International Journal of General Medicine

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Identification of a Four-Gene Signature for Determining the Prognosis of Papillary Thyroid Carcinoma by Integrated Bioinformatics Analysis

Yuting Luo, Rong Chen, Zhikun Ning, Nantao Fu, Minghao Xie

Department of General Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, People's Republic of China

Correspondence: Minghao Xie, Department of General Surgery, The First Affiliated Hospital of Nanchang University, 17 Yongwai Road, Nanchang, Jiangxi, 330006, People's Republic of China, Tel +8613672207521, Email minghao_xie@foxmail.com

Purpose: Although well-differentiated papillary thyroid carcinoma (PTC) has an indolent nature and usually an excellent prognosis, some patients experience disease recurrence or death. The aim of this study was to identify prognostic markers to stratify PTC patients.

Patients and Methods: Eight gene-expression profiles (GSE3467, GSE3678, GSE5364, GSE27155, GSE33630, GSE53157, GSE60542, and GSE104005) were obtained from the Gene Expression Omnibus and used to analyze differentially expressed genes (DEGs) between PTC tissues and non-tumor tissues. Univariable Cox regression survival analysis and Lasso-penalized Cox regression analysis were performed to identify prognostic genes and establish a risk-score model based on the integrated DEGs. Kaplan–Meier (KM) and receiver operating characteristic (ROC) curves were used to validate the prognostic performance of the risk score. A nomogram was constructed based on The Cancer Genome Atlas dataset and Multivariable Cox regression analysis.

Results: A total of 165 upregulated and 207 downregulated DEGs were screened. A four-gene signature including PAPSS2, PCOLCE2, PTX3, and TGFBR3 was identified. The risk-score model showed a strong diagnosis performance for identifying patients with a poor prognosis. KM analysis showed that patients with low risk scores had a significantly more favorable overall survival (OS) than those with high risk scores (p = 0.0002). ROC curves based on the four-gene signature showed better performances in predicting 1-, 3-, and 5-year survival than did the American Joint Committee on Cancer staging system (area under the curve: 0.86 vs 0.84, 0.80 vs 0.63, and 0.79 vs 0.73, respectively). Furthermore, when combined with age and tumor status from the nomogram, the four-gene signature achieved a good performance in guiding postoperative follow-up surveillance of patients with PTC.

Conclusion: The four-gene signature was found to be a novel and reliable biomarker with great potential for clinical application in risk stratification and OS prediction in patients with PTC.

Keywords: gene signature, overall survival, papillary thyroid cancer, prognosis

Introduction

Thyroid carcinoma (THCA) is the most common type of endocrine malignancy and its incidence is increasing.[1] Based on its histopathological characteristics, thyroid carcinoma can be classified into multiple subtypes, such as papillary thyroid carcinoma (PTC), follicular thyroid carcinoma, and anaplastic thyroid carcinoma.[2] PTC is the most common subtype of THCA, comprising approximately 80% of all thyroid malignancies.[3] The prognosis for patients with PTC is highly favorable, although more than 10% of patients eventually experience local or distant disease recurrence.[4] Those patients with potentially poor outcomes must be monitored and administered timely and effective treatments to prolong their survival and improve their quality of life. Therefore, it is necessary to identify effective prognostic biomarkers of PTC to accurately assess their prognosis and to stratify patients.

At present, treatment options for PTC and disease prognosis depend mainly on the histopathological subtype. Some variants of PTC may behave more aggressively than classic PTC. These so-called aggressive variants include the tall-cell variant, the columnar variant, the solid variant, and the recently described hobnail variant. The tall-cell variant is the most common aggressive variant of PTC. The tumor cells have typical nuclear features of PTC, composed of

Received: 7 November 2021
Accepted: 21 January 2022
Published: 4 February 2022

prominent papillary structures lined by cells 2 or 3 times as tall as they are wide, and have abundant eosinophilic (oncocytic-like) cytoplasm. The columnar variant can also be aggressive, particularly in older patients, with larger tumors, presenting symptomatically, and showing a diffusely infiltrative growth, and extrathyroidal extension. The solid variant not only has the same growth pattern with poorly differentiated THCA but also has the nuclear features of PTC. The histologic features of hobnail variant include papillary and micropapillary structures closely lined by cells containing eosinophilic cytoplasm and apically located nuclei with prominent nucleoli. The tumors cells have increased nuclear to cytoplasmic ratios and apically placed nuclei that produce a hobnail like surface bulge.[5] However, aggressive PTC variants remain largely understudied.[6] Some reports have shown that the diffuse-sclerosing variant is an indolent variant of PTC,[7,8] but other data have suggested that the variant is a risk factor for a poor patient prognosis.[9,10] Moreover, in the absence of explicit standards, histopathology definitions vary widely across institutions. There has been considerable disagreement among different investigators concerning the threshold percentage of tall cells (30–70%) that constitute a true tall-cell PTC variant.[11–13] Therefore, the use of histopathological subtyping of PTC in routine clinical practice is limited.

In addition to the histopathological differences, the molecular features of PTC may also vary. Programmed cell death protein 1 (PD-1) is an apoptosis-associated gene that involved in the regulation of the immune response and is one of the most important inhibitory checkpoints.[14] Girolami et al[15] reported that the status of BRAF mutation was significant association with programmed death-ligand 1 (PD-L1) expression in PTC patients. Immunotherapy may be considered in a subset of BRAF mutant PTC. However, the reported rates of BRAF mutation vary significantly between studies. BRAF mutation has been observed in 80–100% tall-cell PTC variant patients. Whereas the BRAF mutation rate is similar between the columnar variant and classic PTC patients. With developments in gene chips and high-throughput sequencing, gene signatures based on mRNA-expression levels have shown great potential for determining PTC prognosis. However, the data from previous studies have often been incomplete or inconsistent with one another. A comprehensive analysis of integrated gene-expression data using bioinformatics methods can overcome the problems caused by heterogeneities in expression studies.

In this study, eight microarray datasets of differentially expressed genes (DEGs) from the Gene Expression Omnibus (GEO) were integrated and used for bioinformatics analyses. Univariable and Lasso–Cox regression analysis were performed to identify overall survival (OS)-related DEGs and propose a prognostic risk-score model to stratify PTC patients. Multivariable Cox regression analysis was used to identify independent prognostic factors for PTC. We identified a four-gene signature as a robust marker with great potential in risk stratification and OS prediction. Furthermore, based on the GeneMANIA database, we analyzed the potential biological functions of this four-gene signature by performing interaction-network and pathway-enrichment analyses. Finally, a prognostic nomogram was constructed based on The Cancer Genome Atlas (TCGA) dataset to guide postoperative follow-up surveillance of patients with PTC.

Materials and Methods

Microarray Preparation and Data Processing

Eight microarray datasets (GSE3467, GSE3678, GSE5364, GSE27155, GSE33630, GSE53157, GSE60542, and GSE104005) were downloaded from the GEO database (https://www.ncbi.nlm.nih.gov/geo/) for DEG analysis, using “thyroid carcinoma” and “Homo sapiens” as search terms. The raw data from these datasets were processed using R language statistical software (version 3.6.1). Background correction was performed by processing the raw data by quartile normalization followed by log2 transformation, in order to obtain normally distributed expression values. The “Limma” package was used to identify the DEGs between PTC tissues and non-tumor tissues.[16] The threshold for statistical significance was set at a log2 fold change (log2FC) of >1 and an adjusted P value of <0.05. Mean expression values were applied for genes whose expression values were studied using multiple probes. The robust rank aggregation (RRA) method was used to identify overlapping DEGs ($P < 0.05$) from the eight GEO datasets.

Constructing a Prognostic Gene Signature

Univariable Cox regression survival analysis was performed based on the overlapping DEGs (where $P < 0.01$ was considered statistically significant). Thereafter, Lasso-penalized Cox regression analysis was performed to construct the prognostic gene signature.[17] The optimal values of the penalty parameter, alpha, were determined using the "glmnet" package of R software.[18] A four-gene prognostic signature with corresponding coefficients was selected. Risk scores were calculated using TCGA data for each patients with thyroid carcinoma, using the "survminer" package of R software, and the patients were divided into two groups based on the median risk score.

TCGA Database Validation and Survival Analysis

Patients with PTC and follow-up times of >30 days in TCGA database were included in the survival analysis. The Kaplan–Meier (KM) curves were analyzed using the "survival" package of R software. Time-dependent receiver operating characteristic (ROC) curve analysis was conducted using the "pROC" and "survivalROC" packages of R.[19] Patients with complete clinical information ($n = 426$) were included in the univariable and multivariable Cox regression analyses. Risk scores and other clinical characteristics (including the age, sex, tumor–node–metastasis classification, tumor location, residual tumor status, American Joint Committee on Cancer [AJCC] pathologic tumor stage, and tumor status at the last follow-up) were analyzed by univariable Cox regression analysis. Variants with a P value of <0.1 in univariable analysis were included in the subsequent multivariable Cox regression model to assess the predictive performance.

