

DNA 甲基化驱动的转录表达特征作为肝癌预后预测标志物的价值

骆红波¹, 曹鹏博², 周钢桥^{1,2}

1. 贵州大学医学院, 贵阳 550025

2. 军事科学院军事医学研究院辐射医学研究所, 蛋白质组学国家重点实验室, 国家蛋白质科学中心(北京), 北京 100850

摘要: 肝细胞癌(hepatocellular carcinoma, 简称肝癌)是最常见的恶性肿瘤之一。DNA 甲基化的异常是恶性肿瘤的特征之一, 并被发现在肝癌等肿瘤的发生发展中发挥重要作用。为了能为肝癌患者提供新的临床预后预测标志物, 本研究首先采用整合组学分析策略在全基因组范围内鉴定与肝癌患者预后相关的 DNA 甲基化驱动的差异表达基因; 然后, 采用 LASSO (least absolute shrinkage and selection operator)分析建立了 10 个最优基因组合的预后预测模型。Cox 比例风险回归分析显示, 在校正临床特征参数后, 此预测模型高风险评分与患者不良预后显著相关, 表明该模型具有潜在的独立预后价值。受试者工作特征(receiver operating characteristic, ROC)曲线分析显示该风险评分模型在预测患者短期和长期预后方面优于其他已被报道的肝癌预后预测模型。基因集富集分析(gene set enrichment analysis, GSEA)表明, 高风险评分与细胞周期和 DNA 损伤修复通路相关。以上结果表明, 本研究构建了一个基于 10 个 DNA 甲基化驱动基因的预后风险评分模型, 该模型可作为肝癌患者的潜在预后生物标志物, 有助于肝癌患者的生存预后评估和治疗策略的指导。

关键词: 肝癌; 转录组; 表观基因组; 预后模型

Prognostic and predictive value of a DNA methylation-driven transcriptional signature in hepatocellular carcinoma

Hongbo Luo¹, Pengbo Cao², Gangqiao Zhou^{1,2}

1. Guizhou University School of Medicine, Guiyang 550025, China

2. State Key Lab of Proteomics, National Center for Protein Sciences (Beijing), Institute of Radiation Medicine, Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 100850, China

Abstract: Hepatocellular carcinoma (HCC) is one of the most common cancers worldwide. DNA methylation alterations are frequently observed in malignant tumours and play critical roles in the development of cancers, including HCC. To

收稿日期: 2020-05-18; 修回日期: 2020-07-10

基金项目: 国家科技重大专项艾滋病和病毒性肝炎等重大传染病防治专项项目(编号: 2018ZX10732202, 2017ZX10203205)资助[Supported by the National S&T Major Project (Nos. 2018ZX10732202, 2017ZX10203205)]

作者简介: 骆红波, 硕士研究生, 专业方向: 肝癌转录组学。E-mail: 965589073@qq.com

通讯作者: 曹鹏博, 博士, 副研究员, 研究方向: 医学遗传与基因组学。E-mail: birchcpb@163.com

周钢桥, 博士, 研究员, 研究方向: 医学遗传与基因组学。E-mail: zhougq114@126.com

DOI: 10.16288/j.ycz.20-139

网络出版时间: 2020/7/13 11:48:46

URI: <https://kns.cnki.net/kcms/detail/11.1913.R.20200710.1712.002.html>

provide novel clinical prognosis biomarkers for HCC patients, we first performed a comprehensive analysis and identified a collection of prognosis-associated genes with DNA methylation-driven expression dysregulation in HCCs. Then, we optimally established a 10-gene prognostic risk score model using the least absolute shrinkage and selection operator (LASSO) analysis. Cox's proportional hazards regression analysis revealed that the high-risk score is significantly associated with poor prognosis after being adjusted by clinical parameters, indicating its potential prognostic value. The receiver operating characteristic curve (ROC) analysis showed that this 10-gene prognostic risk score model outperformed several other publicly available models in predicting both short- and long-term prognosis. Gene set enrichment analysis revealed that the high-risk score is relevantly associated with pathways involved in cell cycle and DNA damage repair. The above results indicate that we have constructed a 10-DNA-methylation-driven-gene prognostic risk score model, which might serve as a potential prognostic biomarker for HCC patients and guide treatment decisions for patients at high risk of tumour progression.

Keywords: hepatocellular carcinoma; transcriptome; epigenome; prognostic model

肝细胞癌(hepatocellular carcinoma; 简称肝癌)是最常见的原发性肝癌,也是全球范围内与癌症相关死亡的主要原因之一^[1]。尽管现代医学的发展和多种治疗策略的结合较好地改善了肝癌患者的临床预后,但由于肝癌的高转移或高复发率,这些患者的长期预后状况仍然较差^[2]。鉴定新的与肝癌复发及生存相关的分子,能为肝癌患者的临床预后预测提供新的候选标志物。

DNA 甲基化(methylation)是最早被发现的 DNA 修饰类型之一。已有研究表明 DNA 甲基化能够引起染色质结构和 DNA 稳定性等发生改变,从而调控基因的表达^[3]。位于启动子区域的异常 DNA 甲基化通常导致抑癌基因的转录沉默或癌基因的高表达,从而促进肿瘤的进展^[4]。因此, DNA 甲基化异常在肝癌等肿瘤的发生发展中发挥重要作用^[5]。

为建立肝癌相关异常 DNA 甲基化所调控基因的预后预测模型,本研究通过肝癌组织的转录组和表观基因组的整合分析,鉴定出一系列与肝癌预后相关的 DNA 甲基化驱动的差异表达基因,并建立了一个高置信的肝癌预后预测模型,为肝癌患者的预后风险分层、预后评估及治疗策略的选择提供了新的参考指标,具有一定的潜在应用价值。

1 材料与方法

1.1 研究对象

本研究主要分为 3 个阶段:候选 DNA 甲基化驱

动的差异表达基因的发掘阶段、模型训练阶段和模型验证阶段,共包括 8 个肝癌队列(1042 个肝癌临床组织样本;表 1,图 1A)。对于涉及患者临床资料的分析,去除生存时间未知和 TNM (tumour node metastasis)分期未知的患者。

发掘阶段:从基因表达汇编数据库(gene expression omnibus, GEO)下载获得 2 个肝癌组织及癌旁配对组织的 DNA 甲基化数据集(GEO: GSE89852 和 GSE54503),以及 2 个肝癌组织及癌旁配对组织的转录组 RNA 测序数据集(GEO: SRP069212 和 SRP118972)。

模型建立阶段:从癌症基因组图谱-肝细胞癌项目(the cancer genome atlas-liver hepatocellular carcinoma, TCGA-LIHC)下载获得 level 3 转录组 RNA 测序数据和 DNA 甲基化数据以及临床信息数据。该队列用于验证发掘阶段所鉴定出的 DNA 甲基化驱动的差异表达基因及建立和训练预后预测模型。

模型验证阶段:包括 3 个独立的肝癌队列。从国际癌症基因组联盟(international cancer genome consortium, ICGC)获得的日本肝癌项目(liver cancer-RIKEN of JP project)的转录组基因表达谱数据集作为第 1 个模型验证队列,从 GEO 下载的 GSE76427 以及 GSE84005 表达谱数据集分别作为第 2 和第 3 个模型验证队列。

