

# 武汉理工大学毕业设计（论文）

## 基于放射性肺损伤芯片数据的预后模型的构建

学院（系）： 信息工程学院

专业班级： 通信 2002 班

学生姓名： 刘 明

指导教师： 王原丽

## 学位论文原创性声明

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名：

年 月 日

## 学位论文版权使用授权书

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构递交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密口，在 年解密后适用本授权书

2、不保密口。

(请在以上相应方框内打“√”)

作者签名： 年 月 日

导师签名： 年 月 日

## 摘要

严重的放射性肺损伤可能会缩短肺腺癌患者的生存期，预后不良。此外，自噬相关基因参与肺腺癌的发生发展，预测放射性肺损伤中自噬相关基因的表达情况在预测肺癌患者生存方面可能具有潜在意义。本文研究了如何利用放射性肺损伤芯片数据来构建一个预测肺腺癌预后的模型。

本文通过对放射性肺损伤芯片实验数据的获取和清洗，并结合统计学方法检测和去除数据的批次效应，利用加权相关网络分析不同基因之间的表达关系，并检测数据的离群值。利用基因差异分析和基因集变异分析筛选出与肺腺癌预后显著的自噬相关基因。

利用 COX 回归和最小绝对收缩和选择算子回归挑选具有统计学意义的基因作为自变量构建预后模型，并挑选健康人类志愿者和肺腺癌患者的基因数据作为测试集，测试了模型的受试者工作特性曲线。根据风险评分模型将肺腺癌患者划分为高风险组和低风险组，并对两个风险组进行对数秩检验。对模型结果的生物信息学过程进行分析，分析了高低风险组在不同特征维度上的分布差异，以及免疫细胞在两个风险组中的分布情况。结果表明风险评分模型具有显著统计学意义。

这项研究不仅揭示了辐射与肺腺癌之间的潜在联系，还为我们提供了一个全新的肺腺癌预后预测工具。

**关键词：**器官芯片；预后模型；生物信息分析；COX 回归

## Abstract

Severe radiation-induced lung injury may shorten the survival time and lead to poor prognosis in patients with lung adenocarcinoma. In addition, autophagy-related genes are involved in the occurrence and development of lung adenocarcinoma, and predicting the expression of autophagy-related genes in radiation-induced lung injury may have potential significance in predicting the survival of lung cancer patients. This paper studies how to use radiation lung injury microarray data to build a model to predict the prognosis of lung adenocarcinoma.

In this paper, we acquire and clean radiation-induced lung injury chip experimental data and combine it with statistical methods to detect and remove batch effects in the data. We use weighted correlation networks to analyze the expression relationships between different genes and detect outliers in the data. Gene difference analysis and gene set variation analysis were used to screen out autophagy-related genes that are significantly related to the prognosis of lung adenocarcinoma.

COX regression and least absolute shrinkage and selection operator regression were used to select statistically significant genes as independent variables to construct a prognostic model, and genetic data from healthy human volunteers and lung adenocarcinoma patients were selected as test sets to test the model's subjects. or operating characteristic curve. Lung adenocarcinoma patients were divided into high-risk and low-risk groups according to the risk score model, and the log-rank test was performed on the two risk groups. The bioinformatics process of the model results was analyzed, and the distribution differences of high and low risk groups in different feature dimensions were analyzed, as well as the distribution of immune cells in the two risk groups. The results show that the risk score model is statistically significant.

This study not only reveals a potential link between radiation and lung adenocarcinoma, but also provides us with a new prognostic tool for lung adenocarcinoma.

**Key Words:** Organ-on-a-chip; prognostic model; bioinformatics analysis; COX regression

# 目 录

第 1 章 绪论 .....	1
1.1 研究背景、目的和意义 .....	1
1.1.1 研究背景 .....	1
1.1.2 研究目的 .....	2
1.1.3 研究意义 .....	2
1.2 相关领域国内外研究现状 .....	2
1.3 本文主要研究内容和组织结构 .....	3
1.3.1 本文主要研究内容 .....	3
1.3.2 本文组织结构 .....	3
第 2 章 放射性肺损伤芯片数据获取筛选 .....	5
2.1 数据处理 .....	5
2.1.1 基因表达数据获取和清洗 .....	5
2.1.2 批次效应的检查和去除 .....	6
2.1.3 离群值检测 .....	8
2.2 构建加权相关网络 .....	9
2.3 筛选差异表达基因 .....	11
2.4 基因富集 .....	13
2.4.1 GO 功能富集 .....	14
2.4.2 KEGG 通路富集 .....	14
2.5 本章小结 .....	15
第 3 章 风险评分模型的建立与测试 .....	16
3.1 风险评分模型的建立 .....	16
3.1.1 单因素 COX 回归模型 .....	16
3.1.2 LASSO 回归模型 .....	17
3.1.3 多因素 COX 回归模型 .....	20
3.2 风险评分模型的测试与性能分析 .....	22

3.2.1 受试者工作特性分析 .....	22
3.2.2 Kaplan-Meier 生存分析 .....	23
3.3 本章小结 .....	28
第 4 章 风险评分结果的生物信息学分析 .....	29
4.1 两个风险组的分布情况分析 .....	29
4.2 两个风险组的分布的免疫浸润分析 .....	30
4.3 本章小结 .....	33
第 5 章 总结与展望 .....	34
5.1 全文工作总结 .....	34
5.2 下一步工作展望 .....	34
参考文献 .....	36
致 谢 .....	39

# 第1章 绪论

## 1.1 研究背景、目的和意义

### 1.1.1 研究背景

在近五十年，肺癌的发病率和死亡率一直居高不下，并且呈现逐年上升的趋势。据统计，肺癌患者的五年生存率只有 15% 左右<sup>[1]</sup>。肺腺癌是非小细胞癌中最常见的亚型，其致病因素主要包括放射性损伤、二手烟、污染物和职业致癌物等<sup>[1]</sup>。其中放射性事故有可能导致放射性肺损伤，如法国铀矿工队列中由于  $\gamma$  射线暴露导致肺癌<sup>[2]</sup>，另一方面肺癌的放射性治疗也会导致放射性损伤的加重<sup>[3]</sup>。

2021 年 12 月，国家工业和信息化部等部门发布发展规划，强调攻关基于新型微型流体控制器、医疗级可穿戴监护装备和人工器官<sup>[4]</sup>。2023 年 9 月，华为盘古结合器官芯片开发了全球首个人体器官芯片医药大模型。2023 年 10 月国家卫健委、发改委等 13 个部门联合发布方案推动癌症防治工作高质量发展，其中涉及深入推进构建分层癌症筛查体系等三项举措<sup>[5]</sup>。2024 年 5 月，中国科学院大连化学物理研究所研究员秦建华牵头，联合多个单位的专家共同成立器官芯片标准起草工作组，形成器官芯片三项团体标准。

多数肺腺癌患者在诊断时已经处于晚期阶段，因而总生存率较低。随着生物信息学和机器学习理论的发展，越来越多的标志物被用于肺腺癌的预测，如肿瘤大小、突变负荷、转移情况等。自噬是由自噬相关基因调控的溶酶体降解过程，可参与多种癌症的发生发展，是肿瘤发展过程中的关键调控因子，如 ATG10 的过量表达与肺癌的不良预后有关，因此能够通过多个自噬相关基因构建风险模型用于癌症的预后模型建立。

器官芯片在近年来得到了广泛的关注，涉及微纳技术和生物医学的结合。器官芯片包含用于操纵和引导微小体积溶液的微流体装置、细如发丝的微通道网络、生长并驻留在微流控芯片中的细胞<sup>[6]</sup>。器官芯片和微生理系统的领域呈现指数级增长，为了更好地了解健康和疾病背后的生理学，并寻找对其进行改善的方法，器官芯片技术的认可度已经远远超出高校实验室的范围，被世界经济论坛评选为 2016 年十大新兴技术之一<sup>[7]</sup>。

集成电路行业的制造工艺也成为器官芯片的推动力，这种方法使用光刻图案制造纳米级别的结构，推动微流体技术的驱动器和传感器向小型化发展，例如通过基于 PDMS 的蚀刻微通道配置软光刻实现大规模生产，并使芯片具有紧凑的尺寸和微通道，可以精确地形成细胞图案并操纵各种流体和化学参数<sup>[8]</sup>。自动化技术集成了一系列片上传感器，如光学传感器、温度传感器以及生物传感器，这使得对器官芯片的微环境实现非侵入式实时监测成为可能<sup>[9]</sup>。美国环境保护署制定了在 2035 年前逐步淘汰使用哺乳动物进行化学品安全性测试的计划，此后任何哺乳动物实验都需要管理员批准<sup>[10]</sup>，从而加速实验向计算机模拟和器官芯片等非动物模型转移<sup>[11]</sup>。

### 1.1.2 研究目的

本研究利用生物信息学、机器学习和器官芯片技术，通过分析放射性肺损伤芯片中与自噬作用相关的基因表达数据，构建肺腺癌风险评分模型，模型不依赖患者的基因数据和临床信息。本研究利用健康人类志愿者和肺腺癌患者样本的基因表达数据和临床信息作为测试集，验证了模型的准确性和可靠性，并对风险评分模型结果的相关生物信息学过程进行分析，以期建立一种全新的基于因果关系预测的肺腺癌预后评估模型。

### 1.1.3 研究意义

医学临床数据在实际使用中存在多重限制，部分非公开数据集在使用时需要通过繁琐且耗时的伦理审批，从而确保参与者的权利和隐私得到充分保护，这一过程通常包括研究目的和方法的评估，并且涉及数据保密性和安全性的严格要求。此外，患者的医疗信息包含高度敏感的个人数据，对其进行保护在法律和伦理上都十分重要。由于数据收集和记录过程中可能存在的偏差，部分临床数据集也难以保证其完整性和准确性。

针对这些问题，本文提出了一种不依赖传统临床样本的方法来构建预后模型，仅通过器官芯片的实验数据和公开的分子生物学信息完成分析，绕开传统临床数据使用中的障碍。本研究所构建的新型预后模型能更准确地识别出高风险的肺腺癌患者，为其提供个性化的治疗方案、改善预后效果，在一定程度上能为临床决策提供有力的数据支持，促进精准医疗的发展。

## 1.2 相关领域国内外研究现状

在医学领域，预后是指疾病或病症的可能过程和结果，是一种预测，以临床经验、流行病学数据和病情本身的内在性质的复杂综合为基础。预后评估通常包含一系列因素，包括个人的年龄、总体健康状况、病情的具体特征、对治疗的反应以及是否存在合并症。

机器学习方法也越来越受欢迎，可以缓解与遗传数据研究相关的困难挑战之一：提取重要且有意义的基因。所以，机器学习和基于统计的模型的整合可以增加该基因作为关键候选基因的证据和置信度<sup>[12]</sup>，利用机器学习对数据进行回归、聚类、降维等操作<sup>[13]</sup><sup>[14]</sup>，使用多个因子来综合评估建模，提高模型的精确性。近年来，医学大数据与人工智能的技术突破极大程度上推进了预后研究的发展，但是由于疾病之间的相互作用相当庞大和复杂，而医疗记录数据通常是有限的。此外，基于神经网络的算法<sup>[15][16]</sup>通常是黑盒算法，因为内部决策过于复杂，无法提供可解释性。

近年来，研究者们已经开发出基于多基因表达的风险评估模型，这些模型在预测不同类型肿瘤的预后方面显示出了广泛的应用潜力，其预后预测性能甚至优于组织病理学诊断和肿瘤分期<sup>[17]</sup>。同时，只针对筛选出的自噬相关基因进行建模，不仅能够降低模型

的复杂度，也在一定程度上缓解了对数据量的求。然而，这种方法仍然需要相当数量的临床样本作为训练集，获取这些数据仍有较大困难。

虽然动物模型已经被用于研究放射性肺损伤和肺癌的发生，但是它们未能模拟临床相关剂量敏感性并概括人类病理生理学的关键特征，价值动物模型成本高、建模时间长和严重的伦理问题，一定程度上限制了该领域的研究。器官芯片技术应运而生。通过使用包含多种细胞类型的肺体外模型可能提供放射性肺损伤更多人类相关的病理生理相关信息。基于器官芯片数据的预后模型的构建可能为肺腺癌的确诊、患病风险预估以及个性化的治疗策略提供全新的视角和应对方案。

## 1.3 本文主要研究内容和组织结构

### 1.3.1 本文主要研究内容

本文从基因表达总库（Gene Expression Omnibus, GEO）收集肺损伤芯片基因数据，从 Gene cards 数据库获取人体自噬相关基因。并对数据进行清洗，处理异常值和离群值。构建加权相关网络，将表达模式相似的基因进行聚类。对基因进行差异分析（Differential Expression Analysis, DEG），将不同处理下的基因表达数据同对照组做统计分析，对表达差异显著的基因进行初步筛选，并进行基因变异集分分析（Gene Set Variation Analysis, GSVA）。

通过进行单因素 COX 回归、最小绝对值收缩和选择算子（Least Absolute Shrinkage and Selection Operator, LASSO）回归筛选对预后有价值的基因，并对其进行多因素 COX 回归分析建立风险评分预后模型。将测试集中的患者分为高风险组和低风险组。采用 Kaplan-Meier 分析和对数秩（log-rank）检验比较两组生存时间的差异，采用受试者工作特性曲线（Receiver Operating Characteristic, ROC）评估模型在测试集中的预测精度。分别在所有自噬相关基因和建立风险评分模型的基因的基础上进行主成分分析（Principal Component Analysis, PCA），探索训练集中两个风险组之间的样本分布。在本研究中，对两个风险组的免疫细胞浸润情况及免疫相关功能进行分析，运用通过估算 RNA 转录本的相对子集识别细胞类型（Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts, CIBERSORT）方法对两组中的 22 种免疫细胞进行定量测定。

### 1.3.2 本文组织结构

第一章总体介绍本文的研究背景、研究目的和研究意义，系统分析国内外对于器官芯片和疾病预后模型的研究进展和方向。

第二章分析放射性肺损伤芯片的结构，进行实验数据的获取和清洗，并结合统计学方法检测并去除数据的批次效应，利用加权相关网络分析不同基因之间的表达关系，并

检测数据的离群值。利用基因差异分析和基因集变异分析筛选出与预后显著相关的基因。

第三章利用风险比例回归模型和 LASSO 回归挑选具有统计学意义的基因作为风险评分模型的自变量，并使用健康人类志愿者和肺腺癌患者的基因数据作为测试集，测试模型的受试者工作特性曲线。根据风险评分模型将肺腺癌患者划分为高风险组和低风险组，并对两个风险组进行 log-rank 检验。

第四章对模型结果的生物信息学过程进行分析，分析高低风险组在不同特征维度上的分布差异，以及免疫细胞在两个风险组中的分布情况。

第五章综述本文的核心研究内容，对研究中存在的不足进行分析，并展望未来研究的方向。

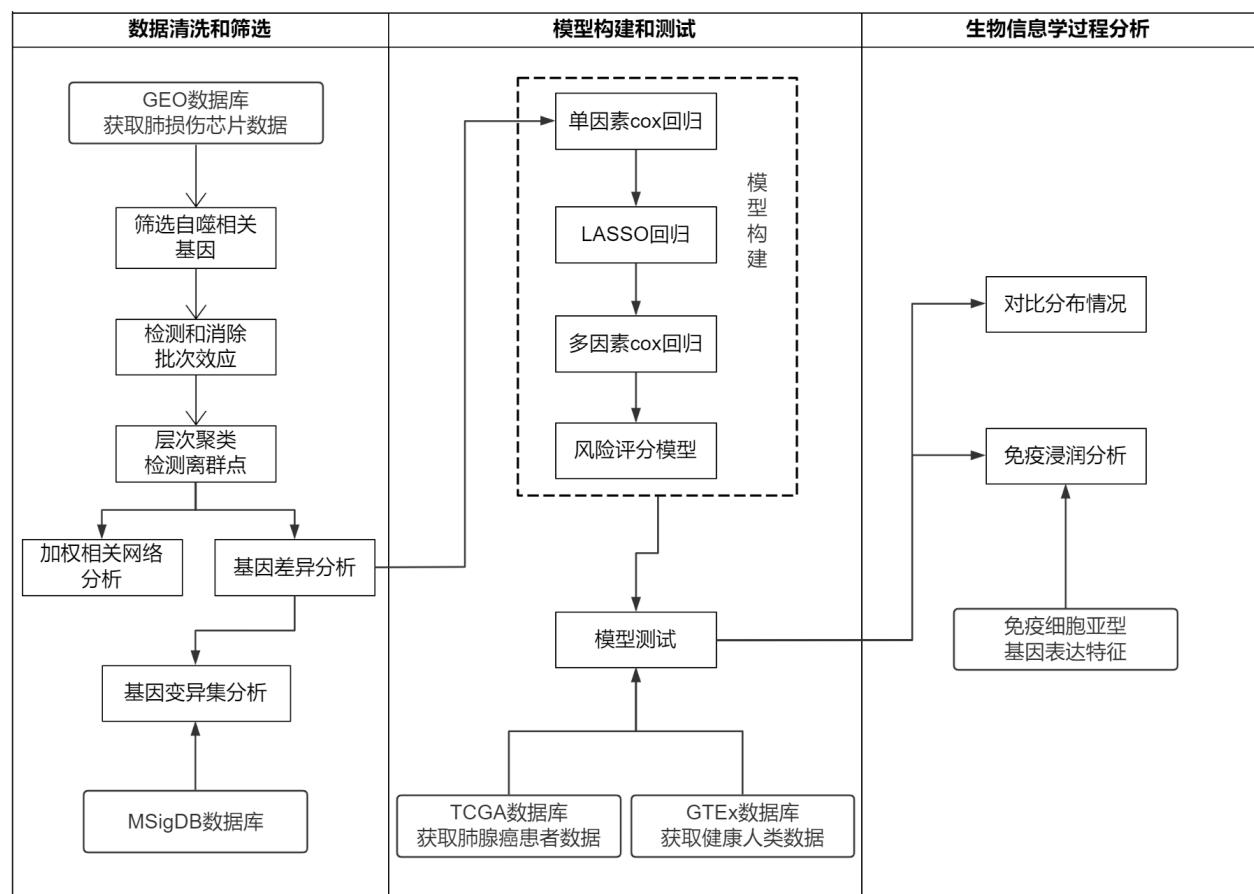


图 1.1 本文组织框图

## 第 2 章 放射性肺损伤芯片数据获取筛选

### 2.1 数据处理

#### 2.1.1 基因表达数据获取和清洗

本文使用的数据来源于受到放射性损伤的器官芯片，肺泡芯片是一种在体外模拟人体肺部结构和功能的装置，一侧排列着人肺微血管内皮细胞，用于模拟肺部微血管的结构，另一侧排列着人原代肺泡上皮细胞，用于模拟气体交换。实验气液界面还包括循环机械装置模拟呼吸运动<sup>[18]</sup>，芯片的示意图如图 2.1 所示。设置两组实验：将芯片放置在 16Gy 的辐射和无辐射条件下下持续 6 小时，以及将芯片放置在 16Gy 的辐射和无辐射条件下下持续 7 天。当这种肺芯片暴露于临床相关剂量的辐射时，可以观察到急性放射性肺损伤的许多生理特征。

