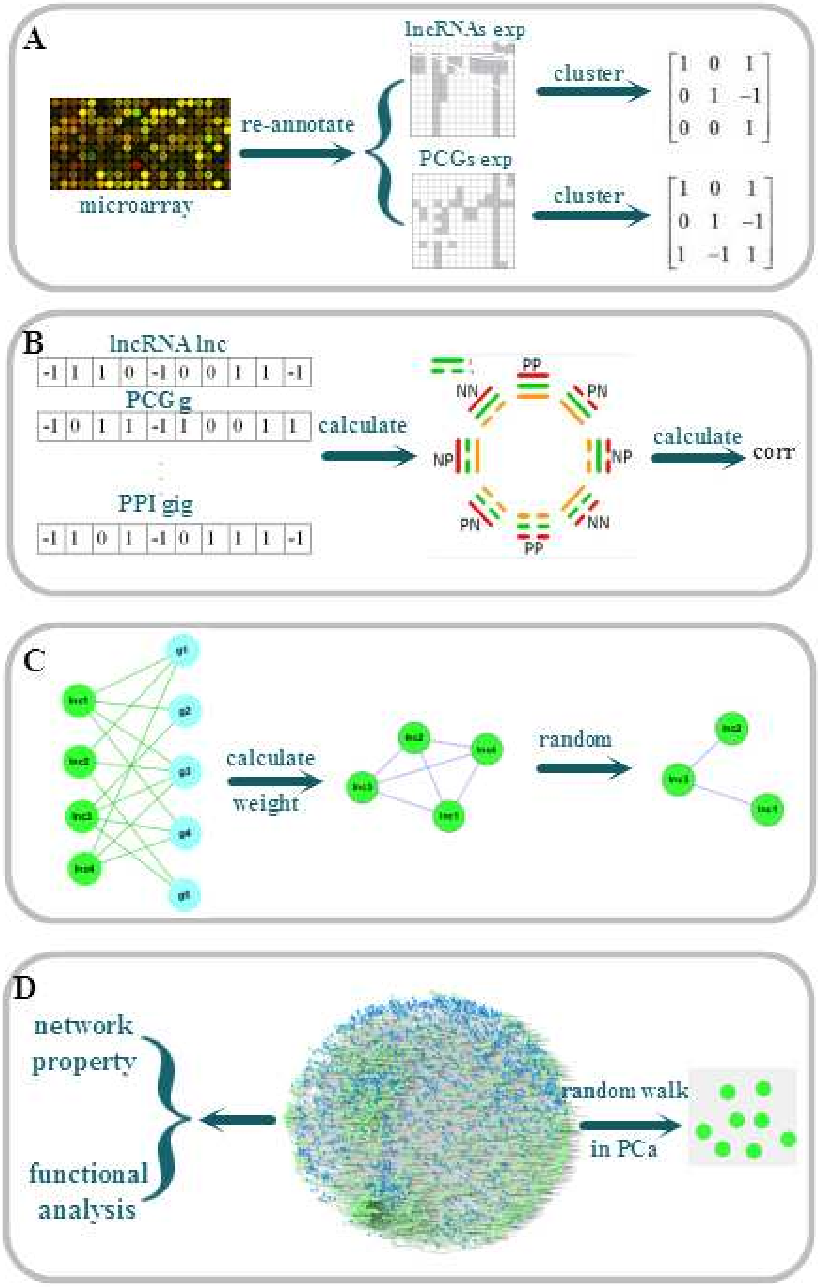

We constructed a lncRNA-PCG bipartite network by sample correlation and identified cancer-related lncRNAs.

Journal Name RSCPublishing

ARTICLE

Cite this: DOI: 10.1039/x0xx00000x

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Construction of lncRNA-PCG bipartite network and identification of cancer-related lncRNAs: A case study in prostate cancer

Rui Zhang [§a], Yongjing Liu [§a], Fujun Qiu [§a], Kening Li [a], Yuanshuai Zhou [a], Desi Shang [a], Yan Xu [*a]

LncRNAs are involved in a wide range of biological processes, such as chromatin remodeling, mRNA splicing, mRNA editing and translation. They can either upregulate or downregulate gene expression, and play key roles in the progression of various human cancers. However, the functional mechanisms of most lncRNAs still remain unknown at present. This paper aims to provide space for the understanding of lncRNAs by proposing a new method to obtain protein-coding genes (PCGs) regulated by lncRNAs, thus identify candidate cancer-related lncRNAs using bioinformatics approaches. This study presents a method based on sample correlation, which is applied to the expression profiles of lncRNAs and PCGs in prostate cancer in combination with protein interaction data to build a lncRNA-PCG bipartite network. Candidate cancer-related lncRNAs were extracted from the bipartite network by using random walk. 14 prostate cancer-related lncRNAs were acquired from LncRNADisease database and MNDR, in which 6 lncRNAs were present in our network. As one of the seed nodes, ENSG00000234741 achieved the highest score among them. Other two cancer-related lncRNAs (ENSG00000225937 and ENSG00000236830) were ranked within top 30. In addition, the top candidate lncRNA ENSG00000261777 shares an intron with DDX19, and interacts with IGF2 P1, indicating its involvement in prostate cancer. In this paper, we described a new method for predicting candidate lncRNA targets, and obtained candidate therapeutic targets using this method. We hope this study will bring a new perspective to future lncRNA studies.

Introduction

There has been a consistent focus on protein-coding genes and their products for a long time, but currently, some of the focus is shifting towards non-coding genes. With the completion of the human genome project, researchers found that protein-coding genes were mapped to only a part of the genome[1]. The research of Kapranov P et al.[2] revealed that the transcribed sequences were an order of magnitude greater than the sequences of characterized exons, which indicated that most of the genome is transcribed to produce large numbers of non-coding RNAs (ncRNAs)[3]. The ncRNA world is constantly expanding, from the small ncRNAs including miRNAs, to the more recently described lncRNAs, which range from 200 nt to 100 kilobases (kb)[4]. Once thought to reside in the 'junk DNA' regions, ncRNAs, especially lncRNAs are emerging as an integral functional component of the mammalian transcriptome[5]. The role of ncRNAs as regulators of gene expression is now widely recognized in various species studied to date[6].

LncRNAs are not only essential for gene regulation, but also appear to be the key players in cancer initiation and progression, boasting many attractive features for both biomarker and therapeutic research[7]. For example, Faghihi MA et al.[8] identified alncRNA, BACE1-AS (BACE1-antisense transcript), in the study of Alzheimer's disease. BACE1 is the enzyme responsible for β-amyloid peptide generation, which is the most common cause of dementia. This indicates that BACE1-AS may be an ideal therapeutic target for Alzheimer's disease. PTENP1(pseudogene of PTEN) can be growth-suppressive by regulating cellular levels of PTEN in several cancers[9], HOTTIP (HOXA transcript at the distal tip) can organate the chromatin domain of the HOXA locus[10] and may be involved in leukemias and hepatocellular carcinoma[11]. Therefore, understanding the relationship between human disease and lncRNA would help to clarify the pathogenesis and pathophysiology, and predictiing disease-related lncRNAs may unveil new drug targets for treatment.

