**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**

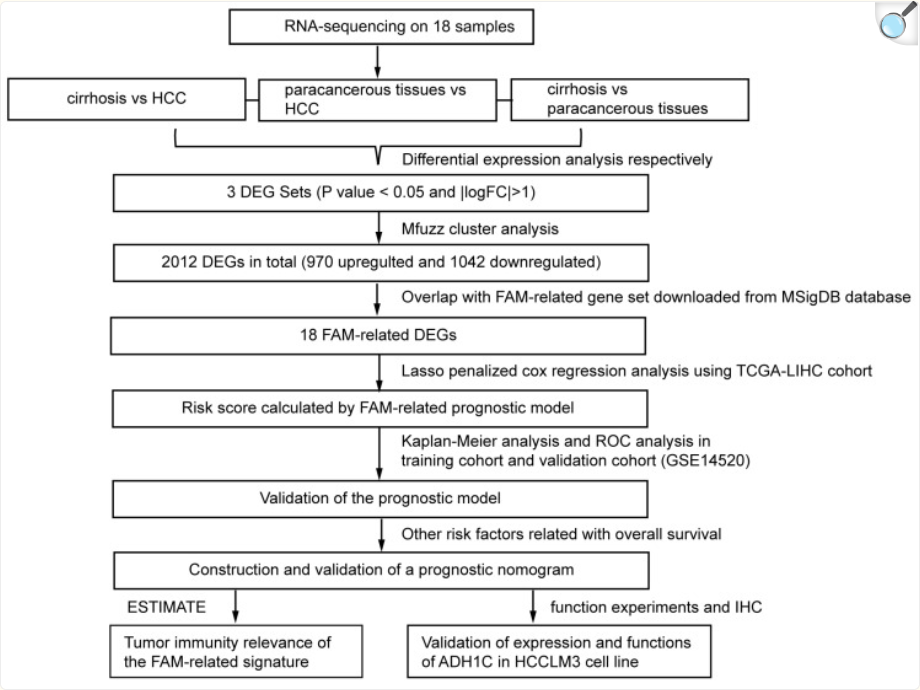
**分析类型：补充分析**

**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。



# 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 过于 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>。 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 adj.P.Val 小于 0.05，|Log2(FC)| 大于 1 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 3.73540696993324 (以 Mfuzz::mestimate 预估) ，得到 8 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。

以 R 包 EFS (1.0.3) (2017, **IF:4**, Q1, BioData Mining)7 筛选关键基因。 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。

数据源自 TCGA-LUSC，筛选 AJCC Stage (ajcc\_pathologic\_stage) 为 Stage I, Stage II 的病人，并且 days\_to\_last\_follow\_up 大于 10 天，且为肿瘤组织的样本。

## 2.7 Survival 生存分析 (Dataset: LUSC)

将 Univariate COX 回归系数用于风险评分计算，根据中位风险评分 0.0797187407744678 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。

## 2.8 COX 回归 (Dataset: PROG)

以 R 包 survival (3.7.0) 做多因素 COX 回归 (survival::coxph)。

## 2.9 GEO 数据获取 (Dataset: LUSC)

以 R 包 GEOquery (2.74.0) 获取 GSE157010 数据集。

## 2.10 Survival 生存分析 (Dataset: GEO\_LUSC)

将 Univariate COX 回归系数用于风险评分计算，根据中位风险评分 0.0418674487761947 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。

## 2.11 estimate 免疫评分 (Dataset: LUSC)

以 R 包 estimate (1.0.13) (2013, **IF:14.7**, Q1, Nature communications)8 预测数据集的 stromal, immune, estimate 得分。 从 TISIDB (, **IF:**, , )9 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 过于 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 adj.P.Val 小于 0.05，|Log2(FC)| 大于 1 的统计结果。

# 3 分析结果

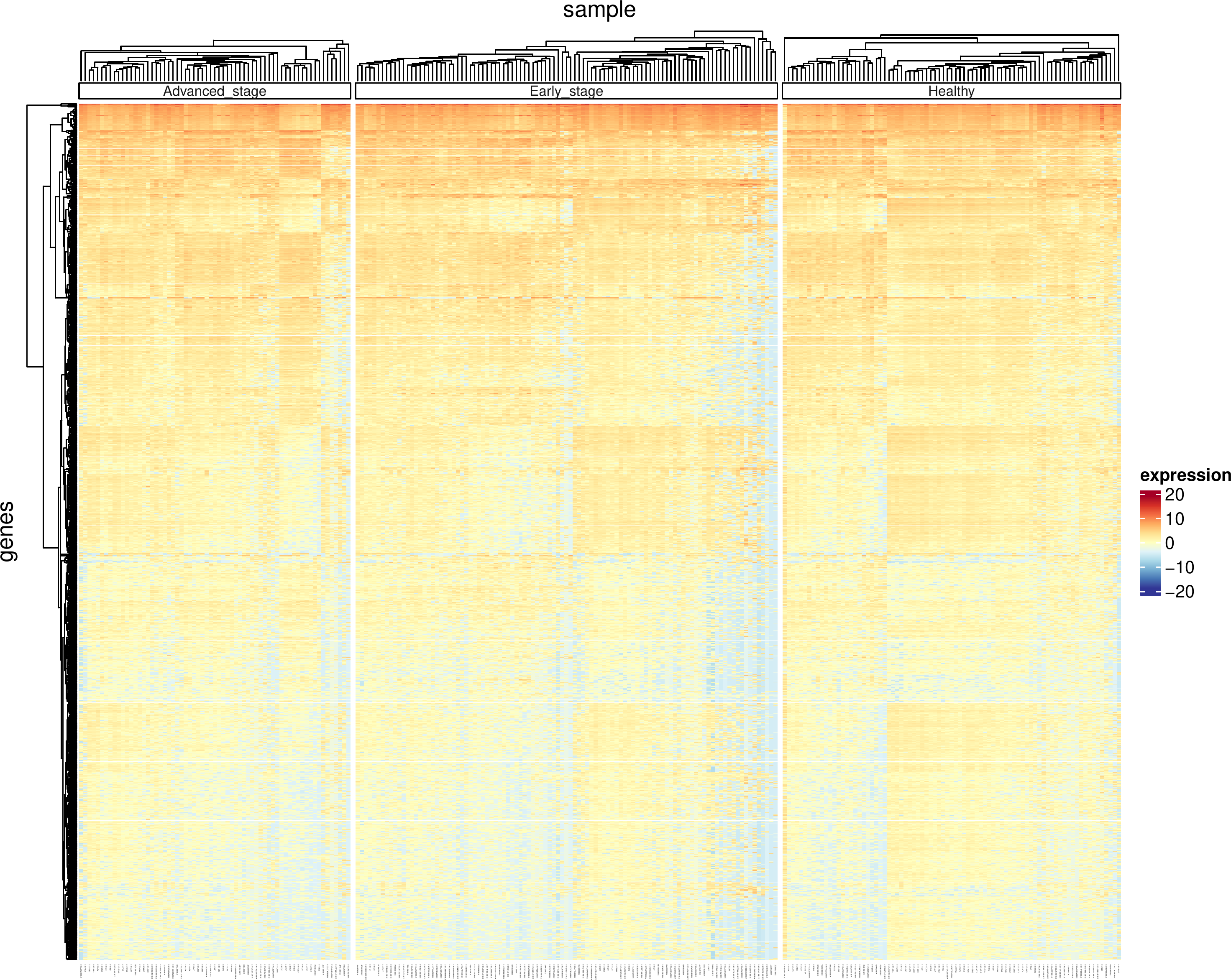
## 3.1 Limma 差异分析 (MRNA)

肝癌 RNA-seq， 共 247 个样本，分 3 组，分别为 Advanced\_stage (65) , Early\_stage (101) , Healthy (81) 。 元数据见 Tab. 。 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. 。 所有 DEGs 表达特征见 Fig. 。 所有上调 DEGs 有 539 个，下调共 781；一共 1278 个 (非重复)。

**Tab.** MRNA metadata

| Sample | Group | Lib.size | Norm.factors | Rownames | Batch |
| --- | --- | --- | --- | --- | --- |
| X180622CMQ... | Early stage | 14419487 | 1 | 180622CMQ-907 | 1806 |
| X180622HLQ... | Early stage | 13558462 | 1 | 180622HLQ-908 | 1806 |
| X180622LSF... | Early stage | 16935778 | 1 | 180622LSF-902 | 1806 |
| X180622SRD... | Early stage | 16297826 | 1 | 180622SRD-906 | 1806 |
| X180622YRQ... | Early stage | 17343112 | 1 | 180622YRQ-903 | 1806 |
| X180623WMC... | Early stage | 16088883 | 1 | 180623WMC-911 | 1806 |
| X180626XMH... | Early stage | 20035739 | 1 | 180626XMH-915 | 1806 |
| X180627CSY... | Early stage | 17851721 | 1 | 180627CSY-918 | 1806 |
| X180627XYJ... | Early stage | 18398673 | 1 | 180627XYJ-917 | 1806 |
| X180628LJH... | Early stage | 11784847 | 1 | 180628LJH-902 | 1806 |
| X180705LLF... | Early stage | 18773735 | 1 | 180705LLF-915 | 1807 |
| X180705WWP... | Early stage | 14747138 | 1 | 180705WWP-912 | 1807 |
| X180705ZQY... | Early stage | 15490310 | 1 | 180705ZQY-911 | 1807 |
| X180705ZZX... | Early stage | 18523030 | 1 | 180705ZZX-913 | 1807 |
| X180707CZM... | Early stage | 21342554 | 1 | 180707CZM-917 | 1807 |
| ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/MRNA-metadata.csv)**



