Development of A Three-Gene Signature Prediction Model for Lymph Node Metastasis in Papillary Thyroid Cancer

Ziwei Huang
Wuhan Union Hospital https://orcid.org/0000-0003-2070-8981
Yuenan Liu
Wuhan Union Hospital
Kehao Le
Zhejiang University School of Medicine Sir Run Run Shaw Hospital
Ming Xu
Wuhan Union Hospital
Wenhui Li
Wuhan Union Hospital
Qiuyang Zhao
Wuhan Union Hospital
Jun Zhou
Wuhan Union Hospital
Yujia Jiang
Wuhan Union Hospital
Wen Yang
Wuhan Union Hospital
Li Yang
Wuhan Union Hospital
Pengfei Yi (yipengfei1986@126.com)
https://orcid.org/0000-0003-1655-9696

Primary research

Keywords: Prediction model, Papillary thyroid cancer, Lymph node metastasis, Transcriptome analysis, WGCNA

Posted Date: July 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-41157/v1

Abstract

Background: Thyroid cancer is one of the most prevalent endocrine cancers with a rising incidence rate over the past years. Papillary thyroid cancer (PTC) is the dominant historical type of thyroid cancer. Early lymph node metastasis happens frequently in PTC. However, some of the lymph node metastasis may be troublesome for detecting because of limited methods.

Methods: Robust rank aggregation afforded us the shared differential expression genes among multiple datasets. Gene ontology analysis was performed to identify potential functions. Weighted gene co-expression network analysis was used to research the correlations between gene expression patterns with clinical characteristic. Protein-protein interaction network was performed to identify the hub genes. The least absolute shrinkage and selection operator and Logistic regression were performed to construct a prediction model.

Results: We developed a three-gene signature prediction model for lymph node metastasis in PTC through transcriptomic analysis. After quality control, we collected 8 microarray datasets from GEO database and an RNA sequencing dataset from TCGA database. We found the transcriptome profiles were correlated with lymph node metastasis and 3 genes were verified to be independent prediction factors towards those statistic approach. Afterwards, we designed a predicable risk score system and effectively confirmed the model in two independent papillary thyroid cancer cohorts.

Conclusions: We recommended a successful predicable model of lymph node metastasis in papillary thyroid cancer patients with moderate accuracy.

Background

Thyroid cancer, one of the most prevalent endocrine cancers with a rising incidence rate over the past years, is the fifth leading incidence of cancer in female [1]. Papillary thyroid cancer (PTC) is the dominant historical type which contributes to approximately 84% of all thyroid cancer [2, 3]. However, through the low mortality and moderate prognosis were frequently mentioned, the recurrence and the complications are still perplexing those PTC patients [4]. Besides, lymphatic invasion, cervical lymph node metastasis, larger size of tumor, increasing diagnosis age and extraordinary enlargement of thyroid tissue increase the progression risk of PTC [5]. Due to early lymph node metastasis happening in PTC frequently [6], early detection and diagnosis are of great value. Currently multiple methods like thyroid and neck ultrasound, CT/MRI and fine-needle aspiration (FNA) for suspicious lateral neck nodes could effectively diagnose thyroid cancer [7, 8]. Yet some of the lymph node metastasis may be troublesome for detecting. We therefore urgently called for a reliable and straightforward approach to determine the possibility of lymph node metastasis.

Recent years, high throughput analysis afford us an advanced and efficient technique of evaluating the molecular disruptions in tumor tissues. For instance, A study of predicting chemotherapy sensitivity in cervical cancer, in which expression levels of 22 total and phosphorylated protein were analyzed in 181 frozen tissue samples, resulted in a model that was capable of predicting patients' chemotherapy sensitivity and assessing clinical outcome [9]. Thus, as more and more cancer sequencing databases are established, there is a great potential for us to acquire and analyze these data and guide clinical decisions with the findings.

In order to discover the predictive value of lymph node metastasis in the papillary thyroid cancer transcriptome, we searched for several RNA-seq as well as gene microarray datasets containing both papillary thyroid cancer samples

and normal thyroid samples to identify differentially expressed genes. Genes with powerful correlation to lymph nodes metastasis were selected as well. Afterwards, we designed a predicable risk score system and effectively confirmed the model in two independent papillary thyroid cancer cohorts. The entire flow of our efforts to identify predictive models was presented in Additional file 1: Fig. S1. Eventually, we recommended a successful predicable model of lymph node metastasis in papillary thyroid cancer patients with moderate accuracy.

Methods

Data collection, normalization and preprocessing

Gene microarray datasets and the associated clinical data of PTC and normal thyroid samples were downloaded from UCSC XENA (https://xenabrowser.net/) and Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/). Totally 8 GEO datasets and 1 The Cancer Genome Atlas (TCGA) dataset were involved in this research. Gene labels in other forms were all normalized into official gene symbols. Furthermore, those samples were proceeded to a quality control step called SimpleAffy to lessen bias. SimpleAffy was used to identify the 3'-to-5' ratios of GAPDH and β-actin. Outliers were excluded from the study.

Differential analysis of gene expression datasets

EdgeR R package afforded us the differential expression analysis between papillary thyroid cancer tissues and normal thyroid tissues. The filter thresholds of *P* value were set as less than 0.05 in integrated analysis. |log2FC (fold change) | was set as more than 0.1 or more than 0.5 in GEO or TCGA microarray separately due to the different sample sizes. A robust rank aggregation (RRA) algorithm was used to determine the overall differentially expressed genes （DEGs） in the 8 GEO datasets in which they were all up-regulated or down-regulated. The most relevant genes could be revealed in the analysis.

Functional annotation analysis

Gene Ontology (GO) analysis was performed on the DEGs in order to identify their potential functions. The process was settled by the clusterProfiler 3.11. The result of three major GO terms called biological process (BP), cellular component (CC) and molecular function (MF) could be visualized in dot plot or pie chart form by using the ggplot2 R package.

Weighted gene co-expression network analysis (WGCNA)

All samples of TCGA underwent a sample clustering test to identify their relationships. Those outliers were eliminated and then the others went through soft-threshold defining procedure. A Scale-free gene co-expression network was attained when the correlation coefficient was 0.85. At this point the soft-thresholding power was equal to 12. Afterward, WGCNA algorithm of R software was performed to construct a scale-free network of all DEGs of TCGA. After the construction, cluster analysis was operated to arrange genes with similar expression patterns into gene modules. The minimum size of each module was set as 30. Besides, the cut height threshold for merging modules was defined as 0.25 and some of the modules could be integrated to analyze. Lately, we constructed the relationships between modules and phenotype. The module–trait association was defined as the correlation of module eigengenes (MEs) and traits. Accordingly, each correlation value and the significance value were calculated separately to reveal the relevant modules, which are intently related to the traits. For each gene modules, gene significance (GS) is described as the correlation level between clinical trait and expression pattern. While module

membership (MM) stands for the correlation level between MEs and expression pattern. In our study, genes with GS score more than 0.8 and MM score more than 0.2 were defined as module genes with high correlation to certain phenotype.

Protein–protein interaction (PPI) network and identification of hub-genes

The DEGs of both GEO and TCGA datasets were uploaded to the STRING tool (http://www.string-db.org/) to look into their correlation and to discover the hub-genes in the gene network. A darker line represents a stronger edge confidence in the network. Then, Cytoscape software was utilized to visualize the outcome of STRING database and to acquire the hub-genes which have a closest relation and have a more considerable function among the DEGs. In our study, the 100 genes with the highest connectivity degree were identified as hub-genes of the PPI network.

Statistical analysis of model acquisition and validation

We randomly separate the TCGA microarray data and relevant clinical data into 2 parts, named a training cohort and a validation cohort. A univariate Logistic regression was applied to DEGs in TCGA training cohort, aid of finding the relationship between those genes and patient's prognostic data. The results were then diminished by using the LASSO algorithm with the R package 'glmnet'. The minimal partial likelihood deviance was carried out as optimal tuning parameter (λ) changed. Also some of the coefficients of gene would reduce to zero. Those genes were excluded and the others were accessed to the multivariate Logistic test. The hazard ratio (HR) and 95% confidence intervals (CI) of each gene would be calculated and only those genes which did not include 1 in the 95% CI were selected as final trait-relative genes. The coefficient of each gene in multivariate Logistic regression model was summed to calculate the risk score of lymph node metastasis. The function could be computed as follows,

$$\text{Risk Score} = \sum_{i=1}^{n} \beta i \times \text{EXPgene(i)}$$

in which βi stands for the Logistic regression coefficient of gene i in the training cohort.