1.2 差异表达基因及差异甲基化基因的鉴定

针对 GEO 下载的 RNA 测序数据,采用 STAR 软件^[6]以 hg19 基因组为参考进行比对,然后采用

表 1 本研究中涉及的所有肝癌队列

Table 1 Samples and datasets used in this study

研究队列	数据集	样本量	数据类型	数据来源
发掘队列	SRP069212	20 例配对的癌和癌旁组织	mRNA 表达	GEO
	SRP118972	12 例癌组织样本和 8 例癌旁组织	mRNA 表达	GEO
	GSE89852	33 例配对的癌和癌旁组织	DNA 甲基化	GEO
	GSE54503	66 例配对的癌和癌旁组织	DNA 甲基化	GEO
模型训练队列	TCGA-LIHC	371 例癌组织和 50 例癌旁组织	mRNA 表达和 DNA 甲基化	TCGA
	ICGC-LIRI-JP	203 例癌组织	mRNA 表达	ICGC
模型验证队列	GSE76427	115 例癌组织	基因表达	GEO
	GSE84005	37 例癌组织	基因表达	GEO

GEO: 基因表达汇编数据库(Gene Expression Omnibus); ICGC-LIRI-JP: 国际癌症基因组联盟日本肝癌项目(international cancer genome consortium liver cancer-RIKEN of JP project); TCGA-LIHC: 癌症基因组图谱-肝细胞癌项目(the cancer genome atlas-liver hepatocellular carcinoma)。

HTseq-count 软件^[7]进行基因表达的定量。采用 DESeq2 软件^[8]进行差异表达分析, 差异表达基因(differently expressed gene, DEG)的筛选标准为 $|\log_2[\text{fold change}]| \geq 1.5$ 且错误发现率(false discovery rate, FDR) <0.05 。

采用 ChAMP 软件^[9]处理 DNA 甲基化数据, 通过 limma 软件包^[10]对肝癌和癌旁间每个 CpG 位点甲基化水平(β)进行 t 检验。差异甲基化的 CpG 位点定义为: $|\Delta\beta| > 0.25$ 且 $\text{FDR} < 0.05$ 。对于匹配到多个 DNA 甲基化探针的基因, 选择甲基化水平倍数变化显著的作为代表^[11]。

利用维恩图获得上述发掘的 2 个候选基因集合(即差异表达基因和差异甲基化基因)间的重叠基因, 即可得到肝癌组织中相对癌旁组织“高甲基化-低表达”和“低甲基化-高表达”的基因, 作为后续 LASSO (least absolute shrinkage and selection operator)回归分析的候选基因。

1.3 预后预测评分模型的建立

基于发掘的候选基因, 采用两个步骤来构建预后预测评分模型。首先, 通过单因素 Cox 回归分析获得与模型训练队列中患者总体生存期(overall survival)相关的候选基因($P < 0.05$), 然后通过具有 Cox 比例风险模型的 LASSO 回归根据最佳惩罚因子(λ)对候选基因进行降维筛选^[12]。LASSO 回归是一种根据 λ 缩小回归系数的方法, 一些系数可能会缩小

为 0, 然后从模型中删除。将具有 Cox 比例风险模型的 LASSO 回归应用于模型训练队列, 然后基于 10 倍交叉验证(10-fold cross validation)确定最佳惩罚因子, 进而估计模型系数。如果系数为 0, 则删除对应的基因, 然后将保留系数不为 0 的基因用于构建预后预测评分模型。每位患者的风险分数可以通过以下公式计算得出:

$$\text{风险评分(Risk score)} = (\text{EXP}_{\text{gene1}} \times \beta_{\text{gene1}}) + (\text{EXP}_{\text{gene2}} \times \beta_{\text{gene2}}) + \dots + (\text{EXP}_{\text{genen}} \times \beta_{\text{genen}})$$

其中 EXP_{gene} 代表某一特定基因的表达水平, β_{gene} 代表 LASSO 回归系数。

以患者总体风险评分的中位数为界, 将队列中的患者划分为高风险组和低风险组。采用单因素和多因素 Cox 回归分析计算该风险模型对总体生存期的风险比例(hazard ratios, HR)。采用时间依赖的受试者工作特征曲线(time-dependent receiver operating characteristic curve, time-dependent ROC)评估该风险模型在预测患者预后方面的性能^[13]。

1.4 列线图的建立

用多因素 Cox 回归构建列线图(nomogram)来量化患者的风险评估, 进而预测患者的临床预后。分析中纳入模型训练队列和模型验证队列共有的临床参数, 包括: 性别、年龄、肿瘤分期和基于预测模型计算的风险评分。列线图通过 rms 工具包构建, 同时采用校准曲线(calibration curves)用来评价该模

型在预测患者生存的准确性^[14]。

1.5 基因集富集分析

基于 KEGG 基因集,采用基因富集分析(gene set enrichment analysis, GSEA)对高、低风险组(基于中位数分组)进行富集分析,鉴定与生存风险相关的信号通路,参考基因集为 MsigDB (6.2 版本),样本置换检验 1000 次, FDR<0.05 作为差异显著性的评价标准。

1.6 统计分析

采用 R 3.4.4 软件进行所有的统计学分析。采用 survival 工具包绘制 Kaplan-Meier 生存曲线,采用 Log-rank 检验计算生存率差异显著性。采用 χ^2 检验计算组间患者临床特征分布差异。对于所有的假设检验, $P<0.05$ 被认为有统计学意义。

2 结果与分析

2.1 鉴定肝癌组织中 DNA 甲基化驱动的差异表达基因

为了鉴定肝癌中 DNA 甲基化变化驱动的差异表达基因,本研究在发掘队列肝癌组织和配对癌旁组织的数据集中进行了全基因组层面的差异表达基因及差异 DNA 甲基化基因的整合分析。通过差异表达分析共鉴定到 547 个基因在肝癌组织中上调表达和 291 个基因在肝癌组织中下调表达(图 1B);通过差异甲基化分析共鉴定到 788 个基因在肝癌组织中高甲基化和 5126 个基因在肝癌组织中低甲基化(图 1B)。通过取这两组基因集合的交集,在发掘队列中共鉴定出 197 个“低甲基化-高表达”基因和 18 个“高甲基化-低表达”基因(图 1B)。在这 215 个基因中,有 163 个基因(占 75.8%)在模型训练队列中被成功地重复,其中包括 153 个“低甲基化-高表达”基因和 10 个“高甲基化-低表达”基因。热图显示这 163 个候选基因在发掘队列和模型训练队列中 DNA 甲基化水平和表达水平高度一致(图 1C)。

2.2 采用 LASSO 回归分析建立由 10 个基因组合的肝癌预后预测评分模型

针对这 163 个基因,首先在模型训练队列中进

行了单因素 Cox 比例风险回归分析,共鉴定出 51 个与肝癌患者预后显著相关的候选基因(均 $P<0.05$)。随后, LASSO 回归进一步降维筛选出其中 10 个有显著性的候选基因组合用于风险模型的构建(MAEL、PRC1、TTC39A、SFN、LPL、STC2、PBK、CDCA8、MYO18B 和 MAPT; 表 2, 图 2A)。除此之外,在模型训练队列中对这 10 个基因的基因组变异频率进行分析,结果显示它们均表现出低频($\leq 10\%$)的点突变或拷贝数变异(表 2),表明表观遗传的改变可能是导致这些基因在肝癌中发生表达异常的主要原因。

