**生信分析报告**

**项目标题： 基于血小板RNA测序数据预测早期肺癌潜在生**   
 **物标志物 ;**

**单 号： BSXG240327 ;**

**分析人员： 黄礼闯 ;**

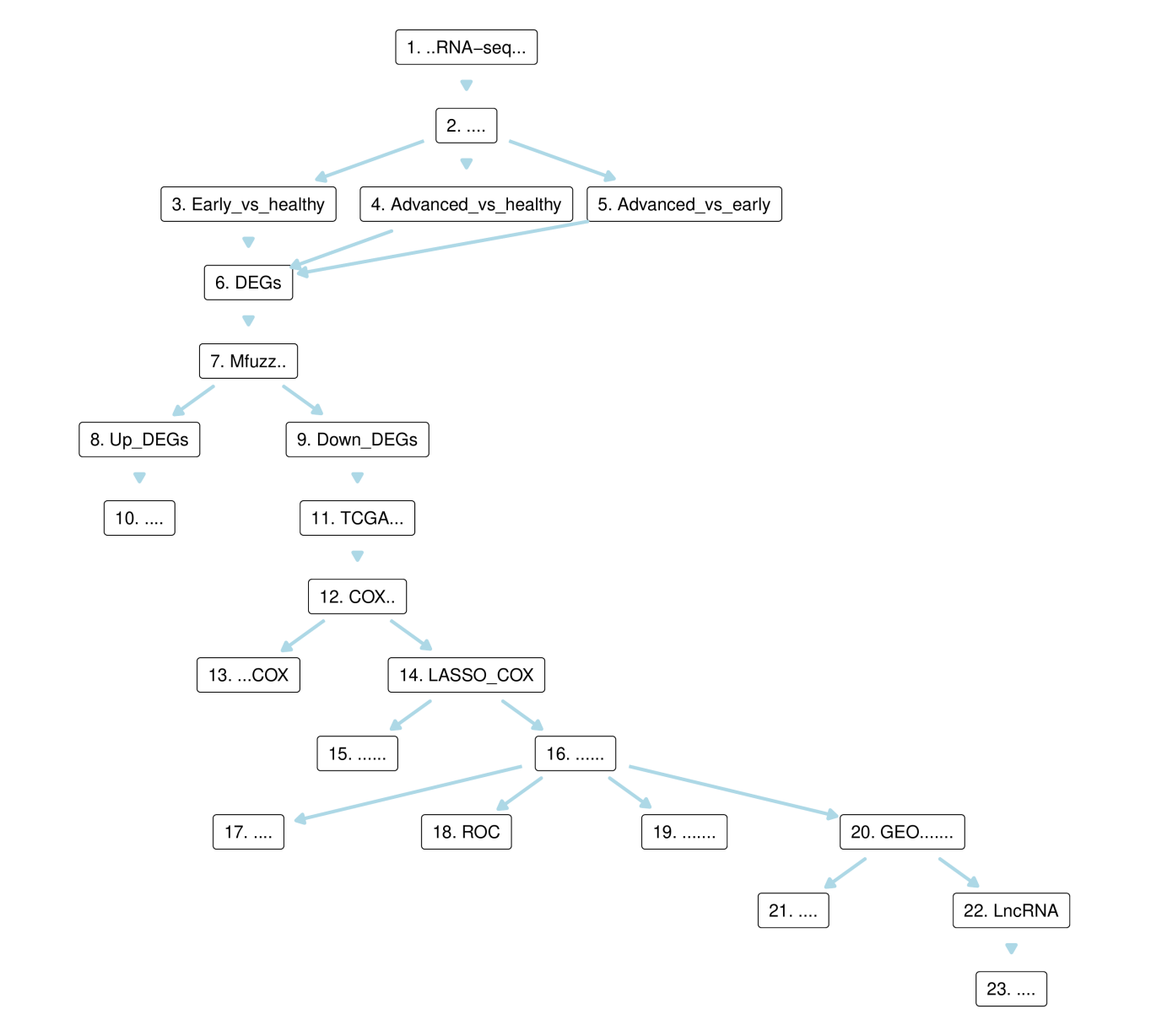
**分析类型： 补充分析 ;**

**委 托 人： 陈立茂 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程

该分析思路与 (2023, **IF:4.8**, Q1, Biomolecules)1 相似。



**Fig.** **1** Route

**(File path: Figure+Table/1.0\_分析流程\_{#abstract}/Route.pdf)**

# 2 材料和方法

## 2.1 数据分析平台

在 Linux pop-os x86\_64 (6.9.3-76060903-generic) 上，使用 R version 4.4.2 (2024-10-31) (<https://www.r-project.org/>) 对数据统计分析与整合分析。

## 2.2 Biomart 基因注释 (Dataset: ALL)

以 R 包 biomaRt (2.62.0) 对基因进行注释，获取各数据库 ID 或注释信息，以备后续分析。

## 2.3 Limma 差异分析 (Dataset: MRNA)

以 R 包 limma (3.62.1) (2005, **IF:**, , )2 edgeR (4.4.0) (, **IF:**, , )3 进行差异分析。以 edgeR::filterByExpr 过滤 count 数量小于 10 的基因。以 edgeR::calcNormFactors，limma::voom 转化 count 数据为 log2 counts-per-million (logCPM)。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。随后，以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) 用于线性分析。 使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 差异分析对比组：Early\_stage vs Healthy, Advanced\_stage vs Healthy, Advanced\_stage vs Early\_stage。以 limma::topTable 提取所有结果，并过滤得到 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

## 2.4 Mfuzz 聚类分析 (Dataset: MRNA)

以 R 包 Mfuzz (2.66.0) (, **IF:**, , )4 对基因聚类分析，设定 fuzzification 参数为 3.6261199327297 (以 Mfuzz::mestimate 预估) ，得到 10 个聚类。

## 2.5 富集分析 (Dataset: MRNA)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)5进行 KEGG 和 GO 富集分析。

## 2.6 TCGA 数据获取 (Dataset: LUSC)

以 R 包 TCGAbiolinks (2.34.0) (2015, **IF:16.6**, Q1, Nucleic Acids Research)6 获取 TCGA 数据集。

## 2.7 TCGA 数据获取 (Dataset: LUAD)

以 R 包 TCGAbiolinks (2.34.0) (2015, **IF:16.6**, Q1, Nucleic Acids Research)6 获取 TCGA-LUAD 数据集。 随后，将 TCGA-LUSC 和 TCGA-LUAD 数据集合并。

以 R 包 survival (3.7.0) 进行单因素 COX 回归 (survival::coxph)。筛选 Pr(>|z|) < .05的基因。在单因素回归得到的基因的基础上， 以 R 包glmnet(4.1.8) 作 lasso 处罚的 cox 回归，以cv.glmnet` 函数作 10 交叉验证获得模型。

筛选 AJCC Stage (ajcc\_pathologic\_stage) 为 Stage I, Stage II 的病人，并且 days\_to\_last\_follow\_up 大于 30 天，且为肿瘤组织的样本 (并去除了来自相同病人的重复样本)。

## 2.8 Survival 生存分析 (Dataset: LUNG)

将 Univariate COX 回归系数用于风险评分计算，根据中位风险评分 0.00671702707171972 将患者分为低危组和高危组。 以 R 包 survival (3.7.0) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。以 R 包 timeROC (0.4) 绘制 1, 3, 5 年生存曲线。

## 2.9 GSE 数据搜索

使用 Entrez Direct (EDirect) <https://www.ncbi.nlm.nih.gov/books/NBK3837/> 搜索 GEO 数据库 (esearch -db gds)，查询信息为: ((lung adenocarcinoma[Description] AND (survival[Description]) AND ((30:1000[Number of Samples]) AND (Homo Sapiens[Organism]) AND (GSE[Entry Type]))。

## 2.10 GEO 数据获取 (Dataset: LUAD)

以 R 包 GEOquery (2.74.0) 获取 GSE87340 数据集。

## 2.11 estimate 免疫评分 (Dataset: LUNG)

以 R 包 estimate (1.0.13) (2013, **IF:14.7**, Q1, Nature communications)7 预测数据集的 stromal, immune, estimate 得分。 从 TISIDB (2019, **IF:4.4**, Q1, Bioinformatics)8 数据库下载的 178 个基因 (genes encoding immunomodulators and chemokines) 比较表达量差异。

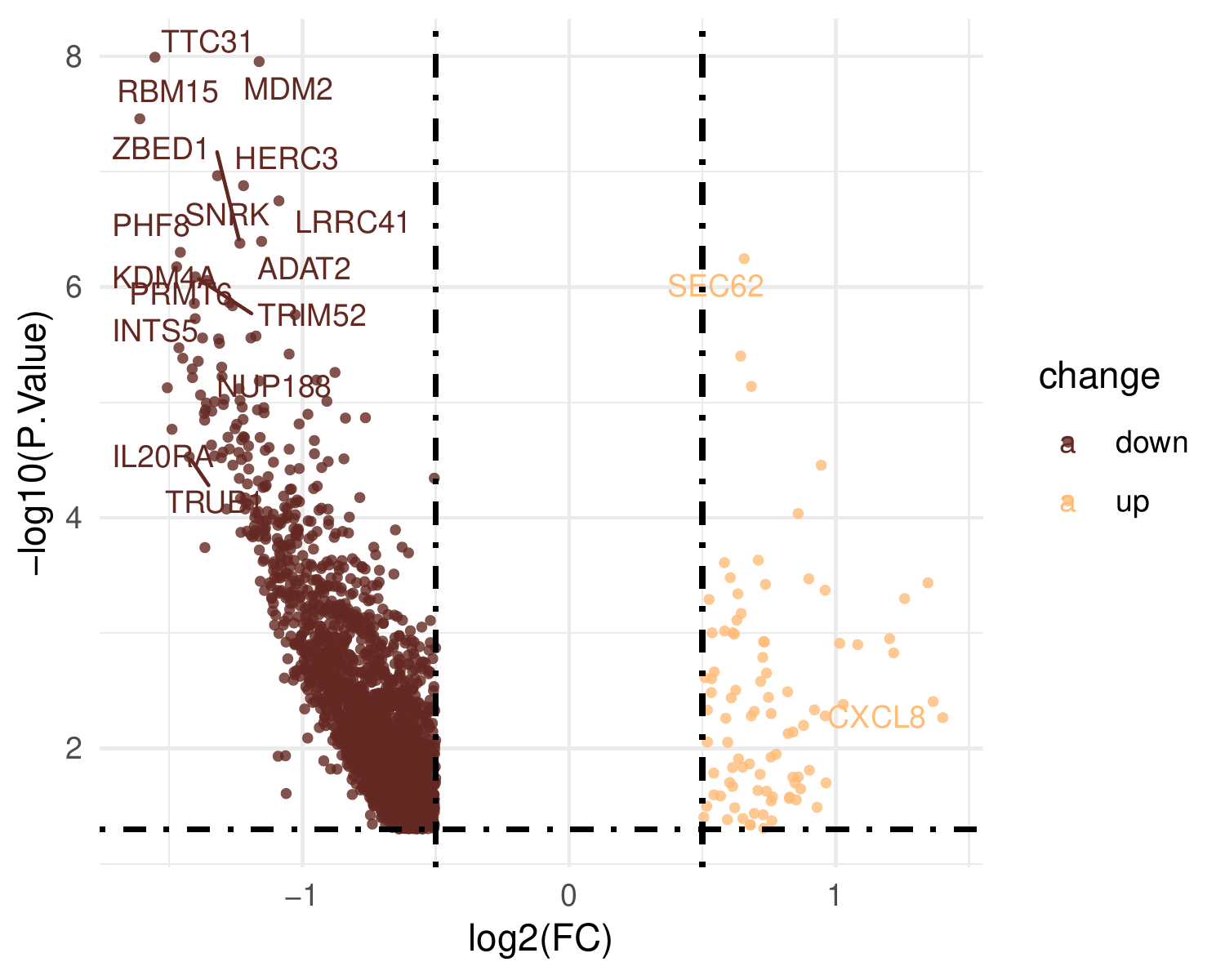
## 2.12 Limma 差异分析 (Dataset: LNCRNA)

以 R 包 limma (3.62.1) (2005, **IF:**, , )2 edgeR (4.4.0) (, **IF:**, , )3 进行差异分析。以 edgeR::filterByExpr 过滤 count 数量小于 10 的基因。以 edgeR::calcNormFactors，limma::voom 转化 count 数据为 log2 counts-per-million (logCPM)。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。随后，以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) 用于线性分析。 使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 差异分析对比组：Early\_stage vs Healthy, Advanced\_stage vs Healthy, Advanced\_stage vs Early\_stage。以 limma::topTable 提取所有结果，并过滤得到 P.Value 小于 0.05，|Log2(FC)| 大于 0.5 的统计结果。

# 3 分析结果

## 3.1 Limma 差异分析 (MRNA)

肝癌 RNA-seq， 共 247 个样本。包含 Advanced\_stage (65) , Early\_stage (101) , Healthy (81) 。 对基因注释后，获取 mRNA 数据差异分析。 差异分析 Early\_stage vs Healthy, Advanced\_stage vs Healthy, Advanced\_stage vs Early\_stage (若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B) 得到的 DEGs 统计见 Fig. **[5](#MRNA-Difference-intersection)** 所有上调 DEGs 有 2483 个，下调共 3779；一共 5497 个 (非重复)。

