**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 ;**

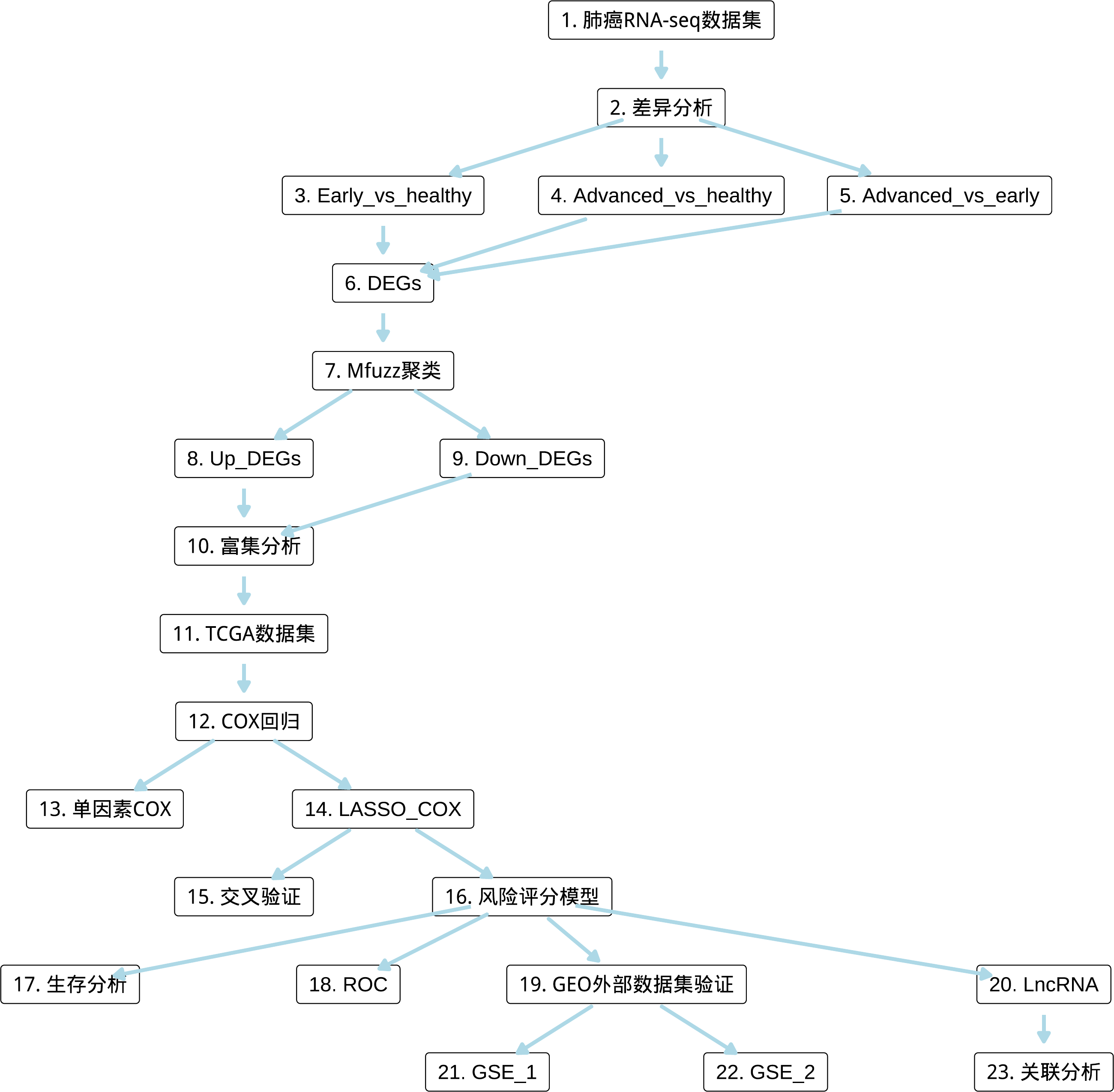
**分析类型： 补充分析 ;**

**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.** 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。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。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)。以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) 用于线性分析。 创建对比矩阵，差异分析：Early\_stage vs Healthy, Advanced\_stage vs Healthy, Advanced\_stage vs Early\_stage。使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 拟合线形模型。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)

根据样本分组 (Advanced\_stage (n=65) , Early\_stage (n=101) , Healthy (n=81) )，取各 DEGs 表达量的平均值。 以 R 包 Mfuzz (2.66.0) (2007, **IF:NA**, NA, Bioinformation)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 ClusterProfiler 富集分析 (Dataset: MRNA)

将 MFuzz 的 ups, downs 聚类分别以 KEGG, GO 富集分析。 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。 以 R 包glmnet(4.1.8) 作 lasso 处罚的 cox 回归，以cv.glmnet` 函数作 5 交叉验证获得模型。

筛选 ajcc\_pathologic\_stage 为 Stage IB, Stage IA, Stage IIB, Stage IIA, Stage II, Stage I, 筛选 days\_to\_last\_follow\_up 大于或等于 30, 筛选 isTumor 为 tumor，最终得到 615 例数据。

## 2.8 Survival 生存分析 (Dataset: LUNG)

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 (Dataset: LUNG)

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 cancer[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: LUNG)

以 R 包 GEOquery (2.74.0) 获取 GSE87340 数据集。 以 GEOquery::getRNASeqData 获取 RNA count 数据以及基因注释。

## 2.11 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。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。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)。以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) 用于线性分析。 创建对比矩阵，差异分析：Early\_stage vs Healthy, Advanced\_stage vs Healthy, Advanced\_stage vs Early\_stage。使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 拟合线形模型。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。

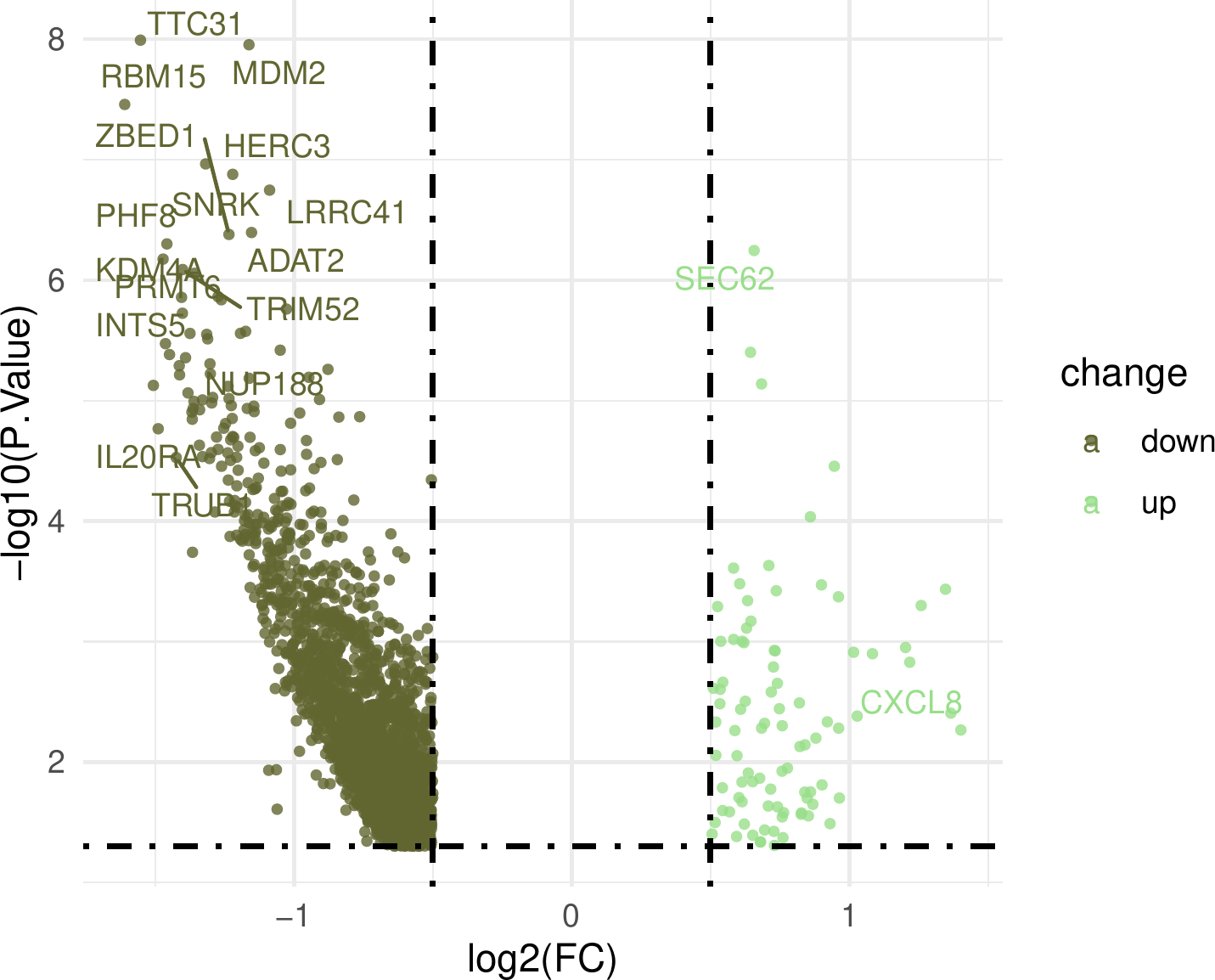
# 3 分析结果

## 3.1 Limma 差异分析 (MRNA)

肺癌 RNA-seq，

样本分组：Advanced\_stage (n=65) , Early\_stage (n=101) , Healthy (n=81) 。 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)。 各组差异分析 DEGs 统计：

* Early\_stage vs Healthy：up (n=90) , down (n=2065) 。
* Advanced\_stage vs Healthy：up (n=1035) , down (n=2283) 。
* Advanced\_stage vs Early\_stage：up (n=2225) , down (n=1058) 。 所有上调 DEGs 共 2483 个，所有下调 DEGs 共 3779 个。所有非重复基因共 5497 个。 见 Fig.