基于这 10 个基因的表达水平及其对应的 LASSO 回归分析的系数,在模型训练队列中建立了预后评分模型: 风险评分(Risk score) = $(-0.1766 \times \text{EXP}_{\text{MAEL}}) + (0.08869 \times \text{EXP}_{\text{PRC1}}) + (0.01347 \times \text{EXP}_{\text{TTC39A}}) + (0.001652 \times \text{EXP}_{\text{SFN}}) + (0.003126 \times \text{EXP}_{\text{LPL}}) + (0.03600 \times \text{EXP}_{\text{STC2}}) + (0.06524 \times \text{EXP}_{\text{PBK}}) + (0.1194 \times \text{EXP}_{\text{CDCA8}}) + (0.1932 \times \text{EXP}_{\text{MYO18B}}) + (0.2597 \times \text{EXP}_{\text{MAPT}})$ 。根据风险评分的中位数,将肝癌患者分为高风险组和低风险组(表 3)。单因素 Cox 比例风险回归分析显示,高风险评分与患者较短的总体生存率显著相关($P<0.0001$; 图 2B)。进一步采用多因素 Cox 分析校正患者的肿瘤分期,结果显示高风险评分仍然是肝癌患者总体生存期的独立风险因素($P<0.0001$; 表 3)。此外,该风险评分预测患者 1 年、2 年、3 年、4 年和 5 年的生存率对应曲线下面积 AUC(areas under the curve)分别为 0.82、0.74、0.72、0.68 和 0.66 (图 2B),提示该模型能较好预测模型队列中患者预后情况。

2.3 基于 10 个基因的肝癌预后预测评分模型在验证队列中表现出良好的预测能力

随后,在另外 3 个独立肝癌队列中验证此预测评分模型的性能。在验证队列 1(包含 203 例肝癌患者)中,高风险评分与肝癌患者较短的总体生存期显著相关($P=0.0015$, 图 2C);在校正患者的肿瘤分期后,高风险评分仍然显示为肝癌患者总体生存期的独立风险因素($P=0.025$; 表 3); ROC 分析显示,虽然预测模型在验证队列 1 中的预测能力略差于发掘队列,但该模型仍具有良好的预测性能,预测模型在此队列中预测患者 1 年、2 年、3 年和 4 年生存率的 AUC 分别为 0.67、0.75、0.68 和 0.63 (图 2C)。在验证队列 2(包含 115 例肝癌患者)中,虽然没有临床

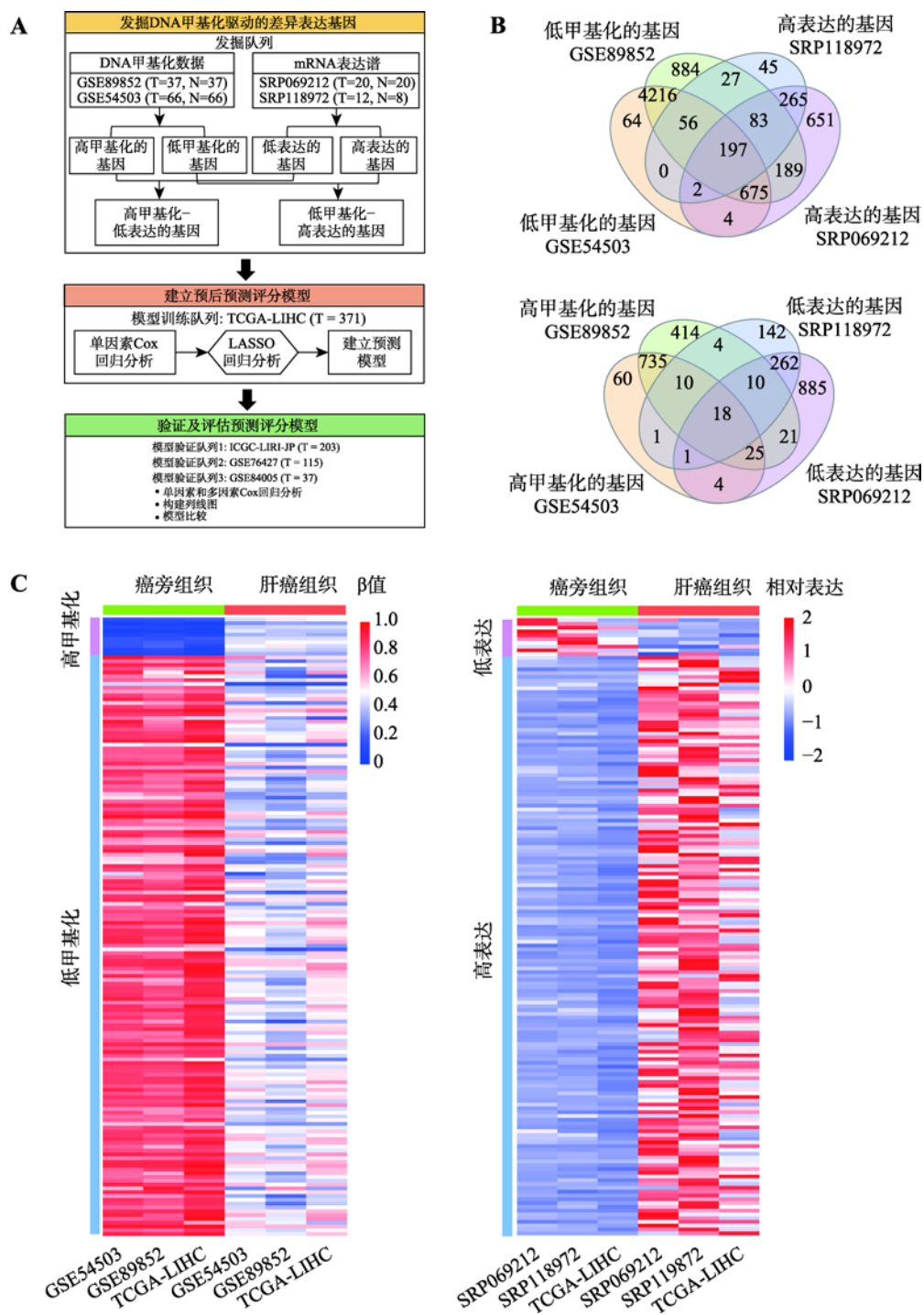


图 1 肝癌中 DNA 甲基化驱动的差异表达基因的鉴定

Fig. 1 Identification of the DNA methylation-driven differentially expressed genes in HCC

A: 研究技术路线图。主要包括候选 DNA 甲基化驱动的差异表达基因的发掘阶段、模型训练阶段和模型验证和评估阶段。B: 在发掘队列中鉴定出的 DNA 甲基化驱动的差异表达基因数量。上图为鉴定出的“低甲基化-高表达”基因数量, 下图为鉴定出“高甲基化-低表达”基因数量。C: 在发掘队列和模型训练队列中 DNA 甲基化驱动的差异表达基因的热图。左图为基因 DNA 甲基化水平热图(平均甲基化水平), 右图为基因表达热图(平均表达水平)。ICGC-LIRI-JP: 国际癌症基因组联盟日本肝癌项目(international cancer genome consortium liver cancer-RIKEN of JP project); N: 肝癌癌旁组织数量; T: 肝癌组织数量; TCGA-LIHC: 癌症基因组图谱-肝细胞癌项目(the cancer genome atlas-liver hepatocellular carcinoma); β : 基因的 DNA 甲基化水平。

