生物信息学系列教程

细胞生物学方向

生物信息学课题教程

基于公共数据库的焦亡基因×肝细胞癌研究全流程

面向零基础新手 • 生信知识补充 • 图文并茂 • 可发表

选题设计 • 数据获取 • 差异表达 • 功能富集 • PPI 网络 • 生存分析 • 论文写作

第 1 版　2025 年

目录

第 00 章导读与学习路线 5

第 01 章细胞生物学知识补充 11

第 02 章生物信息学知识补充 17

第 03 章公共数据库详解 25

第 04 章课题设计与选题 33

第 05 章环境搭建与工具链 39

第 06 章数据获取 45

第 07 章差异表达分析 51

第 08 章可视化与图表 59

第 09 章功能富集分析 67

第 10 章 PPI 网络与 hub 基因 73

第 11 章生存分析与预后模型 79

第 12 章独立验证与拓展 87

第 13 章论文写作与发表 91

第 14 章常见问题与排错 99

第 15 章附录：术语表与资源 107

第 00 章导读与学习路线

欢迎来到《细胞生物学方向 • 生物信息学课题教程》！这是一份专门为**零基础新手**设计的完整入门指南。学完这份教程，你将能独立完成一个 “用公共数据库做细胞生物学方向生信课题” 的完整项目——从选题、下载数据、统计分析、画图，到把它写成一篇可以投稿的论文。

0.1 这本教程是给谁看的？

你是谁	你会得到什么
生物/医学本科生	一套能写进毕业论文或发表成小论文的完整课题流程
低年级研究生	细胞生物学×生物信息学交叉研究的第一块敲门砖
跨专业学习者（计算机/数学背景）	快速补齐细胞生物学常识，理解生信问题的生物学意义
想转型生信的实验科研人员	一套不需要自己测序、用公共数据就能出成果的路线

你不需要：会编程（我们从 R 的第一行命令教起）、懂高深统计（每个概念都用大白话 + 类比讲透）、有实验数据（全部使用公共数据库）。

0.2 学完这本教程，你能做什么？

1. **看懂并复现**一篇典型的 “生信数据挖掘” 论文（这类论文在 Frontiers、BMC、PeerJ 等期刊大量存在）；
2. **独立完成**一个完整课题：选题→下载 TCGA/GEO 数据→差异表达分析→ GO/KEGG 富集→ PPI 网络与 hub 基因→生存分析与风险模型→独立验证；
3. **画出 10+ 种**论文级图表：火山图、热图、PCA、富集气泡图、GSEA、PPI 网络、KM 生存曲线、森林图、ROC、诺模图；
4. **把结果写成**符合 IMRaD 结构的论文初稿，并了解投稿流程与期刊选择；
5. **建立科研素养：**知道什么是多重检验校正、什么是数据窥探偏倚、为什么必须独立验证——这些是审稿人一定会问的问题。

0.3 贯穿全教程的示例课题

我们不讲空洞的理论，而是带着你做一个**真实的、可发表的课题**：

> **《基于公共数据库的细胞焦亡相关基因在肝细胞癌中的预后价值与免疫微环境分析》**

- **科学问题**：细胞焦亡（pyroptosis）相关基因在肝细胞癌（HCC）中表达是否异常？哪些焦亡基因与患者预后相关？能否构建预后风险模型？
- **为什么选焦亡**：焦亡是近年细胞生物学最热的研究方向之一（2015 年 GSDMD 机制阐明、2018 年诺贝尔奖级免疫学热潮），与肿瘤免疫、炎症微环境紧密相关，文献量大、审稿人熟悉、新手容易找到参考文献支撑。
- **为什么选肝细胞癌**：TCGA-LIHC 队列完整（371 肿瘤 + 50 癌旁，含随访），GEO 有 GSE14520 等带生存数据的独立验证队列——**数据免费、队列齐全、验证容易**，非常适合新手。
- **为什么这个模式能发表**：这是一条被大量验证过的论文生产线（“XX 相关基因×癌症×预后模型”），Frontiers in Genetics、BMC Cancer、PeerJ、Journal of Cancer、Aging 等期刊均有大量先例。我们会在第 13 章如实分析它的优缺点与审稿趋势。

0.4 学习路线图（建议 4-6 周，每周 6-10 小时）

周次	内容	章节	产出
第 1 周	细胞生物学基础 + 生信知识补充	01、02	看懂机制图；装好 R/RStudio
第 2 周	数据库认知 + 选题定案 + 环境搭建 + 数据下载	03、04、05、06	确定课题；数据到手
第 3 周	差异表达分析 + 可视化	07、08	火山图、热图、PCA
第 4 周	功能富集 + PPI 网络	09、10	富集气泡图、hub 基因列表
第 5 周	生存分析 + 风险模型 + 独立验证	11、12	KM 曲线、森林图、ROC、验证结果
第 6 周	论文写作 + 投稿准备	13、14、15	论文初稿

时间弹性：急于出结果的读者可以跳过第 01-02 章直接开跑（遇到概念再回头查）；想扎实打基础的读者可以放慢到 8 周。第 15 章附录的术语表随时可查。

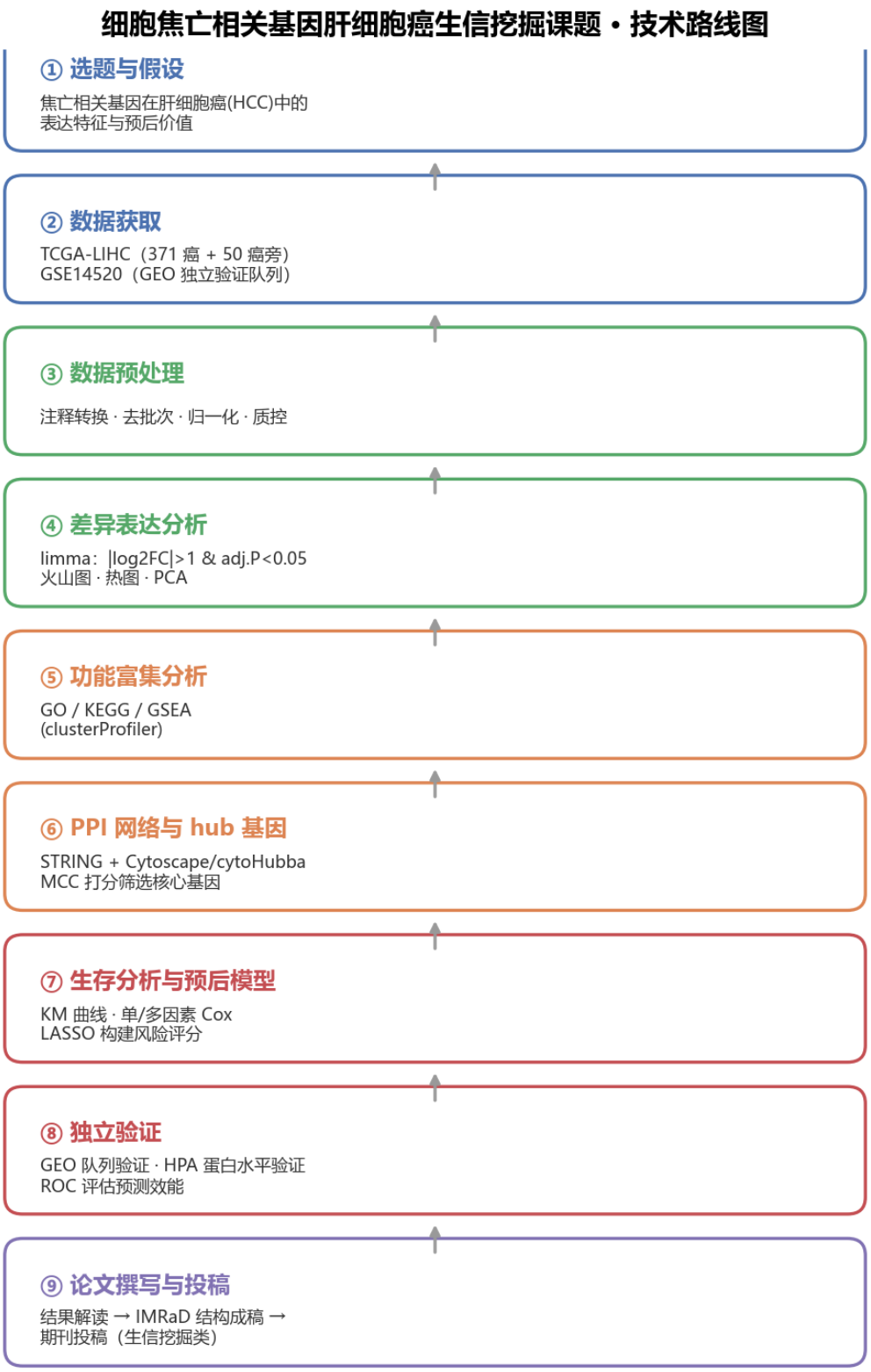
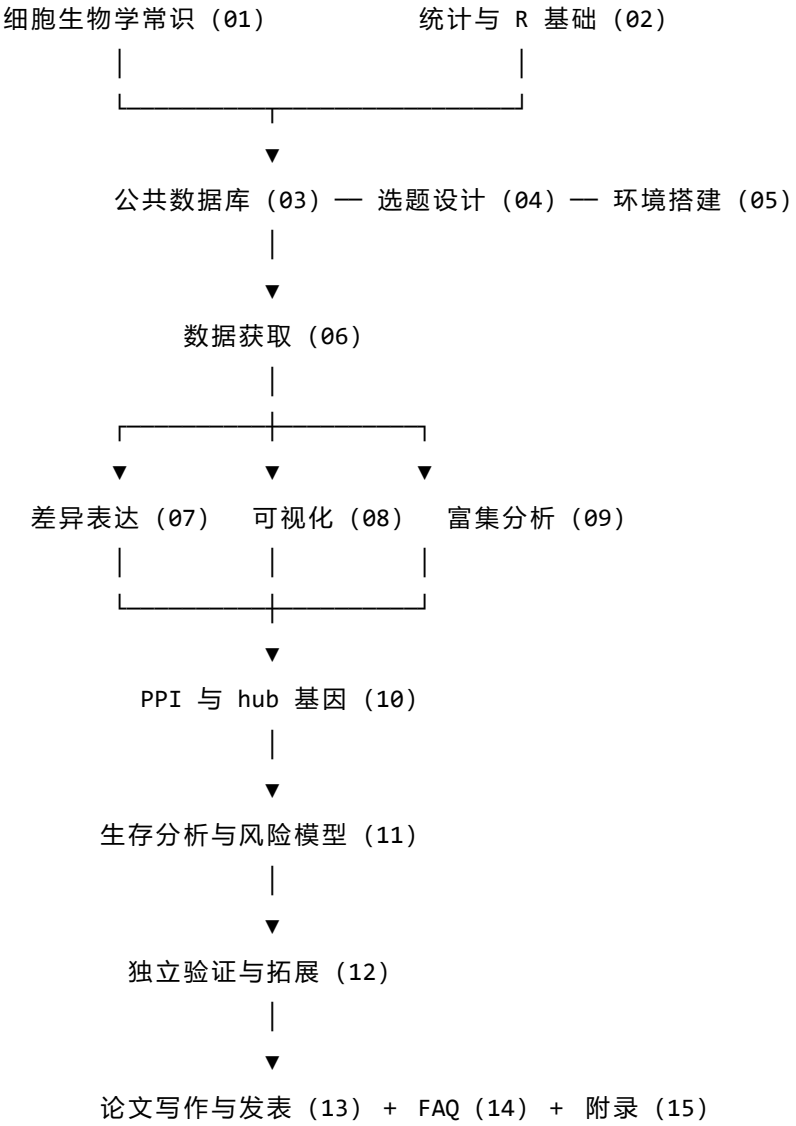


图 1: 图 01 技术路线图（示例）

0.5 教程使用指南

- **每章结构**：本章目标→正文（配图配代码）→本章小结→思考题。
- **代码**：R 代码块可以直接复制到 RStudio 运行（`code/` 目录有整理好的完整脚本）。教程中的图均为**示例图/示意图**（用模拟数据绘制，演示图形样式与解读方法）——你用真实数据跑通后会得到属于自己的图。
- **遇到报错**：先看第 14 章常见问题；解决不了就带着完整报错信息（R 会输出红色错误文本）去搜索，或者问社区——这是所有生信人每天的日常，别怕。
- **科学诚信**：本教程教的是真实、可复现的分析方法。**不要**为了“好看”而挑选截点、篡改数据——审稿人和编辑比你想象的更熟悉这些套路，造假后果极其严重。

0.6 你即将掌握的知识地图



准备好了吗？让我们从“细胞生物学”这门看家学问开始，一步步把基因表达的奥秘变成你可以

发表的数据故事。

第 01 章细胞生物学知识补充

本章为零基础读者补齐 “细胞生物学” 这层背景知识。目标不是背书，而是让后续章节里 “焦亡基因×肝癌” 的每一个名词都听得懂：细胞、基因表达、细胞死亡、焦亡机制、肿瘤微环境。请带着 “这些概念将来会变成哪张数据表” 的想法来读。

本章目标

1. 能用一句话说清 “细胞是生命的基本单位”，并说出细胞生物学在研究什么；
2. 能复述中心法则（DNA → RNA →蛋白质），并理解转录组测序如何把 “基因表达量” 变成数字；
3. 能区分凋亡、坏死、焦亡、铁死亡、自噬性死亡五种细胞死亡方式的核心特征；
4. 能画出细胞焦亡的经典通路（NLRP3 炎症小体→ caspase-1 → GSDMD →膜穿孔→ IL-1β/IL-18 释放），并说出非经典通路的差别；
5. 能用自己的话解释 “焦亡×肝细胞癌” 为什么是一个有临床意义的研究方向。

1. 细胞与细胞生物学是什么

细胞（cell）是生命的基本单位：所有生物——从细菌到人体——都由细胞构成。细胞内部独立完成新陈代谢（metabolism），储存并传递遗传信息，对外界信号做出反应；多细胞生物的身体，就是无数细胞分工协作的 “城市”。人体大约由数十万亿个细胞组成（数量级在 10^{13} 上下，具体数值随估计方法不同有差异，记住数量级即可），其中肝脏主要由肝细胞（hepatocyte）构成——它正是本教程的主角之一。

细胞生物学（cell biology）研究细胞的结构、功能与行为：细胞膜与各种细胞器如何分工（线粒体供能、内质网合成蛋白质、溶酶体降解废物）；细胞如何增殖（proliferation）、分化（differentiation）、通讯（即信号转导，signal transduction）；以及最重要的——细胞如何决定自己的生死。

“死亡” 听起来是终点，但细胞死亡其实是发育、免疫和肿瘤中反复登场的核心角色：胚胎手指间多余的细胞要靠死亡 “雕刻” 出来；被病毒感染的细胞要靠死亡被清除；而癌细胞恰恰是 “拒绝死亡” 的细胞。本章后面会看到：细胞 “怎么死” “死得是否体面”，会直接决定一个肿瘤的命运。

2. 中心法则与基因表达：细胞怎么“读出”自己的说明书

中心法则（central dogma of molecular biology）是分子生物学的“宪法”：遗传信息从 DNA 流向 RNA（这一步叫转录，transcription），再从 RNA 流向蛋白质（这一步叫翻译，translation）。我们可以这样类比：基因（gene）是 DNA 上的一段功能片段，相当于说明书里的一个“章节”；mRNA（信使 RNA，messenger RNA）是章节的“复印件”；蛋白质则是照着复印件“生产出来、真正干活的人”。

关键问题来了：细胞如何控制某个基因“读得有多勤”？答案就是基因表达（gene expression）的强度——转录出的 mRNA 越多，通常说明该基因越活跃。这里必须严谨地补充一句：mRNA 数量并不严格等于蛋白质数量（还存在转录后调控、翻译效率、蛋白质降解等环节），但在高通量测序的尺度上，mRNA 水平是最常用、最可测的“表达量代理指标”。本教程里说“基因 A 在肿瘤里高表达”，指的就是“基因 A 的 mRNA 在肿瘤样本里更多”。

转录组测序（RNA-seq，即 RNA sequencing）如何测量表达量？流程大致是：从样本中提取总 RNA →利用 polyA 尾巴富集 mRNA →逆转录成更稳定的 cDNA →片段化并加上测序接头（adapter）→上机测序得到海量短读段（reads）→把 reads 比对（alignment）到参考基因组→统计每个基因被“覆盖”了多少次。这个计数就是最原始的“表达量”。整个过程见图 02。

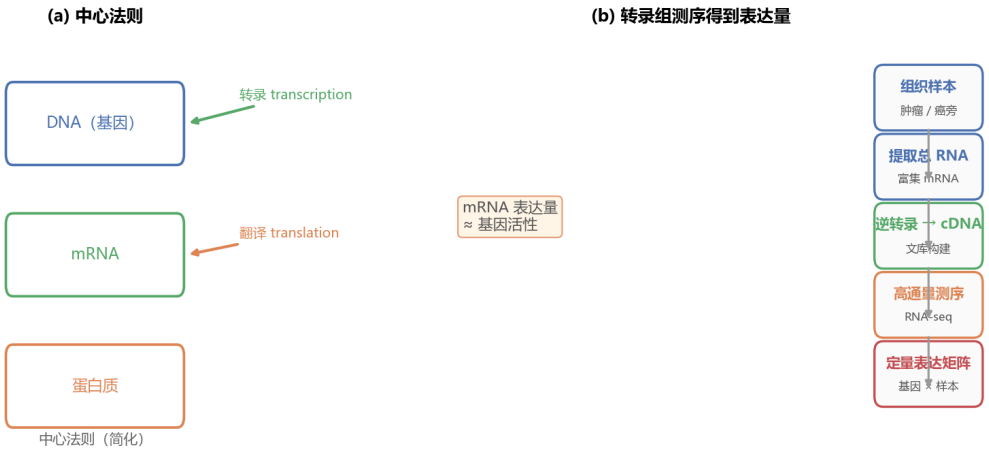


图 2: 图 02 中心法则与转录组测序（示意图）

图 02 为教学示意图（模拟绘制，仅演示概念），左侧展示 DNA → mRNA →蛋白质的信息流向，右侧展示 RNA-seq 如何把 mRNA 变成数字读段计数。读者用真实数据跑出的是自己的图。

3. 细胞死亡方式大家族：凋亡、坏死、焦亡、铁死亡、自噬性死亡

细胞死亡不是一个词，而是一个“大家族”。粗略可分为两大类：意外死亡（如坏死）与程序性死亡（regulated cell death，受基因精确调控的死亡）。五种最常见的死亡方式对比如下（图 03 为示意图）：

死亡方式	英文	核心机制	形态特征	是否引发炎症
凋亡	apoptosis	半胱天冬酶（caspase）级联，细胞“自我拆解”	细胞皱缩、膜出泡、形成凋亡小体，被吞噬清除	通常不引发（“安静地死”）
坏死	necrosis	意外损伤（缺氧、物理/化学损伤）导致膜破裂	细胞肿胀、膜破裂、内容物泄漏	强烈（“爆炸式死”）
焦亡	pyroptosis	gasdermin 家族蛋白在细胞膜上打孔	膜穿孔、细胞肿胀破裂	强烈促炎（“边死边喊救兵”）
铁死亡	ferroptosis	铁依赖性脂质过氧化	线粒体变小、膜完整性尚存	可引发
自噬性死亡	autophagic cell death	自噬（autophagy）过度激活	大量自噬体/自噬溶酶体堆积	较弱

这里有两处需要严谨说明：其一，自噬本身通常是细胞的“自救机制”（把受损的细胞器回收再利用），只有过度或失控的自噬才会导致死亡；其二，凋亡与焦亡虽然都属于程序性死亡，关键区别在于“死法是否安静”——凋亡是拆卸后打包送走，不惊动免疫系统；焦亡是打孔炸开，同时释放促炎因子，相当于“边死边喊救兵”。

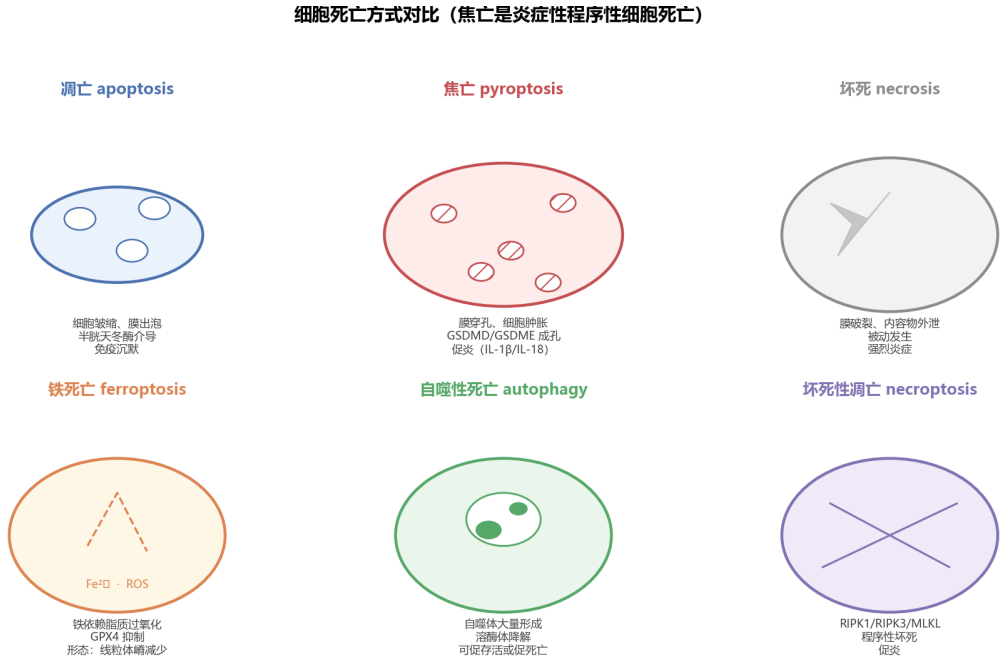


图 3: 图 03 细胞死亡方式对比（示意图）

图 03 为教学示意图（模拟绘制），仅用于直观对比五种死亡方式的形态与炎症特征，

不代表任何真实实验数据。

为什么焦亡近年成为研究热点？主要有三个原因。第一，**机制清晰化**：2015 年邵峰团队在 Nature 杂志报道 gasdermin D（GSDMD）是焦亡的关键执行蛋白，焦亡从此从 “现象” 变成 “可操作的分子通路”；第二，**疾病相关性**：焦亡是感染与炎症性疾病的重要推手，也与肿瘤关系密切；第三，**转化价值**：gasdermin（GSDM）家族蛋白本身成为药物靶点与生物标志物的候选。对肿瘤免疫尤其重要的是：焦亡是一种 “免疫原性” 很强的死亡方式，理论上可以把 “没人理” 的冷肿瘤变成 “被免疫系统围攻” 的热肿瘤——这直接关系到第 5 节的肝癌话题。

4. 细胞焦亡的分子机制详解

焦亡（pyroptosis）的严谨定义：由 gasdermin 家族蛋白介导的、伴随强烈炎症反应的程序性坏死样细胞死亡（lytic cell death）。“坏死样” 指细胞最终会破裂（lysis），“程序性” 指它由特定蛋白精确执行，而不是意外事故。

经典通路（canonical pathway，图 04 左半），一条 “报警→点火→爆破→呼救” 的流水线：

1. **感应**：危险信号（病原体成分如脂多糖 LPS、损伤相关分子、尿酸结晶等）被模式识别受体识别，招募组装 NLRP3 炎症小体（inflammasome）。NLRP3 全名很长，你只需记住它是一个 “报警平台”；
2. **激活**：炎症小体把无活性的 pro-caspase-1 剪切成有活性的 caspase-1（半胱天冬酶-1，一种蛋白酶）；
3. **裂解底物**：caspase-1 同时做两件事——把 GSDMD 裂解成 “N 端成孔片段” 和 “C 端抑制片段”（平时 N 端被 C 端 “锁住”，裂解后解锁），并把无活性的 pro-IL-1 β、pro-IL-18 裂解成成熟的白细胞介素（interleukin）IL-1 β与 IL-18；
4. **打孔与释放**：GSDMD 的 N 端片段在细胞膜上寡聚成环、插入膜内形成孔洞，细胞肿胀破裂，成熟的 IL-1 β/IL-18 连同细胞内其他内容物一起释放，招募并激活免疫细胞，放大炎症。

非经典通路（non-canonical pathway，图 04 右半）：人源的 caspase-4、caspase-5（小鼠中对应为 caspase-11）能直接识别进入胞质的脂多糖（LPS，革兰氏阴性菌外膜成分），被激活后同样裂解 GSDMD 打孔，并间接激活 NLRP3/caspase-1 通路，进一步放大炎症。

> 图 04 为教学示意图（模拟绘制），仅示意经典与非经典两条通路的骨架，省略了大量辅助蛋白细节（如 ASC 接头蛋白等），教学目的而非完整分子图谱。

GSDM 家族还包括 GSDMA、GSDMB、GSDMC、GSDME 等成员。其中 GSDME 能被 “凋亡执行者”caspase-3 裂解——这解释了为什么某些化疗药物会诱导肿瘤细胞发生焦亡样死亡。对初学者而言，先把 GSDMD 这条主线记牢即可。

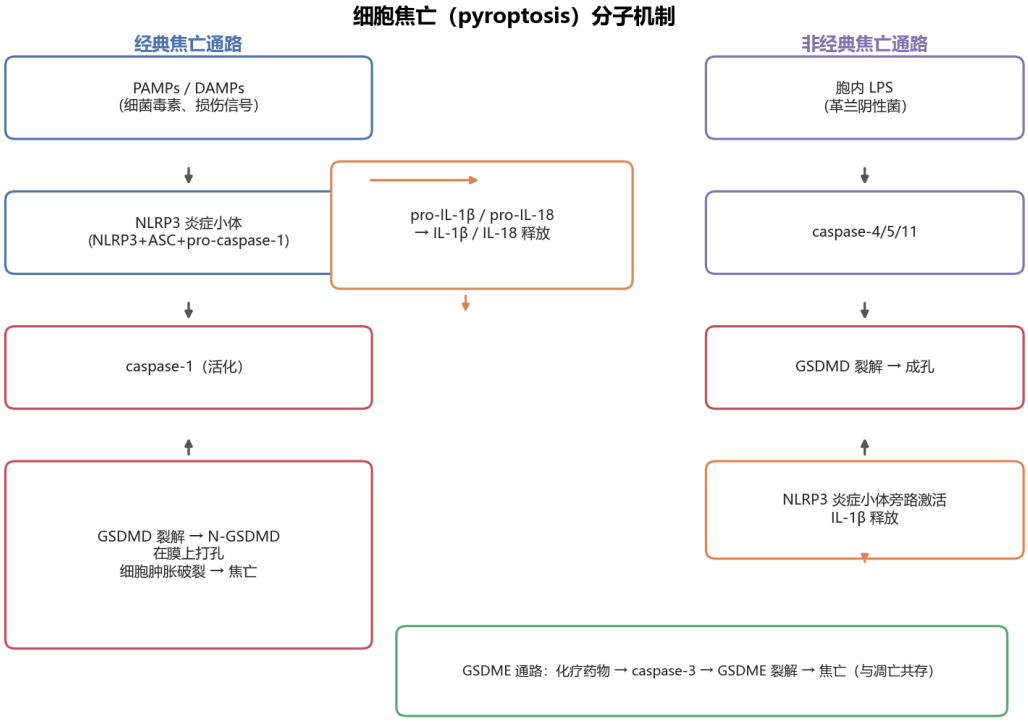


图 4: 图 04 焦亡分子机制（示意图）

5. 肿瘤生物学速览：为什么我们关心“焦亡×肝癌”

肝细胞癌（hepatocellular carcinoma, HCC）是最常见的原发性肝癌类型（约占 75%–85%），也是全球癌症相关死亡的主要原因之一。主要危险因素包括：慢性乙型肝炎病毒（HBV）或丙型肝炎病毒（HCV）感染、酒精性肝病、非酒精性脂肪性肝病（NAFLD，现常称代谢相关脂肪性肝病 MASLD）以及黄曲霉毒素暴露。中国是 HBV 相关性肝癌的高发国家，HCC 因此是国内肿瘤研究的热点。

肿瘤并不只是“一堆疯狂增殖的细胞”。著名肿瘤学家 Hanahan 与 Weinberg 提出“癌症的标志”（hallmarks of cancer），包括：持续增殖信号、逃避生长抑制、抵抗细胞死亡、无限复制潜能、诱导血管生成、激活侵袭转移、重编程能量代谢、逃避免疫摧毁等。注意“抵抗细胞死亡”是其中的标志之一——癌细胞的“抗死能力”是它猖獗的根源，这反过来提示：诱导肿瘤细胞“以正确的方式死亡”（比如焦亡），是潜在的治疗思路。

肿瘤微环境（tumor microenvironment, TME）是肿瘤细胞周围的“小社会”：包括免疫细胞（CD8+ T 细胞、调节性 T 细胞、肿瘤相关巨噬细胞 TAM、NK 细胞等）、成纤维细胞（CAF）、血管内皮细胞、细胞外基质与各类细胞因子。打个比方：免疫细胞是“警察”，肿瘤细胞是“罪犯”，TME 就是警匪之间的博弈场。通常，CD8+ T 细胞浸润多、免疫活跃的肿瘤（俗称“热肿瘤”）预后相对较好，对免疫检查点抑制剂（如 PD-1/PD-L1 抗体）的响应率也更高；而免疫细胞稀少、“冷冰冰”的肿瘤（“冷肿瘤”）则更难被免疫系统清除。

