

Integrated Single-Cell Transcriptomic Analysis of Isonicotinoylation- Related Genes in Cervical Cancer: Diagnosis and Microenvironment Regulation

Zhengchao Yan

Nanxi Mountain Hospital

Sijuan Tang

Affiliated Hospital of Guilin Medical College

Chunhua Tang

Nanxi Mountain Hospital

Minhui He

Nanxi Mountain Hospital

Yuan Lin

Affiliated Hospital of Youjiang Medical University for Nationalities

Minxia Li

Nanxi Mountain Hospital

Shuang Liang

Nanxi Mountain Hospital

Jianbin Yang

Nanxi Mountain Hospital

Fangting Luo

Nanxi Mountain Hospital

Liwu Zhang

Nanxi Mountain Hospital

Jianlin Chen

cj18019000@163.com

Nanxi Mountain Hospital

Research Article

Keywords: cervical squamous cell carcinoma, isocyanation, CDKN2A, XPO1, TST, machine learning, diagnostic decision tree, immune microenvironment, single-cell transcriptome

Posted Date: September 3rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10742446/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Cervical squamous cell carcinoma (CESC) is one of the most common malignancies of the female reproductive system, and its development involves complex gene regulatory networks. Isonicotinoylation, a novel post-translational modification of proteins, has recently garnered increasing attention for its role in regulating tumor metabolism and signal transduction; however, its functional role and clinical significance in cervical cancer remain unclear.

Methods

We integrated multiple CESC transcriptomic datasets from TCGA and GEO (GSE39001, GSE63514, GSE9750, TCGA-CESC, GSE52903) and corrected for batch effects using ComBat. We screened for genes in modules associated with clinical traits using WGCNA, and identified candidate genes by taking the intersection of these modules with differentially expressed genes (DEGs) and the nicotinylation-related gene set (KRG). We selected core genes through cross-validation using three machine learning algorithms—LASSO, SVM-RFE, and Random Forest—and constructed diagnostic receiver operating characteristic (ROC) curves to evaluate their performance. Biological pathways were analyzed using GSEA/GSVA, while immune microenvironment characteristics and molecular subtypes were examined via CIBERSORT and consensus clustering analysis. Single-cell RNA sequencing was used to determine the cellular origins of genes, and regulatory functions were explored using Monocle 2 pseudo-time series analysis, scTenifoldKnn virtual knockout, and CellChat cell-cell communication analysis.

Results

The intersection of DEGs, WGCNA module genes, and KRGs yielded 8 candidate genes; after cross-validation using machine learning, CDKN2A, XPO1, and TST were ultimately identified as core diagnostic genes. The AUC of the three-gene diagnostic panel reached 0.952 (95% CI: 0.916–0.979). These three genes were primarily enriched in cell cycle, DNA replication, and metabolic pathways. Analysis of the immune microenvironment revealed increased infiltration of M0/M1 macrophages and activated memory CD4 + T cells, along with depletion of quiescent mast cells and memory CD4 + T cells; consensus clustering identified three immune subtypes: C1, C2, and C3. Single-cell sequencing revealed that CDKN2A and TST are primarily expressed in epithelial cells, while XPO1 is most highly expressed in macrophages; pseudo-time-series analysis showed that the expression of these three genes dynamically changes throughout development. Virtual knockout analysis indicated that knocking out CDKN2A and XPO1 resulted in approximately 2.0% and 2.8% of genes being differentially regulated, respectively, while knocking out TST had no significant effect. Cell-cell communication analysis revealed that the network was more complex in the CDKN2A low-expression group, while communication was more active in the

XPO1 and TST high-expression groups; the MIF signaling pathway ligand-receptor pairs (APP–CD74, MIF–[CD74 + CXCR4]) showed significant differences across the groups.

Conclusion

CDKN2A, XPO1, and TST are core genes associated with isonitration in CESC; together, they synergistically promote tumor progression by regulating the cell cycle, metabolic reprogramming, and the immune microenvironment. The diagnostic decision tree based on these three genes demonstrates high accuracy, providing new evidence and potential biomarkers for the molecular diagnosis, immunophenotyping, and targeted therapy of CESC.

1 Introduction

Cervical cancer is one of the most common malignancies of the female reproductive system, with approximately 600,000 new cases and 340,000 deaths worldwide each year, representing a particularly heavy disease burden[1]. Persistent infection with high-risk human papillomavirus (HPV) is a necessary condition for the development of cervical cancer; the progression from HPV infection to precancerous lesions and then to invasive cancer typically takes several years to several decades[2]. During this period, host genomic mutations, epigenetic abnormalities, and remodeling of the tumor microenvironment collectively drive disease progression. Although HPV vaccination and screening strategies have significantly reduced incidence rates in developed countries, the five-year survival rate for patients with advanced, recurrent, or metastatic cervical cancer remains below 20%, and treatment options are extremely limited[3]. Therefore, in-depth analysis of the molecular mechanisms underlying cervical cancer and the identification of novel diagnostic biomarkers and therapeutic targets are of great clinical significance[4].

All life processes in eukaryotic cells are precisely regulated by post-translational modifications (PTMs) [5]. By altering protein charge, conformation, stability, and interactions, PTMs rapidly and reversibly regulate key processes such as signal transduction, transcriptional activation, the cell cycle, and metabolic reprogramming[6]. To date, more than 400 types of PTMs have been identified, encompassing classical types such as phosphorylation, acetylation, and ubiquitination[7]. In recent years, high-resolution mass spectrometry has driven the discovery of a series of novel lysine acetylation modifications, including propionylation, crotonylation, succinylation, and 2-hydroxyisobutyryl acetylation. These discoveries have greatly expanded our understanding of how metabolites influence chromatin state and gene transcription through covalent modifications of histones and non-histones, and have provided new perspectives on the cross-talk between metabolic reprogramming and epigenetic regulation[8].

Isonicotinylation (Kinic) is a type of lysine acetylation modification newly identified in recent years. It involves the formation of an acetylation mark on lysine residues using isonicotinyl-coenzyme A—an active metabolite of the first-line anti-tuberculosis drug isoniazid—as the acyl donor[9]. The

establishment of this modification is primarily catalyzed by CBP and P300, while its removal is mediated by HDAC3 and other enzymes. Isonicotinylation can relax chromatin structure, promoting the expression of various oncogenes and the activation of pathways such as PI3K/AKT and TGF β . Subsequent studies have further revealed that non-histones are also subject to isonicotinylation; CBP and KAT5 act as isonicotinytransferases, while HDAC6 and HDAC8 function as deisonicotinytransferases. Isonicotinylation at position K378 of the SMAD3 protein enhances its phosphorylation, activates the TGF β pathway, and promotes the epithelial-mesenchymal transition and invasion of hepatocellular carcinoma cells[10]. The above evidence suggests that isocyanation plays a significant regulatory role in tumor progression. However, to date, systematic studies on isocyanation have almost exclusively focused on hepatocellular carcinoma; its modification profile, regulatory networks, and clinical significance in cervical cancer have not been reported. This knowledge gap provides a clear theoretical motivation for this study.

In recent years, rapid advances in bioinformatics have provided strong support for the discovery of tumor biomarkers. Weighted Gene Co-expression Network Analysis (WGCNA) can identify functional modules and hub genes highly associated with clinical traits at the systems level; cross-validation strategies combining multiple machine learning algorithms can effectively reduce the false-positive rate caused by the bias of a single algorithm and enhance the generalizability of biomarkers; single-cell sequencing technology enables the analysis of gene cellular origin, differentiation trajectories, and microenvironmental cell-cell communication at single-cell resolution[11, 12]. The maturity of these methodological approaches has laid a solid technical foundation for the systematic implementation of this study.

Based on the aforementioned research background and gaps in the literature, this study takes isonitrile-related genes as its entry point. By integrating multiple large-scale transcriptomic datasets and single-cell sequencing data for cervical squamous cell carcinoma (CESC) from the TCGA and GEO databases, and through WGCNA network construction, differential expression analysis, cross-validation of three machine learning models, construction and external validation of diagnostic heatmaps, functional enrichment analysis, analysis of the immune microenvironment and consensus clustering, single-cell localization and pseudo-temporal analysis, virtual knockout simulations, and intercellular communication analysis, we systematically identified core CESC genes associated with isonitration and explored their potential roles in tumor proliferation, metabolic regulation, immune microenvironment remodeling, and intercellular signaling. Our aim is to provide new theoretical foundations and candidate targets for the molecular diagnosis, prognostic stratification, and personalized treatment of cervical squamous cell carcinoma.

2. Materials and Methods

2.1 Data Download and Processing

All data used in this study were obtained from publicly available, freely accessible databases, including The Cancer Genome Atlas (TCGA) and the Gene Expression Omnibus (GEO). The isonitrile-related gene set (KRG) is derived from the list of Kinic-related genes reported by Zhou et al. (see Supplementary Material 1 for details). We downloaded expression profiles and corresponding clinical information for cervical squamous cell carcinoma (CESC) samples from the aforementioned databases. The specific datasets included: GSE39001, GSE63514, GSE9750, and TCGA-CESC (containing 309 tumor samples). In addition, GSE52903 was selected as an external independent validation cohort (containing 300 samples with clinical information). Single-cell RNA sequencing (scRNA-seq) data were obtained from GSE168652, comprising one CESC tumor tissue sample and one control tissue sample. Spatial transcriptomics data were sourced from the 10× Visium platform. All expression profile data were transformed to $\log_2(\text{TPM} + 1)$ prior to subsequent analysis. Gene sets related to isonitration were extracted from published literature. Since all data were publicly available and de-identified, this study did not require additional ethical review.

2.2 Multi-dataset Batch Effect Correction and Integration

To eliminate technical variations across different sequencing platforms and batches, this study utilized the ComBat algorithm from the sva package in R to perform systematic batch effect correction on the three microarray datasets: GSE39001, GSE63514, and GSE9750. The effectiveness of the correction was evaluated by comparing the expression distribution of each sample before and after correction and plotting a principal component analysis (PCA) scatter plot. After correction, the median expression levels and distribution ranges across the datasets became consistent, and samples were uniformly mixed by project source in the PCA plot, confirming that batch effects had been effectively removed.

2.3 Weighted Gene Co-expression Network Analysis (WGCNA)

The weighted gene co-expression network was constructed using the WGCNA package in R. First, the minimum soft threshold (Power = 5 in this study) that first achieved a scale-free R^2 of 0.9 was selected to construct the adjacency matrix, which was then converted into a topological overlap matrix (TOM). Hierarchical clustering was performed using the dynamic tree-cutting algorithm (cutreeDynamic) to identify co-expression modules and calculate the module eigenvectors (Module Eigengenes, MEs) for each module. Pearson correlation analysis was used to evaluate the association between the module eigenvectors and clinical traits (control group vs. CESC group), and the top three modules with the highest correlation coefficients were selected as key modules. Gene significance (GS) was defined as the absolute value of the correlation coefficient between gene expression and disease status, and the distribution of gene significance within each module was compared.

2.4 Identification of Differentially Expressed Genes and Screening of Core Genes

The limma package was used to perform differential expression analysis between CESC tissues and normal tissues, with screening criteria set at $|\log_2 \text{ fold change}| \geq 1$ and a corrected P-value < 0.05 . The intersection of the resulting differentially expressed genes (DEGs), genes from the WGCNA key modules, and the isocyanate-related gene set was calculated. Overlapping genes were visualized using a Venn diagram, and the genes in this intersection were defined as candidate core genes.

2.5 Functional Enrichment Analysis

The clusterProfiler package in R was used to perform Gene Ontology (GO) functional enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis on the candidate core genes, with an adjusted P-value < 0.05 as the significance threshold. The enrichment results were visualized using bar charts and pie charts, respectively.

2.6 Machine Learning for Feature Gene Screening

To further identify robust diagnostic biomarkers from the candidate core genes, this study simultaneously employed three machine learning algorithms for in-depth screening of feature genes. Ultimately, the intersection of the feature genes identified by the three methods was selected as the core diagnostic genes.

