**生信分析报告**

**项目标题： 基于单细胞和普通转录组分析基质体相关基因**   
 **是肺腺癌免疫治疗反应和预后的预测指标 ;**

**单 号： BSLL250322 ;**

**分析人员： 黄礼闯 ;**

**分析类型： 生信分析 ;**

**委 托 人： 田彩平 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程

# 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 UCSCXenaTools 癌症相关数据获取 (Dataset: LUAD)

以 R 包 UCSCXenaTools (1.6.0) (2019, Journal of Open Source Software)1 获取 TcgaTargetGtex 类型数据。

## 2.3 Limma 差异分析 (Dataset: LUAD)

以 R 包 edgeR (4.4.2) ()2 对数据预处理。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 limma (3.62.2) (2005)3 差异分析。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。创建设计矩阵，对比矩阵，差异分析：LUAD vs Normal。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 adj.P.Val 小于 0.05，|Log2(FC)| 大于 1 的统计结果。

## 2.4 ssGSEA 单样本GSEA富集分析 (Dataset: LUAD)

以 R 包 GSVA (2.0.7) 用于 ssGSEA 分析。

## 2.5 WGCNA 分析 (Dataset: LUAD)

以 R 包 WGCNA (1.73) (2008, **IF:2.9**, Q1, BMC Bioinformatics)4 对数据作共表达分析。分析方法参考 <https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/index.html>。 以 WGCNA::pickSoftThreshold 预测最佳 soft thresholding powers。 选择 power 为 11, 以 WGCNA::blockwiseModules 创建共表达网络，检测基因模块。

## 2.6 TCGA 数据获取 (Dataset: LUAD)

以 R 包 TCGAbiolinks (2.35.1) (2015, **IF:16.6**, Q1, Nucleic Acids Research)5 获取 TCGA-LUAD 数据集。

## 2.7 COX 回归 (Dataset: TCGA\_LUAD)

以 R 包 edgeR (4.4.2) ()2 对数据预处理。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 R 包 survival (3.8.3) 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 10 交叉验证获得模型。

## 2.8 Survival 生存分析 (Dataset: TCGA\_LUAD)

以 R 包 survival (3.8.3) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

## 2.9 GEO 数据获取 (Dataset: LUAD\_GSE31210)

以 R 包 GEOquery (2.74.0) 获取 GSE31210 数据集。

## 2.10 Survival 生存分析 (Dataset: LUAD\_GSE31210)

使用 log2 和 limma::normalizeBetweenArrays 对数据标准化。 以 R 包 survival (3.8.3) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

## 2.11 Survival 生存分析 (Dataset: INDIVIDUALS)

以 R 包 edgeR (4.4.2) ()2 对数据预处理。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 R 包 edgeR (4.4.2) ()2 对数据预处理。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)。使用标准化过的基因表达数据。 以 R 包 survival (3.8.3) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

## 2.12 estimate 免疫评分 (Dataset: TCGA\_LUAD\_REGROUP)

以 R 包 estimate (1.0.13) (2013, **IF:14.7**, Q1, Nature communications)6 预测数据集的 stromal, immune, estimate 得分。

## 2.13 GEO 数据获取 (Dataset: SC\_LUAD)

以 R 包 GEOquery (2.74.0) 获取 GSE189357 数据集。

## 2.14 Seurat 集成单细胞数据分析 (Dataset: LUAD)

使用 Seurat R 包 (5.2.1) 进行单细胞数据质量控制 (QC) 和下游分析。依据 <https://satijalab.org/seurat/articles/integration_introduction> 为指导对单细胞数据预处理。 一个细胞至少应有 500 个基因，并且基因数量小于 7500。线粒体基因的比例小于 10%。根据上述条件，获得用于下游分析的高质量细胞。 执行标准 Seurat 分析工作流 (NormalizeData, FindVariableFeatures, ScaleData, RunPCA)。以 ElbowPlot 判断后续分析的 PC 维度。 以 Seurat::IntegrateLayers 集成数据，去除批次效应 (使用 HarmonyIntegration 方法)。在 1-15 PC 维度下，以 Seurat::FindNeighbors 构建 Nearest-neighbor Graph。随后在 1.2 分辨率下，以 Seurat::FindClusters 函数识别细胞群并以 Seurat::RunUMAP 进行 UMAP 聚类。 以 Seurat::FindAllMarkers (LogFC 阈值 0.25; 最小检出率 0.25) 为所有细胞群寻找 Markers。 以 Python 工具 SCSA ((2020, **IF:2.8**, Q2, Frontiers in genetics)7) (<https://github.com/bioinfo-ibms-pumc/SCSA>) 对细胞群注释。

# 3 分析结果

## 3.1 UCSCXenaTools 癌症相关数据获取 (LUAD)

获取 UCSC Xena 的 TcgaTargetGtex 数据。共 801 个数据。样本组织为：Lung (n=801) 。样本类型为：Normal Tissue (n=288) , Primary Tumor (n=513) 。数据来源为：GTEX (n=288) , TCGA (n=513) 。

**Tab.** **1** LUAD all sample metadata

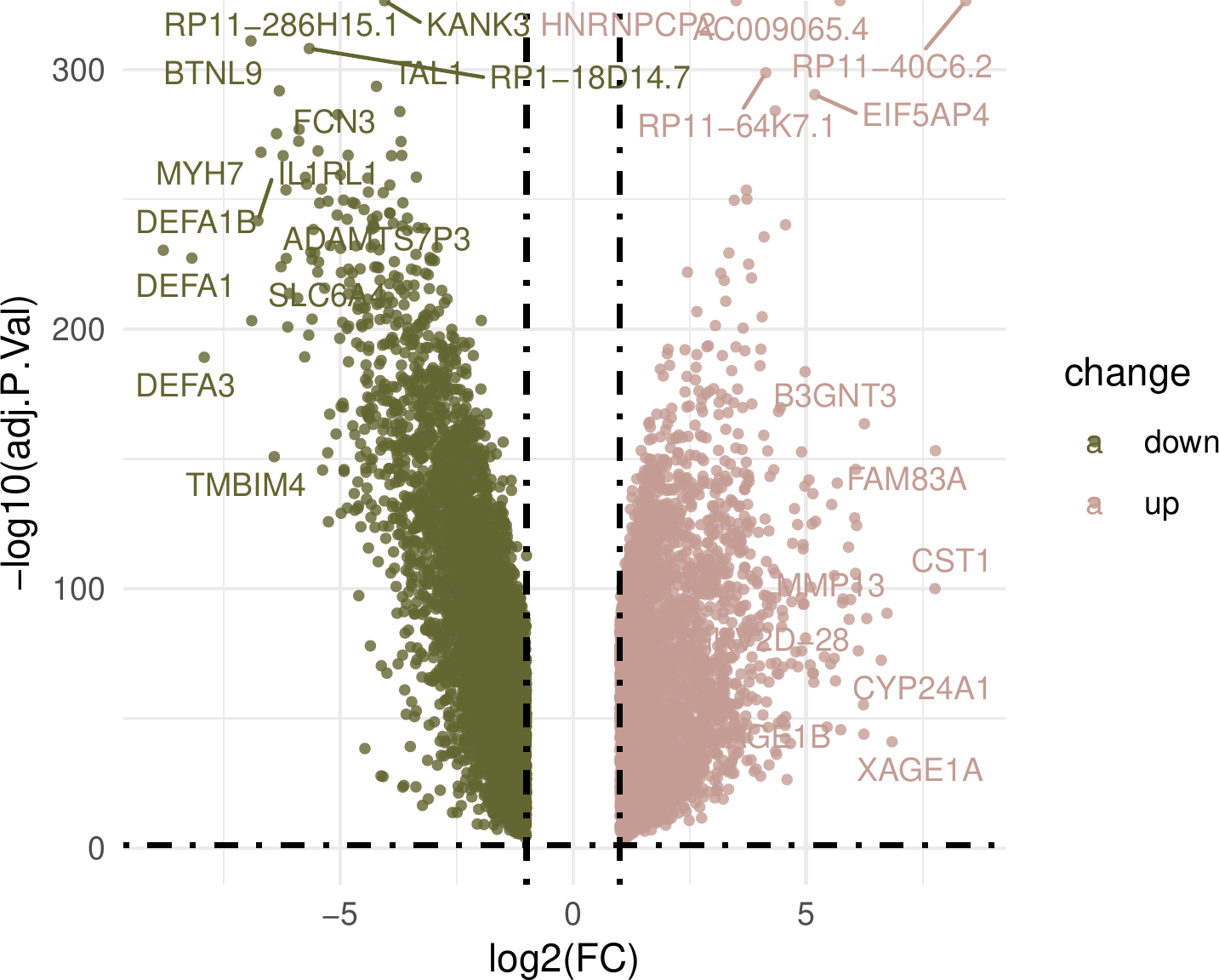
| Sample | Group | Detailed category | Primary disease o... | primary site |
| --- | --- | --- | --- | --- |
| TCGA-55-A4DF-01 | LUAD | Lung Adenocarcinoma | Lung Adenocarcinoma | Lung |
| TCGA-50-5939-01 | LUAD | Lung Adenocarcinoma | Lung Adenocarcinoma | Lung |
| TCGA-69-8254-01 | LUAD | Lung Adenocarcinoma | Lung Adenocarcinoma | Lung |
| TCGA-62-A470-01 | LUAD | Lung Adenocarcinoma | Lung Adenocarcinoma | Lung |
| TCGA-69-7761-01 | LUAD | Lung Adenocarcinoma | Lung Adenocarcinoma | Lung |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.1.0\_UCSCXenaTools\_癌症相关数据获取\_(LUAD)/LUAD-all-sample-metadata.csv)**

Tab. **[1](#LUAD-all-sample-metadata)** 为使用的 XenaDatasets 所有样本的元数据。

## 3.2 Limma 差异分析 (LUAD)

将 TcgaTargetGtex-LUAD 数据转化为 counts 型数据 (原数据为 log2(counts + 1)) 。样本分组：LUAD (n=513) , Normal (n=288) 。以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) (Batch: GTEX (n=288) , TCGA (n=513) )。差异分析：LUAD vs Normal。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。上调或下调 DEGs 统计：up (n=3687) , down (n=4514)



**Fig.** **1** LUAD LUAD vs Normal

**(File path: Figure+Table/3.2.0\_Limma\_差异分析\_(LUAD)/LUAD-LUAD-vs-Normal.pdf)**

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

**(See: Figure+Table/3.2.0\_Limma\_差异分析\_(LUAD)/LUAD-LUAD-vs-Normal-content)**

Fig. **[1](#LUAD-LUAD-vs-Normal)** 为 LUAD - Normal 差异分析火山图。

**Tab.** **2** LUAD data LUAD vs Normal

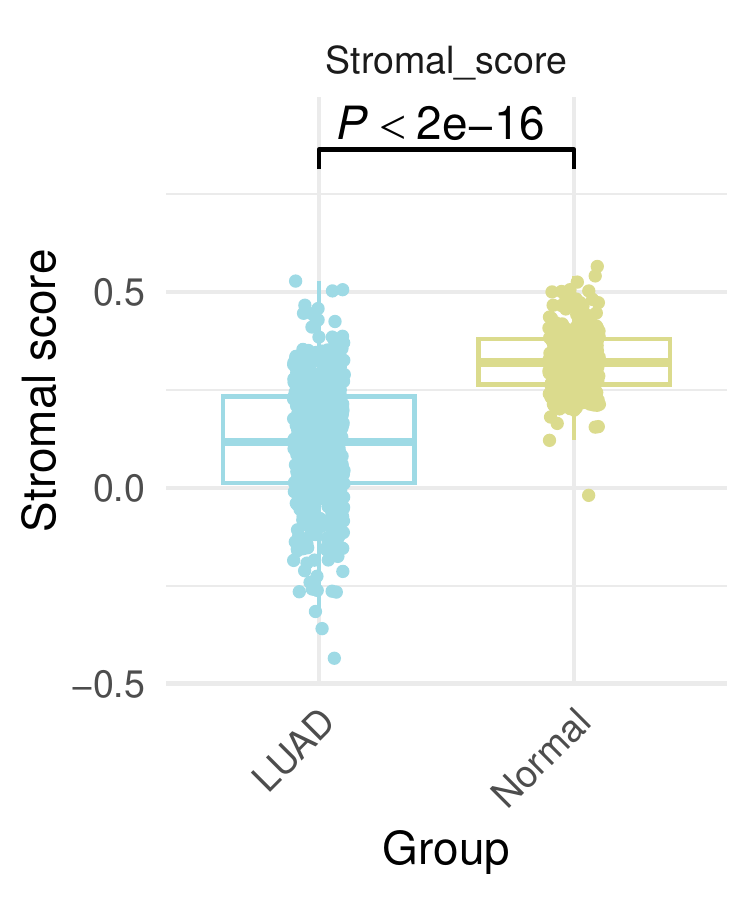
| Hgnc symbol | LogFC | Adj.P.Val | Rownames | Id |
| --- | --- | --- | --- | --- |
| RP11-40C6.2 | 8.425 | 0 | ENSG00000219928.2 | ENSG00000219928.2 |
| AC009065.4 | 5.725 | 0 | ENSG00000279473.1 | ENSG00000279473.1 |
| HNRNPCP2 | 3.498 | 0 | ENSG00000204253.4 | ENSG00000204253.4 |
| KANK3 | -4.045 | 0 | ENSG00000186994.11 | ENSG00000186994.11 |
| RP11-286H15.1 | -6.914 | 8.87199999999773e... | ENSG00000272789.1 | ENSG00000272789.1 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.2.0\_Limma\_差异分析\_(LUAD)/LUAD-data-LUAD-vs-Normal.csv)**

