Pathol. Oncol. Res. (2014) 20:43–50
DOI 10.1007/s12253-013-9618-0

RESEARCH

A New RNA-Seq Method to Detect the Transcription and Non-coding RNA in Prostate Cancer

Xiao-Ming Zhang · Zhong-Wei Ma · Qiang Wang · Jian-Ning Wang · Ji-Wei Yang · Xian-Duo Li · Hao Li · Tong-Yi Men

Received: 19 August 2012 / Accepted: 20 February 2013 / Published online: 17 September 2013

Abstract Prostate cancer is a big killer in many regions especially American men, and this year, the diagnosed rate rises rapidly. We aimed to find the biomarker or any changing in prostate cancer patients. With the development of next generation sequencing, much genomic alteration has been found. Here, basing on the RNA-seq result of human prostate cancer tissue, we tried to find the transcription or non-coding RNA expressed differentially between normal tissue and prostate cancer tissue. 10 T sample data is the RNA-seq data for prostate cancer tissue in this study, we found the differential gene is TFF3-Trefoil factor 3, which was more than seven fold change from prostate cancer tissue to normal tissue, and the most outstanding transcript is C15orf21. Additionally, 9 lncRNAs were found according our method. Finally, we found the many important non-coding RNA related to prostate cancer, some of them were long non-coding RNA (lncRNA).

Keywords Prostate cancer · RNA-seq · Transcription · Non-coding RNA · lncRNA

Xiao-Ming Zhang, Zhong-Wei Ma and Qiang Wang contributed equally to this work as the co-first author.

X.-M. Zhang · J.-N. Wang · J.-W. Yang · X.-D. Li · T.-Y. Men (✉)
Department of Urology, Qianfoshan Hospital Affiliated to Shandong University, No.16766 Jingshi Road, Jinan, Shandong Province 250014, China
e-mail: mentongyish3316@sina.com

Z.-W. Ma · H. Li
Department of Urology, Hospital of Shandong Aluminum Corporation, Zibo 255069, China

Q. Wang
Department of Urology, 309 Hospital of the Chinese People's Liberation Army, Beijing 100091, China

Introduction

Prostate cancer is a type of cancer that develops in the prostate, a gland in the male reproductive system. Detection rate of prostate cancer vary widely across the world, with higher rate in developed countries than in developing countries. It has been the most frequently diagnosed cancer in American men. And this trend rises rapidly in recent years. Among men in the United States, prostate cancer accounts for more than 200,000 new cancer cases and 32,000 deaths annually [1]. These evidence alerts us the importance for researching prostate cancer.

The androgen deprivation therapy yields transient efficacy in prostate cancer sufferer, and there are many patients cannot survive from this deadly killer. As the development of the Next-Generation-Sequencing, many somatic mutations or other genomic alteration has been found, our knowledge about prostate cancer mutation has been expanded. For example, by exon-sequencing of 112 pair prostate cancer tissue this year, Gordon's team not only found the three genes-*MED12*,*FOXA1* and *SPOP* which are always recurrently mutated in prostate cancer patients, but also found a gene fusion [2]. Basing on the Integrating exome copy number analysis, Kenneth identified disruptions of *CHD1* that define a subtype of ETS gene family fusion-negative prostate cancer [3]. All those genomics alteration found by next-generation-sequencing are the potential treatment target in future.

Referring to the use of high-throughput sequencing technologies, RNA-seq, which is short for "Whole Transcriptome Shotgun Sequencing-WTSS", sequence cDNA in order to get information about a sample's RNA content [4],such as gene expression level, new isoform, and so on. As soon as

this technology has published, it has adopted to disease research filed such as cancer [5]. In Mark's study, basing on the RNA-seq result of prostate cancer tissue, they detected non-ETS gene fusions in human prostate cancer. They discovered and characterized seven new cancer-specific gene fusions, two involving the ETS genes *ETV1* and *ERG* [6]. In 2012, aiming to find the ethnic variation, scientific from University of Michigan Medical School also used RNA-seq technology to deeply insight to Chinese prostate cancer patients [7].

