



OPEN

A transcriptomic signature for prostate cancer relapse prediction identified from the differentially expressed genes between TP53 mutant and wild-type tumors

Wensheng Zhang¹✉ & Kun Zhang^{1,2}✉

For prostate cancer (PCa) patients, biochemical recurrence (BCR) is the first sign of disease relapse and the subsequent metastasis. *TP53* mutations are relatively prevalent in advanced PCa forms. We aimed to utilize this knowledge to identify robust transcriptomic signatures for BCR prediction in patients with Gleason score ≥ 7 cancers, which cause most PCa deaths. Using the TCGA-PRAD dataset and the novel data-driven stochastic approach proposed in this study, we identified a 25-gene signature from the genes whose expression in tumors was associated with *TP53* mutation statuses. The predictive strength of the signature was assessed by AUC and Fisher's exact test p-value according to the output of support vector machine-based cross validation. For the TCGA-PRAD dataset, the AUC and p-value were 0.837 and 5×10^{-13} , respectively. For five external datasets, the AUCs and p-values ranged from 0.632 to 0.794 and 6×10^{-2} to 5×10^{-5} , respectively. The signature also performed well in predicting relapse-free survival (RFS). The signature-based transcriptomic risk scores (TRS) explained 28.2% of variation in RFS on average. The combination of TRS and clinicopathologic prognostic factors explained 23–72% of variation in RFS, with a median of 54.5%. Our method and findings are useful for developing new prognostic tools in PCa and other cancers.

The state of an increasing prostate-specific antigen level after radical prostatectomy (RP) or radiation therapy (RT) for localized prostate cancer (PCa) is known as biochemical recurrence (BCR) or biochemical relapse. The rate of BCR following RP was estimated to be 20–40%. BCR is the first sign for disease relapse and subsequent lethal metastases¹, occurring within a wide time span from a few months to over 15 years following the initial therapy². Cases of PCa progression with undetectable or low PSA levels have rarely been observed^{3,4}. In the absence of secondary treatment, patients with BCR have an approximate median period of 5–8 years prior to clinical progression^{2,3}. BCR events usually occur among patients with at least one of the primary and secondary prevalent Gleason patterns (GPs) being graded as 4 or 5. BCR risk and disease-specific mortality increase with the climbing proportions of the GP-4 and GP-5 components, from 3 + 3 with tertiary 4 to 3 + 4, 4 + 3 and 4 + 4, in prostatectomy specimens^{5–7}. Prostate cancer without GP-4 or GP-5 components is unable to metastasize or cause cancer-associated mortality, in addition to having a low risk rate for BCR^{8,9}. BCR and BCR-free survival are significantly associated with overall survival (OS), but they are poorer as surrogate endpoints for OS than metastasis-free survival¹⁰.

The Gleason score (GS) is the sum of the grades of the first and second Gleason patterns of a primary cancer sample. Within the same GS group, individual cancers have heterogeneous molecular mechanisms, implicating varied progression potential. In particular, even patients with $GS \geq 8$ can experience favorable oncological outcomes¹¹. In this regard, stratifying $GS \geq 7$ into relapse risk groups is pivotal to the management of prostate cancer. For example, it can be helpful for scheduling follow-up surveillance after the initial treatment.

¹Bioinformatics Core of Xavier NIH RCMI Center of Cancer Research, Xavier University of Louisiana, New Orleans, LA 70125, USA. ²Department of Computer Science, Xavier University of Louisiana, New Orleans, LA 70125, USA. ✉email: wzhang@xula.edu; kzhang@xula.edu

In past years, several studies have focused on mining predictive marker sets (signatures) for BCR or relapse-free survival (RFS) from high-throughput molecular data resources. The numbers of the involved genes in the identified signatures range from four to a few dozen^{12–16}. While those signatures were valid for the datasets or experimental settings in the reported studies, their robustness and prognostic utility could be uncertain due to drawbacks in the design/analysis, such as the inclusion of low-risk GS-6 cancer samples in the datasets used^{13,14,16} and/or the lack of sufficient validation using external datasets^{14,16}. Moreover, the strategy of identifying a signature for the prediction of cancer relapse from the top differentially expressed genes between tumors and normal samples or between benign tumors and malignant tumors was usually adopted in those studies^{12,14}. As such, overfitting likely arose due to the lack of a mechanism to reduce the risk of its occurrence. An additional challenge is that a molecular signature with a practical application perspective should complement clinicopathologic prognostic factors such as GS in outcome prediction. In a recent report, Wu et al¹² addressed this issue, but our preliminary analysis demonstrated that their result obtained from analyzing the Cancer Genome Atlas Prostate Adenocarcinoma (TCGA-PRAD) dataset could not be confirmed using other datasets (see Results section).

In this study, we aimed to identify a robust predictive gene expression signature for BCR in patients with Gleason score (GS) ≥ 7 cancers. To achieve our goal, a data-driven and biologically informed stochastic approach was developed, which begins with the use of the TCGA-PRAD dataset to identify differentially expressed genes regarding *TP53* mutation status in cancer samples. In the cohort, *TP53* mutations are less frequent than *PTEN* mutations but are more enriched than aberrations in other key oncogenes and/or tumor suppressors such as *KRAS*, *BRAF*, *EGFR* and *MYC*¹⁷. Our methodology conceived with the hypotheses that a *TP53* mutation status-associated prognostic gene signature could be robust, namely, the prediction strength for BCR could keep when the gene expression levels in the datasets of different patient cohorts are measured with varied platforms and experimental settings. Underlying this perception is the fact or finding that the prevalence of somatic *TP53* mutations in advanced forms of PCa is approximately 4 times the quantity in primary cancer^{17,18}, which suggests that the genes with *TP53* mutation status-associated transcription could be enriched with prognostic and etiological factors for cancer relapse.

Material and methods

Signature discovery. *Scheme.* The stochastic approach we have developed for identifying a gene expression signature for BCR is data-driven and biologically informed. It includes three modules (A, B and C parts of Fig. 1). First, using the TCGA-PRAD dataset, the genes differentially expressed between the tumors with somatic *TP53* mutation(s) and those without mutations on the gene are detected. Second, 1000 small (size = 25. See the “Remark” paragraph) subsets of genes are randomly sampled from the output of first modules, and their predictive strengths for BCR are assessed with the area under the receiver operating characteristic curve (AUC) and Fisher’s exact test p-value. For each of the gene subsets (random signatures), the two performance metrics are calculated according to the predicted BCR risk category and decision values of individual subjects from support vector machine-based leave-one-out cross validation (SVM-LOOCV). Third, the results from the second modules are integrated using a “filter” and a “wrapper” to obtain an optimized gene signature.

Filter. The filter works via two operations. OP-1: The gene subsets whose AUCs and p-values meet the cutoffs of ≥ 0.65 and ≤ 0.0001 (see the “Remark” paragraph) are selected. OP-2: The genes with at least three hits in the selected subsets are collected as the “initial” signature.

Wrapper. Suppose the initial signature contains m genes and denote it by a character vector $\mathcal{G} = \{g_1, g_2, \dots, g_m\}$. The wrapper is designed to refine \mathcal{G} and is realized via the following algorithm.

- (1) A numeric vector $A = \{A_{(-1)}, A_{(-2)}, \dots, A_{(-m)}\}$ is calculated, where $A_{(-i)}$ represents the SVM-LOOCV AUC obtained when all the genes in \mathcal{G} except for g_i are used as the features for predicting BCR.
- (2) The elements of vector A are sorted in ascending order to generate $A^* = \{A_1^*, A_2^*, \dots, A_m^*\}$. Correspondingly, the gene vector \mathcal{G} is rearranged to generate $\mathcal{G}^* = \{g_1^*, g_2^*, \dots, g_m^*\}$.
- (3) Specify the lower limit of the size of a desired signature, such as 10, and denote it with k . A numeric vector $B = \{B_{(-1;1)}, B_{(-1;2)}, \dots, B_{(-1;(m-k))}\}$ is calculated, where $B_{(-1;i)}$ represents the SVM-LOOCV AUC obtained when the genes $\{g_{i+1}^*, g_{i+2}^*, \dots, g_m^*\}$ are used as the prediction features. The genes corresponding to individual elements of B are counted, and the numbers are collected into an integer vector C .
- (4) Create scatter plot with C as x-axis and B as y-axis, and model the relationship between C and B with a single-mode smooth-splining curve. The data point adjacent to the mode of the curve from the left side is visually pinpointed, and the corresponding gene set (a subset of \mathcal{G}^*) is determined as the finally selected signature.

