# Prioritizing Cancer lncRNA Modulators via Integrated lncRNA-mRNA Network and Somatic Mutation Data

**Dianshuang Zhou**
Harbin Medical University https://orcid.org/0000-0002-9205-4641
**Xin Li**
Harbin Medical University
**Shipeng Shang**
Harbin Medical University
**Hui Zhi**
Harbin Medical University
**Peng Wang**
Harbin Medical University
**Yue Gao**
Harbin Medical University
**Shangwei Ning** ( ningsw@ems.hrbmu.edu.cn )
Harbin Medical University

**Research Article**

**Keywords:** computational framework, lncRNA-mRNA network, mutations, cancer lncRNA modulators

**Posted Date:** June 22nd, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-636414/v1



**Version of Record:** A version of this preprint was published at Current Bioinformatics on April 21st, 2022.
See the published version at https://doi.org/10.2174/1574893617666220421095601.

# Prioritizing cancer lncRNA modulators via integrated lncRNA-mRNA network and somatic mutation data

Dianshuang Zhou[1+], Xin Li[1], Shipeng Shang[1], Hui Zhi[1], Peng Wang[1], Yue Gao[1], Shangwei Ning[1]*

[1] College of Bioinformatics Science and Technology, Harbin Medical University, 194 Xuefu Road, Harbin 150081, China

# ABSTRACT:

**Background:** Long noncoding RNAs (LncRNAs) represent a large category of functional RNA molecules that play a significant role in human cancers. lncRNAs can be genes modulators to affect the biological process of multiple cancers.

**Methods:** Here, we developed a computational framework that uses lncRNA-mRNA network and mutations in individual genes of 9 cancers from TCGA to prioritize cancer lncRNA modulators. Our method screened risky cancer lncRNA regulators based on integrated multiple lncRNA functional networks and 3 calculation methods in network.

**Results:** Validation analyses revealed that our method was more effective than prioritization based on a single lncRNA network. This method showed high predictive performance and the highest ROC score was 0.836 in breast cancer. It's worth noting that we found that 5 lncRNAs scores were abnormally high and these lncRNAs appeared in 9 cancers. By consulting the literatures, these 5 lncRNAs were experimentally supported lncRNAs. Analyses of prioritizing lncRNAs reveal that these lncRNAs are enriched in various cancer-related biological processes and pathways.

**Conclusions:** Together, these results demonstrated the ability of this method identifying candidate lncRNA molecules and improved insights into the pathogenesis of cancer.



**Keywords:** computational framework, lncRNA-mRNA network, mutations, cancer lncRNA modulators,

# Introduction:

Cancer is a major public health problem across the world and is a leading cause of death in many countries(1). Cancer is a complex disease involving DNA abnormalities, transcriptomic alterations and epigenetic aberrations(2) and whole genome sequencing efforts have uncovered the genomic landscapes of common forms of human cancers(3). The cancer Genome atlas (TCGA) has provided a mass of data of human samples and discover molecular alterations at the DNA and RNA levels(4).

Mutations are important markers of cancer genes and the somatic mutation landscapes and signatures of major cancer types have been reported and stockpiled by international cancer genome projects, such as TCGA and ICGC(5). In recent years, increasing experimentally supported evidence has suggested that lncRNAs as genes modulators affect the process of cancers. For instance, in lung cancer, MALAT1 actively regulates a set of metastasis-associated genes expression, including MIA2, HNF4G and CA2. Moreover, MALAT1 can be used as a valuable prognostic marker and a promising therapeutic target(6). TUG1 can regulate the expression of LIMK2b and then promoted cell growth and chemoresistance of small cell lung cancer(7). NEAT1 transforms the epigenetic landscape of target gene promoters to facilitate transcription and promotes carcinogenic growth(8). These data indicate that researching lncRNAs is an important step in the understanding of cancer mechanisms. However, there is currently no systematic way to explore the lncRNAs which can be genes modulators in multiple cancers.

Network analysis is often used to explore the function of lncRNAs and the relationship between lncRNAs and diseases. For instance, Guo et al. developed a bi-colored network to annotate lncRNA function(9). Currently, Data sets obtained from many lncRNA-related databases can be used for in-depth exploration of lncRNA. TANRIC platform systematically collected data resource that records the expression of lncRNA in 20 human cancers(10). starBase v2.0, RAID V2.0 and NPInter V3.0 stored a huge amount of lncRNA data resources, including lncRNA-mRNA and lncRNA-miRNA interactions(11-13). LncRNADisease and Lnc2Cancer databases manually curated credible lncRNA-disease or lncRNA-caner data, meanwhile, the dysfunction pattern of lncRNAs were annotated (14, 15). However, our knowledge of using these public data to build a network of lncRNAs to further explore cancer lncRNA modulators remains limited.

Here we exploit a cancer lncRNA prioritization computational method based on a lncRNA-mRNA function network and somatic mutation data for 9 cancers (BLCA, BRCA, COAD, GBM, LIHC, LUAD, OV, PRAD, STAD) from TCGA. This study systematically analyzes the result of method, including topological property of our functional network, the perform of method and functional enrichment analysis and biological characteristics of risky cancer lncRNA modulators. We found that most of top ranking lncRNA modulators have been identified cancer lncRNAs by lnc2Cancer database, then these lncRNA modulators also enriched in cancer-related Gene Ontology (GO) terms and pathways. Our method can identify the lncRNA modulators which cannot be identified by differential expression and can be a valuable complement to

experimental studies used in future studies.

# MATERIALS AND METHODS

## Data sources

### Cancer somatic mutation data and golden standard

We used 9 types of cancer somatic mutation data from TCGA: Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Glioblastoma multiforme (GBM), Liver hepatocellular carcinoma (LIHC), Lung adenocarcinoma (LUAD), Ovarian serous cystadenocarcinoma (OV), Stomach adenocarcinoma (STAD).

The golden standard of gene set was obtained from Cancer Genome Census(CGC) database, which includes 616 cancer genes(16). In order to verify the accuracy of the model, we choose the 9 cancers with the most experimentally confirmed data in the Lnc2Cancer database, which includes 148 lncRNAs for 9 cancers.

### The lncRNA-mRNA functional network

In our study, the lncRNA-mRNA functional network was a fusion of lncRNA-mRNA co-expression network, lncRNA-mRNA ceRNA network and lncRNA-protein interaction network. We have used gene expression data from TCGA and lncRNA expression data from TANRIC. Pearson correlation was used to construct our lncRNA-mRNA co-expression network and the lncRNA-mRNA pair was selected if it meets following criteria: corr (lncRNA, mRNA) > 0.8, fdr < 0.05. We used lncRNA-miRNA pairs and mRNA-miRNA pairs to construct our lncRNA-mRNA ceRNA network. lncRNA-miRNA interaction data and mRNA-miRNA interaction data were downloaded from starBase v2.0, NPInter v3.0 and RAID v2.0. According to number of

miRNA shared with lncRNA and mRNA, we used hypergeometric distribution to construct our lncRNA-mRNA ceRNA network. For lncRNA-protein interaction network，we also downloaded lncRNA-protein interaction data from starBase v2.0, NPInter v3.0 and RAID v2.0. The weight of lncRNA-mRNA ceRNA network and lncRNA-protein interaction network was defined as 1. Then we coalesced 3 network and defined the weight of lncRNA-mRNA pairs as (corrlm + Wcelm + Wlplm)/3, where corrlm, Wcelm, Wlplm indicated Pearson correlations for lncRNA-mRNA, the weight in ceRNA network and the weight in interaction network, respectively.

