

肿瘤防治研究

Cancer Research on Prevention and Treatment

基于TCGA数据库的胶质母细胞瘤LncRNA风险预测模型的建立

彭慧, 秦凯, 戴宇翊, 张孟贤, 郭秋云

引用本文:

彭慧, 秦凯, 戴宇, 等. 基于TCGA数据库的胶质母细胞瘤LncRNA风险预测模型的建立[J]. 肿瘤防治研究, 2019, 46(05): 417-420.

PENG Hui, QIN Kai, DAI Yuhong, et al. Establishment of LncRNA Risk Prediction Model for Glioblastoma Based on TCGA Database[J].

Zhong Liu Fang Zhi Yan Jiu, 2019, 46(05): 417-420.

在线阅读 View online: <https://doi.org/10.3971/j.issn.1000-8578.2019.19.0055>

您可能感兴趣的其他文章

Articles you may be interested in

胶质瘤中MKK7与c-Jun磷酸化的表达及其相关性

Expression of MKK7 and Phospho-c-Jun in Glioma and Their Correlation

肿瘤防治研究. 2019, 46(09): 790-795 <https://doi.org/10.3971/j.issn.1000-8578.2019.18.1780>

术前血液学炎症反应标志物在胶质瘤患者预后中的价值

Predictive Value of Preoperative Hematologic Inflammatory Markers in Prognostic of Glioma Patients

肿瘤防治研究. 2019, 46(08): 714-719 <https://doi.org/10.3971/j.issn.1000-8578.2019.19.0163>

基因芯片筛选多形性胶质母细胞瘤差异表达基因和通路

Identification of Differently Expressed Genes and Pathways in Glioblastoma Multiforme Using Microarray

肿瘤防治研究. 2018, 45(7): 441-446 <https://doi.org/10.3971/j.issn.1000-8578.2018.17.1403>

胶质母细胞瘤循环肿瘤细胞的研究进展

Research Progress in Circulating Tumor Cells of Glioblastoma

肿瘤防治研究. 2018, 45(12): 1020-1022 <https://doi.org/10.3971/j.issn.1000-8578.2018.18.0603>

基于GEO数据库的胃肠上皮化生相关基因与通路的生物信息学分析

Bioinformatics Analysis of Gastric Intestinal Metaplasia-related Genes and Pathways Based on GEO Database

肿瘤防治研究. 2018, 45(10): 768-774 <https://doi.org/10.3971/j.issn.1000-8578.2018.17.1651>



杂志官网



微信公众号

基于TCGA数据库的胶质母细胞瘤LncRNA风险预测模型的建立

彭慧, 秦凯, 戴宇翊, 张孟贤, 郭秋云

Establishment of LncRNA Risk Prediction Model for Glioblastoma Based on TCGA Database

PENG Hui, QIN Kai, DAI Yuhong, ZHANG Mengxian, GUO Qiuyun

Department of Oncology, Tongji Hospital of Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Corresponding Author: GUO Qiuyun, E-mail: Happyyun.1985@163.com

Abstract: Objective To establish a risk score model of LncRNA for the prognosis of glioblastoma patients using TCGA database. **Methods** The gene expression profiles and clinical data of glioblastoma and normal nerve tissues in TCGA database were downloaded to screen differentially-expressed LncRNA. The risk score model of LncRNA was screened and established by univariate and multivariate Cox regression models. **Results** The expression profiles of glioblastoma genes were obtained from TCGA database, including 169 glioblastoma tissues and 5 normal nerve tissues. The R software edgeR package was used for differentially-expressed gene analysis ($\log_{2}FC \geq 2$ or ≤ -2 , $FDR < 0.05$, $FDR < 0.05$). A total of 7978 differential expressed genes were obtained, of which 1643 were differential expressed lncRNAs. By univariate and multivariate Cox regression analyses, the prognostic risk model was obtained: Risk score = $0.59 \times \text{NDUFB2-AS1} - 0.41 \times \text{ZEB1-AS1} + 0.31 \times \text{AL139385.1} + 0.21 \times \text{AGAP2-AS1}$. The area under ROC curve (AUC) of the model was 0.864. Risk scores results indicated that the prognosis of patients with high score was worse than that of patients with low score. **Conclusion** The risk prediction models of NDUFB2-AS1, ZEB1-AS1, AL139385.1 and AGAP2-AS1 mentioned above could effectively predict the prognosis of glioblastoma patients and are expected to be used for clinical treatment guidance.