LASSO Regression: Using the R package glmnet, a binary logistic regression model was constructed with the expression levels of the genes in the intersection as the independent variable and sample grouping (Control vs. CESC) as the dependent variable. The parameter alpha was set to 1 to enforce L1 regularization (LASSO). Using 10-fold cross-validation ($n\text{folds} = 10$) and cross-validation deviance ($\text{type.measure} = \text{'deviance'}$) as the evaluation metric, the penalty parameter λ (lambda.min) that minimizes the deviance was selected. Finally, genes with non-zero regression coefficients were retained as the feature genes selected by LASSO.

Support Vector Machine Recursive Feature Elimination (SVM-RFE): A custom svmRFE function (see MRG21.msvmRFE.R) was used, based on a support vector machine with a linear kernel ($\text{kernel} = \text{"linear"}$, cost parameter $\text{cost} = 10$) to obtain multiple sets of weight vectors through k-fold subsampling ($k = 10$). The ratio of the mean to the standard deviation of the squared weights of each feature is calculated as the ranking criterion for recursive feature elimination. In each iteration, if the number of remaining features exceeds 50 ($\text{halve.above} = 50$), half of the features are removed; otherwise, one feature is removed at a time, until all features have been ranked. Subsequently, 10-fold cross-validation ($n\text{fold} = 10$) is used to evaluate the model's classification error for different numbers of features. The number of features that minimizes the cross-validation error is selected, and the features at the corresponding positions are identified as the gene features selected by SVM-RFE.

Random Forest: Using the R package *r* and *omF* or *est*, a Random Forest classification model consisting of 500 decision trees ($n\text{tree} = 500$) is constructed, with the intersection of gene expression levels as input and sample groupings as classification labels. The optimal number of trees (i.e., the point of minimum error) is determined by monitoring the out-of-bag error as the number of trees varies, and the Random Forest is then reconstructed using this number of trees. The importance of each gene was measured by the average decrease in Gini impurity, and genes with an importance score greater than 5 were selected as feature genes for the Random Forest screening. The feature gene lists

output by each of the three algorithms were intersected, and the genes common to all three were identified as core diagnostic genes to ensure the robustness and reproducibility of the screening results.

2.7 Construction and Validation of the Diagnostic Bar Chart Model

In the discovery cohort and external independent validation cohorts (GSE52903 and TCGA-CESC), we compared the expression differences of the three core genes (CDKN2A, XPO1, and TST) between CESC and normal tissues. We plotted receiver operating characteristic (ROC) curves for each individual gene and calculated the area under the curve (AUC) to evaluate their independent diagnostic performance. Based on the expression values of the three genes, a combined diagnostic nomogram was constructed using logistic regression. A calibration plot was plotted to assess the consistency between the nomogram's predicted probabilities and the actual observed probabilities. A ROC analysis was performed on the overall nomogram model to calculate its AUC and 95% confidence interval (CI), which were then compared with those of the single-gene models.

2.8 Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA)

For each core diagnostic gene (CDKN2A, XPO1, TST), using only samples from the disease group (CESC), the samples were divided into high- and low-expression groups based on their median expression levels. GSEA analysis was performed using the R package clusterProfiler, referencing the KEGG gene set (c2.cp.kegg. Hs.symbols.gmt). The log fold change (logFC) between the high- and low-expression groups was used as the ranking metric, with 1,000 permutations. Pathways with a nominal P-value < 0.05 were identified as significantly enriched pathways. The top 5 pathways from the high-expression group (NES > 0) and the low-expression group (NES < 0) were extracted, and GSEA plots were generated for each group. At the same time, the ssGSEA method from the R package GSVA was used to calculate the KEGG pathway enrichment scores for each sample. For each core gene, we compared the differences in GSVA scores for each pathway between the high- and low-expression groups (independent samples t-test). Based on the sign of the t-value and a P < 0.05 threshold, pathways were classified into three categories: upregulated, downregulated, and no significant difference. We selected the pathways with the most significant differences and plotted t-value bar charts to visually illustrate the patterns of pathway activity changes under different gene expression states.

2.9 Assessment of the Immune Microenvironment and Immunomolecular Subtyping

The CIBERSORT algorithm was used to deconvolve the merged expression matrix to obtain the relative abundances of 22 types of immune cells. Control and disease groups were distinguished based on the suffixes in the sample names[13]. Stacked bar charts were plotted to display the composition of immune cells in each group, and box plots were used to compare differences in the abundances of each immune cell type between the two groups. Additionally, only samples from the disease group were retained to

calculate Spearman's correlation coefficients between the abundances of different immune cell types and to plot a correlation heatmap. For the three core diagnostic genes (CDKN2A, XPO1, and TST), we extracted their expression levels in the disease group and performed Spearman correlation analyses with the abundances of the 22 immune cell types one by one. After organizing the correlation coefficients and P-values, we plotted a "lollipop plot." To further elucidate the immunological heterogeneity of CESC, consensus clustering was performed using only the expression data of the core genes from the disease group samples (ConsensusClusterPlus package; clustering algorithm: km (k-means); distance: Euclidean; resampling: reps = 50; pltem = 0.8; maximum number of clusters set to 9). Based on the cumulative distribution function (CDF) curve and changes in the area under the curve, the optimal number of clusters was determined to be $k = 3$, dividing the disease samples into three immunomolecular subtypes: C1, C2, and C3. Principal component analysis (PCA) plots (prcomp) were used to visualize the sample distributions across the three subtypes, and ellipses (veganCovEllipse) were plotted to highlight differences between subtypes. Finally, differences in the abundance of 22 types of immune cells were compared across the three subtypes. The ggboxplot function was used, with superimposed significance tests (Kruskal–Wallis test, labeled with *, **, and ***), to illustrate the immune characteristics of each subtype.

2.10 Single-Cell RNA Sequencing Data Processing and Trajectory Analysis

The GSE168652 single-cell data were processed using the R package Seurat. Eligible cells were selected based on the number of genes ($nFeature_RNA > 50$) and the percentage of mitochondrial genes ($percent.mt < 15$). After standardization using LogNormalize, the top 1,500 highly variable genes were selected for principal component analysis (PCA), and batch effects were corrected based on sample origin using the Harmony algorithm. The top 20 principal components were selected, and cell clustering was performed using a graph-based clustering method (resolution = 0.6), followed by dimension reduction and visualization via t-SNE. The SingleR package was used to integrate multiple reference databases for automatic annotation of cell subsets, yielding identity labels for each cell population. For the three core diagnostic genes (CDKN2A, XPO1, TST), their expression distributions across different cell subsets were visualized at the single-cell level using feature maps, violin plots, and bubble plots. Additionally, the FindMarkers function was used to compare differentially expressed genes between the disease group and the control group across each cell type ($logfc.threshold = 0.1$; selection criteria: $|avg_log2FC| > 0.585$ and adjusted $P < 0.05$). For pseudo-temporal trajectory analysis, the Monocle 2 package was used. Differentially expressed genes identified in each cell subpopulation via Seurat analysis were selected as sequencing genes, and the DDRTree dimension-reduction algorithm was employed to construct cell developmental trajectories. Heatmaps and gene expression curves were used to illustrate the dynamic expression trends of the three core genes along the pseudo-temporal axis.

2.11 Virtual Gene Knockout Simulation and Transcriptomic Perturbation Analysis

To assess the impact of each core gene on gene regulatory networks in the tumor microenvironment, this study utilized the R package `scTenifoldKnn` to perform virtual knockout simulations at the single-cell level. Retaining only samples from the disease group (CESC), we extracted the expression matrix for the target genes and the top 5,000 highly variable genes. We set the mitochondrial content threshold to 0.1 and the minimum number of genes in the library to 500, constructed 10 subnetworks, and randomly sampled 500 cells from each subnetwork to perform virtual knockouts. Compare gene regulatory differences between the knockout group and the control group, and identify significantly differentially regulated genes based on a corrected P-value < 0.05. For each target gene (CDKN2A, XPO1, TST), display the top 20 genes by fold change (FC) in a bar chart, show the proportion of globally significantly differentially expressed genes in a pie chart, and present the distribution characteristics of differentially regulated genes in a volcano plot.

2.12 Intercellular Communication Network Analysis

Based on single-cell transcriptomic data, intercellular communication networks were constructed using the R package `CellChat`. Retaining only samples from the disease group, for each core gene (CDKN2A, XPO1, TST), cells were divided into high-expression and low-expression groups based on the median gene expression level. `CellChat` objects were constructed separately for each group. Using the human `CellChatDB` as the reference database and setting `min.cells = 10` to filter out sparse interactions, the ligand-receptor communication probabilities and signaling pathway activities between cell clusters were calculated. Compare differences in cell communication networks between the high- and low-expression groups, specifically including: comparing the number and strength of interactions between the two groups; visualizing the communication networks among cell populations; plotting cell communication heatmaps; comparing overall differences in signaling pathways; and evaluating the roles of each cell population as senders and receivers within the signaling networks. Focus on the differences in key ligand-receptor pairs sent by epithelial cells to other cell populations under high and low expression states of core genes, and display changes in communication probabilities using bubble plots.

2.13 Statistical Analysis

All statistical analyses were performed using R software (version 4.5.0 or later). Comparisons between two groups were performed using the two-tailed Student's t-test or the Wilcoxon rank-sum test; comparisons among multiple groups were performed using analysis of variance (ANOVA) or the Kruskal–Wallis test, supplemented with appropriate post hoc corrections. Correlation analyses were performed using Pearson's or Spearman's methods. Unless otherwise specified, a P-value < 0.05 was considered statistically significant. All figures were generated using R packages such as `ggplot2` and `pheatmap`.

3. Results

3.1 Identification of Key Functional Modules and Characteristic Genes in Cervical Squamous Cell Carcinoma Based on WGCNA

To eliminate technical biases caused by different sequencing batches, this study first performed systematic batch effect correction on the included gene expression profiles GSE39001, GSE63514, and GSE9750. Prior to batch effect correction, there were obvious batch differences in the expression distribution across datasets; after correction, the median expression levels and distribution ranges of samples across datasets became consistent (Fig. 1A). Furthermore, principal component analysis (PCA) results validated the effectiveness of this correction: prior to correction, samples exhibited significant clustering along the PC1 dimension; after correction, samples were uniformly distributed, indicating that batch effects had been effectively removed (Fig. 1B). Heatmaps showing gene expression differences between the disease group and the normal control group across the three datasets, with color gradients representing relative gene expression levels (red indicates high expression, blue indicates low expression) (Fig. 1C). The similarity among samples is visualized as a dendrogram (Fig. 1E). We selected the soft threshold at which the scale-free R^2 first reached 0.9—i.e., Power = 5—as the optimal parameter for network construction. Based on this threshold, the gene set was converted into a weighted co-expression network, and module eigenvectors (MEs) were calculated. Cluster analysis of the module eigenvectors revealed varying degrees of topological overlap among different modules (Fig. 1F). Subsequently, using the Dynamic Tree Cut algorithm, all gene clusters were partitioned into multiple co-expression modules identified by different colors (Fig. 1G). Module-trait association analysis presented heatmaps showing the correlations between each co-expression module and the control group (Control) and the disease group (CESC). Among these, the brown module (MEbrown) exhibited the strongest association with the CESC disease state (Correlation = -0.72, $P = 1e-40$) (Fig. 1H). Finally, by comparing gene significance across modules, the significance of genes within each module can be visualized more intuitively (Fig. 1I).

