文章编号：2095-2163(2019)05-0030-06 中图分类号：TP311.13 文献标志码：A

# 基于 R 语言的前列腺癌样本的关键基因数据挖掘

孙泽坤，袁钱图，胡建新
（贵州大学 医学院，贵阳 550025）

**摘 要：** 为寻找前列腺癌组织与正常前列腺组织的关键基因，从 Gene Expression Omnibus（GEO）数据库下载前列腺癌样本基因表达谱数据集 GSE69223。进行芯片标准化处理后设置阈值|log2(FC)|>2 且 *pvalue*<0.05 筛选出差异表达的基因，选择其中高表达的 41 个基因进行 GO 和 KEGG 分析，得出 8 个关键基因：FFAR2、THBS4、TRPM4、CLDN3、CLDN8、HPN、PLA2G2A 和 FOLH1 基因。再经 UALCAN 生存分析的到 3 个上调后患者生存可能性降低的基因：FFAR2、HPN 和 FOLH1。得出的 8 个关键基因主要富集在细胞趋化性、细胞-细胞连接、脂肪酸代谢等通路，这些通路与前列腺癌的发生发展有着密切联系。除文献已经报道的与前列腺癌有密切联系的基因外，研究推测：CLDN3、CLDN8 和 FFAR2 基因可能与前列腺癌特别是处于 T2、T3 分期的前列腺癌有着潜在的联系。

**关键词：** R 语言；数据挖掘；前列腺癌；关键基因

## Key gene data mining of Prostate Cancer samples based on R language

SUN Zekun, YUAN Qiantu, HU Jianxin
(College of Medicine, Guizhou University, Guiyang 550025, China)

**【Abstract】** To search for key genes in Prostate Cancer tissues and normal prostate tissues, the gene expression profile data set GSE69223 of Prostate Cancer samples is downloaded from the Gene Expression Omnibus (GEO) database. After the chip standardization treatment, the threshold | log2(FC) | >2 and *pvalue*<0.05 are used to screen out the differentially expressed genes, and 41 genes with high expression are selected for GO and KEGG analysis, and 8 key genes are obtained: FFAR2, THBS4, TRPM4, CLDN3, CLDN8, HPN, PLA2G2A and FOLH1 genes. After UALCAN survival analysis, the genes with reduced survival probability after three up-regulations are: FFAR2, HPN and FOLH1. The eight key genes are mainly enriched in cell chemotaxis, cell-cell junction, and fatty acid metabolism. These pathways are closely related to the development of Prostate Cancer. In addition to the genes already reported in the literature that are closely related to Prostate Cancer, it is hypothesized that the CLDN3, CLDN8, and FFAR2 genes may be potentially associated with Prostate Cancer, particularly Prostate Cancer at T2 and T3 stages.

**【Key words】** R language; data mining; Prostate Cancer; key genes

## 0 引 言

前列腺癌(Prostate Cancer,PCa)是男性常见恶性肿瘤之一,多发于老年男性,同时具有高转移性,且早期没有明显症状,发现可能已经是晚期[1]。据美国癌症协会估计,2018 年美国有大约 164 690 例新发 PCa 病例。同年大约有 29 430 例死于该病,这使其在世界致癌诱因统计榜单中已排至第二位[2]。与大多数其它癌症一样,PCa 病情的发展取决于其扩散,因此局部疾病患者的 5 年生存率几乎为 100%,癌症转移患者的生存率将下降至 28%[3]。中国前列腺癌发病率虽远低于欧美国家,但随着中国社会老龄化程度的逐渐提高、饮食结构及生活习惯的不断改变、诊疗水平及生产工艺的亟待改进等因素,中国前列腺癌的发病率也有逐年上升的趋势[4]。

