European Review for Medical and Pharmacological Sciences 2014; 18: 3504-3510

Both genes and lncRNAs can be used as biomarkers of prostate cancer by using high throughput sequencing data

W.-S. CHENG[1], H. TAO[1], E.-P. HU[1], S. LIU[1], H.-R. CAI[1], X.-L. TAO[1], L. ZHANG[1], J.-J. MAO[2], D.-L. YAN[1]

[1]Department of Urology, Taizhou Municipal Hospital, Taizhou, Zhejiang, China
[2]Department of Infectious Diseases, Taizhou Municipal Hospital, Taizhou, Zhejiang, China

Abstract. – OBJECTIVE: To investigate prostate cancer-related genes and lncRNAs by using a high throughput sequencing dataset.

MATERIALS AND METHODS: RNA-seq data were obtained from the sequencing read archive database, including both benign and malignant tumor samples. After aligning the RNA-seq reads to human genome reference, gene expression profile as well as lncRNA expression profile was obtained. Next, Student's t-test was used to screen both the differentially expressed genes (DEGs) and lncRNAs (DELs) between benign and malignant samples. Finally, Goseq was used to conduct the functional annotation of DEGs.

RESULTS: A total of 7112 DEGs were screened, such as ZNF512B, UCKL1, STMN3, GMEB2, and PTK6. The top 10 enriched functions of DEGs were mainly related to organism development, including multi-cellular development, system development and anatomical structure development. Also, we discovered 26 differentially expressed lncRNAs.

CONCLUSIONS: The analysis used in this study is reliable in screening prostate cancer markers including both genes and lncRNAs by using RNA-seq data, which provides new insight into the understanding of molecular mechanism of prostate cancer.

Keywords
Prostate cancer, RNA-seq, lncRNA, DEG, DEL.

Introduction

Prostate cancer is one type of adenocarcinoma (also named glandular cancer), which begins when normal semen-secreting prostate gland cells become being mutated into cancerous cells. Prostate cancer is the second leading cause of death from cancer in men, and the most common noncutaneous cancer among men in the western world[1]. Most patients with metastatic prostate cancer still die from this disease, despite of transient efficient of androgen deprivation therapy[2]. Thus, it is of great importance to understanding the molecular mechanism of the development of prostate cancer so as to make better treatment to this lethal disease.

Prostate cancer risk has been shown to have a strong genetic component[3]. Several genome-wide association works have identified numerous common variants conferring risk of prostate cancer[4,5]. A number of regions across the genome have been reported to be associated with prostate cancer, including chromosome aberrations[6], gene mutations and gene fusions[7]. Genes, such as HPC1[8], HPCX[9], and p53[10] suggested to be involved in the pathways associated with prostate cancer. Notwithstanding plenty of studies have been performed in order to fully reveal the genetic changes associated with the development of prostate cancer tissues, the full spectrum of prostate cancer genomic alterations remains incompletely characterized, needless to say the lncRNA (long non-coding RNAs) markers that may participate in the prostate cancer genesis.

In the current study, the RNA-data of benign and malignant prostate cancer samples were collected to profile both gene expression and lncRNA expression, following by a statistical test to screen the differentially expressed genes (DGEs) and differentially expressed lncRNAs (DELs), as well as a functional enrichment analysis. Our study may add up to the thorough understanding of the molecular mechanism knowledge of prostate cancer in terms of coding genes and long non-coding RNAs, which will possibly contribute to the revealing the potential mechanism in the process from benign status to malignant status.

Corresponding Author: Dongliang Yan, MD; e-mail: dongliangyan123@126.com

Materials and Methods

RNA-seq Data Acquisition and Analysis

Sequencing data of prostatic cancer were obtained from the Sequencing Read Archive (SRA) database in NCBI website. The data consist of 3 benign tumor samples and 3 malignant tumor samples[11]. Annotation information of the sequencing data was also downloaded from the Illumina GAIIx platform. The accession numbers were SRR073760, SRR073761, and SRR073762 for benign samples, and SRR073769, SRR073770, and SRR073771 for malignant samples. The quality of the sequencing data was evaluated by using a tool called FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Parameters include the GC and AT contents, number of total reads, and average length of the reads.

