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

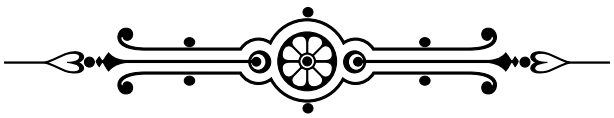
单　　号：BSXG240327

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



Contents

List of Figures

List of Tables

1 分析流程

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

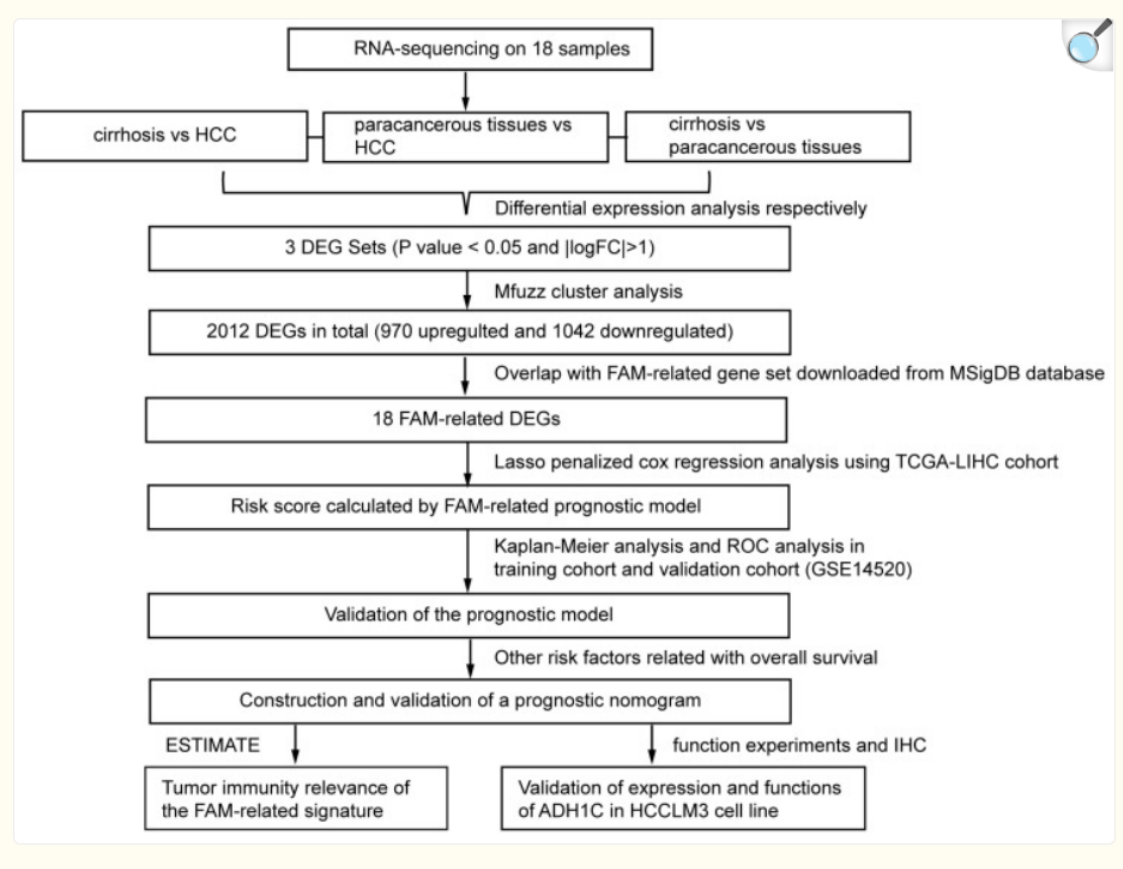


Figure 1: Unnamed chunk 7

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. 1 。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. 6。所有 DEGs 表达特征见 Fig. 2。所有上调 DEGs 有 539 个，下调共 781；一共 1278 个 (非重复)。

Table 1 (下方表格) 为表格 MRNA metadata 概览。

(对应文件为 `Figure+Table/MRNA-metadata.csv`)

> 注：表格共有 247 行 6 列，以下预览的表格可能省略部分数据；含有 247 个唯一 'sample'。

Table 1: 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
...	...	...	...	...	...

Figure 2 (下方图) 为图 MRNA Heatmap of DEGs 概览。

(对应文件为 `Figure+Table/MRNA-Heatmap-of-DEGs.pdf`)

Figure 2: MRNA Heatmap of DEGs

Figure 3 (下方图) 为图 MRNA Early stage vs Healthy 概览。

(对应文件为 `Figure+Table/MRNA-Early-stage-vs-Healthy.pdf`)

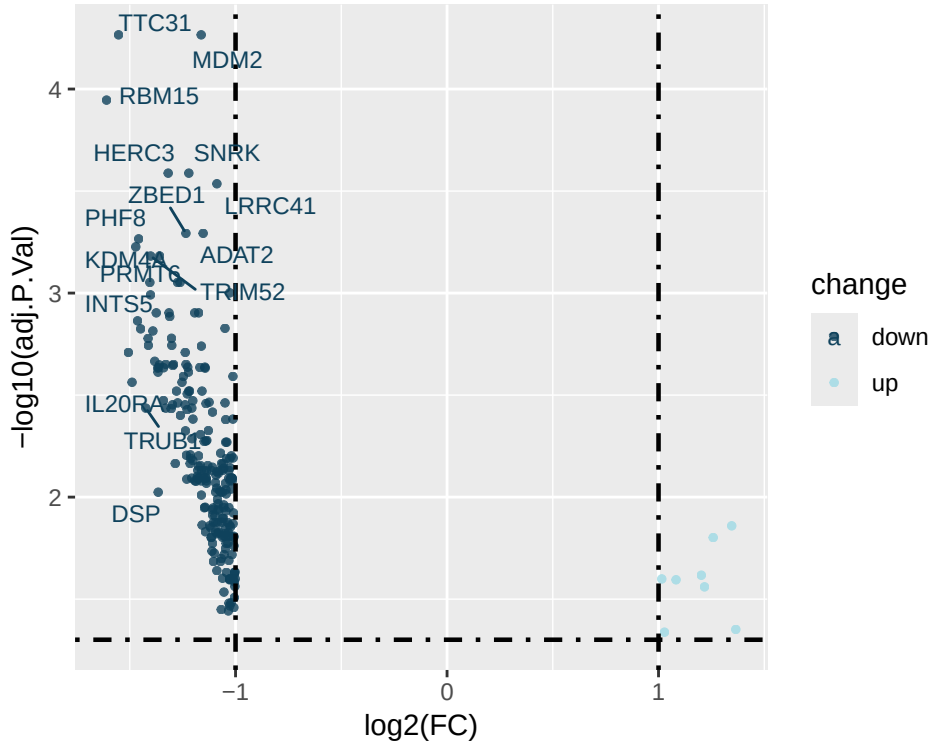


Figure 3: MRNA Early stage vs Healthy

adj.P.Val cut-off :

0.05

Log2(FC) cut-off :

1

(上述信息框内容已保存至 `Figure+Table/MRNA-Early-stage-vs-Healthy-content`)

Figure 4 (下方图) 为图 MRNA Advanced stage vs Healthy 概览。

(对应文件为 `Figure+Table/MRNA-Advanced-stage-vs-Healthy.pdf`)

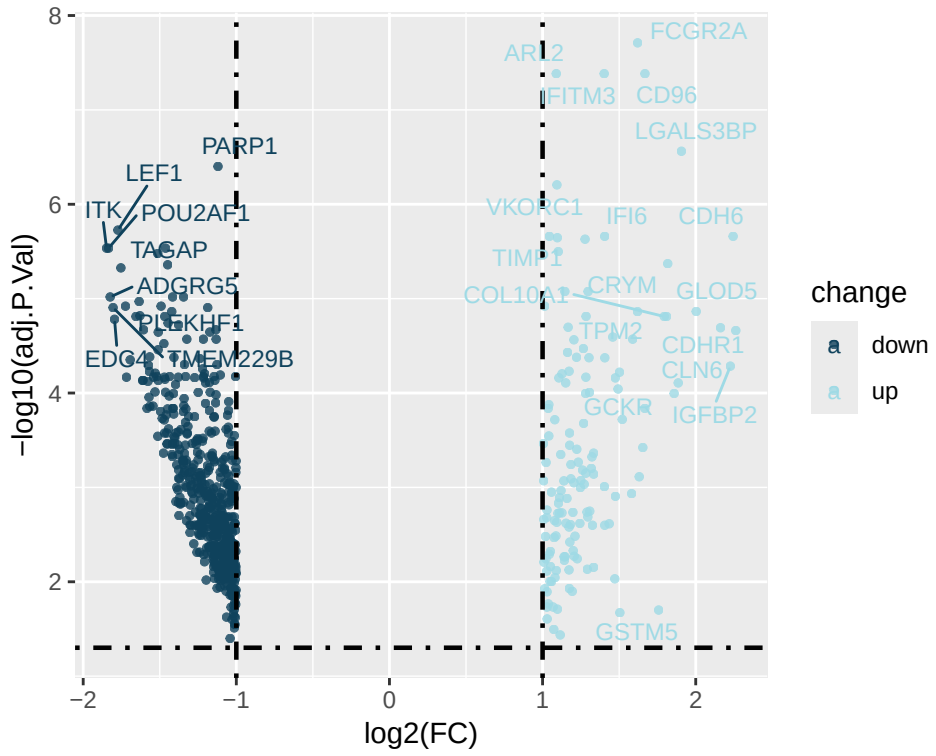


Figure 4: MRNA Advanced stage vs Healthy

adj.P.Val cut-off :

0.05

Log2(FC) cut-off :

1

(上述信息框内容已保存至 `Figure+Table/MRNA-Advanced-stage-vs-Healthy-content`)

Figure 5 (下方图) 为图 MRNA Advanced stage vs Early stage 概览。

(对应文件为 `Figure+Table/MRNA-Advanced-stage-vs-Early-stage.pdf`)