Tab. **[2](#LUAD-data-LUAD-vs-Normal)** 为 LUAD - Normal 差异分析统计表格。

## 3.3 ssGSEA 单样本GSEA富集分析 (LUAD)

从 (2012, Molecular & cellular proteomics : MCP)8 (补充材料) 获取基质体 (matrisome) 数据。将**基因集** (Collagens, ECM Glycoproteins, ECM Regulators, …[n = 6], 来自于Matrisome 基质体相关基因获取[Section: LUAD]) 用于 ssGSEA 富集分析。



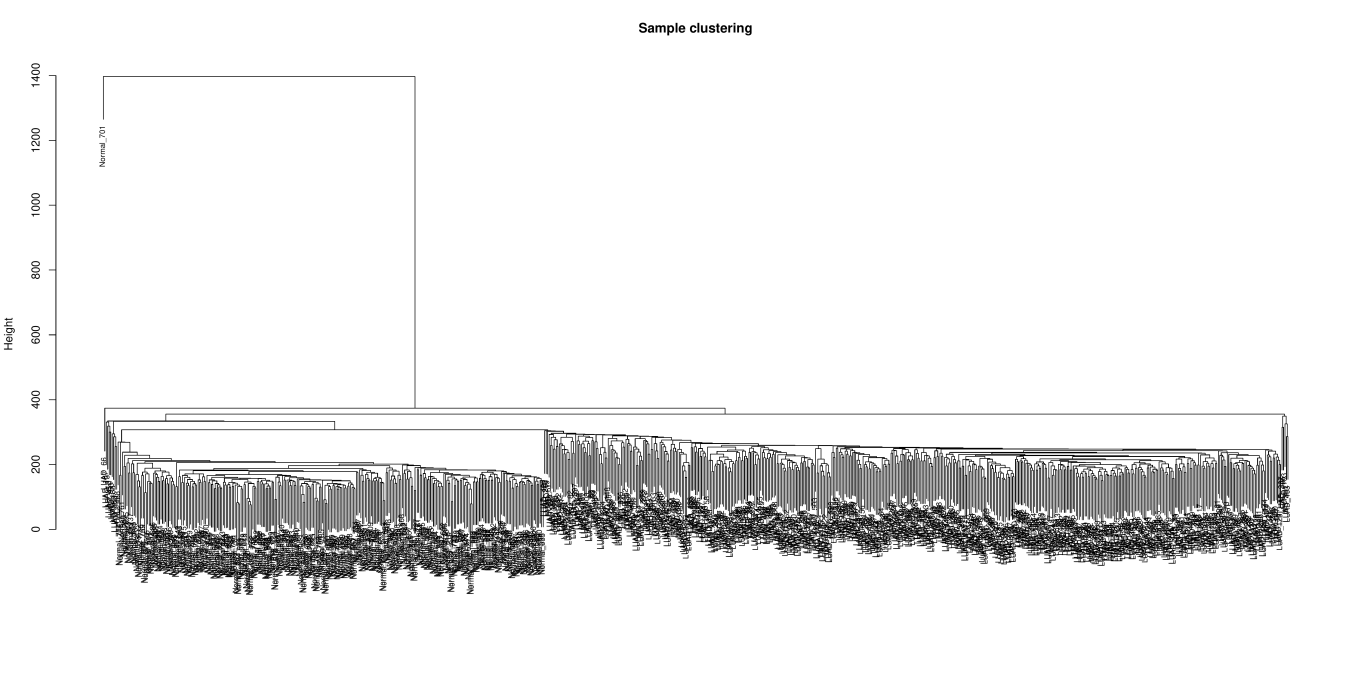
**Fig.** **2** LUAD scores LUAD Normal

**(File path: Figure+Table/3.3.0\_ssGSEA\_单样本GSEA富集分析\_(LUAD)/LUAD-scores-LUAD-Normal.pdf)**

Fig. **[2](#LUAD-scores-LUAD-Normal)** 为各分组 LUAD Normal 基质评分箱形图。

## 3.4 WGCNA 分析 (LUAD)

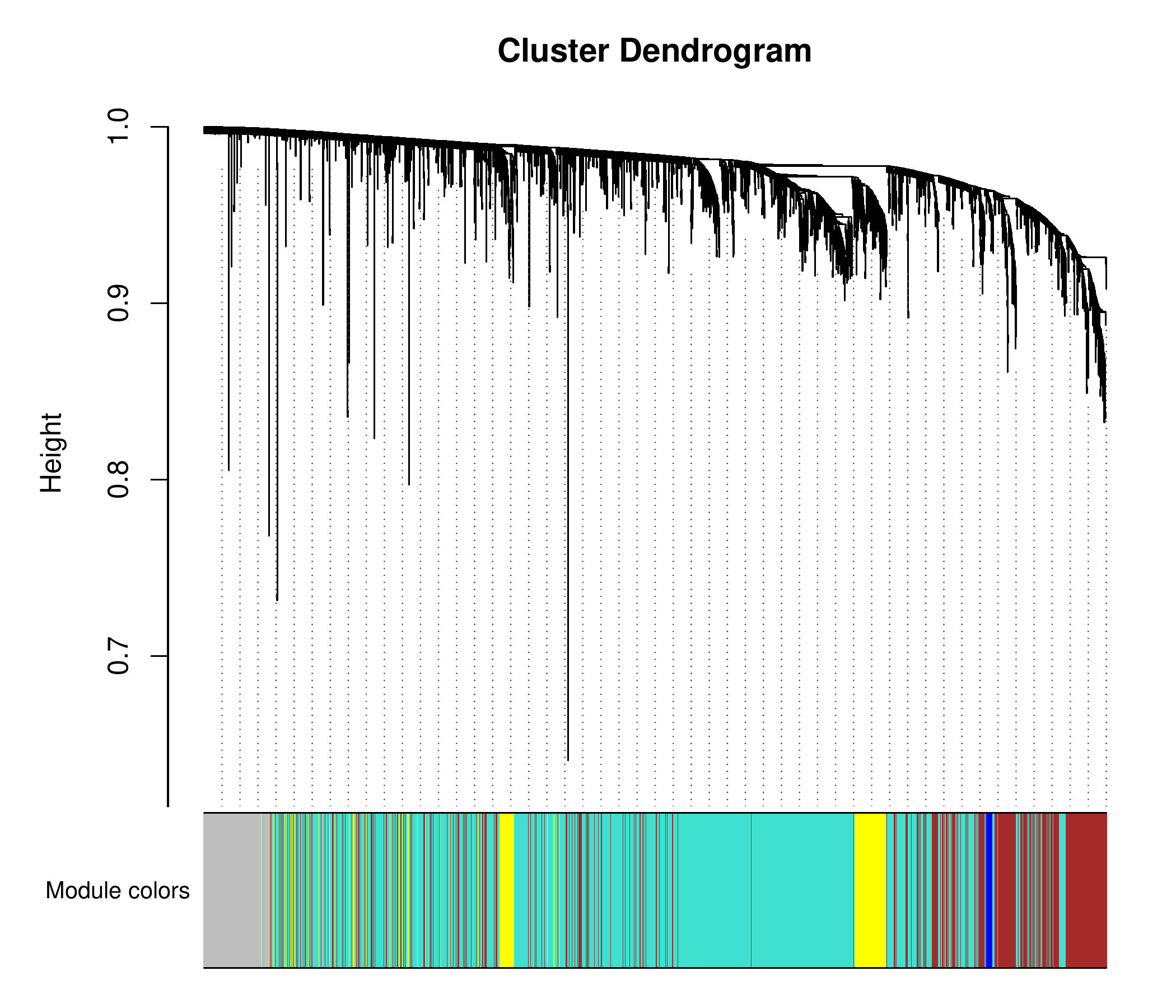
以 WGCNA::cutreeStatic (cutHeight = 600, minSize = 10) 剪切聚类树，滤掉样本 ‘GTEX.SUCS.0626.SM.5CHQE’。以 power 11 (soft thresholding powers) 创建基因共表达模块 (各模块基因数：ME0 (n=10832) , ME1 (n=3921) , ME2 (n=2593) , ME3 (n=1536) , ME4 (n=590) , ME5 (n=382) , ME6 (n=243) , ME7 (n=173) , ME8 (n=126) , ME9 (n=108) , ME10 (n=95) , ME11 (n=95) , ME12 (n=94) , ME13 (n=88) , ME14 (n=88) , ME15 (n=62) , ME16 (n=52) , ME17 (n=51) , ME18 (n=48) , ME19 (n=39) , ME20 (n=36) , ME21 (n=34) , ME22 (n=30) , ME23 (n=28) , ME24 (n=25) )。将 ‘group’ 设置为数值变量 (Normal, LUAD 依次为 1, 2) 与基因共表达模块关联分析。筛选显著关联的共表达模块的基因 (pvalue < 0.05, cor > 0)。



**Fig.** **3** LUAD sample clustering

**(File path: Figure+Table/3.4.0\_WGCNA\_分析\_(LUAD)/LUAD-sample-clustering.pdf)**

Fig. **[3](#LUAD-sample-clustering)** 为样本聚类树



**Fig.** **4** LUAD co expression module

**(File path: Figure+Table/3.4.0\_WGCNA\_分析\_(LUAD)/LUAD-co-expression-module.pdf)**

Fig. **[4](#LUAD-co-expression-module)** 为 WGCNA 创建的网络的基因共表达模块。

**Tab.** **3** LUAD correlation of module with group

| MEs | Cor | Pvalue |
| --- | --- | --- |
| ME0 | 0.9403 | 0 |
| ME1 | 0.9166 | 5.00587312366351e-320 |
| ME3 | -0.9014 | 2.927e-292 |
| ME5 | 0.8478 | 4.772e-222 |
| ME2 | -0.8301 | 1.216e-204 |
| ... | ... | ... |

