International Journal of Biology and Life Sciences
ISSN: 2957-9511 | Vol. 8, No. 3, 2024

Identification of Potential Molecular Mechanisms and Machine Learning Classification Model in Prostate Cancer Progression Based on GEO Datasets and TCGA Dataset

Jiarui Zhang

Art and Science College, Vanderbilt University, Nashville 37235, United States

Abstract: Prostate cancer is a leading cause of cancer-related morbidity and mortality. Identifying key genes involved in its progression is crucial for improving clinical outcomes.This paper used transcriptomic data from the GEO database to analyze differential expression and employed machine learning techniques to identify critical progression genes and develop a grading model. And the results were validated by TCGA dataset. This paper identified 802 key genes associated with prostate cancer progression. The molecular mechanisms are involving immune and cell growth pathways. The developed machine learning grading model demonstrated high AUCs, indicating its potential utility in clinical settings. Those key genes play important roles in prostate cancer.

Keywords: Prostate Cancer; Machine Learning; Molecular Mechanisms.

1. Introduction

Prostate cancer is one of the most prevalent malignancies globally, ranking as the second most common cancer among men and significantly impacting public health (Siegel et al., 2012; Jemal et al., 2010). In many regions, it accounts for a substantial proportion of cancer diagnoses and mortality, particularly among older populations (White et al., 2014). Despite advancements in screening and treatment, challenges remain in clinical diagnosis, often due to the asymptomatic nature of early-stage disease and variability in prostate-specific antigen (PSA) levels (Fenton et al., 2018). These limitations can lead to delays in diagnosis, resulting in progression to advanced stages at the time of detection. Moreover, treatment options, while improving, face limitations in efficacy and adverse effects, prompting concerns regarding personalized care approaches. The heterogeneity of prostate cancer further complicates management, as not all patients respond uniformly to therapies (Tolkach & Kristiansen, 2018). Consequently, there is an urgent need for innovative diagnostic tools and tailored treatment strategies to enhance patient outcomes and mitigate the public health burden associated with this prevalent disease (Viegas et al., 2017).

Transcriptomic analysis plays a pivotal role in advancing our understanding of cancer biology by enabling researchers to examine gene expression patterns across various tumor types and stages (Lim et al., 2023). This approach allows for the identification of critical molecular pathways involved in tumorigenesis, metastasis, and therapeutic resistance, ultimately informing the development of targeted therapies and personalized medicine (Vitelli et al., 2023). The Gene Expression Omnibus (GEO) database serves as a valuable resource for cancer research, offering a vast repository of high-throughput gene expression data. Researchers can access a wealth of datasets that encompass diverse cancer types, experimental conditions, and treatment responses (Barrett & Edgar, 2006).

Identifying key genes associated with prostate cancer progression is crucial for advancing our understanding of the disease and improving clinical outcomes (Schrijver et al., 2018). Transcriptomic analysis provides a powerful tool to explore differential gene expression patterns that may be linked to various stages of cancer progression, from localized disease to metastatic states (Cornford et al., 2021). By integrating these findings, we can pinpoint specific genes that play critical roles in tumor growth, invasion, and therapeutic resistance, thus offering potential targets for intervention (Jung et al., 2007).

Furthermore, the application of machine learning techniques in this context presents an innovative approach to enhance clinical decision-making. By establishing a machine learning-based grading model, we aim to predict disease severity more accurately, utilizing the identified gene signatures as features (Rosenstein et al., 2024). This model can provide clinicians with valuable insights into patient prognosis, facilitating personalized treatment strategies tailored to individual risk profiles. Ultimately, this integrative approach not only enhances our understanding of the molecular underpinnings of prostate cancer but also holds the promise of transforming patient management through improved predictive capabilities.

2. Research Methods

2.1. Datasets and Acquisition

In this study, we sourced the raw messenger RNA (mRNA) expression dataset for prostate cancer from The Cancer Genome Atlas (TCGA) database, accessible at https://portal.gdc.cancer.gov/. This dataset comprises a total of 498 samples, allowing for a comprehensive analysis of the survival effect of genes. Additionally, we incorporated a gene expression dataset from the Gene Expression Omnibus (GEO) database, available at https://www.ncbi.nlm.nih.gov/geo/. Specifically, the series matrix file associated with SRP073789, which carried out RNA-seq based transcriptome profiling using 45 prostate samples with various disease

progression steps such as benign prostate hyperplasia (BPH), primary cancer of prostate (CaP), advanced CaP and CRPC. The data from both TCGA and GEO were freely accessible online, facilitating the identification of differentially expressed genes (DEGs) pertinent.

