5 Gleason score-associated gene signatures serve as novel biomarkers for identifying early recurring events and contributing to early diagnosis for Prostate Adenocarcinoma

Lingyu Zhang
Department of Laboratory Medicine
Yu Li
Department of Biochemistry and Molecular Biology
Weiwei Liu
Department of Laboratory Medicine
Xuchu Wang
Department of Laboratory Medicine
Ying Ping
Department of Laboratory Medicine
Danhua Wang
Department of Laboratory Medicine
Ying Cao
Department of Laboratory Medicine
Yibei Dai
Department of Laboratory Medicine
Zhihua Tao (zrtzh@zju.edu.cn)
The Second Affiliated Hospital of Zhejiang University School of Medicine https://orcid.org/0000-0001-7866-8463

Research

Keywords: Prostate Cancer, Gene signatures, Robust Rank Aggregation, Weighted Gene Co-expression Network Analysis, Disease-Free Interval, Inflammation landscape

Posted Date: July 14th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-41108/v1

Version of Record: A version of this preprint was published on April 30th, 2021. See the published version at https://doi.org/10.7150/jca.52170.

Abstract

Background: Prostate cancer (PCa) recurrence leads to much higher mortality than those without recurring events. Early and accurate laboratory diagnosis is particularly important for early identification of patients at high risk of recurrence and to benefit from additional systemic intervention. This study aimed to develop efficient and accurate Prostate Cancer diagnostic and prognostic biomarkers for the identification of initial tumor new events.

Methods: Gene Expression Omnibus (GEO) datasets and The Cancer Genome Atlas (TCGA) data portal were utilized to obtain differentially expressed genes (DEGs) and clinical trait information in PCa. WGCNA analysis obtained the most relevant clinical traits and genes enriched in several modules. Univariate Cox regression analysis and multivariate Cox proportional hazards (Cox-PH) model was employed to candidate gene signatures related to Disease-Free Interval (DFI). Internal and external cohort was utilized to test and validate the validity, accuracy, and clinical utility of prognostic models.

Results: We constructed and optimized a valid and credible model for predicting patient outcomes, based on 5 Gleason score-associated gene signatures (ZNF695, CENPA, TROAP, BIRC5, KIF20A). Furthermore, ROC and Kaplan-Meier analysis revealed higher diagnostic efficiency for PCa and predictive effectiveness in tumor recurrence and metastasis. Calibration curve also revealed high prediction accuracy in internal TCGA cohort and external GEO cohort. The model was prognostically significant in the stratified cohort, including TNM classification and Gleason score, and was deemed to be an independent PCa prognostic factor, and superioring to other clinicopathological characteristics. On the other hand, we also measured the correlation between gene signatures' expression and inflammation landscape. 5 gene signatures were significantly positively correlated with tumor purity and negatively correlated with the immersion levels of CD8+ T cells.

Conclusions: Our study identified and validated 5 gene signatures as biomarkers for prostate cancer diagnosis, providing an assessment of DFI while predicting tumor progression, possibly providing novel theories for the treatment of prostate cancer.

Introduction

PCa has been taken into account as the second most common malignant tumor in men worldwide, accounting for about 15% of male's new tumors. In Europe and the United States, prostate cancer has reached first place with an incidence rate of 1.051‰[1]. Although the incidence of prostate cancer in China is lower than that in the United States, it is increasing year by year. In 2018, the number of cases has reached 120,000, and it will be supposed to rise to 172,000 in 2022[2]. Prostate-specific antigens (PSA) have been recognized as an effective method for diagnosing prostate cancer and monitoring prognosis over the past few decades. And combined with imaging examination, the diagnostic performance was brought to a higher level. However, there are still limitations, especially at a lower concentration (4–10 ng/mL). The PSA value has a rather weak diagnostic efficiency [3–5].

With the development of chip technology and next-generation high-throughput sequencing (NGS), it has been noted that a variety of novel biomarkers which have been isolated from tumor tissues, serum, and even urine. A considerable number of reports have identified nucleic acid molecules such as mRNAs and ncRNAs (miRNAs, lncRNAs or circRNAs) associated with the pathogenesis and progression in PCa, as well as membrane proteins present in exosomes [6–10]. However, only a few have been approved by the US Food and Drug Administration (FDA) (PSA in 1994, PHI in 2012, and PCA3 in 2012). Ideal diagnostic biomarkers should have high specificity (correctly identify populations without specific diseases; true negative rates), high sensitivity (correctly identify groups with specific diseases; true positive rates), easy to use, highly repeatable, easy to acquire and quantify, with a clear understanding of results. If there are biomarkers available that differentiate patients with different degrees of risk of prostate cancer precisely, and thus identify men with a lower or higher likeliness of prostate cancer-related prevalence and mortality. They can guide more individualized and targeted interventions to avert over-treatment for inert cancers. Therefore, it is necessary to find novel biomarkers for diagnosis and monitoring of PCa.

Detection of gene expression is an effective method for identifying biomarkers because it can assess tumor activity and expression levels in separate tissue types. Potentially distinguishing various molecular subtypes based on gene expression. Even though both proteins and RNA provide information about the activity of a molecule, detection and quantification of RNA are often much easier, even in trace amounts and complex matrix environment. Furthermore, the multiplicity of RNA-based assays is fairly simple, which implies that thousands of potential targets can be evaluated simultaneously using high-throughput assays. RNA-based prostate cancer biomarkers discovery methods include ncRNA analysis, multi-gene expression panels, alterations in the presence of splice variants, and gene fusion transcripts premised on various tumor cell functions. There are almost no symptoms in the primary stage of PCa, and a few are found in randomized physical examinations. Most patients are in the process of the advanced stage before disease-related signs began to emerge, losing the possibility of cure and chance of survival. Most patients with PCa do not die from tumors at the primary site, but rather complications due to tumor spread to the bone marrow and other internal organs. Some evidence suggests that only a quarter of patients with metastatic and invasive PCa can survive more than 5 years after an initial diagnosis of metastasis [11–13]. Notably, invasive prostate cancer boosted into metastatic disease after local treatment, accounting for about two-thirds of the death brought about by prostate cancer, especially CRPC (castration-resistant prostate cancer) [12, 14, 15].

Despite the prevalence of PCa, there is no unambiguous and precise diagnosis or prognostic biomarkers to distinguish this aggressive tumor. Despite the use of serum PSA concentrations as a routine screening technique for PCa, up to 11% of men with PSA 2.0 ng/mL still suffer from PCa, and when PSA is higher than 4 ng/mL and below 10 ng/mL, the diagnosis is challenging and difficult to identify, making it difficult to evaluate the presence or absence of PCa depending on PSA only [11]. The ratio of fPSA (free PSA) to tPSA (total PSA) is an effective way to enhance diagnostic sensitivity and specificity, but it remains limited. Drawing a distinction between low-risk and high-risk PCa patients and improving screening performance becomes difficult and impossible to rely solely on serum levels [16–18]. The new generation of genome detection technology, such as microarray analysis and NGS, deepening our

comprehension of the biological mechanisms of PCa. As a result, the scientific community confronts with opportunities for data explosions, but contemporary challenges for biomarkers discovery and verification. With the improvement of biomarker research methods, combined with low cost and more effective technologies, the potential of personalized genomic diagnosis for clinical decision-making has proved possible in recent years. As one of the most prominent themes in PCa genetics, characteristic changes in the somatic genome of tumor tissue to diagnose disease and predict treatment response are considered a novel proposition. Novel methods include genetic analysis from peripheral blood, germline analysis or DNA/RNA characterization of free circulating nucleic acids or circulating tumor cells (CTCs) [19–23]. Despite the fact that further understanding of the molecular basis of PCa occurrence has produced more prognosis and prediction measures, it still fails to resolve the early identification of invasive PCa. The genetic pathways and/or gene expression panels that forecast PCa prognosis and responses to interventions are promising.

Considering that early diagnosis and prediction can benefit patients in systematic intervention, there is imperative to establish gene signatures based on tumor recurrence in early PCa patients. In this study, we integrated PCa cases with disease-free interval (DFI) data from two independent cohort studies, including TCGA-PRAD and GSE116918, to develop and verify novel personalized genetic signatures. We also investigated clinical and pathological features and immune infiltration landscape. The preliminary construction of prognostic-related gene signatures for PCa patients at an early stage will help clarify the complex underlying mechanism between gene expression and PCa recurrence.

Materials And Methods

Selection of PCa gene expression datasets

All GEO microarray datasets contained in the present study were downloaded from the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/) [24]. The inclusion criteria were as follows: 1) All datasets contained both prostate cancer tissue and corresponding adjacent normal prostate tissue, excluding benign proliferative prostate tissue; 2) the data set contained sufficient sample size; 3) The complete probe names (or probe sequences) and corresponding gene symbols. Based on these criteria, GSE3325 [25], GSE6956 [26], GSE32448 [27], GSE3251 [28], GSE46602 [29], GSE55945 [30], GSE34312 [31], GSE69223 [32], GSE71016 [33], GSE88808 [34] were selected and used for further research. PCa RNA-sequencing profiles and clinical data were consulted on the TCGA database (https://cancergenome.nih.gov/) and utilized in the study.

RRA method for identifying reliable DEGs

As a novel method, Robust Rank Aggregation (RRA) can detect genes that are always better ranked than expected under the null hypothesis of unrelated input and assign a significant score to each gene. The significant score also serves as a rigorous method to reserve statistically relevant genes in the final list. Potential probability models make algorithm parameters free and robust to outliers, noise, and errors. The DEGs analyzed by this method are always better than expected, and a P-value was posted to each DEG.

Bonferroni corrected P values to minimize false positive results[35]. Before gene lists integration and meta-analysis using RRA, the R package “limma” was utilized to normalize the GEO data to reduce errors caused by chip technology after background signal processing and data cleaning, the modified value reflected the true expression level of the genes [36].

Gene function enrichment analyses

R package “Clusterprofiler”[37] was used to investigate the functions and pathways of the candidate DEGs. Gene Ontology (GO) enrichment and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis at the platform was applied to further understanding the biological mechanisms behind DEGs. GO terms or KEGG pathways with P-value < 0.05 and FDR < 0.05 (false discovery rate) were considered statistically significant. The visualization of the results was reached with the R packages "Clusterprofiler". Here we only show the top ten GO terms enriched in molecular function (MF), biological processes (BP), and cell composition (CC).

WGCNA analysis

WGCNA was a method for analyzing gene expression patterns in multiple samples. If a class of genes always has similar expression changes in a physiological process or different tissues, these genes can be clustered and defined as a module and further investigated the relevance between co-expression modules and clinical traits [38]. We obtained the expression profiles and clinical parameters of these DEGs from the TCGA data portal and merged them. In this study, we incorporated seven clinical traits including T grade, N grade, M grade, Diagnosis age, Recurrence, Plymphnode number, and Gleason score. In the R platform, the R package "WGCNA" was utilized to integrate DEGs expression profiles and clinical traits to establish a relationship between weighted gene co-expression modules and clinical traits. In the WGCNA algorithm, the elements in the defined gene co-expression matrix were the weighted values of the correlation coefficients of the genes, and the selection threshold of the weighted value was such that the connections between the genes included in each gene network follow a scale-free network distribution. In our study, the β value is the soft threshold (power), and the scale-free topological fitting index (scale-free R^2, ranging 0 ~ 1) was used to definite the scale-free topology model, and the higher scale-free R^2 value insinuated a better fitting degree. When β value (ranging 1 ~ 40) was at least 6, the corresponding scale-free R2 value was 0.96. In the clustering tree, genes with high absolute correlation were assembled into the invariable co-expression module, and a cluster tree was generated by FlashClust analysis. Genes were divided into distinct gene modules based on TOM-based difference metrics, for relating modules to one another and external clinical traits; for calculating module Gene Significance (GS) and Module Membership (MM) measures to identify significant modules. The relationship between the modules and the 7 clinical traits was visualized by a heat map with P-value < 0.05. The modules with the highest correlation with clinical traits were engaged to explore their biological functions through GO and KEGG analysis.

Construction of Cox-PH regression model and identification of biomarkers based on DFI

According to the introduction of the TCGA database, when patients have a new tumor event whether it is a locoregional recurrence, distant metastasis, biochemical evidence of disease or new primary tumor of cancer, all of the above were defined as positive new tumor events. The number of days from the initial follow-up to the appearance of a new event was defined as DFI. Based on the DFI information in TCGA-PRAD (Prostate adenocarcinoma), univariate Cox regression analysis was performed to analyze the association between the DEGs included in the black module with tumor treatment outcome using the package “survival” in R platform. By performing univariate Cox regression, 113 genes with P-value <0.05 were selected for subsequent analysis. Rely on these prognosis-related DEGs, as previously described, a Cox-PH model (Cox's proportional hazards regression model) was applied after Lasso regression analysis to select the optimal panel of prognostic gene signatures [39-41]. The optimal lambda was definited after running 1,000 stimulations via cross-validation likelihood, and 10 genes were included in the subsequent multivariate Cox-PH regression model. By using the coefficients from multivariate Cox regression as weight, a risk score prediction model based on gene signature expression is set by a linear combination of expression levels of independent gene signatures. The risk-factors scoring algorithm was generated for prognosis prediction as follows: Risk-score = Σ (βmRNA × exprmRNA). Something needs to be noted that βmRNA denotes the Cox-PH coefficient of mRNA while exprmRNA denotes the mRNA expression levels. Based on the risk scoring prediction model, PCa patients were subdivided into low-risk and high-risk groups with the best risk score as the cut-off point. DFI differences between low-risk patients and high-risk patients were calculated by Kaplan-Meier survival curves and compared using a log-rank test. Time-dependent ROC (receiver operating characteristic curves) were used to assess the predictive efficiency of PCa prognosis and the area under the curve (AUC) of the ROC was used to estimate the prediction accuracy of the model. We have incorporated each independent predictors selected by the multivariate Cox-PH regression model to produce a nomogram using the "RMS" package [42]. Risk curves and scatter plots were generated to reflect the risk score and new event risk for each PCa patient in TCGA. GSE116918 [43] served as an external verification cohort to expose the robustness of the prognostic model.

