EBioMedicine 32 (2018) 234–244

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.ebiomedicine.com

Research Paper

RankProd Combined with Genetic Algorithm Optimized Artificial Neural Network Establishes a Diagnostic and Prognostic Prediction Model that Revealed *C1QTNF3* as a Biomarker for Prostate Cancer

Qi Hou [a,b,1], Zhi-Tong Bing [c,d,2], Cheng Hu [e], Mao-Yin Li [e], Ke-Hu Yang [c,d], Zu Mo [f], Xiang-Wei Xie [f], Ji-Lin Liao [f], Yan Lu [b], Shigeo Horie [b], Ming-Wu Lou [a,*]

[a] *Post-Doctoral Research Center, Longgang Central Hospital, Shenzhen Clinical Medical Institute, Guangzhou University of Chinese Medicine, Shenzhen 518116, China*
[b] *Department of Urology, Juntendo University Graduate School of Medicine, Tokyo 1138421, Japan*
[c] *Evidence Based Medicine Center, School of Basic Medical Science, Lanzhou University, Lanzhou 730000, China*
[d] *Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou 730000, China*
[e] *Department of Urology, Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China*
[f] *Department of Urology, Longgang Central Hospital, Shenzhen Clinical Medical Institute, Guangzhou University of Chinese Medicine, Shenzhen 518116, China*

ARTICLE INFO

Article history:
Received 14 March 2018
Received in revised form 8 May 2018
Accepted 8 May 2018
Available online 1 June 2018

Keywords:
RankProd
Artificial neural network
Genetic algorithm
Prostate cancer
Biomarker

ABSTRACT

Prostate cancer (PCa) is the most commonly diagnosed cancer in males in the Western world. Although prostate-specific antigen (PSA) has been widely used as a biomarker for PCa diagnosis, its results can be controversial. Therefore, new biomarkers are needed to enhance the clinical management of PCa. From publicly available microarray data, differentially expressed genes (DEGs) were identified by meta-analysis with RankProd. Genetic algorithm optimized artificial neural network (GA-ANN) was introduced to establish a diagnostic prediction model and to filter candidate genes. The diagnostic and prognostic capability of the prediction model and candidate genes were investigated in both GEO and TCGA datasets. Candidate genes were further validated by qPCR, Western Blot and Tissue microarray. By RankProd meta-analyses, 2306 significantly up- and 1311 down-regulated probes were found in 133 cases and 30 controls microarray data. The overall accuracy rate of the PCa diagnostic prediction model, consisting of a 15-gene signature, reached up to 100% in both the training and test dataset. The prediction model also showed good results for the diagnosis (AUC = 0.953) and prognosis (AUC of 5 years overall survival time = 0.808) of PCa in the TCGA database. The expression levels of three genes, *FABP5*, *C1QTNF3* and *LPHN3*, were validated by qPCR. *C1QTNF3* high expression was further validated in PCa tissue by Western Blot and Tissue microarray. In the GEO datasets, *C1QTNF3* was a good predictor for the diagnosis of PCa (GSE6956: AUC = 0.791; GSE8218: AUC = 0.868; GSE26910: AUC = 0.972). In the TCGA database, *C1QTNF3* was significantly associated with PCa patient recurrence free survival ($P < .001$, AUC = 0.57). In this study, we have developed a diagnostic and prognostic prediction model for PCa. *C1QTNF3* was revealed as a promising biomarker for PCa. This approach can be applied to other high-throughput data from different platforms for the discovery of oncogenes or biomarkers in different kinds of diseases.

1. Introduction

Prostate cancer (PCa) is the most commonly diagnosed cancer in males and one of the leading causes of cancer mortality in the Western world. An estimated 164,690 Americans will be diagnosed with prostate cancer and 29,430 will die of the disease in the United States in 2018 [1]. In recent years the number of diagnosed prostate cancer patients also increased rapidly in developing countries such as China [2]. Here it is one of the ten most common cancers diagnosed in men, with an estimated 60,300 new cases and 26,600 deaths in 2015 [3]. The 5-year relative survival rate of localized prostate cancer patients approaches 100% but sharply decreases to 28% for patients diagnosed at an advanced stage [4]. Therefore, early detection and precise diagnosis for prostate cancer needs to advance further.

At the moment, prostate specific antigen (PSA) testing is widely used for PCa diagnosis at an early organ-confined stage. However, it occasionally leads to unnecessary biopsies due to its poor specificity [5–7]. Therefore, other new biomarkers with high accuracy and specificity are needed to improve diagnosis and prognosis of PCa. Cima et al. employed a proteome method and identified a five-protein signature (GALNTL4,

* Corresponding author.
E-mail address: mingwulou@sina.com (M.-W. Lou).
[1] Equal contributors
[2] Equal contributors

https://doi.org/10.1016/j.ebiom.2018.05.010

FN, AZGP1, GBA and ECM1) in PCa which could be used to improve screening efficacy [8]. Ankerst et al. proposed that detection of PCA3, a PCa-specific gene, combined with PSA testing could improve diagnostic accuracy [9]. In addition, hypermethylation of some critical genes or microRNAs like GSTP1, PITX2, GABRE~miR-452~miR-224, and a methylated site (cg05163709) in Chromosome Y have been proposed as promising biomarkers of PCa [10,11]. Furthermore, a number of commercially available products show potential, but there is still some way to go before there is enough data to convincingly demonstrate the added value of these methods.

High-throughput microarray chip and second-generation sequencing technologies are powerful tools for discovering and studying novel biomarkers for PCa. However, analyses based on high throughput data may encounter the "curse of dimensionality" [12]. This refers to the phenomena that the amount of dependent variables increases greatly while the amount of samples is relative small, resulting in an increase of statistical errors. Fortunately, increasing the sample size and using some machine learning algorithm can effectively improve the problems caused by this "curse" [12,13].

The geometric mean algorithm can integrate ranked lists from various datasets produced by a wide variety of platforms, such as Affymetrix oligonucleotide arrays, two-color cDNA arrays and other custom-made arrays [14,15]. To increase the sample size, RankProd, a non-parametric statistical method which can combine datasets from different origins to increase the power of identification, has been used to datamine various cancers. Suraj Peri et al. integrated various data from kidney tissue microarrays for differential expression of genes (DEG) analysis and identified NF-κB and interferon signatures as clinical features of clear cell renal cell carcinoma (ccRCC) [16]. Many other studies also employed RankProd to extend their sample size [17–19].

In recent years, some machine learning algorithms, such as support vector machines (SVM), principal component analysis (PCA), least absolute shrinkage and selection operator (LASSO) and artificial neural network (ANN) also help with solving the curse of dimensionality. When comparing these models and methods with ANN, the latter shows many advantages when handling high dimensional data by dealing with non-linear relationships in data [20]. ANN is based on a collection of connected units called artificial neurons (analogous to axons in a biological brain). Neurons and synapses may also have a weight that varies as learning proceeds, which can increase or decrease the strength of the signal that it sends downstream. This model is suitable for prediction even when the experimental data are not subjected to Gaussian distribution for it is established simply by construction of multiple layers of artificial neurons and utilizes its network connections to deliver and process the required input information. Due to its advantages in processing defective or non-linear data, ANN is currently widely used in the diagnosis of cancer, survival analysis and estimation of intensive care [21–23]. Genetic algorithm (GA) is a generally evolutionary algorithm that has already been considered appropriate to solve the general optimization problems [24]. GA is widely used in the selection of the variables resulting in the best fit for the ANN models [25,26]. Although many studies reported the use of ANN in classification gene expression microarray, integrating GA with ANN to establish prediction models and filter candidate genes for cancer samples has few reports.

To solve the curse of dimensionality in high-dimensional gene expression data, we were trying to establish a data processing system using these two methods. We firstly integrated data from different independent datasets to expand the sample size by RankProd. Based on that, we employed a GA-ANN model to establish a prediction model and screen for candidate genes of PCa. The combination of RankProd and GA-ANN in this study, allows us to develop a promising processing approach for discovering novel biomarkers and candidate gene patterns for the diagnosis and prognosis prediction of PCa.

2. Materials and Methods

2.1. Data and Sample Collection

Public data was collected from the Gene Expression Omnibus (GEO) dataset and The Cancer Genome Atlas (TCGA) dataset. Only microarray data that met the following criteria were included; (1) Datasets were produced by Genome-wide mRNA expression profiling by microarray, (2) The experimental platform was single-channel; (3) All cases were pathologically diagnosed to be prostate cancer tissues while the controls were identified as para-carcinoma or normal prostate tissues; (4) The minimum number of cases and controls was 3. Finally, available datasets from the following cohort were included. Wallace et al. contained gene expression profiles of primary prostate tumors resected from 33 African-American and 36 European-American patients. It also contained 18 normal prostate tissues from 7 African-American and 11 European-American patients [27]. Wang et al. contained 148 prostate samples [28]. Planche et al. [29] contained 6 prostate cancer and matched normal samples. Taylor et al. contained 218 PCa samples and 149 matched normal samples from patients treated by radical prostatectomy [30]. Ross-Adams et al. contained 99 prostate cancer samples from patients with follow-up data [31]. For the TCGA dataset, we included the TCGA-PRAD which contained 500 PCa patients.