A non-coding RNA (ncRNA) is a function RNA molecule that is not translated into a protein. It contains abundant RNA such as tRNA, miRNA, snoRNA, Piwi-RNA and rRNA and so on. The large number of ncRNA is unknown now, and recently, through many bioinformatics study and new experiment technology, many ncRNA were found, especially some small RNA. After the genome sequencing project have released, this project have revealed an unexpected problem in our understanding of the molecular basis of developmental complexity in the higher organisms: complex organisms have lower numbers of protein coding genes than anticipated. The new role-non-coding RNA have been proved to make the architects of eukaryotic much more complexity [8]. Moreover, miRNA have drew many scientific attention after the Nobel prize for the miRNA discoverer. As the important roles of those small non-coding RNA, such as miRNA, Piwi-Interaction RNA in animal development [9], the long non-coding RNA drew scientific attention either. If the length of ncRNA is greater than 200 bp, we named them long non-coding RNA (lncRNA). This rapid advance filed shows a great potential of their regulation function [10]. In 2011, Howard and his team found that the long non-coding RNA HOTAIR is increased in expression in primary breast tumors and metastases, and HOTAIR expression level in primary tumors is a powerful predictor of eventual metastasis and death [11]. All these findings suggest that non-coding, included miRNA, non-transcript genes and long ncRNAs play active roles in modulating the cancer genome and may be important targets for cancer diagnosis and therapy.

In our study, basing on the RNA-seq result of human prostate cancer tissue, we analysis the data between prostate cancer samples and control samples, aligned them, then assembled the transcripts and finally obtained the transcription and non-coding RNA, which may be important targets for cancer diagnosis and therapy.

Materials and Methods

Data Achievement

Our project is based on the RNA-seq data of a former study's sequencing result [12]. All those data is available on European Nucleotide Archive [13] (ENA; http://www.ebi.ac.uk/ena). It's the primary nucleotide-sequence repository of Europe. ENA collects comprehensive record of the world's *nucleotide* sequencing information, and consists of three main databases: the Sequence Read Archive (SRA), the Trace Archive and EMBL-Bank. When collecting sequencing data, we used the rule bellow: 1) paired-end sequencing; 2) of more than 50 bp length. Those two rules were selected because of our alignment tools. We will explain it later.

Data Preprocessing

According to the preprocessing method of the former study where our data from, we filtered the reads with the following cutoff condition: (1) N-bases number is above and beyond 2 %; (2) the low-quality bases is above and beyond 50 %(Q≤15). Then, we drew base quality distribution to profile the filtering effects.

Alignment, Assemble and Estimate Abundances

The traditional RNA-Seq data analysis method was based on denovo assembling and aligning with reference for sequencing annotation. While this method found the new transcripts only relying on matching different genes between both sides of reads, so it mostly limited the length and numbers of reads, and cannot detected the region of breakpoint.

The new method aligned the genes and cleavage site, and then built the mimetic exon-exon references data using assembling of cleavage site to find differentially expressed genes and transcription as mostly as we can.

It can fix the fragment ends to the different exons to determine which spliceosome is correct, do not need with the previous annotation information.

In this paper, we use this new method for the bioinformatics. There are three steps:

1. the first step, alignment, TopHat [14] is chose to alignment. It aligns reads to genomes using Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

We used hg19 to construct the reference library, with the following condition: 1) minimum intron length is 70; 2) maximum intron length is 500000; 3) tolerance 3 bp deletion/insertion; 4) tolerance two mismatch, samples 10 N and 10 T was mapped and then generated two bam files.

2. We used cufflinks [15] software for the second step—assembling transcripts. Some parameters was set for assemble:1) Mean Inner Distance between Mate

Pairs is 20; 2) Standard Deviation for Inner Distance between Mate Pairs is 20.

3. The third step, we also used Cufflinks estimated the relative abundances of these transcripts based on how many reads support each one. Two normalization methods Quartile and Bias correction are used for improving accuracy of transcript abundance estimates.

Merging Transcripts

The two transcript assembly result of two samples 10 N and 10 T produced were merged by the cufflinks. Mergence conditions: 1) the transcripts have different IDs and the positions are uniform; 2) the transcripts have the intersection of sets with genome mapping; 3) the distance between the transcripts is less than 500 bp. According to these conditions, we got a new transcript that is no redundancy information.

