Research Paper

Identification of microenvironment related potential biomarkers of biochemical recurrence at 3 years after prostatectomy in prostate adenocarcinoma

Xiaoru Sun[1,2], Lu Wang[1,2], Hongkai Li[1,2], Chuandi Jin[1,2], Yuanyuan Yu[1,2], Lei Hou[1,2], Xinhui Liu[1,2], Yifan Yu[1,2], Ran Yan[1,2], Fuzhong Xue[1,2]

[1]Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China
[2]Institute for Medical Dataology, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China

Correspondence to: Fuzhong Xue; **email:** xuefzh@sdu.edu.cn
Keywords: prostate adenocarcinoma, biochemical recurrence, gene expression, tumor microenvironment, targeted maximum likelihood estimation
Received: November 30, 2020 **Accepted:** May 11, 2021 **Published:** June 16, 2021

ABSTRACT

Prostate adenocarcinoma is one of the leading adult malignancies. Identification of multiple causative biomarkers is necessary and helpful for determining the occurrence and prognosis of prostate adenocarcinoma. We aimed to identify the potential prognostic genes in the prostate adenocarcinoma microenvironment and to estimate the causal effects simultaneously. We obtained the gene expression data of prostate adenocarcinoma from TCGA project and identified the differentially expressed genes based on immune-stromal components. Among these genes, 68 were associated with biochemical recurrence at 3 years after prostatectomy in prostate adenocarcinoma. After adjusting for the minimal sets of confounding covariates, 14 genes (*TNFRSF4, ZAP70, ERMN, CXCL5, SPINK6, SLC6A18, CHRM2, TG, CLLU1OS, POSTN, CTSG, NETO1, CEACAM7*, and *IGLV3-22*) related to the microenvironment were identified as prognostic biomarkers using the targeted maximum likelihood estimation. Both the average and individual causal effects were obtained to measure the magnitude of the effect. CIBERSORT and gene set enrichment analyses showed that these prognostic genes were mainly associated with immune responses. *POSTN* and *NETO1* were correlated with androgen receptor expression, a main driver of prostate adenocarcinoma progression. Finally, five genes were validated in another prostate adenocarcinoma cohort (GEO: GSE70770). These findings might lead to the improved prognosis of prostate adenocarcinoma.

INTRODUCTION

Prostate cancer is a common malignant tumor and the leading cause of cancer-related mortality in men [1]. Prostate adenocarcinoma (PRAD) is the most common type of prostate cancer, whereas other types of prostate cancer are relatively rare [2]. The duration between surgery and prostate-specific antigen-defined biochemical recurrence (BCR) (≤ 3 vs. > 3 years) after definitive local therapy is a significant risk factor for defining specific mortality of prostate cancer [3]. Approximately 35% of men who undergo radical prostatectomy have been reported to experience BCR within 10 years [3, 4]. Hence, exploration of new mechanisms of BCR using integrated bioinformatics analysis could be applied to stratify patients at risk and guide the decision-making for treatment.

Many studies have shown that the tumor microenvironment (TME) is implicated in the development and sustained growth, invasion, and metastasis of cancer [5–8]. Infiltrating immune and stromal cells are important components of the TME and have been shown to significantly influence the progression of malignancy [5, 6]. Evidence has suggested that interactions between tumor cells and stroma mediate the development of cancer and tissue preferences for metastasis [9]. In PRAD, the stromal cells express the androgen receptor (AR), which is the main driver of prostate cancer pathogenesis and progression [10, 11]. However, it is still a challenging undertaking to explore the causal effects of gene expressions on the prognosis of patients with PRAD in the TME. In this study, we focused on exploring the prognostic genes in the TME based on the immune and stromal scores, and then detected their causal effects on the PRAD BCR.

Using observational data such as the online databases, the Cancer Genome Atlas (TCGA), to detect causative biomarkers and estimate causal effects is difficult because of the unbalanced distribution of pretreatment variables between treatment groups, henceforth covariates [12, 13]. Several methods have been proposed to overcome these problems, including previously applied propensity scores, inverse probability weighting, and g-computation [14, 15]. These methods rely on the consistent estimations of the exposure or outcome mechanism. In this study, we used targeted maximum likelihood estimation (TMLE), a doubly robust method to detect the prognostic genes and estimate both the average and individual causal effects [16–18]. TMLE is a semi-parametric method that flexibly establishes the causal models using multiple machine learning methods, which requires weaker assumptions than other common models.

When estimating the causal effect of the prognostic gene on the BCR status, controlling too many covariates might result in the poor performance of the estimator [19–21]. Luna et al. proposed the algorithm *CovSel* for covariate selection, which reduced the dimension of the covariate set for estimation of the causal effect [20, 22]. Furthermore, Loh et al. compared *CovSel* with other covariate selection methods, such as collaborative-TMLE and augmented backward elimination, to evaluate their ability to correctly select confounders and control the type I error rate after data-driven covariate selection. They found that *CovSel* selected at least one confounder each time and had an approximate 70% probability to select the sufficient confounders exactly. Additionally, *CovSel* approximately controlled the type I error empirically at the significance level [23]. Thus, we applied the algorithm *CovSel* to select the minimal conditioning set that was sufficient for unbiased effect estimations of the target gene on the BCR.

We accordingly obtained an adult PRAD patient dataset from TCGA project to identify the potential prognostic genes that were related to the TME and caused BCR at 3 years after prostatectomy. Moreover, the causal effects of these genes were estimated for individual therapies. We verified these prognostic genes using the Gene Expression Omnibus (GEO) database. We drew a workflow schematic to illustrate the entire study design for the identification of the prognostic biomarkers in PRAD (Figure 1).

RESULTS

Clinical and pathological characteristics of men with PRAD from TCGA

We obtained the gene expression profiles and clinical information of the PRAD patients with an initial pathologic diagnosis made between 2000 and 2013 from TCGA. A total of 209 patients were included after the exclusion of the subjects according to the exclusion criteria in Methods. Among them, 112 patients (53.6%) had BCR within 3 years after definitive local therapy, whereas 97 patients (46.4%) had no BCR within 3 years (Table 1). Based on the Estimation of STromal and Immune cells in MAlignant Tumour tissues using Expression data (ESTIMATE) algorithm [24], the stromal scores were obtained and distributed between -1,867.0 and 1,789.3, and the immune scores ranged from -1,796.65 to 2962.96. Patients with BCR within 3 years had lower immune and ESTIMATE scores than patients without BCR ($P < 0.05$). We identified radiation therapy, pathological Gleason score and tumor-node-metastasis (TNM) stage as significant risk factors for BCR ($P < 0.05$), which were selected as candidate covariates.

Comparison of gene expression profiles in PRAD according to the immune and stromal scores

To identify the differentially expressed genes (DEGs), we divided the 209 PRAD patients to high and low score groups according to their immune and stromal scores. A total of 112 patients (53.59%) had high stromal scores, and 115 (55.02%) had high immune scores. Based on immune scores, 515 gene expression levels were demonstrated to be increased and 49 genes were decreased in the high score group as compared to the low score group (Figure 2A). Similarly, for the high and low groups based on stromal scores, 882 genes were increased and 43 genes were decreased (Figure 2B). log2|fold change| > 1 and False discovery rate (FDR) < 0.05 were used as the criteria for screening DEGs. Moreover, the Venn diagrams (Figure 2C, 2D)

showed that 716 genes had high expressions in both high immune and stromal score groups, and 17 genes had low expressions. These overlapped DEGs (733 genes in total) might be the determinants of TME status. Thus, we decided to focus on these DEGs for all subsequent analyses.

Correlation between the expression of differentially expressed genes and biochemical recurrence

Results of the univariate logistic regression models showed that 68 DEGs were associated with BCR in PRAD ($P < 0.05$), and thus we selected them as candidate causative genes involved in immune and stromal cells (Figure 3).

Selection of confounding covariates for the minimal sets of confounding covariates

We considered the 68 associated genes along with 3 significant clinical covariates (radiation therapy, pathological Gleason score, and TNM stage) as the candidate confounding covariates. To reduce the dimension of confounding covariates for improved causal effect estimations, we selected the minimal sets of confounders between each candidate causative gene and BCR status using the R package, *CovSel*. W_G is a subset of the candidate confounding covariates that leads to other covariates after removing W_G conditionally independent of the outcome Y given W_G. W_G is illustrated in Supplementary material (Supplementary Figure 1). V_G is a subset of W_G after removing variables conditionally independent of the target gene G given V_G. Figure 4 shows the Pearson correlation coefficients (displayed in color from dark blue to red as correlations from -1 to 1) of the 68 candidate genes and the corresponding 70 candidate confounding covariates of each gene, as well as the selected confounding covariates V_G for each candidate gene (the dark dot of each column).

Selection of prognostic genes and causal effect estimations using targeted maximum likelihood estimation

To explore the prognostic genes of BCR, we estimated the causal effects (the average causal effect (ACE), the marginal odds ratio (MOR), and the individual causal effect (ICE)) of the 68 candidate genes on early-onset BCR using TMLE. Considering the complex network

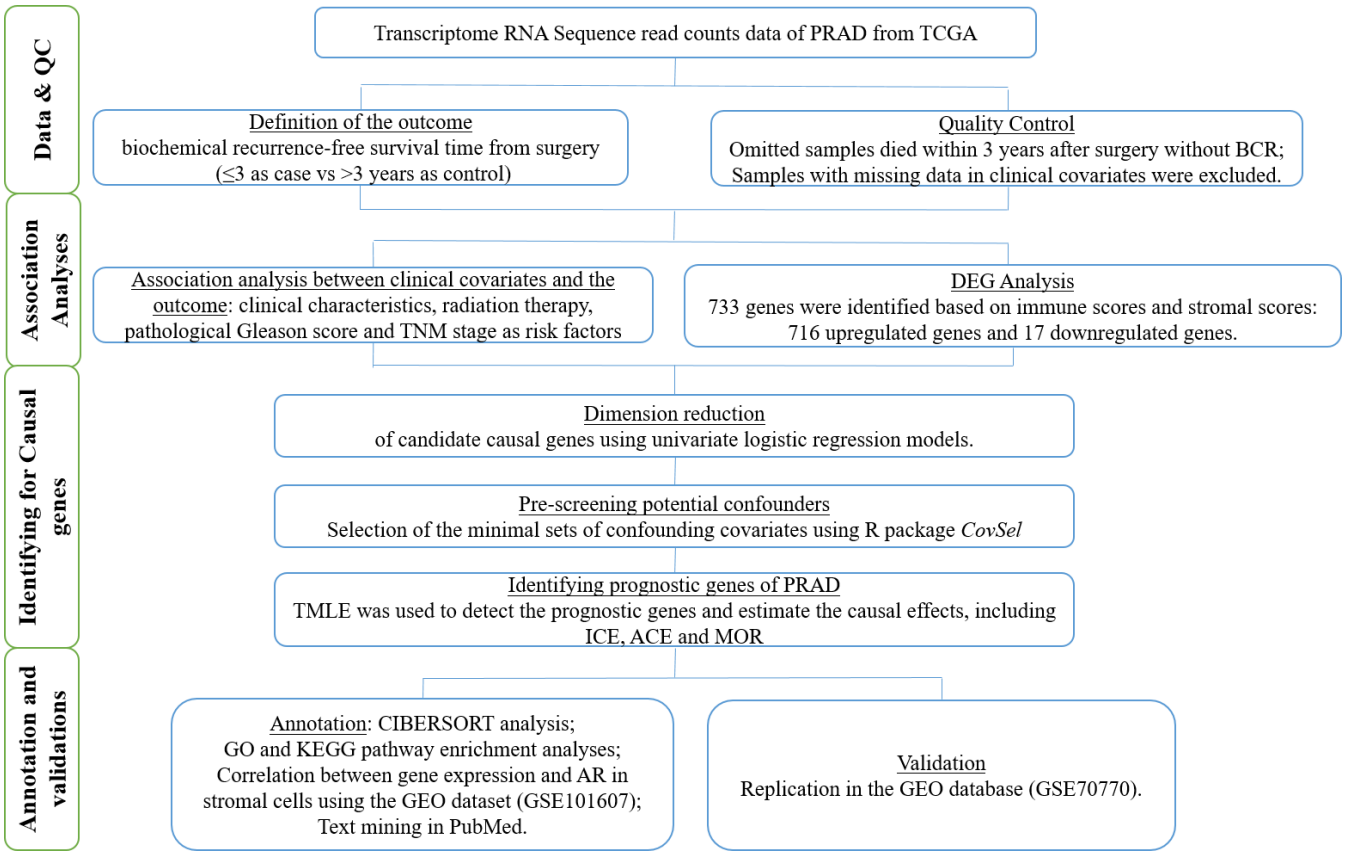


Figure 1. The workflow schematic for identifying prognostic biomarkers in PRAD.