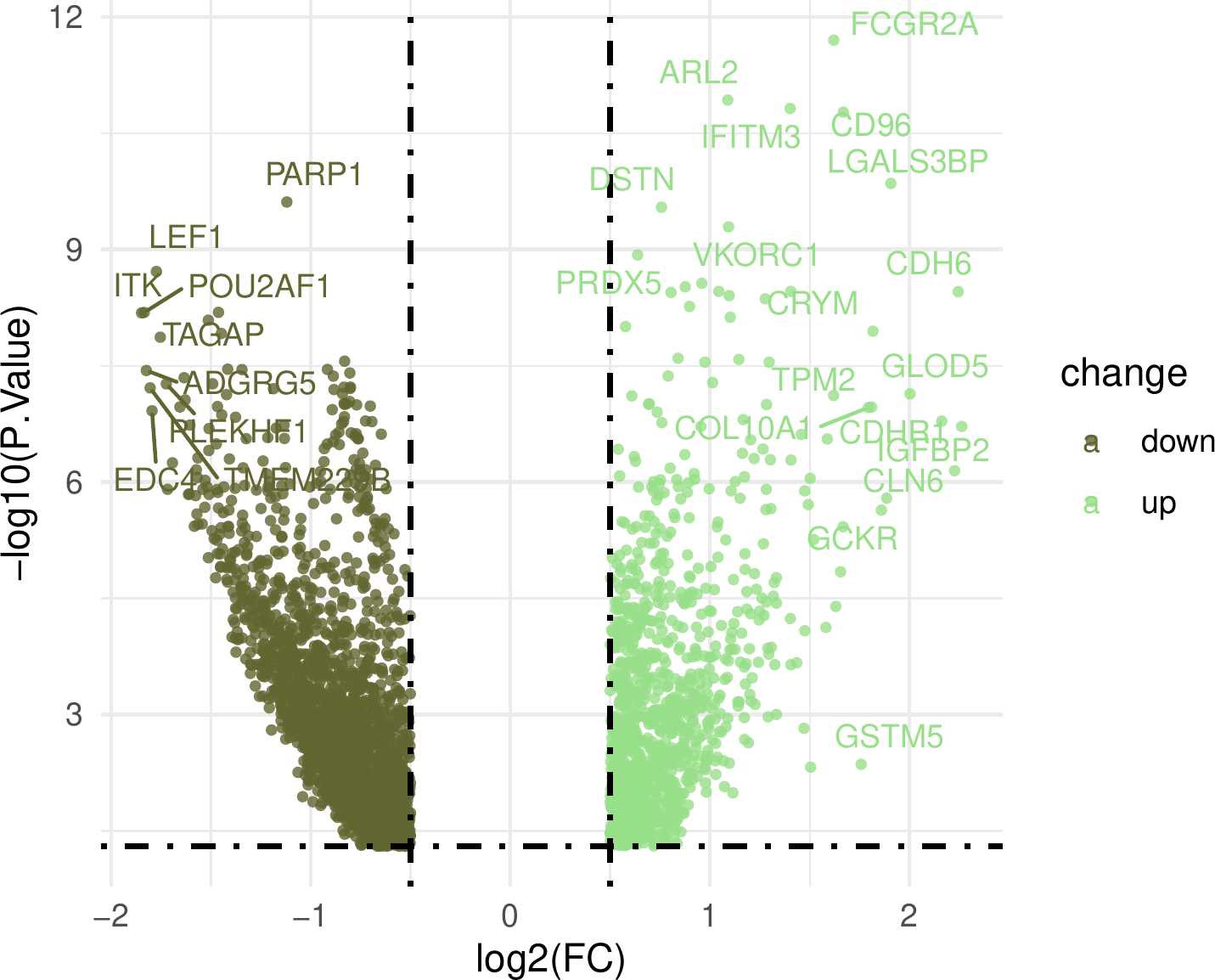


**Fig.** MRNA Early stage vs Healthy

**(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)**

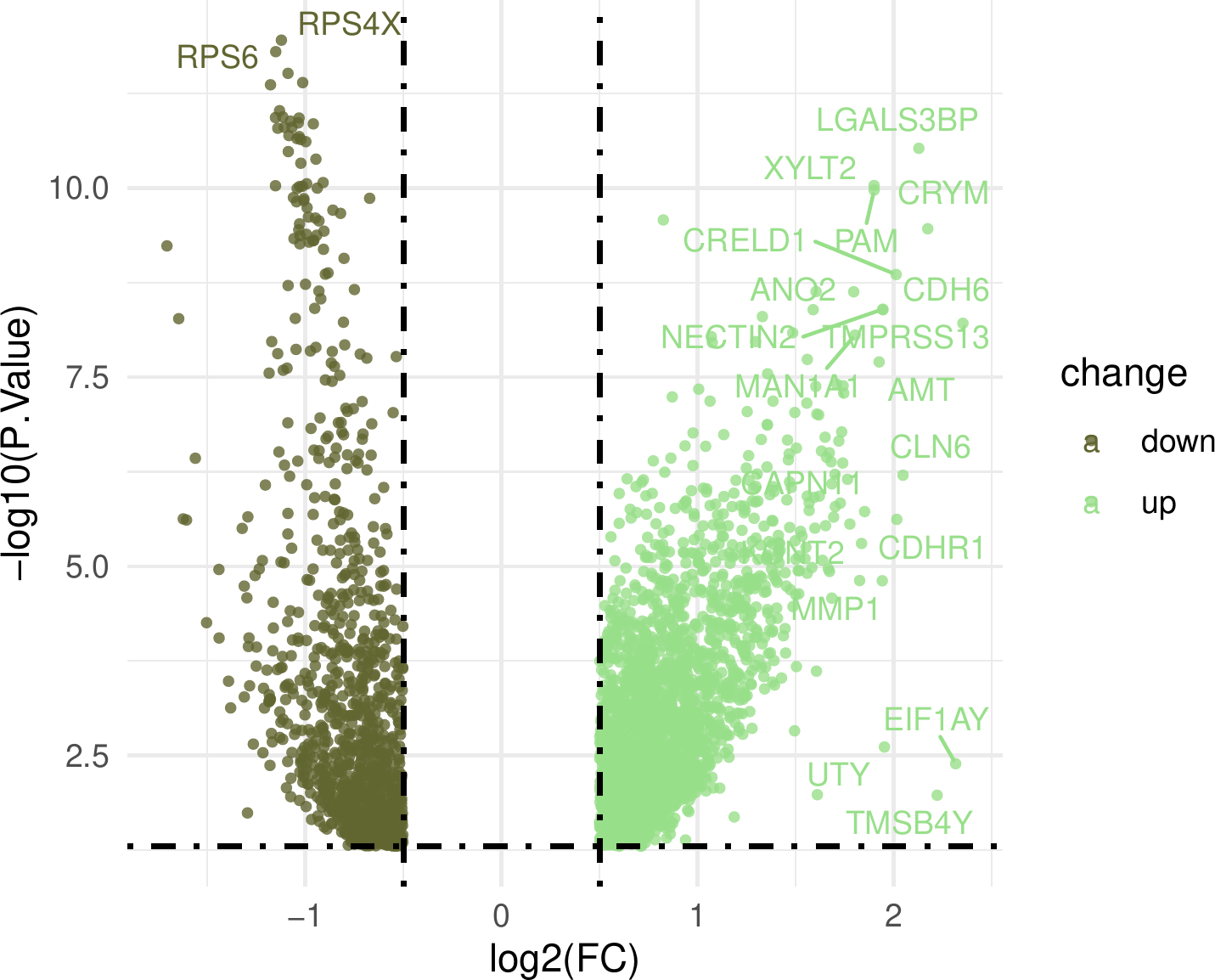


**Fig.** MRNA Advanced stage vs Healthy

**(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)**

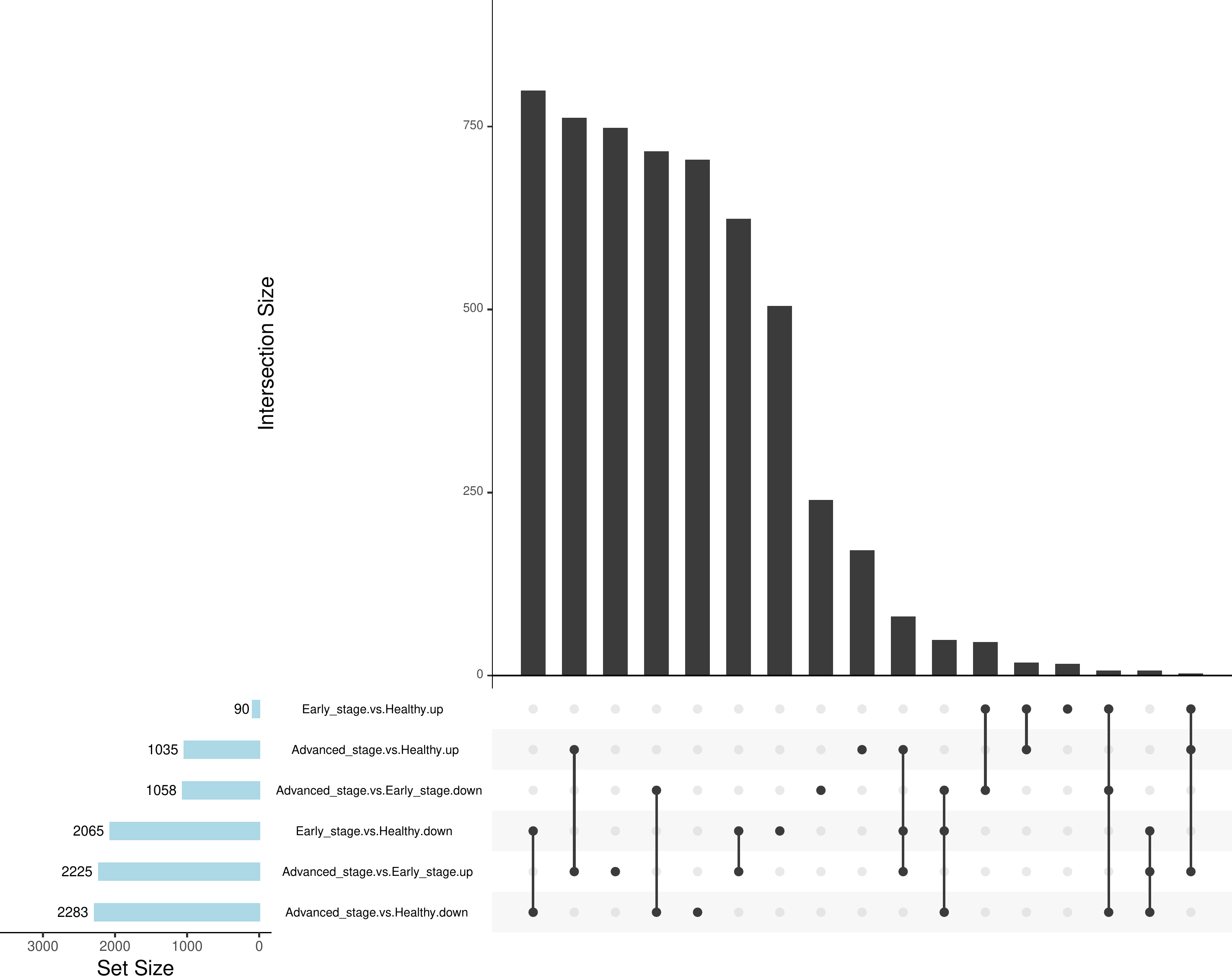


**Fig.** MRNA Advanced stage vs Early stage

**(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.** 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)**

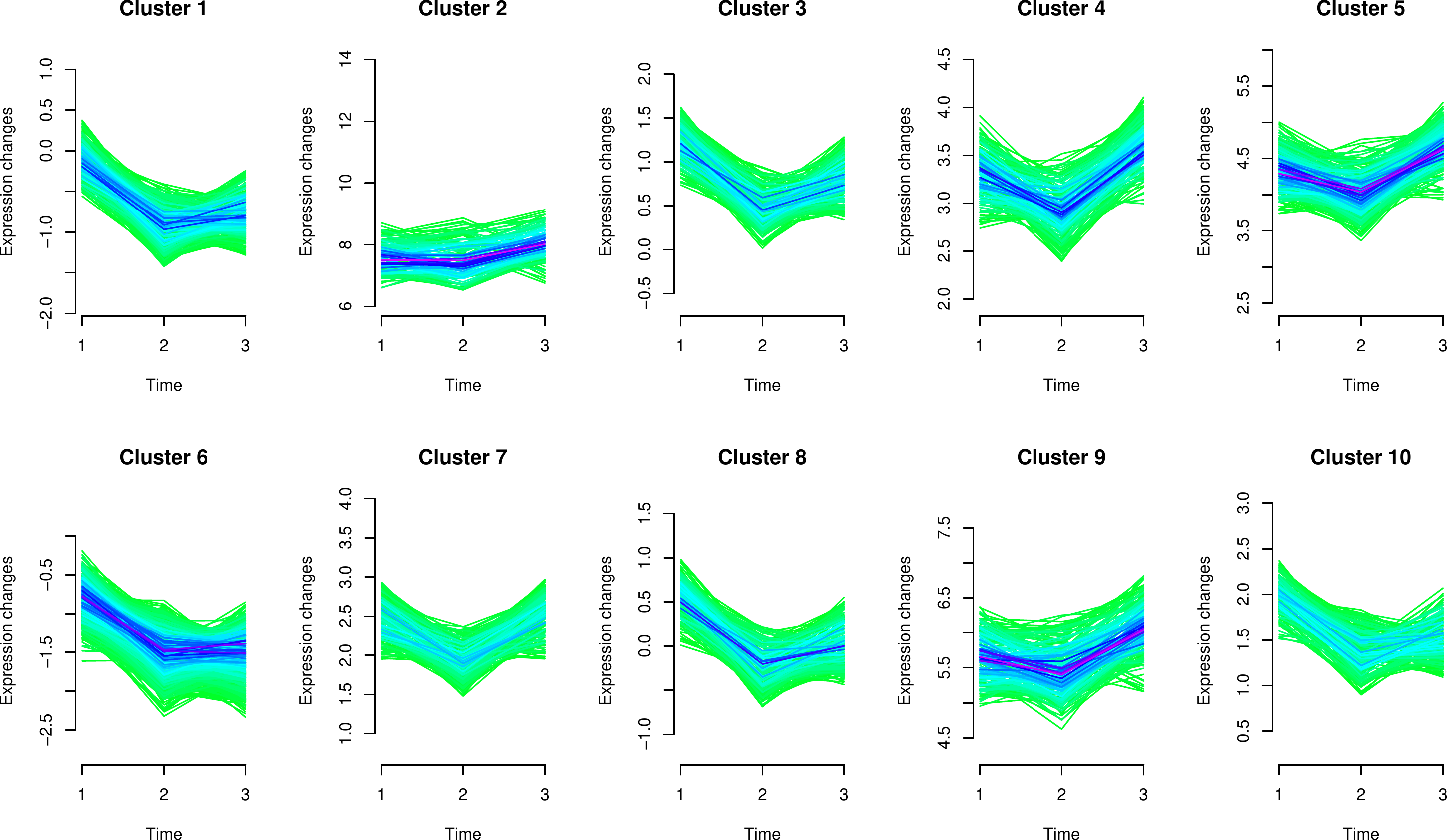
Note: The directory 'Figure+Table/MRNA-data-DEGs' contains 3 files.  
  
