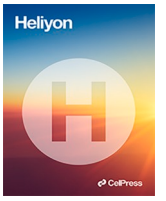
Heliyon 10 (2024) e28878

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Predicting prostate cancer recurrence: Introducing PCRPS, an advanced online web server

Xianya He [1], Sheng Hu [1], Chen Wang, Yongjun Yang, Zhuo Li, Mingqiang Zeng, Guangqing Song, Yuanwei Li **, Qiang Lu *

Department of Urology, Hunan Provincial People's Hospital (The 1st Affiliated Hospital of Hunan Normal University), China

ARTICLE INFO

Keywords:
Prostate cancer
Recurrence
Machine learning
Online web server

ABSTRACT

Background: Prostate cancer (PCa) is one of the leading causes of cancer death in men. About 30% of PCa will develop a biochemical recurrence (BCR) following initial treatment, which significantly contributes to prostate cancer-related deaths. In clinical practice, accurate prediction of PCa recurrence is crucial for making informed treatment decisions. However, the development of reliable models and biomarkers for predicting PCa recurrence remains a challenge. In this study, the aim is to establish an effective and reliable tool for predicting the recurrence of PCa.

Methods: We systematically screened and analyzed potential datasets to predict PCa recurrence. Through quality control analysis, low-quality datasets were removed. Using meta-analysis, differential expression analysis, and feature selection, we identified key genes associated with recurrence. We also evaluated 22 previously published signatures for PCa recurrence prediction. To assess prediction performance, we employed nine machine learning algorithms. We compared the predictive capabilities of models constructed using clinical variables, expression data, and their combinations. Subsequently, we implemented these machine learning models into a user-friendly web server freely accessible to all researchers.

Results: Based on transcriptomic data derived from eight multicenter studies consisting of 733 PCa patients, we screened 23 highly influential genes for predicting prostate cancer recurrence. These genes were used to construct the Prostate Cancer Recurrence Prediction Signature (PCRPS). By comparing with 22 published signatures and four important clinicopathological features, the PCRPS exhibited a robust and significantly improved predictive capability. Among the tested algorithms, Random Forest demonstrated the highest AUC value of 0.72 in predicting PCa recurrence in the testing dataset. To facilitate access and usage of these machine learning models by all researchers and clinicians, we also developed an online web server (https://urology1926.shinyapps.io/PCRPS/) where the PCRPS model can be freely utilized. The tool can also be used to (1) predict the PCa recurrence by clinical information or expression data with high accuracy. (2) provide the possibility of PCa recurrence by nine machine learning algorithms. Furthermore, using the PCRPS scores, we predicted the sensitivity of 22 drugs from GDSC2 and 95 drugs from

* Corresponding author. Department of Urology, Hunan Provincial People's Hospital (The 1st Affiliated Hospital of Hunan Normal University), Changsha, China.

** Corresponding author. Department of Urology, Hunan Provincial People's Hospital (The 1st Affiliated Hospital of Hunan Normal University), Changsha, China.

E-mail addresses: liyuanwei@hunnu.edu.cn (Y. Li), urology@hunnu.edu.cn (Q. Lu).

[1] Xianya He and Sheng Hu contributed equally to this research.

https://doi.org/10.1016/j.heliyon.2024.e28878
Received 27 August 2023; Received in revised form 25 March 2024; Accepted 26 March 2024
Available online 29 March 2024

CTRP2 to the samples. These predictions provide valuable insights into potential drug sensitivities related to the PCRPS score groups.
Conclusion: Overall, our study provides an attractive tool to further guide the clinical management and individualized treatment for PCa.

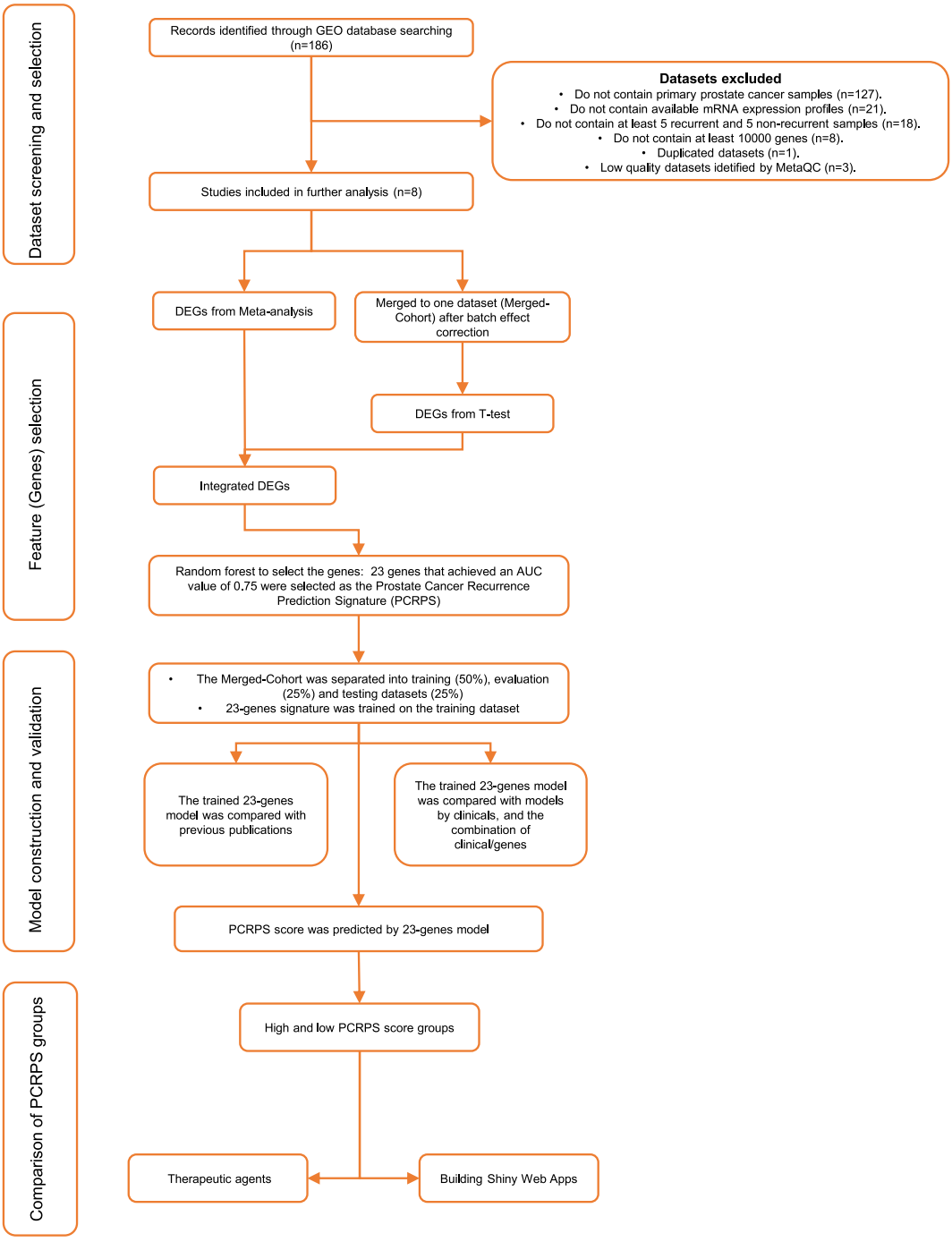


Fig. 1. The workflow of our research.

1. Introduction

Prostate cancer is a significant global public health issue, representing the most common cancer in men worldwide. It affects approximately 1.6 million individuals and ranks as the seventh highest cause of cancer-related mortality among men, resulting in 366,000 deaths globally. As an example, in the year 2019, the United States reported 174,650 new cases of prostate cancer and 31,620 deaths attributed to prostate cancer [1]. Furthermore, the incidence rates of prostate cancer in Asian countries are experiencing a rapid increase [2]. Notably, the disease trajectory of prostate cancer can vary significantly. On one hand, localized prostate cancer exhibits a high five-year survival rate, approaching 100%, primarily managed through surgery [3]. On the other hand, metastatic prostate cancer presents a stark contrast with a 31% survival rate [4]. However, despite the relatively successful management of localized prostate cancer through surgery, such as radical prostatectomy, a considerable proportion of patients, ranging from 20% to 40%, experience biochemical recurrence (BCR) [5], characterized by an increase in serum prostate-specific antigen (PSA) [6]. Furthermore, a significant proportion of BCR tumors are likely to progress to metastatic prostate cancers, which significantly contribute to prostate cancer-related deaths. Hence, there is a pressing demand to construct a new prostate cancer recurrence prediction model that better accounts for enhancing the quality of patient care and outcomes.

Clinical parameters such as the Gleason system, pathological tumor staging, and PSA levels are known to impact the survival and risk of BCR in prostate cancer patients [7]. Various risk classification methods, such as the CAPRA-S score [8] and Stephenson's nomogram [9], have been proposed. However, these methods only partially explain the variability in clinical outcomes, struggling to account for the genetic heterogeneity of the disease and accurately predict recurrence. To improve the evaluation of BCR, molecular biomarkers at the genetic and protein levels have been identified, allowing for the development of risk-predictive signatures that enhance clinical outcomes. Recent advancements in understanding the molecular landscape of early-stage PCa and improved molecular testing methods have led to the development of molecular tissue-based assays, such as Oncotype DX Genomic Prostate Score (12 cancer-related and 5 reference genes) [10] and SigMuc1NW (15-gene signature) [11]. While these assays have prognostic value, they are currently recommended for use with clinical and pathological parameters. Gene signatures formed from aberrant transcriptional patterns, identified through gene chips and high-throughput sequencing, have shown promise in predicting prognosis for prostate cancer patients. Studies have proposed gene signatures, including a 28-gene hypoxia-related signature, as predictors of BCR-free survival [12]. However, independent validations must be performed to ensure the successful implementation of biomarkers into regular clinical practice. Unfortunately, achieving these validations is often challenging and not commonly attained. Thus, there is a need to identify novel signatures that can accurately identify high-risk patients for BCR and use multiple datasets to validate it.

To develop an ideal signature based on 23 prognostic genes, we constructed and validated a Prostate Cancer Recurrence Prediction Signature (PCRPS) via machine-learning algorithms. In 8 multicenter cohorts, PCRPS exhibited robust performance in recurrence prediction. After comparing performance with published signatures of prostate cancer recurrence prediction, our PCRPS also demonstrated stable and dramatically superior predictive capability. In summary, our study offers a valuable reference for the prognostic assessment and personalized therapy of prostate cancer.

