A Novel Metastasis-related Genes Based Signature for Predicting the Progression-free Interval of Patients With Papillary Thyroid Carcinoma

Rui Liu
Peking Union Medical College Hospital Department of General Surgery
Zhen Cao
Peking Union Medical College Hospital Department of General Surgery
Meng-wei Wu
Peking Union Medical College Hospital Department of General Surgery
Xiao-bin Li
Peking Union Medical College Hospital Department of General Surgery
Hong-wei Yuan
Peking Union Medical College Hospital Department of General Surgery
Ziwen Liu (liuziwen@pumch.cn)
Peking Union Medical College Hospital https://orcid.org/0000-0003-1574-8460

Research Article

Keywords: papillary thyroid carcinoma prognostic model, The Cancer Genome Atlas Program, metastasis-related genes, nomogram

Posted Date: January 12th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1221000/v1

Abstract

Background: We aimed to build a novel model with metastasis-related genes (MTGs) signature and relevant clinical parameters for predicting progression-free interval (PFI) after surgery for papillary thyroid carcinoma (PTC).

Methods: We performed a bioinformatic analysis of integrated PTC datasets with the MTGs to identify differentially expressed MTGs (DE-MTGs). Then we generated PFI-related DE-MTGs and established a novel MTGs based signature. After that, we validated the signature on multiple datasets and PTC cell lines. Further, we carried out uni- and multivariate analysis to identify independent prognostic characters. Finally, we established a signature and clinical parameters-based nomogram for predicting the PFI of PTC.

Results: We identified 155 DE-MTGs related to PFI in PTC. The functional enrichment analysis showed that the DE-MTGs were associated with an essential oncogenic process. Consequently, we found a novel 10-gene signature and could distinguish patients with poorer prognoses and predicted PFI accurately. The novel signature had a C-index of 0.76 and the relevant nomogram had a C-index of 0.80. Also, it was closely related to pivotal clinical characters of datasets and invasiveness of cell lines. And the signature was confirmed a significant independent prognostic factor in PTC. Finally, we built a nomogram by including the signature and relevant clinical factors. Validation analysis showed that the nomogram's efficacy was satisfying in predicting PTC's PFI.

Conclusions: The MTG signature and nomogram were closely associated with PTC prognosis and may help clinicians improve the individualized prediction of PFI, especially for high-risk patients after surgery.

Background

Thyroid cancer (TC) has become the most commonly diagnosed endocrine tumor over the past decades[1]. Should the recent trends of TC prevail, it may become the fourth most common cancer in the United States by 2030[2]. The most common and least aggressive histologic type of TC is papillary TC (PTC), comprising 80% of all cases. PTC is characterized by a favorable outcome after adequate surgical removal of the primary tumor and clinically significant lymph nodes[3]. However, one of the primary concerns after the initial surgery is a recurrent disease, which is found to be 5.7% at five years and 9.4% at ten years, as Karl et al. reported in 52,173 PTC surgery patients[4]. Re-operations for the recurrent disease could result in a higher risk of surgical complications[5]. Clinical predictive models such as the American Thyroid Association (ATA) risk stratification have been widely used. However, the clinical and pathological character-based models developed thus far do not reflect individual characteristics at the molecular level. Therefore, novel prognostic tools for guiding personalized surveillance, especially for patients with a high risk of recurrence, are urgently needed. The development of a predictive model based on sensitive biomarkers would facilitate personalized monitoring, which would reduce the possibility of advanced, recurrent diseases in the postoperative follow-up period. Recently, progression in high-

throughput sequencing has led to optimistic expectations about personalized medicine. Signatures based on biomarkers such as mRNA or lncRNA have great potential to predict cancer prognosis[6, 7]. These omics-based models can also reliably predict the prognosis of PTC[8, 9].

Lymph node metastasis (LNM) is one of the significant causes of cancer recurrence[10]. Cells progressing through the metastatic cascade must employ a series of diverse cellular processes[11]. Local invasion and intravasation into the bloodstream, survival in the circulation, and growth in new organ environments are primary features of metastasis[12]. TC frequently metastasizes to the lymph nodes of the neck, and the rate of occult nodal metastases in PTC has been reported to be as high as 60-80%[13, 14]. In 60% to 75% of cases of TC recurrence, the phenomenon occurs in the cervical lymph nodes[15]. Hence, metastasis-related genes (MTGs) based predictive models may be closely related to the metastasis of PTC. Therefore, we previously searched the Human Cancer Metastasis Database (HCMDB), which curates 2183 potential MTGs based on more than 7,000 published pieces of literature[16]. We analyzed four datasets of PTC from Gene Expression Omnibus (GEO) and differentially expressed genes between PTC and normal samples. Then we identified differentially expressed MTGs (DE-MTGs) after the intersection with the experimentally supported MTGs derived from HCMDB. Finally, we proposed a novel 10-gene signature and constructed a nomogram with relevant clinical factors involved.

Methods

Obtain of TCGA-THCA RNA sequencing data and clinical information

We used GDC API to download RNA sequencing data from The Cancer Genome Atlas Thyroid carcinoma (TCGA-THCA) up to 21 Jul 2019, including 507 PTC cases and relevant follow-up information. Transcript per million (TPM) transformation followed by base-2 logarithm normalization was applied. Cases with a follow-up period of less than a month were excluded. We extracted PFI data from University of California Santa Cruz (UCSC) Xena database. Both structural evidence (includes distant metastasis, locoregional recurrence, and new primary tumor) and biochemical evidence of recurrence was defined as progression. We also retrieved clinical and mutational data from the Cbioportal.

Integrated analysis and identification of DE-MTGs

We searched the GEO database to identify DEGs to obtain PTC datasets. The keywords for the search included "Thyroid cancer," "Homo. sapiens," and "Thyroid carcinoma." Only datasets including PTC samples and normal thyroid samples based on the Affymetrix GPL570 Gene Chip (Santa Clara, U.S.) were included. The research focused on "cell lines," "xenografts," "poorly differentiated," and "undifferentiated" was excluded. Cases of childhood PTC, PTC in young adults, and radiation-induced PTC were also excluded. Raw data were normalized using the RMAExpress software[17]. Probe names were transformed to official symbols based on Thermo Fisher Scientific Inc provided annotation file. If more than

one probes to a single gene symbol, then the median value was replaced. DEG lists of four datasets were identified independently with the R package “Limma” with $p < 0.05$, false discovery rate <0.05, and |Log2FC| >1[18]. Reliable DEGs were then identified from
the combination of differential analysis results from four datasets with the "RobustRankAggreg" package of the R software[19]. MTGs were downloaded from the HCMDB. After the intersection with the reliable DEGs, DE-MTGs were generated.

Functional enrichment analysis

We carried out functional enrichment analyses using the "clusterProfiler" package of R to explore the potential enriched function of the DE-MTGs[20, 21]. The Benjamini and Hochberg method was used for FDR correction, define adjusted $p < 0.05$ as statistically significant.

Construction and verification of the novel MTGs based signature

The TCGA-THCA dataset was randomly divided into training and testing datasets in the ratio of 0.8. We used the univariate Cox regression model to identify the DE-MTGs that were significantly associated ($p < 0.05$) with PFI in the training set. The PFI-related DE-MTGs were further included. Then we applied the Least absolute shrinkage and selection operator (LASSO) analysis which is often used in in high-dimensional data to reduce the dimension by penalizing the number of regression coefficients, to further select valid variables using the "glmnet" R package[22]. The “cv.glmnet” function of the package is used to build model. Cross validation used different lambda values to observe the model error. Then cv plot was generated and the best lambda value was chosen. Then a panel of gene signature was found. Based on the median value of risk-scores, the patients were then defined as low- or high-risk. The predictive efficacy of the signature was then assessed with ROC curve, K-M analysis, and C-index by the "timeROC" package and the "survcomp" package of the R software[23].

Gene set enrichment analysis (GSEA) of the 10-gene signature

We explored the potential molecular alterations of the signature by GSEA[24]. 488 PTC samples from the TCGA-THCA dataset were defined as low- or high-risk by the optimal cut-off value generated by X-Tile[25]. GSEA v4.2 has then applied to find the biological alteration in the high-risk group. The gene sets included C2: KEGG[24], C5: GO, and C6: oncogenic signatures. FDR < 0.05 with |NES| > 1 were considered to indicate significant enrichment.

Cell lines culture and lysis

Human PTC cell lines B-CPAP[26] and KTC-1[27] were kindly provided by the National Collection of Authenticated Cell Cultures of the Chinese Academy of Sciences. B-CPAP and KTC-1 were cultivated in RPMI Medium 1640 (Invitrogen, U.S.) with 10% FBS (Gibco, U.S.), Non-essential Amino Acids, Glutamax, and Sodium Pyruvate (Invitrogen, U.S.) added. We used Trizol (Lablead, China) to lysate cells in logarithmic growth phase and isolate RNA following the manufacturer's protocol.

Quantitative real-time Polymerase Chain Reaction (RT-qPCR)

After removing genomic DNA contamination and reverse transcription following the manufacturer's instruction, we conducted RT-qPCR after preparation of cDNA, as described previously with GAPDH mRNA for normalization via the $2^{-\Delta\Delta Ct}$ method[28]. Reagents for removing genomic DNA contamination, reverse transcription and RT-qPCR were purchased from Lablead (Beijing, China). Primers were designed and purified by Sangon (Shanghai, China), sequences as shown in Table 5. Each experiment was repeated four times.

External validation of 10-gene signature in GEO datasets

The expression pattern of MTG-based gene signature from six datasets (GSE29265, GSE33630, GSE76039[29], GSE82208, GSE58545[30] and GSE5364[31]) including PTC, follicular thyroid cancer (FTC), anaplastic thyroid carcinoma (ATC) and poorly differentiated thyroid carcinoma (PDTC) samples were extracted. Each sample's risk-score was generated to evaluate potential clinical utility of the 10-gene signature. P-value of <0.05 as statistically significant. All the GSE datasets were obtained in Gene Expression Omnibus (GEO), the details were listed in Table 1.

Independent prognostic parameters in PTC

We performed uni- and multivariate Cox analyses to find the correlated prognostic parameters in PTC. Clinical parameters included age, gender, BRAFV600E mutation, RAS mutation, extrathyroidal extension, neoplasm size, histological type, anatomic sites of tumors, residual tumor and disease TNM stage. The univariate analysis was performed first, then the factors with $p < 0.2$ were enrolled in multivariate analysis to identify independent ones. A p-value of <0.05 as statistically significant.

Construction of the novel nomogram

After the collinearity diagnosis, a novel stepwise Cox regression model incorporating independent and relevant clinical factors was built and visualized as a nomogram for predicting the 1-, 2-, and 3-year PFI of PTC. The length of each parameter stands for its' weight in regression model. We then evaluated the nomogram's predictive power with the ROC curve, C-index, calibration curve and decision curve analysis (DCA)[32]. The calibration curve was generated by a bootstrap method with 1000 resamples.

Statistical analysis

We used R v3.6.3 and GraphPad Prism 8.4.3 (GraphPad Software, U.S.) for statistical analysis. Comparison of survival curves was analyzed with Log-rank (Mantel-Cox) test. Categorical variables were analyzed using Chi-squared tests. Continuous data were analyzed using unpaired t-tests. A p-value of <0.05 was considered as statistically significant.

Results

Identification of reliable DE-MTGs

We conducted the research according to the flowchart shown in Figure 1. Eventually, four independent datasets (GSE29265, GSE33630[33], GSE35570[34], and GSE60542[35]) with 134 PTC tumor samples and 146 normal thyroid samples were enrolled. Especially for dataset GSE35570, PTC samples derived from Chernobyl radiation-exposed patients were excluded. In all, 587, 851, 1716, and 777 DEGs were classified from GSE29265, GSE33630, GSE35570, and GSE60542, respectively PTC versus normal thyroid samples. 702 DEGs, including 349 up- and 353 down-regulated, were identified with the RRA (Supplementary Table 1). The information of GEO datasets was listed in Table 1. The top 20 upregulated and downregulated DEGs as shown in Figure 2A. We downloaded a list including 1938 experimentally supported MTGs from HCMDB (Supplementary Table 2) to intersect with DEGs. Finally, 155 reliable DE-MTGs were identified, among which 98 were upregulated, and 57 were down-regulated (Figure 2B, Supplementary Table 3).