1 1\_Early\_stage - Healthy.csv  
2 2\_Advanced\_stage - Healthy.csv  
3 3\_Advanced\_stage - Early\_stage.csv

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

## 3.2 Mfuzz 聚类分析 (MRNA)

将所有 DEGs 以 Mfuzz 聚类分析，探究时序变化。

共得到 10 个聚类。 按照 Healthy, Early\_stage, Advanced\_stage 顺序, 在 Mfuzz 聚类中：2 为按时序上调，共 217 个；1, 6 为按时序下调，共 1285 个；其他 DEGs 为离散变化。 见 Fig.



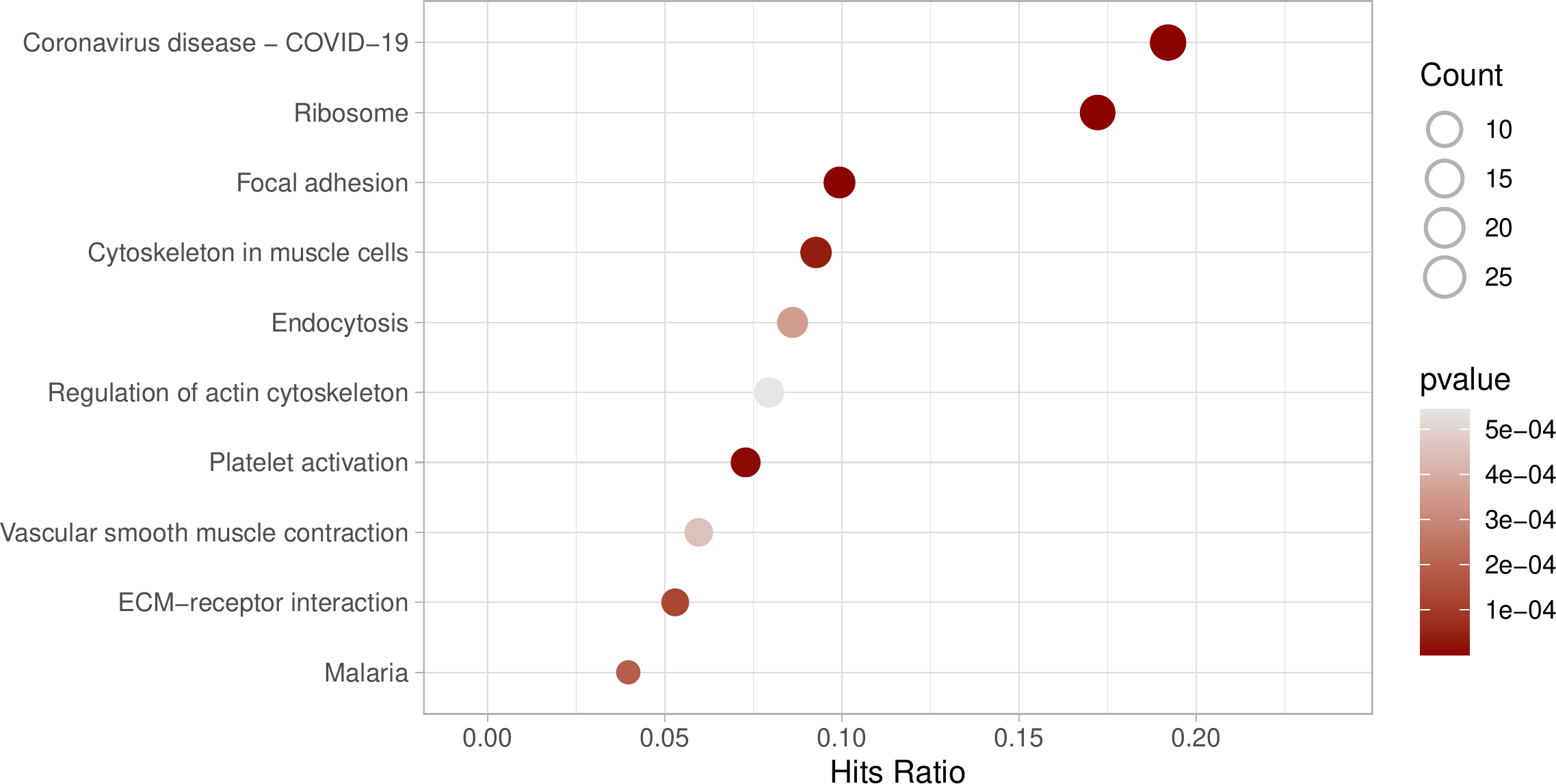
**Fig.** MRNA Mfuzz clusters

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

## 3.3 富集分析 (MRNA)

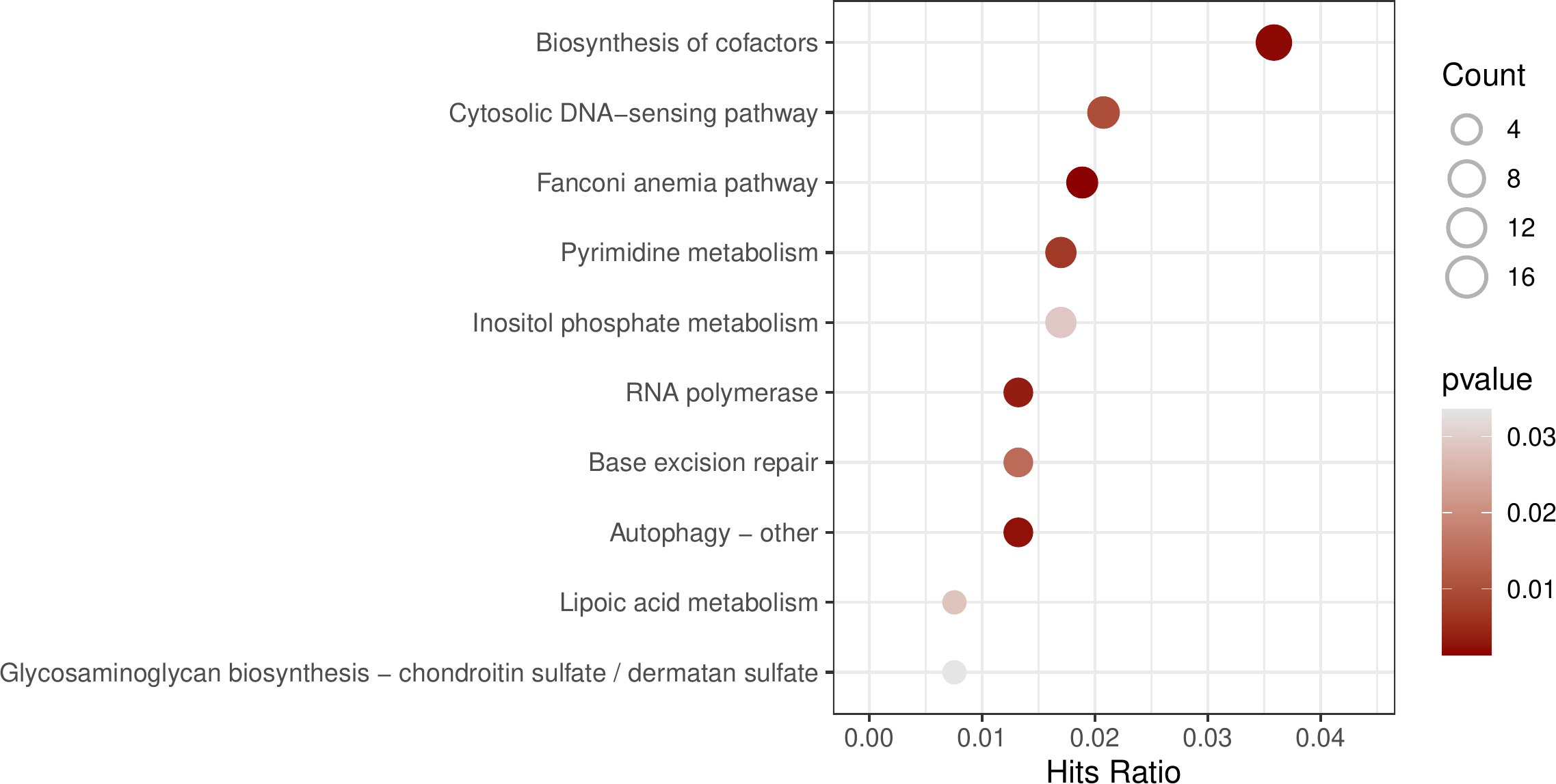
KEGG 见 Fig. , Fig. GO 见 Fig. Fig. 。

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 Fig. 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。



