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

Xiaoru Sun1,2, Lu Wang1,2, Hongkai Li1,2, Chuandi Jin1,2, Yuanyuan Yu1,2, Lei Hou1,2, Xinhui Liu1,2, Yifan Yu1,2, Ran Yan1,2, Fuzhong Xue1,2

1Department of Biostatistics, School of Public Health, Cheeloo College of Medicine, Shandong University, Jinan 250012, Shandong, China
2Institute 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
Copyright: © 2021 Sun et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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.

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

*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, CCLU1OS, POSTN, CTSG, NETO1, CEACAM7, and IGLV3-22) on 12 chromosomes with P-ACE < 0.05 (Table 2). 7 out of these genes (ZAP70, SPINK6, ERMN, CXCL5, SPINK6, SLC6A18, CHRM2, TG, CCLU1OS, POSTN, CTSG, NETO1, CEACAM7, and IGLV3-22)

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.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Chromosome</th>
<th>ACE (95% CI)</th>
<th>P-ACE</th>
<th>MOR (95% CI)</th>
<th>P-MOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>TNFRSF4</td>
<td>1</td>
<td>0.092 (0.022, 0.163)</td>
<td>0.011</td>
<td>1.454 (1.089, 1.940)</td>
<td>0.011</td>
</tr>
<tr>
<td>ZAP70</td>
<td>2</td>
<td>0.116 (0.044, 0.188)</td>
<td>0.002</td>
<td>1.597 (1.191, 2.143)</td>
<td>0.002</td>
</tr>
<tr>
<td>ERMN</td>
<td>2</td>
<td>0.135 (0.050, 0.219)</td>
<td>0.002</td>
<td>1.728 (1.222, 2.444)</td>
<td>0.002</td>
</tr>
<tr>
<td>CXCL5</td>
<td>4</td>
<td>-0.142 (-0.226, -0.059)</td>
<td>0.001</td>
<td>0.562 (0.399, 0.792)</td>
<td>0.001</td>
</tr>
<tr>
<td>SPINK6</td>
<td>5</td>
<td>0.153 (0.047, 0.26)</td>
<td>0.005</td>
<td>1.865 (1.201, 2.897)</td>
<td>0.005</td>
</tr>
<tr>
<td>SLC6A18</td>
<td>5</td>
<td>0.165 (0.017, 0.313)</td>
<td>0.029</td>
<td>1.944 (1.056, 3.578)</td>
<td>0.033</td>
</tr>
<tr>
<td>CHRM2</td>
<td>7</td>
<td>-0.173 (-0.266, -0.079)</td>
<td>&lt;0.001</td>
<td>0.496 (0.337, 0.731)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>TG</td>
<td>8</td>
<td>-0.127 (-0.243, -0.012)</td>
<td>0.031</td>
<td>0.598 (0.372, 0.960)</td>
<td>0.033</td>
</tr>
<tr>
<td>CLLU1OS</td>
<td>12</td>
<td>0.13 (0.053, 0.207)</td>
<td>0.001</td>
<td>1.689 (1.233, 2.312)</td>
<td>0.001</td>
</tr>
<tr>
<td>POSTN</td>
<td>13</td>
<td>0.083 (0.009, 0.156)</td>
<td>0.027</td>
<td>1.398 (1.037, 1.885)</td>
<td>0.028</td>
</tr>
<tr>
<td>CTS5</td>
<td>14</td>
<td>-0.133 (-0.229, -0.038)</td>
<td>0.006</td>
<td>0.584 (0.395, 0.863)</td>
<td>0.007</td>
</tr>
<tr>
<td>NETO1</td>
<td>18</td>
<td>0.137 (0.023, 0.252)</td>
<td>0.019</td>
<td>1.740 (1.090, 2.778)</td>
<td>0.020</td>
</tr>
<tr>
<td>CEACAM7</td>
<td>19</td>
<td>-0.165 (-0.289, -0.042)</td>
<td>0.009</td>
<td>0.511 (0.307, 0.851)</td>
<td>0.010</td>
</tr>
<tr>
<td>IGLV3-22</td>
<td>22</td>
<td>0.166 (0.004, 0.328)</td>
<td>0.044</td>
<td>1.956 (1.001, 3.823)</td>
<td>0.050</td>
</tr>
</tbody>
</table>

*CI: confidence interval.

