Cancer Medicine

WILEY

Cancer Medicine Open Access

RESEARCH ARTICLE OPEN ACCESS

Integrative Single-Cell and Machine Learning Analysis Identifies an EMT-Associated Prognostic Signature for Papillary Thyroid Cancer

Tianfeng Xu[1,2] | Ruonan Sun[1] | Yujie Zhang[3] | Xun Zheng[1]

[1]Division of Thyroid Surgery, Department of General Surgery, West China Hospital, Sichuan University, Chengdu, China | [2]Precision Medicine Center, West China Hospital, Sichuan University, Chengdu, China | [3]Department of Ultrasonography, West China Hospital, Sichuan University, Chengdu, China

Correspondence: Xun Zheng (zhengxun91@wchscu.cn)

Received: 17 October 2025 | **Revised:** 9 February 2026 | **Accepted:** 20 March 2026

Keywords: 101 machine learning algorithm combinations | epithelial mesenchymal transition | papillary thyroid cancer | prognostic genes | single-cell RNA sequencing analysis

ABSTRACT

Background: Epithelial–mesenchymal transition (EMT) plays a critical role in tumor progression; however, the underlying molecular mechanisms of EMT in papillary thyroid carcinoma (PTC) remain incompletely understood. This study aimed to investigate EMT-related mechanisms in PTC using an integrative approach combining single-cell RNA sequencing and machine learning.

Methods: Differentially expressed genes (DEGs) between PTC and normal thyroid tissues were identified, and EMT-related candidate genes were obtained by intersecting DEGs with EMT-related genes (EMT-RGs). Prognostic genes were screened using univariate Cox regression, and a risk model was constructed based on 101 machine learning algorithm combinations. Patients were stratified into high- and low-risk groups (HRG and LRG) according to risk scores, and the model was validated in an internal cohort. Additional analyses included nomogram construction, immune infiltration profiling, tumor mutational burden (TMB) assessment, drug sensitivity prediction, and molecular regulatory network analysis. Prognostic gene expression was further validated in vitro.

Results: Eight EMT-related prognostic genes (TYRO3, E2F1, TNFSF15, TGFBR3, PTX3, FHL2, SNAI1, and WT1) were identified. Patients in the HRG exhibited significantly poorer overall survival than those in the LRG. The nomogram showed good predictive accuracy for survival estimation. Immune infiltration analysis revealed significant differences between risk groups across six immune-related features. Splice site–related mutations were predominantly observed in the LRG but were absent in the HRG. Drug sensitivity analysis indicated higher sensitivity to BIRB.0796 in the LRG, whereas ABT-263, AG-014699, BX-795, and DMOG were more effective in the HRG. Single-cell analysis identified fibroblasts as key cell populations, with FHL2, PTX3, and TGFBR3 showing increased activity during critical differentiation stages. In vitro experiments confirmed expression patterns consistent with bioinformatics findings.

Abbreviations: AUC, Area Under Curve; cor, correlation; DE-circRNAs, differentially expressed circRNAs; DEGs, differentially expressed genes; DE-lncRNAs, differentially expressed lncRNAs; DE-miRNAs, differentially expressed miRNAs; EMT, Epithelial-mesenchymal transition; EMT-RGs, EMT-related genes; GEO, Gene Expression Omnibus; GO, Gene Ontology; HR, Hazard ratio; HRG, high risk group; IC_{50}, half maximal inhibitory concentration; IHC, immunohistochemistry; KEGG, Kyoto Encyclopedia of Genes and Genomes; LRG, low risk group; PH, proportional hazards; PTC, papillary thyroid cancer; RAI, radioactive iodine; ROC, receiver operating characteristic; RT-qPCR, reverse transcription quantitative PCR; scRNA-seq, Single-cell RNA sequencing; TCGA, Cancer Genome Atlas; TIDE, tumor immune dysfunction and exclusion; TIP, tracking tumor immunophenotype; TMB, tumor mutational burden.

Tianfeng Xu and Ruonan Sun contributed equally to this work.

Conclusion: This study identifies eight EMT-related prognostic genes in PTC and highlights their potential value as biomarkers for prognostic evaluation and therapeutic stratification.

1 | Introduction

Papillary thyroid carcinoma (PTC), the most common endocrine malignancy, represents 80%–90% of thyroid cancers [1], with a rapidly rising global incidence attributed to improved diagnostics and environmental factors like radiation exposure during childhood [2, 3]. Although most patients achieve favorable outcomes following surgery and radioactive iodine (RAI) therapy, approximately 10%–15% develop aggressive disease characterized by RAI resistance, recurrence, or distant metastasis—clinical phenotypes that are inadequately predicted by the current TNM staging system [4, 5]. This prognostic uncertainty highlights an urgent need for molecular biomarkers capable of early identification of high-risk patient subsets, with epithelial–mesenchymal transition (EMT)-related pathways emerging as particularly promising candidates due to their established roles in metastatic dissemination [6].

EMT is a conserved biological process in which epithelial cells lose apical–basal polarity and acquire mesenchymal features, thereby gaining migratory and invasive capabilities [7, 8]. Beyond its physiological roles in embryonic development and wound healing, EMT contributes to cancer progression by promoting immune evasion, therapeutic resistance, and metastatic spread [9]. In PTC, EMT activation has been associated with extrathyroidal extension, lymph node metastasis, and tumor dedifferentiation [10]. Key transcriptional regulators, including SNAI1 and TWIST1, repress epithelial markers such as E-cadherin while inducing mesenchymal markers like vimentin [11, 12]. Nevertheless, a systematic and prognostically relevant characterization of EMT-related genes (EMT-RGs) in PTC remains largely lacking.

Single-cell RNA sequencing (scRNA-seq) has revolutionized cancer research by enabling high-resolution dissection of tumor ecosystems at the cellular level [13, 14]. In PTC, scRNA-seq studies have identified rare EMT-transitioning tumor cells, fibroblast-driven stromal remodeling, and immunosuppressive microenvironments dominated by M2 macrophages [10, 15, 16]. most existing studies have focused on descriptive cellular heterogeneity and have not effectively translated these insights into clinically actionable prognostic biomarkers. Integrating scRNA-seq with machine learning provides a powerful opportunity to map EMT-related transcriptional programs across cellular states and to uncover novel therapeutic vulnerabilities.

Conventional prognostic models for PTC often rely on limited gene panels or single-algorithm approaches, which may introduce bias and instability when applied to high-dimensional transcriptomic data [17, 18]. To overcome these limitations, we applied ten distinct machine learning algorithms—including CoxBoost, DeepSurv, and XGBoost—to analyze transcriptomic profiles from 712 patients. This multi-algorithm strategy emphasizes genes with consistent prognostic relevance across diverse computational paradigms, thereby reducing overfitting and improving model robustness and generalizability [19, 20].

To address these challenges, we propose a novel computational–experimental framework that integrates scRNA-seq analysis with 101 combinations of machine learning algorithms. This approach is designed to overcome the limitations of traditional single-model strategies by resolving EMT-driven cellular dynamics at single-cell resolution and prioritizing prognostic genes supported by cross-algorithm consensus. Through this integrative strategy, we aim to establish robust EMT-related prognostic signatures and to identify therapeutic vulnerabilities associated with stromal–immune remodeling. Collectively, our study advances the understanding of EMT in PTC and facilitates the translation of multi-omics discoveries into clinically actionable tools for risk stratification and targeted intervention in thyroid cancer management.

2 | Method

The research workflow of this study was shown in Figure S1.

2.1 | Data Collection

Data on gene expression profiles, clinical characteristics, somatic mutations, and survival outcomes of PTC were obtained from The Cancer Genome Atlas (TCGA) database (http://cancergenome.nih.gov/) (TCGA-THCA; accessed on October 12, 2024). The TCGA-THCA cohort comprised 505 PTC tumor samples, of which 504 had complete survival information, as well as 59 normal thyroid tissue samples. All transcriptomic data were generated using high-throughput mRNA sequencing on the Illumina HiSeq 2000 platform. The TCGA-THCA PTC samples were randomly divided into a training set (TCGA-THCA-train, 70%) and an internal validation set (TCGA-THCA-test, 30%). CircRNA expression data (GSE205733) and single-cell RNA sequencing (scRNA-seq) data (GSE184362) related to PTC were retrieved from the Gene Expression Omnibus (GEO) database. The GSE205733 dataset (platform GPL20301) included four PTC tumor samples and four matched adjacent normal thyroid tissue samples. The GSE184362 dataset (platform GPL24676) consisted of seven PTC tumor samples and was used for downstream single-cell analyses. A total of 1184 EMT-related genes (EMT-RGs) were obtained from the dbEMT database (version 2.0) (https://bioinfo-minzhao.org/dbemt/) (Table S1).

2.2 | Identification of Differentially Expressed Genes (DEGs)

DEGs between PTC and normal thyroid tissues in the TCGA-THCA cohort were identified using the DESeq2 package (v1.34.0) [21] Genes with an absolute $\log_2$ fold change ($|\log_2 FC| > 1$ and an adjusted p value < 0.05 were considered significantly differentially expressed. Volcano plots were generated using the ggplot2 package (v3.4.4) (v 3.4.4) [22], with the top ten upregulated and downregulated genes annotated based on $\log_2 FC$ values. Heatmaps illustrating the top ten

upregulated and downregulated DEGs between PTC and normal tissues were constructed using the ComplexHeatmap package (v2.14.0) [23].

2.3 | Identification and Functional Analysis of Candidate Genes

Candidate genes were identified by intersecting DEGs with EMT-related genes (EMT-RGs) using the ggvenn function implemented in the VennDiagram package (v1.7.3) [24]. Functional enrichment analyses were subsequently performed to explore the biological significance of the candidate genes. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were conducted using the clusterProfiler package (v4.7.1.003) [25] employing the *enrichGO()* and *enrichKEGG()* functions, respectively, with a significance threshold of $p < 0.05$. For KEGG analysis, the top five significantly enriched pathways ranked by ascending p value were visualized using the GOplot package (v1.0.2) [26], while all significantly enriched GO terms were visualized using "ggplot2" package (v 3.4.4). To investigate protein–protein interactions (PPIs) among the candidate genes, a PPI network was constructed using the STRING database (v 12.0) (https://string-db.org/) (Interaction score > 0.4). The resulting network was visualized in Cytoscape (v 3.9.1) [27], with isolated nodes excluded from downstream visualization.

2.4 | Identification of Prognostic Genes

Prognostic genes were identified through a combined strategy integrating univariate Cox regression analysis and multiple machine learning algorithms. Univariate Cox proportional hazards regression was performed on candidate genes in the TCGA-THCA training cohort using the survival package (v3.3–1) [28]. Genes meeting the criteria of $p < 0.1$ and hazard ratio (HR) $\neq 1$ were retained. The proportional hazards (PH) assumption was evaluated using the *cox.zph()* function, and genes satisfying the PH assumption ($p > 0.05$) were considered eligible for subsequent modeling. The results of the univariate Cox analysis were visualized using forest plots generated with the forestplot package (v2.0.1) [29]. To further refine prognostic gene selection, ten machine learning algorithms were applied to the TCGA-THCA training and validation cohorts, generating a total of 101 algorithm combinations with 10-fold cross-validation. The algorithms included stepwise Cox regression (survival, v3.7.0), random survival forest (RSF; randomForestSRC, v3.2.3) [30], elastic net (Enet; glmnet, v4.1.8) (v 4.1.8) [31], least absolute shrinkage and selection operator (LASSO; glmnet, v4.1.8), partial least squares regression for Cox models (PLSR; plsRcox, v1.7.7) [32], CoxBoost (CoxBoost, v1.4) [33], support vector machine–recursive feature elimination (SVM-RFE; survivalsvm, v0.0.5) [34], Ridge regression (glmnet, v4.1.8), supervised principal components (SuperPC; SuperPC, v1.12) [35], and gradient boosting machines (GBM; GBM, v2.1.9) [36].

The 10 machine learning algorithms mentioned above differed in variable selection methods, applicable data types, model interpretability, and assumption conditions. By systematically integrating these algorithms into 101 distinct combinations, the complementary strengths of individual methods were leveraged while mitigating their respective limitations, thereby enhancing the robustness of the analytical framework. Then, the consistency index (C-index) for each model was calculated in TCGA-THCA-train and TCGA-THCA-test, and the models with the highest C-index were identified as optimal models (C-index > 0.6). Genes included in the optimal model were defined as prognostic genes, and patient-specific risk scores were calculated accordingly in both the training and validation cohorts.

2.5 | Construction and Validation of Risk Models

A prognostic risk model was constructed in the TCGA-THCA training cohort based on the risk scores derived from the optimal machine learning model. The formula for calculating the risk score was as follows:

$$\text{Riskscore} = \sum_{i=1}^{n} \text{coef}(\text{gene}_i) * \text{expr}(\text{gene}_i)$$

The risk score for each patient was calculated as the weighted sum of gene expression levels, where *Coef* represents the regression coefficient of each prognostic gene and *expr* denotes the corresponding gene expression level. Patients with available survival information were stratified into a low-risk group (LRG) and a high-risk group (HRG) according to the median risk score.

The distribution of risk scores and survival status was visualized using the survminer package (v0.4.9) [37] and the ggrisk package (v1.3.0) [38]. Kaplan–Meier survival curves were generated using survminer, and survival differences between the HRG and LRG were assessed using the log-rank test. Time-dependent receiver operating characteristic (ROC) curves at 1, 3, and 5 years were constructed using the survivalROC package (v1.0.3) [39] to evaluate the predictive performance of the risk model. The same procedures were applied to the TCGA-THCA validation cohort to assess model robustness.

2.6 | Association Between Risk Scores and Clinical Characteristics

In the TCGA-THCA training cohort, patients were stratified into subgroups according to clinicopathological characteristics, including T stage, N stage, M stage, age, and sex, to investigate the relationship between risk scores and clinical features. Kaplan–Meier survival curves were generated using the survminer package (v0.4.9). Differences in survival outcomes and risk score distributions between the HRG and LRG across clinical subgroups were evaluated using the Wilcoxon rank-sum test.