为什么“焦亡×肝癌”有临床意义？第一，HCC 对传统化疗不敏感，免疫治疗（如 PD-1/PD-L1

抑制剂联合抗血管生成药物）虽已进入一线，但仍有相当比例的患者不响应；第二，焦亡释放损伤相关分子模式（DAMPs，damage-associated molecular patterns）和 IL-1 β/IL-18，能招募并激活免疫细胞，有潜力把 “冷” 肝癌变成 “热” 肝癌，增强抗肿瘤免疫；第三，焦亡基因的表达水平可以作为生物标志物（biomarker），帮助预测患者预后或免疫治疗响应；第四，也是本教程最实际的——TCGA-LIHC 与 GSE14520 等公共数据库里就有肝癌患者的表达谱与生存数据，“焦亡基因在肝癌中表达是否异常、哪些与预后相关” 这个问题可以完全用公开数据回答。这正是本教程示例课题要做的事。

本章小结

- 细胞是生命的基本单位；细胞生物学研究细胞的结构、功能与 “生死决策”。
- 中心法则：DNA → RNA →蛋白质；RNA-seq 通过统计读段计数来测量 mRNA 表达量——表达量是蛋白质产量的 “代理指标”，而非严格等同。
- 细胞死亡分意外死亡与程序性死亡两大类；凋亡安静、焦亡炸裂促炎、坏死是意外、铁死亡靠脂质过氧化、自噬性死亡是 “自救过头”。
- 焦亡经典通路：NLRP3 炎症小体→ caspase-1 → GSDMD 裂解→膜打孔→ IL-1 β/IL-18 释放；非经典通路由 caspase-4/5/11 直接感应 LPS。
- HCC 是常见的原发性肝癌，其微环境中免疫细胞与肿瘤博弈；焦亡有望把冷肿瘤变热、焦亡基因可作预后生物标志物——这是示例课题的科学动机。

思考题

1. 为什么说 “mRNA 表达量” 只是蛋白质产量的 “代理指标”？在什么场景下这个代理可能失真？
2. 凋亡和焦亡都是程序性死亡，为什么焦亡会引发炎症而凋亡通常不会？
3. 如果一个焦亡基因在肝癌中高表达，你能列出至少两种互相矛盾的解释吗？如何用数据帮助区分？
4. “焦亡能抗肿瘤” 与 “慢性炎症促肿瘤” 听起来矛盾，你能用本章知识解释这种 “双刃剑” 现象吗？

第 02 章生物信息学知识补充

本章为零基础读者补齐 “生物信息学” 这层知识：生信是什么、RNA-seq 数据长什么样、新手最容易卡住的统计学概念，以及 R 语言的第一行代码。学完本章，你应该能看懂后续章节里 “limma 差异分析”“FDR 校正”“KM 生存曲线”“ROC” 这些词，并亲手跑出第一个 R 脚本。

本章目标

1. 能用 “图书馆” 类比，向别人解释生物信息学是什么、能做什么；
2. 能说出 RNA-seq 的完整流程，并分清 counts、FPKM、TPM 三个量的含义与用途；
3. 能读懂表达矩阵、临床信息表等常见表格文件的格式；
4. 掌握 p 值、FDR/BH 校正、log2FC、limma、富集分析、生存分析、ROC/AUC 这些统计概念的直觉与严谨定义；
5. 能安装 R 与 RStudio，运行第一个 tidyverse 脚本（读取表达矩阵、画散点图）。

1. 生物信息学是什么、能做什么

把细胞想象成一座图书馆：基因组（genome）是全部藏书——人类基因组约有 30 亿个碱基对（base pair, bp），相当于一套 23 本 “巨著”（23 对染色体）；基因是书里的 “章节”；mRNA 是章节的 “复印本”；蛋白质是读者拿走复印本后做出来的 “成品”。传统生物学家像 “手抄员”，一页一页读；高通量测序（high-throughput sequencing）则像一次性把整座图书馆扫描成电子版——一次实验产生 TB 级（1 TB $\approx 10^{12}$ 字节）的数据，人类不可能逐行阅读。

生物信息学（bioinformatics）就是 “海量图书的智能检索与统计”：用计算机与统计学方法，从海量生物学数据中提取规律。它能做的事包括：序列比对与基因组注释、基因表达定量、差异表达分析（找出两组样本中显著变化的基因）、功能富集分析（这些基因属于哪些通路）、生存分析（哪些基因与患者生存相关）、预后模型构建（用多个基因预测患者风险），以及公共数据库挖掘（把别人已发表的数据拿来回答自己的问题）。

本教程走的是最后一种模式——“dry lab”（纯计算实验），不需要自己做湿实验。这正是公共数据库时代给零基础新手打开的大门：你不需要一台测序仪，也能完成一个完整的课题。

2. 测序与表达矩阵：从 reads 到数字

RNA-seq 的标准流程：样本（肿瘤/癌旁组织）→提取 RNA →建库（逆转录成 cDNA、加接头）→高通量测序产生读段（reads）→质控（去除低质量读段与接头）→比对（alignment）到参考基因组/转录组→定量（统计每个基因被多少读段覆盖）→得到表达矩阵。整体见图 02（示意图）：左侧是中心法则，右侧是测序把 mRNA 变成数字的过程。

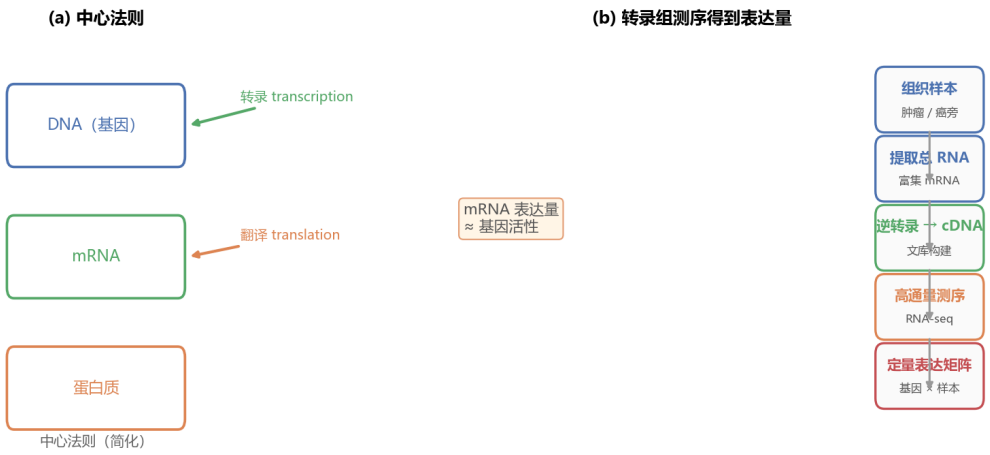


图 5: 图 02 中心法则与转录组测序（示意图）

> 图 02 为教学示意图（模拟绘制），仅演示 “中心法则 + 测序定量” 的概念流程，不代表任何真实实验数据。

测序得到的原始计数叫 counts（读段计数）：每个基因被多少条 read 命中。counts 是整数，但它同时受两个因素干扰：基因越长越容易被命中（长度偏差）；测序越深、所有基因的计数一起变高（文库大小偏差）。因此需要归一化（normalization）才能比较：

- **FPKM**（Fragments Per Kilobase per Million，每千碱基每百万片段）：先按 “每百万读段”（文库大小）归一，再按 “每千碱基”（基因长度）归一。双端测序（paired-end）时统计的是片段（fragment）而非单端读段，所以叫 FPKM；
- **TPM**（Transcripts Per Million，每百万转录本）：先按基因长度归一，再按 “每百万” 归一。FPKM 与 TPM 在数学上可以互相转换，但 TPM 在不同样本间可比性更好，现在更推荐；
- 一句实用口诀：**counts 用于差异分析的统计模型**（limma-voom、DESeq2 等需要 counts 作为输入），**FPKM/TPM 用于展示表达水平和跨样本比较**。

表达矩阵（expression matrix）长什么样？一张表：行是基因（人类约 2 万个蛋白编码基因），列是样本（TCGA-LIHC 约 371 个肿瘤样本 + 50 个癌旁样本），单元格是表达量数字。示意如下：

基因	TCGA-XX-0001（肿瘤）	TCGA-XX-0002（肿瘤）	…	TCGA-XX-00NN（癌旁）
GSDMD	8.21	7.94	…	5.12
NLRP3	4.05	4.66	…	2.30

第一列通常是基因名（gene symbol，如 GSDMD）或 Ensembl ID，后续每列对应一个样本。之后所有分析（差异、富集、生存）都在这张表上展开。补充一点：TCGA 的 RNA-seq 数据提供 FPKM/FPKM-UQ 等归一化值，而 GEO 中有些队列（如 GSE14520）用的是表达芯片（microarray），单位是荧光信号强度而非读段计数——第 06 章会具体处理。

3. 常用数据文件格式

生信里 90% 的数据都是 “表格”，最常见的两种纯文本格式：

- **CSV**（Comma-Separated Values，逗号分隔值）：列与列之间用英文逗号分隔；
- **TSV**（Tab-Separated Values，制表符分隔）：列与列之间用 Tab 分隔（基因名里偶尔会含有逗号，TSV 更保险）。

两者都可以用 Excel/WPS 打开，但文件很大时强烈建议用 R 或专业文本编辑器处理。三个新手常踩的坑：①编码用 UTF-8，否则中文注释会乱码；②缺失值统一记为 NA；③列名不要用空格和特殊符号（用下划线代替）。

教程会用到两类核心表格：

1. **表达矩阵（expression matrix）**：如上节所示，基因×样本；
2. **临床信息表（clinical information table）**：一行一个患者，列包含样本 ID、年龄、性别、TNM 分期、生存时间（overall survival time, OS time，单位月或天）、生存状态（OS status，0 = 存活或删失，1 = 死亡）等。生存分析全靠这张表的最后两列。

GEO 的原始数据包（SOFT/MINiML 格式）和 TCGA 的下载清单会在第 06 章详细介绍；本章先掌握 “表格式文件” 就够用了。

4. 统计分析必备概念（新手最容易卡住的地方）

这一节是本教程的 “统计急救包”。每个概念都先给直觉，再给严谨定义，并说明它在示例课题里出现在哪一步。

0.0.0.1 4.1 p 值

严谨定义：在原假设（null hypothesis，例如 “肿瘤组与癌旁组表达无差异”）为真的前提下，观测到 “当前数据或更极端数据” 的概率。直觉：如果两组其实没有差异，纯靠随机抽样，得到现在这么大差异的概率有多大？这个概率就是 p 值。

新手最容易犯的两个错误：① p 值不是 “两组有差异的概率”，它衡量的是 “证据的强度”，而不是 “效应的大小”；② $p > 0.05$ 不代表 “没有差异”，可能只是样本量太小、检验功效（power）不

足。惯例阈值是 $p < 0.05$，意思是 “如果原假设为真，出现这种结果的概率不到 5%”——这只是约定俗成的门槛，不是数学铁律。

0.0.0.2 4.2 多重检验与 FDR/BH 校正

差异表达分析一次要检验约 2 万个基因。设想所有基因其实都没有差异：按 $p < 0.05$ 的标准，纯随机也会有约 5% 的基因 “碰巧” 显著——也就是约 1000 个假阳性（false positive）。检验做得越多，假阳性越泛滥，这就是多重检验问题（multiple testing problem）。

解决办法是校正（adjustment）。我们控制的是 FDR（False Discovery Rate，错误发现率）：“在被判定为显著的基因里，预期有多少比例其实是假阳性”。最常用的校正是 BH 方法（Benjamini–Hochberg 法），校正后的值叫 adjusted p value（记为 adj.P 或 padj）。本教程的差异基因筛选标准统一为：`|log2FC| > 1 且 adj.P < 0.05 (BH 校正)`。

0.0.0.3 4.3 log2FC

FC 是倍数变化（fold change）：处理组均值 ÷ 对照组均值。FC 有一个毛病：上调 2 倍（$FC = 2$）和下调一半（$FC = 0.5$）在数值上不对称。取以 2 为底的对数后：$\log2(2) = 1$，$\log2(0.5) = -1$，对称且直观。所以 log2FC（log2 fold change）> 1 表示上调超过 2 倍，< -1 表示下调超过 2 倍。

0.0.0.4 4.4 t 检验与 limma 的思想

t 检验（Student’s t-test）比较两组均值是否显著不同，核心思想是：差异（两组均值之差）除以这个差异的 “不确定性”（标准误）。差异越大、数据越稳定（标准误越小），t 值越大，p 值越小。

limma 是生信领域差异分析的 “事实标准” 包：它把每个基因的检验写成一个线性模型（linear model），再用经验贝叶斯（empirical Bayes）方法 “借力”——用全基因组所有基因的信息来稳定单个基因的小样本方差估计，避免 “样本只有 3 个、方差算不准” 的尴尬。对 RNA-seq 的 counts 数据，用 limma-voom 流程（先做 voom 变换再跑 limma）。你现在不需要会推导，只要记住：**limma 的本质是 “更稳的 t 检验”**。

0.0.0.5 4.5 富集分析的统计学思想

拿到一张差异基因列表后，自然的问题：这些基因是不是 “扎堆” 出现在某些通路（如焦亡、IL-1 β 信号、细胞因子）里？富集分析（enrichment analysis）回答的就是这个问题。

它的统计思想是 “抽球问题”：假设全基因组有 N 个基因，其中某通路有 M 个成员；你挑了 n 个差异基因，其中属于该通路的有 k 个。问：如果差异基因是随机挑的，出现 “k 个或更多通路基因” 的概率是多少？这个概率由超几何分布（hypergeometric distribution）给出；等价的实现方式是 2×2 列联表上的 Fisher 精确检验（Fisher’s exact test）或卡方检验（chi-square test）。概

率小，说明差异基因在该通路里“富集”了。常用工具如 clusterProfiler 包（支持 GO 与 KEGG 富集），第 09 章会手把手教。

0.0.0.6 4.6 生存分析基本概念

生存分析（survival analysis）研究“事件发生的时间”——在本课题里事件就是患者死亡。三个核心概念：

- **KM 生存曲线**（Kaplan–Meier curve）：把患者按某个特征（如焦亡基因高/低表达）分组，逐段估计“活到各时间点的比例”，画出阶梯状曲线；用 log-rank 检验比较两条曲线是否有显著差异（给出 p 值）。曲线越高，生存越好；
- **删失**（censoring）：患者随访结束或失访时还没观察到死亡，我们不能说“他活着”，只能说“他至少活到了这个时点”。删失数据仍然有效，KM 法就是为处理删失而设计的；
- **HR 与 Cox 回归**：风险比（hazard ratio, HR）是 Cox 比例风险回归（Cox proportional hazards regression）输出的核心量。直觉：HR = 该组单位时间死亡风险÷对照组。HR $>$ 1 表示该因素增加死亡风险（预后差），HR $<$ 1 表示降低风险（预后好）；通常报告 95% 置信区间（confidence interval, CI），如 HR = 1.8（95% CI 1.2–2.7）。多因素 Cox 则可以同时校正年龄、分期等混杂因素（confounder），看看焦亡基因是否“独立”于这些因素仍有预后价值。

0.0.0.7 4.7 ROC 与 AUC

当我们用某个指标（如风险评分）预测“患者 3 年内会不会死亡”时，怎么评价预测准不准？ROC 曲线（receiver operating characteristic curve，受试者工作特征曲线）横轴是“1 − 特异度”（假阳性率），纵轴是“灵敏度”（真阳性率），曲线上每个点对应一个判断阈值。AUC（Area Under the Curve）是曲线下面积：AUC = 0.5 相当于瞎猜（纯随机），AUC 越接近 1 预测越准，通常 AUC $>$ 0.7 认为有较好的区分能力。它回答的问题是：“把高风险和低风险患者分开的能力有多强？”

5. R 语言零基础入门

0.0.0.8 5.1 为什么生信用 R，RStudio 长什么样

R 是免费开源的统计编程语言，生信圈选它的原因很实在：统计生态最全（t 检验、Cox 回归开箱即用）、有专门的生信包仓库 Bioconductor（limma、DESeq2、clusterProfiler 都在里面）、绘图能力强（ggplot2）、且完全可复现——脚本跑一遍，结果一致，这是论文写作的基本要求。

RStudio 是 R 的“驾驶舱”，界面分四块：左上脚本编辑器（source，写代码的地方，保存为.R 文件）、左下控制台（console，代码运行与结果显示）、右上环境窗格（environment，显示已创建

的变量/数据）、右下文件/绘图/包/帮助窗格（plots 里显示画出的图）。# 开头是注释，不会被运行；Windows 下按 Ctrl+Enter 运行光标所在行。

0.0.0.9 5.2 数据结构与 tidyverse 精神

R 里最常用的两种数据结构：**向量**（vector，一维数据，用 `c()` 创建）和**数据框**（data frame，二维表格，行 = 观测、列 = 变量）。tidyverse 是一套“数据科学全家桶”（ggplot2、dplyr、readr、tidyr 等），它的精神是**整洁数据**（tidy data）：每一行是一个观测，每一列是一个变量，每个单元格是一个值。tidyverse 还有一个标志性符号**`%>%` 管道**（读作“然后”）：把左边的结果传给右边函数的第一个参数，让代码像流水线一样从左往右读。

```
# 向量：一维数据
x <- c(1, 2, 3, 4)

# 数据框：二维表格
df <- data.frame(
  gene  = c("GSDMD", "NLRP3"),
  logFC = c(1.2, -0.8)
)

# 三个最常用的"查看"函数
str(df)   # 查看结构
head(df)  # 查看前几行
dim(df)   # 查看维度（行数 × 列数）
```

0.0.0.10 5.3 第一个脚本：读取表达矩阵、画散点图

下面这个脚本是完整的、可运行的（先给“模拟数据”版本，无需任何文件即可跑通；再给“读取真实文件”版本，第 06 章会教你如何得到那个文件）。安装包需要联网，`install.packages("tidyverse")` 执行一次即可。

```
# ============ 第一个脚本：表达矩阵 + 散点图 ============

# 0) 安装（联网时执行一次即可，tidyverse 是真实存在的 R 包合集）
# install.packages("tidyverse")

# 1) 加载
library(tidyverse)

# 2) 方式 A：用模拟数据先跑通流程（无需任何外部文件）
set.seed(42)  # 固定随机种子，保证每次运行结果一致（可复现）
expr_demo <- tibble(
  sample = 1:50,
```

```
  GSDMD  = rnorm(50, mean = 8, sd = 1.5),   # 模拟 50 个样本的 GSDMD 表达
  NLRP3  = rnorm(50, mean = 8, sd = 1.5),   # 模拟 NLRP3 表达
  group  = rep(c("tumor", "normal"), each = 25)  # 前 25 个是肿瘤，后 25 个是癌旁
)

# 3) 画散点图：X = GSDMD，Y = NLRP3，按分组着色
#    %>%，读作"然后"：把 expr_demo 传给 ggplot 的第一个参数
expr_demo %>%
  ggplot(aes(x = GSDMD, y = NLRP3, color = group)) +
  geom_point(size = 2, alpha = 0.7) +
  labs(title = "GSDMD 与 NLRP3 表达散点图（示例数据）",
       x = "GSDMD 表达量", y = "NLRP3 表达量") +
  theme_minimal()

# 4) 方式 B：读取真实的表达矩阵（第 06 章会教你如何得到这个文件）
# expr <- read_csv("data/expr_matrix.csv")  # 文件存在时取消注释即可运行
# head(expr)
# dim(expr)
```

跑完后，右下角 plots 窗格会出现一张按"肿瘤/癌旁"着色的散点图。如果看到图，恭喜——你已经完成了生信入门的第一公里。后续章节会在这个基础上逐步加入 limma、survival、clusterProfiler 等更专业的包，但"读数据→处理→画图"这个循环，从现在起会一直陪着你。

本章小结

- 生物信息学 = 用计算机与统计学处理海量生物学数据；"dry lab"让新手也能用公共数据做完整课题。
- RNA-seq 流程：样本→建库→测序→质控→比对→定量→表达矩阵；counts 做统计，FPKM/TPM 做展示。
- 表达矩阵（基因×样本）与临床信息表（一行一患者）是本课题的两张核心表格。
- 统计急救包：p 值衡量证据强度而非效应大小；多重检验需 FDR/BH 校正；log2FC 让上下调对称；limma 是"更稳的 t 检验"；富集分析基于超几何分布/Fisher 精确检验；生存分析看 KM 曲线、HR 与 Cox 回归；ROC/AUC 评价预测区分能力。
- R + RStudio + tidyverse 是零基础最平滑的入口；`%>%` 管道让代码像流水线；第一个脚本能画出第一张图。

思考题

1. counts、FPKM、TPM 分别回答什么问题？为什么做差异分析通常用 counts 而不是 TPM？
2. 如果 2 万个基因其实都没有差异，用 $p < 0.05$ 筛选，预期会有大约多少个假阳性？FDR 校

正解决的是什么问题？

3. 一句报告写着”HR = 1.8（95% CI 1.2–2.7, p = 0.003）“，该怎么向非专业人士解读？
4. 一个预后模型的 AUC = 0.85，另一个 AUC = 0.55，分别意味着什么？AUC = 0.5 又代表什么？

第 03 章公共数据库详解

本章属于 “Track B：数据基础”，为第 06 章的数据获取做知识准备。建议先读完第 01、02 章再进入本章。

本章目标

学完本章，你将能够：

1. 说清楚公共生物数据库的数据从哪来、为什么可以免费用于科研；
2. 认识本教程的两大数据来源——TCGA-LIHC 与 GSE14520（GEO），并知道各自的访问入口；
3. 理解 GEO 中 GDS/GSE/GSM/GPL 的层级关系，会用 accession 编号检索数据集；
4. 了解 GO、KEGG、STRING、HPA、GTEx 等常用数据库分别能回答什么问题；
5. 学会根据 “要回答的科学问题” 选择正确的数据库。

3.1 公共数据库生态总览

说明：上图是教程绘制的示意图，仅用于展示数据库之间的分工关系；图中每个数据库的详细内容以官方页面为准。

数据从哪来？世界各地的科研机构（尤其是美国国立卫生研究院（NIH）系统资助的大型项目）按 “数据共享” 要求，把论文背后的数据 “上交” 到公共数据库。以 TCGA 为例：它收集上万例癌症患者的肿瘤组织，完成测序与临床随访后，把表达矩阵、突变、临床表等公开发布，任何研究者都可下载。

为什么免费？这些研究主要由纳税人资助，成果属于公共资源；期刊与基金也要求数据共享。你可以免费下载使用，唯一要求是发表论文时正确引用（见第 13 章）。一句话：**免费但要用，用了要 cite**。

常用数据库按功能分类（对应上图）：

- 原始与表达数据：GEO、TCGA（GDC）、GTEx；
- 临床与多组学整合：TCGA、UCSC Xena、cBioPortal；
- 功能注释与通路：GO、KEGG；
- 蛋白互作与蛋白表达：STRING、HPA；
- 在线分析工具：TIMER、GEPIA。

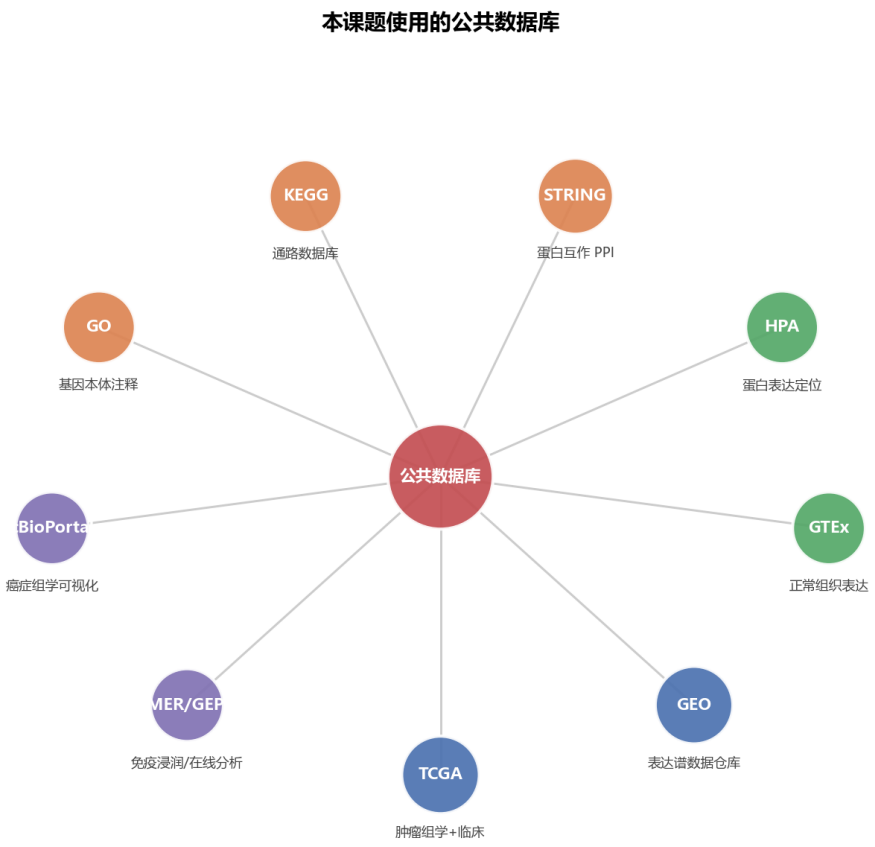


图 6: 图 05 公共数据库全景（示意图，以各数据库官方页面为准）

本教程的核心数据其实只需要两处：**TCGA-LIHC**（发现队列）和 **GSE14520**（验证队列），其余数据库用于辅助分析与验证。

3.2 TCGA：癌症基因组图谱

TCGA（The Cancer Genome Atlas，癌症基因组图谱）由美国国家癌症研究所（NCI）与国家人类基因组研究所（NHGRI）联合发起，自 2006 年起覆盖 33 种癌症类型（以官方发布为准），每种癌症同时提供多组学数据与临床随访信息。

TCGA-LIHC：肝细胞癌（Liver Hepatocellular Carcinoma, LIHC）队列，是肝癌研究中引用最多的公共数据之一。本教程使用的是其 RNA-seq 表达谱与临床随访数据，约 **371 例肿瘤组织 + 50 例癌旁正常组织**（样本数以 TCGA 官方发布为准）。

TCGA 提供哪些数据类型？

- **RNA-seq 表达量**：如 HTSeq-Counts（原始计数）与 HTSeq-FPKM（标准化表达量）；
- **临床随访**：生存状态、生存时间、TNM 分期、年龄、性别等；
- **体细胞突变**（MAF 文件）、拷贝数变异（CNV）、DNA 甲基化。

怎么访问？三个常用入口，用途不同：

1. **GDC Data Portal（官方渠道）**：https://portal.gdc.cancer.gov ，TCGA 数据的官方发布站点，可精确筛选后下载，但对新手来说筛选项较多；
2. **UCSC Xena**：https://xena.ucsc.edu ，把 TCGA 等数据整理成 “基因×样本” 的表达矩阵和 “样本×字段” 的临床表，下载最方便，第 06 章就用它；
3. **cBioPortal**：https://www.cbioportal.org ，网页可视化见长（突变、表达、生存曲线），适合先探索，不必下载。

   注意：TCGA 大部分**处理后的数据**（表达矩阵、临床表）公开可下载；少数**受控数据**（部分原始测序文件、部分临床字段）需要向 dbGaP（基因型与表型数据库）提交申请，以 GDC 官方说明为准。

   本教程只使用 TCGA-LIHC 的表达与临床两类数据，突变、拷贝数、甲基化等其他类型先了解即可。

3.3 GEO：基因表达综合数据库

GEO（Gene Expression Omnibus，基因表达综合数据库）由美国国家生物技术信息中心（NCBI）维护，地址：https://www.ncbi.nlm.nih.gov/geo ，是全世界最大的基因表达数据库。芯片时代的绝大多数表达数据以及大量 RNA-seq 数据都存在这里。

GEO 的四级结构（新手最容易混淆，务必记牢）：

- **GPL（Platform，平台）**：芯片/测序平台，如 GPL3921（Affymetrix 某款人类全基因组芯片）。平台决定 “用什么尺子量表达量”；
- **GSE（Series，系列）**：一项研究的所有样本合在一起，是下载的基本单位，如 GSE14520；
- **GSM（Sample，样本）**：单个样本的表达数据；
- **GDS（DataSet，数据集）**：把某个 GSE 按统一平台整理成便于比较的子集（不是每个 GSE 都有 GDS）。

  本教程直接使用 GSE 层级的系列矩阵，不需要用到 GDS。

一个类比：GSE 是一 “期杂志”，GSM 是其中的 “每篇文章”，GPL 是 “印刷用的纸张规格”，GDS 是杂志社编的 “精选合集”。

GSE14520：本教程的验证队列。它由复旦大学生物医学研究院等团队提交，是肝细胞癌（HCC）表达谱数据集，包含肿瘤与癌旁样本，并**带有生存随访**——这正是它能做 “生存验证” 的原因。其芯片平台为 GPL3921，探针 ID 需映射为基因 Symbol（第 06 章实操）。具体样本数与字段以 GEO 页面为准。

怎么检索？打开 GEO 主页，在搜索框输入 **accession（登录号）**，如 `GSE14520`——它是数据集的 “身份证号”，比名字可靠。页面底部有 “Series Matrix File(s)” 下载入口；R 的 GEOquery 会自动获取这些文件（第 06 章）。

GEO 页面上会看到什么？打开 GSE14520 的页面，主要元素包括：标题与摘要（这个研究做了什么）、“Platforms”（平台，如 GPL3921 的链接）、“Samples”（该系列包含的所有 GSM 样本列表）、以及底部的下载文件区（Series Matrix File(s) 和 Supplementary files）。其中 Series Matrix File 就是我们在 R 里用 GEOquery 读取的文件，格式为文本表格。

3.4 功能注释与通路库：GO 与 KEGG

拿到差异基因列表之后，你想知道 “这些基因在做什么”，就需要功能注释数据库。

GO（Gene Ontology，基因本体）：https://geneontology.org ，用一套受控词汇描述基因功能，分三个分支：

- **BP（Biological Process，生物学过程）**：基因参与的过程，如 “细胞焦亡”（pyroptosis）、“炎症反应”；
- **CC（Cellular Component，细胞组分）**：基因产物所在的细胞位置，如 “质膜” “线粒体”；
- **MF（Molecular Function，分子功能）**：基因产物的分子活性，如 “蛋白结合” “激酶活性”。

