肺腺癌中预后多维转录组分子标签的构建

郑鸿轩[1],张　建[2]
(1. 河南省温县人民医院,河南 焦作 454850;2. 哈尔滨医科大学(大庆)医学信息学院,黑龙江 大庆 163000)

摘　要:[目的] 通过对 TCGA 数据库中肺腺癌数据进行挖掘,构建由编码基因(PCG)、长链非编码 RNA(lncRNA)和小 RNA(microRNA)组成的多维转录组分子标签。[方法] 采用 Cox 风险回归、Kaplan-Meier 法、随机生存森林、ROC 分析等方法,挖掘 TCGA 癌症公共数据库中肺腺癌转录组二代测序数据,筛选预测效能良好的多维转录组分子标签。[结果] 纳入的 397 例肺腺癌患者的平均年龄为 65.67 岁,平均生存时间为 20.77 个月。筛选得到由 ELOVL6、RP11-446E9.2、CTD-2555C10.3、PACERR、hsa-mir-140、hsa-mir-31 和 hsa-mir-582 构成的多维转录组分子标签对肺腺癌患者预后预测效能良好。ROC 分析其预测效能显示,该分子标签 AUC 值为 0.73,大于 TNM 分期的 0.65(测试组:0.68 *vs.* 0.66)。该分子标签能将肺腺癌患者分成高低风险组,生存时间有显著差异(中位生存时间:25.3 个月 *vs.* 85.3 个月,P<0.001;HR=2.36,95%CI:1.88~2.98,199 例)。在测试组 Kaplan-Meier 分析该多维转录分子组标签也能将患者分成高低风险组(中位生存时间:39.8 个月 *vs.* 59.3 个月,P<0.05,198 例)。且多因素 Cox 回归显示该多维转录组分子标签为独立预后因子。[结论] 本研究通过对 TCGA 数据库的挖掘,构建的多维转录组分子模型对肺腺癌患者预后有良好的指示作用,可作为潜在的肺腺癌患者预后指示标签。
关键词:肺腺癌;长链非编码 RNA;小 RNA;预后
中图分类号:R734.2　　**文献标识码**:A　　**文章编号**:1004-0242(2017)10-0820-05
doi:10.11735/j.issn.1004-0242.2017.10.A014

Construction of Multi-dimensional Transcriptom Signature for Prognosis of Lung Adenocarcinoma Patients

ZHENG Hong-xuan[1],ZHANG Jian[2]
(1. Wenxian People's Hospital,Jiaozuo 454850,China;2. Department of Medical Informatics, Daqing Campus,Harbin Medical University,Daqing 163000,China)

Abstract:[Purpose] To construct a multi-dimensional transcriptom signature consisting of protein-coding gene(PCG),long non-coding RNA(lncRNA),microRNA with data-mining the Cancer Genome Atlas (TCGA)public database for prognosis of patients with lung adenocarcinoma.[Methods] Using univariate Cox regression,random survival forest algorithm and ROC analysis,the prognostic markers of lung adenocarcinoma were screened and the multi-dimensional signature was constructed. [Results] The mean age of 397 patients with lung adenocarcinoma was 65.67 years with a mean survival time of 20.77 months.The selected signature was composed by ELOVL6,RP11-446E9.2,CTD-2555C10.3,PACERR,hsa-mir-140,hsa-mir-31,hsa-mir-58,which had highest the area under ROC curve(AUC) in prediction of disease outcome(0.73 Signature *vs.* 0.65 TNM in the training group and 0.68 Signature *vs.* 0.66 TNM in the test group). The patients were divided into high- or low-risk group which were significantly associated with survival of lung adenocarcinoma patients in the training group (median survival:25.3 months *vs.* 85.3 months,P<0.001,HR=2.36,95%CI:1.88~2.98,n=199). The signature was applied to the test group,showing similar prognostic values(median survival:39.8 months *vs.* 59.3 months,P<0.05,n=198). Multivariate Cox regression analysis showed that the signature was an independent prognostic factor for patients with lung adenocarcinoma. [Conclusion] With TCGA data mining,the constructed signature can predict the survival of patients with high accuracy,which may be used as a potential prognostic marker for lung adenocarcinoma.
Key words:lung adenocarcinoma;long non-coding RNA;microRNAs;prognosis

