www.impactjournals.com/oncotarget/ Oncotarget, 2017, Vol. 8, (No. 7), pp: 12041-12051

Research Paper

Systemically identifying and prioritizing risk lncRNAs through integration of pan-cancer phenotype associations

Chaohan Xu[1,*], Rui Qi[1,*], Yanyan Ping[1,*], Jie Li[1], Hongying Zhao[1], Li Wang[1], Michael Yifei Du[3], Yun Xiao[1,2], Xia Li[1]

[1]College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, Heilongjiang, China

[2]Key Laboratory of Cardiovascular Medicine Research, Harbin Medical University, Ministry of Education, China

[3]Weston High School of Massachusetts, Massachusetts, USA

[*]These authors have contributed equally to this work

Correspondence to: *Xia Li,* ***email:*** *lixia@hrbmu.edu.cn*
Yun Xiao, ***email:*** *xiaoyun@ems.hrbmu.edu.cn*

Keywords: *disease phenotype association, risk lncRNA, pan cancer, identification and prioritization*

Received: *July 15, 2016* ***Accepted:*** *December 12, 2016* ***Published:*** *January 05, 2017*

ABSTRACT

LncRNAs have emerged as a major class of regulatory molecules involved in normal cellular physiology and disease, our knowledge of lncRNAs is very limited and it has become a major research challenge in discovering novel disease-related lncRNAs in cancers. Based on the assumption that diverse diseases with similar phenotype associations show similar molecular mechanisms, we presented a pan-cancer network-based prioritization approach to systematically identify disease-specific risk lncRNAs by integrating disease phenotype associations. We applied this strategy to approximately 2800 tumor samples from 14 cancer types for prioritizing disease risk lncRNAs. Our approach yielded an average area under the ROC curve (AUC) of 80.66%, with the highest AUC (98.14%) for medulloblastoma. When evaluated using leave-one-out cross-validation (LOOCV) for prioritization of disease candidate genes, the average AUC score of 97.16% was achieved. Moreover, we demonstrated the robustness as well as the integrative importance of this approach, including disease phenotype associations, known disease genes and the numbers of cancer types. Taking glioblastoma multiforme as a case study, we identified a candidate lncRNA gene *SNHG1* as a novel disease risk factor for disease diagnosis and prognosis. In summary, we provided a novel lncRNA prioritization approach by integrating pan-cancer phenotype associations that could help researchers better understand the important roles of lncRNAs in human cancers.

INTRODUCTION

Long noncoding RNAs (lncRNAs) are a class of non-protein coding transcripts that are longer than 200 nucleotides [1]. They regulates key cellular processes due to their roles in DNA and RNA metabolism and are involved in many complex human diseases including cancer [2, 3]. Systematic studies using high-throughput molecular tools have identified more than 12000 lncRNAs encoded in the human genome with little or no protein-coding capacity (GENCODE Release 23). Cumulative evidence suggests that lncRNAs play crucial roles in tumorigenesis and metastasis. Some lncRNAs, similar to protein-coding genes, can be considered as "oncogenes" or "tumor suppressors" for cancers and are valuable in cancer diagnosis and prognosis [4–7]. However, despite enormous progress made by high-throughput biological detection techniques, the identification of disease related lncRNAs has remained a great challenge for researchers.

Several computational approaches have been proposed to infer novel relationships between lncRNAs and diseases [8–15]. Zhou *et al*. proposed a novel rank-based method (RWRHLD) to prioritize candidate lncRNAs [12]. They constructed a miRNA-associated lncRNA crosstalk network by considering significant co-occurrence of miRNA response elements (MREs) on lncRNA transcripts, a disease–disease similarity network by a directed acyclic graph (DAG) structure and a lncRNA-disease network by using experimentally confirmed lncRNA–disease associations obtained from

lncRNADisease. They integrated these three networks into a heterogeneous network and implemented a random walk with restart on the network for prioritizing candidate disease lncRNAs. They used leave-one-out cross-validation to test the performance of this method based on known experimentally verified lncRNA–disease associations and predict several novel lncRNA–disease associations predicted in ovarian cancer and prostate cancer. In our recent work, we proposed a computational method based on naive Bayesian to identify cancer-related lncRNAs by integrating genome, regulome and transcriptome according to known disease lncRNAs [13]. We totally identified 707 potential cancer-related lncRNAs and demonstrated the performance of the method by ten-fold cross-validation. We found that integration of multi-omic data was necessary to identify cancer-related lncRNAs and our results showed that these candidate lncRNAs tend to exhibit significant differential expression and differential DNA methylation in multiple cancer types. However, the limitation of these studies was the relatively small number of known disease lncRNAs. Subsequently, to improve the limitation, some studies have used known disease miRNAs to help infer disease lncRNAs [12, 14]. Chen *et al*. developed a computational model named Hyper Geometric distribution for LncRNA-Disease Association inference (HGLDA) to prioritize disease candidate lncRNAs [14]. Based on known miRNA-disease relations and experimentally confirmed lncRNA-miRNA interactions detected by CLIP-Seq technology, they used hypergeometric distribution test for each lncRNA-disease pair to detect whether they significantly shared common miRNAs. Those lncRNA-disease pairs with FDR less than 0.05 were selected to be potential lncRNA-disease associations. Moreover, they developed the LFSCM (LncRNA Functional Similarity Calculation based on the information of MiRNA) model to calculate large-scale lncRNA functional similarity by integrating disease semantic similarity, miRNA-disease associations, and miRNA-lncRNA interactions.

Disease associations (namely disease phenotype similarities) can be used to improve the limitation by complementary disease information, which has been successfully applied in prioritization of disease protein-coding genes and miRNAs [16–18]. These studies hypothesized that highly phenotype similar diseases tend to show more close relations, and their relevant genes or miRNAs often reside in the same neighborhood in the interaction networks and form physical or functional modules. Although individual disease may contain only a few information, combination of multiple phenotype similar diseases based on the assumption can provide many additional clues as for the specific disease. Even for those diseases without any known information, such disease association hypothesis can help to reveal some potential risk factors. In our own previous work [17], we presented a miRNA prioritization approach to identify disease-specific miRNAs by using known disease genes and context-dependent miRNA-target interactions derived from matched miRNA and mRNA expression data, independent of known disease miRNAs. Further, we applied this approach to systematically prioritize miRNAs involved in 11 cancer types and yielded an average AUC value of 75.84% based on known disease miRNAs. Due to insufficient disease information on lncRNAs, it is imperative to identify disease-related lncRNAs by curating disease associations or disease knowledge of other risk factors. Additionally, a large number of array-based expression datasets were produced during the past two decades. These array-based expression datasets that have less technical variation and better detection sensitivity can be re-annotated to interrogate lncRNA expression changes when dealing with low-abundance transcripts [19–24]. Array-based datasets simultaneously capture and monitor gene and lncRNA expression in the same cancer samples for diverse cancer types, improves the confirmation process and the quality of identifying lncRNA related genes in the specific contexts [25–27].

