Molecular
BioSystems

Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/ molecularbiosystems

# Identification of module biomarkers from the dysregulated ceRNA-ceRNA interaction network in lung adenocarcinoma †

**Tingting Shao[1], Aiwei Wu[1], Juan Chen[1], Hong Chen[1], Jianping Lu[1], Jing Bai[1], Yongsheng Li[*], Juan Xu[*] and Xia Li[*]**

***Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX***
**DOI: 10.1039/b000000x**

Competitive endogenous RNA (ceRNA) represents a novel layer of gene regulation that plays important roles in the physiology and development of diseases such as cancer and their dysregulation could contribute to cancer pathogenesis. Here, we proposed a computational method to systematically identify genome-wide dysregulated ceRNA-ceRNA interactions by integrating microRNA regulation with expression profiles in cancer and normal tissues by RNA sequencing, as well as considering details of how ceRNAs behavior has changed. These gain or loss dysregulations further assembles into a dysregulated ceRNA-ceRNA network, lncRNAs and pseudogenes are also considered. After applying the method to lung adenocarcinoma, we found that most dysregulations are connected together and formed a lung adenocarcinoma dysregulated ceRNA-ceRNA network (LDCCNet). Our analyses found that ceRNA pairs with gain regulations have consistent expression in cancer, otherwise for loss regulation, it is not necessary. Moreover, ceRNAs with more significant gain regulations (gain ceRNAs) are underwent stronger regulation in cancer, thus their expression are more likely to decrease in cancer, while the expression of loss ceRNAs display the rising trend. Additionally, we found gain and loss ceRNAs as topological key nodes are implicated in the development of cancer. Finally, dysregulated ceRNA modules were identified, which are significantly enriched with known lung cancer microRNAs. We further found that several modules have the power as diagnostic biomarkers even in three independent datasets. For example, the module with lncRNA RP11-457M11.2 as a center is involved in epithelial cell morphogenesis process and provides the average AUC values of 0.95. Our study about LDCCNet opens up the possibility of a new biological mechanism in cancer that could be served as biomarkers for diagnosis.

## Introduction

MicroRNAs (miRNAs) are small single-stranded RNAs of 18-25 nucleotides in length that guide many key biological processes, such as cell proliferation, signal transduction, and apoptosis [1]. MiRNAs can repress their targets through binding to miRNA response elements (MREs) on the 3'UTR of mRNAs, causing translational repression or mRNA degradation [2, 3]. In addition to coding mRNAs, MREs can also be found on non-coding transcripts such as pseudogenes and long non-coding RNAs (lncRNAs). Importantly, each miRNA regulates numerous RNA targets, both coding and non-coding RNAs, and the vast majority of RNA molecules harbor several MREs and are thus repressed by different miRNAs. This target multiplicity has led to the observation that different RNAs might compete for limited pools of miRNAs, thus acting as ceRNAs [4]. Currently, convincing evidences have been provided that ceRNAs are involved in tumorigenesis. For example, Kumar *et al*. have shown that Hmga2 promotes lung cancer progression in human cells and mouse by operating as a ceRNA [5]. In addition, non-coding RNAs are also involved in cancer by acting as target decoys, such as the lncRNA HULC regulating activity of miR-372 to reduce miRNA mediated translational repression of PRKACB in hepatocellular carcinoma, and pseudogene PTENP1 protecting PTEN messenger as a ceRNA [6, 7]. Linc00974 affects KRT19 expression as a ceRNA interacting with miR-642 in hepatocellular carcinoma (HCC) and acts as a biomarker in predicting the growth and metastasis of HCC [8]. These ceRNAs could cross-regulate each other and constitute a ceRNA network allowing RNA to communicate.

Cancers are increasingly modeled by using different biological networks, since analyzing the properties of the entire networks has the potential to rapidly generate new biological hypotheses, such as identifying functionally coherent modules [9]. Most studies have focused on analyzing the ceRNA networks in certain disease or physiological status. A predicted miRNA-mediated ceRNA network in glioblastoma, has been found to regulate canonical oncogenic pathways [10]. Paci *et al.* built normal and cancer networks of lncRNA-mRNA sponge interactions mediated by miRNAs using breast cancer expression data [11]. Additionally, Li *et al*. examined the change of ceRNA networks in different subtypes of prostate cancers based on a previous predicted ceRNA network [12]. Normal ceRNA regulations are needed for the proper functioning of a cell, thus disruption of these ceRNA pairs may promote the development of cancer. Detecting such disruptions can help us better understand the initiation and propagation of cancer, compared to commonly used ceRNA network analysis only under cancer status.

Lung cancer is the leading cause of cancer-related human deaths worldwide [13, 14]. Lung cancer can be classified based on histopathologic findings, and adenocarcinoma is one of most common subtypes [15]. An early and correct diagnosis may warrant immediate therapy to potentially reduce the mortality rate. The dysregulated ceRNA-ceRNA network may provide new hope for exploration of pathogenesis of lung adenocarcinoma and new biomarkers with high accuracy in diagnosis.

Based on the above observation, in the current study, we proposed a computational approach to identify dysregulated ceRNA-ceRNA interactions by integrating miRNA regulation with RNA-seq data in cancer and normal tissues. To reflect the effect of dysregulated ceRNA interactions on the phenotype, differentially expressed ceRNAs were used to build the dysregulated network. At the same time, we introduced both gain and loss ceRNA interactions in the dysregulated network, in order to further understand the roles of ceRNA in cancer progression. After applying this method to lung adenocarcinoma, we found that most dysregulations are connected together and formed a lung adenocarcinoma dysregulated ceRNA-ceRNA network (LDCCNet). Then, the LDCCNet was used to study the roles of coding-genes, lncRNAs and pseudogenes as dysregulated ceRNAs in pathogenesis of lung adenocarcinoma. The properties of dysregulated ceRNA interactions were analyzed and two kinds of special ceRNAs significantly enriched by dysregulated ceRNA interactions (gain and loss ceRNAs) were focused on. Finally, multiple dysregulated ceRNA modules significantly regulated by known lung cancer miRNAs were identified and might serve as novel diagnostic biomarkers. Collectively, our findings provide a comprehensive dysregulated ceRNA-ceRNA network landscape in lung adenocarcinoma and present several new diagnostic biomarkers. The identification of dysregulated ceRNA-ceRNA network not only dictates a promotion of our understanding of gene regulatory networks in cancer but also opens up the possibility of a new biological mechanism that could be served as biomarkers for diagnosis [16].

## Results

### Global properties of the LDCCNet

To systematically explore the influence of dynamic changes of ceRNA regulation to gene expression in lung adenocarcinoma, we first constructed the LDCCNet using computational method developed by us from 72 adenocarcinomas and adjacent normal pairs (see Materials and methods). 4857 dysregulated ceRNA interactions among 1674 protein-coding genes, 46 lncRNAs and 32 pseudogenes were detected (Table 1). In these dysregulated ceRNA interactions, there are 2068 gains and 2789 losses (Table 1). In addition, gain and loss regulations respectively account for 0.07% and 0.1% of all 2,635,721 ceRNA-ceRNA candidates (Table 1). That is to say, a relatively small part of the ceRNA-ceRNA interactions are dysregulated in lung adenocarcinoma. These results show that in lung adenocarcinomas not only some normal ceRNA regulations could disappear or be reduced, but also some new ceRNA regulations could appear, so as to create more favorable conditions for development of cancer cells.

Table 1 - Statistics of the nodes and edges in the LDCCNet

| Network | Node | | | Edge | |
|---|---|---|---|---|---|
| LDCCNet | Coding-genes | lncRNAs | pseudogenes | gain | Loss |
| | 1674 | 46 | 32 | 2068 | 2789 |
| | (622/1052) | (23/24) | (20/12) | (0.07%) | (0.1%) |

**Note**:% represents the percentage of dysregulated ceRNA interactions in genome-wide ceRNA interactions. (a/b) ‘a’ represents the number of ceRNAs with up expression in lung adenocarcinoma, and ‘b’ represents the number of ceRNAs with down expression in lung adenocarcinoma.

