Cao et al. BMC Cancer (2025) 25:105
https://doi.org/10.1186/s12885-025-13483-8

BMC Cancer

RESEARCH Open Access

Prognostic values of intracellular cell-related genes in esophageal cancer and their regulatory mechanisms

Wei Cao[1,2,3†], Dacheng Jin[2,3†], Weirun Min[1,2,3], Haochi Li[1,2,3], Rong Wang[1], Jinlong Zhang[1,2,3] and Yunjiu Gou[2,3*]

Abstract

Esophageal cancer is a grave malignant condition. While radiotherapy, often in conjunction with chemotherapy, serves as a cornerstone in the management of locally advanced or metastatic cases, patient tolerance and treatment resistance frequently hinder its efficacy. Cell-in-cell structures, prevalent in various tumors, have been linked to prognosis. Hence, investigating the prognostic significance and regulatory mechanisms of genes related to these intracellular structures in esophageal cancer is imperative. The Cancer Genome Atlas (TCGA) Esophageal Cancer (ESCA) dataset served as the training set for the analysis. Differentially expressed genes (DEGs) in ESCA samples were identified, with those related to intercellular structures designated cell-in-cell-related differential expression genes (CIC-related DEGs). Cox regression analysis was employed to identify prognostic genes, categorizing samples into high- and low-risk groups based on median risk scores. Validation was conducted using the GSE53624 risk model. Established methodologies included morphological mapping, enrichment analysis, immune infiltration analysis, prognostic gene expression validation, molecular docking, and Reverse Transcription Polymerase Chain Reaction (RT-PCR) validation. Thirty-eight intersecting genes were identified between the disease and normal groups in ESCA samples. Stepwise multivariate Cox analysis pinpointed three prognostic genes: androgen receptor (AR), C-X-C motif chemokine ligand 8 (CXCL8), and epidermal growth factor receptor (EGFR). The risk model's applicability was confirmed in the GSE53624 dataset, revealing eight significantly different immune-related gene sets. Prognostic gene expression validation demonstrated significant differences between the disease and normal groups in both datasets. The proteins corresponding to the three prognostic genes interacted with gefitinib and osimertinib. RT-PCR results corroborated the differential expression of prognostic genes in esophageal cancer tissues. This study identified AR, CXCL8, and EGFR as prognostic genes and demonstrated their molecular interactions with gefitinib and osimertinib, providing a foundation for ESCA diagnosis and treatment.

Keywords Esophageal cancer, Bioinformatics, Cell-in-cell, Molecular docking, Prognostic risk model

[†]Wei Cao and Dacheng Jin contributed equally to this work.

*Correspondence:
Yunjiu Gou
gouyunjiu@163.com
[1]First Clinical Medical School, Gansu University of Chinese Medicine, Lanzhou, China
[2]Chest Clinic Center, Gansu Provincial People's Hospital, Lanzhou, China
[3]First Department of Thoracic Surgery, Gansu Provincial People's Hospital, Lanzhou, China

Introduction

Esophageal cancer (ESCA) ranks as the seventh most common cancer and the sixth leading cause of cancer-related deaths globally [1]. Despite recent advances in comprehensive treatments, the prognosis for esophageal cancer remains dismal [2]. Radiotherapy, a standard treatment modality for ESCA, effectively kills cancer cells, alleviates symptoms, and extends patient survival [3]. Although radiotherapy plays an integral role in the multimodal treatment approach for esophageal cancer, treatment-related toxicity, and patient intolerance can affect its effectiveness, especially when used in conjunction with chemotherapy [4]. Perioperative or palliative radiotherapy often leads to treatment failure and presents a significant challenge in the clinical management of patients with ESCA [5]. Therefore, understanding the molecular mechanisms underpinning radiotherapy tolerance in esophageal cancer and identifying molecular targets for radiotherapy sensitization to develop new therapeutic strategies is imperative.

Cell-in-cell (CIC) structures, wherein one or more living cells internalize into another living cell, forming a "bird's eye cell," have been observed in various tumors and are linked to poor prognosis in cancers such as breast, lung, and pancreatic cancer [6, 7]. Assessing CIC status is an effective method for evaluating prognosis [8]. The formation of CIC structures may be influenced by various factors, including cell adhesion molecules, cytoskeletal reorganization, and the regulation of apoptotic pathways [9]. In some cases, CIC structures may lead to programmed death of phagocytic cells, thereby inhibiting tumor growth; in other cases, they may help tumor cells evade immune recognition and promote tumor progression [10]. Therefore, CIC structures are not only an interesting phenomenon in the process of cancer development but also may become a key indicator for predicting cancer patients' prognosis and guiding therapeutic strategies. Understanding the mechanisms of CIC structures in esophageal cancer is of great clinical significance for developing new therapeutic approaches. This study used the cancer genome atlas (TCGA) ESCA dataset as the training set and differentially expressed gene analysis was used to screen differentially expressed genes related to the CIC structure. The prognostic value of these genes was then analyzed using the Cox proportional hazards model. Subsequently, we validated the applicability of the risk model in the dataset and explored the mechanisms of these prognostic genes in esophageal cancer through multidimensional means, including enrichment analysis, immune infiltration analysis, and Q-reverse transcription polymerase chain reaction (RT-PCR).

Materials and methods

Data extraction

In this study, the TCGA-ESCA RNAseq - HTSeq - Counts matrix was downloaded from the University of California, Santa Cruz (UCSC) Xena database (https://xenabrowser.net/datapages/) to serve as the training set. Concurrently, the corresponding clinical phenotype and survival information data were obtained. The dataset, downloaded in December 2022, comprised 173 samples: 162 disease samples, 11 normal control samples, and 161 disease samples with survival information.

The GSE53624 dataset was downloaded from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), containing transcriptome data of 238 samples, including 119 disease samples (with corresponding survival information) and 119 normal samples, used as a validation set. Additionally, 101 cell-to-cell related genes were sourced from existing literature [6].

Differential expression analysis between disease group and normal group

To identify genes with significant differences in gene expression between the different sample groups, the R software package "DESeq2" (version 3.52.4) [11] was utilized to analyze the differentially expressed genes (DEGs) between ESCA samples and normal group samples in the training set, with screening criteria of $|\log2FoldChange| > 0.5$ and P-value < 0.05. Using the online jvenn tool (http://jvenn.toulouse.inra.fr/app/example.html), DEGs obtained from the analysis were intersected with the 101 cell-related genes sourced from literature. The differentially expressed genes related to cell-in-cell structures were then identified as cell-in-cell-related differential expression genes (CIC-related DEGs).

Risk model construction and prognostic gene analysis

Univariate Cox was performed to obtain signature genes (pvalu < 0.2) for CIC-related DEGs in the disease group of the training set. Subsequent stepwise forward multifactorial Cox analyses were performed, and the step function was used to process the feature genes to obtain prognostic genes (direction = "both"). In this process, the akaike information criterion (AIC) was used as a criterion for model selection, and the current optimal model configuration was determined by continuously adding or removing variables to find the regression equation with the smallest AIC value ($K = 2$). Afterwards, correlations between prognostic genes were calculated using the spearman method and risk models were constructed. We used normalised gene expression data to assess the prognostic value of risk models. Expression levels of three prognostic genes were obtained in the TCGA-ESCA dataset, and DESeqDataSet objects were created using the DESeqDataSetFromMatrix function in the DESeq2

package (version 3.52.4) [11]. Next the low expressed genes were removed and the size factor for each sample was estimated using the estimateSizeFactors function for normalisation and then the counts function was used to return the normalised count data. Finally, the risk score for each patient was calculated using the following formula:

$$\mathbf{Riskscore_{sample}} = \sum_{n=1}^{n} (\boldsymbol{coef}_i * \boldsymbol{x}_i)$$

(Note: Coefi represents the multivariate regression coefficient of the ith gene, xi represents the expression value of the ith gene, and n represents the number of model genes.)

Patients were stratified into high- and low-risk groups based on the median risk score, followed by Kaplan-Meier (K-M) survival curve analysis to compare survival between these groups. The accuracy of the survival curve scoring model was validated using the receiver operating characteristic (ROC) curve, generated by survival ROC package (version 1.42.0) [12]. Additionally, risk curves and expression heat maps of prognostic genes in the high- and low-risk groups were created. The results were further validated using the GSE53624 dataset.

Construction of nomogram model for prognostic factors

Based on the risk score, age, gender, and tumor-node-mMetastasis (TNM) stage, we used the "RMS" package (version 6.0–1) [13] to construct line plots of 1-year and 3-year survival rates for the clinical factors in the risk model. The "pROC" package (version 1.18.0) [14] was used to plot the ROC curves at 1 and 3 years, and the "rmda" package (version 1.6) (https://CRAN.R-project.org/package=rmda) was used to plot the decision curves. The closer the slope of the calibration curve is to 1, the more accurate the prediction will be.

Enrichment analysis of high- and low-risk groups

Using the grouping information for high- and low-risk groups, DESeq2 software (version 3.52.4) [11] analyzed the DEGs between these groups in the training set, with a screening threshold of |log2FoldChange| > 0.5 and P-value < 0.05. The ClusterProfiler package (version 4.4.4) [15] conducted Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses of the DEGs, while the R package "ggplot2" (version 3.3.2) [16] visualized the enrichment results.

