Cancer Cell International

RESEARCH Open Access

Integrating multi-cohort machine learning and clinical sample validation to explore peripheral blood mRNA diagnostic biomarkers for prostate cancer

Xingyu Zhong[1†], Yuxuan Yang[1†], Haodong He[1], Yifan Xiong[1], Mingliang Zhong[1], Shaogang Wang[1*†] and Qidong Xia[1*†]

Abstract

Background The global incidence of prostate cancer (PCa) has been rising annually, and early diagnosis and treatment remain pivotal for improving therapeutic outcomes and patient prognosis. Concurrently, advancements in liquid biopsy technology have facilitated disease diagnosis and monitoring, with its minimally invasive nature and low heterogeneity positioning it as a promising approach for predicting disease progression. However, current liquid biopsy strategies for PCa predominantly rely on prostate-specific antigen (PSA), which lacks specificity and compromises diagnostic accuracy. Thus, there is an urgent need to identify novel liquid biopsy biomarkers to enable early and precise PCa diagnosis.

Methods We integrated 12 machine learning algorithms to construct 113 combinatorial models, screening and validating an optimal PCa diagnostic panel across five datasets from TCGA and GEO databases. Subsequently, the biological feasibility of the selected predictive model was verified in one prostate epithelial cell line and five PCa cell lines. Robust RNA diagnostic targets were further validated for their expression in plasma samples to establish an RNA-based liquid biopsy strategy for PCa. Finally, plasma samples from PCa and benign prostatic hyperplasia (BPH) patients at Wuhan Tongji Hospital were collected to evaluate the strategy's clinical significance.

Results Differential analysis identified 1,071 candidate mRNAs, which were input into the integrated machine learning framework. Among the 113 combinatorial models, the 9-gene diagnostic panel selected by the Stepglm[both] and Enet[alpha = 0.4] algorithms demonstrated the highest diagnostic efficacy (mean AUC = 0.91),

†Co-first authorship: Xingyu Zhong and Yuxuan Yang are contributed equally to this work.

†Co-correspondence authorship: Shaogang Wang and Qidong Xia are contributed equally to this work.

*Correspondence:
Shaogang Wang
sgwangtjm@163.com
Qidong Xia
qidongxia_md@163.com

Full list of author information is available at the end of the article

including *JPH4*, *RASL12*, *AOX1*, *SLC18A2*, *PDZRN4*, *P2RY2*, *B3GNT8*, *KCNQ5*, and *APOBEC3C*. Cell line experiments further validated *AOX1* and *B3GNT8* as robust RNA biomarkers, both exhibiting consistent PCa-specific expression in human plasma samples. In liquid biopsy analyses, *AOX1* and *B3GNT8* outperformed PSA in diagnostic accuracy, achieving a combined AUC of 0.91. Notably, these biomarkers also demonstrated diagnostic utility in patients with ISUP ≤ 2.

Conclusions Through an integrated machine learning approach and clinical validation, we developed an RNA-based diagnostic panel for PCa. Specifically, we identified *AOX1* and *B3GNT8* as novel liquid biopsy biomarkers with promising clinical diagnostic value. These findings provide new targets and insights for early and precise PCa diagnosis.

Keywords Prostate cancer, Liquid biopsy, Cell-free RNA, Biomarker, Precision diagnostic

Introduction

Prostate cancer (PCa), the second most prevalent malignancy in males, demonstrates a steadily increasing global incidence [1–3]. Compared to localized disease, advanced PCa is associated with constrained therapeutic options and diminished survival outcomes, underscoring the imperative for early and accurate detection [4, 5]. Current diagnostic modalities—including digital rectal examination (DRE), prostate-specific antigen (PSA) testing, imaging techniques (e.g., MRI), and histopathological biopsy—are hampered by critical limitations [6]. DRE exhibits substantial operator dependency and low sensitivity for early-stage lesions [7]. Advanced imaging, while improving diagnostic accuracy, requires specialized infrastructure and is too costly for large-scale screening. Although biopsy remains the gold standard for diagnosis, it carries risks of complications and potential tumor dissemination, making it unsuitable for routine use. As a result, liquid biopsy platforms, exemplified by PSA testing, have emerged as a minimally invasive and increasingly preferred tool for PCa surveillance.

Liquid biopsy refers to a disease monitoring method utilizing biological fluids such as blood, urine, and saliva as samples [8, 9]. It offers advantages including minimal invasiveness, low cost, operational simplicity, and reduced heterogeneity, and has been extensively researched and applied in recent years [10, 11]. For PCa diagnosis, serum PSA testing remains the most established liquid biopsy approach, frequently employed even in routine health screenings [12, 13]. However, the low specificity of PSA often leads to overdiagnosis and overtreatment when used in isolation [12, 14]. To address this limitation, researchers have explored various alternative biomarkers, though most remain in the validation phase. For instance, at the protein level, liquid biopsy strategies based on P2PSA and the Prostate Health Index (PHI) have demonstrated superior diagnostic performance compared to PSA [15]. At the genomic level, alterations such as DNA methylation and mutations have been linked to PCa progression and prognosis [16, 17]. Additionally, transcriptomic-level RNAs, which reflect real-time gene expression states and exhibit strong associations with disease initiation and progression, represent promising targets for sensitive and specific diagnosis [18, 19]. In recent years, significant progress has been made in the study of RNA biomarkers in PCa liquid biopsy, with different types of RNA demonstrating potential value in early diagnosis and disease monitoring. For example, prostate cancer antigen 3 (PCA3) is a long non-coding RNA (lncRNA) specifically expressed in prostate cancer cells, and its urine-based detection has been approved by the FDA as an adjunct to PSA testing to improve diagnostic accuracy [20]. Additionally, other lncRNAs, such as SChLAP1 and MALAT1, have shown promising diagnostic and prognostic value in blood and urine samples from PCa patients [21, 22]. In the field of microRNAs (miRNAs), molecules such as miR-141, miR-375 are abnormally expressed in the serum and urine of PCa patients and have been associated with tumor staging, invasiveness, and prognosis [23]. Furthermore, circular RNAs (circRNAs) have recently gained attention, with some circRNAs significantly upregulated in the blood of PCa patients, suggesting their potential as novel biomarkers [24]. Notably, with recent advancements in detection technologies, cell-free RNA (cfRNA) has garnered increasing attention, and its diagnostic value is gradually being validated [25, 26]. Studies have shown that specific expression patterns of cfRNA can distinguish PCa patients from those with benign prostate diseases and demonstrate high clinical value in predicting tumor progression [27]. Additionally, multi-gene RNA-based diagnostic panels have been proposed to further enhance diagnostic specificity and sensitivity. For example, urine-based RNA tests such as SelectMDx and ExoDx Prostate have entered the clinical validation stage [28, 29]. However, current RNA biomarkers for PCa diagnosis predominantly focus on single, mechanistically prominent RNAs directly associated with tumorigenesis, lacking rigorous screening processes, which compromises diagnostic specificity [22, 30]. Although multi-gene panels represent a potential solution, their clinical robustness is often compromised by suboptimal selection methods and insufficient validation across diverse cohorts. Furthermore, the limited application of RNA biomarkers in PCa liquid biopsy has hindered their clinical translation. Therefore, future research should focus on developing large-scale, multi-center validation studies and integrating bioinformatics approaches such as machine learning to establish more stable and reproducible RNA biomarker detection systems, thereby advancing the precision application of PCa liquid biopsy.

