RESEARCH ARTICLE

Identification of Prostate Cancer LncRNAs by RNA-Seq

Cheng-Cheng Hu[1], Ping Gan[1,2], Rui-Ying Zhang[1], Jin-Xia Xue[1], Long-Ke Ran[3]*

Abstract

    <u>Purpose</u>: To identify prostate cancer lncRNAs using a pipeline proposed in this study, which is applicable for the identification of lncRNAs that are differentially expressed in prostate cancer tissues but have a negligible potential to encode proteins. <u>Materials and Methods</u>: We used two publicly available RNA-Seq datasets from normal prostate tissue and prostate cancer. Putative lncRNAs were predicted using the biological technology, then specific lncRNAs of prostate cancer were found by differential expression analysis and co-expression network was constructed by the weighted gene co-expression network analysis. <u>Results</u>: A total of 1,080 lncRNA transcripts were obtained in the RNA-Seq datasets. Three genes (PCA3, C20orf166-AS1 and RP11-267A15.1) showed a significant differential expression in the prostate cancer tissues, and were thus identified as prostate cancer specific lncRNAs. Brown and black modules had significant negative and positive correlations with prostate cancer, respectively. <u>Conclusions</u>: The pipeline proposed in this study is useful for the prediction of prostate cancer specific lncRNAs. Three genes (PCA3, C20orf166-AS1, and RP11-267A15.1) were identified to have a significant differential expression in prostate cancer tissues. However, there have been no published studies to demonstrate the specificity of RP11-267A15.1 in prostate cancer tissues. Thus, the results of this study can provide a new theoretic insight into the identification of prostate cancer specific genes.

Keywords: Long non-coding RNA - RNA-Seq - prostate cancer - bioinformatics

Introduction

    The ENCODE project has recently reported that over 90% of nucleotides in the human genome can be transcribed into RNAs (Birney et al., 2007). However, it is estimated that only a small fraction (about 2%) of these RNAs will be translated into proteins, while the vast majority (98%) of them are non-coding RNAs (ncRNA) without apparent protein-coding capacity. Long non-coding RNAs (lncRNAs) are in general considered as long (>200 nt) transcripts that lack long open reading frames (ORF) and have very low sequence conservation, but otherwise have mRNA-like properties such as poly (A) tails and promoter after splicing (Pauli et al., 2011), thus making it computationally difficult to distinguish them from other genome sequences.

    LncRNAs were once considered to be the transcriptional noise resulting from stochastic transcription, a byproduct of RNA polymerase II during the synthesis of functional RNAs, and they appeared to be more lowly expressed than protein-coding genes (Ponting et al., 2009; Cabili et al., 2011). However, it has become increasingly clear that lncRNAs are involved in regulating a wide variety of important cellular functions, such as gene expression, genome imprinting, recruitment of chromatin modifying machinery, and regulation of X chromosome inactivation (Mercer et al., 2009; Hung et al., 2010; Lee et al., 2012). For instance, lncRNAs such as *Kcnq1ot1* and *Air*, which map to the *Kcnq1* and *Igf2r* imprinted gene clusters respectively, mediate the transcriptional silencing of multiple genes by recruiting the chromatin modifying machinery (Kanduri et al., 2008; Nagano et al., 2008; Korostowski et al., 2012). The X inactive-specific transcript (*Xist*) LncRNA has also been shown to be related to the inactivation of one of the two X chromosomes in female cells (Guttman et al., 2009; Engreitz et al., 2013). The recognition of the important role of LncRNAs has sparked a growing interest in the study of their functionality (Ponjavic et al., 2007; Guttman et al., 2009; Orom et al., 2010; Qiao et al., 2013; Liu et al., 2014).

    RNA Sequencing (RNA-Seq), also called whole transcriptome shotgun sequencing or WTSS, refers to the use of high-throughput sequencing technologies for transcriptome analysis, which allows the sensitive identification of lowly expressed transcripts and is independent of currently available gene annotations (Marioni et al., 2008; Wang et al., 2009). This property makes RNA-Seq a highly desirable method for the detection of novel transcripts including lncRNAs. As such, RNA-Seq has been successfully applied to identify thousands of lncRNAs in human, mouse and various other species (Li et al., 2012; Lv et al., 2013; Zhang et al., 2013; Ye et al., 2014).