Next, we discussed the structure and organization of the LDCCNet. As shown in Figure 1A, most dysregulated ceRNAs in the LDCCNet are connected and form a large connecting subnetwork. We found that a few ceRNAs have a relatively large number of dysregulated ceRNA partners, whereas most ceRNAs have few dysregulated ceRNA partners. Additionally, 70% of ceRNAs in the LDCCNet participate in at least two dysregulated ceRNA interactions. The examination of the degree distribution of the LDCCNet reveals a power law distribution. Therefore, like many large-scale networks, the LDCCNet displays scale-free characteristics (Figure 1B), indicating that the LDCCNet is not random but is characterized by a core set of organizing principles in its structure that distinguish it from randomly linked networks. Lastly, we found the LDCCNet has larger clustering coefficient than random, as expected for module characteristics (Figure 1C and Supplementary Text S1). The scale-free and module characteristics of LDCCNet make dysregulated ceRNA interactions influence each other and effectively exchange dysregulated information both at a global and at a local scale. The dysregulation of one point amplifies affect along veins of the LDCCNet and it may make the LDCCNet more powerful to influence gene expression in lung adenocarcinomas. A recently study also has indicated that indirect interactions critically amplify ceRNA cross-talk [17].

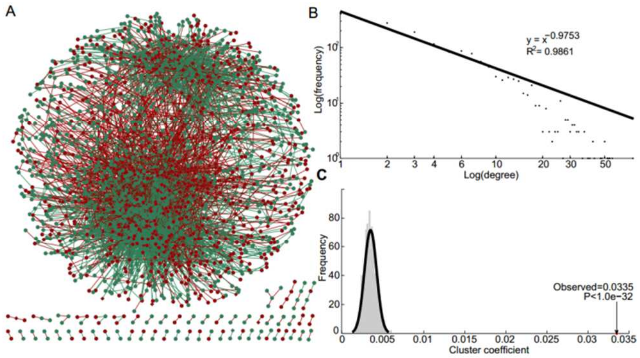

**Figure 1**. The layout of the LDCCNet and its structural features.

(A) The LDCCNet generated by the procedure described in the Methods. A circle node marks coding-gene, a diamond node marks lncRNA, and a hexagon node marks pseudogene. A node colored by green represents low expression in lung adenocarcinoma, and a node colored by red represents high expression in lung adenocarcinoma. A red edge represents a gain dysregulation between two ceRNAs, and a green edge represents a loss dysregulation between two ceRNAs. The width of edge represents the absolute value of PCC difference of a ceRNA pairs Δ*R* (see methods) (B) Degree distribution of the LDCCNet. (C) The cluster coefficient of LDCCNet is higher than random networks.

### Dysregulated ceRNA interactions contributing aberrant expression in cancer

To more deeply explore roles of these emerging dysregulations in cancer, we investigated the contribution of dysregulated ceRNA interactions to aberrant expression of lung adenocarcinoma. As the ceRNAs regulate each other in trans, we first analyzed the distribution of these dysregulated ceRNAs on chromosomes. We found that the ceRNA pairs in 93.24% of dysregulated interactions located on different chromosomes. Known lung cancer genes and their ceRNA regulators also located on different chromosomes. The proportion of the dysregulated ceRNAs on chromosome 6 is the highest (Supplementary Figure S1).

As the trans-regulation of dysregulated ceRNAs is expected to contribute to the cancer aberrant transcriptome, we next investigated the correlation between expression change and dysregulated interactions. The expression pattern of ceRNA pairs was observed and we found the distribution of patterns is significantly different between gain interactions and loss interactions (Figure 2A). As results show that 92% of ceRNA pairs with gain regulations have the same direction of expression change, while for loss regulations, the ceRNA pairs with the same direction of expression change is only 28% and most ceRNA pairs have the opposite direction of expression change (Figure 2A). In other words, a gain interaction may mean that the regulation for the two ceRNAs each other is enhanced, and therefore the ceRNAs have a consistent expression in cancer, while for a loss ceRNA regulation, the ceRNAs are not necessarily with a consistent expression. Moreover, we found the proportion of gain ceRNA pairs co-occurring in the same GO terms is 1.5 times more than that of loss ceRNA pairs by analyzing gain ceRNA pairs and loss ceRNA pairs annotated in GO. This result indicated that gene pairs with gain ceRNA interactions may cooperatively regulate the same biological functions, and furthermore their consistent expression changes are more likely to increase or decrease efficiency of the function, leading to cancer. In addition, to explore how the ceRNA dysregulations affect changes of gene expression in cancer, we focused on two kinds of ceRNAs, “gain ceRNA” and “loss ceRNA” (see Materials and methods). 528 gain ceRNAs and 520 loss ceRNAs were defined. The two kinds of ceRNAs participate in as many as 73% of all the dysregulated ceRNA interactions. The distribution of up-regulated ceRNAs and down-regulated ceRNAs in gain ceRNAs is significantly different with in loss ceRNAs (Figure 2B Fisher's Exact Test p-value =4.3*e-5). There are more down-regulated ceRNAs belonging to gain ceRNAs than loss ceRNAs. However, for loss ceRNAs, the proportion of up-regulated ceRNAs becomes higher than gain ceRNAs. These results suggest that gain ceRNAs get stronger regulation and their expressions are more likely to decrease in cancer, while loss ceRNAs’ expressions have the rising trend. Although ceRNAs can regulate gene expression, we found the extent of expression change has no obvious correlation with the ceRNA dysregulated intensity (Supplementary Figure S2). But some ceRNAs with lots of dysregulated interactions indeed have relatively larger fold change in expression. For example, CAV2 is a major component of the inner surface of caveolae and involved in signal transduction, cellular growth control and apoptosis, which has 48 dysregulated interactions in the LDCCNet. At the same time, expression of CAV2 in lung adenocarcinoma is significantly decreased to about a quarter of the normal expression. These results indicate that dysregulated ceRNAs may play important roles in fine-tuning gene expression.

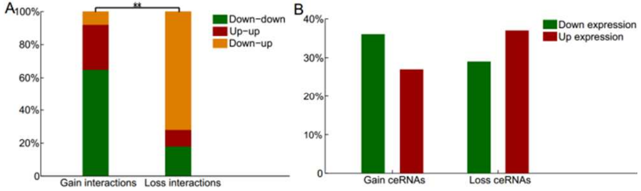

**Figure 2**. Gain and loss ceRNA interactions contributing aberrant expression in lung adenocarcinoma.

(A) The expression pattern of a ceRNA pair is divided into three categories: i. both are up-expression in lung adenocarcinoma (up-up); ii. both are down-expression in lung adenocarcinoma (down-down); iii. one is down-expression and the other is up-expression in lung adenocarcinoma (down-up). The distributions of expression pattern of ceRNA pair respectively in gain and loss ceRNA interactions. **P < 0.005 for Fisher's Exact Test. (B) The distributions of up-expression ceRNAs and down-expression ceRNAs respectively in gain and loss ceRNAs.

### Gain ceRNAs and loss ceRNAs as topological key nodes involving in cancer progression

We further studied topological and functional properties of gain ceRNAs and loss ceRNAs. Firstly, to further inspect whether gain ceRNAs and loss ceRNAs are critical for global network connectivity, we systematically removed either gain or loss ceRNAs from the LDCCNet and analyzed the number of paths between ceRNAs using a topological measure betweenness. In a biological context, betweenness measures the ways of indirect or direct regulation in the regulation network. We found betweenness is more strongly affected by removing gain and loss ceRNAs than other ceRNAs (Kolmogorov-Smirnov test, p-value gain =3.87e-5, p-value loss =2.83e-3) (Figure 3A). Another two topological measures are the number of components and the size of maximum component, which measure global network connectivity and the integrity of a network. We found separate removal of gain or loss ceRNAs produces consistent change, the number of components becomes larger and the main component becomes smaller than removal of other ceRNAs (Figure 3B and Supplementary Figure S3). The sensitivity of these topological measures to removal of gain and loss ceRNAs suggests that the two kinds of ceRNAs are topological key nodes in the LDCCNet and play important roles in ceRNA dysregulations.