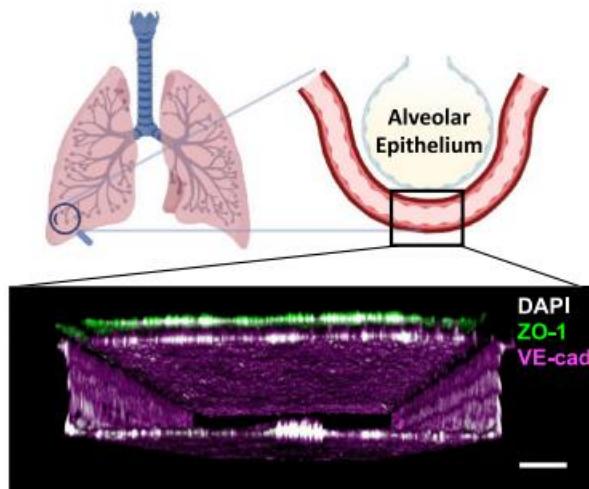
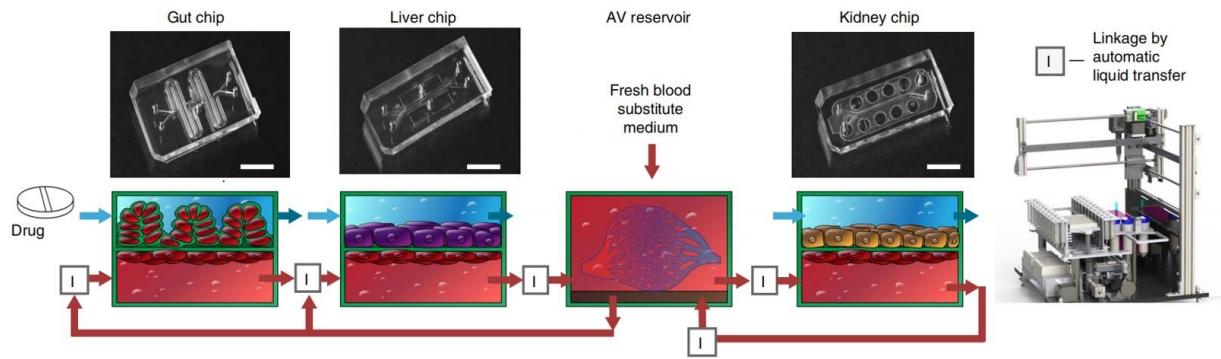


图 2.1 肺芯片示意图<sup>[18]</sup>

为了提高实验模拟人体微环境的精度，检测和分析其他器官在实验中的状态，还可以将多个不同的器官芯片通过流体耦合血管化技术连接起来，构成一个仿真系统，一种连接方式如图 2.2 所示。

这样组成的系统可以通过微流控技术对每个器官芯片实现“血液替代品”的灌注，从而模拟人体的血液循环，进而模拟各种器官之间的相互作用。

本文使用的数据为 GEO 数据库中的肺损伤芯片的转录组测序数据 GSE242840 和 GSE242706，使用“stringr”包和“dplyr”包对数据进行初步清洗，滤除在所有样本中 count 计数均为 0 的低通量基因。

图 2.2 通过流体耦合血管化连接器官芯片<sup>[19]</sup>

### 2.1.2 批次效应的检查和去除

批次效应指的是由于实验条件不同而产生的数据的系统性差异。这种效应极有可能掩盖真正的生物学差异，干扰最终的分析结果，并降低实验的可重复性以及可解释性。由于数据来源于两个批次，所以需要检测数据是否存在批次效应。为了观察两个批次数据的分布情况，首先使用“FactoMineR”包进行 PCA 降维，如图 2.3 所示。

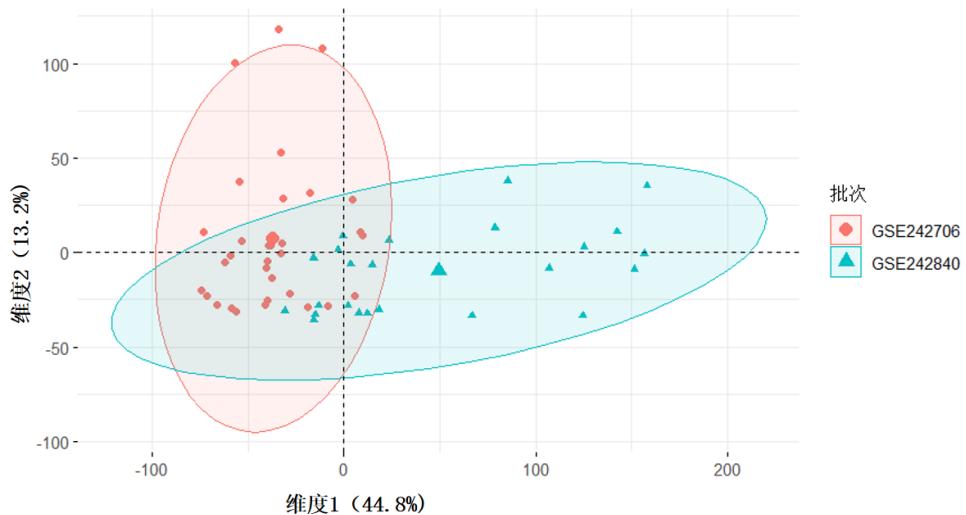


图 2.3 批次数据 PCA 降维后的分布

GSE242840 和 GSE242706 的数据分布出现明显的偏移，所以需要去除其中的批次效应。

将第  $i$  批次样本  $j$  的基因  $g$  的表达量  $Y_{ijg}$  表示为式 (2.1) <sup>[20]</sup>:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg} \quad (2.1)$$

其中  $\alpha_g$  为基因  $g$  在所有样本中的平均表达量， $X$  代表实验设计所制定的矩阵， $\beta_g$  为

对  $X$  矩阵的回归系数，代表不同设计矩阵各个因素下的效应值， $\gamma_{ig}$  代表批次效应的叠加

效应， $\delta_{ig}$ 代表批次效应的乘法效应<sup>[20]</sup>。

式(2.2)为对式(2.1)其进行标准化后的结果。

$$Z_{ijg} = \frac{Y_{ijg} - \hat{\alpha}_g - X\hat{\beta}_g}{\hat{\sigma}_g} \quad (2.2)$$

对于连续型随机变量，可以假定其服从正态分布，可以使用贝叶斯后验来对这些参数进行估计。

$$\begin{aligned} \gamma_{ig} &\sim N(Y_i, \tau_i^2) \\ \delta_{ig}^2 &\sim \text{InverseGamma}(\lambda_i, \theta_i) \end{aligned} \quad (2.3)$$

基因表达量的连续变量形式可以采用式(2.3)的方法进行参数估计，而 count 格式的基因表达量数据为整数，属于离散型随机变量，可以假设其服从负二项分布，如式(2.4)所示。

$$\gamma_{ig} \sim N(\mu_{gij}, \phi_{gi}) \quad (2.4)$$

其中 $\mu_{gij}$ 表示均值， $\phi_{gi}$ 表示分散系数，因此，我们可以建立式(2.5)所示模型。

$$\begin{aligned} \log \mu_{gij} &= \alpha_g + X_j \beta_g + \gamma_{gi} + \log N_j \\ \text{var}(y_{gij}) &= \mu_{gij} + \phi_{gi} \mu_{gij}^2 \end{aligned} \quad (2.5)$$

其中 $\alpha_g$ 表示“阴性”样本计数的期望值的对数， $X_j \beta_g$ 表示由于生物条件改变而对计数期望对数的影响， $X_j$ 可以是样本 J 的生物条件指标， $\beta_g$ 表示对应的回归系数， $N_j$ 表示数据库的大小，即样本 J 所有基因的总数。

根据式(2.5)的模型，可以得到批次效应参数的估计值 $\hat{\gamma}_{gi}$ 和 $\hat{\phi}_{gi}$ ，以及 count 计数均值 $\hat{\mu}_{gij}$ 的拟合期望。假设调整后的数据遵循“无批次效应”的负二项分布 $NB(\mu_{gj}^*, \phi_g^*)$ ，其参数计算公式为式(2.6)。

$$\begin{aligned} \log \mu_{gj}^* &= \log \hat{\mu}_{gij} - \hat{\gamma}_{gi} \\ \phi_g^* &= \frac{1}{N_{batch}} \sum_i \hat{\phi}_{gi} \end{aligned} \quad (2.6)$$

通过寻找与原始数据 $\gamma_{ig}$ 分位数最接近的无批效应分布分位数来计算调整后的数据 $\gamma_{gij}^*$ ，其被估计为 $NB(\mu_{gjj}^*, \phi_g^*)$ 。调整后的 $\gamma_{gij}^*$ 使得 $F^*(y_{gj}^*) = P(y^* \leq y_{gj}^*)$ 最接近 F

$(y_{gij}) = P(y \leq y_{gij})$  的绝对值。对 count 矩阵中的每个值都执行这种映射，从而完成使用“sva”包中的 ComBat\_Seq 实现去批次效应的调整操作。

数据进行去批次效应后的分布图如图 2.4 所示。

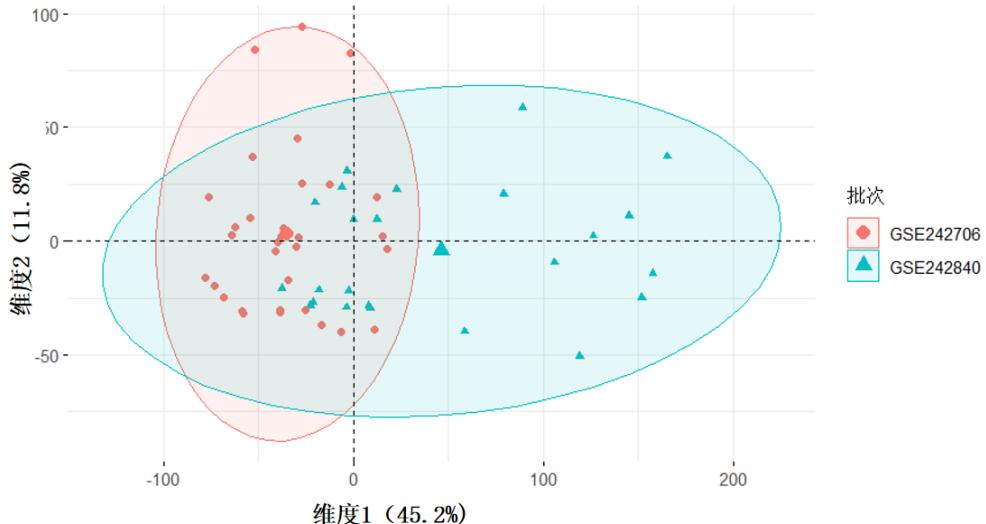


图 2.4 去除批次效应的数据 PCA 降维后的分布

去除批次效应后的两个批次的数据重叠部分面积明显增加，数据去除了与生物状态不相关的干扰效应，只保留放射性损伤所带来的影响。

### 2.1.3 离群值检测

在数据挖掘和统计中，层次聚类的目的是构建聚类层次结构，其合并过程是由贪心算法确定的，即选取欧氏距离最近的两个元素进行合并，在多个点之间的最小距离相等时，则随机选取一对数据点进行合并。在本文中，每个数据点代表一个样本的各个基因表达量。合并过程不断迭代，直到形成一个完整的聚类层次。欧氏距离越短，表明数据点间的相似度越高。通过这种方式，算法逐步构建出一个反映数据点相似性的聚类树。

使用层次聚类的方法来检测样本中的离群值，如果存在样本距离最近的聚类簇过远，则该样本可以被视为一个离群点，所有样本的层次聚类数如图 2.5 所示。图中每个样本距离最近的聚类簇的距离都比较接近，不存在明显的离群点。

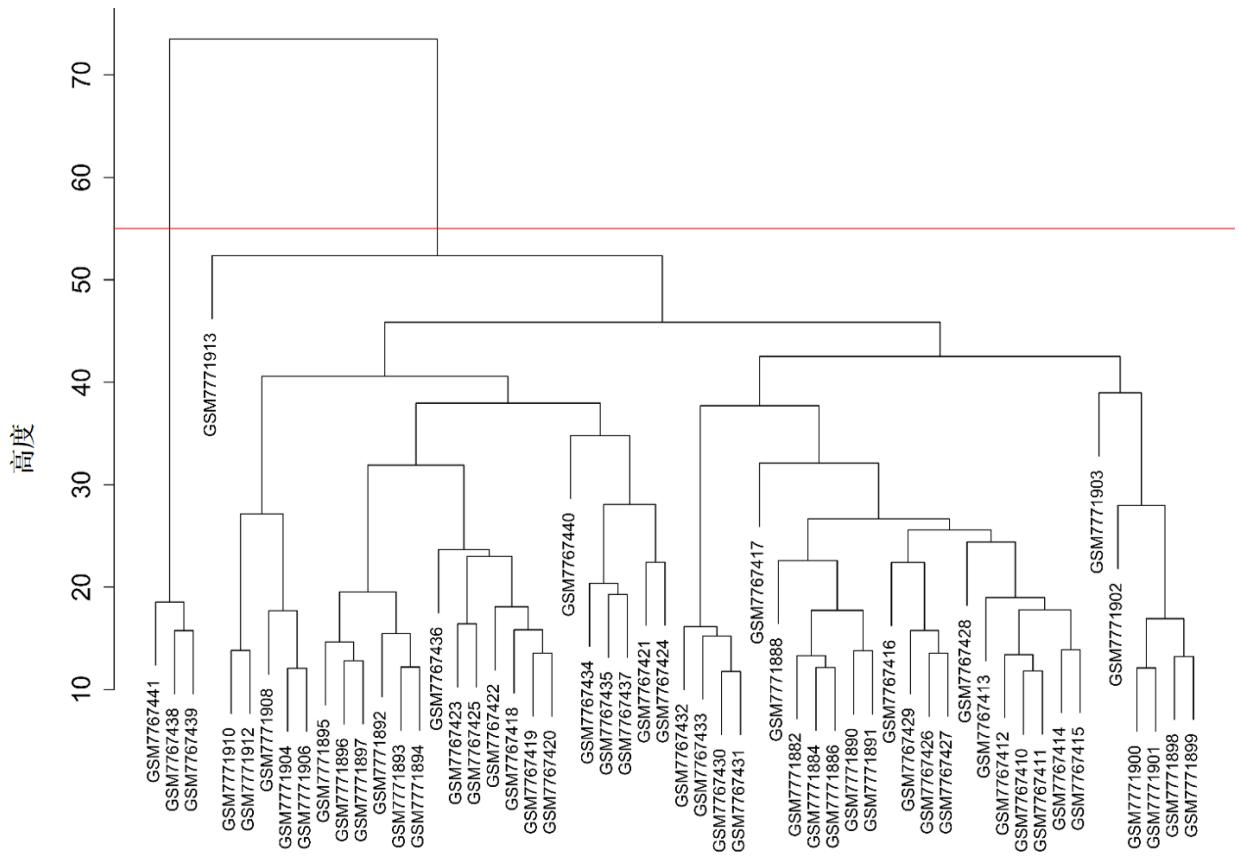


图 2.5 样本层次聚类树

## 2.2 构建加权相关网络

加权相关网络分析，也称为加权基因共表达网络分析，基础是根据基因表达之间的相关性强度构建网络，通过有符号或无符号度量进行量化，利用拓扑特性计算、数据模拟和与外部软件的接口，来分析和解释基因表达的复杂相互作用。