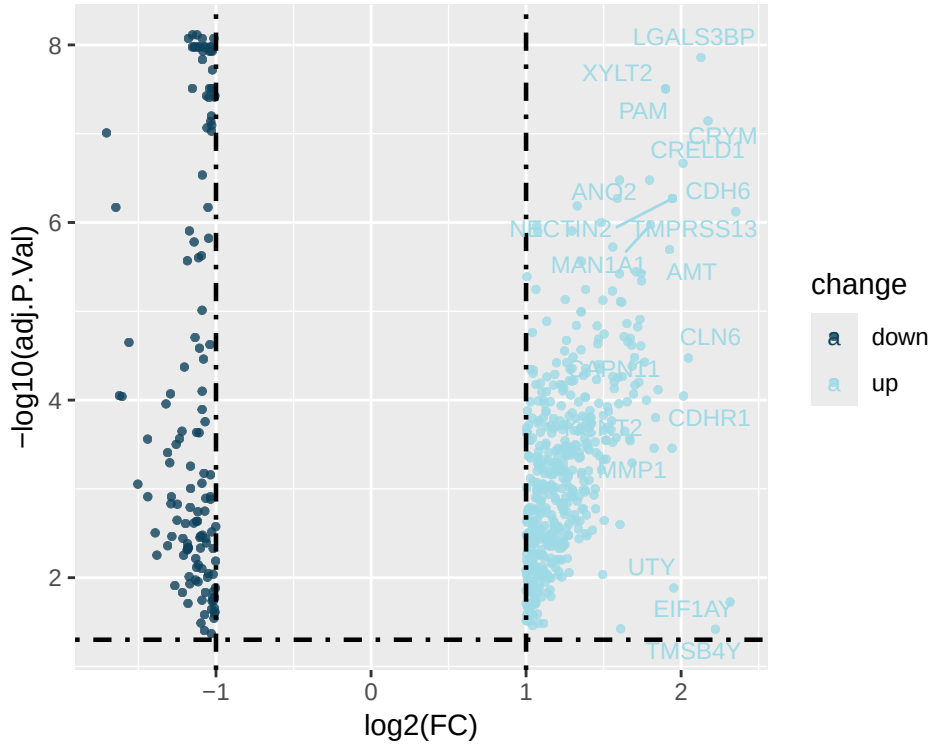


Figure 5: MRNA Advanced stage vs Early stage

adj.P.Val cut-off :

0.05

Log2(FC) cut-off :

1

(上述信息框内容已保存至 `Figure+Table/MRNA-Advanced-stage-vs-Early-stage-content`)

Figure 6 (下方图) 为图 MRNA Difference intersection 概览。

(对应文件为 `Figure+Table/MRNA-Difference-intersection.pdf`)

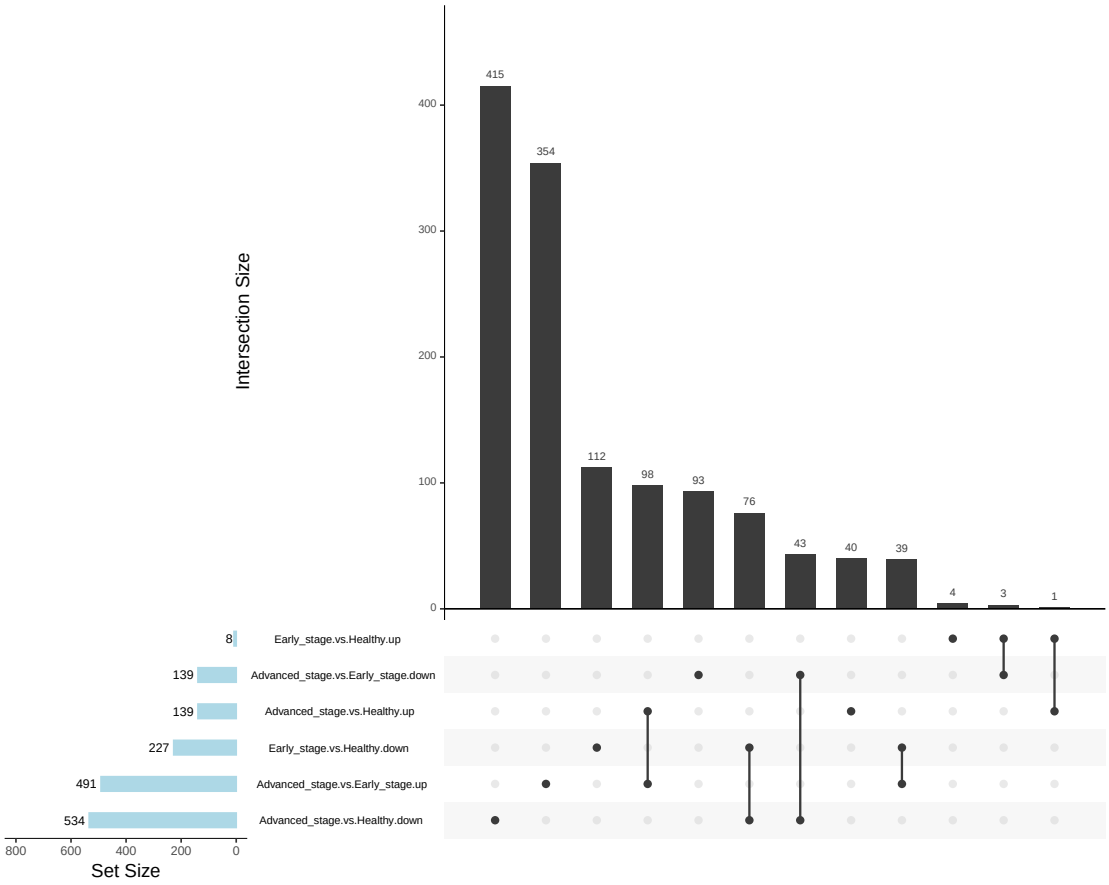


Figure 6: MRNA Difference intersection

All_intersection :

(上述信息框内容已保存至 `Figure+Table/MRNA-Difference-intersection-content`)

'MRNA data DEGs' 数据已全部提供。

(对应文件为 `Figure+Table/MRNA-data-DEGs`)

注：文件夹 Figure+Table/MRNA-data-DEGs 共包含 3 个文件。

1. 1_Early_stage - Healthy.csv
2. 2_Advanced_stage - Healthy.csv
3. 3_Advanced_stage - Early_stage.csv

3.2 Mfuzz 聚类分析 (MRNA)

将上述筛选得的 DEGs 以 Mfuzz 聚类分析。见 Fig. 7。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。。

Figure 7 (下方图) 为图 MRNA Mfuzz clusters 概览。

(对应文件为 `Figure+Table/MRNA-Mfuzz-clusters.pdf`)

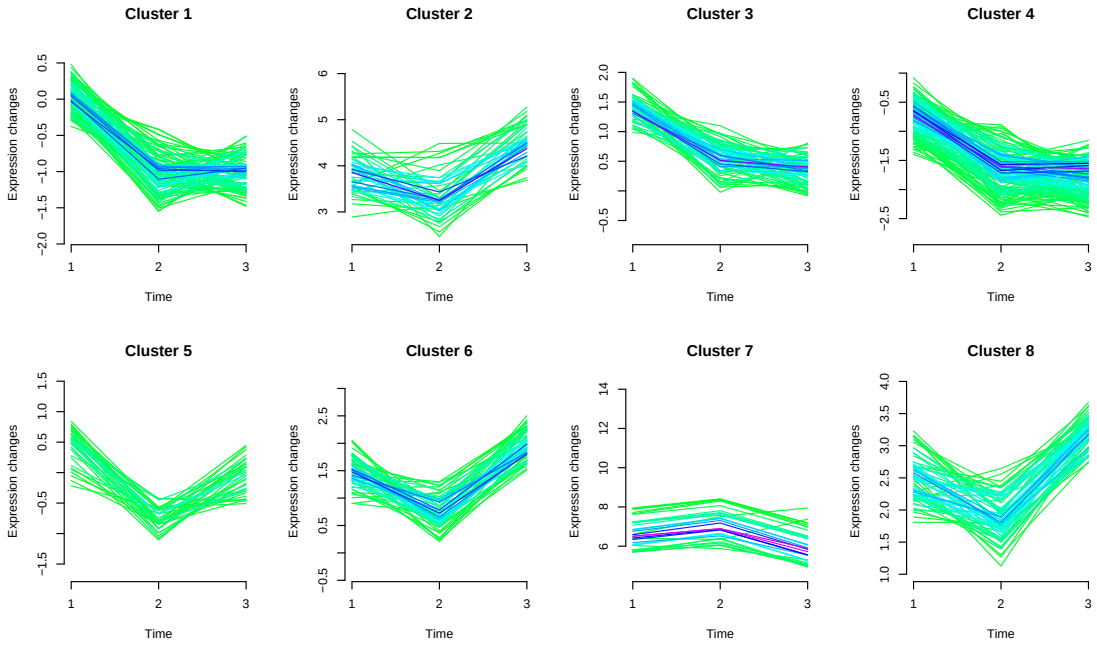


Figure 7: MRNA Mfuzz clusters

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, Fig. 10。GO 见 Fig. 9, Fig. 11。上调组主要富集于 Cellular Processes, Metabolism 相关。下调组富集于 Immune system 相关。

Figure 8 (下方图) 为图 MRNA up KEGG enrichment 概览。

(对应文件为 `Figure+Table/MRNA-up-KEGG-enrichment.pdf`)

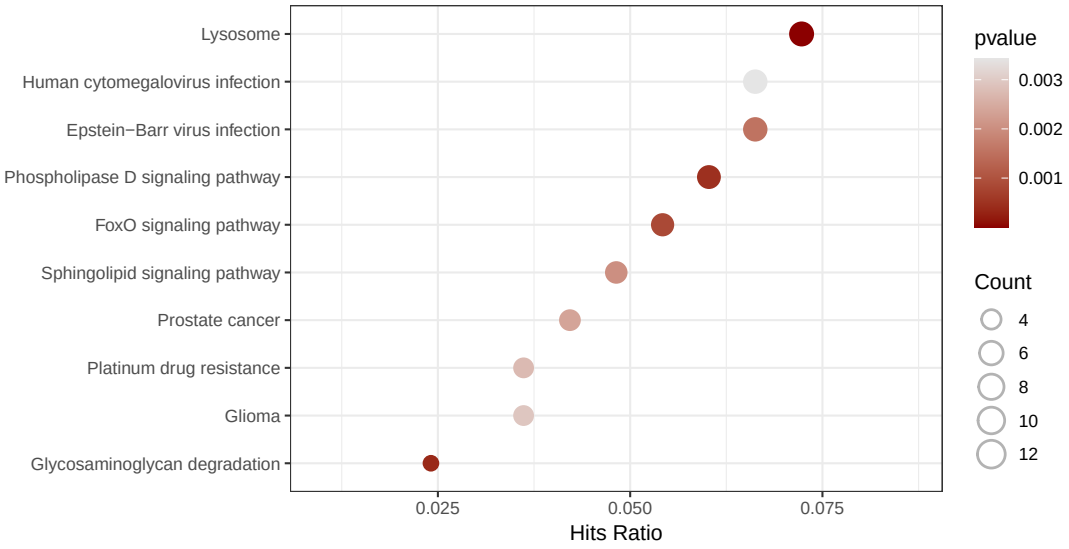


Figure 8: MRNA up KEGG enrichment

Figure 9 (下方图) 为图 MRNA up GO enrichment 概览。

(对应文件为 `Figure+Table/MRNA-up-GO-enrichment.pdf`)

