参赛密码_____

（由组委会填写）

2019 年（第六届）全国大学生统计建模大赛

学校	哈尔滨医科大学

参赛队号		
队员姓名	1.	郭胜男
	2.	徐泽龙
	3.	刘格睿

基于染色体不稳定性和失调 ceRNA 构建

肺腺癌预后标志物识别模型

目录

表格和插图清单

摘要

染色体不稳定在癌症的发展和转移机制中具有一定的推动作用，同时内源性竞争 RNA（ceRNA）互作对关系的失调在肺腺癌中也发挥着重要作用。本研究结合样本染色体不稳定性的特性在 ceRNA 层面对肺腺癌进行探究，旨在构建寻找肺腺癌预后相关标志物的统计模型。

使用构建的模型对肺腺癌染色体不稳定性进行分析，结果表明染色体不稳定性在癌症样本中显著高于癌旁样本，区分两类样本有很高的分类效能（AUC=0.989）。同时染色体不稳定性高的样本倾向有高临床分期和差的预后。

构建染色体不稳定性相关的双权重失调 ceRNA（mRNA-lncRNA）网络，失调网络中边的权重大小代表 ceRNA 的失调程度，点的权重大小代表 RNA 与预后的相关性，在数量上大部分 ceRNA 的失调为 loss，失调程度显著高于 gain 的类型。采用贪婪搜索算法在网络中挖掘双权重较大的模块。对挖掘到的模块进行筛选和整合，得到了由 20 个基因组成的标志物。

在术后无药物治疗或放疗样本和所有样本中，20 个 RNA 高风险降低了患者的预后。在接受药物治疗或化疗的样本中，20 个 RNA 仍然能够较好地预测患者预后，说明其预测效能受治疗影响较小并且可能对患者接受药物治疗或放疗后的预后有一定的影响。相关性检验结果表明，20 个 RNA 风险得分与染色体不稳定性和临床分期呈现显著正相关关系。构建肺腺癌疾病特异生存相关的逐步多因素 cox 风险回归模型，这 20 个 RNA 仍然能够作为肺腺癌患者疾病特异生存显著的独立风险因素，证明预测预后具有较好的独立性和鲁棒性。这 20 个 RNA 显著的富集到染色体不稳定性和癌症相关的功能。以上结果表明，本次研究构建的模型可以准确地识别肺腺癌相关预后标志物。

关键词：肺腺癌；染色体不稳定；失调 ceRNA 网络；临床预后；预后标志物

Abstract

Lung cancer is one of the leading causes of death worldwide among types of cancers with lung adenocarcinoma accounted for a large proportion. Chromosome instability and dys-regulation of ceRNA play a significant role in lung adenocarcinoma. In this research, we integrated characteristic of chromosome instability and dys-regulation of ceRNA with the purpose of a model to identify a valuable prognostic marker in lung adenocarcinoma.

We analyzed the characteristics of chromosome instability in lung adenocarcinoma by our model. The result showed that chromosome instability in tumor samples was significantly higher than those of normal tissue adjacent with the AUC reach up to 0.989, and patients with high chromosome instability risk score tended to have high stage and poor clinical prognosis.

Double weighted chromosome instability related dys-regulated ceRNA network composed by mRNA-lncRNA interactions was constructed in which edge weight and node weight represent the extent of dys-regulation of ceRNA and the correlation of RNA and prognosis, respectively. The overwhelming majority of disorder relationship of ceRNA pairs was loss compared to the type of gain. Greedy search algorithm was used to mining modules which have relatively large weight of edge and node. After screening and integration of the modules, the 20-gene signature was obtained.

The 20 RNA high risk reduced the survival of lung adenocarcinoma patients in all samples and those without postoperative treatment(drug therapy or radiotherapy). In samples received drug therapy or radiotherapy, the 20 RNA still be of value in predicting prognosis, which suggest that the prognostic prediction efficiency of the 20 RNA was stable, affected little by treatment and the 20 RNA may have some effect on the prognosis after drug treatment or radiotherapy. And the result of correlation test showed that the 20 RNA high risk had an outstanding positive correlation with high chromosome instability and poor clinical stage. The 20 RNA Still could be an

independent risk factor, according to the analyzation of disease-specific survival related stepwise multivariate COX regression model for adjusting multicollinearity, which proved the robustness and independence of this signature in predicting prognosis. The 20 RNA significantly enriched with functions related to caner and chromosome instability. The results above suggested that our model can accurately identify lung adenocarcinoma-related prognostic markers.

Keywords:

lung adenocarcinoma; chromosome instability; dys-regulated ceRNA network; clinical outcome; prognostic biomarker

一、问题描述

（一）癌症染色体不稳定性相关研究

大多数人类癌症都存在基因组不稳定性的特征，基因组不稳定性（CIN）被认为能促进其他癌症 hallmarks[37]，其中染色体不稳定性是基因组不稳定性的主要类型，染色体不稳定性在癌前病变和恶性增长中被发现[38]。染色体不稳定性的特点是染色体异常的频率增大，包括染色体整个的或者大段的缺失、获得或结构重排[39]，这些变化会干扰正常的基因结构和功能。染色体不稳定性能够快速积累变异促进癌症发展，并且有助于癌症获得内在的耐药性[40,41]。由于染色体不稳定性与众多类型癌症的发展进程相关，识别和探索染色体不稳定性标志物的研究工作有很多，比如 Rama K.R. Mettu 等研究员识别出由 12 个基因组成的基因组不稳定性的生物标志物能够预测乳腺癌、结直肠肿瘤和卵巢癌的临床预后情况[42]。异常着丝粒和着丝点引起染色体不稳定，通过使染色体错分，导致非整倍体、重排和微核形成。CIN70 标志物被广泛的应用在衡量样本染色体不稳定性程度[45,46]和与其他染色体不稳定性标志物的参照和比较[47]，本篇研究也使用 CIN70 标志物来衡量肺腺癌病人样本染色体不稳定性的程度。

（二） ceRNA 机制

长链非编码 RNA(lncRNA)是其中的一类，通常被定义为长度大于 200 个核苷酸的非编码 RNA[2]，近些年来对 lncRNA 功能的研究表明，lncRNA 能沉默 X 染色体[3–5]、参与基因组印记[6–8]调节等位基因的表达，lncRNA 也表现出细胞特异性表达[9]、定位于亚细胞区室[10]的特点。lncRNA 也与疾病尤其是癌症的发生发展有着很大的相关性[11,12]。

在 lncRNA 发挥的众多功能中，lncRNA 可以参与 ceRNA 的调控机制中。miRNA（微小 RNA）在后转录水平通过诱导沉默复合体降解或抑制目标靶基因的表达量[21]，而 miRNA 的靶基因可以通过竞争共享的 miRNA 来间接调控彼此的表达量，被称为“竞争性内源 RNA”（ceRNA）[22]。ceRNA 串扰基于“miRNA 应答元件”（MRE），可以应用于任何含有 MRE 的 RNA 分子，其中就包括 mRNA、lncRNA、假基因等[22]。基于 ceRNA 假说，lncRNA 可以通过 ceRNA 机制作为 miRNA 海绵与 mRNA 竞争共同的 miRNA 间接调控 mRNA

的表达，从而可能进一步影响 mRNA 编码蛋白质的水平。

由于很多 mRNA 和 lncRNA 含有多个 MRE，很多 miRNA 也靶向多个 mRNA 或 lncRNA，所以 ceRNA 关系互作对能够在复杂网络中起作用。通过 ceRNA 网络图谱的描绘，研究人员在各种癌症间发现 ceRNA 关系对发生了明显的重连，这进一步揭示了每种癌症中保守的和重连的 ceRNA hubs 间存在着激烈的竞争关系，构成了癌症特异性和保守的模块[26]。lncRNA、miRNA、和基因互作被定义为 lncRNA 相关的竞争三元组，在 lncRNA 相关的竞争三元组构成的串扰网络中，疾病相关的三元组表现出与非疾病相关的三元组相比不同的拓扑结构[27]。识别 miRNA 介导的 ceRNA 互作对的方法有很多种[28–32]，其中应用比较广泛的是 ceRNA 互作对需要满足共调控和共表达条件[33,34]，两个 RNA 作为 ceRNA 的前提是首先被显著交叠的 miRNA 所调控，然后在表达水平上满足显著的正相关关系，本篇研究采取了同样的方法识别 ceRNA。

（三） 肺腺癌相关背景

在全世界范围内，肺癌在所有癌症中仍然是发病率和死亡率最高的癌症[1]，肺癌中非小细胞肺癌相比较于小细胞肺癌，癌细胞分裂生长速度和扩散转移速度相对较慢，并且非小细胞肺癌占肺癌的 80%左右，但是约有 75%的非小细胞肺癌患者就医时就已处于中期或晚期，所以 5 年生存率较低。非小细胞肺癌分为肺腺癌、肺鳞癌(鳞状细胞癌)和大细胞癌，肺腺癌大约占到非小细胞肺癌的 50%左右，本研究对肺腺癌进行分析，目的在于识别肺腺癌预后相关的标志物。

二、模型构建流程

（一）模型的数据准备

1.肺腺癌转录组表达谱获取及预处理

肺腺癌表达谱 FPKM（Fragments Per Kilobase Million）和 Count 数据在 TCGA 数据库下载（版本：Data Release9.0， http://cancergenome.nih.gov/），将各样本的表达谱合并，得到行是基因、列是样本的表达谱，一共有 60483 个基因；下载 GENCODE V22[50]人类基因组注释信息（gencode.v22.annotation.gtf）注释 mRNA 和 lncRNA，此版本的参考基因组是 TCGA 数据库处理测序数据注释基因时用到的，所以采用同样的参考基因组。注释 mRNA 转录本时，选择 genebiotype 为“protein_coding”的条目，得到 19184 个 mRNA；注释 lncRNA 转录本时，选择

genebiotype 为“processed_transcript”、“lincRNA”、“3prime_overlapping_ncrna”、“antisense”、“non_coding”、“sense_intronic”和“sense_overlapping”的部分并且长度大于 200 nt 的条目，得到 14820 个 lncRNA（这种注释方法参考了 starBaseV2.0[34] 和 LncAct database[27] 注释 lncRNA 的方法）。

mRNA 和 lncRNA 的 Count 表达谱数据用来做差异分析，识别肺腺癌和癌旁组织间的差异表达基因，还有 CIN 高的病人与 CIN 低的病人两组之间的差异表达基因。

mRNA 和 lncRNA 的 FPKM 表达谱数据用来计算 TPM[51]（Transcripts Per Kilobase Million）值，FPKM 值是对 Count 数据先标准化测序深度再标准化基因长度，而 TPM 是先标准化基因长度后标准化测序深度，最后各个样本中所有基因表达量之和是相等的，TPM 表达谱的基因表达量在样本间是可以直接比较的。用 FPKM 值计算 TPM 的方法为每个基因在每个样本中的表达值除以当前样本的总表达值之和再乘以 10^6。然后分别对得到的 mRNA 和 lncRNA 的 TPM 表达谱进行预处理，表达量在 30%及以上的样本中有缺失的基因进行筛选删除，对剩下的缺失值补极小值 0.05 处理，再对 TPM 表达谱整体进行 log2 转换，预处理后的 TPM 表达谱用来衡量基因表达水平以及识别 ceRNA 互作关系对。

TCGA 肺腺癌样本测序数据的 Count 表达谱被用来识别差异表达的 mRNA 和 lncRNA，其余分析用到的数据是由 FPKM 计算的 TPM 表达值。

2. 肺腺癌临床信息及预后数据的获取和整合

肺腺癌的基本临床信息和预后数据是从 Jianfang Liu 等研究员整合的 TCGA 样本临床信息中直接下载，其中包括 TCGA 样本 barcode、年龄、性别、种族、Stage 分期、OS（总生存）生存状态、OS 生存时间、DSS（疾病特异生存）生存状态、DSS 生存时间、DFI（无病间期）生存状态、DFI 生存时间、PFI（无进展间期）生存状态、PFI 生存时间。在 TCGA 数据库下载原始的 xml 格式的临床数据进行样本合并后提取样本的 T 分期、N 分期、M 分期、术后放疗信息和药物治疗信息，与已有的临床数据根据样本 TCGA barcode 进行合并。

3. 整合实验验证 miRNA 靶向调控信息

miRNA 靶向 mRNA 的信息是从 miRecords[52]（版本：4）、miRTarBase[53]（版本：7.0）和 TarBase[54]（版本：6.0）三个数据库下载，提取人类实验验证的 miRNA 与 mRNA 互作信息合并去重复，得到了包含 2846 个 miRNA 和 18936 个 mRNA 的 388895 条无重复的互作数据。人类 miRNA 和 lncRNA 的靶向互作信

息从 DIANA-LncBase[55]（版本：1.0）、starBase[34]（版本：2.0）、lncRNASNP[56]（版本：2.0）数据库中获取，经过整合去重一共得到了 10318 条无重复的人类 miRNA 和 lncRNA 的靶向信息，其中包括 290 个 miRNA 和 1162 个 lncRNA。