KEGG（Kyoto Encyclopedia of Genes and Genomes，京都基因与基因组百科全书）：https://www.kegg.jp ，把基因映射到代谢与信号通路上，回答 “这些基因富集在哪些通路”，如 NF-κB 信号通路、TNF 信号通路等。

一句话记忆：**GO 管 “功能分类”，KEGG 管 “通路”**。第 09 章会用 clusterProfiler 自动完成

GO/KEGG 富集分析。

在本课题中，GO/KEGG 的作用是把 “焦亡基因集与差异表达基因的交集” 放到功能语境里解释：例如这些基因富集在 “炎症反应”“NF-κB 信号通路” 等条目，就能说明焦亡基因可能通过哪些机制影响 HCC 的进展（第 09 章实操）。

3.5 蛋白互作与蛋白表达：STRING 与 HPA

STRING：https://string-db.org ，蛋白-蛋白互作（protein-protein interaction, PPI）数据库。输入基因列表，它综合实验、文献等证据给出互作网络，第 10 章用它找 hub 基因（关键节点）。

HPA（Human Protein Atlas，人类蛋白图谱）：https://www.proteinatlas.org ，用免疫组化、质谱等手段展示蛋白在人体组织中的真实表达。它提醒我们：**mRNA 表达高≠蛋白表达高**。想验证某个基因在蛋白水平是否真的高表达时，就去查 HPA。

3.6 其他常用数据库

- **GTEx（Genotype-Tissue Expression，基因型-组织表达）**：https://gtexportal.org ，健康人（非肿瘤）各组织的表达基线。想知道 “某个基因在正常肝里应该有多高”，用它做参照；
- **TIMER**：https://timer.cistrome.org ，在线免疫浸润分析工具，输入基因可看它与肿瘤中免疫细胞浸润丰度的相关性（以官网当前版本为准）；
- **GEPIA**：https://gepia.cancer-pku.cn ，北大开发的在线工具，无需下载即可做表达对比与生存分析；
- **UCSC Xena**：https://xena.ucsc.edu ，整合 TCGA、GTEx 等数据，下载与在线分析两用（见 3.2 节）。

3.7 如何选择数据库回答你的科学问题

原则：先有科学问题，再选数据库。下面是一张快速决策表（示意）：

我想知道……	用哪个数据库	数据类型
基因在肿瘤 vs 癌旁中的表达差异	TCGA、GEO、GTEx	表达矩阵
基因与患者生存的关系	TCGA、GEO（带随访）	表达 + 临床
基因有哪些突变	TCGA（GDC / cBioPortal）	突变 MAF
这些基因参与什么通路	GO、KEGG	注释 / 通路
蛋白之间怎么相互作用	STRING	PPI 网络

我想知道……	用哪个数据库	数据类型
蛋白水平是否真的高表达	HPA	蛋白表达
免疫细胞浸润情况	TIMER、GEPIA	在线分析

本教程的选库逻辑：科学问题是 “焦亡相关基因在 HCC 中表达是否异常、哪些与预后相关、能否构建预后风险模型”。因此需要：

1. 一个带肿瘤/癌旁分组 + 生存随访的表达数据集→ **TCGA-LIHC**（发现队列）；
2. 另一个独立数据集做验证→ **GSE14520**（验证队列，同样带生存）。

用两个独立数据集回答同一问题，是生信挖掘论文最常见的证据增强方式（第 12 章会详细讲验证）。

新手常见误区：先下载了一堆数据，再想 “能做什么”——这会导致分析方向混乱。正确顺序是先写下你的科学问题，再决定需要什么类型的数据，最后才去下载。数据不是越多越好，够回答问题即可。

选库不是一次性的：分析过程中发现缺数据（比如缺正常组织表达基线），随时回到这张表补选即可。

本章小结

- 公共数据来自科研机构的共享义务：免费使用，但发表论文时必须正确引用；
- 本教程数据核心是 TCGA-LIHC（发现队列）与 GSE14520（验证队列）；
- GEO 层级：GPL（平台）→ GSE（系列）→ GSM（样本），accession 是检索的 “身份证号”；
- GO 分 BP/CC/MF 管功能分类，KEGG 管通路；STRING 管蛋白互作，HPA 管蛋白水平验证；
- 选库原则：先有科学问题，再对照 “数据类型→数据库” 选库。

思考题

1. 为什么 “mRNA 表达高” 不等于 “蛋白表达高”？如果你想验证某个焦亡基因在蛋白水平的表达，应该用哪个数据库？
2. GSE 和 GSM 有什么区别？为什么下载数据时通常以 GSE 为单位？
3. 如果你的课题想研究 “基因突变频率与免疫治疗疗效”，除了 TCGA，还可能需要哪些数据库？
4. 试着用 accession 检索 GSE14520，看看它的平台（GPL）是什么、样本数有多少，与本章描述是否一致。

第 04 章课题设计与选题

“一个好的生信课题，一半靠分析，一半靠选题。”——生信圈流传的话

本章教你：什么课题能发表、怎么从零设计一个课题、以及如何把“焦亡×肝癌”这个示例课题扩展到属于你自己的版本。

本章目标

1. 理解“生信数据挖掘”类论文的常见类型与发表价值；
2. 掌握选题的四个基本原则（热点、数据、可验证、可发表）；
3. 学会用 PICO 框架把模糊想法变成具体课题；
4. 能独立设计出 1-2 个属于自己的候选课题。

4.1 什么样的生信课题有发表价值？

先破除一个迷思：**“生信课题 = 跑流程”是错的**。能发表的课题，本质上回答了一个**具体、有意义、可验证的科学问题**，跑流程只是实现手段。

按“科学问题的新颖度×工作量”两个维度，新手常见课题可分为三档：

档次	模式	例子	发表难度	适合谁
入门档	单基因/单机制×单癌种挖掘	“GSDMD 在肝癌中的表达与预后”	中低（需要讲出新意）	本科生毕设
进阶级	基因集×单癌种×预后模型	“焦亡相关基因构建肝癌预后模型”（本教程示例）	中（需要独立验证 + 一定工作量）	研究生第一篇文章
挑战级	多组学/单细胞/泛癌×机制闭环	“焦亡在泛癌中的免疫微环境重塑 + 实验验证”	高（需实验或大量计算）	进阶者

本教程选择进阶级：工作量适中、流程完整、有独立验证、有预后模型——这是新手“跳一跳够

得着” 的最佳平衡点。

为什么 “XX 相关基因×癌症×预后模型” 模式能发表？

1. **临床价值明确**：肿瘤预后预测是真实临床需求（医生需要判断哪些患者预后差、需要更积极治疗）；
2. **数据可及**：TCGA 覆盖 33 种癌、GEO 有海量表达谱，全部免费；
3. **机制有故事**：任何生物学过程（焦亡、铁死亡、自噬、衰老、糖酵解……）在肿瘤中都 “应该有” 角色，story 天然存在；
4. **审稿人熟悉**：模式成熟，审稿标准清晰，新手不容易踩 “方法不成立” 的雷。

   诚实提示（写进论文讨论也要用）：这类论文的同质化风险在增加，越来越多的期刊开始要求**独立验证队列**（GEO 验证）甚至**简单实验验证**（qPCR/WB）。所以我们教程把 “独立验证” 作为标配步骤（第 12 章），这既是科学严谨，也是发表策略。

4.2 选题四原则

把任何想法过一遍这四关，能过三关以上就值得做：

原则一：热点（Hot topic）

- 生物学过程要 “热”：焦亡（pyroptosis）、铁死亡（ferroptosis）、自噬（autophagy）、衰老（senescence）、细胞周期（cell cycle）……这些过程有清晰的基因清单、有大量文献、审稿人懂。
- 怎么判断热不热：去 PubMed 搜 “XX-related genes cancer prognosis”，文章数量近 3 年是否持续增长。
- **注意**：热点也有生命周期。2020-2023 年焦亡/铁死亡最热，现在依然可做，但必须加新意（新癌种、新角度、新验证）。

原则二：数据可及（Data availability）

- 目标癌种在 TCGA 有没有队列？（33 种癌全覆盖，但样本量差异大：LIHC 371 例、UVM 80 例）
- GEO 有没有该癌种的**带生存数据**的独立队列？（GSE14520 之于 HCC 是经典选择）
- 没有数据的癌种再好的想法也做不了——**先查数据，再定课题**。

原则三：可验证（Falsifiable）

你的课题必须能回答“是/否”：-（是）可验证：“焦亡相关基因在 HCC 中差异表达，且与预后相关”——能用数据检验；-（否）不可验证：“焦亡在肝癌发生中起重要作用”——太泛，无法证伪。

原则四：可发表（Publishable）

- 工作量是否完整：差异→富集→ PPI →生存→验证，缺一环审稿人就会挑；
- 是否有“增量”：新癌种？新基因集？新分析角度（免疫浸润/单细胞）？完全复制别人做过的同一个癌种同一个基因集很难发表。

4.3 PICO 框架：把想法变成课题

医学研究经典的 PICO 框架，生信挖掘同样适用：

字母	含义	本教程示例
P	Population 人群/疾病	肝细胞癌（HCC）患者
I	Intervention 关注的暴露/因素	细胞焦亡相关基因的表达水平
C	Comparison 比较	肿瘤 vs 癌旁；高表达组 vs 低表达组
O	Outcome 结局	差异表达、富集通路、总生存期（OS）、预后风险

把四格填满，课题一句话就出来了：

> **在肝细胞癌患者中（P），焦亡相关基因的表达（I）相比癌旁组织（C），是否存在差异表达、富集到哪些通路，并与患者总生存期（O）相关？**

试试把你自己的想法套进这个框架——比如把“焦亡”换成“铁死亡”、“肝细胞癌”换成“乳腺癌”，你就有了一个新课题的雏形。

4.4 完整课题设计：本教程示例逐项拆解

4.4.1 研究问题与假设

- **主问题**：焦亡相关基因在 HCC 中的表达与预后价值如何？
- **可操作的子问题**：
  - Q1：焦亡基因在 HCC 肿瘤 vs 癌旁中哪些差异表达？
  - Q2：差异焦亡基因富集在哪些通路？
  - Q3：哪些焦亡基因是 PPI 网络的 hub 节点？

- Q4：hub 焦亡基因能否预测患者生存？能否构建风险模型？
- Q5：风险模型在独立队列中是否仍有效？

4.4.2 变量定义（分析前必须写清楚）

变量	定义	数据来源
结局	总生存期 OS（月）；事件 = 死亡	TCGA-LIHC clinical
分组	Normal / Tumor	TCGA-LIHC 样本类型
表达	RNA-seq FPKM（log2 转换）	UCSC Xena / GDC
焦亡基因集	文献来源的焦亡基因列表（示例：GSDMD、GSDME、CASP1、IL1B、IL18、NLRP3、AIM2、PYCARD、GSDMA/B/C 等；正式研究需给出确切来源文献与列表）	文献补充材料

关键提醒：焦亡基因集的“出处”必须可追溯。真实研究中应从特定文献的补充材料（Supplementary Table）提取完整列表，并在论文方法中注明来源文献。教程演示使用示例列表跑通流程，正式研究请替换为完整、有出处的列表。

4.4.3 统计与分析计划（写进设计文档，防止事后乱改）

1. 差异表达：limma（BH 校正，|log2FC|>1 且 adj.P<0.05）；
2. 富集：clusterProfiler（GO/KEGG，qvalueCutoff=0.05）；GSEA（FDR<0.25 常规阈值）；
3. hub 基因：STRING + Cytoscape cytoHubba（MCC 算法 top10-20）；
4. 生存：中位数分组→ KM + log-rank；单因素 Cox →多因素 Cox 校正；
5. 模型：LASSO（glmnet）选基因构建风险评分；timeROC 评估 1/3/5 年 AUC；
6. 验证：GSE14520 独立队列重复 4-5 步关键分析。

4.4.4 预期结果与备选方案

- **预期**：多数焦亡基因在肿瘤中表达上调，富集于炎症小体/IL-1 β 相关通路；hub 基因（如 GSDMD）高表达与不良预后相关，风险模型 AUC>0.7。
- **备选**：若个别基因方向与预期相反，如实报告（肿瘤中下调也可能有故事：免疫逃逸）；若模型 AUC 低，尝试换基因集/换截点方法（但警惕数据窥探，见 4.6）。

4.5 用“课题设计文档”固定你的计划

动分析前，写一页纸的 design.md（本教程的项目目录 experiments/ 里就有范例结构）：

```
# 课题设计：焦亡相关基因在 HCC 中的预后价值

## 科学问题
（一句话 PICO 表述）

## 数据
- TCGA-LIHC：表达 + 临床（下载日期、版本）
- GSE14520：验证队列（accession、平台）

## 分析步骤（与参数）
1. ...（limma，|log2FC|>1，adj.P<0.05）
2. ...

## 预期结果 / 备选方案
...

## 时间表
...
```

设计文档的价值：**防止分析中“凭感觉改参数”**——所有阈值在动手指之前就定好，这是可复现性（reproducibility）和防数据窥探的第一步。

4.6 学术诚信红线（必读）

1. **不挑结果**：不要反复尝试不同截点/不同基因集直到”P<0.05” 为止（数据窥探 bias / p-hacking）——审稿人问 “为什么选这个截点？” 时你要答得上来，答不上来就是问题；
2. **不隐瞒阴性结果**：生存不显著的基因如实写 “无显著关联”，这也是发现；
3. **基因集有出处**：焦亡基因列表来自哪篇文献必须写清楚，不许 “凭感觉凑”；
4. **独立验证是标配**：训练集上再好的模型，没有独立验证就是空中楼阁。

本章小结

- 能发表的生信课题 = 热点生物学过程 + 可及数据 + 可验证问题 + 完整工作量；
- PICO 框架（人群/因素/比较/结局）能把模糊想法变成具体课题；
- 分析开始前写好设计文档，固定阈值，防止数据窥探；
- “焦亡×肝癌×预后模型” 是本教程示例，你可以用同样的框架换成任何热点过程与癌种；
- 独立验证不是可选项，是发表的基本盘。

思考题

1. 用 PICO 框架设计一个你自己的课题（换一个生物学过程或癌种），写在纸上；
2. 去 PubMed 用 “ferroptosis-related genes prognosis” 检索，观察近 3 年文章数量趋势，判断该方向是否还 “热”；
3. 为什么说 “只报告显著的生存分析结果” 是不道德的？设计文档如何帮助避免这个问题？
4. 你选的癌种在 TCGA 有多少样本？在 GEO 有没有带生存数据的独立队列？（提示：先在脑子里想，第 06 章教你实际查）

第 05 章环境搭建与工具链

本章属于 “Track B：数据基础”。请在第 03 章了解了数据来源之后，先把分析环境搭好，再进入第 06 章的数据获取。

本章目标

学完本章，你将能够：

1. 在自己的 Windows 电脑上安装 R 和 RStudio；
2. 用 `install.packages()` 与 `BiocManager::install()` 安装教程所需的全部 R 包；
3. 认识本教程核心 R 包各自的用途；
4. 用 RStudio Project 组织 “脚本 / 数据 / 结果 / 图” 的项目目录；
5. 用 `sessionInfo()` 记录分析环境，保证结果可复现。

5.1 需要装什么：R 与 RStudio

本教程的分析用 **R 语言**完成，配套使用 **RStudio**（一个更友好的 R 开发界面）。推荐安装 **R 4.x**（如 4.3.x、4.4.x，以官网当前稳定版为准）。

为什么用 R？因为生信挖掘——尤其是 GEO 数据下载、差异表达、富集分析、生存分析——的主流工具链是 R：GEOquery、limma、clusterProfiler、survival 等包在 R 生态里最成熟，论文方法学也几乎都用 R 描述。Python 在深度学习、单细胞分析领域更强，但本教程用不到，属于 “可选装”：学有余力再装，不影响后续任何章节。

> 提示：如果电脑上装过旧版 R（如 3.x），建议先卸载再装新版，避免包版本混乱。

5.2 一步一步安装（Windows 为例）

第 1 步：安装 R。

1. 打开 R 官方下载站 https://cran.r-project.org ，点击 “Download R for Windows” → “install R for the first time”；
2. 下载安装包，双击运行，一路 “下一步”（Next），全部用默认选项即可。默认安装到 `C:\Program Files\R\R-4.x.x`，不用修改；

3. 装完后开始菜单会出现 R 的快捷方式（本教程不直接使用它，我们使用 RStudio）。

第 2 步：安装 RStudio。

1. 打开 https://posit.co/download/rstudio-desktop/ ，下载 Windows 版安装包；
2. 双击运行，同样一路默认安装；
3. 打开 RStudio：它会自动找到电脑上已装的 R。RStudio 界面分为四个窗格：左上脚本编辑区、左下控制台（Console）、右上环境（Environment）、右下文件 / 图 / 帮助。

验证安装：在 RStudio 左下角 Console（控制台）输入：

```
R.version.string
```

回车后能看到类似 `R version 4.4.x (2025-xx-xx)` 的输出，即安装成功。

装完常见问题：如果打开 RStudio 提示找不到 R，多半是 R 未安装或安装不完整，重装 R 后重启 RStudio 即可；如果 RStudio 界面是英文，不影响使用（Tools → Global Options 里可改语言，也可不改）。

5.3 包管理：安装本教程需要的全部 R 包

R 的“包”（package）是别人写好的函数集合。装包有两条命令：

```
# 从 CRAN（R 官方包仓库）安装
install.packages("ggplot2")

# 从 Bioconductor（生物信息学包仓库）安装
# 第一步：先安装包管理器 BiocManager
install.packages("BiocManager")
# 第二步：用 BiocManager 安装 Bioconductor 的包
BiocManager::install("GEOquery")
```

一次性装齐本教程的核心包（把下面整段复制进 Console，回车执行；首次安装可能需要几分钟到几十分钟，请耐心等待，不要中途关闭）：

```
# 1. 包管理器
install.packages("BiocManager")

# 2. Bioconductor 包
BiocManager::install(c("GEOquery", "limma", "clusterProfiler",
                       "org.Hs.eg.db", "survival", "survminer"))

# 3. CRAN 包
```

```
install.packages(c("glmnet", "pheatmap", "ggplot2", "timeROC", "survivalROC"))
```

> 提示：Windows 上装包需要联网。个别包在编译时需要 RTools（https://cran.r-project.org/bin/windows/Rtools/ 下载安装即可），但绝大多数包都有预编译的 Windows 二进制版本，不需要 RTools。
>
> 装包失败先看报错：最常见的是网络问题（可换国内镜像源，如 `install.packages("ggplot2", repos = "https://mirrors.tuna.tsinghua.edu.cn/CRAN/")`）或包名拼写错误。

核心包用途表（本教程会逐章用到，先混个脸熟）：

R 包	用途	主要出现在
GEOquery	下载并读取 GEO 数据（GSE 系列矩阵）	第 06 章
limma	差异表达分析（differential expression analysis, DEA）	第 07 章
clusterProfiler	GO/KEGG 富集分析、GSEA	第 09 章
org.Hs.eg.db	人类基因 ID 注释与转换	第 07/09 章
survival	KM 生存曲线、Cox 回归	第 11 章
survminer	生存分析绘图（ggsurvplot）	第 11 章
glmnet	LASSO 回归，构建预后风险模型	第 11 章
pheatmap	表达热图	第 08 章
ggplot2	通用绘图（火山图等）	第 08 章
timeROC / survivalROC	时间依赖 ROC 曲线（评估模型）	第 11/12 章

使用某个包前要先“加载”：`library(包名)`。如果 `library` 报错“没有这个包”，多半是没装成功，重新执行对应的安装命令即可。

补充一个小知识：`library(包名)` 用于加载包，使其中函数可以直接调用；而 `BiocManager::install(...)` 这种“包名:: 函数”的写法表示“调用某包里的某函数”，不先加载也能用。教程中两种写法都会出现。

5.4 工作目录与项目组织

工作目录（working directory）是 R 读写文件的默认文件夹。可以用 `setwd()` 手动设置：

```
# Windows 路径示例：注意用正斜杠 / 或双反斜杠 \\
setwd("D:/projects/hcc_pyroptosis")
getwd()   # 查看当前工作目录
```

> 新手最常见的坑：文件明明在桌面上，R 却找不到——多半是工作目录不对。**强烈建议用 RStudio Project 代替手动 setwd**，一劳永逸。

建立 RStudio Project：File → New Project → New Directory → New Project，填写项目名（如 `hcc_pyroptosis`），选择存放位置（如 `D:/projects/`），创建后 RStudio 会自动把工作目录设为项目根目录，下次打开项目时也会自动恢复。

相对路径：使用 RStudio Project 后，脚本里的 `data/raw/...` 这类路径就是“相对于项目根目录”的路径，换电脑也能用；而 `C:/Users/xxx/Desktop/...` 这类绝对路径换电脑就要改。本教程一律采用相对路径。

项目内文件夹组织（本教程统一约定，第 06 章会详细说明 data 目录）：

```
hcc_pyroptosis/
├── data/        # 原始数据（下载后只读，不手动改动）
├── scripts/     # R 脚本（01_download.R、02_dea.R ……）
├── results/     # 分析结果（表格、模型输出）
├── figures/     # 图片
└── README.md    # 项目说明与数据登记（见第 06 章）
```

5.5 环境一致性：sessionInfo() 与 renv

生信分析讲究“可复现”（reproducible）：半年后别人（或你自己）重跑脚本，结果应当一致。但 R 包会不断更新，版本不同结果可能不同。这里给出两个由浅入深的办法：

1. **记录 sessionInfo()**（必须做到）：`sessionInfo()` 会把 R 版本、所有已加载包及其版本完整打印出来。把它保存到 results 目录，作为本次分析的环境快照：

```
writeLines(capture.output(sessionInfo()), "results/sessionInfo.txt")
```

2. **用 renv 锁定环境**（进阶，了解即可）：`renv::init()` 初始化，`renv::snapshot()` 把当前所有包的精确版本记录到项目里，之后在别的电脑上 `renv::restore()` 可以一键恢复相同版本。适合需要长期维护的项目。

对新手来说，做到第 1 条（记录 sessionInfo）就已经比大多数教程的要求更严谨了。

什么时候用 renv？项目需要长期维护、多人协作，或投稿前想严格复现时再用 renv；日常练习做到 sessionInfo 记录就足够了。

本章小结

- 装 R 4.x + RStudio，Windows 一路默认安装即可；
- 装包两条命令：`install.packages()`（CRAN）与 `BiocManager::install()`（Bioconductor）；
- 本教程核心包：GEOquery、limma、clusterProfiler、org.Hs.eg.db、survival、survminer、glmnet、pheatmap、ggplot2、timeROC/survivalROC；
- 用 RStudio Project 组织项目，data / scripts / results / figures 分文件夹存放；
- 分析完成后用 `sessionInfo()` 记录环境，进阶可用 renv 锁定包版本。

思考题

1. 为什么本教程选择 R 而不是 Python 作为主分析语言？Python 在什么场景下更合适？
2. `install.packages()` 和 `BiocManager::install()` 有什么区别？为什么 Bioconductor 的包一般不能用第一种命令安装？
3. 在 Console 中运行 `setwd("C:/Users/你的用户名/Desktop")` 后，`getwd()` 会返回什么？为什么教程推荐用 RStudio Project 而不是 setwd？
4. 动手实践：在你自己的电脑上完成 R 与 RStudio 的安装，并成功运行 `library(ggplot2)`。

第 06 章数据获取

本章属于 “Track B：数据基础”。环境搭好（第 05 章）之后，我们正式开始下载本课题的两套核心数据：**TCGA-LIHC**（发现队列）与 **GSE14520**（验证队列）。

本章目标

学完本章，你将能够：

1. 从 UCSC Xena 网页下载 TCGA-LIHC 的表达矩阵与临床表（并了解 GDC 的下载思路）；
2. 用 R 的 GEOquery 包下载 GSE14520，提取表达矩阵与临床信息；
3. 理解 GSM → GSE 的层级关系，以及探针 ID 如何映射到基因 Symbol；
4. 按规范组织 `data/` 目录（原始数据只读、脚本与结果分开）；
5. 对下载的数据做最简单的质控初检（维度、缺失值、分组列）。

> 重要提示：本章所有下载都需要**联网**，且 TCGA/GEO 文件较大，下载可能较慢；教学演示时也可以只取部分样本（见 6.2 节提示）。

6.1 下载 TCGA-LIHC：推荐用 UCSC Xena

开始前请先做两件事：一是在 `data/raw/` 下建好 `TCGA-LIHC/` 与 `GSE14520/` 两个文件夹（见 6.3 节）；二是确认网络稳定——TCGA 表达矩阵文件较大，中途断网可能导致下载不完整，建议在网速较好的时段下载。

TCGA 数据可以从官方 GDC 下载，但对新手来说，**UCSC Xena**（https://xena.ucsc.edu ）更友好：它把 TCGA 整理成 “基因×样本” 的表达矩阵和 “样本×字段” 的临床表，点几下就能下载。

UCSC Xena 网页下载步骤（以当前界面为例；网站界面会更新，以官网为准）：

1. 打开 https://xena.ucsc.edu ，点击 **Launch Xena** 进入数据浏览界面（Xena Browser）；
2. 在左侧数据列表中找到 **TCGA Liver Cancer (LIHC)**，点击展开；
3. 勾选表达数据：选择 **HTSeq - FPKM**（FPKM 是一种按基因长度和测序深度标准化的表达量，适合基因间比较），这是本教程后续分析使用的表达矩阵；如果需要做差异表达用的原始计数，可另下载 **HTSeq - Counts**；
4. 勾选临床数据：选择 **phenotype**（临床表），里面包含生存状态、生存时间、TNM 分期等字段；

5. 使用页面的 **Download** 功能分别下载表达矩阵与临床表。

下载到的是什么文件？

- **表达矩阵**：一个压缩的 TSV 文本文件（.gz），第一列是基因 ID（如 ENSG000001…… 或基因 Symbol），之后每一列是一个样本。样本名形如 `TCGA-BC-A10Q-01A`，其中 `-01` 一般代表肿瘤组织、`-11` 代表癌旁正常组织（TCGA 样本类型编码规则，以官方说明为准）；
- **临床表**：样本×字段的表格，每个样本一行，包含生存和临床信息。

FPKM 还是 Counts？本教程选用 HTSeq-FPKM：它按基因长度与测序深度做了标准化，适合比较基因之间的表达水平，也是 limma 差异表达分析（第 07 章）的常见输入（通常先做 log2 转换）。HTSeq-Counts 是原始计数，更适合 DESeq2 / edgeR 等专门工具，本教程暂不涉及。

GDC 简述（了解即可）：https://portal.gdc.cancer.gov → Repository →左侧筛选（Cancer Type 选 Liver Hepatocellular Carcinoma；Data Category 选 Transcriptome Profiling；Experimental Strategy 选 RNA-Seq）→把符合条件的文件加入购物车→下载表达矩阵与临床文件。GDC 更权威，但筛选项多、文件格式多样，新手阶段先用 Xena。

下载后检查：表达矩阵文件通常有几十 MB（解压后更大），临床表只有几百 KB。下载完成后建议先确认文件能正常打开：用记事本或 Excel 打开前几行看看格式（注意：大文件用 Excel 打开会很慢，可用 R 的 `read.table` 或 `data.table::fread` 读取）。文件命名建议加上日期，例如 `LIHC_HTSeqFPKM_2025-01-15.txt.gz`。如果只是想练习完整流程，也可以先用较小的临床表文件熟悉格式。

6.2 用 R 的 GEOquery 下载 GSE14520

GSE14520 存放在 GEO 中。GEO 提供了官方 R 包 **GEOquery**，一条命令即可下载整个数据集。

代码 6-1：下载并读取 GSE14520

```r
# 第 05 章已安装：BiocManager::install("GEOquery")
library(GEOquery)

# 下载 GSE14520 的 Series Matrix 文件（需联网，视网速可能需要数分钟）
gse <- getGEO("GSE14520", GSEMatrix = TRUE)

# gse 是一个列表：一个 GPL 平台对应一个元素
length(gse)   # 一般等于该 GSE 使用的平台数量
gse[[1]]      # 第一个平台的 ExpressionSet 对象

# 提取表达矩阵：行是探针 ID，列是样本（GSM 编号）
expr <- exprs(gse[[1]])
dim(expr)     # 查看维度，例如 2 万余个探针 × 数百个样本
```

```
# 提取样本信息（临床 / 分组 / 随访）
pdata <- pData(gse[[1]])
colnames(pdata)  # 先看有哪些列，再决定怎么用
head(pdata)      # 查看前几行
```

ExpressionSet 是什么？ GEOquery 把每个平台的完整数据封装成一个 ExpressionSet 对象，它像一个“文件夹”，里面装了三样东西：表达矩阵（用 `exprs()` 取出）、样本信息（用 `pData()` 取出）、探针注释（用 `fData()` 取出）。理解这个结构后，GEOquery 的日常用法就只剩这三个函数了。

解释几个关键点：

- `getGEO("GSE14520", GSEMatrix = TRUE)` 下载的是 GEO 官方整理的 Series Matrix 文件；返回的 `gse` 是**列表**，每个元素对应一个平台（GPL）的 ExpressionSet 对象。同一个 GSE 如果用了不同芯片平台，就会有多个元素，所以我们用 `gse[[1]]` 取第一个；
- `exprs()` 取出表达矩阵（ExpressionSet 的“表达”槽），`pData()` 取出样本信息（phenotype data 槽）——这是 GEOquery 最常用的两个函数；
- **GSM → GSE 层级回顾**（第 03 章）：GSE14520 是一个“系列”，包含成百上千个 GSM 样本；`expr` 的列名就是 GSM 编号，`pdata` 的每一行对应一个 GSM；
- 样本分组信息通常在 `pdata` 的 `characteristics_ch1` 等列中，例如 `tissue: tumor` / `tissue: normal`；GSE14520 还带有生存随访字段（这是它能做生存验证的前提）。**不同数据集的列名不同，务必先用 `colnames(pdata)` 查看实际列名**，再写代码提取。