We also analyzed 69 primary prostate cancer patients and paired adjacent normal tissues by a tissue microarray obtained from Shanghai Outdo Biotech, China (Supplementary file 1: Table S1). Another 28 independent PCa and paired adjacent normal tissues were analyzed by qPCR and western blot (Supplementary file 2: Table S2). All fresh tissues were obtained with informed consent from patients hospitalized at the Department of Urology, Longgang Central Hospital and the Department of Urology, Third Affiliated Hospital of Sun Yat-Sen University. All tissue specimens were confirmed by pathology and immediately frozen in liquid nitrogen. All experiments in this study were approved by the ethics committee of Longgang Central Hospital and Third Affiliated Hospital of Sun Yat-Sen University.

2.2. Individual Participant Data Processing

In order to integrate microarray data from different platforms, meta-analysis was carried out by RankProd. The annotation files corresponding to the types of microarrays were downloaded from the official Affymetrix website. To pre-process Affymetrix microarray data, RMAExpress1.0.5 was introduced for background adjustments, normalization was done by Quantile and summarization by Median Polish. The output files were composed of the normalized expression values of every probe. Shared probes were extracted from different platforms using Perl 5.10 and RankProd package installed in R (v3.4.0) was run. Probe signals with percentage of false prediction (pfp) value lower than 0.05 would be considered as DEGs. GO enrichment and KEGG analysis were carried out using clusterProfiler package in R (v3.4.0) [32].

2.3. Development of GA-ANN PCa Prediction Model

After acquiring the DEG list, we constructed the ANN model in MATLAB (MathWorks, Massachusetts, USA) by setting the clinical phenotype of 163 microarray samples as the output variable (normal or cancer patients) and the expression values of the top 500 up- and down-regulated probes as the input variables. A training set was built with 100 randomly selected microarray samples and the other 63 microarray samples were used as a test set. The model was composed of 3 layers with 1000 nodes as the input layer (each representing an expression value of a probe) and 1 node as the output layer (the clinical phenotype). We set the maximum recursive time to 100 and the threshold of mean square error to 0.005. The weight-corrected learning rate was 0.1 and the transfer function from input layer to hidden layer was tansig while purelin was configured as the transfer function from

hidden layer to output layer. In terms of optimization by GA, the number of initial population was 100 and the maximum evolutionary generation was 50. During each round of calculation, GA-ANN randomly selected the useful input variables keeping the computational accuracy stable. Therefore, the number of input variables could reduce nearly half every round. After 6 rounds of calculations, 15 candidate input variables (probes) were obtained.

2.4. Diagnosis Assay for 15-Gene Signature in Independent Dataset

The prediction accuracy was calculated both in the training and test set. We employed the genes from the TCGA cohort to assay their relative risk and capacity of diagnosis by logistic regression. This test was performed on "glm" function in R software. Then, a linear model was constructed by combining the gene expressions. A coefficient of logistic regression and index by combination was assigned to each sample. Finally, the area under curve (AUC) of the receiver operating characteristic (ROC) curves was employed to estimate the performance of the model with the "ROCR" package in R.

2.5. Prognostic Index of 15-Gene Signature in Prognosis of Survival of PCa

A prognostic index (PI) [33] was constructed as an integrated indicator of the 15 candidate genes selected by the ANN model for each PCa patient. The PI was calculated as a linear combination of the expression value of the genes weighted by univariate Cox regression coefficients. The standard form of PI was defined as follow:

$$\text{Prognostic index(PI)} = \sum_{i} (\beta_i \times X_i)$$

β_i is the regression coefficient of the ith variable and X_i is the value of the ith variable. For the form of PI, X_i is the log2-transformed expression value of each mRNA and β_i is the univariate Cox regression coefficient of the ith RNA.

2.6. Investigation of Diagnosis and Prognosis Capacity of C1QTNF3 in PCa

The capacity of *C1QTNF3* to diagnose PCa was evaluated by measuring the AUC of the ROC curves using the "ROCR" package in R. To integrate and combine the results from three *C1QTNF3* probes, the "aggregate" function of R was applied. Differential expression of *C1QTNF3* in tumor and normal tissue was computed by the "limma" package in R. Logistic regression was measured using the "glm" function in R (Version is 3.4.0). *C1QTNF3* was validated by analyzing available PCa samples in the TCGA database with the cBioPortal web tool (http://www.cbioportal.org/index.do) [34]. Survival analysis was calculated automatically by this tool.

2.7. Quantitative Real-Time PCR Analysis (qRT-PCR)

Total RNA was extracted from patients' tissues samples with TRIzol reagent (Invitrogen, USA) and treated with DNase I (Merck, Sigma, USA). A total of 2 μg of RNA was reverse transcribed into cDNA with oligo (dT) primers using the cDNA synthesis kit (Takara, Japan). Quantitative PCR was performed in 20 μl reactions using SYBR Green qPCR Master Mix (Takara, Japan) according to manufacturer's instruction. β-actin mRNA levels were used for normalization. The following primers were used to amplify a 110-bp PCR product for *C1QTNF3*: forward, 5′- GGCAACACA GTCTTCAGCAT-3′; reverse, 5′- ATTCGCAGCCAAACCTCATC-3′, a 98-bp PCR product for *FABP5*: forward, 5′-AGATGGTGCATTGGTTCAGC-3′; reverse, 3′-TCATGACACACTCCACCACT-5′, a 115-bp PCR product for *LPHN3*: forward, 5′-CACACCTTCCATCAGCATCG-3′; reverse, 3′-GGCTGC TTTGCTATCTGTCC-5′ and a 120-bp PCR product for *ACTB*: forward, 5′-ACTCTTCCAGCCTTCCTTCC-3′; reverse, 5′- CGTACAGGTCTTTGCGGATG-3′. The PCR amplification program was as follow; initial denaturing at 95 °C for 10 min, and then denature at 95 °C for 10 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 45 s. The mRNA level of *C1QTNF3*, *FABP5*, and *LPHN3* was measured using the Applied Biosystems 7500 Real-Time PCR System (ABI, USA). Measurements were repeated 3 times and relative quantification analysis was performed using the comparative CT (2^-ΔΔCT) method.

2.8. Western Blot

Cancer or paired normal tissue (0.2 g) was crushed in liquid nitrogen and lysed in RIPA lysis buffer (CelLytic, Sigma-Aldrich) in the presence of a proteinase inhibitor cocktail (Merck Millipore, USA). Total protein extracts were separated by SDS-PAGE and transferred to PVDF membranes (Merck Millipore, USA). Immunoblotting was done with rabbit polyclonal antibody against C1QTNF3 (ab36870, Abcam, 1:1500 dilution) in accordance with the manufacturer's instruction. Signals were visualized using enhanced chemiluminescent substrate (ECL, BioRad, Richmond, CA, USA) and the Western Breeze chromogenic detection system (Invitrogen).

2.9. Tissue Microarray and Immunohistochemistry Staining

The human PCa tissue microarray (HPro-Ade180PG-02; Shanghai Outdo Biotech, China) was constructed with formalin-fixed, paraffin-embedded PCa tissues and paired adjacent normal tissues. Immunohistochemistry staining was performed by Shanghai Outdo Biotech Co., Ltd. Tissue microarray sections were blocked with goat serum, incubated with anti-C1QTNF3 (ab36870, Abcam, 1:200 dilution), deparaffinized, rehydrated, and subjected to heat-induced antigen retrieval, as previously described [35]. The expression of *C1QTNF3* in each tissue was semi quantitatively graded by two independent pathologists according to staining intensity (0, negative, 1, weakly positive; 2, moderately positive; or 3, strongly positive).

3. Results

3.1. Raw Microarray Data and Overall Processing Methodology Description

The microarray assays enrolled in this study included GSE6956 [27], GSE8218 [28], GSE26910 [29], GSE21032 [30] and GSE70769 [31]. GSE6956 is composed of 69 primary prostate tumors and 18 non-tumor prostate tissues (Platform: GPL571 Affymetrix Human Genome U133A 2.0 Array). GSE8218 contained 148 prostate samples with various amount of different cell types of which 10 were normal (Platform: Affymetrix Human Genome U133A Array). By filtering out unidentified cellular component samples from GSE8218, we obtained 133 cases and 30 controls in total. GSE26910 contained 6 samples of stroma surrounding invasive prostate primary tumors and 6 matched samples of normal stroma samples (Platform: GPL570 Affymetrix Human Genome U133 plus 2.0 Array) and was used for validation. A detailed description of GEO datasets is available in Supplementary file 3: Table S3. Gene expression profile from the TCGA was used for verification. After screening the clinical data (excluding NA in survival time), 466 patients' samples were selected. Furthermore, two independent datasets were used to validate prognostic capability of the 15-gene signature. From the Taylor dataset, gene expression and follow-up data of 140 patients with primary prostate cancer were collected from GSE21032. The Ross-Adams dataset was collected from GSE70769. After matching gene expression and clinical data samples, 92 primary prostate cancer patients were included for analysis. The flowchart in Fig. 1 shows the data analysis process.