加权相关网络分析还可以研究共表达模块间的关系、比较不同网络的拓扑结构，并进行数据提炼、聚类分析、特征选择，以及综合分析互补的基因组数据<sup>[21]</sup>。加权相关网络分析的网络方法适合于整合互补的基因组数据，使得基于网络的元分析技术成为可能。

将 count 数据转化成 TPM 格式后转化为相关系数矩阵，转化方式为式 (2.7)。

$$\begin{aligned} S_{BC}^{\text{unsigned}} &= |\text{cor}(x_B, x_C)| \\ S_{BC}^{\text{signed}} &= 0.5 + 0.5\text{cor}(x_B, x_C) \end{aligned} \quad (2.7)$$

其中  $\text{cor}(x_B, x_C)$  表示  $x_B$  和  $x_C$  的相关系数， $S_{BC}^{\text{unsigned}}$  表示不区分正相关和负相关的转化值， $S_{BC}^{\text{signed}}$  表示区分正相关和负相关的转化值，本模型的转化值采用  $S_{BC}^{\text{signed}}$ 。

得到相关系数矩阵后再将其转化为邻接矩阵，转化关系为式 (2.8)。

$$a_{ij} = \text{power}(s_{ij}, \beta) \equiv |s_{ij}|^\beta \quad (2.8)$$

在加权相关网络分析中，人为设定的阈值可能引入主观偏差，所以引入了软阈值方法。软阈值通过使用幂函数将相关性矩阵转换为邻接矩阵，其中需要确定幂参数 $\beta$ 。该网络需要更接近于无尺度网络的特性，且在此过程中应尽可能地保留原始数据的连通性信息<sup>[22]</sup>。加权相关网络分析使用拓扑重叠指标作为距离，其综合了两个基因之间的直接相互作用以及它们与网络中其他基因的相互作用强度。这一指标是评估网络中基因相互连接性的一个稳健方法。通过将拓扑重叠指标作为输入，应用平均连锁层级聚类方法，可以对基因进行系统的分类。采用动态树切割方法处理聚类结果，从而确定网络中具有相似表达模式的基因集合<sup>[23]</sup>。该部分使用“WGCNA”包和“DESeq2”包完成。

图 2.6 左侧纵坐标是无尺度网络的评价指标 $r^2$ ，图 2.6 右侧纵坐标是平均连通性，通常选择 $r^2$ 首次达到 0.9 时的 $\beta$ 值。通过网络拓扑分析确定软阈值，结果如图 2.6 所示，图中左侧红色横线标记了无尺度网络的评价指标 $r^2$ 为 0.9 的位置，其上方第一个点的 $\beta$ 取值为 12。

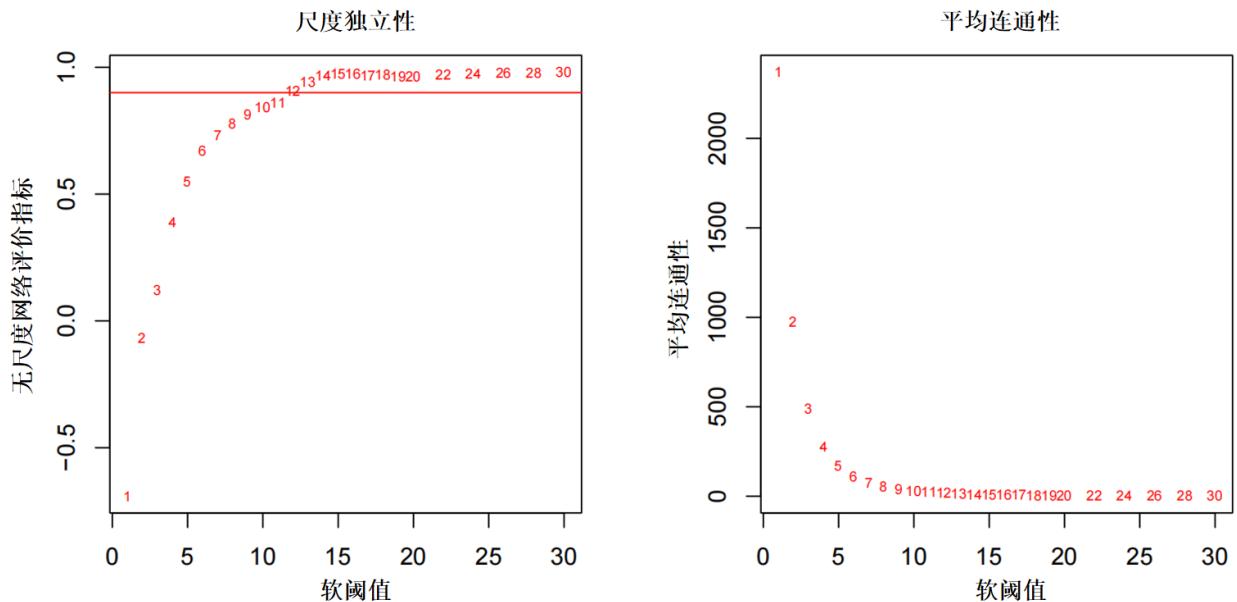


图 2.6 无尺度网络评价指标（左）平均连通性（右）

将结果带入式 (2.8)，即可将相关性矩阵转换为邻接矩阵。相关系数主要用于衡量两个变量之间的线性关系，而拓扑重叠矩阵（Topological Overlap Matrix, TOM）则是考虑到中间节点连接关系的间接计算方法，能够揭示网络中节点的整体连接模式，可以更全面地理解基因表达数据的复杂交互作用。由邻接矩阵构建拓扑重叠矩阵的方法为式 (2.9)。

$$\text{TOM}_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} \quad (2.9)$$

根据拓扑重叠矩阵对基因进行聚类，结果如图 2.7 所示。其中下方第一条彩色条带表示通过动态树切割方法确定的模块划分，第二条彩色条带表示以树高阈值为 0.25 进行相似模块合并后的结果。

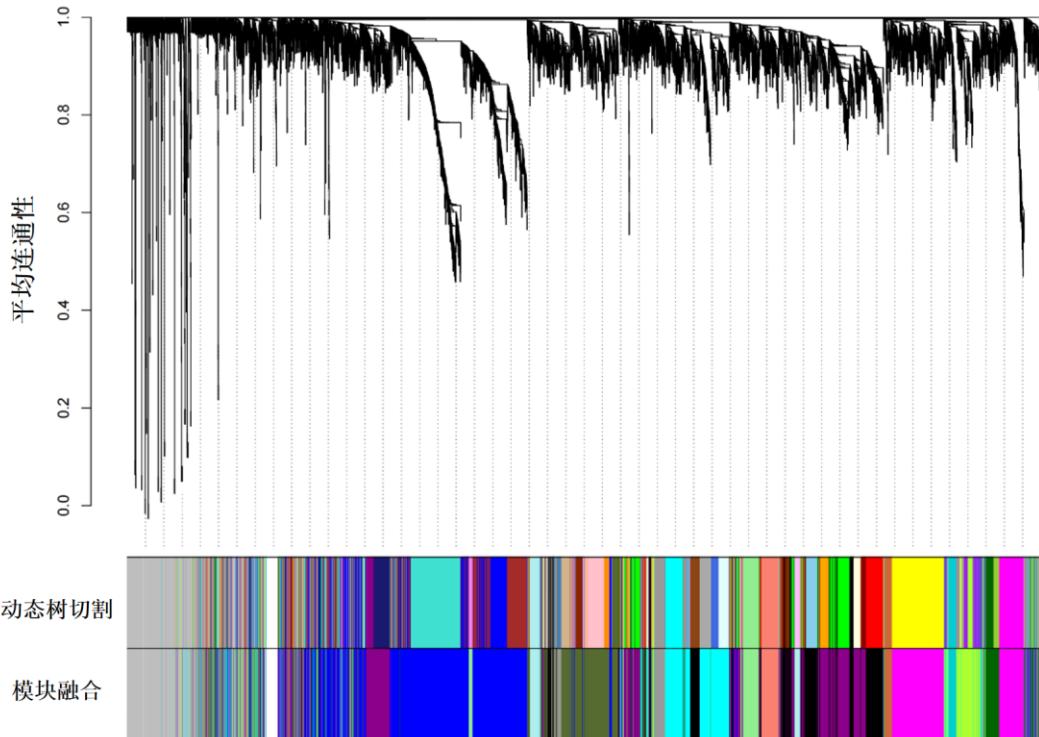


图 2.7 基因聚类及模块融合结果

### 2.3 筛选差异表达基因

首先利用实验信息对数据进行标注，GSE242840 数据中受辐射损伤和未受辐射损伤的样本各有 12 例，GSE242706 数据中受辐射损伤和未受辐射损伤的样本各有 16 例。将辐射损伤的数据标注为“Treat”，未受辐射损伤的数据标注为“CK”，设置此处的对比矩阵  $C=[1, -1]$ 。由式 (2.1) 可知，本研究中的基因表达量为线性模型，所以对其进行线性拟合，其回归系数的估计系数的标准误差为式 (2.10)。

$$\sigma_{N_j} = \frac{\sigma}{\sqrt{N_j}} \quad (2.10)$$

其中  $\sigma$  表示系数的标准差，可以表示为式 (2.11)。

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N_j} (\beta_{jg} - \bar{\beta}_g)^2}{N_j}} \quad (2.11)$$

利用经验贝叶斯矫正估计的参数及其标准误差<sup>[24]</sup>，修正后的系数标准误差更加稳定，即使是基于少量样本的估计也更可靠。因此，使用经验贝叶斯方法可以提高统计测试的功效。

基因表达的差异倍数 Fold Change 的计算方法为式 (2.12)。

$$FC = \frac{\bar{x}_{\text{Treat}}}{\bar{x}_{\text{CK}}} \quad (2.12)$$

识别两种处理下表达差异显著的基因，使用式 (2.13) 所示的 T 检验方法计算 T 统计量。

$$t = \frac{\bar{x}_{\text{Treat}} - \bar{x}_{\text{CK}}}{\sqrt{\frac{s_{\text{Treat}}^2}{n} + \frac{s_{\text{CK}}^2}{n}}} \quad (2.13)$$

使用“limma”、“dplyr”和“edge”包完成差异分析，并根据 t 分布计算显著性 p 值，设置差异表达比值  $|\log_2 FC| > 0.5$ ，矫正后的显著性水平  $p < 0.05$ ，得到 278 个上调基因和 75 个下调基因。使用“ggplot2”和“ggrepel”包绘制其火山图，如图 2.8 所示，图中上调基因被标记为右侧红色散点，下调基因被标记为左侧蓝色散点。

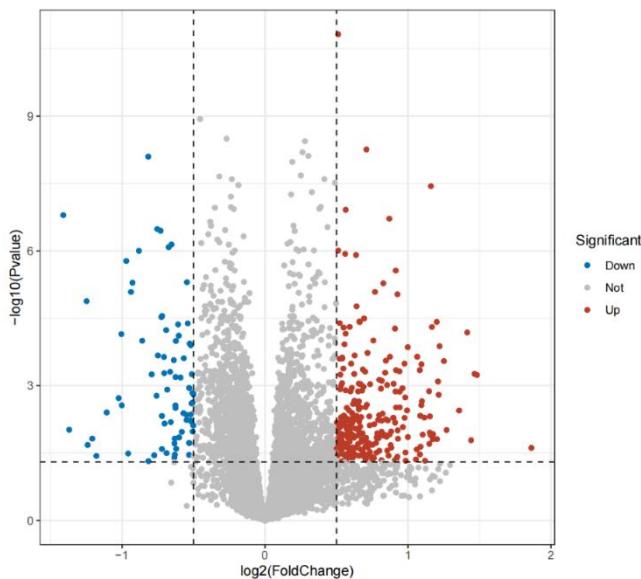


图 2.8 差异基因火山图

使用“pheatmap”包绘制上调基因和下调基因的热图，如图 2.9 所示。

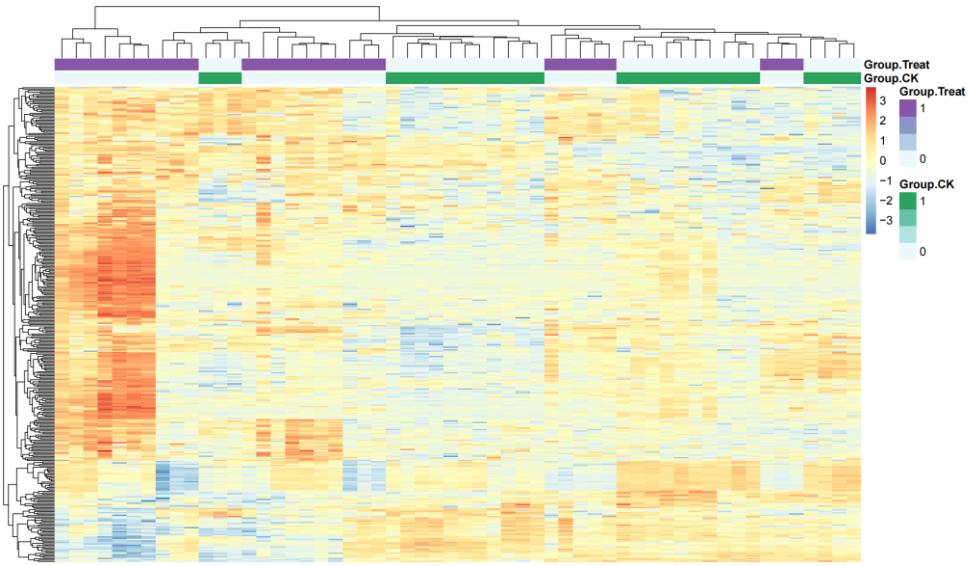


图 2.9 差异表达基因热图

## 2.4 基因富集

基因集变异分析用于估计样本群体中基因集的变异性或活跃度。与传统的单基因分析方法不同，GSVA 将分析的焦点从单个基因转移到基因集，如生物学通路或功能类别，从而提供了一种在样本间评估基因集富集的方法。

首先利用泊松核函数计算 count 数据的表达程度  $\hat{F}_r$ <sup>[25]</sup>，如式 (2.14) 所示。

$$Z_{jg} = \hat{F}_r(Y_{jg}) = \frac{1}{n} \sum_{k=1}^M \sum_{y=0}^{Y_{jg}} \frac{e^{-(Y_{gk}+r)} (Y_{gk} + r)^y}{y!} \quad (2.14)$$

式中 M 表示样本总数，r 取 0.5，对  $Z_{jg}$  进行排序，并将排序后的结果记为  $Z_{j(g)}$ ，并定义  $r_{jg} = |p/2 - z_{j(g)}|$ ，由此得到富集分数  $v_{jk}$ ，如式 (2.15) 所示。

$$v_{jk}(\ell) = \frac{\sum_{g=1}^{\ell} |r_{jg}|^\tau I(h_{(g)} \in \xi_k)}{\sum_{g=1}^p |r_{jg}|^\tau I(h_{(g)} \in \xi_k)} - \frac{\sum_{g=1}^{\ell} I(h_{(g)} \notin \xi_k)}{p - |\xi_k|} \quad (2.15)$$

式中  $\tau$  表示随机游走尾部权重的参数， $\xi_k$  表示第 k 个基因集， $|\xi_k|$  表示该基因集中的基因总数， $I(h_{(g)} \in \xi_k)$  表示基因是否属于  $\xi_k$  的指示函数，p 表示数据集的基因总数。

根据富集分数能够计算出每个基因集的 GSVA 分数  $ES_{jk}^{\text{diff}}$ ，如式 (2.16) 所示。

$$ES_{jk}^{\text{diff}} = |ES_{jk}^+| - |ES_{jk}^-| = \max_{\ell=1,\dots,p} (0, v_{jk}(\ell)) - \min_{\ell=1,\dots,p} (0, v_{jk}(\ell)) \quad (2.16)$$

## 2.4.1 GO 功能富集

基因本体论 (Gene Ontology, GO) 是计算生物学领域的一个综合框架，旨在代表我们对多种生物体内基因功能的集体认知。GO 富集分析利用基因本体分类系统解释基因集，根据功能特征将基因分配到预定义的分区中。这种功能分析可用于阐明与癌变状况相关的潜在细胞机制。

首先使用“msigdb”包从 MSigDB 数据库下载 C5 基因集，使用“GSVA”包和“GSEABase”包进行分析，得到火山图如图 2.10 所示，共计有 20 个上调基因集和 34 个下调基因集。其中 GOBP\_DCMP\_CATABOLIC\_PROCESS、GOBP\_DUMP\_CATABOLIC\_PROCESS、GOBP\_UMP\_CATABOLIC\_PROCESS 等都与细胞自噬相关，说明自噬相关基因对模型有显著影响。