By calculating risk score of each sample, we could sort them and divide the cohort into low risk group and high risk group by the median value. A student's t test of two groups in comparing the trait demonstrated that the model could predict the risk effectively. Meanwhile, the receiver operating characteristic (ROC) curves was used to verify the sensitivity and specificity of lymph node metastasis related model risk prediction. After the model construction, the corresponding approach above was applied to TCGA validation cohort as well as GEO independent cohort in turn for risk predicting system confirmation.

Results

Identification of DEGs and associated GO analysis in GEO datasets

Firstly, all the human tissue microarrays including papillary thyroid carcinoma samples and paired/unpaired normal thyroid samples were extracted from the GEO database. In order to make sure the quality of the research, SimpleAffy R package was utilized to determine the 3'-to-5' ratios of β-actin and GAPDH (Additional file 2: Fig. S2). After excluding seven tumor samples and 2 normal samples in three datasets, a total of 187 tumor samples and 119 normal control samples in eight datasets (GSE33630 [10, 11], GSE60542 [12], GSE66783 [13], GSE5364 [14],

GSE129562 [15], GSE97001 [16], GSE3467 [17] and GSE27155 [18, 19]) were included this research (Table 1). After performing differential analysis on the 8 datasets by the edgeR and robust rank aggregation algorithm, the DEGs between the normal tissues and PTC tissues of each dataset were obtained, with an adjusted *P* value < 0.05 and |log2FC (fold change) | > 0.1 as the cut-offs. A total of 531 over-expressed genes and 474 suppressed genes were discovered in PTC tissues. The top ten genes that were overexpressed or suppressed are listed by a heatmap in Fig. 1a. GO analysis was performed on the DEGs in order to identify potential functions. Those DEGs were enriched in GO terms including 500 BPs, 70 CCs and 28 MFs with the cut-off value set as adjust *P* value < 0.01 and q value < 0.01. TOP 10 GO terms of each category were displayed in Fig. 1b. Among them, extracellular structure organization and extracellular matrix organization, cell-substrate adhesion and renal system development were three significant aspects in BP category. The first two functions are also remarkable in CC term. The most enriched MF terms were sulfur compound binding, glycosaminoglycan binding and serine-type peptidase activity.

Identification of DEGs and associated GO analysis in TCGA dataset

The TCGA-THCA dataset consisting of 56 normal thyroid tissues and 497 thyroid cancer tissues were selected by integrated analysis (*P* value < 0.05 and |log2FC | > 0.5 set as cut-offs), from which 6492 mRNAs with steady differentially expressed patterns were identified by edgeR analysis. Similarly, these DEGs were analyzed by GO functional enrichment analysis to find out the potential biological functions. The BP term were shown in a pie chart (Fig. 2a). Activation of MAPK activity and leukocyte mediated cytoxicity are the top 2 enriched terms of BPs. While CCs and MFs were exhibited in Fig. 2b. Each category only shows the top ten functions.

Discovery of a strong correlation between lymph node metastasis and DEGs in TCGA dataset by WGCNA algorithm

For DEGs of thyroid cancer tissues and normal thyroid tissues in the TCGA dataset, we used WGCNA algorithm to find out which clinical traits they are significantly related to. Firstly, after screening the clinical characteristics and sample tree, a total of 482 thyroid cancer samples were included in the study. In addition, the sample tree and the clinical characteristics of each sample are also shown in Additional file 3: Fig. S3a. Then, through the soft threshold screening (Additional file 3: Fig. S3b), a gene co-expression network with the scale-free characteristics was established, in which the evaluation parameter R^2 was set as 0.85. After the network was established, genes with similar expression level were grouped in the same module and displayed in the form of a hierarchical clustering diagram (Additional file 3: Fig. S3c). Different colors indicated separate modules. It should be noted that the genes in gray color were not classified into any modules. After that, the clinical characteristics of the samples were taken out and analyzed for correlation with each module (Fig. 3). We can find that the brown module (r =-0.34, p=3e-14) and the turquoise module (r= 0.35, p=6e-15) were highly correlated with lymph node metastasis. By calculating the MM value of these two modules (cut-off set as MM *P* value < 0.05 and Gene MM value > 0.8), a total of 106 genes that were closely related to the brown module and the turquoise module were identified for further analysis. Since the results of WGCNA strongly implied the relevance of these DEGs to lymph node metastasis, we subsequently concentrated on lymph node metastasis and attempted to explore models that could accurately predict lymph node metastasis in PTC.

Potential hub-genes relating to lymph node metastasis revealed by PPI network

The shared DEGs in both GEO datasets and TCGA dataset were considered as important genes for papillary thyroid cancer. To narrow the scope and identify the much meaningful genes among them, PPI network analysis was used to find key regulatory genes in these genes. STRING database (https://string-db.org) was used to identify the PPI network of these 599 shared DEGs. Fig. 5a showed the overall PPI regulatory network of these differential genes.

through the PPI network connection score, one cluster with the highest score was displayed in Fig. 5b. After all genes going through PPI analysis, some potential hub-genes can be found based on the score. The first 100 hub-genes were extracted as key genes. What's more, those hub-genes were intersected with the WGCNA module genes mentioned above (Fig. 5c). A total of 9 genes were identified as important DEGs and related to lymph node metastasis.

construction a risk scoring system by logistic regression analysis and LASSO algorithm

We randomly divided the thyroid cancer samples in the TCGA database into two parts. The first cohort named training cohort was utilized to find the risk scoring system of lymph nodes metastasis. Meanwhile, the other one named validation cohort was used to confirm the system. Also, an independent GEO cohort with clinical traits was enrolled to verify the model. Their clinical features were listed in Table 2. Univariate logistic analysis was performed on the 9 genes to find out if they related to lymph nodes metastasis (Fig. 4d). A predictive gene was defined as a hazard ratio (HR) and 95% confidence interval (CI) greater than or less than 1 and *P* value less than 0.05. It is gratifying that these nine genes were all found to be associated with lymph node metastasis after the analysis. Among them, EPHB3, also called EPH Receptor B3; MET, one of the receptor tyrosine kinase; ICAM1, also named Intercellular Adhesion Molecule 1; SERPINA1, also called Serpin Family A Member 1 and FN1, also named Fibronectin 1 are the single risk factors for lymph nodes metastasis, while ITPR1, also named Inositol 1,4,5-Trisphosphate Receptor Type 1; PPARGC1A,also called PPARG Coactivator 1 Alpha; GNA14, also named G Protein Subunit Alpha 14 and BCL2,a apoptosis regulator were found as protective factors in this trait. In order to compress the model and identity the key genes, A least absolute shrinkage and selection operator (LASSO) regression model was then used to test these 9 Genes. In the LASSO model, when the value of λ increases, more coefficients (genes) will be set to zero, meaning those variables could be remove from the model due to their shrinking property (Fig. 5a). Six genes model satisfied the minimum partial likelihood deviance due to the ridge regression. The minimum log(λ) was -3.67 at this status. Finally, these six genes were subjected to multivariate logistic analysis in order to find the final risk prediction model. The result indicated that MET, ITPR1, and BCL2 were independent prognostic factors for lymph nodes metastasis (Fig. 5c, Table 3). The score of risk estimation model could be calculated as: expression of ITPR11 × 0.589 + expression of MET × 0.841 - expression of BCL2 × 0.786. The factors stand for the respective multivariate Logistic regression coefficients.

Verification of the prediction system through validation cohorts from TCGA and GEO datasets

TCGA training cohort used for model construction were separate into high-risk group and low-risk group based on each sample's risk score (Fig. 6a). The median risk score (3.879) was set as the cut-off. For two groups, a significant difference of lymph nodes metastasis was clearly shown in heatmap (Fig. 6b). A student's t test also demonstrated that the high-risk group had a higher frequency of lymph nodes metastasis than the low-risk group (Fig. 6c) .The receiver operating characteristic (ROC) curve shows the efficiency of the prediction model, in which the area under the curve (AUC) was 0.744 (Fig. 6d). In order to verify the prediction accuracy of the risk prediction model, we used multiple cohorts to verify. TCGA validation cohort has been operated by the same protocol (Fig. 6e). Obviously, the result also showed that a high risk score could be riskier to lymph nodes metastasis by a heatmap (Fig. 6f) and student's t test (Fig. 6g). The AUC over validation model is 0.711 (Fig. 6h). At the same time, we conducted an independent verification in another cohort combined from two GEO dataset (GSE3467 and GSE60542) with the same platform (GPL570) and corresponding clinical data, in order to show that the risk scoring system is generally applicable (Fig. 7a). In the heatmap, a strong tendency was discovered of which higher risk score indicated higher frequency of lymph nodes metastasis (Fig. 7b). A student's t test verifies the point (Fig. 7d).