2.2. Functional Enrichment Analysis of the DEGs

To elucidate the biological significance of the identified DEGs, we conducted functional enrichment analysis using the Metascape database (www.metascape.org). This comprehensive platform enables the exploration of functional correlations among gene sets, integrating various types of biological data. We performed Gene Ontology (GO) analysis to identify processes, cellular components, and molecular functions associated with the DEGs, alongside Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis to map the genes onto specific metabolic and signaling pathways. Statistical significance was determined using criteria that included a minimum overlap of three genes and a p-value threshold of ≤ 0.01, ensuring that only robust associations were reported.

2.3. Machine Learning Grading Model

To differentiate among the four grades of tumor classification, we employed several machine learning algorithms, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), XGBoost. Each algorithm was selected based on its unique strengths in handling complex data patterns and classifying high-dimensional datasets.

The Random Forest algorithm, known for its robustness and ability to reduce overfitting through ensemble learning, was utilized to build a set of decision trees for accurate grade prediction. The SVM algorithm was employed to find the optimal hyperplane that separates the classes in high-dimensional space, making it particularly effective for this type of classification task. KNN was included for its simplicity and effectiveness in capturing local patterns within the data. XGBoost, which leverages gradient boosting techniques, was used for its high performance and efficiency in handling large datasets.

To ensure the reliability of our results, we conducted cross-validation and AUC calculation for each algorithm.

2.4. Statistical Analysis

All statistical analyses were conducted using R software (version 3.6), which offers a robust platform for data analysis and visualization. To identify prognostic genes, we utilized Cox regression analysis, allowing us to assess the impact of gene expression on patient survival outcomes. Survival analysis was performed using the Kaplan-Meier method, which provides insights into the survival distribution of different patient cohorts. Comparisons between groups were conducted using the Wilcoxon test, a non-parametric statistical test suitable for comparing two independent samples. In our analyses, a p-value threshold of < 0.05 was deemed statistically significant, underscoring the importance of rigor in our statistical evaluations.

3. Results

3.1. Multivariate Analysis

We performed a multivariate analysis using the UMAP (Uniform Manifold Approximation and Projection) dimensionality reduction algorithm to compare normal samples with tumor samples from three distinct stages of disease progression. The analysis revealed a significant separation between tumor samples and normal samples, with normal samples predominantly clustering in the upper left quadrant and tumor samples occupying the lower right quadrant of the UMAP plot. Notably, samples from typical prostate cancer exhibited a closer proximity to normal samples, while samples from more advanced stages of the disease progressively migrated toward the lower right quadrant, indicating a distinct alteration in the underlying biological state as the disease progresses (Figure 1).