肺腺癌严重危害着人类健康，已成为世界上发病率和死亡率最高的恶性肿瘤，其特点表现为恶性程度高、易复发和转移。目前对肺腺癌的复发和转移治疗效果不理想,导致肺腺癌患者生存率非常低,预后很差[1]。提高肺腺癌转移的诊断和治疗是降低其死亡率的主要手段。这迫使科研工作者和临床医师

收稿日期:2016-12-02;修回日期:2017-02-28
通讯作者:张建,E-mail:hmudqzj@163.com

不断努力寻找肺腺癌侵袭转移相关的分子标志物。目前研究认为癌症的发生发展是多阶段、多因素共同作用基因表达调控发生改变的过程[2,3]。近些年研究者报道用高通量测序技术分析了肺腺癌的分化和预后，得到了一些肺腺癌分化和预后的相关 PCG、lncRNA、microRNA[4-9]。诸多研究报道 microRNAs 和 lncRNA 可对 PCGs 转录表达进行调控，因此 microRNAs、lncRNAs 联合 PCGs 表达能多维度地反映肿瘤发生发展过程，从而更好地指示患者的预后[10-12]，而关于从转录组多维度开发指示预后的分子标签研究很少。为更好地从多维度水平揭示基因转录表达与肺腺癌患者预后的关系，本文通过挖掘公共数据，结合生物信息学算法，构建筛选了多维转录组分子模型，对肺腺癌患者预后有良好的指示作用。

1 资料与方法

1.1 数据收集

数据资料收集从 TCGA 数据库（https://tcga-data.nci.nih.gov/tcga/）下载并预处理肺腺癌癌数据集的 PCG 表达 RNASEqV2 数据和 microRNA 数据和相关临床资料。从（http://ibl.mdanderson.org/）得到对应肺腺癌 lncRNA 表达数据。通过对三套表达谱矩阵去空值处理，保留了在 70%患者中均表达的 PCG、lncRNA、microRNA。采用 K-mean 法进行空值替代。得到有 14 896 个 PCG，7174 个 lncRNA，416 个 microRNA，总计 397 例样本的多维转录组表达矩阵，并随机分成训练集和测试集两组，进行模型构建。

1.2 加权拟合多基因构建预测总生存期模型风险得分算法

在训练数据集中，我们使用单变量 Cox 回归分析评估每个 PCG、lncRNA、microRNA 表达水平与患者的预后关系。随后，我们加权拟合得到一个模型来估计模型对预后风险指示，公式如下：

$$\text{Risk Score (RS)} = \sum^{N}_{i=1} (\text{Exp} * \text{Coef})$$ [13,14]

其中 N 是 PCG、lncRNA、microRNA 总个数，Exp 是对应 PCG、lncRNA、microRNA 的表达量，Coef 为单因素 Cox 回归系数。Risk Score (RS)是多节点的加权拟合后的风险得分。

1.3 统计学方法

使用 SPSS17.0 软件进行统计学分析。临床病理参数相关性分析，组间比较采用 χ^2 检验及 Fisher 确切概率法，生存分析采用 Kaplan–Meier 和 Log-rank 检验法。高通量计算如 Cox 风险回归、Kaplan–Meier、随机生存森林、ROC 分析采用 survivalROC、survival、randomForestSRC 等 R 包计算分析，R 包均来自 BioconductorR（http://www.bioconductor.org/）。P<0.05 为差异有统计学意义。