Table 1. Clinical and pathological characteristics of 209 PARD patients from TCGA.

Clinic pathologic variable	BCR (n=112)	BCR-free (n=97)	Total (n=209)	P value
Age (years)	60.74 (6.84)	61.51 (6.09)	61.1 (6.50)	0.397
Stromal scores	-533.82 (459.27)	-402.32 (607.96)	-472.79 (536.16)	0.076
Immune scores	-471.57 (587.37)	-263.78 (789.91)	-375.14 (694.90)	0.03
ESTIMATE scores	-1,005.4 (930.04)	-666.11 (1,296.03)	-847.93 (1,124.99)	0.028
Weight	197.5 (181.27)	316.25 (430.36)	267.11 (353.25)	0.072
Radiation therapy				
Yes	11 (9.82)	22 (22.68)	33 (15.79)	0.019
No	101 (90.18)	75 (77.32)	176 (84.21)	
Gleason score				
6	6 (5.36)	1 (1.03)	7 (3.35)	
7	67 (59.82)	25 (25.77)	92 (44.02)	< 0.001
8~9	39 (34.82)	71 (73.20)	110 (52.63)	
TNM stage				
I	5 (4.46)	1 (1.03)	6 (2.87)	
II	37 (33.04)	9 (9.28)	46 (22.01)	< 0.001
III	54 (48.21)	59 (60.82)	113 (54.07)	
IV	16 (14.29)	28 (28.87)	44 (21.05)	

*Data are mean ± SD and frequency (percent) for numeric and category variables, respectively.

and interactions among these genes, we collaborated the Super Learner with TMLE and obtained a weighted causal effect of these models. Accordingly, we identified 14 prognostic genes (*TNFRSF4*, *ZAP70*, *ERMN*, *CXCL5*, *SPINK6*, *SLC6A18*, *CHRM2*, *TG*, *CLLU1OS*, *POSTN*, *CTSG*, *NETO1*, *CEACAM7*, and *IGLV3-22*) on 12 chromosomes with *P*-ACE < 0.05 (Table 2). 7 out of these genes (*ZAP70*, *SPINK6*,

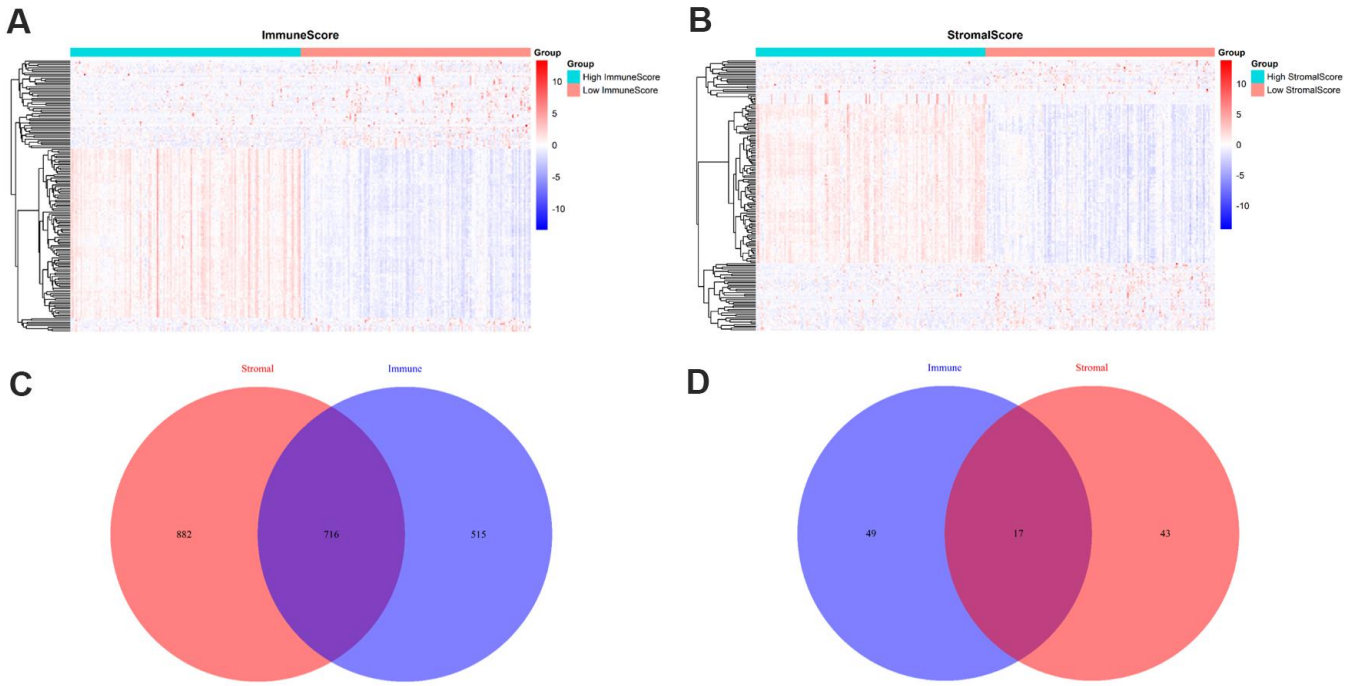


Figure 2. Comparison of gene expression profiles with immune and stromal scores. (**A**) A heat map of DEGs between the high and low immune score groups; (**B**) A heat map of DEGs between the high and low stromal score groups; (**C**) Venn diagrams showing the number of high expression and (**D**) low expression of DEGs in both immune and stromal score groups.

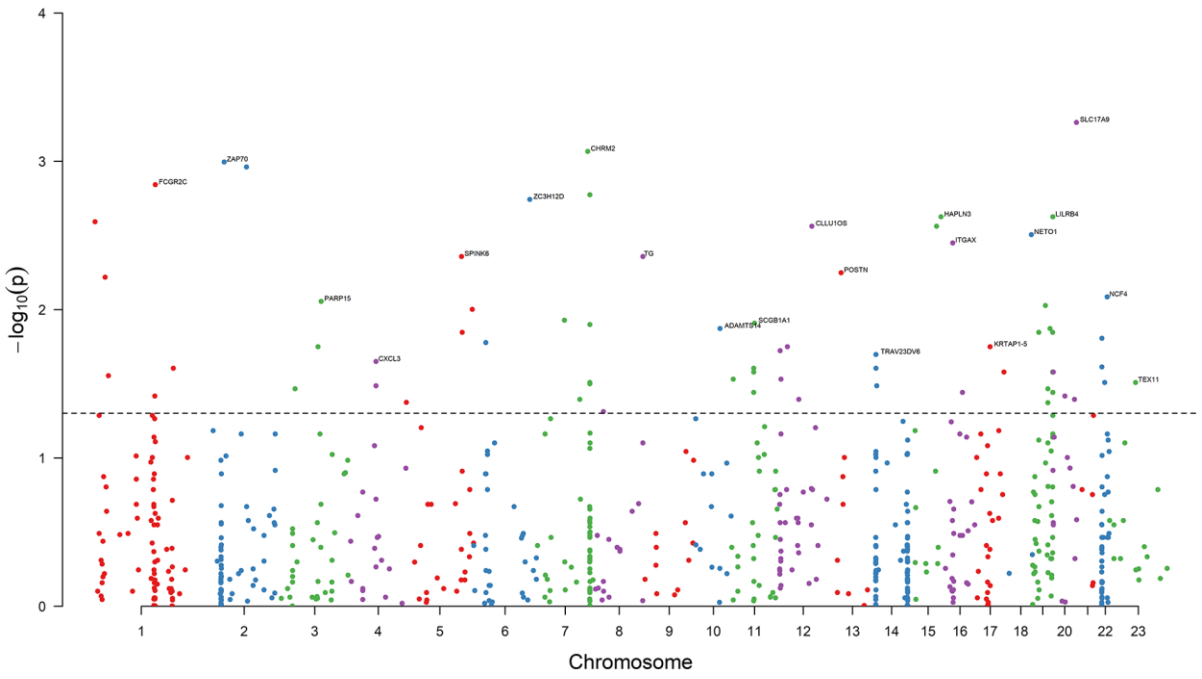


Figure 3. Results of association analysis. The dashed black line is the bound of $P = 0.05$. 68 DEGs are associated with the PRAD BCR ($P < 0.05$). The top genes with the minimum P value on each chromosome are annotated. Chromosome 23 is denoted Chromosome X.

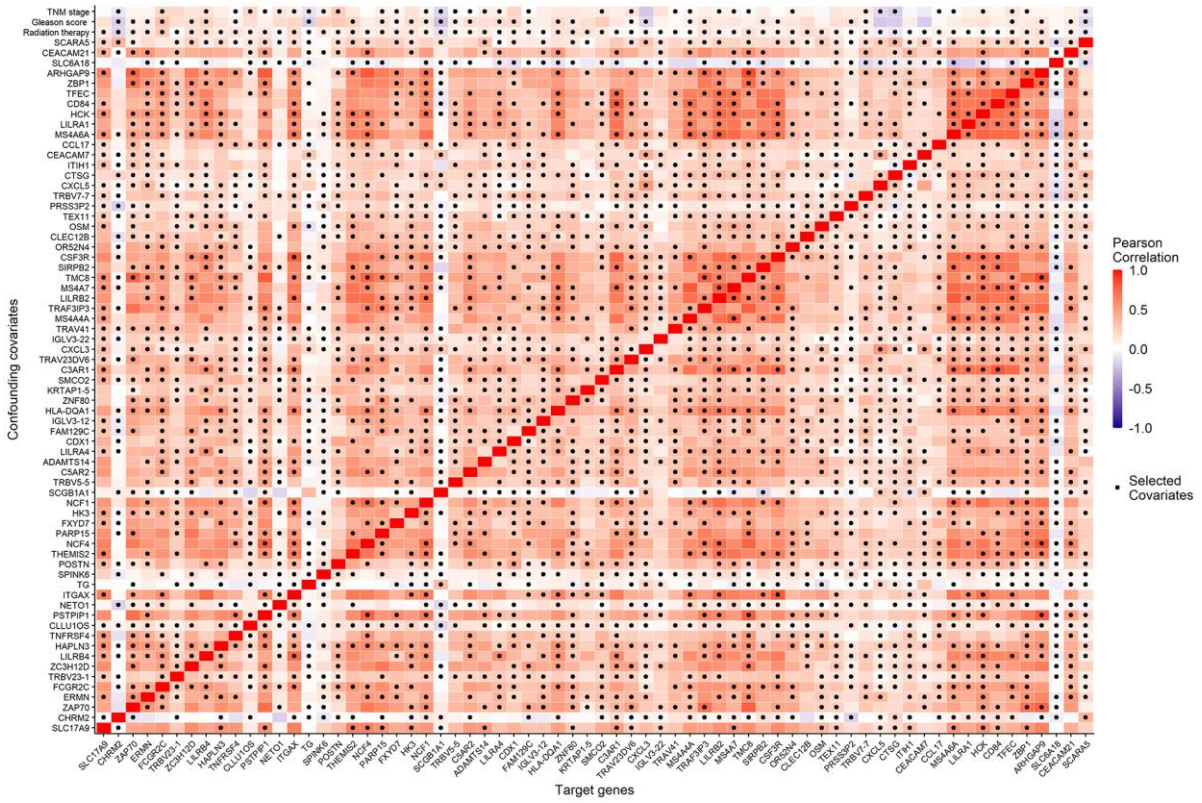


Figure 4. The Pearson correlation coefficients (the corresponding color) of the 68 candidate genes (horizontal axis) and the 70 candidate confounding covariates (vertical axis), as well as the minimal confounding covariate set of each candidate gene (the dark dot of each column).

