European Journal of Medical Research

RESEARCH

Open Access

Identification of thyroid cancer biomarkers using WGCNA and machine learning

Gaofeng Hu[1,2†], Wenyuan Niu[2,3†], Jiaming Ge[1,2†], Jie Xuan[1,2], Yanyang Liu[1,2], Mengjia Li[1,2], Huize Shen[2,3], Shang Ma[2*], Yuanqiang Li[2*] and Qinglin Li[1,2*]

Abstract

Objective The incidence of thyroid cancer (TC) is increasing in China, largely due to overdiagnosis from widespread screening and improved ultrasound technology. Identifying precise TC biomarkers is crucial for accurate diagnosis and effective treatment.

Methods TC patient data were obtained from TCGA. DEGs were analyzed using DESeq2, and WGCNA identified gene modules associated with TC. Machine learning algorithms (XGBoost, LASSO, RF) identified key biomarkers, with ROC and AUC > 0.95 indicating strong diagnostic performance. Immune cell infiltration and biomarker correlation were analyzed using CIBERSORT.

Results Four key genes (*P4HA2*, *TFF3*, *RPS6KA5*, *EYA1*) were found as potential biomarkers. High *P4HA2* expression was associated with suppressed anti-tumor immune responses and promoted disease progression. In vitro studies showed that *P4HA2* upregulation increased TC cell growth and migration, while its suppression reduced these activities.

Conclusion Through bioinformatics and experimental validation, we identified *P4HA2* as a key potential thyroid cancer biomarker. This finding provides new molecular targets for diagnosis and treatment. *P4HA2* has the potential to be a diagnostic or therapeutic target, which could have significant implications for improving clinical outcomes in thyroid cancer patients.

Keywords Thyroid cancer, Biomarkers, Machine learning, *P4HA2*

Introduction

Neoplasms remain the main killer worldwide [1–4]. According to recent statistics, in 2024, the global incidence of thyroid cancer (TC) is ranked 7th among 185 countries, marking a notable rise from 11th place in 2022 [5, 6]. Furthermore, the incidence of thyroid cancer ranks third among all malignant tumors in China. Thyroid cancer is primarily categorized into three histological types: among these, differentiated thyroid cancer (DTC) is the most prevalent pathological type, constituting approximately 90% of malignant thyroid tumors, with females experiencing a substantially higher incidence than males [7].

In its early stages, TC typically presents with no overt symptoms; however, recent advancements in the

†Gaofeng Hu, Wenyuan Niu and Jiaming Ge have contributed equally to the work.

*Correspondence:
Shang Ma
1226609433@qq.com
Yuanqiang Li
yuanqiangli@hotmail.com
Qinglin Li
qinglin200886@126.com
[1] Wenzhou Medical University, Wenzhou, Zhejiang, China
[2] Zhejiang Cancer Hospital, Hangzhou, Zhejiang, China
[3] School of Pharmaceutical Sciences, Zhejiang Chinese Medical University, Hangzhou, Zhejiang, China

diagnosis and treatment of TC have facilitated earlier detection. Thyroid nodules are generally identified through palpation and ultrasonography during routine thyroid cancer screenings of the general population, with 30–40% of cases being detected via palpation [8–10]. Although the overall prognosis for TC is generally favorable, the risk of tumor metastasis and recurrence remains substantial. Approximately 30% of patients with DTC experience recurrence and distant metastasis. Within this subgroup, about one-third progressively lose their sensitivity to radioiodine (RAI) therapy due to reduced uptake of radioiodine by cancer cells, ultimately developing into radioiodine-refractory differentiated thyroid cancer (RAIR-DTC) [7, 11, 12]. The 10-year survival probability for RAIR-DTC is a mere 10%, with a median survival duration spanning from 3 to 5 years. In addition, the disease-specific mortality rate for anaplastic thyroid cancer (ATC) is close to 100%, with a median survival of only 4 months after diagnosis. TC significantly affects both the physical and mental health of individuals and imposes considerable economic and social burdens on society. Consequently, there is an urgent need to identify reliable biomarkers to enhance the detection and treatment of this disease.

In the medical domain, the utilization of machine learning technologies is garnering increasing scholarly attention. Machine learning can extract valuable insights from extensive medical datasets, thereby facilitating more precise diagnoses, treatment plans, and predictive models for clinicians and researchers [13]. In our study, we employed three machine learning algorithms: LASSO (least absolute shrinkage and selection operator), RF (Random Forest), and XGBoost (extreme gradient boosting), to identify characteristic genes associated with thyroid cancer. LASSO regression is a linear model that is extensively used for feature selection and regression analysis in the medical field. It incorporates a penalty term λ, which reduces model complexity and enhances generalizability [14]. RF models, an ensemble learning technique, construct multiple decision trees for classification or regression purposes. XGBoost, an advanced gradient boosting algorithm, enhances predictive accuracy by integrating decision trees and is prevalently utilized in machine learning applications. In the medical context, Random Forest is extensively applied in areas such as disease risk assessment, survival analysis, and image recognition [15].

In our study, we screened biomarkers for TC based on the TCGA-THCA database. We used three machine learning methods: LASSO, RF, and XGBoost. We further experimentally validated the functions of the screened biomarkers (Fig. 1). Our research offers innovative approaches for diagnosing thyroid cancer and identifies novel therapeutic targets for its treatment. Our study represents the first multi-omics integrated analysis to reveal the dual role of *P4HA2* in thyroid cancer (pro-cancer and immunosuppressive), providing novel insights into its potential as a diagnostic and therapeutic target.

Materials and methods

Data acquisition

From the TCGA (https://portal.gdc.cancer.gov/) [16–18] and the Gene Expression Omnibus (GEO) [19–21] we retrieved and downloaded transcriptomic data for TC along with corresponding clinical information. In the THCA cohort derived from the TCGA database, this study encompassed a total of 572 samples, consisting of 59 normal thyroid tissue samples and 513 thyroid cancer tissue samples. Furthermore, two GEO datasets based on the Affymetrix GPL570 platform were utilized: GSE29265, comprising 10 ATC samples, 59 TC samples, and 45 normal samples; and GSE33630, which includes 60 TC samples and 45 normal samples [22, 23].

Differential expression genes screening

We identified DEGs between TC and normal tissues from the THCA cohort using the "DESeq2" package (version 1.44.0) in R software (version 4.4.1), employing a threshold of a false discovery rate (FDR) < 0.05 and an absolute log2 fold change (|log2 FC|) > 1, where log2 FC < − 1 were classified as down-regulated, and log2 FC > 1 were classified as up-regulated. Visualization of DEGs was conducted through volcano plots and heat map generated with the "ggplot2" package (version 3.5.1).

WGCNA

In this study, we utilized the R package "WGCNA" to perform WGCNA. The TCGA dataset was divided into cancerous and adjacent normal tissue groups to identify modules most relevant to the dataset. We computed Pearson correlation coefficients between genes and determined a soft threshold of 6 to ensure that the gene network conformed to a scale-free topology. This selection was based on the evaluation of the Scale-Free Topology Model Fit Index (signed R^2) and Mean Connectivity. Specifically, we chose the first soft threshold that surpassed an R^2 value of 0.8 for the scale-free topology model fit, while the corresponding Mean Connectivity approached zero, ensuring the sparsity and biological significance of the network [24]. A hierarchical clustering dendrogram was generated, with distinct branches representing different gene modules. Genes with similar expression patterns were grouped into modules, each containing a minimum of 30 genes. Subsequently, analogous modules were combined at a cut height threshold of 0.25. By overlapping the core genes within the core module with

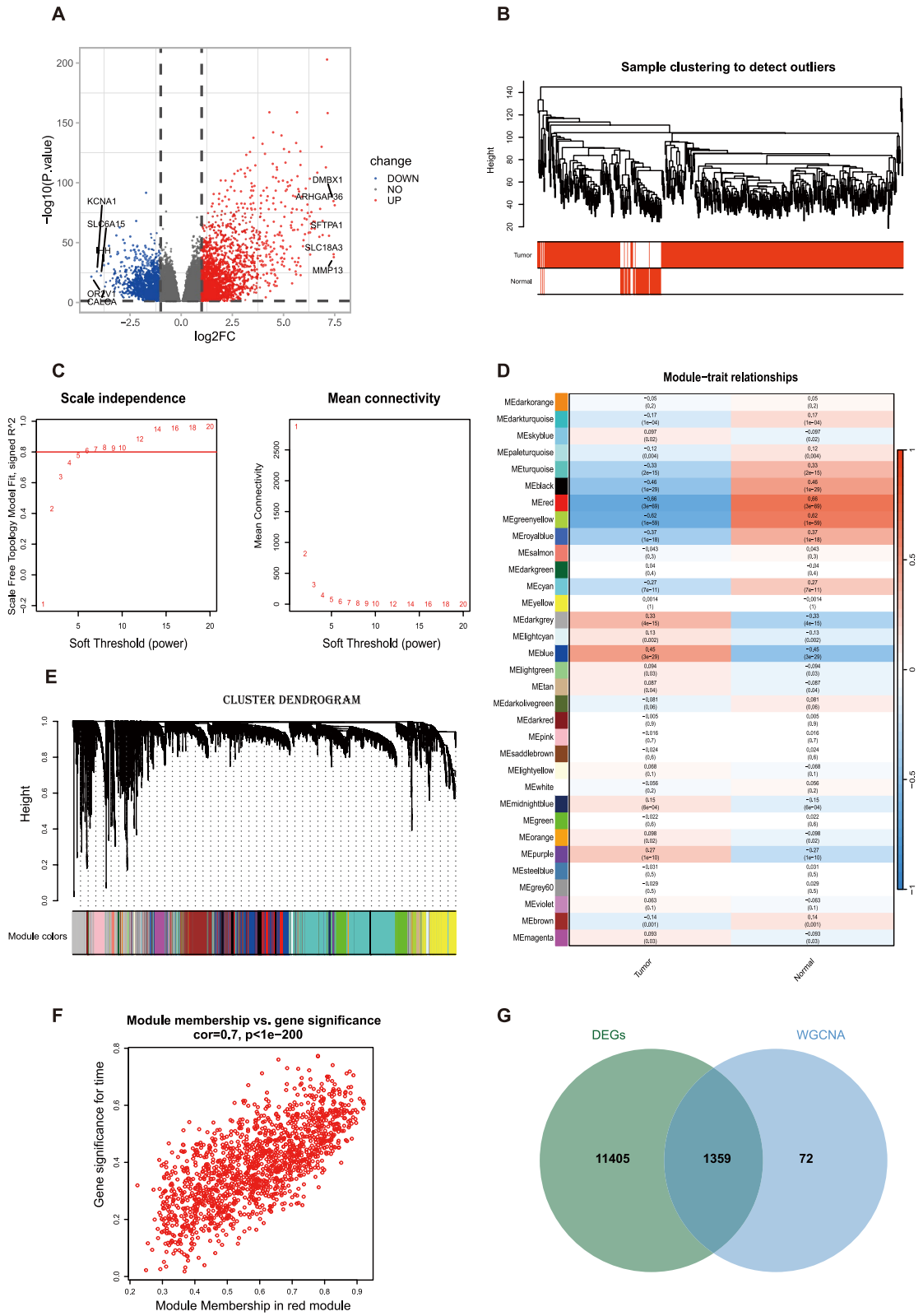

Fig. 1 Integrated genomic analysis revealing key genes and modules in thyroid cancer. **A** Volcano plot highlighting differentially expressed genes between thyroid cancer and normal tissues. **B** Sample clustering and outlier detection. **C** Selection of soft-thresholding power for WGCNA analysis. **D** Heatmap of module–trait relationships in WGCNA analysis. **E** Dynamic tree cut for WGCNA module identification. **F** Correlation of the red module with thyroid cancer traits. **G** Intersection of differentially expressed genes and genes within the most relevant WGCNA module

previously identified DEGs, we identified 1359 genes as potential hub genes [25].

PPI network analysis

All 1359 genes were subsequently used as input for protein–protein interaction (PPI) network analysis. We utilized STRINGdb (version:2.16.4), setting the protein interaction score threshold at 905 to ensure that only proteins with significant interactions were included. To identify key hub genes, we employed the Maximum Clique Centrality (MCC) method, which focuses on the most densely connected subsets of nodes within the network. Utilizing the functionality of the STRING database, we created a protein–protein interaction network that visually illustrates the relationships between proteins. This analysis ultimately identified 291 key hub genes, which were characterized by their centrality within the network and their involvement in the most significant interactions, as determined by the MCC approach.