表 2 10 个最优基因的 Cox 比例风险回归分析结果、LASSO 回归系数和基因组变异频率

Table 2 The results of Cox proportional hazard model analysis, LASSO regression coefficients and genetic alteration of the 10 optimal genes

基因	HR (95% CI)	P 值	LASSO 系数	基因组变异频率(%)
CDCA8	2.21 (1.54~3.01)	<0.0001	0.1194	0.0
PRC1	1.85 (1.29~2.54)	0.0005	0.08869	0.3
MAPT	1.72 (1.21~2.46)	0.0021	0.2597	1.7
SFN	1.72 (1.21~2.45)	0.0021	0.001652	0.3
STC2	1.73 (1.22~2.43)	0.0021	0.03600	0.8
MYO18B	1.69 (1.19~2.37)	0.0031	0.1932	4.0
PBK	1.67 (1.17~2.35)	0.0036	0.06524	6.0
MAEL	1.55 (1.09~2.17)	0.013	-0.1766	10.0
TTC39A	1.55 (1.09~2.17)	0.013	-0.01347	1.4
LPL	1.55 (1.10~2.18)	0.014	0.003126	7.0

基因组变异频率数据来源于 cBioportal 数据库(<https://www.cbioportal.org/>) 癌症基因组图谱-肝细胞癌项目, 包括点突变频率及拷贝数变异频率。CI: 置信区间(confidence interval); HR: 风险比例(hazard ratio)。

特征参数表现出显著的预后价值(表 3), 但高风险评分也与肝癌患者较短的总体生存期显著相关($P=0.0014$; 图 2C); 风险评分预测 1 年、2 年、3 年、4 年和 5 年生存率的 AUC 分别为 0.63、0.60、0.63、0.69 和 0.69 (图 2C)。在验证队列 3 (包含 37 例肝癌患者)中, 虽然没有临床特征参数表现出显著的预后价值, 但是风险评分仍然与肝癌患者的总体生存期显著相关($P=0.0016$; 图 2C) (表 3); 且其预测肝癌患者 1 年、2 年、3 年和 4 年生存率的 AUC 分别达到 0.79、0.71、0.66 和 0.57 (图 2C)。综上, 本研究建立的预测评分模型可以显著地将肝癌患者分为预后高风险组和低风险组。

2.4 整合的临床参数与预后预测评分模型建立列线图

为了进一步建立可应用于预测肝癌患者总体生存率预测的直接定量方法, 在模型训练队列中, 将预测模型评分和临床特征参数进行多因素 Cox 比例风险回归分析构建列线图(图 3A)。列线图显示, 相比较其他临床特征参数, 预后预测评分模型贡献了最大的风险值(范围为 0~100) (图 3A), 提示此预测模型在所有列线图变量中的作用最为显著。由于本研究中有 2 个验证队列生存超过 4 年的样本例数较少, 为了避免结果的偏倚, 校正曲线图只计算 3 年

生存期预测的准确性。结果可见, 通过列线图构建的模型能够较好地预测肝癌患者的 3 年生存状态(图 3B)。

2.5 基于 10 个基因的肝癌预后预测评分模型优于其他模型

为了进一步评估此模型的预测性能, 本研究比较了该预测模型与上述建立的列线图、肿瘤分期系统以及 3 个已被报道的肝癌预后模型[Zheng 等^[15]的 4 基因模型(*SPINK1*、*TXNRD1*、*LCAT*、*PZP*)、Yang 等^[16]的 3 基因模型(*SPP2*、*CDC37LI*、*ECHDC2*)和 Long 等^[17]的 2 基因模型(*SPP1*、*LCAT*)]的预测性能。有趣的是, 无论是在模型训练队列还是在 3 个验证队列中, 本研究建立的 10 个基因的预后预测模型与列线图对患者总体生存期的预测性能大致相同(图 4)。经对比发现, 此 10 个基因的预后预测模型和列线图在模型训练队列中比其他 3 个已被报道的模型和肿瘤分期系统表现出更好的预测性能(图 4A)。在 3 个验证队列中, 本研究的模型预测 1 年总体生存期(验证队列 1~3 的 AUC 分别为 0.67、0.63 和 0.79), 3 年总体生存期(验证队列 1~3 的 AUC 分别为 0.68、0.63 和 0.66)和 5 年总体生存期(验证队列 2 的 AUC 为 0.69)的结果表明 10 个基因的预后预测模型在预测短期和长期生存期方面均表现出良好的性能(图 4B)。