Table 2. The prognostic genes of PRAD BCR and their corresponded causal effects.

Gene	Chromosome	ACE (95% CI)	*P*-ACE	MOR (95% CI)	*P*-MOR
TNFRSF4	1	0.092 (0.022, 0.163)	0.011	1.454 (1.089, 1.940)	0.011
ZAP70	2	0.116 (0.044, 0.188)	0.002	1.597 (1.191, 2.143)	0.002
ERMN	2	0.135 (0.05, 0.219)	0.002	1.728 (1.222, 2.444)	0.002
CXCL5	4	-0.142 (-0.226, -0.059)	0.001	0.562 (0.399, 0.792)	0.001
SPINK6	5	0.153 (0.047, 0.26)	0.005	1.865 (1.201, 2.897)	0.005
SLC6A18	5	0.165 (0.017, 0.313)	0.029	1.944 (1.056, 3.578)	0.033
CHRM2	7	-0.173 (-0.266, -0.079)	<0.001	0.496 (0.337, 0.731)	<0.001
TG	8	-0.127 (-0.243, -0.012)	0.031	0.598 (0.372, 0.960)	0.033
CLLU1OS	12	0.13 (0.053, 0.207)	0.001	1.689 (1.233, 2.312)	0.001
POSTN	13	0.083 (0.009, 0.156)	0.027	1.398 (1.037, 1.885)	0.028
CTSG	14	-0.133 (-0.229, -0.038)	0.006	0.584 (0.395, 0.863)	0.007
NETO1	18	0.137 (0.023, 0.252)	0.019	1.740 (1.090, 2.778)	0.020
CEACAM7	19	-0.165 (-0.289, -0.042)	0.009	0.511 (0.307, 0.851)	0.010
IGLV3-22	22	0.166 (0.004, 0.328)	0.044	1.956 (1.001, 3.823)	0.050

*CI: confidence interval.

CHRM2, *TG*, *CLLU1OS*, *POSTN*, and *NETO1*) were the top associated genes on 7 chromosomes. Among these prognostic genes, 9 genes (*TNFRSF4*, *ZAP70*, *ERMN*, *SPINK6*, *SLC6A18*, *CLLU1OS*, *POSTN*, *NETO1*, and *IGLV3-22*) were unfavorable prognostic genes, and 5 favorable prognostic genes (*CXCL5*, *CHRM2*, *TG*, *CTSG*, and *CEACAM7*).

Figure 5 shows the individual treatment effects of each prognostic gene on BCR. The unfavorable genes were a

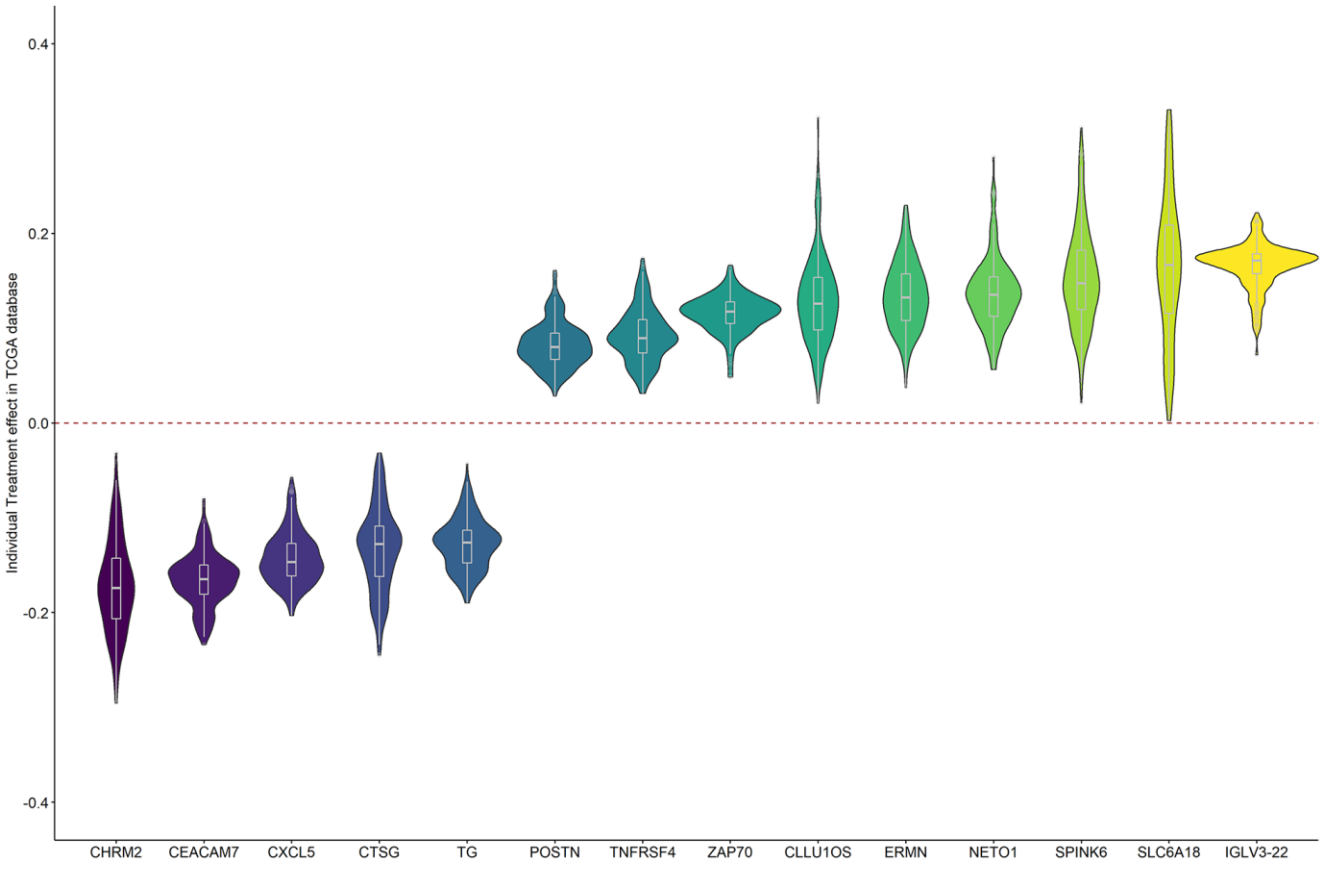


Figure 5. The individual causal effects of the 14 prognostic genes on PRAD BCR.

risk for all patients, but with a range of positive causal effects. Similar results were obtained for the favorable genes.

Functional analysis of the prognostic genes

Figure 6 illustrates the estimated proportion of tumor-infiltrating immune components in the PRAD samples using the Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT) analysis [25]. Among these immune cell profiles, CD4 memory resting T cells occupied the largest proportion in the PRAD samples. The Wilcoxon rank-sum test showed that only two types of tumor-infiltrating immune cells (plasma cells and follicular helper T cells) were not associated with the expression levels of the 14 prognostic genes.

Among the 14 identified prognostic genes, 6 genes (*TNFRSF4*, *ZAP70*, *CXCL5*, *CHRM2*, *CTSG*, and *IGLV3-22*) were reported to be associated with promoting antitumor immunity in previously published papers, 4 genes (*SPINK6*, *POSTN*, *CLLU1OS*, and *CEACAM7*) were associated with the metastasis of tumor cells, and 4 genes (*ERMN*, *SLC6A18*, *TG*, and *NETO1*) were associated with other diseases.

Gene ontology term and Kyoto encyclopedia of genes and genomes pathway analysis of the prognostic genes

To outline the potential functions of the prognostic genes, we performed a functional enrichment analysis of the 14 prognostic genes. The results of gene ontology (GO) term enrichment analysis suggested strong correlations of these genes with immune responses (Figure 7). Kyoto encyclopedia of genes and genomes (KEGG) pathway analysis revealed the significant enrichment of four pathways. In particular, we detected that *ZAP70* was enriched in primary immunodeficiency, *TNFRSF4*, and *CXCL5* were enriched in cytokine-cytokine receptor interaction, *CTSG* was enriched in the renin-angiotensin system, and the *CHRM2* and *CTSG* were enriched in neuroactive ligand-receptor interaction (Figure 8). These results further illustrated that the two

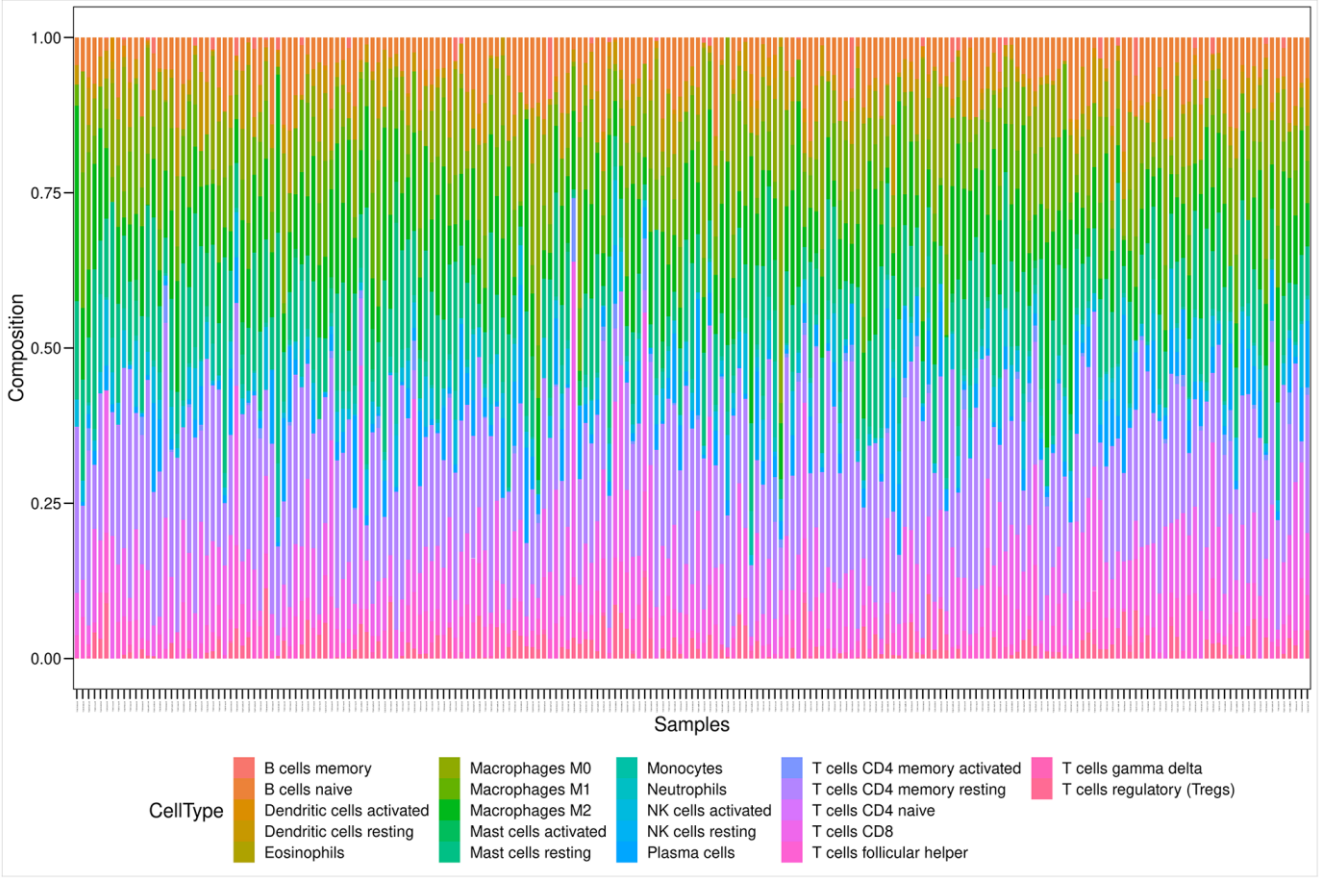


Figure 6. Bar plot showing the proportion of 22 kinds of TIC profiles in PRAD tumor samples. Rows and columns represent immune cell compositions and samples, respectively.

pathways derived from the KEGG analysis were associated with immune responses.