研究可知,R 语言是由 Ihaka 和 Gentleman 教授联合开发的一种计算机语言[5],现已经主要应用于数据处理、统计计算、数学建模、数据可视化等多个领域,是一款开源、免费、自由的面向对象的编程软件,并已拥有 Linux、(Mac)OS X、Windows 等多个版本。R 语言使用的拓展包(packages)可根据用户需要自由开发,同时还可供使用者免费下载[6]。随着计算机技术及高通量测序技术的发展,生物芯片已然成为临床样本分析的一种有效方法,为疾病预测、

**基金项目：** 贵州省科技计划项目(黔科合基础[2019]1208 号)。

**作者简介：** 孙泽坤(1996-),男,硕士研究生,主要研究方向:医学信息处理; 袁钱图(1994-),男,硕士研究生,主要研究方向:光谱数据采集与处理; 胡建新(1970-),男,博士,教授,硕士生导师,主要研究方向:男性不育症的病因和发病机制。

**通讯作者：** 胡建新 Email:982312329@qq.com

**收稿日期：** 2019-07-10

分子诊断、新药开发发挥着强有力的助益作用[7-8]。本研究采用了基于 R 语言的芯片分析方法来研究前列腺癌与正常前列腺组织之间的基因差异，从 GEO 数据（https://www.ncbi.nlm.nih.gov/geo/）下载基因表达谱数据集 GSE69223 后对样本进行质量检测，数据清洗后设定阈值 | log2（FC）| >2，*pvalue*<0.05，筛选出差异表达基因（FC：fold change 基因倍数变化），对其中的上调基因进行 KEGG 和 GO 分析以及 UALACN（http://ualcan.path.uab.edu/）生存分析，从而发现了一些前列腺癌、特别是处于 T2、T3 分期的前列腺癌的关键基因，对研究前列腺癌的分子诊断、抗前列腺癌药物候选靶点提供了有益参考。

## 1　材料与方法

### 1.1　材料

芯片数据集 GSE69223 及芯片平台数据 GPL570 从 GEO 数据库（https://www.ncbi.nlm.nih.gov/geo/）下载得到，R 语言版本为 R3.6。除内置程序包外，其余拓展包下载自 https://cran.r-project.org/及 http://bioconductor.org/packages。

### 1.2　实验方法

1.2.1　数据获取及数据清洗

GSE69223 基因表达谱芯片数据由美国 Affymetrix 公司制作，使用芯片平台为 GPL570。数据集 GSE69223 包括 15 个正常前列腺组织样本以及 15 个前列腺癌组织样本。下载 txt 格式的原始数据，使用 R 语言获取表达矩阵、分组信息、表型数据，过滤掉没有基因名对应的探针以及对应某个基因名的多个探针。

1.2.2　聚类分析和 PCA 分析

使用 R 语言中的 dist 和 hclust 函数对 30 个样品进行聚类分析，初步判断 15 个正常样本与 15 个前列腺癌样本的差异，用以检测该数据集是否具有数据挖掘的潜力。再对样本进行主成分分析（PCA），用以判断是否有潜在因子影响两者之间的差异性。

1.2.3　获得表达差异基因

用 T 检验获得包含基因名、LogFC、*pvalue* 等信息的数据框，以 | log2（FC）| >2，*pvalue*<0.05 为阈值筛选出差异基因，并规定 LogFC>2 为上调，LogFC<−2 为下调。

1.2.4　差异表达基因的 KEGG 分析和 GO 分析

使用 R 语言中的 clusterProfiler 包对差异表达基因中的上调基因进行 KEGG 分析和 GO 富集分析。找出该基因的功能和富集的 KEGG 信号通路等信息。

1.2.5　生存分析

将经 KEGG 分析和 GO 分析的上调差异基因上传到 UALACN（http://ualcan.path.uab.edu/），选择 prostate adenocarcinoma（前列腺腺癌）进行生存分析，获得差异基因与生存时间之间的关系。