**Fig.** MRNA Heatmap of DEGs

**(File path: Figure+Table/MRNA-Heatmap-of-DEGs.pdf)**

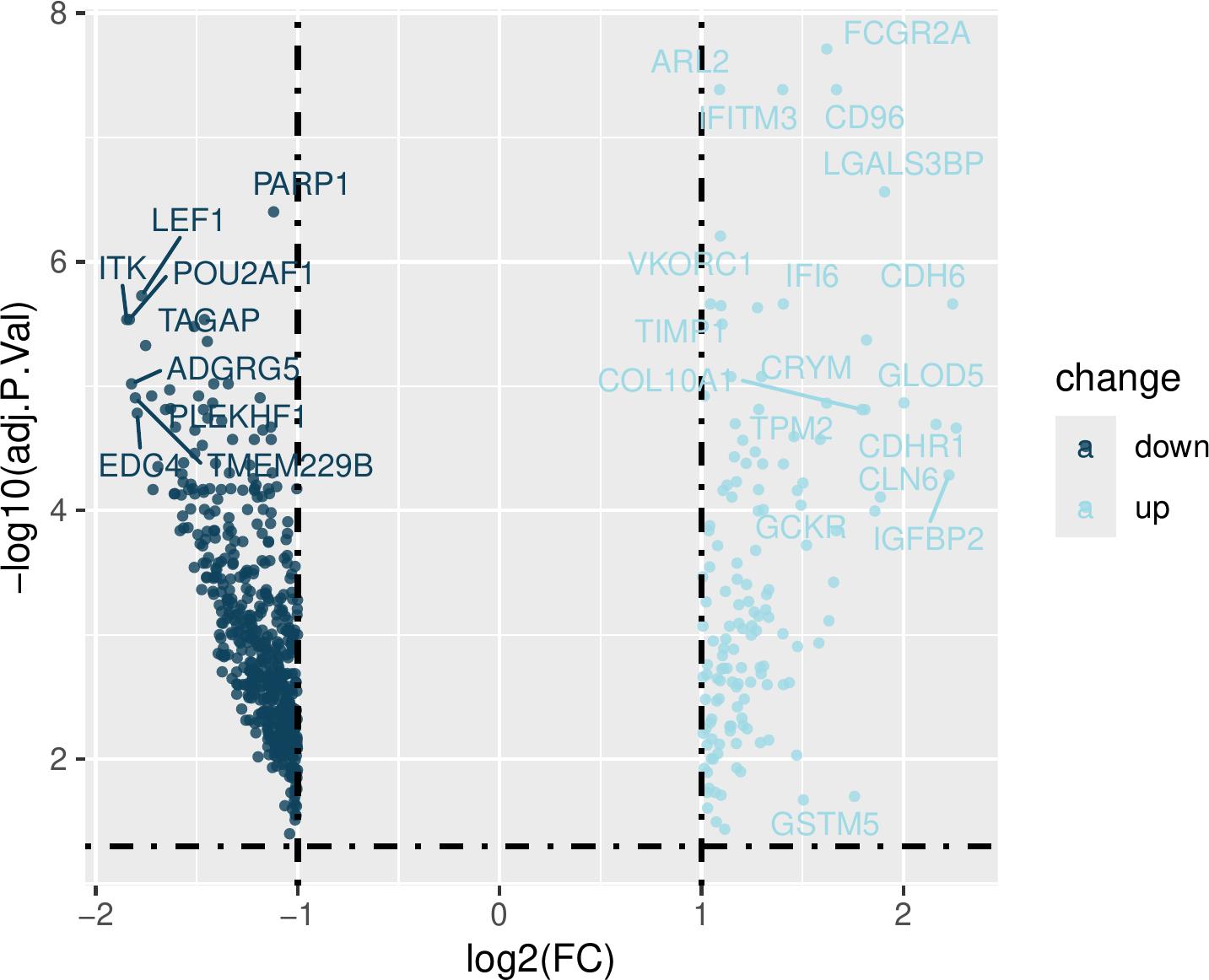


**Fig.** MRNA Early stage vs Healthy

**(File path: Figure+Table/MRNA-Early-stage-vs-Healthy.pdf)**

* adj.P.Val cut-off: 0.05
* Log2(FC) cut-off: 1

**(See: Figure+Table/MRNA-Early-stage-vs-Healthy-content)**

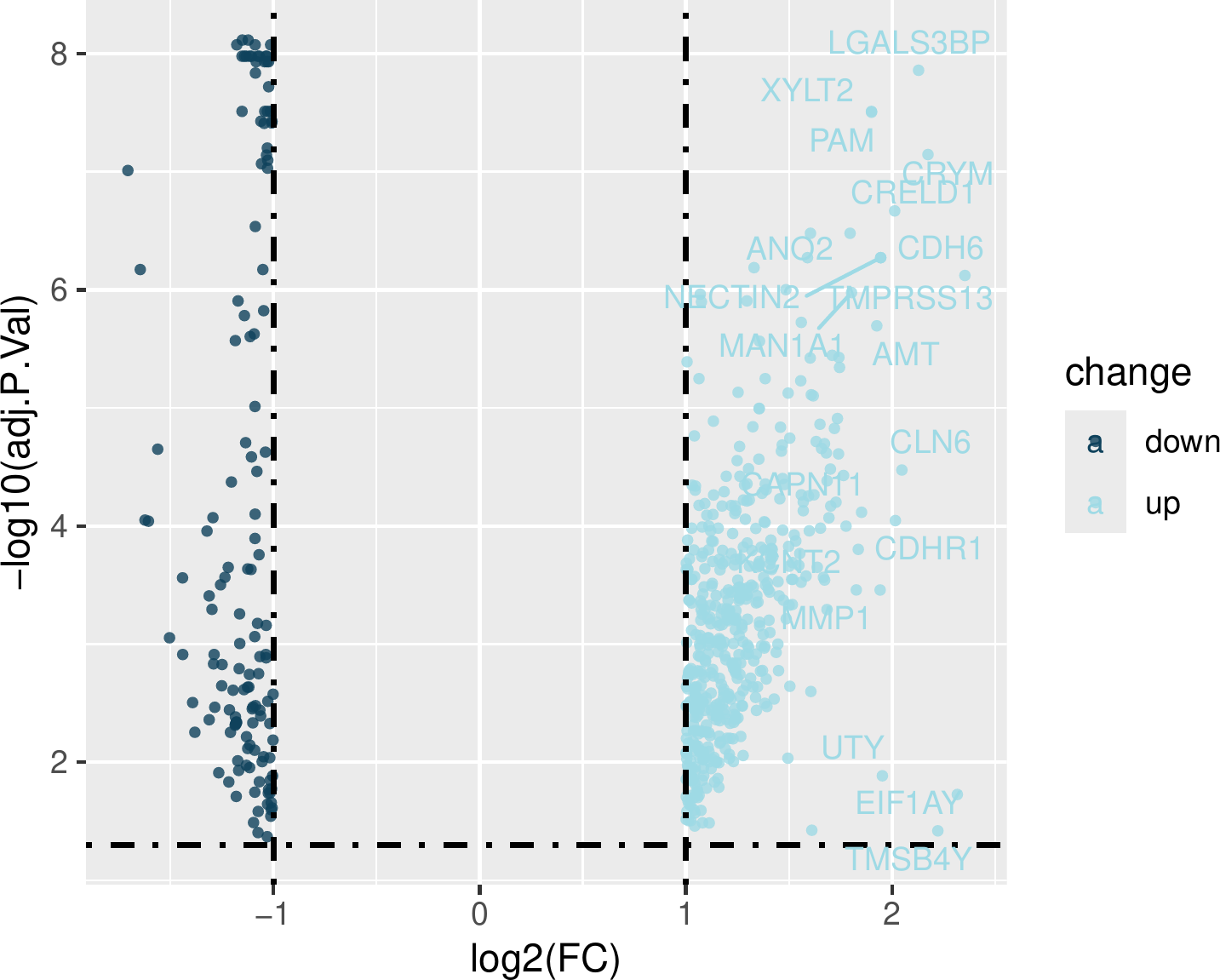


**Fig.** MRNA Advanced stage vs Healthy

**(File path: Figure+Table/MRNA-Advanced-stage-vs-Healthy.pdf)**

* adj.P.Val cut-off: 0.05
* Log2(FC) cut-off: 1

**(See: Figure+Table/MRNA-Advanced-stage-vs-Healthy-content)**

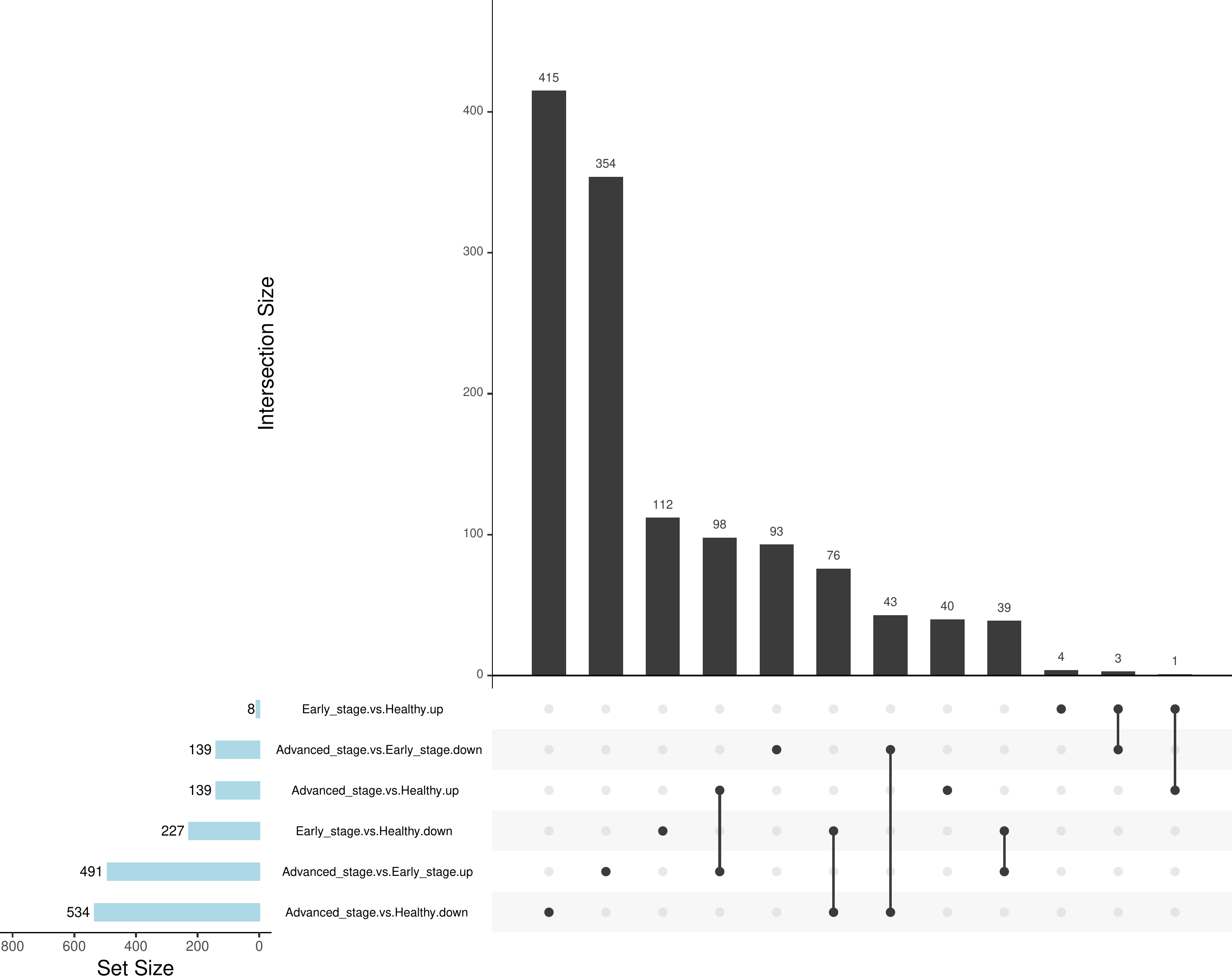


**Fig.** MRNA Advanced stage vs Early stage

**(File path: Figure+Table/MRNA-Advanced-stage-vs-Early-stage.pdf)**

* adj.P.Val cut-off: 0.05
* Log2(FC) cut-off: 1

**(See: Figure+Table/MRNA-Advanced-stage-vs-Early-stage-content)**



**Fig.** MRNA Difference intersection

**(File path: Figure+Table/MRNA-Difference-intersection.pdf)**

* All\_intersection:

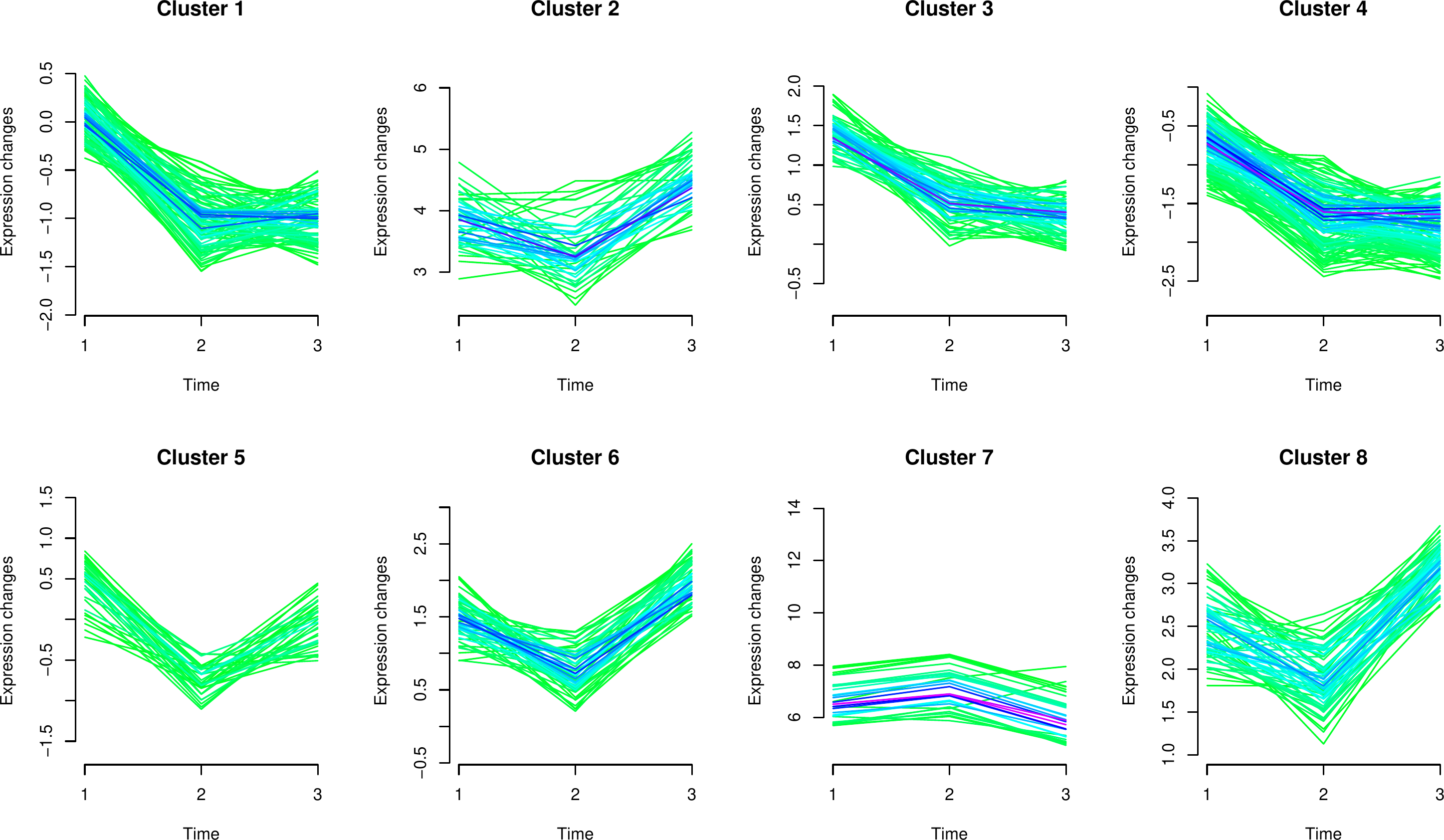
**(See: Figure+Table/MRNA-Difference-intersection-content)**

