The Research of Medical Science Review
ISSN: 3007-1208 & 3007-1216 Volume 4, Issue 4, 2026

COMPUTATIONAL REDISCOVERY OF PROSTATE-SPECIFIC ANTIGEN AS A PAN-CANCER TRANSCRIPTOMIC DISCRIMINATOR: A VARIANCE-BASED ANALYTICAL FRAMEWORK

Mateen Ahmad[1], Dr. Malik Amer Atta[*2], Mudassir Khan[3], Zile Huma[4]

[1]*MS Scholar, Department of Biotechnology, COMSATS University Islamabad, Abbottabad Campus, Pakistan*
[*2]*Assistant Professor, Institute of Education & Research, Quaid-E-Azam Campus, Gomal University, D.I.Khan, Khyber Pakhtunkhwa, Pakistan*
[3]*MS Scholar, Department of Biotechnology, COMSATS University Islamabad, Abbottabad Campus, Pakistan*
[4]*MS Scholar, Department of Computing, ABASYN University, Islamabad Campus, Islamabad, Pakistan*

abmateenofficial@gmail.com[1], maatta@gu.edu.pk[*2], mudasir12122@gmail.com[3], humaaly1989@gmail.com[4]

DOI: https://doi.org/10.5281/zenodo.19843962

Keywords
KLK3 Gene Expression, Prostate-Specific Antigen, Pan-Cancer Transcriptomics, RNA Sequencing, Tissue-Of-Origin Biomarkers, Variance-Based Gene Selection, Non-Parametric Statistics, The Cancer Genome Atlas, Cancer Classification.

Article History
Received on 15 Mar 2026
Accepted on 12 Apr 2026
Published on 28 Apr 2026

Corresponding Author: *

Dr. Malik Amer Atta*
Assistant Professor, Institute of Education & Research, Quaid-E-Azam Campus, Gomal University, D.I.Khan, Khyber Pakhtunkhwa, Pakistan

Abstract

Background: *Tumor gene expression is largely determined by tissue of origin, as shown by TCGA RNA sequencing studies. The majority of computational studies are more towards classification than interpretability, and there is a missing link in the statistical validation of the genes that can contribute to inter-cancer variability. Although KLK3 is a famous tissue-specific biomarker, its formal recognition as the most variable and discriminative gene across the types of cancer has not been shown in a hypothesis-driven and assumption-tested model.*

Methods: *UCSC Xena provided the gene expression data of 801 tumor samples (PRAD, n=136; BRCA, n=300; LUAD, n=141; KIRC, n=146; COAD, n=78). The variance was computed with 20,531 genes that were cancer-free. Shapiro-Wilk and Levene tests were used to confirm the violation of normality and equal variance, hence the use of Kruskal-Wallis H as the main omnibus test and Dunn as the second test to compare directional PRAD with the Bonferroni correction and Welch t-tests. Sex-linked genes (RPS4Y1, XIST) were deleted.*

Results: *KLK3 was the most variable gene (variance = 44.76), followed by KLK2 (36.36) and SFTPB (34.50). Kruskal-Wallis confirmed highly significant differential expression ($H = 495.12$, $p = 7.62 \times 10^{-106}$, $\eta^2 = 0.617$). Welch t-tests (adjusted $p < 0.001$; Cohen's d: 9.59-22.01) confirmed that KLK3 overexpression was confirmed in PRAD relative to all comparators. Dunn post-hoc had significant results in 9 out of ten of the comparisons; the only not significant one was that of KIRC and LUAD (adjusted $p = 1.000$).*

Conclusion: *This is the first assumption-verified computational validation that KLK3 is the most varying and discriminative gene among a five-cancer panel, recovered by a completely unsupervised approach based on variance, which supports interpretable and*

data-driven frameworks of pan-cancer biomarker discovery and tissue-of-origin classification.

INTRODUCTION

Cancer is a diverse set of diseases associated with uncontrolled cell growth and tissue invasion, and it is one of the major causes of morbidity and mortality worldwide, with 18.1 million new cases and 9.6 million deaths estimated in 2018 alone (Bray et al., 2018). Cancer, in contrast to infectious diseases, is the result of internal somatic changes (genetic mutations and epigenetic dysregulation) that disrupt normal cellular homeostasis. More importantly, gene expression patterns of individual tumors are highly dependent on the type of tissue from which the tumor is formed, and each type of cancer leaves behind a unique transcriptomic print that can be identified computationally by comparing large-scale RNA sequencing samples.