Remark. The parameters used in the discovery procedure were set by referring to the results of a preliminary study. These results include (1) a random signature consisting of 10–200 genes rarely demonstrated a prediction strength of $\text{AUC} \geq 0.70$ and $p \leq 0.00001$; and (2) among numbers 10, 25, 50, 100 and 200, the second was the best as the size of the gene subset for efficiently selecting candidate prognostic signatures that could meet the modestly specified performance criteria.

After the signature for BCR prediction was identified, its robustness and prognostic utility were future evaluated by the analyses outlined in the parts D1 and D2 of the Fig. 1.

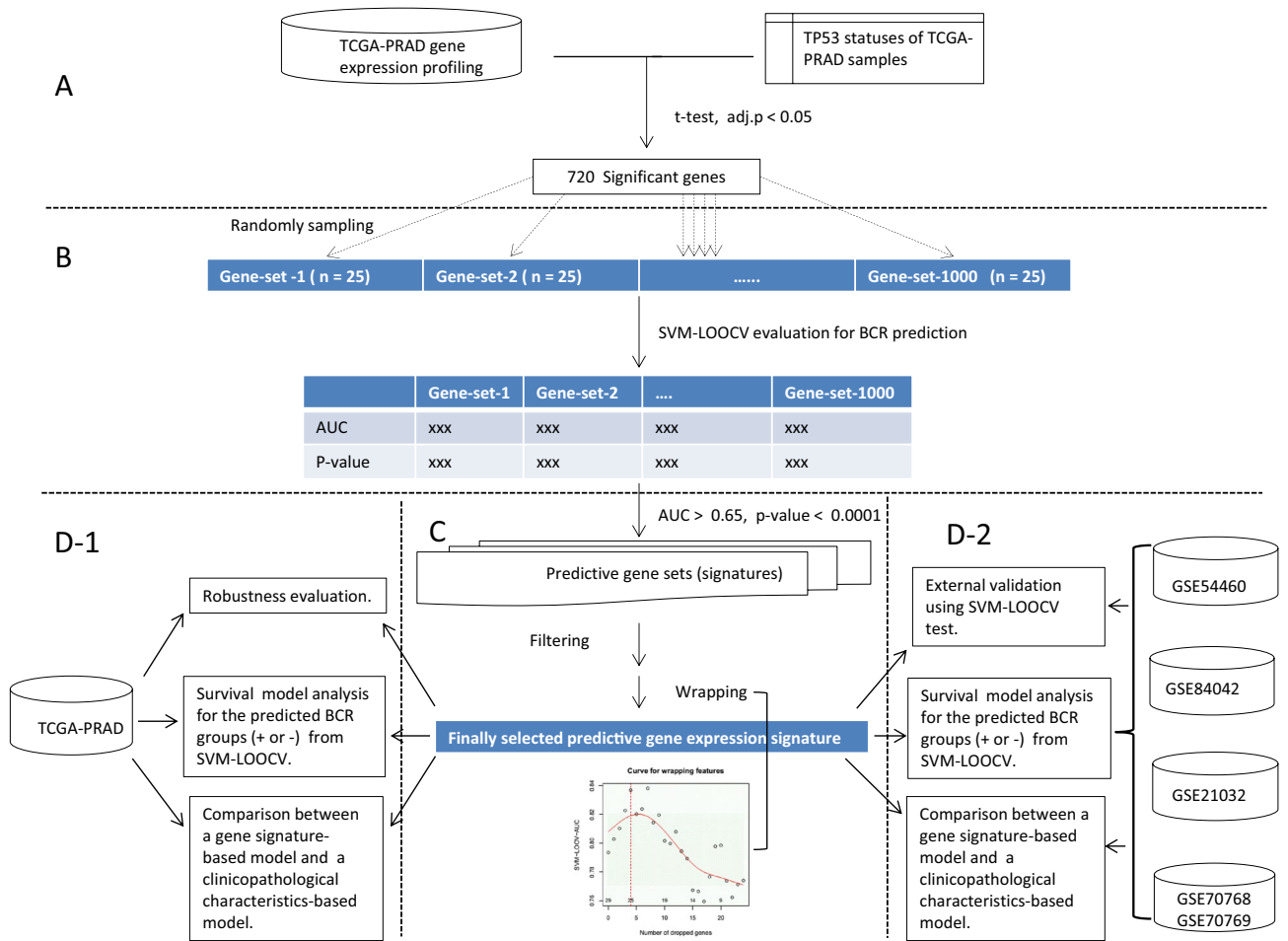


Figure 1. Flow chart of the identification approach (A,B,C) and performance/utility evaluation (D-1, D-2) of a TP53 mutation status-associated predictive transcriptomic signature for BCR. In A, the top differentially expressed genes (DEGs) regarding TP53 statuses were identified. In B, 1000 small subsets of the DEGs were randomly sampled and their predictive strengths for BCR were assessed by SVM-based cross validation. In C, the results from B step were integrated by a “filter” and a novel “wrapper” to obtain an optimized gene signature. In D1) and D2), the performance of the finally selected signature for BCR prediction and the clinical utility were evaluated in the TCGA dataset and five external datasets using statistical and machine learning methods. See the main text for a more detailed explanation.

Statistics and machine learning methods. SVM-LOOCV. First, based on the clinical outcomes regarding binary BCR occurrence, the N cancer cases in a cohort are divided into two classes: BCR⁻ (“-1” group) and BCR⁺ (“1” group). The labels of these cancer cases are then saved in a vector $Y = (y_1, y_2, \dots, y_i, \dots, y_N)$, where $y_i \in (-1, 1)$. After that, the (assumedly unknown) class of a leave-out tumor i is predicted from its gene expression profiling (\vec{x}_i) of the signature genes by the SVM model, which is trained on the data $(\{\vec{x}_j; y_j\})$ of the other $N-1$ samples. That is,

$$\hat{z}_i = \text{sign}(t_i), t_i = \sum_{j \in S} a_j y_j k\left(\vec{x}_j, \vec{x}_i\right) + b, S = \{1, 2, 3, \dots, N\}$$

In the equations, \hat{z}_i denotes the predicted category (1 or -1) for the i th sample; t_i is the decision value, $k\left(\vec{x}_j, \vec{x}_i\right)$ is the kernel function, and $\{a_j\}$ and b are the model parameters decided in the previous training process. Third, by summarizing the true label vector Y and the output label (i.e., predicted label) vector $\hat{Z} = (\hat{z}_1, \hat{z}_2, \hat{z}_3, \dots, \hat{z}_N)$, a 2×2 contingency table is generated, on which Fisher’s test of independence is performed. Finally, by combining the true label vector Y and an assemblage of tumor sample classifications based on the vector of decision values $T = (t_1, t_2, t_3, \dots, t_N)$, i.e. the transcriptomic risk scores estimated from SVM-LOOCV, and serially changed cutoffs, a receiver operating characteristic curve is generated and the AUC is calculated.

Data ID	Sample sizes	Sample partition on Gleason pattern (primary + second)				Sample partition on clinical T-stage				BCR %	Interquartile of ages at diag	
		3 + 4	4 + 3	3 + 5, 4 + 4 5, 5 + 4 5		T1	T2	T3	NA		Q2	Q3
TCGA-PRAD	366	114	83	169		124	133	48	61	13.6	57.0	66.5
GSE54460	95	56	24	15		10	67	18	0	53.7	56.7	66.3
GSE84042	57	40	17	0		0	28	29	0	24.6	56.2	63.9
GSE21032	89	53	21	15		43	42	4	0	28.1	54.3	61.8
GSE70768	95	65	21	9		47	32	14	2	20.0	56.0	65.0
GSE70769	70	36	19	15		26	32	9	3	62.9	NA	NA

Table 1. Summary of the used datasets regarding clinicopathologic characteristics of patients.