    LncRNAs have attracted increasing scientific attention

*[1]Laboratory of Biomedical Engineering, [2]Laboratory of Medical Physics, [3]Department of Bioinformatics, Chongqing Medical University, Chongqing, China *For correspondence: ranlk@sina.com*

over the past few years. A considerable number of studies have shown that there is a significant change in the expression of some lncRNAs in human tumor cells, thus lncRNAs have the potential to be used as a useful diagnosis marker and treatment targets for cancers. The main purpose of this study is to identify prostate cancer-specific lncRNAs using two publicly available RNA-Seq datasets form normal prostate tissues and prostate cancer tissues. First, lncRNAs with a length of <200 nt, the ORF length of >300 nt and single exon were filtered out; Then, we scored the coding potential of all transcripts using the phylogenetic codon substitution frequency (PhyloCSF) (Lin et al., 2011), and searched the Pfam protein families database (Finn et al., 2008). After stringent filtering out of the putative protein-coding potential transcripts, a confident set of 1080 lncRNA transcripts were obtained; At last, prostate cancer specific gene co-expression networks were constructed by the differential expression (DE) analysis and weighted gene co-expression network analysis (WGCNA), and we identified three lncRNAs that showed a significant differential expression in prostate cancer tissues, including prostate cancer antigen (PCA3), C20orf166-AS1 and RP11-267A15.1. The results also showed that the brown and black gene modules had a significant negative and positive correlation with prostate cancer, respectively.

Materials and Methods

Preparation of the prostate cancer RNA-Seq datasets

We used two publicly available RNA-Seq datasets from normal prostate tissues and prostate cancer tissues, each consisting of 5 samples. The datasets were downloaded from the NCBI GEO database with the accession number of GSE22260, and sequenced by using paired-end sequencing on an Illumina Genome Analyzer II (GAII). About 10 million to 16 million of reads were obtained in each sample, resulting in a total of about 300 million of reads with an average length of 36 bp.

Prediction of prostate cancer lncRNAs

In this study, transcripts were reconstructed using genome guided methods. An integrative pipeline to map, reconstruct, and determine the coding potential of lincRNAs was proposed in this study, which was illustrated schematically in Figure 1 (Martin et al., 2011): First, Tophat was used to map the RNA-Seq reads in each sample to the human GRCh37 reference genome (Trapnell et al., 2009), and 76.9% of the total RNA-Seq reads in each sample have been successfully mapped to the reference genome. Then, Cufflinks was used to assemble these aligned reads into transcripts based on the known gene annotation (Trapnell et al., 2010), and the assembled transcripts were annotated and grouped into different categories using the Cuffcompare program from the Cufflinks package. Small ncRNAs were filtered out using a length threshold of 200 nt, and putative protein-coding RNAs were also filtered out using an ORF length threshold of 300 nt. Thus, transcripts with a length of > 200 nt and ORF length of < 300 nt were retained as candidate lncRNAs, and then further screened

using codon substitution frequencies (CSF) to distinguish protein-coding genes from non-coding genes. Transcripts with a PhyloCSF score <0 were considered as non-coding genes. At last, transcripts that encoded any of the protein domains cataloged in the Pfam database were excluded, and transcripts with gene significance of 1 was filtered out.

Differential expression analysis

The central goal of differential expression analysis is to identify genes that change in abundance between different experimental conditions. In this study, we used Cuffdiff (a module in CuffLinks package), to detect differentially expressed lncRNAs between the normal prostate tissues and prostate cancer tissues (Trapnell et al., 2012). Those genes differentially expressed in prostate cancer tissues might be prostate cancer specific gene.