2.7 | Independent Prognostic Analysis and Nomogram Construction

To evaluate the significance of independent prognostic factors in patients with PTC, variables associated with prognosis were incorporated to construct a nomogram model. In the TCGA-THCA

training cohort, T stage, risk score, N stage, age, M stage, and gender were initially assessed using univariate Cox proportional hazards regression analysis ($p < 0.05$), with verification of the PH assumption ($p > 0.05$), implemented using the *survival* package (version 3.3–1). Subsequently, variables meeting the above criteria were further analyzed using multivariate Cox regression analysis ($p < 0.05$), and the PH assumption was again confirmed ($p > 0.05$) using the same package.

Based on the independent prognostic factors identified, a nomogram model was constructed in PTC samples with available survival information from the TCGA-THCA training cohort using the *rms* package (datadist() function; version 6.2–0) [40]. In the nomogram, each variable was assigned a corresponding score, and the sum of these scores generated a total point value. Higher total points were associated with a lower probability of patient survival. The predictive performance of the nomogram was evaluated using calibration curves generated with the *rms* package (calibrate() function; version 6.2–0) and time-dependent receiver operating characteristic (ROC) curves constructed using the *timeROC* package (version 0.4) [41]. The area under the curve (AUC) values for 1-, 3-, and 5-year survival exceeded 0.6, indicating satisfactory predictive accuracy of the model.

2.8 | Gene Set Variation Analysis (GSVA)

Gene set variation analysis (GSVA) was performed in PTC samples with available survival information from the TCGA-THCA training cohort using the *GSVA* package (gsva() function; version 1.46.0) [42] to evaluate pathway-level differences between the HRG and LRG. The reference gene set used for analysis was *c2.cp.v2023.2.Hs.symbols.gmt*, obtained from the Molecular Signatures Database (MSigDB) (https://www.gsea-msigdb.org/gsea/msigdb). Following GSVA, enrichment scores representing the relative activity of different pathways were calculated for each sample. Differential pathway enrichment between the HRG and LRG was subsequently assessed using the *DESeq2* package (version 1.34.0). Pathways with an adjusted p value < 0.05 were considered significantly different and were identified as key functional pathways.

2.9 | Immune Infiltration and ESTIMATE Score Analysis

To investigate immune cell infiltration in PTC, the CIBERSORT algorithm was applied using the *deconvo_tme()* function to estimate the relative abundances of 22 immune cell types [43] in PTC samples with available survival information from the TCGA-THCA training cohort. Differences in immune cell infiltration between the HRG and LRG were assessed using the Wilcoxon rank-sum test, with $p < 0.05$ considered statistically significant. The results were visualized using the *ggplot2* package (version 3.4.4). In addition, the *psych* package (version 2.1.6) [44] was used to analyze correlations between differentially infiltrated immune cells and the risk score, with |correlation coefficient (cor)| > 0.3 and $p < 0.05$ defined as the threshold for significance. Furthermore, the ESTIMATE algorithm was employed to evaluate the stromal and immune components of the TME in PTC samples from the TCGA-THCA training cohort. Stromal score, immune score, and ESTIMATE score were calculated using the *estimate* package (estimateScore() function; version 1.0.13) [45].

The stromal score reflects the presence of stromal cells, the immune score represents the level of immune cell infiltration, and the ESTIMATE score, as a composite index of the former two, indicates the overall proportion of non-tumor cells within the TME. These scores enabled a comprehensive characterization of the tumor microenvironment. Correlations between the ESTIMATE-related scores and the risk score were subsequently analyzed using the *psych* package (version 2.1.6), with |cor| > 0.3 and $p < 0.05$ considered statistically significant.

2.10 | Copy Number Variations (CNV) and Tumor Mutational Burden (TMB) Analyses

To characterize differences in tumor mutation load between the HRG and LRG, TMB was calculated in the TCGA-THCA training cohort based on somatic mutation data. TMB was defined as the number of somatic mutations per megabase of the effective coding region, according to the formula: TMB = *somatic mutations/L*, where L represents the effective coverage area. TMB values were calculated and visualized using the *maftools* package (tcga_load() function; version 2.18.0) [46].

In addition, CNV analysis was performed to explore CNV-related genomic alterations in PTC and to evaluate their potential impact on prognosis between the HRG and LRG in the TCGA-THCA training cohort. Significant amplified and deleted genomic regions were identified using GISTIC 2.0. The CNV landscape was subsequently visualized using the *gisticChromPlot()* function from the *maftools* package (version 2.18.0).

2.11 | Cancer Immune Cycle and Tumor Immune Dysfunction and Exclusion (TIDE) Analyses

In PTC samples with available survival information from the TCGA-THCA training cohort, differences in cancer immunity-related processes between the LRG and HRG were evaluated using the cancer immune cycle framework. The seven-step cancer immune cycle pathways were obtained from the Tracking Tumor Immunophenotype (TIP) database (http://biocc.hrbmu.edu.cn/TIP/). To further assess potential differences in immunotherapy responsiveness between the HRG and LRG, TIDE scores were calculated using the TIDE algorithm. Comparisons of cancer immune cycle pathway activities and TIDE scores between the two risk groups were conducted using the Wilcoxon rank-sum test, with $p < 0.05$ considered statistically significant. In addition, Spearman correlation analysis was performed to evaluate the association between TIDE scores and risk scores, with |correlation coefficient (cor)| > 0.3 and $p < 0.05$ defined as statistically significant. All results were visualized using the *ggplot2* package (version 3.4.4).

2.12 | Drug Sensitivity Analysis

To evaluate differences in drug sensitivity between the LRG and HRG in the TCGA-THCA training cohort, PTC-related therapeutic agents were obtained from the Genomics of Drug Sensitivity in Cancer (GDSC) database (v 12.0) (https://www.cancerrxgene.org/). The half-maximal inhibitory concentration (IC50) for each patient sample in the training cohort was predicted using the *pRRophetic* package (pRRopheticPredict() function; version 0.5) [47]. Differences in IC50 values between the LRG and HRG were assessed using the Wilcoxon rank-sum test, with $p < 0.05$ considered statistically significant. The results were visualized using the *ggplot2* package (version 3.4.4).

2.13 | Construction of the ceRNA Network

To elucidate the molecular regulatory mechanisms of prognostic genes in PTC, a ceRNA network incorporating lncRNAs, miRNAs, circRNAs, and mRNAs was constructed. Expression data for lncRNAs and miRNAs were obtained from the TCGA-THCA training cohort, and differentially expressed lncRNAs (DE-lncRNAs) and miRNAs (DE-miRNAs) were identified using the *DESeq2* package (version 1.34.0), with $|\log_2 \text{fold change}| > 1$ and $p < 0.05$ as the thresholds. Differentially expressed circRNAs (DE-circRNAs) were identified from the GSE205733 dataset using the *limma* package (version 3.50.1) [48]. MiRNAs targeting prognostic genes were predicted using the miRNet database (https://www.mirnet.ca/), and key miRNAs were determined by intersecting the predicted miRNAs with DE-miRNAs. Subsequently, lncRNAs and circRNAs interacting with the key miRNAs were retrieved from miRNet and intersected with DE-lncRNAs and DE-circRNAs to identify key lncRNAs and circRNAs. All intersections were performed using the *VennDiagram* package (version 1.7.3). The lncRNA–miRNA–mRNA and circRNA–miRNA–mRNA networks were visualized using Cytoscape (version 3.9.1). In addition, the top five regulatory axes with the highest number of predicted interactions were selected for ROC analysis using the *pROC* package (version 1.18.0) [49] to evaluate diagnostic performance, demonstrating an AUC > 0.6 considered indicative of acceptable accuracy.

2.14 | Analysis of Prognostic Gene Expression

To investigate expression differences of prognostic genes between PTC tissues and normal controls, differential expression analysis was performed using the *DESeq2* package (version 1.34.0) in the TCGA-THCA dataset, with $p < 0.05$ considered statistically significant. In addition, immunohistochemistry (IHC) staining images of proteins encoded by prognostic genes in PTC and normal thyroid tissues were retrieved from the HPA database (https://www.proteinatlas.org/) to further validate gene expression patterns at the protein level.

2.15 | The scRNA-Seq Analysis

Single-cell RNA sequencing data from the GSE184362 dataset were processed using the *Seurat* package (version 5.0.1) [50]. Quality control filtering was performed to remove cells with fewer than 200 detected genes, genes expressed in fewer than three cells, cells with fewer than 400 or more than 5000 detected features (nFeature_RNA), cells with fewer than 300 or more than 20,000 total counts (nCount_RNA), and cells with mitochondrial gene expression exceeding 5% (percent.mt). The filtered data were normalized using the *NormalizeData* function (LogNormalize method, scale factor = 10,000). Highly variable genes were identified using the *FindVariableFeatures* function and used for downstream analyses, with the top 10 most variable genes visualized using the *LabelPoints* function. Data scaling was performed using the *ScaleData* function, and statistically significant principal components (PCs) were identified using the *JackStraw* and *JackStrawPlot* functions. Principal component analysis (PCA) was conducted using the *RunPCA* function based on the selected highly variable genes, and the optimal number of PCs was determined using the *ElbowPlot* function. PCs with $p < 0.05$ were retained for subsequent analyses. Dimensionality reduction was performed using the Uniform Manifold Approximation and Projection (UMAP) algorithm, and cell clustering was conducted using the *FindClusters* function with a resolution of 0.4. Cell clusters were annotated based on known marker genes (Table S2), and marker gene expression across cell types was visualized using bubble plots. The relative proportions of different cell types in PTC samples were visualized using the *ggplot2* package (version 3.4.4).

The prognostic risk model was subsequently applied to the single-cell dataset to calculate a RiskScore for each cell. The spatial distribution of RiskScore across the single-cell landscape was visualized using FeaturePlot. Next, the average RiskScore per sample and per major cell type was computed and displayed in a heatmap. Finally, the expression and distribution of the prognostic genes at single-cell resolution were illustrated through UMAP projection and a bubble plot, which collectively helped identify key cell types. To investigate ligand–receptor interactions between key cell populations and other cell types, cell–cell communication analysis was performed using the *CellChat* package (version 1.6.1) [51]. To further explore cellular heterogeneity within key cell populations, these cells were re-clustered using the same dimensionality reduction and clustering parameters described above. Differentially distributed subclusters were identified and visualized using UMAP plots and subsequently subjected to pseudotime trajectory analysis using the *Monocle* package (version 2.26.0) [52] to infer developmental trajectories.

2.16 | Clinical Sample Validation

The expression difference of prognostic genes was verified by reverse transcription quantitative PCR (RT-qPCR) between the PTC and the control. Five paired tissue samples, including five PTC tissues and matched normal thyroid tissues, were obtained from West China Hospital. Written informed consent was obtained from all participants, and the study was approved by the institutional ethics committee (approval No. 2300068267). Firstly, the total RNA of 5 pairs of tissue samples was derived by TRizol reagent (Ambion, USA). The RNA concentrations were calculated via NanoPhotometer N50. Secondly, mRNA was reverse transcribed into cDNA utilizing a test kit (Servicebio, Wuhan, China). Finally, the

RT-qPCR was conducted. The reagents, conditions, and primers required for the experiment were in the attached table (Table S3). The expression levels of prognostic genes were calculated by $2^{-\Delta\Delta Ct}$ between control and PTC samples. The internal reference gene was GAPDH, which was employed to normalize the results calculated by GraphPad Prism (v 5.0) [53]. The expression differences between PTC and control samples were measured by *t*-test in the RT-qPCR experiment ($p < 0.05$).

2.17 | Statistical Analysis

All statistical analyses and visualizations were performed using R software (v 4.2.2). For differential expression analysis between PTC and normal tissues in the TCGA-THCA and GSE205733 cohorts, raw RNA-seq count data were processed using the DESeq2 package (v 1.34.0). In analyses focusing on overall gene expression patterns, intergroup trends, or multi-dimensional associations, gene expression levels were standardized using FPKM values. For the scRNA-seq dataset (GSE184362), data normalization was conducted with the Seurat package (v 5.0.1) using the LogNormalize method (scale. factor = 10,000), followed by scaling via the ScaleData function to regress out confounding sources of variation.

For comparisons between two groups, the two-sided Wilcoxon rank-sum test was applied for non-normally distributed data, while the two-sided independent samples *t*-test was used for data meeting normality assumptions (assessed by the Shapiro–Wilk test), with a significance level of 0.05. Comparisons among three or more groups were performed using the Kruskal–Wallis test. Survival differences between risk groups were evaluated with the two-sided log-rank test. In pathway enrichment analyses, the false discovery rate (FDR) was controlled using the Benjamini–Hochberg (BH) correction method. A BH-adjusted *p*-value < 0.05 was considered statistically significant. In figures, significance levels were denoted as follows: $^{****}p < 0.0001$; $^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.

3 | Results

3.1 | Functional Analysis of Candidate Genes

A total of 7174 DEGs were identified in PTC, including 3459 upregulated and 3715 downregulated genes. The volcano plot highlights the top ten upregulated and downregulated genes (Figure 1A), while the heatmap illustrates the expression patterns of these genes between tumor and normal tissues (Figure 1B). By intersecting the DEGs with EMT-RGs, 264 candidate genes were obtained (Figure 1C). GO analysis identified 1786 significantly enriched terms, comprising 1677 BPs, 30 CCs and 79 MFs (Table S4). The BP was mainly enriched in mesenchyme development and taxis (Figure S2A), the MF was mainly enriched in basement membrane and Golgi lumen (Figure S2B), and the MF was mainly enriched in chemokine activity and cytokine activity (Figure S2C). KEGG analysis enriched 77 significant pathways (Table S5), mainly including ECM-receptor interaction, PI3K-Akt signaling pathway, TGF-beta signaling pathway and Wnt signaling pathway (Figure S2D). They directly participated in the EMT process by regulating extracellular matrix remodeling, activation of transcription factors (such as SNAI1), and dynamics of cell adhesion. PPI network indicated that there was no interaction between the proteins encoded by the 22 candidate genes (Figure S2E).