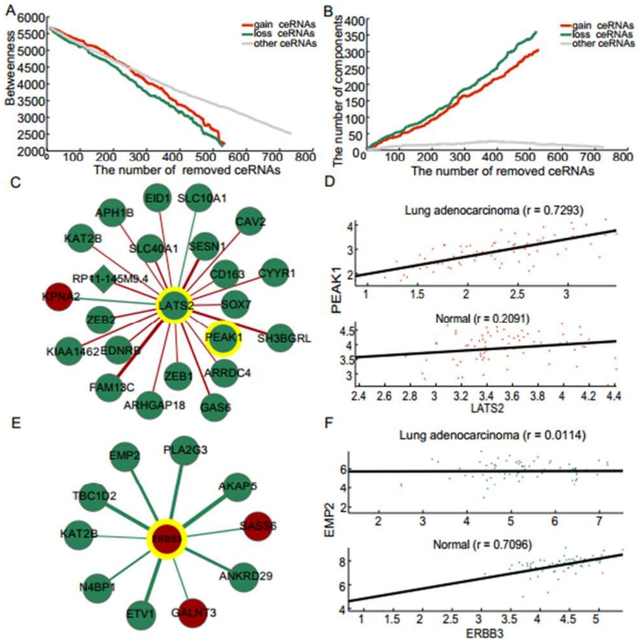

**Figure 3**. Gain ceRNAs and loss ceRNAs.

(A) Network betweenness remaining after removing the gain ceRNAs, loss ceRNAs and other ceRNAs. (B)The number of components remaining after removing the gain ceRNAs, loss ceRNAs and other ceRNAs.(C) An example of gain ceRNA LATS2 with all its dysregulated interactions in the LDCCNet. A node with yellow border represents known lung cancer gene. (D) One example of ceRNA pair (LATS2 versus PEAK1) whose expression variation across individuals (x- and y- axis) reveals a gain ceRNA interaction in lung adenocarcinoma. (E) An example of loss ceRNA ERBB3 with all its dysregulated interactions in the LDCCNet. A node with yellow border represents known lung cancer gene. (F) One example of ceRNA pair (ERBB3 versus EMP2) whose expression variation across individuals (x- and y- axis) reveals a loss ceRNA interaction in lung adenocarcinoma.

Next, we found that 56% of the known lung cancer genes in the LDCCNet belong to the gain or loss ceRNAs. But the distributions of known lung cancer genes in the two types of ceRNAs are not significantly different (Fisher's Exact Test p-value =0.28), suggesting that both gain and loss ceRNA regulations contribute to lung adenocarcinoma progression. For example, the well-known lung cancer gene LATS2 is a gain ceRNA, and gains 22 ceRNA regulations in lung adenocarcinoma (Figure 3C). These regulations significantly change in cancer compared with normal, such as LATS2 and PEAK1, which are more significantly positively correlated in cancer than normal (Figure 3D). These dysregulated ceRNA-ceRNA interactions are significantly mediated by lung cancer-related miRNAs (Hypergeometric test p-value=2.12*e-8). LATS2 is serine/threonine protein kinase belonging to the LATS tumor suppressor family. It has been shown that down-regulation of the LATS2 gene was observed in most non-small cell carcinoma but was not related to any mutation or polymorphism [18]. So gain ceRNA regulations at post-transcriptional level could be an inducement for down-regulation of LATS2. Another known lung cancer gene, ERBB3, is a member of the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases and has been recognized showing strong association with malignant proliferation [19]. ERBB3 is a loss ceRNA and losses 10 ceRNA regulations in lung adenocarcinoma (Figure 3E). These regulations work well in normal while are disrupted in cancer. Figure 3F highlights an example of the pairwise correlation between ERBB3 and EMP2 showing obviously disrupted co-expression patterns in lung adenocarcinoma. EMP2 has been recognized as a putative tumor suppressor gene in certain model systems and down-regulated in lung adenocarcinoma [20]. There are four known lung cancer-related miRNAs involved in these loss dysregulations. One of the miRNAs is hsa-mir-372 which has been found highly expressed in lung cancer, and it participates in 70% of loss dysregulations [21]. With confidence that the gain and loss ceRNAs play important roles in lung adenocarcinoma, we next interrogated which biological processes may be subjected to ceRNA-mediated regulation. In Gene Ontology analysis done systematically for gain and loss ceRNAs, respectively, biological processes related to cancer are significantly enriched. Remarkably, we found an overrepresentation of both gain and loss ceRNAs involved in cell cycle (Table 2). At the same time, gain ceRNAs also specifically participate in cell adhesion, cell proliferation and vasculogenesis (Table 2). The hallmarks of cancer are shared in common by most and perhaps all human tumor types, defined as acquired functional capabilities that allow cancer cells to survive, proliferate, and disseminate[22]. A list of GO terms related to the hallmarks of cancer was obtained from a previous study[23], and genes annotated in these GO terms were considered as the known gene sets involved in these cancer hallmarks. After performing function enrichment, we found that gain ceRNAs or loss ceRNAs were significantly enriched in most of the hallmarks (Supplementary Table S1). This analysis revealed that ceRNA regulations as a new regulatory mechanism might be involved in most cancer hallmarks in lung adenocarcinoma, such as antigrowth signals, reprogramming energy metabolism, tissue invasion and metastasis and so on. Lastly, we also mapped the hallmark genes onto the LDCCNet to generate the hallmark networks proposed by Wang et[24, 25]. Then, we identified a ceRNA associated survival network, in which all the ceRNAs involves the functions to sustain chronic proliferation, resist to cell death, and resist inhibitory signals (Supplementary Figure S4). Such a network represents the ceRNA regulation mechanism for lung adenocarcinoma cell survival and proliferation. Additionally, another ceRNA associated hallmark network, the EMT network, was identified in which all the ceRNAs participate in tissue invasion and metastasis (Supplementary Figure S5).

Table2 - Functional enrichment of gain ceRNAs and loss ceRNAs based on GO biological process terms

| Category of ceRNAs | Term name | Count | FDR |
|---|---|---|---|
| Gain ceRNAs | mitotic cell cycle | 35 | 0 |
| | cell adhesion | 33 | 0.000014 |
| | negative regulation of cell proliferation | 25 | 0.000262 |
| | regulation of cell cycle | 10 | 0.00032 |
| | cell proliferation | 23 | 0.000753 |
| | mitosis | 16 | 0.001248 |
| | negative regulation of tumor necrosis factor production | 6 | 0.001625 |
| | immune response | 22 | 0.001928 |
| | M phase of mitotic cell cycle | 10 | 0.002263 |
| | vasculogenesis | 7 | 0.00733 |
| Loss ceRNAs | mitotic cell cycle | 23 | 0.002706 |
| | G1 phase of mitotic cell cycle | 6 | 0.033278 |
| | M phase of mitotic cell cycle | 9 | 0.033278 |
| | neuromuscular junction development | 5 | 0.033278 |
| | brown fat cell differentiation | 5 | 0.033278 |