Functional enrichment analysis

We analyzed the potential function and pathway enrichment of the 155 DE-MTGs (Figure 3A-D). In terms of BPs, the 155 DE-MTGs were mainly enriched in the cellular matrix organization, extracellular structure organization, and cell-substrate adhesion (Figure 3A). In terms of CCs, the DE-MTGs identified were significantly enriched in the extracellular matrix, focal adhesion, and membrane raft (Figure 3B). In terms of MFs, the DE-MTGs identified were significantly enriched in receptor-ligand activity, signaling receptor activity, and so on. Pathway analysis further revealed that the 155 DE-MTGs were mainly enriched in the proteoglycans in cancer, PI3K-Akt signaling pathways, and so on (Figure 3D).

Identification and establishment of a 10-gene signature

After excluding cases in which PFI was $\leq$ 30 days, 488 PTC samples with complete follow-up information were finally included in the analysis. The baseline clinical characteristics are as shown in Table 2. Twenty-eight PFI-related DE-MTGs were identified in training set using the Cox proportional-hazards model. Forest plots of the logfc, P-value and hazard ratio of each item are shown in Figure 4A. The potential protein-protein interaction network of the 28 items was explored using the STRING

database as shown in Figure 4B[36]. A novel gene signature consisting of the following 10 DE-MTGs was constructed and listed in Table 4. The formula to calculate the risk score was as follows: β1 × gene-one's expression + β2 × gene-two's expression + ····· + βn × gene-N's expression, where β is the corresponding correlation coefficient. The patients with the high-risk have shorter PFIs by using the Log-rank analysis with the training, testing, and total ($P < 0.05$; Figure 5 A-C). The correlations between risk scores and recurrences are presented in Figure 5 D-F. In the training set, the AUCs for PFI prediction was 0.799 (1-year), 0.781 (2-year), and 0.737 (3-year), respectively (Figure 5G). The C-index was 0.752. In the testing set, the AUCs were 0.850, 0.750, and 0.649 (Figure 5H). The C-index was 0.750. In the total set, AUCs were 0.812, 0.772, and 0.717 (Figure 5I). The C-index was 0.748. After optimizing the training set by enrolling all the 488 cases, thirty-one PFI-related DE-MTGs were identified, then the optimized 10-gene signature was generated as shown in and Figure 6 and Table 4. In the total set, the AUCs for PFI prediction was 0.836 (1-year), 0.775 (2-year), and 0.736 (3-year), respectively (Figure 6D). The C-index was 0.756. Collectively, our results indicated that the 10-gene signature functions well in the PFI forecast for PTC.

GSEA

To seek the potential alteration underlying the 10-gene signature, we conducted GSEA in PTC from the TCGA-THCA (Figure 7A–C). For KEGG pathways, the molecular alterations in the high-risk group samples were related to the homologous recombination, cell cycle, and DNA replication. For the oncological signatures, including the AKT_UP.V1_DN, MTOR_UP.VI_DN, and CYCLIN_D1_UP.VI_UP, were related. GSEA results were presented in Supplementary Table 4.

RT-qPCR analysis of MTGs based signatures in PTC cell lines

The relative 10 gene expression level of MTGs signature in B-CPAP and KTC-1 cell were generated through RT-qPCR quantification. In 10 genes of MTGs signature, the expression level of KISS1R, EZH2, and FAM3B were higher in KTC-1 than B-CPAP, and the expression level of FBLN5 and SDPR were higher in B-CPAP than KTC-1, the differences were statistically significant ($P<0.05$). The differences of TGFBR3, ARHGDIB, SDC2, SOD3 and CCL14 were not statistically significant ($P>0.05$), as shown in Figure 8A. The MTGs signature scores of KTC-1 were higher than B-CPAP, the difference was statistically significant ($P<0.05$), as shown in Figure 8B.

Clinical correlation of the MTGs based signature

In groups divided by a series of pivotal clinical and pathological characters, patients who were with mutated BRAF status, advanced extra-invasion existence, advanced T stage had higher risk-scores, as shown in Figure 9B-E. Patients who were with lymph nodes metastasis, residual tumor, uni-focality, tall cell histological type and advanced disease stages also had higher risk-scores, as shown in Figure 9F-J, the differences were statistically significant ($P<0.05$).

Verification of the novel 10-gene signature in external datasets

We validated the pattern of 10-gene signature in six external GEO datasets and compared the risk-scores between ATC/PDTC/PTC/FTC. For 10-gene signature, in the datasets GSE 29265 and GSE 33630, risk-scores were higher in ATC samples compared to PTC samples ($p < 0.01$, 0.0001, respectively), as shown in Figure 10A, B. In the dataset GSE 76039, risk-scores were higher in ATC samples than PDTC samples ($p < 0.01$), as shown in Figure 10C. In GSE82208, the difference of risk-scores between follicular adenoma and FTC was statistically significant ($P<0.0001$), as shown in Figure 10D. Also, in GSE58545 and GSE5364, the difference of risk-scores between normal thyroid samples and PTC samples was statistically significant ($P<0.0001$), as shown in Figure 10E, F. The AUCs of diagnostic ROC for differentiating sample's type were also presented.

Prognosis-associated factors of the PFI in PTC

488 TCGA PTC Cases were enrolled to identify prognostic factors. Univariate analysis showed that age, T stage, M stage, AJCC stage, neoplasm size, histological type and risk-scores were significantly PFI-related ($p < 0.05$) (as shown in Table 3). A total of 408 cases with complete information were then enrolled in multivariate analysis, and the risk-score ($p < 0.001$) was the only independent factor for prognosis, as shown in Table 3.

Construction and validation of the predictive nomogram

A nomogram for predicting the AUCs of PTC was built with a stepwise Cox regression model (Figure 11A). Parameters including risk-score, RAS mutation, neoplasm size, residual tumor, age, TNM stage and histological type were incorporated in the nomogram. For the histological type, the tall cell variant was defined as aggressive[37]. The calibration curve and DCA showed that the nomogram functions well in predicting the PFIs (Figure 11B, C). The AUCs for 1-, 2-, and 3-year PFIs were 0.940, 0.793 and 0.779, respectively (Figure 11D).

Discussion

Most patients with PTC achieve a relatively good prognosis. However, persistent disease or recurrences are observed in 5%-20% of patients, associated with severe complications following re-operation or other therapies[38]. For patients with a low risk of recurrence, prolonged thyroid-stimulating hormone suppression therapy may cause multiple adverse effects such as osteoporosis or osteopenia and cardiac comorbidities like atrial fibrillation[39]. Considering the relatively excellent prognosis, the development of novel diagnostic tools with high sensitivity and specificity seems to have greater clinical significance than the exploration of neoadjuvant therapies. Traditional staging systems such as the ATA risk

stratification system allow evaluation of recurrence risk with a stratified population rather than individualized risk, which indicates that a group of patients sharing the same clinical and pathological characteristics would have the same chance of recurrence[40]. However, the biological mechanisms underlying PTC progression are highly complex and heterogeneous and require a more accurate and personalized prediction model based on biomarkers at the molecular level. Therefore, specified gene signatures would predict the metastatic and recurrent potential of tumors effectively.

The incidence of PTCs has been continually increasing; however, the mortality rate has not changed substantially, which is may because most PTCs diagnosed incidentally are low-risk papillary thyroid microcarcinomas (PTMCs). Except for tumors with high-risk features such as extrathyroidal extension, clinically evident LNM(+), and particular aggressive types, active surveillance appears to be safe[41]. It can replace immediate surgery for low-risk PTC[42]. In general, active surveillance begins when patients are diagnosed with low-risk PTC by ultrasound examination of fine-needle aspiration biopsy (FNAB). Since PTCs involve complex biological mechanisms, the decision to perform active surveillance based on signatures followed by FNAB with micro-assessing technique such as droplet digital PCR[43] (ddPCR) would be safer than assessments based on superficial clinical or imaging characteristics. Therefore, patients with a higher risk-score but with a low risk of clinical features would be treated more rationally.

PTC patients with cervical LNM are usually at a high risk of recurrence and have a poorer prognosis as PTC incidence has increased rapidly in recent years[44]. Therefore, LNM is a significant reason for locally advanced and recurrent diseases, which motivated us to focus on differentially expressed MTGs derived from HCMDB, which annotated about 2,000 potential MTGs based on more than 7,000 published pieces of literature. We identified 155, then reduced the variables to 31 DE-MTGs that were PFI-related to PTC. GO enrichment analysis showed that DE-MTGs were enriched in cell adhesion, cellular matrix organization, and cell-substrate junction, consistent with the definition of MTGs, which have been proven to be associated with cancer metastasis. A novel 10-gene signature was then established and proved to be an independent prognostic factor of PTC. The high-risk patients were with a significantly shorter PFI than those with a low-risk scores. Among the 10 genes, EZH2, KISS1R were upregulated and associated with shorter PFI (HR>1), whereas ARHGDIB, CCL14, FAM3B, FBLN5, SDC2, SDPR, SOD3 and TGFBR3 were downregulated and associated with better PFI (HR<1). In the identified 10 genes, several were previously proved to be associated with PTC progression through experiments. For example, SDPR functions as a tumor suppressor gene in papillary thyroid cancer[45], stromal SOD3 had a stimulatory effect on thyroid cancer cell growth and an inhibitory effect on cancer cell migration[46]. Thus, these genes have the potential to predict metastasis and recurrence in PTC.

Besides, we explored the potential molecular alteration by the 10-gene signature using GSEA. GSEA, which is based on careful consideration of all differential genes' roles, can help reveal the complex behavior of genes in health and disease more accurately. In contrast, traditional strategies, including KEGG or GO, focus on identifying individual genes that exhibit differences[47]. Multiple gene expression alterations in the high-risk group were involved in tumor biology processes, such as homologous recombination, and cell cycle pathways. Thus, the potential mechanisms underlying patients' poor PFI in

the high-risk group could be elucidated. Further, we validated the MTGs signature panel on two PTC cell lines with different genetic background. The B-CPAP which originated from primary low risk PTC, grows slow in cultivation and with a relatively low invasiveness. However, the KTC-1 which originated from advanced metastatic PTC and refractory to radio iodine, grows rapidly and with a relatively higher invasiveness. The higher expressed genes and risk-score of KTC-1 than B-CPAP also partly addressed the MTGs signature's efficacy in predicting PTC invasiveness. In the aspects of the clinical correlation, we found that patients with advanced or worse clinical status such as mutated BRAF, advanced extra-invasion existence, advanced TNM stages, residual tumor, tall cell histological type had higher risk-scores, which strongly demonstrated the clinical efficacy of the MTGs signature. To our knowledge, patients with ATC and PDTC only have a mean survival after diagnosis of 0.5 and 3.2 years, respectively, and de-differentiation is a significant reason for the highly malignant degree[29]. The results of significantly higher gene risk scores in ATC samples could partly confirm our conjecture. Further, we found that the risk-scores significantly differs between follicular adenoma from FTC, revealing a pivotal clinical value in FNA diagnosis which remains challenging in differentiating FTC with benign follicular thyroid adenoma[48].

Nomograms are widely used since the ability to present the numerical probability of a particular clinical event by integrating prognostic variables[49]. Nomograms including a risk score based on gene signatures and clinicopathological parameters can predict prognosis more precisely after surgery. Moreover, numerical results are more comfortable for patients to understand than the traditional staging system. To the best our knowledge, the MTGs based signature and the relevant nomogram achieved the highest AUC for predicting PFI in TCGA-THCA cohort ever and has not been reported yet. The limited number of genes made it practical and economically feasible than whole-genome sequencing. Moreover, since the significantly worse clinical outcomes with PDTC/ATC than DTC, the unique advantage in discovering the de-differentiation potential of TC made the 10-gene signature feasible in individualized follow-up.