Survival analysis. The association between the predicted BCR groups and relapse-free survival (RFS) was evaluated with the p-value from the log-rank test. The performance comparison between transcriptomic risk scores and clinicopathologic prognostic factors was conducted using Cox-PH regression models. The explained variation (R^2), i.e., the proportion of variability in the outcome variable RFS explained by the explanatory variable(s), was calculated using Royston's method¹⁹. The goodness-of-fit of a survival model was also evaluated with Schwarz's Bayesian information criterion (BIC). When picking from several models, the model with a lower BIC value is generally preferred.

Software and application notes. Statistical/computational analysis was completed using the relevant functions in the R packages "stats", "01,071", "AUC", "survival" and "survMisc", as well as our labor-owned R codes. In the implementation of the *SVM()* function, except for the specially noted cases, a linear kernel was used, the class weights were specified as the reciprocals of the ratios between the "1" samples to the "-1" samples in the training set, and defaults for the hyperparameter *cost* and *gamma* were held on. The p-value from two-tailed Fisher's exact test was calculated in evaluating the randomly sampled signatures (gene subsets), and the p-value from one-tailed Fisher's exact test was calculated in evaluating the finally identified signature. In the analyses where BCR was treated as a binary endpoint, the time from the initial PCa diagnosis to relapse for a BCR⁺ sample and to the end of follow-up for a BCR⁻ sample were not considered.

Data. The clinical data of the six cohorts used as the discovery set (the TCGA-PRAD cohort) or testing/validation cohorts (the GSE54460 and others) in this study are summarized in Table 1. While these cohorts were filtered by the criterion of $GS \geq 7$ preceding our advanced analysis, we still denoted them with the IDs given in the source database. The following is a brief description of the gene expression datasets of those cohorts.

Discovery dataset. The level-3 gene expression dataset (version 2) of TCGA-PRAD samples was downloaded from the Genomic Data Commons Data Portal. The TCGA group performed RNA-Seq experiments on an Illumina HiSeq platform and estimated the gene expression levels by transcripts per million (TPM) values using an expectation maximization method and RSEM software²⁰. Log₂ transformation on this dataset was performed preceding the analysis.

Testing (validation) datasets. All the gene expression datasets (the series matrices) of the testing cohorts were downloaded from the Gene Expression Omnibus (GEO) database. The authors of GSE54460²¹ performed RNA-Seq experiments with the Illumina HiSeq 2000 platform, mapped short reads on the human genome hg19 assembly using TopHat and Bowtie software, and estimated gene expression levels with fragments per kilobase million (FPKM) values using Cufflinks software. The authors of GSE84042²² performed microarray experiments with Affymetrix Human Gene 2.0 ST array and preprocessed expression intensities using the robust multichip average (RMA) algorithm²³ and log₂ transformation. The authors of GSE21032²⁴ performed microarray experiments using Affymetrix Human Exon 1.0 ST Array and preprocessed expression intensities using RMA and quantile normalization. We performed log₂ transformation on the downloaded dataset, which contained transcript (Refseq RNA) expression levels. For a gene with two or multiple transcript IDs, we chose the one whose expression levels across samples had the largest interquartile range (IRQ) as the representative. The raw data of GSE70768 and GSE70769 were generated by the same authors using Illumina HumanHT-12 V4.0 Expression BeadChip¹⁶. We first downloaded the two matrices of the non-normalized expression levels of the two cohorts and removed the columns for the samples that would not be used in our analysis. Then, quantile normalization and log₂ transformation were applied. For a gene with two or multiple probes, we chose the one whose expression levels across samples had the largest IRQ in the GSE70768 cohort as the representative. Finally, we homogenized the two normalized expression matrices to make them have the same global 75% quantile.

Result

TP53 mutation status-associated (TP53-mut-ass) genes. We selected 720 TP53-mut-ass genes via the following procedure. First, based on the entire expression matrix of the TCGA-PRAD cohort, the genes unexpressed in at least half the samples were filtered out. The expression levels of those excluded genes were typically very low, even in the samples where the quantities were not zero. As such, we assume they may be not activated in prostate tissue actually and could hardly save as reliable prognostic features. Second, a t-test for the

Symbol	Name	Relevance with cancer/tumor/patient and references
<i>CDKN1A</i>	cyclin dependent kinase inhibitor 1A	Variants; advanced PCa ²⁵
<i>DDB1</i>	damage specific DNA binding protein 1	Apoptosis, chemo-resistance regulation and progression; multiple cancer types ²⁶⁻²⁸
<i>EIF5A2</i>	eukaryotic translation initiation factor 5A2	Cell growth, metastasis, chemotherapy resistance; multiple cancer types ^{29,30}
<i>GCDH</i>	glutaryl-CoA dehydrogenase	
<i>GK3P</i>	glycerol kinase 3 pseudogene	
<i>KIAA0196</i>	Strumpellin	Amplified and overexpressed; PCa ³¹
<i>LARP4B</i>	La ribonucleoprotein 4B	Cell migration and invasion; PCa ³²
<i>NAA50 (NAT13)</i>	N-alpha-acetyltransferase 50, NatE catalytic subunit	
<i>NDUFA9</i>	NADH: ubiquinone oxidoreductase subunit A9	Cell Proliferation, Metastasis; breast cancer ³³
<i>NFATC3</i>	nuclear factor of activated T cells 3	Tumor growth, cell proliferation and migration; astroglioma ³⁴
<i>NUMB</i>	NUMB endocytic adaptor protein	Invasion, metastasis, migration; melanoma ³⁵ , colon cancer ³⁶
<i>OIP5</i>	Opa interacting protein 5	Growth, metastasis and drug-resistance; bladder cancer ³⁷
<i>PFKFB2</i>	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 2	Glycolysis, cell proliferation; pancreatic cancer ³⁸
<i>PLEKHF2</i>	pleckstrin homology and FYVE domain containing 2	Amplified, Survival; PCa ³⁹
<i>RASIP1</i>	Ras interacting protein 1	Cell migration; non-small-cell lung cancer cells lines ⁴⁰
<i>RNF167</i>	ring finger protein 167	Activates mTORC1 and promotes tumorigenesis; breast and liver cancer cell lines. ⁴¹
<i>SELENBP1</i>	selenium binding protein 1	Tumor growth, progression, survival; lung cancer ⁴² , PCa ⁴³
<i>SLC45A3</i>	solute carrier family 45 member 3	SLC45A3-ERG fusion, survival; PCa ⁴³
<i>SMC4</i>	structural maintenance of chromosomes 4	TGFβ/Smad signaling, cell invasion; glioma cells ⁴⁴ , PCa ⁴⁵
<i>TMEM87A</i>	transmembrane protein 87A	Cell proliferation and metastasis; gastric cancer ⁴⁶
<i>UBXN2B</i>	UBX domain protein 2B	
<i>SRSF10 (SFRS13A)</i>	serine and arginine rich splicing factor 10	Maintenance of oncogenic features; colon cancer cells ⁴⁷
<i>C3orf67</i>	Chromosome 3 open reading frame 67	
<i>C14orf169 (NO66)</i>	Chromosome 14 open reading frame 169	Osteolytic lesions, invasion and metastasis; PCa ⁴⁸
<i>LOC678655 (CD27-AS1)</i>	CD27 antisense RNA 1	Progression; acute myeloid leukemia ⁴⁹

Table 2. Signature genes and their relevance to PCa and/or other cancers.

difference in the average expression levels between *TP53*-mutant and wild-type $GS \geq 7$ samples was performed to scan the ~17,700 genes that remained after the first step. Third, the p-values estimated in the second step were adjusted by the BH method and adj. $p < 0.05$ was used as the cutoff for gene selection.