Analysis Transcripts Expression

Combined the assemble transcripts and the alignment produced by Tophat, we computed the expression value of every transcripts. Traditional expression value was represented by RPKM [16], it means the reads number of one gene per million reads, considering the impact on reads count of sequencing depth. At the same time, because the reads are pair-end, we can connect the pair reads to rebuild the fragment input to sequencer. Basing on the RPKM algorithm, we computed the fragment count, and got the FPKM value. It is more reliable to substitute the RPKM with the expression value [17].

Finding Significant Transcripts

As we can imagine, transcripts must have some significant different FPKM value between two samples. So, we combined the FPKM in two samples according to transcripts, calculated the fold change value of them, and computed the *p*-value. Then, we used these two feature value of each transcripts to plot volcano picture. After that, we can get the significance boundary to define the transcript if differentially expressed or not.

Results

Summary of Raw RNA-seq Data

The RNA-seq data which is complete transcriptomic landscape of prostate cancer in the Chinese population were downloaded from ENA. Basing on the rule we described before, we finally chose two sample-10 N and 10 T for our analysis, which are pair-end sequenced, and of 90 bp length. Detail information is shown in Table 1. 10 N sample data is the RNA-seq data for normal tissue, and 10 T sample data is the RNA-seq data for prostate cancer tissue.

Prepossessing Result of Sequencing Data

To evaluate the prepossessing method we used, we drew box plot picture of bases quality through whole reads before and after prepossessing. Figure 1 showed the distribution of bases quality map before and after filtering (Fig. 1). Certainly, the upper half part is the distribution of bases quality map of raw data, the lower half part is that of preprocessing data. The black line in each box represents the median quality score. The information this picture tells us: (1) The fluctuating of bases quality is lower in prepossessed data than in raw data, which suggested that the filter method was worked; (2) The overall data are distributing in the part more than Q15, the median value is in more than Q34 and focus on more than Q36. Consequently, after preprocessing, the quality of reads has improved significantly. The data of preprocessing is used for all our following analysis. Table 2 showed the statistics result of data before and after preprocessed (Table 2).

Alignment and Assemble

We used TopHat for sequences alignment, and Cufflink for transcripts assembling. We thought our method which aligns first is of great potential to make use of the RNA-seq data as

Table 1 Sample information table

Sample name	Type	Library	Data size	ENA ID	Download address
10 T	Tumor	Pair-end	12G base	ERR031018	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR031/ERR031018/ERR031018_1.fastq.gz
					ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR031/ERR031018/ERR031018_2.fastq.gz
10 N	Normal	Pair-end	12G base	ERR031017	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR031/ERR031017/ERR031017_1.fastq.gz
					ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR031/ERR031017/ERR031017_2.fastq.gz