Analysis of immune infiltration in the high- and low-risk groups

In this study, we used the gene set file immune-v29.gmt containing 29 immune cell types, which provides information about the gene sets associated with each immune cell. The cell types are: B cells naive, B cells memory, plasma cells, T cells CD8, T cells CD4 naive, T cells CD4 memory resting, T cells CD4 memory activated, T cells follicular helper, T cells regulatory (Tregs), T cells gamma delta, NK cells resting, natural killer (NK) cells activated, monocytes, macrophages M0, macrophages M1, macrophages M2, dendritic cells resting, dendritic cells activated, mast cells resting, mast cells activated, eosinophils, and neutrophils.

We utilized the getGmt function in the GSEABase package (version 1.64.0) (https://doi.org/10.18129/B9.bioc.GSEABase) to extract gene set information from the immune-v29.gmt file. This step was performed in preparation for gene set enrichment analysis (GSEA) for subsequent immune infiltration analysis.

Based on the categorical information of the high- and low-risk groups, we used the single-sample gene set enrichment analysis (ssGSEA) method in the training set to analyze the immune infiltration between the two groups. The ssGSEA is a modified GSEA method that allows for gene set enrichment analysis of individual samples to assess the activity levels of each immune cell type in different risk groups. With this approach, we looked at the composition of immune cells in the training set and identified immune cell types that were differentially expressed between the high- and low-risk groups. In addition to ssGSEA analysis, we further quantified the relative abundance of immune cells using the CIBERSORT algorithm, which accurately estimates the composition and distribution of different cell types from complex mixed tissue samples through non-negative matrix factorization (NMF) and linear regression techniques that utilize known cell type-specific gene expression patterns. The CIBERSORT algorithm provides a powerful tool to gain insight into immune cell infiltration in the tumor microenvironment, which is important for unraveling disease mechanisms and evaluating treatment response.

Prognostic gene expression validation

Expression matrices of prognostic genes were extracted from the training and validation sets, which were subsequently normalised. Then, expression analyses were performed and t-tests were used to compare the two sets of samples: one group of tumor samples and the other group of control samples. The main purpose of this comparison is to determine whether there are significant differences in the expression levels of specific prognostic genes during tumor development and progression. Through this differential expression analysis, we aim to identify genes that may play a key role in disease prognosis.

Table 1 Primers and their sequences

Primers	Sequences
AR F	AGGCAGTGTCGGTGTCCATG
AR R	CCTTTGGTGTAACCTCCCTTGA
CXCL8 F	TCTGCAGCTCTGTGTGAAGG
CXCL8 R	TTCTCAGCCCTCTTCAAAAACT
EGFR F	GCCAAGGCACGAGTAACAAGC
EGFR R	AGGGCAATGAGGACATAACCAG
GAPDH F	CGAAGGTGGAGTCAACGGATTT
GAPDH R	ATGGGTGGAATCATATTGGAAC

Molecular docking analysis

The crystal structures of the prognostic genes were obtained from the protein data bank (PDB) database (https://www.rcsb.org/), and molecular docking with gefitinib and osimertinib was executed using AutoDock software (version 4.2) [17].

Reverse transcription polymerase chain reaction (RT-PCR)

Five pairs of frozen esophageal cancer samples, along with five pairs of cancer tissues and five pairs of adjacent tissues, were collected from the Department of Thoracic Surgery at Gansu Provincial People's Hospital. From each sample, 50 mg of tissue was used for total ribonucleic acid (RNA) extraction. The SureScript First-Strand complementary deoxyribonucleic acid (cDNA) Synthesis Kit facilitated reverse transcription and detection. Relative RNA quantities were calculated using the comparative cycle threshold (Ct) method ($2^{-\triangle\triangle CT}$). The primers and their sequences are detailed in Table 1.

Statistical analysis

Statistical analyses and graphical representations were conducted using R software (version 4.2.0) and GraphPad Prism (version 8.0). Data were expressed as mean ± standard deviation, and differences were evaluated using the Wilcoxon rank-sum test. Multiple corrections were not performed in order to maintain high sensitivity and not miss signals that might be biologically significant. A P-value of less than 0.05 was considered statistically significant.

Results

Identification of CIC-related DEGs

The distribution of DEGs is illustrated in Fig. 1A and B. A total of 12,985 DEGs were identified, comprising 6,617

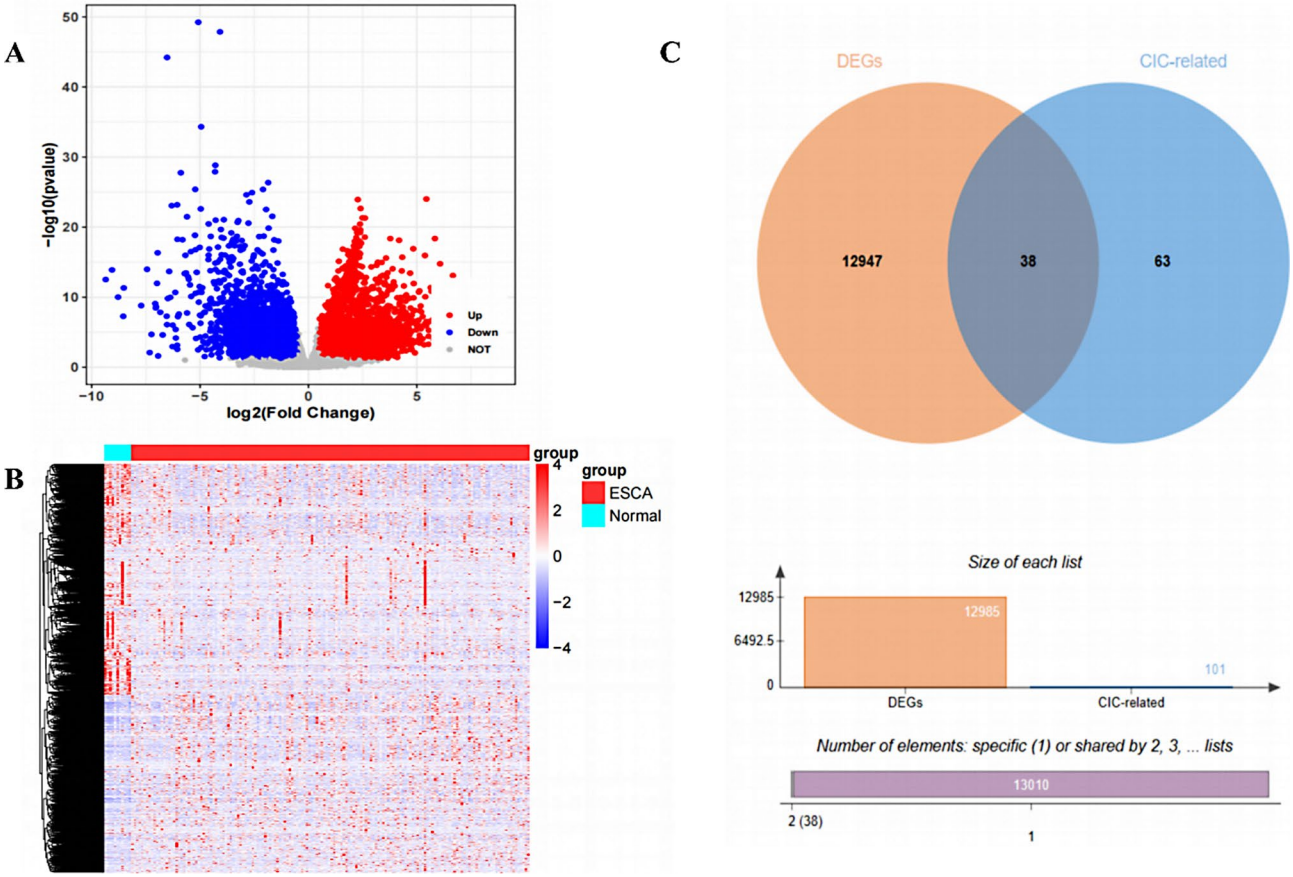

Fig. 1 Analysis of differentially expressed genes. ((**a**) Volcano plot of differentially expressed genes. (**b**) Heat map of DEGs between sample groups. (**c**) Venn diagram showing the intersection of DEGs and related genes between cells

up-regulated and 6,368 down-regulated genes. Taking the intersection of DEGs and 101 cell-related genes, a total of 38 genes were identified as CIC-related DEGs (Figue. 1 C, Supplementary Table 1).

Construction of risk models

Univariate COX regression analysis

The TCGA-ESCA dataset was utilized as the training set to perform univariate Cox proportional hazards regression analysis on 38 CIC-related DEGs, with a univariate P-value threshold of 0.2, identifying a total of five genes. The results and corresponding forest plots are presented in Table 2; Fig. 2A, respectively.

Multivariate COX regression analysis

Three prognostic genes (androgen receptor (AR), C-X-C motif chemokine ligand 8 (CXCL8), and epidermal growth factor receptor (EGFR)) were identified through stepwise multivariate Cox analysis using the step function. The multivariate coefficient (coef) for each prognostic gene was calculated to construct the survival risk model. The multivariate results and corresponding forest plots are displayed in Table 3; Fig. 2a-b, respectively.

In the ESCA group of the training set, the expression matrix of the prognostic genes (AR, CXCL8, EGFR) was extracted. The correlation between these prognostic genes was calculated using the "spearman" method, and a correlation chord diagram was generated using the R software "circlize" package (version 0.4.15) [18]. The results are illustrated in Fig. 2c.