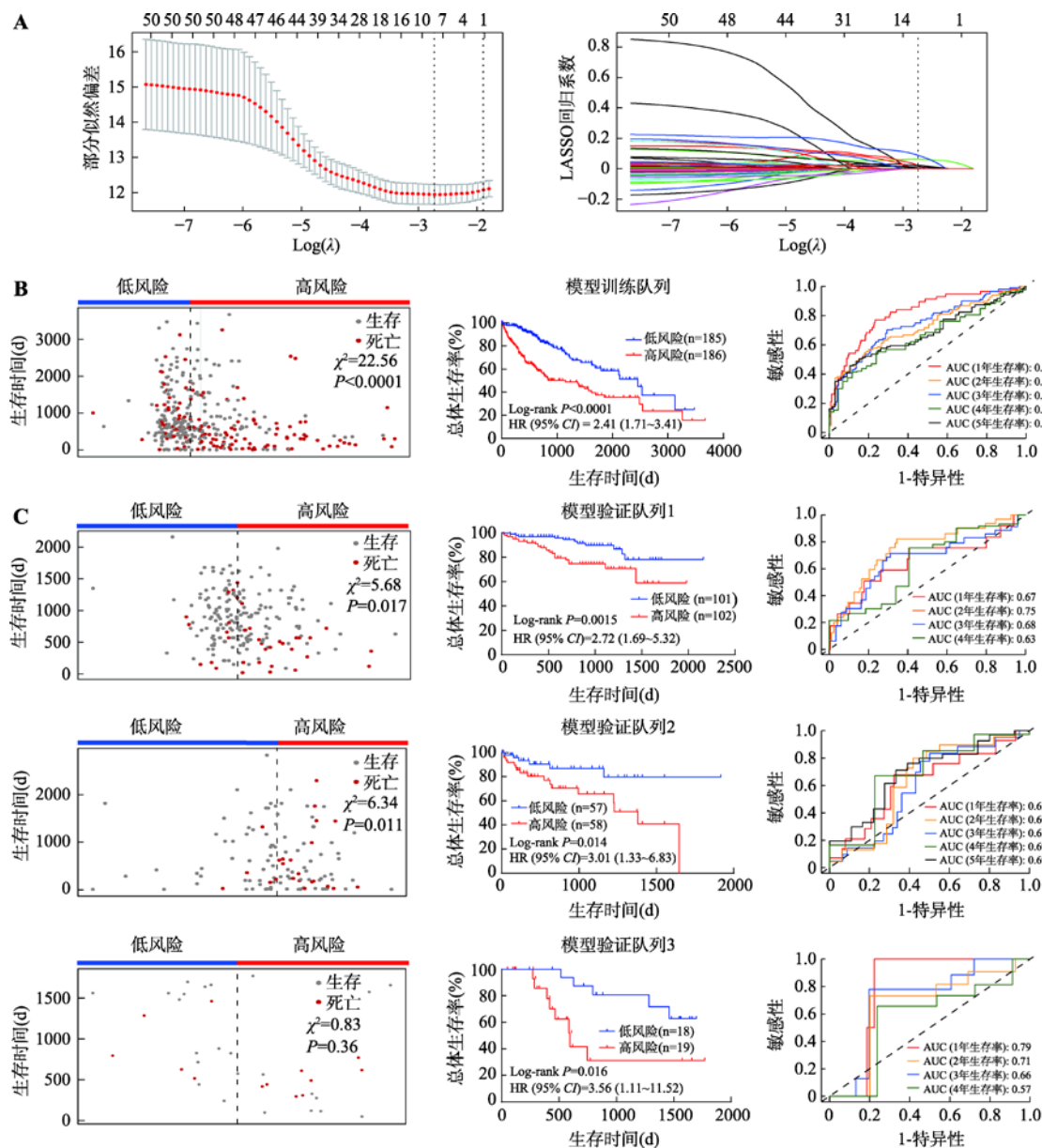


图2 建立和验证10个基因的预后评分模型

Fig. 2 Construction and validation of the 10-gene prognostic risk score model

A: 使用 LASSO 回归分析和 10 倍交叉验证构建预后预测评分模型。左图为基于最小原则(minimum criteria)采用 10 倍交叉验证对 LASSO 模型进行调参, 通过 LASSO 回归交叉验证计算的部分似然偏差(partial likelihood deviance)被绘制为 $\log(\lambda)$ 的函数。y 轴表示部分似然偏差, x 轴表示 $\log(\lambda)$, 沿 x 轴上方的数字表示预测变量的平均数量, 红点表示具有给定 λ 的每个模型的平均偏差值, 穿过红点的竖线表示偏差的上限和下限, 垂直虚线分别表示最小误差的 λ 值和最大 λ 值。右图为 51 个预后基因的 LASSO 系数分布, 垂直虚线表示采用 10 倍交叉验证选取的基因数, 当基因数为 10 时, 部分似然偏差为最小值, 对应最小 λ 值。B: 基于 LASSO 系数和基因表达在模型训练队列建立预后模型。左图为模型训练队列中肝癌患者的风险评分及生存时间的散点图, 中间图为模型训练队列中不同风险评分组(中位数分组)患者的生存曲线图, 右图为采用 ROC 曲线分析评估预测模型对训练队列中患者生存率的预测性能。C: 在模型验证队列中验证预后模型。左图为模型验证队列中肝癌患者的风险评分及生存时间的散点图, 中间图为模型验证队列中不同风险评分组(中位数分组)患者的生存曲线图, 右图为 ROC 曲线分析评估预测模型对验证队列中患者生存率的预测性能。由于验证队列 1 和 3 中患者达到 5 年生存的数量较少, 所以并未对患者 5 年生存率进行 ROC 分析。卡方检验(χ^2)用于评价组间患者生存分布差异; 组间生存差异采用 Log-rank 方法进行比较; ROC 分析用于评估模型预测性能。AUC: 曲线下面积(area under curve); CI: 置信区间(confidence interval); HR: 风险比例(hazard ratios); LASSO: 最小绝对值收敛和选择算子(least absolute shrinkage and selection operator); ROC: 受试者工作特征曲线(receiver operating characteristic curve)。

表 3 肝癌患者总体生存期相关变量的单因素和多因素生存分析
Table 3 Univariate and multivariate survival analyses of variables associated with overall survival in HCC patients

临床特征	模型训练队列(n=371)			模型验证队列 1 (n=203)			模型验证队列 2 (n=115)			模型验证队列 3 (n=37)		
	单因素分析		多因素分析	单因素分析		多因素分析	单因素分析		多因素分析	单因素分析		多因素分析
	HR (95% CI)	P	HR (95% CI)	HR (95% CI)	P	HR (95% CI)	HR (95% CI)	P	HR (95% CI)	HR (95% CI)	P	HR (95% CI)
年龄(岁)												
≤60	1.23	0.24	-	0.93	0.87	-	1.29	0.37	-	1.21	0.083	-
>60	(0.87~1.73)			(0.42~2.09)			(0.56~2.94)			(0.31~4.71)		
性别												
女	1.22	0.76	-	1.72	0.13	-	1.53	0.35	-	3.70	0.17	-
男	(0.85~1.75)			(0.75~3.85)			(0.45~5.22)			(1.02~13.32)		
TNM 分期												
I / II	2.91	<0.001	2.05	0.0002	2.81	0.0015	1.72	0.19	2.24	0.55	-	1.51
III / IV	(1.88~4.50)		(1.40~3.02)		(1.38~5.71)		(0.76~3.90)		(0.79~6.32)			(0.41~5.51)
风险评分												
低	2.41	<0.001	2.33	<0.001	2.72	0.0015	2.32	0.025	3.01	0.014	3.01	0.014
高	(1.71~3.41)		(1.56~3.50)		(1.69~5.32)		(1.11~4.85)		(1.33~6.83)		(1.33~6.83)	(1.11~11.52)

患者按风险分数中位数被分为高风险组和低风险组。-：没有进行多因素 Cox 风险比例模型分析；CI：置信区间(confidence interval)；HR：风险比例(hazard ratio)；TNM：肿瘤淋巴结转移(tumor node metastasis)。

