

## Research Article

# Predicting the Functions of Long Noncoding RNAs Using RNA-Seq Based on Bayesian Network

Yun Xiao,<sup>1,2</sup> Yanling Lv,<sup>1</sup> Hongying Zhao,<sup>1</sup> Yonghui Gong,<sup>1</sup> Jing Hu,<sup>1</sup> Feng Li,<sup>1</sup> Jinyuan Xu,<sup>1</sup> Jing Bai,<sup>1</sup> Fulong Yu,<sup>1</sup> and Xia Li<sup>1</sup>

<sup>1</sup>College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang 150086, China

<sup>2</sup>Key Laboratory of Cardiovascular Medicine Research, Harbin Medical University, Ministry of Education, Harbin, Heilongjiang 150086, China

Correspondence should be addressed to Xia Li; [lixia@hrbmu.edu.cn](mailto:lixia@hrbmu.edu.cn)

Received 15 November 2014; Revised 5 February 2015; Accepted 6 February 2015

Academic Editor: Tatsuya Akutsu

Copyright © 2015 Yun Xiao et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Long noncoding RNAs (lncRNAs) have been shown to play key roles in various biological processes. However, functions of most lncRNAs are poorly characterized. Here, we represent a framework to predict functions of lncRNAs through construction of a regulatory network between lncRNAs and protein-coding genes. Using RNA-seq data, the transcript profiles of lncRNAs and protein-coding genes are constructed. Using the Bayesian network method, a regulatory network, which implies dependency relations between lncRNAs and protein-coding genes, was built. In combining protein interaction network, highly connected coding genes linked by a given lncRNA were subsequently used to predict functions of the lncRNA through functional enrichment. Application of our method to prostate RNA-seq data showed that 762 lncRNAs in the constructed regulatory network were assigned functions. We found that lncRNAs are involved in diverse biological processes, such as tissue development or embryo development (e.g., nervous system development and mesoderm development). By comparison with functions inferred using the neighboring gene-based method and functions determined using lncRNA knockdown experiments, our method can provide comparable predicted functions of lncRNAs. Overall, our method can be applied to emerging RNA-seq data, which will help researchers identify complex relations between lncRNAs and coding genes and reveal important functions of lncRNAs.

## 1. Introduction

There are only ~1% of human transcripts encoding proteins [1], and a large fraction of transcripts is long noncoding RNAs (lncRNAs), which are an unknown component of mammalian genomes [2]. lncRNAs are spliced, polyadenylated ranging from 200 bp to more than 10 kb [3–5]. They are transcribed from genome regions that are known to lack protein-coding genes, open reading frames, and other properties necessary to be translated into proteins [6, 7]. Recent studies showed that lncRNAs play key roles in many important biological processes, such as the development of vertebrates, cell differentiation, and immune responses, and are related to complex human diseases [2, 8–12]. lncRNAs can have diverse functions in gene regulation, especially in the epigenetic control of chromatin [13–16]. The most

famous example is the inactive X chromosome through cis-acting of XIST lncRNA [17]. In addition to cis-regulation, lncRNAs can also act in trans to regulate gene expression [7]. For example, Rinn et al. found that HOTAIR acted in trans to repress HOXD locus transcription [5]. Despite many interesting findings of a few lncRNAs, generalizing these findings to thousands of lncRNAs is difficult. More importantly, the functions of most lncRNAs are largely unknown in comparison to small noncoding RNAs (i.e., microRNAs) [18]. Therefore, predicting functions of lncRNAs remains a greatly substantial challenge.

Currently, there have been some significant efforts applied to identify lncRNAs and explore their functions. Guttman et al. performed loss-of-function experiments on many large intergenic noncoding RNAs expressed in mouse embryonic stem cells and characterized the effects on gene

expression [8]. They demonstrated that these noncoding RNAs (ncRNAs) play key roles in the control of embryonic stem cell state. However, these knockdown experiments are time consuming and labor intensive [19]. Some researchers attempt to predict functions of lncRNAs by means of different biological information, such as sequences or genomic positions of lncRNAs. For example, Bellucci et al. utilized lncRNA secondary structure propensities inferred based on sequence information to predict their associated proteins [20] and hence provide a potential way to predict functions of lncRNAs. Wamstad et al. determined GO enrichment for the two nearest neighboring protein-coding genes relative to lncRNAs and found the involvement of lncRNAs in development, morphogenesis, and transcriptional processes [13]. In addition, studies of protein-coding genes have revealed that the relations between mRNAs with similar half-lives have closely related physiological functions, raising the possibility that the half-lives of noncoding RNAs also can be used to identify their functions [21, 22]. Recently, researchers used reannotation microarray expression data to identify lncRNAs and predicted functions of lncRNAs based on coding-noncoding gene coexpression network [23]. However, reannotation microarray strongly depends on the design of the probes [24, 25].

RNA-sequencing (RNA-seq) performs whole transcriptome sequencing and quantifying gene expression with dynamic range, which overcomes the shortcomings of the microarray technology and has already been widely used for studying model organisms and human [9, 26–28]. Cabili et al. defined a reference catalog of more than 8000 human long intergenic noncoding RNAs from RNA-seq data [11], most of which were not previously described. Recent great advances in RNA-seq and computational methods for reconstructing transcriptome offer a wonderful opportunity to annotate and characterize lncRNAs, and a large number of lncRNAs have been discovered using RNA-seq [11, 29, 30]. Therefore, abundant RNA-seq data allow us to comprehensively identify and quantify lncRNAs (also protein-coding genes) and enable us to study the important roles of lncRNAs in various biological processes.

Here, we used RNA-seq of 58 prostate samples to identify lncRNAs and protein-coding genes and construct transcript profiles of lncRNAs and protein-coding genes, respectively. Based on the Bayesian network method, a regulatory network for capturing relations from lncRNAs to protein-coding genes was constructed. Protein-coding gene modules linked with each lncRNA from the regulatory network were identified through mapping to protein interaction network, and its functions were subsequently predicted. A total of 762 lncRNAs were assigned functions. Consistent with previous reports, many lncRNAs are widely involved in the development, cell cycle, metabolism, and other biological processes.

## 2. Materials and Methods

**2.1. RNA-Seq Data Sets.** Fifty-eight prostate samples [31] were detected using RNA-seq, including 42 prostate cancer samples and 16 benign samples. The alignment BAM files that were available at the Gene Expression Omnibus (GEO)

database with accession number GSE25183 were directly used for subsequent analysis. We obtained other RNA-seq data of 30 prostate cancer samples (GSE22260) [32]. In addition, the raw RNA-seq data of 32 breast cancer samples was downloaded from GEO with accession number GSE45419 [33]. We also used our previously detected RNA-seq data of brain tissues from 38 psychiatric and normal samples that has been deposited at the Sequence Read Archive (SRA) database (accession number SRP035524) [34].

**2.2. Protein-Protein Interaction (PPI) Network.** Protein-protein interaction network can offer a global view to understand gene functions and various cellular processes. The protein interaction network was obtained from the Human Protein Reference Database (HPRD). We extracted the maximum component of the protein interaction network, which contained 36900 interactions and 9219 genes.

**2.3. Construction of Transcript Profiles of lncRNAs and Protein-Coding Genes.** The sequenced reads were mapped to human reference genomic utilizing TopHat [35]. Then, we used Cufflinks [36] to assemble exonic and splice-junction readings into transcripts using their alignment results from TopHat and estimated transcript abundances in fragments per kilobase of exon per million fragments mapped (FPKM) by parsimonious allocations of readings to the transcripts. Subsequently, we used known annotation information from UCSC and Ensembl database to identify ncRNAs, lincRNAs, and protein-coding genes. We combined known ncRNAs and lincRNAs to obtain more comprehensive annotation of lncRNAs (Figure 1).

The assembly results were classified based on the following.

- (i) Genes have at least 90% overlap with known annotation of pseudogenes, which are considered as pseudogenes in our study.
- (ii) Genes that do not pass the above step are then compared with the annotation of known noncoding RNAs and those which have at least 90% overlap with known noncoding RNAs are retained as ncRNAs.
- (iii) Genes that do not pass the above steps and have at least 90% overlap with long intergenic noncoding gene are retained as lincRNAs.
- (iv) The set of genes following the above steps was then compared with the annotation set of known protein-coding genes, if one gene has 60% overlap with known protein-coding genes considering coding gene.
- (v) The remaining genes are unannotated and are thus excluded from our study.

In order to explore potential lncRNA-gene relations, lncRNAs and protein-coding genes were considered for further analysis only if they were expressed in at least 50 samples. Finally, we reserved these lncRNAs longer than 200 bp and constructed transcript profiles of lncRNAs and protein-coding genes. FPKM of lncRNAs and protein-coding genes were set to 0 when they were not present in some samples.