There were limitations to our study. First, the primary source of RNA sequencing data and clinical information was the TCGA program, in which the source of samples was from North American people. When applying the model to patients from different countries or regions, possible deviations or biases would occur. Second, due to the lack of a large independent dataset of PTC with complete follow-up information, we validated the nomogram's power on the TCGA dataset itself, then we carried out experimental and GEO datasets validation. Future validation of external datasets with complete follow-up data is necessary. Finally, we didn't compare our classifier and nomogram with the latest ATA risk stratification since the available information in TCGA-THCA was evaluated based on the 2009 version. Further comparison is required to validate the nomogram's efficacy with the newest ATA risk system.

Conclusion

We built a novel 10-gene signature, then established a nomogram combining the signature and relevant clinical and pathological factors for predicting PTC PFI. The efficacy of novel MTGs signature and

relevant nomogram was satisfying. It may be helpful for individualized active and postoperative surveillance strategies.

Abbreviations

AUC, area under the curve; PTC, papillary thyroid carcinoma; ATC, anaplastic thyroid cancer; PDTC, poorly differentiated thyroid cancer; DE-MTGs, differentially expressed MTGs; PFI, progression-free interval; HCMDB, Human Cancer Metastasis Database; ROC, receiver operating characteristic; C-index, concordance index; GO, Gene Ontology; BPs, biological processes; CCs, cellular components; MFs, molecular functions; KEGG, Kyoto Encyclopedia of Genes and Genomes; ATA, American Thyroid Association; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas Program; EMT, epithelial-to-mesenchymal transition; RRA, robust rank aggregation; droplet digital PCR, ddPCR; American Joint Committee on Cancer; ANOVA, one-way analysis of variance, Least absolute shrinkage and selection operator, LASSO.

Declarations

Acknowledgments

We thank Dr.Wei Ge for her helpful suggestions on the manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

None.

Availability of data and materials

We obtained all the datasets from the TCGA (https://portal.gdc.cancer.gov/), the UCSCxena (https://xenabrowser.net/datapages/), and the Cbioportal database (http://www.cbioportal.org/). The databases are open accessed and available for the public.

Funding

This research was supported by the National Natural Science Foundation of China [grant number: 82172727], Nature Science Foundation of Beijing [grant number: 7202164], and CAMS Innovation Fund for Medical Sciences (CIFMS) [grant number: 2021-I2M-1-002].

Authors' contributions

RL and ZC designed the study and obtained the data; RL and ZC carried out experiment and analyzed data: RL and MW wrote the manuscript; HY and ZL revised and approved the manuscript.

References

1. C. La Vecchia, M. Malvezzi, C. Bosetti, W. Garavello, P. Bertuccio, F. Levi, E. Negri, Thyroid cancer mortality and incidence: a global overview. Int J Cancer **136**, 2187–2195 (2015). https://doi.org/10.1002/ijc.29251
2. L. Rahib, B.D. Smith, R. Aizenberg, A.B. Rosenzweig, J.M. Fleshman, L.M. Matrisian (2014) Projecting Cancer Incidence and Deaths to 2030: The Unexpected Burden of Thyroid, Liver, and Pancreas Cancers in the United States. Cancer Res 74:2913–2921. https://doi.org/10/f56ftb
3. H. Lim, S.S. Devesa, J.A. Sosa, D. Check, C.M. Kitahara, Trends in Thyroid Cancer Incidence and Mortality in the United States, 1974-2013. Jama-J Am Med Assoc **317**, 1338–1348 (2017). https://doi.org/10.1001/jama.2017.2719
4. K.Y. Bilimoria, D.J. Bentrem, C.Y. Ko, A.K. Stewart, D.P. Winchester, M.S. Talamonti, C. Sturgeon, Extent of surgery affects survival for papillary thyroid cancer. Ann. Surg. **246**, 375–384 (2007). https://doi.org/10.1097/SLA.0b013e31814697d9
5. B.R. Haugen, E.K. Alexander, K.C. Bible, G.M. Doherty, S.J. Mandel, Y.E. Nikiforov, F. Pacini, G.W. Randolph, A.M. Sawka, M. Schlumberger, K.G. Schuff, S.I. Sherman, J.A. Sosa, D.L. Steward, R.M. Tuttle, L. Wartofsky, 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid Off J Am Thyroid Assoc **26**, 1–133 (2016). https://doi.org/10.1089/thy.2015.0020
6. M. Wu, X. Li, T. Zhang, Z. Liu, Y. Zhao, Identification of a Nine-Gene Signature and Establishment of a Prognostic Nomogram Predicting Overall Survival of Pancreatic Cancer. Front Oncol **9**, 996 (2019). https://doi.org/10.3389/fonc.2019.00996
7. Y. Du, Y. Gao, Development and validation of a novel pseudogene pair-based prognostic signature for prediction of overall survival in patients with hepatocellular carcinoma. BMC Cancer **20**, 887 (2020). https://doi.org/10.1186/s12885-020-07391-2
8. M. Wu, H. Yuan, X. Li, Q. Liao, Z. Liu (2019) Identification of a Five-Gene Signature and Establishment of a Prognostic Nomogram to Predict Progression-Free Interval of Papillary Thyroid Carcinoma. Front Endocrinol 10:. https://doi.org/10.3389/fendo.2019.00790
9. P. Lin, Y. Guo, L. Shi, X. Li, H. Yang, Y. He, Q. Li, Y. Dang, K. Wei, G. Chen, Development of a prognostic index based on an immunogenomic landscape analysis of papillary thyroid cancer. Aging **11**, 480–500 (2019). https://doi.org/10.18632/aging.101754
10. P.S. Steeg, Targeting metastasis. Nat. Rev. Cancer **16**, 201–218 (2016). https://doi.org/10.1038/nrc.2016.25

11. K. Karamanou, M. Franchi, D. Vynios, S. Brézillon, Epithelial-to-mesenchymal transition and invadopodia markers in breast cancer: Lumican a key regulator. Semin Cancer Biol **62**, 125–133 (2020). https://doi.org/10.1016/j.semcancer.2019.08.003
12. J. Ko, M.M. Winslow, J. Sage, Mechanisms of small cell lung cancer metastasis. EMBO Mol Med e13122 (2020). https://doi.org/10.15252/emmm.202013122
13. N. Wada, Q.-Y. Duh, K. Sugino, H. Iwasaki, K. Kameyama, T. Mimura, K. Ito, H. Takami, Y. Takanashi, Lymph Node Metastasis From 259 Papillary Thyroid Microcarcinomas. Ann. Surg. **237**, 399–407 (2003). https://doi.org/10.1097/01.SLA.0000055273.58908.19
14. P. Asimakopoulos, A.R. Shaha, I.J. Nixon, J.P. Shah, G.W. Randolph, P. Angelos, M.E. Zafereo, L.P. Kowalski, D.M. Hartl, K.D. Olsen, J.P. Rodrigo, V. Vander Poorten, A.A. Mäkitie, A. Sanabria, C. Suárez, M. Quer, F.J. Civantos, K.T. Robbins, O. Guntinas-Lichius, M. Hamoir, A. Rinaldo, A. Ferlito, Management of the Neck in Well-Differentiated Thyroid Cancer. Curr Oncol Rep **23**, 1 (2020). https://doi.org/10.1007/s11912-020-00997-6
15. J.C. Watkinson, J.A. Franklyn, J.F.C. Olliff, Detection and surgical treatment of cervical lymph nodes in differentiated thyroid cancer. Thyroid Off J Am Thyroid Assoc **16**, 187–194 (2006). https://doi.org/10.1089/thy.2006.16.187
16. G. Zheng, Y. Ma, Y. Zou, A. Yin, W. Li, D. Dong, HCMDB: the human cancer metastasis database. Nucleic Acids Res **46**, D950–D955 (2018). https://doi.org/10.1093/nar/gkx1008
17. B.M. Bolstad, R.A. Irizarry, M. Astrand, T.P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinforma Oxf Engl **19**, 185–193 (2003). https://doi.org/10.1093/bioinformatics/19.2.185
18. M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res **43**, e47 (2015). https://doi.org/10.1093/nar/gkv007
19. R. Kolde, S. Laur, P. Adler, J. Vilo, Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics **28**, 573–580 (2012). https://doi.org/10.1093/bioinformatics/btr709
20. G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters. Omics J Integr Biol **16**, 284–287 (2012). https://doi.org/10.1089/omi.2011.0118
21. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28**, 27–30 (2000). https://doi.org/10.1093/nar/28.1.27
22. J. Friedman, T. Hastie, R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. **33**, 1–22 (2010)
23. M.S. Schröder, A.C. Culhane, J. Quackenbush, B. Haibe-Kains, survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. Bioinformatics **27**, 3206–3208 (2011). https://doi.org/10.1093/bioinformatics/btr511
24. A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, J.P. Mesirov, Gene set enrichment analysis: A knowledge-based

approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A **102**, 15545–15550 (2005). https://doi.org/10.1073/pnas.0506580102

25. R.L. Camp, M. Dolled-Filhart, D.L. Rimm, X-tile: a new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res Off J Am Assoc Cancer Res **10**, 7252–7259 (2004). https://doi.org/10.1158/1078-0432.CCR-04-0713
26. F. Arturi, D. Russo, J.M. Bidart, D. Scarpelli, M. Schlumberger, S. Filetti (2001) Expression pattern of the pendrin and sodium/iodide symporter genes in human thyroid carcinoma cell lines and human thyroid tumors. Eur J Endocrinol 145:129–135. https://doi.org/10/dt6tmf
27. J. Kurebayashi, K. Tanaka, T. Otsuki, T. Moriya, H. Kunisue, M. Uno, H. Sonoo, All-trans-retinoic acid modulates expression levels of thyroglobulin and cytokines in a new human poorly differentiated papillary thyroid carcinoma cell line, KTC-1. J Clin Endocrinol Metab **85**, 2889–2896 (2000). https://doi.org/10.1210/jcem.85.8.6732
28. Z. Zhang, H.-Y. Zhang, Y. Zhang, H. Li (2019) Inactivation of the Ras/MAPK/PPARγ signaling axis alleviates diabetic mellitus-induced erectile dysfunction through suppression of corpus cavernosal endothelial cell apoptosis by inhibiting HMGCS2 expression. Endocrine 63:615–631. https://doi.org/10/gnxhzm
29. I. Landa, T. Ibrahimpasic, L. Boucai, R. Sinha, J.A. Knauf, R.H. Shah, S. Dogan, J.C. Ricarte-Filho, G.P. Krishnamoorthy, B. Xu, N. Schultz, M.F. Berger, C. Sander, B.S. Taylor, R. Ghossein, I. Ganly, J.A. Fagin, Genomic and transcriptomic hallmarks of poorly differentiated and anaplastic thyroid cancers. J. Clin. Invest. **126**, 1052–1066 (2016). https://doi.org/10.1172/JCI85271
30. D. Rusinek, M. Swierniak, E. Chmielik, M. Kowal, M. Kowalska, R. Cyplinska, A. Czarniecka, W. Piglowski, J. Korfanty, M. Chekan, J. Krajewska, S. Szpak-Ulczok, M. Jarzab, W. Widlak, B. Jarzab, BRAFV600E-Associated Gene Expression Profile: Early Changes in the Transcriptome, Based on a Transgenic Mouse Model of Papillary Thyroid Carcinoma. PloS One **10**, e0143688 (2015). https://doi.org/10.1371/journal.pone.0143688
31. K. Yu, K. Ganesan, L.K. Tan, M. Laban, J. Wu, X.D. Zhao, H. Li, C.H.W. Leung, Y. Zhu, C.L. Wei, S.C. Hooi, L. Miller, P. Tan (2008) A Precisely Regulated Gene Expression Cassette Potently Modulates Metastasis and Survival in Multiple Solid Cancers. PLoS Genet 4:e1000129. https://doi.org/10/fg8ncf
32. A.J. Vickers, E.B. Elkin (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Mak Int J Soc Med Decis Mak 26:565–574. https://doi.org/10/cgv59z
33. G. Dom, M. Tarabichi, K. Unger, G. Thomas, M. Oczko-Wojciechowska, T. Bogdanova, B. Jarzab, J.E. Dumont, V. Detours, C. Maenhaut, A gene expression signature distinguishes normal tissues of sporadic and radiation-induced papillary thyroid carcinomas. Br J Cancer **107**, 994–1000 (2012). https://doi.org/10.1038/bjc.2012.302
34. D. Handkiewicz-Junak, M. Swierniak, D. Rusinek, M. Oczko-Wojciechowska, G. Dom, C. Maenhaut, K. Unger, V. Detours, T. Bogdanova, G. Thomas, I. Likhtarov, R. Jaksik, M. Kowalska, E. Chmielik, M.