The Cancer Genome Atlas (TCGA) has played a significant role in this aspect by profiling over 20,000 primary tumor samples in 33 types of cancer (Weinstein et al., 2013). The systematically unsupervised analyses always indicate that tumors cluster by tissue lineage and that those genes whose cross-cancer expression variance is the highest have the most tissue-restricted expression patterns (Hoadley et al., 2014; Hoadley et al., 2018), which is the biological explanation of why we choose genes by their cross-cancer expression variance as the dimensionality reduction strategy employed in pan-cancer studies. KLK3 - the gene of Prostate Specific Antigen (PSA) is one of the most interesting tissue-specific markers. PSA is an almost exclusively serine protease that is under the control of the androgen receptor (Diamandis & Yousef, 2002; Riegman et al., 1991), and no suggested prostate biomarker has yet matched its cross-tissue specificity. The KLK3 locus is located in a 15-gene cluster with KLK2 in chromosome 19q13, both of which are regulated by androgen receptors (Avgeris & Scorilas, 2016), and variants of KLK3 have been associated with the development and spread of prostate cancer (Meyers et al., 2024).

There has been an increasing use of machine learning in pan-cancer transcriptomic classification. Deep learning models trained on TCGA data have reached classification accuracy of over 97% with KLK3 and other prostate-associated genes always appearing at the top of the list of most discriminative features (Keup et al., 2023). Random forests, support vector machines, and gradient boosting classifiers also find tissue-specific genes in the top discriminators (Chen et al., 2021), and the variance-based feature selection method has proven to maintain the performance of classifiers but reduce dimensionality significantly (Zhao et al., 2020). It is interesting to note that sex-linked transcripts like RPS4Y1 and XIST are often found in the high-variance gene set in cohorts of mixed sex, driven by demographic composition and not cancer biology, and are important to annotate to eliminate confounding tissue-of-origin effects. In spite of such a literature review, a PubMed and Google Scholar search revealed no research that formally justified KLK3 as the most variable and discriminative gene across cancer types in an assumption-tested, hypothesis-driven computational framework.

Research Gap

Although much pan-cancer transcriptomic work has been done, the mechanisms underlying inter-cancer differences have been subordinated to classification, rather than biological interpretability. Despite KLK3 being a well-established prostate cancer biomarker, its position as a dominant statistical discriminant in a variety of malignancies has never been strictly tested in an assumption-tested computational context. This paper fills such a gap.

Objectives of the Study

The present study was conducted with the following objectives:

To determine and statistically confirm that KLK3 is the most transcriptomically variable and discriminative gene across the five distinct cancer types by Kruskal-Wallis, Dunn's, post-hoc, with Bonferroni, and Welch t-tests on 20,531 gene-profiled tumor samples (n=801).

To select genes by unsupervised variance-based gene selection and to cross-reference the top-ranked genes with known literature in molecular biology, it is necessary to ensure that computational results derived using data are interpretable biologically and based on known tissue-specific patterns of expression.

Delimitations

Only five types of cancer were evaluated in this study, using a single normalization pipeline, but with unequal group samples, and without protein-level or functional validation. Although these limit the scope of generalizability, they do not invalidate the fundamental conclusion that unsupervised selection of variance can recapitulate previously validated tissue-specific biomarkers with high statistical accuracy, on their own.

MATERIALS AND METHODS

Data Source, Preprocessing, and Quality Control

The UCSC Xena TCGA Pan-Cancer (PANCAN) data on transcriptomic profiles of 801 primary tumors (Weinstein et al., 2013) was used, including five malignancies: BRCA (n=300), KIRC (n=146), LUAD (n=141), PRAD (n=136), and COAD (n=78). Phenotypic metadata was combined with the levels of gene expression of 20,531 features using sample identifiers. The resulting matrix had no missing values, did not need imputation, and was $\log_2$-transformed RSEM counts that were pre-normalized by the UCSC Xena PANCAN pipeline and directly used in downstream analyses.

Variance-Based Gene Selection and Annotation

All 20,531 genes were computed in a completely non-supervised way, regardless of the label of the type of cancer. Ranking of genes was done based on the declining variance; the 50 top genes were visualized through a heatmap, and the top 10 were chosen to be biologically interpreted by matching identifiers to the official HGNC symbols using PANCAN RNASeqV2 reference annotation. RPS4Y1 (rank 4) and XIST (rank 7) were kept in the ranking as a transparency but not interpreted in terms of tissue-specificity, as their variance is seen to be due to cohort sex composition and not cancer biology.