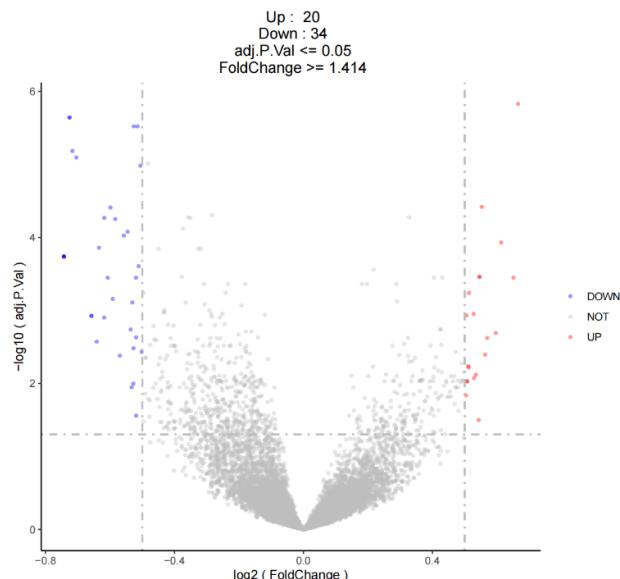


图 2.10 GO 富集差异火山图

GSVA 分数的热图如图 2.11 所示，在首次聚为 2 类时，其中一类全部为对照组，另一类中包含了全部 28 组实验组和 6 组对照组，说明实验组和对照组在基因集表达上具有显著差异。

## 2.4.2 KEGG 通路富集

京都基因与基因组百科全书 (Kyoto Encyclopedia of Genes and Genomes, KEGG) 富集分析用于识别在特定生物学条件下显著富集的代谢通路。这种分析利用 KEGG 数据库，可以使用 MSigDB 中的 C2 基因集。类似地进行 KEGG 通路富集，结果如图 2.12 所示。可以发现信号通路主要聚集在 DNA 复制、细胞生长和代谢等通路相关。

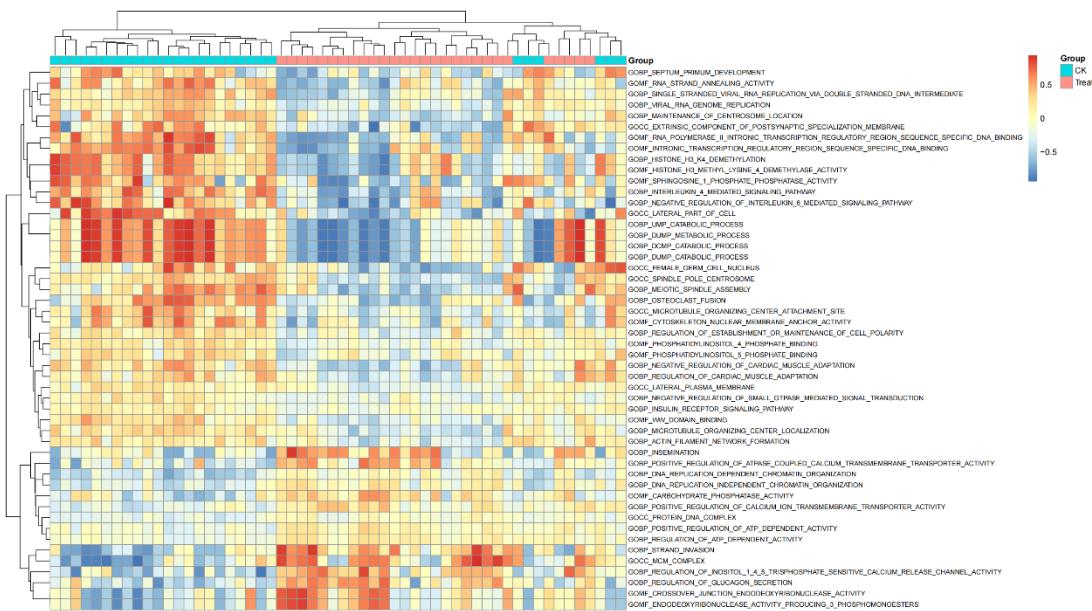


图 2.11 GO 富集热图

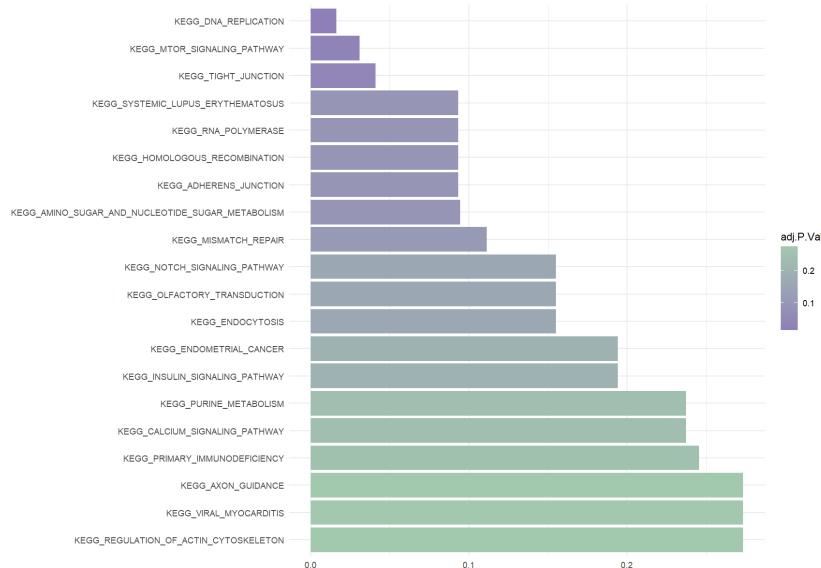


图 2.12 KEGG 富集分析

## 2.5 本章小结

本章对放射性肺损伤芯片数据进行预处理和分析。首先，从 GEO 数据库获取肺损伤芯片转录组数据，对其进行清洗，并通过统计学方法去除批次效应。利用层次聚类检测离群值，并构建加权相关网络，以识别基因表达的复杂相互作用。通过线性拟合和经验贝叶斯方法筛选差异表达基因。最后，进行基因富集分析，揭示实验组和对照组的显著差异。这些步骤确保了数据的质量，为深入分析奠定了基础。

## 第3章 风险评分模型的建立与测试

### 3.1 风险评分模型的建立

#### 3.1.1 单因素 COX 回归模型

生存分析的核心目标是模拟和描述从时间到事件的数据，分析从一个明确定义的时间起点到一个或多个事件发生的持续时间。生存模型将某些事件发生之前经过的时间与可能与该时间量相关的一个或多个协变量相关联。比例风险模型的协变量中单位增加的独特效应相对于风险率是乘法的。

通常将风险率函数表示为式 (3.1) [26]。

$$h(t, X) = h_0(t) \cdot f(X) \quad (3.1)$$

其中  $h(t, X)$  表示  $t$  时刻的风险率函数， $h_0(t)$  表示  $t$  时刻的基准风险率函数， $f(X)$  表示协变量函数。

COX 比例风险回归模型被称为半参数模型，其基准风险率函数未知，可以表示为式 (3.2)。

$$h(t, X) = h_0(t) \cdot e^{(\beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m)} \quad (3.2)$$

其中  $\beta_1, \beta_2, \dots, \beta_m$  为自变量的偏回归系数。除  $h_0(t)$  分布无明确的假定外，其余参数可以通过样本的实际观察来估计。

对式 (3.2) 进行对数变换可以得到式 (3.3)。

$$\ln \left[ \frac{h(t, X)}{h_0(t)} \right] = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m \quad (3.3)$$

在单因素情况下，只有一个协变量  $x$ ，COX 回归模型可以被表示为式 (3.4)。

$$h(t, X) = h_0(t) \cdot e^{(\beta x)} \quad (3.4)$$

通过式 (3.4) 可以得到暴露与未暴露的相对危险度为式 (3.5)。

$$RR = \frac{h(t, X | x = 1)}{h(t, X | x = 0)} = \frac{h_0(t) \cdot e^{(\beta \times 1)}}{h_0(t) \cdot e^{(\beta \times 0)}} = e^\beta \quad (3.5)$$

显然  $\beta_j > 0$  则  $PR > 0$ ，表示该因素为危险因素； $\beta_j = 0$ ，则  $PR = 0$ ，表示该因素为无关因素； $\beta_j < 0$ ，则  $PR < 0$ ，表示该因素为保护因素。

COX 回归模型与一般传统的回归分析不同，模型的重点是危害函数，即在个体存活

至时间  $t$  时，事件在时间  $t$  发生的瞬时风险，而传统回归模型通常不对危险函数建模。在参数估计完成后，可以对这两个函数进行估计，并据此计算出每个时间点的生存率。

进行 COX 回归分析需要满足两个假设条件，即比例风险假设和对数线性假设。由于基准风险函数的分布不确定，所以不能使用一般的最大似然估计模型的参数，应使用偏似然估计。模型的假设检验通常采用 Wald 检验，其统计量为式 (3.6)。

$$\chi^2 = \left( \frac{\beta_j}{S_{\beta_j}} \right)^2 \quad (3.6)$$

通过式 (3.6) 可以检验模型中的协变量是否应具有统计学意义以及是否该被剔除。

常见的 COX 回归模型的因变量采用患者的生存时间作为回归的因变量，但由于本模型所使用的数据为微流控器官芯片的测序数据，缺少临床的生存时间，因而将回归的因变量设置为受到放射性射线照射的时间。使用“survival”包，在显著性水平  $\alpha=0.05$  上进行检验，共计得到  $p$  值小于 0.05 的基因 258 个，回归结果的部分森林图如图 3.1 所示。

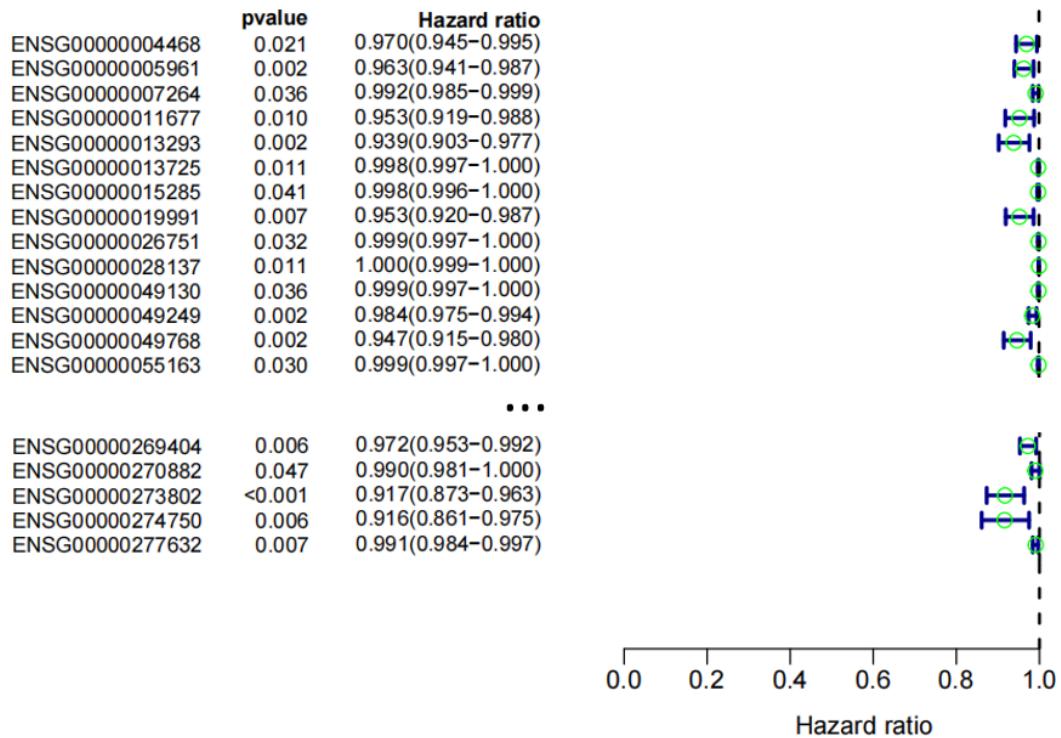


图 3.1 单因素 COX 回归森林图（部分）

### 3.1.2 LASSO 回归模型

由于单因素 COX 回归分析不考虑变量之间的相关性，使用 LASSO 回归进行进一步筛选<sup>[27][28]</sup>。LASSO 回归则可以在考虑变量相关性的同时进行特征选择，通过对系数加

上正则化项 L1 惩罚来推动部分系数收缩至零，从而实现自动特征选择，剔除不重要的变量，减少模型复杂度。且在实践中，单因素 COX 回归可能会出现过度拟合的情况，而 LASSO 回归减少模型的参数，增加其泛化能力。LASSO 回归与岭回归的区别在于正则化项的不同，其在目标函数  $Q(\beta)$  后添加 1-范数，即式 (3.7) [29]。

$$Q(\beta) = \|y - X\beta\|^2 + \lambda \|\beta\|_1 \quad (3.7)$$

这种正则项使得 LASSO 回归可以将系数压缩为 0，1-范数为向量内各元素的绝对值之和，可以表示为式 (3.8)。

$$\|X\|_1 = \sum_{i=1}^n |x_i| \quad (3.8)$$

而岭回归所采用的是 2-范数也被称为欧几里得范数，为开方后的元素的平方和，如式 (3.9) 所示。

$$\|X\|_2 = \left( \sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}} = \sqrt{\sum_{i=1}^n x_i^2} \quad (3.9)$$

LASSO 回归的优化算法模型可以写作式 (3.10)。

$$\begin{aligned} & \arg \min \|y - X\beta\|^2 \\ & \text{s.t. } \sum |\beta_j| \leq s \end{aligned} \quad (3.10)$$

式 (3.10) 的约束条件不是连续可导的，导致不能使用常规的机器学习算法进行求解，首先初始化位置点，如式 (3.11) 所示。

$$\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_p^{(0)}) \quad (3.11)$$

坐标轴下降算法不需要损失函数的导数。在 p 维的条件下，参数  $\theta$  为一个 p 维向量，只改变其中的 1 个参数，固定其余的 p-1 个参数，寻找使得损失函数  $J(\theta)$  达到最小的点，对 p 个参数都进行一次寻优，迭代过程如式 (3.12) 所示。

$$\begin{aligned} \theta_1^{(k)} &= \arg \min_{\theta_1} J(\theta_1, \theta_2^{(k-1)}, \theta_3^{(k-1)}, \dots, \theta_p^{(k-1)}) \\ \theta_2^{(k)} &= \arg \min_{\theta_2} J(\theta_1^{(k)}, \theta_2, \theta_3^{(k-1)}, \dots, \theta_p^{(k-1)}) \\ \theta_3^{(k)} &= \arg \min_{\theta_3} J(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3, \dots, \theta_p^{(k-1)}) \\ &\dots \\ \theta_p^{(k)} &= \arg \min_{\theta_p} J(\theta_1^{(k)}, \theta_2^{(k)}, \theta_3^{(k)}, \dots, \theta_p) \end{aligned} \quad (3.12)$$

如果所有参数 $\theta_j^{(k)}$ 的改变量都小于预先设定的阈值则结束计算，否则继续迭代直至符合条件。经过上述迭代后，必然能够得到一个局部最优解，又由于损失函数 $J(\theta)$ 为凸函数，其任意局部最优解必定是全局最优解。

使用“glmnet”包和“foreign”包进行 LASSO 回归分析，得到随 $\log(\lambda)$ 增加，各个系数的变化曲线如图 3.2 所示。可以看到随着 $\lambda$ 增加，部分变量系数变为 0，即实现了参数数量的压缩，对数据实现进一步筛选。LASSO 回归的交叉验证如图 3.3 所示，从图中能得到模型的均方根误差变化情况，其中左侧竖线标记均方根误差最小时的 $\log(\lambda)$ 取值，随 $\log(\lambda)$ 增加，此时的模型的非零参数个数为 41；右侧竖线标记均方根误差最小值加一个标准差时的 $\log(\lambda)$ 取值，此时的模型的非零参数个数为 30。

选择模型均方根误差最小时的参数取值，使用“pROC”包可以绘制 LASSO 回归对于芯片数据的 ROC 曲线如图 3.4 所示。ROC 曲线是每个阈值设置下真阳性率与假阳性率的关系图。ROC 曲线下面积（Area Under the Curve, AUC）越大，表示分类器的准确性越高，性能越好。AUC<0.6 代表低区分度，0.6~0.75 代表中等区分度，AUC>0.75 代表高区分度，本模型 AUC 为 0.796>0.75，说明模型数据呈现高区分度。

