参赛队号：

2021年（第七届）全国大学生统计建模大赛

参赛学校：	哈尔滨医科大学
论文题目：	**联合DNA甲基化和基因表达谱构建肺腺癌早期预后风险评估模型**
参赛队员：	吕晨，周新羽，林孔选
指导老师：	张春龙

联合 DNA 甲基化和基因表达谱构建肺腺癌早期预后风险评估模型

目录

表格和插图清单

摘要

肺腺癌是最常见的非小细胞肺癌类型。其临床治疗普遍依据肿瘤分期与分级。然而仅依赖患者病理检查得到的分期结果设计肿瘤治疗方案，疗效往往不尽人意。本研究从早期肺腺癌患者的临床分期出发，通过整合其基因组与表观组信息，将同一分期不同患者的肿瘤预后风险等级细化。旨在构建肺腺癌早期预后风险评估模型，以辅助临床做出更加准确的诊断以及合理化用药指导。为给不同分期患者提供特异的预后风险评估，我们设计平行实验，基于高通量表达谱数据以及重建的子通路，分析预测出肺腺癌的风险子通路活性得分，并构建子通路水平风险打分模型。

我们识别出基因表达谱中的差异表达基因（DEG）以及 DNA 甲基化谱中的差异甲基化基因(DMG)，依据基因交集构建差异甲基化与差异基因表达的关联，再利用皮尔森相关系数计算相关性，找到与自身甲基化水平显著负相关的疾病风险基因。将我们识别出的风险基因映射到构建好的 PPI 网络，利用 Random Walk 算法进行基因扩散，为网络中的基因打分。通过 SubpathwayMiner 方法，我们识别出富集的子通路，并通过随机扰动子通路内部的基因，筛选出显著的子通路。对于筛选到的子通路，我们利用 ssGSEA 方法对其进行活性打分。继而为建立子通路活性与样本临床分期的关联，我们使用 Cox 回归方法构建出子通路水平风险打分模型，并根据训练集数据计算出风险打分 cutoff 值，使用验证集验证该模型的有效性。

首先，我们对肺腺癌 I 期的样本进行模型有效性验证。依据 cutoff 值将验证集分为高风险与低风险两类，结合样本生存时间，我们画出了两组样本的生存曲线，观察到低风险组样本生存曲线显著高于高风险组，且具有统计学显著性。我们又对 stageII 期的模型进行了显著性验证，结果进一步佐证了我们模型的有效性。以上结果表明，本次研究构建的肺腺癌预后评估模型，可以有效地划分早

期肺腺癌同一分期不同风险的患者，在一定程度上能够辅助医生的诊断以及合理化用药。

关键词：肺腺癌；多组学数据；临床预后；子通路预后标志

Abstract

Lung adenocarcinoma is the most common non-small cell lung cancer, and its clinical treatment is generally based on tumor stage and grade. However, it is often unsatisfactory to design treatment strategies according to the staging results of pathological examination. Our study set out from the clinical stages of patients with lung adenocarcinoma, compared the transcriptome and epigenome differences between the cancer and normal groups, and integrated the data to detail the risk grade of tumor prognosis of different patients in stageI and stageII. Our study is aim to establish a risk assessment model for early prognosis for lung adenocarcinoma, so as to assist the clinic in making more accurate diagnosis and rationalizing medication guidance.

In order to provide a early prognosis risk assessment model for patients at stageI and stageII, we designed parallel experiments to scored the activities of risk subpathways in stageI and stageII , which was based on high-throughput expression profile data and reconstructed subpathways. Futhermore, we constructed a risk scoring model at the level of the subpathways.

We identified differentially expressed genes (DEG) in gene expression profiles and differentially methylated genes (DMG) in DNA methylation profiles. Next, we constructed associations between differential methylation and differential gene expression based on the gene

intersection, and then calculated the correlations using Pearson correlation coefficients.After that,we found risk genes that were significantly correlated with their methylation levels. The risk genes we identified were mapped to the constructed PPI network, and gene diffusion was performed using the Random Walk algorithm to score the genes in the network. We ranked the gene scores and selected the high-scoring genes for further analysis.With the use of subpathwayMiner method, we identified the enriched subpathways and screened the significant subpathways by randomly perturbing the genes inside the subpathways.For the screened subpathways, we scored their activities using the ssGSEA method. Subsequently, to associate subpathway activity and clinical staging of the samples, we constructed a subpathway level risk scoring model using the Cox regression method and calculated a cutoff values which can distinguish the risk of stages, and validated the model using the validation set.

First, we validated the model of stage I lung Adenocarcinoma. According to the cutoff value, the validation set was divided into two groups: high-risk Group and low-risk Group. Combined with the survival time of the samples, the survival curves of the two groups were drawn.We observed that the survival curves of the low-risk group were significantly higher than those of the high-risk Group.We also carried on the remarkable verification separately to the stageII model, the result further confirmed our model validity. These results suggest that

the stage-specific prognostic assessment model developed in this study can effectively classify patients with different risks at the same stage of Lung Adenocarcinoma, and to some extent assist doctors in diagnosis and rational use of drugs.

Keywords:

lung adenocarcinoma; multiomics; clinical prognosis; prognostic markers of subpathway

一、问题描述

（一）数据新动能的统计测度

随着数据产业的蓬勃发展，我们迎来了生物数据大爆发时代，大量测序数据呈井喷式增长，为生命科学的研究带来了更多的契机。生命科学的研究最终要反馈到临床，将研究成果运用到患者身上。但是这些复杂的测序数据对于临床医生和患者来说是难以理解的，我们需要构建一种统计测度，将人类基因组等复杂数据中的有效信息提取出来。如何才能找到生物大数据里隐藏的统计测度？如何挖掘人类“DNA 密码”，找到治疗人类疾病的钥匙？

通过对各种测序数据的研究以及统计学方法的运用，从样本表达谱中找出显著差异的标志物，构建模型，目的在于基于统计测度给出让人明白的定量结论。在这个模型中，样本特征符合形成总体定量结论的要求，同时能够体现样本的个体差异，也就是构建的模型既具有普适性，同时对于不同样本，又具有特异性。

（二）肺腺癌治疗相关背景

非小细胞肺癌是世界范围内一种常见的恶性肿瘤，约占肺癌总数的 85%[1]。非小细胞肺癌包括肺腺癌，肺鳞癌和大细胞癌，其中肺腺癌约占到非小细胞肺癌的百分之五十左右。因肺腺癌早期无特异性症状，与一般呼吸系统疾病症状相似，经常被患者忽视，很难实现早期诊断，并且由于现阶段缺乏有效的治疗手段，肺腺癌已成为目前死亡率最高的癌症类型之一。临床常常基于肺癌患者不同 TNM 阶段选择相应的治疗手段，而研究人员们也开始着眼于疾病更深层次的探索，投入更大的精力为寻找有助于推断预后风险的临床特征与生物分子标记[2-6]。在目前的临床实践中，依据肺腺癌分型和患者临床表现制定患者治疗方案依旧是肺腺癌治疗的首要策略[7]。因此，进一步探讨肺腺癌预后的生物学机制，建立早期风险评估系统，识别关键的治疗靶点，是改善肺腺癌患者预后的关键。

（三）当前肺腺癌治疗的局限性

对于 I 期非小细胞肺癌的治疗标准是治疗性切除术，当前 NCCN（肿瘤学临床治疗指南）不建议对于 IA 期肺腺癌 R0 术后患者使用化疗，但是研究表明大约有百分之二十到百分之四十的 IA 期非小细胞肺癌患者在术后 5 年内发生了复发[8-10]。如今，尽管癌症诊断技术和分子治疗手段取得快速进展，但肺腺癌患者的 5 年生存率仍较低[11]。因此，对 I 期患者进行预后风险评估进而选取适当有效的临床治疗方式极其重要，根据预测的预后评分对患者进行风险分级，可以为风险较高的患者定制辅助治疗和检测成像[12]。同样，对于 II 期的患者进行预后风险评估也能在一定程度上预知并及时干预治疗，防止肿瘤的进一步恶化。本研究对肺腺癌进行分析，目的在于识别肺腺癌早期预后相关的子通路，构建通路水平的预后风险评估模型。