Key words: Glioblastoma; LncRNA; TCGA database; Cox regression model

摘要: 目的 利用TCGA数据库建立胶质母细胞瘤患者预后的LncRNA风险评分模型。**方法** 下载TCGA数据库中胶质母细胞瘤及正常神经组织的基因表达谱数据、临床相关数据, 筛选差异表达LncRNA, 采用单因素和多因素Cox风险回归模型筛选和建立LncRNA预后模型。**结果** 从TCGA数据库中得到了169份胶质母细胞瘤组织和5份正常神经组织的基因表达谱, 使用R语言edgeR包进行差异基因分析 ($\log_{2}FC \geq 2$ 或 ≤ -2 , $FDR < 0.05$) 得到差异基因7 978个, 其中差异LncRNA 1 643个。单因素Cox分析及多因素Cox回归分析得到基于4个LncRNA的多因素预后风险模型: 风险评分 = $0.59 \times \text{NDUFB2-AS1} - 0.41 \times \text{ZEB1-AS1} + 0.31 \times \text{AL139385.1} + 0.21 \times \text{AGAP2-AS1}$ 。模型的ROC曲线下面积AUC=0.864。患者风险评分结果提示高分患者预后较低分患者差。**结论** NDUFB2-AS1、ZEB1-AS1、AL139385.1和AGAP2-AS1的风险预测模型可有效预测胶质母细胞瘤患者的预后, 有望用于指导临床治疗。

关键词: 胶质母细胞瘤; LncRNA; TCGA数据库; Cox回归模型

中图分类号: R730.246

文献标识码: A

开放科学(资源服务)标识码(OSID):



0 引言

胶质瘤起源于神经外胚层分化而来的胶质细

胞, 约占颅内原发肿瘤的70%, 是中枢神经系统最常见的原发性肿瘤。世界卫生组织 (WHO) 将胶质瘤分为 I~IV 级, 级别越高, 恶性程度越高, 其中IV级胶质母细胞瘤 (Glioblastoma, GBM) 占有胶质瘤的50%, 恶性程度最高, 经过手术、放疗、化疗等综合治疗后, 生存期仍仅有12~15月^[1]。深入了解胶质母细胞瘤的分子机制是改善患者预后和识别新的预后生物标志物的关键。

收稿日期: 2019-01-15; 修回日期: 2019-02-22

基金项目: 国家自然科学基金 (81772680)

作者单位: 430030 武汉, 华中科技大学同济医学院附属同济医院肿瘤中心

通信作者: 郭秋云, E-mail: Happyyun.1985@163.com

作者简介: 彭慧 (1984-), 女, 硕士, 主治医师, 主要从事肿瘤精准治疗方面的研究

长非编码RNA (LncRNA) 是长度超过200个核苷酸的转录物, 这些核苷酸不翻译成蛋白质, 但在基因转录和mRNA翻译中起调节作用^[2-4]。越来越多的与癌症发生发展有关的LncRNA被筛选出来^[5]。世界上最大的肿瘤数据存储和分析网站肿瘤基因组图谱 (Cancer Genome Atlas, TCGA) 至今收录了26种组织类型, 共计33个癌种的11 000多名患者的肿瘤基因数据, 包括蛋白编码基因、LncRNA、microRNA以及表观遗传学数据^[6]。

本研究利用TCGA数据库研究胶质母细胞瘤与正常组织的差异表达LncRNA, 从中筛选与胶质母细胞瘤预后相关的LncRNA, 建立胶质母细胞瘤预后风险评估的LncRNA模型, 并在胶质母细胞瘤患者中进行验证。

1 资料与方法

1.1 数据来源

2018年12月20日从TCGA数据库中下载所有胶质母细胞瘤患者的基因表达谱数据及相关的临床数据。

1.2 方法

对下载的基因表达谱数据进行注释, 区分其中的蛋白编码基因及LncRNA, 使用R语言软件edgeR包对下载的数据进行数据标准化, 并筛选在胶质母细胞瘤组织和正常组织样本中差异表达的LncRNA, 筛选条件为 $\log FC \geq 2$ 或 ≤ -2 , $FDR < 0.05$ 。对差异LncRNA采用单因素Cox回归分析, 依据 $P < 0.001$ 筛选影响患者生存的差异LncRNA, 纳入多因素Cox回归分析建立LncRNA预后风险评估模型, 并计算每位患者的风险评分 (risk score, RS), 依据评分中位值将患者分为高、低风险组, 并进一步验证风险评估模型。