Effect of prognostic genes on survival

The risk value for each patient was calculated, and 161 patients with survival information in the training set were stratified into high- and low-risk groups based on the median risk value (1.047788). The high-risk group consisted of 80 samples, while the low-risk group included 81 samples. As shown in Fig. 2d, survival analysis revealed a significant difference in survival between the high- and low-risk groups in the TCGA-ESCA dataset ($P < 0.05$).

ROC curve and risk curve

The multivariate Cox model was employed to calculate the RiskScore. Using the survival ROC package (version 1.42.0) [12], false positive and true positive rates were computed, and the ROC curve was plotted based on these results. The area under curve (AUC) was calculated, as depicted in Fig. 2e, with AUC values of 0.74 for 1-year and 0.62 for 3-year survival, demonstrating the efficacy of the risk regression model as a prognostic tool. The risk curve, shown in Fig. 2g, stratifies samples into high- and low-risk groups based on the median value.

Table 2 Results of univariate analysis

id	pvalue	HR	HR.95 L	HR.95 H
EGFR	0.025465674	0.816222106	0.683036621	0.975377463
AR	0.032093467	0.816018744	0.677565483	0.982763448
CXCL8	0.051096995	1.154841387	0.999313898	1.334574283
HTT	0.06904872	0.699816164	0.476290936	1.028242669
TP53	0.164783007	0.867781691	0.710418302	1.060002342

Expression heatmap of prognostic genes in the high- and low-risk groups

The R package "pheatmap" (version 1.0.12) (https://rdrr.io/cran/pheatmap/) was utilized to visualize the expression levels of prognostic genes in the high- and low-risk groups. The resulting heatmap is displayed in Fig. 2f.

Survival risk model validation

Effect of prognostic genes on survival

The risk value for each patient was calculated, and 119 patients in the training set were stratified into high- and low-risk groups based on the median risk value (0.9625565), resulting in 59 samples in the high-risk group and 60 samples in the low-risk group. The survival analysis results for these groups are presented in Fig. 3a, demonstrating a significant survival difference between the high- and low-risk groups in the GSE53624 dataset ($P < 0.05$).

ROC curve and risk curve

The multivariate Cox model was employed to calculate the RiskScore. Using the survival ROC package (version 1.42.0) [12], false positive and true positive rates were computed, and these results were used to plot the ROC curve. The AUC values, shown in Fig. 3b, were both above 0.6 for 1-year and 3-year survival, indicating the efficacy of the constructed risk regression model as a prognostic tool. As illustrated in Fig. 3c, samples were stratified into high- and low-risk groups based on the median value.

Expression heatmap of prognostic genes in the high- and low-risk groups

The R package "pheatmap" (version 1.0.12) (https://rdrr.io/cran/pheatmap/) was utilized to visualize the expression levels of prognostic genes in the high- and low-risk groups. The resulting heatmap is displayed in Fig. 3d.

Construction of nomogram model for prognostic factors

Using the "RMS" package (version 6.0–1) [13], survival nomograms for 1 and 3 years were constructed based on risk score, age, sex, and TNM stage. Each factor corresponds to a score, and the total score is the sum of these individual scores. The 1-year and 3-year survival rates were predicted based on the total score, with higher scores indicating lower survival rates. The nomogram (Fig. 4a) and calibration curve (Fig. 4b) were developed.

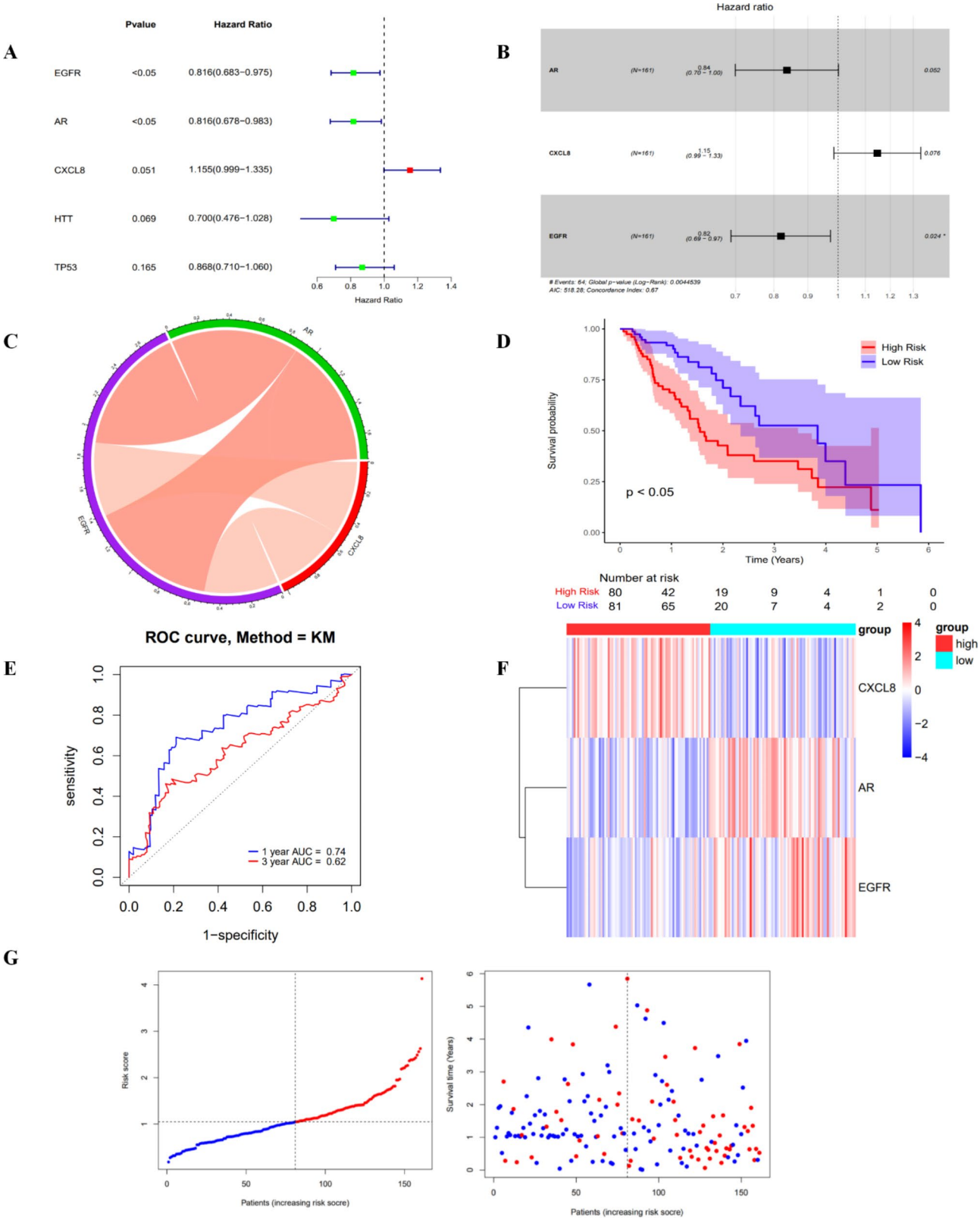

Fig. 2 Construction of the risk model (**a**) Forest plot of univariate COX results. (**b**) Forest plot of multivariate COX results. (**c**) Chord diagram depicting the correlation between prognostic genes, with line color indicating the strength of the correlation (redder lines denote stronger positive correlations). (**d**) K-M survival curve of RiskScore, indicating a significant difference ($P < 0.05$) between the high-risk and low-risk groups. (**e**) ROC curve evaluating the effectiveness of the risk model. (**f**) Heatmap showing the expression of prognostic genes between high- and low-risk groups. (g**g**) Risk curve of the validation set for high- and low-risk groups

Table 3 Results of multivariate analysis

id	coef	HR	HR.95 L	HR.95 H	pvalue
AR	-0.178187505	0.836785507	0.699135927	1.001536265	0.051989207
CXCL8	0.137065624	1.14690341	0.985768463	1.334377678	0.075996366
EGFR	-0.199620941	0.819041159	0.688642196	0.97413203	0.024059355