（四）多组学与肺腺癌

由于肿瘤的分子变化先于临床变化，研究人员们希望能够找到关键的分子生物标记物，能够实现较为准确地预测肺腺癌患者的预后情况以及癌症的复发，并在一定程度上实现个性化治疗。DNA 甲基化属于表观组范畴，是哺乳动物基因组中的一种重要的表观遗传修饰，经常发生在 CpG 岛，是基因和 microRNA 表达调控[13]中的一个重要机制。DNA 甲基化异常在肺腺癌等癌症中经常发生[14, 15]，并且，这种异常的程度会随着肺腺癌的进展不断积累[16]。由于 DNA 甲基化具有稳定并且容易被定性、定量检测的特点，与突变、拷贝数变异(CNV)和基因表达或 microRNA 表达状态相比[17-19]，DNA 甲基化被认为对肺腺癌早期检测起到更好的标记效果。同时，基因表达谱的变化能直接的反应出癌旁与肺腺癌样本之间的差异，前期研究者常用基因表达谱挖掘与疾病相关的标志物，并围绕该标志物进行一系列的生物学研究。

尽管基于单一组学数据进行的肿瘤研究已经发现了许多致癌的标志性因子，但生命的发展机理是一个多层次、多水平和多功能的复杂结构系统，仅通过单一层面的研究往往不够充分，所以现阶段研究人员们更倾向于整合多组学数据，如基因组、表观组、蛋白质组、代谢组等进行更全面研究。DNA 甲基化水平能影响个体的基因表达模式、基因组的稳定性以及机体的生长发育过程，还可以从多个角度造成肿瘤发展、转移以及恶化。本研究通过整合转录组和表观组信息，联合基因表达谱和 DNA 甲基化数据，对肺腺癌进行研究，更加深入地挖掘转录组和表观组中与肺腺癌预后风险相关的标志物。

二．本研究的整体流程

本研究通过对 stageI 和 stageII 的患者分别构建肺腺癌预后风险模型，对构建模型的有效性进行验证，并探究不同分期肺腺癌预后风险模型在子通路水平上的差异（图 4）。

本研究的整体流程主要围绕以下三方面展开，首先，初步筛选肺腺癌 stageI 和 stageII 期标志物。其次，识别肺腺癌风险子通路并刻画活性状态。最后，构建并评估 stage 特异的肺腺癌早期预后风险模型。

i)从基因组学 RNA-seq 数据和表观组学甲基化数据出发，初步筛选肺腺癌早期 stage 特异的标志物，该步骤筛选出的标志物在基因组层面和表观组层面都是显著差异的，并认为这些标志物是潜在的肺腺癌风险基因。

ii)将肺腺癌风险基因映射到全局 PPI 网络，先在基因水平上进行分析，给风险等级更高的基因赋予更高的权重，然后利用已知的通路功能注释信息，重构子通路，紧接着筛选出显著的子通路，利用 ssGSEA 方法以及样本表达谱和 DNA 甲基化谱数据，得到各个样本在不同子通路中的活性得分值。

iii)利用样本子通路活性谱以及患者临床生存数据，构建 stage 特异的肺腺癌早期预后风险模型。

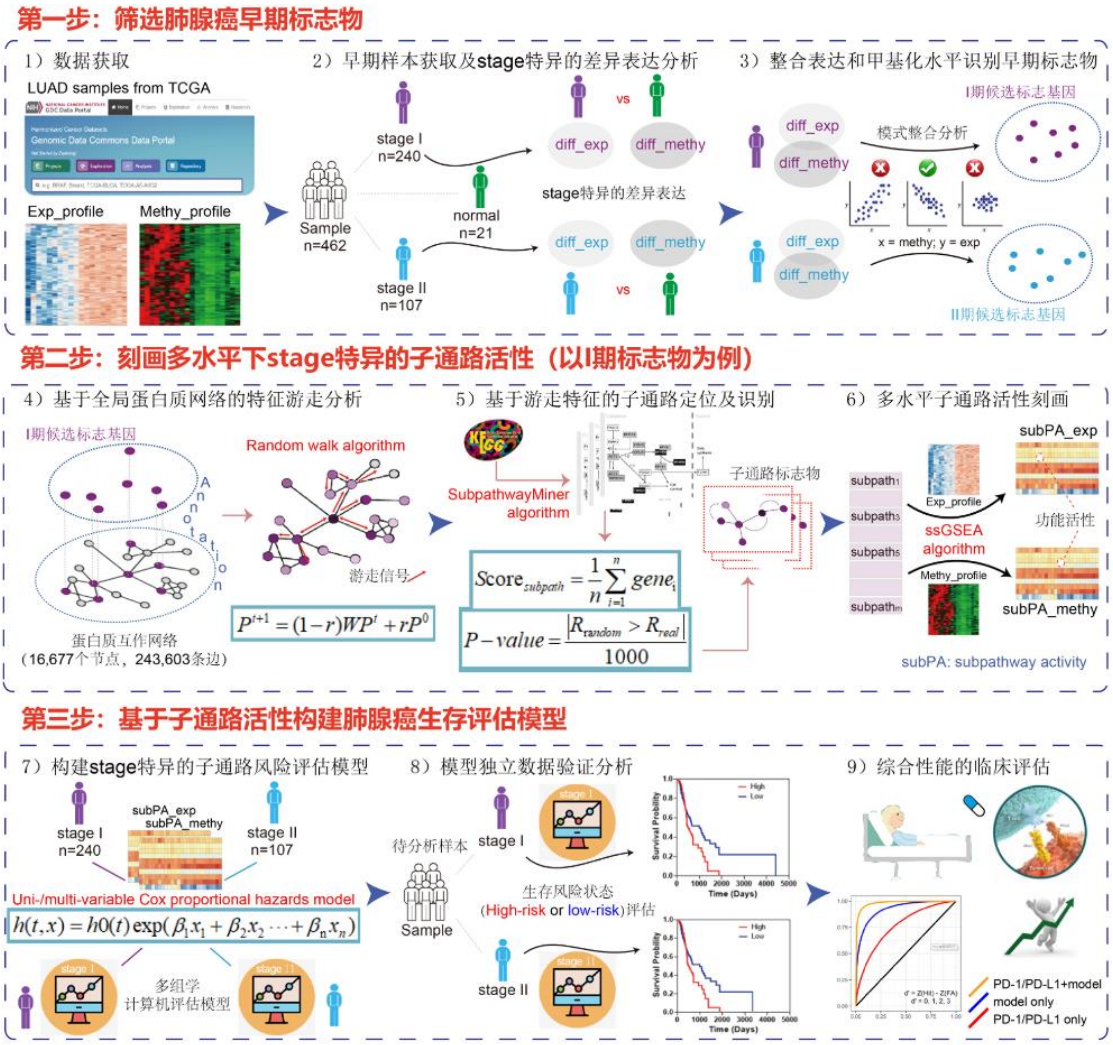

图 1　流程图

三．模型构建流程

（一）模型数据的准备

1.肺腺癌 RNA-seq count 数据获取及预处理

依据临床分期信息，只选取其中既有生存信息又有表达谱数据与甲基化谱的样本。肺腺癌病人样本 RNA-seq count 数据从 UCSC 数据库上获得（https://xenabrowser.net/），由于该数据对原始的 count 值进行了$log_2(count+1)$的处理，所以这里对 count 值取2^x-1进行还原，将探针转化为 geneSymbol,对重复基因的 count 值取均值并且向下取整。

2.肺腺癌 DNA 甲基化数据获取及预处理

从 UCSC 上下载获取 TCGA-LUAD.methylation450.tsv 甲基化谱数据。首先，对甲基化谱数据的样本进行筛选以及分组，接下来构建癌旁样本和 I、II 期肿瘤样本的甲基化矩阵，其中每一列代表一个样本，每一行代表一个探针。然后，对甲基化数据谱进行预处理操作，首先删除 80%缺失的探针，然后利用 k 近邻算法[20]，这里选择 5 个邻居的平均值来填充缺失值，接下来用到 ChAMP R 包进行数据处理分析[21]。ChAMP 通过加载公式或 minfi 加载功能获取 IDAT 数据，再综合考虑检测到的 P 值，染色体定位，单核苷酸多态在探针中的发生以及杂交现象，对探针进行筛选。同时，ChAMP 也将标准化、功能归一化作为 beta-mixture 分位标准化的可选项，对甲基化数据处理批量效应和混杂因素。用 ChAMP R 包对甲基化谱进行过滤，过滤掉 p-value>0.01 的探针和非 GpC 位点的探针，以及所有 SNP 相关的、X 和 Y 染色体上、和 muti-hit 探针。之后，对过滤后的数据进行标准化操作用于后续的差异分析。