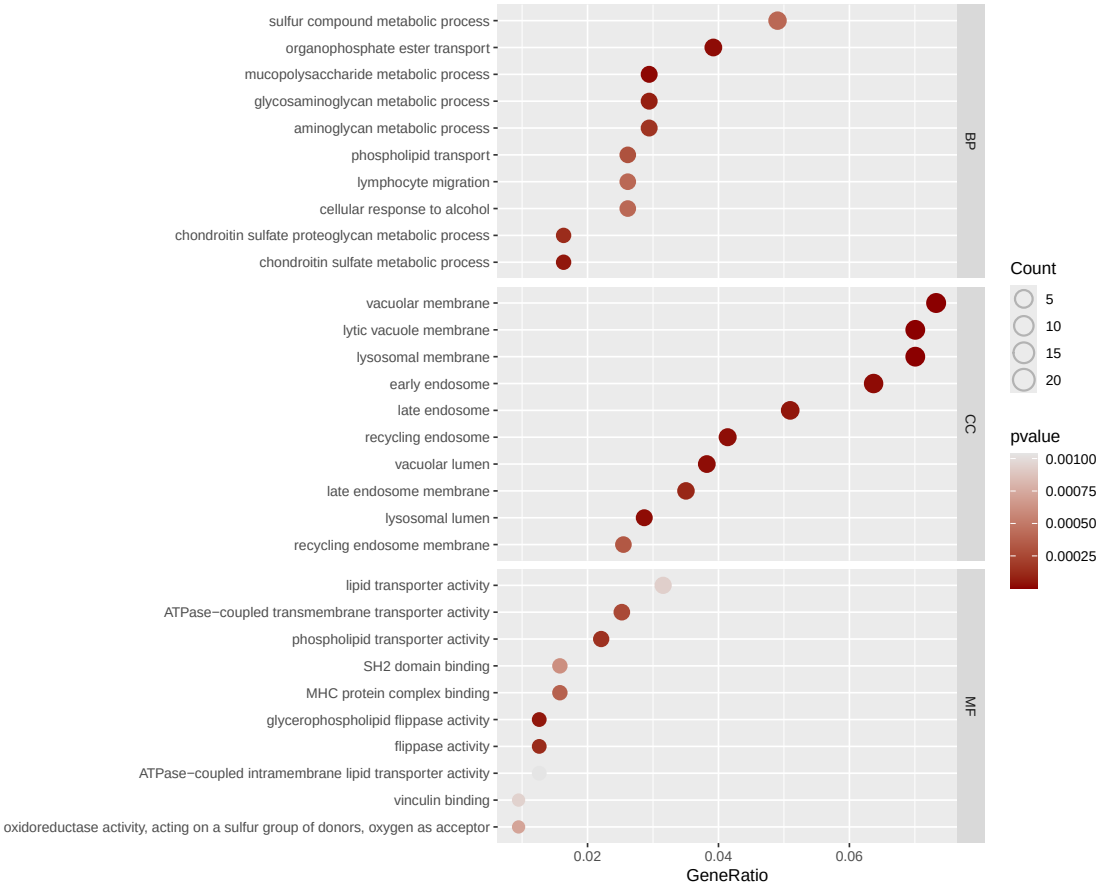


Figure 9: MRNA up GO enrichment

Figure 10 (下方图) 为图 MRNA down KEGG enrichment 概览。

(对应文件为 `Figure+Table/MRNA-down-KEGG-enrichment.pdf`)

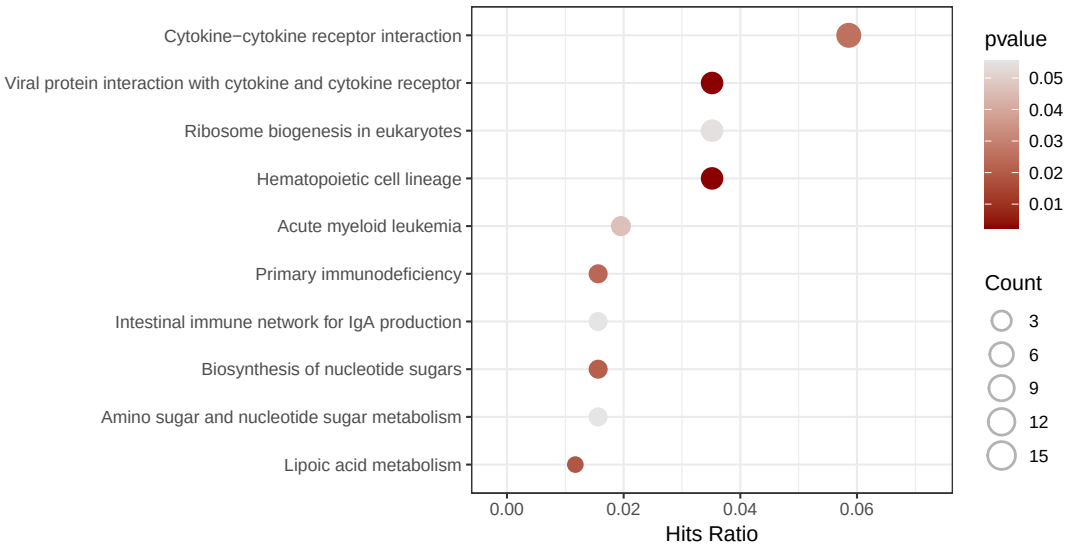


Figure 10: MRNA down KEGG enrichment

Figure 11 (下方图) 为图 MRNA down GO enrichment 概览。

(对应文件为 `Figure+Table/MRNA-down-GO-enrichment.pdf`)

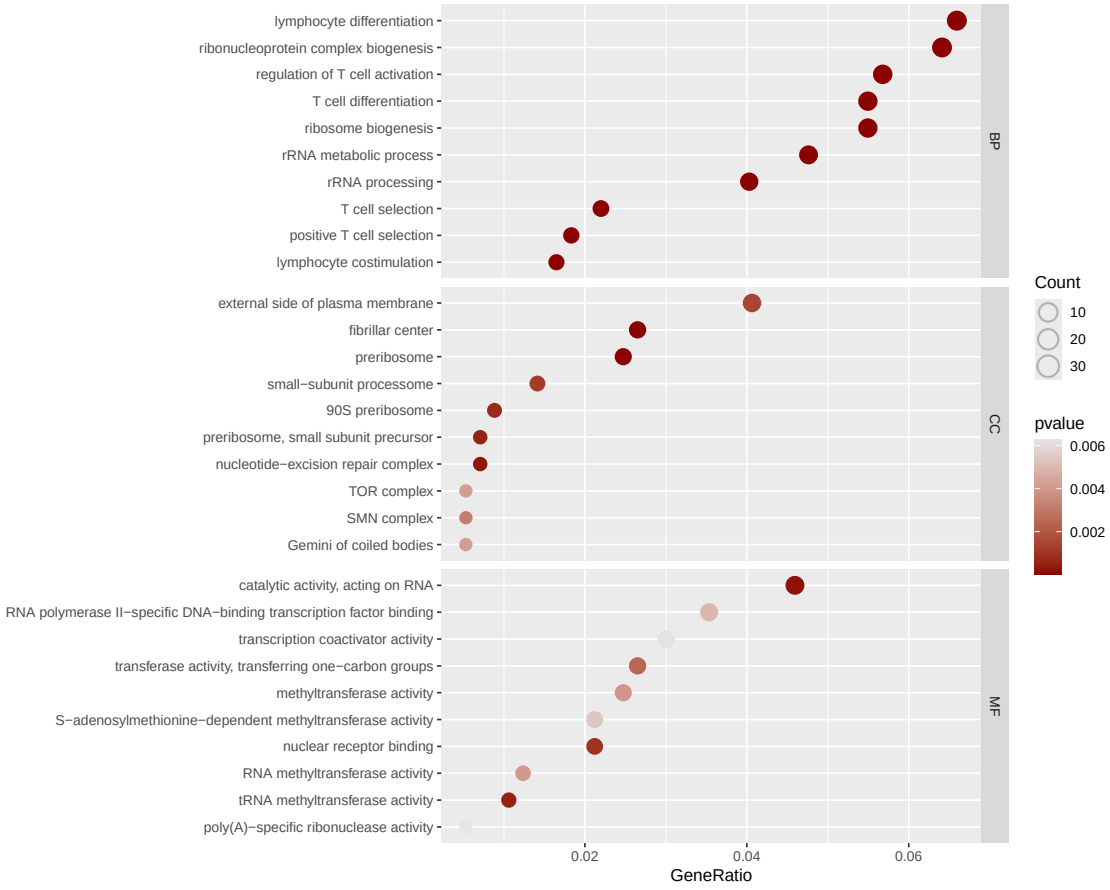


Figure 11: MRNA down GO enrichment

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. 3。

数据特征如下：

Data Frame Summary
Dimensions: 296 x 6
Duplicates: 93

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	age_at_index [integer]	Mean (sd) : 67.4 (8.5) min < med < max: 39 < 69 < 84 IQR (CV) : 12 (0.1)	40 distinct values	: : : : : : : . : : : : : : . . : : : : : : : :	291 (98.3%)	5 (1.7%)
2	vital_status [character]	1. Alive 2. Dead	228 (77.0%) 68 (23.0%)	IIIIIIIIIIIIIII IIII	296 (100.0%)	0 (0.0%)
3	gender [character]	1. female 2. male	78 (26.4%) 218 (73.6%)	IIIII IIIIIIIIIIIIII	296 (100.0%)	0 (0.0%)
4	tumor_grade [character]	1. Not Reported	296 (100.0%)	IIIIIIIIIIIIIIIIIIII	296 (100.0%)	0 (0.0%)
5	ajcc_pathologic_stage [character]	1. Stage IA 2. Stage IB 3. Stage II 4. Stage IIA 5. Stage IIB	72 (24.3%) 107 (36.1%) 1 (0.3%) 53 (17.9%) 63 (21.3%)	IIII IIIIIII III IIII	296 (100.0%)	0 (0.0%)
6	classification_of_tumor [character]	1. not reported	296 (100.0%)	IIIIIIIIIIIIIIIIIIII	296 (100.0%)	0 (0.0%)

将 LUSC 数据 (count) 标准化后 (同 MRNA 的方法)，以生存状态为指标 (Fig. 12)，以 EFS 算法，进行 Feature selection, 得到 Top 30 基因, 统计得分见 Fig. 13。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. 4。共 9 个基因:。

Table 3 (下方表格) 为表格 LUSC metadata 概览。

(对应文件为 `Figure+Table/LUSC-metadata.xlsx`)

> 注: 表格共有 296 行 87 列, 以下预览的表格可能省略部分数据; 含有 296 个唯一 ‘sample’。

Table 3: LUSC metadata

sample	group	lib.size	norm.f...	barcode	patient	shortL...	defini...	sample.....9	sample.....10	...
TCGA-1...	Dead	372553761		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Alive	434925071		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Dead	397148181		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Dead	399758341		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Alive	407479171		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Alive	444292151		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Alive	217350121		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-1...	Alive	615553201		TCGA-1...	TCGA-1...	TP	Primar...	TCGA-1...	01	...
TCGA-2...	Dead	485268831		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
TCGA-2...	Alive	604430271		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
TCGA-2...	Alive	602266461		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
TCGA-2...	Alive	583659241		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
TCGA-2...	Dead	586316011		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
TCGA-2...	Dead	540988351		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
TCGA-2...	Alive	549216961		TCGA-2...	TCGA-2...	TP	Primar...	TCGA-2...	01	...
...	...	...	...	...	...	...	...	...	...	...