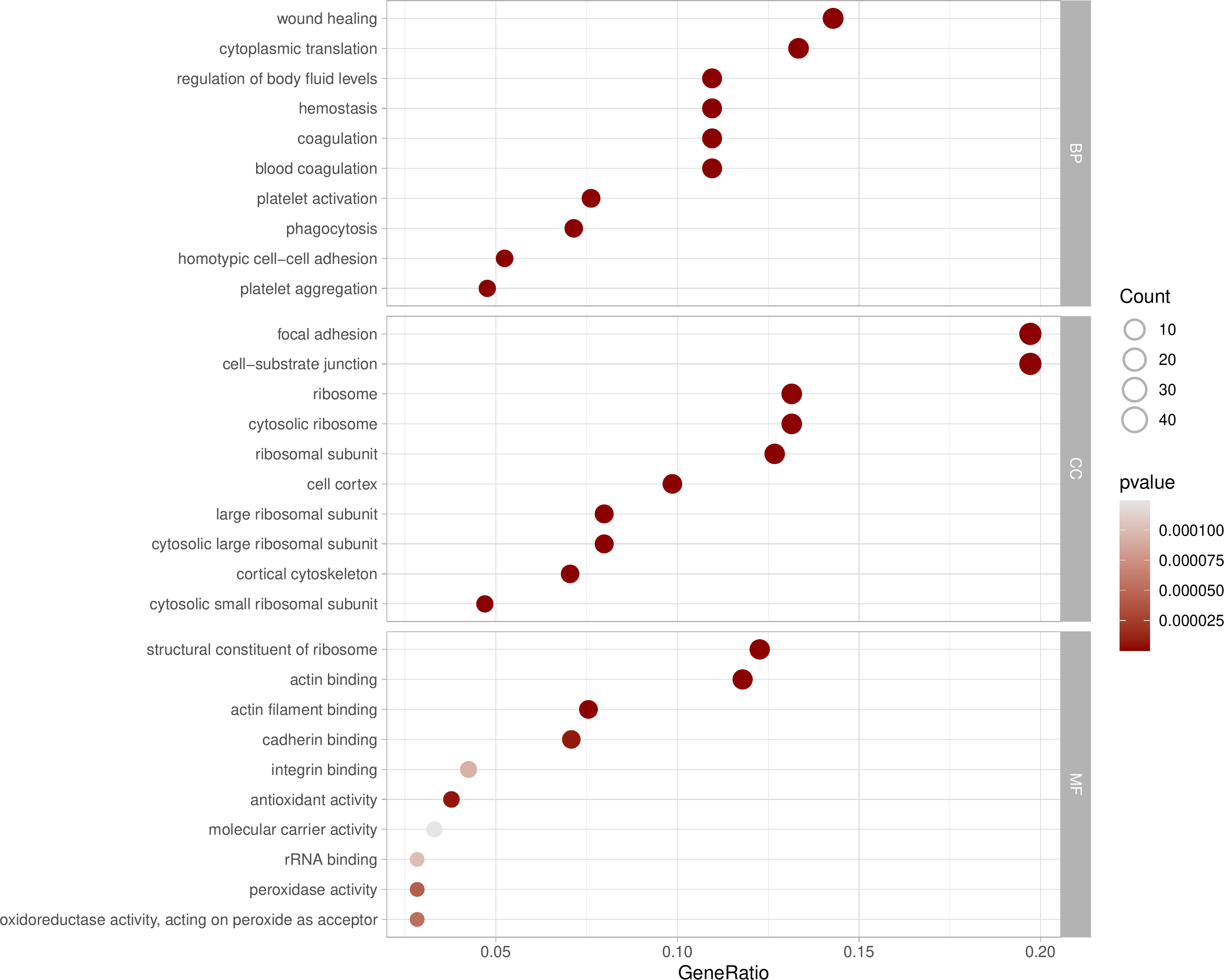
**Fig.** MRNA ups KEGG enrichment

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



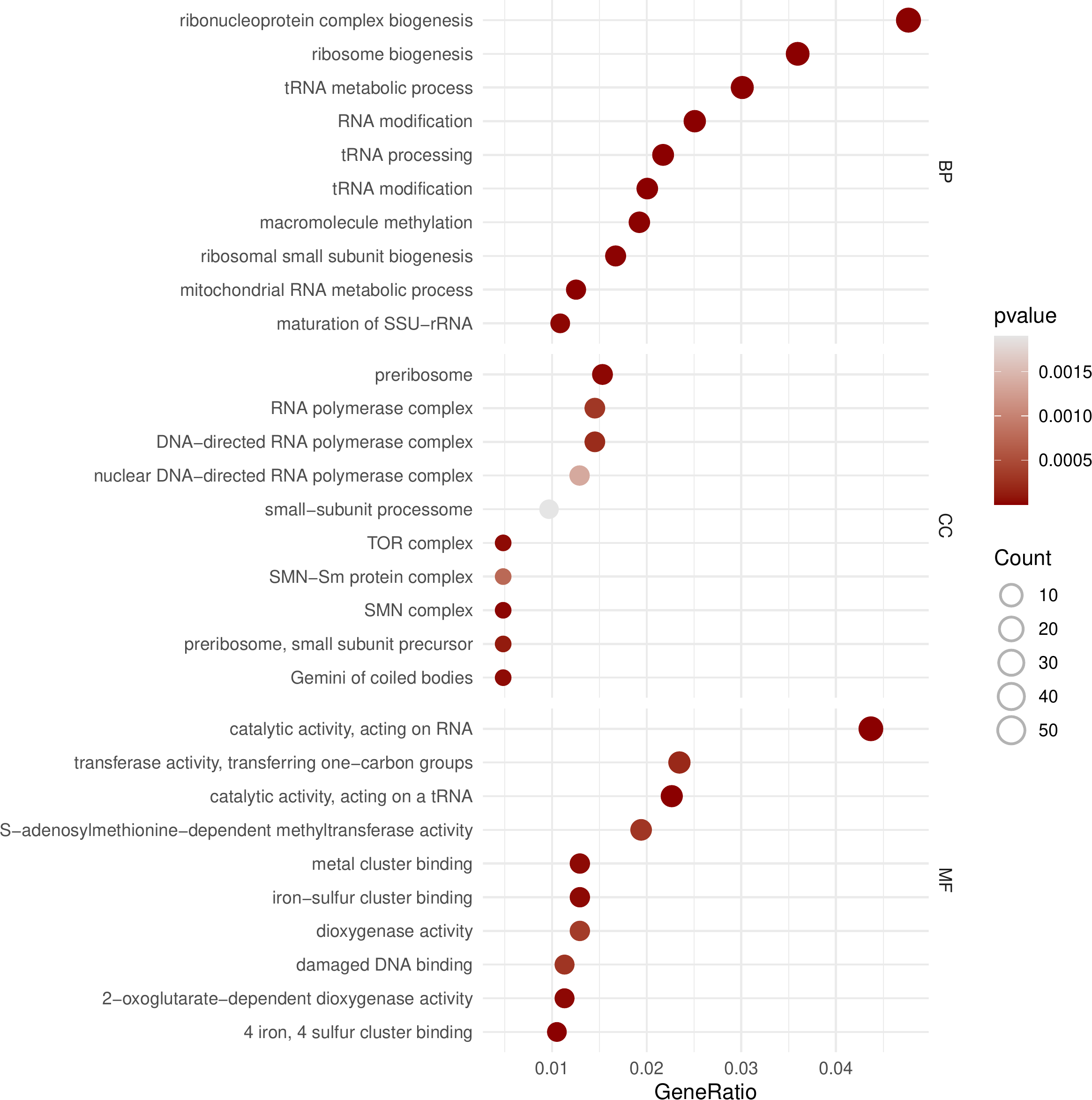
**Fig.** MRNA downs KEGG enrichment

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



**Fig.** MRNA ups GO enrichment

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



**Fig.** 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)

将 TCGA.LUAD (n=589) , TCGA.LUSC (n=501) 数据集合并。

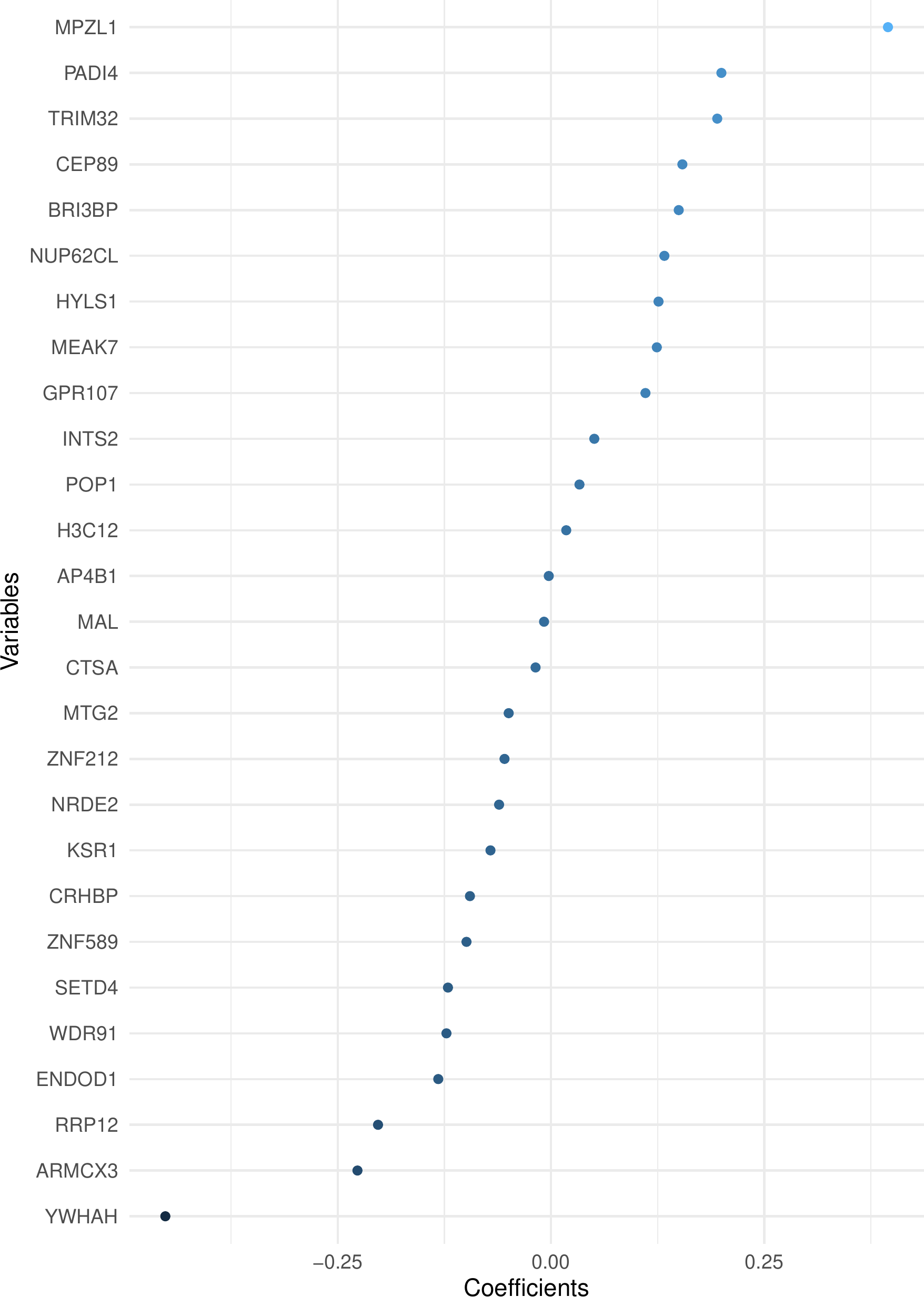
## 3.5 COX 回归 (LUNG)

筛选 ajcc\_pathologic\_stage 为 Stage IB, Stage IA, Stage IIB, Stage IIA, Stage II, Stage I, 筛选 days\_to\_last\_follow\_up 大于或等于 30, 筛选 isTumor 为 tumor，最终得到 615 例数据。样本分组：Alive (n=499) , Dead (n=105) 。 共 1470 个基因在数据集 TCGA-LUSC, TCGA-LUAD 中找到 (根据基因名匹配)。以 TCGA (dataset: LUNG) 的基因表达数据 (经过标准化，见方法章节)，将 Mfuzz (dataset: MRNA) 聚类结果中的聚类 2, 1, 6 (up, down DEGs, n=1502) 以 COX 回归分析。所有数据生存状态 (去除生存状态未知的数据)，(Alive (n=499) , Dead (n=105) )。执行单因素 COX 回归，筛选 P 值 < 0.05，共筛选到 61 个基因。在单因素回归得到的基因的基础上，使用 glmnet::cv.glmnet 作 5 倍交叉验证 (评估方式为 Partial Likelihood Deviance)，筛选 lambda 值。lambda.min, lambda.1se 值分别为 0.014, 0.051 (R 随机种子为 987456)。对应的特征数 (基因数) 分别为 27, 0。