3.肺腺癌临床信息及预后数据的获取和整合

从 UCSC 上下载获取 TCGA-LUAD.GDC_phenotype.tsv 样本信息文件。文件包括了 TCGA 样本性别、最初的病理诊断年龄、种族、TNM 分期（T 为原发肿瘤、N 为淋巴结转移、M 为远端转移）、附加药物治疗、附加放射治疗、新的肿瘤事件后是否初步治疗、barcode、OS（总生存）生存状态、OS 生存时间、DSS（疾病特异生存）、DSS 生存时间等。本项目只取“Primary Tumor”原发性肿瘤样本和“Solid Tissue Normal”癌旁样本进行分析。对于肺腺癌样本，我们按照临床分期信息，将其分为一期、二期。经数据整理后，一共收集到 462 个样本，其中一期样本有 240 个，二期样本有 107 个,癌旁样本有 21 个。本研究后续用到的模型验证集也来自于该样本文件，我们在总样本的不同分期样本中分别随机抽取三分之一作为验证集。以下研究我们将依据样本分期，分别进行 stage 特异的样本与癌旁样本的分析比较。

（二）模型构建方法

1.识别差异表达基因与差异甲基化基因

对经预处理的甲基化表达谱数据求差异甲基化基因，我们利用 ChAMP R 包求出差异甲基化探针（$P<0.05$），ChAMP 是利用甲基化探针与 geneSymbol 对应数据，删除没有对应 geneSymbol 的差异探针，然后用 geneSymbol 替代探针，对于多个探针对应一个 geneSymbol 的情况，我们对其求平均值。这样就实现了从探针到基因的转换，并将结果以热图的形式呈现（图 2）。其中，我们从不同分期样本中随机抽取 23 个样本，并对差异甲基化基因进行聚类。

肺腺癌病人样本 RNA-seq count 数据用 R 语言 DESeq2 包[22]进行识别分析。DESeq2 为 RNA-seq 数据的基因分析提供可全面且通用的解决方案。DESeq2 方法检测并纠正了在离散度对平均表达水平依赖关系的模型中过低的离散度估计。以离散度均值为基准建立模型，使用离散度估计值的平均绝对偏差来减少离群的影响，整体实现收缩估计。收缩估计方法与最大似然方法相比，大大增强了分析结果的稳定性和重现性。

校正后的 P 值小于 0.05，log2FC 大于 1 的基因为显著高表达基因；校正后的 P 值小于 0.05，log2FC 小于-1 的基因为显著低表达基因。将得到的每期差异表达基因结果可视化（图 3）。

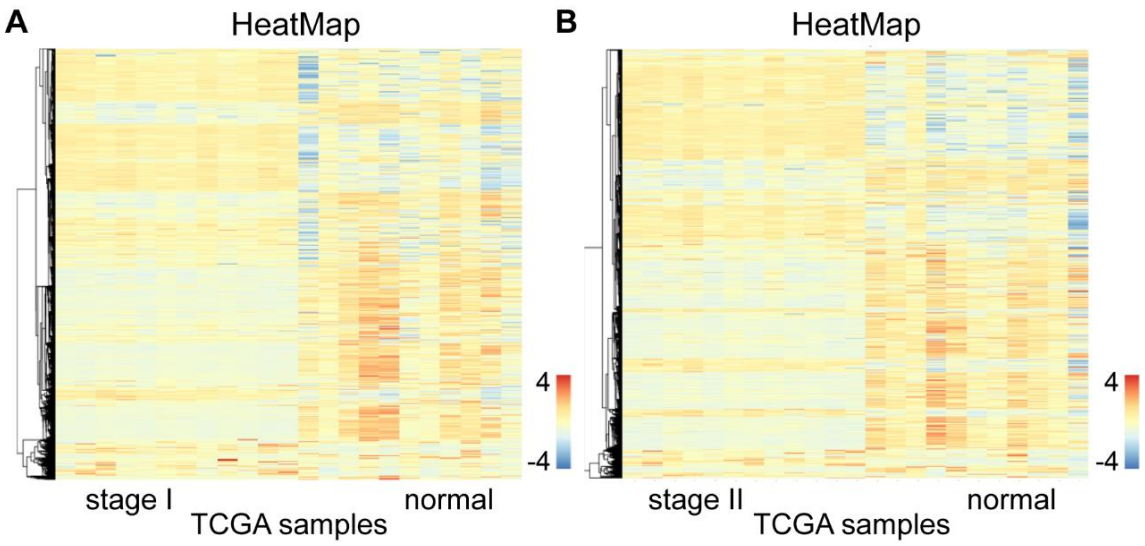

图 2　差异甲基化热图

（A） 是 I 期样本差异甲基化热图；（B）是 II 期样本差异甲基化热图；蓝色代表低甲基化，红色代表高甲基化

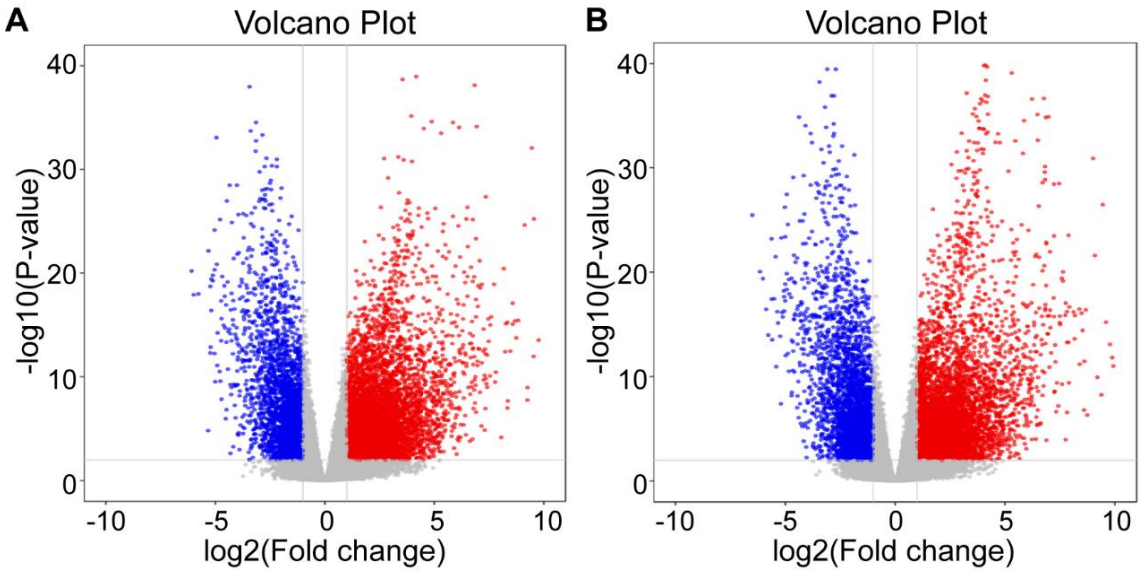

图 3　各分期基因的差异表达情况

（A）为 stageI 期样本与癌旁样本的差异表达基因，（B）stageII 期样本与癌旁样本的差异表达基因。图中的红色圆圈表示癌症样本中差异上调的基因，蓝色圆圈表示癌症样本中差异下调的基因，灰色圆圈表示的基因在癌症样本与癌旁样本中的表达可能存在差异，但认为无统计学的显著性。

