Scientific Reports

https://doi.org/10.1038/s41598-026-41052-x

Article in Press

Shared patterns of dysregulated gene expression across squamous cell carcinomas unveil predictors for prognosis and drug sensitivity

Received: 25 June 2025

Accepted: 17 February 2026

Published online: 10 March 2026

Cite this article as: Wang D., Li X., Zhou J. *et al.* Shared patterns of dysregulated gene expression across squamous cell carcinomas unveil predictors for prognosis and drug sensitivity. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-41052-x

Danke Wang, Xu Li, Jiaqi Zhou, Huangbo Yuan, Peipei Gao, Yucan Li, Yixin Zeng, Chen Suo & Xingdong Chen

Shared patterns of dysregulated gene expression across squamous cell carcinomas unveil predictors for prognosis and drug sensitivity

Authors

Danke Wang[1], Xu Li[1], Jiaqi Zhou[1], Huangbo Yuan[1], Peipei Gao[1], Yucan Li[1], Yixin Zeng[1], Chen Suo[2,3,4]*, Xingdong Chen[1,2,5,6]*

Affiliations

[1]State Key Laboratory of Genetic Engineering, Human Phenome Institute, Zhangjiang Fudan International Innovation Center, School of Life Science, Fudan University, Shanghai, 20032, China

[2]Fudan University Taizhou Institute of Health Sciences, Taizhou, Jiangsu, 225300, China

[3]Department of Epidemiology, School of Public Health, Fudan University, Shanghai, 20032, China.

[4]Shanghai Institute of Infectious Disease and Biosecurity, Fudan University, Shanghai, 20032, China

[5]National Clinical Research Center for Aging and Medicine, Huashan Hospital, Fudan University, Shanghai, 200040, China

[6]Yiwu Research Institute of Fudan University, Yiwu, Zhejiang, 322000, China

* Correspondence:
xingdongchen@fudan.edu.cn (Xingdong Chen);
suochen@fudan.edu.cn (Chen Suo)

Abstract

Despite extensive multi-omics studies on squamous cell carcinomas (SCCs) across different organs, the shared transcriptional regulatory mechanisms that driving SCC remain unclear. This study systematically identified common and distinct transcriptomic alterations in SCCs, highlighting key genes and pathways with prognostic and therapeutic relevance. By integrating large-scale gene expression data from SCC tumors and adjacent normal tissues, we revealed dysregulated gene expression patterns (DGEPs) and quantified their similarity across SCCs through correlation and regression analyses. Gene co-expression network analysis identified SCC-associated modules and hub genes, whose biological and clinical significance was further explored through subtype analysis and prognostic modeling. Our findings show that SCCs from the head and neck, esophagus, and cervix share highly similar DGEPs and regulatory networks, whereas lung and skin SCCs exhibit more distinct molecular characteristics. Key processes such as epithelial-mesenchymal transition, extracellular matrix remodeling, and immune-related pathways were strongly linked to SCC prognosis. Moreover, a six-gene prognostic signature (*COL1A1, MMP1, SERPINE1, KRT6A, IGF2BP3,* and *SPP1*) demonstrated robust predictive power for clinical outcomes and therapy response. These findings provide insights into SCC progression and potential therapeutic targets.

Keywords: squamous cell carcinomas, transcriptome, epithelial-mesenchymal transition, immunosuppression, survival, drug

Introduction

Squamous cell carcinoma (SCC) is one of the most common types of cancer, primarily occurring in the skin, oral cavity, nasopharynx, esophagus, lung, and cervix. Epidemiological studies have shown that squamous cell carcinomas (SCCs) arising from multiple organ sites account for approximately one-fifth of all cancer deaths, with the total number exceeding two million[1-3]. SCCs share some of common environmental risk factors, such as smoking, alcohol consumption, and HPV infection[4-8]. Despite originating from different organs, SCCs share a same pathological mechanism of malignant transformation of squamous epithelial cells, which results in similar histological characteristics across these cancers[8]. With the advancement of sequencing technologies, accumulating evidence has highlighted significant molecular similarities among SCCs, including DNA mutations, DNA methylation, RNA expression, and alternative splicing events, and these molecular features are notably

distinct when compared to other types of solid tumors[8-14].

Several studies have conducted integrative analyses of SCCs from DNA mutation, immune-related gene expression to proteomics[15-17]. Campbell et al.[15] identified shared DNA mutations and methylation features in SCCs by integrating multi-omics data. The associated genes primarily regulate squamous cell stemness, progression, differentiation, and genome integrity. Based on the expression of immune-related genes, Li et al.[16] identified six reproducible immune subtypes of cross-organ SCCs and demonstrated that these immune subtypes can be used to guide the prognosis of SCC. By comparing proteomic data from common and rare SCCs, Song et al.[17] found that lipid metabolism reprogramming and the expression of transcription factors (RUNX2, FOXO1) were significantly different between these two types of SCC. These studies have offered valuable insights into the molecular foundations of SCC across various organs, emphasizing the importance of squamous cell stemness, immune responses, and lipid metabolism in SCCs.

However, current analyses of cross-organ SCCs have yet to incorporate adjacent normal tissues to explore dysregulated gene expression patterns (DGEPs), and a more detailed characterization of pan-SCCs gene co-expression network features is still needed. To fill this research gap, we utilized a large-scale gene expression data from SCC tumor and normal adjacent tissues (NAT) to systematically characterize the DGEPs (Tumor vs. Normal) across five common types of SCCs, including lung squamous cell carcinoma (LUSC), head and neck squamous cell carcinoma (HNSC), esophageal squamous cell carcinoma (ESCC), cutaneous squamous cell carcinoma (CSCC), and cervical squamous cell carcinoma (CESC). To explore commonalities in gene expression regulatory mechanism across SCCs associated with tumor progression, we quantified the similarities of DGEPs between different SCC types. By constructing a gene co-expression network that integrates gene expression data from cross-organ SCC tumor and adjacent normal tissues, we were able to identify gene co-expression patterns with broad relevance to SCC, highlighting hub genes and molecular pathways implicated in its pathogenesis. This approach also enabled us to uncover more detailed regulatory patterns of gene expression in the development of individual SCC, shedding light on their unique molecular signatures. By leveraging the hub genes identified in this network, we classified cross-organ SCCs into distinct subtypes and explored genes and pathways associated with patient survival. Ultimately, the prognostic model constructed using survival-associated genes not only enabled the differentiation of SCC patients with varying prognoses, but also provided insights into the potential sensitivity of

patients to chemotherapeutic agents. These findings provide valuable insights for the identification of novel therapeutic targets, facilitating the development of precision therapies, and hold the potential to improve the prognosis and extend the survival of SCC patients.

Results

Study design

This study aims to demonstrate the similarity of DGEP in cross-organ SCCs, revealed shared and distinct transcriptomic alterations, and pinpoint key genes and signaling pathways involved. Ultimately, this study aims to provide computational evidence linking these molecular alterations to clinical outcomes in SCC. The overall study design is shown in **Figure 1**.

We collected gene expression data from both tumor and NAT across five major types of SCC, including LUSC, HNSC, ESCC, CSCC, and CESC. The gene expression microarray chip data (**Table S1**) were obtained from the Gene Expression Omnibus (GEO) database and served as the discovery dataset. RNA sequencing (RNA-seq) data (**Table S2**), downloaded from the European Nucleotide Archive (ENA) database, were used as the validation dataset. And, RNA-seq data from cancer tissue of SCCs, along with matched clinical messages, were retrieved from The Cancer Genome Atlas (TCGA) database to investigate clinical relevance.

First, we conducted differential gene expression analysis to identify DGEPs across the five types of SCCs. To assess the similarity in DGEP between SCCs, we calculated Spearman's correlation coefficient (ρ) based on the log2 fold-change (log2FC) values of the overlapped genes[18]. For better contextualization of the observed similarities, we included lung adenocarcinoma (LUAD) and esophageal adenocarcinoma (EAC) as controls. To further investigate the fine regulation of genes and biological pathways in cross-organ SCCs, we constructed a gene co-expression network to identify gene modules that are shared across SCCs. Next, we investigated the regulatory relationships of gene modules in SCCs, identifying the hub genes and pathways involved. And then, based on hub genes identified within the modules, we performed subtype analysis on SCCs with highly similar DGEPs, highlighting critical pathways and genes that may play a pivotal role in the prognosis of SCC. Finally, we used the survival-related genes found above to construct a prognosis prediction model and further validated its role in predicting drug sensitivity.

Squamous cell carcinomas across organs share more similar

dysregulated gene expression patterns as contrast to adenocarcinoma.

After rigorous preprocessing of individual datasets, we combined microarray datasets from the same type of SCC and corrected for batch effects (**Figure S1**). To investigate whether SCCs from different organ site exhibit similar DGEPs, we first performed differential gene expression analysis to identify the dysregulation of genes (**Data Table S1**). We then conducted a correlation analysis using the log2FC values of overlapped genes to assess the degree of similarity in these DGEPs between SCC pairs.

The five types of SCCs display a shared DGEP (**Figure 2A**), with all SCC pairs showing a Spearman's coefficient $\rho \geq 0.32$ (1000 times permutations, $p < 0.001$). To confirm the robustness of this correlation, we applied different batch correction methods, and the results remained consistent (**Figure S2A and S2B**). After Z-score standardization of log2FC profiles to remove scale and variance effects, strong Spearman's correlations between SCC pairs were consistently preserved (**Figure S2C and S2D**), supporting that the observed similarities reflect robust concordance in relative transcriptomic dysregulation patterns. Among the SCC pairs, the similarities in DGEPs of ESCC_HNSC ($\rho = 0.64$) and ESCC_CESC ($\rho = 0.63$) were higher than those of LUAD_LUSC ($\rho = 0.55$) and EAC_ESCC ($\rho = 0.52$); the lowest similarity was observed between ESCC and CSCC ($\rho = 0.32$). Compared to ESCC_HNSC and ESCC_CESC, the similarity between ESCC and EAC was lower, while the similarity between LUSC and LUAD was higher than that between LUSC with any other type of SCC (Figure 2A, **Figure S3**). Furthermore, the similarity between all SCC pairs was significantly higher than that observed between LUAD and EAC ($\rho = 0.1$, $p < 0.05$, Figure 2A).

Regression analysis of the log2FC values of overlapped genes between ESCC with LUSC, CSCC, CESC, and HNSC showed that the slopes were 5.4, 3.1, 1.3, and 0.93, respectively, indicating that the severity of transcriptomic dysregulation decreased in this following order: LUSC > CSCC > CESC > ESCC ≈ HNSC (**Figure 2B**). Further threshold-based filtering of differentially expressed genes (DEGs, $|\log2FC| > 1$, FDR < 0.05) showed that LUSC, with the highest dysregulation, had the most DEGs, whereas HNSC and ESCC had fewer (**Figure 2C**).

Finally, we used an independent SCCs RNA-seq dataset (Table S2) as a validation set to verify that the similarities of DGEP between SCC pairs we obtained is of general significance. The same analysis process applied to the discovery set was also performed on the validation set after standardizing the data through pre-processing

and correcting for batch effects (**Figure S4**, **Data Table S1**). Based on a rank-based comparison of SCC-pair correlation coefficients between the discovery and validation sets, we observed a strong concordance between the two sets (Spearman's $\rho = 0.67$, $p = 0.032$; **Figure 2D**), which indicates that the SCC similarities of DGEPs are of general significance. At the same time, the DGEPs of these five types of SCCs were also verified. Regression analysis was performed for each type of SCC using the overlapped-genes log2FC values obtained from both the discovery and validation datasets. The results show that DGEPs are highly correlated between these two sets in five types of SCCs (all Pearson correlation coefficient $\rho > 0.65$, $p < 2.2e{-}16$, **Figure S5**). Furthermore, we validated the DEGs of SCCs found in the discovery set (**Table S3**, all the Overlap Odds Ratio > 11.9 and all Overlap $p < 1.75E\text{-}49$, Hypergeometric test).

In summary, the five types of SCCs from different organs exhibit significantly correlated DGEPs (all $\rho \geq 0.32$, permutation test, $p < 0.001$), indicating that they share remarkably similar gene expression regulatory mechanisms in the process of tumor progression. Notably, the DGEPs among ESCC, HNSC and CESC were more similar to each other than that observed between adenocarcinoma and SCC within the same organ (LUAD vs. LUSC, ESCC vs. EAC); while, LUSC exhibited lower similarity in DGEP with other types of SCC than with LUAD. Moreover, we validated the DGEPs and the observed similarities of these patterns across SCCs are of broad and general significance.

Network analysis reveals co-expressed gene modules in squamous cell carcinomas

The above results indicate that SCCs from different organs share similar DEGPs at a global profile. To further elucidate the hub genes and pathways driving these dysregulation patterns and to better understand the gene co-expression landscape in SCCs, we performed a robust weighted gene co-expression network analysis (rWGCNA).

To assess the robustness of the co-expression network, we performed a series of sensitivity analyses. First, the WGCNA network was reconstructed using multiple parameter settings, and the disease-associated modules identified in the primary analysis were consistently retained across most parameter combinations (**Figure S7A**). In addition, we applied a robust WGCNA (rWGCNA) framework by repeatedly resampling two-thirds of the samples for network construction. Across 100 resampling iterations, the disease-associated modules remained stable in the majority of cases, indicating that the identified modules were not driven by specific parameter choices or outlier samples (**Figure S7B**). Ultimately, the

resulting network comprised nine modules, designated CD1 through CD9 (**Figure 3A**).

