Short Communications

https://doi.org/10.7124/bc.000B40
UDC577.218+616.65

G.V. Gerashchenko, V.I. Kashuba
Institute of Molecular Biology and Genetics, NAS of Ukraine
150, Akademika Zabolotnoho Str., Kyiv, Ukraine, 03143
g.v.gerashchenko@edu.imbg.org.ua

VALIDATION OF PROSTATE CANCER MOLECULAR SUBTYPING APPROACH, BASED ON THE CLUSTER ANALYSIS OF CANCER CELL AND TUMOR MICROENVIRONMENT GENE EXPRESSION PATTERNS

Aim. *To verify the previously studied gene sets associated with cancer, tumor microenvironment, and lipid metabolism to identify molecular subtypes of prostate cancer by analyzing the prostate cancer gene expression data from the TCGA database.* ***Methods.*** *Analysis of TCGA RNA-sequencing based gene expression data of 490 prostate cancer samples of 55 previously studied genes. Statistical and K-means clustering methods were used for molecular subtyping of prostate cancer samples.* ***Results.*** *Cluster analysis revealed two and three potentially significant clusters of prostate cancer samples based on the expression levels of three gene groups which is 27 cancer-associated genes, 23 tumor microenvironment related genes and 5 lipid metabolism genes. Among three clusters, the first one has the most aggressive prostate cancer samples and has elevated levels of mesenchymal markers and high levels of inflammation markers and tumor microenvironment elements. The second and third clusters of prostate cancer samples showed signs of presumably luminal and basal subtypes with lower levels of inflammation markers. The highest level of correlation in the distribution of samples across clusters was found for cancer-associated and tumor microenvironment-related gene groups.* ***Conclusions.*** *The results revealed correlations and a high degree of dispersion in the expression of the studied genes, which made it possible to identify several molecular clusters. A more detailed statistical analysis is needed to determine clinicaly relevant molecular subtypes and to establish the most significant expression markers in biological modules of the studied genes for the diagnosis, prognosis, and effective treatment of prostate cancer.*

Keywords: *prostate cancer, gene expression patterns, TCGA, cluster analysis, molecular subtypes, prostate cancer-associated genes, tumor microenvironment-related genes, lipid metabolism genes.*

Citation: Gerashchenko G.V., Kashuba V.I. (2026) Validation of prostate cancer molecular subtyping approach, based on the cluster analysis of cancer cell and tumor microenvironment gene expression patterns. *Biopolymers & Cell*, 2(42), 150—156. https://doi.org/10.7124/bc.000B40

Introduction

The assessment of the expression pattern of tumor-associated genes plays an important role in the treatment of patients with prostate cancer, in particular, in the diagnosis of various tumors, predicting the course of the disease, assessing tumor aggressiveness, predicting patient survival, the risk of disease recurrence, predicting resistance to therapy, *etc.* [1, 2].These fundamental and clinical tasks are relevant to prostate cancer, which is one of the most common types of cancer among men worldwide [3].

Currently, a significant number of studies were aimed at prostate tumors molecular characterizing and subtyping at the genetic, epigenetic, and transcriptomic levels using modern technologies such as microchips and next-generation sequencing [4, 5]. This revealed several different molecular classifications of prostate cancer based on the genomic and transcriptomic characteristics of prostate cancer cells [6, 7].

The current findings indicate that carcinogenesis is influenced by the molecular properties of tumor cells as well as the characteristics of normal epithelial, stromal, and immune cells of the host. These characteristics are reflected in the specific composition of the tumor microenvironment and the state of the immune system [8, 9, 10]. Our hypothesis for analyzing gene expression and the tumor transcriptome is to identify genes that are specifically expressed in different types of tumor cells (marker genes) and cancer metabolic pathways.

We previously conducted a relative expression study of more than 60 genes/transcripts using qPCR. The gene groups included genes associated with epithelial-mesenchymal transition, genes associated with prostate cancer, markers of fibroblasts and tumor-associated fibroblasts, markers of lymphocytes and inflammation, markers of macrophages and tumor-associated macrophages, and lipid metabolism genes for molecular profiling of prostate cancer samples and the identification of some molecular clusters [11—15]. The objective of this study is to validate the obtained data on a larger group of prostate cancer samples.