2 结 果

2.1 构建并筛选多维转录组分子模型

首先我们对所有 PCGs、lncRNAs、microRNAs 进行 Cox 单因素分析，得到与肺腺癌预后相关的其中 1291 个 PCG，521 个 lncRNA ，47 个 microRNA，总 1859 个与预后相关的因子，并分别进行随机生存森林筛选，降维后得到 4 个 PCGs，5 个 lncRNAs，4 个 microRNAs，接着进行拟合加权全排列后得到 8191 个多维转录组模型，筛选后的得到由 ELOVL6、PACERR、RP11-446E9.2、CTD-2555C10.3、hsa-mir-140、hsa-mir-31、hsa-mir-58 组合成的多维转录组分子标签（Table 1），预后指示效能最好（AUC=0.72）。

Table 1 PCGs，lncRNAs and microRNAs in the prognostic expression signature，and their univariable Cox association with prognosis

Ensembl ID	Gene symbol	HR [a]	*P*-value [a]	Gene expression level association with poor prognosis	Chromosome location
ENSG00000170522	ELOVL6	1.42	0.00	Low	8:23702451–23706598 [–]
ENSG00000259230	CTD-2555C10.3	3.67	0.00	Low	14:102545254–102555826 [+]
ENSG00000271722	RP11–446E9.2	3.25	0.02	Low	8: 56013487–56014168 [+]
ENSG00000273129	PACERR	1.58	0.02	Low	1:186680622–186681446 [+]
hsa-miR-140	MIR140	0.45	0.00	High	16: 69933081–69933180 [+]
hsa-miR-31	MIR31	1.40	0.00	Low	9: 21512115–21512185[–]
hsa-miR-582	MIR582	1.18	0.00	Low	5: 59703606–59703703 [–]

Note：a：derived from the univariable Cox regression analysis in the training set.

2.2 筛选的多转录组模型对肺腺癌患者预后指示

在训练集中总 199 例患者，使用该分子模型的风险得分中位数作为分界点，患者被分为高风险组(100 例)或低风险组(99 例)。高危组患者中位生存时间(25.3 个月)明显比低风险组(85.3 个月)短($P<0.05$)(Figure 1A)。高危组患者 5 年生存率不到 20%，而低风险组在 50%以上。风险得分越高，预后越差。并且与临床属性分析中，该模型与 N 分期、TNM 分期相关(Table 2)。构建预后模型在测试数据集有相同指示 (中位生存时间：39.8 个月 *vs.* 59.3 个月，$P<0.05$)。验证了预后生存预测稳定性(Figure 2A)。多因素 Cox 回归显示该多维转录组分子标签为预后独立因子(Table 3)。

2.3 筛选的模型与 TNM 的联合分析

采用 ROC 分析预测效能，得到筛选的多维转录组模型比 TNM 系统分期预测效能更好，并且通过 Cox 单因素系数拟合后，发现该分子标签联合 TNM 分期可进一步提高预测效能(训练集：$AUC_{联合}=0.75$ *vs.* $AUC_{分子标签}=0.73$ *vs.* $AUC_{TNM}=0.65$，测试集：$AUC_{联合}=0.71$ *vs.* $AUC_{分子标签}=0.68$ *vs.* $AUC_{TNM}=0.66$(Figure1B、2B)。