Identification of optimal diagnostic gene biomarkers by machine learning

RF is a machine learning technique that builds an ensemble of decision trees to perform classification or regression tasks, improving the model's accuracy and robustness by aggregating predictions [26]. In our research, we utilized the "randomForest" R package(version:4.7–1.2) to develop the RF model. To determine the optimal number of trees (ntree), we conducted a grid search over a range of values and selected 151 based on minimizing the error rate across the entire dataset, which was further validated through a tenfold cross-validation process. We focused on genes with a feature importance score exceeding 25.

LASSO logistic regression is a data analysis method that employs L1 regularization to penalize regression coefficients, thereby simplifying the model, reducing multicollinearity, and preventing overfitting [27]. We implemented the LASSO logistic regression model using the "glmnet" R package (version:4.1–8) with tenfold cross-validation (nfolds = 10) to assess model stability and select the optimal regularization parameter λ ("lambda.min"). Model performance was evaluated using a binomial distribution and default settings. By selecting the regularization parameter λ that minimized the criteria, we extracted coefficients from the LASSO model and identified 33 candidate genes.

Furthermore, a dataset was created that integrates the expression profiles of the selected genes with their associated clinical features for analysis using XGBoost. XGBoost, an efficient ensemble learning method, optimizes a regularized objective function using a gradient boosting framework [28]. Through model training and feature importance evaluation by "XGBoost" (version: 1.7.8.1), 42 candidate genes were identified from an initial set of 291 hub genes. These candidates are potentially pivotal in the onset and progression of the disease. The XGBoost model was trained using default parameters for max_depth (5), eta (0.3), and nround (25), which are commonly recommended to balance model complexity and performance, and its stability was assessed through a fivefold cross-validation process.

Ultimately, biomarkers were identified through an integrative approach employing LASSO regression, Random Forest, and XGBoost methodologies. The diagnostic efficacy of these biomarkers was assessed via ROC curve analysis, with the AUC calculated using the "pROC" R package. A Venn diagram was utilized to depict the overlap of results obtained from LASSO, Random Forest, and XGBoost. Heatmaps were created using the "pheatmap" (version 1.0.12) packages.

Development and validation of a prognostic model for thyroid cancer based on multivariate Cox regression and nomogram analysis

We employed the survminer package (version 0.5.0) to conduct a multivariate Cox regression analysis. Using the expression data and regression coefficients of the genes *TFF3*, *EYA1*, *RPS6KA5*, and *P4HA2*, we calculated a risk score for each sample. Subsequently, the samples were stratified into low- and high-risk subgroups in a 3:1 ratio based on their risk scores. We then utilized the survival package to visualize the distribution of survival times between the high-risk and low-risk groups. Additionally, we constructed time-dependent ROC curves and calculated the AUC at 1-, 3-, and 5-year survival time points using the timeROC package (version 0.4).

We further performed a multivariate Cox regression analysis to assess the independent prognostic significance of clinical stage, gender, race, age, and risk score in TC. By integrating these variables, we determined the independence of each factor and its impact on clinical outcomes. Moreover, using the rms package (version 6.8–2), we developed a nomogram to facilitate the prediction of 1-, 3-, and 5-year survival probabilities for TC patients based on individual characteristics. The diagnostic performance of the model was evaluated using the pROC package.

Immune infiltration analysis

Immune cell infiltration was assessed using the CIBERSORT algorithm for estimating relative subsets of RNA transcripts along with correlation analysis between infiltrating immune cells and biomarkers. For the immune infiltration analysis, the input data was standardized by converting FPKM data to TPM (transcripts per million).

This standardization method ensures that the gene expression data are on a comparable scale across different samples, which is crucial for accurate CIBERSORT analysis. Patients were divided into *P4HA2*-high and *P4HA2*-low expression groups based on the median expression level of *P4HA2*. The CIBERSORT algorithm was then used to determine the infiltration levels of immune cells in the *P4HA2*-high and *P4HA2*-low groups. The relative abundance of each of the 22 immune cell subtypes was calculated for each sample, with 1000 iterations performed per sample [29].

Functional enrichment analysis

Conduct differential expression analysis between the *P4HA2*-high and *P4HA2*-low groups to pinpoint upregulated genes, which will subsequently be analyzed using gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses to uncover their biological roles. We employed the "clusterProfiler" R package (version:4.12.6) to perform GO and KEGG enrichment analysis, evaluating biological processes (BP), molecular functions (MF), cellular components (CC), and gene-associated signaling pathways. Following this, the findings from the GO and KEGG pathway analyses were depicted using the ggplot2 (version: 3.5.1) package. An adjusted p-value of less than 0.05 was deemed to indicate statistical significance.

Cell culture and cell transfection

The normal human thyroid cell line Nthy-ori 3–1 and human thyroid cancer cell lines FTC238, IHH4, FTC133, BCPAP, TCO-1, TPC-1, 8505C, CAL62, SW579, K1, TTA1 were all sourced from Zhejiang Provincial People's Hospital (Zhejiang, China). Nthy-ori 3–1, IHH4, BCPAP, and TPC-1 cells were grown in RPMI-1640 complete medium, which was enriched with 10% (vol/vol) fetal bovine serum (FBS) and 1% (vol/vol) antibiotics (100 U/mL penicillin and streptomycin) (Thermo, USA). FTC238 cells were maintained in a DMEM/F-12(HAM)1:1 complete medium, also supplemented with 10% (vol/vol) FBS and 1% (vol/vol) antibiotics (100 U/mL penicillin and streptomycin) (Thermo, USA). FTC133, TCO-1, 8505C, CAL62, SW579, K1, and TTA1 cells were cultured in DMEM complete medium, containing 10% (vol/vol) FBS and 1% (vol/vol) antibiotics (100 U/mL penicillin and streptomycin) (Thermo, USA). All these cell lines were incubated in a humidified environment at 37 °C with 5% CO_2.

To assess the effects of the P4HA2 gene on the biological characteristics of thyroid cancer cell lines, cells from the FTC238, TPC-1, and K1 lines were plated at a density of 4×10^5 cells per well in a 6-well cell culture plate, 14 h before transfection. Transfection was performed using P4HA2 siRNA/overexpression vectors with Lipofectamine 3000 (Invitrogen, USA). siRNA was transfected at a concentration of 50 nM according to the instructions, and cells were collected 48 h later. The siRNA sequences are as follows: genOFF™ st-h-*P4HA2*_001: GAAGGT GACTACCGAACAA; genOFF™ st-h-*P4HA2*_002: GCC GAATTCTTCACCTCTA; genOFF™ st-h-*P4HA2*_003: GCTGCAATTTGGCCTAAGA (RiboBio, China). After an 8-h transfection period, the medium was changed to a complete medium, and the cells were cultured for an additional 40 h in the incubator before being harvested for further examination.

RT qPCR

TRIzol reagent (Invitrogen, USA) was used to extract RNA from the cultured cells. Using the HiScript II Q RT SuperMix from Vazyme, China, the extracted RNA was converted into complementary DNA (cDNA). Quantitative real-time PCR (RT qPCR) was conducted using the ChamQ Universal SYBR qPCR Master Mix (Vazyme, China) on a BioRad CFX96 Touch™ Real-Time PCR Detection System (BioRad, CA). Threshold cycle (Ct) values were obtained and analyzed using the BioRad CFX Manager Software version 5.0. In the experiment, the human primer sequences are detailed in Table 1, and 18S ribosomal RNA (18S) was used as the reference gene for normalization purposes.

Western blot

Protein extraction from TPC-1, FTC238, and K1 cells was carried out using RIPA lysis buffer, followed by separation of protein samples via SDS-PAGE. Subsequent protein transfer onto polyvinylidene fluoride (PVDF) membranes was performed. Blocking with 5% non-fat milk was followed by an overnight incubation of primary antibodies at 4 °C. The membranes were then washed thrice with Tris-Buffered Saline Tween-20 (TBST), incubated with secondary antibodies at room temperature for 1 h, and detection was carried out using the Amersham ImageQuant 800 system (Cytiva, Tokyo, Japan). Following three washes with TBST, the membranes were incubated with secondary antibodies at room temperature for

Table 1 Primers used for RT qPCR in this study

Gene	Forward primer (5′-3′)	Reverse primer (5′-3′)
P4HA2	AAAAGGGCATCGTGGACAGA	CACAGAGGGAAGTGTCTGGG
TBC1D4	CCCGAGAGGTGATCCTGGT	TAACATCAGGAACCTGGCTGG
PRKCQ	CCGTGCTCGCTCCAGG	ACATCTGCCCGTTCTCTGATT
PLSCR4	AGGCCACTCGCAGCTTG	ATAGGCAAGCCTCCTGGGTA
18s	CCTTTGCCATCACTGCCATT	CACACGTTCCACCTCATCCTC

one hour, and detection was conducted using the Amersham ImageQuant 800 system (Cytiva, Tokyo, Japan).

CCK-8 assay for cell proliferation

Healthy log-phase cells were placed into 96-well plates at a concentration of 1×10^5 cells per well. Following a 12-h incubation period to ensure full cell adhesion, the cells were transfected with either overexpression plasmids or siRNA. Multiple replicates were prepared for each experimental group, which included the following: Vector (empty vector control), OE (*P4HA2* overexpression group), si-NC (nonspecific siRNA control group), and si-*P4HA2* (*P4HA2*-specific siRNA knockdown group). At 0, 24, 48, 72, and 96 h post-treatment, DMEM medium mixed with CCK-8 reagent (Vazyme, China) at a 10:1 ratio was introduced into each well. Following a 2-h incubation in a humidified incubator, the absorbance of the plates was measured at 450 nm with a Tecan Spark multimode microplate reader from Tecan, Switzerland. The readings recorded at 0, 24, 48, 72, and 96 h were used to evaluate and calculate the cell proliferation capacity.

Transwell-based cell migration studies

A transwell migration assay was employed in this study to evaluate how well cells can migrate. The experimental procedure was as follows: first, cells were washed twice with phosphate-buffered saline (PBS) to remove surface residues. Trypsin was used to digest the cells, creating a suspension of single cells. To eliminate the influence of cell proliferation on migratory capacity, cells were resuspended in a serum-free medium. Cell density was carefully calibrated to 5×10^5 cells/mL. The assay employed transwell chambers with an 8 μm pore size in 24-well plates from Corning. Before the experiment, the chambers were hydrated with serum-free DMEM medium for 30 min to ensure membrane moisture. Following hydration, the medium was aspirated, and 200 μL of cell suspension without serum was introduced into the upper chamber. Concurrently, 600 μL of complete medium with 10% FBS was added to the lower chamber. In a 5% CO_2 incubator set at 37 °C, cells were incubated for 48 h to encourage migration through the transwell membrane, with migrated cells attaching to the membrane's underside. After the experiment concluded, the transwell chambers were taken out and carefully rinsed with PBS. The cells were then fixed using 4% paraformaldehyde for 15 min to maintain their structure. Following fixation, the cells were stained with 0.1% crystal violet for 15 min to facilitate observation under a microscope. The upper surface of the membrane was gently wiped with a cotton swab to remove excess crystal violet, and the chambers were air-dried. The underside of the transwell membrane was then observed using an inverted microscope, with five random fields photographed and counted. Each experimental group was conducted three times to guarantee the reliability of the results. This rigorous protocol allowed for an accurate assessment of cell migratory ability, providing essential data for subsequent biological research.

Wound healing assay

During this experiment, cells were placed into 6-well plates. Overnight incubation of the plates at 37 °C with 5% CO_2 allowed the cells to reach full confluence. A straight scratch was made across the cell monolayer using a sterile 200 μL pipette tip, creating a uniform wound. The wells were gently washed three times with PBS to remove floating cells and debris. Serum-free medium was then added to eliminate the influence of cell proliferation on migration results.

According to the experimental design, the following groups were established: Vector (empty vector control), OE (*P4HA2* overexpression), si-NC (nonspecific siRNA control), and si-*P4HA2* (*P4HA2*-specific siRNA knockdown). Migration at the scratch site was documented at 0 h and 24 h (or at specified time points) using an inverted microscope to capture images of the same field of view. Each experiment was performed in triplicate to ensure reliability. The scratch closure was quantified using ImageJ software by measuring changes in the wound area. The cell migration rate was calculated using the following formula:

$$\text{Migration rate} = \frac{\text{Initial wound area} - \text{final wound area}}{\text{Initial wound area}} \times 100\%.$$

This procedure allowed for accurate assessment of cell migration and provided a reliable basis for further analysis.

In vitro colony forming assay

In this experiment, cells were extracted from the logarithmic growth phase and treated with trypsin–EDTA solution for digestion. The cells were then dispersed into a single-cell suspension using gentle pipetting. The cell suspension was subsequently diluted to 500 cells per well and seeded into 6-well plates, ensuring adequate space for individual colonies to form. The plates were gently shaken to achieve uniform cell distribution. The plates were incubated in a 37 °C, 5% CO_2 environment for a period of 7 to 14 days. During the cultural period, the formation of colonies was regularly monitored. Once the colonies reached a size visible to the naked eye, fixation was performed.