Materials and Methods

RNA-Seq data analysis

Normalized gene expression levels (NGE) of 490 prostate cancer samples and 52 paired conventionally normal tissues were brought from TCGA data for which RNA-Seq data were available for representative Prostate adenocarcinomas (TCGA-PRAD) sample set. https://portal.gdc.cancer.gov/ and https://www.cell.com/cell/fulltext/S0092-8674 (15)01339-2. The sample was selected randomly to reflect elements of the general population. The data were processed with a modified version of CrossHub (https://sourceforge.net/projects/crosshub/), a tool for the multiway analysis of TCGA transcriptomic and genomic data. Read counts data were downloaded from the TCGA data portal (https://portal.gdc.cancer.gov/) and normalized using the TMM method and then recalculated for 1 million library size.

Statistical analysis

The Kolmogorov-Smirnov test was used to analyze the normality of distribution. Descriptive statistics methods were used to calculate group statistical parameters. To identify correlations between the expression of the studied genes, the Spearman rank correlation test was used. Numiqo statistics calculator were used to generate the Elbow plot for initial identification of cluster numbers (https://numiqo.com/statistics-calculator/cluster). The K-Mean clustering was applied for prostate cancer subtyping and statistical analysis were performed by STASISTICA 10 as describe earlier [13, 14]. The Benjamini-Hochberg procedure with false discovery rate (FDR) 0.10—0.25 was used when multiple comparisons were performed [16]. A difference with $p < 0.05$ was considered significant.

Results and Discussion

Our previous studies of the relative expression of more than 60 gene transcripts in tumors, conditionally normal tissues, and prostate adenomas were performed using quantitative PCR [11—15]. In this

Fig. 1. Expression patterns of cancer-associated genes in three clusters of prostate cancer samples: *1 — AR*; *2 — PSA*; *3 — PCA3*; *4 — NKX3-1*; *5 — KRT18*; *6 — CDH1*; *7 — PTEN*; *8 — GCR*; *9 — ESR1*; *10 — ESR2*; *11 — VIM*; *12 — FN1*; *13 — OCLN*; *14 —PRLR*; *15 — XIAP*; *16 — MKI67*; *17 — MMP9*; *18 — MMP2*; *19 — PRL*; *20 — HOTAIR*; *21 — IGF1R*; *22 — INSR*; *23 — CASP3*; *24 — VDR*; *25 — TMPRSS2*; *26 — ERG*; *27 — CDH2*

study, normalized gene expressions of 55 genes were utilized, as some transcripts were absent in TCGA-PRAD database, including some non-coding RNAs and the *TMPRSS2-ERG* fusion transcript. Instead, the normalized expression le vels of two genes, *TMPRSS2* and *ERG*, were analyzed. The studied genes were divided into three big groups: tumor-associated genes, tumor microenvironment-related genes, and lipid metabolism genes.

As demonstrated by the analysis of descriptive statistics of 490 prostate cancer samples, most genes exhibited high level of dispersion in prostate tumors. These findings are consistent with those from our previous studies. An assessment of Spearman's correlation coefficients between normalized gene expression levels in TCGA-PRAD samples, adjusted for multiple comparisons using FDR, revealed a number of significant correlations between the expressions of the studied genes with $p < 0.001$ and $q < 0.001$. Specifically, the strongest positive correlations were identified between the expressions of the *VIM-MMP2* ($r^s = 0.811$), *MMP2-S100A4* ($r^s = 0.681$), *CD68-CD163* ($r^s = 0.780$), *CD163-IL2RA* ($r^s = 0.783$), *CCR4-IL2RA* ($r^s = 0.694$), and *CCL22-CTLA4* ($r^s = 0.688$). This supports the feasibility of further analysis at the level of individual genes and as a set of functional modules reflecting the tumor-epithelial, stromal-immune, and metabolic components of TCGA-PRAD samples.

The analysis conducted to select the optimal number of clusters using the Elbow plot demonstrated a distortion point at two clusters and minor changes as the number increased for all three gene groups. We performed K-Means clustering of 490 prostate cancer samples for two clusters and found no significant difference in expression for one-third of the genes in each group. The subsequent decision was made to implement clustering for three distinct clusters, aligning with the approach employed in prior studies.