Furthermore, gene expressions of each sample tissue were displayed in Fig. 7c as a heatmap. Finally, we performed a ROC analysis and a meaningful result was revealed with 0.6842 of AUC value. Accordingly, the three-genes model was verified in multiple microarray and all showed a great difference. It is a reliable, accurate and independent predictive appliance for determining lymph nodes metastasis in papillary thyroid cancer patients.

Discussion

Genomic analysis has an extraordinarily critical role in tumor research. For example, Paik S et al. derived a 21-genes based recurrence scoring system from a prospective analysis of multiple gene expression levels in a breast cancer population [20]. Besides, transcriptome mapping could also be applied for the determination of molecular subtypes of tumors. It has already been implemented in colorectal cancer [21], breast cancer [22], prostate cancer [23], and pancreatic cancer [24], which has a facilitating effect on clinical decision-making. Moreover, genomic studies provide an insight into the tumor immune microenvironment. Xu M et al. calculated the corresponding immune infiltration scores from the expression of immune-related molecules in breast cancer specimens and the immune score was found to perform a detrimental effect in overall survival and recurrence-free survival [25]. Chakladar J et al. also analyzed immune-related genes in PTC through the combined application of genomics and transcriptomics [26]. In the present research, we investigated multiple independent datasets of PTC, and successfully established a three-gene (MET, ITPR1 and BCL2) prediction model for lymph node metastasis in PTC patients.

In our study, MET and ITPR1 expression in the predictive model was a risk factor for lymph node metastasis, whereas BCL2 was a protective factor. Previous study has reported that BCL2 was found to be highly expressed only in poorly differentiated tumors [27]. Moreover, another study also showed that a lower expression level of BCL2 could act as an early sign of oncogenesis and be a reason for the favorable prognosis [28], suggesting that BCL2 may act as a protective factor in thyroid cancer. As for ITPR1, one study demonstrated that up-expression of ITPR1 shelters renal cancer cells against natural killer cells [29]. Another study showed that ITPR1 could enhance paclitaxel toxicity in breast cancer [30]. We discovered ITPR1 is a risk factor for lymph node metastasis in thyroid cancer, but the biological mechanism still under solving. As a heterodimeric transmembrane receptor tyrosine kinase, MET mediates the activation of multiple signaling pathways, including PI3K/AKT, Ras-Rac/Rho and phospholipase C-γ pathways [31]. Significantly higher level of MET was detected in PTC [32, 33], non-small cell lung cancer [34], bladder cancer [35] and oral cancer [36]. Those results all add up to a better proof of MET as a cancer-promoting factor and can be corroborated with our findings. In conclusion, based on the available studies, BCL2 and MET match the results more accurately, while ITPR1 in thyroid cancer has been less studied and needs to be further explored.

The Robust rank aggregation algorithm was utilized to analyze multiple gene sets integrally and identify the common DEGs. This algorithm compensates for the limitations of the previous single data set analysis and minimizes bias. Since PTC samples are generally small, the RRA algorithm could be quite helpful. WGCNA is an effective method for describing associations of gene expression patterns with clinical phenotypes. In thyroid cancer research, WGCNA was widely used [37, 38]. Based on the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement, studies developing new prediction models should always go through an internal validation (also called self-validation) to quantify the predictive appearance. Also, it is firmly recommended to appraise the model in other data (also called external validation) after developing a prediction model [39]. We then applied our prediction model to TCGA self-validation cohort and one independent GEO cohort. The outcomes of two dataset strongly fit in the model.

The innovative feature of our study is that the combined application of RRA algorithm, WGCNA analysis and PPI methods was first used to analyze the correlation between PTC and lymph node metastasis. The degradation of samples may cause bias of the results. Therefore, we performed quality control on each sample in the dataset. Tissues that did not meet the requirements (degradation occurred) were excluded from the follow-up experiment. Furthermore, the predictive model of PTC lymph node metastasis was uniquely established, which might be useful for therapeutic decision-making and clinical monitoring. However, the inadequacy of this study is that the prognostic data in the TCGA database for papillary thyroid cancer are quite good and we failed to find significant differences in prognosis, while the GEO dataset was unable to find prognosis information. Therefore, it is regrettable that prognosis cannot be measured.

For PTC, lymph node metastasis may arise at an early stage and there is a potential risk for skip metastasis [40], the mechanism of which is currently not clear. Routine diagnosis methods such as neck CT/MRI and neck lymph node ultrasound [7, 8] may not be able to fully detect the development of lymph node metastasis. Patients who develop lymph node metastases are likely to require postoperative I_{131} radiation therapy and may have a higher recurrence rate and lower survival rate [41]. Therefore, the outcome of this study may have a better role in predicting lymph node metastasis in PTC patients. Patients in the low-risk group have a lower likelihood of lymph node metastasis. With current conventional methods of measuring expression level, such as RT-qPCR or immunohistochemistry, we can easily, accurately and economically obtain risk scores for this patient, thus allowing the model to be better applied in clinical practice.

Conclusions

Summarily, this study is a highly scientific and accurate method with successful predictive significance for determining lymph node metastasis in PTC patient. Among the model, 2 genes were identified as risk factors and 1 genes was protective factor in PTC patients. It might be potential treatment targets and urged for further research.

Abbreviations

PTC, Papillary Thyroid Cancer; CT, Computed Tomography; MRI, Magnetic Resonance Imaging; GEO, Gene Expression Omnibus; TCGA, The Cancer Genome Atlas; DEGs, Differentially Expressed Genes; RRA, Robust Rank Aggregation; GO, Gene Ontology; WGCNA, Weighted Gene Co-expression Network Analysis; MEs, Module Eigengenes; GS,Gene Significance; MM, Module Membership; PPI, Protein–Protein Interaction; HR, Hazard Ratio; CI, Confidence Intervals; ROC, Receiver Operating Characteristic; FC, Fold Change; BP, Biological Process; CC, cellular component; MF, molecular functions.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

PY, ZH, and YL conceived and arranged the experiments, and wrote the manuscript. KL, QZ, WL, MX, JZ, YJ, WY and LY collected and analyzed the data. All authors read and approved the final version of the manuscript.

Funding

Not applicable.

Availability of data and materials

Gene microarray datasets and the associated clinical data of PTC and normal thyroid samples in our study were all publicly available. The data of PTC patients from TCGA were downloaded from UCSC XENA (https://xenabrowser.net/). Additionally, Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/) afford us 8 cohorts (GSE33630, GSE60542, GSE66783, GSE5364, GSE129562, GSE97001, GSE3467 and GSE27155) for further research in the study.

Ethics approval and consent to participate

This study was approved by Ethics Committee of Huazhong University of science and technology (HUST).

Consent for publication

All authors have agreed for this publication.

Competing interests

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2020**. *CA Cancer J Clin* 2020, **70**(1):7-30.
2. Geraldo MV, Kimura ET: **Integrated Analysis of Thyroid Cancer Public Datasets Reveals Role of Post-Transcriptional Regulation on Tumor Progression by Targeting of Immune System Mediators**. *PLoS One* 2015, **10**(11):e0141726.
3. Aschebrook-Kilfoy B, Kaplan EL, Chiu BCH, Angelos P, Grogan RH: **The Acceleration in Papillary Thyroid Cancer Incidence Rates is Similar Among Racial and Ethnic Groups in the United States**. *Ann Surg Oncol* 2013, **20**(8):2746-2753.
4. Chrisoulidou A, Boudina M, Tzemailas A, Doumala E, Iliadou PK, Patakiouta F, Pazaitou-Panayiotou K: **Histological subtype is the most important determinant of survival in metastatic papillary thyroid cancer**. *Thyroid Res* 2011, **4**(1):12.
5. Cheng Q, Li X, Acharya CR, Hyslop T, Sosa JA: **A novel integrative risk index of papillary thyroid cancer progression combining genomic alterations and clinical factors**. *Oncotarget* 2017, **8**(10):16690-16703.
6. Moo TA, McGill J, Allendorf J, Lee J, Fahey T, 3rd, Zarnegar R: **Impact of prophylactic central neck lymph node dissection on early recurrence in papillary thyroid carcinoma**. *World J Surg* 2010, **34**(6):1187-1191.
7. Grani G, Ramundo V, Falcone R, Lamartina L, Montesano T, Biffoni M, Giacomelli L, Sponziello M, Verrienti A, Schlumberger M *et al*: **Thyroid Cancer Patients With No Evidence of Disease: The Need for Repeat Neck Ultrasound**. *J Clin Endocrinol Metab* 2019, **104**(11):4981-4989.
8. Torlontano M, Attard M, Crocetti U, Tumino S, Bruno R, Costante G, D'Azzo G, Meringolo D, Ferretti E, Sacco R *et al*: **Follow-up of low risk patients with papillary thyroid cancer: role of neck ultrasonography in detecting lymph node metastases**. *J Clin Endocrinol Metab* 2004, **89**(7):3402-3407.