External validation of the expression levels of 5 gene signatures

Gene signatures expression in different tissues and different Gleason scores was verified using the TCGA dataset and the protein expression pattern was verified in Human Protein Atlas [44]. The GSE32269 [45] dataset was used to evaluate the expression of these 5 gene signatures in primary prostate cancer and metastatic prostate cancer. The ROC curve was used to discriminate prostate cancer from normal prostate tissue. The selection of the cutoff threshold was based on the maximization of Youden's index (Sensitivity + Specificity -1). A P-value of less than 0.05 was regarded as significant, and all statistical tests were two-sided.

Gene set enrichment analysis

Based on the optimal separation, we separated the TCGA-PRAD sample into high-risk and low-risk groups. The software GSEA was used to identify differential gene expression patterns between the two

subgroups, with a cutoff point of <0.05 for both P-value and FDR. The top four enriched gene sets in the high-risk group and low-risk group are displayed in the results, respectively in the Gene Oncology gene set and Kyoto Encyclopedia of Genes and Genomes gene set.

Evaluation of relationships between gene signatures and the immune microenvironment

By utilizing CIBERSORT（http://cibersort.stanford.edu/）[46] and TIMER (https://cistrome.shinyapps.io/timer/) [47] (two genes expression-based deconvolution algorithm). The CIBERSORT algorithm was an analytical tool that utilized standardized RNA-seq data to measure changes in immune cell abundance and obtain the proportion of various types of immune cells from tissue samples. CIBERSORT provides 22 kinds of cells including monocytes, NK T cells, B cells, T cells, and so on. TIMER is a comprehensive resource for the systematic analysis of immune infiltration in multiple cancer types. The abundance of six immune infiltrates (B cells, CD4+ T cells, CD8+ T cells, Neutrophils, Macrophages, and Dendritic cells) was valued by the statistical method, which is validated using pathological estimation.

Statistical Analysis

Results are expressed as mean ± standard error of the mean. After comparing the two groups, an unpaired Student's t-test was conducted to determine statistical significance. When the data is normally distributed, the two-way t-test is used to determine the group comparison using unequal variance or paired t-test; otherwise, a two-sided Wilcoxon test is used. The multiple hypothesis tests were performed using the Benjamini and Hochberg methods unless otherwise stated. Statistical analysis was performed using RStudio v1.2.1335 (RStudio Inc.) and GraphPad Prism software v8.0 (GraphPad Software Inc.). During the experiment and outcome evaluation, $P < 0.05$ was deemed to indicate a statistically significant difference.

Results

Incorporating GEO datasets and screening DEGs using RRA method

To describe our study more clearly, a flow chart of the analytical procedure was developed in Fig. S1. To be more universally applicable to the GEO datasets included in this study, we picked the 10 most representative GEO datasets and used them for subsequent RRA analysis. The important characteristics of these datasets such as ID, platform, the number of raw probes, and the number of samples are shown in Table 1. To allow each group measurement or measurement under numerous experimental conditions to be compared with each other, non-experimental differences between measurements such as product preparation, hybridization processes, or hybridization signal processing are eliminated. We first normalized the GEO dataset, and these results are illustrated in Fig. S2a-j. We used $|logFC| \geq 1$ and P-value < 0.05 as screening criteria. Depending upon RRA analysis, a total of 1128 up-regulated genes and 962 down-regulated DEGs were identified. Up-regulated genes and down-regulated DEGs ranked in the top 20 are given in Fig. 1 based on logFC values. The top 400 DEGs based on logFC (200 up-regulated genes and 200 down-regulated genes) were used for GO and KEGG analysis. Top 10 GO terms that enrich in

molecular function (MF), biological processes (BP), and cell composition (CC) are shown in Fig. 2a. As the picture shows, these genes are remarkably enriched into the three GO terms of the muscle system process, extracellular matrix, and the actin-binding. As for the KEGG pathway, DEGs are significantly enriched human T-cell leukemia virus 1 infection, Epstein-Barr virus infection, and Ras signaling pathway, etc (Fig. 2b).

Wgcna Analysis And Identification Key Module

The up-regulated genes ranked in the top 5000 in the RRA analysis were selected on the basis of WGCNA analysis. To determine key modules involved in these DEGs and clinical traits in PCa, we merged clinical traits and expression profiles of prostate cancer from TCGA, these clinical traits include TNM grade, diagnosis age, recurrence, number of metastatic lymph nodes, and Gleason score. Samples dendrogram and traits heatmap were revealed in Fig. 3a. Independence and the average connectivity degree of the co-expression modules were established by power value (β) and scale R2 value. By setting a soft-thresholding power at 6 (scale-free R^2 = 0.96) (Fig. 3b-c) and cutting height as 0.20 (Fig. 3d). Therefore, we define the adjacency matrix using a soft threshold to construct and identify distinct co-expression gene modules in PCa samples. Inspired by the TOM-based dissimilarity measure, a clustering tree diagram of all selected genes was developed. These identified co-expression modules are assigned in different colors (Fig. 3e). The interactions of these co-expression modules were evaluated with the Pearson correlation coefficient (Fig. 3f). From the heat map of the module feature correlation, we found that the black module occupied the highest correlation with the Gleason score (correlation coefficient = 0.41, P = 4E-24) (Fig. 4a). Most interestingly, a scatter plot of gene significance (GS) and module membership (MM) was attracted in the co-expressed black color module. The results consistently revealed that MM in the black module significantly correlated with Plymphnodes Number (correlation coefficient = 0.7, P = 3e-32), Recurrence (correlation coefficient = 0.76, P = 8.6e-41), and Gleason score (correlation coefficient = 0.92, P = 1.5e-86) (Fig. 4b). We extracted a total of 227 genes from the black module. To reveal the hypothetical biological functions of these genes, we performed GO and KEGG analyses based on these genes in the black module. We detected significant enrichment of these genes in several GO terms in chromosome segregation, mitotic nuclear division, nuclear division, organelle fission, mitotic sister chromatid segregation-22 (Fig. 4c). As for the KEGG pathway analysis, these genes are allocated into three terms, Cell cycle, Tight junction, and Leukocyte transendothelial migration (Fig. 4d).

Identifying Hub Genes Associated With Dfi

Given the results in Fig. 4, we speculated that genes in the black module may be closely related to the tumor outcome (Gleason score, recurrence, lymph node metastasis, etc.) in PCa. To reduce the scale of the number of genes in the black module, we used a univariate regression analysis based on DFI to calculate the association between the expression level of each gene and the patient's DFI. In order to avert the overfitting of the predicted model, the Lasso regression was performed to screen the genes associated with DFI. Supported by the hazard ratio according to P-value < 0.05, 113 genes were included in the Lasso regression model. According to the results of Cross-validation for tuning parameter selection

in the proportional hazards model, 5 genes were incorporated into the multivariate Cox regression model (Fig. 5a-b). In multivariate regression analysis, the hazard ratio of these 5 genes was based on the risk score of the new event status and time shown in Fig. 5c (Concordance Index = 0.78). We also constructed an inclusive nomogram containing hub genes expression (log2) and prognosis status, providing clinicians with an effective tool to predict the risk of new events among PCa patients (Fig. 5d). For the prognostic model, 495 PCa samples from TCGA were randomly classified into two subsets, 30% of which were used for the discovery cohort, and the other 70% were used as the validation cohort. In order to evaluate the robustness of the prognostic model, the Cox-PH model based on 5 gene signatures was used to test in the discovery cohort, validation cohort, and total cohort. Consistent with the discovery cohort and validation cohort, risk curves and scatter plots were generated to show the risk score and new event risk for each PCa patient in the total cohort. Patient clinicopathologic characteristics were revealed in Table 2. Patients in high-risk groups have a higher probability of new events than those in low-risk groups (Fig. 6a-f). Then, the Kaplan-Meier curve revealed that the prediction model of risk score had reliable discrimination ($P < 0.001$), and patients with high-risk scores have a shorter disease-free interval (Fig. 6g-i). To assess the reliability of the risk prognostic model, we mapped the time-dependent ROC curve to assess DFI in PCa patients for three or five years. Results showed that, for predicting recurrence at 3 and 5 years, the 5 gene signatures had an area under the curve (AUC) values of 0.784, and 0.758, respectively (Fig. 6j-l). With the above results, we finally determined that 5 gene signatures (ZNF695, CENPA, TROAP, BIRC5, KIF20A) are closely related to the tumor outcome of prostate cancer and can be used as independent prognosis risk factors.

5 Gleason score-associated gene signatures are independent risk factors for tumor recurrence in PRAD

Based on the previous prediction model, the calibration curve shows good agreement between prediction and observation (Fig. 7a-b). Various clinical factors affected the prognosis of PRAD patients. In order to verify whether the 5 gene signatures can predict the prognosis of patients independently of other clinical factors. Complete clinical characteristics carried by the TCGA-PRAD data set (including patient age, T classification, N classification, M classification, Stage classification, cancer status, and New tumor event) in Table 3. The correlation between different risk stratification and clinicopathological characteristics according to risk score was using the chi-square test. In addition to the patient's age, other clinical characteristics are significantly associated with poor prognosis. Forest plots for univariate and multivariate Cox regression analysis showed that the risk score can be independent of other clinical characteristics and superior to them (Fig. 7c-d, f). The distribution of clinicopathological characteristics and expression of gene signatures in low-risk and high-risk groups was displayed in Fig. 7e. KM-plot showed that patients in the T3-4 classification have a shorter Disease-free interval, compared to T1-2 ($P < 0.001$)(Fig. 8b). Similarly, patients with higher pathological grades have shorter Disease-free Interval, compared with lower-grade pathological classification (N1 vs N0, $P < 0.001$; M1 vs M0, $P = 0.016$; Gleason > 7 vs Gleason < = 7, $P < 0.001$) (Fig. 8c-f). Age cannot be an independent factor in predicting patient risk, Age cannot be an independent factor in predicting patient risk, as shown in Fig. 8a. To verify the authenticity of the above prognostic model, we constructed another prognostic model using the external GEO cohort (GSE116918) (248 samples). According to the TCGA cut-off value, the samples were divided

into low-risk (n = 146) and high-risk groups (n = 102) according to the risk score (Fig. 9a-b). K-M survival analysis shows that low-risk patients have longer disease-free intervals, compared to high-risk groups (Fig. 9c). Time-independent ROC curve and calibration curve show that the prediction model has high accuracy, for predicting tumor new events at 3 (AUC = 0.837) and 5 years (AUC = 0.857) in Fig. 9d-f. In order to ascertain the universal applicability of the prediction model, we conducted a survival analysis by stratified analysis based on the entire TCGA-PRAD cohort. In the two tiers of Age < = 60 and > 60, the risk-free interval of the high-risk group was considerably lower than that of the low-risk group (Fig. 10a-b). Similar significant results were revealed in different TNM classifications and Gleason scores stratification in Fig. 10c-j.

Biological phenotype and inflammatory landscape related to risk scoring model

To identify the biological pathways and processes in different risk stratifications, we performed GSEA to clarify the biological function of the prognostic model. As shown in Fig. 11a-b, for the GO and KEGG pathways, genes that are highly expressed in the high-risk group show significant enrichment in multiple biological processes, such as meiotic chromosome segregation, chromatin remodeling at centromere, homologous chromosome segregation, cell cycle, homologous recombination, DNA replication, etc.

Validation of 5 gene signatures in PCa

We got expression profiles of PCa samples from TCGA to verify the relationship between these five gene signatures and clinical traits. Unsurprisingly, these five gene signatures are highly expressed in prostate cancer tissue, as opposed to the corresponding normal prostate tissue (Fig. 12a). Validation in the Human Protein Atlas revealed that the protein levels of these 5 gene signatures were significantly higher in PCa tissues than those in paracancerous normal tissues (Fig. 12b). Built on the results of the WGCNA analysis, we can conclude that the 5 gene signatures from the black module had the highest correlation with the Gleason score in PCa. We assessed the expression of five gene signatures in five discrete levels of Gleason score (6, 7, 8, 9, 10), indicating that the higher the level of expression represented the higher the Gleason score (Fig. 12c). Additional independent GEO microchip dataset GSE32269 was used to assess the expression levels of these five gene signatures in primary and metastatic tumors (bone metastasis or lymph node metastasis), and more interestingly, as it was shown in Fig. 12d, these genes had higher expression levels in metastatic tumors. Perhaps these results indicate that higher levels of expression confer a tendency for tumor cells to metastasize from in-situ to a distant location. Regarding prognosis, Kaplan-Meier curves show that higher expression of these genes is considerably associated with poor DFI (Fig. 12e). Notably, all of these gene signatures have higher diagnostic efficiency with higher AUC (Fig. 12f). We obtain the cutoff value (log2) of each gene by calculating the maximum Yuden index. It is gratifying that these five gene signatures have high specificity and sensitivity at the set threshold based on the cutoff value, and all these results are shown in Table 3. When these five gene signatures are combined, the diagnostic model displays a very high specificity and sensitivity in distinguishing normal prostate and prostate cancer tissues, with an AUC of up to 0.9473 (95% CI = 0.9149 ~ 9698) (Fig. S3).