1.3 统计学方法

应用R3.5.2软件进行统计学分析及相应图形绘制, edgeR包筛选差异基因, Survival包进行单因素和多因素Cox比例回归模型筛选, 并建立多基因预后模型。使用Survival ROC包计算受试者工作特征 (ROC) 曲线评价模型的有效性, 并计算曲线下面积 (AUC)。

2 结果

2.1 差异LncRNA的筛选

从TCGA数据库中下载得到胶质母细胞瘤表基因表达谱矩阵, 包含肿瘤组织169份、正常组织标本5份, 共33 800个基因的表达, 其中LncRNA 14 143个。使用R语言edgeR包进行差异基因分析 ($\log FC \geq 2$ 或 ≤ -2 , $FDR < 0.05$) 得到差异基因7 978个, 其中差异LncRNA 1 643个。根据FDR值排序前10位差异表达LncRNA, 见表1。

表1 FDR排名前10的LncRNA

Table1 Top 10 LncRNA according to FDR

	logFC	logCPM	P	FDR
AC011995.2	-5.816912399	-2.320054164	7.61E-39	1.68E-36
LINC01168	-4.837945786	-1.945451207	2.77E-38	5.66E-36
SNAI3-AS1	-2.503533495	1.980730105	3.89E-38	7.60E-36
PWAR6	-3.41292033	4.45594543	1.68E-34	2.18E-32
AP003559.1	-4.235815438	-2.183154082	5.83E-33	6.91E-31
AC134312.1	-4.874822891	-0.383504201	2.86E-32	3.16E-30
AC124312.2	-4.090474071	-0.520088797	5.43E-31	5.12E-29
SLC26A4-AS1	-5.000173322	1.983245626	1.53E-30	1.35E-28
AC005329.3	-3.830420839	-1.942245203	2.18E-30	1.87E-28
PART1	-3.959052682	1.961438102	4.59E-30	3.78E-28

Notes: FC: fold change; CPM: counts per million; FDR: false discovery rate

2.2 LncRNA预后模型的建立

对筛选出的1 643个胶质母细胞瘤和正常组织的差异LncRNA进行单因素Cox回归分析, 依据 $P < 0.001$ 筛选得到影响患者生存的5个LncRNA, 分别是ZEB1-AS1、AGAP2-AS1、NDUFB2-AS1、AL139385.1和AC022148.1, 将上述5个基因纳入多因素Cox回归分析, 构建模型得到基于4个LncRNA的多因素预后风险模型: 风险得分 $= 0.59 \times \text{NDUFB2-AS1} - 0.41 \times \text{ZEB1-AS1} + 0.31 \times \text{AL139385.1} + 0.21 \times \text{AGAP2-AS1}$, 见表2、图1。

2.3 LncRNA预后模型的评估

K-M生存分析显示高风险组预后显著差于低风险组 ($P = 7.707 \times 10^{-8}$), 见图2。绘制模型的ROC曲线, 见图3, 曲线下面积AUC=0.864, 模

表2 差异LncRNA的单因素和多因素Cox回归分析

Table2 Univariate and multivariate Cox regression analyses of differentially-expressed LncRNA

Gene	Univariate regression model			Multivariate regression model		
	HR	coef	P	HR	se(coef)	P
ZEB1-AS1	0.535978433	-0.410460142	0.000270518	0.663344947	0.178876105	0.021752521
AGAP2-AS1	1.212915823	0.209167474	0.000271363	1.232651419	0.058778135	0.000372846
NDUFB2-AS1	2.023827506	0.586924983	0.000399033	1.798449641	0.191582719	0.002187210
AL139385.1	1.498844279	0.305184139	0.000626320	1.356874834	0.119789831	0.010844713
AC022148.1	0.794270837		0.000945392			