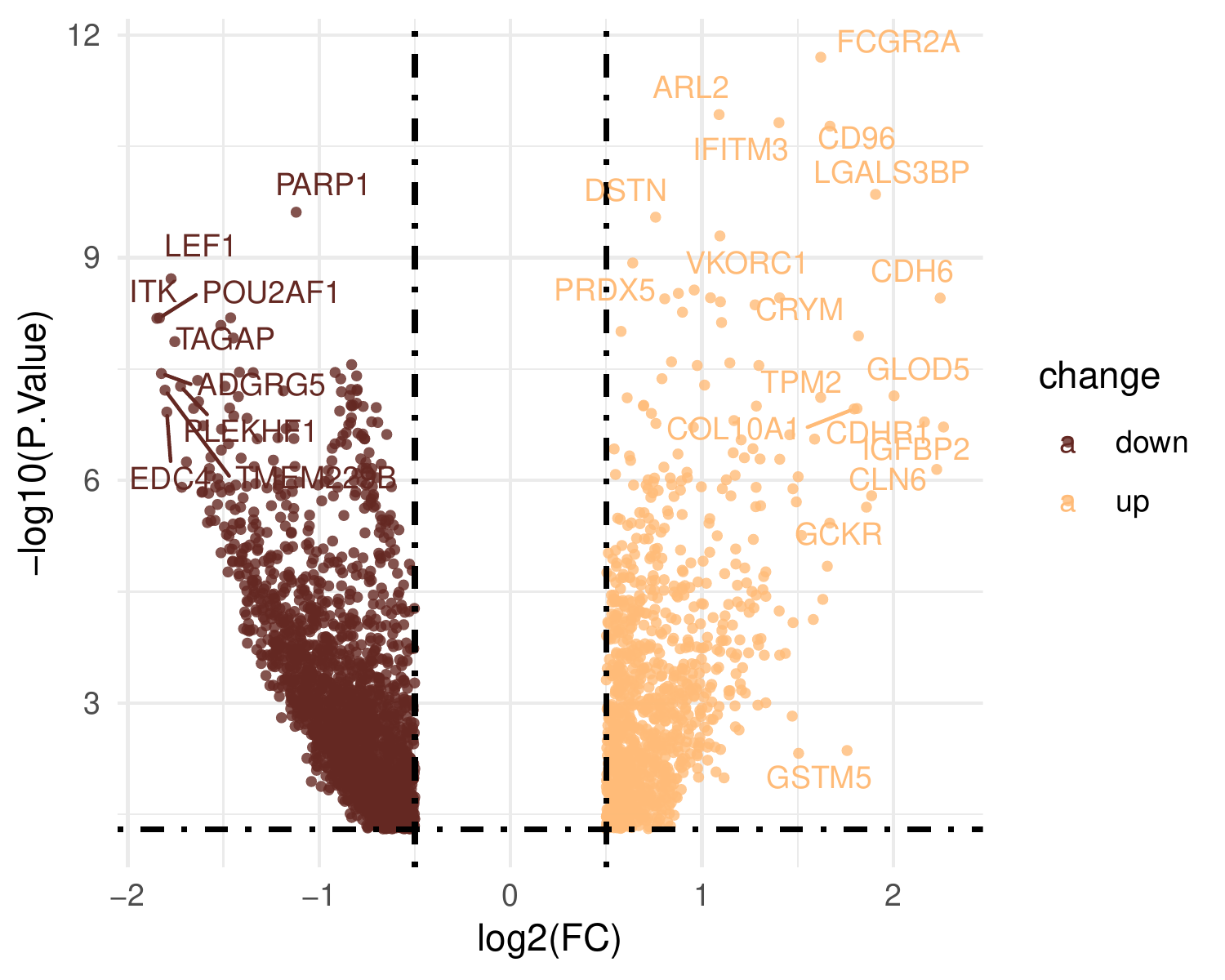


**Fig.** **2** MRNA Early stage vs Healthy

**(File path: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Early-stage-vs-Healthy.pdf)**

* P.Value cut-off: 0.05
* Log2(FC) cut-off: 0.5

**(See: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Early-stage-vs-Healthy-content)**

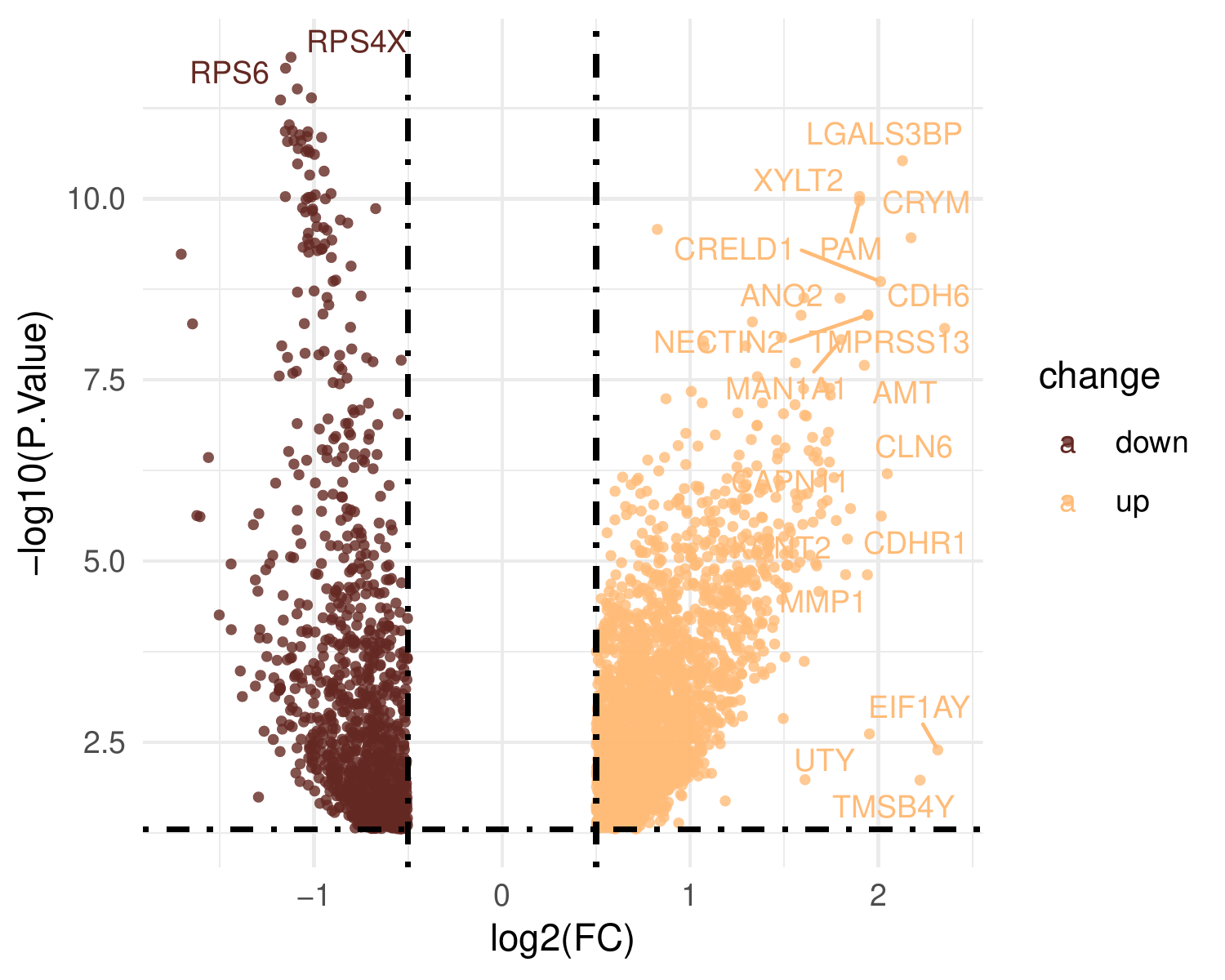


**Fig.** **3** MRNA Advanced stage vs Healthy

**(File path: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Advanced-stage-vs-Healthy.pdf)**

* P.Value cut-off: 0.05
* Log2(FC) cut-off: 0.5

**(See: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Advanced-stage-vs-Healthy-content)**

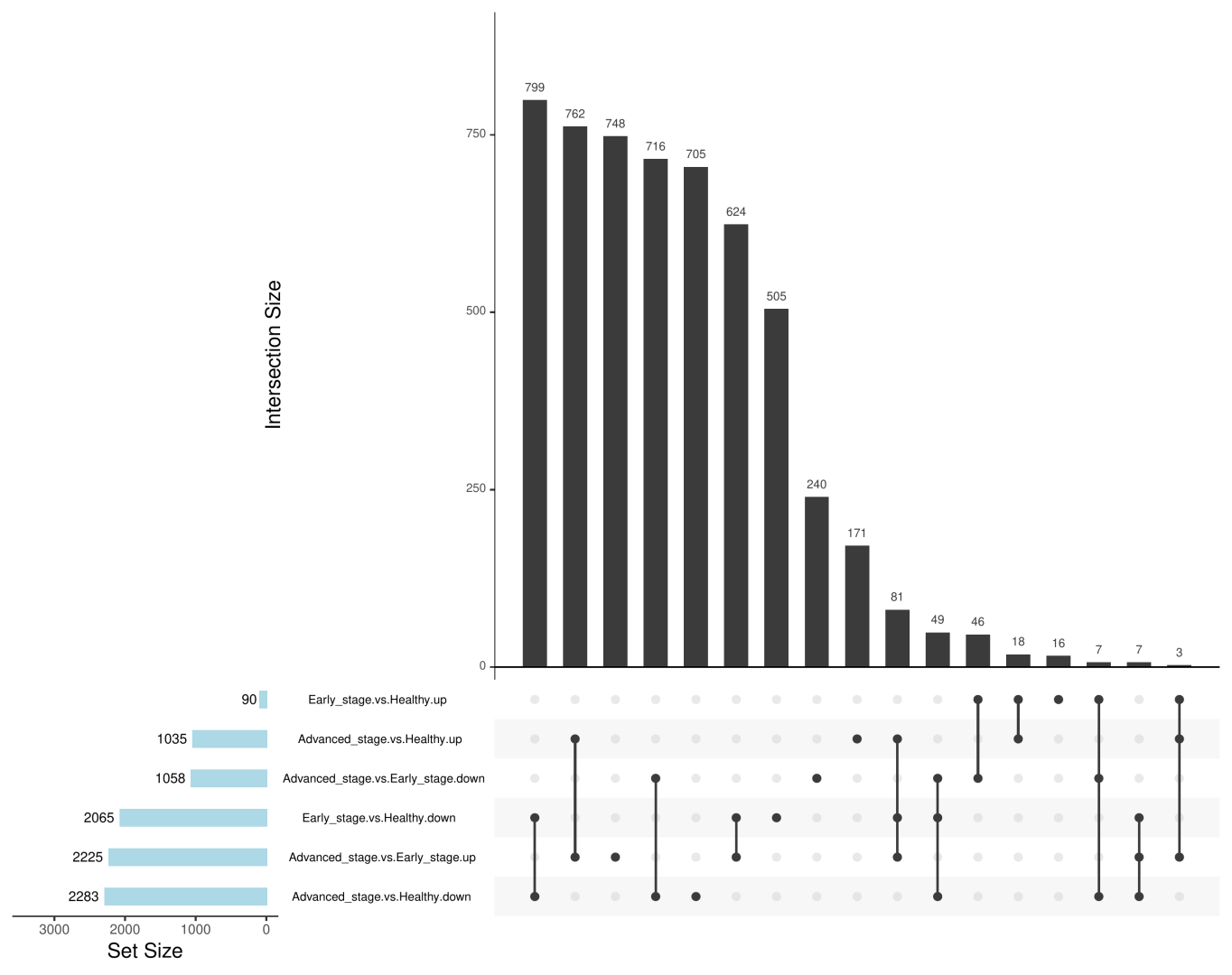


**Fig.** **4** MRNA Advanced stage vs Early stage

**(File path: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Advanced-stage-vs-Early-stage.pdf)**

* P.Value cut-off: 0.05
* Log2(FC) cut-off: 0.5

**(See: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Advanced-stage-vs-Early-stage-content)**



**Fig.** **5** MRNA Difference intersection

**(File path: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Difference-intersection.pdf)**

* All\_intersection:

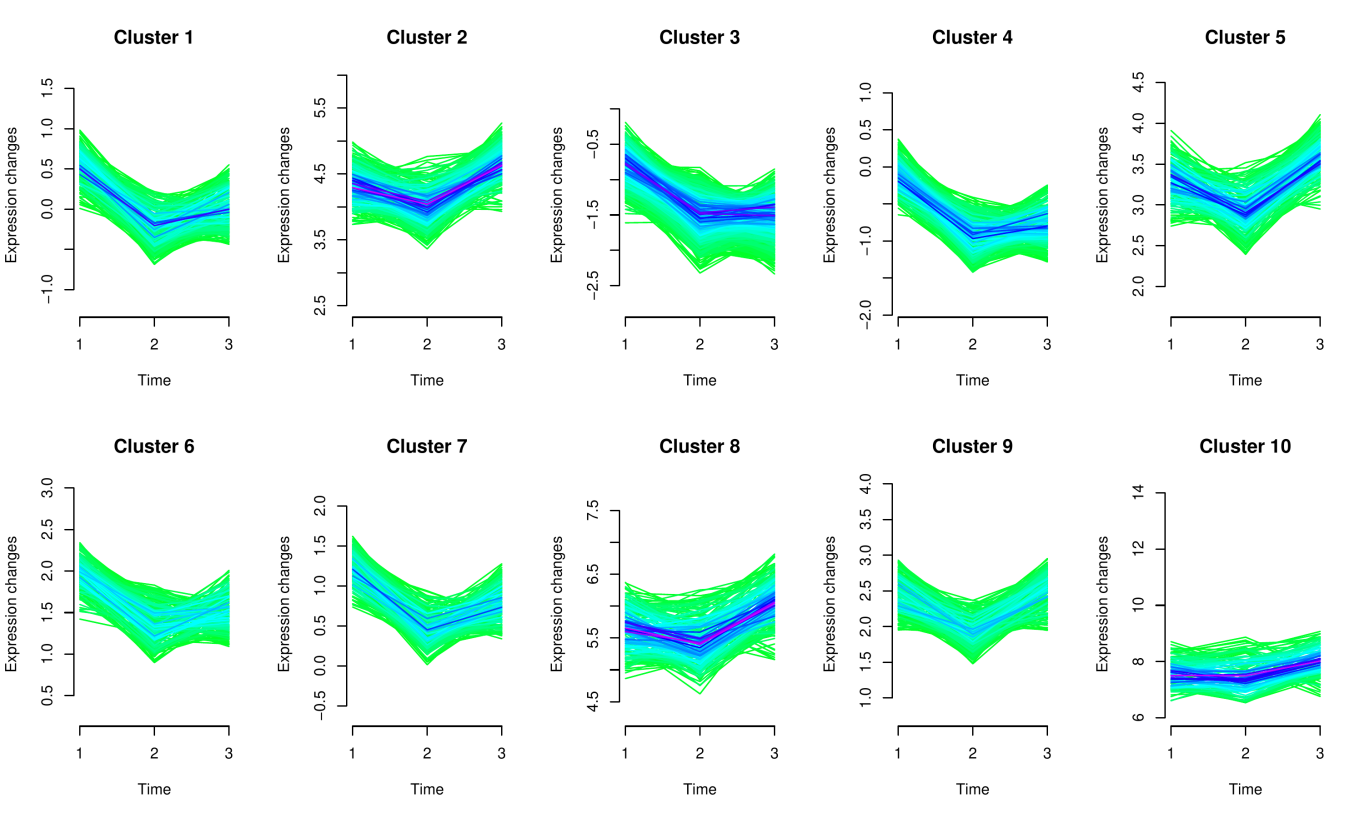
**(See: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-Difference-intersection-content)**

Note: The directory 'Figure+Table/MRNA-data-DEGs' contains 3 files.  
  
1 1\_Early\_stage - Healthy.csv  
2 2\_Advanced\_stage - Healthy.csv  
3 3\_Advanced\_stage - Early\_stage.csv

**(File path: Figure+Table/3.1\_Limma\_差异分析\_(MRNA)/MRNA-data-DEGs)**

## 3.2 Mfuzz 聚类分析 (MRNA)

将上述筛选得的 DEGs 以 Mfuzz 聚类分析。 见 按照 Healthy, Early\_stage, Advanced\_stage 顺序, 在 Mfuzz 聚类中，8, 10 为按时序上调，共 549 个， 3, 4 为按时序下调，共 1278 个， 其他基因为离散变化。。



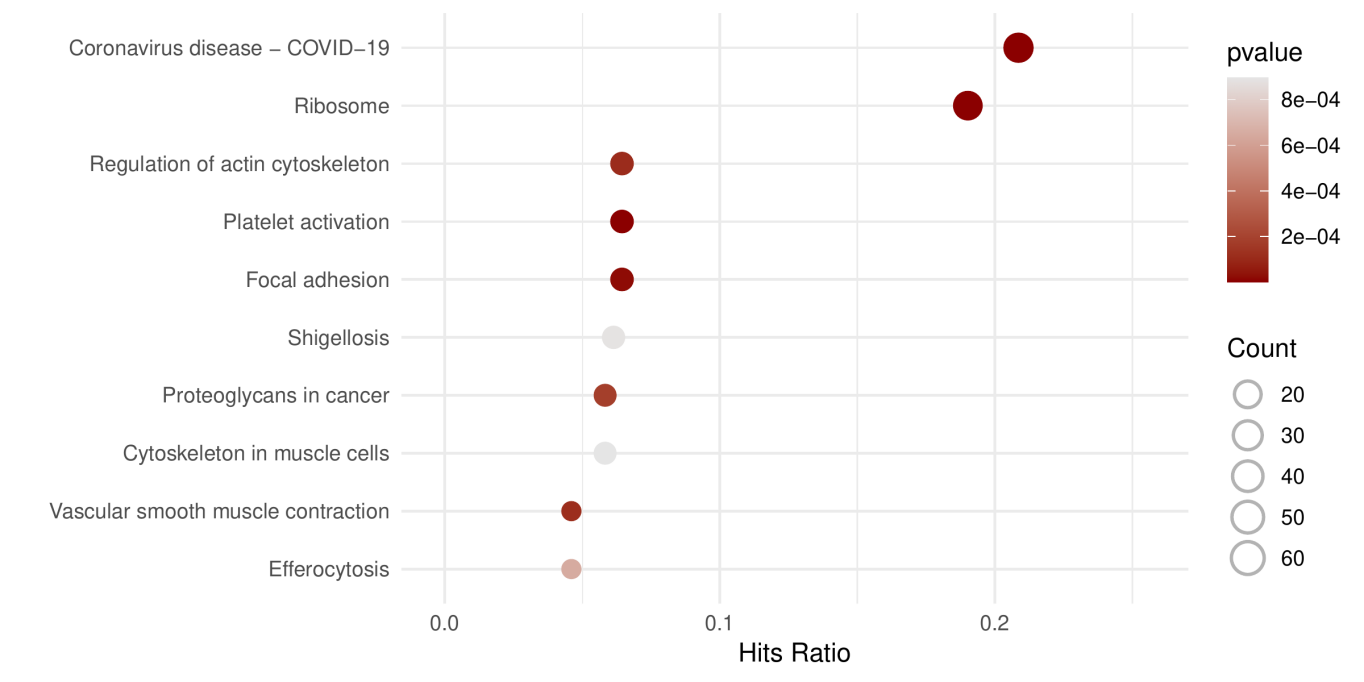
**Fig.** **6** MRNA Mfuzz clusters

**(File path: Figure+Table/3.2\_Mfuzz\_聚类分析\_(MRNA)/MRNA-Mfuzz-clusters.pdf)**

## 3.3 富集分析 (MRNA)

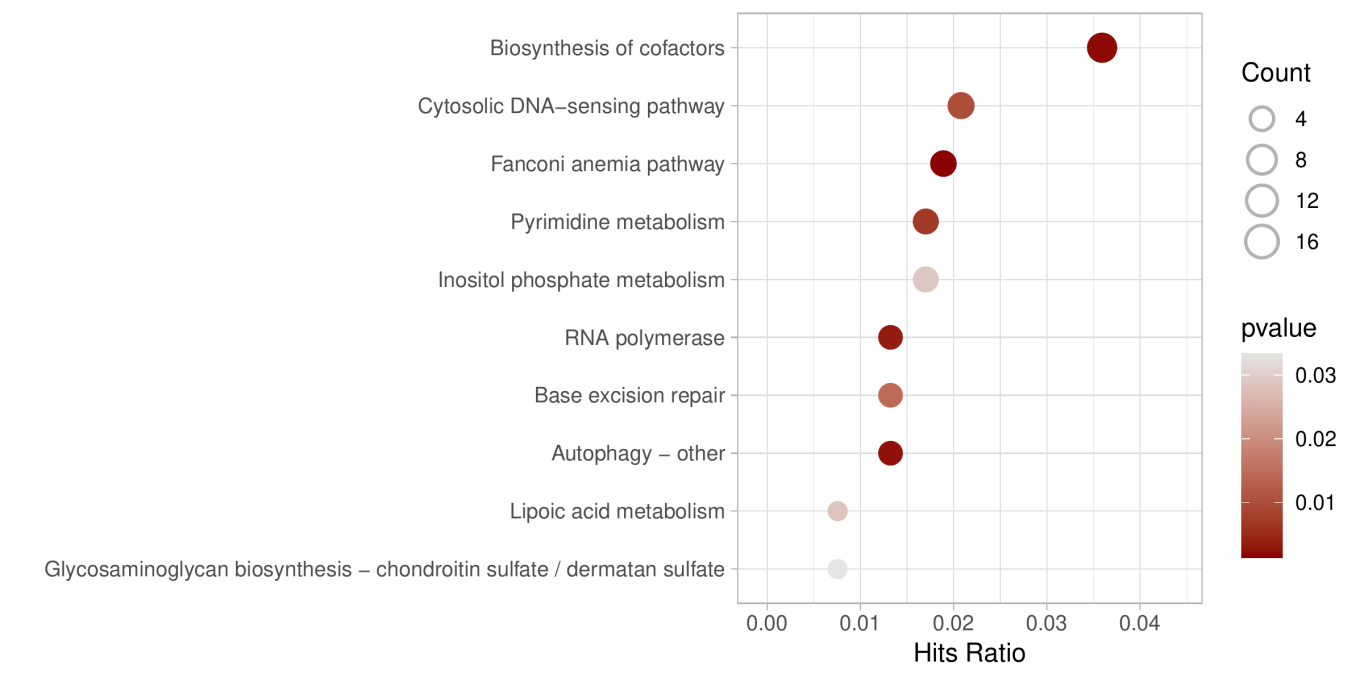
将 MFuzz 上调聚类与下调聚类分别以 KEGG 富集分析。 KEGG 见 Fig. **[8](#MRNA-downs-KEGG-enrichment)**, Fig. **[7](#MRNA-ups-KEGG-enrichment)** GO 见 Fig. **[10](#MRNA-downs-GO-enrichment)** Fig. **[9](#MRNA-ups-GO-enrichment)**。

可以发现，上调与下调富集的首要通路，都与遗传信息的处理相关。 在 中，下调的趋势更明显， 暗示 Biosynthesis of cofactors 等代谢相关的通路，可能更与肺癌的早期发展相关。