Therefore, the aim of our study was to generate an lncRNA computational approach to systematically prioritize and identify candidate disease risk lncRNAs by integrating disease phenotype associations. We interrogated lncRNA expression in thousands of tumor samples and constructed a gene and lncRNA co-expression pan-cancer network (GLCPN) for 14 cancer types. Utilization of known disease genes as seeds independently of disease lncRNAs, we used random walk method to prioritize candidate disease lncRNAs for each cancer type. The average AUC score is 80.66% for prioritization of candidate disease lncRNAs and 97.16% for protein-coding genes. Our results show that through the integration of disease phenotype associations, the lncRNA prioritization performance can be improved, especially for some diseases with few or without known disease lncRNAs.

RESULTS

Construction of GLCPN using the pan-cancer data

Through comprehensively searching "Affymetrix Human Exon 1.0 ST array" in GEO and ArrayExpress databases, forty-three array-based expression studies consisting of 2828 disease samples from fourteen cancer types were identified for our study. The cancer types included bladder cancer (BLC), breast cancer (BC), hepatocellular carcinoma (HCC), gastric cancer (GC), glioblastoma multiforme (GBM), renal cell carcinoma (RCC), medulloblastoma (MB), melanoma (MM), prostate cancer (PC), lung cancer (LC), lymphoblastic leukemia (LL), neuroblastoma (NB), cervical cancer (CC) and ovarian cancer (OC) (Supplementary Table 1).

Through re-annotation, 18376 unique genes and 10092 lncRNAs covered by at least four probes were obtained (Supplementary Figure 1A). After repurposing the expression datasets to probe lncRNA expression for each cancer study, lncRNA expression datasets that had the same number of disease samples as the gene expression datasets were generated. We found that the expression levels of lncRNAs in fourteen cancer types were generally lower than genes (Supplementary Figure 1B), which is consistent with previously reported re-annotation studies [21, 25, 28].

Based on the assumption that genes and lncRNAs with similar expression have similar functions, we constructed disease-specific co-expression sub-networks for each cancer type according to the gene-gene and lncRNA-gene co-expression associations (Figure 1). These fourteen sub-networks were further integrated into a GLCPN. Protein interactions derived from STRING database were also incorporated into the GLCPN. Finally, the pan-cancer network that was constructed included 29071 nodes and 159132861 edges (Supplementary Table 2). The co-expression frequency in the fourteen cancer types or the protein interaction probability obtained from the STRING database was used to weight the edges. This weighted functional network was used for the following prioritization of risk lncRNAs.

Prioritization of disease risk lncRNAs by integrating of disease phenotype associations

To efficiently prioritize candidate disease lncRNAs, we proposed a method based on the random walk that used known disease genes and phenotype similarities to quantify the links between known disease genes and candidate disease lncRNAs in the GLPCN. For one cancer type, a prediction score for each candidate lncRNA was computed (see Methods). Finally, fourteen candidate lncRNA lists represented the prioritization results of fourteen cancer types were generated.

To further investigate the performance of our approach in prioritization of genes, we then performed the LOOCV analysis. Since only one known disease gene respectively can be found in CC and MM, we applied LOOCV to other twelve expression studies. The average AUC score of twelve cancers can reach 97.16%, strongly supporting that our prioritization approach has good prioritization performance (Figure 2A). During the leave-one-out cross validation, we found that all known disease genes in twelve cancer types were ranked in the top 40 out of 18979 genes (0.22%) in the corresponding candidate disease gene lists. For example, known disease genes *BRCA2, CDH1, IDH1, CDKN2A, NME1* and *FGFR3* frequently occurred at the top one in seven cancer types (including BC, GC, GBM, MB, MM, NB and OC), even though only two or three known-disease genes existed in NB, GBM and MB gene lists. To further evaluate the performance of our approach in prioritization of lncRNAs, we extracted known disease lncRNAs of the fourteen cancer types from the LncRNADisease database and computed their AUC scores. The average AUC of fourteen cancers was 80.66% (Figure 2B), suggesting that our methodology efficiently prioritized and identified cancer related risk lncRNAs.

Furthermore, we also investigated the overall distribution of known disease lncRNAs at the top of candidate lists (Figure 2C). Some recently identified disease lncRNAs like *MYCNOS, CDKN2B-AS1, WT1-AS, IGF2-AS* and *GAS5* that play important roles in NB, GBM, LL, RCC and PC [29–33], were ranked at the top of the candidate lists. Intriguingly, we found that more than half of the known disease lncRNAs were ranked at the top 10% of prioritization lists in ten cancer types, namely, BC, HCC, GBM, MB, MM, PC, LC, NB, CC and OC. Moreover, we also selected five representative disease lncRNAs (*HOTAIR, MALAT1, H19, MEG3* and *TUG1*) that play key roles as oncogenic molecules associated with various cancers [34–37] to investigate

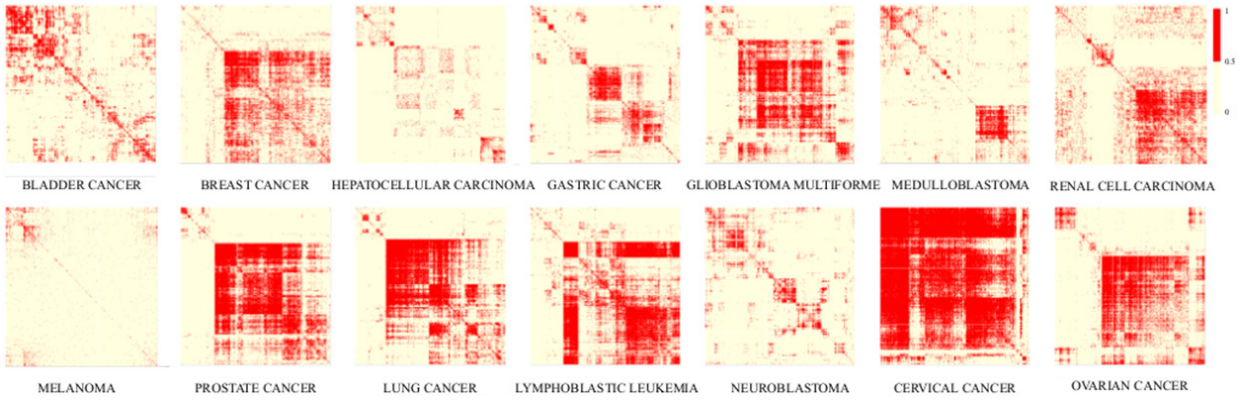

Figure 1: Heat maps of co-expression relationships in the fourteen cancer-types. Clustering maps showing existence of gene-gene or lncRNA-gene co-expression relationships among different cancer types that can be used to construct disease-specific sub-networks.

their occurrences at the top 5%-20% of the candidate lists (Supplementary Figure 2A). We found that *HOTAIR, H19* and *TUG1* almost appeared in all cancer types.