Association of gene signatures expression with tumor immune infiltration

We hypothesized that different prostate tissues will have diverse immune cell infiltration components, and CIBERSORT and TIMER use unique genetic features to assess immune cell infiltration abundance in each prostate cancer sample. Fig. 13a displayed the proportion of 22 immune cells calculated by the CIBERSORT algorithm in PCa tissues based on TCGA-PRAD. Among the 22 immune cell types, the violin plot (Fig. 13b) depicted results of the Wilcoxon rank-sum test, which showed that the infiltration of T cells CD8 (P = 0.026), T cells CD4 memory resting (P=0.047), NK T cells activated (0.037), Macrophages M0 (P=0.044), and Mast cells resting (P=0.033) were significantly different between the high-risk group and the low-risk group. We applied the TIMER tool to investigate tumor purity and the infiltrating immune cell landscape of PCa (B cells, CD4+ T cells, CD8+ T cells, neutrophils, macrophages, and dendritic cells) in the context of mRNA expression of the 5 gene signatures. Notably, 5 gene signatures of ZNF695, CENPA, TROAP, BIRC5, and KIF20A were significantly correlated with tumor purity and negatively correlated with CD8+ T cells, and had a lower negative correlation with macrophages. Conversely, no or weak associations were observed among these 5 genes and infiltration of B cells, CD4+ T cells, neutrophils, and dendritic cells (Fig. 13c).

Discussion

Prostate cancer, with the morbidity and mortality in Europe and the United States, accounted for the top three of all cancers, while in China, morbidity and mortality accounted for the top ten. However, these Fig.ures have shown a significant upward trend as the population ages and the diet changes [48–50]. Despite the fact that further comprehension of the molecular basis of PCa tumorigenesis has produced more diagnostic and therapeutic measures, early diagnosis and prognosis evaluation of PCa, especially tumor recurrence, has not been well addressed. As a heterogeneous tumor with a multifaceted mechanism, the imbalance of some key genes will lead to different or even opposite prognostic effects on different treatments in PCa. A large number of studies using microarrays and RNA-seq to discover novel biomarkers and therapeutic targets for PCa are an effective method. Therefore, we focus on the value of genomic markers in the individualized prediction of PCa consequences and responses to various therapeutic interventions. As our study has shown, genes that were consistently differentially expressed in prostate cancer tissues and corresponding normal tissues in 10 GEO databases. WGCNA and hierarchical clustering analysis determined that some genes are highly correlated with the clinical traits of multiple PCa in the black module. Based on Disease-Free Interval and new tumor events, the univariate and multivariate Cox proportional hazard regression identified 5 DFI-related gene signatures (ZNF695, CENPA, TROAP, BIRC5, KIF20A), these genes can discriminate differences between normal and tumor tissues, as well as differences between primary and metastatic prostate cancer. ROC curves indicated that these 5 gene signatures may be potential biomarkers for the diagnosis of PCa with higher sensitivity and specificity, and can be considered as objective indicators for the prognosis of PCa patients.

ZNF695 is responsible for encoding a class of zinc finger proteins of unknown function, and there are rare reports on it, especially in prostate cancer [51]. The pathophysiological function of ZNF695 in normal cells and tumor cells is still not clear. In this study, we provided a series of evidence that ZNF695 is

overexpressed in PCa and may lead to tumor progression and metastasis. CENPA, a histone H3 variant, is a fundamental determinant of centromere identity and is thought to play a central role in guiding kinetic assembly and centromere function, Participating in the regulation of chromosome segregation during cell division [52, 53]. Ashish B. R et al reported that CENPA is increased in breast cancer tissue and associated with shorter DFS (Disease-free survival) [54]. While CENPA overexpression-mediated pRb depletion is involved in the development and progression of retinoblastoma [55].TROAP has been reported to be dysregulated in various tumors such as breast cancer, liver cancer, prostate cancer, and gastric cancer, and participates in the promotion of tumor cell proliferation and distant metastasis through multiple signaling pathways such as WNT3/survivin [56–60]. Our study confirms that TROAP may play a major role in the involvement of PCa metastasis and is also an independent predictor of risk for new prostate cancer events. BIRC5 (Survivin) is a dual cell function protein that directly regulates apoptosis and filament-division of embryonic cells during embryo development, and directly regulates apoptosis and filament-division of cancer cells during tumor occurrence and metastasis. It has been recorded to be closely related to the adverse prognosis of malignant peripheral nerve sheath tumors, renal cell carcinoma, lung adenocarcinoma and ovarian cancer [61–64]. Kinesin family member 20A (KIF20A) is a mitochondrial-related kinesin (MCAK) and is the most representative member of the kinesin-6. It participates in microtubule disaggregation, bipolar spindle formation, and chromosome segregation, which regulates mitosis and the cell-cycle [65, 66]. A reduction or loss of KIF20A expression will disrupt the normal mitotic process. All of these are regarded as potential causes of tumorigenesis. In fact, the role of KIF20A in numerous tumors has been previously reported, such as tongue cancer, colorectal cancer, breast cancer, and the like [67–70]. Our research suggests that the combination of the 5 gene signatures can be used as an indicator for risk stratification, contributing to predict the prognosis of PCa. The identified prognostic biomarkers may provide basic information about individualized treatment decision-making for individual patients and improve treatment outcomes.

For some time, a large number of convincing data indicate that the microenvironment in which cancer cells are located plays a pivotal role in the development and progression of cancer. Even microenvironmental changes are important factors leading to tumorigenesis and may affect development. The most common metastatic site for prostate cancer is bone, and the incidence of advanced disease is 65–80% [71, 72]. Owing to the complex interactions between the microenvironment and tumor cells, the skeletal microenvironment makes it easy to metastasize to castration-resistant prostate cancer (CRPC). Bone marrow, besides cell precursors, contains different types of recirculating mature immune cells, including Dendritic cells (DC), macrophages, different T and B lymphocyte subsets, myeloid-derived suppressor cells (MDSCs), and NK cells. Some of these leukocytes participate in the pathogen clearance and anti-tumor processes [73, 74]. We utilized CIBERSORT and TIMER to assess the immune infiltration landscape between different risk stratifications. Some tumor immune-associated leukocyte subsets are significantly dysregulated in PCa with a higher risk of recurrence and metastasis, especially CD8 + T cells. As an important immunomodulator, elevated CD8 + tumor-infiltrating immune cells correlate with prolonged survival in several types of tumors [75–77]. Our study confirms that abnormal expression of these 5 gene signatures is accompanied by loss of CD8 + T cells in PCa samples

with high risk, plays an important role in tumor immune escape and the formation of tumor microenvironment, and these results may provide new insights into interventions for early tumor recurrence.

Besides, we also evaluated the abnormal expression of these five gene signatures in other types of tumors based on the TCGA data portal, perhaps these genes play a critical role in the development of multiple tumors (Fig. S3). As for whether these genes can be used as biomarkers for other tumors, further exploration is needed.

Conclusion

Overall, the analytic methods used in this study enabled us to identify a specific set of gene signatures for the diagnosis of PCa and the definition of patient outcomes through a comprehensive analysis of different bioinformatics datasets. However, the results obtained from this study represent only the starting point for determining the effective marker for PCa. Therefore, further experimental and functional studies are required to complete a large number of samples to assess the expression levels of these biomarkers and to verify their predictive effects on PCa and specific mechanisms.

Abbreviations

GEO: Gene expression omnibus database; TCGA : The Cancer Genome Atlas; DEGs: Differentially expressed genes; AUC: Area Under roc Curve; CI: Confidence Interval; RRA: Robust Rank Aggregation; WGCNA: Weighted Gene Co-expression Network Analysis; DFI: Disease-Free Interval; MSigDB: Molecular signatures database; GSEA: Gene set enrichment analysis.

Declarations

Acknowledgements

Not applicable.

Funding

The present study was supported by grants from the from the National Natural Science Foundation of China Youth Science Foundation Project (Grant nos. 81802571); Zhejiang Medical and Health Science and Technology Project (2019RC039); the National Natural Science Foundation of China (Grant nos. 81902156) and the Natural Science Key Project of Bengbu Medical College (No.BYKY2019012ZD).

Conflict of interest disclosures

All authors participating in the study stated that they had no competing economic interests.

Authors contributions

Lingyu Zhang, Yu Li and Zhihua Tao conceived and designed the study. Yu Li searched a large number of databases and incorporated a series of datasets available for the study. Lingyu Zhang, Weiwei Liu analyzed the datasets and was responsible for the writing of this manuscript. Lingyu Zhang, Yu Li and Xuchu Wang were under the responsibility of the production of Fig.ures. Lingyu Zhang, Ying Ping, Danhua Wang, Ying Cao, Yibei Dai and Pan Yu has searched a large number of literature and were responsible for reference compilation. Zhihua Tao gave a lot of guidance on manuscript writing. All authors reviewed and considered the final manuscript.

Consent for publication

Not applicable.

Data availability statement

The analysis datasets generated by the current study can be obtained from the corresponding author for reasonable reasons.

Ethics approval and consent to participate

Since the identities of patients in the TCGA and GEO databases cannot be identified, no approval and informed consent from the institutional review board is required.

Consent for publication

Not applicable.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer J Clin. 2018;68:394–424.
2. Committee CAAU. Chinese experts consensus on the treatment of metastatic prostate cancer 2018 edition. Chinese Journal of Surgery. 2018;56:646–52.
3. Cooperberg MR, Brooks JD, Faino AV, Newcomb LF, Kearns JT, Carroll PR, et al. Refined Analysis of Prostate-specific Antigen Kinetics to Predict Prostate Cancer Active Surveillance Outcomes. Eur Urol. 2018;74:211–7.
4. Thomsen FB, Brasso K, Berg KD, Gerds TA, Johansson J, Angelsen A, et al. Association between PSA kinetics and cancer-specific mortality in patients with localised prostate cancer: analysis of the placebo arm of the SPCG-6 study. Ann Oncol. 2016;27:460–6.
5. Mahal BA, Yang DD, Wang NQ, Alshalalfa M, Davicioni E, Choeurng V, et al. Clinical and Genomic Characterization of Low–Prostate-specific Antigen, High-grade Prostate Cancer. Eur Urol. 2018;74:146–54.

6. Alhasan AH, Scott AW, Wu JJ, Feng G, Meeks JJ, Thaxton CS, et al. Circulating microRNA signature for the diagnosis of very high-risk prostate cancer. PNAS. 2016;113:10655–60.
7. Ylipaa A, Kivinummi K, Kohvakka A, Annala M, Latonen L, Scaravilli M, et al. Transcriptome Sequencing Reveals PCAT5 as a Novel ERG-Regulated Long Noncoding RNA in Prostate Cancer. Cancer Res. 2015;75:4026–31.
8. Haldrup C, Lynnerup A, Storebjerg TM, Vang S, Wild P, Visakorpi T, et al. Large-scale evaluation of SLC18A2 in prostate cancer reveals diagnostic and prognostic biomarker potential at three molecular levels. Mol Oncol. 2016;10:825–37.
9. Xu Y, Deng J, Wang G, Zhu Y. Long non-coding RNAs in prostate cancer: Functional roles and clinical implications. Cancer Lett. 2019;464:37–55.
10. Wang Y, Ji J, Wang B, Chen H, Yang Z, Wang K, et al. Tumor-Derived Exosomal Long Noncoding RNAs as Promising Diagnostic Biomarkers for Prostate Cancer. Cell Physiol Biochem. 2018;46:532–45.
11. Siegel RL, Miller KD, Jemal A. Cancer statistics. 2015. CA: A Cancer J Clin. 2015;65:5–29.
12. Robert B. Den KYEJ, Adam P. Dicker CDLL. Genomic Classifier Identifies Men With Adverse Pathology After Radical Prostatectomy Who Benefit From Adjuvant Radiation Therapy. J Clin Oncol. 2015;33:944–55.
13. Radka Stoyanova MTYT, Nicholas Erho YBSP, Gillies AAP. Prostate cancer radiomics and the promise of radiogenomics. Transl Cancer Res. 2016.
14. Wu JB, Shao C, Li X, Li Q, Hu P, Shi C, et al. Monoamine oxidase A mediates prostate tumorigenesis and cancer metastasis. J Clin Invest. 2014;124:2891–908.
15. Popiolek M, Rider JR, Andrén O, Andersson S, Holmberg L, Adami H, et al. Natural History of Early, Localized Prostate Cancer: A Final Report from Three Decades of Follow-up. Eur Urol. 2013;63:428–35.
16. Center MM, Jemal A, Lortet-Tieulent J, Ward E, Ferlay J, Brawley O, et al. International Variation in Prostate Cancer Incidence and Mortality Rates. Eur Urol. 2012;61:1079–92.
17. Heijnsdijk E, Kinderen A, Wever E, Isma G, Roobol-Bouts M, Koning H. Overdetection, overtreatment and costs in prostate-specific antigen screening for prostate cancer. Brit J Cancer. 2009;101:1833–8.
18. Lacher DA, Hughes JP. Total, free, and complexed prostate-specific antigen levels among US men, 2007–2010. Clin Chim Acta. 2015;448:220–7.
19. Narayan VM. A critical appraisal of biomarkers in prostate cancer. World J Urol. 2019.
20. Yang M, Park JY. DNA methylation in promoter region as biomarkers in prostate cancer. Methods Mol Biol. 2012;863:67–109.
21. Jerónimo C, Bastian PJ, Bjartell A, Carbone GM, Catto JWF, Clark SJ, et al. Epigenetics in Prostate Cancer: Biologic and Clinical Relevance. Eur Urol. 2011;60:753–66.
22. Barbieri CE, Bangma CH, Bjartell A, Catto JWF, Culig Z, Grönberg H, et al. The Mutational Landscape of Prostate Cancer. Eur Urol. 2013;64:567–76.