### Diagnosis power of dysregulated ceRNA modules

It is widely accepted that the occurrence of cancer is carried out through the concerted activity of many genes [26]. Functionally coherent modules can be as biomarkers to distinguish cancer from normal. We applied affinity propagation clustering approach to detect modules in the LDCCNet [27]. Then the modules were further required to be composed of more than 4 ceRNAs and significantly enriched by known lung cancer miRNAs. We finally detected 35 modules (Supplementary Table S2). For example, Figure 4A shows a module with FGF2 as center, which is enriched by gain ceRNA interactions. In human cancer, FGF-2 promotes cancer cell proliferation, migration, and invasiveness, as well as angiogenesis and the cycling of cancer stem cells [28]. The dysregulated ceRNA interactions in this module are significantly regulated by many well-known lung cancer-related miRNAs and function analysis reveals that this module is involved in the regulation of epithelial cell proliferation. As known that, lung adenocarcinoma is mainly originated from the bronchial epithelial cells, and these results suggest that the dysregulated ceRNA module may contribute to development of lung adenocarcinoma [29]. Additionally, the lncRNA MALAT1 in the module, is up-regulated in several malignancies, and has been implicated in non-small cell lung cancer [30]. The loss of the ceRNA regulation for FGF2 to MALAT1 may cause the over-expression of MALAT1. Then, we inspected the potential of this module as a biomarker for early detection. Hierarchical clustering analysis based on the expression of ceRNAs in the module clearly separates cancer from normal tissue samples, independently of clinic pathologic characteristics (Supplementary Figure S6). The performance of the module is evaluated by area under the receiver operation characteristic (ROC) curve (AUC). ROC curves of the module gives AUC values of 0.966 in the LAC72s dataset (Figure 4B). Lastly, to confirm the robustness of this module as biomarker, three independent datasets were analyzed, including one RNA-seq LACtcga, two microarray LAC107 and LAC58s datasets. We found most cancers are able to be distinguished from benign adjacent tissues in these three datasets and the AUCs are 0.967, 0.974 and 0.916 respectively in the LACtcga, LAC107 and LAC58s datasets (Figure 4B). Moreover, we also used the re-sampling tests to estimate the robustness of biomarkers [31]. We generated 20 subsets by randomly selecting 19 paired lung adenocarcinomas and adjacent non-tumor samples (25% of all samples) from LAC72s, with < 30% sample overlap among these subsets. Then, we calculated the AUCs of the biomarker in each subset respectively and found that on the whole the AUCs do not vary greatly from one subset to another, although some AUCs' value become lower (Supplementary Figure S7A). These results suggested that the biomarker is relatively robust. Another module with a loss ceRNA ZBTB4 as center, is mainly involved in differentiation, metabolism and regulation of cell proliferation (Figure 4C). It has been shown down-regulation of ZBTB4 relative to normal tissue occurs in lung carcinoma and could promote cell cycle arrest in response to activation of p53 and suppress apoptosis through regulation of P21CIP1 [32]. At the same time, a lot of well- known lung cancer-related miRNAs mediate the dysregulation of this module, such as hsa-mir-139, hsa-mir-196 and hsa-mir-20a. Then, we also evaluated its capacity for distinguishing between cancer and normal. As the Supplementary Figure S8 shown that normal samples and cancer samples are clearly grouped in two distinct sub-branches in LAC72s dataset, and AUC is as high as 0.983 (Figure 4D). The module is also powerful as biomarker in three independent datasets, and the AUCs are 0.976, 0.984 and 0.888 respectively in the LACtcga, LAC107 and LAC58s datasets (Figure 4D). The re-sampling tests also showed the robustness of this module (Supplementary Figure S7 B).

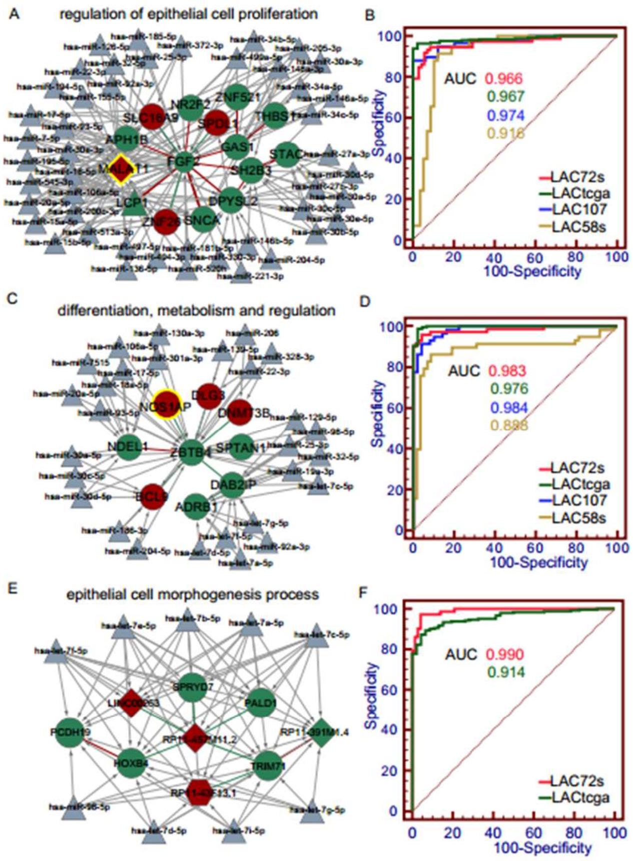

**Figure 4**. ceRNA modules as biomarkers.

(A, C, E) Examples of ceRNA modules. Triangle represents lung cancer miRNA. A gray edge represents a regulation from a miRNA to one of its targets. A node with yellow border represents known lung cancer gene. (B, D, F) ROCs of the ceRNA module biomarkers in “LAC72s”, “LACtcga”, “LAC107” and “LAC58s” expression dataset, respectively.

Non-coding RNAs have been shown to play a critical role in tumorigenesis and we indeed found some lncRNAs and pseudogenes are involved in the lung adenocarcinoma ceRNA dysregulations. Non-coding RNAs are very promising as diagnostic biomarkers. And, as expected, we explored a module mainly composed of non-coding RNAs, including three lncRNAs and one pseudogene (Figure 4E). Most of ceRNAs in this module are connected by loss interactions. The hub of this module is RP11-457M11.2, a loss ceRNA. RP11-457M11.2 is significantly higher expressed in lung adenocarcinoma and the average expression value of it is 4.13*e9 (FPKM), which is two times of the average expression value of all lncRNAs in lung adenocarcinoma. These dysregulations among non-conding RNAs and coding RNAs in the module are mediated by some known lung cancer miRNAs. Most of the miRNAs belong to let-7 family postulated to function as tumor suppressor, expression of which is reduced in non-small cell lung cancer [33, 34]. Coding genes in the module are mainly involved in epithelial cell morphogenesis process, suggesting that these non-coding RNAs may participate in the regulation of epithelial cell function to promote the development of cancer by acting as ceRNAs. Then, hierarchical clustering analysis recommended this module as another promising biomarker for lung adenocarcinoma, and the AUCs also reach up to 0.99 and 0.914 in the LAC72s and LACtcga datasets, respectively (Figure 4F and Supplementary Figure S9). Because the two independent datasets are microarray data without non-coding RNA expression, we only validated it in TCGA dataset. This module also showed robust by the re-sampling tests (Supplementary Figure S7 C).

Currently, several published studies have dedicated in identifying diagnostic biomarkers in lung adenocarcinoma based on gene expression, miRNA expression, DNA methylation, etc. For example, Melissa Rotunno et al found eight genes as a biomarker with AUC=0.81 based on gene expression data[35]. Zeng XL et al discovered that low expression of miR-143 and high expression of miR-150 for distinguishing lung adenocarcinoma patients from healthy subjects respectively with AUC=0.885 and AUC=0.834 [36]. In addition, Anglim PP et al identified a panel of DNA methylation markers for lung cancer ranged from a modest AUC =0.75 for PITX2 to a much better AUC=0.9 for GDNF[37]. Thus, the prediction power of ceRNA modules we identified is relatively high compared with biomarkers from these published studies. In our study, we integrated multiple types of information to indentify the biomarkers, including competitive relationship among the mRNAs, mRNA expression, and miRNA-target regulations. We proposed that the integrated information may contribute to the higher AUC values of ceRNA modules in our study.