2. Materials and methods

2.1. Expression data extraction

Fig. 1 shows our research workflow schematic. We extensively reviewed publicly available Gene Expression Omnibus (GEO) datasets on December 31, 2022. The following keywords were added to our search strategy: (Recurrence OR Relapse) AND (Prostate OR Prostatic) AND (Cancer OR Tumor OR Neoplasia OR Neoplasm OR Malignancy OR Carcinoma OR Adenocarcinoma OR PCa) AND (Gene expression OR Expression). The entry type was restricted to "series," and the organism was filtered by "Homo sapiens".

The gene expression datasets used in our analyses underwent the following filtering steps: (1) Exclusion of datasets without prostate cancer samples. (2) Exclusion of datasets lacking mRNA expression data. (3) Exclusion of datasets with fewer than 5 samples in both recurrent and non-recurrent groups. (4) Exclusion of datasets with fewer than 10,000 genes having available expression data. (5) Exclusion of duplicate datasets.

2.2. Data processing and quality control

For the seven datasets (GSE25136, GSE40272, GSE46602, GSE70768, GSE70769, GSE89317, GSE116918), the downloaded expression data was normalized using the min-max normalization method. As for the datasets GSE54460, GSE54691, GSE111177, and GSE216490, the downloaded expression data was normalized using the normalizeBetweenArrays and min-max normalization methods. The reason for using the normalizeBetweenArrays method on these datasets was due to significant differences observed among samples before normalization. The normalizeBetweenArrays function provided by the "limma" package is employed to execute between-array normalization on gene expression data [13]. This normalization method addresses discrepancies in total intensity and distribution among arrays, effectively removing such variations.

Given the diversity of platforms from which the datasets were derived, a meta-analysis approach was utilized to identify common differential genes across the datasets. This meta-analysis technique enhances the statistical power and improves the reliability of the results. In order to mitigate potential biases arising from platform differences, all datasets underwent quality control using the criteria established in the MetaQC package [14]. Data quality control, or QC, plays a vital role in bioinformatics analysis to ensure the quality, reliability, and accuracy of datasets. In this study, the MetaQC method was employed, which provides a comprehensive assessment of

microarray data across multiple studies. This method aids in establishing inclusion/exclusion criteria for genomic meta-analysis. The MetaQC package was utilized to conduct QC procedures on the datasets, with lower-quality datasets being filtered out based on the determined criteria. Six statistical indicators were employed by the MetaQC software to assess a dataset's quality: internal quality control (IQC), external quality control (EQC), accuracy quality control of different expression genes (AQCg), accuracy quality control of pathways (AQCp), consistency quality control of genes (CQCg), and consistency quality control of pathways (CQCp). Furthermore, a mean rank score, derived from the combination of these six indicators, was generated to provide an overall evaluation of dataset quality.

2.3. Meta-analysis

The MetaDE package encompasses 12 primary meta-analysis methods for differential expression analysis. The eight selected datasets comprise 239 recurrent and 494 non-recurrent samples. The "MetaDE" function was deployed to identify differentially expressed genes (DEGs), with Fisher's exact test selected as the meta-analysis method within the package. The threshold for identifying DEGs was a false discovery rate (FDR) below 0.05.

2.4. Batch effect correction and DEGs

The normalized expression data were adjusted for batch effects employing the ComBat function from the sva R package, which uses an empirical Bayes model [15]. During this process, we considered the sample type, distinguishing between recurrent and non-recurrent cases, as a crucial variable to account for the biological variability between these two categories. To evaluate the remaining batch effects, we performed principal component analysis (PCA) on the transcriptome profiles. By utilizing the ComBat function, we successfully merged the normalized expression data from the eight cohorts into a single Merged-Cohort, ensuring consistency across the datasets. This allowed us to effectively address batch effects and create a unified dataset for further analysis.

In the Merged-Cohort, we utilized the T-test to identify differentially expressed genes (DEGs). A false discovery rate (FDR) threshold of less than 0.05 was employed to determine the significance of these DEGs. We then integrated the DEGs obtained from the meta-analysis and the DEGs identified through the T-test, and retained them for subsequent analysis. This integration step ensured that we captured a comprehensive set of DEGs for further investigation. In the next step, we employed Enrichr, a powerful online tool for gene set enrichment analysis, to identify enriched Gene Ontology (GO), Hallmark, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [16]. This approach allowed us to gain insights into the functional annotations and biological pathways associated with the identified DEGs. We further selected and plotted the top five pathways according to their adjusted p-values for detailed visualization.

2.5. Feature selection via machine learning

To identify diagnostic biomarkers for recurrence, we employed Random Forest (RDF) as a feature selection method. RDF is a randomized algorithm that aims to improve model accuracy by generating multiple decision trees from a single training set, thereby mitigating overfitting. In our study, we used the Merged-Cohort, which incorporated the data from the 8 GEO cohorts after removing batch effects, for analysis. We divided the Merged-Cohort into three subsets: a training cohort consisting of 50% of the samples, an evaluation dataset with 25% of the samples, and a testing dataset with the remaining 25% of the samples. The training cohort was used to train the Random Forest model, while the evaluation dataset served to assess the performance of the model during the feature selection process. Finally, the testing dataset was used to evaluate the predictive performance of the selected features. Using RDF, we identified relevant features (genes) with higher importance, which were ranked based on their 'MeanDecreaseGini' value. The Mean Decrease Gini score measures how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. A higher Mean Decrease Gini score indicates a higher importance of the variable in the model. To ensure the robustness of our feature selection process, we employed tenfold cross-validation. We selected the top genes that achieved an Area Under the Curve (AUC) value of 0.77 or higher in the evaluation dataset as characteristic genes, indicating their potential as diagnostic biomarkers for recurrence. The selected genes were named as Prostate Cancer Recurrence Prediction Signature (PCRPS) in our study.

2.6. Collection and calculation of published signatures

To assess and compare the predictive accuracy of our PCRPS signature with existing published signatures, such as the Gleason score-related signature and immune signature, we performed a thorough search on PubMed for articles discussing prognostic models until March 1, 2023. The primary objective of this search was to collect pertinent gene expression data from these articles across eight GEO cohorts and the Merged-Cohort. Subsequently, these datasets were utilized to evaluate the prognostic ability of the published signatures in predicting the recurrence of prostate cancer (PCa). Our evaluation was primarily based on the calculation of Area Under the Curve (AUC) values. To initiate the analysis, we randomly split cohort into equal-sized training and testing groups. The gene expression data obtained from the articles in the eight GEO cohorts and the Merged-Cohort were employed to train random forest (RDF) models based on gene expression data. Subsequently, we evaluated the performance of these models using the testing samples. This systematic process enabled us to evaluate the prognostic accuracy of the published signatures in predicting the recurrence of prostate cancer. By following this procedure, we were able to comprehensively compare the predictive performance of our signature with that of previously published signatures.

In this study, we thoroughly examined 23 articles that explored gene signatures in their prediction of prostate cancer recurrence. The articles included in this study presented various gene signatures for predicting the recurrence of prostate cancer. Here is a summary of the articles and their corresponding gene signatures: (1) Article 1: This study proposed a gene signature consisting of 10 prognostic genes for predicting prostate cancer recurrence [17]. Article 2 focused on identifying differentially expressed genes in prostate cancer and constructing a gene signature using six genes for predicting recurrence [18]. In Article 3, a gene signature comprising six specific genes was proposed for predicting recurrence in prostate cancer [19]. Article 4 presented a gene signature composed of eight genes for predicting recurrence in prostate cancer [20]. Article 5 introduced a gene signature that was derived from four Gleason score-related genes for predicting recurrence in prostate cancer [21]. Article 6 described a gene signature consisting of 18 immune-related genes for predicting recurrence in prostate cancer [22]. Article 7 reported a gene signature comprising three angiogenesis-related genes for predicting recurrence in prostate cancer [23]. Article 8 presented a gene signature consisting of six immune-related genes for predicting recurrence in prostate cancer [24]. Article 9 proposed a gene signature comprising five genes associated with progression-free survival for predicting recurrence in prostate cancer [25]. Article 10 introduced a gene signature that incorporated three differentially expressed genes related to the Gleason score for predicting recurrence in prostate cancer [26]. Article 11 presented a gene signature consisting of four specific genes for predicting recurrence in prostate cancer [27]. Article 12 described a gene signature composed of 9 specific genes for predicting recurrence in prostate cancer [28]. Article 13 reported a gene signature comprising 10 specific genes for predicting recurrence in prostate cancer [29]. Article 14 presented a gene signature consisting of five genes associated with the Gleason score for predicting recurrence in prostate cancer [30]. Article 15 introduced a gene signature comprised of ten genes for predicting recurrence in prostate cancer [31]. Article 16 proposed a gene signature composed of three genes for predicting the recurrence of prostate cancer [32]. Article 17 described a gene signature consisting of 28 hypoxia-related prognostic genes for predicting recurrence in prostate cancer [12]. Article 18 presented a 15-gene signature known as SigMuc1NW for predicting recurrence in prostate cancer [11]. Article 19 proposed a gene signature comprised of 10 specific genes for predicting recurrence in prostate cancer [33]. Article 20 introduced a gene signature consisting of 22 specific genes for predicting recurrence in prostate cancer [34]. Article 21 presented the Oncotype DX Prostate Cancer Assay, a gene signature of 17 specific genes, for predicting recurrence in prostate cancer [10]. Article 22 introduced a gene signature of 31 specific genes involved in predefined cell cycle progression for predicting recurrence in prostate cancer [35]. Article 23 is the Prostate Cancer Recurrence Prediction Signature (PCRPS) generated in this study.

2.7. Comparison of prediction performance of our signature with clinical variables

This section compares the predictive performance of our 23-gene signature with traditional clinical variables. It is important to recognize that clinical variables have long been utilized as key prognostic indicators in prostate cancer. Common factors include patient age, stage, and biochemical markers such as prostate-specific antigen (PSA) levels. However, they can sometimes fall short in their predictive capabilities, particularly when dealing with multifactorial and highly heterogeneous diseases like cancer. Our analyses aimed to explore this aspect further, contrasting the predictive performance of our 23-gene signature with these established clinical variables. By comparing the two, we sought to discern whether our gene signature could enhance predictive accuracy or even potentially replace the use of some traditional clinical parameters. This would not only deepen our understanding of the disease biology but could also pave the way for more precise, personalized treatment strategies for patients with prostate cancer.