3.2 Machine Learning-Based Identification and Functional Annotation of Core Genes

To accurately identify key driver genes and potential biomarkers for CESC, this study integrated differential expression analysis (DEGs), co-expression network analysis (WGCNA), and isonitration-related genes (KRG), and used Venn diagrams to identify the intersection, resulting in the identification of 8 candidate core genes with overlapping significance (Fig. 2A). To elucidate the potential biological functions of these 8 overlapping genes, we performed GO functional enrichment and KEGG pathway enrichment analyses. The results indicated that these genes are primarily involved in biological processes such as heterochromatin, positive regulation of macrophage apoptosis, and negative regulation of T-cell differentiation, and they showed high enrichment in the cell cycle signaling pathway (Figs. 2B–C). We performed in-depth screening of candidate genes using three machine learning algorithms: LASSO regression, Random Forest (RF), and Support Vector Machines (SVM). Through 10-fold cross-validation (10-fold CV) and parameter tuning, the LASSO model determined the optimal λ value and the best feature combination (Fig. 2D–E). In the Random Forest algorithm, the model error rate decreased rapidly and then leveled off as the number of decision trees increased; simultaneously, key genes with high diagnostic value, such as CDKN2A and TST, were identified based on feature importance

scores (Fig. 2F). Finally, we took the intersection of the signature genes identified by the three models mentioned above; the results showed that three genes were retained by all three machine learning algorithms (Fig. 2G). This finding not only fully validates the robustness of these genes as diagnostic biomarkers for CESC but also provides highly reliable core targets for subsequent mechanistic studies and clinical translation.

3.3 Validation of Candidate Gene Diagnostic Performance and Construction of a Linear Model

To validate the clinical diagnostic potential of the three core genes (CDKN2A, XPO1, and TST) identified through machine learning, this study conducted validation of expression differences and assessment of diagnostic performance across multiple datasets. In the discovery cohort, the expression levels of CDKN2A and XPO1 in CESC tissues were both significantly higher than those in the normal control group, while TST expression in CESC was significantly lower than in the normal group (all $P < 0.001$; Fig. 3A). Analysis of single-gene receiver operating characteristic (ROC) curves showed that the areas under the curve (AUC) for distinguishing the disease group from the control group using CDKN2A, XPO1, and TST were 0.927, 0.796, and 0.820, respectively, with CDKN2A exhibiting the highest single-gene diagnostic performance (Fig. 3B). To further evaluate its robustness, we replicated the above results in two independent external validation cohorts. The same differences were observed in the GSE52903 dataset (Fig. 3C). In the TCGA database, the expression levels of CDKN2A and XPO1 in tumor samples were also significantly higher than those in normal samples; however, the P-value for the expression difference of TST was not statistically significant (Figs. 3D–F). Given the limitations of single-gene diagnosis, this study constructed a combined diagnostic decision tree based on the expression profiles of these three genes. Calibration curve analysis showed that the probability of CESC occurrence predicted by the decision tree was highly consistent with the actual observed probability (Fig. 3G–H). When the decision curve was evaluated as a comprehensive diagnostic model using ROC analysis (Fig. 3I), the model demonstrated excellent predictive performance in distinguishing tumor from normal samples, with an AUC of 0.952 (95% CI: 0.916–0.979), significantly outperforming any single-gene diagnostic model. These results confirm that the diagnostic receiver operating characteristic (ROC) curve constructed using this three-gene combination has high potential for clinical translation.

3.4 KEGG Pathway Enrichment Analysis of Potential Molecular Mechanisms of Core Diagnostic Genes

To further elucidate the potential biological functions and regulatory networks of the three core diagnostic genes—CDKN2A, XPO1, and TST—in cervical squamous cell carcinoma (CESC), this study conducted gene set enrichment analysis (GSEA) and gene set variation analysis (GSVA). GSEA results showed that CDKN2A was significantly enriched in proliferation-related pathways, such as the cell cycle and DNA replication, in both the high- and low-expression subgroups; In addition to the aforementioned proliferation-related pathways, the XPO1 high-expression subgroup was also significantly positively enriched in the Mismatch repair and Spliceosome pathways, whereas the low-expression subgroup was

primarily enriched in pathways such as arachidonic acid metabolism, linoleic acid metabolism, and cytochrome P450 xenometabolism; The high-expression subgroup of TST was primarily associated with metabolic pathways such as arachidonic acid and glutathione (Fig. 4A–C). GSVA further quantified the differences in pathway activity among different expression subgroups; the bar charts show that CDKN2A expression levels were significantly positively correlated with the cell cycle and DNA replication pathways; XPO1 was significantly activated in signaling pathways such as selenocysteine metabolism, cancer, and the cell cycle; TST, meanwhile, showed a significant positive correlation with arachidonic acid and histidine metabolic pathways (Fig. 4D–F). In summary, these three candidate genes may synergistically contribute to the onset and progression of CESC by regulating different cell proliferation- and metabolism-related signaling pathways.

3.5 Characteristics of the Immune Microenvironment and CESC Immune Subtype Classification

To investigate the immunological significance of core diagnostic genes in cervical squamous cell carcinoma (CESC), this study used the CIBERSORT algorithm to systematically evaluate differences in the abundance of 22 immune cell types between CESC and normal tissues. The results showed that the infiltration proportions of Macrophages M0, Macrophages M1, and activated CD4 memory T cells were significantly elevated in CESC tissues, while immune cells such as resting mast cells and resting CD4 memory T cells exhibited a significant depletion or decline (Fig. 5A–B). A heatmap of correlations among immune cells illustrated the relationships between cell subsets (Fig. 5C). Further analysis of the correlation between core genes and immune infiltration levels revealed that CDKN2A expression was significantly positively correlated with Macrophages M1 and memory B cells, and negatively correlated with naïve B cells and resting CD4 memory T cells. XPO1 expression was positively correlated with Macrophages M1 and negatively correlated with neutrophils. TST expression was significantly positively correlated with CD8 T cells and activated NK cells, and significantly negatively correlated with M1 macrophages (Fig. 5D). Given the significant inter-individual heterogeneity of the immune microenvironment in CESC patients, this study performed consensus clustering based on the abundance matrix of the aforementioned 22 immune cell types. Changes in the cumulative distribution function (CDF) curve and the area under the curve indicated that the optimal number of clusters was 3 (Fig. 5E). Based on the consensus clustering matrix and the results of dimension reduction via principal component analysis (Figs. 5F–G), all samples were clearly divided into three immunomolecular subtypes: C1, C2, and C3. Further comparison of immune cell infiltration differences among the three subtypes revealed distinct immune cell enrichment patterns: the C2 subtype exhibited a significantly higher proportion of M1 macrophages, while the C3 subtype showed a high proportion of activated dendritic cells (Fig. 5H). These results indicate that subtyping based on immunological characteristics aids in the precise identification of patients' complex immune states and provides an important reference for the selection of subsequent immunotherapy strategies.

3.6 Single-Cell Transcriptomics-Driven Cell-of-Origin Tracing and Dynamic Trajectory Reconstruction of

Diagnostic Genes

To precisely elucidate the cellular origins of key diagnostic genes (CDKN2A, XPO1, TST) and their expression profiles in the tumor microenvironment (TME) at the single-cell level, this study analyzed the single-cell RNA sequencing dataset GSE168652 and performed single-cell RNA sequencing and downstream analyses. Through dimensionality reduction and clustering, we identified five major cell types: epithelial cells, fibroblasts, endothelial cells, CD8 + T cells, and macrophages (Fig. 6A), which can be further subdivided into 16 cell subpopulations (Fig. 6B). The signature genes and functional enrichment heatmaps for each cell type not only validated the reliability of the annotations and the expression trends of cell-associated genes but also revealed the specific functions of each cell population in processes such as matrix remodeling, immune chemotaxis, and humoral immunity (Fig. 6C). Building on this, we focused on examining the cell-level expression profiles of three core diagnostic genes. Bubble plots and violin plots indicate that CDKN2A is primarily expressed in epithelial cells; XPO1 is widely expressed across various cell types, but its expression is relatively highest in macrophages; whereas TST exhibits relatively specific high expression in epithelial cells (Fig. 6D–F). Furthermore, to explore the dynamic evolution of cellular states, we constructed pseudotime analysis trajectories (Fig. 6G) and used pseudotime heatmaps and gene trajectory plots to reveal the dynamic expression changes of CDKN2A, XPO1, and TST during the developmental evolution of different cell subpopulations. The pseudotime trajectories for CDKN2A and TST first declined, then rose, and finally declined again, while XPO1 exhibited an upward trend (Fig. 6H–I).

3.7 Virtual Knockout of CDKN2A, XPO1, and TST Induces Differential Transcriptional Disturbances

To assess the impact of different target genes on gene regulatory networks in the tumor microenvironment, we performed virtual knockout simulations based on single-cell transcriptomics for the CDKN2A, XPO1, and TST genes, respectively, and compared the resulting differential regulation profiles. The top 20 significantly differentially regulated genes, as shown by fold change (FC), revealed that the knockdown of each target gene triggered distinctly different patterns of downstream gene changes. In the CDKN2A group, genes such as CD2, PERP, and CSTB ranked among the top in terms of FC; whereas in the XPO1 and TST models, the TRAC and TRBC2 genes both exhibited extremely significant upregulation. In particular, in the TST virtual knockout model, the fold changes (FC) for these two genes were significantly higher than those for all other genes, reaching over 2,000-fold (Figure A). Analysis of global gene expression changes using pie charts revealed that in the CDKN2A group, only approximately 2.0% of genes exhibited significant differential regulation ($p_{\text{adj}} < 0.05$); in the XPO1 group, this proportion was approximately 2.8%. In contrast, the TST virtual knockout did not induce any globally significant transcriptional changes (0% of genes were significant) (Figure B). In the CDKN2A knockout group, a majority of genes exceeded the significance threshold (red dashed line), including NEAT1, FERMT1, IL1RN, and other core genes; the XPO1 knockout group similarly induced widespread gene disruption, such as PSMB9, HSPE1, and BIRC5. In stark contrast to the two groups mentioned above,

only two genes—TRBC2 and TRAC—in the TST group had scatterplots above the significance threshold (Figure C).

3.8 Different Expression States Predict the Remodeling of Intercellular Communication Networks in the Microenvironment

By comparing the intercellular communication networks in groups with high and low expression of target genes (Figures A–C), we observed distinctly different communication architecture patterns. In the predictive model, the CDKN2A low-expression group exhibited a more complex and stronger communication network, while the XPO1 and TST high-expression groups also exhibited more complex and stronger communication networks; in particular, there was a significant increase in interaction links between fibroblasts, macrophages, and CD8 + T cells. To further elucidate the specific molecular basis underlying these macroscopic network differences, we extracted key ligand-receptor interaction pairs from the high- and low-expression groups for visualization analysis. The results of the bubble plots (Figures D–F) suggest that differences in communication probabilities among the three groups of high- and low-expression genes are primarily concentrated in specific signaling pathways. Specifically, the MIF signaling pathway (e.g., APP-CD74, MIF-[CD74 + CXCR4], and MIF-[CD74 + CD44]) exhibited significant differences across multiple group comparisons. Compared with the low-expression group, the communication probability of the aforementioned ligand-receptor pairs was significantly higher in the CDKN2A high-expression group, with strong statistical significance.

4 Discussion

By integrating large-scale transcriptomic data from multiple platforms with single-cell sequencing data, this study systematically identified three core genes—CDKN2A, XPO1, and TST—associated with isonitrile modification in C ESCs and explored their potential regulatory roles in tumorigenesis, immune microenvironment remodeling, and intercellular communication. To our knowledge, this is the first report to link isocyanate modification to cervical cancer and to conduct a systematic bioinformatics study on this topic.

The CDKN2A gene encodes two cell cycle repressor proteins, p16INK4a and p14ARF, which play a central role in RB–E2F cell cycle checkpoint regulation[14]. In HPV-associated cervical cancer, the E7 protein of high-risk HPV lifts the inhibition of E2F by degrading the Rb protein, which in turn induces the overexpression of p16INK4a, making CDKN2A one of the most classic auxiliary diagnostic markers for cervical precancerous lesions and invasive carcinoma[15]. In this study, CDKN2A was significantly overexpressed in C ESC tissues (AUC = 0.927). A recent multi-omics integrative analysis similarly confirmed that CDKN2A is a core biomarker of cell cycle dysregulation in cervical cancer, supporting its clinical value as a diagnostic and prognostic biomarker[16]. Notably, CDKN2A was included in the analysis as a member of the isocyanate-modified gene set, suggesting that its expression or function may be regulated by isocyanate modification. Recent studies in hepatocellular carcinoma have found

that isocyanate modification can activate the PI3K/Akt/mTOR signaling pathway, while CDKN2A has been shown to inhibit the proliferation and invasion of cervical cancer cells via the LDHA-mediated AKT/mTOR pathway[9, 17]. GSEA results revealed that the CDKN2A-high-expression subgroup was significantly enriched in the cell cycle and DNA replication pathways, further validating its central role in driving tumor cell proliferation. Immune-related analyses revealed that CDKN2A expression is positively correlated with M1-type macrophages and negatively correlated with resting memory CD4 + T cells, consistent with recent multi-omics analyses indicating tumor-type-specific interactions between CDKN2A expression and immune cell subsets[16]. Tumors with high CDKN2A expression may be in an immune state characterized by “inflammation activation but T-cell exhaustion,” suggesting that p16INK4a, as a tumor-associated autoantigen, may serve as a potential target for immunotherapy.