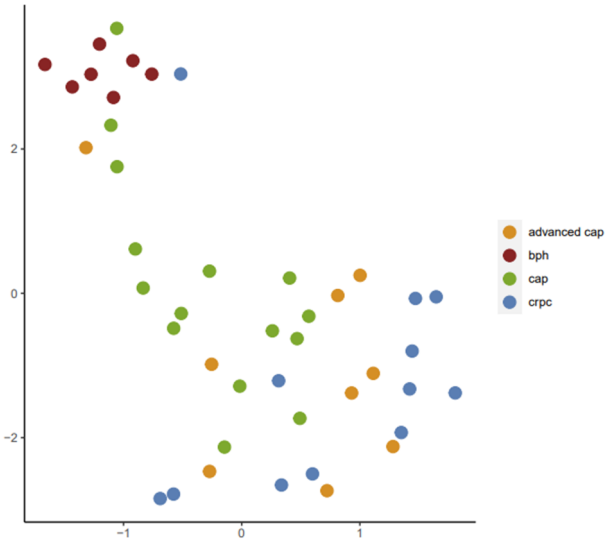


Fig 1. UMAP plot of the four types of samples

3.2. Univariate Analysis

Using the Wilcoxon rank-sum test, we identified a substantial number of differentially expressed genes between tumor and normal samples (Figure 2A-C). It is evident that there are minimal differences in gene expression between typical prostate cancer and normal samples; however, as the disease advances to more severe stages, such as advanced prostate cancer or castration-resistant prostate cancer (CRPC), the differential expression of genes becomes increasingly pronounced. Specifically, in the comparison between typical prostate cancer and normal samples, 5597 genes such as SERPINE2, NR1H4 were found to be upregulated, while 6877 genes such as FGR, TFPI was downregulated. In advanced prostate cancer versus normal samples, 5081 genes such as LRRK2 was upregulated, and 13456 genes including SLC25A13, CD22 was downregulated. Similarly, in the comparison between CRPC and normal samples, we observed upregulation of 5015 genes and downregulation of 13170 genes. The intersection of differentially expressed genes across all three comparisons yielded a total of 2,701 downregulated genes and 3,137 upregulated genes (Figure 2D).

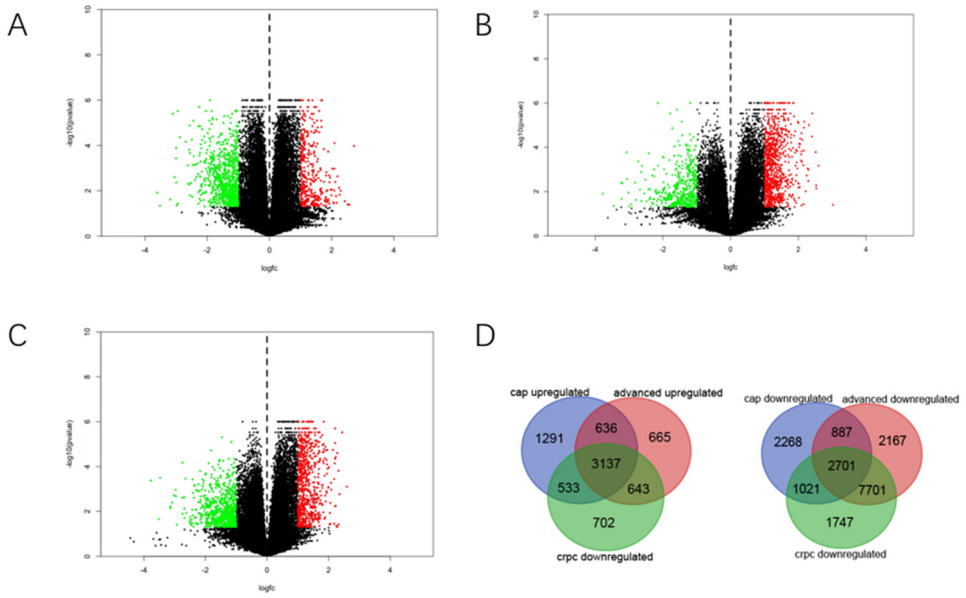


Fig 2. Results of univariate analysis. A-C. volcano plot of three comparison. D. venn plot of co-upregulated genes and co-downregulated genes

3.3. Pathway Enrichment Analysis

Through Gene Ontology (GO) pathway enrichment analysis, we further elucidated the biological pathways associated with the upregulated genes. Notably, the enriched pathways included cell population proliferation, tissue morphogenesis, adaptive immune response, regulation of cell activation, cytoskeleton organization in muscle cells, hemostasis, positive regulation of locomotion, and actin filament-based processes (Figure 3A). Conversely, downregulated pathways encompassed mitochondrial RNA metabolic processes, DNA metabolic processes, tRNA metabolic processes, carbohydrate derivative biosynthetic processes, chromosome organization, mitochondrial organization, cell cycle regulation, rRNA processing, ribosome biogenesis, and translation. These pathways exhibit substantial interconnections, suggesting a complex regulatory network influenced by the tumor microenvironment (Figure 3B). These pathways are interconnected and form several large clusters (Figure 3A, B).