（二）模型构建方法

1. 识别差异表达的 mRNA 和 lncRNA

肺腺癌病人样本和癌旁组织样本与 CIN 高的病人组别和 CIN 低的病人之间显著差异表达的 mRNA 和 lncRNA 用 R 语言 DESeq2[57]包进行识别分析。mRNA 和 lncRNA 的 Count 表达谱作为输入数据，校正后的 P 值小于 0.05，log2 FC 大于 1 的基因为显著高表达基因；校正后的 P 值小于 0.05，log2 FC 小于-1 的基因为显著低表达基因。

2. StepMiner 二分类方法计算 CIN 得分

StepMiner 方法最早是用于研究时间进程的微阵列数据，比如细胞受到外界刺激或给药后基因表达层面的反应状态，它能够找到最符合基因表达模式转换的模型：一个转换（在某个时间点上升或下降）或两个转换（先上升后下降或先下降后上升），并且找到状态转换发生的时间以及转换时的表达值阈值[58]。后来有研究用 StepMiner 方法将一个基因的表达水平在不同样本中分成这个基因的阳性和阴性两组（或表达高和表达低两组），方法是将所有样本按照这个基因的表达水平从低到高排序，模拟一个转换模式在某个点上升的情况，找到基因表达的阈值从而将样本分为阳性和阴性表达两组（或高表达和低表达两组）[59,60]。本次研究应用 StepMiner 方法的目的是找到对数值型向量进行二分类的阈值，将数值向量从低到高排序后用穷举法遍历每一个值，找到合适的阈值将向量分成高低两组时当前的信噪比（signal-to-noise ratio）最高[61]。具体步骤为：

（1）给定随机变量 X：X_1，X_2，X_3…X_n，

将随机变量从低到高排序：$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \ldots \leq X_{(n)}$；

（2）穷举法遍历 $X_{(1)}$，$X_{(2)}$，$X_{(3)}$，…，$X_{(n)}$，找到使信噪比最大的位置 t’，

信噪比计算公式：

$$\text{SNR}=\frac{\text{Signal}}{\text{Noise}}=\frac{\sum_{i=1}^{n}(f(i)-\mu)^{2}}{\sum_{i=1}^{n}(f(i)-x_{(i)})^{2}} \quad (1)$$

$$f(i)=\mu_1 I(i\le t)+\mu_2 I(i>t) \quad (1\le t<n)$$

I 为指示函数，符合括号中的条件时 I=1，不符合括号中的条件时 I=0；

μ ：$X_{(1)}$，$X_{(2)}$，$X_{(3)}$，…，$X_{(n)}$的均值；

μ_1：$X_{(1)}$，$X_{(2)}$，$X_{(3)}$，…，$X_{(t)}$的均值；

μ_2：$X_{(t+1)}$，$X_{(t+2)}$，$X_{(t+3)}$，…，$X_{(n)}$的均值

（3）$\text{Threshold}=\frac{1}{2(x(t')+x(t'+1))}$，若存在两个位置都使信噪比的值最大，取这两个位置的均值作为阈值。

StepMiner 方法能够识别数值向量二分类的阈值从而将数值向量分成两组，其特点是组间差异大，组内差异小。在本研究中 StepMiner 应用于根据样本 CIN 得分将样本分成 CIN 高和 CIN 低的两组样本，以及根据样本的预后风险得分将样本分成高风险组和低风险组。

3. 构建 CIN 相关失调 ceRNA 双权重网络

首先在肺腺癌 CIN 低的样本组中识别 mRNA 和 lncRNA 组成的 ceRNA 互作对。在肺腺癌样本中相对癌旁样本中显著差异表达的 mRNA 和 lncRNA 被视为是与肺腺癌相关的基因，作为 ceRNA 对的识别范围。

给定一组 mRNA 和 lncRNA 时，用累计超几何分布检验两个 RNA 是否被显著共享的 miRNA 靶向调控，fdr 值小于 0.05 的 mRNA 和 lncRNA 符合显著共调控关系，公式为：

$$\text{p}=1-F(x\mid N,K,M)=\frac{1-\sum_{i=0}^{x-1}C(K,i)\times C(N-K,M-i)}{C(N,M)} \quad (2)$$

p：超几何富集检验的显著性 p 值；

x：当前 mRNA 和 lncRNA 被共同的 miRNA 调控的个数；

N：背景，所有人类 miRNA 的个数；

K：调控当前 mRNA 的 miRNA 的个数；

M：调控当前 lncRNA 的 miRNA 的个数；

对符合共调控的 mRNA-lncRNA 对进行相关性检验，皮尔森相关系数大于 0，fdr 值小于 0.05 的 mRNA 和 lncRNA 作为显著共表达互作对。同时满足显著共调控和共表达条件的 mRNA 和 lncRNA，被识别为 ceRNA 互作对。

基于在 CIN 低的肺腺癌样本中识别的 mRNA 和 lncRNA 的 ceRNA 互作对构建失调网络的基本构架，然后计算点和边的权重。

（1） 失调网络边权重计算方法为：

$$\text{edgeweight(e)}=\varphi^{-1}[1-2\times(1-\varphi(|X|))] \tag{3}$$

φ 是标准正态分布函数。

Fisher 转换：

$$\text{F(x)}=\frac{1}{2}\ln\frac{1+x}{1-x} \tag{4}$$

Fisher 检验：

$$\text{X}=\frac{F(r_{CIN\_High})-F(r_{CIN\_Low})}{\sqrt{\frac{1}{n_{CIN\_High}-3}+\frac{1}{n_{CIN\_low}-3}}} \tag{5}$$

定义一个新的统计量 X，X 近似符合标准正态分布

给定一对 mRNA 和 lncRNA 组成的 ceRNA 互作对，n_{CIN_High} 和 n_{CIN_Low} 分别代表 CIN 高的组别和 CIN 低的组别中的样本个数；r_{CIN_High} 和 r_{CIN_Low} 分别代表 ceRNA 互作对在 CIN 高的组别和 CIN 低的组别中的皮尔森相关系数，CIN 高样本中 ceRNA 互作对的相关性检验的 p 值大于 0.05，将它的相关系数赋值为 0，认为这对 mRNA 和 lncRNA 的表达水平不存在相关性。

（2）失调网络点权重的计算方法为：

$$\text{nodeweight(v)} = \varphi^{-1}(1-p) \tag{6}$$

网络中每个基因对疾病特异生存（DSS）的预后价值进行评估，计算单因素 cox 风险回归分析模型的显著性 p 值，为了排除临床上肺腺癌患者接受术后治疗（药物治疗或放疗）对预后的影响，这里选择的是术后未接受治疗的样本。

在肺腺癌患者 CIN 相关的 mRNA 和 lncRNA 组成的具有双权重的失调 ceRNA 网络中，边权值的意义是权值越大，当前这对 ceRNA 在 CIN 高样本中皮尔森相关系数相比于 CIN 低样本的改变量越大；点的权值越大，代表当前 mRNA 或 lncRNA 与预后的相关性越大。

4. 挖掘失调 ceRNA 网络中点和边权重相对大的模块

基于 CIN 相关的失调 ceRNA 网络，边的权值越大代表 ceRNA 在两组样本中失调程度越大，点的权值越大代表当前 mRNA 或 lncRNA 和 DSS 的相关性越大。实现挖掘点和边双权重大的模块，采用 R 包 dmGWAS_3.0 的方法[62]。对于给定的模块，模块打分的计算公式为：

$$\text{S}=\lambda\frac{\sum\text{edgeweight(e)}}{n_e}+(1-\lambda)\frac{\sum\text{nodeweight(v)}}{n_v} \tag{7}$$

n_e 代表给定模块中的边的个数，n_v 代表给定模块中的点的个数，参数 λ 是 0 到 1 之间的数值来衡量模块打分时边和点权重的比重，S 衡量的是 ceRNA 的皮尔森相关系数在两个组中的改变程度和基因与 DSS 的相关性。λ 取默认值，计算方法为：随机在背景网络中取 1000 个子网计算 mr 值，λ 的取值为：

$$\lambda = \frac{1}{(1+\text{median(mr)})} \tag{8}$$

$$\text{mr}=\left|\frac{\sum\text{edgeweight(e)}}{n_e}\Big/\frac{\sum\text{nodeweight(v)}}{n_v}\right| \tag{9}$$

挖掘双权重相对大的模块采用的是贪婪搜索算法，具体步骤为：

1） 初始设置一个由单个基因组成的模块 M，计算模块 M 的模块得分 S_M；

2） 在网络中模块 M 周围所有的第一个邻居中找到点 $N_{\max}$，当点 $N_{\max}$ 加入模块 M 中重新计算模块得分时的增量最大。

3） 如果点 $N_{\max}$ 加入模块后重新计算模块得分的增量大于 $r \times S_M$，将点 $N_{\max}$ 加入到模块中。参数 r 是决定模块得分增量的数值，根据之前研究的推荐 r 取值为 0.1[63]。

4） 在模块周围最短路径 d=1 的范围内进行贪婪搜索，重复前面三个步骤直到没有新的点加入到模块之中。

最后对挖掘到的所有模块得分进行标准化，对于每个给定的模块 M，M 中包含 K 个点，在背景网络中随机抽取 10000 次大小相同的子网，计算子网打分和这些随机子网打分的均值 μ 和标准差 σ，模块标准化得分为：

$$S_N = \frac{(S_M - \mu)}{\sigma} \qquad (10)$$

5. 多基因的样本预后风险分组

给定一组基因：gene(1)，gene(2)，gene(3)...gene(n)，构建多因素 cox 风险回归模型。每一个样本的风险得分为：

$$\sum_{i=1}^{n} \beta_i \times \mathrm{E}_{gene(i)} \qquad (11)$$

n 是基因个数，β_i 是第 i 个基因在多因素 cox 风险回归模型中的回归系数，$\mathrm{E}_{gene(i)}$ 是第 i 个基因在当前样本的表达量。用 StepMiner 方法找到样本预后风险得分阈值，从而将样本分成预后高风险组和低风险两组。

三、模型的效能分析

（一） CIN 标志物基因的差异表达

使用 CIN70 标志物计算了 513 个肺腺癌病人样本的 CIN 风险得分，然后将样本的 CIN 风险得分从低到高依次排列，用 StepMiner 方法计算阈值将病人样本分成两组：CIN 高的组别有 289 个病人样本，CIN 低的组别有 224 个病人样本。

用 DESeq2 R 包识别在肺腺癌与癌旁样本相比显著差异表达的 mRNA 和 lncRNA，以及 CIN 高与 CIN 低样本相比显著差异表达的 mRNA 和 lncRNA（表-1），fdr 小于 0.05，log2 FC>1 的基因为显著高表达基因；fdr 值小于 0.05，log2 FC<-1 的基因为显著低表达基因。肺腺癌样本对比癌旁样本：mRNA 显著差异高、低表达分别为 3400 个，1975 个；lncRNA 显著差异高、低表达分别为 2680 个，1085 个；CIN 高的样本与 CIN 低的样本相比 mRNA 显著差异高、低表达分别为 1224 个，741 个；lncRNA 显著差异高、低表达分别为 1209 个，309 个。CIN70 标志物的基因中有 52 个基因显著差异高表达，0 个基因显著低表达；这 52 个基因中有 42 个基因在 CIN 高的样本比 CIN 低的样本显著差异高表达，0 个显著低表达，结果表明这两种情况都显著差异高表达的 CIN70 标志物中的基因其差异程度普遍较高（图-1）。

以上结果表明，CIN70 标志物中的基因在肺腺癌样本和 CIN 高的样本中总体上趋向于高表达。

表-1 显著差异表达的 mRNA 和 lncRNA

	LUAD vs adjacent	CIN_High vs CIN_Low
mRNA up-regulated	3400	1224
mRNA down-regulated	1975	741
lncRNA up-regulated	2860	1209
mRNA down -regulated	1085	309

注：LUAD vs adjacent: 肺腺癌相对癌旁样本的差异表达 mRNA 和 lncRNA；CIN_High vs CIN_Low: CIN 高相比 CIN 低样本的差异表达 mRNA 和 lncRNA