23. Choudhury AD, Eeles R, Freedland SJ, Isaacs WB, Pomerantz MM, Schalken JA, et al. The Role of Genetic Markers in the Management of Prostate Cancer. Eur Urol. 2012;62:577–87.
24. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. 2012;41:D991-5.
25. Varambally S, Yu J, Laxman B, Rhodes DR, Mehra R, Tomlins SA, et al. Integrative genomic and proteomic analysis of prostate cancer reveals signatures of metastatic progression. Cancer Cell. 2005;8:393–406.
26. Wallace TA, Prueitt RL, Yi M, Howe TM, Gillespie JW, Yfantis HG, et al. Tumor Immunobiological Differences in Prostate Cancer between African-American and European-American Men. Cancer Res. 2008;68:927–36.
27. Derosa CA, Furusato B, Shaheduzzaman S, Srikantan V, Wang Z, Chen Y, et al. Elevated osteonectin/SPARC expression in primary prostate cancer predicts metastatic progression. Prostate Cancer Prostatic Dis. 2012;15:150–6.
28. Kuner R, Fälth M, Pressinotti NC, Brase JC, Puig SB, Metzger J, et al. The maternal embryonic leucine zipper kinase (MELK) is upregulated in high-grade prostate cancer. J Mol Med (Berl). 2013;91:237–48.
29. Mortensen MM, Høyer S, Lynnerup A, Ørntoft TF, Sørensen KD, Borre M, et al. Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy. Sci Rep-Uk. 2015;5.
30. Arredouani MS, Lu B, Bhasin M, Eljanne M, Yue W, Mosquera JM, et al. Identification of the Transcription Factor Single-Minded Homologue 2 as a Potential Biomarker and Immunotherapy Target in Prostate Cancer. Clin Cancer Res. 2009;15:5794–802.
31. Ashida S, Orloff MS, Bebek G, Zhang L, Zheng P, Peehl DM, et al. Integrated Analysis Reveals Critical Genomic Regions in Prostate Tumor Microenvironment Associated with Clinicopathologic Phenotypes. Clin Cancer Res. 2012;18:1578–87.
32. Meller S, Meyer HA, Bethan B, Dietrich D, Maldonado SG, Lein M, et al. Integration of tissue metabolomics, transcriptomics and immunohistochemistry reveals ERG- and gleason score-specific metabolomic alterations in prostate cancer. Oncotarget. 2016;7:1421–38.
33. Zhang L, Wang J, Wang Y, Zhang Y, Castro P, Shao L, et al. MNX1 Is Oncogenically Upregulated in African-American Prostate Cancer. Cancer Res. 2016;76:6290–8.
34. Ding Y, Wu H, Warden C, Steele L, Liu X, Iterson MV, et al. Gene Expression Differences in Prostate Cancers between Young and Old Men. Plos Genet. 2016;12:e1006477.
35. Kolde R, Laur S, Adler P, Vilo J. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics. 2012;28:573–80.
36. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43:e47.
37. Yu G, Wang L, Han Y, He Q. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS. 2012;16:284–7.

38. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. Bmc Bioinformatics. 2008;9:559.
39. Huang S, Yee C, Ching T, Yu H, Garmire LX. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. Plos Comput Biol. 2014;10:e1003851.
40. Meng T, Huang R, Zeng Z, Huang Z, Yin H, Jiao C, et al. Identification of Prognostic and Metastatic Alternative Splicing Signatures in Kidney Renal Clear Cell Carcinoma. Front Bioeng Biotech. 2019;7.
41. Lian H, Han YP, Zhang YC, Zhao Y, Yan S, Li QF, et al. Integrative analysis of gene expression and DNA methylation through one-class logistic regression machine learning identifies stemness features in medulloblastoma. Mol Oncol. 2019;13:2227–45.
42. Frank E. JH. Rms: Regression Modeling Strategies. R Package version 3.4-0.
43. SM SJCAL, SO WSM, DM H. M, et al. Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy. Ann Oncol. 2018;29:215–22.
44. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, et al. A pathology atlas of the human cancer transcriptome. Science. 2017;357:n2507.
45. Cai C, Wang H, He HH, Chen S, He L, Ma F, et al. ERG induces androgen receptor-mediated regulation of SOX9 in prostate cancer. J Clin Invest. 2013;123:1109–22.
46. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
47. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, et al. TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells. Cancer Res. 2017;77:e108-10.
48. Yip I, Aronson W, Heber D. Nutritional approaches to the prevention of prostate cancer progression. Adv Exp Med Biol. 1996;399:173–81.
49. Saunders LR, Verdin E. Sirtuins: critical regulators at the crossroads between cancer and aging. Oncogene. 2007;26:5489–504.
50. Tseng CH. Diabetes and Risk of Prostate Cancer: A study using the National Health Insurance. Diabetes Care. 2011;34:616–21.
51. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database. 2016;2016:w100.
52. Fujita R, Otake K, Arimura Y, Horikoshi N, Miya Y, Shiga T, et al. Stable complex formation of CENP-B with the CENP-A nucleosome. Nucleic Acids Res. 2015;43:4909–22.
53. Blower MD, Karpen GH. The role of Drosophila CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. Nat Cell Biol. 2001;3:730–9.
54. Rajput AB, Hu N, Varma S, Chen C, Ding K, Park PC, et al. Immunohistochemical Assessment of Expression of Centromere Protein—A (CENPA) in Human Invasive Breast Cancer. Cancers.

2011;3:4212–27.

55. Amato A, Schillaci T, Lentini L, Di Leonardo A. CENPA overexpression promotes genome instability in pRb-depleted human cells. Mol Cancer. 2009;8:119.
56. Jing K, Mao Q, Ma P. Decreased expression of TROAP suppresses cellular proliferation, migration and invasion in gastric cancer. Mol Med Rep. 2018;18:3020–6.
57. Ye X, Lv H. MicroRNA-519d-3p inhibits cell proliferation and migration by targeting TROAP in colorectal cancer. Biomed Pharmacother. 2018;105:879–86.
58. Hu H, Xu L, Chen Y, Luo S, Wu Y, Xu S, et al. The Upregulation of Trophinin-Associated Protein (TROAP) Predicts a Poor Prognosis in Hepatocellular Carcinoma. J Cancer. 2019;10:957–67.
59. Li K, Zhang R, Wei M, Zhao L, Wang Y, Feng X, et al. TROAP Promotes Breast Cancer Proliferation and Metastasis. Biomed Res Int. 2019;2019:1–8.
60. Ye J, Chu C, Chen M, Shi Z, Gan S, Qu F, et al. TROAP regulates prostate cancer progression via the WNT3/survivin signalling pathways. Oncol Rep. 2019;41:1169–79.
61. Lin TY, Chan HH, Chen SH, Sarvagalla S, Chen PS, Coumar MS, et al. BIRC5/Survivin is a novel ATG12-ATG5 conjugate interactor and an autophagy-induced DNA damage suppressor in human cancer and mouse embryonic fibroblast cells. Autophagy. 2019:1–18.
62. Zhang H, Li W, Gu W, Yan Y, Yao X, Zheng J. MALAT1 accelerates the development and progression of renal cell carcinoma by decreasing the expression of miR-203 and promoting the expression ofBIRC5. Cell Proliferat. 2019;52.
63. Kolberg M, Høland M, Lind GE, Ågesen TH, Skotheim RI, Sundby Hall K, et al. Protein expression of BIRC5, TK1, and TOP2A in malignant peripheral nerve sheath tumours - A prognostic test after surgical resection. Mol Oncol. 2015;9:1129–39.
64. Cao Y, Zhu W, Chen W, Wu J, Hou G, Li Y. Prognostic Value of BIRC5 in Lung Adenocarcinoma Lacking EGFR, KRAS, and ALK Mutations by Integrated Bioinformatics Analysis. Dis Markers. 2019;2019:1–12.
65. Bai Y, Xiong L, Zhu M, Yang Z, Zhao J, Tang H. Co-expression network analysis identified KIF2C in association with progression and prognosis in lung adenocarcinoma. Cancer Biomark. 2019;24:371–82.
66. Gan H, Lin L, Hu N, Yang Y, Gao Y, Pei Y, et al. KIF2C exerts an oncogenic role in nonsmall cell lung cancer and is negatively regulated by miR-325-3p. Cell Biochem Funct. 2019;37:424–31.
67. Bendre S, Rondelet A, Hall C, Schmidt N, Lin Y, Brouhard GJ, et al. GTSE1 tunes microtubule stability for chromosome alignment and segregation by inhibiting the microtubule depolymerase MCAK. The Journal of Cell Biology. 2016;215:631–47.
68. Shimo A, Tanikawa C, Nishidate T, Lin M, Matsuda K, Park J, et al. Involvement of kinesin family member 2C/mitotic centromere-associated kinesin overexpression in mammary carcinogenesis. Cancer Sci. 2007:1426707130.

69. Ishikawa K, Kamohara Y, Tanaka F, Haraguchi N, Mimori K, Inoue H, et al. Mitotic centromere-associated kinesin is a novel marker for prognosis and lymph node metastasis in colorectal cancer. Brit J Cancer. 2008;98:1824–9.
70. Wang C, Xiang F, Li Y, Xing X, Wang N, Chi J, et al. Relation between the expression of mitotic centromere–associated kinesin and the progression of squamous cell carcinoma of the tongue. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology. 2014;117:353 – 60.
71. Weilbaecher KN, Guise TA, McCauley LK. Cancer to bone: a fatal attraction. Nat Rev Cancer. 2011;11:411–25.
72. Tyekucheva S, Bowden M, Bango C, Giunchi F, Huang Y, Zhou C, et al. Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. Nat Commun. 2017;8.
73. Logothetis C, Morris MJ, Den R, Coleman RE. Current perspectives on bone metastases in castrate-resistant prostate cancer. Cancer Metast Rev. 2018;37:189–96.
74. Roato I, Vitale M. The Uncovered Role of Immune Cells and NK Cells in the Regulation of Bone Metastasis. Front Endocrinol. 2019;10.
75. Piersma SJ, Jordanova ES, van Poelgeest MIE, Kwappenberg KMC, van der Hulst JM, Drijfhout JW, et al. High Number of Intraepithelial CD8 ± Tumor-Infiltrating Lymphocytes Is Associated with the Absence of Lymph Node Metastases in Patients with Large Early-Stage Cervical Cancer. Cancer Res. 2007;67:354–61.
76. Kmiecik J, Poli A, Brons NHC, Waha A, Eide GE, Enger P, et al. Elevated CD3 + and CD8 + tumor-infiltrating immune cells correlate with prolonged survival in glioblastoma patients despite integrated immunosuppressive mechanisms in the tumor microenvironment and at the systemic level. J Neuroimmunol. 2013;264:71–83.
77. Kim ST, Jeong H, Woo OH, Seo JH, Kim A, Lee ES, et al. Tumor-infiltrating Lymphocytes, Tumor Characteristics, and Recurrence in Patients With Early Breast Cancer. Am J Clin Oncol. 2013;36:224–31.

Tables

TABLE 1 Characteristics of the GEO datasets.

GEOset ID	Contributors	Platform ID	Samples	Number of rows per platform
GSE3325	Varambally S, et al	GPL570	6N 13T	54675
GSE6956	Wallace TA, et al	GPL1571	20N 69T	22277
GSE32448	Derosa CA, et al	GPL570	40N 40T	54675
GSE32571	Kuner R, et al	GPL6947	39N 59T	48652
GSE46602	Mortensen MM, et al	GPL570	14N 36T	54675
GSE55945	Arredouani MS, et al	GPL570	8N 13T	54675
GSE34312	Ashida S, et al	GPL6884	10N 10T	48803
GSE69223	Meller S, et al	GPL570	15N 15T	54675
GSE71016	Zhang L, et al	GPL16699	47N 48T	62976
GSE88808	Ding Y, et al	GPL22571	49N 49T	20260

Abbreviations: GEO: Gene Expression Omnibus; T: tumor samples; N: paracancerous normal samples.