Most biological processes could be very complicated as they involve DNAs, RNAs and proteins. Recent studies have shown the ability of lncRNAs to participate in the biological processes by regulating splicing, translation, apoptosis and cell cycle. We can see it this way: lncRNA function as a peculiar biological component, it may interact with proteins, together maintaining the complicated biological processes. Zhao Y et al.[12] constructed a coding-non-coding gene co-expression network to predict the functions of lncRNAs. Zhao's work used Pearson correlation coefficients (PCC) to calculate the correlations between lncRNA-lncRNA, lncRNA-coding gene and coding gene-coding gene, but they considered only the expression relationships between lncRNAs and coding genes, the genes that interact with coding genes were neglected.

In a biological network, the alteration of the expression of PCGs will indirectly affect the expression of their interacting genes. Therefore, lncRNAs may not only regulate the expression of their target genes, but might induce the expression alteration of the genes that interact with the target genes. Especially in a disease state, the alteration (copy number and methylation) of a lncRNA will lead to its upregulation or downregulation, thus disrupting the expression of its target genes. Similar to the butterfly effect, the change started from a few genes, and finally become a whole-genome level event, and result in disease. Based on this assumption, this study considered the expression relationships between PCGs with first-order interactions, and proposed a new approach, using sample correlation to predict lncRNA-PCG pairs and using random walk to search candidate cancer-related lncRNAs in the lncRNA-PCG bipartite network. We analyzed the candidate lncRNAs, and validated their close relationship with prostate cancer.

Materials and Methods

Data Sources

Expression profiling data

We obtained the expression profile from the study by X Shirley Liu et al.[13], which comprised a total of 10,207 lncRNAs and 18,319 PCGs by re-annotating the dataset of GSE21034 from the GEO database. GSE21034 was generated by the MSKCC Prostate Oncogenome Project[14], containing 29 normal adjacent tissues, 131 primary tumor tissues and 19 metastatic tissues, as well as four prostate cell lines, with exon array data. The 29 normal adjacent tissues and 150 cancer tissueswere included in the study.Adifferential expression analysis was carried out on the data using SAM algorithm[15] to identify lncRNAs and PCGs with statistically significant changes in expression between normal adjacent and cancer tissues. In the end, we got 2114 differentially expressed lncRNAs and 4489 differentially expressed PCGs(fold change>1.2).

Protein-protein interaction data

Human protein–protein interaction(PPI) data was downloaded from the BioGRID database(http://thebiogrid.org/)[16] (downloaded on October 2013), which contains both experimental and predicted PPI data. The redundant PPIs were removed, resulting in 213155 interactions.

Seed lncRNAs

The random walk[17] is an iterative procedure starting at a given seed node(s), at each step, the walker moves from its current node to a randomly selected neighbor based on the probabilities of the edges between two nodes. The seed lncRNAs need to be set in advance, in this study, the driver lncRNA spredicted by X Shirley Liu et al.[13]were chosen as the seed lncRNAs. The driver lncRNAs in X Shirley Liu et al. study was defined as the lncRNAs showed significant and concordant expression changes in tumor samples with a corresponding somatic copynumber gain or loss compared to the other samples. The seven identified driver lncRNAs were ENSG00000225951, ENSG00000228544, ENSG00000233016, ENSG00000234741, ENSG00000235999, ENSG00000228288 and ENSG00000231806, which are used as seed nodes to predict new candidate lncRNAs.

Known lncRNAs associated with prostate cancer

Experimental validated prostate cancer-related lncRNAs were obtained from lncRNADisease[18] and MNDR[56]. lncRNADisease (http://cmbi.bjmu.edu.cn/lncrnadisease)[18] is a manually curated lncRNA-disease relations database, which has recorded approximately 480 high quality experimentally verified lncRNA-disease associations. It also curated 478 entries of lncRNA interacting partners at various molecular levels, including protein, RNA, miRNA and DNA. Using the key words ‘prostate cancer’, we acquired lncRNAs and corresponding dysfunction types and references from lncRNA Disease, and identified 12 known lncRNAs associated with prostate cancer. MNDR(www.rna-society.org/mndr/)[56] is a manually curated diverse ncRNA-disease repository by integrating evidence in three mammals(Homo sapiens,Mus musculus and Rattus norvegicus). It also recorded more than 800 experimentally verified lncRNA-disease associations. Using the key words ‘prostate cancer’, we acquired 17 lncRNAs associated with prostate cancer in Homo sapiens, in which 10 lncRNAs with a corresponding Ensembl name were used in subsequent analysis. Using the data obtained from the two databases, we finally identified 14 known lncRNAs associated with prostate cancer(Table 1).

Methods

In previous studies, many methods was used to measure the correlation between two genes. For example, Yamanishi Y et al.[19]applied a generalized kernel canonical correlation analysis to extract correlated gene clusters from multiple genomic data. Costes SV et al.[20]adapted PCC to quantitatively measure the protein-protein colocalization in live cells. This study presents a method for the identification of cancer related lncRNAs by constructing a lncRNA-PCG bipartite network base on sample correlation. The procedures are illustrated via a flow chart in Fig. 1.

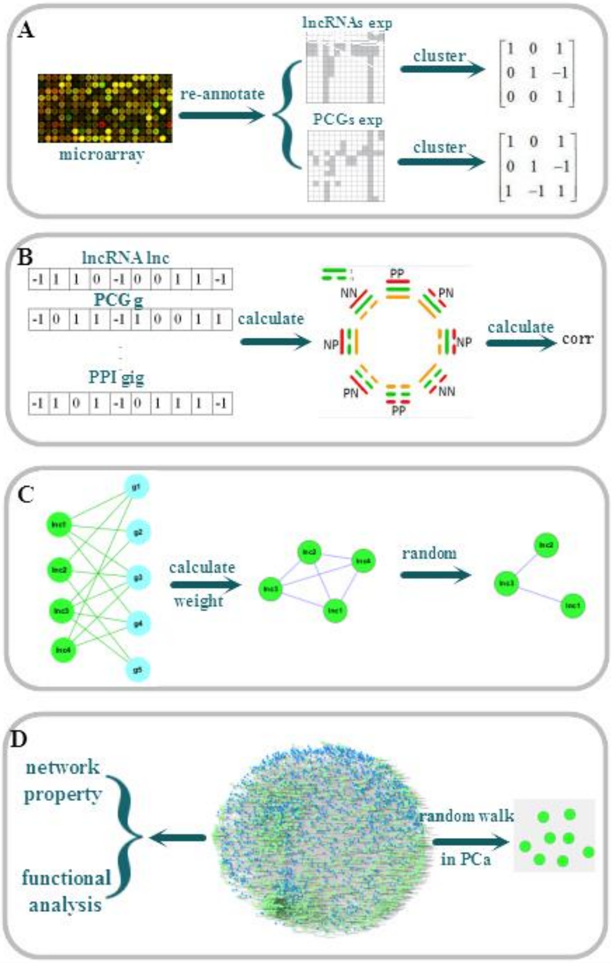

Fig.1 A flowchart of research. This research mainly consists of four steps: (A) Obtain the expression profiles of lncRNAs and PCGs by probe re-annotation, and use cluster analysis to discretize expression values; (B) Calculate the correlation between lncRNAs and PCGs according to the expression patterns of lncRNAs, PCGs and their interacting genes; (C) Calculate the correlation of lncRNA-lncRNA pairs according to the correlation between lncRNAs and PCGs and build a bipartite network; (D) Topology analysis and module analysis on the bipartite network to explore the biological significance of the network, and random walk on the network to identify candidate cancer lncRNAs.