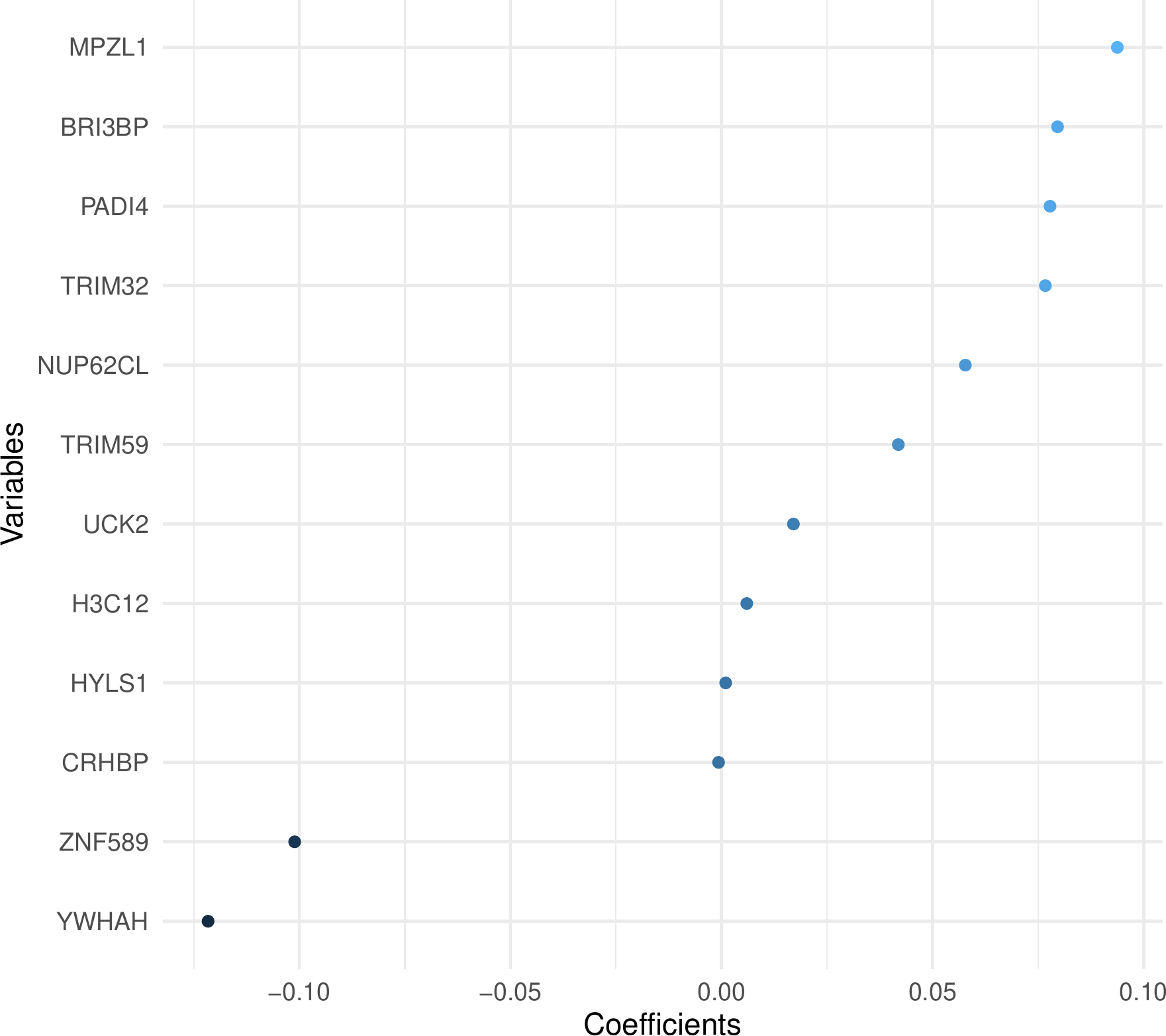
**Fig.** LUNG lasso COX model

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



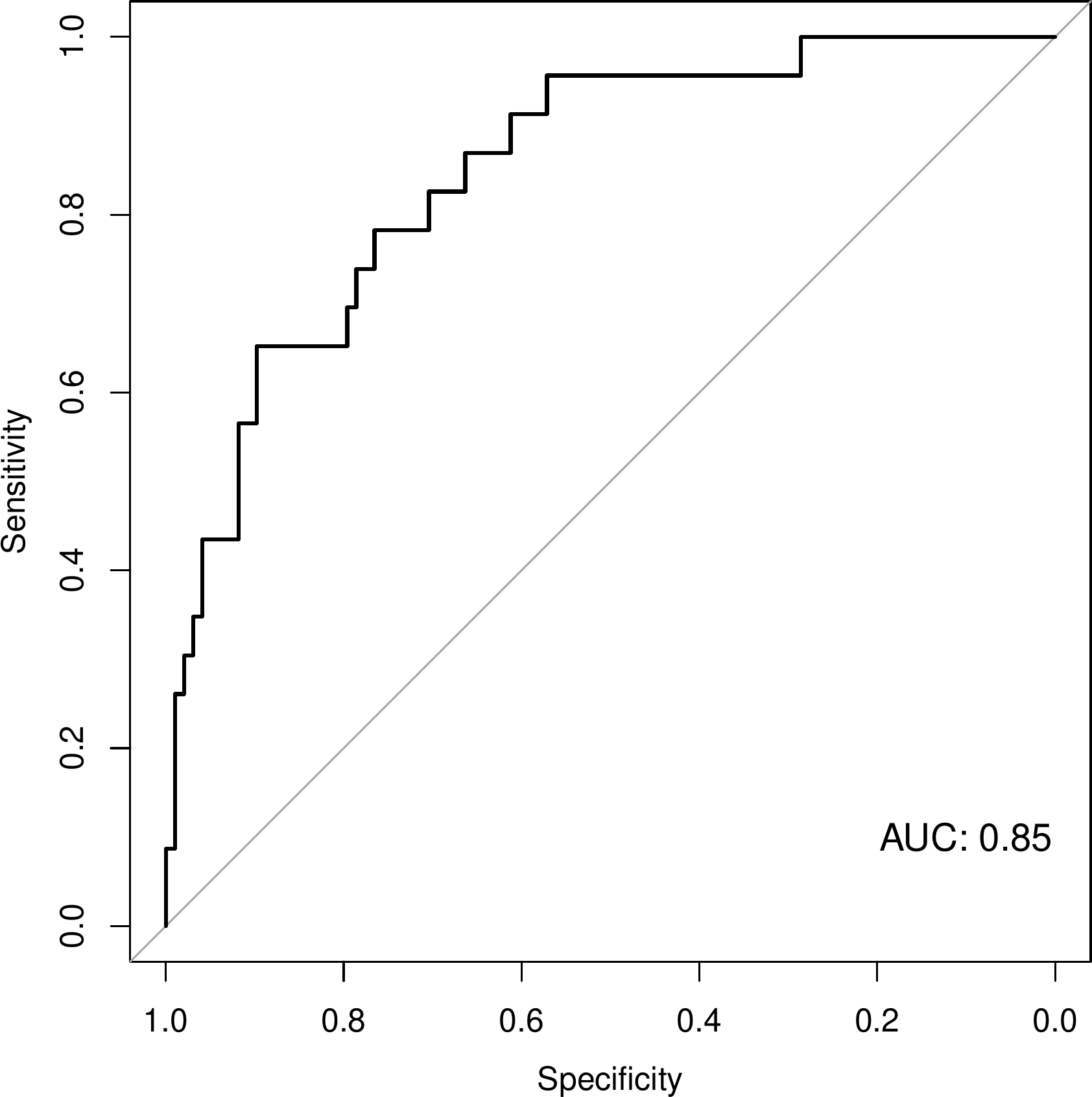
**Fig.** LUNG lasso COX coeffients lambda min

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



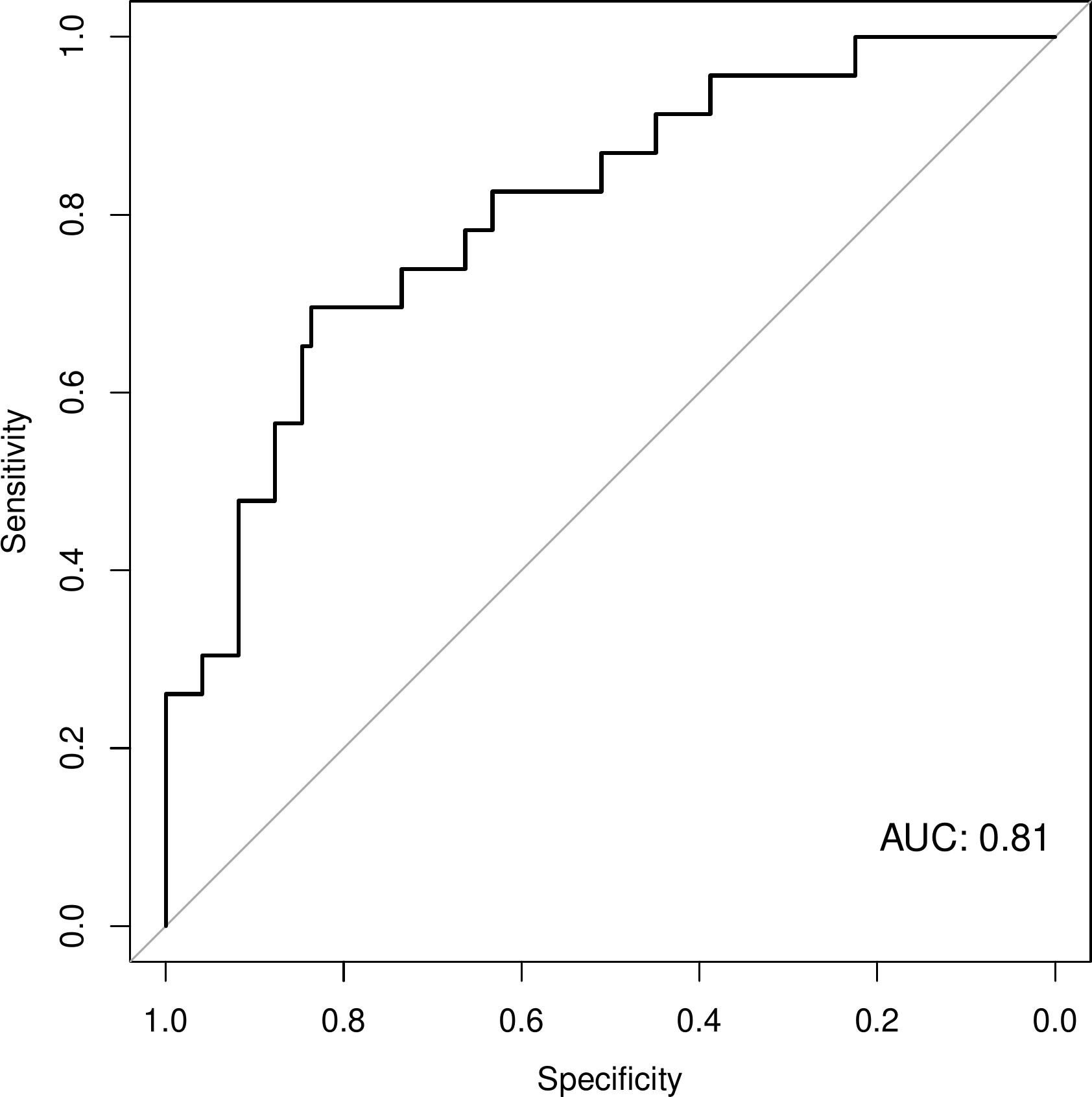
**Fig.** LUNG lasso COX coeffients lambda 1se