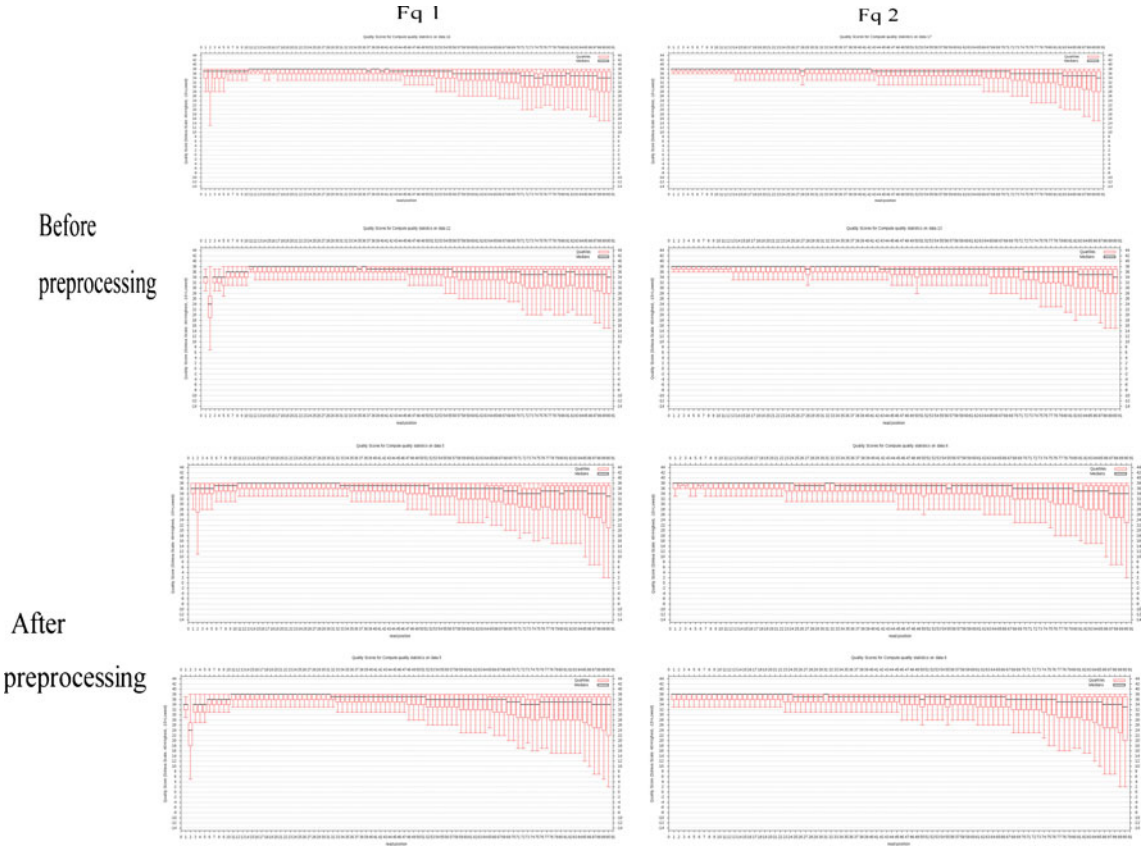

Fig. 1 The distribution of bases quality about before and after processing map

many as we can. After the assemble result came out, we merged the "neighbor" transcripts as method session commented, and got the merging result of all transcripts. For example, if transcript A in sample 10 N is overlapped with transcript B in sample 10 T, we merged them for the convenient comparing. Finally, samples 10 N and 10 T get about 400,000 and 230,000 transcripts, respectively.

FPKM Distribution

To profiling the expression level of each transcript, we calculated an average fragments per kilo base of transcript per million fragments mapped (FPKM). According the FPKM calculation foundation described before, we got the FPKM value of all transcripts. Figure 2 is the density distribution mapping of the FPKM of every transcript (Fig. 2). As we can see, 10 T samples have higher FPKM value than 10 N samples. It seems that cancer samples are always of greater expression level than the normal samples. 10 T samples have two peak value of FPKM distribution. The first peak in 0.7–0.8 log10(FPKM), which cannot find in samples 10 N. The second peak is shared with two samples in almost 0 value. Figure 3 is the box plot of the FPKM of the all transcripts of two samples (Fig. 3). In this picture, we can understand the distribution much better. Samples 10 N have median value under 0 log10 (FPKM), and have no outstanding outliers. But in samples 10 T, the median value is increased upon 0, and has many outstanding outliers. To further analysis those outlier transcripts, we tried to find the boundary to distinguish differential transcripts.

Significant Transcripts

By calculating the *p*-value and fold change with FPKM between two samples, we got all differential level of all related transcripts. Figure 4 is the volcano picture, which

Table 2 The statistics result of reads about before and after processing

Sample	Before preprocessing reads number	After preprocessing reads number	Reads length
10 N	34536162*2	31500516*2	90 bp
10 T	34007787*2	30925707*2	90 bp

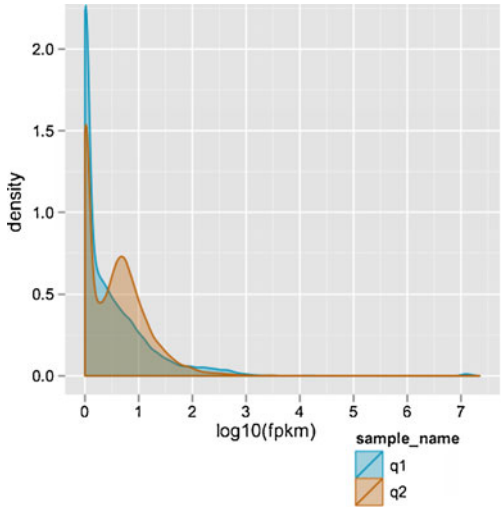

Fig. 2 The density distribution mapping of FPKM(Q1:10 N Q2:10 T)

