Dynamical network analysis reveals key microRNAs in progressive stages of lung cancer

Chao Kong1,2,3, Yu-Xiang Yao3, Zhi-Tong Bing4,5,6, Bing-Hui Guo7, Liang Huang4,3, Zhi-Gang Huang1,2,*, Ying-Cheng Lai8,9

1 The Key Laboratory of Biomedical Information Engineering of Ministry of Education, Institute of Health and Rehabilitation Science, School of Life Science and Technology, Xi’an Jiaotong University, The Key Laboratory of Neuro-informatics & Rehabilitation Engineering of Ministry of Civil Affairs, Xi’an, Shaanxi, P. R. China, 2 National Engineering Research Center for Healthcare Devices, Guangzhou, Guangdong, P. R. China, 3 Institute of Computational Physics and Complex Systems, School of Physical Science and Technology, Lanzhou University, Lanzhou, China, 4 Evidence Based Medicine Center, School of Basic Medical Science of Lanzhou University, Lanzhou, China, 5 Key Laboratory of Evidence Based Medicine and Knowledge Translation of Gansu Province, Lanzhou, China, 6 Department of Computational Physics, Institute of Modern Physics, Chinese Academy of Sciences, Lanzhou, China, 7 Beijing Advanced Innovation Center for Big Data and Brain Computing, LMIB and School of Mathematics and System Sciences, Beihang University, Beijing, China, 8 School of Electrical, Computer and Energy Engineering, Arizona State University, Tempe, Arizona, United States of America, 9 Department of Physics, Arizona State University, Tempe, Arizona, United States of America

* huangzg@xjtu.edu.cn

Abstract

Non-coding RNAs are fundamental to the competing endogenous RNA (CeRNA) hypothesis in oncology. Previous work focused on static CeRNA networks. We construct and analyze CeRNA networks for four sequential stages of lung adenocarcinoma (LUAD) based on multi-omics data of long non-coding RNAs (lncRNAs), microRNAs and mRNAs. We find that the networks possess a two-level bipartite structure: common competing endogenous network (CCEN) composed of an invariant set of microRNAs over all the stages and stage-dependent, unique competing endogenous networks (UCENs). A systematic enrichment analysis of the pathways of the mRNAs in CCEN reveals that they are strongly associated with cancer development. We also find that the microRNA-linked mRNAs from UCENs have a higher enrichment efficiency. A key finding is six microRNAs from CCEN that impact patient survival at all stages, and four microRNAs that affect the survival from a specific stage. The ten microRNAs can then serve as potential biomarkers and prognostic tools for LUAD.

Author summary

Lung cancer is the leading cause of cancer-related human deaths worldwide. Lung adenocarcinoma is one of the most common subtypes, and has more pronounced genomic variations than other lung cancer subtypes. A milestone discovery in cancer research is the roles played by non-coding RNAs which have been identified as the oncogenic drivers...
and tumor suppressors. In cancer development, non-coding RNAs form an inseparable unity of RNA-level regulating networks in the intracellular environment, and the dynamical interplay and competition among different types of RNAs are playing a pivotal role. We have developed a quantitative approach to reconstructing the the mutual regulation networks of RNAs for the progressive stages of lung adenocarcinoma at the post-transcriptional level. Our analysis revealed the emergence of two characteristically distinct types of networks that possess a two-level bipartite structure, and we uncovered a number of key genes that affect or even determine the survival of patients at each stage. Our work establishes a more comprehensive gene-data analysis framework than previous ones, not only providing a tool to probe more deeply into the mechanisms of cancer evolution than previously possible but also having the potential to lead to more effective biomarkers and drug targets for lung cancer.

Introduction

Lung cancer is the leading cause of cancer-related human deaths worldwide [1]. Approximately 85% of the lung cancer cases can be classified as non small-cell lung cancer (NSCLC), among which lung adenocarcinoma (LUAD) is one of the most common subtypes [2]. The mechanisms behind cancer evolution are extremely complex, impeding accurate and reliable prognosis as well as effective treatment [3]. A standard existing approach to monitoring tumor progress and detecting/ascertaining the underlying mechanism is mRNA transcriptomics, which has led to a large number of significant prognostic biomarkers and therapeutic targets [4–6]. When combined with other omics information, such as microRNA, methylation and epigenetic data, mRNA transcriptomics has provided valuable insights into the mechanisms of lung cancers [7–10].

A milestone discovery in cancer research is the roles played by non-coding RNAs (ncRNAs). In the general developmental and disease contexts, at the mRNA level, ncRNAs have been found in key regulators of physiological functions [11–15]. Particularly relevant to cancer, ncRNAs have been identified as the oncogenic drivers and tumor suppressors in major cancer types [16]. The discovery of the pivotal role of ncRNAs in cancer has led to the paradigmatic, competing endogenous RNA (CeRNA) hypothesis [17–19]: in cancer emergence and development, microRNAs, mRNAs, and ncRNAs form an inseparable unity of RNA-level regulating network in the intracellular environment, collectively known as the CeRNA network. For example, it has been known [8, 18] that RNA-induced silencing complex (RISC) can be produced through binding of microRNA response elements (MREs) at RNAs 3'UTR, causing inactivation of the target mRNA, but this mechanism is also present in non-coding RNAs, where the target RNA usually has multiple MREs. There has been increasing evidence for the fundamental roles played by CeRNAs in biological systems [16, 18] and, as a result, studying cancer-related CeRNA networks based on RNA-sequence expression data has gained momentum. For example, dysregulated CeRNA-CeRNA interactions in the CeRNA networks of LUAD were analyzed [20, 21], suggesting that the gain or loss of CeRNAs can be used in functional analysis and as potential diagnostic biomarkers. The difference in the expression profiles between early and late stages in CeRNA networks of LUAD was noticed and some LUAD specific, long non-coding RNAs (lncRNAs) with their functional enrichment and clinical features were uncovered [22]. The possible roles played by lncRNAs through CeRNA networks in other types of cancer such as liver cancer and papillary thyroid cancer were also studied [5].
In existing studies of CeRNA networks, a commonly practiced methodology is to identify some differentially expressed lncRNAs, mRNAs and microRNAs based on absolute fold changes and values of the false positive ratio. Consequently, information used to construct the underlying CeRNA networks and to reveal the network functions with clinical implications is directly from data bases without any dynamical ingredient. The resulting CeRNA networks are thus simply a combination of different kinds of RNAs, whereas the dynamical interplay and competition among different types of RNAs were completely ignored. To remedy this deficiency has motivated our work.