XPO1 (exportin-1/CRM1) encodes a key nuclear export receptor responsible for the nuclear-cytoplasmic transport of numerous tumor suppressor proteins, including p53 and FOXO1[18]. XPO1 is upregulated in various solid tumors, including ovarian cancer, glioma, osteosarcoma, pancreatic cancer, and cervical cancer[19]. In cervical cancer, previous studies have shown that XPO1 regulates the nuclear export of cell cycle and tumor suppressor genes by binding to Rad21; inhibition of XPO1 suppresses the stemness and radiation resistance of CD44-positive cervical cancer cells[20]. In this study, XPO1 was significantly overexpressed in C ESCs (AUC = 0.796), consistent with the upregulation trend observed in a study on nicotinylation in hepatocellular carcinoma—in which XPO1 was also identified as one of the significantly upregulated genes in the high-risk subgroup[21]. GSEA revealed that the XPO1-high-expression subgroup was significantly enriched in mismatch repair and spliceosome pathways, suggesting that XPO1 may promote cervical cancer progression by maintaining genomic stability and RNA processing integrity—a functional dimension that has received little attention in previous cervical cancer studies. Single-cell sequencing revealed that XPO1 expression was relatively highest in macrophages; this finding is particularly significant—it suggests that XPO1 not only exerts its effects through intrinsic mechanisms within tumor cells but may also participate in the remodeling of the immune microenvironment by regulating the function of tumor-associated macrophages[22, 23]. In clinical research, the XPO1 inhibitor selinexor has demonstrated monotherapy activity in Phase II trials for recurrent gynecologic malignancies (including cervical cancer), with a median overall survival of 5.0 months among patients[24]. Given the correlation pattern showing that XPO1 is positively correlated with M1-type macrophages and negatively correlated with neutrophils, it is hypothesized that tumors with high XPO1 expression may shape an immunosuppressive microenvironment by modulating the polarization state of macrophages.

TST (thiosulfate sulfotransferase) is a mitochondria-localized enzyme that plays a key role in sulfur metabolism, redox balance, and the maintenance of iron-sulfur proteins. Dysregulation of TST expression has been shown to be associated with various pathological conditions, including metabolic diseases, neurological disorders, and cancer[25]. Proteinome databases indicate that TST serves as a marker of favorable prognosis in renal cell carcinoma and urothelial carcinoma, and that TST is associated with various cancer types, including hepatocellular carcinoma, non-small cell lung cancer, and cervical cancer. Pan-cancer single-cell studies have also revealed that TST is differentially expressed

in tumor-derived epithelial cells. This study found that TST is significantly underexpressed in CESC (AUC = 0.820), and GSEA analysis revealed that the TST-high expression subgroup is primarily enriched in the arachidonic acid and glutathione metabolic pathways. This finding has important functional implications: downregulation of TST may lead to a decrease in the ability to scavenge reactive oxygen species within tumor cells and an accumulation of pro-inflammatory metabolites, thereby creating an inflammatory microenvironment conducive to tumor progression. Notably, TST has not been identified as a hub gene in studies on isonitration in hepatocellular carcinoma, suggesting that the association between TST and isonitration may be tumor-type-specific[9]. The uniqueness of TST is further evident in virtual knockout simulations—no significant global transcriptomic perturbations were detected following TST knockout, in stark contrast to CDKN2A (approximately 2.0%) and XPO1 (approximately 2.8%). This result suggests that the functional regulation of TST may be more focused on the metabolite level rather than direct transcriptional regulatory networks[26]. Given that metabolic reprogramming has been identified as a key feature of cervical cancer, an integrated analysis of metabolomics and transcriptomics may be a more appropriate research approach for metabolic enzyme genes such as TST[27].

Analysis of intercellular communication further revealed the differential remodeling of ligand-receptor networks in the tumor microenvironment by the expression states of these three genes. This study found that the CDKN2A low-expression group exhibited a more complex and intense communication network, whereas the XPO1 and TST high-expression groups displayed more active communication, suggesting that different gene expression states may influence the overall landscape of the tumor microenvironment by differentially regulating intercellular signaling. Notably, several key ligand-receptor pairs in the MIF signaling pathway (APP–CD74, MIF–[CD74 + CXCR4], and MIF–[CD74 + CD44]) exhibited significant differences across multiple group comparisons[28]. MIF (macrophage migration inhibitory factor) is a multifunctional pro-inflammatory cytokine and tumor-promoting factor that is highly expressed in various malignant tumors and is associated with poor prognosis[29]. Upon binding to CD74, MIF activates the NF- κ B and ERK signaling pathways, promoting tumor proliferation, angiogenesis, and the establishment of an immunosuppressive microenvironment[30]. Previous studies have confirmed that MIF promotes tumor invasion and migration in cervical cancer, and MIF gene polymorphisms are closely associated with cervical cancer risk and the infiltration of CD74 + tumor-associated macrophages[31–34]. In this study, compared with the low-expression group, the communication probability of the aforementioned ligand-receptor pair was significantly higher in the CDKN2A high-expression group, suggesting that CDKN2A may promote malignant interactions between tumor cells and immune cells by upregulating the activity of the MIF signaling pathway. Furthermore, tumor-associated macrophages have been shown to be the primary source of MIF in the cervical cancer microenvironment, which is consistent with the finding in this study that XPO1 is highly expressed in macrophages[33, 34]. Together, these results suggest that isocyanurate-related genes may reshape the pattern of intercellular communication in the tumor microenvironment by regulating the MIF-CD74 signaling axis, providing a new perspective on understanding the non-autonomous functions of the three core genes in CECs.

The expression trends of the three genes were markedly different—CDKN2A and XPO1 were upregulated, while TST was downregulated—but the diagnostic receiver operating characteristic (ROC) curve constructed using all three genes achieved an AUC of 0.952, significantly outperforming any single gene. Previous studies have reported various cervical cancer diagnostic models based on multigene combinations; for example, a four-gene diagnostic model integrating CCND1, TRIP13, MYBL2, and GNB4, with an AUC of 0.989; a model based on genetic algorithms and support vector machines (GA-SVM) achieving an accuracy of approximately 99%; and a computational model based on p16/Ki67 expression with an AUC of 0.96 in an independent test set[35–37]. Compared with the aforementioned models, the line graph developed in this study demonstrates a competitive advantage in diagnostic accuracy. Furthermore, the three genes synergistically drive the progression of cervical cancer from three distinct dimensions—proliferation control (CDKN2A), molecular transport (XPO1), and metabolic adaptation (TST)—and this multidimensional functional complementarity may be the underlying reason for the exceptional performance of the combined diagnostic model.

This study opens up new avenues for research into the role of isocyanation in cervical cancer, and its theoretical significance is primarily reflected on three levels. First, it marks the first time that the scope of isocyanation research has been expanded from liver cancer to cervical cancer, filling a knowledge gap regarding this type of modification in the field of gynecologic oncology. Second, through single-cell-level cellular localization and virtual knockout simulations, this study provides a new perspective for understanding the non-autonomous functions of isocyanate-modified genes in the tumor microenvironment—in particular, the high expression of XPO1 in macrophages, the unique phenotype following TST knockout, and the differential regulation of the MIF signaling pathway all suggest that isocyanate modification may indirectly regulate tumor progression by influencing immune cell function and intercellular communication[38]. Third, the MIF signaling pathway exhibited significant differences across multiple group comparisons, suggesting that isonitrile-modified genes may reshape the pattern of intercellular communication in the tumor microenvironment by regulating the MIF-CD74 signaling axis. This finding aligns with recent reports identifying the MIF signaling pathway as a key oncogenic pathway in HPV-associated cancers[39]. At the clinical translation level, the diagnostic flowchart constructed based on the three genes demonstrated excellent diagnostic performance, and clinical experience with XPO1 inhibitors in gynecological cancers provides a theoretical basis for expanding their application in cervical cancer[24].

This study has several limitations. First, all analyses are based on retrospective data from public databases, and there is a lack of prospective clinical cohorts and experimental validation. Second, there is a lack of direct experimental evidence for isonitrile modification itself; further studies are needed to identify the specific protein targets and modification sites of isonitrile modification in cervical cancer tissues and cell lines using mass spectrometry and specific antibodies. Third, both virtual knockout simulations and cell-to-cell communication predictions are computational inferences and require functional validation via CRISPR-Cas9 gene editing and in vivo animal models. Fourth, the single-cell data included only one tumor and one control sample; this small sample size may limit a comprehensive analysis of tumor microenvironment heterogeneity.

Looking ahead, the following directions warrant priority exploration: (1) Identify the proteome of isocyanate-modified proteins in cervical cancer tissues using immunoprecipitation-mass spectrometry (IP-MS) to determine whether CDKN2A, XPO1, and TST are directly subject to isocyanate modification; (2) Establish cell models with knockout or overexpression of these three genes, and combine metabolomics and transcriptomics to validate their roles in cervical cancer proliferation, migration, and metabolic reprogramming; (3) Validate the cell-type-specific expression patterns of these three genes and their association with immune cell infiltration in a larger single-cell sequencing cohort; (4) Further explore the therapeutic efficacy of the XPO1 inhibitor selinexor in preclinical models of cervical cancer and its potential for combination with immunotherapy, with a particular focus on its effects on the MIF signaling pathway and cellular communication within the tumor microenvironment. In summary, this study provides potential candidate targets and a theoretical basis for the molecular diagnosis and targeted therapy of cervical cancer.

5. Conclusions

The isocyanation-related diagnostic model based on the three genes CDKN2A, XPO1, and TST developed in this study provides an efficient computational tool for risk assessment of cervical squamous cell carcinoma and is closely associated with tumor cell cycle regulation, metabolic reprogramming, and immune microenvironment remodeling. This model may serve as a potential reference framework for the exploration of molecular diagnosis, risk stratification, and personalized treatment strategies for cervical cancer; however, its clinical utility still requires further validation in independent, large-scale prospective cohorts and real-world clinical settings.

Declarations

Data Availability

All datasets used in this study are publicly available. The CESC transcriptomic and clinical data are from the Cancer Genome Atlas (TCGA-CESC) and can be accessed at <https://portal.gdc.cancer.gov/>. The GEO datasets GSE39001, GSE63514, GSE9750, GSE52903, and GSE168652 are available at <https://www.ncbi.nlm.nih.gov/geo/>. The nicotinylation-related gene set (KRG) was extracted from the list of Kinic-related genes reported by Zhou et al. in a published paper [npj Precis Oncol, 2026;10:132 (Table S1)]. Other relevant data are provided in the main text and supplementary materials. Analysis scripts may be reasonably requested from the corresponding author.

Acknowledgements

Not applicable.

Funding

None.

Clinical trial number

Not applicable.

Author information

Authors and Affiliations

Department of Clinical Laboratory, Nanxishan Hospital of Guangxi Zhuang Autonomous Region, Guilin 541002

Zhengchao Yan, Chuanhua Tang, Minhui He, Minxia Li, Shuang Liang, Jianbin Yang, Fangting Luo, Jianlin Chen.

Department of Laboratory Medicine, The Second Affiliated Hospital of Guilin Medical University, Guilin 541100, China.

Sijuan Tang

Education Department, Liuzhou Hospital Affiliated to Youjiang Medical University for Nationalities, Liuzhou 545007, China.

Yuan Lin

Department of Obstetrics and Gynecology, Nanxishan Hospital of Guangxi Zhuang Autonomous Region, Guilin 541002, China.