## Discussion

In this study, we constructed the LDCCNet using a computational method. As general biological networks, the LDCCNet is scale-free, and modularity. There are abundant indirect interactions where two linked ceRNAs are also connected through a third ceRNA in the LDCCNet which greatly amplify dysregulated ceRNA cross-talk. Then, we found that gain ceRNA regulations may be one of the ways to guarantee the consistent expression of ceRNA pairs in cancer. Moreover, gain ceRNAs under stronger gain regulation are more likely to be lowly expressed in cancer while loss ceRNAs’ expression has the rising trend. We concluded that aberrant expression partly mediated by dysregulated ceRNAs in cancer may be controlled by these rules. Gain and loss ceRNAs participate in lots of cancer-related biological processes. Finally, several dysregulated ceRNA modules significantly regulated by known lung cancer miRNAs are identified and have the power to distinguish cancer from normal samples. In addition, another independent expression data from TCGA was used to re-construct the dysregulated ceRNA-ceRNA network and we found that patterns and the characteristics of the dysregulated ceRNA-ceRNA network are robust (details in Supplementary Text S2).

Increasing evidences point that non-coding RNAs can operate cellular processes through a variety of mechanisms and the aberrant regulation of non-coding RNAs might contribute to cancer phenotypes [38]. In our results, we discovered that lncRNAs and pseudogenes can potentially interact with miRNAs and act as ceRNAs to affect the competing RNAs’ level underpinning cancer development. By systematically analyzing their regulatory features in three different groups of RNAs, coding genes, lncRNAs and pseudogenes, we found that coding RNAs are under more wide and strict miRNA regulation compared with non-coding RNAs (Figure 5A and B). From the view of miRNAs, we observed that a few miRNAs exhibit extensive regulation to non-coding RNAs, whereas many miRNAs interact with a relatively small number of non-coding RNAs, especially for pseudogenes (Figure 5C). In addition, if a miRNA tends to regulate many coding genes, the miRNA also exhibits strong regulation to non-coding RNAs (Figure 5D). Some non-coding ceRNAs indeed play a key role in the LDCCNet. For example, heat shock 60kDa protein 1 pseudogene 1(HSPD1-2P), is a hub in the LDCCNet. At the same time, HSPD1-2P is a highly expressed pseudogene in lung adenocarcinoma. Most of the genes regulated by HSPD1-2P are involved in cell cycle and up-expressed in lung adenocarcinoma. These dysregulated ceRNA interactions may make the cell cycle more active to promote the development of lung adenocarcinoma. One of HSPD1-2P ceRNA partners is HLJ1 as a tumor suppressor in lung adenocarcinoma by inhibiting cell proliferation [39]. The ceRNA regulation of HSPD1-2p to HLJ1 losses in lung adenocarcinoma and HLJ1 expression is lower in lung adenocarcinoma than adjacent normal tissues. This loss ceRNA regulation may be as a potential cause for down-expression of HLJ1.

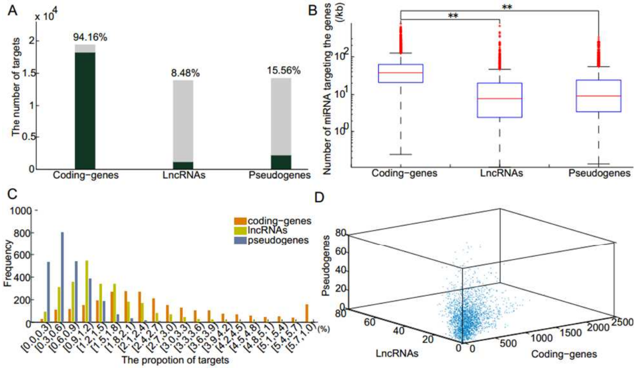

**Figure 5**. Coding-genes, lncRNAs, and pseudogenes targeted by miRNAs.

(A) Stacking barplots representing the percentages of coding-genes, lncRNAs, and pseudogenes regulated by miRNAs respectively. (B) Box plots representing the number of miRNA targeting coding-genes, lncRNAs, and pseudogenes per KB respectively. **P < 0.005 for Wilcoxon rank sum test (C) The distributions of the propotion of targets (coding-genes, lncRNAs, and pseudogenes) for miRNAs. (D) Three-dimensional scatter plot displays relationship among the number of coding-gene targets, lncRNA targets, and pseudogene targets for miRNAs.

Some studies have analyzed topology characteristics of known disease genes in different networks. Here, we also analyzed the properties of known lung cancer genes. The LDCCNet contains 84 known lung cancer genes. These cancer genes in the LDCCNet have higher clustering coefficients than random conditions (Supplementary Figure S10A). Additionally, the average shortest path length among lung cancer genes is 4.2639, significantly lower than random conditions, indicating that lung cancer genes are close proximity each other in the LDCCNet (Supplementary Figure S10B). Based on these characteristics of known cancer genes in the LDCCNet, we ranked the ceRNAs in the LDCCNet by integrating the number of known lung cancer genes in neighbors and the distance to the known lung cancer genes to provide clues for predicting lung cancer gene candidates (Supplementary Table S3 and Supplementary Text S1). The smaller the rank of a ceRNA is, the more possibly the ceRNA serves as a lung cancer gene. Top-ranked ceRNAs are significantly more often associated with known lung cancer genes than expected (Supplementary Figure S11). CAV2 ranks fifth and it is involved in essential cellular functions, including cellular growth control and apoptosis as described above. CAV2 is a hub ceRNA in LDCCNet and the expression of it is significantly down-regulated in lung adenocarcinoma with 0.22 fold change (FDR=2.95e-60). Thus, CAV2 may be as a potential lung cancer gene and its roles are worth to be further studied in lung adenocarcinoma. In addition, we also extended lung cancer gene set from integrating the genes with frequency of mutations >1% based on the Level 4 data from Firehose (http://gdac.broadinstitute.org/) and lung cancer genes from three reports[40-42]. These lung cancer genes in the LDCCNet also have higher clustering coefficients than random conditions (Supplementary Figure S12A). At the same time, the average shortest path length among these lung cancer genes is 4.3164, significantly lower than random conditions (Supplementary Figure S12B). We also predicted lung cancer genes in the LDCCNet using the same method. We found that top-ranked ceRNAs are also significantly more often associated with known lung cancer genes than expected (Supplementary Figure S13). These analysis results indicated that the LDCCNet can provide clues for predicting lung cancer gene candidates.

The change of ceRNA interactions might have direct relationship with the expression of ceRNA itself, the miRNA expression and the regulation between miRNAs and ceRNAs. The competitive interactions are mediated by miRNAs, thus the expression changes of miRNAs might influence the ceRNA interactions. For example, the expression of hsa-mir-192 is up-regulated about 21 times by analyzing TCGA miRNA expression data, which might mediate two loss ceRNA interactions, KLHL15 and CXCL2, LACTB2 and ARHGAP11A. On the other hand, for a dys-regulated ceRNA pair, the expression changes of one ceRNA could influence the expression of its competitive interactor and cause the dysregulation of the ceRNA interaction. Increasing evidences have shown that several levels of genomic alterations could regulate the ceRNAs' abundance in cancer, such as DNA deletions, amplifications or chromosomal translocations. In the LDCCNet, CDKN2B as a cell growth regulator has six dysregulations, all of which are loss ceRNA interactions. CDKN2B is frequently deleted in lung adenocarcinoma, accounting for 56% of the samples in TCGA. Additionally, the regulation alterations (gain or loss) between miRNAs and target mRNAs could also cause the ceRNA dysregulations, such as 3'UTR shortening, mutations in MREs and so on. Totally, changes of one of these levels might bring out the dysregualtion of ceRNA interactions.

## Conclusions

In conclusion, our study provides a global view of dysregulatory ceRNA-ceRNA network in lung adenocarcinoma by a new computational pipeline. Additionally, dysregulated ceRNA interactions are found contributing to the aberrant expression in cancer. Several dysregulated ceRNA modules are identified and have the power to distinguish cancer from normal samples. Although experimental validation of these dysregulated ceRNA interactions will be required to further estimate their roles in cancer, our findings in this work can serve as a significant foundation for further investigating the pathogenesis of caner and developing biomarkers.