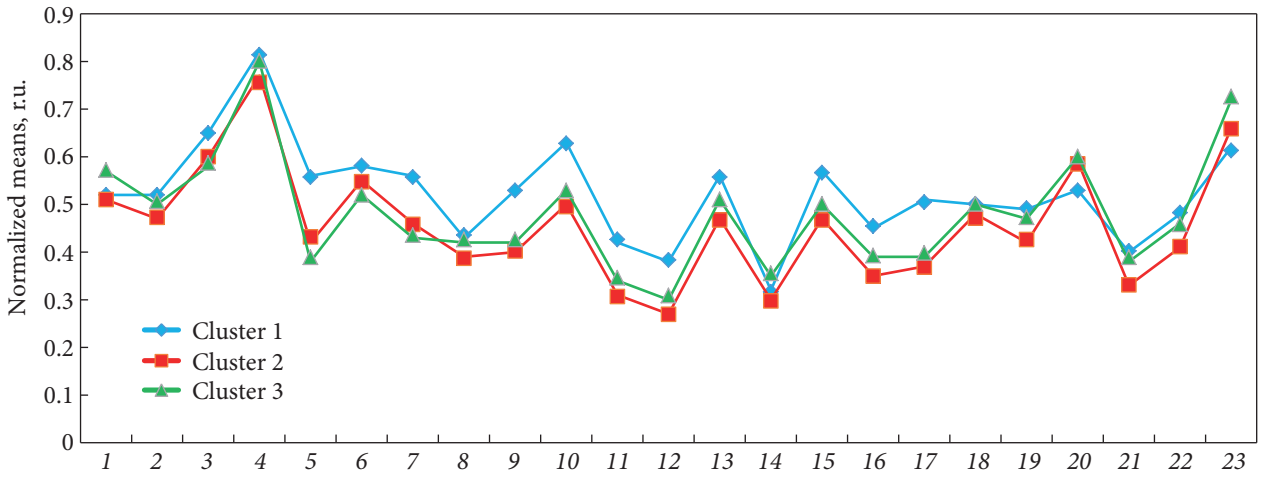

Fig. 2. Expression patterns of tumor microenvironment genes in three clusters of prostate cancer samples: *1 — ACTA2*; *2 — CXCL12*; *3 — CXCL14*; *4 — CTGF*; *5 — FAP*; *6 — HIF1A*; *7 — THY1*; *8 — S100A4*; *9 — CD68*; *10 — CD163*; *11 — CCR4*; *12 — CCL17*; *13 — CCL22*; *14 —NOS2*; *15 — CIAS1*; *16 — CTLA4*; *17 — IL2RA*; *18 — HLA-G*; *19 — IRF1*; *20 — IL1R1*; *21 — IL1RL1*; *22 — KLRK1*; *23 — MSMB*

K-mean clustering was performed for maximal initial distances measured by Euclidean distances. Two categorical variables from clinical pathological characteristics of prostate cancer samples — stages and Gleason scores were used for clustering analysis. As a result of the clustering analysis of the normalized expression levels of 27 tumor-associated genes in prostate cancer samples, three distinct clusters with the highest cluster cost and maximal initial distances were identified (Fig. 1). The first gene cluster included the samples with predominantly Gleason scores of 8—9, while the second and third clusters included the samples with predominantly Gleason scores of 6—7. It is important to note that in this group, genes exhibit significantly different levels of expression. The analysis revealed the presence of both highly expressed genes, including *CDH1, PTEN, PSA, PCA3, TMPRSS2*, and genes with low expression, such as *PRL* and *HOTAIR*.

The analysis included both ANOVA for continuous variables and Kruskal-Wallis and Dunn-Bonferroni post hoc tests. These analyses revealed significant differences between at least two of the three clusters ($p < 0.05$), with the exception of only the *PRL* and *MMP9* genes.

The cluster analysis of a group of 23 tumor microenvironment-related genes of 490 prostate cancer samples also revealed three clusters of prostate cancer samples (Fig. 2). The distribution of samples by clusters demonstrated a similar outcome to that observed in the group of tumor-associated genes: cluster 1 included samples with a Gleason score predominantly of 8—9, while clusters 2—3 included samples with a Gleason score of 6—7. ANOVA analysis for continuous variables and Kruskal-Wallis and Dunn-Bonferroni post hoc tests revealed significant differences ($p < 0.05$) between at least two clusters for all genes except *HLA-G*. Spearman's test for correlations between gene expression levels in this group in prostate cancer samples also revealed a large number of significant positive and negative correlations with $p < 0.01$.