TABLE 2 Baseline characteristics of patients in TCGA cohorts

Clinical Traits	Risk Score		c^2	*P* value
	High Risk n (%)	Low Risk n (%)		
Age			7.689	0.056
≤50	141(43.65%)	182(56.35%)		
>50	393(52.89%)	350(47.11%)		
T			10.74	0.0132
T1	124(44.44%)	155(55.56%)		
T2	308(49.92%)	309(50.08%)		
T3	76(56.30%)	59(43.70%)		
T4	26(68.42%)	12(31.58%)		
N			11.85	0.0079
N0	226(44.66%)	280(55.34%)		
N1	183(53.04%)	162(46.96%)		
N2	86(58.11%)	62(51.89%)		
N3	39(55.71%)	31(44.29%)		
M			5.094	0.024
M0	498(48.54%)	528(51.46%)		
M1	29(65.91%)	15(34.09%)		
Satge			15.30	0.0016
I	74(40.88%)	107(59.12%)		
II	315(50.97%)	303(49.03%)		
III	127(51.21%)	121(48.79%)		
IV	18(81.82%)	4(18.18%)		
Cancer Status			30.35	<0.001
Tumor Free	457(95.61%)	21(4.39%)		
With Tumor	398(85.04%)	70(14.96%)		
New Tumor Event			5.237	0.0221
No	56(42.75%)	75(57.25%)		
Yes	501(53.41%)	437(46.59%)		

TABLE 3 Visualization of specificity, sensitivity, and cutoff values of 5 gene signatures as diagnostic and prognostic markers

Gene	AUC	Std. Error	95% CI	Cutoff (log2)	Sensitivity%	Specificity%
NF695	0.7868	0.03956	0.7093 ~ 0.8643	3.566	67.8	84.62
NEPA	0.8740	0.03153	0.8122 ~ 0.9358	4.413	81.53	86.54
ROAP	0.8923	0.02976	0.8340 ~ 0.9506	5.499	89.36	86.29
IRC5	0.8864	0.02865	0.8302 ~ 0.9425	6.930	85.4	84.62
EF20A	0.8506	0.03188	0.7881 ~ 0.9130	3.349	83.73	75.00

Abbreviations: **AUC: Area Under roc Curve; CI: Confidence Interval**

Figures

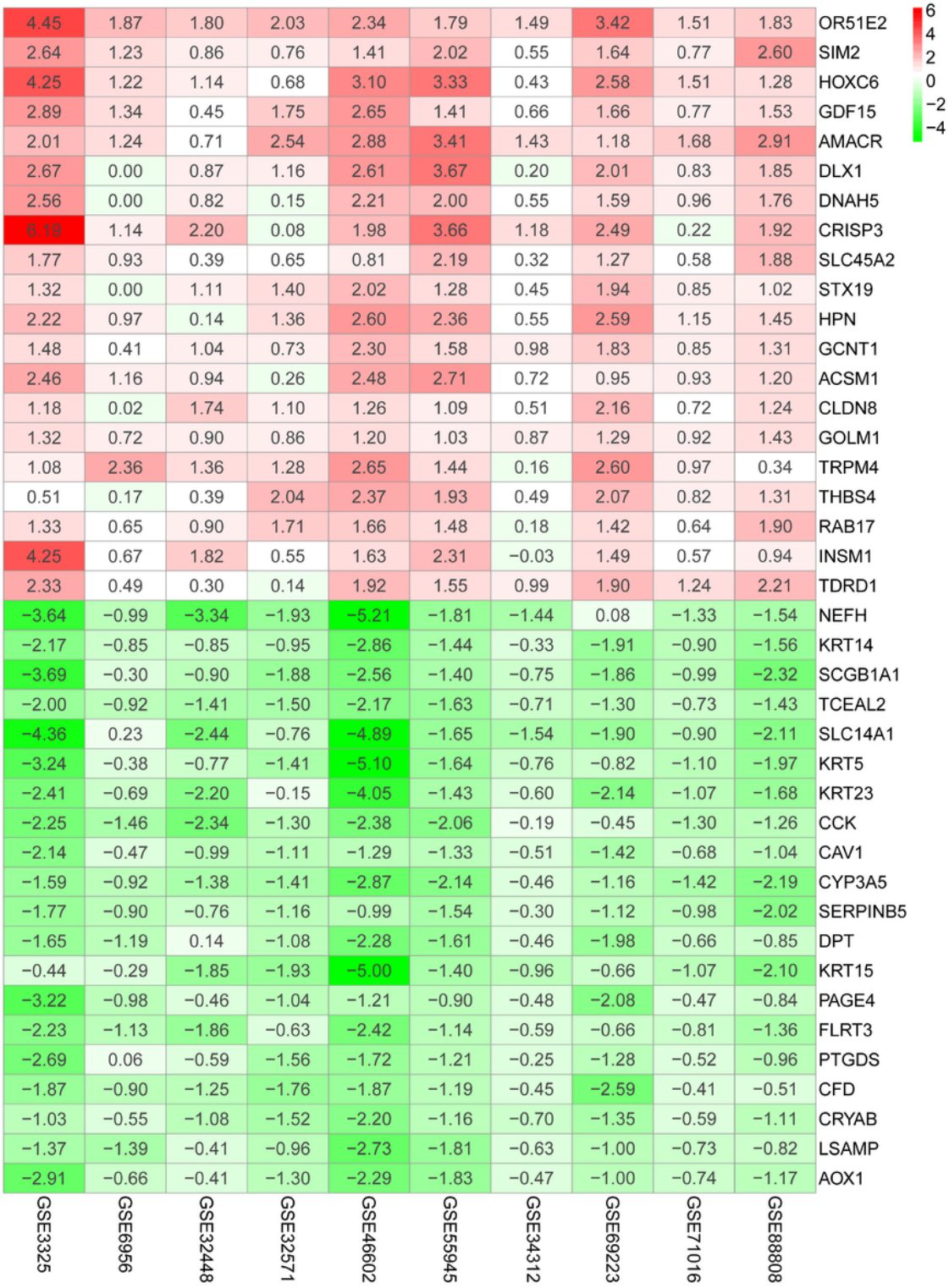

Figure 1

Heatmap shows top 20 DEGs in up-regulated and down-regulated genes based on RRA DEGs were defined with P-value < 0.05 and |logFC| ≥ 1.

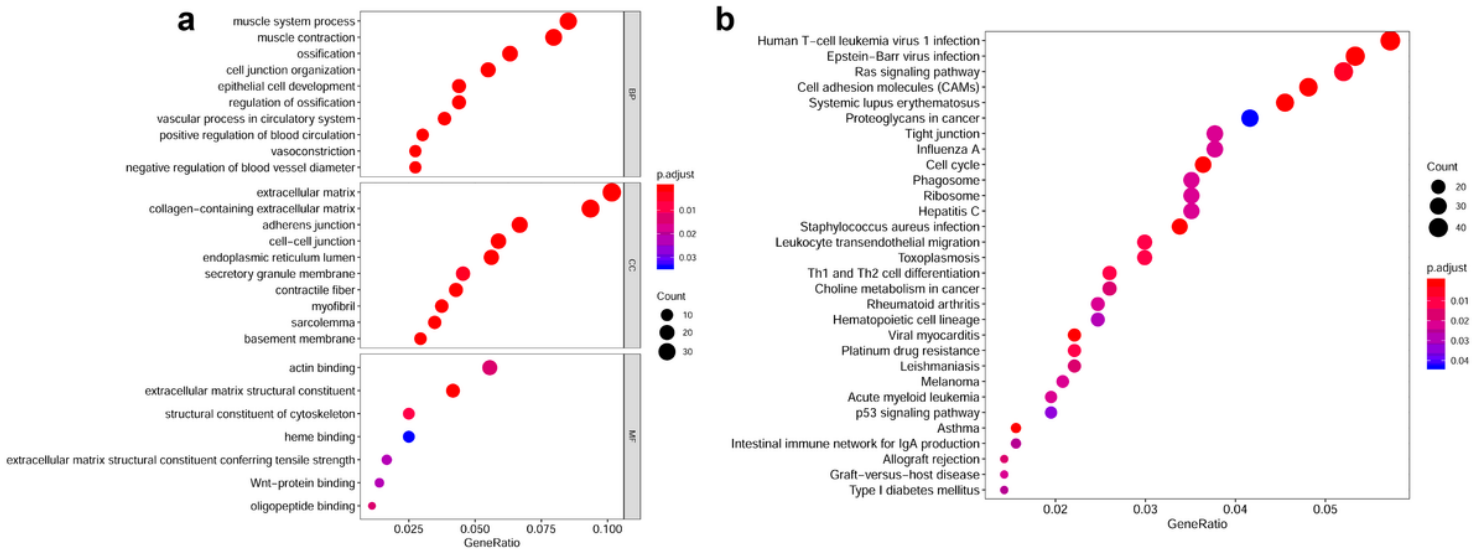

Figure 2

Gene function enrichment analyses of DEGs. a GO terms enrichment analysis of DEGs in MF, BP, and CC. b KEGG pathways enrichment analysis of DEGs.

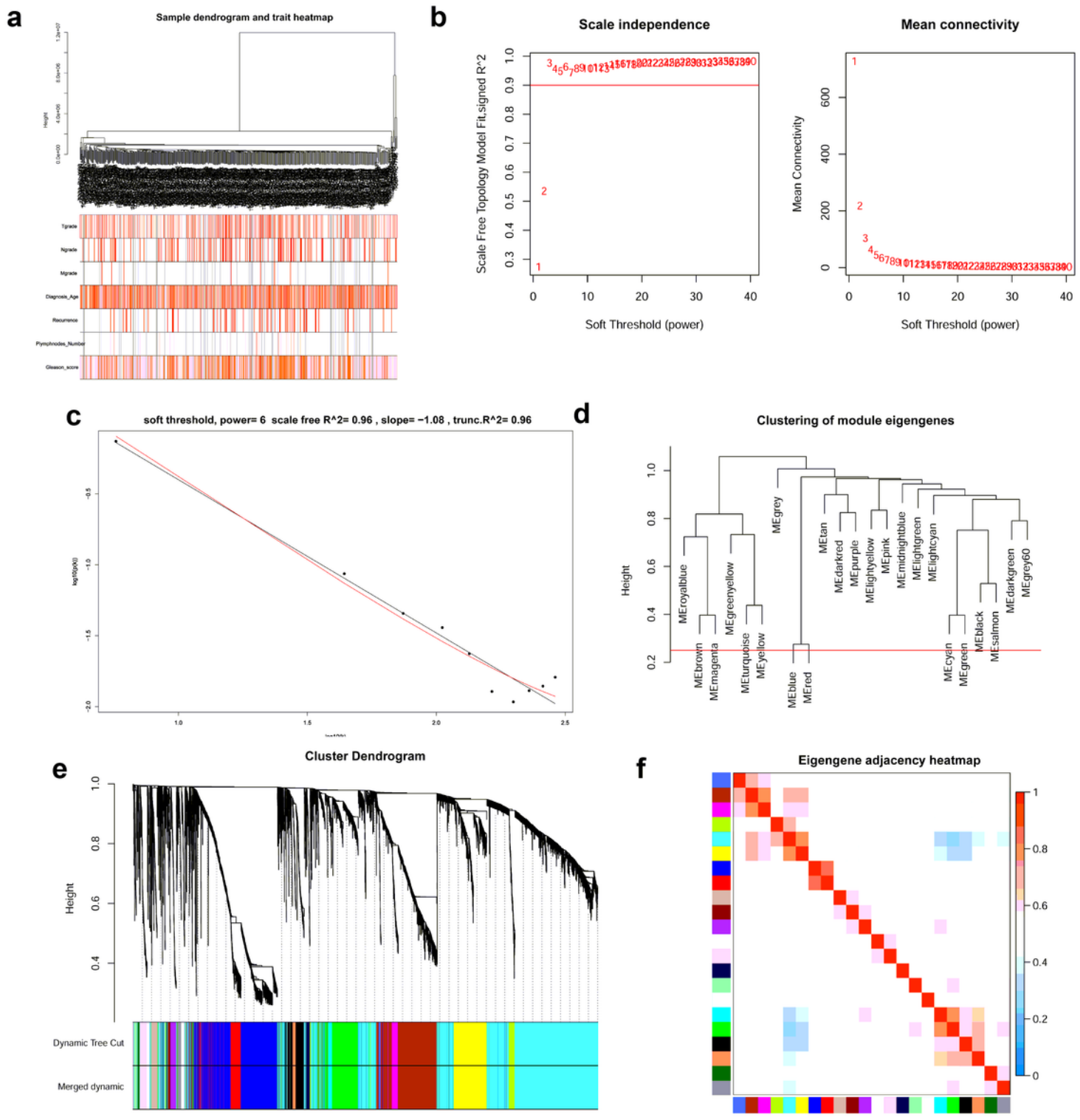

Figure 3

WGCNA analysis for DEGs. a Samples dendrogram and traits heatmap between TCGA-PRAD samples and clinical traits. b Analysis of the average connectivity and scale-free fit index by setting unequal soft-thresholding powers. c The scale-free R2 reached its maximum value when setting soft-thresholding power at 6. d Clustering of module eigengenes. The red line indicates cut height (0.20). e Dendrogram of

all DEGs clustered based on TOM. f Analysis of the relationship of co-expression modules based on the pearson correlation coefficient.

