肝细胞癌患者生存预后相关长链非编码RNA(LncRNA)的生物信息学分析*

吴良银[a]，李文丽[b]，刘　俊[a]　(粤北人民医院 a. 检验科；b. 生殖医学中心，广东韶关　512025)

摘　要：目的　通过癌症基因图谱(the cancer genome atlas，TCGA)数据库，利用生物信息学方法构建与肝细胞癌患者生存预后相关的长链非编码RNA (LncRNA)筛选模型，建立肝细胞癌诊断和预后相关的LncRNA，为研究肝细胞癌的发生发展及预后提供新的研究思路，有望成为新的治疗靶点。**方法**　从癌症基因图谱(TCGA)数据库中下载424例肝细胞癌患者的RNA表达谱数据和相应的临床资料，其中包括374例肿瘤组织和50例正常对照。通过R语言的edgeR包，获取在正常组织和肝细胞癌组织中差异表达的LncRNA，并使用单因素Cox回归分析筛选出与预后相关的LncRNA。用lasso回归和多因素回归分析建立与肝细胞癌预后相关LncRNA模型，绘制受试者工作特征曲线，验证其模型的可靠性。**结果**　在TCGA数据库中用单因素生存分析筛选出899个与生存预后相关的LncRNA，进一步用多因素Cox回归，筛选出5个与生存预后最显著相关的LncRNA，并构建肝细胞癌预后模型，其受试者工作特征曲线的曲线下面积表明该模型对肝细胞癌的三年和五年生存率的评估有很好地准确性，可能是肝细胞癌发生发展中的关键基因，为后期临床及动物实验提供研究方向和依据。**结论**　研究确定了肝细胞癌中与生存预后显著相关的5个LncRNA，包括B3GALT5-AS，KC877982.1，MIR7-3JG，PGM5P3-AS1和TBX5-AS1($P<0.01$)。其表达特征与患者的存活风险具有较高的关联性，联合检测这5个LncRNA能较准确地预测肝细胞癌患者的三年和五年生存率。

关键词：肝细胞癌；长链非编码RNA(LncRNA)；靶基因；生物信息学
中图分类号：R735.7；R730.43　**文献标志码：**A　**文章编号：**1671-7414(2019)04-018-04
doi：10.3969/**j. issn.** 1671-7414.2019.04.005

Bioinformatics Analysis of Long-Chain Non-Coding RNA Related to Survival and Prognosis in Patients with Hepatocellular Carcinoma

WU Liang-yin[a], LI Wen-li[b], LIU Jun[a]　(a. *Department of Clinical Laboratory*; b. *Reproductive Center, Northern Guangdong People's Hospital, Guangdong Shaoguan* 512025, *China*)

Abstract: Objective　The cancer genome atlas (TCGA) database was used to construct a long-chain non-coding RNA (LncRNA) screening model related to survival and prognosis of patients with hepatocellular carcinoma (HCC) by bioinformatics method. The LncRNA related to diagnosis and prognosis of HCC was established, which provided a new research idea for the occurrence, development and prognosis of HCC and was expected to become a new therapeutic target. **Methods**　RNA expression profiles and clinical data of 424 patients with hepatocellular carcinoma (HCC) were downloaded from the cancer gene map (TCGA) database, including 374 tumors and 50 normal controls. LncRNA differentially expressed in normal tissues and hepatocellular carcinoma tissues was obtained by edgeR package of R language, and LncRNA related to prognosis was screened by single factor Cox regression analysis. LncRNA model related to prognosis of hepatocellular carcinoma was established by lasso regression and multivariate regression analysis. The working characteristic curve of subjects was drawn to verify the reliability of the model. **Results**　899 LncRNAs related to survival and prognosis were screened out by single factor survival analysis in TCGA database. Five LncRNAs most significantly related to survival and prognosis were screened by multi-factor Cox regression. A prognostic model of hepatocellular carcinoma was constructed. The area under the curve of the working characteristic curve of the subjects showed that the model was very accurate in evaluating the three-year and five-year survival rates of hepatocellular carcinoma. Sex may be a key gene in the occurrence and development of hepatocellular carcinoma, providing research direction and basis for later clinical and animal experiments. **Conclusion**　Five LncRNAs were identified to be significantly associated with survival and prognosis in HCC, including B3GALT5-AS, KC877982.1, MIR7-3JG, PGM5P3-AS1 and TBX5-AS1 ($P<0.01$). The expression characteristics of LncRNA had a high correlation with the survival risk of patients. The combined detection of these five LncRNA can accurately predict the 3-year and 5-year survival rates of patients with hepatocellular carcinoma.