9. Choi CH, Chung JY, Kang JH, Paik ES, Lee YY, Park W, Byeon SJ, Chung EJ, Kim BG, Hewitt SM *et al*: **Chemoradiotherapy response prediction model by proteomic expressional profiling in patients with locally advanced cervical cancer**. *Gynecol Oncol* 2020, **157**(2):437-443.
10. Tomas G, Tarabichi M, Gacquer D, Hebrant A, Dom G, Dumont JE, Keutgen X, Fahey TJ, 3rd, Maenhaut C, Detours V: **A general method to derive robust organ-specific gene expression-based differentiation indices: application to thyroid cancer diagnostic**. *Oncogene* 2012, **31**(41):4490-4498.
11. Dom G, Tarabichi M, Unger K, Thomas G, Oczko-Wojciechowska M, Bogdanova T, Jarzab B, Dumont JE, Detours V, Maenhaut C: **A gene expression signature distinguishes normal tissues of sporadic and radiation-induced papillary thyroid carcinomas**. *Br J Cancer* 2012, **107**(6):994-1000.
12. Tarabichi M, Saiselet M, Tresallet C, Hoang C, Larsimont D, Andry G, Maenhaut C, Detours V: **Revisiting the transcriptional analysis of primary tumours and associated nodal metastases with enhanced biological and statistical controls: application to thyroid cancer**. *Br J Cancer* 2015, **112**(10):1665-1674.
13. Lan X, Zhang H, Wang Z, Dong W, Sun W, Shao L, Zhang T, Zhang D: **Genome-wide analysis of long noncoding RNA expression profile in papillary thyroid carcinoma**. *Gene* 2015, **569**(1):109-117.
14. Yu K, Ganesan K, Tan LK, Laban M, Wu J, Zhao XD, Li H, Leung CH, Zhu Y, Wei CL *et al*: **A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers**. *PLoS Genet* 2008, **4**(7):e1000129.
15. Lee S, Bae JS, Jung CK, Chung WY: **Extensive lymphatic spread of papillary thyroid microcarcinoma is associated with an increase in expression of genes involved in epithelial-mesenchymal transition and cancer stem cell-like properties**. *Cancer Med* 2019, **8**(15):6528-6537.
16. Iacobas DA, Tuli NY, Iacobas S, Rasamny JK, Moscatello A, Geliebter J, Tiwari RK: **Gene master regulators of papillary and anaplastic thyroid cancers**. *Oncotarget* 2018, **9**(2):2410-2424.
17. He H, Jazdzewski K, Li W, Liyanarachchi S, Nagy R, Volinia S, Calin GA, Liu CG, Franssila K, Suster S *et al*: **The role of microRNA genes in papillary thyroid carcinoma**. *Proc Natl Acad Sci U S A* 2005, **102**(52):19075-19080.
18. Giordano TJ, Kuick R, Thomas DG, Misek DE, Vinco M, Sanders D, Zhu Z, Ciampi R, Roh M, Shedden K *et al*: **Molecular classification of papillary thyroid carcinoma: distinct BRAF, RAS, and RET/PTC mutation-specific gene expression profiles discovered by DNA microarray analysis**. *Oncogene* 2005, **24**(44):6646-6656.
19. Giordano TJ, Au AY, Kuick R, Thomas DG, Rhodes DR, Wilhelm KG, Jr., Vinco M, Misek DE, Sanders D, Zhu Z *et al*: **Delineation, functional validation, and bioinformatic evaluation of gene expression in thyroid follicular carcinomas with the PAX8-PPARG translocation**. *Clin Cancer Res* 2006, **12**(7 Pt 1):1983-1993.
20. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, Baehner FL, Walker MG, Watson D, Park T *et al*: **A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer**. *N Engl J Med* 2004, **351**(27):2817-2826.
21. Guinney J, Dienstmann R, Wang X, de Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P *et al*: **The consensus molecular subtypes of colorectal cancer**. *Nat Med* 2015, **21**(11):1350-1356.
22. Lehmann BD, Bauer JA, Chen X, Sanders ME, Chakravarthy AB, Shyr Y, Pietenpol JA: **Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies**. *J Clin Invest* 2011, **121**(7):2750-2767.
23. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, Ferrari M, Egevad L, Rayford W, Bergerheim U *et al*: **Gene expression profiling identifies clinically relevant subtypes of prostate cancer**. *Proc Natl Acad Sci U S A* 2004, **101**(3):811-816.

24. Bailey P, Chang DK, Nones K, Johns AL, Patch AM, Gingras MC, Miller DK, Christ AN, Bruxner TJ, Quinn MC *et al*: **Genomic analyses identify molecular subtypes of pancreatic cancer**. *Nature* 2016, **531**(7592):47-52.
25. Xu M, Li Y, Li W, Zhao Q, Zhang Q, Le K, Huang Z, Yi P: **Immune and Stroma Related Genes in Breast Cancer: A Comprehensive Analysis of Tumor Microenvironment Based on the Cancer Genome Atlas (TCGA) Database**. *Front Med (Lausanne)* 2020, **7**:64.
26. Chakladar J, Chu M, Gnanasekar A, Rosenberg KF, Tsai JC, Wong LM, Ongkeko WM: **Computational analysis of immune-associated genomic and transcriptomic elements differentiating papillary thyroid cancer subtypes**. *Cancer Research* 2019, **79**(13).
27. Soda G, Antonaci A, Bosco D, Nardoni S, Melis M: **Expression of bcl-2, c-erbB-2, p53, and p21 (waf1-cip1) protein in thyroid carcinomas**. *J Exp Clin Cancer Res* 1999, **18**(3):363-367.
28. Aksoy M, Giles Y, Kapran Y, Terzioglu T, Tezelman S: **Expression of bcl-2 in papillary thyroid cancers and its prognostic value**. *Acta Chir Belg* 2005, **105**(6):644-648.
29. Messai Y, Noman MZ, Hasmim M, Janji B, Tittarelli A, Boutet M, Baud V, Viry E, Billot K, Nanbakhsh A *et al*: **ITPR1 protects renal cancer cells against natural killer cells by inducing autophagy**. *Cancer Res* 2014, **74**(23):6820-6832.
30. Xu S, Wang P, Zhang J, Wu H, Sui S, Zhang J, Wang Q, Qiao K, Yang W, Xu H *et al*: **Ai-lncRNA EGOT enhancing autophagy sensitizes paclitaxel cytotoxicity via upregulation of ITPR1 expression by RNA-RNA and RNA-protein interactions in human cancer**. *Mol Cancer* 2019, **18**(1):89.
31. Birchmeier C, Birchmeier W, Gherardi E, Vande Woude GF: **Met, metastasis, motility and more**. *Nat Rev Mol Cell Biol* 2003, **4**(12):915-925.
32. Chitikova Z, Pusztaszeri M, Makhlouf AM, Berczy M, Delucinge-Vivier C, Triponez F, Meyer P, Philippe J, Dibner C: **Identification of new biomarkers for human papillary thyroid carcinoma employing NanoString analysis**. *Oncotarget* 2015, **6**(13):10978-10993.
33. Wang G, Cai C, Chen L: **MicroRNA-3666 Regulates Thyroid Carcinoma Cell Proliferation via MET**. *Cell Physiol Biochem* 2016, **38**(3):1030-1039.
34. Lutterbach B, Zeng Q, Davis LJ, Hatch H, Hang G, Kohl NE, Gibbs JB, Pan BS: **Lung cancer cell lines harboring MET gene amplification are dependent on Met for growth and survival**. *Cancer Res* 2007, **67**(5):2081-2088.
35. Shintani T, Kusuhara Y, Daizumoto K, Dondoo TO, Yamamoto H, Mori H, Fukawa T, Nakatsuji H, Fukumori T, Takahashi M *et al*: **The Involvement of Hepatocyte Growth Factor-MET-Matrix Metalloproteinase 1 Signaling in Bladder Cancer Invasiveness and Proliferation. Effect of the MET Inhibitor, Cabozantinib (XL184), on Bladder Cancer Cells**. *Urology* 2017, **101**:169 e167-169 e113.
36. Saintigny P, William WN, Jr., Foy JP, Papadimitrakopoulou V, Lang W, Zhang L, Fan YH, Feng L, Kim ES, El-Naggar AK *et al*: **Met Receptor Tyrosine Kinase and Chemoprevention of Oral Cancer**. *J Natl Cancer Inst* 2018, **110**(3).
37. Zhai T, Muhanhali D, Jia X, Wu Z, Cai Z, Ling Y: **Identification of gene co-expression modules and hub genes associated with lymph node metastasis of papillary thyroid cancer**. *Endocrine* 2019, **66**(3):573-584.
38. Tang X, Huang X, Wang D, Yan R, Lu F, Cheng C, Li Y, Xu J: **Identifying gene modules of thyroid cancer associated with pathological stage by weighted gene co-expression network analysis**. *Gene* 2019, **704**:142-148.
39. Collins GS, Reitsma JB, Altman DG, Moons KG: **Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement**. *BMJ* 2015, **350**:g7594.