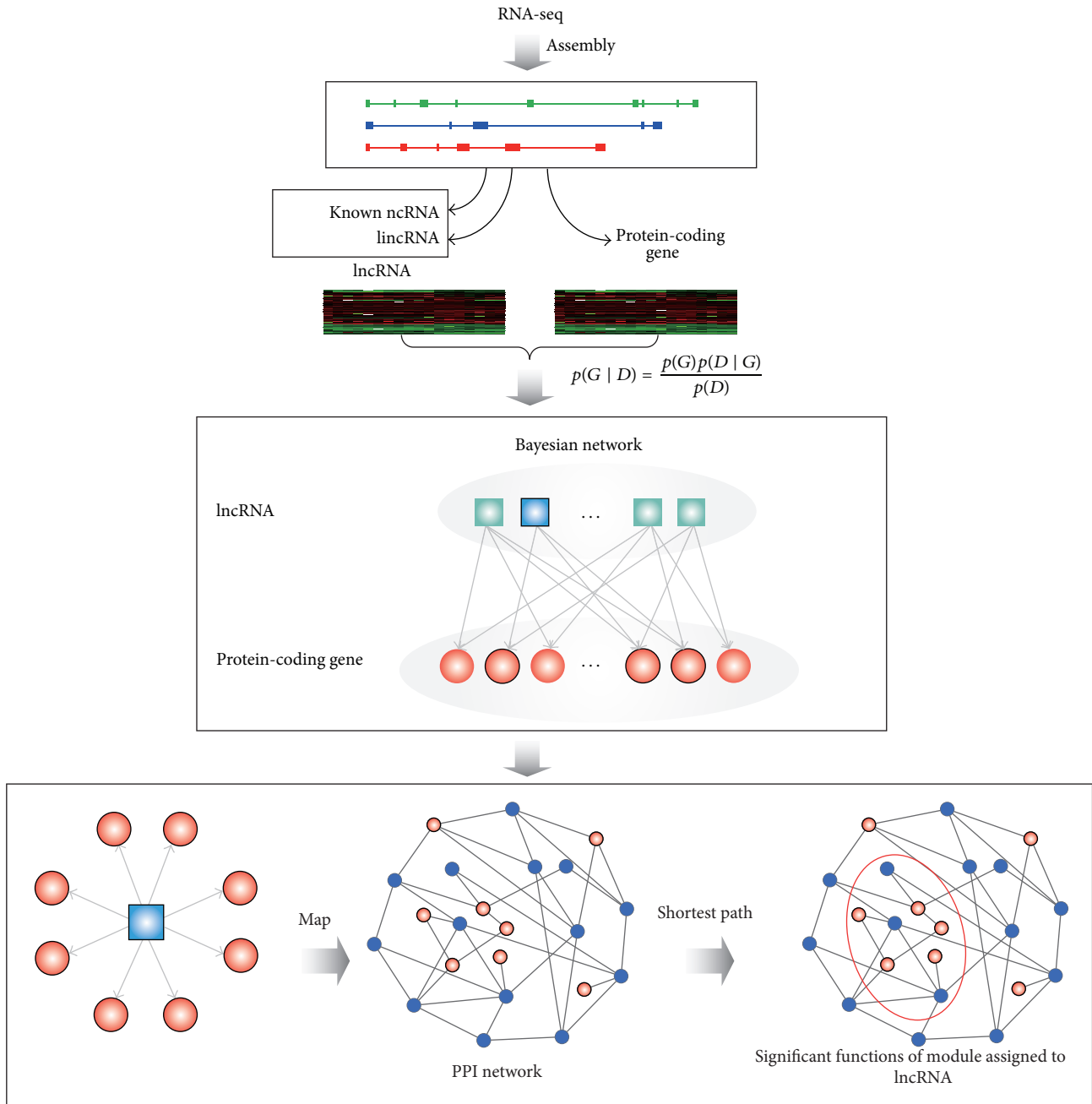


FIGURE 1: The workflow of functional prediction of lncRNAs. We identified lncRNAs and protein-coding genes from RNA-seq data of 58 prostate cancer samples and created transcript profiles of lncRNAs and protein-coding genes for construction of the regulatory network between lncRNAs and protein-coding genes based on the Bayesian network method. To predict functions of individual lncRNA in the regulatory network, we mapped its linked protein-coding genes onto human PPI network and mined highly connected modules, which was subsequently used to predict functions by functional enrichment ( $P$  value  $< 0.05$ ).

2.4. *Construction of Bayesian Network.* In this study, we used Bayesian network to reveal regulatory relationships between lncRNAs and protein-coding genes, which has been widely used for discovering gene regulatory networks [37–41]. Bayesian network represents a joint probability distribution as a directed acyclic graph. It consists of two components.

The first component,  $G$ , is a directed acyclic graph (DAG) whose vertices represent the random variables  $u_1, u_2, \dots, u_N$  and whose edges correspond to dependencies between variables. The second component describes a conditional distribution for each variable which is only dependent on its parent vertices. These two components specify a unique

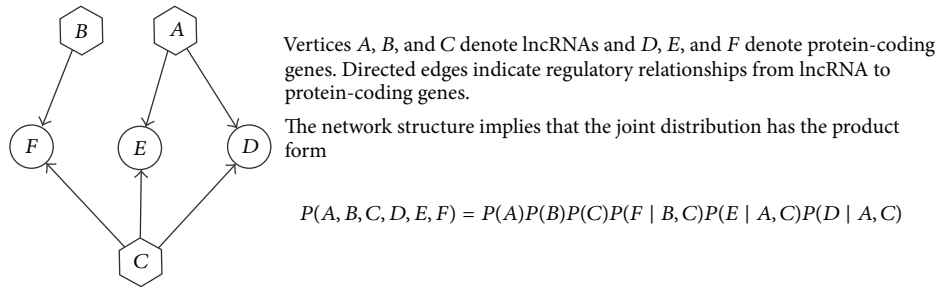


FIGURE 2: An example of a simple Bayesian network structure.

distribution on  $u_1, u_2, \dots, u_N$ . This joint distribution can be decomposed into a product of conditional probabilities based on the graphical structure:

$$p(u_1, u_2, \dots, u_N) = \prod_{i=1}^N p(u_i | \text{Pa}^G(u_i)), \quad (1)$$

where  $\text{Pa}^G(u_i)$  is the set of parents of  $u_i$  in  $G$ . Figure 2 shows an example of a Bayesian network  $G$  and the joint probability distribution it implies.

To construct the regulatory network, we discretized the expression levels of each lncRNA and protein-coding gene from continuous values into two categories (high expression and low expression) by Hartemink's pairwise mutual information method [42]. Then these probability distributions in formula (1) can be computed by counting the frequencies of different combinations. In the process of Bayesian network structure learning, the most likely graph  $G$  for a given data set  $D$  can be inferred by searching for the optimal graph based on a Bayesian scoring metric. As both structure and parameters of the Bayesian network are typically unknown [43], we thus employed the commonly used approximate Bayesian scoring metric, Bayesian information criteria (BIC) [44, 45]. The BIC scoring function can be defined as follows:

$$\text{BICscore}(G, D) = \log L(u_1, u_2, \dots, u_N) - \frac{d}{2} \log n, \quad (2)$$

where  $L(u_1, u_2, \dots, u_N)$  is the likelihood of the data  $D$  according to estimated parameters and structure  $G$ ,  $n$  is the sample size of the data set, and  $d$  is the number of parameters. Finally, the graph space was explored using the greedy hill-climbing algorithm with random restarts to get the most likely graph  $G$  for lncRNAs and protein-coding genes. Pseudocode and illustration for hill-climbing algorithm are shown in Figure 3.

The construction of Bayesian network was generated using  $R$  package bnlearn [45].

**2.5. Prediction of the Functions of lncRNAs by Network Modules.** In general, genes with high interconnections tend to have more similar functions [46]. Therefore, predicting the function of lncRNAs based on their directly linked protein-coding genes together can benefit from network modular strategies [47, 48] because lncRNAs may be involved in multiple biological processes.

We integrate the protein-protein interaction (PPI) network and then predict the functions of lncRNAs using network modules derived from the PPI network. For each lncRNA in the regulatory network constructed based on the Bayesian network method, its directly linked protein-coding genes were mapped onto the human PPI network. Of these genes in the PPI network, we computed the shortest path lengths between any two genes and then used the dynamic cutting tree [49] to mine gene modules. The significant functions enriched by each module ( $P$  value  $< 0.05$ ) were identified. These functions were regarded to be associated with the lncRNA.

### 3. Result

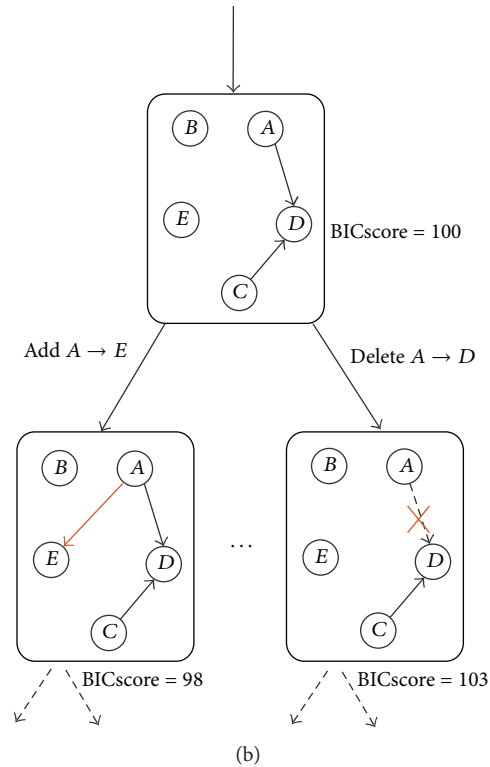
**3.1. Construction of Transcript Profiles of lncRNAs and Protein-Coding Genes Using RNA-Seq.** RNA-seq provides an accuracy and dynamic characterization of the whole cell transcriptome, including different types of RNAs, such as mRNAs and lncRNAs. A total of 58 prostate samples were detected using RNA-seq [31]. The size of 250–300 bp polyA-RNA fragments was selected to construct libraries, which were sequenced using single-end and paired-end on an Illumina Genome Analyzer I and Genome Analyzer II flow cell. In total, approximately 300 million readings were mapped to human reference genome hg18 using TopHat with default parameters. The alignment BAM files that were available at the Gene Expression Omnibus (GEO) database with accession number GSE25183 were directly used for subsequent analysis (Figure 1).