The process began by removing the medium and gently washing the cells with PBS buffer. Next, 4% paraformaldehyde was introduced into each well to fix the cells at

room temperature for 10 to 15 min. After the fixation process was finished, the cells were rinsed again with PBS buffer, followed by the addition of a 0.1%crystal violet staining solution for 10 to 30 min at room temperature. Following staining, the cells were carefully rinsed with PBS to eliminate any surplus dye until the solution became clear. After the colonies dried, they were examined under a microscope, and the colonies were counted. This process could be done manually or automated using ImageJ analysis software.

Immunofluorescence

In the experiment, cells were used to prepare a suspension with a concentration of 5×10^5 cells/mL. A suitable amount of culture medium was added to the position on a 24-well plate where the coverslips would be placed. The coverslips were gently placed in the wells, and cells were seeded onto the coverslips. Once the cells reached an appropriate density, the coverslips were removed and fixed with 4% paraformaldehyde for 15 min. After fixation, the coverslips were washed with PBS buffer and blocked with 500 μL of blocking solution for 60 min. During this time, the primary antibody solution was prepared using antibody diluent.

Once blocking was complete, the coverslips were rinsed with PBS buffer and left with the primary antibody at 4 °C overnight. The next day, they were washed three times with PBS for 5 min each to eliminate any unbound primary antibody. Then, an Alexa Fluor 488-conjugated secondary antibody (Thermo Fisher Scientific, USA) was added and incubated at room temperature for 1 h in the dark. After incubation, the coverslips were washed three times with PBS for 5 min each to remove any unbound secondary antibody. Ultimately, the coverslips were affixed using an anti-fade mounting medium that included DAPI (ProLong™ Diamond, Invitrogen), taking care to avoid the formation of any air bubbles during the procedure.

Fluorescent signals were detected and recorded using a confocal microscope (NIKON, Japan) at the appropriate wavelengths. The fluorescence intensity or the proportion of positive cells was measured using ImageJ software to evaluate the expression levels of the cell markers.

Statistical analysis

Data are expressed as mean ± standard deviation (SD). Statistical analyses were conducted using R software (R version 4.4.0) and GraphPad Prism 9.5. Comparisons between the two groups were conducted using a Student's t-test. To control the false discovery rate and account for multiple testing, we applied the Benjamini–Hochberg method for multiple test correction. In this study, a P-value < 0.05 was considered statistically significant.

Result

Identification of TC-correlated genes with DEGs and WGCNA

During the analysis of the TCGA-THCA dataset, we identified DEGs in TC samples ($n=512$) and normal samples ($n=59$). To visually represent these DEGs, we constructed a volcano plot (Fig. 1A). Comprehensive data are presented in Supplementary Table S1. Additionally, we conducted WGCNA analysis on the complete transcriptome dataset.

In the initial stage of WGCNA, we constructed a sample clustering dendrogram to identify and remove outlier samples (Fig. 1B). To convert Pearson correlation coefficients into a weighted adjacency matrix, we determined an optimal soft-thresholding exponent (β) to make the network topology more closely align with a scale-free distribution (Fig. 1C). Using the adjacency matrix, we calculated the topological overlap matrix and generated a gene dynamic tree cut dendrogram (Fig. 1E).

Subsequently, we identified the module most strongly correlated with TC and normal samples, the red module (correlation coefficient = 0.66, p < 3e−49) (Fig. 1D, F). Ultimately, a comparative analysis was conducted involving the differentially expressed genes (DEGs) ($n=12{,}764$) and the genes within the red module ($n=1431$) (Fig. 1G), resulting in the identification of 1359 common genes. These genes, listed in detail in Table S2, will be further investigated in subsequent studies.

Identification of key genes associated with TC

In this investigation, we performed a comprehensive analysis of the PPI network involving the 1,359 overlapping genes to pinpoint the central hub genes. These hub genes were identified based on their substantial interaction connectivity, which suggests their critical regulatory roles within the network. Through the construction of the PPI network, we identified 291 hub genes characterized by high interaction frequency, suggesting their potential roles as core regulators within the network (Fig. 2A). These key nodes can be seen in Supplementary Table S3.

To narrow down the list of candidate genes, we utilized three machine-learning techniques for feature extraction. First, the XGBoost method, an ensemble learning method based on gradient-boosted decision trees, was used to assess feature importance, resulting in the identification of 42 genes with significant contributions (Fig. 2B). Second, we applied the LASSO regression algorithm, which performs feature selection through L1 regularization and successfully identified 33 genes (Fig. 2C, D). Lastly,

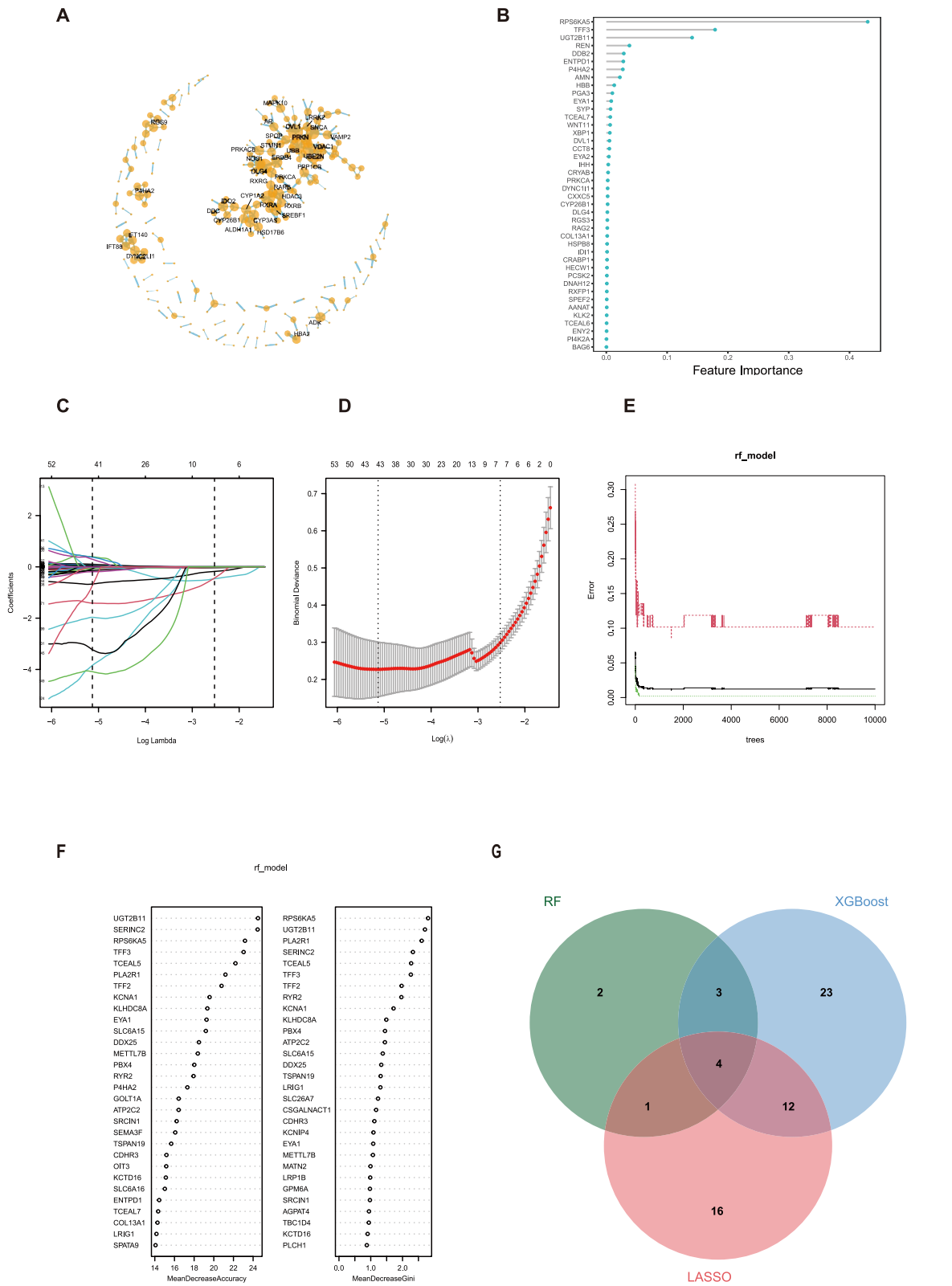

Fig. 2 Advanced analytics of key genes and machine learning insights. **A** Protein–protein interaction (PPI) network analysis of genes identified in Fig. 1F intersection. **B** Feature importance derived from XGBoost model. **C** Coefficient path plot from LASSO regression analysis. **D** Cross-validation curve from LASSO regression analysis. **E** Error curve of Random Forest model. **F** Feature importance from Random Forest model. **G** Intersection of hub genes from PPI network and genes identified by machine learning models

the RF algorithm, an ensemble method combining multiple decision trees to improve predictive accuracy, was utilized. Based on feature importance scores, we selected the top 10 genes with scores greater than 25 as key feature genes (Fig. 2E, F). These genes are presented in Supplementary Table S4-6.

By integrating the results from these three machine learning methods, we identified four genes as the final consensus feature genes (*TFF3*, *EYA1*, *RPS6KA5*, and *P4HA2*) (Fig. 2G). These genes were selected not only based on their high interaction frequency in the PPI network, but also on their importance scores across multiple machine learning models. These findings highlight the pivotal roles of these genes as regulatory hubs within the network, offering new insights into the molecular mechanisms underlying network regulation. Moreover, these genes represent potential targets for future biological research and clinical applications.

Of the four hub genes, *TFF3*, *EYA1*, and *RPS6KA5* were lowly expressed in cancerous tissues, whereas *P4HA2* demonstrated upregulated expression in the cancer tissues (Fig. 3A). To assess the diagnostic efficacy of each gene, we validated their performance within the TCGA-THCA dataset, revealing that the AUC for each gene surpassed 0.95 (Fig. 3B), indicating their potential for effective diagnosis of TC. Furthermore, in the external validation using the GSE29265 and GSE33630 datasets, we consistently observed the downregulation of *TFF3*, *EYA1*, and *RPS6KA5*, and the upregulation of *P4HA2* in thyroid cancer tissues (Fig. 3C, D). These results highlight the possibility of using these genes as diagnostic markers for TC, with their expression patterns robustly associating with disease status across independent datasets.

Development and validation of a prognostic risk model for thyroid cancer

Subsequently, we constructed a thyroid cancer risk model based on the four hub genes and calculated the risk score using a multivariate Cox regression model (Fig. S1A). The risk score formula is as follows: 0.0001230044**TFF3* + 0.0094348493**P4HA2* + 0.7060780955**RPS6KA5* + 2.5469380036**EYA1*. Figure S1B shows that the high-risk group had significantly worse survival outcomes (P = 0.00058). The time-dependent ROC curves for 1-year, 3-year, and 5-year survival demonstrated AUC values of 0.708, 0.799, and 0.682, respectively (Fig. S1C). Multivariate Cox regression analysis indicated that age is an independent prognostic factor for thyroid cancer, while the risk score did not qualify as an independent prognostic factor (Fig. S1D). Furthermore, a nomogram model integrating clinical features and the risk score was developed, which exhibited a high diagnostic capacity with an AUC value of 0.946 (Fig. S1E and Fig. S1F). Although the risk score based on these four genes effectively assessed survival outcomes, it could not serve as an independent prognostic factor.

Elevated expression of P4HA2 in TC cell lines

To validate the bioinformatics-derived expression patterns, we employed RT qPCR and Western blot techniques to evaluate the expression of *P4HA2* at both mRNA and protein levels in normal thyroid cells (Nthy-ori3-1) as well as in nine different thyroid cancer cell lines. Our findings revealed markedly elevated mRNA and protein levels of *P4HA2* in FTC238, TPC-1, and K1 cell lines in comparison to Nthy-ori3-1 cells (Fig. 4A, B). As a result, these cell lines were chosen for additional in vitro experiments.

Optimization of P4HA2 knockdown and overexpression efficiency

To explore the optimal experimental conditions for subsequent research, we conducted transient transfection experiments using the TPC-1 cell line. The goal was to screen for specific siRNAs and determine the optimal working concentration of the *P4HA2* overexpression plasmid. Western blotting was utilized to assess the knockdown efficiency of various siRNAs, which showed that si-*P4HA2*-2 had the highest knockdown efficiency (Fig. 4C). Moreover, results from both Western blot and immunofluorescence assays confirmed that at a transfection concentration of 1 μg/mL, the *P4HA2* overexpression plasmid achieved more effective protein overexpression in the TPC-1 cell line (Fig. 4D, E). These findings provide critical parameters and conditions for future experiments.