Notes: HR: hazard ratio; coef: coefficient

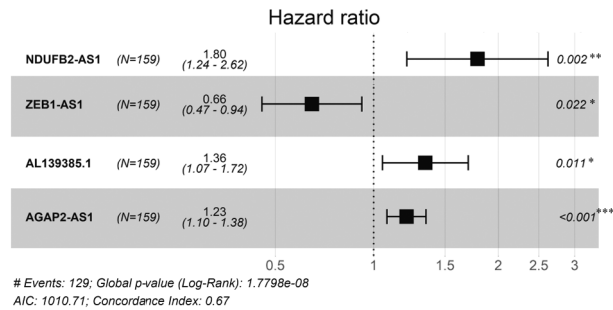


图1 多因素Cox回归模型中4个LncRNA森林图

Figure1 Forest maps of four LncRNA in multivariate Cox regression model

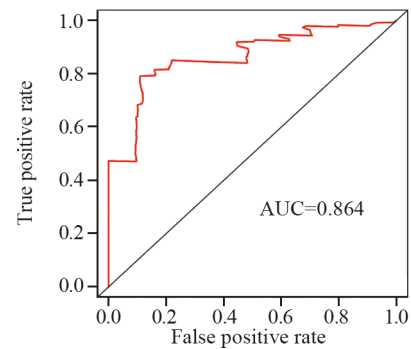


图3 多因素Cox分析模型ROC曲线

Figure3 ROC curve of multivariate Cox analysis model

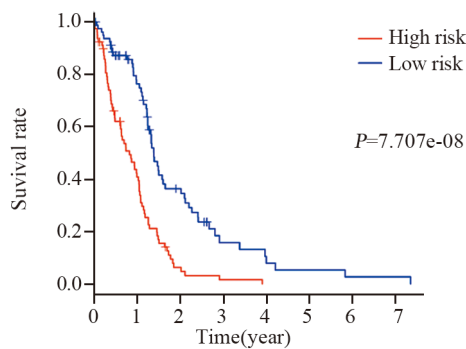


图2 高低风险组的K-M生存分析

Figure2 K-M survival analysis of high and low risk groups

型具有较好的灵敏度和特异性。多因素Cox模型中RS评分与相关LncRNA表达热图及生存时间点图,见图4,可见随着RS值得升高,患者生存时间呈缩短趋势,且死亡患者(图中红点显示)明显增多。

3 讨论

近年来的研究显示,LncRNA可以通过染色质修饰、转录和翻译过程参与基因调控,在基因调控网络中具有关键性作用。其在各种生理病理过程中,例如上皮-间充质转化、组织再生和肿瘤发生等,具有广泛作用^[5,7-9]。研究证实,不受管制的LncRNA表达谱是癌症的一个新特征,其丰度与肿瘤侵袭性和患者预后具有显著相关性^[10-11]。

本研究从TCGA数据库中胶质母细胞瘤表达谱数据和临床生存数据进行统计分析得到基于4个LncRNA的多因素预后风险模型:风险得分 $RS=0.59 \times NDUFB2-AS1 - 0.41 \times ZEB1-AS1 + 0.31 \times AL139385.1 + 0.21 \times AGAP2-AS1$ 。其中,ZEB1-AS1为保护性LncRNA,其余为危险性LncRNA。这4个LncRNA在其他肿瘤中的研究较少,Tian等^[12]研究发现AGAP2-AS1在胶质母细胞瘤中表达高于邻

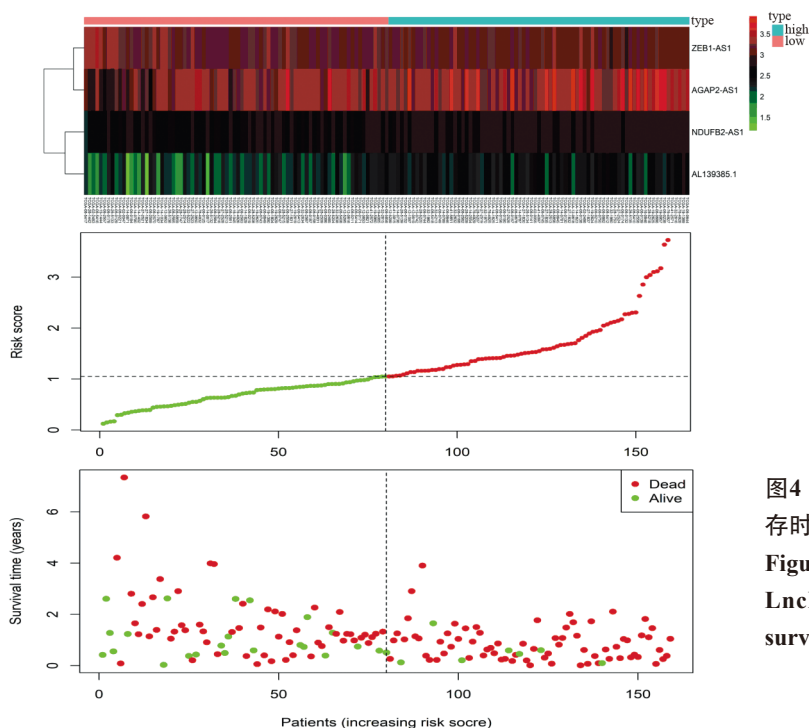


图4 风险评分与相关LncRNA表达热图及生存时间点图

Figure4 Risk score and heat map of related LncRNA expression and scattered plots of survival time