The correlation between modules is shown in **Figure 3B**. For example, CD1 is positively correlated with CD9 and negatively correlated with CD2, CD3, CD7 and CD8. Further correlation analysis between modules and SCCs showed that CD1 was positively correlated with all five types of SCCs and was enriched in biological processes related to ribosome biogenesis and DNA replication. CD3 and CD7 were negatively correlated with all the five SCCs and enriched in biological processes such as actomyosin structure organization, lipid modification and cell matrix adhesion. The CD2 module was not significantly regulated in LUSC, but was significantly negatively correlated with the other four SCCs and enriched in various metabolism-related biological pathways. CD4 was positively correlated with LUSC and CSCC, but negatively correlated with ESCC, HNSC and CESC, mainly enriched in aerobic respiration biological process. CD8 was negatively correlated with LUSC and positively correlated with CSCC, ESCC, and CESC, and was enriched in angiogenesis and developmental regulation preprocesses. CD5, CD6, and CD9 were negatively correlated with LUSC and positively correlated with the other four SCCs, and were enriched in biological processes such as immunity, superoxide metabolism, and response to viruses. The relationships between modules and five types of SCCs are depicted in **Figure 3C**, and the enrichment results are shown in **Figure 3D**. In general, the gene co-expression status of HNSC, CESC, and ESCC is closer, and the regulation directions of the nine modules are consistent. CSCC has only one module (CD4) with opposite regulation trends compared with these three types of SCCs (HNSC, CESC, and ESCC), while LUSC has five modules (including CD4, CD6, CD8, CD9 and CD5) with opposite regulation trends. This result is consistent with the similarities between SCC pairs obtained by comparing DGEPs in the above.

In the SCC gene co-expression network, hub genes are defined as those with the top 10% connectivity within each module (**Data Table S2**), including *BIRC5*, *CDCA8*, *PPL*, *EVPL*, *SASH1*, *NMU*, et al. A total of 441 hub genes were identified, which were predominantly enriched in biological pathways related to DNA replication, the cell cycle, and endothelial cell proliferation (Data Table S2). Notably, several genes commonly mutated in SCC genomes were included, such as *SOX2*, *TP63*, *ZNF750*, *COL1A1*, *KRT5*, and *MCM7*[8,10,15].

Identification of squamous cell carcinoma subtypes based on hub genes

Based on the expression profiles of hub genes that identified in the gene co-expression network, we identified four cross-organ-SCCs (HNSC, CESC and ESCC) subtypes in the TCGA dataset, named Subtype1, Subtype2, Subtype3 and Subtype4 (**Data Table S3**). The above results demonstrated that CESC, HNSC, and ESCC exhibit remarkable similarity in DGEPs and gene co-expression status. Therefore, these three types of SCCs were combined for the subsequent analysis. **Figure 4A** shows the expression status of the top 100 hub genes with the highest standard deviation across samples. CD1-derived genes such as *SOX2*, *SPP1*, *COL1A1*, *TP63* and *FAP* are highly expressed in Subtype1 and Subtype4; CD2, CD3, and CD4 derived hub genes are highly expressed in Subtype1, including genes *CRNN*, *SLURP1*, *KLK13*, *SCEL*, *SPRR3* and *SPRR1A* et. al. Overall, the hub genes are all expressed at a lower level in Subtype3 (Figure 4A). Significant differences in age, sex, and HPV infection status were observed among the four SCC subtypes (Figure 4A and **Table S4**), with Subtype3 showing a pronounced enrichment of CESC and HPV-positive SCC cases.

In order to figure out whether our SCC subtypes are significantly correlated with those obtained in previous studies[15,16], we compared our subtypes with these previous classifications or subtypes. Campbell et al.[15] identified four types of SCC clusters based on MDSC (myeloid derived suppressor cell)-related signatures, miRNA, copy number variation, and DNA methylation data. Comparing our subtypes with these four types of clusters respectively, the results showed that our subtypes was significantly correlated with these four types of clusters (chi-square test, all the $p < 0.001$). Another study[16] identified six immune subtypes of SCC based on the expression of immune-related genes. Comparison showed that this immune subtype was also significantly correlated with subtypes found in our research (chi-square test, $p < 2.2e\text{-}16$). The comparison results showed that the SCC subtypes identified in this present research through the expression of the hub genes were not only potentially related to immune status, but also correlated with other omics data features. This indicates that the hub genes we identified are of great significance for SCC.

Using the same approach and parameters that conducted in TCGA dataset (**Figures S8A and S8B**), we validated the subtypes of SCC in the GEO microarray SCCs dataset. The results confirmed that the SCC subtypes identified in our study are broadly applicable. The gene expression profiles of Subtype1 and Subtype2 across different modules were consistent between the TCGA and GEO datasets (**Figures S8C and S8D**). However, due to the limited sample size, the gene expression characteristics of Subtype3 and

Subtype4 were less distinct in the GEO dataset (Figure S8D).

Four Subtypes of squamous cell carcinoma have distinct prognoses, revealing relationships between EMT and immune.

To further characterize the clinical relevance of the identified subtypes, we performed Kaplan–Meier survival analyses across the four SCC subtypes. Significant differences in overall survival were observed among the subtypes (**Figure 5A** and **Figure S9A**, log-rank p = 0.00019). Notably, Subtype3 consistently exhibited a significantly more favorable prognosis compared with each of the other subtypes. Pairwise comparisons revealed that patients classified as Subtype3 had significantly improved survival relative to Subtype1 (HR = 1.55, 95% CI: 1.14–2.09), Subtype2 (HR = 1.79, 95% CI: 1.32–2.42), and Subtype4 (HR = 2.23, 95% CI: 1.34–3.72) (Figure S9A). Given that Subtype3 is enriched for CESC and HPV-positive SCCs, we further assessed whether the observed survival differences associated with Subtype3 could be accounted for by organ origin or HPV status. Among HPV-positive patients, non-Subtype 3 cases exhibited significantly worse overall survival compared with Subtype3 cases (HR = 2.05, 95% CI: 1.00–4.19). In addition, although non-CESC patients within Subtype 3 showed poorer survival than CESC Subtype 3 patients (HR = 2.19, 95% CI: 1.29–3.09), the survival difference between non-CESC non-Subtype3 patients and CESC Subtype3 patients was even greater (HR = 2.35, 95% CI: 1.66–3.32), indicating that survival stratification was more pronounced across subtypes than within Subtype3 (**Figure S9B**).

By comparing the subtype with better survival and the subtypes with worse survival, we found that the subtypes with worse survival (Subtype1, Subtype2 and Subtype4) were mainly enriched in biological processes such as epidermal development, epidermal cell differentiation, extracellular matrix (ECM) organization, and epithelial-mesenchymal transition (EMT), while the Subtype3 with better survival was mainly enriched in immune-related biological pathways (**Figure 5B**, Data Table S3). The EMT process and immune-related process play a vital role in the proliferation and development of SCC[19-21]. Therefore, we checked the expression status of transcription factors (*SNAI1, SNAI2, TWIST1* and *ZEB2*) and marker gene *VIM* of EMT process in the four subtypes[22]. The expression of these genes in Subtype3 was significantly lower than that in Subtype1, Subtype2, and Subtype4 (**Figure 5C**). Comparing the immune cell infiltration scores of these four subtypes, we found that CD4+ T cells, CD8+ T cells, and activated dendritic cells had the highest infiltration level in Subtype3, and the lowest in Subtype4 (**Figure 5D**).

Based on the above results, it was found that the high expression

of EMT transcription factors and marker gene was often accompanied with a lower infiltration of anti-tumor immune cells (CD4+ T cells, CD8+ T cells, and activated dendritic cells) in SCC subtypes (Figures 5C and 5D). Therefore, we further performed a correlation analysis (Spearman's correlation) between the expression of EMT transcription factors (or marker gene) and immune cell infiltration score in SCC patients. We found that the gene expression of *SNAI2* was significantly negatively correlated with the infiltration of anti-tumor immune cells such as CD4+ T cells ($p < 0.01$), CD8+ T cells ($p < 0.001$), activated dendritic cells ($p < 0.001$), activated B cells ($p < 0.01$) and effector memory CD8+ T cells ($p < 0.001$); and the expression of *TWIST1* was also significantly negatively correlated with the infiltration of CD8+ T cells ($p < 0.05$) (**Figure 5E**). The expression of *VIM* and *SNAI1* was significantly negatively correlated with the infiltration of mature dendritic cells (all $p < 0.05$), and significantly positively correlated with other immune cells, such as MDSC and regulatory T cell; the expression of *ZEB2* and *VIM* was significantly negatively correlated with the infiltration of neutrophils ($p < 0.01$) (Figure 5E). The infiltration of immune cells is crucial for tumor patients, as it can impact tumor progression in multiple ways, subsequently influencing patient survival[23-26]. Thus, we further validated that the expression level of genes *SNAI2* and *TWIST1* was negatively correlated with the survival of SCC patients (**Figure 5F**, all $p < 0.001$), while, the expression of genes *SNAI1*, *VIM*, and *ZEB2* was not significantly associated with the survival of SCC patients (**Figure S9C**, all $p > 0.01$).

Identification of a six-genes signature that can predict the prognosis and drug sensitivity of SCC patients.

Considering the hub genes and DEGs between Subtype4 and Subtype3, we found that the expression levels of six genes were significantly negatively correlated with the survival of SCC patients, including *COL1A1, MMP1, SERPINE1, KRT6A, IGF2BP3* and *SPP1* (**Figure 6A**). And these six genes were overexpressed in Subtype4 (Subtype4 vs. Subtype3, **Figure 6B**), which may be one of the reasons for the worse survival of Subtype4. These six genes are mainly involved in regulation of cell-substrate adhesion, regulation of response to wounding, and collagen metabolic processes **(Figure 6C)**.

Furthermore, we used lasso regression to confirmed that the six-gene signature was associated with the prognosis of SCC. To assess the robustness of feature selection, we evaluated gene selection frequencies across repeated cross-validation and a range of lambda values in the LASSO Cox model. Several genes exhibited consistent

selection across lambda values, whereas others were selected only under weaker penalization, indicating differential stability of candidate genes (**Figure S10**). Therefore, we constructed a prognostic prediction model using the expression levels of these six genes in the train dataset. The model was employed to calculate the risk score for each SCC patient, and patients were ranked in descending order based on their risk scores. The top 25% of patients were classified as the High-Risk group, while the bottom 25% were classified as the Low-Risk group (**Data Table S4**). Survival analysis conducted in the training dataset demonstrated that patients in the High-Risk group had significantly poorer survival outcomes compared to those in the Low-Risk group (log-rank $p < 0.0001$; HR = 2.62, 95% CI: 1.71–4.02, **Figure 7A**). And, the same results were obtained in the test dataset (log-rank $p = 0.0014$; HR = 4.21, 95% CI: 1.83–9.68, **Figure 7B**) and the independent external validation dataset (log-rank $p = 0.022$; HR = 1.89, 95% CI: 1.09–3.28, **Figure 7C**). To further evaluate the prognostic relevance of the six-gene signature, we next assessed the association between the continuous risk score and patient survival using Cox proportional hazards models (**Table 1**). The risk score was significantly associated with increased mortality risk in both the training dataset (HR = 2.457, 95% CI: 1.745–3.458; $P < 0.001$) and the test dataset (HR = 2.453, 95% CI: 1.440–4.178; $P = 0.001$), with moderate discriminative ability as reflected by the C-index values. In the external test dataset, the continuous risk score showed a weaker association with survival (HR = 1.101, 95% CI: 1.000–1.213; $P = 0.051$), suggesting reduced prognostic resolution when modeling risk as a continuous variable in this dataset. In addition, we evaluated an alternative risk stratification strategy using the median risk score as the cutoff. While this approach retained prognostic significance in the training cohort and marginal significance in the test cohort, it failed to distinguish survival outcomes in the external test dataset (**Figure S11A**). These results indicate that the six-gene signature demonstrates its strongest and most robust prognostic performance when applied to the identification of patients at extreme risk, rather than uniform risk stratification across the entire SCC population.

Importantly, given that Subtype3 was enriched for CESC and HPV-positive SCCs, we further examined whether the observed prognostic value was driven by CESC-specific composition. When restricting the analysis to non-CESC SCC patients only, the six-gene signature remained significantly associated with patient survival, with High-Risk patients exhibiting markedly poorer outcomes than Low-Risk patients (**Figure S11B**). This result demonstrates that the prognostic utility of the six-gene signature is not solely attributable

to CESC origin or HPV-associated biology, but instead reflects shared prognostically relevant transcriptional programs across SCCs from diverse anatomical sites.