Keywords: hepatocellular carcinoma; long-chain non-coding RNA; target gene; bioinformatics

肝癌作为一种高发的恶性肿瘤，发病率呈现上　升的趋势，目前肝癌已成为全球恶性肿瘤死亡原因

* 基金项目：韶关市卫生计生科研项目(Y19046)。
作者简介：吴良银(1980－)，男，硕士研究生，主管检验师，从事临床医学检验与分子诊断研究，E-mail：53537972@qq.com。
通讯作者：刘　俊，硕士研究生，从事临床医学检验与分子诊断研究，E-mail：liuyu8566@126.com。

之一，中国每年新发肝癌的病人数占全球新发肝癌近一半。原发性肝癌主要分为三型，其中肝细胞癌是最主要的一种亚型，约占原发性肝癌的85%。肝细胞癌引起的死亡率较高，特别是在发展中国家，如亚洲、非洲地区。虽然当今医学诊疗技术不断提高，联合检测甲胎蛋白(AFP)、铁蛋白(FER)和血清肿瘤特异性生长因子(TSGF)可提高肝脏恶性肿瘤诊断的准确度[1]，以及其他检测标志物，但肝癌的早期难发现，肝癌切除术后转移率和复发率较高，严重影响着肝癌病人的治疗效果和生存期。因此，在加强肝癌早期诊断研究的同时，阐述引起肝癌发生发展及生存预后的具体机制，进而探索新的干预措施，增加患者的存活率和生活质量对疾病的预防和治疗具有重大意义。

肝细胞癌病理学复杂，随着生物信息学的发展，大量的基因测序数据被上传到公共数据库中，例如癌症基因组图谱(the cancer genome atlas，TCGA)数据库，它为我们进一步研究肝细胞癌发生发展及生存预后的分子机制提供了很好的工具，以筛选出新的预后标志物和分子治疗靶点。

长链非编码RNA(LncRNA)是长度大于200个核苷酸的非编码RNA，虽然缺少有效的开放式阅读框不编码蛋白，但有着复杂的生物学功能，在生物体内发挥着非常重要的作用，也是近期研究的热点。越来越多的研究集中在肝细胞癌中涉及的致病LncRNA，这可能有利于更好地理解肝细胞癌发生和发展的分子机制。LncRNA也被证明是肝细胞癌的重要调节因子，它们在肝细胞癌发展中占有重要地位。对于LncRNA参与调控肝细胞癌不同阶段的研究也是一个令人兴奋和快速发展的领域，其中LncRNA-H19，HOTAIR和MALAT1等已经证明其表达和肝癌的发生发展及预后相关[2]。肝细胞癌是一种高度异质化的疾病，其发生发展是多基因参与，多步骤调控的过程，目前的研究认识还不足以为早期疾病的诊断和预后风险评估提供详实的证据。

本研究中提取了癌症基因组图谱(TCGA)数据库中含有临床信息的所有LncRNA表达样本，包括50个对照和374个肝癌组织，首先筛选出有差异表达的LncRNA。进一步对差异基因做单因素生存分析，筛选对预后有影响的基因，随后用这些基因做一千次lasso回归进一步筛选有显著影响的LncRNA，最后又对降维后的LncRNA做了多因素的Cox回归，最终筛选出在预后中有显著差异的LncRNA，为后期临床及动物实验提供研究方向和依据。