3.2. DEG Identification by RankProd

RankProd was performed to detect DEGs. When we restricted the conditions to $pfp \approx 0$, the number of up-regulated probes dropped to 2306 and down-regulated to 1311. The top 500 up- and down-

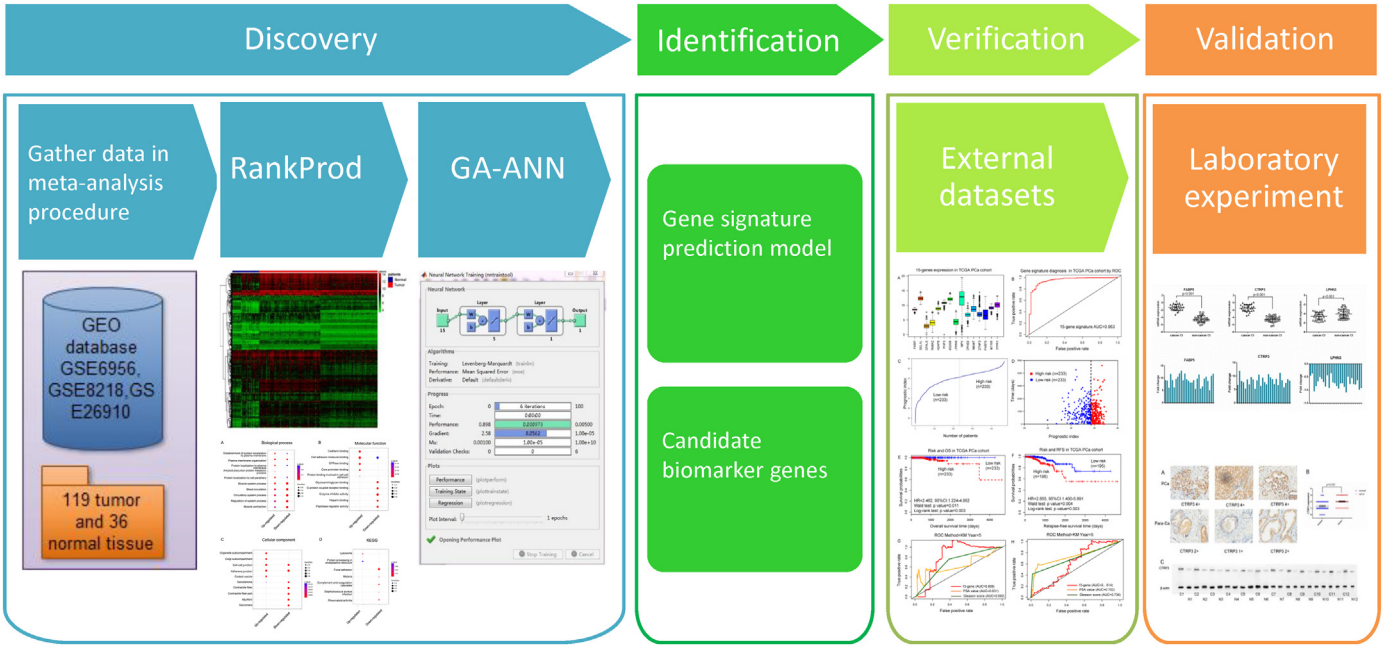

Fig. 1. Flowchart for the systematic analysis and validation of key genes in PCa.

regulated probes are shown in a heatmap plot (Fig. 2) (Supplementary file 4: Table S4). The whole DEG lists are available in additional file (Supplementary file 5: Table S5).

3.3. Gene Ontology Enrichment and Pathway Enrichment

Differential expressed genes were annotated using the ClusterProfiler package. GO and KEGG analysis indicated that up-regulated genes enriched in pathways were obviously different from down-regulated genes (Supplementary file 6: Table S6). For example, KEGG analysis showed that the up-regulated genes related to cancer pathways included proteins that were involved in protein processing in the endoplasmic reticulum and lysosome, while downregulated genes were mainly involved in cancer pathways including focal adhesion, and complement and coagulation cascades (Fig. 3).

3.4. GA-ANN Screening for Candidate PCa Biomarker Genes

After we obtained the DEGs from RankProd, we adjusted the number of nodes in the hidden layer to improve the prediction accuracy of the ANN model. As this can also lead to dramatic complexity of the neural network and an increase in modeling duration, we fixed the optimal

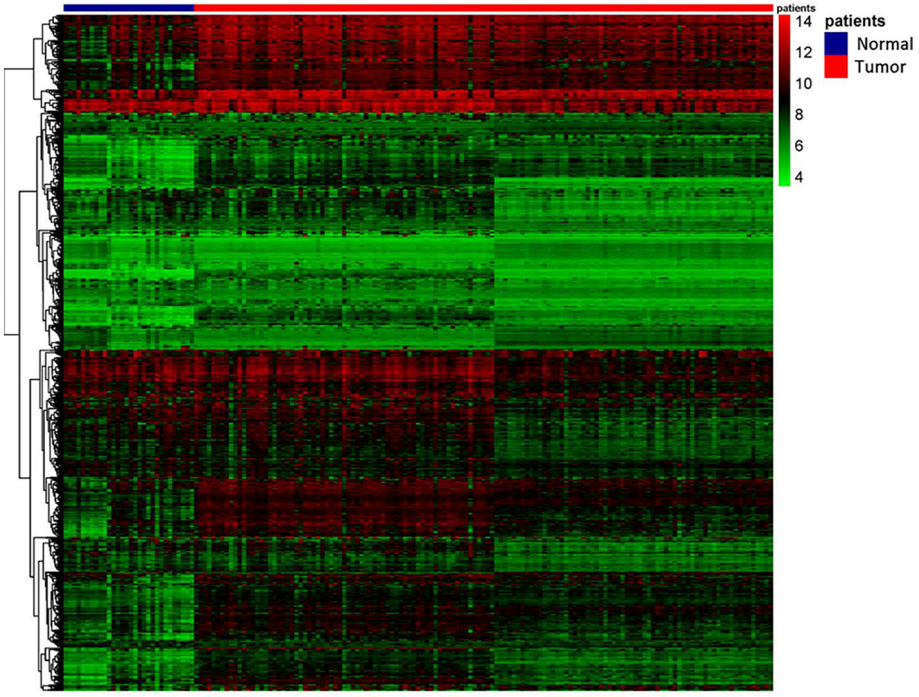

Fig. 2. Heatmap plot of top 1000 differentially expressed genes (DEG) from Rankprod. The blue shade represented normal tissue and red shade represented patients tumor tissue.

