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

W.-S. CHENG¹, H. TAO¹, E.-P. HU¹, S. LIU¹, H.-R. CAI¹, X.-L. TAO¹, L. ZHANG¹, J.-J. MAO², D.-L. YAN¹

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

Abstract. – OBJECTIVE: To investigate prostate cancer-related genes and IncRNAs 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 IncRNA expression profile was obtained. Next, Student’s t-test was used to screen both the differentially expressed genes (DEGs) and IncRNAs (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 IncRNAs.
CONCLUSIONS: The analysis used in this study is reliable in screening prostate cancer markers including both genes and IncRNAs by using RNA-seq data, which provides new insight into the understanding of molecular mechanism of prostate cancer.

Keywords
Prostate cancer, RNA-seq, IncRNA, 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 world¹. Most patients with metastatic prostate cancer still die from this disease, despite of transient efficient of androgen deprivation therapy². 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 component³. Several genome-wide association works have identified numerous common variants conferring risk of prostate cancer⁴,⁵. A number of regions across the genome have been reported to be associated with prostate cancer, including chromosome aberrations⁶, gene mutations and gene fusions⁷. Genes, such as HPC1⁸, HPCX⁹, and p53¹⁰ 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 IncRNA (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 IncRNA expression, following by a statistical test to screen the differentially expressed genes (DGEs) and differentially expressed IncRNAs (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.
Both genes and IncRNAs can be used as biomarkers of prostate cancer

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 GAIx 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 IncRNA, of which the annotation was downloaded from Human Body Map IncRNAs database [13]. RPKM was also used to normalize the expression of IncRNA across all samples. IncRNAs with RPKM >1 was defined as expressed IncRNA.

**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 IncRNA 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 IncRNA 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 IncRNA 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 IncRNA 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,
IncRNAs are also candidate molecular markers that can distinguish prostate samples from benign samples, indicating the involvement of IncRNAs 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 2A. 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 IncRNAs (DELs). In the end, we obtained 26 significant DELs out of 725 expressed IncRNAs. The p value distribution for IncRNA is shown in Figure 2B.

**Table 1. Statistics of the raw RNA-seq data.**

<table>
<thead>
<tr>
<th>Sample name</th>
<th>GC content</th>
<th>AT content</th>
<th>Total reads</th>
<th>Mapped reads</th>
<th>Mapping ratio</th>
<th>Average length</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR073760</td>
<td>53%</td>
<td>47%</td>
<td>5300188</td>
<td>4823172</td>
<td>91%</td>
<td>35</td>
</tr>
<tr>
<td>SRR073761</td>
<td>53%</td>
<td>47%</td>
<td>5347764</td>
<td>4759514</td>
<td>89%</td>
<td>35</td>
</tr>
<tr>
<td>SRR073762</td>
<td>54%</td>
<td>46%</td>
<td>4778245</td>
<td>4300420</td>
<td>90%</td>
<td>35</td>
</tr>
<tr>
<td>SRR073769</td>
<td>53%</td>
<td>47%</td>
<td>8175900</td>
<td>7358312</td>
<td>90%</td>
<td>35</td>
</tr>
<tr>
<td>SRR073770</td>
<td>52%</td>
<td>48%</td>
<td>5372814</td>
<td>4855540</td>
<td>90%</td>
<td>35</td>
</tr>
<tr>
<td>SRR073771</td>
<td>53%</td>
<td>47%</td>
<td>5210292</td>
<td>4637162</td>
<td>89%</td>
<td>35</td>
</tr>
</tbody>
</table>

![Figure 1](image1.png)

**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 IncRNA expression level.
Both genes and lncRNAs can be used as biomarkers of prostate cancer.

**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 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. 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. GAM/ZNF512B is a vertebrate-specific developmental

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

<table>
<thead>
<tr>
<th>Category</th>
<th>Corrected p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0032501 multicellular organismal process</td>
<td>5.67E-35</td>
</tr>
<tr>
<td>GO:0048731 system development</td>
<td>3.92E-31</td>
</tr>
<tr>
<td>GO:0007275 multicellular organismal development</td>
<td>4.69E-30</td>
</tr>
<tr>
<td>GO:0005575 cellular component</td>
<td>5.45E-30</td>
</tr>
<tr>
<td>GO:0032502 developmental process</td>
<td>1.10E-28</td>
</tr>
<tr>
<td>GO:0048856 anatomical structure</td>
<td>4.06E-27</td>
</tr>
<tr>
<td>GO:0007155 cell adhesion</td>
<td>9.77E-26</td>
</tr>
<tr>
<td>GO:0022610 biological adhesion</td>
<td>9.77E-26</td>
</tr>
<tr>
<td>GO:0005886 plasma membrane</td>
<td>1.02E-24</td>
</tr>
<tr>
<td>GO:0048513 organ development</td>
<td>2.53E-24</td>
</tr>
</tbody>
</table>
ZNF512 is a biological marker indicative of an occurrence of metastasis in breast cancer patients. UCKL1 (Uridine-Cytidine Kinase 1-like 1) is reported to be potential positive breast cancer markers. Substantially higher levels of STMN3 (stathmin-like 3) is observed in metastatic ovarian cancer tissues. STMN3 facilitates tubulin depolymerization and is regulated during mitosis. Germline alterations in STMN3 might affect tubulin binding and, thus, affect mitotic segregation. GME (glucocorticoid modulatory element) is in the modulation of glucocorticoid receptors transcriptional properties. GMEB2 (GME-binding proteins 2) is reported as a down-regulated gene in a clear cell carcinoma. 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. The expression of PTK6 is found to be increased in androgen-independent prostate cancer cells using Q-reverse transcription-PCR. Besides PTK6, 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. 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 as
well the cellular component. Function of cell adhesion was also found to be abnormal in developing prostate\textsuperscript{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


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.


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.