We hold the belief that cancer development and evolution are fundamentally a dynamic process. Manifested in the underlying CeRNA networks, it is unlikely that the intrinsic interactions, interplay and the network structure remain static during cancer development. To better understand cancer and to identify more effective biomarkers, the dynamical aspects of the CeRNA networks must be taken into account. This is in line with the field of network physiology [23–27], where transitions and reorganization occur in the networks of physiological/organ systems in the human body on larger spatial and temporal scales, which can be constructed using readily accessible continuous time series. In the CeRNA network analysis, it is infeasible to collect RNA data in continuous time. Nonetheless, with the presently available gene data, we are able to incorporate the time axis into the analysis but only for a limited set of time intervals. In particular, we focus on the four stages of LUAD progression to construct, based on both RNA expression and clinical data of LUAD from the cancer genome atlas (TCGA), the lncRNA-microRNA-mRNA CeRNA network for each stage. For any given stage, our quantitative approach to reconstructing the CeRNA network consists of analyzing differentially expressed RNAs, matching microRNA targets by base complementary pairing, and selecting the negative correlation by invoking the CeRNA hypothesis. We then calculate the fold changes and the average expression level of RNAs in the CeRNA networks for the four stages and carry out Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses to validate the results. Our analysis reveals the emergence of two characteristically distinct types of networks that play an important role in the progression of LUAD from stage I to stage IV: one is common to all four stages, which we name as the common competing endogenous RNA network (CCEN), and another is unique for each stage, which we call as the unique competing endogenous RNA networks (UCENs). Analyzing the properties of CCEN and UCENs, we uncover a number of key genes that affect or even determine the survival of patients at each stage of LUAD: six microRNAs from the CCEN which affect the survival of LUAD patients at all stages, and four other microRNAs that influence the survival at each specific stage. Our work establishes a more comprehensive gene-data analysis framework than previous ones, not only providing a tool to probe more deeply into the mechanisms of cancer evolution than previously possible but also having the potential to lead to more effective biomarkers and drug targets for LUAD as well as other types of cancer.

Results

Reconstructed lncRNA-microRNA-mRNA CeRNA networks and doubly bipartite structure

As described in Materials and Methods, to reconstruct the lncRNA-microRNA-mRNA CeRNA networks, we first obtain the differentially expressed RNAs by comparing the gene expression values in the LUAD samples with those from the normal samples of lncRNAs, microRNAs and mRNAs. We then apply the principle of base complementary pairing to match the microRNAs with lncRNAs and mRNAs, all differentially expressed, leading to values of the correlation between microRNA-lncRNA and microRNA-mRNA nodes.
only the negatively correlated pairs of microRNA-lncRNAs and microRNA-mRNAs, we obtain the final CeRNA networks. (S1 Table presents more details of the network reconstruction process).

Fig 1 presents examples of the reconstructed lncRNA-microRNA-mRNA CeRNA networks for the four stages of LUAD. A common feature for the networks in all four states is that direct connections exist only between microRNAs (red nodes) and lncRNAs (green nodes) or between the microRNAs and mRNAs (blue nodes). The connections between lncRNAs and
mRNAs are thus indirect: they occur through the microRNAs. Regarding the lncRNAs, mRNAs, and microRNAs as three entities (or "supernodes"), we see that the connection between the first two supernodes is through the third one, which is characteristic of mutualistic or bipartite type of interactions. The mutualism occurs at a finer scale because the subnetwork of microRNAs-lncRNAs (or that of microRNAs-mRNAs) also has a bipartite structure: there are no direct links among the microRNAs or among the lncRNAs, i.e., any connection between two nodes of the same color must be indirect and through a node of a different color. Such bipartite network structure arises commonly in ecology, e.g., the pollinator-plant mutualistic networks [28–36]. The striking feature here is that the lncRNA-microRNA-mRNA CeRNA networks for all four stages of LUAD possess a doubly bipartite structure: mutualism occurs at two distinct scales.

**Changes in fold-change value and expression**

Compared with RNAs from the direct differential expression analysis, the differentially expressed RNAs of the final CeRNA networks obtained through matching the differentially expressed microRNA targets and retaining only those RNA pairs with negative correlation have higher values of the expression of microRNAs and fold change, as shown in Fig 2. Our

![Fig 2](https://doi.org/10.1371/journal.pcbi.1007793.g002)
construction of the CeRNA networks thus result in more deviated RNAs relative to normal sample expression of LUAD.

**Validation of network reconstruction: Enrichment analysis of mRNAs in CeRNA networks**

Are the lncRNA-microRNA-mRNA CeRNA networks so constructed meaningful? To address this question, we perform gene ontology and KEGG pathway gene enrichment analysis of mRNAs in the CeRNA networks at the four stages of LUAD, with results in Fig 3. We find that the mRNAs in the CeRNA networks at all four stages are strongly correlated with blood vessel development (−log₁₀ P > 20), tissue morphogenesis (−log₁₀ P > 20) and regulation of cell adhesion (−log₁₀ P > 15), where P is the P-value (Materials and methods). Strong correlation
\((-\log_{10} P > 10)\) also exists between the mRNAs and factors such as extracellular matrix organization, cell-substrate adhesion, response to growth factor, mesenchyme development, developmental growth and negative regulation of cell proliferation. Remarkably, beyond the “static” information provided by the conventional gene ontology analysis of the four stages of LUAD, our CeRNA networks give rise to a dynamic scenario for tumor progression as the TNM stage deteriorates (see Materials and methods for the meaning of TNM). For example, mRNAs are more correlated with coronary vasculature development and less correlated with substance-dependent cell spreading in the early than the late two stages. In addition, mRNAs are more correlated with positive regulation of cell death and negative regulation of cell cycle \((-\log_{10} P > 7)\) and less correlated with lung morphogenesis \((-\log_{10} P < 6)\) in the late stages than the early three stages of LUAD.