Gene co-expression network analysis

Gene co-expression network analysis has proven useful in detecting clusters (referred to as modules) of co-expressed genes, and investigating the relationship between gene networks and phenotypic traits of interest to the researchers at the transcriptome level. In this study, the weighted gene co-expression network analysis (WGCNA) was used to investigate the association between the predicted lncRNA modules and prostate cancer (Langfelder et al., 2008).

Results

Reconstruction of transcriptomes

Current transcriptome assembly strategies fall into two broad categories depending on whether or not a reference genome is available: genome-guided and genome-independent de novo assembly (Garber et al., 2011). As the de novo assembly depends critically on sequencing depth and length, the genome-guided assembly was used in this study for the reconstruction of transcriptomes. Tophat was used to map the RNA-Seq reads to human GRCh37 reference genome, and about 76.9% (220 hundred million out of a total of 290 hundred million) of all reads were successfully mapped to the reference genome. Cufflinks was subsequently used to assemble these aligned reads into transcripts. However, a potential problem of the genome-guided assembly used in this study was that the assembled transcripts contained not only the annotated
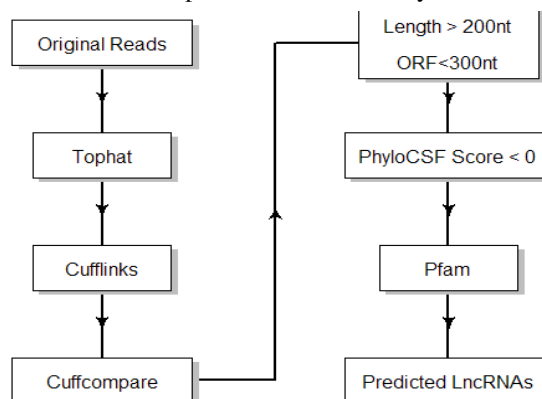

Figure 1. Pipeline for the Identification of lncRNAs Based on RNA-Seq

genes, but also novel genes. These assembled transcripts were annotated and grouped into different categories using the Cuffcompare program, which yielded a total of 157,926 transcripts.

Prediction results of lncRNA

It is clear that lncRNAs have two distinct characteristics. First, lncRNAs, by definition, have a length of >200nt; Second, the ORF length of protein-coding mRNAs is generally greater than 300 bases or 100 amino acids, thus those genes whose ORF length is smaller than 300 bases have a very small probability to code proteins. Accordingly in this study, lncRNAs with a length of <200nt and the ORF length of >300nt were filtered out. In addition, to obtain a reliable lncRNA dataset, the transcripts with only one exon were also filtered out. At last, a stringent set of 6941 candidate lncRNAs were obtained.

We used the PhyloCSF program to score the protein-coding potential of the candidate transcripts, and reported the resulting log-likelihood ratios in units of decibans that quantified the probability of transcript to be present in the protein-coding or non-coding modules. Transcripts with a PhyloCSF score of 0 were considered to have an equal probability of coding proteins, thus the PhyloCSF threshold was set to be less than 0 in this study. As a result, a total of 1776 transcripts were retained. Then we searched the Pfam database to identify transcripts that coded any of the protein domains cataloged in the Pfam database, and transcripts with gene significance of 1 was filtered out. By now, a total of 1080 long non-coding transcripts were obtained in the prostate cancer RNA-Seq dataset.