STATISTICAL ANALYSIS

Software, Assumptions, and Significance Threshold

Python (3.12.7) was used to do all of the statistical computations, using Scipy, Statsmodels, and scikit-posthocs packages and a significance threshold of 0.05. The Shapiro-Wilk test (scipy.stats.shapiro) was used to test normality of Klk3 expression in each type of cancer, and the Levene test (scipy.stats.levene) was used to test homogeneity of variances. Neither assumption was met; therefore, non-parametric and heteroscedasticity robust methods were selected as the main method of analysis.

Inferential Tests and Post Hoc Comparisons

All pair-wise comparisons of cancer-types were done with Dunn test with Bonferroni correction (scikit_ posthocs.posthoc_dunn, p adjusted = 'bonferroni'). Since Kruskal–Wallis shows the significance of the post-hoc procedure is non-parametric, the test to use is Dunn, since it works in the same rank-based model without having the conditions of normality and equal variances. Directional pairwise comparisons between PRAD and each comparator cancer type were also performed with Welch's t-test (scipy.stats.ttest_ind, equal_var = False), with Bonferroni error correction performed on the four comparisons.

Effect Size Estimation and Visualization

The Kruskal-Wallis test was used to compute Eta squared (η^2) = (Hk + 1) / (n k) and Cohen's d = (group means difference)/pooled standard deviation to find pairwise comparisons. Matplotlib and Seaborn were used to generate a visualization that included a sample distribution bar chart, a heatmap of the top 50 most variable genes, a mean expression bar chart of the top 10 variable genes across all cancer types, and a KLK3 expression boxplot.

Data and Code Availability

All gene expression data are publicly accessible through the UCSC Xena platform (https://xenabrowser.net) under the TCGA PANCAN IlluminaHiSeq RNASeqV2 dataset. The complete Python analysis pipeline is open source

and available at https://github.com/abmateen612/gene-expression-cancer-analysis, comprising a Jupyter Notebook with all procedures from data loading through statistical testing and visualization.

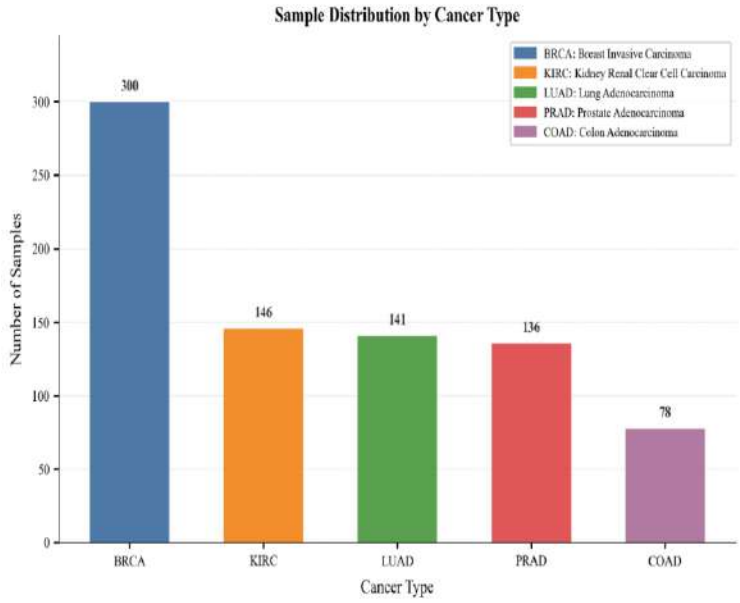
Ethics Statement

This analysis made use of publicly available de-identified TCGA data through the UCSC Xena platform. The secondary analysis of such data did not need any extra ethical approval or patient consent.

RESULTS

Gene Ranking and Dataset Overview

The dataset comprised 801 tumor samples across five cancer types with 20,531 genes and no missing values. Mean $\log_2$ RSEM expression was comparable across all types, indicating no global expression bias.