The KEGG pathway enrichment analysis reveals that the mRNAs associated with pathway in cancer \((-\log_{10} P = 2.98, 4.75, 7.77, \text{ and } 10.2\) for the four stages, respectively) and cell cycle \((-\log_{10} P = 3.72, 4.11, 4.23, \text{ and } 8.54\) for the four stages, respectively) have stronger correlation in early than late TNM stages of LUAD. Additionally, some cancer-related pathways in the late stages are more significant than in the early stages, such as PI3K-Akt signaling pathway, focal adhesion, gap junction, transcriptional misregulation in cancer, proteoglycans in cancer, Rap1 signaling pathway, miRNAs in cancer, EGFR tyrosine kinase inhibitor resistance, and circadian entrainment. On the contrary, ECM-receptor interaction as well as protein digestion and absorption are less relevant to the LUAD late stages than to the early stages.

Taken together, the enrichment analysis of mRNAs reveals that our reconstructed lncRNA-microRNA-mRNA CeRNA networks exhibit a close correspondence to the development of LUAD and deterioration of physiological indicators from stage I to stage IV, validating our reconstruction method.

**CCEN and UCEN analysis**

To better understand the relationship between CeRNA networks and TNM stages of LUAD, we extract the CCEN and UCENs from the lncRNA-microRNA-mRNA CeRNA networks, as shown in Fig 4.

Fig 5A shows the results of gene ontology enrichment analysis. We see that the mRNAs of the UCENs are correlated with blood vessel development, tissue morphogenesis, regulation of cell adhesion, extracellular matrix organization, cell-substrate adhesion and regulation of growth \((-\log_{10} P > 5)\). Consistent with the results of enrichment analysis of the mRNAs in the CeRNA networks (Fig 3), the UCENs can also distinguish the biological characteristics between early and later stages of the cancer. In particular, the UCENs from the late stages are more correlated with regulation of growth, negative regulation of cell adhesion, positive regulation of cell death, substrate adhesion-dependent cell spreading, regulation of cell division and vascular process in circulatory system than those in the early stage. However, extracellular matrix organization and mesenchyme development in the fourth stages do not exhibit significant changes.

From the KEGG pathway enrichment analysis (Fig 5B), we find that the CCEN network is relevant to pathways in cancer, small cell lung cancer, PI3K-Akt signaling pathway, ECM-receptor interaction, focal adhesion and cell cycle, while all these pathways belonging to the later two stages show a stronger correlation than those in the early two stages.

**Enrichment efficiency of microRNA-linked mRNAs in UCENs**

Our reconstructed CeRNA networks, UCEN, and CCEN have 868, 585, and 283 mRNAs altogether, respectively. In comparison, in the UCEN there are 201 microRNA-linked mRNAs.
Fig 4. Representative CCEN and UCENs. (A) CCEN extracted from the lncRNA-microRNA-mRNA CeRNA networks of the four stages of LUAD. (B-E) UCENs corresponding to the four stages. The green, red, and blue nodes represent lncRNAs, microRNAs, and mRNAs, respectively. The edges represent mutual regulation between microRNA-lncRNA and between microRNA-mRNA from the microRNA target data bases.

https://doi.org/10.1371/journal.pcbi.1007793.g004
We find that they offer the best average enrichment efficiency in blood vessel development, regulation of cell adhesion, circulatory system process, tissue morphogenesis and pathways in cancer, as shown in Fig 6. The microRNA-associated CeRNA networks with higher enrichment efficiency thus exhibit stronger correlation with LUAD development than other types of CeRNA networks.

**MicroRNAs in CCEN with a significantly influence on patient survival**

We perform the Kaplan-Meier analysis [39] and find that there are six microRNAs that have a significant effect on patient survival for all stages of LUAD, as shown in Fig 7. They are hsa-
mir-9, hsa-mir-21, hsa-mir-31, hsa-mir-148a, hsa-mir-195, and hsa-mir-375. For each stage, there is a specific microRNA that affects the survival rate: hsa-mir-19a (P-value, 0.048), hsa-mir-196b (P-value, 0.058), hsa-mir-194 (P-value, 0.061), and hsa-mir-144 (P-value, 0.012) for stages I-IV, respectively, as shown in Fig 8. To further support for this finding, we also calculate the four microRNAs survival curves of LUAD patients at each stage, and find that hsa-mir-19a, hsa-mir-196b, hsa-mir-194 and hsa-mir-144 have little effect on the survival at other stages except for stages I (S1 Fig), II (S2 Fig), III (S3 Fig) and IV (S4 Fig), respectively.
Conclusion

LUAD has more pronounced genomic variations [40] than other lung cancer subtypes, which are rarely caused by a few genetic changes, rendering necessary articulating alternative methodologies in order to obtain a reasonable understanding of the complicated mechanisms behind the evolution of LUAD. Analyses based on the CeRNA hypothesis provide a promising approach to understanding mutual regulation at the post-transcriptional level [9, 10, 41].

We have reconstructed the CeRNA networks for IncRNAs, microRNAs and mRNAs corresponding to the four stages of LUAD. Support for the validity of the microRNA-associated CeRNA networks is obtained by comparing variations in the fold-change values and expression levels, as well as by enriching mRNAs to cancer-related items in gene ontology and

Fig 7. Survival curves of microRNAs from Kaplan-Meier analysis. Six microRNAs in CCEN that have a significant effect on the survival of LUAD patient survival in all four stages. Log-rank tests are used to analyze the Kaplan-Meier survival curve.

https://doi.org/10.1371/journal.pcbi.1007793.g007
KEGG pathways. Through a systematic enrichment efficiency analysis on mRNAs linked by microRNA in the UCENs and mRNAs selected directly from CeRNA networks (Fig 6), we have identified the key underpinning RNA molecules for LUAD. Especially, by partitioning the network, we find that some specific microRNAs in the CCEN and UCENs have significant influence on the survival rate of LUAD patients (Figs 7 and 8). By performing KEGG pathway enrichment analysis on the mRNA of the CeRNA networks, we find that these mRNAs are not only related to cancer pathways as well as cell cycles and microRNAs in cancer, but there is also a reduction in the degree of the corresponding correlation as the patient state deteriorates from stage I to stage IV of LUAD. This is strong indication that these mRNAs and their CeRNA networks are involved in LUAD and its evolution.