2.识别肺腺癌 stage 特异的候选风险基因

然后，对差异表达基因与处理好的甲基化数据的基因取交集，利用皮尔森相关系数求出 DNA 甲基化与 RNA-seq 表达呈显著负相关的基因(cor<-0.3,p<0.05)作为疾病风险基因。这里我们仅展示 stageI、stageII 样本中，基因表达与 DNA 甲基化呈显著负相关程度最强的基因（图 4），图中 stageI 与 stageII 的风险基因相同为 AKR7A3。AKR7A3(Aldo-Keto 还原酶家族成员 A3)是一种蛋白质编码基因。与该基因相关通路有药物代谢，新陈代谢通路和细胞色素 P450。GO（基因本体论）注释中该基因具有电子转移活性和 aldo-keto 还原酶(辅酶 ii)活性。OMIM 数据库查询结果显示 AKR7A3 主要参与醛和酮的解毒。

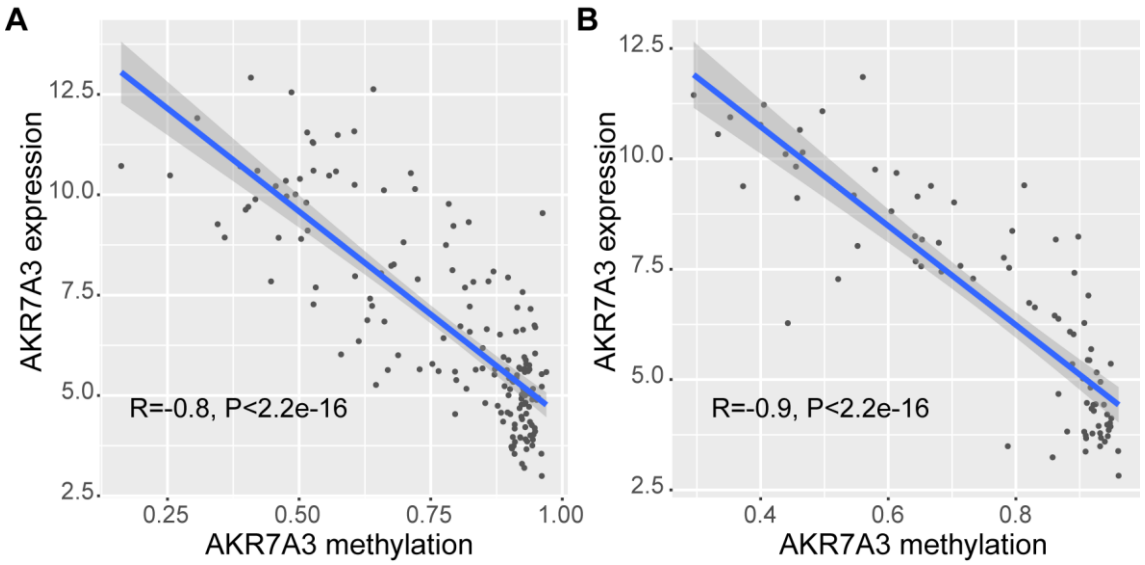

图 4　对潜在风险基因的差异甲基化和差异表达水平计算相关性

图中表示了部分潜在风险基因的相关性分析图。（A）部分表示的是潜在风险基因 AKR7A3 在 Stage I 肺腺癌样本中基因表达与 DNA 甲基化水平的相关性，每一个圆圈代表一个样本，（B）部分表示的是潜在风险基因 AKR7A3 在 StageII 肺腺癌样本中基因表达与 DNA 甲基化水平的相关性。

3.基于 PPI 网络风险基因的随机游走分析

我们从文献中获取基因互作关系[23]，并构建 PPI 网络，将风险基因映射到全局 PPI 网络，删除未注释到网络中的风险基因，然后利用 Random Walk（随机游走算法）[24]对网络中的基因进行扩散，我们以风险基因作为随机游走算法的种子，这一步之后，网络中的每一个基因都会有一个得分值，得分值越高的基因，风险程度越高，得分值越低，说明该基因风险程度越低。该方法不仅考虑了候选基因数目的影响区别，还考虑了 PPI 互作网络中节点的拓扑特异性。

4.注释识别重构的子通路并评估通路显著性

我们从 R 包 SubpathwayMiner 中获得子通路数据，其中包括子通路的注释功能以及子通路中包含的基因集，其中包含 343 个通路，以及相应的 1980 个子通路。与 DAVID 和 GSEA 基因功能富集不同，该方法着重于在子通路水平上进行功能富集，因为通路是由子通路构成的，而生物学异常分析往往需要对通路进行更精细的识别，也就是识别显著的子通路，通路的局部区域（子通路）往往与疾病更加相关。前期研究已经发现癌症的发生和转移往往与多个子通路的交互异常有关，跟正常通路相比，更多相关的异常通路关系密切的子通路在癌症的发生发展中更为重要。SubpathwayMiner 的研发人员使用距离相似法从总通路中定位子通路区域[25]。

我们从 SubpathwayMiner 中获得子通路数据（子通路的注释功能以及子通路中的基因集），先将随机游走结果中基因得分值对应到每个子通路上，计算每个子通路中基因得分的均值，然后通过随机扰动子通路里的基因（N=1000），对于每个子通路每次扰动后会计算一次基因得分的均值，统计随机得分大于真实得分的次数，除以 1000 为 p 值（显著性），本项目认为 p-value<=0.05 的子通路为显著的子通路。其中 stageI 筛选到 181 个子通路，stageII 筛选到 294 个子通路。这一部分，我们将较为显著的子通路从总的通路集中筛选出来，收集得到的这些子通路被认为是潜在与肺腺癌早期预后相关的标志物，我们将利用这些标志物进一步运用回归方法构建模型。

5.表达和甲基化水平下子通路活性评估

筛选出显著子通路后，我们进一步利用 R 包 GSVA 进行子通路活性的刻画评估。GSVA（Gene Set Variation Analysis）为基因集变异分析，主要用来评估转录组基因集富集结果，是一种非参数的无监督分析方法，其核心思想是，通过将不同样本间基因的表达矩阵转化为基因集在样本间的表达量矩阵，来评估不同的通路在样本间是否富集。作为一种分析方法，它主要是通过研究感兴趣的基因集在不同样本间的差异，从功能水平的角度来解释导致表型差异的原因。其中用到的 ssGSEA 方法计算每个样本在各个子通路中的富集得分，也就是每个子通路的活性得分。

这里需要各期样本的基因表达谱和 DNA 甲基化谱，以及子通路里的基因集。如果某一子通路中的基因不能被完全覆盖，就不选取该子通路。首先利用基因表达谱和子通路基因集构建出一个表达水平上子通路活性谱，又利用 DNA 甲基化谱和子通路基因集构建出一个甲基化水平上子通路活性谱。之后，根据样本将两者进行合并，得到最终的子通路活性谱，上半部分属于基因表达水平，下半部分属于甲基化水平，其中每一列为一个样本，每一行为一个通路，其中的数值代表通路活性得分。

经分析，stageI 期的风险通路共有个 360（其中基于表达谱的有 181 个，基于甲基化谱的有 179 个）,stageII 期的风险通路有 588 个（其中基于表达谱的有 294 个，基于甲基化谱的有 294 个）。

6.构建子通路水平风险打分模型

结合样本生存数据（样本生存时间与生存状态）与子通路活性谱，我们利用Cox 回归分析构建了 Stage 特异的早期肺腺癌风险打分模型。Cox 比例风险模型（Cox proportional-hazards model，也称为 Cox 回归），是由英国统计学家 D.R.Cox 在 1972 年提出的用于慢性病和肿瘤的预后分析模型。它的优点是可以使用多因素分析方法，不用考虑生存时间分布，而且还能利用截尾数据。首先，我们利用单因素 Cox 回归分析进行子通路初步筛选。对于 I 期特异的肺腺癌风险子通路，在 $P< 0.05$ 的约束下，我们筛选到了共 10 个子通路。将筛选出的子通路进一步用 Cox 多因素回归拟合，赋予相应子通路的系数(表 1)。基于以上信息，我们构建一个子通路活性水平肺腺癌 I 期风险打分模型。

运用同样的方法，我们对 II 期特异的风险子通路进行单因素 Cox 回归分析，评估通路的生存相关性，进一步筛选到共 25 个子通路。将筛选出的子通路用 Cox 多因素回归拟合，赋予 II 期子通路的系数(表 3)，基于以上信息，我们构建了肺腺癌 II 期风险打分模型。