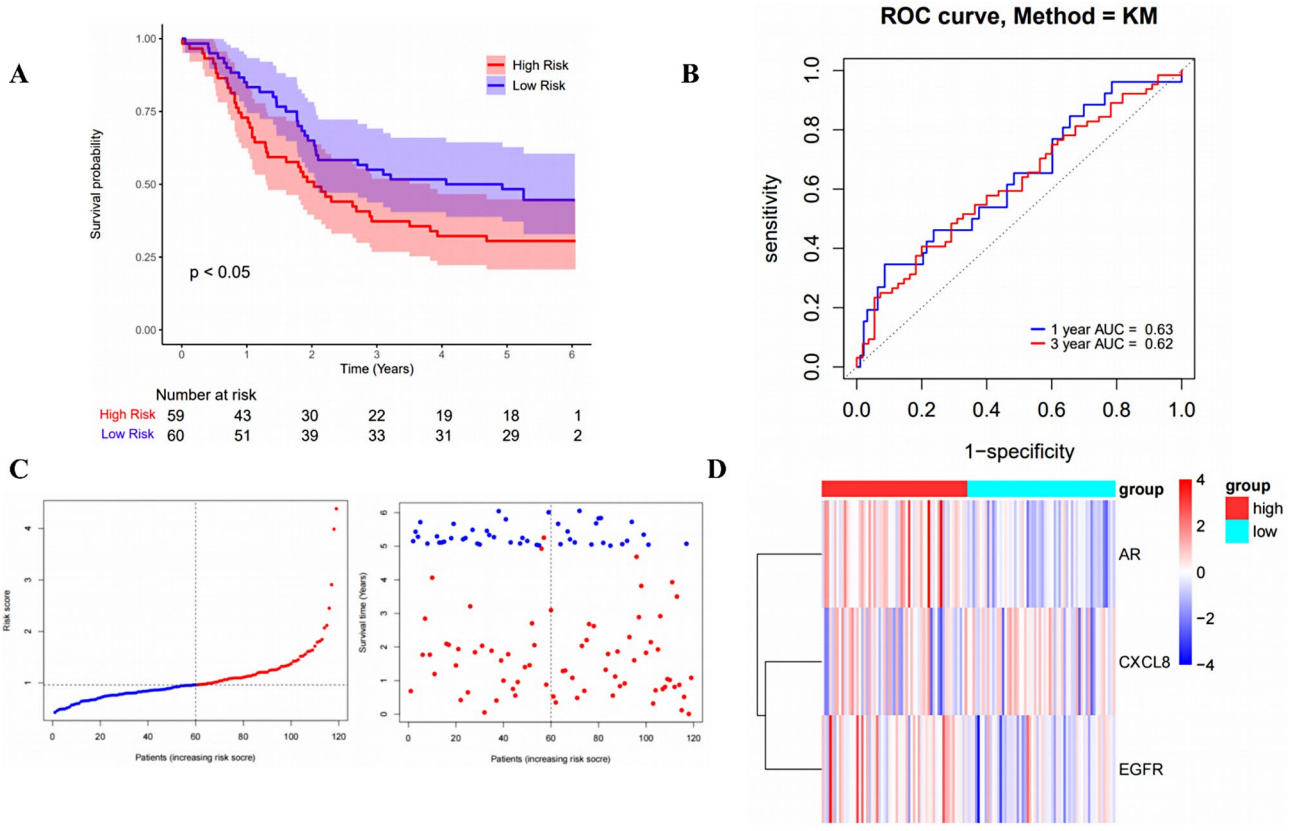

Fig. 3 Survival risk model validation. (**a**) Kaplan-Meier (**K-M**) survival curve of RiskScore in the validation set. The P-value < 0.05 indicates a significant difference between the high- and low-risk groups. (**b**) ROC curve of the validation set, used to assess the effectiveness of the risk model. (**c**) Risk curve of the validation set, depicting high- and low-risk groups. (d) Expression heatmap of prognostic genes in the high- and low-risk groups

The 1-year and 3-year ROC curves of the nomogram were generated using the "pROC" package (version 1.18.0) [14], which showed that the AUC values for the 1-year and 3-year ROC curves of the nomogram were both greater than 0.7, underscoring the model's high prognostic accuracy (Fig. 4c). The decision curve analysis (DCA) was performed using the "rmda" package (version 1.6) (https://CRAN.R-project.org/package=rmda), which revealed that within the high-risk threshold range of 0–1, the nomogram model provided significant clinical benefit, surpassing the clinical benefits of individual factors such as "age," "riskScore," "T," "N," "M," and "gender" (Fig. 4d). The calibration curve demonstrated a minimal error between the actual and predicted risks, indicating high predictive accuracy of the nomogram model for ESCA.

Enrichment analysis of high- and low-risk groups

Based on the expression matrix in the TCGA-ESCA dataset, 80 high-risk samples and 81 low-risk samples were extracted. The "DESeq2" package (version 3.52.4) [11] was employed for differential expression analysis with screening criteria set at |log2FoldChange| > 0.5 and P-value < 0.05. This analysis identified 5,727 significantly differentially expressed genes between the high- and low-risk groups, including 2,965 up-regulated and 2,762 down-regulated genes. Figure 5a displays the volcano plot, illustrating the distribution of these DEGs, while Fig. 5b presents a heat map depicting their expression across sample groups.

The ClusterProfiler package (version 4.4.4) [15] in R was used to perform functional enrichment analyses (GO and KEGG) of the DEGs between the high- and low-risk groups. The screening thresholds were set at P-value < 0.05, identifying a total of 887 GO Biological

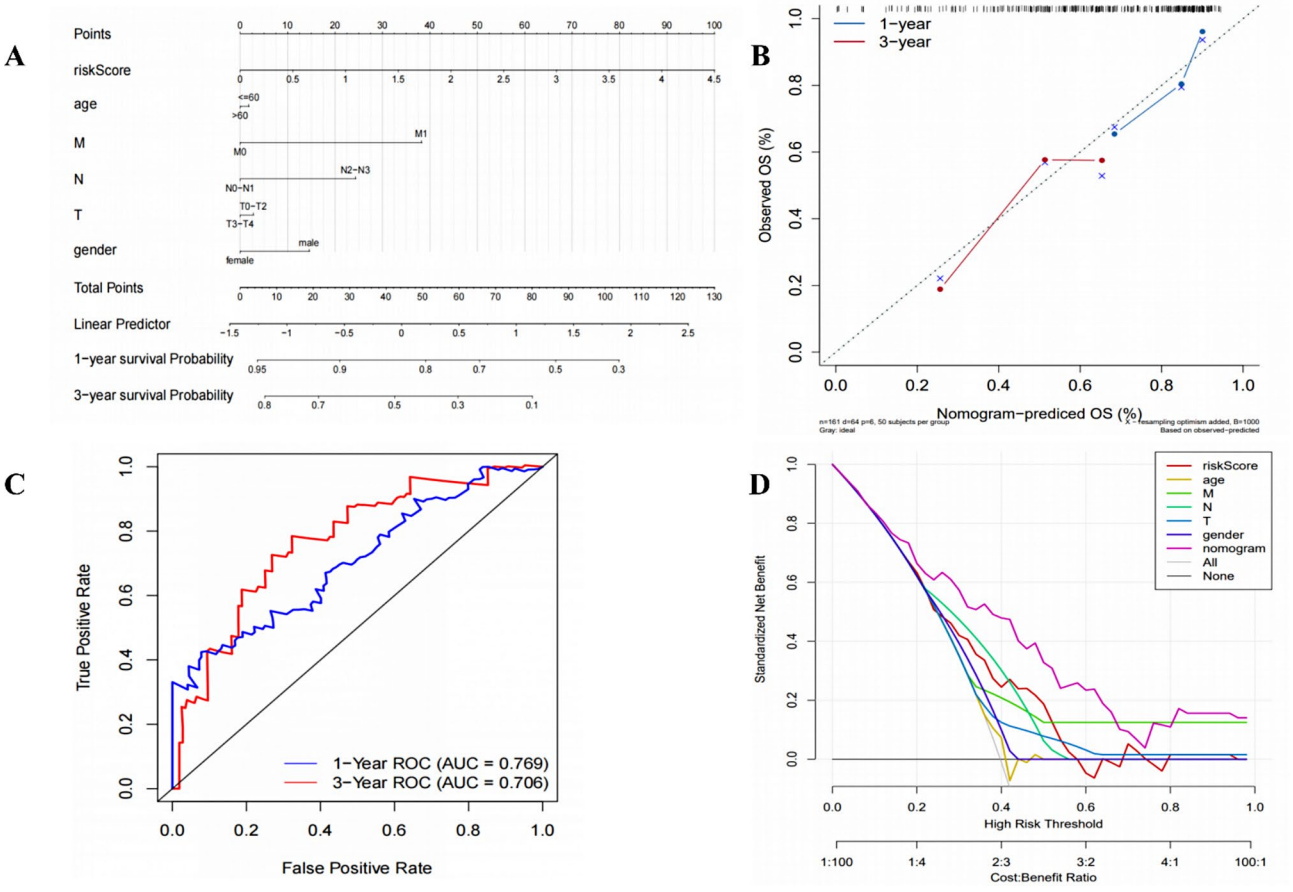

Fig. 4 Construction of a nomogram model for prognostic factors. (**a**) Nomogram model incorporating clinical factors. (**b**) Calibration curve evaluating the predictive accuracy of the nomogram model. A slope closer to 1 indicates higher predictive accuracy. (**c**) ROC curves depicting 1-year and 3-year survival rates as predicted by the nomogram model. (**d**) Decision Curve Analysis (DCA) curves evaluating the clinical applicability of the nomogram model. The nomogram curve surpasses the gray line, as well as the "age," "riskScore," "T," "N," "M," and "gender" curves, indicating the nomogram model's benefit within the high-risk threshold range of 0–1

Processes (BP), 67 GO Cellular Components (CC), 178 GO Molecular Functions (MF), and 53 KEGG pathways. The enrichment results were visualized using the R package "ggplot2," (version 3.3.2) [16] as shown in Fig. 5c and d. The top 10 descriptions of GO and KEGG were selected for display, ranked by P-value.

Analysis of immune infiltration in high- and low-risk groups

Based on the expression matrix and grouping information of high- and low-risk groups in the training set (TCGA-ESCA), ssGSEA was performed to determine the expression levels of 29 immune gene sets (comprising 16 immune cells and 13 immune-related pathways) in each sample. The t-test revealed eight significantly different immune-related gene sets in the training set: B cells, iDCs, Macrophages, Mast cells, NK cells, T cell co-inhibition, T helper cells, and Type II IFN Response. The results are depicted in Fig. 6a.