Figure 12 (下方图) 为图 LUSC Group distribution 概览。

(对应文件为 `Figure+Table/LUSC-Group-distribution.pdf`)

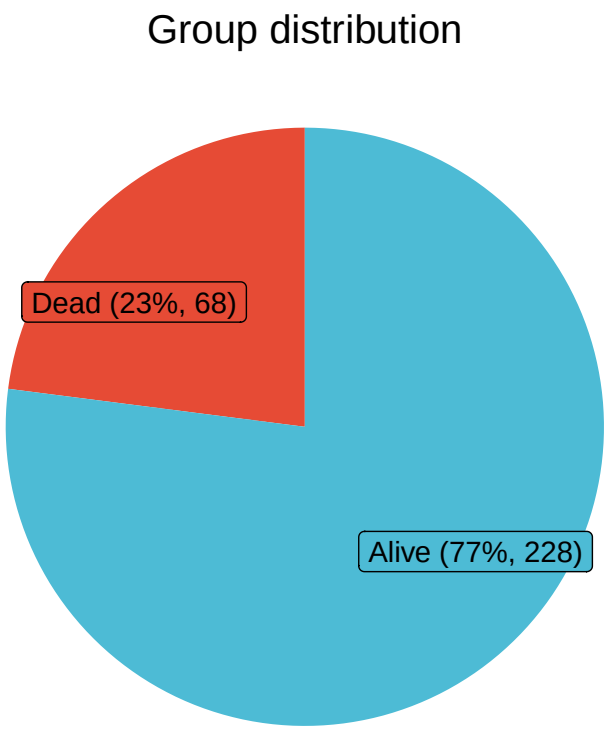


Figure 12: LUSC Group distribution

Figure 13 (下方图) 为图 LUSC Top Features Selected By EFS 概览。

(对应文件为 `Figure+Table/LUSC-Top-Features-Selected-By-EFS.pdf`)

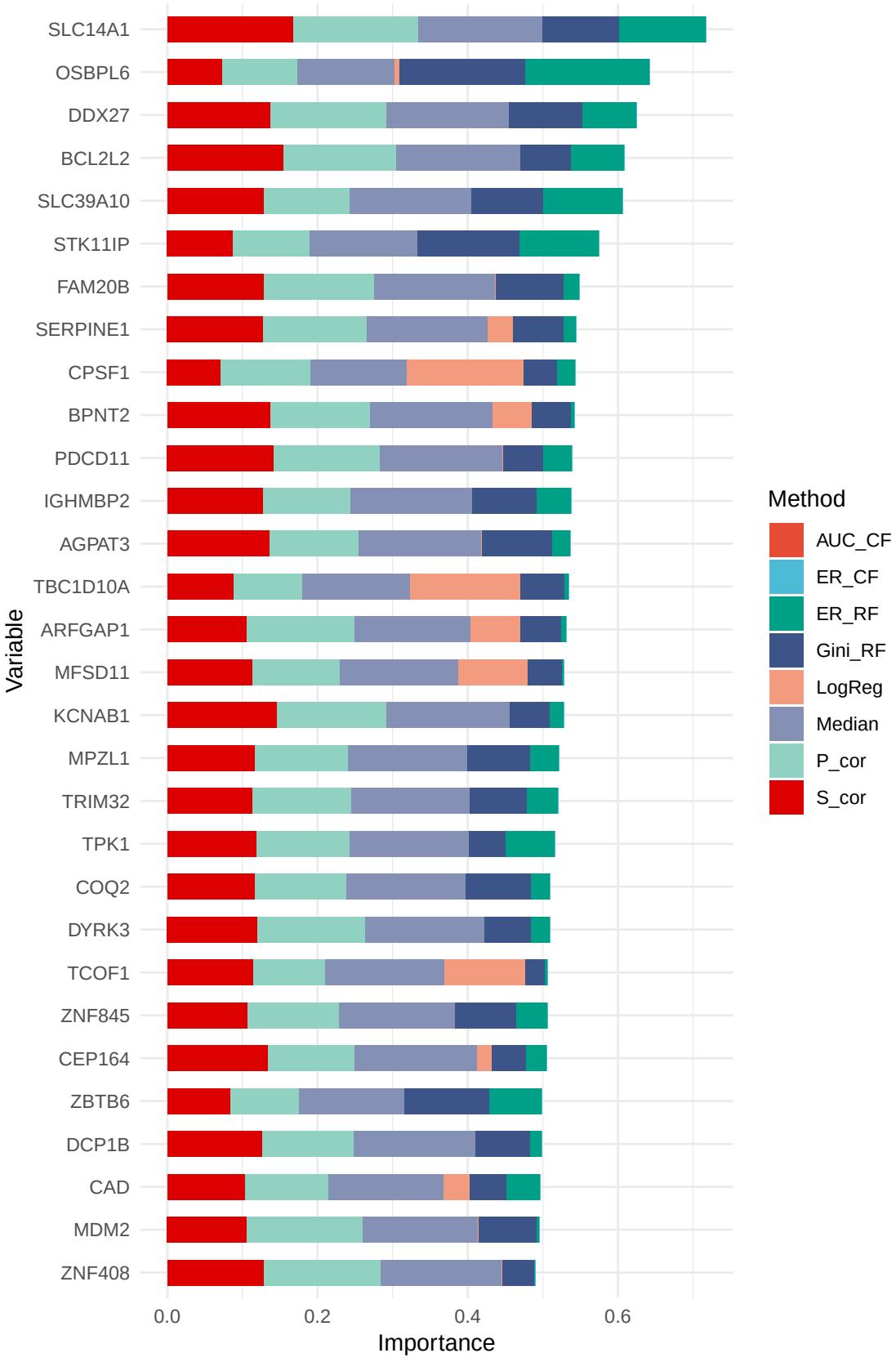

SLC14A1
OSBPL6
DDX27
BCL2L2
SLC39A10
STK11IP
FAM20B
SERPINE1
CPSF1
BPNT2
PDCD11
IGHMBP2
AGPAT3
TBC1D10A
ARFGAP1
MFSD11
KCNAB1
MPZL1
TRIM32
TPK1
COQ2
DYRK3
TCOF1
ZNF845
CEP164
ZBTB6
DCP1B
CAD
MDM2
ZNF408
Variable
0.0
0.2
0.4
0.6
Importance
Method
AUC_CF
ER_CF
ER_RF
Gini_RF
LogReg
Median
P_cor
S_cor

Figure 13: LUSC Top Features Selected By EFS

Table 4 (下方表格) 为表格 LUSC Uni COX cofficients filtered by EFS 概览。

(对应文件为 `Figure+Table/LUSC-Uni-COX-cofficients-filtered-by-EFS.csv`)

> 注：表格共有 9 行 7 列，以下预览的表格可能省略部分数据；含有 9 个唯一 'feature'。

Table 4: 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...

3.6 Survival 生存分析 (LUSC)

这些基因表达特征如 Fig. 14 热图所示。

建立预后特征，构建风险评分：

$$Score = \sum (expr(Gene) \times coef)$$

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。AUC 见 Fig. 16。第 1，3，5 年存活的患者，风险评分显著较低。

Figure 14 (下方图) 为图 LUSC risk score related genes heatmap 概览。

(对应文件为 `Figure+Table/LUSC-risk-score-related-genes-heatmap.pdf`)

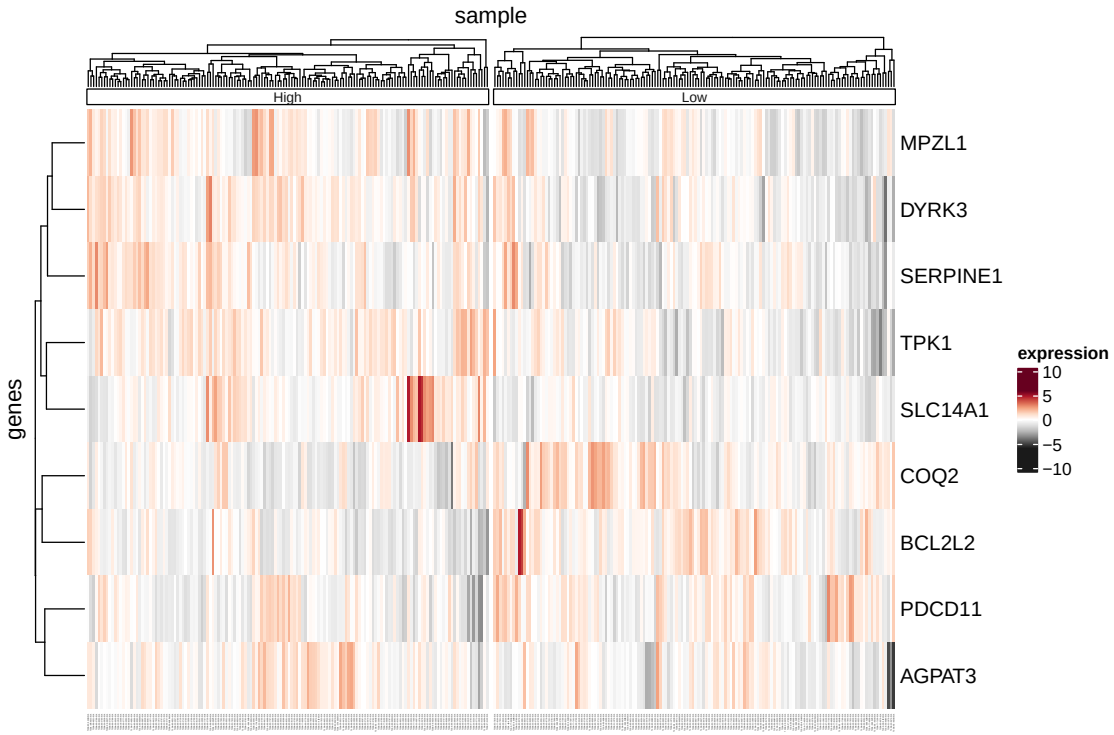


Figure 14: LUSC risk score related genes heatmap

Figure 15 (下方图) 为图 LUSC survival curve of risk score 概览。

(对应文件为 `Figure+Table/LUSC-survival-curve-of-risk-score.pdf`)