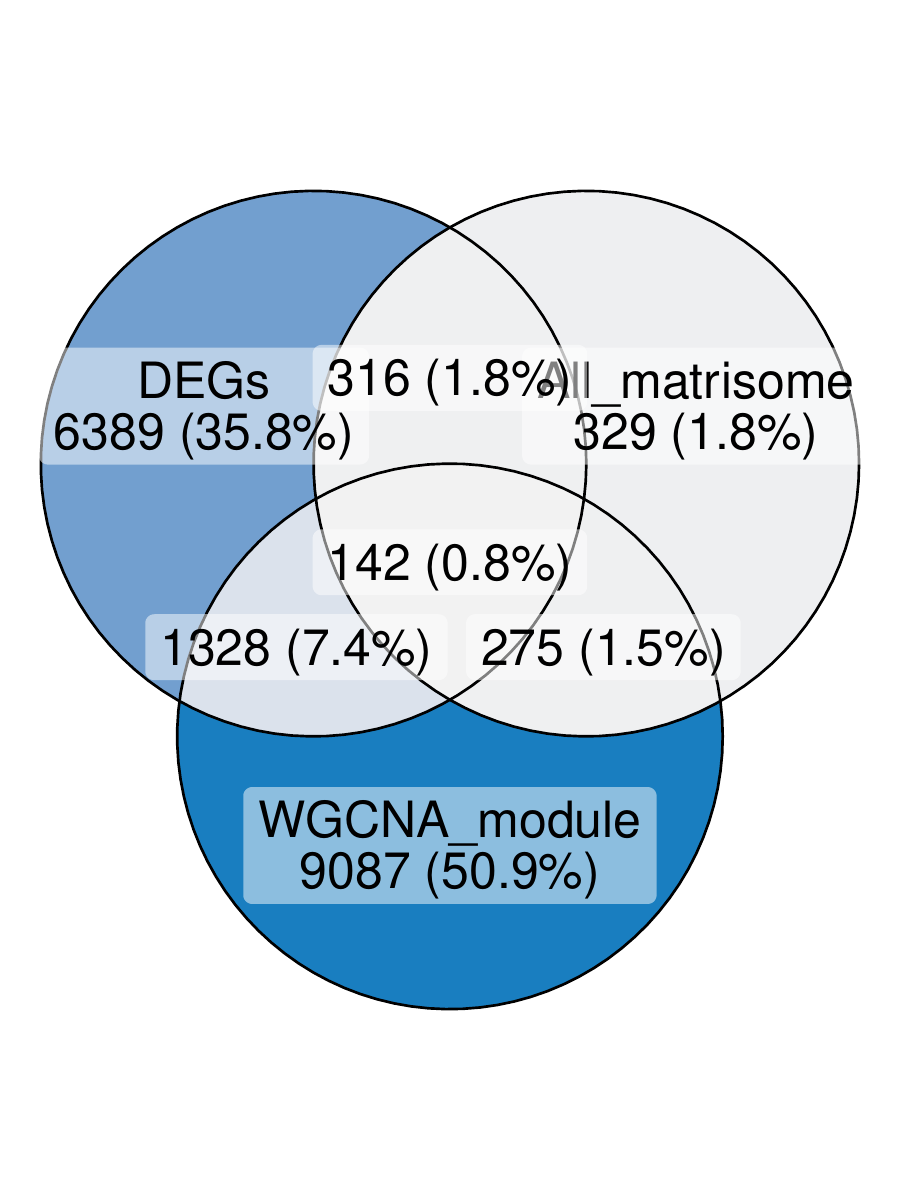
**(File path: Figure+Table/3.4.0\_WGCNA\_分析\_(LUAD)/LUAD-correlation-of-module-with-group.csv)**

Tab. **[3](#LUAD-correlation-of-module-with-group)** 为共表达模块与 group 的关联性。

## 3.5 汇总: DEGs + All\_matrisome + WGCNA\_module (LUAD)

数据集为：

* **基因集** (LUAD - Normal, 来自于Limma 差异分析[Section: LUAD])
* **基因集** (Collagens, ECM Glycoproteins, ECM Regulators, …[n = 6], 来自于Matrisome 基质体相关基因获取[Section: LUAD])
* **基因集** (ME0, 来自于WGCNA 与 group 显著关联的共表达模块的基因[Section: LUAD])



**Fig.** **5** LUAD Intersection of DEGs with All matrisome with WGCNA module

**(File path: Figure+Table/3.5.0\_汇总:\_DEGs\_+\_All\_matrisome\_+\_WGCNA\_module\_(LUAD)/LUAD-Intersection-of-DEGs-with-All-matrisome-with-WGCNA-module.pdf)**

* All\_intersection: MMP13, IGFBPL1, ZPLD1, WNT10A, PLAU, S100A2, ADAM28, IL12B, IL23A, MUC5B, CST2, KAZALD1, TSKU, MEGF10, MMP9, OIT3, CXCL14, MUC13, S100A6, MUC3A, MMP10, MEP1A, WNT4, CCL8, SERPINB5, MUC4, BTC, MUC16, SCUBE3, EGF, MMP16, CRISPLD1, CCL20, INHBB, FGL1, MMP3, CCL28, LGI2, PODNL1, SPOCK3, S100P, COL9A2, GDNF, AMBP, VWDE, NELL1, SCUBE2, GDF15, IGFBP2, CSF2, SMOC1, IL11, ELFN2, ITIH2, PLXNB3, SERPIND1, FGB, COMP, MUC5AC, FGA, SERPINB4, FGFBP1, MUC6, PRSS3, FGG, CTSE, SERPINB3, COL6A2, EMILIN1, CXCL3,…

**(See: Figure+Table/3.5.0\_汇总:\_DEGs\_+\_All\_matrisome\_+\_WGCNA\_module\_(LUAD)/LUAD-Intersection-of-DEGs-with-All-matrisome-with-WGCNA-module-content)**

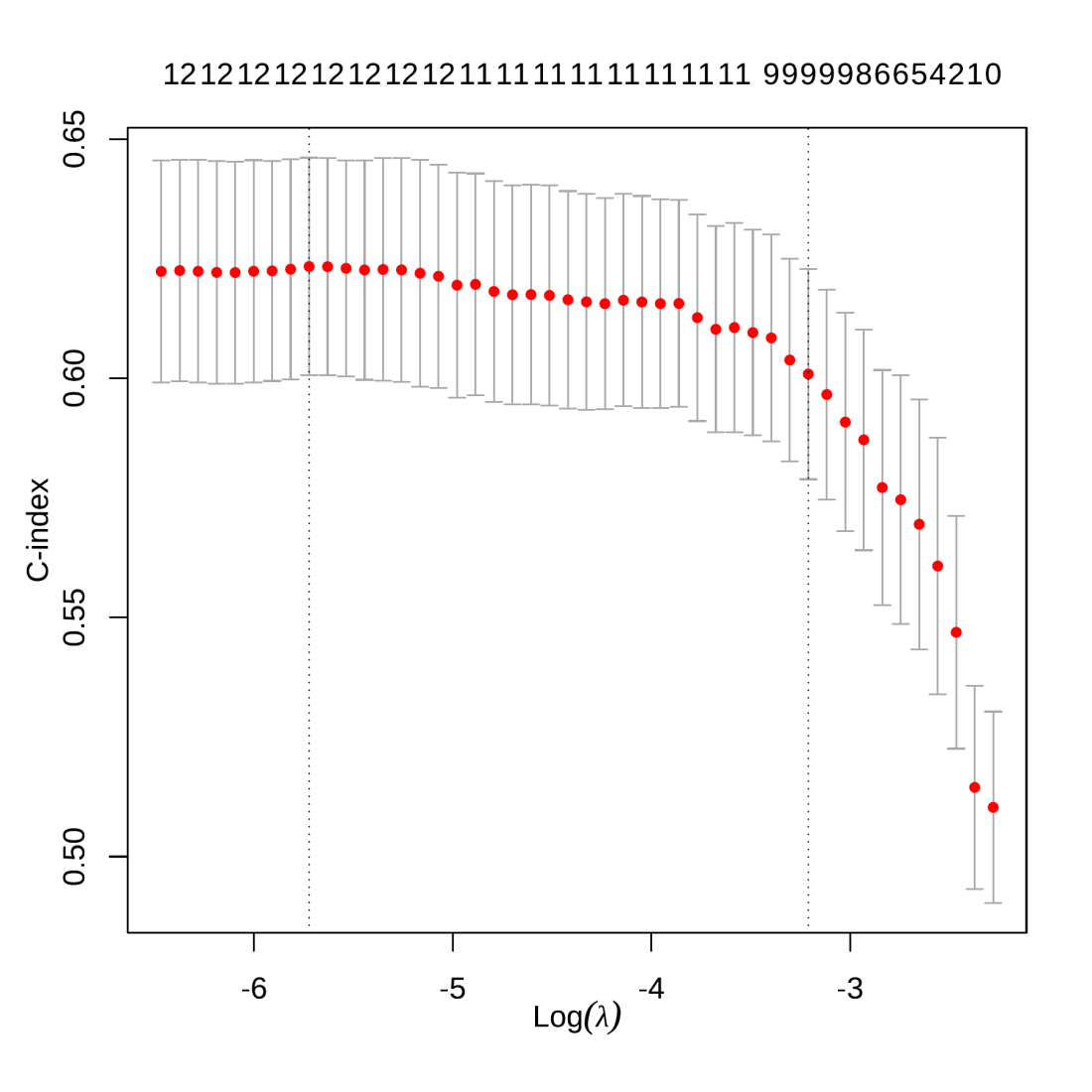
Fig. **[5](#LUAD-Intersection-of-DEGs-with-All-matrisome-with-WGCNA-module)** 为 DEGs, All\_matrisome, WGCNA\_module 各自交集。

## 3.6 TCGA 数据获取 (LUAD)

获取 TCGA-LUAD (RNA, clinical) 数据。

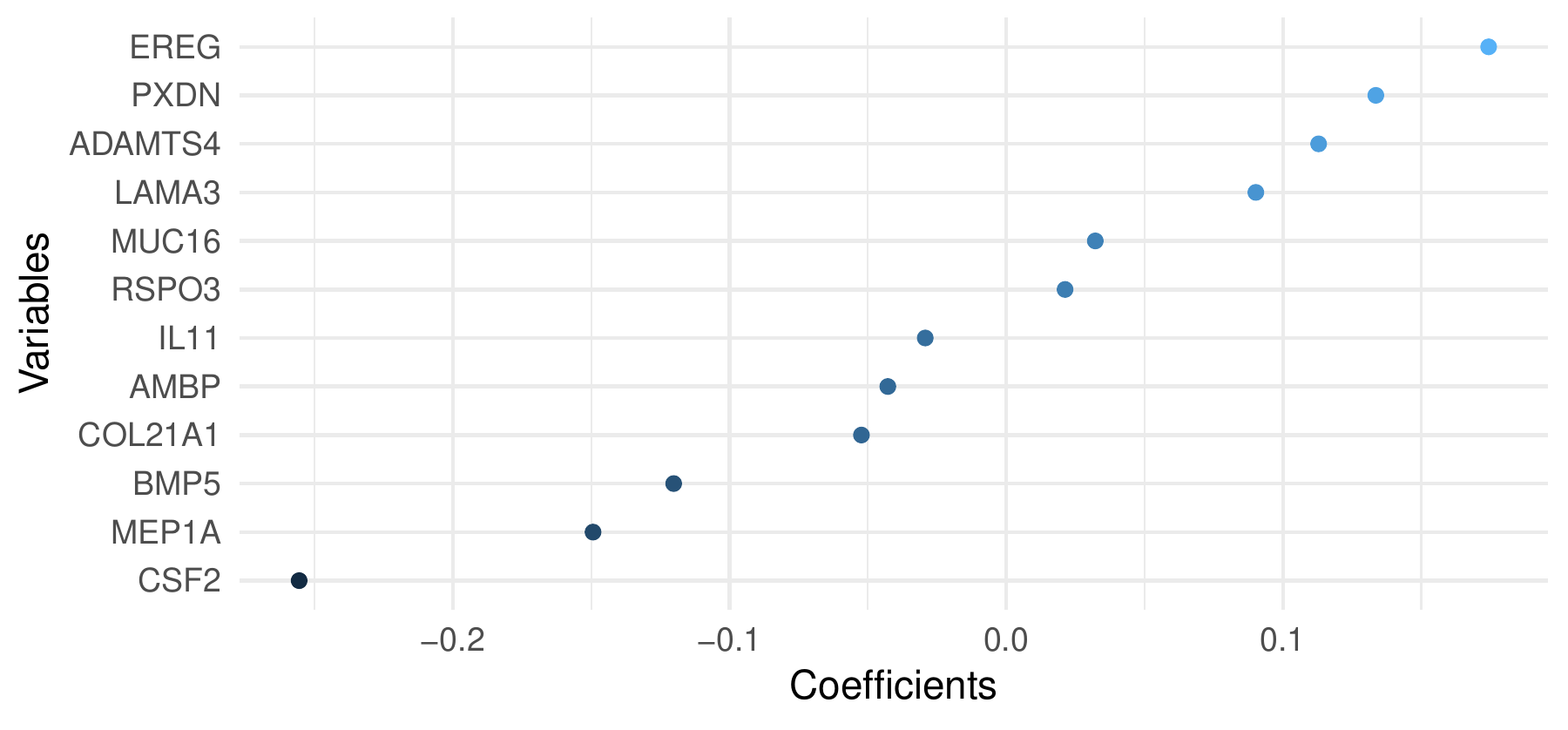
## 3.7 COX 回归 (TCGA\_LUAD)