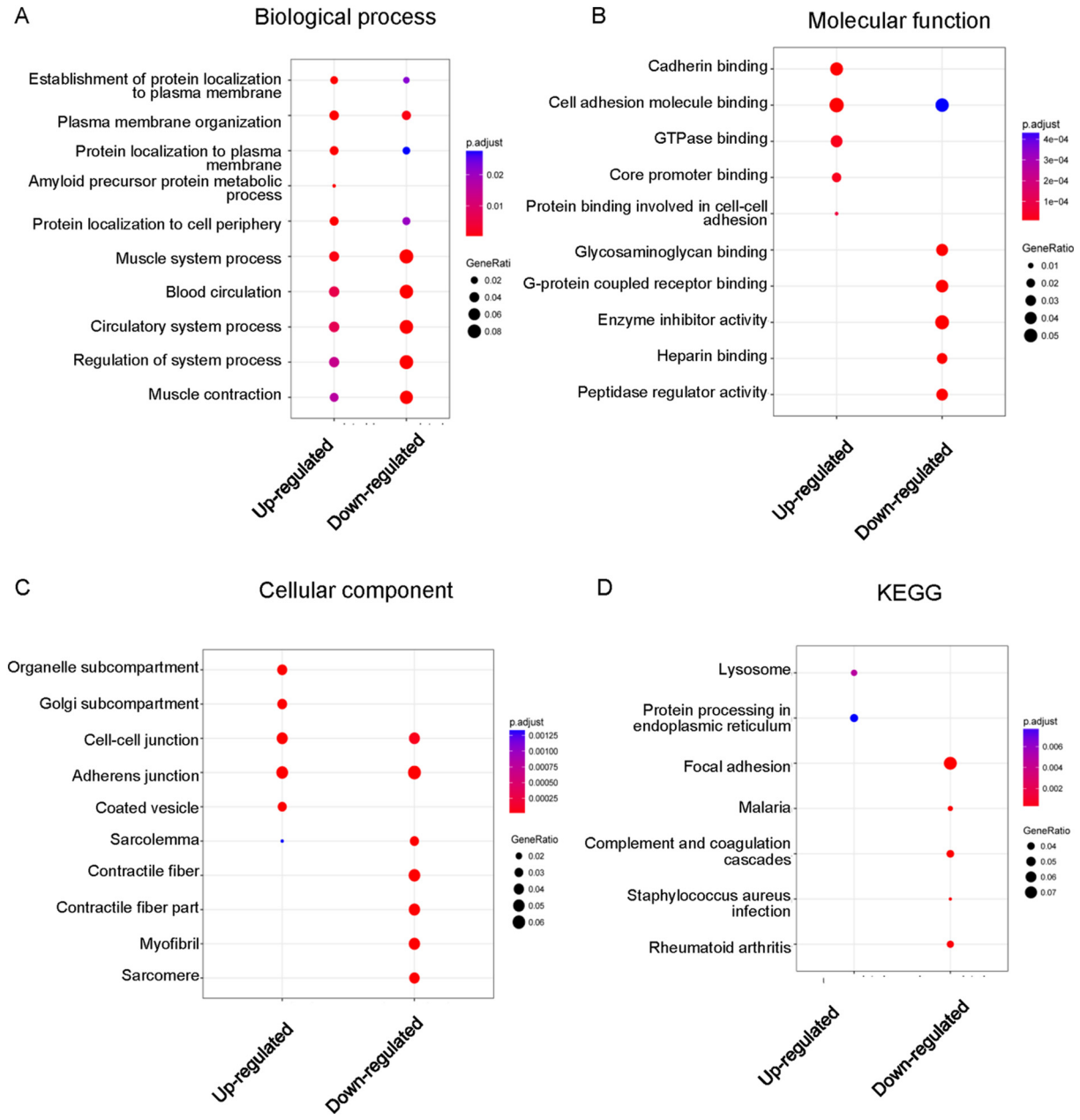

Fig. 3. GO enrichment and KEGG pathway analysis for up and down-regulated genes in PCa (a) Biological process (b) Molecular function (c) Cellular component (d) KEGG pathway.

number of nodes in the hidden layer at 5 (Fig. 4a, b). Furthermore, we noticed a significant advantage of genetic algorithm optimized ANN over general ANN in the performance of prediction and modeling duration. The prediction accuracy of both the training and test set reached 100% with high modeling speed (1.326 s). The process of training and testing are listed in Table 1.

Finally, we obtained 15 genes (Table 2) as a minimum candidate gene list to let the ANN model predict whether a prostate sample was normal or tumor tissue (Fig. 4c, d, e).

3.5. Diagnosis and Prognosis Capacity of Candidate Genes (15-gene signature) for PCa in TCGA Datasets

The expression distribution of the 15 genes in the TCGA PCa cohort is shown by boxplot (Fig. 5a). The *P* values, hazard ratios (HR) and coefficients of the 15 genes in overall survival prediction model for TCGA cohort are listed (Table 2). These genes have a high AUC value (0.953), which represents the high diagnosis capacity in this model (Fig. 5b). The patients in the TCGA PCa cohort were ranked according to the PI. Using the median value of PI as the cutoff, 466 patients were divided into 2 groups: a high-risk group with 233 patients and low-risk group with 233 patients (Supplementary file 7: Fig. S1). The PI was significantly associated with PCa patient 5 years overall survival (OS) (Fig. 5c) and recurrence-free survival (RFS) (Fig. 5d). The survival rates of the high-risk group in OS and RFS were both significantly lower than that of the low-risk group (log-rank *P* value = .003). We used the 5 years OS and RFS survival rate to compare the prognostic capacity of the 15-gene signature model, PSA screening and the Gleason score. The Gleason score is the most popular pathology grade for PCa and is a measure of how likely the tumor will grow invasively. The results suggested that the 15-gene signature was the best index for predicting PCa in 5 years OS with an

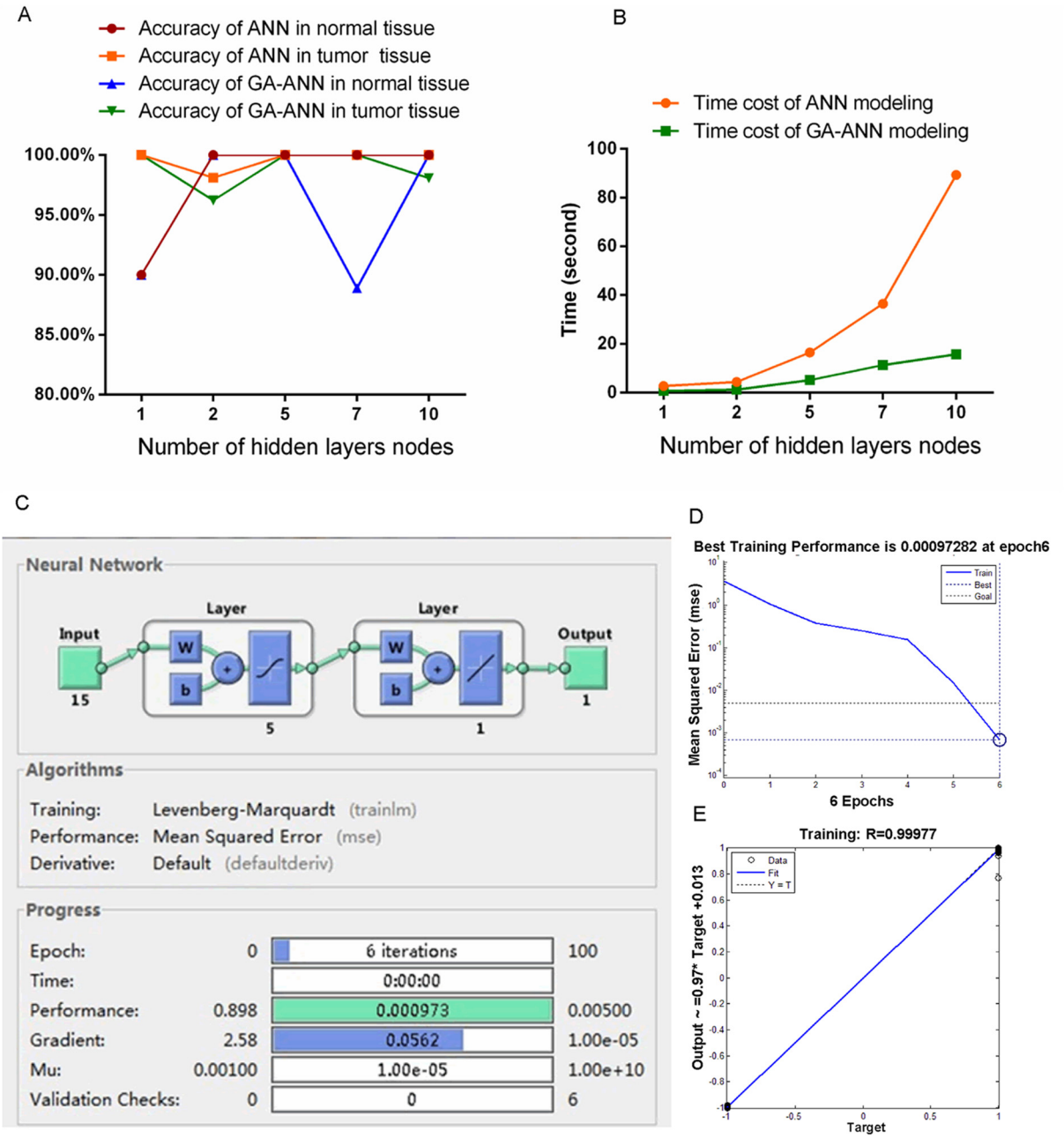

Fig. 4. ANN model training process (a) The number of hidden layer nodes affects the accuracy of the ANN and GA-ANN model (b) The number of hidden layer nodes affects the modeling time in the ANN and GA-ANN model. (c) The configuration of the final ANN (d) The plot of mean squared error in training ANN. After six epochs, the mean squared error of prediction model trained by ANN descends below the threshold 0.005 (e) The regression plot shows the relationship between outputs of prediction model trained by ANN and targets. The regression plot suggests the training of prediction model is perfect; the outputs are nearly equal to the targets.

AUC = 0.808. The AUC of PSA screening and the Gleason score were 0.631 and 0.692 respectively (Fig. 5e). As for the 5 years RFS, AUCs of the three indicators demonstrated that the 15-gene signature model (AUC = 0.614), PSA screening (AUC = 0.702) and the Gleason score (AUC = 0.740) have similar capacity, and the Gleason score performed best (Fig. 5f).