Cluster analysis of a group of five lipid metabolism genes also revealed three clusters of prostate cancer samples (Fig. 3). The first cluster included samples with predominantly Gleason scores of 8—9, while the second and third clusters included samples with predominantly Gleason scores of 6—7. ANOVA analysis for continuous variables and Kruskal-Wallis and Dunn-Bonferroni post hoc tests revealed significant differences ($p < 0.05$)

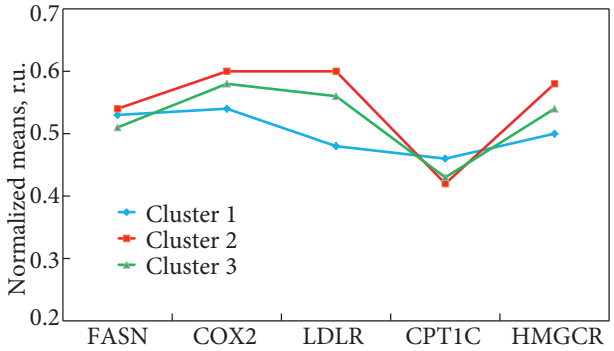

Fig. 3. Expression patterns of lipid metabolism genes in three molecular profiles of prostate cancer samples

Table 1. **Spearman rank correlations coefficients (r^s) for three sample Clusters of three gene groups**

Groups	PrCaAssoc	TM
PrCaAssoc	1.00000	
TM	**0.76737**	1.00000
LipidMet	*0.46279*	*0.41887*

Note: PrCaAssoc — cancer-associated genes, TM — tumor microenvironment genes, LipidMet — genes of lipid metabolism; $p < 0.001$ — *italics*; $p < 0.0001$ — **bold**

between at least two clusters for four of the five genes, with the exception of *FASN*.

Spearman's correlation analysis between three groups of genes based on the distribution of tumor samples across three identified clusters (Table 1) revealed moderate positive significant correlations ($r^s = 0.46279$ and $r^s = 0.41887$, $p < 0.001$) for the lipid metabolism and cancer-associated and tumor microenvironment group genes, respectively. Meanwhile, for the cancer-associated and tumor microenvironment gene groups, a high positive correlation of sample distribution ($r^s = 0.76737$, $p < 0.0001$) was found between the three identified molecular clusters of gene expression indicators.

While these groups exhibit a high degree of correlation with one another, some prostate cancer samples exhibit different molecular characteristics across various gene groups. This could be critical for diagnosis and treatment.

The results of the clustering analysis presented in this study, which was conducted on a more representative sample of prostate cancer cases, also enabled the identification of three tumor clusters for each gene group. The prostate cancer samples from cluster 1 exhibited higher Gleason scores, indicating a higher degree of aggression. This is confirmed by lower expression value for several epithelial cell markers, particularly *PSA, PCA3, NKX3-1*. Conversely, these samples showed high expression levels of mesenchymal markers, including *VIM, MMP9, MMP2* and *CDH2*. In addition, the prostate cancer samples of cluster 1 are characterized by high expression levels of the tumor microenvironment markers characteristic of cancer-associated fibroblasts, tumor-associated macrophages, and pro-inflammatory markers of immune cells, in particular *FAP, THY1, HIF1A, CTLA4, CIAS1, CCR4* and others. The low levels of expression of the lipid metabolism genes *COX2, LDLR* and *HMGCR* in prostate cancer samples from cluster 1 may indicate reduced sensitivity of tumors to inhibitors of these proteins [17].

The second and third clusters of prostate cancer samples, which generally have lower Gleason scores, showed a more pronounced difference in the expression of tumor-associated genes, in particular *AR, KRT18, GCR, OCLN, PRLR, XIAP, IGF1R* and *ERG*. Meanwhile, the expression patterns of tumor microenvironment genes have similar profiles to those of lipid metabolism genes.