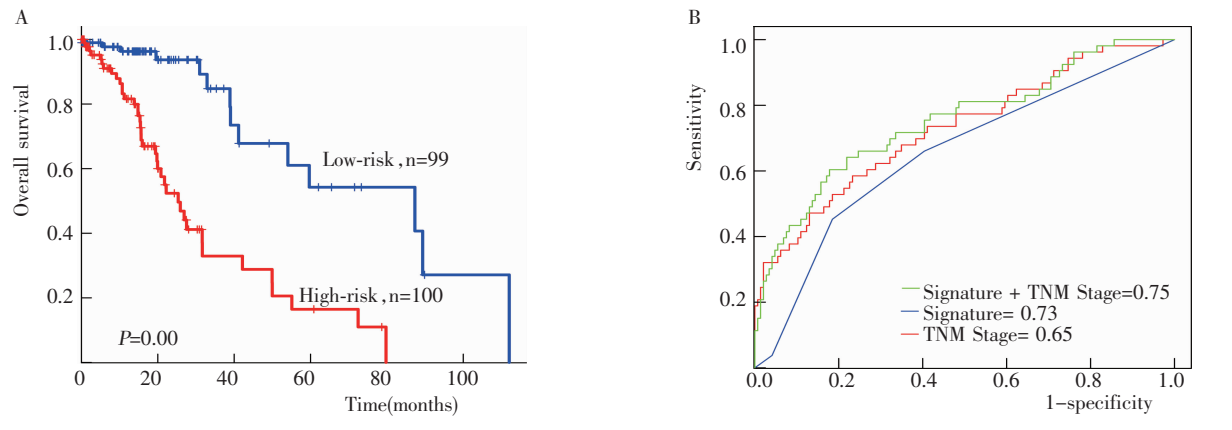
3 讨 论

肺腺癌是严重危害人类健康的恶性肿瘤之一，发病较隐匿，易转移，这就是其预后差的主要原因。临床上缺乏能预测肺腺癌转移的分子标志物及控

Table 2 Association of the PCG-lncRNA-microRNA signature with clinicopathological characteristics in lung adenocarcinoma patients (n=199)

Variables	PCG-lncRNA-microRNA signature		*P*
	Low risk*	High risk*	
Age(years)			
≤67	52	57	0.52
>67	48	42	
Sex			
Female	54	47	0.44
Male	46	52	
Histologic grade			
G_1	66	66	
G_2	4	3	0.93
G_3	30	30	
Primary tumor			
T_1	39	28	
T_2	51	52	0.15
T_3	8	12	
T_4	2	7	
Regional lymph nodes			
N_0	69	55	
N_1	12	20	0.02
N_2	15	24	
N_3	4	0	
pTNM stage			
Ⅰ	63	42	
Ⅱ	19	24	0.02
Ⅲ	14	29	
Ⅳ	4	4	

Note：*：low risk ≤median of risk score；high risk>median of risk score.

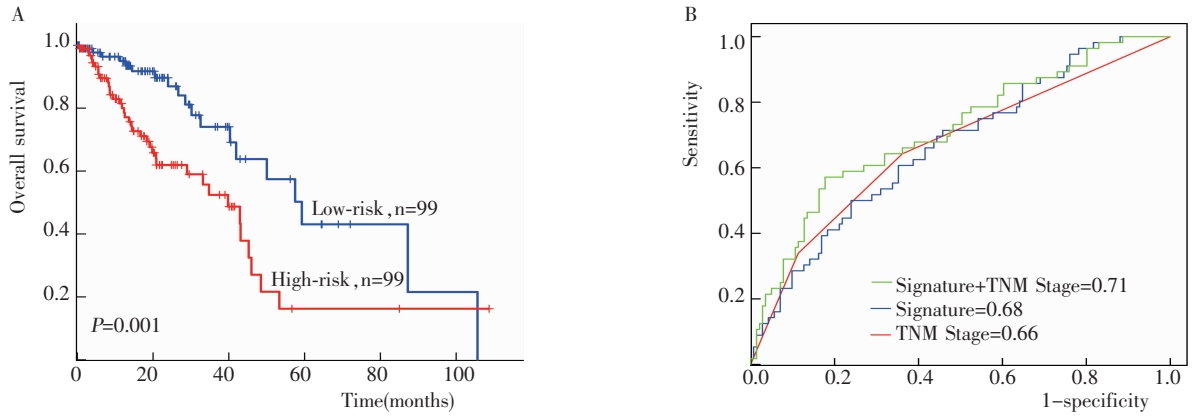


Note：A：Kaplan−Meier survival curves of patients classified into high- and low-risk groups using the PCG-lncRNA signature in the training and test dataset. *P* values were calculated by Log-rank test. B：ROC analysis was used to compare survival prediction power between the PCG-lncRNA-microRNA signature and TNM stage.