To construct transcript profiles of lncRNAs and protein-coding genes, these mapped readings were assembled into transcripts using Cufflinks [36] with default parameters (a maximum intronic length of 300 kb; minor isoforms with abundance less than 10% of the major isoform). For all 58 samples, 1.69 million transcripts were generated from Cufflinks corresponding to 59404 genes (Supplemental Table S1 available online at <http://dx.doi.org/10.1155/2015/839590>). We annotated these millions of transcripts generated by Cufflinks based on known genomic annotation information from different databases (e.g., UCSC and Ensembl). These genomic annotations are composed of the following: (1) 18921 protein-coding genes from UCSC database coding gene track; (2) 37584 ncRNAs from type of noncoding transcripts in Ensembl database; (3) 8669 pseudogenes obtained through

```

Procedure  $G = \text{BIChillclimb}(D)$ 
(1)  $E \leftarrow \emptyset$ 
(2)  $T \leftarrow \text{Probability Tables}(E, D)$ 
(3)  $G \leftarrow (U, E, T)$ 
(4)  $\text{score} \leftarrow -\infty$ 
(5) do:
  (a)  $\text{maxscore} \leftarrow \text{score}$ 
  (b) for each lncRNA protein-coding gene pair  $(X, Y)$  do
  (c)   for each  $G' \in \{E \cup \{X \rightarrow Y\}, E - \{X \rightarrow Y\}\}$ 
  (d)      $T' \leftarrow \text{Probability Tables}(E', D)$ 
  (e)      $G' \leftarrow (U, E', T')$ 
  (f)      $\text{newscore} \leftarrow \text{BICscore}(G', D)$ 
  (g)     if  $\text{newscore} > \text{score}$  then
            $G \leftarrow G'$ 
            $\text{score} \leftarrow \text{newscore}$ 
(6) while  $\text{score} > \text{maxscore}$ 
(7) Return  $G$ 
    
```

(a)



(b)

FIGURE 3: Pseudocode (a) and illustration (b) of a Bayesian network structure hill-climbing search procedure.  $E, G, D$ , and  $U$ , respectively, are edge sets, Bayesian graph, the training data, and a subset of vertices in Pseudocode (a).

combination of the annotation of pseudogenes in UCSC and Ensembl database; and (4) 21552 lincRNAs from the UCSC lincRNA track.

Based on known annotation information of protein-coding genes, ncRNAs, pseudogenes, and lincRNAs, transcripts in each sample were classified into sets of protein-coding genes, lincRNAs, known ncRNAs, pseudogenes, and unannotated transcripts. ncRNAs with length greater than 200 bp and lincRNAs were both regarded as lincRNAs. We identified 7843 lincRNAs (6267 from the ncRNA annotation and 1576 from the lincRNA annotation) and 15305 protein-coding genes from the 58 prostate samples according to four filtering rules (details in the method section). We found large variance of lincRNAs across all samples; that is, more than 60.7% of lincRNAs (4763 of 7843) occur at just a few samples ( $\leq 10$ ) and only 1355 lincRNAs (17.3%) are present at more than 50 samples. Interestingly, the binary map showing the presence and absence of lincRNAs was able to successfully distinguish normal, metastatic, and localized prostate samples (Figure 4(a)), suggesting the existence of phenotype-specific lincRNAs. In fact, we observed that some lincRNAs trended to be expressed only in metastatic prostate cancer samples and some are expressed only in localized prostate cancer samples. To characterize potential lincRNA-gene relations, lincRNAs and protein-coding genes were obtained only if they were expressed in at least 50 samples. Finally, we obtained 1355 lincRNAs and 8644 protein-coding genes, which were subsequently used to construct a lincRNA-gene regulatory network.

**3.2. Construction of lincRNA-Gene Regulatory Network Based on Bayesian Network Method.** Network analysis offers an efficient method of functional annotation of various biological molecules [50, 51]. Using transcript profiles of lincRNAs and protein-coding genes derived from RNA-seq data, we constructed a lincRNA-gene regulatory network based on the Bayesian network method. We discretized transcript profiles of lincRNAs and protein-coding genes using Hartemink's pairwise mutual information method. Then, using a hill-climbing greedy search on the space of the directed acyclic graph, the optimal network matching transcript profiles of lincRNAs and protein-coding genes were identified. There are 20957 edges referring to 9999 nodes composed of 1355 lincRNAs and 8644 protein-coding genes in the regulatory network containing the dependency relationships between lincRNAs protein-coding genes (Figure 4(b)). We analyzed the distribution of lincRNA degree in the regulatory network (Figure 4(c)), finding that the degree of most lincRNAs is small (mean degree of all lincRNAs is 15.46); only several lincRNAs have large degree (maximum degree of lincRNA is 135). Also, we observed that one lincRNA can connect many protein-coding genes, and one protein-coding gene can also be connected by several lincRNAs.

Previous studies have suggested that lincRNAs can act in cis to activate or silence transcription of genes, such as cis-acting of Xist [17]. However, there is mounting evidence showing that lincRNAs can act in trans, such as HOTAIR silencing HOXD locus [5]. Therefore, we sought to analyze whether lincRNAs tend to affect protein-coding genes in cis

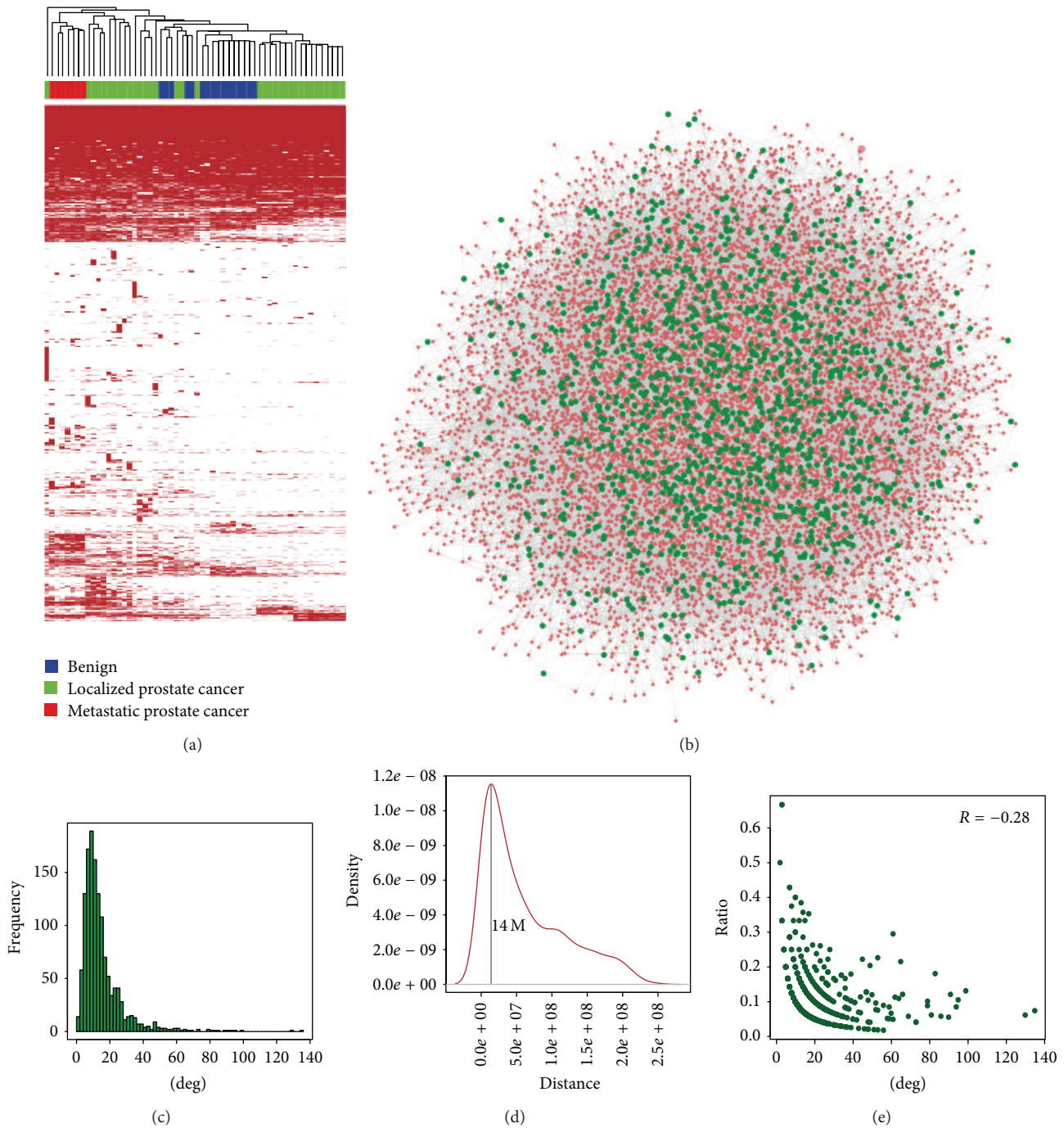


FIGURE 4: (a) A binary map showing the presence (red) and absence (white) of lncRNAs in 58 samples can distinguish prostate cancer samples from benign samples and differentiate localized prostate cancer samples from metastatic samples. (b) The regulatory network between lncRNAs and protein-coding genes was constructed using RNA-seq data of 58 prostate samples based on the Bayesian network method. Red nodes represent protein-coding genes, and green nodes represent lncRNAs. (c) Distribution of degree of lncRNAs in the regulatory network. (d) The distance between lncRNAs and their linked protein-coding genes. (e) The ratio of protein-coding genes located on the same chromosome with their linking lncRNAs.