  小提示：getGEO 会把下载的 Series Matrix 文件缓存在当前工作目录，重复运行时直接读取缓存，速度更快。如果下载中断，删除缓存文件后重新运行即可。

教学演示提示：GSE14520 样本数较多，教学时可以先只取部分样本跑通流程：

```
# 示例：只保留前 20 个样本（教学演示用；正式分析请用全部样本）
expr_sub <- expr[, 1:20]
```

关于临床列：GSE14520 的 `pData` 中，样本分组列（tumor / normal）和生存列（生存时间、生存状态）的列名在不同版本中可能不同，常见命名有 `characteristics_ch1`、survival time、survival status 等。第 11 章做生存分析前，我们会先把这些列整理成统一格式：时间列必须是数值（如 12.5 个月），状态列必须是 0/1（0 = 删失 censored，1 = 事件发生）。如果状态列是文字（如 alive / dead），需要先转换为 0/1。

探针 ID →基因 Symbol 的映射（平台注释）

GSE14520 使用的芯片平台是 **GPL3921**（Affymetrix 人类全基因组芯片，具体型号以 GEO 页面为准），表达矩阵的行名是探针 ID（形如 `1552777_at`），不是基因名。要得到基因 Symbol，需要平台注释。三种常用做法：

方法 A（最简单）：Series Matrix 文件本身常带注释列，直接查看：

```
fData(gse[[1]])                # 探针注释表
colnames(fData(gse[[1]]))      # 看看有哪些注释列（如 Gene Symbol）
```

方法 B（从 GEO 下载 GPL 注释表）：

```
gpl <- getGEO("GPL3921")       # 下载平台注释
colnames(Table(gpl))           # 查看注释表的列，通常有 ID、Gene Symbol 等
head(Table(gpl))               # 查看前几行示例
```

方法 C（Bioconductor 注释包）：GPL3921 这类 Affymetrix 平台一般有对应的注释包（如 hthgu133a.db，具体以 Bioconductor 当前文档为准），可用 `mapIds()` 做探针→ Symbol 的映射。第 07 章预处理时会给出完整映射代码，这里先理解原理：**探针 ID 必须靠平台注释变成基因 Symbol，分析才有生物学意义**。

无论用哪种方法，最终目的都是得到一张 "探针 ID →基因 Symbol" 的对照表，然后把它合并到表达矩阵上，把同一基因对应的多条探针合并（通常取平均值或最大值，第 07 章会给出处理重复基因的完整代码）。这一步是芯片数据分析的关键转折点：**从此以后，你的分析对象从 "探针" 变成了 "基因"**，后面所有分析（差异表达、富集、生存）都在基因层面进行。

6.3 数据文件怎么组织

下载到的原始文件不要乱放，按第 05 章的约定组织（这是本教程所有章节的统一约定）：

```
hcc_pyroptosis/
├── data/
│   ├── raw/                   # 原始下载文件（只读，永不手动修改）
│   │   ├── TCGA-LIHC/         # Xena 下载的表达矩阵与临床表
│   │   └── GSE14520/          # GEO 下载的系列矩阵文件
│   └── processed/             # 处理后的数据（第 07 章输出）
├── scripts/                   # 01_download.R、02_dea.R ……
├── results/                   # 分析结果表
└── figures/                   # 图片
```

把 Xena 下载的文件放进 `data/raw/TCGA-LIHC/`；把 GEOquery 下载的对象保存到本地，避免以后反复联网下载。`processed/` 目录则留给第 07 章的预处理产物（如探针映射为基因 Symbol、去重、log2 转换后的表达矩阵）——**原始文件永远只读，一切处理都在脚本里可重现**：

```
# 保存 GSE14520 原始对象到本地（约几十 MB）
save(gse, file = "data/raw/GSE14520/gse14520_raw.RData")
```

为什么原始数据必须只读？分析的可信度来自“可追溯”：任何处理都应当通过脚本完成，而不是手动修改原始文件。一旦你在 Excel 里手动改过原始矩阵，别人（包括半年后的你自己）就无法知道你改了什么，结果也就无法复现。记住一句话：**原始数据只读，脚本重现一切**。

6.4 数据质控初检

下载完成后先别急着分析，做三件最基础的检查（每件都是 R 一行命令）：

```
# 1. 表达矩阵维度：行数（基因/探针）× 列数（样本）
dim(expr)

# 2. 缺失值数量：缺失太多说明数据质量差，需要处理
sum(is.na(expr))

# 3. 样本分组：看看有哪些组别、每组多少样本
#    注意：先 colnames(pdata) 确认分组列名，再替换下面的列名
table(pdata$characteristics_ch1)
```

这三个检查回答三个问题：**数据有多大？缺失多不多？样本怎么分组？**如果缺失值很多（例如数以万计），第 07 章会教你处理（过滤 / 补缺 / 剔除），这里先记录观察即可。把检查结果写进 `results/` 下的记录文件，作为分析日志的一部分：

```
# 记录质控结果（示例），供日后核对
writeLines(c(
  paste(" 探针数:", dim(expr)[1], " 样本数:", dim(expr)[2]),
  paste(" 缺失值数:", sum(is.na(expr)))
), "results/00_QC_note.txt")
```

两个容易踩的坑：一是表达矩阵的列名（GSM 编号）与 `pData` 的行名必须一一对应，分组前最好确认顺序一致（`all(colnames(expr) == rownames(pdata))` 返回 TRUE 才放心）；二是分组列里可能有大小写或写法差异（如 tumor 与 Tumor），用 `unique()` 查看所有取值后再写分组逻辑。

另外，质控的阈值没有统一标准，关键是如实记录并解释你的选择——这本身就是严谨的做法。

6.5 数据登记意识

做科研要能“说清楚数据从哪来”。建议在 `data/README.md`（或 data 目录下一个文本文件）里做登记：

```
# 数据登记表
- TCGA-LIHC：来源 UCSC Xena；表达：HTSeq-FPKM；临床：phenotype；
```

```
  accession/版本: TCGA-LIHC (以 Xena 页面标注为准); 下载日期: 2025-XX-XX
- GSE14520: 来源 GEO; 平台: GPL3921;
  accession: GSE14520; 下载日期: 2025-XX-XX
```

只要记住三要素——**accession 编号、下载日期、版本/平台**——以后写论文方法学、或被审稿人追问数据来源时，都能答得上来。（一句话登记就够了，但必须记。）

为什么登记很重要？公共数据库的版本会更新（例如 GEO 偶尔补充样本、Xena 更新注释版本），如果不记录下载日期和版本，以后重跑分析可能得到与论文不一致的结果。另外，审稿人或读者要求 “提供数据来源” 时，登记表就是你的证据。

本章小结

- TCGA-LIHC 用 UCSC Xena 网页下载（HTSeq-FPKM 表达 + phenotype 临床）；GDC 是更权威但更繁琐的备选；
- GSE14520 用 `getGEO("GSE14520", GSEMatrix = TRUE)` 一行下载，`exprs()` 取表达矩阵、`pData()` 取临床信息；探针 ID 需用平台注释（GPL3921）映射为基因 Symbol；
- 原始数据放 `data/raw/`（只读），脚本、结果、图分开存放；
- 下载后先做三维度检查：`dim()`、`sum(is.na())`、`table(分组列)`；
- 登记 accession、下载日期、版本，是数据管理的底线。

思考题

1. 为什么表达矩阵的列名是 GSM 编号而不是患者编号？患者编号信息一般藏在 `pData` 的哪一列？
2. 如果 `getGEO("GSE14520")` 返回的 `gse` 有多个元素，说明什么？用 `gse[[2]]` 取出来看看与第一个有什么不同。
3. “探针 ID” 和 “基因 Symbol” 有什么区别？为什么分析前必须做映射？
4. 在 `pData(gse[[1]])` 中找一找 GSE14520 的生存相关字段（如生存时间、生存状态），想一想后续 KM 生存分析需要哪两列。

第 07 章差异表达分析

本章目标

学完本章，你将能够：

1. 用一句话说清差异表达分析（differential expression analysis, DEA）回答什么问题，以及它在整个课题流程中的位置；
2. 整理“表达矩阵 + 分组信息”两类输入，并用 `model.matrix()` 构建设计矩阵；
3. 用 limma 包跑通 log2 转换→ lmFit → eBayes → topTable 的完整流程，并读懂结果表的每一列；
4. 用 `|log2FC| > 1 且 adj.P < 0.05` 的标准筛选差异基因，能解释为什么不能只看 p 值；
5. 把差异基因与焦亡基因集取交集，导出 CSV 结果表，为后续富集分析与预后分析提供输入。

正文

1. 差异表达分析是什么

在肝细胞癌（hepatocellular carcinoma, HCC）课题中，我们最想知道的一件事是：与癌旁正常组织相比，肿瘤组织里“谁变了”。差异表达分析（differential expression analysis, DEA）就是系统地回答这个问题的统计方法：对每一个基因，比较它在两组样本（如 Normal vs Tumor）中的平均表达水平，检验差异是否显著，并估计变化幅度。

一句话版本：**DEA 给每个基因算两个数——变化幅度（fold change, FC）和显著性（p 值/校正后 p 值），然后按统一标准筛选出“变化大且可信”的基因。**

打个比方，表达谱像一张“基因点名册”，DEA 就是逐个点名、看谁在两组之间“冒头”或“缩头”。差异基因列表是整个课题的“原材料”，后面的功能分析（第 09 章）、网络分析（第 10 章）和生存分析（第 11 章）都建立在它之上。

2. 数据准备：表达矩阵与分组信息

DEA 需要两类输入：

- **表达矩阵**：行是基因，列是样本，值是表达量；
- **分组信息**：每个样本属于哪个组（Normal 还是 Tumor）。

本教程课题使用 TCGA-LIHC（肝细胞癌）的 RNA-seq 表达谱，约 371 个肿瘤样本 + 50 个癌旁正常样本（数据下载与整理见第 06 章）。注意：TCGA 的 RNA-seq 原始数据是整数 counts，公开渠道常以 TPM/FPKM 等归一化形式提供。**本章教学演示用 log2 转换后的表达量跑通流程；若手头是 counts，建议用 limma 的 voom() 或 limma-trend 处理**——新手常踩的坑。

在 R 中，分组信息是一个和表达矩阵列顺序一一对应的因子向量：

```
group <- factor(c(rep("Normal", 50), rep("Tumor", 371)))
```

limma 需要把分组信息转成**设计矩阵**（design matrix），它告诉模型“每个样本属于哪个组、我们要比较哪两组”：

```
design <- model.matrix(~0 + group)      # 不设截距，每组一个系数
colnames(design) <- levels(group)       # 列名改为 Normal / Tumor
```

新手先记住：`model.matrix()` 就是把分组因子翻译成一张 0/1 矩阵——行是样本，列是组，某样本属于某组则对应位置为 1。真正要比较的两组之差，再用 `makeContrasts()` 显式声明：

```
contrast.matrix <- makeContrasts(Tumor - Normal, levels = design)
```

3. limma 流程：三件套与关键参数

limma（linear models for microarray data）是 Bioconductor 上最常用的差异分析包。它的核心思想是：对每个基因拟合一个线性模型，再用**经验贝叶斯（empirical Bayes）**方法借用全基因组信息来稳定单个基因的方差估计。标准流程是“三件套”：

```
library(limma)
expr_log2 <- log2(expr + 1)                  # TPM 表达量加 1 再取 log2，避免 log(0)

fit  <- lmFit(expr_log2, design)             # 第一步：对每个基因拟合线性模型
fit2 <- contrasts.fit(fit, contrast.matrix)  # 第二步：取出我们关心的对比 Tumor - Normal
fit2 <- eBayes(fit2)                         # 第三步：经验贝叶斯方差收缩，得到统计量
res  <- topTable(fit2, coef = 1, number = Inf, sort.by = "none")
```

逐步解释：

- `lmFit()`：逐基因做线性回归，得到每个基因在每个组系数下的均值估计与标准误；
- `contrasts.fit()`：把系数组合成我们关心的对比（Tumor 减去 Normal），得到每个基因的 logFC 与标准误；
- `eBayes()`：把全基因组的信息“借”给每个基因，让方差估计更稳定——这是 limma 相对普通 t 检验的核心优势；

- `topTable()`：汇总成结果表。**`number = Inf` 表示输出全部基因**（默认只输出前 10 个，新手常在这里“丢基因”）；`sort.by = "none"` 保持原始基因顺序，方便和表达矩阵对应。

`topTable` 输出的结果表包含以下关键列：

列名	含义
logFC	log2 变化倍数：肿瘤相对癌旁的表达倍数（log2 尺度）
AveExpr	该基因在所有样本中的平均表达量
t	该对比的 t 统计量
P.Value	未校正的 p 值
adj.P.Val	BH 校正后的 p 值（即 FDR 估计值）
B	该基因“确实有差异”的对数后验优势（log-odds）

其中 `logFC` 和 `adj.P.Val` 是我们筛选差异基因最关心的两列。

4. 筛选标准：为什么不能只用 p 值

本教程采用的筛选标准（也是生信论文最常用的标准）：

$$|\mathbf{log2FC}| > 1 \text{ 且 } \mathbf{adj.P} < 0.05$$

- `|log2FC| > 1`：变化倍数超过 2 倍（FC >2 或 FC <0.5）。这保证差异**有生物学意义**——一个基因只变了 10% 可能不值得关注；
- `adj.P < 0.05`：保证差异**统计上可信**。

为什么不能只看 p 值？因为一次要比较成千上万个基因：即使两组完全没差异，按 $p<0.05$ 的标准，也会有约 5% 的基因“纯靠运气”显著。假设检测 20,000 个基因，就会随机出现约 1,000 个假阳性。这就是**多重检验问题**（multiple testing）。

解决方法是对 p 值做校正。本教程用 **Benjamini–Hochberg（BH）校正**：把 p 值从小到大排序、逐级调整，得到 adj.P.Val，它控制的是**假发现率**（false discovery rate, FDR）。直观理解：adj.P <0.05 意味着“被判定为显著的这批基因里，预期约 5% 是假阳性”。R 里的等价写法是 `p.adjust(p, method = "BH")`，topTable 默认已做这一步。

筛选代码：

```
deg_idx <- abs(res$logFC) > 1 & res$adj.P.Val < 0.05
deg     <- res[deg_idx, ]
table(res$adj.P.Val < 0.05, abs(res$logFC) > 1)   # 交叉计数，看两个条件各自筛掉多少
```

5. 结果解读：上调/下调与焦亡基因集交集

拿到差异基因后，先看整体格局：

```
up   <- deg[deg$logFC > 0, ]   # 上调：肿瘤中表达升高
down <- deg[deg$logFC < 0, ]   # 下调：肿瘤中表达降低
nrow(up); nrow(down)           # 分别计数
```

注意方向约定：`logFC > 0` 表示在 Tumor − Normal 中升高，即**肿瘤中上调**，反之则下调。

下一步回到课题主题——细胞焦亡。把差异基因与 “焦亡基因集” 取交集，得到 “在 HCC 中异常表达的焦亡基因”，这正是后续生存分析与免疫分析的核心对象：

```
pyroptosis_genes <- c("GSDMD", "GSDME", "CASP1", "IL1B", "IL18",
                      "NLRP3", "AIM2", "PYCARD")   # 示例列表
intersect(rownames(deg), pyroptosis_genes)
```

重要说明：焦亡基因集的 “权威版本” 需要从文献（焦亡机制综述或相关生信论文的补充材料）中获取，不同论文收录的基因从十几个到几十个不等。本教程的列表仅为演示用示例；正式课题中请注明基因集来源并核对基因名拼写（如是否收录 GSDMB、GSDMC、CASP4/5 等）。

6. 输出结果：保存 CSV

把结果保存下来，供第 08、09 章继续使用：

```
dir.create("results", showWarnings = FALSE)
write.csv(res, "results/DEA_TCGA_LIHC_all.csv")    # 全部基因的结果表
write.csv(deg, "results/DEA_TCGA_LIHC_DEGs.csv")   # 筛选后的差异基因
write.csv(intersect(rownames(deg), pyroptosis_genes),
          "results/pyroptosis_DEGs.csv")           # 焦亡相关的差异基因
```

保存时务必保留 “未筛选” 的全基因结果表：第 09 章做 GSEA 时需要对所有基因按 logFC 排序，只用差异基因子集会丢失大量信息。

完整可运行示例：用模拟数据跑通上述全部流程（`set.seed` 保证可复现；真实分析时把 `expr`、`group` 换成第 06 章得到的真实数据即可）：

```
set.seed(2024)
n_gene <- 2000; n_normal <- 30; n_tumor <- 30
expr <- matrix(rnbinom(n_gene * (n_normal + n_tumor), size = 10, mu = 100),
               nrow = n_gene)
rownames(expr) <- paste0("Gene", 1:n_gene)
expr[1:200, (n_normal + 1):(n_normal + n_tumor)] <-   # 模拟 200 个上调基因（约 4 倍）
```

```
  expr[1:200, (n_normal + 1):(n_normal + n_tumor)] * 4
group <- factor(c(rep("Normal", n_normal), rep("Tumor", n_tumor)))
# ……随后依次执行第 3 节的 design / limma / 筛选代码即可看到结果……
```

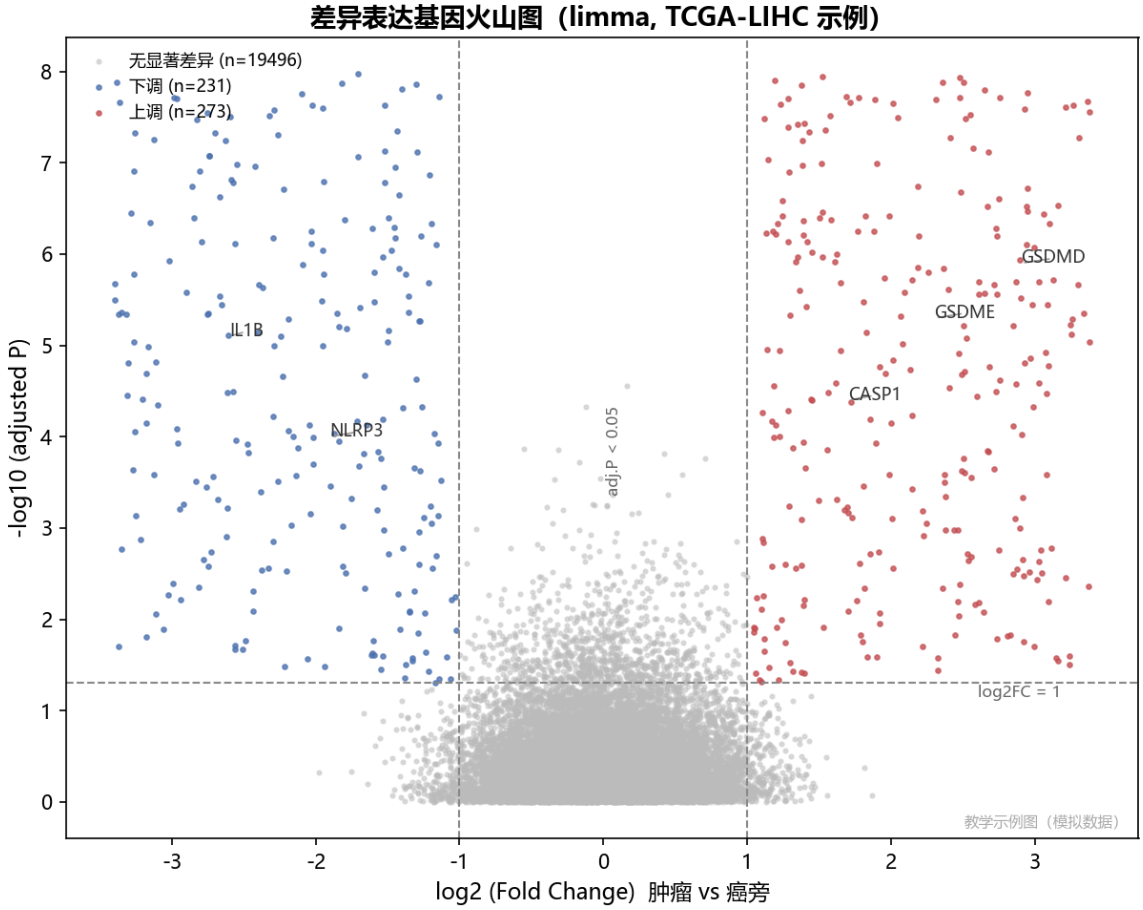


图 7: 图 06 差异表达火山图（示例，模拟数据）

上图为火山图示例（第 08 章会教怎么画、怎么读）：每个点是一个基因，横轴 log2FC，纵轴 −log10(adj.P.Val)，越靠左右两侧高处，越是“变化大且显著”的差异基因。下图 PCA 图则从样本层面展示 Normal 与 Tumor 的整体分离，说明分组确实能解释表达差异（两张图均为模拟数据示意，读者用真实数据跑脚本后得到自己的图）。

本章小结

- DEA 逐基因比较两组表达，输出变化幅度（logFC）与显著性（adj.P.Val），回答“肿瘤 vs 癌旁哪些基因变了”。
- 输入是表达矩阵 + 分组信息；分组用 `model.matrix()` 转成设计矩阵，用 `makeContrasts()` 声明比较。
- limma 三件套：`lmFit` → `contrasts.fit` → `eBayes`，最后 `topTable` 出表；`number = Inf` 别

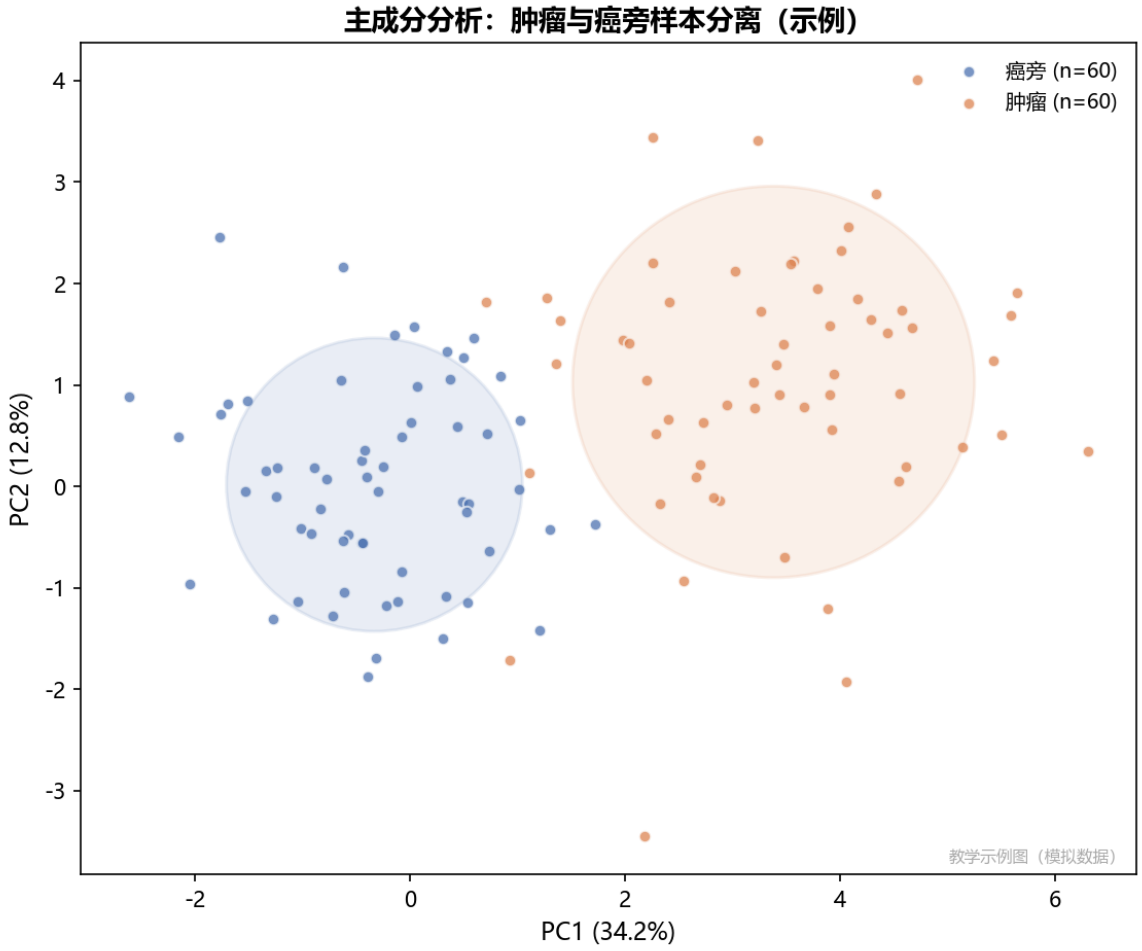


图 8: 图 08 PCA 图（示例，模拟数据）

忘。

- 筛选标准 `|log2FC| > 1 且 adj.P < 0.05`：前者保证幅度，后者（BH 校正）控制多重检验下的假阳性，不能只看 p 值。
- 与焦亡基因集取交集得到课题核心基因；基因集要注明来源。
- 全基因结果表也要保存，第 09 章 GSEA 要用。

思考题

1. `topTable` 为什么默认只返回 10 行？`number = Inf` 和 `sort.by = "none"` 各解决什么问题？
2. 如果两组样本量都只有 3 个，limma 的经验贝叶斯校正还可靠吗？为什么？
3. 假设一个基因 p 值极小（1e-10）但 log2FC = 0.2，你会把它算作差异基因吗？为什么？
4. 焦亡基因集来自不同论文时结果可能不同，你认为论文里应该如何报告基因集的选取与版本？

第 08 章可视化与图表

本章目标

学完本章，你将能够：

1. 理解 ggplot2 的图层语法：数据、映射（aes）、几何对象（geom）、主题（theme）；
2. 独立画出并正确解读火山图（volcano plot），并会做发表级美化；
3. 用 pheatmap 画带分组注释条和聚类树的热图（heatmap）；
4. 用 prcomp + ggplot2 画 PCA 图并加置信椭圆，判断样本整体分离；
5. 掌握论文图表的规范：字号、分辨率（dpi = 300）、图注（Figure legend）写法。

正文

1. ggplot2 绘图哲学：一图一图层

R 基础绘图（`plot()`）像“在白纸上直接涂色”，而 ggplot2 像“搭积木”：先告诉它“数据是什么、把数据的哪些列映射到图形的哪些元素”，再一层一层往上叠加几何对象和主题。它的核心是三件事：

- **数据**：一个整洁的数据框（data.frame）；
- **映射 `aes()`**：把数据列映射到图形属性——x 坐标、y 坐标、颜色、形状、大小；
- **几何对象 `geom_*()`**：决定画什么形状——点 `geom_point()`、线 `geom_line()`、柱 `geom_col()`、箱线 `geom_boxplot()` 等。

新手三步画图法：

```
# 第一步：准备数据（通常整理成长格式：一列一个变量、一行一个观测）
# 第二步：ggplot(data, aes(x=..., y=...)) 建立画布与映射
# 第三步：+ geom_xxx() 加图层，+ theme_xxx() 调主题，+ ggsave() 存图
ggplot(data, aes(x = xvar, y = yvar, color = group)) +
  geom_point() +
  theme_bw()
```

两个常见坑：一是 `aes()` 里写的是数据框的**列名**而不是外面的向量；二是图层叠加用 **`+` 号**而不是管道符`%>%`。记住这两点，ggplot2 就成功了一半。

2. 火山图：一张图看懂“谁在变”

火山图是差异表达分析的标准配图：每个点是一个基因，横轴 log2FC，纵轴 $-\log10(\text{adj.P.Val})$。纵轴取负对数后，p 值越小点越靠上，所以“显著且差异大”的基因分布在左右两侧高处，整张图呈火山喷发状——因此得名。

最直接的画法（输入是第 07 章的 `res` 结果表）：

```
library(ggplot2)
res$change <- ifelse(res$adj.P.Val < 0.05 & res$logFC > 1, "up",
              ifelse(res$adj.P.Val < 0.05 & res$logFC < -1, "down", "ns"))

ggplot(res, aes(x = logFC, y = -log10(adj.P.Val))) +
  geom_point(aes(color = change), size = 0.8) +
  scale_color_manual(values = c(up = "#d7191c", down = "#2c7bb6", ns = "grey80"),
                     labels = c(up = " 上调", down = " 下调", ns = " 不显著")) +
  geom_vline(xintercept = c(-1, 1), linetype = "dashed", linewidth = 0.4) +
  geom_hline(yintercept = -log10(0.05), linetype = "dashed", linewidth = 0.4) +
  labs(x = "log2(fold change)", y = "-log10(adjusted P)",
       title = "TCGA-LIHC 肿瘤 vs 癌旁差异表达火山图") +
  theme_bw() +
  theme(legend.position = "top")
ggsave("figures/fig06_volcano.png", width = 6, height = 5, dpi = 300)
```

怎么读（四象限含义）：

- 右上象限（$\text{log2FC} > 1$ 且 $\text{adj.P} < 0.05$）：显著**上调**基因；
- 左上象限（$\text{log2FC} < -1$ 且 $\text{adj.P} < 0.05$）：显著**下调**基因；
- 两条竖线之间或横线以下的点：不满足任一标准，**不显著**。

想要“论文级”效果，也可以用现成包 EnhancedVolcano（基于 ggplot2，自动标注显著基因名并避开标签重叠）：

```
# BiocManager::install("EnhancedVolcano")
library(EnhancedVolcano)
EnhancedVolcano(res,
                lab = rownames(res),      # 标注的基因名
                x = "logFC", y = "adj.P.Val",
                pCutoff = 0.05, FCcutoff = 1,
                title = "TCGA-LIHC DEGs",
                pointSize = 1.5, labSize = 3)
ggsave("figures/fig06_volcano.png", width = 7, height = 6, dpi = 300)
```

如果只想标注重点基因（比如我们的焦亡基因），用 ggrepel 包只给它们加标签：

```
res$label <- ifelse(rownames(res) %in% pyroptosis_genes, rownames(res), "")
ggplot(res, aes(x = logFC, y = -log10(adj.P.Val))) +
  geom_point(aes(color = change), size = 0.8) +
  ggrepel::geom_text_repel(aes(label = label), max.overlaps = 20) +
  theme_bw()
```