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/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 Fig. 。按照 Healthy, Early\_stage, Advanced\_stage 顺序, 在 Mfuzz 聚类中，6, 8 为按时序上调，共 325 个，1, 3, 4 为按时序下调，共 590 个。其他基因为离散变化。。

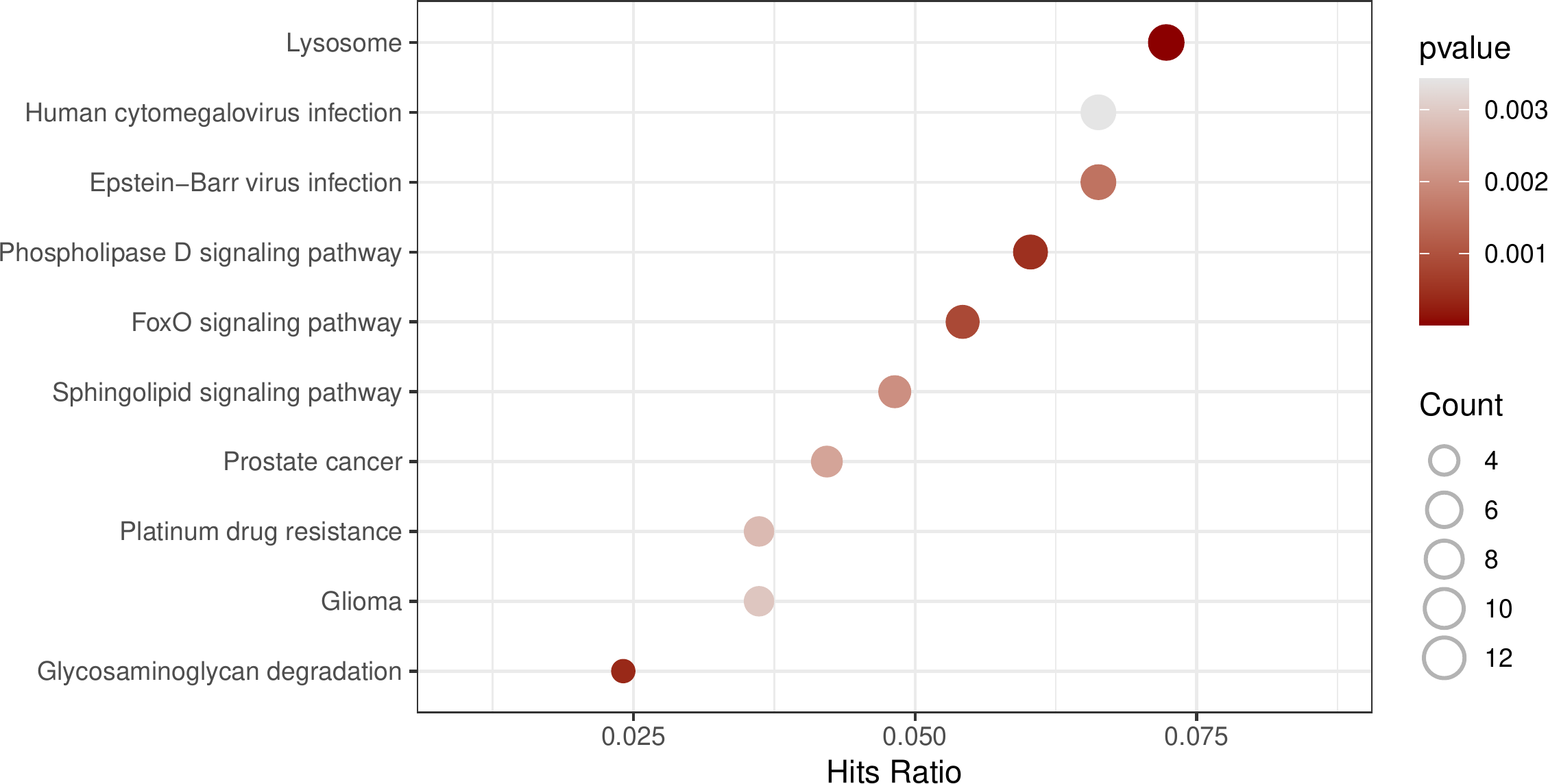


**Fig.** MRNA Mfuzz clusters

**(File path: Figure+Table/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. , Fig. 。 GO 见 Fig. , Fig. 。 上调组主要富集于 Cellular Processes, Metabolism 相关。 下调组富集于 Immune system 相关。