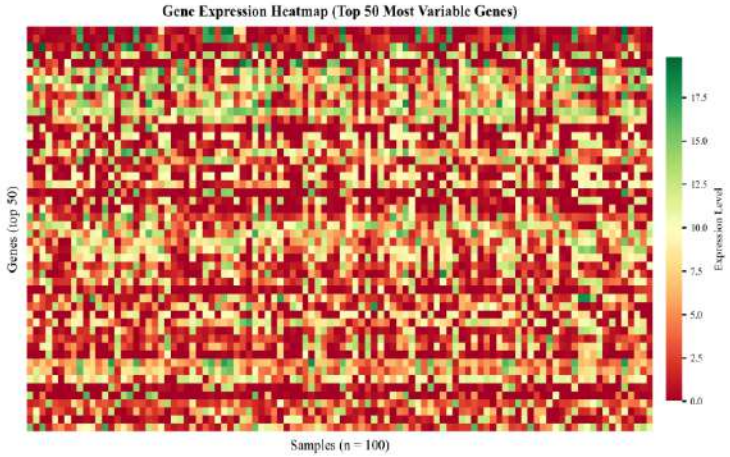
Figure#1: Tumor sample distribution across five cancer types (n = 801).

Unsupervised variance ranking identified KLK3 as the most variable gene (variance = 44.76), followed by KLK2 (36.36) and SFTPB (34.50). All top-ranked genes corresponded to known tissue-specific expression programs, confirming the biological fidelity of the approach. RPS4Y1 and XIST were excluded from tissue specificity interpretation as their variance reflects sex composition rather than cancer biology.

Table#1: The top 10 most variable genes in transcriptomically variable genes were identified by variance-based ranking across 801 tumor samples.

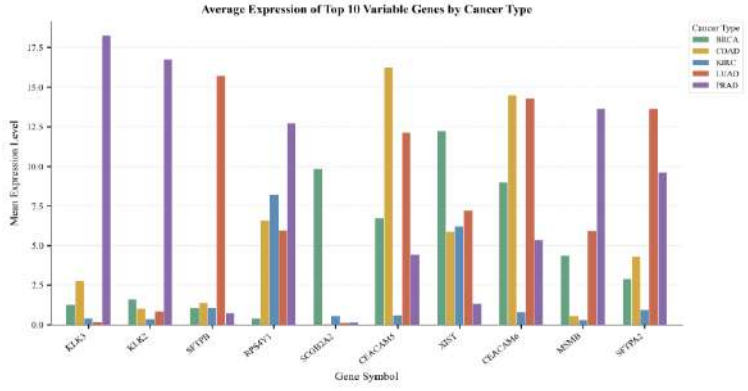
Rank	Gene Identifier	Gene Symbol	Biological Significance	Variance
1	gene_9176	KLK3	Prostate-specific serine protease (PSA)	44.76
2	gene_9175	KLK2	Prostate-specific kallikrein-related peptidase	36.36
3	gene_15898	SFTPB	Pulmonary surfactant protein B (lung-specific)	34.50
4	gene_15301	RPS4Y1	Y-chromosome ribosomal protein (male-specific)	33.46

5	gene_15589	SCGB2A2	Mammoglobin (breast-specific secretory protein)	31.33
6	gene_3540	CEACAM5	Carcinoembryonic antigen (colon-associated)	30.59
7	gene_19661	XIST	X-chromosome inactivation transcript (sex-specific)	30.08
8	gene_3541	CEACAM6	Carcinoembryonic antigen-related cell adhesion molecule	28.72
9	gene_11250	MSMB	Prostate-specific microseminoprotein-beta	26.52
10	gene_15897	SFTPA2	Pulmonary surfactant protein A2 (lung-specific)	26.02



Figure#2: Heatmap of the top 50 most transcriptomically variable genes (n = 100 samples).

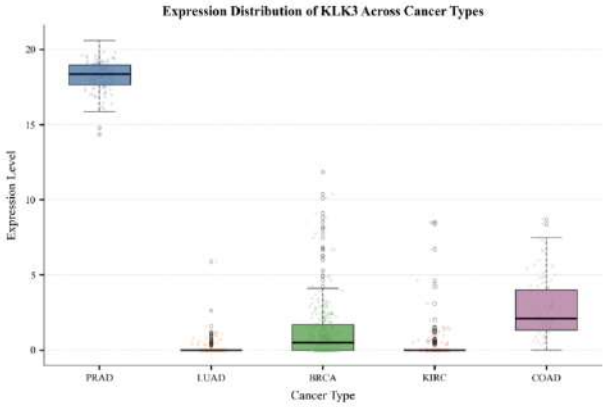
Mean KLK3 expression in PRAD (18.24) was approximately 97 times higher than in LUAD (0.19), consistent across virtually all PRAD samples.