Identifying lncRNA-PCG associations

In this study, the log2-transformed expression values were used for the differentially expressed lncRNAs and PCGs for 150 samples. The expressions of each lncRNA is recognized as a vector, and K-means clustering (K=3) was applied on the vectors. Discrete values were then assigned to the three groups: 1, -1, 0 for the group with higher expression levels, lower expression levels and moderate expression levels, respectively. A similar process is repeated for each PCG. Two matrices were then constucted, one is the lncRNA matrix, in which lncRNAs_{ij} means the discretized expression value of the i-th lncRNA in the j-th sample; and the other is the PCG matrix, PCGs_{ij} means the discretized expression value of the i-th PCGs in the j-th sample (Fig.1 A).

Table 1 Prostate cancer-associated lncRNAs

Gene symbol	Ensembl name	Chr	Start	End
GAS5	ENSG00000234741	chr1	173833039	173837125
PCGEM1	ENSG00000227418	chr2	193614571	193641625
TERC	ENSG00000270141	chr3	169482398	169482848
PCAT1	ENSG00000253438	chr8	128025399	128033259
PVT1	ENSG00000249859	chr8	128806779	129113499
CDKN2B-AS1	ENSG00000240498	chr9	21994790	22121096
PCA3	ENSG00000225937	chr9	79379354	79402465
IGF2-AS	ENSG00000099869	chr11	2161758	2169896
MALAT1	ENSG00000251562	chr11	65265233	65273940
C1QTNF9B-AS1	ENSG00000205861	chr13	24463028	24466242
MEG3	ENSG00000214548	chr14	101292445	101327363
CBR3-AS1	ENSG00000236830	chr21	37504065	37528606
HOTAIR	ENSG00000228630	chr12	54356096	54362515
XIST	ENSG00000229807	chrX	73040495	73072588

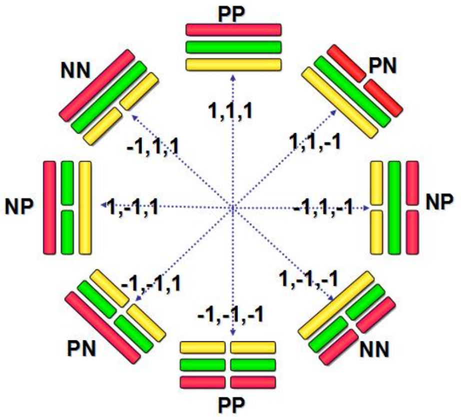

Fig.2 The definition of *PP*, *PN*, *NP* and *NN*. The study defines the vector v = (lnc, g, gig) to measure the expression correlation of lncRNAs, PCGs and PCGs interacting with the former PCGs, and then define four statistics (*PP*, *PN*, *NP* and *NN*) to calculate the number of samples with different types of regulatory patterns. In the figure, the pink bar represents the expression of *lnc*, green represents the expression of *g*, yellow represents the expression of *gig*. In addition, the unbroken bar represents high expression and the broken bar represents low expression. As an example, *PP* means the number of samples with a simultaneously high expression or low expression of *lnc*, *g* and *gig*, that is, the number of samples with v = (1,1,1) or v = (-1, -1, -1).

Let *LNC* be the row vectors in the lncRNAs matrix, *G* be the row vectors in the PCG matrix and *GIG* be the vectors of PCGs interacting with *G* in the PPI network. According to the expression status of *LNC*, *G* and *GIG* in each sample ({*lnc*,*g*,*gig*}denoted as *V*), we defined four statistics, *PP*, *PN*, *NP* and *NN*. As shown in figure Fig.1 B, without considering normal expression, that is, a value of 0 is not taken into consideration, the values of *V* can be divided into eight categories, which may merge into four classes *PP*, *PN*, *NP* and *NN* (Fig.2) depending on the regulation pattern of *lnc*, *g* and *gig*.

PP is the number of samples with simultaneously high expression or low expression of *lnc*, *g* and *gig*. This situation can be explained in a biological regulatory network, a lncRNA upregulates (downregulates) the expression of its target genes, inducing their high expression (low expression), and subsequently causing the high expression (low expression) of genes that interact with them. Similarly, *PN* means the number of samples with a positive correlation of *lnc* and *g* and a negative correlation of *g* and *gig; NP* means the number of samples with a negative correlation of *lnc* and *g*as well as a negative correlation of *g* and *gig*; *NN* means the number of samples with a negative correlation of *lnc* and *g* while a positive correlation of *g* and *gig*. The correlation value of a lncRNA-PCG pair is defined by the following formula:

$$\mathrm{Corr}(LNC,G,GIG)=\frac{\max(PP,PN,NP,NN)}{N} \qquad (1)$$

$$\mathrm{avgCorr}(LNC,G)=\frac{\sum_{i=1}^{k}\mathrm{Corr}(LNC,G,GIG_i)}{k} \qquad (2)$$

In which *N* denotes the number of samples. The number of genes that interact with *G* is represent as *k*, *k* may be more than one, so the relevance of *LNC* and *G* is the mean Corr value for *k GIGs*. The lncRNAs with *k*=0 were excluded from this study. Consistent with the principles of PCC, the more similar the expression pattern of lncRN. As towards the expression pattern of PCGs, the higher correlation they may have. Thus, our method can be used as a measurement of the relationship between the regulatory relationships of lncRNAs and PCGs under certain circumstances. Different from PCC, this method considers the interaction between gene set {*GIGi*} and PCGs, using the expression of the genes with first-order linkage to PCGs in PPI Network.

We used the Skewness and kurtosis normality test methods[21] to test the frequency distribution of association value(avgCorr) for normality. The results showed that its frequency distribution follows a normal distribution(*p* value<0.05) (Fig.3). The pairs with top 5% of avgCorr were taken as significant lncRNA-PCG pairs, and used to calculate the correlation of lncRNA-lncRNA pairs.

Calculating lncRNA-lncRNA correlation

MicroRNAs tend to exert the same or similar functions by the inhibition of common target genes in a coordinated manner[22]. Based on this, we believe that the functional similarity of lncRNAs is related to the number of common target genes. Therefore,the similarity score between two lncRNAs (*lnc*₁, *lnc*₂) is defined by the following formula:

$$W(lnc_1, lnc_2) = \sum \frac{2}{1/c_1 + 1/c_2} \times \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \qquad (3)$$

In which S_1 and S_2 are the gene sets associate with LNC_1 and LNC_2 respectively, c_1= avgCorr (LNC_1， *Gi*), c_2= avgCorr (LNC_2， *Gi*), *Gi* means the common genes inS_1 and S_2.We employ the harmonic mean instead of the arithmetic mean, as when only one of the lncRNAs is strongly related to the gene set, the result of the arithmetic average would also be large, thus distorting the actual correlation between the two lncRNAs.