## Scoring scheme of genes

We defined $N_i$ for the number of non-synonymous mutations of a gene from the somatic mutation data. Meanwhile, we screened the differential expression genes according to the gene expression data and got the *P-value* of Student's t-test of each gene. The *P-value* was transformed by a standardization method which named *scale function* in R:

$$np_i = 1 - {}^{1}/_{1 + e^{scale(-log10(p_i))}}$$

Where $np_i$ represents the normalization score of differential expression gene(i), $p_i$ represents the P-value of differential expression gene(i).

We formulated the score of gene to use mutation occurrences and $np_i$:

$$G_i = M_i * (1 + np_i)$$

Where $G_i$ represents the final score of gene(i), $M_i$ represents the mutational occurrences of gene(i).

## Scoring scheme of lncRNAs

We designed three different ways to use gene's score of direct neighbors and edge weights in our network for prioritizing lncRNAs: the first computational method named

"Smax" was defined that lncRNA's score is the biggest score of its direct gene's score multiply by edge weights:

$$L_j = max(G_i * w(i,j))$$

the second computational method named "Ssum" was defined that lncRNA's score is the sum score of its direct gene's score multiply by edge weights:

$$L_j = \sum_{i=1}^{N} G_i * w(i,j)$$

Where $L_i$ represents the score of lncRNA(j), $w(i,j)$ represents the edge weights of lncRNA(j) and gene(i).

the third computational method named "NWsum" was defined that lncRNA's score is the sum score of its direct gene's score divide by the number of gene's direct neighbors:

$$L_j = \sum_{i=1}^{N} \frac{G_i}{n_i}$$

Where $n_i$ represents the number of gene(i)'s direct neighbors.

For example, if a lncRNA had 5 direct neighbors and the number of 5 genes' direct were 2,4,1,1,3, we can obtain the score for this lncRNA:

$$\mathrm{L} = \frac{G(1)}{2} + \frac{G(2)}{4} + \frac{G(3)}{1} + \frac{G(4)}{1} + \frac{G(5)}{3}$$

## A summary of validated cancer-related lncRNAs

Literature mining is an effective way to collect "gold-standard" for a large number of disease-related molecules because of experimental methods, such as Western blot, Luciferase reporter assay. In this study, we used a set of validated lncRNAs for 9 cancers from Lnc2Cancer database(http://www.bio-bigdata.net/lnc2cancer/), which contains cancer-related lncRNAs based on experiment by thousands of articles. Due to the emergence of a lot of new data after the database published, we added the latest data

to test our method through manually collecting lncRNA-cancer associations.

## Calculation of network topology property and survival analysis

Degree centrality and betweenness centrality are two important indicators in the nature of network topology. Generally, the larger the node degree of a node is, the higher the degree of centrality of the node is, and the more important the node is in the network; betweenness centrality is equal to the number of shortest paths from each node to all others that pass through this node, as an important global geometric quantity, betweenness centrality reflects the role and influence of the corresponding node in the entire network. Degree centrality and betweenness centrality was calculated using the R package “igraph” (17).

A Kaplan-Meier survival analysis was performed using the clinical data from TCGA, and statistical significance was assessed using the log-rank test. The survival curve was drawn using the R package "survival". All analyses were performed on the R 3.6.0 framework.

# Result

## Overview

A general workflow of method is given in Fig. 1. To prioritize lncRNA molecules, the first step was to score genes based on somatic mutation data and gene expression data. The second step was to integrate three lncRNA-mRNA networks: lncRNA-mRNA co-expression, lncRNA-mRNA ceRNA network, lncRNA-protein interaction network. The third step was to score lncRNAs based on genes' score and integrated lncRNA-mRNA network. We used three methods to score lncRNAs and then sorted them according to lncRNAs' score. The higher the ranking, the more likely it is to become a risk cancer lncRNA.

**Global properties of the lncRNA-mRNA function network**

There was an average of 378768 edges, including 19883 coding genes and 12150 lncRNAs in our network (Fig. 2A). In this network, we first calculated the degree centrality and found a small number of nodes had very high degree. The degree distribution of the network obeys a power law distribution (Fig. 2B). Then we compared the degree and betweenness (Fig. 2C) of cancer lncRNA with candidate lncRNA. The degree and betweenness of cancer lncRNA were significantly higher than the candidate lncRNA and the P-value by Wilcoxon test was less than 0.001. These results indicated "cancer lncRNA nodes" is a key factor in the network and plays a regulatory role for a large number of genes, which is consistent with the previous research results(18).

## Performance of method

First, we use genes' score which were calculated based on somatic mutation data and gene expression data for 9 cancers from TCGA and a protein–protein interaction network which named STRING v10 to appraise the method when used to prioritizing cancer genes. The golden standard was a high-confidence gene set form the Cancer Genome Census database (CGC), including 616 cancer driver genes.

To assess the performance of our method, ROC analysis was executed for each type of cancer and the AUC value used to determine the quality of the method. For instance, the AUC values which were calculated based on 3 methods and cancer genes annotated by CGC were 0.909, 0.896, 0.843, respectively (Fig. 3A). This indicated that our method showed high predictive performance when using PPI network and mutated genes to predict driving genes.

For prioritizing cancer lncRNA, we used genes' score and lncRNA-mRNA functional network. To evaluate the performance of our method, we select the experimentally supported cancer lncRNAs as the golden standard from Lnc2Cancer database. We first compared the values of AUC using a single subnet with the integrated network, finding that the values of AUC which used lncRNA-mRNA co-expression network significantly below the values of AUC which used lncRNA-mRNA ceRNA network and lncRNA-protein network. Meanwhile, the values of AUC which used integrated network showed the best performance (Fig. 3B). This result showed that using a single network for scoring lncRNA may not be the best method and consolidated multiple networks can improve prediction performance. We also compared the performance of the three methods, finding that all three methods show high predictive

performance. The median values of AUC for 3 methods were more than 0.77. NWsum algorithm got the highest AUC values and the median values of AUC was 0.822 (Fig. 3C).

The robustness analysis was performed by deleting some network nodes. We randomly deleted 10% and 20% of the network nodes and their connected edges. Using incomplete networks and NWsum algorithm, we recalculate the AUC values, finding that the AUC values of all cancer had no prominent change. The most variable cancer was Glioblastoma multiforme (GBM), however, the level of change did not exceed 10%. These results showed that although the absence of the network would have a certain impact on the performance of the method, the effect was small (Fig. 3D). This was because that the key nodes in the network had very high degrees and betweenness, even if a part of the network was missing, it would not affect the result. All the results indicated our network was robust.

## Analysis of high-risk cancer lncRNA modulators

Duo to the performance of NWsum algorithm was slightly stronger than the other two algorithms, we use the lncRNA predicted by the NWsum algorithm to analyze.