3.2 | Prognostic Value of the Risk Model

Thirteen survival related genes were identified by univariate Cox regression analysis (Figure 1D, Figure S3A). Among multiple machine learning combinations, the CoxBoost plus SuperPC model demonstrated optimal performance (CoxBoost: C-index = 0.877, SuperPC: C-index = 0.823). Based on this model, eight prognostic genes, including TYRO3, E2F1, TNFSF15, TGFBR3, PTX3, FHL2, SNAI1, and WT1, were selected to construct the risk model (Figure 1E). Furthermore, in the TCGA-THCA-train cohort, PTC patients with a risk score > -0.0457 were classified into the HRG (185 PTC patients), and those with a risk score ≤ -0.0457 were classified into the LRG (185 PTC patients) (Figure 1F). The survival curve indicated that patients in HRG had a lower survival rate ($p = 0.00053$) (Figure 1G). The AUC values of ROC were greater than 0.8, which indicated that the model had high prediction accuracy (Figure 1H). Similarly, in the TCGA-THCA-test, the prognostic value of the model was verified (Figure S3B–D).

Analysis of clinical characteristics showed that patients older than 60 years, female patients, those with T3 or T4 stage disease, N0 stage, and M0 stage constituted the majority in both risk groups (Figure S4A,B). Comparison of risk scores across clinical subgroups revealed significant differences only between T1 and T2 stages ($p < 0.05$) (Figure S4C). Subgroup survival analyses further demonstrated significant survival differences between the high risk and low risk groups in patients aged over 60 years, in both sexes, in T3 and T4 stages, and across M stages ($p < 0.05$) (Figure S4D).

3.3 | Construction of the Nomogram

Univariate Cox regression analysis demonstrated that M stage, risk score, and age were significantly associated with overall survival in patients with PTC (Figure 2A, Figure S5A). These variables remained statistically significant in multivariate Cox regression analysis (Figure 2B, Figure S5B) and were therefore incorporated into the nomogram model (Figure 2C). Calibration curves for 1, 3, and 5 years showed good agreement between predicted and observed survival probabilities, with slopes close to 1, indicating high predictive accuracy of the nomogram (Figure 2D). Receiver operating characteristic analysis further confirmed the robust discriminative ability of the model, with area under the curve values exceeding 0.9 (Figure 2E).

3.4 | Enrichment Pathway, Immune Microenvironment, and Immunotherapy in PTC

According to GSVA analysis, there were 1822 different pathways in HRG, which included 1056 upregulated pathways and 766

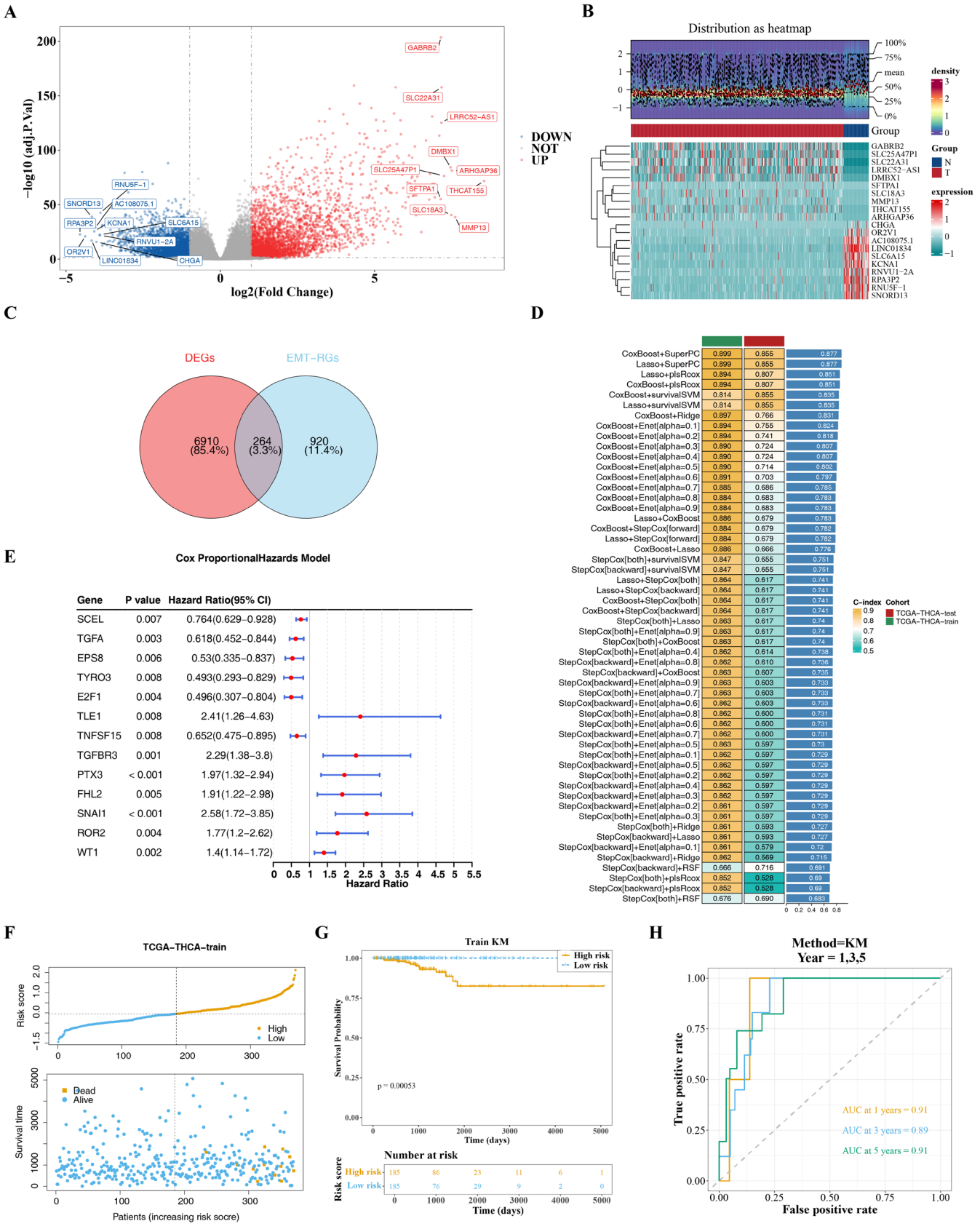

FIGURE 1 | Legend on next page.

FIGURE 1 | Acquisition of 8 prognostic genes. (A) Volcano plot of differentially expressed genes (DEGs). The y-axis represents $-\log_{10}$ (adj.*p* value), and the x-axis represents the fold change, expressed as $\log_2$FC. Red dots represent upregulated genes, blue dots represent downregulated genes, and gray dots represent genes with nonsignificant differences. The names of the top 10 upregulated and downregulated genes were labeled. (B) Heatmap of the DEGs labeled in the volcano plot. "T" represents the PTC group, and "N" represents the control group. Red indicates high expression, and green indicates low expression. (C) The intersection of DEGs and EMT-RGs was used as candidate genes. (D) Forest plot of univariate Cox regression analysis. (E) 101 machine learning algorithm combinations. The left side shows the combinations of machine learning algorithms, the middle shows the C-index values corresponding to TCGA-THCA-train and TCGA-THCA-test, and the right side shows the average C-index. (F) Risk curve and scatter plot of PTC patients in TCGA-THCA-train. Blue represents the LRG ($n=185$), yellow represents the HRG ($n=185$), and the median risk score is −0.0457. (G) KM curve of PTC patients in the LRG and HRG within TCGA-THCA-train. The x-axis represents overall survival time (days), and the y-axis represents survival probability. (H) ROC curve of the TCGA-THCA-train model. The x-axis represents specificity, and the y-axis represents sensitivity. The area enclosed by the curve and the x-axis is called the AUC. The closer the AUC value is to 1, the better the diagnostic performance.

downregulated pathways (Table S6). After adjustment for multiple testing, the top five upregulated and downregulated pathways were visualized, among which "tethering of late endosomes and lysosomes" and "BMP signaling pathway BMP antagonist" were prominently enriched (Figure 3A). These pathways may represent critical regulatory nodes involved in mesenchymal epithelial transition phenotypes and may contribute to the reversibility of invasion and metastasis in PTC. The relative infiltration levels of 22 immune cell types were estimated using the CIBERSORT algorithm (Figure 3B). Six immune cell populations, including M1 macrophages, M0 macrophages, M2 macrophages, eosinophils, neutrophils, and memory B cells, exhibited significant differences between the two risk groups (Figure 3C). However, no strong correlations were observed between individual immune cell proportions and risk scores, as none reached a Spearman correlation coefficient greater than 0.3 (Figure 3D). ESTIMATE analysis revealed that the ESTIMATE score, immune score, and stromal score were all significantly higher in the HRG than in the LRG ($p<0.01$) (Figure 3E). Correlation analysis further demonstrated that only the stromal score showed a significant positive correlation with the risk score (cor = 0.4, $p=2.2\times10^{-16}$) (Figure 3F), suggesting a predominant contribution of stromal components to risk stratification. Analysis of the cancer immune cycle revealed significant differences between the two groups in step 1, step 7, and most components of step 4 (Figure 3G). In addition, the TIDE score was significantly higher in the high risk group than in the low risk group ($p<0.0001$) and showed a strong positive correlation with the risk score (cor = 0.36, $p=5.6\times10^{-13}$) (Figure 3H), indicating a more pronounced immune dysfunction and exclusion phenotype in high risk PTC patients.

3.5 | Tumor Mutations and Drug Sensitivity in the HRG and LRG

According to TMB analysis, the BRAF (52%), NRAS (7%), and MUC16 (3%) were the most common mutations in HRG (Figure 4A). However, the BRAF (71%), NRAS (8%), and TG (6%) were the most common mutations in LRG (Figure 4B). Notably, splice-site mutations were observed exclusively in the LRG but were absent in the HRG. CNV analysis further demonstrated marked genomic differences between the two groups. The HRG showed copy number gains on chromosome 5 and copy number losses on chromosomes 2, 8, 10, 11, and 16 (Figure 4C). In contrast, the LRG exhibited copy number gains only, primarily on chromosomes 1, 4, and 22, without significant deletions (Figure 4D).

Drug sensitivity analysis identified 52 antitumor agents with significantly different predicted responses between the two groups. Among these, 37 agents showed higher sensitivity in the HRG, whereas 15 agents were more sensitive in the LRG (Figure 4E). The top five agents with the most significant differences were further illustrated. Except for BIRB.0796, which demonstrated greater sensitivity in the LRG, the remaining four agents (ABT.263, AG.014699, BX.795, and DMOG) were more sensitive in the HRG (Figure 4F).

3.6 | Diagnostic Value of Regulatory Molecules for PTC

To further explore the regulatory mechanisms underlying prognostic gene expression in PTC, a competing endogenous RNA network was constructed. Differential expression analysis identified 2564 differentially expressed lncRNAs, 64 differentially expressed miRNAs, and 185 differentially expressed circRNAs (Figure S6A). Through database-based intersection analysis, 12 key lncRNAs, 10 key miRNAs, and 39 key circRNAs were selected for downstream investigation (Figure S6B). Based on these regulatory relationships, lncRNA–miRNA–mRNA networks (e.g., *LIFR-AS1–hsa-miR-342–E2F1*) and circRNA–miRNA–mRNA networks (e.g., *ZEB1–hsa-miR-342–E2F1*) were constructed (Figure S6C,D). ROC analysis of the top five regulatory axes demonstrated that both the prognostic genes and key circRNAs exhibited favorable diagnostic performance for PTC (Figure S6E).

3.7 | Expression of Prognostic Genes in PTC-Related Cells

ScRNA-seq data analysis yielded 23,050 genes and 63,128 cells after quality control (Figure S7A,B). The top 2000 highly variable genes were identified, and the ten most variable genes were visualized (Figure S7C). Principal component analysis showed that the first 50 PCs were statistically significant, while the elbow plot indicated stabilization after 30 PCs; therefore, 30 PCs were selected for downstream analyses (Figure S7D). Based on these PCs, 25 distinct cell clusters were identified in the GSE184362 dataset (Figure S7E).