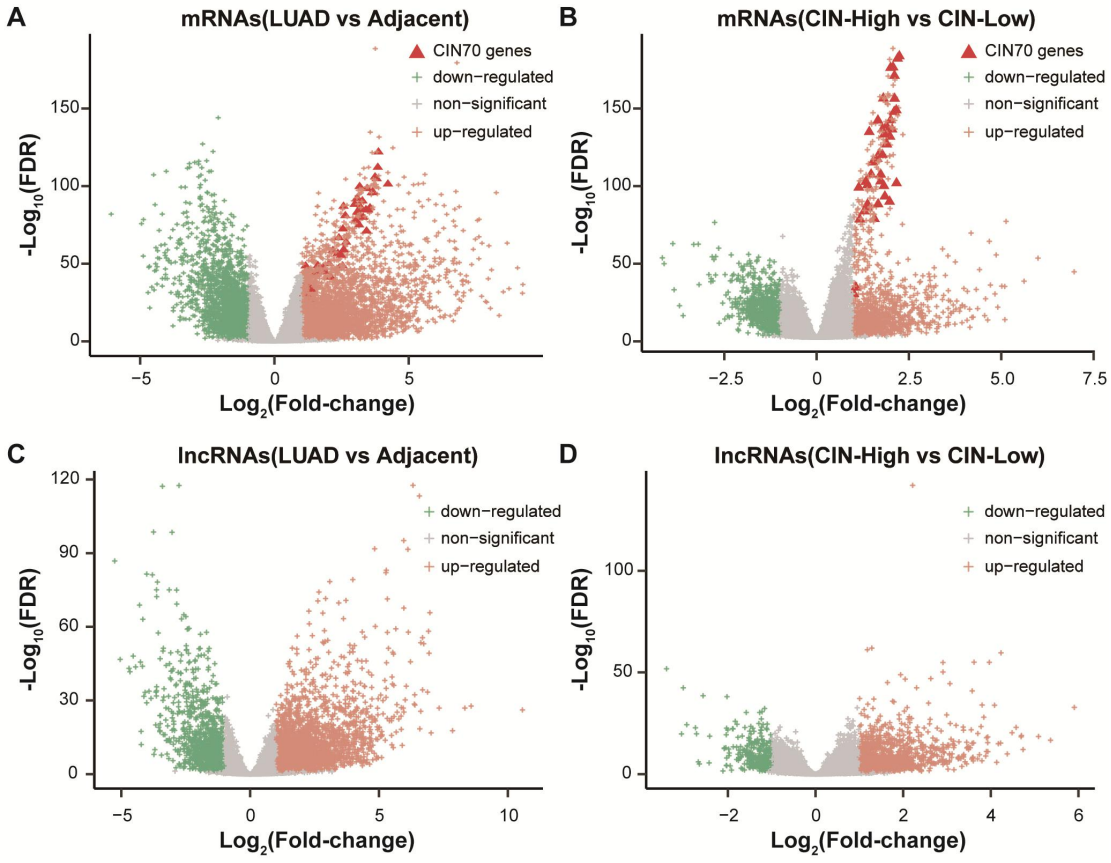

图-1 差异表达 mRNA 和 lncRNA 的火山图

图注：(A)肺腺癌样本相比于癌旁组织差异表达的 mRNA。(B)CIN 高的样本相比于 CIN 低的样本差异表达的 mRNA。(C)肺腺癌样本相比于癌旁组织差异表达的 lncRNA。(D)CIN 高的样本相比于 CIN 低的样本差异表达的 lncRNA。X 轴代表 log2 转换后的 fold-change 值，Y 轴表示-log10 转换后的 fdr 值。绿色加号代表显著差异低表达，橙色加号代表显著差异高表达，灰色代表非显著差异表达，红色三角代表 CIN70 标志物中的基因。

（二）肺腺癌与癌旁 CIN 分类效能

在肺腺癌和癌旁配对的 57 对样本中，肺腺癌 CIN 风险得分结果为 350.44±60.00，显著高于配对的癌旁样本（214.1201±22.98），Wilcoxon 秩和检验 $p=5.3e^{-11}$（图-2 A）。用 CIN 风险得分作为区分肺腺癌与癌旁样本的分类器，ROC 曲线下面积 AUC 达到了 0.989，分类效能非常高（图-2 B）。

肺腺癌相比癌旁样本 CIN 程度更高，此结果与癌症普遍具有基因组不稳定性的特征一致。

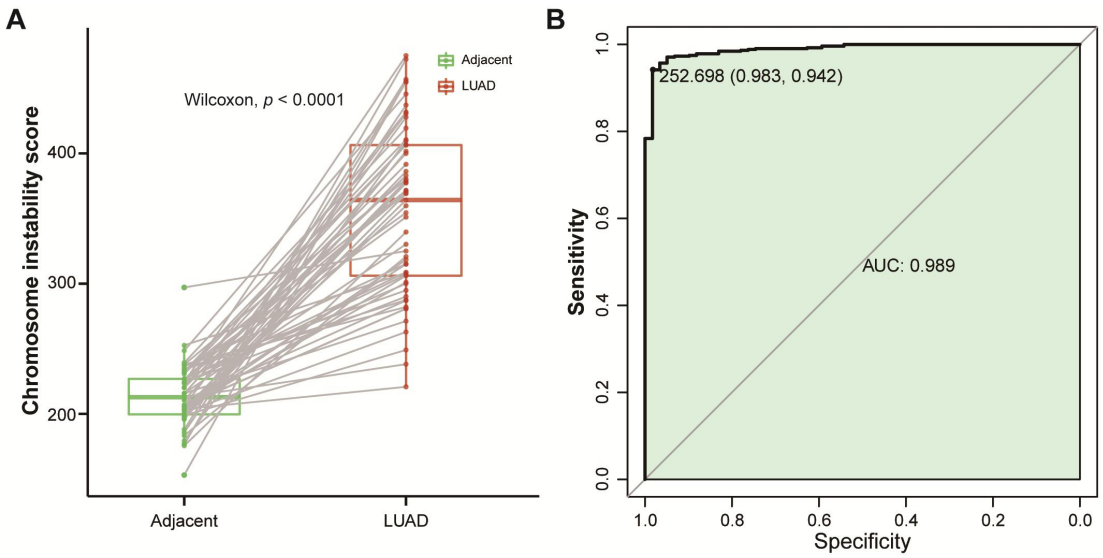

图-2 CIN 在肺腺癌样本和癌旁样本中的分类效能

图注：（A）肺腺癌与癌旁配对样本的染色体不稳定性风险得分的箱线图。绿色柱表示癌旁组织，红色柱表示肺腺癌组织，X 轴表示组织类型，Y 轴表示 CIN 风险得分。（B）CIN 风险得分的 ROC 曲线。X 轴表示特异性，Y 轴表示灵敏度，绿色多边形表示 ROC 曲线下的面积（AUC）。

（三） 不同临床分期 CIN 的差异

比较肺腺癌样本中不同分期样本的 CIN 风险得分，两组数据的检验用 Wilcoxon 秩和检验，三组以上总体检验用 Kruskal-Wallis 检验。在 T 分期中，T_2 分期样本的 CIN 风险得分显著高于 T_1 分期的样本，总体四个分期检验同样是显著的（Kruskal-Wallis 检验，p=0.00033）（图-3 A）。在 N 分期中，N_1 分期 CIN 风险得分显著高于 N_0 分期，N_2 分期 CIN 风险得分显著高于 N_0 分期，总体四个分期检验同样是显著的（Kruskal-Wallis 检验，p=0.0098）（图-3 B）。在 M 分期中，M_1 分期 CIN 风险得分显著高于 M_0 分期（Wilcoxon 秩和检验，p=0.044）（图-3 C）。在总的临床分期中，StageII 分期 CIN 风险得分显著高于 StageI 分期，StageIII 分期 CIN 风险得分显著高于 StageI 分期，StageIV 分期 CIN 风险得分显著高于 StageI 分期，总体四个分期检验都是显著的（Kruskal-Wallis 检验，p=0.00073）（图-3 D）。

总之，染色体不稳定高与差的临床 TNM 分期和总分期是相关的。

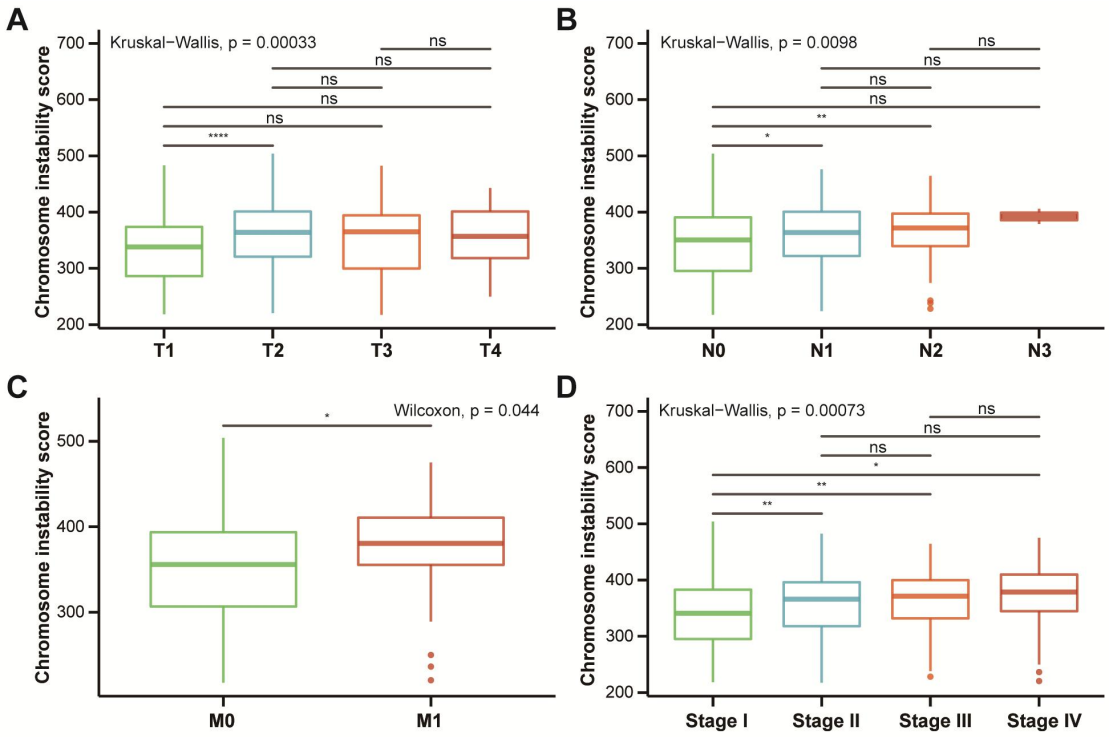

图 3 TNM 分期和临床分期中 CIN 差异

图注:（A）样本不同 T 分期的 CIN 风险得分差异。（B）样本不同 N 分期的 CIN 风险得分差异。（C）样本不同 M 分期的 CIN 风险得分差异。（D）样本不同 Stage 分期的 CIN 风险得分差异。X 轴表示分期，Y 轴表示 CIN 风险得分。

显著性水平：*代表 p<0.05，**代表 p<0.01，***代表 p<0.001，ns 代表 p>0.05。

（四）模型分类结果的预后效果评估

分析肺腺癌患者四种临床预后在 CIN 高和 CIN 低样本的差异，对两组样本的生存时间的比较方法用的是 Kaplan–Meier 检验，用 CIN 风险得分对四种临床数据样本分别进行分组，CIN 高低两组的阈值用 StepMiner 方法计算。从四种不同的预后指标（总生存 OS、疾病特异生存 DSS、无病间期 DFI 和无进展间期 PFI）对生存时间进行了评估。

如图 4-1 所示，CIN 高的样本的总生存显著低于 CIN 低的样本（log-rank p = 0.00069，单因素 cox 风险回归分析 HR = 1.68，95%CI：1.24-2.26），CIN 高的样本的 5 年总生存同样显著低于 CIN 低的样本（log-rank p = 0.00068，单因素 cox 风险回归分析 HR = 1.71，95%CI：1.25-2.33）。在疾病特异生存的预后分析中，CIN 高的样本的 DSS 显著低于 CIN 低的样本（log-rank p = 0.00028，单因素 cox 风险回归分析 HR = 2.06，95%CI：1.38-3.06）（图 4-2 A），CIN 高的样本的 5

年 DSS 同样显著低于 CIN 低的样本（log-rank p < 0.0001，单因素 cox 风险回归分析 HR = 2.25，95%CI：1.48-3.4）（图 4-2 B）。肺腺癌患者 CIN 高的样本无疾病间期比 CIN 低的样本显著短（log-rank p = 0.032，单因素 cox 风险回归分析 HR = 1.61，95%CI：1.04-2.5）（图 4-3 A），肺腺癌患者 CIN 高的样本 5 年内的无疾病间期同样比 CIN 低的样本显著短（log-rank p = 0.037，单因素 cox 风险回归分析 HR = 1.62，95%CI：1.03-2.54）（图 4-3 B）。无进展期的预后分析结论和前面相同，同样是 CIN 高的患者的预后比 CIN 低的样本显著的更差（log-rank p = 0.0063，单因素 cox 风险回归分析 HR = 1.47，95%CI：1.11-1.95）（图 4-4 B），5 年无进展期的预后分析结论同样是一致的（log-rank p = 0.0051，单因素 cox 风险回归分析 HR = 1.5，95%CI：1.13-1.99）（图 4-4 B）。这些结果已整合与表 2 中。结果表明，构建的模型可以准确地识别出与预后显著相关的样本，且在不同指标的预后结果中都获得了良好的结果，分类效果好。

表-2 预后效果评估

预后指标	log-rank p	HR	95% CI
OS	0.00069	1.68	1.24-2.26
5 年 OS	0.00068	1.71	1.38-3.06
DSS	0.00028	2.06	1.38-3.06
5 年 DSS	< 0.0001	2.25	1.48-3.4
DFI	0.032	1.61	1.04-2.5
5 年 DFI	0.037	1.62	1.03-2.54
PFI	0.0063	1.47	1.11-1.95
5 年 PFI	0.0051	1.5	1.13-1.99

对四种临床预后指标的分析表明，CIN 高和 CIN 低样本相比有更差的预后（OS、DSS、DFI 和 PFI），即构建的模型能准确地分类出预后效果不同的两组样本。