2.8. Construction and selection of machine learning-based models

To construct a consensus prognosis model for prostate cancer recurrence, we followed the workflow described below. (1) Selection of Classical Algorithms: We used nine classical algorithms, namely Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Decision Trees (DCT), Random Forests (RDF), Gradient Boosting Machines (GBM), K-Nearest Neighbors (KNN), Naive Bayes (NVB), Multilayer Perceptron (MTP), Extreme Gradient Boosting (XGB). By using these algorithms, we aimed to select the most robust and accurate model for predicting recurrence. (2) Evaluation of Model Performance: We evaluated the performance of each algorithm using Area Under the Curve (AUC). The AUC values were calculated using the testing dataset, which served as an independent set of samples not used during model training. (3) Selection of Best Model: Based on the AUC values obtained from the testing dataset, we selected the model with the best performance. (4) Prediction of Recurrence Possibility Score: The final selected model was utilized to predict the recurrence possibility score for all samples. This score, named the PCRPS Score, represents the predicted likelihood of recurrence for each sample. (5) Categorization of Patients: To facilitate risk stratification and clinical decision-making, patients were categorized into high and low PCRPS groups. This categorization was performed based on the median value of the PCRPS scores. Patients with PCRPS scores above the median were assigned to the high PCRPS group, indicating a higher probability of recurrence, while those with scores below the median were assigned to the low PCRPS group, suggesting a lower probability of recurrence.

2.9. Building shiny web apps in R

In recent years, the Shiny package has emerged as a powerful resource for developing interactive web-based tools [36]. These tools, noted for their accessibility, interactivity, robustness, and relative ease of development, have become a dynamic platform for data management and analysis across a multitude of research domains. Studies have utilized Shiny to host online web servers for cancer prediction [37]. Our study showed the extension of machine learning models to the medical field through Shiny apps, which can process gene expression or clinical data as input. This novel application, once fed with the appropriate input file, empowers our models

to analyze the gene expression, thereby enabling the accurate prediction of the recurrence of prostate cancer in the respective samples.

2.10. Development and validation of potential therapeutic agents

Drug Sensitivity Database, such as GDSC2 (Genomics of Drug Sensitivity in Cancer) [38] and CTRP (The Cancer Therapeutics Response Portal) [39] were used in drug prediction. The expression data of 805 cell lines and their drug sensitivity data to 198 compounds were available from GDC2. Similarly, CTRP connects the genetic features of 829 cancer cell lines to drug sensitivity of 545 compounds. OncoPredict is an R package for predicting *in vivo* or cancer patient drug response and biomarkers from cell line screening data [40]. The drug sensitivities (IC50) for each drug and each patient were predicted by the expression data of Merged-Cohort. The Wilcoxon rank-sum tests were performed between the drug sensitivities of samples from low and high PCPRS score groups. The log2 fold change (FC) and p-values were calculated, and drugs with log2FC < -0.2 and p-value < 0.05 were selected.

3. Results

3.1. Included datasets

Our workflow is presented in Fig. 1. A total of 11 microarray datasets were identified in the literature search based on our pre-specified criteria. The details and characteristics of each dataset, including the sources and sample sizes, are provided in Table 1. Among the gene expression datasets analyzed in this study, GSE25136 and GSE46602 were generated using the Affymetrix platform, while GSE54460, GSE70768, GSE70769, and GSE111177 employed the Illumina platform. Additionally, GSE54691 and GSE89317 were performed using the Agilent platform, while the remaining datasets utilized other platforms. The microarray raw data for the eleven datasets was obtained from their respective platforms. For the datasets GSE25136, GSE40272, GSE46602, GSE70768, GSE70769, GSE89317, and GSE116918, the downloaded expression data was normalized using the min-max normalization method. On the other hand, for the datasets GSE54460, GSE54691, GSE111177, and GSE216490, the downloaded expression data was normalized using a combination of the normalizeBetweenArrays and min-max normalization methods. The reason for the additional use of normalizeBetweenArrays was that GSE54460, GSE54691, GSE111177, and GSE216490 had more significant differences between samples in each dataset. By combining the normalizeBetweenArrays method with min-max normalization, it helps to ensure that the expression data is appropriately normalized and comparable across samples within these datasets. The normalization results of these 11 datasets were shown in the boxplots (Supplementary Fig. 1 A–K). The boxplots demonstrated that the expression values of samples within each dataset were consistently distributed at similar levels. This observation suggests that the internal differences within the dataset, such as batch effects or other sources of variability, have been effectively eliminated through the normalization process. Principal component analysis (PCA) of recurrent and non-recurrent samples from all datasets were shown in Supplementary Fig. 2 A–K.

3.2. QC of the microarray data

Table 2 and Fig. 2A present the results of quality control (QC) for the eleven microarray datasets. To establish a threshold for determining the quality of the datasets, the 25th percentile was calculated for each QC metric. Datasets that showed values lower than at least two of the 25th percentile metrics were classified as low-quality datasets. This approach effectively identifies datasets that deviate significantly from the majority in terms of QC metrics, indicating potential issues or lower data quality. In line with this, Dataset11 (GSE216490), Dataset6 (GSE70768), Dataset4 (GSE54460), Dataset3 (GSE46602), and Dataset7 (GSE70769) demonstrated strong performance in both internal quality control (IQC) and external quality control (EQC) metrics. Conversely, Dataset1 (GSE25136), Dataset9 (GSE111177), and Dataset5 (GSE54691) exhibited poorer performance in IQC and EQC. The rankings of QC

Table 1
Characteristics of datasets.

Accession/ ID	Platform	Submission date (Year)	PMID	Number of non-recurrent (Non) tissues	Number of recurrent (Rec) tissues	Reference
GSE25136	GPL96	2010	19343730	40	39	[41]
GSE40272	GPL9497/ GPL15971/ GPL15972/ GPL15973/	2012	22349817	63	19	[42]
GSE46602	GPL570	2013	26522007	14	22	[43]
GSE54460	GPL11154	2014	24713434	48	46	[44]
GSE54691	GPL8737	2014	25024180	80	24	[45]
GSE70768	GPL10558	2015	26501111	93	19	[46]
GSE70769	GPL10558	2015	26501111	48	45	[46]
GSE89317	GPL14550	2016	28977898	8	8	[47]
GSE111177	GPL16791	2018	30314329	10	14	[48]
GSE116918	GPL25318	2018	29045551	192	56	[49]
GSE216490	GPL10999	2022	36329628	28	24	[50]

Table 2
QC scores of the 11 datasets, lower values and rank indicate lower quality.

	IQC	EQC	AQCg	AQCp	CQCg	CQCp	Average Rank
GSE25136	0.0002	0.69	0	0	0.87	410	4.08
GSE40272	2.86	2.74	0	0	0.77	410	4.92
GSE46602	6.59	6.93	7.56	0	8.58	410	8.67
GSE54460	6.75	6.63	0.4	0	0.007	410	6.83
GSE54691	0.39	0.86	0	0	3.84	410	4.75
GSE70768	7.12	6.57	0	0	3.85	410	7.25
GSE70769	5.09	4.97	0	0	5.32	410	6.75
GSE89317	3.76	3.13	0	0	0.43	410	5.08
GSE111177	1.34	1.43	0	0	0.11	410	4.25
GSE116918	4.48	3.87	2.79	0	0.91	410	6.83
GSE216490	8.08	8.30	0	0	0.0001	410	6.58