Cell clusters were annotated using canonical marker genes, resulting in the identification of six major PTC-associated cell types: B cells, myeloid cells, endothelial cells, fibroblasts, T and NK cells, and thyrocytes (Figure 5A). The expression patterns of marker

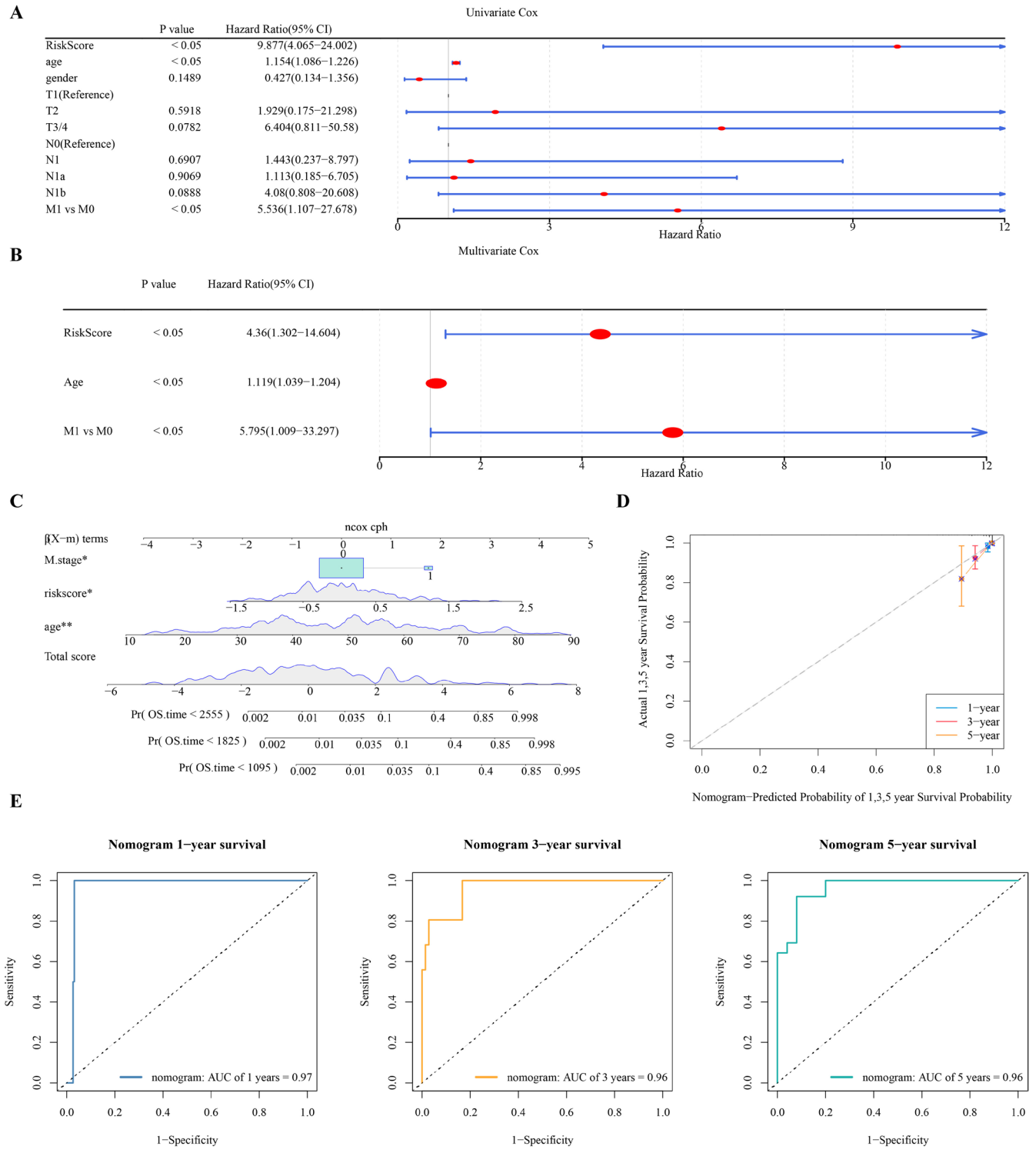

FIGURE 2 | Development and validation of a nomogram for predicting PTC prognosis. (A) Forest plot of univariate Cox regression analysis for risk scores and clinical features. (B) Forest plot of multivariate Cox regression analysis. (C) Nomogram model constructed from independent prognostic factors, predicting the 1, 3, and 5 years mortality probabilities of PTC patients. (D) Calibration curve of the nomogram. The x-axis represents the predicted event rate, and the y-axis represents the observed actual event rate. The dashed line indicates the perfect calibration state, and the colored lines represent the prediction accuracy for 1, 3, and 5 years. (E) ROC curves of the nomogram for 1, 3, and 5 years.

genes across clusters and the relative proportions of each cell type are shown in Figure S7F,G. Comparative analysis revealed that most prognostic genes exhibited higher expression levels in fibroblasts than in other cell types (Figure 5B, Figure S7H). Furthermore, risk scores exhibited marked heterogeneity across different cell types, with endothelial cells and fibroblasts displaying significantly higher average risk scores compared to other cell populations (Figure 5C). These findings suggest that fibroblasts may play a central role in PTC progression and were therefore defined as the key cell type for subsequent analyses.

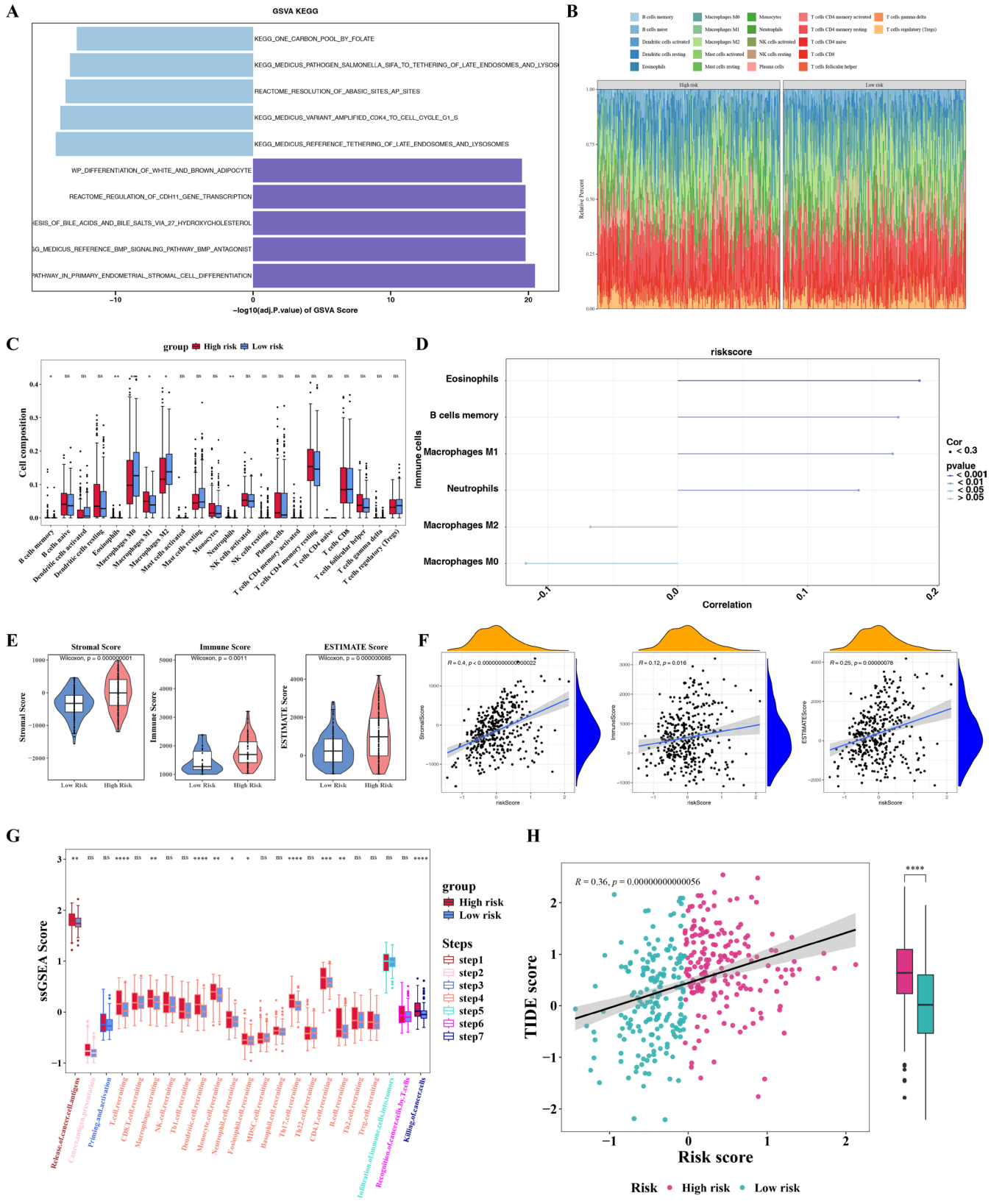

FIGURE 3 | Pathway analysis and immune cell infiltration with risk score correlation in PTC. (A) Top 10 results of GSVA. Blue represents activation in the LRG, and purple represents activation in the HRG. (B) Relative infiltration abundance of immune cells in PTC patients of the LRG and HRG. (C) Differences in the relative infiltration abundance of immune cells between the LRG and HRG. (D) Correlation between risk scores and differentially infiltrated immune cells. The x-axis represents the correlation coefficient. The darker the color of the line, the smaller the *p*-value; the larger the circle at the end of the line, the higher the correlation. (E) Differences in stromal score, immune score, and ESTIMATE score between the LRG and HRG. (F) Correlation between stromal/immune/ESTIMATE scores and risk score. (G) Differences in ssGSEA scores of immune phenotypes/processes between the LRG and HRG. (H) Correlation between TIDE score and risk score.

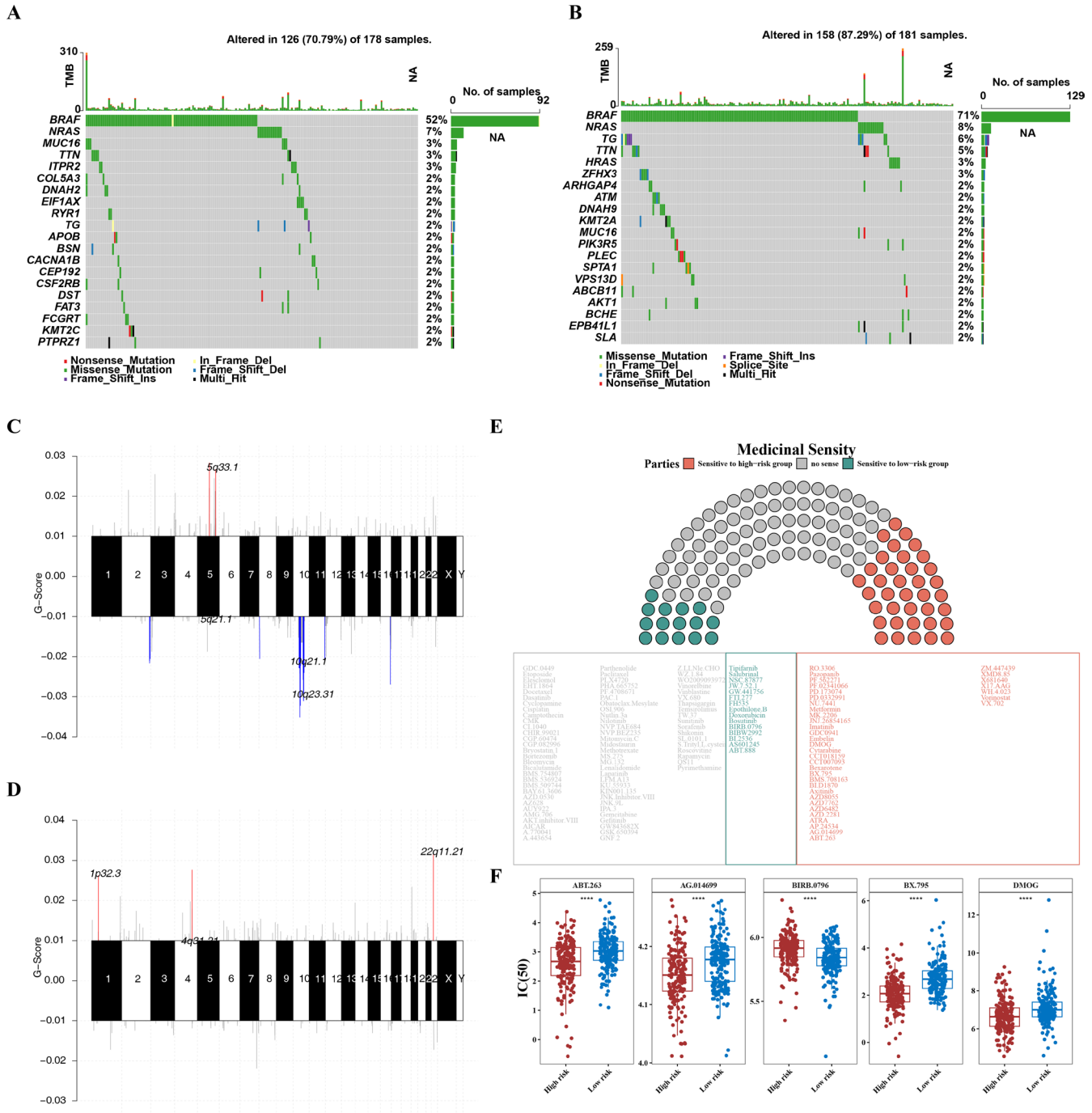

FIGURE 4 | Integrated analysis of genomic profiles and drug sensitivity between HRG and LRG in PTC. (A) The top 20 genes with the highest mutation frequencies in the HRG. (B) The top 20 genes with the highest mutation frequencies in the LRG. (C) CNV distribution map of the HRG. Red regions indicate the frequency of copy number gains (amplifications), and blue regions indicate the frequency of copy number losses (deletions). (D) CNV distribution map of the LRG. (E) Differences in drug sensitivity between patients in the LRG and HRG. Green represents drugs more sensitive to the LRG, red represents drugs more sensitive to the HRG, and gray represents no significant difference. (F) The top 5 most significant drugs.

Cell–cell communication analysis indicated that myeloid cells exhibited the strongest interactions with fibroblasts and thyrocytes, whereas interactions between B cells and thyrocytes were the most prominent overall (Figure 5D). Besides, B cells and thyrocytes interacted through MIF—(CD74 + CXCR4) (Figure 5E). Fibroblasts were further subdivided into 11 distinct subclusters (Figure 5F). Pseudotime trajectory analysis revealed a progressive differentiation process from early to late stages (Figure 5G). Early-stage fibroblasts (stage 1) were mainly composed of clusters 0 and 1, intermediate stages (stages 2 and 3) were dominated by clusters 3 and 4, and late stages (stages 4 and 5) were primarily represented by clusters 2 and 9, while clusters 5 and 7 were present across all stages. Gene expression dynamics showed that FHL2 was predominantly active during early differentiation, whereas PTX3 and TGFBR3 were upregulated at later stages, while the remaining prognostic genes displayed relatively stable expression throughout fibroblast differentiation (Figure 5H).