Fig 8. Survival curves of microRNAs from Kaplan-Meier analysis. Four microRNAs in the CeRNA networks, each affecting the survival of a specific stage. Log-rank tests are used to analyze the Kaplan-Meier survival curve.

https://doi.org/10.1371/journal.pcbi.1007793.g008
Through the Kaplan–Meier survival analysis [39], we find six microRNAs that markedly affect the patient survival rate for all four stages of LUAD. The first is hsa-mir-9 reported in recent studies, which is involved in the regulation of NSCLC-related eukaryotic translation initiation factor 5A2, TGFBR2, Lamin and other biomolecules [42–44]. Another microRNA is hsa-mir-21, the first oncogenic microRNA discovered [45], which is associated with pulmonary fibrosis [46], inhibits PTEN, promotes growth and invasion [47], affects cell proliferation, migration, and survival [48, 49]. The third one is hsa-mir-31, which is involved in the inhibition of specific tumor suppressants in human lung cancer [50], affects the survival of LUAD patients [51] and participates in the regulation of lung cancer treatment [52, 53]. The fourth one is hsa-mir-148a, which inhibits the metastasis, proliferation and pathogenesis of NSCLC [54–56]. The fifth one is hsa-mir-195, which participates in the process of affecting lung cancer by regulating MYB, cyclin D3, Ailanthone [57–59]. The sixth one is hsa-mir-375, which affects YAP1 molecule [60] in lung cancer and is involved in the regulation of multiple lung cancer stages [61] as a target of Claudin-1 in NSCLC [62]. It is also involved in mouse alveoli inhibition of Wnt/b-catenin pathway [63]. These previous studies suggest that the six microRNAs that we have successfully identified through our quantitative analysis of the CeRNA networks of LUAD are indeed significant for cancer development, validating and demonstrating the power of our framework of dynamical network analysis.

We have also identified four microRNAs that affect the survival of patients at a specific stage of LUAD. In particular, hsa-mir-19a affects the survival at the first stage, which regulates biological molecules such as grape seed procyanidin, MTUS1, c-Met, and FOXP1 to affect lung cancer [64–67]. MicroRNA hsa-mir-196b influences the survival of patients at the second stage of LUAD, which regulates targets Homeobox A9 and Runx2, etc. [68, 69] and acts on early stage lung cancer [70]. MicroRNA hsa-mir-194 has an effect on the survival of patients at the third stage of LUAD, which can serve as a target for human nuclear distribution protein C [71] and constitute a negative feedback loop with Cullin 4B in carcinogenesis [72]. Finally, hsa-mir-144 impacts the patient survival at the fourth stage of LUAD, which regulates EZH2, TIGAR, Lico A, etc. that affect lung cancer [73–75].

The emerging field of network physiology and medicine [23–27] focuses on the associations between network structures and physiologic states, which relies heavily on measurements from biomedical experiments on large spatial and temporal scales. In this sense, our findings demonstrate an association between CeRNA network and physiologic dysfunction of the lung organ.

While CeRNA network analysis can play a pivotal role in the study of post-transcriptional gene regulation in cancer and disease treatment, there are limitations. For example, for the analysis to be valid, the relative concentrations of CeRNAs in the cellular microenvironment should not be too heterogeneous. There can also be defects and errors in the existing microRNA targets data bases.

Another issue concerns the use of possible surrogate test, which is useful for assessing the effects of different data sources on the results. However, our work is to analyze the CeRNA network constructed from genomics data collected from TCGA and clinical data from a large number of LUAD samples in a step-by-step manner. If we use surrogate test to obtain paired signals from different LUAD samples to construct the CeRNA network, there will be many possible CeRNA networks, making it extremely difficult (if not impossible) to determine the networks for identifying biomarkers. Another aspect of our work is a focus on the variations among the progressive stages of lung cancer, and the number of sequential CeRNA networks constructed for different LUAD stages is limited. Our approach is then to use the gene expression, the FC values of the CeRNA network nodes, and gene enrichment analysis to validate the network. For surrogate networks, the meanings of these tests are not clear. We have
demonstrated that our approach does lead to a number of microRNA biomarkers, and we have carried out indirect validation to test the effectiveness of the targets, which includes the Kaplan-Meier survival analysis based on combining the time event correspondence to the clinical physiological data of the sample patient and a priori knowledge or expert system verification.

Taken together, based on the CeRNA hypothesis, we have developed a framework to reconstruct the CeRNA networks for the four evolutionary stages of LUAD. Our analysis has yielded some prognostic markers closely related to the survival of LUAD patients. The comparative study of the CeRNA networks for the four stages of LUAD provides new insights into understanding cancer mechanisms and identifying targets for better drugs.

Materials and methods

Clinical and gene expression data of LUAD

A total of 877 samples with the corresponding lncRNA, microRNA and mRNA expression profile data from the cancer genome atlas (TCGA) were used. Especially, the numbers of lncRNA, microRNA and mRNA profiles are 29095, 1881 and 25527, respectively. Our analysis requires pre-processing to match the RNA expression data with the patient clinical data, which has excluded 269 samples. We follow the Tumor-Node-Metastasis (TNM) staging criteria for malignant tumors to classify LUAD into four stages. In particular, TNM is a standardized classification system established by the International Association for the Study of Lung Cancer to describe the development of lung cancer in terms of size and spread, where "T" describes the size of the tumor and any spread of the cancer into nearby tissues, "N" denotes the spread of the cancer to nearby lymph nodes, and "M" stands for metastasis, i.e., the spread of the cancer to other parts of the body [76]. For our datasets, we have that, of the remaining 608 samples, 288, 124, 85, and 26 are labeled as stages I, II, III, and IV, respectively, whereas 85 are normal. (More details about the classification criteria of LUAD can be found in S1 Appendix). We perform the match pre-processing on the gene expression profiles as well, resulting in 14460, 1881 and 24991 lncRNA, microRNA and mRNA profiles, respectively. We integrate and further match the clinical physiological data of the samples with either the lncRNA and mRNA data or the microRNA data. For match with the lncRNA and mRNA data, we have a total of 575 cases, where 26, 284, 122, 84, and 59 samples belong to stages I, II, III, IV, and normal, respectively. For match with the microRNA data, there are 24, 279, 122, 85, and 46 samples corresponding to stages I-IV LUAD and normal cases, respectively. (More details are listed in S2 Table).