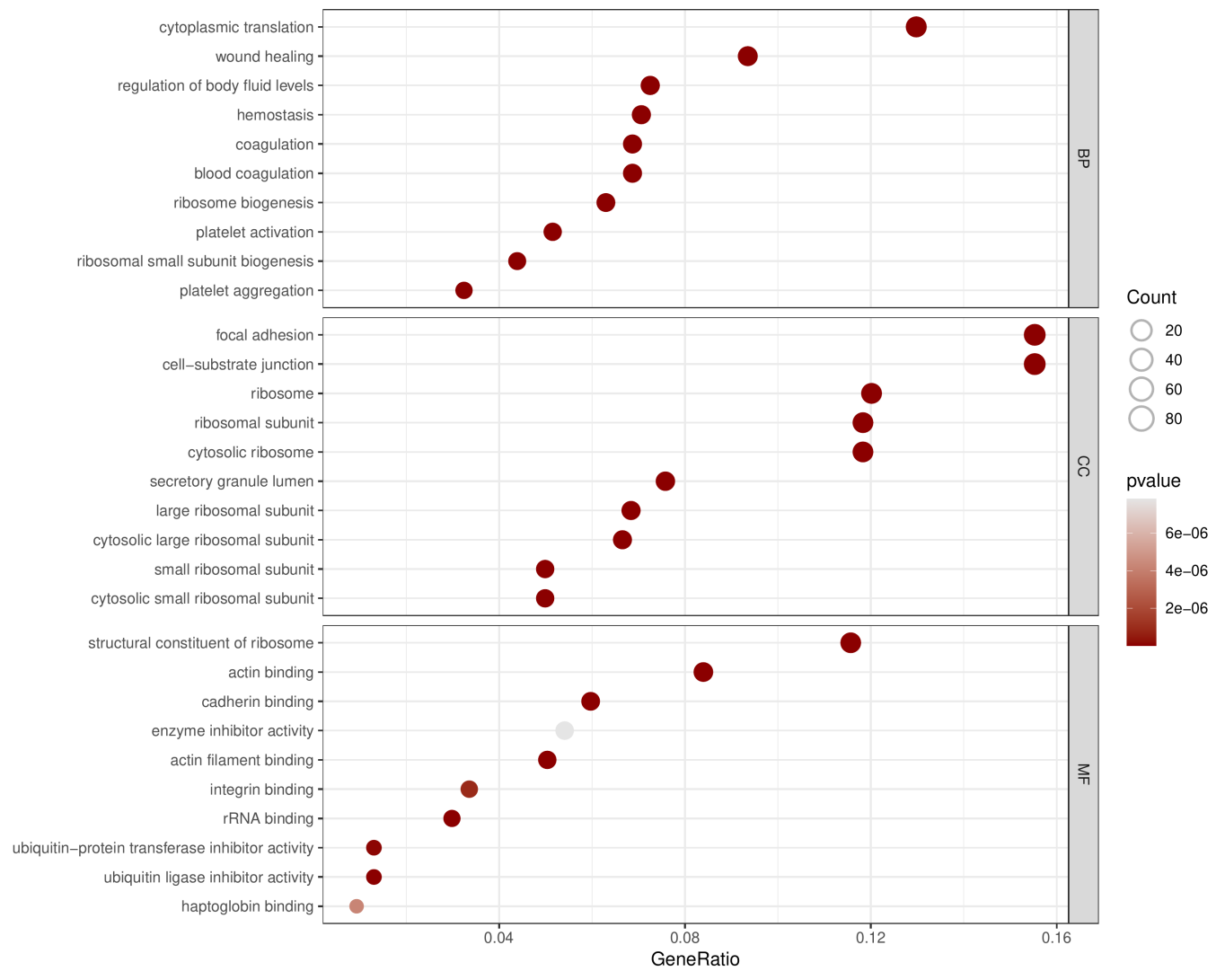
**Fig.** **7** MRNA ups KEGG enrichment

**(File path: Figure+Table/3.3\_富集分析\_(MRNA)/MRNA-ups-KEGG-enrichment.pdf)**



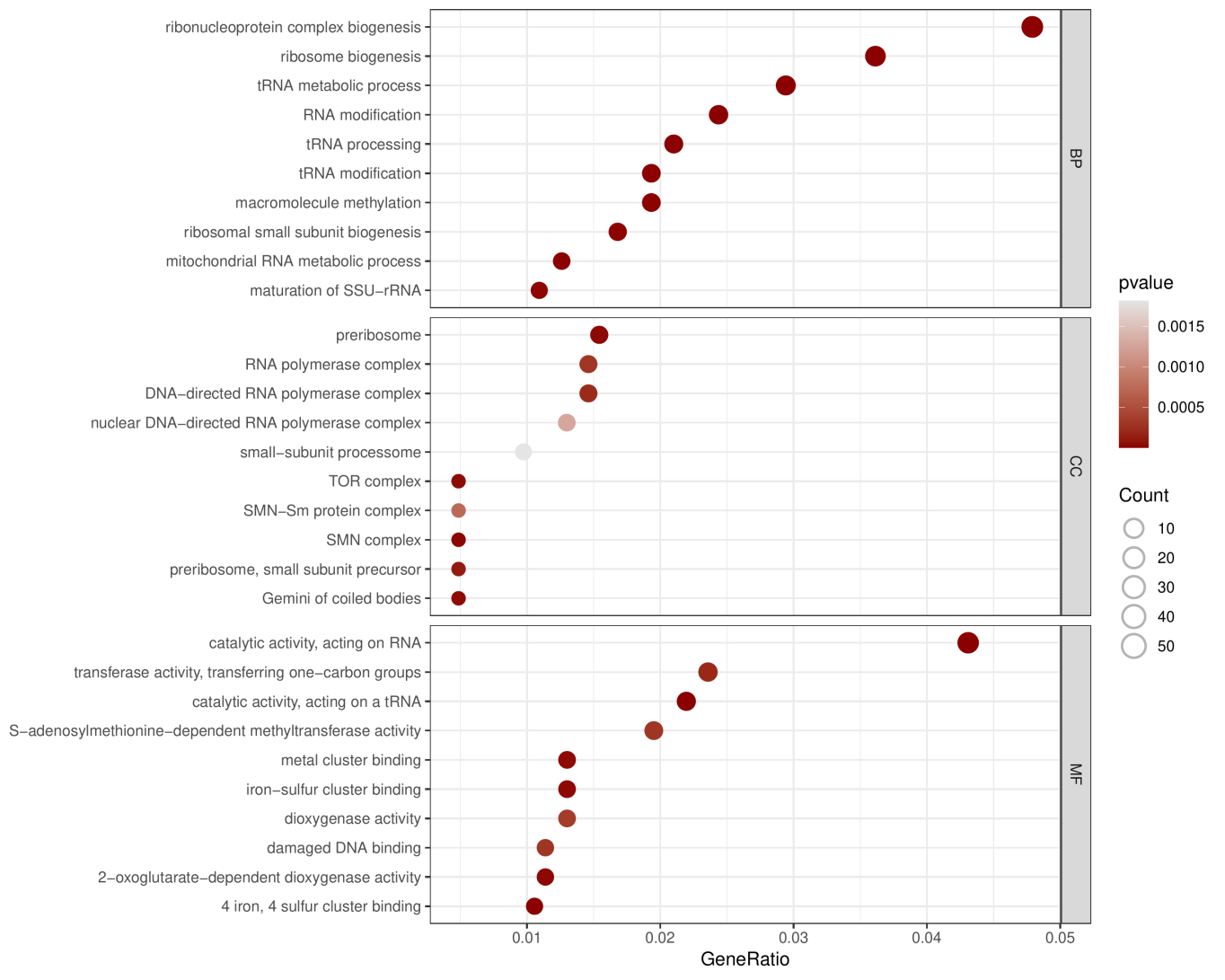
**Fig.** **8** MRNA downs KEGG enrichment

**(File path: Figure+Table/3.3\_富集分析\_(MRNA)/MRNA-downs-KEGG-enrichment.pdf)**



**Fig.** **9** MRNA ups GO enrichment

**(File path: Figure+Table/3.3\_富集分析\_(MRNA)/MRNA-ups-GO-enrichment.pdf)**



**Fig.** **10** MRNA downs GO enrichment

**(File path: Figure+Table/3.3\_富集分析\_(MRNA)/MRNA-downs-GO-enrichment.pdf)**

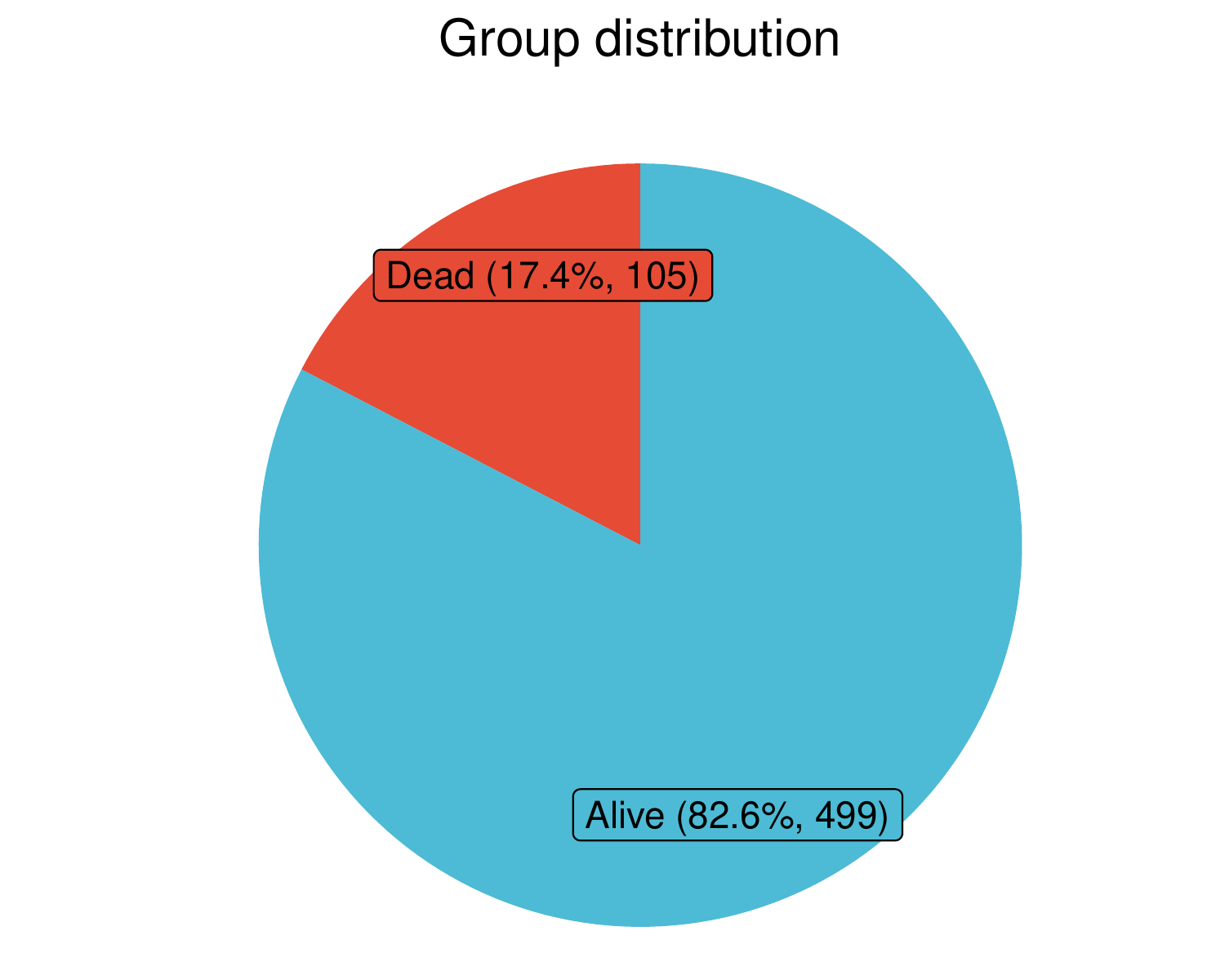
## 3.4 TCGA 数据获取 (LUSC, LUAD)

获取 TCGA-LUSC, TCGA-LUAD 数据，并将其合并，用于临床数据分析和预后模型建立。

## 3.5 COX 回归 (LUNG)

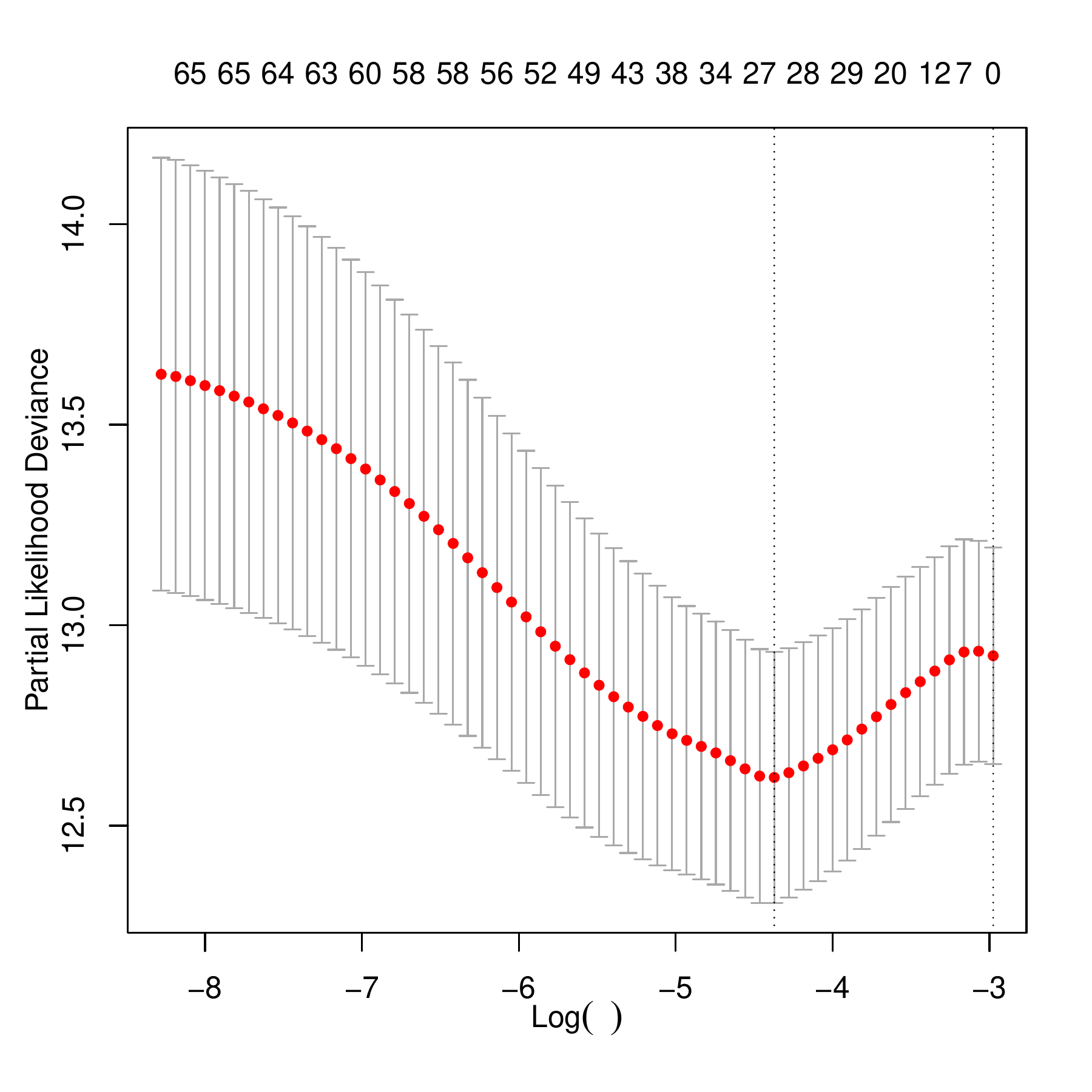
筛选 AJCC Stage (ajcc\_pathologic\_stage) 为 Stage I, Stage II 的病人，并且 days\_to\_last\_follow\_up 大于 30 天，且为肿瘤组织的样本。共 604 个样本。

