.1.2. Epidemiology and pathogenicity of HPV infection

Human papillomavirus (HPV) infection stands as the most prevalent sexually transmitted infection, primarily transmitted through skin-to-skin contact. A study in 2014 underscored those individuals with at least one sexual partner have an average lifetime risk of HPV infection exceeding 80% [9]. Vaccination serves as a preventive measure against HPV infection, offering benefits not only to the vaccinated individual but also to their future sexual partners by curbing HPV transmission [10]. However, the COVID-19 pandemic has led to a loss of over 25% in HPV vaccine coverage achieved in 20[11].

The majority of HPV infections are asymptomatic, and approximately 90% of them resolve spontaneously within two years. Nevertheless, HPV infections can manifest as warts or precancerous lesions, elevating the risk of cancer, notably in the vulva, vagina, penis, and occasionally the anus [12].

According to the World Health Organization (WHO) report, in 2020, there were 604,127 new cases of cervical cancer (CESC) and 341,831 related deaths. An alarming 70% of CESC patients and CESC-related fatalities were found to be positive for HPV16 and HPV18 [12]. While most HPV infections of the cervix are swiftly cleared by the immune system without causing carcinogenesis in normal cervical cells, in certain cases, HPV disrupts host gene expression by deregulating DNA methylation, histone modification, and transcription factors to evade immune clearance [13]. Proteins E6 and E7, by binding to p53 and pRB, compel host cells to accumulate damaged DNA that cannot be repaired. Consequently, an accumulation of mutations occurs, leading to the transformation of cells into fully developed cancerous cells [14].

In addition to CESC, high-risk HPV strains are implicated in head and neck squamous cell carcinoma (HNSCC), anal squamous cell carcinoma (ASCC), and other genitourinary cancers [15, 16]. As per the American Cancer Society (ACS) report, HNSCC is expected to affect around 1.2% of both men and women at some point in their lives, with an annual incidence of 11.5 new cases per 100,000 individuals. HNSCC is projected to account for 2.8% of all new cancer cases in 2022 [17]. Over 90% of ASCC cases are attributed to HPV infection, with HPV16 being the predominant type. Although ASCC comprises only about 4% of lower gastrointestinal cancers [18], the incidence of anal cancer is on the rise, particularly in developed countries, where it is increasing by 1-3% annually [19].

1.2. DNA methylation

Pathophysiological investigations into HPV virus-induced carcinogenesis have revealed a pivotal process: the integration of HPV genomes into the host cell chromosomes disrupts the genomic stability of vulnerable host cells, subsequently triggering a secondary epigenetic reprogramming [20]. This phenomenon is characterized by the overexpression of E6 and E7 genes, which, in turn, stimulate the expression and activity of DNA methyltransferase I (DNMTI). This stimulation results in consequential hypermethylation within the host cells [21, 22]. Building on these findings, it has been hypothesized that in all HPV-related cancers, the maintenance of stemness-like differentiation status in epithelial cells is contingent upon the ongoing hypermethylation induced by E6 and E7, a process that plays a pivotal role in cancer progression [20].

Advancements in technology have significantly enhanced our capacity to explore DNA methylation on a global scale, thanks to profiling methods and computational approaches for processing extensive datasets. These capabilities transcend the limitations of traditional experiment-based studies, enabling comprehensive examinations of DNA methylation across various disease progressions [23]. However, most existing studies focused solely on DNA methylation anomalies in HPV-related cancers, often overlooking downstream changes in the transcriptome. This limitation somewhat diminishes the significance of their findings. To gain a profound understanding of the biological implications of aberrantly methylated genes, it is essential to integrate transcriptomic data analysis. Transcriptomic changes are not solely influenced by DNA methylation but are also impacted by other epigenetic modifications, including histone acetylation and microRNA alterations. Only through such integrated analyses can we unravel the intricate biological underpinnings of these epigenetic modifications.

1.3. Genomic testing and precision medicines

Precision oncology represents an evolving field with the goal of delivering personalized cancer treatments. It encompasses a broad spectrum of patient-specific data, including clinical information, lifestyle factors, genetic markers, and additional biomarker data. A notable success story in precision oncology revolves around the determination of the human epidermal growth factor receptor (HER)-2 status in breast cancer patients. Initially identified as a prognostic marker, HER-2 positivity was associated with a more aggressive disease course. Subsequent clinical trials demonstrated the efficacy of the monoclonal antibody trastuzumab against HER-2, which is now used to treat a specific subgroup of HER-2-positive patients [24]. The core principle of precision oncology lies in the characterization of inter- and intra-tumoral heterogeneity, recognized as a key factor in

therapeutic success and preventing disease relapse. However, before the significant advancements in next-generation sequencing (NGS) technology and the proliferation of bioinformatic tools, these endeavors were constrained by the limitations of traditional research methods.

Over the past decade, NGS techniques have been extensively employed to investigate gene expression patterns at the population level. Single-cell RNA sequencing (scRNA-seq) is a specialized methodology utilizing next-generation sequencing to unveil the global gene expression profile of individual cells. This breakthrough enables the exploration of previously undiscovered heterogeneity within cell populations. Computational analyses leveraging NGS and scRNA-seq data, coupled with functional enrichment analysis, enhance our understanding of various facets of tumor heterogeneity. This approach facilitates the identification of key genes driving tumor progression. With this data, prognostic models can be developed, combining multiple predictors to estimate the likelihood of survival or disease recurrence [12, 24]. Furthermore, the application of drug prioritization algorithms allows for the precise matching of potential anti-tumor medications to the genomic alterations and transcriptome characteristics of diverse cancer types.

In this study, we conducted comprehensive bioinformatic analyses by integrating DNA methylation and RNA sequencing data, encompassing both single-cell RNA-seq and bulk RNA-seq data, in the context of HPV-related cancers. Our objective was to identify pivotal genes implicated in the progression of HPV-related cancers. Subsequently, we performed survival analyses utilizing the TCGA database to validate the clinical significance of these identified genes. Drugs exhibiting a positive sensitivity correlation with these genes were further selected, and in vitro experiments were conducted to validate these findings. Through this research, we aspire to provide fresh insights into the disease mechanisms underlying HPV-related cancers. Our work lays the groundwork for the development of precise molecular targeted therapies and offers valuable insights for clinical decision-making.

2. Methods

2.1. Data collection

Single-cell RNA-seq data were acquired from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo/), which included the following samples:

2 HPV-positive cervical squamous cell carcinoma (CESC) samples (GSM5236544, GSM5236545)

2 HPV-negative CESC samples (GSM5236546, GSM5236547)

8 HPV-negative skin squamous cell carcinoma (SCC) samples (GSM4138147, GSM4138149, GSM4138151, GSM4138153, GSM4138155, GSM4138157, GSM4138159, GSM4138161)

18 HPV-positive SCC samples (GSE4138111, GSM4138113, GSM4138115, GSM4138117, GSM4138119, GSM4138121, GSM4138123, GSM4138125, GSM4138127, GSM4138129, GSM4138131, GSM4138133, GSM4138135, GSM4138137, GSM4138139, GSM4138141, GSM4138143, GSM4138145)

3 normal skin samples (GSM5364334, GSM5364335, GSM5364336)

Transcriptome profiling data and corresponding clinical data were collected from the GEO database, which included:

18 HPV-positive CESC samples (GSM155645, GSM155646, GSM155647, GSM155648, GSM155649, GSM155650, GSM155651, GSM155652, GSM155654, GSM155656, GSM155657, GSM155658, GSM155659, GSM155660, GSM155661, GSM155662, GSM155663, GSM155664)

2 HPV-negative CESC samples (GSM155653, GSM155655)

121 anal squamous cell carcinoma (ASCC) DNA methylation samples

13 adjacent grade 3 anal intraepithelial neoplasia (AIN3) DNA methylation samples

9 adjacent normal mucosae DNA methylation samples (GSE186858)

9 CESC DNA methylation samples

9 cervical intraepithelial neoplasia grade 3 (CIN3) DNA methylation samples

10 adjacent normal cervical DNA methylation samples

12 wart samples and paired normal skin samples (GSE136347)

Treeman syndrome sample and 3 paired control samples (GSM4136969, GSM4136984, GSM4136998, GSM4137012)

5 condyloma acuminata samples and 3 normal skin samples (GSE140662)

Additionally, TCGA-CESC samples and TCGA-skin cutaneous melanoma (SKCM) samples were obtained from the Cancer Genome Atlas (TCGA) project, with samples lacking clinical information or having histological types other than cervical squamous cell carcinoma being excluded. A total of 237 CESC samples were selected for analysis. Data related to ferroptosis, oxidative stress, and necroptosis were collected from the FerrDb database [25] and GeneCards [26]. Skin cancer samples were downloaded from the International Cancer Genome Consortium (ICGC) [27].

2.2. Single cell RNA-seq Data processing

2.2.1. Single cell RNA-seq data pre-processing

To assess the proportions of mitochondrial and ribosomal genes expressed in each cell, we utilized the PercentageFeatureSet function within the Seurat package. An upper threshold of 5% for the percentage of mitochondrial count was established, leading to the exclusion of cells exceeding this threshold [28].

After the cells were filtered, the data were normalized using the LogNormalize method in the Global Scaling Normalization package. This process consists of dividing the feature count of each cell by the total count of that cell, multiplying by 10^4, and then performing a natural logarithmic transformation [29].

Subsequently, the individual samples were merged into a unified dataset through the merge function. To facilitate the integration of multiple datasets, we employed the FindIntegrationAnchors and IntegrateData functions [28].

2.2.2. Highly Variable Features Identification

Highly variable features, denoting characteristics with pronounced expression differences between cells, were computed from the scRNA-seq dataset. These highly variable genes serve to emphasize biological signals, playing a pivotal role in downstream single-cell dataset analyses. To identify these highly variable features, we applied a variance stabilizing transformation (vst) model [30].

2.2.3. Principal component analysis

In the downstream analysis of high-dimensional data, linear dimensionality reduction is a pivotal step. It involves taking high-dimensional data as input and projecting it into an optimal low-dimensional space. Principal component analysis (PCA) is a commonly used technique that captures the covariance of features. In PCA, data are selected and extracted based on increasing variance, and the label "Principal Component" (PC) is assigned to those features exhibiting the highest variation. To identify the most informative components, we selected the top 20 PCs using visualization techniques such as the JackStraw plot (with 100 replicates) and the Elbow plot.

2.2.4. Cell-clustering

During this phase, a community detection algorithm was employed to cluster the cells. We utilized the R tool FindCluster() with the parameter "resolution = 1" to control the number of clusters generated. It's important to note that one of the significant limitations of PCA is its inability to capture non-linear variance. To address this limitation, we applied non-linear dimensionality reduction techniques, namely Uniform Manifold Approximation and Projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE). These techniques were used to visualize the clustering of single cells in a lower-dimensional space, providing a more comprehensive representation.

2.2.5. Biomarkers of clusters

We identified cluster-specific marker genes by utilizing the FindMarkers() tool within the Seurat package. This involved employing the default non-parametric Wilcoxon rank sum test with Bonferroni correction. To annotate the cell groups based on the differentially expressed genes (DEGs) between clusters, we referred to a well-established cellular marker reference obtained from the R package Celldex [31]. Cell annotation was facilitated using the R package SingleR [31].

2.2.6. Single-cell trajectories

We employed the Monocle3 R package to analyze single-cell trajectories with default settings. This approach assumes that a one-dimensional 'time' parameter can effectively describe high-dimensional expression values, allowing us to discover transitions in cell

states [32]. The clusters obtained earlier were loaded into the analysis. The "differentialGeneTest" function was used to identify genes that exhibited variations along with the pseudotime, and these genes were visualized using the plot pseudotime heatmap function. Subsequently, they were categorized into subgroups based on their expression patterns.

To further investigate cell branching, we conducted Branch Expression Analysis Modeling (BEAM) analysis to identify genes that played a role in dividing cells into branches. The genes identified through BEAM analysis were grouped and presented using the plot genes branched heatmap() function.

2.3. Bulk RNA-seq Data processing

2.3.1. Bulk RNA-seq Data pre-processing

Gene expression quantification is essential for a comprehensive analysis of differential gene expression across multiple conditions. To achieve this, we utilized transcripts per million (TPM) for bulk RNA-seq quantification. GFF3 annotation files were obtained from GENCODE (https://www.gencodegenes.org/human/). With the gene length data from GFF3 and the count data from bulk RNA-seq, TPM values were calculated for each gene using the formula TPMi = (Ni / Li) * 1,000,000 / sum(N0 / L0 + … + Nm / Lm), where Ni represents the read counts mapped to gene I, Li refers to the sum of exon lengths for gene I, and m represents the total number of genes.

To address missing entries in the bulk RNA-seq matrix, we performed matrix completion. Rows with more than X% of unavailable (NA) values (genes) and columns with more than Y% of NA values (samples) were excluded. Subsequently, we used the impute.knn function from the R package impute [33] to fill in missing values by considering K neighbors. Furthermore, various data normalization techniques were applied, including quartile normalization, median normalization, log2 transformation, row z-score transformation, column z-score transformation, as well as row and column z-score transformations.

These steps ensured a comprehensive and standardized analysis of gene expression across conditions.

2.3.2. Immune Estimate

The estimation of immune scores was conducted using the ESTIMATE (Estimation of Stromal and Immune cells in Malignant Tumors using Expression data) method [34]. This approach specifically targets the stromal and immune cells that constitute the major non-tumor components within a tumor sample. It identifies distinct features associated with stromal and immune cell infiltration within the tumor tissue. By employing single-sample gene set enrichment analysis (ssGSEA), the ESTIMATE package calculated stromal and immune scores based on markers related to stromal tissue and immune cell infiltration. The combined score, referred to as the "ESTIMATE score," was derived from these two scores. This methodology enabled the determination of the proportions of stromal and immune cells in tumor samples using gene expression profiles.

For the analysis of immune cell proportions within bulk RNA-seq samples, we employed the Cell type Identification by Estimating Relative Subsets of RNA Transcripts (CIBERSORT) method [35]. We utilized the LM22 signature matrix, available for download at https://cibersortx.stanford.edu/download.php. LM22 comprises 547 genes that effectively distinguish 22 distinct mature human hematopoietic populations isolated from peripheral blood or in vitro culture conditions. These populations encompass various T cell types, including naive and memory B cells, plasma cells, NK cells, as well as subsets of myeloid cells. Through the use of the support vector regression tool, we accurately estimated the immune composition of tumor samples [36].

2.3.3. DEGs analysis

To perform the analysis of Differentially Expressed Genes (DEGs), we utilized the "limma" R package [37]. In the case of CESC samples, considering that the difference between HPV-positive and HPV-negative samples was not significant, we set the threshold for DEGs as $|\log2\text{fold-change (FC)}| > 1.2$ and $p < 0.05$. Similarly, for the DEGs between normal skin, AIN3, and ASCC samples, we applied a cutoff value of $|\log2FC| > 1.5$ and $p < 0.05$. When comparing other disease sample groups with the healthy group, we employed a threshold of $|\log2FC| > 2$ and $p < 0.05$.

These criteria ensured the identification of significant gene expression differences in the respective sample comparisons.

2.3.4. Weighted correlation network analysis (WGCNA)

The WGCNA package was employed for weighted correlation network analysis [38]. WGCNA is a systems biology approach that constructs a weighted gene co-expression network to describe patterns of gene associations across samples. This approach helps identify highly synergistic gene sets, potential biomarker genes, or therapeutic targets based on genomic endogeneity and associations with phenotypes. Unlike focusing solely on differentially expressed genes, WGCNA leverages information from thousands, or even nearly 10,000 of the most variable genes, or all genes, to identify sets of genes of interest and perform significant association analyses with phenotypes. WGCNA transforms numerous gene-to-phenotype associations into a smaller number of gene-set-to-phenotype associations, mitigating the need for multiple hypothesis testing corrections.

In this study, WGCNA was used to link gene modules with immune features. The system selected genes for further investigation based on the top 25% of variance. Pearson's correlation matrices were computed using the samples.

In the unsigned network, genes represent nodes, and the interactions between nodes represent edges. The edge properties were calculated as abs(cor(gene_x, gene_y))^power, where the appropriate power value was determined using the pickSoftThreshold function in the WGCNA package. This calculation strengthens strong correlations while weakening negative correlations, aligning the correlation values with the characteristics of a scale-free network, and making them more biologically meaningful. Additionally, a topological overlap measure (TOM) matrix was created using the weighted adjacency matrix to quantify the quality of connectedness in the network. A clustering dendrogram of the TOM matrix was generated using average linkage hierarchical clustering. The threshold for merging comparable modules was set to 0.25, and the minimum gene module size was set to 30 to ensure appropriate modules. Co-expression modules consist of genes with a high degree of topological overlap, often indicating stronger co-expression within the same module.

Two different techniques were used to identify key modules related to immune features. The module eigengene (ME), representing the first principal component of a module, explains the module's expression pattern in each sample. Module membership (MM) measures the reliability of a gene belonging to a module, expressed as the correlation coefficient between genes and module eigengenes. Finally, modules were assessed for their correlation with immune scores and proportions to identify crucial immune modules.

2.3.5. Molecular Risk Model Construction

The Least Absolute Shrinkage and Selection Operator (Lasso) is a linear regression method that incorporates L1-regularization, which encourages some of the learned features to have a weight of 0, resulting in sparsity and feature selection. The fundamental concept behind Lasso is to minimize the sum of squared residuals while imposing a constraint that the sum of the absolute values of the regression coefficients is less than a constant. This constraint enables certain regression coefficients to be exactly 0, leading to a more interpretable model [39].

Given that the number of genes is much larger than the number of observations (i.e., patients), Lasso is used to address the issue of covariance. Cox risk regression was performed using the `coxph()` function from the survival package. A p-value of 0.05 was considered significant for association with survival or relapse [40].

Lasso employs a penalty function to compress certain coefficients and set some of them to zero, resulting in a refined model with the advantage of subset shrinkage. To effectively handle multicollinearity in regression analysis and simultaneously select variables while estimating parameters, a biased estimation for multicollinear data processing is employed. The analysis was conducted using the glmnet package, utilizing the 10-fold cross-validation method and Lasso Cox regression to construct the model [39].

2.3.6. Functional Enrichment Analysis

Gene Ontology (GO) [41], Kyoto Encyclopedia of Genes and Genomes (KEGG) [42], and Metascape database [43] enrichment analyses were conducted to identify the biological processes and pathways that are most closely associated with hub genes, using a significance cutoff of $p < 0.05$. These analyses help reveal the functional context and pathways in which the identified hub genes are involved, providing insights into their roles in the biological system.

2.3.7. Protein-protein interaction (PPI)

The investigation of the biochemistry of individual proteins and the creation of a comprehensive picture of protein interactions on a proteomic scale rely on the ability to identify protein-protein interactions. In this study, the STRING database was used to identify co-expressed genes and to construct protein interaction networks [44]. Additionally, Cytoscape's molecular complex detection (MCODE) algorithm was employed to

identify key proteins within these large protein interaction networks [45]. Parameters such as degree cutoff, node score cutoff, K-core, and maximum depth were set to identify strongly interacting nodes in the clusters of interest. This approach helps identify important proteins and their interactions in complex biological networks.

2.3.8. Cancer therapeutics response analysis

The correlation between chemical sensitivity and gene expression levels was assessed using data from the Cancer Therapeutics Response Portal (CTRP) V2 [46]. The CTRP data matrix contains normalized AUC (Area Under the Curve) values for each chemical in each cell line. To ensure robust statistical analysis, compounds that were profiled in at least 2/3 of the entire collection of solid cancer cell lines and cell lines that were profiled with at least 50% of the compounds were included in the lineage/histotype-specific studies. Primary cancer types with more than 5 cell lines profiled for chemical sensitivity were considered.

To investigate the relationship between chemical sensitivity and gene expression, Mann-Whitney-Wilcoxon tests were conducted to compare cancer cell lines from each type of cancer with other solid cancer cell line collections. To account for multiple-test correction, the Benjamini-Hochberg correction approach was applied to adjust the statistical significances [47]. Furthermore, compound sensitivity-gene expression correlation analysis was performed, and outliers were identified using an interquartile multiplier of 1.5 with the web tools available in CTRP. This analysis aims to uncover associations between gene expression and the sensitivity of cancer cell lines to specific compounds.

2.3.9. HPV-Associated Pan Cancer Analysis

The Gene Expression Profiling Interactive Analysis 2021 (GEPIA2021) is a web-based program that allows for the examination of the expression of hub genes and overall survival in cancers associated with HPV [48]. It addresses the limitations of bulk RNA-seq data, which lacks the resolution to provide information at the level of individual cell types and their proportions. GEPIA2021 integrates bulk expression data from The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) datasets. Additionally, it employs deconvolution methods such as CIBERSORT, EPIC, and quanTIseq to estimate the absolute percentage of different cell types in each sample.

2.4. DNA Methylation Data processing

2.4.1. Methylation Profiling and Data Analysis

The analysis of methylation raw data was conducted using the Chip Analysis Methylation Pipeline (champ) R package [49]. Champ is a comprehensive methylation analysis package that includes various features for quality control and the identification of differentially methylated probes (DMPs), differentially methylated regions (DMRs), and differentially methylated blocks (DMBs).

Here's an overview of the specific steps taken in the methylation data analysis using champ:

1. Quality Control: Probes that met certain criteria were filtered out. This included probes with a detection p-value greater than 0.01, probes with fewer than 3 beads in at least 5% of samples per probe, and probes located on the sex chromosome.

2. Normalization: The differentially methylated probes were normalized using the champ.norm() function to account for potential biases in the methylation data.

3. Identification of Differentially Methylated Probes (DMPs): The champ.DMP() function was used to calculate the methylation differences (β value) of probes. DMPs were selected based on criteria such as a |logFC| (log-fold change) greater than 0.2 and a p-value less than 0.05.

2.4.2. Function annotation

In the analysis of the differentially methylated probes (DMPs) identified in the previous step, several database resources were used for functional annotation and pathway analysis. Here's an overview of the methods used:

1. Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Analysis: The DMPs were annotated with KEGG pathway information. KEGG is a valuable resource for understanding the biological pathways and functions associated with genes and their epigenetic regulation. This analysis helped identify which KEGG pathways were enriched or affected by the DMPs.

2. FerrDb and GeneCard Database: Additional annotations were obtained from the FerrDb and GeneCard databases. FerrDb is a resource specifically focused on ferritin-related genes and their regulation. GeneCard provides comprehensive information on various genes, including their functions, interactions, and relevance to diseases. Data from GeneCard with a relevance score above the median value were selected for mapping.

2.4.3. Immune analysis

In this study, we conducted a comprehensive analysis of immune cell proportions and their relationship with gene expression [35]. We began by identifying genes associated with immune cell proportions, which were obtained from Differentially Methylated Probes (DMP), using the Cell type Identification by Estimating Relative Subsets Of known RNA Transcripts (CIBERSORT) method. Subsequently, we employed the Seurat R package for identification of genes related to the different immune infiltration patterns identified by CIBERSORT [28]. An upper bound threshold for the percentage of mitochondrial count (5%) was defined, and the cells above the upper bound were filtered out [28]. Data normalization was carried out after cell filtering, utilizing the global-scaling normalization package LogNormalize, which divides the specific feature counts of each cell by the overall counts of that cell, divides it by 10^4, and then performs a natural log-transformation [29]. The samples were then merged into a single data set using the merge function. The FindIntegrationAnchors function was used to find the anchors, and the IntegrateData function was used to integrate multiple data sets [30]. Community detection algorithm was applied for clustering the cells, R tool FindCluster() and the parameter "resolution = 1" for controlling the number of clusters. Non-linear dimensional reduction technique Uniform Manifold Approximation and Projection (UMAP) and t-distributed stochastic neighbor embedding (t-SNE) were performed to visualize single cell clustering in low-dimension. The cluster-specific marker genes were obtained using the Findmarkers () tool in the Seurat package with default non-parametric Wilcoxon rank sum test as well as Bon-ferroni correction. The well-known cellular marker reference was obtained from R package Celldex, cells were automatically annotated with R package SingleR [31]. Plot1cell R package [50] was used to visualize and quantify the scRNA data.

2.5. Molecular docking

Autodock Vina, a computational tool for protein-ligand docking analysis, was utilized to assess the binding affinities and interaction mechanisms between drug candidates and

their respective targets [51, 52]. The molecular structure of afatinib (PubChem 10184653) was obtained from PubChem Compound (https://pubchem.ncbi.nlm.nih.gov/) [53]. The 3D coordinates of CDC42 (PDB ID: 1AJE; Resolution: Not Applicable) [54] and EGFR (PDB ID: 6VH4; Resolution: 2.80 Å) [55] were sourced from the Protein Data Bank (PDB) (http://www.rcsb.org/). To prepare the protein and ligand data for docking analysis, all water molecules were removed, and polar hydrogen atoms were added, followed by conversion to the PDBQT format. The docking grid box was positioned at the center of the proteins to enable unrestricted molecular mobility and encompass the domains of each protein.

In addition, rigid protein-protein docking using ZDOCK was carried out to explore the interactions between CDC42 and EGFR [56]. Protein structural domains in PDB format were obtained from the same database. The 3D coordinates of the CDC42-GTPase-effector interface (PDB ID: 5UPL; Resolution: 3.00 Å) [57] and the EGFR-afatinib complex (PDB ID: 4G5J; Resolution: 2.80 Å) [58] were retrieved from the PDB. The ZDOCK module was employed to identify potential docking sites and calculate ZDOCK scores, shedding light on the binding interactions between these two proteins.

2.6. *In vitro* experiment

2.6.1. Cell lines and culture conditions

The A431 epidermoid carcinoma cell line and HaCaT keratinocytes cell line were procured from the American Type Culture Collection (ATCC, Wesel, Germany). These cell lines were maintained in DMEM (Sigma-Aldrich, Schnelldorf, Germany) media supplemented with 10% Fetal Bovine Serum (FBS) (Sigma-Aldrich, Schnelldorf, Germany) and 1% penicillin-streptomycin (Sigma-Aldrich, Schnelldorf, Germany) at a temperature of 37 °C in a humidified incubator with 5% CO2. Subculturing was performed at regular intervals with a split ratio of 1:4, typically twice a week. Trypsin-EDTA solution (Sigma-Aldrich, Schnelldorf, Germany) was employed for the subcultivation process.