Liwu Zhang

Contributions

Zhengchao Yan and Sijuan Tang contributed equally to this work. Zhengchao Yan led the conceptualization, formal analysis, investigation, methodology, software development, visualization, and wrote the original draft. Sijuan Tang performed data curation, formal analysis, investigation, visualization, and co-wrote the original draft. Chunhua Tang, Minxia Li, and Fangting Luo were responsible for data curation, investigation, and validation, with Fangting Luo additionally contributing to visualization. Minhui He carried out investigation and methodology, while Shuang Liang also assisted with visualization. Yuan Lin and Jianbin Yang handled formal analysis, software, and visualization. Liwu Zhang and Jianlin Chen jointly conceived the study, acquired funding, administered the project, provided resources, supervised the research, and critically revised the manuscript. All authors reviewed and approved the final version.

Corresponding author

Correspondence to Jianlin Chen and Liwu Zhang.

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Singh D, et al. Global estimates of incidence and mortality of cervical cancer in 2020: a baseline analysis of the WHO Global Cervical Cancer Elimination Initiative. *Lancet Glob Health*. 2023;11:e197–206.
2. Wilailak S, Kengsakul M, Kehoe S. Worldwide initiatives to eliminate cervical cancer. *Int J Gynaecol Obstet*. 2021;155(Suppl 1):102–6.
3. Baba SK, et al. Human papilloma virus (HPV) mediated cancers: an insightful update. *J Transl Med*. 2025;23:483.
4. Boldeanu L et al. Cervical Cancer in the Era of HPV: Translating Molecular Mechanisms into Preventive Public Health Action. *Int J Mol Sci* 26(2025).
5. Ma L, Luan Y, Lu L. Analyze the Diversity and Function of Immune Cells in the Tumor Microenvironment From the Perspective of Single-Cell RNA Sequencing. *Cancer Med*. 2025;14:e70622.
6. Xu N, Shao J, Jiang L, Zhao H, Zheng W. Post-translational modifications as hubs of amino acid metabolism in cancer: From oncogenic signaling to therapeutic targeting. *Biochim Biophys Acta Rev Cancer*. 2026;1881:189653.
7. Lothrop AP, Torres MP, Fuchs SM. Deciphering post-translational modification codes. *FEBS Lett*. 2013;587:1247–57.
8. Xu H, Wu M, Ma X, Huang W, Xu Y. Function and Mechanism of Novel Histone Posttranslational Modifications in Health and Disease. *Biomed Res Int* 2021, 6635225 (2021).
9. Jiang Y, et al. Isonicotinylation is a histone mark induced by the anti-tuberculosis first-line drug isoniazid. *Nat Commun*. 2021;12:5548.
10. Li Y, et al. Global isonicotinylome analysis identified SMAD3 isonicotinylates promotes liver cancer cell epithelial-mesenchymal transition and invasion. *iScience*. 2024;27:110775.

11. Cesaro G et al. Advances and challenges in cell-cell communication inference: a comprehensive review of tools, resources, and future directions. *Brief Bioinform* 26(2025).
12. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*. 2008;9:559.
13. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods*. 2015;12:453–7.
14. Sherr CJ. The INK4a/ARF network in tumour suppression. *Nat Rev Mol Cell Biol*. 2001;2:731–7.
15. Klaes R, et al. Overexpression of p16(INK4A) as a specific marker for dysplastic and neoplastic epithelial cells of the cervix uteri. *Int J Cancer*. 2001;92:276–84.
16. Elsayim R et al. Multi-Omics Analysis of CDKN2A (p16INK4a) in Cervical Carcinoma in the Context of Human Papillomavirus and in Endometrial Carcinoma. *Genes (Basel)* 17(2026).
17. Luan Y, Zhang W, Xie J, Mao J. CDKN2A inhibits cell proliferation and invasion in cervical cancer through LDHA-mediated AKT/mTOR pathway. *Clin Transl Oncol*. 2021;23:222–8.
18. Li YQ, Fang Z, Zhang W, Rao GW, Zheng Q. Targeting XPO1 for fighting relapsed/refractory diseases: The research progress of XPO1 inhibitors. *Bioorg Chem*. 2025;154:108073.
19. Özdaş S, Canatar İ. Targeting of nucleo–cytoplasmic transport factor exportin 1 in malignancy (Review). *Med Int (Lond)*. 2022;2:2.
20. Xia L, et al. The effect of aberrant expression and genetic polymorphisms of Rad21 on cervical cancer biology. *Cancer Med*. 2018;7:3393–405.
21. Zhou J et al. Kinic index: an artificial intelligence-driven predictive model and multitarget drug discovery framework for hepatocellular carcinoma patients. *NPJ Precis Oncol* 10(2026).
22. Fisher JG, et al. Modulation of anti-tumour immunity by XPO1 inhibitors. *Explor Target Antitumor Ther*. 2025;6:1002310.
23. Kady N, et al. The GATA-3-dependent transcriptome and tumor microenvironment are regulated by eIF4E and XPO1 in T-cell lymphomas. *Blood*. 2025;145:597–611.
24. Vergote IB, et al. Phase 2 study of the Exportin 1 inhibitor selinexor in patients with recurrent gynecological malignancies. *Gynecol Oncol*. 2020;156:308–14.
25. Luo Y, et al. Thiosulphate sulfurtransferase: Biological roles and therapeutic potential. *Redox Biol*. 2025;82:103595.
26. Osorio D, et al. scTenifoldKnk: An efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. *Patterns (N Y)*. 2022;3:100434.
27. Li B, Sui L. Metabolic reprogramming in cervical cancer and metabolomics perspectives. *Nutr Metab (Lond)*. 2021;18:93.
28. Xie W, Dong Y, Lv J, Wang X. Macrophage migration inhibitory factor superfamily in tumor metabolism: mechanistic insights and therapeutic potential. *Biochem Pharmacol*. 2026;246:117720.
29. Youness RA, et al. Macrophage migration inhibitory factor (MIF) and the tumor ecosystem: a tale of inflammation, immune escape, and tumor growth. *Front Immunol*. 2025;16:1636839.

30. Mendoza-Reinoso V et al. Bone Marrow Macrophages Induce Inflammation by Efferocytosis of Apoptotic Prostate Cancer Cells via HIF-1 α Stabilization. *Cells* 11(2022).
31. Zhang Q, Wang M, Wu S. Macrophage migration inhibitory factor in inflammasome formation and macrophage recruitment by cervical squamous cell carcinoma cells. *Oncol Lett.* 2025;29:164.
32. Cheng RJ, Deng WG, Niu CB, Li YY, Fu Y. Expression of macrophage migration inhibitory factor and CD74 in cervical squamous cell carcinoma. *Int J Gynecol Cancer.* 2011;21:1004–12.
33. Wu S, et al. Macrophage migration inhibitory factor promoter polymorphisms (-794CATT(5–7)) as potential biomarker for early-stage cervical cancer. *J Obstet Gynaecol Res.* 2017;43:571–9.
34. Wang Z et al. Targeting tumor-associated macrophage-derived CD74 improves efficacy of neoadjuvant chemotherapy in combination with PD-1 blockade for cervical cancer. *J Immunother Cancer* 12(2024).
35. Zhang H et al. Integrated bioinformatics, machine learning, and experimental validation identify a four-gene diagnostic signature for cervical cancer associated with PI3K/AKT signaling. *Sci Rep* 16(2026).
36. Nezamabadi Farahani L, Kazemnejad A, Afrasiabi M, Tapak L. Silico Transcriptomic Analysis for Identification of Potential Diagnostic and Prognostic Biomarkers and Therapeutic Targets in Cervical Cancer using a Hybrid Genetic Algorithm-Support Vector Machine Approach. *Arch Iran Med.* 2025;28:677–86.
37. Gisca T et al. A Prospective Study on the Progression, Recurrence, and Regression of Cervical Lesions: Assessing Various Screening Approaches. *J Clin Med* 13(2024).
38. Fan J, Ryu V, Qin Z, Dai L. Roles of Macrophage Migration Inhibitory Factor (MIF) Signaling Pathway in Oncovirus Infection and Virus-Associated Cancers. *Viruses* 17(2025).
39. Huang S, et al. Single-cell analysis reveals the regulatory role of ECT2 in HPV-driven cervical carcinogenesis process. *Infect Agent Cancer.* 2025;20:66.

Figures

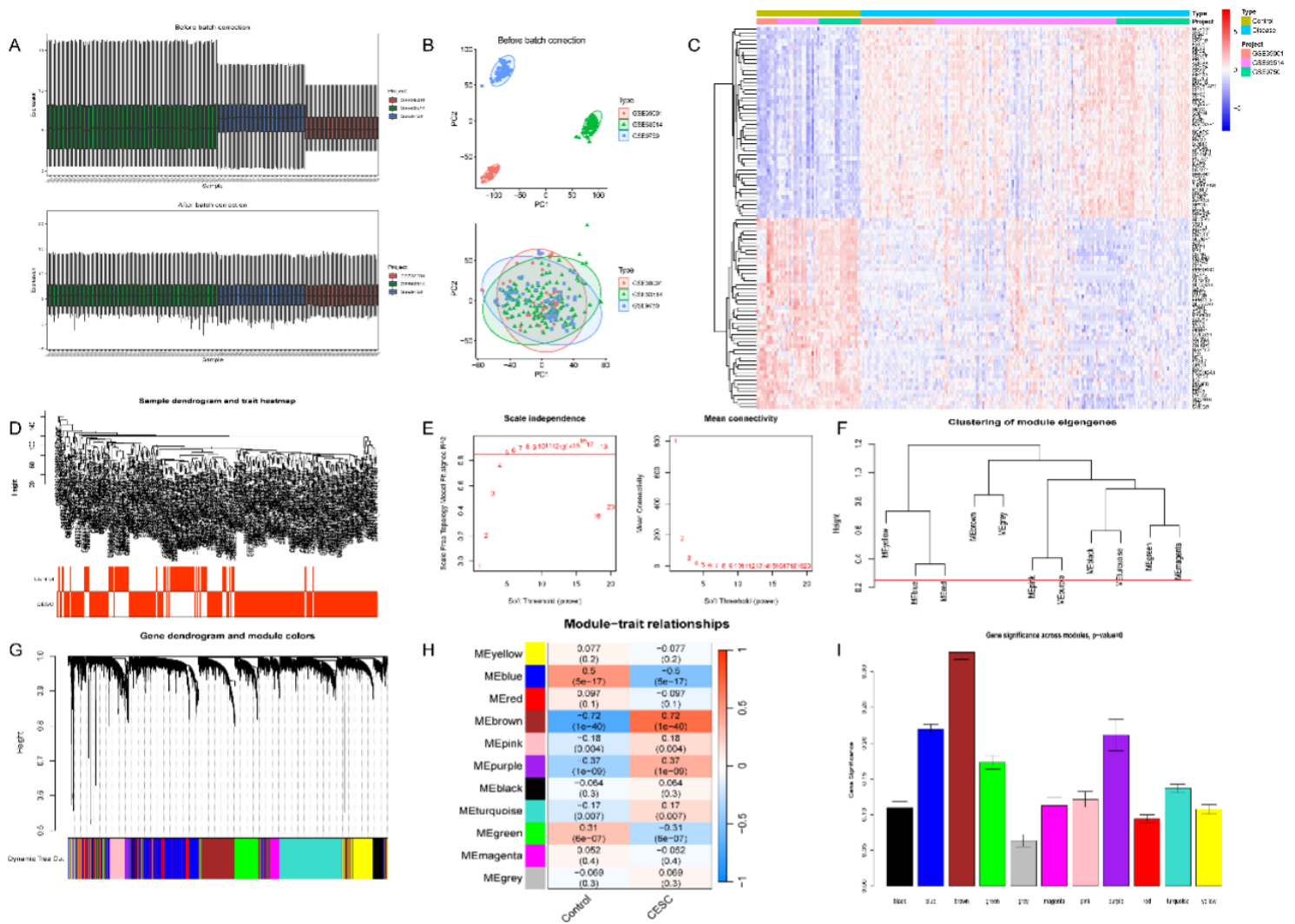


Figure 1

Batch effect correction and WGCNA integration analysis across multiple datasets. (A) Comparison of gene expression distributions in the three microarray datasets GSE39001, GSE63514, and GSE9750 before and after batch effect correction; technical background differences among samples were effectively eliminated after correction. (B) Principal component analysis (PCA) scatter plots of expression profiles before and after correction, reflecting the mixing of samples after batch effects were removed. (C) Heatmap of whole-gene expression after correction; the color gradient represents expression levels (red for high, blue for low), and the annotation bar at the top identifies disease groups (Control/CESC) and data sources. (D) Sample clustering tree and corresponding clinical trait heatmap, demonstrating expression similarity among samples and associations with disease groups. (E) Soft-threshold analysis plot; based on the criterion that the scale-free topological fit index (Scale-free R^2) first reached 0.9, the optimal soft threshold was determined to be 5. (F) Hierarchical clustering tree of characteristic genes (MEs) in co-expression modules. (G) Gene phylogenetic tree and co-expression modules partitioned by Dynamic Tree Cut; the color bands below represent different modules. (H) Module-trait association heatmap; the values in the matrix represent the Pearson correlation coefficients

between each module and the two groups, with significance P-values in parentheses. (I) Statistical comparison of gene significance within each co-expression module, revealing the overall distribution of association strength between each module and the phenotype