## Materials and methods

### Genome annotation

Human genome (hg19) was used as a reference genome. GencodeV19 annotation file was obtained from Gencode [43]. Protein-coding transcripts, lncRNA transcripts and pseudogene transcripts were respectively defined following Li *et al* [44]. Meanwhile, we extracted 3'UTR sequences of protein-coding genes and DNA sequences of the whole human genome from the UCSC. The lncRNA and pseudogene sequences were respectively obtained based on the genomic coordinates of lncRNA and pseudogene transcripts. When multiple annotated 3'UTR/lncRNA/pseudogene sequences were available for a protein-coding gene/lncRNA/pseudogene, the longest one was chosen for analyses. Known mature miRNA sequences of Human were downloaded from miRBaseV20 [45].

### Expression profile datasets

Raw data from previously sequenced lung adenocarcinoma RNA-Seq datasets were downloaded: from GEO under accession number GSE40419 denoted by “LAC72s” [15] including 72 paired samples of lung adenocarcinomas and adjacent normal tissues. Sequence reads were aligned to the human genome (hg19) using the TopHatV2.0.9 with default parameters [46]. The resulting alignment data from TopHat were then fed to Cufflinks V2.1.1 to perform transcript assembly and abundance estimation [47]. Gene expression FPKM (Fragments Per Kilobase of exon per million fragments mapped) values were calculated with Cufflinks using gencodeV19 annotation file. Genes with FPKM > 0.001 were considered as expressed, and those genes expressed in less than 90% of tumor or tumor-adjacent normal tissues were filtered out for the following analysis.

Three independent datasets were used to test the power of modules as biomarkers. Level 3 IlluminaHiseq gene datasets of human lung tissues were obtained from TCGA, denoted by “LACtcga”, including 40 paired lung adenocarcinomas and adjacent non-tumor tissues, plus an additional 6 unmatched adjacent non-tumor tissues and 203 unmatched lung adenocarcinomas. Normalized HG-U133A array dataset with 58 adenocarcinoma and 49 non-tumor tissue samples was downloaded from GEO (GSE10072), and denoted by “LAC107”. Another dataset was also downloaded from GEO (GSE32863), denoted by “LAC58s”, which is the expression data of 58 paired lung adenocarcinoma and adjacent non-tumor lung fresh frozen tissues profiled by Illumina HumanWG-6 v3.0 expression beadchip.

### Lung cancer genes, lncRNAs and miRNAs

We widely collected experiment validated the lung cancer-related genes, lncRNAs and miRNAs from different databases. Lung cancer genes were collected by integrating LuGenD [48], OMIM [49], TSGDB [50], and methycancer [51]. Lung cancer lncRNAs were from LncRNADisease [52] and White *et al* [53]. In addition, lung cancer miRNAs were collected by integrating miRCancer [54], miR2Disease [55], HMDD [56] and OncomiRDB [57].

### miRNA target prediction

miRanda with default parameters was used to identify miRNA target sites in the 3’UTR of coding transcripts and full length of lncRNA transcripts and pseudogene transcripts [58]. All the CLIP-identified AGO binding sites peaks of 26 human AGO1/2 CLIP-seq datasets were directly downloaded from starBase v2.0 [44]. The target sites from miRanda that resided within any entry of the AgoCLIP peaks were considered as CLIP-supported sites. One miRNA was considered to regulate one target if and only if the miRNA had at least one CLIP-supported site in the target.

### Identifying the dysregulated ceRNA-ceRNA interactions

In order to identify dysregulated ceRNA-ceRNA interactions contributing aberrant expression in disease, we developed a computational method by integrating miRNA regulation with the expression profiles in disease and normal. Firstly, candidate ceRNA-ceRNA interactions were predicted by sharing significantly more miRNAs. Secondly, dysregulated ceRNA-ceRNA interactions were identified in the context of disease. Finally, ceRNAs in dysregulated ceRNA-ceRNA interactions were required to be differentially expressed in disease.

Concretely, the computational approach encompasses the following three steps:

i. Predicting candidate ceRNA-ceRNA interactions

For a given RNA pair (RNA A and RNA B), we first identified miRNAs that regulate the two RNAs, and then hypergeometric test was used to measure whether these two RNAs significantly share miRNAs. The probability P is calculated as according to

$$P = 1 - F(x|N,K,M) = 1 - \sum_{t=0}^{x-1} \frac{\binom{K}{t}\binom{N-K}{M-t}}{\binom{N}{M}}$$

where N is the number of all human miRNAs (default background distribution), K represents the total number of miRNAs regulating A, M represents the total number of miRNAs regulating B, x is the number of shared miRNAs between A and B and is required to be at least three. We controlled for multiple hypotheses using the false discovery rate (FDR), and only pairs passing an FDR of 0.05 were considered to be significantly co-regulated and as candidate ceRNA pairs.

ii. Identifying dysregulated ceRNA-ceRNA interactions in the context of disease

We used the change of correlation of the ceRNA pair’s expression in cancer samples compared with normal samples to determine the extent of dysregulation. Then, if the ceRNA pair’s expression for cancer samples is more obvious positive correlation than normal samples, the ceRNA pair is defined as “gain” dysregulation. If the ceRNA pair’s expression for normal samples is more obvious positive correlation than cancer samples, the ceRNA pair is defined as “loss” dysregulation. The Pearson correlation coefficients (PCC) of between the expression profiles of each ceRNA pair were calculated to measure the correlation. We defined the PCC difference of a ceRNA pair between cancer samples and normal samples:

$$\Delta R = corr_{cancer}(A,B) - corr_{normal}(A,B)$$

$corr_{cancer}(A,B)$ is the PCC estimated from the cancer samples between A and B, while $corr_{normal}(A,B)$ is estimated from the normal samples.

To determine whether $\Delta R$ was statistically significant, randomization tests were performed and realized by permuting the cancer/normal labels of all samples. For each ceRNA pair, the $\Delta R$ was calculated after permutation and the procedure was repeated 1,000 times. The significant P-value of each $\Delta R$ was given as the frequency of the $\Delta R$ values in random conditions, which was greater than the value in the real condition. A bonferroni-corrected p-value of 0.05 was used as threshold.

In order to further identify ceRNA interactions with dominant change, we required that for the gain ceRNA interaction in cancer, $\Delta R$ should be significant, $\Delta R > 0.5$ and $corr_{cancer}(A, B)$ is significantly positive and for the loss ceRNA interaction in cancer, $\Delta R$ should be significant, $\Delta R < -0.5$ and $corr_{normal}(A, B)$ is significantly positive.

iii. Requiring ceRNAs with differential expression patterns

To reflect the regulation effect of dysregulated ceRNA interactions on the phenotype, we further required that ceRNAs were differentially expressed. EdgeR was used to estimate differential expression with default parameters [59]. ceRNAs with a FDR < 0.005 and a fold change > 1.5 were considered to be differentially expressed.

After identifying all significantly dysregulated ceRNA pairs, we generated the dysregulated ceRNA-ceRNA network. Vertices in the network represent dysregulated ceRNAs with differential expression. An undirected edge between the ceRNA pair exists if their relationship is significantly dysregulated (gain or loss), and the weight of the edge is set to the absolute value of PCC difference.

### Gain ceRNAs and Loss ceRNAs

The ceRNAs which were significantly enriched with gain ceRNA interactions were defined as gain ceRNAs and the ceRNAs with significant loss ceRNA interactions were defined as loss ceRNAs. The hypergeometric test was used to measure statistically significant (Supplementary Text S1). A ceRNA with bonferroni-corrected p-value<0.05 was considered to be gain or loss ceRNA.