2.6.2. Cell transfection

The HPV E6/E7 expressing plasmid (p1321 HPV16 E6/E7, Addgene #8641) was generously provided by Prof. Peter Howley [59]. LB agar was prepared by mixing 5.0 g yeast extract, 10.0 g peptone from casein, 10.0 g sodium chloride, and 12.0 g agar-agar in 1 liter of distilled water, followed by autoclaving at 121°C for 15 minutes. Subsequently, 100 µg/ml Ampicillin was added to the molten agar mixture. Bacterial stab containing

the plasmid was streaked onto an LB agar plate and grown overnight at 37°C in a humidified incubator with 5% CO2. Single colonies were isolated, and Ampicillin-resistant colonies were selected on LB agar and cultured overnight in LB/ampicillin medium at 37°C in a humidified incubator with 5% CO2.

Plasmid DNA was extracted from the bacterial culture using ethanol precipitation. To the bacterial culture transformed with plasmid, 2.5 volumes of 99% ethanol and 10% of a volume of 3M sodium acetate (pH 4.8) were added. The mixture was inverted to mix and placed at -80°C for 30 minutes. After centrifugation at 12,000 rpm for 30 minutes at 4°C, the supernatant was discarded, and the pellet was air-dried for 20 minutes. The pellet was then washed with 500 µl of cold 70% ethanol and centrifuged at 12,000 rpm for 5 minutes at room temperature. After removing the washing supernatant, the tube was air-dried for 20 minutes, and the dry DNA was resuspended in TE solution (10 mM Tris-HCl pH 8.0 and 0.1 mM EDTA).

Transfection of the plasmid DNA was carried out using X-tremeGENE 9 DNA transfection reagent (Roche, Mannheim, Germany). The plasmid DNA solution was diluted to a concentration of 1.0 µg/µl, and A431/HaCaT cells with a low passage number were selected. Cells were seeded in the logarithmic growth phase, and the cell concentration was accurately quantified using a Countess 3 automated cell counter (Thermo Fisher Scientific, Planegg, Germany). Transfections were performed 24 hours after cell seeding. The X-tremeGENE 9 DNA transfection reagent was diluted with serum-free DMEM to a concentration of 3 µl reagent/100 µl DMEM (3:1 ratio). Subsequently, 1 µg of DNA was mixed with 100 µl of the diluted X-tremeGENE 9 DNA transfection reagent, and the DNA transfection reagent complex was incubated for 20 minutes at room temperature. For 96-well plates, 5 µl of the DNA transfection reagent complex was added to each well, and for 10-cm dishes, 500 µl of the complex was used. The cells were incubated for 24 hours before further analysis.

2.6.3. Cell viability assay

Following the 24-hour transfection, HaCaT and A431 cells were subjected to treatment with various concentrations (0µM, 0.1µM, 1µM, 10µM) of afatinib (SML3109, Sigma-Aldrich, Schnelldorf, Germany). Afatinib was identified as a compound exhibiting significant correlation with genes playing crucial roles in the carcinogenesis of keratinocytes subsequent to HPV infection, as detailed in section 2.3.8. Untransfected groups were seeded and treated under identical conditions concurrently. In the case of the transfected groups mixed with the DNA transfection reagent complex, the untransfected

groups received the transfection reagent without DNA. Subsequently, cell viability was assessed using the Water-Soluble Tetrazolium 1 (WST-1) assay (Sigma-Aldrich, Schnelldorf, Germany). Specifically, 10 µl of the WST-1 reagent was introduced into each well within 96-well plates. Following a 4-hour incubation at 37°C with 5% CO2, the absorbance of the samples was measured using a plate reader (Spectra MR, Dynex Technologies, Chantilly, USA) at a wavelength of 440 nm.

2.6.4. Protein extraction

HaCaT and A431 cells were seeded in 10 cm dishes and allowed to incubate overnight, and the grouping for subsequent experiments mirrored that of the cell viability assay. After exposing the cells to varying concentrations (0µM, 0.1µM, 1µM, 10µM) of afatinib for 24 hours, protein extraction was carried out using RIPA lysis and extraction buffer (89901, Thermo Fisher Scientific, Planegg, Germany) along with protease inhibitors set (Roche, Mannheim, Germany).

During protein extraction, the cells were maintained on ice, with culture medium being removed and the cells washed twice with cold PBS. Subsequently, 1 ml of cold RIPA lysis buffer was added for every 5*10^6 cells, and the plate was swirled occasionally for 5 minutes. The lysate was then collected and transferred to a microcentrifuge tube, followed by centrifugation at 14,000 × g for 15 minutes to remove cell debris. The supernatant was transferred to a new tube for analysis via the Bradford protein assay. The total protein concentration was determined using the Bradford assay (Bio-Rad Laboratories GmbH, Feldkirchen, Germany), which relies on Coomassie Brilliant Blue G-250 dye. This dye undergoes a color change from red to blue upon binding to proteins, which can be quantified spectroscopically. A dye reagent was prepared by diluting 1 part dye reagent concentrate with 4 parts deionized water, and it was filtered to remove any particles. A protein standard was prepared with five dilutions ranging from 0.25 to 1.4 mg/ml. Subsequently, 50 µl of each standard and test sample solution was pipetted into a 96-well plate, and 200 µl of the diluted Bradford reagent was added to each well. After a 5-minute incubation at room temperature, the absorbance of each well was measured using a plate reader (Spectra MR, Dynex Technologies, Chantilly, USA) at a wavelength of 595 nm. The obtained absorbance of the test solution was subjected to regression calculations and multiplied by the dilution factor to determine the protein quantity.

2.6.5. Western Blot

For SDS-PAGE gel preparation, a 10% separating gel was created by mixing 4 ml of deionized water, 3.3 ml of 30% acrylamide, 2.5 ml of 1.5 M Tris-HCl (pH 8.8), 0.1 ml of 10%

SDS, 0.1 ml of 10% ammonium persulfate (APS), and 0.01 ml of TEMED. Additionally, 5% stacking gels were prepared using 2.7 ml of deionized water, 0.67 ml of 30% acrylamide, 0.5 ml of 1 M Tris-HCl (pH 8.8), 0.04 ml of 10% SDS, 0.04 ml of 10% ammonium persulfate (APS), and 0.004 ml of TEMED after 30 minutes.

A Tris-glycine-methanol transfer buffer was created by diluting 10× Tris-glycine buffer with methanol and water to achieve a final solution containing 25 mM Tris, 192 mM glycine, and 0.1% SDS. A protein ladder (Sigma-Aldrich, Schnelldorf, Germany) was utilized to determine molecular mass (5 µl). Each well of the SDS-PAGE gel was loaded with 10 µl of cell lysate and 2 µl of 6× loading buffer (Sigma-Aldrich, Schnelldorf, Germany). Electrophoresis was conducted at 80 V for 50 minutes, followed by 120 V until the marker proteins reached the bottom of the gel.

Protein samples were subsequently transferred onto polyvinylidene fluoride (PVDF) membranes (Sigma-Aldrich, Schnelldorf, Germany) using a transfer buffer composed of 12 mM Tris and 96 mM glycine. The PVDF membranes were activated with methanol for 5 minutes. A Filter Paper Sandwich (Thermo Fisher Scientific, Planegg, Germany) configuration (sponge-filter paper-gel-membrane-filter paper-sponge) was assembled in the transfer tank, and any air bubbles were removed. The transfer was performed at 200 mA for 90 minutes on ice.

After completing the transfer, the membrane was removed from the filter paper sandwich. The PVDF membranes were blocked with 5% skim milk in TBST (20 mM Tris, 150 mM NaCl, and 0.1% Tween 20) for 2 hours at room temperature after a 10-minute wash with PBS. Subsequently, the membranes were washed three times for 5 minutes each with TBST. Primary antibodies targeting CDC42 (HPA069590, Sigma-Aldrich, Schnelldorf, Germany), EGFR (AMAB90816, Sigma-Aldrich, Schnelldorf, Germany), and GAPDH (#2118, Cell Signaling Technology, USA) were applied and allowed to incubate overnight at 4 °C. The membranes were washed three times for 5 minutes each with TBST the following day. Secondary antibodies were applied to the membranes and incubated at room temperature for 1 hour. Afterward, the membrane was washed again three times for 5 minutes each with TBST. Finally, protein bands were visualized using the ChemiDoc Imaging Systems (Bio-Rad Laboratories GmbH, Feldkirchen, Germany).

3. Results

3.1. Heterogeneities between HPV-pos. and HPV-neg. CESC

3.1.1. ScRNA-seq analysis of HPV-pos. and HPV-neg. CESC

CESC is the most known high-risk HPV-associated cancer. Hence, CESC samples were used for initial exploration to find out whether there are heterogeneities between HPV-pos. and HPV-neg. cancers. In the analysis of HPV-pos. and HPV-neg. CESC scRNA-seq samples, scRNA data were clustered into 19 groups due to 11,899 marker genes (Fig 1A, B), top 5 marker genes of each cluster were used for heatmap visualization. (Fig 1C).

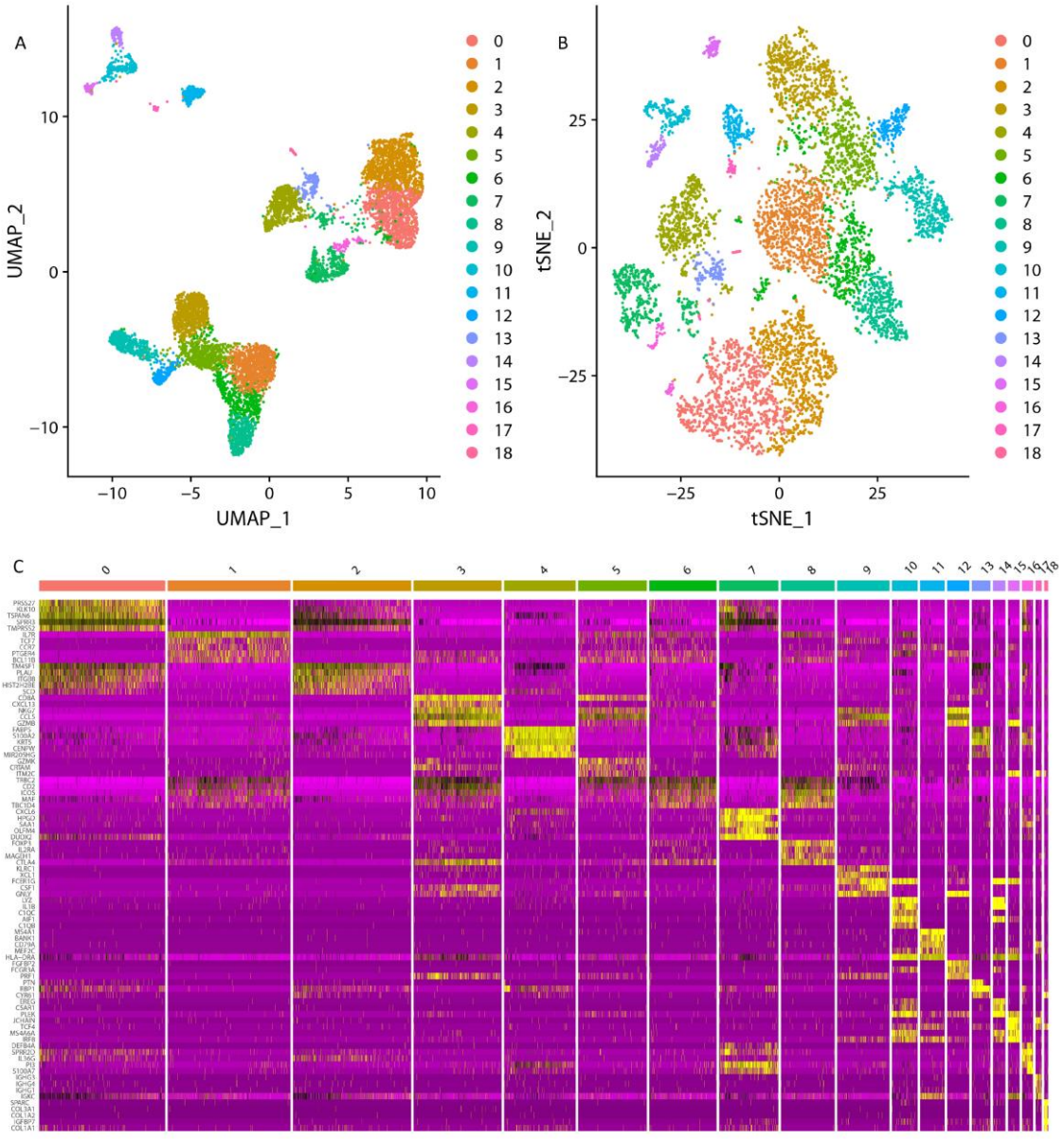


Figure 1 The dimension reduction of CESC scRNA-seq.

A. Separate clusters according to UAMP.
B. Separate clusters according to tSNE.
C. Top five marker genes of 19 clusters.

In the CESC scRNA-seq dataset, cell annotations revealed the presence of nine distinct cell types (Fig 2A), which were determined based on variations in marker gene expression levels (Fig 2B, C). Notably, t-SNE dimensionality reduction illustrated a reduction in CD8+ T cell and B cell clusters, accompanied by an increase in Treg cell, CD4+ T cell, and epithelial cell clusters within HPV-positive CESC (Fig 2A). Pseudotime trajectory analysis further elucidated the intricate process of cell differentiation (Fig 2D, E), underscoring the marked heterogeneity in cell development between HPV-positive and HPV-negative CESC. Pseudotime plots depicted the estimated lineage trajectories for each cluster based on the scRNA-seq data, with dots representing individual cells and black lines representing cell differentiation paths. Due to the limited sample size and the absence of matched clinical data in the scRNA-seq dataset, complementary analyses were conducted using bulk RNA-seq data to explore HPV-related heterogeneity. Given the observed immune differences in the scRNA-seq data, additional immune analyses were conducted using bulk RNA-seq data.

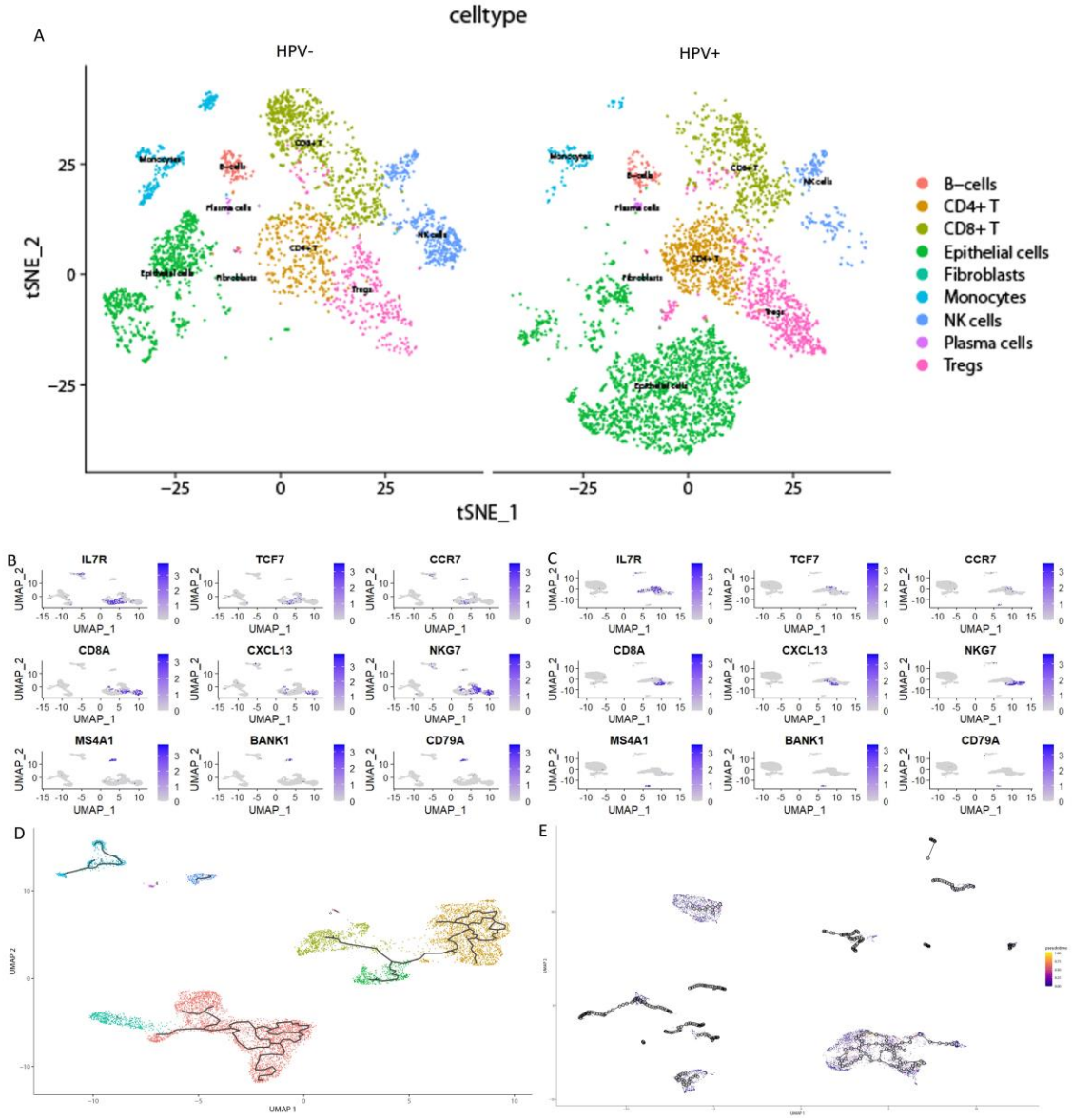


Figure 2 The heterogeneity in HPV-pos. and HPV-neg. CESC scRNA-seq.

A. Cell type annotation of HPV-pos. and HPV-neg. group.
B. Top marker genes of $CD8^+$T, $CD4^+$T, and B cells in HPV-neg. group.
C. Top marker genes of $CD8^+$T, $CD4^+$T, and B cells in HPV-pos. group.
D. Pseudotime and trajectory analysis of HPV-neg. group.
E. Pseudotime and trajectory analysis of HPV-pos. group.

3.1.2. Bulk RNA-seq analysis of HPV-pos. and HPV-neg. CESC

Utilizing the ESTIMATE algorithm with bulk RNA-seq data enabled an in-depth exploration of immunological signatures between HPV-positive and HPV-negative CESC groups, capitalizing on the advantages of large-scale integration with matched clinical infor-

mation. While no significant differences were observed in stromal cell infiltration between HPV-positive and HPV-negative CESC (Fig 3A), disparities in immune cell infiltration were evident. Subsequently, CIBERSORT analysis was conducted to delineate the immune cell landscape, revealing the proportions of 22 distinct immune cell types (Fig 3B). These proportions, in conjunction with clinical data from bulk RNA-seq, were employed in overall survival (OS) analysis (Fig 4). Notably, OS analysis indicated no significant associations between survival probability and the expression levels of various immune cell types, including activated dendritic cells, resting dendritic cells, eosinophils, macrophages M0, macrophages M1, macrophages M2, resting mast cells, neutrophils, activated NK cells, resting NK cells, plasma cells, resting CD4+ T memory cells, naïve CD4+ T cells, follicular helper T cells, and gamma delta T cells. Consequently, these immune cell clusters were excluded from subsequent analyses due to their lack of statistical impact on the survival outcomes of CESC patients. Although low expression levels of activated mast cells were associated with favorable CESC OS (P = 7.7e-4), they were not further investigated, as scRNA-seq data revealed no differences in the abundance of activated mast cells between HPV-positive and HPV-negative CESC. Furthermore, genes within the epithelial cell cluster were excluded from analysis due to their absence among the 22 immune cell types. Genes from the naïve B cells and CD8+ T cells clusters were selected for further investigation, as Kaplan–Meier estimator (KM) curves indicated a lower survival probability for patients with reduced abundances of CD8+ T cells and naïve B cells (Fig 3C, D).

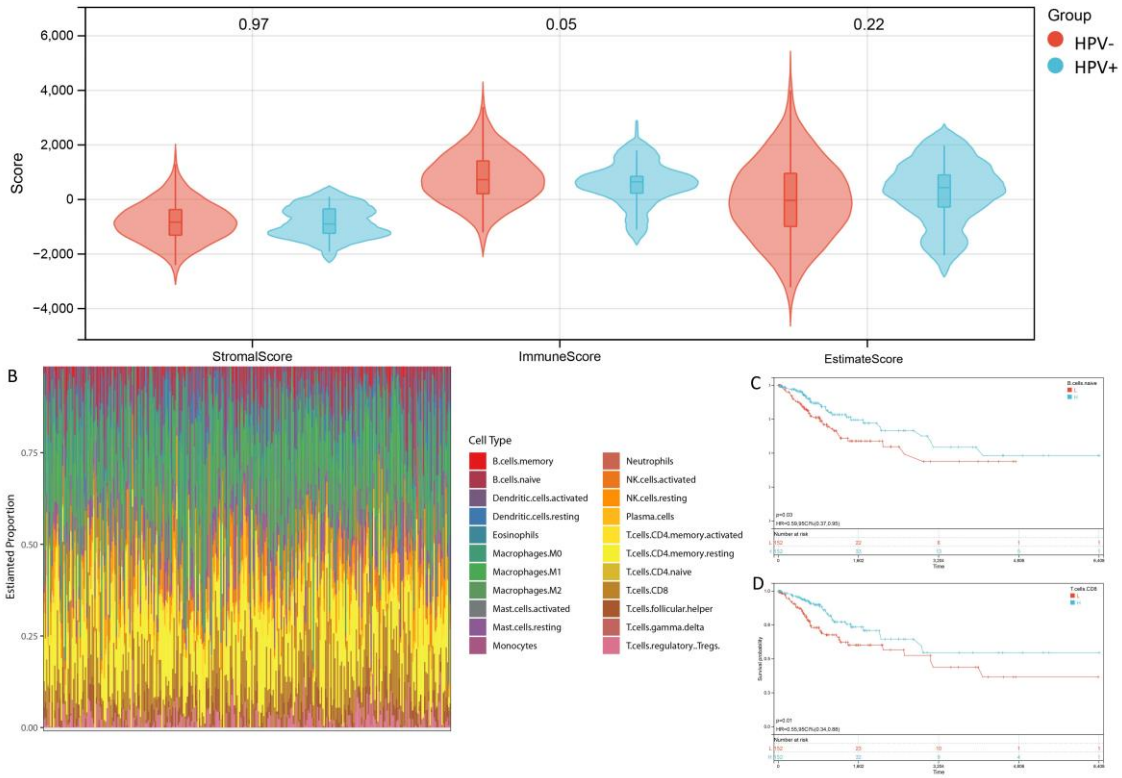


Figure 3 Character of immune infiltration in bulk RNA-seq.

A. Stromal score, Immune Score and Estimate Score of HPV-pos. and HPV-neg. bulk RNA-seq samples.
B. Bar charts of 22 immune cell proportions in bulk RNA-seq samples.
C. KM curves of different expression naïve B cells.
D. KM curves of different expression $CD8^+$ T cells.

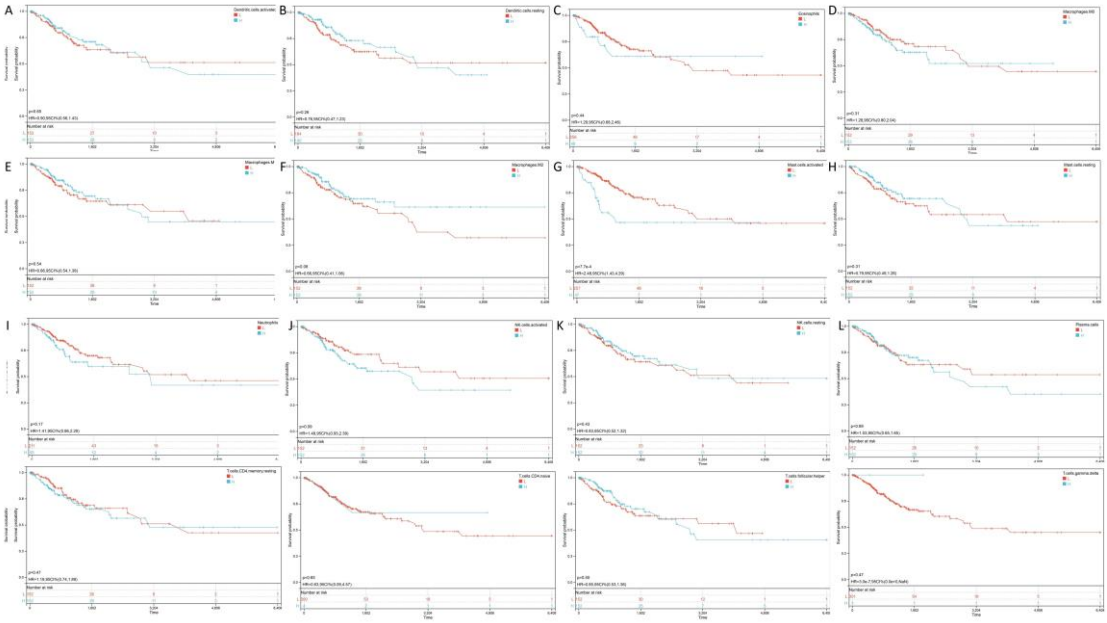


Figure 4 Overall survival analysis with different immune cell types.

A. KM curves of different expression activated dendritic cells.
B. KM curves of different expression resting dendritic cells.
C. KM curves of different expression eosinophils cells.
D. KM curves of different expression macrophages M0 cells.
E. KM curves of different expression macrophages M1 cells.
F. KM curves of different expression macrophages M2 cells.
G. KM curves of different expression activated mast cells.
H. KM curves of different expression resting mast cells.
I. KM curves of different expression neutrophils cells.
J. KM curves of different expression activated NK cells.
K. KM curves of different expression resting NK cells.
L. KM curves of different expression plasma cells.
M. KM curves of different expression resting $CD4^+$ T memory cells.
N. KM curves of different expression naïve $CD4^+$ T cells.
O. KM curves of different expression follicular helper T cells.
P. KM curves of different expression gamma delta T cells.