筛选 isTumor 为 “tumor”，最终得到 530 例数据。将**基因集** (MMP13, IGFBPL1, ZPLD1, …[n = 142], 来自于Venn 交集[Section: LUAD]) 用于模型建立。共 138 个基因在数据集 TCGA-LUAD 中找到 (根据基因名 Symbol 以及 ALIAS (org.Hs.eg.db, 3.20.0) 匹配)。所有数据生存状态 (去除生存状态未知的数据)，(Alive (n=340) , Dead (n=189) )。执行单因素 COX 回归，筛选 P 值 < 0.05，且满足 PH 假定 (P > 0.05)，共筛选到 30 个基因。在单因素回归得到的基因 (P < 0.01) 的基础上，使用 glmnet::cv.glmnet 作 10 倍交叉验证 (评估方式为 C-index)，筛选 lambda 值。lambda.min, lambda.1se 值分别为 0.003, 0.04 (R 随机种子为 987456)。对应的特征数 (基因数) 分别为 12, 9。



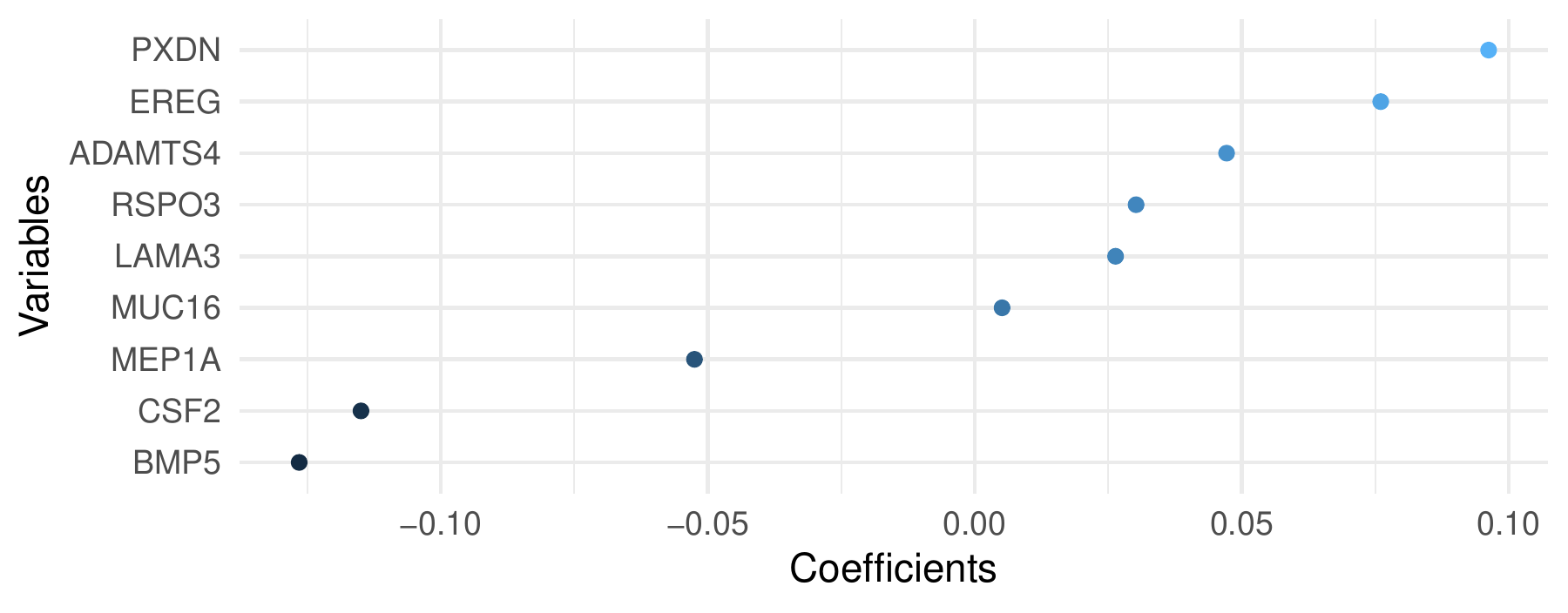
**Fig.** **6** TCGA LUAD lasso COX model

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_LUAD)/TCGA-LUAD-lasso-COX-model.pdf)**



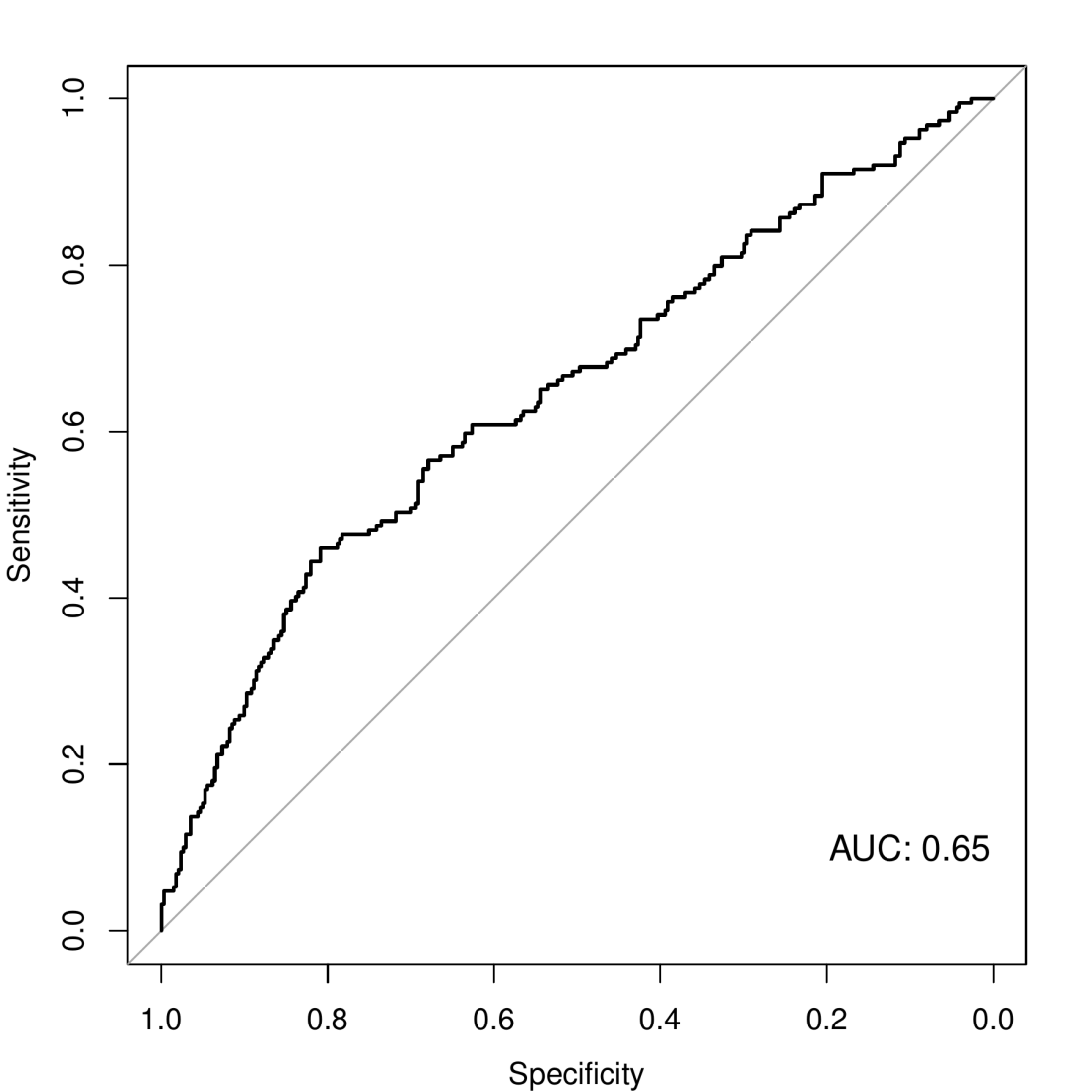
**Fig.** **7** TCGA LUAD lasso COX coeffients lambda min

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_LUAD)/TCGA-LUAD-lasso-COX-coeffients-lambda-min.pdf)**



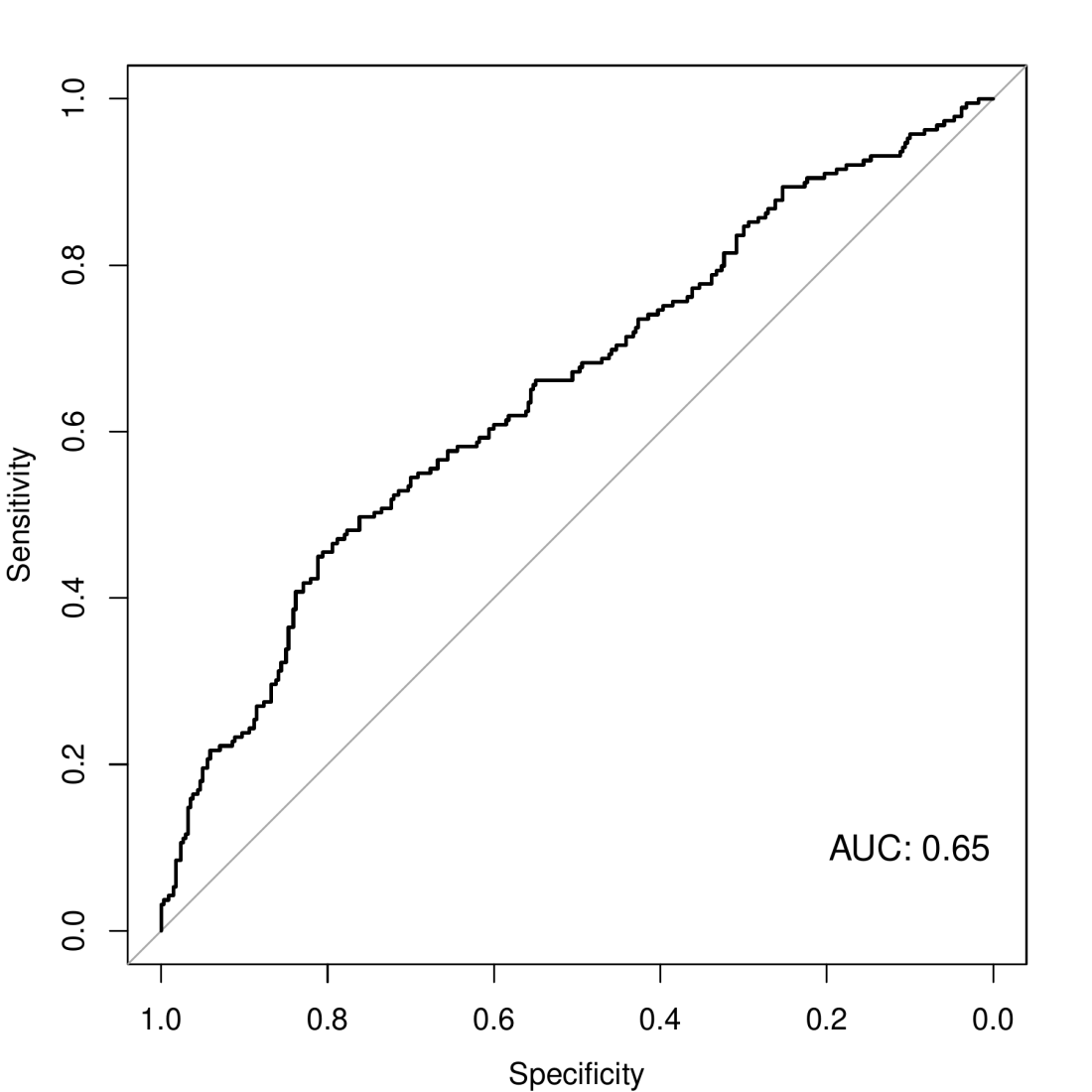
**Fig.** **8** TCGA LUAD lasso COX coeffients lambda 1se