## 2　结果与分析

### 2.1　数据获取及数据清洗

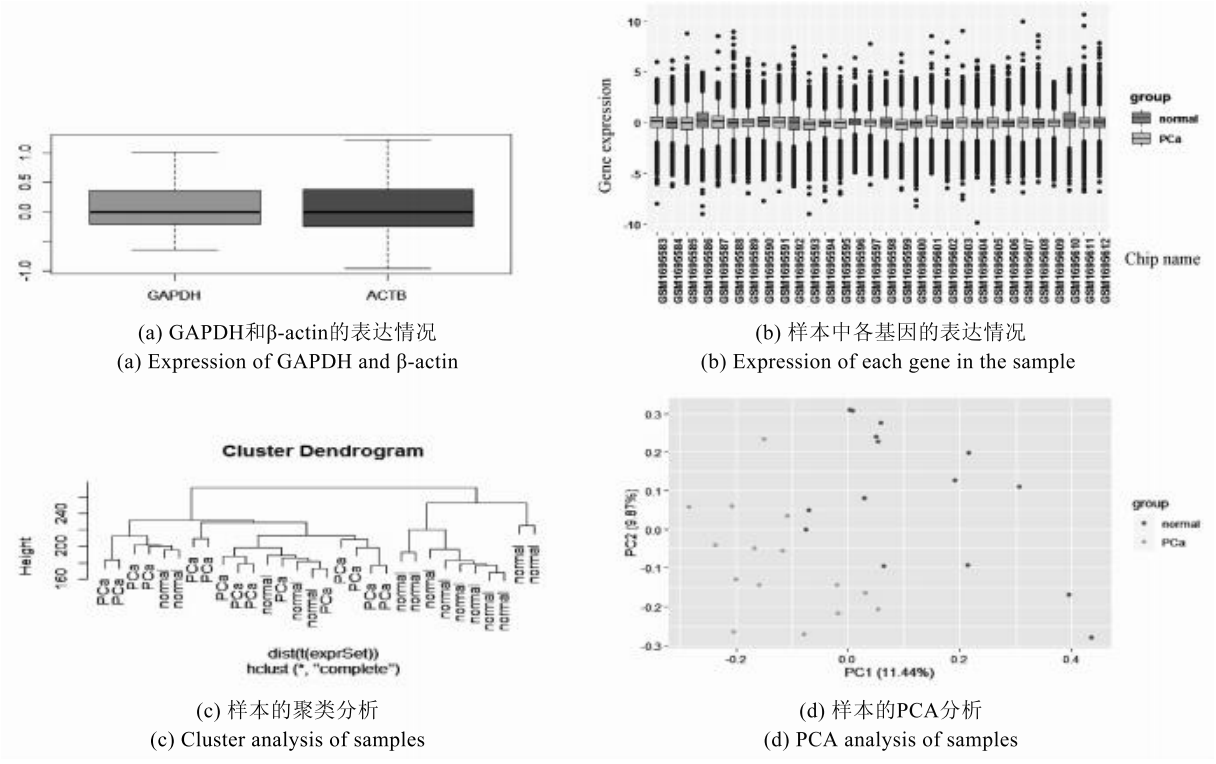
数据集包括 15 个正常前列腺样本以及 15 个前列腺癌样本的、共 54 675 个基因。通过数据清洗及标准化过程，可得与探针具有一一对应关系的基因有 23 521 个。为检验基因表达量的准确性，研究绘制了管家基因（GAPDH）以及 β-actin 的箱型图（见图 1（a）），发现两者的表达量平均值都在 0 附近，这表明此数据集中的基因表达未出现异常，在误差允许范围内可进行后续分析。将 30 个样本纳入分析范围，图 1（b）展示了各样本中基因的表达情况。