3.1.3. Identification of immune-associated genes in HPV-pos. and HPV-neg. CESC DEGs

To identify genes associated with naïve B cells and CD8+ T cells, a Weighted Gene Co-expression Network Analysis (WGCNA) was conducted. Initially, in the limma analysis, 6,317 Differentially Expressed Genes (DEGs) were identified in HPV-positive and HPV-negative CESC using the criteria of $|\log2FC| > 1.2$ and $p < 0.05$ (Fig 5A), with red spots indicating up-regulated genes and green spots indicating down-regulated genes (in HPV-positive CESC compared to HPV-negative CESC). Subsequently, a WGCNA analysis was performed using a matrix comprising the identified DEGs (Fig 5B). This analysis revealed a total of 11 immune-associated modules within the scale-free WGCNA co-expression networks (Fig 5C). In the upper left of the squares, the color denoted the correlation value with the module and the immune cell type, with red representing a positive correlation and green indicating a negative correlation. The color in the bottom right of the squares represented the p-value. Notably, the black module displayed a significant positive association with the naïve B cell score (r=0.37, p=1.8e-11), while the red module exhibited the highest correlation with the CD8+ T cell score (r=0.46, p=2.8e-17). These findings suggested that the key components of the black and red modules corresponded to CD8+ T cells and naïve B cells, respectively. Consequently, 249 genes meeting the criteria of MM>0.8 and GS>0.1 were extracted from these two modules.

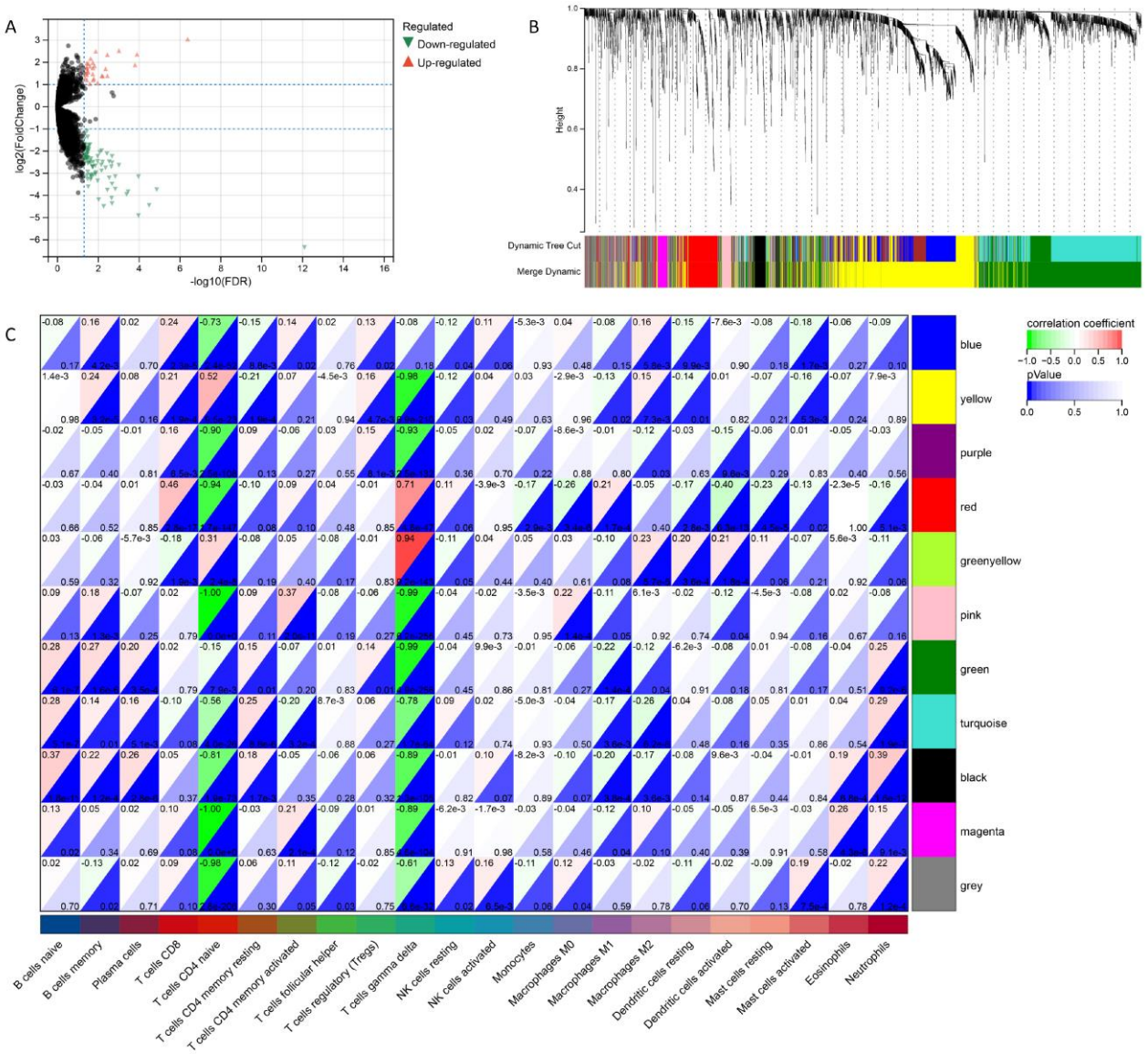


Figure 5 WGCNA analysis for CESC DEGs

A. Volcano plot of CESC DEGs.
B. Clustering results of minded modules in DEGs dataset.
C. The correlation between different modules and the proportions of 22 immune cells.

3.1.4. CESC survival risk model based on immune-associated genes

The 249 genes were designated as hub genes due to their representation of CD8+ T cells and naïve B cells, as well as their differential expression in HPV-positive and HPV-negative CESC. Given that these hub genes exhibited altered expression in response to HPV infection, no distinction was made based on HPV infection status in the development of the prognostic model. Initially, a univariate Cox proportional hazard model was employed to assess the hub genes. Out of these, 200 genes showed significant associations with overall survival in CESC ($p < 0.05$), with 39 genes considered to increase hazard (Hazard Ratio > 1), while the remaining 161 genes were identified as reducing hazard (Hazard Ratio < 1) (Fig 6A). To address multicollinearity issues and reduce the number of genes in the risk model, LASSO regression was applied (Fig 6B). Following LASSO-Cox regression, an internal 10-fold cross-validation was conducted to optimize the model (Fig 6C), resulting in an optimal lambda value of 0.0285755570440733. Nine genes (IKZF3, APOBEC3H, JAK3, CLECL1, FOXP3, CD6, CLEC2D, LINC00158, PILRA) were selected to construct the CESC survival risk model. The final 9-gene survival signature is as follows:

RiskScore = -0.110045311944626 * IKZF3 - 0.180200464406653 * APOBEC3H - 0.009738330408125074 * JAK3 - 0.100574502555081 * CLECL1 - 0.0109656212011594 * FOXP3 - 0.00569462037861635 * CD6 - 0.00590561916507251 * CLEC2D - 0.279175668959309 * LINC00158 - 0.00677990623712316 * PILRA

Since all nine genes in the CESC survival risk model have negative coefficients, they are considered favorable; lower expression levels of these genes result in higher RiskScores (HR = 2.71). CESC bulk RNA-seq samples were classified into high-risk and low-risk groups based on median standardization of RiskScore (Fig 6D). Notably, significant differences between these two groups were observed ($p = 4.0e-5$) in the KM curves. Furthermore, CESC bulk RNA-seq samples were ranked by calculated RiskScore from lowest to highest (Fig 6E), revealing that lower RiskScores were associated with decreased mortality and increased survival probabilities in patients with elevated expression of IKZF3, APOBEC3H, JAK3, CLECL1, FOXP3, CD6, CLEC2D, LINC00158, and PILRA. Additionally, receiver operating characteristic (ROC) curves confirmed the sensitivity and specificity of the overall survival model (Fig 6F), with AUC values of 0.74, 0.68, and 0.71 for the 1-, 3-, and 5-year prognoses, respectively, substantiating the model's accuracy in predicting the overall survival of CESC patients.

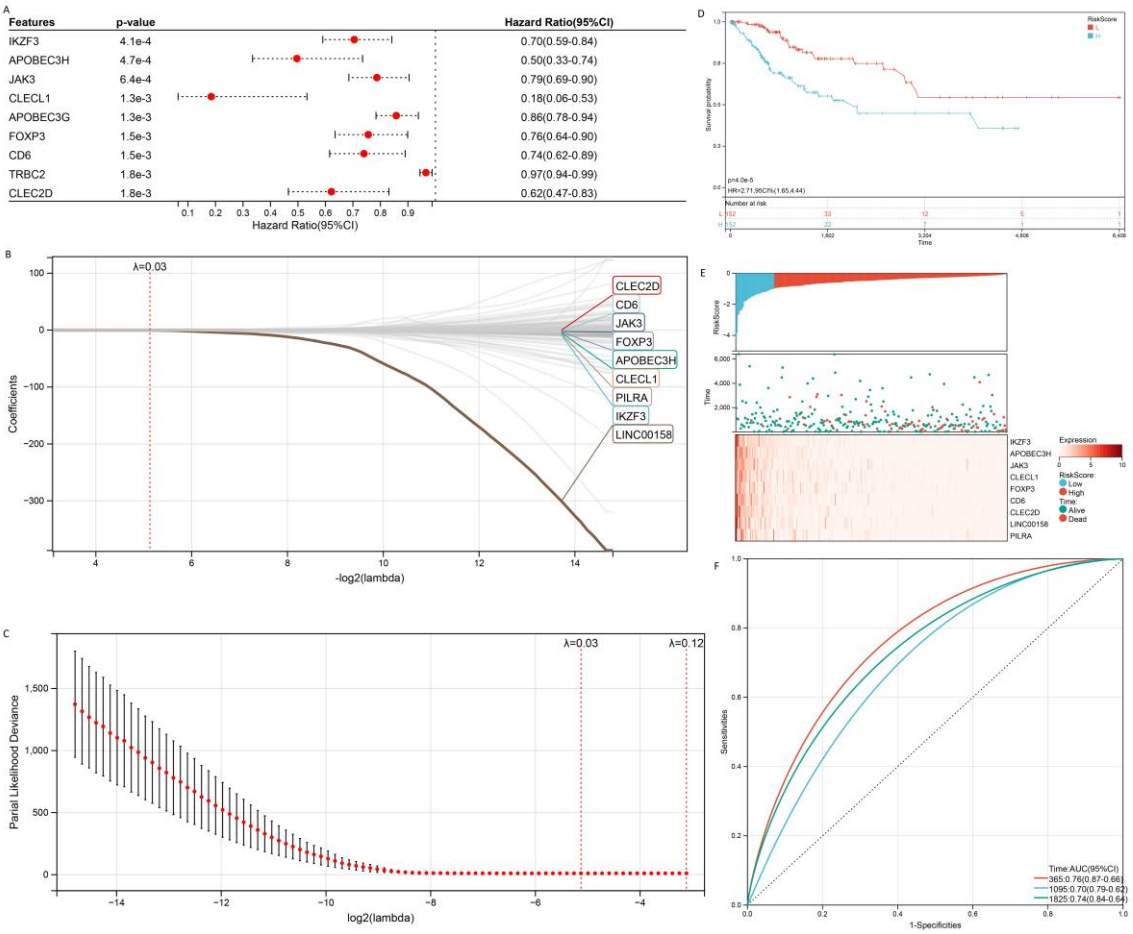


Figure 6 Construction of the Cox and LASSO Cox OS model

A. Top nine survival-related genes in univariate Cox model.
B. The trajectory of each independent variable in the survival risk model: The independent variable lambda's log2 value is shown on the horizontal axis, while its coefficient is shown on the vertical axis.
C. The confidence interval for each lambda in the survival risk model.
D. KM curves of different RiskScore in the survival risk model.
E. RiskScore and survival time with survival status profile and expression levels of the 9 gene signatures.
F. ROC curve for the survival risk model.

3.1.5. CESC relapse risk model based on immune-associated genes

The process for creating a relapse model followed a similar procedure as before, but specifically focused on relapse data. In the univariate Cox model, 139 genes were significantly associated with relapse (log-rank test $p < 0.05$, Figure 7A), including 54 genes with increased hazard ($HR > 1$) and 85 genes with reduced hazard ($HR < 1$). The optimal lambda value for the relapse model was found to be 0.0283647620242562. For the relapse risk model, 7 genes (ARHGAP30, DOK2, ICOS, JAK3, FOXP3, IKZF3, HEATR9) were selected (Fig 7B, C), resulting in the following 7-gene relapse signature:

RiskScore = -0.0210123392211442 * ARHGAP30 - 0.0139776933948472 * DOK2 - 0.0654058831289185 * ICOS - 0.00439045899467425 * JAK3 - 0.0303215128344747 * FOXP3 - 0.0203955365226131 * IKZF3 - 1.57542330970538 * HEATR9

Similar to the survival model, all 7 genes in the CESC relapse risk model had negative coefficients, classifying them as favorable (HR = 2.06). Notably, the low RiskScore group exhibited a significantly lower relapse probability compared to the high RiskScore group (Fig 7D). Down-regulation of these genes was identified as a potential risk factor for relapse in patients with decreased expression of ARHGAP30, DOK2, ICOS, JAK3, FOXP3, IKZF3, and HEATR9 (Fig 7E). The ROC curves demonstrated AUC values of 0.76 (1-year relapse), 0.70 (3-year relapse), and 0.74 (5-year relapse), affirming the model's impressive predictive ability for relapse in CESC patients (Fig 7F).

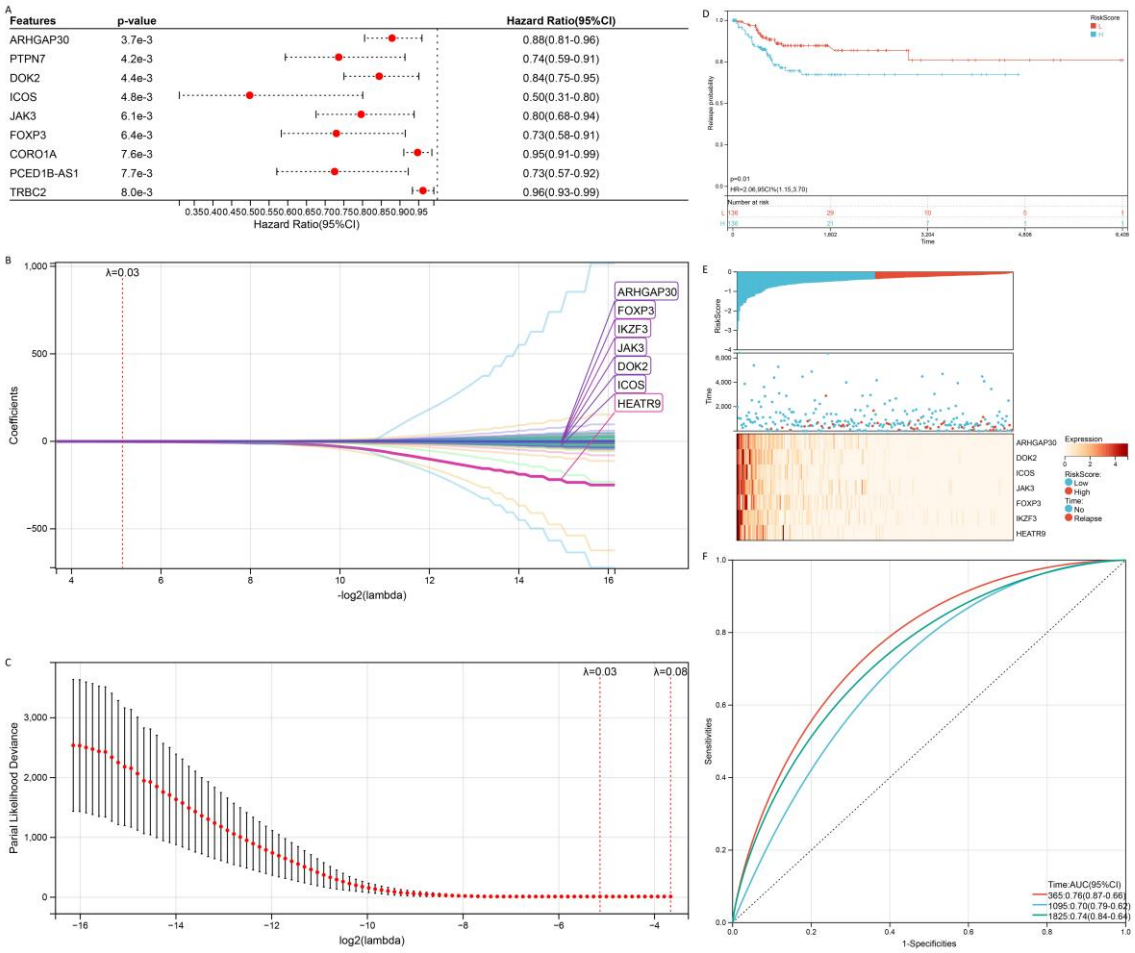


Figure 7 Construction of the Cox and LASSO Cox relapse model

A. Top 9 relapse-related genes in univariate Cox model.

B. The trajectory of each independent variable in the survival risk model: The in-dependent variable lambda's log2 value is shown on the horizontal axis, while its coefficient is shown on the vertical axis.

C. The confidence interval for each lambda in the relapse risk model.

D. KM curves of different RiskScore in the relapse risk model.

E. RiskScore and relapse time with relapse status profile and expression levels of the 7 gene signatures.

F. ROC curve for the relapse risk model.

3.1.6. Functional Enrichment Analysis

GO functional and KEGG pathway analyses were conducted to explore the non-immunological functions of the hub genes, which exhibited the strongest correlation with CD8+ T cells or B cells in the WGCNA. The results indicated that genes essential for both the survival risk model and the relapse risk model were primarily enriched in immunological functions, while their impact on non-immunological processes appeared negligible (Fig 8A, B, C). However, it's worth noting that no significant hits were identified in the KEGG pathway analysis due to the limited number of genes in the relapse risk model.

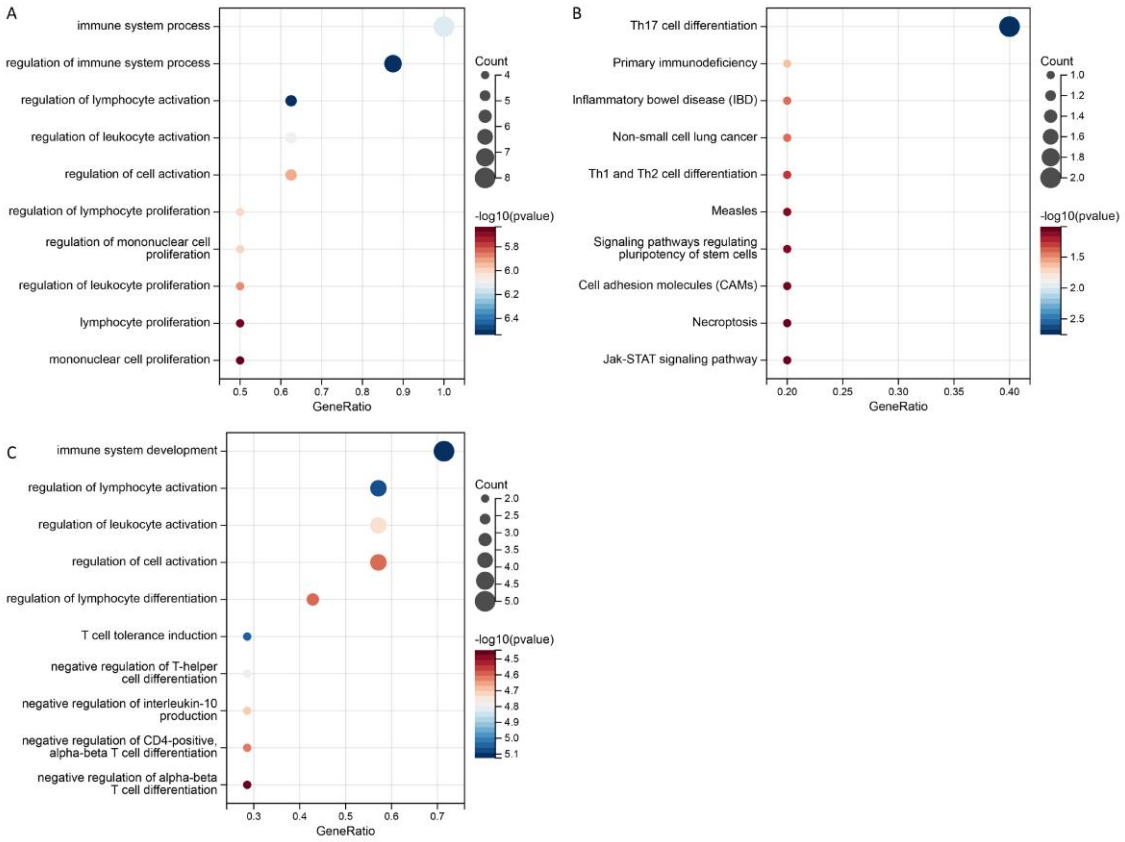


Figure 8 Gene set enrichment analysis of the hub genes.

A. GO enrichment in the survival risk model.
B. KEGG pathways in survival risk model.
C. GO enrichment in the relapse risk model.

3.1.7. HPV-Associated Pan Cancer Analysis

The shared presence of IKZF3, FOXP3, and JAK3 in both the survival and relapse risk models suggests their potential significance in cancerogenesis. To assess whether these

genes play similar roles in other HPV-associated cancers, a pan-cancer analysis was conducted (Fig 9A). Among the selected HPV-associated cancers, including bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), prostate adenocarcinoma (PRAD), and head and neck squamous cell carcinoma (HNSCC), IKZF3 showed a favorable impact on BLCA (Fig 9B) and HNSCC (Fig 9C) survival. However, its expression had no significant effect on the survival probability of patients with COAD (Fig 9D) and PRAD (Fig 9E). On the other hand, FOXP3 expression did not significantly influence the survival of BLCA (Fig 9F), PRAD (Fig 9G), and COAD (Fig 9H) patients. Interestingly, higher FOXP3 expression was associated with improved survival rates in HNSCC (Fig 9I). Elevated JAK3 expression in HNSCC patients correlated with longer survival (Fig 9J), while JAK3 expression levels in COAD (Fig 9 K), PRAD (Fig L), and BLCA (Fig 9M) had no significant impact.

These hub genes exhibit a similar trend in the scRNA-seq analysis of HPV-positive (Fig 9N) and HPV-negative (Fig 9O) HNSCC as they do in HPV-positive and HPV-negative CESC. Furthermore, quantifying the expression profiles of these genes in HNSCC bulk RNA-seq data reveals noticeable differences (Fig 9P).

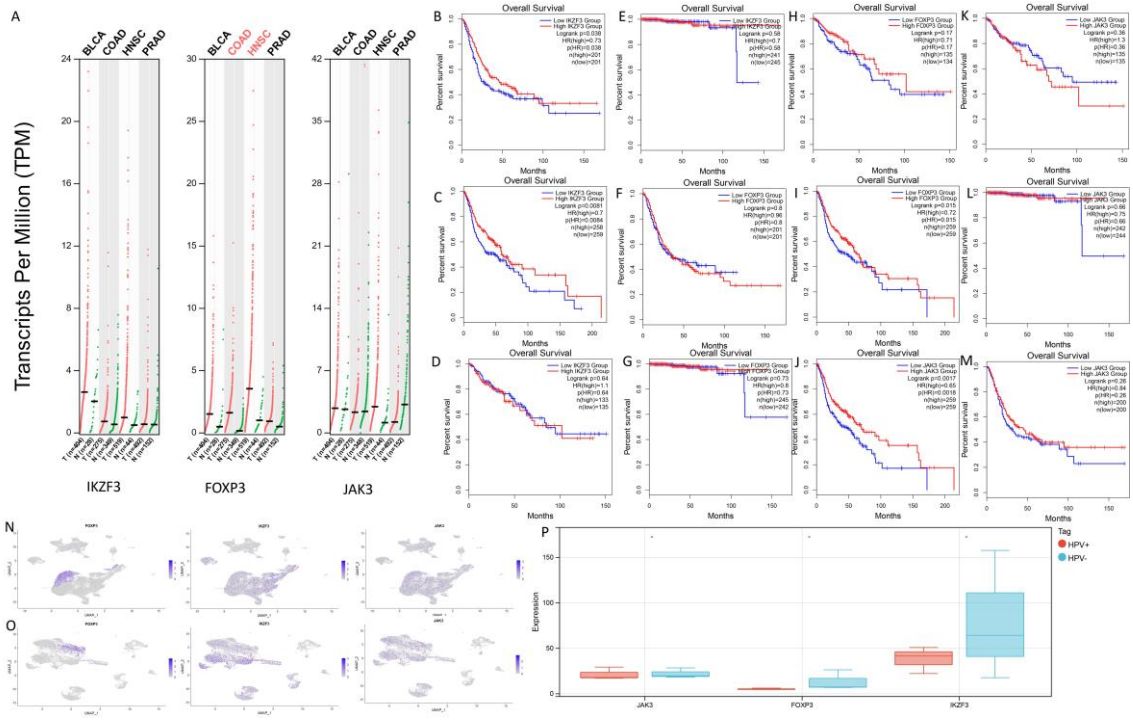


Figure 9 Expression levels and survival analysis in HPV-associated cancers.

A. Expression levels of IKZF3, FOXP3, and JAK3 in BLCA, HNSC, PRAD, and COAD.
B. Survival analysis in BLCA comparing the high- and low-IKZF3 groups.
C. Survival analysis in HNSCC comparing the high- and low-IKZF3 groups.
D. Survival analysis in COAD comparing the high- and low-IKZF3 groups.

E. Survival analysis in PRAD comparing the high- and low-IKZF3 groups.
F. Survival analysis in BLCA comparing the high- and low-FOXP3 groups.
G. Survival analysis in PRAD comparing the high- and low-FOXP3 groups.
H. Survival analysis in COAD comparing the high- and low-FOXP3 groups.
I. Survival analysis in HNSCC comparing the high- and low-FOXP3 groups.
J. Survival analysis in HNSCC comparing the high- and low-JAK3 groups.
K. Survival analysis in COAD comparing the high- and low-JAK3 groups.
L. Survival analysis in PRAD comparing the high- and low-JAK3 groups.
M. Survival analysis in BLCA comparing the high- and low-JAK3 groups.
N. Survival analysis in HNSCC comparing the high- and low-JAK3 groups.
O. Expressions of FOXP3, IKZF3, JAK3 in scRNA-seq HPV-neg. HNSCC.
P. Expressions of FOXP3, IKZF3, JAK3 in scRNA-seq HPV-pos. HNSCC.
Q. Quantification of FOXP3, IKZF3, JAK3 in bulk RNA-seq HPV-pos. and HPV-neg. HNSCC.