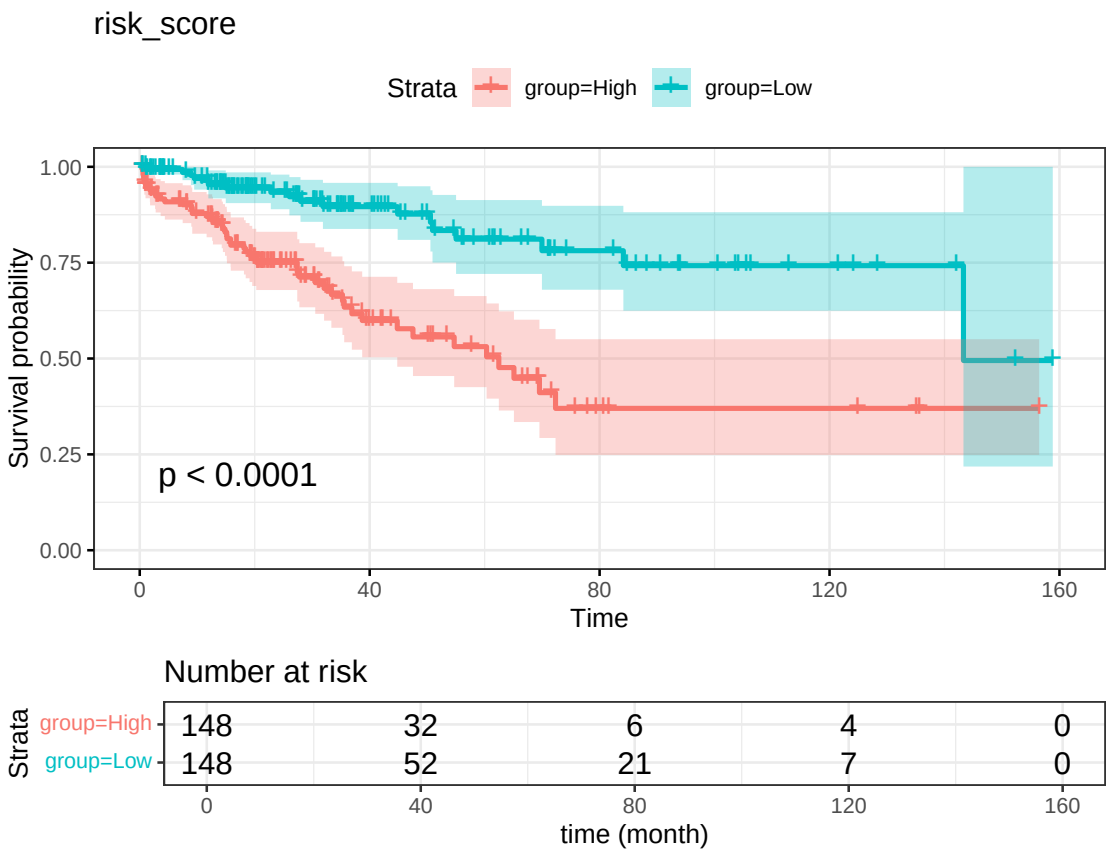


Figure 15: LUSC survival curve of risk score

Figure 16 (下方图) 为图 LUSC time ROC 概览。

(对应文件为 `Figure+Table/LUSC-time-ROC.pdf`)

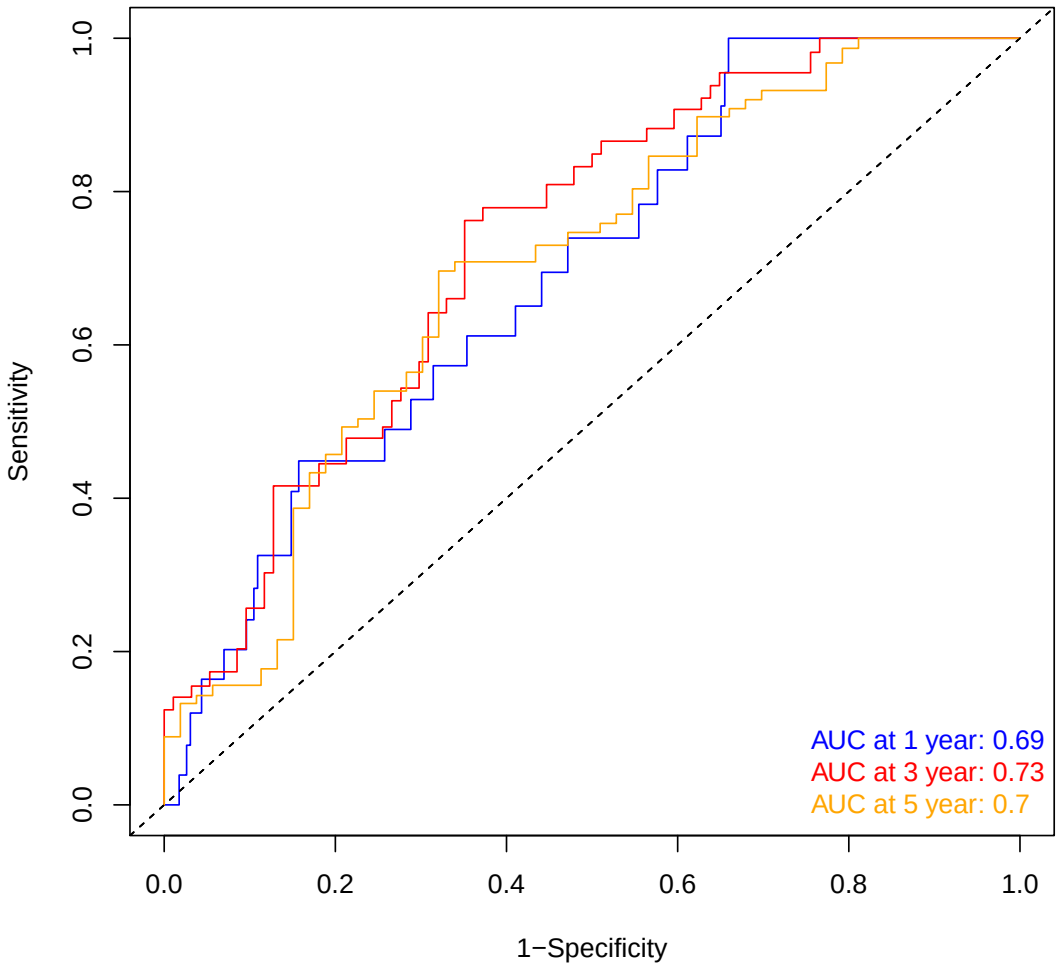


Figure 16: LUSC time ROC

Figure 17 (下方图) 为图 LUSC boxplot of risk score 概览。

(对应文件为 `Figure+Table/LUSC-boxplot-of-risk-score.pdf`)

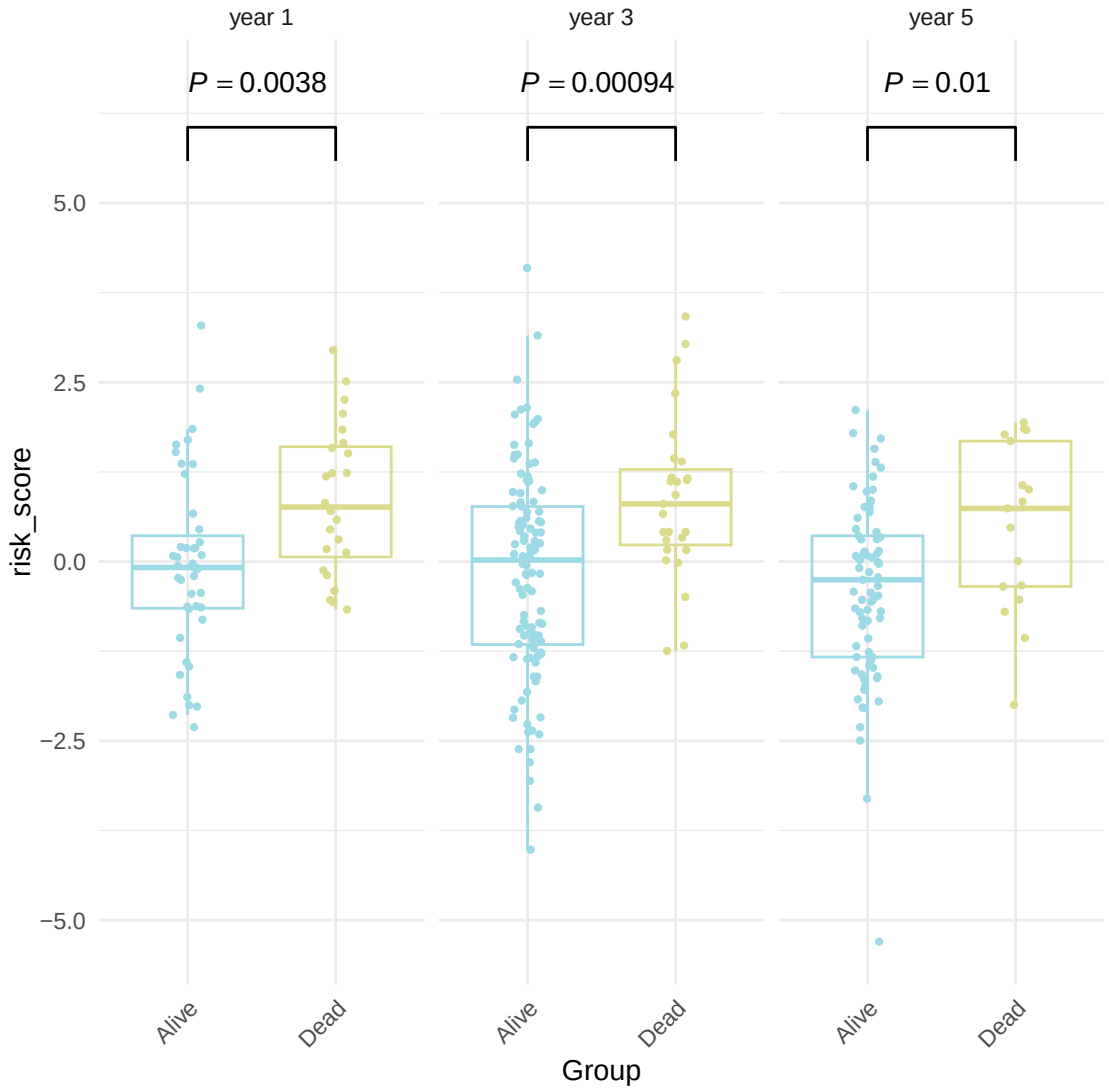


Figure 17: LUSC boxplot of risk score

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) 。数据特征如下：

Data Frame Summary
Dimensions: 291 x 5
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	rownames [character]	1. TCGA-18-3408-01A 2. TCGA-18-3409-01A 3. TCGA-18-3415-01A 4. TCGA-18-3416-01A 5. TCGA-18-3419-01A 6. TCGA-18-3421-01A 7. TCGA-18-4721-01A 8. TCGA-18-5592-01A 9. TCGA-21-1071-01A 10. TCGA-21-1072-01A [281 others]	1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 1 (0.3%) 281 (96.6%)	IIIIIIIIIIIIIIIIII	291 (100.0%)	0 (0.0%)
2	age [character]	1. <= 65 2. 65	111 (38.1%) 180 (61.9%)	IIIIIII IIIIIIIIIII	291 (100.0%)	0 (0.0%)
3	gender [character]	1. female 2. male	77 (26.5%) 214 (73.5%)	IIIII IIIIIIIIIIII	291 (100.0%)	0 (0.0%)
4	ajcc_pathologic_stage [character]	1. Stage I 2. Stage II	176 (60.5%) 115 (39.5%)	IIIIIIIIIII IIIIIII	291 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
5	risk_score [numeric]	Mean (sd) : 0 (1.4) min < med < max: -5.3 < 0.1 < 4.1 IQR (CV) : 1.8 (122)	291 distinct values	: : : : : : . : : : : : . : : : : : .	291 (100.0%)	0 (0.0%)