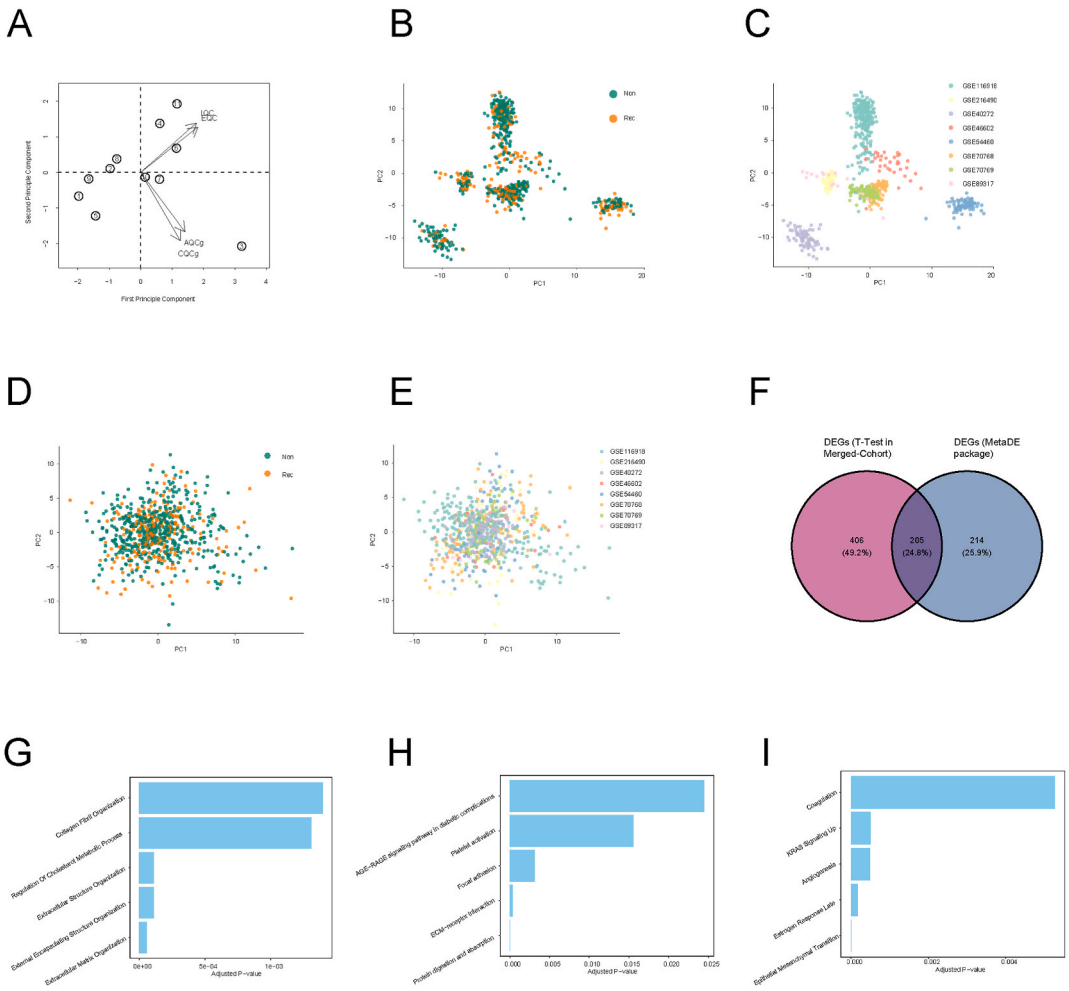


Fig. 2. PCA plot illustrating the QC results of the 11 datasets. The datasets are denoted as 1–11, corresponding to Table 2. (**A**) The X-axis represents the 1st principal component, while the Y-axis represents the 2nd component. Arrows project the six QC measures for each dataset onto the subspace of the first two principal components. Circles with numbers indicate the datasets, with the numbers corresponding to the serial numbers in Table 2 (Dataset1: GSE25136, Dataset2: GSE40272, Dataset3: GSE46602, Dataset4: GSE54460, Dataset5: GSE54691, Dataset6: GSE70768, Dataset7: GSE70769, Dataset8: GSE89317, Dataset9: GSE111177, Dataset10: GSE116918, Dataset11: GSE216490). (**B**) PCA result of recurrent (Rec) and non-recurrent (Non) samples in the Merged-Cohort before batch effect correction. (**C**) PCA result of samples from different datasets in the Merged-Cohort before batch effect correction. (**D**) PCA result of non-recurrent samples in the Merged-Cohort after batch effect correction. (**E**) PCA result of samples from different datasets in the Merged-Cohort after batch effect correction. (**F**) Venn plot depicting the intersected DEGs. Enriched GO-BP (**G**), KEGG (**H**), and Hallmark (**I**) pathways of the DEGs.

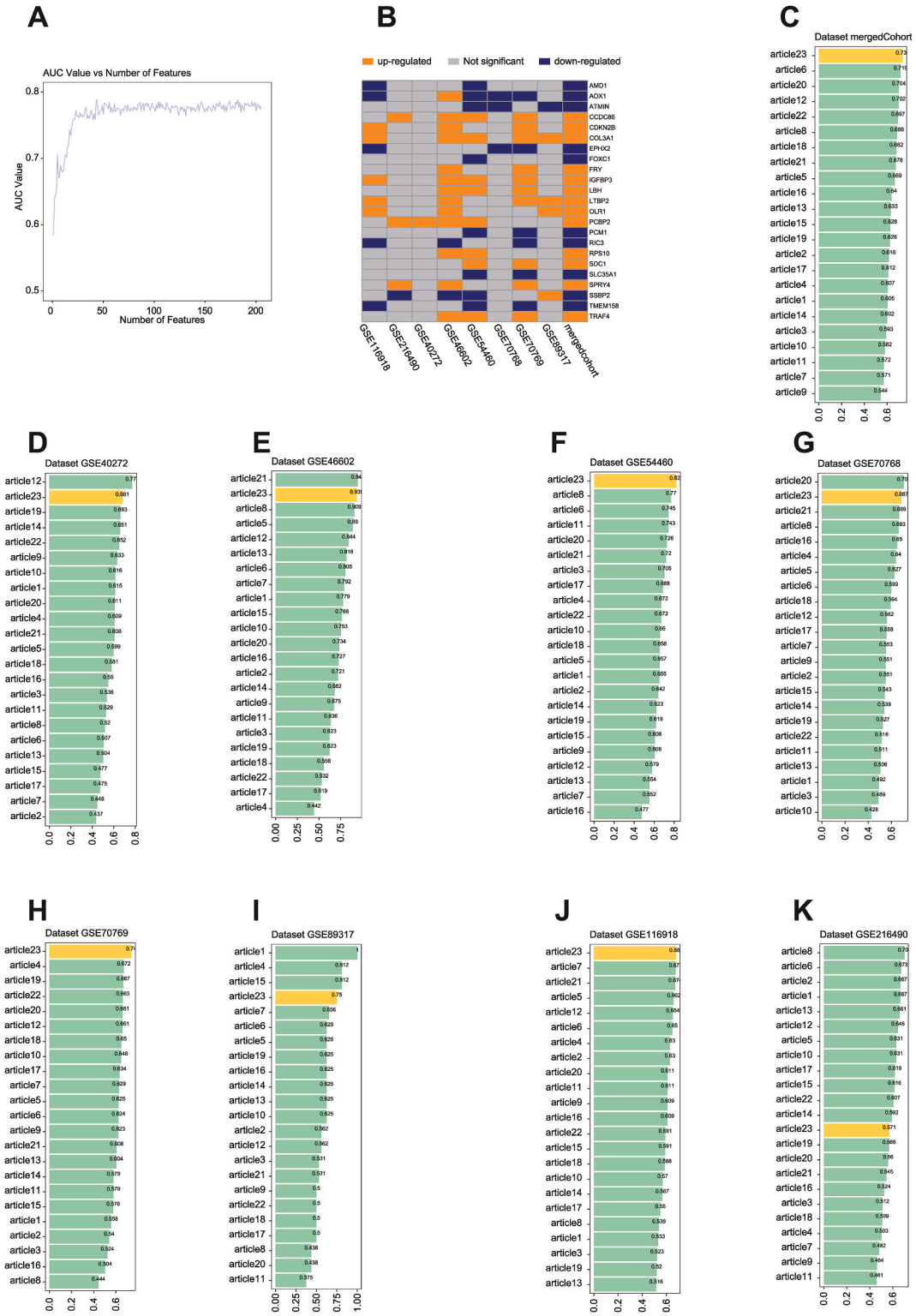


Fig. 3. Expression of genes and the predictive performance of signatures. (**A**) The selection of genes for predicting prostate cancer recurrence. (**B**) The T-test results of genes in GEO cohorts and Merged-Cohort. Yellow: up-regulated genes; Blue: down-regulated genes; Gray: Not significant genes. AUC values of signatures from articles for predicting the recurrence of prostate cancer in Merged-Cohort (**C**), GSE40272 (**D**), GSE46602 (**E**), GSE54460 (**F**), GSE70768 (**G**), GSE70769 (**H**), GSE89317 (**I**), GSE116918 (**J**), and GSE216490 (**K**). The article23 with the yellow color is the Prostate Cancer Recurrence Prediction Signature (PCRPS). Other 22 signatures come from 22 published articles.

scores and their positions in the PCA biplot (Fig. 2A) confirmed that Dataset1 (GSE25136), Dataset9 (GSE111177), and Dataset5 (GSE54691) were classified as low-quality datasets. Consequently, these three datasets were excluded from further analyses. The remaining eight datasets, which exhibited satisfactory performance in terms of QC metrics, were chosen for further analysis.

3.3. Meta-analysis and DEGs

The meta-analysis was conducted using the MetaDE package on the eight selected datasets, which consisted of 239 recurrent and 494 non-recurrent samples. Applying a false discovery rate (FDR) threshold of less than 0.05, a total of 419 differentially expressed genes (DEGs) were identified using the Fisher meta-analysis method implemented in the MetaDE package. Supplementary Fig. 3 displays the expression profiles of these significant genes obtained from the MetaDE analysis.

By applying the ComBat function to the normalized expression data, we effectively corrected for potential batch effects. This correction process allowed us to mitigate any biases introduced by batch variations and merge the individual datasets into a unified dataset named the Merged-Cohort. During this phase, we considered the type of sample (either recurrent or non-recurrent) as a crucial variable, ensuring the preservation of the biological variation inherent in the two categories of samples. We discerned significant discrepancies across various datasets (Fig. 2B and C). The PCA confirmed that batch effects between the datasets had been effectively alleviated, with the primary components elucidating more about the biological variances in the sample types rather than residual batch effects (Fig. 2D and E). This testament underscores the efficacy of our batch effect correction methodology.

After the batch effect correction, within the Merged-Cohort, we executed a T-test analysis between 239 recurrent and 494 non-recurrent samples. Utilizing the FDR threshold of less than 0.05, we were able to identify a total of 611 DEGs. A Venn diagram illustrated that 205 genes were shared between the DEGs from meta-analysis and T-test (Fig. 2F). These findings provide valuable insight into the biological differences between the two classes of samples and highlight the importance of efficient batch effect correction strategies in gene expression studies.

The enrichment analysis highlighted the significant involvement of specific biological functions and pathways (Fig. 2G–I). GO-BP results presented enrichment in “Extracellular Matrix Organization”, “Extracellular Structure Organization”, “External Encapsulating Structure Organization”, “Regulation of Cholesterol Metabolic Process”, and “Collagen Fibril Organization” (Fig. 2G). The KEGG pathways such as “Protein digestion and absorption”, “ECM-receptor interaction”, “Focal adhesion”, “Platelet activation”, and “AGE-RAGE signaling pathway in diabetic complications” showed considerable enrichment (Fig. 2H). Furthermore, in the Hallmark results, processes like Epithelial Mesenchymal Transition (EMT), Estrogen Response Late pathway, Angiogenesis, KRAS Signaling Up, and Coagulation exhibited enrichments (Fig. 2I). These results hint at these pathways’ role in regulating extracellular matrix components, cholesterol metabolism, and blood coagulation mechanisms. The enrichment analysis underscores the potential roles of the analyzed gene set in various significant biological functions and pathways, providing guidance for further exploration and understanding.

3.4. Identification of diagnostic biomarkers for recurrence prediction

We utilized the random forest (RDF) algorithm to select diagnostic biomarkers that indicate the likelihood of recurrence. The RDF algorithm is a randomization algorithm that addresses overfitting issues by constructing multiple decision trees from a single training set, thereby improving the accuracy of the model. By employing RDF, we were able to identify relevant features with higher importance and rank them based on the ‘MeanDecreaseGini’ value. In our study, we trained random forest models using a training dataset and evaluated their performance using an evaluation dataset. We identified the top genes that achieved an AUC value of 0.77 in the evaluation dataset as characteristic genes. We selected the gene combination that yielded the minimum number of genes while achieving an AUC value of 0.77. Our results demonstrated that an AUC value of 0.77 was attained when 23 features were selected (Fig. 3A). The importance values of 23 genes, represented by “MeanDecreaseGini”, were plotted in Supplementary Fig. 4. We recorded the p-values of the T-test analysis of these 23 genes in 8 GEO cohorts and the Merged-Cohort, respectively. The T-test results of these 23 genes in 8 GEO cohorts and the Merged-Cohort are shown in Fig. 3B. Based on the T-test results from the Merged-Cohort, we found that among the 23 genes, 10 were down-regulated, and 13 were up-regulated in recurrent prostate cancer samples. The performance of our model was evaluated using the training, evaluation, and testing datasets, with AUC values of 1, 0.772, and 0.714, respectively. This subset of characteristic genes, identified through the RDF algorithm, has the potential to serve as reliable diagnostic biomarkers for recurrence. The selected 23 genes were used as Prostate Cancer Recurrence Prediction Signature (PCRPS).