Transcriptomic signature for BCR. We identified a 25-gene transcriptomic signature from the *TP53*-mut-ass genes for predicting BCR. Coincidentally, the number of member genes was equal to the size of a random signature tested in the second module of our approach. It was also within the range of sizes of PCa prognostic gene signatures identified by other research groups¹²⁻¹⁶. The signature was not enriched with the genes in any gene ontology (GO) term or KEGG pathway, as demonstrated by a functional enrichment analysis using the DAVID tool/database³⁰. However, it was enriched ($p = 0.02$, from Fisher's exact test) with the cancer gene census (downloaded on April 16, 2020) established by the Catalog of Somatic Mutations in Cancer (<https://cancer.sanger.ac.uk/cosmic>). The cancer genes contained in the census included *CDKN1A*, *LARP4B* and *SLC45A3*. Previous studies showed that *LARP4B* inhibited the migration and invasion of prostate cancer cells³², *SLC45A3* downregulation was significantly associated with shorter PSA-free survival times, and the expression of *SLC45A3* protein was downregulated through *SLC45A3-ERG* fusion⁵¹. Moreover, a recent study demonstrated that another 16 genes in the signature were relevant to the formation and progression of prostate cancer and/or other cancers (Table 2). For example, Zhao et al. found that *SMC4* knockdown reduced migration and/or invasion of cancer cells and that outlier expression of the gene was significantly associated with poor PCa prognosis⁴⁵, and Jiang et al. showed that overexpression of *SMC4* activated *TGFβ/Smad* signaling and promoted an aggressive phenotype in glioma cells⁴⁴.

Predicting BCR. The performance of the identified signature for BCR prediction was finally evaluated using the TCGA-PRAD dataset and validated using five external datasets (Fig. 2).

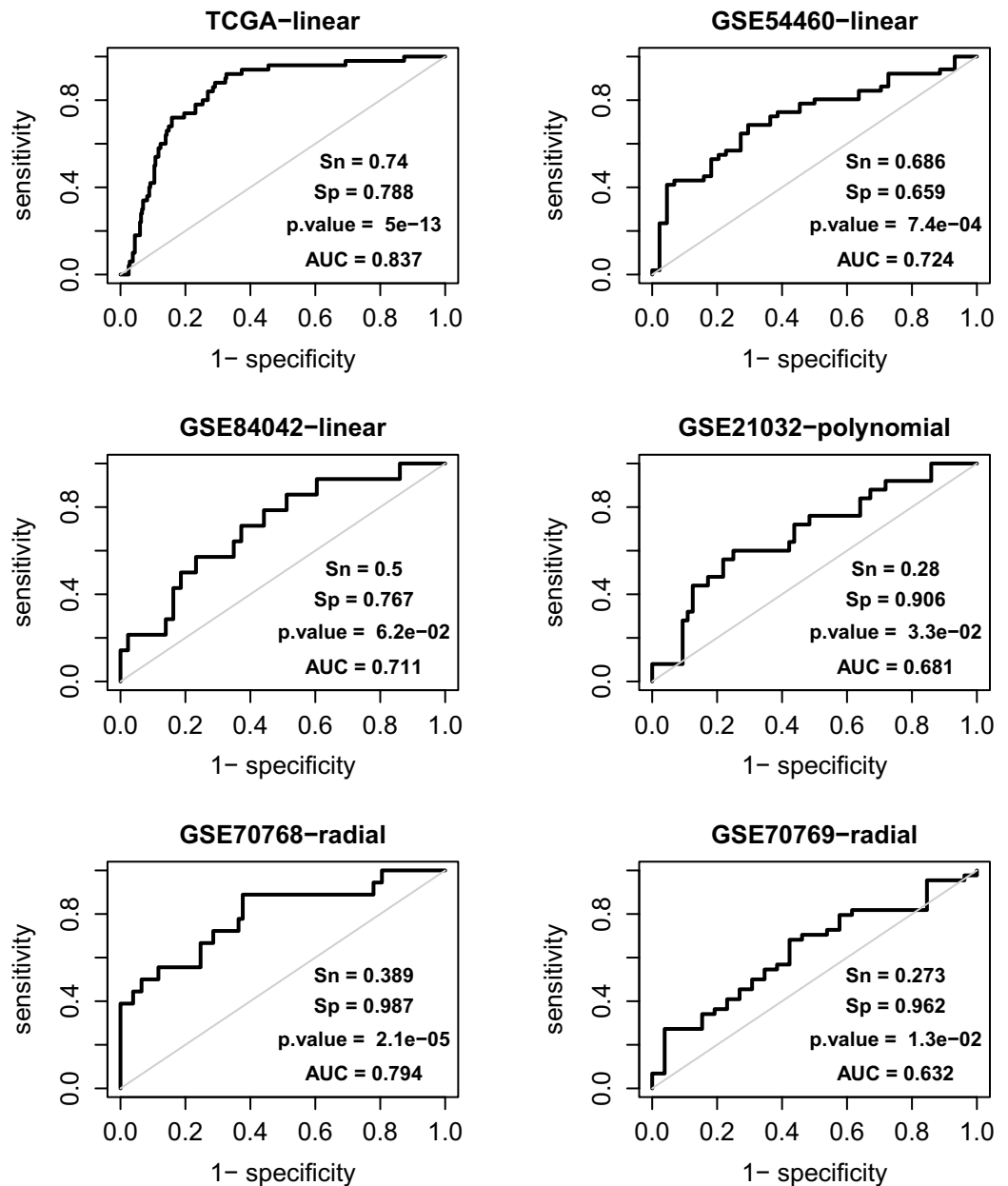


Figure 2. The performance of the TP53 mutation status-associated transcriptomic signature for BCR prediction in the discovery dataset TCGA (-PRAD) and five external datasets, i.e., the GSE54460 and others. The “-linear”, “-polynomial” and “-radial” indicate the kernel functions used in the SVM models. The output BCR label and decision value, i.e. the transcriptomic risk score (TRS) of a patient in GSE70769 was predicted by the model trained using the GSE70768 dataset. For the patients in other cohorts, the labels and scores were predicted via LOOCV. Together with the actual BCR labels, the output BCR labels and TRSs are used to calculate a 2×2 contingency table for estimating the p-value and to generate the ROC curve, respectively. Sn and Sp denote sensitivity and specificity, respectively.

While SVM-LOOCV, a strategy appropriate for accurately estimating the performance of a classification model, was adopted, testing on the TCGA-PRAD cohort was subject to overfitting. Namely, the obtained prediction strength (AUC=0.837; p-value= 5×10^{-13}) could be overestimated for the patient population(s) represented by this cohort. The reason was that the gene signature was discovered in the same dataset. In this regard, we performed a complementary analysis to verify the prediction robustness, i.e., to demonstrate that the observed performance was not due to the specific data body exactly consisting of the expression profiling and BCR statuses of all 366 focused samples. Briefly, 500 working sets with size $n=330$ (366×0.9) were generated from the TCGA-PRAD dataset by non-replacement sampling, and on each of them, the SVM-LOOCV AUC was estimated after its expression matrix was altered by artificial noise. The noises were introduced by a two-step procedure: (i) the expression levels of the signature genes were rescaled by z-transformation, and (ii) random

noises $x \sim \text{norm}(0, 0.1)$ were added to the standardized expression metrics (in this setting, the variance of the introduced noise was equal to 10% data variance). The results from this analysis showed that the 0.05 and 0.95 quantiles of the 500 AUC values were 0.688 and 0.803, respectively, and only 0.6% of them were less than 0.65.

The results from GSE54460 and GSE84042 clearly validated the signature. First, both of the two SVM-LOOCV AUC values were over 0.70. Second, in establishing the SVM model, a simple linear kernel was used, and no hyperparameters were tuned. For the GSE21032 dataset, the output from the polynomial-kernel (with defaults for other hyperparameters) SVM model was somewhat better than that from the linear-kernel model (0.681 vs 0.652). For the GSE70768 and GSE70769 datasets, tuned hyperparameters (i.e., *cost* and *gamma*) were required to obtain prediction strength. In such a scenario, the final performance had to be assessed on an independent dataset. Due to the limited cohort sizes and the imbalance in BCR statuses, we could not partition either of the GSE70768 and GSE70769 datasets into a substantial training subset and a substantial testing subset. As such, we addressed these two datasets by considering the former as the training set and the latter as the testing set, the same setting used in the original study¹⁶. With the optimized *cost* and *gamma* (0.105 and two-thirds the inverse of the feature dimension), the training set had an SVM-LOOCV AUC of 0.794 and a p-value of 2.1×10^{-5} . When the model trained with all the samples in the training set was used to predict the samples in the testing set, the external validation AUC and p-value were 0.632 and 1.3×10^{-2} , respectively.