**(File path: Figure+Table/3.5\_COX\_回归\_(LUNG)/LUNG-lasso-COX-coeffients-lambda-1se.pdf)**



**Fig.** LUNG lasso COX ROC lambda min

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



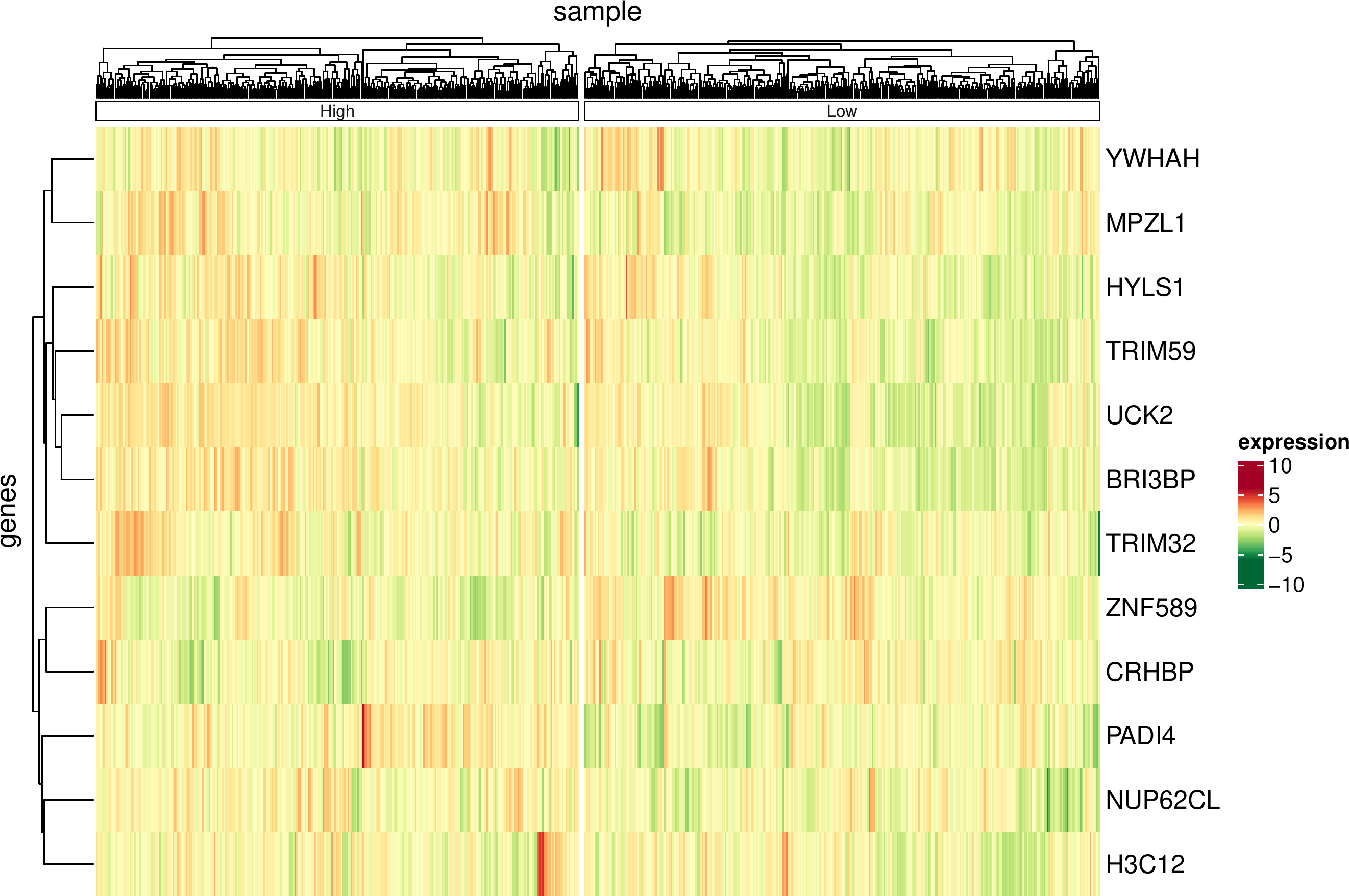
**Fig.** LUNG lasso COX ROC lambda 1se

**(File path: Figure+Table/3.5\_COX\_回归\_(LUNG)/LUNG-lasso-COX-ROC-lambda-1se.pdf)**

## 3.6 Survival 生存分析 (LUNG)

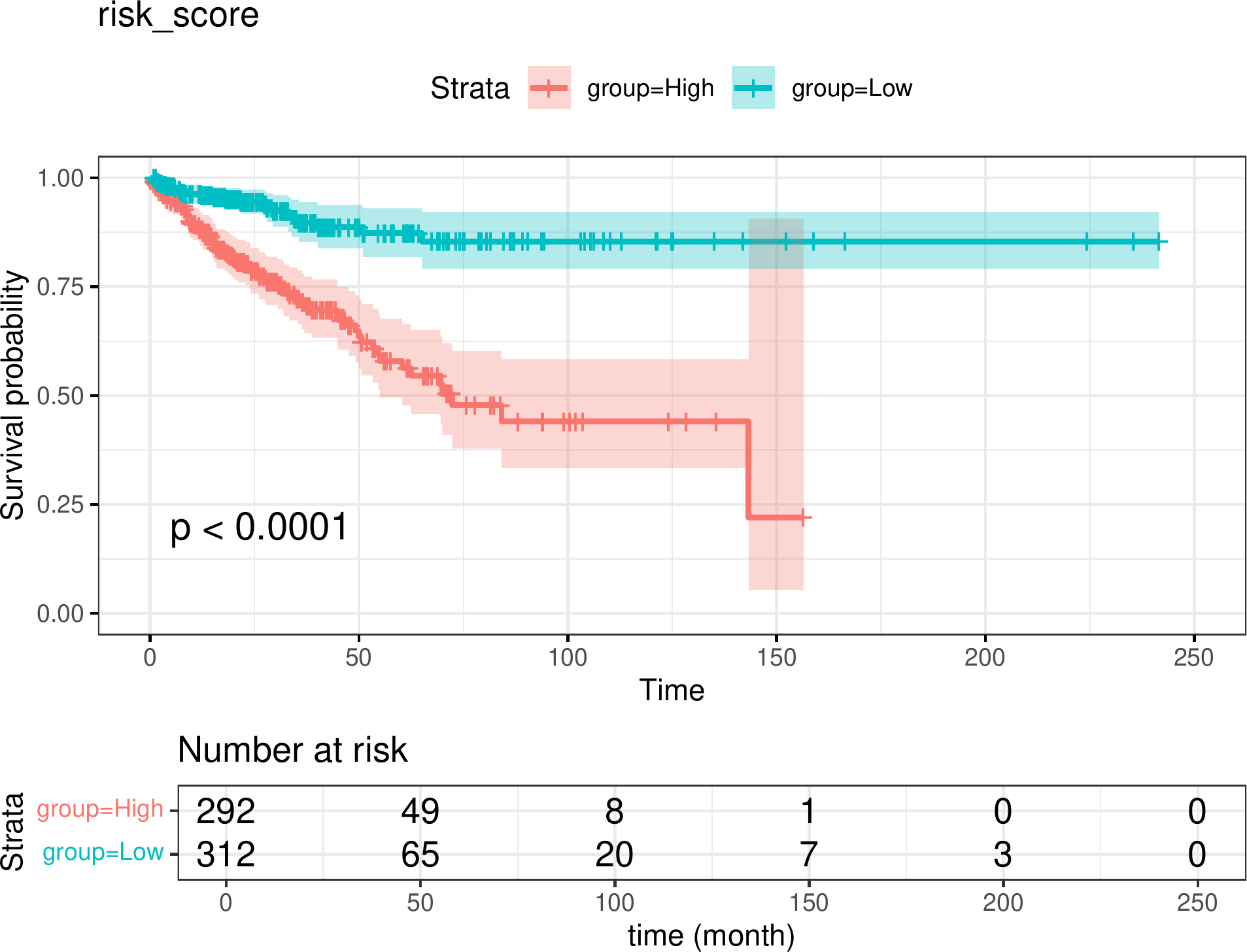
选择 lambda.1se 时得到的特征集，包含 12 个基因， 分别为: TRIM32, YWHAH, UCK2, CRHBP, PADI4, ZNF589, BRI3BP, H3C12, MPZL1, NUP62CL, HYLS1, TRIM59。以回归系数构建风险评分模型。

按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 0.287250134868389) (High (n=292) , Low (n=312) )， 随后进行生存分析。



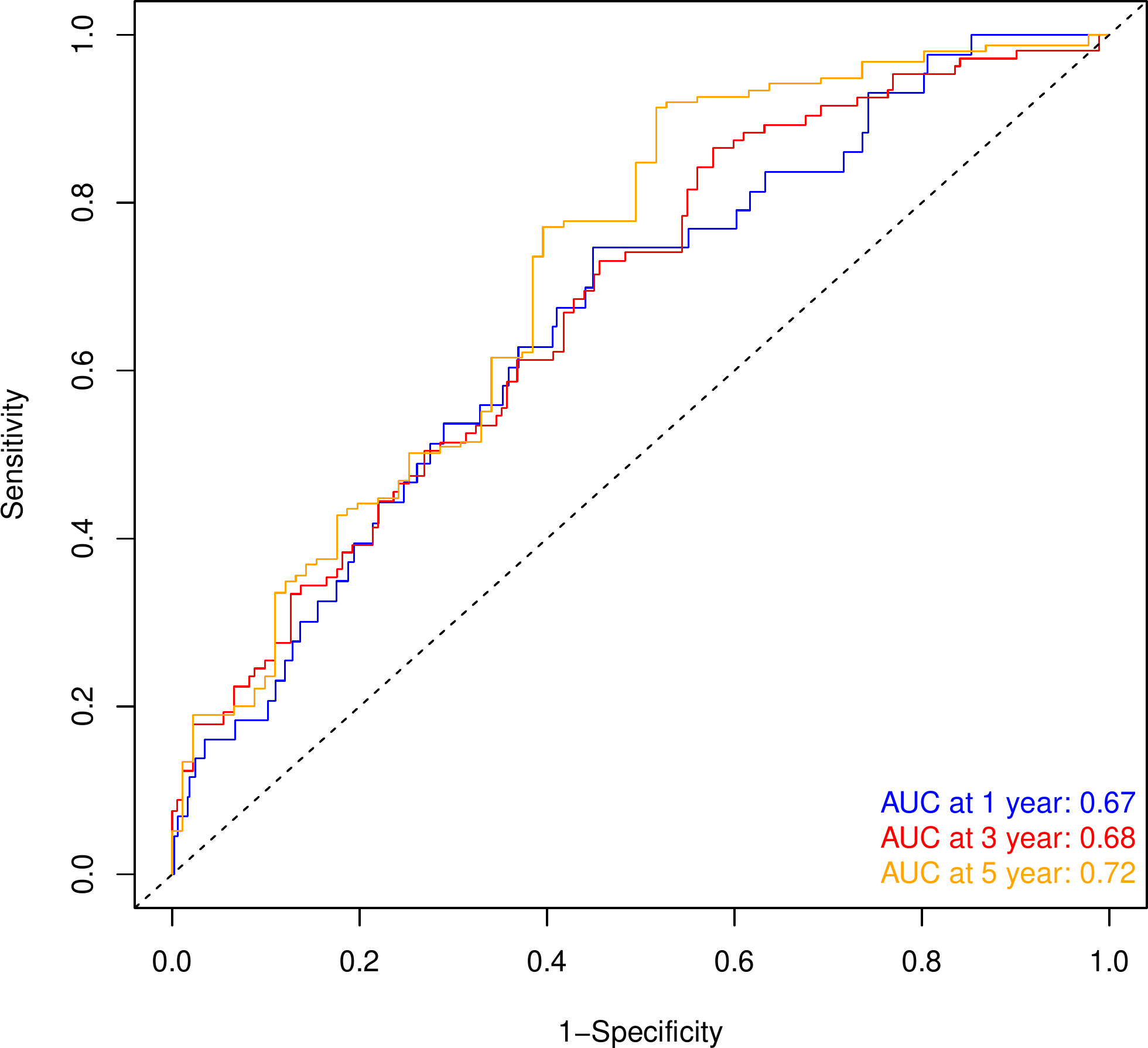
**Fig.** LUNG risk score related genes heatmap

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



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

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

## 3.7 GSE 数据搜索 (LUNG)

以 Entrez Direct (EDirect) 搜索 GEO 数据库 (检索条件见方法章节) 。筛选 gdsType 为 Expression profiling by high throughput sequencing，最终得到 25 例数据。以 GEOquery 获取 GSE 数据集，检查元数据。 GSE87340, GSE198048, GSE256047 含有生存信息。 GSE198048 不包含生存事件 (Dead) 。 因此，以 GSE87340, GSE256047 验证。

## 3.8 GEO 数据获取 (GEO\_LUNG2)

* Data Source ID: GSE256047
* data\_processing: Sequences were trimmed for adaptor sequence using Trim Galore
* data\_processing.1: Trimmed reads were mapped to GRCH37/hg19 using Salmon aligner
* data\_processing.2: Transcript abundance was caputured using tximport pipeline and data was compared to previously published cohort of HPV-positive oropharyngeal cancers
* data\_processing.3: Batch correction was performed using DESeq2
* (Others): …