Constructing a Predictive Nomogram

The "rms" package of R software was used to construct a predictive nomogram based on independent prognostic parameters used to predict the probabilities of 3-, 5- and 8-year OS. The time-dependent area under the ROC curve (AUC) was calculated to determine the discriminatory ability of the nomogram. A calibration curve was used to visualize the performance of the nomogram. The patients were stratified into two groups according to the median nomogram score. The KM survival curves for the different groups were then plotted.

Functional-Enrichment Analysis

Pathway-enrichment analysis was carried out using the Kyoto Encyclopedia of Genes and Genomes (KEGG, http://www.genome.jp/kegg/) and Enrichr (http://amp.pharm.mssm.edu/Enrichr/) databases. Gene ontology (GO, http://www.geneontology.org/) was used to explore the biological functions of the four-gene signature. GO functional annotation included biological process (BP), cellular component (CC), and molecular function (MF) terms.

Statistical Analysis

The data generated in this study are expressed as the mean ± standard deviation. R software (version 3.6.1) was used for all statistical analyses. Categorical variables were analyzed by the χ^2 test or Fisher's exact test. Continuous variables were analyzed using Student's *t*-test for paired samples. Two groups of boxplots were analyzed using the Wilcoxon test. A heatmap was generated to visualize the DEGs using the "pheatmap" package of R software. KM survival curves and hazard ratios (HRs) with 95% confidence intervals (CIs) were constructed using log–rank tests. Correlations among the individual genes in the signature were assessed with Pearson correlation coefficients. $P < 0.05$ was considered to reflect a statistically significant difference.

Ethics Approval

The study was reviewed and approved by the Institutional Review Board and the Ethics Committee of the First Affiliated Hospital of Nanchang University, Nanchang, China.

Results

Identification of DEGs

In this study, multistep analysis was performed to explore key DEGs in patients with PTC, and their biological functions and potentials for clinical application were studied using integrated bioinformatics methods. The overall study workflow is shown in Figure 1. The eight microarray datasets studied were normalized to increase the reliability of the data (Figure 2A and Supplementary Figure 1). Hierarchical clustering analysis revealed different DEG profiles in the tumor and matched non-tumor tissues (Figure 2B and Supplementary Figure 2). The volcano plots shown in Figure 2C and Supplementary Figure 3 represent the distributions of DEGs identified from each of the eight datasets. Integrated analysis of the eight datasets using the RRA method identified 372 overlapping DEGs, including 165 upregulated and 207 downregulated genes. The top 20 overlapping upregulated and downregulated DEGs in the eight datasets are shown in Figure 2D.

Screening and Expression Validation of Key Genes

Univariable Cox regression analysis revealed that 36 of the 372 DEGs identified correlated significantly with OS (Figure 3A). Lasso-penalized Cox regression analysis was then performed to identify a four-gene prognostic signature, consisting of 3′-phosphoadenosine 5′-phosphosulfate synthase 2 (PAPSS2), procollagen C-endopeptidase enhancer 2 (PCOLCE2), pentraxin 3 (PTX3), and transforming growth factor beta receptor 3 (TGFBR3) (Figure 3B). The four genes were significantly downregulated in THCA tissues, compared to their expression levels in normal tissues (Figure 3C). To validate expression differences in the four genes between tumor and non-tumor tissues, 512 THCA tissues and 59 normal tissues were compared using Gene Expression Profiling Interactive Analysis (http://gepia.cancer-pku.cn/). As shown in Figure 3D, the mRNA expression levels of the four genes were decreased in THCA tissues.

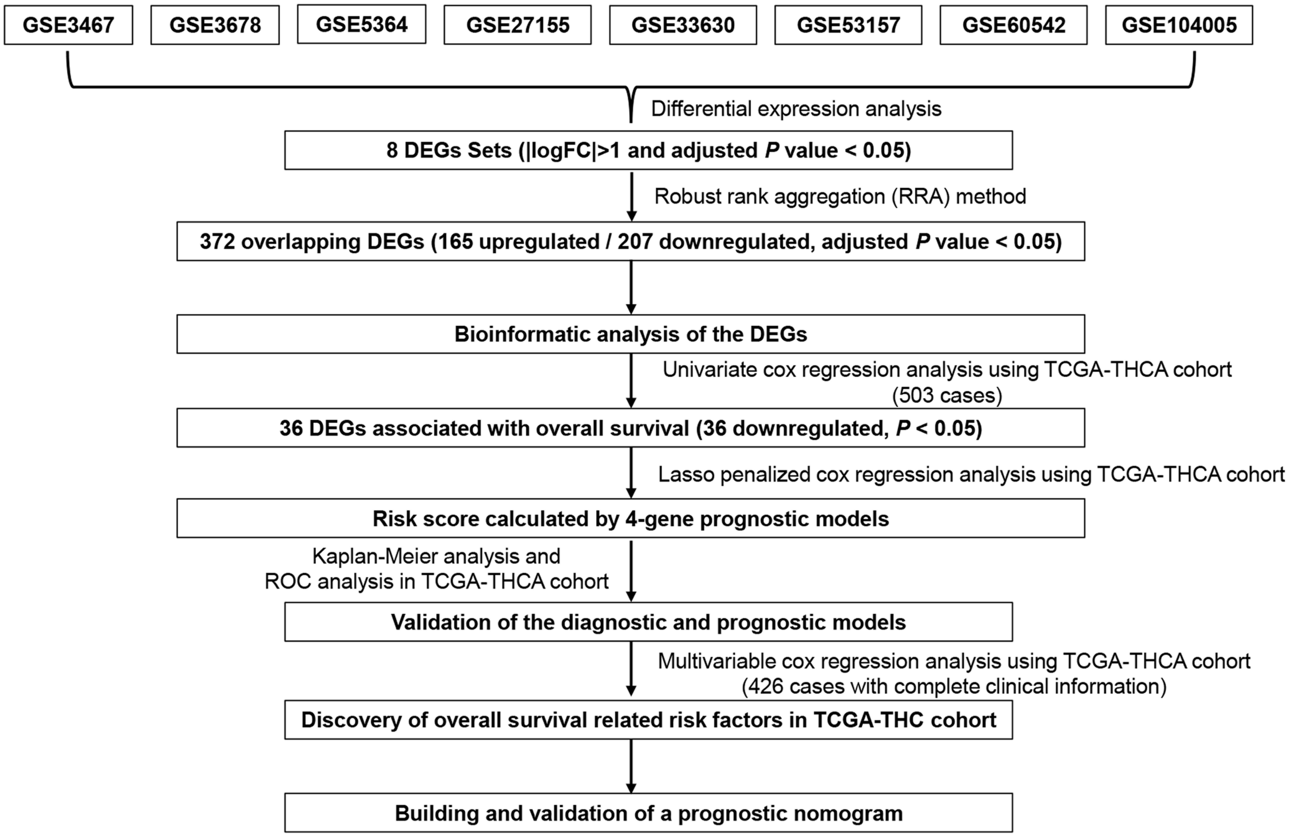
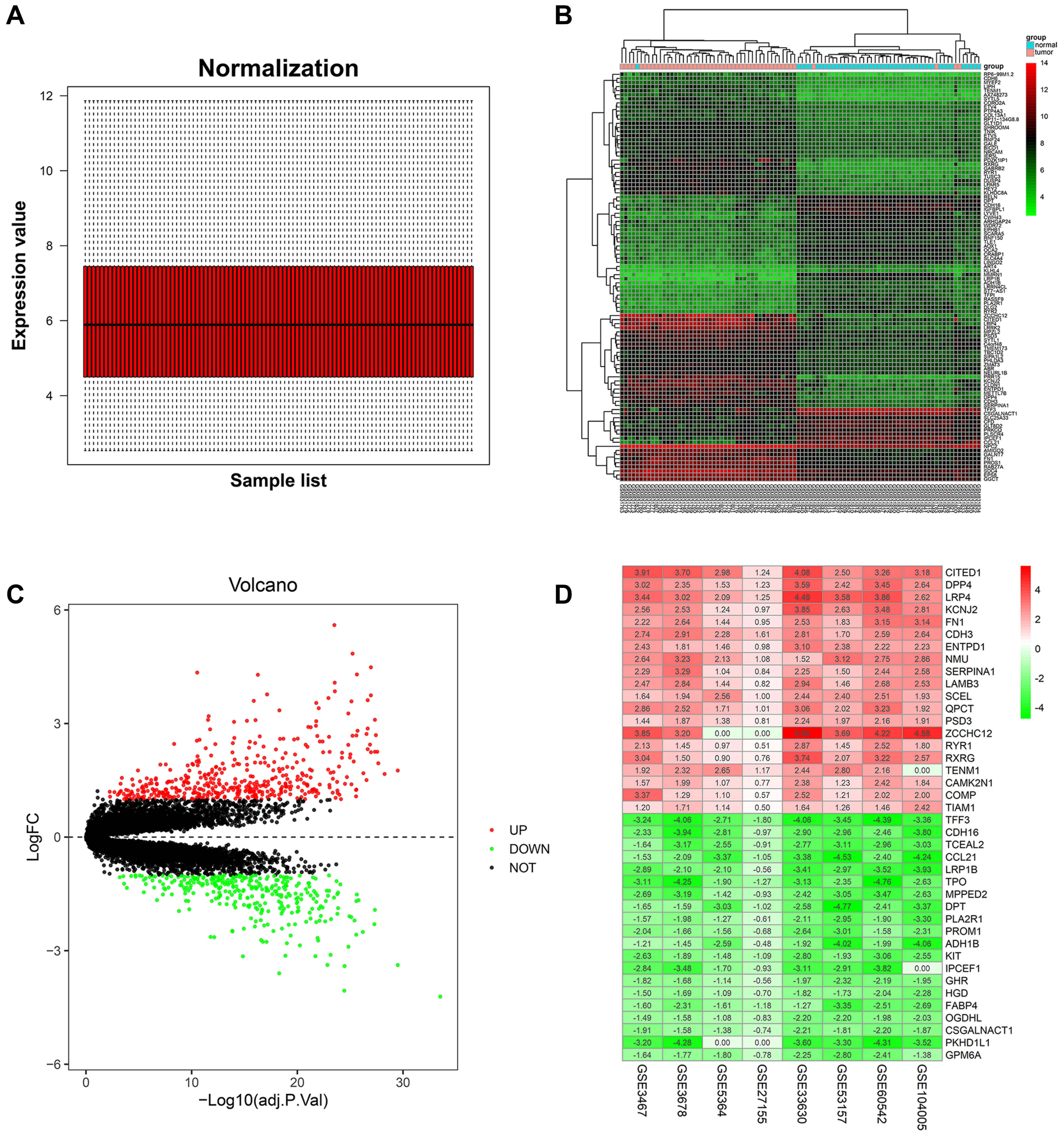