To further validate the performance of the 15-gene signature, two independent datasets of Taylor et al. and Ross-Adams et al. were employed. The results showed that the 15-gene signature performed well in both independent datasets. PI calculated from the 15-gene signature can significantly classify patients into low- and high-risk (Taylor cohort: HR = 2.893, *p* value = .003, AUC = 0.74; Ross-Adams cohort: HR = 1.886, *p* value = .03, AUC = 0.67). The Kaplan-Meier and ROC curve for the two datasets are shown in Supplementary file 8 (Fig. S2).

3.6. Validation of Candidate Genes in Prostate Cancer Tissues

From the 15 genes, we selected 3 genes (*FABP5*, *C1QTNF3* and *LPHN3*) for further analysis (Table 2). To analyze mRNA levels of *FABP5*, *C1QTNF3* and *LPHN3*, qRT-PCR was performed on tissues from 28 prostate cancer patients. The mRNA level of *FABP5* and *C1QTNF3* was in all prostate cancer tissues higher than in the paired adjacent

Table 1
Table of parameter of ANN, accuracy rate of prediction and number of genes filtered from GA-ANN.

Nodes of Input layer	popu = 100, gen = 50				popu = 50, gen = 50	popu = 20, gen = 50	15-gene signature
	$S = 1000$	$S = 478$	$S = 229$	$S = 122$	$S = 54$	$S = 27$	
Training(normal/tumor)	100(22/78)	100(20/80)	100(19/81)	100(16/84)	100(17/83)	100(21/79)	100(22/78)
Testing(normal/tumor)	63(8/55)	63(10/53)	63(11/52)	63(14/49)	63(13/50)	63(9/54)	63(8/55)
Testing results of ANN							
Normal(accuracy rate)	8(100%)	9(90%)	11(100%)	14(100%)	10(76.92%)	9(100%)	8(100%)
Tumor(accuracy rate)	55(100%)	52(98.11%)	52(100%)	49(100%)	50(100%)	54(100%)	55(100%)
Time cost for modeling	23.525 s	4.3524 s	2.5428 s	3.2448 s	1.3416 s	1.9032 s	1.326 s
Testing results of GA-ANN							
Normal(accuracy rate)	8(100%)	10(100%)	11(100%)	14(100%)	13(100%)	9(100%)	
Tumor(accuracy rate)	55(100%)	52(98.11%)	52(100%)	48(97.96%)	50(100%)	54(100%)	
Time cost for modeling	3.822 s	1.5132 s	0.96721 s	0.93601 s	0.5616 s	0.6864 s	
Iterations	2	3	3	6	3	6	6
Candidate genes	478	229	122	54	27	15	

normal tissues (average fold change of 9.19 ($P < .01$) and 8.23 ($P < .01$) respectively, Fig. 6) and the level of *LPHN3* was lower (average fold change of 0.72 ($P < .01$), Fig. 6).

Immunohistochemical staining of a tissue microarray containing an additional 69 pairs of PCa and their paired adjacent normal tissues showed that C1QTNF3 levels were significantly higher in PCa tissues when compared with the paired adjacent normal tissues ($P < .001$, Fig. 7a, b). Western blot assay for another 28 paired tissues samples was performed to confirm the protein levels of C1QTNF3. All PCa tissues showed a higher C1QTNF3 protein expression than in the paired adjacent normal tissues (Fig. 7c). Overall, these data suggest that C1QTNF3 is constantly overexpressed in PCa.

3.7. Diagnostic and Prognostic Capacity of C1QTNF3 for PCa Prediction in Various Datasets

C1QTNF3 expression was analyzed in GEO datasets GSE6956, GSE8218 and GSE26910 (Fig. 8a). The AUC of the ROC curve showed that *C1QTNF3* showed good performance on diagnosis for PCa in all three datasets (GSE6956: OR = 1.253, 95% CI: 0.872–1.636, $P = .001$, AUC = 0.791; GSE8218: OR = 2.848, 95% CI: 1.365–4.331, $P = .055$, AUC = 0.868; GSE26910: OR = 5.332, 95% CI: 2.062–8.602, $P = .105$, AUC = 0.972, Fig. 8b). Additionally, we also tested the prognostic ability of *C1QTNF3* in the TCGA dataset. The results showed that *C1QTNF3* overexpression is closely associated with recurrence-free survival time ($P < .001$, AUC = 0.57) (Fig. 8c, d).

Table 2
The *P* values, HR and coefficients of 15 genes in overall survival prediction model for the TCGA cohort.

Gene symbol	*P* value	HR (95% CI)	Coefficient
FKRP	0.727	1.032 (0.865–1.231)	0.031
SEL1L	0.580	1.073(0.837–1.375)	0.070
IGFALS	0.742	1.027(0.877–1.202)	0.026
PNMA2	0.801	1.015(0.901–1.144)	0.015
ARHGAP8	0.693	1.015(0.944–1.090)	0.014
PHF3	0.284	1.093(0.929–1.287)	0.089
HMG20B	0.338	1.205(0.822–1.766)	0.187
LPHN3	0.869	0.991(0.895–1.098)	−0.009
NPY	0.372	0.989(0.966–1.013)	−0.011
EPHB2	0.225	0.925(0.815–1.049)	−0.078
NNMT	0.670	0.991(0.950–1.034)	−0.009
C1QTNF3	0.247	0.940(0.848–1.043)	−0.061
FABP5	0.485	0.986(0.947–1.026)	−0.014
MTRR	0.144	1.069(0.978–1.168)	0.066
ITPR1	0.322	1.028(0.973–1.086)	0.028

4. Discussion

Prediction and diagnosis is the most important step in PCa management for patients. In order to screen candidate biomarkers which may be helpful for diagnoses and prognosis for PCa, we have combined RankProd with GA-ANN to create a prediction model. This process could also provide a general framework for rational cancer gene signature discovery based on high throughput data. To datamine oncogenes, biomarkers or gene signature prediction models for prostate cancer, high throughput data from microarray or next generation sequencing is a fundamental source. Data processing approach plays a crucial part in such studies. Since the ANN model can fit any nonlinear function it has more advantages in processing high-throughput data. At the moment, depth neural networks have been applied to a variety of artificial intelligence applications. In the future, neural networks are bound to be used more in molecular medicine.

In this study, a 15-gene signature was identified by our data processing system that exhibited a great capacity for diagnosis and prognosis of PCa. The AUCs of the 15 genes signature showed a perfect diagnostic ability in PCa gene expression samples from datasets from both GEO and TCGA. Although the genes individual were not significant in the 5-year OS prognostic test, the 15-gene signature can effectively classify PCa patients into high- and low-risk groups, and showed a good prediction of the 5-year survival rate in the PCa cohort from TCGA, Taylor et al. and Ross-Adams et al. Other studies also tried to establish prediction models for diagnosis or prognosis of PCa. Cima et al. combined bioinformatic prioritization with targeted proteomics and machine learning to build predictive regression models for tissue PTEN status and diagnosis and grading of PCa [8]. Wu et al. constructed a 32-gene signature model which could predict PSA recurrence of post-radical prostatectomy patients via PCA coupling with Cox regression [36]. In 2014, Bismar et al. used a singular value decomposition (SVD) method to identify an ETS transcription factor (EGR) relative 10 genes signature for establishing a prognostic prediction model to predict patients' clinical outcome [37]. Our work here provides a different approach to establish prediction model and select the candidate oncogenes or biomarkers.