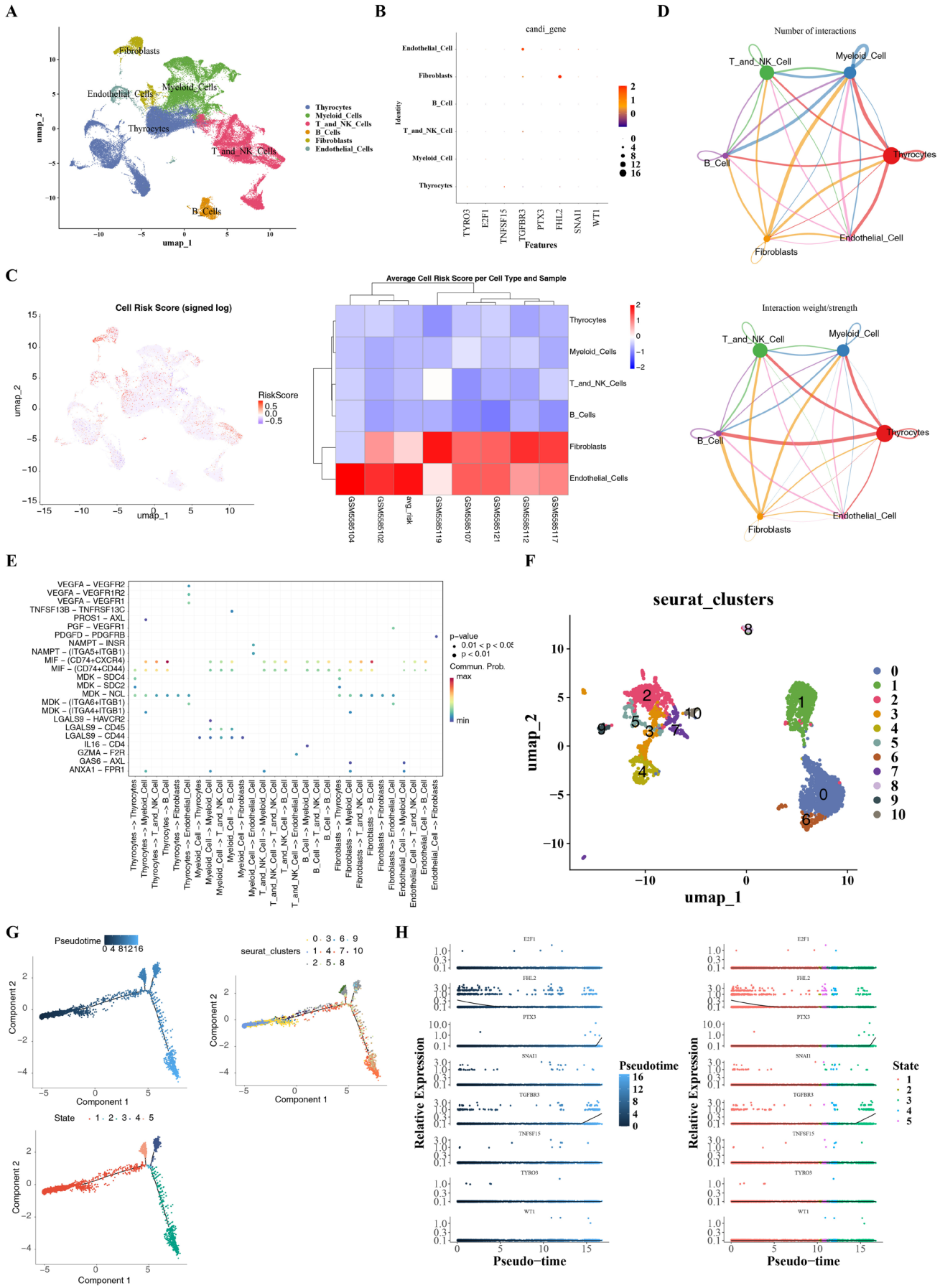

FIGURE 5 | Legend on next page.

FIGURE 5 | Single-cell RNA sequencing analysis. (A) Annotation of 25 cell clusters into six cell types. (B) Expression levels of prognostic genes in different cell types. Larger and redder circles indicate higher expression levels. (C) Risk scores across different cells, with red indicating higher risk and blue indicating lower risk. (D) Number/strength of cell communications. Larger circles indicate a greater number/strength of cells; thicker lines indicate a higher number/strength of interactions. (E) Diagram of cell communication ligand-receptor pairs. The x-axis represents detailed types and directions of cell interactions; the y-axis represents ligand-receptor pairs. A redder color indicates a higher interaction probability. (F) UMAP plot of secondary clustering of fibroblasts. (G) Differentiation trajectory of fibroblasts. (H) Expression of prognostic genes in the differentiation trajectory of fibroblasts.

3.8 | Validation of Prognostic Gene Expression

Bulk expression analysis demonstrated that E2F1, TNFSF15, TYRO3, and WT1 were upregulated in PTC samples, whereas FHL2, PTX3, SNAI1, and TGFBR3 showed higher expression in normal control tissues (Figure 6A). Protein-level expression patterns of the prognostic genes were further validated using immunohistochemistry data (Figure 6B). RT-qPCR analysis of clinical samples largely corroborated the bioinformatic findings. Significant expression differences were observed for most prognostic genes between PTC and control tissues ($p < 0.05$), with the exception of TNFSF15 and TYRO3, for which the differences did not reach statistical significance (Figure 6C).

4 | Discussion

In this study, we performed an integrated bioinformatics analysis to investigate the prognostic role of EMT-related genes in PTC. Initially, differential expression analysis between PTC and normal tissues identified 7174 DEGs. Intersection with curated EMT-related gene sets yielded 264 candidate genes, which were significantly enriched in pathways associated with cancer progression and EMT. Prognostic screening using univariate Cox regression, followed by machine learning–based feature selection, identified an eight-gene signature that was used to construct a risk stratification model. This model effectively separated patients into distinct risk groups, with higher risk scores significantly associated with poorer overall survival. Multivariate Cox analysis further demonstrated that the risk score, together with age and M stage, served as independent prognostic factors, enabling the construction of a clinically applicable nomogram for survival prediction.

Comprehensive characterization of the risk groups revealed marked differences in the tumor immune microenvironment, including the infiltration of six immune cell types, stromal/immune scores (ESTIMATE)-derived stromal and immune scores, and TIDE scores. Differences in tumor mutational burden and somatic mutation landscapes were also observed between the two groups. Single-cell RNA sequencing analysis identified fibroblasts as key contributors to PTC progression, and pseudotime trajectory analysis highlighted dynamic expression of three prognostic genes—*FHL2*, *PTX3*, and *TGFBR3*—during fibroblast differentiation. RT-qPCR validation using clinical samples confirmed the expression patterns of most prognostic genes, consistent with bioinformatic predictions, although *TNFSF15* and *TYRO3* did not reach statistical significance. Protein-level evidence from public databases further supported these transcriptomic findings.

Although PTC is generally indolent, a subset of patients develops aggressive disease characterized by radioiodine resistance and metastatic behavior, phenotypes that are increasingly associated with EMT [10]. Conventional prognostic systems, such as TNM staging, fail to capture the molecular heterogeneity underlying EMT-driven progression, highlighting the need for biomarkers that reflect TME dynamics. In this study, we integrated single-cell RNA sequencing with 101 machine learning algorithm combinations to identify eight EMT-related genes (*TYRO3, E2F1, TNFSF15, TGFBR3, PTX3, FHL2, SNAI1*, and *WT1*) that stratify patients into high- and low-risk groups with distinct clinical outcomes. These genes are collectively associated with key processes in EMT—including transcriptional regulation, stromal remodeling, and immune modulation—and may thus offer a multifaceted perspective on PTC progression. Among these genes, *SNAI1* plays a central role as a canonical EMT transcription factor. *SNAI1* encodes a zinc finger DNA-binding protein that promotes EMT by repressing E-cadherin expression, thereby facilitating tumor invasion and dissemination. Aberrant *SNAI1* expression has been implicated in multiple pathological conditions, including renal cell carcinoma, triple-negative breast cancer, and hepatic fibrosis [54–56]. Consistent with previous reports in PTC [57] and other malignancies such as breast cancer [58], our results demonstrated that *SNAI1* overexpression was significantly associated with lymph node metastasis (HR = 2.1, $p = 0.008$). Notably, emerging evidence indicates that enhanced autophagy can inhibit EMT and tumor metastasis by promoting degradation of nuclear *SNAI1*, suggesting a potential therapeutic vulnerability [59].

The *WT1* gene encodes a transcription factor essential for the regulation of cell proliferation, differentiation, and apoptosis. In pathological contexts, chromosomal rearrangements may generate *WT1* fusion proteins, leading to loss of tumor-suppressive function and acquisition of oncogenic properties [60]. Such alterations promote malignant progression through dysregulation of downstream targets, including *IGF2* and *BCL2* [61]. In our study, *WT1*, together with *SNAI1*, emerged as a novel EMT activator in PTC-associated fibroblasts, implicating stromal reprogramming and extracellular matrix deposition as key drivers of tumor aggressiveness [62, 63].

This stromal interplay is further reinforced by *TGFBR3*, a transmembrane proteoglycan that functions as a co-receptor for TGF-β superfamily signaling by facilitating ligand presentation to *TGFBR1* and *TGFBR2*. Frequent downregulation of *TGFBR3* across multiple malignancies has been associated with enhanced tumor invasion and metastasis [64, 65], supporting its tumor-suppressive role through inhibition of SMAD-independent pathways such as MAPK/ERK signaling [66]. In the present study, *TGFBR3* was enriched in late-stage fibroblasts, where it may

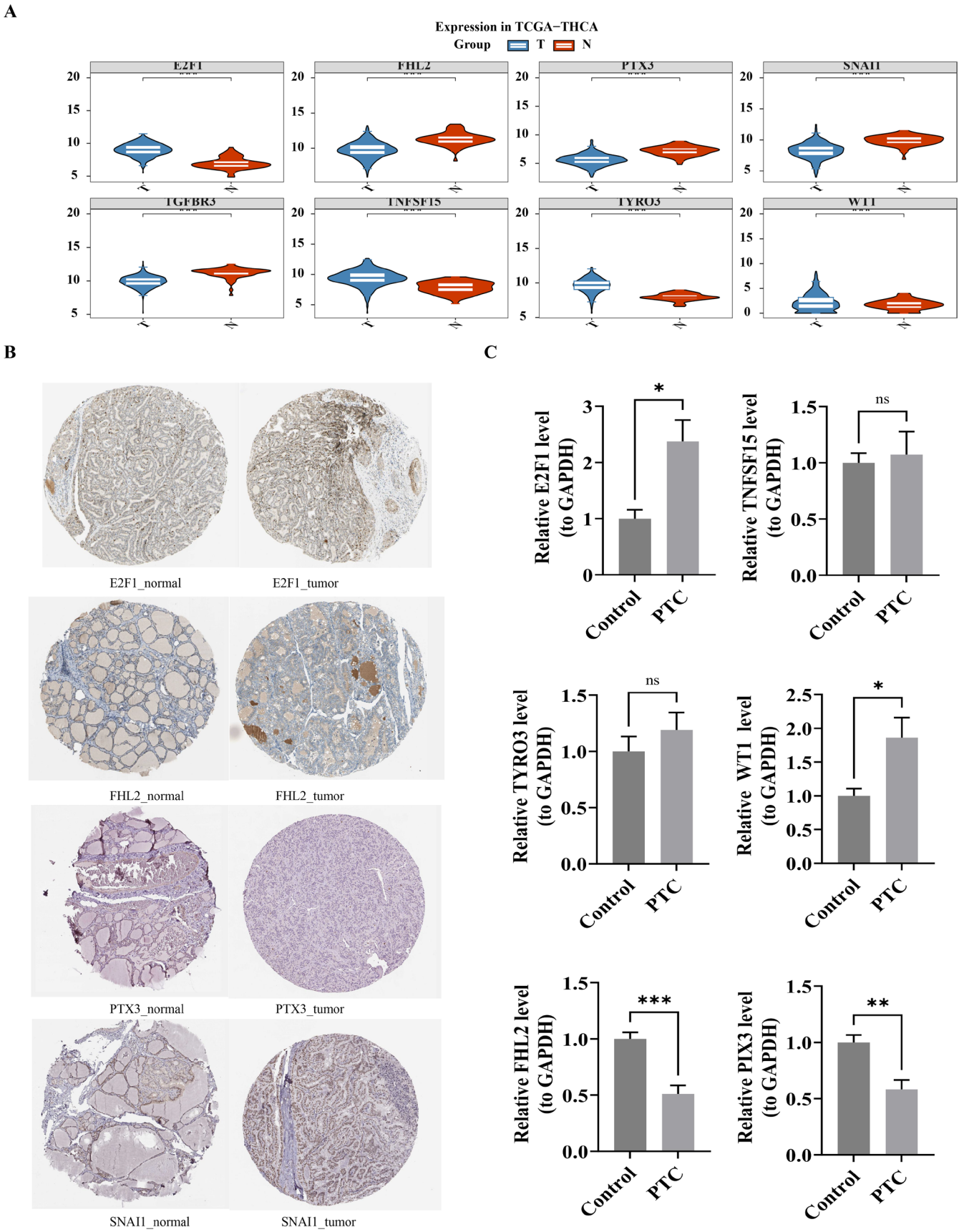

FIGURE 6 | Expression comparison and validation of prognostic genes in PTC patients and control group. (A) Differential expression of mRNA levels of prognostic genes between the PTC group ($n = 505$) and the control group ($n = 59$) in the TCGA-THCA dataset. (B) Differential expression of protein levels of prognostic genes between the PTC group and the control group in the HPA database. (C) Differential expression of mRNA levels of prognostic genes between the PTC group ($n = 5$) and the control group ($n = 5$) in clinical samples. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

contribute to fibroblast activation and the formation of an immunosuppressive stromal niche [67]. Consistent with our findings, *TGFBR3*-mediated promotion of stromal fibrosis has also been reported in oral cancer, where it drives collagen deposition and immune exclusion [68].