GSE3467
GSE3678
GSE5364
GSE27155
GSE33630
GSE53157
GSE60542
GSE104005
Differential expression analysis
8 DEGs Sets (|logFC|>1 and adjusted P value < 0.05)
Robust rank aggregation (RRA) method
372 overlapping DEGs (165 upregulated / 207 downregulated, adjusted P value < 0.05)
Bioinformatic analysis of the DEGs
Univariate cox regression analysis using TCGA-THCA cohort (503 cases)
36 DEGs associated with overall survival (36 downregulated, P < 0.05)
Lasso penalized cox regression analysis using TCGA-THCA cohort
Risk score calculated by 4-gene prognostic models
Kaplan-Meier analysis and ROC analysis in TCGA-THCA cohort
Validation of the diagnostic and prognostic models
Multivariable cox regression analysis using TCGA-THCA cohort (426 cases with complete clinical information)
Discovery of overall survival related risk factors in TCGA-THC cohort
Building and validation of a prognostic nomogram

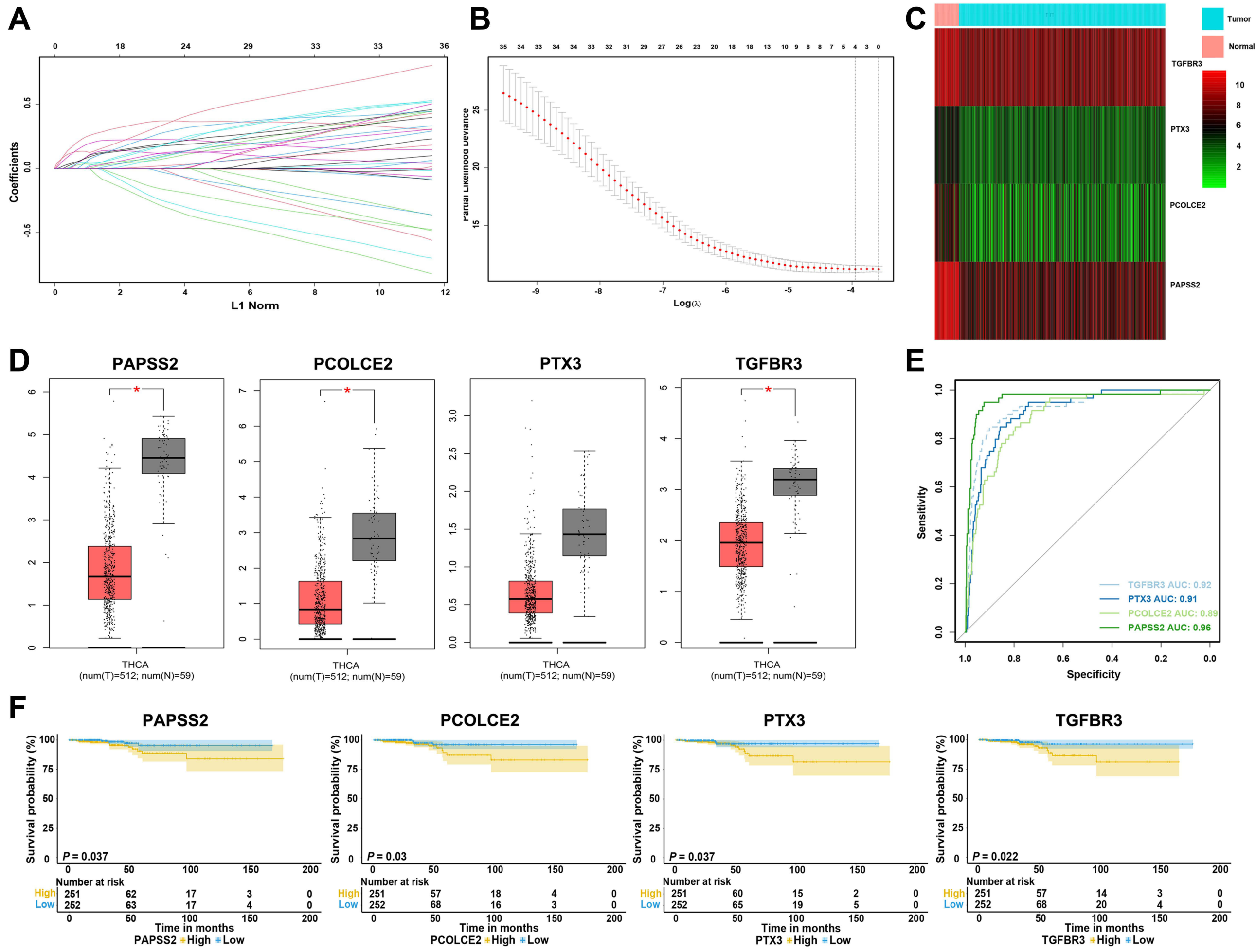
Figure 1 Flowchart showing the workflow used for the identification, bioinformatic analysis, and validation of the prognostic gene signature of PTC.