Finally, through comprehensive drug sensitivity analysis of SCC patients (Data Table S4), we observed that High-Risk and Low-Risk groups exhibited marked differential responses to conventional SCC chemotherapeutic agents (**Figure 7D**). The Low-Risk group demonstrated greater sensitivity to cisplatin, afatinib, gemcitabine, and irinotecan ($p < 0.05$ for all agents), whereas the High-Risk group showed higher responsiveness to vinblastine and vinorelbine ($p < 0.05$ for both agents). Moreover, the expression levels of these six genes and the calculated risk score were significantly correlated with the drug sensitivity of SCC patients to the six aforementioned agents (**Figure 7E**). Specifically, the risk score exhibited a significant negative correlation with the sensitivity to cisplatin, afatinib, gemcitabine, and irinotecan, while it was significantly positively correlated with the sensitivity to vinblastine and vinorelbine.

In conclusion, the six-gene signature shows consistent discriminatory ability among patients at the extremes of the risk distribution, enabling the identification of clinically distinct High- and Low-Risk groups and underscoring its potential relevance for risk-adapted therapeutic stratification. The distinct drug response profiles between risk groups suggest this molecular signature could serve as a potential biomarker for guiding personalized chemotherapy regimens in SCC management.

Discussion

Previous studies have detailed the molecular similarities across SCCs, focusing on DNA mutations, DNA methylation, and immune gene expression, thereby revealing common features that apply to SCC[15-17]. However, the associated gene expression regulatory mechanisms underlying the tumor progression of cross-organ SCC remain largely unexplored. In this context, we integrated gene expression data based on a large-scale dataset of tumor and adjacent normal tissue samples from cross-organ SCCs. We evaluated the DGEP in different SCCs by performing differential gene expression analysis. Similarities of DGEP between different types of SCC were assessed by correlation analysis of log2FC values of the overlapped genes. Our results showed that SCCs across organs have widely shared DGEPs. We then constructed a cross-organ SCC gene co-expression network, identifying co-expression patterns that are widely associated with SCCs and uncovering hub genes and pathways involved in its pathogenesis. Based on the hub genes

identified in the co-expression network, we classified cross-organ SCC into four distinct subtypes. These four subtypes have significantly different prognoses. Comparison of subtypes with better survival and those with worse survival highlighted several key pathways—such as ECM organization, EMT, and immune-related processes—that may explain the survival differences between these subtypes. Notably, we observed that the expression of the EMT-related transcription factors *SNAI2* and *TWIST1* was associated with reduced infiltration of anti-tumor immune cells, including $CD8^+$ and $CD4^+$ T cells, in SCCs. Finally, by analyzing the expression of hub genes across the four subtypes, we identified six survival-associated genes. We utilized these six (*COL1A1*, *MMP1, SERPINE1, KRT6A, IGF2BP3* and *SPP1*) of these genes to construct a prognostic prediction model, which effectively distinguishes between extreme high-risk and low-risk SCC patients with significantly different survival time. Furthermore, we validated the potential application of this model in predicting the sensitivity of SCC patients to chemotherapeutic agents.

Cross-organ SCCs exhibit significant shared gene expression dysregulation patterns. Despite originating from distinct anatomical sites, these five types of SCCs (HNSC, ESCC, CESC, CSCC and LUSC) demonstrate overlapping molecular signatures during tumor progression. This commonality likely arises from shared internal carcinogenic mechanisms—such as DNA mutations and the activation or suppression of specific signaling pathways—along with similar external environmental exposures, including tobacco smoke, alcohol consumption, and viral infections[4-7,10]. Additionally, these cancers arise from a common squamous epithelial cell lineage and often exhibit comparable tumor microenvironments. Several molecular alterations are consistently observed across these SCCs, including mutations in the *TP53* gene, copy number variations in *SOX2* and *TP63*, activation of the NOTCH, PI3K-AKT, Wnt/β-catenin, and EMT pathways, these alterations are widely recognized as central to the pathogenesis of SCCs[15,27-29].

However, despite the shared molecular features, our study quantifies the degree of similarity between these cancers and reveals notable variability. The DEGPs of ESCC, HNSC, CESC, and CSCC are more similar to each other, while LUSC exhibits the least similarity to the other SCCs. Interestingly, comparing to EAC, ESCC is more similar to other SCCs, underscoring important distinctions between these two cancer types. Similarly, the risk factors and endemic regions for ESCC and EAC differ significantly[13,30,31]. ESCC is primarily associated with dietary habits and exposure to carcinogens, and is most prevalent in African and Asian

populations[30]. In contrast, EAC is often linked to the development of esophageal epithelial changes due to chronic gastroesophageal reflux disease and is more commonly seen in Western countries[31]. These factors may contribute to the differences in gene expression dysregulation patterns of ESCC and EAC. Prior studies[10,14] have suggested that LUSC shares more closer molecular features with other SCCs than with LUAD, while, our analysis indicates that LUSC exhibits greater DGEP similarity to LUAD than to other SCCs. This result suggests that molecular characterization based solely on tumor tissue type may not fully capture the complex and subtle biological processes underlying tumor progression. Incorporating adjacent tissues into the analysis could provide a more comprehensive and accurate understanding of this process.

The construction of network revealed shared gene co-expression modules across different SCCs offering valuable insights into the regulation of specific biological pathways. These co-expression gene modules were predominantly enriched in processes related to DNA replication, fatty acid metabolism, adaptive immune response and the regulation of angiogenesis. Increasing evidence indicates that dysregulated fatty acid metabolism, encompassing both anabolic and catabolic pathways, plays a role in tumor progression by sustaining cancer stem–like cell populations and promoting malignant phenotypes such as metastasis, therapeutic resistance, and recurrence[32]. Tumor angiogenesis plays a central role in cancer development by supporting metabolic demands and disease progression, and emerging molecular insights beyond the VEGF pathway underscore its significance as a persistent and evolving therapeutic target[33]. Altogether, these enrichment patterns suggest that the identified co-expression modules capture core biological processes that are essential for SCC pathogenesis and disease progression.

The regulation of these co-expressed gene modules varies across the five types of SCCs. Overall, the module regulation patterns of ESCC, HNSC, and CESC are more similar to each other, followed by CSCC, while LUSC exhibits distinct regulatory patterns compared to the other SCC types. These findings align with observed similarities in overall transcriptomic dysregulation across these SCCs. LUSC is primarily distinguished from other SCCs by its involvement in processes such as aerobic respiration, adaptive immunity, superoxide metabolism, angiogenesis regulation, and viral response. Notably, CSCC and LUSC share similar regulatory mechanisms in aerobic respiration, a biological process that sets them apart from the other three SCCs. The hub genes identified through this network hold significant clinical relevance. These include not only well-

established SCC-associated mutations in genes such as *SOX2*, *TP63*, *COL1A1, KRT5*and *ZNF750*[8,10,15], which are frequently mutated and shared across multiple SCC types, but also contribute to the identification of four distinct SCC subtypes with unique clinical characteristics. In addition, these subtypes may be linked to other omics features of SCC, indicating that they could have broader implications for the molecular understanding of these SCCs. Collectively, these hub genes appear to occupy central regulatory positions in SCC, reflecting shared molecular programs that underpin tumor development and clinical heterogeneity.

By comparing subtypes with significantly different survival outcomes, we focused on key biological processes—namely, ECM organization, EMT, and immune-related pathways—that may associate with the differential survival observed in SCC subtypes. The role of immune-related processes in tumor progression is well-established, as immune cell infiltration in tumor tissues can influence tumor proliferation and development through mechanisms such as anti-tumor immunity and immune evasion[21,24]. Previous studies have highlighted the critical involvement of the EMT process in creating tumor immunosuppressive microenvironment[34,35]. Our study further demonstrated that EMT-related transcription factors *ZEB2* and *SNAI1*, are positively correlated with the infiltration of immunosuppressive cells, such as Regulatory T cell and MDSC. And, the expression of *SNAI2* and *TWIST1* was significantly negatively correlated with the infiltration of anti-tumor immune cells and with poor survival in SCC patients. Earlier research has shown that SNAI2 promotes breast cancer proliferation by maintaining stem cell-like properties[36], and it is also a critical driver of proliferation and metastasis in head and neck squamous cell carcinoma[37]. Likewise, Twist1 facilitates the onset and progression of lung cancer by activating the Wnt/β-catenin signaling pathway[38]. Our findings extend these observations by highlighting the roles of SNAI2 and TWIST1 in modulating immune cell infiltration, which subsequently influences the survival outcomes of SCC patients.

The prognostic prediction model, constructed based on the expression profiles of six hub genes (*COL1A1, MMP1, SERPINE1, KRT6A, IGF2BP3,* and *SPP1*), demonstrated robust efficacy in stratifying SCC patients with divergent clinical outcomes across both internal and external validation datasets. Notably, the model showed particular strength in distinguishing patients at the extremes of prognostic risk, effectively identifying subsets with markedly poor or favorable outcomes. Furthermore, subsequent validation studies confirmed the model's significant potential in predicting therapeutic response to chemotherapeutic agents. These

findings collectively indicate that this six-gene signature represents a clinically potential biomarker with translational utility for optimizing personalized chemotherapy strategies in SCC treatment paradigms.

It should be noted that the use of NAT as the reference for identifying DGEPs, while effective in controlling for inter-individual genetic background, may introduce inherent limitations in the context of SCC. Owing to the well-recognized Field Cancerization effect in SCC, which arises from chronic exposure to carcinogens such as tobacco and alcohol, NAT samples may already harbor pre-neoplastic genetic or epigenetic alterations[39,40]. As a consequence, transcriptional changes associated with the earliest stages of tumor initiation may be partially attenuated or filtered out when NAT is used as the primary control. This potential bias suggests that the DGEPs and co-expression modules identified in the present study are more likely to capture molecular programs related to tumor progression, invasion, and clinical outcome rather than the initial oncogenic drivers of SCC development. Accordingly, the "universality" of the proposed prognostic model should be interpreted as reflecting conserved progression-associated transcriptional features across SCCs from different anatomical sites, rather than universal mechanisms of tumor initiation. This methodological consideration defines the biological scope of our findings and provides an important context for their interpretation.

Beyond this specific methodological aspect, several additional limitations of the present study should also be acknowledged. First, the analyses were primarily based on publicly available datasets, which may introduce inherent biases related to sample selection, data heterogeneity, and differences in sequencing platforms. In particular, the validation of molecular subtypes may be compromised by the limited number of SCC samples available in the GEO microarray datasets, potentially affecting the robustness and generalizability of subtype classification. Second, the conclusions regarding the association between EMT-related transcription factors and immune cell infiltration, as well as the identification of prognostic genes, were derived mainly from computational analyses and lack independent wet-lab experimental validation, which is essential to confirm these observations and clarify underlying biological mechanisms. Third, although the prognostic prediction model was evaluated using internal test and external test strategies, the possibility of model overfitting cannot be completely excluded. Moreover, due to the limited availability of detailed clinical prognostic information in public datasets, the predictive performance of the model could not be further assessed in larger and

more diverse cohorts. Future studies incorporating prospective clinical data and experimental validation will be required to strengthen the clinical relevance of our findings.

In summary, our study quantified the similarity of the transcriptome dysregulation landscape of cross-organ SCC by comparing the gene expression of tumor and NAT tissues, and explained the specific hub genes and pathways involved by constructing a gene co-expression network. The prognostic prediction model we built based on hub genes can well distinguish patients with significantly different survival, and can help clinicians to develop more personalized chemotherapy strategies in the future. Overall, our findings provide a dual perspective for understanding the progression mechanism of SCC through both molecular mechanisms and clinical applicability, offering a valuable framework for translational medicine that bridges multi-omics exploration with precision medicine interventions.

Materials and Methods

Data collection

Raw microarray gene expression data of SCCs cancer and paired normal adjacent tissues (NAT) were downloaded from the GEO (https://www.ncbi.nlm.nih.gov/geo/) database. The datasets included LUSC (n = 246, T&N = 128&118)[41-45], HNSC (n = 131, T&N = 72&59)[46-49], ESCC (n = 288, T&N = 144&144)[50-55], CESC (n = 104, T&N = 63&41)[56,57], and CSCC (n = 86, T&N = 40&46)[58,59], as well as LUAD (n = 52, T&N = 26&26)[60-63] and EAC (n = 100, T&N = 73&27)[64]. Detailed information for these microarray datasets is provided in Table S1.

Raw RNA sequencing (RNA-seq) data of SCCs cancer and paired NAT were downloaded from ENA (https://www.ebi.ac.uk/) database. These datasets included LUSC (n = 12, T&N = 6&6)[65], HNSC (n = 72, T&N = 37&35)[66,67], ESCC (n = 78, T&N = 37&41) [65,68-70], CESC (n = 23, T&N = 15&8)[65,71], and CSCC (n = 48, T&N = 26&22)[72-74]. Detailed information for the RNA-seq datasets is provided in Table S2.

All raw gene expression datasets obtained from the GEO and ENA databases adhered to the following inclusion criteria: 1) both case and control samples were available within each study; 2) patients had not received any treatment prior to sample collection; and 3) raw data were publicly accessible for download.

Gene expression data of tumor samples along with detailed clinical information of SCC patients, were downloaded from TCGA (https://portal.gdc.cancer.gov/) database (Projects TCGA-HNSC, TCGA-CESC and TCGA-ESCC) and GSE53625[75] dataset from GEO

database. The TCGA dataset included three types of SCCs: HNSC (n = 443), ESCC (n = 89), and CESC (n = 170); the GSE53625 includes 179 ESCC tumor samples. The clinical data comprised patient information such as age, gender, tumor stage, survival status, and survival time.