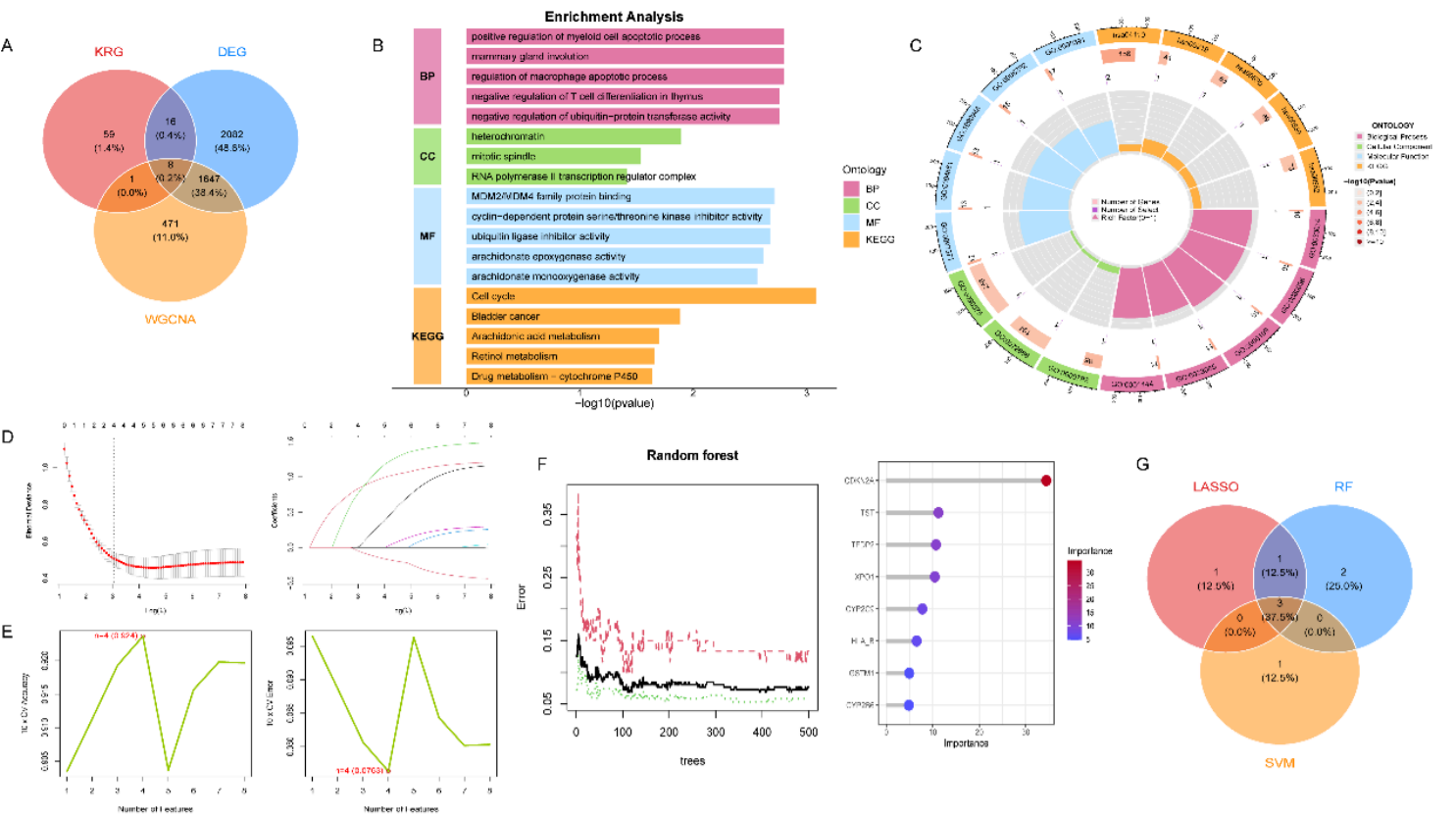


Figure 2

Identification of key isonitration genes, functional enrichment analysis, and cross-validation of machine learning algorithms. (A) A Venn diagram showing the intersection among differentially expressed genes (DEGs), genes from key modules identified by weighted gene co-expression network analysis (WGCNA), and isonitration-related genes (KRGs); a total of 8 overlapping genes were identified. (B–C) Results of GO and KEGG functional enrichment analyses for the overlapping genes. Figure B is a bar chart showing the top-ranked entries categorized by biological process (BP), cellular component (CC), molecular function (MF), and KEGG pathway; Figure C is a pie chart displaying the specific number of genes in each enriched pathway, their classification, and significance levels. (D) The feature selection process for the optimal LASSO regression model; the x-axis represents the logarithm of λ , the y-axis represents cross-validation bias, and the vertical dashed line indicates the λ value and number of features corresponding to the optimal model. (E) A line graph showing model evaluation metrics (accuracy and AUC) under 10-fold cross-validation (10-fold CV) as a function of the number of included features, used to determine the optimal feature combination for the model. (F) Results of the RF algorithm: The left figure shows a curve where the model error rate converges as the number of decision trees increases; the right figure shows a

ranking of the Gini importance scores for each feature gene. (G) A Venn diagram illustrating the intersection of feature genes selected by the three machine learning algorithms (LASSO, RF, and SVM)

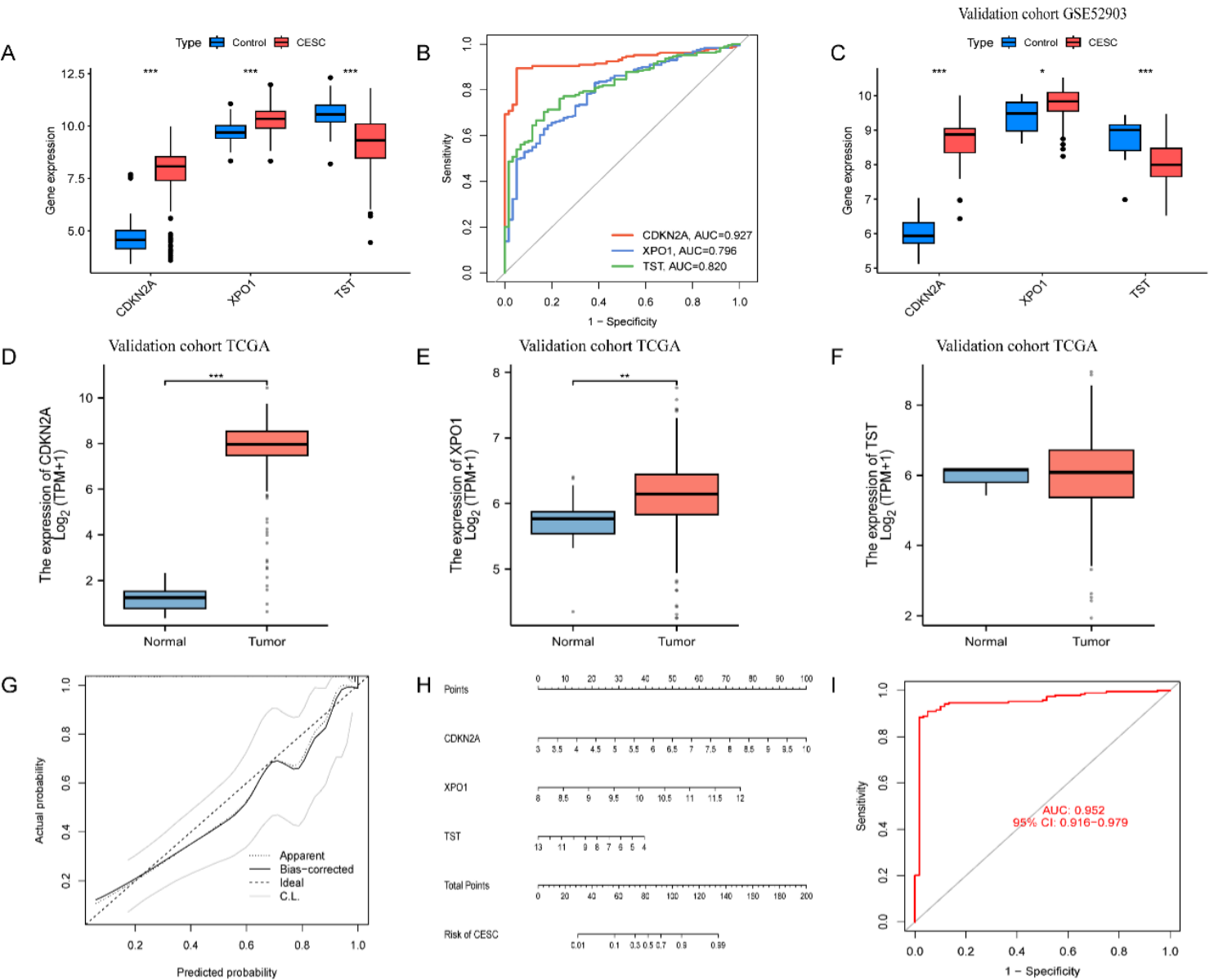


Figure 3

Construction and validation of diagnostic ROC curves. (A) Box-and-whisker plots showing differences in the expression of the CDKN2A, XPO1, and TST genes between the Control and CESC groups in the discovery cohort (**P < 0.01, ***P < 0.001). (B) ROC curves for the three genes when used individually as diagnostic markers in the discovery cohort; the area under the curve (AUC) for each gene is labeled in the legend. (C) Differences in expression levels of the three genes between the Control and CESC groups in the external validation dataset GSE52903 (*P < 0.05, ***P < 0.001). (D–F) From left to right: comparisons of mRNA expression levels of the CDKN2A, XPO1, and TST genes in Normal and Tumor samples from the TCGA validation cohort (**P < 0.01, ***P < 0.001). (G) Calibration plot of the diagnostic receiver operating characteristic (ROC) curve, with predicted probability on the x-axis and actual probability on the y-axis;

the dashed line represents “Ideal,” while the solid and dotted lines represent “Bias-corrected” and “Apparent,” respectively. (H) A CESC diagnostic receiver operating characteristic (ROC) plot constructed based on the expression levels of the three genes CDKN2A, XPO1, and TST; “Total Points” are calculated by summing the scores for each gene and are then mapped to the corresponding “Risk of CESC.” (I) The overall ROC curve for the diagnostic receiver operating characteristic (ROC) model demonstrates the model’s overall ability to distinguish between cancer and normal samples (AUC = 0.952, 95% CI = 0.916–0.979)