GSE3467	GSE3678	GSE5364	GSE27155	GSE33630	GSE53157	GSE60542	GSE104005	
3.91	3.70	2.98	1.24	4.08	2.50	3.26	3.18	CITED1
3.02	2.35	1.53	1.23	3.59	2.42	3.45	2.64	DPP4
3.44	3.02	2.09	1.25	4.49	3.58	3.86	2.62	LRP4
2.56	2.53	1.24	0.97	3.85	2.63	3.48	2.81	KCNJ2
2.22	2.64	1.44	0.95	2.53	1.83	3.15	3.14	FN1
2.74	2.91	2.28	1.61	2.81	1.70	2.59	2.64	CDH3
2.43	1.81	1.46	0.98	3.10	2.38	2.22	2.23	ENTPD1
2.64	3.23	2.13	1.08	1.52	3.12	2.75	2.86	NMU
2.29	3.29	1.04	0.84	2.25	1.50	2.44	2.58	SERPINA1
2.47	2.84	1.44	0.82	2.94	1.46	2.68	2.53	LAMB3
1.64	1.94	2.56	1.00	2.44	2.40	2.51	1.93	SCEL
2.86	2.52	1.71	1.01	3.06	2.02	3.23	1.92	QPCT
1.44	1.87	1.38	0.81	2.24	1.97	2.16	1.91	PSD3
3.85	3.20	0.00	0.00	5.60	3.69	4.22	4.58	ZCCHC12
2.13	1.45	0.97	0.51	2.87	1.45	2.52	1.80	RYR1
3.04	1.50	0.90	0.76	3.74	2.07	3.22	2.57	RXRG
1.92	2.32	2.65	1.17	2.44	2.80	2.16	0.00	TENM1
1.57	1.99	1.07	0.77	2.38	1.23	2.42	1.84	CAMK2N1
3.37	1.29	1.10	0.57	2.52	1.21	2.02	2.00	COMP
1.20	1.71	1.14	0.50	1.64	1.26	1.46	2.42	TIAM1
-3.24	-4.06	-2.71	-1.80	-4.06	-3.45	-4.39	-3.36	TFF3
-2.33	-3.94	-2.81	-0.97	-2.90	-2.96	-2.46	-3.80	CDH16
-1.64	-3.17	-2.55	-0.91	-2.77	-3.11	-2.96	-3.03	TCEAL2
-1.53	-2.09	-3.37	-1.05	-3.38	-4.53	-2.40	-4.24	CCL21
-2.89	-2.10	-2.10	-0.56	-3.41	-2.97	-3.52	-3.93	LRP1B
-3.11	-4.25	-1.90	-1.27	-3.13	-2.35	-4.76	-2.63	TPO
-2.69	-3.19	-1.42	-0.93	-2.42	-3.05	-3.47	-2.63	MPPED2
-1.65	-1.59	-3.03	-1.02	-2.58	-4.77	-2.41	-3.37	DPT
-1.57	-1.98	-1.27	-0.61	-2.11	-2.95	-1.90	-3.30	PLA2R1
-2.04	-1.66	-1.56	-0.68	-2.64	-3.01	-1.58	-2.31	PROM1
-1.21	-1.45	-2.59	-0.48	-1.92	-4.02	-1.99	-4.06	ADH1B
-2.63	-1.89	-1.48	-1.09	-2.80	-1.93	-3.06	-2.55	KIT
-2.84	-3.48	-1.70	-0.93	-3.11	-2.91	-3.82	0.00	IPCEF1
-1.82	-1.68	-1.14	-0.56	-1.97	-2.32	-2.19	-1.95	GHR
-1.50	-1.69	-1.09	-0.70	-1.82	-1.73	-2.04	-2.28	HGD
-1.60	-2.31	-1.61	-1.18	-1.27	-3.35	-2.51	-2.69	FABP4
-1.49	-1.58	-1.08	-0.83	-2.20	-2.20	-1.98	-2.03	OGDHL
-1.91	-1.58	-1.38	-0.74	-2.21	-1.81	-2.20	-1.87	CSGALNACT1
-3.20	-4.28	0.00	0.00	-3.60	-3.30	-4.31	-3.52	PKHD1L1
-1.64	-1.77	-1.80	-0.78	-2.25	-2.80	-2.41	-1.38	GPM6A

Figure 2 Identification of DEGs in PTC between tumor and normal tissues. (**A**) Normalization of the GSE33630 dataset. (**B**) A representative heatmap of dataset GSE33630 showing that DEGs can effectively differentiate tumors from normal tissues (red represents higher expression and green represents lower expression). (**C**) Volcano plot of the GSE33630 dataset (green: downregulated genes; black: not differentially expressed; red: upregulated genes). (**D**) Heat map of the top 20 upregulated (red) and downregulated (green) DEGs of each expression microarray. The number in each column represents the LogFC value.

In addition, 503 PTC patients with a follow-up period of >30 days were included in subsequent analysis (two TCGA-THCA patients were excluded because of a diagnosis of follicular carcinoma or poorly differentiated oncocytic carcinoma). The baseline characteristics of the patients are presented in Table 1. All four genes showed good correlations with PTC prognosis, as shown by ROC analysis (TGFBR3 AUC = 0.92, PTX3 AUC = 0.91, PCOLCE2 AUC = 0.89, and PAPSS2 AUC = 0.96) (Figure 3E). Furthermore, survival analyses were performed to verify correlations between gene-expression levels and prognosis. The patient cohorts were stratified into low- and high-risk



Number at risk	0	50	100	150	200
PAPSS2 High	251	62	17	3	0
PAPSS2 Low	252	63	17	4	0
PCOLCE2 High	251	57	18	4	0
PCOLCE2 Low	252	68	16	3	0
PTX3 High	251	60	15	2	0
PTX3 Low	252	65	19	5	0
TGFBR3 High	251	57	14	3	0
TGFBR3 Low	252	68	20	4	0

Figure 3 Identification of genes significantly correlated with PTC prognosis. (**A** and **B**) The four-gene prognostic signature was identified by univariate and Lasso–Cox regression analysis. (**C**) Heat map showing expression differences for the four genes between THCA and normal tissues. (**D**) Expression differences in the four genes were validated in THCA and normal tissues with GEPIA. (**E**) ROC analysis for the four genes in patients with PTC. (**F**) KM survival curves for the four genes in patients with PTC. *P < 0.05.

groups according to the median expression of each gene. KM survival curves showed that patients in the high-risk group (with low expression of PAPSS2, PCOLCE2, PTX3, or TGFBR3) had a higher mortality rate than those in the low-risk group (Figure 3F).

Correlations Between Key Genes and Clinical Characteristics

The expression levels of the four hub genes were analyzed according to various clinical characteristics, including the age, tumor (T) classification, node (N) classification, metastasis (M) classification, and tumor status at the time of last follow-up. Primarily, the expression levels of the four key genes were significantly lower in patients with PTC, when compared with those in healthy subjects (Figure 4). However, their expression levels did not differ according to the presence of absence of extrathyroidal invasion (Figure 4A). TGFBR3-expression levels were significantly higher in patients with PTC and lymph node metastasis, when compared with those in patients without lymph node metastasis (Figure 4B). The expression levels of the four genes were significantly lower in patients with distant metastasis compared with those in patients without distant metastasis (Figure 4C). PCOLCE2 levels were significantly lower in younger patients (< 55 years) compared with those in older patients (≥ 55 years), as shown in Figure 4D. PTX3-expressions levels were significantly lower in the patients surviving with tumors than in tumor-free patients (Figure 4E).

Table 1 Baseline Characteristics of Patients with PTC, Using TCGA Data

Clinical Feature	Number (%)	Clinical Feature	Number (%)
Follow-up time (months)	40.0 ± 32.5	**AJCC stage**	
Survival status		Stage I	283 (56.3%)
Alive	487 (96.8%)	Stage II	51 (10.1%)
Dead	16 (3.2%)	Stage III	112 (22.3%)
Age (year)	47.3 ± 15.8	Stage IV	55 (10.9%)
Sex		Not available	2 (0.4%)
Male	135 (26.8%)	**T classification**	
Female	368 (73.2%)	T1	143 (28.4%)
Pharmaceutical treatment		T2	165 (32.8%)
No	218 (43.3%)	T3	170 (33.8%)
Yes	17 (3.4%)	T4	23 (4.6%)
Not available	268 (53.3%)	TX	2 (0.4%)
Radiation treatment		**N classification**	
No	78 (15.5%)	N0	229 (45.5%)
Yes	156 (31.0%)	N1	224 (44.6%)
Not available	269 (53.5%)	NX	50 (9.9%)
Tumor status		**M classification**	
Tumor free	410 (81.5%)	M0	281 (55.9%)
With tumor	50 (9.9%)	M1	9 (1.8%)
Not available	43 (8.6%)	MX	212 (42.1%)
Laterality		Not available	1 (0.2%)
Left lobe	176 (35.0%)	**Residual tumor**	
Right lobe	213 (42.3%)	RO	385 (76.5%)
Isthmus	22 (4.4%)	R1	52 (10.3%)
Bilateral	86 (17.1%)	R2	4 (0.8%)
Not available	6 (1.2%)	RX	30 (6.0%)
Extrathyroidal extension		Not available	32 (6.4%)
None	332 (66.0%)	**Relapse**	
Minimal (T3)	134 (26.6%)	Disease-free	443 (88.1%)
Moderate/advanced (T4a)	18 (3.6%)	Recurred	46 (9.1%)
Very advanced (T4b)	1 (0.2%)	Not available	14 (2.8%)
Not available	18 (3.6%)		

Abbreviation: AJCC stage, American Joint Committee on Cancer pathologic tumor stage.

Development of the Four-Gene Prognostic Signature

The coefficient values and expression levels of the four genes were determined to calculate the risk core for each patient using the following formula: (0.17363 × PAPSS2 level) + (0.00064 × PCOLCE2 level) + (0.07881 × PTX3 level) + (0.14243 × TGFBR3 level). Subsequently, the included 503 TCGA-PTC patients were divided into high-risk (n = 251) and low-risk (n = 252) groups, according to the median risk score of 2.759 (Figure 5A). KM survival-curve analysis showed that the high-risk group had a worse OS than the low-risk group (p = 0.0002; Figure 5B). The calibration plots for 1-, 3-, and 5-year survival predictions showed that the four-gene signature had satisfactory predictive performances for patients with PTC (Figure 5C). Time-dependent ROC-curve analysis was performed to evaluate the sensitivity and specificity of the four-gene risk scores. The AUCs of the four-gene signature were significantly greater than those of the AJCC stage at 1 year (0.86 vs 0.84), 3 years (0.80 vs 0.63), and 5 years (0.79 vs 0.73) (Figure 5D).