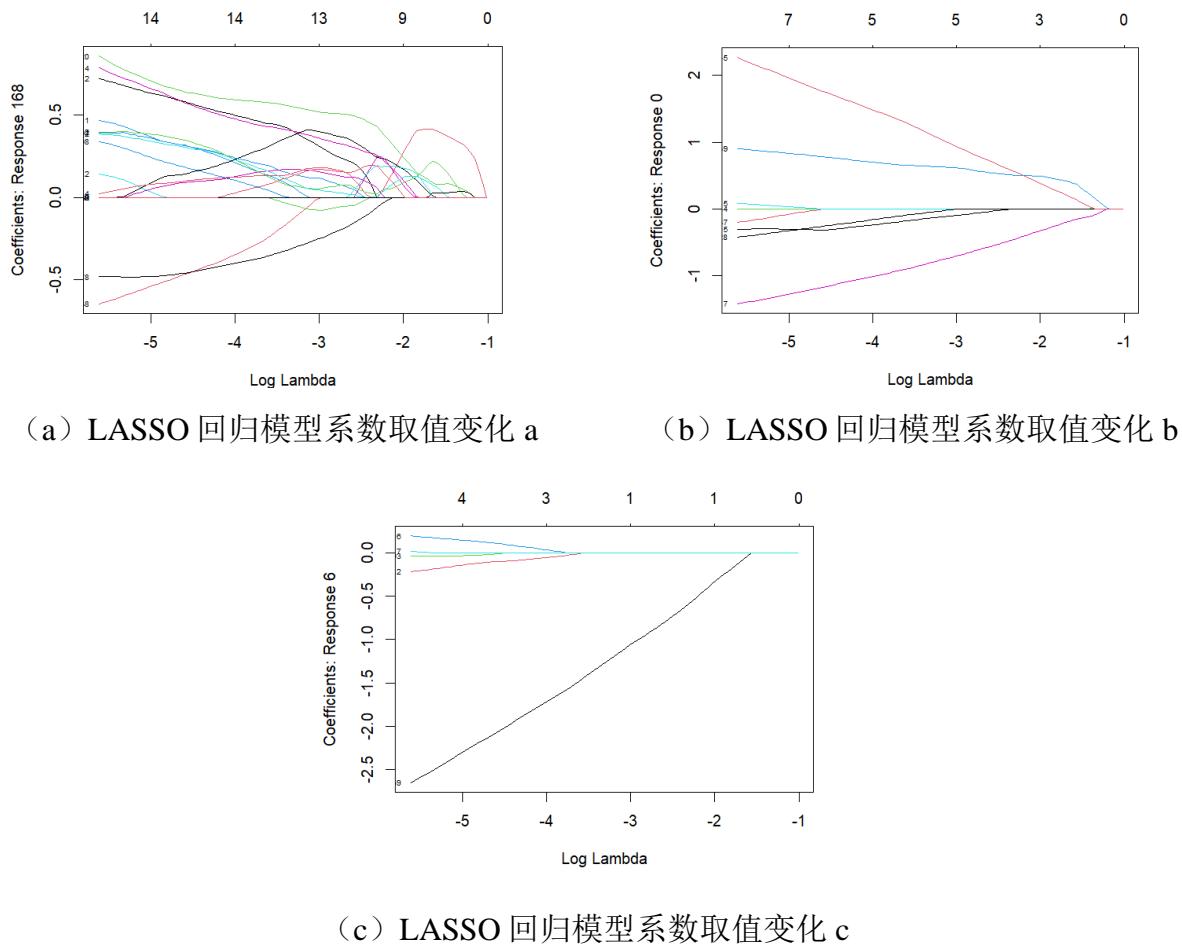


图 3.2 LASSO 回归模型系数取值随 $\log(\lambda)$ 变化图

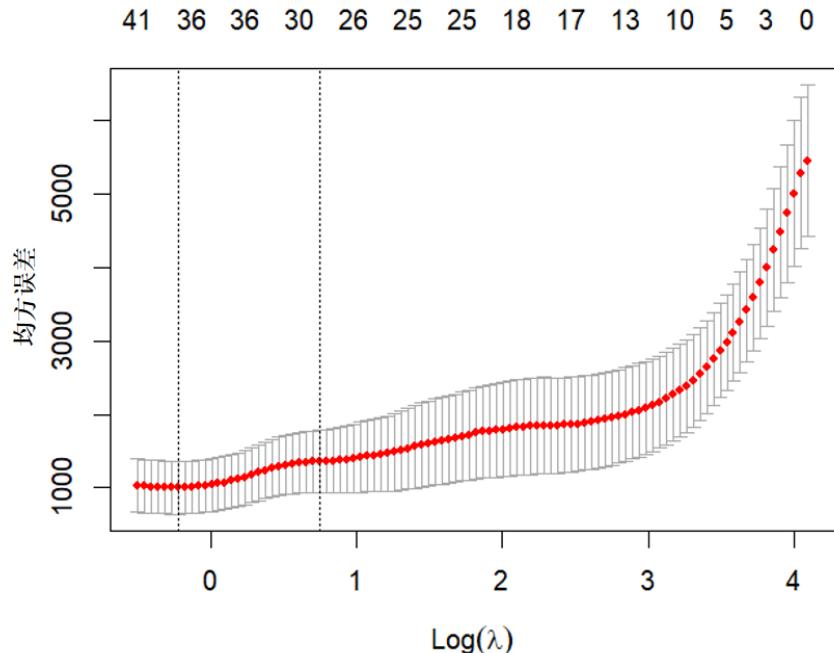


图 3.3 LASSO 回归交叉验证图

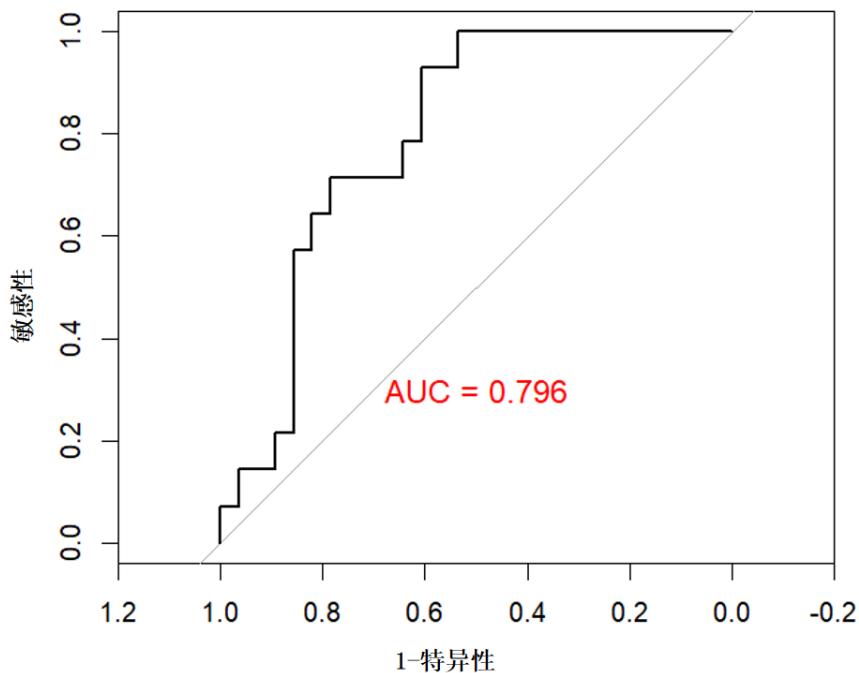


图 3.4 LASSO 回归 ROC 曲线

### 3.1.3 多因素 COX 回归模型

选取 LASSO 回归中的最佳建模基因进行多因素 COX 回归，其原理与单因素 COX 回归类似，基本形式可以表示为式 (3.1)，回归的结果如图 3.5 所示。可以发现在单因素

回归中  $p\text{-value} < 0.05$  的绝大多数基因的多因素回归  $p\text{-value}$  大于 0.05。导致这种现象的原因是两种 COX 回归中  $p\text{-value}$  的意义略有不同，多因素回归中的  $p\text{-value}$  是在保持其他变量不变的情况下，当前变量的统计学意义。而单因素分析没有考虑其他潜在的混杂变量，这可能会夸大单个变量的效应。其次，多重共线性的存在可能会影响多因素模型中变量的显著性，尤其是当相关的预测变量在模型中同时出现时，即在加权相关网络建立部分所考虑的基因表达一般不是孤立起作用的。

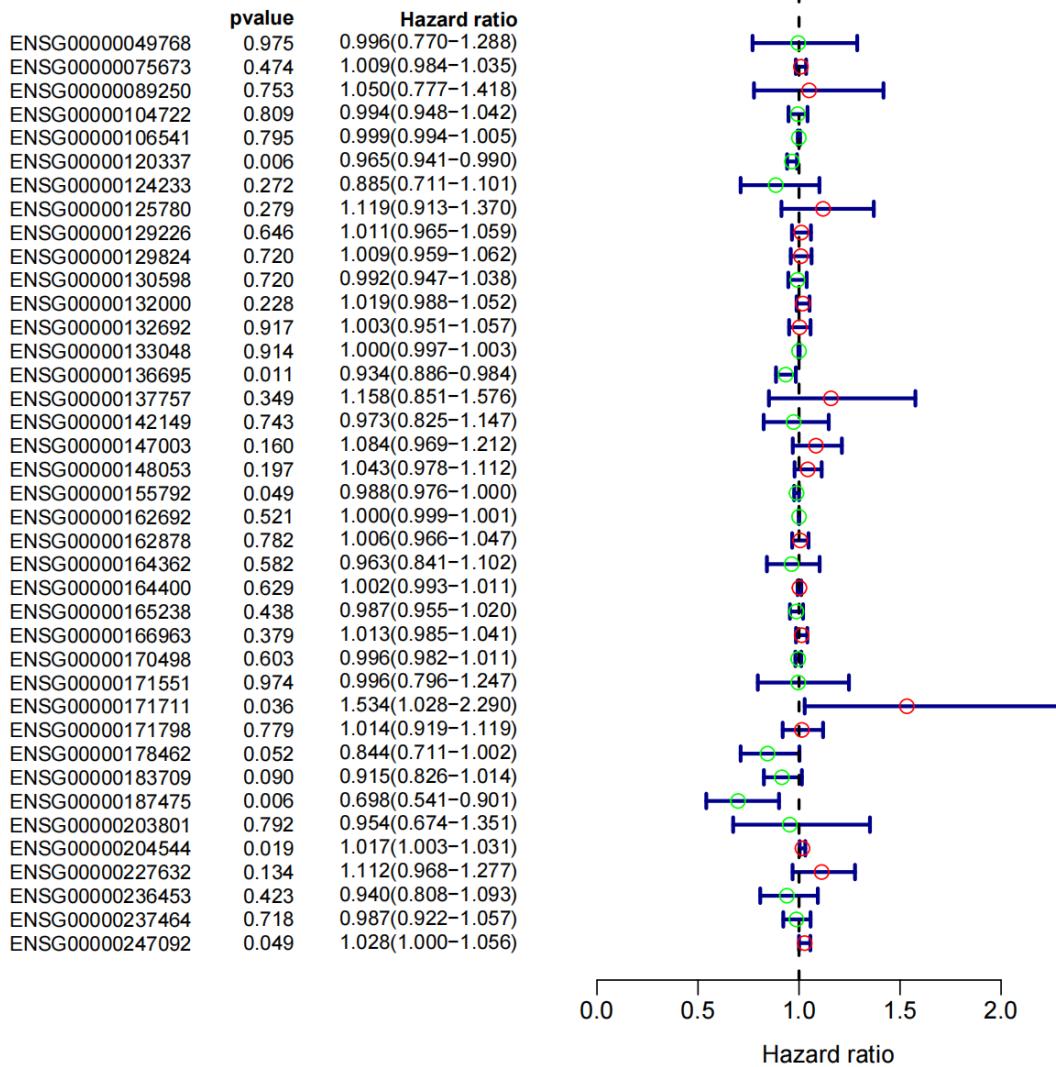


图 3.5 多因素 COX 回归森林图

对于以生存时间为因变量进行的 COX 回归分析，风险比例（Hazard Ratio, HR）在大于 1 时，取值越高表示此因素的危险因素特征越显著；其取值小于 1 时，则值越地表示此因素的保护因素特征越显著，这种情况下的风险评分模型可以表示为式 (3.13)。

$$RiskScore = \sum_{i=0}^N (HR_i - 1) \quad (3.13)$$

式(3.13)中N为非零系数的个数，表明临床患者的风险评分与其生存时间呈现负相关，而本模型中的因变量为受辐射照射的时间，显然风险评分与辐射照射时间为正相关，故可得到本模型中的风险评分模型(3.14):

$$RiskScore = - \sum_{i=0}^N (HR_i - 1) \quad (3.14)$$

## 3.2 风险评分模型的测试与性能分析

### 3.2.1 受试者工作特性分析

首先从基因型-组织表达(Genotype-Tissue Expression, GTEx)数据库上下载共计7862例健康人体的基因数据，GTEx项目可用于研究人类不同组织的特异性基因表达和调节。

肺腺癌患者的基因数据则可以通过癌症基因组图谱(The Cancer Genome Atlas, TCGA)得到。TCGA是癌症基因组学领域的一个开创性项目，在改变癌症研究、诊断和治疗的格局方面发挥了重要作用。从该数据上下载完整的LUAD-TCGA项目作为癌症参考数据，数据中包含600例肺腺癌患者的基因表达数据。加州大学圣克鲁兹分校的Xena是一个强大的癌症组学分析平台，可以从该平台上下载LUAD-TCGA项目的患者临床信息，包括患者的生存时间、距离死亡的时间、最终状态、是否吸烟等相关临床数据。

对健康人类个体和肺腺癌患者的基因表达数据进行风险评分，并利用评分结果对其进行二值分类。二值分类器的评估可以使用受试者工作特征曲线进行评估，其结果如图3.6所示。

采用类似的方法，从组成风险评分模型的自变量中选取四个基因，单独观察其ROC曲线，结果如图3.7所示。可以发现所选取的基因只有ENSG00000166963没有分类性能外，其余的ENSG00000132000、ENSG00000133048、ENSG00000171551分类性能均比较理想，曲线均保持在左上角。

本模型对于真实临床数据测试集的ROC曲线的AUC为0.793，仅比训练集低0.003，区分度性能较好。

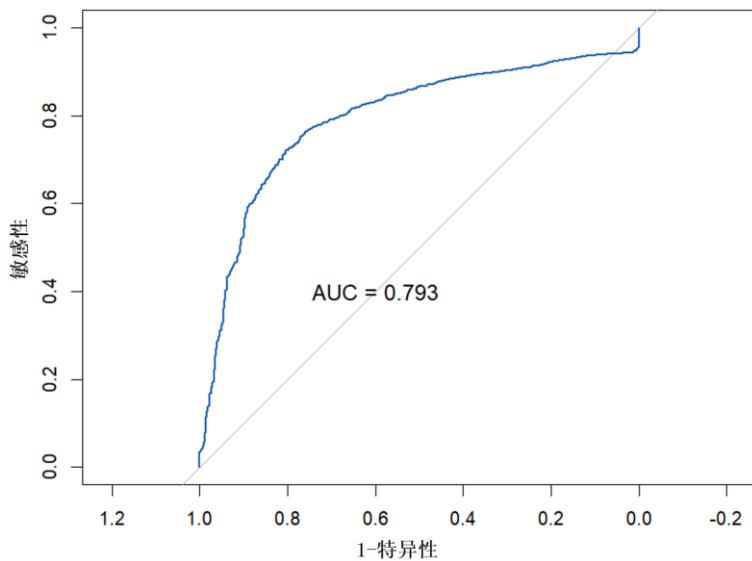


图 3.6 风险评分模型 ROC 曲线

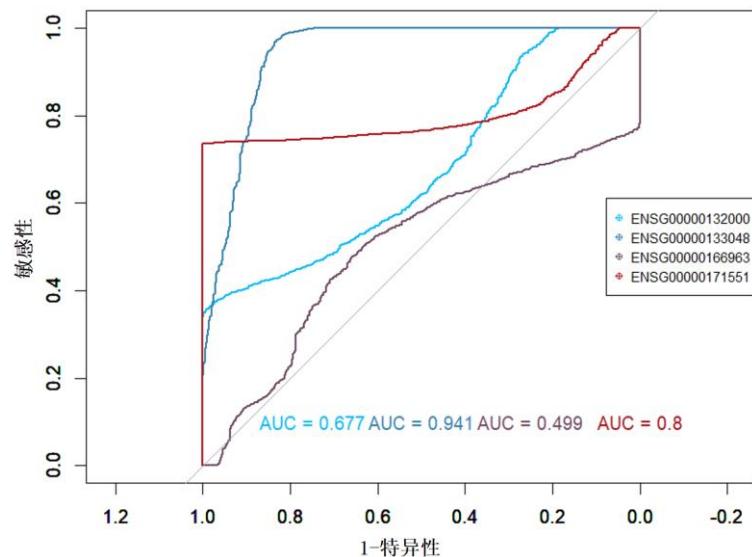


图 3.7 单因素评分 ROC 曲线

### 3. 2. 2 Kaplan-Meier 生存分析

Kaplan-Meier 法通常简称为 KM 法，该方法通过生存时间和终止状态（如死亡或失访）的数据，生成生存曲线，以此来估计在特定时间点上的生存概率。KM 法适用于比较两组或多组患者的生存情况，因此在临床研究中得到了广泛应用。

KM 生存分析的核心是生存函数  $s(t)$ ，它表示在时间  $t$  之前生存的概率。当数据中存在删失情况时，KM 法能够处理不完全的信息，用于从生命周期数据中估计生存函数。生存率是通过对各个观察时间点的生存率进行累积乘积得到的，这样得到的生存率是在