This study establishes a machine learning framework integrating 113 combinatorial algorithms to identify robust mRNA signatures using TCGA and four GEO cohorts, followed by clinical validation in peripheral blood specimens. Our work pioneers a novel diagnostic paradigm that may enable non-invasive and precise PCa detection through transcriptomic liquid biopsy, ultimately informing personalized therapeutic strategies.

Methods

Collection and processing of public databases

Transcriptomic data from a total of 1,096 PCa patients across five cohorts (TCGA-PRAD, GSE94767, GSE200879, GSE229904, and GSE246067) were collected from the TCGA and GEO databases to construct and validate our diagnostic signature. Among these, 502 patients from TCGA were designated as the training set for panel construction, while 594 patients from the four GEO datasets served as the validation set to evaluate diagnostic performance.

The RNA-seq read count matrix from TCGA, comprising 59,428 coding RNAs, was retrieved. Differential expression analysis was performed using the R packages "DESeq2," "edgeR," and "limma," with thresholds set at $|\text{logFC}| > 1.5$ and p-value < 0.01. Genes identified as differentially expressed by all three methods were selected and visualized via Venn diagrams. Gene expression matrices from the four GEO datasets were intersected with the TCGA-PRAD results to identify common genes across all five datasets.

Diagnostic panel generation via integrated machine learning approaches

We employed 12 machine learning algorithms to generate 113 combinatorial configurations, integrating feature selection and predictive modeling methods, including Lasso, Ridge, Elastic Net (Enet), Stepglm, SVM, glmBoost, LDA, plsRglm, RandomForest, GBM, XGBoost, and NaiveBayes. A two-step approach was implemented, where one algorithm was used for feature selection and another for model construction. Feature selection was performed using LASSO, Ridge, and Elastic Net, with the optimal λ parameter determined through 10-fold cross-validation to minimize the mean squared error. In predictive modeling, SVM with an RBF kernel was applied, and hyperparameters (C and γ) were optimized through grid search. The RandomForest model was built using 1000 decision trees. XGBoost hyperparameters were fine-tuned using Bayesian optimization, with a learning rate of 0.1 and a maximum tree depth of 2. Model performance was evaluated by calculating the area under the receiver operating characteristic curve (AUC) on external validation datasets. The optimal model was defined as the one achieving the highest average AUC across multiple datasets.

Cell lines processing

One prostate epithelial cell line (RWPE-1) and five PCa cell lines (22RV1, C4-2, DU 145, LNCaP, PC-3) were obtained from the American Type Culture Collection (ATCC, VA, USA). All cell lines were cultured in RPMI-1640 medium (Gibco, USA) supplemented with 10% fetal bovine serum (FBS, Gibco, USA) and 1% penicillin/streptomycin (Gibco, USA) at 37 °C in a humidified incubator with 5% CO_2. To minimize genetic drift, cells were maintained between passages 5 and 10. Cell line authentication was confirmed via STR profiling (PowerPlex 21 PCR Kit, Promega, USA) and cross-checked with ATCC and DSMZ databases. All cell lines tested negative for mycoplasma (MycoAlert™, Lonza, Switzerland) and bacterial contamination before use. Total RNA was extracted using the RNAsimple Total RNA Kit (Tiangen Biotech, China) and stored at – 80 °C for further analysis.

Collection and processing of human blood samples

From December 2024 to February 2025, we prospectively collected plasma samples from patients newly diagnosed with PCa or benign prostatic hyperplasia (BPH) at Tongji Hospital (Wuhan, China). All participants underwent PSA testing and needle biopsy, with exclusion criteria including infectious diseases and other malignancies. Preoperative samples were collected and centrifuged within 1 h (1,000× g, 15 min). The isolated plasma was stored at -80 °C for no more than 40 days. Clinical data, including age, PSA levels, and pathological biopsy results, were recorded. Plasma cell-free RNA was extracted using the miRNeasy Serum/Plasma Advanced Kit (QIAGEN, Germany), and stored at – 80 °C (no more than 40 days) (Figure S1). The study was approved by the Ethics Committee of Tongji Hospital (Wuhan, China) (TJ-IRB202407023), with informed consent obtained from all participants.

Quantitative real-time PCR (qRT-PCR)

qRT-PCR was performed using 2× Hieff® Ultra-Rapid HotStart PCR Master Mix (YEASEN, China) following the manufacturer's protocols. Primer sequences for the nine target RNAs and GAPDH are provided in Table S1. Table S2 and S3 shows the results of qRT-PCR.

Data analysis

All statistical graphs were generated using GraphPad Prism 8.0.1. ROC and regression analyses were performed in SPSS 29. Group mean differences were assessed via Student's t-test (two groups) or one-way ANOVA (multiple groups). $p < 0.05$ was considered statistically significant.