Previous studies have indicated that on average lncRNAs had smaller length and fewer exons than that of protein coding transcripts in mammalian cells (2.9 exons and a transcript length of 1 kb for lincRNAs *vs* 10.7 exons and a transcript length of 2.9 kb for protein-coding transcripts) (Guttman et al., 2010; Cabili et al., 2011). To examine whether all the lncRNAs predicted as described previously had the same characteristics, these lncRNAs were compared with 39906 protein-coding transcripts and 9656 non-coding transcripts in the RefSeq database, and the results were shown in Figure 2. It clearly showed that the average length of the predicted lncRNA transcripts was about one third of that of protein-coding transcripts (1.0
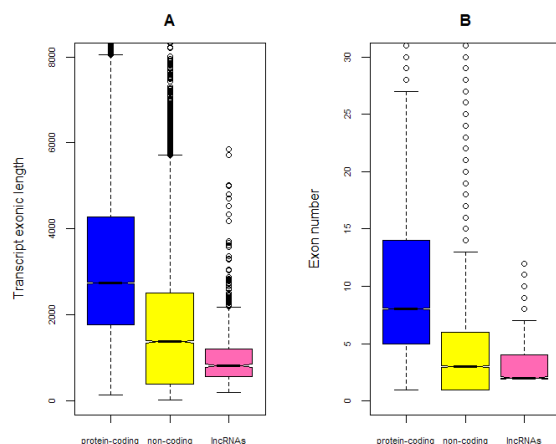
kb *vs* 3.4 kb), as shown in Figure 2A. In addition, they also had fewer exons (2.0 *vs* 6.3), as shown in Figure 2B. Thus, the results of this study agreed well with previous studies on the lncRNAs.

Differential expression

Cuffdiff was used to detect differentially expressed lncRNAs between the normal prostate tissues and prostate cancer tissues (Trapnell et al., 2010), where the false discovery rate (FDR) was set to be 0.05. Figure 3 showed that there were 12 genes that had significant differential expression in the prostate cancer tissues. However, due to the possibility of false positive identifications of differentially expressed genes, it is possible that not all of the 12 genes that had significant differential expression were the lncRNAs. Further analysis showed that of the 12 genes, only PCA3, C20orf166-AS1, RP11-267A15.1 were prostate cancer specific lncRNAs.

Figure 4 showed that PCA3 and RP11-267A15.1 showed a highly specific expression in the prostate cancer cells, but low expression in normal prostate cells. And C20orf166-AS1PCA3 showed contrary result. PCA3 gene has been identified as a prostate specific gene highly over-expressed in prostate cancer by several other methods, such as Northern blot (Bussemakers et al., 1999) and real time quantitative reverse transcription-polymerase chain reaction (RT-PCR) (Landers et al., 2005). It has also proven to be a potential target for the treatment of prostate cancer. C20orf166-AS1 is a large intergenic noncoding RNA (lincRNA) that shows a high expression in normal prostate tissues, whereas relatively low expression in the prostate cancer tissues. C20orf166-AS1 has also been found to be significantly associated with prostate cancer or prostatitis using other methods (Kimura et al., 2006; Eeles et al., 2013). RP11-267A15.1 is an antisense RNA that also shows a highly specific expression in the prostate
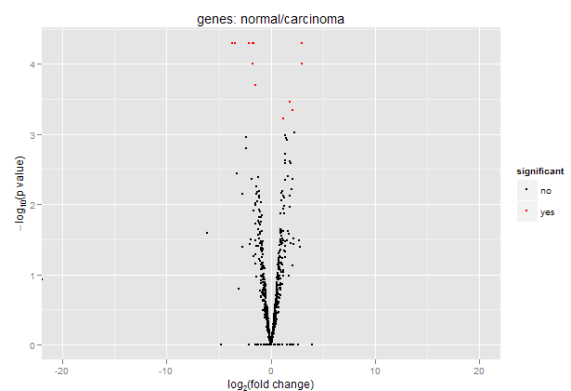
Figure 3. LncRNAs with Significant Differential Expression between Prostate Cancer Tissues and Normal Prostate Tissues
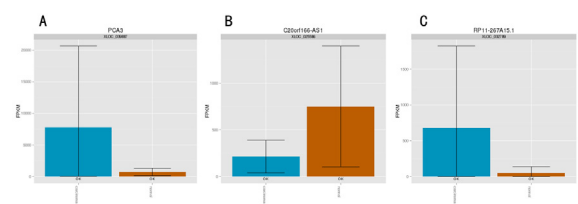
Figure 2. Comparisons of Transcript Length and Exon Number

Figure 4. The FPKM of PCA3, C20orf166-AS1 and RP11-267A15.1 between the Two Groups of Samples