The analysis of HPV-positive and HPV-negative CESC samples has unveiled significant heterogeneities between these two groups. Moreover, the survival and relapse models have highlighted a consistent protective role for genes with altered expression due to HPV infection in various HPV-associated cancers. Given the limited number of HPV-negative CESC samples, the subsequent steps of this study were not confined to CESC alone. Various HPV types, including those other than HPV16 and HPV18, were included in order to obtain a comprehensive understanding of the impact of HPV infection on these diseases.

3.2. Transcriptome changes in skin low-risk HPV infection

3.2.1. DEGs and immune infiltration analysis in low-risk HPV infection

The genetic alterations induced by low-risk HPV infection were explored using samples from warts, condyloma acuminata, treeman syndrome, and normal skin. In comparison to normal skin samples, wart samples exhibited up-regulation of 9,333 genes and down-regulation of 2,416 genes (Fig 10A), while treeman syndrome samples showed up-regulation of 2,710 genes and down-regulation of 2,212 genes (Fig 10B). In condyloma acuminata samples, 172 genes were up-regulated, and 114 genes were down-regulated when compared to normal skin samples (Fig 10C). Subsequently, CIBERSORT analysis was performed to characterize the impact of low-risk HPV types on immune infiltration. In wart samples, there was an increase in macrophage M0 cells, neutrophils, activated

dendritic cells, plasma cells, activated mast cells, and resting NK cells, along with a decrease in resting mast cells, resting dendritic cells, activated NK cells, macrophage M1 cells, and macrophage M2 cells when compared to normal skin (Fig 10D). In treeman syndrome samples, an increase in macrophage M2 cells and a decrease in activated dendritic cells were observed (Fig 10E). Condyloma acuminata samples showed a decrease in neutrophils and CD8+ T cells (Fig 10F). These differences were compiled into an ensemble for further analyses.

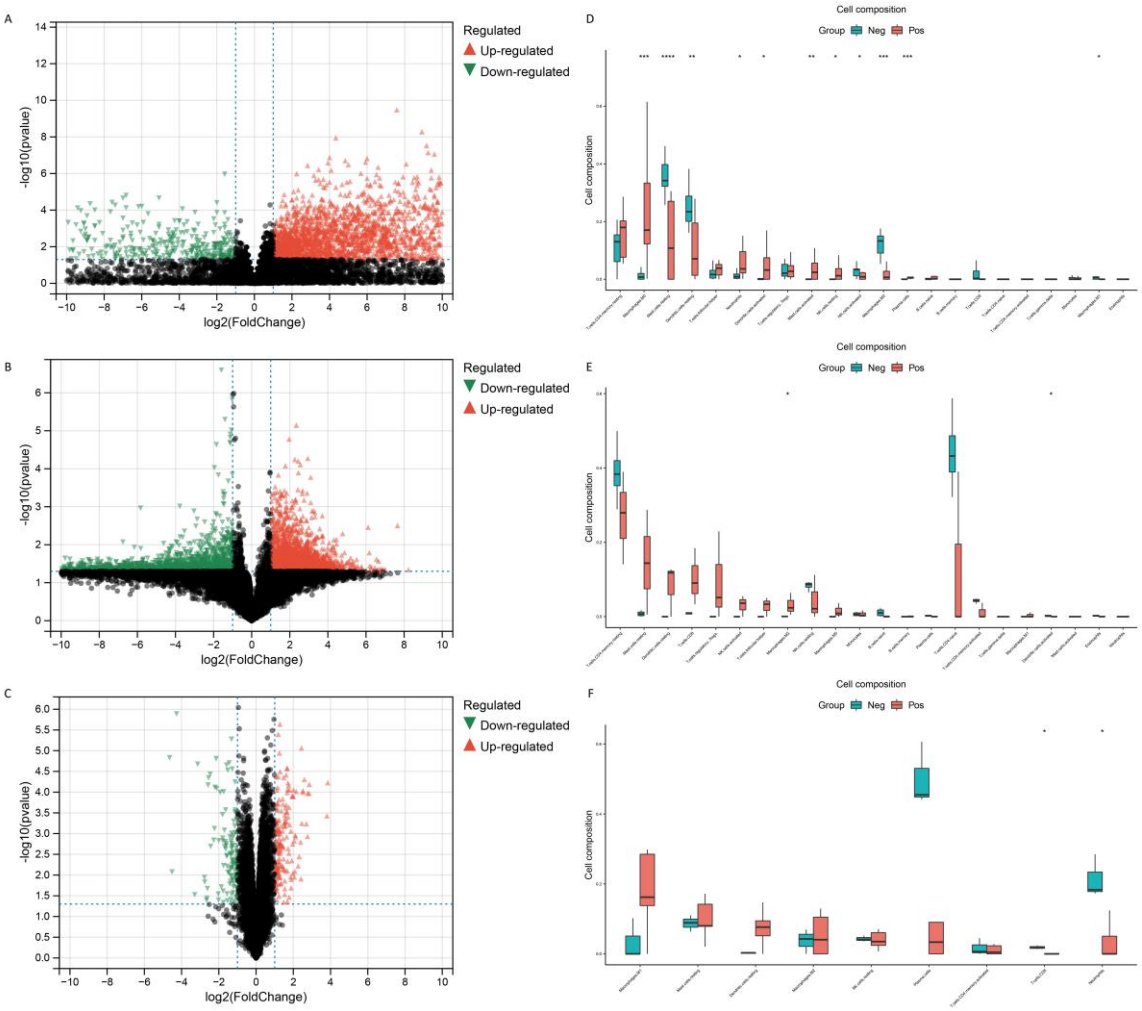


Figure 10 DEGs and different immune infiltration of low-risk HPV infection

A. Volcano plot of DEGs in wart samples vs normal samples.
B. Volcano plot of DEGs in treeman sample vs normal samples.
C. Volcano plot of DEGs in condyloma acuminata samples vs normal samples.
D. Different immune infiltration in wart samples and normal samples. Asterisks represent statistical significance ((*p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001)

E. Different immune infiltration in treeman sample and normal samples.
F. Different immune infiltration in condyloma acuminata samples and normal samples.

3.2.2. Identification of hub genes in low-risk HPV infection

To address the issue of an excessive number of genes in the ensemble that cannot be effectively used for subsequent analysis, a Weighted Gene Co-expression Network Analysis (WGCNA) was performed to reduce the dimensionality of the Differentially Expressed Genes (DEGs) [35]. In this process, the 22 immune cell types were grouped into 11 major leukocyte types to create more manageable modules. A total of 23 immune-associated modules were identified within the WGCNA networks (Fig 11).

Specifically, the following modules were selected for further analysis based on their associations with specific immune cell scores: the purple module (associated with macrophage cell scores, r=0.28, p=3.2e-9), the ivory module (associated with neutrophils, r=0.48, p=3.2e-27), the brown module (associated with CD8+ T cells, r=0.16, p=1.0e-3), the green module (associated with dendritic cells, r=0.38, p=1.2e-16), and the royal blue module (associated with mast cells, r=0.39, p=8.5e-17). It's worth noting that many genes in these modules exhibited low immune cell specificity but were highly associated with immune marker genes.

Next, the Metascape database was utilized to explore the functional roles of genes with low immune cell specificity. Immune-associated pathways were excluded from the analysis [60] (Fig 11C). The enriched pathways primarily included human diseases, human papillomavirus infection, metabolic pathways, and pathways in cancer. Additionally, pathways related to ferroptosis, oxidative stress, and necroptosis were enriched. These pathways have gained recognition in recent years for their close association with HPV infection and tumorigenesis [60]. Therefore, differentially expressed genes (DEGs) related to ferroptosis, oxidative stress, and necroptosis were also included in the subsequent analyses.

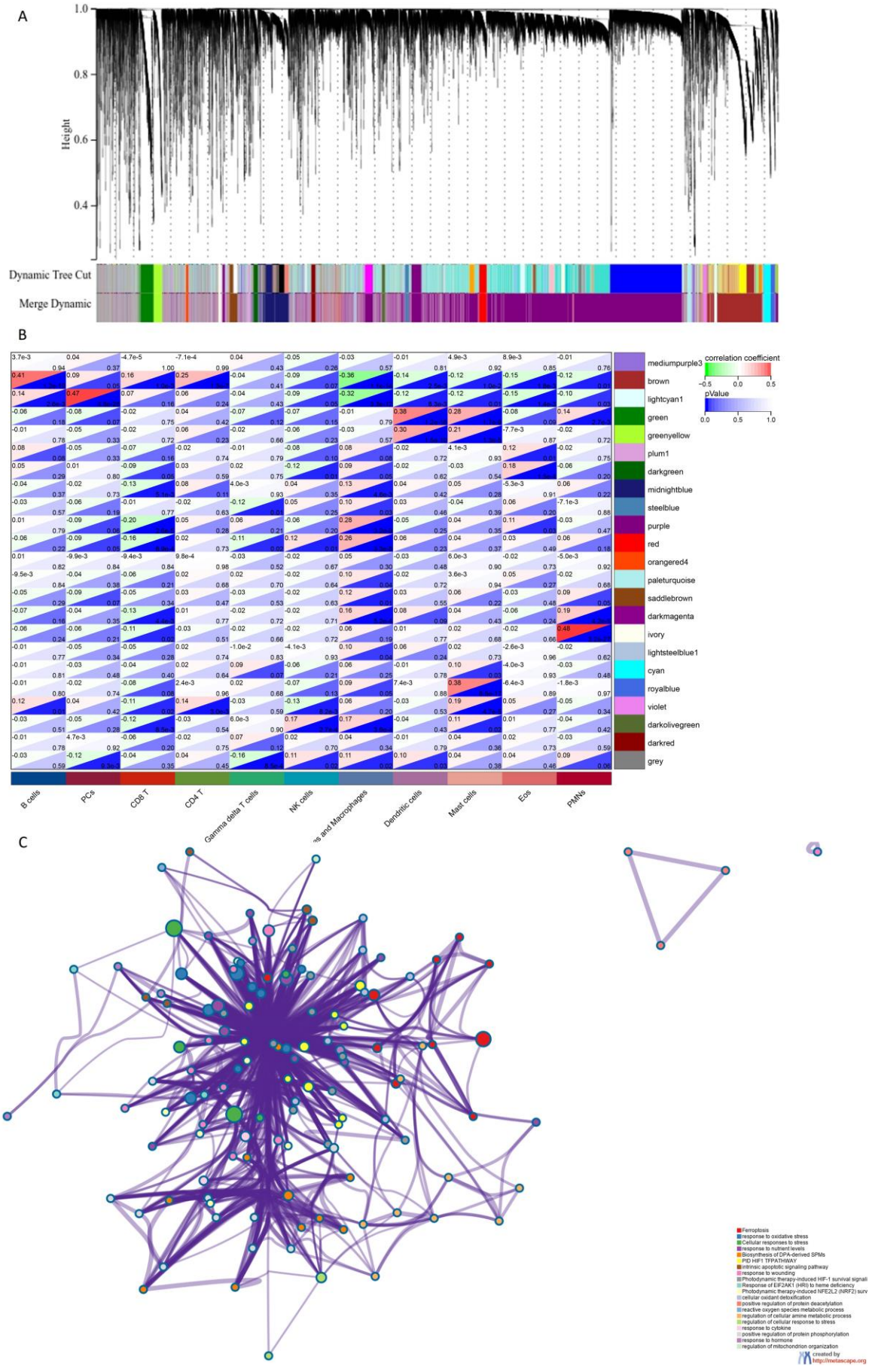

A
Height
1.0
0.8
0.6
0.4
Dynamic Tree Cut
Merge Dynamic
B
mediumpurple3
brown
lightcyan1
green
greenyellow
plum1
darkgreen
midnightblue
steelblue
purple
red
orangered4
paleturquoise
saddlebrown
darkmagenta
ivory
lightsteelblue1
cyan
royalblue
violet
darkolivegreen
darkred
grey
correlation coefficient
pValue
B cells
PCs
CD8 T
CD4 T
Gamma delta T cells
NK cells
Dendritic cells
Mast cells
Eos
PMNs
C
Ferroptosis
response to oxidative stress
Cellular responses to stress
response to nutrient levels
Biosynthesis of DPA-derived SPMs
PID HIF1 TFPATHWAY
intrinsic apoptotic signaling pathway
response to wounding
Photodynamic therapy-induced HIF-1 survival signali
Response of EIF2AK1 (HRI) to heme deficiency
Photodynamic therapy-induced NFE2L2 (NRF2) surv
cellular oxidant detoxification
positive regulation of protein deacetylation
reactive oxygen species metabolic process
regulation of cellular amine metabolic process
regulation of cellular response to stress
response to cytokine
positive regulation of protein phosphorylation
response to hormone
regulation of mitochondrion organization
created by
http://metascape.org

Figure 11 WGCNA analysis for low-risk HPV infection DEGs

A. Clustering results of minded modules in DEGs dataset.
B. The correlation between different modules and the proportions of 11 immune cells.
C. Metascape analysis for low immune cell specificity DEGs.

3.2.3. Association of low-risk HPV-altered genes with skin cancer

In the analysis of benign skin lesion samples with low-risk HPV infection, the enrichment of pathways related to cancer suggests that gene alterations in these samples might be associated with the development of skin cancer. To further investigate these genes, immune-associated genes were identified from the WGCNA analysis. Additionally, genes associated with ferroptosis, oxidative stress, and necroptosis were obtained by mapping differentially expressed genes (DEGs) with the FerrDb and Gene-Cards databases.

The relationship between these genes with different specificities and overall survival in skin cancer was analyzed separately using univariate Cox proportional hazard models. To address multicollinearity, LASSO-cox regression was performed. The model was internally validated through 10-fold cross-validation, and external validation was carried out using data from the ICGC database.

The optimal lambda value for the model based on immune-associated genes was determined to be 0.103105993256383 (Fig 12). This model included 18 genes: ZNF703, CLN6, HMOX2, LINC01138, AC099343.4, SPATA13, UBE2O, TTC39C, PKIG, MYO10, ACSL4, FOXK1, HEXD, STAMBPL1, AP003721.4, CREG1, HSPA7, and AC012213.1. The final 18-gene relapse signature is represented by the following RiskScore formula:

RiskScore = 0.000447096284873543 * ZNF703 + 0.000496930564037156 * CLN6 + 0.00198313033500234 * HMOX2 - 0.0140447939837885 * LINC01138 - 0.0227729288000317 * AC099343.4 - 0.00166467347841079 * SPATA13 + 0.000775173016576975 * UBE2O - 0.00545438753998658 * TTC39C - 0.000169884439727878 * PKIG + 2.08968514258868e-05 * MYO10 - 0.00075962394768721 * ACSL4 + 0.000530366943891432 * FOXK1 - 0.00356655757176417 * HEXD - 0.000171228577629836 * STAMBPL1 - 0.0034310030038542 * AP003721.4 - 0.000169331851861702 * CREG1 - 0.000145982841095522 * HSPA7 - 0.0419117546482515 * AC012213.1

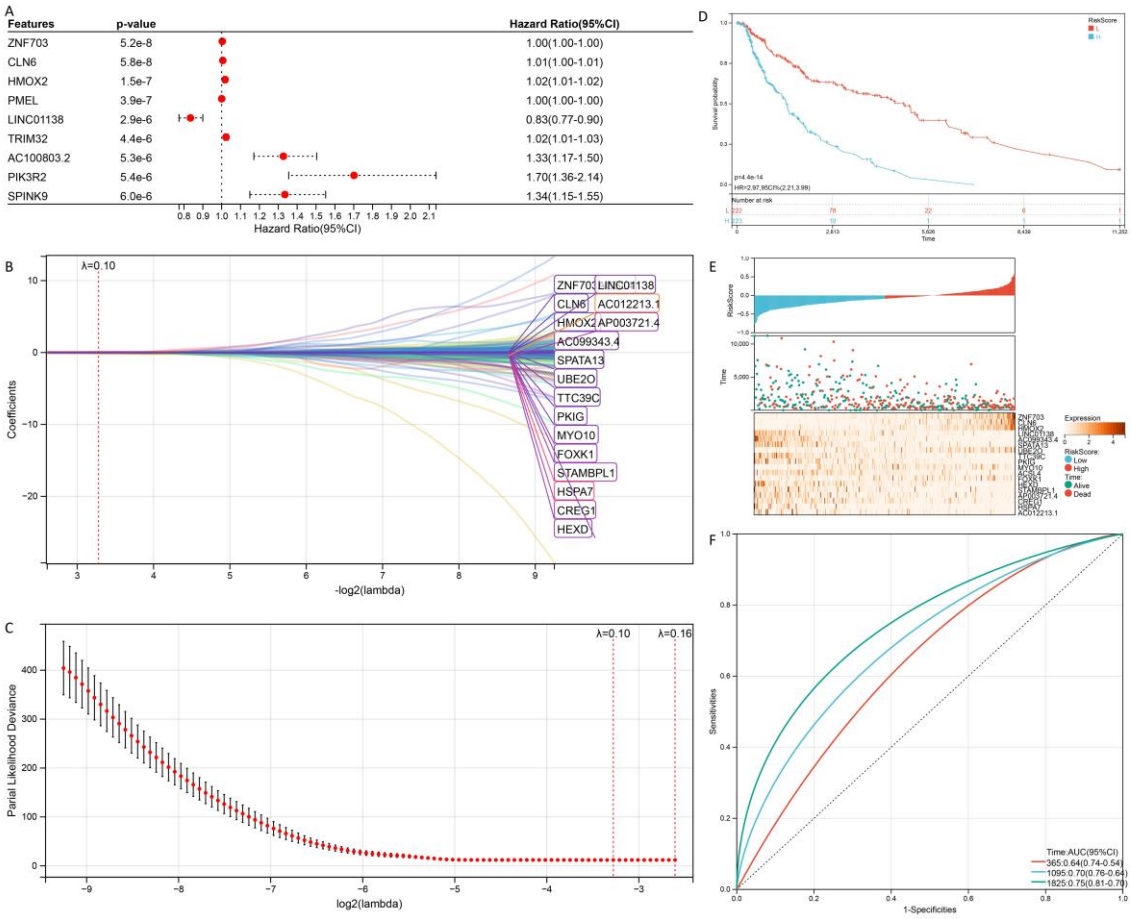


Figure 12 Construction of the immune-associated Cox and LASSO Cox model

A. Top 9 immune-associated genes in univariate Cox model.
B. The trajectory of each independent variable in the immune-associated survival risk model.
C. The confidence interval for each lambda in the immune-associated survival risk model.
D. KM curves of different RiskScore in the immune-associated survival risk model.
E. RiskScore and survival time with survival status profile and expression levels of the 18 gene signatures.
F. ROC curve for the immune-associated risk model.

In the model based on ferroptosis-associated genes, the optimal lambda value was determined to be 0.016512977172364 (Fig 13). This model included 18 genes: NCF2, ANGPTL7, CHAC1, SLC7A11, JDP2, VEGFA, HSD17B11, HMOX1, FTL, DRD4, MAP3K5, SLC2A1, ALOX5, ALOX15, HBA1, NNMT, HIC1, and AURKA. The final 18-gene relapse signature is represented by the following RiskScore formula:

RiskScore = -0.00938989420423927 * NCF2 - 0.000472378323572829 * ANGPTL7 + 0.00259420769181918 * CHAC1 + 0.00269077908427012 * SLC7A11 - 0.00200395718658391 * JDP2 - 0.000501434252325555 * VEGFA - 0.0010393709159819 * HSD17B11 + 0.000295359464052238 * HMOX1 + 8.16942125872471e-06 * FTL - 0.0344816232911417 * DRD4 - 0.00072296935563573 * MAP3K5 + 0.0011747907951178 * SLC2A1 - 0.00829774914015909 * ALOX5 + 0.00341581793741788 * ALOX15 + 0.0300329250092999 * HBA1 - 0.000737359738304914 * NNMT + 0.0043877495412482 * HIC1 + 0.00177997795502466 * AURKA

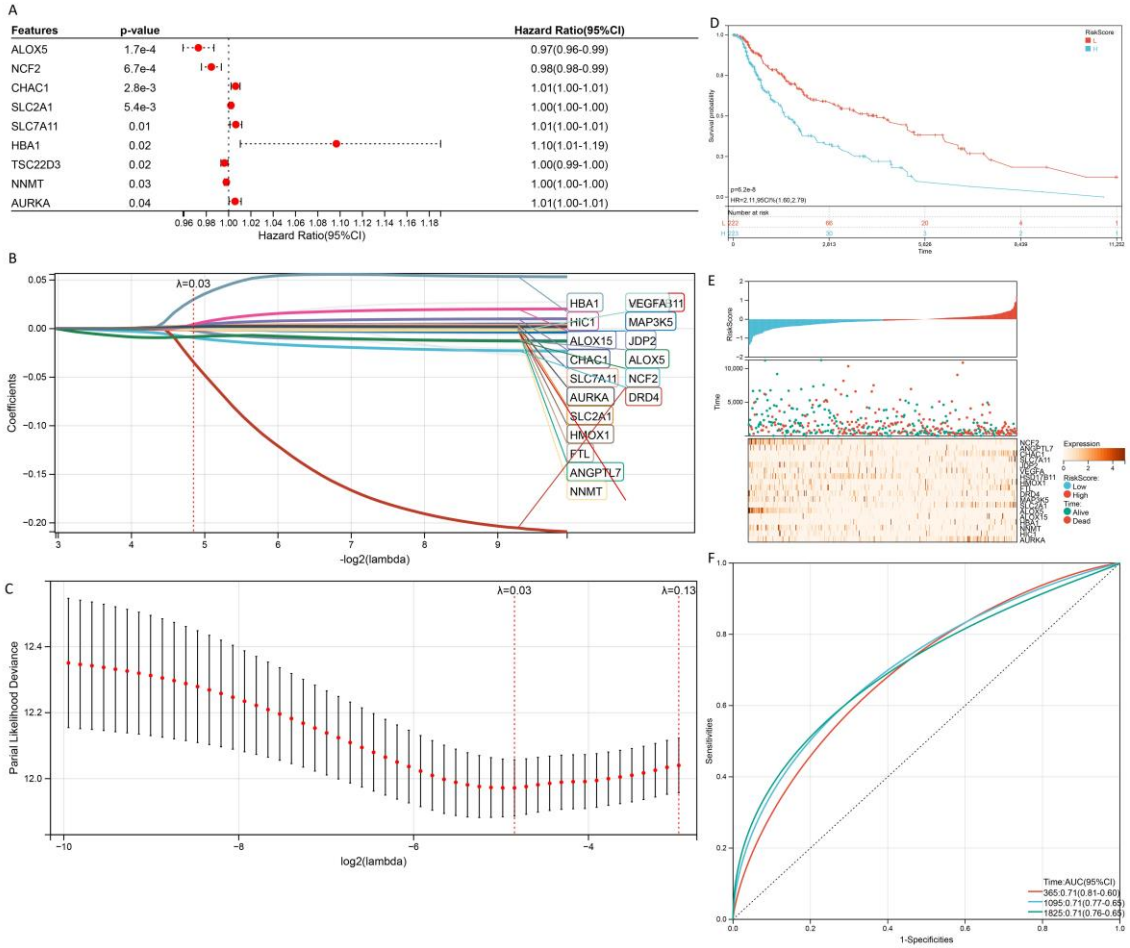


Figure 13 Construction of the ferroptosis-associated Cox and LASSO Cox model

A. Top 9 ferroptosis-associated genes in univariate Cox model.
B. The trajectory of each independent variable in the ferroptosis-associated survival risk model.
C. The confidence interval for each lambda in the ferroptosis-associated survival risk model.
D. KM curves of different RiskScore in the ferroptosis-associated survival risk model.

E. RiskScore and survival time with survival status profile and expression levels of the 18 gene signatures.
F. ROC curve for the ferroptosis-associated risk model.