### 2.2　聚类分析和 PCA 分析

为初步判断 30 个样本中的前列腺正常样本（normal）与前列腺癌样本的差异，研究对样本进行了聚类分析和 PCA 分析。分析结果表明，样本中的某些基因的差异表达，可作为前列腺癌的诊断依据。在此次聚类分析中，有 10 个正常样本与前列腺癌样本分开，准确度达到 66.7%，但此数据集中样本总量为 30 个，分组数据较少，用聚类分析只能初步揭示正常样品与前列腺癌样品具有差异性（见图 1（c））。进一步地，对样品进行 PCA 分析。结果表明，主成分 1 对样本差异性的贡献率为 11.44%，主成分 2 对样本差异性的贡献率为 9.87%，通过主成分 1（PC1）和主成分 2（PC2）可以将前列腺正常样本与前列腺癌样本较好的进行区分（见图 1（d））。

### 2.3　获得表达差异基因

通过 T 检验，得到包含基因名、log2（FC）以及 *pvalue* 的数据框，设定阈值 *pvalue*<0.05，log2（FC）>2 以及 log2（FC）<−2 的基因，并规定 log2（FC）>2 的基因为上调基因，log2（FC）<−2 的基因为下调基因，得到 101 个下调基因和 41 个上调基因（见图 2（a））。选择前列腺癌较正常前列腺组织中的上调基因 41 个，导出其基因名及 *pvalue* 详见表 1。




(a) GAPDH和β-actin的表达情况
(a) Expression of GAPDH and β-actin

(b) 样本中各基因的表达情况
(b) Expression of each gene in the sample

(c) 样本的聚类分析
(c) Cluster analysis of samples

(d) 样本的PCA分析
(d) PCA analysis of samples

**图1 样本的聚类分析与 PCA 分析**

**Fig. 1 Cluster analysis and PCA analysis of samples**

**表1 log2(FC)>2, *pvalue*<0.05 的上调基因**

**Tab. 1 Upregulated genes with log2(FC) >2 and *pvalue*<0.05**

| 基因号 | *pvalue* | 基因号 | *pvalue* |
|---|---|---|---|
| 217111_at | 0.001 359 038 | 204934_s_at | 1.41E-08 |
| 209424_s_at | 0.000 338 451 | 1563742_at | 4.03E-06 |
| 219659_at | 0.014 960 872 | 239319_at | 3.48E-06 |
| 219521_at | 5.82E-07 | 1557598_at | 3.66E-06 |
| 230896_at | 5.23E-05 | 215323_at | 3.97E-05 |
| 223484_at | 2.14E-06 | 214614_at | 0.002 575 023 |
| 220638_s_at | 4.57E-07 | 237168_at | 0.000 553 978 |
| 205937_at | 3.91E-06 | 1553808_a_at | 0.006 503 746 |
| 203953_s_at | 2.09E-06 | 221424_s_at | 1.21E-05 |
| 214598_at | 1.95E-06 | 203649_s_at | 1.02E-05 |
| 210262_at | 0.002 547 755 | 219926_at | 2.76E-05 |
| 207147_at | 0.002 559 901 | 215806_x_at | 3.01E-06 |
| 232191_at | 0.003 249 907 | 206004_at | 3.11E-05 |
| 207260_at | 3.85E-06 | 204776_at | 4.69E-07 |
| 221345_at | 5.68E-07 | 205347_s_at | 2.00E-08 |
| 220584_at | 0.000 607 937 | 215047_at | 9.07E-06 |
| 217483_at | 0.000 590 980 | 219360_s_at | 3.25E-09 |
| 205860_x_at | 7.18E-05 | 237350_at | 1.10E-05 |
| 211303_x_at | 0.000 104 107 | 236448_at | 0.000 284 464 |
| 221582_at | 4.96E-08 | 223642_at | 0.002 786 843 |
| 206858_s_at | 0.000 129 713 | | |