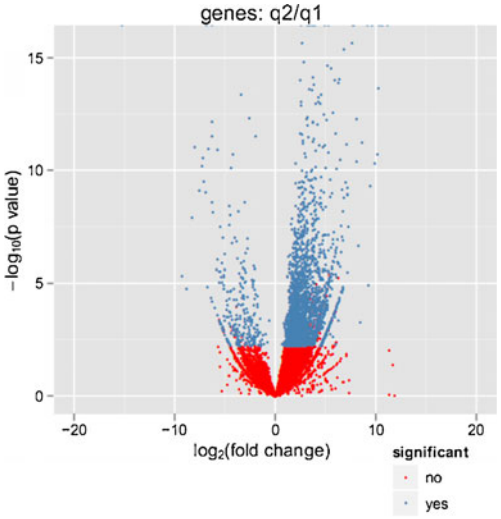

Fig. 4 Volcano picture of two samples

reflects the different situation of related transcripts between two samples (Fig. 4).

According to the information of Fig. 4 showed, we set the following boundary to distinguish differential transcriptions:

1) FPKM is more than three in both of two samples
2) |log2(fold_change)|>2;
3) *P*-value<0.006.

According to the above conditions, we got 197 significant transcripts (supplement), and there are 17 transcripts are non-coding transcripts. See Tables 3 and 4.

New lncRNA Discovery

To deeply analysis the other non-coding region, we focused on the long non coding RNA. We selected the assembling transcripts with over 200 bp length long, and located them on all human genes. The assembling transcripts cannot located in any of human genes are what we called lncRNA. Finally, we found that 36 lncRNAs are significant differential lncRNA shown in Table 5.

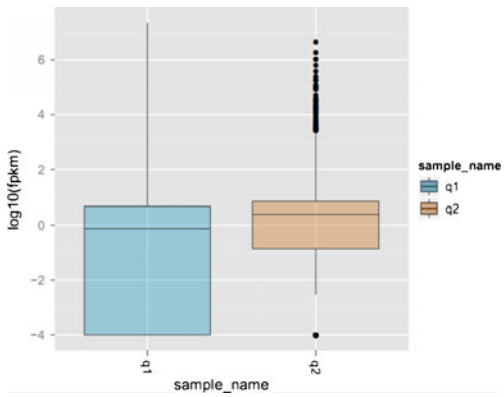

Fig. 3 Boxplot of FPKM in two samples (Q1:10 N Q2:10 T)

Discussion

Differential Coding Transcripts

As we can see in Table 4, the most differential gene is TFF3-Trefoil factor 3, which was more than 7 fold change from prostate cancer tissue to normal tissue. Some cDNA expression array analysis reveals that TFF3 may over express in prostate cancer patients. Recently, many studies have reported the strong relationship between gene TFF3 and prostate cancer. In 2004, immunohistochemistry was performed on a prostate cancer tissue microarray containing tumor tissue samples from 246 primary radical retro pubic prostatectomy cases with antibodies specific for TFF3, and Reiter's team ensured that the up-expressed situation of TFF3 were found in those tumor sample [18]. Then, in 2008, Arul's team announced that they have processed qPCR on seven prostate cancer biomarker, and found that TFF3 was a biomarker truly [19]. Now, our project has confirmed it. What all we human should do is developing the diagnosis kit for prostate early detecting. And interesting, we found the gene TFF1 was also in our Top 10 differential genes. But in our list, TFF1 has an opposite trend with TFF3, down-expressed in prostate cancer patients. In the many former study, most of them said that TFF1 (ps2 protein) was an up-expressed gene in prostate tumor. The family trefoil factor, included TFF1, TFF2, TFF3, are all over-expressed in prostate tumor, and the genes in this family are so differentially expressed in plasma levels in patients with advanced prostate cancer [20]. But shahid collected 95 malignant