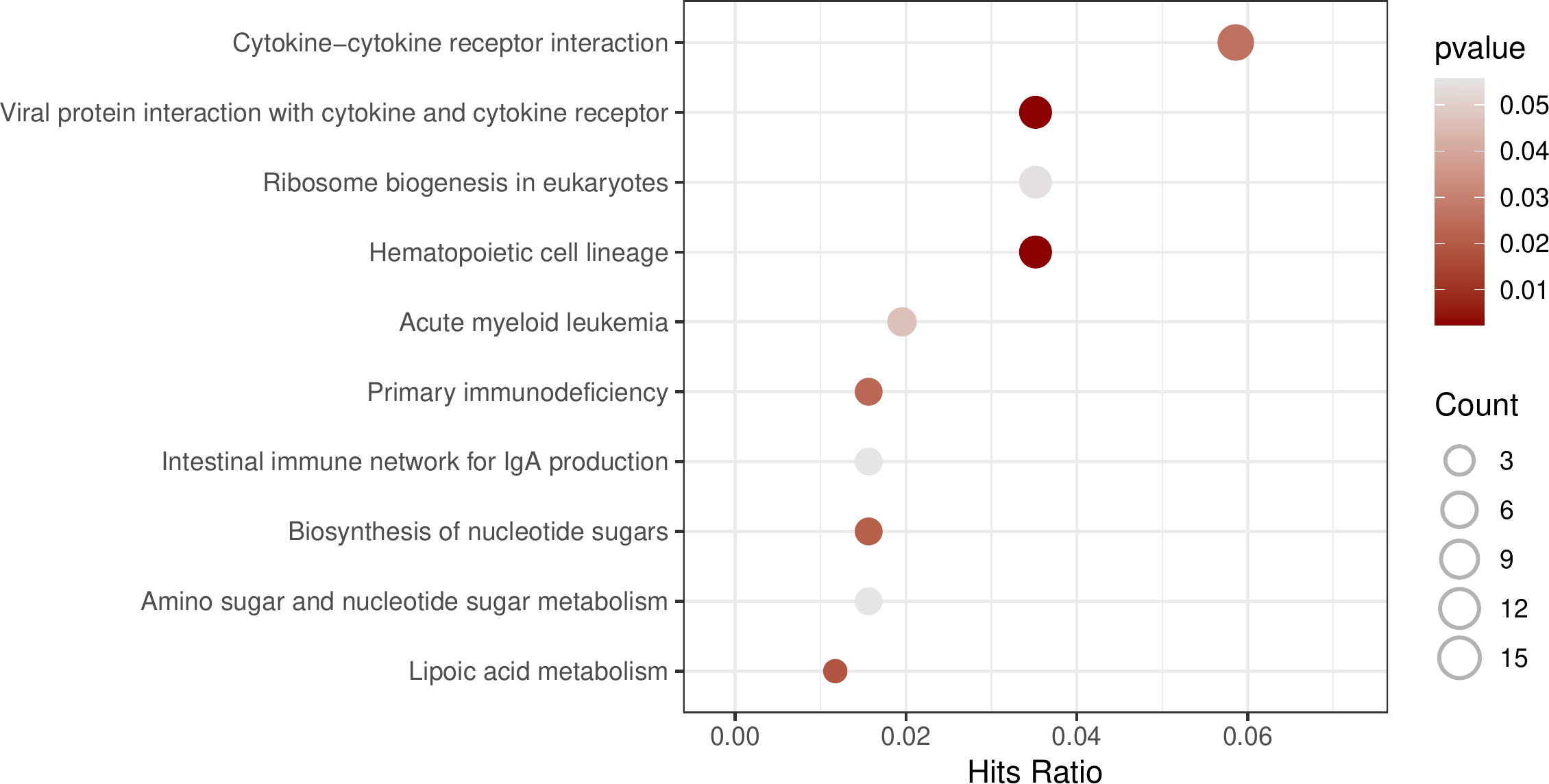
**Fig.** MRNA up KEGG enrichment

**(File path: Figure+Table/MRNA-up-KEGG-enrichment.pdf)**



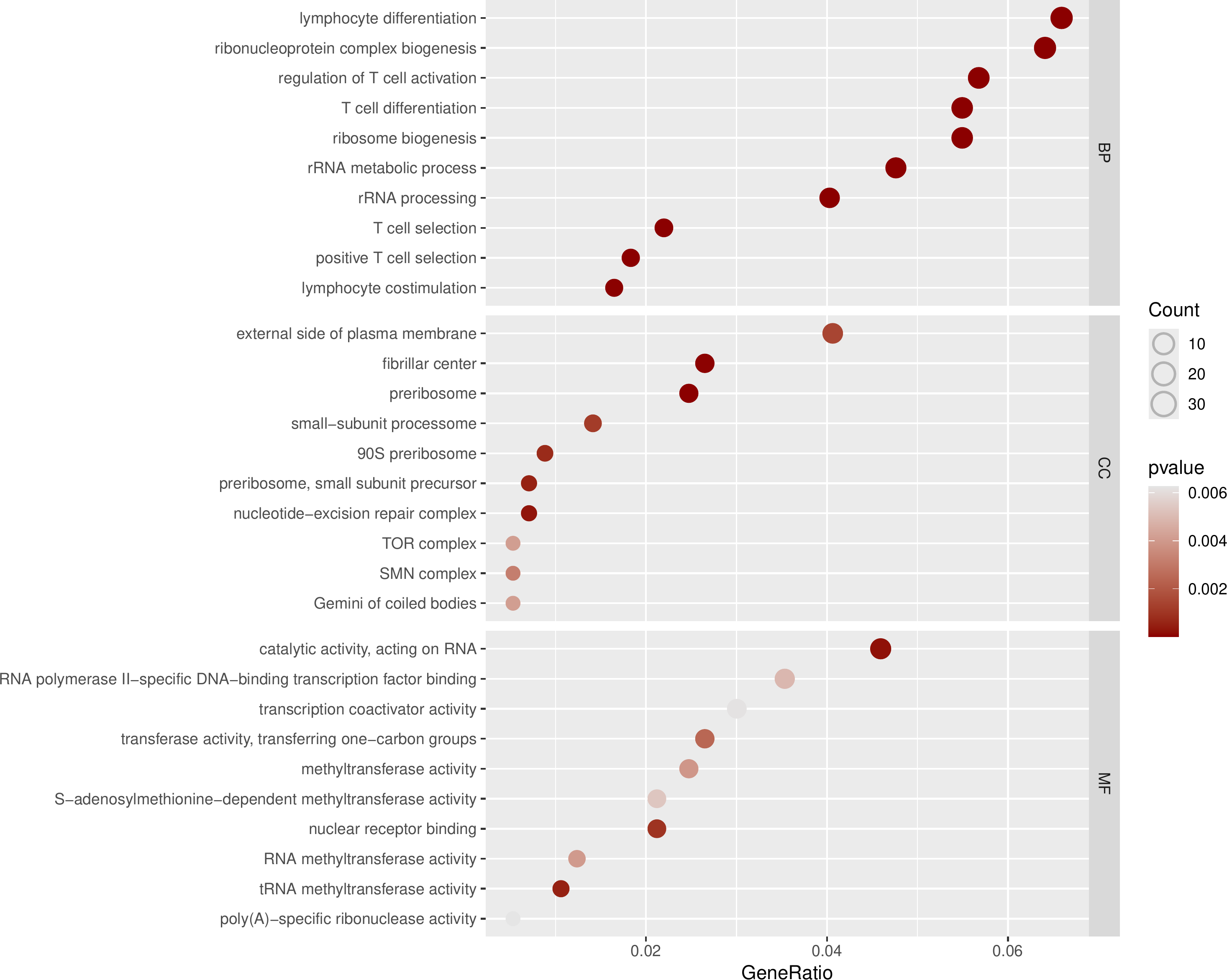
**Fig.** MRNA up GO enrichment

**(File path: Figure+Table/MRNA-up-GO-enrichment.pdf)**



**Fig.** MRNA down KEGG enrichment

**(File path: Figure+Table/MRNA-down-KEGG-enrichment.pdf)**



**Fig.** MRNA down GO enrichment

**(File path: Figure+Table/MRNA-down-GO-enrichment.pdf)**

## 3.4 TCGA 数据获取 (LUSC)

获取 TCGA-LUSC 数据，用于临床数据分析和预后模型建立。

## 3.5 COX 回归 (LUSC)

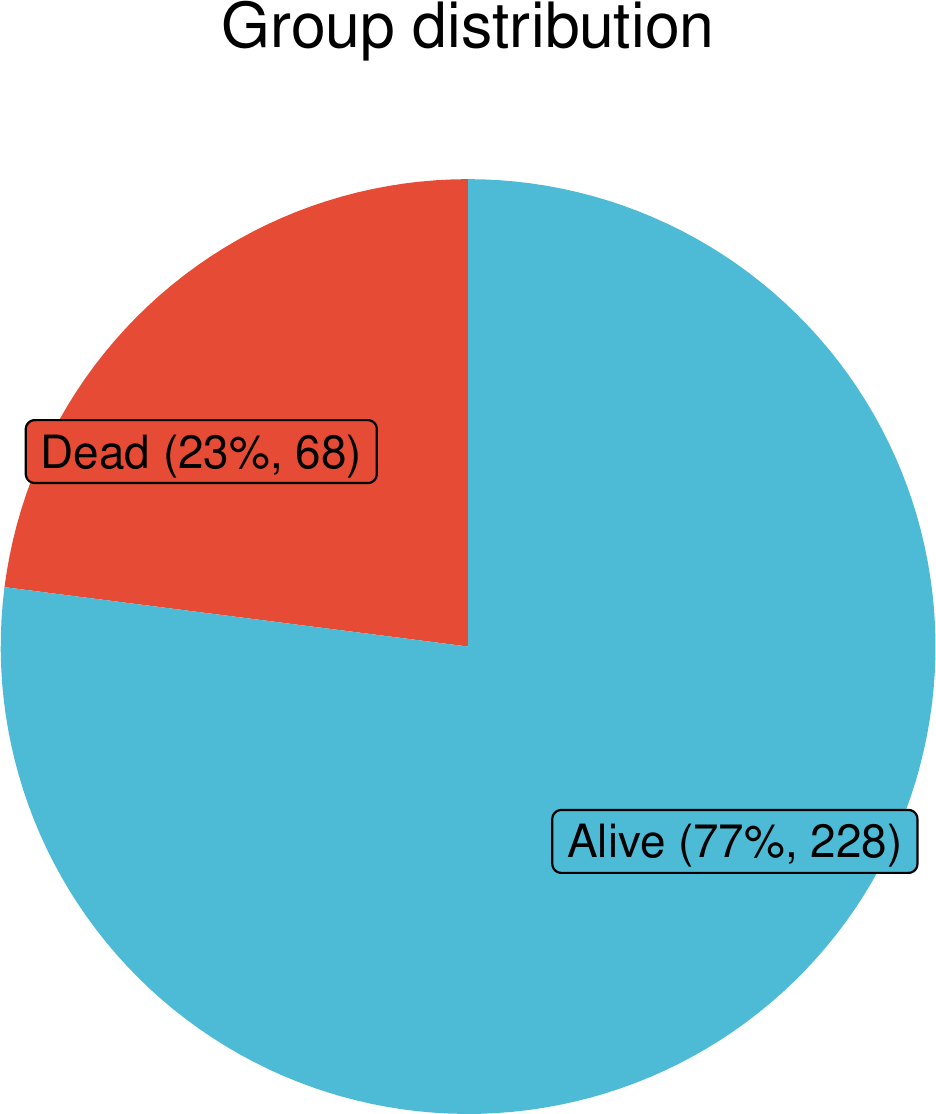
数据源自 TCGA-LUSC，筛选 AJCC Stage (ajcc\_pathologic\_stage) 为 Stage I, Stage II 的病人，并且 days\_to\_last\_follow\_up 大于 10 天，且为肿瘤组织的样本。所用样本的元数据见 Tab. 。

将 LUSC 数据 (count) 标准化后 (同 MRNA 的方法)，以生存状态为指标 (Fig. )，以 EFS 算法，进行 Feature selection, 得到 Top 30 基因, 统计得分见 Fig. 。 随后，以单因素 COX 回归，筛选能显著预测生存结局的基因。EFS 与单因素 COX 回归结果如 Tab. 。共 9 个基因：SERPINE1, BCL2L2, SLC14A1, DYRK3, PDCD11, AGPAT3, COQ2, TPK1, MPZL1。

**Tab.** LUSC metadata

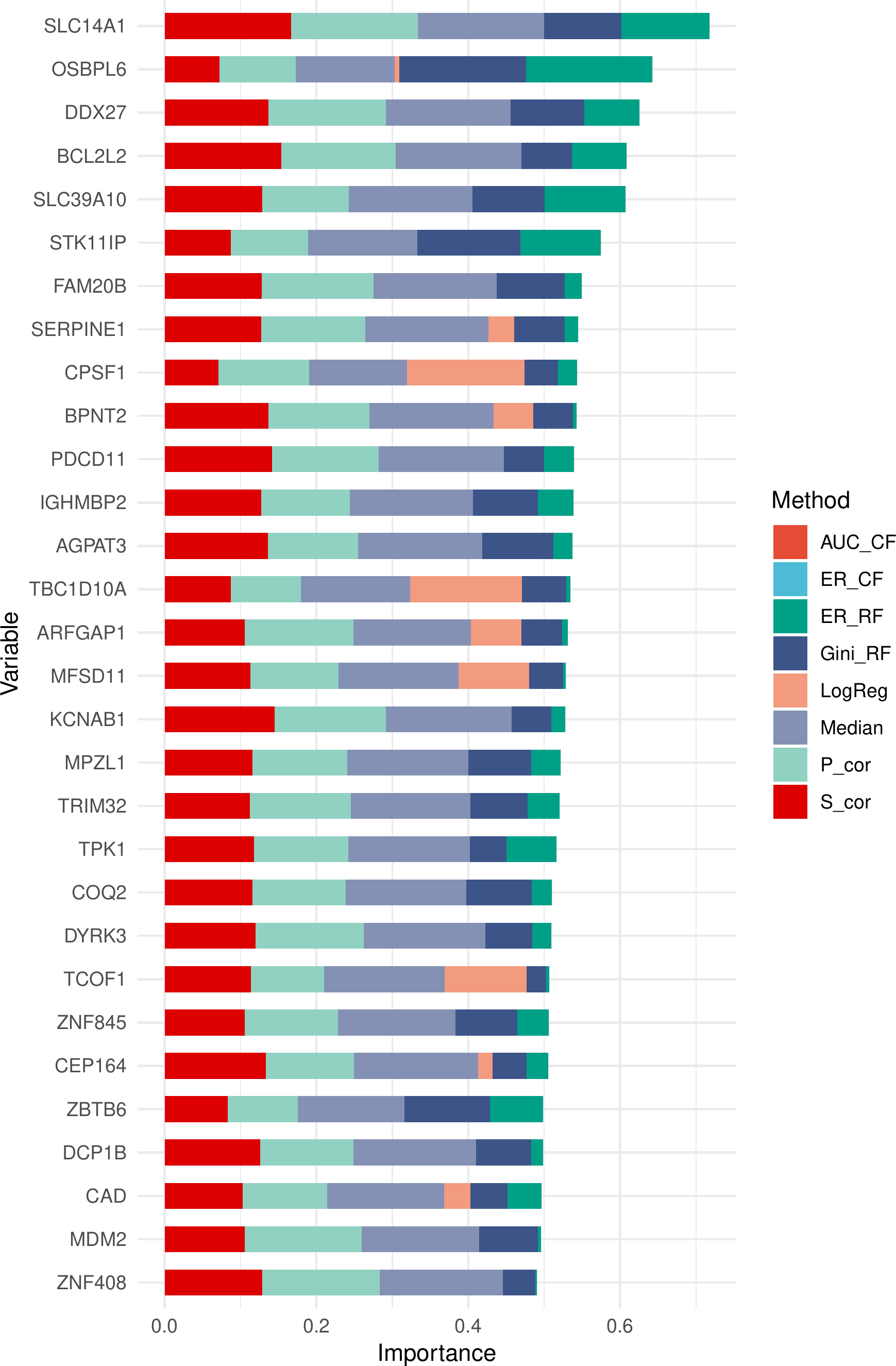
| Sample | Group | Lib.size | Norm.f... | Barcode | Patient | ShortL... | Defini... | Sample......9 | Sample......10 | ..\* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TCGA-1... | Dead | 37255376 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Alive | 43492507 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Dead | 39714818 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Dead | 39975834 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Alive | 40747917 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Alive | 44429215 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Alive | 21735012 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-1... | Alive | 61555320 | 1 | TCGA-1... | TCGA-1... | TP | Primar... | TCGA-1... | 01 | ... |
| TCGA-2... | Dead | 48526883 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| TCGA-2... | Alive | 60443027 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| TCGA-2... | Alive | 60226646 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| TCGA-2... | Alive | 58365924 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| TCGA-2... | Dead | 58631601 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| TCGA-2... | Dead | 54098835 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| TCGA-2... | Alive | 54921696 | 1 | TCGA-2... | TCGA-2... | TP | Primar... | TCGA-2... | 01 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/LUSC-metadata.xlsx)**



**Fig.** LUSC Group distribution

**(File path: Figure+Table/LUSC-Group-distribution.pdf)**



**Fig.** LUSC Top Features Selected By EFS

**(File path: Figure+Table/LUSC-Top-Features-Selected-By-EFS.pdf)**

**Tab.** LUSC Uni COX cofficients filtered by EFS

| Feature | Coef | Exp(coef) | Se(coef) | Z | Pvalue | P.adjust |
| --- | --- | --- | --- | --- | --- | --- |
| SERPINE1 | 0.24390416... | 1.27622202... | 0.11338208... | 2.15117024... | 0.03146276... | 0.83624230... |
| BCL2L2 | -0.4283650... | 0.65157349... | 0.14369845... | -2.9809998... | 0.00287308... | 0.83624230... |
| SLC14A1 | 0.45789329... | 1.58074031... | 0.09645783... | 4.74708263... | 2.06371657... | 0.00184702... |
| DYRK3 | 0.29492209... | 1.34302172... | 0.13055192... | 2.25904045... | 0.02388086... | 0.83624230... |
| PDCD11 | -0.2666409... | 0.76594805... | 0.11389728... | -2.3410647... | 0.01922883... | 0.83624230... |
| AGPAT3 | 0.27364386... | 1.31474648... | 0.12352324... | 2.21532281... | 0.02673791... | 0.83624230... |
| COQ2 | -0.2876134... | 0.75005145... | 0.12333293... | -2.3320085... | 0.01970024... | 0.83624230... |
| TPK1 | 0.31038423... | 1.36394908... | 0.13272344... | 2.33857877... | 0.01935724... | 0.83624230... |
| MPZL1 | 0.30363800... | 1.35477853... | 0.11695141... | 2.59627472... | 0.00942406... | 0.83624230... |

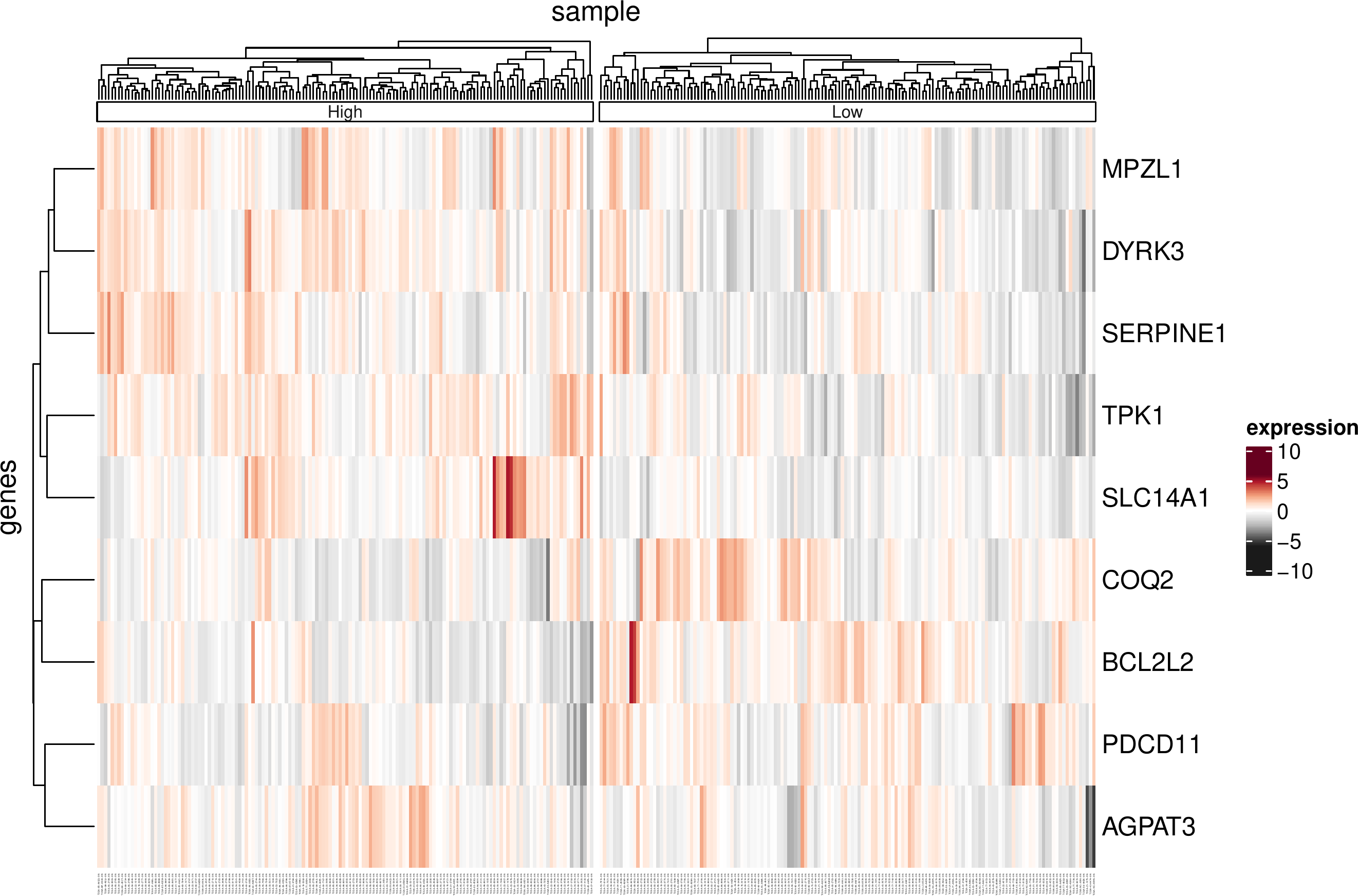
**(File path: Figure+Table/LUSC-Uni-COX-cofficients-filtered-by-EFS.csv)**