## 2.4 KEGG 和 GO 分析

将得到的 41 个基因利用超几何分布原理在 KEGG 和 GO 数据库中进行比对,得到富集结果见表2、表3。GO 分析结果表明,前列腺癌细胞较前列腺正常细胞上调的差异基因主要富集的细胞活动过程有:白细胞迁移、细胞趋化性、细胞粘附、外肽酶活性、细胞 - 细胞连接。FFAR2、THBS4、TRPM4、CLDN3、CLDN8 以及 HPN 被富集到多条通路。FFAR2、HBS4 和 TRPM4 三个基因与白细胞迁移与细胞趋化性有关,白细胞迁移,可能导致前列腺癌组织中的白细胞增多,白细胞产生白介素,调控多种生理生化反应。该样本中前列腺癌样本集中于 T2、T3 分期,该时期的前列腺癌存在转移潜能,因此可能与细胞趋化性有关。CLDN3 和 CLDN8 是 Claudin 家族基因,该基因编码的蛋白由 Shoichiro Tsukita 及其同事在 1998 年发现,是细胞紧密连接的重要分子,已有报道称 Claudin-1 在结肠癌、Claudin-10 在肝细胞癌、Claudin-18 在胃癌中具有一定的临床价值[9-11]。HPN 基因又叫 Hepsin 基因,该基因编码一种 II 型跨膜丝氨酸蛋白酶,该蛋白酶可能参与多种细胞功能,包括凝血和维持细胞形态。编码蛋白的表达与癌症,尤其是前列腺癌的生长和发展有

关[12]。KEGG 富集分析结果表明，差异表达的基因主要集中在紧密连接信号通路、多种生物分子代谢信号通路（在此列举一条 α-亚麻酸代谢信号通路）、细胞粘附分子（CAMs）信号通路、黏着力信号通路、维生素消化吸收信号通路。除 GO 分析结果涉及的基因外，KEGG 分析中还出现了 PLA2G2A、FOLH1 两个基因。PLA2G2A 基因编码的蛋白是磷脂酶 A2 家族（PLA 2）的成员。该基因产物属于Ⅱ类，含有分泌型 PLA 2，这是一种低分子质量的胞外酶，需要钙离子进行催化。也可催化磷酸甘油中 sn-2 脂肪酸酰基酯键的水解，释放游离脂肪酸和溶血磷脂，并参与生物膜磷脂代谢的调控[13]。同时，通过富集的结果来看，该基因还参与其他生物大分子如亚油酸代谢、脂肪消化吸收、醚脂代谢、花生四烯酸代谢、甘油磷脂代谢。而脂肪酸的氧化代谢过程已被证实与前列腺癌的发生和发展有着密切联系[14]。FOLH1 基因编码属于 M28 肽酶家族的Ⅱ型跨膜糖蛋白。该蛋白以谷氨酸羧肽酶的形式存在于不同的替代底物上，包括营养叶酸和神经肽 N-乙酰-1-天冬氨酰-1-谷氨酸，在前列腺、中枢神经、外周神经系统和肾脏等多种组织中均有表达。在前列腺中，该基因编码的蛋白质（PSMA）在癌细胞中被上调，并被用作前列腺癌的有效诊断和预后指标[15]。

**表 2　GO 富集分析**

**Tab. 2　GO enrichment analysis**

| 功能 | 分类 | 基因数 | 基因名 | *P* 值 |
|---|---|---|---|---|
| leukocyte migration | BP | 4 | FFAR2<br>NKX2-3<br>THBS4<br>TRPM4 | 0.007 818 805 |
| cell chemotaxis | BP | 3 | FFAR2<br>THBS4<br>TRPM4 | 0.012 029 327 |
| calcium-independent cell-cell adhesion via plasma membrane cell-adhesion molecules | BP | 2 | CLDN3<br>CLDN8 | 0.000 944 841 |
| exopeptidase activity | MF | 3 | FOLH1<br>FOLH1B<br>HPN | 0.000 740 883 |
| cell-cell junction | CC | 3 | CLDN3<br>CLDN8<br>HPN | 0.028 800 572 |