Table 3 Top 10 differential transcripts

Gene	Value_1	Value_2	|log2 (fold_change)|
TFF3:NM_003226	3.86646	775.093	7.64721
SCGB3A1:NM_052863	135.952	3.75705	5.17735
NUDT8:NM_001243750	14.8653	448.795	4.91604
PSCA:NM_005672	211.443	9.64236	4.45474
CRABP2:NM_001199723	99.2394	4.58465	4.43603
RPS28:NM_001031	223.819	11.1042	4.33315
ANPEP:NM_001150	20.6093	348.262	4.07881
TFF1:NM_003225	543.156	33.8014	4.00621
HPN:NM_002151,NM_182983	5.23802	83.6546	3.99735
ELK4:NM_001973,NM_021795	3.81245	60.5403	3.98911

prostatic specimens from primary adenocarcinoma, performed immunohistochemical staining, he found that there was no significant correlation between TFF1 expression and the stage of disease, but TFF1 expression in prostate cancer significantly correlates with histological grade and the neuroendocrine differentiation [21]. So, although the TFF1 trend in our analysis is opposite with some other studies, this study reveals us that TFF1 can be a biomarker, but only for some stage of prostate cancer. Because TFF1 maybe reflects a contradiction expression level in different prostate cancer stage.

Differential Non-Coding Genes

Why we concern about the non-coding genes? The non-coding genes are always some pseudogene, or some function-unknown open reading frame. Many of them cannot be related to disease, especially cancer. But if we found them differentially over-expressed, we can say that gene has a great potential to be related to in the disease, for example prostate cancer in our project. Among the 17 transcripts we found, only two of them are down-expressed. The most outstanding transcript is NR_022014, one transcript for gene

Table 4 Difference non-coding transcripts

Transcript	Chromosome	Start	End	Length	Value_1	Value_2	log2(foldchange)	*p*-value
NR_022014	chr15	45770497	45850625	80128	5.06864	70.9165	3.80645	1.62E-14
NR_046211	chr14	106109426	106331644	222218	389.782	35.9914	−3.43694	4.57E-14
NR_038342	chr4	79892901	80229953	337052	8.98933	137.257	3.93253	2.13E-10
NR_027180								
NR_029684	chr5	148786439	148812563	26124	3.78058	27.4959	2.86254	3.61E-09
NR_029686								
NR_027786	chr22	42896584	42978017	81433	9.66855	185.334	4.26069	5.73E-08
NR_033322	chr7	75039623	75115568	75945	3.55121	20.2116	2.5088	3.24E-07
NR_015422	chr1	246939311	246956050	16739	3.07995	15.6054	2.34107	6.17E-06
NR_024103	chr17	46800531	46806494	5963	53.1348	5.69143	−3.22279	6.68E-06
NR_026811								
NR_033936	chr15	83130032	83182930	52898	3.55694	15.5705	2.13011	4.95E-05
NR_024448	chr22	23980672	24059610	78938	3.3211	30.6129	3.20441	6.53E-05
NR_002809	chr12	122233171	122241390	8219	3.7568	28.7025	2.9336	7.52E-05
NR_033874	chr4	102268933	102270040	1107	23.4628	140.186	2.57889	0.00025
NR_033853	chr11	58695101	58825925	130824	6.8126	30.7738	2.17543	0.00028
NR_028272	chr11	65190268	65194003	3735	14.4089	102.226	2.82673	0.00060
NR_024054								
NR_029426	chr5	69423288	69586004	162716	6.86276	27.6591	2.0109	0.00082
NR_036447	chr16	16411465	16444465	33000	5.04424	21.8833	2.11712	0.00202
NR_033968	chr5	70503779	70555122	51343	3.86906	18.529	2.25973	0.00281