Figure 1 The PCG-lncRNA-microRNA signature predicted overall survival of patients with lung adenocarcinoma patients and compared survival prediction power between the PCG-lncRNA-microRNA signature and TNM stage in the training set(n=199)

Table 3 Univariable and multivariable Cox regression analysis of the association between PCG-lncRNA-microRNA signature and survival of lung adenocarcinoma patients (n=199)

Variables		HR	95%CI		*P*
			Lower	Upper	
Univariable analysis					
Age	>67 *vs.*≤67	1.12	0.64	1.94	0.70
Gender	Female *vs.* male	1.01	0.59	1.75	0.96
pTNM stage	Ⅳ *vs.* Ⅰ+Ⅱ+Ⅲ	1.49	1.15	1.93	0.00
PCG-lncRNA-microRNA signature	High risk *vs.* low risk	2.37	1.88	2.98	0.00
Multivariable analysis					
Age	>67 *vs.*≤67	1.23	0.69	2.20	0.48
Gender	Female *vs.* male	0.86	0.48	1.52	0.60
pTNM stage	Ⅳ *vs.* Ⅰ+Ⅱ+Ⅲ	1.41	1.05	1.90	0.02
PCG-lncRNA-microRNA signature	High risk *vs.* low risk	2.31	1.82	2.92	0.00



Note: A: Kaplan–Meier survival curves of patients classified into high- and low-risk groups using the PCG-lncRNA signature in the training and test dataset. *P* values were calculated by Log-rank test. B: ROC analysis was used to compare survival prediction power between the PCG-lncRNA-microRNA signature and TNM stage.

Figure 2 The PCG-lncRNA-microRNA signature predicted overall survival of patients with lung adenocarcinoma patients and compared survival prediction power between the PCG-lncRNA-microRNA signature and TNM stage in the test set(n=198)

制其转移的治疗靶标，因此寻找有价值的肺腺癌转移相关分子，对临床诊断、预后预测具有重要意义。

多项研究显示，PCG、lncRNA、microRNA 已经对癌症患者预后有指示作用[15~17]，但联合起来却缺乏相应研究探索，本研究通过建立一套系统的表达谱筛选流程能更好地揭示多维转录组与临床预后的关系，更加精确地反映基因表达调控关系，从而更好地指示患者生存。

本研究的不足之处在于，TCGA 数据库提供的 PCG 水平的数据可能无法完全代表基因表达，并且注释出来的 lncRNA、microRNA 较少，不能完整地呈现基因表达调控系统。但是肿瘤样本中测序检测到的可能更加能反映患者病理生理状态。

综上所述，本研究首次在肺腺癌中，联合 PCG、lncRNA、microRNA 作为临床预后分子标签，为筛选潜在效能更好地指示肺腺癌患者预后的分子标志物提供了新方向。

参考文献：

[1] Little AG, Gay EG, Gaspar LE, et al. National survey of non-small cell lung cancer in the United States: epidemiology, pathology and patterns of care[J]. Lung Cancer, 2007, 57(3): 253–260.

[2] Cancer immunotherapy: multi-pronged tumour attack [J]. Nature, 2016, 538(7626): 431.

[3] Tian X, Azpurua J, Hine C, et al. High-molecular-mass hyaluronan mediates the cancer resistance of the naked