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. 6。

Table 6 (下方表格) 为表格 META Coefficients Of COX 概览。

(对应文件为 `Figure+Table/META-Coefficients-Of-COX.csv`)

> 注：表格共有 4 行 5 列，以下预览的表格可能省略部分数据；含有 4 个唯一 'feature'。

Table 6: 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

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。数据特征如下：

Data Frame Summary

Dimensions: 202 x 14

Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	sample [character]	1. GSM4750621 2. GSM4750622 3. GSM4750623 4. GSM4750624 5. GSM4750625 6. GSM4750626 7. GSM4750627 8. GSM4750628 9. GSM4750630 10. GSM4750631 [192 others]	1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 192 (95.0%)	IIIIIIIIIIIIIIIIIII	202 (100.0%)	0 (0.0%)
2	vital_status [character]	1. Alive 2. Dead	107 (53.0%) 95 (47.0%)	IIIIIIIIII IIIIIIIII	202 (100.0%)	0 (0.0%)
3	group [character]	1. Alive 2. Dead	107 (53.0%) 95 (47.0%)	IIIIIIIIII IIIIIIIII	202 (100.0%)	0 (0.0%)
4	days_to_last_follow_up [numeric]	Mean (sd) : 1403 (697.3) min < med < max: 4.9 < 1669.8 < 3062.5 IQR (CV) : 1111.8 (0.5)	186 distinct values	: : : : : . : : : : : : .	202 (100.0%)	0 (0.0%)
5	stage [character]	1. T1 2. T2	63 (31.2%) 139 (68.8%)	IIIIII IIIIIIIIIIIII	202 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
6	rownames [character]	1. GSM4750621 2. GSM4750622 3. GSM4750623 4. GSM4750624 5. GSM4750625 6. GSM4750626 7. GSM4750627 8. GSM4750628 9. GSM4750630 10. GSM4750631 [192 others]	1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 192 (95.0%)	IIIIIIIIIIIIIIIIII	202 (100.0%)	0 (0.0%)
7	title [character]	1. CAD_NA379PT_RNA_2115A_F4 2. CAD_NA380PT_RNA_2115A_H1 3. CAD_NA381PT_RNA_2115A_C3 4. CAD_NA382PT_RNA_2115A_G4 5. CAD_NA383PT_RNA_2115A_E7 6. CAD_NA384PT_RNA_2115A_E4 7. CAD_NA385PT_RNA_2115A_G5 8. CAD_NA386PT_RNA_2116A_B1_ 9. CAD_NA388PT_RNA_2116A_D5 10. CAD_NA389PT_RNA_2116A_D12 [192 others]	1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 192 (95.0%)	IIIIIIIIIIIIIIIIII	202 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
8	age.ch1 [character]	1. 63 2. 64 3. 65 4. 71 5. 70 6. 59 7. 61 8. 66 9. 68 10. 72 [29 others]	12 (5.9%) 12 (5.9%) 12 (5.9%) 10 (5.0%) 9 (4.5%) 9 (4.5%) 9 (4.5%) 9 (4.5%) 99 (49.0%)	I I I I IIIIIIIII	202 (100.0%)	0 (0.0%)
9	diagnosis.ch1 [character]	1. Squamous Cell Carcinoma o	202 (100.0%)	IIIIIIIIIIIIIIIIIII	202 (100.0%)	0 (0.0%)
10	os_event.ch1 [character]	1. 0 2. 1	107 (53.0%) 95 (47.0%)	IIIIIIIII IIIIIIIII	202 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
11	os_mo.ch1 [character]	1. 55.6931506849315 2. 60.6904109589041 3. 69.041095890411 4. 15.0904109589041 5. 60.1972602739726 6. 60.2958904109589 7. 60.7890410958904 8. 61.0520547945205 9. 61.0849315068493 10. 61.7095890410959 [176 others]	3 (1.5%) 3 (1.5%) 3 (1.5%) 2 (1.0%) 2 (1.0%) 2 (1.0%) 2 (1.0%) 2 (1.0%) 2 (1.0%) 179 (88.6%)	IIIIIIIIIIIIIIII	202 (100.0%)	0 (0.0%)
12	Sex.ch1 [character]	1. Female 2. Male	71 (35.1%) 131 (64.9%)	IIIIIII IIIIIIIIIII	202 (100.0%)	0 (0.0%)
13	Stage.ch1 [character]	1. T1a 2. T1b 3. T2a 4. T2b	23 (11.4%) 40 (19.8%) 74 (36.6%) 65 (32.2%)	II III IIIIIII IIIIII	202 (100.0%)	0 (0.0%)
14	tissue.ch1 [character]	1. lung	202 (100.0%)	IIIIIIIIIIIIIIIIIII	202 (100.0%)	0 (0.0%)

Table 8 (下方表格) 为表格 LUSC GSE157010 metadata 概览。

(对应文件为 `Figure+Table/LUSC-GSE157010-metadata.csv`)

注: 表格共有 202 行 14 列, 以下预览的表格可能省略部分数据; 含有 202 个唯一 'sample'。

Table 8: 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
...	...	...	...	...	...	...	...	...	...

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. 18。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. 21，高风险组与低风险组显著差异。ROC 曲线见 Fig. 20。第 1，3，5 年风险评分差异见 Fig. 19 。

Figure 18 (下方图) 为图 GEO LUSC risk score related genes heatmap 概览。

(对应文件为 `Figure+Table/GEO-LUSC-risk-score-related-genes-heatmap.pdf`)

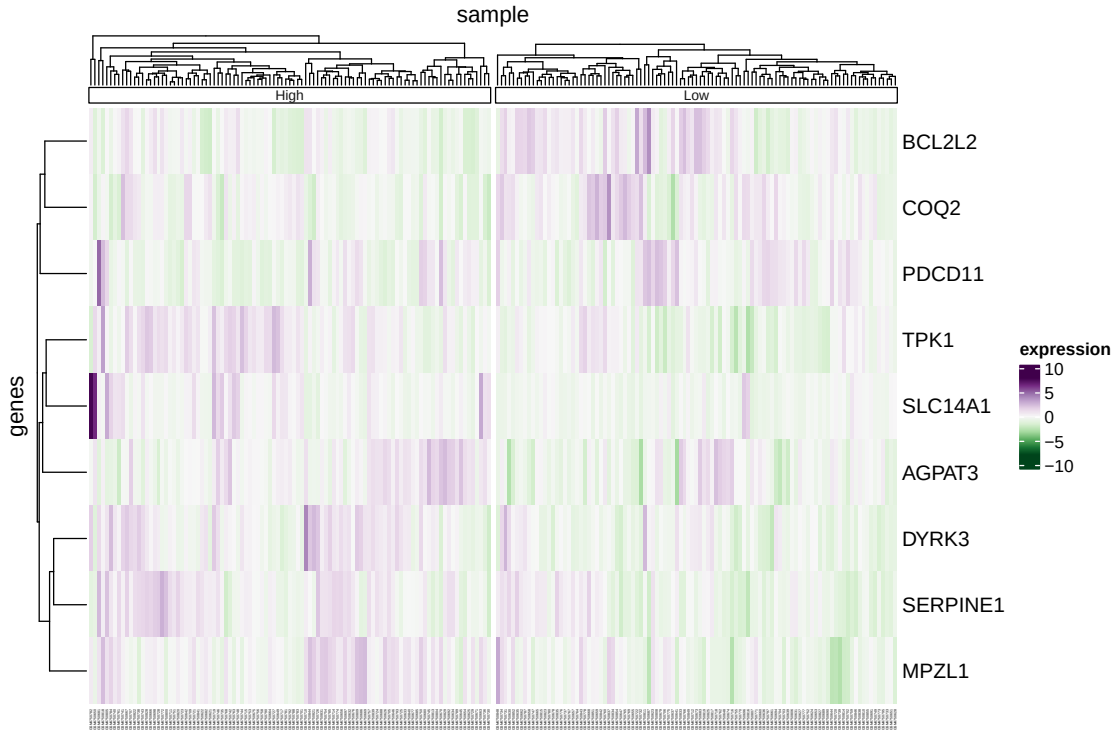


Figure 18: GEO LUSC risk score related genes heatmap

Figure 19 (下方图) 为图 GEO LUSC boxplot of risk score 概览。

(对应文件为 `Figure+Table/GEO-LUSC-boxplot-of-risk-score.pdf`)