Framework of analysis

Our articulated framework of dynamical network analysis combines the following methods of analysis: differential expression analysis, match with microRNA-mRNA and microRNA-lncRNA interaction data, identification of negative correlation between microRNA-mRNA and microRNA-lncRNA expressions, CeRNA network partition, gene ontology and KEGG enrichment analysis, and survival analysis. A flow chart of these methods is illustrated in Fig 9. In the following, each of the methods is described.

Differential expression analysis

We use the R-language "Limma" package [77] to analyze the differentially expressed lncRNA, microRNA and mRNA profile data [78] in the four stages of LUAD, with the threshold of log2 FC (log2 fold change) absolute value for filtering the three types RNAs set as one, while ensuring that their P-value (t-statistic [79], see S2 Appendix for detail) is less than 0.05. Fold changes
greater than one correspond to an up-regulated gene, while those less than minus one are associated with a down-regulated gene. We then obtain the numbers of lncRNAs, microRNAs and mRNAs and the up or down regulated expression genes in the four stages of LUAD, as listed in S3 Table.

The individual RNA behavior can be assessed from the volcano map of RNA expression. Fold Change characterizes the relative expression level of interested samples to that of the control samples. For log2 Diese Change = 1, the individual RNA expression level equals the group average of RNAs. We have provided the volcano maps for the expressions of lncRNA (S5 Fig), microRNA (S6 Fig) and mRNA (S7 Fig) at the four LUAD stages.

**Data of microRNA-mRNA and microRNA-lncRNA interactions**

The base of complementary pairing matching relationship data in the RNA sequences for microRNA-mRNA and microRNA-lncRNA are downloaded from the websites of TarBase (version 6.0) [80], miRTarBase (version 6.1) [81], and miRecords (version 4) [82]. The differentially expressed microRNAs are matched with the microRNA-mRNA and microRNA-lncRNA interaction relationships, following which the lncRNAs and mRNAs targeted by these
microRNAs are selected for matching the differentially expressed lncRNAs and mRNAs. A detailed workflow illustrating the analysis of microRNA-mRNA and microRNA-lncRNA interactions is shown in Fig 10, with more details in S4 Table.

**Negative correlation between microRNA-mRNA and microRNA-lncRNA expressions**

Based on the CeRNA hypothesis [17, 18], we pick out only those microRNA-mRNA and microRNA-lncRNA interaction pairs with negative correlation calculated from the Pearson correlation of the RNA expression data, where the threshold in the correlation coefficient is set to be -0.1. (See S5 Table for more details).

**Partition of CeRNA network**

We select the identical microRNAs among the LUAD four stages as well as the mRNAs and lncRNAs connected by these microRNAs in the CeRNA network as constituting the CCEN, where the numbers of identical lncRNAs, microRNAs and mRNAs emerging in the LUAD CeRNA networks of all four stages are 29, 53 and 283, respectively.

Correspondingly, we compare the one-of-a-kind microRNAs of a specified LUAD stage with those in other stages as well as the mRNAs and lncRNAs connected by those microRNAs in the CeRNA networks, which constitute the UCEN. The numbers of (lncRNAs, microRNAs, mRNAs) in the UCEN networks associated with stages I-IV are (27, 5, 197), (30, 2, 236), (41, 8, 312), and (36, 4, 335), respectively.
Gene ontology and KEGG enrichment analysis

We perform gene ontology and KEGG enrichment analysis for mRNAs of the LUAD CeRNA networks, taking into account the various biological processes associated with gene ontology.

Gene enrichment analysis is a widely used approach to identifying biological connections. In Figs 3 and 5, we implement the hypergeometric model to assess whether the number of selected genes in the CeRNA networks associated with lung cancer is larger than that which can be expected purely by chance. In particular, the P-value determines whether any term annotates a specified list of genes at frequency greater than that which can be expected by chance, as determined by the hypergeometric distribution:

\[
p = 1 - \sum_{i=0}^{k-1} \frac{\binom{N-M}{n-i} \binom{M}{i}}{\binom{N}{n}},
\]

where \(N\) is the total number of genes in the background distribution, \(M\) is the number of genes within that distribution that are annotated (either directly or indirectly) to the node of interest, \(n\) is the size of the list of genes of interest, and \(k\) is the number of genes within that list which are annotated to the node. The background distribution by default is all the genes that have an annotation. Then, the outcome of a statistical hypothesis test for the hypergeometric distribution in the gene enrichment analysis, divided by the total number \(N\) of genes in the analysis, defines the enrichment efficiency \(\eta\):

\[
\eta = -\log_{10} P/N.
\]

Survival analysis

We use the Kaplan-Meier method [39], a non-parametric method, to estimate the survival probability from the observed survival data. The survival probability as a function of time is calculated according to

\[
S(t_i) = S(t_{i-1})(1 - d_i/n_i),
\]

where \(n_i\) is the number of patients who were alive before time \(t_i\) and \(d_i\) is the number of death events at \(t_i\). The estimated probability is a step function that changes value only at the time of each event.

Log-rank tests are used to carry out a univariate analysis of the Kaplan-Meier survival curve, which belong to the chi-square test, where all time points of the sample survival information are equal (i.e., with weight setting to one).

Computational packages

We carry out data processing and statistical analysis using R language (v.3.5.1), analyze differentially expressed RNAs by using the “Limma” package [77], and plot the heatmap of the enrichment analysis and the Kaplan-Meier survival curves using the “heatmap,” “survival,” and “survminer” packages. The various CeRNA networks are visualized via Cytoscape (v3.6.1). Finally, we perform the gene ontology and KEGG functional enrichment analysis using the online tool Metascape (http://metascape.org).