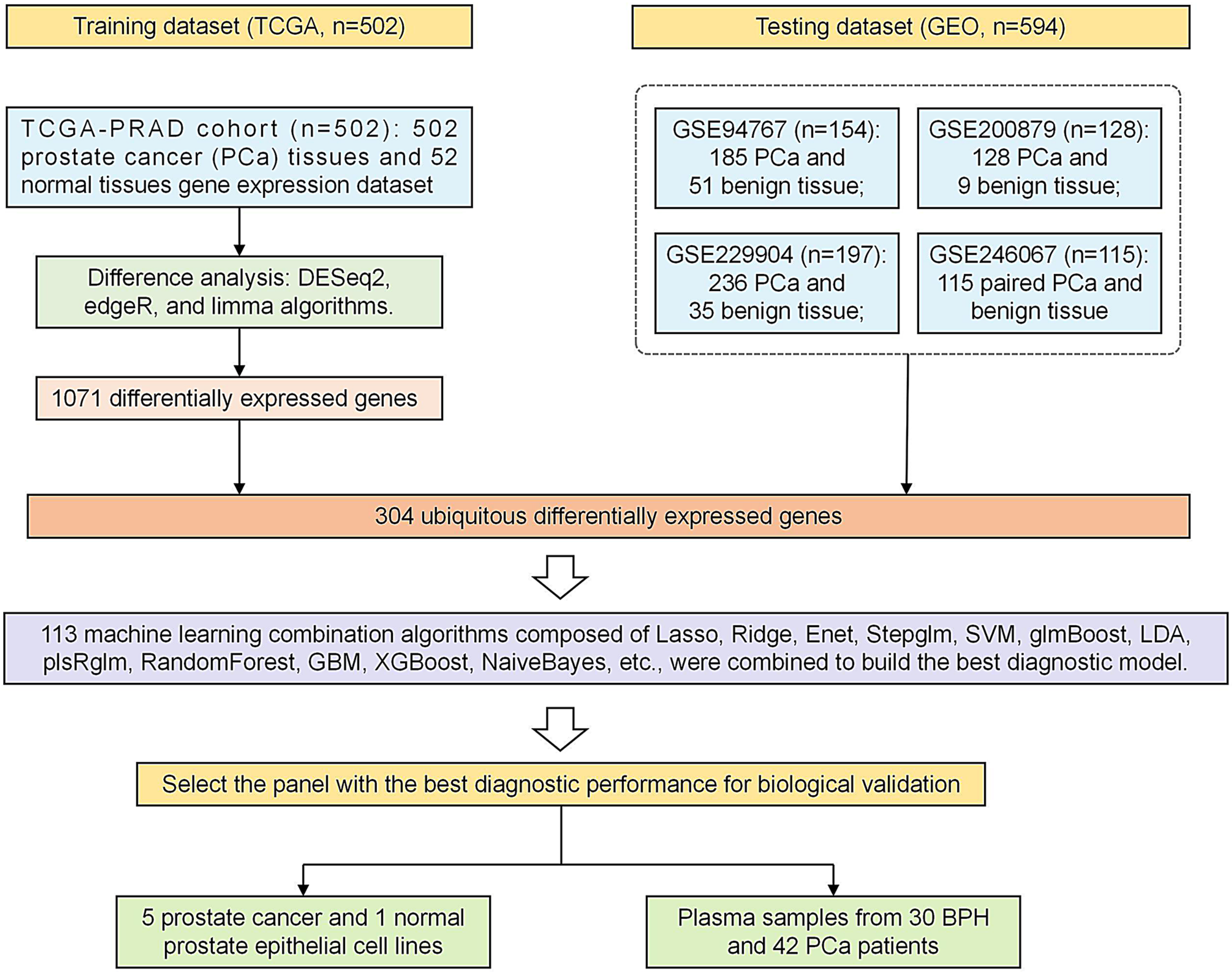
Results

Screening of differentially expressed genes in prostate cancer

The workflow of this study is illustrated in Fig. 1. To identify PCa-specific biomarkers, we first performed differential expression analysis on transcriptomic data from TCGA. After gene symbol annotation, 59,428 genes were included. Three methods—DESeq2, edgeR, and limma—were employed for differential gene screening. To balance the number of candidate genes (the inclusion of an excessive number of genes can exponentially increase the computational burden on machine learning algorithms and complicate the final diagnostic panel) and stability of differential expression, thresholds were set at |log2 fold change (logFC)| > 1.5 and p-value < 0.01. With DESeq2, edgeR, and limma, 3,060, 2,743, and 1,469 differentially expressed genes (DEGs) were identified, respectively. To ensure reliability, the intersection of 1,071 genes from three methods was selected for further analysis. The detailed visualization of the DEGs is provided in Figure S2 and Figure S3.

Integrated construction of the prostate cancer diagnostic panel

Direct utilization of a large number of DEGs as a diagnostic panel is impractical. Thus, we further refined these variables to construct an optimal predictive model. Twelve machine learning algorithms—Lasso, Ridge, Elastic Net (Enet), Stepwise Generalized Linear Model (Stepglm), Support Vector Machine (SVM), Generalized Linear Model Boosting (glmBoost), Linear Discriminant Analysis (LDA), Partial Least Squares Regression


Training dataset (TCGA, n=502)

TCGA-PRAD cohort (n=502): 502 prostate cancer (PCa) tissues and 52 normal tissues gene expression dataset

Difference analysis: DESeq2, edgeR, and limma algorithms.

1071 differentially expressed genes

Testing dataset (GEO, n=594)

GSE94767 (n=154): 185 PCa and 51 benign tissue;

GSE200879 (n=128): 128 PCa and 9 benign tissue;

GSE229904 (n=197): 236 PCa and 35 benign tissue;

GSE246067 (n=115): 115 paired PCa and benign tissue

304 ubiquitous differentially expressed genes

113 machine learning combination algorithms composed of Lasso, Ridge, Enet, Stepglm, SVM, glmBoost, LDA, plsRglm, RandomForest, GBM, XGBoost, NaiveBayes, etc., were combined to build the best diagnostic model.

Select the panel with the best diagnostic performance for biological validation

5 prostate cancer and 1 normal prostate epithelial cell lines

Plasma samples from 30 BPH and 42 PCa patients

Fig. 1 Flowchart of integrated machine learning and clinical sample validation for screening PCa liquid biopsy biomarkers

with GLM (plsRglm), Random Forest, Gradient Boosting Machine (GBM), eXtreme Gradient Boosting (XGBoost), and Naive Bayes—were systematically combined to generate 113 algorithmic configurations. Using TCGA data as the training set and four large GEO datasets (more than 100 patients each) for validation, the Stepglm[both] + Enet[alpha = 0.4] combination yielded a 9-gene diagnostic panel with optimal performance, achieving a mean AUC of 0.913 across all five datasets (Fig. 2a). Notably, all nine genes in this panel—*JPH4*, *RASL12*, *AOX1*, *SLC18A2*, *PDZRN4*, *P2RY2*, *B3GNT8*, *KCNQ5*, and *APOBEC3C*—exhibited tissue-specific downregulation in prostate cancer, suggesting that "protective role"-associated negative markers may hold unique diagnostic value for PCa. Subsequent differential expression analysis in paired clinical samples further validated the robustness of this 9-gene signature (Fig. 2b-j).

Validation of the gene signature in cell lines

To evaluate the biological applicability of the diagnostic panel, we validated the expression of the nine candidate genes in one prostate epithelial cell line (RWPE-1) and five PCa cell lines (22RV1, C4-2, DU 145, LNCaP, and PC-3). The detailed results of genes expression can be found in Fig. 3. While all genes showed significant downregulation in at least two prostate cancer cell lines, only *AOX1* and *B3GNT8* exhibited consistent low expression across all tested PCa lines. Notably, divergent or opposing expression patterns were observed among the remaining cell lines (excluding LNCaP), highlighting transcriptional heterogeneity across PCa subtypes of distinct origins.

RNA biomarkers for liquid biopsy and clinical significance

Based on these findings, *AOX1* and *B3GNT8* were confirmed as stably downregulated in PCa, the specific operational workflow can be seen in Fig. 4a. To evaluate their