From the 15-gene signature model, we selected 3 candidate genes (*FABP5*, *LPHN3* and *C1QTNF3*) for further analysis. *FABP5* has been demonstrated as a target of PCa in previous studies [38,39]. *Latrophilin 3* (*LPHN3*) is a brain-specific member of the G-protein coupled receptor family associated to both attention-deficit/hyperactivity disorder (ADHD) genetic susceptibility and methylphenidate (MPH) pharmacogenetics [40] and was down regulated in PCa. Interestingly, we found C1q and tumor necrosis factor related protein 3 (*C1QTNF3*, alias as *CTRP3*), a highly expressed gene in PCa tissue, in the 15-gene list. In addition, the three probes (209424_s_at, 209426_s_at, 209425_at)

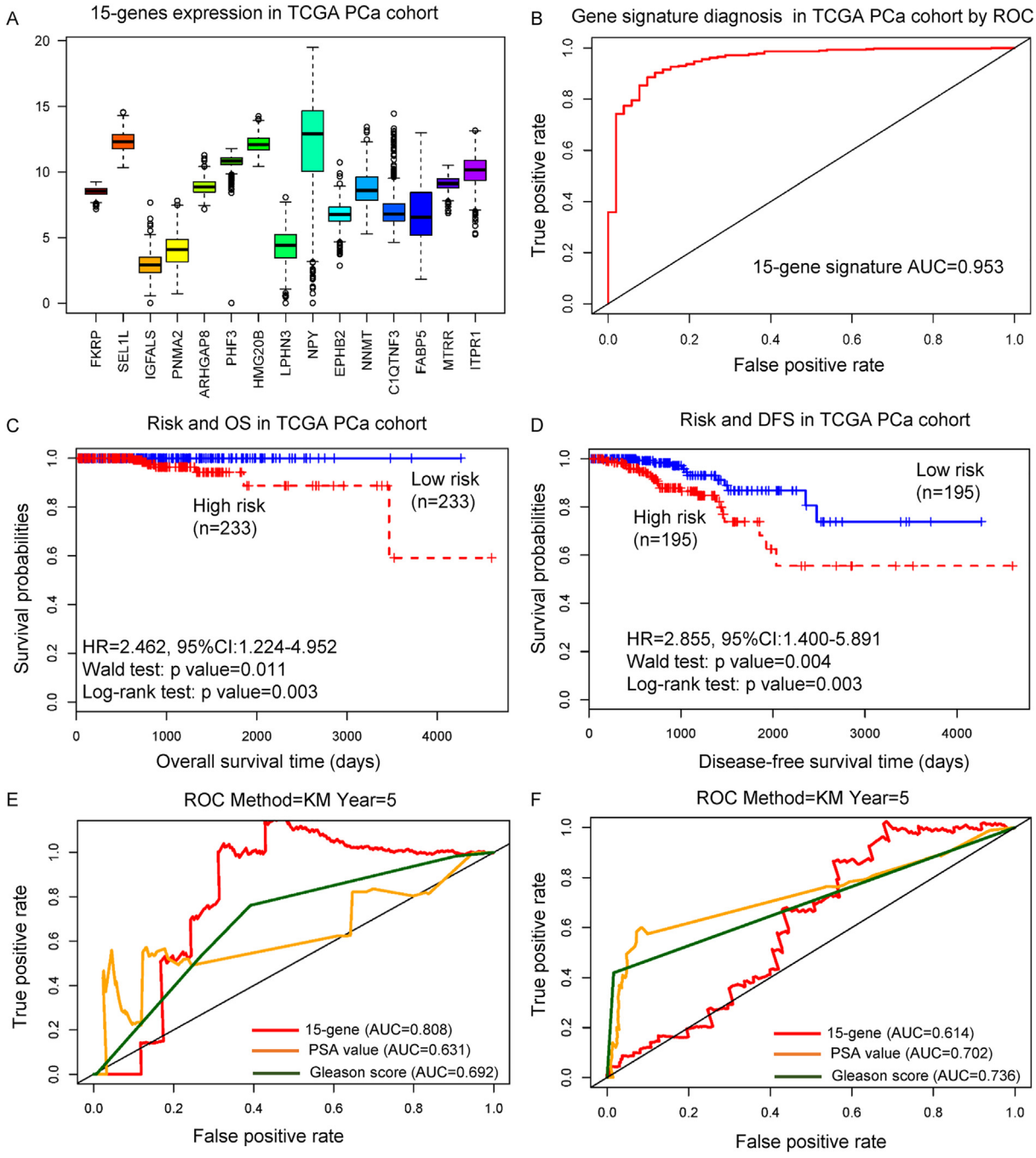

Fig. 5. Diagnostic and prognostic capacity of the 15-gene signature for PCa in TCGA dataset. (a) 15-gene expression value distribution in TCGA PCa cohort by boxplots. The line within the box indicates the median value; the box spans the interquartile range. (b) ROC curve for the 15-gene signature for PCa diagnosis (c) Kaplan-Meier curves for the low- and high-risk groups separated by the PI of the 15-gene signature in the TCGA PCa cohort. Significant differences in overall survival between the 2 groups were analyzed by log-rank test ($P = .003$). (d) Kaplan-Meier curves for the low-risk and high-risk groups of the 15-gene signature in the TCGA PCa cohort. Significant differences in DFS between the two groups were determined by the log-rank test ($P = .003$). (e) ROC curves for the prediction of the 5 years overall survival among the 15-gene signature model, PSA screening and the Gleason score. (f) ROC curves for the 5 years DFS among 15-gene signature, PSA screening and the Gleason score.

standing for *C1QTNF3* respectively ranked the first, third and fourth place among the significantly up-regulated probes. These hints have inspired us to explore the role of *C1QTNF3* as a susceptibility gene in PCa. Our previous studies have demonstrated that C1QTNF3 stimulated proliferation and anti-apoptosis in prostate cells through the protein kinase C signaling pathway [41]. Furthermore, C1QTNF3 regulated 14-3-3 sigma and GLRX3 which has functions in various kinds of tumors as well as in prostate cancer [41]. It suggests that *C1QTNF3* may promote the transformation from prostate cells to malignant cells.

To confirm our findings, we validated *C1QTNF3* by qPCR, Western Blot and tissue microarray. We were able to draw the conclusion that *C1QTNF3* was significantly overexpressed in prostate tumor tissues. Besides, in the GEO dataset, the AUCs of ROC demonstrated *C1QTNF3* had good performance for PCa diagnosis. The TCGA dataset showed that *C1QTNF3* expression is closely associated with DFS time of prostate cancer patients. These results together indicate that *C1QTNF3*, as a biomarker for diagnosis and prognosis of PCa, has a high reliability and accuracy.

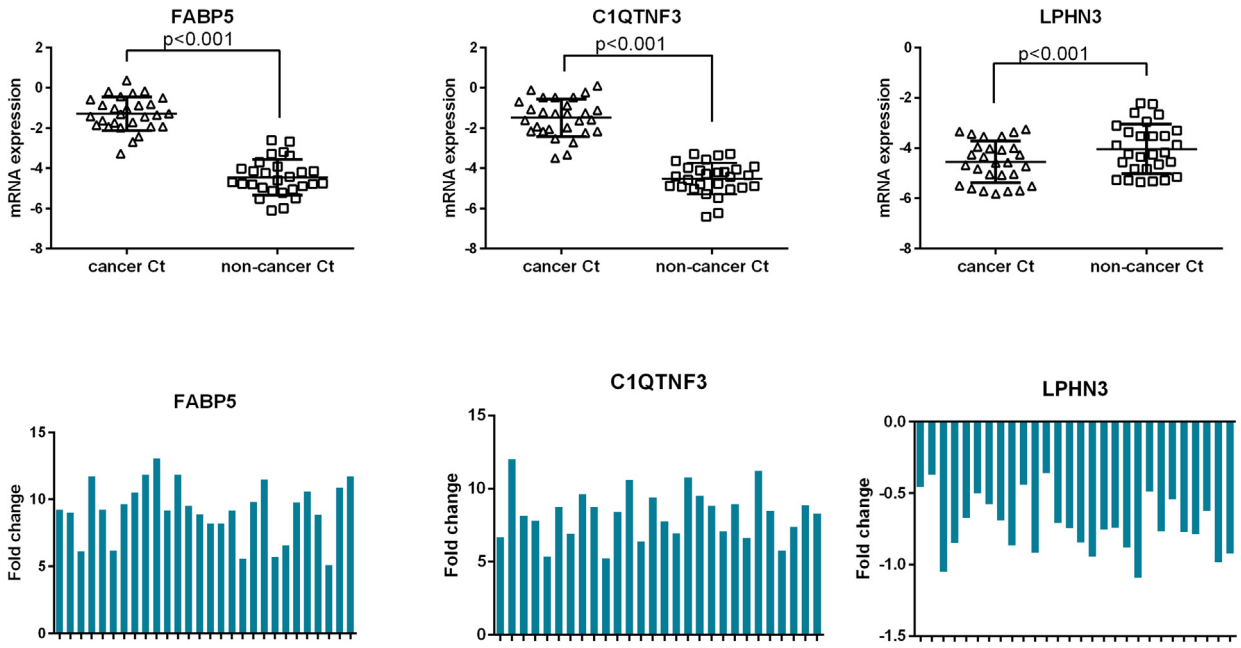

Fig. 6. qPCR assay of *FABP5*, *C1QTNF3* and *LPHN3* genes in PCa and normal adjacent tissues. Scatter diagram of the gene expression and fold changed distribution of gene expression in different samples.

The data processing approach presented here provides a new view for the discovery of biomarkers with the aim of promoting diagnostic and prognostic prediction of PCa. This approach can also be applied to other high-throughput data for the discovery of oncogenes or biomarkers in different kinds of diseases and different platforms. In our study, we have established a diagnostic and prognostic prediction model and revealed *C1QTNF3* as a promising biomarker for prostate cancer. However, more studies are warranted to determine the roles of the 15-gene signature prediction model and *C1QTNF3* for PCa.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ebiom.2018.05.010.