**Top-rank lncRNAs consist with known cancer lncRNAs**

Currently, there have been some studies of cancer related lncRNAs by experimental methods, although the number is not large, it is still the most powerful standard to verify the performance of our method. According to the scores calculated by the NWsum algorithm, we found that 5 lncRNAs scores were abnormally high and the top 5 lncRNAs which were MALAT1, NEAT1, FENDRR, CRNDE, TUG1 all appeared in 9 cancers. By consulting the literatures, these five lncRNAs were experimentally

supported lncRNAs and associated with the development of multiple cancers (Supplementary Table S1). LncRNA MALAT1 had been confirmed with all cancers we researched and many pathogenic mechanisms had been described, including influencing the metastasis phenotype of lung cancer cells(6), inducing epithelial-to-mesenchymal transition duo to the dysfunction of MALAT1 in bladder cancer(19), promoting tumor growth in colorectal cancer(20) and so on. So far, the relationships between LncRNA FENDRR and cancers were not so much. It had been confirmed that decreased expression of FENDRR regulates gastric cancer cell metastasis(21) and FENDRR as a tumor suppressor gene in NSCLC inhibits cell proliferation and induces apoptosis. Meanwhile, researcher found FENDRR was able to regulate heart and body wall development in the mouse(22). So lncRNAs FENDRR may prove to be risk factors for other cancers in future studies.

## Common character and special character cancer lncRNA modulators

Previous studies have shown that some lncRNAs function in many types of cancer, and some lncRNAs have certain tumor specificities. We chose the top 20 lncRNAs and their corresponding scores (excluding the 5 lncRNAs mentioned above) and standardized scores for each cancer. At last, we got 75 lncRNAs and their standardized scores for 9 cancers. Based on the score of these lncRNAs, a global risk-evaluation score profile was constructed and clustered by the k-means method (Fig.4A). 18 lncRNAs scored higher in 9 cancers and 13 lncRNAs only scored higher in colon adenocarcinoma (COAD). To characterize the function of these lncRNAs, a gene set enrichment analysis web server called “Enrichr” was used for pathway annotations, including Gene

Ontology biological process, KEGG pathway, Reactome pathway, WikiPathways(23). Common character lncRNAs were annotated the pathways which associated with multiple cancers. The biological process of cell (mitochondrion localization and apoptotic cell clearance), known cancer pathway (WNT pathway) and cancer-related protein (SMO and GTP) were highly represented. To the special character cancer lncRNAs, BMP signaling pathway which was a developmental pathway and a potential therapeutic target was been enriched (Fig.4B). Specifically, these lncRNAs enriched to the corresponding diseases (Neoplasm of the colon, intestinal polyposis and Neoplasm of the rectum) by using a tool named Human Phenotype Ontology (Fig.4C).

In Common character lncRNAs, 5 lncRNAs (ZNF518A, HCG18, FGD5-AS1, TSIX, RP11-553L6.5) were not verified by literatures to this day. We extracted the interaction genes and miRNAs of these 5 lncRNAs and compared with the confirmed evidence. 399 miRNAs and 2855 genes were extracted, respectively. The experimentally supported miRNAs for 9 cancers were from HMDD 2.0 database(24) and genes for 9 cancers were from Cancer Genome Census database (CGC). The number of experimentally supported miRNAs and genes were 134 and 616 (Fig.4D). We found 96 miRNAs and 166 genes exhibited both interaction of association and confirmed evidence, including miR-106a, let-7b, FUS, EWS, etc. (Fig.4E) High overlap indicated they had a significant association with 9 cancers ($P < 0.001$, $P < 0.001$). We thus speculated that these 5 lncRNAs were likely to be a high-risk clinical factor and further experiments need to be carried out.

## Biological characteristics of prioritizing lncRNA modulators

lncRNAs are diverse and involved in a variety of biological processes. In the process of transcription, some special features of lncRNA are formed and these features are closely linked with the function of lncRNA. We divided the top 2000 lncRNAs into 3 levels (Lv1: 1~50, Lv2: 51~500, Lv3: 501~2000) to observe the differences in biological characteristics. Previous studies have shown that exons are one of the important genomic characteristics of lncRNAs(25). The average number of exons for Lv1 was 1.74-fold that of Lv2(Wilcoxon test, $P = 0.039$) and the average number of exons for Lv2 was 1.78-fold that of Lv3($P < 0.001$) (Fig.5A). One of the important functions of lncRNA is its ability which bind to miRNA competitively to regulate gene expression. The number of miRNA binding sites on lncRNA reflects the ability of lncRNAs to bind miRNAs. The density of miRNA binding sites was computed by two prediction methods: miRanda and TargetScan. The average number of miRNA binding sites for Lv1 was 1.42-fold that of Lv2 ($P < 0.001$) and the average number of miRNA binding sites for Lv2 was 1.32-fold that of Lv3 ($P < 0.001$) (Fig.5B). Previous research shows some specific human lncRNAs which different evolutionary conservation beyond primates but have proven to be both functional and therapeutically relevant(26). The UCSC phyloP score was used to calculate the conservation score of lncRNAs. Duo to the conserved state belongs to the corresponding coding gene instead of lncRNA, we excluded antisense lncRNAs when we calculated the conservation scores. The average scores for Lv1 were 3.2-fold that of Lv2 ($P = 0.003$) and the average scores for Lv2 was 1.42-fold that of Lv3 ($P = 0.013$) (Fig.5C). Some famous lncRNA, such as HOTAIR and PCAT7, were significantly upregulated in various cancers. We got the lncRNA

expression profile from TANRIC database and calculated the expression values of 3 levels. The average expression values for Lv1 were 13.2-fold that of Lv2 ($P < 0.001$) and the average expression values for Lv2 was 3.29-fold that of Lv3 ($P < 0.001$) (Fig.5D). In general, top lncRNAs have more significant biological characteristics than others candidate lncRNAs.

## Case study: Stomach Cancer

Stomach cancer is one of the most common cancers in the digestive system and the second most common cause of cancer death in the world. Researchers found some lncRNAs affected the occurrence and development of stomach cancer. Hao Li et.al. presented overexpression of H19 promoted the features of GC including proliferation, migration, invasion and metastasis(27). Yongchao Liu et.al. presented lncRNA GAS5/YBX1/p21 pathway may become a useful therapeutic method since the YBX1 protein level was reduced by down-regulation of GAS5 and decreased p21 expression, thus abolished G1 phase cell cycle arrest in stomach cancer(28). So, we propose a case study of stomach cancer to investigate whether our methods can find lncRNA molecules of stomach cancer. We selected the top 20 lncRNAs for sample analysis.

First, we compared the differential expression lncRNAs in samples of stomach cancer by using Student's t-test with the top 20 lncRNAs which we selected (Supplementary Table S2). Thirteen of the top 20 lncRNAs were differential expression. LncRNA TINCR which cannot be identified by differential expression was identified by our method. This indicated our method could identify cancer lncRNAs that were not identified by differential expression method.

For the top 20 lncRNAs that were not be verified by experiment, we use the way of literature retrieval to mine their functions one by one. ZNF518A was recorded in top 20 Co-Expressed gene's Protein-protein interaction Network Plot of Cancer Cell Metabolism Gene DB(29). LINC00657 could suppresses hepatocellular carcinoma cell growth and breast cancer cell growth(30, 31). HCG18 may serve as a potential biomarker in breast cancer and lung cancer and as an immune-related lncRNA to affect the survival of glioma patients(32, 33). XIST and TSIX, as cancer-immune biomarkers, manipulated PD-L1 expression in BC cell lines, leading to lymph node metastasis. For other lncRNAs that were not be verified by experiment, there was currently no content related to tumor or tumor pathogenesis. They may become new directions for future research.

To verify whether the top 20 lncRNAs are closely related to gastric cancer, a functional annotation of these lncRNAs was carried out by “Enrichr” web server. The results manifested these lncRNAs enriched many critical biological functions and pathways, including S33 mutants of beta-catenin, Beta-catenin phosphorylation cascade, Toll Like Receptor signaling, TGF beta signaling pathway, etc. The result of functional annotation proved these lncRNAs affected the occurrence and development of stomach cancer.