Reads Alignment to Human Genome and RNA Expression Quantification

The sequencing data were aligned to the human genome (hg19) downloaded from UCSC (University of California Santa Cruz) using TopHat[12]. The parameter was set as –g 1, and only the sequences with unique location were left to quantify RNA expression. Ensemble gene annotation was used to annotate the gene expression profiles across all 6 samples, and genes with at least one read count were defined as expressed genes. The raw expression file was normalized using the RPKM (Reads Per Kilobase of exon per Million mapped reads) method. We used the same procedure to profile the expression matrix of lncRNA, of which the annotation was downloaded from Human Body Map lincRNAs database [13]. RPKM was also used to normalize the expression of lncRNA across all samples. lncRNAs with RPKM >1 was defined as expressed lncRNA.

Principle Component Analysis and Hierarchical Cluster Analysis

We performed a principle component analysis by using R's 'princomp' method in 'stats' package for gene expression profile and lncRNA expression profile respectively. Benign samples and cancerous samples are labeled with different colors. Further, we also made a hierarchical cluster analysis for both gene expression and lncRNA expression by using the Euclidian distance.

DEGs and DELs Analysis

Aiming to screen prostate cancer associated RNA molecules including coding genes and long non-coding RNAs, we performed a Student's T-test for gene expression file and lncRNA expression profile separately to distinguish differentially expressed molecules between benign samples and malignant samples. To evaluate the false discovery rate (FDR), we shuffled the benign samples and malignant samples 100,000 times randomly and redid the analysis to calculate the random p value distribution.

Functional Enrichment Analysis of DEGs

For RNA sequencing data, length of different genes varies largely, and this could be a potential factor that could influence the robustness of GO enrichment analysis by introducing bias. Goseq [14] was suggested to be able to eliminate the bias caused by gene length difference. Hence we used this tool to analyze the enriched Gene Ontology terms and infer the functions of DEGs. GO terms with a p value less than 0.05 after Bonferroni correction were defined as significantly enriched GO.

Results

RNA-seq Data Alignment and Quality Analysis

After aligning the sequencing reads to human genome (hg19), we found that in average about 90% percent of the reads can be mapped back to human genome reference (Table I), indicating a good mapping ratio as well as a great usable data ratio in the downstream analysis. Also, we analyzed the quality of the sequencing data to investigate parameters such as GC content, total reads number, median reads length and average base quality value (Table I).

Gene Expression Analysis

Principal component analysis and hierarchical cluster analysis were performed to compare the gene expression difference of benign the malignant samples in a global view. As shown in Figure 1 A and B, benign and malignant samples are well separated on gene expression level indicated by both PCA plot and hierarchical cluster plot, indicating that prostatic tumor samples can distinguish themselves from normal samples on gene expression level. Interestingly, we also observed that on lncRNA expression level, this effect also exists. Specifically, 3 cancer samples are well separated from the other 3 normal samples from both PCA plot (Figure 1 C) and hierarchical cluster plot (Figure 1 D). These evidences suggest that besides genes,

Table I. Statistics of the raw RNA-seq data.

Sample name	GC content	AT content	Total reads	Mapped reads	Mapping ratio	Average length
SRR073760	53%	47%	5300188	4823172	91%	35
SRR073761	53%	47%	5347764	4759514	89%	35
SRR073762	54%	46%	4778245	4300420	90%	35
SRR073769	53%	47%	8175900	7358312	90%	35
SRR073770	52%	48%	5372814	4835540	90%	35
SRR073771	53%	47%	5210292	4637162	89%	35

lncRNAs are also candidate molecular markers that can distinguish prostate samples from benign samples, indicating the involvement of lncRNAs in the regulation and development of prostate cancer.

DEGs and DELs Screening Analysis

In order to detect genes involved in the process from benign to malignant, Student's *t*-test was conducted to identify the DEGs. In total, 7112 DEGs were identified when setting the threshold of $p < 0.0001$. Typical DEGs include ZNF512B, UCKL1, STMN3, GMEB2, and PTK6. The distribution of DEGs p values is shown in Figure 2 A. The randomization test (see materials and methods) shows that the false discovery rate (FDR) is lower than 0.01, suggesting these DEGs are more true signals than background noise. Further, we used the same test to screen differentially expressed lncRNAs (DELs). In the end, we obtained 26 significant DELs out of 725 expressed lncRNAs. The p value distribution for lncRNA is shown in Figure 2 B.