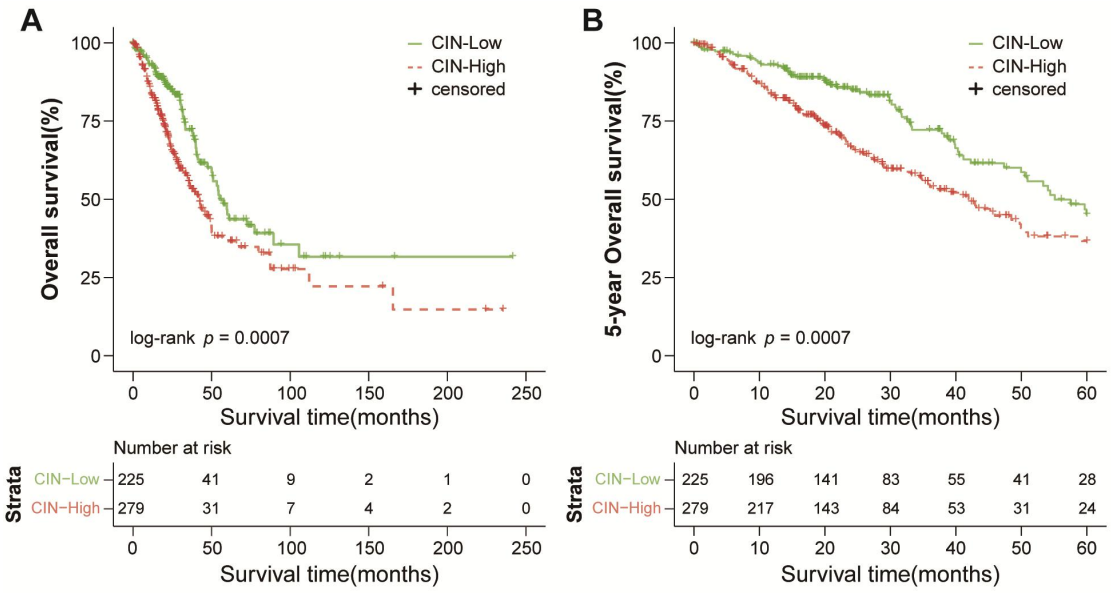

图 4-1 CIN 高和 CIN 低的样本 OS 预后分析

图注：（A）CIN 高和 CIN 低的两组样本的 OS 差异。（B）CIN 高和 CIN 低的两组样本的 5 年 OS 差异。绿色实线代表 CIN 低的样本，红色虚线代表 CIN 高的样本，+代表数据删失。

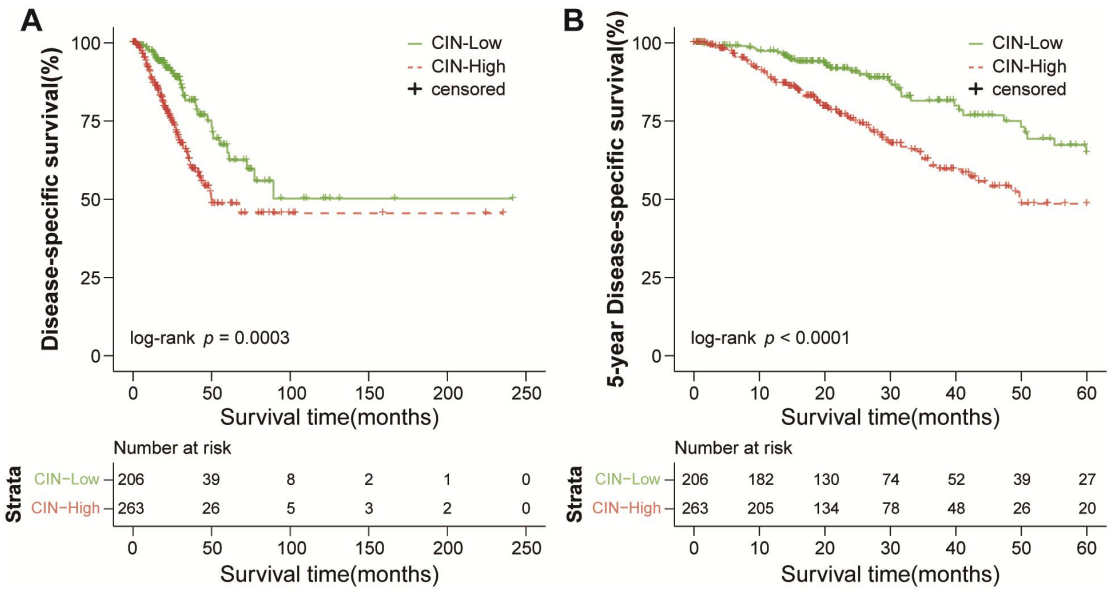

图 4-2 CIN 高和 CIN 低的样本 DSS 预后分析

图注：（A）CIN 高和 CIN 低的两组样本的 DSS 差异。（B）CIN 高和 CIN 低的两组样本的 5 年 DSS 差异。绿色实线代表 CIN 低的样本，红色虚线代表 CIN 高的样本，+代表数据删失。

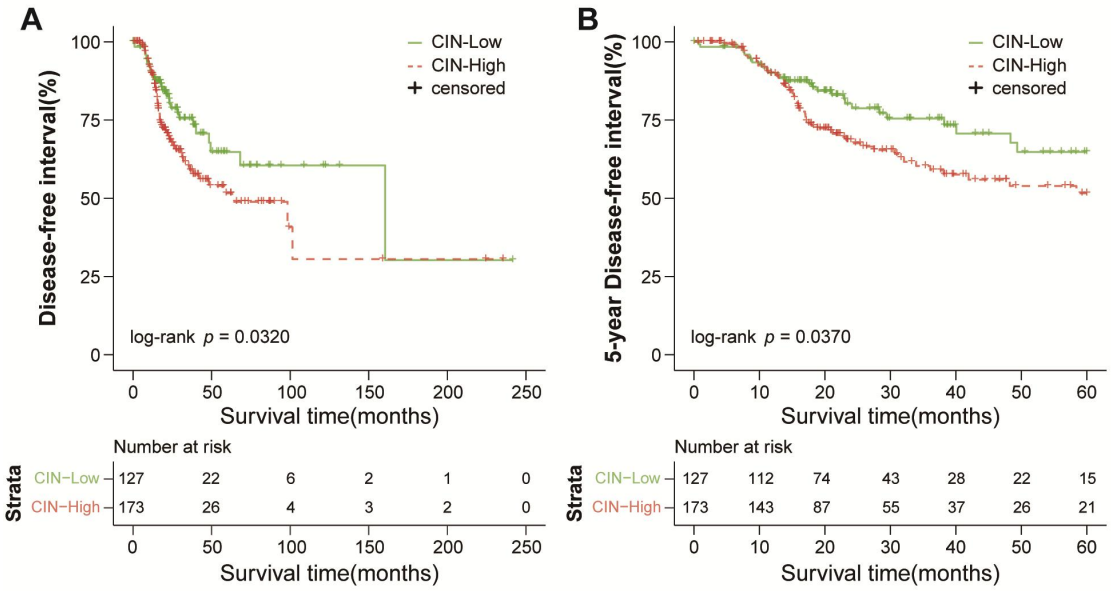

图 4-3 CIN 高和 CIN 低的样本 DFI 预后分析

图注：（A）CIN 高和 CIN 低的两组样本的 DFI 差异。（B）CIN 高和 CIN 低的两组样本的 5 年 DFI 差异。绿色实线代表 CIN 低的样本，红色虚线代 CIN 高的样本，+代表数据删失。

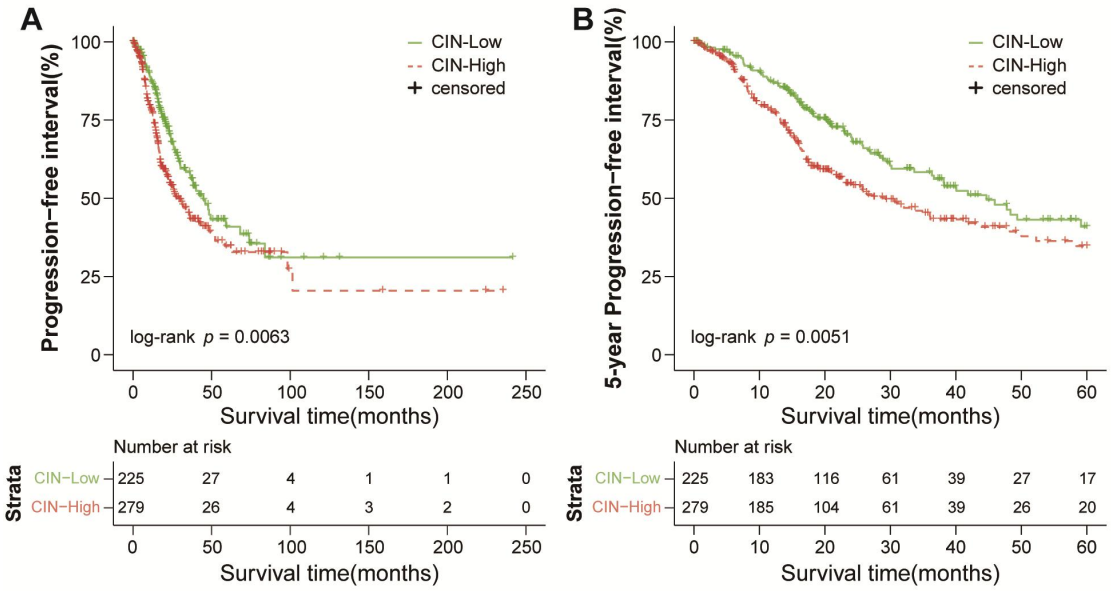

图 4-4 CIN 和 CIN 低的样本 PFI 预后分析

图注：（A）CIN 高和 CIN 低的两组样本的 PFI 差异。（B）CIN 高和 CIN 低的两组样本的 5 年 PFI 差异。绿色实线代表 CIN 低的样本，红色虚线代表 CIN 高的样本，+代表数据删失。

四、模型的应用及特征筛选

（一）CIN 低样本中 ceRNA 的识别

用 R 包 DESeq2 对肺腺癌的 Count 表达谱进行差异分析，识别出肺腺癌样本相比癌旁样本差异表达的 mRNA 和 lncRNA，这些 mRNA 和 lncRNA 视为可能与肺腺癌相关的基因，将这些 RNA 作为识别 CIN 低样本中的 ceRNA 关系对的范围。从 miRecords、miRTarBase 和 TarBase 三个数据库下载并整合了 miRNA-mRNA 实验验证的互作信息，miRNA 和 lncRNA 的靶向互作信息是从 DIANA-LncBase、starBase、lncRNASNP 数据库中获得，一共得到了 388895 条 miRNA 和 mRNA 无重复的互作数据，10318 条无重复的 miRNA 和 lncRNA 实验验证的互作数据。遍历 mRNA 和 lncRNA 对，用累计超几何富集检验筛选显著共调控的 mRNA 和 lncRNA 对（fdr<0.05），并且满足显著共表达（皮尔森相关系数大于 0，fdr<0.05）的条件作为 ceRNA。

在 CIN 低的样本中一共识别了 2536 对 mRNA 和 lncRNA 组成的 ceRNA 关系对，包含 854 个 mRNA 和 120 个 lncRNA。

（二）CIN 高样本中失调 ceRNA 的识别

对于在 CIN 低的样本中识别的 mRNA 和 lncRNA 组成的 ceRNA 对，在 CIN 高的样本中重新进行相关性分析，计算相关系数和相关性检验显著性 p 值。如果相关性检验的 p 值大于 0.05，就认为这对 mRNA 和 lncRNA 的表达水平不存在相关性，将相关系数赋值为 0，2536 对 ceRNA 对中有 744 对 ceRNA 在 CIN 高的样本中不再具有相关性，即 1792 对具有相关性。

对 ceRNA 在两类样本中的失调类型做了定义：如果 ceRNA 对在 CIN 高的样本中相关系数变得更大了，这样的 ceRNA 失调的类型定义为 gain 类型；如果 ceRNA 对在 CIN 高的样本中相关系数变得更小或者没有相关性（相关系数等于 0），这样的 ceRNA 失调的类型定义为 loss 类型。统计分析 2536 对 ceRNA 的失调情况，发现 1811 对 ceRNA 都是 loss 类型的，相关性降低或消失，大约占 71.4%，其中包括 744 对在 CIN 高的样本中不再具有相关性的失调 ceRNA；另外 725 对 ceRNA 失调类型为 gain 类型，占 28.6%，这部分 mRNA 和 lncRNA 的相关性在 CIN 高的样本中变得更强。

对 ceRNA 失调程度进行了基本的统计（表-3），通过对 ceRNA 失调类型和失调程度分析，失调 ceRNA 网络中 loss 类型的失调数量和程度都显著高于 gain 类型。

表-3 ceRNA 失调分析

失调关系	数量	CIN	相关系数（均值±标准差）	相关性改变量（绝对值）	p value
gain	725	高	0.21±0.079	0.09±0.066	2.2e-16
		低	0.18±0.076		
loss	1811	高	0.11±0.11	0.03±0.029	
		低	0.21±0.079		