40. Machens A, Holzhausen HJ, Dralle H: **Skip metastases in thyroid cancer leaping the central lymph node compartment**. *Arch Surg* 2004, **139**(1):43-45.

41. Podnos YD, Smith D, Wagman LD, Ellenhorn JD: **The implication of lymph node metastasis on survival in patients with well-differentiated thyroid cancer**. *Am Surg* 2005, **71**(9):731-734.

Tables

Table 1
Information of enrolled PTC patients from 8 GEO datasets after quality control.

Country	Organization	Series	Platform	Normal	Tumor	Quality control	Publication
Belgium	ULB	GSE33630	GPL570	45	47	Excluded 2 tumor samples	(Collins et al. 2015, Dom et al. 2012)
Belgium	IRIBHM	GSE60542	GPL570	30	33	Passed	(Tarabichi et al. 2015)
China	The First Hospital of China Medical University	GSE66783	GPL19850	5	5	Passed	(Lan et al. 2015)
Singapore	National Cancer Centre	GSE5364	GPL96	16	35	Passed	(Yu et al. 2008)
South Korea	The Catholic University of Korea	GSE129562	GPL10558	8	8	Passed	(Lee et al. 2019)
USA	Center for Computational Systems Biology	GSE97001	GPL10332	4	4	Passed	(Iacobas et al. 2018)
USA	Ohio State University	GSE3467	GPL570	7	5	Excluded 4 tumor and 2 normal samples	(He et al. 2005)
USA	University of Michigan	GSE27155	GPL96	4	50	Excluded 1 tumor sample	((Giordano et al. 2006; Giordano et al. 2005)

Table 2

Clinical pathological characteristics of patients in the training, self-validation cohorts and the independent GEO cohort.

Characteristics	TCGA training cohort	TCGA validation cohort	GEO validation cohort
	(N = 248)	(N = 249)	(N = 42)
Age at initial diagnosis (year)	46.3 ± 15.5	48.5 ± 16.1	45.11 ± 13.5
Gender			
Male	82(33.06%)	52(20.88%)	19(45.24%)
Female	166(66.94%)	197(79.12%)	23(54.76%)
Pathologic T			
T1 or Tx	69(27.82%)	75(30.12%)	9(21.43%)
T2	81(32.66%)	81(32.53%)	4(9.52%)
T3	87(35.08%)	82(32.93%)	24(57.12%)
T4	11(4.44%)	11(4.42%)	5(11.90%)
Pathologic N			
N0 or Nx	139(56.05%)	137(55.02%)	19(45.24%)
N1	109(43.95%)	112(44.98%)	23(54.76%)
Pathologic M			
M0 or Mx	244(98.39%)	245(98.39%)	39(92.86%)
M1	4(1.61%)	4(1.61%)	3(7.12%)
Tumor stage			
Stage I	144(58.06%)	135(54.22%)	19(45.24%)
Stage II	27(10.89%)	25(10.04%)	0(0.00%)
Stage III	52(20.97%)	59(23.69%)	10(23.81%)
Stage IV	24(9.68%)	29(11.65%)	3(7.14%)
Not report	1(0.40%)	1(0.40%)	10(23.81%)
Overall survival status			NA
Alive	242(97.58%)	239(95.98%)	
Dead	6(2.42%)	10(4.02%)	

Table 3
The three independent prediction factors of lymph node metastasis in papillary thyroid cancer.

Entrez ID	Gene	multivariate Cox regression analysis				Eight GEO datasets		TCGA dataset	
		Coefficient	Hazard ratio	confidence interval (95%)	*P* value	Log_2FC	Adjusted *P* value	Log_2FC	Adjusted *P* value
3708	ITPR1	0.589	1.801	1.133–2.864	0.013	0.322	3.532e-05	1.161	3.343e-24
4233	MET	0.841	2.320	1.606–3.350	7.000e-06	0.686	2.194e-08	0.620	4.746e-23
596	BCL2	-0.786	0.456	0.271–0.767	0.030	-0.342	2.736e-04	-0.689	5.406e-26

Figures

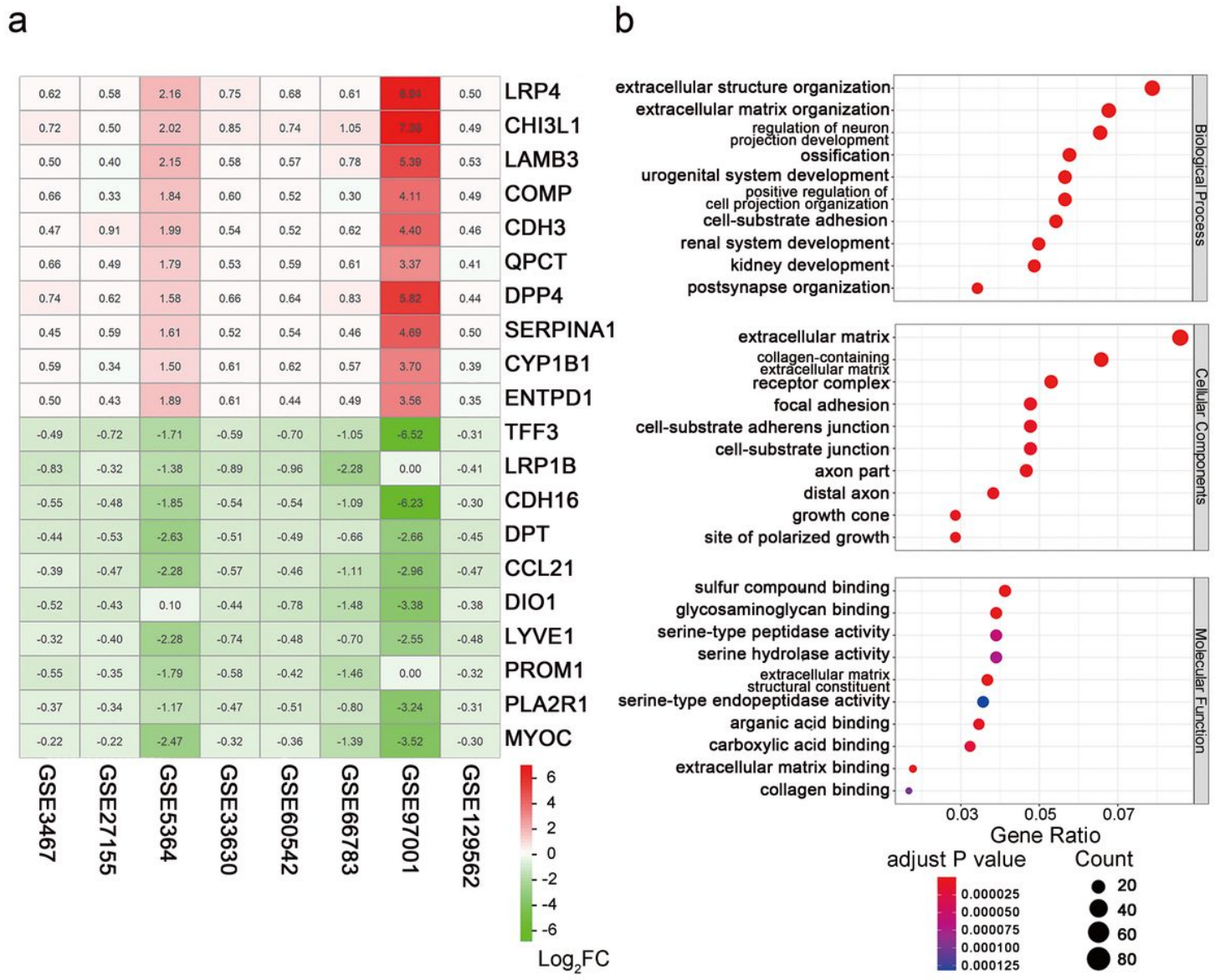

Figure 1

Identification of DEGs in GEO datasets along with their GO analysis. a TOP 10 up-regulated and down-regulated genes in eight GEO datasets were displayed as a heat map. Each grid represents the differential expression of these genes in each dataset with the log2FC inset, while the red color representing up-regulated genes and the green color representing down-regulated genes. b Significantly enriched GO terms of DEGs in GEO datasets. Each category including Biological process, Cellular Components and Molecular Functions were shown in separate charts. A bigger size of circle informed more genes were enriched. Respectively, the color of each circle represented the adjust P value.