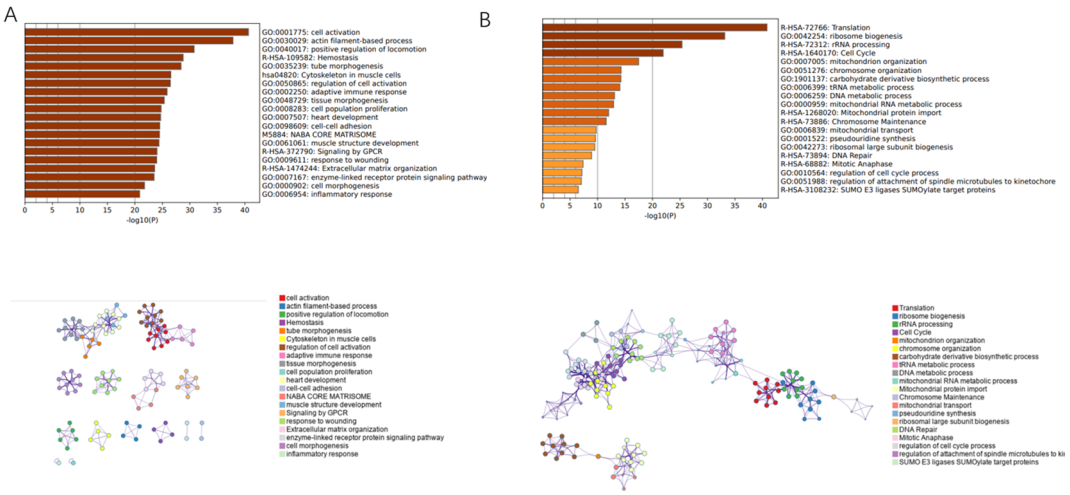


Fig 3. Results of Pathway Enrichment Analysis

3.4. Machine Learning Modeling

We selected the top five upregulated and downregulated pathways from the previously identified pathways that could cluster together, resulting in a total of 802 key differentially expressed genes for machine learning modeling. Four algorithms were employed: k-nearest neighbors (KNN), random forest, support vector machine (SVM), and XGBoost (Figure 4A-D). Notably, XGBoost demonstrated the best performance, while the other algorithms exhibited significant error rates. Further analysis using the XGBoost algorithm revealed high area under the curve (AUC) values for most classifications, except for the "advanced cap" category, likely due to its close resemblance to the "cap" category (Figure 4E). Key contributing genes in this model included TMEM98, SLC25A13, and FGR, which are important hub genes in the development of prostate cancer (Figure 4F).

3.5. The Survival Analysis of Hub Genes in TCGA Dataset

We conducted a prognostic analysis using a cohort of 498

samples from the TCGA dataset. Both univariate and multivariate analyses indicated that ITGA2B is a risk factor, while CX3CL1 serves as a protective factor, independent of other clinical variables (Figure 5A, B). The performance of the Nomogram model is validated by its proximity to the calibration curve, suggesting robust predictive capabilities. The Nomogram was constructed based on these significant differential genes. In the Nomogram, each variable is represented by a line segment marked with scales that indicate potential value ranges, with the length of each segment reflecting the variable's impact on prognosis. The "Point" in the diagram represents individual scores corresponding to the variable's different values, while "Total Point" indicates the sum of these individual scores. The "1, 2, 3-year survival Prob" illustrates the survival probabilities at 1, 2, and 3 years. By summing the scores of clinical indicators, we can predict the patient's survival rates over the next one, two, and three years. The constructed model demonstrates excellent predictive efficacy (Figure 5C, D).