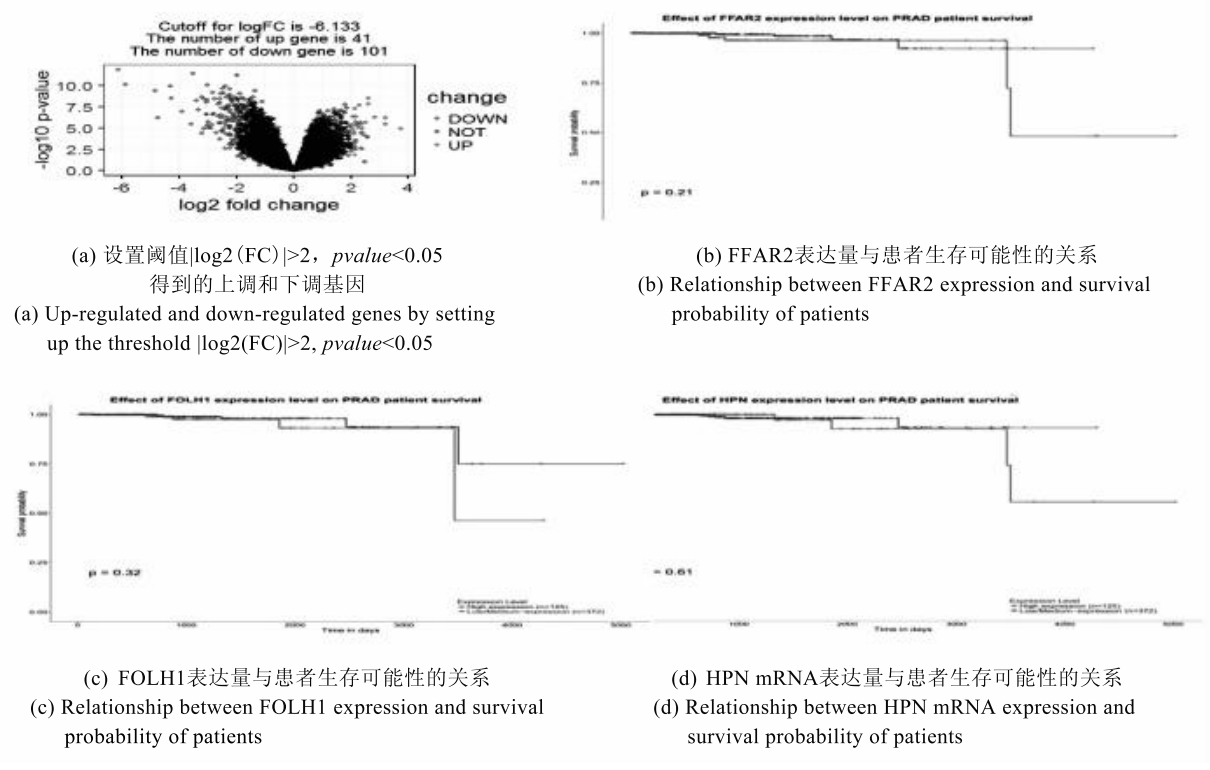
**表 3　KEGG 富集分析**

**Tab. 3　KEGG enrichment analysis**

| 通路 ID | 通路名称 | 基因数 | *P* 值 | 基因 |
|---|---|---|---|---|
| hsa04530 | Tight junction | 2 | 0.045 566 413 | CLDN3　CLDN8 |
| hsa00592 | alpha-Linolenic acid metabolism | 1 | 0.049 629 717 | PLA2G2A |
| hsa04514 | Cell adhesion molecules (CAMs) | 2 | 0.034 536 177 | CLDN3　CLDN8 |
| hsa04510 | Focal adhesion | 1 | 0.336 184 490 | THBS4 |
| hsa04977 | Vitamin digestion and absorption | 1 | 0.047 689 690 | FOLH1 |