FHL2 is a multifunctional scaffold protein belonging to the LIM domain family. Through its four LIM domains, *FHL2* dynamically interacts with oncogenic and tumor-suppressive proteins to regulate transcriptional activity, cytoskeletal remodeling, and integrin-mediated signaling [69]. Previous studies have demonstrated that *FHL2* activates the Wnt/β-catenin pathway and mediates TGF-β1–induced epithelial–mesenchymal transdifferentiation in renal tubular epithelial cells [70]. Notably, *FHL2* exhibits context-dependent functional duality: it suppresses proliferation and invasion in breast cancer by inhibiting inhibitor of differentiation protein 3 [71], while in hepatocellular carcinoma, *FHL2–p53* interactions promote apoptosis via upregulation of *PARP12* [72]. In our study, *FHL2* expression peaked during early fibroblast differentiation in PTC and was associated with enhanced migratory capacity through integrin signaling, paralleling its pro-metastatic role in lung adenocarcinoma [73]. Moreover, KEGG pathway analysis revealed significant enrichment of the Wnt signaling pathway, suggesting a similar contribution of this axis to EMT regulation in PTC.

PTX3 is a soluble pattern-recognition receptor that plays a critical role in innate immunity by binding exogenous ligands, activating the complement cascade, and promoting effector mechanisms such as phagocytosis [74]. Beyond host defense, *PTX3* modulates tumor immune surveillance through interactions with extracellular matrix components and regulation of macrophage recruitment [75]. In this study, *PTX3* expression predominated during later stages of fibroblast differentiation and was positively correlated with M2 macrophage infiltration in PTC, indicating a role in establishing an immunosuppressive microenvironment. This observation is consistent with findings in colorectal cancer, where *PTX3* facilitates recruitment of tumor-associated macrophages (TAMS) and promotes immune evasion [76].

The interplay between EMT and immune evasion in PTC is partially mediated by *TYRO3* and *TNFSF15*. *TYRO3*, a member of the TAM family of transmembrane receptor tyrosine kinases, is widely expressed, with notable activity in ovarian, neural, and immune tissues, where it regulates immune tolerance through modulation of dendritic cell maturation and promotes tumor progression by driving EMT [77]. In this study, *TYRO3* expression was significantly associated with increased PD-L1 levels in high-risk PTC patients ($p < 0.001$), suggesting its potential as a therapeutic target for combination immunotherapy [78]. Consistent with this finding, inhibition of *TYRO3* has been shown to synergize with anti–PD-1 therapy in non–small cell lung cancer, resulting in enhanced T-cell infiltration and antitumor immunity [79].

TNFSF15 is a cytokine of the tumor necrosis factor ligand family characterized by a conserved TNF homology domain and is involved in the regulation of angiogenesis, inflammation, and apoptosis through activation of the NF-κB and MAPK pathways. Although *TNFSF15* suppresses pathological angiogenesis by competitively binding VEGFR2 and inhibiting endothelial cell proliferation [80], our results suggest that it may paradoxically sustain EMT in PTC via NF-κB activation. This functional duality may account for its borderline validation significance in clinical samples ($p = 0.06$). Similar context-dependent protumorigenic effects of *TNFSF15* have been reported in renal cell carcinoma, where it promotes tumor survival through NF-κB–mediated cytokine production [81].

Based on these findings, we developed a prognostic risk model incorporating eight EMT-related genes. The risk score was calculated using a CoxBoost + SuperPC algorithm, achieving a concordance index of 0.877. Patients were stratified into HRG and LRG groups according to the median risk score. The model demonstrated robust prognostic performance in both the TCGA-THCA training and validation cohorts, with HRG patients exhibiting significantly poorer overall survival than those in the LRG. Integration of the risk score with age and M stage enabled construction of a nomogram with improved predictive accuracy.

From a clinical perspective, expression levels of the prognostic genes can be readily assessed by RT-qPCR using routinely processed tumor specimens, including postoperative paraffin-embedded tissues or biopsy samples, with GAPDH as an internal reference. Combined with patient age and M stage, the nomogram allows estimation of 5-year survival probability in PTC. Using the optimal cutoff value (−0.0457), patients can be stratified into risk categories: HRG patients may benefit from intensified postoperative surveillance and consideration of adjuvant targeted therapies (e.g., *SNAI1* inhibition), whereas LRG patients may continue standard follow-up to avoid overtreatment. Nonetheless, multicenter prospective studies are required to validate the model and establish population-specific cutoff values.

Our risk stratification model revealed pronounced heterogeneity in the immune landscape between high risk and low risk PTC patients. Six immune cell populations, including memory B cells, eosinophils, M0, M1, and M2 macrophages, as well as neutrophils, exhibited significantly differential infiltration patterns. Elevated TIDE scores in the high-risk group indicate a profoundly immunosuppressive TME, characterized by T-cell dysfunction and enrichment of myeloid-derived suppressor cells (MDSCs), a pattern consistent with EMT-high lung adenocarcinoma [82]. Mechanistically, this immunosuppressive shift is supported by M2 macrophage accumulation and concomitant M1 depletion, resulting in an IL-10/TGF-β–dominant milieu that promotes MDSC expansion. Notably, eosinophils displayed paradoxical protumorigenic reprogramming via activation of the IL-33/ST2–MMP9 axis.

The strong positive correlation between stromal scores and risk scores further implicates fibroblast-driven extracellular matrix remodeling mediated by the FHL2, PTX3, and TGFBR3 axis as a key determinant of immune suppression. These observations are supported by single cell trajectory analyses, which revealed dynamic crosstalk between fibroblasts and neutrophils mediated through IL 8 and CXCR2 signaling. This interaction may act synergistically with MDSCs to promote the establishment of metastatic niches. From a clinical perspective, the dual

association of the risk score with stromal activation and T cell exhaustion underscores its potential utility as a biomarker for combination therapeutic strategies targeting both fibroblast populations and myeloid-derived suppressor cells. However, the lack of spatial resolution in the current dataset limits definitive localization of MDSCs, highlighting the need for validation using multiplex immunohistochemistry or spatial transcriptomic approaches.

Single-cell transcriptomic analysis further delineated the cellular dynamics underlying EMT at unprecedented resolution. Among 63,128 analyzed cells, fibroblasts emerged as central regulators of stromal–tumor crosstalk, exhibiting distinct differentiation trajectories marked by dynamic expression of FHL2 and TGFBR3. Early-stage fibroblasts (Stage 1) demonstrated elevated FHL2 expression, coinciding with migratory activation, whereas late-stage fibroblasts (Stage 3) upregulated TGFBR3 and PTX3, thereby promoting ECM remodeling and M2 macrophage recruitment—a mechanism previously described in pancreatic ductal adenocarcinoma [83].

Fibroblasts thus orchestrate PTC progression through temporally and spatially coordinated mechanisms. During tumor initiation, FHL2high fibroblasts prime the premalignant niche by secreting matrix metalloproteinases and hepatocyte growth factor, disrupting basement membrane integrity and stimulating thyrocyte proliferation. Concurrent Ephrin–Eph signaling between fibroblasts and epithelial cells induces epigenetic reprogramming that facilitates malignant transformation [84]. As tumors advance, fibroblasts adopt a TGFBR3high/PTX3high immunosuppressive phenotype: LOXL2-mediated ECM stiffening activates YAP/TAZ signaling in tumor cells to sustain EMT [85], while PTX3 and TGFBR3 promote recruitment of M2 macrophages and neutrophils, respectively, thereby reinforcing immune evasion [83]. During metastatic dissemination, fibroblasts contribute to the formation of premetastatic niches through exosomal transfer of SNAI1 mRNA and systemic secretion of fibronectin and versican, which recruit myeloid cells via the MIF–CD74/CXCR4 axis [86]. These mechanisms parallel observations in hepatocellular carcinoma [87], underscoring a conserved role of stromal fibroblasts in driving EMT-associated immune escape across malignancies.

The stage specific functions of fibroblasts provide opportunities for precision therapeutic intervention. In early stage tumors, inhibition of the FHL2 integrin beta 1 axis may effectively suppress fibroblast mediated migratory priming [88]. In contrast, advanced disease is more likely to require combined targeting of PTX3 and TGFBR3 together with CCR2 blockade to disrupt immune suppressive signaling networks and myeloid cell recruitment [89]. In addition, nanoparticle mediated delivery of small interfering RNA targeting fibroblast derived exosomes represents a promising strategy to impair pre metastatic niche formation and long range stromal communication [90]. Current limitations in spatial resolution highlight the need for advanced technologies such as Visium spatial transcriptomics to precisely map fibroblast thyrocyte interactions at invasive tumor fronts. Future clinical trials should incorporate stratification based on fibroblast activation signatures to validate stage adapted therapeutic strategies, thereby positioning fibroblast reprogramming as a central component of precision oncology for PTC [91].

The prognostic capacity of the identified gene set was strengthened through the integration of one hundred and one machine learning algorithm combinations, a strategy specifically designed to mitigate bias associated with single model approaches. By cross validating LASSO, XGBoost, and CoxBoost frameworks, we prioritized genes that demonstrated stable and consistent predictive performance across diverse computational paradigms, resulting in superior model accuracy with an area under the curve exceeding 0.9 when compared with existing prognostic tools. Similar multi algorithm strategies have been successfully applied in glioblastoma to derive robust and reproducible prognostic signatures [92]. Furthermore, the nomogram incorporating the risk score together with patient age and M stage provided enhanced prognostic resolution, particularly for patients with intermediate TNM stages who are inadequately stratified by conventional staging systems. Notably, individuals classified as intermediate TNM stage but assigned to the high risk group exhibited a 3.5-fold increase in mortality risk, underscoring the clinical relevance of the proposed model.

In summary, this study establishes a novel prognostic framework for PTC through the integration of single cell transcriptomic profiling and machine learning based modeling. Three major advances were achieved. First, we systematically identified eight EMT-related hub genes including TYRO3, E2F1, TNFSF15, TGFBR3, PTX3, FHL2, SNAI1, and WT1 that cooperatively drive tumor progression through transcriptional reprogramming, stromal remodeling, and immune modulation. Second, incorporation of one hundred and one machine learning algorithm combinations markedly improved risk stratification performance, achieving a five year survival prediction area under the curve of 0.91 and outperforming the traditional TNM staging system. Most importantly, we delineated fibroblasts as central regulators of the PTC microenvironment, providing a mechanistic rationale for fibroblast targeted interventions such as FHL2 and TGFBR3 inhibition as well as risk adapted immunotherapeutic strategies.

Despite the integrative approach, several limitations should be considered. First, the findings are primarily based on retrospective cohorts (e.g., TCGA), which may introduce selection bias. Although validation in an independent dataset and partial confirmation by RT-qPCR using clinical samples help mitigate this concern, further prospective validation is warranted. Second, while scRNA-seq delineated cellular heterogeneity, it lacks spatial context. The dynamic crosstalk between tumor cells and fibroblasts, particularly at the invasive front, remains to be spatially resolved. Third, due to constraints in time and resources, we could not perform functional experiments (e.g., gene perturbation) to establish causality or comprehensively validate protein-level discrepancies (e.g., TNFSF15/TYRO3).

To address these gaps, future work will include: (1) expanding clinical sample sizes for robust validation of the gene signature; (2) applying spatial transcriptomics (e.g., Visium) to map EMT hotspots and cellular interactions in situ; (3) employing proteomic profiling to complement transcriptomic findings; and (4) establishing in vitro and in vivo models using CRISPR-mediated

editing to functionally interrogate key genes (e.g., via dual-luciferase and Co-IP assays). These steps will strengthen the mechanistic foundation and accelerate the translation of our findings toward precision strategies for high-risk PTC.

5 | Conclusion

This study integrates single-cell RNA sequencing with 101 machine-learning algorithm combinations to identify eight EMT–related prognostic genes in PTC and to construct a robust risk stratification model. The model effectively separates patients into distinct survival groups. High-risk patients are characterized by an immunosuppressive tumor microenvironment and fibroblast-driven EMT dynamics, which were further validated experimentally. This integrative framework provides actionable insights for precision-targeted therapies and highlights the need for future multicenter validation to support clinical translation.

Author Contributions

Xun Zheng: conceptualization, methodology, writing – review and editing, formal analysis, funding acquisition, writing – original draft. **Tianfeng Xu and Ruonan Sun:** investigation, writing – original draft, formal analysis. **Yujie Zhang:** data curation, methodology, software.

Acknowledgements

The authors are indebted to the patients who participated in this study by providing clinical samples.

Funding

This work was supported by the "Qimingxing" Research Fund for Young Talents of West China Hospital (No. HXQMX0105).

Ethics Statement

The informed consent form was signed and filled out by all participants, while the ethical approval agency was No. 2300068267 (West China Hospital, Sichuan University).

Consent

The authors have nothing to report.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. C. Georgiades, "Slow Growth, Excellent Prognosis: The Treatment and Overtreatment of Papillary Thyroid Cancer," *Radiology* 311, no. 1 (2024): e240207.

2. N. Arroyo, K. J. L. Bell, V. Hsiao, et al., "Prevalence of Subclinical Papillary Thyroid Cancer by Age: Meta-Analysis of Autopsy Studies," *Journal of Clinical Endocrinology and Metabolism* 107, no. 10 (2022): 2945–2952.

3. M. L. Iglesias, A. Schmidt, A. A. Ghuzlan, et al., "Radiation Exposure and Thyroid Cancer: A Review," *Archives of Endocrinology and Metabolism* 61, no. 2 (2017): 180–187.