Prognostic gene expression validation

The expression matrices of the prognostic genes (AR, CXCL8, EGFR) were extracted from both the training set (TCGA-ESCA) and the validation set (GSE53624). Utilizing the grouping information from each dataset, a t-test was performed. The R software package "ggplot2" (version 3.3.2) [16] was employed to create boxplots of the prognostic gene expressions in both datasets. Significant differences were observed between the disease group and the normal group for the three prognostic genes in both datasets, with consistent trends in the expression of the remaining two genes, except for EGFR. The results are depicted in Fig. 6b and c.

Molecular docking analysis

Docking analysis was conducted for three prognostic genes (AR, CXCL8, EGFR) with two drugs (gefitinib and osimertinib). The crystal structures for AR, CXCL8, and EGFR were retrieved from the PDB database (https://ww

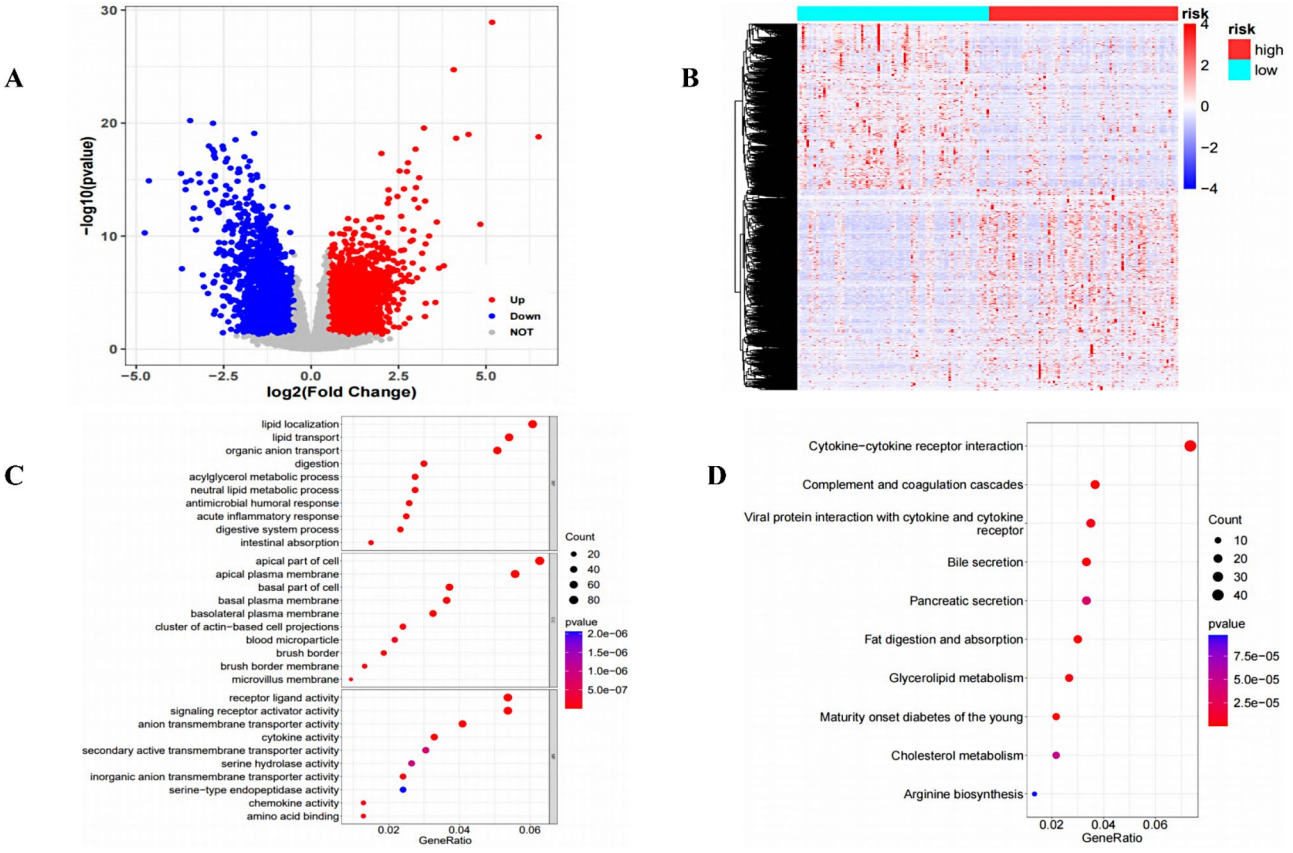

Fig. 5 Enrichment analysis of high- and low-risk groups. (**c**) Volcano plot of DEGs between high- and low-risk groups in the training set. (**b**) Heat map of DEGss between high- and low-risk groups in the training set. (**c**) Bubble plot of GO TOP10 enrichment results of DEGs between high- and low-risk groups. The vertical axis represents the enriched GO terms, while the horizontal axis shows the GeneRatio, which indicates the proportion of genes related to each term among the total DEGs. (**d**) Bubble plot of KEGGTOP10 enrichment results of DEGs between high- and low-risk groups. The ordinate represents the enriched KEGG pathways, and the abscissa shows the GeneRatio

w.rcsb.org/) with PDB IDs AR-1T7R, CXCL8-7JNY, and EGFR-1IP0, respectively.

Molecular docking using AutoDock revealed specific interactions: [1] Docking between AR-1T7R and gefitinib formed two hydrogen bonds and involved two residues; [2] Docking between AR-1T7R and osimertinib formed one hydrogen bond and involved one residue; [3] Docking between CXCL8-7JNY and gefitinib formed three hydrogen bonds and involved two residues; [4] Docking between CXCL8-7JNY and osimertinib formed three hydrogen bonds and involved two residues; [5] Docking between EGFR-1IP0 and gefitinib formed two hydrogen bonds and involved two residues; [6] Docking between EGFR-1IP0 and osimertinib formed four hydrogen bonds and involved four residues.

All three prognostic genes exhibited interactions with both drugs. Figure 7a and b illustrate the docking results for EGFR-1IP0 with gefitinib and osimertinib, respectively, while additional molecular docking results are provided in the supplementary material.

RT-PCR

Five pairs of frozen esophageal cancer samples were obtained from the Department of Thoracic Surgery, Gansu Provincial People's Hospital, for RT-PCR experiments. As shown in Fig. 8a-c, AR expression was significantly lower in esophageal cancer tissues ($P<0.0001$), while CXCL8 and EGFR expressions were significantly higher ($P<0.05$). These RT-PCR results were consistent with previous analyses.

Discussion

Esophageal cancer arises from the abnormal proliferation of esophageal squamous or glandular epithelium and is among the most prevalent malignant tumors worldwide [19]. It ranks sixth in cancer mortality [20]. The early symptoms of esophageal cancer are often subtle, leading to diagnosis at advanced stages; 38% of cases are diagnosed in the late stage, missing the optimal treatment window. Currently, the 5-year survival rate for patients with esophageal cancer stands at 10% [21]. Hence, understanding the mechanisms underlying esophageal cancer development and exploring new treatment methods