Knockdown of P4HA2 inhibits TC cells proliferation in vitro

To delve deeper into the function of *P4HA2* in TC cell lines, we utilized si-*P4HA2*-2 to transiently knock down the level of expression of *P4HA2* in FTC238, TPC-1, and K1 cells. The knockdown efficiency was confirmed at both mRNA and protein levels through RT qPCR, Western blot, and immunofluorescence analyses (Fig. 5A–C). Subsequent CCK-8 and colony formation assays showed a notable decrease in the rate of proliferation of FTC238, TPC-1, and K1 cells with *P4HA2* knockdown. Moreover, the number of colonies formed by cells with *P4HA2* deficiency was significantly fewer compared to control cells under identical culture conditions (Fig. 5D, E). These findings suggest that *P4HA2* knockdown effectively suppresses the in vitro proliferative capacity of TC cell lines.

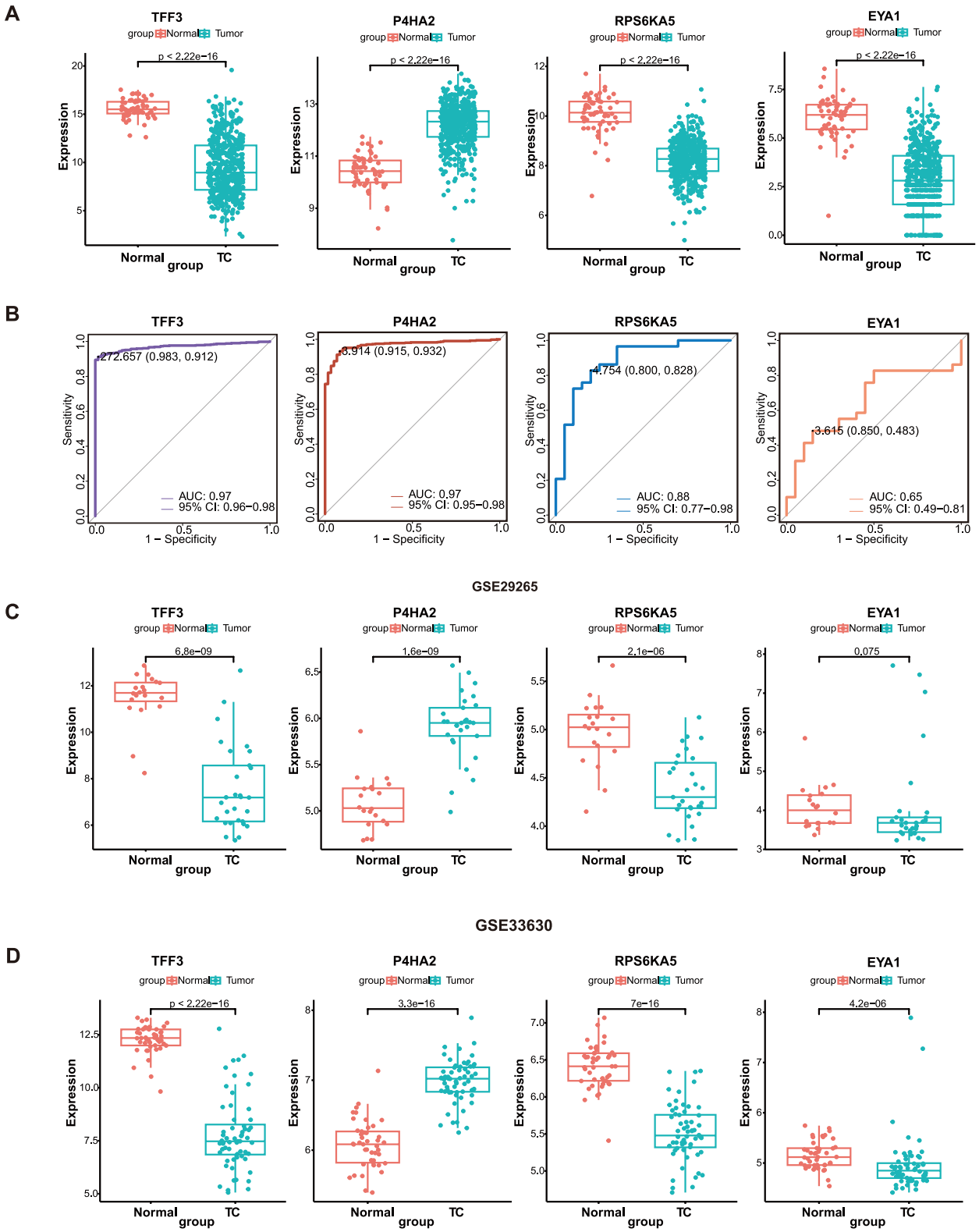

Fig. 3 Expression levels of the four genes in the training set. **A** receiver operating characteristic (ROC) analysis of the core genes in the training set. **B** The area under the ROC curve (AUC) for key gene markers. **C** Expression levels of the four genes in GSE29265. **D** Expression levels of the four genes in GSE33630

Knockdown of P4HA2 inhibits TC cell migration in vitro

To investigate the effects of *P4HA2* knockdown on additional biological activities of TC cells, we carried out scratch wound healing and transwell migration tests. In the scratch wound closure experiment, transient knockdown of *P4HA2* significantly impaired the migratory ability of FTC238, TPC-1, and K1 cell lines in vitro (Fig. 6A). Quantitative analysis of the wound closure area and migration speed demonstrated a significant association between *P4HA2* expression levels and the migration ability of TC cells. These results were further corroborated by transwell migration assays, which demonstrated that *P4HA2-knockout cell lines* exhibited a substantially reduced migration capacity under conditions simulating in vivo migration (Fig. 6B). Collectively, these findings indicate that *P4HA2* knockdown effectively inhibits the migratory capacity of TC cell lines in vitro.

Overexpression of P4HA2 promotes TC cell proliferation and migration in vitro

To further verify the impact of *P4HA2* expression on the biological characteristics of TC cells, we employed a transient transfection strategy to introduce a *P4HA2* overexpression plasmid into FTC238, TPC-1, and K1 cell lines. Successful overexpression of *P4HA2* at the mRNA and protein levels was validated through RT qPCR, Western blot, and immunofluorescence assays (Fig. 7A–C). Subsequent CCK-8 and colony formation experiments showed that the overexpression of *P4HA2* markedly increased the proliferation rate and colony formation capability of FTC238, TPC-1, and K1 cells relative to the control groups (Fig. 7D, E). Additionally, scratch wound closure and transwell migration experiments indicated that increased *P4HA2* expression substantially enhanced the migration ability of TC cells in vitro (Fig. 8A, B). Taken together, these results demonstrate that *P4HA2* overexpression markedly enhances the in vitro proliferation and migration capabilities of TC cells.

GO, KEGG and immune cell infiltration in P4HA2-HIGH

To clarify the mechanism of *P4HA2* in TC, we divided the TC tissues in TCGA-THCA into two groups, *P4HA2*-Low and *P4HA2*-High, according to the expression of *P4HA2*. To explore the correlation between *P4HA2* and the TC immune microenvironment, we performed CIBERSORT algorithm on *P4HA2*-Low and *P4HA2*-High groups. The findings revealed that, in contrast to the *P4HA2*-low group, the *P4HA2*-high group showed markedly higher infiltration levels of resting dendritic cells, M0 macrophages, M2 macrophages, resting mast cells, and regulatory T cells (Tregs). In contrast, the infiltration levels of memory B cells, M1 macrophages, neutrophils, plasma cells, and CD8 + T cells were significantly reduced (Fig. 9A). Furthermore, *P4HA2* expression levels were positively correlated with resting dendritic cells, regulatory T cells, M0 and M2 macrophages, while negatively correlated with CD8 + T cells, CD4 + T cells, natural killer (NK) cells, and M1 macrophages (Fig. 9B). To provide a more comprehensive view, we have meticulously annotated the correlations between *P4HA2* and various immune cells, complemented by their respective P-values (Fig. 9C).

Next, we analyzed the DEGs of the *P4HA2*-High and *P4HA2*-Low groups using the R package "Limma". Additionally, these DEGs were subjected to GO and KEGG enrichment analysis. The KEGG analysis further indicated that the DEGs were enriched in pathways related to cancer progression, such as *bacterial invasion of epithelial cells*, *adherens junctions*, *lysosome function*, *tight junctions*, *proteoglycans in cancer*, and various bacterial infection pathways. These pathways, which encompass cell interactions, signal transduction, and intracellular structural regulation, indicate a significant role in facilitating cancer development and metastasis (Fig. 9D). The GO analysis indicated that the high-expression cohort of the *P4HA2* gene was enriched in several oncogenic pathways, which are integral to critical biological processes such as intercellular signaling, protein functionality, and organelle trafficking. Notably, the pathways implicated include those associated with cell–matrix interactions, such as focal adhesion and cell–matrix binding, which are pivotal for cellular adhesion, migration, and invasion—parameters intrinsically linked to tumorigenesis and cancer spread (Fig. 9E). The enrichment of pathways related to the lysosomal membrane and endoplasmic reticulum further underscores the significance of intracellular membrane structures in the context of material transport, degradation, and signal transduction within the cell. Additionally, the enrichment of pathways involving cytoplasmic small GTPase binding and GTPase binding

(See figure on next page.)

Fig. 4 *P4HA2* functions as an oncogene in thyroid cancer. **A** RT qPCR analysis quantifies *P4HA2* mRNA expression in thyroid cancer cell lines compared to normal thyroid cells. **B** Western blot analysis evaluates *P4HA2* protein expression in thyroid cancer cell lines and normal thyroid cells. **C** Western blot analysis confirms the knockdown efficiency of *P4HA2* using siRNA in TPC-1 cells. **D** Western blot analysis examines the efficiency of *P4HA2* overexpression across a concentration gradient in TPC-1 cells. **E** Immunofluorescence (IF) analysis validates the efficiency of *P4HA2* overexpression across a concentration gradient in TPC-1 cells. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$ (scar bar = 100 μm)

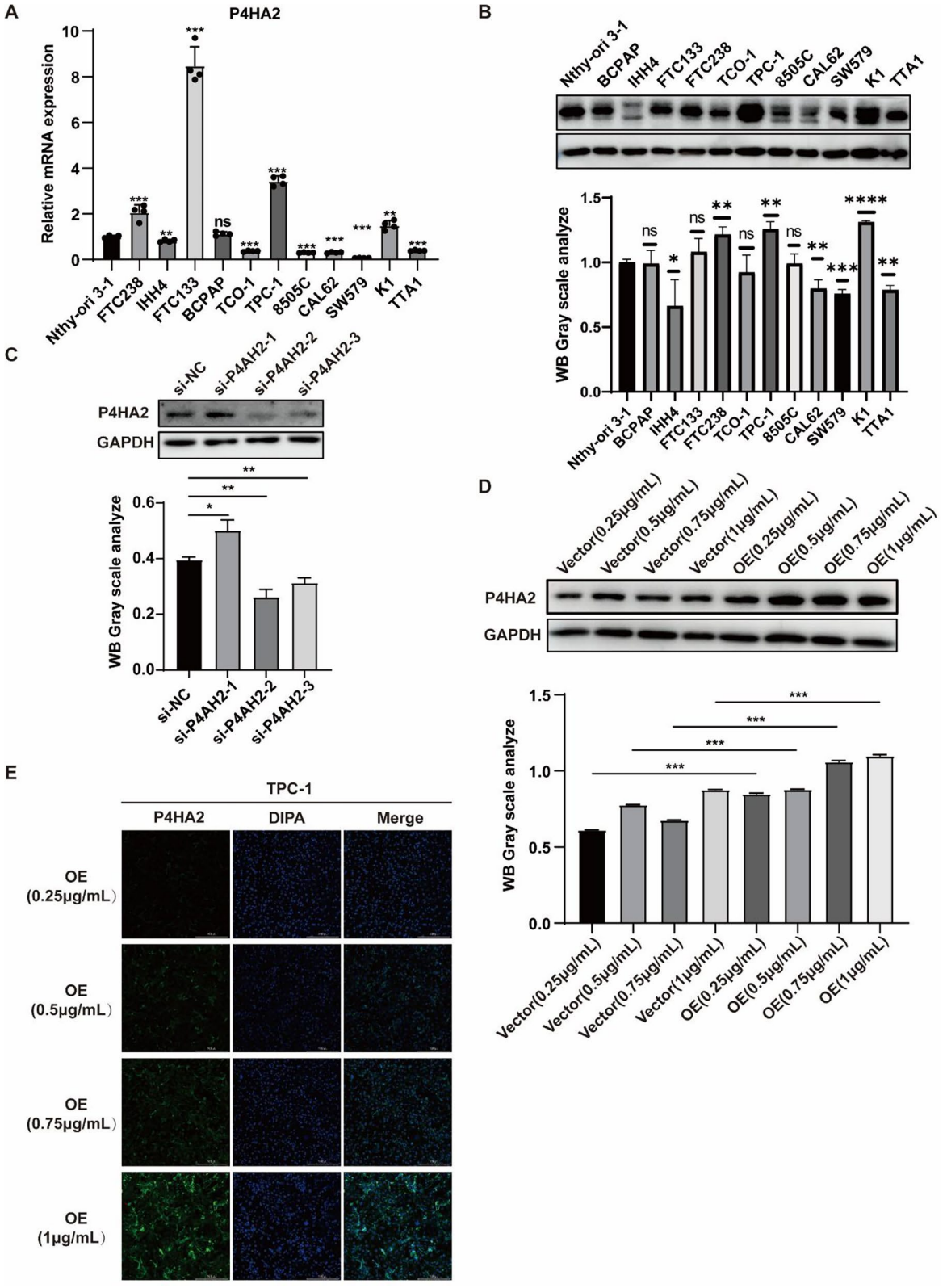

Fig. 4 (See legend on previous page.)

suggests an active role in cell signaling, potentially tied to proliferative and metastatic signaling cascades. Together, these pathways offer valuable insights into the potential role of *P4HA2* in oncogenic mechanisms. The relevant details can be found in Supplementary Table S7, S8.