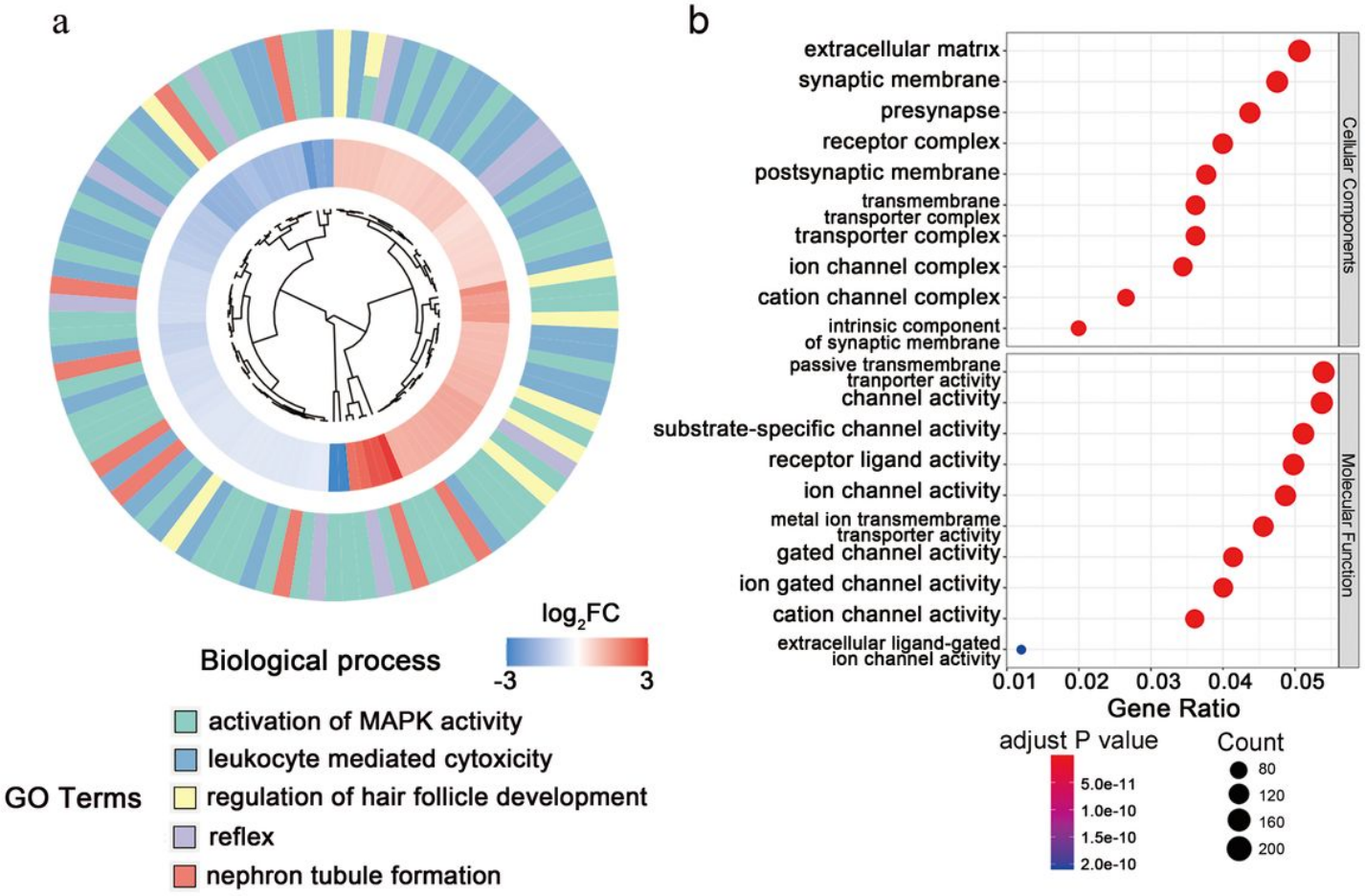

Figure 2

Plots of the enriched GO terms of DEGs in TCGA dataset. a Biological process, one of the GO categories, was presented in the form of a circle chart. TOP 5 enriched terms were listed below. b Cellular Components and Molecular Functions were shown in dot plot. The size of circle represented the numbers which genes in this term.

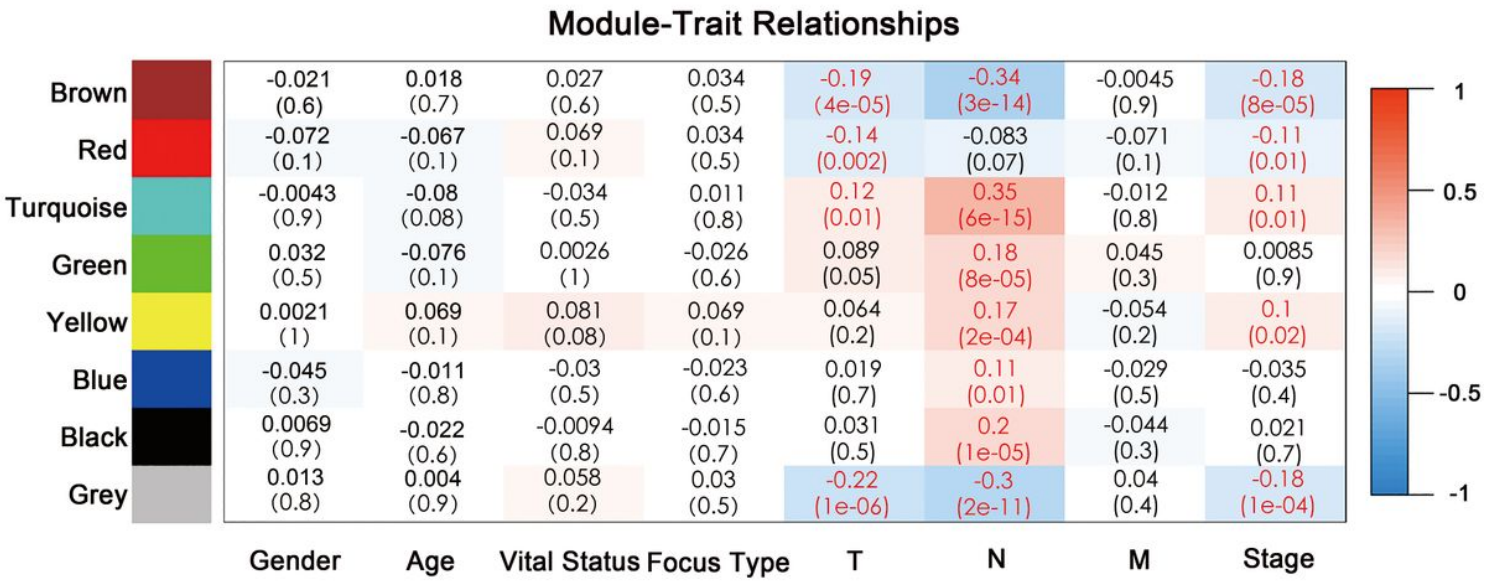

Figure 3

A strong correlation between lymph node metastasis and DEGs found by WGCNA. The Pearson correlation coefficient was shown in the box and the P value was shown in the brackets below. Values with significant differences were marked in red color. Abbreviations: T, Primary tumor; N, Regional lymph nodes; M, Metastasis.