Predicting RFS. We further evaluated the prognostic performance of the identified signature by testing the association between the stratification of disease relapse-free survival and the predicted BCR partition via SVM-LOOCV. Suppose pre-BCR⁺ and pre-BCR⁻ represent the predicted “positive” and “negative” groups, respectively. The results of survival analysis showed that for each cohort, the Kaplan–Meier curve of the pre-BCR⁻ group was better than the curve of pre-BCR⁺ (Fig. 3). The p-values from the log-rank test of the differences ranged from 2.8×10^{-2} (for the GSE84042 cohort) to 3.3×10^{-13} (for the TCGA-PRAD cohort).

Practical utility: compared with clinicopathologic factors. We deciphered the practical utility of the identified gene signature by comparing its capability to stratify patient RFS with that of clinicopathologic prognostic factors (CPFs). The considered CPFs consisted of Gleason patterns (GPs), T stages (T1, T2, T3) and patient ages at initial PCa diagnosis. In particular, regarding GPs, we partitioned patients into three categories, i.e., 3+4, 4+3 and the other. On each dataset (cohort), the analysis was conducted by running three Cox-PH regression models. The first (M-1) included the transcriptomic risk score (TRS), which was estimated by the decision value outputted from SVM-LOOCV for individual subjects, as the only explanatory variable. The second (M-2) and third (M-3) included the three CPFs or both the TRS and CPFs as the explanatory variables, respectively. The performance and validity of a model were assessed by the explained variation (R^2), BIC and global log rank p-value.

Except for M2 in GSE84042, all three models were significant ($p < 0.02$) in the six cohorts. M-1 performed better than M-2 in GSE54460, GSE84042 and GSE70768 in terms of the higher R^2 values and/or lower BIC values but was poorer in the other three cohorts. All the R^2 values from M-3 were higher than those from both M-1 and M-2, while its BIC value in GSE21032 was slightly higher than the score of M-2. From the statistics itemized in Table 3, we derived two conclusions about the prognostic utility of our gene signature. First, TRS could explain 9–60% of the variation in RFS, with an average of 28.2%. Second, TRS could replace and/or complement CPFs to predict RFS, and the combination of TRS and CPFs could explain 23–72% of the variation in RFS, with a median of 54.5%.

Reevaluating the documented signatures. To demonstrate the relative advantage of our signature, we reevaluated the prognostic performance of five transcriptomic signatures identified by other researchers. Here, we provide a brief description of those signatures before discussing the results. The methods used in the identification are outlined in Supplementary Text 1.

Wu’s signature. The 10-gene signature was identified for predicting BCR in the $GS \geq 7$ patient set¹². The dataset of 414 TCGA-PRAD prostate adenocarcinoma samples (including 37 $GS = 6$ samples) was used as the discovery (training) set.

Li’s signature. The signature consisted of 74 gene pairs from a combination of 60 genes¹³. It was identified to predict BCR, regardless of Gleason scores or patterns. The entire GSE21032 ($N = 131$) cohort was used as the discovery set. The 60 genes were considered individual features in our analysis. Li et al.’s work also included the analysis of two datasets (i.e., GSE46602 and GSE40272) that were not addressed in our study. The reasons were that there were only 20 $GS \geq 7$ samples in the former, and Gleason score information is unavailable in the latter.

Komisarof’s signature. The signature included four cooperation response (to oncogenic mutations) genes^{14,52}. It was identified to predict BCR, regardless of Gleason scores or patterns. The discovery cohort consisted of 32 samples.

Erho’s signature. The signature consisted of 22 features (genome fragments) located on the coding or noncoding regions of 19 genes¹⁵. It was identified to predict early prostate cancer metastasis and is being used in the commercial Decipher Prostate Genomic Test (<https://decipherbio.com/>). The discovery cohorts contained 359 samples.

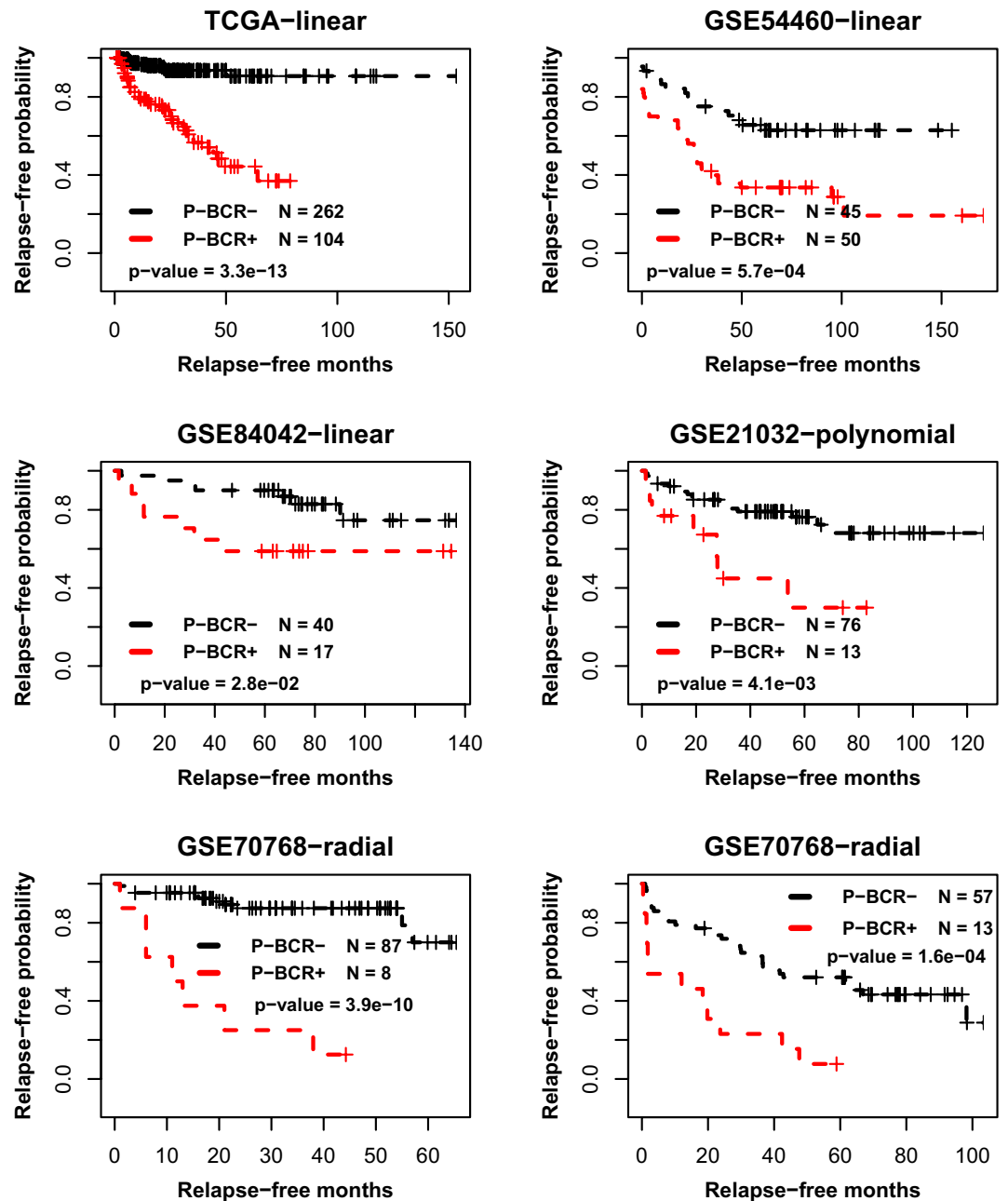


Figure 3. The association between RFS stratification and the BCR partition predicted using the TP53 mutation status-associated prognostic transcriptomic signature in the discovery dataset TCGA (-PRAD) and five external datasets, i.e., the GSE54460 and others. The “-linear”, “-polynomial” and “-radial” indicate the kernel functions used in the SVM models. The output BCR label (pre-BCR⁺ and pre-BCR⁻) of a patient in GSE70769 is predicted by the model trained using the GSE70768 dataset. For the patients in other cohorts, the labels are predicted via LOOCV. The survival profiles of pre-BCR⁺ and pre-BCR⁻ samples are depicted by red and black Kaplan-Meier curves, respectively.