Figure#3: Mean expression of the top 10 variable genes across five cancer types.

Assumption Testing and Omnibus Analysis

Shapiro-Wilk confirmed non-normality across all five cancer types, and Levene's test confirmed unequal variances ($W = 23.64$, $p < 0.001$), justifying the Kruskal-Wallis H test as the primary omnibus test. Results confirmed highly significant differential KLK3 expression ($H = 495.12$, $p = 7.62 \times 10^{-106}$, $\eta^2 = 0.617$), indicating that 61.7% of expression rank variance is attributable to cancer type. One-way ANOVA ($F(4, 796) = 3463.55$, $p < 0.001$) is retained as supplementary analysis.



Figure#4: ***KLK3 expression distribution across five cancer types; PRAD shows markedly elevated expression.***

Pairwise Comparisons and Effect Sizes

Welch t tests with Bonferroni correction confirmed significant KLK3 overexpression in PRAD against all comparators, with Cohen's d values of 9.59 to 22.01 indicating differences of exceptional magnitude.

Table#2: Results of a pairwise Welch t-test of KLK3 expression in PRAD vs all other forms of cancer and Bonferroni correction.

Comparison	t-statistic	p-value	Bonferroni-corrected p-value	Cohen's d	Interpretation
PRAD vs LUAD	182.42	1.37×10^{-244}	< 0.001	10.7561	Exceptional
PRAD vs KIRC	131.19	7.14×10^{-247}	< 0.001	15.5624	Exceptional
PRAD vs BRCA	118.49	≈ 0.00	< 0.001	22.0133	Exceptional
PRAD vs COAD	62.44	1.33×10^{-80}	< 0.001	9.5898	Exceptional

Dunn's post hoc test identified significant differences in nine of ten pairwise comparisons. The sole non-significant result, KIRC vs LUAD (adjusted $p = 1.000$), is analytically coherent as both are near-zero KLK3 expressors.

Table#3: Dunn's post-hoc test results with Bonferroni correction for KLK3 expression across all pairwise cancer type combinations ($\alpha = 0.05$).

Group 1	Group 2	Adjusted p-value	Significant
BRCA	COAD	2.54×10^{-8}	Yes
BRCA	KIRC	1.13×10^{-6}	Yes
BRCA	LUAD	2.02×10^{-8}	Yes
BRCA	PRAD	7.32×10^{-57}	Yes
COAD	KIRC	3.05×10^{-19}	Yes

COAD	LUAD	2.85×10^{-21}	Yes
COAD	PRAD	2.33×10^{-9}	Yes
KIRC	LUAD	1.000	No
KIRC	PRAD	1.29×10^{-74}	Yes
LUAD	PRAD	1.50×10^{-78}	Yes

DISCUSSION

Principal Finding, Statistical Robustness, and Novelty

With 801 pan-cancer tumor samples, KLK3 was selected by a variance-based gene selection methodology out of 20,531 candidate genes that had no biological input. At this point, though, the tissue-specificity of KLK3 is already established (Riegman et al., 1991; Diamandis and Yousef, 2002); the main finding of this research is formally validating it as the top-ranked discriminating gene in an assumption-tested multi-cancer model. The near-exclusive overexpression of KLK3 in PRAD by the Kruskal-Wallis framework (η^2=0.617) was justified by violated assumptions of normality and homogeneity, and the 97-fold difference in the mean expression (Cohen's d values of 9.59-22.01) is an indication of extraordinary biological significance. The only non-significant pairwise result, KIRC vs. LUAD, is analytically coherent as both cancers are near-zero KLK3 expressors.

Biological Validation and Discriminability of the Variance-Based Framework

The η^2 of 0.617 confirms that the variance of KLK3 expression is not due to within-group heterogeneity but because of cancer type and thus, the KLK3 is discriminative. KLK2, SFTPB, SCGB2A2, CEACAM5, and SFTPA2 were independently ranked as tissue-specific markers to confirm the biological fidelity of the pipeline to four other cancer lineages. Proper sex-linked transcripts (RPS4Y1, XIST) identification and exclusion are additional indications of the framework's resistance to demographic confounding, in line with the finding that pan-cancer transcriptomes are mainly structured around tissue-of-origin signals (Hoadley et al., 2014, 2018).