Quality Control and Normalization

These two types of microarray chip data undergo corresponding preprocessing process. For each Affymetrix microarray chip dataset, the preprocess steps were (1) RMA normalized with the *affy* package[76] in R, including background correction, log2 transformation, and quantile normalization; (2) remove outliers (outliers are defined as samples with standardized sample network connectivity Z scores < -2); (3) regress all available biological and technical covariates, except for diagnostic group. The Agilent microarray chip datasets preprocess steps were (1) normalized with *limma* package[77] in R (read.maimages, backgroundCorrect, normalizeBetweenArrays); and the steps (2) and (3) are the same as Affymetrix chip dataset. All these preprocess steps were performed in R.

There are four steps in RNA-seq data preprocessing. (1) Quality control: FastQC[78] was used to assess the quality of sequencing data; low-quality reads were removed with the tool Trim Galore; employing Cutadapt to remove reads with atypical GC content at the beginning of reads; Using the Bowtie[79] to remove rRNA from samples. (2) Alignment: align the reads to the GRCh38 reference genome using HISAT2[80]. (3) Quantify gene expression counts with featureCounts[81]. (4) Standardized count data to TPM format, and log2 transformed. Preprocess steps (1), (2) and (3) were performed in Linux system; step (4) was performed in R.

After preprocessing, all datasets of the same type of SCC from the same technique were merged and the batch effect was eliminated using the ComBat function of the *sva* package[82] in R.

Differential gene expression analysis and correlation analysis of DGEP

Differential gene expression analysis was performed using two complementary statistical frameworks to ensure robustness against different strategies for handling batch effects. In the first approach, batch effects were corrected using the ComBat function implemented in the *sva* package, followed by differential expression analysis based on linear models using the *limma* package[77].

The second method, we constructed a linear mixed-effects model with the *nlme* package[83] in R. The linear mixed-effects model accounts for both fixed effects (such as sample type and batch effects)

and random effects (such as individual differences). Gene expression was modeled as:

$$y_{\{ij\}} = \beta_0 + \beta_1 \text{SampleType}_{\{ij\}} + \beta_2 \text{Batch}_{\{ij\}} + u_i + \epsilon_{\{ij\}}$$

where $y_{\{ij\}}$ denotes the expression level of a given gene in sample j from individual i; *SampleType* (e.g., tumor vs. normal) and *Batch* were treated as fixed effects; u_i represents a subject-specific random intercept accounting for within-individual correlation; and $\epsilon_{\{ij\}}$ denotes the residual error. Differential expression was assessed based on the statistical significance of the fixed effect associated with sample type.

These differential gene expression analysis methods integrated research data from multiple sources and calculated meta-analytic log2FC for each gene and carcinoma. Genes were then filtered to include only those were present in all studies for subsequent analysis, a total of 8816 genes.

The Spearman's correlation coefficient (ρ) was calculated using the log2FC values of each cancer pair to assess the similarity in DGEPs across different carcinoma types. A permutation test was employed to evaluate the statistical significance of the ρ values between carcinoma pairs. For each individual carcinoma study, we randomized the case/control status 1,000 times and re-performed the linear mixed-effects model meta-analysis, as described in the previous section. Each permutation generated log2FC values for each carcinoma, which were subsequently used to calculate ρ. This process was repeated 1,000 times to generate a null distribution of ρ values for each carcinoma pair.

The differentially expressed genes (DEGs) in each type of carcinoma are defined as $|\log2FC| > 1$ and $FDR < 0.05$.

Gene co-expression network analysis

In order to reveal the systematic gene expression behavior of SCCs, we performed Weighted Gene Co-expression Network Analysis (WGCNA). Individual microarray datasets (after preprocessing) were combined together using the 8816 genes present across all studies. And the batch effect was mitigated by ComBat function.

The network was constructed with WGCNA package[84] in R with a signed network and biweight midcorrelation to reduce sensitivity to outliers. Signed networks and dynamic tree cut methods were employed, as they have been shown to yield more biologically meaningful modules compared to unsigned networks and static tree cut approaches[85]. The soft-thresholding power was selected according to the scale-free topology criterion. Specifically, the smallest power at which the scale-free topology fit index reached $R^2 >$

0.8 was chosen (power = 6), ensuring approximate scale-free network properties while preserving sufficient network connectivity.

Modules were identified using the dynamic tree cut algorithm with the following parameters: minimum module size = 50, deepSplit = 4, merge cut height = 0.1, and pamStage = FALSE. A minimum module size of 50 was selected to reduce the likelihood of identifying small, noise-driven clusters, which are more common when using lower thresholds. A relatively high deepSplit value was used to allow finer module delineation, which is supported by the large sample size analyzed in this study. Modules with highly similar eigengenes were merged using a cut height of 0.1 to avoid redundant modules.

To assess the dependence of module detection on parameter choice, we systematically reconstructed co-expression networks across a broad parameter space, including pamStage = TRUE/FALSE, minClusterSize = 30/50/100, cutHeight = 0.1/0.2, and deepSplit = 0-4. The modules reported in this study were consistently preserved across the majority of parameter combinations, indicating that the identified modules were not driven by a specific parameter setting (Figure S7A).

In addition, module robustness was further evaluated by resampling-based analysis. Using the final parameter set (pamStage = FALSE, minClusterSize = 50, cutHeight = 0.1, deepSplit = 4), two-thirds of the samples were randomly selected to reconstruct the network, and this procedure was repeated 100 times. The reproducibility of key modules across resampled networks further supports the stability and robustness of the identified co-expression structure (Figure S7B).

Modules were labeled by color and assigned unique identifiers (CD#). Genes that did not cluster into any specific module were assigned to the grey module (CD0). Genes within each module were prioritized based on their module membership (kME), defined as correlation to the module eigengene. In this study, the genes with top 10% connectivity within each module are defined as hub gene.

The correlation analysis between each module and SCC was performed with Pearson correlation analysis.

Subtype analysis

Based on the hub genes found in co-expression network analysis, the subtype analysis was performed with the consensusClusterPlus[86] package in R, and the SCCs RNA-seq data downloaded from TCGA database was used. Prior to clustering, gene expression values were median-centered for each gene. Hierarchical clustering was applied as the base clustering algorithm, using Pearson correlation as the distance metric. Consensus clustering was conducted with a maximum of 10 clusters (maxK = 10), 50 resampling iterations, a

resampling proportion of 80% of samples (pItem = 0.8), and all features included in each iteration (pFeature = 1). A fixed random seed was used to ensure reproducibility. According to the Delta Area Plot, we ultimately chose k = 4, at which point the area under the CDF curve did not increase significantly.

Immune cell infiltration analysis of subtypes was performed with *GSVA* package[87] in R. The immune cell marker genes constituted the background gene set for the immune infiltration analysis[88].

Candidate gene selection and prognostic model construction

To identify robust prognostic genes in squamous cell carcinoma (SCC) and construct a biologically interpretable survival model, we employed a stepwise gene prioritization strategy integrating gene co-expression network, expression variability, subtype-specific expression patterns, and penalized regression.

First, hub genes were ranked according to the standard deviation (SD) of expression across the TCGA-SCC cohort, and the top 100 most variable hub genes were retained. In parallel, differential expression analysis was conducted between Subtype4 and Subtype 3 to identify genes significantly upregulated in Subtype4 (log2FC >1, FDR < 0.05), which may be associated with poorer clinical outcomes. The intersection of highly variable network hub genes and Subtype 4–upregulated genes yielded a candidate gene set consisting of 31 genes, which was used as input for subsequent modeling.

Second, to reduce redundancy and address multicollinearity among these 31 candidate genes, a Cox proportional hazards model with LASSO regularization was applied using the glmnet[89] package (alpha = 1). Ten-fold cross-validation was used to select the penalty parameter (lambda). To evaluate the robustness of feature selection, repeated cross-validation and lambda sensitivity analyses were performed, and gene selection frequencies across multiple lambda values were calculated.

Importantly, genes prioritized by LASSO were not automatically regarded as the final prognostic markers. Instead, candidate genes were further evaluated based on survival analyses, as well as biological relevance to SCC-associated processes such as extracellular matrix remodeling, epithelial differentiation, and tumor–microenvironment interactions. Based on this integrative assessment, six genes (*COL1A1, MMP1, SERPINE1, KRT6A, IGF2BP3* and *SPP1*) were selected to construct the final prognostic model.

A multivariate Cox proportional hazards model was then fitted using these six genes in the TCGA-SCC training cohort. The dataset was randomly divided into a training set and an internal test set at a

ratio of 7:3. Model performance was subsequently evaluated in the internal test cohort and validated in an independent external dataset (GSE53625[75]).

Drug sensitivity analysis

The drug sensitivity analysis of patients in the TCGA-SCC dataset to anticancer drugs was performed with the *oncoPredict*[90] package in R. Gene expression data from the GDSC2 training set were obtained from preprocessed datasets in which expression values were RMA-normalized and log-transformed. Drug response data were transformed back to the linear scale prior to model fitting. TCGA-SCC gene expression data were TPM-normalized and log-transformed, and were used as the test dataset. Gene identifiers were matched between the TCGA-SCC and GDSC2 datasets prior to analysis. Batch effects between the GDSC2 training data and TCGA-SCC dataset were corrected using an empirical Bayes method implemented in oncoPredict. No additional power transformation of phenotype data was applied. Genes with low expression variability were removed during data homogenization using a variance threshold of 0.2, and only drugs with a minimum of 20 training samples were retained for prediction. The GDSC2 datasets were downloaded from the OSF repository https://osf.io/c6tfx/.

Survival analysis and enrichment analysis

Survival analysis was performed with R packages *survival*[91] and *survminer*[92], and using the RNA-seq data of SCC tumor samples from TCGA database. Survival analysis was performed between different subtypes, between “High” and “Low” expression groups of hub genes and between “High Risk” and “Low Risk” groups. The “High” group represents patients whose expression levels are higher than the median value, and the “Low” group represents patients whose expression levels are lower than the median value. The “High Risk” group represented patients with risk scores in the top 25%, and the “Low Risk” group represented patients with risk scores in the bottom 25%.

The GO enrichment analysis was performed with the R package *clusterProfiler*[93] based on the GO database. The significantly enriched pathways are defined as FDR < 0.05.

Statistical analysis

Hypergeometric test was used to test whether the validation dataset replicates the differentially expressed genes found in the discovery dataset. Student's t test was used to compare the immune cell infiltration scores and the expression of EMT transcription factors of different subtypes. Chi-square test was used to verify whether our subtypes were significantly associated with the

classifications in other studies. $P < 0.05$ was considered statistically significant. All analyses were performed in R or Linux.

Acknowledgments:

Thanks to the efforts of all people involved in this work. Thanks to the GEO, ENA and TCGA databases for providing valuable datasets.

Funding:

We acknowledge financial supports from the National Natural Science Foundation of China (grant numbers: 82073637, 82122060, 82473700), Science and Technology Innovation 2030 Major Projects (grant number: 2023ZD0510000), National Key Research and Development program of China (grant number: 2023YFC2508001), Shanghai Municipal Science and Technology Major Project (grant numbers: ZD2021CY001, 2023SHZDZX02).

Author contributions:

Conceptualization: X.C., D.W., X.L.

Methodology: D.W., X.L., J.Z., Y.H., P.G., C.S., X.C.

Data collection and analysis: D.W.

Supervision: X.C., C.S.

Writing—original draft: D.W.

Writing—review & editing: D.W., X.L., J.Z., Y.H., P.G., Y.L., YZ., C.S., X.C.

Competing interests:

Authors declare that they have no competing interests.

Consent for publication:

Not applicable.

Ethics approval and consent to participate:

Not applicable.

Data and codes availability:

All data used in this study can be downloaded from the public database. Raw gene expression microarray data can be downloaded from GEO database, accession numbers are recorded in Table S1. Raw RNA-seq data can be download from ENA (accession numbers are recorded in Table S2). Gene expression data matrix of tumor samples and clinical messages of SCC patients can be downloaded from TCGA (Projects of TCGA-ESCA, TCGA-CESC, TCGA-HNSC) database and GEO database (accession number: GSE53625[75]). Data

Table S1-S4, preprocessed data and code required to reproduce the analyses presented in this study are publicly available at https://github.com/WangDanke/Shared_DGEP_SCCs. The repository contains comprehensive scripts covering raw data preprocessing, downstream analyses, and figure generation, as well as the finalized processed data matrices used in key analyses.

Supplementary Materials:
The supplementary materials contain Figure S1 to Figure S11, Table S1 to Table S4, and Data Table S1 to Data Table S3.