4. N. Qu, D. Chen, B. Ma, et al., "Integrated Proteogenomic and Metabolomic Characterization of Papillary Thyroid Cancer With Different Recurrence Risks," *Nature Communications* 15, no. 1 (2024): 3175.

5. J. M. Oh and B. C. Ahn, "Molecular Mechanisms of Radioactive Iodine Refractoriness in Differentiated Thyroid Cancer: Impaired Sodium Iodide Symporter (NIS) Expression Owing to Altered Signaling Pathway Activity and Intracellular Localization of NIS," *Theranostics* 11, no. 13 (2021): 6251–6277.

6. M. Taki, K. Abiko, M. Ukita, et al., "Tumor Immune Microenvironment During Epithelial-Mesenchymal Transition," *Clinical Cancer Research* 27, no. 17 (2021): 4669–4679.

7. R. Zhao and P. A. Trainor, "Epithelial to Mesenchymal Transition During Mammalian Neural Crest Cell Delamination," *Seminars in Cell & Developmental Biology* 138 (2023): 54–67.

8. J. Sarrand and M. S. Soyfoo, "Involvement of Epithelial-Mesenchymal Transition (EMT) in Autoimmune Diseases," *International Journal of Molecular Sciences* 24, no. 19 (2023): 14481.

9. N. Zhang, A. S. Ng, S. Cai, Q. Li, L. Yang, and D. Kerr, "Novel Therapeutic Strategies: Targeting Epithelial-Mesenchymal Transition in Colorectal Cancer," *Lancet Oncology* 22, no. 8 (2021): e358–e368.

10. W. Pu, X. Shi, P. Yu, et al., "Single-Cell Transcriptomic Analysis of the Tumor Ecosystems Underlying Initiation and Progression of Papillary Thyroid Carcinoma," *Nature Communications* 12, no. 1 (2021): 6058.

11. H. E, L. Zhang, Z. Yang, et al., "SNAI1 Promotes Epithelial-Mesenchymal Transition and Maintains Cancer Stem Cell-Like Properties in Thymic Epithelial Tumors Through the PIK3R2/p-EphA2 Axis," *Journal of Experimental & Clinical Cancer Research* 43, no. 1 (2024): 324.

12. T. Guan, M. Li, Y. Song, et al., "Phosphorylation of USP29 by CDK1 Governs TWIST1 Stability and Oncogenic Functions," *Advanced Science (Weinheim, Baden-Württemberg, Germany)* 10, no. 11 (2023): e2205873.

13. B. Zhang, K. Zeng, R. C. Guan, et al., "Single-Cell RNA-Seq Analysis Reveals Macrophages Are Involved in the Pathogenesis of Human Sporadic Acute Type A Aortic Dissection," *Biomolecules* 13, no. 2 (2023): 399.

14. H. Tan, W. Wang, C. Zhou, et al., "Single-Cell RNA-Seq Uncovers Dynamic Processes Orchestrated by RNA-Binding Protein DDX43 in Chromatin Remodeling During Spermiogenesis," *Nature Communications* 14, no. 1 (2023): 2499.

15. T. Wang, J. Shi, L. Li, et al., "Single-Cell Transcriptome Analysis Reveals Inter-Tumor Heterogeneity in Bilateral Papillary Thyroid Carcinoma," *Frontiers in Immunology* 13 (2022): 840811.

16. J. Ning, X. Hou, J. Hao, et al., "METTL3 Inhibition Induced by M2 Macrophage-Derived Extracellular Vesicles Drives Anti-PD-1 Therapy Resistance via M6A-CD70-Mediated Immune Suppression in Thyroid Cancer," *Cell Death and Differentiation* 30, no. 10 (2023): 2265–2279.

17. G. J. Xu, M. A. Loberg, J. N. Gallant, et al., "Molecular Signature Incorporating the Immune Microenvironment Enhances Thyroid Cancer Outcome Prediction," *Cell Genomics* 3, no. 10 (2023): 100409.

18. M. Schlumberger and S. Leboulleux, "Current Practice in Patients With Differentiated Thyroid Cancer," *Nature Reviews. Endocrinology* 17, no. 3 (2021): 176–188.

19. I. S. Forrest, B. O. Petrazzini, A. Duffy, et al., "Machine Learning-Based Marker for Coronary Artery Disease: Derivation and Validation in Two Longitudinal Cohorts," *Lancet* 401, no. 10372 (2023): 215–225.

20. N. Zhang, H. Zhang, W. Wu, et al., "Machine Learning-Based Identification of Tumor-Infiltrating Immune Cell-Associated lncRNAs for Improving Outcomes and Immunotherapy Responses in Patients With Low-Grade Glioma," *Theranostics* 12, no. 13 (2022): 5931–5948.

21. M. I. Love, W. Huber, and S. Anders, "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data With DESeq2," *Genome Biology* 15, no. 12 (2014): 550.

22. E. K. Gustavsson, D. Zhang, R. H. Reynolds, S. Garcia-Ruiz, and M. Ryten, "Ggtranscript: An R Package for the Visualization and Interpretation of Transcript Isoforms Using ggplot2," *Bioinformatics* 38, no. 15 (2022): 3844–3846.

23. Z. Gu, R. Eils, and M. Schlesner, "Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data," *Bioinformatics* 32, no. 18 (2016): 2847–2849.

24. C. H. Gao, G. Yu, and P. Cai, "ggVennDiagram: An Intuitive, Easy-to-Use, and Highly Customizable R Package to Generate Venn Diagram," *Frontiers in Genetics* 12 (2021): 706907.

25. T. Wu, E. Hu, S. Xu, et al., "clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data," *Innovation (Cambridge (Mass.))* 2, no. 3 (2021): 100141.

26. W. Walter, F. Sanchez-Cabo, and M. Ricote, "GOplot: An R Package for Visually Combining Expression Data With Functional Analysis," *Bioinformatics* 31, no. 17 (2015): 2912–2914.

27. P. Shannon, A. Markiel, O. Ozier, et al., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research* 13, no. 11 (2003): 2498–2504.

28. J. Lei, T. Qu, L. Cha, et al., "Clinicopathological Characteristics of Pheochromocytoma/Paraganglioma and Screening of Prognostic Markers," *Journal of Surgical Oncology* 128, no. 4 (2023): 510–518.

29. H. Jiang, Y. Li, and T. Wang, "Comparison of Billroth I, Billroth II, and Roux-En-Y Reconstructions Following Distal Gastrectomy: A Systematic Review and Network Meta-Analysis," *Cirugia Española* 99, no. 6 (2021): 412–420.

30. P. Zhao, H. Zhen, H. Zhao, Y. Huang, and B. Cao, "Identification of Hub Genes and Potential Molecular Mechanisms Related to Radiotherapy Sensitivity in Rectal Cancer Based on Multiple Datasets," *Journal of Translational Medicine* 21, no. 1 (2023): 176.

31. Y. Liu, Y. Zhao, S. Zhang, et al., "Developing a Prognosis and Chemotherapy Evaluating Model for Colon Adenocarcinoma Based on Mitotic Catastrophe-Related Genes," *Scientific Reports* 14, no. 1 (2024): 1655.

32. Z. Wang, C. Ma, Q. Teng, et al., "Identification of a Ferroptosis-Related Gene Signature Predicting Recurrence in Stage II/III Colorectal Cancer Based on Machine Learning Algorithms," *Frontiers in Pharmacology* 14 (2023): 1260697.

33. F. Rong, H. Xiang, L. Qian, Y. Xue, K. Ji, and R. Yin, "Machine Learning for Prediction of Outcomes in Cardiogenic Shock," *Frontiers in Cardiovascular Medicine* 9 (2022): 849688.

34. D. Ding, L. Wang, Y. Zhang, K. Shi, and Y. Shen, "Machine Learning Developed a Programmed Cell Death Signature for Predicting Prognosis and Immunotherapy Benefits in Lung Adenocarcinoma," *Translational Oncology* 38 (2023): 101784.

35. P. Zhang, S. Pei, L. Wu, et al., "Integrating Multiple Machine Learning Methods to Construct Glutamine Metabolism-Related Signatures in Lung Adenocarcinoma," *Frontiers in Endocrinology* 14 (2023): 1196372.

36. K. Li, S. Yao, Z. Zhang, et al., "Efficient Gradient Boosting for Prognostic Biomarker Discovery," *Bioinformatics* 38, no. 6 (2022): 1631–1638.

37. T. T. Liu, R. Li, C. Huo, et al., "Identification of CDK2-Related Immune Forecast Model and ceRNA in Lung Adenocarcinoma, a Pan-Cancer Analysis," *Frontiers in Cell and Development Biology* 9 (2021): 682002.

38. Y. Fang, S. Huang, L. Han, S. Wang, and B. Xiong, "Comprehensive Analysis of Peritoneal Metastasis Sequencing Data to Identify LINC00924 as a Prognostic Biomarker in Gastric Cancer," *Cancer Management and Research* 13 (2021): 5599–5611.

39. P. J. Heagerty, T. Lumley, and M. S. Pepe, "Time-Dependent ROC Curves for Censored Survival Data and a Diagnostic Marker," *Biometrics* 56, no. 2 (2000): 337–344.

40. J. Xu, T. Yang, F. Wu, T. Chen, A. Wang, and S. Hou, "A Nomogram for Predicting Prognosis of Patients With Cervical Cerclage," *Heliyon* 9, no. 11 (2023): e21147.

41. M. Ding, X. Ran, S. Qian, et al., "Clinical and Therapeutical Significances of the Cluster and Signature Based on Oxidative Stress for Osteosarcoma," *Aging (Albany NY)* 15, no. 24 (2023): 15360–15381.

42. S. Hanzelmann, R. Castelo, and J. Guinney, "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data," *BMC Bioinformatics* 14 (2013): 7.

43. X. Li, M. Zeng, J. Liu, et al., "Identifying Potential Biomarkers for the Diagnosis and Treatment of IgA Nephropathy Based on Bioinformatics Analysis," *BMC Medical Genomics* 16, no. 1 (2023): 63.

44. S. Orifjon, J. Jammatov, C. Sousa, R. Barros, O. Vasconcelos, and P. Rodrigues, "Translation and Adaptation of the Adult Developmental Coordination Disorder/Dyspraxia Checklist (ADC) Into Asian Uzbekistan," *Sports (Basel)* 11, no. 7 (2023): 135.

45. K. Yoshihara, M. Shahmoradgoli, E. Martinez, et al., "Inferring Tumour Purity and Stromal and Immune Cell Admixture From Expression Data," *Nature Communications* 4 (2013): 2612.

46. A. Mayakonda, D. C. Lin, Y. Assenov, C. Plass, and H. P. Koeffler, "Maftools: Efficient and Comprehensive Analysis of Somatic Variants in Cancer," *Genome Research* 28, no. 11 (2018): 1747–1756.

47. P. Geeleher, N. Cox, and R. S. Huang, "pRRophetic: An R Package for Prediction of Clinical Chemotherapeutic Response From Tumor Gene Expression Levels," *PLoS One* 9, no. 9 (2014): e107468.

48. M. E. Ritchie, B. Phipson, D. Wu, et al., "Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies," *Nucleic Acids Research* 43, no. 7 (2015): e47.

49. X. Robin, N. Turck, A. Hainard, et al., "pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves," *BMC Bioinformatics* 12 (2011): 77.

50. Z. Tan, X. Chen, J. Zuo, S. Fu, H. Wang, and J. Wang, "Comprehensive Analysis of scRNA-Seq and Bulk RNA-Seq Reveals Dynamic Changes in the Tumor Immune Microenvironment of Bladder Cancer and Establishes a Prognostic Model," *Journal of Translational Medicine* 21, no. 1 (2023): 223.

51. S. Jin, C. F. Guerrero-Juarez, L. Zhang, et al., "Inference and Analysis of Cell-Cell Communication Using CellChat," *Nature Communications* 12, no. 1 (2021): 1088.

52. X. Qiu, Q. Mao, Y. Tang, et al., "Reversed Graph Embedding Resolves Complex Single-Cell Trajectories," *Nature Methods* 14, no. 10 (2017): 979–982.

53. J. Chang, H. Wu, J. Wu, et al., "Constructing a Novel Mitochondrial-Related Gene Signature for Evaluating the Tumor Immune Microenvironment and Predicting Survival in Stomach Adenocarcinoma," *Journal of Translational Medicine* 21, no. 1 (2023): 191.

54. P. Wang, Q. Kang, W. S. Wu, and L. Rui, "Hepatic Snai1 and Snai2 Promote Liver Regeneration and Suppress Liver Fibrosis in Mice," *Cell Reports* 43, no. 3 (2024): 113875.

55. C. Tsirigoti, M. M. Ali, V. Maturi, C. H. Heldin, and A. Moustakas, "Loss of SNAI1 Induces Cellular Plasticity in Invasive Triple-Negative Breast Cancer Cells," *Cell Death & Disease* 13, no. 9 (2022): 832.

56. Y. Wang, Y. C. Feng, Y. Gan, et al., "LncRNA MILIP Links YBX1 to Translational Activation of Snai1 and Promotes Metastasis in Clear Cell Renal Cell Carcinoma," *Journal of Experimental & Clinical Cancer Research* 41, no. 1 (2022): 260.

57. Z. Liao, Y. Cheng, H. Zhang, et al., "A Novel Prognostic Signature and Immune Microenvironment Characteristics Associated With Disulfidptosis in Papillary Thyroid Carcinoma Based on Single-Cell RNA Sequencing," *Frontiers in Cell and Development Biology* 11 (2023): 1308352.

58. Z. Qiu, B. Dong, W. Guo, et al., "STK39 Promotes Breast Cancer Invasion and Metastasis by Increasing SNAI1 Activity Upon Phosphorylation," *Theranostics* 11, no. 16 (2021): 7658–7670.