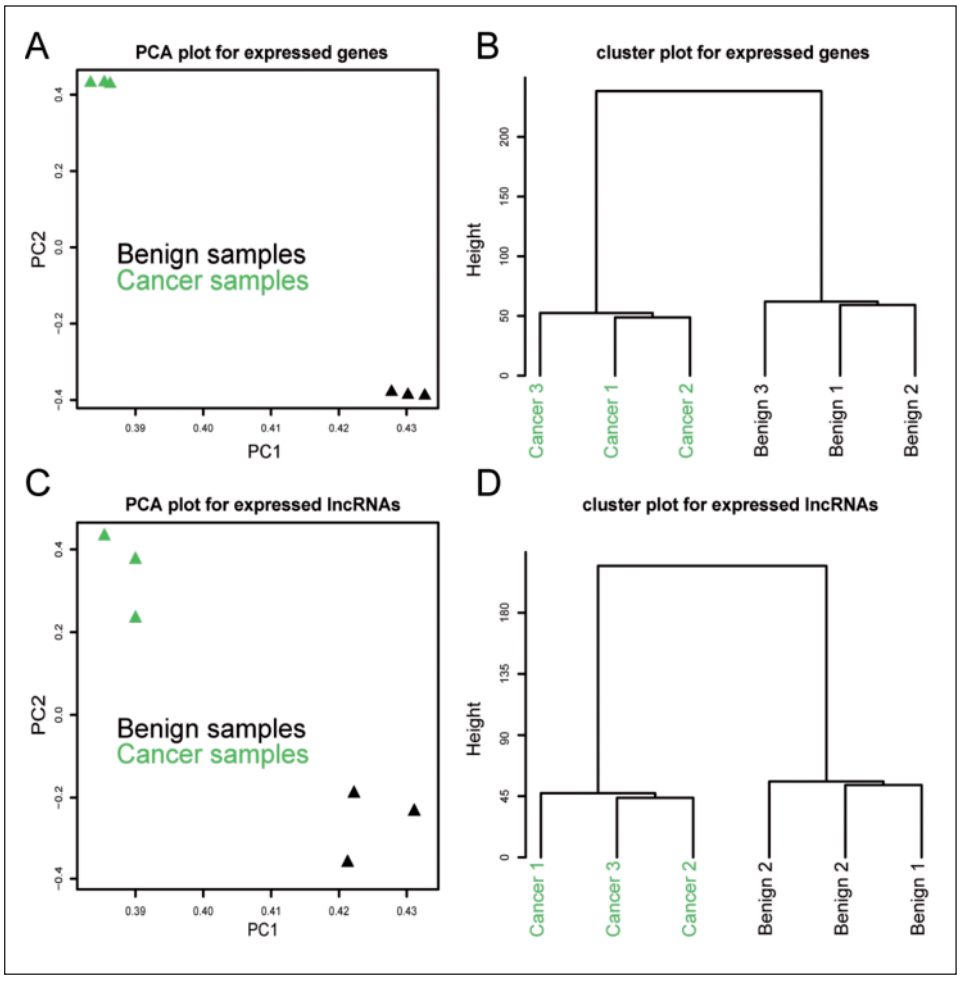

Figure 1. Principle component analysis (**A**) and hierarchical cluster analysis (**B**) of 6 samples including 3 benign samples and 3 cancer samples on gene expression level. Principle component analysis (**C**) and hierarchical cluster analysis (**D**) of 6 samples on lnc RNA expression level.