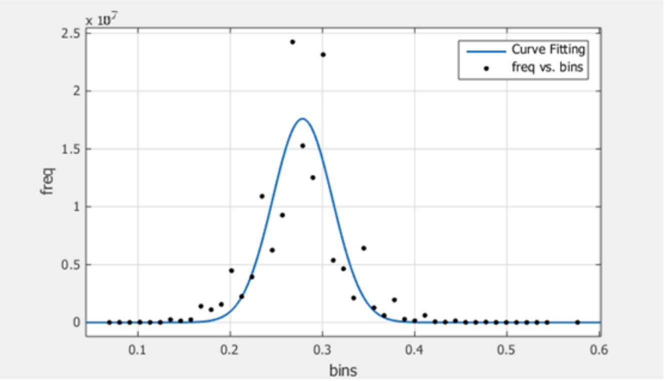

Fig.3 The frequency distribution of avgCorr, the horizontal axis denotes the correlation value and the vertical axis denotes the frequency. As can be observed, the result follows a normal distribution.

Constructing lncRNA-PCG networks and identifying candidate lncRNAs of cancer

We performed 1000 permutations and assessed the statistical significance (*p*<0.05) of lncRNA-lncRNA interactions by comparing the actual similarity scores to the random ones. Meanwhile, correlations between the expression of PCG pairs were calculated by PCC and those absent in the PPI network were excluded. In the end, these relations were integrated to construct the lncRNA-PCG bipartite network. The network consists of three types of links: lncRNA-lncRNA, lncRNA-PCG, PCG-PCG. In this study, we used an R package ‘RWOAG’[17] to identify candidate lncRNAs of cancer by random walk on the bipartite network.

Results

In this paper, the two node sets in the lncRNA-PCG bipartite network are lncRNA set and PCG set, respectively. LncRNAs and PCGs were linked by calculating the avgCorr value, then the top 5% of avgCorr value were chosen as significant correlation values (denoted as LNC2COD). In addition, the relationships between lncRNAs were calculated by the formula (3), the *p*-value is calculated as the fraction of 1000 random permutations in which the avgCorr value is greater than that in actual data, and the results with a *p*-value less than 0.05 were remained as the final lncRNA-lncRNA relationship. At the same time, the PCC was also applied to compute the relationships between PCGs, and only the final PCG-PCG relationships with an absolute value greater than 0.7 were included. The final PCG pairs are these genes common in both sets (the PPI network and LNC2COD set). By integrating the associations between lncRNA-PCG, lncRNA-lncRNA and PCG-PCG pairs, we constructed a lncRNA-PCG bipartite network containing 3165 nodes (1981 lncRNAs and 1184 PCGs) and 180,358 edges (Fig.4). Further analysis revealed that the degree distribution of the bipartite network followed a power-law distribution(Fig.5), and the vast majority of nodes have degrees less than 100, wherein nodes with degrees less than 20 accounted for about half of the network. Only 11 nodes have degrees more than 500, indicating that the majority of lncRNAs interact with only a few PCGs/lncRNAs, whereas a few lncRNAs interact with a large number of PCGs/lncRNAs.

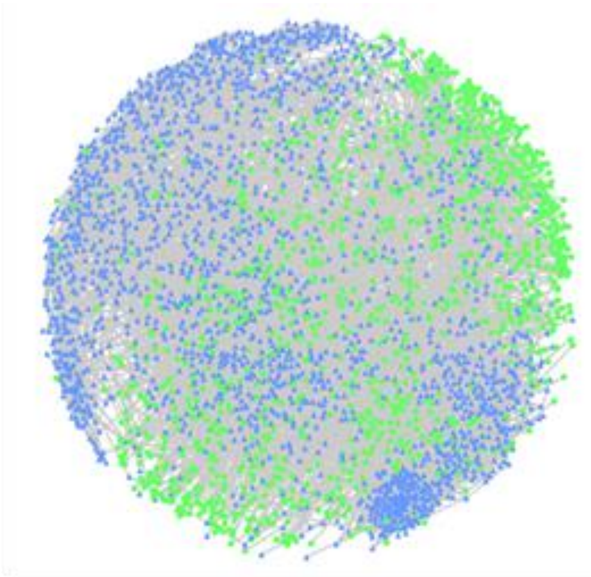

Fig.4TheLncRNA-PCG bipartite network. The blue triangles indicate lncRNAs, green circles indicate PCGs.

Mining cancer functional modules

Biological systems contain a variety of different functions, which are usually formed as distinct subsystems to ensure the independence of the function, and these different subsystems may combine as a whole biological network through some specific connection components. In the network, the independence of the various subsystems can be understood as the network modularity, nodes are closely connected together to subserve a particular function in the internal module, and only particular nodes were used to maintain subsystem connectivity between the modules. In this study, the MCODE plugin in Cytoscape were used to divide the functional modules of the binary network, and identified a total of 61 modules. Preliminary analysis of these modules found that 29 modules consist of less than 20 nodes, the composition of these modules is unitary, always consists of only one type of component (lncRNA or PCG). For example, the module 28(XLOC_001526, XLOC_006800, ENSG00000249945, ENSG00000258285, XLOC_011152) is comprised of five lncRNAs; the module3 (IL18, CANT1, CXorf57, ENSG00000260236, HAUS3, ENSG00000255680, S100A14, ENSG00000263165) is composed of three lncRNAs and five PCGs. This kind of modules is generally recognized as motifs in network analysis. We focus on the modules with a node number equal to or greater than 20, and performed functional annotation and structure analysis on these modules. Since the function of most lncRNAs were not clear, we only conducted functional annotation analysis of PCGs in the module by g:profiler[23], which has a simple, user-friendly web interface with powerful visualization for capturing Gene Ontology (GO), pathway, or transcription factor binding site enrichments down to individual gene levels. The annotations of PCGs in a module were obtained by putting the gene list into the g: GOSt module.

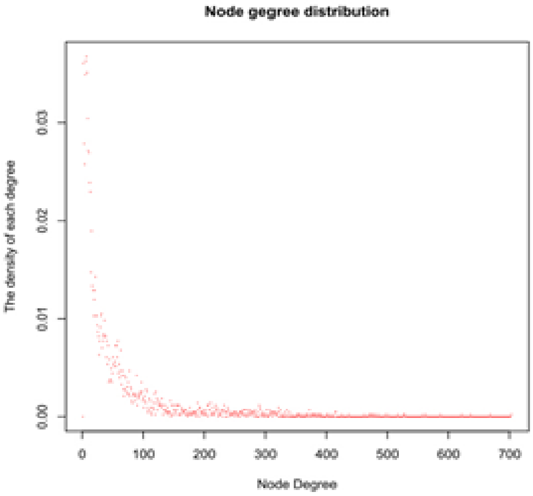

Fig.5 Distribution of the node degree in bipartite network, as can be seen from the figure the degree distributions of the bipartite networks followed a power-law distribution.