Translational Implications and Future Directions

The unsupervised variance-based feature selection can obtain a set of validated tissue-specific biomarkers on publicly available data, providing an interpretable alternative to black-box classification models (Keup et al., 2023). It can be directly applied to cancer of unknown primary origin (Moran et al., 2016; Conway et al., 2019), and the future application involving the survival analysis, sex-stratified variance, and external cohort validation would enhance its translational application.

CONCLUSIONS

The following conclusions were drawn.

The first objective was met with unsupervised variance-based selection, which computationally rediscovered KLK3 as the most transcriptomically variable and statistically discriminative gene across five cancer types. The Kruskal-Wallis, Dunn post-hoc with Bonferroni correction, and Welch t-tests all agreed that KLK3 expression was extraordinarily and significantly increased in prostate adenocarcinoma compared to any other type of comparator cancer. *(Aligned with Objective#1)*

In response to the second objective, the highest-ranking co-genes (KLK2, SFTPB, SCGB2A2, CEACAM5, and SFTPA2) were well-defined tissue programs of expression, providing support to the biological interpretability of the variance-based framework with respect to existing literature in molecular biology. The principled exclusion of sex-linked transcripts RPS4Y1 and XIST further validated that the pipeline at least captures actual cancer biology as opposed to demographic confounding, and that variance selection provides a reproducible basis on which to discover pan-cancer biomarkers and classify tissues-of-origin. *(Aligned with Objective#2)*

REFERENCES

Avgeris M, Scorilas A (2016). Kallikrein-related peptidases (KLKs) as emerging therapeutic targets: focus on prostate cancer and skin pathologies. Expert Opinion on Therapeutic Targets 20(7): 801–818. https://doi.org/10.1517/14728222.2016.1147560

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide. CA: A Cancer Journal for Clinicians 68(6): 394–424. https://doi.org/10.3322/caac.21492

Chen S, Zhou W, Tu J, Li J, Wang B, Mo X, Tian G, Lv K, Huang Z (2021). A novel XGBoost method to infer the primary lesion of 20 solid tumor types from gene expression data. Frontiers in Genetics 12: 632761. https://doi.org/10.3389/fgene.2021.632761

Cohen J (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Lawrence Erlbaum Associates, Hillsdale, NJ, USA.

Conway AM, Pearce SP, Clipson A, Hill SM, Eranki A, Verjee S, Blackhall F (2019). Molecular characterisation and liquid biomarkers in carcinoma of unknown primary. British Journal of Cancer 120(2): 141–153. https://doi.org/10.1038/s41416-018-0376-1

Diamandis EP, Yousef GM (2002). Human tissue kallikreins: a family of new cancer biomarkers. Clinical Chemistry 48(8): 1198–1205. https://doi.org/10.1093/clinchem/48.8.1198

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Benz CC (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell 173(2): 291–304. https://doi.org/10.1016/j.cell.2018.03.022

Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, Stuart JM (2014). Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158(4): 929–944. https://doi.org/10.1016/j.cell.2014.06.049

Keup C, Kolberg HC, Kimmig R, Fehm T (2023). Machine learning for pan-cancer classification based on RNA sequencing data. Frontiers in Molecular Biosciences 10: 1285795. https://doi.org/10.3389/fmolb.2023.1285795

Meyers TJ, Lynch JA, Gao A, Agiri F, Pridgen K, Seibert T, Hauger R (2024). Effect of variants in the KLK3 gene on prostate cancer survival and time to metastases. Journal of Clinical Oncology 42(4 suppl): 219. https://doi.org/10.1200/JCO.2024.42.4_suppl.219

Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, Esteller M (2016). Epigenetic profiling to classify cancer of unknown primary. The Lancet Oncology 17(10): 1386–1395. https://doi.org/10.1016/S1470-2045(16)30297-2

Riegman PH, Vlietstra RJ, van der Korput JA, Brinkmann AO, Trapman J (1991). The promoter of the prostate-specific antigen gene contains a functional androgen-responsive element. Molecular Endocrinology 5(12): 1921–1930. https://doi.org/10.1210/mend-5-12-1921

Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Stuart JM (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics 45(10): 1113–1120. https://doi.org/10.1038/ng.2764

Zhao Z, Li S, Du L, Wang C, Zhang K (2020). Hierarchical integration of multi-omics data for predicting cancer tissue-of-origin. Briefings in Bioinformatics 21(6): 2213–2224. https://doi.org/10.1093/bib/bbz137