## 2.5　生存分析

在 UALCAN 得到的生存分析结果中，研究发现，在候选的 8 个基因中，有 5 个基因的高表达组的生存可能较高，而 FFAR2、FOLH1、HPN 高表达组的生存可能性较低（见图 2(b)～(d)）。其中，已经有文献报道 HPV 编码的蛋白与前列腺癌有关[12]，FOLH1 基因编码的蛋白已成为前列腺癌的肿瘤标志物[11]，而在相同的数据库相同样本的情况下，FFAR2 组的 *P* 值最小，差异最为显著，因此研究推断，FFAR2 基因与前列腺癌有较大关联性。




(a) 设置阈值|log2(FC)|>2，*pvalue*<0.05 得到的上调和下调基因

(a) Up-regulated and down-regulated genes by setting up the threshold |log2(FC)|>2, *pvalue*<0.05

(b) FFAR2表达量与患者生存可能性的关系

(b) Relationship between FFAR2 expression and survival probability of patients

(c) FOLH1表达量与患者生存可能性的关系

(c) Relationship between FOLH1 expression and survival probability of patients

(d) HPN mRNA表达量与患者生存可能性的关系

(d) Relationship between HPN mRNA expression and survival probability of patients

**图2 相关基因表达的生存分析**

**Fig. 2 Survival analysis of related gene expressions**

## 3 结束语

R语言作为一种操作简单、免费、开源的编程语言，适用于多种操作系统，为使用者提供了极大的方便。此次研究从GSE69223基因表达谱数据集中获取了54 675个基因，设定阈值 *pvalue*<0.05，log2(FC)>2，筛选出其中的41个上调基因，并对这些基因进行了KEGG分析和GO分析，获得8个关键基因FFAR2、THBS4、TRPM4、CLDN3、CLDN8、HPN、PLA2G2A以及FOLH1。其中，FFAR2、THBS4、TRPM4三个基因与细胞趋化性相关，查看该数据集的表型数据发现，肿瘤样本全部处于T2、T3时期，此3个基因的上调，印证了该分期的前列腺癌继续发展可能转移的事实。CLDN3和CLDN8属于Claudin家族基因，该基因编码的蛋白是细胞紧密连接的重要分子，已有报道称Claudin-1在结肠癌、Claudin-10在肝细胞癌、Claudin-18在胃癌中具有一定的临床价值，因此研究推测CLDN3和CLDN8两个基因可能与前列腺癌有潜在联系。PLA2G2A参与多种脂类大分子代谢，而脂肪酸的氧化代谢过程已被证实与前列腺癌的发生发展有着密切联系。HPN和FOLH1已被文献报道与前列腺癌有着密切联系，并且FOLH1编码的蛋白(PSMA)还被用作前列腺癌的肿瘤标志物，在前列腺癌的诊断和预后中起着不可替代的作用。通过生存分析，研究还发现这8个关键基因中，FFAR2、HPN以及FOLH1三个基因的高表达会减低患者生存可能性，除文献已经报道的HPN核FOLH1基因外，本文再次经过分析推测后指出，FFAR2基因与前列腺癌的发生及发展有着潜在的关联性。但要明确其具体机制，却还需展开进一步研究。

## 参考文献

[1] SHI Wei, DONG Li, BAO Junsheng. Progress in the studies of prostate cancer related molecules [J]. National Journal of Andrology, 2015, 21(4):357-362.

[2] America Cancer Society. Cancer Information, Answers, and Hope [EB/OL]. https://www.cancer.org/cancer/prostate-cancer/about/key statistics.html.

[3] MILLER K D, SIEGEL R L, LIN C C, et al. Cancer treatment and survivorship statistics,2016[J]. CA Cancer J Clin. 2016,66(4):271-289.

[4] 万克松. 手术去势间断联合抗雄激素药物治疗晚期前列腺癌临床疗效研究[D]. 广州:南方医科大学, 2012.

[5] IHAKA R, GENTLEMAN R. R: A language for data analysis and graphics [J]. Journal of Computational and Graphical Statistics,1996, 5(3):299-314.

(下转第39页)