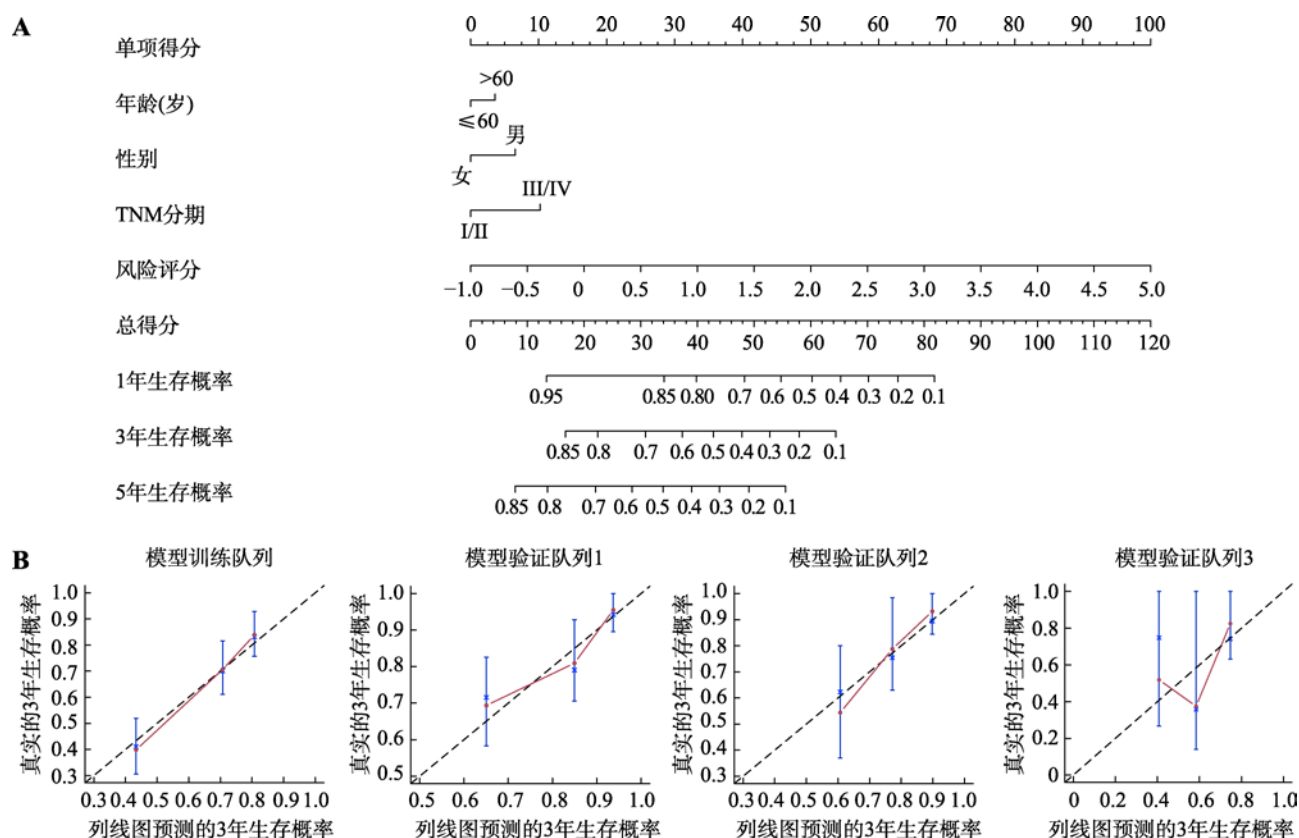


图3 基于列线图评估肝癌患者的预后

Fig. 3 The nomogram for predicting the overall survival of HCC patients

A: 列线图预测模型训练队列中患者1年、3年和5年的生存情况; B: 校正曲线描绘列线图预测患者3年生存与实际情况之间的一致性。粉红色实线表示列线图的预测性能, 45°斜线表示一种理想的校正模型。TNM: 肿瘤淋巴结转移(tumor node metastasis)。

综上所述, 本研究建立的预后评分模型在预测肝癌患者总体生存期方面显示出与整合临床特征参数的列线图相似的能力, 且优于传统的肿瘤分期系统和其他已被报道的模型。

2.6 不同的风险评分组具有生物学差异

采用基因集富集分析(gene set enrichment analysis, GSEA)探索与高风险评分或低风险评分相关的生物学信号通路(图 5A), 结果显示高风险组显著富集于与 DNA 复制、细胞周期和 DNA 修复相关的信号通路(图 5B), 表明高风险评分反映了增殖信号和 DNA 修复信号传导的异常, 这些通路的异常已经被报道是肿瘤发生发展的驱动力^[18]。相反地, 低风险组显著富集于与补体和凝血级联、脂肪酸、药物和丙酸酯代谢等肝代谢途径(图 5B), 提示低风险评分的患者保留了相对正常的肝功能, 这可以保护肝癌患者免受各种药物引起的持续药物毒性。

3 讨论

缺乏有效和可靠的预后生物标志物或预测模型仍然是改善肝癌患者临床结局的主要挑战。已有研究表明, 表观遗传异常和基因组异常变化是导致肝癌发生发展的重要原因^[19]。因此, 迫切需要有效和可靠的表观遗传生物标志物作为肝癌患者预后预测指标和治疗靶标。本研究基于肝癌组织的转录组与表观基因组, 采用整合组学的分析策略, 发现了一系列与肝癌预后相关的 DNA 甲基化驱动的差异表达基因, 并建立了一个高置信的肝癌预后预测模型, 为肝癌患者的生存预后评估提供了候选参考标准, 将有助于指导临床治疗策略的选择。

生物标志物不仅可以作为肿瘤患者预后预测工具, 而且对测量治疗反应, 监测肿瘤复发以及指导临床决策具有重要意义。近来, 已经提出了几个基于 DNA 甲基化驱动的基因建立的肝癌预后模型, 包

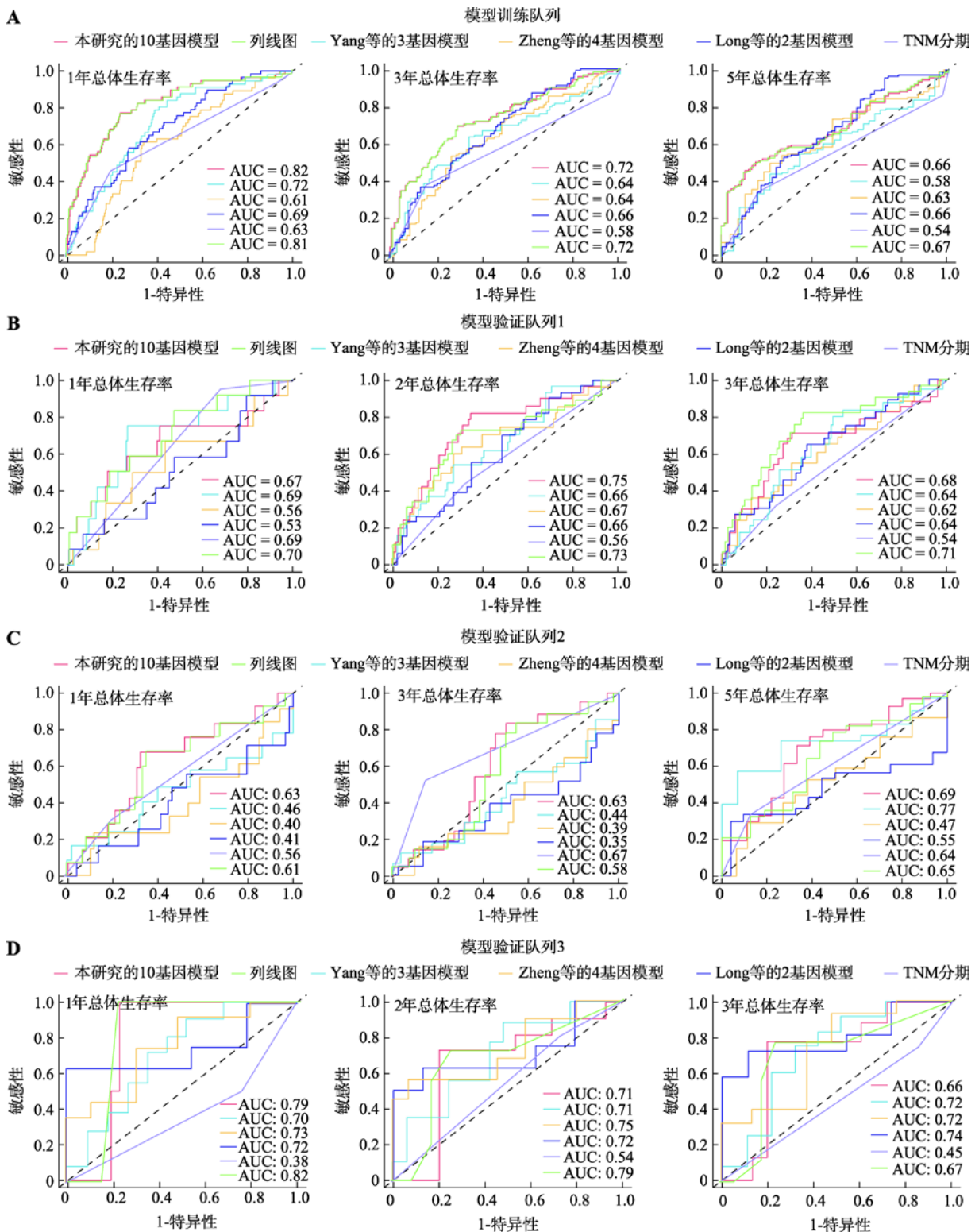


图 4 基于 10 个基因的肝癌预后预测评分模型的预测性能优于其他模型

Fig. 4 Performance comparison of our 10-gene prognostic risk score model with the other models