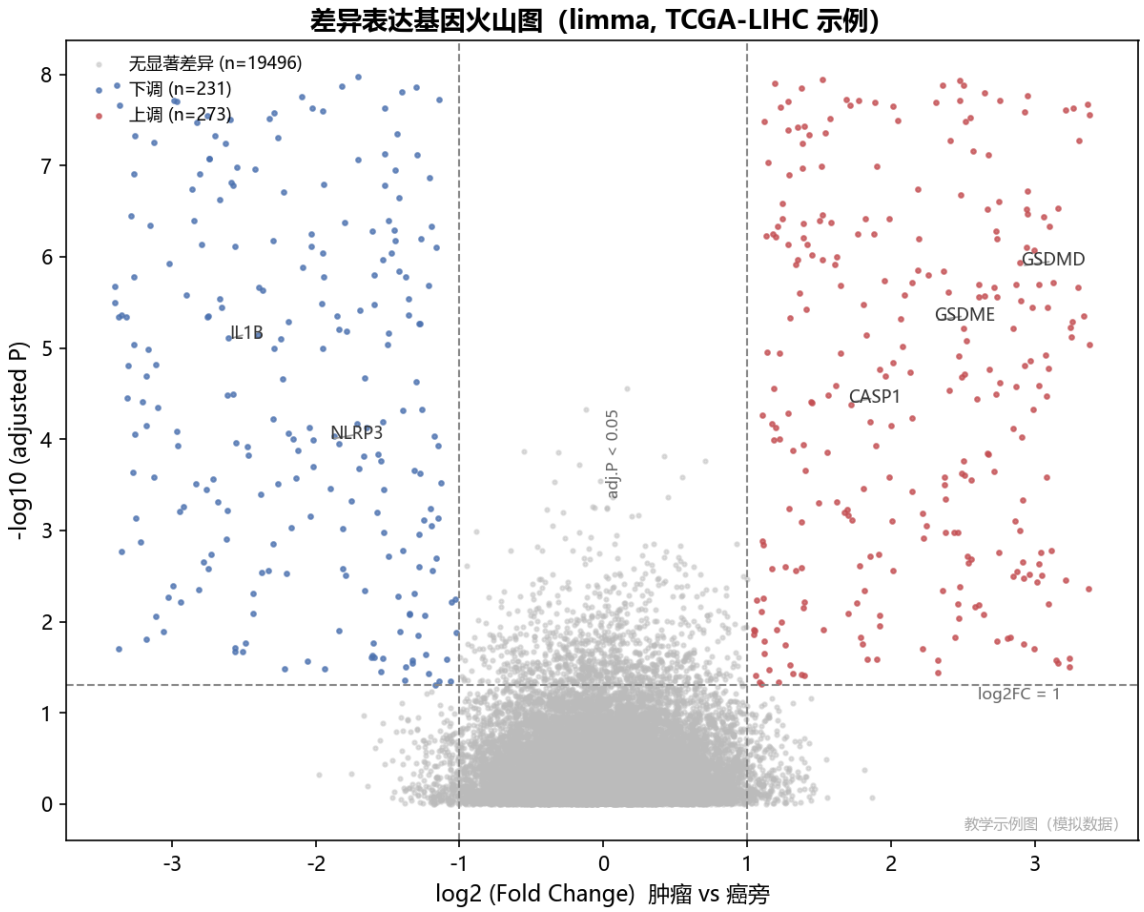


图 9: 图 06 差异表达火山图（示例，模拟数据）

3. 热图：表达模式的整体视图

热图把 “基因×样本” 的表达量画成颜色矩阵：行是基因、列是样本，颜色深浅代表表达高低。它适合回答 “这批基因在不同样本里有没有清晰的分组模式”。用 pheatmap 画差异基因（或焦亡差异基因）的表达热图：

```
# BiocManager::install("pheatmap")
library(pheatmap)

mat <- expr_log2[rownames(deg), ]              # 差异基因子集
mat <- t(scale(t(mat)))                        # 对每个基因做 z-score 标准化
```

```r
anno <- data.frame(group = group)              # 样本注释数据框
rownames(anno) <- colnames(expr_log2)          # 行名必须与样本名一致

png("figures/fig07_heatmap.png", width = 8, height = 6, units = "in", res = 300)
pheatmap(mat,
         scale = "row",                        # 按行（基因）标准化
         cluster_rows = TRUE, cluster_cols = TRUE,   # 行/列层次聚类
         annotation_col = anno,                # 顶部画分组注释条
         show_rownames = TRUE, show_colnames = FALSE,
         color = colorRampPalette(c("navy", "white", "firebrick3"))(100),
         main = " 差异基因表达热图（示例）")
dev.off()
```

需要理解的三点：

- **z-score 标准化**：对每个基因，把表达值减去该基因的均值、再除以标准差，使每个基因在行内以 0 为中心。否则高表达基因会永远偏红、低表达基因永远偏蓝，看不出“相对变化”。（`scale = "row"` 与 `t(scale(t(mat)))` 等价，二选一即可。）
- **行/列聚类**：pheatmap 默认对行和列做层次聚类，把表达模式相近的基因/样本放在相邻位置。如果 Normal 与 Tumor 在列聚类中各自成团，说明差异表达模式清晰——这是热图要传达的核心信息。
- **注释条**：`annotation_col` 是一个数据框，行名必须与表达矩阵的列名（样本名）完全一致，列名就是注释项的名称（如 group）。颜色映射可用 `annotation_colors` 参数自定义。

4. PCA 图：样本层面的“全家福”

PCA（主成分分析，principal component analysis）把高维表达数据压缩成少数几个“主成分”，让我们能在二维平面上一眼看出样本的整体关系。注意视角切换：差异分析看**基因**，PCA 看**样本**——它回答“肿瘤样本和癌旁样本在整体表达模式上是否分开”。

```r
pca <- prcomp(t(expr_log2), scale. = TRUE)      # 注意转置：行 = 样本、列 = 基因
pca_sum <- summary(pca)$importance               # 各主成分解释的方差比例
pca_df <- data.frame(PC1 = pca$x[, 1], PC2 = pca$x[, 2], group = group)

ggplot(pca_df, aes(x = PC1, y = PC2, color = group)) +
  geom_point(size = 2, alpha = 0.8) +
  stat_ellipse(level = 0.95) +                   # 95% 置信椭圆
  labs(x = paste0("PC1 (", round(pca_sum[2, 1] * 100, 1), "%)"),
       y = paste0("PC2 (", round(pca_sum[2, 2] * 100, 1), "%)")) +
  theme_bw()
ggsave("figures/fig08_pca.png", width = 6, height = 5, dpi = 300)
```

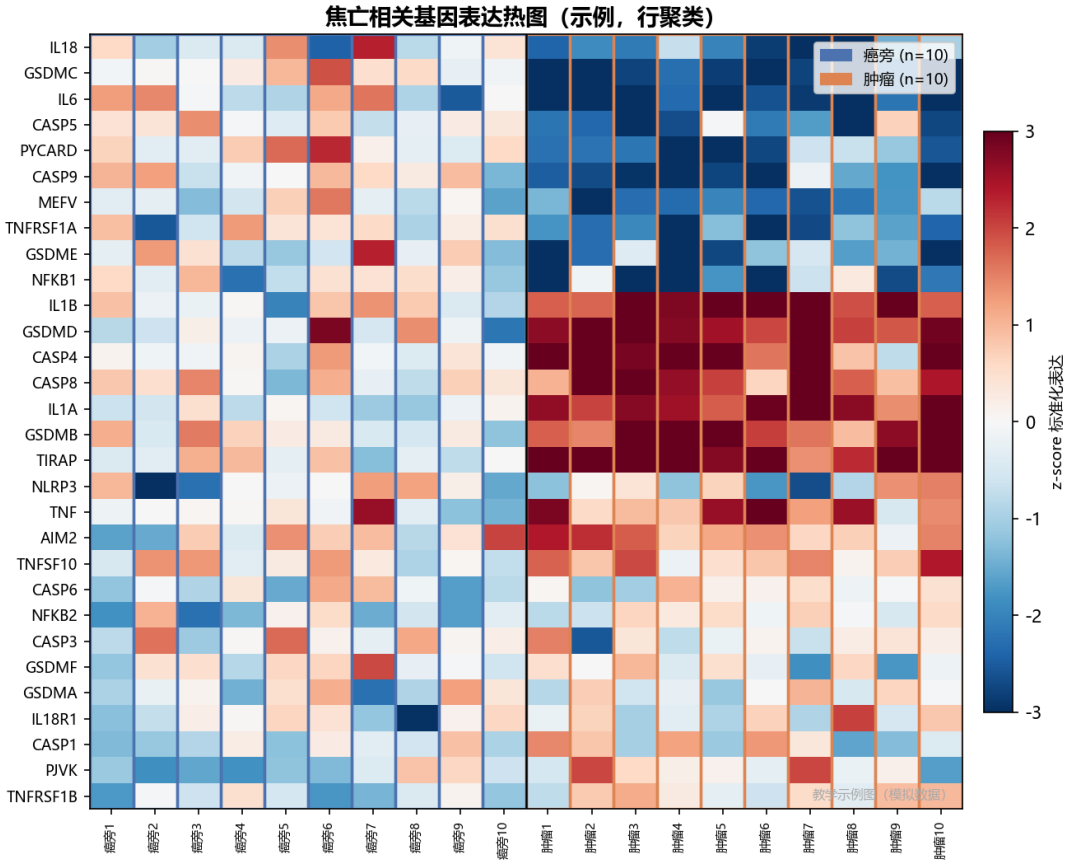


图 10: 图 07 差异基因表达热图（示例，模拟数据）

要点：

- **一定要转置**：`prcomp()` 要求行是观测（样本）、列是变量（基因），而表达矩阵是“基因×样本”，所以先 `t()`。
- `scale. = TRUE`：对每个基因标准化，防止高表达基因主导主成分。
- 轴标签写”PC1 (xx%)“：括号里是主成分解释的方差比例，说明这张图捕捉了数据多少信息。
- `stat_ellipse(level = 0.95)`：给每组画 95% 置信椭圆，帮助判断两组是否真的分开。
- PCA 的另一个用途是**初步检查批次效应**：如果同批次的样本聚在一起、而不是按分组聚在一起，就要警惕技术性差异混入了生物学差异。

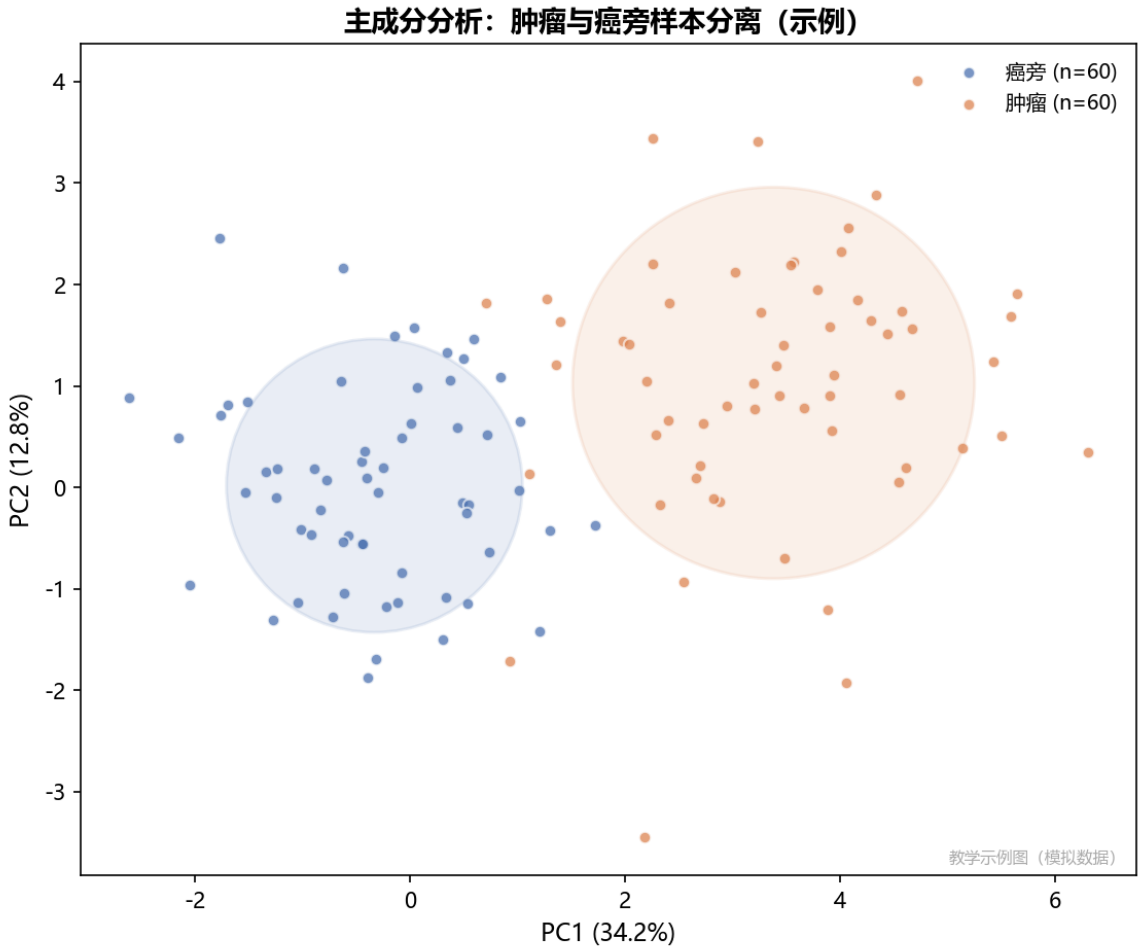


图 11: 图 08 PCA 图（示例，模拟数据）

5. 图表规范：从“能看”到“能发表”

论文图表和课堂作业图的差距往往在细节：

- **字号**：正文图建议基础字号 8–12 pt（`theme_bw(base_size = 12)`），保证缩印后仍可读；图内文字不小于 7 pt。
- **分辨率**：位图保存用 `ggsave(..., dpi = 300)`，这是多数期刊的底线要求；矢量图（PDF/

SVG）更好，任意缩放不模糊。

- **尺寸**：按期刊栏宽设置（单栏约 8.5 cm、双栏约 17.5 cm），先定尺寸再调字号，避免文字被压缩。
- **图注（Figure legend）**：写在图的下方，格式为 “图号 + 一句话标题 + 数据与方法说明 + 统计说明”。例如：“图 1. TCGA-LIHC 肿瘤与癌旁差异表达火山图。横轴为 log2 变化倍数，纵轴为 BH 校正后 p 值的负对数；红色为上调基因（$|\log2FC| > 1$ 且 $adj.P < 0.05$），蓝色为下调基因，灰色为不显著。” 图注要能脱离正文独立读懂。
- **一图一义**：每张图必须有独立含义，同一信息不要既画图又列表；图与图之间避免重复；配色考虑色盲友好（如 viridis 色系）。

本章小结

- ggplot2 的语法 = 数据 + `aes()` 映射 + `geom_*()` 图层 + `theme_*()` 主题，一层层用 `+` 叠加。
- 火山图：每点一基因，横轴 log2FC、纵轴 $-\log10(adj.P)$，四象限对应上调/下调/不显著；EnhancedVolcano 可自动标注，ggrepel 可只标重点基因。
- 热图：pheatmap 的 `scale = "row"` 做 z-score、`cluster_rows/cols` 聚类、`annotation_col` 加分组注释条。
- PCA：`prcomp(t(expr), scale. = TRUE)` + `stat_ellipse()`，从样本层面看分离与批次。
- 发表级规范：base_size 8–12、dpi = 300、图注独立可读、一图一义。

思考题

1. 热图里为什么通常先做 z-score 标准化？不做会出现什么误导？
2. PCA 图里两组完全混在一起，可能有哪些原因？下一步你会怎么排查？
3. 火山图的纵轴为什么用 $-\log10(adj.P.Val)$ 而不是直接用 adj.P.Val？
4. 一篇论文配了 8 张图，审稿人说 “图太多且重复”，你会如何取舍？

第 09 章功能富集分析

本章目标

学完本章，你将能够：

1. 理解富集分析的思想与统计学原理（超几何分布 / 费舍尔精确检验）；
2. 用 clusterProfiler 做 GO 富集，并分清 BP、CC、MF 三个层面；
3. 用 clusterProfiler 做 KEGG 通路富集，并了解 pathview 通路图；
4. 画出气泡图 / 柱状图并正确解读；
5. 理解 ORA 与 GSEA 的区别，能跑通 GSEA 并解读富集分数（ES）曲线。

正文

1. 富集分析是什么：从“一堆基因”到“一条通路”

第 07 章我们拿到了几百上千个差异基因，但“这些基因在功能上意味着什么？”——单个基因名看不出名堂。**功能富集分析**（functional enrichment analysis）回答的正是这个问题：给定一堆基因，看它们是否显著地集中在某些已知的功能类别或通路上。

统计思想用“袋子抽球”类比：想象一个袋子里有 N 个球，代表基因组里所有被测到的基因，其中 M 个是红球，代表“某条通路里的基因”。你随机抽了 n 个球（我们的差异基因），发现里面有 k 个红球：k 这么大是纯属巧合，还是说明这条通路真的和我们的基因列表有关？

这正是**超几何分布**（hypergeometric distribution）描述的“不放回抽球”概率问题；等价地，也可以用一张 2 × 2 列联表做**费舍尔精确检验**（Fisher's exact test）算出“纯属巧合”的 p 值。富集分析就是逐条通路（或 GO 条目）做这个检验，然后同样用 BH 校正控制多重检验。这一类方法统称 ORA（over-representation analysis，过度代表分析）。

两个容易踩的坑：

- **背景基因集（universe）**：默认是所有被测到的基因，而不是全基因组。背景选错了，结果会系统性偏差——用表达矩阵覆盖的基因当背景最稳妥；
- **输入列表的质量**：富集分析只在输入列表可靠时才有意义，所以务必使用第 07 章严格筛选出的差异基因。

2. GO 富集：三个层面

GO（Gene Ontology，基因本体）把基因功能分成三个层面：

- **BP**（biological process，生物学过程）：如炎症反应、细胞程序性死亡；
- **CC**（cellular component，细胞组分）：如线粒体、质膜；
- **MF**（molecular function，分子功能）：如半胱氨酸型内肽酶活性。

clusterProfiler 是 Bioconductor 上最主流的富集工具。标准流程如下：

```
# BiocManager::install("clusterProfiler")
# BiocManager::install("org.Hs.eg.db")
library(clusterProfiler)
library(org.Hs.eg.db)

deg_entrez <- bitr(rownames(deg), fromType = "SYMBOL",
                   toType = "ENTREZID", OrgDb = org.Hs.eg.db)$ENTREZID

ego <- enrichGO(gene = deg_entrez,
                OrgDb = org.Hs.eg.db,
                keyType = "ENTREZID",      # 输入基因 ID 的类型
                ont = "BP",                # 可选 BP / CC / MF / ALL
                pAdjustMethod = "BH",
                pvalueCutoff = 0.05,       # 条目自身 p 值阈值
                qvalueCutoff = 0.2,        # q 值阈值（更严格）
                readable = TRUE)           # 结果中显示基因符号而非 Entrez ID

head(as.data.frame(ego))
```

关键参数解释：

- `bitr()`：**ID 转换**。enrichGO 默认接受 Entrez ID，而差异基因通常是基因符号（SYMBOL），需先转换（`fromType` / `toType` 声明来源与目标类型）；
- `keyType`：告诉 enrichGO 输入基因的 ID 类型；
- `pvalueCutoff` 与 `qvalueCutoff`：q 值是 p 值经多重检验校正后的版本（同第 07 章 adj.P 的思路）。clusterProfiler 默认 `qvalueCutoff = 0.2`，较宽松，富集检验通常只要求“信号存在”；
- `readable = TRUE`：把结果中的 Entrez ID 换回基因符号，方便阅读。

结果表的关键列：ID（GO 条目号）、Description（功能描述）、GeneRatio（差异基因中属于该条目的比例）、BgRatio（背景基因中属于该条目的比例）、pvalue、p.adjust、qvalue、geneID（参与基因）。

3. KEGG 富集：通路数据库

KEGG（Kyoto Encyclopedia of Genes and Genomes，京都基因与基因组百科全书）是**通路数据库**，它的条目是“通路图”而不是功能词条。KEGG 富集同样基于超几何检验，但基因要用 Entrez ID，并指定物种代码（人类为 `hsa`）：

```
ekegg <- enrichKEGG(gene = deg_entrez,
                    organism = "hsa",      # 人类
                    keyType = "kegg",
                    pvalueCutoff = 0.05,
                    pAdjustMethod = "BH")
head(as.data.frame(ekegg))
```

如果结果为空（新手常见），先放宽 `pvalueCutoff` 看有没有信号，再检查基因 ID 是否转换正确。KEGG 通路图可以用 **pathview** 绘制：它把差异基因的 logFC 着色到通路图上（红 = 上调、绿 = 下调），直观展示基因在通路中的位置：

```
# BiocManager::install("pathview")
library(pathview)
fc <- setNames(res$logFC, rownames(res))
fc_map <- bitr(names(fc), fromType = "SYMBOL", toType = "ENTREZID", OrgDb = org.Hs.eg.db)
fc_vec <- fc[fc_map$SYMBOL]; names(fc_vec) <- fc_map$ENTREZID   # 重复 ID 可去重，教程略
pathview(gene.data = fc_vec,
         pathway.id = "hsa04668",   # 示例：TNF 信号通路
         species = "hsa", out.suffix = "deg")
```

注意：pathview 需联网从 KEGG 下载通路图；KEGG 有使用条款（学术用途免费），论文中要正确引用。

4. 结果可视化：气泡图与柱状图

clusterProfiler 的结果对象自带绘图函数，最常用的是**气泡图**（dotplot）与**柱状图**（barplot）：

```
dotplot(ego, showCategory = 15)   # 横轴 GeneRatio，点大小 = 基因数，颜色 = q 值
barplot(ego, showCategory = 15)   # 柱长 = 基因数/比例，颜色 = 校正 p 值
```

- **气泡图怎么读**：横轴 GeneRatio（差异基因中属于该条目的比例），纵轴是富集到的条目名，点越大包含的基因越多，颜色越红越显著。
- **柱状图怎么读**：柱长代表包含的基因数（或 GeneRatio），颜色代表显著性。

实际论文常把 GO 的 BP/CC/MF 三个层面并排展示，或用 `enrichplot::cnetplot()` 画“条目—基因”连接网络图。下图是富集结果示例（模拟数据绘制，仅示意样式）：

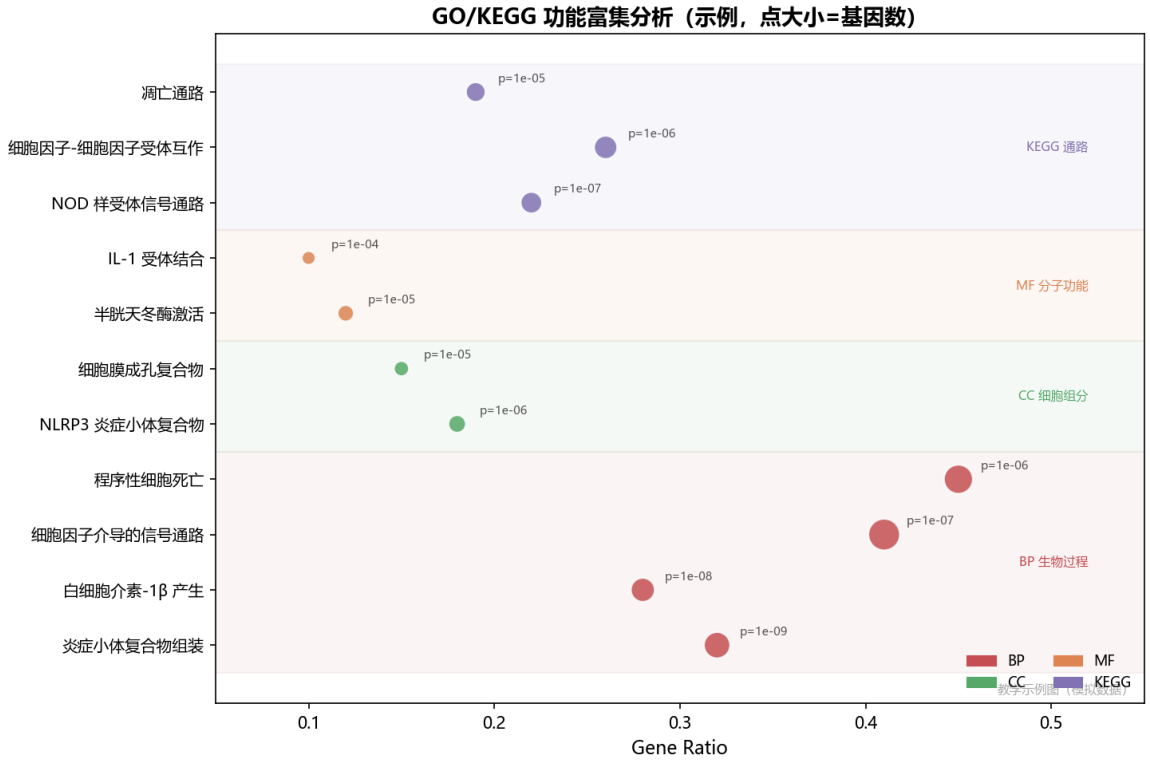


图 12: 图 09 GO/KEGG 富集气泡图（示例，模拟数据）

5. GSEA 进阶：用全部基因排序

ORA 有一个明显局限：它需要先设阈值筛出差异基因，阈值两端的基因被“一刀切”，且只回答“是否富集”，不回答“富集的方向”。**基因集富集分析**（gene set enrichment analysis, GSEA）换个思路：不筛基因，而是把**所有**基因按与表型的关联强度（如 log2FC）从高到低排序，再看某条通路里的基因是整体偏前（激活/上调）还是偏后（抑制/下调）。

两者的区别一句话：**ORA 用“取出来的差异基因”，GSEA 用“全部基因的排序”。**

GSEA 的核心统计量是**富集分数**（enrichment score, ES）：从排序顶端往下走，遇到通路基因加分、遇到非通路基因减分，累计曲线偏离 0 的最大值就是 ES。ES 为正表示通路基因集中在排序前段（整体上调）。之后用置换检验评估显著性，得到 NES（标准化 ES）和 FDR。

clusterProfiler 中跑 GSEA（以 KEGG 为例）：

```
genelist <- sort(setNames(res$logFC, rownames(res)), decreasing = TRUE)  # 全部基因按 logFC 降
gl_map <- bitr(names(genelist), fromType = "SYMBOL", toType = "ENTREZID", OrgDb = org.Hs.eg.d )
genelist <- genelist[gl_map$SYMBOL]
names(genelist) <- gl_map$ENTREZID

gsea_res <- gseKEGG(geneList = genelist,
                    organism = "hsa",
                    minGSSize = 10, maxGSSize = 500,
                    pvalueCutoff = 0.25,      # GSEA 常用较宽松的阈值
```

```
                    seed = TRUE)
head(as.data.frame(gsea_res))

gseaplot2(gsea_res, geneSetID = 1, title = " 示例通路 GSEA 曲线")
ggsave("figures/fig10_gsea.png", width = 7, height = 5, dpi = 300)
```

`gseaplot2` 的输出包含三部分：顶部是排序基因的位置标记（黑色竖线代表该通路基因）；中部是 ES 累积曲线（富集分数随排序位置的变化）；底部是排序度量值（如 logFC）。曲线在**左侧**冲高，说明通路基因整体偏前（上调）。**leading edge**（前沿子集）指曲线达到最大偏移之前的那段基因——“驱动富集” 的核心基因，值得重点关注。

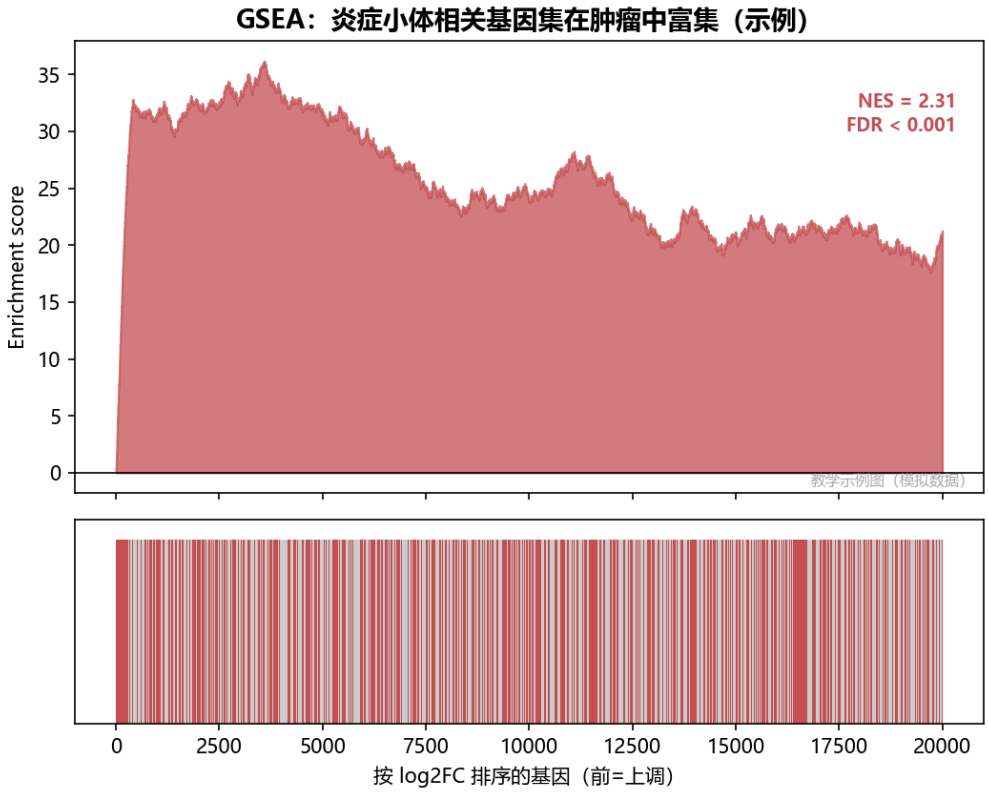


图 13: 图 10 GSEA 富集曲线（示例，模拟数据）

GO 版本的 GSEA 用法类似：`gseGO(geneList = genelist, OrgDb = org.Hs.eg.db, ont = "BP", ...)`。fgsea 包是 GSEA 的更快独立实现，进阶可选。GSEA 需要全部基因的排序向量，这正是第 07 章保留全基因结果表的原因。