Discussion

Main interpretation

Thyroid cancer is a malignant neoplasm originating in the thyroid gland, and its global incidence has markedly increased in recent years. According to recent data from the International Agency for Research on Cancer (IARC) and GLOBOCAN 2022 statistics, thyroid cancer ranks as the seventh most prevalent malignancy worldwide and the fifth most common among women [6, 30, 31]. Notably, in 2022, China accounted for an estimated 466,000 new cases of thyroid cancer, representing more than half of the global total [32, 33].

Although fine-needle aspiration biopsy (FNAB) under ultrasound guidance is currently the primary method for assessing thyroid nodules, its diagnostic accuracy and consistency are contingent upon the operator's experience and technical proficiency, resulting in inter-individual variability in interpretation [34–36]. Furthermore, early diagnosis remains challenging due to the small size, concealed location, or atypical pathological features of certain lesions. When FNAB results are inconclusive, additional molecular marker testing may be necessary to support the diagnosis [37]. Currently, the molecular markers most frequently utilized in clinical practice include BRAF mutations, RAS mutations, RET/PTC rearrangements, and PAX8/PPARγ fusion genes [38–40]. Among these, The BRAF V600E mutation is predominantly associated with PTC, with meta-analyses reporting a specificity of up to 99%. However, its sensitivity is relatively low, approximately 60%, indicating that not all thyroid cancer patients exhibit BRAF mutations [41]. While the BRAF V600E mutation is closely linked to PTC, it has a limited association with other thyroid cancer subtypes, such as follicular and medullary carcinomas [42]. Therefore, the BRAF V600E mutation cannot be considered a universal biomarker for all thyroid cancer types, and its clinical utility is inherently restricted [43, 44]. RAS mutations, on the other hand, display complex biological mechanisms and have low diagnostic sensitivity and specificity. These mutations are more frequently observed in follicular thyroid carcinoma (FTC) than in PTC [45]. Furthermore, given that RAS mutations occur across a variety of cancer types, their specificity is inadequate for use as a standalone diagnostic tool [46]. RET/PTC rearrangements are detected in about 10–20%of PTC cases, suggesting that a substantial proportion of PTC patients do not exhibit this genetic alteration, thus limiting its effectiveness as a universal biomarker [47, 48]. Additionally, RET/PTC rearrangements are more commonly detected in radiation-induced PTC, further restricting their relevance as biomarkers for PTC cases not associated with radiation exposure [49]. Some studies have also suggested a correlation between RET/PTC rearrangements and multifocality in thyroid cancer, potentially reducing their diagnostic utility as standalone markers [50, 51].

In recent years, considerable research has been focused on discovering new biomarkers for thyroid cancer. Multi-omics investigations have uncovered precise molecular markers. For example, Montero-Conde et al. [52]. identified new prognostic markers related to chromatin spatial organization at the 5pter and TERT loci through RNA sequencing of 106 tumor samples, establishing TRER and TREC as independent prognostic indicators. Similarly, Shi et al. employed exome-wide sequencing, RNA profiling, DNA methylation analysis arrays, proteomics, and phosphoproteomics to develop a comprehensive multi-omics atlas of 102 medullary thyroid carcinoma (MTC) samples. The study identified novel driver genes, including BRAF and NF1, and delineated three molecularly heterogeneous subtypes of medullary thyroid carcinoma (MTC) through proteomics-based stratification. Additionally, two members of the tenascin family, TNC and TNXB, emerged as potential prognostic biomarkers for MTC [53]. As high-throughput molecular biology technologies advance rapidly, there is a growing emphasis on the role of non-coding RNAs in thyroid cancer research. These molecules, encompassing microRNAs (miRNAs) and circular RNAs (circRNAs), are characterized by enriched expression and stability, rendering them advantageous for biomarker development [54]. CircRNAs, in particular, have demonstrated potential as biomarkers;

(See figure on next page.)

Fig. 5 Knockdown of P4HA2 inhibits the in vitro proliferation of TC cells. **A** Western blot analysis was used to evaluate the knockdown efficiency of *P4HA2* protein in FTC238, TPC-1, and K1 cells. **B** RT qPCR analysis assessed the knockdown efficiency of *P4HA2* mRNA in FTC238, TPC-1, and K1 cells. **C** Immunofluorescence (IF) analysis confirmed the knockdown efficiency of *P4HA2* in FTC238, TPC-1, and K1 cell lines (scar bar = 100μm). **D** The CCK-8 assay was performed to evaluate the effect of *P4HA2* knockdown on the proliferation rate of FTC238, TPC-1, and K1 cells in vitro. **E** Colony formation assays measured changes in colony formation capacity in FTC238, TPC-1, and K1 cells following *P4HA2* knockdown. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$

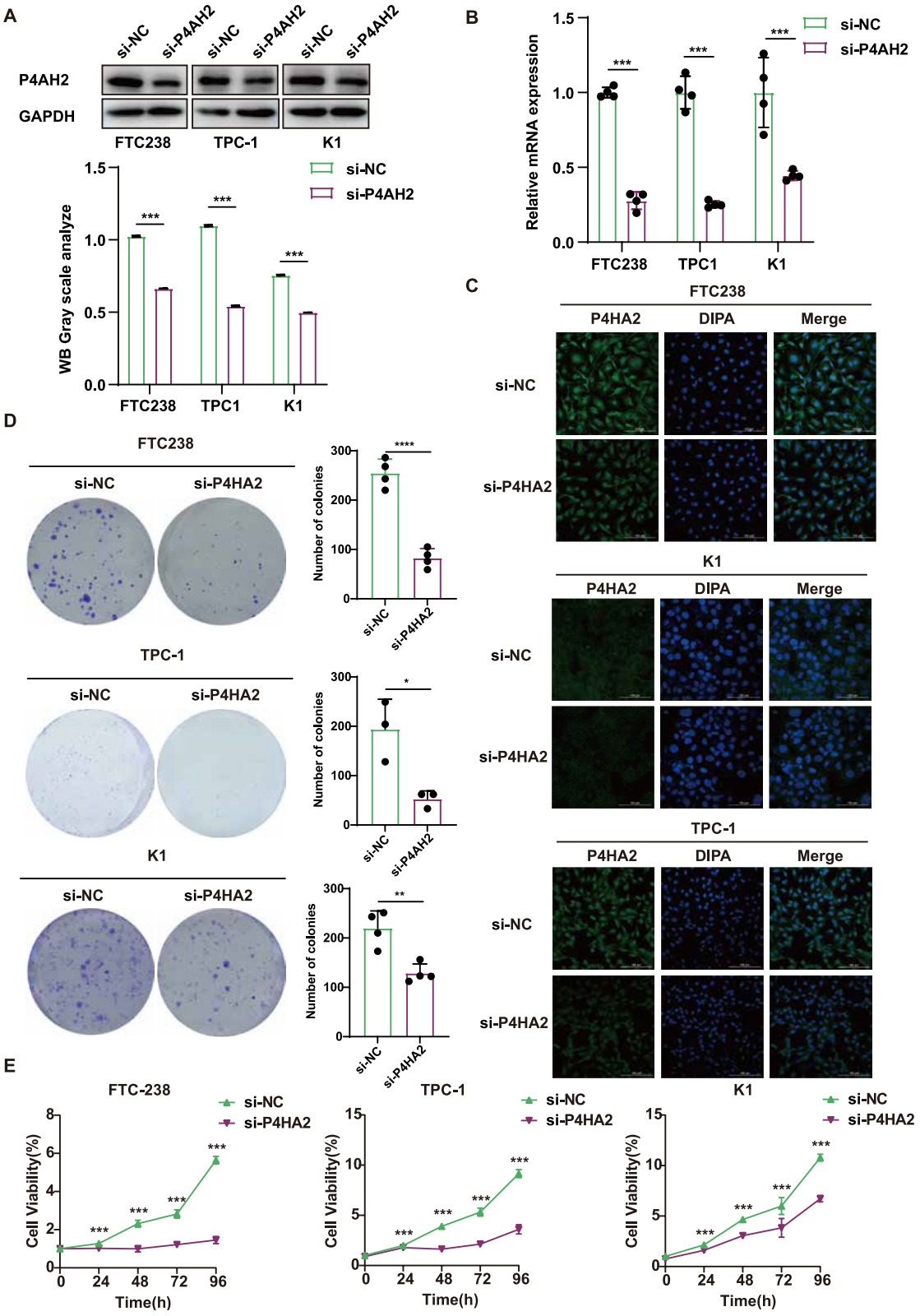

Fig. 5 (See legend on previous page.)

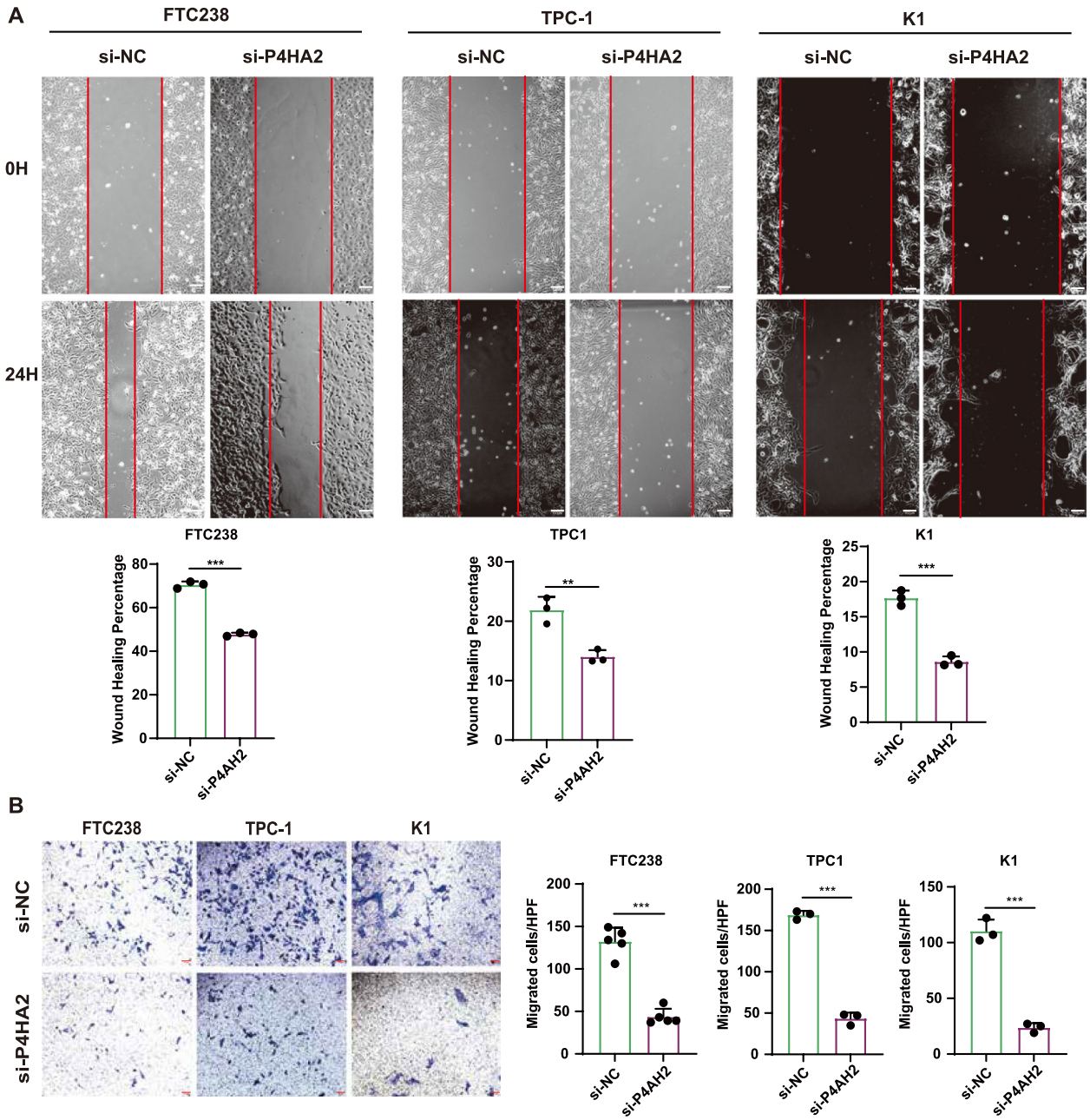

Fig. 6 Knockdown of *P4HA2* inhibits the in vitro migration of TC cells. **A** Wound healing assays were performed to evaluate alterations in the migration capacities of FTC238, TPC-1, and K1 cells following *P4HA2* knockdown. **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001 (scar bar = 100 μm). **B** Transwell migration assays were used to assess the impact of *P4HA2* knockdown on the migratory abilities of FTC238, TPC-1, and K1 cell lines (scar bar = 100 μm)

(See figure on next page.)