59. S. Zada, J. S. Hwang, M. Ahmed, T. H. Lai, T. M. Pham, and D. R. Kim, "Control of the Epithelial-to-Mesenchymal Transition and Cancer Metastasis by Autophagy-Dependent SNAI1 Degradation," *Cells* 8, no. 2 (2019): 129.

60. L. Boublikova, V. Bakardjieva-Mihaylova, K. Skvarova Kramarzova, et al., "Wilms Tumor Gene 1 (WT1), TP53, RAS/BRAF and KIT Aberrations in Testicular Germ Cell Tumors," *Cancer Letters* 376, no. 2 (2016): 367–376.

61. S. Ariyaratana and D. M. Loeb, "The Role of the Wilms Tumour Gene (WT1) in Normal and Malignant Haematopoiesis," *Expert Reviews in Molecular Medicine* 9, no. 14 (2007): 1–17.

62. J. Park, D. H. Kim, S. R. Shah, et al., "Switch-Like Enhancement of Epithelial-Mesenchymal Transition by YAP Through Feedback Regulation of WT1 and Rho-Family GTPases," *Nature Communications* 10, no. 1 (2019): 2797.

63. X. Chen, S. Lin, Y. Lin, et al., "BRAF-Activated WT1 Contributes to Cancer Growth and Regulates Autophagy and Apoptosis in Papillary Thyroid Carcinoma," *Journal of Translational Medicine* 20, no. 1 (2022): 79.

64. C. C. Wang, S. S. Bajikar, L. Jamal, K. A. Atkins, and K. A. Janes, "A Time- and Matrix-Dependent TGFBR3-JUND-KRT5 Regulatory Circuit in Single Breast Epithelial Cells and Basal-Like Premalignancies," *Nature Cell Biology* 16, no. 4 (2014): 345–356.

65. L. Luo, L. L. Zhang, W. Tao, T. L. Xia, and L. Y. Li, "Prediction of Potential Prognostic Biomarkers in Metastatic Prostate Cancer Based on a Circular RNA-Mediated Competing Endogenous RNA Regulatory Network," *PLoS One* 16, no. 12 (2021): e0260983.

66. A. Vander Ark, J. Cao, and X. Li, "TGF-Beta Receptors: In and Beyond TGF-Beta Signaling," *Cellular Signalling* 52 (2018): 112–120.

67. X. Fang, D. Chen, X. Yang, et al., "Cancer Associated Fibroblasts-Derived SULF1 Promotes Gastric Cancer Metastasis and CDDP Resistance Through the TGFBR3-Mediated TGF-Beta Signaling Pathway," *Cell Death Discov* 10, no. 1 (2024): 111.

68. W. Y. Fang, Y. Z. Kuo, J. Y. Chang, et al., "The Tumor Suppressor TGFBR3 Blocks Lymph Node Metastasis in Head and Neck Cancer," *Cancers (Basel)* 12, no. 6 (2020): 1375.

69. J. Zhang, Q. Zeng, and M. She, "The Roles of FHL2 in Cancer," *Clinical and Experimental Medicine* 23, no. 7 (2023): 3113–3124.

70. Y. Duan, Y. Qiu, X. Huang, C. Dai, J. Yang, and W. He, "Deletion of FHL2 in Fibroblasts Attenuates Fibroblasts Activation and Kidney Fibrosis via Restraining TGF-beta1-Induced Wnt/Beta-Catenin Signaling," *Journal of Molecular Medicine (Berlin, Germany)* 98, no. 2 (2020): 291–307.

71. Y. H. Chen, Z. Q. Wu, Y. L. Zhao, Y. L. Si, M. Z. Guo, and W. D. Han, "FHL2 Inhibits the Id3-Promoted Proliferation and Invasive Growth of Human MCF-7 Breast Cancer Cells," *Chinese Medical Journal* 125, no. 13 (2012): 2329–2333.

72. C. Shao, Y. Qiu, J. Liu, et al., "PARP12 (ARTD12) Suppresses Hepatocellular Carcinoma Metastasis Through Interacting With FHL2 and Regulating Its Stability," *Cell Death & Disease* 9, no. 9 (2018): 856.

73. R. Kanzaki, S. Reid, P. Bolivar, et al., "FHL2 Expression by Cancer-Associated Fibroblasts Promotes Metastasis and Angiogenesis in Lung Adenocarcinoma," *International Journal of Cancer* 156, no. 2 (2025): 431–446.

74. A. Doni, A. Mantovani, B. Bottazzi, and R. C. Russo, "PTX3 Regulation of Inflammation, Hemostatic Response, Tissue Repair, and Resolution of Fibrosis Favors a Role in Limiting Idiopathic Pulmonary Fibrosis," *Frontiers in Immunology* 12 (2021): 676702.

75. H. Zhang, Y. Wang, Y. Zhao, et al., "PTX3 Mediates the Infiltration, Migration, and Inflammation-Resolving-Polarization of Macrophages in Glioblastoma," *CNS Neuroscience & Therapeutics* 28, no. 11 (2022): 1748–1766.

76. F. W. Chen, Y. L. Wu, C. C. Cheng, et al., "Inactivation of Pentraxin 3 Suppresses M2-Like Macrophage Activity and Immunosuppression in Colon Cancer," *Journal of Biomedical Science* 31, no. 1 (2024): 10.

77. N. Y. Kim, H. Y. Lee, and C. Lee, "Metformin Targets Axl and Tyro3 Receptor Tyrosine Kinases to Inhibit Cell Proliferation and Overcome Chemoresistance in Ovarian Cancer Cells," *International Journal of Oncology* 47, no. 1 (2015): 353–360.

78. K. V. Myers, S. R. Amend, and K. J. Pienta, "Targeting Tyro3, Axl and MerTK (TAM Receptors): Implications for Macrophages in the Tumor Microenvironment," *Molecular Cancer* 18, no. 1 (2019): 94.

79. Z. Jiang, S. O. Lim, M. Yan, et al., "TYRO3 Induces Anti-PD-1/PD-L1 Therapy Resistance by Limiting Innate Immunity and Tumoral Ferroptosis," *Journal of Clinical Investigation* 131, no. 8 (2021): e139434.

80. G. L. Yang, Z. Zhao, T. T. Qin, et al., "TNFSF15 Inhibits VEGF-Stimulated Vascular Hyperpermeability by Inducing VEGFR2 Dephosphorylation," *FASEB Journal* 31, no. 5 (2017): 2001–2012.

81. R. S. Al-Lamki, J. Wang, J. S. Pober, and J. R. Bradley, "Co-Expression and Functional Interactions of Death Receptor 3 and E-Selectin in Clear Cell Renal Cell Carcinoma," *American Journal of Pathology* 192, no. 4 (2022): 722–736.

82. M. Gao, M. Wang, S. Zhou, et al., "Machine Learning-Based Prognostic Model of Lactylation-Related Genes for Predicting Prognosis and Immune Infiltration in Patients With Lung Adenocarcinoma," *Cancer Cell International* 24, no. 1 (2024): 400.

83. Z. Yin, T. Ma, B. Huang, et al., "Macrophage-Derived Exosomal MicroRNA-501-3p Promotes Progression of Pancreatic Ductal Adenocarcinoma Through the TGFBR3-Mediated TGF-Beta Signaling Pathway," *Journal of Experimental & Clinical Cancer Research* 38, no. 1 (2019): 310.

84. J. B. McCarthy, D. El-Ashry, and E. A. Turley, "Hyaluronan, Cancer-Associated Fibroblasts and the Tumor Microenvironment in Malignant Progression," *Frontiers in Cell and Development Biology* 6 (2018): 48.

85. M. Alonso-Nocelo, L. Ruiz-Canas, P. Sancho, et al., "Macrophages Direct Cancer Cells Through a LOXL2-Mediated Metastatic Cascade in Pancreatic Ductal Adenocarcinoma," *Gut* 72, no. 2 (2023): 345–359.

86. W. Chen, X. Yu, H. Li, et al., "Single-Cell RNA-Seq Reveals MIF-(CD74 + CXCR4) Dependent Inhibition of Macrophages in Metastatic Papillary Thyroid Carcinoma," *Oral Oncology* 148 (2024): 106654.

87. Z. Cheng, S. Li, S. Yang, et al., "Endoplasmic Reticulum Stress Promotes Hepatocellular Carcinoma by Modulating Immunity: A Study Based on Artificial Neural Networks and Single-Cell Sequencing," *Journal of Translational Medicine* 22, no. 1 (2024): 658.

88. R. Yuan, Q. Fan, X. Liang, et al., "Cucurbitacin B Inhibits TGF-beta1-Induced Epithelial-Mesenchymal Transition (EMT) in NSCLC

Through Regulating ROS and PI3K/Akt/mTOR Pathways," *Chinese Medicine* 17, no. 1 (2022): 24.

89. A. Subuddhi, A. Uosef, D. Zou, H. V. Ubelaker, R. M. Ghobrial, and M. Kloc, "Comparative Transcriptome Profile of Mouse Macrophages Treated With the RhoA/Rock Pathway Inhibitors Y27632, Fingolimod (Gilenya), and Rezurock (Belumosudil, SLx-2119)," *International Immunopharmacology* 118 (2023): 110017.

90. X. Cun, J. Chen, M. Li, et al., "Tumor-Associated Fibroblast-Targeted Regulation and Deep Tumor Delivery of Chemotherapeutic Drugs With a Multifunctional Size-Switchable Nanoparticle," *ACS Applied Materials & Interfaces* 11, no. 43 (2019): 39545–39559.

91. C. Li, A. F. Teixeira, H. J. Zhu, and P. Ten Dijke, "Cancer Associated-Fibroblast-Derived Exosomes in Cancer Progression," *Molecular Cancer* 20, no. 1 (2021): 154.

92. H. Qin, A. Abulaiti, A. Maimaiti, et al., "Integrated Machine Learning Survival Framework Develops a Prognostic Model Based on Inter-Crosstalk Definition of Mitochondrial Function and Cell Death Patterns in a Large Multicenter Cohort for Lower-Grade Glioma," *Journal of Translational Medicine* 21, no. 1 (2023): 588.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Workflow of this study. **Figure S2:** Exploration of the biological functions of 264 candidate genes. (A) Results of the BP analysis in GO. Gray dots represent genes, yellow dots represent different function names, and different colored lines represent the pathways enriched by the genes. The larger the yellow dot, the greater the number of genes enriched in the function. (B) Results of the CC analysis in GO. (C) Results of the MF analysis in GO. (D) Results of the KEGG enrichment analysis. The x-axis represents the number of genes, and the y-axis represents the KEGG pathways. The size of the dots indicates the number of genes enriched in the pathway, and the redder the color of the dot, the smaller the significance *p*-value. (E) PPI network. Yellow nodes represent candidate genes, and the connecting lines represent interaction relationships. **Figure S3:** (A) Schoenfeld residual plot showing the time-dependent effect of candidate genes identified by univariate Cox analysis. The y-axis (Beta (t)) reflects the change in candidate gene coefficients over time (x-axis), with the x-axis representing time. The dashed lines indicate confidence intervals. (B) Risk distribution of PTC patients in TCGA-THCA-test, with a median risk score of −0.007852. (C) KM curves of the LRG ($n = 67$) and HRG ($n = 67$) in TCGA-THCA-test. (D) ROC curves of the TCGA-THCA-test model for 1, 3, and 5 years. **Figure S4:** Clinical features, risk scores, and survival differences between HRG and LRG in PTC. (A) Proportion of PTC patients in different clinical characteristic subgroups within the HRG. (B) Proportion of PTC patients in different clinical characteristic subgroups within the LRG. (C) Differences in risk scores among different clinical subgroups. $^*p < 0.05$, ns (not significant) $p > 0.05$. (D) KM curves of PTC patients in different clinical characteristic subgroups. **Figure S5:** PH assumption test. (A) Results of the PH assumption test for univariate Cox regression analysis. (B) Results of the PH assumption test for multivariate Cox regression analysis. **Figure S6:** (A) Volcano plot of DE-lncRNAs, DE-miRNAs, DE-circRNAs. (B) The intersection of lncRNAs/circRNAs/miRNAs predicted by the miRnet database and DE-lncRNAs/DE-circRNAs/DE-miRNAs in the dataset. (C) lncRNAs-miRNAs-mRNAs network. (D) circRNAs-miRNAs-mRNAs network. (E) ROC curves of the top 5 lncRNAs/circRNAs/miRNSAs. **Figure S7:** (A, B) Single-cell datasets before and after quality control. From left to right are violin plots of nFeature RNA (number of genes in a single cell), nCount RNA (total number of all mRNA molecules detected in a single cell), and percent.mt (proportion of mitochondrial gene expression to total gene expression in the cell). (C) The top 2000 highly variable genes. Red dots represent individual highly variable genes; the x-axis shows average expression levels, and the y-axis shows standardized variance. (D) Significance of the top 50 PCs. Quantile–quantile (Q–Q) plot of *p*-values for PCs from single-cell transcriptomic analysis of PTC. The x-axis represents empirical *p*-values, while the y-axis shows theoretical *p*-values under the null hypothesis. Colored points denote individual PCs with their corresponding *p*-values. Deviations from the diagonal dashed line indicate PCs harboring significant biological signals, potentially linked to EMT dynamics, stromal remodeling, or immune microenvironment features in PTC. And scree plot of the standard deviation distribution of PCs in PCA. (E) UMAP plot of cell clustering. (F) Bubble plot of marker genes for cell annotation. (G) Abundance of each cell type in PTC patients. (H) UMAP plot showing the expression of prognostic genes in different cell types. **Table S1:** The 1184 EMT-RGs. **Table S2:** Marker genes used for cell annotation. **Table S3:** Information on reagents, conditions, and primers used in qPCR experiments. **Table S4:** Complete results of the GO enrichment analysis for the 264 candidate genes. **Table S5:** Complete results of the KEGG enrichment analysis for the 264 candidate genes. **Table S6:** Complete results of the GSEA enrichment analysis for the eight prognostic genes.