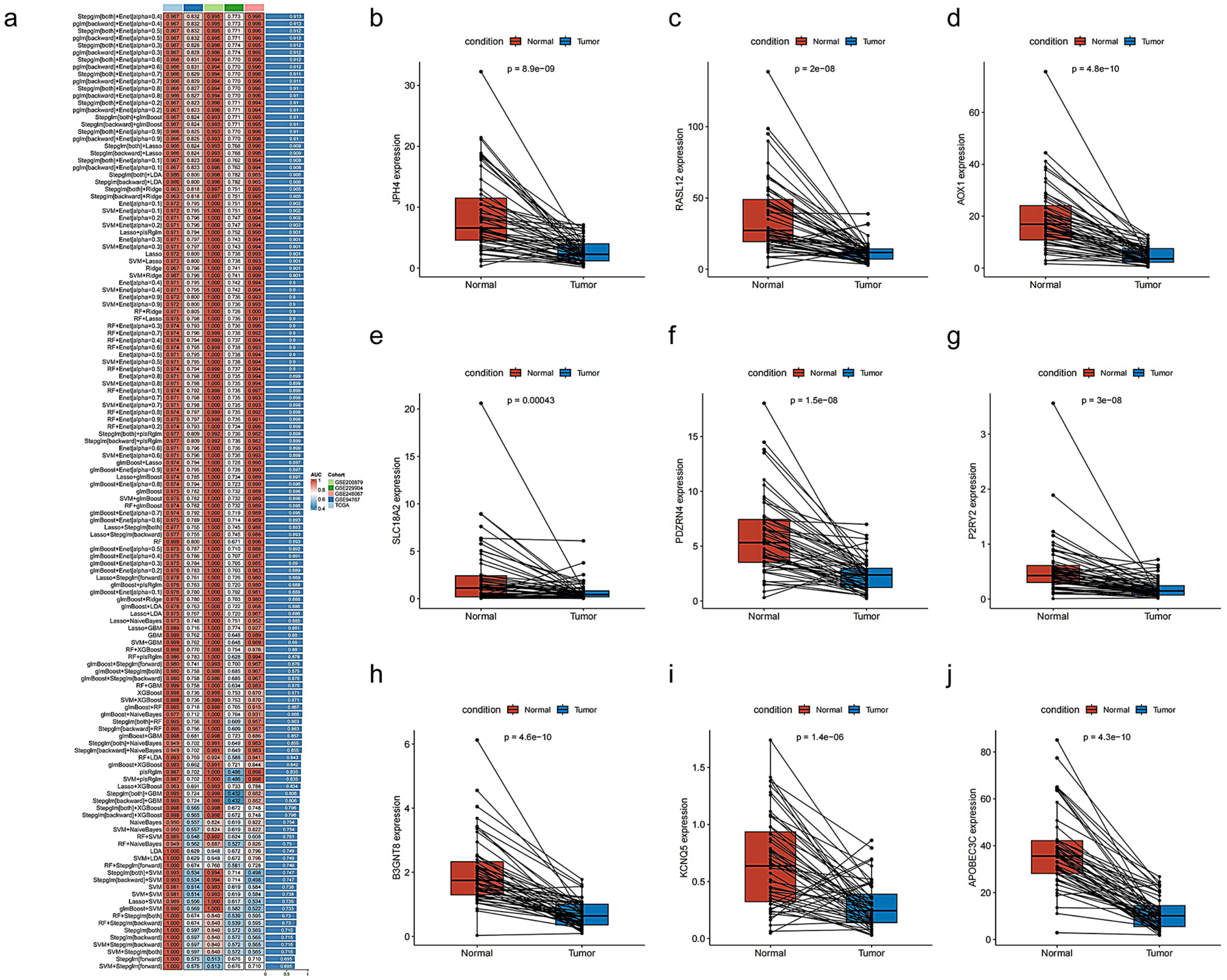


Fig. 2 (**a**) Performance of 113 combinatorial machine learning algorithms in constructing PCa classification models, with AUC values calculated in training and validation sets. (**b**-**j**) Validation of the 9-gene diagnostic panel in TCGA-PRAD paired samples

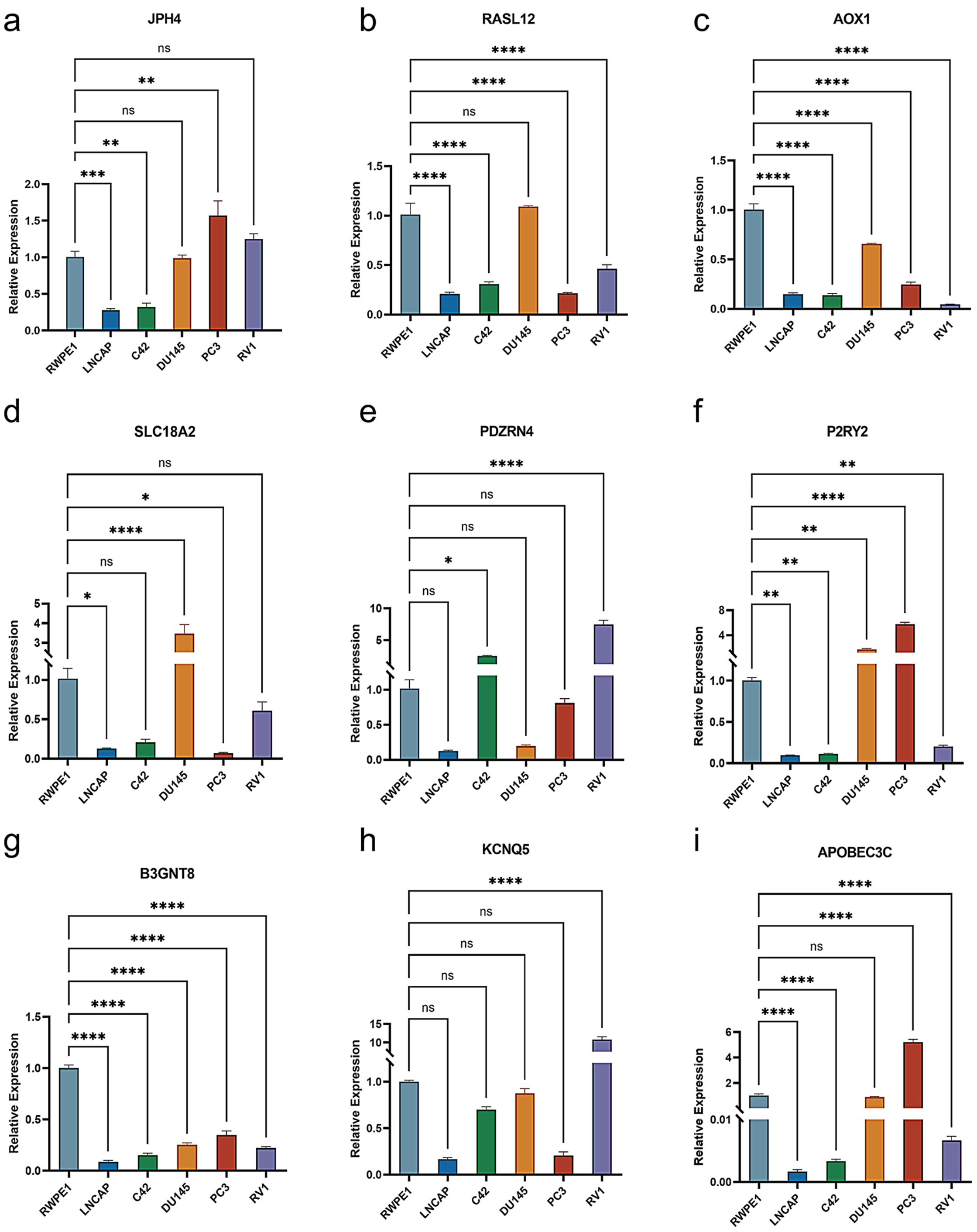


Fig. 3 Gene expression validation of *JPH4* (**a**), *RASL12* (**b**), *AOX1* (**c**), *SLC18A2* (**d**), *PDZRN4* (**e**), *P2RY2* (**f**), *B3GNT8* (**g**), *KCNQ5* (**h**), and *APOBEC3C* (**i**) in one prostate epithelial cell line and five PCa cell lines. $^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$, $^{****}p<0.0001$

clinical diagnostic utility, we further investigated their expression in plasma using blood samples from 30 BPH and 42 PCa patients (Table 1). Plasma cell-free RNA was extracted, and RT-PCR revealed significant downregulation of *AOX1* and *B3GNT8* in PCa plasma (Fig. 4b-c), supporting their potential as liquid biopsy biomarkers. Notably, to validate the robustness of our diagnostic approach, we incorporated 26 RNA samples (10 BPH and 16 PCa) with prolonged storage durations exceeding 30 days. These long-term preserved samples demonstrated gene expression profiles consistent with fresh specimens (Figure S4). ROC curve analysis demonstrated diagnostic AUCs of 0.88 (*AOX1*) and 0.79 (*B3GNT8*), both surpassing the performance of PSA (AUC = 0.66). Notably,