Four-Gene Signature Interaction Network, Functional Annotation, and Pathway-Enrichment Analysis

To further explore the biological functions of these four genes, the Enrichr tool was utilized to analyze GO functions, KEGG pathways, and WikiPathways enrichment. The top three enriched GO terms and pathways are shown in Figure 6.

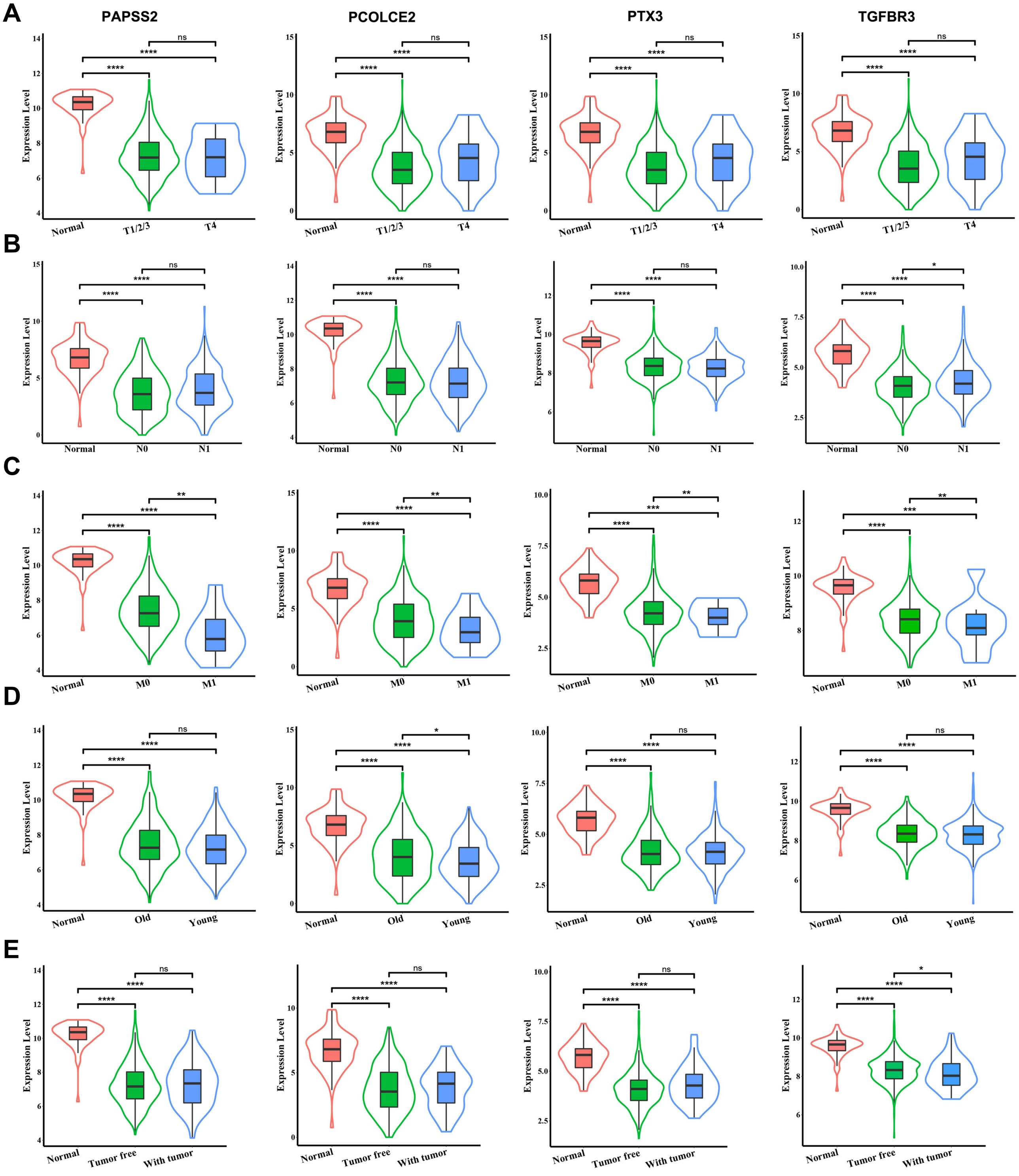


Figure 4 Expression of four key genes in subgroups of patients with PTC, stratified according to their clinical characteristics. The violin plots present quantitative analysis of the relative expression levels of four key genes in normal individuals and patients with PTC. (**A** and **C**) Patients with PTC were divided into two groups according to presence or absence of extrathyroidal invasion (**A**), lymph node metastasis (**B**), or distant metastasis (**C**). (**D** and **E**) An age of ≥55 years and survival with tumors were analyzed as group variables for patients with PTC. ns: P > 0.05; *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001.

GO BP analysis showed that the four genes were significantly enriched for "glycoprotein metabolism," "cardiac muscle cell proliferation," and "striated muscle cell proliferation." CC analysis showed that "tertiary granule lumen," "specific granule lumen," and "specific granule" were among the most highly enriched subcategories. Enrichment analysis of MF

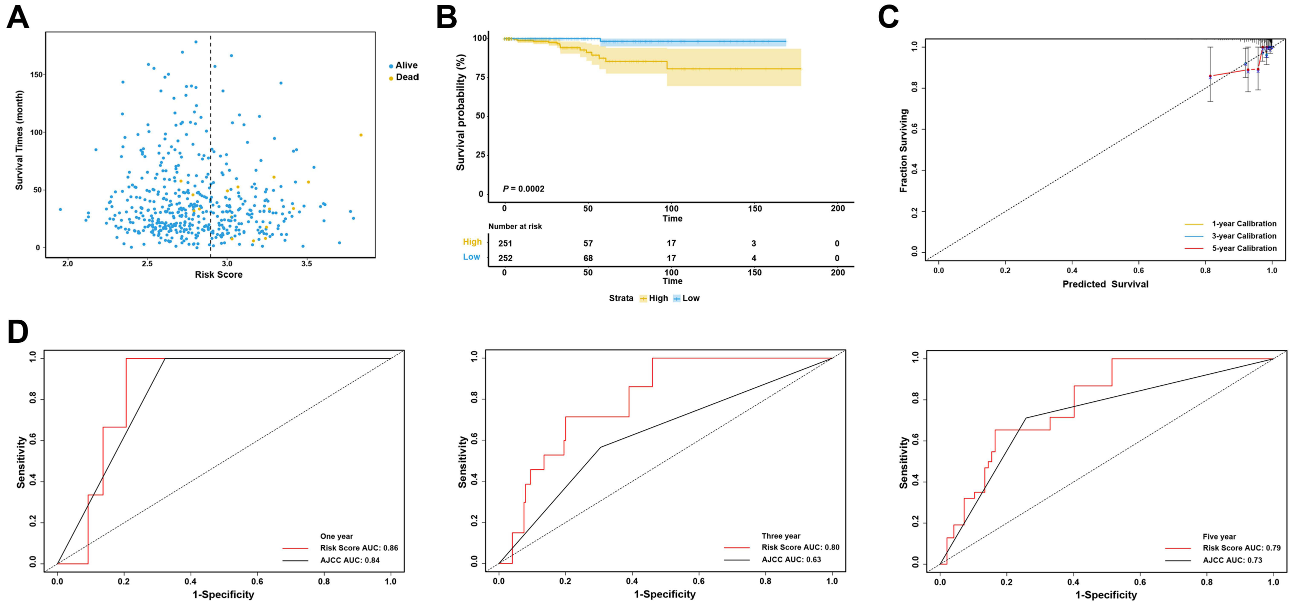


Figure 5 Establishment and evaluation of the four-gene signature in TCGA-PTC patients. (**A**) Distributions of the risk scores and survival data. Patients that were alive or dead at the study endpoint are represented with blue and yellow dots, respectively. (**B**) KM survival curves of the risk-score model for 503 TCGA-PTC patients. (**C**) Calibration plot of the four-gene signature for predicting the survival probability at 1, 3, or 5 years after the initial treatment. (**D**) Time-dependent ROC curves of the risk-score model for predicting 1-, 3-, and 5-year OS rates, compared with the results obtained using the AJCC staging system.