Furthermore, we investigated whether top 20 lncRNAs can act as an independent prognostic factor. Clinical information for stomach Cancer was obtained from TCGA. We used lncRNA expression data and clinical information to performed a Kaplan–Meier survival analysis. Then we obtained 2 lncRNAs which significantly related to the

patient's survival (NEAT1, P = 0.0309; LINC00657, P = 0.0296). Our results displayed that high or low expression of lncRNA could significantly divide patients into two categories based on survival time. Drawing a comparison between the miRNAs and genes which were interaction with the 2 lncRNAs and stomach cancer related miRNAs and genes, many experimentally confirmed gastric cancer miRNAs and genes appear in miRNAs and genes that interact with the 2 lncRNAs ($P < 0.001$). The stomach cancer miRNAs and genes linked by the 2 lncRNAs are displayed using Cytoscape (34).

In summary, our method can accurately identify stomach cancer-associated lncRNA molecules, furthermore, our method can identify lncRNAs which can’t be identified by the method of differential expression.

# Discussion

lncRNA is a regulator of gene expression to act the translation of genetic codons into protein sequences (35). In previous studies, researchers were more inclined to develop methods which explore driver coding genes for cancers, such as PolyPhen-2, MutSigCV and HotNet2. However, lncRNA as a key factor that can influence tumor initiation, growth and metastasis of cancer cells (36), a few methods which prioritize candidate cancer lncRNA molecules were exploited. At the same time, there were currently not many data for detecting lncRNA mutation sites, so we attempted to use the data from somatic mutations of TCGA to screen for cancer risk lncRNAs. We first utilized expression data and public lncRNA, miRNA, gene interaction data to construct our lncRNA-mRNA network. Then the gene's score was obtained by combining the occurrences of somatic mutations and the differential expression level. Here, we present an approach for identifying cancer lncRNA molecules using our lncRNA-mRNA functional network and three computing method in direct neighbors for nine cancers, providing a view for predicting cancer lncRNAs.

Using a co-expression network alone could not obtain good predictive results. We found that the correlation coefficient of co-expression between most experiments supported lncRNA and gene were not very high. In fact, lncRNA-mRNA co-expression does not mean that the two have a regulatory relationship. When we choose a higher Pearson correlation coefficient, the experimentally supported lncRNAs will be filtered out in lncRNA-mRNA co-expression network, so we join the lncRNA-mRNA ceRNA network and the lncRNA-protein interaction network. The ROC analysis results

indicated using lncRNA-protein interaction network could get the best result in three networks. This shows that the experimentally supported interaction network has higher accuracy than the network obtained by the calculation method. In our integrated network, the degree and median of experimentally supported cancer lncRNAs were significantly higher than other lncRNAs.

In our predicted results, in addition to using ROC analysis to evaluate the overall prediction results, we also chose the top ranked lncRNA for analysis. By comparing biological characteristics, lncRNA ranks and characteristic scores had a consistent trend. For more accurate analysis, we identified 5 lncRNAs (MALAT1, NEAT1, FENDRR, CRNDE, TUG1) which ranked in the top 5 of 9 cancers. These 5 lncRNAs had been documented in many literatures and were closely related to various cancers. In STAD study, we found this method make accurate and complement lncRNA found by differential expression as the high false positives and missing parts cancer lncRNA of the differential expression method. A mass of experimentally supported cancer genes and cancer miRNAs which connect with our predicted lncRNA also show the accuracy of our method.

Currently, experimentally supported cancer lncRNAs were small number. Some lncRNAs have been confirmed in common cancers only. Using golden standard to evaluate predictions is an effective method, so we select 9 cancers which have more confirmed lncRNAs. With the continuous improvement of experimental technology, an increasing number of confirmed cancer lncRNAs will be found and our method will also apply to other cancers. Patients with the same tumor may have different reactions

using the same drug. The relationship of lncRNA and drugs is also the research direction of many researchers. Ehsan Malek et al. present lncRNA expression is associated with drug resistance and base on the function and structure of lncRNA to avoid drug resistance by using appropriate drugs(37). In future research, we make an attempt to use a data from a patient's somatic mutation to obtain more effective therapies.

## Ethics approval and consent to participate

Not applicable

## Authors' contributions

Conception and design of the work, SW Ning and DS Zhou; acquisition and analysis of data, DS Zhou; writing, reviewing and editing the paper, DS Zhou, X Li, SP Shang, H Zhi, P Wang, Y Gao. All authors have read and agreed to the published version of the manuscript and to have agreed to both be personally accountable for the author's contributions and ensure to answer any questions related to the accuracy or integrity of any part of the work. All authors read and approved the final manuscript.

## Funding

This work was supported by National Natural Science Foundation of China [32070672], Postdoctoral Scientific Research Development Fund, and Yu Weihan Outstanding Youth Training Fund of Harbin Medical University.

## Competing interests

The authors declare no conficts of interest regarding this work.

## Acknowledgements

Not applicable

## Availability of data and material

Not applicable

## Consent for publication

Not applicable

# References


1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2020. CA: a cancer journal for clinicians. 2020;70(1):7-30.
2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646-74.
3. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. Science. 2013;339(6127):1546-58.
4. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature genetics. 2013;45(10):1113-20.
5. Cheng F, Zhao J, Zhao Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. Briefings in bioinformatics. 2016;17(4):642-56.
6. Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Res. 2013;73(3):1180-9.
7. Niu Y, Ma F, Huang W, Fang S, Li M, Wei T, et al. Long non-coding RNA TUG1 is involved in cell growth and chemoresistance of small cell lung cancer by regulating LIMK2b via EZH2. Molecular cancer. 2017;16(1):5.
8. Chakravarty D, Sboner A, Nair SS, Giannopoulou E, Li R, Hennig S, et al. The oestrogen receptor alpha-regulated lncRNA NEAT1 is a critical modulator of prostate cancer. Nat Commun. 2014;5:5383.
9. Guo X, Gao L, Liao Q, Xiao H, Ma X, Yang X, et al. Long non-coding RNAs function annotation: a global prediction method based on bi-colored networks. Nucleic acids research. 2013;41(2):e35.
10. Li J, Han L, Roebuck P, Diao L, Liu L, Yuan Y, et al. TANRIC: An Interactive Open Platform to Explore the Function of lncRNAs in Cancer. Cancer Res. 2015;75(18):3728-37.
11. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic acids research. 2014;42(Database issue):D92-7.
12. Yi Y, Zhao Y, Li C, Zhang L, Huang H, Li Y, et al. RAID v2.0: an updated resource of RNA-associated interactions across organisms. Nucleic acids research. 2017;45(D1):D115-D8.
13. Hao Y, Wu W, Li H, Yuan J, Luo J, Zhao Y, et al. NPInter v3.0: an upgraded database of noncoding RNA-associated interactions. Database : the journal of biological databases and curation. 2016;2016.
14. Chen G, Wang Z, Wang D, Qiu C, Liu M, Chen X, et al. LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research. 2013;41(Database issue):D983-6.
15. Ning S, Zhang J, Wang P, Zhi H, Wang J, Liu Y, et al. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. Nucleic acids research. 2016;44(D1):D980-5.
16. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, et al. A census of human cancer genes. Nature reviews Cancer. 2004;4(3):177-83.
17. Mora A, Donaldson IM. iRefR: an R package to manipulate the iRefIndex consolidated protein interaction database. BMC bioinformatics. 2011;12:455.
18. Schmitt AM, Chang HY. Long Noncoding RNAs in Cancer Pathways. Cancer cell. 2016;29(4):452-63.
19. Ying L, Chen Q, Wang Y, Zhou Z, Huang Y, Qiu F. Upregulated MALAT-1 contributes to bladder cancer cell migration by inducing epithelial-to-mesenchymal transition. Molecular bioSystems. 2012;8(9):2289-94.
20. Ji Q, Zhang L, Liu X, Zhou L, Wang W, Han Z, et al. Long non-coding RNA MALAT1 promotes tumour growth and metastasis in colorectal cancer through binding to SFPQ and releasing oncogene PTBP2 from SFPQ/PTBP2 complex. British journal of cancer. 2014;111(4):736-48.