### Functional analysis of ceRNAs

Hypergeometric test was used to identify the significantly overrepresented biological function categories of a specific ceRNA set based on the GO database [60]. Functional categories with a bonferroni-corrected p-value less than 0.05 and annotated by at least three ceRNAs were considered in our analyses. We used the same method to perform enrichment analysis for each hallmark of cancer.

### Identification of diagnostic biomarkers

In order to evaluate the potential of dysregulated ceRNA modules as biomarkers, the scoring classifier was constructed. For each ceRNA module, we first performed Z-score transformation on the expression levels across the samples for each member of the module and then summarized the Z-scores as the integrated expression signature (score). Then, the samples could be divided into two classes (normal and tumor) by choosing a cutoff. And the receiver operation characteristic (ROC) curve was used for classifier evaluation which was drawn by plotting sensitivity against the false-positive rate. This procedure was performed by the R package ROCR [61]. In addition, unsupervised hierarchical cluster analysis was carried out to visually display the classification performance. Complete linkage and 1-Pearson correlation as a distance measure were used to do hierarchical clustering of the expression profile.

## Conflict of interest

The authors declare no conflict of interest.

## Authors' contributions

XL, JX and YL designed the study, ST and AW carried out the data analysis and drafted the manuscript. JC and HC performed the module analyses and drafted the manuscript. JL and JB performed the function enrichment analyses. All authors read and approved the final manuscript.

## Acknowledgements

This work was supported 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, and 61203264], the China Postdoctoral Science Foundation [Grant No.2014T70364, 2015M571436 and LBH-Z14134], Natural Science Foundation of Heilongjiang Province [QC2015020], WeihanYu Youth Science Fund Project of Harbin Medical University and Harbin Special Funds of Innovative Talents on Science and Technology Research Project [Grant No. RC2015QN003080] and the Innovation Research Fund for Graduate Students of Heilongjiang province [YJSCX2012-251HLJ].

## Notes and references

*College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China. Fax: 86-451-86615922; Tel: 86-451-86615922; E-mail: lixia@hrbmu.edu.cn,xujuanbiocc@ems.hrbmu.edu.cn, liyongsheng@ems.hrbmu.edu.cn*

† Electronic Supplementary Information (ESI) available: [Supplementary data.doc, Supplementary Table S1.xls, Supplementary Table S2.xls, Supplementary Table S3.xls]. See DOI: 10.1039/b000000x/

1. J. Xu, C. X. Li, J. Y. Lv, Y. S. Li, Y. Xiao, T. T. Shao, X. Huo, X. Li, Y. Zou, Q. L. Han, X. Li, L. H. Wang and H. Ren, *Molecular cancer therapeutics*, 2011, **10**, 1857-1866.
2. D. Baek, J. Villen, C. Shin, F. D. Camargo, S. P. Gygi and D. P. Bartel, *Nature*, 2008, **455**, 64-71.
3. D. P. Bartel, *Cell*, 2009, **136**, 215-233.
4. L. Salmena, L. Poliseno, Y. Tay, L. Kats and P. P. Pandolfi, *Cell*, 2011, **146**, 353-358.
5. M. S. Kumar, E. Armenteros-Monterroso, P. East, P. Chakravorty, N. Matthews, M. M. Winslow and J. Downward, *Nature*, 2014, **505**, 212-217.
6. L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman and P. P. Pandolfi, *Nature*, 2010, **465**, 1033-1038.
7. J. Wang, X. Liu, H. Wu, P. Ni, Z. Gu, Y. Qiao, N. Chen, F. Sun and Q. Fan, *Nucleic Acids Res*, 2010, **38**, 5366-5383.
8. J. Tang, H. Zhuo, X. Zhang, R. Jiang, J. Ji, L. Deng, X. Qian, F. Zhang and B. Sun, *Cell death & disease*, 2014, **5**, e1549.
9. M. Narayanan, J. L. Huynh, K. Wang, X. Yang, S. Yoo, J. McElwee, B. Zhang, C. Zhang, J. R. Lamb, T. Xie, C. Suver, C. Molony, S. Melquist, A. D. Johnson, G. Fan, D. J. Stone, E. E. Schadt, P. Casaccia, V. Emilsson and J. Zhu, *Molecular systems biology*, 2014, **10**, 743.
10. P. Sumazin, X. Yang, H. S. Chiu, W. J. Chung, A. Iyer, D. Llobet-Navas, P. Rajbhandari, M. Bansal, P. Guarnieri, J. Silva and A. Califano, *Cell*, 2011, **147**, 370-381.
11. P. Paci, T. Colombo and L. Farina, *BMC systems biology*, 2014, **8**, 83.