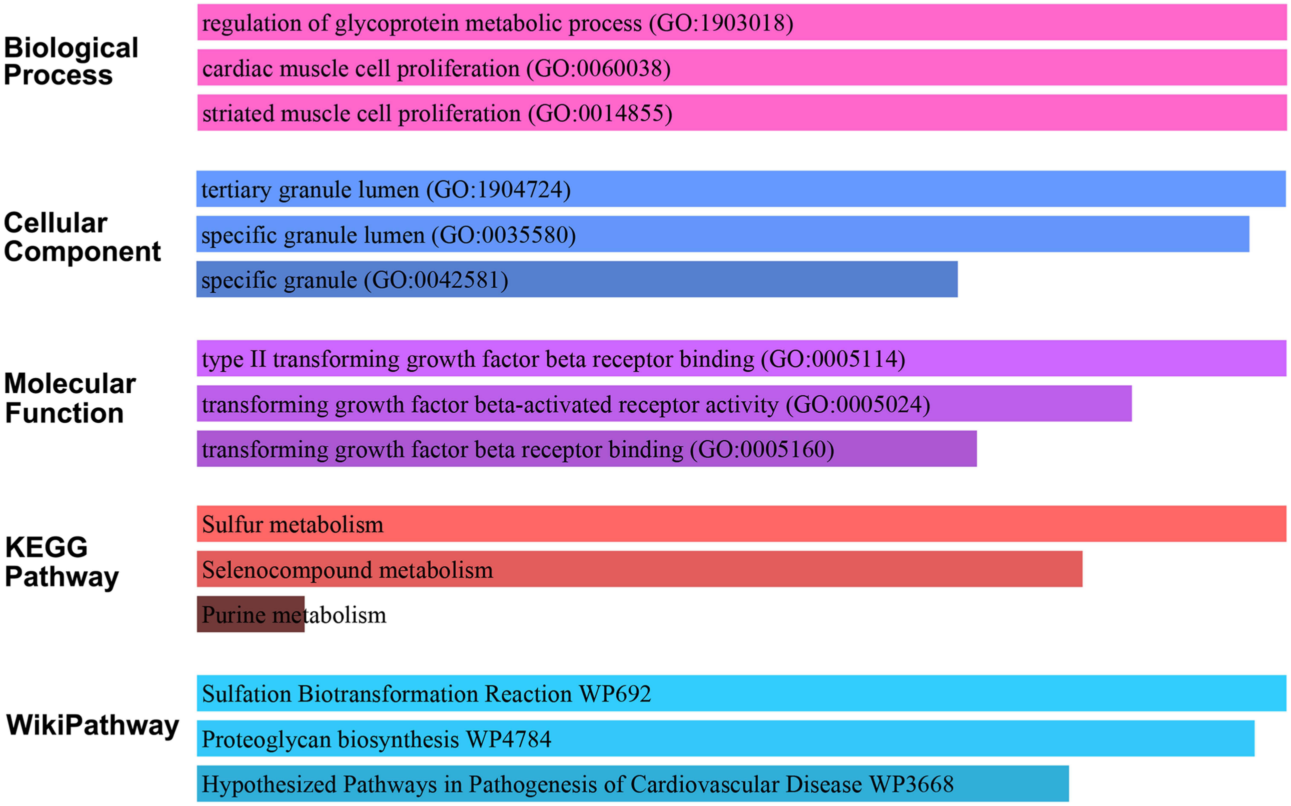


Figure 6 Enrichment analysis of the four key genes. The bar chart visualizes the top three enriched terms and their P values. The longer the bar, the smaller the P-value.

terms revealed that the four key genes were mainly enriched for “type II transforming growth factor beta (TGF-β) receptor binding,” “TGF-β-activated receptor activity,” and “TGF-β receptor activity.” KEGG pathway-enrichment analysis revealed that the four key genes were significantly enriched in terms of “sulfur metabolism,” “selenocompound metabolism,” and “purine metabolism.” WikiPathways analysis revealed an enrichment for the four key genes in terms of the “sulfation biotransformation reaction,” “proteoglycan biosynthesis,” and “hypothesized pathways in pathogenesis of cardiovascular disease” categories.

Construction and Verification of a Predictive Nomogram

Data from 425 TCGA-PTC patients with complete clinical information were used to build a prognostic nomogram. Univariate Cox regression analysis showed that the risk score, age, presence of residual tumors, AJCC stage, and tumor status were significantly associated with patient survival. However, multivariate Cox regression analyses showed that only the risk score, age, and tumor status were independent prognostic factors for patients with PTC (Table 2).

To develop a quantitative method for determining the prognosis of patients with PTC in clinical settings, we established a nomogram by integrating the age, tumor status, and four-gene signature risk score (Figure 7A). Subsequently, the patients with PTC were divided into groups with a high risk score (n = 212) or a low risk score (n = 213), based on median of the nomogram (score = 68) (Figure 7B). Calibration plots showed that the nomogram performed well in predicting OS in patients with PTC (Figure 7C). The KM curves showed a significant difference in the OS between the two groups. Those with higher scores had a significantly lower OS period (p = 0.00042; Figure 7D). The AUCs of the 3-, 5-, and 8-year OS predictions for the nomogram were 1, 0.97, and 0.94, respectively (Figure 7E).

Table 2 Univariate and Multivariate Cox Regression Analysis of TCGA Data from Patients with PTC (n = 425)

Characteristics	**Univariate Cox**		**Multivariate Cox**	
	HR (95% CI)	***P***	**HR (95% CI)**	***P***
Risk score				
Low	1		1	
High	12.814 (1.573–104.38)	**0.017**	12.103 (1.226–119.469)	**0.033**
Age (years)				
<55	1		1	
≥55	30.020 (3.671–245.503)	**0.002**	29.126 (2.653–319.759)	**0.006**
Sex				
Female	1		–	–
Male	3.174 (0.789–12.773)	0.104	–	–
Extrathyroid extension				
None	1		–	–
Minimal–advanced	1.101 (0.262–4.622)	0.895	–	–
Laterality				
Unilateral	1			
Isthmus	0.038 (0–43,447.345)	0.646	–	–
Bilateral	0.036 (0–838.309)	0.518	–	–
Residual tumor				
R0	1		1	
Non-R0	5.491 (1.372–21.967)	**0.016**	2.941 (0.563–15.358)	0.201
AJCC stage				
I + II	1		1	
III + IV	4.501 (1.071–18.921)	**0.040**	0.122 (0.100–1.556)	0.105
Tumor status				
Tumor-free	1		1	
With tumor	18.057 (4.305–75.740)	**<0.001**	18.822 (1.845–191.987)	**0.013**

Note: Significant p-values (< 0.05) are bolded.
Abbreviations: HR, hazard ratios; CI, confidence intervals; AJCC stage, American Joint Committee on Cancer pathologic tumor stage.

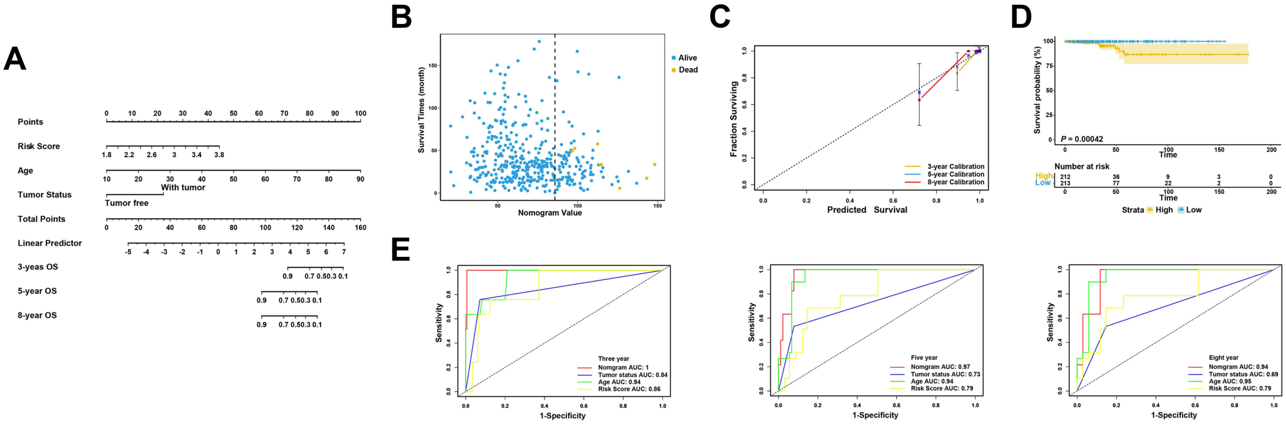


Figure 7 Construction of a nomogram for assessing survival and associations between the four-gene risk score and clinical characteristics. (**A**) The nomogram was constructed by combining the observed clinical characteristics with the four-gene risk score. (**B**) Distribution of the nomogram scores and survival data. Patients that were alive or dead at the study endpoint are represented with blue or yellow dots, respectively. (**C**) Calibration plot of the nomogram for predicting survival probabilities at 3, 5, or 8 years after treatment. (**D**) KM survival curves of the nomogram in 415 TCGA-PTC patients. (**E**) Time-dependent ROC curve of the nomogram for 3, 5, and 8-year OS predictions.