or in trans. For each lncRNA in the regulatory network, we extracted its linked protein-coding genes and examined how frequent the linked protein-coding genes located at the same chromosome as the lncRNA. We found only a small part of protein-coding genes located at the same chromosomes as

their associated lncRNAs. We further found that lncRNAs affect protein-coding genes with distance around 10–20 Mb with a peak of 14 M (Figure 4(d)), which is more than 100 kb used by analysis of lncRNAs based on neighboring genes [8]. Interestingly, only 15 (1.11%) of 1355 lncRNAs linked

with protein-coding genes within 100 kb were identified, suggesting that lncRNAs may act on broader regions in cis, although most genes are probably affected by lncRNAs in trans. Additionally, we found that lncRNAs with larger degrees connect less frequently with protein-coding genes on the same chromosome (Figure 4(e)), suggesting that some hub lncRNAs may function dependently on their effects in trans.

Considering the existence of disease and normal samples in these 58 prostate samples, we used the significance analysis of microarrays (SAM) method [52] to identify differentially expressed lncRNAs and protein-coding genes (false discovery rate,  $FDR \leq 0.05$ ). Of 1355 lncRNAs, 510 (37.6%) showed differential expression with 351 upregulated and 159 downregulated lncRNAs. Of 8644 protein-coding genes, 3821 (44.2%) showed differential expression with 2295 upregulated and 1526 downregulated coding genes. Among 20957 edges in the regulatory network, there are 2167 edges which show consistent upregulation, 885 edges with consistent downregulation, 838 edges with upregulated lncRNAs and downregulated genes, and 983 edges with downregulated lncRNAs and upregulated coding genes (Supplemental Figure S1A). As expected, upregulated lncRNAs tend to connect with upregulated protein-coding genes (Wilcoxon's rank sum test,  $P$  value  $< 2.2e - 16$ , Supplemental Figure S1B), but not vice versa.

In addition, recent studies showed that dysregulated lncRNAs contribute to many human diseases [31, 53, 54]. A possible hypothesis is that dysfunction lncRNAs may destroy some known disease genes, which in turn induce the development of disease. Therefore, we assessed whether disease-related lncRNAs are connected with known disease genes in the regulatory network. Through manual literature searching, we found 12 disease-associated lncRNAs (including ANRIL, DGCR5, GAS5, H19, Malat1, NEAT1, TUG1, Zfas1, ncRAN, DLEU2, Sox2ot, PTENP1, and PlncRNA-1) [53] in the network and 7655 disease genes derived from OMIM. We did not find obvious difference between the disease and nondisease genes linked by disease-associated lncRNAs (Figure 5). By comparing the mean degrees of nondisease-associated lncRNAs and disease-associated lncRNAs, we found that the mean degree of disease-associated lncRNAs (mean degree 12.31) is lower than nondisease-associated lncRNAs (mean degree 15.5).

**3.3. Predicting Functions of lncRNAs Based on Network Modules.** To predict the functions of lncRNAs, we applied a module-based method that has been extensively used to predict functions of gene sets through integrating our inferred lncRNA-gene regulatory network and protein-protein interactions. For each lncRNA in the regulatory network, their linked protein-coding genes were mapped onto the human PPI network. On the basis of the PPI network, we computed the shortest path lengths between any two genes to mine modules. The significant functions enriched by each module ( $P$  value  $< 0.05$ ) were regarded to be associated with the lncRNA and assigned to the lncRNA.

Of the 1355 lncRNAs in the network, 762 were assigned with enriched functions. The other lncRNAs were not

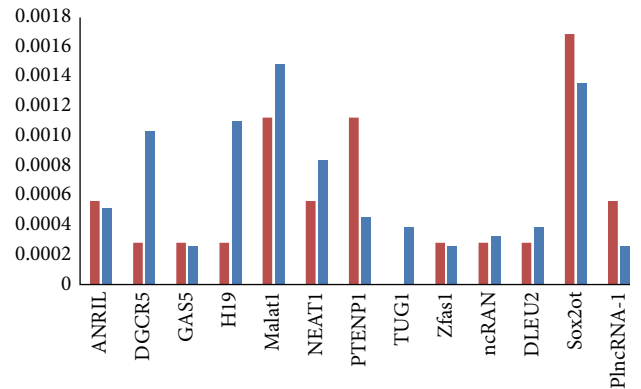


FIGURE 5: There are 12 disease-associated lncRNAs in the regulatory network. Red bar denotes the proportion of disease genes linked by the disease-associated lncRNAs in all disease genes derived from OMIM. Blue bar denotes the proportion of nondisease genes linked by the disease-associated lncRNAs in the protein-coding genes derived from UCSC.

because their linked protein-coding genes either cannot form modules in the PPI network or are not significantly involved in any biological processes. Consistent with previous studies, many of these lncRNAs were found to be associated with development, metabolism, and some fundamental cellular functions (e.g., cell cycle, signal transduction, and transcription) [9, 23]. Using our method can assign functions to known prostate-associated lncRNAs. For example, one lncRNA named PlncRNA-1 has been demonstrated to be related to prostate cancer. A recent study suggests that silencing of PlncRNA-1 significantly reduced cell proliferation and induced apoptosis [54]. Consistently, through our prediction method, we annotated PlncRNA-1 with functions of cell cycle. In the network, we also found that PlncRNA-1 can affect POFUT1 gene involved in Notch signaling pathway, which is required for normal prostatic epithelial cell proliferation and differentiation [55].

By using relationships between lncRNAs and protein-coding genes from the regulatory network, we can also annotate poorly characterized lncRNAs with novel functions. For example, Malat1 lncRNA linked 27 protein-coding genes in the network, which form two modules in the PPI network. Genes in the two modules were significantly involved in biological processes, including cerebellar cortex formation, nucleosome assembly, cell cycle, transcription elongation, and cell-cell signaling (Figure 6(a)). Consistently, previous studies have suggested that Malat1 lncRNA is a component of nuclear bodies and may be associated with the cerebellum of human alcoholics, depletion of which resulted in aberrant mitosis increased cell death [56–59]. The lncRNA Malat1 is dispensable for mouse development [60], which consists of our prediction functions of embryonic morphogenesis. In addition, we found several novel functions of Malat1 lncRNA, such as cell aging, histone modification, and metabolic process. Another lncRNA named NEAT1 linked with 15 protein-coding genes in the network forms one module in PPI network form, which was significantly involved in