In the annotation results of modules, there are 14 modules with significant functional annotation. Module 3 annotated to the function such as positive regulation of insulin secretion(GO:0032024), positive regulation of peptide hormone secretion(GO:0090277), B cell receptor signaling pathway (GO:0050853), and positive regulation of peptide secretion(GO:0002793). Hsing AW et al.[24] has shown that men in the highest tertile of insulin levels had a 2.56-fold risk of prostate cancer compared with men in the lowest tertile, indicating the high serum insulin levels may associate with an increased risk of prostate cancer. Giovannucci E et al.[25] has shown that excessive energy intake tends to increase circulating levels of insulin and free insulin-like growth factor-1 (IGF-I), which may increase risk of some cancers that are common in Western countries. By the study of 3270 incident cancers, they found that higher plasma C-peptide levels were associated with an increased risk of cancer. B cell antigen receptor can signal transduction of cell proliferation and apoptosis, its abnormal can cause defect of the humoral immune response or autoimmunity. Abnormal changes in B cell receptor signaling pathway was likely to cause multiple immune diseases, such as Chronic lymphocytic leukemia[26]and B cell lymphoma[27]. A new study[28] suggested that B cells are not only related to autoimmune disease, but may be associated with the development of prostate cancer. They found that the presence of increased B-cell tumor infiltrating lymphocytes (TILs) was seen in mouse prostate cancer, and the mean intra-tumoral B cell density was higher than in the extra-tumoral region of each prostatectomy section. These studies confirmed the close relationship between functions of the uncovered modules and prostate cancer. Therefore, we suspect that the lncRNAs and PCGs in the module may be associated with prostate cancer, they could be candidate markers of prostate cancer. In addition, module 19 was annotated to the activation of transmembrane receptor protein tyrosine kinase activity (GO:0007171). Tyrosine kinases are known to play central roles in the growth and differentiation of normal and tumor cells[29], and multiple tyrosine kinase transmembrane receptors, such as EphA2 [30], Ack1[31] and VEGFR-2[32], have a tight relationship with the occurrence of prostate cancer, suggesting the biological significance of the uncovered modules. The results of these studies indicate that the modules uncovered in this study have a tight relationship with the occurrence and development of prostate cancer. The results also show that lncRNAs and PCGs together play a role in promoting the deterioration of prostate cancer through interaction. The mechanisms in prostate cancer can be further investigated by analyzing the interaction in the modules.

Candidate lncRNAs of prostate cancer

The research of X Shirley Liuet al.[13]identified seven prostate cancer candidate driver lncRNAs, in which six lncRNAs are in our lncRNA-PCG bipartite network. These six driver lncRNAs were used as seed nodes in the random walk process for scoring each lncRNAs. Then we performed 1000 permutations that shuffled the nodes and links in the bipartite network, the lncRNAs with P values less than 0.05 were identified as new cancer candidate lncRNAs. Finally we got a total of 218 candidate lncRNAs(Supplementary File 1). The top 30 results of random walk are shown in table 2 (seed nodes removed).We tested the accuracy of the candidate of lncRNAs according to the LncRNA Disease database[18] and also confirmed ENSG00000261777,XLOC_006441 and XLOC_003422 as cancer-related by literature.

Six lncRNAs in the bipartite network are included in the lncRNA Disease database thus have been verified to be associated with prostate cancer. They are ENSG00000236830, ENSG00000234741, ENSG00000214548, ENSG00000225937, ENSG00000253438andENSG00000249859, respectively. By the random walk algorithm, we ranked the six lncRNAs in the bipartite network. ENSG00000234741 has the highest score, as one of the seed nodes. Three lncRNAs were included in our result, what is more, ENSG00000225937(gene name PCA3) and ENSG00000236830(gene name CBR3-AS1) ranked 6th(p value is 0.006) and 27th(p value is 0.006)respectively, which illustrated the accuracy of the method. In addition, the P values ofENSG00000214548 and ENSG00000249859 was 0.167 and 0.071, their insignificance may be due to the few amount of PCGs associating with them.

In this study, ENSG00000261777,XLOC_006441 and XLOC_003422 are used as example for further analysis.ENSG00000261777(other names are OTTHUMG00000177044.1 or RP11-529K1.2) ranked first among candidate lncRNAs, located on chr16:70349543-70380650, with a length of 1472 base pairs. The location also contains a methylation site (chr16:70380105-70381766) in peripheral blood mononuclear cell lines. Chen SS et al. [33]showed that DNA methylation plays an important role in the process of the occurrence and development of multiple cancers, and methylation can be used as biomarkers to predict the occurrence of cancer. Other results suggested that the methylation of lncRNAs has an important relationship with disease[34,35]. Therefore, the methylation of ENSG00000261777 may have close relation with the incidence of prostate cancer.

In the lncRNome database (http://genome.igib.res.in/lncRNome)[36], we found an interaction between ENSG00000261777 and PCG IGF2 P1, which have been confirmed by experiment. IGF2(Insulin Like Growth Factor 2) belongs to the insulin family and encodes a polypeptide Growth Factor, it plays an essential role in growth and development. The gene also associate with a variety of diseases such as hypoglycemic coma and Russell-Silver syndrome[37]. Lubik Amy et al.[38] showed that IGF2 contribute to prostate cancer progression via de novo steroidogenesis. SachinBhusari et al. [39] showed that loss of IGF2 imprinting can be employed to identify prostate cancer in patients, in other words, IGF2 can be used as molecular markers for prostate cancer. From the perspective that ENSG00000261777 is involved in the regulation of IGF2, we speculated that the lncRNA may is one of the candidate targets for prostate cancer treatment.

In addition, the exact position of ENSG00000261777 was identified using the UCSC Genome Browser, and there is an overlap with DDX19(DEXD/H-box RNA Helicase, OMIM ID:605812). As shown in Fig.5, the lncRNA appears to be a Long intergenic noncoding RNA (lincRNA). DDX19 is a member of the DEXD/H-box RNA helicase family, it is a recognized RNA helicase which involved in a variety of cell processes, including RNA secondary structure transformation, mitochondrial RNA splicing, assembly of ribosomes and spliceosomes, maintaining the stability of the mRNA and mRNA degradation, etc. Herwig Schüler et al. [40] showed that the N-terminal extension of DDX19 shares significant homology with that of DDX25/GRTH, which is a testis-specific protein with a key regulatory role in spermatogenesis [41],indicating the possible association of DDX19 and the formation of sperm cells. Research has shown that some lincRNAs may act in cis and affect the gene expression of their chromosomal neighborhood[42-44]. Therefore, it's likely that the lincRNA may overlap in function with disease gene DDX19.In X Shirley Liu's study, no copy number change of ENSG00000261777 was observed, but a structure variation of that region is recorded in the DGV (Database of Genomic Variants)[45] Track in UCSC (Fig.6).In CNVD (Copy Number Variation in Disease)(http://bioinfo.hrbmu.edu.cn/CNVD) database[46], the copy number variation of the genomic region (chr16:70349543-70380650) is found to be related to breast cancer[47], suggesting ENSG00000261777's involvement in cancers.