**(见Figure+Table/3.8\_GEO\_数据获取\_(GEO\_LUNG2)/LUNG2-GSE256047-content)**

筛选 vital\_status 为 Alive, Dead, 筛选 days\_to\_last\_follow\_up 大于 10，最终得到 45 例数据。

## 3.9 GEO 数据获取 (GEO\_LUNG)

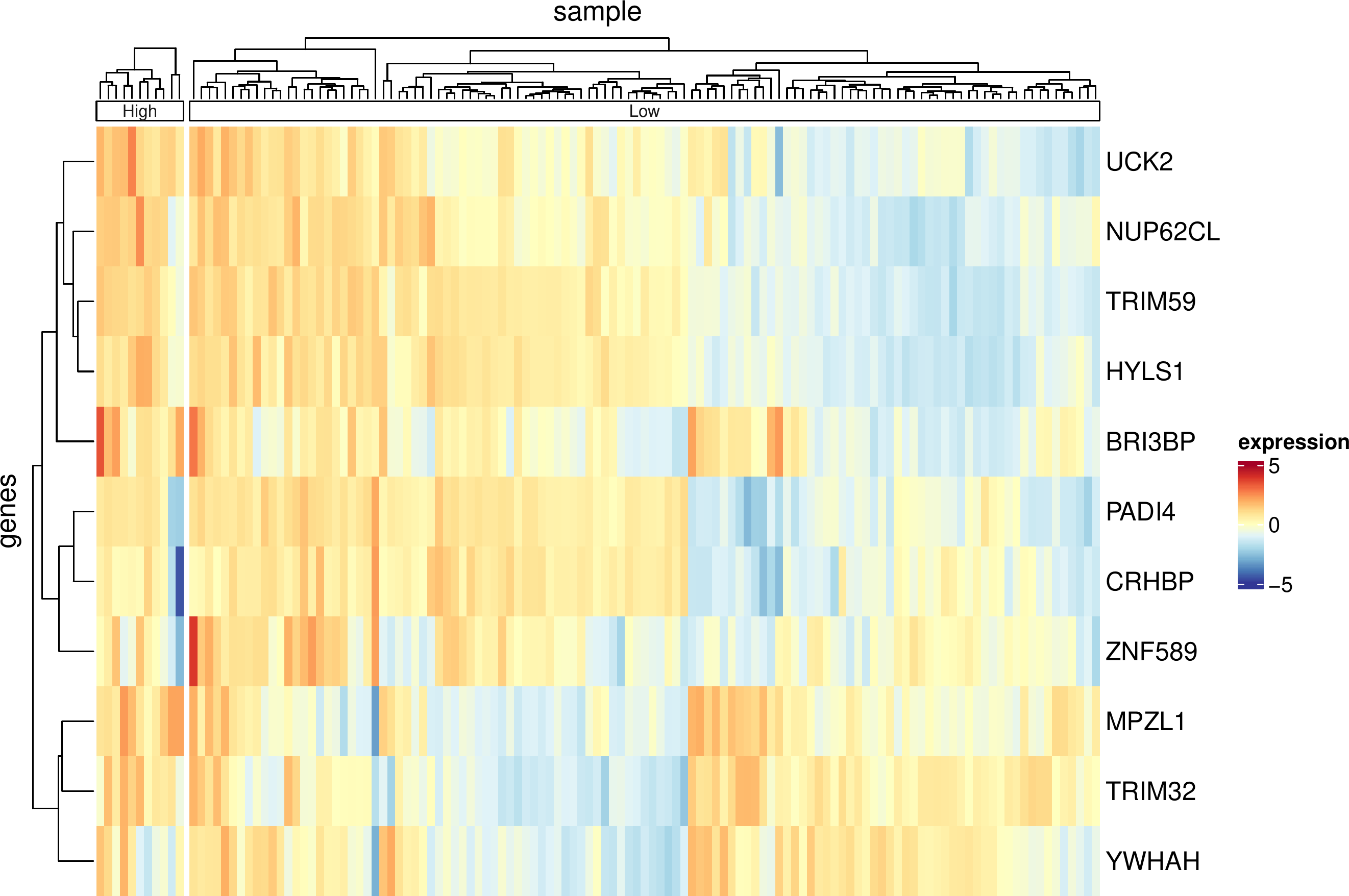
* Data Source ID: GSE87340
* data\_processing: Internal MAPRSeq pipeline (<http://bioinformaticstools.mayo.edu/research/maprseq/>)
* data\_processing.1: RPKM normalized, Log2 transformed
* data\_processing.2: Genome\_build: hg19
* data\_processing.3: Supplementary\_files\_format\_and\_content: txt, raw and normalized expression

**(见Figure+Table/3.9\_GEO\_数据获取\_(GEO\_LUNG)/LUNG-GSE87340-content)**

筛选 days\_to\_last\_follow\_up 大于 10，未过滤掉任何数据 (n = 54)。样本分组：Alive (n=40) , Dead (n=14) 。

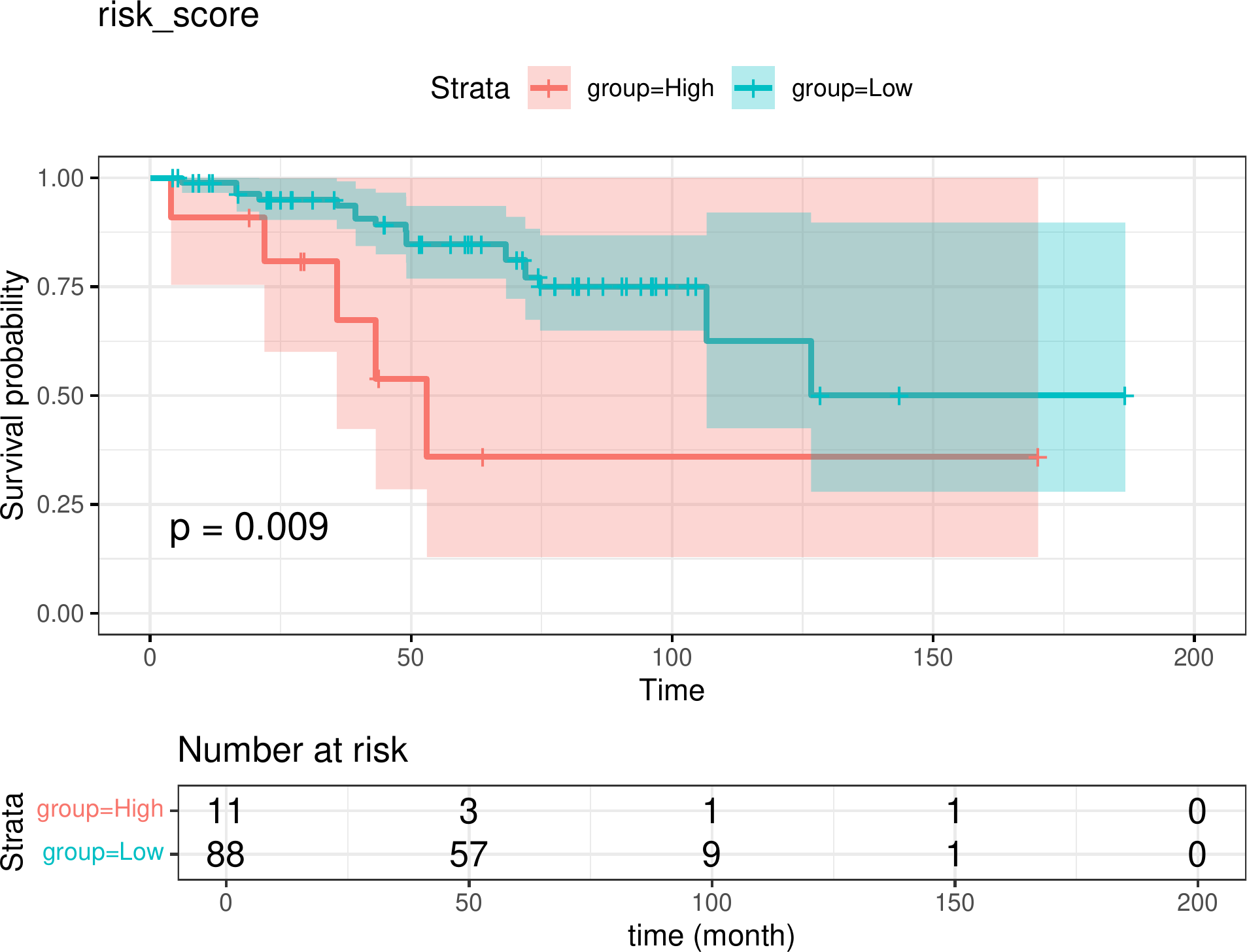
## 3.10 Survival 生存分析 (GEO\_LUNG) (验证)

将数据集合并 () 。所有数据生存状态，(Alive (n=76) , Dead (n=23) )。 将风险评分模型运用于数据集 (dataset: GEO\_LUNG)。该数据集 (dataset: GEO\_LUNG) 不包含基因：H3C12，后续计算评分时去除。按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 0.355654110070428) (High (n=11) , Low (n=115) )， 随后进行生存分析。



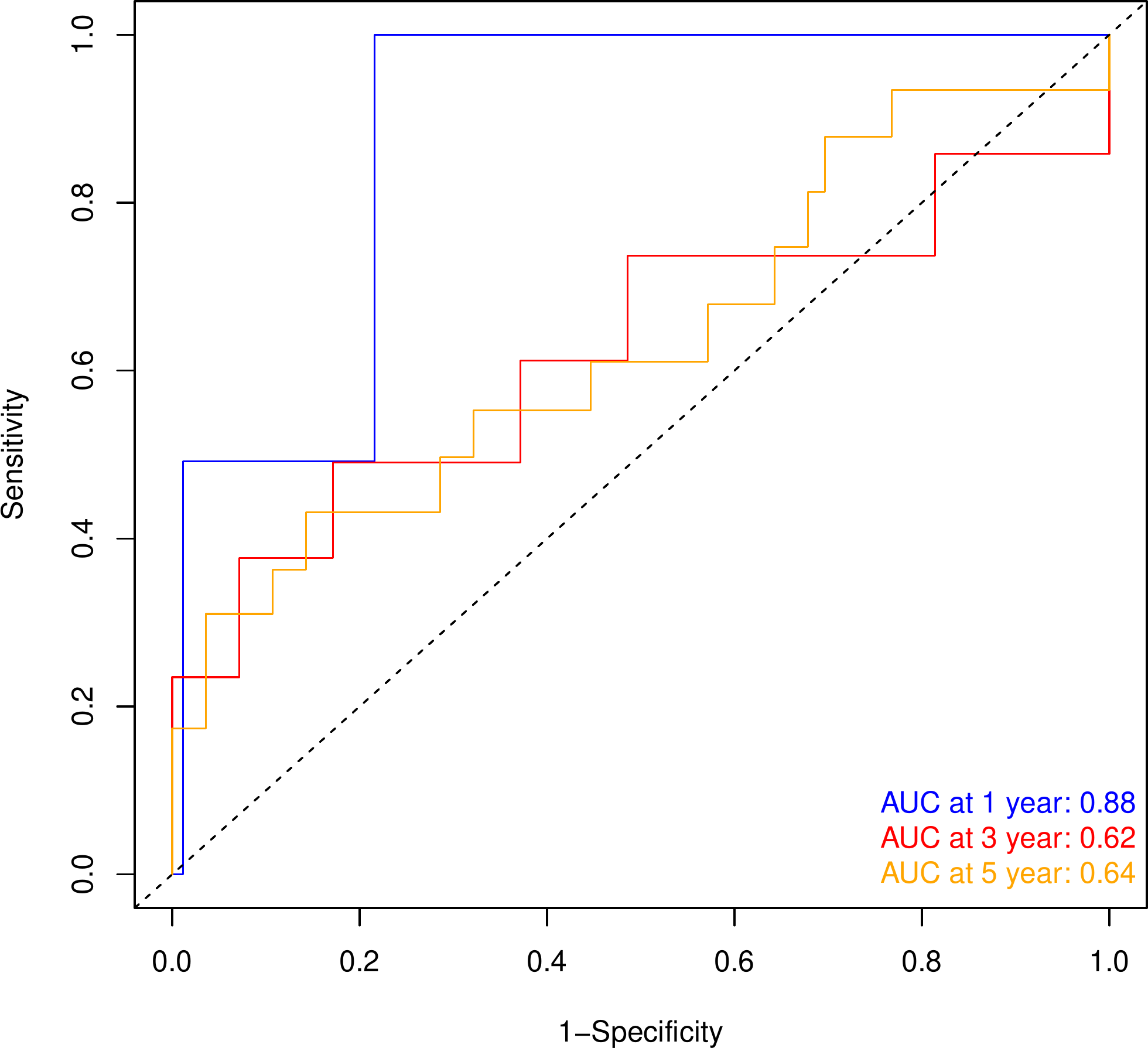
**Fig.** GEO LUAD risk score related genes heatmap