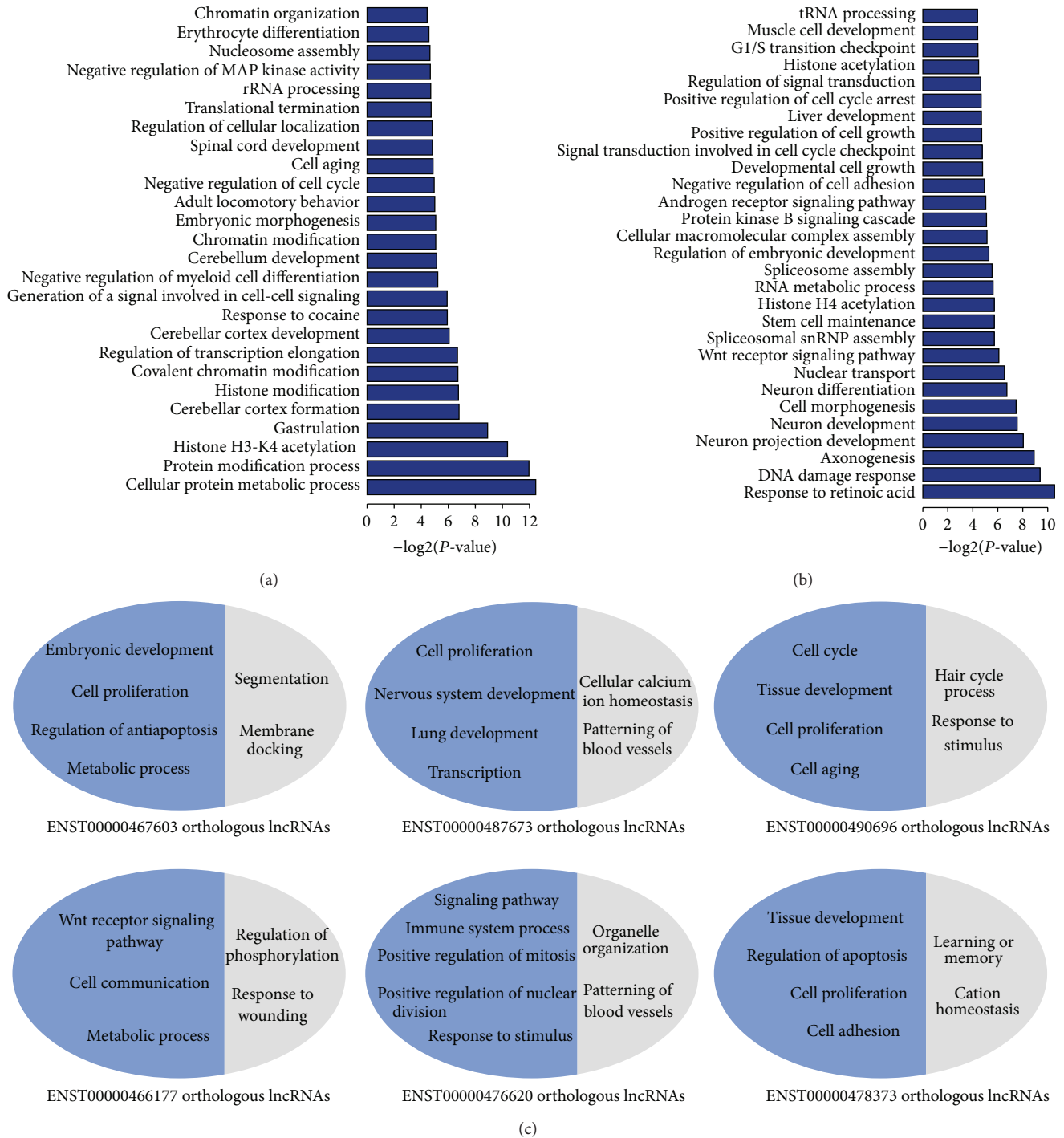


FIGURE 6: (a) Functional enrichment of Malat1 lncRNA. (b) Functional enrichment of NEAT1 lncRNA. (c) Functions predicted by our method overlapping with those determined in the lncRNA knockdown experiments. Blue color represents overlapping functions, and gray color represents functions only predicted by knockdown experiments.

functions related to neuron projection development, cell differentiation, DNA damage response, cell cycle, Wnt receptor signaling pathway, and nuclear transport (Figure 6(b)). Previous studies have suggested that NEAT1 lncRNA has an important structural role in the nuclear paraspeckles [61] and it plays important roles in Huntington's disease by disrupting neuron differentiation [62], which coincides with our

predicted functions. A recent study found that NEAT1 might be a general feature of differentiation [62–64], supporting our predicted function of cell differentiation. Interestingly, novel functions, such as cell cycle and DNA damage response, suggest that NEAT1 lncRNA may be associated with the parthenogenesis of cancer. Further experiments could help to elucidate its roles in cancer.



Furthermore, we analyzed the affected biological processes of lncRNAs through their nearest neighboring genes. We identified the nearest neighboring protein-coding genes of these 1355 lncRNAs in the network and completed the functional enrichment using all of these protein-coding genes. The set of neighboring genes is significantly involved in 263 GO terms ( $P$  value  $< 0.05$ ), such as metabolic process, development, and cell cycle, consistent with previous reports [11, 65]. Of these enriched GO terms, 89.3% were also found using our method. For each lncRNA in the network, we examined the overlaps between GO terms associated with its nearest protein-coding gene and ones predicted using our method. We found that 132 of 762 lncRNAs have at least one shared GO term, but the overlapping GO terms only occupy a small proportion of terms predicted using our method.

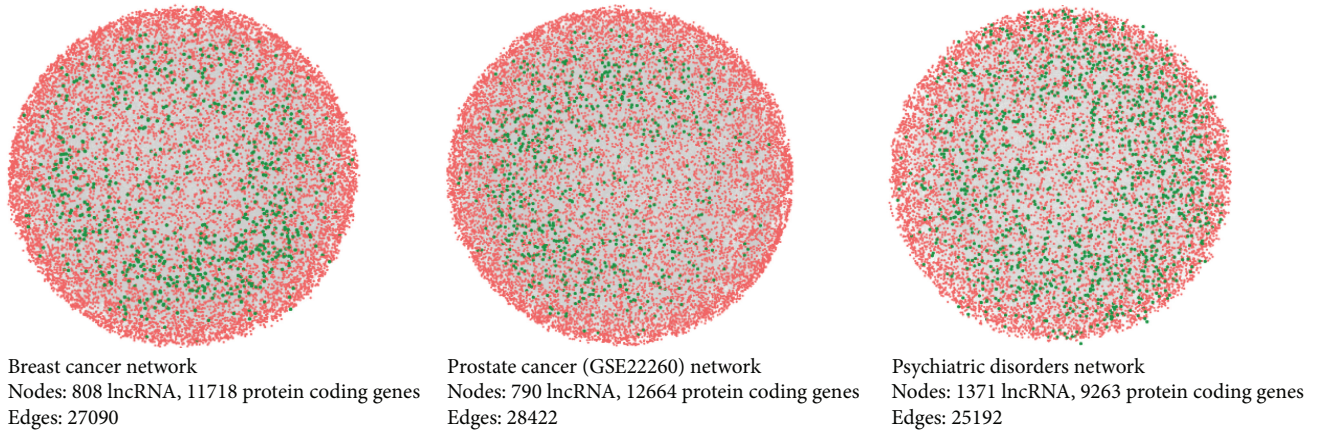
Through knockdown experiments of lncRNAs and subsequent microarray-based expression profiling, a recent study systematically analyzed functions of 147 lncRNAs in mice [8]. To further evaluate the performance of our approach, we searched for orthologous lncRNAs with knockdown data in mice and evaluated the overlapping of functions identified between our approach and knockdown-based experiments. Based on sequence alignment, 48 orthologous lncRNAs were obtained. For each lncRNA, its knockdown expression data were used to determine the affected genes by differential expression analysis (fold change  $> 2$  and  $t$ -test with FDR  $< 0.05$ ). We then performed GO enrichment analysis based on the affected genes for the determination of its functions. Among these 48 lncRNAs, we found that 38 lncRNAs show overlapping of functions identified between our approach and knockdown-based experiments; 4 were not involved in any functions based on knockdown expression data, and 5 were not based on our approach. Also, we found that our approach can capture many important functions (e.g., development, cell proliferation, and cell differentiation) that were also confirmed by knockdown-based experiments (Figure 6(c), Supplemental Figure S2). For example, by analyzing the knockdown expression data of ENST00000467603 orthologous lncRNA in mice, we found the orthologous lncRNA involved in embryonic development, cell proliferation, and cell death processes (Figure 6(c)) that were also predicted using our approach. In particular, we found that the lncRNA connects with the IRF6 gene in the regulatory network, which encodes a member of the interferon regulatory transcription factor (IRF) family involved in the development process [66, 67] and regulation of cell proliferation [68]. Consistently, the knockdown of its orthologous lncRNA in mice significantly affected the expression level of the ortholog gene of IRF6. By analyzing the knockdown expression data of another lncRNA named ENST00000487673 orthologous lncRNA in mice, we found the orthologous lncRNA involved in cell proliferation, lung development, and tissue morphogenesis processes, which were also predicted using our method. Moreover, in the regulatory network, the ENST00000487673 lncRNA was found to link with SLC39A6 gene, an essential cofactor for hundreds of enzymes, encoding one member of the SLC39 family and involved in differentiation and development [69, 70]. Notably, knockdown of its orthologous lncRNA in mice significantly affected

the expression level of SLC39A8 gene, whose orthologous gene in human together with SLC39A6 belong to the SLC39 family [69].

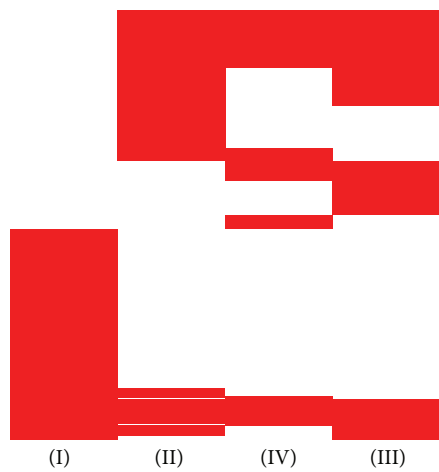
**3.4. Application of the Method to Other RNA-Seq Data.** We applied our method to three other RNA-seq data sets including breast cancer, prostate cancer, and psychiatric disorders and obtained their corresponding regulatory networks. There were 27090 edges referring to 12526 nodes composed of 808 lncRNAs and 11718 protein-coding genes in the regulatory network of breast cancer, 28422 edges referring to 13454 nodes composed of 790 lncRNAs and 12664 protein-coding genes in prostate cancer, and 25192 edges referring to 10634 nodes composed of 1371 lncRNAs and 9263 protein-coding genes in psychiatric disorders (Figure 7(a)). For each regulatory network, we predicted the functions of lncRNAs by using network modules derived from the PPI network. A total of 668, 648, and 717 lncRNAs were assigned with enriched functions in breast cancer, prostate cancer, and psychiatric disorders, respectively. We found that some lncRNAs commonly appeared in these four RNA-seq data sets (Figure 7(b)), and some lncRNAs were present in only one RNA-seq data set, in line with high tissue specificity of lncRNAs [11]. As expected, the two RNA-seq data sets of prostate cancer shared more lncRNAs than others (Figure 7(b)). To compare the functions of lncRNAs among these four RNA-seq data sets, we obtained 85 common lncRNAs. We found that lncRNAs can be involved in the same functions among these four RNA-seq data sets and, notably, they can also be enriched in some tissue-related functions (Figure 7(c)). For example, lncRNA ENST00000448587 (known as TINCR) was found to be enriched in heart development among these four data sets (Figure 7(d)). In particular, we found that the lncRNA was also involved in regulation of cell differentiation in three data sets including breast cancer, prostate cancer, and psychiatric disorders, which was consistent with a previous study that TINCR can control tissue differentiation [71]. Moreover, we found that the lncRNA was enriched in tissue-related functions, such as synaptic transmission and axon ensheathment in psychiatric disorders and response to steroid hormone in prostate cancer. In these four RNA-seq data sets, lncRNA ENST00000411553 was enriched in apoptotic process, suggesting its importance in the development of diseases. We also found that ENST00000411553 was involved in tissue-related functions such as neuron differentiation, axonogenesis, and neuron development in psychiatric disorders, hormone-mediated signaling pathway in prostate cancer, and regulation of cell migration in breast cancer (Figure 7(e)).