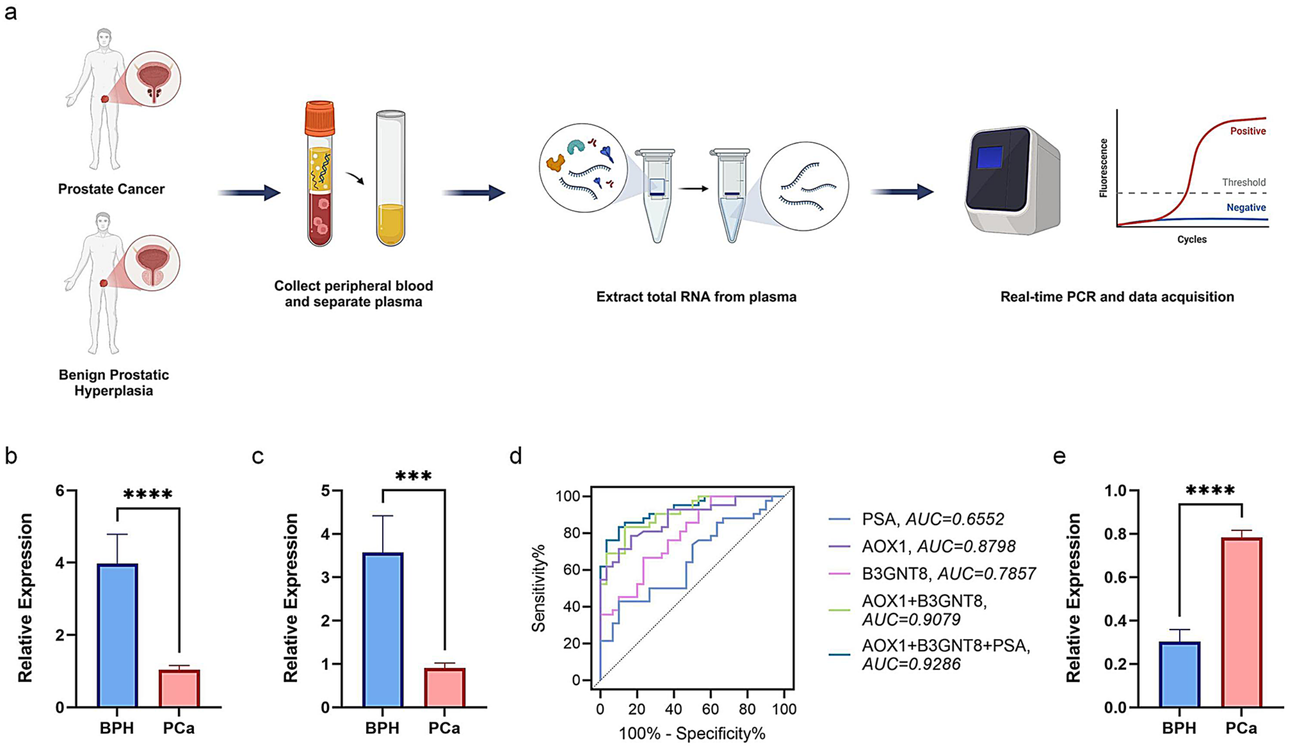


Fig. 4 (**a**) Workflow of liquid biopsy analysis using PCa plasma samples. (**b**-**c**) Validation of *AOX1* (**b**) and *B3GNT8* (**c**) expression in human plasma. (**d**) ROC curves comparing diagnostic performance of liquid biopsy biomarkers. (**e**) Combined expression levels of *AOX1* and *B3GNT8* in plasma from BPH and PCa patients. $^{***}p<0.001$, $^{****}p<0.0001$

Table 1 Clinical characteristics of prostate cancer (PCa) and benign prostatic hyperplasia (BPH) patients

	PCa	BPH
Number of patients	42	30
Age (year), median (range)	67.33 (46–75)	66.27 (51–87)
PSA (ng/mL)		
0–4	0	3
4–10	25	20
> 10	17	7
ISUP		
1	2	
2	19	
3	10	
4	6	
5	5	

combining both biomarkers elevated the AUC to 0.91, reinforcing their synergistic diagnostic power (Fig. 4d).

To further assess their clinical value, we analyzed their expression in early-stage PCa (ISUP grade ≤ 2). While neither gene alone achieved specificity for PCa detection, their combined expression showed significant differential expression between BPH and early PC, suggesting their utility in PCa's early (low-risk stage) diagnosis (Fig. 4e and S5). We further investigated the diagnostic performance of these biomarkers in patients with PSA levels 4–10 ng/mL (commonly regarded as the "gray zone"). Among 25 PCa and 20 BPH cases, *AOX1* and *B3GNT8* maintained significantly lower expression levels in PCa patients (Figure S6a-b). Within this PSA range, while PSA itself showed limited diagnostic value (AUC = 0.57), both *AOX1* (AUC = 0.86) and *B3GNT8* (AUC = 0.81) demonstrated superior discriminatory capacity. Their combination achieved enhanced diagnostic performance (AUC = 0.89), showing complementary diagnostic significance to PSA (Figure S6c). Furthermore, in clinically significant PCa (csPCa), both *AOX1* and *B3GNT8* demonstrated higher diagnostic accuracy than conventional PSA testing. And *AOX1* exhibited a significant correlation with ISUP grading ($p<0.01$) (Fig. 5), suggesting its potential as a predictive biomarker for PCa risk stratification.