CHRMT2, TG, CLLU1OS, POSTN, and NET01) 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, CTS5, and CEACAM7).

Figure 5 shows the individual treatment effects of each prognostic gene on BCR. The unfavorable genes were a
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, CLLU10S, 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...
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.
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.

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

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

**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 years 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 χ² tests (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. \( G = \{g_1, g_2, \ldots, g_\ell\} \) 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, \( \{G \setminus g_i\} + X \) denoted the complete covariate vector of the target gene \( g_i \) and \( Y \). We assumed that \( W_G \) was the subset of confounding covariates \( \{(G \setminus g_i) + X\} \), which satisfied \( Y \indep \{(G \setminus g_i) + X\}|W_G \). \( V_G \) was the minimal set of confounding covariates \( W_G \) satisfying \( G \perp W_G|V_G \), 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 \( \{E[Y(1)] × E[1 − Y(0)]\} /\{E[1 − Y(1)] × E[Y(0)]\} \), 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 \( E_Y \{E_G [E_Y (Y | G = 1, V_G) − E_Y (Y | G = 0, V_G)] \} \) 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, \( E(Y | G, V_G) \), which was used to predict the potential outcomes. Second, we evaluated the exposure mechanism \( P(G = 1|V_G) \) to update the estimation of \( E(Y | G, V_G) \). We eventually used the final updated estimation of \( E(Y | G, V_G) \) 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


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

   https://doi.org/10.1038/nm.3394
   PMID:24202395

   https://doi.org/10.18632/aging.101415
   PMID:29676997

   https://doi.org/10.1155/2020/5019793
   PMID:32509861

   https://doi.org/10.1038/nrc2618
   PMID:19279573

   https://doi.org/10.1016/j.ccell.2017.06.003
   PMID:28697342

    https://doi.org/10.1038/srep37085
    PMID:27848972

    https://doi.org/10.1210/er.2002-0032
    PMID:15082523

    https://doi.org/10.1093/ije/15.3.413
    PMID:3771081

    https://doi.org/10.1097/00001648-199901000-00008
    PMID:9888278

    https://doi.org/10.1001/jama.2015.13480
    PMID:26501539

    https://doi.org/10.1093/ije/dyw323
    PMID:28039382

    https://doi.org/10.2202/1557-4679.104

    https://doi.org/10.1093/aje/kww165
    PMID:27941068

    https://doi.org/10.1002/sim.7628
    PMID:29687470

    https://doi.org/10.7326/0003-4819-127-8_part_2-19971015-00064
    PMID:9382394

    https://doi.org/10.1093/biomet/asr041

    https://doi.org/10.1186/s12874-017-0449-7
    PMID:29281984

    https://doi.org/10.18637/jss.v068.i01

    https://doi.org/10.1002/sim.8792
    PMID:33150645

    https://doi.org/10.1038/ncomms3612
    PMID:24113773


43. Lo WL, Shah NH, Ahsan N, Horkova V, Stepanek O, Salomon AR, Kuriyan J, Weiss A. Lck promotes Zap70-
https://doi.org/10.1038/s41590-018-0131-1 
PMID: 29915297

https://doi.org/10.1182/blood-2005-04-1718 
PMID: 16332969

https://doi.org/10.1016/j.bbrc.2015.11.093 
PMID: 26620225

https://doi.org/10.1593/neo.07976 PMID: 18320069

https://doi.org/10.1158/2159-8290.CD-15-0224 
PMID: 26701088

https://doi.org/10.1007/s12035-019-1521-6 
PMID: 30729426

https://doi.org/10.1002/jcb.29332 PMID: 31452262

https://doi.org/10.1016/j.trecan.2016.10.010 PMID: 28626801

https://doi.org/10.1007/s00262-003-0409-4 PMID: 12920480

https://doi.org/10.1016/s0304-3835(98)00379-6 
PMID: 10408906

https://doi.org/10.1158/1078-0432.CCR-18-0133 
PMID: 29945996

https://doi.org/10.1182/blood-2007-12-130229 
PMID: 18326815

https://doi.org/10.1073/pnas.1913810117 
PMID: 32047037

https://doi.org/10.1016/j.biochi.2007.10.009 
PMID: 18021746

https://doi.org/10.1097/ALN.0000000000000828 
PMID: 26270939

https://doi.org/10.1038/onc.2012.204 
PMID: 22641217


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.