In the model based on oxidative stress-associated genes, the optimal lambda value was determined to be 0.016512977172364 (Fig 14). This model included 43 genes: HRNR, UGT1A7, SPRR2B, KIT, TGM3, HMOX2, PITX1, PSMB8, STAT1, PLAAT4, RABIF, HLA-DQB1, PAEP, KPNA2, ICAM1, KCNN4, CLIC2, HEXD, IRF7, TM6SF2, MZB1, MEAK7, ADORA2A, FCER1G, CAP2, S100A11, SNX22, RTN1, KCNMB1, NEFM, DOCK9, MED12L, NOX4, IFI27, FGF13, WNT1, CD33, IQGAP2, UNC13A, PSCA, ABCB4, WWTR1, and PCSK1. The final 43-gene relapse signature is represented by the following RiskScore formula:

RiskScore = 0.0159716926846922 * HRNR + 0.108696999312889 * UGT1A7 + 0.00100154577184382 * SPRR2B + 0.000281405166826053 * KIT + 0.000614857020909774 * TGM3 + 0.00186458220852995 * HMOX2 + 0.00358982603802565 * PITX1 - 0.000262977849183458 * PSMB8 - 0.00024566538943957 * STAT1 - 0.000499686021817832 * PLAAT4 + 0.00513672254515595 * RABIF - 9.65773499678376e-05 * HLA-DQB1 + 1.13712788414208e-05 * PAEP + 0.000888337787254083 * KPNA2 - 0.000572284523885771 * ICAM1 - 0.000720372269384761 * KCNN4 - 0.00271567021722054 * CLIC2 - 0.00855204231985427 * HEXD - 3.69585688325384e-05 * IRF7 + 0.0672400620745375 * TM6SF2 - 0.00011572610524236 * MZB1 + 0.00217099443046481 * MEAK7 - 0.0655107652306773 * ADORA2A - 0.000130763104916954 * FCER1G + 0.00238916558287936 * CAP2 + 8.455153079688e-06 * S100A11 - 0.00080750734003968 * SNX22 - 0.00887381802525652 * RTN1 - 0.0194093598950961 * KCNMB1 + 0.00199095439803593 * NEFM + 0.000396657527955989 * DOCK9 - 0.145183403561008 * MED12L - 0.00221609711218496 * NOX4 - 0.0001609032171123 * IFI27 - 0.01251629634583 * FGF13 - 0.00204177776550964 * WNT1 - 0.00110634263981778 * CD33 - 0.000204348990660977 * IQGAP2 - 0.0015563390828012 * UNC13A - 6.21504474707566e-05 * PSCA - 0.00054863662562471 * ABCB4 - 1.8846670226815e-05 * WWTR1 - 0.000627142494598778 * PCSK1

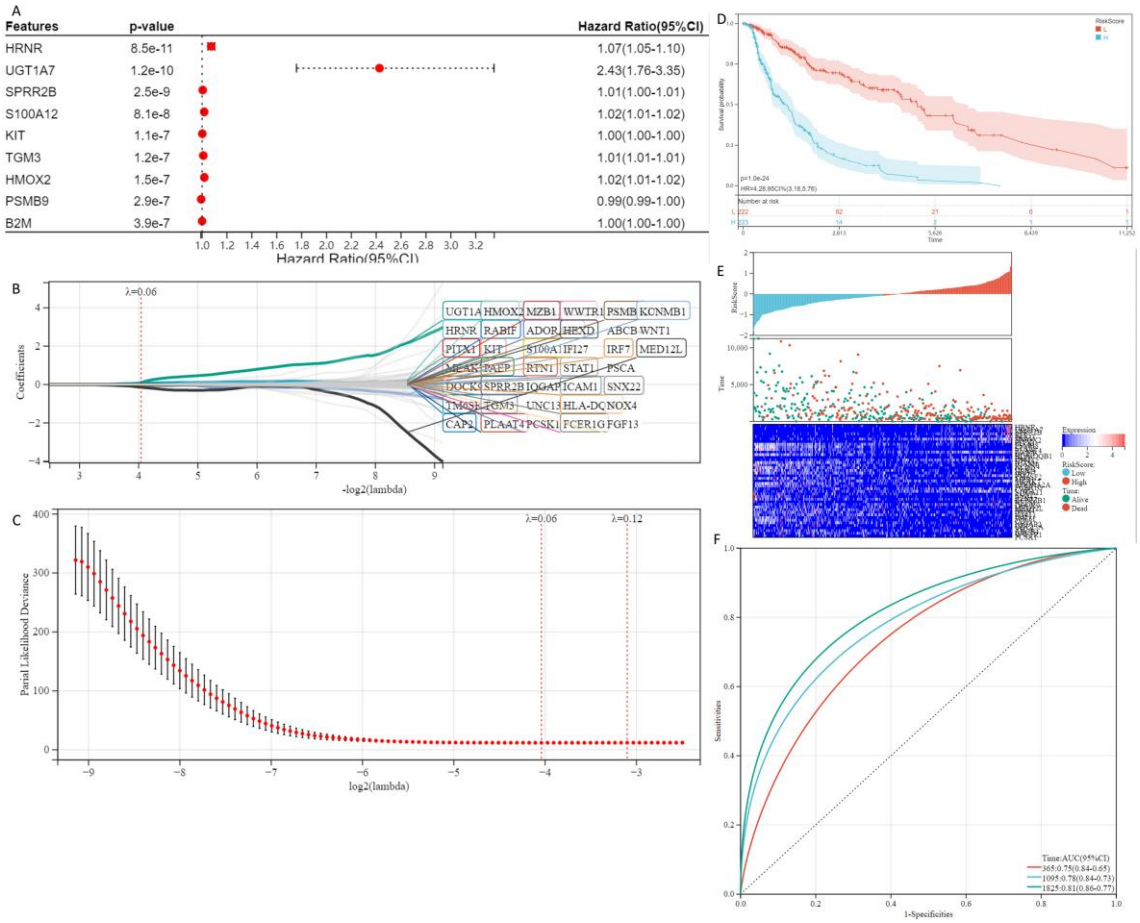


Figure 14 Construction of the oxidative stress-associated Cox and LASSO Cox model

A. Top 9 oxidative stress-associated genes in univariate Cox model.
B. The trajectory of each independent variable in the oxidative stress-associated survival risk model.
C. The confidence interval for each lambda in the oxidative stress-associated survival risk model.
D. KM curves of different RiskScore in the oxidative stress-associated survival risk model.
E. RiskScore and survival time with survival status profile and expression levels of the 43 gene signatures.
F. ROC curve for the oxidative stress-associated risk model.

In the model based on necroptosis-associated genes, the optimal lambda value was determined to be 0.0426492157852511 (Fig 15). This model includes 16 genes: HRNR, CD74, CLEC7A, KRT16, TUBB4A, TUBA1C, GSDMD, NOX4, NME8, KRT19, PELI1, COPS2, ACHE, HSPA4, FLNC, and DAPK1. The final 16-gene relapse signature is represented by the following RiskScore formula:

RiskScore = 0.042540605234936 * HRNR - 4.77265188341723e-05 * CD74 - 0.0132404452836909 * CLEC7A + 8.15839193972035e-06 * KRT16 + 0.0001478075116376 * TUBB4A + 0.00209534884027515 * TUBA1C - 0.00427908155832343 * GSDMD - 0.00914688320536871 * NOX4 - 0.0271025523792678 * NME8 + 0.000978209322398353 * KRT19 - 0.00157133493894347 * PELI1 - 0.00553120824438331 * COPS2 - 0.00254324843510403 * ACHE + 0.000895966855034512 * HSPA4 + 0.000396142731535711 * FLNC + 0.00189832090895527 * DAPK1

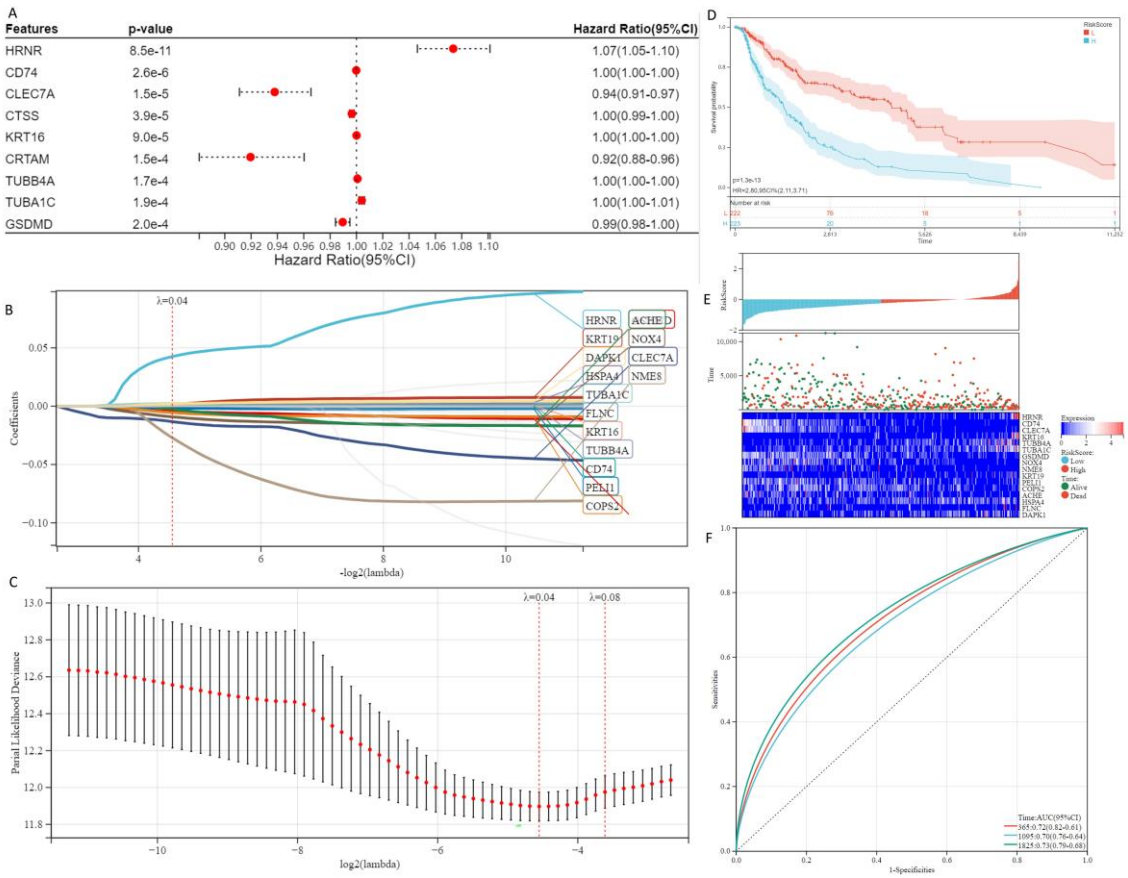


Figure 15 Construction of the necroptosis-associated Cox and LASSO Cox model

A. Top 9 necroptosis-associated genes in univariate Cox model.
B. The trajectory of each independent variable in the necroptosis-associated survival risk model.
C. The confidence interval for each lambda in the necroptosis-associated survival risk model.
D. KM curves of different RiskScore in the necroptosis-associated survival risk model.

E. RiskScore and survival time with survival status profile and expression levels of the 16 gene signatures.
F. ROC curve for the necroptosis-associated risk model.

The internal validation of the survival risk model based on oxidative stress showed the highest AUC values, with 0.75, 0.78, and 0.81 for the 1-, 3-, and 5-year prognoses, respectively. External validation also yielded similar results, with AUC values of 0.67, 0.69, and 0.74 for the 1-, 3-, and 5-year prognoses (Fig 16A). The other modules also exhibited AUC values above 0.5 in external data validation (Fig 16B, C, D), indicating that the genes altered in low-risk HPV infection skin lesions are associated with the progression of skin cancer. These results demonstrate the potential utility of these gene signatures in predicting the prognosis of skin cancer patients.

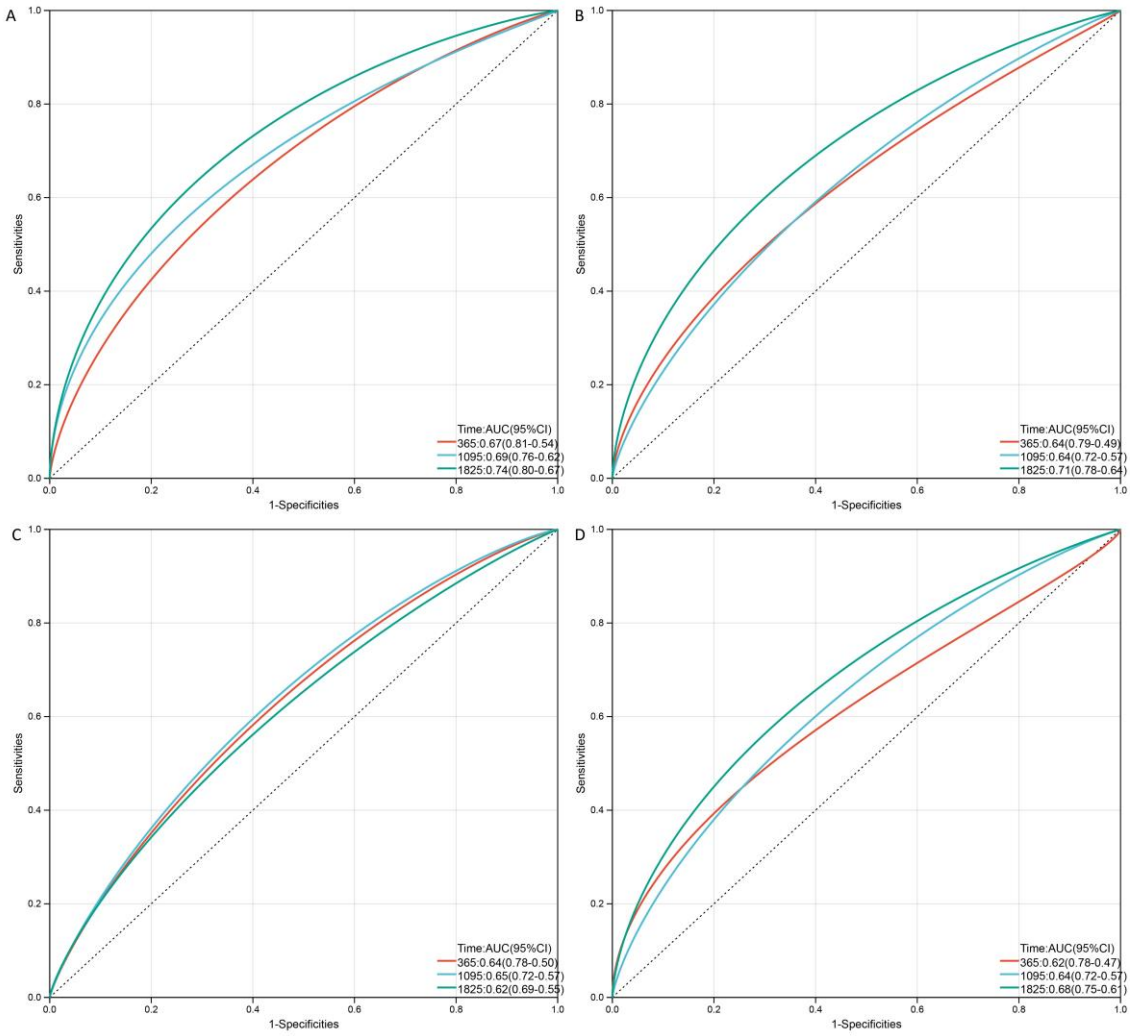


Figure 16 External validation from ICGC database.

A. ROC curve for the oxidative stress-associated risk model.
B. ROC curve for the immune-associated risk model.
C. ROC curve for the ferroptosis-associated risk model.
D. ROC curve for the necroptosis-associated risk model.

3.2.4. Relationship between low-risk HPV-altered genes and cancer therapeutics response

All the models had AUC values greater than 0.5, suggesting that all genes in the models were highly correlated with the prognosis of skin cancer. A protein-protein interaction (PPI) network analysis was performed using all the genes in the survival risk model (Fig 17A), resulting in 86 nodes and 84 edges. Subsequently, the Cytoscape software plug-in MCODE was employed with specific settings (degree cutoff = 2, node score cutoff = 0.2, K-core = 2, and maximum depth = 100) to identify the most significant genes, which included VEGFA, STAT1, HMOX1, ICAM1, HSPA4, KIT, and NOX4 (Fig 17B).

To assess the relationship between clinical features and gene expression levels, CTRP V2 was utilized. Among the identified genes, ICAM1 and KIT were downregulated in the lesions, so drugs with a negative correlation with their gene expression levels were selected. On the other hand, VEGFA, STAT1, HMOX1, HSPA4, and NOX4 were upregulated in the lesions, and drugs were chosen that showed a positive correlation with their gene expression levels. Based on the results of drug sensitivity correlated with gene expression levels, signaling pathways with altered genes were identified (Fig 17C).

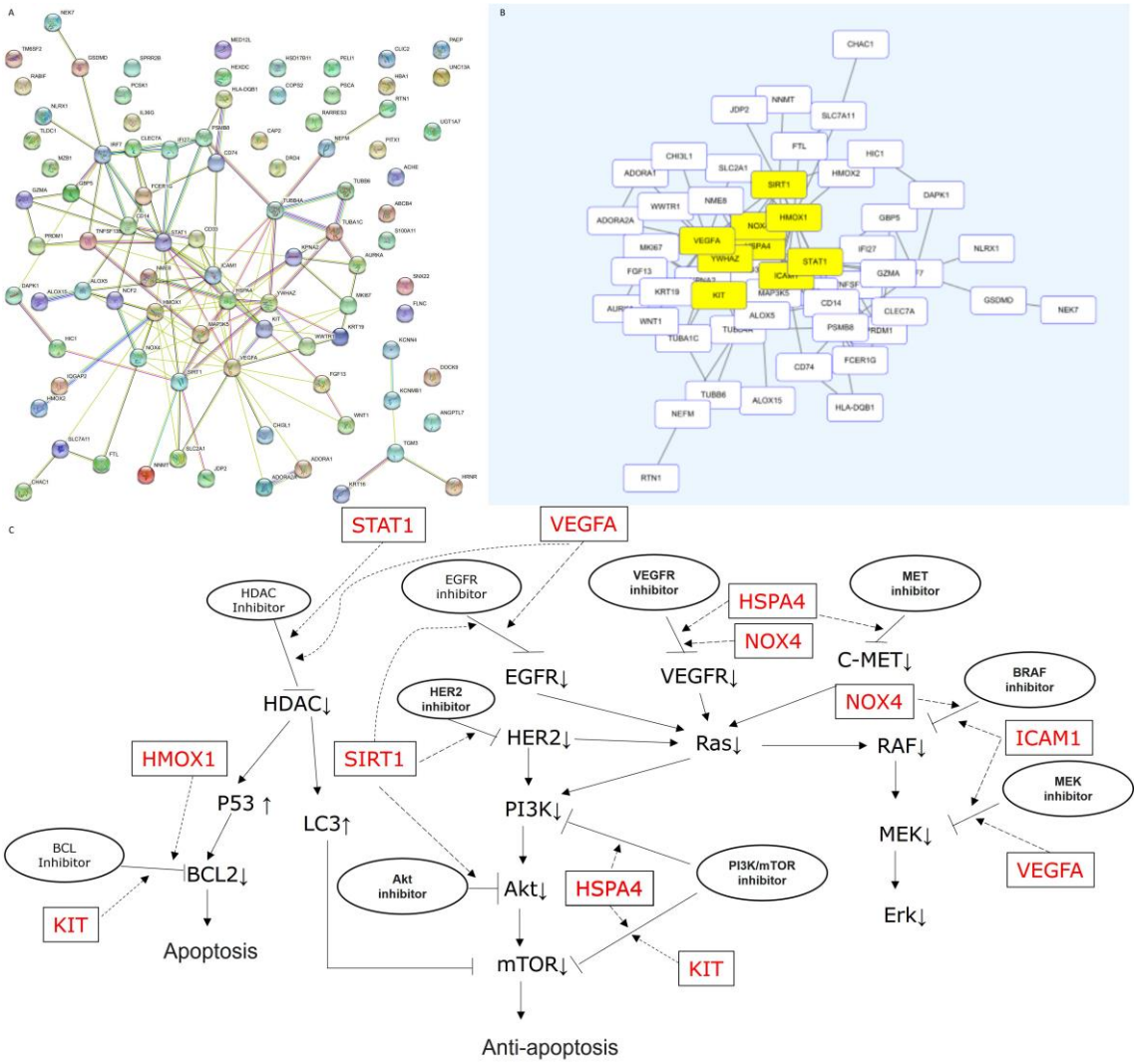


Figure 17 Relationship between the core genes of the PPI network and cancer therapeutics response

A. PPI network of genes in survival risk models.
B. Selection of core genes in cytoscape plug-in MCODE.
C. signaling pathways with core genes.

The study revealed that the expression of core genes was altered in HPV-infected skin lesions and had an impact on the prognosis of skin cancer, suggesting the potential influence of low-risk HPV infection. However, the association between skin cancer and HPV infection was not confirmed, and the observed gene alterations could be attributed to skin tissue lesions. Therefore, further analysis of drug sensitivity related to these genetic alterations was not conducted. The next phase of the study aims to identify genes that are influenced by disease progression and regulated by HPV infection in a type of cancer that is definitively affected by HPV infection.

3.3. Identification of genes associated with carcinogenesis and HPV infection

3.3.1. Identify DMPs between normal, precancerous lesion and cancer samples

The analysis began with 390,065 probes that passed quality control. Principal component analysis (PCA) was conducted to compare the β-values for all samples (Fig 18A). The separation of the normal anal group and normal cervical group indicated meaningful grouping, but AIN3, ASCC, and CESC samples were closely clustered, making it difficult to distinguish between them. This pattern was also evident in the heatmap of the correlation matrix (Fig 18B), where samples from AIN3, ASCC, and CESC could not be clearly separated. Similarly, CESC samples were mixed with ASCC samples in the visualization of sample similarity based on the top 1000 most variable probes. It was observed that there was a significant overlap of differentially methylated probes (DMPs) between AIN3, ASCC, and CESC samples, while CIN3 samples were more similar to normal cervical samples. Consequently, the analysis focused on genes with overlapping DMPs between the different tumor groups, excluding intraepithelial neoplasia groups due to their distinct similarity.

Among the 390,065 probes, 44,137 (11.31%) showed differential methylation levels in CESC and adjacent normal tissues, while 11,440 (2.9%) probes exhibited differences in ASCC compared to normal tissues. In the comparison to the normal cervical group, 25,300 (57.4%) probes were hypermethylated, and 18,837 (42.6%) probes were hypomethylated. In contrast, compared to the normal anal group, 9,726 (85.0%) probes were hypermethylated, and 1,714 (15.0%) probes were hypomethylated in the ASCC group. The distribution of DMPs varied among genomic locations, with the majority enriched in open sea regions (Fig 18C). Furthermore, both hypermethylated and hypomethylated DMPs were predominantly distributed on Chromosome 1 (Fig 18D). A Venn plot was used to quantify the overlapping hypermethylated and hypomethylated DMPs between ASCC and CESC (Fig 18E). The trends in methylation were similar in ASCC and CESC, with only 3 probes exhibiting contrasting methylation patterns. In total, 7,633 DMPs were hypermethylated, while 1,024 DMPs were hypomethylated in both tumors.

Due to the insufficient number of overlapping genes in the analysis of differentially methylated regions (DMRs), DMPs were employed for the subsequent analysis.

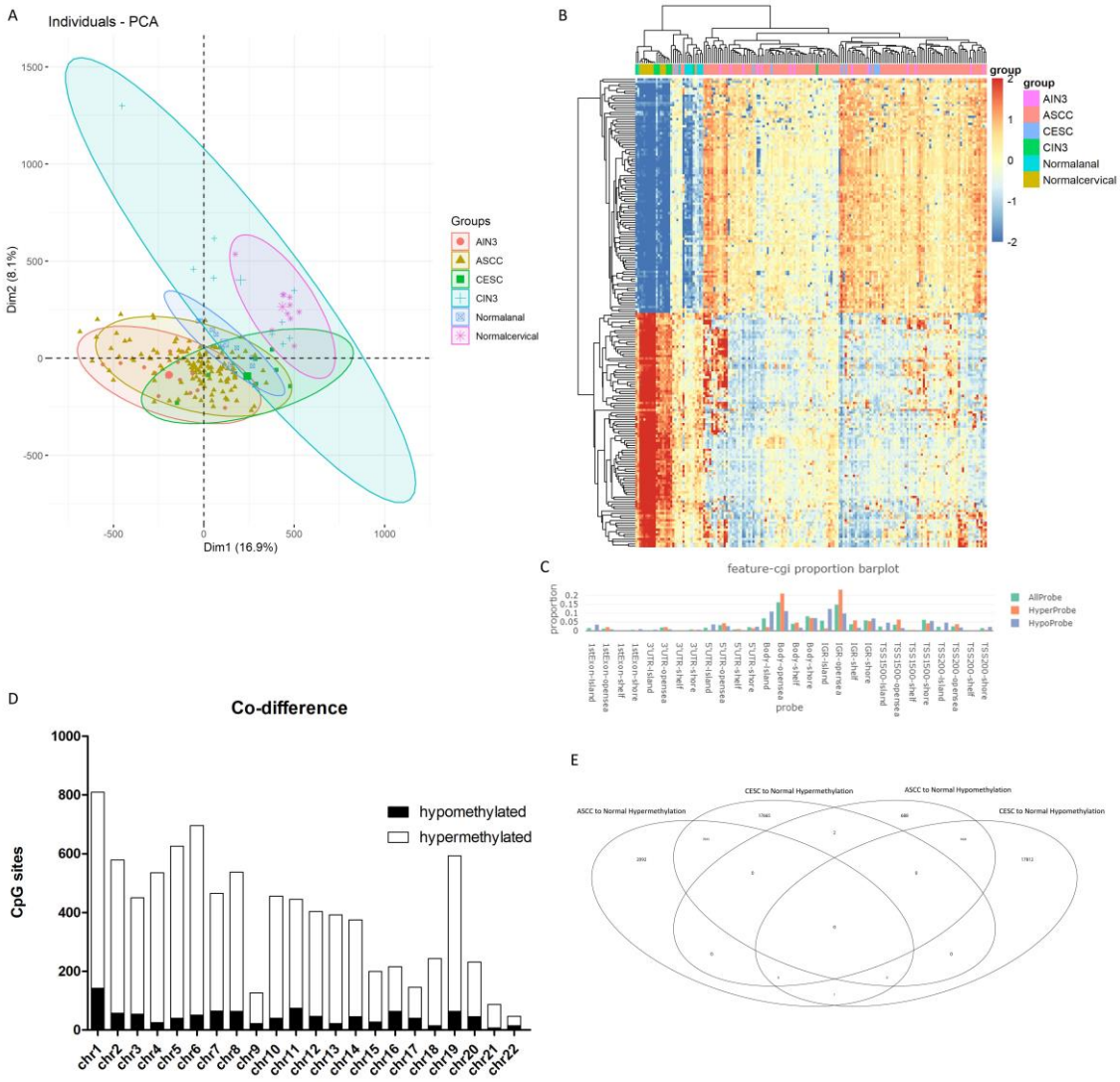


Figure 18 Overview of DNA methylation data.

A. Individuals plot of PCA, samples were represented as follows: red, AIN3. Brown, ASCC. Green, CESC. Blue cross, CIN3. Blue square, Normal cervical. Purple, Normal anal.
B. Heatmap of Top 200 DMPs, samples were represented as: Purple, AIN3. Red, ASCC. Blue, CESC. Green, CIN3. Light blue normal anal. Brown, normal cervical.
C. Proportion of the feature of CpG-islands, orange represented for hypermethylated probes, blue represented for hypomethylated probes.
D. The overlapped DMPs between CESC and ASCC, X axis represented for different chromosome, Y axis represented for number of DMPs, black bar represented for hypomethylated DMPs, white bar represented for hypermethylated DMPs.
E. Quantities of overlapped DMPs.

In the next step, KEGG analysis was conducted on the 7,633 hypermethylated DMPs and 1,024 hypomethylated DMPs (Fig 19A). The red bars represent genes associated with hypermethylated DMPs ($p < 0.05$), while the blue bars represent genes with hypomethylated DMPs ($p < 0.05$). Many of these sites were associated with carcinogenesis and the immune system. Furthermore, pathways related to HPV infection (173/3,314), oxidative stress (15/3,314), ferroptosis (17/3,314), and necroptosis (48/3,314) were also enriched. Oxidative stress, ferroptosis, and necroptosis processes are linked with infection and can occur in immune cells and some epithelial cells. It was hypothesized that these processes could be a result of HPV infection and contribute to carcinogenesis. Therefore, genes annotated with these functions were selected for further analysis.