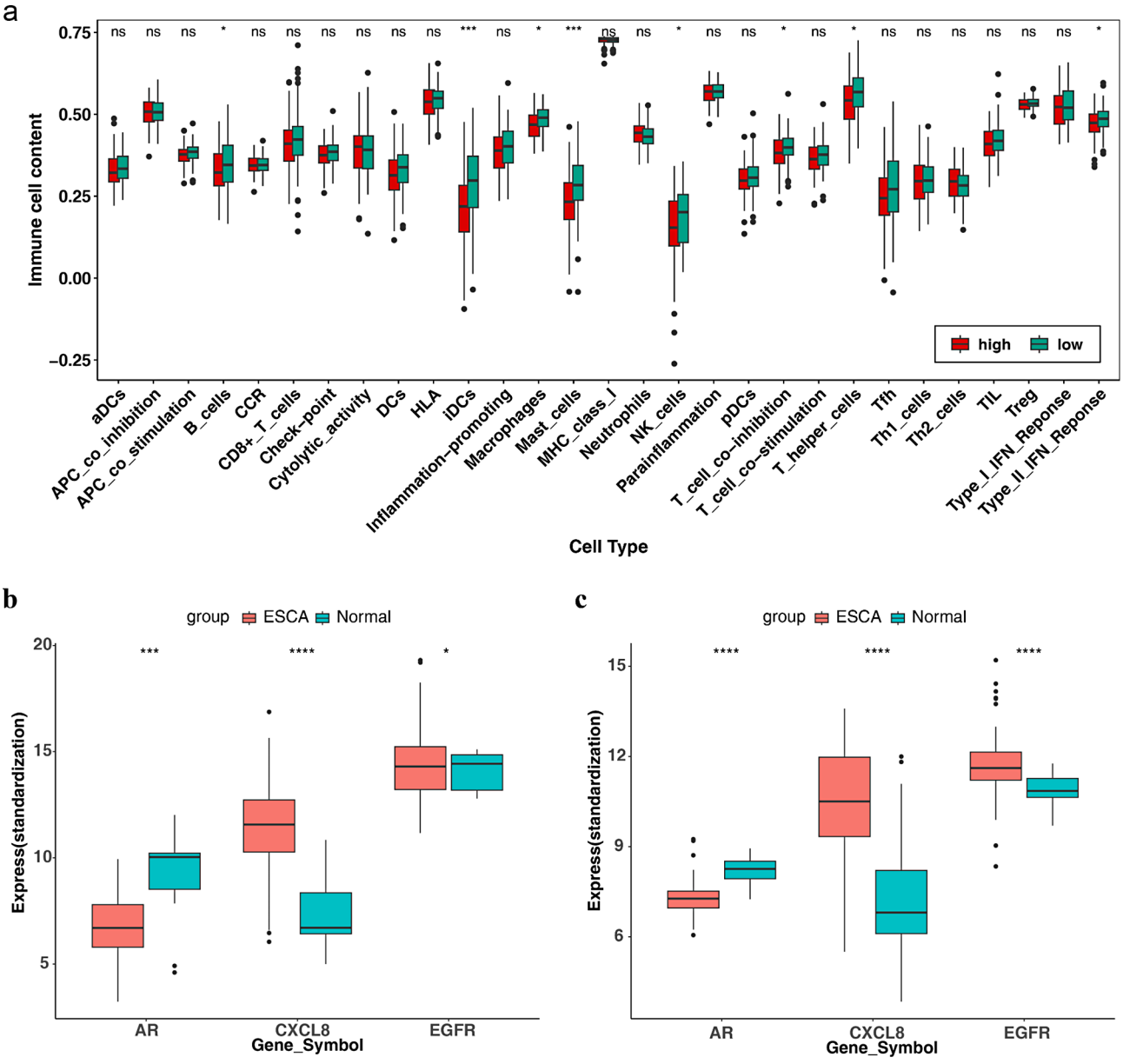

Fig. 6 Immune infiltration analysis and prognostic gene expression visualization. (**a**) Box plot illustrating the abundance of 29 immune gene sets in the high- and low-risk groups of the training set. The vertical axis denotes the infiltration level, while the horizontal axis represents the immune cell gene sets. (**b**) Box plot showing the expression levels of prognostic genes in the training set. (**c**) Box plot showing the expression levels of prognostic genes in the validation set. (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$)

is essential. During tumor cell culture, some large cells with internal vacuoles and multiple nuclei, indicating one cell inside another (cell-in-cell structures), have been observed in various tumors and shown to correlate with prognosis [8]. Studying the prognostic value and regulatory mechanisms of intracellular structure-related genes in esophageal cancer is therefore essential [22]. Cell-in-cell (CIC) structures have been noted in various tumor types and are associated with poor prognosis. For instance, a clinical study linked CIC structures with aggressive features and cancer-related mortality in tongue cancer [23]. However, the role of CIC structures in esophageal cancer remains unclear.

The GDC TCGA Esophageal Cancer (ESCA) dataset was utilized as the training set for this study, while the GSE53624 dataset was downloaded from the GEO database and used as the validation set. A total of 101 cell-related genes were sourced from existing literature. Based on these data, a series of bioinformatics analyses were conducted.

First, differential expression analysis was performed between the disease and normal groups. The intersection

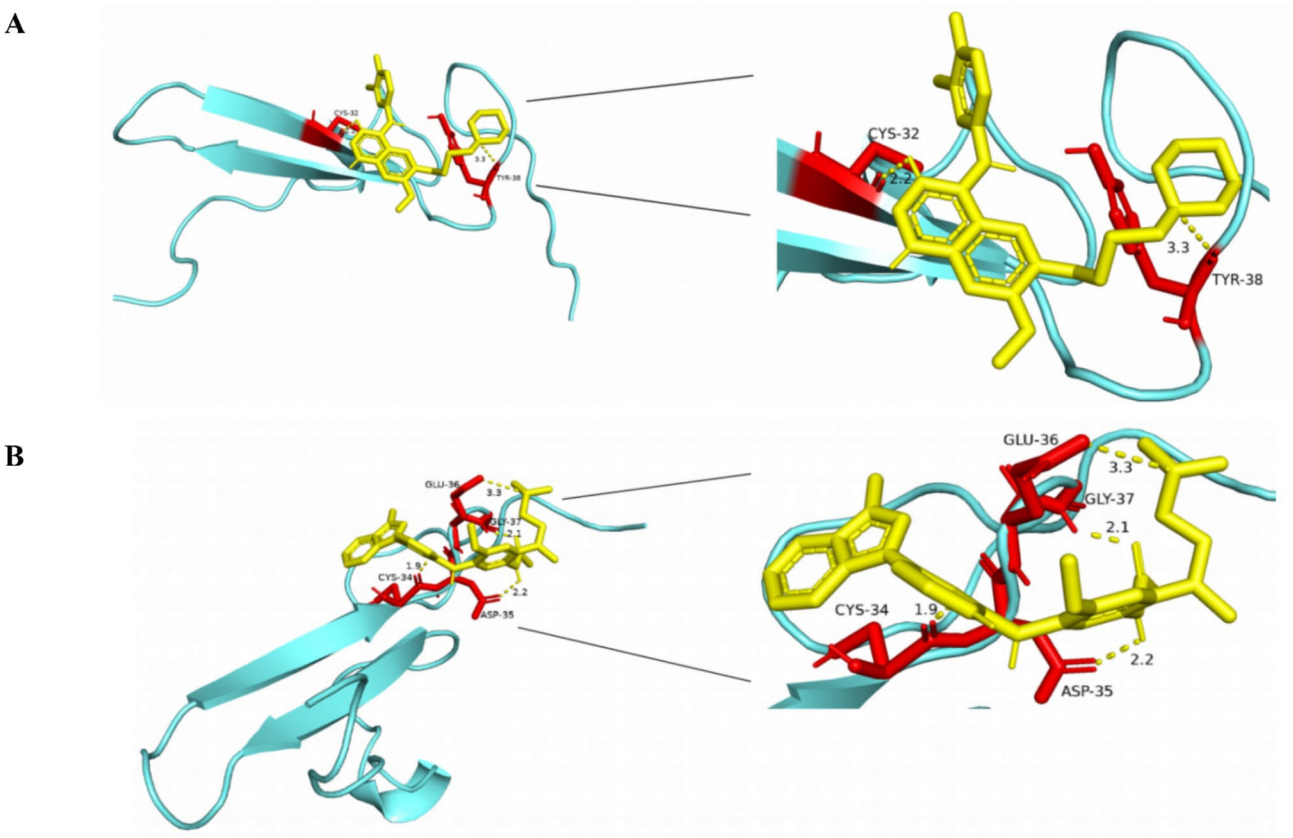

Fig. 7 Molecular docking analysis. (**a**) Docking results of 1IP0 with gefitinib. (**b**) Docking results of 1IP0 with Osimertinib. In the figure, the cyan ring model represents the active molecule 1IP0. The stick structures near 1IP0 indicate amin o acid residues with hydrogen bond interactions with the active molecule. The yellow dotted lines depict the hydrogen bonds formed between the active molecule and the amino acid residues, while the red parts highlight the amino acid residues

of the DEGs and the 101 cell-related genes yielded 38 intersection genes, designated as CIC-related DEGs. These 38 genes underwent univariate and multivariate analyses, identifying three prognostic genes: AR, CXCL8, and EGFR. This study employed the TCGA esophageal carcinoma (ESCA) dataset as the training set and downloaded the GSE53624 dataset from the GEO database as the validation set. Three prognostic genes related to intercellular communication (AR, CXCL8, and EGFR) were screened out through univariate and stepwise multivariate Cox regression analyses, and a prognostic model for esophageal carcinoma was constructed. Subsequently, based on the risk scores of patients in the prognostic model, they were classified into high-risk and low-risk groups. Immune infiltration analysis was carried out to investigate changes such as the distribution of immune cells within different risk groups. Additionally, functional enrichment analysis was performed to offer novel insights and grounds for the prognosis judgment and treatment strategy of esophageal carcinoma. Currently, multivariate Cox regression analysis is one of the frequently utilized approaches in constructing prognostic models [24–26], featuring strong applicability and the ability to effectively handle cases with missing survival time data, rendering the analysis results more reliable. In this study, the stepwise multivariate Cox regression analysis was adopted. It can automatically select the independent variables that have a significant impact on survival time and to a certain extent, reduce the influence of multicollinearity, thereby enhancing the predictive performance of the prognostic model and providing robust scientific evidence and new perspectives for the prognosis determination and treatment strategy formulation of esophageal carcinoma.

AR is a type I nuclear receptor related to cell classification, differentiation, apoptosis, proliferation, and angiogenesis [27, 28]. Studies have pointed out that during the process of prostate tissue transforming from a benign state to a malignant one, along with the structural and genomic changes in stromal cells, the expression level of AR shows a gradual downward trend [29]. Furthermore, when exploring the role of AR in breast cancer, researchers found that AR is expressed in normal breast tissue, but its expression level gradually decreases during the transition from ductal carcinoma in situ (DCIS) to invasive cancer as the disease progresses [30]. Currently, research on AR in cancer mainly focuses on prostate