6. 结果解读注意

- **读懂通路名的生物学含义**：比如富集到”cytokine-cytokine receptor interaction”（细胞因子—细胞因子受体相互作用）提示炎症信号活跃，与焦亡（促炎性细胞死亡）的课题主题吻

合。不要只报编号。

- **富集≠因果**：某通路“富集”只说明通路里的基因在你的列表中出现得比随机多，不代表通路一定被激活或抑制；方向性问题要靠 GSEA 回答。
- **避免过度解读**：一个 GO 条目描述很长时，看它包含的具体基因（geneID 列）来核实；显著但只包含 1–2 个基因的条目通常不可靠。GO 条目有父子层级，可能造成“重复富集”，别当成独立发现。
- **分层报告**：GO 的 BP/CC/MF、KEGG、GSEA 回答不同问题，论文方法部分要交代各自的输入基因、数据库版本与阈值。

本章小结

- 富集分析用超几何分布 / 费舍尔精确检验回答“我的基因是否集中在某些通路”，记得做 BH 校正；背景基因集要选对。
- GO 分 BP（过程）/CC（组分）/MF（功能）三层；clusterProfiler 流程：`bitr` 转 ID → `enrichGO` → `dotplot`。
- KEGG 用 `enrichKEGG(gene, organism = "hsa")`，通路图可用 pathview 把 logFC 着色上去。
- ORA 用差异基因子集；GSEA 用全部基因的排序，输出 ES 曲线，能回答富集方向。
- 解读时看 geneID 核实、区分相关与因果、报告时明确输入与阈值。

思考题

1. ORA 里“背景基因集”选错会怎样？比如用全基因组做背景，而你的表达矩阵只覆盖 2 万个基因。
2. 为什么 GSEA 不需要先筛差异基因？它和 ORA 各适合什么场景？
3. 富集到"NF-kappa B signaling pathway"（q = 0.01，含 5 个基因）和富集到"Apoptosis"（q = 0.05，含 40 个基因），你会优先报告哪个？为什么？
4. 如果我们的焦亡基因只有 8 个，直接拿它们做富集分析合理吗？怎样做更稳妥？

第 10 章 PPI 网络与 hub 基因

本章目标

学完本章，你将能够：

1. 用自己的话说清楚什么是蛋白质-蛋白质相互作用（PPI）网络，以及为什么“连接最多的节点”常被当作候选关键基因；
2. 在 STRING（string-db.org）网页上输入基因列表、选择物种 Homo sapiens、导出互作关系文件（TSV）；
3. 安装并基本使用 Cytoscape，把 STRING 网络导入、布局、美化；
4. 用 cytoHubba 插件按 MCC / Degree 算法筛选出 top hub 基因；
5. 知道如何用 R 的 igraph 做最简单的替代分析，并理解 hub 基因如何衔接下一章的生存分析。

正文

10.1 什么是 PPI 网络，为什么要找 hub 基因

先打个比方。想象一个社交网络：每个人是一个点，朋友关系是两点之间的连线。有的人朋友特别多，消息靠他们一传十、十传百，信息流动高度依赖这些人。蛋白质也类似：一个细胞里有成千上万种蛋白质，它们并不是各干各的，而是通过物理结合、共同参与某个通路等方式“打交道”。把“谁和谁存在相互作用”画成图——节点（node）是蛋白质，边（edge）是相互作用——就得到了蛋白质-蛋白质相互作用（protein-protein interaction, PPI）网络。

hub 基因（hub gene）指网络中连线特别多的节点，通常对应在信号传导、细胞命运决定中起枢纽作用的蛋白。肿瘤研究里有一个朴素但常用的假设：hub 基因一旦异常，会通过它的大量互作伙伴把影响“放大”出去，因此更可能是关键基因（candidate key gene）。在我们这个焦亡课题里，PPI 网络的作用是**缩小候选范围**：第 7 章的差异基因、焦亡基因列表往往有几十上百个，不可能全部做生存分析；优先挑出“网络位置重要”的 hub 基因来检验，既减少多重检验负担，也让故事有逻辑。

需要提醒：hub ≠已证实的驱动基因。它只是“网络位置重要”的线索，后续还要靠生存分析、独立验证乃至实验来检验。

10.2 STRING 数据库：网页操作

STRING（Search Tool for the Retrieval of Interacting Genes/Proteins，string-db.org）是目前最常用的 PPI 数据库之一。它整合了多种证据来源（实验、数据库注释、共表达、文本挖掘等），为每对互作打一个综合分数（combined score，取值 0~1），分数越高越可信。

操作步骤（网页版，全部是鼠标操作，不需要写代码）：

1. 打开 string-db.org；
2. 在输入方式中选择"List of names"（列表输入），把我们前面得到的焦亡基因列表粘贴进去，每行一个基因符号（如 GSDME、GSDMD、CASP1）；
3. 选择物种：Organism 下拉框选 **Homo sapiens**（人类）；
4. 点击 SEARCH，在匹配确认页核对基因都认对了，点 CONTINUE，得到网络图；
5. 点 Settings（设置），把置信度（confidence）设为 **medium confidence（0.400）**——这是常用默认阈值，兼顾互作数量与可信度；想更严格可以试 0.7，但网络会明显变稀疏；
6. 勾选 hide disconnected nodes（隐藏孤立节点），让图更干净；
7. 可选设置：Network type（网络类型）有 full network（全部证据）与 physical subnetwork（仅物理结合证据）两个选项——如果只想看 “蛋白是否真的结合”，可选 physical，但边数会明显变少；默认 full network 即可；
8. 导出数据：点 Export（导出）→选 “as simple tabular text output (TSV)”，下载互作关系表，里面至少包含 node1、node2、combined_score 等列，这是后续 Cytoscape 和 R 分析的输入。

两点提醒：第一，STRING 的分数里包含 textmining（文本挖掘）证据——从文献里 “共现” 扒出来的关联，可能有噪音，解读时要留意；第二，网页上截的网络图只能当示意图，正式分析请用导出的 TSV 数据。

10.3 Cytoscape 入门

Cytoscape（cytoscape.org）是开源的网络可视化桌面软件（基于 Java，Windows/Mac 都需先装 Java 运行环境），也是 PPI 分析论文里最常见的出图工具。

基本流程：

1. 从官网下载安装 Cytoscape，按向导完成安装；
2. 打开软件：File → Import → Network from File，选择 STRING 导出的 TSV；导入向导会自动把前两列识别为边的两端（两个互作的蛋白），其余列作为边的属性；
3. 点 OK 后出现网络。默认布局较乱，用 Layout（布局）菜单换布局：常用 yFiles Organic（有机布局）或 Prefuse Force Directed（力导向布局），让互作密集的节点聚在一起；
4. 美化：在 Style（样式）面板里把节点大小、颜色映射到 Degree（度，即连接数）——连接越多的节点越大越红，hub 一目了然；
5. 导出图片：File → Export → Network to Image，存成 PNG，供论文或教程使用。

10.4 cytoHubba：筛选 hub 基因

网络“看着像”还不够，我们需要量化排序。cytoHubba 是 Cytoscape 的插件（App），提供多种节点重要性打分算法。步骤：

1. 安装：Cytoscape 菜单 Apps → App Manager →搜索 cytoHubba → Install；
2. 打开导入的网络。建议先取出**最大连通分量**（最大的连成一片的子网络，避免孤立小团块干扰打分）再分析：可用 Select 过滤或菜单 Tools → Network Analyzer 辅助；
3. 运行：Apps → cytoHubba，选择要打分的网络；
4. 算法选 **MCC**（Maximal Clique Centrality，最大团中心性）或 **Degree**（度），Top N 填 10 或 20，点 Submit；
5. 结果按分数排序并可在网络里高亮前 N 个节点——这些就是候选 hub 基因。

经验上 MCC 比单纯 Degree 更稳健（它同时考虑了节点所在的小团体结构），但不同算法结果会有差异，可把两种算法的 top 列表取交集。cytoHubba 的结果可以直接截图，也可以从面板复制基因列表，供下一步生存分析使用。

下图是 PPI 网络与 hub 基因的**示例图**（模拟数据绘制，仅演示图形样式与 hub 高亮方式；读者用自己数据跑 STRING + Cytoscape 后得到的是自己的图）：

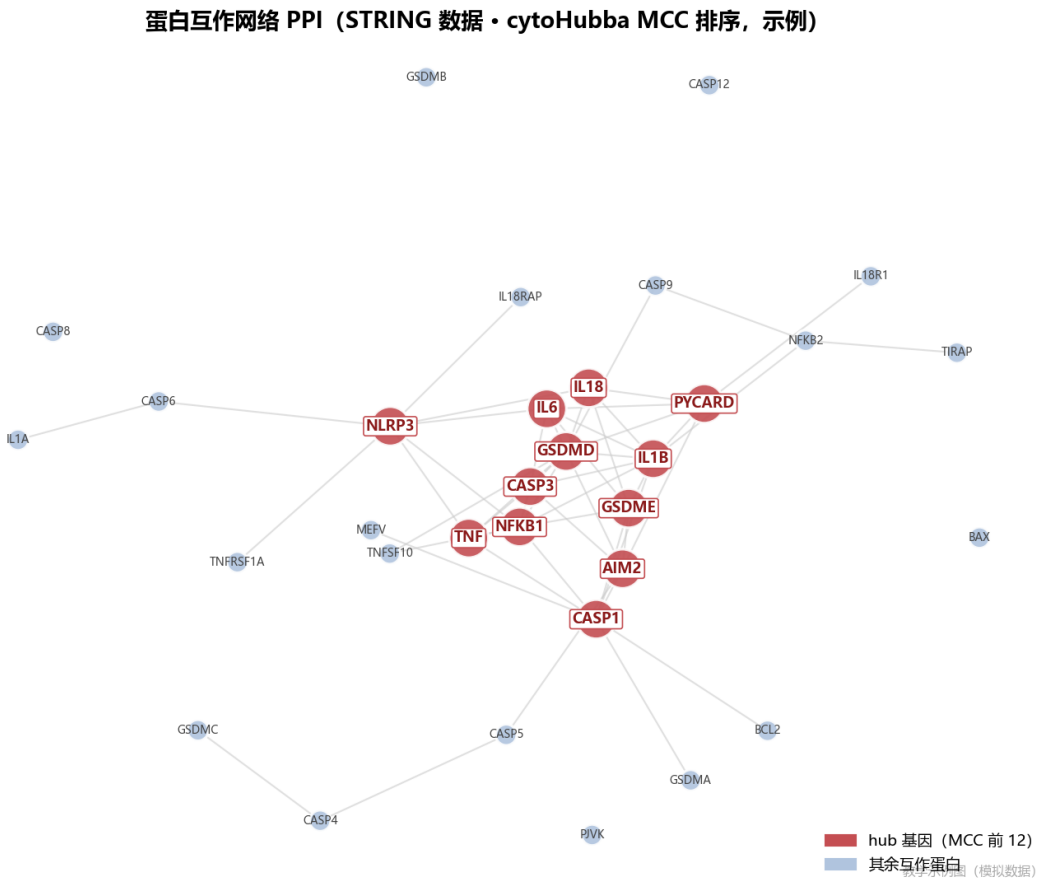


图 14: 图 11 PPI 网络与 hub 基因示意（示例图，模拟数据，仅演示样式）

10.5 替代方案：用 R 的 igraph 做简单 PPI 分析（可选）

如果暂时不想装 Cytoscape，R 包 igraph 可以做最基本的分析——读入 STRING 的 TSV、计算度、画简单网络：

```r
# install.packages("igraph")
library(igraph)

# 读入 STRING 导出的 TSV（含 node1, node2 等列）
ppi <- read.delim("string_interactions.tsv", stringsAsFactors = FALSE)

# 建无向图：每条边连接两个蛋白
g <- graph_from_data_frame(ppi[, c("node1", "node2")], directed = FALSE)

# 度：每个节点连接的边数
deg <- degree(g)
sort(deg, decreasing = TRUE)[1:10]   # 连接最多的前 10 个基因

# 简单可视化（教学演示用；正式发表建议用 Cytoscape 精修出图）
V(g)$size <- deg * 2
plot(g, layout = layout_with_fr(g), vertex.label.cex = 0.6)
```

igraph 是“够用但简陋”的替代：能算度、介数（betweenness）等指标，但交互式调整和出图质量不如 Cytoscape。正式分析建议两者结合：igraph 算指标，Cytoscape 出图。

10.6 衔接：hub 基因与生存分析

拿到 hub 基因后，核心问题变成：这些“网络里重要”的基因，是否真的与患者预后有关？比如某个 hub 基因高表达的患者是否活得更短？这需要回到带生存随访的数据（TCGA-LIHC 的临床信息里有生存时间 OS.time 与结局 OS），对每个 hub 基因做 KM 生存分析和 Cox 回归——这正是下一章的内容。换句话说，PPI 网络在这里充当“漏斗”：从几十上百个差异/焦亡基因中先筛出 hub 基因，再逐一检验预后价值，最后用 LASSO 把这些基因压缩成一个风险模型。

本章小结

- PPI 网络把蛋白质互作画成“节点-边”图，hub 基因是连线密集的关键节点，是候选关键基因的线索而非定论；
- STRING 网页版流程：输入基因列表→选 Homo sapiens →置信度设 medium 0.4 →导出 TSV；网页截图只作示意图，正式分析用导出的数据；
- Cytoscape 是主流网络可视化软件：导入 TSV →换布局→按度映射节点大小/颜色→导出 PNG；

- cytoHubba 用 MCC / Degree 等算法量化节点重要性，取 top 10/20 作为 hub 基因候选；
- 不装软件时可用 R 的 igraph 快速算度并画简单网络；
- hub 基因是 “缩小候选池” 的手段，下一步用生存分析检验其预后价值。

思考题

1. 为什么 STRING 的 confidence 设得越高，网络里的节点和边通常越少？0.4 与 0.7 各适合什么场景？
2. 如果一个基因在 STRING 里 “连接很多”，能直接说它是致癌基因吗？还需要哪些证据？
3. cytoHubba 的 Degree 与 MCC 排序结果不一致时，你倾向信哪个？为什么？
4. 假如你的差异基因列表有 200 个基因，直接用全部基因逐个做生存分析会有什么问题？（提示：多重检验）

第 11 章生存分析与预后模型

本章目标

学完本章，你将能够：

1. 理解随访数据的两要素（生存时间 + 结局事件）与删失（censoring）的含义；
2. 用 survival + survminer 画 KM 生存曲线并做 log-rank 检验；
3. 把连续表达值分成高/低组（中位数分组、surv_cutpoint），并理解截点选择的偏倚风险；
4. 用 coxph 做单因素/多因素 Cox 回归，解读 HR 与 95% CI，画森林图；
5. 用 glmnet 的 LASSO 筛选基因并构建风险评分模型；
6. 用 timeROC 评估 1/3/5 年 AUC，用 rms 包画诺模图；
7. 说出生存分析常见坑（样本量、删失比例、数据窥探、多重检验）。

正文

11.1 生存分析基础：随访、删失、KM 曲线与 log-rank

生存分析（survival analysis）处理的是“到某个事件发生还要多久”的数据。在肿瘤队列里，我们关心的是患者从入组到死亡（或复发）的时间。每个患者都有两个关键变量：

- 生存时间（survival time，如 OS.time）：随访了多少时间（月或天）；
- 结局（status / event）：到研究结束时是否发生了事件（死亡记为 1，未发生记为 0）。

难点在**删失（censoring）**：有些患者到研究结束还活着，有些中途失访，还有些死于其他原因——我们只知道“他至少活了这么久”，并不知道确切的事件时间。这类数据不能扔掉：生存分析的价值正在于正确处理删失，删失患者贡献的是“到删失时点为止仍然存活”这条信息。

KM 曲线（Kaplan-Meier curve）：把患者按时间排序，逐段估计存活概率，画成阶梯状曲线。把高表达组和低表达组的两条曲线画在一起，可以直观比较两组的生存差异。**log-rank 检验**（log-rank test）：检验两条曲线是否显著不同的非参数方法，输出 p 值——$p < 0.05$ 表示两组生存差异有统计学意义。注意：p 值只回答“有没有差异”，不回答差异多大；效应大小要看曲线分开的程度和后面的 HR。

还有两个常用概念：**中位生存时间**（median survival）指存活概率降到 50% 所需的时间，可以从 KM 曲线上读取（曲线与 0.5 水平线的交点），是报告生存结果时最常引用的数字之一；**随访时间**（follow-up time）指队列观察的时长跨度，通常用中位随访时间描述，它大致决定了 KM 曲

线能可靠延伸到多远——随访太短，长时点上的曲线就没有数据支撑。

下图是 KM 曲线的**示例图**（模拟数据绘制，仅演示样式与 p 值标注方式）：

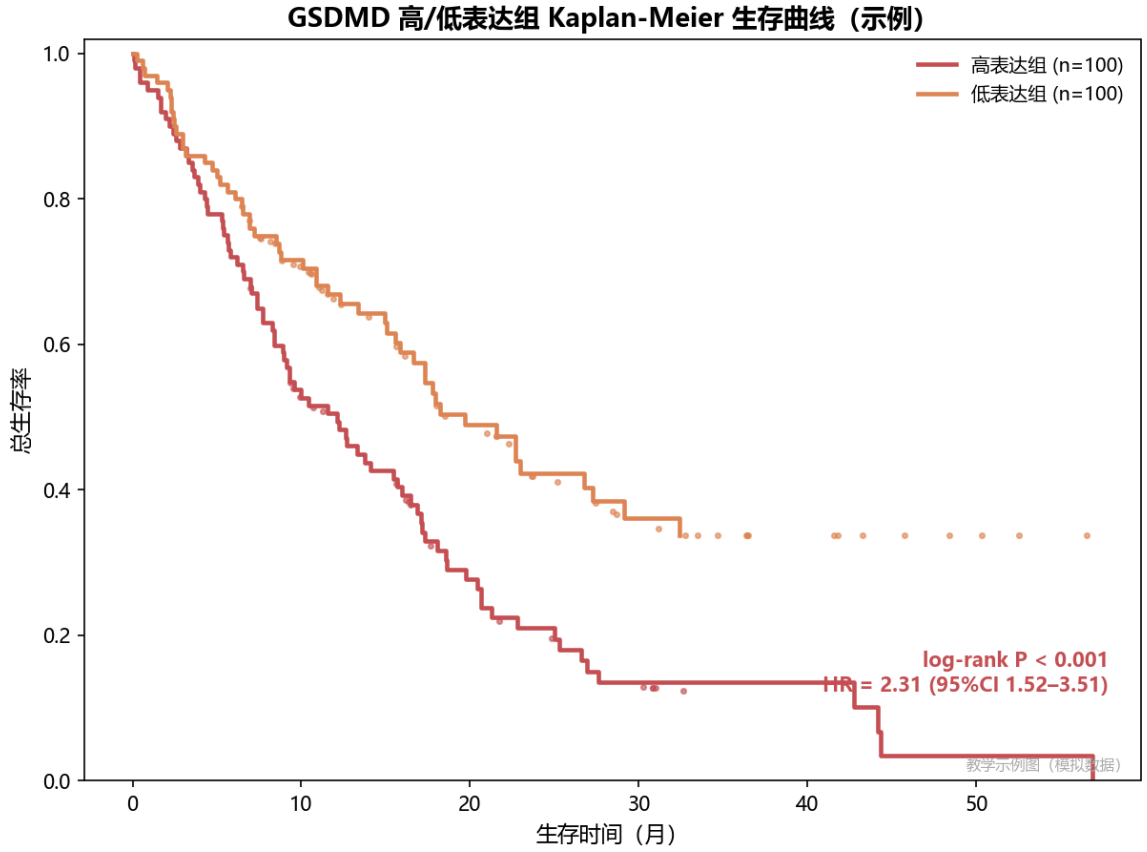


图 15: 图 12 KM 生存曲线示例（示例图，模拟数据）

11.2 表达分组：连续值怎么变成高/低组

基因表达是连续值，而 KM 曲线和 log-rank 需要分组。常用的做法有两种：

1. **中位数分组（median split）**：表达高于全体中位数 = 高表达组，否则低表达组。简单、可复现、默认推荐。
2. **最佳截点（optimal cutpoint）**：用 `survminer::surv_cutpoint` 在所有可能的截点里挑一个让两组差异最显著的阈值。

方法 2 看似“更聪明”，但有一个重要风险：**数据窥探（data snooping）**。在同一份数据上“先找最显著的截点、再用它检验”，p 值会被系统性低估，假阳性升高；而且这个截点是针对当前队列量身定做的，换一个队列往往不成立。因此建议：默认用中位数分组；如果要用 surv_cutpoint，就把它当作探索性结果，必须在独立验证队列中确认，并在论文里如实说明截点来源。另外，分组用的表达值应是标准化、可比的值（如 TPM/FPKM 或 log 变换后），原始 count 不宜直接比较。

11.3 单基因生存分析：survival + survminer

以 TCGA-LIHC 为例：假设已有表达矩阵 `expr`（行 = 基因，列 = 样本）和临床信息 `clin`（含 OS.time 与 OS 列），分析某个焦亡基因（如 GSDME）的代码如下：

```r
# 依赖包：survival, survminer
# install.packages(c("survival", "survminer"))
library(survival)
library(survminer)

# 组装数据：生存时间、结局、目标基因表达（注意样本顺序与 clin 一致）
dat <- data.frame(
  time   = clin$OS.time,                  # 生存时间（单位与后续 Cox/ROC 保持一致）
  status = clin$OS,                       # 0 = 删失/存活，1 = 死亡
  expr   = as.numeric(expr["GSDME", ])
)

# 中位数分组
dat$group <- ifelse(dat$expr > median(dat$expr), "High", "Low")
dat$group <- factor(dat$group, levels = c("Low", "High"))

# KM 拟合 + 绘图（pval = TRUE 显示 log-rank p 值，risk.table 显示风险人数）
fit <- survfit(Surv(time, status) ~ group, data = dat)
ggsurvplot(fit, data = dat, pval = TRUE, risk.table = TRUE,
           xlab = "Time (months)", ylab = "Overall survival",
           legend.title = "GSDME")
```

结果解读：$p < 0.05$ 说明该基因的表达高低与生存相关；看曲线中段（中位生存时间附近）两组分开的幅度判断方向——高表达组的曲线更低，说明高表达预后更差。对几十个 hub 基因逐个跑一遍，得到一个“候选预后基因”列表。

11.4 单因素/多因素 Cox 回归：HR、95% CI 与森林图

KM 只能回答“有没有差异”，Cox 回归（Cox proportional hazards regression）能回答“差多少”。它把某一时刻的死亡风险建模为风险比（hazard ratio, HR）：$HR = 1$ 表示无差异；$HR > 1$ 表示该变量对应的风险更高（预后更差）；$HR < 1$ 相反。报告时给出 95% 置信区间（95% CI），区间不跨越 1 即认为显著。

- **单因素 Cox（univariate）**：模型里只放一个变量（如基因表达），回答“这个基因单独看是否与预后相关”；
- **多因素 Cox（multivariate）**：同时放入基因和临床协变量，**校正（adjust for）**混杂因素——临床上通常校正年龄（age）、性别（gender）、肿瘤分期（stage / TNM）。多因素中基因仍显著，说明它的预后价值不依赖于这些临床因素，即“独立预后因素”。

```r
# 单因素：只看 GSDME
fit1 <- coxph(Surv(time, status) ~ expr, data = dat)
summary(fit1)                 # 看 coef、exp(coef) = HR、Pr(>|z|)

# 多因素：校正年龄、性别、分期
dat$age    <- clin$age
dat$gender <- clin$gender
dat$stage  <- clin$stage
fit2 <- coxph(Surv(time, status) ~ expr + age + gender + stage, data = dat)
summary(fit2)

# 森林图（forest plot）：一行一个变量，点 = HR 估计，横线 = 95% CI
ggforest(fit2, data = dat)
```

下图是 Cox 森林图的**示例图**（模拟数据，仅演示样式）：

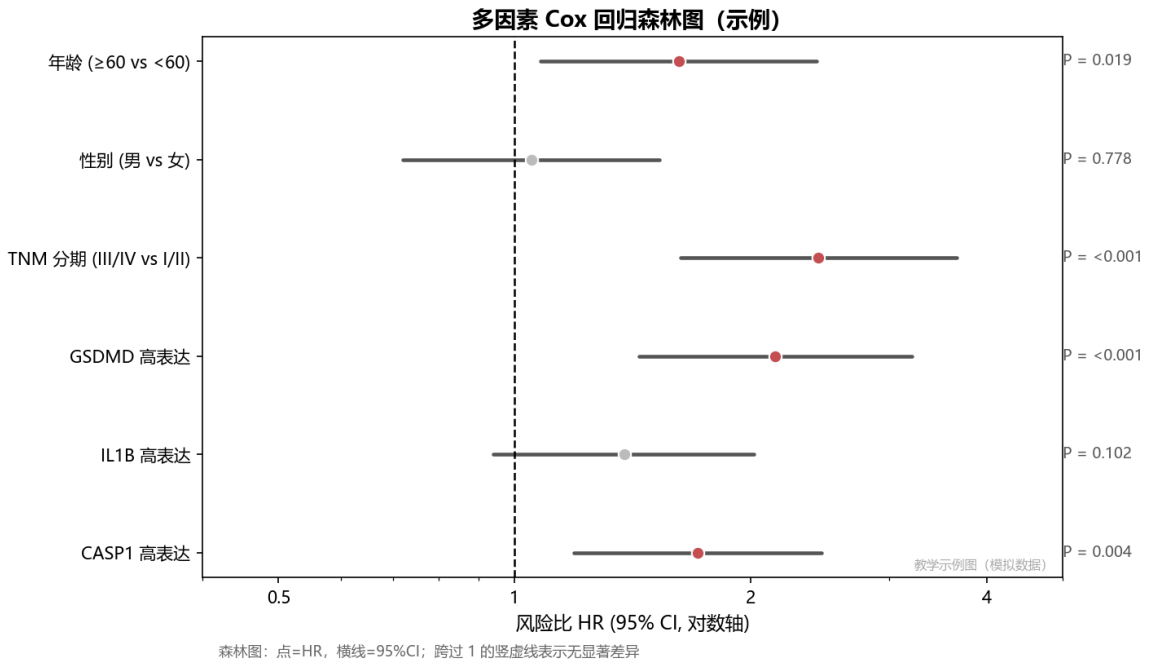


图 16: 图 13 Cox 森林图示例（示例图，模拟数据）

解读要点：森林图每一行一个变量，圆点是 HR 点估计，横线是 95% CI；横线不穿过中间的竖线（HR = 1）表示该变量显著；“基因在多因素校正后仍显著”是论文里最常引用的结论。

还有一个前提需要交代：Cox 模型依赖**比例风险假设**（proportional hazards assumption）——各组之间的风险比随时间大致恒定。可以用 `cox.zph(fit2)` 做检验，$p < 0.05$ 提示假设可能不成立，此时该变量的 HR 解读要谨慎（分层分析或时变系数是进阶处理）。大多数预后论文会报告这个检验，审稿人也常问。

注意：多因素模型需要足够的**事件数**（即死亡人数）。经验规则是每个预测变量至少约 10 个事件（events per variable, EPV）；事件太少而变量太多时，系数估计极不稳定——这正是下一节

用 LASSO 压缩变量的动机之一。

11.5 LASSO 构建多基因风险模型

逐个基因做单因素分析有两个问题：多重检验（几十个基因各测一次，假阳性累积）和共线性（焦亡基因常共表达，变量冗余）。LASSO（Least Absolute Shrinkage and Selection Operator，glmnet 包）是一种正则化回归：它给系数加 L1 惩罚，把不重要的变量系数直接压到 0，自动完成“选基因”，同时抑制过拟合。模型的形式为：

风险评分（risk score）= Σ (基因系数×基因表达量)

构建流程的关键意识是**训练集 / 验证集划分**：先在训练集上选基因、定系数，再到验证集检验，绝不能把验证集的信息泄漏进建模。

```r
# install.packages("glmnet")
library(glmnet)

# x: 候选基因表达矩阵（行 = 样本，列 = 基因；行顺序须与 y 一致）
X <- t(as.matrix(expr[c("GSDME", "GSDMD", "GSDMB", "CASP1", "IL1B", "IL18"), ]))
y <- cbind(time = clin$OS.time, status = clin$OS)   # glmnet 的 Cox 族要求两列矩阵

# 7:3 划分训练 / 验证
set.seed(2024)                                      # 固定随机种子，保证可复现
idx <- sample(1:nrow(X), floor(0.7 * nrow(X)))
trainX <- X[idx, ]; trainY <- y[idx, ]
testX  <- X[-idx, ]; testY  <- y[-idx, ]

# 10 折交叉验证选择最优 lambda（alpha = 1 即 LASSO，family = "cox" 即 Cox 生存模型）
cvfit <- cv.glmnet(trainX, trainY, family = "cox", alpha = 1)
plot(cvfit)                # 两条虚线对应 lambda.min（最小偏差）与 lambda.1se（最简模型）
cvfit$lambda.min

# 取出系数非 0 的基因（即被选入模型的基因）
coefs <- coef(cvfit, s = "lambda.min")
genes <- rownames(coefs)[coefs[, 1] != 0]

# 计算风险评分（type = "link" 返回线性预测值，即 Σ 系数×表达）
risk_train <- as.numeric(predict(cvfit, newx = trainX, s = "lambda.min", type = "link"))
risk_test  <- as.numeric(predict(cvfit, newx = testX,  s = "lambda.min", type = "link"))
```

lambda 的选择：`lambda.min` 取交叉验证偏差最小的点；`lambda.1se` 取“偏差在最小值一个标准误以内、但模型更简约（基因更少）”的点。论文常见 `lambda.min`，追求简约可用 `lambda.1se`。得到风险评分后，通常再按中位数把患者分成高危 / 低危组，画两组 KM 曲线验证区分度——这一步训练集、验证集都要做。