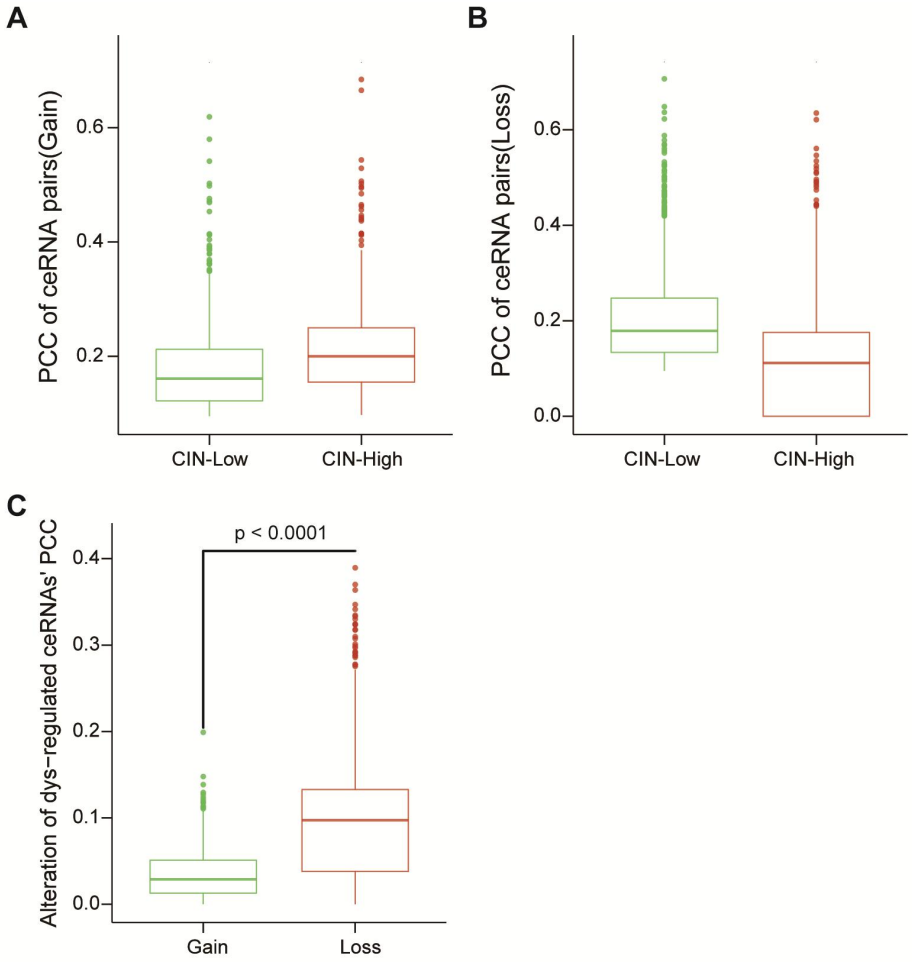

图-5 ceRNA 模型失调程度的统计

图注：（A）CIN 高和 CIN 低的两组样本中失调类型为 gain 的 ceRNA 对的相关系数。（B）CIN 高和 CIN 低的两组样本中失调类型为 loss 的 ceRNA 对的相关系数。绿色代表 CIN 低的样本，红色代表 CIN 高的样本。（C）ceRNA 失调类型为 gain 和 loss 类型的 ceRNA 对失调程度（绝对值）。绿色代表 gain 的 ceRNA 对，红色代表 loss 的 ceRNA 对。

（三）构建 CIN 相关双权重失调 ceRNA 网络

基于 CIN 低样本中识别的 mRNA 和 lncRNA 组成的 ceRNA 关系对构建染色体不稳定相关的失调 ceRNA 网络。失调 ceRNA 网络边的权重是利用 ceRNA 对在 CIN 高相比于 CIN 低的样本中皮尔森相关系数的改变程度，通过转换进行计算（详见材料与方法），网络边权重的值越大，ceRNA 在两类样本中相关性改变的程度越大。

对组成网络的每一个 mRNA 和 lncRNA 进行单因素 cox 风险回归分析，由于术后药物治疗和放疗是肺腺癌常见的治疗手段，可能会对肺癌的预后造成一定的影响，为了控制治疗的影响，选择没有接受术后治疗的 269 个样本，同时由于疾病特异生存（DSS）比总生存（OS）更能反映癌症对预后的影响，所以预后信息选择疾病特异生存。对网络中所有的点进行单因素 cox 风险回归分析，计算 mRNA 或 lncRNA 与预后 DSS 相关性的显著性 p 值，从而计算点的权重（详见材料与方法）。网络中的点与 DSS 单因素 cox 风险回归的显著性 p 值越小，点的权值越大，代表点的表达情况对预测预后 DSS 的贡献就越大。

构建染色体不稳定相关双权重 ceRNA 网络，网络包含点 974 个（mRNA 854 个，lncRNA 120 个，表-4），边 2536 条（图-6）。lncRNA 的平均度约为 21.13，mRNA 的平均度约为 2.97，lncRNA 的度显著高于 mRNA 的度（$p < 2.2e-16$，Wilcoxon 秩和检验）。正是因为很多 lncRNA 有很高的度，他们可以通过 ceRNA 机制间接调控很多 mRNA 的表达水平，进而可能影响蛋白质的翻译发挥生物学功能。

表-4 网络中节点数量与平均度

节点类型	数量	平均度	Wilcoxon pvalue
mRNA	854	2.97	2.2e-16
lncRNA	120	21.13	

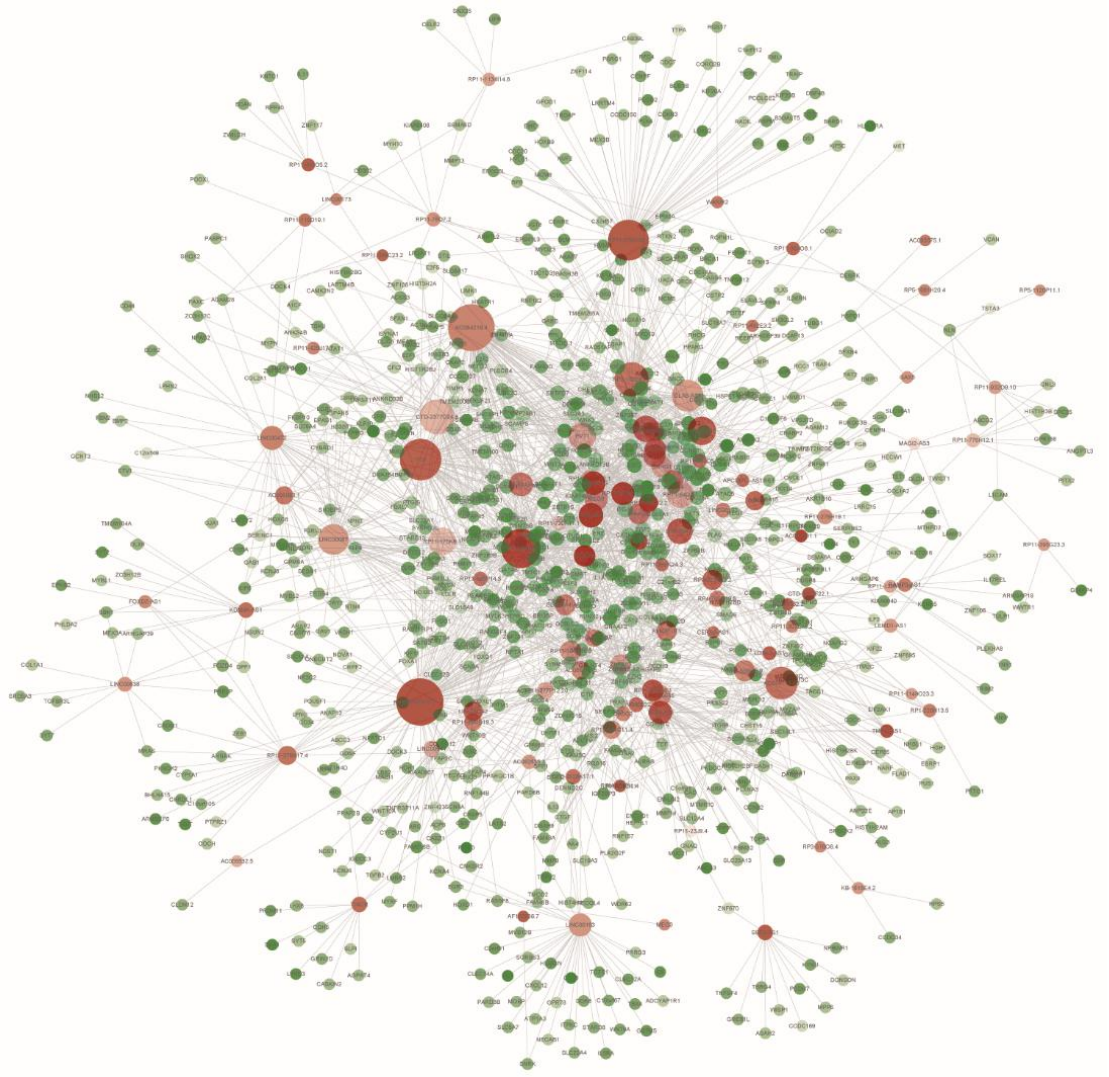

图-6 mRNA 和 lncRNA 构成的 CIN 相关双权重失调 ceRNA 网络模型

图注：绿色点代表 mRNA，红色点代表 lncRNA。失调网络中点的大小代表点的度，点的透明度代表点的权重大小，边的宽度代表边的权重大小（网络图由 Cytoscape[64]绘制）。

（四） CIN 相关失调 ceRNA 网络模块挖掘

对于构建的 mRNA 和 lncRNA 组成的 CIN 相关的具有点和边双权重的失调 ceRNA 网络，边的权值越高，代表在 CIN 低的样本中识别的 ceRNA 对在 CIN 高的样本中改变程度更大；点的权值越高，代表当前 mRNA 或 lncRNA 对预测预后 DSS 的贡献和价值越大。对网络模块挖掘的出发点是，挖掘到的模块具有点和边的权重都相对较大，这样的模块中 ceRNA 失调程度更大并且具有更好的预后潜能。

采用 R 包 dmGWAS_3.0 的方法进行模块挖掘，这个算法能够平衡点和边的

权重为模块进行打分，在已有模块的邻居点进行贪婪搜索，新加入的点能够使模块打分增加，直到模块打分增量不符合条件为止。挖掘模块的参数设置为：平衡点和边权重的参数λ为默认值（详细计算方法见材料与方法），经过计算得到λ的值为 0.42，在已有模块点的最短路径参数 d 等于 1 的范围内进行贪婪搜索，贪婪搜索模块打分的增量参数设置为 r=0.1。通过这种方法一共得到了 712 个模块。

（五）识别结果筛选与整合

分析 712 个模块，每一个模块包含 RNA 的个数为 5 到 13 个不等，这些模块一共覆盖了网络中 832 个点，而网络中一共包括 974 个点（由 854 个 mRNA 和 120 个 lncRNA 构成），由此可见挖掘到的这些模块对网络的覆盖度非常高，所以后续将对这些模块进行进一步的筛选，筛选出 ceRNA 对在 CIN 高和低两部分样本中改变程度相对更大和与预后 DSS 更相关的 6 个模块（表-5）。经过观察发现，6 个模块的 RNA 有大量的重复，且都含有基因 ANLN，说明这些模块在失调 ceRNA 网络中可以通过基因 ANLN 相连，在失调 ceRNA 网络背景中提取出由这些模块组成的子网，其中包含 20 个 RNA（图-7）。

在失调 ceRNA 网络中挖掘模块经过筛选和整合得到了由 20 个 RNA 组成的子网数据集，接下来会对数据集进行分析，旨在评估其作为预后标志物的潜能。

表-5 筛选后结果包含的 RNA

Module	Symbol of mRNAs and lncRNAs
Module1	ANLN,RHOV,CEP55,ADAM12,CDCA4,LINC00511
Module2	ANLN,CEP55,HMGA1,CDCA4,LINC00511,AC011893.3
Module3	ANLN,RHOV,CEP55,C1QTNF6,FAT2,CDCA4,LINC00511
Module4	ANLN,RHOV,CEP55,C1QTNF6,CDCA4,HIST2H2BE,LINC00511,LINC00473
Module5	HMGA2,ANLN,RHOV,CIDEC,CDK1,RAD51,HOXA5,RP3-523K23.2,LINC00511,LINC00152
Module6	ANLN,RHOV,CEP55,ADAM12,CDCA4,LINC00511,RP11-276H19.1