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_LUAD)/TCGA-LUAD-lasso-COX-coeffients-lambda-1se.pdf)**



**Fig.** **9** TCGA LUAD lasso COX ROC lambda min

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_LUAD)/TCGA-LUAD-lasso-COX-ROC-lambda-min.pdf)**



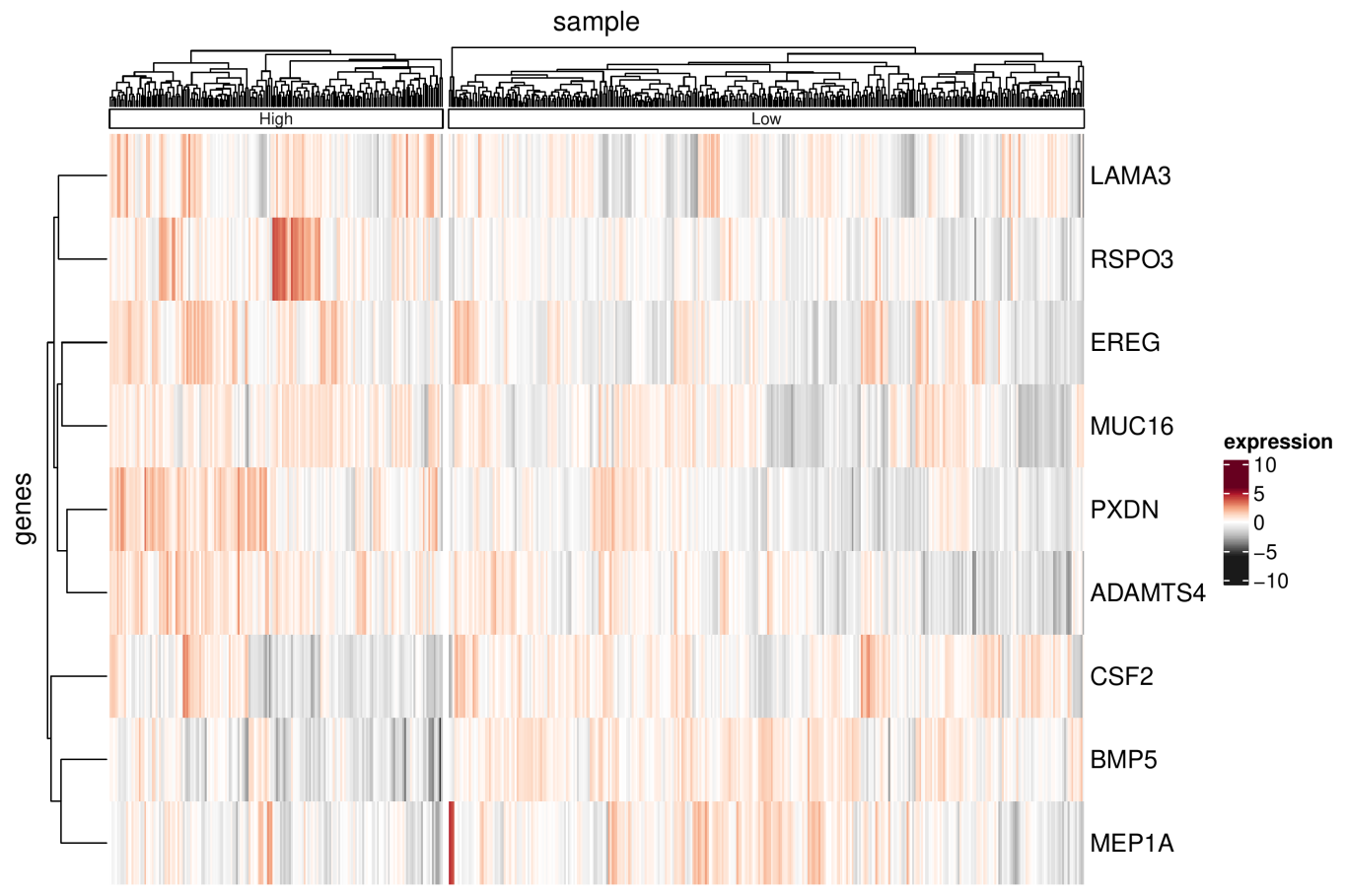
**Fig.** **10** TCGA LUAD lasso COX ROC lambda 1se

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_LUAD)/TCGA-LUAD-lasso-COX-ROC-lambda-1se.pdf)**

## 3.8 Survival 生存分析 (TCGA\_LUAD)

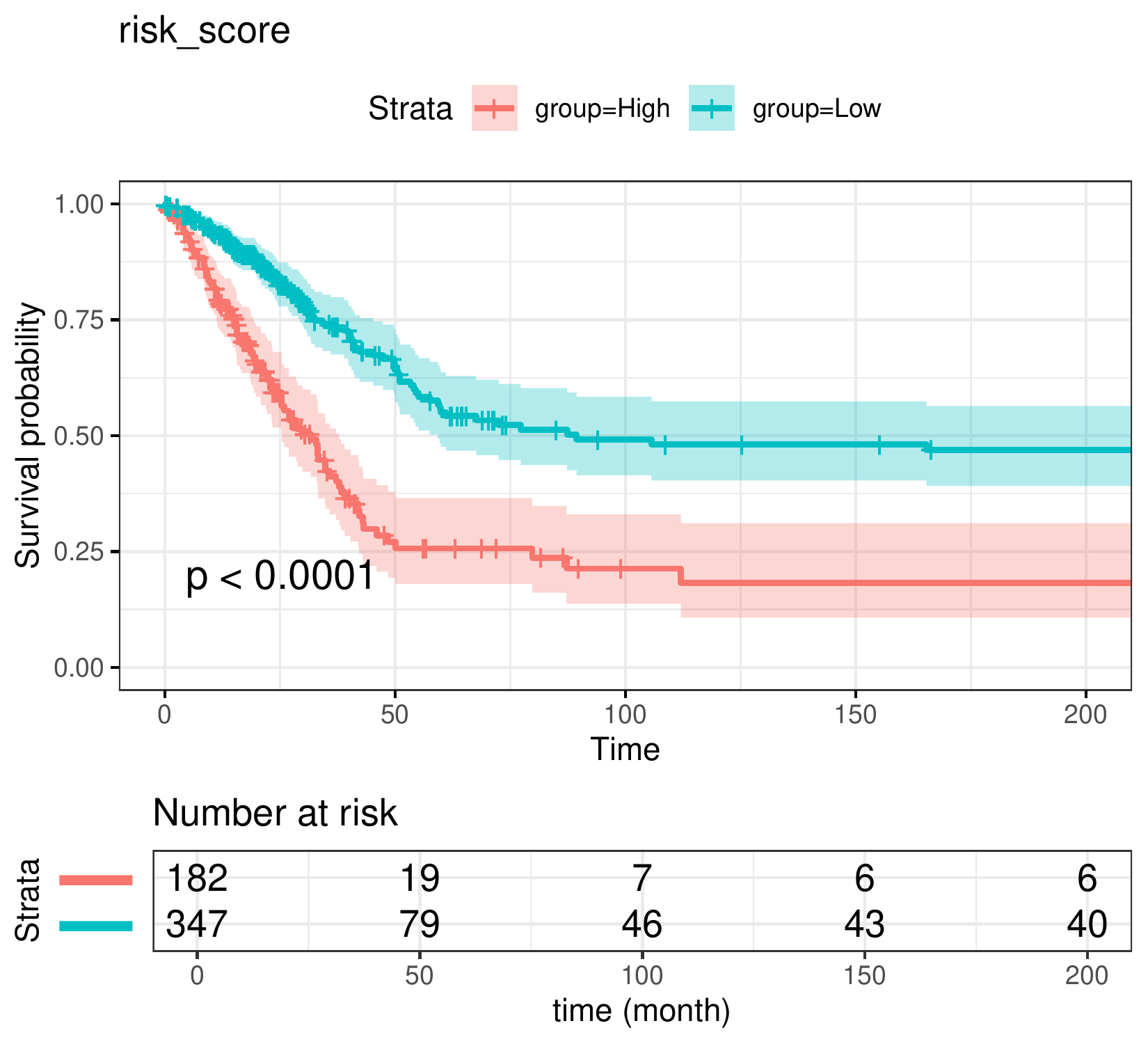
选择 lambda.1se 时得到的特征集，包含 9 个基因， 分别为: ADAMTS4, BMP5, CSF2, EREG, LAMA3, MEP1A, MUC16, PXDN, RSPO3。以回归系数构建风险评分模型。

按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 0.0625877926451681) (High (n=182) , Low (n=347) )， 随后进行生存分析。risk\_score 生存分析 P value < 0.05, 且满足 ph 假设检验 (P > 0.05)。



**Fig.** **11** TCGA LUAD risk score related genes heatmap

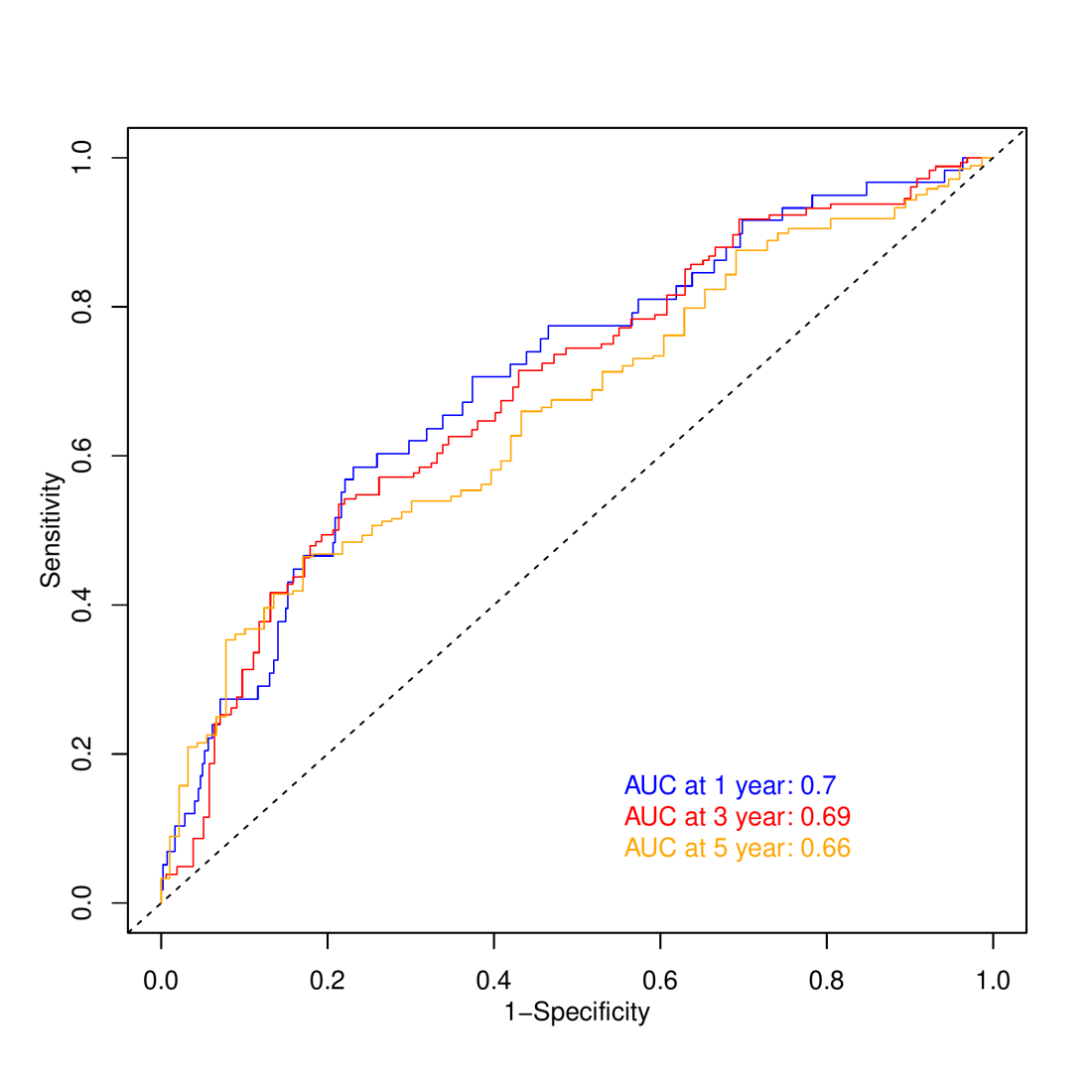
**(File path: Figure+Table/3.8.0\_Survival\_生存分析\_(TCGA\_LUAD)/TCGA-LUAD-risk-score-related-genes-heatmap.pdf)**