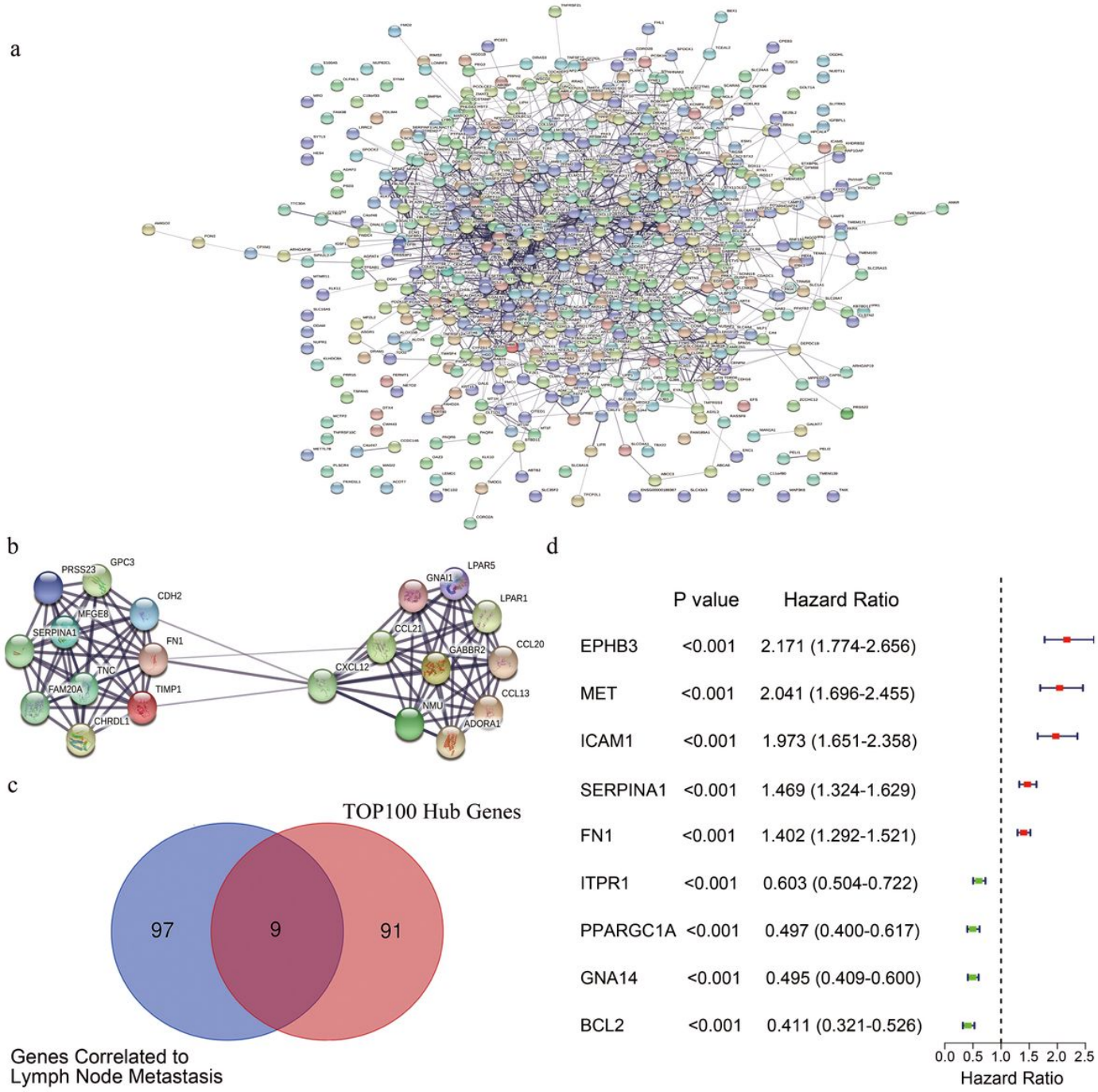

Figure 4

Revealing of potential hub-genes relating to lymph node metastasis. a The overall PPI network made of 599 DEGs in both TCGA and GEO datasets. b One of the clusters that had the highest score in overall PPI network. c Venn gram revealed the intersection between top 100 genes in PPI network and the genes with high relationship to modules that strongly related to lymph node metastasis. d Forest illustration of the univariate logistic analysis. P value and Hazard ratio of each gene were shown on the left side. Hazard ratio less than 1 indicated that the gene was a protective factor. Meanwhile a bigger-than-1's hazard ratio indicated a harmful factor.

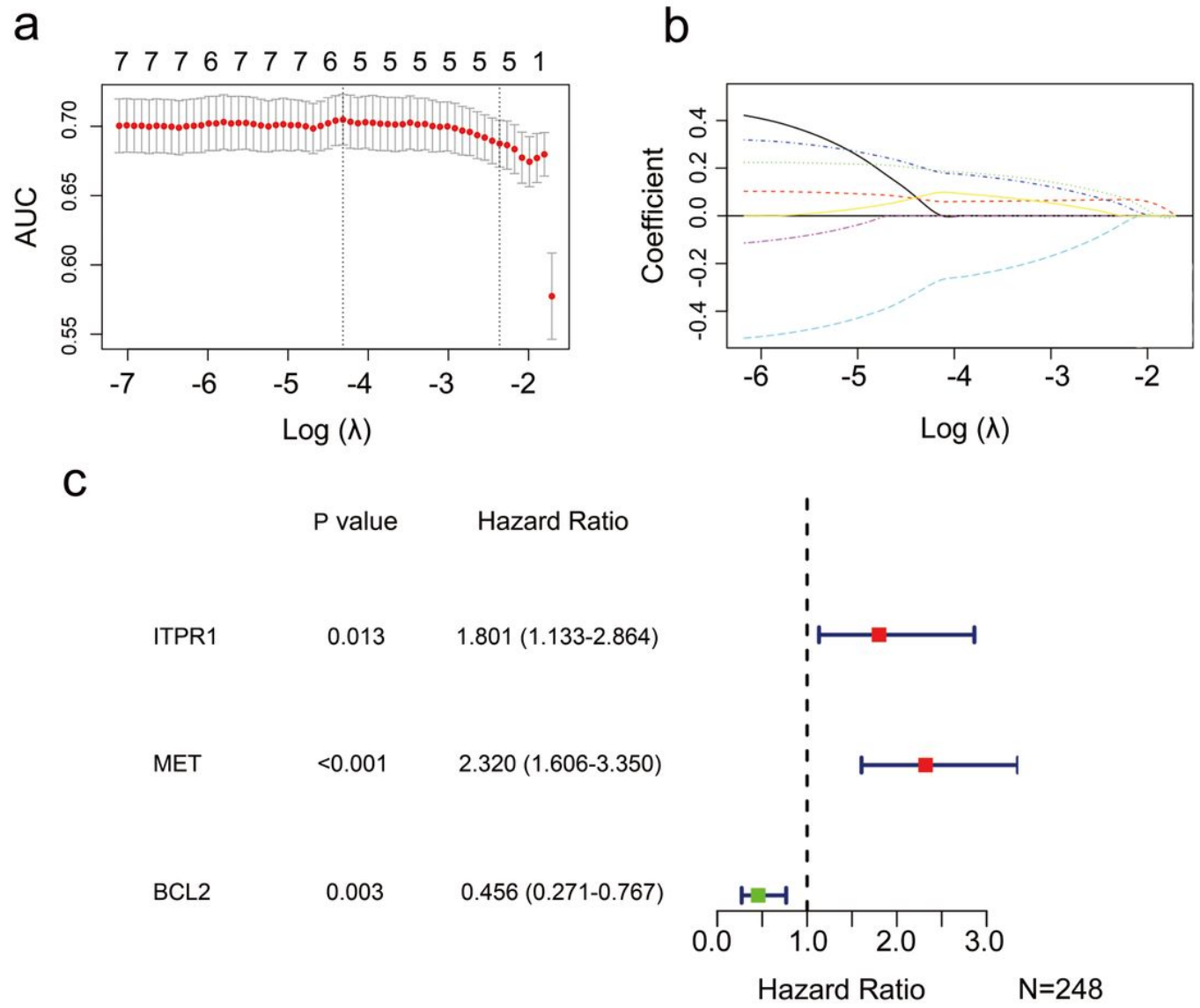

Figure 5

Drawing three genes most relevant to lymph node metastasis by LASSO algorithm and multivariate Logistic regression. a The optimal parameter (λ) was chosen by cross validation. The dashed vertical line on the left intersects over the best log λ, corresponding to the maximum value of AUC. b LASSO coefficient plot of 9 DEGs. Each curve represents a coefficient and the x-axis represents the regularization penalty parameter. Those coefficients that do not become zero as x changes are included in the LASSO regression model. c HRs and 95% CIs of the three genes based on multivariate Logistic regression analysis of the training cohort from TCGA-THCA dataset.