XLOC_006441 (transcript name is TCONS_00013170) is the 8th among candidate lncRNAs of prostate cancer. In Linc2GO(http://www.bioinfo.tsinghua.edu.cn/~liuke/Linc2GO/index.html)[48] database, the lncRNA was annotated to cortisol metabolic process(GO:0034650),cortisol biosynthetic process(GO:0034651), system development (GO:0048731),anatomical structure development (GO:0048856), aldosterone metabolic process(GO:0032341), aldosterone biosynthetic process(GO:0032342), metanephric collecting duct development(GO:0072205), regulation of synaptic plasticity(GO:0048167), mineralocorticoid biosynthetic process (GO:0006705), mineralocorticoid metabolic process (GO:0008212) and collecting duct development(GO:0072044).Multiple studies have shown that cortisol levels can be used as a sign to evaluate the therapeutic effectiveness of prostate cancer [49, 50], providing support for a either a direct or indirect relationship between prostate cancer and the synthesis and metabolism of cortisol. XLOC_006441 was annotated to the functions of cortisol metabolic process and cortisol biosynthetic process, indicating the lncRNA may be a candidate for molecular marker for prostate cancer. We identified a downstream gene (FIGNL1) adjacent to XLOC_006441, Jingsong et al.[57]showed that, FIGNL1 could bind to RAD51 and take part in the homologous recombination process，which is critical for DNA repair, genomic stability, and the prevention of developmental disorders and cancer. The protein complex containing FIGNL1 and KIAA0146 may also have diagnostic and therapeutic potential. The expression of XLOC_006441 may affect the function of FIGNL1, which is also a further evidence of the robustness of the method in this study. XLOC_003422(transcript name is TCONS_00008737) is the 17th among candidate lncRNAs of prostate cancer. In Linc2GO[48] database, the lncRNA was annotated to regulation of small GTPase mediated signal transduction (GO:0051056) and regulation of biological quality (GO:0051056). The cell migration and invasion of prostate cancer is regulated strongly by members of the Ras superfamily of small GTPases, especially the Rho family[51]. Prostate cancer cells can be treated with inhibitors of RhoA signaling or transfect with siRNAs targeting RhoA[52]. These documents proved that small GTPase is involved in the development of prostate cancer. While the function of XLOC_003422 is regulation of the small GTPase mediated signal transduction, we speculate that the lncRNA may induce invasion in prostate cancer cells through the signal pathway, it can be used as a therapeutic target for prostate cancer.

In summary, the candidate lncRNAs uncovered in this study were well validated in the lncRNA Disease database, and we tested a number of candidate lncRNAs by surveying literature for evidence, all the results showed that the candidate lncRNAs are closely associated with prostate cancer, indicating theirpotential as therapeutic targets or molecular markers.

Discussion

This paper acquired the expression profiles of lncRNAs and PCGs by re-annotation of array probes. Based on sample correlation and protein interaction, we constructed a high-quality lncRNA-PCG bipartite network, and identfied candidate cancer-related lncRNAs.

Previous studies always investigate the relation between lncRNAs and PCGs by expression correlation, for example, Zhao Y et al.[12] applied PCC to construct their network. Unlike them, our study applied a correlation measurement considering not only the relationships between lncRNAs and PCGs, but also the protein interaction data. Given a lncRNA-PCG pair, their correlation cannot be ensured only by their expressions, with the noise inherent in microarray data[53-55]. To rectify the situation, protein interaction was incorporated, the correlation between lnc and g was adjusted according to the correlation between *lnc* and *{gig}*, the gene set that functionally related to g. The review of Jeremy et al.[58] showed several possible ways of lncRNA affecting PCGs. The lncRNAs can either promote gene expression by inducing chromatin remodeling, or suppress gene expression by inhibiting RNA polymerase II recruitment. The antisense transcripts of lncRNAs is able to hybridize to the precursor mRNAs of PCGs, thus intervene the alternative splicing of PCGs, or generate endogenous siRNAs to modulate the expression of PCGs. As long noncoding RNAs can function via numerous paradigms, it is hard to track the complex interactions between lncRNAs and PCGs. The functional survey of the lncRNA world has just started, only a few of them have known mechanisms and functions. Taken together, it's reasonable that not enough evidence could be obtained. However, the relation between ENSG00000225937 (PCA3) and FGF8 we identified was verified by the study of Ferreira et al.[59], which partly indicated the feasibility of our method.

However, a major flaw with this method is, till now, there is no complete protein interaction data. Many interactions are predicted by softwares and algorithms, and some of the data are biased, contributing to the inaccuracy of the bipartite network and the lncRNA-PCG pairs. However, the picture of protein interaction is geting more and more complete, we believe that one day our method will effectively predict the relationships between lncRNAs and PCGs.

Furthermore, as we applied statistical method, there was certain requirement for sample size. In the preliminary experiment, we used the expression profile of >8000 human lncRNAs from

Moran N et al.[44], the result correlation values didn't show a normal distribution. But with the increase of sample size, the significance of the distribution of correlation values gradually improves.

To validate the usefulness of this method, we acquired lncRNAs that were confirmed to be prostate cancer related from lncRNA Disease database, and 6 out of 12 were in our bipartite network. As one of the seed node in the random walk, ENSG00000234741 had the highest score among the six lncRNAs, ENSG00000225937, ENSG00000236830 and ENSG00000253438 were also selected as candidate. Other seven lncRNAs didn't show up in the network, in which six lncRNAs wasn't differentially expressed. It may be due to that the lncRNA Disease database collect all the lncRNAs that may be involved in prostate cancer, but since prostate cancer is highly heterogeneous, some of prostate cancer-related lncRNAs are altered only in particular samples, which need further investigation. Besides, although ENSG00000227418 was differentially expressed, only one gene (OMP) was significantly correlated with it, the mere amount of correlating PCGs resulted in a failure in the filtering process.

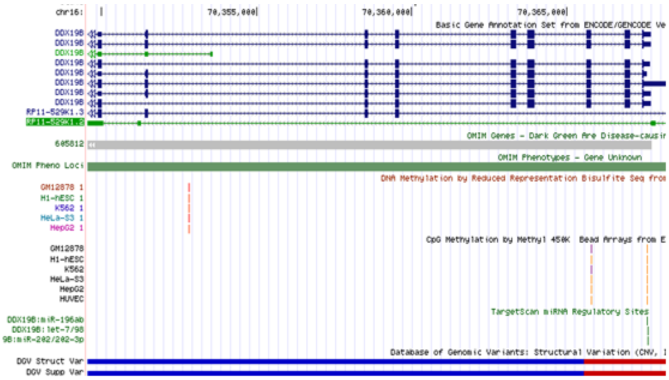
chr16:
70,355,000|
70,360,000|
70,365,000|
Basic Gene Annotation Set from ENCODE/GENCODE Ve
DDX19B
DDX19B
DDX19B
DDX19B
DDX19B
DDX19B
DDX19B
DDX19B
RP11-529K1.3
RP11-529K1.2
OMIM Genes - Dark Green Are Disease-causir
605812
OMIM Phenotypes - Gene Unknown
OMIM Pheno Loci
DNA Methylation by Reduced Representation Bisulfite Seq from
GM12878 1
H1-hESC 1
K562 1
HeLa-S3 1
HepG2 1
CpG Methylation by Methyl 450K Bead Arrays from E
GM12878
H1-hESC
K562
HeLa-S3
HepG2
HUVEC
TargetScan miRNA Regulatory Sites
DDX19B:miR-196ab
DDX19B:let-7/98
9B:miR-202/202-3p
Database of Genomic Variants: Structural Variation (CNV, I
DGV Struct Var
DGV Supp Var

Fig.6 The track of ENSG00000261777 in UCSC genome browser. As can be seen, the lncRNA contains multiple methylation sites and transcription factor binding sites, it overlaps with DDX19, and the track of DGV StructVar shows its copy number variation.