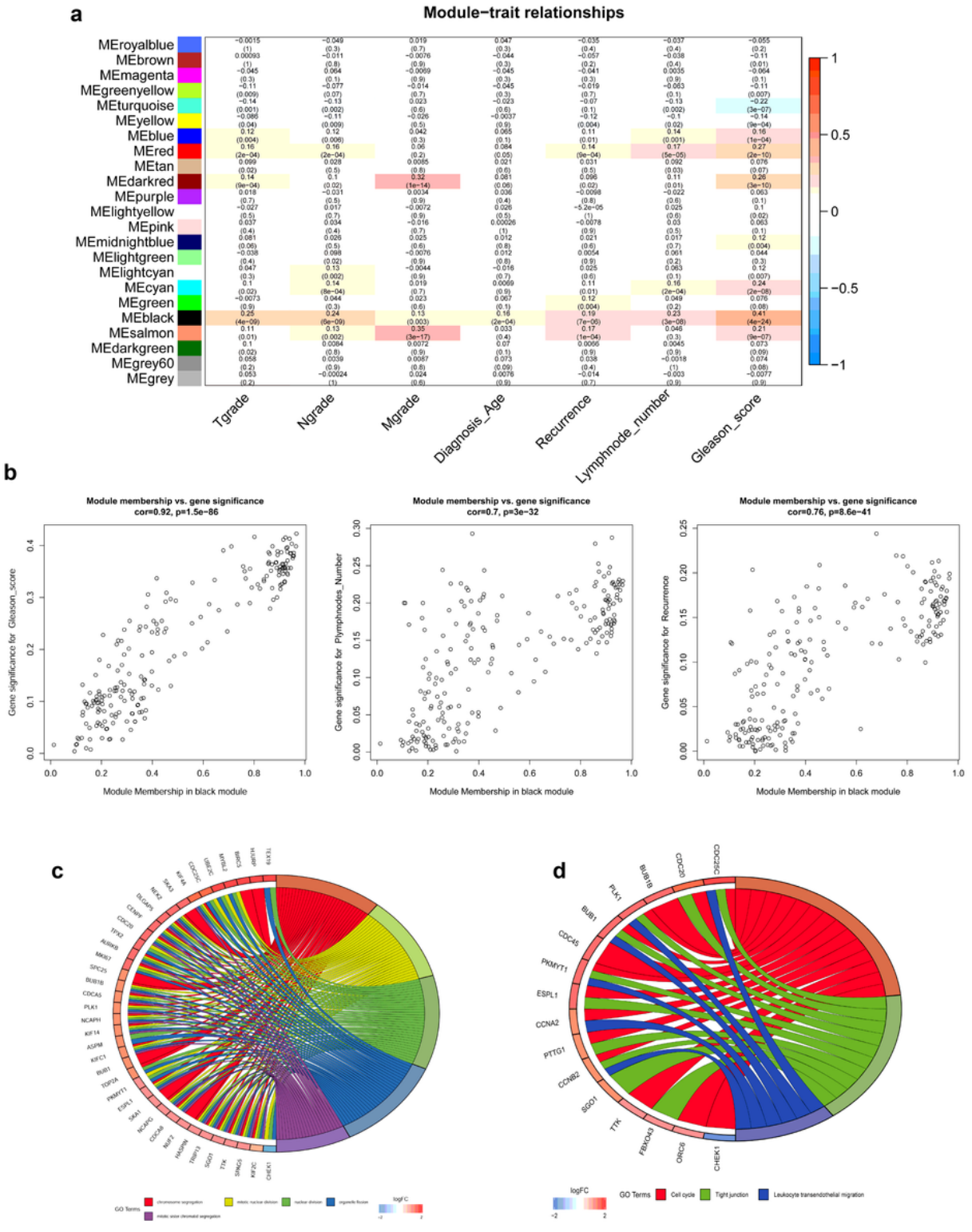

Figure 4

Identification of key genes and functional annotation of the black module. a The correlation between modules and the clinical traits. b Scatter plot of module eigengenes in the black module related with Plymphnodes Number (correlation coefficient = 0.7), Recurrence (correlation coefficient = 0.76), and

Gleason score (correlation coefficient = 0.92). c Chord plot depicted the relationship between genes and GO terms of molecular function. d Chord plot indicated the relationship between genes and KEGG pathways.

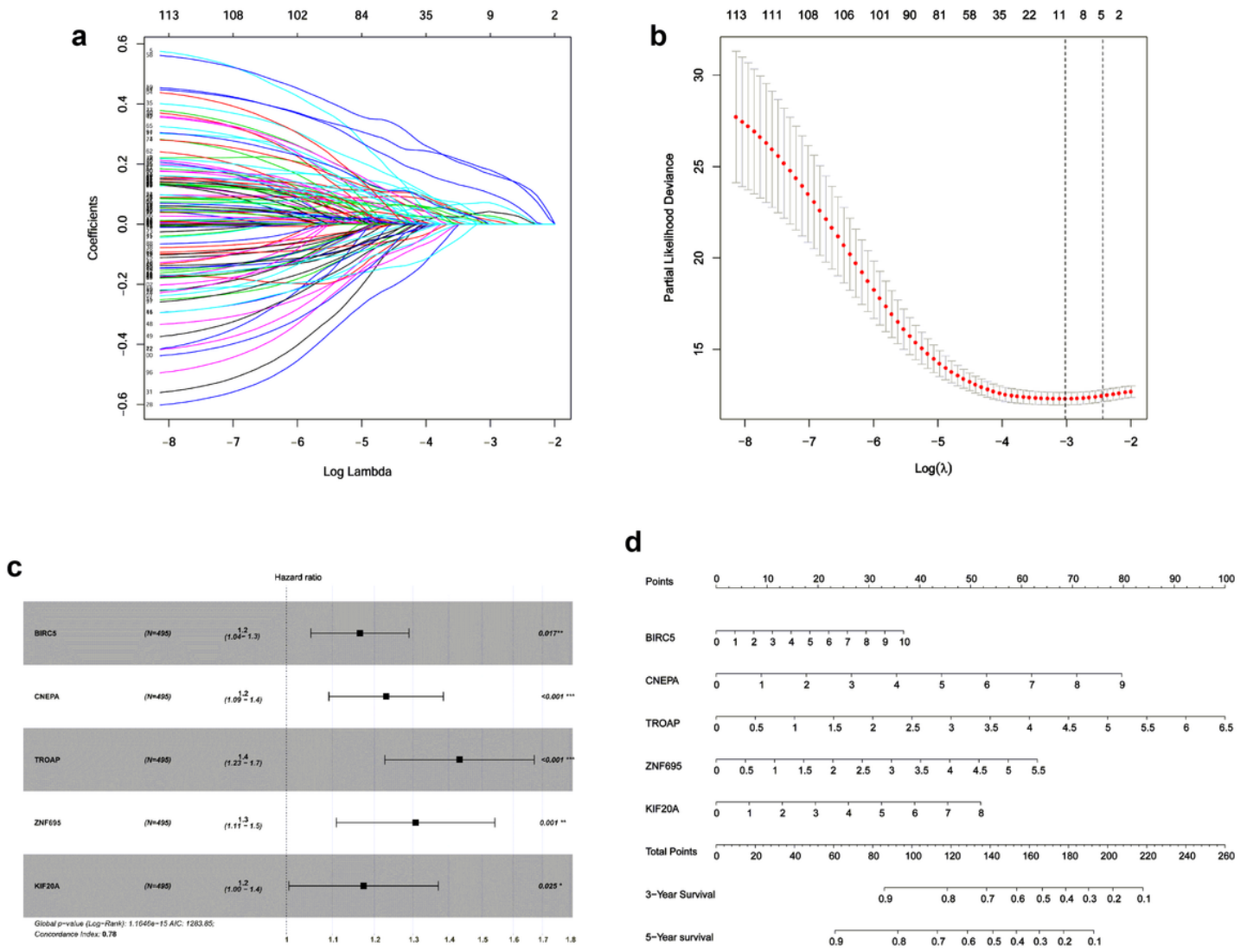

Figure 5

Identification of gene signatures associated with disease-free interval based on Cox-PH regression model. a Lasso coefficient profiles of the 495 progression- associated events in Pca. b Selection of the tuning parameter (λ) in the LASSO model through 10-fold cross-validation procedure was plotted as a function of log(λ). c-d Construction of multivariate Cox-PH regression model and ZNF695, CENPA, TROAP, BIRC5, KIF20A were considered significant and used to construct a prognostic model

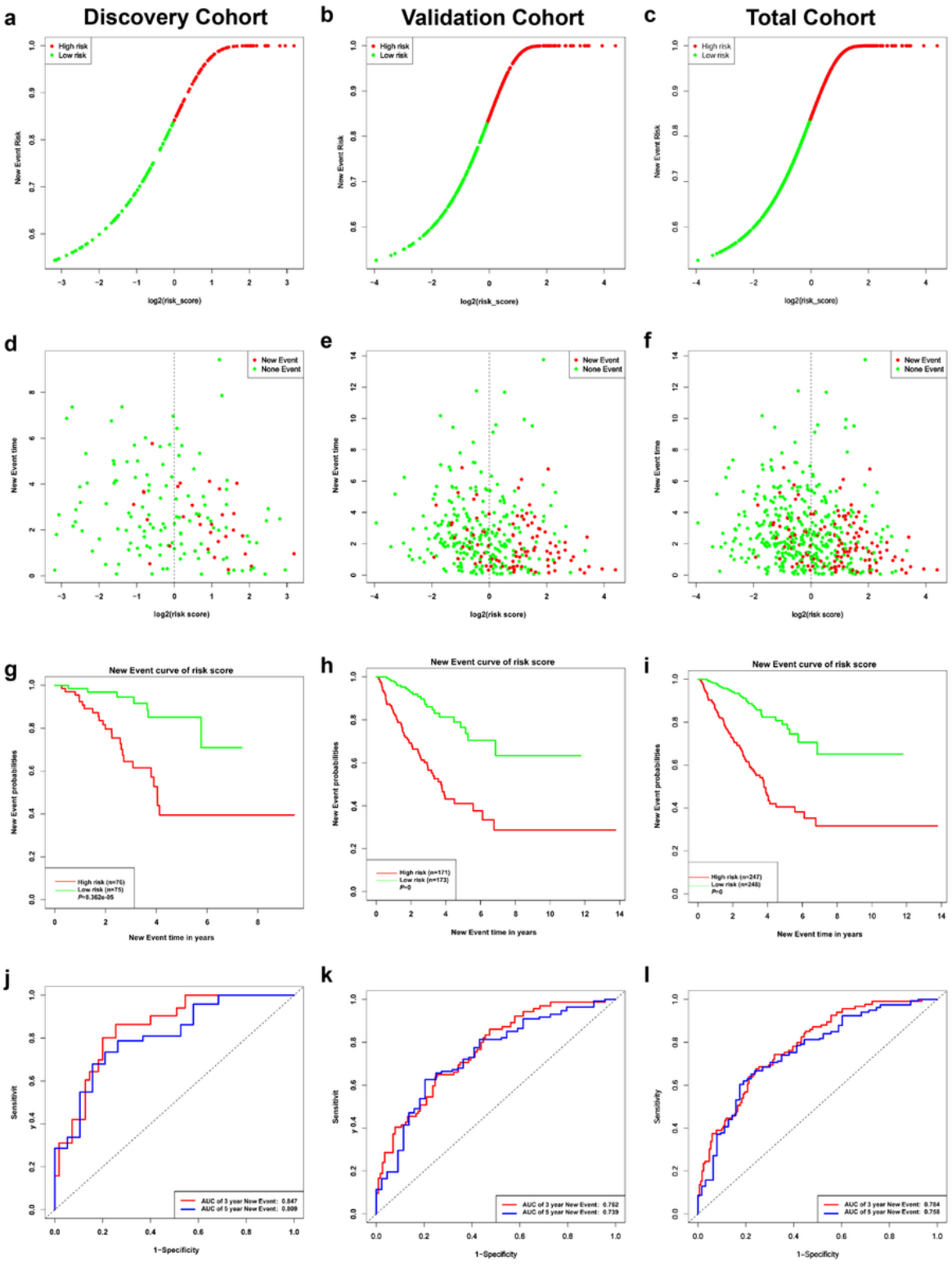

Figure 6

Risk curves (a-c) and scatter plots (d-f) inplied the risk score and new event risk for each PCa patient. g-i KM survival curves revealed that the prediction model of risk score had good discrimination and patients with high-risk scores have shorter disease-free interval. j-l According to the prognostic model, the ROC curve has a higher efficiency in predicting 3 or 5 years of DFI in PCa patients