mole rat[J]. Nature,2013,499(7458):346–349.
[4] Li X,Shi Y,Yin Z,et al. An eight-miRNA signature as a potential biomarker for predicting survival in lung adenocarcinoma[J]. J Transl Med,2014,12:159.
[5] Gao X,Wu Y,Yu W,et al. Identification of a seven-miRNA signature as prognostic biomarker for lung squamous cell carcinoma[J]. Oncotarget,2016,7(49):81670–81679.
[6] Sun Y,Hou L,Yang Y,et al. Two-gene signature improves the discriminatory power of IASLC/ATS/ERS classification to predict the survival of patients with early-stage lung adenocarcinoma[J]. Oncol Targets Ther,2016,9:4583–4591.
[7] Krzystanek M,Moldvay J,Szuts D,et al. A robust prognostic gene expression signature for early stage lung adenocarcinoma[J]. Biomark Res,2016,4:4.
[8] Sui J,Li YH,Zhang YQ,et al. Integrated analysis of long non-coding RNA associated ceRNA network reveals potential lncRNA biomarkers in human lung adenocarcinoma[J]. Int J Oncol,2016,49(5):2023–2036.
[9] Zhou M,Xu W,Yue X,et al. Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma[J]. Oncotarget,2016,7(20):29720–29738.
[10] He L,He X,Lim LP,et al. A microRNA component of the p53 tumour suppressor network[J]. Nature,2007,447(7148):1130–1134.
[11] Guo JC,Li CQ,Wang QY,et al. Protein-coding genes combined with long non-coding RNAs predict prognosis in esophageal squamous cell carcinoma patients as a novel clinical multi-dimensional signature[J]. Mol Biosyst,2016,12(11):3467–3477.
[12] Wu H,Liu J,Li W,et al. LncRNA-HOTAIR promotes TNF-alpha production in cardiomyocytes of LPS-induced sepsis mice by activating NF-kappaB pathway[J]. Biochem Biophys Res Commun,2016,471(1):240–246.
[13] Sun LL,Wu JY,Wu ZY,et al. A three-gene signature and clinical outcome in esophageal squamous cell carcinoma [J]. Int J Cancer,2015,136(6):E569–E577.
[14] Cao HH,Zhang SY,Shen JH,et al. A three-protein signature and clinical outcome in esophageal squamous cell carcinoma[J]. Oncotarget,2015,6(7):5435–5448.
[15] Li J,Chen Z,Tian L,et al. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma [J]. Gut,2014,63(11):1700–1710.
[16] Schmid F,Wang Q,Huska MR,et al. SPON2,a newly identified target gene of MACC1,drives colorectal cancer metastasis in mice and is prognostic for colorectal cancer patient survival[J]. Oncogene,2016,35(46):5942–5952.
[17] Liffers ST,Munding JB,Vogt M,et al. MicroRNA-148a is down-regulated in human pancreatic ductal adenocarcinomas and regulates cell survival by targeting CDC25B [J]. Lab Invest,2011,91(10):1472–1479.

作者/通讯作者校对文稿须知

作者/通讯作者自校拟发排校样稿,是期刊出版工作中不可缺少的重要环节,也是确保期刊质量的重要手段。特此重申,请作者/通讯作者务必按以下要求进行校对:

1. 首先全面校对全文,对编辑提出的校样稿中需特别注意校对及需补充的内容,必须予以改正或解释。

2. 所有需修改和补充的内容,均请用红笔将正确的字符书写清楚(避免使用不规范的汉字);必须改动的字符,直接在校样稿的空白处写出,所增删字数最好相符。

3. 文题、作者、单位名称、邮政编码、通讯作者等信息,务必确认无误。

4. 对正文文字(包括外文字母及大小写)、标点符号、数据、图表、计量单位、参考文献等应认真细致逐一校对;请用规范的通用药品名称(不用商品名)和医学名词,认真核查并使用标准计量单位及药物剂量。

5. 参考文献缺项的部分,应按本刊规定的著录格式进行补充。请作者务必认真核实所引用文献是否正确,并核查正文中角码是否与文后所列参考文献序号对应。

6. 校对完毕请作者/通讯作者签名,并在规定的日期内将校样稿寄回编辑部。如有要求补充的资料,也需一并寄回。

7. 由于出版周期的限制,如作者/通讯作者不能在规定时间校对寄回,请及时联系本刊编辑部说明原因,否则可能造成该文稿延期出版,或者取消刊发。

《中国肿瘤》编辑部