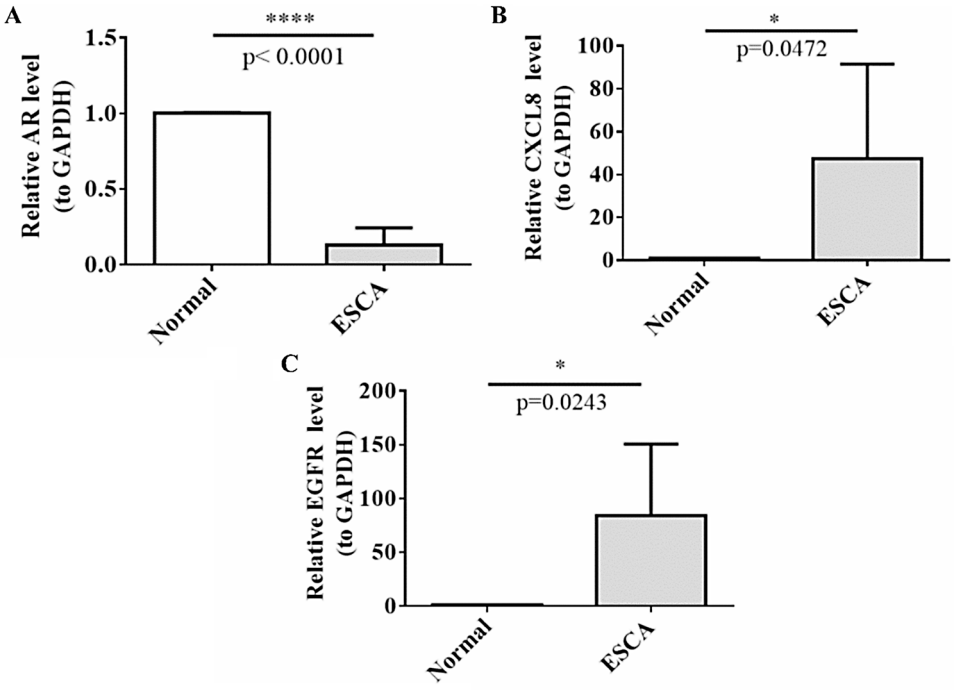

Fig. 8 RT-PCR results. (**a**) AR expression level in cancer tissues and adjacent tissues (**b**) CXCL8 expression level in cancer tissues and adjacent tissues (**c**) EGFR expression level in cancer tissues and adjacent tissues (*$P < 0.05$, **** $P < 0.0001$)

cancer [31, 32], while studies on AR in esophageal cancer are still scarce. The multivariate Cox analysis in this study revealed that AR is a protective factor for ESCA, and its expression level in ESCA patients is significantly lower than that in the control group. This finding provides new clues for our understanding of the role of AR in ESCA, but its specific mechanism of action in ESCA still requires further research to be revealed.

CXCL8 and EGFR were significantly up-regulated in esophageal cancer. CXCL8, also known as interleukin-8 (IL-8), acts as a neutrophil-activating factor and is primarily produced by neutrophils, monocytes, macrophages, T cells, epithelial cells, and endothelial cells. This gene shows a similar expression trend in ovarian, breast, and liver cancers [33–35] and head and neck squamous cell carcinoma and holds considerable prognostic and diagnostic value [36].

Silencing CXCL8 in esophageal cancer cells can inhibit their proliferation and invasion [37]. Research by Helen et al. revealed that CXCL8 expression is closely linked to epithelial-mesenchymal transition and cell-matrix vascularization, promoting tumor cell invasion and metastasis, thereby explaining CXCL8's predictive ability for tumor metastasis [38].

EGFR predominantly resides in the stroma, epidermis, and certain smooth muscle cells. Upon activation, EGFR can initiate the downstream mitogen-activated protein kinase (MAPK) signaling pathway, inhibit cell apoptosis and DNA repair, accelerate cell invasion [39], and contribute to the progression of lung, colorectal, and breast cancers [40–42]. The study discovered that EGFR is a crucial downstream target of NAT10. Overexpression of EGFR in ESCA cells with NAT10 deficiency could partially restore the migration and invasion capabilities of NAT10-deficient ESCA cells and promote the oncogenic function of NAT10 to facilitate ESCA progression [43]. In this study, it was found that the expression of EGFR significantly increased in ESCA patients. Therefore, we can infer that the increased expression of EGFR might promote the disease progression of ESCA by enhancing the migration and invasion abilities of ESCA cells.

In this study, 29 immune-related gene sets were selected. Immune infiltration analysis of high- and low-risk groups identified eight significantly different immune-related gene sets in the training set: B cells, iDCs, Macrophages, Mast cells, NK cells, T cell co-inhibition, T helper cells, and Type II IFN Response. IDCs exist in peripheral tissues and are key targets for stimulating tumor immunity, specializing in antigen capture [44, 45], and are related to the tolerance and induction of regulatory T cells (Tregs) [46]. This study discovered that the infiltration of iDCs cells decreased in the high-risk group, which might weaken the body's ability to recognize and present tumor-related antigens, thereby influencing the anti-tumor effect of the immune system. Hence, enhancing the function of iDCs or promoting their infiltration into tumor sites might offer significant insights for developing novel tumor immunotherapy strategies.

In recent years, molecular docking methods have emerged as a pivotal technique in computer-aided drug research. Molecular docking facilitates drug design by analyzing the characteristics of receptors and their interactions with drug molecules, employing theoretical simulation to study molecular interactions (such as ligands and receptors) and predict their binding modes and affinities [47].

AutoDock, an open-source molecular simulation software, is primarily utilized for ligand-protein molecular docking. It employs a semi-flexible docking method that allows conformational changes in small molecules, using binding free energy as the basis for evaluating docking results [48]. Molecular docking of the three prognostic genes with gefitinib and osimertinib revealed interactions, suggesting that these genes may influence drug efficacy or that the drugs may affect gene expression.

Osimertinib, a third-generation EGFRr-tyrosine kinase inhibitor (TKI), is used in the treatment of lung cancer [49]. However, its use can lead to severe complications, including lung injury and interstitial lung disease (ILD), potentially exacerbating pulmonary inflammation due to impaired epithelial healing [50]. Gefitinib, a specific small molecule EGFR tyrosine kinase inhibitor, inhibits esophageal cancer cell growth by blocking EGFR tyrosine kinase signal transduction and phosphorylation. It also impedes tumor angiogenesis and metastasis, serving as a therapeutic agent for various cancers [51]. Gefitinib's inhibitory effects on metastatic non-small cell lung cancer, ovarian cancer, and other tumors are well-documented [52]. Our study discovered that gefitinib and osimertinib not only interact with EGFR but also with AR and CXCL8, which may have significant implications for future targeted drug development.

Conclusion

In this study, bioinformatics analysis of the TCGA-ESCA transcriptome data identified three prognostic genes: AR, CXCL8, and EGFR, which were further validated by RT-PCR. Molecular docking of these prognostic genes with gefitinib and osimertinib revealed significant interactions, suggesting that further attention and in-depth studies are essential to understanding the pathogenesis of the disease and enhancing therapeutic outcomes. Consequently, this research provides a valuable reference for the future diagnosis, mechanism exploration, and treatment of esophageal cancer. However, Owing to the imbalance of disease and control sample sizes in the dataset, in future studies, multiple covariates (such as age, gender, disease stage, etc.) need to be incorporated in the DESeq2 analysis stage for adjustment, in the hope of reducing confounding effects and revealing the differences in gene expression between disease and control samples more accurately.Additionally, the impact of the "cell-in-cell" structure on the biological behavior of esophageal cancer cells and the body's immune response requires confirmation through in vivo studies. Finally, the molecular docking results need further validation through subsequent related experiments.

Abbreviations

ESCA	Esophageal cancer
CIC	Cell-in-cell
DEGs	Differentially expressed genes
K-M	Kaplan-Meier curve
ROC	Receiver operating characteristic
KEGG	Kyoto Encyclopedia of Genes and Genomes
GO	Gene Ontology Enrichment Analysis
AUC	Area Under Curve
CIC-related genes	Cell-in-cell-related genes
TCGA	The Cancer Genome Atlas
GEO	Gene Expression Omnibus
CIC-related DEGs	Cell-in-cell-related differential expression genes
Tregs	T cells regulatory
NMF	Non-negative matrix factorization
RT-PCR	Reverse Transcription Polymerase Chain Reaction

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13483-8.

Supplementary Material 1

Acknowledgements
Not Applicable.

Author contributions
W.D. and W.H. wrote the main manuscript text and R.J. prepared Figs. 1, 2, 3, 4, 5, 6, 7 and 8. Y. was responsible for the study design. All authors reviewed the manuscript.

Funding
This study was supported by grants from the Gansu Provincial People's Hospital. Research Funding (22GSSYD-30; 22GSSYD-25, 22GSSYC-9), Gansu Youth Science and Technology Fund (22JR11RA241), Gansu province health committee project (GSWSKY2020-50), and Science and Technology Department of Gansu Province (22YF7FA095).

Data availability
The data that supports the findings of this study are available in the supplementary material of this article.

Declarations

Ethics approval and consent to participate
The authors confirm that all methods were conducted according to the principles of the Declaration of Helsinki and were approved by the Gansu Province People's Hospital Ethics Committee for examination and approval. Written informed consents have been obtained from all subjects.

Consent for publication
Not Applicable.

Competing interests
The authors declare no competing interests.