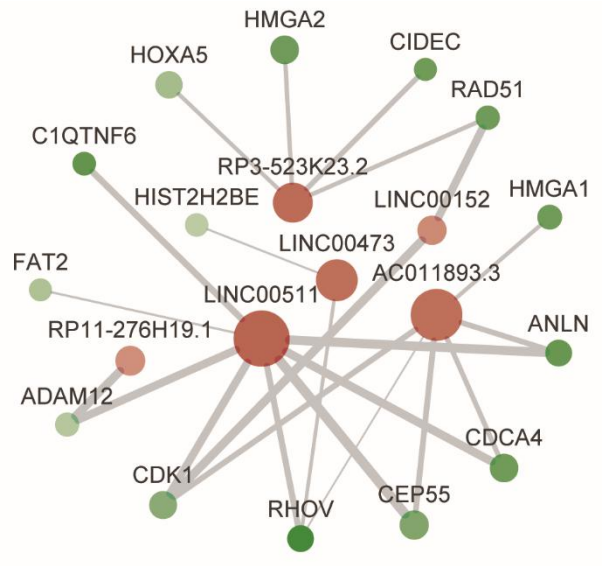

图-7 筛选后模块构成的子网数据集

图注：绿色点代表 mRNA，红色点代表 lncRNA。失调网络中点的大小代表点的度，点的透明度代表点的权重大小，边的宽度代表边的权重大小。

五、基于特征构建 cox 风险回归模型及识别结果分析

（一） 构建多因素 cox 风险回归模型

为了更准确评估数据集中的 RNA 对 DSS 的影响，排除术后药物治疗（主要包括化疗、靶向分子治疗、免疫治疗等）或放疗对预后的影响，选取未做治疗的样本构建多因素 cox 风险回归模型，计算每一个 RNA 的风险回归系数（表-6）。20 个 RNA 中 11 个 RNA 风险回归系数大于 0 时，趋向于风险因素；9 个 RNA 风险回归系数小于 0 时，趋向于保护因素。根据未做治疗的样本计算的风险回归系数计算所有样本的风险得分，每一个样本的风险得分为：当前样本中 20 个 RNA 的风险回归系数和表达值乘积的加和。

表-6 数据集中 RNA 的回归系数

RNA	coefficient
HMGA2	0.04405576
ANLN	0.50774076
RHOV	0.18207198
CIDEC	0.14330953
C1QTNF6	0.35484176
CDCA4	0.26342992
RP3-523K23.2	0.03374881
LINC00511	0.10120939
LINC00152	0.04444016
LINC00473	0.06508567
RP11-276H19.1	0.05766278
CEP55	-0.02017802
ADAM12	-0.12566402
CDK1	-0.62392838
HMGA1	-0.14241773
RAD51	-0.03105128
FAT2	-0.05172098
HIST2H2BE	-0.16015944
HOXA5	-0.36387485
AC011893.3	-0.06241368

由于治疗手段(术后放疗、药物治疗)会对肺腺癌患者 DSS 产生一定的影响，在评估 20 个 RNA 对预后的影响时将治疗手段纳入到考虑之中。

排除治疗对 DSS 的影响，在 215 个术后无治疗的样本中，用 StepMiner 方法将样本根据数据集风险得分成两组：高风险组 109 个样本和低风险组 160 个样本，高风险组样本的 DSS 比低风险组显著差，log rank p = 3.41e-06（HR = 4.07，95%CI：2.15-7.73）（图-8A），说明排除治疗对 DSS 的影响时，样本数据集风险得分对预后有着非常显著的影响。

考虑到治疗对 DSS 的影响，选取所有的 469 个样本，将样本分为高风险组 195 个样本和低风险组 274 个样本。生存分析结果表明，考虑到治疗对 DSS 的影响，高风险组样本仍然有着比低风险组样本更差的预后（log rank p = 9.17e-071，HR = 1.99，95%CI：1.71-3.63），数据集对预后的影响的显著性是稳定的（图-8B）。

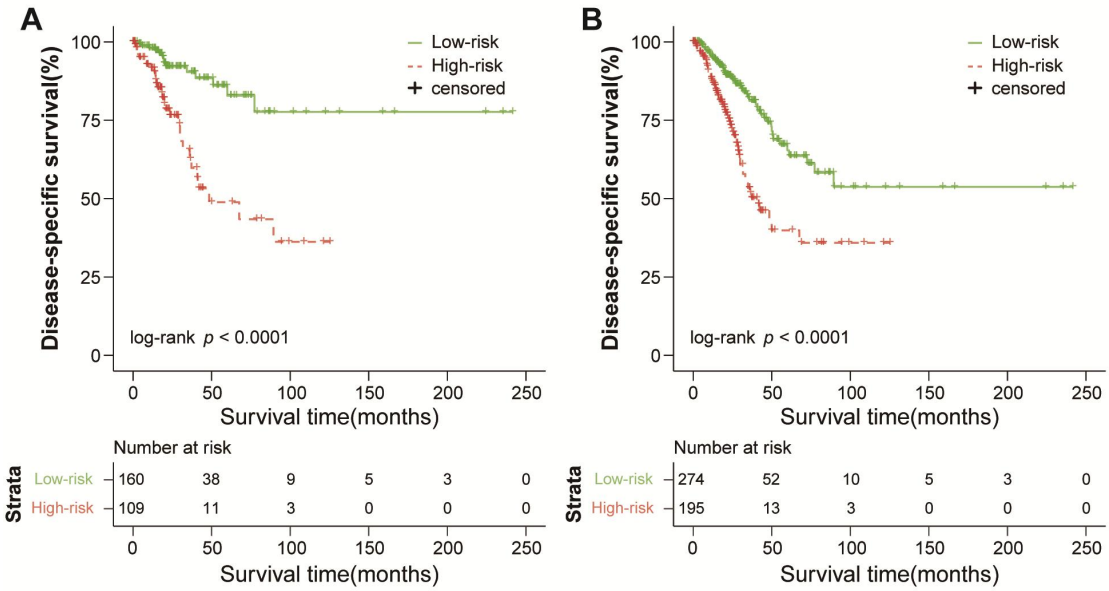

图-8 高风险样本和低风险样本 DSS 预后分析

图注：（A）未接受治疗样本高风险组和低风险组 DSS 生存差异。（B）所有样本高风险组和低风险组 DSS 生存差异。绿色实线代表低风险样本，红色虚线代表高风险样本，+代表数据删失。

在接受药物治疗的样本中，风险得分高和低两组样本进行生存分析结果表明，高风险样本比低风险样本预后情况差（图-9A，边缘显著，log rank p = 0.059，

HR = 1.73，95%CI：0.97-3.06）；在接受放疗的患者能更显著区分高低风险组的生存（图-9B，log rank p = 0.0061，HR = 2.35，95%CI：1.26-4.38）。

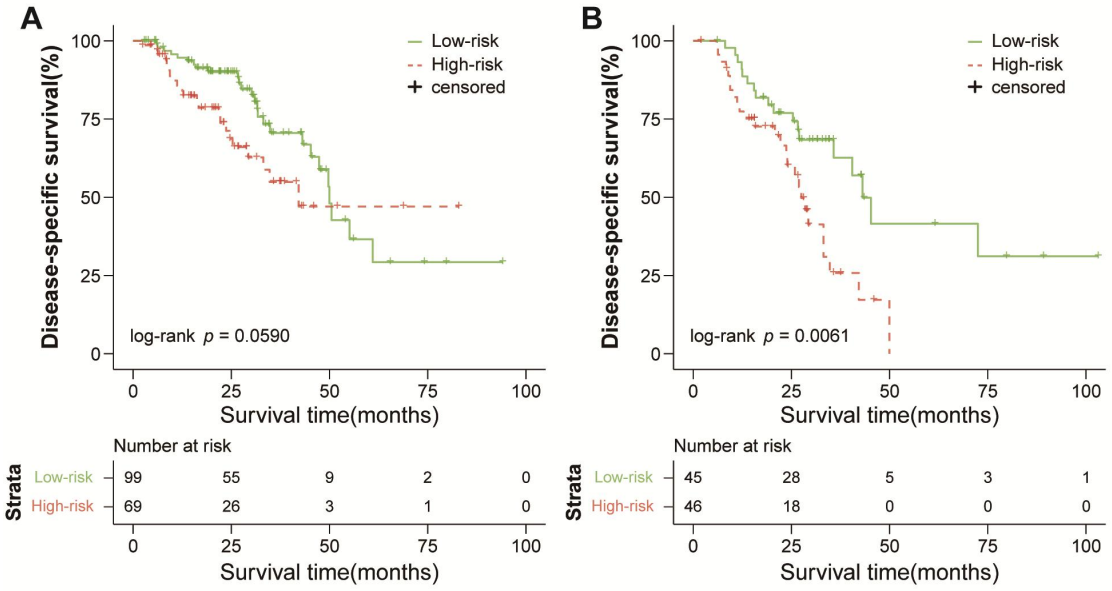

图-9 治疗后样本数据集高风险样本和低风险样本 DSS 预后分析

图注：（A）接受药物治疗的样本高风险组和低风险组 DSS 生存差异。（B）接受放疗的样本高风险组和低风险组 DSS 生存差异。绿色实线代表低风险样本，红色虚线代表高风险样本，+代表数据删失。

在无治疗样本中和所有样本的生存分析结果表明，20 个 RNA 高风险的患者相比较低风险的患者都有显著更差的预后。在术后有治疗的样本中，20 个 RNA 仍然能够很好的区分患者生存，可以从两个方面对这个结果进行解释：数据集的 20 个 RNA 对患者预后的预测的能力具有鲁棒性和独立性，受术后治疗影响较小；而低风险可能会提高患者对药物治疗或放疗的敏感性，这样的患者接受治疗可能更好的提升预后。这些结果表明构建的统计模型得到的结果可以良好地区分肺腺癌预后结果。

（二）数据集预后独立性评估

数据集的 20 个 RNA 风险得分、CIN 风险得分、临床分期与吸烟量可能对肺腺癌患者 DSS 产生影响，对这些因素进行相关性分析。每两组变量之间计算相关系数和显著性 p 值（斯皮尔曼秩相关）。

吸烟量与风险得分和临床分期并无显著的相关性（p>0.05），和 CIN 风险得分呈现正相关关系（图-10），相关系数为 0.10，显著性 p 值为 0.066，边缘显著，和已有研究发现的吸烟能与染色体不稳定性存在显著相关的结论是一致的[66–69]。

风险得分、CIN 风险得分和临床分期之间，两两显著正相关（图-10）。样本数据集的 20 个 RNA 风险得分和 CIN 风险得分的相关系数为 0.19，显著性 p 值为 7.79e-04，有着极显著的正相关关系；样本数据集的 20 个 RNA 风险得分和临床分期的相关系数为 0.23，显著性 p 值为 4.30e-05，同样有着极强的显著正相关关系；样本 CIN 风险得分和临床分期的相关系数为 0.15，显著性 p 值为 0.0093。

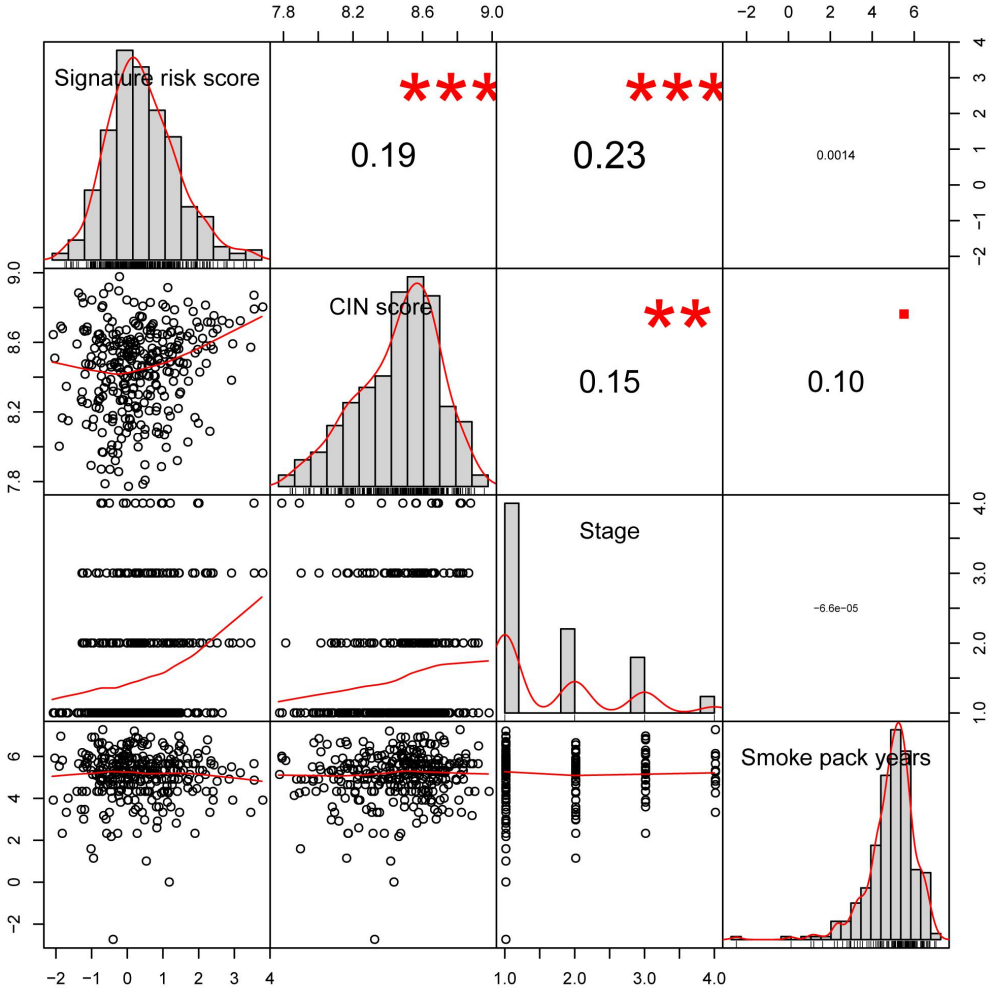

图-10 数据集的 20 个 RNA 风险得分、CIN 风险得分、临床分期和吸烟量的相关性分析

图注：显著性水平： .代表 0.05<p<0.1，*代表 p<0.05，**代表 p<0.01， ***代表 p<0.001

（三）逐步多因素 cox 比例风险回归

数据集的 20 个 RNA 对肺腺癌患者 DSS 的影响受治疗影响小，考虑到样本的年龄、性别、临床分期、吸烟量、CIN 风险得分对预后的影响，探究数据集的 20 个 RNA 是否能作为预测预后的独立因素。由上述相关性分析结果可知，肺腺癌患者的 20 个 RNA 风险得分、CIN 风险得分与临床分期呈现两两显著正相关性，吸烟量和 CIN 风险得分之间呈现边缘显著的正相关关系。由于变量之间存在多重共线性问题，采用逐步多因素 Cox 风险回归分析，评估变量对样本预后的独立性，逐步回归的特点是留下能够解释更多因变量变异的自变量组合。多因素 Cox 逐步回归分析结果表明，患者的年龄、性别、吸烟量和 CIN 风险得分对 DSS 无显著影响（p>0.05），只有临床分期（HR=1.270，95%CI：1.001-1.611）和 20 个 RNA 的风险得分（HR=1.573，95%CI：1.278-1.935）对预后有着独立显著的影响，倾向于独立的风险因素，并且 20 个 RNA 风险得分的显著性 p 值（p = 1.87e-05）比临床分期（p = 0.0493）更显著（表-7）。