3.5. Re-evaluation of previously 22 published signatures in PCa recurrence prediction

The rapid proliferation of high-throughput sequencing technologies has offered an invaluable way to examine stratified management and precision treatment of tumors. This has led to an eruption of research in recent years, particularly in prostate cancer. Innovative machine-learning algorithms such as LASSO have been employed to construct numerous prognostic signatures. These signatures, built upon vast amounts of high-quality data, have displayed significant efficacy in predicting the recurrence of prostate cancer. As part of our comprehensive study, we compiled 22 published mRNA prognostic signatures, facilitating a detailed comparison of the predictive accuracy between our proposed signature and existing ones, as depicted in Fig. 3C–K. We subsequently evaluated the predictive capabilities of our PCRPS signature (article23) alongside these 22 established gene signatures from 22 articles (article1 to article22), employing AUC values across eight distinct GEO cohorts and the Merged-Cohort. Intriguingly, our AUC results highlighted a unique feature: only our signature demonstrated consistent statistical significance across all eight cohorts and the Merged-Cohort. To facilitate an equitable performance comparison between our model and others, we utilized the raw and unadjusted expression data

before the batch correction. Our signature consistently outperformed the other models, evidencing its distinctly superior accuracy across nearly all cohorts examined. It clinched the top spot in four cohorts and a respectable second place in three others, underlining its robustness and reliability in predicting outcomes. In summary, our findings strongly advocate for the predictive strength of our 23-gene PCRPS signature. These comprehensive analyses suggest that PCRPS harbors robust predictive power for recurrence in prostate cancer patients, offering potentially invaluable insights for personalized treatment approaches and strategic therapeutic interventions.

3.6. Performance of signature by genes, clinical settings, and their combination

Next, we compared the prediction power of PCRPS with clinical variables. Samples from each of the GSE40272, GSE46602, GSE70768, and GSE116918 datasets were separately divided into training and testing datasets in a random and even manner. These four GEO datasets were selected because they include available clinical data. We used the expression profiles of these 23 genes from PCRPS, along with the clinical settings (Age, Gleason score, PSA, and stage) of the samples. Within the training cohort, we employed nine algorithms through ten-fold cross-validation to build prediction models, subsequently calculating the average AUC values of each algorithm across the testing cohorts. The AUC values of GEO cohorts and Merged-Cohort showed that our model had good prediction performance (Fig. 4A–D). The AUCs of clinical settings in GSE40272, GSE46602, GSE70768, and GSE116918 were 0.65, 1.0, 0.54, and 0.54. The AUCs of PCRPS in GSE40272, GSE46602, GSE70768, and GSE116918 were 0.92, 0.92, 0.66, and 0.73. PCRPS signature had a superior ability to predict PCa recurrence in clinical settings. As shown in Fig. 4E, in the testing dataset of Merged-Cohort, the highest AUC values for clinical settings, PCRPS, and their combinations were 0.63, 0.72, and 0.76. These results indicate that PCRPS outperforms clinical settings alone in predicting PCa recurrence, underscoring the significance of integrating clinical variables with gene expression profiles for improved prediction accuracy.

3.7. Software usage

In the model of predicting recurrence by clinical variables, the user needs to input the values of the following clinical variables. (1) Age: Please enter the patient's age in years. (2) Gleason Score: If the patient has a Gleason score of 3 + 4, please enter the value as 7. (3) PSA Value: Please enter the patient's PSA (Prostate-Specific Antigen) value, which is a numeric measurement usually ranging from 2 to 40 ng/ml (4) Stage: Please enter the numeric value of the patient's pathological stage, which can range from 1 to 4. In the model of predicting recurrence by expression values of 23 genes, the user needs to input the normalized values of gene expression (range 0–1). In

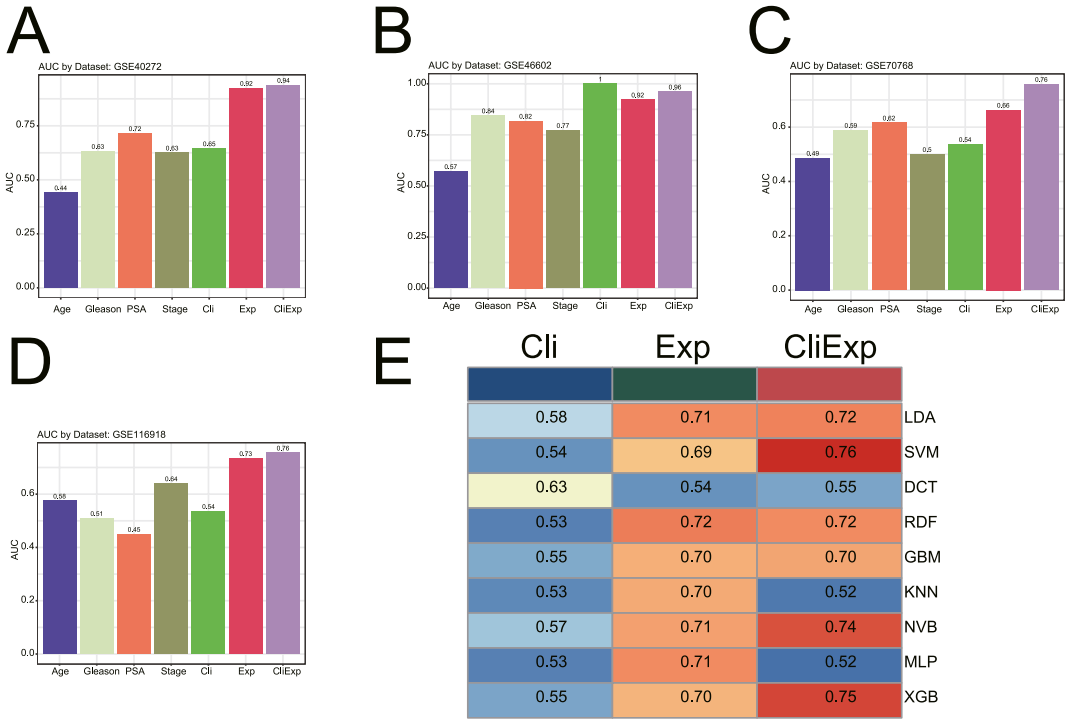


Fig. 4. Machine learning models by signature (expression, Exp), clinical variables (Cli) and their combination (CliExp). The predictive performance of signature was compared with common clinical variables in the GSE40272 (**A**), GSE46602 (**B**), GSE70768 (**C**), GSE116918 (**D**). (**E**) The AUC values of 9 machine-learning algorithm combinations in the testing dataset of Merged-Cohort by clinical data, expression data and their combination. Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Decision Trees (DCT), Random Forests (RDF), Gradient Boosting Machines (GBM), K-Nearest Neighbors (KNN), Naive Bayes (NVB), Multilayer Perceptron (MTP), eXtreme Gradient Boosting (XGB).

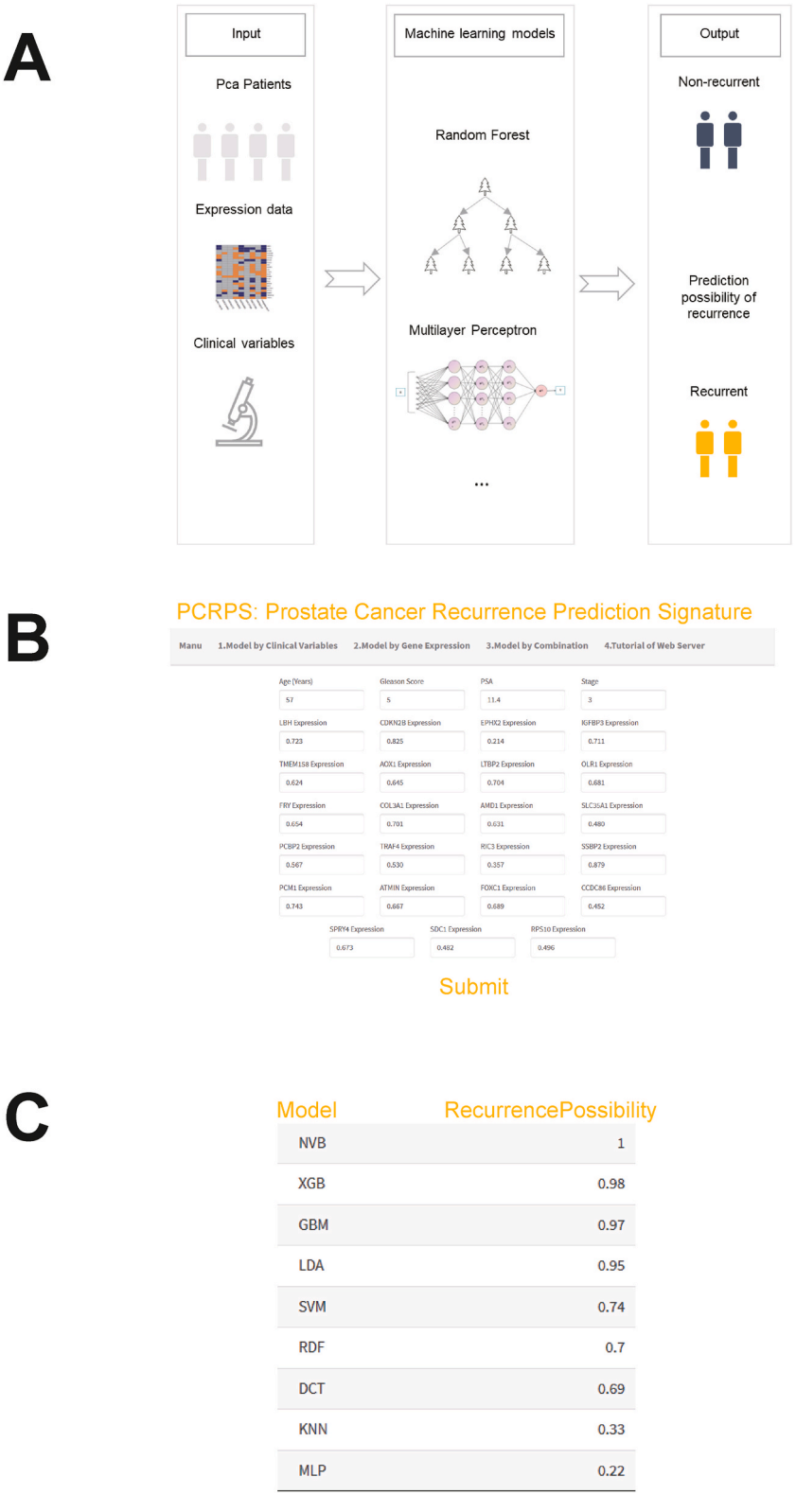


Fig. 5. The workflow (**A**) and case study (**B–C**) of PCRPS web server.

the model of predicting recurrence by the combination of clinical variables and gene expression, the user just needs to follow the same rules of input values. At present, PCRPS will output the predicted possibility score of recurrence of nine different machine learning algorithms. The workflow of our web server was shown in Fig. 5A. We randomly selected a sample (GSM1133136) with recurrence and uploaded the clinical variables and expression values of 23 genes to the web server (Fig. 5B). The prediction result was shown in Fig. 5C and recurrence possibility with a higher of 0.5 means recurrence. Seven of nine machine learning models successfully predicted the recurrence (>0.5). The web server could be freely used at https://urology1926.shinyapps.io/PCRPS/.