References

1 Bray, F. *et al.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229-263 (2024). https://doi.org/10.3322/caac.21834

2 Kudelka, M. R., Lavin, Y., Sun, S. & Fuchs, E. Molecular and cellular dynamics of squamous cell carcinomas across tissues. *Genes Dev* (2024). https://doi.org/10.1101/gad.351990.124

3 Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249 (2021). https://doi.org/10.3322/caac.21660

4 Doorbar, J., Egawa, N., Griffin, H., Kranjec, C. & Murakami, I. Human papillomavirus molecular biology and disease association. *Rev Med Virol* **25 Suppl 1**, 2-23 (2015). https://doi.org/10.1002/rmv.1822

5 Grob, J. J. *et al.* Pembrolizumab Monotherapy for Recurrent or Metastatic Cutaneous Squamous Cell Carcinoma: A Single-Arm Phase II Trial (KEYNOTE-629). *J Clin Oncol* **38**, 2916-2925 (2020). https://doi.org/10.1200/JCO.19.03054

6 Paz-Ares, L. *et al.* Pembrolizumab plus Chemotherapy for Squamous Non-Small-Cell Lung Cancer. *N Engl J Med* **379**, 2040-2051 (2018). https://doi.org/10.1056/NEJMoa1810865

7 Sheikh, M. *et al.* Individual and Combined Effects of Environmental Risk Factors for Esophageal Cancer Based on Results From the Golestan Cohort Study. *Gastroenterology* **156**, 1416-1427 (2019). https://doi.org/10.1053/j.gastro.2018.12.024

8 Dotto, G. P. & Rustgi, A. K. Squamous Cell Cancers: A Unified Perspective on Biology and Genetics. *Cancer Cell* **29**, 622-637 (2016). https://doi.org/10.1016/j.ccell.2016.04.004

9 Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304 e296 (2018). https://doi.org/10.1016/j.cell.2018.03.022

10 Hoadley, K. A. *et al.* Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* **158**, 929-944 (2014). https://doi.org/10.1016/j.cell.2014.06.049

11 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169-175 (2017). https://doi.org/10.1038/nature20805

12 Zhang, L. *et al.* Integrated single-cell RNA sequencing analysis reveals distinct cellular and transcriptional modules associated with survival in lung cancer. *Signal Transduct Target Ther* **7**, 9 (2022). https://doi.org/10.1038/s41392-021-00824-9

13 Zhang, X., Wang, Y. & Meng, L. Comparative genomic analysis of esophageal squamous cell carcinoma and adenocarcinoma: New opportunities towards molecularly targeted therapy. *Acta Pharm Sin B* **12**, 1054-1067 (2022). https://doi.org/10.1016/j.apsb.2021.09.028

14 Kahles, A. *et al.* Comprehensive Analysis of Alternative Splicing Across Tumors from 8,705 Patients. *Cancer Cell* **34**, 211-224 e216 (2018). https://doi.org/10.1016/j.ccell.2018.07.001

15 Campbell, J. D. *et al.* Genomic, Pathway Network, and Immunologic Features Distinguishing Squamous Carcinomas. *Cell Rep* **23**, 194-212 e196 (2018). https://doi.org/10.1016/j.celrep.2018.03.063

16 Li, B., Cui, Y., Nambiar, D. K., Sunwoo, J. B. & Li, R. The Immune Subtypes and Landscape of Squamous Cell Carcinoma. *Clin Cancer Res* **25**, 3528-3537 (2019). https://doi.org/10.1158/1078-0432.CCR-18-4085

17 Song, Q. *et al.* Proteomic analysis reveals key differences between squamous cell carcinomas and adenocarcinomas across multiple tissues. *Nat Commun* **13**, 4167 (2022). https://doi.org/10.1038/s41467-022-31719-0

18 Gandal, M. J. *et al.* Shared molecular neuropathology across major psychiatric disorders parallels polygenic overlap. *Science* **359**, 693-697 (2018). https://doi.org/10.1126/science.aad6469

19 Chien, M. H. *et al.* Cyclic increase in the ADAMTS1-L1CAM-EGFR axis promotes the EMT and cervical lymph node metastasis of oral squamous cell carcinoma. *Cell Death Dis* **15**, 82 (2024). https://doi.org/10.1038/s41419-024-06452-9

20 Liao, C. *et al.* Partial EMT in Squamous Cell Carcinoma: A Snapshot. *Int J Biol Sci* **17**, 3036-3047 (2021). https://doi.org/10.7150/ijbs.61566

21 Jiang, Y. *et al.* Reciprocal inhibition between TP63 and STAT1 regulates anti-tumor immune response through interferon-gamma signaling in squamous cancer. *Nat Commun* **15**, 2484 (2024). https://doi.org/10.1038/s41467-024-46785-9

22 Lu, W. & Kang, Y. Epithelial-Mesenchymal Plasticity in Cancer Progression and Metastasis. *Dev Cell* **49**, 361-374 (2019). https://doi.org/10.1016/j.devcel.2019.04.010

23 Mellman, I., Chen, D. S., Powles, T. & Turley, S. J. The cancer-immunity cycle: Indication, genotype, and immunotype. *Immunity* **56**, 2188-2205 (2023). https://doi.org/10.1016/j.immuni.2023.09.011

24 Park, J., Hsueh, P. C., Li, Z. & Ho, P. C. Microenvironment-driven metabolic

adaptations guiding CD8(+) T cell anti-tumor immunity. *Immunity* **56**, 32-42 (2023). https://doi.org/10.1016/j.immuni.2022.12.008

25 Virassamy, B. *et al.* Intratumoral CD8(+) T cells with a tissue-resident memory phenotype mediate local immunity and immune checkpoint responses in breast cancer. *Cancer Cell* **41**, 585-601 e588 (2023). https://doi.org/10.1016/j.ccell.2023.01.004

26 Chu, Y. *et al.* Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nat Med* **29**, 1550-1562 (2023). https://doi.org/10.1038/s41591-023-02371-y

27 Cancer Genome Atlas, N. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576-582 (2015). https://doi.org/10.1038/nature14129

28 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012). https://doi.org/10.1038/nature11404

29 Song, Y. *et al.* Identification of genomic alterations in oesophageal squamous cell cancer. *Nature* **509**, 91-95 (2014). https://doi.org/10.1038/nature13176

30 Murphy, G. *et al.* International cancer seminars: a focus on esophageal squamous cell carcinoma. *Ann Oncol* **28**, 2086-2093 (2017). https://doi.org/10.1093/annonc/mdx279

31 Rubenstein, J. H. & Shaheen, N. J. Epidemiology, Diagnosis, and Management of Esophageal Adenocarcinoma. *Gastroenterology* **149**, 302-317 e301 (2015). https://doi.org/10.1053/j.gastro.2015.04.053

32 Kuo, C. Y. & Ann, D. K. When fats commit crimes: fatty acid metabolism, cancer stemness and therapeutic resistance. *Cancer Commun (Lond)* **38**, 47 (2018). https://doi.org/10.1186/s40880-018-0317-9

33 Carmeliet, P. & Jain, R. K. Molecular mechanisms and clinical applications of angiogenesis. *Nature* **473**, 298-307 (2011). https://doi.org/10.1038/nature10144

34 Taki, M. *et al.* Tumor Immune Microenvironment during Epithelial-Mesenchymal Transition. *Clin Cancer Res* **27**, 4669-4679 (2021). https://doi.org/10.1158/1078-0432.CCR-20-4459

35 Wang, X., Eichhorn, P. J. A. & Thiery, J. P. TGF-beta, EMT, and resistance to anti-cancer treatment. *Semin Cancer Biol* **97**, 1-11 (2023). https://doi.org/10.1016/j.semcancer.2023.10.004

36 Guo, W. *et al.* Slug and Sox9 cooperatively determine the mammary stem cell state. *Cell* **148**, 1015-1028 (2012). https://doi.org/10.1016/j.cell.2012.02.008

37 Zhang, M. *et al.* FOSL1 promotes metastasis of head and neck squamous cell carcinoma through super-enhancer-driven transcription program. *Mol Ther* **29**, 2583-2600 (2021). https://doi.org/10.1016/j.ymthe.2021.03.024

38 Pan, J. *et al.* lncRNA JPX/miR-33a-5p/Twist1 axis regulates tumorigenesis and metastasis of lung cancer by activating Wnt/beta-catenin signaling. *Mol*

Cancer **19**, 9 (2020). https://doi.org/10.1186/s12943-020-1133-9

39 Curtius, K., Wright, N. A. & Graham, T. A. An evolutionary perspective on field cancerization. *Nat Rev Cancer* **18**, 19-32 (2018). https://doi.org/10.1038/nrc.2017.102

40 Slaughter, D. P., Southwick, H. W. & Smejkal, W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer* **6**, 963-968 (1953). https://doi.org/10.1002/1097-0142(195309)6:5<963::aid-cncr2820060515>3.0.co;2-q

41 Cheng, Z. *et al.* circTP63 functions as a ceRNA to promote lung squamous cell carcinoma progression by upregulating FOXM1. *Nat Commun* **10**, 3200 (2019). https://doi.org/10.1038/s41467-019-11162-4

42 Li, P. *et al.* Proliferation genes in lung development associated with the prognosis of lung adenocarcinoma but not squamous cell carcinoma. *Cancer Sci* **109**, 308-316 (2018). https://doi.org/10.1111/cas.13456

43 Marwitz, S. *et al.* Downregulation of the TGFbeta Pseudoreceptor BAMBI in Non-Small Cell Lung Cancer Enhances TGFbeta Signaling and Invasion. *Cancer Res* **76**, 3785-3801 (2016). https://doi.org/10.1158/0008-5472.CAN-15-1326

44 Sanchez-Palencia, A. *et al.* Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int J Cancer* **129**, 355-364 (2011). https://doi.org/10.1002/ijc.25704

45 Tong, R. *et al.* Decreased Interferon Alpha/Beta Signature Associated with Human Lung Tumorigenesis. *J Interferon Cytokine Res* **35**, 963-968 (2015). https://doi.org/10.1089/jir.2015.0061

46 Ambatipudi, S. *et al.* Genome-wide expression and copy number analysis identifies driver genes in gingivobuccal cancers. *Genes Chromosomes Cancer* **51**, 161-173 (2012). https://doi.org/10.1002/gcc.20940

47 Bayir, O. *et al.* Differentially expressed genes related to lymph node metastasis in advanced laryngeal squamous cell cancers. *Oncol Lett* **24**, 409 (2022). https://doi.org/10.3892/ol.2022.13529

48 Oshima, S. *et al.* Identification of Tumor Suppressive Genes Regulated by miR-31-5p and miR-31-3p in Head and Neck Squamous Cell Carcinoma. *Int J Mol Sci* **22** (2021). https://doi.org/10.3390/ijms22126199

49 Reis, P. P. *et al.* A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. *BMC Cancer* **11**, 437 (2011). https://doi.org/10.1186/1471-2407-11-437

50 Erkizan, H. V. *et al.* African-American esophageal squamous cell carcinoma expression profile reveals dysregulation of stress response and detox networks. *BMC Cancer* **17**, 426 (2017). https://doi.org/10.1186/s12885-017-3423-1

51 Hu, N. *et al.* Genome wide analysis of DNA copy number neutral loss of heterozygosity (CNNLOH) and its relation to gene expression in esophageal squamous cell carcinoma. *BMC Genomics* **11**, 576 (2010). https://doi.org/10.1186/1471-2164-11-576

52 Hu, N. *et al.* Integrative genomics analysis of genes with biallelic loss and its relation to the expression of mRNA and micro-RNA in esophageal squamous cell carcinoma. *BMC Genomics* **16**, 732 (2015). https://doi.org/10.1186/s12864-015-1919-0

53 Lee, J. J. *et al.* Hypoxia activates the cyclooxygenase-2-prostaglandin E synthase axis. *Carcinogenesis* **31**, 427-434 (2010). https://doi.org/10.1093/carcin/bgp326

54 Ming, X. Y. *et al.* RHCG Suppresses Tumorigenicity and Metastasis in Esophageal Squamous Cell Carcinoma via Inhibiting NF-kappaB Signaling and MMP1 Expression. *Theranostics* **8**, 185-198 (2018). https://doi.org/10.7150/thno.21383

55 Su, H. *et al.* Global gene expression profiling and validation in esophageal squamous cell carcinoma and its association with clinical phenotypes. *Clin Cancer Res* **17**, 2955-2966 (2011). https://doi.org/10.1158/1078-0432.CCR-10-2724

56 Scotto, L. *et al.* Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. *Genes Chromosomes Cancer* **47**, 755-765 (2008). https://doi.org/10.1002/gcc.20577

57 Zhai, Y. *et al.* Gene expression analysis of preinvasive and invasive cervical squamous cell carcinomas identifies HOXC10 as a key mediator of invasion. *Cancer Res* **67**, 10163-10172 (2007). https://doi.org/10.1158/0008-5472.CAN-07-2056

58 Garcia-Diez, I. *et al.* Transcriptome and cytogenetic profiling analysis of matched in situ/invasive cutaneous squamous cell carcinomas from immunocompetent patients. *Genes Chromosomes Cancer* **58**, 164-174 (2019). https://doi.org/10.1002/gcc.22712

59 Lambert, S. R. *et al.* Key differences identified between actinic keratosis and cutaneous squamous cell carcinoma by transcriptome profiling. *Br J Cancer* **110**, 520-529 (2014). https://doi.org/10.1038/bjc.2013.760

60 He, F. *et al.* Microarray profiling of differentially expressed lncRNAs and mRNAs in lung adenocarcinomas and bioinformatics analysis. *Cancer Med* **9**, 7717-7728 (2020). https://doi.org/10.1002/cam4.3369