Evaluation of the robustness and the integration importance of lncRNA prioritization approach

The principle of our lncRNA prioritization approach depended on the disease phenotype associations among diverse cancer types, as well as the topological similarities between known disease genes and context-specific co-expression genes of lncRNAs in the GLCPN. Therefore, it was important to evaluate the contribution of these factors to the performance of our lncRNA prioritization approach.

Evaluation of the influence of disease phenotype associations

The fourteen disease phenotype associations can be used to characterize the relationship between diseases and provide the opportunity for us to elucidate the pathogenesis mechanisms of diseases in the crosstalk pattern (Figure 3A and Supplementary Table 3). To evaluate the importance of disease phenotype associations, we prioritized risk lncRNAs in diverse cancer types without utilizing any phenotype associations. The average AUC based on known disease lncRNAs from LncRNADisease was 78.78%, lower than the AUC score (80.66%) with the inclusion of disease phenotype associations (Figure 3B and 3D). Notably, the AUC score for LL dropped from 69.4% to 45.73%, suggesting that the disease phenotype associations can be efficiently used to supply the incomplete information of some diseases and improve the overall performance of lncRNA prioritization.

To further evaluate the influence of disease phenotype similarity, we permuted the phenotype similarity matrix and recomputed the prioritized scores for all candidate lncRNAs (Figure 3C and 3D). The average AUC score was 61.21%. Taken together, the decreased performances in prioritization by remove or permute phenotype associations supported that the disease phenotype association was one of necessary and indispensable factors for the lncRNA prioritization in our approach. It also showed that the disease phenotype associations could efficiently complement the incomplete disease information in individual cancer types and thus provide more power to identify cancer-related lncRNAs through pan-cancer analysis.

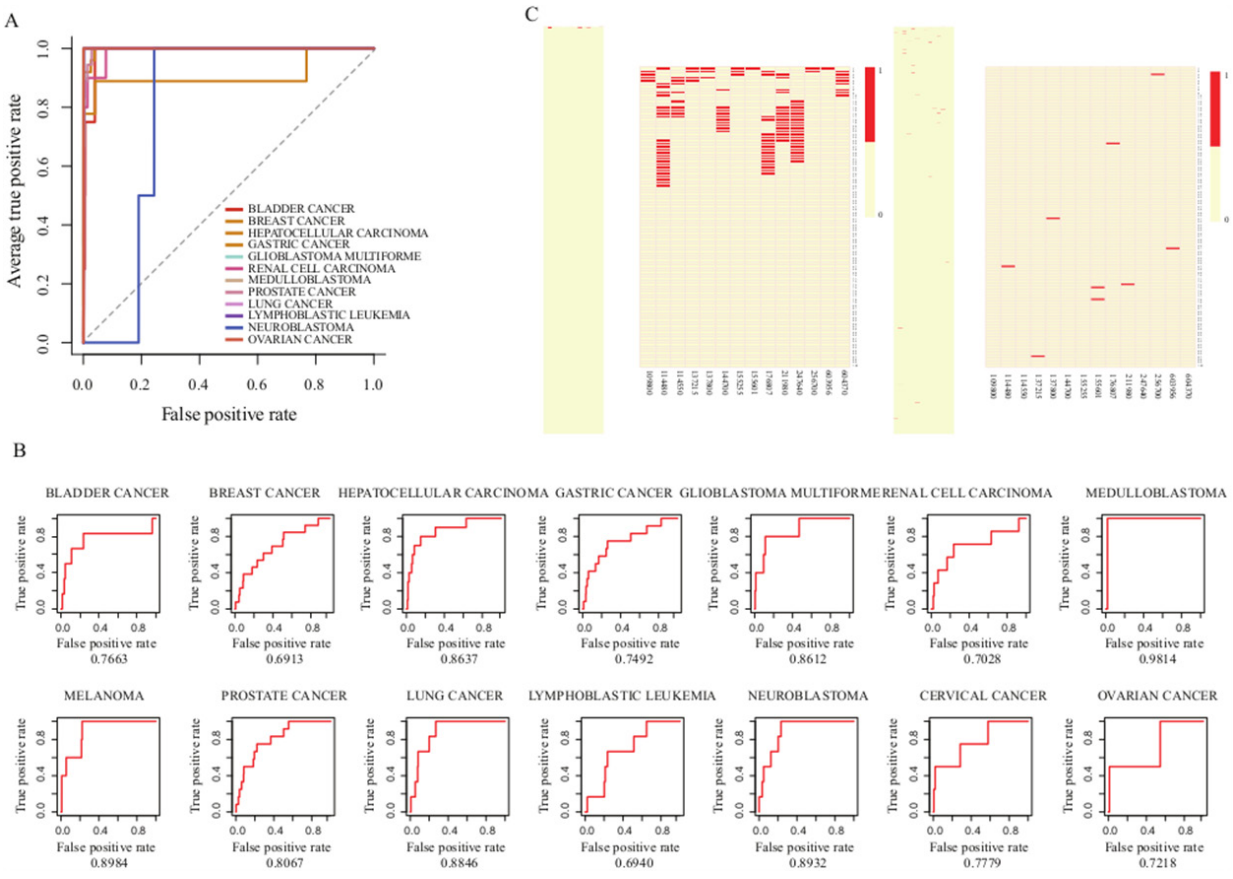

Figure 2: Evaluation of the performance of our lncRNA prioritization approach. A. The ROC curves of gene prioritization results by LOOCV **B.** The ROC curves of lncRNA prioritization results. **C.** Top 100 ranks of known disease genes (left) and lncRNAs (right) after prioritization.