## 3.6 Survival 生存分析 (LUSC)

这些基因表达特征如 Fig. 热图所示。

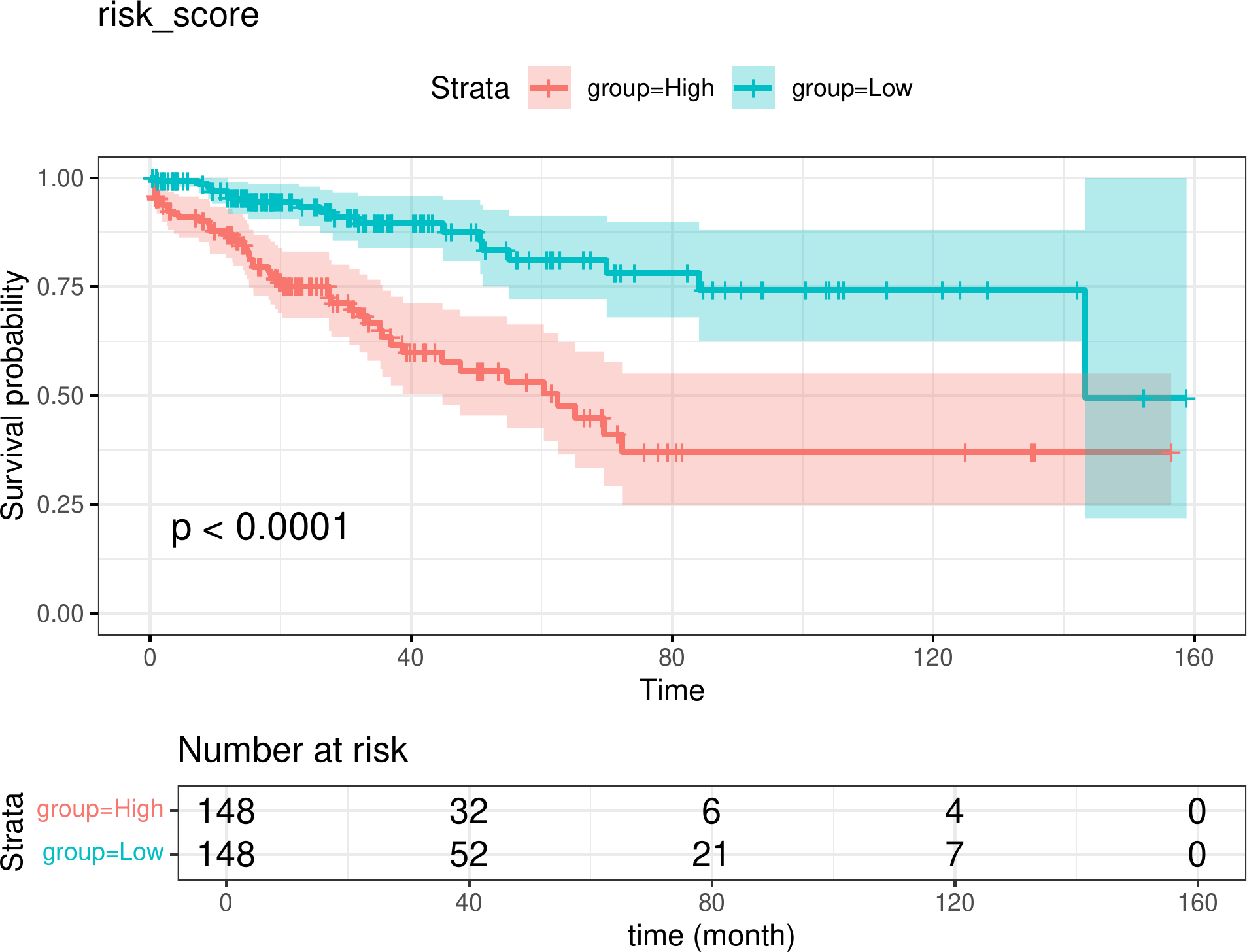
建立预后特征，构建风险评分：

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. 。 AUC 见 Fig. 。 第 1，3，5 年存活的患者，风险评分显著较低。



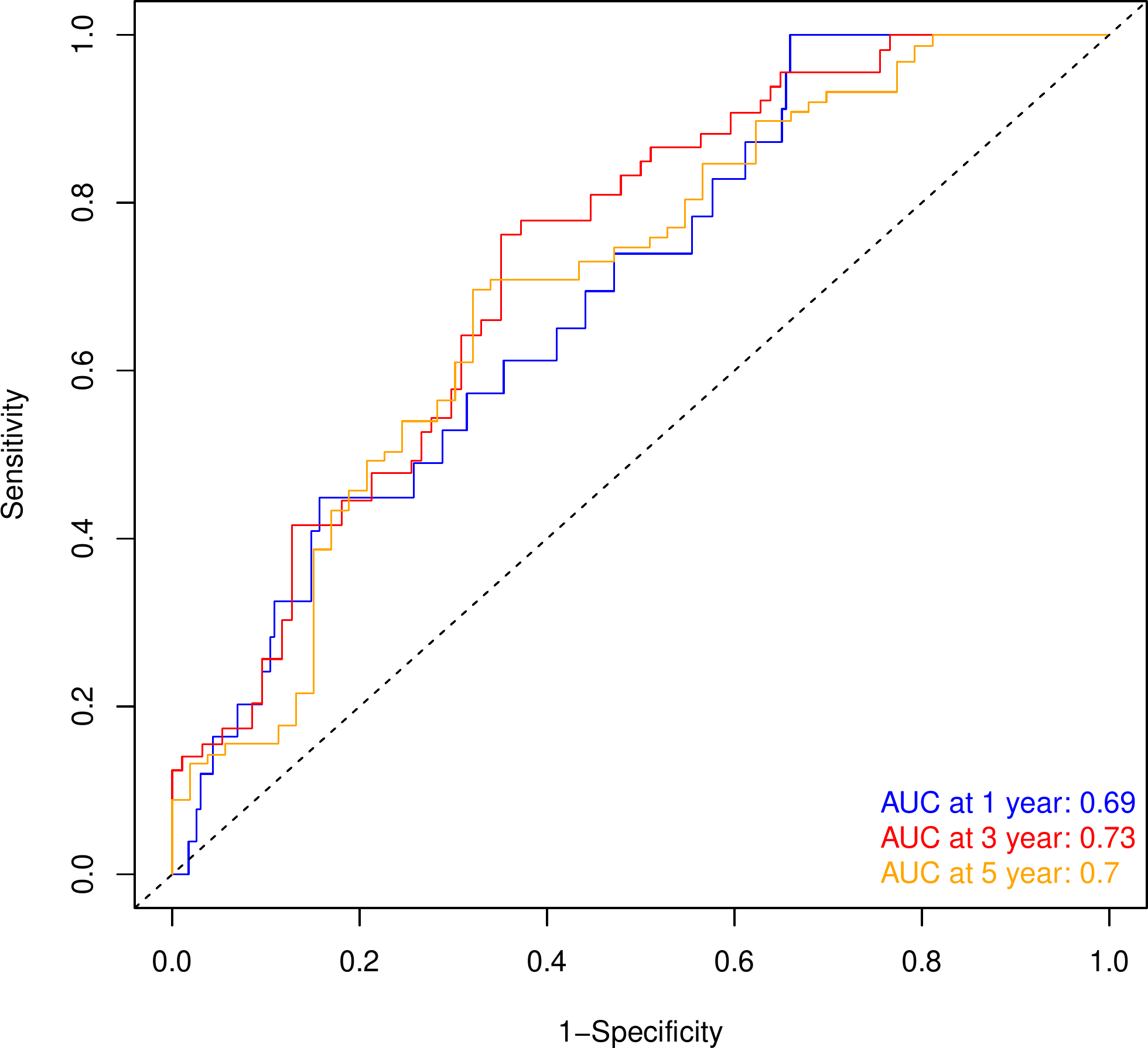
**Fig.** LUSC risk score related genes heatmap

**(File path: Figure+Table/LUSC-risk-score-related-genes-heatmap.pdf)**



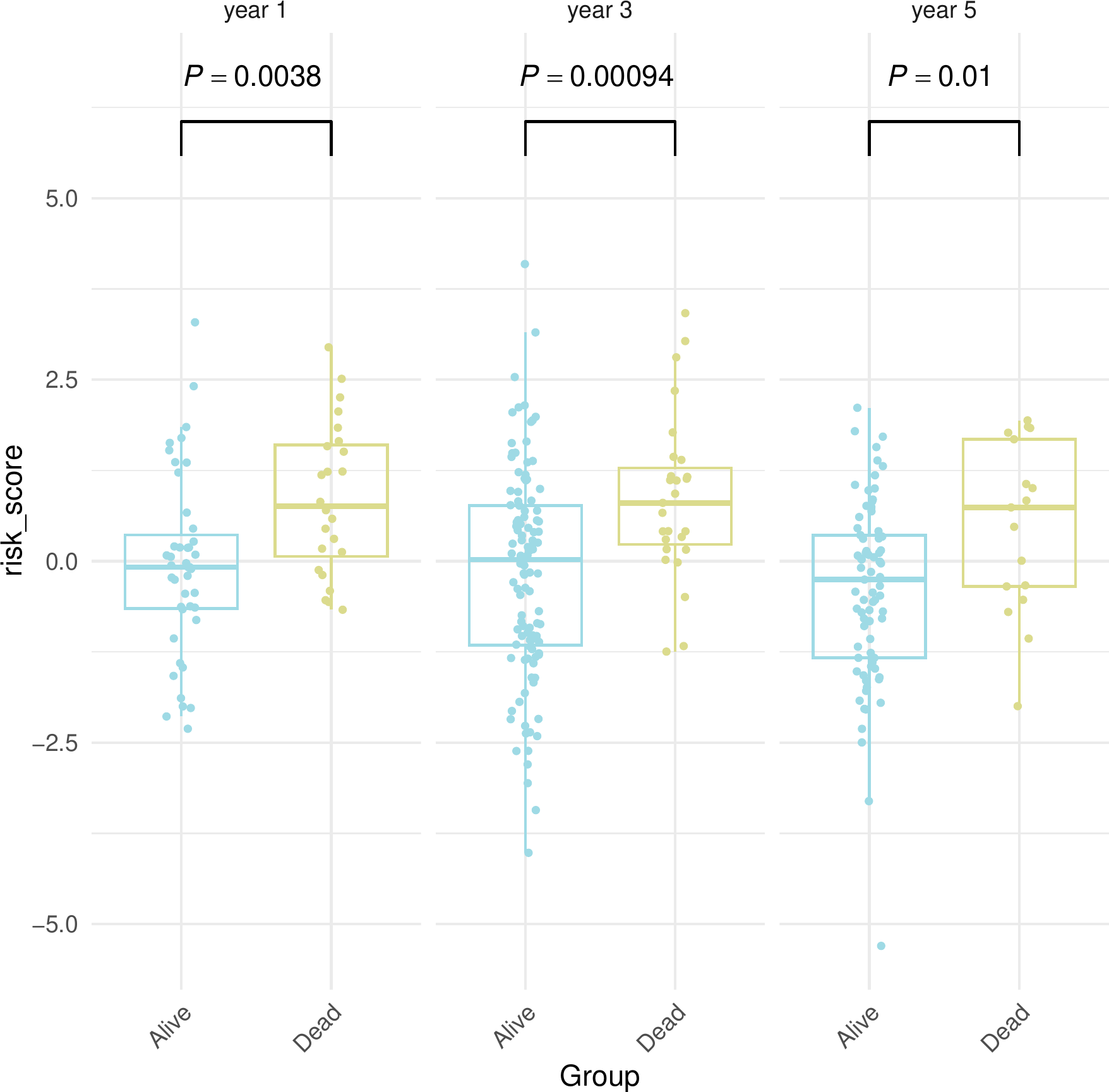
**Fig.** LUSC survival curve of risk score