Received: 22 May 2024 / Accepted: 9 January 2025
Published online: 20 January 2025

References

1. Lin EW, Karakasheva TA, Hicks PD, Bass AJ, Rustgi AK. The tumor microenvironment in esophageal cancer. Oncogene. 2016;35(41):5337–49.
2. Wang M, Sun X, Xin H, Wen Z, Cheng Y. SPP1 promotes radiation resistance through JAK2/STAT3 pathway in esophageal carcinoma. Cancer Med. 2022;11(23):4526–43.
3. Matsuda S, Takeuchi M, Kawakubo H, Kitagawa Y. Lymph node metastatic patterns and the development of multidisciplinary treatment for esophageal cancer. Dis Esophagus. 2023;36(4):doad006.
4. Sardaro A, Ferrari C, Carbonara R, Altini C, Lavelli V, Rubini G. Synergism between Immunotherapy and Radiotherapy in Esophageal Cancer: an overview of current knowledge and future perspectives. Cancer Biother Radiopharm. 2021;36(2):123–32.
5. Shitara K, Ajani JA, Moehler M, Garrido M, Gallardo C, Shen L, et al. Nivolumab plus chemotherapy or ipilimumab in gastro-oesophageal cancer. Nature. 2022;603(7903):942–8.
6. Song J, Ruze R, Chen Y, Xu R, Yin X, Wang C, et al. Construction of a novel model based on cell-in-cell-related genes and validation of KRT7 as a biomarker for predicting survival and immune microenvironment in pancreatic cancer. BMC Cancer. 2022;22(1):894.
7. Song J, Xu R, Zhang H, Xue X, Ruze R, Chen Y, et al. Cell-in-cell-mediated entosis reveals a progressive mechanism in pancreatic Cancer. Gastroenterology. 2023;165(6):1505–e152120.
8. Almangush A, Mäkitie AA, Hagström J, Haglund C, Kowalski LP, Nieminen P, et al. Cell-in-cell phenomenon associates with aggressive characteristics and cancer-related mortality in early oral tongue cancer. BMC Cancer. 2020;20(1):843.
9. de Oliveira GAP, Rangel LP, Costa DC, Silva JL. Misfolding, aggregation, and disordered segments in c-Abl and p53 in Human Cancer. Front Oncol. 2015;5:97.
10. Kerr JF, Wyllie AH, Currie AR. Apoptosis: a basic biological phenomenon with wide-ranging implications in tissue kinetics. Br J Cancer. 1972;26(4):239–57.
11. Love MI, Huber W, Anders S. Moderated estimation of Fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.
12. Heagerty PJ, Lumley T, Pepe MS. Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics. 2000;56(2):337–44.
13. Levinson N. The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction[M].1946.
14. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S + to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.
15. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16(5):284–7.
16. Wickham H, Chang W. RStudio. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. In Book of Abstracts, 2016.
17. Forli S, Botta M. Lennard-Jones potential and dummy atom settings to overcome the AUTODOCK limitation in treating flexible ring systems. J Chem Inf Model. 2007;47(4):1481–92.
18. Gu Z, Gu L, Eils R, Schlesner M, Brors B. Circlize implements and enhances circular visualization in R. Bioinformatics. 2014;30(19):2811–2.
19. Rogers JE, Sewastjanow-Silva M, Waters RE, Ajani JA. Esophageal cancer: emerging therapeutics. Expert Opin Ther Targets. 2022;26(2):107–17.
20. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. CA Cancer J Clin. 2022;72(1):7–33.
21. Huang FL, Yu SJ. Esophageal cancer: risk factors, genetic association, and treatment. Asian J Surg. 2018;41(3):210–5.
22. Fais S, Overholtzer M. Cell-in-cell phenomena in cancer. Nat Rev Cancer. 2018;18(12):758–66.
23. Siquara da Rocha L, de O, Souza BS, de Lambert F. Gurgel Rocha C De A. Cell-in-cell events in oral squamous cell carcinoma. Front Oncol. 2022;12:931092.
24. Liu J, Meng H, Nie S, Sun Y, Jiang P, Li S, et al. Identification of a prognostic signature of epithelial ovarian cancer based on tumor immune microenvironment exploration. Genomics. 2020;112(6):4827–41.
25. Liu J, Geng R, Ni S, Cai L, Yang S, Shao F, et al. Pyroptosis-related lncRNAs are potential biomarkers for predicting prognoses and immune responses in patients with UCEC. Mol Ther Nucleic Acids. 2022;27:1036–55.
26. Liu J, Chen C, Wang Y, Qian C, Wei J, Xing Y, et al. Comprehensive of N1-Methyladenosine modifications patterns and immunological characteristics in Ovarian Cancer. Front Immunol. 2021;12:746647.
27. Anestis A, Zoi I, Papavassiliou AG, Karamouzis MV. Androgen receptor in breast Cancer-Clinical and Preclinical Research insights. Molecules. 2020;25(2):358.
28. Venema CM, Bense RD, Steenbruggen TG, Nienhuis HH, Qiu SQ, van Kruchten M, et al. Consideration of breast cancer subtype in targeting the androgen receptor. Pharmacol Ther. 2019;200:135–47.
29. Aurilio G, Cimadamore A, Mazzucchelli R, Lopez-Beltran A, Verri E, Scarpelli M, et al. Androgen receptor signaling pathway in prostate Cancer: from Genetics to Clinical Applications. Cells. 2020;9(12):2653.
30. Salvi S, Bonafè M, Bravaccini S. Androgen receptor in breast cancer: a wolf in sheep's clothing? A lesson from prostate cancer. Semin Cancer Biol. 2020;60:132–7.
31. Özturan D, Morova T, Lack NA. Androgen receptor-mediated transcription in prostate Cancer. Cells. 2022;11(5):898.
32. Lumahan LEV, Arif M, Whitener AE, Yi P. Regulating androgen receptor function in prostate Cancer: exploring the diversity of post-translational modifications. Cells. 2024;13(2):191.
33. Fu X, Wang Q, Du H, Hao H. CXCL8 and the peritoneal metastasis of ovarian and gastric cancer. Front Immunol. 2023;14:1159061.
34. Mishra A, Suman KH, Nair N, Majeed J, Tripathi V. An updated review on the role of the CXCL8-CXCR1/2 axis in the progression and metastasis of breast cancer. Mol Biol Rep. 2021;48(9):6551–61.
35. Yang S, Wang H, Qin C, Sun H, Han Y. Up-regulation of CXCL8 expression is associated with a poor prognosis and enhances tumor cell malignant behaviors in liver cancer. Biosci Rep. 2020;40(8):BSR20201169.
36. Li Y, Wu T, Gong S, Zhou H, Yu L, Liang M, et al. Analysis of the prognosis and therapeutic value of the CXC Chemokine Family in Head and Neck squamous cell carcinoma. Front Oncol. 2020;10:570736.
37. Yue D, Liu S, Zhang T, Wang Y, Qin G, Chen X, et al. NEDD9 promotes cancer stemness by recruiting myeloid-derived suppressor cells via CXCL8 in esophageal squamous cell carcinoma. Cancer Biol Med. 2021;18(3):705–20.
38. Ha H, Debnath B, Neamati N. Role of the CXCL8-CXCR1/2 Axis in Cancer and Inflammatory diseases. Theranostics. 2017;7(6):1543–88.
39. Levantini E, Maroni G, Del Re M, Tenen DG. EGFR signaling pathway as therapeutic target in human cancers. Semin Cancer Biol. 2022;85:253–75.
40. Oxnard GR, Chen R, Pharr JC, Koeller DR, Bertram AA, Dahlberg SE, et al. Germline EGFR mutations and familial Lung Cancer. J Clin Oncol. 2023;41(34):5274–84.
41. Cheng WL, Feng PH, Lee KY, Chen KY, Sun WL, Van Hiep N, et al. The role of EREG/EGFR pathway in Tumor Progression. Int J Mol Sci. 2021;22(23):12828.
42. M N, V M, F M, P P. Crosstalk between CXCR4/ACKR3 and EGFR Signaling in Breast Cancer Cells. International journal of molecular sciences [Internet]. 2022 Jun 10 [cited 2024 May 16];23(19). Available from: https://pubmed.ncbi.nlm.nih.gov/36233192/
43. Wei W, Zhang S, Han H, Wang X, Zheng S, Wang Z, et al. NAT10-mediated ac4C tRNA modification promotes EGFR mRNA translation and gefitinib resistance in cancer. Cell Rep. 2023;42(7):112810.
44. Tiberio L, Del Prete A, Schioppa T, Sozio F, Bosisio D, Sozzani S. Chemokine and chemotactic signals in dendritic cell migration. Cell Mol Immunol. 2018;15(4):346–52.
45. Kim R, Emi M, Tanabe K. Functional roles of immature dendritic cells in impaired immunity of solid tumour and their targeted strategies for provoking tumour immunity. Clin Exp Immunol. 2006;146(2):189–96.
46. Ferlazzo G, Moretta L. Dendritic cell editing by natural killer cells. Crit Rev Oncog. 2014;19(1–2):67–75.
47. Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. Machine-learning methods for ligand-protein molecular docking. Drug Discov Today. 2022;27(1):151–64.
48. 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.
49. Cheng Y, He Y, Li W, Zhang HL, Zhou Q, Wang B, et al. Osimertinib Versus Comparator EGFR TKI as First-Line treatment for EGFR-Mutated Advanced NSCLC: FLAURA China, a Randomized Study. Target Oncol. 2021;16(2):165–76.
50. Ohmori T, Yamaoka T, Ando K, Kusumoto S, Kishino Y, Manabe R, et al. Molecular and clinical features of EGFR-TKI-Associated Lung Injury. Int J Mol Sci. 2021;22(2):792.
51. Noronha V, Patil VM, Joshi A, Menon N, Chougule A, Mahajan A, et al. Gefitinib Versus Gefitinib Plus Pemetrexed and Carboplatin Chemotherapy in EGFR-Mutated Lung Cancer. J Clin Oncol. 2020;38(2):124–36.

52. Murphy M, Stordal B. Erlotinib or Gefitinib for the treatment of relapsed platinum pretreated non-small cell lung cancer and ovarian cancer: a systematic review. Drug Resist Updat. 2011;14(3):177–90.

Publisher's note