Discussion

In this study, 372 overlapping DEGs between THCA tissues and non-tumor tissues were identified from eight GEO datasets after integrating the datasets using the RRA method. Thirty-six prognosis-related genes were sifted out from TCGA datasets using univariable Cox regression, which were then subjected to Lasso regression with 10-fold cross-validation. Finally, we screened four key genes, including PAPSS2, PCOLCE2, PTX3, and TGFBR3, which are known to be involved in regulating cellular metabolism and participate in PTC development. In each case, the four-gene signature risk score was calculated based on the model described above, which was used to stratify the patients into high- or low-risk groups. The four-gene signature was identified as an independent prognostic factor of PTC, according to KM-survival curve analysis and ROC analysis. Furthermore, the four-gene signature was superior to the AJCC staging system in predicting 1-, 3-, and 5-year OS. In addition, the four-gene signature was independent of other clinical factors and performed better in predicting OS when combined with the age and tumor status in a nomogram. Collectively, these findings demonstrate that the four-gene signature can be valuable to patients with PTC and thyroid surgeons because it can help evaluate the risk for tumor-related death after surgical treatment and guide clinical treatment decisions.

Bioinformatics analysis has enabled the discovery of new tumor biomarkers, and extensive research has been conducted using microarrays and RNA-sequencing. The results of clinical practice based on gene-expression profiles have shown that high-throughput methods for identifying gene signatures might show promise in helping determine effective therapies.[20,21] Several studies have explored the prognostic roles of gene signatures in predicting THCA outcomes.[22–25] However, only a few studies focused specifically on the prognosis of PTC.[26–29] Ren et al[26] identified six genes (CTGF, CYR61, CHRDL1, OGN, FGF13, and CDH3) as a prognostic signature for PTC. Wang et al[27] identified eight candidate genes (FN1, CCND1, CDH2, CXCL12, MET, IRS1, DCN, and FMOD) as potential prognostic indicators for PTC. In our study, we identified a four genes (PAPSS2, PCOLCE2, PTX3, and TGFBR3) signature as a robust marker with great potential in risk stratification and OS prediction. The causes leading to the inconsistent results may be due to the inclusion of more gene-expression profiles than that of the previous study and the exclusion of the patients without PTC. Furthermore, another difference between these previous studies and our current study is that we analyzed correlations between the key genes and the clinical characteristics of PTC. Our four-gene signature showed higher 1-, 3-, and 5-year AUCs than the above gene signatures, indicating that our results are more robust and powerful.

The four genes identified in this study were downregulated in PTC tissues. The PAPSS2 gene encodes 3′-phosphoadenosine 5′-phosphosulfate synthase 2, which is involved in several biological processes,[30] including the sulfation conjugation of xenobiotic compounds.[31] PAPSS2 has been reported to be expressed at low levels in prostate cancer[32] and colon cancer[33] tissues. PCOLCE2 promotes the enzymatic cleavage of type-I procollagen to yield mature, structured fibrils.

Previous data showed that PCOLCE2 was mainly expressed in the adult heart and developing cartilage tissues.[34] However, Wu et al[35] reported lower PCOLCE2 expression in patients with nasopharyngeal carcinoma than in non-tumor populations. PTX3 has been reported as a key homeostatic component that functions at the crossroads between innate immunity, inflammation, tissue repair, and cancer.[36] PTX3 was identified as a component of the humoral arm of innate immunity and an extrinsic tumor-suppressor gene that tames tumor-promoting inflammation.[37] As a tumor-suppressor gene, TGFBR3 has been related to the development of various cancers. The loss of TGFBR3 expression promoted the progression of pancreatic cancer,[38] hepatocellular cancer,[39] and ovarian cancer.[40] Similar to the findings of previous studies, the decreased expression of these four genes were confirmed in our study by integrated bioinformatics analysis in tumor tissues from patients with PTC. Our results indicated that low expression of these four genes may be related to PTC development.

To understand the potential mechanisms whereby the four-gene signature affected the prognosis of patients with PTC, we analyzed the correlations between their expression levels and the associated clinical characteristics. No differences were found in their expression levels between patients with or without extrathyroid extension. However, in patients with distant metastasis, the expression levels of the four genes were significantly lower, compared with those in patients with no evidence of distant metastasis. These results indicate that these four genes might relate to the molecular mechanism of distant metastasis of PTC cells, but not local invasion. Furthermore, our results show that lower PCOLCE2 expression was associated with an earlier age of PTC onset, suggesting that the loss of PCOLCE2 gene expression might accelerate tumorigenesis. In addition, PTX3 expression were significantly lower in patients surviving with tumors than in tumor-free patients, revealing a potential correlation between PTX3 and tumor recurrence.

GO analysis showed that the four genes were significantly enriched for the "glycoprotein metabolism," "tertiary granule lumen," and "TGF-β receptor binding" terms. WikiPathways and KEGG-pathway analyses revealed that the four genes were significantly enriched for "sulfur metabolism" and "sulfation biotransformation reaction." These enriched signaling pathways serve vital catalytic roles in the development and progression of malignant disease. Changes in protein glycosylation are increasingly being recognized as important modifications associated with cancer etiology. Mammadova–Bach et al[41] reported that the genetic deficiency of platelet glycoprotein VI in mice correlated with decreased experimental and spontaneous metastasis of colon and breast cancer cells. TGF-β signaling has been associated with the progression of many cancers, such as colorectal cancer[42] and breast cancer.[43] The findings of earlier studies suggest that sulfur metabolism constituted the cellular antioxidant system, mediated intercellular and intracellular signaling, and facilitated the epigenetic regulation of gene expression, all of which can contribute to tumorigenesis.[44] These enriched pathways of the four genes in patients with PTC might provide insight into the molecular mechanisms underlying poor prognosis in the high-risk group (with low expression of the four genes).

In this study, we not only demonstrated the validity and applicability of the four-gene signature for determining the prognosis of patients with PTC through multiple bioinformatics analysis, but we also analyzed relationships between the signature and clinicopathological features. However, this study also had several limitations. First, this study was a retrospective analysis of data deposited in public databases; thus, the potentials for selection bias and confounding bias were inevitable. Second, additional in vitro and in vivo functional experiments need to be performed to further understand the biological roles of the four-gene signature in PTC. Therefore, additional work is required to further explore the potential relationship between this four-gene signature and PTC.

Conclusion

In conclusion, we identified a prognostic four-gene signature via comprehensive bioinformatics analysis with GEO and TCGA-THCA cohorts. The four-gene signature could be used to effectively stratify patients with PTC into high- and low-risk groups and independently predict their OS. Our findings indicate that the four-gene signature may help facilitate personalized cancer management in clinical settings. We therefore recommend using this classifier as a molecular diagnostic test to evaluate the prognostic risk in patients with PTC.

Data Acquisition

Data can be retrieved from the TCGA and GEO databases (GEO database https://www.ncbi.nlm.nih.gov/geo/, TCGA database https://portal.gdc.cancer.gov/). The codes used in this study can be obtained from the corresponding author upon reasonable request.

Acknowledgments

The authors would like to acknowledge the Gene Expression Omnibus database for providing raw data in this study.