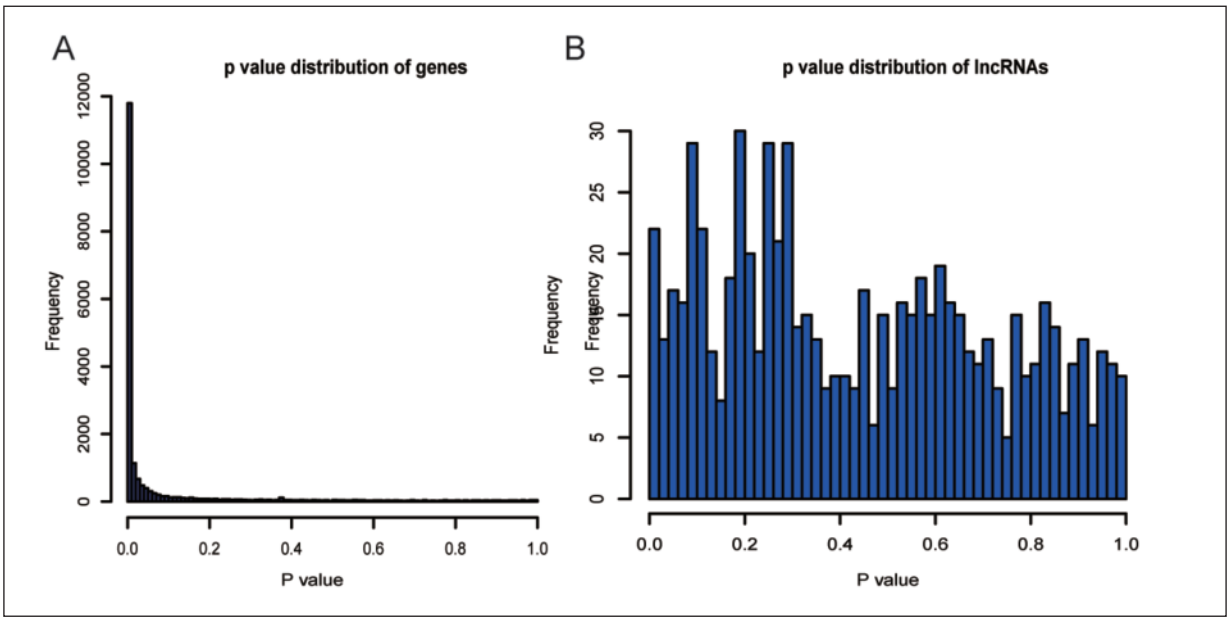

Figure 2. Histogram of p-value of the results of Student's test for both (**A**) expressed genes and (**B**) expressed lncRNAs.

Functional Annotation of DEGs and DELs

To investigate the biological functions of these DEGs, we performed an enrichment test of Gene Ontology terms to obtain the significantly enriched functions. By performing this analysis on Bioconductor[15] platform using 'Goseq' package (see materials and methods). Specifically, gene length of DEGs was rectified firstly, and then was used as input for function enrichment. The result showed that a total of 194 GO terms were enriched, suggesting they may be associated with prostatic cancer. The top significant functions are showed in Table II. Also, we got 26 significantly differentially expressed lncRNAs. The neighboring genes suggest these lncRNA might have relations with genes such as BARHL2, TRPM1, NAIF1, etc.

Discussion

Prostate cancer is a genetically complex disease which is regulated by numerous molecular processes including oncogene activation or suppressor gene inactivation[16]. To investigate the prostate cancer-related genes, sequencing data of benign and malignant prostatic cancer samples were obtained from the SRA database from National Center for Biotechnology Information (NCBI). After aligning the data to human genome and gene annotation, the gene expression profile and lncRNA expression profile were obtained. Student's t-test identified a total of 7112 DEGs and 26 DELs in the process from benign to malignant. Functional annotation revealed that these DEGs main associated with organism development, including multicellular development, system development and anatomical structure development.

ZNF512B, UCKL1, STMN3, GMEB2, and PTK6 were identified as DEGs in this study. ZNF512B (Zinc figure protein 512B) is a transcription factor, which encodes an important positive regulator of TGFβ signaling[17]. GAM/ZNF512B is a vertebrate-specific developmental

Table II. Top 10 enriched functions of differentially expressed genes.

Category	Corrected *p* value
GO:0032501 multicellular organismal process	5.67E-35
GO:0048731 system development	3.92E-31
GO:0007275 multicellular organismal development	4.69E-30
GO:0005575 cellular component	5.45E-30
GO:0032502 developmental process	1.10E-28
GO:0048856 anatomical structure development	4.06E-27
GO:0007155 cell adhesion	9.77E-26
GO:0022610 biological adhesion	9.77E-26
GO:0005886 plasma membrane	1.02E-24
GO:0048513 organ development	2.53E-24

regulator[18]. ZNF512 is a biological marker indicative of an occurrence of metastasis in breast cancer patients[19]. UCKL1 (Uridine-Cytidine Kinase 1-like 1) is reported to be potential positive breast cancer markers[20]. Substantially higher levels of STMN3 (stathmin-like 3) is observed in metastatic ovarian cancer tissues[21]. STMN3 facilitates tubulin depolymerization and is regulated during mitosis[22]. Germline alterations in STMN3 might affect tubulin binding and, thus, affect mitotic segregation[23]. GME (glucocorticoid modulatory element) is in the modulation of glucocorticoid receptors transcriptional properties[24]. GMEB2 (GME-binding proteins 2) is reported as a down-regulated gene in a clear cell carcinoma[25]. PTK6 (protein tyrosine kinase 6) is a member of a distinct family of kinases that id evolutionarily related to the SRC family of tyrosine kinases, and its expression is detected in a large proportion of human mammary gland tumors[26]. The expression of PTK6 is found to be increased in androgen-independent prostate cancer cells using Q-reverse transcription-PCR[27]. Besides PKT6, the other genes have currently no relations with prostate cancer been reported. Our study indicated that the identified DEGs in the study exert important roles in the process from benign to malignant in prostate cancer.