61 Jiang, N. *et al.* HIF-1a-regulated miR-1275 maintains stem cell-like phenotypes and promotes the progression of LUAD by simultaneously activating Wnt/beta-catenin and Notch signaling. *Theranostics* **10**, 2553-2570 (2020). https://doi.org/10.7150/thno.41120

62 Liang, J. *et al.* Mex3a interacts with LAMA2 to promote lung adenocarcinoma metastasis via PI3K/AKT pathway. *Cell Death Dis* **11**, 614 (2020). https://doi.org/10.1038/s41419-020-02858-3

63 Xu, L. *et al.* SPINK1 promotes cell growth and metastasis of lung adenocarcinoma and acts as a novel prognostic biomarker. *BMB Rep* **51**, 648-653 (2018). https://doi.org/10.5483/BMBRep.2018.51.12.205

64 Wang, Q., Ma, C. & Kemmner, W. Wdr66 is a novel marker for risk

stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma. *BMC Cancer* **13**, 137 (2013). https://doi.org/10.1186/1471-2407-13-137

65 Liu, L. *et al.* Analysis of Bulk RNA Sequencing Data Reveals Novel Transcription Factors Associated With Immune Infiltration Among Multiple Cancers. *Front Immunol* **12**, 644350 (2021). https://doi.org/10.3389/fimmu.2021.644350

66 Satgunaseelan, L. *et al.* Oral Squamous Cell Carcinoma in Young Patients Show Higher Rates of EGFR Amplification: Implications for Novel Personalized Therapy. *Front Oncol* **11**, 750852 (2021). https://doi.org/10.3389/fonc.2021.750852

67 Wan, Z. *et al.* Integrative Multi-Omics Analysis Reveals Candidate Biomarkers for Oral Squamous Cell Carcinoma. *Front Oncol* **11**, 794146 (2021). https://doi.org/10.3389/fonc.2021.794146

68 Cao, W. *et al.* Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat Commun* **11**, 3675 (2020). https://doi.org/10.1038/s41467-020-17227-z

69 Chen, S. W. *et al.* The Clinical Significance and Potential Molecular Mechanism of PTTG1 in Esophageal Squamous Cell Carcinoma. *Front Genet* **11**, 583085 (2020). https://doi.org/10.3389/fgene.2020.583085

70 Lin, Z. *et al.* Prognostic Value of SPOCD1 in Esophageal Squamous Cell Carcinoma: A Comprehensive Study Based on Bioinformatics and Validation. *Front Genet* **13**, 872026 (2022). https://doi.org/10.3389/fgene.2022.872026

71 Kurmyshkina, O. V., Dobrynin, P. V., Kovchur, P. I. & Volkova, T. O. Sequencing-based transcriptome analysis reveals diversification of immune response- and angiogenesis-related expression patterns of early-stage cervical carcinoma as compared with high-grade CIN. *Front Immunol* **14**, 1215607 (2023). https://doi.org/10.3389/fimmu.2023.1215607

72 Chitsazzadeh, V. *et al.* Cross-species identification of genomic drivers of squamous cell carcinoma development across preneoplastic intermediates. *Nat Commun* **7**, 12601 (2016). https://doi.org/10.1038/ncomms12601

73 Das Mahapatra, K. *et al.* A comprehensive analysis of coding and non-coding transcriptomic changes in cutaneous squamous cell carcinoma. *Sci Rep* **10**, 3637 (2020). https://doi.org/10.1038/s41598-020-59660-6

74 Srivastava, A. *et al.* MAB21L4 Deficiency Drives Squamous Cell Carcinoma via Activation of RET. *Cancer Res* **82**, 3143-3157 (2022). https://doi.org/10.1158/0008-5472.CAN-22-0047

75 Li, J. *et al.* LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut* **63**, 1700-1710 (2014). https://doi.org/10.1136/gutjnl-2013-305806

76 Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* **20**, 307-315 (2004). https://doi.org/10.1093/bioinformatics/btg405

77 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-

sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015). https://doi.org/10.1093/nar/gkv007

78 SA Bittencourt, S. A. B., S. A., S Bittencourt a. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinformatics* (2010).

79 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009). https://doi.org/10.1186/gb-2009-10-3-r25

80 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915 (2019). https://doi.org/10.1038/s41587-019-0201-4

81 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014). https://doi.org/10.1093/bioinformatics/btt656

82 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-883 (2012). https://doi.org/10.1093/bioinformatics/bts034

83 Pinheiro JC, B. D. *Mixed-Effects Models in S and S-PLUS*. (Springer-Verlag, 2000).

84 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008). https://doi.org/10.1186/1471-2105-9-559

85 Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719-720 (2008). https://doi.org/10.1093/bioinformatics/btm563

86 Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572-1573 (2010). https://doi.org/10.1093/bioinformatics/btq170

87 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013). https://doi.org/10.1186/1471-2105-14-7

88 Charoentong, P. *et al.* Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell Rep* **18**, 248-262 (2017). https://doi.org/10.1016/j.celrep.2016.12.019

89 Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).

90 Maeser, D., Gruener, R. F. & Huang, R. S. oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. *Brief Bioinform* **22** (2021). https://doi.org/10.1093/bib/bbab260

91 T, T. A Package for Survival Analysis in R. (2023).

92 Kassambara A, K. M., Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. (2021).

93 Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation (Camb)* **2**, 100141 (2021). https://doi.org/10.1016/j.xinn.2021.100141

Figures and Table1

Figure 1

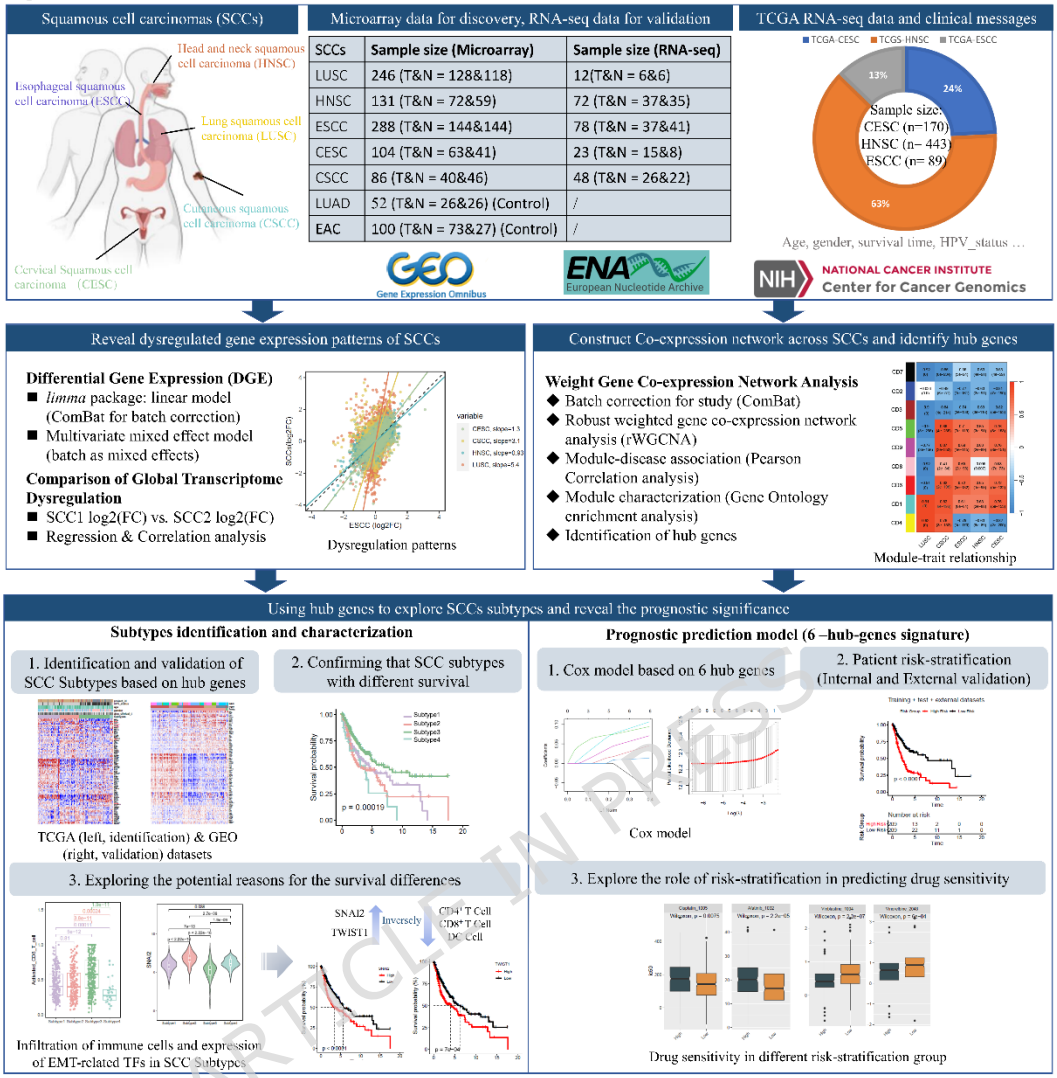

Figure 1. The overall design and workflow of this study. A total of 1790 tumor and normal samples of SCC patients were included in this study. The shared dysregulated gene expression patterns (DGEPs) of SCC were derived and validated in GEO microarray dataset and ENA RNA-seq dataset. And, the co-expression network across SCCs was constructed in GEO microarray dataset. Finally, the clinical relevance of hub genes found in the network was validated in the TCGA-SCC RNA-seq dataset.

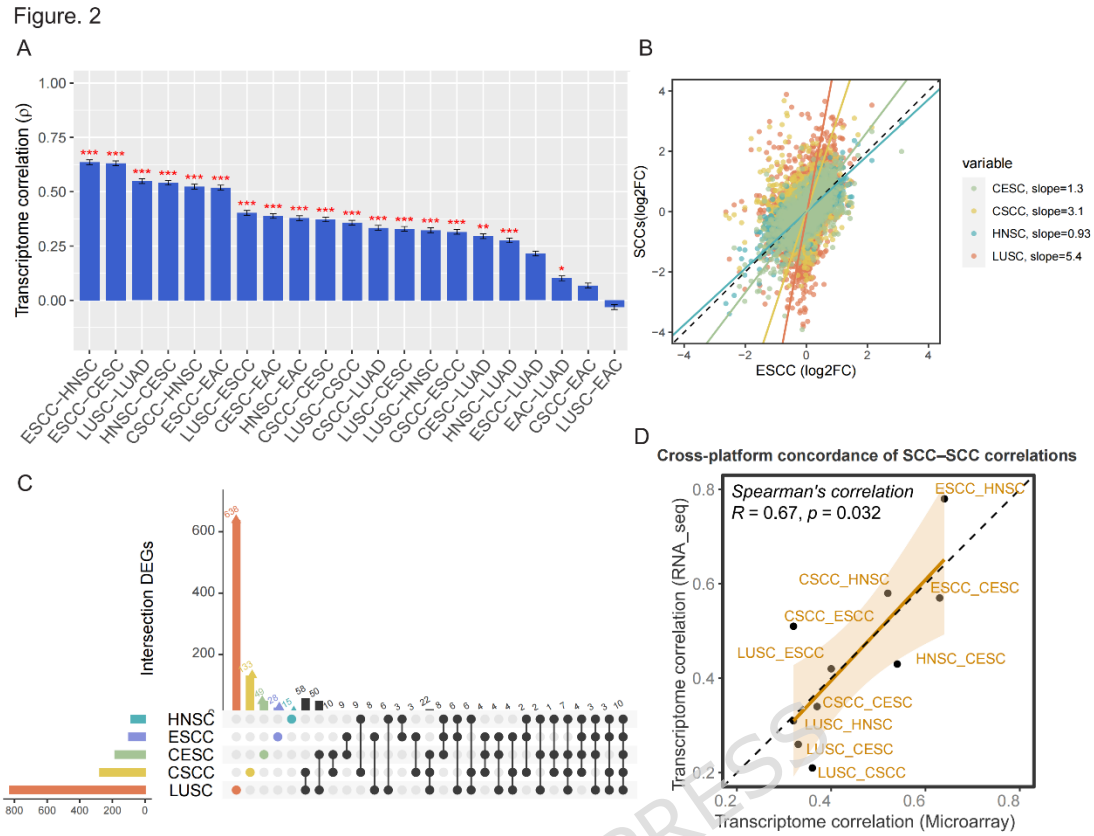

Figure 2. Shared dysregulated gene expression patterns (DGEPs) across SCCs.

(A) Similarity in DGEPs of carcinoma pairs was measured by the Spearman's correlation coefficient of the differential gene expression log2FC values of the overlapped genes. Conducted permutation test using the Bootstrap method (resampled 1000 times), ***p < 0.001, **p < 0.01, *p < 0.05.

(B) Comparison of the degrees of gene expression dysregulation across SCCs, and the severity is LUSC > CSCC > CESC > ESCC ≈ HNSC. The linear relationship between SCCs was estimated by principal component analysis, and measured the severity by comparing slopes.

(C) The interaction sets of differential expression genes (DEGs, |log2FC| >1, FDR < 0.05) among SCCs.