A: 模型训练队列中预测性能的比较; B: 模型验证队列中预测性能的比较。采用 ROC 曲线分析比较本研究的 10 个基因预测模型与其他 5 个模型(列线图模型、TNM 阶段、Zheng 等^[15]的 4 基因模型、Yang 等^[16]的 3 基因模型和 Long 等^[17]的 2 基因模型)的预测性能。

AUC: 曲线下面积(area under curve); ROC: 受试者工作特征曲线(receiver operating characteristic curve)。

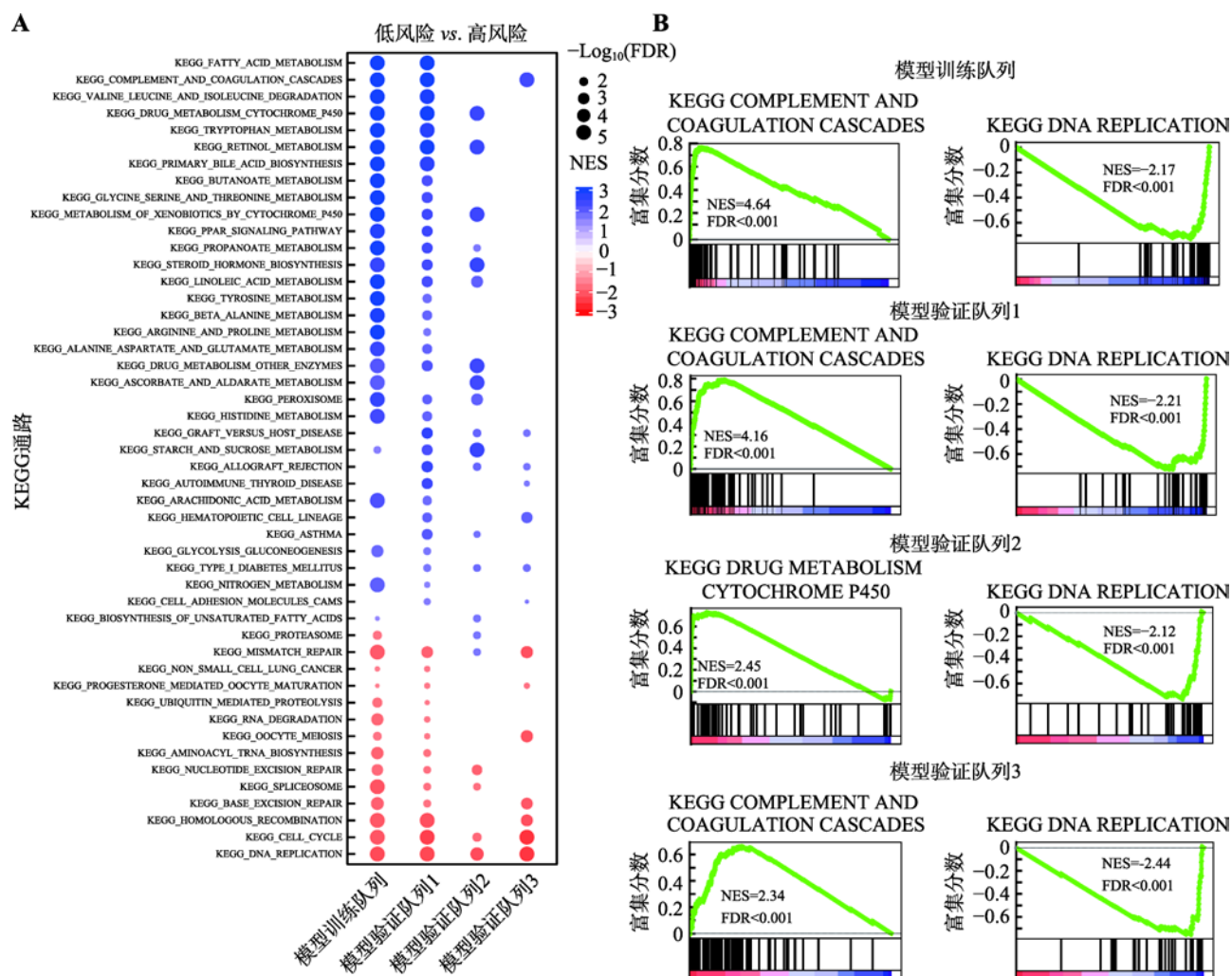


图 5 模型训练队列和验证队列中的基因集富集分析

Fig. 5 Gene set enrichment analysis in model training cohort and validation cohort

A: 在模型训练队列和验证队列中的基因 GSEA 结果展示。蓝色和红色点分别表示在低风险组和高风险组(中位数分组)中显著富集的基因集, 点的大小表示为 P 值, 基因集至少在 2 个队列中显著富集($FDR < 0.05$)才被绘制。B: 在模型训练队列和验证队列中最显著富集基因集的 GSEA 图。FDR: 错误发现率(false discovery rate); GSEA: 基因集富集分析(gene set enrichment analysis); NES: 标准化富集分数(normalized enrichment score)。

括 Zheng 等^[15]的 4 基因模型、Yang 等^[16]的 3 基因模型和 Long 等^[17]的 2 基因模型。但是, 这些模型大多数是在小型队列中建立或缺乏独立队列验证。本研究使用大规模的模型训练队列(即 TCGA-LIHC, 371 例肝癌组织样本)建立了 10 个基因的预后预测评分模型, 并在两个公共的独立大规模验证队列(即 ICGC-LIRI-JP, 203 例肝癌组织样本; GSE76427, 115 例肝癌组织样本)中进一步验证了该模型。与那些已报道的预测模型和传统的 TNM 分期系统相比, 本研究的 10 个基因预后预测评分模型显示出更优的预测能力。此外, 本研究的 10 个基因预后预测评

分模型的预测能力与整合有风险评分模型和临床特征的列线图预测能力相似。因此, 本研究建立的 10 个基因预后预测评分模型为预测肝癌患者的预后提供一种简单而准确的方法。

肝癌是一种高异质性的肿瘤, 探索与肝癌进展有关的失调基因可能有助于改善治疗策略和预后。在本研究的预后预测评分模型中, 这 10 个特征基因在肝癌组织中均呈现低 DNA 甲基化引起的高表达。GSEA 分析表明, 高风险评分与 DNA 复制、细胞增殖和 DNA 修复等通路改变有关。在这 10 个基因中, 其中 8 个已被报道在肝癌组织中高表达。在