**Fig.** **12** TCGA LUAD survival curve of risk score

**(File path: Figure+Table/3.8.0\_Survival\_生存分析\_(TCGA\_LUAD)/TCGA-LUAD-survival-curve-of-risk-score.pdf)**

Fig. **[12](#TCGA-LUAD-survival-curve-of-risk-score)** 为 risk\_score 生存曲线。



**Fig.** **13** TCGA LUAD time ROC

**(File path: Figure+Table/3.8.0\_Survival\_生存分析\_(TCGA\_LUAD)/TCGA-LUAD-time-ROC.pdf)**

Fig. **[13](#TCGA-LUAD-time-ROC)** 为 risk\_score 1, 3, 5 年生存分析 ROC 曲线。

**Tab.** **4** TCGA LUAD Survival PValue

| Name | Pvalue | Ph | Group low survival |
| --- | --- | --- | --- |
| Risk score | 1.815e-12 | 0.2509 | High |

**(File path: Figure+Table/3.8.0\_Survival\_生存分析\_(TCGA\_LUAD)/TCGA-LUAD-Survival-PValue.csv)**

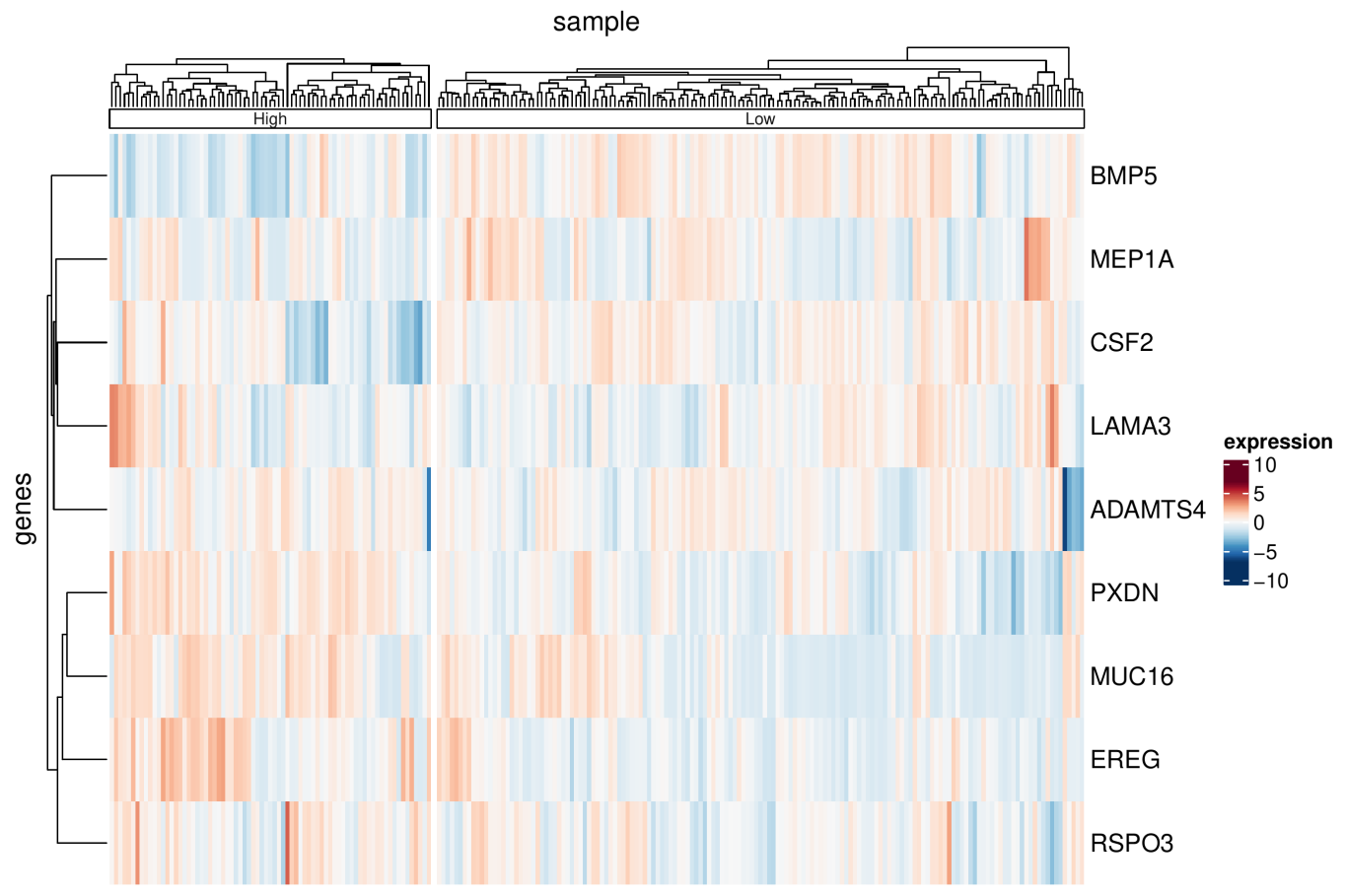
## 3.9 外部验证

### 3.9.1 GEO 数据获取 (LUAD\_GSE31210)

以 GEOquery 获取 GSE31210 的数据信息。

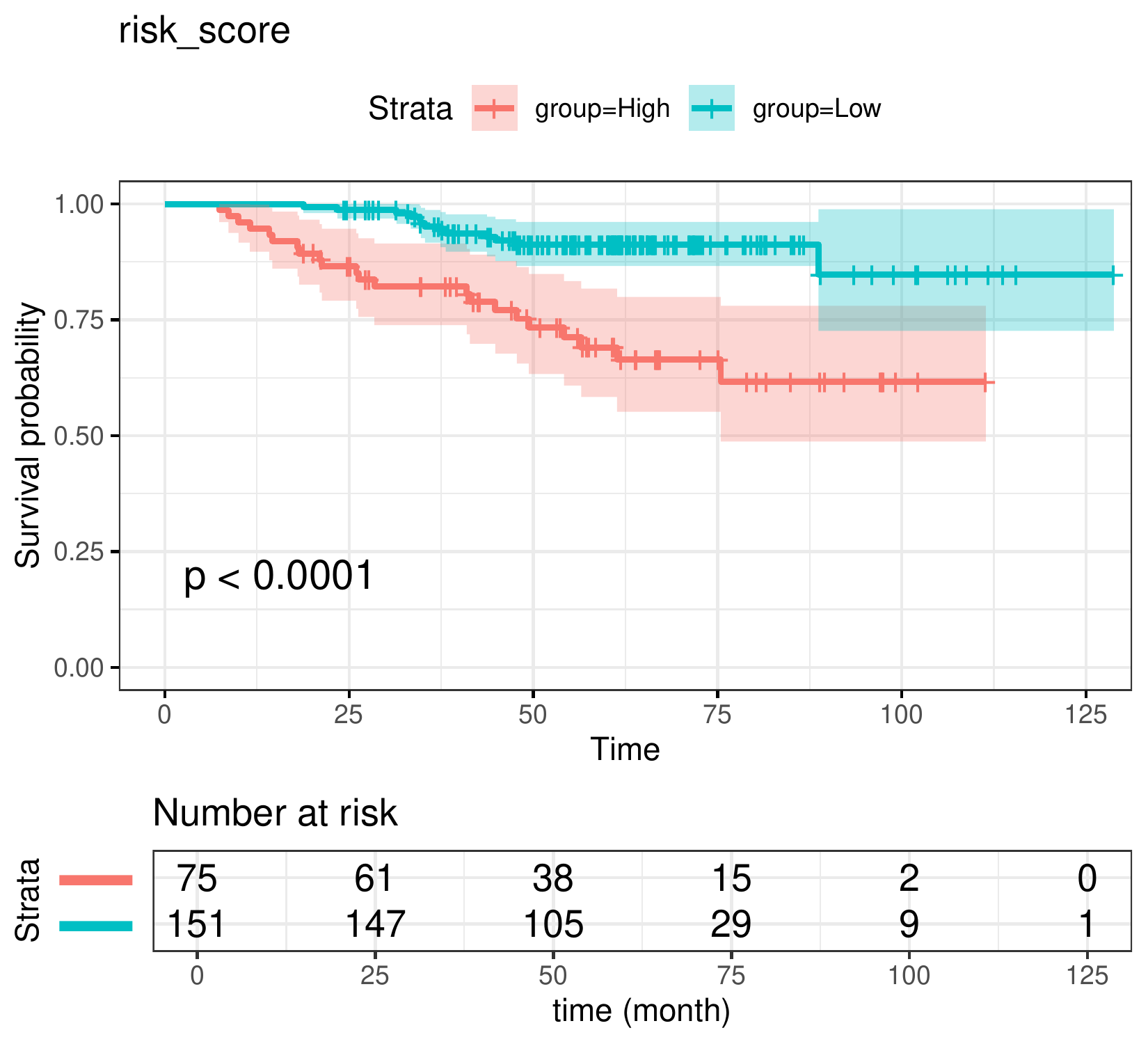
### 3.9.2 Survival 生存分析 (LUAD\_GSE31210)

按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 0.0968699367720758) (High (n=75) , Low (n=151) )， 随后进行生存分析。risk\_score 生存分析 P value < 0.05, 且满足 ph 假设检验 (P > 0.05)。



**Fig.** **14** Risk score related genes heatmap

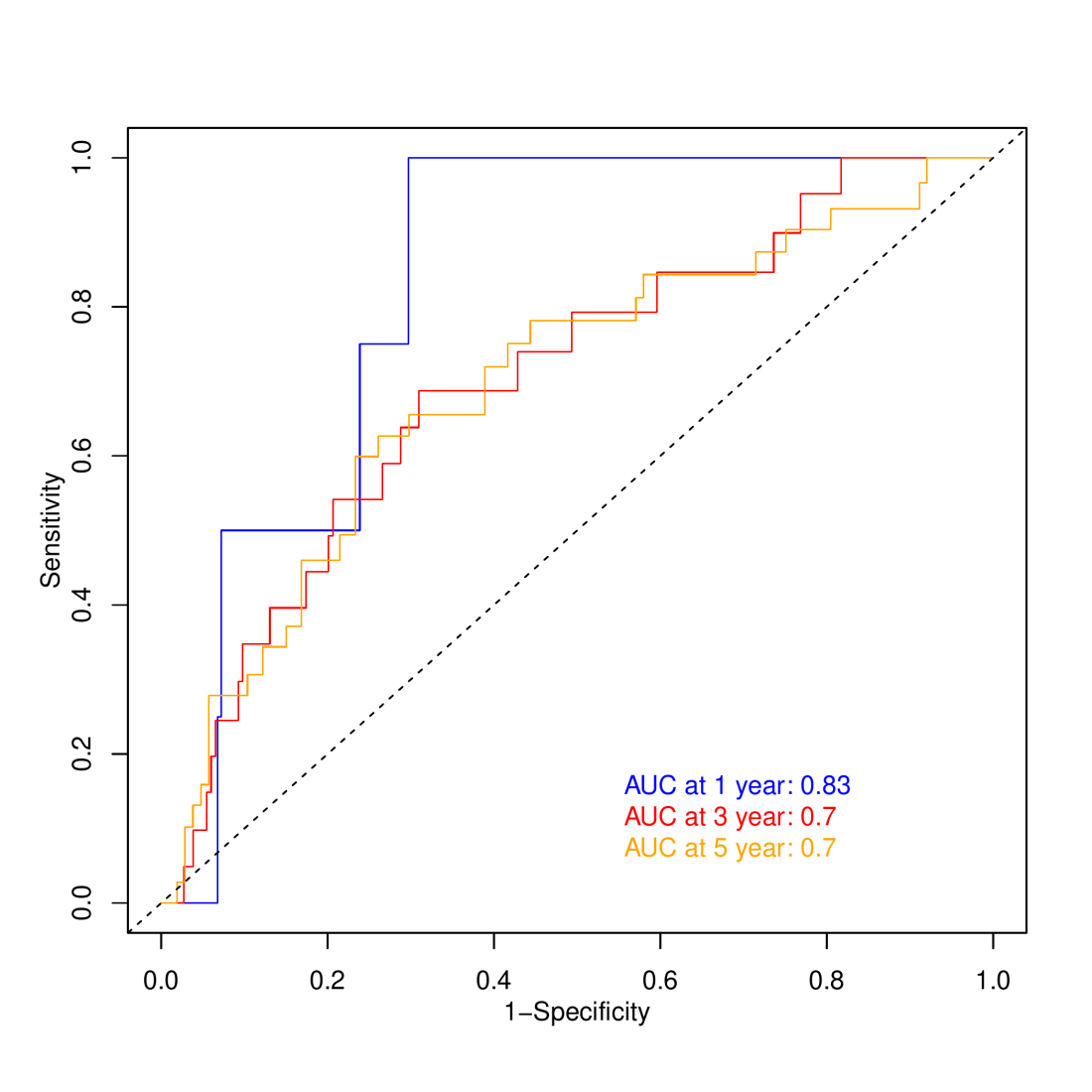
**(File path: Figure+Table/3.9.2\_Survival\_生存分析\_(LUAD\_GSE31210)/Risk-score-related-genes-heatmap.pdf)**