Discussion

PCa is one of the most common malignant tumors worldwide. With population aging and advancements in detection methods, its incidence continues to rise [31]. Although PSA screening is widely used in clinical practice, its specificity and sensitivity are relatively low, often leading to over-biopsy and overtreatment [32, 33]. Therefore, there is an urgent need for more precise molecular diagnostic biomarkers. The emergence of transcriptome analysis has provided important molecular insights into the development and progression of PCa. However,

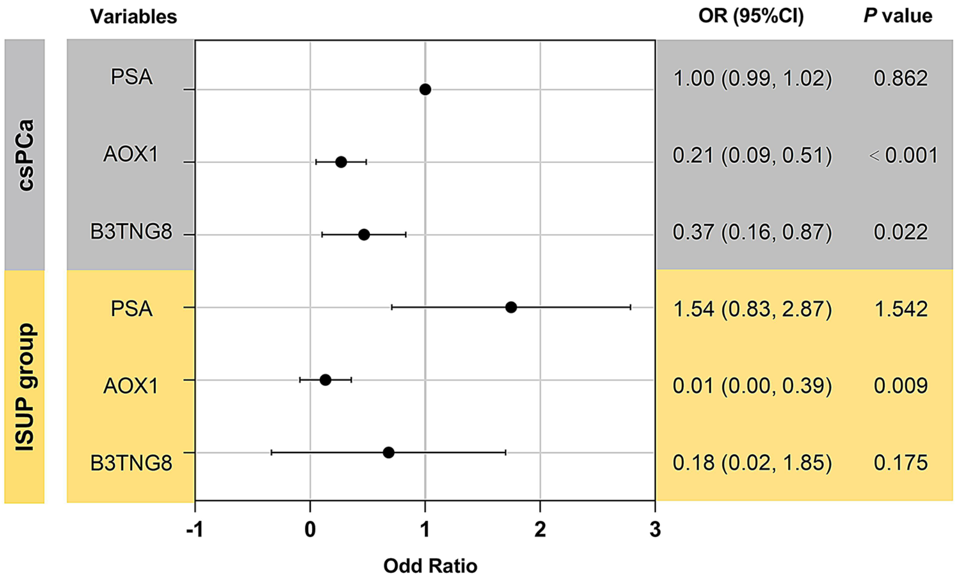


Fig. 5 Logistic regression analysis and forest plot evaluating the diagnostic value of PSA, *AOX1*, and *B3GNT8* for clinically significant PCa and ISUP grading

previous studies have largely relied on single-cohort data or traditional statistical methods, making it difficult to achieve reproducible validation in independent datasets, thereby limiting the generalizability and clinical applicability of the models.

In this study, we leveraged multi-cohort transcriptome data from TCGA and GEO and innovatively integrated 113 machine learning algorithm combinations to optimize feature selection, developing a robust and efficient PCa diagnostic gene panel. This approach overcomes the limitations of previous studies that relied heavily on single algorithms or subjective gene selection. To further validate its robustness, we performed external validation across four independent GEO datasets. The results demonstrated consistently high diagnostic performance, with AUC values exceeding 0.75 in all datasets and surpassing 0.90 in some cases, significantly outperforming single-gene markers and traditional PSA testing [34]. Further feature selection analysis identified *AOX1* and *B3GNT8* as key genes, both of which exhibited significant differential expression between PCa and benign tissues. Notably, both genes showed consistent downregulation in PCa across all tested cell lines (22RV1, C4-2, DU 145, LNCaP, and PC-3), highlighting their stable expression patterns and their potential as reliable biomarkers for PCa diagnosis.

Previous studies further highlight the potential role of *AOX1* and *B3GNT8* downregulation in PCa progression. These genes are implicated in tumor development through distinct biological mechanisms: *AOX1* plays a crucial role in cell metabolism, redox homeostasis, and tumor microenvironment regulation, while *B3GNT8* is involved in glycosylation modifications, cell signaling, and tumor microenvironment remodeling. Their downregulation has been linked to PCa development, although the exact mechanisms remain under investigation. Specifically, *AOX1* (Aldehyde Oxidase 1) is significantly downregulated in prostate cancer, which may disrupt androgen metabolism and oxidative stress balance, thereby contributing to tumor progression [35]. Moreover, *AOX1* downregulation has been confirmed to result from methylation modifications [36] and is closely associated with tumor dedifferentiation, increased invasiveness, and poor prognosis [37]. A potential mechanism involves the activation of tryptophan metabolism, which may facilitate tumor immune evasion and the acquisition of drug resistance [38]. On the other hand, *B3GNT8* (β-1,3-N-acetylglucosaminyltransferase 8), a key glycosyltransferase, plays a crucial role in glycosylation modifications, cell signaling transduction, and tumor microenvironment remodeling. Aberrant glycosylation has been recognized as an important molecular feature of prostate cancer [39, 40]. However, research on B3GNT8 in prostate cancer remains limited. Existing studies have primarily focused on colorectal cancer, where *B3GNT8* is found to be widely upregulated [41]. Further mechanistic investigations suggest that *B3GNT8*-mediated aberrant glycosylation can regulate key signaling pathways, promoting tumor cell survival, drug resistance, and invasion [42]. However, some studies have also suggested that *B3GNT8* may exert protective effects against biological aging [43]. These findings indicate that the function of *B3GNT8* may be tumor-specific, and its precise role and underlying mechanisms in prostate cancer require further investigation.

To further validate the generalizability of our findings, particularly their applicability to the Chinese population, we collected plasma samples from hospitalized patients at Wuhan Tongji Hospital for independent validation and assessed the clinical potential of this gene panel in non-invasive PCa detection. Plasma analysis revealed that *AOX1* and *B3GNT8* expression levels were significantly lower in PCa patients compared to healthy controls, consistent with their expression patterns in tissue samples. Compared to traditional PSA-based plasma biomarkers, the combined detection of *AOX1*, *B3GNT8*, and PSA achieved an AUC consistently above 0.90, significantly enhancing the precise diagnostic capability for PCa. Notably, our subgroup analysis specifically focusing on the PSA "gray zone" patients further revealed their supplementary diagnostic value to PSA, highlighting their potential for early prostate cancer detection. Overall, the *AOX1*+*B3GNT8* panel demonstrated superior diagnostic performance over single-gene or PSA testing (AUC range: 0.66–0.79) and enabled more precise PCa diagnosis through both tissue and plasma detection. This discovery not only deepens our understanding of the molecular mechanisms underlying PCa but also offers a novel strategy for precise detection and non-invasive screening, holding significant clinical translational potential.

Despite its multiple strengths, this study has several limitations. First, our plasma sample size was relatively small ($n=72$), and future studies should expand the cohort to assess the stability of this non-invasive detection method. Second, the biological mechanisms linking *AOX1* and *B3GNT8* to PCa progression remain incompletely understood, necessitating further functional studies. Moreover, long-term follow-up has not been conducted in this cohort, precluding further assessment of the prognostic value of the identified biomarkers. Future research should focus on multi-omics integration, larger clinical studies, functional validation, and AI-assisted diagnostics to facilitate the clinical translation of our findings. In conclusion, this study leveraged machine learning-optimized multi-cohort transcriptomic analysis to identify an accurate PCa diagnostic gene panel, laying a critical foundation for precise diagnosis and personalized screening of PCa.

Conclusion

In this study, we constructed a 9-gene PCa classification model using integrated machine learning algorithms, which exhibited robust diagnostic performance. Furthermore, *AOX1* and *B3GNT8* were validated as PCa-specific RNA biomarkers in plasma samples, demonstrating predictive value for both diagnosis and stratification. The two genes showed higher diagnostic accuracy than PSA in csPCa and could serve as complementary biomarkers to enhance PSA-based screening. Additionally, *AOX1* exhibited a significant correlation with ISUP grading, suggesting its potential for PCa risk stratification. Overall, this work establishes a reliable RNA-based diagnostic framework for PCa and proposes novel targets to advance liquid biopsy applications in clinical practice.