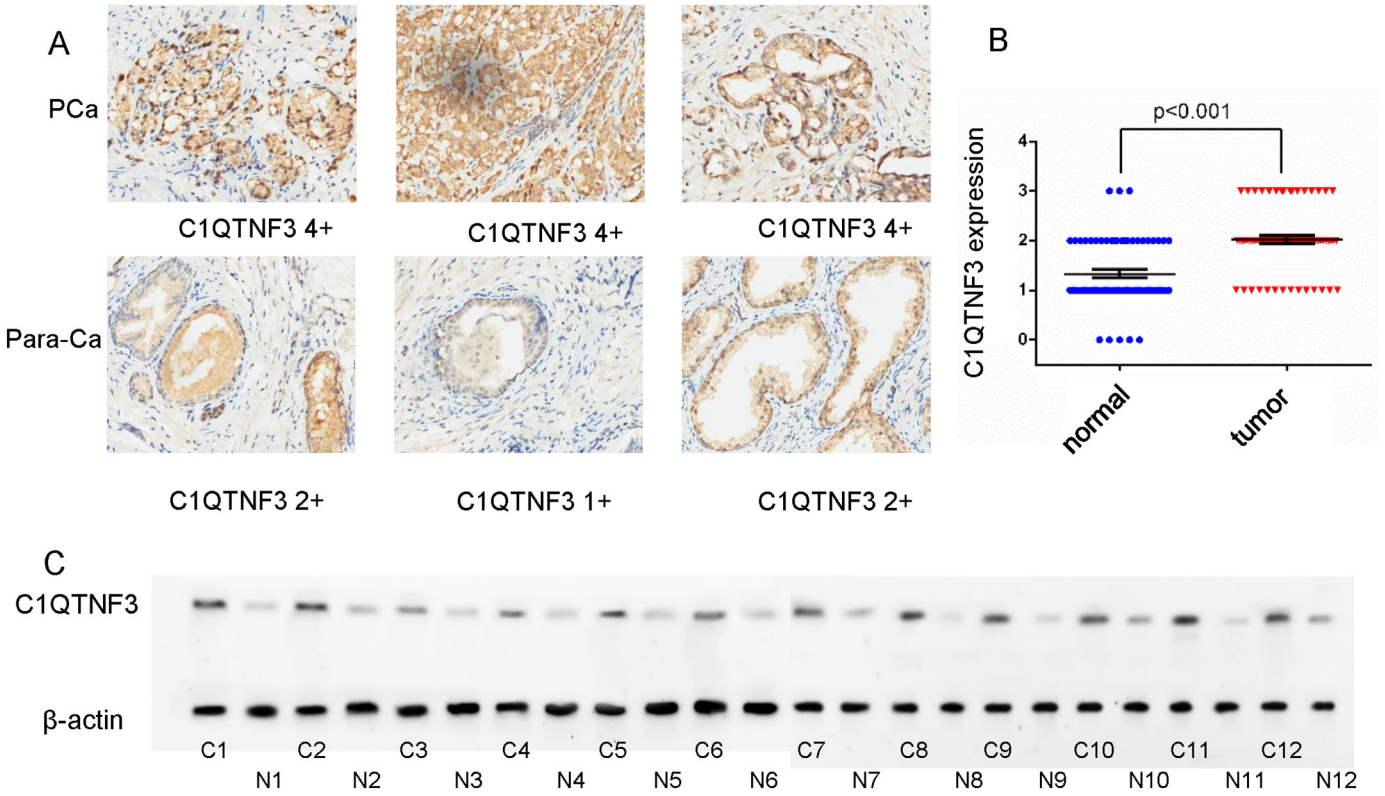

Fig. 7. Validation of C1QTNF3 expression in tissues microarray and Western blot. (a). Pathological sections of PCa and para-carcinoma tissue. (b) C1QTNF3 expression in tumor tissue was significantly increased when compared with para-carcinoma tissues. (c) C1QTNF3 expression assayed by Western blot.

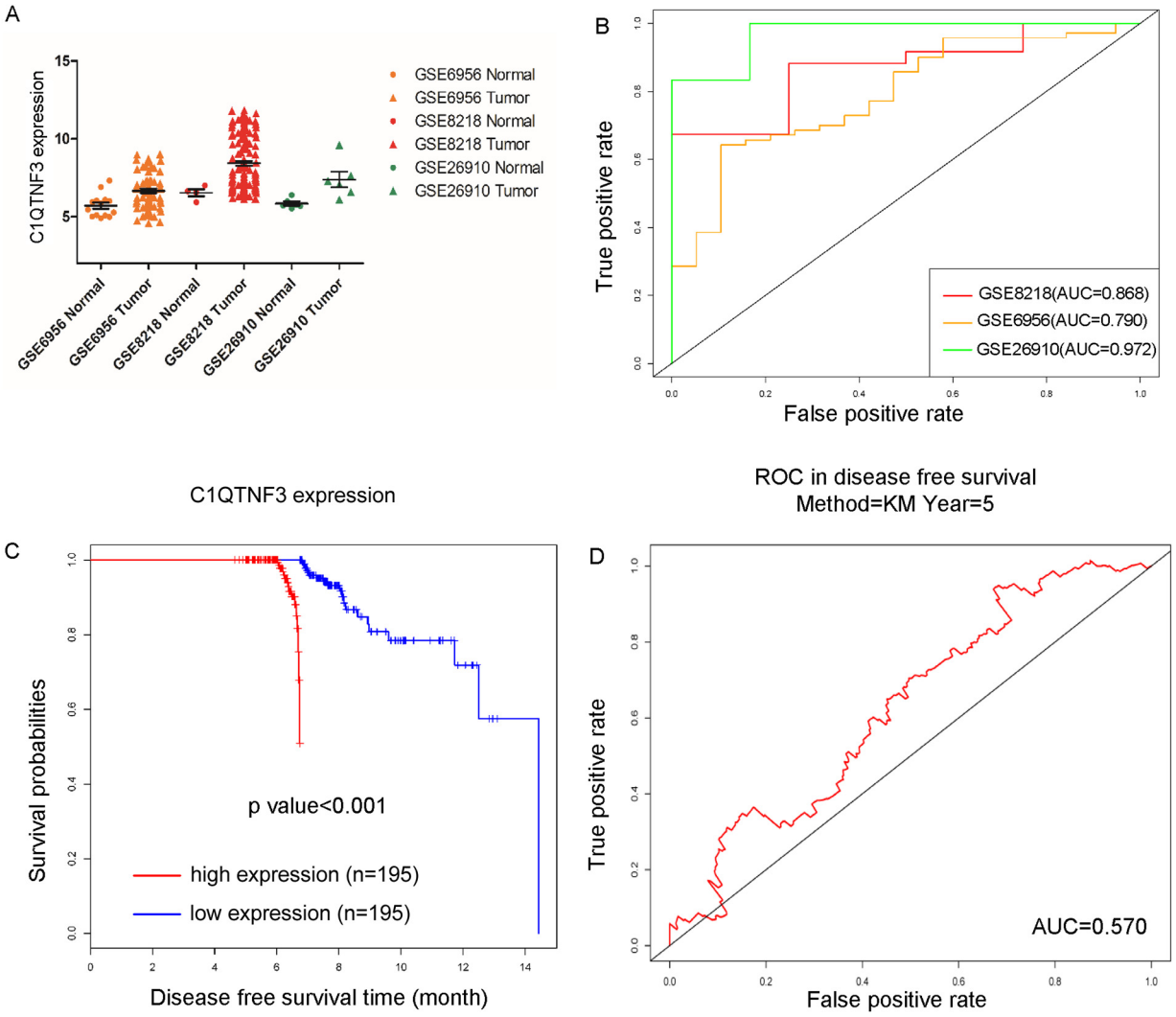

Fig. 8. Diagnostic and prognostic capacity of *C1QTNF3* for PCa in GEO and TCGA datasets. (a) *C1QTNF3* expression in three datasets (GSE6956, GSE8218, GSE26910) of GEO database. (b) The AUC of the ROC curve showed diagnostic capacity of *C1QTNF3* for PCa in GSE6956, GSE8218 and GSE26910 datasets. (c) *C1QTNF3* expression was associated with DFS time (log-rank test, *P* value < .001). (d) AUC of *C1QTNF3* in DFS time is 0.57.

Competing Interests

All authors declare that they have no competing interests.