Supporting information

S1 Appendix. TNM stages.
(PDF)
S2 Appendix. Statistical analysis and P-value. We have used four methods of statistical analysis involving hypothesis testing to calculate the P-values.
(PDF)

S1 Fig. Survival curves of hsa-mir-19a. The P-values of hsa-mir-19a in the Kaplan-Meier survival curves are 0.24, 0.95, 0.13 for stages II, III and IV, respectively. Log-rank tests are used to analysis of Kaplan-Meier survival curve.
(TIF)

S2 Fig. Survival curves of hsa-mir-196b. The P-values of hsa-mir-196b in the Kaplan-Meier survival curves are 0.58, 0.7, 0.99 for stages I, III and IV, respectively.
(TIF)

S3 Fig. Survival curves of hsa-mir-194. The P-values of hsa-mir-194 in the Kaplan-Meier survival curves are 0.87, 0.92, 0.6 for stages I, II and IV, respectively.
(TIF)

S4 Fig. Survival curves of hsa-mir-144. The P-values of hsa-mir-144 in the Kaplan-Meier survival curves are 0.9, 0.35, 0.58 for stages I, II and III, respectively.
(TIF)

S5 Fig. Individual lncRNA expressions. Volcano map of the lncRNA expression level of four stages of LUAD samples. The x-axis is log₂ Fold–Change (FC value), and the y-axis is the P-values from the differential expression analysis.
(TIF)

S6 Fig. Individual microRNA expressions. Volcano map of the microRNA expression level of four stages of LUAD samples.
(TIF)

S7 Fig. Individual mRNA expressions. Volcano map of the mRNA expression level of four stages of LUAD samples.
(TIF)

S1 Table. The number of nodes and edges of CeRNA networks. Detailed parameter values of the reconstructed lncRNA-microRNA-mRNA CeRNA networks.
(PDF)

S2 Table. Distribution of clinical samples in gene expression data of LUAD. Clinical physiological data of the four LUAD stages matched with the lncRNA and mRNA expression data.
(PDF)

S3 Table. The number of differential expression (DE) gene. Differentially expressed lncRNA, microRNA and mRNA profile data in the four stages of LUAD.
(PDF)

S4 Table. The number of lncRNAs or mRNAs targeted by microRNAs and their relationships. The microRNA-mRNA and microRNA-lncRNA interactions relationships obtained by using the base of complementary pairing matching relationship data in the RNA sequences.
(PDF)

S5 Table. The number of lncRNAs or mRNAs selected by introducing negative correlation. The Pearson correlation of the RNA expression data between microRNA-mRNA and microRNA-lncRNA expressions is calculated.
(PDF)
S1 Data. Figure data. (RAR)

Acknowledgments
We thank Prof. Celso Grebogi and Prof. Lei Yang for helpful discussions.

Author Contributions
Conceptualization: Zhi-Tong Bing, Bing-Hui Guo, Liang Huang, Zi-Gang Huang, Ying-Cheng Lai.

Data curation: Chao Kong, Zhi-Tong Bing.

Formal analysis: Chao Kong, Zhi-Tong Bing, Bing-Hui Guo, Zi-Gang Huang.

Funding acquisition: Zhi-Tong Bing, Bing-Hui Guo, Zi-Gang Huang.

Investigation: Chao Kong, Yu-Xiang Yao, Zhi-Tong Bing.

Methodology: Yu-Xiang Yao, Zhi-Tong Bing, Zi-Gang Huang.

Project administration: Zhi-Tong Bing, Bing-Hui Guo, Zi-Gang Huang.

Resources: Zhi-Tong Bing, Bing-Hui Guo, Liang Huang, Zi-Gang Huang.

Validation: Chao Kong, Zi-Gang Huang, Ying-Cheng Lai.

Visualization: Chao Kong.

Writing – original draft: Chao Kong.

Writing – review & editing: Yu-Xiang Yao, Liang Huang, Zi-Gang Huang, Ying-Cheng Lai.

References


Supplementary Information for

Dynamical network analysis reveals key microRNAs in progressive stages of lung cancer

Chao Kong, Yu-Xiang Yao, Zhi-Tong Bing, Bing-Hui Guo, Liang Huang, Zi-Gang Huang, and Ying-Cheng Lai

Corresponding author: Zi-Gang Huang

Contents

Supplementary Notes

Note 1: Survival curves ............................................................... 1
Note 2: TNM stages ................................................................. 1
Note 3: Statistical analysis and P-value ......................................... 2
Note 4: Individual RNA expressions in the volcano map ................. 2

Supplementary Figures 3

Supplementary Tables 7
Supplementary Notes

**Note 1: Survival curves**

Corroborating the results in Fig. 8 in the main text, we calculate the microRNA-based survival curves of LUAD patients at each stage. As described in the main text, microRNAs hsa-mir-19a, hsa-mir-196b, hsa-mir-194 and hsa-mir-144 affect the survival of LUAD patients in stages I, II, III and IV, respectively.

Further support for this finding is presented in Supplementary Figs. 1-4, which show that the four microRNAs have little effect on the survival of LUAD patients at other stages except for stages I, II, III and IV, respectively. More specifically, the P-values of hsa-mir-19a in the Kaplan-Meier survival curves are 0.24, 0.95, 0.13 for stages II, III and IV, respectively. For hsa-mir-196b, the P-values are 0.58, 0.7, 0.99 for stages I, III and IV, respectively. For hsa-mir-194, the P-values are 0.87, 0.92, 0.6 for stages I, II and IV, respectively. For hsa-mir-144, the P-values are 0.9, 0.35, 0.58 for stages I, II and III, respectively.

**Note 2: TNM stages**

Lung Adenocarcinoma (LUAD) is a kind of lung cancer that is a malignant tumor. The International Association for the Study of Lung Cancer established the Lung Cancer Staging Project in 1998 and updated the Tumor-Node-Metastasis (TNM) classification system of malignant tumors, which forms the basis for classifying lung cancer into distinct stages. The current (seventh) edition classifies all histotypes of lung cancer after evaluation of the outcomes from an extensive worldwide database of patients.