Table 5 Significant lncRNA

Gene_ID	Locus	Q1 FPKM	Q2 FPKM	log2 (fold_change)	*p*-value
XLOC_017565	chr2:96502828-96593024	3.45934	21.0665	2.60639	2.54E-12
XLOC_020722	chr22:16226489-16231476	4.27748	121.458	4.82756	1.01E-09
XLOC_029109	chr7:100609801-100611622	105.792	4.28654	−4.62527	3.81E-09
XLOC_009482	chr14:19612666-19616445	9.09186	154.358	4.08556	7.28E-08
XLOC_016246	chr19:16011127-16011911	17.8498	314.085	4.13718	7.80E-08
XLOC_011528	chr16:74423901-74426024	28.5411	202.088	2.82387	8.07E-08
XLOC_024017	chr4:80748355-80799530	6.97652	101.441	3.862	2.29E-07
XLOC_017554	chr2:89156672-89161069	1629.39	376.649	−2.11303	1.74E-06
XLOC_009483	chr14:19631005-19631820	3.15919	64.6857	4.35582	4.66E-06
XLOC_013771	chr17:7967573-7971851	5.82135	57.5181	3.30459	4.97E-06
XLOC_020229	chr22:23241800-23243587	1074.74	123.706	−3.11901	3.61E-05
XLOC_024018	chr4:80802969-80804296	9.66583	76.6322	2.98699	8.05E-05
XLOC_022634	chr3:131962287-131963333	54.5552	6.39125	−3.09355	0.000144
XLOC_013949	chr17:46811997-46821760	68.7657	6.23488	−3.46325	0.000297
XLOC_009794	chr14:103674935-103675179	4.06342	74.0403	4.18755	0.000776
XLOC_009480	chr14:19605935-19606209	56.2565	399.723	2.82891	0.000839
XLOC_032885	chr9:32946905-32946998	1.99E+06	118051	−4.07376	0.001102
XLOC_002954	chr1:160864686-160866290	21.0322	93.5645	2.15336	0.001893
XLOC_020051	chr21:29780073-29791593	6.15813	27.418	2.15456	0.001952
XLOC_031343	chr8:1920327-1920524	15.9728	538.328	5.07479	0.002031
XLOC_018034	chr2:8992758-8994585	3.79683	21.1956	2.4809	0.00232
XLOC_020135	chr21:42653619-42654457	48.5836	8.46814	−2.52035	0.002354
XLOC_030098	chr7:63965435-63967371	4.09327	21.8106	2.41371	0.002564
XLOC_013816	chr17:16520352-16520763	15.194	98.4542	2.69595	0.002768
XLOC_030062	chr7:42896688-42897634	25.2703	3.75698	−2.7498	0.00299
XLOC_003179	chr1:224133263-224218601	4.76378	24.3255	2.35229	0.003358
XLOC_004375	chr10:76849265-76849359	1.04E+06	38595.4	−4.74549	0.0036
XLOC_007770	chr12:53126359-53126453	1.04E+06	38595.4	−4.74549	0.0036
XLOC_007939	chr12:95247417-95248603	3.12413	19.403	2.63476	0.003945
XLOC_019632	chr20:62258515-62260109	4.50218	22.9355	2.34889	0.004009
XLOC_022442	chr3:106561532-106639852	3.18728	40.0901	3.65285	0.004303
XLOC_033208	chr9:140445512-140446124	10.6154	56.9655	2.42393	0.004931
XLOC_023830	chr4:8356425-8358602	5.21159	23.0769	2.14665	0.005253
XLOC_009564	chr14:38377517-38378622	4.53556	24.2341	2.41769	0.005641
XLOC_004217	chr10:20011553-20011812	37.6673	3.28617	−3.51883	0.005953
XLOC_029689	chr7:102135797-102136645	458.542	38.452	−3.57593	0.005963

C15orf21. We detected this gene is 3 fold up change in prostate cancer with P=1.62E-14, fitted the result of a former study by Arul in 2007 [22]. In his result, C15orf21 showed over-expressed in prostate cancer with significance *p*-value in prostate cancer with P=3.4*10E-6, which be confirmed by our project.