In this study, Goseq, an application for performing gene oncology analysis on RNA-seq data was used. Goseq can markedly change the results, highlighting categories more consistent with the known biology[14]. The enriched functions of DEGs were mainly related to organismal development, suggesting that the organismal development related biological processes were changed in prostate cancer samples. Multicellular organismal process and system development have been reported to be changed in prostate cancer[28], as

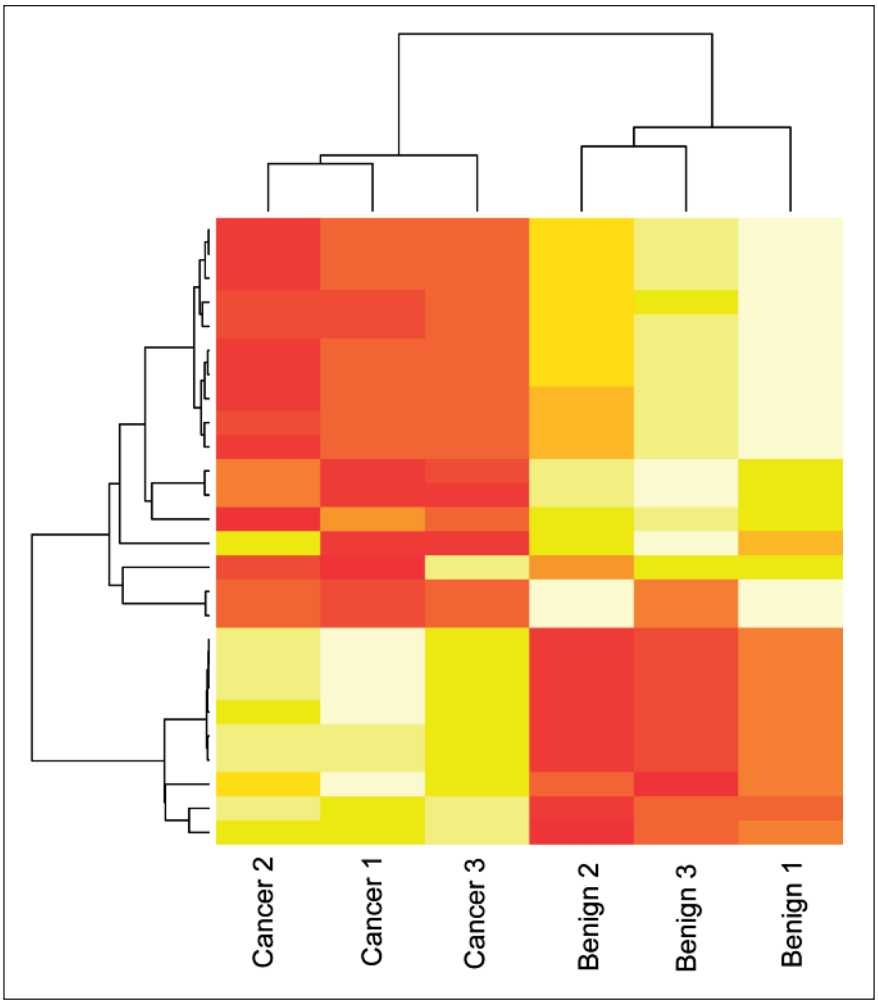

Figure 3. Heatmap plot of 26 DELs across all six samples by using the z-transformed expression as the input.

well the cellular component. Function of cell adhesion was also found to be abnormal in developing prostate[29]. Thus, our study results were consistent with the previous studies.

Further, we discovered 26 differentially expressed lncRNAs between benign samples and prostate samples, of which one third of them are down-regulated and two third of them are up-regulated. The neighboring genes suggest these candidate biomarkers associated with prostate cancer development may participate in the similar pathways as gene like BARHL2, TRPM1, NAIF1 do, indicating a potential molecular mechanism for the prostate cancer development.