Analysis of the correlations between prognostic genes and androgen receptor in stromal cells

To explore the correlations between AR and the identified genes, we compared the expression of genes in the AR- and non-AR-driven groups using the GEO dataset (GSE101607). Our results showed that *POSTN* was significantly overexpressed in non-AR compared with AR-driven samples (Figure 9A), consistent with the findings of Cattrini et al. [26]. *NETO1* was overexpressed in AR-driven samples (Figure 9B). Although we did not detect any association between the expression of *CXCL5* and AR (Figure 9C), AR signaling has been reported to promote PRAD progression via modulation the AKT-NF-κB-*CXCL5* signaling [10].

Validation in the GEO database

We further verified the 14 prognostic genes in an additional PRAD cohort obtained from the GEO database. We downloaded and analyzed the gene expression data of 203 PRAD cases in the GSE70770 dataset. A total of five genes (*ZAP70*, *CXCL5*, *SPINK6*, *CHRM2*, and *TG*) were significantly associated with early-onset BCR in PRAD (Table 3). The results of individual causal effects indicated that *CHRM2* and *SPINK6* might have different functions in different individuals with different features (Supplementary Figure 2).

DISCUSSION

The aim of this study was to identify TME-related biomarkers, implicated in the development of BCR after prostatectomy, using the semi-parametric targeted approach, TMLE. TME is known to be comprised of a complex mixture of tumor-associated fibroblasts, infiltrating immune cells, endothelial cells, extracellular matrix proteins, and signaling molecules, such as cytokines [27–29]. Both immune and stromal cells have been proposed to be valuable for tumor diagnosis and prognosis evaluation. Similar to many other solid tumor types, prostate cancer is characterized by a rich tumor-

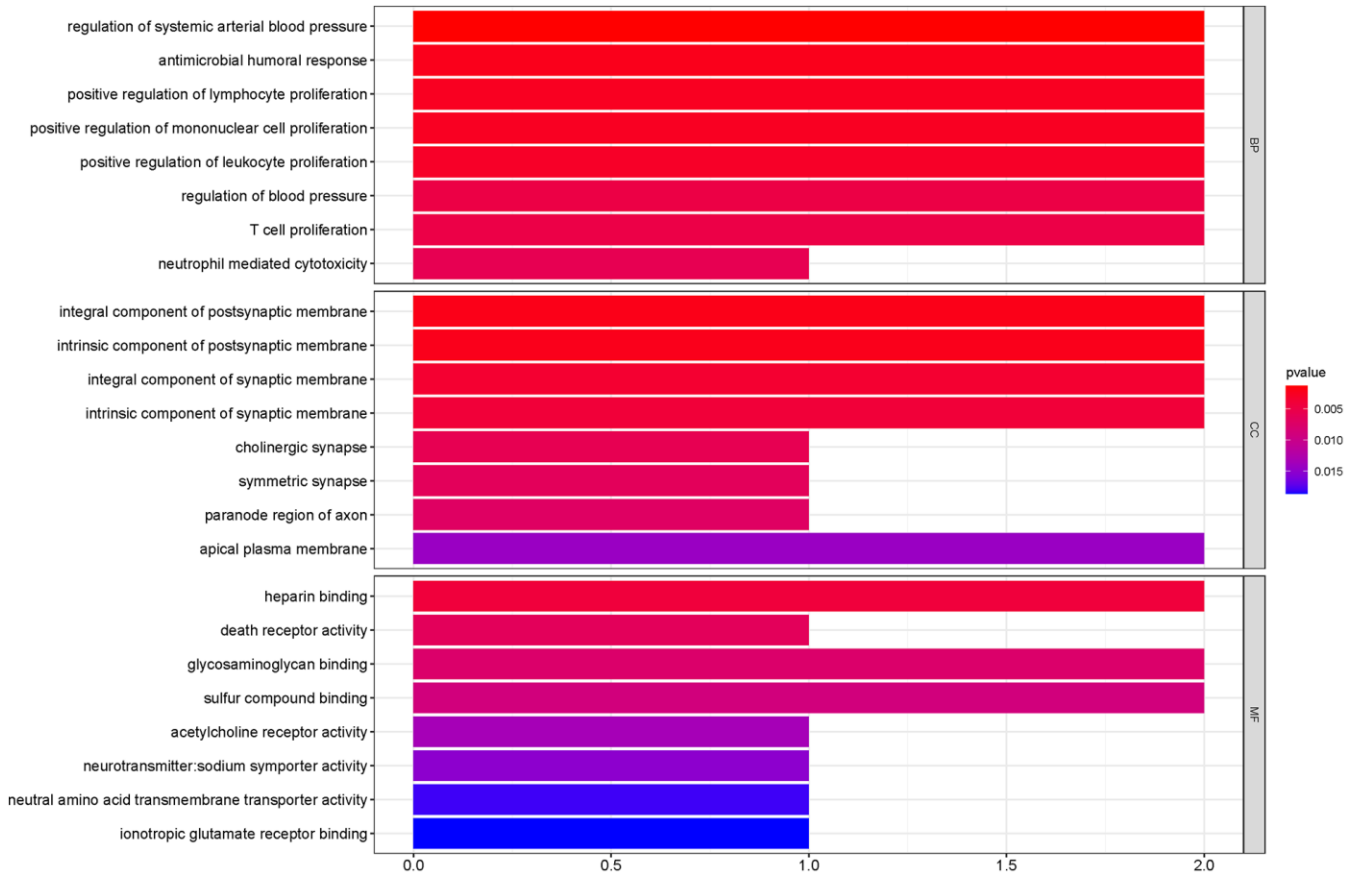


Figure 7. GO term enrichment analyses of the prognostic genes. The main GO terms (*P* values < 0.05) are shown for Biological process (BP), Cellular component (CC) and Molecular function (MF) respectively.

stroma interaction network that forms the TME [27–29]. Our results also revealed that the TME (immune and stromal scores) was associated with early-onset BCR of patients with PRAD.

In addition, BCR serves as an indicator of the early stages of relapse, as local recurrence and distant metastasis might occur after BCR [30]. Furthermore, as BCR within 3 years of surgery is a critical node for

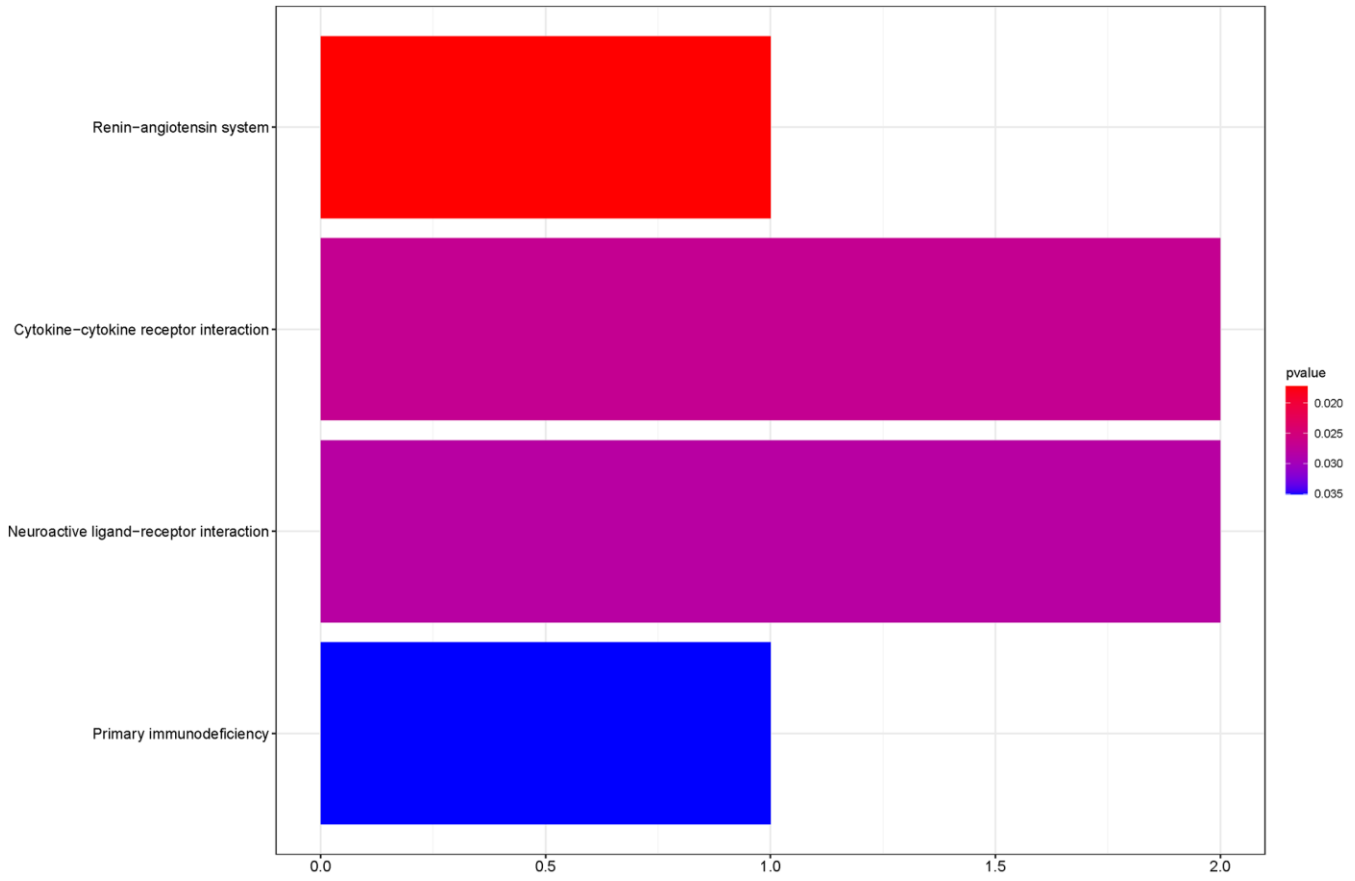


Figure 8. KEGG pathway analyses of the prognostic genes.