近正常脑组织,且高表达者总生存期更短。对胶质母细胞瘤细胞敲降AGAP2-AS1后,细胞的增殖侵袭能力明显受到抑制。Qi等^[13]在胃癌组织和细胞系中得到相似的结论,且研究显示转录因子SP1可以激活AGAP2-AS1的表达,AGAP2-AS1通过与LSD1和EZH2相互作用并抑制CDKN1A(P21)和E-钙黏蛋白转录而发挥致癌作用。Li等^[14]研究显示AGAP2-AS1在非小细胞肺癌中高表达,可能通过抑制肿瘤抑制因子LATS2和KLF2转录而作为癌基因发挥作用。

本研究经过TCGA胶质母细胞瘤表达谱数据库的挖掘,筛选出差异常表达LncRNA并成功构建风险预测模型,用于计算胶质母细胞瘤患者的风险评分,可以较好地反应患者的预后,且模型有较好的敏感度和特异性,为胶质母细胞瘤的临床预后判断提供帮助,并为基础研究提供更多可供选择的生物标志物。然而本研究尚存在许多不足之处,首先模型的预测能力仍需大量多中心的循证医学证据证实,其次纳入模型的LncRNA在生物体内的功能尚不明确,其在胶质母细胞瘤发生发展中起的作用尚缺乏实验证据的支持。

参考文献:

- [1] Balca-Silva J, Matias D, Carmo AD, *et al.* Cellular and molecular mechanisms of glioblastoma malignancy: Implications in resistance and therapeutic strategies[J]. *Semin Cancer Biol*, 2018, pii: S1044-579X(18)30072-5.
- [2] Trimarchi T, Bilal E, Ntziachristos P, *et al.* Genome-wide mapping and characterization of Notch-regulated long noncoding RNAs in acute leukemia[J]. *Cell*, 2014, 158(3): 593-606.
- [3] Wu H, Yang L, Chen LL. The Diversity of Long Noncoding RNAs and Their Generation[J]. *Trends Genet*, 2017, 33(8): 540-52.
- [4] Rynkeviciene R, Simiene J, Strainiene E, *et al.* Non-Coding RNAs in Glioma[J]. *Cancers (Basel)*, 2018, 11(1): pii: E17.
- [5] Lin C, Yang L. Long Noncoding RNA in Cancer: Wiring Signaling Circuitry[J]. *Trends Cell Biol*, 2018, 28(4): 287-301.
- [6] Deng M, Brägelmann J, Schultze JL, *et al.* Web-TCGA: an online platform for integrated analysis of molecular cancer data sets[J]. *BMC Bioinformatics*, 2016, 17: 72.
- [7] Qian X, Zhao J, Yeung PY, *et al.* Revealing lncRNA Structures and Interactions by Sequencing-Based Approaches[J]. *Trends Biochem Sci*, 2019, 44(1): 33-52.
- [8] Kopp F, Mendell JT. Functional Classification and Experimental Dissection of Long Noncoding RNAs[J]. *Cell*, 2018, 172(3): 393-407.
- [9] Adams BD, Parsons C, Walker L, *et al.* Targeting noncoding RNAs in disease[J]. *J Clin Invest*, 2017, 127(3): 761-71.
- [10] Sahu A, Singhal U, Chinnaiyan AM. Long noncoding RNAs in cancer: from function to translation[J]. *Trends Cancer*, 2015, 1(2): 93-109.
- [11] Huarte M. The emerging role of lncRNAs in cancer[J]. *Nat Med*, 2015, 21(11): 1253-61.
- [12] Tian Y, Zheng Y, Dong X. AGAP2-AS1 serves as an oncogenic lncRNA and prognostic biomarker in glioblastoma multiforme[J]. *J Cell Biochem*, 2019, 120(6): 9056-62.
- [13] Qi F, Liu X, Wu H, *et al.* Long noncoding AGAP2-AS1 is activated by SP1 and promotes cell proliferation and invasion in gastric cancer[J]. *J Hematol Oncol*, 2017, 10(1): 48.
- [14] Li T, Xie J, Shen C, *et al.* Upregulation of long noncoding RNA ZEB1-AS1 promotes tumor metastasis and predicts poor prognosis in hepatocellular carcinoma[J]. *Oncogene*, 2016, 35(12): 1575-84.

[编辑校对: 尤婷婷]

作者贡献:

彭 慧: 数据分析、文章撰写

秦 凯: 数据收集下载

戴宇翔: 数据整理

张孟贤: 数据分析指导

郭秋云: 数据分析