Fig. 7 Overexpression of *P4HA2* promotes the in vitro proliferation of TC cells. **A** Western blot analysis was used to evaluate the overexpression efficiency of *P4HA2* protein in FTC238,TPC-1, and K1 cells. **B** RT qPCR analysis assessed the overexpression efficiency of *P4HA2* mRNA in FTC238,TPC-1,and K1 cells. **C** Immunofluorescence(IF)analysis confirmed the overexpression efficiency of *P4HA2* in FTC238,TPC-1,and K1 cell lines (scar bar = 100μm). **D** The CCK-8 assay was performed to evaluate the effect of *P4HA2* overexpression on the proliferation rate of FTC238,TPC-1,and K1 cells in vitro. **E** Colony formation assays measured changes in colony formation capacity in FTC238, TPC-1, and K1 cells following *P4HA2* overexpression. **P* < 0.05, ***P* < 0.01, ****P* < 0.001, *****P* < 0.0001

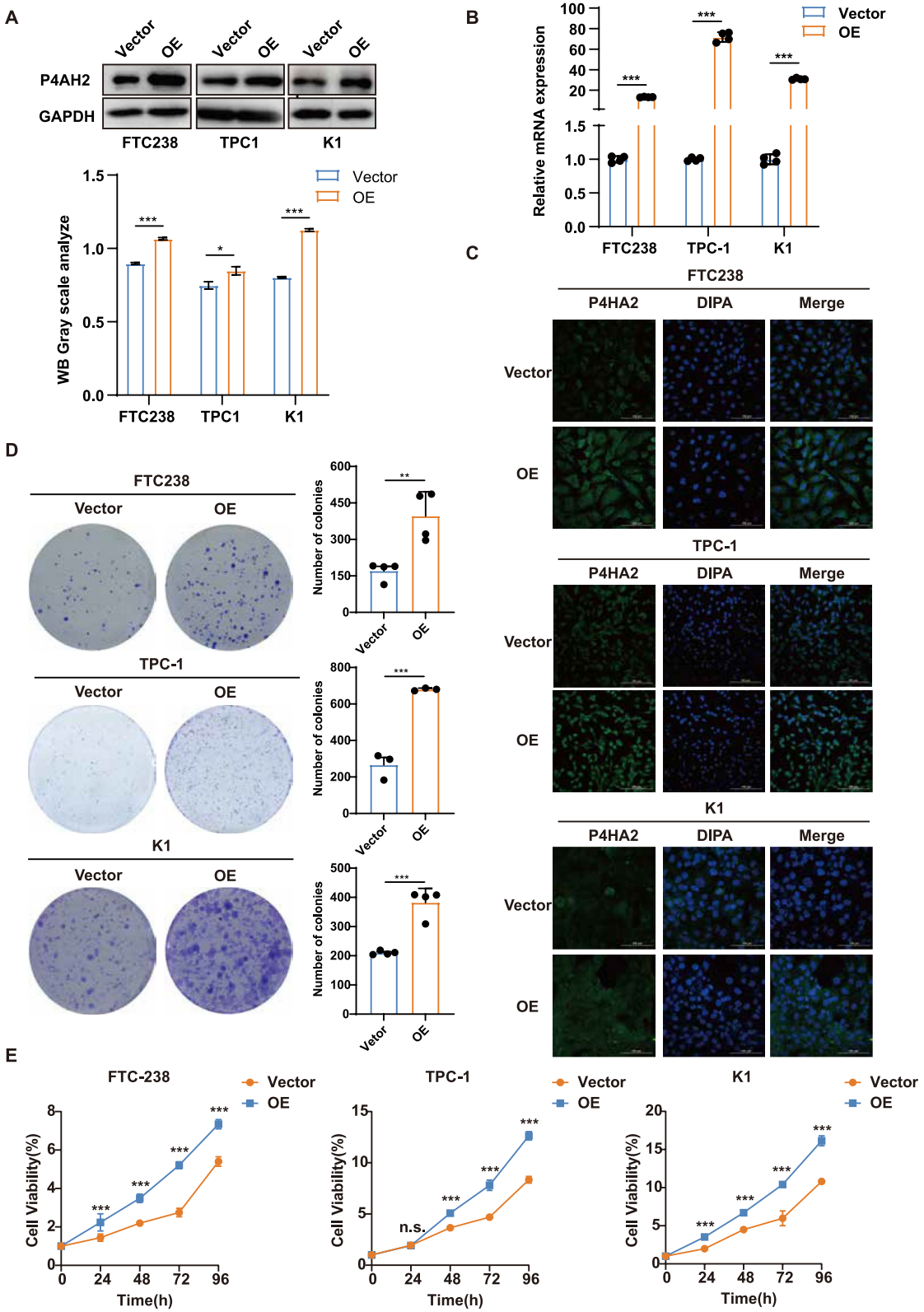

Fig. 7 (See legend on previous page.)

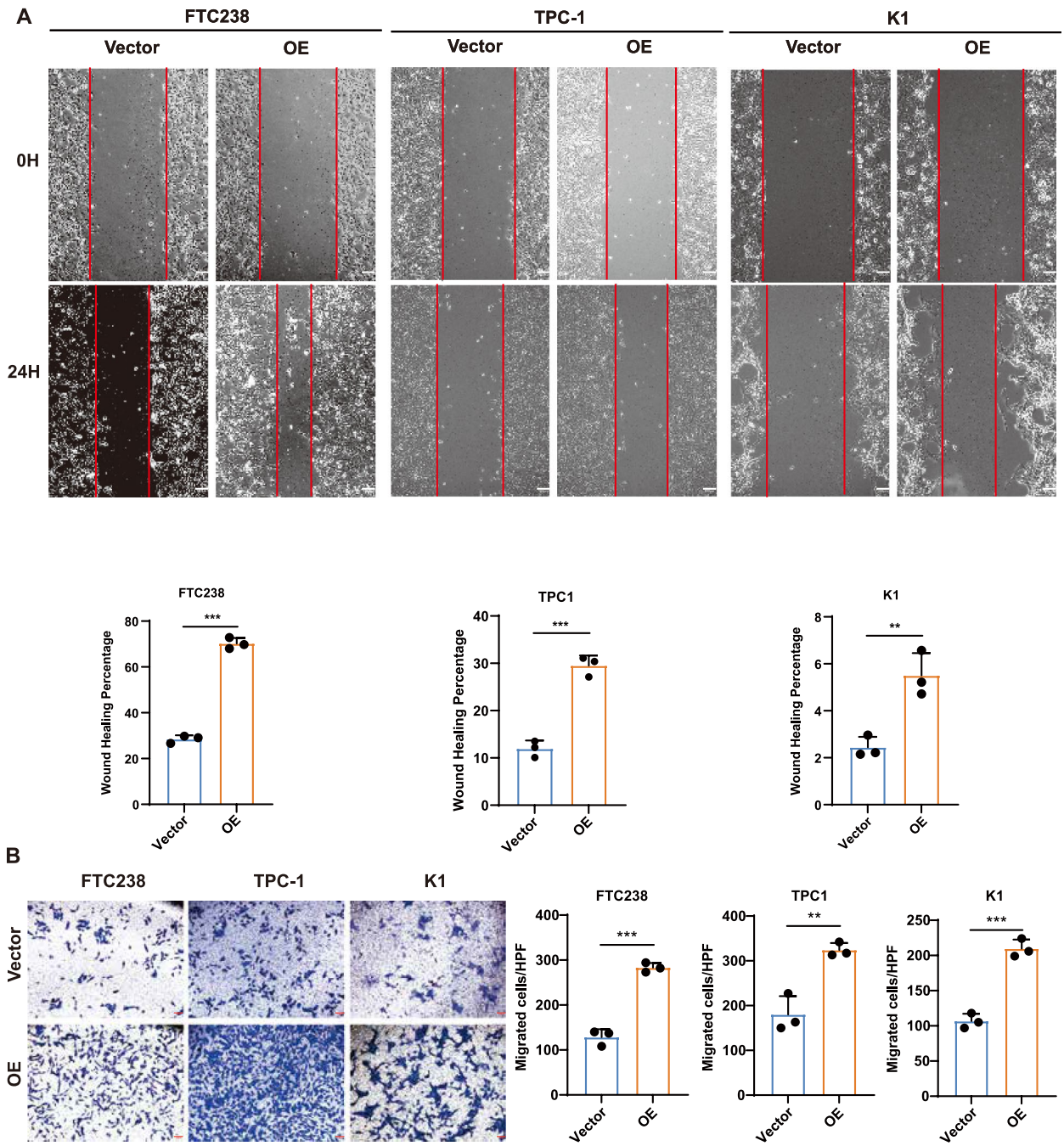

Fig. 8 Overexpression of *P4HA2* promotes the in vitro migration of TC cells. **A** Scratch wound healing assays were performed to assess changes in the migration capacity of FTC238, TPC-1, and K1 cells following *P4HA2* overexpression. *$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, ****$P < 0.0001$ (scar bar = 100μm). **B** Transwell migration assays were conducted to evaluate the effect of *P4HA2* overexpression on the migration ability of FTC238, TPC-1, and K1 cells

for instance, circFAT1 and circRNA_102171 have been associated with the proliferation and invasiveness of PTC [55]. Nonetheless, further investigations are required to substantiate their clinical efficacy and reliability. Systematic reviews and meta-analyses have highlighted markers such as HBME-1, Galectin-3 (Gal-3), and Cytokeratin-19 (CK19) as valuable auxiliary molecular markers for PTC [56, 57]. While these markers have shown utility in the diagnosis and treatment of thyroid cancer, their effectiveness is constrained by limitations in sensitivity,