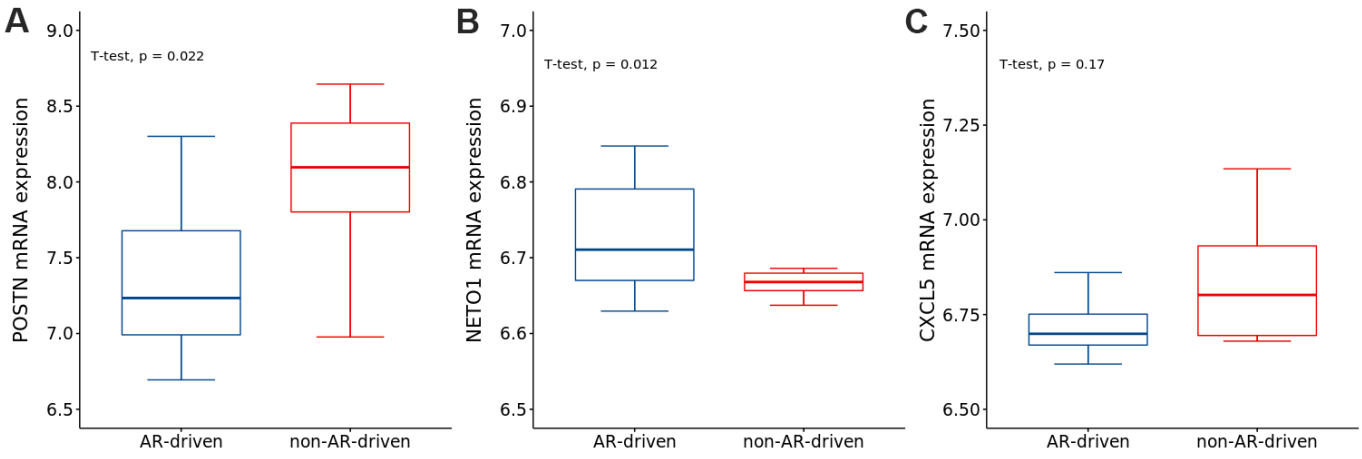


Figure 9. The expressions of (**A**) *POSTN*, (**B**) *NETO1* and (**C**) *CXCL5* in AR- and non-AR-driven groups using GSE101607. *T*-test was used to measure the difference between the two groups.

Table 3. The validated genes in GEO database and their corresponded causal effects.

Gene	Chromosome	ACE (95% CI)	*P*-ACE	MOR (95% CI)	*P*-MOR
ZAP70	2	0.288 (0.226, 0.35)	<0.001	4.099 (2.895, 5.805)	<0.001
CXCL5	4	-0.14 (-0.256, -0.023)	0.018	0.425 (0.226, 0.798)	0.008
SPINK6	5	0.364 (0.294, 0.434)	<0.001	5.797 (3.881, 8.658)	<0.001
CHRM2	7	-0.211 (-0.271, -0.151)	<0.001	0.34 (0.247, 0.469)	<0.001
TG	8	-0.221 (-0.325, -0.116)	<0.001	0.199 (0.094, 0.423)	<0.001

prostate-specific mortality [3], it is necessary to identify the risk factors for 3-year BCR in the TME.

Using the traditional approaches, such as linear or logistic regression models, confounding factors and complex associations among covariates might bias the results and lead to fallacious conclusions. Whereas, robust TMLE was demonstrated to help reduce the risk of spurious findings [17]. Although TMLE optimizes the bias-variance tradeoff for the estimated causal effects, a rough trend could still be observed for the individual effects of patients. Based on this strategy, we identified 14 genes involved in the prognosis of PRAD.

TNFRSF4 (also known as *OX40* or *CD134*) is a member of the tumor necrosis factor receptor superfamily, subserving co-stimulatory functions of T-cells during infection [31–33]. It is predominantly and transiently expressed by both human CD4+ and CD8+ T cells [32]. Studies have shown that regulatory T cells express more *TNFRSF4* than conventional CD4+ T cells in multiple human tumors [34]. Several anti-*TNFRSF4* agonistic monoclonal antibodies are currently being tested in early-phase cancer clinical trials [31]. The expression of *TNFRSF4* on tumor-infiltrating lymphocytes (TILs) has been studied in different tumor types, such as breast cancer, melanoma, B-cell lymphoma and head and neck cancers [35–40]. In colon cancer, the high expression of *TNFRSF4* in TILs, mesenteric lymph nodes, or invasive margin lymphoid aggregates was reported to correlate with better overall survival [40]. In our study, *TNFRSF4* was expressed at high levels in the BCR group, indicating that *TNFRSF4* could be a marker of BCR status in PRAD.

ZAP70, a 70 kDa tyrosine kinase of the Syk family, has been reported to significantly promote tumor angiogenesis and immunosuppression in cancer cell lines or samples [41–43]. Richardson et al. reported that *ZAP70* was activated in response to migratory and survival signals in B-cell chronic lymphocytic leukemia [44]. Fu et al. found that *ZAP70* was overexpressed in prostate cancer cell lines and tissues, facilitating prostate cancer cell migration and invasion [45]. MiR-631 was shown to target the 3′-UTR of *ZAP70* mRNA and inhibit the expression of *ZAP70*, thereby inhibiting prostate cancer cell migration and invasion [45]. Our results showed that *ZAP70* was mapped to the primary immunodeficiency pathway in KEGG and expressed at high levels in patients with BCR. Combined with previous studies, we speculated that *ZAP70* might not only be an important regulator of cancer metastasis but also useful in predicting the BCR in patients with prostate cancer.

CXCL5 is a proangiogenic CXC-type chemokine known to act as an inflammatory mediator and a powerful attractant for granulocytic immune cells. It is secreted by both immune (neutrophils, monocytes, and macrophages) and nonimmune (epithelial, endothelial, and fibroblastic) cell types [46]. Wang et al. pointed out that CXCL5 was a cancer-secreted chemokine that attracted *CXCR2*-expressing myeloid-derived suppressor cells (MDSCs) and, correspondingly, pharmacological inhibition of CXCR2 impeded tumor progression [47]. Biological, molecular and pharmacological analyses established that a Yap1-mediated CXCL5-CXCR2 signaling axis recruits MDSCs into the TME [47]. Moreover, AR signaling was shown to promote the progression of PRAD via modulation of the AKT-NF-κB-*CXCL5* signaling. Our results showed that *CXCL5* was a prognostic gene, suggesting that the inflammatory mediator, *CXCL5*, might be a potentially protective prognostic factor for prostate cancer.

CHRM2, which mediates various cellular responses, has been demonstrated to be a significant marker of cognitive flexibility [48] and a potential therapeutic target for gastric cancer [49]. We found that *CHRM2* was a significant gene of the 3-year PRAD BCR. Validation using the GEO dataset confirmed our findings. We speculated that *CHRM2* might be a prognostic factor for prostate cancer.

IGLV3-22, which is a membrane-bound or secreted glycoprotein produced by B lymphocytes, belongs to the immunoglobulin lambda variable 3 (IGLV3) family. B lymphocytes are important cell types involved in the immune response of mammals [50]. Reports have shown that 40% of tumor-infiltrating lymphocytes in some patients with breast cancer were B cells [51, 52], suggesting the critical roles of these cells in modulating

tumor responses [50]. Likewise, *IGLV3-21*, an important paralog of *IGLV3-22*, was confirmed to be a risk factor for chronic lymphocytic leukemia [53–55]. Our results showed that *IGLV3-22* was a causative gene of PRAD and indicated that patients with higher expression of *IGLV3-22* might have a worse prognosis.

CTSG is a serine protease of the chymotrypsin family that is stored in the primary (azurophil) granules of polymorphonuclear neutrophils [56]. Proteases are known to increase peripheral and central inflammation by regulating the chemotaxis of immune cells and the production of cytokines and chemokines. Previous studies have suggested *CTSG* as a potential marker of chronic pain after surgery [57], granulopoiesis or leukemogenesis [58]. In this study, we found that *CTSG* was a protective causative gene of BCR in patients with PRAD. Targeting and suppression of *CTSG* was shown to potentially inhibit the antitumor immunity.

POSTN is a 90-kDa extracellular matrix protein that interacts with multiple integrins to coordinate a variety of cellular processes, including epithelial-to-mesenchymal transition and cell migration [59, 60]. Stromal *POSTN* has been shown to participate in the regulation of cancer stem cell maintenance and expansion during metastatic colonization [61]. Researchers have reported that *POSTN* functions as a progression-associated and prognostic biomarker in glioma via the induction of invasive and proliferative phenotypes [62]. In our study, we found that *POSTN* was an unfavorable biomarker of PRAD BCR.

CLLU1OS is located on chromosome 12q22. The 12q21.33-12q22 region is dense with the expressed sequence tags derived from the germinal center B cells and CLL cells and is highly accessible for transcription in the B cells [63]. Accordingly, we detected a high expression of *CLLU1OS* in the BCR group, indicating that *CLLU1OS* was a risk factor for prostate cancer.

SPINK6, which is overexpressed in tumors and highly metastatic nasopharyngeal carcinoma cells, has been reported as an independent unfavorable prognostic factor [64]. It has been reported to act as a functional regulator of nasopharyngeal carcinoma metastasis via the EGFR signaling. Our results showed that SPINK6 was a risk gene for 3-year BCR in patients with PRAD.

CEACAM7 is a human cellular adhesion protein that belongs to the immunoglobulin superfamily. It has been reported to have low expression in colorectal cancers [65]. Our results showed low expression of *CEACAM7* in the BCR group, suggesting a putative role in the initiation and progression of prostate cancer.

ERMN is an essential gene involved in cytoskeletal rearrangements during myelinogenesis [66], and acts as a primary target of the disrupted folate metabolism [67]. *SLC6A18* is a specific transporter for neurotransmitters, amino acids, and osmolytes such as betaine, taurine, and creatine [68]. *TG* expresses the protein precursor of thyroid hormones, which are essential for the growth, development, and control of metabolism in vertebrates [69, 70]. *NETO1* has been found to be abnormally expressed in human carcinomas [71]. Higher expression of *NETO1* in epithelial ovarian cancer tissue samples has been reported to lead to worse overall survival and a higher probability of bowel metastases [71, 72]. The associations between these four genes and prostate cancer warrant further investigations.

The interaction between PRAD and TME might have serious effects on tumor evolution, further influencing tumor resistance, recurrence, and overall prognosis. Wang et al. provided a detailed description of the mechanism by which the activation of tumor-inherent genes altered TME [62]. The present study focused on the genetic characteristics of the TME, stimulating the development of PRAD. Our results might provide a basis for further studies on the role of TME in PRAD. However, the underlying mechanism remains unclear. Eventually, the putative role of these genes in the prognosis of prostate cancer would require further evaluation in future studies.

In summary, we used TCGA dataset to identify the potential prognostic genes in PRAD. Using the associations between the immune/stromal scores and the prognosis of PRAD, we revealed a set of genes related to the TME and the BCR of PRAD. These findings might facilitate the prognosis of PRAD. Based on our study, previously neglected genes could be used as biomarkers for PRAD. Finally, further studies of these genes could provide a more comprehensive understanding of the potential relationship between the prognosis of PRAD and the TME.

MATERIALS AND METHODS

Data collection

The transcriptome RNA sequence read count data of 550 patients with PRAD and their corresponding clinical profiles were obtained from TCGA (https://portal.gdc.cancer.gov). We collected data from patients with primary prostate tumors and defined the outcome by BCR-free survival time from surgery (≤3 as case vs. >3 years as control). Patients who died within 3 years after surgery without BCR were excluded.

Clinical characteristics of the patients including data of age at initial pathologic diagnosis, pathological Gleason score, clinical TNM stage, and radiation therapy were also collected. TNM stage for each individual was classified according to the eighth edition of the American Joint Committee on Cancer TNM staging manual. Samples with missing clinical information were excluded from this study.

For validation, an additional PRAD cohort of 293 patients was obtained with the accession number GSE70770 from the GEO database (https://www.ncbi.nlm.nih.gov/gds). Gene expression data and clinical data were also downloaded. The same exclusion criteria as those used in TCGA were applied to this cohort.