Abbreviations

AOX1	Aldehyde Oxidase 1
AUC	Area Under The Curve
B3GNT8	β-1,3-N-acetylglucosaminyltransferase 8
BPH	Benign Prostatic Hyperplasia
cfRNA	cell-free RNA
csPCa	clinically significant PCa
DEGs	Differentially Expressed Genes
DRE	Digital Rectal Examination
Enet	Elastic Net
GBM	Gradient Boosting Machine
glmBoost	Generalized Linear Model Boosting
LDA	Linear Discriminant Analysis
PCa	Prostate Cancer
PHI	Prostate Health Index
plsRglm	Partial Least Squares Regression with GLM
PSA	Prostate-Specific Antigen
qRT-PCR	Quantitative Real-Time PCR
Stepglm	Stepwise Generalized Linear Model
SVM	Support Vector Machine
XGBoost	eXtreme Gradient Boosting

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12935-025-03788-w.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

XZ and QX conceived the study design and performed the machine learning analysis. XZ and YY conducted the qRT-PCR experiments and drafted the original manuscript. HH collected and processed patient blood samples. YX and MZ performed statistical analysis and interpreted clinical data. SW and QX reviewed and edited the manuscript.

Funding

This research is supported by grants from Hubei Urology Minimally Invasive Therapy Clinical Medical Center (SCZ2023015 and SCZ2023017) and grants from Natural Science Foundation of Hubei Province (ZRMS2020002466).

Data availability

Data is provided within the manuscript or supplementary information files.

Declarations

Ethics approval and consent to participate

Informed consent and approval were obtained from all the patients and the Ethics Committee of Tongji Hospital (TJ-IRB202407023).

Consent for publication

All authors approved the final version of the manuscript and the submission to this journal.

Competing interests

The authors declare no competing interests.

Author details

[1]Department and Institute of Urology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, No. 1095 Jiefang Avenue, Wuhan 430030, P.R. China