其中 stageI 期风险打分模型中包含的子通路分别富集到了以下的通路中（表 2）。stageII 期风险打分模型中包含的子通路分别富集到了以下的通路中（附表 1）。

表 1　特征通路系数（I 期）

	SubPathWay	Coef	P
1	Exppath:04064_18	-0.769150	0.003
2	Exppath:04380_22	-0.513843	0.027
3	Exppath:04520_8	0.103206	0.044
4	Exppath:04660_10	-0.132797	0
5	Exppath:05100_11	-1.548765	0.016
6	Exppath:05132_1	-1.758473	0.047
7	Exppath:05168_1	-1.978194	0.027
8	Exppath:05200_31	-1.731284	0.026
9	Exppath:05222_9	4.209545	0.021
10	Methpath:05223_6	-2.398497	0.041

注：该表展示了回归分析后得到的，具有 I 期肺腺癌生存相关性的特征通路，以及其影响系数。P 值评估了对应特征性征对整体模型影响的显著性效果。

表 2　生存相关的风险子通路（I 期）

PathwayId	PathwayName
path:04660	T cell receptor signaling pathway
path:04064	NF-kappa B signaling pathway
path:05100	Bacterial invasion of epithelial cells
path:05222	Small cell lung cancer
path:05200	*Pathways in cancer
path:04380	Osteoclast differentiation
path:05168	*Herpes simplex infection
path:05223	*Non-small cell lung cancer
path:04520	*Adherens junction
path:05132	Salmonella infection

注：该表展示了通过 Cox 回归方法最终映射到的：与 I 期肺腺癌患者生存相关的风险通路。

其中用 *标注的通路为在 I 期和 II 期均风险的通路，其余未标注的子通路为 I 期肺腺癌患者与生存相关的特异性通路。

表 3　特征通路系数（II 期）

	SubPathWay	Coef	P
1	Exppath:04010_13	2.443268	0.016
2	Exppath:04510_18	-2.373545	0.048
3	Methpath:04011_1	0.640159	0.011
4	Methpath:04060_38	0.258341	0.004
5	Methpath:04060_62	0.247733	0.026
6	Methpath:04066_6	1.021489	0.03
7	Methpath:04520_4	0.245641	0.046
8	Methpath:04623_2	3.183853	0.006
9	Methpath:04623_3	-1.903804	0.012
10	Methpath:04970_2	-1.880446	0.026
11	Methpath:04971_3	8.304757	0.045
12	Methpath:04971_4	-9.296025	0.038
13	Methpath:05110_2	1.098957	0.029
14	Methpath:05160_13	0.868261	0.031
15	Methpath:05160_2	5.069147	0.004
16	Methpath:05166_19	-3.147393	0.022
17	Methpath:05168_14	-6.239419	0.001
18	Methpath:05200_13	2.702755	0.02
19	Methpath:05200_20	-2.56684	0.027
20	Methpath:05211_1	1.447350	0.007
21	Methpath:05211_2	-18.779025	0.017
22	Methpath:05211_7	15.480469	0.004
23	Methpath:05212_2	-5.086483	0.025
24	Methpath:05220_4	0.291145	0.032
25	Methpath:05223_6	-0.370321	0.012

注：该表展示了回归分析后得到的，具有 II 期肺腺癌生存相关性的特征通路，以及其影响系数。P 值评估了对应特征性征对整体模型影响的显著性效果。

四．模型有效性验证

（一）依据风险打分取得区分各分期风险性的 cutoff 值

基于以上构建的子通路水平风险打分模型，我们进一步将其应用到测试集的肺腺癌各期（I、II）样本上，以获取各分期中所有样本的子通路水平风险得分。对于每一分期的患者风险得分，我们使用训练集计算其每个样本风险得分的中位数，作为该分期划分患者风险性（高风险或低风险）的指标，此处我们将其命名为 cutoff 值。考虑到中位风险得分的稳健性与有效性，我们认为 cutoff 值对于评估不同分期患者的疾病风险有重要的生物学意义。我们得到 I 期的 cutoff 值为 -4.477421 ，II 期的 cutoff 值为 -8.295978 ，利用 cutoff 值可以衡量患者预后风险。

（二）结合病人生存信息，评估模型有效性

为进一步验证 cutoff 值具有显著的风险划分效果，即以上设计模型的有效性，我们以 I 期患者为例，将验证集中 I 期患者以 cutoff 值为界分为两组(高风险组和低风险组）。分组之后，我们通过样本的临床信息获取其相应的生存数据（生存时间以及生存状态）作为考察信息，调用 survival 包，以构建 K-M(Kaplan-Meier) 生存曲线。

可视化 I 期高风险组与低风险组的生存状态后，我们可以清楚地观察到两组人群的整体生存时间上的差异(p=0.014)（图 5）。我们又利用 cox 单因素进行验证其 p=0.0173,通过其显著性指数的审核（P<0.05），进一步印证了我们的猜测，即 I 期构建的风险打分模型具有显著的区分患者风险性的效果。

同样地，我们对肺腺癌 II 期样本对象进行 cutoff 预测以及分组查看。从而进一步构建生存曲线（图 6）。结果显示该曲线区分高风险队列和低风险队列的显著性指数为 p=0.083 , 我们又利用 cox 单因素进行验证其 p=0.0947, p 值接近显著。经分析可能是由于 stageII 期样本量太少，导致结果不理想。但是换一个角度看，stageII 期样本只有 36 个样本，能够得到 0.083 的显著性水平，表明 stageII 期模型还是有一定意义的。基于以上两组实验，证明了我们的模型在区分早期肺腺癌患者风险上具有一定的显著性，stageI 期效果显著，stageII 期效果接近显著。

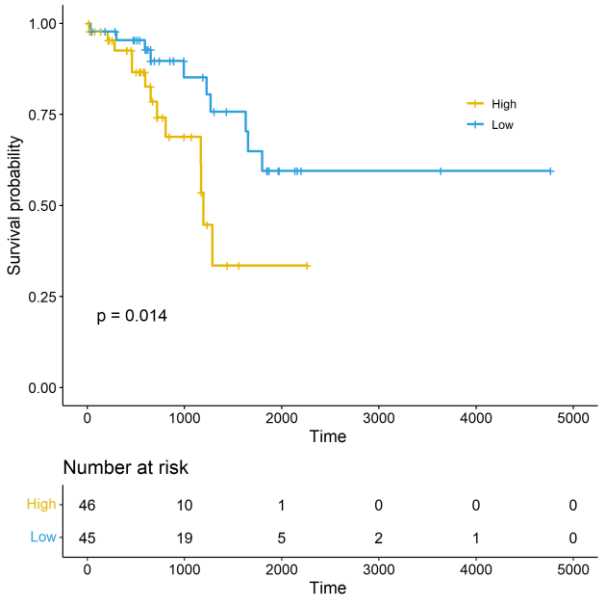

图 5　stageI 期患者生存曲线

图中黄色代表高风险样本，蓝色代表低风险样本，横轴代表生存时间（day），纵轴代表样本的生存率。在同一时间点，不同风险等级的患者明显分开，其中低风险样本的生存率明显高于高风险样本的生存率，并且经 log-rank 检验，其 p 值非常显著。