Acknowledgement

This study was supported by grants from the National Post-doctoral Science Fund (2016M602451), the Natural Science Fund of Guangdong Province (2015A030310384), the Science and Technology Fund of Guangdong Province (2016A020215026), the Medical research Fund of Guangdong Province (C2015025), the Science and Technology Fund of Shenzhen City (JCYJ20160427101148065), the research project of Health and Family Planning Commission of Shenzhen City (20150524) and the Medical Science and Technology Program of Shenzhen Longgang District (20160606163318972). The funders had no involvement in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors' Contributions

QH conceptualized the project and contributed to experimental design, all data analysis and wrote the first draft of the manuscript. ZB contributed to processing, analysis, and interpretation of the data. CH, ML, ZM, XX and YL contributed to sample acquisition and experiment. KY and JL contributed to experimental design. SH and ML contributed to guide the experimental design, data analysis, and manuscript writing. All authors read and approved the final manuscript.

References

[1] Siegel RL, Miller KD, Di AJ. Cancer statistics, 2018. CA Cancer J Clin 2018;68.
[2] Zhang W, Xiang YB, Liu ZW, Fang RR, Ruan ZX, Sun L, et al. Trends analysis of common urologic neoplasm incidence of elderly people in Shanghai, 1973–1999. Chinese J Cancer 2004;23:555.
[3] Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, et al. Cancer statistics in China, 2015. CA Cancer J Clin 2016;66:115.
[4] Miller KD, Siegel RL, Lin CC, Mariotto AB, Kramer JL, Rowland JH, et al. Cancer treatment and survivorship statistics, 2016. CA Cancer J Clin 2016;66:271.
[5] Andriole GL, Grubb III RL, Buys SS, Chia D, Church TR, Fouad MN, et al. Mortality results from a randomized prostate-cancer screening trial. N Engl J Med 2009;360:1310.
[6] Loeb S, Gashti SN, Catalona WJ. Exclusion of inflammation in the differential diagnosis of an elevated prostate-specific antigen (PSA). Urol Oncol-Semin Orig Investig 2009;27:64–6.
[7] Stattin P, Vickers AJ, Sjoberg DD, Johansson R, Granfors T, Johansson M, et al. Improving the specificity of screening for lethal prostate Cancer using prostate-specific antigen and a panel of Kallikrein markers: a nested case-control study. Eur Urol 2015;68:207–13.
[8] Cima I, Schiess R, Wild P, Kaelin M, Schuffler P, Lange V, et al. Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. Proc Natl Acad Sci U S A 2011;108:3342–7.

[9] Ankerst DP, Groskopf J, Day JR, Blase A, Rittenhouse H, Pollock BH, et al. Predicting prostate cancer risk through incorporation of prostate cancer gene 3. J Urol 2008; 180:1303–8.
[10] Strand SH, Orntoft TF, Sorensen KD. Prognostic DNA methylation markers for prostate Cancer. Int J Mol Sci 2014;15:16544–76.
[11] Yao LS, Ren SC, Zhang MJ, Du FX, Zhu YS, Yu H, et al. Identification of specific DNA methylation sites on the Y-chromosome as biomarker in prostate cancer. Oncotarget 2015;6:40611–21.
[12] Nanni L, Brahnam S, Lumini A. Combining multiple approaches for gene microarray classification. Bioinformatics 2012;28:1151–7.
[13] Bengio S, Bengio Y. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Trans Neural Netw 2000;11:550–7.
[14] Del Carratore F, Jankevics A, Eisinga R, Heskes T, Hong FX, Breitling R. RankProd 2.0: a refactored bioconductor package for detecting differentially expressed features in molecular profiling datasets. Bioinformatics 2017;33:2774–5.
[15] Hong FX, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 2006;22:2825–7.
[16] Peri S, Devarajan K, Yang DH, Knudson AG, Balachandran S. Meta-analysis identifies NF-kappaB as a therapeutic target in renal cancer. PLoS One 2013;8:e76746.
[17] Chow YP, Alias H, Jamal R. Meta-analysis of gene expression in relapsed childhood B-acute lymphoblastic leukemia. BMC Cancer 2017;17:120.
[18] Lee YS, Kim JK, Park TH, Kim YR, Myeong HS, Kwon K, et al. Systematic identification of novel biomarker signatures associated with acquired erlotinib resistance in cancer cells. Mol Cell Toxicol 2016;12:139–48.
[19] Sun Y, Yuan K, Zhang P, Ma R, Zhang QW, Tian XS. Crosstalk analysis of pathways in breast cancer using a network model based on overlapping differentially expressed genes. Exp Ther Med 2015;10:743–8.
[20] Bourquin J, Schmidli H, van Hoogevest P, Leuenberger H. Advantages of Artificial Neural Networks (ANNs) as alternative modelling technique for data sets showing non-linear relationships using data from a galenical study on a solid dosage form. Eur J Pharm Sci 1998;7:5–16.
[21] Peng JH, Fang YJ, Li CX, Ou QJ, Jiang W, Lu SX, et al. A scoring system based on artificial neural network for predicting 10-year survival in stage II a colon cancer patients after radical surgery. Oncotarget 2016;7:22939–47.
[22] Rau HH, Hsu CY, Lin YA, Atique S, Fuad A, Wei LM, et al. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network. Comput Methods Prog Biomed 2016;125:58–65.
[23] Saadah LM, Chedid FD, Sohail MR, Nazzal YM, Al Kaabi MR, Rahmani AY. Palivizumab prophylaxis during nosocomial outbreaks of respiratory syncytial virus in a neonatal intensive care unit: predicting effectiveness with an artificial neural network model. Pharmacotherapy 2014;34:251–9.
[24] Haupt RL, Haupt SE, Haupt SE. Practical Genetic Algorithms. , vol. 2New York: Wiley; 1998.
[25] Chen Y-C, Yang W-W, Chiu H-W. Artificial neural network prediction for cancer survival time by gene expression data. Paper Presented at: Bioinformatics and Biomedical Engineering, 2009 ICBBE 2009 3rd International Conference on (IEEE).; 2009.
[26] Wu JH, Mei JA, Wen SX, Liao SY, Chen JC, Shen Y. A self-adaptive genetic algorithm-artificial neural network algorithm with leave-one-out cross validation for descriptor selection in QSAR study. J Comput Chem 2010;31:1956–68.
[27] 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.
[28] Wang YP, Xia XQ, Jia ZY, Sawyers A, Yao HZ, Wang-Rodriquez J, et al. In silico estimates of tissue components in surgical samples based on expression profiling data. Cancer Res 2010;70:6448–55.
[29] Planche A, Bacac M, Provero P, Fusco C, Delorenzi M, Stehle JC, et al. Identification of prognostic molecular features in the reactive stroma of human breast and prostate Cancer. PLoS One 2011;6:e18640.
[30] Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, Carver BS, et al. Integrative genomic profiling of human prostate cancer. Cancer Cell 2010;18:11–22.
[31] Rossadams H, Lamb AD, Dunning MJ, Halim S, Lindberg J, Massie CM, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study. Ebiomedicine 2015;2:1133–44.
[32] Yu GC, Wang LG, Han YY, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS J Integr Biol 2012;16:284–7.
[33] Lee SJ, Lindquist K, Segal MR, Covinsky KE. Development and validation of a prognostic index for 4-year mortality in older adults. J Am Med Assoc 2006;295:801–8.
[34] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6:pl1.
[35] Huang J, Zheng DL, Qin FS, Cheng N, Chen H, Wan BB, et al. Genetic and epigenetic silencing of SCARA5 may contribute to human hepatocellular carcinoma by activating FAK signaling. J Clin Investig 2010;120:223–41.
[36] Wu CL, Schroeder BE, Ma XJ, Cutie CJ, Wu SL, Salunga R, et al. Development and validation of a 32-gene prognostic index for prostate cancer progression. Proc Natl Acad Sci U S A 2013;110:6121–6.
[37] Bismar TA, Alshalalfa M, Petersen LF, Teng LH, Gerke T, Bakkar A, et al. Interrogation of ERG gene rearrangements in prostate cancer identifies a prognostic 10-gene signature with relevant implication to patients' clinical outcome. BJU Int 2014;113:309–19.
[38] Forootan SS, Bao ZZ, Forootan FS, Kamalian L, Zhang Y, Bee A, et al. Atelocollagen-delivered siRNA targeting the FABP5 gene as an experimental therapy for prostate cancer in mouse xenografts. Int J Oncol 2010;36:69–76.
[39] Lin JF, Xu J, Tian HY, Gao X, Chen QX, Gu Q, et al. Identification of candidate prostate cancer biomarkers in prostate needle biopsy specimens using proteomic analysis. Int J Cancer 2007;121:2596–605.
[40] Bruxel EM, Salatino-Oliveira A, Akutagava-Martins GC, Tovo-Rodrigues L, Genro JP, Zeni CP, et al. LPHN3 and attention-deficit/hyperactivity disorder: a susceptibility and pharmacogenetic study. Genes Brain Behav 2015;14:419–27.
[41] Hou Q, Lin J, Huang W, Li M, Feng J, Mao X. CTRP3 stimulates proliferation and anti-apoptosis of prostate cells through PKC signaling pathways. PLoS One 2015;10: e0134006.