21. Xu TP, Huang MD, Xia R, Liu XX, Sun M, Yin L, et al. Decreased expression of the long non-coding RNA FENDRR is associated with poor prognosis in gastric cancer and FENDRR regulates gastric cancer cell metastasis by affecting fibronectin1 expression. Journal of hematology & oncology. 2014;7:63.
22. Grote P, Wittler L, Hendrix D, Koch F, Wahrisch S, Beisaw A, et al. The tissue-specific lncRNA Fendrr is an essential regulator of heart and body wall development in the mouse. Developmental cell. 2013;24(2):206-14.
23. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research. 2016;44(W1):W90-7.
24. Li Y, Qiu C, Tu J, Geng B, Yang J, Jiang T, et al. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. Nucleic acids research. 2014;42(Database issue):D1070-4.
25. Zhang YC, Liao JY, Li ZY, Yu Y, Zhang JP, Li QF, et al. Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. Genome biology. 2014;15(12):512.
26. Johnsson P, Lipovich L, Grander D, Morris KV. Evolutionary conservation of long non-coding RNAs; sequence, structure, function. Biochimica et biophysica acta. 2014;1840(3):1063-71.
27. Li H, Yu B, Li J, Su L, Yan M, Zhu Z, et al. Overexpression of lncRNA H19 enhances carcinogenesis and metastasis of gastric cancer. Oncotarget. 2014;5(8):2318-29.
28. Liu Y, Zhao J, Zhang W, Gan J, Hu C, Huang G, et al. lncRNA GAS5 enhances G1 cell cycle arrest via binding to YBX1 to regulate p21 expression in stomach cancer. Scientific reports. 2015;5:10159.
29. Kim P, Cheng F, Zhao J, Zhao Z. ccmGDB: a database for cancer cell metabolism genes. Nucleic acids research. 2016;44(D1):D959-68.
30. Hu B, Cai H, Zheng R, Yang S, Zhou Z, Tu J. Long non-coding RNA 657 suppresses hepatocellular carcinoma cell growth by acting as a molecular sponge of miR-106a-5p to regulate PTEN expression. The international journal of biochemistry & cell biology. 2017;92:34-42.
31. Liu H, Li J, Koirala P, Ding X, Chen B, Wang Y, et al. Long non-coding RNAs as prognostic markers in human breast cancer. Oncotarget. 2016;7(15):20584-96.
32. Ding X, Zhang Y, Yang H, Mao W, Chen B, Yang S, et al. Long non-coding RNAs may serve as biomarkers in breast cancer combined with primary lung cancer. Oncotarget. 2017;8(35):58210-21.
33. Wang W, Zhao Z, Yang F, Wang H, Wu F, Liang T, et al. An immune-related lncRNA signature for patients with anaplastic gliomas. Journal of neuro-oncology. 2018;136(2):263-71.
34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research. 2003;13(11):2498-504.
35. Quinn JJ, Chang HY. Unique features of long non-coding RNA biogenesis and function. Nat Rev Genet. 2016;17(1):47-62.
36. Yang G, Lu X, Yuan L. LncRNA: a link between RNA and cancer. Biochimica et biophysica acta. 2014;1839(11):1097-109.
37. Malek E, Jagannathan S, Driscoll JJ. Correlation of long non-coding RNA expression with metastasis, drug resistance and clinical outcome in cancer. Oncotarget. 2014;5(18):8027-38.

## Figure legends

**Fig.1** The workflow for prioritizing cancer lncRNA modulators. Step 1: Score the mutated gene. Step 2: Construction of lncRNA-mRNA functional network. Step 3: Score the lncRNA modulators by combining gene's score and network.

**Fig.2** lncRNA-mRNA functional network and properties. (A) A global network consisting of 19883 coding genes and 12150 lncRNAs. (B) The degree of lncRNA in the network. (C) The betweenness of lncRNA in the network.

**Fig.3** Assessment of predictive power for our method. (A) The performance of our method for cancer genes. (B) The performance of four networks for lncRNA modulators. (C) The performance of three methods for lncRNA modulators. (D) The performance of our method with part of edges deleted.

**Fig.4** Functional Analysis of high-risk cancer lncRNA modulators. (A) The heatmap of 80 lncRNA scores for 9 cancers. (B) Functional annotation for common character lncRNA modulators. (C) Functional annotation for cancer-specific lncRNA modulators. (D) Venn diagram showing the number of intersections between neighbor miRNAs and genes and cancer miRNAs and genes. (E) Schematic representation of lncRNA-linked cancer miRNAs and genes in the network.

**Fig.5** Comparison of biological characteristics of different levels of lncRNA modulators, including(A) number of exons, (B)number of miRNA binding sites, (C) conservation score and (D)lncRNA expression.

**Fig.6** Analysis of top20 stomach cancer lncRNA modulators. (A) 4 lncRNAs were significantly associated with survival. (B) Functional annotation for top 20 lncRNA modulators. (C) The number of intersections between neighbor miRNAs and genes and cancers miRNA and genes for 4 lncRNAs. (D) The subnet of the cancer genes linked to these lncRNAs in the network. (E) The subnet of the cancer miRNAs linked to these lncRNAs in the network.