Construction of tumor microenvironment

The immune and stromal components in TME were calculated based on the ESTIMATE algorithm [24] using the R package estimate. Immune, stromal, and ESTIMATE scores, corresponding to the levels of immune cells, stromal cells, and the sum of both, respectively, were obtained. These three scores were applied to assess the infiltration level of immune and stromal cells and tumor purity in tumor tissues.

Identification of differentially expressed genes based on immune scores and stromal scores

All patients were classified into two groups based on the mediation of immune and stromal scores to explore the correlation between gene expression profiles and immune or stromal scores. We used the R package *DESeq2* to perform differentiation analysis of gene expression, and DEGs were generated by comparing the high and low score groups. DEGs with log2|fold change| > 1 and adjusted P value < 0.05, after Benjamini–Hochberg false discovery rate [73], were considered significant.

Statistical analysis of associated genes and clinical covariates

The univariate logistic regression models were used to determine the DEGs to be considered as candidate causative genes of BCR involved in the immune and stromal cells. According to its median expression level, each gene in tumor samples was grouped into high- or low-expression groups. Significant clinical covariates for BCR status, which were selected as candidate covariates in subsequent analysis, were examined using *t*-tests (age at initial pathologic diagnosis and weight) and χ^2tests (radiation therapy of patients, pathological Gleason score and clinical TNM stage).

Selection of the minimal sets of confounding covariates

To select the minimal sets of confounding covariates between candidate causative genes and the BCR status, the R package *CovSel* was used to screen all candidate causative genes and clinical covariates. $G = \{g_1, g_2, \ldots, g_n\}$ denoted the binary candidate causative genes, $X = \{x_1, x_2, \ldots, x_m\}$ the selected clinical covariates, and Y the outcome 3-year BCR in patients with PRAD. In addition, $\{(G\backslash g_i) + X\}$ denoted the complete covariate vector of the target gene g_i and Y. We assumed that W_G was the subset of confounding covariates $\{(G\backslash g_i) + X\}$, which satisfied $Y \perp \{(G\backslash g_i) + X\}\backslash W_G | W_G$. V_G was the minimal set of confounding covariates W_G satisfying $G \perp W_G\backslash V_G | V_G$, and the final confounding set of the target gene g_i and Y.

Estimation of the causal effects of genes on biochemical recurrence

Causal effects are commonly defined in potential outcomes [74], that is, the average causal effect (ACE), defined as $E[Y(1) - Y(0)]$, the marginal odds ratio (MOR), defined as $\{E[Y(1)] \times E[1 - Y(0)]\} / \{E[1 - Y(1)] \times E[Y(0)]\}$, and the individual causal effect (ICE), defined as $Y(1) - Y(0)$, where $Y(1)$ is the outcome under exposure (G =1) and $Y(0)$ is the outcome when unexposed (G = 0).

To causally interpret the causal effects, we put forward the following assumptions. i) Stable-unit-treatment-value assumption: the potential outcomes of a given individual will not be affected by their exposure status. ii) No unmeasured confounders: all the covariates altering the exposure and outcome are measured, formulated as $(Y(1), Y(0)) \perp G | V_G$. iii) Positivity: every individual has a non-zero probability conditioning on the covariates within strata of G, which could be formulated as $0 < P(G = 1|V_G) < 1$. Thus, the ACE could be defined as $E_V [E(Y|G = 1, V_G) - E(Y|G = 0, V_G)]$ with the observed dataset.

In this study, TMLE was used to detect the causative genes and estimate the causal effects, including ACE, MOR, and ICE. TMLE used two steps to target the optimal bias-variance tradeoff and obtain the target parameters. First, we conditionally estimated the expectation of the outcome with both exposures and confounders, $E(Y \mid G, V_G)$, which was used to predict the potential outcomes. Second, we evaluated the exposure mechanism $P(G = 1 \mid V_G)$ to update the estimation of $E(Y \mid G, V_G)$. We eventually used the final updated estimation of $E(Y \mid G, V_G)$ to predict a pair of potential outcomes for each individual, and calculate both ACE and ICE. The ICE was calculated as the

difference between these pairs of each individual, whereas ACE was the average difference of ICE. Then, we calculated MOR as the ratio of the two odds of BCR occurring in the high- and low-expression groups. Combining TMLE with Super Learner, we selected five models (logistic regression model with or without interaction, elastic net regression, BART, and random forest, using the R functions *glm*, *glm interaction*, *glmnet*, *bartMachine*, and *randomforest*, respectively) to build both outcome and propensity score models to improve the robustness and precision of our estimates.

Functional annotation and analysis

To elucidate the biological functions of the prognostic genes with the immune microenvironment, we performed CIBERSORT analyses to estimate the proportion of tumor-infiltrating immune components in PRAD samples. The Wilcoxon rank-sum test was used to determine the association between the expression level of the prognostic genes and 22 types of immune cell profiles. Statistical significance was set at $P < 0.05$.

We performed GO and KEGG pathway enrichment analyses to investigate the shared biological functions among the identified genes that were common among the high and low immune/stromal score groups. The enrichment analyses were performed using the R packages *clusterProfiler*, *enrichplot*, and *ggplot2*. Only terms with adjusted $P < 0.05$ by FDR were considered as significantly enriched. To explore the correlations between AR and the identified genes, we compared the expression of genes in AR- and non-AR-driven groups using the GEO dataset (GSE101607).

All statistical analyses were performed using the software R 3.6.2 from CRAN (http://cran.r-project.org/). *P* values were corrected using the Benjamini–Hochberg method to control the FDR for multiple testing when appropriate [73].

Abbreviations

PRAD: prostate adenocarcinoma; TCGA: the Cancer Genome Atlas; GEO: Gene Expression Omnibus; GO: gene ontology; KEGG: the Kyoto Encyclopedia of Genes and Genomes; BCR: biochemical recurrence; TME: tumor microenvironment; DEG: differentially expressed gene; TMLE: targeted maximum likelihood estimation; AR: androgen receptor; FDR: false discovery rate.

AUTHOR CONTRIBUTIONS

F.X. and X.S. jointly conceived the idea and designed the study. X.S. was involved in the data collection and analysis, as well as the major contributor in writing the manuscript. L.W. aided in the collation of data used in this study. C.J. helped with the functional analysis in this study. H.L., Y.Y., L.H., X.L., Y.Y. and R.Y. were contributed to discussion, reviewed and edited the manuscript. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

The authors gratefully appreciate the dataset released by the TCGA project and GEO database. We would like to thank Editage (https://www.editage.com/) for English language editing.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant 81773547 and 82003557), Shandong Provincial Natural Science Foundation of China (ZR2019ZD02) and Shandong Provincial Key Research and Development project (2018CXGC1210).

REFERENCES

1. Kumar S, Singh R, Malik S, Manne U, Mishra M. Prostate cancer health disparities: An immuno-biological perspective. Cancer Lett. 2018; 414:153–65. https://doi.org/10.1016/j.canlet.2017.11.011 PMID:29154974

2. Grignon DJ. Unusual subtypes of prostate cancer. Mod Pathol. 2004; 17:316–27. https://doi.org/10.1038/modpathol.3800052 PMID:14976541

3. Freedland SJ, Humphreys EB, Mangold LA, Eisenberger M, Dorey FJ, Walsh PC, Partin AW. Risk of prostate cancer-specific mortality following biochemical recurrence after radical prostatectomy. JAMA. 2005; 294:433–39. https://doi.org/10.1001/jama.294.4.433 PMID:16046649

4. Han M, Partin AW, Pound CR, Epstein JI, Walsh PC. Long-term biochemical disease-free and cancer-specific survival following anatomic radical retropubic prostatectomy. The 15-year Johns Hopkins experience. Urol Clin North Am. 2001; 28:555–65. https://doi.org/10.1016/s0094-0143(05)70163-4 PMID:11590814

5. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. Nat Med. 2013; 19:1423–37. https://doi.org/10.1038/nm.3394 PMID:24202395

6. Jia D, Li S, Li D, Xue H, Yang D, Liu Y. Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. Aging (Albany NY). 2018; 10: 592–605. https://doi.org/10.18632/aging.101415 PMID:29676997

7. Zhao X, Hu D, Li J, Zhao G, Tang W, Cheng H. Database Mining of Genes of Prognostic Value for the Prostate Adenocarcinoma Microenvironment Using the Cancer Gene Atlas. Biomed Res Int. 2020; 2020:5019793. https://doi.org/10.1155/2020/5019793 PMID:32509861

8. Joyce JA, Pollard JW. Microenvironmental regulation of metastasis. Nat Rev Cancer. 2009; 9:239–52. https://doi.org/10.1038/nrc2618 PMID:19279573

9. Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, deCarvalho AC, Lyu S, Li P, Li Y, Barthel F, Cho HJ, Lin YH, et al. Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. Cancer Cell. 2017; 32:42–56.e6. https://doi.org/10.1016/j.ccell.2017.06.003 PMID:28697342

10. Guan Z, Li C, Fan J, He D, Li L. Androgen receptor (AR) signaling promotes RCC progression via increased endothelial cell proliferation and recruitment by modulating AKT → NF-κB → CXCL5 signaling. Sci Rep. 2016; 6:37085. https://doi.org/10.1038/srep37085 PMID:27848972

11. Heinlein CA, Chang C. Androgen receptor in prostate cancer. Endocr Rev. 2004; 25:276–308. https://doi.org/10.1210/er.2002-0032 PMID:15082523

12. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol. 1986; 15:413–19. https://doi.org/10.1093/ije/15.3.413 PMID:3771081

13. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999; 10:37–48. https://doi.org/10.1097/00001648-199901000-00008 PMID:9888278

14. Haukoos JS, Lewis RJ. The Propensity Score. JAMA. 2015; 314:1637–38. https://doi.org/10.1001/jama.2015.13480 PMID:26501539

15. Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017; 46:756–62. https://doi.org/10.1093/ije/dyw323 PMID:28039382

16. van der Laan MJ, Rubin D. Targeted Maximum Likelihood Learning. Int J Biostat. 2006; 2:11. https://doi.org/10.2202/1557-4679.104

17. Schuler MS, Rose S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. Am J Epidemiol. 2017; 185:65–73. https://doi.org/10.1093/aje/kww165 PMID:27941068

18. Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. Stat Med. 2018; 37:2530–46. https://doi.org/10.1002/sim.7628 PMID:29687470

19. Rubin DB. Estimating causal effects from large data sets using propensity scores. Ann Intern Med. 1997; 127:757–63. https://doi.org/10.7326/0003-4819-127-8_part_2-199710151-00064 PMID:9382394

20. De Luna X, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect. Biometrika. 2011; 98:861–75. https://doi.org/10.1093/biomet/asr041

21. Yu Y, Li H, Sun X, Su P, Wang T, Liu Y, Yuan Z, Liu Y, Xue F. The alarming problems of confounding equivalence using logistic regression models in the perspective of causal diagrams. BMC Med Res Methodol. 2017; 17:177. https://doi.org/10.1186/s12874-017-0449-7 PMID:29281984

22. Häggström J, Persson E, Waernbaum I, de Luna X. CovSel: An R Package for Covariate Selection When Estimating Average Causal Effects. J Stat Softw. 2015; 68:68. https://doi.org/10.18637/jss.v068.i01

23. Loh WW, Vansteelandt S. Confounder selection strategies targeting stable treatment effect estimators. Stat Med. 2021; 40:607–30. https://doi.org/10.1002/sim.8792 PMID:33150645

24. Yoshihara K, Shahmoradgoli M, Martínez E, Vegesna R, Kim H, Torres-Garcia W, Treviño V, Shen H, Laird PW, Levine DA, Carter SL, Getz G, Stemke-Hale K, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013; 4:2612. https://doi.org/10.1038/ncomms3612 PMID:24113773

25. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015; 12:453–57.

https://doi.org/10.1038/nmeth.3337
PMID:25822800

26. Cattrini C, Barboro P, Rubagotti A, Zinoli L, Zanardi E, Capaia M, Boccardo F. Integrative Analysis of Periostin in Primary and Advanced Prostate Cancer. Transl Oncol. 2020; 13:100789. https://doi.org/10.1016/j.tranon.2020.100789 PMID:32416542

27. Karlou M, Tzelepi V, Efstathiou E. Therapeutic targeting of the prostate cancer microenvironment. Nat Rev Urol. 2010; 7:494–509. https://doi.org/10.1038/nrurol.2010.134 PMID:20818327

28. Junttila MR, de Sauvage FJ. Influence of tumour microenvironment heterogeneity on therapeutic response. Nature. 2013; 501:346–54. https://doi.org/10.1038/nature12626 PMID:24048067

29. Hanahan D, Coussens LM. Accessories to the crime: functions of cells recruited to the tumor microenvironment. Cancer Cell. 2012; 21:309–22. https://doi.org/10.1016/j.ccr.2012.02.022 PMID:22439926

30. Shen MM, Abate-Shen C. Molecular genetics of prostate cancer: new prospects for old challenges. Genes Dev. 2010; 24:1967–2000. https://doi.org/10.1101/gad.1965810 PMID:20844012

31. Aspeslagh S, Postel-Vinay S, Rusakiewicz S, Soria JC, Zitvogel L, Marabelle A. Rationale for anti-OX40 cancer immunotherapy. Eur J Cancer. 2016; 52:50–66. https://doi.org/10.1016/j.ejca.2015.08.021 PMID:26645943

32. Buchan SL, Rogel A, Al-Shamkhani A. The immunobiology of CD27 and OX40 and their potential as targets for cancer immunotherapy. Blood. 2018; 131:39–48. https://doi.org/10.1182/blood-2017-07-741025 PMID:29118006

33. Fujita T, Ukyo N, Hori T, Uchiyama T. Functional characterization of OX40 expressed on human CD8+ T cells. Immunol Lett. 2006; 106:27–33. https://doi.org/10.1016/j.imlet.2006.04.001 PMID:16750861

34. Redmond WL, Gough MJ, Weinberg AD. Ligation of the OX40 co-stimulatory receptor reverses self-Ag and tumor-induced CD8 T-cell anergy *in vivo*. Eur J Immunol. 2009; 39:2184–94. https://doi.org/10.1002/eji.200939348 PMID:19672905

35. Sarff M, Edwards D, Dhungel B, Wegmann KW, Corless C, Weinberg AD, Vetto JT. OX40 (CD134) expression in sentinel lymph nodes correlates with prognostic features of primary melanomas. Am J Surg. 2008; 195:621–25. https://doi.org/10.1016/j.amjsurg.2007.12.036 PMID:18374895

36. Xie F, Wang Q, Chen Y, Gu Y, Mao H, Zeng W, Zhang X. Costimulatory molecule OX40/OX40L expression in ductal carcinoma *in situ* and invasive ductal carcinoma of breast: an immunohistochemistry-based pilot study. Pathol Res Pract. 2010; 206:735–39. https://doi.org/10.1016/j.prp.2010.05.016 PMID:20634005

37. Vetto JT, Lum S, Morris A, Sicotte M, Davis J, Lemon M, Weinberg A. Presence of the T-cell activation marker OX-40 on tumor infiltrating lymphocytes and draining lymph node cells from patients with melanoma and head and neck cancers. Am J Surg. 1997; 174:258–65. https://doi.org/10.1016/s0002-9610(97)00139-6 PMID:9324133

38. Morris ES, MacDonald KP, Kuns RD, Morris HM, Banovic T, Don AL, Rowe V, Wilson YA, Raffelt NC, Engwerda CR, Burman AC, Markey KA, Godfrey DI, et al. Induction of natural killer T cell-dependent alloreactivity by administration of granulocyte colony-stimulating factor after bone marrow transplantation. Nat Med. 2009; 15:436–41. https://doi.org/10.1038/nm.1948 PMID:19330008

39. Marabelle A, Kohrt H, Sagiv-Barfi I, Ajami B, Axtell RC, Zhou G, Rajapaksa R, Green MR, Torchia J, Brody J, Luong R, Rosenblum MD, Steinman L, et al. Depleting tumor-specific Tregs at a single site eradicates disseminated tumors. J Clin Invest. 2013; 123:2447–63. https://doi.org/10.1172/JCI64859 PMID:23728179

40. Petty JK, He K, Corless CL, Vetto JT, Weinberg AD. Survival in human colorectal cancer correlates with expression of the T-cell costimulatory molecule OX-40 (CD134). Am J Surg. 2002; 183:512–18. https://doi.org/10.1016/s0002-9610(02)00831-0 PMID:12034383

41. Braiman A, Isakov N. The Role of Crk Adaptor Proteins in T-Cell Adhesion and Migration. Front Immunol. 2015; 6:509. https://doi.org/10.3389/fimmu.2015.00509 PMID:26500649

42. Au-Yeung BB, Shah NH, Shen L, Weiss A. ZAP-70 in Signaling, Biology, and Disease. Annu Rev Immunol. 2018; 36:127–56. https://doi.org/10.1146/annurev-immunol-042617-053335 PMID:29237129

43. Lo WL, Shah NH, Ahsan N, Horkova V, Stepanek O, Salomon AR, Kuriyan J, Weiss A. Lck promotes Zap70-

dependent LAT phosphorylation by bridging Zap70 to LAT. Nat Immunol. 2018; 19:733–41. https://doi.org/10.1038/s41590-018-0131-1 PMID:29915297

44. Richardson SJ, Matthews C, Catherwood MA, Alexander HD, Carey BS, Farrugia J, Gardiner A, Mould S, Oscier D, Copplestone JA, Prentice AG. ZAP-70 expression is associated with enhanced ability to respond to migratory and survival signals in B-cell chronic lymphocytic leukemia (B-CLL). Blood. 2006; 107:3584–92. https://doi.org/10.1182/blood-2005-04-1718 PMID:16332969

45. Fu D, Liu B, Zang LE, Jiang H. MiR-631/ZAP70: A novel axis in the migration and invasion of prostate cancer cells. Biochem Biophys Res Commun. 2016; 469: 345–51. https://doi.org/10.1016/j.bbrc.2015.11.093 PMID:26620225

46. Begley LA, Kasina S, Mehra R, Adsule S, Admon AJ, Lonigro RJ, Chinnaiyan AM, Macoska JA. CXCL5 promotes prostate cancer progression. Neoplasia. 2008; 10:244–54. https://doi.org/10.1593/neo.07976 PMID:18320069

47. Wang G, Lu X, Dey P, Deng P, Wu CC, Jiang S, Fang Z, Zhao K, Konaparthi R, Hua S, Zhang J, Li-Ning-Tapia EM, Kapoor A, et al. Targeting YAP-Dependent MDSC Infiltration Impairs Tumor Progression. Cancer Discov. 2016; 6:80–95. https://doi.org/10.1158/2159-8290.CD-15-0224 PMID:26701088

48. Zink N, Bensmann W, Arning L, Stock AK, Beste C. CHRM2 Genotype Affects Inhibitory Control Mechanisms During Cognitive Flexibility. Mol Neurobiol. 2019; 56:6134–41. https://doi.org/10.1007/s12035-019-1521-6 PMID:30729426

49. Wang J, Ding Y, Wu Y, Wang X. Identification of the complex regulatory relationships related to gastric cancer from lncRNA-miRNA-mRNA network. J Cell Biochem. 2020; 121:876–87. https://doi.org/10.1002/jcb.29332 PMID:31452262

50. Yuen GJ, Demissie E, Pillai S. B lymphocytes and cancer: a love-hate relationship. Trends Cancer. 2016; 2: 747–57. https://doi.org/10.1016/j.trecan.2016.10.010 PMID:28626801

51. Coronella-Wood JA, Hersh EM. Naturally occurring B-cell responses to breast cancer. Cancer Immunol Immunother. 2003; 52:715–38. https://doi.org/10.1007/s00262-003-0409-4 PMID:12920480

52. Marsigliante S, Biscozzo L, Marra A, Nicolardi G, Leo G, Lobreglio GB, Storelli C. Computerised counting of tumour infiltrating lymphocytes in 90 breast cancer specimens. Cancer Lett. 1999; 139:33–41. https://doi.org/10.1016/s0304-3835(98)00379-6 PMID:10408906

53. Stamatopoulos B, Smith T, Crompot E, Pieters K, Clifford R, Mraz M, Robbe P, Burns A, Timbs A, Bruce D, Hillmen P, Meuleman N, Mineur P, et al. The Light Chain IgLV3-21 Defines a New Poor Prognostic Subgroup in Chronic Lymphocytic Leukemia: Results of a Multicenter Study. Clin Cancer Res. 2018; 24: 5048–57. https://doi.org/10.1158/1078-0432.CCR-18-0133 PMID:29945996

54. Ghia EM, Jain S, Widhopf GF 2nd, Rassenti LZ, Keating MJ, Wierda WG, Gribben JG, Brown JR, Rai KR, Byrd JC, Kay NE, Greaves AW, Kipps TJ. Use of IGHV3-21 in chronic lymphocytic leukemia is associated with high-risk disease and reflects antigen-driven, post-germinal center leukemogenic selection. Blood. 2008; 111:5101–08. https://doi.org/10.1182/blood-2007-12-130229 PMID:18326815

55. Maity PC, Bilal M, Koning MT, Young M, van Bergen CA, Renna V, Nicolò A, Datta M, Gentner-Göbel E, Barendse RS, Somers SF, de Groen RA, Vermaat JS, et al. IGLV3-21 * 01 is an inherited risk factor for CLL through the acquisition of a single-point mutation enabling autonomous BCR signaling. Proc Natl Acad Sci USA. 2020; 117:4320–27. https://doi.org/10.1073/pnas.1913810117 PMID:32047037

56. Korkmaz B, Moreau T, Gauthier F. Neutrophil elastase, proteinase 3 and cathepsin G: physicochemical properties, activity and physiopathological functions. Biochimie. 2008; 90:227–42. https://doi.org/10.1016/j.biochi.2007.10.009 PMID:18021746

57. Liu X, Tian Y, Meng Z, Chen Y, Ho IH, Choy KW, Lichtner P, Wong SH, Yu J, Gin T, Wu WK, Cheng CH, Chan MT. Up-regulation of Cathepsin G in the Development of Chronic Postsurgical Pain: An Experimental and Clinical Genetic Study. Anesthesiology. 2015; 123:838–50. https://doi.org/10.1097/ALN.0000000000000828 PMID:26270939

58. Jin W, Wu K, Li YZ, Yang WT, Zou B, Zhang F, Zhang J, Wang KK. AML1-ETO targets and suppresses cathepsin G, a serine protease, which is able to degrade AML1-ETO in t(8;21) acute myeloid leukemia. Oncogene. 2013; 32:1978–87. https://doi.org/10.1038/onc.2012.204 PMID:22641217

59. Park SY, Piao Y, Jeong KJ, Dong J, de Groot JF. Periostin (POSTN) Regulates Tumor Resistance to Antiangiogenic Therapy in Glioma Models. Mol Cancer Ther. 2016; 15:2187–97. https://doi.org/10.1158/1535-7163.MCT-15-0427 PMID:27307601

60. Norris RA, Potts JD, Yost MJ, Junor L, Brooks T, Tan H, Hoffman S, Hart MM, Kern MJ, Damon B, Markwald RR, Goodwin RL. Periostin promotes a fibroblastic lineage pathway in atrioventricular valve progenitor cells. Dev Dyn. 2009; 238:1052–63. https://doi.org/10.1002/dvdy.21933 PMID:19334280