Evaluation of the influence of the number of cancer types

Disease phenotype associations can be used to bridge relationships among different cancer types and efficiently improve the performance of our prioritization approach. Therefore, we sought to determine the performance of our lncRNA prioritization approach by varying the number of cancer types. Towards this, we randomly selected 3, 6, 9 and 12 cancers from the original fourteen cancer types to re-compute prioritization scores for candidate disease lncRNAs. We found that upon increasing the number of cancer types for analysis, the average AUC scores increased from 73.49% to 80.01% (Supplementary Figure 2B). This suggested that utilization of more diseases with their phenotype associations can facilitate the improvement of prioritization of candidate disease lncRNAs.

Evaluation of the influence of known disease genes

Our lncRNA prioritization approach only relied upon known disease related protein-coding genes without the requirement of known disease lncRNAs. Therefore, we evaluated the efficiency of known disease genes by randomly selecting the same number of non-disease-associated genes as known disease genes for each cancer type. Owing to the lack of the non-disease gene set, we obtained a total of 43899 human genes from the NCBI Gene database and 15229 known disease genes from the OMIM database. A non-disease gene set containing 28670 genes was then generated. Equal numbers of non-disease genes for each cancer were randomly selected 1000 times and used for prioritization. We obtained an average AUC score of 59.21% that was significantly lower than the prioritization result based on known disease genes (p<0.001).

Evaluation of prediction performance of our prioritization approach

To assess the prediction performance of our lncRNA prioritization approach, we prioritized candidate disease lncRNAs for each cancer type only using information of the other cancer types through disease phenotype similarities. Surprisingly, the average AUC was 81.63%, supporting that our lncRNA prioritization approach has superior performance in predicting of potential risk lncRNAs (Supplementary Figure 2C). All of the validation results showed that our prioritization approach has a good ability in identification of known disease lncRNAs and genes (Supplementary Figure 2D), even for some diseases with little or without known disease information.

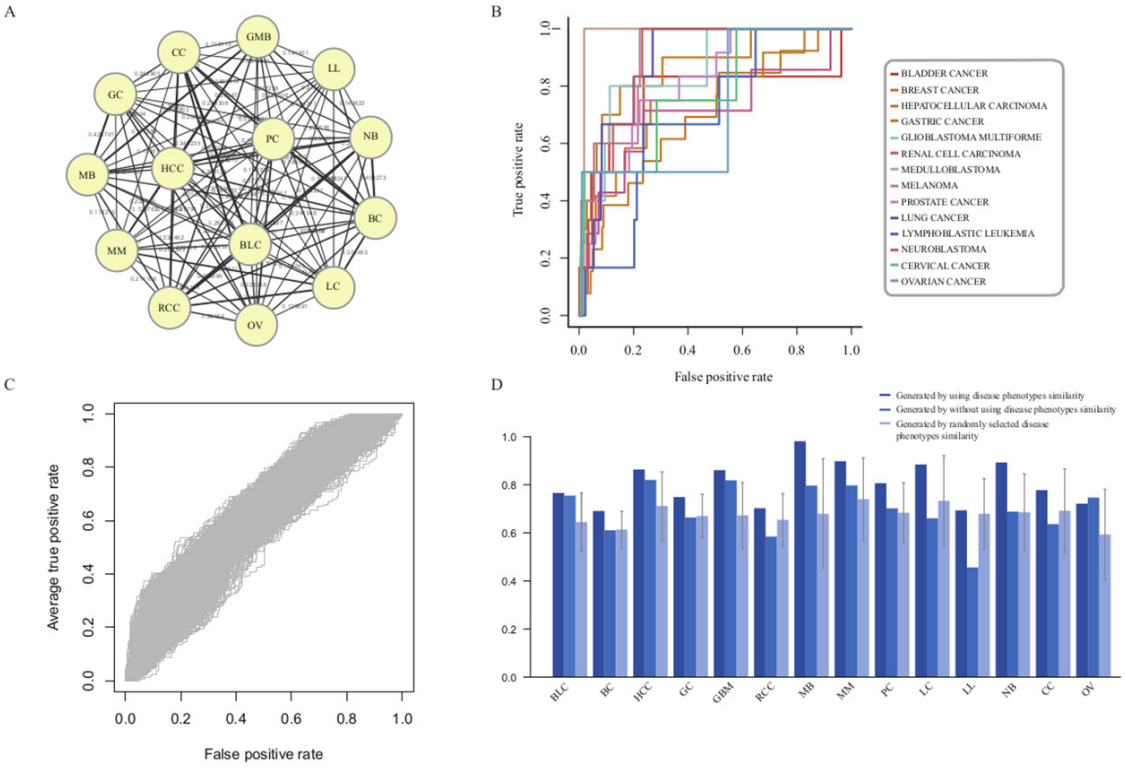

Figure 3: Evaluation of the influence of disease phenotype associations. A. Fourteen disease phenotype association network. **B.** The ROC curves of lncRNA prioritization results generated by excluding disease phenotype associations. **C.** The ROC curves generated by randomly selecting disease phenotype associations with 1000 repetitions. **D.** The comparison results of lncRNA prioritization generated by either using, excluding or permuting disease phenotype associations.

A case study of GBM

GBM is a highly aggressive brain cancer with extremely poor prognostic outcome despite intensive treatment regimes. Taking GBM as a case study, we used our approach to prioritize risk genes and lncRNAs associated with GBM. Through LOOCV, we found all known disease genes rank at the top of the candidate disease gene list, in which *IDH1, ERBB2*, and *TP53* are ranked at 1^{th}, 2^{nd} and 20^{th}, respectively (Supplementary Table 4). We then extracted the top 20 candidate disease genes and carried out function enrichment analysis using the remaining seventeen unknown disease genes by DAVID (Benjamini test, $p<0.01$) [38]. We found that these candidate genes were significantly enriched in cancer-related GO functions, including "regulation of cell proliferation", "regulation of apoptosis", "regulation of programmed cell death", "regulation of cell death" and "regulation of nucleocytoplasmic transport" (Figure 4A and Supplementary Table 5) and KEGG pathways, including "pathways in cancer", "melanoma", "bladder cancer", "non-small cell lung cancer" and "glioma". These results suggested that the seventeen candidate disease genes may play crucial roles in GBM tumorigenesis (Figure 4A and Supplementary Table 5).