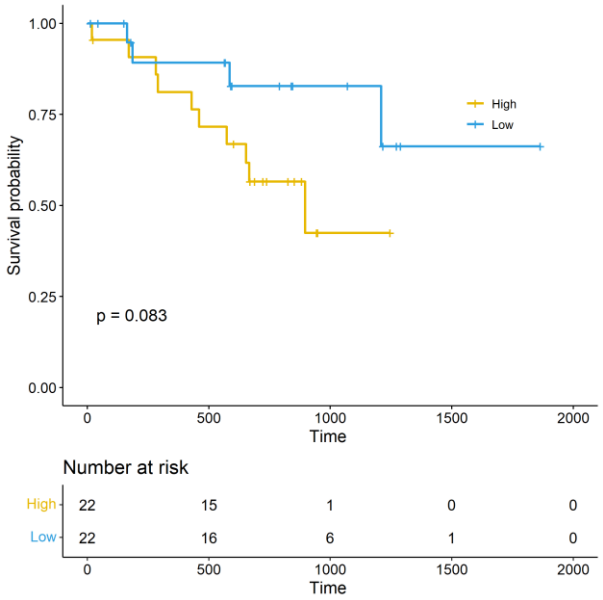

图 6　stageII 期患者生存曲线

图中黄色代表高风险样本，蓝色代表低风险样本，横轴代表生存时间（day），纵轴代表样本的生存率。

五．分析不同风险打分模型的功能差异

（一）肺腺癌早期构建模型的通路层面的差异

根据 cox 回归筛选出的 stageI 和 stageII 子通路（如表 3 和附表 1），发现 stageI 和 stageII 期子通路都富集到了如下通路中（4 个）（图 7A）：癌症中的通路（Pathways in cancer path:05200）、单纯疱疹感染（Herpes simplex infection path:05168）、非小细胞肺癌（Non-small cell lung cancer path:05223）、Adherens junction(path:04520)。其中 stageI 和 stageII 共有的子通路是非小细胞肺癌通路（图 7B、图 8）。在 stageI 期中 T 细胞受体信号通路（T cell receptor signaling pathway path:04660）最为显著，其次是 Nf-kappa b 信号通路，然后是细菌侵袭上皮细胞的信号通路（Bacterial invasion of epithelial cells path:05100）。而在 stageII 期中 Herpes simplex infection 通路最显著，在 stageI 期中没有被筛选出的细胞因子-细胞因子受体相互作用（Cytokine-cytokine receptor interaction path:04060），而疟疾（Malaria path:05144）为第三显著的结果。

T 细胞受体在 T 细胞行使功能以及免疫突出形成方面都起到重要作用。它是联系 T 细胞和抗原提呈细胞的桥梁。而 T 细胞受体的激活促进一些信号级联，从而调节一些细胞因子的产生、细胞生存、增值以及分化的过程，因而达到决定细胞命运的目的。解释 stageI 期患者 T 细胞受体的激活，启动一系列级联反应以及产生一些细胞因子，攻击肿瘤细胞。

单纯疱疹（Herpes simplex）是一种常见的病毒性感染，表现为局部水泡，会一次或多次出现在大多数人的一生中。诱发因素为两种类型的单纯疱疹病毒（HSV）之一，任何一种感染都会到达至任一部位的皮肤或粘膜。为了更好地进行免疫逃逸以及在再次激活时抵御免疫攻击，HSV-1 具有多重免疫机制以对抗免疫反应，促进病毒在宿主体内更好的生存和繁殖，这与肿瘤的免疫逃逸有相似之

处。核转录因子是炎症及免疫反应调控的枢纽，它能调控多种粘附因子，细胞因子及凋亡相关的基因的表达，并且参与了肿瘤发生的分子机制。近年来的研究发现单纯疱疹病毒能够利用核转录因子的持续激活，促进病毒的复制和抑制凋亡。StageII 期患者 Herpes simplex 通路的显著激活，可能与肿瘤的免疫逃逸相关。

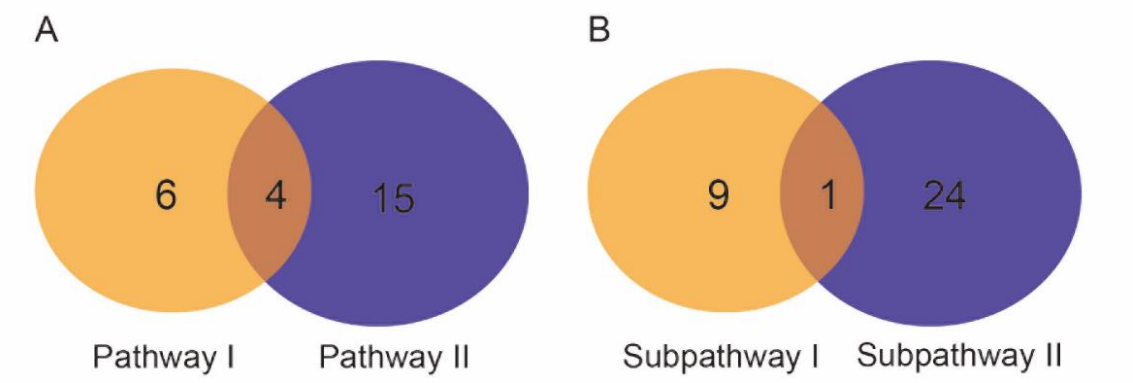

图 7 一二期模型中通路以及子通路的比较

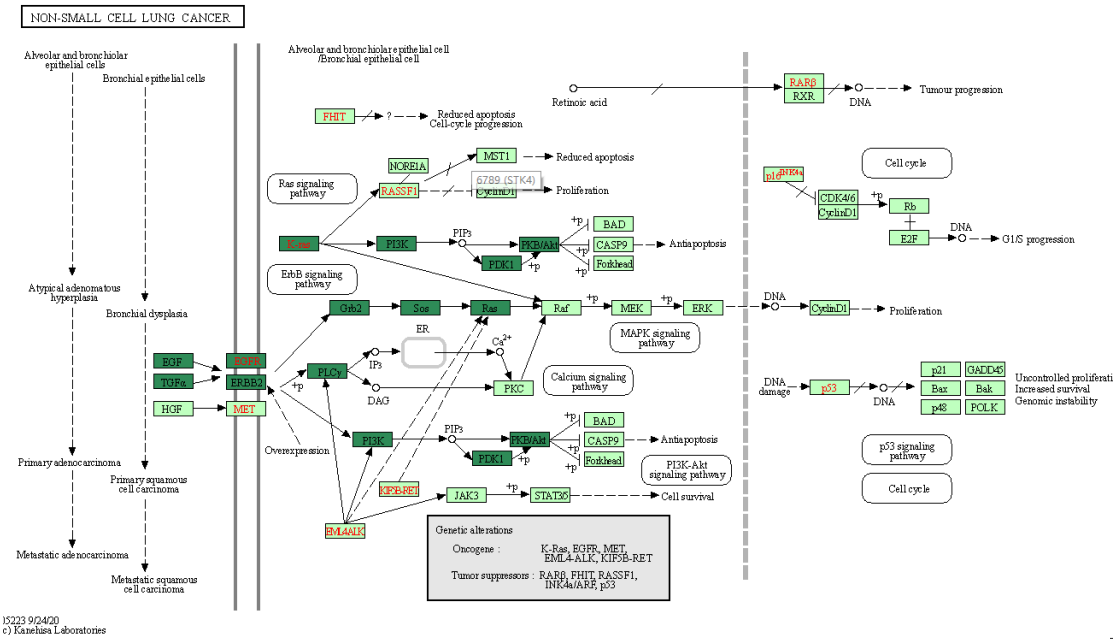

图 8 非小细胞肺癌通路中的子通路

图中深绿色部分代表非小细胞肺癌通路中的子通路结构，该通路图来自于 KEGG 数据库（https://www.kegg.jp/pathway/hsa05223）

六．分析预后标志物的可靠性

（一）免疫特征信息验证模型可靠性

了解肿瘤的发生和转移过程对抗癌疗法的发展至关重要。实现这一点的方法之一是通过研究上皮性伤口愈合（Wound Healing），伤口愈合的生理过程与癌症进展非常相似。哺乳动物的伤口愈合反应涉及一系列精密的过程和分子事件，由特定信号通路的短暂激活严格调控。准确控制这些事件是至关重要的，如果未能在正确的时间启动关键步骤，将会延迟伤口的愈合，并导致慢性伤口。有研究表明伤口愈合的异常启动过程，可能会促进癌症进展[26]。前期相关研究人员提出了“癌症是一个过度愈合的伤口”的观点。许多对伤口愈合至关重要的信号通路和分子机制都与癌细胞增殖或转移有关[27, 28]。