这 8 个基因中, *MALE*、*LPL*、*MYO18B* 和 *CDCA8* 已被报道与 DNA 损伤响应和细胞周期检查点密切相关^[20~22]; 同时, *PRC1*、*STC2*、*SFN* 和 *PBK* 被发现参与调控多条与肿瘤进展相关的激酶通路^[23~25]。其他两个基因 *MAPT* 和 *TTC39A* 尚未有与肿瘤相关的研究报道。但是, *MAPT* 作为微管相关蛋白可以参与细胞的迁移^[26]; *TTC39A* 可以通过泛素化降解胆固醇受体 *LXR* 参与调节高密度脂蛋白的代谢^[27]。因此, 本研究首次鉴定发现 *MAPT* 和 *TTC39A* 可以作为肝癌患者的预后标志物, 它们可能在肝癌发生发展中发挥重要作用, 具有成为新候选靶标的潜力。

综上所述, 本研究通过表观基因组和转录组整合分析建立的预后预测评分模型为肝癌患者预后评估提供了一种简单而准确的方法, 为患者生存的预测和治疗策略的选择提供指导。

参考文献(References):

- [1] Villanueva A. Hepatocellular carcinoma. *N Engl J Med*, 2019, 380(15): 1450–1462. [DOI]
- [2] Samonakis DN, Kouroumalis EA. Systemic treatment for hepatocellular carcinoma: still unmet expectations. *World J Hepatol*, 2017, 9(2): 80–90. [DOI]
- [3] Zhang JW, Xu Q, Li GL. Epigenetics in the genesis and development of cancers. *Hereditas(Beijing)*, 2019, 41(7): 567–581.
张竞文, 续倩, 李国亮. 癌症发生发展中的表观遗传学研究. *遗传*, 2019, 41(7): 567–581. [DOI]
- [4] Sun LY, Li XY, Sun ZW. Progress of epigenetics and its therapeutic application in hepatocellular carcinoma. *Hereditas(Beijing)*, 2015, 37(6): 517–527.
孙凌云, 李星逾, 孙志为. 原发性肝癌的表观遗传学及其治疗. *遗传*, 2015, 37(06): 517–527. [DOI]
- [5] Tan AC, Jimeno A, Lin SH, Wheelhouse J, Chan F, Solomon A, Rajeshkumar NV, Rubio-Viqueira B, Hidalgo M. Characterizing DNA methylation patterns in pancreatic cancer genome. *Mol Oncol*, 2009, 3(5–6): 425–438. [DOI]
- [6] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 2013, 29(1): 15–21. [DOI]
- [7] Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods*, 2015, 12(4): 357–360. [DOI]
- [8] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, 2014, 15(12): 550. [DOI]
- [9] Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S. ChAMP: 450k chip analysis methylation pipeline. *Bioinformatics*, 2014, 30(3): 428–430. [DOI]
- [10] Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, Shi W, Smyth GK. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*, 2015, 43(7): e47. [DOI]
- [11] Mah WC, Thurnherr T, Chow PKH, Chung AYF, Ooi LLPJ, Toh HC, Teh BT, Sauntharajah Y, Lee CGL. Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. *PLoS One*, 2014, 9(8): e104158. [DOI]
- [12] Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, 2010, 33(1): 1–22. [DOI]
- [13] Blanche P, Dartigues JF, Jacqmin-Gadda H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*, 2013, 32(30): 5381–5397. [DOI]
- [14] Steyerberg EWS, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*, 2014, 35(29): 1925–1931. [DOI]
- [15] Zheng YJ, Liu YL, Zhao SF, Zheng ZT, Shen CY, An L, Yuan YL. Large-scale analysis reveals a novel risk score to predict overall survival in hepatocellular carcinoma. *Cancer Manag Res*, 2018, (10): 6079–6096. [DOI]
- [16] Yang Y, Lu Q, Shao XJ, Mo BH, Nie XQ, Liu W, Chen XH, Tang Y, Deng YC, Yan J. Development of a three-gene prognostic signature for hepatitis b virus associated hepatocellular carcinoma based on integrated transcriptomic analysis. *J Cancer*, 2018, 9(11): 1989–2002. [DOI]
- [17] Long JY, Chen PP, Lin JZ, Bai Y, Yang X, Bian J, Lin Y, Wang DX, Yang XB, Zheng YC, Sang XT, Zhao HT. DNA methylation-driven genes for constructing diagnostic, prognostic, and recurrence models for hepatocellular carcinoma. *Theranostics*, 2019, 9(24): 7251–7267. [DOI]
- [18] Torgovnick A, Schumacher B. DNA repair mechanisms in cancer development and therapy. *Front Genet*, 2015, 6: 157. [DOI]
- [19] Feitelson MA. Parallel epigenetic and genetic changes in the pathogenesis of hepatitis virus-associated hepatocellular

- carcinoma. *Cancer Lett*, 2006, 239(1): 10–20. [DOI]
- [20] Liu LL, Dai YD, Chen JN, Zeng TT, Li Y, Chen LL, Zhu YH, Li JC, Li Y, Ma S, Xie D, Yuan YF, Guan XY. Maelstrom promotes hepatocellular carcinoma metastasis by inducing epithelial-mesenchymal transition by way of Akt/GSK-3 β /Snail signaling. *Hepatology*, 2014, 59(2): 531–543. [DOI]
- [21] Zhang ZY, Zhu JF, Huang YS, Li WB, Cheng HQ. MYO18B promotes hepatocellular carcinoma progression by activating PI3K/AKT/mTOR signaling pathway. *Diagn Pathol*, 2018, 13(1): 85. [DOI]
- [22] Cao D, Song XH, Che L, Li XL, Pilo MG, Vidili G, Porcu A, Solinas A, Cigliano A, Pes GM, Ribback S, Dombrowski F, Chen X, Li L, Calvisi DF. Both de novo synthesized and exogenous fatty acids support the growth of hepatocellular carcinoma cells. *Liver Int*, 2017, 37(1): 80–89. [DOI]
- [23] Liu P, Atkinson SJ, Akbareian SE, Zhou ZG, Munsterberg A, Robinson SD, Bao YP. Sulforaphane exerts anti-angiogenesis effects against hepatocellular carcinoma through inhibition of STAT3/HIF-1 α /VEGF signalling. *Sci Rep*, 2017, 7(1): 12651. [DOI]
- [24] Wang HX, Wu KJ, Sun Y, Li YD, Wu MY, Qiao Q, Wei YJ, Han ZG, Cai B. STC2 is upregulated in hepatocellular carcinoma and promotes cell proliferation and migration *in vitro*. *BMB Rep*, 2012, 45(11): 629–634. [DOI]
- [25] Chen JX, Rajasekaran M, Xia HP, Zhang XQ, Kong SN, Sekar K, Seshachalam VP, Deivasigamani A, Goh BKP, Ooi LL, Hong WJ, Hui KM. The microtubule-associated protein PRC1 promotes early recurrence of hepatocellular carcinoma in association with the Wnt/ β -catenin signalling pathway. *Gut*, 2016, 65(9): 1522–1534. [DOI]
- [26] Desai A, Mitchison TJ. Microtubule polymerization dynamics. *Annu Rev Cell Dev Biol*, 1997, 13: 83–117. [DOI]
- [27] Safran M, Dalah I, Alexander J, Rosen N, Stein TI, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D. GeneCards Version 3: the human gene integrator. *Database (Oxford)*, 2010, 2010:baq020. [DOI]

(责任编辑: 方向东)