(D) Validation of the similarity between SCCs, each point represents the Spearman's correlation of tumor–normal log2FC profiles between a pair of SCC types, calculated independently in the microarray and RNA-seq datasets. The X-axis shows the Spearman's correlation coefficients of the SCC pairs based on microarray dataset (discovery); Y-axis shows the Spearman's correlation coefficients of the SCC pairs based on RNA-seq dataset (validation). A significant positive concordance was observed between platforms, with a

Spearman's correlation coefficient of R = 0.67 (p = 0.032).

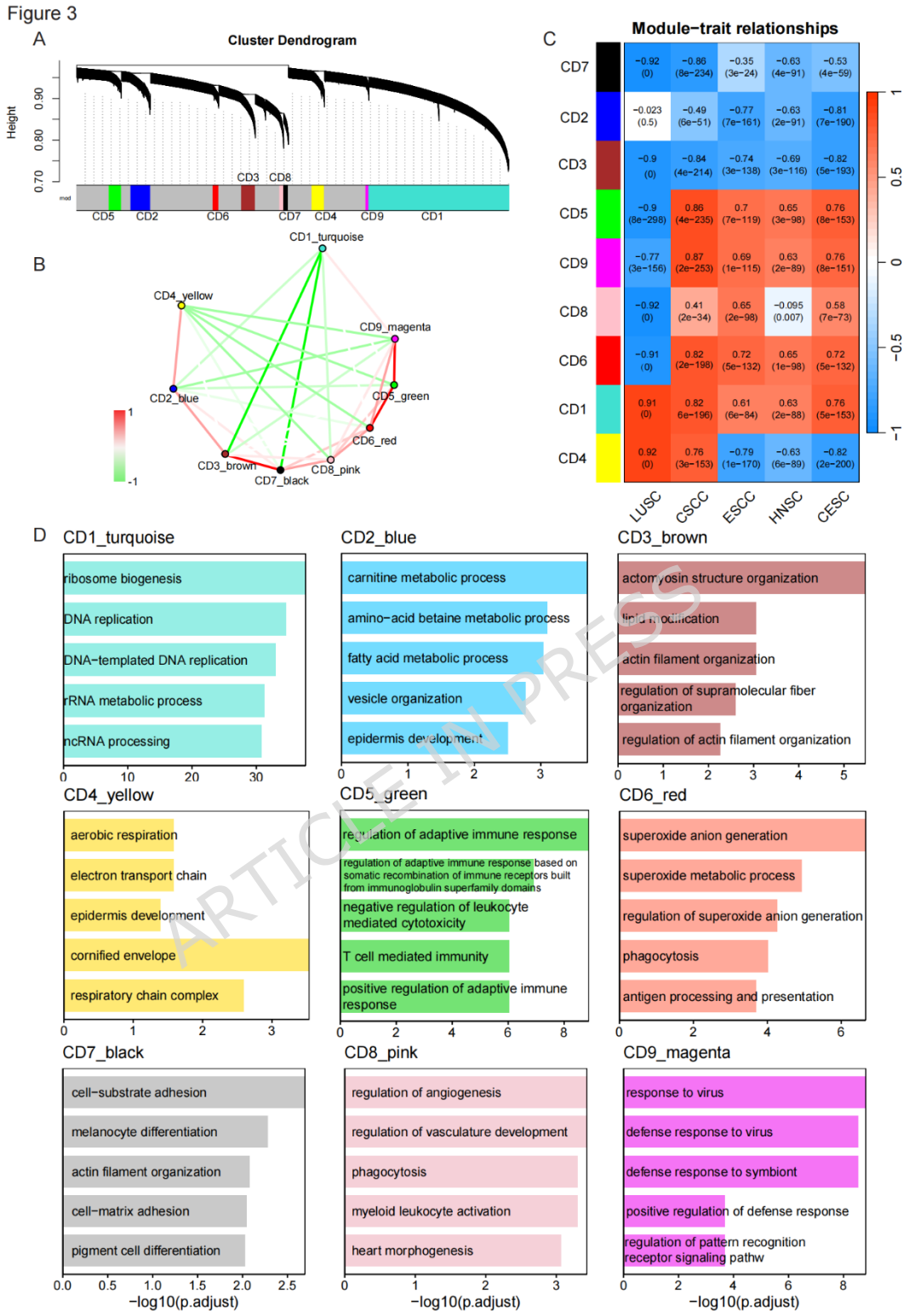

Figure 3. Network analysis revealed modules of co-expressed genes across SCCs.

(A) Network dendrogram that construct with WGCNA method, nine modules of co-expressed genes are identified.

(B) Network diagram demonstrates relationships between modules. The color of edge represents correlation coefficient.
(C) Correlations of modules with SCCs, the Pearson correlation coefficients and the p values are showing in each box. CD1 is positively correlated with all SCCs; CD3 and CD7 are negatively correlated with all SCCs; while, CD2, CD4, CD6, CD7, CD8 and CD9 show different directions of correlation with different SCC.
(D) Enrichment of biological pathways in nine modules. CD1 enriched in DNA replication and rRNA metabolic pathways; CD2 enriched in carnitine metabolic process and fatty acid metabolic process; CD3 enriched in actomyosin structure organization and actin filament organization pathways; CD4 enriched in aerobic respiration and epidermis development processes; CD5 enriched in immune-related pathways; CD6 enriched in superoxide synthesis and metabolism processes; CD7 enriched in cell−substrate adhesion and melanocyte differentiation processes; CD8 enriched in angiogenesis-related pathways; CD9 enriched in pathways related to response to external virus.

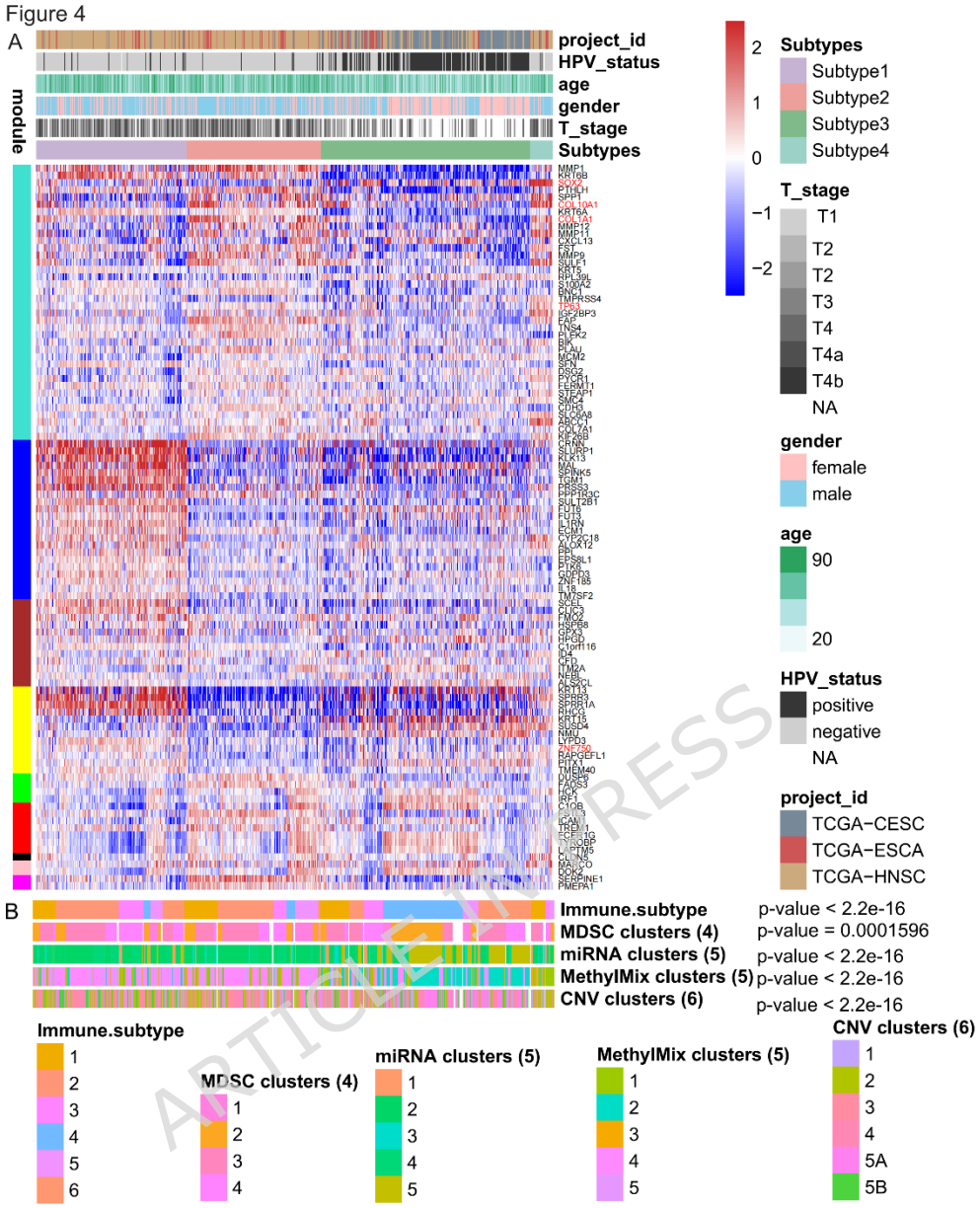

Figure 4. The profiles of our subtypes.

(A) The expression profiles of the top 100 hub genes with the largest standard deviation in the four subtypes. The clinical information of the samples is also displayed, including HPV status, age, gender and clinical T stage.

(B) Comparisons of our SCCs subtypes with those in previous studies. These four types of clusters (MDSC clusters, miRNA clusters, MethylMix clusters, CNV clusters) were obtained by Campbell et al.[15] based on MDSC (myeloid derived suppressor cell)-related signatures, miRNA, copy number variation, and DNA methylation data, repectively. Based on the expression profiles of immune-related genes, Li et al.[16] identified Immune.subtypes. The associations of our subtypes with those clusters and subtypes were

annotated. Chi-square test was used, all the p < 0.001.

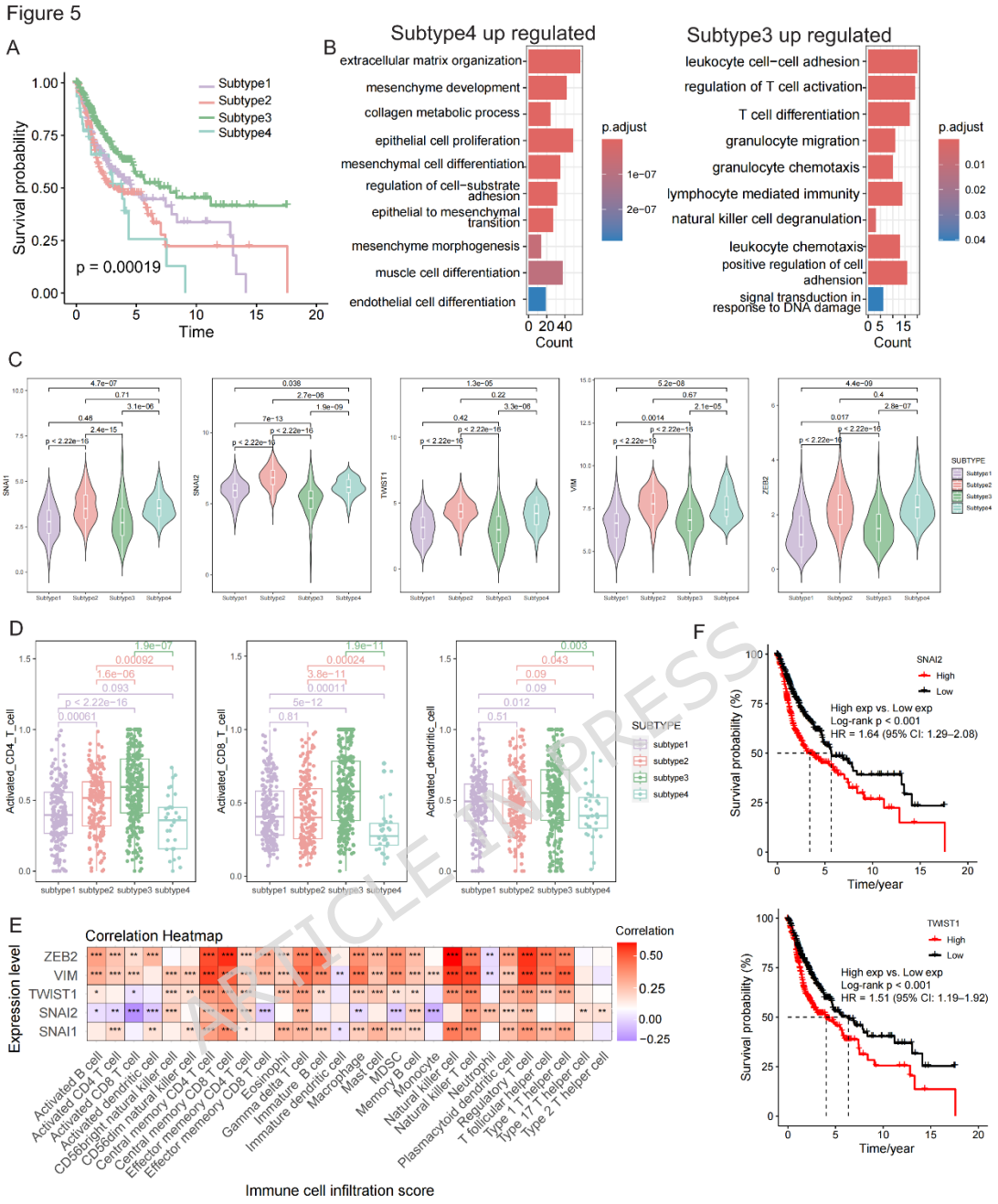

Figure 5. Characteristics of Subtypes. (A) Survival significantly differs among four subtypes, $p = 0.00019$. Kaplan–Meier curves of all patients stratified by Subtypes. Subtype 3 has a better prognosis than Subtype1, Subtype2, and Subtype4. p value was calculated by the log-rank test among subtypes.