**Fig.** **15** LUAD GSE31210 survival curve of risk score

**(File path: Figure+Table/3.9.2\_Survival\_生存分析\_(LUAD\_GSE31210)/LUAD-GSE31210-survival-curve-of-risk-score.pdf)**

Fig. **[15](#LUAD-GSE31210-survival-curve-of-risk-score)** 为 risk\_score 生存曲线。



**Fig.** **16** LUAD GSE31210 time ROC

**(File path: Figure+Table/3.9.2\_Survival\_生存分析\_(LUAD\_GSE31210)/LUAD-GSE31210-time-ROC.pdf)**

Fig. **[16](#LUAD-GSE31210-time-ROC)** 为 risk\_score 1, 3, 5 年生存分析 ROC 曲线。

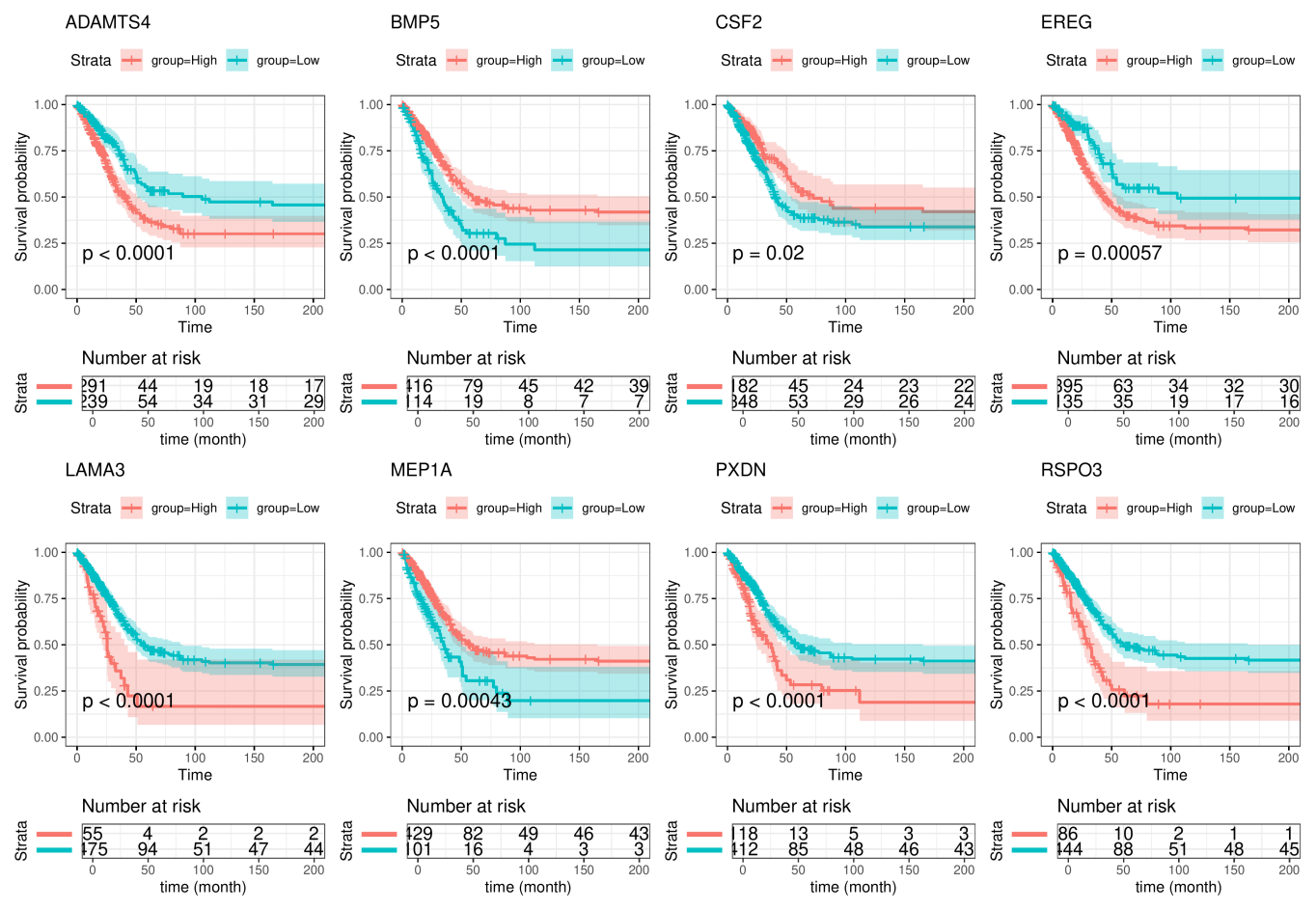
**Tab.** **5** LUAD GSE31210 Significant Survival PValue

| Name | Pvalue | Ph | Group low survival |
| --- | --- | --- | --- |
| Risk score | 1.558e-05 | 0.6373 | High |

**(File path: Figure+Table/3.9.2\_Survival\_生存分析\_(LUAD\_GSE31210)/LUAD-GSE31210-Significant-Survival-PValue.csv)**

## 3.10 Survival 生存分析 (INDIVIDUALS)

(对预后模型中的各个基因做了生存分析) 筛选 isTumor 为 “tumor”，最终得到 530 例数据。对**基因集** (ADAMTS4, BMP5, CSF2, …[n = 9], 来自于Survival 生存分析[Section: TCGA\_LUAD]) 进行Survival 生存分析。按 survminer::surv\_cutpoint 计算的 cutoff，将样本分为 Low 和 High 风险组。生存数据为TCGA-LUAD，使用标准化过的基因表达数据。根据元数据信息 (即临床数据) ，去除了生存状态未知的样例。根据 P value < 0.05, 共筛到 8 个特征。 分别为 ADAMTS4, BMP5, CSF2, EREG, LAMA3, MEP1A, PXDN, RSPO3。



**Fig.** **17** INDIVIDUALS all significant genes survival curves

**(File path: Figure+Table/3.10.0\_Survival\_生存分析\_(INDIVIDUALS)/INDIVIDUALS-all-significant-genes-survival-curves.pdf)**

Fig. **[17](#INDIVIDUALS-all-significant-genes-survival-curves)** 为所有基因的生存曲线图 (显著的)。

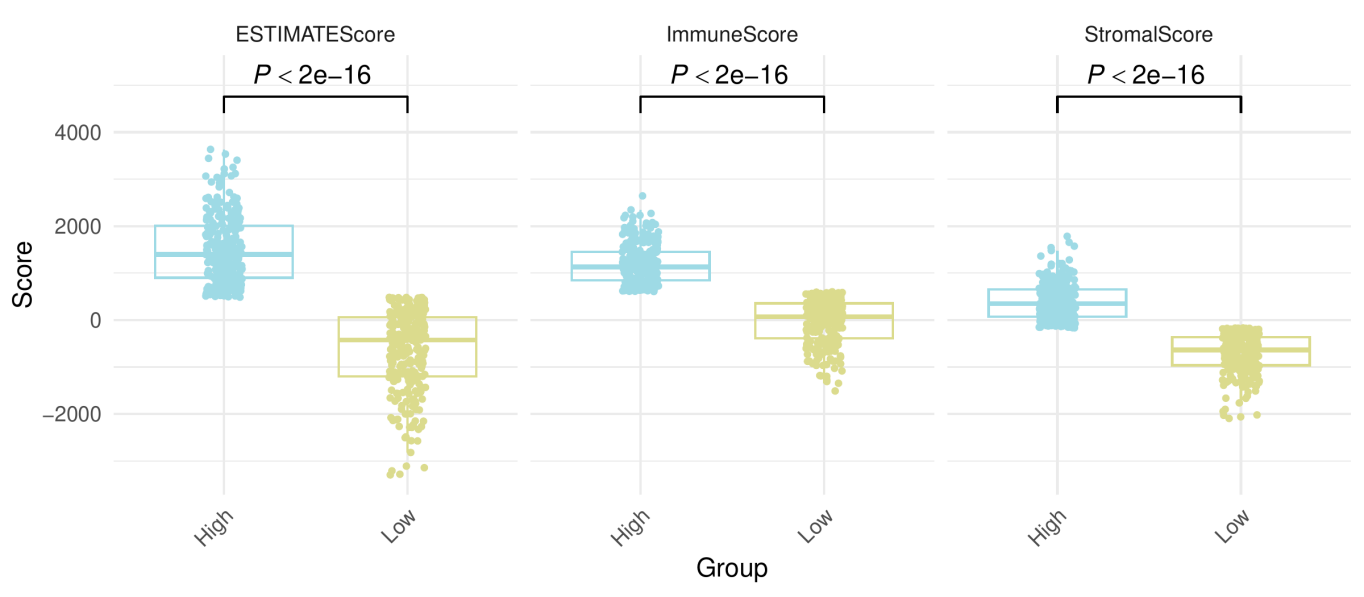
**Tab.** **6** INDIVIDUALS Significant Survival PValue

| Name | Pvalue | Ph | Group low survival |
| --- | --- | --- | --- |
| ADAMTS4 | 4.752e-05 | 0.3269 | High |
| BMP5 | 5.719e-05 | 0.7797 | Low |
| CSF2 | 0.02017 | 0.06617 | Low |
| EREG | 0.0005732 | 0.7556 | High |
| LAMA3 | 2.323e-05 | 0.6952 | High |
| ... | ... | ... | ... |

**(File path: Figure+Table/3.10.0\_Survival\_生存分析\_(INDIVIDUALS)/INDIVIDUALS-Significant-Survival-PValue.csv)**

## 3.11 estimate 免疫评分 (TCGA\_LUAD\_REGROUP)

使用 Survival 分析 (TCGA\_LUAD) 时定义的分组数据 (根据 risk\_score 的表达，分为 High (n=182) , Low (n=347) )。以 estimate 进行免疫评分分析。



**Fig.** **18** TCGA LUAD REGROUP Group boxplot

**(File path: Figure+Table/3.11.0\_estimate\_免疫评分\_(TCGA\_LUAD\_REGROUP)/TCGA-LUAD-REGROUP-Group-boxplot.pdf)**

Fig. **[18](#TCGA-LUAD-REGROUP-Group-boxplot)** 为 Group 评分差异箱形图。

**Tab.** **7** TCGA LUAD REGROUP immune Scores

| NAME | Sample | Score | Group |
| --- | --- | --- | --- |
| StromalScore | TCGA-44-2665-01A | 1205 | High |
| StromalScore | TCGA-97-7937-01A | -529.7 | Low |
| StromalScore | TCGA-44-2665-01B | 936.9 | High |
| StromalScore | TCGA-64-5774-01A | -872.8 | Low |
| StromalScore | TCGA-78-7220-01A | -1046 | Low |
| ... | ... | ... | ... |