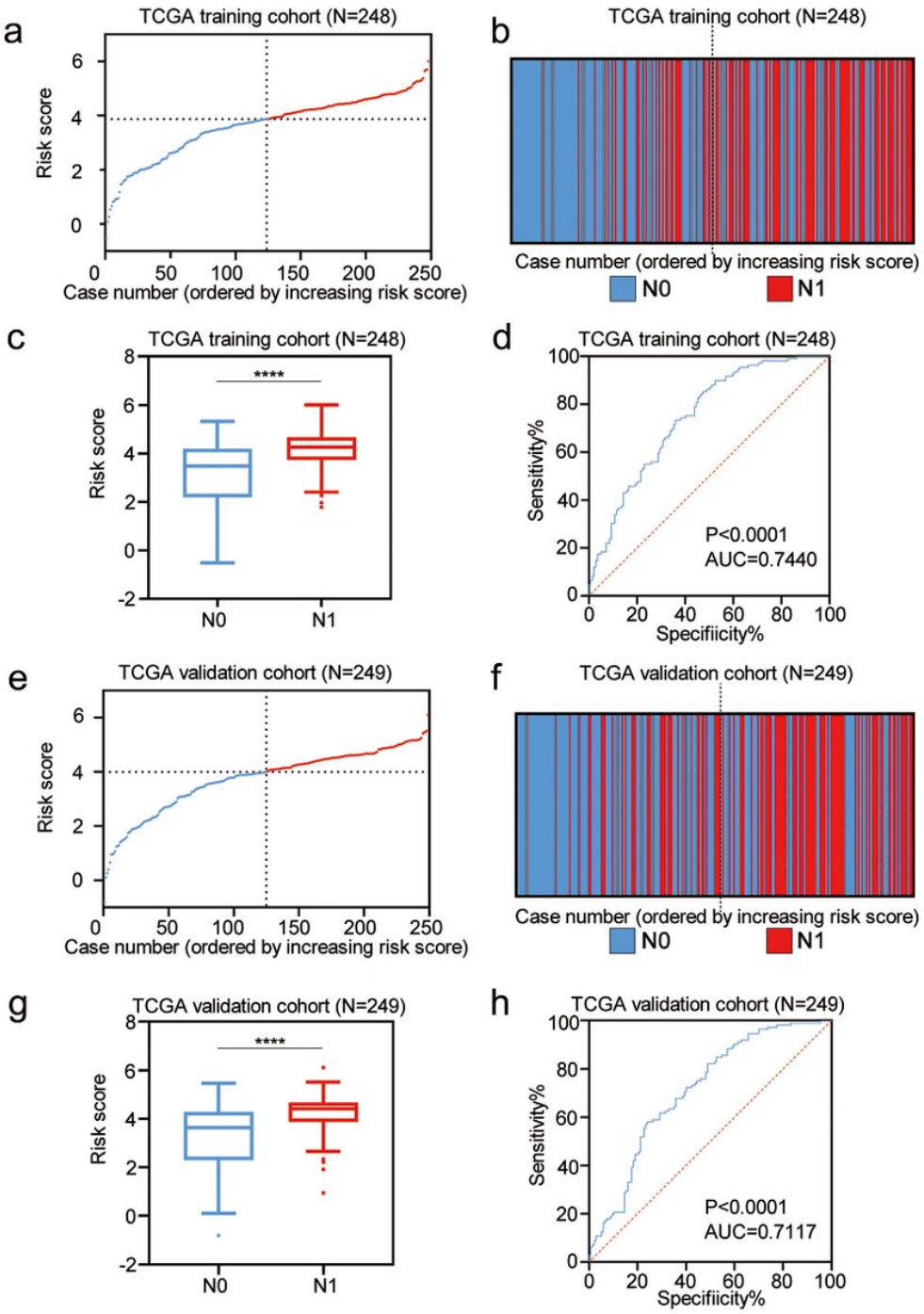

Figure 6

Obtainment of the risk score system from TCGA cohort. The median risk score, which served as the cut-off value for dividing the high-risk and low-risk groups, was presented as a horizontal dashed line. The vertical dashed line separated patients on the standard of the high-risk (red) and low-risk (blue) (a, e). The distribution of lymph node metastasis in the training (b) and validation (f) cohorts from TCGA was illustrated. Patients with lymph node metastasis were shown in red, while patients without lymph node metastasis were shown in blue. The result of student's t test of patients predicted to be at risk for poor outcomes in the training (c) and validation (g) cohorts from TCGA. The number of patients remaining at a particular timepoint was shown at the bottom. ROC curves for predicting Lymph node metastasis in the training (d) and validation (h) cohorts from TCGA. ****, P<0.0001. Error bars indicate mean ± SD.

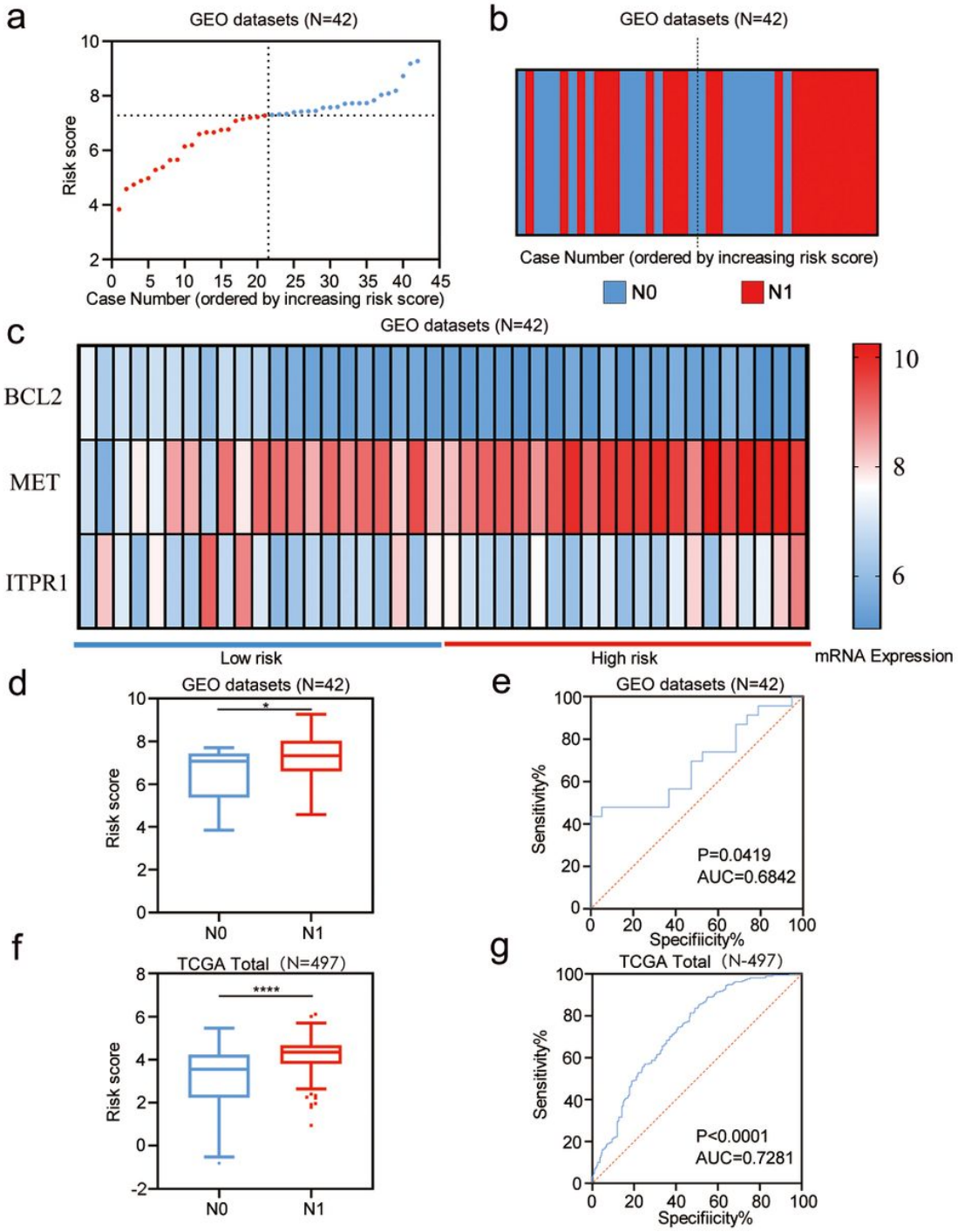

Figure 7

Verification of the risk scoring system through the GEO independent datasets. a The distribution of risk scores in independent GEO dataset. b The distribution of patients with (red) or without (blue) lymph node metastasis. Patients were ranked in ascending order of risk score. c The expression of the three genes in valuating system of all the papillary thyroid cancer patients in GEO datasets. Student's t test of the risk score between N0 and N1 patients in GEO dataset (d) and TCGA total dataset (f). ROC curves for predicting Lymph node metastasis in the GEO dataset (e) and TCGA total dataset (g). *, P value<0.05, ****, P value<0.0001. Error bars indicate mean ± SD.

Supplementary Files

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

- AdditionalFigureLegends.docx
- Additionalfig3.tif
- Additionalfig2.tif
- Additionalfig1.tif