## 4. Discussion

Although human genome encodes thousands of lncRNAs, only a few lncRNAs have been functionally characterized, and most functions of lncRNAs remain unknown. Here, we proposed an integrative framework to systematic prediction of lncRNA function. Using combination of a large number of RNA-seq data sets, we constructed a lncRNA-gene regulatory network based on the Bayesian network method and then

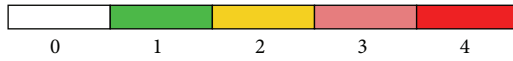
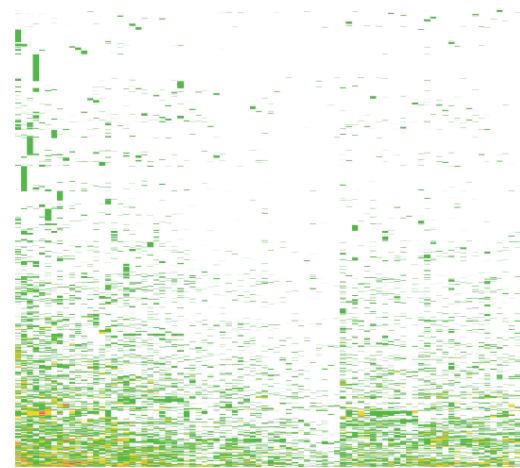


(a)



(I) Psychiatric disorders (III) Prostate cancer (GSE22260)  
(II) Breast cancer (IV) Prostate cancer (GSE25183)

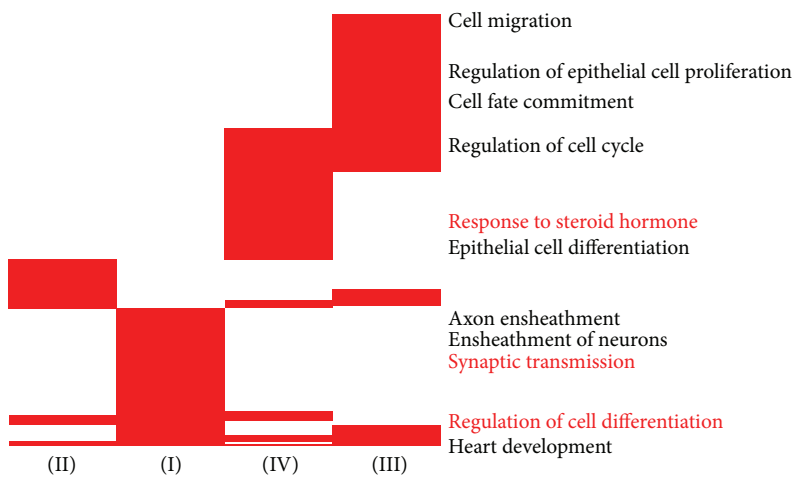
(b)



(c)

ENST00000448587

ENST00000448587



(I) Psychiatric disorders (III) Prostate cancer (GSE22260)  
(II) Breast cancer (IV) Prostate cancer (GSE25183)

(d)

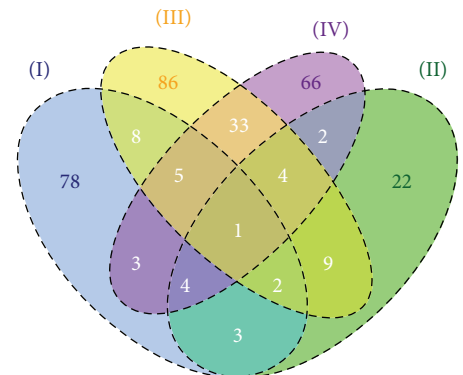


FIGURE 7: Continued.

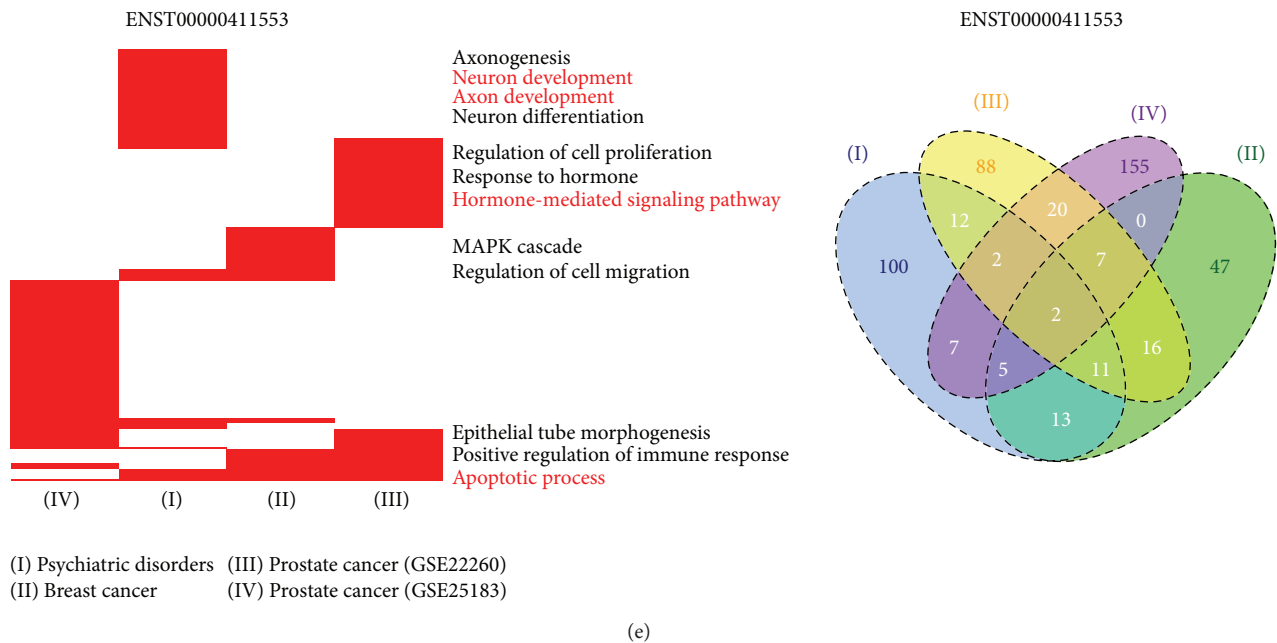


FIGURE 7: (a) Regulatory networks between lncRNAs and protein-coding genes were, respectively, constructed using RNA-seq data of breast cancer, prostate cancer, and psychiatric disorders based on the Bayesian network method. Red nodes represent protein-coding genes, green nodes represent lncRNAs, and edges represent regulatory relationships. (b) Heatmap representing the GO terms of lncRNAs in four data sets. For each lncRNA, detected GO terms are indicated in red. I denotes psychiatric disorders, II denotes breast cancer, III denotes prostate cancer (GSE22260), and IV denotes prostate cancer (GSE25183). (c) Heatmap shows the shared functions of lncRNAs in four data sets (columns). Color key represents the number of data sets sharing the same GO terms for each lncRNA (rows). (d) Functional heatmap (left panel) representing the GO terms (rows) of ENST00000448587 in four data sets (columns). Detected GO terms in each data set are indicated in red. Venn diagram (right panel) showing the number of overlapped GO terms of ENST00000448587 in four data sets. (e) Functional heatmap (left panel) representing the GO terms (rows) of ENST00000411553 in four data sets (columns). Detected GO terms in each data set are indicated in red. Venn diagram (right panel) showing the number of overlapped GO terms of ENST00000411553 in four data sets.

predicted lncRNA functions using a module-based strategy by integrating protein interaction network. Our results show that lncRNAs are involved in diverse biological processes, such as development, metabolism, and differentiation, consistent with many previous studies.

Recently, some researchers used reannotation microarray expression data to predict functions of lncRNAs [23, 72]. However, microarray is greatly dependent on designed probes and hence cannot comprehensively characterize dynamic and relatively low expression of lncRNAs [24, 25]. Also, lncRNAs have strong tissue specificity [9, 11], and many lncRNAs are not identified at present [29]. RNA-seq has the ability to capture the expression levels of genome-wide transcripts, including ones with extremely low expression levels, which thus provides a more precise measurement of levels of transcripts with great dynamic range in comparison to microarray [27, 50]. Expression levels of lncRNAs detected using RNA-seq cannot be reproduced using microarray because of their low correlation [73]. More importantly, RNA-seq data can be used to identify known and novel lncRNAs [11, 29] and quantify their transcript abundance [27] in a specific condition [31]. Therefore, our approach—utilization of RNA-seq data—has the ability to characterize condition-specific lncRNAs and mRNAs, which can further help to systematically depict their potential relations.