TNM characterizes the size and location of the tumor, the location of cancer in the lymph nodes and to where the cancer has spread (Metastases). In particular, the letter “T” stands for primary tumor:

- **“T1a”** – tumor size ≤ 2cm;
- **“T1b”** – tumor size > 2 ~ 3cm;
- **“T2a”** – tumor size > 3 ~ 5cm;
- **“T2b”** – tumor size > 5 ~ 7cm;
- **“T3”** – tumor size > 7cm and/or multiple tumor nodules in the same lobe;
- **“T4”** – multiple tumor nodules (of any size) in the same lung but in a different lobe.

The letter “N” denotes the spread of the cancer to nearby lymph nodes:

- **“N0”** – no regional lymph node metastasis;
- **“N1”** – metastasis in ipsilateral peribronchial and/or ipsilateral hilar lymph nodes as well as intrapulmonary nodes including involvement by direct extension;
- **“N2”** – metastasis in ipsilateral mediastinal and/or subcarinal lymph node(s);
- **“N3”** – metastasis in contralateral mediastinal, contralateral hilar, ipsilateral or contralateral scalene, or supraclavicular lymph node(s).
The letter “M” describes distant metastasis:

“**M1a**” – malignant pleural or pericardial effusions and/or separate tumor nodules in the contralateral lung;

“**M2b**” – distant metastasis in extrathoracic organs.

Based on the seventh edition of the TNM classification system described above, we list the stage groupings for non-small-cell lung cancer in Supplementary Table 6. In order to have sufficient data amount for statistical analysis of each LUAD stage, in our work we combine stage Ia and Ib and label them as stage I, and do the same for Stages II, III, and IV.

### Note 3: Statistical analysis and P-value

We have used four methods of statistical analysis involving hypothesis testing to calculate the P-values. The four methods are as follows.

- In our construction of the CeRNA network, RNAs are processed through a differentially expressed analysis, where the basic statistic used for significance assessment is the moderated t-statistic [1], which is computed for each contrast. This has the same interpretation as an ordinary t-statistic except that the standard errors have been moderated across genes, i.e., squeezed towards a common value, using a simple Bayesian model. This has the effect of borrowing information from the ensemble of genes to enhance inference about each individual gene. Moderated t-statistics lead to P-values in the same way that ordinary t-statistics do except that the degrees of freedom are increased, reflecting the greater reliability associated with the smoothed standard errors.

- The two-sample Wilcoxon test (also known as “Mann-Whitney” test) [2, 3] is applied to vectors of the RNA expression data, with the following notations for statistical significance: ns – p > 0.05; * – p ≤ 0.05; ** – p ≤ 0.01; *** – p ≤ 0.001; **** – p ≤ 0.0001.

- Gene enrichment analysis as described in MATERIALS AND METHODS and implemented in Figs. 3 and 5 in the main text.

- Log-rank tests as described at the end of subsection entitled “Framework of analysis” in MATERIALS AND METHODS in the main text.

### Note 4: Individual RNA expressions in the volcano map

Fold Change characterizes the relative expression level of interested samples to that of the control samples. For log<sub>2</sub> Fold-Change = 1, the individual RNA expression level equals the group average of RNAs. The individual RNA behavior can be assessed from the volcano map of RNA expression. The volcano maps for the expressions of lncRNA, microRNA and mRNA at the four LUAD stages are shown in Supplementary Figs. 5-7.
Supplementary Figures

Supplementary Figure 1. Survival curves of microRNA hsa-mir-19a in stages II, III and IV. Log-rank tests are used to analysis of Kaplan-Meier survival curve.
Supplementary Figure 2. Survival curves of microRNA hsa-mir-196b in stages I, III and IV. Log-rank tests are used to analysis of Kaplan-Meier survival curve. Log-rank tests are used to analysis of Kaplan-Meier survival curve.

Supplementary Figure 3. Survival curves of microRNA hsa-mir-194 in stages I, II and IV. Log-rank tests are used to analysis of Kaplan-Meier survival curve.
**Supplementary Figure 4.** Survival curves of microRNA hsa-mir-144 in stages I, II and III. Log-rank tests are used to analysis of Kaplan-Meier survival curve.

**Supplementary Figure 5.** Volcano map of the lncRNA expression level of four stages of LUAD samples. The $x$-axis is $\log_2$ Fold-Change (FC value), and the $y$-axis is the P-values from the differential expression analysis.
**Supplementary Figure 6.** Volcano map of the microRNA expression level of four stages of LUAD samples. The $x$-axis is $\log_2$ Fold-Change (FC value), and the $y$-axis is the P-values from the differential expression analysis.

**Supplementary Figure 7.** Volcano map of the mRNA expression level of four stages of LUAD samples. The $x$-axis is $\log_2$ Fold-Change (FC value), and the $y$-axis is the P-values from the differential expression analysis.
Supplementary Tables

**Supplementary Table 1.** Detailed parameter values of the reconstructed lncRNA-microRNA-mRNA CeRNA networks. The rows denoted as stages I, II, III, and IV correspond to the networks at the four stages of LUAD. The columns “total,” “dif,” and “same” represent the total number of nodes, the number of unique nodes, and the number of common nodes, respectively, for the networks at the four stages.

<table>
<thead>
<tr>
<th></th>
<th>IncRNA</th>
<th>MicroRNA</th>
<th>mRNA</th>
<th>Total</th>
<th>Dif</th>
<th>Same</th>
<th>Total</th>
<th>Dif</th>
<th>Same</th>
<th>All Nodes</th>
<th>All Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I</td>
<td>56</td>
<td>27</td>
<td>58</td>
<td>2</td>
<td>480</td>
<td>197</td>
<td>594</td>
<td>1633</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage II</td>
<td>59</td>
<td>30</td>
<td>55</td>
<td>2</td>
<td>519</td>
<td>236</td>
<td>633</td>
<td>1603</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage III</td>
<td>70</td>
<td>41</td>
<td>61</td>
<td>8</td>
<td>595</td>
<td>312</td>
<td>726</td>
<td>1952</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stage IV</td>
<td>65</td>
<td>36</td>
<td>57</td>
<td>4</td>
<td>618</td>
<td>335</td>
<td>740</td>
<td>1945</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Supplementary Table 2.** Clinical physiological data of the four LUAD stages matched with the lncRNA and mRNA expression data. Listed are the numbers of four LUAD samples and normal samples. The rows denoted as stages I, II, III, and IV represent the numbers of samples in the networks in the four stages of LUAD. The row “control” gives the number of normal samples.