Table 1. Correlation Coefficients between Module Eigengenes and Disease Status

Module	Turquoise	Magenta	Brown	Green	Yellow	Blue	Purple	Pink	Black	Red	Grey
Coefficient	-0.3	-0.37	-0.92	-0.63	-0.41	0.3	0.4	0.47	0.7	0.21	-0.0097
P value	0.39	0.29	0.00019	0.051	0.24	0.39	0.25	0.17	0.025	0.56	0.98
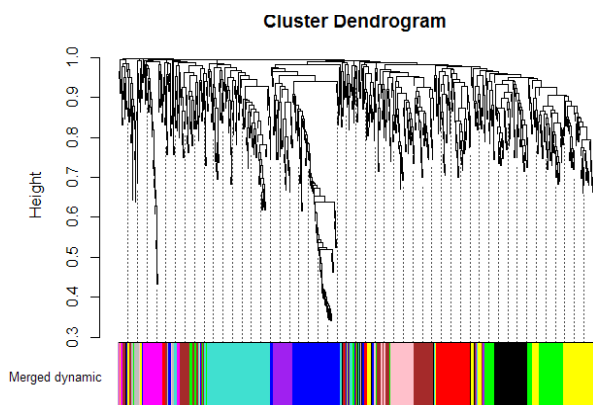
Figure 5. Cluster Dendrogram

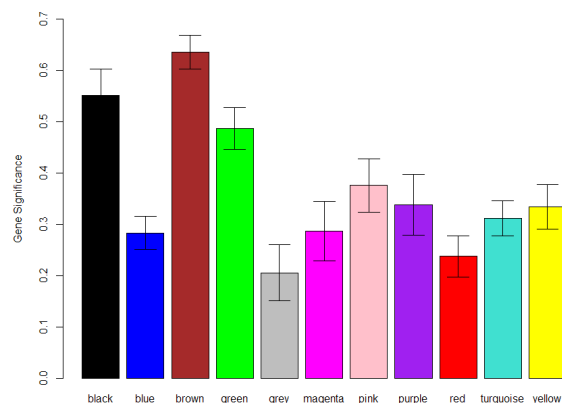**Figure 6. The Average Gene Significance of Each Module**

cancer cells, but low or no expression in normal prostate tissues. However, to the best of our knowledge, there have been no published studies to demonstrate the specificity of RP11-267A15.1 in prostate cancer tissues. Thus, the results of this study can provide a new theoretic insight into the identification of prostate cancer specific genes.

Gene co-expression network analysis

The lincRNA genes were hierarchically clustered into groups with respect to their expression profiles, and the modules in the resulting dendrogram were identified using the Dynamic Tree Cut method, whereby the highly interconnected genes were assigned to the same module. In this study, the association between lncRNA gene modules and prostate cancer was investigated using the WGCNA method, and the minimum number of genes in each module was set to be 30. The resulting cluster dendrogram was shown in Figure 5, where each branch of the dendrogram represented a gene, while the color-coded module membership was displayed in the color bars below the dendrograms using the Dynamic Tree Cut method. In this study, the lncRNAs were assigned to 11 modules.

The correlation coefficients between the module eigengenes and the disease status of prostate cancer patients were calculated, and the results were shown in Table 1. The results clearly showed that the brown module had a significant negative correlation with prostate cancer (r=-0.92, $p<0.01$), indicating that the genes in this module appeared to have an inhibitory effect on prostate cancer. In addition, the black module had a significant positive correlation with prostate cancer (r=0.70, $p<0.05$), indicating that the genes in this module had a promoting effect on prostate cancer.

Module significance can also be used to indicate the relationship between the modules and prostate cancer. The average gene significance of each module was presented in the bar graph in Figure 6. It showed that the brown module had the maximum average gene significance, followed by the black module, further indicating that these two modules had a significant correlation with prostate

cancer. Further analysis of the differentially expressed genes showed that C20orf166-AS1 and RP11-267A15.1 were present in the brown module, further indicating that these two lncRNAs had an inhibitory effect on the prostate cancer.