3.8. Searching for potential therapeutic agents for the high-group

Investigating the potential therapeutic agents is crucial for providing drugs for patients who are predicted to be recurrent by PSRPS. In Fig. 4E, the RDF has the highest AUC values (0.72) in predicting PCa recurrence by expression profiles, so this model was selected as the final model for calculating the PCRPS score. PCRPS score comes from the prediction possibility score for recurrence by random forest model, and it ranges from 0 to 1. The higher value of the PCRPS score represents a higher possibility of recurrence. We investigated the correlations between the PCRPS score and clinical variables (Supplementary Figs. 5A–D). For age and PSA, there are no significant correlations with the PCRPS score. For stage, there is a positive correlation in the PCRPS score with the increasing of stage. For Gleason, although there is a significant p-value, we do not find a clear correlation.

We next evaluated the differences in drug susceptibility between the high and low PCRPS score groups. The differential analysis demonstrated that IC50 of 22 drugs from GDSC2 were significantly lower in high-PCRPS score group in reference to low-PCRPS score group. Differential analysis demonstrated that IC50 of 95 drugs from CTRP2 were significantly lower in the high-PCRPS score group in reference to the low-PCRPS score group. The results from GDSC2 showed that the high-PCRPS group was more sensitive to Vorinostat, Axitinib, AZD7762, 5-Fluorouracil, and Tozasertib (Supplementary Table 1). The results from CTRP2 showed that the high-PCRPS group was more sensitive to Betulinic acid, Teniposide, Sildenafil, Simvastatin, and PRIMA-1 (Supplementary Table 2).

4. Discussion

Biochemical recurrence develops in almost one-third of men with prostate cancer after treatment [51]. The lack of accessible biomarkers for screening, stratified management, and prognostic follow-up has posed an urgent challenge for clinicians and researchers, potentially leading to excessive or insufficient treatment. We have devised a signature for forecasting biochemical recurrence in prostate cancer, employing gene expression data and machine learning techniques. We developed and validated a predictive model of 23 genes using a combination of 9 machine-learning algorithms across 8 independent multicenter cohorts. In comparison to several commonly used clinicopathological features and previously published signatures for predicting recurrence in prostate cancer, our model exhibited strong and superior predictive ability. Additionally, our research confirmed that the high PRCPS score group displayed higher tumor stages and worse prognosis. Consequently, our model holds the potential to serve as a reliable platform for facilitating personalized decision-making in clinical settings. Prostate cancer patients require optimal biomarkers that can accurately estimate prognosis and treatment effectiveness. However, the conventional TNM staging system is insufficient for this purpose in the era of precision medicine.

Predictive indicators for prostate cancer have been developed in the past few years, however, nearly all of them depend on a particular pathway, such as immunity, metabolism [52], and m6A methylation [53]. These indicators neglected information on other biological mechanisms that were critical to the development of prostate cancer. In this study, based on DEGs and feature selection analysis, we further obtained 23 genes via random forest. We have chosen nine different machine learning algorithms to reduce inaccuracies caused by model selection. Ultimately, we finally obtained a 23-gene signature termed PRCPS through the random forest. The results of ROC curve indicated that PRCPS had excellent predictive performance in the testing cohorts of GEO and Merged-Cohorts. Moreover, compared with 22 published prostate cancer recurrence prediction signatures, PRCPS exhibited distinctly superior accuracy than other signatures in almost all cohorts, revealing the robustness of PRCPS. It was observed that most of these signatures utilized around 10 genes for predicting recurrence. However, relying on only 10 genes may not provide sufficient robustness in prediction accuracy. On the other hand, some signatures employed more than 30 genes, but this approach could limit the clinical utility of the signature. In contrast, our signature comprised 23 genes, ensuring robustness and clinical usability. Overall, our study demonstrated the superior prediction performance of our signature compared to the published signatures and common clinical variables. The inclusion of a larger number of genes in our signature balanced robustness and clinical utility, overcoming the limitations of using either too few or too many genes. Furthermore, compared to common clinical and molecular characteristics such as TNM stage, grade, and PSA, our PRCPS exhibited significantly improved accuracy in predicting recurrence. This indicates that our signature outperformed these clinical variables in terms of predictive ability. The limited performance of the clinical variables emphasizes the challenges associated with relying solely on them for recurrence prediction.

By stratifying prostate cancer patients into high and low PRCPS groups, we found that there were no significant differences in terms of age or gender. However, the high PRCPS group displayed more advanced TNM stages, which contributed to a worse prognosis to some extent. These findings highlight the potential of PRCPS as a biomarker with broad applicability in prostate cancer. Shiny is an open-source R package that provides an elegant and powerful web framework for building web applications using R. Recently, Shiny was adopted by some researchers to build websites for medical and biology researchers. For example, Shimada built an easy-to-use browser, shinyDepMap, to identify targetable cancer genes and their functional connections [54]. A user-friendly shiny app was built for prediction of the overall survival of spinal metastatic disease [55]. In our study, we implanted machine learning models by clinical settings, PCRPS and their combinations with https://urology1926.shinyapps.io/PCRPS/. The tool also can be used to (1)

predict the PCa recurrence by clinical information or expression data with high accuracy, and (2) provide the possibility of PCa recurrence by 9 machine learning algorithms.

This study differs from previous studies in several key aspects. Firstly, we conducted a systematic collection of data from 11 large multicenter cohorts and used 8 of them to compare the accuracy of PCRPS in predicting prostate cancer (PCa) recurrence. This comparison involved evaluating published articles and considering clinical settings. Secondly, our PCRPS differs from current prognostic models that focus on specific pathways. Instead, we based our approach on an analysis of over 8000 intersection genes derived from the 11 cohorts. This comprehensive gene selection aimed to avoid the omission of other essential biological processes involved in the initiation and progression of PCa. To address potential biases stemming from personal preferences in modeling, we utilized nine machine-learning algorithms and selected the most accurate model based on their performance. While we made every effort to ensure the rigor and comprehensiveness of our research, it is important to acknowledge certain limitations. Firstly, although we collected data from 11 independent multicenter cohorts, further validation through prospective studies is warranted. Secondly, the specific roles of the 23 genes included in PCRPS in the context of PCa remain to be fully elucidated, necessitating additional functional experimental validation in future studies. Finally, conducting further clinical trials will be essential to confirm the therapeutic efficacy of compounds in PCa patients identified as having a high PCRPS score.

5. Conclusion

In conclusion, we developed PCRPS, the first user-friendly web application to predict prostate cancer recurrence. Using 23 genes, we constructed and validated a consensus prognostic signature using machine-learning algorithms. PCRPS demonstrated superior predictive capabilities compared to important clinicopathological features and previously published signatures. Our PCRPS also showed significant clinical implications for the management and personalized treatment of prostate cancer. In summary, our study offers a valuable reference for the prognostic assessment and personalized therapy of prostate cancer.

Data availability statement

Public datasets were analyzed in this study, available for access at the National Center for Biotechnology Information (NCBI) website: https://www.ncbi.nlm.nih.gov. The expression and recurrence data can be retrieved by searching for the dataset IDs: GSE25136, GSE40272, GSE46602, GSE70768, GSE70769, GSE89317, GSE116918, GSE54460, GSE54691, GSE111177, and GSE216490. Additionally, the code developed for generating the main dataset and associated results is accessible on GitHub at the following URL: https://github.com/PRADpre/PCRPS.

Fundings

This study was supported by Hunan Provincial Inclusive Policy and Innovative Environment Construction Plan (2020SK50902) and National Clinical Key Professional Sections Construction Plan (Urology).

Ethics approval

All data used in this study is publicly available, and ethics approval is not needed in our study.

CRediT authorship contribution statement

Xianya He: Writing – original draft, Methodology, Data curation. **Sheng Hu:** Writing – original draft, Methodology, Data curation. **Chen Wang:** Data curation. **Yongjun Yang:** Data curation. **Zhuo Li:** Data curation. **Mingqiang Zeng:** Data curation. **Guangqing Song:** Data curation. **Yuanwei Li:** Writing – review & editing, Supervision. **Qiang Lu:** Writing – review & editing, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e28878.

References

[1] R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2019, CA A Cancer J. Clin. 69 (2019) 7–34, https://doi.org/10.3322/caac.21551.

[2] Y. Zhu, H.-K. Wang, Y.-Y. Qu, D.-W. Ye, Prostate cancer in East Asia: evolving trend over the last decade, Asian J. Androl. 17 (2015) 48–57, https://doi.org/10.4103/1008-682X.132780.

[3] L.A. Gurski, A.K. Jha, C. Zhang, X. Jia, M.C. Farach-Carson, Hyaluronic acid-based hydrogels as 3D matrices for in vitro evaluation of chemotherapeutic drugs using poorly adherent prostate cancer cells, Biomaterials 30 (2009) 6076–6085, https://doi.org/10.1016/j.biomaterials.2009.07.054.