Knezevic-Klein’s signature. The 17-gene signature was identified to predict clinical recurrence, prostate cancer death, and adverse pathology^{53,54}, and is being used in the commercial Oncotype DX Genomic Prostate Score Test (<https://www.oncotypeiq.com/en-US>). The discovery cohort consisted of 441 patients.

Performance summary. Among these five signatures, only Knezevic-Klein’s signature demonstrated prediction strength for BCR on all the six datasets in the SVM-LOOCV reevaluation (Supplementary Fig. S1, S2, S3, S4, S5). In particular, the BCR statuses of the samples in GSE84042 could not be predicted by any one of the other four signatures. However, Komisarof’s signature in GSE54460 and Li’s signature in GSE21032 had higher AUCs (0.766 vs 0.724 and 0.859 vs 0.681) and lower p-values (1.2×10^{-4} vs 7.4×10^{-4} and 1.2×10^{-6} vs 3.3×10^{-2}) than our signature, respectively. Here, we specifically noted the following three points. First, due to the preceding

Data ID	M-1, TRS			M-2, CPFs			M-3, TRS + CPFs		
	R ²	BIC [‡]	p-value	R ²	BIC	p-value	R ²	BIC	p-value
TCGA-PRAD	0.43	475.18	1.1 × 10 ⁻¹⁰	0.44	421.18	5.8 × 10 ⁻⁷	0.63	403.45	2.2 × 10 ⁻¹¹
GSE54460	0.12	417.59	5.8 × 10 ⁻⁴	0.16	429.32	7.5 × 10 ⁻³	0.23	426.79	7.6 × 10 ⁻⁴
GSE84042	0.36	99.29	2.2 × 10 ⁻³	0.17	109.63	2.6 × 10 ⁻¹	0.51	104.67	1.3 × 10 ⁻²
GSE21032	0.09	205.6	7.4 × 10 ⁻³	0.58	192.68	8.5 × 10 ⁻⁹	0.58	195.52	2.1 × 10 ⁻⁸
GSE70768	0.60	126.27	9.2 × 10 ⁻⁸	0.41	143.26	1.9 × 10 ⁻³	0.72	130.37	9.8 × 10 ⁻⁸
GSE70769	0.09	327.67	9.0 × 10 ⁻³	0.17	308.64	9.9 × 10 ⁻³	0.25	306.2	1.2 × 10 ⁻³

Table 3. Results from Cox regression model analysis. † † The three models (M-1, M-2 and M-3) are specified by the included predictor variable(s) for cancer relapse-free survival. TRS: transcriptomic risk score. CPFs: clinicopathologic prognostic factors. See the main text for a more detailed description. ‡ BIC: Bayesian Information Criterion.

information leak in the identification process, the good performance of Li's signature in GSE21032 might be overestimated. Second, when Li's signature was applied to the GSE21032 dataset, the polynomial-kernel SVM worked better than the linear-kernel SVM, similar to the situation in evaluating our signature. Third, the modest significant performance (AUC = 0.627 or 0.644, and $p < 0.1$) of Wu's and Erho's signatures on the GSE70768 dataset was obtained after the post hoc optimization of the hyperparameter *cost* and *gamma* of the SVM model, again similar to the situation in evaluating our signature. The last two points indirectly verified the reliability of the observed performance of our signature on these two datasets.

Extended analysis. In this subsection, we demonstrated that the approach proposed in this study was also efficient for identifying prognostic signatures from the gene set defined by a specific cancer progression-related biological theme. The activity of genes involved in immunology pathways is such a theme^{55–57}.

We identified an immune signature of 16 genes (*SPARC*, *IFNAR2*, *FOXQ1*, *G3BP1*, *IKBIP*, *BAT1*, *AZINI*, *ZDHHC17*, *RRAS*, *DOK7*, *DMRTA1*, *ACTG1*, *AGFG1*, *M6PR*, *MED7*, and *PSAPL1*) from ~ 1200 immune function- or regulation-related genes⁵⁸. A couple of minor modifications were made in the method implementation: (i) the number of random signatures to be tested was increased from 1000 to 2000 and (ii) a gene to be selected to the initial signature had to be included in 2 (rather than 3) random signatures that met the criteria of AUC (> 0.65) and p-value (< 0.0001) in predicting the BCR statuses of the TCGA-PRAD samples.

This immune signature was evaluated on the six datasets by the same methods and model settings as used in the aforementioned tests. The results showed that the expression profiles of the signature genes could predict both BCR and RFS (Supplementary Fig. S6, S7). Except for GSE84042, in which it lacked prediction strength, its performance was comparable with the signature extracted from the *TP53*-mut-ass genes.

Discussion

In this study, using the data generated by the TCGA-PRAD group, we identified a transcriptomic signature to predict BCR in patients with Gleason score ≥ 7 prostate cancer according to gene expression levels measured from prostatectomy specimens. The 25-gene signature was a small portion of 720 genes that were differentially expressed (FDR < 0.05) between the samples with somatic *TP53* mutation(s) and those without *TP53* mutation. However, the prognostic signature is not tied to the mutation statuses and/or expression levels of *TP53* itself and *TP53* mutation analysis is not a pre-requisite for the expected utilization. The signature was evaluated on the discovery dataset and five external datasets, demonstrating robust prognostic performance not only for predicting BCR but also for stratifying RFS. The risk scores derived from the signature by SVM-LOOCV explained 9–63% of variation in RFS and could complement clinicopathologic prognostic factors.

The advantage of our *TP53* mutation status-associated signature was clearly shown in the comparison with those (i.e., Li's, Wu's, Komisarof's, Erho's and Knezevic-Klein's) presented in recent literature and/or used in commercial prognostic tools regarding their performance on the same datasets. While Knezevic-Klein's signature could predict BCR in all the six datasets, its performance was somewhat poorer than our signature in terms of AUCs and p-values, in general. Nevertheless, the reevaluation results of the five external signatures should be carefully scrutinized. For example, because Erho's signature was identified to predict metastasis, which is preceded by BCR but is not equivalent to BCR, its deficient performance in predicting BCR should not alleviate the potential prognostic utility. Moreover, in the original study, Li's signature was identified and assessed by their paired-gene expression-based tree model, which might partially explain the poor robustness when it was tested using SVM-LOOCV.

The member genes of our signature were complementary to each other in predicting BCR, although no enrichment relationship was found between them (as a while) and any cancer pathway. Previous studies reported that the activity (expression) of 80% of these genes was related to cancer cell invasion, cancer progression and/or patient outcomes (Table 2). However, according to our additional analysis with univariate regression models, none of them alone could consistently predict BCR and/or RFS in all the six datasets. Even the combination of the five genes, including *LARP4B*, *PLEKHF2*, *SMC4*, *SLC45A3* and *NO66*^{32,39,44,45,48,51}, whose clinical relevance and prognostic implications were observed in prostate cancer, had very limited prediction strength for BCR and RFS (not reported in the Results section). In this regard, we perceived that due to the complicated clinical, genetic, pathological and demographical heterogeneity of prostate cancer, a patient cohort may have specific

gene expression prognostic factors, and a robust transcriptomic signature should include the genes that could cover such heterogeneity.

Our signature was identified using the data-driven and biologically informed stochastic approach developed in this study. It was characterized with a couple of points. First, a biological information source (i.e., somatic *TP53* mutation profile) critical to cancer progression was used to select a set of candidate signature genes. Naturally, this was a step of introducing external information (knowledge), which has been adopted in our and others' published studies^{59,60}. Second, the ranks of candidate genes regarding their differences in expression levels between BCR-positive and BCR-negative samples were not considered in any steps. Alternatively, the chance of a gene being selected into the initial (and final) signature was determined by the post hoc performance of the random signatures that contained it. As such, the member genes of the final signature did not exclusively consist of the discovery set-specific top BCR-distinguishing genes. Our expectation that such a design could reduce the risk of overfitting was verified by the aforementioned results.