The variation in the expression of epithelial and stromal cell marker genes across different clusters, notably clusters 2 and 3, could suggest the presence of distinct cancer subtypes, such as luminal and basal, which may require different treatment approaches. The research has indicated that distinguishing between Luminal and Basal subtypes of prostate cancer can impact a patient's prognosis and response to therapy [18]. With regard to the initial group of samples exhibiting the most aggressive tumors, it is challenging to discuss these subtypes (basal or luminal) due to the elevated levels of expression observed in tumor microenvironment genes and pro-inflammatory markers [19]. Although this cluster is closer

to the basal subtype in terms of tumor-associated gene expression characteristics, further analysis is necessary to determine the implications of this finding. It is important to note that the K-means clustering method, like any other analytical method, has certain limitations that prevent the identification of both the most relevant and significant markers among the analyzed genes and the insignificant markers that may introduce so-called noise into the analysis. Therefore, a more in-depth analysis of the results is needed to identify the strongest and most significant markers for molecular tumor subtyping.

Conclusion

Analysis of normalized expression data of 490 prostate cancer samples from the TCGA database of the set of cancer-associated, tumor microenvironment-related, and lipid metabolism genes we previously studied by qPCR in small sample number of Ukrainian patients showed significant levels of expression dispersion for many genes. The cluster analysis of these gene groups revealed two and three clusters, which are potential molecular subtypes of prostate cancer, with high levels of correlation between these gene groups. A more detailed bioinformatic, statistical, and molecular biological analysis of the results is necessary to characterize the identified molecular subtypes and their potential clinical significance. It is also necessary to establish the most significant features and biological gene modules of the expression of the studied genes for the diagnosis, prognosis, and effective treatment of prostate cancer.

Acknowledgments. This research was conducted within the budgetary research project of the Department of Molecular Oncogenetics of IMBG. This study was supported by an IMBG Simons Foundation Grant for Ukrainian institutions № SFI-PD-Ukraine-00017453 [G.G, V.K.].

Conflict of Interest. The authors declare that they have no conflicts of interest with the contents of this article.

REFERENCES

1. *Jiménez N, Reig Ò, Marín-Aguilera M, et al., and Mellado B.* Transcriptional Profile Associated with Clinical Outcomes in Metastatic Hormone-Sensitive Prostate Cancer Treated with Androgen Deprivation and Docetaxel. *Cancers (Basel).* 2022; **14**(19):4757.
2. *Hu D, Jiang L, Luo S, et al., and Tang W.* Development of an autophagy-related gene expression signature for prognosis prediction in prostate cancer patients. *J Transl Med.* 2020; **18**(1):160.
3. *Bray F, Laversanne M, Sung H, et al., and Jemal A.* Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024; **74**(3):229—63.
4. *Ren S, Wei GH, Liu D, et al., and Sun Y.* Whole-genome and Transcriptome Sequencing of Prostate Cancer Identify New Genetic Alterations Driving Disease Progression. *Eur Urol.* 2018; **73**(3):322—39.
5. *Sattari M, Rauhala H, Latonen L, et al., and Visakorpi T.* Identification of protein-coding genes associated with metastatic prostate cancer. *Endocr Relat Cancer.* 2025; **32**(7):e250070.
6. *Aggarwal R, Rydzewski NR, Zhang L, et al., and Zhao SG.* Prognosis Associated With Luminal and Basal Subtypes of Metastatic Prostate Cancer. *JAMA Oncol.* 2021; **7**(11):1644—52.
7. *Cancer Genome Atlas Research Network.* The Molecular Taxonomy of Primary Prostate Cancer. *Cell.* 2015; **163**(4):1011—25.
8. *Hanahan D, Weinberg RA.* The hallmarks of cancer. *Cell.* 2000; **100**(1):57—70.
9. *Hanahan D, Weinberg RA.* Hallmarks of cancer: the next generation. *Cell.* 2011; **144**(5):646—74.
10. *Pickup MW, Mouw JK, Weaver VM.* The extracellular matrix modulates the hallmarks of cancer. *EMBO Rep.* 2014; **15**(12):1243—53.
11. *Gerashchenko GV, Mankovska OS, Dmitriev AA, et al., and Kashuba VI.* Expression of epithelial-mesenchymal transition-related genes in prostate tumours. *Biopolym Cell.* 2017; **33**(5):335—55.
12. *Gerashchenko GV, Mevs LV, Chashchina LI, et al., and Kashuba VI.* Expression of steroid and peptide hormone receptors, metabolic enzymes and EMT-related genes in prostate tumors in relation to the presence of the TMPRSS2/ERG fusion. *Exp Oncol.* 2018; **40**(2):101—8.