[4] P. Dodla, V. Bhoopalan, S.K. Khoo, C. Miranti, S. Sridhar, Gene expression analysis of human prostate cell lines with and without tumor metastasis suppressor CD82, BMC Cancer 20 (2020) 1211, https://doi.org/10.1186/s12885-020-07675-7.

[5] B.Z. McCormick, A.M. Mahmoud, S.B. Williams, J.W. Davis, Biochemical recurrence after radical prostatectomy: current status of its use as a treatment endpoint and early management strategies, Indian J. Urol. IJU J. Urol. Soc. India 35 (2019) 6–17, https://doi.org/10.4103/iju.IJU_355_18.

[6] P. Cornford, J. Bellmunt, M. Bolla, E. Briers, M. De Santis, T. Gross, A.M. Henry, S. Joniau, T.B. Lam, M.D. Mason, H.G. van der Poel, T.H. van der Kwast, O. Rouvière, T. Wiegel, N. Mottet, EAU-ESTRO-SIOG guidelines on prostate cancer. Part II: treatment of relapsing, metastatic, and castration-resistant prostate cancer, Eur. Urol. 71 (2017) 630–642, https://doi.org/10.1016/j.eururo.2016.08.002.

[7] C. Hoey, M. Ahmed, A. Fotouhi Ghiam, D. Vesprini, X. Huang, K. Commisso, A. Commisso, J. Ray, E. Fokas, D.A. Loblaw, H.H. He, S.K. Liu, Circulating miRNAs as non-invasive biomarkers to predict aggressive prostate cancer after radical prostatectomy, J. Transl. Med. 17 (2019) 173, https://doi.org/10.1186/s12967-019-1920-5.

[8] M.R. Cooperberg, J.F. Hilton, P.R. Carroll, The CAPRA-S score: a straightforward tool for improved prediction of outcomes after radical prostatectomy, Cancer 117 (2011) 5039–5046, https://doi.org/10.1002/cncr.26169.

[9] A.J. Stephenson, P.T. Scardino, J.A. Eastham, F.J. Bianco, Z.A. Dotan, P.A. Fearn, M.W. Kattan, Preoperative nomogram predicting the 10-year probability of prostate cancer recurrence after radical prostatectomy, J. Natl. Cancer Inst. 98 (2006) 715–717, https://doi.org/10.1093/jnci/djj190.

[10] D. Knezevic, A.D. Goddard, N. Natraj, D.B. Cherbavaz, K.M. Clark-Langone, J. Snable, D. Watson, S.M. Falzarano, C. Magi-Galluzzi, E.A. Klein, C. Quale, Analytical validation of the Oncotype DX prostate cancer assay – a clinical RT-PCR assay optimized for prostate needle biopsies, BMC Genom. 14 (2013) 690, https://doi.org/10.1186/1471-2164-14-690.

[11] Y. Jiang, W. Mei, Y. Gu, X. Lin, L. He, H. Zeng, F. Wei, X. Wan, H. Yang, P. Major, D. Tang, Construction of a set of novel and robust gene expression signatures predicting prostate cancer recurrence, Mol. Oncol. 12 (2018) 1559–1578, https://doi.org/10.1002/1878-0261.12359.

[12] L. Yang, D. Roberts, M. Takhar, N. Erho, B.A.S. Bibby, N. Thiruthaneeswaran, V. Bhandari, W.-C. Cheng, S. Haider, A.M.B. McCorry, D. McArt, S. Jain, M. Alshalalfa, A. Ross, E. Schaffer, R.B. Den, R. Jeffrey Karnes, E. Klein, P.J. Hoskin, S.J. Freedland, A.D. Lamb, D.E. Neal, F.M. Buffa, R.G. Bristow, P.C. Boutros, E. Davicioni, A. Choudhury, C.M.L. West, Development and validation of a 28-gene hypoxia-related prognostic signature for localized prostate cancer, EBioMedicine 31 (2018) 182–189, https://doi.org/10.1016/j.ebiom.2018.04.019.

[13] M.E. Ritchie, B. Phipson, D. Wu, Y. Hu, C.W. Law, W. Shi, G.K. Smyth, Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (2015) e47, https://doi.org/10.1093/nar/gkv007.

[14] D.D. Kang, E. Sibille, N. Kaminski, G.C. Tseng, MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis, Nucleic Acids Res. 40 (2012) e15, https://doi.org/10.1093/nar/gkr1071.

[15] J.T. Leek, W.E. Johnson, H.S. Parker, A.E. Jaffe, J.D. Storey, The sva package for removing batch effects and other unwanted variation in high-throughput experiments, Bioinformatics 28 (2012) 882–883, https://doi.org/10.1093/bioinformatics/bts034.

[16] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (2013) 128, https://doi.org/10.1186/1471-2105-14-128.

[17] R. Zhou, Y. Feng, J. Ye, Z. Han, Y. Liang, Q. Chen, X. Xu, Y. Huang, Z. Jia, W. Zhong, Prediction of biochemical recurrence-free survival of prostate cancer patients leveraging multiple gene expression profiles in tumor microenvironment, Front. Oncol. 11 (2021) 632571, https://doi.org/10.3389/fonc.2021.632571.

[18] F. Li, J.-P. Ji, Y. Xu, R.-L. Liu, Identification a novel set of 6 differential expressed genes in prostate cancer that can potentially predict biochemical recurrence after curative surgery, Clin. Transl. Oncol. 21 (2019) 1067–1075, https://doi.org/10.1007/s12094-018-02029-z.

[19] A. Krzyzanowska, S. Barron, D.F. Higgins, T. Loughman, A. O'Neill, K.M. Sheehan, C.-J.A. Wang, B. Fender, L. McGuire, J. Fay, A. O'Grady, D. O'Leary, R. W. Watson, A. Bjartell, W.M. Gallagher, Development, validation, and clinical utility of a six-gene signature to predict aggressive prostate cancer, Eur. Urol. Focus 0 (2023), https://doi.org/10.1016/j.euf.2023.04.004.

[20] J. Chu, N. Li, W. Gai, Identification of genes that predict the biochemical recurrence of prostate cancer, Oncol. Lett. 16 (2018) 3447–3452, https://doi.org/10.3892/ol.2018.9106.

[21] Y. Yimamu, X. Yang, J. Chen, C. Luo, W. Xiao, H. Guan, D. Wang, The development of a Gleason score-related gene signature for predicting the prognosis of prostate cancer, J. Clin. Med. 11 (2022) 7164, https://doi.org/10.3390/jcm11237164.

[22] M. Fu, Q. Wang, H. Wang, Y. Dai, J. Wang, W. Kang, Z. Cui, X. Jin, Immune-related genes are prognostic markers for prostate cancer recurrence, Front. Genet. 12 (2021) 639642, https://doi.org/10.3389/fgene.2021.639642.

[23] B. Fan, Y. Wang, X. Zheng, X. Zhang, Z. Zhang, X. Hu, A novel angiogenesis-related gene signature to predict biochemical recurrence of patients with prostate cancer following radical therapy, JAMA Oncol. 2022 (2022) 2448428, https://doi.org/10.1155/2022/2448428.

[24] J. Luan, Q. Zhang, L. Song, Y. Wang, C. Ji, R. Cong, Q. Zheng, Z. Xu, J. Xia, N. Song, Identification and validation of a six immune-related gene signature for prediction of biochemical recurrence in localized prostate cancer following radical prostatectomy, Transl. Androl. Urol. 10 (2021) 1018–1029, https://doi.org/10.21037/tau-20-1231.

[25] Z. Mou, J. Spencer, B. Knight, J. John, P. McCullagh, J.S. McGrath, L.W. Harries, Gene expression analysis reveals a 5-gene signature for progression-free survival in prostate cancer, Front. Oncol. 12 (2022) 914078, https://doi.org/10.3389/fonc.2022.914078.

[26] Y. Wang, Z. Yang, A Gleason score-related outcome model for human prostate cancer: a comprehensive study based on weighted gene co-expression network analysis, Cancer Cell Int. 20 (2020) 159, https://doi.org/10.1186/s12935-020-01230-x.

[27] Z. Song, F. Chao, Z. Zhuo, Z. Ma, W. Li, G. Chen, Identification of hub genes in prostate cancer using robust rank aggregation and weighted gene co-expression network analysis, Aging 11 (2019) 4736–4756, https://doi.org/10.18632/aging.102087.

[28] X. Wu, D. Lv, M. Eftekhar, A. Khan, C. Cai, Z. Zhao, D. Gu, Y. Liu, A new risk stratification system of prostate cancer to identify high-risk biochemical recurrence patients, Transl. Androl. Urol. 9 (2020) 2572–2586, https://doi.org/10.21037/tau-20-1019.

[29] X. Wu, D. Lv, M. Lei, C. Cai, Z. Zhao, M. Eftekhar, D. Gu, Y. Liu, A 10-gene signature as a predictor of biochemical recurrence after radical prostatectomy in patients with prostate cancer and a Gleason score $\geq$7, Oncol. Lett. 20 (2020) 2906–2918, https://doi.org/10.3892/ol.2020.11830.

[30] L. Zhang, Y. Li, X. Wang, Y. Ping, D. Wang, Y. Cao, Y. Dai, W. Liu, Z. Tao, Five-gene signature associating with Gleason score serve as novel biomarkers for identifying early recurring events and contributing to early diagnosis for Prostate Adenocarcinoma, J. Cancer 12 (2021) 3626–3647, https://doi.org/10.7150/jca.52170.

[31] Q. Long, B.A. Johnson, A.O. Osunkoya, Y.-H. Lai, W. Zhou, M. Abramovitz, M. Xia, M.B. Bouzyk, R.K. Nam, L. Sugar, A. Stanimirovic, D.J. Williams, B. R. Leyland-Jones, A.K. Seth, J.A. Petros, C.S. Moreno, Protein-coding and MicroRNA biomarkers of recurrence of prostate cancer following radical prostatectomy, Am. J. Pathol. 179 (2011) 46–54, https://doi.org/10.1016/j.ajpath.2011.03.008.

[32] F. Söderdahl, L.-D. Xu, J. Bring, M. Häggman, A novel risk score (P-score) based on a three-gene signature, for estimating the risk of prostate cancer-specific mortality, res. Rep, Urol. Times 14 (2022) 203–217, https://doi.org/10.2147/RRU.S358169.