Another novel component of our approach is the algorithm denoted by “wrapper”, which was designed to filter out the genes that erode the prediction strength that can be expected from a reduced signature without them. The wrapper somewhat resembles the backward variable selection (BVS) procedure usually adopted by a multivariable regression analysis, in which the initially selected features (explanatory variables) are ranked based on their statistical significance level for explaining the outcome variable and the feature-dropping process begins from the least significant one. Meanwhile, it conceptually differentiates from BVS in that the benefit from dropping a potentially redundant feature is assessed by the AUC gain outputted from SVM-LOOCV rather than the R^2 (or its variants) and $-\log_{10}(\text{p-value})$ gain from fitting a reduced (linear) model. Therefore, the wrapper actually combines the model optimization of a regression analysis and the model validation step that is highly desired for avoiding overfitting. In this aspect, the wrapper is more similar to the recursive feature elimination (RFE) algorithm proposed in Ref⁶¹. However, in RFE, redundant features are iteratively dropped due to their relatively minor contribution to the SVM classifier rather than their negative impact on the expected prediction strength.

In the “Extended analysis” subsection of the Results section, we identified an alternative prognostic signature via a modified implementation of our approach, which consisted of 16 immune-related genes. By this, we demonstrated that the approach was also efficient for extracting prognostic signatures from a gene set under a specific cancer progression-related biological theme. Here, we further note that it can be adapted as a general feature selection method, contributing to more general applications of high-throughput data such as the molecular prediction of the subtypes and progression stages of a disease. The adaptation can be worked out by replacing the *TP53*-mut-ass gene set, or immune gene set, with any one that is enriched with the molecular diagnostic/prognostic factors regarding the focused clinical trait in a specific study.

Data availability

The used TCGA and GEO datasets reside at <https://portal.gdc.cancer.gov/> and <https://www.ncbi.nlm.nih.gov/geo/>, respectively.

Received: 5 November 2021; Accepted: 7 June 2022

Published online: 22 June 2022

References

- Paller, C. J. & Antonarakis, E. S. Management of biochemically recurrent prostate cancer after local therapy: Evolving standards of care and new directions. *Clin. Adv. Hematol. Oncol.* **11**, 14–23 (2013).
- Popiolek, M. *et al.* Natural history of early, localized prostate cancer: a final report from three decades of follow-up. *Eur. Urol.* **63**, 428–435. <https://doi.org/10.1016/j.eururo.2012.10.002> (2013).
- Pound, C. R. *et al.* Natural history of progression after PSA elevation following radical prostatectomy. *JAMA* **281**, 1591–1597. <https://doi.org/10.1001/jama.281.17.1591> (1999).
- Leibovici, D. *et al.* Prostate cancer progression in the presence of undetectable or low serum prostate-specific antigen level. *Cancer* **109**, 198–204. <https://doi.org/10.1002/cncr.22372> (2007).
- Lavery, H. J. & Droller, M. J. Do Gleason patterns 3 and 4 prostate cancer represent separate disease states?. *J. Urol.* **188**, 1667–1675. <https://doi.org/10.1016/j.juro.2012.07.055> (2012).
- Stamey, T. A., McNeal, J. E., Yemoto, C. M., Sigal, B. M. & Johnstone, I. M. Biological determinants of cancer progression in men with prostate cancer. *JAMA* **281**, 1395–1400. <https://doi.org/10.1001/jama.281.15.1395> (1999).
- Cheng, L., Davidson, D. D., Lin, H. & Koch, M. O. Percentage of Gleason pattern 4 and 5 predicts survival after radical prostatectomy. *Cancer* **110**, 1967–1972. <https://doi.org/10.1002/cncr.23004> (2007).
- Ross, H. M. *et al.* Do adenocarcinomas of the prostate with Gleason score (GS) ≤ 6 have the potential to metastasize to lymph nodes?. *Am. J. Surg. Pathol.* **36**, 1346–1352. <https://doi.org/10.1097/PAS.0b013e3182556dcd> (2012).
- Esserman, L. J. *et al.* Addressing overdiagnosis and overtreatment in cancer: a prescription for change. *Lancet Oncol.* **15**, e234–242. [https://doi.org/10.1016/S1470-2045\(13\)70598-9](https://doi.org/10.1016/S1470-2045(13)70598-9) (2014).
- Gharzai, L. A. *et al.* Intermediate clinical endpoints for surrogacy in localised prostate cancer: an aggregate meta-analysis. *Lancet Oncol.* **22**, 402–410. [https://doi.org/10.1016/S1470-2045\(20\)30730-0](https://doi.org/10.1016/S1470-2045(20)30730-0) (2021).
- Beauval, J. B. *et al.* Biochemical recurrence-free survival and pathological outcomes after radical prostatectomy for high-risk prostate cancer. *BMC Urol.* **16**, 26. <https://doi.org/10.1186/s12894-016-0146-6> (2016).
- Wu, X. *et al.* A 10-gene signature as a predictor of biochemical recurrence after radical prostatectomy in patients with prostate cancer and a Gleason score ≥ 7 . *Oncol Lett.* **20**, 2906–2918. <https://doi.org/10.3892/ol.2020.11830> (2020).
- Li, X. *et al.* A qualitative transcriptional signature for predicting the biochemical recurrence risk of prostate cancer patients after radical prostatectomy. *Prostate* **80**, 376–387. <https://doi.org/10.1002/pros.23952> (2020).
- Komisarov, J. *et al.* A four gene signature predictive of recurrent prostate cancer. *Oncotarget* **8**, 3430–3440. <https://doi.org/10.18632/oncotarget.13837> (2017).
- Erho, N. *et al.* Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy. *PLoS ONE* **8**, e66855. <https://doi.org/10.1371/journal.pone.0066855> (2013).
- Ross-Adams, H. *et al.* Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine* **2**, 1133–1144. <https://doi.org/10.1016/j.ebiom.2015.07.017> (2015).