将数据 (count) 标准化后 (同 MRNA 的方法)，以生存状态为指标 (Fig. **[11](#Group-distribution)**)，单因素 COX 回归，筛选能显著预测生存结局的基因。 在单因素 COX 回归的基础上，以 Lasso 处罚的多因素 COX 回归构建模型，十倍交叉验证，选择最佳 lambda 值， 见 Fig. **[12](#LUNG-lasso-COX-model)**，选择 lambda.min， 确定模型系数，见 Tab. **[1](#LUNG-sig-Multivariate-Cox-Coefficients)**



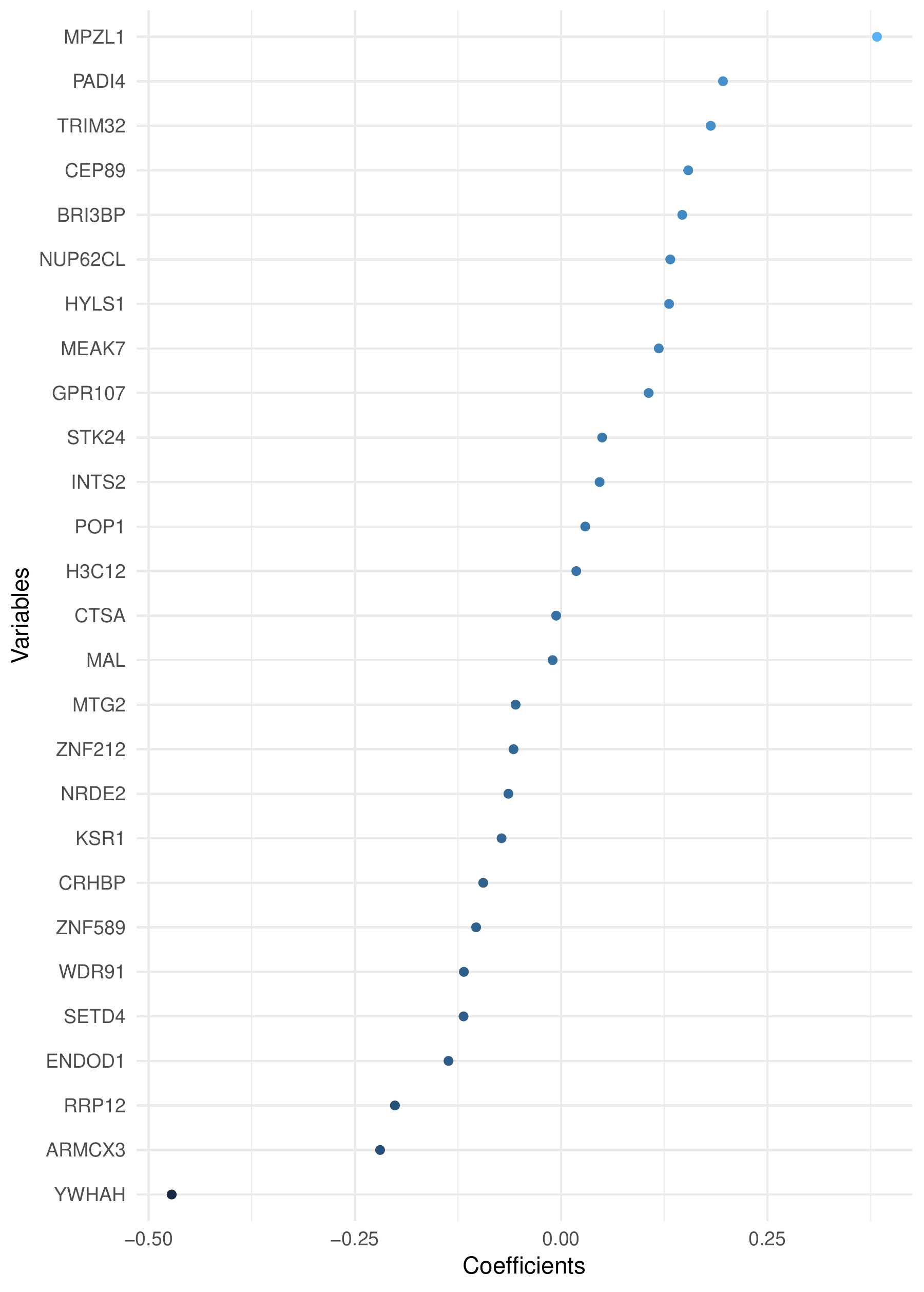
**Fig.** **11** Group distribution

**(File path: Figure+Table/3.5\_COX\_回归\_(LUNG)/Group-distribution.pdf)**



**Fig.** **12** LUNG lasso COX model

**(File path: Figure+Table/3.5\_COX\_回归\_(LUNG)/LUNG-lasso-COX-model.pdf)**



**Fig.** **13** LUNG lasso COX coeffients

**(File path: Figure+Table/3.5\_COX\_回归\_(LUNG)/LUNG-lasso-COX-coeffients.pdf)**

**Tab.** **1** LUNG sig Multivariate Cox Coefficients

| Feature | Coef |
| --- | --- |
| RRP12 | -0.201464208652924 |
| CTSA | -0.0060851143741419 |
| MTG2 | -0.0552036433279771 |
| ARMCX3 | -0.219510556361052 |
| STK24 | 0.0496255102834819 |
| ... | ... |

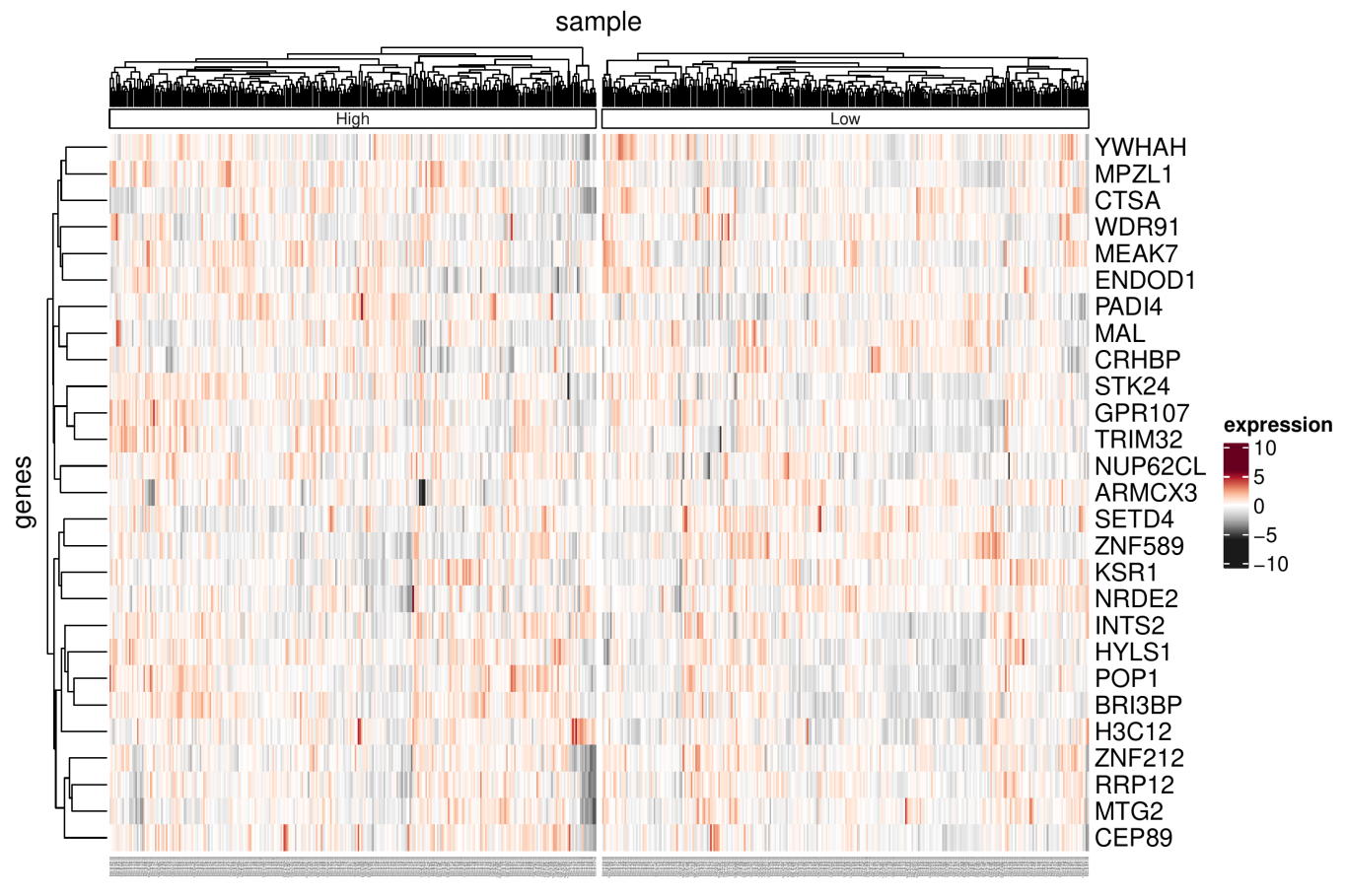
**(File path: Figure+Table/3.5\_COX\_回归\_(LUNG)/LUNG-sig-Multivariate-Cox-Coefficients.csv)**

## 3.6 Survival 生存分析 (LUNG)

这些基因表达特征如热图所示Fig. **[14](#LUNG-risk-score-related-genes-heatmap)**，

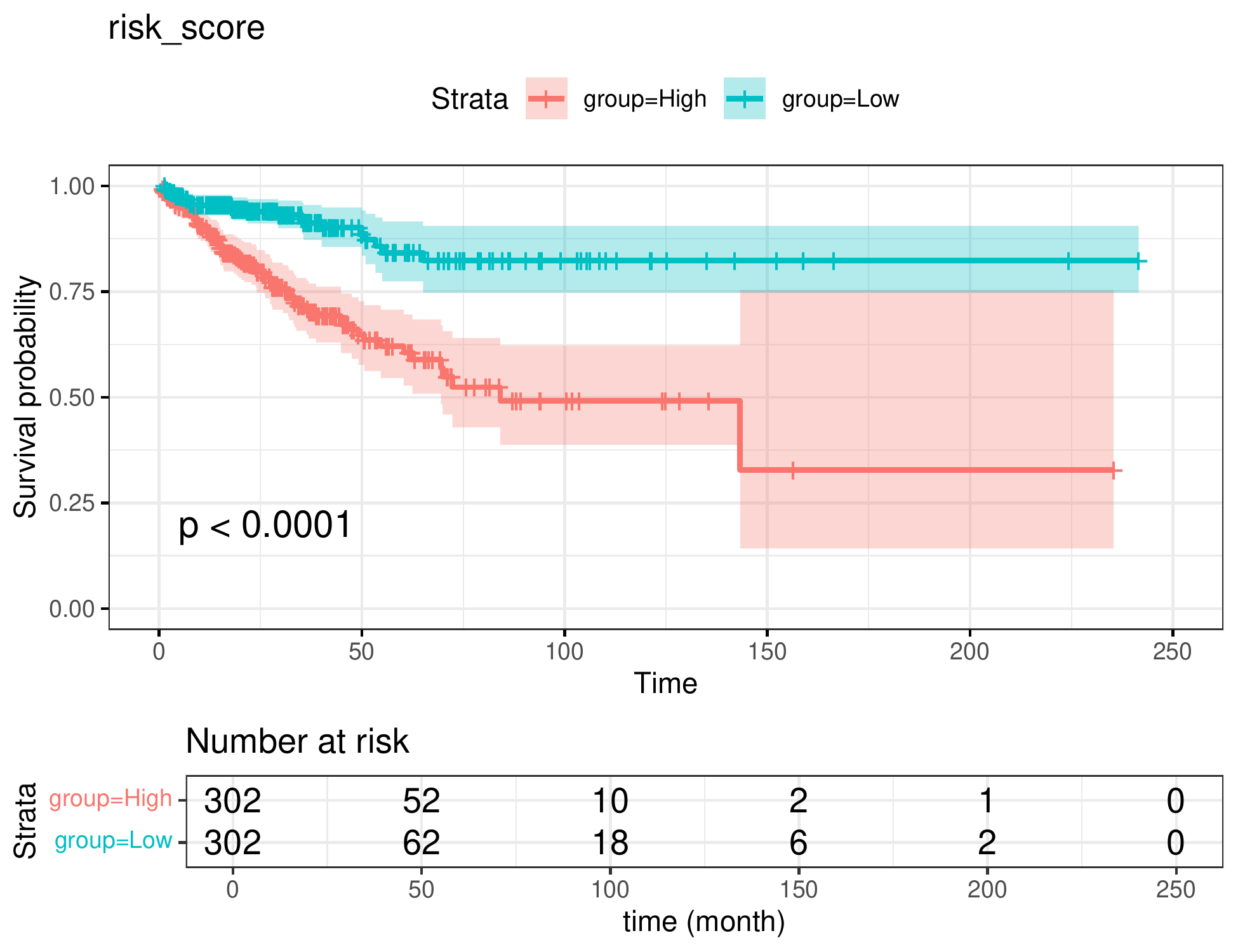
建立预后特征，构建风险评分：

按中位风险评分，将病例分为 Low 和 High 风险组，随后进行生存分析， 见 Fig. **[15](#LUNG-survival-curve-of-risk-score)**。 此外，第 1，3，5 年存活的患者，风险评分显著较低Fig. **[17](#LUNG-boxplot-of-risk-score)**， AUC 见Fig. **[16](#LUNG-time-ROC)** 。