Considering that the KEGG database is extensive but may have omissions in details, such as the omission of inducers, promoters, and drivers in the mapping of ferroptosis, additional annotation was performed using CIBERSORT, FerrDb, GeneCard, and other specific methods. As a result, 264 genes were identified as oxidative stress-associated, 23 genes were ferroptosis-associated, and 65 genes were necroptosis-associated. CIBERSORT was used to analyze the immune proportions of the genes with overlapping methylations (Fig 19B), revealing differences mainly in T-reg cells, CD4+ T cells, CD8+ T cells, dendritic cells, and monocytes (Fig 19C). This result was obtained from CESC due to the absence of RNA data and clinical data for ASCC.

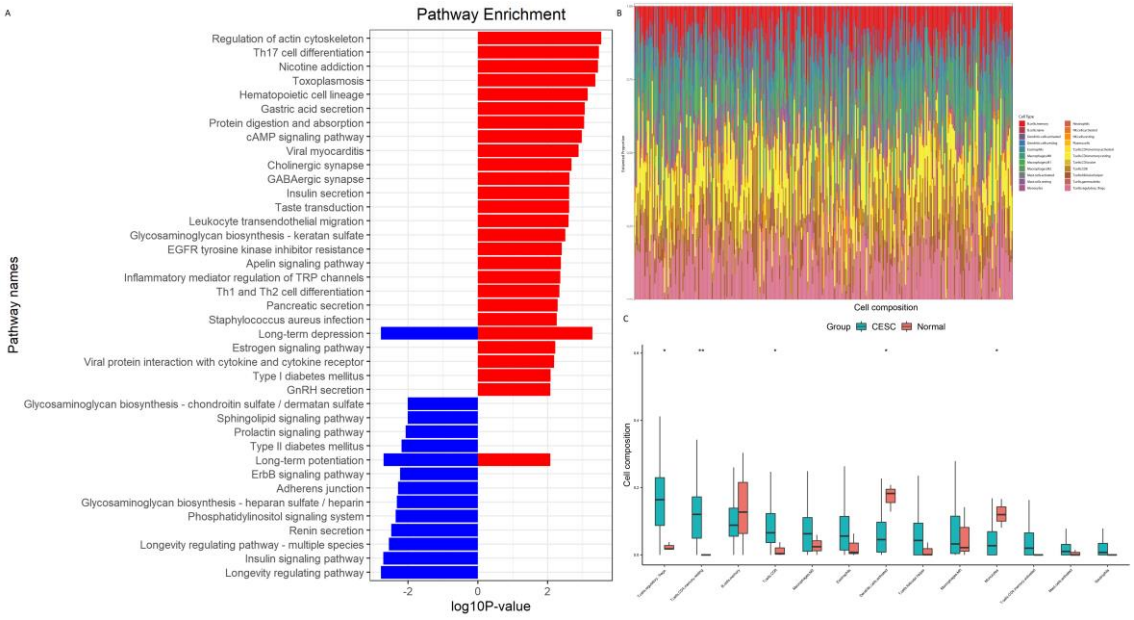


Figure 19 Function annotations and difference in immune infiltration of DMPs

A. KEGG analysis of the DMPs, red bars represented the genes with hypermethylated DMPs, blue bars represented the genes with hypomethylated DMPs.
B. Immune infiltration base on the gene with DMPs, each column represented one sample, different color represented different proportions of immune cell type.
C. Immune cell types with significant differences in the enrichment of genes for DMPs. （*p < 0.05, **p < 0.01）

3.3.2. Identification of key genes

Differential immune infiltration information was obtained through CIBERSORT analysis, but it was important to connect these immune alterations with specific genes. To achieve this, genes related to these immune alterations were identified from the scRNA data.

To analyze the immune identities and functions of the cells, the scRNA data from CESC were first clustered and visualized, resulting in 21 clusters using the Seurat FindCluster() function, and UMAP and tSNE algorithms from the Seurat R package (Fig 20A). Marker genes were obtained from these 21 clusters using the Findmarker() function, and the top 3 marker genes of each cluster were used for heatmap visualization to show the differential abundance of marker gene expression (Fig 20B). Reference data from Celldex were obtained, and the SingleR algorithm was used to automatically annotate clusters with well-known cellular markers (Fig 20C). Clusters that were annotated as specific cell types including T-reg cells (Cluster 4, marker: FOXP3), CD4+ T cells (Clusters 5, 7, 18, marker: IL7R), CD8+ T cells (Cluster 2, marker: CD8A), dendritic cells (Cluster 11, 13, 16, marker: LYZ), and monocytes (Cluster 8, 17, marker: CXCL6) were selected for further analysis.

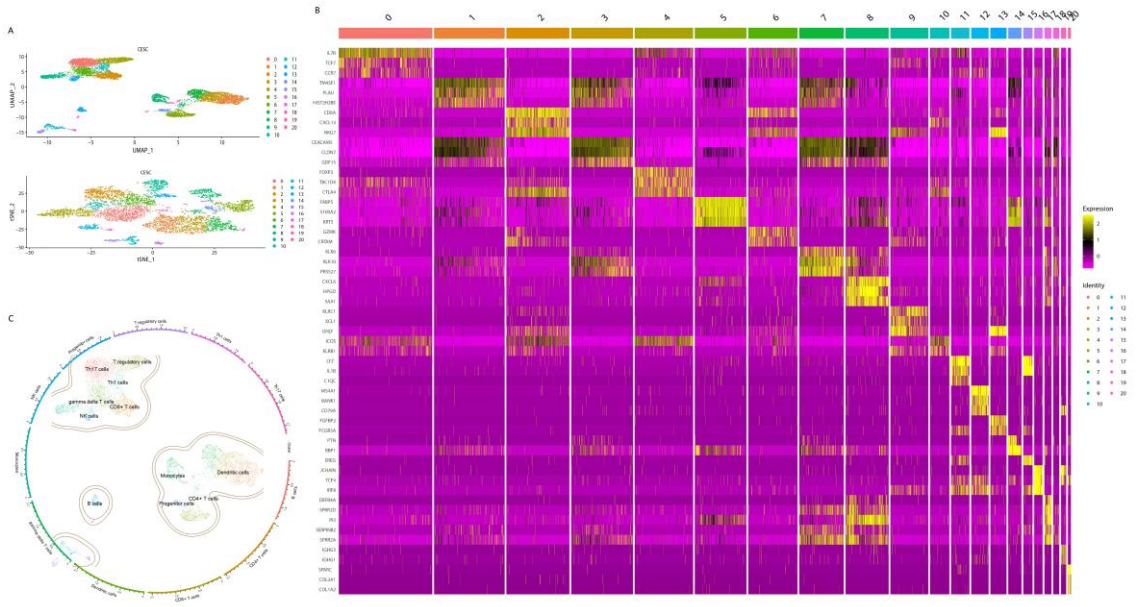


Figure 20 Clusters, marker genes and annotations of CESC scRNA-seq data.

A. The dimension reduction of CESC scRNA-seq. Visualization of separate clusters based on UAMP and tSNE.
B. Heatmap of the most significant marker genes, each cluster showed top 3 marker genes.
C. Clusters annotated by immune cell markers.

A co-expression network was constructed using the new matrix, and a soft threshold power β of 6 was selected to ensure a scale-free network (Fig 21A). Subsequently, 14 distinct gene modules were generated based on hierarchical clustering dendrogram (Fig 21B). Clinical features "Early group" and "Advanced group" were input as previously defined, and several modules including the pink module (r=-0.13, p=0.04), purple module (r=-0.14, p=0.04), salmon module (r=-0.13, p=0.04), blue module (r=-0.13, p=0.05), and grey module (r=-0.14, p=0.03) showed significant correlations with progression stage (Fig 21C).

To assess the potential biological function of these modules, a correlation between gene significance (GS) and module membership (MM) was created. Scatter plots were used to show the correlation of two modules as representations (Fig 21D, E). The modules demonstrated a significant association between GS and MM, indicating that the genes in those modules are not only co-expressed but also positively linked to clinical features.

All modules associated with clinical features were included in the study, resulting in a large set of 652 genes obtained from different modules. This data, including gene information (nodes) and weight from the adjacency matrix (edges), was visualized using Cytoscape.

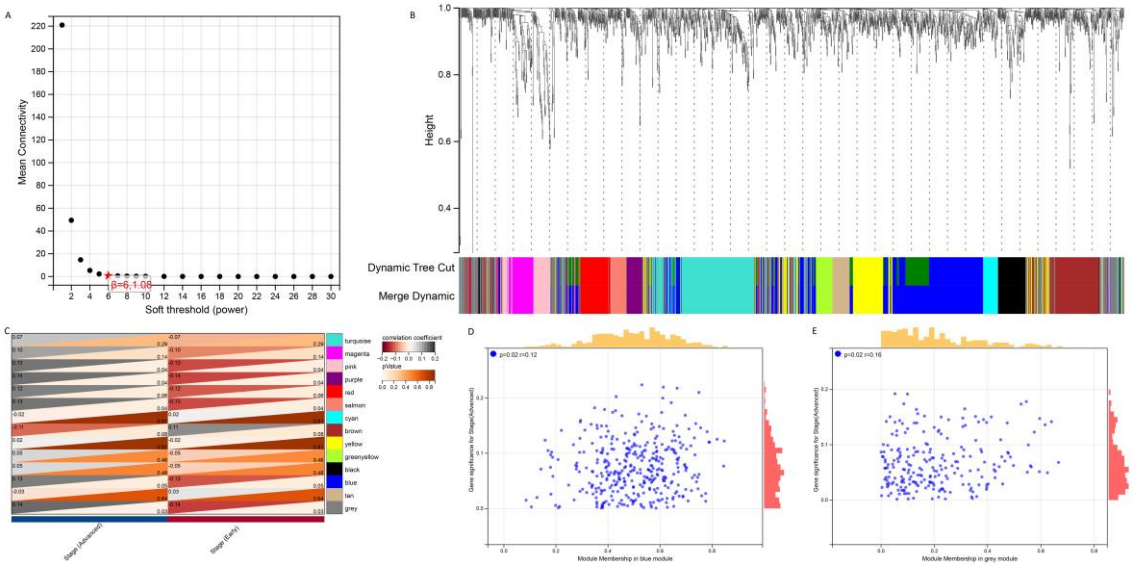


Figure 21 WGCNA analysis of the correlation between gene modules and clinical features

A. The mean connectivity and scale-free topology index for each power value between 1 to 20. Investigation of the mean connectivity (degree, Y axis) for different soft-thresholding powers (X axis).
B. Dendrogram of genes in new matrix (Associated to ferroptosis, necroptosis, oxidative stress, HPV infection pathway, T-reg cells, CD4+ T cells, CD8+ T cells, dendritic cells and monocytes) clustered based on a dissimilarity measure (1-TOM). Densely linked, highly co-expressed genes are grouped together on the dendrogram's branches.
C. Correlations of modules and clinical feature. Each row corresponded to a module, The number in the upper left corner represents the correlation, and the number in the lower right corner represents the p-value.
D. Scatter plot of module membership (MM) vs. gene significance (GS) in blue modules. MM presents the correlation between gene expression and each module eigengene. GS represents the association between gene expression and each trait. In both modules, GS and MM have a high correlation.
E. Scatter plot of module membership (MM) vs. gene significance (GS) in grey modules.

In the co-expression network analysis using Cytoscape plugin MCODE, highly interconnected regions in the network were identified, with nodes having more interconnected neighbors achieving higher scores. The highest region detected by MCODE revealed the top 10 highest scoring genes (Fig 22A), which included PSMD11, RB1, SAE1, TAF15, TFDP1, CORO1C, JOSD1, CDC42, KPNA2, and NUP62.

Subsequently, the relationship between the expression of these genes and CESC survival was analyzed using the GEPIA2021 online tool (Fig 22B, C, D, E, F, G, H, I, J, K). Among these highly co-expressed genes associated with progression, the expression of CDC42 was found to be significantly correlated with the survival of CESC (Hazard Ratio: 1.6, p=0.045), indicating that patients with higher expression levels of CDC42 had a worse prognosis. Therefore, CDC42 was selected as the target for downstream analysis.

All the connections of CDC42 were listed in the network (Fig 22L), where green diamonds represented genes associated with ferroptosis, purple ellipses represented genes clustered in immune cell clusters of scRNA, orange rectangles represented genes related to necroptosis, and yellow triangles represented genes annotated with oxidative stress.

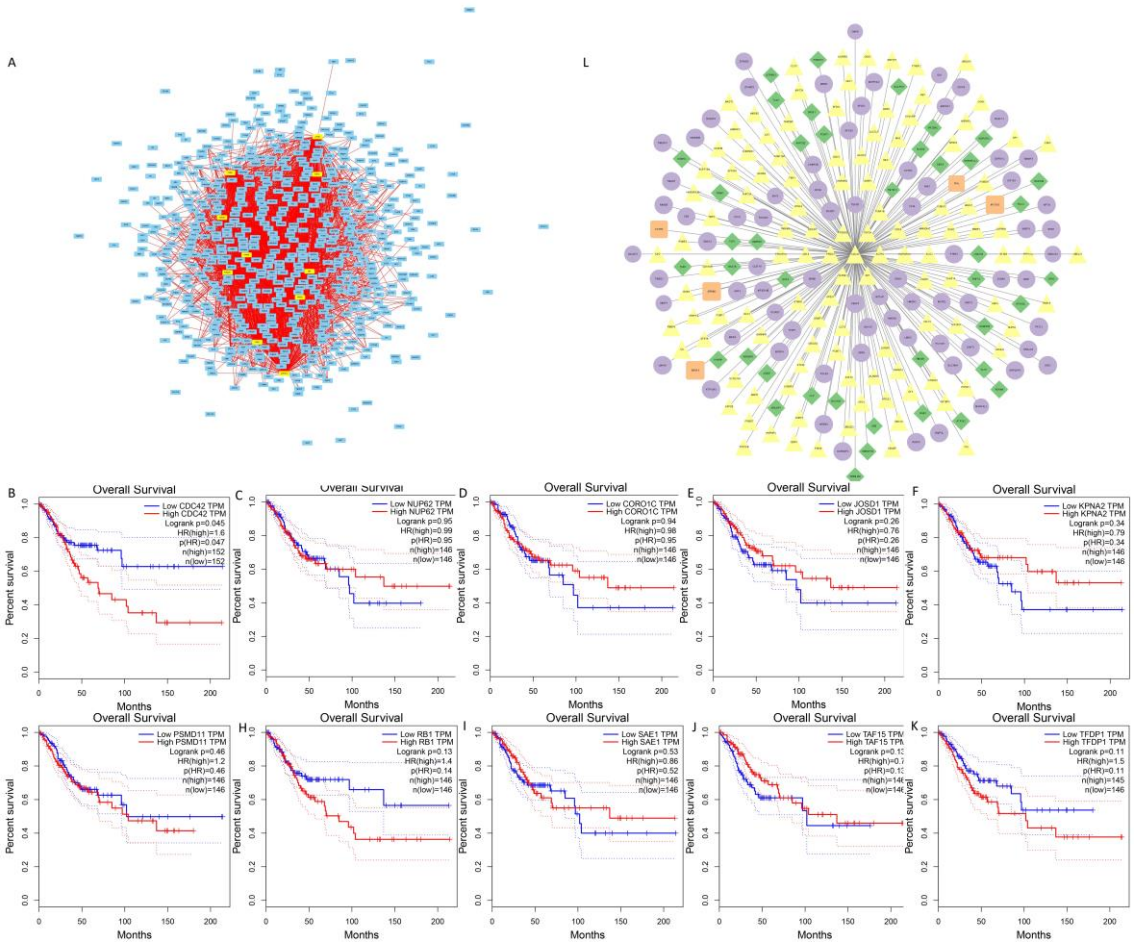


Figure 22 Identification of hub genes

A. Top 10 MCODE scoring genes were highlighted in the co-expression network.
B. KM curves (Kaplan–Meier estimator) showed the correlation between CESC overall survival and CDC42 expression.
C. KM curves showed the correlation between CESC overall survival and NUP62 expression.
D. KM curves showed the correlation between CESC overall survival and CORO1C expression.
E. KM curves showed the correlation between CESC overall survival and JOSD1 expression.
F. KM curves showed the correlation between CESC overall survival and KPNA2 expression.
G. KM curves showed the correlation between CESC overall survival and PSMD11 expression.
H. KM curves showed the correlation between CESC overall survival and RB1 expression.
I. KM curves showed the correlation between CESC overall survival and SAE1 expression.
J. KM curves showed the correlation between CESC overall survival and TAF15 expression.
K. KM curves showed the correlation between CESC overall survival and TFDP1 expression.
L. Hub gene CDC42 with its highly co-expressed neighbors. Green diamonds represented genes associated with ferroptosis, purple ellipses represented genes clustered in immune cell clusters of scRNA, orange rectangles represented genes related to necroptosis, and yellow triangles represented genes annotated with oxidative stress.

The role of CDC42 in HPV infection and carcinogenesis was investigated through DNA methylation and gene expression analysis. Three differentially methylated probes (DMPs) in CDC42, namely cg08608952, cg13962372, and cg23019935, were identified. The volcano plot showed the differential methylation of CDC42 between CESC and normal cervical tissue (Fig 23A) and between CESC and CIN3 (Fig 23B). Hypermethylation of the three DMPs was significant in the CESC group compared to the normal cervical group, while it was only slightly significant in CIN3 (Fig 23C).

In RNA profiling, the expression of CDC42 was found to be higher in CESC compared to normal cervical tissue (Fig 23D). This upregulation of CDC42 in CESC was also observed in scRNA-seq data, where CDC42 was upregulated in CESC samples compared to normal

samples (Fig 23E), indicating that the upregulation of CDC42 may be driven by carcinogenesis.

To assess the impact of HPV infection on CDC42 expression, CDC42 expression was quantified in HPV-positive CESC, HPV-negative CESC, and normal groups (Fig 23F). In the CESC scRNA-seq data, CDC42 was found in cluster 1 (Dendritic cells), cluster 3 (Dendritic cells), and cluster 7 (CD4+ T cells). Violin plots showed the expression of CDC42 in dendritic cells and CD4+ T cells between the three groups. Compared to normal tissues, CDC42 expression was lower, while among CESC samples, it was higher in the HPV-positive group than in the HPV-negative group, though not significantly. This suggests that the upregulation of CDC42 may be caused by both carcinogenesis and HPV infection.

Furthermore, a correlation analysis between CDC42 expression level and chemical sensitivities was performed using the CTRP database. The results revealed a significant positive correlation between the expression of CDC42 and 14 chemical compounds, with the epidermal growth factor receptor (EGFR) inhibitor afatinib being the most significant one (Fig 23G). In vitro experiments, including cell proliferation, cell viability, and western blot analysis, were conducted to validate these findings.

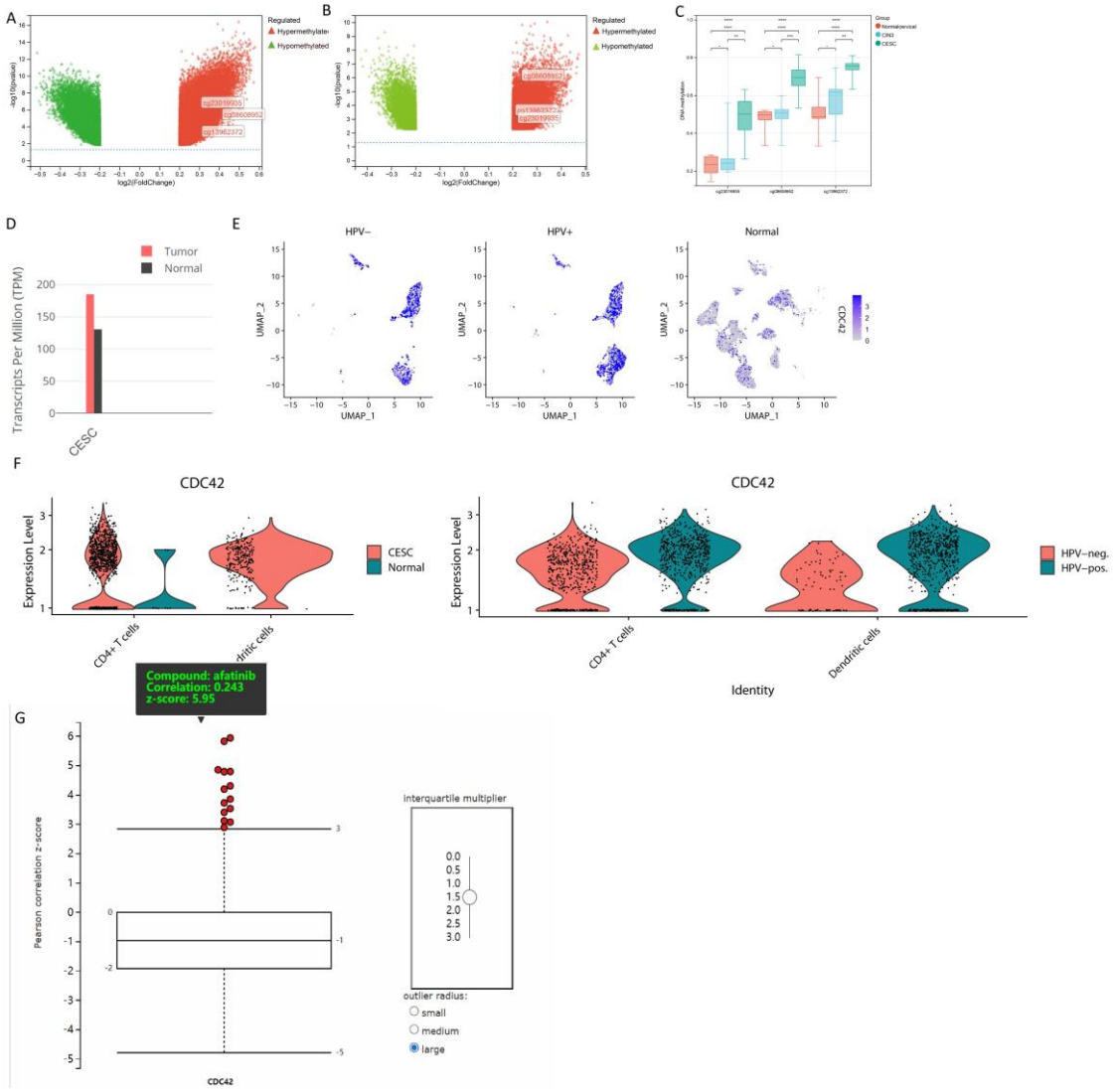


Figure 23 Gene pathway and process enrichment analysis and WGCNA analysis.

A. Volcano plot of the DMPs between CESC and normal samples. Red plot corresponded to hypermethylated probe, green plot corresponded to hypomethylated probes.
B. Volcano plot of the DMPs between CESC and CIN3 samples.
C. β value of CDC42 DMPs (cg08608952, cg13962372 and cg23019935) in different groups.
D. RNA expression of CDC42 in CESC and normal samples.
E. Colors single cells on a dimensional reduction plot according to the expression of CDC42.
F. Quantifies of CDC42 in CD4+ T cells and dendritic cells between CESC and normal samples, HPV-pos. CESC and HPV-neg. CESC.
G. Correlation analysis of CDC42 expression level and chemical compounds sensitives, afatinib showed the significantly and positively correlation.

3.3.3. Cell proliferation, cell viability and western blot

WST-1 assays were conducted on HaCaT and A431 cells with or without HPV16 E6 and E7 transfection in the presence of various concentrations (0 M, 0.1 M, 1 M, and 10 M) of afatinib to assess changes in cell viability. The spectrophotometric readings of cells exposed to afatinib at a concentration of 0 M were used as the reference standard (100%) in the figure (Fig 24A, C). Similarly, the spectrophotometric readings of cells without transfection and afatinib were set as the standard to determine the effect of HPV transfection on cell proliferation (Fig 24B, D).

The results showed that cell proliferations of HaCaT and A431 were increased after transfection with HPV16 E6 and E7 ($p < 0.05$). In HaCaT cells, afatinib did not reach the IC50 at 10 µM, whereas HaCaT cells transfected with HPV had a lower IC50 at 1 to 10 µM. Similar results were observed in A431 cells, where the IC50 was between 1-10 µM without transfection and there was a lower IC50 after transfection with HPV.

Cell proliferation and cell viability assays confirmed that transfection with HPV16 E6 and E7 can make cells more sensitive to afatinib. However, the proteins altered in this process remained unclear. After treating the cells with afatinib for 24 hours (48 hours following transfection), protein expressions were measured (Fig 24E). Transfection with HPV16 E6 and E7 enhanced the effects of afatinib. Quantitative analysis with Image J showed lower expression of pEGFR in the transfected HPV group at the same afatinib concentration, while CDC42 was up-regulated in the transfected HPV group (Fig 24F). It's important to note that these results should be interpreted as a general trend and not as precise quantification. The Western blot analysis could not clarify the specific mechanism, and no replicate experiments were performed, so the results were not statistically significant.