17. Cancer Genome Atlas Research, N. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* **163**, 1011–1025. <https://doi.org/10.1016/j.cell.2015.10.025> (2015).
18. Robinson, D. *et al.* Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell* **162**, 454. <https://doi.org/10.1016/j.cell.2015.06.053> (2015).
19. Royston, P. Explained variation for survival models. *Stand. Genom. Sci.* **6**, 83–96 (2006).
20. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinf.* **12**, 323. <https://doi.org/10.1186/1471-2105-12-323> (2011).
21. Long, Q. *et al.* Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer Res.* **74**, 3228–3237. <https://doi.org/10.1158/0008-5472.CAN-13-2699> (2014).
22. Fraser, M. *et al.* Genomic hallmarks of localized, non-indolent prostate cancer. *Nature* **541**, 359–364. <https://doi.org/10.1038/nature20788> (2017).
23. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264. <https://doi.org/10.1093/biostatistics/4.2.249> (2003).
24. Taylor, B. S. *et al.* Integrative genomic profiling of human prostate cancer. *Cancer Cell* **18**, 11–22. <https://doi.org/10.1016/j.ccr.2010.05.026> (2010).
25. Kibel, A. S. *et al.* CDKN1A and CDKN1B polymorphisms and risk of advanced prostate carcinoma. *Cancer Res.* **63**, 2033–2036 (2003).
26. Zhang, Y. *et al.* Role of Damage DNA-Binding Protein 1 in Pancreatic Cancer Progression and Chemoresistance. *Cancers (Basel)* **11**, 1. <https://doi.org/10.3390/cancers11121998> (2019).
27. Yamaji, S. *et al.* Hepatocyte-specific deletion of DDB1 induces liver regeneration and tumorigenesis. *Proc. Natl. Acad. Sci. USA* **107**, 22237–22242. <https://doi.org/10.1073/pnas.1015793108> (2010).
28. Liu, Z. *et al.* TTF-1/NKX2-1 binds to DDB1 and confers replication stress resistance to lung adenocarcinomas. *Oncogene* **36**, 3740–3748. <https://doi.org/10.1038/ncr.2016.524> (2017).
29. Cao, T. T. *et al.* Eukaryotic translation initiation factor 5A2 promotes metabolic reprogramming in hepatocellular carcinoma cells. *Carcinogenesis* **38**, 94–104. <https://doi.org/10.1093/carcin/bgw119> (2017).
30. Fujimura, K. *et al.* A hypusine-eIF5A-PEAK1 switch regulates the pathogenesis of pancreatic cancer. *Cancer Res.* **74**, 6671–6681. <https://doi.org/10.1158/0008-5472.CAN-14-1031> (2014).
31. Porkka, K. P., Tammela, T. L., Vessella, R. L. & Visakorpi, T. RAD21 and KIAA0196 at 8q24 are amplified and overexpressed in prostate cancer. *Genes Chromosomes Cancer* **39**, 1–10. <https://doi.org/10.1002/gcc.10289> (2004).
32. Seetharaman, S., Flemming, E., Shen, J., Conte, M. R. & Ridley, A. J. The RNA-binding protein LARP4 regulates cancer cell migration and invasion. *Cytoskeleton (Hoboken)* **73**, 680–690. <https://doi.org/10.1002/cm.21336> (2016).
33. Li, L. D. *et al.* Down-Regulation of NDUFB9 Promotes Breast Cancer Cell Proliferation, Metastasis by Mediating Mitochondrial Metabolism. *PLoS ONE* **10**, e0144441. <https://doi.org/10.1371/journal.pone.0144441> (2015).
34. Urso, K. *et al.* NFATc3 controls tumour growth by regulating proliferation and migration of human astroglia cells. *Sci. Rep.* **9**, 9361. <https://doi.org/10.1038/s41598-019-45731-w> (2019).
35. Hristova, D. M. *et al.* Numb as a Therapeutic Target for Melanoma. *J. Invest. Dermatol.* <https://doi.org/10.1016/j.jid.2021.11.027> (2021).
36. Yang, Y. *et al.* Numb inhibits migration and promotes proliferation of colon cancer cells via RhoA/ROCK signaling pathway repression. *Exp. Cell Res.* **411**, 113004. <https://doi.org/10.1016/j.yexcr.2021.113004> (2022).
37. Wang, D. *et al.* OIP5 Promotes Growth, Metastasis and Chemoresistance to Cisplatin in Bladder Cancer Cells. *J. Cancer* **9**, 4684–4695. <https://doi.org/10.7150/jca.27381> (2018).
38. Ozcan, S. C. *et al.* PFKFB2 regulates glycolysis and proliferation in pancreatic cancer cells. *Mol. Cell Biochem.* **470**, 115–129. <https://doi.org/10.1007/s11010-020-03751-5> (2020).
39. Shamsara, E. & Shamsara, J. Bioinformatics analysis of the genes involved in the extension of prostate cancer to adjacent lymph nodes by supervised and unsupervised machine learning methods: The role of SPAG1 and PLEKHF2. *Genomics* **112**, 3871–3882. <https://doi.org/10.1016/j.ygeno.2020.06.035> (2020).
40. Chen, Y. *et al.* Rasip1 is a RUNX1 target gene and promotes migration of NSCLC cells. *Cancer Manag. Res.* **10**, 4537–4552. <https://doi.org/10.2147/CMAR.S168438> (2018).
41. Li, T. *et al.* RNF167 activates mTORC1 and promotes tumorigenesis by targeting CASTOR1 for ubiquitination and degradation. *Nat. Commun.* **12**, 1055. <https://doi.org/10.1038/s41467-021-21206-3> (2021).
42. Caswell, D. R. *et al.* Tumor Suppressor Activity of Selenbp1, a Direct Nkx2-1 Target, Lung Adenocarcinoma. *Mol. Cancer Res.* **16**, 1737–1749. <https://doi.org/10.1158/1541-7786.MCR-18-0392> (2018).
43. Ansong, E. *et al.* Evidence that selenium binding protein 1 is a tumor suppressor in prostate cancer. *PLoS ONE* **10**, e0127295. <https://doi.org/10.1371/journal.pone.0127295> (2015).
44. Jiang, L. *et al.* Overexpression of SMC4 activates TGFbeta/Smad signaling and promotes aggressive phenotype in glioma cells. *Oncogenesis* **6**, e301. <https://doi.org/10.1038/oncsis.2017.8> (2017).
45. Zhao, S. G. *et al.* The Landscape of Prognostic Outlier Genes in High-Risk Prostate Cancer. *Clin. Cancer Res.* **22**, 1777–1786. <https://doi.org/10.1158/1078-0432.CCR-15-1250> (2016).
46. Wang, H. *et al.* Circular RNA TMEM87A promotes cell proliferation and metastasis of gastric cancer by elevating ULK1 via sponging miR-142-5p. *J. Gastroenterol.* **56**, 125–138. <https://doi.org/10.1007/s00535-020-01744-1> (2021).
47. Zhou, X. *et al.* BCLAF1 and its splicing regulator SRSF10 regulate the tumorigenic potential of colon cancer cells. *Nat. Commun.* **5**, 4581. <https://doi.org/10.1038/ncomms5581> (2014).
48. Sinha, K. M. *et al.* Oncogenic and osteolytic functions of histone demethylase NO66 in castration-resistant prostate cancer. *Oncogene* **38**, 5038–5049. <https://doi.org/10.1038/s41388-019-0774-x> (2019).
49. Tao, Y. *et al.* LncRNA CD27-AS1 promotes acute myeloid leukemia progression through the miR-224-5p/PBX3 signaling circuit. *Cell Death Dis.* **12**, 510. <https://doi.org/10.1038/s41419-021-03767-9> (2021).
50. da Huang, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57. <https://doi.org/10.1038/nprot.2008.211> (2009).
51. Perner, S. *et al.* Loss of SLC45A3 protein (prostein) expression in prostate cancer is associated with SLC45A3-ERG gene rearrangement and an unfavorable clinical course. *Int. J. Cancer* **132**, 807–812. <https://doi.org/10.1002/ijc.27733> (2013).
52. McMurray, H. R. *et al.* Synergistic response to oncogenic mutations defines gene class critical to cancer phenotype. *Nature* **453**, 1112–1116. <https://doi.org/10.1038/nature06973> (2008).
53. Klein, E. A. *et al.* A 17-gene assay to predict prostate cancer aggressiveness in the context of Gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur. Urol.* **66**, 550–560. <https://doi.org/10.1016/j.eururo.2014.05.004> (2014).
54. Knezevic, D. *et al.* Analytical validation of the Oncotype DX prostate cancer assay—A clinical RT-PCR assay optimized for prostate needle biopsies. *BMC Genom.* **14**, 690. <https://doi.org/10.1186/1471-2164-14-690> (2013).
55. Rui, X., Shao, S., Wang, L. & Leng, J. Identification of recurrence marker associated with immune infiltration in prostate cancer with radical resection and build prognostic nomogram. *BMC Cancer* **19**, 1179. <https://doi.org/10.1186/s12885-019-6391-9> (2019).
56. Bedognetti, D., Hendrickx, W., Marincola, F. M. & Miller, L. D. Prognostic and predictive immune gene signatures in breast cancer. *Curr. Opin. Oncol.* **27**, 433–444. <https://doi.org/10.1097/CCO.000000000000234> (2015).

57. Babar, L. *et al.* Prognostic immune markers for recurrence and survival in locally advanced esophageal adenocarcinoma. *Oncotarget* **10**, 4546–4555. <https://doi.org/10.18632/oncotarget.27052> (2019).
58. Awasthi, S. *et al.* Comparative genomics reveals distinct immune-oncologic pathways in african american men with prostate cancer. *Clin. Cancer Res.* **27**, 320–329. <https://doi.org/10.1158/1078-0432.CCR-20-2925> (2021).
59. Xu, F. *et al.* A TP53-associated gene signature for prediction of prognosis and therapeutic responses in lung squamous cell carcinoma. *Oncoimmunology* **9**, 1731943. <https://doi.org/10.1080/2162402X.2020.1731943> (2020).
60. Zhang, W., Flemington, E. K. & Zhang, K. Driver gene mutations based clustering of tumors: methods and applications. *Bioinformatics* **34**, i404–i411. <https://doi.org/10.1093/bioinformatics/bty232> (2018).
61. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl. Acad. Sci. USA* **98**, 15149–15154. <https://doi.org/10.1073/pnas.211566398> (2001).

Acknowledgements

This research is supported by the NIH grant U54MD007595 (WZ and KZ). The authors thank the editor and reviewers for their constructive comments, which greatly improved this paper.

Author contributions

Study conceiving and designing: W.Z., K.Z. Experiments performing: W.Z. Data analysis: W.Z., K.Z. Writing: W.Z., K.Z. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-14436-y>.

Correspondence and requests for materials should be addressed to W.Z. or K.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022