Discussion

The advent of RNA-seq technologies has greatly facilitated our understanding of the transcription of human and animal genome. As comparison with chromosome marker analysis, RNA-seq provides a more direct approach to transcriptome profiling, and the major advantage of this approach is that it is an open biological technology that makes it possible to reconstruct the whole transcriptions. At present, high-throughput sequencing followed by bioinformatic analysis has became the mainstream method for the prediction and screening of lncRNAs. In this study, we used the paired-end reads and genome guided method to reconstruct the transcripts, which could help to improve the accuracy of the reconstruction of transcriptome. In addition, we also used PhyloCSF and Pfam database to identify the coding potential of the putative lncRNA transcripts. However, there are many other methods besides the ones used in this study to improve the prediction accuracy of lncRNAs. For instance, the original low-quality reads can be filtered out or truncated, or more filtering methods, such as the Coding Potential Calculator (CPC) (Kong et al., 2007) and the Conserved Sequence Tags CST miner (Furuno et al., 2003), could also be used to obtain a more stringent result in the prediction process.

In this study, a pipeline is designed for the identification of prostate cancer specific lncRNAs that are differentially expressed in prostate cancer tissues but have a negligible potential to encode proteins. A total of 1080 lncRNAs were obtained in the RNA-Seq datasets from normal prostate tissues and prostate cancer tissues. The comparison between these lncRNA transcripts and the NCBI RefSeq database, a high-quality non-redundant protein database,

showed that the average length of the lncRNA transcripts was about one third of that of protein coding transcripts (1.0 kb *vs* 3.4 kb), and they also had fewer exons (2.0 *vs* 6.3). It is well known that lncRNAs are characterized by small transcript length and small number of exons, thus the results indicate that the pipeline proposed in this study is useful for the prediction of prostate cancer specific lncRNAs.

Three genes (PCA3, C20orf166-AS1 and RP11-267A15.1) were identified as prostate cancer specific lncRNAs in this study. PCA3 has proven to be a potential target for the treatment of prostate cancer, and C20orf166-AS1 has also been suggested to be significantly associated with prostate cancer or prostatitis. However, to the best of our knowledge, there have been no published studies to demonstrate the specificity of RP11-267A15.1 in prostate cancer tissues. Thus the results of this study can provide a new theoretic insight into the identification of prostate cancer specific genes. In addition, the association analysis showed that C20orf166-AS1 and RP11-267A15.1 were present in the brown module and had a significant negative correlation with prostate cancer, indicating that these two lncRNA genes had an inhibitory effect on the prostate cancer.

To better understand the relationship between the gene or gene modules involved in this study and the prostate cancer, more studies are needed to use bioinformatic tools to analyze the functional annotation and gene functions, and to further investigate their biological significance. In addition, the real-time fluorescent quantitation RT-PCR can also be used to verify these lncRNAs with significant differential expression in the prostate cancer tissues.

References

Birney E, Stamatoyannopoulos JA, Dutta A, et al (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799-16.

Bussemakers MJ, Van Bokhoven A, Verhaegh GW, et al (1999). DD3: a new prostate-specific gene highly overexpressed in prostate cancer. *Cancer Res*, **59**, 5975-9.

Cabili MN, Trapnell C, Goff L, et al (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, 25, 1915-27.

Eeles RA, Olama AA, Benlloch S, et al (2013). Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet*, **45**, 385-91.

Engreitz JM, Pandya-Jones A, McDonel P, et al (2013). The *Xist* lncRNA Exploits Three-Dimensional Genome Architecture to Spread Across the X Chromosome. *Science*, **341**, 1237973.

Finn RD, Tate J, Mistry J, et al (2008). The Pfam protein families database. *Nucleic Acids Res*, **36**, 281-8.

Furuno M, Kasukawa T, Saito R, et al (2003). CDS Annotation in Full-Length cDNA Sequence. *Genome Research*, **13**, 1478-87.

Garber M, Grabherr MG, Guttman M, et al (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Meth*, **8**, 469-77.

Guttman M, Amit I, Garber M, et al (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223-7.