<table>
<thead>
<tr>
<th></th>
<th>mRNA and lncRNA Data</th>
<th>microRNA Data</th>
<th>All Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>59</td>
<td>46</td>
<td>85</td>
</tr>
<tr>
<td>Stage I</td>
<td>284</td>
<td>279</td>
<td>288</td>
</tr>
<tr>
<td>Stage II</td>
<td>122</td>
<td>122</td>
<td>124</td>
</tr>
<tr>
<td>Stage III</td>
<td>84</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>Stage IV</td>
<td>26</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Total</td>
<td>575</td>
<td>556</td>
<td>608</td>
</tr>
</tbody>
</table>
**Supplementary Table 3.** Differentially expressed lncRNA, microRNA and mRNA profile data in the four stages of LUAD. The threshold of the absolute value of log$_2$ FC (log$_2$ fold change) for filtering the three types of RNAs is set to be one, ensuring that their P-values are less than 0.05 in the differential expression analysis. Fold changes greater than one correspond to up-regulated genes, while those less than minus one are associated with down-regulated genes. The rows Stages I, II, III, and IV represent the networks in the four stages of LUAD.

<table>
<thead>
<tr>
<th>Stages</th>
<th>DE lncRNAs</th>
<th>DE microRNAs</th>
<th>DE mRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>up</td>
<td>down</td>
</tr>
<tr>
<td>Stage I</td>
<td>104</td>
<td>56</td>
<td>48</td>
</tr>
<tr>
<td>Stage II</td>
<td>118</td>
<td>75</td>
<td>43</td>
</tr>
<tr>
<td>Stage III</td>
<td>131</td>
<td>80</td>
<td>51</td>
</tr>
<tr>
<td>Stage IV</td>
<td>119</td>
<td>69</td>
<td>50</td>
</tr>
</tbody>
</table>

**Supplementary Table 4.** The microRNA-mRNA and microRNA-lncRNA interactions relationships obtained by using the base of complementary pairing matching relationship data in the RNA sequences. The columns “total,” “up” and “down” represent the numbers of all, up-regulated and down-regulated RNAs in the four stages of LUAD. The columns miRNA-lncRNA and miRNA-mRNA list the numbers of the edges in the microRNA-lncRNA and microRNA-mRNA subnetworks, respectively, in the networks of four stages of LUAD.

<table>
<thead>
<tr>
<th>Stages</th>
<th>DE lncRNA</th>
<th>DE microRNAs</th>
<th>DE mRNAs</th>
<th>miRNA-lncRNA</th>
<th>miRNA-mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>total</td>
<td>up</td>
<td>down</td>
<td>total</td>
<td>up</td>
</tr>
<tr>
<td>Stage I</td>
<td>77</td>
<td>44</td>
<td>33</td>
<td>58</td>
<td>12</td>
</tr>
<tr>
<td>Stage II</td>
<td>89</td>
<td>59</td>
<td>30</td>
<td>57</td>
<td>13</td>
</tr>
<tr>
<td>Stage III</td>
<td>99</td>
<td>61</td>
<td>38</td>
<td>61</td>
<td>13</td>
</tr>
<tr>
<td>Stage IV</td>
<td>93</td>
<td>53</td>
<td>40</td>
<td>58</td>
<td>12</td>
</tr>
</tbody>
</table>
Supplementary Table 5. The Pearson correlation of the RNA expression data between microRNA-mRNA and microRNA-lncRNA expressions is calculated. The threshold of the correlation coefficient is set to be -0.1. The columns of miRNA-lncRNA and miRNA-mRNA list the numbers of edges for the microRNA-lncRNA and microRNA-mRNA subnetworks, respectively, in the networks of four stages of LUAD. The rows stage I, II, III, and IV correspond to the networks of the four stages of LUAD.

<table>
<thead>
<tr>
<th></th>
<th>DE IncRNA</th>
<th>DE miRNA</th>
<th>DE mRNA</th>
<th>miRNA-lncRNA</th>
<th>miRNA-mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I</td>
<td>56</td>
<td>58</td>
<td>480</td>
<td>202</td>
<td>1431</td>
</tr>
<tr>
<td>Stage II</td>
<td>59</td>
<td>55</td>
<td>519</td>
<td>206</td>
<td>1397</td>
</tr>
<tr>
<td>Stage III</td>
<td>70</td>
<td>61</td>
<td>595</td>
<td>276</td>
<td>1676</td>
</tr>
<tr>
<td>Stage IV</td>
<td>65</td>
<td>57</td>
<td>618</td>
<td>292</td>
<td>1653</td>
</tr>
</tbody>
</table>

Supplementary Table 6. Stage groupings for non-small-cell lung cancer

<table>
<thead>
<tr>
<th>Tumour (T)</th>
<th>Node (N)</th>
<th>Metastasis (M)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ia</td>
<td>T1a or T1b</td>
<td>N0</td>
</tr>
<tr>
<td>Ib</td>
<td>T2a</td>
<td>N0</td>
</tr>
<tr>
<td></td>
<td>T1a or T1b</td>
<td>N1</td>
</tr>
<tr>
<td>IIa</td>
<td>T2a</td>
<td>N1</td>
</tr>
<tr>
<td></td>
<td>T2b</td>
<td>N0</td>
</tr>
<tr>
<td>IIb</td>
<td>T2b</td>
<td>N1</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>N0</td>
</tr>
<tr>
<td>IIIa</td>
<td>T1 or T2</td>
<td>N2</td>
</tr>
<tr>
<td></td>
<td>T3</td>
<td>N1 or N2</td>
</tr>
<tr>
<td></td>
<td>T4</td>
<td>N0 or N1</td>
</tr>
<tr>
<td>IIIb</td>
<td>T4</td>
<td>N2</td>
</tr>
<tr>
<td></td>
<td>any T</td>
<td>N3</td>
</tr>
<tr>
<td>IV</td>
<td>any T</td>
<td>any N</td>
</tr>
<tr>
<td></td>
<td>any T</td>
<td>any N</td>
</tr>
</tbody>
</table>
References