At last, we investigated candidate lncRNAs ENSG00000261777, XLOC_006441 and XLOC_003422, functional analysis and literature evidence confirmed their direct/indirect relation with prostate cancer. Notably, the top ranked lncRNA ENSG00000261777 interact with IGF2 P1, their binding at chr16:70351274 has been experimental validated[36]. Recent studies showed that IGF2 contribute to prostate cancer progression via de novo steroidogenesis[38], and IGF2 can be used as molecular markers for prostate cancer[39]. These studies suggested the role of IGF2 in the initiation and progression of prostate cancer. As ENSG00000261777 can regulate IGF2, we are inclined to believe that ENSG00000261777 may be an essential factor impacting on the initiation and progression of prostate cancer. So, it is reasonable to speculate that the lncRNAs uncovered in this study is the candidate lncRNAs that relate to prostate cancer.

Generally speaking, our study used a new method to construct a lncRNA-coding gene bipartite network, then performed a large-scale prediction of candidate cancer-related lncRNAs. Our candidate lncRNAs are proved to be significantly related to cancer, this result will provide a new source for further research. With the development of technology, the data source of lncRNA expression will no more be limited to re-annotation, more accurate data may be obtained by RNA-Seq and specified lncRNA array. We hope in the near future to apply our method to other cancer data, and identify more candidate cancer-related lncRNAs.

Table 2 The top 30 lncRNAs results of random walk

lncRNA name	Random score	P value
ENSG00000261777	0.001972558	0.003
ENSG00000257117	0.001554664	0.007
ENSG00000226950	0.001544459	0.007
ENSG00000224298	0.001518964	0.004
XLOC_007883	0.001505695	0.006
ENSG00000225937	0.001472041	0.006
ENSG00000224525	0.001468371	0.008
XLOC_006441	0.001073719	0.005
ENSG00000261744	0.001048421	0.002
ENSG00000230082	0.001042155	0.008
ENSG00000257298	0.001029042	0.005
ENSG00000258727	0.001012561	0.005
ENSG00000230918	0.001009477	0.011
ENSG00000230387	0.001007374	0.003
ENSG00000242687	0.001002077	0.002
XLOC_004110	0.001002005	0.007
XLOC_003422	0.000999801	0.006
XLOC_011936	0.000999146	0.006
XLOC_011152	0.000995304	0.004
ENSG00000228392	0.000646705	0.008
ENSG00000215841	0.000641354	0.009
XLOC_001181	0.000633649	0.004
ENSG00000227646	0.000627846	0.005
ENSG00000215190	0.000618964	0.007
ENSG00000229862	0.000617786	0.004
XLOC_000881	0.000617511	0.01
ENSG00000236830	0.000617147	0.006
XLOC_012045	0.000615553	0.009
ENSG00000254100	0.000613928	0.007
ENSG00000217576	0.000613386	0.004

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 81372492), the Natural Science Foundation of Heilongjiang Province (Grant No. D201116), and part by Scientific Research Fund of Heilongjiang Provincial Education Department (Grant No. 12541278)

Notes and references

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China.

*Corresponding author E-mail: xuyan@ems.hrbmu.edu.cn.