1 材料和方法

1.1 数据来源和处理 从癌症基因组图谱(TCGA)数据库(https://tcga-data.nci.nih.gov/tcga/)中查询肝细胞癌的LncRNA表达数据。纳入的标准是：①病理类型是肝细胞癌；②可获得病人的总体生存时间；③可获得标准化LncRNA的表达数据。筛选后，获得来自TCGA数据库中的LIHC数据集中的表达数据和相应的临床数据。在此项研究中我们根据病人的生存时间来构建预后模型，所以本研究排除了临床数据不完整或总体生存率不到1个月的标本，最终有374例肝细胞癌患者和50例对照参与了本研究。用R软件的edgeR包(http://www.bioconductor.org/packages/release/bioc/html/edgeR.html)进行差异基因的筛选，筛选标准是$|\log_2 FC| \geqslant 1$，adjust $P<0.05$，并对差异基因进行标准化处理。

1.2 方法 构建LncRNA风险模型，对于筛选出的差异表达的LncRNA，采用单因素Cox回归模型研究了LncRNA表达水平与患者整体生存的关系，鉴定标准是 $P<0.05$。为了防止结果过度拟合，进行了lasso降维处理，筛选出在进行单因素Cox回归后更具有生存相关性的LncRNA。Lasso是一种常用的高维指标回归方法，该方法通过缩小回归系数，通过施加与其大小呈正比的惩罚，对参与肝细胞癌患者预后的LncRNA进行次级选择。Lasso回归是在R软件glmnet包下进行构建的。通过变量选择和收缩，进一步筛选了与生存预后相关的LncRNA。随后，再对经过lasso回归筛选的LncRNA进行多因素Cox回归，根据所选LncRNA的个体表达水平计算了每个肝细胞癌患者的风险评分，其公式为 $\text{riskscore}=\sum_{i=1}^{n}\beta i \times \exp(Gi)$，其中 n 为纳入基因的个数，$\exp(Gi)$ 代表基因i的标准化表达，βi 代表基因i的系数。我们将风险评分的中位数设为截断值，将患者分为风险评分≥中位数的高风险组和风险评分<中位数的低风险组。

1.3 统计学分析 评估LncRNA风险模型，用R软件的survival包和timeROC包绘制受试者工作特征曲线(ROC曲线)，通过计算ROC曲线下的面积(AUC)对LncRNA风险模型进行评估。对于生存分析，我们应用Kaplan-Meier方法计算不同风险组的总体生存时间，进行对数秩检验以检查统计学显著性，使用 $P<0.05$ 确定统计学显著性，所有分析均在R软件中进行。

2 结果

差异LncRNA的鉴定：从癌症基因组图谱(TCGA)数据库中下载肝细胞癌的表达数据，包括50例对照和374例癌症组织，以及对应的临床随访信息。通过对比50例肝细胞癌正常组织和

374 例肝癌组织，我们筛选出 1 292 个表达有显著差异的 LncRNA，其中包括 1 212 个在肝癌组织中上调的 LncRNA 和 80 个下调的 LncRNA。

构建 LncRNA 的风险模型：在 1 292 个差异表达的 LncRNA 中，进一步做单因素生存分析计算每一个 LncRNA 的预后影响的显著性，其中 $P<0.05$ 的有 899 个 LncRNA。随后，又利用 lasso 回归进一步降维处理筛选出 31 个 LncRNA，进一步用多因素 Cox 回归，筛选出 5 个与生存预后最显著相关的 LncRNA，包括 B3GALT5-AS，KC877982.1，MIR7-3JG，PGM5P3-AS1 和 TBX5-AS1，其 $P<0.01$，见图 1。