For the lncRNA prioritization result, we extracted the top 5% of lncRNAs and found all known disease lncRNAs, *CDKN2B-AS1* and *H19*, in the candidate disease lncRNA list. To further validate GBM-related lncRNAs with high-confidence, we chose the top 20 lncRNAs and investigated their potential functions according to LncRNA2Function (Benjamini test, $p=0.01$, Supplementary Table 6). We found that the lncRNA-related genes were significantly enriched in GBM-related GO terms and KEGG pathways. The GO terms included "transmission of nerve impulse", "multicellular organismal signaling", "synaptic transmission", "cell-cell signaling", "neurological system process", "system process" and "nervous system development" etc. (Figure 4A). The KEGG enrichment analysis included "neuronal system", "transmission across chemical synapses", "neuroactive ligand-receptor interaction", "neurotransmitter release cycle" and "transmembrane transport of small molecules" pathways etc. (Figure 4A).

Next, we investigated whether these candidate lncRNAs were independent prognostic factors for survival. Towards this, we obtained two public expression datasets from the GEO database that contained 80 and 263 GBM samples (GSE7696 and GSE16011), respectively, and performed the same re-annotation process as described in Method. We found 2673 lncRNAs that were then subjected to survival analysis based on which we identified, three lncRNAs including *ENSG00000267519, ENSG00000255717* and *ENSG00000263731* that significantly correlated with survival of GBM (Figure 4B). Interestingly, high expression of the lncRNA gene *ENSG00000255717*, namely *SNHG1*, correlated with poor prognosis in both of the two datasets. Previously, Cao *et al.* identified abnormal expression of *SNHG1* in gastric cancer [39]. Our findings suggested that these potential lncRNAs may promote the development of GBM and could serve as novel prognostic markers for GBM, once verified.

DISCUSSION

Although a large number of lncRNAs have been identified in the human genome over the past decade

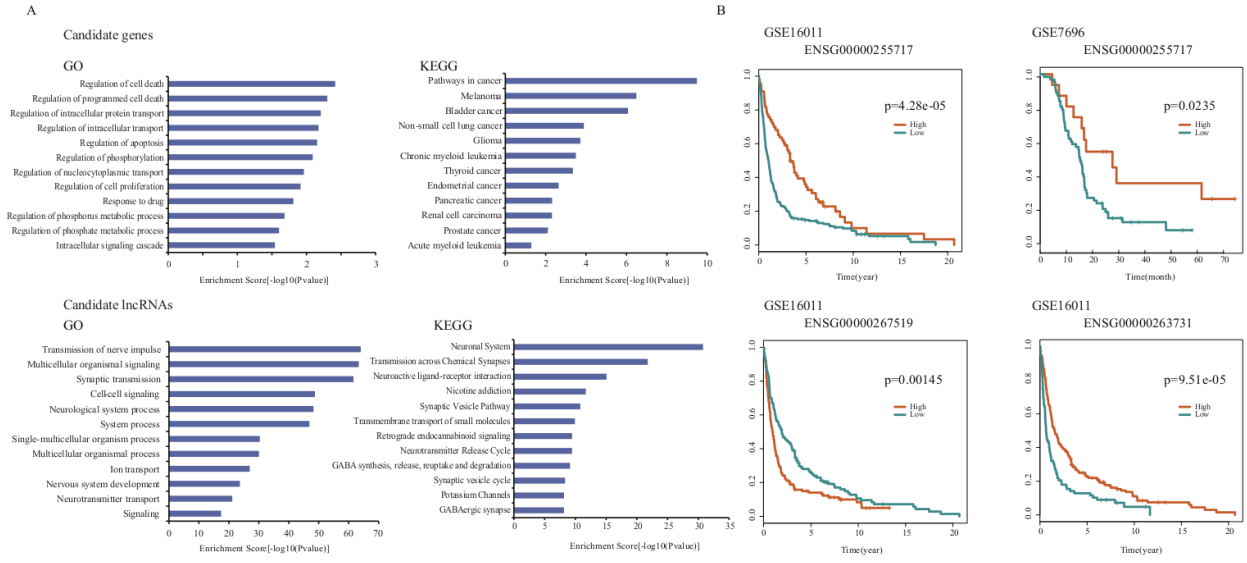

Figure 4: The prioritization results in the case study of GBM. A. The GO and KEGG enrichment analysis results for top 17 non-disease candidate genes and top 20 candidate lncRNAs of GBM. **B.** Survival analysis results of three candidate lncRNAs in GSE7696 and GSE16011.

[20–22, 27]. Only a few lncRNAs have been verified to be associated with diseases. How to integrate different biological datasets to accurately predict risk lncRNAs has become a critical issue for understanding disease mechanisms at the lncRNA level.

Based on the assumption that different diseases with similar phenotype associations involve similar molecular mechanisms, several studies have demonstrated that disease phenotype associations can help link different diseases with common genes and/or miRNAs [17, 18, 21, 40]. Disease phenotype associations have been widely used to benefit the systematic identification of disease-related protein-coding genes or miRNAs and facilitate in-depth understanding of their pathogenesis in human cancers. In this study, we developed a prioritization approach that was based on disease phenotype associations and systematically identified disease risk lncRNAs through integration of large-scale array-based expression datasets. Through collecting and re-annotating Affymetrix Human Exon 1.0 ST array datasets, we obtained 2818 samples with matched gene and lncRNA expressions in fourteen cancer types. We then constructed a GLCPN by using the pan-cancer datasets and found that our prioritization strategy was efficient in identifying candidate disease lncRNAs apart from being cost-effective. Our prioritization results showed that the top ranked lncRNAs or genes have high probabilities of being bona fide disease-related lncRNAs or genes.

The majority of disease candidate lncRNA prioritization approaches utilized known disease lncRNA information to predict disease and lncRNA associations [8–11, 13, 15]. However, only a few disease-related lncRNAs have been identified, and this limited information results in incomplete training sets during prioritization and hence can influence the performance in previous lncRNA prioritization approaches. Some other lncRNA prioritization approaches were designed by integration of other information, such as predictive and experimentally validated lncRNA-miRNA interactions [12, 14]. Such information provided the additional ability to measure the relationships between lncRNAs and diseases. Notably, the numbers of these known biological associations are relatively limited. LncRNA–miRNA interactions experimentally confirmed by molecular biological technologies in starBase v2.0 database refer to 1114 lncRNAs and 132 miRNAs. Such incomplete