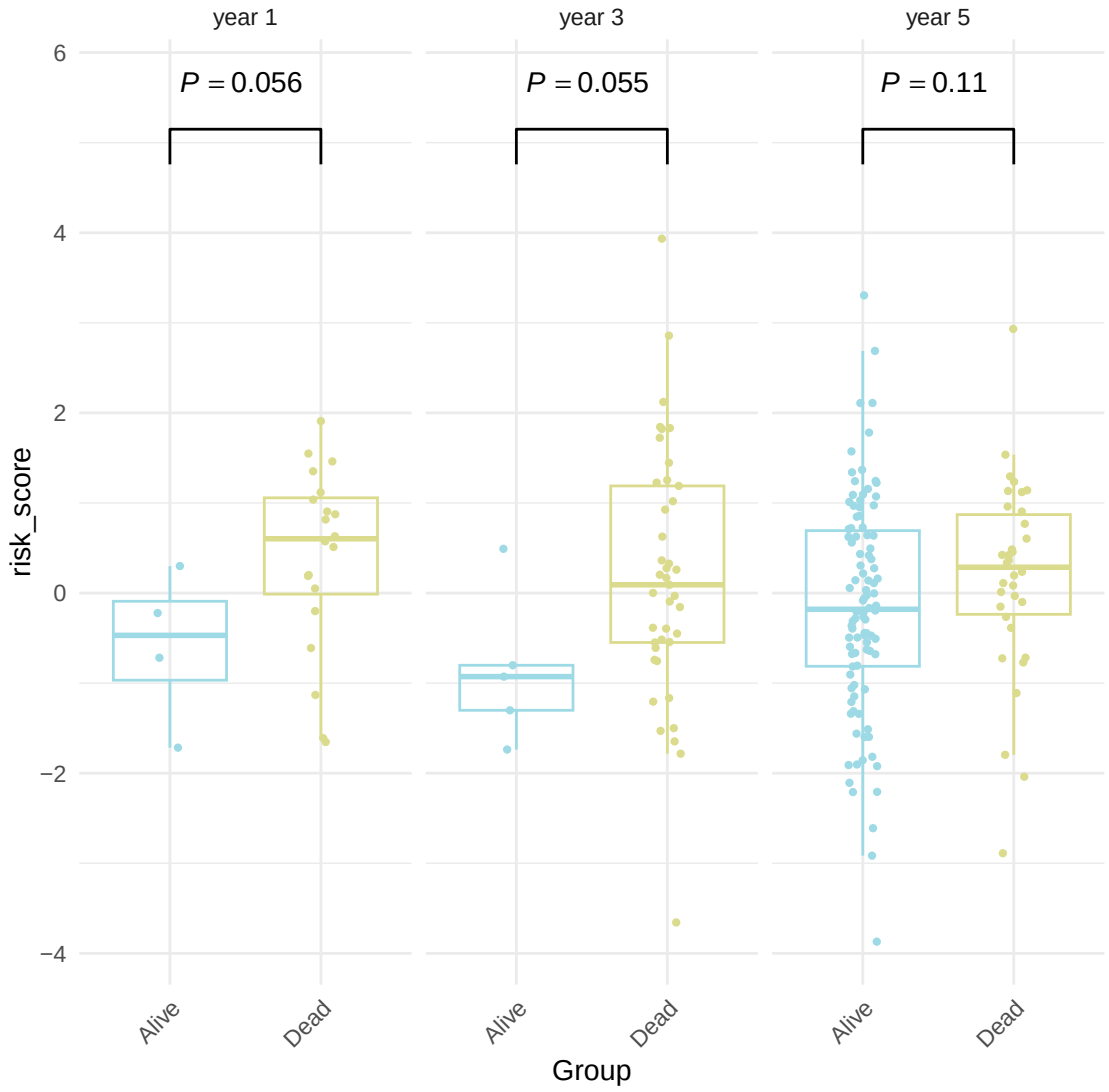


Figure 19: GEO LUSC boxplot of risk score

Figure 20 (下方图) 为图 GEO LUSC time ROC 概览。

(对应文件为 `Figure+Table/GEO-LUSC-time-ROC.pdf`)

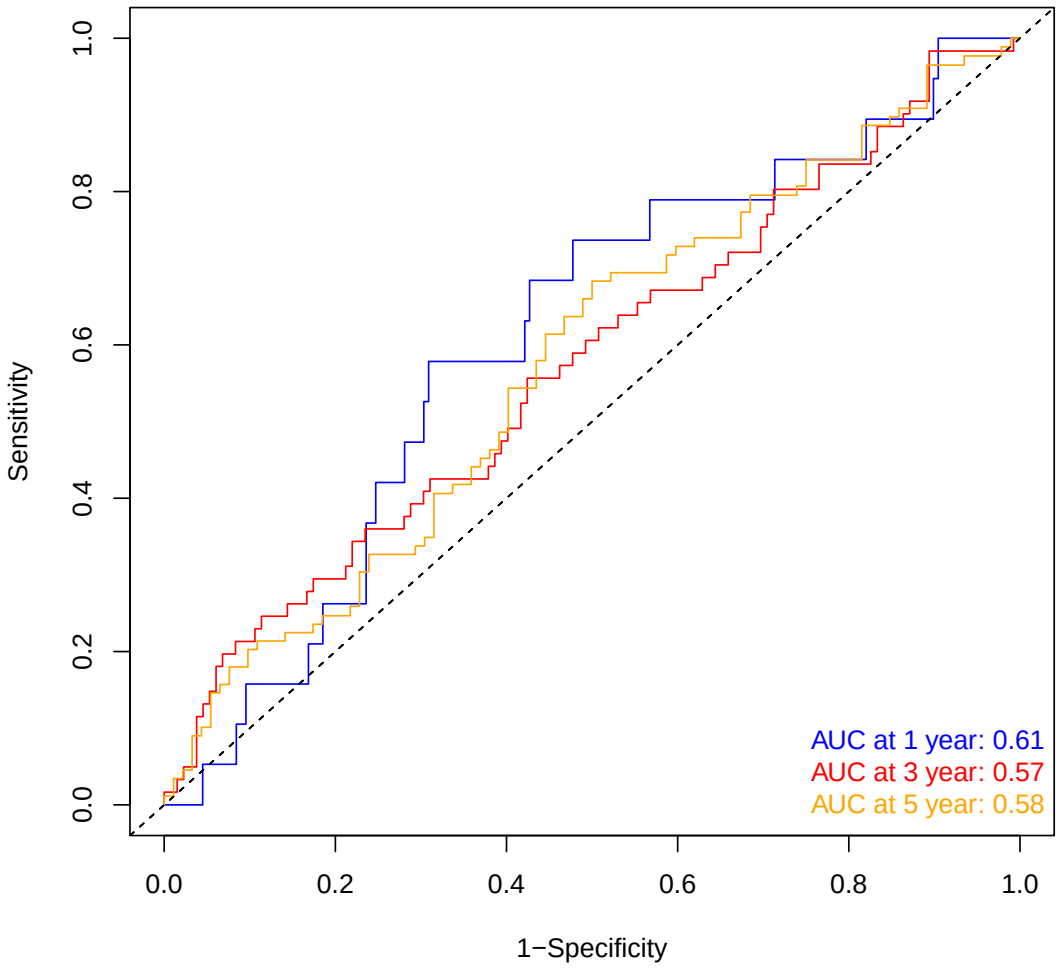


Figure 20: GEO LUSC time ROC

Figure 21 (下方图) 为图 GEO LUSC survival curve of risk score 概览。

(对应文件为 `Figure+Table/GEO-LUSC-survival-curve-of-risk-score.pdf`)

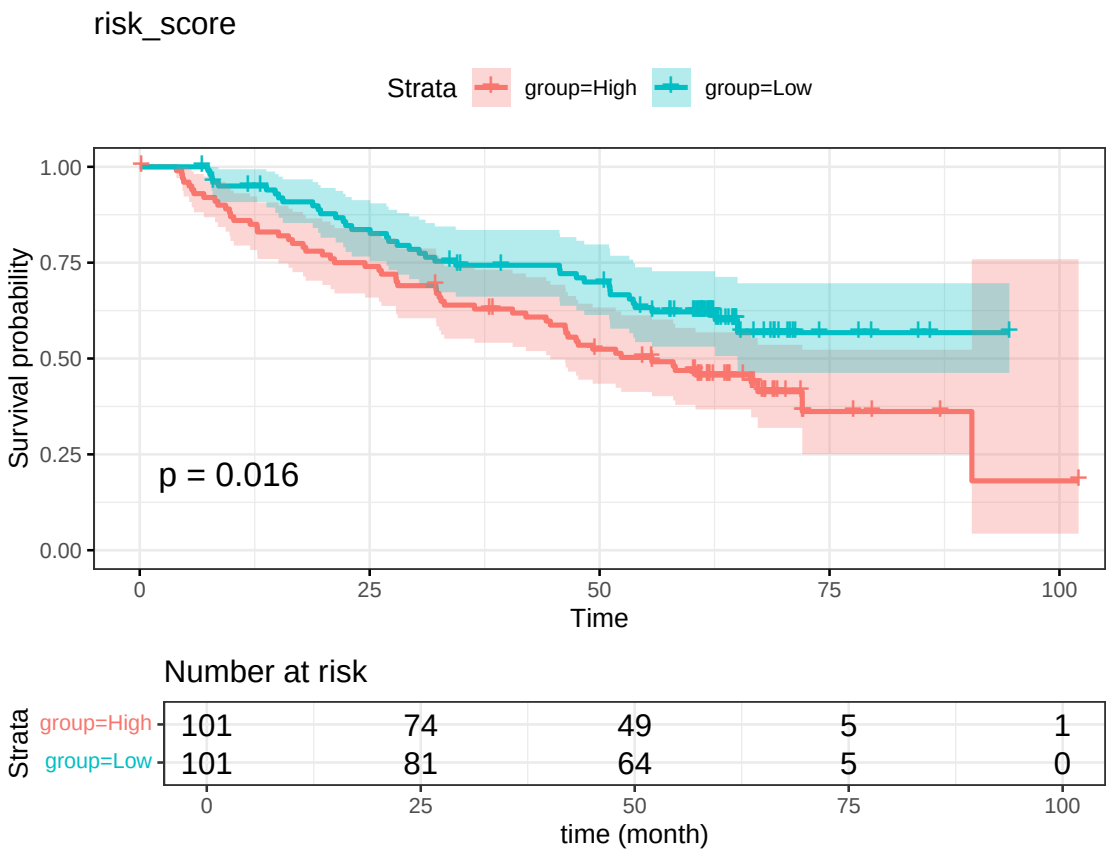


Figure 21: GEO LUSC survival curve of risk score

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. 22。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. 23。

Figure 22 (下方图) 为图 LUSC immune Scores Plot 概览。

(对应文件为 `Figure+Table/LUSC-immune-Scores-Plot.pdf`)

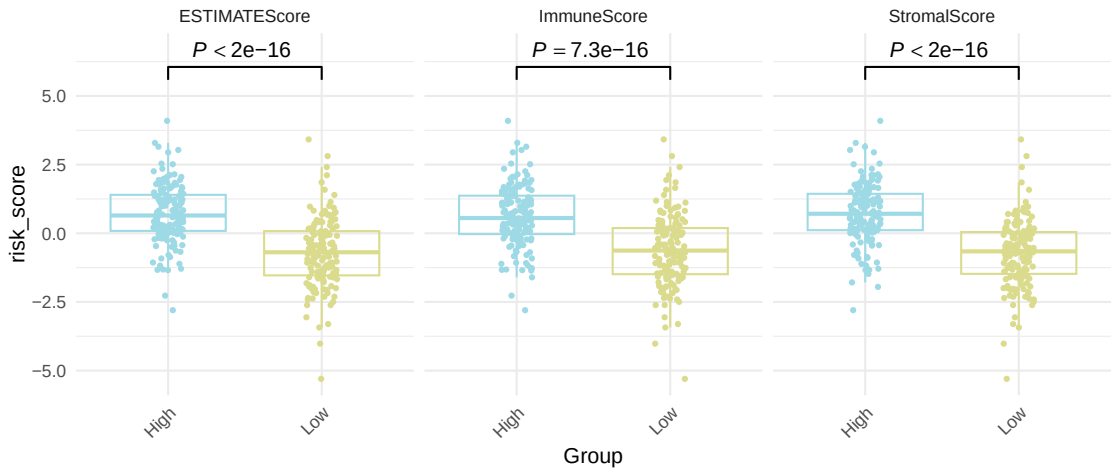


Figure 22: LUSC immune Scores Plot

Figure 23 (下方图) 为图 LUSC Top10 Immune Related Genes 概览。

(对应文件为 `Figure+Table/LUSC-Top10-Immune-Related-Genes.pdf`)