New lncRNA Discovery

Large intergenic non-coding RNAs (lincRNAs) are emerging as key regulators of diverse cellular processes. Determining the function of individual lincRNAs remains a challenge. In 2011, John Rinn from Broad Institute used RNA-seq to produce the most complement catalogue of lincRNA [23] crossing 24 tissues, included prostate cancer tissue. So, in this catalogue, we can find their result of prostate cancer related lncRNA. As shown in Table 5, red highlight part represents the lncRNAs related with prostate cancer has been published, 9 lncRNAs were found according our method; 3 blue highlight lncRNAs have been published but don't find the relationship with prostate cancer, other 24 lncRNAs are significant in this project. So,

there is a huge possibility that the 24 lncRNAs are related with the prostate cancer.

Interesting

When we queried these lncRNA regions on UCSC to get the average conservation score of each candidate or putative lncRNA, most of them are reflecting a very low score. We image that lncRNA are not "rubbish" any more, so they should be conservative across mammal. But why they are always so low conservational score? Can it explain us that, lncRNA are not so conservative and change acutely across mammal? All these questions are waiting to be explored.

Conflict of Interest The authors have no conflict of interest to declare.

References

1. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C (2011) The genomic complexity of primary human prostate cancer. Nature 470(7333):214–220
2. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, White TA, Stojanov P, Van Allen E, Stransky N (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat Genet 44(6):685–689
3. Grasso CS, Wu YM, Robinson DR, Cao X, Dhanasekaran SM, Khan AP, Quist MJ, Jing X, Lonigro RJ, Brenner JC (2012) The mutational landscape of lethal castration-resistant prostate cancer. Nature 487(7406):239–243
4. Morin RD, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh TJ, McDonald H, Varhol R, Jones SJM, Marra MA (2008) Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45(1):81–94
5. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM (2009) Transcriptome sequencing to detect gene fusions in cancer. Nature 458(7234):97–101
6. Pflueger D, Terry S, Sboner A, Habegger L, Esgueva R, Lin PC, Svensson MA, Kitabayashi N, Moss BJ, MacDonald TY (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Genome Res 21(1):56–67
7. Sahu A, Iyer MK, Chinnaiyan AM (2012) Insights into Chinese prostate cancer with RNA-seq. Cell Res 22(5):786–788
8. Mattick JS (2001) Non-coding RNAs: the architects of eukaryotic complexity. EMBO reports 2(11):986–991
9. Stefani G, Slack FJ (2008) Small non-coding RNAs in animal development. Nat Rev Mol Cell Biol 9(3):219–230
10. Mercer TR, Dinger ME, Mattick JS (2009) Long non-coding RNAs: insights into functions. Nat Rev Genet 10(3):155–159
11. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL (2010) Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature 464(7291):1071–1076
12. Ren S, Peng Z, Mao JH, Yu Y, Yin C, Gao X, Cui Z, Zhang J, Yi K, Xu W (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res 22(5):806–821
13. Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R (2011) The European nucleotide archive. Nucleic Acids Res 39(suppl 1):D28–D31
14. Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25(9):1105–1111
15. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7(3):562–578
16. Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-Seq. Bioinformatics 25(8):1026–1032
17. Toung JM, Morley M, Li M, Cheung VG (2011) RNA-sequence analysis of human B-cells. Genome Res 21(6):991–998
18. Garraway IP, Seligson D, Said J, Horvath S, Reiter RE (2004) Trefoil factor 3 is overexpressed in human prostate cancer. Prostate 61(3):209–214
19. Laxman B, Morris DS, Yu J, Siddiqui J, Cao J, Mehra R, Lonigro RJ, Tsodikov A, Wei JT, Tomlins SA (2008) A first-generation multiplex biomarker analysis of urine for the early detection of prostate cancer. Cancer Res 68(3):645
20. Vestergaard EM, Borre M, Poulsen SS, Nexø E, Tørring N (2006) Plasma levels of trefoil factors are increased in patients with advanced prostate cancer. Clin Cancer Res 12(3):807–812
21. Ather MH, Abbas F, Faruqui N, Israr M, Pervez S (2004) Expression of pS2 in prostate cancer correlates with grade and Chromogranin A expression but not with stage. BMC Urol 4(1):14
22. Tomlins SA, Laxman B, Dhanasekaran SM, Helgeson BE, Cao X, Morris DS, Menon A, Jing X, Cao Q, Han B (2007) Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer. Nature 448(7153):595–599
23. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25(18):1915–1927