With the Bayesian network method as our basis, we constructed a lncRNA-gene regulatory network using transcript profiles of lncRNAs and protein-coding genes generated from RNA-seq data. The Bayesian network allows us to discover causal relations between lncRNAs and genes by capturing properties of conditional independence between variables [37]. It also allows us to handle noise and focus on dependency relationships with strong signals in observed data. It has been widely used for building a variety of regulatory networks [37, 40, 74].

Subsequently, we applied a module-based strategy through combination of the lncRNA-gene regulatory network and protein interaction network. Such modular method has been widely used in prediction of molecular functions because of the prevalence of modular organization of biological networks [51]. lncRNAs may also exert specific functions by regulating function-related genes or by regulating key genes, which in turn affect downstream function-related genes. Using the module-based method, genes connected by a given lncRNA in the regulatory network were divided into coherent groups of genes that show tight connections in protein interaction network. Thus, our method can effectively predict lncRNA functions by considering not only causal relations between lncRNAs and protein-coding genes but also functional associations between genes in protein interaction network.

Furthermore, taking into account cis-acting of lncRNAs, several studies used their neighboring protein-coding genes to annotate the function of lncRNAs. However, it is difficult to determine the genomic range of cis-acting for lncRNAs. That is, there is absence of a unified criterion for the establishment of neighboring protein-coding genes of lncRNAs. Previous studies used different distances to search for their neighboring genes, such as 10 kb [28, 75] and 300 kb [8], and also used one (or two) nearest neighboring protein-coding gene(s) to describe the function of lncRNAs [11, 13]. Moreover, it is difficult to predict functions of individual lncRNAs dependent on one or a few neighboring genes. More importantly, a recent loss-of-function experiment has been used to investigate the effects of lncRNAs on protein-coding gene expression [8]. They found that only 2 of 147 lincRNAs function in cis and most lincRNAs affect gene expression in trans. The trans-acting of lncRNAs will hinder the function characterization of lncRNAs based on neighboring protein-coding genes. In comparison with neighboring gene-based function prediction of lncRNAs, we found a part of lncRNAs with functions predicted using our method overlapping with those predicted by their neighboring genes. Furthermore, we found that our method can also capture most functions of orthologous lncRNAs in mouse determined by lncRNA knockdown experiments. Obviously, our method is not restricted to potential cis-acting of lncRNAs and thus can be used to explore more extensive functions for lncRNAs.

In addition, in the regulatory network, 8644 (45.7%) of 18921 protein-coding genes are linked by lncRNAs, suggesting broad effects of lncRNAs on protein-coding genes and therefore supporting their important roles in biology [8]. Interestingly, we found 10 of the 12 disease-associated lncRNAs linked with at least one disease protein-coding gene in the regulatory network, suggesting that disease lncRNAs may contribute to the pathogenesis of disease by regulating some known disease genes. Their regulatory relations may enable us to discover novel disease lncRNAs using known disease genes.

## 5. Conclusion

In conclusion, we proposed a framework that integrates RNA-seq data and PPI network based on Bayesian network method to comprehensively characterize the functions of lncRNAs. By applying our method to RNA-seq data from prostate samples, we performed a large-scale functional prediction of lncRNAs and analyzed the features of regulatory relations between lncRNAs and protein-coding genes. Our study demonstrated that RNA-seq combining with PPI network based on Bayesian network method is a powerful method for functional analysis of poorly characterized ncRNAs and can be further used for mining functions of ncRNAs in other conditions.

## Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

## Authors' Contribution

Yun Xiao and Yanling Lv contributed equally to this work.

## Acknowledgments

This work was supported in part by the National High Technology Research and Development Program of China (863 Program, Grant no. 2014AA021102), the National Program on Key Basic Research Project (973 Program, Grant no. 2014CB910504), the National Natural Science Foundation of China (Grant nos. 91439117, 61473106, 61170154, and 31200997), the National Science Foundation of Heilongjiang Province (Grant no. C201207), Wu Lien-Teh Youth Science Fund Project of Harbin Medical University (Grant no. WLD-QN1407), and Key Laboratory of Cardiovascular Medicine Research (Harbin Medical University), Ministry of Education.

## References

- [1] U. A. Ørom, T. Derrien, M. Beringer et al., "Long noncoding RNAs with enhancer-like function in human cells," *Cell*, vol. 143, no. 1, pp. 46–58, 2010.
- [2] C. P. Ponting, P. L. Oliver, and W. Reik, "Evolution and functions of long noncoding RNAs," *Cell*, vol. 136, no. 4, pp. 629–641, 2009.
- [3] Y. Hayashizaki, "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, 2005.
- [4] P. Kapranov, J. Drenkow, J. Cheng et al., "Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays," *Genome Research*, vol. 15, no. 7, pp. 987–997, 2005.
- [5] J. L. Rinn, M. Kertesz, J. K. Wang et al., "Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs," *Cell*, vol. 129, no. 7, pp. 1311–1323, 2007.
- [6] M. Baker, "Long noncoding RNAs: the search for function," *Nature Methods*, vol. 8, no. 5, pp. 379–383, 2011.
- [7] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs," *Nature*, vol. 482, no. 7385, pp. 339–346, 2012.
- [8] M. Guttman, J. Donaghey, B. W. Carey et al., "LincRNAs act in the circuitry controlling pluripotency and differentiation," *Nature*, vol. 477, no. 7364, pp. 295–300, 2011.
- [9] I. Ulitsky, A. Shkumatava, C. H. Jan, H. Sive, and D. P. Bartel, "Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution," *Cell*, vol. 147, no. 7, pp. 1537–1550, 2011.
- [10] R. J. Taft, K. C. Pang, T. R. Mercer, M. Dinger, and J. S. Mattick, "Non-coding RNAs: regulators of disease," *Journal of Pathology*, vol. 220, no. 2, pp. 126–139, 2010.
- [11] M. N. Cabili, C. Trapnell, L. Goff et al., "Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses," *Genes & Development*, vol. 25, no. 18, pp. 1915–1927, 2011.
- [12] T. Zhao, J. Xu, L. Liu et al., "Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features," *Molecular BioSystems*, vol. 11, no. 1, pp. 126–136, 2015.
- [13] J. A. Wamstad, J. M. Alexander, R. M. Truty et al., "Dynamic and coordinated epigenetic regulation of developmental transitions in the cardiac lineage," *Cell*, vol. 151, no. 1, pp. 206–220, 2012.