**(File path: Figure+Table/LUSC-survival-curve-of-risk-score.pdf)**



**Fig.** LUSC time ROC

**(File path: Figure+Table/LUSC-time-ROC.pdf)**



**Fig.** LUSC boxplot of risk score

**(File path: Figure+Table/LUSC-boxplot-of-risk-score.pdf)**

## 3.7 COX 回归 (Prognosis)

进一步通过单因素和多因素 COX 回归的方式评估了包括风险评分在内的4项预后特征 (smoking, treatment 等其他数据缺失值较多，不易处理) 。

单因素和多因素分析结果，风险评分是诊断早期肺癌预后的独立风险指标，见 Tab. 。

**Tab.** META Coefficients Of COX

| Feature | Uni coefficients | Uni p | Multi coefficients | Multi p |
| --- | --- | --- | --- | --- |
| Age (>65/<=64) | 0.215192027477178 | 0.41311128570459 | 0.101704964279124 | 0.709950395815607 |
| Gender (female/male) | 0.183214670354876 | 0.523084334023976 | 0.36615958534245 | 0.209021149635656 |
| AJCC stage (I/II) | 0.0144776502394665 | 0.954293406973733 | 0.295933974836866 | 0.256805285385573 |
| Risk score | 0.54781086968627 | 6.06868708145961e-09 | 0.565964036073118 | 1.45743179043085e-09 |

**(File path: Figure+Table/META-Coefficients-Of-COX.csv)**

## 3.8 GEO 数据获取 (GEO\_LUSC)

为了验证预后特征在不同数据平台上的性能，这里获取了 GEO 数据平台的早期肺癌数据 (GSE157010，微阵列数据)，并筛选了 Stage 为 I，II 阶段的病例。

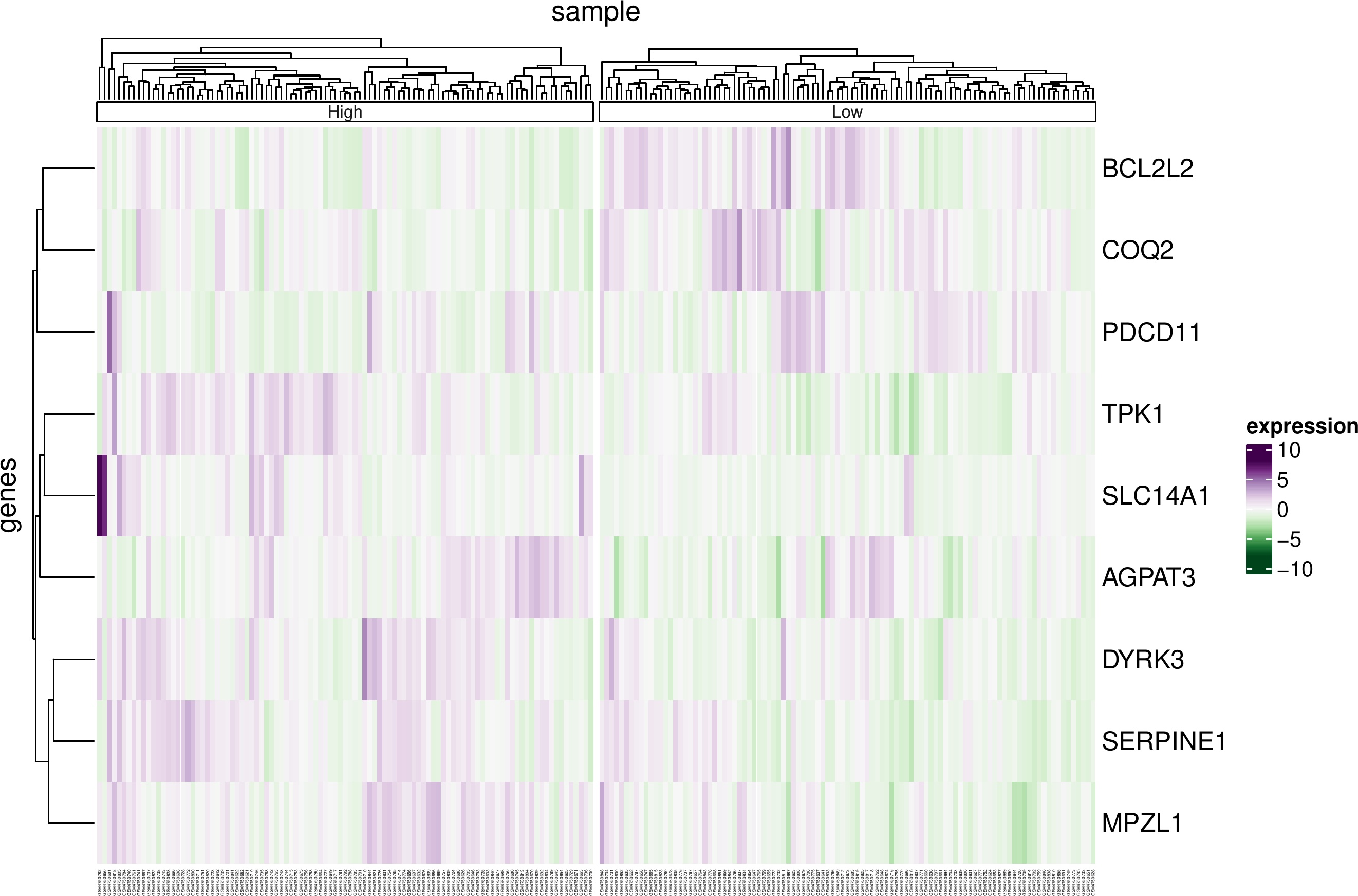
**Tab.** LUSC GSE157010 metadata

| Sample | Vital ... | Group | Days t... | Stage | Rownames | Title | Age.ch1 | Diagno... | Os eve... |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GSM475... | Dead | Dead | 1259.5... | T2 | GSM475... | CAD NA... | 63 | Squamo... | 1 |
| GSM475... | Dead | Dead | 993.20... | T2 | GSM475... | CAD NA... | 78 | Squamo... | 1 |
| GSM475... | Alive | Alive | 2071.2... | T2 | GSM475... | CAD NA... | 68 | Squamo... | 0 |
| GSM475... | Alive | Alive | 2836.6... | T2 | GSM475... | CAD NA... | 71 | Squamo... | 0 |
| GSM475... | Dead | Dead | 452.71... | T2 | GSM475... | CAD NA... | 83 | Squamo... | 1 |
| GSM475... | Dead | Dead | 835.39... | T2 | GSM475... | CAD NA... | 63 | Squamo... | 1 |
| GSM475... | Alive | Alive | 1945.9... | T2 | GSM475... | CAD NA... | 73 | Squamo... | 0 |
| GSM475... | Dead | Dead | 234.73... | T2 | GSM475... | CAD NA... | 71 | Squamo... | 1 |
| GSM475... | Alive | Alive | 1045.4... | T2 | GSM475... | CAD NA... | 76 | Squamo... | 0 |
| GSM475... | Dead | Dead | 581.91... | T2 | GSM475... | CAD NA... | 72 | Squamo... | 1 |
| GSM475... | Alive | Alive | 1919.3... | T2 | GSM475... | CAD NA... | 78 | Squamo... | 0 |
| GSM475... | Dead | Dead | 982.35... | T2 | GSM475... | CAD NA... | 72 | Squamo... | 1 |
| GSM475... | Dead | Dead | 1091.8... | T2 | GSM475... | CAD NA... | 71 | Squamo... | 1 |
| GSM475... | Dead | Dead | 671.67... | T2 | GSM475... | CAD NA... | 68 | Squamo... | 1 |
| GSM475... | Alive | Alive | 1880.8... | T2 | GSM475... | CAD NA... | 59 | Squamo... | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/LUSC-GSE157010-metadata.csv)**

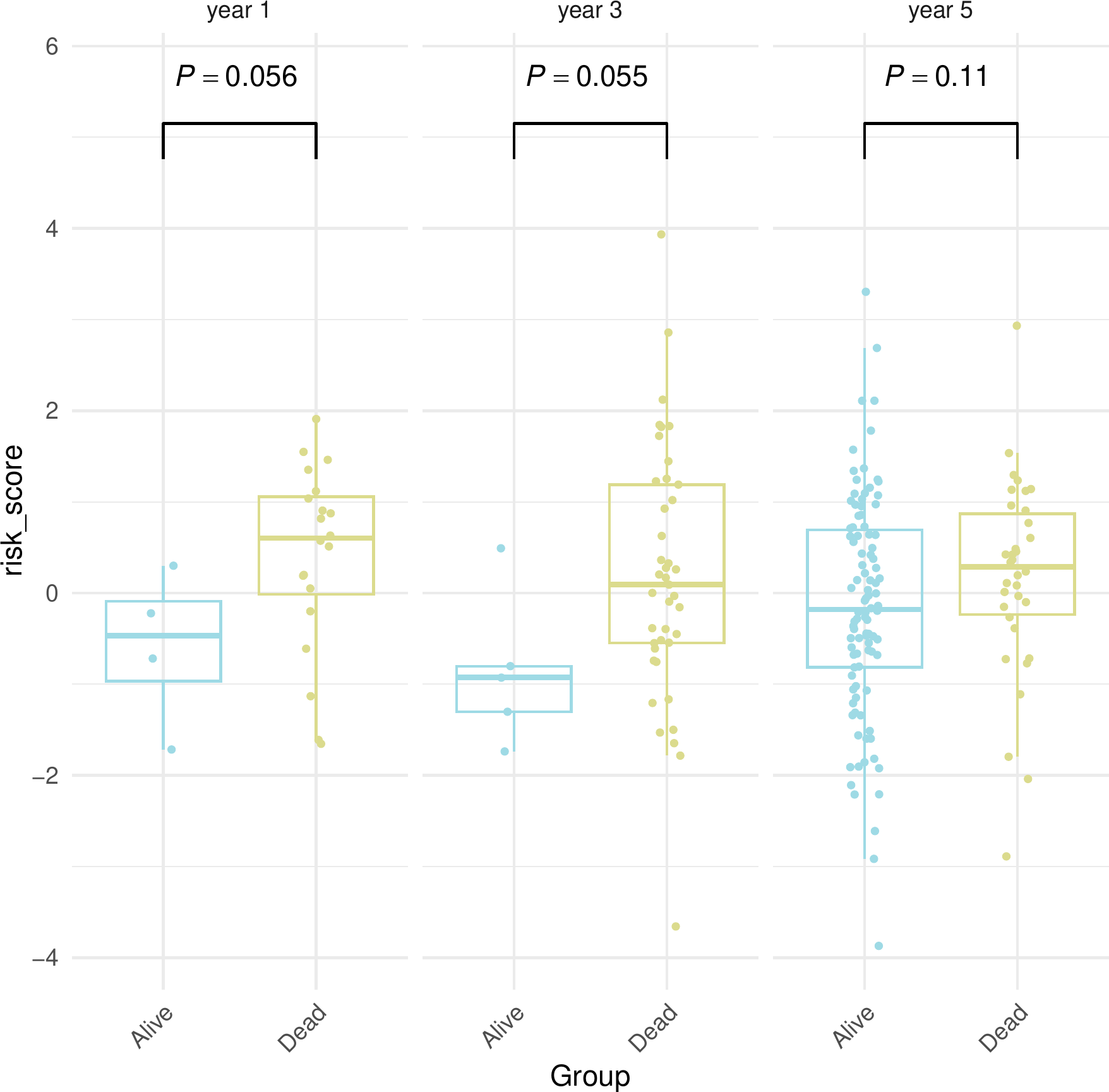
## 3.9 Survival 生存分析 (GEO\_LUSC)

GEO 数据集中，风险评分基因集表达特征见 Fig. 。 将 GEO 数据集按相同的方式处理，并计算风险评分， 生存结果见 Fig. ，高风险组与低风险组显著差异。 ROC 曲线见 Fig. 。 第 1，3，5 年风险评分差异见 Fig. 。