每个观察时间点上存活的概率，形成了生存曲线，其曲线图是一系列水平递减的阶梯。

在医疗领域中，KM 生存分析被广泛应用于比较不同治疗方法、不同患者群体或其他因素对生存时间的影响。生存曲线直观地反映了不同情况下患者生存差异，是临床研究中重要的工具之一。而一般情况下的疾病生存情况并非单一因素起效，因此可以将风险评分模型的分数作为 KM 生存分析的自变量，以此综合分析多因素作用下的生存情况。

对于 n 个观察对象，每个对象都有一个生存时间  $t_i$  ( $i=1,2,\dots,n$ ) 及一个二元事件指示器  $\delta_i$ ，事件指示器取值为 1 时代表时间发生，取值为 0 时代表截尾。在本模型中的事件即为患者死亡，截尾表示患者在此生存时间下仍然存活<sup>[30]</sup>。对观察时间进行排序，得到升序的观察时间序列： $t_1 < t_2 < \dots < t_n$ 。可以计算生存函数为式 (3.15)。

$$S_i = S_{i-1} \times \left(1 - \frac{d_i}{n_i}\right) \quad (3.15)$$

式 (3.15) 中  $d_i$  表示对于观察时间  $t_i$  发生事件的个体数， $n_i$  表示在此事件前存活的个体总数， $S_i$  表示生存率。

对于 KM 分析中不相交的两条生存曲线可以进行对数秩检验，以验证其统计学意义。对数秩检验有 Z 检验和卡方检验两种，现在一般采用后者<sup>[31]</sup>。

令 i 时刻分组 j 中个体死亡数量为  $d_{j,i}$ ，则有 i 时刻的死亡总数  $d_i$  为  $d_{1,i} + d_{2,i}$ ，且 i 时刻分组 j 的理论死亡数可以表示为式 (3.16)。

$$\begin{cases} T_{i,j} = \frac{n_{j,i}d_i}{n_j}, j = 1,2 \\ T_j = T_{1,i} + T_{2,i} \end{cases} \quad (3.16)$$

由于  $d_{j,i}$  服从超几何分布  $d_{j,i} \sim h(n_i, n_{j,i}, d_j)$ ，实际死亡数的方差可以表示为式 (3.17)。

$$V_{j,i} = \frac{n_{j,i}d_j(n_i - d_j)(n_i - n_{j,i})}{n_j^2 d_j} \quad (3.17)$$

在样本数量足够大时，其趋近于正态统计量式 (3.18)。

$$\frac{\sum_{i=1}^k (d_{j,i} - T_{j,i})}{\sqrt{\sum_{i=1}^k V_{j,i}}} \rightarrow N(0,1) \quad (3.18)$$

对式 (3.18) 取平方，其服从自由度为 1 的卡方分布，如式 (3.19) 所示。

$$\chi^2 = \frac{\left(\sum_{i=1}^k (d_{j,i} - T_{j,i})\right)^2}{\sum_{i=1}^k V_{j,i}} \rightarrow \chi^2(1) \quad (3.19)$$

清洗 LAUD-TCGA 项目中临床信息缺失的样本后使用“survminer”包进行生存分析，使用最大选择秩统计法得到最佳截断值，如图 3.8 所示。将肺腺癌患者分为低风险组和高风险组，分别由 532 和 55 个样本，得到 Kaplan-Meier 曲线如图 3.9 所示。图中蓝色曲线代表低风险组，黄色曲线代表高风险组，可以发现高风险组曲线始终在低风险曲线下方，说明同等条件下高风险组患者存活率低于低风险组。

取显著性水平  $\alpha=0.05$ , 95% 分位数  $\chi^2_{0.95,1}=3.84$ , 得到  $p$  值为  $0.033<0.5$ , 不能拒绝原假设，可以认为高风险组和低风险组的总体分布不同，说明风险评分模型具有统计学意义。

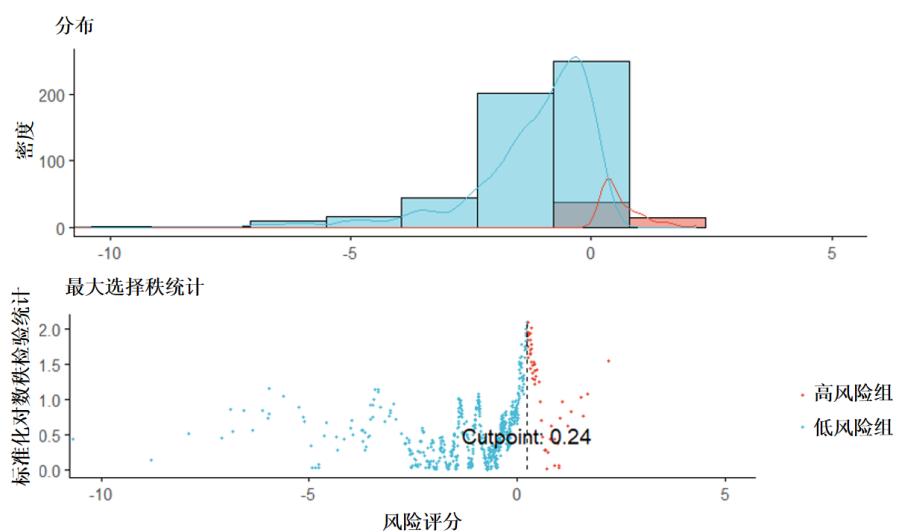


图 3.8 最大选择秩统计法选取截断点

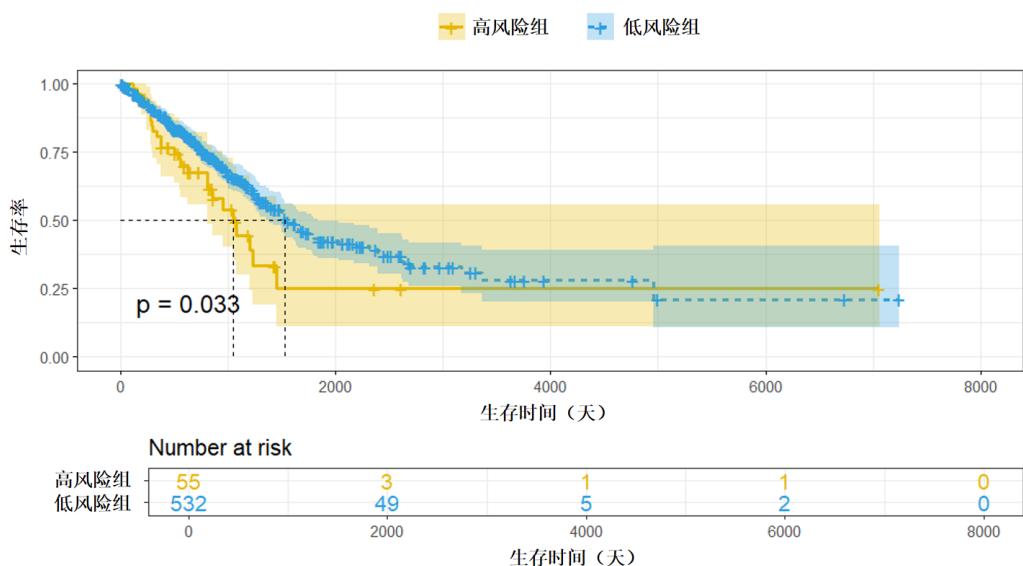


图 3.9 Kaplan-Meier 生存分析图

为了进一步分析单因素 ROC 曲线中存在的差异，同样对这四个基因进行 KM 分析，其结果如图 3.10 所示，发现根据四种基因进行分组的 log-rank 检验 p 值分别为 0.0067、0.0057、0.066、<0.0001，仅 ENSG00000166963 组未通过检验，与 ROC 曲线指示的结果相符。ENSG00000166963 对分类性能和生存分析均不存在显著作用，但被纳入了风险评分模型，其中一种原因是样本数量较小。在筛选部分中使用了差异分析和单因素 COX 回归两种方法，分别使用了 t 检验和卡方检验验证显著性水平，而样本数量会在一定程度上影响“第二类错误”的发生概率。

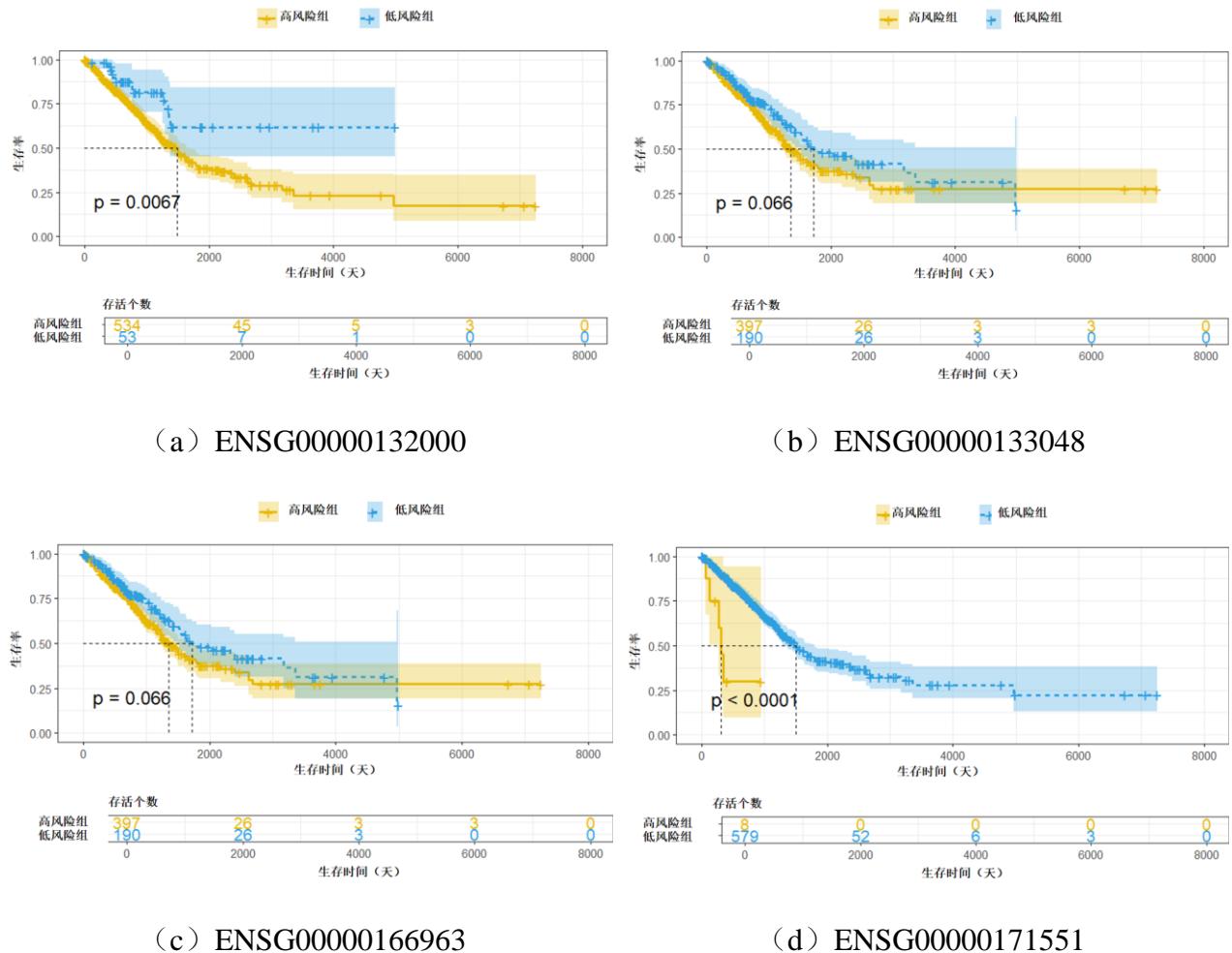


图 3.10 单因素 KM 分析曲线

T 检验发生“第二类错误”的概率可以表示为式 (3.20)。

$$\begin{aligned} \beta &= P(\text{未能拒绝 } H_0 \mid H_0 \text{ 为假}) \\ &= 1 - P(\text{未能拒绝 } H_0 \mid H_0 \text{ 为假}) = \Phi\left(\frac{t_{1-\alpha/2} + \sqrt{n} \cdot \delta}{\sqrt{2}}\right) \end{aligned} \quad (3.20)$$

其中 $\Phi$ 是标准正态分布函数， $\delta$ 是零假设和真实总体之间的差值，或称效应大小。以零假设下的总体均值为 0，真实的总体均值为 1，总体标准差为 1 的情况为例，设置模拟次数为 1000，使用“pwr”包可以得到“第二类错误”概率随样本数量变化的曲线，如图 3.11 所示。

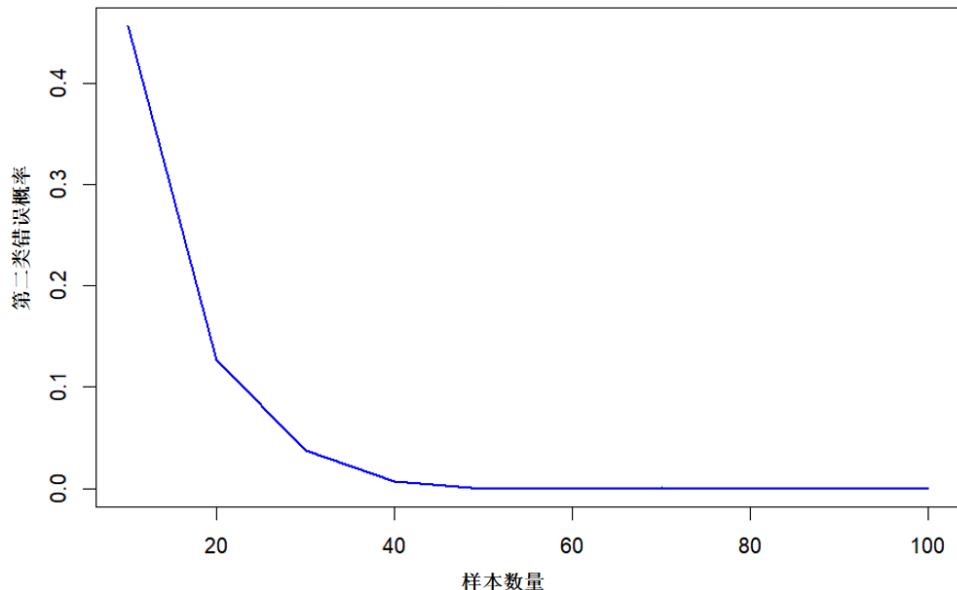


图 3.11 样本数量对 t 检验的影响

卡方检测的第二类错误概率可以表示为式 (3.21)。

$$\begin{aligned}\beta &= P(\text{未能拒绝 } H_0 \mid H_0 \text{ 为假}) \\ &= 1 - P(\text{未能拒绝 } H_0 \mid H_0 \text{ 为假}) = 1 - P(\chi^2 \geq \chi^2_{1-\alpha}(df))\end{aligned}\quad (3.21)$$

其中 $\chi^2_{1-\alpha}(df)$ 是自由度为 $df$ 的卡方分布上的临界值。设置自由度为 1，效应大小为 0.2，使用“pwr”包进行计算，发生“第二类错误”概率随样本数量变化的曲线，如图 3.12 所示。在两种假设检验中，发生“第一类错误”的概率是由显著性水平决定的，不受到样本数量的影响，而发生“第二类错误”的概率则随着样本数量的增加而下降。所以纳入 ENSG00000166963 这类对模型影响并不显著的自变量，即对应原假设为假，但不能未能拒绝原假设的现象，这一现象会在样本数量增加时逐步消失。

此外，由于训练集中的因变量是受辐射照射时间，而辐射所导致的基因表达改变并不一定能够显著作用于肺腺癌的预后。

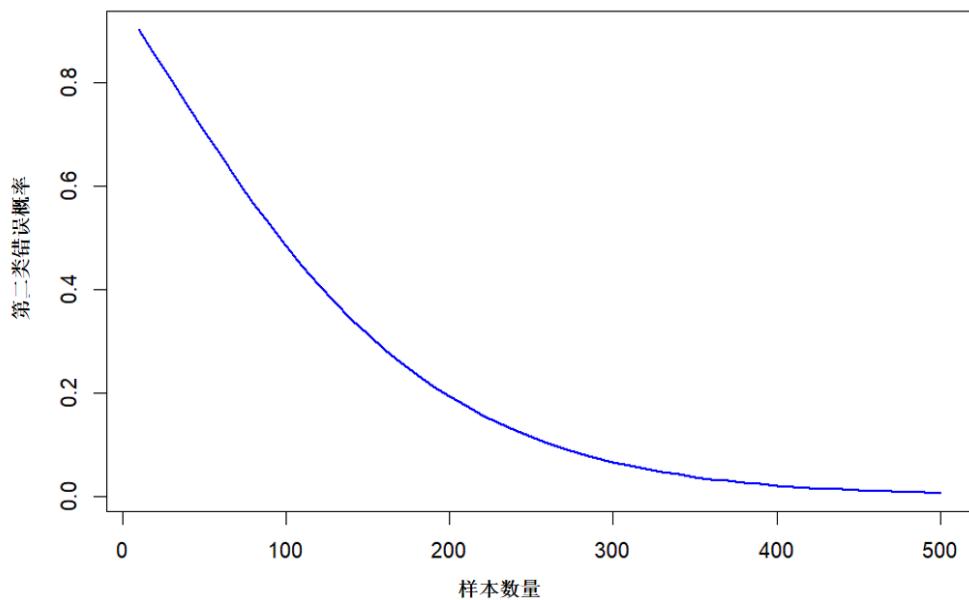


图 3.12 样本数量对卡方检验的影响

### 3.3 本章小结

本章使用单因素 COX 回归模型、LASSO 回归模型和多因素 COX 回归模型建立风险评分模型，并使用 ROC 曲线和 KM 分析对其进行测试和性能分析。本文首先使用单因素 COX 回归评估单个协变量对模型的影响，并使用 LASSO 回归进一步对基因进行筛选、降低模型的复杂度。使用多因素 COX 回归模型整合多个变量，并根据风险比率建立风险评分模型。在模型测试部分，使用 ROC 曲线评估模型区分肺腺癌患者和健康人类的性能，使用 KM 分析比较模型对肺腺癌患者高低风险组划分的效果。最后，本章分析了样本数量对模型显著性检验的影响。