13. *Gerashchenko GV, Grygoruk OV, Kononenko OA, et al., and Kashuba VI.* Expression pattern of genes associated with tumor microenvironment in prostate cancer. *Exp Oncol.* 2018; **40**(4):315—22.
14. *Gerashchenko GV, Kononenko OA, Bondarenko YuM, et al., and Kashuba VI.* Expression patterns of genes that regulate lipid metabolism in prostate tumors. *Biopolym Cell.* 2018; **34**(6):445—60.
15. *Gerashchenko GV, Kononenko OA, Bondarenko YuM, et al., and Kashuba VI.* Expression patterns of the various PDCD1 and PDL1 isoforms in prostate tumors. *Biopolym Cell.* 2022; **38**(3):169—85.
16. *Benjamini Y, Hochberg Y.* Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol.* 1995; **57**:289—300.
17. *Gerashchenko GV, Chashchina LI, Rynditch AV, Kashuba VI.* The gene expression pattern as a tool for assessment of components of microenvironment and response to anti-cancer therapy of prostate tumors. *Dopov Akad Nauk Ukr.* 2019; **4**:86—93.
18. *Zhao SG, Chang SL, Erho N, et al., and Feng FY.* Associations of Luminal and Basal Subtyping of Prostate Cancer With Prognosis and Response to Androgen Deprivation Therapy. *JAMA Oncol.* 2017; **3**(12):1663—72.
19. *Becht E, de Reyniès A, Giraldo NA, et al., and Fridman WH.* Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clin Cancer Res.* 2016; **22**(16):4057—66.

Received: 28.05.2026
Accepted: 16.06.2026
Published: 25.06.2026

Г.В. Геращенко, В.І. Кашуба

Інститут молекулярної біології і генетики НАН України
150, вул. Академіка Заболотного, Київ, Україна, 03143
g.v.gerashchenko@edu.imbg.org.ua

ВАЛІДАЦІЯ ПІДХОДУ МОЛЕКУЛЯРНОГО СУБТИПУВАННЯ РАКУ ПЕРЕДМІХУРОВОЇ ЗАЛОЗИ НА ОСНОВІ КЛАСТЕРНОГО АНАЛІЗУ ПАТЕРНІВ ЕКСПРЕСІЇ ГЕНІВ РАКОВИХ КЛІТИН ТА ПУХЛИННОГО МІКРООТОЧЕННЯ

Мета. Перевірити раніше вивчені набори генів, пов›язаних з раком, мікрооточенням пухлини та метаболізмом ліпідів, для можливості ідентифікації молекулярних підтипів раку передміхурової залози шляхом аналізу даних експресії генів раку передміхурової залози, отриманих з бази TCGA. ***Методи.*** Аналіз даних експресії генів 490 зразків раку передміхурової залози на основі РНК-секвенування з бази TCGA для 55 раніше досліджених генів. Для молекулярного субтипування зразків раку передміхурової залози було використано статистичні та кластеризаційні методи. ***Результати.*** За допомогою кластерного аналізу виявлено 2 та 3 потенційно значущі кластери зразків раку передміхурової залози як за рівнями відносної експресії груп рак-асоційованих (27 генів), маркерів пухлинного мікрооточення (23 гени) та генів ліпідного метаболізму (5 генів). Серед трьох кластерів, перший містить зразки з найбільш агресивними пухлинами, має підвищені рівні експресії маркерів мезенхімальних клітин та високий рівень маркерів запалення та елементів пухлинного мікрооточення. Другий та третій кластери зразків пухлин мають ознаки вірогідно люмінального та базального підтипів з нижчими рівнями маркерів запалення. Найвищий рівень кореляції у розподілі зразків по кластерах було виявлено для груп генів, пов'язаних з раком та мікрооточенням пухлини. ***Висновки.*** Отримані результати показали наявність кореляцій та високого рівня дисперсії експресії досліджуваних генів, що дозволило виявити кілька молекулярних кластерів. Небхідно провести більш глибокий аналіз для визначення клінічно значущих молекулярних підтипів та встановлення найбільш важливих маркерів експресії в біологічних модулях досліджуваних генів для діагностики, прогнозування та ефективного лікування раку передміхурової залози.

***Ключові слова*:** рак передміхурової залози, патерни експресії генів, TCGA, кластерний аналіз, молекулярні підтипи, гени, пов›язані з раком передміхурової залози, гени, пов›язані з мікрооточенням пухлини, гени ліпідного метаболізму.