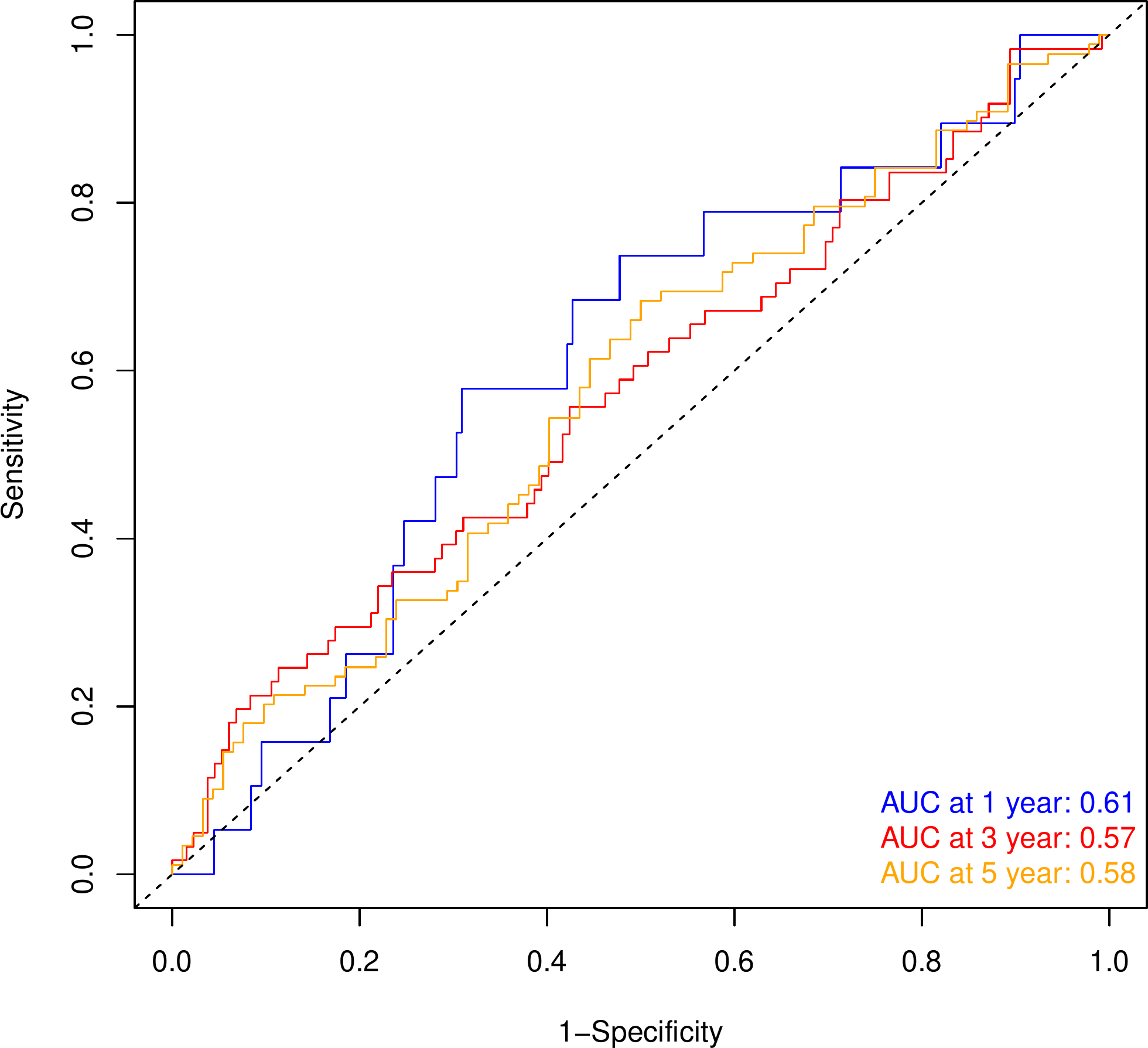
**Fig.** GEO LUSC risk score related genes heatmap