总结数据集的 20 个 RNA 与预后相关的分析，在术后无治疗样本、接受治疗样本和所有样本的分析中，数据集单因素与预后生存分析的结果表明，其高风险显著降低了患者生存；考虑到其他对 DSS 可能产生影响的因素时，这 20 个 RNA 仍然可以作为独立的风险因素，说明数据集的 20 个 RNA 对预测肺腺癌患者的疾病特异生存具有很高的鲁棒性，有作为肺腺癌疾病特异生存预后标志物的潜能。

表-7 逐步多因素 cox 比例风险回归

	coefficients	HR	95%CI	pvalue
Stage	0.2388	1.270	1.001-1.611	0.0493
20RNA risk score	0.4529	1.573	1.278-1.935	1.87e-05

对数据集中的 14 个基因进行基因功能富集分析（具体结果见附录），功能富集分析结果表明，数据集中的基因显著富集到了癌症相关的功能通路，通过查阅已发表的 RNA 与癌症相关的研究，为这 20 个 RNA 可能作为预后标志物提供了一些佐证。

功能分析结果表明，数据集中的 20 个基因显著富集到染色体不稳定相关和癌症相关的功能通路，为这些作为潜在的肺腺癌预后标志物提供了进一步的证据支持。同时也证明了模型可以在肺腺癌样本中准确地识别出癌症预后标志物，这一模型同样也可以扩展到其他癌症中以识别其他癌症的生物预后标志物。

六、总结

本研究通过对肺腺癌患者样本的染色体不稳定性（CIN）进行打分，发现癌症样本的 CIN 显著高于癌旁样本，用 CIN 风险得分能够区分肺腺癌样本和癌旁样本，且有着非常好的分类效能（AUC 为 0.989）。同时，癌症样本高 CIN 与高临床分期和差的预后相关。

构建 CIN 相关失调 ceRNA（mRNA-lncRNA）双权重网络，寻找与 CIN 相关和与预后相关的 ceRNA，网络边的权重越大代表 ceRNA 互作对在 CIN 高相比于 CIN 低的样本，其失调程度越大；点的权重越大代表 RNA 与疾病特异生存越相关。用贪婪搜索算法在失调网络中挖掘点和边双权重大的模块，对模块进行筛选与整合，得到了由 14 个 mRNA 和 6 个 lncRNA 组成的子网。

这 20 个 RNA 对肺腺癌患者疾病特异生存有着稳定的影响，无论是在术后无治疗（药物治疗或放疗）、所有样本还是有治疗样本中，其高风险显著降低了患者预后。分析 20 个 RNA 的特性，发现 20 个 RNA 风险得分与 CIN 和临床分期都呈现显著正相关关系，为了解决变量多重共线性问题，采用逐步多因素 Cox 风险回归分析，此模型得到的结果作为预后风险因素仍然有着显著并独立的影响，其预测效能具有很好的鲁棒性。分析 20 个 RNA 的功能，发现它显著富集到 CIN 和癌症相关功能，进一步表明 20 个 RNA 可能作为预后相关的标志物。

综上所述，本研究基于肺腺癌样本染色体不稳定性的特性，并在 ceRNA 层面构建模型，识别肺腺癌预后相关的标志物时，选取的疾病特异生存，比总生存更能准确刻画疾病对预后的影响，分析标志物预后效能的独立性采用逐步回归的方法。结果分析表明，得到的 20 个 RNA 与肺腺癌预后有着显著并且稳定的相关性，能作为预测肺腺癌患者预后的潜在标志物，验证了模型的分类效能，同时这一模型也可以应用到其他癌症中寻找其他癌症的预后标志物。

参考文献

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* (2018). doi:10.3322/caac.21492
2. Kung, J. T. Y., Colognori, D. & Lee, J. T. Long noncoding RNAs: Past, present, and future. *Genetics* **193**, 651–669 (2013).
3. Lyon, M. F. Gene action in the X-chromosome of the mouse (mus musculus L.). *Nature* (1961). doi:10.1038/190372a0
4. Lee, J. T. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nature Reviews Molecular Cell Biology* (2011). doi:10.1038/nrm3231
5. Brown, C. J. *et al.* Localization of the X inactivation centre on the human X chromosome in Xq13. *Nature* (1991). doi:10.1038/349082a0
6. Lyle, R. *et al.* The imprinted antisense RNA at the lgf2r locus overlaps but does not imprint Mas1. *Nat. Genet.* (2000). doi:10.1038/75546
7. Schmidt, J. V., Matteson, P. G., Jones, B. K., Guan, X. J. & Tilghman, S. M. The Dlk1 and Gtl2 genes are linked and reciprocally imprinted. *Genes Dev.* (2000).
8. Rocha, S. T. da, Edwards, C. A., Ito, M., Ogata, T. & Ferguson-Smith, A. C. Genomic imprinting at the mammalian Dlk1-Dio3 domain. *Trends in Genetics* (2008). doi:10.1016/j.tig.2008.03.011
9. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* (2015). doi:10.1038/ng.3192
10. Mas-Ponte, D. *et al.* LncATLAS database for subcellular localization of long noncoding RNAs. *RNA* (2017). doi:10.1261/rna.060814.117
11. Lin, R. *et al.* Control of RNA processing by a large non-coding RNA over-expressed in carcinomas. *FEBS Lett.* (2011). doi:10.1016/j.febslet.2011.01.030
12. Ji, P. *et al.* MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* (2003). doi:10.1038/sj.onc.1206928
13. Tano, K. *et al.* MALAT-1 enhances cell motility of lung adenocarcinoma cells by influencing the expression of motility-related genes. *FEBS Lett.* (2010). doi:10.1016/j.febslet.2010.10.008
14. Kogo, R. *et al.* Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers. *Cancer Res.* (2011). doi:10.1158/0008-5472.CAN-11-1021
15. Niinuma, T. *et al.* Upregulation of miR-196a and HOTAIR drive malignant character in gastrointestinal stromal tumors. *Cancer Res.* (2012). doi:10.1158/0008-5472.CAN-11-1803
16. Gupta, R. A. *et al.* Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* (2010). doi:10.1038/nature08975
17. Kim, K. *et al.* HOTAIR is a negative prognostic factor and exhibits pro-oncogenic activity in pancreatic cancer. *Oncogene* (2013). doi:10.1038/onc.2012.193
18. Zhang, J. *et al.* Overexpression of FAM83H-AS1 indicates poor patient survival and knockdown impairs cell proliferation and invasion via MET/EGFR signaling in lung cancer. *Sci. Rep.* (2017). doi:10.1038/srep42819
19. Wang, C. *et al.* A cancer-testis non-coding RNA LIN28B-AS1 activates driver gene LIN28B by interacting with IGF2BP1 in lung adenocarcinoma. *Oncogene* (2018). doi:10.1038/s41388-018-0548-x
20. Lu, X. *et al.* A Novel Long Non-Coding RNA, SOX21-AS1, Indicates a Poor Prognosis and Promotes Lung Adenocarcinoma Proliferation. *Cell. Physiol.*

Biochem. (2017). doi:10.1159/000479543
21. Bartel, D. P. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* (2009). doi:10.1016/j.cell.2009.01.002
22. Salmena, L., Poliseno, L., Tay, Y., Kats, L. & Pandolfi, P. P. A ceRNA hypothesis: The rosetta stone of a hidden RNA language? *Cell* (2011). doi:10.1016/j.cell.2011.07.014
23. Cesana, M. *et al.* A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* (2011). doi:10.1016/j.cell.2011.09.028
24. Wang, J. *et al.* CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res.* (2010). doi:10.1093/nar/gkq285
25. Li, H. *et al.* Long non-coding RNA PVT1-5 promotes cell proliferation by regulating miR-126/SLC7A5 axis in lung cancer. *Biochem. Biophys. Res. Commun.* (2018). doi:10.1016/j.bbrc.2017.12.114
26. Xu, J. *et al.* The mRNA related ceRNA-ceRNA landscape and significance across 20 major cancer types. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv853
27. Wang, P. *et al.* Identification of lncRNA-associated competing triplets reveals global patterns and prognostic markers for cancer. *Nucleic Acids Res.* **43**, 3478–3489 (2015).
28. Li, Y. *et al.* Systematic review of computational methods for identifying miRNA-mediated RNA-RNA crosstalk. *Brief. Bioinform.* (2017). doi:10.1093/bib/bbx137
29. Xu, J. *et al.* Extensive ceRNA-ceRNA interaction networks mediated by miRNAs regulate development in multiple rhesus tissues. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw587
30. Paci, P., Colombo, T. & Farina, L. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst. Biol.* (2014). doi:10.1186/1752-0509-8-83
31. Zhang, Y. *et al.* Comprehensive characterization of lncRNA-mRNA related ceRNA network across 12 major cancers. *Oncotarget* (2016). doi:10.18632/oncotarget.11637
32. Sumazin, P. *et al.* An extensive MicroRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* (2011). doi:10.1016/j.cell.2011.09.041
33. Zhou, X., Liu, J. & Wang, W. Construction and investigation of breast - cancer - specific ceRNA network based on the mRNA and miRNA expression data. *IET Syst. Biol.* (2013). doi:10.1049/iet-syb.2013.0025
34. Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. StarBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **42**, 92–97 (2014).
35. Zhang, G. *et al.* Characterization of dysregulated lncRNA-mRNA network based on ceRNA hypothesis to reveal the occurrence and recurrence of myocardial infarction. *Cell death Discov.* (2018). doi:10.1038/s41420-018-0036-7
36. Shao, T. *et al.* Identification of module biomarkers from the dysregulated ceRNA-ceRNA interaction network in lung adenocarcinoma. *Mol. Biosyst.* (2015). doi:10.1039/c5mb00364d
37. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* (2011). doi:10.1016/j.cell.2011.02.013
38. Pihan, G. A., Wallace, J., Zhou, Y. & Doxsey, S. J. Centrosome abnormalities and chromosome instability occur together in pre-invasive carcinomas. *Cancer Res.* (2003).
39. Geigl, J. B., Obenauf, A. C., Schwarzbraun, T. & Speicher, M. R. Defining