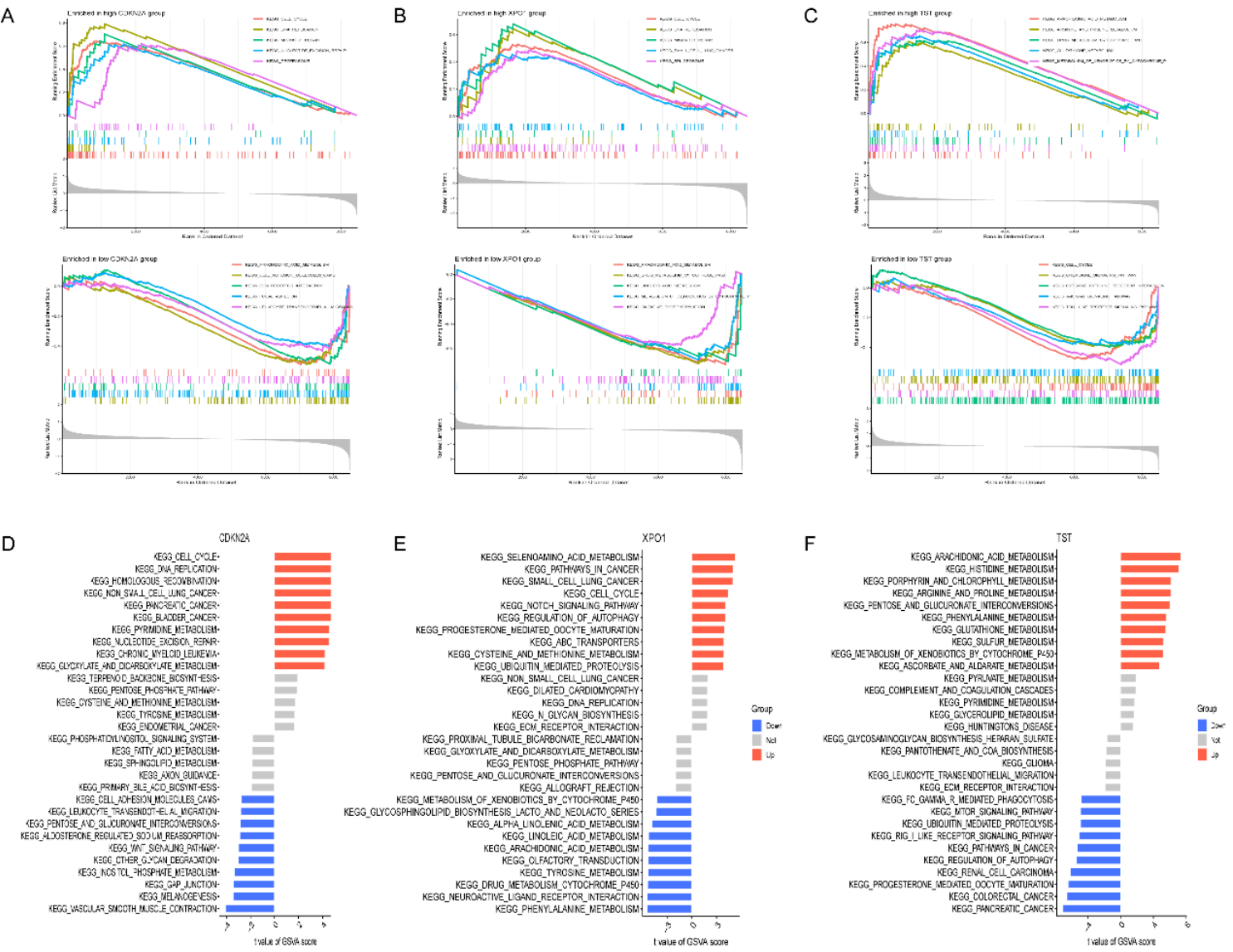


Figure 4

GSEA and GSVA functional pathway enrichment analyses. (A–C) Gene set enrichment analysis results for the three core genes. The figures show the KEGG signaling pathways significantly enriched in the high-expression group (top row) and low-expression group (bottom row) for CDKN2A, XPO1, and TST (C), respectively. (D–F) Gene Set Variation Analysis results. The bar charts show the differences in GSVA scores between the high- and low-expression subsets for CDKN2A, XPO1, and TST, respectively (the x-axis represents t-values). Red bars (Up) indicate that a pathway is significantly enriched in the high-

expression group, blue bars (Down) indicate that a pathway is enriched in the low-expression group, and gray bars indicate that the difference did not reach statistical significance

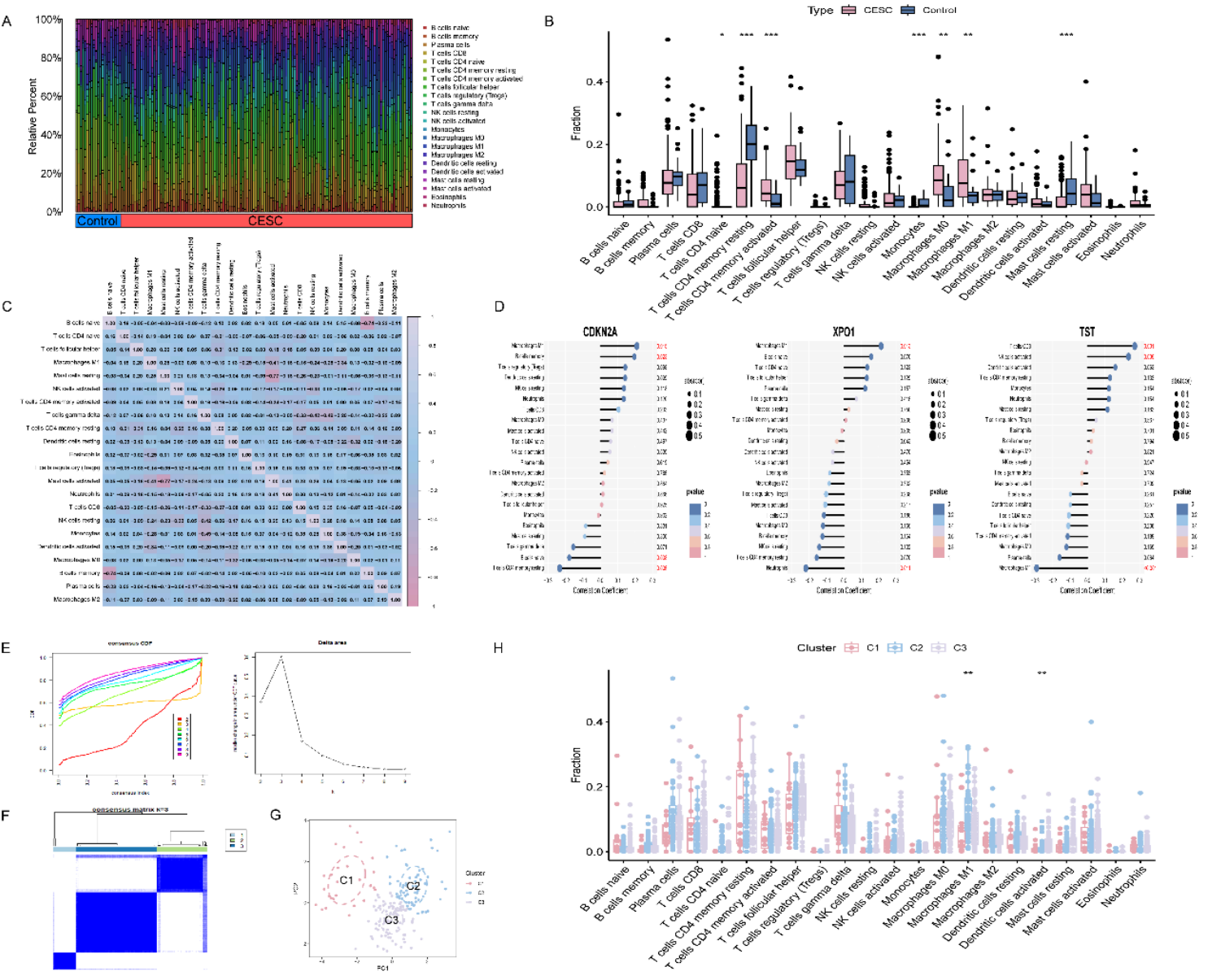


Figure 5

Characteristics of the CESC immune microenvironment and molecular subtyping. (A) A stacked bar chart shows the relative proportions of 22 immune cell types across different samples in the CESC group and the control group. (B) Box plots compare the statistical differences in the infiltration abundance of various immune cell types between the CESC group and the control group (*P < 0.05, **P < 0.01, ***P < 0.001). (C) A heatmap showing correlations among the abundances of the 22 immune cell types, with red and blue representing positive and negative correlations, respectively. (D) A lollipop plot illustrating the Spearman correlations between three core genes (CDKN2A, XPO1, and TST) and the infiltration abundances of various immune cell types. Dot size represents the p-value, and the color gradient represents the correlation coefficient. (E) Cumulative distribution function (CDF) curve (left) and

line plot of the area under the CDF curve (right), used to determine the optimal number of clusters (k = 3) for immune subtyping. (F) Heatmap of the consensus clustering matrix with k = 3, clearly illustrating the sample distribution across the three subtypes: C1, C2, and C3. (G) Principal component analysis (PCA) scatter plot based on the three molecular subtypes, used to assess differences in sample distribution among subtypes. (H) Box plots showing statistical comparisons of the abundance of 22 types of immune cell infiltration among the three immune molecular subtypes (C1, C2, and C3)