Jarzab, A. Swierniak, B. Jarzab, Gene signature of the post-Chernobyl papillary thyroid cancer. Eur J Nucl Med Mol Imaging **43**, 1267–1277 (2016). https://doi.org/10.1007/s00259-015-3303-3

35. M. Tarabichi, M. Saiselet, C. Trésallet, C. Hoang, D. Larsimont, G. Andry, C. Maenhaut, V. Detours, Revisiting the transcriptional analysis of primary tumours and associated nodal metastases with enhanced biological and statistical controls: application to thyroid cancer. Br J Cancer **112**, 1665–1674 (2015). https://doi.org/10.1038/bjc.2014.665
36. D. Szklarczyk, A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N.T. Doncheva, J.H. Morris, P. Bork, L.J. Jensen, C. von Mering, STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res **47**, D607–D613 (2019). https://doi.org/10.1093/nar/gky1131
37. S. Cartwright, A. Fingeret, Contemporary evaluation and management of tall cell variant of papillary thyroid carcinoma. Curr. Opin. Endocrinol. Diabetes Obes. **27**, 351–357 (2020). https://doi.org/10.1097/MED.0000000000000559
38. H. Wong, K.P. Wong, T. Yau, V. Tang, R. Leung, J. Chiu, B.H.H. Lang, Is there a role for unstimulated thyroglobulin velocity in predicting recurrence in papillary thyroid carcinoma patients with detectable thyroglobulin after radioiodine ablation? Ann Surg Oncol **19**, 3479–3485 (2012). https://doi.org/10.1245/s10434-012-2391-6
39. B. Schmidbauer, K. Menhart, D. Hellwig, J. Grosse (2017) Differentiated Thyroid Cancer-Treatment: State of the Art. Int. J. Mol. Sci. 18:. https://doi.org/10.3390/ijms18061292
40. D.S. Cooper, G.M. Doherty, B.R. Haugen, R.T. Kloos, S.L. Lee, S.J. Mandel, E.L. Mazzaferri, B. McIver, F. Pacini, M. Schlumberger, S.I. Sherman, D.L. Steward, R.M. Tuttle, Revised American Thyroid Association Management Guidelines for Patients with Thyroid Nodules and Differentiated Thyroid Cancer. Thyroid **19**, 1167–1214 (2009). https://doi.org/10.1089/thy.2009.0110
41. Y. Ito, A. Miyauchi, H. Oda, Low-risk papillary microcarcinoma of the thyroid: A review of active surveillance trials. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol **44**, 307–315 (2018). https://doi.org/10.1016/j.ejso.2017.03.004
42. B. Saravana-Bawan, A. Bajwa, J. Paterson, T. McMullen, Active surveillance of low-risk papillary thyroid cancer: A meta-analysis. Surgery **167**, 46–55 (2020). https://doi.org/10.1016/j.surg.2019.03.040
43. I.M. Cazacu, A. Semaan, B. Stephens, D.B. Swartzlander, P.A. Guerrero, B.S. Singh, C.V. Lungulescu, M.M. Danciulescu, I.F. Cherciu Harbiyeli, I. Streata, C. Popescu, A. Saftoiu, S. Roy-Chowdhuri, A. Maitra, M.S. Bhutani, Diagnostic value of digital droplet polymerase chain reaction and digital multiplexed detection of single-nucleotide variants in pancreatic cytology specimens collected by EUS-guided FNA. Gastrointest Endosc **93**, 1142–1151.e2 (2021). https://doi.org/10.1016/j.gie.2020.09.051
44. X. Yu, X. Song, W. Sun, S. Zhao, J. Zhao, Y.-G. Wang, Independent Risk Factors Predicting Central Lymph Node Metastasis in Papillary Thyroid Microcarcinoma. Horm Metab Res Horm Stoffwechselforschung Horm Metab **49**, 201–207 (2017). https://doi.org/10.1055/s-0043-101917

45. Q.-X. Wang, E.-D. Chen, Y.-F. Cai, Y.-L. Zhou, S.-Y. Dong, X.-H. Zhang, O.-C. Wang, Q. Li, Serum deprivation response functions as a tumor suppressor gene in papillary thyroid cancer. Clin Genet **96**, 418–428 (2019). https://doi.org/10.1111/cge.13609
46. A. Parascandolo, F. Rappa, F. Cappello, J. Kim, D.A. Cantu, H. Chen, G. Mazzoccoli, P. Hematti, M.D. Castellone, M. Salvatore, M.O. Laukkanen (2017) Extracellular Superoxide Dismutase Expression in Papillary Thyroid Cancer Mesenchymal Stem/Stromal Cells Modulates Cancer Cell Growth and Migration. Sci Rep 7:41416. https://doi.org/10/f9qdmp
47. A. Bild, P.G. Febbo, Application of a priori established gene sets to discover biologically important differential expression in microarray data. Proc Natl Acad Sci U S A **102**, 15278 (2005). https://doi.org/10.1073/pnas.0507477102
48. N.A. Cipriani, S. Nagar, S.P. Kaplan, M.G. White, T. Antic, P.M. Sadow, B. Aschebrook-Kilfoy, P. Angelos, E.L. Kaplan, R.H. Grogan (2015) Follicular Thyroid Carcinoma: How Have Histologic Diagnoses Changed in the Last Half-Century and What Are the Prognostic Implications? Thyroid Off J Am Thyroid Assoc 25:1209–1216. https://doi.org/10/f7xkw5
49. V.P. Balachandran, M. Gonen, J.J. Smith, R.P. DeMatteo, Nomograms in Oncology – More than Meets the Eye. Lancet Oncol. **16**, e173–e180 (2015). https://doi.org/10.1016/S1470-2045(14)71116-7

Tables

Table 1. GEO datasets for analysis and validation of MTGs signature

Datasets	References	Platform	Samples	Usage
GSE35570	34	Affymetrix Human Genome U133 Plus 2.0 Array	65PTCs, 51Normals	Differential analysis
GSE29265	Contributed by Tomas G, et, al.	Affymetrix Human Genome U133 Plus 2.0 Array	20PTCs, 20Normals, 9ATCs	Differential analysis & external Validation
GSE33630	33	Affymetrix Human Genome U133 Plus 2.0 Array	49PTCs, 45Normals, 11ATCs	Differential analysis & external Validation
GSE60542	35	Affymetrix Human Genome U133 Plus 2.0 Array	33PTCs, 30Normals	Differential analysis
GSE76039	29	Affymetrix Human Genome U133 Plus 2.0 Array	20ATCs, 17PDTCs	External Validation
GSE58545	30	Affymetrix Human Genome U133A Array	27PTCs, 18Normals	External Validation
GSE82208	Contributed by Wojtas B, et, al.	Affymetrix Human Genome U133 Plus 2.0 Array	25Adenomas, 27FTCs	External Validation
GSE5364	31	Affymetrix Human Genome U133A Array	35PTCs, 16Normals	External Validation

Table 2. Baseline characters of 488 cases enrolled from TCGA-THCA cohort

Characteristic	Training set	Testing set	p
n	380	108	
Progression, n (%)			0.901
Free	341 (69.9%)	98 (20.1%)	
Progression	39 (8%)	10 (2%)	
RAS_status, n (%)			0.602
Mutated	48 (9.8%)	11 (2.3%)	
Wild type	332 (68%)	97 (19.9%)	
BRAF_status, n (%)			0.482
Mutated	212 (43.4%)	65 (13.3%)	
Wild type	168 (34.4%)	43 (8.8%)	
Extrathyroid_extension, n (%)			0.756
Minimal (T3)	102 (21.7%)	28 (5.9%)	
Moderate/Advanced (T4)	13 (2.8%)	5 (1.1%)	
None	255 (54.1%)	68 (14.4%)	
Histological_type, n (%)			0.348
Classical/usual	268 (54.9%)	83 (17%)	
Follicular	84 (17.2%)	17 (3.5%)	
Tall Cell	28 (5.7%)	8 (1.6%)	
Neoplasm_focus_type, n (%)			0.101
Multifocal	178 (37.2%)	40 (8.4%)	
Unifocal	195 (40.8%)	65 (13.6%)	
Anatomic_site, n (%)			0.113
Bilateral	69 (14.3%)	12 (2.5%)	
Isthmus	19 (3.9%)	3 (0.6%)	
Unilateral	286 (59.3%)	93 (19.3%)	
Residual_tumor, n (%)			0.677
R0	291 (68.3%)	80 (18.8%)	

R1	38 (8.9%)	13 (3.1%)	
R2	3 (0.7%)	1 (0.2%)	
Ajcc_stage, n (%)			0.716
Stage I	215 (44.2%)	58 (11.9%)	
Stage II	37 (7.6%)	14 (2.9%)	
Stage III	84 (17.3%)	26 (5.3%)	
Stage IV	42 (8.6%)	10 (2.1%)	
M_stage, n (%)			0.691
M0	372 (76.4%)	107 (22%)	
M1	7 (1.4%)	1 (0.2%)	
N_stage, n (%)			0.378
N0	180 (41.1%)	45 (10.3%)	
N1	162 (37%)	51 (11.6%)	
T_stage, n (%)			0.960
T1	111 (22.8%)	30 (6.2%)	
T2	126 (25.9%)	35 (7.2%)	
T3	125 (25.7%)	38 (7.8%)	
T4	16 (3.3%)	5 (1%)	
Gender, n (%)			0.097
Female	286 (58.6%)	72 (14.8%)	
Male	94 (19.3%)	36 (7.4%)	
Age, n (%)			1.000
<55	253 (51.8%)	72 (14.8%)	
≥55	127 (26%)	36 (7.4%)	
Follow-up time, meidan (IQR)	888.5 (496.5, 1463.75)	927 (460, 1340.75)	0.671

Table 3. Uni- and multivariate Cox analysis in TCGA-THCA cohort

Characteristics	Total(N)	Univariate analysis		Multivariate analysis	
		Hazard ratio (95% CI)	P value	Hazard ratio (95% CI)	P value
Riskscore	488	4.892 (2.879-8.313)	**<0.001**	3.447 (1.931-6.151)	**<0.001**
BRAF_status	488				
Wild type	211	Reference			
Mutated	277	1.455 (0.801-2.645)	0.218		
RAS_status	488				
Wild type	429	Reference			
Mutated	59	1.640 (0.768-3.504)	0.201		
Extrathyroid_extension	471				
None/Minimal	453	Reference			
Moderate/Advanced (T4)	18	2.100 (0.754-5.847)	0.156	0.744 (0.209-2.649)	0.648
Neoplasm_size	474				
<2cm	153	Reference			
≥2cm	321	3.914 (1.547-9.901)	**0.004**	2.593 (0.989-6.799)	0.053
Histological_type	488				
Classical/Follicular	452	Reference			
Tall Cell	36	2.417 (1.084-5.389)	**0.031**	1.379 (0.538-3.538)	0.503
Anatomic_site	482				
Unilateral	379	Reference			
Bilateral	81	1.101 (0.513-2.365)	0.805		
Isthmus	22	0.413 (0.057-3.004)	0.382		
M_stage	487				
M0	479	Reference			

M1	8	5.630 (2.021-15.687)	**<0.001**	1.665 (0.467-5.942)	0.432
Ajcc_stage	486				
Stage I/II	324	Reference			
Stage III/IV	162	2.753 (1.567-4.839)	**<0.001**	1.384 (0.538-3.559)	0.500
N_stage	438				
N0	225	Reference			
N1	213	1.736 (0.950-3.172)	0.073	1.177 (0.599-2.310)	0.636
T_stage	486				
T1/T2	302	Reference			
T3/T4	184	2.806 (1.569-5.018)	**<0.001**	1.275 (0.609-2.670)	0.520
Gender	488				
FEMALE	358	Reference			
MALE	130	1.747 (0.977-3.124)	0.060	1.108 (0.573-2.146)	0.760
Age	488				
≥55	163	Reference			
<55	325	0.443 (0.252-0.776)	**0.004**	0.669 (0.281-1.590)	0.363