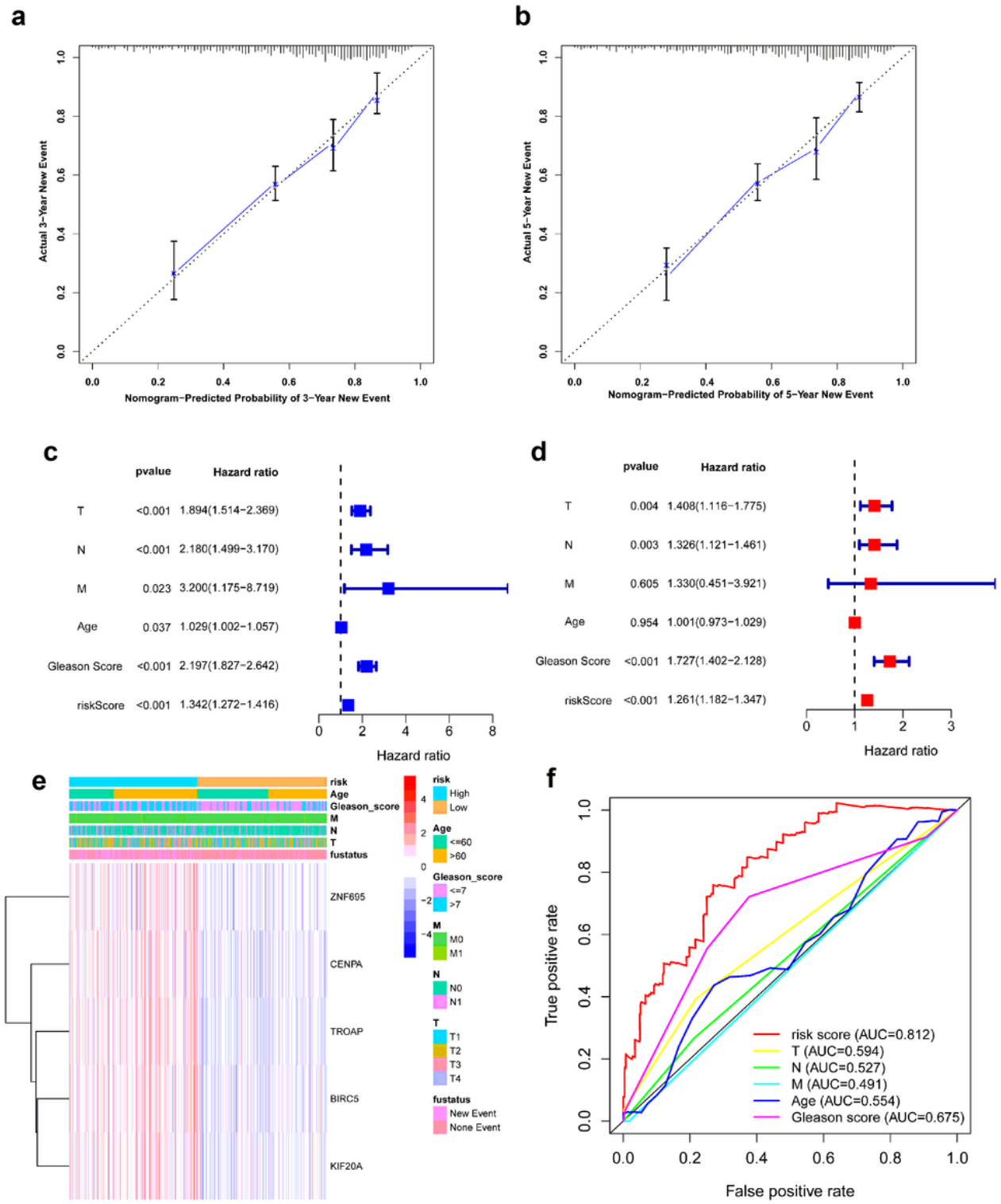

Figure 7

a-b Calibration curves of the nomogram for predicting the probability of DFI at 3 and 5 year. c-d The univariate and multivariate Cox regression analysis of risk score, age, Gleason score, and TNM stage. e Distribution characteristics of expression profiles of 5 gene signatures in different risk and clinicopathological groups. f Multiline ROC curves showed the superiority of 5-gene panel based on a 10-fold cross-validation, for predicting tumor new event.

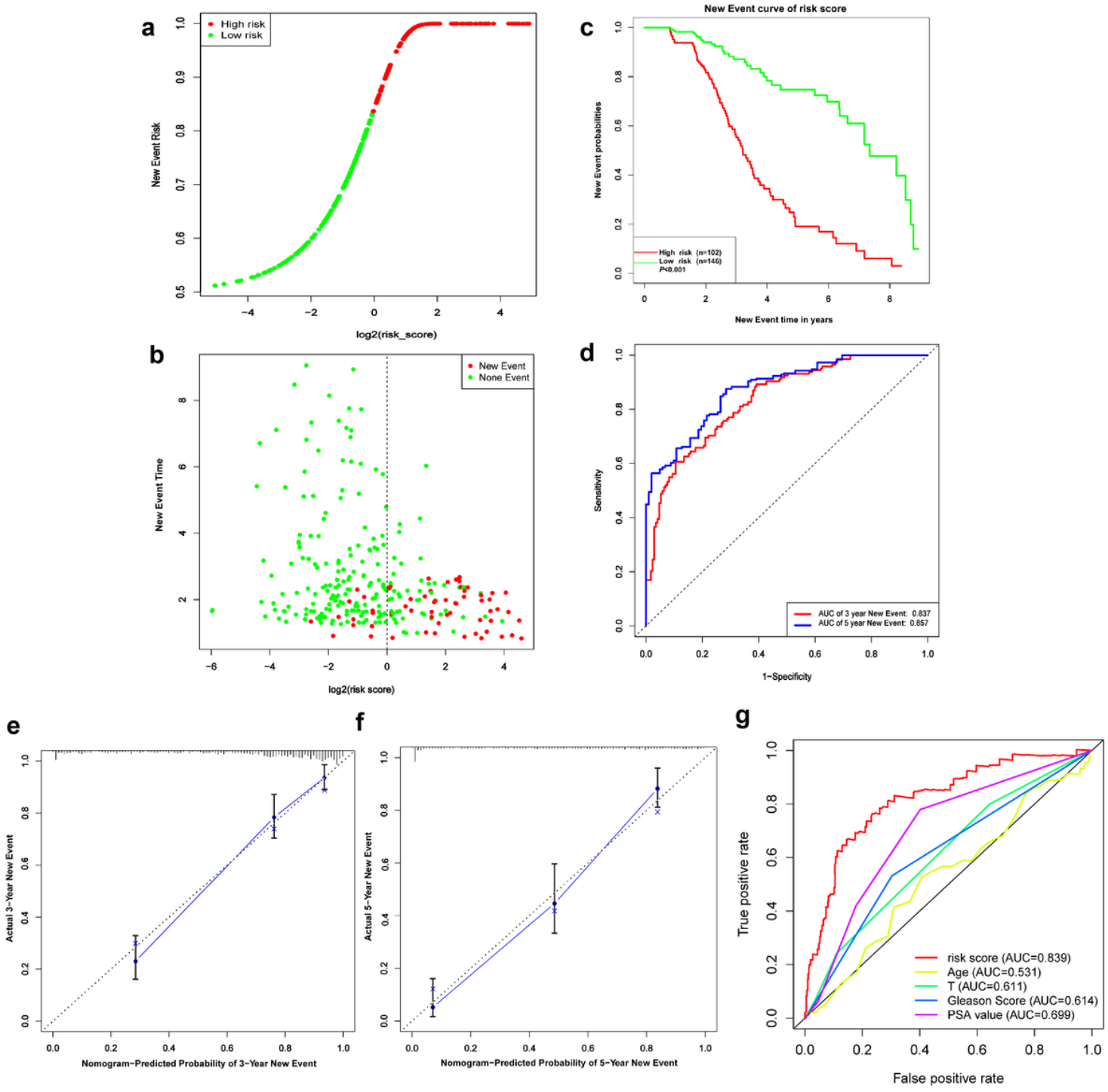

Figure 8

a-e KM survival curves shows that T classification (P<0.001), N classification (P<0.001), M classification (P=0.016) and Gleason score (P<0.001) are independent risk factors, except for age (P=0.167)

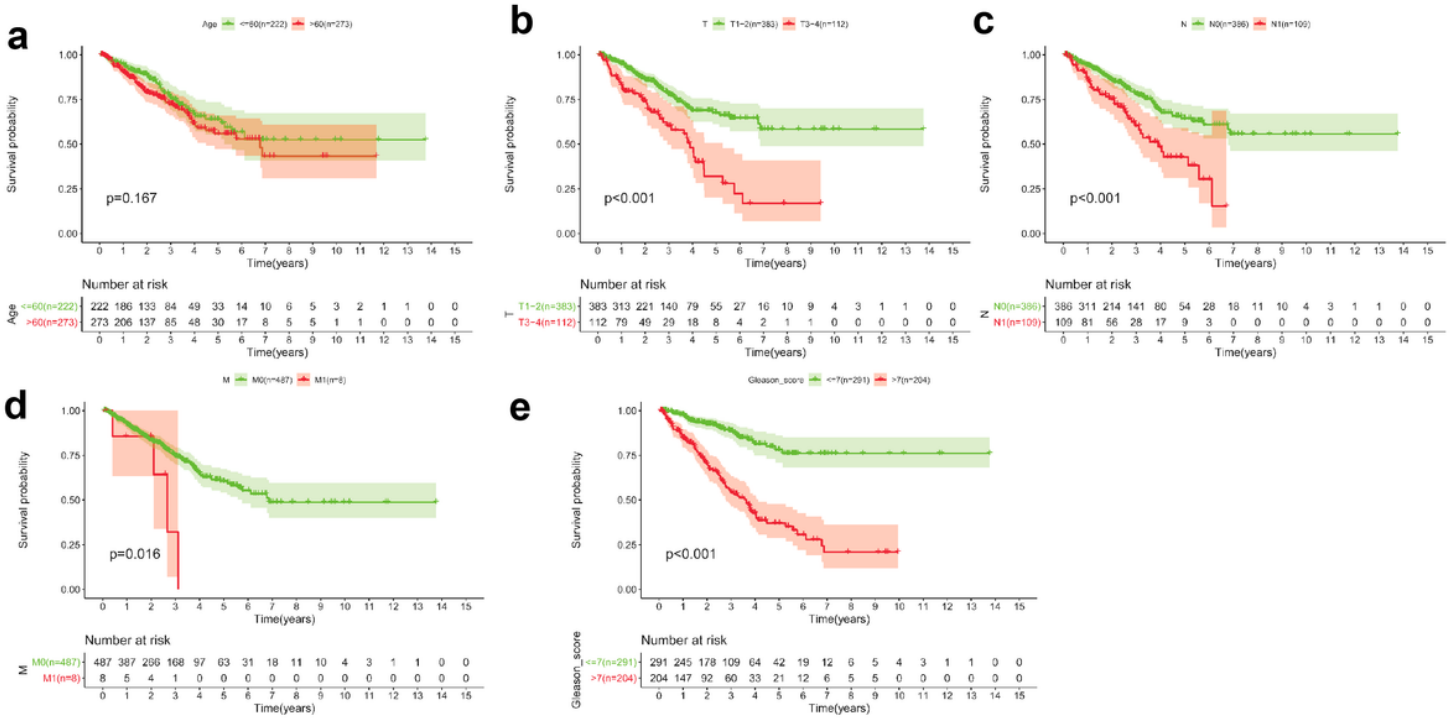

Figure 9

a-d Risk score, KM survival curves and time-dependent ROC curves of DFI in GSE116918 validation cohort. (e-f) Calibration curves of the nomogram for predicting the probability of DFI at 3 and 5 year (g) Multiline ROC curves showed the superiority of 5-gene panel than Age, T classification, Gleason score, and PSA value

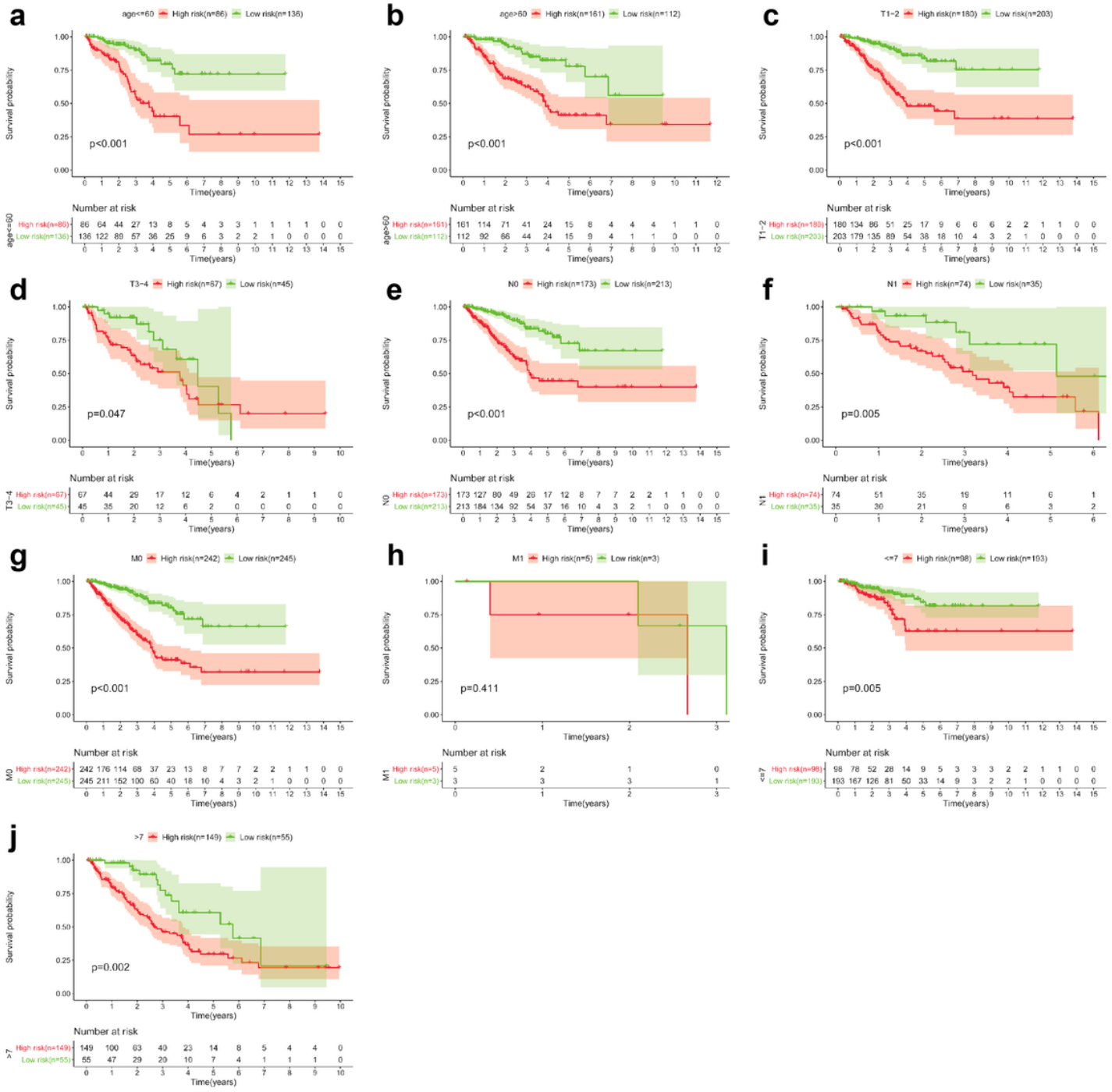

Figure 10

KM survival curves for the high and low risk groups stratified by clinicopathological variables. Age (a, b), T classification (c, d), N classification (e, f), M classification (g, h), and Gleason score (i, j)