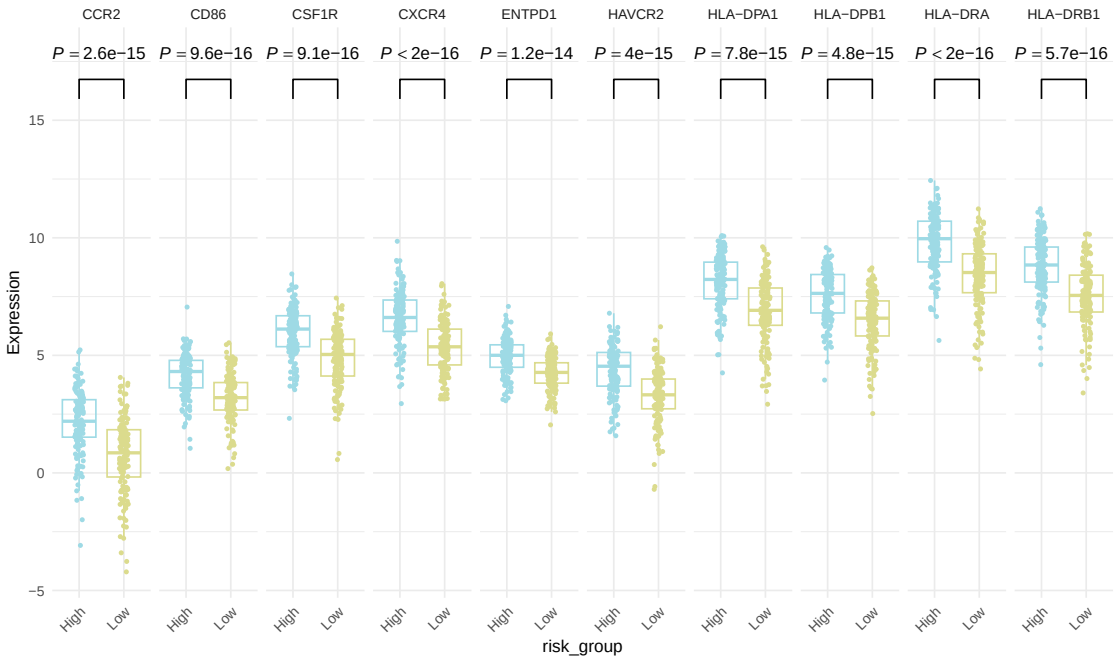


Figure 23: LUSC Top10 Immune Related Genes

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. 24。所有上调 DEGs 有 539 个，下调共 781；一共 1278 个 (非重复)。。

'LNCRNA DEGs data' 数据已全部提供。

(对应文件为 `Figure+Table/LNCRNA-DEGs-data`)

> 注：文件夹 Figure+Table/LNCRNA-DEGs-data 共包含 3 个文件。
>
> 1. 1_Early_stage - Healthy.csv
> 2. 2_Advanced_stage - Healthy.csv
> 3. 3_Advanced_stage - Early_stage.csv

Figure 24 (下方图) 为图 LNCRNA Difference intersection 概览。

(对应文件为 `Figure+Table/LNCRNA-Difference-intersection.pdf`)

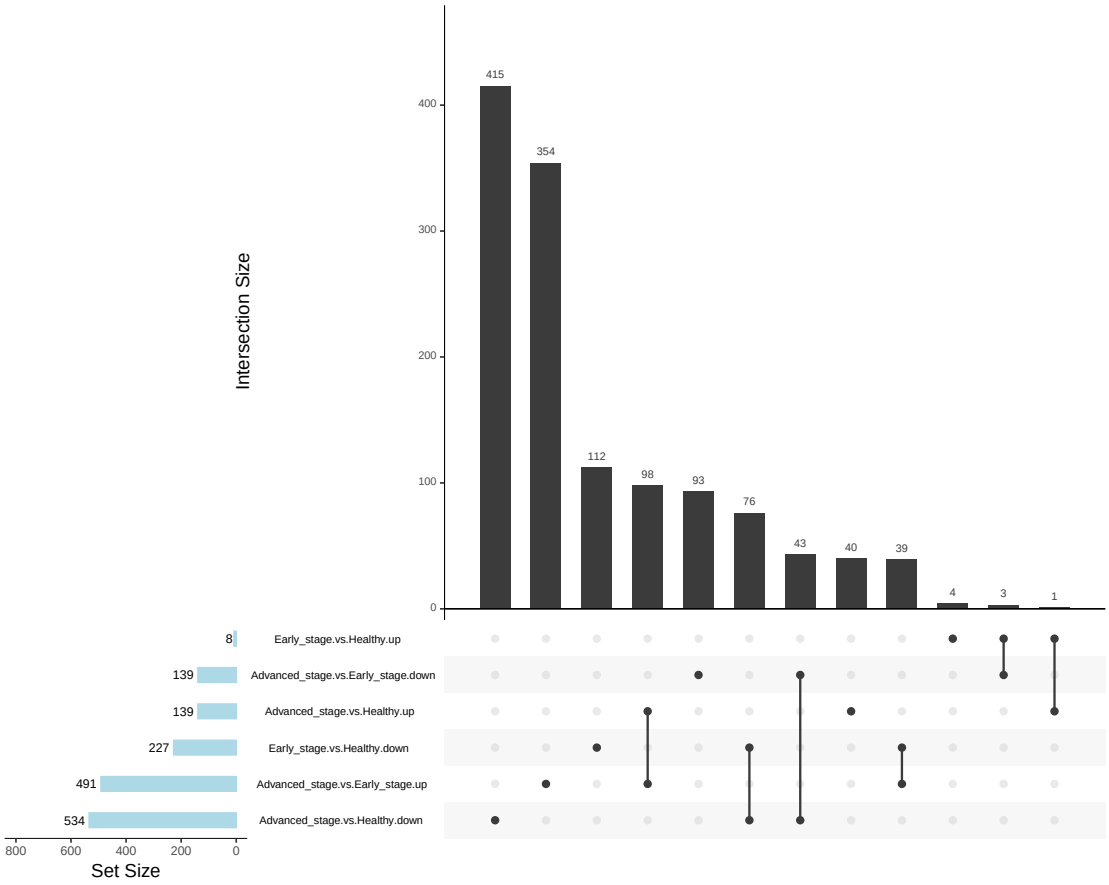


Figure 24: LNCRNA Difference intersection

All_intersection :

(上述信息框内容已保存至 `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. 25。共包含 4 个 mRNA，4 个 lncRNA，52 对关联关系。

Figure 25 (下方图) 为图 Significant Correlation mrna lncRNA 概览。

(对应文件为 `Figure+Table/Significant-Correlation-mrna-lncRNA.pdf`)

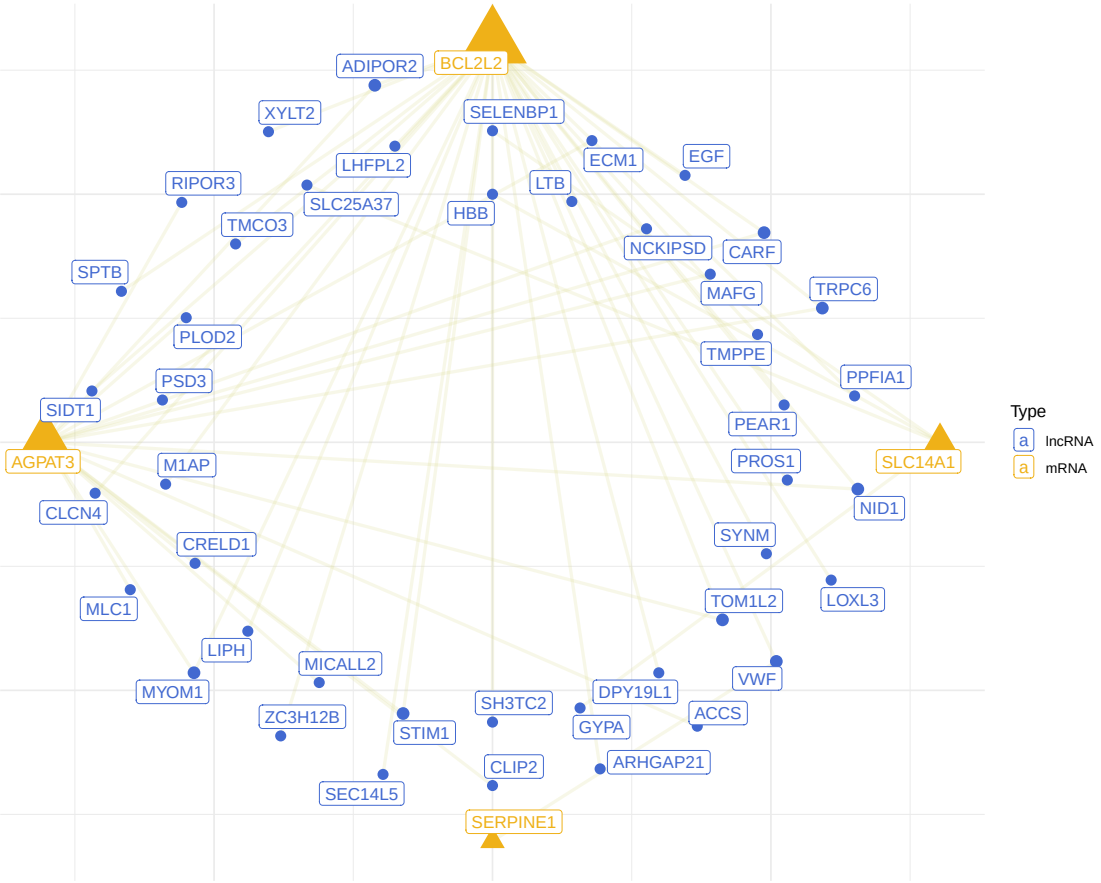


Figure 25: Significant Correlation mrna lncRNA

Table 9 (下方表格) 为表格 Significant correlation 概览。

(对应文件为 `Figure+Table/Significant-correlation.csv`)

注：表格共有 52 行 7 列，以下预览的表格可能省略部分数据；含有 4 个唯一 'mRNA'。

Table 9: 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	**
...	...	...	...	...	...	...

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.

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).