**(File path: Figure+Table/3.10\_Survival\_生存分析\_(GEO\_LUNG)\_(验证)/GEO-LUAD-risk-score-related-genes-heatmap.pdf)**



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

**(File path: Figure+Table/3.10\_Survival\_生存分析\_(GEO\_LUNG)\_(验证)/GEO-LUAD-survival-curve-of-risk-score.pdf)**



**Fig.** GEO LUAD time ROC

**(File path: Figure+Table/3.10\_Survival\_生存分析\_(GEO\_LUNG)\_(验证)/GEO-LUAD-time-ROC.pdf)**

## 3.11 Limma 差异分析 (LNCRNA)

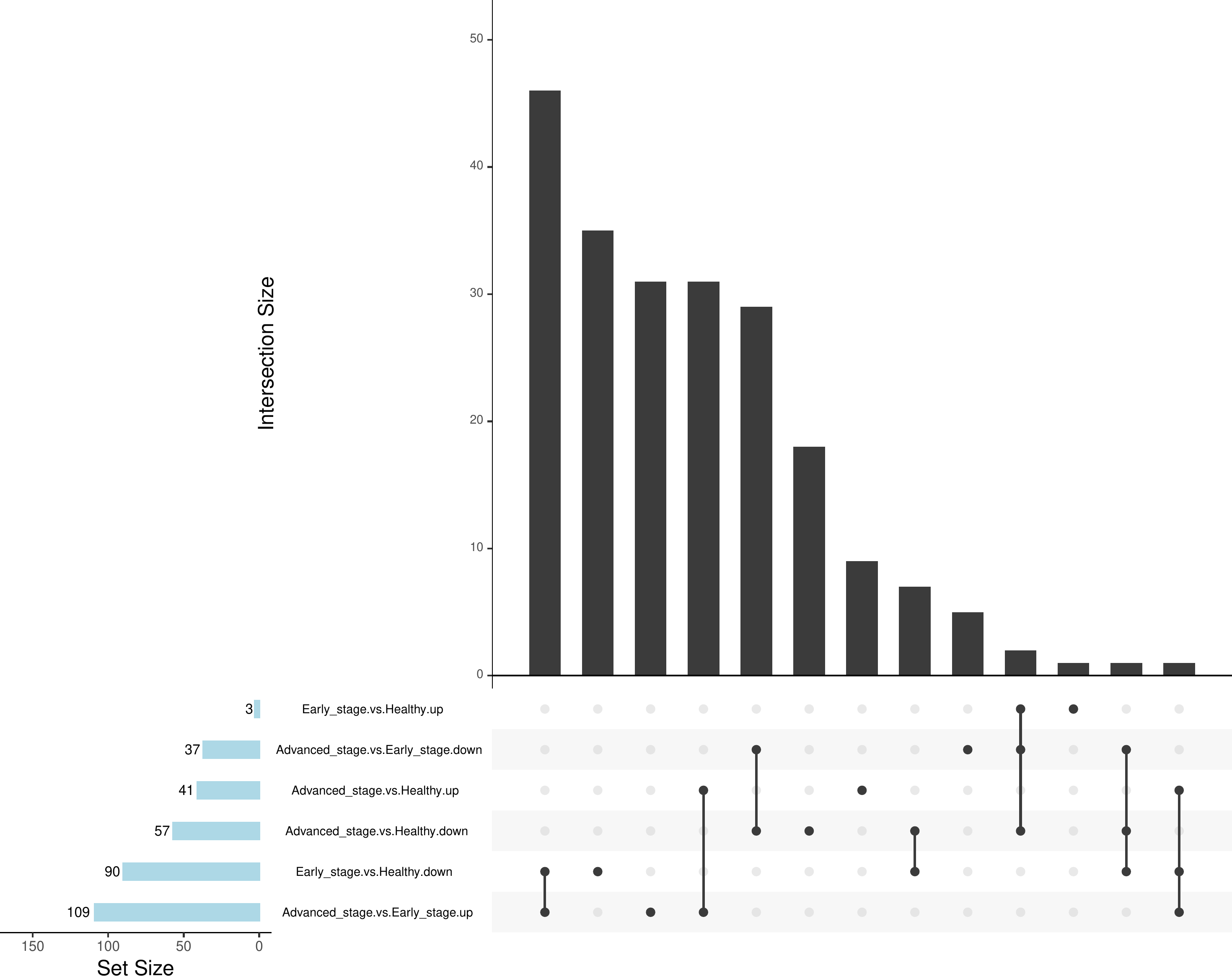
长链非编码RNA（lncRNA）在基因调控和癌症发展中起着重要作用。

样本分组：Advanced\_stage (n=65) , Early\_stage (n=101) , Healthy (n=81) 。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)。各组差异分析 DEGs 统计：

* Early\_stage vs Healthy：up (n=3) , down (n=90) 。
* Advanced\_stage vs Healthy：up (n=41) , down (n=57) 。
* Advanced\_stage vs Early\_stage：up (n=109) , down (n=37) 。 所有上调 DEGs 共 121 个，所有下调 DEGs 共 144 个。所有非重复基因共 216 个。

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/3.11\_Limma\_差异分析\_(LNCRNA)/LNCRNA-DEGs-data)**



**Fig.** LNCRNA Difference intersection

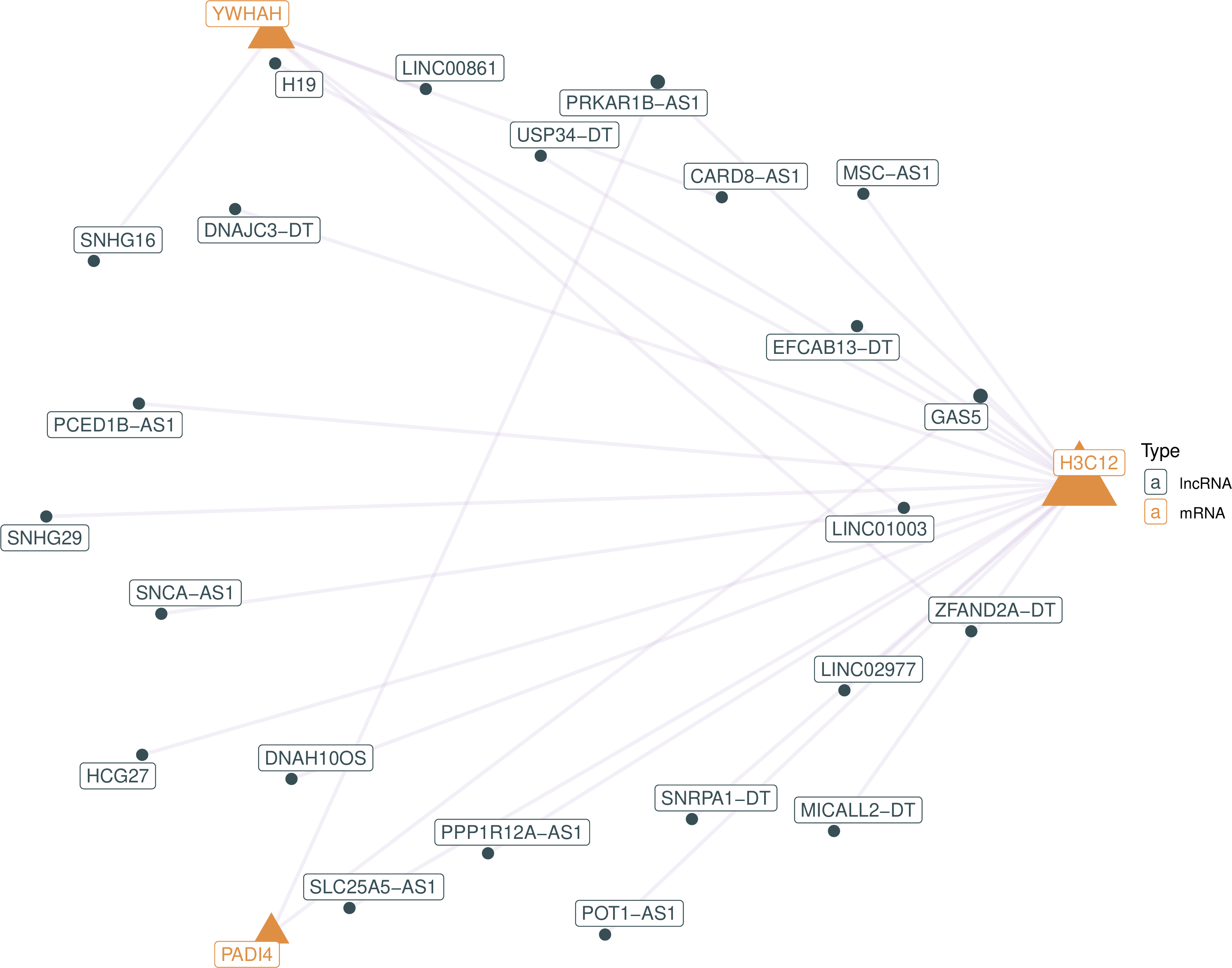
**(File path: Figure+Table/3.11\_Limma\_差异分析\_(LNCRNA)/LNCRNA-Difference-intersection.pdf)**

* All\_intersection:

**(See: Figure+Table/3.11\_Limma\_差异分析\_(LNCRNA)/LNCRNA-Difference-intersection-content)**

## 3.12 关联分析 (MRNA, LNCRNA)

将两组相同样品来源的数据集 (dataset: MRNA, LNCRNA)) 关联分析。将基因集 (A -> B) (A:12, B:216) 关联分析，共得到 1221 个显著的基因对 (P < 0.05)。 筛选 |cor| 大于 0.4, 筛选 pvalue 小于 0.001，最终得到 25 例数据。



**Fig.** Significant Correlation mrna lncRNA

**(File path: Figure+Table/3.12\_关联分析\_(MRNA,\_LNCRNA)/Significant-Correlation-mrna-lncRNA.pdf)**

**Tab.** Significant correlation

| MRNA | LncRNA | Cor | Pvalue | -log2(... | Signif... | Sign | |cor| |
| --- | --- | --- | --- | --- | --- | --- | --- |
| H3C12 | USP34-DT | 0.41 | 0 | 16.61 | < 0.001 | \*\* | 0.41 |
| YWHAH | CARD8-AS1 | -0.58 | 0 | 16.61 | < 0.001 | \*\* | 0.58 |
| H3C12 | EFCAB1... | 0.44 | 0 | 16.61 | < 0.001 | \*\* | 0.44 |
| YWHAH | LINC01003 | 0.61 | 0 | 16.61 | < 0.001 | \*\* | 0.61 |
| H3C12 | LINC02977 | 0.46 | 0 | 16.61 | < 0.001 | \*\* | 0.46 |
| ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.12\_关联分析\_(MRNA,\_LNCRNA)/Significant-correlation.csv)**

## 3.13 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 TRIM32, YWHAH, UCK2, CRHBP, PADI4, ZNF589, BRI3BP, H3C12, MPZL1, NUP62CL, HYLS1, TRIM59， 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，因为训练于 TCGA 的 RNA-seq 数据集。 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，本次验证尽可能涵盖了整个 GEO 数据库。

# 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).