Table 4. MTGs based signatures from different ratio of training set

Training set of 380 cases		
Symbol	Gene name	Coefficient
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit	0.634261
SDPR	caveolae associated protein 2	-0.06451
CCL14	C-C motif chemokine ligand 14	-0.23073
STARD13	StAR related lipid transfer domain containing 13	-0.05099
DEPDC1B	DEP domain containing 1B	0.075524
SOD3	superoxide dismutase 3	-0.05601
TGFBR3	transforming growth factor beta receptor 3	-0.16766
FBLN5	fibulin 5	-0.01901
ARHGDIB	Rho GDP dissociation inhibitor beta	-0.18682
KISS1R	KISS1 receptor	0.076108

Training set of 488 cases		
Symbol	Gene name	Coefficient
ARHGDIB	Rho GDP dissociation inhibitor beta	-0.06602
CCL14	C-C motif chemokine ligand 14	-0.12939
EZH2	enhancer of zeste 2 polycomb repressive complex 2 subunit	0.610733
FAM3B	FAM3 metabolism regulating signaling molecule B	-0.10611
FBLN5	fibulin 5	-0.13797
KISS1R	KISS1 receptor	0.013126
SDC2	Syndecan 2	-0.00671
SDPR	caveolae associated protein 2	-0.13063
SOD3	superoxide dismutase 3	-0.08979
TGFBR3	transforming growth factor beta receptor 3	-0.01407

Figures

Table 5. Sequences of primers used for RT-qPCR experiment.		
Target gene	Primer	Sequence (5'-3')
GAPDH	forward	AGTCCCTGCCACACTCAG
	reverse	TACTTTATTGATGGTACATGACAAGG
ARHGDIB	forward	GCAAGCTCAATTATAAGCCTCC
	reverse	CTTCCAGATCTCCAGTAAGGTC
CCL14	forward	GGACATGAAGGAGAACTGAGTGACC
	reverse	TGAGAGTTAGCGGTGGGTGGAG
EZH2	forward	AAATCAGAGTACATGCGACTGA
	reverse	GTATCCTTCGCTGTTTCCATTC
FBLN5	forward	CTGTGACCCAGGATATGAACTT
	reverse	TTGTAAATTGTAGCACGTCTGC
FAM3B	forward	CCTATGCCTACAGGTTACTCAG
	reverse	TAAACTTTGTCATCGGTCCAGA
KISS1R	forward	CTGGTCATCTACGTCATCTGC
	reverse	ACTCATGGCGGTCAGAGT
SDC2	forward	GCTGATGAGGATGTAGAGAGTC
	reverse	GTATATTCAGCGTCGTGGTTTC
SDPR	forward	ACATCGACTTGACTATTGTGGA
	reverse	CTACCCTCATAGCGTACCTTCT
SOD3	forward	GGAGTGGATCCGAGACATGTA
	reverse	CGAAGAAGGCGTCGAGCTT
TGFBR3	forward	TCCTCTGAATGGCTGCGGTACTC
	reverse	GGCTGGAACCTGTATCACAATGGAG

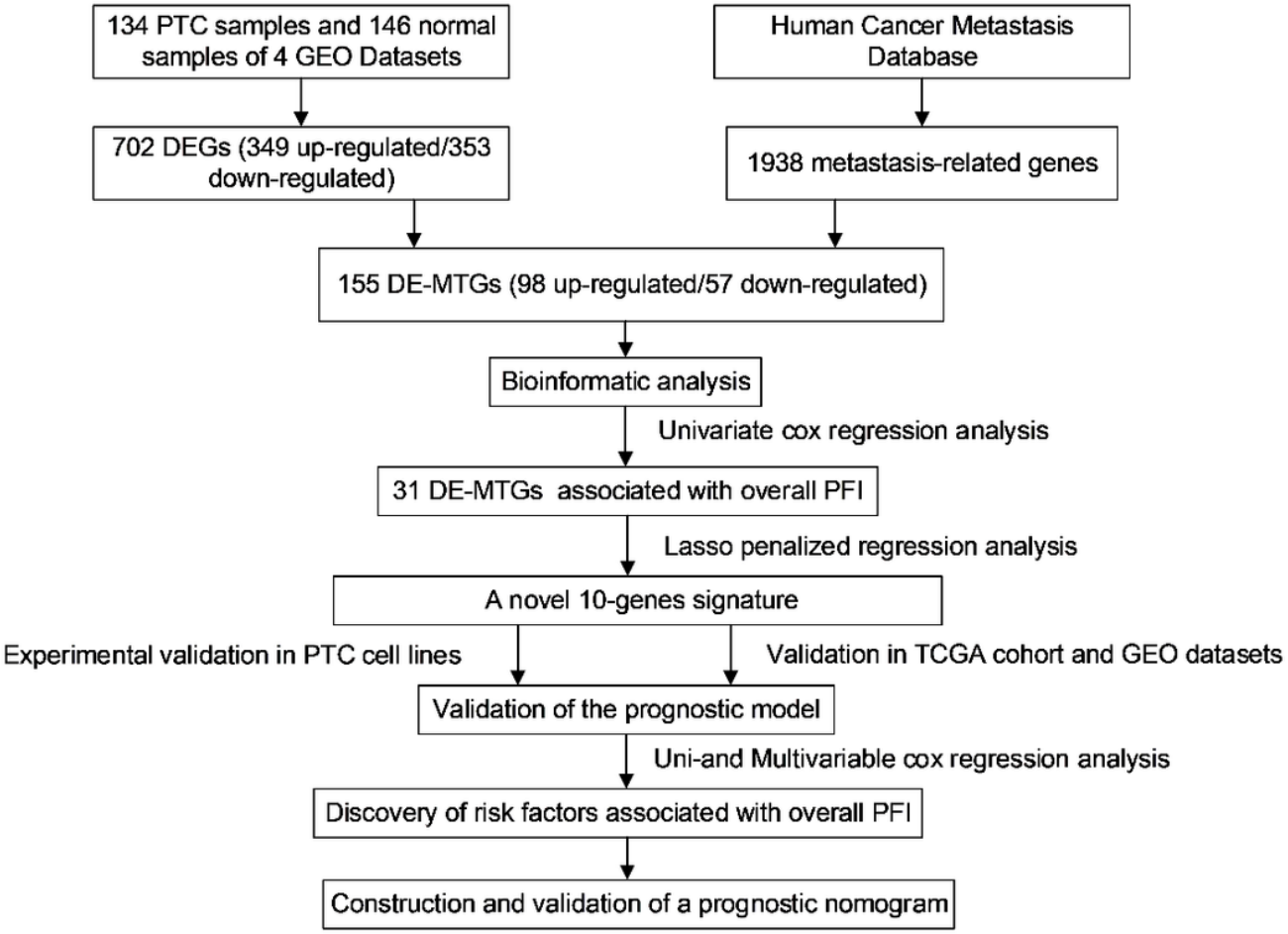

Figure 1

Flowchart describing the process of establishment and validation of the novel 10-gene signature and prognostic nomogram.

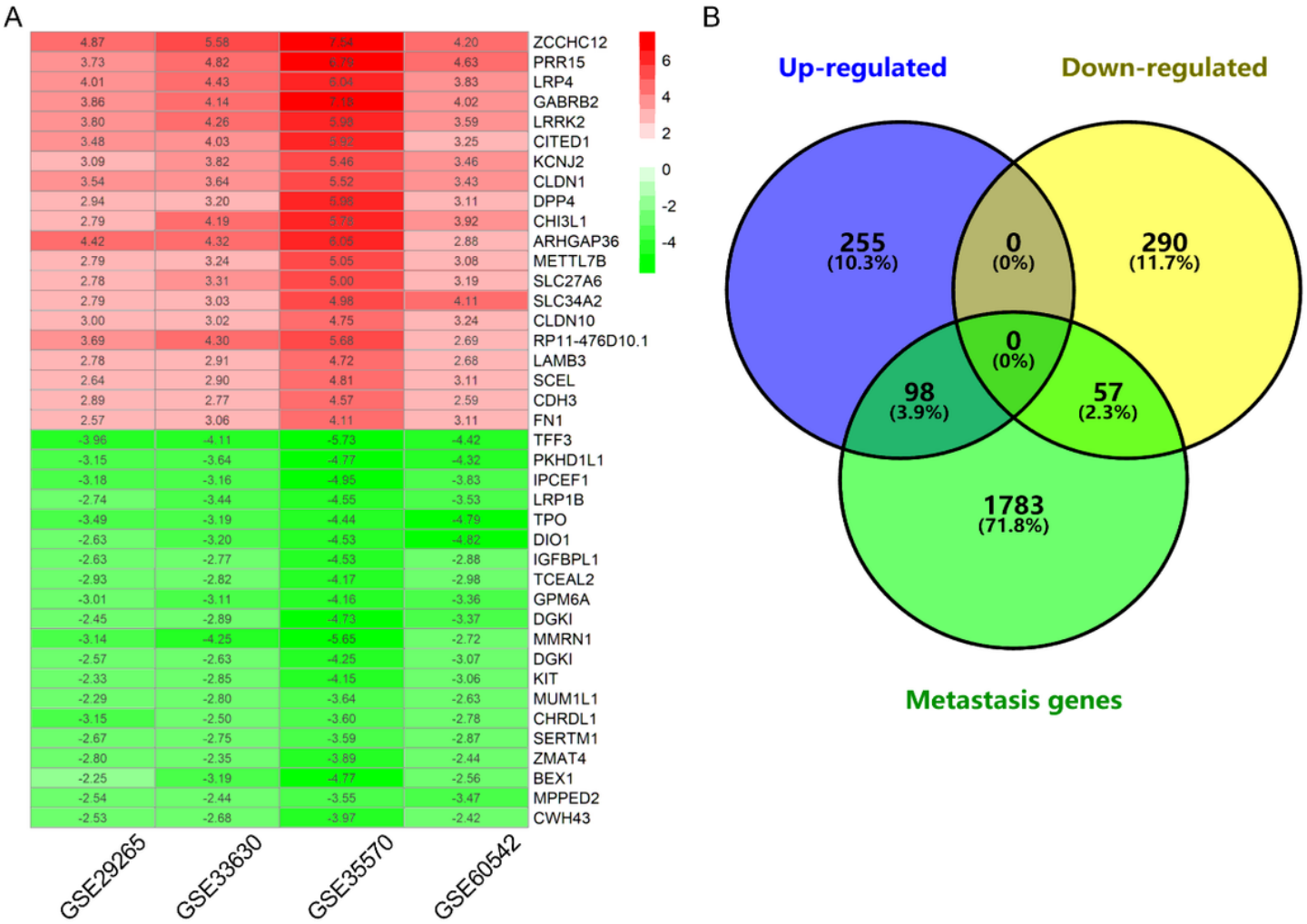

Figure 2

Identification of DE-MTGs in PTC. (A) Heatmap presenting the top 20 upregulated and downregulated DEGs in PTC after integrated analysis of the 4 GEO datasets using the RRA method. (B) 155 DE-MTGs, including 98 upregulated and 57 downregulated genes, were identified based on the intersection between GEO and THCA datasets.