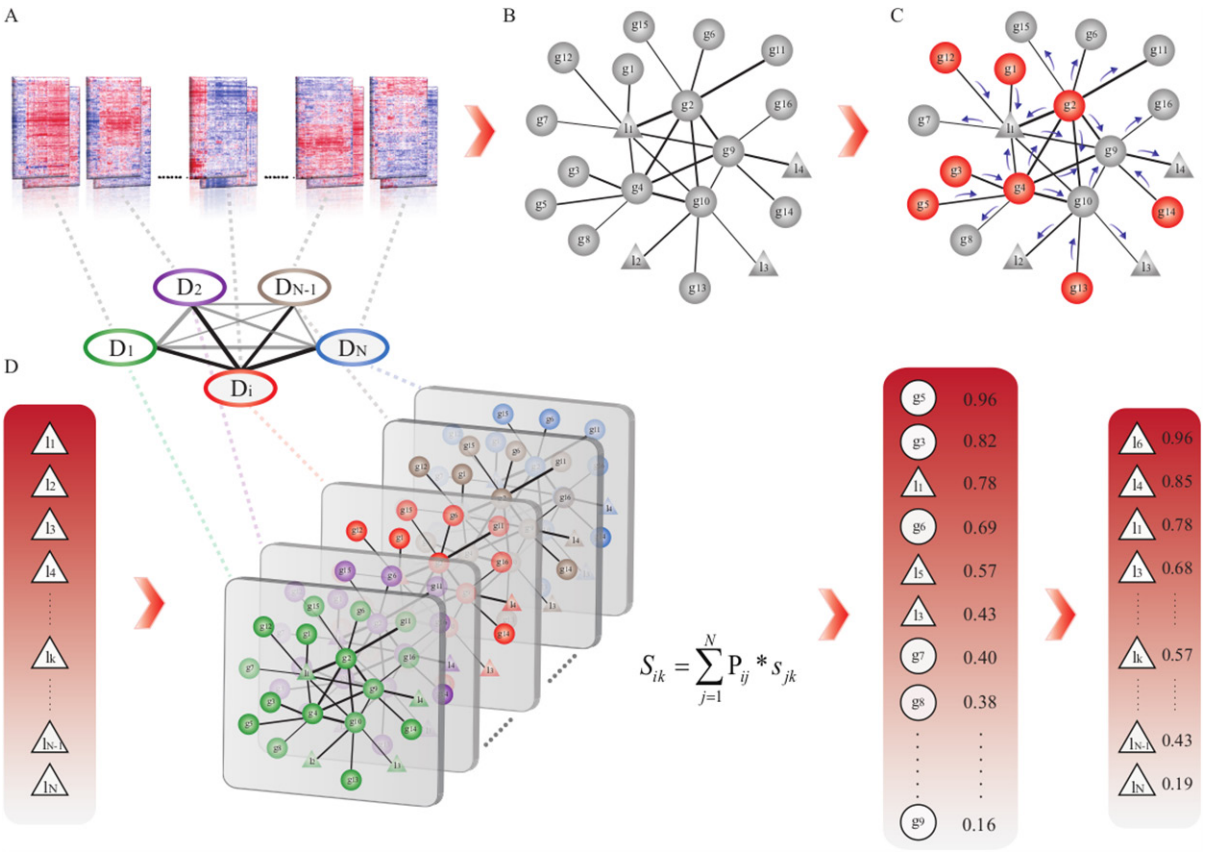

Figure 5: The workflow of prioritization of risk lncRNAs through integration of disease phenotype associations. A. Array-based expression data collection and re-annotation. **B.** Construction of the gene and lncRNA co-expression pan-cancer network (GLCPN). **C.** Application of the random walk method to predict scores for all candidates according to known disease genes. **D.** Integration of prediction scores by disease phenotype associations and generation of disease candidate lncRNA lists for prioritization.

information will limit the power in prioritizing or predicting lncRNAs that are potential associated with diseases. Relative to known disease lncRNAs and miRNAs, more disease protein-coding genes have been identified and confirmed. In contrast to previous methods [8–14], our method required only the knowledge of known disease genes for prioritization and did not depend on known disease-related lncRNAs. This enabled us to prioritize more comprehensive risk lncRNAs associated with a specific disease even in diseases without any known disease-related lncRNAs. Moreover, our approach was based on disease phenotype associations that can reduce the influence of the limited numbers of known disease genes and are effective in the prioritization of disease risk lncRNAs. The strategy utilized in our approach can help to advance the understanding of lncRNA function in cancer etiology. In summary, we presented an integrated prioritization approach for systematically prioritizing risk lncRNAs associated with human disease. This approach can be used to facilitate the identification of disease-related lncRNAs and to increase the understanding of lncRNA-mediated pathogenesis. Using our approach, we performed overall prioritization of the risk lncRNAs for fourteen cancer types, which provided testable hypotheses to guide further experiments.

MATERIALS AND METHODS

Array-based expression data collection and re-annotation

To satisfy the requirement that re-annotated lncRNAs have the broad coverage across the whole genome and microarray platforms designed by distinct expression studies are consistent, we selected "Affymetrix Human Exon 1.0 ST Array" as the research platform and collected this kind of expression datasets from GEO and ArrayExpress databases [41]. In order to ensure the sufficient sample size, we selected cancer studies with five or more disease samples to be considered and used for the further analysis. After widespread screening, forty-three studies with 2828 disease samples were identified. They were associated with fourteen cancer types namely, bladder cancer (BLC), breast cancer (BC), hepatocellular carcinoma (HCC), gastric cancer (GC), glioblastoma multiforme (GBM), renal cell carcinoma (RCC), medulloblastoma (MB), melanoma (MM), prostate cancer (PC), lung cancer (LC), lymphoblastic leukemia (LL), neuroblastoma (NB), cervical cancer (CC) and ovarian cancer (OC) (Supplementary Table 1).

Construction of lncRNA expression through re-annotation

We applied a custom pipeline to re-annotate Affymetrix Human Exon 1.0 ST Array taking advantage of its huge amount of probes annotated to thousands of lncRNAs. The probe sequences were downloaded from the manufacturer's website (http://www.affymetrix.com) and were then uniquely mapped to the human genome (hg19) by Bowtie without mismatch. Through using BEDTools (http://code.google.com/p/bedtools), probes completely falling into exons of lncRNAs but without overlapping with protein-coding genes were remained. Expression values of one lncRNA gene detected by at least four probes were averaged. All the expression data was log2-transformed and uniformly normalized by the quantile normalization approach. Finally, lncRNA expression datasets were constructed for all cancer types.