转化生长因子(TGF-b)是一种多功能细胞因子，属于转化生长因子超家族，它参与各种生物过程，包括细胞增殖、免疫监测、血管生成、胚胎干细胞硬化和分化等。激活的 TGF-b 复合物与其他因子形成丝氨酸/苏氨酸激酶复合物，随后与 TGF-b 受体结合，刺激下游磷酸化信号级联，最终会导致不同底物和调控蛋白的激活[29]。异常高水平的 TGF-b 配体存在于各种恶性实体中，其中 TGF-b 增加与肿瘤进展和不良预后相关。鉴于 TGF-b 和伤口愈合在肿瘤发生发展中的关键作用，我们利用这两个因素验证我们模型的可靠性[30]。

研究人员利用 TCGA 汇编的数据，对包括 33 种不同癌症的 10 万种肿瘤进行了广泛的免疫原性分析[31]，鉴别出了六种免疫亚型，包括：伤口愈合，IFN-γ Dominant，炎症，淋巴细胞衰竭，免疫静息，TGF-β Dominant。研究者进一步研究了不同癌症亚型之间巨噬细胞或淋巴细胞特征、肿瘤异质性程度、细胞增殖、免疫调节基因的表达，以及预后等指标的差异。我们从文献中提取出 TCGA 肺腺癌样本集免疫细胞相关信息，选取其中的 Wound Healing 和 TGF-beta Response 两个免疫特征，对模型进行可靠性进行验证。

首先利用风险打分模型将验证集分为高风险组和低风险组，利用文献中提供的 Wound Healing 和 TGF-beta Response 两个免疫特征，对高低风险样本进行显著性检验（t 检验），并绘制箱式图将结果可视化（图 9）。StageI 期和 stageII 期高风险样的 Wound Healing 和 TGF-beta Response 均值都高于低风险组。

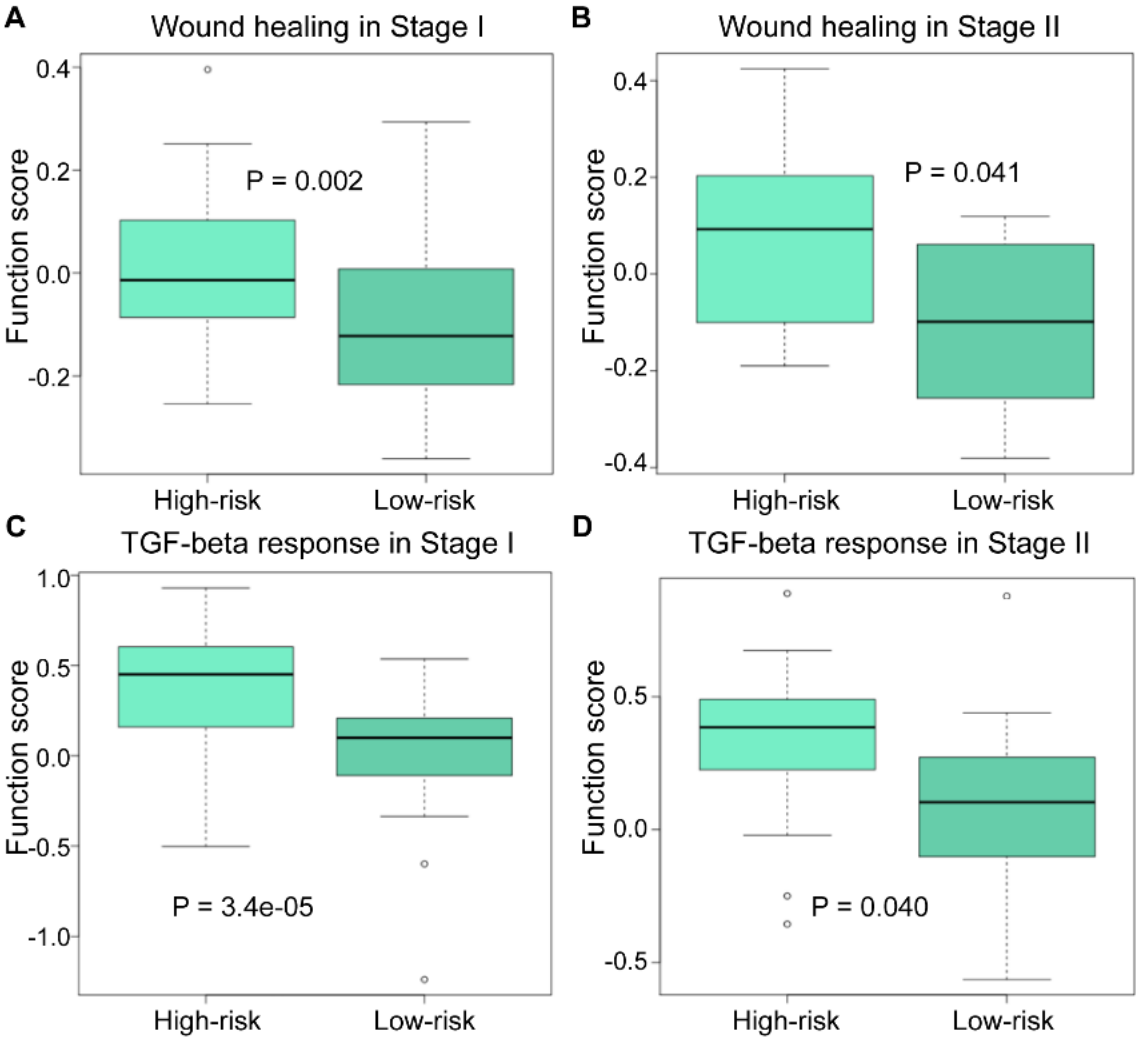

图 9　两个免疫特征在高低风险组中的比较

A 图与图 B 代表 Wound Healing 特征在 stageI 和 stageII 期中高低风险组的差异，C 图与 D 图代表 TGF-beta Response 特征在 stageI 和 stageII 期中高低风险组的差异。

七．总结与展望

（一）总结

对于肺腺癌早期（stageI,stageII 分期）的样本对象，首先通过 bulk RNA-Seq 数据，探究其与癌旁样本基因表达上的差异。肿瘤在发生发展的过程中，经常会调控许多本应沉默的基因高表达，使一些本应活跃的基因沉默。因此，选择在该分期下差异表达的基因，可以较为有效的评估其差异状态。以相似的策略处理样本的表达谱数据，获取差异甲基化的基因。将差异甲基化与差异基因关联，分析其相关性，筛选差异甲基化影响下的差异表达基因。这些基因与该分期下样本的差异甲基化与差异表达状态都相关，可以认为是有一定风险的潜在生物标志物。

将获取的风险基因映射到高质量的蛋白质-蛋白质互作网络（PPI）中，以探究其网络的拓扑性质。随机游走策略可以有效评估网络基因的整体风险性，并可将其风险性以得分的形式表现。在赋予网络基因风险得分后，通过 SubpathwayMiner 方法可以较为准确的识别到风险基因富集的子通路，通路的富集得分通过对其相关基因的风险值求平均数而得出。子通路的富集分数越高，其对肺腺癌各分期患者预后影响的程度也越大。对于某一风险子通路，比较该通路风险得分与随机扰动 1000 次的子通路得分比较，其显著性可以进一步验证所研究通路的重要意义。

stageI 期筛选得到的风险子通路共有 181 个，stageII 期有 294 个。基于各通路的相关基因集与已知基因表达谱以及 DNA 甲基化谱，联合构建具有基因表达或甲基化水平特征的通路活性打分模型。联合基因表达谱以及 DNA 甲基化谱可以更全面地描述子通路活性。

采用单因素及多因素 Cox 回归方法分析活性值。由于通路活性具有差异性，Cox 回归可全面地评估子通路对疾病独立的影响以及联合效应。最终获得了可对肺腺癌各分期风险性估计的打分模型。

将该打分模型应用于训练样本中，得到了划分 stageI 期与 stageII 期风险性的 cutoff 得分。中位值的选取使该分类具有更稳健且全局性的效果。生存分析用于验证分类的效能，结果显示具有很好的鲁棒性。

本文应用生信分析方法，基于已有的样本数据展开讨论并构建风险评估模型。由于模型的特殊性，可获取的样本数有限，但足够支撑模型的构建，且已验证其效能。该模型也可应用在其他肺腺癌风险性预测与评估上，具有临床指导的作用。