## 第4章 风险评分结果的生物信息学分析

### 4.1 两个风险组的分布情况分析

首先基于所有自噬相关数据对两个风险组进行 PCA 分析，得到分布图如图 4.1 所示，发现两个风险组高度重叠，且在图中所示的两个维度上高风险组将低风险组完全包围，不能体现两个风险组在特征空间中的差异。

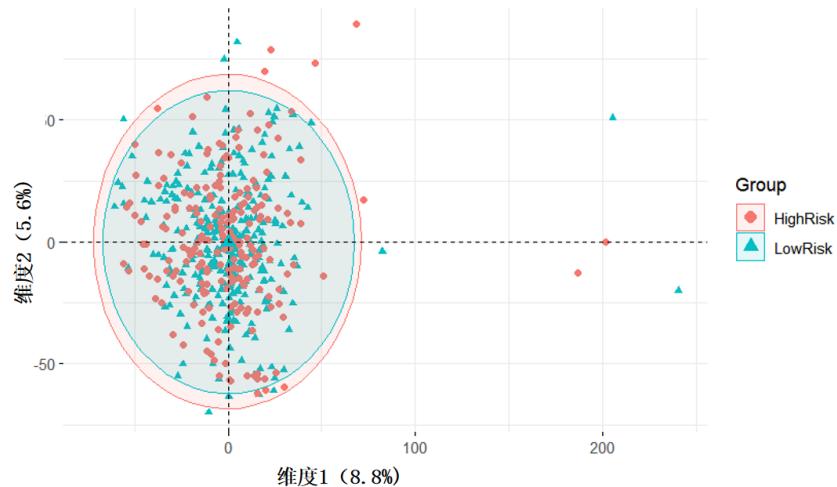


图 4.1 基于所有自噬相关基因进行 PCA 分析

基于风险评分模型所选取的特征因素基因再次进行 PCA 分析，得到的分布图如图 4.2 所示，可以发现两个风险组虽然仍有重叠区域，但两个主成分上均有各自独有的区域，区分度具有一定程度的提升。

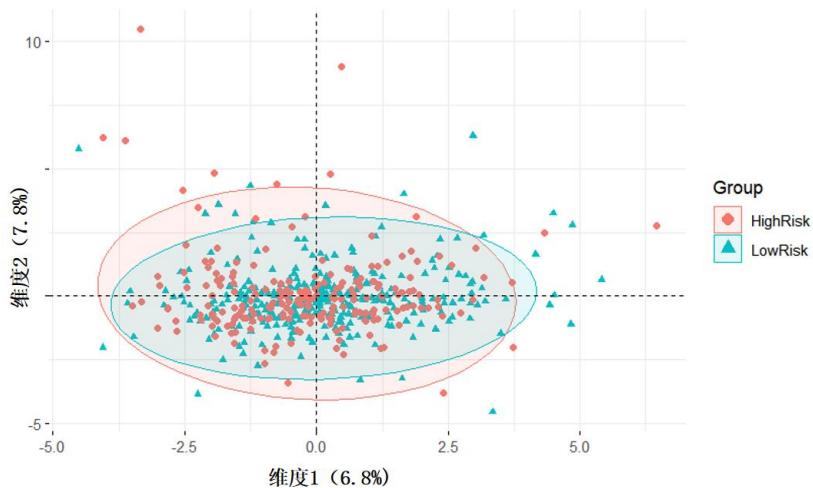


图 4.2 基于用于构建评分模型的自噬相关基因进行 PCA 分析

## 4.2 两个风险组的分布的免疫浸润分析

在现代肿瘤学研究中，免疫浸润分析已成为一项关键技术，它允许研究者定量评估肿瘤微环境（TME）中的免疫细胞组成和活性。这种分析对于揭示肿瘤免疫微环境的复杂性、指导免疫治疗策略的制定以及预测患者的治疗反应具有重要意义。免疫浸润分析通过先进的生物信息学方法对肿瘤组织中的免疫细胞亚群进行定量评估。这些方法基于特定的免疫细胞标记基因或去卷积算法，从而估计肿瘤样本中各种免疫细胞的相对丰度。

通过免疫浸润分析，可以识别出肿瘤微环境中的免疫沉寂与免疫活跃的区域。随着单细胞测序技术的发展，免疫浸润分析的精度和分辨率得到了显著提高。这使得研究者能够在更细致的水平上理解 TME 的异质性，为个体化医疗提供了更为精确的生物标志物。

CIBERSORT 是一种先进的生物信息学工具，根据线性支持向量回归原理，利用一组已知的免疫细胞特征基因预先训练的特征基因表达谱模型分析混合细胞样本中各个免疫细胞亚群的比例。通过这个模型，CIBERSORT 可以利用去卷积的方式准确地从混合的基因表达数据中分离出各个免疫细胞亚群的表达信号<sup>[32]</sup>。CIBERSORT 相较于其他类似工具，在处理噪声数据方面表现出更高的精确度。

本文所采用的 CIBERSORTx 于 2019 年由斯坦福大学的助理教授 Aaron M. Newman 提出，CIBERSORTx 开发了一种自适应滤波的算法来消除噪音，以降低其对吸引表达估计的影响<sup>[33]</sup>。可以从图 4.3 可以观察到自适应滤波前后 FACS 纯化和 MACS5 纯化的相关性显著上升。图中  $r$  代表皮尔逊相关系数， $\rho$  代表斯皮尔曼相关系数，对角斜线表示线性回归的结果。

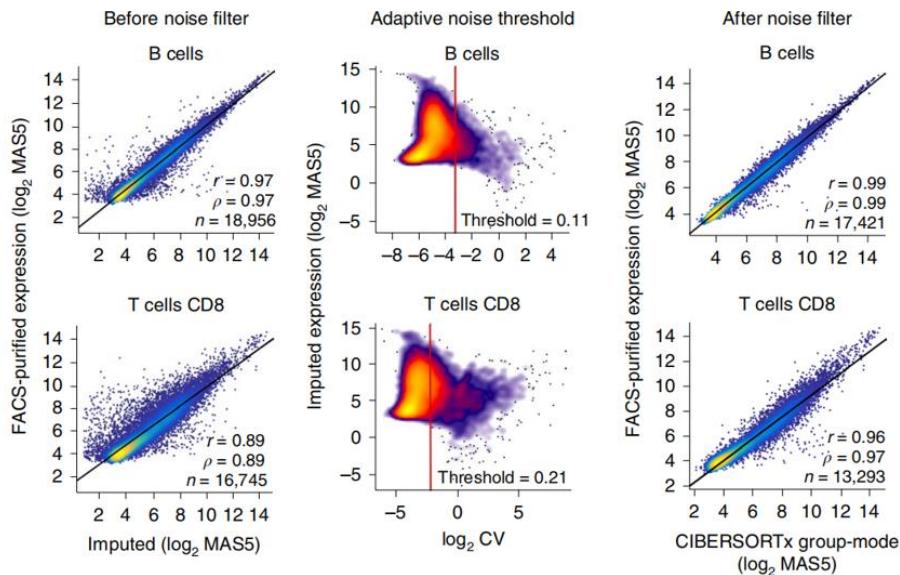


图 4.3 B 细胞和 T 细胞自适应滤波前后对比<sup>[32]</sup>

该方法目前仅支持在线输入数据，由于低风险组样本数量过多，因而将其切分为 5 个批次作为混合矩阵输入，签名矩阵文件使用默认的 LM22 文件。

选取高低风险组中的 15 个样本，使用“ComplexHeatmap”包绘制其热图，使用“g gpubr”包绘制堆叠条柱图，如图 4.4~4.7 所示，可以观察到高低风险组在部分免疫细胞分布上存在差异。

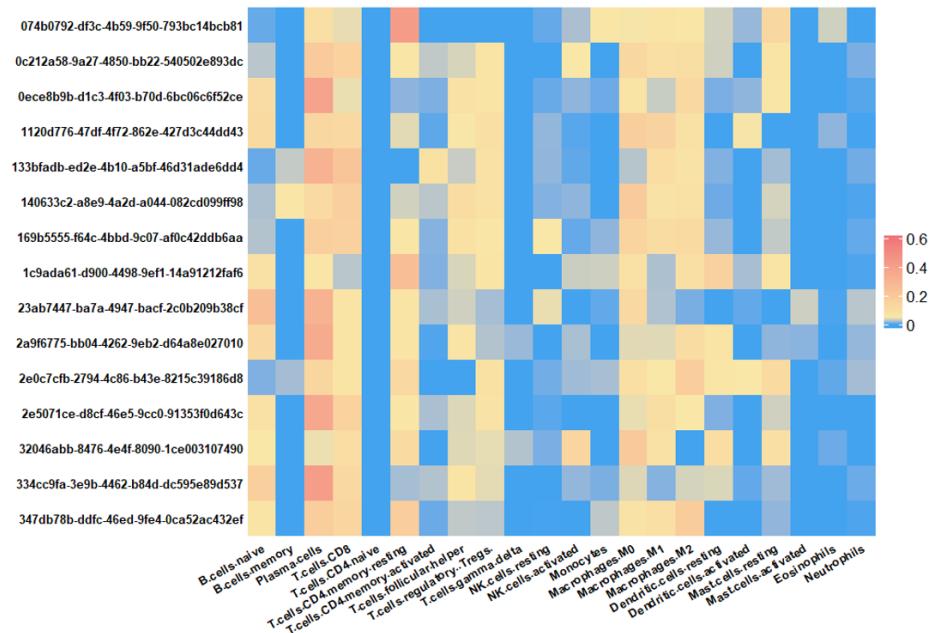


图 4.4 高风险组中 15 个样本免疫浸润热图

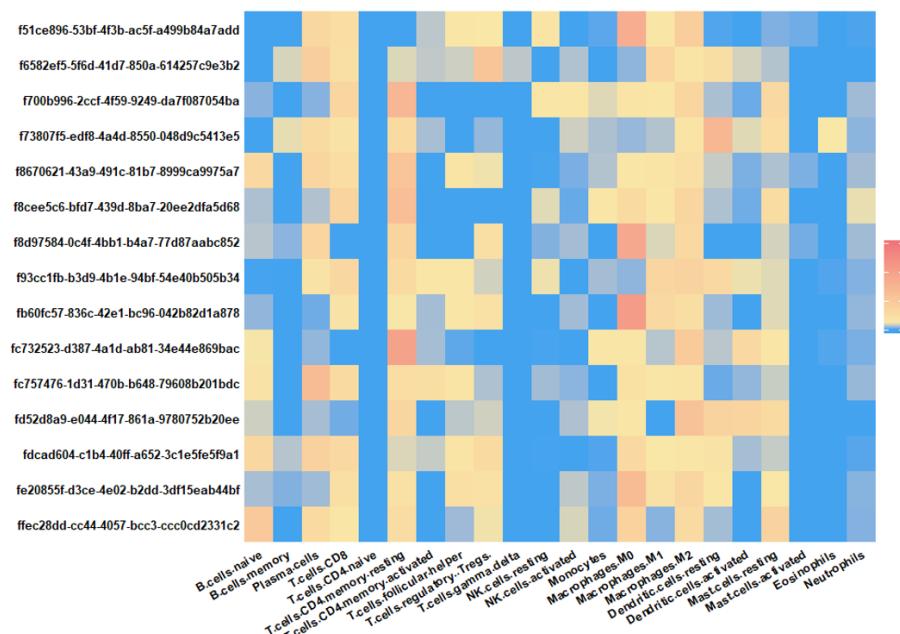


图 4.5 低风险组中 15 个样本免疫浸润热图

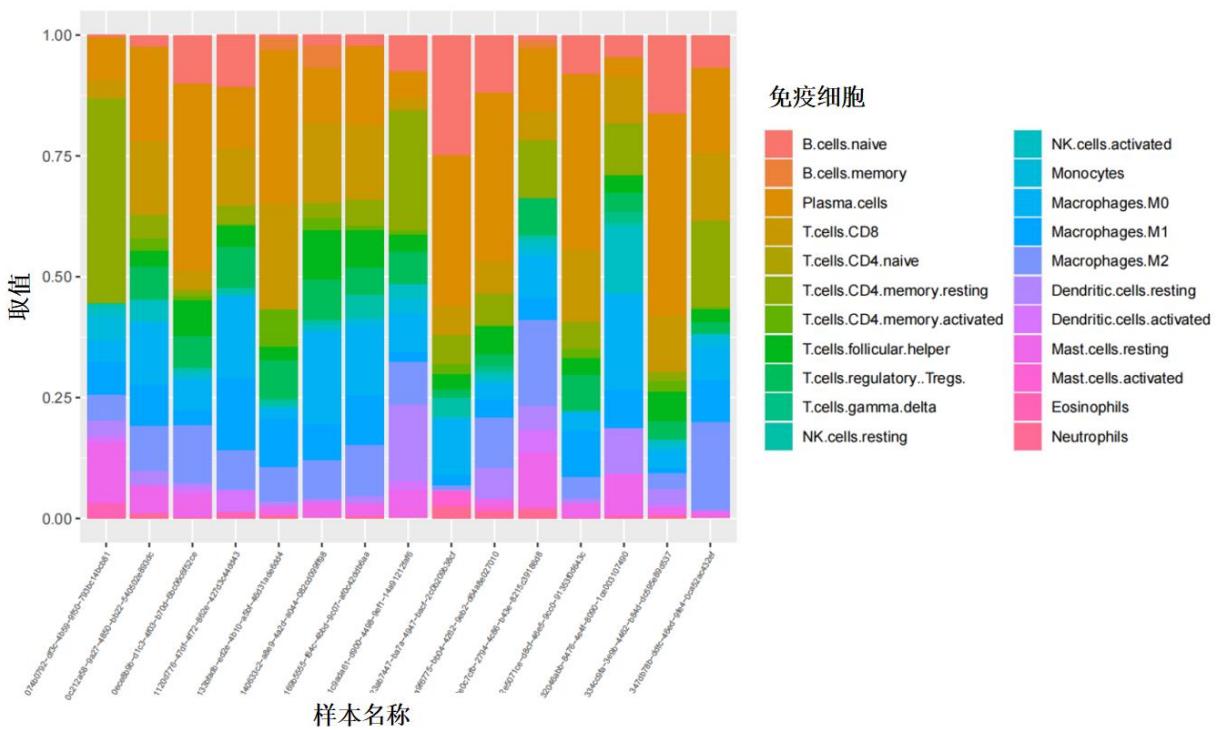


图 4.6 高风险组中 15 个样本免疫浸润堆叠条柱图

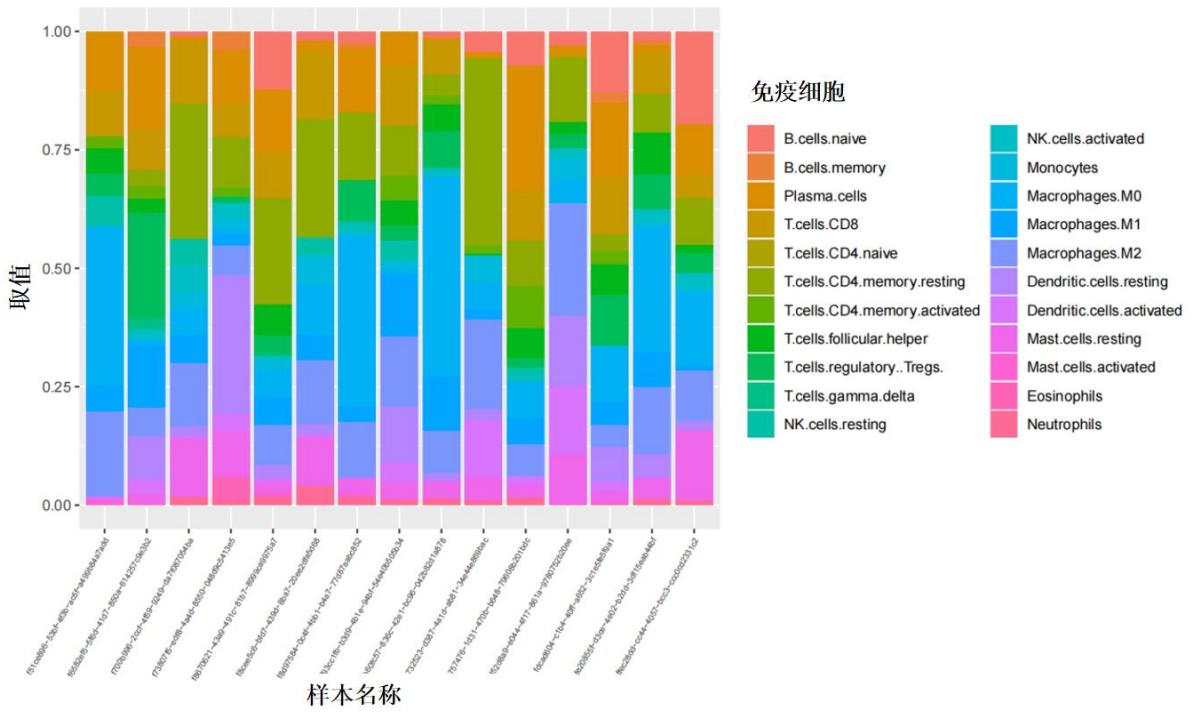


图 4.7 低风险组中 15 个样本免疫浸润堆叠条柱图

为了分析全部样本之间的差异，对免疫浸润矩阵进行合并整理，绘制两个风险组的箱线图如图 4.8 所示，发现两个风险组在 plasma.cell 等细胞上表达具有显著差异。