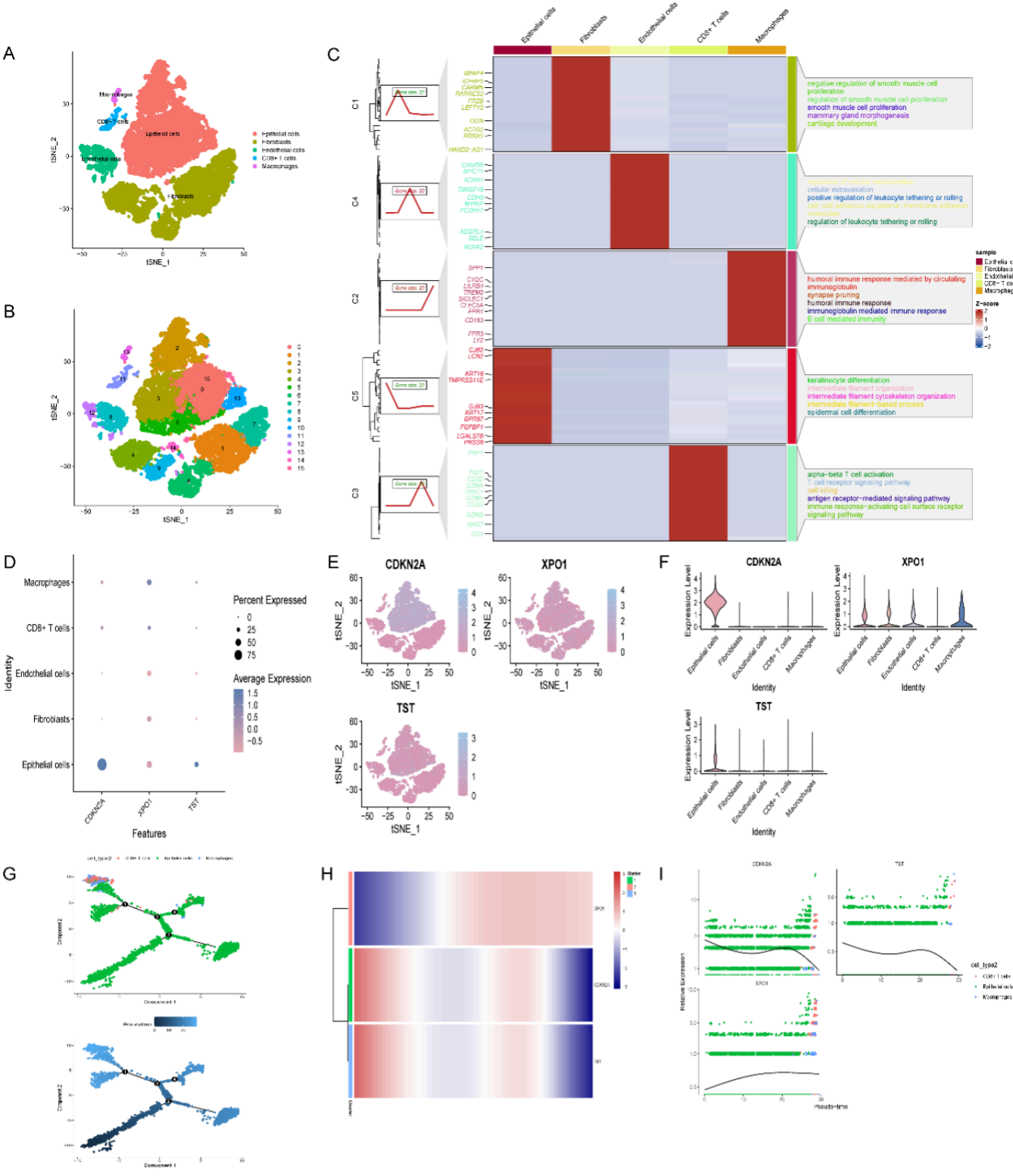


Figure 6

Identification of key diagnostic gene-based cell subpopulations and pseudo-temporal trajectory analysis based on single-cell RNA sequencing. (A) A t-SNE plot showing the distribution of cervical cancer cell subpopulations. (B) A t-SNE plot presenting the annotation results for cervical cancer cell subpopulations. (C) Heatmaps of cell-type-specific marker genes and functional enrichment; the left side shows the expression profiles of each subpopulation, while the right side lists the corresponding specific biological processes. (D) A bubble plot showing the expression proportions (dot size) and average expression levels (dot color) of core genes (CDKN2A, XPO1, TST) across five major cell types. (E) A feature expression plot of core genes (CDKN2A, XPO1, TST) on the t-SNE plot, visually reflecting the spatiotemporal distribution of gene expression. (F) A violin plot shows a comparison of the absolute expression levels of the three genes—CDKN2A, XPO1, and TST—across the five major cell types. (G) Pseudo-temporal developmental trajectory plots of cells, illustrating the evolution of cell states and their arrangement paths along the pseudo-temporal dimension. (H) A pseudo-time heatmap showing the expression dynamics of different genes over the pseudo-time axis (from left to right), with color gradients representing gene expression levels (red for high expression, blue for low expression). (I) Expression curves of core genes along pseudo-time trajectories, illustrating the dynamic trends in their expression levels throughout the developmental trajectory

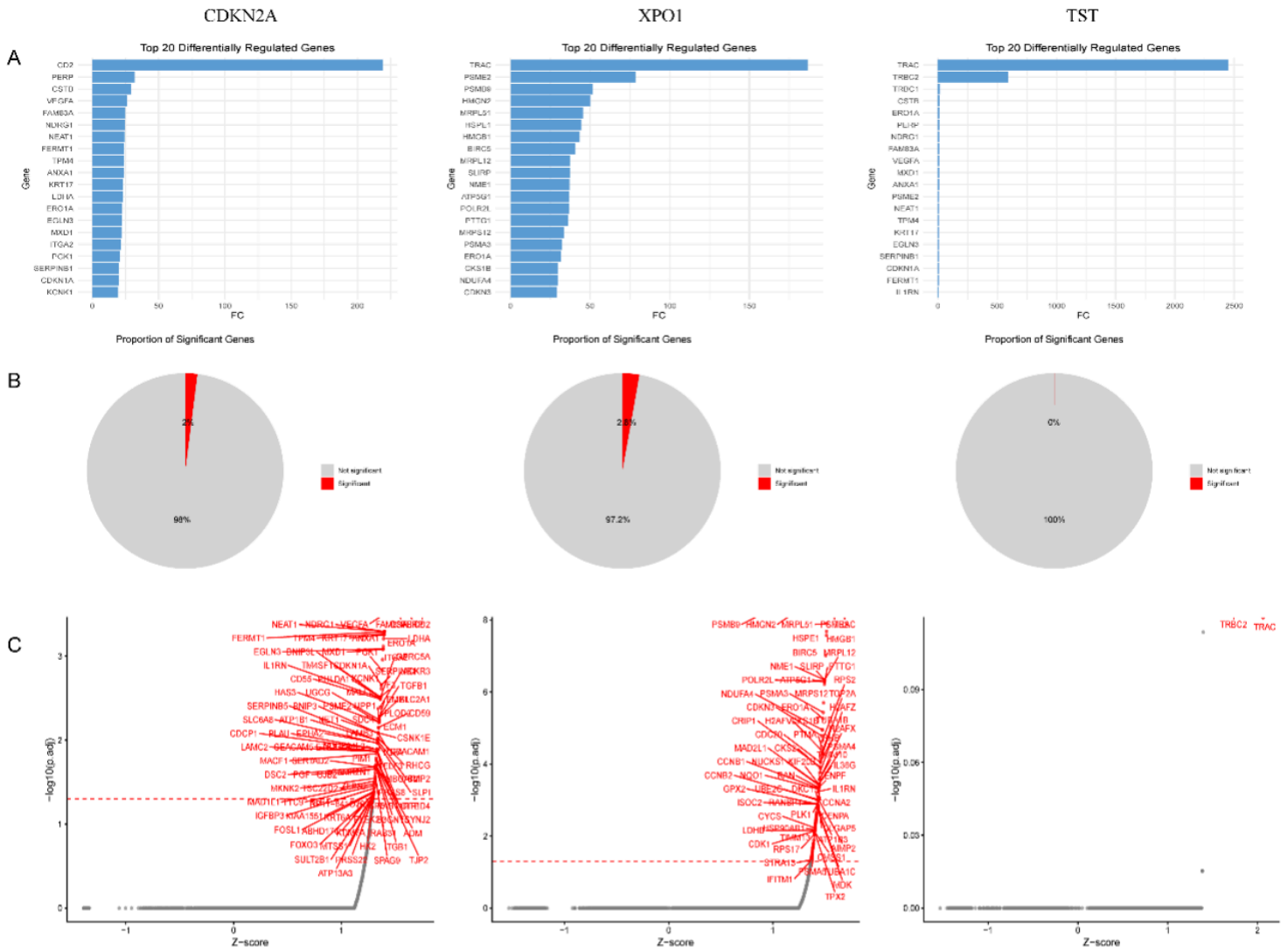


Figure 7

Analysis of transcriptional differential regulation following virtual knockout of the CDKN2A, XPO1, and TST genes. (A) Bar charts showing the top 20 differentially regulated genes, sorted by fold change (FC), following virtual knockout of the CDKN2A, XPO1, and TST genes, respectively. The x-axis represents FC values, and the y-axis represents gene names. (B) Pie charts showing the proportions of significantly differentially regulated genes (red sectors, $p_{adj} < 0.05$) and non-significantly differentially regulated genes (gray sectors, $p_{adj} \geq 0.05$) among all genes following virtual knockout of the three genes, respectively. The numerical values in the center of each pie chart represent the percentage for the corresponding category. (C) Volcano plots showing differentially regulated genes following the virtual knockout of each of the three genes. The x-axis represents the Z-score (standardized deviation), and the y-axis represents the negative logarithm of the adjusted p-value ($-\log_{10}(p_{adj})$). The red horizontal dashed line in the figure represents the threshold for statistical significance ($p_{adj} = 0.05$). Points above the threshold line indicate significantly differentially regulated genes.

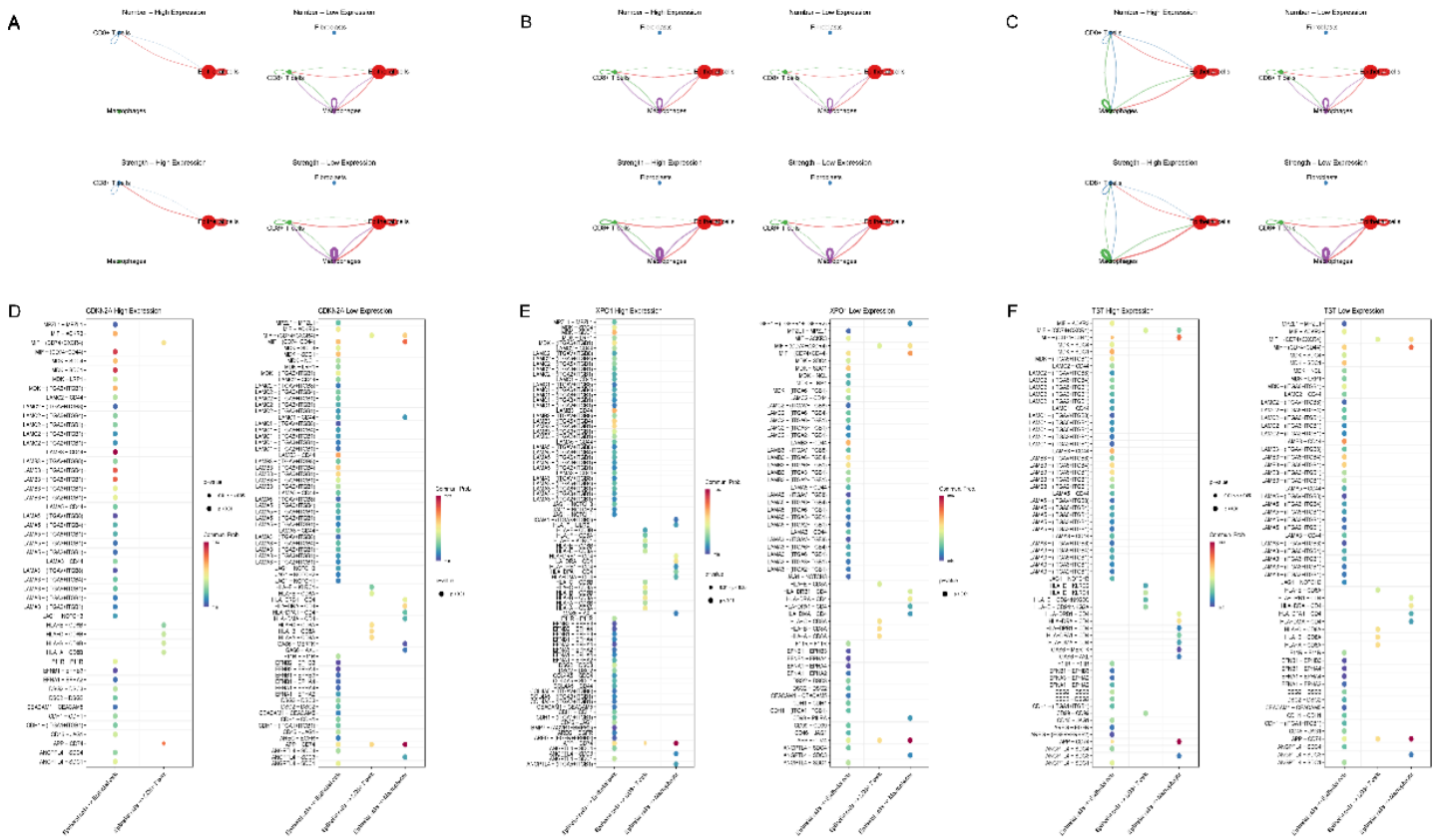


Figure 8

Effects of CDKN2A, XPO1, and TST expression profiles, based on single-cell transcriptomic data, on intercellular communication networks and ligand-receptor interactions (A–C). Cells were divided into high- and low-expression groups based on the median expression levels of the CDKN2A (A), XPO1 (B), and TST (C) genes into high-expression and low-expression groups, showing the global intercellular communication networks among major cell populations (CD8+ T cells, macrophages, fibroblasts, and endothelial cells). The top row (Number) indicates the number of predicted interactions between different cell populations, while the bottom row (Strength) indicates the predicted interaction strength. The size of circular nodes represents the proportion of cells in that cell population, and the thickness of the lines connecting nodes reflects the number or strength of interactions. (D–F) show the communication probabilities of key ligand-receptor interaction pairs in the high- and low-expression groups for the CDKN2A (D), XPO1 (E), and TST (F) genes, respectively. On the Y-axis, the figure represents ligand-receptor molecular pairs, while the X-axis represents the cell types involved in the interactions. The color gradient of the bubbles ranges from blue to red, indicating communication probability from low to high; the diameter of the bubbles represents statistical significance (p-value)