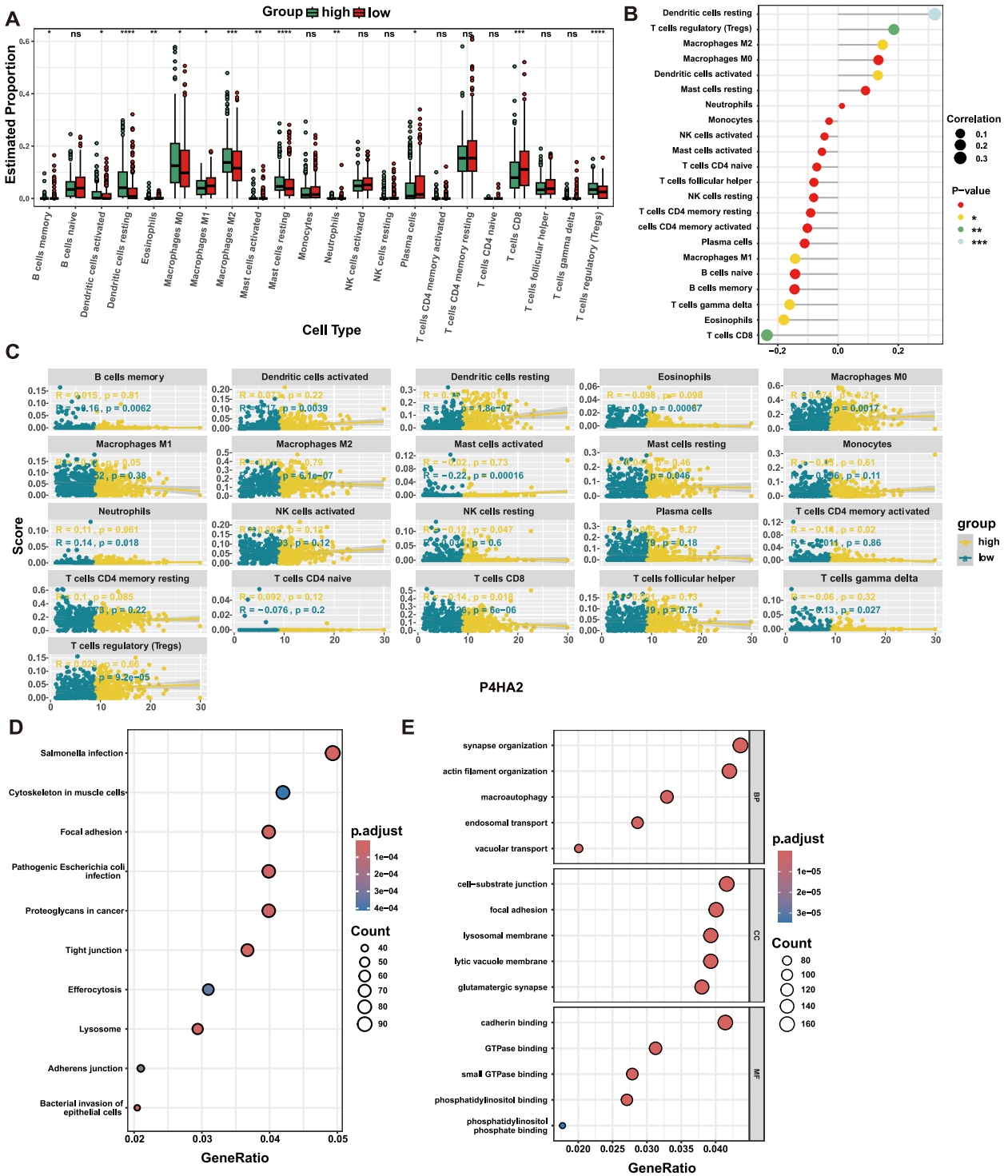

Fig. 9 GO, KEGG and immune infiltration in P4HA2-high group. **A** Immune cell infiltration profiles in high and low P4HA2 expression groups assessed using CIBERSORT. **B** Correlation between P4HA2 expression levels and various immune cell types evaluated through CIBERSORT analysis. **C** Comprehensive associations between P4HA2 expression and individual immune cell populations identified by CIBERSORT. D KEGG analysis. E GO analysis

specificity, and a lack of universal applicability across all thyroid cancer subtypes. This highlights the pressing clinical need for more effective diagnostic tools.

This study employs transcriptome sequencing data and advanced bioinformatics techniques to perform a comprehensive analysis of gene expression profiles in thyroid cancer, with the objective of identifying novel diagnostic and prognostic biomarkers. The initial phase involved extracting pertinent data from the TCGA database and conducting differential expression analysis between thyroid cancer tissues and adjacent normal tissues. Utilizing WGCNA, gene modules most strongly associated with cancerous and non-cancerous groups were identified. Subsequently, the intersection of DEGs and module-specific genes underwent PPI analysis, resulting in the identification of over 200 key node genes with confidence scores exceeding 905. Through the application of three distinct machine learning algorithms, four genes with notable diagnostic and prognostic potential were identified: *TFF3*, *EYA1*, *RPS6KA5*, and *P4HA2*. It was observed that *TFF3*, *RPS6KA5*, and *RPS6KA5* are downregulated in thyroid cancer tissues, whereas *P4HA2* is upregulated. Consequently, *P4HA2* was selected for further investigation. The identification of these biomarkers offers promising prospects for enhancing early molecular diagnosis and gene-targeted therapy in thyroid cancer, potentially improving diagnostic precision and therapeutic specificity.

Recent studies have demonstrated that *P4HA2* is significantly overexpressed in various tumors, thereby accelerating their malignant progression, including in B-cell lymphoma, breast cancer, and prostate cancer [58, 59]. Notably, during tumor progression, abnormal collagen deposition can facilitate the invasion and metastasis of tumor cells [60]. Collagen deposition can alter the physical and chemical properties of the extracellular matrix (ECM), such as its stiffness and porosity. These changes can influence the behavior of tumor cells, including their migration and invasion capabilities. For example, a stiffer ECM can provide more resistance to cell movement, but it can also activate certain signaling pathways that promote cell proliferation and survival [59, 61, 62].

Recent research has found that *P4HA2* promotes collagen deposition and ECM stiffening through the hydroxylation of collagen precursors, playing a crucial role in regulating collagen deposition and signaling pathways in tumor progression. *P4HA2* regulates cancer cell behavior by enhancing collagen synthesis and is closely associated with hypoxia, collagen deposition, glycolysis, and the migration and invasion of cancer cells [63–65]. Studies have shown that a stiffer ECM can enhance tumor cell invasiveness through mechanical signaling (e.g., activation of YAP/TAZ) and promote epithelial–mesenchymal transition (EMT) [66, 67]. For instance, in breast cancer, increased collagen deposition is significantly correlated with elevated tumor stiffness, which activates the integrin–FAK signaling pathway, thereby driving cell migration [68, 69]. In hepatocellular carcinoma, overexpression of *P4HA2* significantly increases collagen deposition, activates the PI3K/AKT/mTOR signaling pathway, and promotes tumor cell proliferation, migration, and invasion [70]. Abnormal collagen deposition can also form a physical barrier that impedes immune cell infiltration and regulates the immune microenvironment.

Overexpression of *P4HA2* reduces the infiltration of $CD8^+$T cells into the tumor stroma and promotes the recruitment of immunosuppressive cells (such as Tregs and M2-type macrophages), thereby inhibiting the antitumor immune response [71]. A study by the team of Zhang Hongtao at Soochow University revealed that *P4HA2* hydroxylates mTOR kinase, enhancing its stability and activating the downstream PI3K/AKT/mTOR signaling axis, which promotes the proliferation of lung adenocarcinoma cells. Knockdown of *P4HA2* in combination with mTOR inhibitors (e.g., AZD-8055) can synergistically inhibit tumor growth [72]. Activation of the PI3K/AKT/mTOR pathway can lead to a series of downstream effects that promote tumor progression. For example, it can upregulate the expression of Cyclin D1, a key cell cycle regulatory protein that drives cells from the G1/G0 phase to the S phase, thereby promoting cell proliferation. Additionally, it can increase the expression of Survivin, an inhibitor of apoptosis protein that inhibits apoptosis and promotes cell survival [73].

In lung adenocarcinoma (LUAD), *P4HA2* activates the mTOR signaling pathway by hydroxylating the key proline residue P2341 in mTOR kinase, which subsequently affects the phosphorylation levels of S6K-T389 and AKT-S473, both of which are crucial for tumor cell growth [72]. In other contexts, a study by the team of Jiang Wei at Fudan University found that *P4HA2* hydroxylates SUFU, a core negative regulator of the Hedgehog signaling pathway, in cancer-associated fibroblasts, promoting its dissociation from the KIF7 complex and thereby activating Hedgehog signaling [74]. This activation leads to the release of paracrine growth factors that promote the malignant proliferation of B-cell lymphoma cells. Knockout of *P4HA2* significantly delays tumor growth in mouse lymphoma models [75]. In diffuse large B-cell lymphoma (DLBCL), *P4HA2* hydroxylates the immune negative regulator Carabin, leading to its ubiquitination and degradation, thereby relieving the inhibition of the Ras/ERK pathway and promoting tumor proliferation [75]. Additionally, studies have shown that *P4HA2* expression levels are closely related to the proliferation and migration capabilities of glioblastoma multiforme (GBM) cells.

Notably, low expression of *P4HA2* in glioma stem cells (GSCs) is associated with prolonged patient survival, suggesting that *P4HA2* may act as a regulatory switch in the transition from GSCs to GBM cells [76].

Nonetheless, the association between *P4HA2* and thyroid cancer remains insufficiently understood. In this study, we demonstrated that *P4HA2* promotes malignant phenotypes, such as proliferation and migration, in thyroid cancer by overexpressing and silencing *P4HA2* in thyroid cancer cells. Additionally, we investigated the potential mechanisms underlying this function through bioinformatics analysis. Utilizing immune infiltration analysis, we confirmed that anti-tumor immunity is significantly suppressed in thyroid cancer tissues with elevated *P4HA2* expression, as evidenced by the down-regulation of $CD8^+$T cells and M1-type macrophages, alongside the up-regulation of regulatory T cells (Tregs) and M2-type macrophages. Furthermore, GO and KEGG analyses revealed that high *P4HA2* expression promotes extracellular matrix formation, thereby accelerating the malignant progression of thyroid cancer. In conclusion, *P4HA2* appears to facilitate proliferation, invasion, and immune evasion through mechanisms such as enhancing collagen deposition, producing specific collagen subtypes, and interacting with the tumor immune microenvironment.

This study employed transcriptome sequencing data and bioinformatics techniques to conduct a comprehensive analysis of thyroid cancer gene expression profiles, successfully identifying four key biomarkers: *TFF3*, *EYA1*, *RPS6KA5*, and *P4HA2*. Notably, the elevated expression of *P4HA2* contributes to the malignant progression of thyroid cancer, a finding corroborated by experimental validation. *P4HA2* has the potential to be a liquid biopsy biomarker, such as through serum or exosome detection, which could offer a non-invasive approach for early diagnosis and prognosis assessment. However, there are limitations to this study. The lack of in vivo experiments, such as mouse models, limits the comprehensive understanding of the biological functions and mechanisms of *P4HA2* in thyroid cancer development and progression. Future research should focus on conducting in vivo studies to further validate the role of *P4HA2* and explore its potential as a therapeutic target. This research provides novel molecular markers for the diagnosis of thyroid cancer and identifies new targets for its precise treatment.

Limitations

The TCGA database, despite its large sample size, may exhibit biases related to geographical and population selection. Machine learning algorithms, such as XGBoost, LASSO, and RF,utilized for feature selection and model construction, can be influenced by factors like data feature distribution and parameter settings. Additionally, the study lacks in vivo experimental validation.

Abbreviations

TC	Thyroid cancer
ATC	Anaplastic thyroid carcinoma
AUC	Area under the ROC curve
BP	Biological processes
DTC	Differentiated thyroid cancer
ECM	Extracellular matrix
FDR	False discovery rate
FTC	Follicular thyroid carcinoma
DEG	Differentially expressed gene
RIA	Radioimmunoassay
TCGA	The Cancer Genome Atlas
ROC	Receiver operating characteristic
WGCNA	Weighted gene co-expression network analysis
MTC	Medullary thyroid cancer
GEO	Gene Expression Omnibus
LASSO	The least absolute shrinkage and selection operator
IHC	Immunohistochemistry
MCC	Maximum clique centrality
PPI	Protein–protein interaction
RF	Random forest
TOM	Topological overlap matrix
XGBoost	Extreme gradient boosting
GO	Gene ontology
Tregs	Regulatory T cells
IARC	International agency for research on cancer
KEGG	Kyoto encyclopedia of genes and genomes
NK	Natural killer cell
MF	Molecular functions
FNAB	Fine-needle aspiration biopsy
GBM	Glioblastoma multiforme
OSCC	Oral squamous cell carcinoma
CC	Cellular components
WHO	World Health Organization
miRNAs	MicroRNAs
circRNAs	Circular RNAs
Gal-3	Galectin-3
CK19	Cytokeratin-19
GSCs	Glioma stem cells
LUAD	Lung adenocarcinoma

Acknowledgements

We express our gratitude to all individuals who participated in this study.

Author contributions

Gaofeng Hu, Wenyuan Niu and Jiaming Ge: Conceived the original ideas and completed the experimental part of this manuscript, responsible for the collection, organization of the data, and drafting the initial manuscript. Yanyang Liu, Mengjia Li, Huize Shen, Jie Xuan, Jiaming Ge: Contributed to the data collection and data analysis. Yuanqiang Li: Responsible for the overall design and conception of the study.Qinglin Li: Provided financial support for the research project and supervised the progress of the research. Gaofeng Hu, Wenyuan Niu and Jiaming Ge contributed equally to this article, and all authors have approved the final submitted manuscript.

Funding

This work was supported by the Zhejiang Provincial Basic Public Welfare Research Plan (LR24H270001), the National Natural Science Foundation of China (82173346), the National Natural Science Foundation of China (82474317), and the Traditional Chinese Medicine Science and Technology Project of Zhejiang Province (2020ZQ005, 2019ZZ004).

Data availability
The datasets used and analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
The patient data used in this study were derived from public datasets, hence no ethics committee approval or participant consent was required.

Consent for publication
Consent for publication has been obtained from all authors.