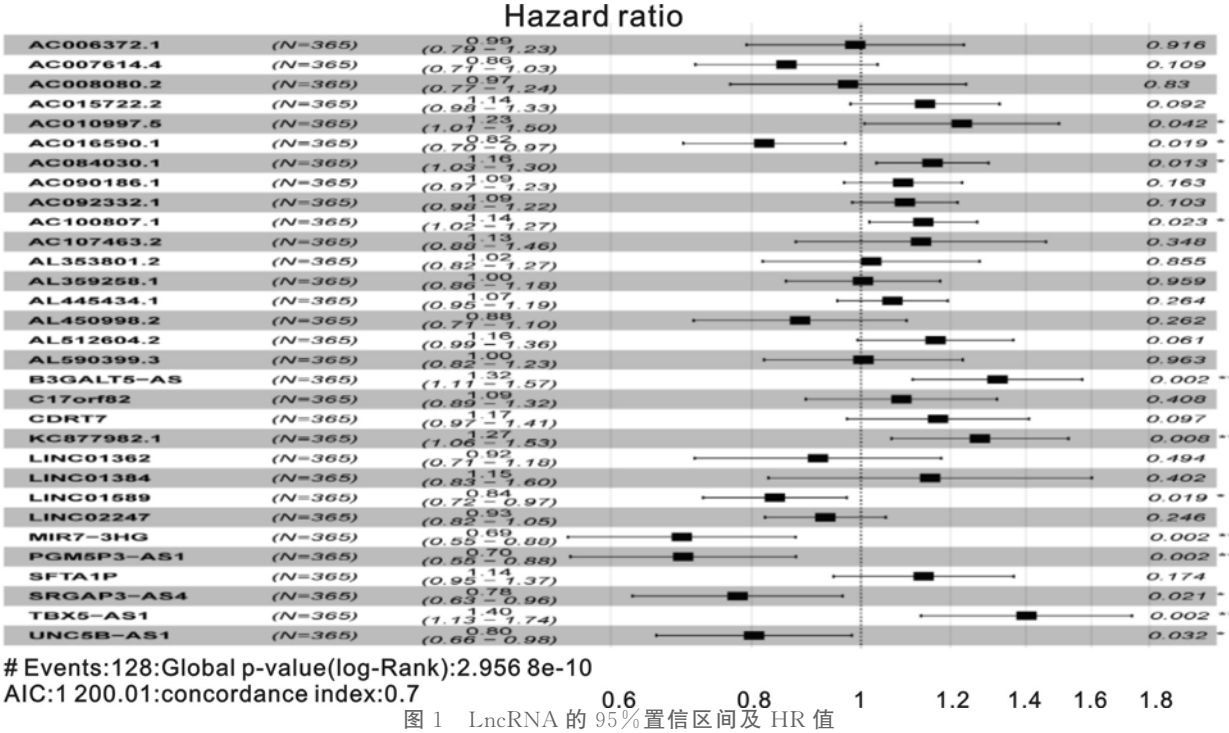


图 1　LncRNA 的 95%置信区间及 HR 值

风险模型的评估：根据 5 个 LncRNA 的多因素预后结果进行风险评分，分为低风险组和高风险组，其 K-M 生存风险曲线显示高风险组具有较差的预后，其 P 值<0.001（图 2）。同时，绘制 ROC 曲线用 AUC 面积来评价我们的预测模型，结果显示其预测三年总体生存率的曲线下面积为 0.801，其预测五年总体生存率的曲线下面积为 0.823。其结果反映出我们建立的预测模型具有较好的准确性（图 3）。说明我们构建筛选出来的 5 个与生存预后显著相关的 LncRNA（B3GALT5-AS，KC877982.1，MIR7-3JG，PGM5P3-AS1 和 TBX5-AS）表达特征与患者的存活风险具有较高的关联性，联合检测这五个 LncRNA 能较准确地预测肝细胞癌患者的三年和五年生存率。