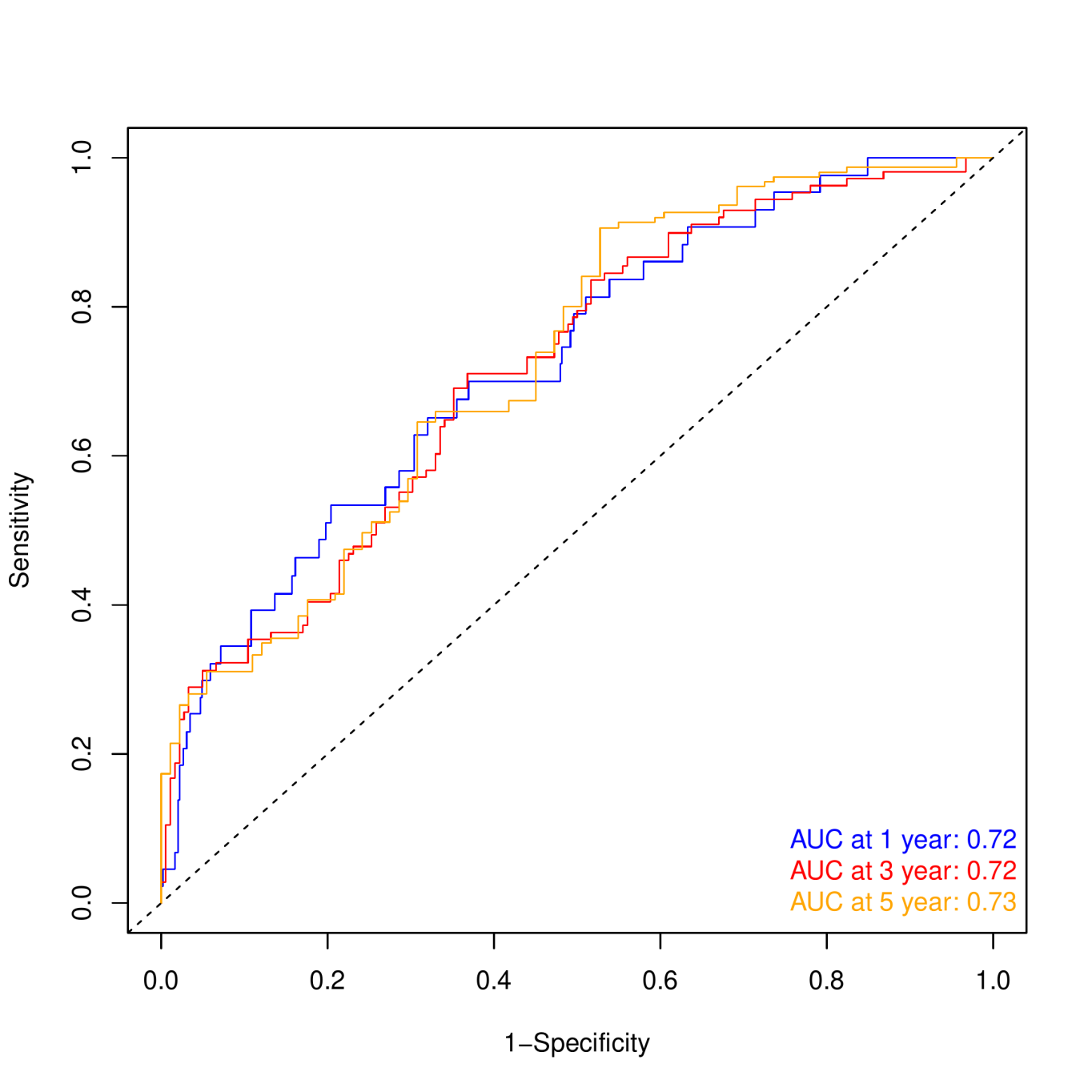
**Fig.** **14** LUNG risk score related genes heatmap

**(File path: Figure+Table/3.6\_Survival\_生存分析\_(LUNG)/LUNG-risk-score-related-genes-heatmap.pdf)**



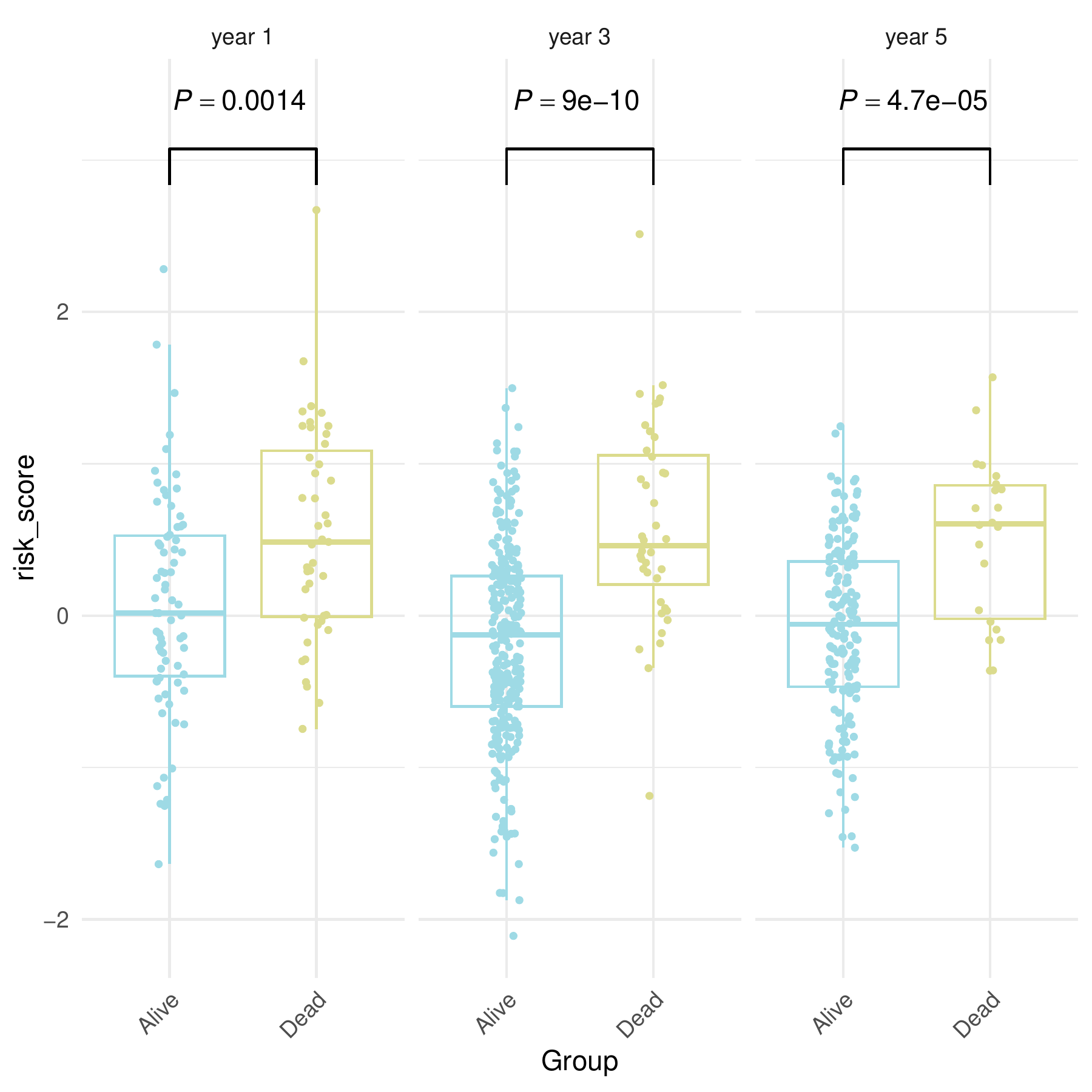
**Fig.** **15** LUNG survival curve of risk score

**(File path: Figure+Table/3.6\_Survival\_生存分析\_(LUNG)/LUNG-survival-curve-of-risk-score.pdf)**



**Fig.** **16** LUNG time ROC

**(File path: Figure+Table/3.6\_Survival\_生存分析\_(LUNG)/LUNG-time-ROC.pdf)**



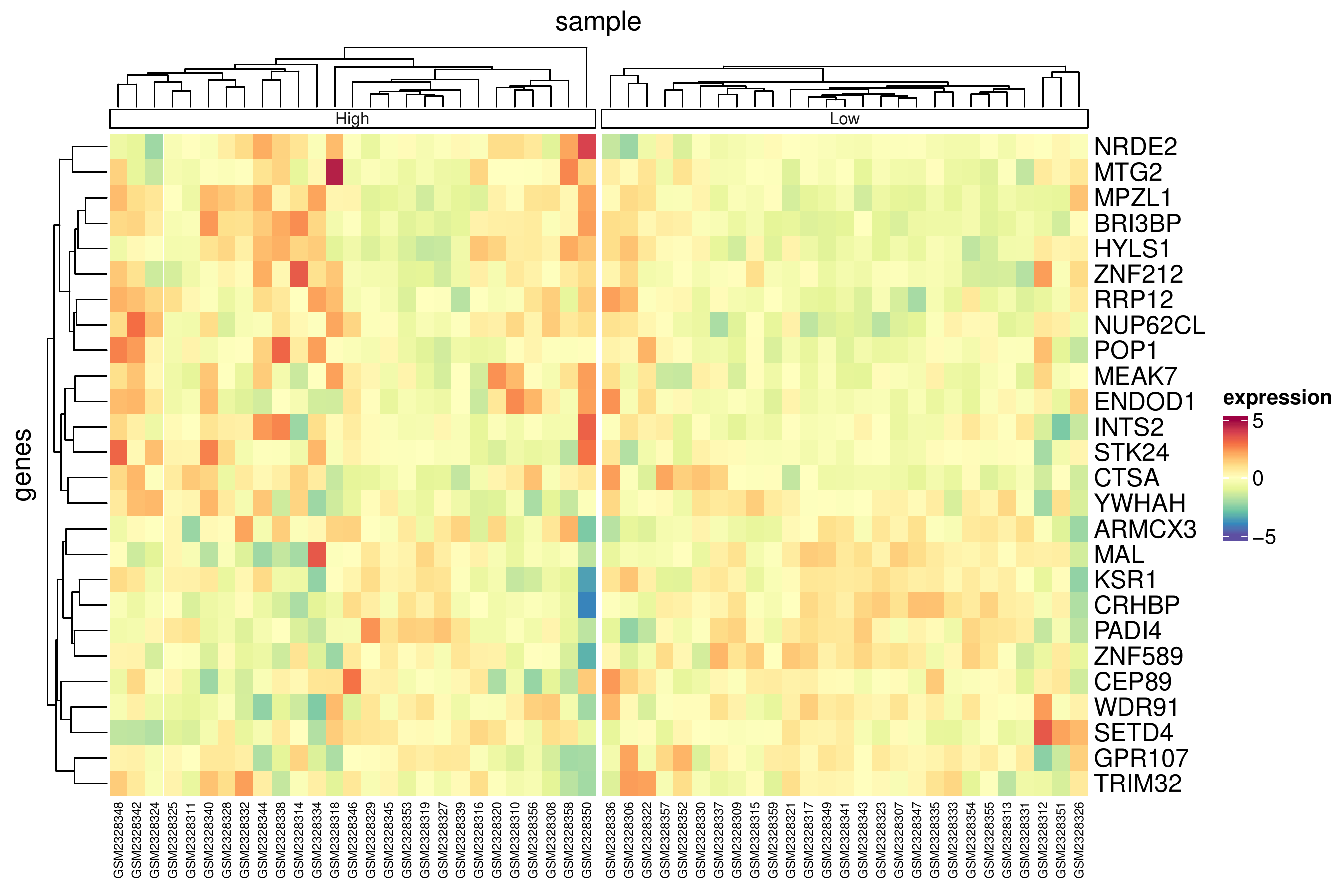
**Fig.** **17** LUNG boxplot of risk score