§ These authors contributed equally to this work

1. Shabalina, S.A. and N.A. Spiridonov, *The mammalian transcriptome and the function of non-coding DNA sequences*.Genome Biol, 2004.**5**(4): p. 105.
2. Kapranov, P., et al., *Large-scale transcriptional activity in chromosomes 21 and 22*.Science, 2002.**296**(5569): p. 916-9.
3. Mercer, T.R., M.E. Dinger, and J.S. Mattick, *Long non-coding RNAs: insights into functions.* Nat Rev Genet, 2009. **10**(3): p. 155-9.
4. Costa, F.F., *Non-coding RNAs: Meet thy masters*.Bioessays, 2010.**32**(7): p. 599-608.
5. Gibb, E.A., et al., *Human cancer long non-coding RNA transcriptomes*.PLoS One, 2011.**6**(10): p. e25915.
6. Brosnan, C.A. and O. Voinnet, *The long and the short of noncoding RNAs*.CurrOpin Cell Biol, 2009.**21**(3): p. 416-25.
7. Bolton, E.M., et al., *Noncoding RNAs in prostate cancer: the long and the short of it*.Clin Cancer Res, 2014.**20**(1): p. 35-43.
8. Faghihi, M.A., et al., *Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase*.Nat Med, 2008.**14**(7): p. 723-30.
9. Poliseno, L., et al., *A coding-independent function of gene and pseudogene mRNAs regulates tumour biology*.Nature, 2010.**465**(7301): p. 1033-8.
10. Wang, K.C., et al., *A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression*.Nature, 2011.**472**(7341): p. 120-4.
11. Quagliata, L., et al., *Long noncoding RNA HOTTIP/HOXA13 expression is associated with disease progression and predicts outcome in hepatocellular carcinoma patients*.Hepatology, 2014.**59**(3): p. 911-23.
12. Liao, Q., et al., *Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network*.Nucleic Acids Res, 2011.**39**(9): p. 3864-78.
13. Du, Z., et al., *Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer*.Nat StructMolBiol, 2013.**20**(7): p. 908-13.
14. Taylor, B.S., et al., *Integrative genomic profiling of human prostate cancer*.Cancer Cell, 2010.**18**(1): p. 11-22.
15. Tusher, V.G., R. Tibshirani, and G. Chu, *Significance analysis of microarrays applied to the ionizing radiation response*.ProcNatlAcadSci U S A, 2001. **98**(9): p. 5116-21.
16. Stark, C., et al., *BioGRID: a general repository for interaction datasets*.Nucleic Acids Res, 2006.**34**(Database issue): p. D535-9.
17. Kohler, S., et al., *Walking the interactome for prioritization of candidate disease genes*.Am J Hum Genet, 2008. **82**(4): p. 949-58.
18. Chen, G., et al., *LncRNADisease: a database for long-non-coding RNA-associated diseases*.Nucleic Acids Res, 2013.**41**(Database issue): p. D983-6.
19. Yamanishi, Y., et al., *Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis*.Bioinformatics, 2003.**19 Suppl 1**: p. i323-30.
20. Costes, S.V., et al., *Automatic and quantitative measurement of protein-protein colocalization in live cells*.Biophys J, 2004. **86**(6): p. 3993-4003.
21. Kim, H.Y., *Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis*.Restor Dent Endod, 2013.**38**(1): p. 52-4.
22. Jiang, Q., et al., *Prioritization of disease microRNAs through a human phenome-microRNAome network*. BMC SystBiol, 2010. **4 Suppl 1**: p. S2.
23. Reimand, J., et al., *g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments*.Nucleic Acids Res, 2007.**35**(Web Server issue): p. W193-200.
24. Hsing, A.W., et al., *Prostate cancer risk and serum levels of insulin and leptin: a population-based study*. J Natl Cancer Inst, 2001. **93**(10): p. 783-9.
25. Giovannucci, E., et al., *Height, predictors of C-peptide and cancer risk in men*.Int J Epidemiol, 2004. **33**(1): p. 217-25.
26. Herman, S.E., et al., *Ibrutinib inhibits BCR and NF-kappaB signaling and reduces tumor proliferation in tissue-resident cells of patients with CLL*.Blood, 2014.**123**(21): p. 3286-95.
27. Deeb, S.J., et al., *N-linked glycosylation enrichment for in-depth cell surface proteomics of diffuse large B-cell lymphoma subtypes*.Mol Cell Proteomics, 2014.**13**(1): p. 240-51.
28. Woo, J.R., et al., *Tumor infiltrating B-cells are increased in prostate cancer tissue*.J Transl Med, 2014.**12**: p. 30.
29. Robinson, D., et al., *A tyrosine kinase profile of prostate carcinoma*.ProcNatlAcadSci U S A, 1996. **93**(12): p. 5958-62.
30. Zeng, G., et al., *High-level expression of EphA2 receptor tyrosine kinase in prostatic intraepithelial neoplasia.* Am J Pathol, 2003.**163**(6): p. 2271-6.
31. Mahajan, N.P., et al., *Activated tyrosine kinase Ack1 promotes prostate tumorigenesis: role of Ack1 in polyubiquitination of tumor suppressor Wwox*.Cancer Res, 2005.**65**(22): p. 10514-23.
32. Krupitskaya, Y. and H.A. Wakelee, *Ramucirumab, a fully human mAb to the transmembrane signaling tyrosine kinase VEGFR-2 for the potential treatment of cancer*.CurrOpinInvestig Drugs, 2009.**10**(6): p. 597-605.
33. Chen, S.S., et al., *Epigenetic changes during disease progression in a murine model of human chronic lymphocytic leukemia*.ProcNatlAcadSci U S A, 2009. **106**(32): p. 13433-8.
34. Greer, E.L. and Y. Shi, *Histone methylation: a dynamic mark in health, disease and inheritance*. Nat Rev Genet, 2012. **13**(5): p. 343-57.
35. Wu, S.C., E.M. Kallin, and Y. Zhang, *Role of H3K27 methylation in the regulation of lncRNA expression*.Cell Res, 2010.**20**(10): p. 1109-16.
36. Bhartiya, D., et al., *lncRNome: a comprehensive knowledgebase of human long noncoding RNAs*. Database (Oxford), 2013. **2013**: p. bat034.
37. Safran, M., et al., *GeneCards Version 3: the human gene integrator*. Database (Oxford), 2010. **2010**: p. baq020.
38. Lubik, A.A., et al., *IGF2 increases de novo steroidogenesis in prostate cancer cells*.EndocrRelat Cancer, 2013.**20**(2): p. 173-86.
39. Bhusari, S., et al., *Insulin-like growth factor-2 (IGF2) loss of imprinting marks a field defect within human prostates containing cancer*.Prostate, 2011.**71**(15): p. 1621-30.

40. Collins, R., et al., *The DEXD/H-box RNA helicase DDX19 is regulated by an {alpha}-helical switch.* J BiolChem, 2009. **284**(16): p. 10296-300.
41. Dufau, M.L. and C.H. Tsai-Morris, *Gonadotropin-regulated testicular helicase (GRTH/DDX25): an essential regulator of spermatogenesis.*Trends EndocrinolMetab, 2007.**18**(8): p. 314-20.
42. Ponjavic, J., et al., *Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain.*PLoS Genet, 2009. **5**(8): p. e1000617.
43. Orom, U.A., et al., *Long noncoding RNAs with enhancer-like function in human cells.*Cell, 2010.**143**(1): p. 46-58.
44. Cabili, M.N., et al., *Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses.* Genes Dev, 2011. **25**(18): p. 1915-27.
45. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.*Nucleic Acids Res, 2014.**42**(Database issue): p. D986-92.
46. Qiu, F., et al., *CNVD: text mining-based copy number variation in disease database.* Hum Mutat, 2012. **33**(11): p. E2375-81.
47. Curtis, C., et al., *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.*Nature, 2012.**486**(7403): p. 346-52.
48. Liu, K., et al., *Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis.*Bioinformatics, 2013.**29**(17): p. 2221-2.
49. Hsiao, C.P., et al., *High perceived stress is linked to afternoon cortisol levels and greater symptom distress in patients with localized prostate cancer.*Cancer Nurs, 2011.**34**(6): p. 470-8.
50. Thomas, K.S., et al., *Post-traumatic disorder symptoms and blunted diurnal cortisol production in partners of prostate cancer patients.*Psychoneuroendocrinology, 2012.**37**(8): p. 1181-90.
51. Roy, M., H.J. Kung, and P.M. Ghosh, *Statins and prostate cancer: role of cholesterol inhibition vs. prevention of small GTP-binding proteins.*Am J Cancer Res, 2011. **1**(4): p. 542-61.
52. Schmidt, L.J., et al., *RhoA as a mediator of clinically relevant androgen action in prostate cancer cells*.MolEndocrinol, 2012.**26**(5): p. 716-35.
53. Tu, Y., G. Stolovitzky, and U. Klein, *Quantitative noise analysis for gene expression microarray experiments.*ProcNatlAcadSci U S A, 2002. **99**(22): p. 14031-6.
54. Marshall, E., *Getting the noise out of gene arrays.*Science, 2004.**306**(5696): p. 630-1.
55. Klebanov, L. and A. Yakovlev, *How high is the level of technical noise in microarray data?*Biol Direct, 2007. **2**: p. 9.
56. Wang, Y., et al., Mammalian ncRNA-disease repository: a global view of ncRNA-mediated disease network. Cell Death Dis, 2013. 4: p. e765.
57. Yuan, J. and J. Chen, FIGNL1-containing protein complex is required for efficient homologous recombination repair. Proc Natl Acad Sci U S A, 2013. 110(26): p. 10640-5.
58. Wilusz, J.E., H. Sunwoo, and D.L. Spector, Long noncoding RNAs: functional surprises from the RNA world. Genes Dev, 2009. 23(13): p. 1494-504.
59. Ferreira, L.B., et al., PCA3 noncoding RNA is involved in the control of prostate-cancer cell survival and modulates androgen receptor signaling. BMC Cancer, 2012. 12: p. 507