**(File path: Figure+Table/GEO-LUSC-risk-score-related-genes-heatmap.pdf)**



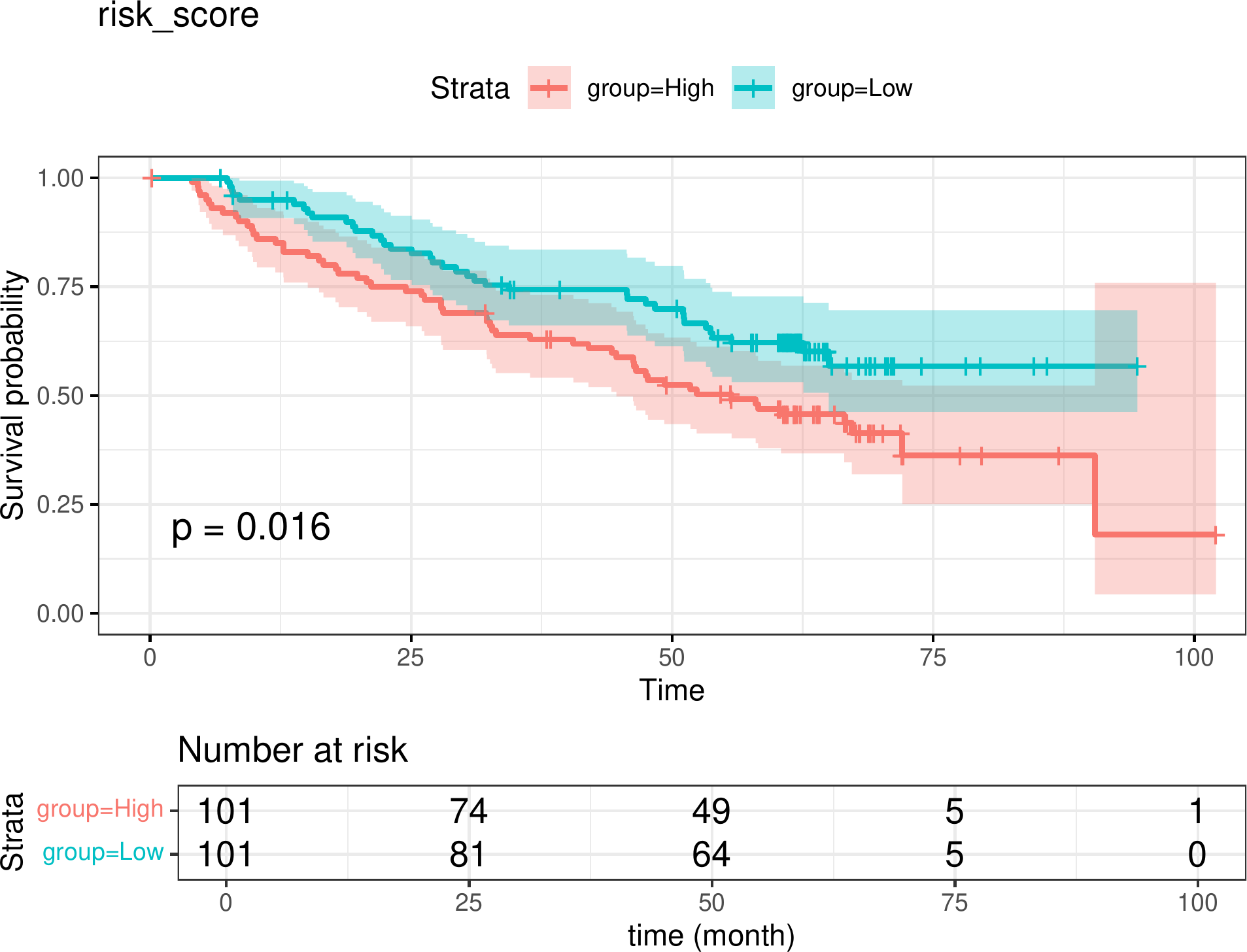
**Fig.** GEO LUSC boxplot of risk score

**(File path: Figure+Table/GEO-LUSC-boxplot-of-risk-score.pdf)**



**Fig.** GEO LUSC time ROC

**(File path: Figure+Table/GEO-LUSC-time-ROC.pdf)**

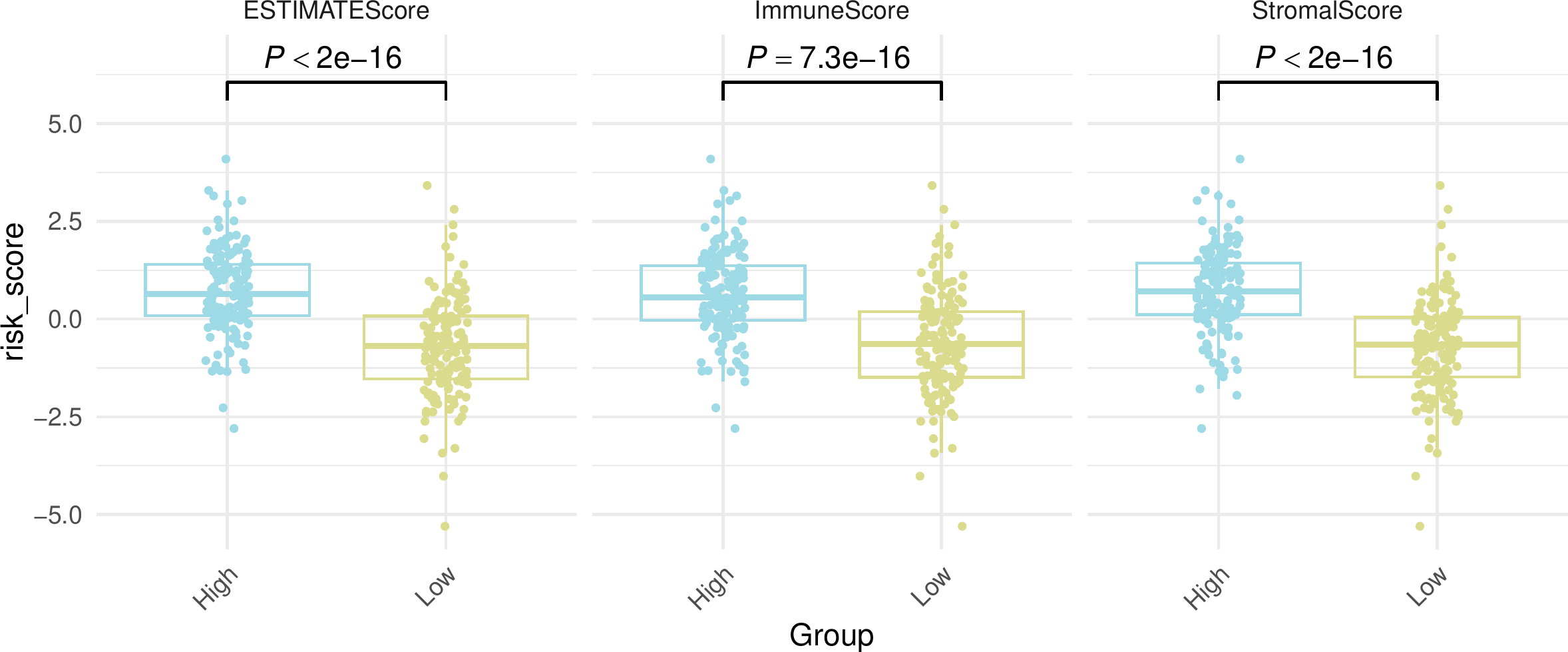


**Fig.** GEO LUSC survival curve of risk score

**(File path: Figure+Table/GEO-LUSC-survival-curve-of-risk-score.pdf)**

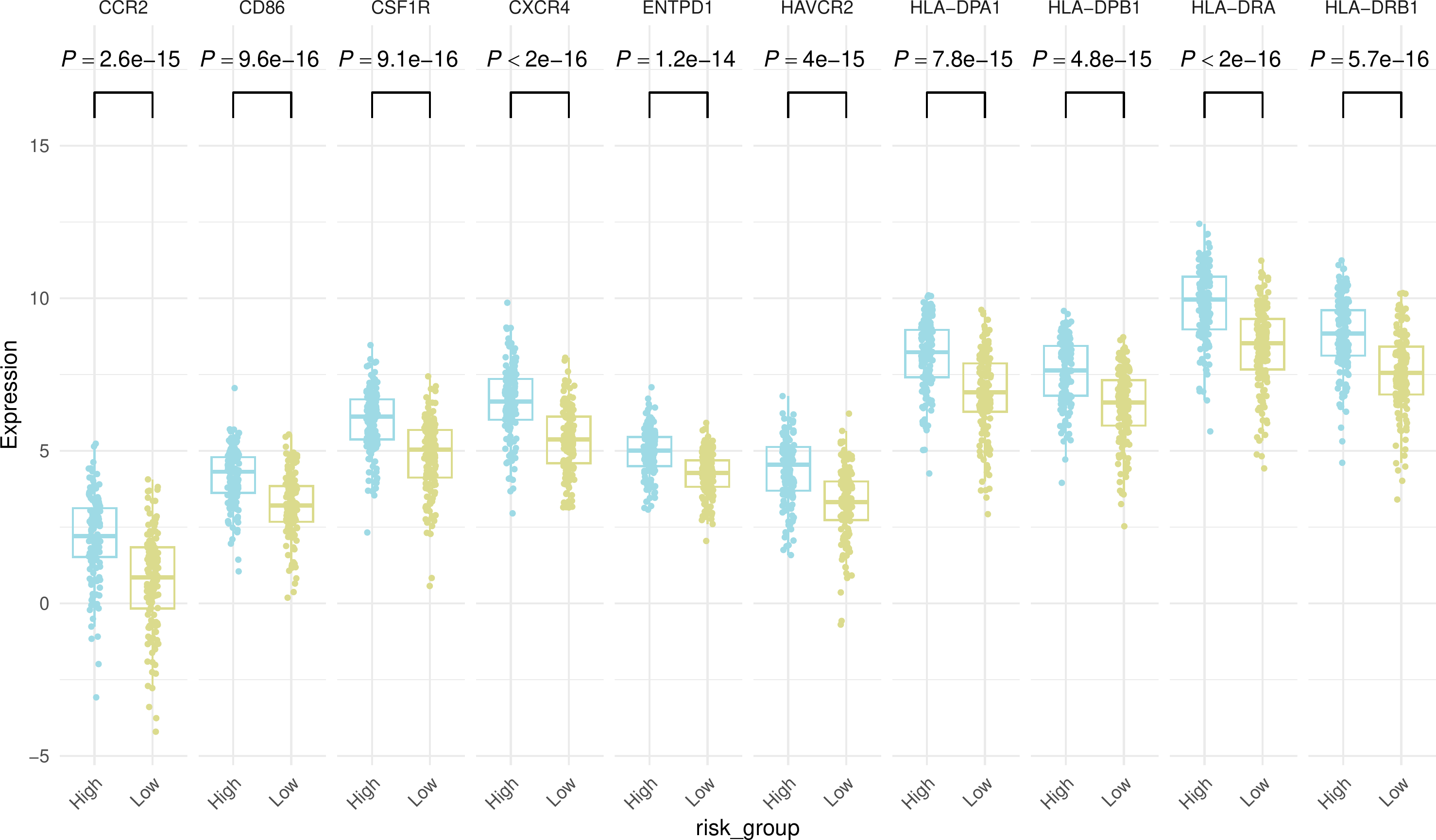
## 3.10 estimate 免疫评分 (LUSC)

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，免疫评分和 ESTIMATE 评分较低的患者具有较高的风险评分，见 Fig. 。 此外，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 119 个表达存在差异 (p.value < 0.05)。 前 10 个基因见 Fig. 。



**Fig.** LUSC immune Scores Plot

**(File path: Figure+Table/LUSC-immune-Scores-Plot.pdf)**



**Fig.** LUSC Top10 Immune Related Genes

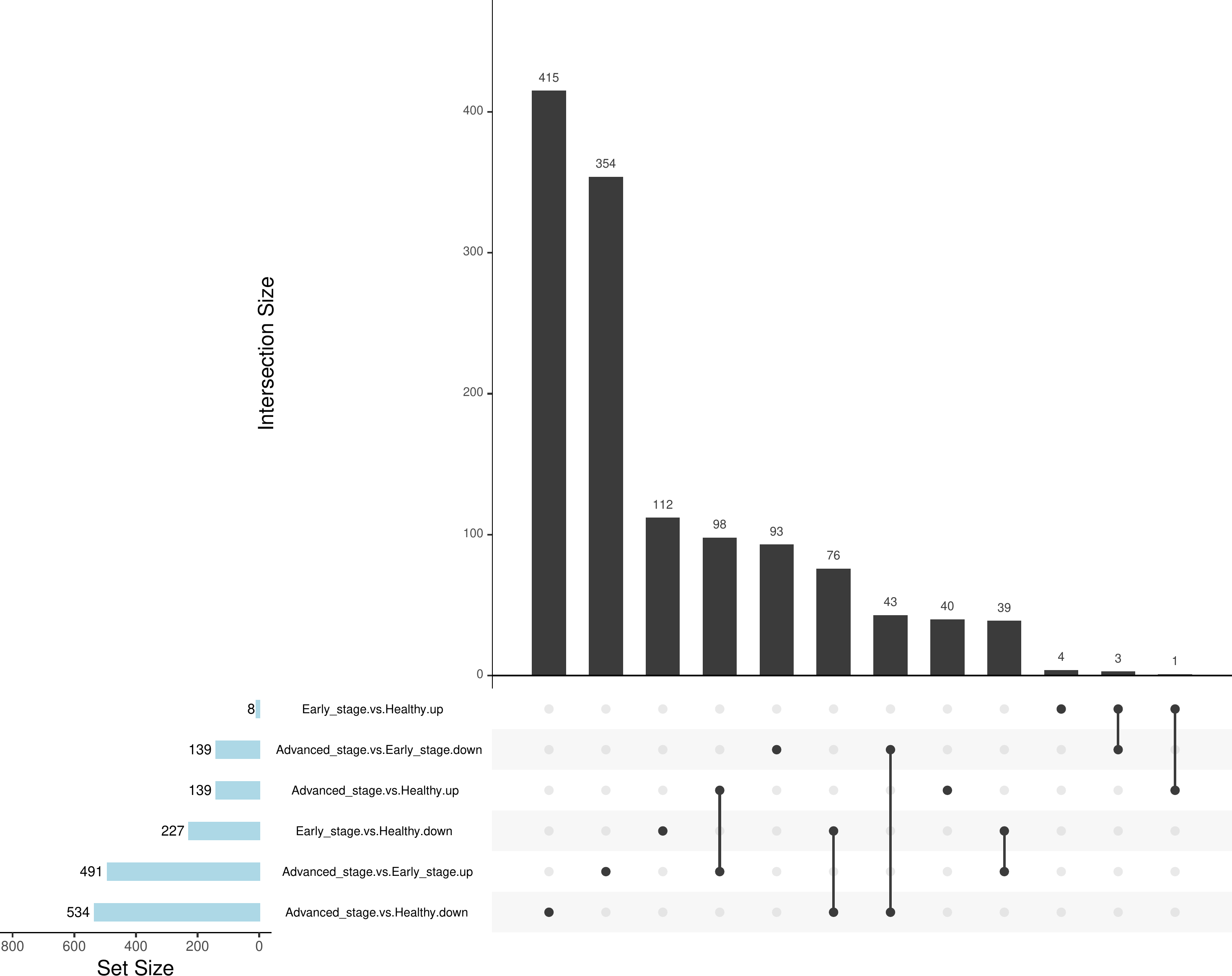
**(File path: Figure+Table/LUSC-Top10-Immune-Related-Genes.pdf)**

## 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。 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. 。 所有上调 DEGs 有 539 个，下调共 781；一共 1278 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/LNCRNA-DEGs-data)**



**Fig.** LNCRNA Difference intersection

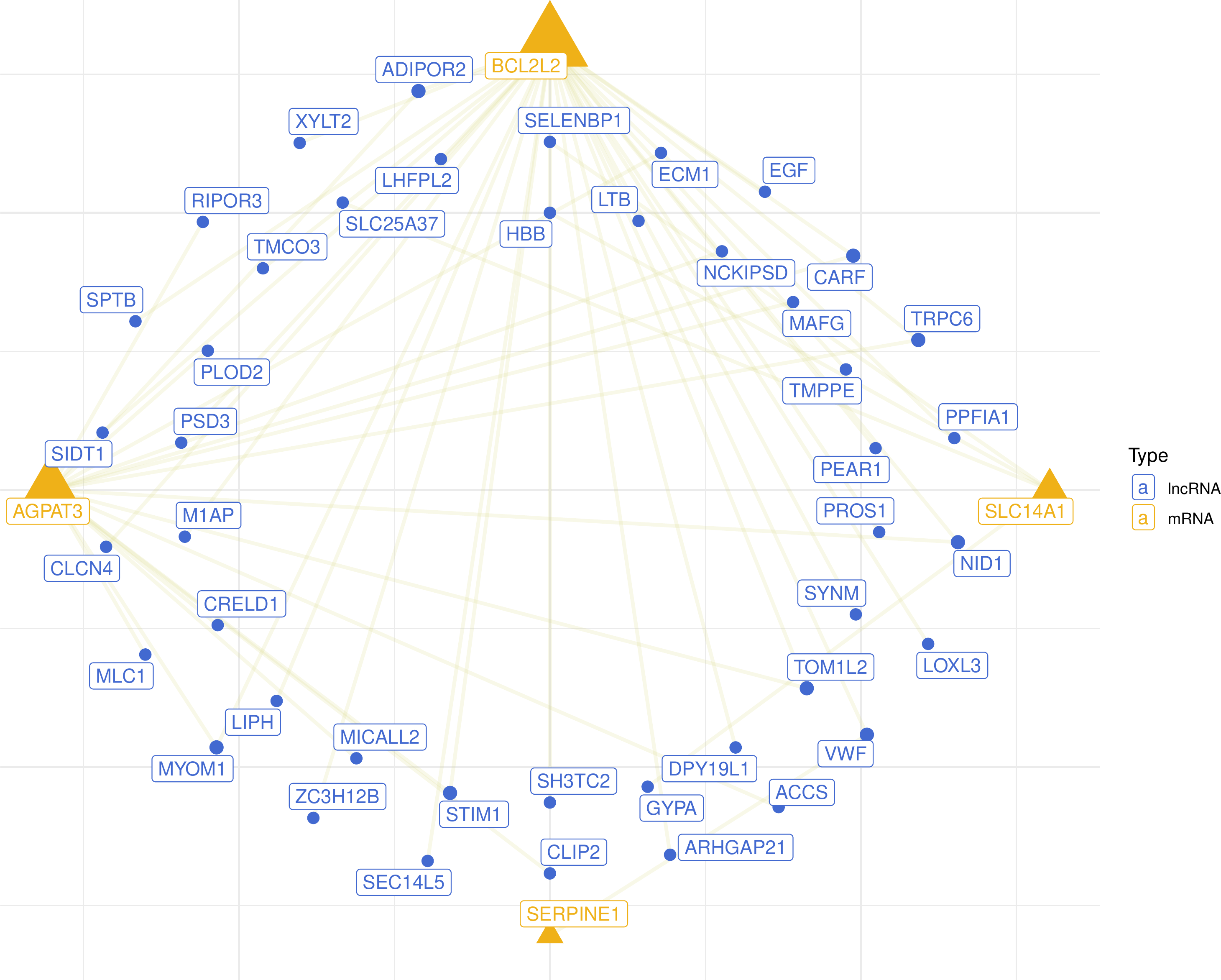
**(File path: Figure+Table/LNCRNA-Difference-intersection.pdf)**

* All\_intersection:

**(See: Figure+Table/LNCRNA-Difference-intersection-content)**

## 3.12 关联分析 (MRNA, LNCRNA)

将相关系数 > 0.6 和 p < 0.001 设定为识别相关阈值，最终建立网络图见 Fig. 。 共包含 4 个 mRNA，4 个 lncRNA，52 对关联关系。



**Fig.** Significant Correlation mrna lncRNA

**(File path: Figure+Table/Significant-Correlation-mrna-lncRNA.pdf)**

**Tab.** Significant correlation

| MRNA | LncRNA | Cor | Pvalue | -log2(P.va... | Significant | Sign |
| --- | --- | --- | --- | --- | --- | --- |
| SLC14A1 | HBB | 0.61 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | LTB | -0.63 | 0 | 16.6096404... | < 0.001 | \*\* |
| AGPAT3 | NCKIPSD | 0.71 | 0 | 16.6096404... | < 0.001 | \*\* |
| AGPAT3 | MAFG | 0.62 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | TMPPE | 0.65 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | PEAR1 | 0.61 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | PROS1 | 0.63 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | SYNM | 0.63 | 0 | 16.6096404... | < 0.001 | \*\* |
| AGPAT3 | TOM1L2 | 0.61 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | TOM1L2 | 0.61 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | DPY19L1 | 0.62 | 0 | 16.6096404... | < 0.001 | \*\* |
| SLC14A1 | GYPA | 0.66 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | SH3TC2 | 0.62 | 0 | 16.6096404... | < 0.001 | \*\* |
| AGPAT3 | STIM1 | 0.61 | 0 | 16.6096404... | < 0.001 | \*\* |
| BCL2L2 | STIM1 | 0.63 | 0 | 16.6096404... | < 0.001 | \*\* |
| ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/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 ， 可预测肺癌 LUSC 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， 我们未能更多的验证。 后续评估发现，该风险评分与免疫微环境 (根据 ESTIMATE 评分) 显著相关。

# 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. Neumann, U., Genze, N. & Heider, D. EFS: An ensemble feature selection tool implemented as r-package and web-application. *BioData Mining* **10**, 21 (2017).

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

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