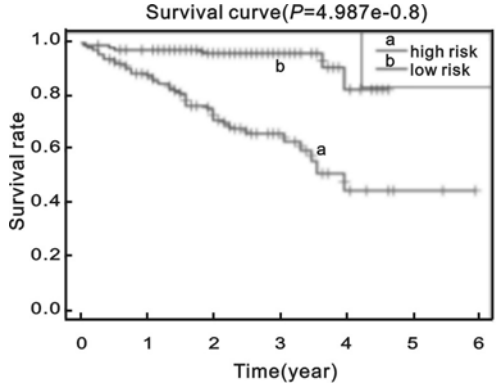
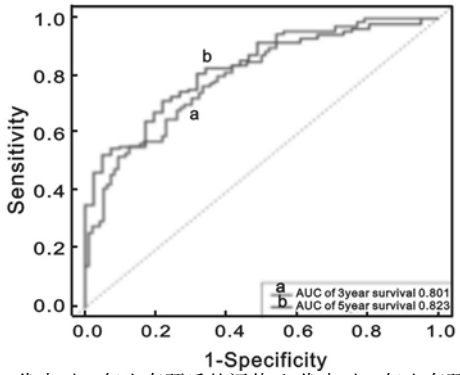


图 2　根据 5 个 LncRNA 的多因素预后结果进行分类成高低风险组的 K-M 生存风险曲线



a 代表对 3 年生存预后的评估，b 代表对 5 年生存预后的评估。

图 3　受试者特征曲线

3　**讨论**　大部分肝癌患者在确诊时已经处于中晚期，无法进行手术时，只能采取传统的化疗药物治疗，但由于肝癌具有较强的化疗抵抗性，治疗效果不明显。对于实施手术切除术的病人，肝癌的高复发率也会导致不好的预后。在过去的十几年中，已经证明 LncRNA 在癌症发生和肿瘤进展中起重要作用。一些 LncRNA 被认为在评估肝细胞癌患者的预后中是有效的，例如 UCA1 和 RP11-

466I1.1[3-4]。本研究利用生物信息学方法提取含有临床信息的LncRNA表达信息，研究肝细胞癌患者生存预后相关LncRNA，为肝细胞癌的发生发展及预后提供了新的研究思路，也希望能为肝细胞癌治疗提供新靶点。

WANG等[5]人的研究表明B3GALT5-AS通过调节miR-203介导结肠癌的上皮间质转化，通过激活B3GALT5-AS/miR-203轴可能是结肠癌肝转移的潜在治疗策略，本研究结果也表明B3GALT5-AS在肝细胞癌患者的预后中有意义。ZHANG等[6]人建立了肝细胞癌患者生存时间的风险评分系统，其研究结果表明PGM5P3-AS1与肝细胞癌的诊断和预后显著相关，这与本研究的预测结果一致。QIAO等[7]人的研究表明TBX5-AS1的表达与不吸烟女性肺癌的预后不良有关，我们的预测结果表明TBX5-AS1与肝细胞癌的不良预后相关，其说明TBX5-AS1的表达与癌症的发生发展有关系。KC877982.1和MIR7-3HG的表达也与肝细胞癌的不良预后相关，而且目前还未见报道这两个LncRNA在其它疾病中的作用。

在本研究中，选取了大样本，通过构建与生存预后相关的模型来筛选关键LncRNA，并对模型进行了评估，采用单因素Cox回归，lasso回归以及多回素Cox回归筛选出12个与生存预后显著相关的LncRNA，其中B3GALT5-AS，KC877982.1，MIR7-3HG，PGM5P3-AS1和TBX5-AS1最为显著，其$P<0.01$，选取这5个LncRNA构建了预后模型，并且还对模型的预测能力进行了评估，显示其预测三年总体生存率及其预测五年总体生存率的都有很好的准确性，为临床应用提供有力证据。本研究将这5个LncRNA组合成一个验证小组，并对肝细胞癌患者总体生存率有较准确的预测性。