- [14] K. C. Wang and H. Y. Chang, "Molecular mechanisms of long noncoding RNAs," *Molecular Cell*, vol. 43, no. 6, pp. 904–914, 2011.
- [15] L. da Sacco, A. Baldassarre, and A. Masotti, "Bioinformatics tools and novel challenges in long non-coding RNAs (lncRNAs) functional analysis," *International Journal of Molecular Sciences*, vol. 13, no. 1, pp. 97–114, 2012.
- [16] T. Zhao, J. Xu, L. Liu et al., "Computational identification of epigenetically regulated lncRNAs and their associated genes based on integrating genomic data," *FEBS Letters*, vol. 589, no. 4, pp. 521–531, 2015.
- [17] S. Augui, E. P. Nora, and E. Heard, "Regulation of X-chromosome inactivation by the X-inactivation centre," *Nature Reviews Genetics*, vol. 12, no. 6, pp. 429–442, 2011.
- [18] Y. Xiao, J. Guan, Y. Ping et al., "Prioritizing cancer-related key miRNA-target interactions by integrative genomics," *Nucleic Acids Research*, vol. 40, no. 16, pp. 7653–7665, 2012.
- [19] D. Chakraborty, D. Kappei, M. Theis et al., "Combined RNAi and localization for functionally dissecting long noncoding RNAs," *Nature Methods*, vol. 9, no. 4, pp. 360–362, 2012.
- [20] M. Bellucci, F. Agostini, M. Masin, and G. G. Tartaglia, "Predicting protein associations with long noncoding RNAs," *Nature Methods*, vol. 8, no. 6, pp. 444–445, 2011.
- [21] H. Tani, R. Mizutani, K. A. Salam et al., "Genome-wide determination of RNA stability reveals hundreds of short-lived non-coding transcripts in mammals," *Genome Research*, vol. 22, no. 5, pp. 947–956, 2012.
- [22] M. B. Clark, R. L. Johnston, M. Inostroza-Ponta et al., "Genome-wide analysis of long noncoding RNA stability," *Genome Research*, vol. 22, no. 5, pp. 885–898, 2012.
- [23] Q. Liao, C. Liu, X. Yuan et al., "Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network," *Nucleic Acids Research*, vol. 39, no. 9, pp. 3864–3878, 2011.
- [24] S. Flibotte and D. G. Moerman, "Experimental analysis of oligonucleotide microarray design criteria to detect deletions by comparative genomic hybridization," *BMC Genomics*, vol. 9, article 497, 2008.
- [25] F. Duan, M. A. Pauley, E. R. Spindel, L. Zhang, and R. B. Norgren Jr., "Large scale analysis of positional effects of single-base mismatches on microarray gene expression data," *BioData Mining*, vol. 3, no. 1, article 2, 2010.
- [26] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.
- [27] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.
- [28] A. Pauli, E. Valen, M. F. Lin et al., "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis," *Genome Research*, vol. 22, no. 3, pp. 577–591, 2012.
- [29] J. Harrow, A. Frankish, J. M. Gonzalez et al., "GENCODE: the reference human genome annotation for the ENCODE project," *Genome Research*, vol. 22, no. 9, pp. 1760–1774, 2012.
- [30] S. R. Atkinson, S. Marguerat, and J. Bähler, "Exploring long non-coding RNAs through sequencing," *Seminars in Cell and Developmental Biology*, vol. 23, no. 2, pp. 200–205, 2012.
- [31] J. R. Prensner, M. K. Iyer, O. A. Balbin et al., "Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression," *Nature Biotechnology*, vol. 29, no. 8, pp. 742–749, 2011.
- [32] K. Kannan, L. Wang, J. Wang, M. M. Ittmann, W. Li, and L. Yen, "Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 22, pp. 9172–9177, 2011.
- [33] K. R. Kalari, B. M. Necela, X. Tang et al., "An integrated model of the transcriptome of HER2-positive breast cancer," *PLoS ONE*, vol. 8, no. 11, Article ID e79298, 2013.
- [34] Y. Xiao, C. Camarillo, Y. Ping et al., "The DNA methylome and transcriptome of different brain regions in schizophrenia and bipolar disorder," *PLoS ONE*, vol. 9, no. 4, Article ID e95875, 2014.
- [35] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, 2009.
- [36] C. Trapnell, B. A. Williams, G. Pertea et al., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation," *Nature Biotechnology*, vol. 28, no. 5, pp. 511–515, 2010.
- [37] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.
- [38] K. Gifford and H. A. John, *Principled computational methods for the validation and discovery of genetic regulatory networks [Ph.D. thesis]*, Massachusetts Institute of Technology, Cambridge, Mass, USA, 2001.
- [39] Y. Tamada, S. Kim, H. Bannai et al., "Estimating gene networks from gene expression data by combining Bayesian network model with promoter element detection," *Bioinformatics*, vol. 19, supplement 2, pp. ii227–ii236, 2003.
- [40] B. Liu, J. Li, A. Tsykin, L. Liu, A. B. Gaur, and G. J. Goodall, "Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy," *BMC Bioinformatics*, vol. 10, article 408, 2009.
- [41] Y. Wang, X.-S. Zhang, and Y. Xia, "Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data," *Nucleic Acids Research*, vol. 37, no. 18, pp. 5943–5958, 2009.
- [42] A. J. Hartemink, *Principled Computational Methods for the Validation and Discovery of Genetic Regulatory Networks*, Massachusetts Institute of Technology, 2001.
- [43] C. J. Needham, J. R. Bradford, A. J. Bulpitt, and D. R. Westhead, "A primer on learning in Bayesian networks for computational biology," *PLoS Computational Biology*, vol. 3, no. 8, article e129, 2007.
- [44] W. Lam and F. Bacchus, "Learning Bayesian belief networks: an approach based on the MDL principle," *Computational Intelligence*, vol. 10, no. 3, pp. 269–293, 1994.
- [45] M. Scutari, "Learning Bayesian networks with the bnlearn R Package," *Journal of Statistical Software*, vol. 35, no. 3, pp. 1–22, 2010.
- [46] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, supplement, no. 6761, pp. C47–C52, 1999.
- [47] S. Pu, J. Vlasblom, A. Emili, J. Greenblatt, and S. J. Wodak, "Identifying functional modules in the physical interactome of *Saccharomyces cerevisiae*," *Proteomics*, vol. 7, no. 6, pp. 944–960, 2007.
- [48] A. J. Enright, S. van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic Acids Research*, vol. 30, no. 7, pp. 1575–1584, 2002.

- [49] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, 2008.
- [50] O. D. Iancu, S. Kawane, D. Bottomly, R. Searles, R. Hitzemann, and S. McWeeney, "Utilizing RNA-Seq data for de novo co-expression network inference," *Bioinformatics*, vol. 28, no. 12, pp. 1592–1597, 2012.
- [51] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, no. 1, article 88, 2007.
- [52] I. B. Jeffery, D. G. Higgins, and A. C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," *BMC Bioinformatics*, vol. 7, article 359, 2006.
- [53] O. Wapinski and H. Y. Chang, "Long noncoding RNAs and human disease," *Trends in Cell Biology*, vol. 21, no. 6, pp. 354–361, 2011.
- [54] Z. Cui, S. Ren, J. Lu et al., "The prostate cancer-up-regulated long noncoding RNA PlncRNA-1 modulates apoptosis and proliferation through reciprocal regulation of androgen receptor," *Urologic Oncology: Seminars and Original Investigations*, vol. 31, no. 7, pp. 1117–1123, 2013.
- [55] X.-D. Wang, C. C. Leow, J. Zha et al., "Notch signaling is required for normal prostatic epithelial cell proliferation and differentiation," *Developmental Biology*, vol. 290, no. 1, pp. 66–80, 2006.
- [56] S. Nakagawa, J. Y. Ip, G. Shioi et al., "Malat1 is not an essential component of nuclear speckles in mice," *RNA*, vol. 18, no. 8, pp. 1487–1499, 2012.
- [57] J. Y. Ip and S. Nakagawa, "Long non-coding RNAs in nuclear bodies," *Development Growth and Differentiation*, vol. 54, no. 1, pp. 44–54, 2012.
- [58] D. Bernard, K. V. Prasanth, V. Tripathi et al., "A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression," *The EMBO Journal*, vol. 29, no. 18, pp. 3082–3093, 2010.
- [59] V. Tripathi, J. D. Ellis, Z. Shen et al., "The nuclear-retained noncoding RNA MALAT1 regulates alternative splicing by modulating SR splicing factor phosphorylation," *Molecular Cell*, vol. 39, no. 6, pp. 925–938, 2010.
- [60] B. Zhang, G. Arun, Y. S. Mao et al., "The lncRNA malat1 is dispensable for mouse development but its transcription plays a cis-regulatory role in the adult," *Cell Reports*, vol. 2, no. 1, pp. 111–123, 2012.
- [61] Y. T. Sasaki, T. Ideue, M. Sano, T. Mituyama, and T. Hirose, "MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 8, pp. 2525–2530, 2009.
- [62] T. R. Mercer, I. A. Qureshi, S. Gokhan et al., "Long noncoding RNAs in neuronal-glia fate specification and oligodendrocyte lineage maturation," *BMC Neuroscience*, vol. 11, article 14, 2010.
- [63] H. Sunwoo, M. E. Dinger, J. E. Wilusz, P. P. Amaral, J. S. Mattick, and D. L. Spector, "Men epsilon/beta nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles," *Genome Research*, vol. 19, no. 3, pp. 347–359, 2009.
- [64] L.-L. Chen and G. G. Carmichael, "Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: functional role of a nuclear noncoding RNA," *Molecular Cell*, vol. 35, no. 4, pp. 467–478, 2009.
- [65] M. Guttman, I. Amit, M. Garber et al., "Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals," *Nature*, vol. 458, no. 7235, pp. 223–227, 2009.
- [66] C. R. Ingraham, A. Kinoshita, S. Kondo et al., "Abnormal skin, limb and craniofacial morphogenesis in mice deficient for interferon regulatory factor 6 (*Irf6*)," *Nature Genetics*, vol. 38, no. 11, pp. 1335–1340, 2006.
- [67] F. Moretti, B. Marinari, N. L. Iacono et al., "A regulatory feedback loop involving p63 and IRF6 links the pathogenesis of 2 genetically different human ectodermal dysplasias," *The Journal of Clinical Investigation*, vol. 120, no. 5, pp. 1570–1577, 2010.
- [68] C. M. Bailey, D. E. Abbott, N. V. Margaryan, Z. Khalkhali-Ellis, and M. J. C. Hendrix, "Interferon regulatory factor 6 promotes cell cycle arrest and is regulated by the proteasome in a cell cycle-dependent manner," *Molecular and Cellular Biology*, vol. 28, no. 7, pp. 2235–2243, 2008.
- [69] E. A. Bellomo, G. Meur, and G. A. Rutter, "Glucose regulates free cytosolic Zn<sup>2+</sup> concentration, Slc39 (ZIP), and metallothionein gene expression in primary pancreatic islet beta-cells," *The Journal of Biological Chemistry*, vol. 286, no. 29, pp. 25778–25789, 2011.
- [70] K. M. Taylor and R. I. Nicholson, "The LZT proteins; the LIV-1 subfamily of zinc transporters," *Biochimica et Biophysica Acta—Biomembranes*, vol. 1611, no. 1–2, pp. 16–30, 2003.
- [71] M. Kretz, Z. Siprashvili, C. Chu et al., "Control of somatic tissue differentiation by the long non-coding RNA TINCR," *Nature*, vol. 493, no. 7431, pp. 231–235, 2013.
- [72] X. Guo, L. Gao, Q. Liao et al., "Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks," *Nucleic Acids Research*, vol. 41, no. 2, article e35, 2013.
- [73] T. Derrien, R. Johnson, G. Bussotti et al., "The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression," *Genome Research*, vol. 22, no. 9, pp. 1775–1789, 2012.
- [74] S.-Q. Wang and H.-X. Li, "Bayesian inference based modelling for gene transcriptional dynamics by integrating multiple source of knowledge," *BMC Systems Biology*, vol. 6, no. 1, article S3, 2012.
- [75] H. Jia, M. Osak, G. K. Bogu, L. W. Stanton, R. Johnson, and L. Lipovich, "Genome-wide computational identification and manual annotation of human long noncoding RNA genes," *RNA*, vol. 16, no. 8, pp. 1478–1487, 2010.