（二）展望

尽管如今测序事业及相关行业已经展现蓬勃发展之势，但由于数据库权限、涉及患者个人隐私、测序成本高、数据来源有限等因素，使得能够获取到的样本数据有限。本模型的构建涉及到多组学数据，为筛选来自同一样本的甲基化和表达谱数据，已进行大量检索并获得足以支撑模型的样本量，且已验证其效能。我们相信，客观上的限制会随着测序技术的逐步推广以及测序成本的下降而得到改善。期待在不远的将来，有更多的测序数据可供研究人员分析，生物医学大样本量的拟合模型将会成为精准医疗的“利器”。

参考文献

1. Li, H., et al., *Genome-wide analysis of the hypoxia-related DNA methylation-driven genes in lung adenocarcinoma progression.* Biosci Rep, 2020. **40**(2).
2. Lopez Guerra, J.L., et al., *Risk factors for local and regional recurrence in patients with resected N0-N1 non-small-cell lung cancer, with implications for patient selection for adjuvant radiation therapy.* Ann Oncol, 2013. **24**(1): p. 67-74.
3. Wang, X., et al., *Prediction of recurrence in early stage non-small cell lung cancer using computer extracted nuclear features from digital H&E images.* Sci Rep, 2017. **7**(1): p. 13543.
4. Wu, C.F., et al., *Recurrence Risk Factors Analysis for Stage I Non-small Cell Lung Cancer.* Medicine (Baltimore), 2015. **94**(32): p. e1337.
5. Zhang, Y., et al., *A clinicopathologic prediction model for postoperative recurrence in stage Ia non-small cell lung cancer.* J Thorac Cardiovasc Surg, 2014. **148**(4): p. 1193-9.
6. Zhang, Y., et al., *Development and Validation of Web-Based Nomograms to Precisely Predict Conditional Risk of Site-Specific Recurrence for Patients With Completely Resected Non-small Cell Lung Cancer: A Multiinstitutional Study.* Chest, 2018. **154**(3): p. 501-511.
7. Kerr, K.M. and M.C. Nicolson, *Prognostic factors in resected lung carcinomas.* EJC Suppl, 2013. **11**(2): p. 137-49.
8. Birim, O., et al., *Survival after pathological stage IA nonsmall cell lung cancer: tumor size matters.* Ann Thorac Surg, 2005. **79**(4): p. 1137-41.
9. Gajra, A., et al., *Impact of tumor size on survival in stage IA non-small cell lung cancer: a case for subdividing stage IA disease.* Lung Cancer, 2003. **42**(1): p. 51-57.
10. Port, J.L., et al., *Tumor size predicts survival within stage IA non-small cell lung cancer.* Chest, 2003. **124**(5): p. 1828-33.
11. Zhang, L., Z. Zhang, and Z. Yu, *Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma.* J Transl Med, 2019. **17**(1): p. 423.
12. Thornblade, L.W., et al., *Challenges in Predicting Recurrence After Resection of Node-Negative Non-Small Cell Lung Cancer.* Ann Thorac Surg, 2018. **106**(5): p. 1460-1467.
13. He, Y., et al., *Hypomethylation of the hsa-miR-191 locus causes high expression of hsa-mir-191 and promotes the epithelial-to-mesenchymal transition in hepatocellular carcinoma.* Neoplasia, 2011. **13**(9): p. 841-53.
14. Shen, N., et al., *A Diagnostic Panel of DNA Methylation Biomarkers for Lung Adenocarcinoma.* Front Oncol, 2019. **9**: p. 1281.
15. Shi, B., et al., *Genetic and epigenetic regulation of major histocompatibility complex class I gene expression in bovine trophoblast cells.* Am J Reprod Immunol, 2018. **79**(1).
16. He, W., et al., *Aberrant CpG-methylation affects genes expression predicting survival in lung adenocarcinoma.* Cancer Med, 2018. **7**(11): p. 5716-5726.
17. Goncalves, C.S., et al., *WNT6 is a novel oncogenic prognostic biomarker in human glioblastoma.* Theranostics, 2018. **8**(17): p. 4805-4823.
18. Shi, J., et al., *Serum miR-626 and miR-5100 are Promising Prognosis Predictors for Oral Squamous Cell Carcinoma.* Theranostics, 2019. **9**(4): p. 920-931.
19. Wang, H., et al., *Long noncoding RNA miR503HG, a prognostic indicator, inhibits tumor metastasis by regulating the HNRNPA2B1/NF-kappaB pathway in hepatocellular carcinoma.* Theranostics, 2018. **8**(10): p. 2814-2829.
20. Zhang, Z., *Introduction to machine learning: k-nearest neighbors.* Ann Transl Med, 2016. **4**(11): p. 218.
21. Tian, Y., et al., *ChAMP: updated methylation analysis pipeline for Illumina*

BeadChips. Bioinformatics, 2017. **33**(24): p. 3982-3984.
22. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
23. Cheng, F., I.A. Kovacs, and A.L. Barabasi, *Network-based prediction of drug combinations.* Nat Commun, 2019. **10**(1): p. 1197.
24. Musco, C., H.H. Su, and N.A. Lynch, *Ant-inspired density estimation via random walks.* Proc Natl Acad Sci U S A, 2017. **114**(40): p. 10534-10541.
25. Li, C., et al., *SubpathwayMiner: a software package for flexible identification of pathways.* Nucleic Acids Res, 2009. **37**(19): p. e131.
26. Sundaram GM, Quah S, Sampath P. Cancer: the dark side of wound healing. FEBS J. 2018 Dec;285(24):4516-4534. doi: 10.1111/febs.14586. Epub 2018 Jun 25. PMID: 29905002.
27. Rybinski, B., J. Franco-Barraza, and E. Cukierman, *The wound healing, chronic fibrosis, and cancer progression triad.* Physiol Genomics, 2014. **46**(7): p. 223-44.
28. Schafer, M. and S. Werner, *Cancer as an overhealing wound: an old hypothesis revisited.* Nat Rev Mol Cell Biol, 2008. **9**(8): p. 628-38.
29. Massague, J., *TGFbeta signalling in context.* Nat Rev Mol Cell Biol, 2012. **13**(10): p. 616-30.
30. Tao, S., et al., *TGF-beta/Smads Signaling Affects Radiation Response and Prolongs Survival by Regulating DNA Repair Genes in Malignant Glioma.* DNA Cell Biol, 2018. **37**(11): p. 909-916.
31. Thorsson, V., et al., *The Immune Landscape of Cancer.* Immunity, 2018. **48**(4): p. 812-830 e14.

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　生存相关的风险子通路（II 期）

PathwayId	PathwayName
path:05168	*Herpes simplex infection
path:04060	Cytokine-cytokine receptor interaction
path:04623	Cytosolic DNA-sensing pathway
path:05160	Hepatitis C
path:05211	Renal cell carcinoma
path:04011	MAPK signaling pathway - yeast
path:05223	*Non-small cell lung cancer
path:04010	MAPK signaling pathway
path:05200	*Pathways in cancer
path:05166	HTLV-I infection
path:05212	Pancreatic cancer
path:04970	Salivary secretion
path:05110	Vibrio cholerae infection
path:04066	HIF-1 signaling pathway
path:05220	Chronic myeloid leukemia
path:04971	Gastric acid secretion
path:04520	*Adherens junction

path:04510	Focal adhesion

注：该表展示了通过 Cox 回归方法最终映射到的：与 II 期肺腺癌患者生存相关的风险通路。

其中用 *标注的通路为在 I 期和 II 期均风险的通路，其余未标注的子通路为 II 期肺腺癌患者与生存相关的特异性通路。

致谢

从开始拿到论文题目到论文的顺利完成，一直都离不开老师、同学、朋友给我热情的帮助，在这里请接受我诚挚的谢意!

此论文是在张老师的耐心点拨和诚恳建议下，一次又一次的修改，最终完成的。从选题到论文的最终完成，老师都给予我细心的指导和不懈的支持。老师在我论文撰写的前期给了我很好的指引和帮助，并且在撰写的后期在百忙之中还抽时间一次又一次的给我提建议，帮助我论文的完善。“经师易得，人师难求”，希望借此机会向张老师表示最衷心的感谢!

同时，还要真挚地感谢李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。