肝细胞癌的发生是受表观遗传和多基因变化调控的复杂过程[8]。LncRNA AB209371的表达促进肝细胞癌的上皮间质转化，PREX2基因的体细胞突变促进了细胞增殖，并且与肝细胞癌的侵袭相关[9]。已知的几条致癌信号通路如PI3K/AKT和NF-κB都与肝细胞癌的发生发展相关[10-11]。本研究从TCGA数据库中鉴定出异常表达的关键LncRNA，构建了基于肝细胞癌生存预后的一个风险评估模型，用于预测关键LncRNA的表达与肝细胞癌患者不良预后的关系。结果表明B3GALT5-AS，KC877982.1，MIR7-3JG，PGM5P 3-AS1和TBX5-AS1表达特征与患者的存活风险具有较高的关联性，联合检测这五个LncRNA能较准确地预测肝细胞癌患者的三年和五年生存率，而且其有可能成为肝细胞癌治疗的潜在靶点。后续我们会通过临床及动物实验来进一步验证本研究结果的准确性。

参考文献：

[1] 颜丽，魏莲花，齐发梅，等. 血清TSGF，AFP，CEA和FER联合检测在肝脏恶性肿瘤诊断中的应用价值[J]. 现代检验医学杂志，2018，33(5)：24-26，141.
YAN Li，WEI Lianhua，QI Famei，et al. Significance of combined detection of TSGF，AFP，CEA and FER in the diagnosis of hepatic malignancy[J]. Journal of Modern Laboratory Medicine，2018，33(5)：24-26，141.

[2] LI Haiyan，AN Jiahui，WU Mengying，et al. LncRNA HOTAIR promotes human liver cancer stem cell malignant growth through downregulation of SETD2[J]. Oncotarget，2015，6(29)：27847-27864.

[3] ZHENG Zhikun，PANG Cui，YANG Yang，et al. Serum long noncoding RNA urothelial carcinoma-associated 1：A novel biomarker for diagnosis and prognosis of hepatocellular carcinoma[J]. J Int Med Res，2018，46(1)：348-356.

[4] ZHANG Junyong，ZHANG Di，ZHAO Qi，et al. A distinctively expressed long noncoding RNA，RP11-466I1.1，may serve as a prognostic biomarker in hepatocellular carcinoma[J]. Cancer Med，2018，7(7)：2960-2968.

[5] WANG Liang，WEI Zhewei，WU Kaiming，et al. Long noncoding RNA B3GALT5-AS1 suppresses colon cancer liver metastasis via repressing microRNA-203[J]. Aging (Albany NY)，2018，10(12)：3662-3682.

[6] ZHANG Yaqing，WANG Yong，LIU Huaidong，et al. Six genes as potential diagnosis and prognosis biomarkers for hepatocellular carcinoma through data mining[J]. J Cell Physiol，2018：1-6. https://doi.org/10.1002/JCP.27664.

[7] QIAO Fang，LI Na，LI Wei. Integrative bioinformatics analysis reveals potential long non-coding RNA biomarkers and analysis of function in non-smoking females with lung cancer[J]. Med Sci Monit，2018，24：5771-5778.

[8] LANGE N，DUFOUR JF. Changing epidemiology of HCC：how to screen and identify patients at Risk[J]. Dig Dis Sci，2019，64(4)：903-909.

[9] YANG Minghui，YEN Chiahung，CHEN Yenfu，et al. Somatic mutations of PREX2 gene in patients with hepatocellular carcinoma[J]. Sci Rep，2019，9(1)：2552.

[10] SUN Xiangjun，WANG Mingchun，ZHANG Fenghua，et al. Inhibition of NET-1 suppresses proliferation and promotes apoptosis of hepatocellular carcinoma cells by activating the PI3K/AKT signaling pathway[J]. EXP Ther Med. 2019，17(3)：2334-2340.

[11] QIAN Fuliang，HU Qingqing，TIAN Yali，et al. ING4 suppresses hepatocellular carcinoma via a NF-κB/miR-155/FOXO3a signaling axis[J]. Int J Biol Sci，2019，15(2)：369-385.

收稿日期：2019-03-28 修回日期：2019-05-22