[33] H. Abou-Ouf, M. Alshalalfa, M. Takhar, N. Erho, B. Donnelly, E. Davicioni, R.J. Karnes, T.A. Bismar, Validation of a 10-gene molecular signature for predicting biochemical recurrence and clinical metastasis in localized prostate cancer, J. Cancer Res. Clin. Oncol. 144 (2018) 883–891, https://doi.org/10.1007/s00432-018-2615-7.

[34] N. Erho, A. Crisan, I.A. Vergara, A.P. Mitra, M. Ghadessi, C. Buerki, E.J. Bergstralh, T. Kollmeyer, S. Fink, Z. Haddad, B. Zimmermann, T. Sierocinski, K. V. Ballman, T.J. Triche, P.C. Black, R.J. Karnes, G. Klee, E. Davicioni, R.B. Jenkins, Discovery and validation of a prostate cancer genomic classifier that predicts early metastasis following radical prostatectomy, PLoS One 8 (2013) e66855, https://doi.org/10.1371/journal.pone.0066855.

[35] J. Cuzick, G.P. Swanson, G. Fisher, A.R. Brothman, D.M. Berney, J.E. Reid, D. Mesher, V.O. Speights, E. Stankiewicz, C.S. Foster, H. Møller, P. Scardino, J. D. Warren, J. Park, A. Younus, D.D. Flake, S. Wagner, A. Gutin, J.S. Lanchbury, S. Stone, Prognostic value of an RNA expression signature derived from cell cycle proliferation genes for recurrence and death from prostate cancer: a retrospective study in two cohorts, Lancet Oncol. 12 (2011) 245–255, https://doi.org/10.1016/S1470-2045(10)70295-3.

[36] W. Chang, J. Cheng, J.J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, B. Borges, RStudio, jQuery F. (jQuery library and jQuery U. library), jQuery contributors (jQuery library; authors listed in inst/www /shared/jquery-AUTHORS.txt), jQuery U. contributors (jQuery U. library; authors listed in inst/www /shared/jqueryui/AUTHORS.txt), M.O. (Bootstrap library), J.T. (Bootstrap library), B. contributors (Bootstrap library), Twitter, I. (Bootstrap library), P.N.K. (Bootstrap accessibility plugin), V.T. (Bootstrap accessibility plugin), D.L. (Bootstrap accessibility plugin), S.C. (Bootstrap accessibility plugin), C.O. (Bootstrap accessibility plugin), PayPal, I. (Bootstrap accessibility plugin), S.P. (Bootstrap-datepicker library), A.R. (Bootstrap-datepicker library), B.R. (selectize js library), S.B. (selectize-plugin-a11y library), D.I. (ion rangeSlider library), S.S. (Javascript strftime library), S.L. (DataTables library), J.F. (showdown js library), J.G. (showdown js library), I.S. (highlight js library), R.C.T. (tar implementation from R), shiny, Web Application Framework for R, 2023. https://cran.r-project.org/web/packages/shiny/index.html. (Accessed 4 August 2023).

[37] Z. Chen, M. Wang, R.L. De Wilde, R. Feng, M. Su, L.A. Torres-de la Roche, W. Shi, A machine learning model to predict the triple negative breast cancer immune subtype, Front. Immunol. 12 (2021) 749459, https://doi.org/10.3389/fimmu.2021.749459.

[38] W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, N. Bindal, D. Beare, J.A. Smith, I.R. Thompson, S. Ramaswamy, P.A. Futreal, D.A. Haber, M.R. Stratton, C. Benes, U. McDermott, M.J. Garnett, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells, Nucleic Acids Res. 41 (2013) D955–D961, https://doi.org/10.1093/nar/gks1111.

[39] A. Basu, N.E. Bodycombe, J.H. Cheah, E.V. Price, K. Liu, G.I. Schaefer, R.Y. Ebright, M.L. Stewart, D. Ito, S. Wang, A.L. Bracha, T. Liefeld, M. Wawer, J.C. Gilbert, A.J. Wilson, N. Stransky, G.V. Kryukov, V. Dancik, J. Barretina, L.A. Garraway, C.S.-Y. Hon, B. Munoz, J.A. Bittker, B.R. Stockwell, D. Khabele, A.M. Stern, P. A. Clemons, A.F. Shamji, S.L. Schreiber, An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules, Cell 154 (2013) 1151–1161, https://doi.org/10.1016/j.cell.2013.08.003.

[40] D. Maeser, R.F. Gruener, R.S. Huang, oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data, Briefings Bioinf. 22 (2021) bbab260, https://doi.org/10.1093/bib/bbab260.

[41] Y. Sun, S. Goodison, Optimizing molecular signatures for predicting prostate cancer recurrence, Prostate 69 (2009) 1119–1127, https://doi.org/10.1002/pros.20961.

[42] Z.G. Gulzar, J.K. McKenney, J.D. Brooks, Increased expression of NuSAP in recurrent prostate cancer is mediated by E2F1, Oncogene 32 (2013) 70–77, https://doi.org/10.1038/onc.2012.27.

[43] M.M. Mortensen, S. Høyer, A.-S. Lynnerup, T.F. Ørntoft, K.D. Sørensen, M. Borre, L. Dyrskjøt, Expression profiling of prostate cancer tissue delineates genes associated with recurrence after prostatectomy, Sci. Rep. 5 (2015) 16018, https://doi.org/10.1038/srep16018.

[44] Q. Long, J. Xu, A.O. Osunkoya, S. Sannigrahi, B.A. Johnson, W. Zhou, T. Gillespie, J.Y. Park, R.K. Nam, L. Sugar, A. Stanimirovic, A.K. Seth, J.A. Petros, C. S. Moreno, Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence, Cancer Res. 74 (2014) 3228–3237, https://doi.org/10.1158/0008-5472.CAN-13-2699.

[45] H. Hieronymus, N. Schultz, A. Gopalan, B.S. Carver, M.T. Chang, Y. Xiao, A. Heguy, K. Huberman, M. Bernstein, M. Assel, R. Murali, A. Vickers, P.T. Scardino, C. Sander, V. Reuter, B.S. Taylor, C.L. Sawyers, Copy number alteration burden predicts prostate cancer relapse, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 11139–11144, https://doi.org/10.1073/pnas.1411446111.

[46] H. Ross-Adams, A.D. Lamb, M.J. Dunning, S. Halim, J. Lindberg, C.M. Massie, L.A. Egevad, R. Russell, A. Ramos-Montoya, S.L. Vowler, N.L. Sharma, J. Kay, H. Whitaker, J. Clark, R. Hurst, V.J. Gnanapragasam, N.C. Shah, A.Y. Warren, C.S. Cooper, A.G. Lynch, R. Stark, I.G. Mills, H. Grönberg, D.E. Neal, Integration of copy number and transcriptomics provides risk stratification in prostate cancer: a discovery and validation cohort study, EBioMedicine 2 (2015) 1133–1144, https://doi.org/10.1016/j.ebiom.2015.07.017.

[47] R. Demidenko, K. Daniunaite, A. Bakavicius, R. Sabaliauskaite, A. Skeberdyte, D. Petroska, A. Laurinavicius, F. Jankevicius, J.R. Lazutka, S. Jarmalaite, Decreased expression of MT1E is a potential biomarker of prostate cancer progression, Oncotarget 8 (2017) 61709–61718, https://doi.org/10.18632/oncotarget.18683.

[48] N.V. Sharma, K.L. Pellegrini, V. Ouellet, F.O. Giuste, S. Ramalingam, K. Watanabe, E. Adam-Granger, L. Fossouo, S. You, M.R. Freeman, P. Vertino, K. Conneely, A.O. Osunkoya, D. Trudel, A.-M. Mes-Masson, J.A. Petros, F. Saad, C.S. Moreno, Identification of the transcription factor relationships associated with androgen deprivation therapy response and metastatic progression in prostate cancer, Cancers 10 (2018) 379, https://doi.org/10.3390/cancers10100379.

[49] S. Jain, C.A. Lyons, S.M. Walker, S. McQuaid, S.O. Hynes, D.M. Mitchell, B. Pang, G.E. Logan, A.M. McCavigan, D. O'Rourke, D.G. McArt, S.S. McDade, I. G. Mills, K.M. Prise, L.A. Knight, C.J. Steele, P.W. Medlow, V. Berge, B. Katz, D.A. Loblaw, D.P. Harkin, J.A. James, J.M. O'Sullivan, R.D. Kennedy, D.J. Waugh, Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy, Ann. Oncol. 29 (2018) 215–222, https://doi.org/10.1093/annonc/mdx637.

[50] N.L. Acosta-Vega, R. Varela, J.A. Mesa, J. Garai, M.C. Baddoo, A. Gómez-Gutiérrez, S.J. Serrano-Gómez, M.N. Lemus, M.L. Serrano, J. Zabaleta, A.L. Combita, M. C. Sanabria-Salas, Metabolic pathways enriched according to ERG status are associated with biochemical recurrence in Hispanic/Latino patients with prostate cancer, Cancer Med. 12 (2022) 4306–4320, https://doi.org/10.1002/cam4.5301.

[51] N.I. Simon, C. Parker, T.A. Hope, C.J. Paller, Best approaches and updates for prostate cancer biochemical recurrence, Am. Soc. Clin. Oncol. Educ. Book Am. Soc. Clin. Oncol. Annu. Meet 42 (2022) 1–8, https://doi.org/10.1200/EDBK_351033.

[52] L. Zhou, R. Fan, Y. Luo, C. Zhang, D. Jia, R. Wang, Y. Zeng, M. Ren, K. Du, W. Pan, J. Yang, F. Tian, C. Gu, A metabolism-related gene landscape predicts prostate cancer recurrence and treatment response, Front. Immunol. 13 (2022) 837991, https://doi.org/10.3389/fimmu.2022.837991.

[53] Y. Liang, X. Zhang, C. Ma, J. Hu, m6A methylation regulators are predictive biomarkers for tumour metastasis in prostate cancer, Cancers 14 (2022) 4035, https://doi.org/10.3390/cancers14164035.

[54] K. Shimada, J.A. Bachman, J.L. Muhlich, T.J. Mitchison, shinyDepMap, a tool to identify targetable cancer genes and their functional connections from Cancer Dependency Map data, Elife 10 (n.d.) e57116. https://doi.org/10.7554/eLife.57116..

[55] X. He, Y. Jiao, X. Yang, Y. Hu, A novel prediction tool for overall survival of patients living with spinal metastatic disease, World Neurosurg. 144 (2020) e824–e836, https://doi.org/10.1016/j.wneu.2020.09.081.