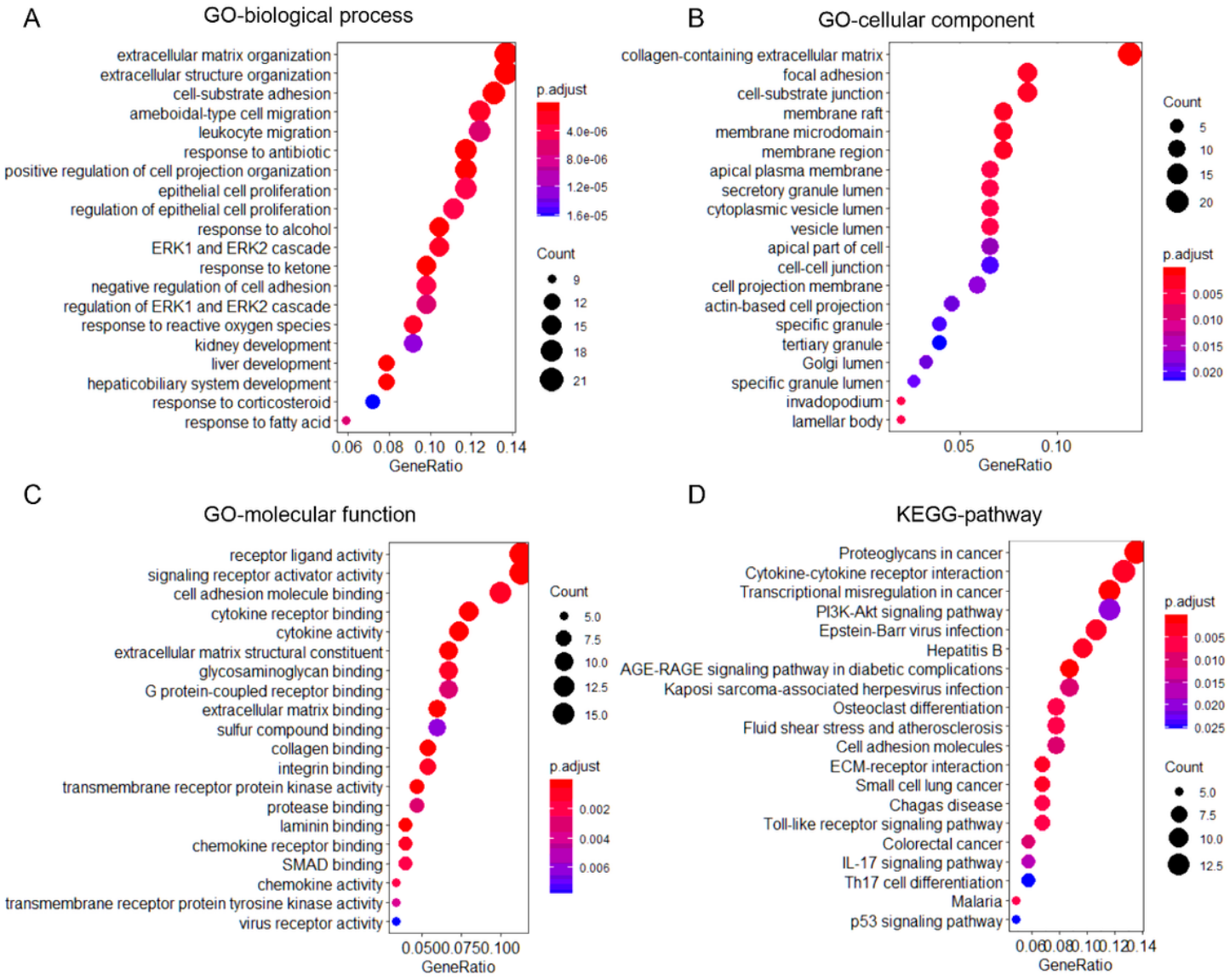

Figure 3

Functional enrichment analysis of the 155 DE-MTGs. (A) The top 20 enriched gene ontology (GO) biological process (BP) terms of the DE-MTGs. (B) The cellular components (CC) terms of the DE-MTGs. (C) The molecular function (MF) terms of the DE-MTGs. (D) The KEGG pathway terms of the DE-MTGs[24].

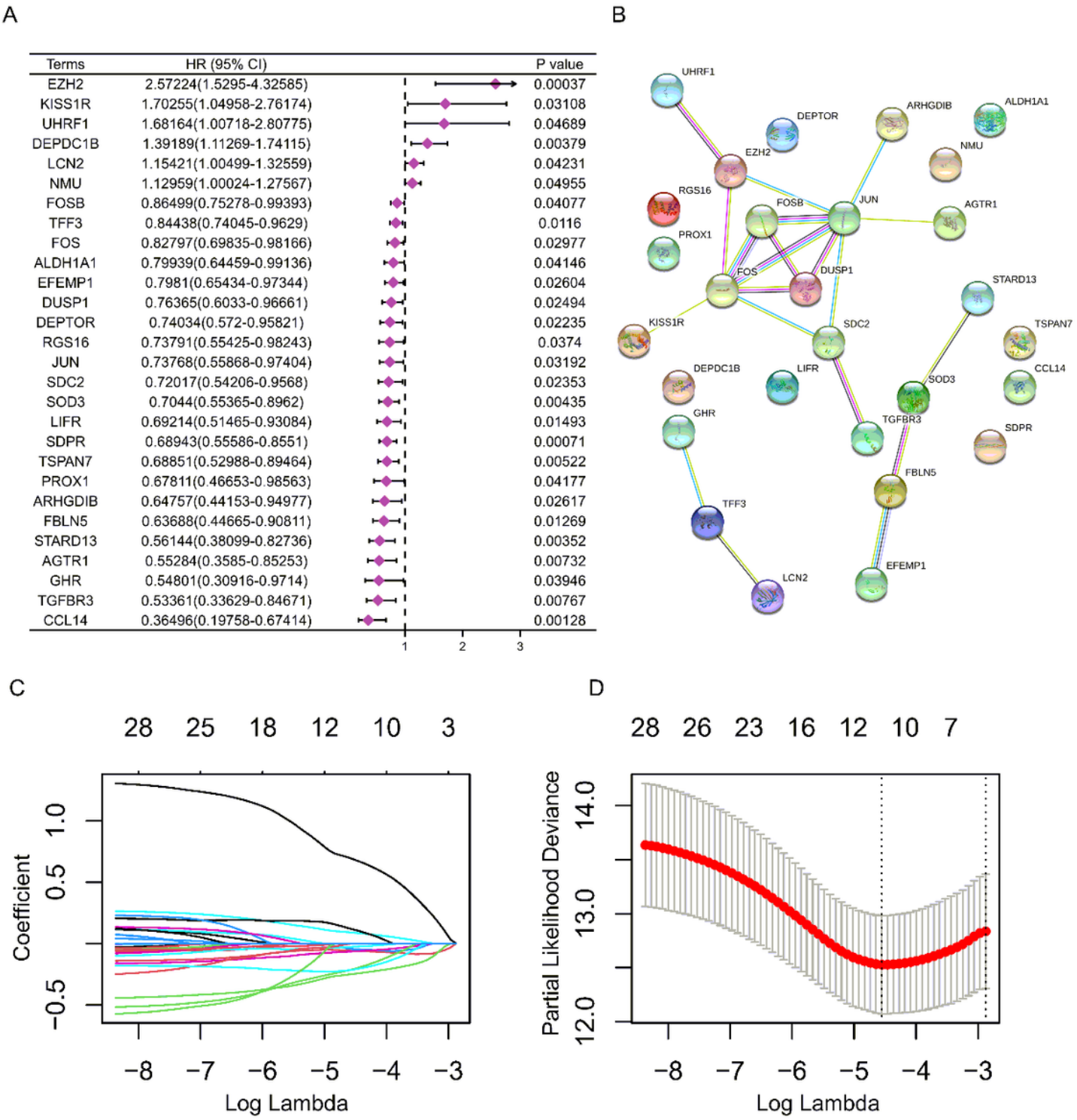

Figure 4

Differential expression level and hazard ratios (HR) of the 28 DE-MTGs in training set. (A) Forest plot with hazard ratios (HR) representing the predictive values of the 28 DE-MTGs that were PFI-related in PTC. (B) Protein-Protein Interaction network of the 28 DE-MTGs. (C) LASSO coefficient profiles of the 28 DE-MTGs. (D) Lasso deviance profiles of the 28 DE-MTGs. The lambda selection criterion was based on the value of lambda giving a minimum mean cross-validation error; lambda min = 0.01394.

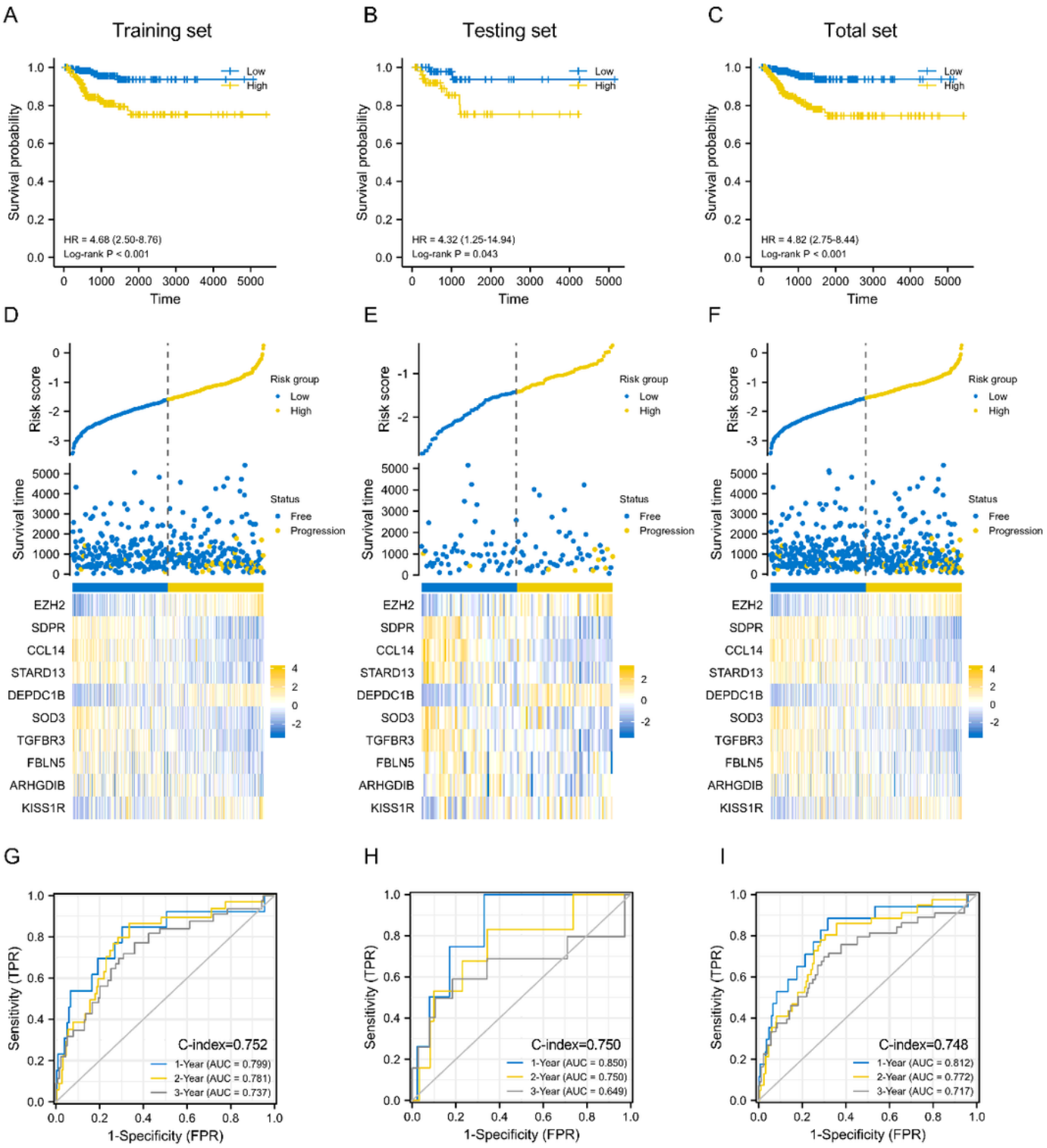

Figure 5

Evaluation of the efficacy of the 10-gene signature in the TCGA-THCA dataset. The dataset was randomly divided into the training set, and the validation set in 0.8 ratio. (A-C) Log-rank survival curves of the 10-gene signature represented PFI proportions at different timepoints. Patients from the training/validation/total sets are defined as "high risk" or "low risk" according to the median value of riskscores. (D-F) Relationship between the gene risk score (up) and recurrence status of patients of

high/low-risk (down) in training/validation/total TCGA-THCA dataset. (G-I) Time-dependent ROC for the predictions of PFI for the 10-gene signature in the training/validation/total sets.