Construction of a GLCPN across pan-cancer datasets

Guilt by association implies that genes or lncRNAs with similar expression patterns under multiple experimental conditions have a high probability of sharing similar functions or being involved in common biological pathways [42, 43]. Therefore, we constructed a GLCPN from all the cancer datasets we had collected. We used the Pearson correlation coefficient to quantify the relations between or within the genes and the lncRNAs from the expression datasets in all the 14 cancer types. For each cancer type, gene-gene and gene-lncRNA pairs with co-expression scores greater than 0.8 and 0.7, respectively, were combined to formed a disease-specific network in each dataset as performed in previous studies [44–46]. Furthermore, we integrated all associated disease-specific networks into one GLCPN. The edge weights of GLCPN were assigned by the frequencies in the different cancer types. We further integrated protein-protein interaction (PPI) relationships (17649 nodes and 2079530 edges) obtained from a publicly available database STRING database [47] into the GLCPN. The normalized interaction probability scores obtained from the STRING database were considered as the weights.

Prioritization of risk lncRNAs and genes through integration of disease phenotype associations

To efficiently prioritize risk lncRNAs and genes in different cancer types based on the GLCPN, we applied the random walk method to calculate prediction score for all candidate lncRNAs and genes in each cancer type. By considering the known disease genes as seed nodes for any queried cancer type 'i' (Figure 5), we utilized the random walk method to compute prediction scores for each node (gene or lncRNA) in the GLCPN. By assuming that diverse diseases with phenotype associations show similar molecular mechanisms, we further combined disease phenotype similarity scores with the prediction scores of lncRNAs or genes into a unique prioritization score S_{ij} by:

$$S_{ik} = \sum_{j=1}^{N} P_{ij} * s_{jk}$$

Where P_{ij} represents the disease phenotype similarity score between cancer type i and j, and S_{jk} represents the corresponding prediction score for candidate lncRNA (or gene) k in cancer type j (Figure 5). Disease phenotype similarity scores were derived from the MimMiner tool, which calculates the correlation scores of 5080 known disease phenotypes through text mining analysis [48]. The candidate disease genes and lncRNAs were then ranked according to the prediction scores.

Evaluation of the robustness and the integration importance of our prioritization approach

We evaluated the performance of our prioritization approach by known disease lncRNAs and genes by using the ROC curve analysis, and the leave-one-out cross-validation (LOOCV) was carried out to assess the gene prioritization performance. Known causal genes and lncRNAs were extracted from the Online Mendelian Inheritance in Man (OMIM) [49] and the LncRNADisease database [50].

To evaluate the robustness and the integration importance of our prioritization approach, we accepted the evaluation strategies by leaving out or permuting relevant influence factors, included disease phenotype associations, the number of cancer types and known disease genes, and interrogated the changes in the prioritization results. Finally, we assessed the prediction performance of our prioritization approach in identifying disease-related lncRNAs and genes for each cancer type by only using information from other diseases.

Gene or lncRNA functional enrichment analysis

Functional enrichment analysis of candidate genes was performed by using the DAVID bioinformatics tool (http://david.abcc.ncifcrf.gov/conversion.jsp). For lncRNAs, we used LncRNA2Function (http://mlg.hit.edu.cn/lncrna2function) to perform function characterization [51]. All statistical analyses were performed using the R software package (http://www.r-project.org).

ACKNOWLEDGMENTS AND FUNDING

This work was supported by the National High Technology Research and Development Program of China [863 Program, Grant Nos. 2014AA021102], the National Program on Key Basic Research Project [973 Program, Grant Nos. 2014CB910504], the National Natural Science Foundation of China [Grant Nos. 91439117, 61473106, 61573122, 31601076], Wu lien-teh youth science fund project of Harbin medical university [Grant Nos. WLD-QN1407]. The Health Department Science Foundation of Heilongjiang Province (Grant Nos. 2013128), the Education Department Science Foundation of Heilongjiang Province (Grant Nos. 12541415), the Postdoctoral project of Heilongjiang Province (Grant Nos. LBH-Z14130), the Harbin Municipal Science and Technology Project (Grant Nos. RC2014QN003035).

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. Ota T, Suzuki Y, Nishikawa T, Otsuki T, Sugiyama T, Irie R, Wakamatsu A, Hayashi K, Sato H, Nagai K, Kimura K, Makita H, Sekine M, Obayashi M, Nishi T, Shibahara T, et al. Complete sequencing and characterization of 21,243 full-length human cDNAs. Nature genetics. 2004; 36:40-45.
2. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: functional surprises from the RNA world. Genes & development. 2009; 23:1494-1504.
3. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007; 447:799-816.
4. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: regulators of disease. The Journal of pathology. 2010; 220:126-139.
5. Esteller M. Non-coding RNAs in human disease. Nature reviews Genetics. 2011; 12:861-874.
6. Yuan J, Yue H, Zhang M, Luo J, Liu L, Wu W, Xiao T, Chen X, Chen X, Zhang D, Xing R, Tong X, Wu N, Zhao J, Lu Y, Guo M, et al. Transcriptional profiling analysis and functional prediction of long noncoding RNAs in cancer. Oncotarget. 2016; 7:8131-8142. doi: 10.18632/oncotarget.6993.
7. Zhao W, Luo J, Jiao S. Comprehensive characterization of cancer subtype associated long non-coding RNAs and their clinical implications. Scientific reports. 2014; 4:6591.
8. Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics. 2013; 29:2617-2624.
9. Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network. Molecular bioSystems. 2014; 10:2074-2081.
10. Li J, Gao C, Wang Y, Ma W, Tu J, Wang J, Chen Z, Kong W, Cui Q. A bioinformatics method for predicting long

noncoding RNAs associated with vascular disease. Science China Life sciences. 2014; 57:852-857.