Author Contributions

All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This study is supported by Educational Department Foundation of Jiangxi Province (No. GJJ180053).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Siegel RL, Miller KD, Fuchs HE, et al. Cancer statistics, 2021. *CA Cancer J Clin*. 2021;71(1):7–33. doi:10.3322/caac.21654
2. Zou M, Baitei EY, Alzahrani AS, et al. Concomitant RAS, RET/ PTC, or BRAF mutations in advanced stage of papillary thyroid carcinoma. *Thyroid*. 2014;24(8):1256–1266. doi:10.1089/thy.2013.0610
3. Lundgren CI, Hall P, Dickman PW, et al. Clinically significant prognostic factors for differentiated thyroid carcinoma: a population-based, nested case-control study. *Cancer-Am Cancer Soc*. 2006;106(3):524–531. doi:10.1002/cncr.21653
4. Grant CS. Papillary thyroid cancer: strategies for optimal individualized surgical management. *Clin Ther*. 2014;36(7):1117–1126. doi:10.1016/j.clinthera.2014.03.016
5. Nath MC, Erickson LA. Aggressive variants of papillary thyroid carcinoma: hobnail, tall cell, columnar, and solid. *Adv Anat Pathol*. 2018;25(3):172–179. doi:10.1097/PAP.0000000000000184
6. Sywak M, Pasieka JL, Ogilvie T. A review of thyroid cancer with intermediate differentiation. *J Surg Oncol*. 2004;86(1):44–54. doi:10.1002/jso.20044
7. Regalbuto C, Malandrino P, Tumminia A, et al. A diffuse sclerosing variant of papillary thyroid carcinoma: clinical and pathologic features and outcomes of 34 consecutive cases. *Thyroid*. 2011;21(4):383–389. doi:10.1089/thy.2010.0331
8. Lam AK, Lo CY. Diffuse sclerosing variant of papillary carcinoma of the thyroid: a 35-year comparative study at a single institution. *Ann Surg Oncol*. 2006;13(2):176–181. doi:10.1245/ASO.2006.03.062
9. Vuong HG, Kondo T, Pham TQ, et al. Prognostic significance of diffuse sclerosing variant papillary thyroid carcinoma: a systematic review and meta-analysis. *Eur J Endocrinol*. 2017;176(4):433–441. doi:10.1530/EJE-16-0863
10. Malandrino P, Russo M, Regalbuto C, et al. Outcome of the diffuse sclerosing variant of papillary thyroid cancer: a meta-analysis. *Thyroid*. 2016;26(9):1285–1292. doi:10.1089/thy.2016.0168
11. Axelsson TA, Hrafnkelsson J, Olafsdottir EJ, et al. Tall cell variant of papillary thyroid carcinoma: a population-based study in Iceland. *Thyroid*. 2015;25(2):216–220. doi:10.1089/thy.2014.0075
12. Liu Z, Zeng W, Chen T, et al. A comparison of the clinicopathological features and prognoses of the classical and the tall cell variant of papillary thyroid cancer: a meta-analysis. *Oncotarget*. 2017;8(4):6222–6232. doi:10.18632/oncotarget.14055
13. Longheu A, Canu GL, Cappellacci F, et al. Tall cell variant versus conventional papillary thyroid carcinoma: a retrospective analysis in 351 consecutive patients. *J Clin Med*. 2020;10(1). doi:10.3390/jcm10010070
14. Munari E, Mariotti FR, Quatrini L, et al. PD-1/PD-L1 in cancer: pathophysiological, diagnostic and therapeutic aspects. *Int J Mol Sci*. 2021;22(10). doi:10.3390/ijms22105123
15. Girolami I, Pantanowitz L, Mete O, et al. Programmed Death-Ligand 1 (PD-L1) is a potential biomarker of disease-free survival in papillary thyroid carcinoma: a systematic review and meta-analysis of PD-L1 immunoexpression in follicular epithelial derived thyroid carcinoma. *Endocr Pathol*. 2020;31(3):291–300. doi:10.1007/s12022-020-09630-5
16. Ritchie ME, Phipson B, Wu D, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43(7):e47. doi:10.1093/nar/gkv007
17. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385–395. doi:10.1002/(sici)1097-0258(19970228)16:4<385:aid-sim380>3.0.co;2-3
18. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.

19. Robin X, Turck N, Hainard A, et al. PROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinform*. 2011:1277. doi:10.1186/1471-2105-12-77
20. Kopetz S, Tabernero J, Rosenberg R, et al. Genomic classifier coloprint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *Oncologist*. 2015;20(2):127–133. doi:10.1634/theoncologist.2014-0325
21. Li X, Xie M, Yin S, et al. Identification and validation of a six immune-related genes signature for predicting prognosis in patients with stage II colorectal cancer. *Front Genet*. 2021:12666003. doi:10.3389/fgene.2021.666003
22. Shen Y, Dong S, Liu J, et al. Identification of potential biomarkers for thyroid cancer using bioinformatics strategy: a study based on GEO datasets. *Biomed Res Int*. 2020:20209710421. doi:10.1155/2020/9710421
23. Liu L, He C, Zhou Q, et al. Identification of key genes and pathways of thyroid cancer by integrated bioinformatics analysis. *J Cell Physiol*. 2019;234(12):23647–23657. doi:10.1002/jcp.28932
24. Zhang B, Chen Z, Wang Y, et al. Integrated bioinformatics analysis for the identification of key genes and signaling pathways in thyroid carcinoma. *Exp Ther Med*. 2021;21(4):298. doi:10.3892/etm.2021.9729
25. Pan Y, Wu L, He S, et al. Identification of hub genes in thyroid carcinoma to predict prognosis by integrated bioinformatics analysis. *Bioengineered*. 2021;12(1):2928–2940. doi:10.1080/21655979.2021.1940615
26. Ren H, Liu X, Li F, et al. Identification of a six gene prognosis signature for papillary thyroid cancer using multi-omics methods and bioinformatics analysis. *Front Oncol*. 2021;11:624421. doi:10.3389/fonc.2021.624421
27. Wan Y, Zhang X, Leng H, et al. Identifying hub genes of papillary thyroid carcinoma in the TCGA and GEO database using bioinformatics analysis. *Peerj*. 2020;8:e9120. doi:10.7717/peerj.9120
28. Zhang S, Wang Q, Han Q, et al. Identification and analysis of genes associated with papillary thyroid carcinoma by bioinformatics methods. *Biosci Rep*. 2019;39(4). doi:10.1042/BSR20190083
29. Liu Y, Gao S, Jin Y, et al. Bioinformatics analysis to screen key genes in papillary thyroid carcinoma. *Oncol Lett*. 2020;19(1):195–204. doi:10.3892/ol.2019.11100
30. Faiyaz UHM, King LM, Krakow D, et al. Mutations in orthologous genes in human spondyloepimetaphyseal dysplasia and the brachymorphic mouse. *Nat Genet*. 1998;20(2):157–162. doi:10.1038/2458
31. Xu ZH, Freimuth RR, Eckloff B, et al. Human 3'-phosphoadenosine 5'-phosphosulfate synthetase 2 (PAPSS2) pharmacogenetics: gene resequencing, genetic polymorphisms and functional characterization of variant allozymes. *Pharmacogenetics*. 2002;12(1):11–21. doi:10.1097/00008571-200201000-00003
32. Ibeawuchi C, Schmidt H, Voss R, et al. Exploring prostate cancer genome reveals simultaneous losses of PTEN, FAS and PAPSS2 in patients with PSA recurrence after radical prostatectomy. *Int J Mol Sci*. 2015;16(2):3856–3869. doi:10.3390/ijms16023856
33. Xu P, Xi Y, Zhu J, et al. Intestinal sulfation is essential to protect against colitis and colonic carcinogenesis. *Gastroenterology*. 2021;161 (1):271–286. doi:10.1053/j.gastro.2021.03.048
34. Steiglitz BM, Keene DR, Greenspan DS. PCOLCE2 encodes a functional procollagen C-proteinase enhancer (PCPE2) that is a collagen-binding protein differing in distribution of expression and post-translational modification from the previously described PCPE1. *J Biol Chem*. 2002;277 (51):49820–49830. doi:10.1074/jbc.M209891200
35. Wu ZH, Zhou T, Sun HY. DNA methylation-based diagnostic and prognostic biomarkers of nasopharyngeal carcinoma patients. *Medicine*. 2020;99 (24):e20682. doi:10.1097/MD.0000000000020682
36. Garlanda C, Bottazzi B, Magrini E, et al. PTX3, a humoral pattern recognition molecule, in innate immunity, tissue repair, and cancer. *Physiol Rev*. 2018;98(2):623–639. doi:10.1152/physrev.00016.2017
37. Rubino M, Kunderfranco P, Basso G, et al. Epigenetic regulation of the extrinsic oncosuppressor PTX3 gene in inflammation and cancer. *Oncoimmunology*. 2017;6(7):e1333215. doi:10.1080/2162402X.2017.1333215
38. Gordon KJ, Dong M, Chislock EM, et al. Loss of type III transforming growth factor beta receptor expression increases motility and invasiveness associated with epithelial to mesenchymal transition during pancreatic cancer progression. *Carcinogenesis*. 2008;29(2):252–262. doi:10.1093/carcin/bgm249
39. Zhang S, Sun WY, Wu JJ, et al. Decreased expression of the type III TGF-beta receptor enhances metastasis and invasion in hepatocellullar carcinoma progression. *Oncol Rep*. 2016;35(4):2373–2381. doi:10.3892/or.2016.4615
40. Hempel N, How T, Dong M, et al. Loss of betaglycan expression in ovarian cancer: role in motility and invasion. *Cancer Res*. 2007;67 (11):5231–5238. doi:10.1158/0008-5472.CAN-07-0035
41. Mammadova-Bach E, Gil-Pulido J, Sarukhanyan E, et al. Platelet glycoprotein VI promotes metastasis through interaction with cancer cell-derived galectin-3. *Blood*. 2020;135(14):1146–1160. doi:10.1182/blood.2019002649
42. Miguchi M, Hinoi T, Shimomura M, et al. Gasdermin c is upregulated by inactivation of transforming growth factor beta receptor type II in the presence of mutated apc, promoting colorectal cancer proliferation. *PLoS One*. 2016;11(11):e166422. doi:10.1371/journal.pone.0166422
43. Tu CF, Wu MY, Lin YC, et al. FUT8 promotes breast cancer cell invasiveness by remodeling TGF-beta receptor core fucosylation. *Breast Cancer Res*. 2017;19(1):111. doi:10.1186/s13058-017-0904-8
44. Ward NP, DeNicola GM. Sulfur metabolism and its contribution to malignancy. *Int Rev Cell Mol Biol*. 2019;347:39–103. doi:10.1016/bs.ircmb.2019.05.001