# Figures

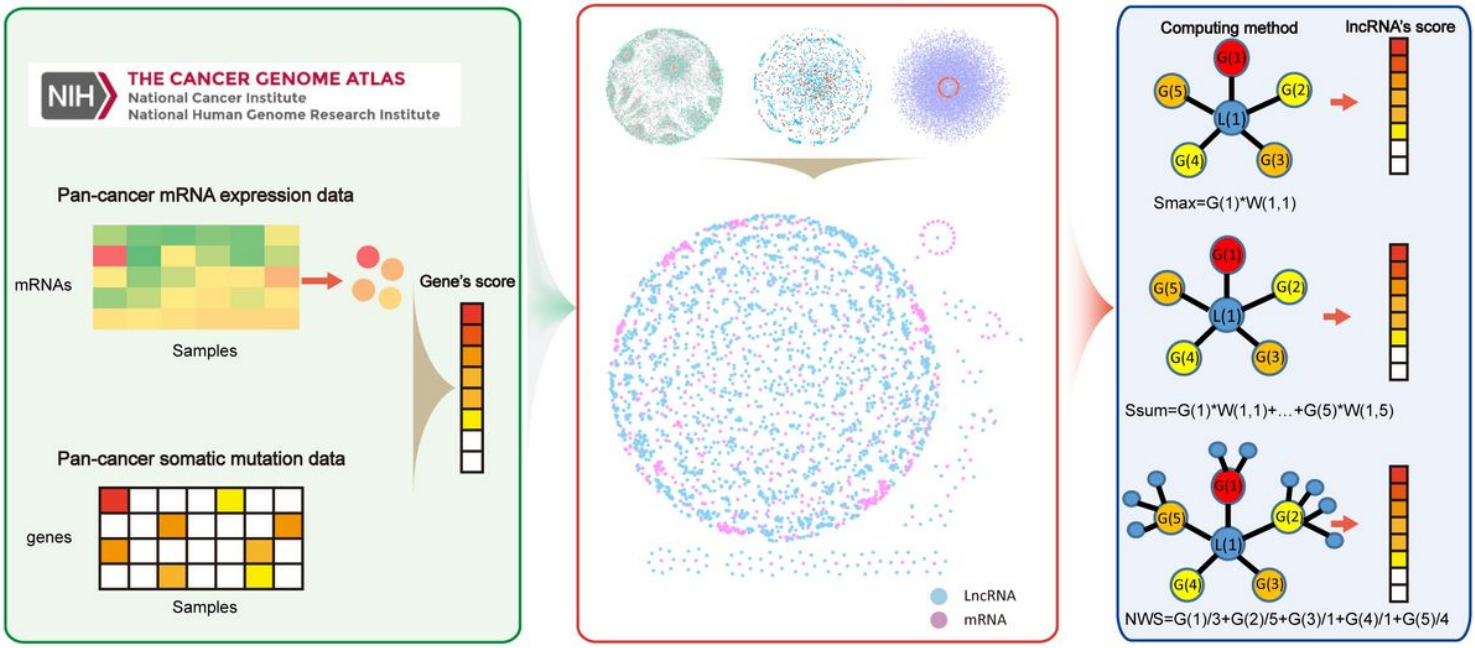



**Figure 1**

The workflow for prioritizing cancer lncRNA modulators. Step 1: Score the mutated gene. Step 2: Construction of lncRNA-mRNA functional network. Step 3: Score the lncRNA modulators by combining gene's score and network.

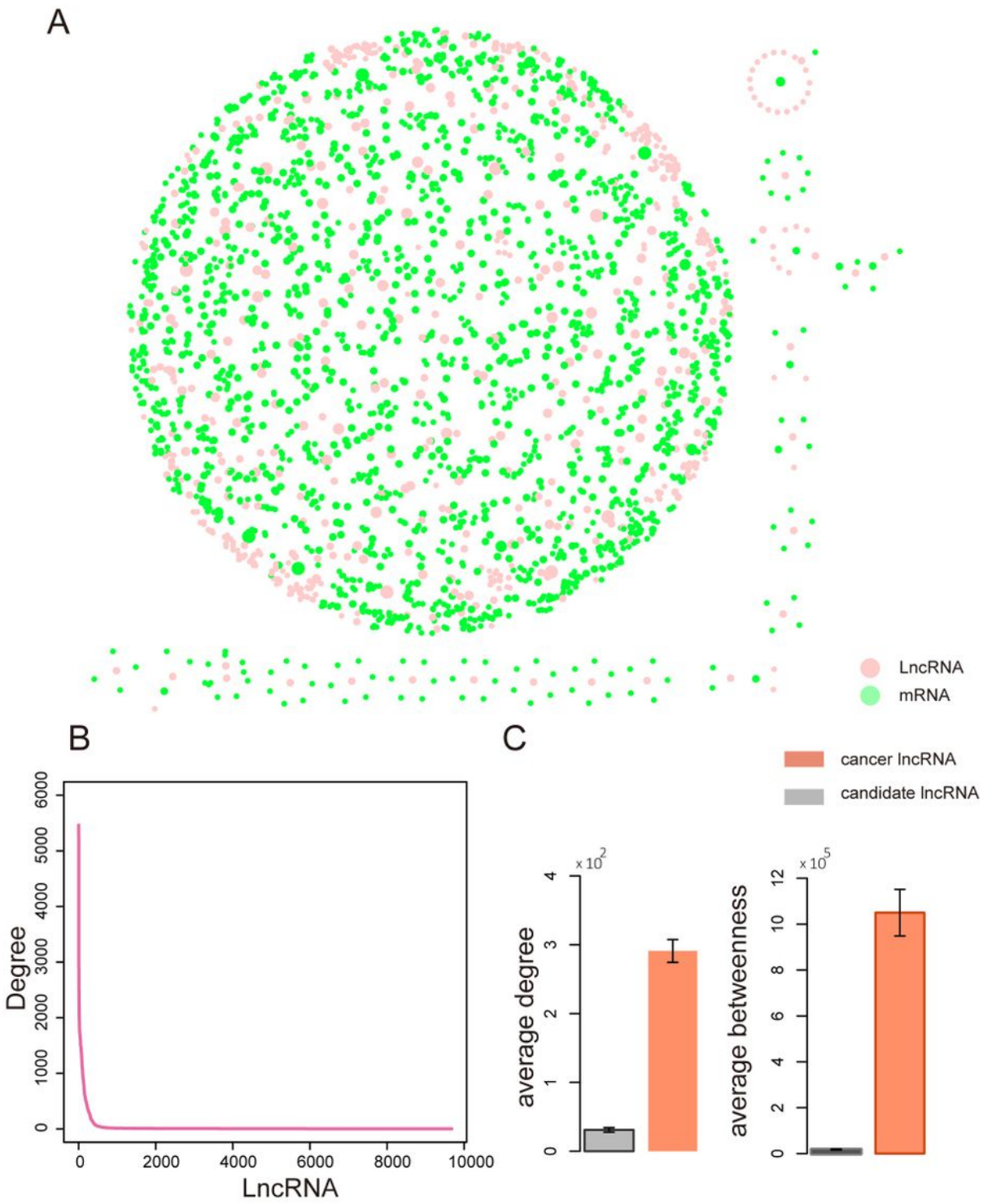
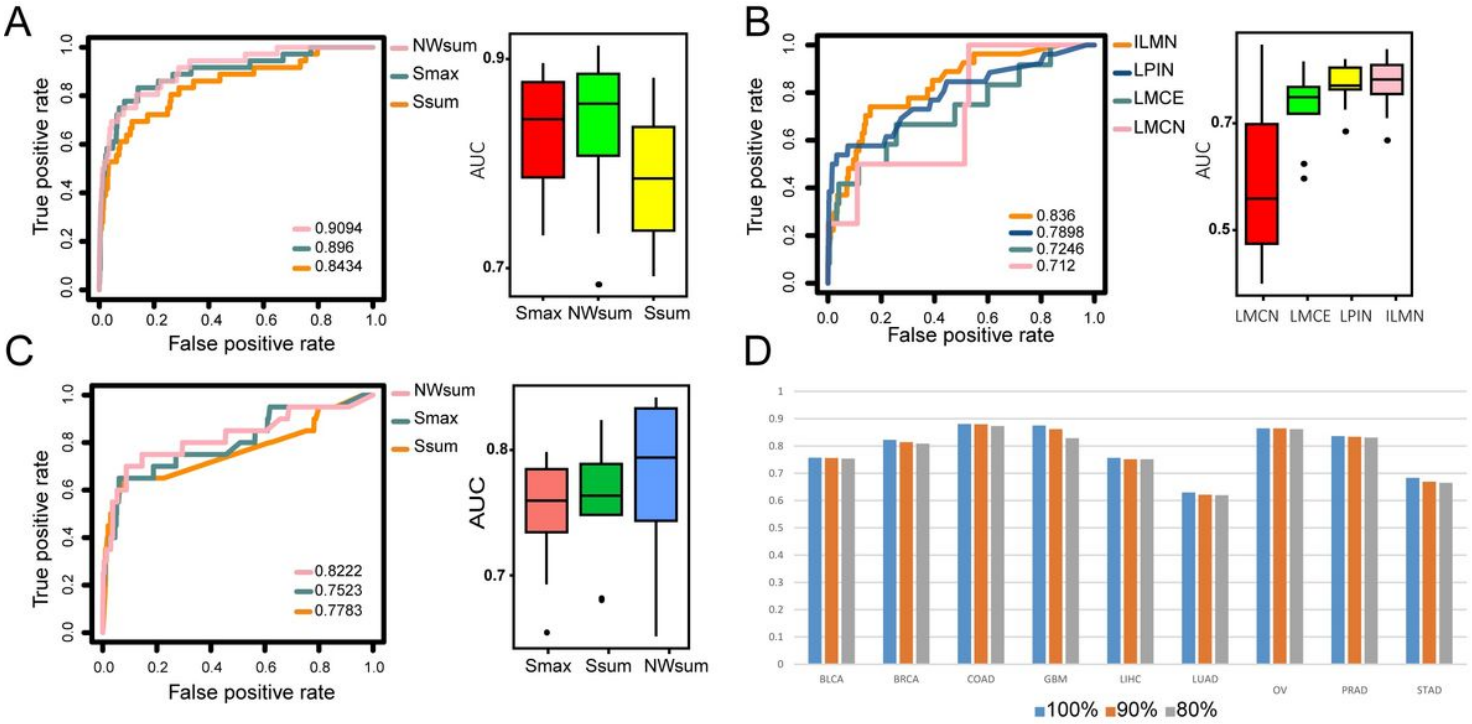