11. Yang X, Gao L, Guo X, Shi X, Wu H, Song F, Wang B. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases. PloS one. 2014; 9:e87797.
12. Zhou M, Wang X, Li J, Hao D, Wang Z, Shi H, Han L, Zhou H, Sun J. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network. Molecular bioSystems. 2015; 11:760-769.
13. Zhao T, Xu J, Liu L, Bai J, Xu C, Xiao Y, Li X, Zhang L. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. Molecular bioSystems. 2015; 11:126-136.
14. Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Scientific reports. 2015; 5:13186.
15. Chen X, Yan CC, Luo C, Ji W, Zhang Y, Dai Q. Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. Scientific reports. 2015; 5:11338.
16. Chen X, Yan CC, Zhang X, You ZH, Deng L, Liu Y, Zhang Y, Dai Q. WBSMDA: Within and Between Score for MiRNA-Disease Association prediction. Scientific reports. 2016; 6:21106.
17. Xu C, Ping Y, Li X, Zhao H, Wang L, Fan H, Xiao Y, Li X. Prioritizing candidate disease miRNAs by integrating phenotype associations of multiple diseases with matched miRNA and mRNA expression profiles. Molecular bioSystems. 2014; 10:2800-2809.
18. Xiao Y, Xu C, Ping Y, Guan J, Fan H, Li Y, Li X. Differential expression pattern-based prioritization of candidate genes through integrating disease-specific expression data. Genomics. 2011; 98:64-71.
19. Gellert P, Ponomareva Y, Braun T, Uchida S. Noncoder: a web interface for exon array-based detection of long non-coding RNAs. Nucleic acids research. 2013; 41:e20.
20. Michelhaugh SK, Lipovich L, Blythe J, Jia H, Kapatos G, Bannon MJ. Mining Affymetrix microarray data for long non-coding RNAs: altered expression in the nucleus accumbens of heroin abusers. Journal of neurochemistry. 2011; 116:459-466.
21. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H, Skogerbo G, Wu Z, Zhao Y. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. Nucleic acids research. 2011; 39:3864-3878.
22. Mercer TR, Dinger ME, Sunkin SM, Mehler MF, Mattick JS. Specific expression of long noncoding RNAs in the mouse brain. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:716-721.
23. Johnson R. Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiology of disease. 2012; 46:245-254.
24. Zhang X, Sun S, Pu JK, Tsang AC, Lee D, Man VO, Lui WM, Wong ST, Leung GK. Long non-coding RNA expression profiles predict clinical phenotypes in glioma. Neurobiology of disease. 2012; 48:1-8.
25. Du Z, Fei T, Verhaak RG, Su Z, Zhang Y, Brown M, Chen Y, Liu XS. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. Nature structural & molecular biology. 2013; 20:908-913.
26. Liu MX, Chen X, Chen G, Cui QH, Yan GY. A computational framework to infer human disease-associated long noncoding RNAs. PloS one. 2014; 9:e84408.
27. Tang JY, Lee JC, Chang YT, Hou MF, Huang HW, Liaw CC, Chang HW. Long noncoding RNAs-related diseases, cancers, and drugs. TheScientificWorldJournal. 2013; 2013:943539.
28. Guo X, Gao L, Liao Q, Xiao H, Ma X, Yang X, Luo H, Zhao G, Bu D, Jiao F, Shao Q, Chen R, Zhao Y. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic acids research. 2013; 41:e35.
29. Jacobs JF, van Bokhoven H, van Leeuwen FN, Hulsbergen-van de Kaa CA, de Vries IJ, Adema GJ, Hoogerbrugge PM, de Brouwer AP. Regulation of MYCN expression in human neuroblastoma cells. BMC cancer. 2009; 9:239.
30. Pasmant E, Sabbagh A, Vidaud M, Bieche I. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. FASEB journal. 2011; 25:444-448.
31. Dallosso AR, Hancock AL, Malik S, Salpekar A, King-Underwood L, Pritchard-Jones K, Peters J, Moorwood K, Ward A, Malik KT, Brown KW. Alternately spliced WT1 antisense transcripts interact with WT1 sense RNA and show epigenetic and splicing defects in cancer. Rna. 2007; 13:2287-2299.
32. Yang JM, Chen WS, Liu ZP, Luo YH, Liu WW. Effects of insulin-like growth factors-IR and -IIR antisense gene transfection on the biological behaviors of SMMC-7721 human hepatoma cells. Journal of gastroenterology and hepatology. 2003; 18:296-301.
33. Martens-Uzunova ES, Bottcher R, Croce CM, Jenster G, Visakorpi T, Calin GA. Long noncoding RNA in prostate, bladder, and kidney cancer. European urology. 2014; 65:1140-1151.
34. Cai B, Song XQ, Cai JP, Zhang S. HOTAIR: a cancer-related long non-coding RNA. Neoplasma. 2014; 61:379-391.
35. Hajjari M, Salavaty A. HOTAIR: an oncogenic long non-coding RNA in different cancers. Cancer biology & medicine. 2015; 12:1-9.
36. Matouk IJ, DeGroot N, Mezan S, Ayesh S, Abu-lail R, Hochberg A, Galun E. The H19 non-coding RNA is essential for human tumor growth. PloS one. 2007; 2:e845.

37. Yoshimizu T, Miroglio A, Ripoche MA, Gabory A, Vernucci M, Riccio A, Colnot S, Godard C, Terris B, Jammes H, Dandolo L. The H19 locus acts *in vivo* as a tumor suppressor. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:12417-12422.

38. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology. 2003; 4:P3.

39. Cao WJ, Wu HL, He BS, Zhang YS, Zhang ZY. Analysis of long non-coding RNA expression profiles in gastric cancer. World journal of gastroenterology. 2013; 19:3658-3664.

40. Xu W, Jiang X, Hu X, Li G. Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization. BMC medical genomics. 2014; 7:S1.

41. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, et al. ArrayExpress update—simplifying data submissions. Nucleic acids research. 2015; 43:D1113-1116.

42. Lee HK, Hsu AK, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome research. 2004; 14:1085-1094.

43. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America. 1998; 95:14863-14868.

44. Saviozzi S, Ceppi P, Novello S, Ghio P, Lo Iacono M, Borasio P, Cambieri A, Volante M, Papotti M, Calogero RA, Scagliotti GV. Non-small cell lung cancer exhibits transcript overexpression of genes associated with homologous recombination and DNA replication pathways. Cancer research. 2009; 69:3390-3396.

45. Jordan IK, Marino-Ramirez L, Wolf YI, Koonin EV. Conservation and coevolution in the scale-free human gene coexpression network. Molecular biology and evolution. 2004; 21:2058-2070.

46. Gatti DM, Zhao N, Chesler EJ, Bradford BU, Shabalin AA, Yordanova R, Lu L, Rusyn I. Sex-specific gene expression in the BXD mouse liver. Physiological genomics. 2010; 42:456-468.

47. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2015; 43:D447-452.

48. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. European journal of human genetics. 2006; 14:535-542.

49. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic acids research. 2015; 43:D789-798.

50. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, Zhang Q, Yan G, Cui Q. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research. 2013; 41:D983-986.

51. Jiang Q, Ma R, Wang J, Wu X, Jin S, Peng J, Tan R, Zhang T, Li Y, Wang Y. LncRNA2Function: a comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. BMC genomics. 2015; 16:S2.