61. Malanchi I, Santamaria-Martínez A, Susanto E, Peng H, Lehr HA, Delaloye JF, Huelsken J. Interactions between cancer stem cells and their niche govern metastatic colonization. Nature. 2011; 481:85–89. https://doi.org/10.1038/nature10694 PMID:22158103

62. Wang H, Wang Y, Jiang C. Stromal protein periostin identified as a progression associated and prognostic biomarker in glioma via inducing an invasive and proliferative phenotype. Int J Oncol. 2013; 42:1716–24. https://doi.org/10.3892/ijo.2013.1847 PMID:23467707

63. Buhl AM, Jurlander J, Jørgensen FS, Ottesen AM, Cowland JB, Gjerdrum LM, Hansen BV, Leffers H. Identification of a gene on chromosome 12q22 uniquely overexpressed in chronic lymphocytic leukemia. Blood. 2006; 107:2904–11. https://doi.org/10.1182/blood-2005-07-2615 PMID:16339396

64. Zheng LS, Yang JP, Cao Y, Peng LX, Sun R, Xie P, Wang MY, Meng DF, Luo DH, Zou X, Chen MY, Mai HQ, Guo L, et al. SPINK6 Promotes Metastasis of Nasopharyngeal Carcinoma via Binding and Activation of Epithelial Growth Factor Receptor. Cancer Res. 2017; 77:579–89. https://doi.org/10.1158/0008-5472.CAN-16-1281 PMID:27671677

65. Bonsor DA, Beckett D, Sundberg EJ. Structure of the N-terminal dimerization domain of CEACAM7. Acta Crystallogr F Struct Biol Commun. 2015; 71:1169–75. https://doi.org/10.1107/S2053230X15013576 PMID:26323304

66. Salek Esfahani B, Gharesouran J, Ghafouri-Fard S, Talebian S, Arsang-Jang S, Omrani MD, Taheri M, Rezazadeh M. Down-regulation of ERMN expression in relapsing remitting multiple sclerosis. Metab Brain Dis. 2019; 34:1261–66. https://doi.org/10.1007/s11011-019-00429-w PMID:31123898

67. Zhang L, Xue Z, Liu Q, Liu Y, Xi S, Cheng Y, Li J, Yan J, Shen Y, Xiao C, Xie Z, Qiu Z, Jiang H. Disrupted folate metabolism with anesthesia leads to myelination deficits mediated by epigenetic regulation of ERMN. EBioMedicine. 2019; 43:473–86. https://doi.org/10.1016/j.ebiom.2019.04.048 PMID:31060905

68. Matsumoto K, Shimodaira M, Nakagawa T, Nakayama T, Nakazato T, Izumi Y, Soma M, Matsumoto K, Sato N, Aoi N. Association study: SLC6A18 gene and myocardial infarction. Clin Biochem. 2011; 44:789–94. https://doi.org/10.1016/j.clinbiochem.2011.03.031 PMID:21420947

69. Citterio CE, Targovnik HM, Arvan P. The role of thyroglobulin in thyroid hormonogenesis. Nat Rev Endocrinol. 2019; 15:323–38. https://doi.org/10.1038/s41574-019-0184-8 PMID:30886364

70. Coscia F, Taler-Verčič A, Chang VT, Sinn L, O'Reilly FJ, Izoré T, Renko M, Berger I, Rappsilber J, Turk D, Löwe J. The structure of human thyroglobulin. Nature. 2020; 578:627–30. https://doi.org/10.1038/s41586-020-1995-4 PMID:32025030

71. Xu Y, Wang W, Chen J, Mao H, Liu Y, Gu S, Liu Q, Xi Q, Shi W. High neuropilin and tolloid-like 1 expression associated with metastasis and poor survival in epithelial ovarian cancer via regulation of actin cytoskeleton. J Cell Mol Med. 2020; 24:9114–24. https://doi.org/10.1111/jcmm.15547 PMID:32638511

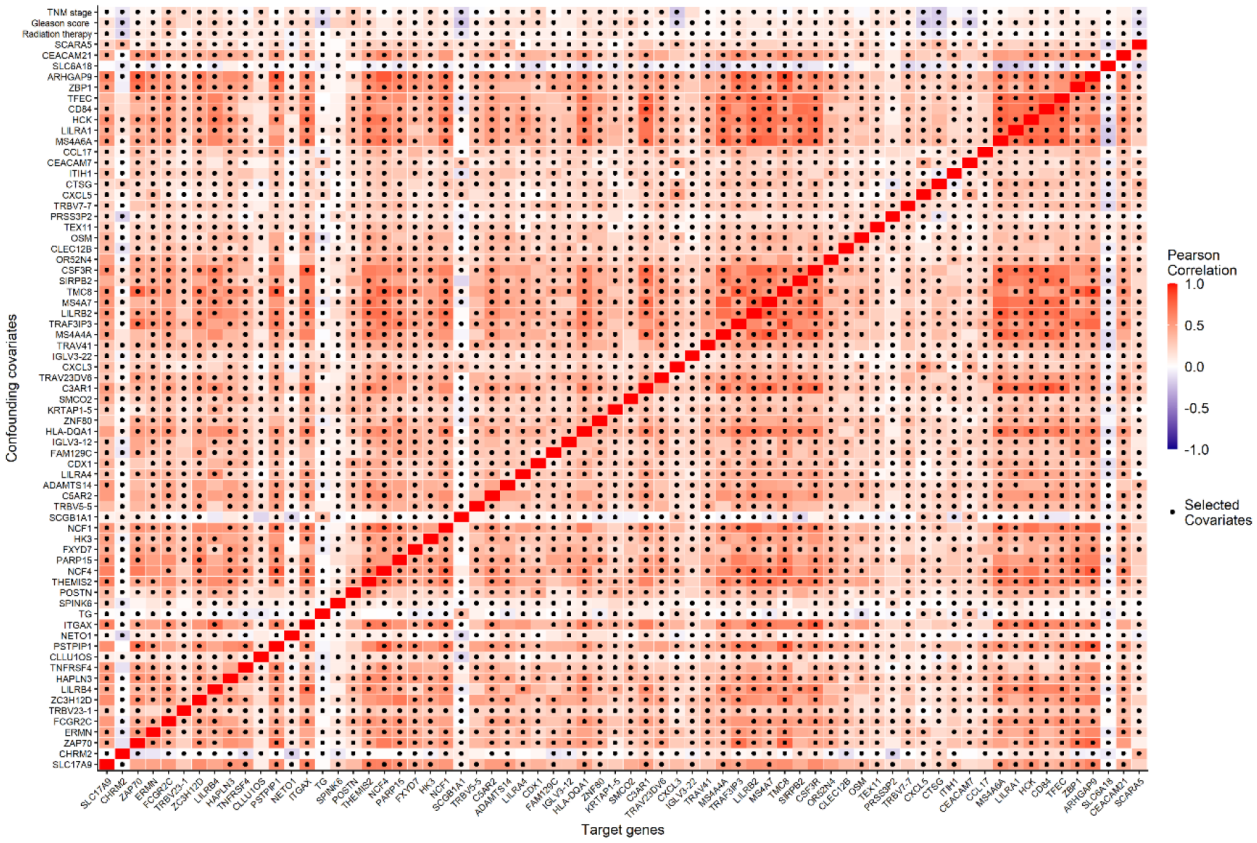
72. Mariani A, Wang C, Oberg AL, Riska SM, Torres M, Kumka J, Multinu F, Sagar G, Roy D, Jung DB, Zhang Q, Grassi T, Visscher DW, et al. Genes associated with bowel metastases in ovarian cancer. Gynecol Oncol. 2019; 154:495–504. https://doi.org/10.1016/j.ygyno.2019.06.010 PMID:31204077

73. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - A Practical and Powerful Approach To Multiple Testing. J R Stat Soc B. 1995; 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

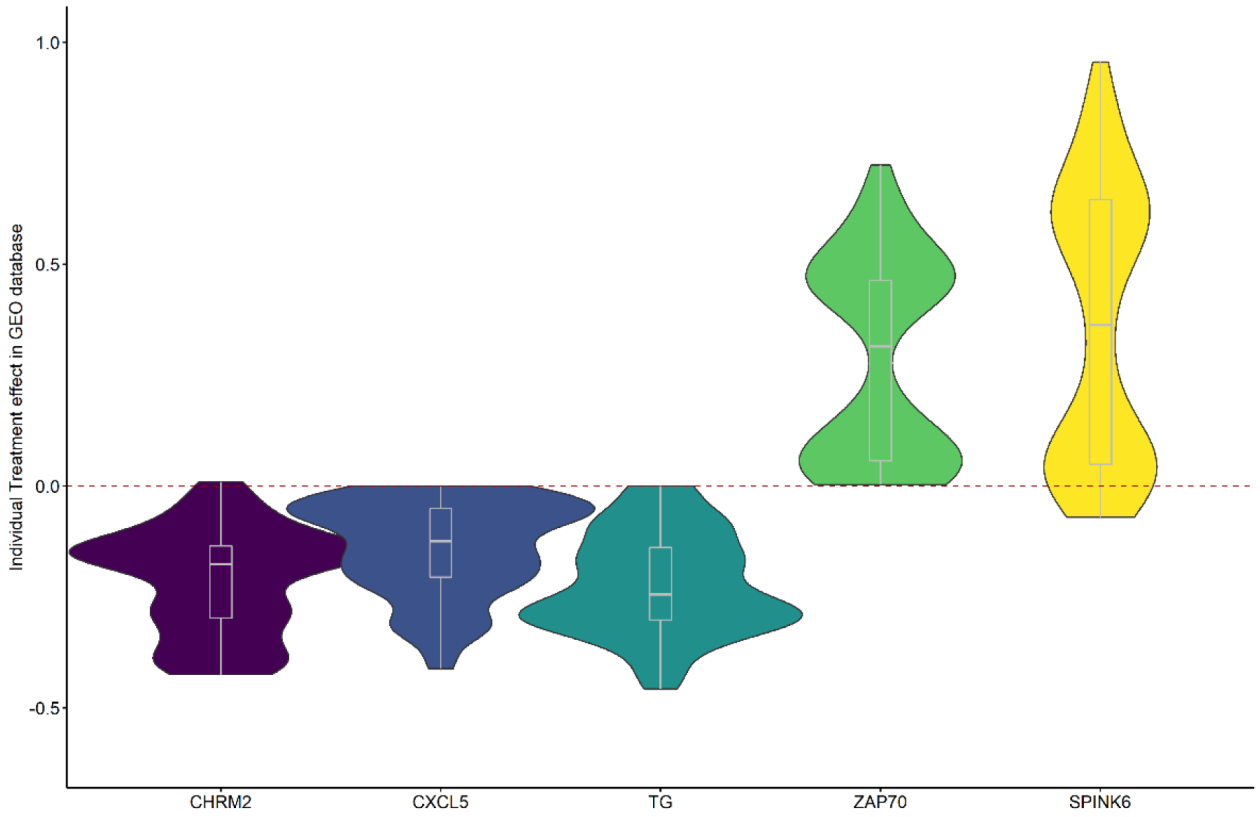
74. Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman and Hall/CRC. 2019.

SUPPLEMENTARY MATERIALS

Supplementary Figures



Supplementary Figure 1. The Pearson correlation coefficients (the corresponding color) of the 68 candidate genes (horizontal axis) and the 70 candidate confounding covariates (vertical axis), as well as the covariate set W_G that the candidate confounding covariates after removing variables W_G would be conditionally independent of outcome Y given W_G (the dark dot of each column).



Supplementary Figure 2. The individual causal effects of the validated genes in the GEO dataset.