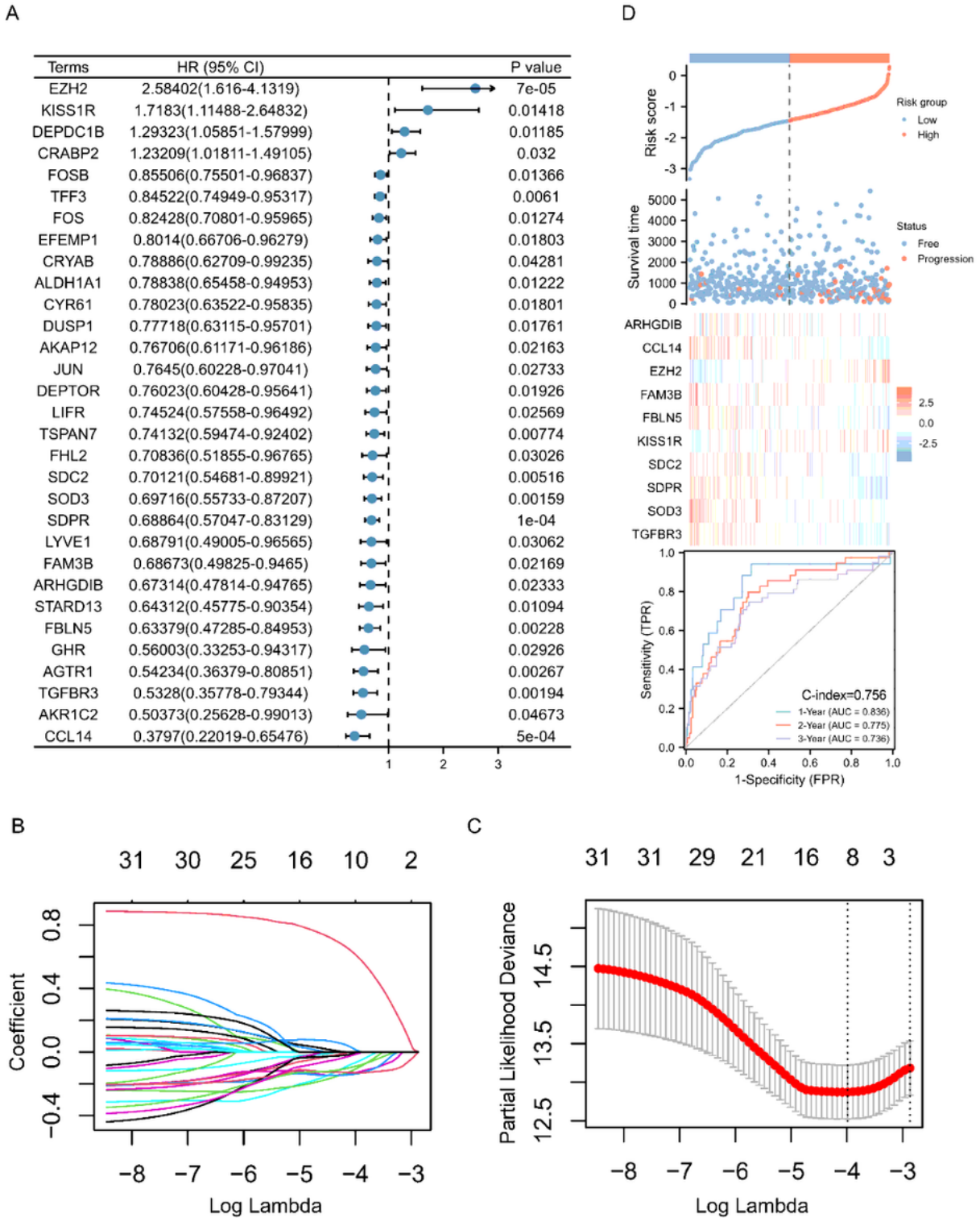

Figure 6

Generation and evaluation of the efficacy of the optimized 10-gene signature in the total TCGA-THCA dataset. All the dataset was enrolled into the training set. (A) Forest plot with hazard ratios (HR)

representing the predictive values of the 18 DE-MTGs that were PFI-related in PTC. (B) LASSO coefficient profiles of the 18 DE-MTGs. (C) Lasso deviance profiles of the 28 DE-MTGs. Lambda min = 0.01851. (D) Relationship between the gene risk score and recurrence status of patients of high/low risk in total set divided by median value. Time-dependent ROC for the predictions of PFI for the optimized 10-gene signature in the total set.

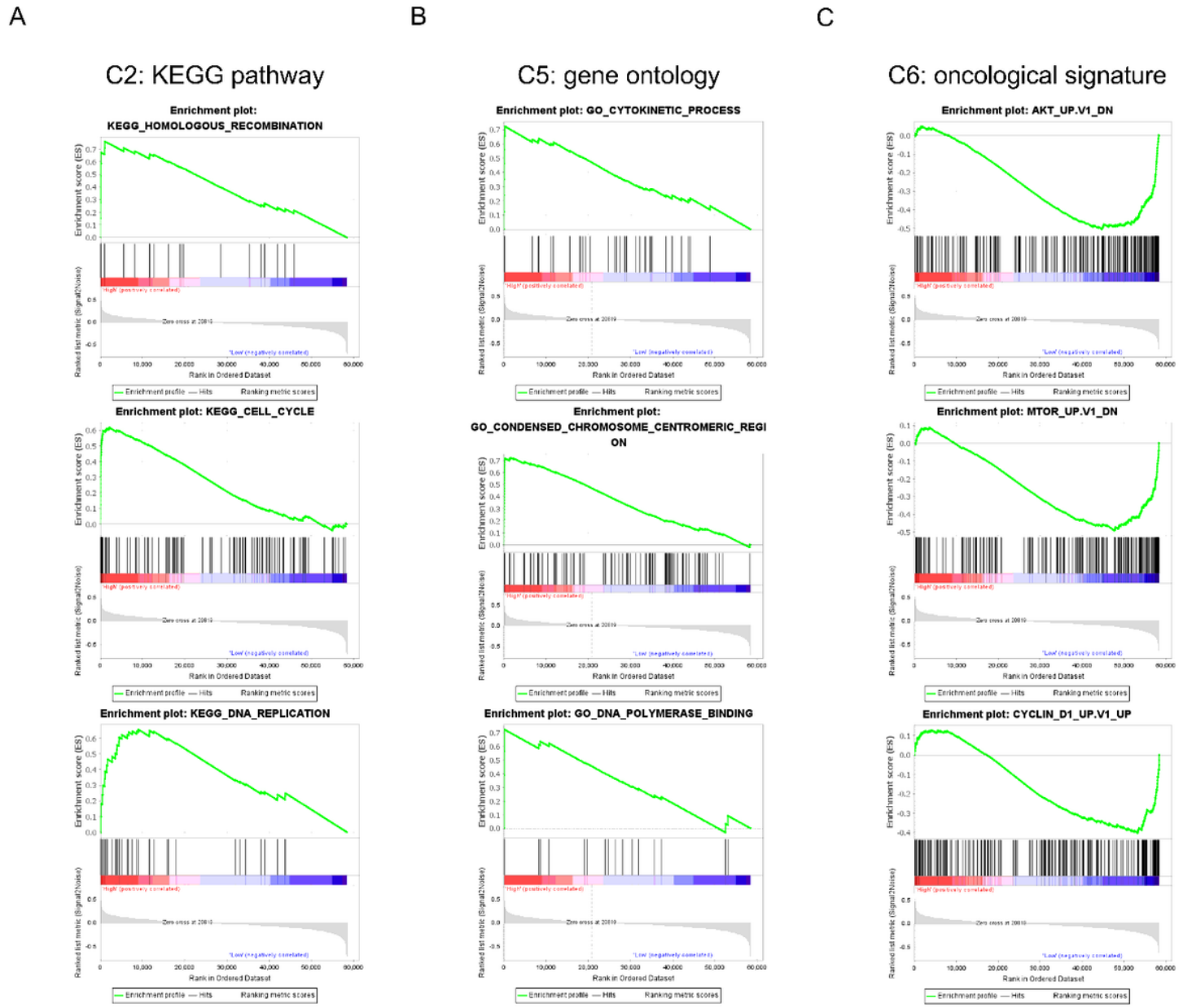

Figure 7

Gene set enrichment analysis (GSEA) analysis of the 10-gene signature. (A-C) Representative signaling pathways, biological functions, and oncogenic signatures significantly enriched in the high-risk group identified by GSEA.

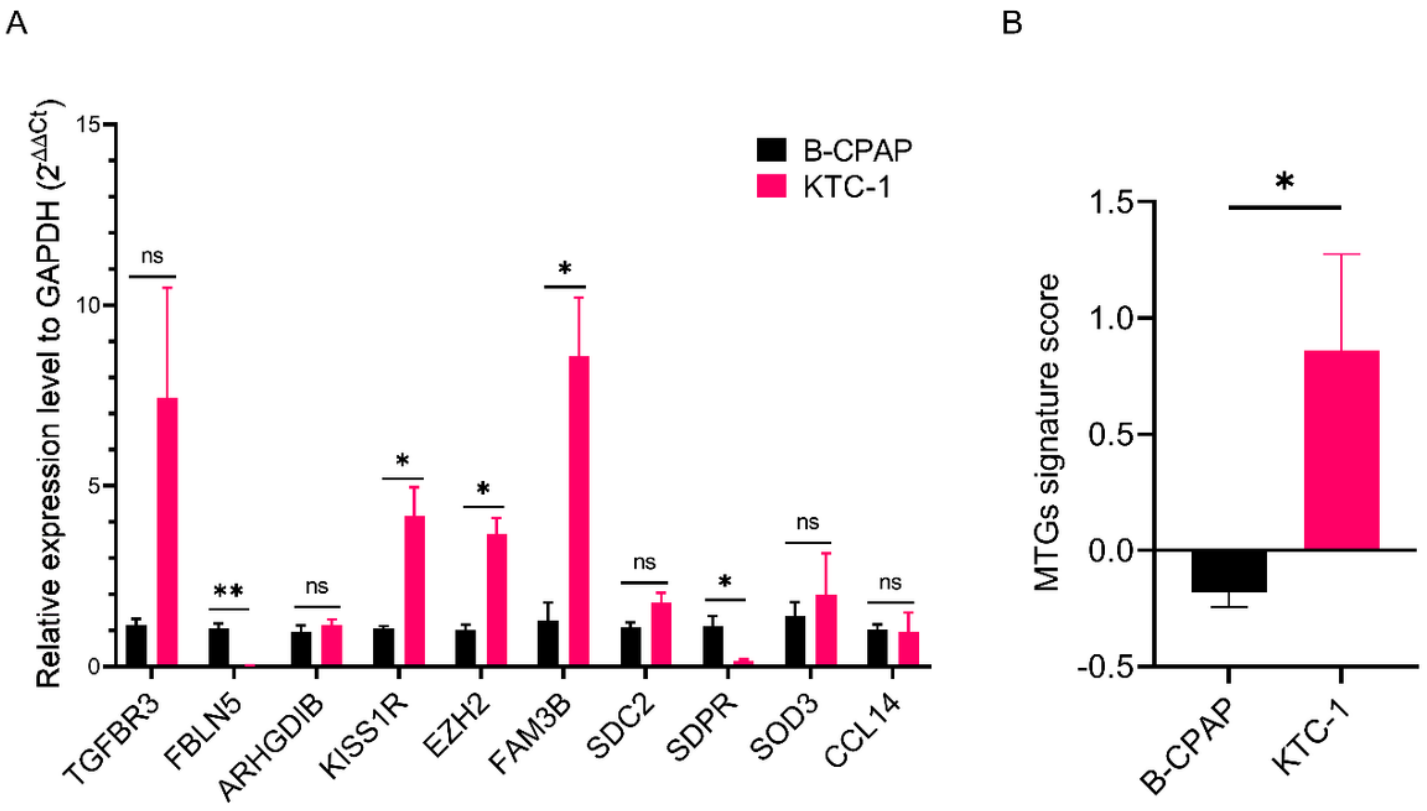

Figure 8

RT-qPCR quantification analysis of MTGs based 10-gene signatures in PTC cell lines. (A) Relative expression level of 10 genes to GAPDH ($2^{-\Delta\Delta CT}$) in B-CPAP and KTC-1 cell line. (B) Riskscores of PTC cell lines calculated by MTGs signature. Data are presented as Mean with SEM. Unpaired t test (n = 4), *P < .05, **P < 0.01.