Competing interests
The authors declare no competing interests.

Received: 15 January 2025 Accepted: 17 March 2025
Published online: 05 April 2025

References

1. Wu X, et al. MicroRNA: another pharmacological avenue for colorectal cancer? Front Cell Dev Biol. 2020;8:812.
2. Li S, et al. Exosomes: another intercellular lipometabolic communication mediators in digestive system neoplasms? Cytokine Growth Factor Rev. 2023;73:93–100.
3. Hu C, et al. Oleanolic acid induces autophagy and apoptosis via the AMPK–mTOR signaling pathway in colon cancer. J Oncol. 2021;2021:8281718.
4. Zhang X, et al. Research progress on the interaction between oxidative stress and platelets: another avenue for cancer? Pharmacol Res. 2023;191: 106777.
5. Ferlay J et al. Global cancer observatory: cancer today. 2024. https://gco.iarc.who.int/today.
6. Bray F, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2024;74(3):229–63.
7. Liu Y, et al. Radioiodine therapy in advanced differentiated thyroid cancer: resistance and overcoming strategy. Drug Resist Updat. 2023;68: 100939.
8. Chen DW, et al. Thyroid cancer. Lancet. 2023;401(10387):1531–44.
9. Sajisevi M, et al. Evaluating the rising incidence of thyroid cancer and thyroid nodule detection modes: a multinational, multi-institutional analysis. JAMA Otolaryngol Head Neck Surg. 2022;148(9):811–8.
10. Rahman ST, et al. Understanding pathways to the diagnosis of thyroid cancer: are there ways we can reduce over-diagnosis? Thyroid. 2019;29(3):341–8.
11. Liu Y, et al. Molecular mechanisms of thyroid cancer: a competing endogenous RNA (ceRNA) point of view. Biomed Pharmacother. 2022;146: 112251.
12. Shen H, et al. Radioiodine-refractory differentiated thyroid cancer: molecular mechanisms and therapeutic strategies for radioiodine resistance. Drug Resist Updat. 2024;72: 101013.
13. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30.
14. Kang J, et al. LASSO-based machine learning algorithm for prediction of lymph node metastasis in T1 colorectal cancer. Cancer Res Treat. 2021;53(3):773–83.
15. Dai P, et al. Retrospective study on the influencing factors and prediction of hospitalization expenses for chronic renal failure in China based on random forest and LASSO regression. Front Public Health. 2021;9: 678276.
16. Wang C, et al. Deciphering the value of anoikis-related genes in prognosis, immune microenvironment, and drug sensitivity of laryngeal squamous cell carcinoma. Pathol Res Pract. 2025;268: 155849.
17. Wang Y, et al. Prognostic value of CDKN2A in head and neck squamous cell carcinoma via pathomics and machine learning. J Cell Mol Med. 2024;28(9): e18394.
18. Guo Q, et al. The regulatory network and potential role of LINC00973-miRNA-mRNA ceRNA in the progression of non-small-cell lung cancer. Front Immunol. 2021;12: 684807.
19. Cai LH, et al. DDB2 and MDM2 genes are promising markers for radiation diagnosis and estimation of radiation dose independent of trauma and burns. Funct Integr Genomics. 2023;23(4):294.
20. He S, et al. Signatures of 4 autophagy-related genes as diagnostic markers of MDD and their correlation with immune infiltration. J Affect Disord. 2021;295:11–20.
21. Sun M, et al. A nine-lncRNA signature predicts distant relapse-free survival of HER2-negative breast cancer patients receiving taxane and anthracycline-based neoadjuvant chemotherapy. Biochem Pharmacol. 2021;189: 114285.
22. Dom G, et al. A gene expression signature distinguishes normal tissues of sporadic and radiation-induced papillary thyroid carcinomas. Br J Cancer. 2012;107(6):994–1000.
23. Tomás G, et al. A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic. Oncogene. 2012;31(41):4490–8.
24. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
25. Bardou P, et al. jvenn: an interactive Venn diagram viewer. BMC Bioinform. 2014;15(1):293.
26. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
27. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol). 1996;58(1):267–88.
28. Chen, T. and C. Guestrin. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International conference on knowledge discovery and data mining. 2016. San Francisco: ACM.
29. Newman AM, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12(5):453–7.
30. Cao W, et al. Comparative study of cancer profiles between 2020 and 2022 using global cancer statistics (GLOBOCAN). J Natl Cancer Cent. 2024;4(2):128–34.
31. Pizzato M, et al. The epidemiological landscape of thyroid cancer worldwide: GLOBOCAN estimates for incidence and mortality rates in 2020. Lancet Diabetes Endocrinol. 2022;10(4):264–72.
32. Xia C, et al. Cancer statistics in China and United states, 2022: profiles, trends, and determinants. Chin Med J (Engl). 2022;135(5):584–90.
33. Han B, et al. Cancer incidence and mortality in China, 2022. J Natl Cancer Cent. 2024;4(1):47–53.
34. Anderson JB, Webb AJ. Fine-needle aspiration biopsy and the diagnosis of thyroid cancer. Br J Surg. 1987;74(4):292–6.
35. Redman R, et al. The impact of assessing specimen adequacy and number of needle passes for fine-needle aspiration biopsy of thyroid nodules. Thyroid. 2006;16(1):55–60.
36. Gursoy A, et al. Fine-needle aspiration biopsy of thyroid nodules: comparison of diagnostic performance of experienced and inexperienced physicians. Endocr Pract. 2010;16(6):986–91.
37. Baloch Z, et al. Role of repeat fine-needle aspiration biopsy (FNAB) in the management of thyroid nodules. Diagn Cytopathol. 2003;29(4):203–6.
38. Hajjar R, et al. The correlation between biological markers and prognosis in thyroid cancer. Biomedicines. 2024;12(12):2826.
39. Guo M, et al. Advances in targeted therapy and biomarker research in thyroid cancer. Front Endocrinol (Lausanne). 2024;15:1372553.
40. Wu Y, Zhang Z. Editorial: advances in targeted therapy and biomarker research for endocrine-related cancers. Front Endocrinol. 2024;15:1533623.
41. Jia Y, et al. Diagnostic value of B-RAF(V600E) in difficult-to-diagnose thyroid nodules using fine-needle aspiration: systematic review and meta-analysis. Diagn Cytopathol. 2014;42(1):94–101.
42. Zhou B, et al. Correlation between BRAF(V600E) mutation and aggressive biological behavior of papillary thyroid carcinoma. Zhonghua Yi Xue Za Zhi. 2023;103(14):1060–3.
43. Lubitz CC, et al. Circulating BRAF(V600E) levels correlate with treatment in patients with thyroid carcinoma. Thyroid. 2018;28(3):328–39.
44. Niedziela E, et al. Detection of the BRAF(V600E) mutation in circulating free nucleic acids as a biomarker of thyroid cancer: a review. J Clin Med. 2024;13(18):5396.
45. Macerola E, et al. Molecular genetics of follicular-derived thyroid cancer. Cancers (Basel). 2021;13(5):1139.

46. Medici M, et al. Long- versus short-interval follow-up of cytologically benign thyroid nodules: a prospective cohort study. BMC Med. 2016;14:11.
47. Nikiforov YE. RET/PTC rearrangement—a link between Hashimoto's thyroiditis and thyroid cancer…or not. J Clin Endocrinol Metab. 2006;91(6):2040–2.
48. Nikiforov YE. RET/PTC rearrangement in thyroid tumors. Endocr Pathol. 2002;13(1):3–16.
49. Elisei R, et al. RET/PTC rearrangements in thyroid nodules: studies in irradiated and not irradiated, malignant and benign thyroid lesions in children and adults. J Clin Endocrinol Metab. 2001;86(7):3211–6.
50. Sugg SL, et al. Distinct multiple RET/PTC gene rearrangements in multifocal papillary thyroid neoplasia 1. J Clin Endocrinol Metab. 1998;83(11):4116–22.
51. Zhang X, et al. RET/PTC rearrangement affects multifocal formation of papillary thyroid carcinoma. Zhonghua Er Bi Yan Hou Tou Jing Wai Ke Za Zhi. 2017;52(6):435–9.
52. Montero-Conde C, et al. Comprehensive molecular analysis of immortalization hallmarks in thyroid cancer reveals new prognostic markers. Clin Transl Med. 2022;12(8): e1001.
53. Shi X, et al. Integrated proteogenomic characterization of medullary thyroid carcinoma. Cell Discov. 2022;8(1):120.
54. Zhu G, et al. CircRNA: a novel potential strategy to treat thyroid cancer (review). Int J Mol Med. 2021;48(5):1.
55. Bi W, et al. CircRNA circRNA_102171 promotes papillary thyroid cancer progression through modulating CTNNBIP1-dependent activation of β-catenin pathway. J Exp Clin Cancer Res. 2018;37(1):275.
56. Mohan U, et al. Systematic review and meta-analysis to identify the immunocytochemical markers effective in delineating benign from malignant thyroid lesions in FNAC samples. Endocr Pathol. 2022;33(2):243–56.
57. Turan Z, Erkılıç S. TROP2: a potential marker in diagnosis of thyroid neoplasms. Irish J Med Sci. 2023;192(1):99–103.
58. Wu Y, et al. P4HA2 promotes cell proliferation and migration in glioblastoma. Oncol Lett. 2021;22(2):601.
59. Xiong G, et al. Prolyl-4-hydroxylase α subunit 2 promotes breast cancer progression and metastasis by regulating collagen deposition. BMC Cancer. 2014;14:1.
60. Xu S, et al. The role of collagen in cancer: from bench to bedside. J Transl Med. 2019;17(1):309.
61. Lin J, et al. P4HA2 is associated with prognosis, promotes proliferation, invasion, migration and EMT in glioma. bioRxiv. 2020. https://doi.org/10.1101/2020.02.05.935221.
62. Wang S, et al. CEBPB upregulates P4HA2 to promote the malignant biological behavior in IDH1 wildtype glioma. Faseb J. 2023;37(4): e22848.
63. Shi R, et al. Collagen prolyl 4-hydroxylases modify tumor progression. Acta Biochim Biophys Sin (Shanghai). 2021;53(7):805–14.
64. Song M, et al. Increased expression of collagen prolyl hydroxylases in ovarian cancer is associated with cancer growth and metastasis. Am J Cancer Res. 2023;13(12):6051.
65. Lee S-A, et al. Biophysical interplay between extracellular matrix remodeling and hypoxia signaling in regulating cancer metastasis. Front Cell Dev Biol. 2024;12:1335636.
66. Glaviano A, et al. PI3K/AKT/mTOR signaling transduction pathway and targeted therapies in cancer. Mol Cancer. 2023;22(1):138.
67. Calvo F, et al. Mechanotransduction and YAP-dependent matrix remodelling is required for the generation and maintenance of cancer-associated fibroblasts. Nat Cell Biol. 2013;15(6):637–46.
68. Browne IM, et al. Optimal targeting of PI3K–AKT and mTOR in advanced oestrogen receptor-positive breast cancer. Lancet Oncol. 2024;25(4):e139–51.
69. Burstein HJ, et al. Endocrine and targeted therapy for hormone receptor-positive, human epidermal growth factor receptor 2-negative metastatic breast cancer-capivasertib-fulvestrant: ASCO rapid recommendation update. J Clin Oncol. 2024;42(12):1450–3.
70. Shang L, et al. P4HA2 promotes occurrence and progression of liver cancer by regulating the PI3K/Akt/mTOR signaling pathway. Nan Fang Yi Ke Da Xue Xue Bao. 2022;42(5):665–72.
71. Wu Y-L, et al. P4HA2 contributes to head and neck squamous cell carcinoma progression and EMT through PI3K/AKT signaling pathway. Med Oncol. 2024;41(6):163.
72. Jin E, et al. P4HA2 activates mTOR via hydroxylation and targeting P4HA2-mTOR inhibits lung adenocarcinoma cell growth. Oncogene. 2024;43(24):1813–23.
73. Chi Z, et al. P4HA2 promotes proliferation, invasion, and metastasis through regulation of the PI3K/AKT signaling pathway in oral squamous cell carcinoma. Sci Rep. 2024;14(1):15023.
74. Li Q, et al. P4HA2 hydroxylates SUFU to regulate the paracrine Hedgehog signaling and promote B-cell lymphoma progression. Leukemia. 2024;38(8):1751–63.
75. Jiang W, et al. Prolyl 4-hydroxylase 2 promotes B-cell lymphoma progression via hydroxylation of Carabin. Blood. 2018;131(12):1325–36.
76. Lin J, et al. P4HA2 promotes epithelial-to-mesenchymal transition and glioma malignancy through the collagen-dependent PI3K/AKT pathway. J Oncol. 2021;2021:1406853.

Publisher's Note