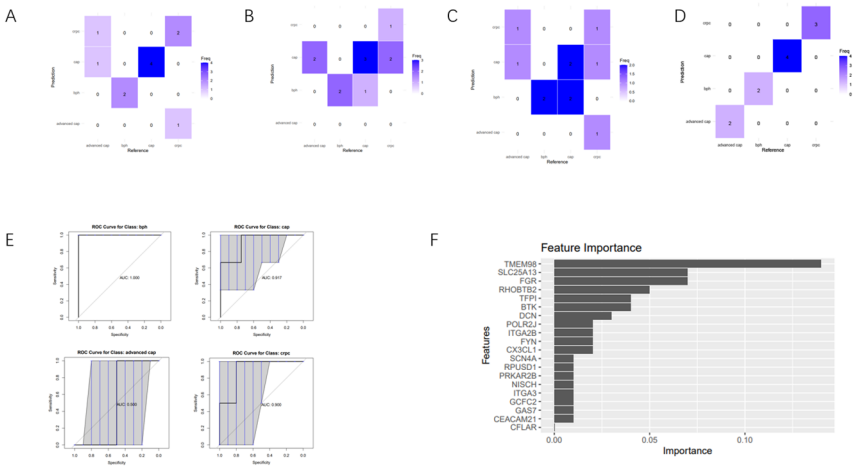


Fig 4. Results of Machine Learning Modeling

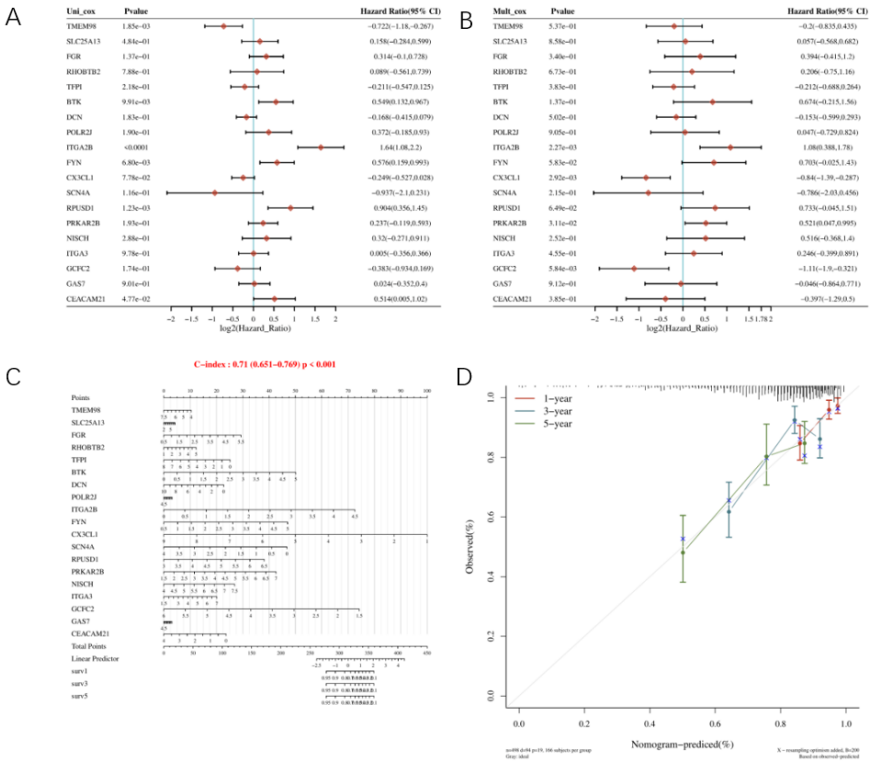


Fig 5. Univariate and multivariate Cox analyses provide the p-values, hazard ratios (HR), and confidence intervals (CI) for gene expression and clinical characteristics. The nomogram can predict the 1-year, 2-year, and 3-year overall survival of XX patients. The calibration curve of the overall survival nomogram model in the discovery cohort is shown, where the diagonal dashed line represents the ideal nomogram, and the blue, red, and orange lines represent the observed 1-year, 2-year, and 3-year nomograms, respectively

4. Discussion

The differential expression analysis performed on transcriptomic data from the GEO database revealed 802 key genes associated with prostate cancer progression. This finding aligns with the established understanding that gene expression changes are critical in cancer development and progression. Notably, our results indicated that these genes are involved in vital molecular pathways related to immune response and cell growth, suggesting that immune modulation and cell proliferation are crucial aspects of prostate cancer biology. The identification of these pathways not only enhances our comprehension of the disease but also points to potential therapeutic targets.

The application of machine learning techniques to develop a grading model for prostate cancer represents a significant innovation in clinical oncology. By integrating various algorithms—Random Forest, SVM, KNN, and XGBoost—we achieved a robust model with high AUC scores, particularly with the XGBoost algorithm, which demonstrated superior predictive accuracy. This performance underscores the potential of machine learning to assist in clinical decision-making by providing precise stratification of disease severity based on gene expression profiles. The model's ability to identify patients at varying risks can facilitate more tailored treatment approaches, thus optimizing patient management and improving outcomes.