'chromosomal instability'. *Trends Genet.* (2008). doi:10.1016/j.tig.2007.11.006
40. Lee, A. J. X. *et al.* Chromosomal instability confers intrinsic multidrug resistance. *Cancer Res.* (2011). doi:10.1158/0008-5472.CAN-10-3604
41. Swanton, C. *et al.* Regulators of Mitotic Arrest and Ceramide Metabolism Are Determinants of Sensitivity to Paclitaxel and Other Chemotherapeutic Drugs. *Cancer Cell* (2007). doi:10.1016/j.ccr.2007.04.011
42. Mettu, R. K. R., Wan, Y. W., Habermann, J. K., Ried, T. & Guo, N. L. A 12-gene genomic instability signature predicts clinical outcomes in multiple cancer types. *Int. J. Biol. Markers* (2010). doi:10.5301/JBM.2010.6079
43. Zhang, W. *et al.* Centromere and kinetochore gene misexpression predicts cancer patient survival and response to radiotherapy and chemotherapy. *Nat. Commun.* (2016). doi:10.1038/ncomms12619
44. Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat. Genet.* **38**, 1043–1048 (2006).
45. Roylance, R. *et al.* Relationship of extreme chromosomal instability with long-term survival in a retrospective analysis of primary breast cancer. *Cancer Epidemiol. Biomarkers Prev.* (2011). doi:10.1158/1055-9965.EPI-11-0343
46. Birkbak, N. J. *et al.* Paradoxical relationship between chromosomal instability and survival outcome in cancer. *Cancer Res.* (2011). doi:10.1158/0008-5472.CAN-10-3667
47. Beinert, H. (1954). BIOLOGICAL OXIDATIONSl, 2, (1). *et al.* DARS2 protects against neuroinflammation and apoptotic neuronal loss, but is dispensable for myelin producing cells. *Hum. Mol. Genet.* (2017). doi:10.1093/hmg/ddx307
48. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Wspolczesna Onkologia* (2015). doi:10.5114/wo.2014.47136
49. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Article Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. 400–416 (2018). doi:10.1016/j.cell.2018.02.052
50. Harrow, J. *et al.* GENCODE: The reference human genome annotation for the ENCODE project. *Genome Res.* **22**, 1760–1774 (2012).
51. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
52. Xiao, F. *et al.* miRecords: An integrated resource for microRNA-target interactions. *Nucleic Acids Res.* **37**, 105–110 (2009).
53. Chou, C. *et al.* miRTarBase update 2018 : a resource for experimentally validated microRNA-target interactions. **46**, 296–302 (2018).
54. Vergoulis, T. *et al.* TarBase 6.0: Capturing the exponential growth of miRNA targets with experimental support. *Nucleic Acids Res.* **40**, 222–229 (2012).
55. Paraskevopoulou, M. D. *et al.* DIANA-LncBase: Experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.* (2013). doi:10.1093/nar/gks1246
56. Miao, Y. R., Liu, W., Zhang, Q. & Guo, A. Y. LncRNASNP2: An updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.* **46**, D276–D280 (2018).
57. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).
58. Sahoo, D., Dill, D. L., Tibshirani, R. & Plevritis, S. K. Extracting binary signals from microarray time-course data. **35**, 3705–3712 (2007).
59. Dalerba, P. *et al.* Single-cell dissection of transcriptional heterogeneity in

human colon tumors. *Nat. Biotechnol.* (2011). doi:10.1038/nbt.2038
60. Dalerba, P. *et al.* CDX2 as a Prognostic Biomarker in Stage II and Stage III Colon Cancer. *N. Engl. J. Med.* (2016). doi:10.1056/nejmoa1506597
61. Liu, Y., Ji, Y. & Qiu, P. Identification of thresholds for dichotomizing DNA methylation data. 1–8 (2013).
62. Wang, Q., Yu, H., Zhao, Z. & Jia, P. EW-dmGWAS: Edge-weighted dense module search for genome-wide association studies and gene expression profiles. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv150
63. Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. dmGWAS: Dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btq615
64. Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P. L. & Ideker, T. Cytoscape 2.8: New features for data integration and network visualization. *Bioinformatics* (2011). doi:10.1093/bioinformatics/btq675
65. Freeman, J. R., Chu, S., Hsu, T. & Huang, Y.-T. Epigenome-wide association study of smoking and DNA methylation in non-small cell lung neoplasms. *Oncotarget* **7**, (2016).
66. Li, E. *et al.* Continual exposure to cigarette smoke extracts induces tumor-like transformation of human nontumor bronchial epithelial cells in a microfluidic chip. *J. Thorac. Oncol.* (2014). doi:10.1097/JTO.0000000000000219
67. Saletta, F. *et al.* Exposure to the tobacco smoke constituent 4-aminobiphenyl induces chromosomal instability in human cancer cells. *Cancer Res.* (2007). doi:10.1158/0008-5472.CAN-06-4420
68. De La Chica, R. A., Ribas, I., Giraldo, J., Egozcue, J. & Fuster, C. Chromosomal instability in amniocytes from fetuses of mothers who smoke. *J. Am. Med. Assoc.* (2005). doi:10.1001/jama.293.10.1212
69. Huang, J., Okuka, M., McLean, M., Keefe, D. L. & Liu, L. Telomere susceptibility to cigarette smoke-induced oxidative damage and chromosomal instability of mouse embryos in vitro. *Free Radic. Biol. Med.* (2010). doi:10.1016/j.freeradbiomed.2010.03.026
70. Ni, Z. *et al.* Ailanthone inhibits non-small cell lung cancer cell growth through repressing DNA replication via downregulating RPA1. *Br. J. Cancer* **117**, 1621–1630 (2017).
71. Wang, W. *et al.* Cathepsin L activated by mutant p53 and Egr-1 promotes ionizing radiation-induced EMT in human NSCLC. *J. Exp. Clin. Cancer Res.* (2019). doi:10.1186/s13046-019-1054-x
72. Castellano, J. J. *et al.* LincRNA-p21 Impacts Prognosis in Resected Non-Small Cell Lung Cancer Patients through Angiogenesis Regulation. *J. Thorac. Oncol.* **11**, 2173–2182 (2016).
73. Gao, X. *et al.* HMGA2 regulates lung cancer proliferation and metastasis. *Thorac. Cancer* (2017). doi:10.1111/1759-7714.12476
74. Barletta, J. A., Yeap, B. Y. & Chirieac, L. R. Prognostic significance of grading in lung adenocarcinoma. *Cancer* (2010). doi:10.1002/cncr.24831
75. Shepelev, M. V. & Korobko, I. V. The RHOV gene is overexpressed in human non-small cell lung cancer. *Cancer Genet.* (2013). doi:10.1016/j.cancergen.2013.10.006
76. Takeuchi, T., Adachi, Y. & Nagayama, T. Expression of a secretory protein C1qTNF6, a C1qTNF family member, in hepatocellular carcinoma. *Anal. Cell. Pathol.* (2011). doi:10.3233/ACP-2011-009
77. Xu, Y., Wu, X., Li, F., Huang, D. & Zhu, W. CDCA4, a downstream gene of the Nrf2 signaling pathway, regulates cell proliferation and apoptosis in the MCF‐7/ADM human breast cancer cell line. *Mol. Med. Rep.* **17**, 1507–1512 (2018).
78. Xing, C. *et al.* ΔNp63α induces the expression of FAT2 and Slug to promote tumor invasion. *Oncotarget* (2016).

doi:10.18632/oncotarget.8696
79. Teo, W. W. *et al.* HOXA5 determines cell fate transition and impedes tumor initiation and progression in breast cancer through regulation of E-cadherin and CD24. *Oncogene* (2016). doi:10.1038/onc.2016.95
80. Ding, J., Yang, C. & Yang, S. LINC00511 interacts with miR-765 and modulates tongue squamous cell carcinoma progression by targeting LAMC2. *J. Oral Pathol. Med.* **47**, 468–476 (2018).
81. Yan, L., Wu, X., Liu, Y. & Xian, W. LncRNA Linc00511 promotes osteosarcoma cell proliferation and migration through sponging miR-765. *J. Cell. Biochem.* (2018). doi:10.1002/jcb.27999
82. Lu, G. *et al.* Long noncoding RNA LINC00511 contributes to breast cancer tumourigenesis and stemness by inducing the miR-185-3p/E2F1/Nanog axis. *J. Exp. Clin. Cancer Res.* **37**, 289 (2018).
83. Sun, C.-C. *et al.* Long Intergenic Noncoding RNA 00511 Acts as an Oncogene in Non-small-cell Lung Cancer by Binding to EZH2 and Suppressing p57. *Mol. Ther. Nucleic Acids* **5**, e385 (2016).
84. Yu, M. *et al.* Linc00152 promotes malignant progression of glioma stem cells by regulating miR-103a-3p/FEZF1/CDC25A pathway. *Mol. Cancer* **16**, 110 (2017).
85. Cai, Q. *et al.* Long non-coding RNA LINC00152 promotes gallbladder cancer metastasis and epithelial-mesenchymal transition by regulating HIF-1α via miR-138. *Open Biol.* **7**, (2017).
86. Feng, S. *et al.* Overexpression of LINC00152 correlates with poor patient survival and knockdown impairs cell proliferation in lung cancer. *Sci. Rep.* **7**, 2982 (2017).
87. Zhang, P.-P. *et al.* Linc00152 promotes Cancer Cell Proliferation and Invasion and Predicts Poor Prognosis in Lung adenocarcinoma. *J. Cancer* **8**, 2042–2050 (2017).
88. Wang, L., Zhang, X., Sheng, L., Qiu, C. & Luo, R. LINC00473 promotes the Taxol resistance via miR-15a in colorectal cancer. *Biosci. Rep.* **38**, (2018).
89. Shi, C. *et al.* The long noncoding RNA LINC00473, a target of microRNA 34a, promotes tumorigenesis by inhibiting ILF2 degradation in cervical cancer. *Am. J. Cancer Res.* **7**, 2157–2168 (2017).
90. Chen, Z. *et al.* cAMP/CREB-regulated LINC00473 marks LKB1-inactivated lung cancer and mediates tumor growth. *J. Clin. Invest.* **126**, 2267–79 (2016).

附 录

本论文中使用到的程序具体内容因篇幅原因，已提供在数据包中。

附表 1 20 个基因生物学过程的功能富集

Go terms	Overlap.G	Path.G	pvalue
DNA conformation change	5	292	2.12e-06
DNA packaging	4	208	1.71e-05
heterochromatin assembly	2	10	2.62e-05
heterochromatin organization	2	16	6.96e-05
positive regulation of cellular senescence	2	17	7.89e-05
cell aging	3	111	8.37e-05
positive regulation of cell aging	2	19	9.91e-05
mitotic G2 DNA damage checkpoint	2	23	0.000146368
chromatin assembly	3	164	0.000265791
mitotic G2/M transition checkpoint	2	31	0.000268042
chromosome condensation	2	34	0.000322941
response to X-ray	2	35	0.000342358
histone phosphorylation	2	37	0.000382864
chromatin assembly or disassembly	3	190	0.000409331
G2 DNA damage checkpoint	2	39	0.000425593
base-excision repair	2	43	0.000517698
regulation of cellular senescence	2	45	0.000567062
cortical cytoskeleton organization	2	50	0.00070008
regulation of cell aging	2	53	0.00078645
regulation of double-strand break repair	2	58	0.000941263
mitotic cytokinesis	2	67	0.001253881
cellular response to ionizing radiation	2	67	0.001253881
protein-DNA complex subunit organization	3	283	0.00130167
cellular senescence	2	71	0.001406719
DNA replication	3	296	0.001481052
aging	3	298	0.001509962
DNA duplex unwinding	2	75	0.00156804
cytoskeleton-dependent cytokinesis	2	81	0.001825825
DNA geometric change	2	85	0.002008148
positive regulation of cell cycle arrest	2	88	0.002150352

cellular response to drug	3	340	0.002201576
negative regulation of cell cycle process	3	350	0.002390956
negative regulation of G2/M transition of mitotic cell cycle	2	97	0.00260484
regulation of angiogenesis	3	366	0.002714629
regulation of DNA repair	2	100	0.002765563
mitotic DNA damage checkpoint	2	107	0.003158362
negative regulation of cell cycle G2/M phase transition	2	109	0.003275136
regulation of vasculature development	3	396	0.003392364
regulation of DNA metabolic process	3	399	0.003465358
mitotic DNA integrity checkpoint	2	113	0.003514708
regulation of cell cycle arrest	2	113	0.003514708

附表 2 20 个 RNA 细胞组成的功能富集

Go terms	Overlap.G	Path.G	pvalue
Midbody	3	166	0.000233
cell division site part	2	64	0.001022
heterochromatin	2	73	0.001327
cell division site	2	81	0.001631
nuclear chromosome, telomeric region	2	124	0.003768

附表 3 20 个 RNA 分子功能富集

Go terms	Overlap.G	Path.G	pvalue
AT DNA binding	2	10	2.65e-05
DNA-(apurinic or apyrimidinic site) endonuclease activity	2	13	4.59e-05
catalytic activity, acting on DNA	3	197	0.000462988

注：Go terms:20 个 RNA 显著富集到的 GO 功能；Overlap.G：20 个 RNA 与 GO 功能显著交叠的基因个数；Path.G：GO 功能包含的基因个数；pvalue：20 个 RNA 与 GO 功能基因超几何富集检验的显著性 p 值。

致　谢

首先要感谢我们的指导教师崔颖老师，无论在学习还是生活中，崔颖老师都给予了我们极大的帮助。崔颖老师严谨的工作态度、一丝不苟的学术作风都是我们学习的榜样，对我工作上的督促及鼓励都给了我们极大的信心。崔颖老师生活中十分地平易近人、和蔼可亲，让我们感受到了亲人般的温暖。在工作和生活中，崔颖老师都极大地感染并激励了我们，让我们受益匪浅，并且使我们在未来的工作和学习中更加地积极乐观。

我们还要感谢科室的其他老师：李孔宁教授、徐建凯老师和鲁健平老师在生活和科研中对我们的帮助。感谢我们的同学以及科室的师兄师姐师弟师妹们在科研和生活中对我们提供的帮助，在大家共同营造的认真严谨的科研环境和轻松愉快的科室氛围中，我们不仅学到了知识，同时也结交了很多真挚的朋友。

还要真挚地感谢李霞院长及生物信息科学与技术学院的所有老师们，感谢你们创造出来的浓浓的生物信息的学习与科研氛围。我们更感激学校为我们提供的这次参加统计建模大赛的机会，让我们的学习能力得到了进一步的提高。