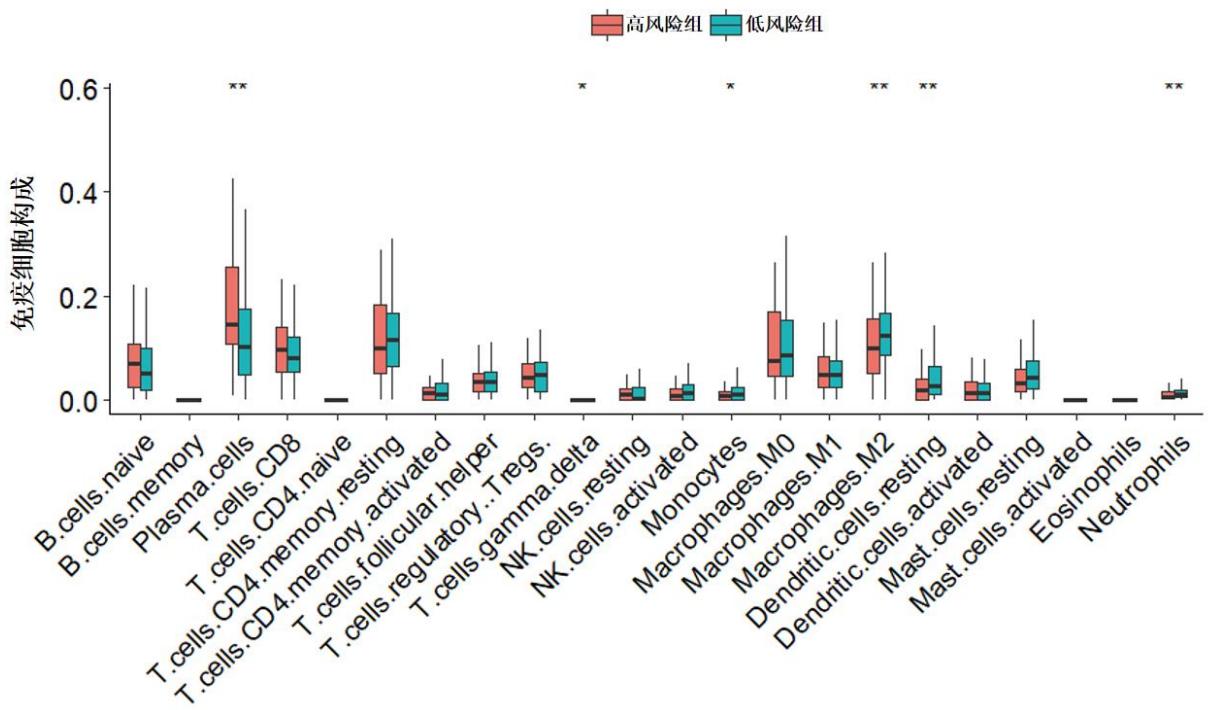


图 4.8 高低风险组免疫浸润分析

### 4.3 本章小结

本章使用生物信息学方法揭示了两个风险组在不同特征下的分布差异以及免疫浸润情况。首先使用 PCA 分析展示基于评分模型所选取的基因，两个风险组在特征空间中的区分度得到一定程度的提升。免疫浸润分析证实了高低风险组在肿瘤微环境中免疫细胞组成上的显著差异，尤其是在 plasma.cell 等细胞类型上。风险评分结果的生物信息学分析为肺腺癌的生物标志物发现和免疫治疗目标的识别提供了重要线索。

## 第5章 总结与展望

### 5.1 全文工作总结

利用器官芯片能够模拟人体器官和结构的功能，将其应用于疾病的预后分析，可以有效帮助建立癌症的风险评分模型。本文通过对机器学习以及生物信息分析相关理论和技术的研究，完成了基于放射性肺损伤芯片数据的风险评分模型建立。

本文完成的工作如下：

(1) 比较两个批次数据在所有特征上的分布以检测其批次效应，并对批次效应进行消除。利用加权相关网络分析了基因间的相互作用关系，并对离群值进行检测，利用基因差异分析、基因变异集分析、风险比例回归模型和 LASSO 回归模型筛选对预后具有显著价值的基因。

(2) 根据多因素风险比例回归模型建立了风险评分模型，并利用外部健康人体数据集和肺腺癌患者数据集对模型进行测试。证明模型对健康志愿者和肺癌患者的二分类受试者工作特性曲线性能符合要求，肺癌患者的高低风险组 Kaplan-Meier 生存分析曲线的 log-rank 检验具有显著的统计学意义。

(3) 利用 PCA 分析研究了高低风险组在所有自噬相关基因上的分布和所筛选基因上的分布之间的区别，验证了筛选过程的有效性。利用生物信息分析中的 CIBERSORTx 法免疫浸润分析研究了高低风险组的 22 种免疫细胞在组织中分布的差异。

### 5.2 下一步工作展望

利用器官芯片及其相关技术建立预后模型，能够显著降低获取风险评分模型数据的难度，便于对患者的客观检测和评估，提高预后的针对性和科学性。本文完成了基于肺损伤芯片中自噬相关基因的肺腺癌预后模型的构建，并对结果的生物信息学过程进行研究。但本文的研究内容存在一些缺陷，后期可以进行以下改进：

(1) 提高受试者工作特性曲线的精度。本文的训练集数据来源于公开的 GEO 数据集，在合并两个批次数据后也仅有 56 个样本，在一定程度上限制了模型的精度。后续可以在此基础上补充实验，提高训练集的样本数量。

(2) 提高模型的泛化能力。放射性损伤并不是肺腺癌的唯一致病因素，仅仅利用受放射性损伤的芯片数据建立预后模型会限制模型的泛化能力，降低模型在测试集中的精度。后续可以融合其他致病因素，如增加香烟、致癌化学品、病毒等损害的芯片数据，利用每一个致病因素构建一个专家网络，并利用混合专家模型进行预后模型构建。对每个测试集数据进行评分前，首先通过一个门网络，为每个专家网络分配系数，最终的评分结果由各个专家网络综合生成，这样的模型具有高度的可解释性。

(3) 将模型应用于个性化医疗。对于常见病，可以利用已有的临床数据作为训练集构建基于因果关系预测风险评分模型。利用肺腺癌患者的癌细胞制备芯片，并利用微流控技术将芯片置于各种致病因素及治疗环境中，将器官芯片数据作为测试数据，得到不同治疗方案下的风险评分差异，为肺腺癌患者提供最佳的定制化治疗方法。

## 参考文献

- [1] 周静, 王心悦, 李兆娜, 等. 肺腺癌自噬相关基因预后风险评分模型构建及验证[J]. 中国肺癌杂志, 2021, 24(08): 557-566.
- [2] M B , C G , E R , et al. A Bayesian hierarchical approach to account for left-censored and missing radiation doses prone to classical measurement error when analyzing lung cancer mortality due to  $\gamma$ -ray exposure in the French cohort of uranium miners[J]. Radiation and environmental biophysics, 2020, 59(3): 423-437.
- [3] N M , T S , V S , et al. CYTOGENETIC DAMAGES IN LUNG CANCER PATIENTS TREATED BY EXTERNAL RADIATION THERAPY[J]. Probl Radiac Med Radiobiol, 2019, 24: 411-425.
- [4] 工业和信息化部, 国家卫生健康委员会, 国家发展和改革委员会, 等. 十部门关于印发《“十四五”医疗装备产业发展规划》的通知[EB/OL]. 中华人民共和国中央人民政府 2021(2021-12-21)[2024-05-16]. [https://www.gov.cn/zhengce/zhengceku/2021-12/28/content\\_5664991.htm](https://www.gov.cn/zhengce/zhengceku/2021-12/28/content_5664991.htm).
- [5] 国家卫生健康委, 国家发展改革委, 教育部, 等. 关于印发健康中国行动—癌症防治行动实施方案（2023—2030 年）的通知[EB/OL]. 中华人民共和国中央人民政府 2023(2023-10-30)[2024-05-23]. [https://www.gov.cn/zhengce/zhengceku/202311/content\\_6915380.htm](https://www.gov.cn/zhengce/zhengceku/202311/content_6915380.htm).
- [6] Leung Chak Ming, de Haan Pim, Ronaldson Bouchard Kacey, et al. A guide to the organ-on-a-chip[J]. Nature Reviews Methods Primers, 2022, 2(1): 1-29.
- [7] 秦建华, 张敏, 于浩, 等. 人体器官芯片[J]. 中国科学院院刊, 2017, 32(12): 1281-1289.
- [8] Chao M, Yansong P, Hongtong L, et al. Organ-on-a-Chip: A New Paradigm for Drug Development[J]. Trends in Pharmacological Sciences, 2020, 42(2): 119-133.
- [9] Low Lucie A, Mummery Christine, Berridge Brian R, et al. Organs-on-chips: into the next decade[J]. Nature reviews Drug discovery, 2020, 20(5): 345-361.
- [10] Grimm D . EPA plan to end animal testing splits scientists[J]. Science, 2019, 365(6459): 1231-1231.
- [11] Berivan C , Christina K , Mubashir N , et al. Multi-Organ-on-Chips for Testing Small-Molecule Drugs: Challenges and Perspectives[J]. Pharmaceutics, 2021, 13(10): 1657-1657.
- [12] M. A. M. Hasan, M. Maniruzzaman, J. Shin. Gene Expression and Metadata Based Identification of Key Genes for Hepatocellular Carcinoma Using Machine Learning and Statistical Models[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2023, 20(6): 3786-3799.
- [13] Peng Y , Feng Z , Dong N , et al. Fused Sparse Network Learning for Longitudinal Analysis of Mild Cognitive Impairment[J]. IEEE transactions on cybernetics, 2021, 51(1): 233-246.
- [14] Xiangtao L , Ka-Chun W . Evolutionary Multiobjective Clustering and Its Applications to Patient

- Stratification[J]. IEEE transactions on cybernetics, 2018, 49(5): 1680-1693.
- [15] Cheng H , Koc L , Harmsen J , et al. Wide Deep Learning for Recommender Systems[J]. 2016, DLRS 2016: 7-10.
- [16] Xu Z , Zhang Q , Yip F S P . Predicting post-discharge self-harm incidents using disease comorbidity networks: A retrospective machine learning study[J]. Journal of Affective Disorders, 2020, 277(prepubish): 402-409.
- [17] Chen W, Ou M, Tang D, et al. Identification and validation of immunerelated gene prognostic signature for hepatocellular carcinoma[J]. Immunol Res, 2020, 13(1): 1-14.
- [18] Evan W J, Cheng L, Ariel R. Adjusting batch effects in microarray expression data using empirical Bayes methods[J]. Biostatistics (Oxford, England), 2007, 8(1): 118-127.
- [19] Anna H , M B M , Debarun D , et al. Quantitative prediction of human pharmacokinetic responses to drugs via fluidically coupled vascularized organ chips[J]. Nature biomedical engineering, 2020, 4(4): 421-436.
- [20] Dasgupta Queeny, Jiang Amanda, Wen Amy M. et al. A human lung alveolus-on-a-chip model of acute radiation-induced lung injury[J]. Nature Communications, 2023, 14(1): 6506-6519.
- [21] Peter L , Steve H . Eigengene networks for studying the relationships between co-expression modules[J]. BMC Systems Biology, 2007, 1(1): 54.
- [22] Bin Z , Steve H . A general framework for weighted gene co-expression network analysis[J]. Statistical applications in genetics and molecular biology, 2005, 4: 17.
- [23] Peter L , Bin Z , Steve H . Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R[J]. Bioinformatics (Oxford, England), 2008, 24(5): 719-720.
- [24] 蒋杰, 王立春. 线性模型参数向量的近似贝叶斯估计[J]. 高校应用数学学报 A 辑, 2022, 37(01): 1-14.
- [25] Sonja H, Robert C, Justin G. GSVA: gene set variation analysis for microarray and RNA-seq data[J]. BMC bioinformatics, 2013, 14(1): 7.
- [26] D. J K, E. D S. Fifty Years of the COX Model[J]. Annual Review of Statistics and Its Application, 2023, 10: 1-23.
- [27] Tingchuan X, Yinghong W, Changjun Z. A risk model based on 10 ferroptosis regulators and markers established by LASSO-regularized linear COX regression has a good prognostic value for ovarian cancer patients[J]. Diagnostic pathology, 2024, 19(1): 4.
- [28] TIBSHIRANI, R. THE LASSO METHOD FOR VARIABLE SELECTION IN THE COX MODEL[J]. Statist Med, 1997, 16: 385-395.

- [29] Robert T. Regression Shrinkage and Selection Via the Lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 2018, 58(1): 267-288.
- [30] Stalpers A J L, Kaplan L E, Edward L. Kaplan and the Kaplan-Meier Survival Curve[J]. BSHM Bulletin: Journal of the British Society for the History of Mathematics, 2018, 33(2): 109-135.
- [31] Bland M J, Altaian G D. The logrank test[J]. BMJ: British Medical Journal, 2004, 328(7447): 1412.
- [32] Binbin C, S M K, Long C L, et al. Profiling Tumor Infiltrating Immune Cells with CIBERSORT[J]. Methods in molecular biology(Clifton, N. J. ), 2018, 1711: 243-259.
- [33] M A N, B C S, Long C L, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry[J]. Nature biotechnology, 2019, 37(7): 773-782.

## 致 谢

“立正的时候眼睛盯住‘心至楼’三个字”，好似一瞬，我就从军训走到了毕业设计。坐在图书馆撰写毕业论文的我似乎还能从窗户看见在博学广场站军姿的我。

首先感谢王原丽老师和杜庆国老师在课程设计过程中一直为我提供指导和帮助，课程设计能够按照进度正常完成离不开老师的悉心指导。感谢朱力学长和张磊学姐为我检查和修改开题报告、文献翻译，并且牺牲个人时间耐心解答我的各种问题。感谢刘嘉悦师兄为我确定论文选题，帮助我理清了论文的整体思路和提纲，解决我的种种疑问，耐心修改论文的初稿。感谢秦建华老师和王鹏老师帮助我了解微流控器官芯片，学习这一技术的原理和应用。特别感谢马力老师在百忙之中阅读我的论文，耐心细致地进行批注，并亲自指导修改。

感谢母校和信息学院的培养，祝愿母校早日完成“建设让人民满意、让世人仰慕的优秀大学”的目标。感谢班主任张小梅老师三年以来的辛苦付出，感谢讲授专业课的付琴老师、许建霞老师、傅雪蕾老师、李瑞芳老师、刘雪冬老师、黄铮老师、杨福宝老师、刘皓春老师、方艺霖老师、尹勇老师、胡辑伟老师、撒继铭老师、李平安老师、孟伟老师、李昌振老师、洪建勋老师、魏洪涛老师、沈增帧老师、唐静老师、张琪老师的辛勤教导，以及辅导员张林老师、周杰老师、韩昊锟老师、刘浩杰老师在生活和学习上对我的帮助。感谢在比赛中为我提供指导、陪伴我挑灯夜战的王永圣老师、黄小为老师、陈建业老师、谢良老师、詹金鹏老师。方寸之地、寥寥数语，实难道尽我对各位老师的感谢，回想起在学校求学的点滴，我遇到的每一位老师都学识渊博、平易近人，给我提供了极大的帮助。

感谢我的父母家人对我一如既往的支持和关爱，在任何时候都会给予我鼓励，让我得以顺利完成学业。感谢我的室友祝瑜、苗立同、陆仁涛，我们一直共同学习，在毕设期间也相互陪伴、克服各种困难，感谢信息学院的其他同学们，和你们共同度过了最快乐的四年。

最后，衷心感谢论文的评阅老师百忙之中对我的论文进行阅读和评审。