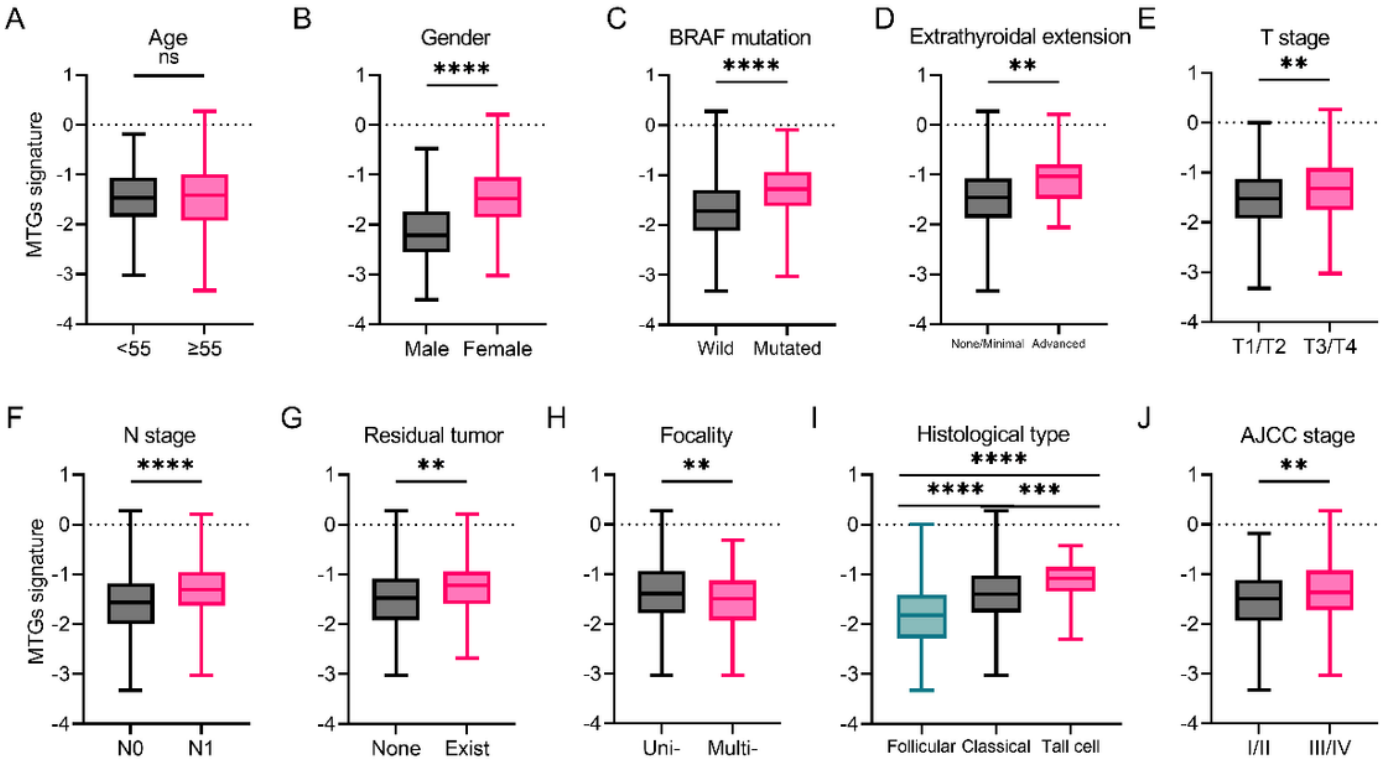

Figure 9

Correlations between MTGs signature with clinical characters in TGCA-THCA cohort. (A-J) The distribution of the signature scores according to different status of age, gender, BRAF mutation, extrathyroidal invasion, residual tumor, focality type, histological type and TNM disease stage in the TCGA-THCA dataset. Unpaired t test, *P < .05 ***P < 0.001.

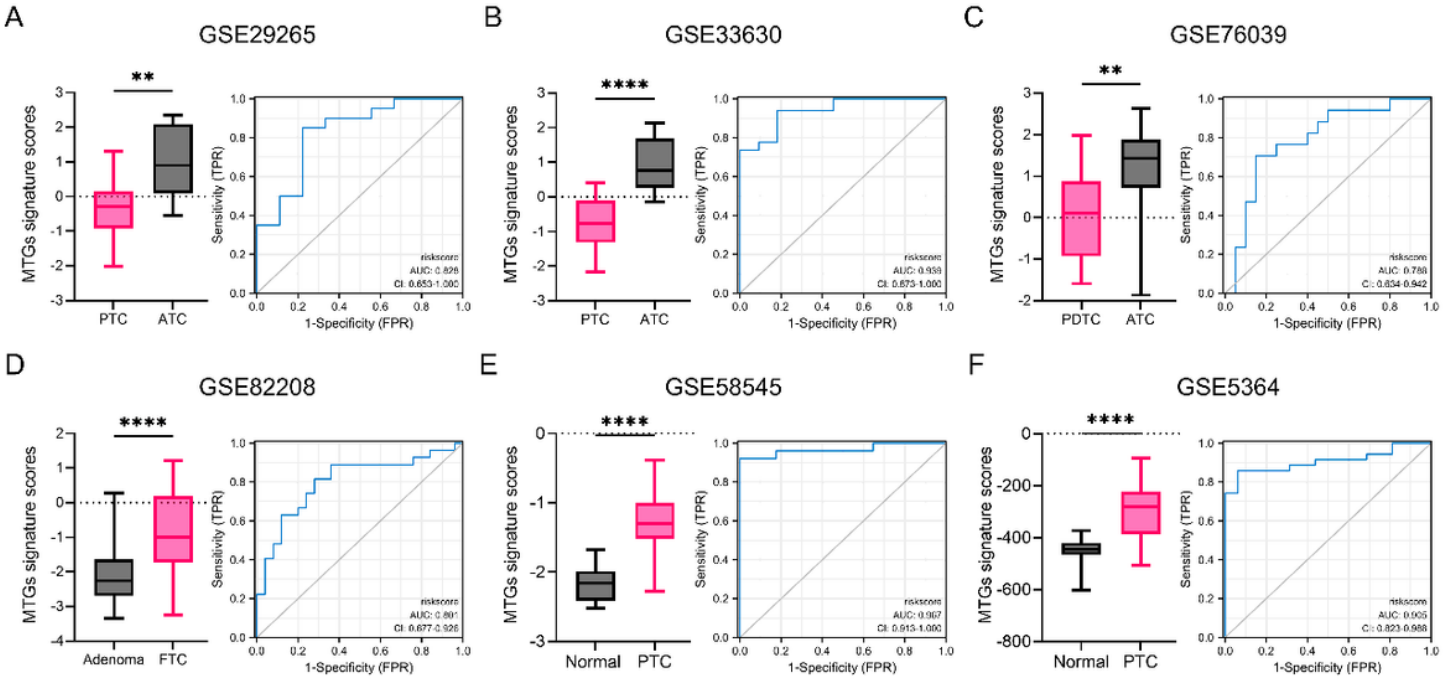

Figure 10

External clinical validation of the MTGs signature. Box plot and ROC curve revealing the discriminative value of MTGs signature. (A-C) Risk-scores of ATC/PDTC versus PTC samples from GSE 29265, GSE 33630 and GSE 76039. (D) Risk-scores of follicular adenoma samples versus FTC samples in GSE 82208. (E, F) Risk-scores of normal thyroid tissues versus PTC tissues in GSE58545 and GSE5364. Unpaired t test, **P < 0.01, ****P < 0.0001.

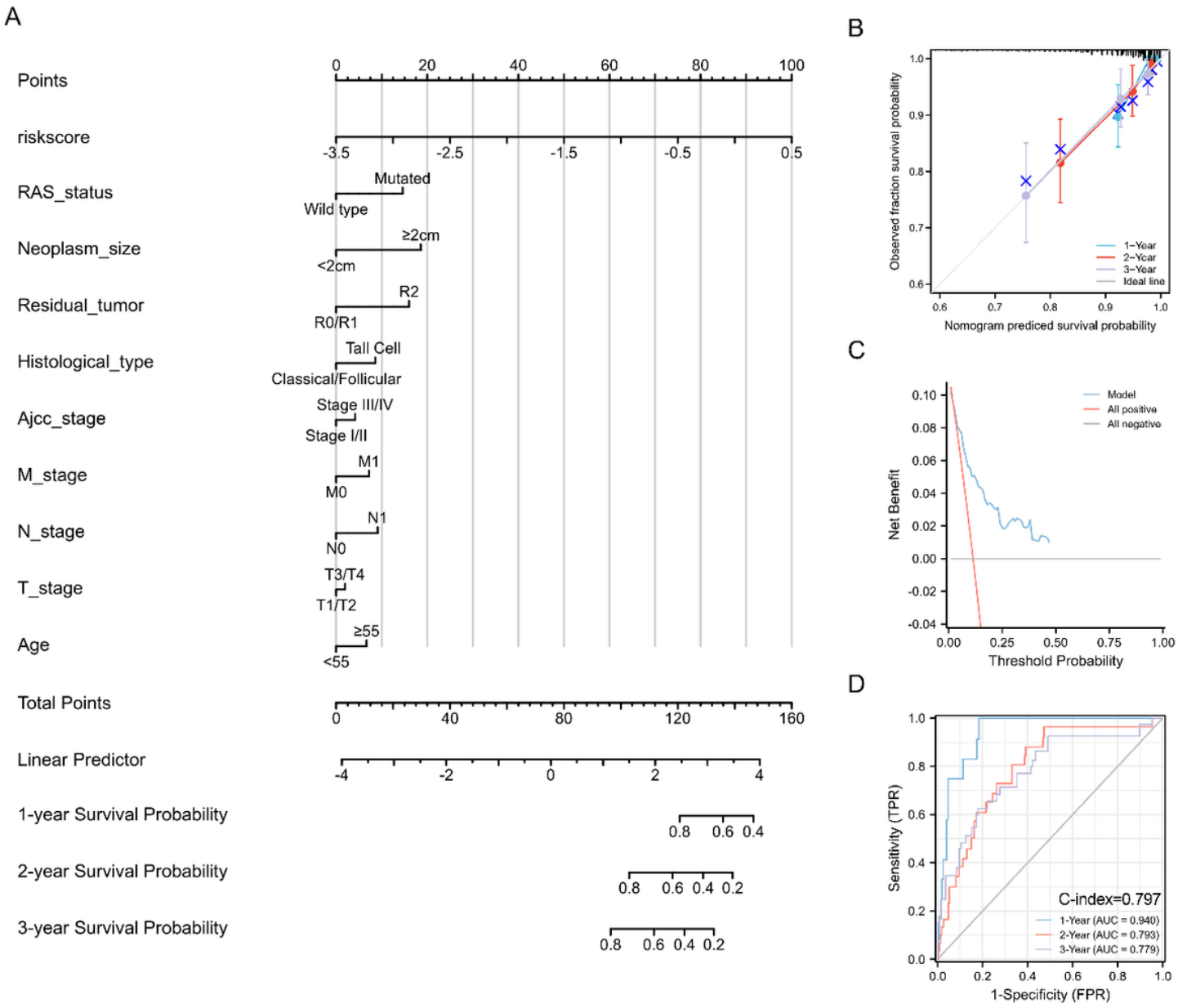

Figure 11

Construction and validation of the nomogram in predicting PFI of PTC in the TCGA-THCA dataset. (A) A nomogram based on the 10-gene signature and relevant clinical features for forecasting the PFI of PTC. (B) The calibration curve for internal validation of the nomogram. (C) The DCA curve showing the clinical

utility of the nomogram in 3-year PFI. (D) Time-dependent ROC for predicting the 1-, 2- and 3-year PFI of PTC.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 20211228supplementaryfile.pdf