Our survival analysis utilizing the TCGA dataset further corroborates the importance of the identified genes. The univariate and multivariate analyses indicated that ITGA2B is a significant risk factor, while CX3CL1 serves as a protective factor, both of which operate independently of other clinical variables. The development of a Nomogram based on these key genes provides a practical tool for clinicians, offering a visual representation of prognosis that can guide treatment decisions. The model's calibration curve indicates its reliability in predicting patient survival at one, two, and three years, thereby reinforcing its clinical applicability.

The intricate regulatory networks highlighted by our functional enrichment analyses illustrate the complexity of prostate cancer biology. The pathways enriched in our study not only emphasize the role of immune response and cell proliferation but also point to potential vulnerabilities within the cancer cells that could be exploited for therapeutic benefit. The downregulated pathways associated with metabolic processes suggest that advanced prostate cancer may exhibit altered metabolic profiles, a feature that could be targeted by novel therapeutic strategies.

References

[1] Jemal, A., Siegel, R., Xu, J., Ward, E., 2010. Cancer statistics, 2010.CA Cancer J. Clin. 60, 277e300.

[2] White MC, Holman DM, Boehm JE, Peipins LA, Grossman M, Henley J. Age and cancer risk: a potentially modifiable relationship. Am J Prev Med. 2014;46(3 Suppl 1):S7–15. doi: 10. 1016/j.amepre.2013.10.029.

[3] Fenton JJ, Weyrich MS, Durbin S, Liu Y, Bang H, Melnikow J. Prostate-specific antigen–based screening for prostate cancer. JAMA. 2018 May 8;319(18):1914. doi:10.1001/ jama. 2018. 3712.

[4] Tolkach Y, Kristiansen G. The heterogeneity of prostate cancer: A practical approach. Pathobiology. 2018;85(1–2):108–16. doi:10.1159/000477852.

[5] Viegas S, Ladeira C, Costa-Veiga A, Perelman J, Gajski G. Forgotten public health impacts of cancer – an overview. Archives of Industrial Hygiene and Toxicology. 2017 Dec 20;68(4):287–97. doi:10.1515/aiht-2017-68-3005.

[6] Lim J, Kurzrock R, Nishizaki D, Miyashita H, Adashek JJ, Lee S, et al. Pan-Cancer analysis of tim-3 transcriptomic expression reveals high levels in pancreatic cancer and interpatient heterogeneity. Cancer Medicine. 2023 Dec 22;13(1). doi:10. 1002/ cam4.6844.

[7] Vitelli V, Fleischer T, Ankill J, Arjas E, Frigessi A, Kristensen VN, et al. Transcriptomic Pan-Cancer Analysis using rank-based bayesian inference. Molecular Oncology. 2023 Jan 23;17(4):548–63. doi:10.1002/1878-0261.13354.

[8] Schrijver WAME, Selenica P, Lee JY, et al. Mutation profiling of key cancer genes in primary breast cancers and their distant metastases. Cancer Res. 2018;78(12):3112–21. doi:10.1158/ 0008-5472.CAN-17-2310.

[9] Barrett T, Edgar R. [19] gene expression omnibus: Microarray data storage, submission, retrieval, and analysis. Methods in Enzymology. 2006;352–69. doi:10.1016/s0076-6879(06) 110 19-8.

[10] Cornford P, van den Bergh RCN, Briers E, Van den Broeck T, Cumberbatch MG, De Santis M, et al. Eau-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer. part II—2020 update: Treatment of relapsing and metastatic prostate cancer. European Urology. 2021 Feb;79(2):263–82. doi:10.1016/j. eururo. 2020.09.046.

[11] Jung E, Moon H, Cho BI, et al. Galectin-1 expression in cancer-associated stromal cells correlates with tumor invasiveness and progression in breast cancer. Int J Cancer. 2007;120(11):2331–8. doi:10.1002/ijc.22434.

[12] Rosenstein BS, Rattay T, Kang J, eds. Machine learning and artificial intelligence in radiation oncology: a guide for clinicians. 1st ed. Stacy Masucci; 2024.