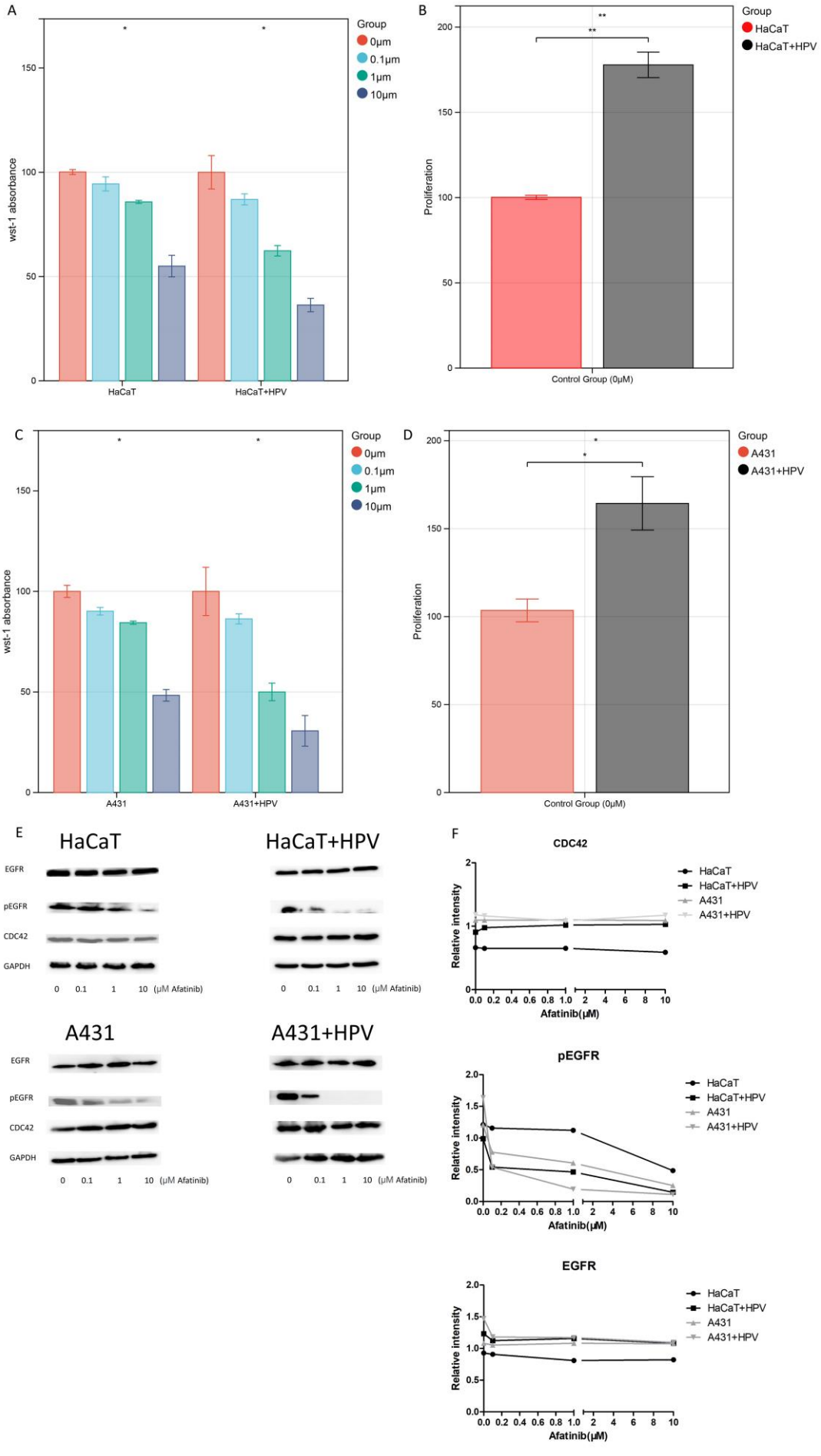

A
wst-1 absorbance
Group
0μm
0.1μm
1μm
10μm
HaCaT
HaCaT+HPV
B
Proliferation
Group
HaCaT
HaCaT+HPV
Control Group (0μM)
C
wst-1 absorbance
A431
A431+HPV
D
Proliferation
Group
A431
A431+HPV
Control Group (0μM)
E
HaCaT
HaCaT+HPV
A431
A431+HPV
EGFR
pEGFR
CDC42
GAPDH
0 0.1 1 10 (μM Afatinib)
F
CDC42
pEGFR
EGFR
Relative intensity
Afatinib(μM)
HaCaT
HaCaT+HPV
A431
A431+HPV

Figure 24 Cell proliferation activities and Western blot

A. WST-1 cell viability assay of HaCaT cell, 0 μM afatinib groups were set as 100%.
B. WST-1 cell proliferation test of HaCaT cells transfected with HPV E6 and E7, un-transfected and untreated HaCaT cells groups were set as 100%.
C. WST-1 cell viability assay of A431 cell, 0 μM afatinib groups were set as 100%.
D. WST-1 cell proliferation test of A431 cells transfected with HPV E6 and E7, un-transfected and untreated A431 cells groups were set as 100%.
E. Western blot of EGFR, pEGFR and CDC42.
F. Quantitative analysis of the western blot results. Experiment was performed only once, so there was not statistical analysis.

3.3.4. Molecular docking

Molecular docking analysis was performed to investigate the binding of afatinib with CDC42 and EGFR. The binding poses and interactions of afatinib with CDC42 and EGFR were acquired, and binding energy was calculated for each interaction (Fig 25A, B). The results showed that afatinib bound to both CDC42 and EGFR through apparent hydrogen bonds and strong electrostatic interactions. Afatinib also occupied the hydrophobic pockets of CDC42 and EGFR (Fig 25C, D).

Binding energy values less than -7.0 kcal/mol indicate strong binding activity of the ligand to the receptor. There were 9 binding models of afatinib to EGFR, and all of them had low binding energies, ranging from -8.027 kcal/mol to -9.186 kcal/mol, indicating highly stable binding. In contrast, the only binding model of afatinib to CDC42 had a much lower binding energy of -81.793 kcal/mol, indicating a higher affinity. The figure displayed the interaction forces as dashed lines, and the amino acids involved in the interaction were shown in a ball-and-stick model, illustrating the super-strong interaction force between afatinib and CDC42.

Although the affinity of afatinib for both CDC42 and EGFR was confirmed, it remained unclear how CDC42 acted on EGFR after binding to afatinib. Attempts were made to calculate CDC42-afatinib-EGFR interactions, but no interactions could be calculated. However, in the PDB database, it was found that CDC42 could bind to the extended GTPase-effector interface. Therefore, a combination of CDC42-GTPase-effector interface-EGFR-afatinib was constructed, and ZDOCK was used for protein structure docking. The top 10 best ZDOCK scores for the CDC42-GTPase-effector interface-EGFR-afatinib complex were obtained, and they all had high scores, indicating strong docking (Fig 25E).

The interaction involved hydrogen bond links with amino acid sites such as GLU 171, ASN167, LEU165, VAL 168, PHE 169, and LYS 166-GTPase-effector interface GLU1005, ARG 489, LEU 475. GTPase-effector interface formed hydrogen bond links with amino acid sites such as GLN303, ASN996-PRO 975, ALA 972 EGFR (Fig 25F). EGFR was bound to afatinib through amino acids such as ASN 808, HIS 988, PRO 848, forming hydrogen bonds and strong interactions (Fig 25G). Comprehensive analysis indicated that the CDC42-GTPase-effector interface-EGFR-afatinib complex formed a stable docking model.

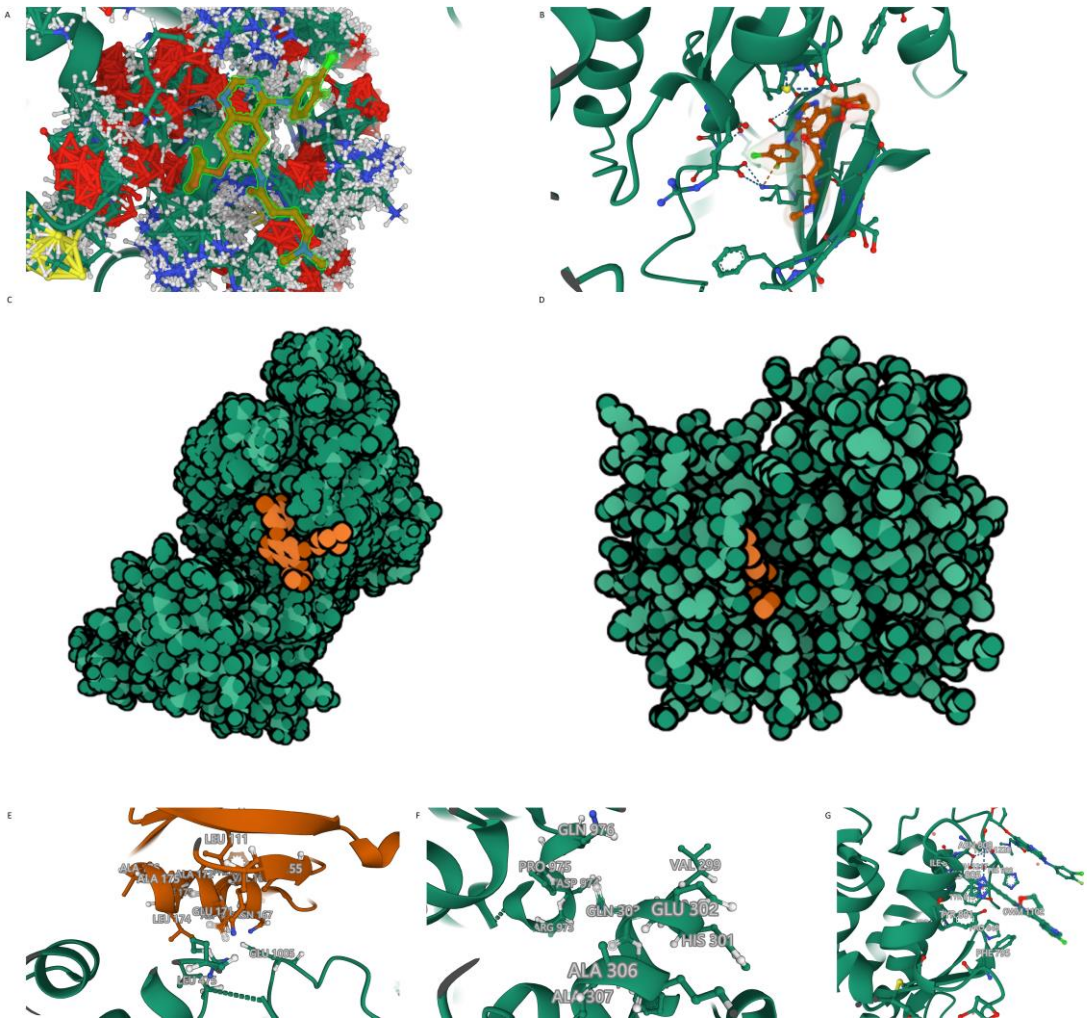


Figure 25 Molecular docking and Rigid protein–protein docking

A. Binding mode of afatinib to CDC42. CDC42 was set as ball-and-stick model with gaussian volume. The amino acids involved in the interaction are shown in a ball-and-stick model, the amino acids did not involve in the interaction were showed as cartoon.
B. Binding mode of afatinib to EGFR.
C. The Molecule of the Month feature used cartoon illustrations to demonstrate

the overlay of the crystal structures of afatinib and CDC42.

D. The Molecule of the Month feature used cartoon illustrations to demonstrate the overlay of the crystal structures of afatinib and EGFR.
E. CDC42 formed hydrogen bonds with the extended GTPase-effector interface amino acid sites.
F. Extended GTPase-effector interface formed hydrogen bonds with EGFR amino acid sites.
G. Binding mode of afatinib to CDC42- Extended GTPase-effector interface -EGFR.

4. Discussion

4.1. Heterogeneities of tumors caused by high risk-HPV infection

While HPV-associated lesions are known to carry an elevated risk of progressing to malignant tumors [61], there has been a noticeable scarcity of studies delving into the comparative aspects of immune cell infiltration, gene expression profiles, and tumor heterogeneity between HPV-pos. and HPV-neg. cancers. This study underscores the potential of single-cell transcriptomics as a potent tool for dissecting the intricacies of heterogeneity within HPV-pos. and HPV-neg. cancers. This multifaceted approach not only enhances our comprehension of the influence of HPV infection on cancer biology but also lends itself to the identification of prospective gene markers with diagnostic, prognostic, and therapeutic significance [12].

Our initial analysis honed in on CESC samples, unequivocally linked to high-risk HPV infection, to corroborate the presence of discernible heterogeneity between HPV-pos. and HPV-neg. cases. Our underlying hypothesis posits that HPV's impact on the host extends from manifest transcriptional alterations to distinct patterns of immune cell infiltration. Consequently, our investigative journey embarked on scrutinizing transcriptional variances between HPV-pos. and HPV-neg. samples, leveraging a combination of scRNA-seq and bulk RNA sequencing.

Our findings brought to light distinct gene expression profiles within HPV-pos. samples. Notably, CD8+ T cells and B cells displayed a propensity for downregulation, juxtaposed with the upregulation of T-reg, $CD4^+$ T cells, and epithelial cells among HPV-pos. CESC patients. Remarkably, these discernible patterns of gene expression align harmoniously with observations from prior investigations into immunotherapeutic strategies for CESC [62]. This concurrence is particularly evident in the pivotal roles played by adaptive immunity activation and the ensuing generation of helper T cells, which, in turn, catalyze the production of $CD8^+$ cytotoxic T cells, specially attuned to target viral early proteins—an essential facet of HPV clearance in basal epithelial cells. Furthermore, the orchestration of T-helper cells serves to amplify B cell activation, culminating in enhanced defense mechanisms against subsequent infections, both at mucosal and systemic levels.

Following the observation of differential immune infiltration between HPV-pos. and HPV-neg. samples, our investigation delved deeper into the landscape of tumor immune infiltration through CIBERSORT analysis of bulk RNA sequencing samples coupled with survival data. This examination unveiled a noteworthy finding: samples characterized by

the infiltration of $CD8^+$ cells and B cells exhibited significantly prolonged survival durations.

In light of this, we embarked on a quest to identify differential genes, which included genes with low immune cell specificity, and bore the strongest association with CD8+ cells and B cells. This endeavor was facilitated through the employment of WGCNA. Subsequently, these genes served as the foundation for constructing survival and relapse prediction models. Our quest culminated in the identification of three genes—JAK3, FOXP3, and IKZF3—exhibiting reduced expression in the HPV-pos. group. These genes emerged as protective factors, heralding enhanced prospects for both survival and reduced risk of recurrence. Notably, heightened expression of these three genes corresponded to lower RiskScores within the model, indicative of a diminished risk of mortality and relapse.

The observed differences between HPV-pos. and HPV-neg. HNSCC samples align with the increased expression of these three pivotal genes, thereby signifying a higher likelihood of survival among HNSCC patients. However, it's noteworthy that no other potential HPV-associated tumors yielded analogous results. In fact, only IKZF3 was identified as a favorable factor in BLCA. This outcome resonates with the insights of Jørgensen KR and Jensen JB, who emphasized the challenges in establishing a conclusive role of HPV in BLCA across its various histological subtypes [63]. In COAD and PRAD, our analysis did not identify any hub genes capable of correlating gene expression levels with survival outcomes. This observation resonates with the findings of KHOURY et al. [64], who pointed out that RNA-Seq data did not detect any HPV transcripts in prostate or colon adenocarcinomas. Hence, it becomes evident that while HPV infection exerts an impact on hub genes, elevated hub gene expression in non-HPV-infected tumors does not inherently confer a protective effect against adverse clinical outcomes.

Considerable research has explored the roles of FOXP3, JAK3, and IKZF3 within the realm of tumor immunology. FOXP3, deemed the most specific and reliable biomarker for T-reg cells, has yielded conflicting results. In the context of patient survival across various cancer types, the infiltration of tumors by FOXP3+ Tregs has been regarded as unfavorable. Recent research has introduced Helios as a novel Treg marker, indicating that the immunosuppressive function of Tregs is only exerted when both FOXP3 and Helios are expressed [65]. Interestingly, FOXP3 has emerged as a favorable prognostic factor in the study of HPV-positive tonsillar squamous cell carcinoma [65, 66]. Additionally, the Foxp3 gene has been identified as an immune regulator capable of suppressing oncogenes while activating tumor suppressor genes [67]. A reduction in FOXP3 expression may compromise its ability to repress oncogenes and activate tumor suppressors [68, 69].

JAK3, a tyrosine kinase belonging to the Janus family of kinases, is predominantly expressed in T cells and NK cells [70]. Studies have indicated that HPV-infected individuals exhibit significant combined immunological dysfunction associated with JAK3 deficiency [71]. Moreover, JAK3 mutations have been detected in HPV-positive Squamous Cell Carcinomas of the Oropharynx (OP-SCC) [72] and Cervical Precancers [73]. IKZF3, a transcription factor crucial for regulating B cell proliferation and differentiation, has been implicated as a target for HPV integration [74]. In investigations involving OPSCC, IKZF3 RNA was detected in HPV-positive cell lines, and immune cell infiltration exhibited substantial IKZF3 positivity, irrespective of the tumor cells' expression pattern. Notably, IKZF3 demonstrated a protective effect with a hazard ratio of 0.38 in this study [75]. This research has introduced two signatures based on the heterogeneity of CESC, potentially applicable for prognostic evaluation, relapse prediction, and the identification of novel immunotherapy targets.

In conclusion, the analysis of scRNA-seq and bulk RNA-seq data from both HPV-positive and HPV-negative CESC samples has unveiled a spectrum of heterogeneity encompassing transcriptional states and immune infiltration. Leveraging this polymorphism, we have identified genes associated with prognosis that are influenced by HPV infection. Using these genes, we have constructed survival and relapse prediction models through LASSO Cox regression analysis. Remarkably, the distribution and protective effects of IKZF3, FOXP3, and JAK3 within both risk models closely mirror their roles in HPV-associated HNSCC, suggesting that HPV infection modulates the expression of specific hub genes. However, it's worth noting that the genetic mechanisms underlying both signatures remain unexplored. Future research should delve into the molecular intricacies of signature genes, providing experimental evidence to facilitate their application in risk prediction and the treatment of HPV-associated cancers.

4.2. Transcriptome changes caused by skin low-risk HPV infection

The analysis of CESC not only revealed heterogeneity in transcriptional states and immune infiltration between HPV-pos. and HPV-neg. cancers but also indicated that HPV infection led to changes in specific hub genes. However, it remained unclear whether these effects were exclusive to high-risk HPV infections. To address this question, an analysis of low-risk HPV infections was conducted.

The analysis of low-risk HPV began by examining samples from various sources, including condyloma acuminata, treeman syndrome, and normal skin samples, alongside wart samples. It was hypothesized that low-risk HPV might have similar effects to high-risk

HPV. Consequently, the analysis of low-risk HPV infections commenced by identifying DEGs and assessing immune infiltration.

In low-risk HPV-pos. samples, a reduction in CD8+ T cells, consistent with high-risk HPV infections, was observed. Furthermore, alterations were noted in various immune cell populations, including macrophages, neutrophils, dendritic cells, plasma cells, mast cells, and NK cells. Prior research has indicated that low-risk HPV infections can induce the transformation of macrophages from an M1 to an M2 phenotype, contributing to immunosuppression [76]. Additionally, FOXP3+ Tregs have been implicated in this process [77], aligning with the findings from the analysis of high-risk HPV infections. Further analysis of different low-risk HPV-infected samples highlighted a significant infiltration of dendritic cells, particularly plasmacytoid dendritic cells (pDCs). pDCs are a specialized subset of dendritic cells known for their robust production of type I interferons (IFNs) and their crucial role in antiviral responses [78].

While previous studies have predominantly examined the association of neutrophils, plasma cells, mast cells, and NK cells with high-risk HPV infections, this analysis focused on their relationship with low-risk HPV infections. These immune cell alterations have been frequently linked to high-risk HPV infections in various cancers, including penile cancer[79], head and neck squamous cell carcinoma (HNSCC)[80, 81], squamous cell carcinoma (SCC)[82], and cervical squamous cell carcinoma (CESC)[83, 84]. The unexpected changes in immune infiltration and the involvement of genes with low immune cell specificity prompted further characterization of genes related to these cell types using functional assays.

Surprisingly, the analysis revealed enrichment of pathways associated with cancer, human diseases, and human papillomavirus (HPV) infection, suggesting that genetic changes induced by low-risk HPV may share similarities with alterations caused by high-risk HPV infections. Additionally, pathways related to tumor processes such as ferroptosis, oxidative stress, and necroptosis were also enriched. Epithelial tissues, being the primary target of HPV infection, are highly exposed to oxidative stress from various sources [85]. High-risk HPV infections have been shown to induce oxidative stress, perpetuate the inflammatory process, generate reactive oxygen species (ROS), cause DNA damage, and create a favorable environment for malignant transformation [86]. Furthermore, cells exhibit increased resistance to oxidative stress after undergoing cancerization, and HPV E7 promotes uncontrolled proliferation of infected cells [86]. High-risk HPV infections also induce oxidative stress, leading to reduced SLC7A11 expression and increased sensitivity to ferroptosis [60]. Ferroptosis is generally associated with high-

risk HPV infections, as squamous intraepithelial lesions progress into squamous cell carcinoma following such infections, leading to the accumulation of intracellular Fe-(II) [87]. Moreover, HPV-induced oxidative stress can induce necroptosis, but high-risk HPV infections render cells resistant to necroptosis by downregulating RIPK3 [88]. RIPK3 expression controls macrophage M1 markers, which have been associated with the prognosis of HPV-related cancers [89].

Despite the strong predictability of the models based on these pathways, it is challenging to determine which group of pathways is the most critical. Therefore, a PPI network incorporating all genes from the risk models was constructed to gain further insights.

Seven genes were obtained from the hub cluster of the protein-protein interaction (PPI) network. STAT1 [90], ICAM1 [91], NOX4 [92], and KIT [93] are associated with oxidative stress. ICAM1 is highly expressed in neutrophils [94], while STAT1 is specifically expressed in single-cell cluster T cells [95]. KIT is enriched in granulocytes [96] and NK cells [97], while NOX4 is detected in non-immune cells. HSPA4 is a low immune-specific gene associated with necroptosis [98]. Low immune-specific VEGFA and monocyte-specific HMOX1 are associated with ferroptosis [99]. According to the CTRP V2 database, alterations in the expression of these seven hub genes are associated with different compound sensitivities. Integration of major inhibitors revealed that the Ras-ERK and PI3K-mTOR pathways are primarily involved.

This step offers the advantage of integrating transcriptomic alterations caused by multiple low-risk HPV infections, enabling the discovery of potential associations of genes in low-risk HPV skin lesions with skin cancer through predictive models and functional analysis. However, there is a notable disadvantage. Genetic alterations caused by different low-risk HPV types were analyzed together, potentially obscuring disease-specific alterations or alterations specific to a particular HPV type. For instance, some studies have reported that low-risk HPV infections could be associated with a 1-3% risk of cancer progression [100], a small change that may easily be overlooked without a large sample size.

If the genetic alterations originate from skin lesions, the results are understandable since both high-risk and low-risk HPV types can be found on the skin of immunocompetent individuals without symptomatic lesions [100]. However, the presence of other diseases may make patients more susceptible to HPV infections and the development of precancerous lesions in sun-exposed areas [101]. If the genetic alterations are attributed

to low-risk HPV infections, it raises the question of whether the carcinogenic role of high-risk HPV infections should also be attributed to immune alterations. This would imply altered pathways, including those related to cancer, oxidative stress, ferroptosis, and necroptosis, rather than solely focusing on immune changes, as in the CESC analysis.

The question remains, and it may be worthwhile to address it in the next step by investigating a disease where both precancerous lesions and cancer are affected by HPV infection. The findings from the first two analyses were also mapped to transcriptome profiling data to determine if the target genes were altered with HPV infection and the carcinogenesis of keratinocytes.

4.3. Commonalities of high-risk and low-risk HPV infections

CESC and ASCC, along with their precancerous lesions CIN3 and AIN3, were selected for further investigation to elucidate the hypothesis. Additionally, this study aimed to address the treatment of HPV-associated cancer. Despite significant advancements in chemoradiotherapy, there have been no substantial changes in the management of ASCC and CESC treatment in recent years. The identification of biomarkers that support clinical decision-making could be highly beneficial. However, there is a limited amount of research on precision medicine for ASCC and CESC, highlighting the need for increased focus to enhance our understanding of these diseases [102, 103].

In the preceding steps, HPV-positive and HPV-negative cervical cancer samples were analyzed to illustrate the heterogeneity from the transcriptional state to immune infiltration. Similar trends were observed in other HPV-related cancers, such as HNSCC. Furthermore, the analysis of low-risk HPV infections revealed that, in addition to immune effects similar to high-risk HPV infections, low-risk HPV also impacts pathways related to ferroptosis, necroptosis, oxidative stress, and other cancer-related pathways. This suggests that benign skin lesions exhibit genetic alterations similar to those seen in skin cancer. Consequently, this step aims to achieve a more profound molecular characterization of HPV-associated cancers and to propose novel potential treatment options.

Through a comparison of the methylation β-values of probes, we identified 7,633 DMPs that were hypermethylated and 1,024 DMPs that were hypomethylated in both ASCC and CESC. These DMPs were associated with genes primarily involved in carcinogenesis

and immune function. CIBERSORT analysis revealed that genes with overlapping methylation patterns were predominantly differentially expressed in T-reg cells, CD4+ T cells, CD8+ T cells, dendritic cells, and monocytes. Genes related to these immune cell types were extracted from CESC single-cell RNA-seq data. Additionally, we identified 264 genes associated with oxidative stress, 23 genes associated with ferroptosis, and 65 genes associated with necroptosis based on functional annotations. Using these genes from CESC RNA-seq data, we constructed a new matrix that included expression profiles and clinical data. WGCNA analysis of this new matrix demonstrated that 652 of these genes were significantly correlated with clinical TNM stage, with CDC42 showing high co-expression and a significant association with overall survival in CESC.

Further network analysis in Cytoscape of the genes connected to CDC42 revealed that most of these connections were related to oxidative stress, and oxidative stress-related genes were prominent among the hub genes identified in low-risk HPV infections. While the results of immune analysis partially agreed, it is plausible to assume that immune alterations occur separately in different types of HPV infections. However, the more likely commonality appears to be alterations in the oxidative stress pathway.

CDC42, a pivotal regulator of cell cycle progression, has been firmly implicated in the pathogenesis of various cancer types [104]. Notably, CDC42 tends to be overexpressed in a multitude of malignancies, a phenomenon often associated with unfavorable clinical outcomes [104]. The present study has unveiled a notable increase in CDC42 expression levels that parallel disease progression, a trend consistent with observations in other cancer types. Intriguingly, both high-risk HPV E6 and low-risk HPV E6 have been identified as activators of CDC42, with the former demonstrating a more pronounced effect on CDC42 activation levels [105]. To further elucidate these findings, in vitro experiments were conducted using normal keratinocytes (HaCaT) and squamous cell carcinoma cells (A431), which revealed heightened CDC42 expression subsequent to transfection with HPV E6/7. Consequently, it can be postulated that the upregulation of CDC42 in both CESC and ASCC can be attributed to epithelial cell carcinogenesis facilitated by HPV infection.