12. L. Li, D. Wang, M. Xue, X. Mi, Y. Liang and P. Wang, *Scientific reports*, 2014, **4**, 5406.
13. A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward and D. Forman, *CA: a cancer journal for clinicians*, 2011, **61**, 69-90.
14. N. Cancer Genome Atlas Research, *Nature*, 2014, **511**, 543-550.
15. J. S. Seo, Y. S. Ju, W. C. Lee, J. Y. Shin, J. K. Lee, T. Bleazard, J. Lee, Y. J. Jung, J. O. Kim, J. Y. Shin, S. B. Yu, J. Kim, E. R. Lee, C. H. Kang, I. K. Park, H. Rhee, S. H. Lee, J. I. Kim, J. H. Kang and Y. T. Kim, *Genome research*, 2012, **22**, 2109-2119.
16. X. Su, J. Xing, Z. Wang, L. Chen, M. Cui and B. Jiang, *Chinese journal of cancer research = Chung-kuo yen cheng yen chiu*, 2013, **25**, 235-239.
17. U. Ala, F. A. Karreth, C. Bosia, A. Pagnani, R. Taulli, V. Leopold, Y. Tay, P. Provero, R. Zecchina and P. P. Pandolfi, *Proc Natl Acad Sci U S A*, 2013, **110**, 7154-7159.
18. M. Strazisar, V. Mlakar and D. Glavac, *Lung cancer*, 2009, **64**, 257-262.
19. S. V. Sharma, D. W. Bell, J. Settleman and D. A. Haber, *Nature reviews. Cancer*, 2007, **7**, 169-181.
20. S. Wang, X. Li, Z. G. Li, J. Lu, W. Y. Fang, Y. Q. Ding and K. T. Yao, *International journal of cancer. Journal international du cancer*, 2011, **128**, 753-762.
21. S. L. Yu, H. Y. Chen, G. C. Chang, C. Y. Chen, H. W. Chen, S. Singh, C. L. Cheng, C. J. Yu, Y. C. Lee, H. S. Chen, T. J. Su, C. C. Chiang, H. N. Li, Q. S. Hong, H. Y. Su, C. C. Chen, W. J. Chen, C. C. Liu, W. K. Chan, W. J. Chen, K. C. Li, J. J. Chen and P. C. Yang, *Cancer cell*, 2008, **13**, 48-57.
22. D. Hanahan and R. A. Weinberg, *Cell*, 2011, **144**, 646-674.
23. C. L. Plaisier, M. Pan and N. S. Baliga, *Genome research*, 2012, **22**, 2302-2314.
24. E. Wang, N. Zaman, S. McGee, J. S. Milanese, A. Masoudi-Nejad and M. O'Connor-McCourt, *Seminars in cancer biology*, 2015, **30**, 4-12.
25. N. Zaman, L. Li, M. L. Jaramillo, Z. Sun, C. Tibiche, M. Banville, C. Collins, M. Trifiro, M. Paliouras, A. Nantel, M. O'Connor-McCourt and E. Wang, *Cell reports*, 2013, **5**, 216-223.
26. E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller and N. Friedman, *Nat Genet*, 2003, **34**, 166-176.
27. B. J. Frey and D. Dueck, *Science*, 2007, **315**, 972-976.
28. F. Kottakis, C. Polytarchou, P. Foltopoulou, I. Sanidas, S. C. Kampranis and P. N. Tsichlis, *Molecular cell*, 2011, **43**, 285-298.
29. D. Y. Lu, X. H. Mao, Y. H. Zhou, X. L. Yan, W. P. Wang, Y. B. Zheng, J. J. Xiao, P. Zhang, J. G. Wang, N. Ashwani, W. L. Ding, H. Jiang, Y. Shang and M. H. Wang, *Asian Pacific journal of cancer prevention : APJCP*, 2014, **15**, 5249-5252.
30. P. Ji, S. Diederichs, W. Wang, S. Boing, R. Metzger, P. M. Schneider, N. Tidow, B. Brandt, H. Buerger, E. Bulk, M. Thomas, W. E. Berdel, H. Serve and C. Muller-Tidow, *Oncogene*, 2003, **22**, 8031-8041.
31. J. Li, A. E. Lenferink, Y. Deng, C. Collins, Q. Cui, E. O. Purisima, M. D. O'Connor-McCourt and E. Wang, *Nat Commun*, 2010, **1**, 34.
32. A. Weber, J. Marquardt, D. Elzi, N. Forster, S. Starke, A. Glaum, D. Yamada, P. A. Defossez, J. Delrow, R. N. Eisenman, H. Christiansen and M. Eilers, *The EMBO journal*, 2008, **27**, 1563-1574.
33. A. Esquela-Kerscher, P. Trang, J. F. Wiggins, L. Patrawala, A. Cheng, L. Ford, J. B. Weidhaas, D. Brown, A. G. Bader and F. J. Slack, *Cell cycle*, 2008, **7**, 759-764.
34. M. S. Kumar, S. J. Erkeland, R. E. Pester, C. Y. Chen, M. S. Ebert, P. A. Sharp and T. Jacks, *Proceedings of the National Academy of Sciences of the United States of America*, 2008, **105**, 3903-3908.
35. M. Rotunno, N. Hu, H. Su, C. Wang, A. M. Goldstein, A. W. Bergen, D. Consonni, A. C. Pesatori, P. A. Bertazzi, S. Wacholder, J. Shih, N. E. Caporaso, P. R. Taylor and M. T. Landi, *Cancer Prev Res (Phila)*, 2011, **4**, 1599-1608.
36. X. L. Zeng, S. Y. Zhang, J. F. Zheng, H. Yuan and Y. Wang, *Chin Med J (Engl)*, 2013, **126**, 4510-4516.
37. P. P. Anglim, J. S. Galler, M. N. Koss, J. A. Hagen, S. Turla, M. Campan, D. J. Weisenberger, P. W. Laird, K. D. Siegmund and I. A. Laird-Offringa, *Mol Cancer*, 2008, **7**, 62.
38. T. Huang, A. Alvarez, B. Hu and S. Y. Cheng, *Chinese journal of cancer*, 2013, **32**, 582-593.
39. M. F. Tsai, C. C. Wang, G. C. Chang, C. Y. Chen, H. Y. Chen, C. L. Cheng, Y. P. Yang, C. Y. Wu, F. Y. Shih, C. C. Liu, H. P. Lin, Y. S. Jou, S. C. Lin, C. W. Lin, W. J. Chen, W. K. Chan, J. J. Chen and P. C. Yang, *Journal of the National Cancer Institute*, 2006, **98**, 825-838.
40. E. C. de Bruin, N. McGranahan, R. Mitter, M. Salm, D. C. Wedge, L. Yates, M. Jamal-Hanjani, S. Shafi, N. Murugaesu, A. J. Rowan, E. Gronroos, M. A. Muhammad, S. Horswell, M. Gerlinger, I. Varela, D. Jones, J. Marshall, T. Voet, P. Van Loo, D. M. Rassl, R. C. Rintoul, S. M. Janes, S. M. Lee, M. Forster, T. Ahmad, D. Lawrence, M. Falzon, A. Capitanio, T. T. Harkins, C. C. Lee, W. Tom, E. Teefe, S. C. Chen, S. Begum, A. Rabinowitz, B. Phillimore, B. Spencer-Dene, G. Stamp, Z. Szallasi, N. Matthews, A. Stewart, P. Campbell and C. Swanton, *Science*, 2014, **346**, 251-256.
41. J. Zhang, J. Fujimoto, J. Zhang, D. C. Wedge, X. Song, J. Zhang, S. Seth, C. W. Chow, Y. Cao, C. Gumbs, K. A. Gold, N. Kalhor, L. Little, H. Mahadeshwar, C. Moran, A. Protopopov, H. Sun, J. Tang, X. Wu, Y. Ye, W. N. William, J. J. Lee, J. V. Heymach, W. K. Hong, S. Swisher, Wistuba, II and P. A. Futreal, *Science*, 2014, **346**, 256-259.
42. R. Califano, A. Abidin, N. U. Tariq, P. Economopoulou, G. Metro and G. Mountzios, *Cancer Treat Rev*, 2015, **41**, 401-411.
43. J. Harrow, A. Frankish, J. M. Gonzalez, E. Tapanari, M. Diekhans, F. Kokocinski, B. L. Aken, D. Barrell, A. Zadissa, S. Searle, I. Barnes, A. Bignell, V. Boychenko, T. Hunt, M. Kay, G. Mukherjee, J. Rajan, G. Despacio-Reyes, G. Saunders, C. Steward, R. Harte, M. Lin, C. Howald, A. Tanzer, T. Derrien, J. Chrast, N. Walters, S. Balasubramanian, B. Pei, M. Tress, J. M. Rodriguez, I. Ezkurdia, J. van Baren, M. Brent, D. Haussler, M. Kellis, A. Valencia, A. Reymond, M. Gerstein, R. Guigo and T. J. Hubbard, *Genome research*, 2012, **22**, 1760-1774.
44. J. H. Li, S. Liu, H. Zhou, L. H. Qu and J. H. Yang, *Nucleic acids research*, 2014, **42**, D92-97.
45. A. Kozomara and S. Griffiths-Jones, *Nucleic Acids Res*, 2014, **42**, D68-73.
46. C. Trapnell, L. Pachter and S. L. Salzberg, *Bioinformatics*, 2009, **25**, 1105-1111.
47. C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold and L. Pachter, *Nature biotechnology*, 2010, **28**, 511-515.
48. Lung Cancer Gene Database. http://www.bioinformatics.org/lugend/.
49. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic acids research*, 2005, **33**, D514-517.
50. Y. Yang and L. M. Fu, *Bioinformatics*, 2003, **19**, 2311-2312.
51. X. He, S. Chang, J. Zhang, Q. Zhao, H. Xiang, K. Kusonmano, L. Yang, Z. S. Sun, H. Yang and J. Wang, *Nucleic acids research*, 2008, **36**, D836-841.
52. G. Chen, Z. Wang, D. Wang, C. Qiu, M. Liu, X. Chen, Q. Zhang, G. Yan and Q. Cui, *Nucleic acids research*, 2013, **41**, D983-986.
53. N. M. White, C. R. Cabanski, J. M. Silva-Fisher, H. X. Dang, R. Govindan and C. A. Maher, *Genome biology*, 2014, **15**, 429.
54. B. Xie, Q. Ding, H. Han and D. Wu, *Bioinformatics*, 2013, **29**, 638-644.
55. Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu, *Nucleic acids research*, 2009, **37**, D98-104.
56. M. Lu, Q. Zhang, M. Deng, J. Miao, Y. Guo, W. Gao and Q. Cui, *PloS one*, 2008, **3**, e3420.
57. D. Wang, J. Gu, T. Wang and Z. Ding, *Bioinformatics*, 2014, **30**, 2237-2238.
58. A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander and D. S. Marks, *Genome biology*, 2003, **5**, R1.
59. M. D. Robinson, D. J. McCarthy and G. K. Smyth, *Bioinformatics*, 2010, **26**, 139-140.
60. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock, *Nat Genet*, 2000, **25**, 25-29.
61. T. Sing, O. Sander, N. Beerenwinkel and T. Lengauer, *Bioinformatics*, 2005, **21**, 3940-3941.