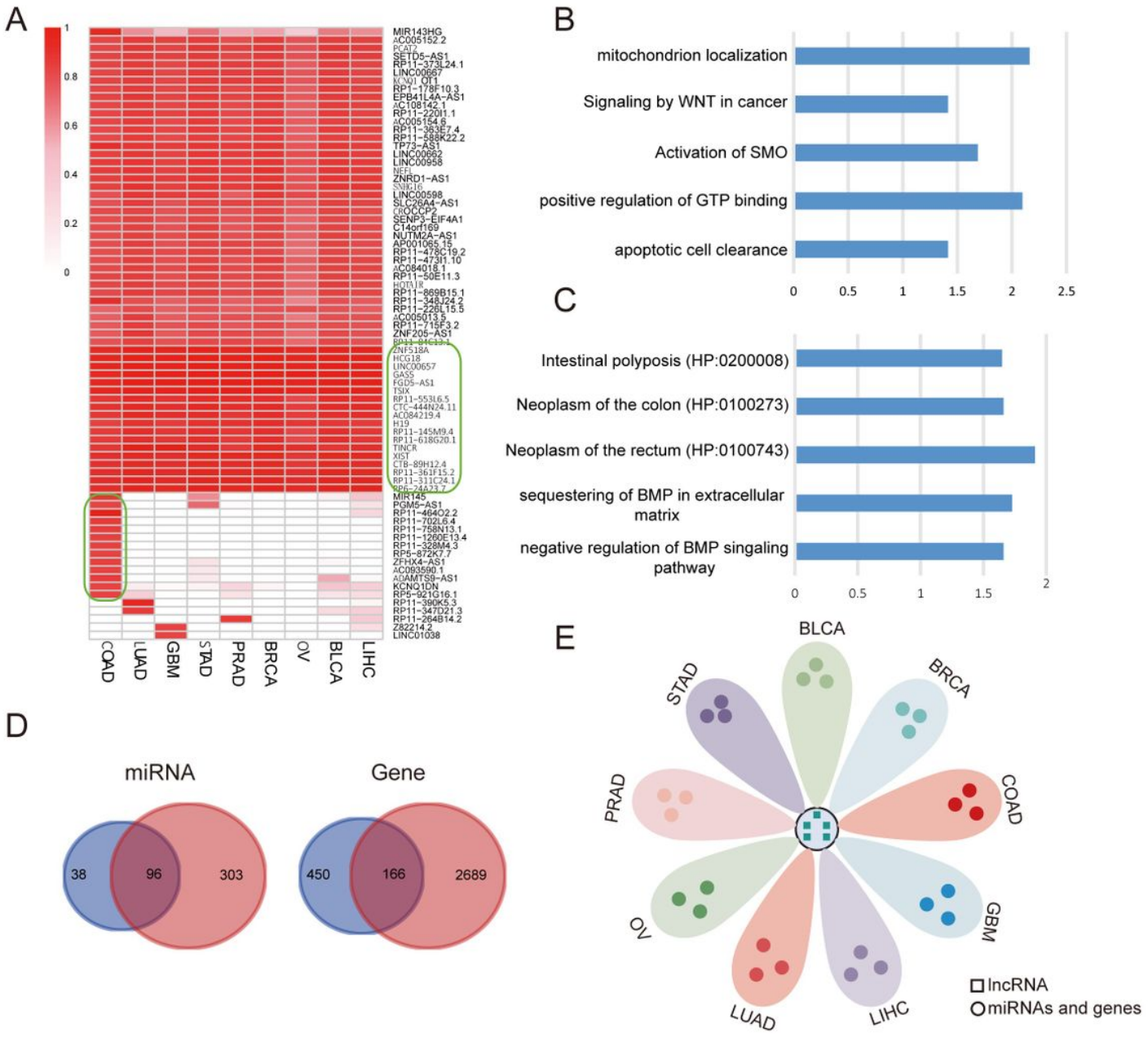

**Figure 2**

lncRNA-mRNA functional network and properties. (A) A global network consisting of 19883 coding genes and 12150 lncRNAs. (B) The degree of lncRNA in the network. (C) The betweenness of lncRNA in the network.




## Figure 3

Assessment of predictive power for our method. (A) The performance of our method for cancer genes. (B) The performance of four networks for lncRNA modulators. (C) The performance of three methods for lncRNA modulators. (D) The performance of our method with part of edges deleted.




## Figure 4

Functional Analysis of high-risk cancer lncRNA modulators. (A) The heatmap of 80 lncRNA scores for 9 cancers. (B) Functional annotation for common character lncRNA modulators. (C) Functional annotation for cancer-specific lncRNA modulators. (D) Venn diagram showing the number of intersections between neighbor miRNAs and genes and cancer miRNAs and genes. (E) Schematic representation of lncRNA-linked cancer miRNAs and genes in the network.