(B) GO ontology results of DEGs (Subtype4 vs. Subtype3). The epithelial to mesenchymal transition and extracellular matrix organization pathways are enriched in Subtype4; the immune related pathways are enriched in Subtype3.

(C) The gene expression values of transcription factors (or marker gene) (*SNAI1, SNAI2, TWIST1, VIM, ZEB2*) for epithelial-

mesenchymal transition (EMT) in four subtypes. Subtype3 has the lowest expression level of EMT transcription factors. Violin plots represent the kernel density of expression values, with overlaid boxplots indicating the median and interquartile range (IQR). Pairwise comparisons between subtypes were conducted using two-sided Student's t-tests. Statistical significance is indicated by connecting brackets.

(D) The infiltration scores of $CD4^+$ T cells, $CD8^+$ T cells, positive DC cells, and neutrophils in four subtypes. Subtype 3 has the highest immune cell infiltration scores of CD4+ T cells, CD8+ T cells, and positive DC cells. Boxplots represent the median and IQR, with individual samples shown as jittered points. Pairwise comparisons between subtypes were conducted using two-sided Student's t-tests. Statistical significance is indicated by connecting brackets.

(E) Correlation between the expression levels of EMT-related transcription factors or marker genes and immune cell infiltration scores. Spearman's correlation coefficient (ρ) is shown; ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

(F) The Expression levels of genes *SNAI2* and *TWIST1* were negatively correlated with survival of SCC patients (all $p < 0.001$). The x-axis represents survival rate, y -axis represents survival time (years). The "High" group represents patients whose expression levels are higher than the median value, and the "Low" group represents patients whose expression levels are lower than the median value. Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated using univariable Cox proportional hazards models, with the low-exp group as the reference

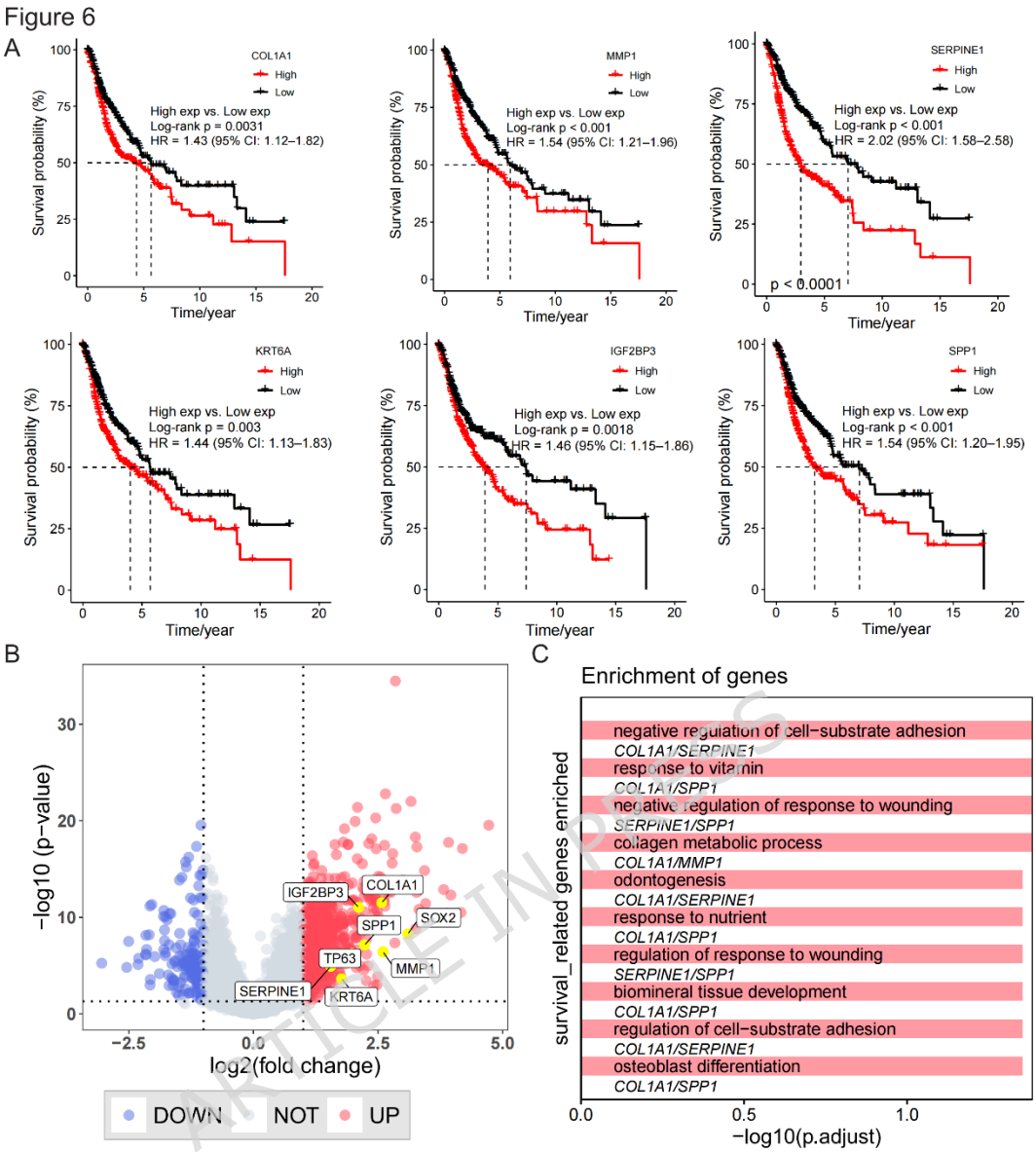

Figure 6. Uncovering prognostic targets for SCC.
(A) Kaplan–Meier curves showing that the high expression levels of these six hub genes (*COL1A1, MMP1, SERPINE1, KRT6A, IGF2BP3,* and *SPP1*) were significantly associated with worse survival of SCC patients. The "High" group represents patients whose expression levels are higher than the median value, and the "Low" group represents patients whose expression levels are lower than the median value. Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated using univariable Cox proportional hazards models, with the low-exp group as the reference.
(B) Volcano plot of differential expression genes (DEGs, |log2FC| >1, FDR < 0.05) of Subtype4 vs. Subtype3. The survival-related genes (*COL1A1, MMP1, SERPINE1, KRT6A, IGF2BP3,* and *SPP1*) and SCC common mutation genes (*SOX2* and *TP63*) are highly expressed in

Subtype4.
(C) The GO enrichment results. The six genes are mainly involved in negative regulation of cell-substrate adhesion, negative regulation of response to wounding, and collagen metabolic processes.

Figure 7

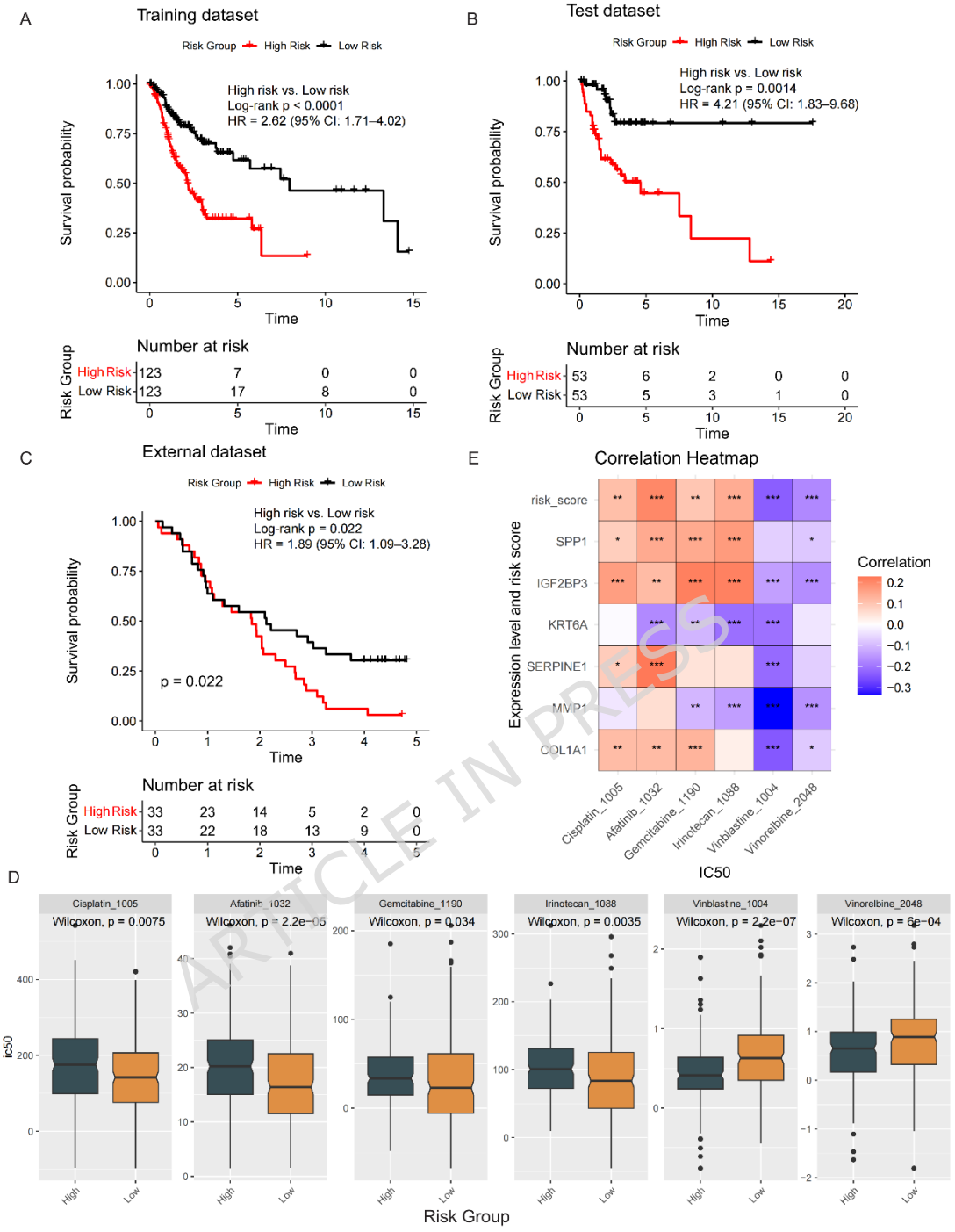

Figure 7. Prognostic and drug response stratification based on the six-gene signature in SCC patients

(A-C) Kaplan–Meier survival analyses of SCC patients stratified by the six-gene risk signature. (A) Training dataset. High-risk patients exhibited significantly worse overall survival compared with low-risk patients (log-rank $p < 0.0001$; HR = 2.62, 95% CI: 1.71–4.02); (B) Internal test dataset. Consistent survival separation was observed between high- and low-risk groups (log-rank $p = 0.0014$; HR = 4.21, 95% CI: 1.83–9.68). (C) External test dataset. The six-gene signature

remained associated with overall survival, with high-risk patients showing poorer outcomes than low-risk patients (log-rank $p = 0.022$; HR = 1.89, 95% CI: 1.09–3.28). The “High Risk” group represented patients with risk scores in the top 25%, and the “Low Risk” group represented patients with risk scores in the bottom 25%. Hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated using univariable Cox proportional hazards models, with the low-risk group as the reference. The numbers of patients at risk are shown below each Kaplan–Meier curve.

(D) IC50 of conventional SCC chemotherapeutic agents (cisplatin, afatinib, gemcitabine, irinotecan, vinblastine, vinorelbine) in High and Low risk groups. Boxplots represent the median and interquartile range, and statistical significance between groups was evaluated using a two-sided Wilcoxon rank-sum test. IC50: half-maximal inhibitory concentration.

(E) The correlation heatmap of these six genes expression levels / (risk score) and IC50 of six chemotherapeutic agents, Spearman’s correlation, ***p < 0.001, **p < 0.01, *p < 0.05.

Table 1. Prognostic Performance of the Continuous Risk Score

Dataset	No. of Patients (Events)	HR† (95% CI)	P-value	C-index
Train	490 (199)	**2.457 (1.745–3.458)**	**<0.001**	0.605
Test	210 (73)	**2.453 (1.440–4.178)**	**0.001**	0.617
External test	131 (104)	1.101 (1.000–1.213)	0.051	0.526

†HR, hazard ratio; per 1-unit increase in the risk score. The C-index was calculated to evaluate discriminative ability.