一个实现细节：glmnet 默认会把每个变量标准化后再计算（standardize = TRUE），因此各基因表达量级的差异（如有的基因 TPM 上百万、有的只有几十）不会直接扭曲系数；但训练集与验证集的预处理必须完全一致。解读风险评分方向时注意系数符号——系数为正的基因高表达会推高风险评分，两者方向要一致，别把“高危”解释反了。

11.6 模型评估：ROC / AUC 与诺模图

时间依赖 ROC（time-dependent ROC）：生存数据里“结局是否发生”取决于观察时长，不能直接用普通 ROC。`timeROC` 包可以在指定时间点（如 1、3、5 年）计算 AUC——AUC 表示“随机抽一个患者，风险评分能多大程度区分他在该时间点是否死亡”。0.5 = 纯随机，越接近 1 越好；肿瘤预后模型做到 0.65~0.75 已算不错，不要期待 0.9+。

```
# install.packages("timeROC")
library(timeROC)

# 教学演示：把训练/验证的风险评分与生存数据拼回全体（真实分析请保持向量对齐）
risk_all <- c(risk_train, risk_test)
t_all <- c(clin$OS.time[idx], clin$OS.time[-idx])
s_all <- c(clin$OS[idx], clin$OS[-idx])

ROC <- timeROC(T = t_all, delta = s_all, marker = risk_all, cause = 1,
               weighting = "marginal",
               times = c(365, 1095, 1825),   # 1/3/5 年（天），单位必须与 T 一致
               iid = TRUE)
ROC$AUC
plot(ROC, time = 365, col = "red")
```

诺模图（nomogram，列线图）：把 Cox 模型变成可手算的工具——每个变量按取值给分，把分数相加得到总分，再对应到 1/3/5 年生存概率。用 rms 包：

```
# install.packages("rms")
library(rms)
dd <- datadist(dat); options(datadist = "dd")   # rms 需要先设定数据分布

# rms 版 Cox（必须 x = TRUE, y = TRUE, surv = TRUE 才能算生存概率）
fit_cph <- cph(Surv(time, status) ~ expr + age + stage, data = dat,
               x = TRUE, y = TRUE, surv = TRUE)

# 1/3/5 年生存概率函数（时间单位与数据一致）
surv <- Survival(fit_cph)
nom <- nomogram(fit_cph,
                fun = list(function(x) surv(365, x),
                           function(x) surv(1095, x),
                           function(x) surv(1825, x)),
```

```
                funlabel = c("1-year OS", "3-year OS", "5-year OS"))
plot(nom)
```

下图是 ROC 曲线与诺模图的**示例图**（模拟数据，仅演示样式）：

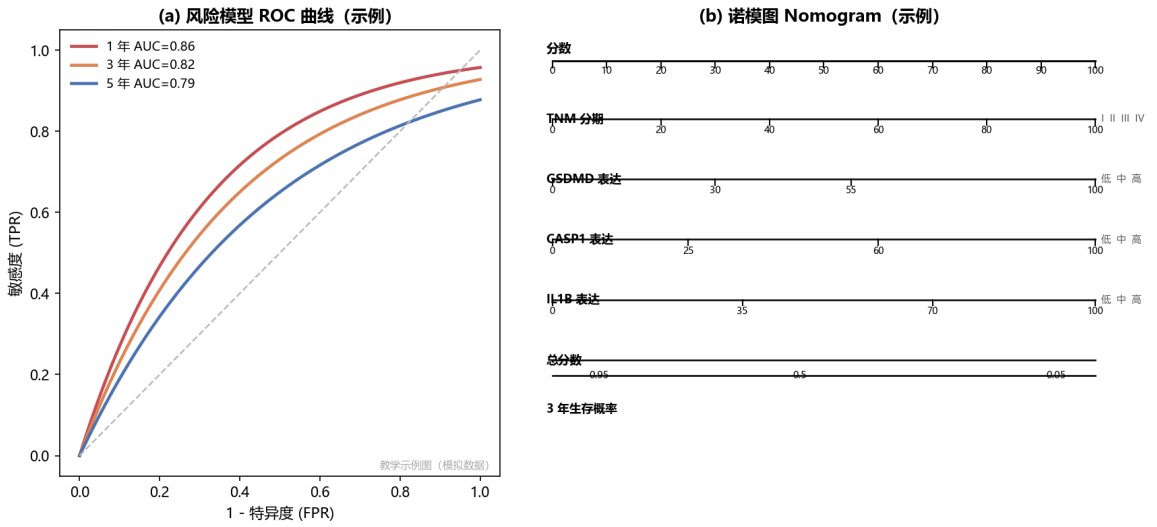


图 17: 图 14 ROC 曲线与诺模图示例（示例图，模拟数据）

进阶评估还有：校准曲线（calibration，比较预测概率与实际观察，`rms::calibrate`）和决策曲线分析（DCA，评估模型在临床决策中的净获益），高分论文常配图。另外建议把训练集与验证集的 AUC 放在一起报告——两者差距过大提示过拟合，差距小且都高于 0.6 才是“模型真的有用”的证据。

11.7 生存分析常见坑

1. **样本量与事件数**：限制统计功效的往往不是样本总数而是事件数（死亡/复发人数）。事件太少时 HR 估计不稳，多因素模型尤其危险（EPV $\geqslant$ 10 的经验规则）。
2. **删失比例过高**：如果研究结束时大部分患者仍存活（删失率高），KM 曲线的右尾（长随访时间）只有少数患者支撑，那一段曲线不可靠，解读要谨慎。
3. **截点选择偏倚（数据窥探 bias）**：`surv_cutpoint` 在同一数据上“挑最显著截点再检验”会夸大显著性。默认用中位数分组；探索性截点必须在独立队列验证。
4. **多重检验**：对几十上百个基因逐一做 KM/Cox，$p < 0.05$ 的基因里必然混入假阳性（做 100 次检验、$\alpha = 0.05$ 时平均约 5 个假阳性）。对策：报告 FDR 校正、用 LASSO 统一选择、最关键的是独立验证。
5. **只看 p 值不看效应量**：p 值显著但 HR 接近 1（如 1.05）的结果几乎没有临床意义，报告时一定要给出 HR 与 95% CI，并结合曲线分开的程度一起判断。
6. **时间单位不一致**：TCGA 的生存时间常以“天”为单位；混用月/天会让 AUC 时间点、诺模图全乱，建模前统一单位。
7. **信息泄漏**：先用全部数据选基因、定截点，再在同一份数据上“验证”——这不是验证。验

证必须发生在真正没有参与建模的独立数据上（下一章详述）。

本章小结

- 生存分析处理 “时间 + 事件” 数据，删失信息必须保留，不能丢弃；
- KM 曲线 + log-rank 回答 “两组有没有生存差异”，Cox 回归回答 “风险差多少”（HR 与 95% CI）；
- 表达分组默认用中位数；最佳截点有数据窥探风险，须在独立队列验证；
- 多因素 Cox 校正年龄、性别、分期等混杂；事件数不足时慎用多变量模型；
- LASSO（glmnet，family = “cox”）自动选基因并给出风险评分 Σ (系数×表达)，建模前必须先划分训练/验证集；
- timeROC 评估 1/3/5 年 AUC，rms 画诺模图与校准曲线；
- 常见坑：事件数不足、删失过高、截点偏倚、多重检验、只看 p 值不看效应量、单位不一致、信息泄漏。

思考题

1. 为什么说 “删失患者的信息不能扔掉”？如果一个队列 90% 的患者到研究结束时还活着，KM 曲线的哪一段最不可靠？
2. HR = 1.8（95% CI：1.2–2.7）如何向非统计背景的读者解释？
3. surv_cutpoint 找出的截点在训练集显著、在验证集不显著，最可能的原因是什么？
4. LASSO 相比 “逐个基因做单因素 Cox 再挑显著的” 有什么优势？lambda.min 和 lambda.1se 你倾向选哪个，为什么？

第 12 章独立验证与拓展

本章目标

学完本章，你将能够：

1. 说明为什么独立验证是生信论文的“生死线”，能区分过拟合与真信号；
2. 用 GEOquery 获取 GSE14520，独立验证 hub 基因的表达与生存关联、验证风险模型；
3. 在 HPA（人类蛋白图谱）查询 hub 基因蛋白水平的免疫组化染色，并正确截图引用；
4. 理解免疫浸润分析（CIBERSORT / ssGSEA）的基本思想及其与焦亡的关联（进阶内容）；
5. 了解单细胞分析、实验验证等拓展方向，并正确认识纯生信挖掘论文的审稿趋势。

正文

12.1 为什么必须独立验证

前面几章的所有分析——差异表达、PPI、KM 生存、LASSO 建模——都是在同一份 TCGA-LIHC 数据上完成的。任何统计建模都会“记住”训练数据里的噪声，尤其当候选基因很多、还用了最优截点时，模型在 TCGA 上表现好是“应该的”，说明不了推广价值。这就是**过拟合（overfitting）**。

独立验证（independent validation）就是用一份从未参与建模的数据（外部队列）检验同样的结论：如果 hub 基因的高低表达分组、风险模型的高低危分组在验证集里依然显著、AUC 依然可观，审稿人才会相信这是真信号而不是数据巧合。在现在的审稿环境里，“只在训练集显著”几乎必然被质疑；而“训练集发现 + 独立队列验证”是最低限度的证据链，也是本章的主题。

12.2 GEO 独立队列验证：GSE14520

GSE14520 是复旦大学中山医院的 HBV 相关肝细胞癌队列（Affymetrix U133A 2.0 芯片，含约 240 例 HCC 患者的肿瘤与癌旁样本），带有随访信息，是 HCC 生信论文最常用的独立验证队列之一。下载方法见第 6 章（需要联网）。

代码思路（教学演示，真实运行需联网下载）：

```
library(GEOquery)
gse <- getGEO("GSE14520", GSEMatrix = TRUE, getGPL = TRUE)[[1]]

expr  <- exprs(gse)     # 探针 × 样本 的表达矩阵
```

```r
pdata <- pData(gse)      # 样本信息：用 colnames() 查看列名，
                         # 找到生存时间与结局列（随访信息在 pData 中）
```

拿到数据后按三步验证：

1. **探针→基因名映射**：用平台注释包（U133A 2.0 对应 hgu133a2.db）把探针 ID 转成基因符号；一个基因对应多个探针时合并（取均值或方差最大的探针）；
2. **表达验证**：hub 基因在肿瘤 vs 癌旁中的差异方向是否与 TCGA 一致（箱线图 + wilcoxon 检验）；
3. **生存验证**：用与 TCGA 相同的分组方式（如中位数）在同一基因上分组，画 KM 曲线做 log-rank 检验；对 LASSO 风险模型，把 TCGA 训练好的系数**原封不动**套到 GSE14520 对应基因的表达上算风险评分，再检验高低危组的生存差异与 AUC——系数必须来自训练集、不得在验证集上重新拟合，否则验证就失效了。

（教学演示说明：上面是思路与关键代码骨架，正式分析还要处理缺失值、样本匹配、时间单位统一等细节。）

12.3 蛋白水平验证：HPA 查免疫组化

mRNA 表达不等于蛋白表达。人类蛋白图谱（Human Protein Atlas，proteinatlas.org，简称 HPA）收录了数万种蛋白在正常组织与多种癌症中的免疫组化（immunohistochemistry, IHC）染色图像，可免费在线查看，是 “mRNA 之外再加一层证据” 的常用手段。操作步骤：

1. 打开 proteinatlas.org，搜索 hub 基因（如 GSDME）；
2. 看 **Pathology（病理）**页签，找 Liver cancer（肝细胞癌）条目下的染色图；**Tissue（组织）**页签可看正常肝组织的染色；
3. 对比正常肝细胞与 HCC 组织的染色强度（negative / weak / moderate / strong）和阳性比例，判断蛋白水平是否有差异、方向是否与 mRNA 一致；
4. **截图引用**：截图前记录基因名、数据来源（Pathology/Tissue）、查看日期；论文中注明来源 “Human Protein Atlas (proteinatlas.org)” 与访问日期。HPA 的数据库内容以知识共享许可发布，可署名引用，但请遵守其使用条款，不要自行改图。

注意：HPA 每个基因的 HCC 染色样本数有限，IHC 染色只能作为支持性证据（正文或补充材料的一张图），不能替代定量实验。

12.4 免疫浸润分析简介（进阶内容）

肿瘤微环境（tumor microenvironment）中浸润的免疫细胞类型与比例，与预后和治疗反应密切相关。细胞焦亡本身就是一种促炎性细胞死亡（见第 1 章），会释放炎症因子、招募免疫细胞，因此 “焦亡基因×免疫浸润” 是这类论文常见的配套分析。两种主流思路：

- **CIBERSORT**：用特征基因矩阵（LM22，对应 22 种免疫细胞）对整体表达谱做**反卷积**

（**deconvolution**），估算每个样本中各种免疫细胞的比例。官方版本需在其网站注册，对数据标准化有要求，门槛较高；

- **ssGSEA**（单样本基因集富集分析，GSVA 包实现）：对每个样本计算一组免疫细胞标记基因集的富集分数，分数高低代表该细胞类型的相对丰度。实现简单、免费可复现，是新手首选：

```
# BiocManager::install("GSVA")
library(GSVA)
# 免疫细胞标记基因集（例如从文献整理的 marker 列表，元素为基因符号向量）
# gs <- readRDS("immune_markers.rds")
# score <- gsva(expr, gs, method = "ssgsea")   # 行 = 样本（新旧版本参数写法略有差异）
```

拿到浸润分数后，可以做的分析有：各免疫细胞在高危/低危组间的差异（wilcoxon 检验 + 箱线图）、浸润分数与风险评分的相关性（cor.test）、焦亡基因表达与免疫浸润的相关热图等。诚实说明：本章只是概念入门；CIBERSORT 的完整流程与免疫结果的解读属于进阶内容，新手应先把“独立验证”做扎实，再决定是否加这一节。

12.5 临床特征关联（可选）

锦上添花的描述性分析：检验 hub 基因表达与临床特征的关系——例如表达在临床分期（stage I–IV）或病理分级（grade）之间是否有差异（多组比较用 Kruskal-Wallis 检验），与性别、HBV 感染等二分类特征的关系（两组比较用 Wilcoxon/Mann-Whitney 检验）。结果用箱线图/小提琴图展示。注意两点：这类分析是描述性的，不能当作因果证据；分组样本量不均衡时结果容易误导，显著性解释要克制。

12.6 扩展方向与实验闭环

1. **单细胞测序（scRNA-seq）：**在 HCC 单细胞数据中（可在 GEO 以 hepatocellular carcinoma + single cell 检索，有多个公开数据集）看焦亡基因主要在哪类细胞（如肿瘤细胞、巨噬细胞）中表达，比 bulk 数据精细得多。学习成本高，通常作为论文的补充分析，属进阶方向。
2. **实验验证（qPCR / Western blot）：**在细胞系或临床样本中检测 hub 基因的 mRNA/蛋白表达，或用敲低/过表达观察细胞表型——这是证据链的“闭环”。如实说明：湿实验需要实验室条件，不是所有生信学习者都能完成，它是**加分项**而非必需项。
3. **关于纯生信论文的审稿趋势，必须诚实：**早期（约 2015–2020 年）“公共数据库挖掘 + 生信分析”就能发表不错的期刊；近年来 Frontiers in Genetics、BMC Cancer 等期刊仍会接收挖掘类论文，但审稿人普遍要求至少有一份独立验证队列，越来越多期刊要求补充实验验证（哪怕简单的 qPCR）。高分期刊（如 Cancer Cell、Gut 等）几乎不可能接受纯挖掘论文。对新手而言，合理路径是：公共数据挖掘（本教程全部内容）→独立验证（GSE14520 + HPA）→有条件再补实验→选择适合的期刊投稿。不要相信“保证发表”的承诺，发表难度在逐年上升。

本章小结

- 独立验证是区分 “过拟合” 与 “真信号” 的关键，也是审稿的基本要求；
- GSE14520 可做表达与生存的独立验证；风险模型验证时系数必须沿用训练集、不得重新拟合；
- HPA 提供蛋白水平的 IHC 证据，截图引用需注明来源与访问日期；
- CIBERSORT / ssGSEA 可估算免疫浸润，焦亡与炎症免疫天然关联，属进阶内容；
- 临床特征关联是可选描述性分析，解读需克制；
- 单细胞与湿实验是拓展方向：实验验证是加分项；纯生信挖掘论文的门槛逐年提高，独立验证是底线。

思考题

1. 为什么 “在验证集上重新拟合系数再评估” 会让独立验证失效？请举一个具体例子。
2. HPA 染色显示蛋白高表达，而 TCGA 的 mRNA 显示低表达，可能的原因有哪些？（提示：翻译调控、样本异质性）
3. 如果你的风险模型在 GSE14520 中 AUC 只有 0.55，你会从哪些方向排查？（提示：基因重叠、系数、分组方式）
4. 一篇只做数据挖掘、没有任何独立验证的论文，放在今天会收到怎样的审稿反馈？

第 13 章论文写作与发表

分析做完了、图有了，但“论文”才是成果的最终形态。这一章把“从结果到可投稿论文”的每一步拆开讲透：结构、写作、图表、投稿、审稿回复。

本章目标

1. 掌握 IMRaD 结构与“生信挖掘”论文的写作模板；
2. 学会把分析结果组织成图表（Figure 1-5）并写图注；
3. 了解期刊选择与投稿流程（含 cover letter）；
4. 掌握审稿意见回复的基本原则，避开常见拒稿理由。

13.1 论文结构：IMRaD 与“生信挖掘”论文的对应

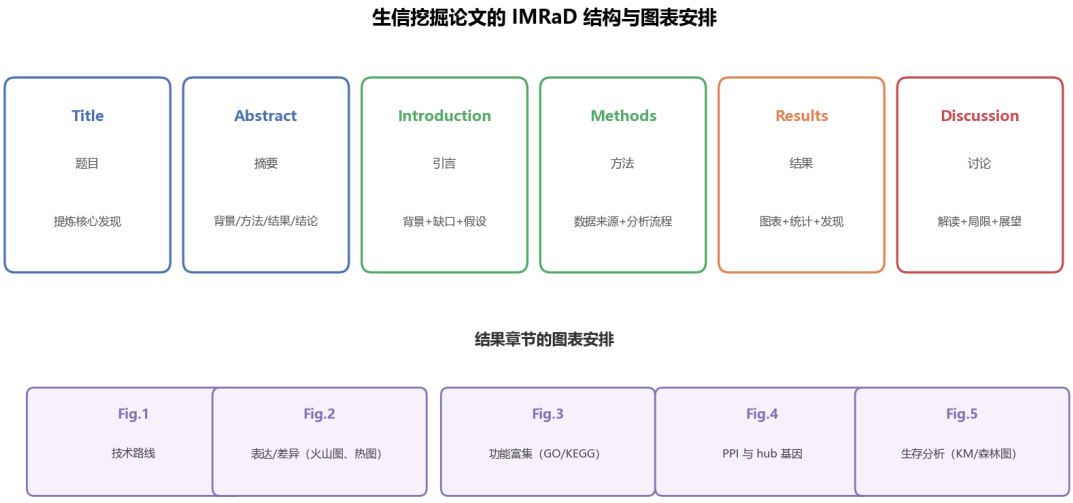


图 18: 图 15 论文结构与图表安排（示意图）

几乎所有生物医学论文都遵循 IMRaD：**I**ntroduction（引言）、**M**ethods（方法）、**R**esults（结果）、**a**nd **D**iscussion（讨论），外加 Title/Abstract。生信挖掘论文的每部分装什么：

部分	字数参考	装什么
Title 题目	1 句	“Identification and validation of pyroptosis-related genes … in hepatocellular carcinoma”
Abstract 摘要	200-300 词	背景 1 句→方法 3-4 句→结果 4-5 句→结论 1-2 句
Introduction 引言	500-800 词	疾病负担→焦亡机制与研究热度→现有缺口→本研究假设与工作
Methods 方法	600-1000 词	数据来源（accession!）→每步分析工具 + 参数→统计方法
Results 结果	1000-1500 词	按 Figure 顺序组织：每段 = 一个图 + 一个发现 + 一个统计数字
Discussion 讨论	700-1000 词	主要发现重述→与文献对比→机制解释（推测要标注）→局限→展望
Conclusion 结论	2-3 句	一句话收束

13.2 论文图表：你的 Figure 就是你的论文

13.2.1 生信挖掘论文的经典图表配置

图号	内容	对应教程章节	分析工具
Figure 1	技术路线图 + 研究设计	04	绘图软件/PPT
Figure 2	差异表达：火山图 + 热图 + PCA	07、08	ggplot2/pheatmap
Figure 3	功能富集：GO/KEGG 气泡图 + GSEA	09	clusterProfiler
Figure 4	PPI 网络 + hub 基因	10	Cytoscape
Figure 5	生存分析：KM + 森林图 + ROC + 诺模图	11	survminer/rms
Figure 6（加分）	独立验证 + 免疫浸润	12	ggplot2/CIBERSORT
Table 1	患者临床特征表	11	R 表格
Table 2	多因素 Cox 回归结果	11	R 表格

13.2.2 图表要求（投稿前自查）

- **每张图必须有独立信息**：删掉任何一张图都不完整，才说明每张都有用；
- **图注（legend）规范**：说明统计方法、样本量（n=）、误差棒含义、显著标记（*P<0.05, **P<0.01）；
- **分辨率**：位图 300 dpi（TIFF/PNG），矢量图（PDF/SVG）更佳；字号不小于 6pt；
- **配色**：色盲友好（避免红绿对比，用蓝-橙），本教程图即采用蓝/橙主色调；
- **坐标轴与单位**：完整标注，不要截断坐标轴误导读者。

本教程的示例图都是教学演示（模拟数据）。你的论文图要用真实数据重新生成，并在 `figures/` 里按 Figure 1-6 编号归档，配 `analyses/` 里的生成脚本——**图可复现**是投稿加分项。

13.3 逐部分写作指南

13.3.1 Title 与 Abstract

Title 模板（新手直接套）：

> Identification and validation of [process]-related genes as prognostic biomarkers in [cancer] based on public databases

示例：*Identification and validation of pyroptosis-related genes as prognostic biomarkers in hepatocellular carcinoma based on public databases*

Abstract 四段式（300 词内）：- Background：1-2 句（疾病 + 焦亡背景 + 为什么重要）；- Methods：数据来源（TCGA-LIHC、GSE14520）+ 分析方法链（一句话）；- Results：最关键的 4-5 个数字（多少个差异基因、几个 hub 基因、HR、AUC）；- Conclusion：本研究构建了 XX 预后模型，为 XX 提供潜在生物标志物与治疗靶点。

13.3.2 Introduction（漏斗结构）

宽：HCC 是全球高发恶性肿瘤，预后差……（引用流行病学数据）
窄：焦亡是一种炎症性程序性细胞死亡，近年发现其在肿瘤发生发展中扮演双重角色……
更窄：然而，焦亡相关基因在 HCC 中的综合表达模式与预后价值尚缺乏系统分析
最窄：因此，本研究基于 TCGA 与 GEO 公共数据，系统评估焦亡相关基因在 HCC 中的表达、功能与预后价值，并构建预后风险模型。

技巧：每句都有文献支撑（[1][2][3]）；最后一段明确写出 “本研究首次/系统性地……”（要诚实，不要过度宣称）。

13.3.3 Methods（审稿人最先看的部分！）

可复现是硬标准，逐条写清：

1. **数据获取**：TCGA-LIHC（GDC/UCSC Xena，下载日期，版本）；GSE14520（GEO accession、平台 GPL）；样本数量（371 tumor + 50 normal；验证队列 N 例）；
2. **差异表达**：limma 包（版本）；归一化方法；筛选阈值 |log2FC|>1、adj.P<0.05（BH）；
3. **富集**：clusterProfiler；GO（BP/CC/MF）+ KEGG（hsa）；GSEA（排序指标、FDR 阈值）；
4. **PPI**：STRING v11.5（confidence 0.4）；Cytoscape 3.x + cytoHubba（MCC）；
5. **生存**：中位数分组；KM+log-rank；单/多因素 Cox（校正变量列出）；LASSO（glmnet，10 折交叉验证）；
6. **模型评估**：timeROC 1/3/5 年 AUC；
7. **验证**：GSE14520 重复关键分析；
8. **统计软件**：R 4.x + 包版本列表（sessionInfo() 附录）。

13.3.4 Results（按图组织，一段一图一发现）

写作公式：**“我们用 XX 方法分析了 XX 数据（Figure X）。结果显示……（数字），这表明……（一句解读）”**

示例段（配 Figure 2）：

> 我们首先比较了焦亡相关基因在 HCC 肿瘤与癌旁组织中的表达差异（Figure 2A）。结果显示，共有 14 个焦亡相关基因显著差异表达，其中 11 个上调、3 个下调（|log2FC|>1, adj.P<0.05）。热图与 PCA 分析显示肿瘤与癌旁样本可清晰分离（Figure 2B, 2C）。

纪律：结果只陈述“是什么”，不解释“为什么”（那是讨论的事）；数字必须与你的分析输出一致（写完用结果表核对一遍）。

13.3.5 Discussion（四层结构）

1. **重述主要发现**（1 段，不要复制 Abstract）；
2. **与文献对话**：我们的结果与 XX 研究一致/不一致，可能原因；
3. **机制解释（标注推测）**：如“GSDMD 高表达可能通过促进炎症微环境重塑影响预后（推测，需实验验证）”；
4. **局限性与展望**（必写，审稿人加分项）：
   - 纯生信、无实验验证；
   - TCGA/GEO 队列的固有偏倚（种族、治疗史不明）；
   - 风险模型的临床应用需前瞻性验证；
   - 未来方向：单细胞测序、功能实验、多中心队列。

13.4 期刊选择与投稿流程

13.4.1 怎么选期刊

生信挖掘类论文的可选期刊（如实介绍，具体以投稿时官网要求为准）：

期刊类型	例子（均为真实存在的期刊）	特点
生信友好 SCI	Frontiers in Genetics、Frontiers in Oncology、BMC Cancer、PeerJ、Journal of Cancer、Aging、Cancer Cell International	接受纯生信但越来越多要求独立验证
综合性期刊	Scientific Reports、PLOS ONE	门槛看方法严谨性
中文核心	中华肿瘤防治杂志、中国细胞生物学学报、生物信息学等（按年份以官网目录为准）	中文写作，适合毕业要求

选择步骤：1. 用你论文的 Title 去 PubMed 搜 3-5 篇**最相似**的文章，看它们发在哪；2. 打开目标期刊官网：确认 **scope（范围）**、审稿周期、是否收纯生信、版面费（APC）；3. 看 “作者指南”（Author Guidelines）的字数/图表/参考文献格式要求；4. 备选 3 个期刊，按顺序投（很多期刊拒稿后可转投，但**禁止一稿多投**）。

13.4.2 投稿流程（以多数在线投稿系统为例）

准备投稿材料 → 注册期刊投稿系统 → 上传文件（Manuscript + Figures + Tables + Cover letter）
→ 编辑初审（desk reject 高风险期，约 1-2 周）→ 送外审（1-3 个月）
→ 大修/小修（major/minor revision）→ 修回 → 接收（accept）→ 校样（proof）→ 发表

Cover letter 模板（300 词内）：

> Dear Editor, We submit our manuscript entitled “……” for consideration in [Journal]. [1 段：研究背景与问题] [1 段：我们的发现与意义] [1 句：本文为原创，未一稿多投，无利益冲突，所有作者同意投稿] Sincerely, …

13.4.3 审稿意见回复（Rebuttal）黄金法则

1. **逐条回复**（Response to Reviewers）：把每条意见复制→回”Thank you for this comment.”→说明修改内容（标出正文修改位置，如 “We have revised the Discussion, Page 8, Lines 245-251”）；

2. **能改就改**：补分析（如审稿人要求独立验证——你已经有了）、补文献、改措辞；
3. **改不了就解释**：如 “我们没有实验条件，已在 Discussion 局限中明确说明”（态度诚恳通常可接受）；
4. **语气永远礼貌**，即使意见不合理——回复审稿人也是在回复潜在读者；
5. 修回时附**修改标记版**（tracked changes 或高亮）与**清洁版**。

13.5 常见拒稿原因（提前自查）

拒稿原因	如何避免（本教程对应章节）
无新意/复制他人	选题加增量（04）；文献调研充分（03）
无独立验证	第 12 章 GSE14520 验证是标配
统计方法错误/未校正	第 02、07 章（BH 校正等）
图质量差/图注不全	第 08 章 + 13.2
方法不可复现	13.3.3（accession、版本、参数全写）
过度解读/无局限	13.3.5 局限必写
语言问题（英文期刊）	润色（可请人/工具），投稿前通读

本章小结

- 生信论文 = IMRaD 结构 + Figure 1-6 图表体系 + 可复现 Methods；
- Results 一段一图一发现，Discussion 区分结果与推测、必写局限；
- 选刊先搜相似文章，确认 scope 与是否收纯生信；备选 3 刊依次投；
- 独立验证 + 方法可复现 + 局限诚实，是减少拒稿的三张保险牌；
- 投稿是 “被拒-修改-再投” 的常态，把审稿意见当免费同行评审。