**(File path: Figure+Table/3.6\_Survival\_生存分析\_(LUNG)/LUNG-boxplot-of-risk-score.pdf)**

## 3.7 GEO 数据获取 (GEO\_LUAD)

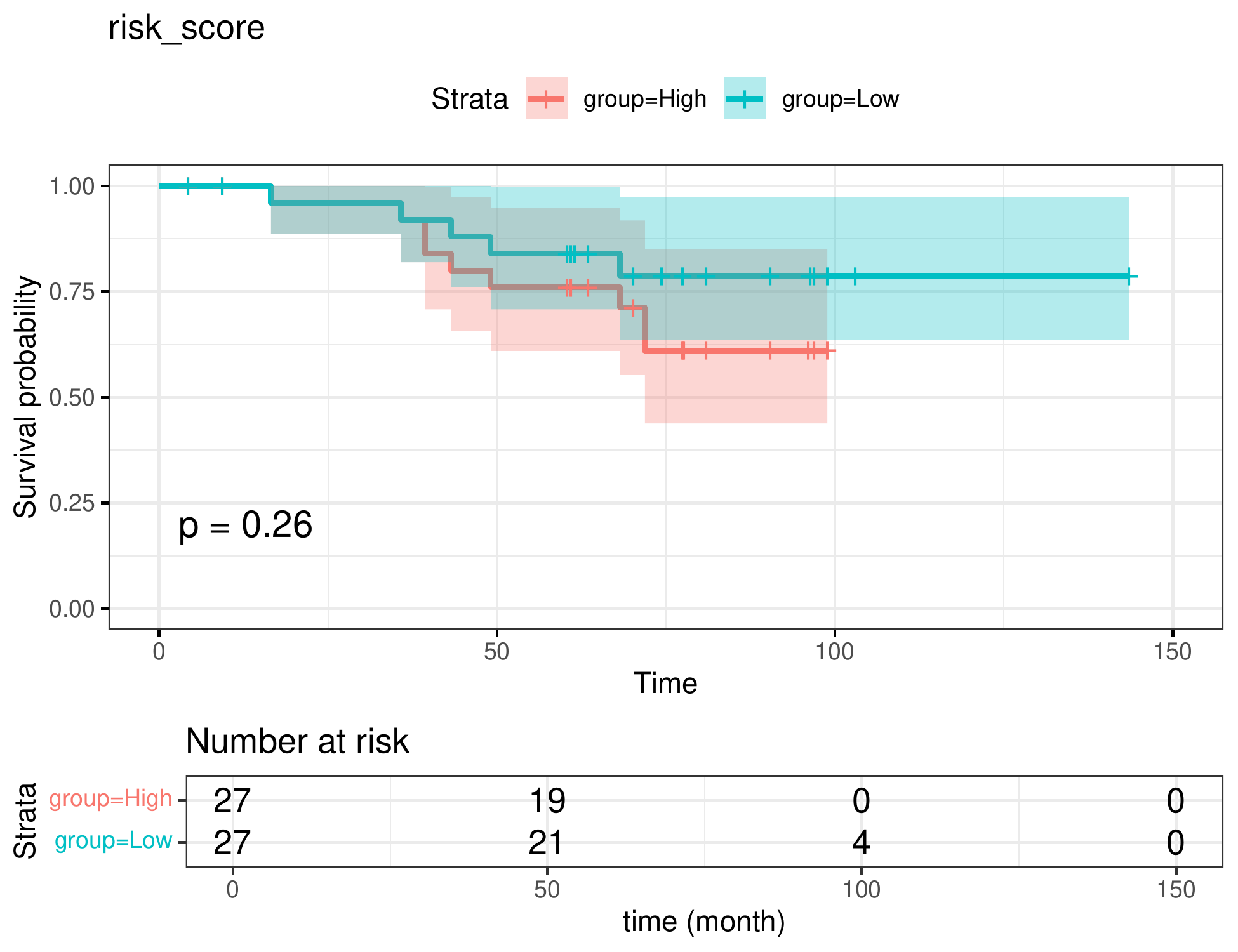
使用 EDirect 检索所有包含 LUAD 数据后，进一步以 ‘Expression profiling by high throughput sequencing’ 为条件 (考虑到 RNAseq 与微阵列数据的差异性)，最后，我们找到了一个可用的验证数据集: GSE87340。 该数据集包含 54 个样本。

风险评分相关基因表达见 Fig. **[18](#GEO-LUAD-risk-score-related-genes-heatmap)** (注，该 GEO 数据集不含 H3C12 基因) 生存曲线见 Fig. **[19](#GEO-LUAD-survival-curve-of-risk-score)**，高风险评分生存率低于低风险评分组，但碍于于样本量较小，未显著。ROC (五年) 见 Fig. **[20](#GEO-LUAD-time-ROC)** 。



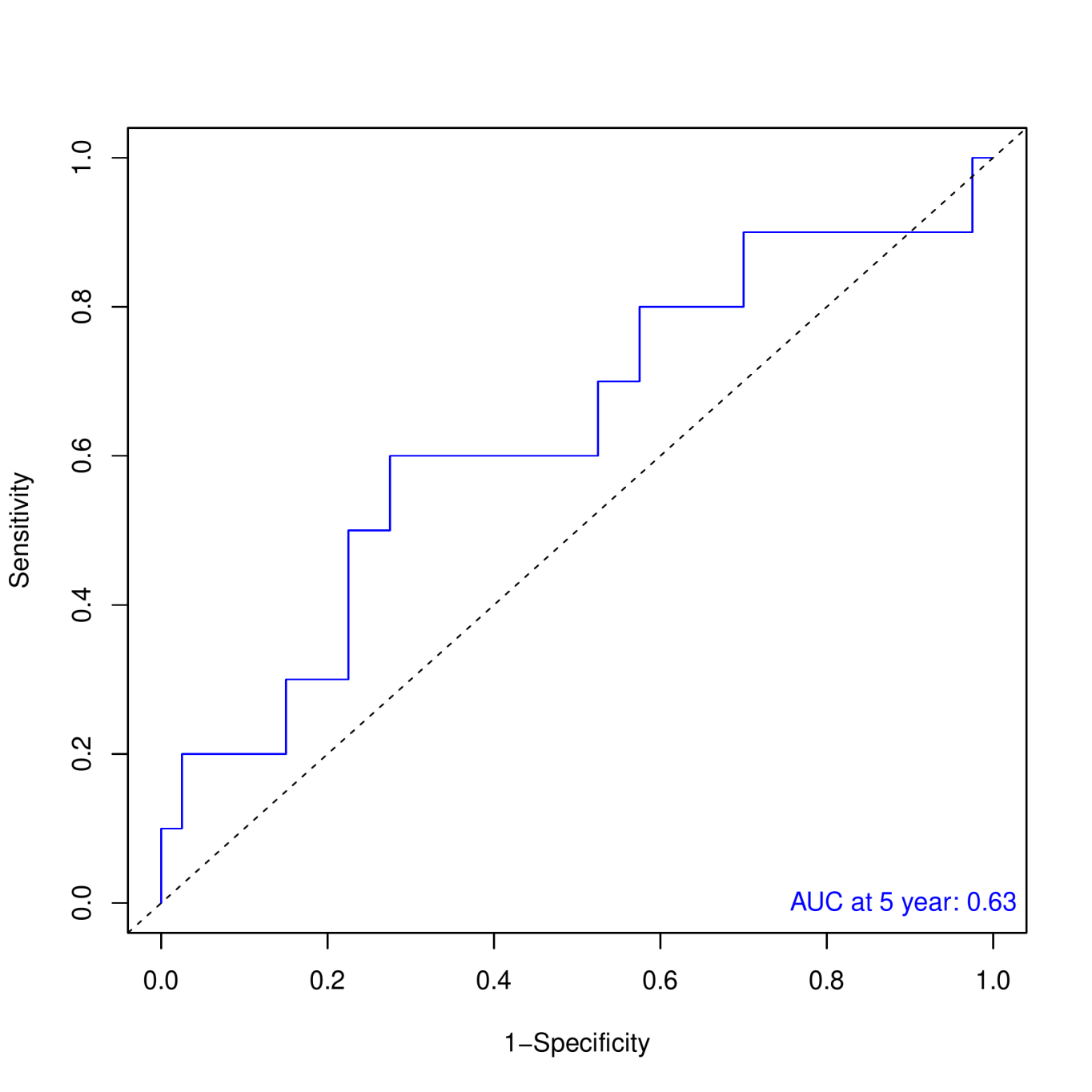
**Fig.** **18** GEO LUAD risk score related genes heatmap

**(File path: Figure+Table/3.7\_GEO\_数据获取\_(GEO\_LUAD)/GEO-LUAD-risk-score-related-genes-heatmap.pdf)**



**Fig.** **19** GEO LUAD survival curve of risk score

**(File path: Figure+Table/3.7\_GEO\_数据获取\_(GEO\_LUAD)/GEO-LUAD-survival-curve-of-risk-score.pdf)**

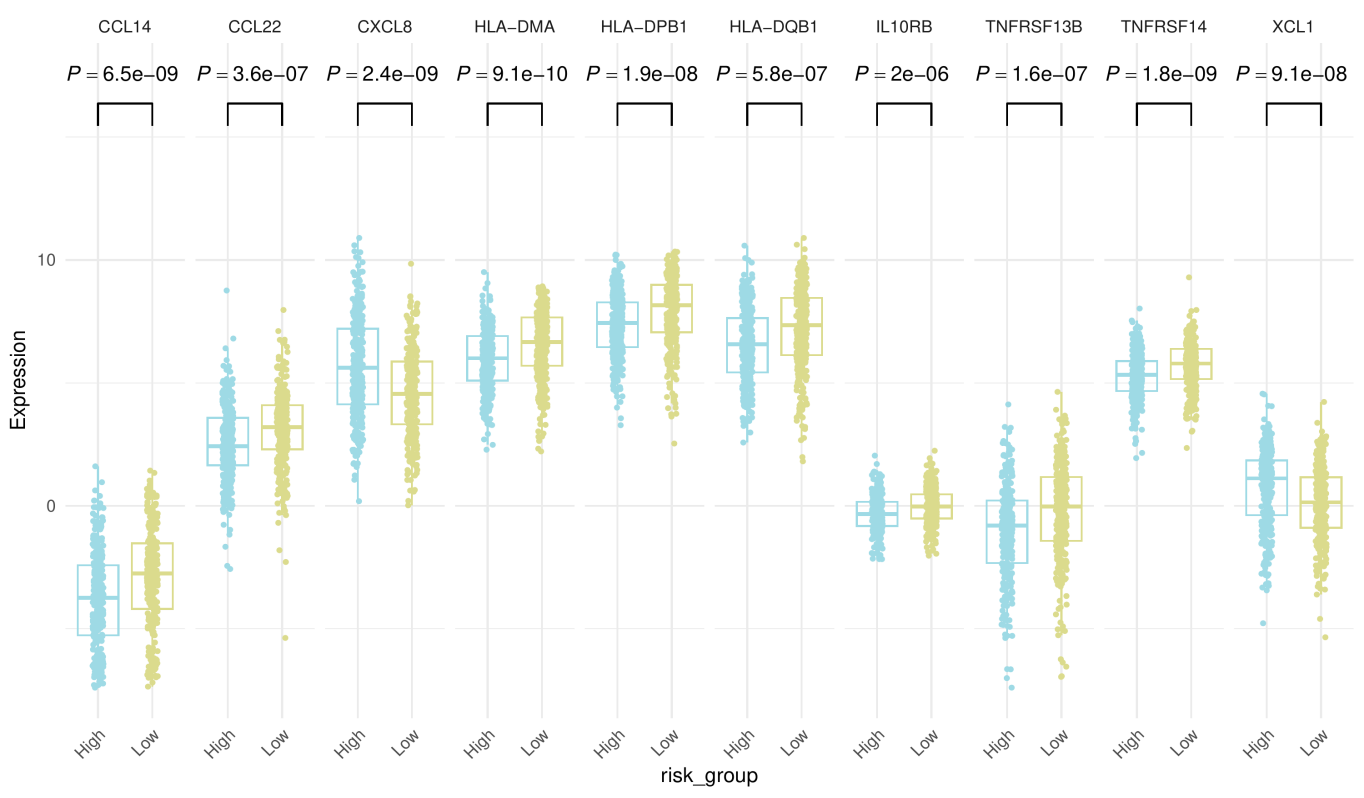


**Fig.** **20** GEO LUAD time ROC

**(File path: Figure+Table/3.7\_GEO\_数据获取\_(GEO\_LUAD)/GEO-LUAD-time-ROC.pdf)**

## 3.8 estimate 免疫评分 (LUNG)

为了探索标记与肿瘤免疫微环境之间的关系，我们对来自 TCGA LUSC 的数据进行了 ESTIMATE 计算免疫评分、ESTIMATE 评分和stromal 评分。根据评分结果，将病例分为 High 组和 Low 组。 还比较了高危组和低危组之间编码免疫调节剂和趋化因子的基因的表达情况。从 TISIDB 数据库下载的 178 个基因中，有 127 个可以在 TCGA 表达矩阵中找到，两组之间有 79 个表达存在差异 (p.value < 0.05)。 前 10 个基因见 Fig. **[21](#LUSC-Top10-Immune-Related-Genes)**