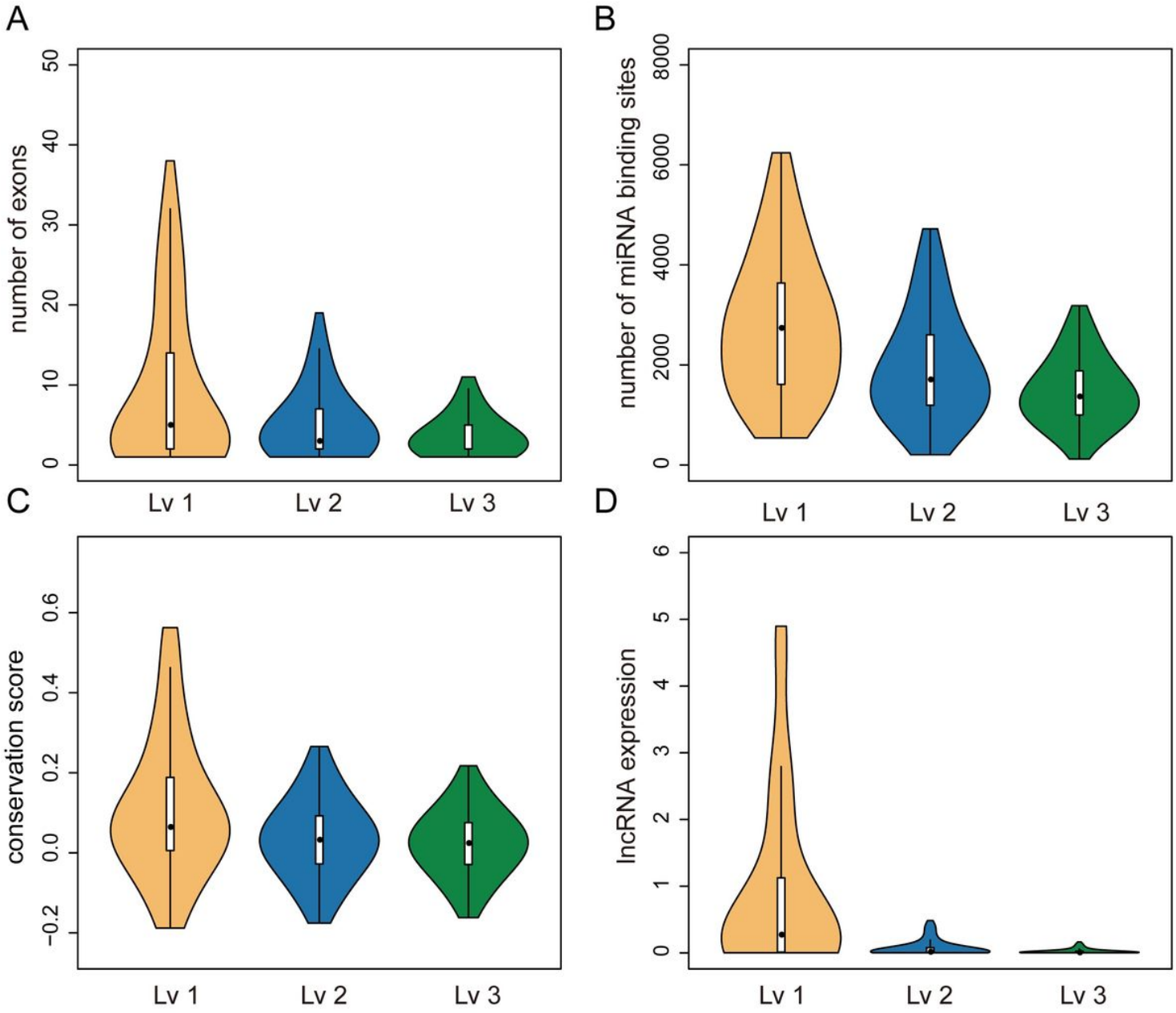
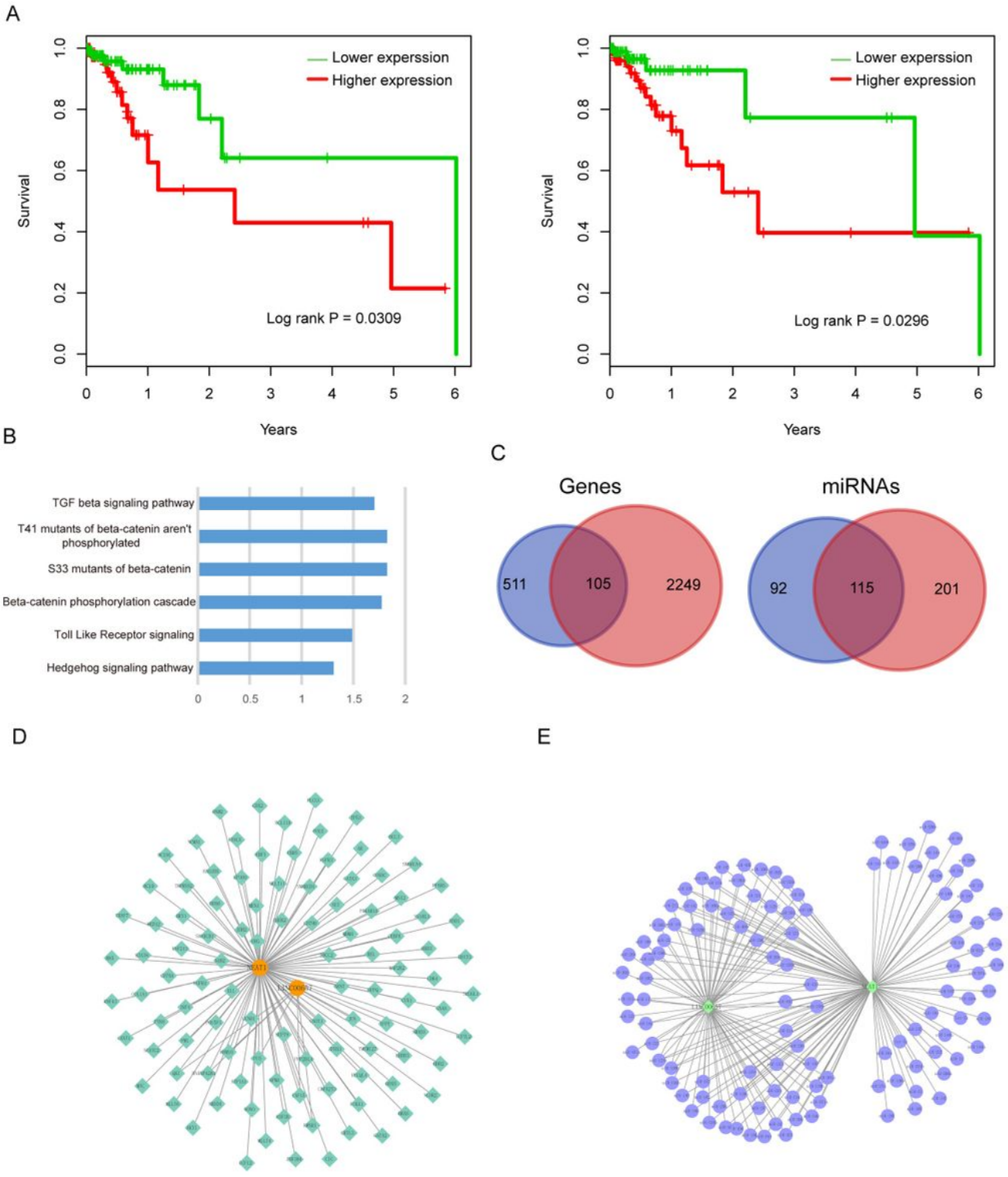



**Figure 5**

Comparison of biological characteristics of different levels of lncRNA modulators, including(A) number of exons, (B)number of miRNA binding sites, (C) conservation score and (D)lncRNA expression.




## Figure 6

Analysis of top20 stomach cancer lncRNA modulators. (A) 4 lncRNAs were significantly associated with survival. (B) Functional annotation for top 20 lncRNA modulators. (C) The number of intersections between neighbor miRNAs and genes and cancers miRNA and genes for 4 lncRNAs. (D) The subnet of the cancer genes linked to these lncRNAs in the network. (E) The subnet of the cancer miRNAs linked to these lncRNAs in the network.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTablesS1.xlsx
- SupplementaryTablesS2.xlsx