思考题

1. 用你自己的课题，写出一段符合 “漏斗结构” 的 Introduction（150 词）；
2. 为你的 Figure 5（生存分析图）写一段规范的图注（含统计方法、n、显著标记说明）；
3. 找出你课题最大的 3 个局限性，写成 Discussion 的”Limitations” 段落；
4. 去 PubMed 找 2 篇与你课题最相似的文章，判断它们投在哪个期刊、为何被接受。

下一站：写完初稿别急着投——先做一轮自查（第 14 章常见问题），并对照本教程的 “自查清单” 逐项核对，然后自信地点击”Submit”。

第 14 章常见问题与排错

本章覆盖从装环境到投稿最常遇到的 9 类问题，每条按 “问题→原因→解决” 展开。遇到报错时，请先读报错信息的第一行与最后几行，再对号入座。

本章目标

学完本章，你将能够：

1. 独立解决 R 包安装、GEO 数据下载等环境类问题；
2. 看懂表达矩阵、差异分析、绘图中最常见的报错并定位原因；
3. 面对 “生存分析不显著”“富集结果为空” 等结果类问题，给出规范、诚实的处理方案；
4. 掌握内存管理与数据降采样的基本技巧；
5. 知道如何回应审稿人对 “缺少实验验证” 的质疑，守住科研诚信底线。

14.1 环境与安装

问题 1：安装包失败

问题： `install.packages()` 或 `BiocManager::install()` 报错、装不上包。

原因： 最常见有三类——①网络问题：CRAN/Bioconductor 服务器在境外，校园网或公司网络常常很慢甚至连不上；②包来源搞错：Bioconductor 的包用 `install.packages()` 装，会提示 “package ‘limma’ is not available for this version of R”；③依赖缺失：报错尾部写着 “ERROR: dependency ‘xxx’ is not available”；④ R 版本过旧：新包普遍要求 R ≥ 4.x。

解决：

1. 先分清包属于哪个仓库：CRAN 包用 `install.packages()`；Bioconductor 包先 `install.packages("BiocManager")`，再 `BiocManager::install("limma")`；GitHub 包用 `remotes::install_github("用户名/仓库名")`。
2. 换国内镜像加速：`install.packages("ggplot2", repos = "https://mirrors.tuna.tsinghua.edu.cn/CRAN/")`，或写入 `options(repos = c(CRAN = "https://mirrors.tuna.tsinghua.edu.cn/CRAN/"))` 一劳永逸。
3. 缺依赖就按报错顺序先装依赖；提示 “需要更高版本 R” 就升级 R 到当前稳定版（≥ 4.3）。
4. Windows 下遇到编译类报错（如 igraph、Matrix 提示需要本地编译），通常是没装 Rtools；

到 CRAN 下载对应版本安装后重启 R。优先使用官方预编译的二进制包可避开多数编译问题。

5. 实在装不上不必死磕：先把报错最后几行复制到搜索引擎，多数问题都有现成答案；少数包可用功能相近的替代包。

问题 2：GEOquery 下载慢/失败

问题： `getGEO("GSE14520")` 转圈半天、报 "Timeout of 60 seconds was reached"，或下载到一半断掉。

原因： GEO 服务器位于美国，跨洋网络不稳定；而 GSE 的 series matrix 文件可能有几十 MB（GSE14520 含 400+ 样本），R 默认约 60 秒的超时很容易不够用。

解决：

1. 加大超时：`options(timeout = 600)`，下载大文件前必改。
2. 缓存到本地：`getGEO("GSE14520", destdir = "data/")`，文件会存到 `data/`，之后重复读取不再联网。
3. 失败就重试，或换个时段；网络抖动是常态，不是你的代码问题。
4. 终极办法：用浏览器打开 GEO 官网的 GSE 页面，手动下载 series matrix 压缩包放到 `data/`，再 `getGEO(filename = "data/GSE14520_series_matrix.txt.gz")` 离线读取——把 "下载" 和 "分析" 分离后，网络问题就不再阻塞分析。
5. TCGA 数据同理：文件更大，建议用 TCGAbiolinks 的 query/download 接口分步下载，并确认断点续传。

14.2 数据与差异分析

问题 3：探针 ID 无法映射到基因

问题： 表达矩阵行名是 `202763_at` 这类探针 ID，与基因符号对不上；或一个基因对应多条探针，不知该取哪个。

原因： 芯片平台的探针与基因是多对多关系（一条探针可能命中多个转录本，一个基因被多条探针覆盖），映射关系存放在平台（GPL）对应的注释包中；平台错配是 "映射结果几乎全 NA" 的头号原因。

解决：

1. 从 series matrix 里找到`!Series_platform_id`，确定 GPL 编号（如 GPL3921 = Affymetrix HG-U133A 芯片），安装对应注释包：`BiocManager::install("hgu133a.db")`。
2. 用 `AnnotationDbi::select(hgu133a.db, keys = 探针 ID, columns = "SYMBOL", keytype = "PROBEID")` 得到探针→基因符号的映射。

3. 多探针对一基因：按”collapseByGene” 的思想聚合——教学上常用 `aggregate()` 按基因符号取均值（也可取最大信号），先把 NA 过滤掉再聚合。
4. RNA-seq 数据没有探针问题，但若行名是 Ensembl ID，仍需转换：用 `clusterProfiler::bitr()` 或 biomaRt 转成 gene symbol。

问题 4：差异分析报错

问题：limma 报 “contrasts can be applied only to factors”、结果全是 NA，或倍数变化方向与预期相反。

原因与解决：

1. **分组方向反了：**limma 按因子水平的顺序比较，第一个水平是参照组。用 `factor(group, levels = c(”normal”, ”tumor”))` 明确指定，再看 `colnames(fit)` 确认比较方向后再下结论。
2. **表达值不是数值：**读入的矩阵可能被当成 character（逗号、引号处理不当）。用 `str()` 检查，`as.matrix()` 后确认 `mode()` 为 “numeric”。
3. **存在 NA：**limma 不接受 NA。先 `na.omit()` 或补全，再过滤掉全零、全低表达的行。
4. **组内样本太少：**每组至少 3 个样本，limma 才能稳健估计方差；某组只有 1 个样本时先合并分组或放弃该比较。
5. **counts 数据不能直接喂 limma：**RNA-seq 原始计数需先 `edgeR::DGEList()` → `calcNormFactors()` → `voom()` 转换，再进 limma；或直接改用 edgeR/DESeq2。
6. 报错别慌：把 design 公式、contrasts 定义和 `str()` 输出贴给搜索引擎，绝大多数是上述五类之一。

14.3 可视化

问题 5：热图/火山图画不出来

问题：pheatmap 报错、火山图一片空白、ggplot 报 “Aesthetics must be either length 1 or the same as the data”。

原因与解决：

1. pheatmap 要求输入为纯数值矩阵：传 `as.matrix(expr)`，行名是基因、列名是样本，不要带其他列；行名重复会报错，先 `make.unique()` 或按基因聚合。
2. 火山图数据要有 logFC、adj.P、gene 三列，先 `str()` 确认列名与类型；若取 `-log10(adj.P)` 时 P 值全为 NA（limma 对某些行无法估计），图自然空白——先过滤 NA 再画。
3. ggplot 常见坑：列名大小写写错、把外部向量直接塞进 `aes()`（aes 里只能写数据框的列名）、scale 函数参数拼错。逐行检查映射即可。
4. 中文显示：默认绘图设备对中文支持差，会报字体警告或显示方块。图内统一用英文标签，

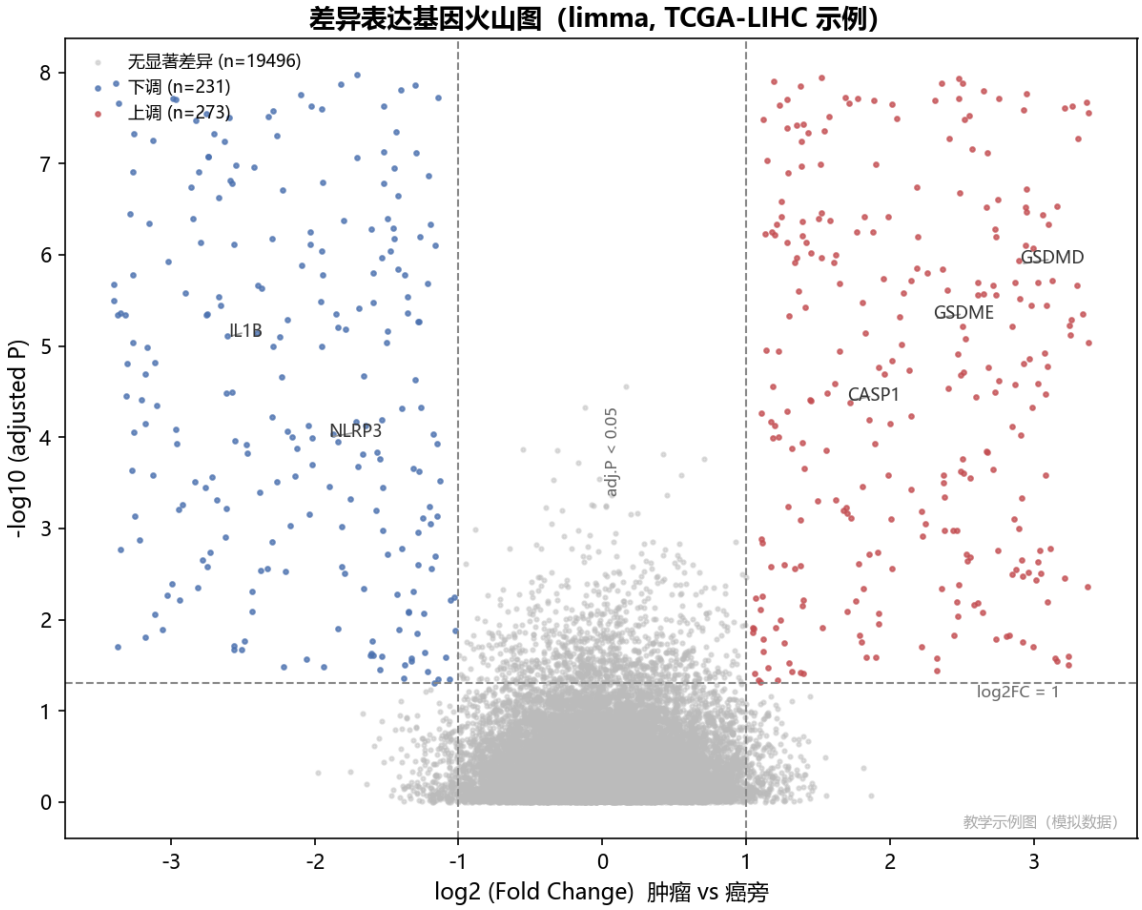


图 19: 图 06 差异表达火山图（示例，模拟数据）

或加载 showtext 包配中文字体。

5. 调试技巧：先画最简散点图确认数据没问题，再逐步加颜色、标签、主题；报错信息最后几行是定位关键。

14.4 结果解读

问题 6：生存分析结果“不显著”怎么办

问题：KM 曲线 $P > 0.05$，看起来“没戏了”。

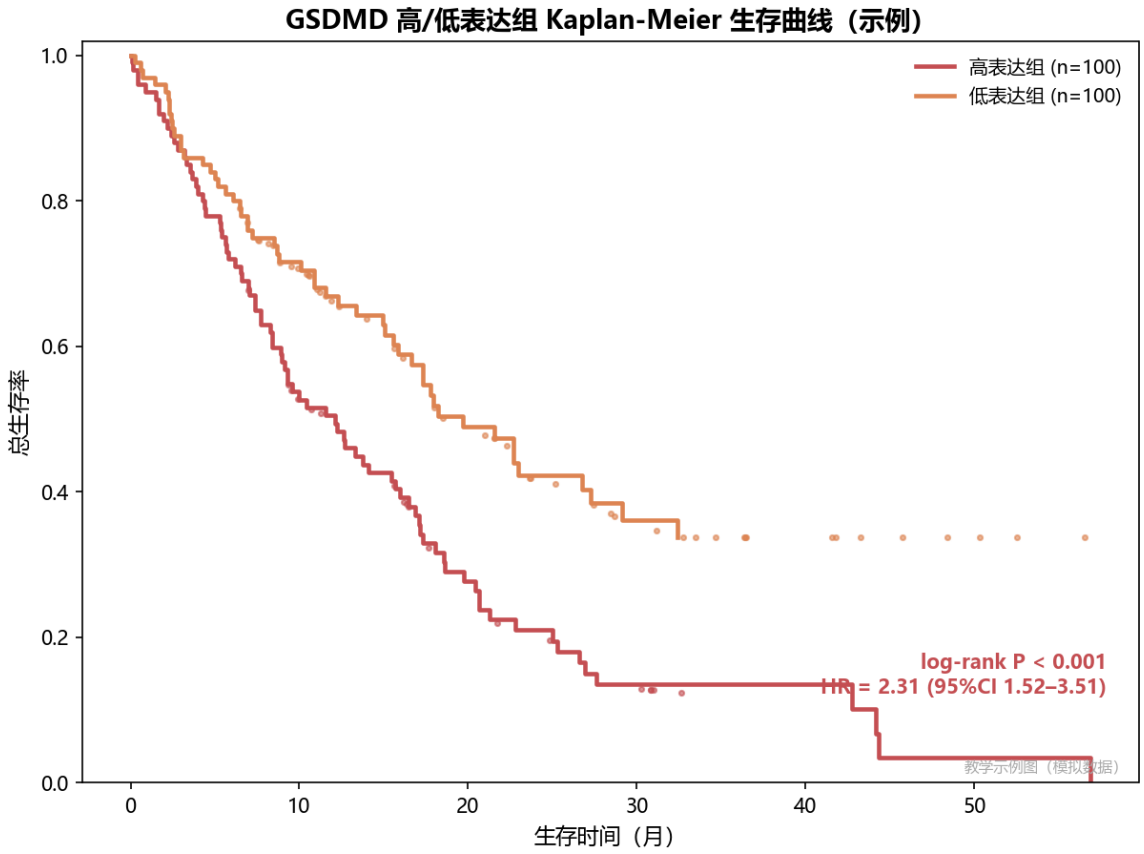


图 20: 图 12 KM 生存曲线（示例，模拟数据）

原因：样本量小、事件数少（删失比例过高时检验功效不足）、分组方式不佳，或该基因在本队列中确实与预后无关。

规范应对：

1. 先查硬条件：报告事件数与删失比例；事件太少时即使真有关也检不出。
2. 再查分组方式：中位数分组只是默认做法，可用 `survminer::surv_cutpoint()` 按数据自动找最佳截点，并在 Methods 里写清楚。
3. 换一个视角：用连续型 Cox 回归看 HR（风险比），比二分法更充分利用数据。
4. **警惕 p-hacking：**不要在几十个截点里挑 P 值最小的那个当“发现”而不做任何校正；也

不要在检验过的 20 个基因里只报显著的那几个。规范做法是预先定好截点规则、报告检验过的基因总数，把阴性结果如实写入正文或补充材料。

5. 不显著也是结论：写“该基因在 TCGA-LIHC 队列中与总生存无显著关联（$P = 0.3$，log-rank 检验）”本身就是诚实的科学内容。

问题 7：富集分析结果为空/通路太少

问题：enrichGO/enrichKEGG 报 “No gene can be mapped”，或返回 0 条通路、通路少得可怜。

原因：最常见是基因 ID 类型不匹配——clusterProfiler 默认要求 ENTREZID，而你输入的是 gene symbol；其次是差异基因太少、阈值过严；此外 KEGG 收录的通路偏经典代谢与信号通路，免疫相关基因常富集在 GO 条目而 KEGG 通路少，这并不代表分析错了。

解决：

1. 先转 ID：`bitr(genes, fromType = "SYMBOL", toType = "ENTREZID", OrgDb = org.Hs.eg.db)`，再用转换结果做 `enrichGO(OrgDb = org.Hs.eg.db, keyType = "ENTREZID")`。
2. 检查物种：org.Hs.eg.db 是人类注释包，别拿它映射小鼠基因；enrichKEGG 记得设 `organism = "hsa"`。
3. 阈值放宽：差异基因只有几十个时，可降低 |log2FC|（如 0.5）或只用未校正 P 值筛选，并在论文中说明；或改用 GSEA——它不需要先筛差异基因，用全部基因的表达变化排序做富集，结果更稳健。
4. GO 与 KEGG 分开报告，KEGG 通路少不代表分析失败。

14.5 性能与诚信

问题 8：内存不足/电脑卡顿

问题：RStudio 卡死，报 “cannot allocate vector of size …”。

原因：表达矩阵本身不大（371 样本× 2 万基因仅几十 MB），但中间对象（DGEList、voom 结果、绘图对象）层层复制，叠加浏览器占用就容易撑爆内存。

解决：

1. 对象瘦身：只保留需要的基因（先取焦亡基因子集再做后续分析）；用完的中间变量及时 `rm()` 并 `gc()`。
2. 分块运行：下载→存 RDS，差异分析→存 RDS，绘图时再读 RDS；重启 R 后从断点继续，而不是一个脚本从头跑到尾。
3. 数据降采样：TCGA 用 TCGAbiolinks 只取 LIHC 的 RNA-seq 与临床，不必全库下载；GEO 用 destdir 缓存避免重复下载。

4. 关掉不用的浏览器标签；矩阵类运算（如免疫浸润反卷积）实在跑不动时，可考虑学校集群或云服务器，R 脚本迁移成本很低。
5. Windows 上通常不需要手动调内存参数，问题几乎总在 “对象太多”，而不是 R 本身的限制。

问题 9：审稿人质疑 “没有实验验证”

问题：“你只有生信分析，没有实验验证，结论可靠吗？”

回应思路：不是硬杠，而是补证据 + 说实话。

1. **独立验证**：用第二个公共队列（如 GSE14520）复现 TCGA-LIHC 的预后模型——这正是本教程第 12 章做的事，也是挖掘型论文最重要的证据。
2. **多层面交叉验证**：转录组（TCGA/GEO）＋蛋白层面（HPA 免疫组化）＋免疫浸润（CIBERSORT），多个独立数据源互相印证。
3. **机制讨论**：结合已发表的焦亡机制文献，解释 hub 基因在通路中的位置，把统计关联翻译成生物学解释（明确标注为推测）。
4. **如实写 limitations**：在 Discussion 末尾写明 “本研究为基于公共数据库的发现性研究（hypothesis-generating），结论有待 qPCR/WB 等湿实验验证”，并把湿实验列为未来工作——不承诺未做的实验。
5. 审稿人要的是 “你知道自己结论的边界”，态度诚恳、方法透明通常就能过关。

本章小结

- 报错先看第一行和最后几行，把原文复制到搜索引擎，比自己瞎猜高效得多。
- 环境问题三件套：分清包来源（CRAN/Bioconductor/GitHub）、换国内镜像、按依赖顺序安装。
- 数据问题的核心是 “类型与格式”：先 `str()` 看结构，再谈分析。
- 结果类问题先查统计前提（事件数、样本量、ID 类型），再考虑换方法；不显著、富集为空都要如实报告。
- 科研诚信是底线：不挑 P 值、不隐瞒阴性结果、不虚构实验。

思考题

1. 你用 `install.packages("limma")` 报错 “package ‘limma’ is not available”，而同学用 `BiocManager::install("limma")` 装上了——问题出在哪？
2. 火山图横坐标全为 0，可能是什么数据问题？你会按什么顺序排查？
3. 你的 KM 曲线 $P = 0.08$。审稿人可能怎么看？你会在论文中如何报告这个结果？
4. 富集分析返回 0 条通路，列出至少三种可能原因及对应的检查方法。

第 15 章附录：术语表与资源

本章目标

1. 遇到陌生术语时能快速查表，看懂论文与教程中的概念；
2. 知道官方文档、教材与社区资源的去处，能自主查漏补缺；
3. 理解本教程所有图的属性，知道如何用自己的数据复现。

15.1 术语表

下表按分析流程顺序排列，覆盖本教程出现的核心术语；缩写均给出英文全称。

中文术语	英文	一句话解释
转录组	transcriptome	细胞或组织在特定状态下全部 RNA 的总和，反映“哪些基因在表达、表达多少”。
差异表达分析	differential expression analysis (DEA)	比较两组（如肿瘤 vs 癌旁）基因表达水平差异是否具有统计学意义。
log2 倍数变化	log2 fold change (log2FC)	两组表达均值之比取以 2 为底的对数：log2FC = 1 表示上调 2 倍，-1 表示下调一半。
校正后 P 值	adjusted P value (adj.P)	经多重检验校正后的 P 值，比原始 P 值更保守，用于控制假阳性。
错误发现率	false discovery rate (FDR)	在所有“被判定为显著”的结果中，预期假阳性的比例。
BH 校正	Benjamini-Hochberg correction	最常用的 FDR 控制方法，本教程筛选差异基因统一采用（adj.P $<$ 0.05）。

中文术语	英文	一句话解释
limma	Linear Models for Microarray Data	Bioconductor 经典差异分析包，线性模型 + 经验贝叶斯估计方差；RNA-seq 数据配合 voom 使用。
富集分析	enrichment analysis	检验一组基因（如差异基因）是否在某个功能/通路类别中异常集中。
基因本体	Gene Ontology (GO)	标准化的基因功能注释体系，分生物学过程（BP）、细胞组分（CC）、分子功能（MF）三类。
KEGG	Kyoto Encyclopedia of Genes and Genomes	收录代谢与信号通路的数据库，enrichKEGG 用于通路富集。
基因集富集分析	Gene Set Enrichment Analysis (GSEA)	不设阈值，用全部基因按表达变化排序，检验某基因集是否整体富集。
蛋白-蛋白相互作用	protein-protein interaction (PPI)	蛋白间直接或间接结合的相互作用关系网络。
hub 基因	hub gene	PPI 网络中连接度最高、处于网络中心位置的关键基因。
STRING	STRING database	在线 PPI 数据库，输入基因列表即可构建相互作用网络（string-db.org）。
Cytoscape	Cytoscape	可视化与分析生物网络的桌面软件，常用于美化 PPI 图。
生存曲线	Kaplan-Meier (KM) curve	以生存概率对时间作图，展示高/低表达组随时间存活差异的曲线。
log-rank 检验	log-rank test	比较两条 KM 曲线差异是否显著的非参数检验。
Cox 回归	Cox proportional hazards regression	生存分析最常用的回归模型，可纳入多个协变量并输出风险比（HR）。

中文术语	英文	一句话解释
风险比	hazard ratio (HR)	某因素每变化一个单位（或某组相对参照组）事件风险的倍数，HR >1 表示风险升高。
LASSO	Least Absolute Shrinkage and Selection Operator	带 L1 惩罚的回归，自动筛选变量并压缩系数，常用于构建预后风险模型。
受试者工作特征曲线	receiver operating characteristic (ROC)	以假阳性率为横轴、真阳性率为纵轴，评价模型区分能力的曲线。
曲线下面积	area under the curve (AUC)	ROC 曲线下的面积，0.5 表示随机猜测，越接近 1 区分能力越强。
诺模图	nomogram	把多因素模型画成“算分表”，用总分预测个体生存概率的可视化工具。
TCGA	The Cancer Genome Atlas	大型癌症多组学数据库，含 RNA-seq、突变、临床随访（portal.gdc.cancer.gov）。
GEO	Gene Expression Omnibus	NCBI 的基因表达数据库，存储海量芯片与测序数据（ncbi.nlm.nih.gov/geo）。
GSE	GEO Series	GEO 中一个系列编号，代表一项研究/一个数据集（如 GSE14520）。
GSM	GEO Sample	GEO 中单个样本的编号。
GPL	GEO Platform	GEO 中平台的编号（如 GPL3921 = Affymetrix HG-U133A 芯片）。
FPKM	Fragments Per Kilobase per Million	每百万片段中每千碱基外显子上的片段数，RNA-seq 归一化单位（正逐渐被 TPM 取代）。
TPM	Transcripts Per Million	每百万转录本中的转录本数，RNA-seq 最常用的归一化单位，跨样本可比。

中文术语	英文	一句话解释
原始计数	counts	比对到每个基因上的原始读段数，edgeR/DESeq2 直接基于它建模。
批次效应	batch effect	因实验批次（时间、平台、操作者不同）引入的系统性表达差异，可用 limma::removeBatchEffect 等校正。
删失	censoring	生存数据中研究对象在观察期内未发生事件（仍存活或失访），其真实事件时间未知。
免疫浸润	immune infiltration	肿瘤微环境中各类免疫细胞（T 细胞、巨噬细胞等）的组成与丰度。
CIBERSORT	CIBERSORT	通过反卷积算法从 bulk 转录组估计 22 种免疫细胞相对比例的工具。
定量 PCR	quantitative PCR (qPCR)	湿实验定量基因 mRNA 表达的金标准方法，常用于验证生信发现。
蛋白印迹	Western blot (WB)	检测蛋白表达量的经典湿实验方法。
人类蛋白图谱	Human Protein Atlas (HPA)	提供蛋白在组织/细胞中表达与定位的免疫组化数据库（proteinatlas.org）。
GTEx	Genotype-Tissue Expression	正常人体多组织表达数据库，常与肿瘤数据对照（gtexportal.org）。
cBioPortal	cBioPortal	交互式探索 TCGA 等肿瘤组学数据的门户（cbioportal.org）。
UCSC Xena	UCSC Xena	多组学数据可视化平台，可下载整理好的 TCGA/GTEx 表达矩阵（xenabrowser.net）。

中文术语	英文	一句话解释
Bioconductor	Bioconductor	面向生信分析的 R 包官方仓库（bioconductor.org）。
BiocManager	BiocManager	安装 Bioconductor 包的 R 包，用法：`BiocManager::install("limma")`。
tidyverse	tidyverse	风格统一的 R 数据科学包集合（ggplot2、dplyr、tidyr 等），本教程数据处理多用它。
管道符	pipe operator (%>%)	把左侧结果传给右侧函数第一个参数的运算符，让代码像流水线一样从左读到右。
表达矩阵	expression matrix	行是基因、列是样本的数值矩阵，是绝大多数下游分析的输入。
注释包	annotation package	存放平台/物种注释关系（探针与基因、ID 与功能）的 Bioconductor 包，如 hgu133a.db、org.Hs.eg.db。

15.2 学习资源清单

官方资源

资源	类型	一句话说明
R 官网（r-project.org）	官方	R 语言官方发布页与手册，下载 R 与查阅文档。
RStudio / Posit（posit.co）	官方	最常用的 R 集成开发环境（IDE），下载 RStudio Desktop。
Bioconductor（bioconductor.org）	官方	生信 R 包官方仓库，含安装说明与包文档（vignette）。

资源	类型	一句话说明
TCGA GDC 门户（portal.gdc.cancer.gov）	官方	TCGA 数据官方下载门户，获取 RNA-seq 表达与临床随访数据。
GEO（ncbi.nlm.nih.gov/geo）	官方	NCBI 基因表达数据库，检索与下载 GSE 数据集。
STRING（string-db.org）	官方	PPI 网络构建与结果导出的官方站点。
HPA（proteinatlas.org）	官方	蛋白表达与定位数据库，可查 hub 基因的蛋白水平证据。
GTEx（gtexportal.org）	官方	正常组织表达数据库，可作为肿瘤数据的对照。
cBioPortal（cbioportal.org）	官方	探索 TCGA 等数据的交互式门户，适合快速查看单基因概况。
UCSC Xena（xenabrowser.net）	官方	下载整理好的 TCGA/GTEx 表达矩阵，省去自行解析的麻烦。

教材与教程

资源	类型	一句话说明
R for Data Science（r4ds.hadley.nz；中文版《R 数据科学》）	教材	Hadley Wickham 等所著，tidyverse 数据处理与可视化的经典入门书。
Bioconductor 官方工作流（bioconductor.org/help/workflows）	教程	官方维护的端到端分析流程，RNA-seq、差异表达等均有标准做法。
生信技能树（微信/知乎/B 站搜索 “生信技能树”）	社区资源	中文生信社区，提供大量新手教程与答疑；属社区资源，注意甄别时效性。
B 站生信入门系列视频	社区资源	中文视频课程，搜索 “生信入门”“R 语言基础” 可找到系列教程（泛称，建议选更新近、口碑好的）。

资源	类型	一句话说明
《R 语言实战》（R in Action 中文版）	教材	中文 R 入门书，适合零基础通读（按需选用，非必需）。

文献检索

资源	类型	一句话说明
焦亡（pyroptosis）相关综述	文献	在 PubMed（pubmed.ncbi.nlm.nih.gov）以 “pyroptosis”[Title] AND review 检索最新综述，了解焦亡机制与疾病研究进展；本教程不预设具体文献，请按关键词自行检索最新版本。

15.3 致谢与使用说明

- 本教程所有图（`figures/` 目录）均为**教学示例图**，使用模拟数据绘制，仅用于演示图形样式与解读方法，不代表任何真实研究结果。
- 用教程中的 R 脚本分析你自己的真实数据（如 TCGA-LIHC 与 GSE14520）后，得到的是**你自己的结果与图**，可直接用于论文。
- 教程中的统计阈值（$|\log2FC| > 1$、adj.P < 0.05、BH 校正）是教学默认值，正式发表前请结合领域惯例与审稿要求说明。
- 引用本教程的方法时，请注明数据来源（TCGA/GEO 的 accession）与分析工具版本，保证结果可复现。

本章小结

- 术语表按分析流程排列，遇到不懂的缩写先查表；缩写与全称、直觉解释一一对应。
- 官方资源优先，社区资源作补充；所有资源均可通过搜索引擎直达官网域名。
- 教程图是示例图，论文中的图必须来自你自己的真实分析。

思考题

1. 从术语表中挑出 5 个与 “预后模型” 直接相关的术语，说明它们之间的逻辑关系。

2. 为什么论文要求给出数据 accession（如 GSE14520）与软件版本？这与“可复现性”有什么关系？

结束语

到这里，本教程的全部内容就结束了：从细胞生物学与生信基础出发，你走过了数据库检索、差异分析、富集分析、PPI 网络、生存模型、独立验证，直到论文写作与投稿。现在，你已经具备完成一个完整生信课题的能力——去创造你的第一个 Figure 1 吧！