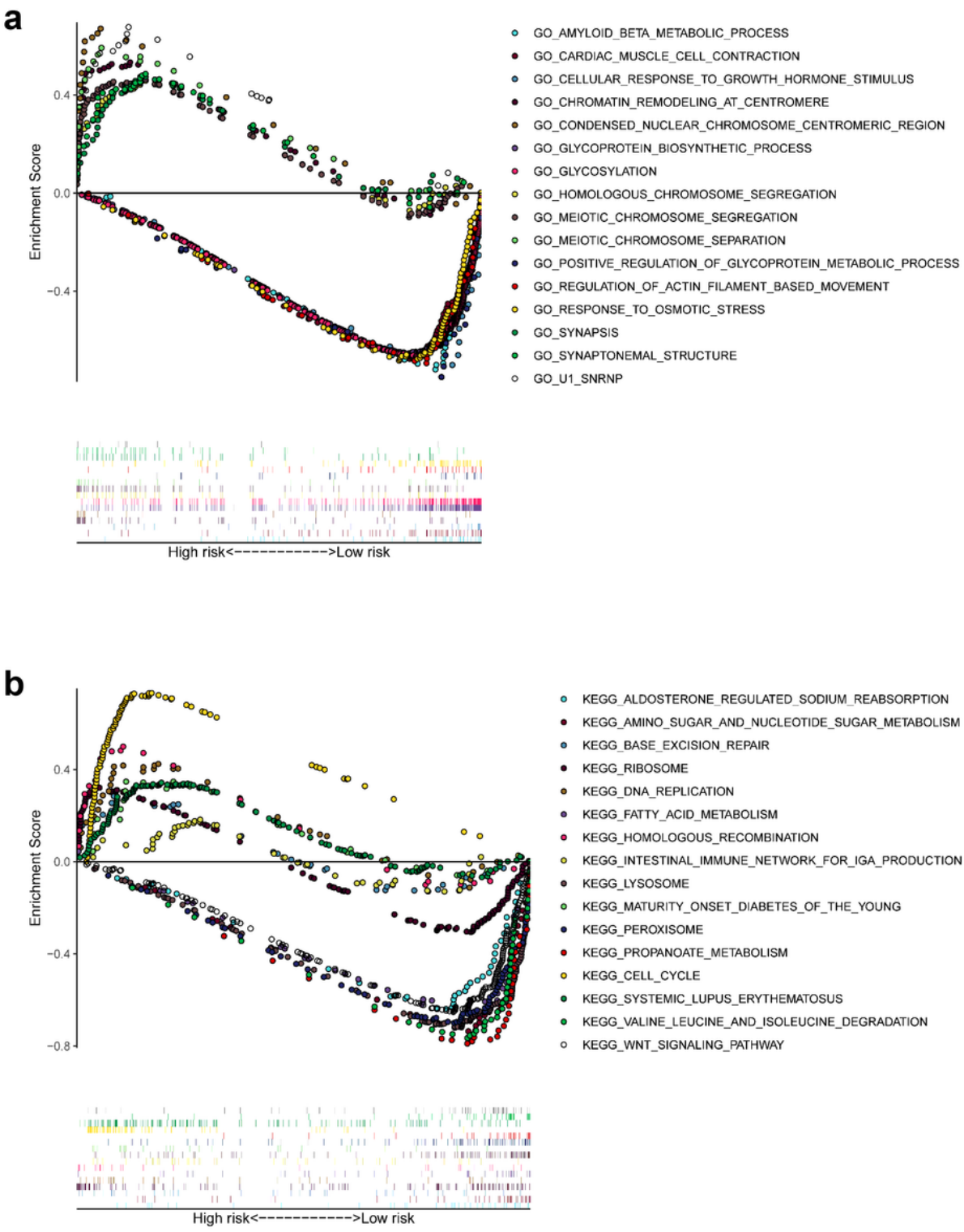

Figure 11

GSEA delineates biological pathways and processes between high and low risk using gene sets of Gene Oncology (a) and Kyoto Encyclopedia of Genes and Genomes (b). Each run was performed with 1000 permutations.

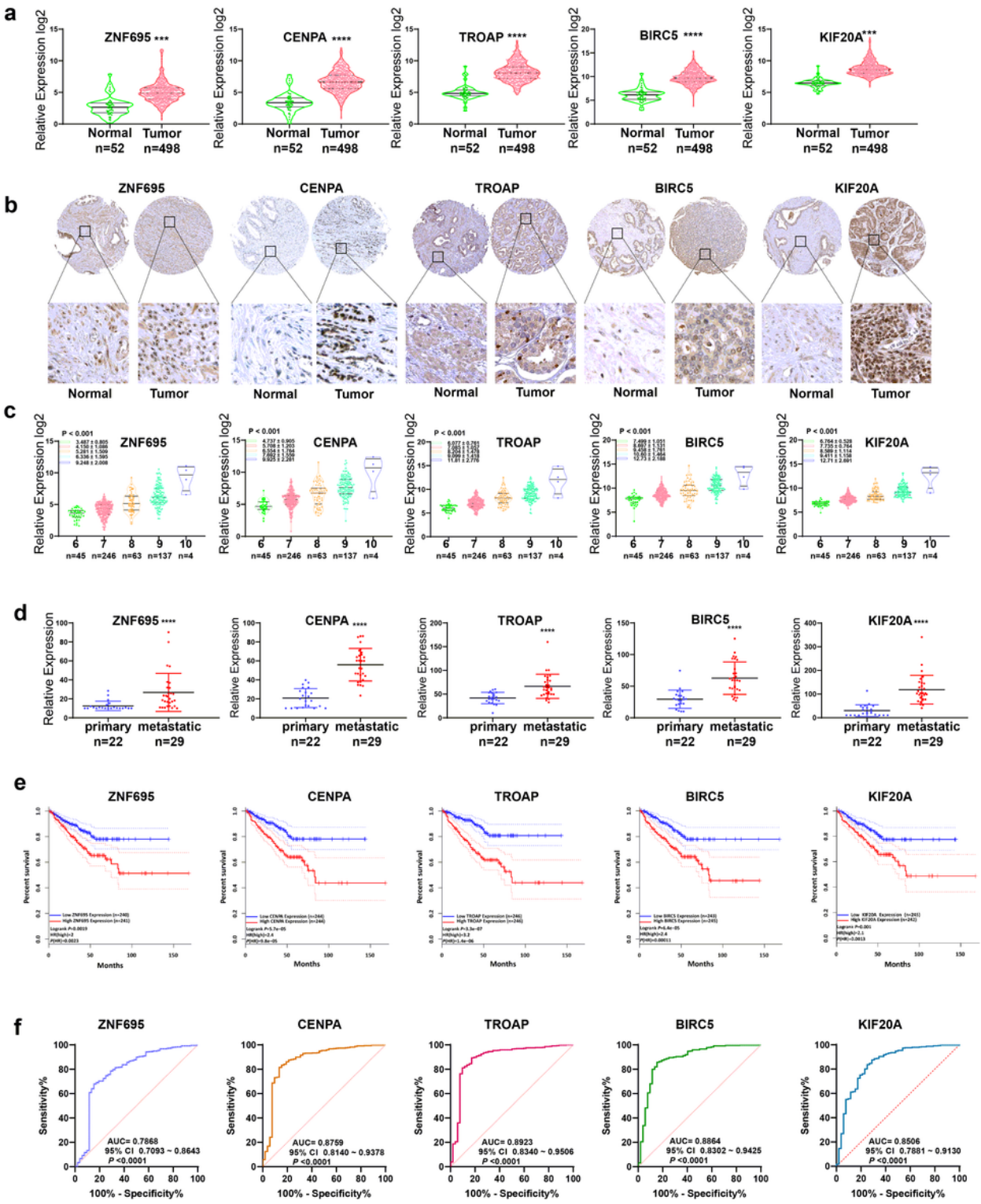

Figure 12

External validation of 5 gene signatures in Pca. a, b Differentially expression of these 5 gene signatures in PCa tissue and paracancerous normal samples at mRNA and protein level. c The expression levels of these 5 gene signatures increased with increasing Gleason score. d The expression levels of these 5 gene signatures in primary and metastatic tumors. e Kaplan-Meier survival curves shown that higher

expression of these genes was significantly associated with poor PFS. f ROC curve implied higher diagnostic efficiency of these 5 gene signatures.

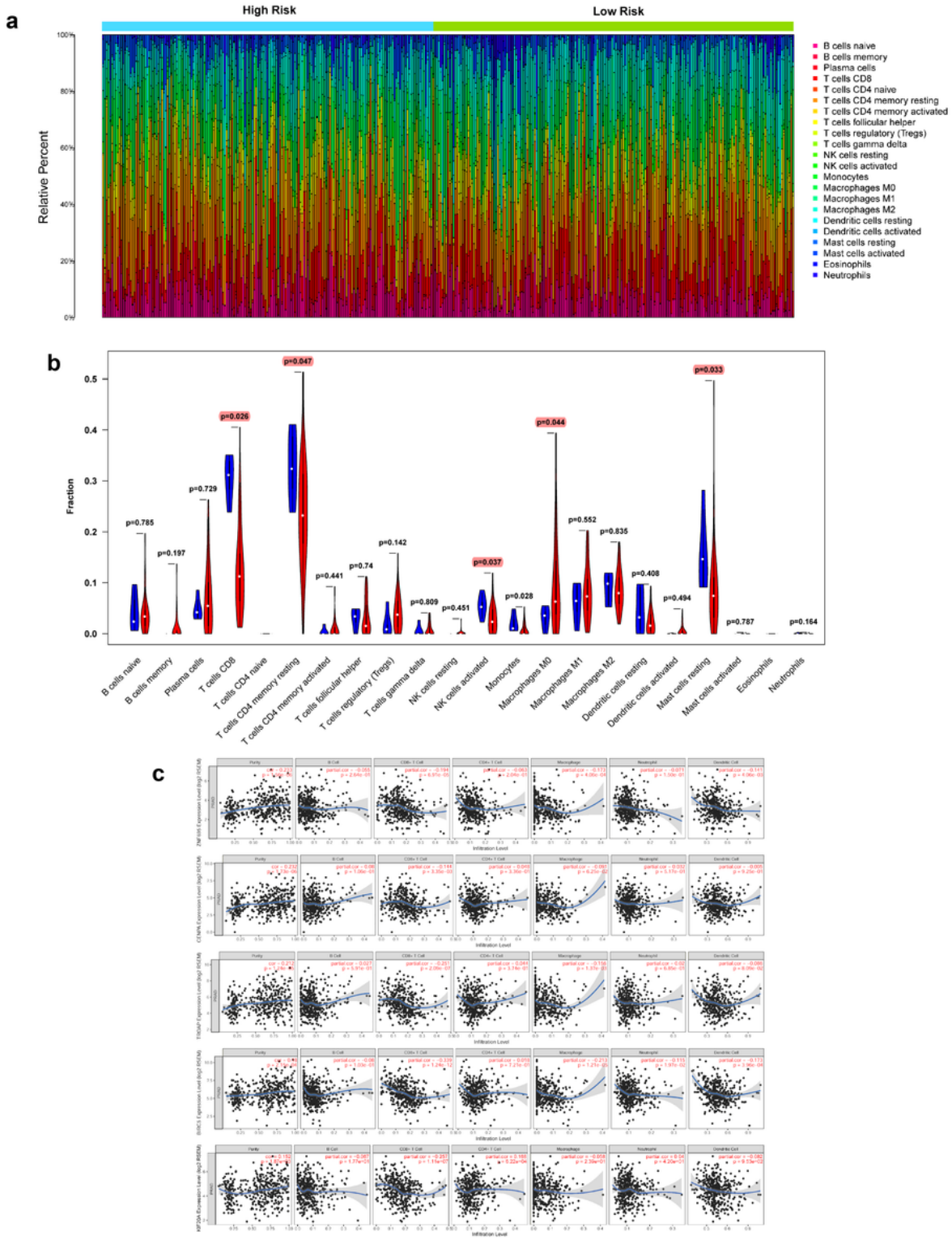

Figure 13

Evaluation of relationships between gene signatures and the immune microenvironment. a Assessment of the proportion of 22 immune cells calculated by the CIBERSORT algorithm in PCa tissues based on normalized expression in TCGA-PRAD. b Violin plot depicted the infiltration of significantly different

subsetst (T cells CD8, T cells CD4 memory resting, NK T cells activated, Macrophages M0, and Mast cells resting) between the high and low risk. c TIMER algorithm implied these 5 gene signatures were positively correlated with tumor purity and negatively correlated with CD8+ T cells.

Supplementary Files

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

- FigureS4.tif
- FigureS3.tif
- FigureS2.tif
- FigureS1.tif