However, it is worth noting that developing a selective inhibitor for CDC42 remains an intricate challenge, primarily due to its structural resemblance to other GTPases within the Rho family and the broader Ras superfamily [106]. An intriguing observation emerged from our analysis, demonstrating a positive correlation between elevated CDC42 expression levels and heightened sensitivity to the EGFR inhibitor afatinib, as evidenced by data obtained from the CTRP database. The apparent conundrum regarding

CDC42's high affinity for the EGFR inhibitor prompted us to employ molecular docking techniques in an attempt to elucidate potential intermolecular interactions. Surprisingly, afatinib exhibited a higher binding affinity for CDC42, underscored by a lower binding energy. Nonetheless, this observation left us grappling with the physical conundrum, given that CDC42 is localized within microtubules and cytokinetic bridges, while EGFR primarily resides within the plasma membrane and cell junctions. Such spatial incongruities do not readily explain their direct interaction.

To unravel this perplexity, we delved deeper into the molecular intricacies. It emerged that upon binding, an extended GTPase-effector interface formed, linking CDC42 to EGFR in a stable docking model when afatinib was introduced. This finding suggests that the augmented sensitivity to afatinib resulting from CDC42 upregulation likely involves a more complex mechanistic interplay. Remarkably, CDC42, a gene closely associated with oxidative stress, has been shown to interact with the γ subunit of the coatomer protein complex, thus facilitating the accumulation of EGFR [107]. This interaction triggers the production of reactive oxygen species (ROS), often seen in response to EGFR overexpression, mutational activation, or the abnormal activation of downstream effectors upon exposure to oxidative stress. ROS, in turn, promote cell proliferation [108].

In vitro experiments further substantiated these findings by revealing that afatinib-induced apoptosis in HaCaT/A431 cells was notably enhanced in conjunction with elevated CDC42 expression levels, resulting in the more efficient repression of EGFR signaling. Consequently, it is plausible to posit that HPV infection and epithelial cell carcinogenesis contribute to the upregulation of CDC42 in CESC, a phenomenon that holds promise as a potential biomarker for EGFR-targeted inhibitor therapy.

While afatinib is primarily indicated for the treatment of non-small cell lung cancer (NSCLC) in individuals harboring EGFR mutations [109], its selection for this study was motivated by the absence of specific EGFR inhibitors tailored for the treatment of HPV-related squamous cell carcinoma. Although the precise mechanistic details underlying the interaction between afatinib and CDC42 remain elusive, it is worth noting that one CESC patient derived notable benefit from afatinib. However, the levels of CDC42 expression in this particular patient with EGFR amplification remain undisclosed [110].

In summary, our comprehensive analysis has illuminated both shared characteristics and nuanced distinctions between high-risk and low-risk HPV infections. While these distinct

HPV types engender diverse alterations in immune cell subtypes, a common thread emerges in their shared impact on oxidative stress pathways. Our investigation has further delved into the intricacies of oxidative stress alterations, shedding light on the underlying mechanisms of carcinogenesis in the context of HPV infection.

Notably, the upregulation of CDC42 has emerged as a significant contributor to the increased sensitivity of cancer cells to the EGFR inhibitor afatinib. This finding suggests a promising avenue for therapeutic intervention, emphasizing the potential utility of CDC42 as a biomarker for guiding EGFR-targeted inhibitor therapy. The proposed specific interaction involving CDC42, the GTPase-effector interface, EGFR, and afatinib presents an intriguing direction for future research and clinical exploration.

Moving forward, it is imperative for future investigations to shift their focus towards intricate in vivo models, which can provide invaluable insights into the practical application of the CDC42-afatinib-EGFR combination. Additionally, the development of specific oxidative stress inhibitory compounds tailored to HPV infection and epithelial cell carcinogenesis represents a promising frontier in the quest for more effective therapeutic strategies.

References

1. Milner DA. Diagnostic pathology: infectious diseases E-Book: Elsevier Health Sciences; 2019.

2. Ganguly N, Parihar SP. Human papillomavirus E6 and E7 oncoproteins as risk factors for tumorigenesis. J Biosci. 2009;34(1):113-23.

3. Humans IWGotEoCRt, Meeting IWGotEoCRtH, Organization WH, Cancer IAfRo. Human papillomaviruses: World Health Organization; 2007.

4. Gong Y, Sui L, Li Y. Recombination in Papillomavirus: Controversy and Possibility. Virus Res. 2022;314:198756.

5. Getinet M, Gelaw B, Sisay A, Mahmoud EA, Assefa A. Prevalence and predictors of Pap smear cervical epithelial cell abnormality among HIV-positive and negative women attending gynecological examination in cervical cancer screening center at Debre Markos referral hospital, East Gojjam, Northwest Ethiopia. BMC Clin Pathol. 2015;15:16.

6. Munoz N, Bosch FX, de Sanjose S, Herrero R, Castellsague X, Shah KV, et al. Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med. 2003;348(6):518-27.

7. Munger K, Howley PM. Human papillomavirus immortalization and transformation functions. Virus Res. 2002;89(2):213-28.

8. Yeo-Teh NSL, Ito Y, Jha S. High-Risk Human Papillomaviral Oncogenes E6 and E7 Target Key Cellular Pathways to Achieve Oncogenesis. Int J Mol Sci. 2018;19(6).

9. Chesson HW, Dunne EF, Hariri S, Markowitz LE. The estimated lifetime probability of acquiring human papillomavirus in the United States. Sex Transm Dis. 2014;41(11):660-4.

10. Meites E, Szilagyi PG, Chesson HW, Unger ER, Romero JR, Markowitz LE. Human papillomavirus vaccination for adults: updated recommendations of the Advisory Committee on Immunization Practices. Wiley Online Library; 2019. p. 3202-6.

11. Muhoza P, Danovaro-Holliday MC, Diallo MS, Murphy P, Sodha SV, Requejo JH, et al. Routine vaccination coverage—Worldwide, 2020. Morbidity and Mortality Weekly Report. 2021;70(43):1495.

12. Wei E, Reisinger A, Li J, French LE, Clanner-Engelshofen B, Reinholz M. Integration of scRNA-Seq and TCGA RNA-Seq to Analyze the Heterogeneity of HPV+ and HPV- Cervical Cancer Immune Cells and Establish Molecular Risk Models. Front Oncol. 2022;12:860900.

13. Westrich JA, Warren CJ, Pyeon D. Evasion of host immune defenses by human papillomavirus. Virus Res. 2017;231:21-33.

14. Burd EM. Human papillomavirus and cervical cancer. Clin Microbiol Rev. 2003;16(1):1-17.

15. Johnson DE, Burtness B, Leemans CR, Lui VWY, Bauman JE, Grandis JR. Head and neck squamous cell carcinoma. Nat Rev Dis Primers. 2020;6(1):92.

16. Lowy DR, Schiller JT. Reducing HPV-associated cancer globally. Cancer prevention research. 2012;5(1):18-23.

17. Institute NC. Cancer stat facts: oral cavity and pharynx cancer. 2019.

18. Rosenblatt A, de Campos Guidi HG. Human papillomavirus: A practical guide for urologists: Springer Science & Business Media; 2009.

19. Grulich AE, Poynten IM, Machalek DA, Jin F, Templeton DJ, Hillman RJ. The epidemiology of anal cancer. Sex Health. 2012;9(6):504-8.

20. von Knebel Doeberitz M, Prigge ES. Role of DNA methylation in HPV associated lesions. Papillomavirus Res. 2019;7:180-3.

21. Au Yeung CL, Tsang WP, Tsang TY, Co NN, Yau PL, Kwok TT. HPV-16 E6 upregulation of DNMT1 through repression of tumor suppressor p53. Oncol Rep. 2010;24(6):1599-604.

22. Burgers WA, Blanchon L, Pradhan S, de Launoit Y, Kouzarides T, Fuks F. Viral oncoproteins target the DNA methyltransferases. Oncogene. 2007;26(11):1650-5.

23. Angarica VE, Del Sol A. Bioinformatics Tools for Genome-Wide Epigenetic Research. Adv Exp Med Biol. 2017;978:489-512.

24. Konig IR, Fuchs O, Hansen G, von Mutius E, Kopp MV. What is precision medicine? Eur Respir J. 2017;50(4).

25. Zhou N, Bao J. FerrDb: a manually curated resource for regulators and markers of ferroptosis and ferroptosis-disease associations. Database (Oxford). 2020;2020.

26. Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, et al. PathCards: multi-source consolidation of human biological pathways. Database (Oxford). 2015;2015.

27. International Cancer Genome C, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, et al. International network of cancer genome projects. Nature. 2010;464(7291):993-8.

28. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29.

29. Seth S, Mallik S, Bhadra T, Zhao Z. Dimensionality Reduction and Louvain Agglomerative Hierarchical Clustering for Cluster-Specified Frequent Biomarker Discovery in Single-Cell Sequencing Data. Front Genet. 2022;13:828479.

30. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888-902 e21.

31. Aran D, Looney AP, Liu L, Wu E, Fong V, Hsu A, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nat Immunol. 2019;20(2):163-72.

32. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496-502.

33. Hastie T, Tibshirani R, Narasimhan B, Chu G. impute: impute: Imputation for microarray data. R package version; 2018.

34. Yoshihara K, Shahmoradgoli M, Martinez E, Vegesna R, Kim H, Torres-Garcia W, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612.

35. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453-7.

36. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol. 2018;1711:243-59.

37. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.

38. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17.

39. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med. 1997;16(4):385-95.

40. Therneau TM, editor Extending the Cox model. Proceedings of the First Seattle symposium in biostatistics; 1997: Springer.

41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25-9.

42. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27-30.

43. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun. 2019;10(1):1523.

44. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 2009;37(Database issue):D412-6.

45. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics. 2003;4:2.

46. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12(2):109-16.

47. Zou Y, Palte MJ, Deik AA, Li H, Eaton JK, Wang W, et al. A GPX4-dependent cancer cell state underlies the clear-cell morphology and confers sensitivity to ferroptosis. Nat Commun. 2019;10(1):1617.

48. Li C, Tang Z, Zhang W, Ye Z, Liu F. GEPIA2021: integrating multiple deconvolution-based analysis into GEPIA. Nucleic Acids Res. 2021;49(W1):W242-W6.

49. Tian Y, Morris TJ, Webster AP, Yang Z, Beck S, Feber A, et al. ChAMP: updated methylation analysis pipeline for Illumina BeadChips. Bioinformatics. 2017;33(24):3982-4.

50. Wu H, Gonzalez Villalobos R, Yao X, Reilly D, Chen T, Rankin M, et al. Mapping the single-cell transcriptomic response of murine diabetic kidney disease to therapies. Cell Metab. 2022;34(7):1064-78 e6.

51. Morris GM, Huey R, Olson AJ. Using AutoDock for ligand-receptor docking. Curr Protoc Bioinformatics. 2008;Chapter 8:Unit 8 14.

52. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. J Chem Inf Model. 2021;61(8):3891-8.

53. Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, et al. PubChem BioAssay: 2017 update. Nucleic Acids Res. 2017;45(D1):D955-D63.

54. Feltham JL, Dotsch V, Raza S, Manor D, Cerione RA, Sutcliffe MJ, et al. Definition of the switch surface in the solution structure of Cdc42Hs. Biochemistry. 1997;36(29):8755-66.

55. Heppner DE, Gunther M, Wittlinger F, Laufer SA, Eck MJ. Structural Basis for EGFR Mutant Inhibition by Trisubstituted Imidazole Inhibitors. J Med Chem. 2020;63(8):4293-305.

56. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics. 2014;30(12):1771-3.

57. Ha BH, Boggon TJ. CDC42 binds PAK4 via an extended GTPase-effector interface. Proc Natl Acad Sci U S A. 2018;115(3):531-6.

58. Solca F, Dahl G, Zoephel A, Bader G, Sanderson M, Klein C, et al. Target binding properties and cellular activity of afatinib (BIBW 2992), an irreversible ErbB family blocker. J Pharmacol Exp Ther. 2012;343(2):342-50.

59. Munger K, Phelps WC, Bubb V, Howley PM, Schlegel R. The E6 and E7 genes of the human papillomavirus type 16 together are necessary and sufficient for transformation of primary human keratinocytes. J Virol. 1989;63(10):4417-21.

60. Cruz-Gregorio A, Aranda-Rivera AK, Ortega-Lozano AJ, Pedraza-Chaverri J, Mendoza-Hoffmann F. Lipid metabolism and oxidative stress in HPV-related cancers. Free Radic Biol Med. 2021;172:226-36.

61. Does H, Does H. Genital HPV Infection–CDC Fact Sheet.

62. Vici P, Mariani L, Pizzuti L, Sergi D, Di Lauro L, Vizza E, et al. Immunologic treatments for precancerous lesions and uterine cervical cancer. J Exp Clin Cancer Res. 2014;33:29.

63. Jorgensen KR, Jensen JB. Human papillomavirus and urinary bladder cancer revisited. APMIS. 2020;128(2):72-9.

64. Khoury JD, Tannir NM, Williams MD, Chen Y, Yao H, Zhang J, et al. Landscape of DNA virus associations across human malignant cancers: analysis of 3,775 cases using RNA-Seq. J Virol. 2013;87(16):8916-26.

65. Park K, Cho KJ, Lee M, Yoon DH, Kim SB. Importance of FOXP3 in prognosis and its relationship with p16 in tonsillar squamous cell carcinoma. Anticancer Res. 2013;33(12):5667-73.

66. Molling JW, de Gruijl TD, Glim J, Moreno M, Rozendaal L, Meijer CJ, et al. CD4(+)CD25hi regulatory T-cell frequency correlates with persistence of human papillomavirus type 16 and T helper cell responses in patients with cervical intraepithelial neoplasia. Int J Cancer. 2007;121(8):1749-55.

67. Zuo T, Wang L, Morrison C, Chang X, Zhang H, Li W, et al. FOXP3 is an X-linked breast cancer suppressor gene and an important repressor of the HER-2/ErbB2 oncogene. Cell. 2007;129(7):1275-86.

68. Li W, Wang L, Katoh H, Liu R, Zheng P, Liu Y. Identification of a tumor suppressor relay between the FOXP3 and the Hippo pathways in breast and prostate cancers. Cancer Res. 2011;71(6):2162-71.

69. Cezar-Dos-Santos F, Ferreira RS, Okuyama NCM, Trugilo KP, Sena MM, Pereira ER, et al. FOXP3 immunoregulatory gene variants are independent predictors of human papillomavirus infection and cervical cancer precursor lesions. J Cancer Res Clin Oncol. 2019;145(8):2013-25.

70. Leonard WJ, O'Shea JJ. Jaks and STATs: biological implications. Annu Rev Immunol. 1998;16:293-322.

71. Laffort C, Le Deist F, Favre M, Caillat-Zucman S, Radford-Weiss I, Debre M, et al. Severe cutaneous papillomavirus disease after haemopoietic stem-cell transplantation in patients with severe combined immune deficiency caused by common gammac cytokine receptor subunit or JAK-3 deficiency. Lancet. 2004;363(9426):2051-4.

72. Dogan S, Xu B, Middha S, Vanderbilt CM, Bowman AS, Migliacci J, et al. Identification of prognostic molecular biomarkers in 157 HPV-positive and HPV-negative squamous cell carcinomas of the oropharynx. Int J Cancer. 2019;145(11):3152-62.

73. Regauer S, Reich O, Kashofer K. Cervical Precancers Originate From Infected Proliferating Reserve Cells: A Comparative Histologic and Genetic Study of Thin and Thick High-grade Squamous Intraepithelial Lesions. Am J Surg Pathol. 2022;46(4):519-27.

74. Rusan M, Li YY, Hammerman PS. Genomic landscape of human papillomavirus-associated cancers. Clin Cancer Res. 2015;21(9):2009-19.

75. Gleber-Netto FO, Rao X, Guo T, Xi Y, Gao M, Shen L, et al. Variations in HPV function are associated with survival in squamous cell carcinoma. JCI Insight. 2019;4(1).

76. Liao LC, Hu B, Zhang SP. Macrophages participate in the immunosuppression of condyloma acuminatum through the PD-1/PD-L1 signaling pathway. J Chin Med Assoc. 2019;82(5):413-8.

77. Tiemessen MM, Jagger AL, Evans HG, van Herwijnen MJ, John S, Taams LS. CD4+CD25+Foxp3+ regulatory T cells induce alternative activation of human monocytes/macrophages. Proc Natl Acad Sci U S A. 2007;104(49):19446-51.

78. Saadeh D, Kurban M, Abbas O. Plasmacytoid dendritic cells and type I interferon in the immunological response against warts. Clin Exp Dermatol. 2017;42(8):857-62.

79. Bernhard MC, Zwick A, Mohr T, Gasparoni G, Khalmurzaev O, Matveev VB, et al. The HPV and p63 Status in Penile Cancer Are Linked with the Infiltration and Therapeutic Availability of Neutrophils. Mol Cancer Ther. 2021;20(2):423-37.

80. Al-Sahaf S, Hendawi NB, Ollington B, Bolt R, Ottewell PD, Hunter KD, et al. Increased Abundance of Tumour-Associated Neutrophils in HPV-Negative Compared to HPV-Positive Oropharyngeal Squamous Cell Carcinoma Is Mediated by IL-1R Signalling. Front Oral Health. 2021;2:604565.

81. Wieland A, Patel MR, Cardenas MA, Eberhardt CS, Hudson WH, Obeng RC, et al. Defining HPV-specific B cell responses in patients with head and neck cancer. Nature. 2021;597(7875):274-8.

82. Lisco A, Hsu AP, Dimitrova D, Proctor DM, Mace EM, Ye P, et al. Treatment of Relapsing HPV Diseases by Restored Function of Natural Killer Cells. N Engl J Med. 2021;385(10):921-9.

83. Gutierrez-Hoya A, Soto-Cruz I. NK Cell Regulation in Cervical Cancer and Strategies for Immunotherapy. Cells. 2021;10(11).

84. Li C, Hua K. Dissecting the Single-Cell Transcriptome Network of Immune Environment Underlying Cervical Premalignant Lesion, Cervical Cancer and Metastatic Lymph Nodes. Front Immunol. 2022;13:897366.

85. De Marco F. Oxidative stress and HPV carcinogenesis. Viruses. 2013;5(2):708-31.

86. Georgescu SR, Mitran CI, Mitran MI, Caruntu C, Sarbu MI, Matei C, et al. New Insights in the Pathogenesis of HPV Infection and the Associated Carcinogenic Processes: The Role of Chronic Inflammation and Oxidative Stress. J Immunol Res. 2018;2018:5315816.

87. Wang T, Gong M, Cao Y, Zhao C, Lu Y, Zhou Y, et al. Persistent ferroptosis promotes cervical squamous intraepithelial lesion development and oncogenesis by regulating KRAS expression in patients with high risk-HPV infection. Cell Death Discov. 2022;8(1):201.

88. Ma W, Tummers B, van Esch EM, Goedemans R, Melief CJ, Meyers C, et al. Human Papillomavirus Downregulates the Expression of IFITM1 and RIPK3 to Escape from IFNgamma- and TNFalpha-Mediated Antiproliferative Effects and Necroptosis. Front Immunol. 2016;7:496.

89. Li L, Yu S, Zang C. Low Necroptosis Process Predicts Poor Treatment Outcome of Human Papillomavirus Positive Cervical Cancers by Decreasing Tumor-Associated Macrophages M1 Polarization. Gynecol Obstet Invest. 2018;83(3):259-67.

90. Totten SP, Im YK, Cepeda Canedo E, Najyb O, Nguyen A, Hebert S, et al. STAT1 potentiates oxidative stress revealing a targetable vulnerability that increases phenformin efficacy in breast cancer. Nat Commun. 2021;12(1):3299.

91. Ksiazek K, Mikula-Pietrasik J, Catar R, Dworacki G, Winckiewicz M, Frydrychowicz M, et al. Oxidative stress-dependent increase in ICAM-1 expression promotes adhesion of colorectal and pancreatic cancers to the senescent peritoneal mesothelium. Int J Cancer. 2010;127(2):293-303.

92. Kuroda J, Ago T, Matsushima S, Zhai P, Schneider MD, Sadoshima J. NADPH oxidase 4 (Nox4) is a major source of oxidative stress in the failing heart. Proc Natl Acad Sci U S A. 2010;107(35):15565-70.

93. Fleming AM, Zhu J, Ding Y, Burrows CJ. Location dependence of the transcriptional response of a potential G-quadruplex in gene promoters under oxidative stress. Nucleic Acids Res. 2019;47(10):5049-60.

94. Yang L, Froio RM, Sciuto TE, Dvorak AM, Alon R, Luscinskas FW. ICAM-1 regulates neutrophil adhesion and transcellular migration of TNF-alpha-activated vascular endothelium under flow. Blood. 2005;106(2):584-92.

95. Quigley M, Huang X, Yang Y. STAT1 signaling in CD8 T cells is required for their clonal expansion and memory formation following viral infection in vivo. J Immunol. 2008;180(4):2158-64.

96. Chen J, Yanuck RR, 3rd, Abbondanzo SL, Chu WS, Aguilera NS. c-Kit (CD117) reactivity in extramedullary myeloid tumor/granulocytic sarcoma. Arch Pathol Lab Med. 2001;125(11):1448-52.

97. Colucci F, Di Santo JP. The receptor tyrosine kinase c-kit provides a critical signal for survival, expansion, and maturation of mouse natural killer cells. Blood. 2000;95(3):984-91.

98. Xiao G, Zhuang W, Wang T, Lian G, Luo L, Ye C, et al. Transcriptomic analysis identifies Toll-like and Nod-like pathways and necroptosis in pulmonary arterial hypertension. J Cell Mol Med. 2020;24(19):11409-21.

99. Chiang SK, Chen SE, Chang LC. The Role of HO-1 and Its Crosstalk with Oxidative Stress in Cancer Cell Survival. Cells. 2021;10(9).

100.Egawa N, Doorbar J. The low-risk papillomaviruses. Virus Res. 2017;231:119-27.

101.Schierbeck J, Vestergaard T, Bygum A. Skin Cancer Associated Genodermatoses: A Literature Review. Acta Derm Venereol. 2019;99(4):360-9.

102.Moreno V, Garcia-Carbonero R, Maurel J, Feliu J. Phase 1 study of cetuximab in combination with 5-fluorouracil, cisplatin, and radiotherapy in patients with locally advanced anal canal carcinoma. Cancer. 2014;120(3):454-6.

103.Deutsch E, Lemanski C, Pignon JP, Levy A, Delarochefordiere A, Martel-Lafay I, et al. Unexpected toxicity of cetuximab combined with conventional chemoradiotherapy in patients with locally advanced anal cancer: results of the UNICANCER ACCORD 16 phase II trial. Ann Oncol. 2013;24(11):2834-8.

104.Stengel K, Zheng Y. Cdc42 in oncogenic transformation, invasion, and tumorigenesis. Cell Signal. 2011;23(9):1415-23.

105.Oliver AW, He X, Borthwick K, Donne AJ, Hampson L, Hampson IN. The HPV16 E6 binding protein Tip-1 interacts with ARHGEF16, which activates Cdc42. Br J Cancer. 2011;104(2):324-31.

106.Murphy NP, Mott HR, Owen D. Progress in the therapeutic inhibition of Cdc42 signalling. Biochem Soc Trans. 2021;49(3):1443-56.

107.Wang XY, Gan MX, Li Y, Zhan WH, Han TY, Han XJ, et al. Cdc42 induces EGF receptor protein accumulation and promotes EGF receptor nuclear transport and cellular transformation. FEBS Lett. 2015;589(2):255-62.

108.Smirnova IS, Chang S, Forsthuber TG. Prosurvival and proapoptotic functions of ERK1/2 activation in murine thymocytes in vitro. Cell Immunol. 2010;261(1):29-36.

109.Vavala T. Role of afatinib in the treatment of advanced lung squamous cell carcinoma. Clin Pharmacol. 2017;9:147-57.

110.Chen Q, Huang Y, Shao L, Han-Zhang H, Yang F, Wang Y, et al. An EGFR-Amplified Cervical Squamous Cell Carcinoma Patient with Pulmonary Metastasis Benefits from Afatinib: A Case Report. Onco Targets Ther. 2020;13:1845-9.

Acknowledgements

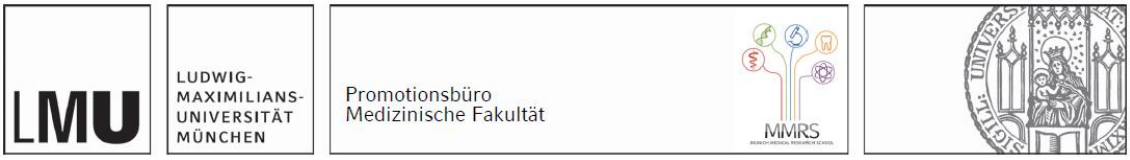
I am indebted to my dear motherland and China Scholarship Council.

I would like to thank Prof. Dr. Markus Reinholz for the opportunity to do my MD project at the Department of Dermatology and Allergology of the Ludwig-Maximilians-University, his always inspiration, and encouragement (and delicious coffee and cookies) to help me complete this work. Likewise, these thanks go to Dr. Benjamin Clanner-engelshoven for the great cooperation, guidance, and support in experiment. Our time smoking on the roof was unforgettable. Special thanks also go to Prof. Dr. med. Lars French, who has always supported this work with great interest. Communication with the knowledgeable professor always gave me new insights.

A thank you also goes to lab colleagues who always supported me with words and deeds, provided amusement and an outstanding atmosphere in the lab, named: Jiahua Li, Philipp Anand, Tobias Nellessen, Luka Ständer, Amin Reisinger, Dr. Takashi Satoh, Claudia Kammerbauer. More than a lab colleague, Philipp Anand and Jiahua Li are my best friends I made in Munich, I really appreciate the beers we had together. Special mention also goes to my friend and roommate Hao Yu and Huijin Yang, it was a great two years of being together. Wish them to finish their degrees and get rich soon.

Last but not least, I would also like to thank my family, my friends, willing to endure long distances apart for me to pursue my doctor degree.

Affidavit



Affidavit

Wei Erdong
_____
Surname, first name

_____
Street

_____
Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Heterogeneity analysis of low-risk HPV infection and high-risk HPV infection, HPV-positive and HPV-negative cancers

……

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

MINNEAPOLIS,20.11.2023__
place, date

___Erdong Wei_____
Signature doctoral candidate