**Fig.** **21** LUSC Top10 Immune Related Genes

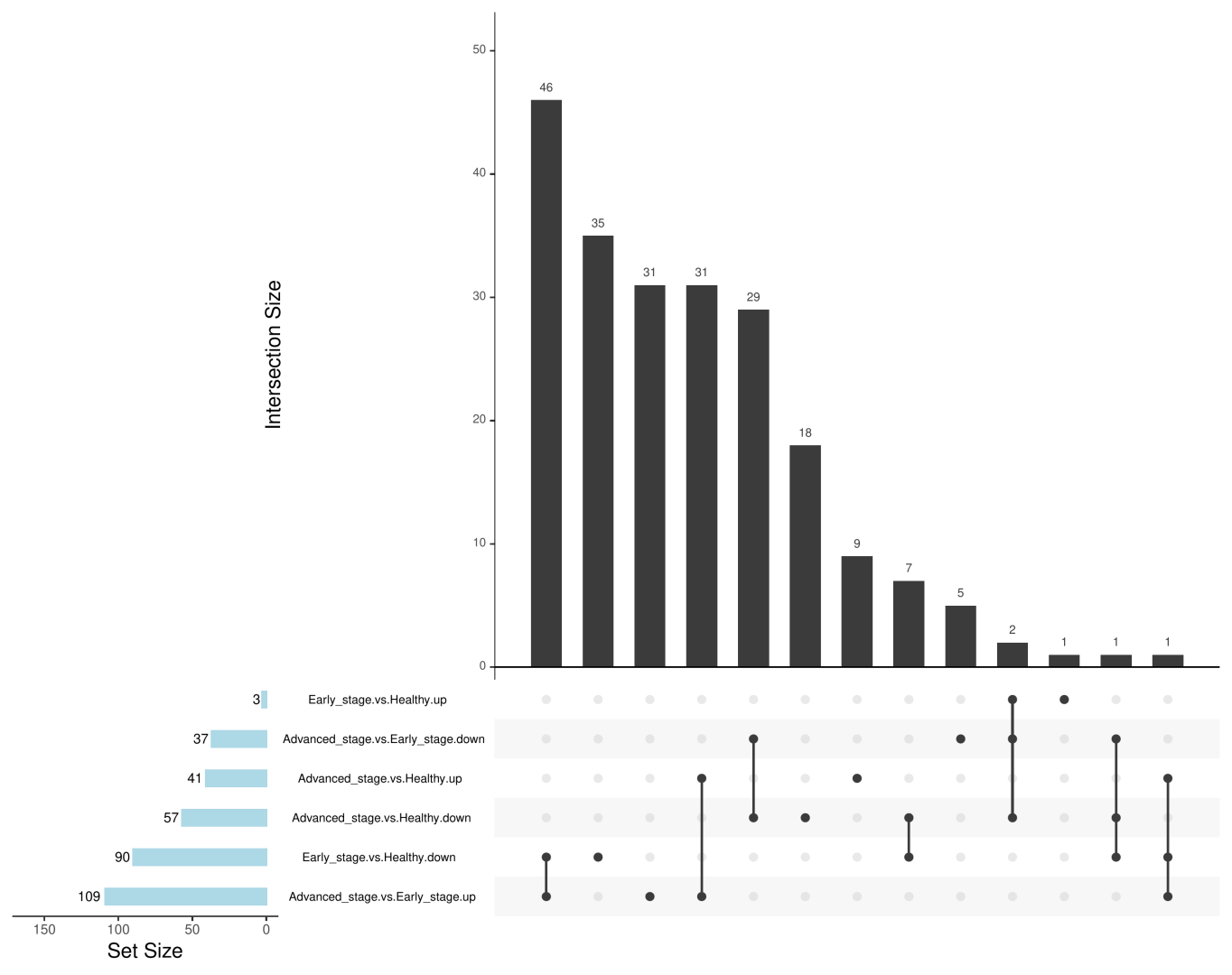
**(File path: Figure+Table/3.8\_estimate\_免疫评分\_(LUNG)/LUSC-Top10-Immune-Related-Genes.pdf)**

## 3.9 Limma 差异分析 (LNCRNA)

长链非编码RNA（lncRNA）在基因调控和癌症发展中起着重要作用。 这里对 lncRNA 做了差异分析，并与 mRNA 关联分析。 差异分析 Early\_stage vs Healthy, Advanced\_stage vs Healthy, Advanced\_stage vs Early\_stage (若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。 得到的 DEGs 统计见 Fig. **[22](#LNCRNA-Difference-intersection)** 所有上调 DEGs 有 121 个，下调共 144；一共 216 个 (非重复)。。

Note: The directory 'Figure+Table/LNCRNA-DEGs-data' contains 3 files.  
  
1 1\_Early\_stage - Healthy.csv  
2 2\_Advanced\_stage - Healthy.csv  
3 3\_Advanced\_stage - Early\_stage.csv

**(File path: Figure+Table/3.9\_Limma\_差异分析\_(LNCRNA)/LNCRNA-DEGs-data)**



**Fig.** **22** LNCRNA Difference intersection

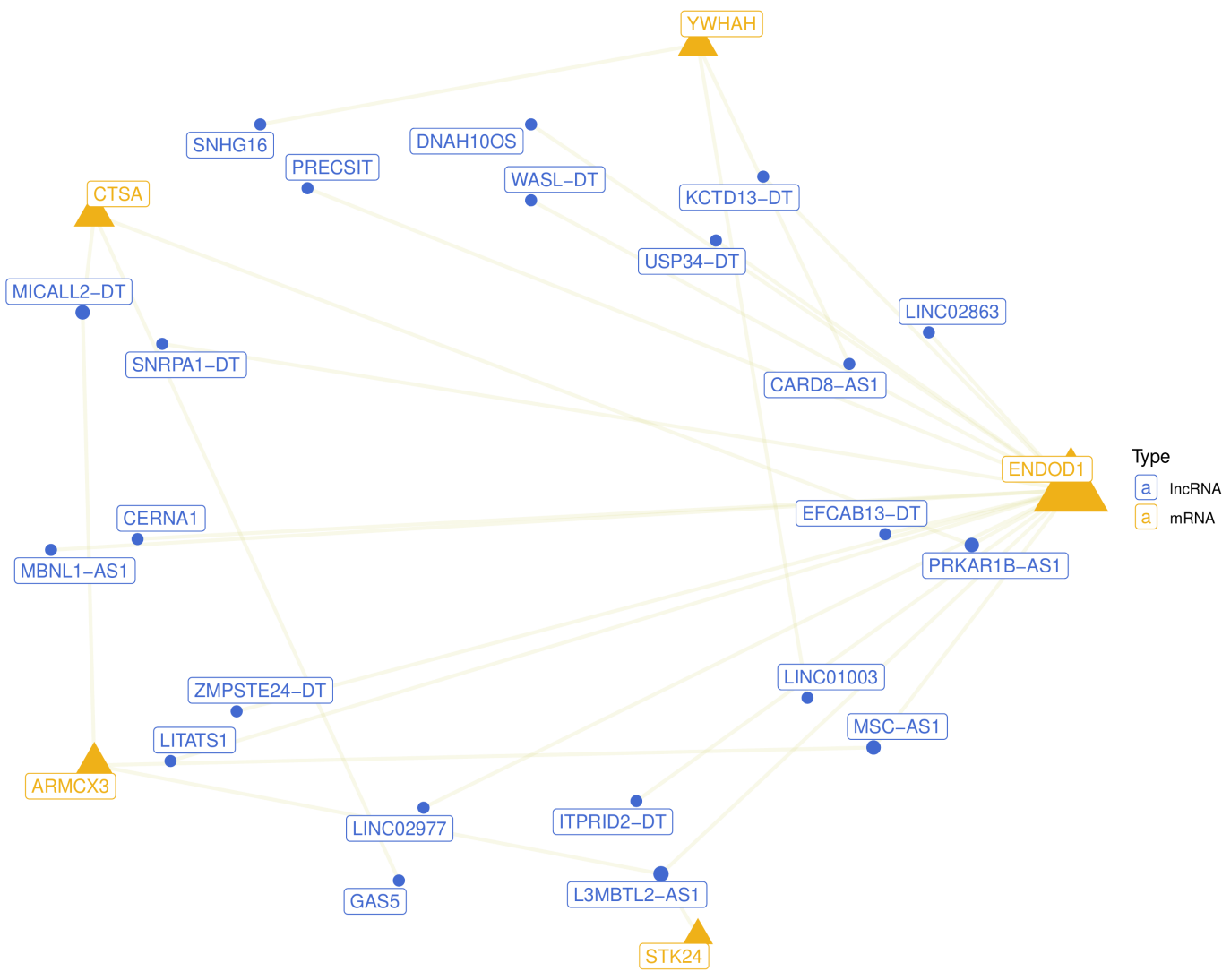
**(File path: Figure+Table/3.9\_Limma\_差异分析\_(LNCRNA)/LNCRNA-Difference-intersection.pdf)**

* All\_intersection:

**(See: Figure+Table/3.9\_Limma\_差异分析\_(LNCRNA)/LNCRNA-Difference-intersection-content)**

## 3.10 关联分析 (MRNA, LNCRNA)

将相关系数 > 0.5 和 p < 0.001 设定为识别相关阈值，最终建立网络图见 Fig. **[23](#Significant-Correlation-mrna-lncRNA)** 共包含 5 个 mRNA， 5 个 lncRNA， 27 对关联关系。



**Fig.** **23** Significant Correlation mrna lncRNA

**(File path: Figure+Table/3.10\_关联分析\_(MRNA,\_LNCRNA)/Significant-Correlation-mrna-lncRNA.pdf)**

**Tab.** **2** Significant correlation

| MRNA | LncRNA | Cor | Pvalue | -log2(P.va... | Significant | Sign |
| --- | --- | --- | --- | --- | --- | --- |
| ENDOD1 | WASL-DT | 0.55 | 0 | 16.6096404... | < 0.001 | \*\* |
| ENDOD1 | USP34-DT | 0.56 | 0 | 16.6096404... | < 0.001 | \*\* |
| YWHAH | CARD8-AS1 | -0.58 | 0 | 16.6096404... | < 0.001 | \*\* |
| ENDOD1 | EFCAB13-DT | 0.64 | 0 | 16.6096404... | < 0.001 | \*\* |
| YWHAH | LINC01003 | 0.61 | 0 | 16.6096404... | < 0.001 | \*\* |
| ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.10\_关联分析\_(MRNA,\_LNCRNA)/Significant-correlation.csv)**

## 3.11 实验验证

请参考 (2023, **IF:4.8**, Q1, Biomolecules)1

# 4 总结

本研究为肺癌早期诊断建立了预后的独立风险指标，这些基因是 RRP12, CTSA, MTG2, ARMCX3, STK24, POP1, WDR91, INTS2, TRIM32, NRDE2, CEP89, YWHAH, MEAK7, KSR1, CRHBP, GPR107, ENDOD1, PADI4, ZNF589, ZNF212, MAL, BRI3BP, SETD4, H3C12, MPZL1, NUP62CL, HYLS1， 可预测肺癌 (包括 LUSC, LUAD) 中，Sage I、II 的预后疗效。 该风险评分对于 RNA-seq 可能有更敏感的评估，因为我们在 GEO 的微阵列数据集中，High 组与 Low 组 的风险评分差异不如 TCGA 显著。由于 GEO 中，包含生存结局和详细临床数据记录的数据集不多， 仅对 LUSC 的数据集进行了验证，而未能找到 LUAD 的 RNA-seq 数据集进行验证 (对 GEO 的搜索策略涵盖了整个 GEO 数据集：以 EDirect 检索得到 GSE ID，所有以 GEOquery 获取数据集的元数据，仅包含生存数据的数据集可用于验证) 。 诊断模型的构建，采用了 LASSO 处罚的 COX 回归，并采用了十倍交叉验证的方式选择参数，一定程度上反映了模型的鲁棒性。

# Reference

1. Wang, H. *et al.* HCC: RNA-sequencing in cirrhosis. *Biomolecules* **13**, (2023).

2. Smyth, G. K. Limma: Linear models for microarray data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer-Verlag, 2005). doi:[10.1007/0-387-29362-0\_23](https://doi.org/10.1007/0-387-29362-0_23).

3. Chen, Y., McCarthy, D., Ritchie, M., Robinson, M. & Smyth, G. EdgeR: Differential analysis of sequence read count data users guide. 119.

4. Kumar, L. & E Futschik, M. Mfuzz: A software package for soft clustering of microarray data. *Bioinformation* **2**, 5–7 (2007).

5. Wu, T. *et al.* ClusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, (2021).

6. Colaprico, A. *et al.* TCGAbiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research* **44**, (2015).

7. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nature communications* **4**, (2013).

8. Ru, B. *et al.* TISIDB: An integrated repository portal for tumorimmune system interactions. *Bioinformatics* **35**, 4200–4202 (2019).