Conclusions

Using RNA-seq data, DEGs such as ZNF512B, UCKL1, STMN3, GMEB2, and PTK6 were identified between benign and malignant prostate cancer, the main enriched functions of them were mainly related with organismal development. Also, we discovered 26 differentially expressed long non-coding RNAs which are candidate biomarkers for prostate cancer as well.

References

1) Kantoff PW, Higano CS, Shore ND, Berger ER, Small EJ, Penson DF, Redfern CH, Ferrari AC, Dreicer R, Sims RB, Xu Y, Frohlich MW, Schellhammer PF; for the IMPACT Study Investigators. Sipuleucel-T immunotherapy for castration-resistant prostate cancer. N Engl J Med 2010; 363: 411-422.
2) Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, Onofrio R, Carter SL, Park K, Habegger L, Ambrogio L, Fennell T, Parkin M, Saksena G, Voet D, Ramos AH, Pugh TJ, Wilkinson J, Fisher S, Winckler W, Mahan S, Ardlie K, Baldwin J, Simons JW, Kitabayashi N, MacDonald TY, Kantoff PW, Chin L, Gabriel SB, Gerstein MB, Golub TR, Meyerson M, Tewari A, Lander ES, Getz G, Rubin MA, Garraway LA. The genomic complexity of primary human prostate cancer. Nature 2011; 470: 214-220.
3) Amundadottir LT, Thorvaldsson S, Gudbjartsson DF, Sulem P, Kristjansson K, Arnason S, Gulcher JR, Bjornsson J, Kong A, Thorsteinsdottir U, Stefansson K. Cancer as a complex phenotype. pattern of cancer distribution within and beyond the nuclear family. PLoS Med 2004; 1: e65.
4) Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, Benediktsdottir KR, Sigurdsson A, Magnusson OT, Gudjonsson SA, Magnusdottir DN, Johannsdottir H, Helgadottir HT, Stacey SN, Jonasdottir A, Olafsdottir SB, Thorleifsson G, Jonasson JG, Tryggvadottir L, Navarrete S, Fuertes F, Helfand BT, Hu Q, Csiki IE, Mates IN, Jinga V, Aben KK, van Oort IM, Vermeulen SH, Donovan JL, Hamdy FC, Ng CF, Chiu PK, Lau KM, Ng MC, Gulcher JR, Kong A, Catalona WJ, Mayordomo JI, Einarsson GV, Barkardottir RB, Jonsson E, Mates D, Neal DE, Kiemeney LA, Thorsteinsdottir U, Rafnar T, Stefansson K. A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet 2012; 44: 1326-1329.
5) Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, Quist MJ, Jing X, Lonigro RJ, Brenner JC, Asangani IA, Ateeq B, Chun SY, Siddiqui J, Sam L, Anstett M, Mehra R, Prensner JR, Palanisamy N, Ryslik GA, Vandin F, Raphael BJ, Kunju LP, Rhodes DR, Pienta KJ, Chinnaiyan AM, Tomlins SA. The mutational landscape of lethal castration-resistant prostate cancer. Nature 2012; 487: 239-243.
6) Qian J, Jenkins RB, Bostwick DG. Genetic and chromosomal alterations in prostatic intraepithelial neoplasia and carcinoma detected by fluorescence in situ hybridization. Eur Urol 1999; 35: 479-483.
7) Mani RS, Tomlins SA, Callahan K, Ghosh A, Nyati MK, Varambally S, Palanisamy N, Chinnaiyan AM. Induced chromosomal proximity and gene fusions in prostate cancer. Science 2009; 326: 1230.
8) Sood R, Bonner TI, Makalowska I, Stephan DA, Robbins CM, Connors TD, Morgenbesser SD, Su K, Faruque MU, Pinkett H, Graham C, Baxevanis AD, Klinger KW, Landes GM, Trent JM, Carpten JD. Cloning and characterization of 13 novel transcripts and the human RGS8 gene from the 1q25 region encompassing the hereditary prostate cancer (HPC1) locus. Genomics 2001; 73: 211-222.
9) Lange EM, Chen H, Brierley K, Perrone EE, Bock CH, Gillanders E, Ray ME, Cooney KA. Linkage analysis of 153 prostate cancer families over a 30-cM region containing the putative susceptibility locus HPCX. Clin Cancer Res 1999; 5: 4013-4020.
10) Moul JW. Angiogenesis, p53, bcl-2 and Ki-67 in the progression of prostate cancer after radical prostatectomy. Eur Urol 1999; 35: 399-407.
11) Prensner JR, Iyer MK, Balbin OA, Dhanasekaran SM, Cao Q, Brenner JC, Laxman B, Asangani IA, Grasso CS, Kominsky HD, Cao X, Jing X, Wang X, Siddiqui J, Wei JT, Robinson D, Iyer HK, Palanisamy N, Maher CA, Chinnaiyan AM. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 2011; 29: 742-749.
12) Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 2009; 25: 1105-1111.
13) Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 2011; 25: 1915-1927.
14) Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq. accounting for selection bias. Genome Biol 2010; 11: R14.

15) Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JY, Zhang J. Bioconductor. open software development for computational biology and bioinformatics. Genome Biol 2004; 5: R80.

16) Ulkus L, Wu M, Cramer SD. Stem cell models for functional validation of prostate cancer genes. In. Stem Cells and Prostate Cancer. Springer, 2013; pp. 149-173.

17) Galbiati M, Onesto E, Zito A, Crippa V, Rusmini P, Mariotti R, Bentivoglio M, Bendotti C, Poletti A. The anabolic/androgenic steroid nandrolone exacerbates gene expression modifications induced by mutant SOD1 in muscles of mice models of amyotrophic lateral sclerosis. Pharmacol Res 2012; 65: 221-230.

18) Tili E, Michaille JJ. Resveratrol, MicroRNAs, Inflammation, and Cancer. J Nucleic Acids 2011; 2011: 102431.

19) Lidereau R, Driouch K, Landemaine T. Method for predicting the occurrence of metastasis in breast cancer patients. In: Google Patents, 2008.

20) Geiger T, Madden SF, Gallagher WM, Cox J, Mann M. Proteomic portrait of human breast cancer progression identifies novel prognostic markers. Cancer Res 2012; 72: 2428-2439.

21) Walter-Yohrling J, Cao X, Callahan M, Weber W, Morgenbesser S, Madden SL, Wang C, Teicher BA. Identification of genes expressed in malignant cells that promote invasion. Cancer Res 2003; 63: 8939-8947.

22) Charbaut E, Curmi PA, Ozon S, Lachkar S, Redeker V, Sobel A. Stathmin family proteins display specific molecular and tubulin binding properties. J Biol Chem 2001; 276: 16146-16154.

23) Jenkins RB, Wrensch MR, Johnson D, Fridley BL, Decker PA, Xiao Y, Kollmeyer TM, Rynearson AL, Fink S, Rice T, McCoy LS, Halder C, Kosel ML, Giannini C, Tihan T, O'Neill BP, Lachance DH, Yang P, Wiemels J, Wiencke JK. Distinct germ line polymorphisms underlie glioma morphologic heterogeneity. Cancer Genet 2011; 204: 13-18.

24) Kaul S, Blackford JA, Jr., Cho S, Simons SS, Jr. Ubc9 is a novel modulator of the induction properties of glucocorticoid receptors. J Biol Chem 2002; 277: 12541-12549.

25) Jutras S, Bachvarova M, Keita M, Bascands JL, Mes-Masson AM, Stewart JM, Gera L, Bachvarov D. Strong cytotoxic effect of the bradykinin antagonist BKM-570 in ovarian cancer cells--analysis of the molecular mechanisms of its antiproliferative action. FEBS J 2010; 277: 5146-5160.

26) Brauer PM, Tyner AL. Building a better understanding of the intracellular tyrosine kinase PTK6-BRK by BRK. Biochim Biophys Acta 2010; 1806: 66-73.

27) Singh AP, Bafna S, Chaudhary K, Venkatraman G, Smith L, Eudy JD, Johansson SL, Lin MF, Batra SK. Genome-wide expression profiling reveals transcriptomic variation and perturbed gene networks in androgen-dependent and androgen-independent prostate cancer cells. Cancer Lett 2008; 259: 28-38.

28) Yano K. Gene expression correlation analysis predicts involvement of high- and low-confidence risk genes in different stages of prostate carcinogenesis. Prostate 2010; 70: 1746-1759.

29) Pritchard CC, Nelson PS. Gene expression profiling in the developing prostate. Differentiation 2008; 76: 624-640.