Received: 4 March 2025 / Accepted: 12 April 2025
Published online: 22 April 2025

References

1. Cornford P, van den Bergh RCN, Briers E, Van den Broeck T, Brunckhorst O, Darraugh J, Eberli D, De Meerleer G, De Santis M, Farolfi A, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate Cancer-2024 update. Part I: screening, diagnosis, and local treatment with curative intent. Eur Urol. 2024;86(2):148–63.
2. Schaeffer EM, Srinivas S, Adra N, An Y, Barocas D, Bitting R, Bryce A, Chapin B, Cheng HH, D'Amico AV, et al. Prostate cancer, version 4.2023, NCCN clinical practice guidelines in oncology. J Natl Compr Canc Netw. 2023;21(10):1067–96.
3. James ND, Tannock I, N'Dow J, Feng F, Gillessen S, Ali SA, Trujillo B, Al-Lazikani B, Attard G, Bray F, et al. The lancet commission on prostate cancer: planning for the surge in cases. Lancet. 2024;403(10437):1683–722.
4. Wei JT, Barocas D, Carlsson S, Coakley F, Eggener S, Etzioni R, Fine SW, Han M, Kim SK, Kirkby E, et al. Early detection of prostate cancer: AUA/SUO guideline part I: prostate Cancer screening. J Urol. 2023;210(1):46–53.
5. Papachristodoulou A, Abate-Shen C. Precision intervention for prostate cancer: Re-evaluating who is at risk. Cancer Lett. 2022;538:215709.
6. Sarkar S, Gogoi M, Mahato M, Joshi AB, Baruah AJ, Kodgire P, Boruah P. Biosensors for detection of prostate cancer: a review. Biomed Microdevices. 2022;24(4):32.
7. Matsukawa A, Yanagisawa T, Bekku K, Kardoust Parizi M, Laukhtina E, Klemm J, Chiujdea S, Mori K, Kimura S, Fazekas T, et al. Comparing the performance of digital rectal examination and prostate-specific antigen as a screening test for prostate cancer: A systematic review and Meta-analysis. Eur Urol Oncol. 2024;7(4):697–704.
8. Nikanjam M, Kato S, Kurzrock R. Liquid biopsy: current technology and clinical applications. J Hematol Oncol. 2022;15(1):131.
9. Ma L, Guo H, Zhao Y, Liu Z, Wang C, Bu J, Sun T, Wei J. Liquid biopsy in cancer current: status, challenges and future prospects. Signal Transduct Target Ther. 2024;9(1):336.
10. Citarella A, Besharat ZM, Trocchianesi S, Autilio TM, Verrienti A, Catanzaro G, Splendiani E, Spinello Z, Cantara S, Zavattari P, et al. Circulating cell-free DNA (cfDNA) in patients with medullary thyroid carcinoma is characterized by specific fragmentation and methylation changes with diagnostic value. Biomark Res. 2023;11(1):82.
11. Baca SC, Seo JH, Davidsohn MP, Fortunato B, Semaan K, Sotudian S, Lakshminarayanan G, Diossy M, Qiu X, El Zarif T, et al. Liquid biopsy epigenomic profiling for cancer subtyping. Nat Med. 2023;29(11):2737–41.
12. Leenen RCA, Venderbos LDF, Helleman J, Gómez Rivas J, Vynckier P, Annemans L, Chloupková R, Májek O, Briers E, Vasilyeva V, et al. Prostate Cancer early detection in the European union and UK. Eur Urol. 2025;87(3):326–39.
13. Van Poppel H, Roobol MJ, Chapple CR, Catto JWF, N'Dow J, Sønksen J, Stenzl A, Wirth M. Prostate-specific antigen testing as part of a Risk-Adapted early detection strategy for prostate cancer: European association of urology position and recommendations for 2021. Eur Urol. 2021;80(6):703–11.
14. Pinsky PF, Parnes H. Screening for prostate Cancer. N Engl J Med. 2023;388(15):1405–14.
15. Agnello L, Vidali M, Giglio RV, Gambino CM, Ciaccio AM, Lo Sasso B, Ciaccio M. Prostate health index (PHI) as a reliable biomarker for prostate cancer: a systematic review and meta-analysis. Clin Chem Lab Med. 2022;60(8):1261–77.
16. Fonseca NM, Maurice-Dror C, Herberts C, Tu W, Fan W, Murtha AJ, Kollmannsberger C, Kwan EM, Parekh K, Schönlau E, et al. Prediction of plasma ctdna fraction and prognostic implications of liquid biopsy in advanced prostate cancer. Nat Commun. 2024;15(1):1828.
17. Herberts C, Annala M, Sipola J, Ng SWS, Chen XE, Nurminen A, Korhonen OV, Munzur AD, Beja K, Schönlau E, et al. Deep whole-genome ctdna chronology of treatment-resistant prostate cancer. Nature. 2022;608(7921):199–208.
18. Wang H, Meng Q, Qian J, Li M, Gu C, Yang Y. Review: RNA-based diagnostic markers discovery and therapeutic targets development in cancer. Pharmacol Ther. 2022;234:108123.
19. Zaher HS, Mosammaparast N. RNA damage responses in cellular homeostasis, genome stability, and disease. Annu Rev Pathol. 2025;20(1):433–57.
20. Wu D, Ni J, Beretov J, Cozzi P, Willcox M, Wasinger V, Walsh B, Graham P, Li Y. Urinary biomarkers in prostate cancer detection and monitoring progression. Crit Rev Oncol Hematol. 2017;118:15–26.
21. Mehra R, Udager AM, Ahearn TU, Cao X, Feng FY, Loda M, Petimar JS, Kantoff P, Mucci LA, Chinnaiyan AM. Overexpression of the long Non-coding RNA SChLAP1 independently predicts lethal prostate Cancer. Eur Urol. 2016;70(4):549–52.
22. Goyal B, Yadav SRM, Awasthee N, Gupta S, Kunnumakkara AB, Gupta SC. Diagnostic, prognostic, and therapeutic significance of long non-coding RNA MALAT1 in cancer. Biochim Biophys Acta Rev Cancer. 2021;1875(2):188502.
23. Martínez-González LJ, Sánchez-Conde V, González-Cabezuelo JM, Antunez-Rodríguez A, Andrés-León E, Robles-Fernandez I, Lorente JA, Vázquez-Alonso F, Alvarez-Cubero MJ. Identification of MicroRNAs as viable aggressiveness biomarkers for prostate Cancer. Biomedicines 2021, 9(6).
24. Zhang ZH, Wang Y, Zhang Y, Zheng SF, Feng T, Tian X, Abudurexiti M, Wang ZD, Zhu WK, Su JQ, et al. The function and mechanisms of action of circular RNAs in urologic Cancer. Mol Cancer. 2023;22(1):61.
25. Loy C, Ahmann L, De Vlaminck I, Gu W. Liquid biopsy based on Cell-Free DNA and RNA. Annu Rev Biomed Eng. 2024;26(1):169–95.
26. Moufarrej MN, Vorperian SK, Wong RJ, Campos AA, Quaintance CC, Sit RV, Tan M, Detweiler AM, Mekonen H, Neff NF, et al. Early prediction of preeclampsia in pregnancy with cell-free RNA. Nature. 2022;602(7898):689–94.
27. Mugoni V, Ciani Y, Nardella C, Demichelis F. Circulating RNAs in prostate cancer patients. Cancer Lett. 2022;524:57–69.
28. Hendriks RJ, van der Leest MMG, Israël B, Hannink G, YantiSetiasti A, Cornel EB, Hulsbergen-van de Kaa CA, Klaver OS, Sedelaar JPM, Van Criekinge W, et al. Clinical use of the SelectMDx urinary-biomarker test with or without MpMRI in prostate cancer diagnosis: a prospective, multicenter study in biopsy-naïve men. Prostate Cancer Prostatic Dis. 2021;24(4):1110–9.
29. Tutrone R, Lowentritt B, Neuman B, Donovan MJ, Hallmark E, Cole TJ, Yao Y, Biesecker C, Kumar S, Verma V, et al. ExoDx prostate test as a predictor of outcomes of high-grade prostate cancer - an interim analysis. Prostate Cancer Prostatic Dis. 2023;26(3):596–601.
30. Broomfield J, Kalofonou M, Pataillot-Meakin T, Powell SM, Fernandes RC, Moser N, Bevan CL, Georgiou P. Detection of YAP1 and AR-V7 mRNA for prostate Cancer prognosis using an ISFET Lab-On-Chip platform. ACS Sens. 2022;7(11):3389–98.
31. Moschini M, Carroll PR, Eggener SE, Epstein JI, Graefen M, Montironi R, Parker C. Low-risk prostate cancer: identification, management, and outcomes. Eur Urol. 2017;72(2):238–49.
32. Catalona WJ, Partin AW, Slawin KM, Brawer MK, Flanigan RC, Patel A, Richie JP, deKernion JB, Walsh PC, Scardino PT, et al. Use of the percentage of free prostate-specific antigen to enhance differentiation of prostate cancer from benign prostatic disease: a prospective multicenter clinical trial. JAMA. 1998;279(19):1542–7.
33. Etzioni R, Penson DF, Legler JM, di Tommaso D, Boer R, Gann PH, Feuer EJ. Overdiagnosis due to Prostate-specific antigen screening: lessons from U.S. Prostate cancer incidence trends. J Natl Cancer Inst. 2002;94(13):981–90.
34. Chiu PK, Shen X, Wang G, Ho CL, Leung CH, Ng CF, Choi KS, Teoh JY. Enhancement of prostate cancer diagnosis by machine learning techniques: an algorithm development and validation study. Prostate Cancer Prostatic Dis. 2022;25(4):672–6.
35. Mo X, Yuan K, Hu D, Huang C, Luo J, Liu H, Li Y. Identification and validation of immune-related hub genes based on machine learning in prostate cancer and AOX1 is an oxidative stress-related biomarker. Front Oncol. 2023;13:1179212.
36. Haldrup C, Mundbjerg K, Vestergaard EM, Lamy P, Wild P, Schulz WA, Arsov C, Visakorpi T, Borre M, Høyer S, et al. DNA methylation signatures for prediction of biochemical recurrence after radical prostatectomy of clinically localized prostate cancer. J Clin Oncol. 2013;31(26):3250–8.
37. Strand SH, Orntoft TF, Sorensen KD. Prognostic DNA methylation markers for prostate cancer. Int J Mol Sci. 2014;15(9):16544–76.
38. Vantaku V, Putluri V, Bader DA, Maity S, Ma J, Arnold JM, Rajapakshe K, Donepudi SR, von Rundstedt FC, Devarakonda V, et al. Epigenetic loss of AOX1 expression via EZH2 leads to metabolic deregulations and promotes bladder cancer progression. Oncogene. 2020;39(40):6265–85.
39. Scott E, Hodgson K, Calle B, Turner H, Cheung K, Bermudez A, Marques FJG, Pye H, Yo EC, Islam K, et al. Upregulation of GALNT7 in prostate cancer modifies O-glycosylation and promotes tumour growth. Oncogene. 2023;42(12):926–37.
40. Xu X, Peng Q, Jiang X, Tan S, Yang W, Han Y, Oyang L, Lin J, Shen M, Wang J, et al. Altered glycosylation in cancer: molecular functions and therapeutic potential. Cancer Commun (Lond). 2024;44(11):1316–36.
41. Shen L, Yu M, Xu X, Gao L, Ni J, Luo Z, Wu S. Knockdown of β3GnT8 reverses 5-fluorouracil resistance in human colorectal cancer cells via

Inhibition the biosynthesis of polylactosamine-type N-glycans. Int J Oncol. 2014;45(6):2560–8.

42. Xie P, Mo JL, Liu JH, Li X, Tan LM, Zhang W, Zhou HH, Liu ZQ. Pharmacogenomics of 5-fluorouracil in colorectal cancer: review and update. Cell Oncol (Dordr). 2020;43(6):989–1001.
43. Mao R, Li J, Xiao W. Identification of prospective aging drug targets via Mendelian randomization analysis. Aging Cell. 2024;23(7):e14171.

Publisher's note