Guttman M, Garber M, Levin JZ, Donaghey J, et al (2010). Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multiexonic structure of lincRNAs. *Nat Biotechnol*, **28**, 503-10.

Hung T, Chang HY (2010). Long noncoding RNA in genome regulation: prospects and mechanisms. *RNA Biol*, **7**, 582-5.

Kanduri C (2008). Functional insights into long antisense noncoding RNA *Kcnq1ot1* mediated bidirectional silencing. *RNA Biol*, **5**, 208-11.

Kimura K, Wakamatsu A, Suzuki Y, et al (2006). Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. *Genome Res*, **16**, 55-65.

Kong L, Zhang Y, Ye ZQ, et al (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Research*, **35**, 345-9.

Korostowski L, Sedlak N, Engel N (2012). The *Kcnq1ot1* long non-coding RNA affects chromatin conformation and expression of *Kcnq1*, but does not regulate its imprinting in the developing heart. *PLoS Genet*, **8**, 1002956.

Landers KA, Burger MJ, Tebay MA, et al (2005). Use of multiple biomarkers for amolecular diagnosis of prostate cancer. *Int J Cancer*, **114**, 950-6.

Langfelder P, Horvath S (2008). WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**, 559.

Lee JT (2012). Epigenetic regulation by long noncoding RNAs. *Science*, **338**, 1435-9.

Li T, Wang S, Wu R, Zhou X, Zhu D, Zhang Y (2012). Identification of long non-protein coding RNAs in chicken skeletal muscle using next generation sequencing. *Genomics*, **99**, 292-8.

Lin MF, Jungreis I, Kellis M (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, 275-82.

Liu JH, Chen G, Dang YW, Li CJ, Luo DZ (2014). Expression and prognostic significance of lncRNA MALAT1 in pancreatic cancer tissues. *Asian Pac J Cancer Prev*, **15**, 2971-7.

Lv J, Cui W, Liu H, et al (2013). Identification and characterization of long non-coding RNAs related to mouse embryonic brain development fromavailable transcriptomic data. *PLoS One*, **8**, 71152.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, **18**, 1509-17.

Martin JA, Wang Z (2011). Next-generation transcriptome assembly. *Nat Rev Genet*, **12**, 671-82.

Mercer TR, Dinger ME, Mattick JS (2009). Long non-coding RNAs: insights into functions. *Nat Rev Genet*, **10**, 155-9.

Nagano T, Mitchell JA, Sanz LA, et al (2008). The *Air* noncoding RNA epigenetically silences transcription by targeting G9a to chromatin. *Science*, **322**, 1717-20.

Orom UA, Derrien T, Beringer M, et al (2010). Long noncoding RNAs with enhancer-like function in human cells. **Cell**, **143**, 46-58.

Pauli A, Rinn JL, Schier AF (2011). Non-coding RNAs as regulators of embryogenesis. *Nat Rev Genet*, **12**, 136-49.

Ponjavic J, Ponting CP, Lunter G (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res*, **17**, 556-65.

Ponting CP, Oliver L, Reik W (2009). Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629-41.

Qiao HP, Gao WS, Huo JX, Yang ZS (2013). Long non-coding RNA GAS5 functions as a tumor suppressor in renal cell carcinoma. *Asian Pac J Cancer Prev*, **14**, 1077-82.

Trapnell C, Pachter L, Salzberg SL (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105-11.

Trapnell C, Roberts A, Goff L, et al (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc*, **7**, 562-78.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, **28**, 511-5.

Wang Z, Gerstein M, Snyder M (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57-63.

Ye N, Wang B, Quan ZF, et al (2014). Functional Roles of Long Non-coding RNA in Human Breast Cancer. *Asian Pac J Cancer Prev*, **15**, 5993-7.

Zhang Q, Geng PL, Yin P, et al (2013). Down-regulation of long non-coding RNA TUG1 inhibits osteosarcoma cell proliferation and promotes apoptosis. *Asian Pac J Cancer Prev*, **14**, 2311-5.