**(File path: Figure+Table/3.11.0\_estimate\_免疫评分\_(TCGA\_LUAD\_REGROUP)/TCGA-LUAD-REGROUP-immune-Scores.csv)**

## 3.12 GEO 数据获取 (SC\_LUAD)

以 GEOquery 获取 GSE189357 的数据信息。

## 3.13 Seurat 集成单细胞数据分析 (LUAD)

读取 TD1, TD2, TD3, TD4, TD5, TD6, TD7, TD8, TD9 样本的数据集。前期质量控制，一个细胞至少应有 500 个基因，并且基因数量小于 7500。线粒体基因的比例小于 10%。过滤后，所有样本共包含 107378 个细胞用于后续分析。数据归一化，PCA 聚类 (Seurat 标准工作流，见方法章节) 后，绘制 PC standard deviations 图。去除批次效应后 (详见方法章节) ，在 1-15 PC 维度，1.2 分辨率下，对细胞群 UMAP 聚类。计算所有细胞群的 Marker。使用特异性 Marker，以 SCSA 对细胞群注释。

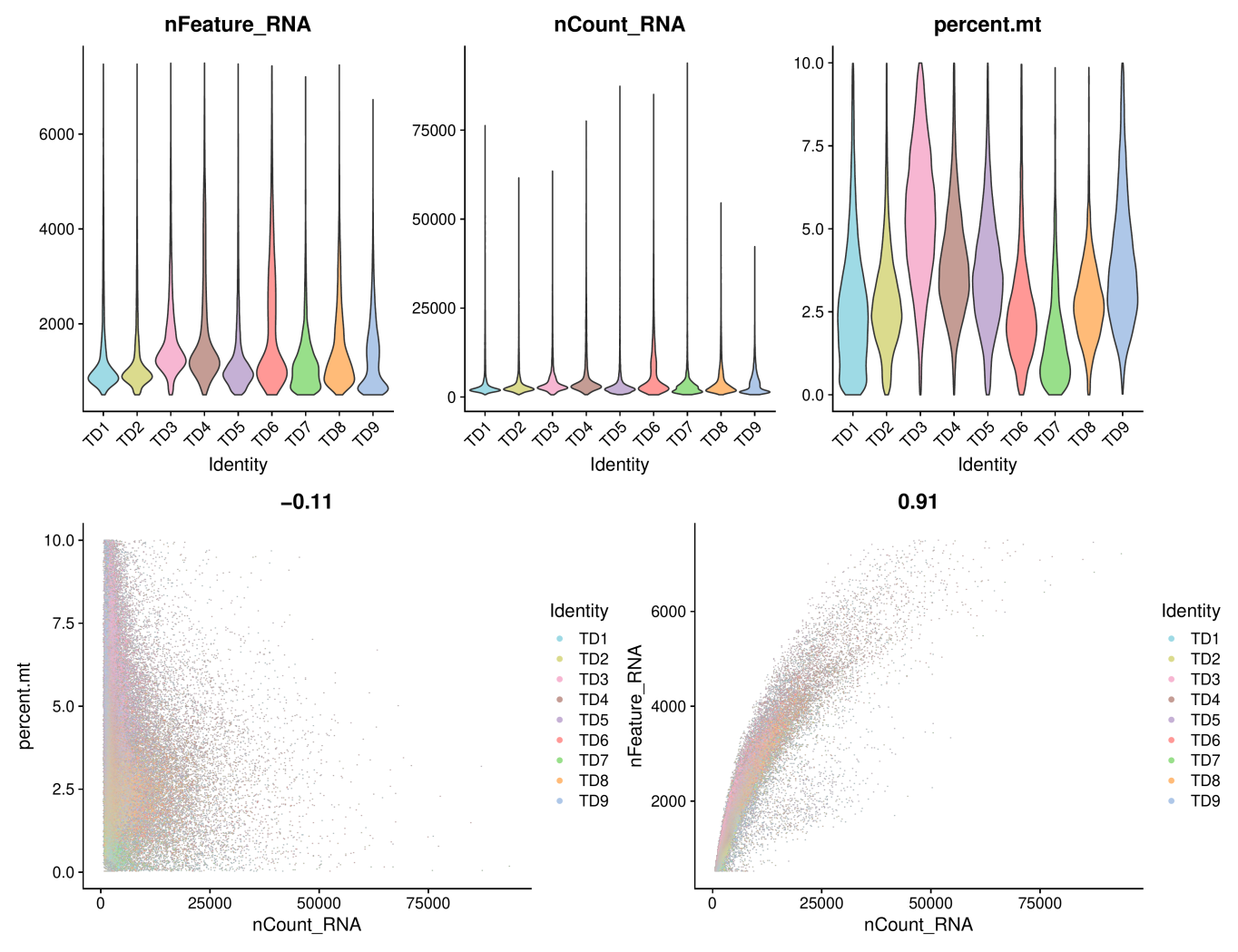
(Marker 来自于原研究文献 PMID: GSE189357)



**Fig.** **19** Pre Quality control

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/Pre-Quality-control.pdf)**

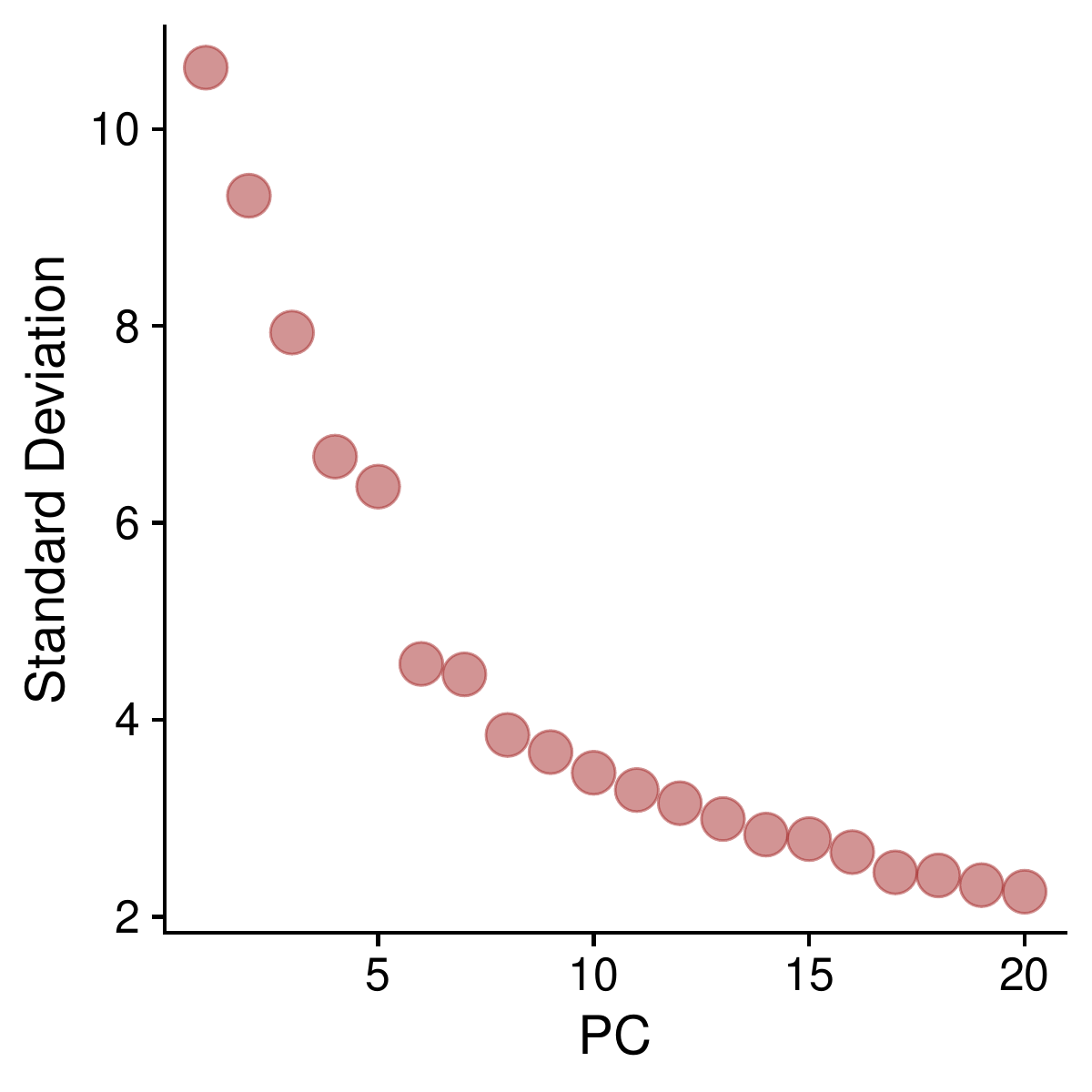
Fig. **[19](#Pre-Quality-control)** 为 QC (质量控制) 图 (数据过滤前) 。



**Fig.** **20** LUAD After Quality control

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-After-Quality-control.pdf)**

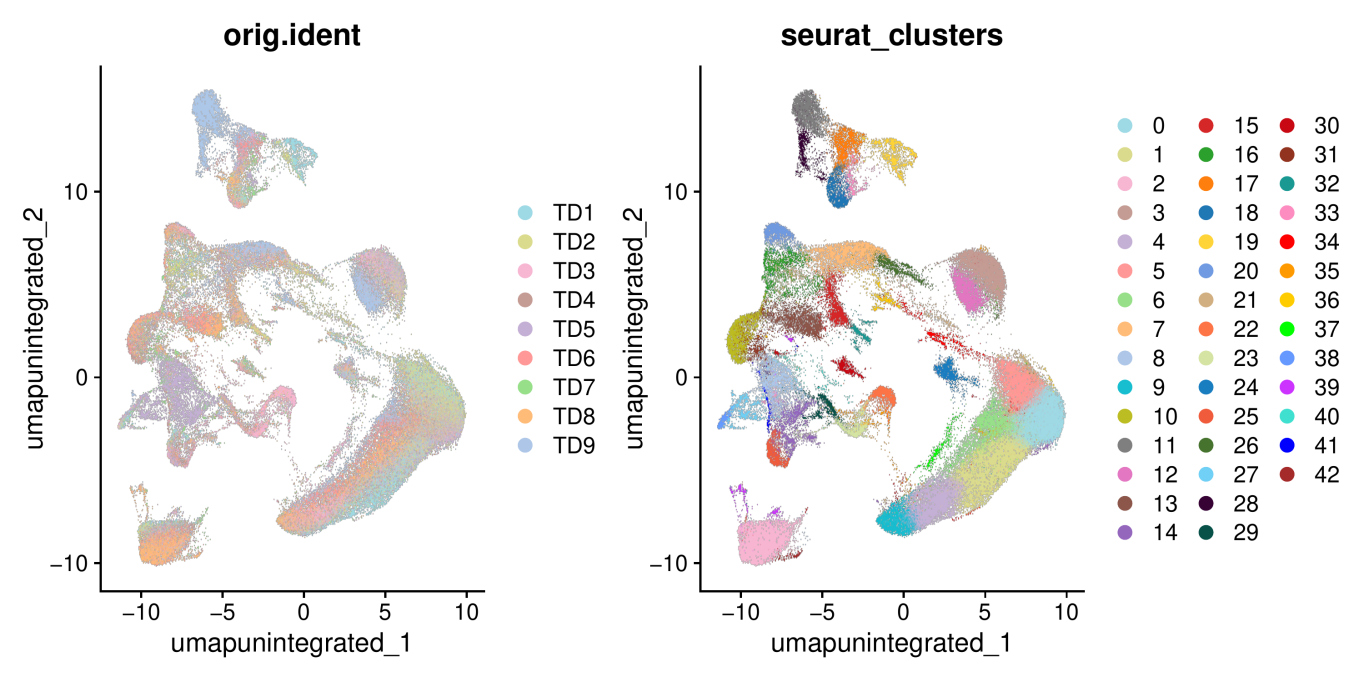
Fig. **[20](#LUAD-After-Quality-control)** 为数据过滤后的 QC 图。



**Fig.** **21** LUAD Standard deviations of PCs

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-Standard-deviations-of-PCs.pdf)**

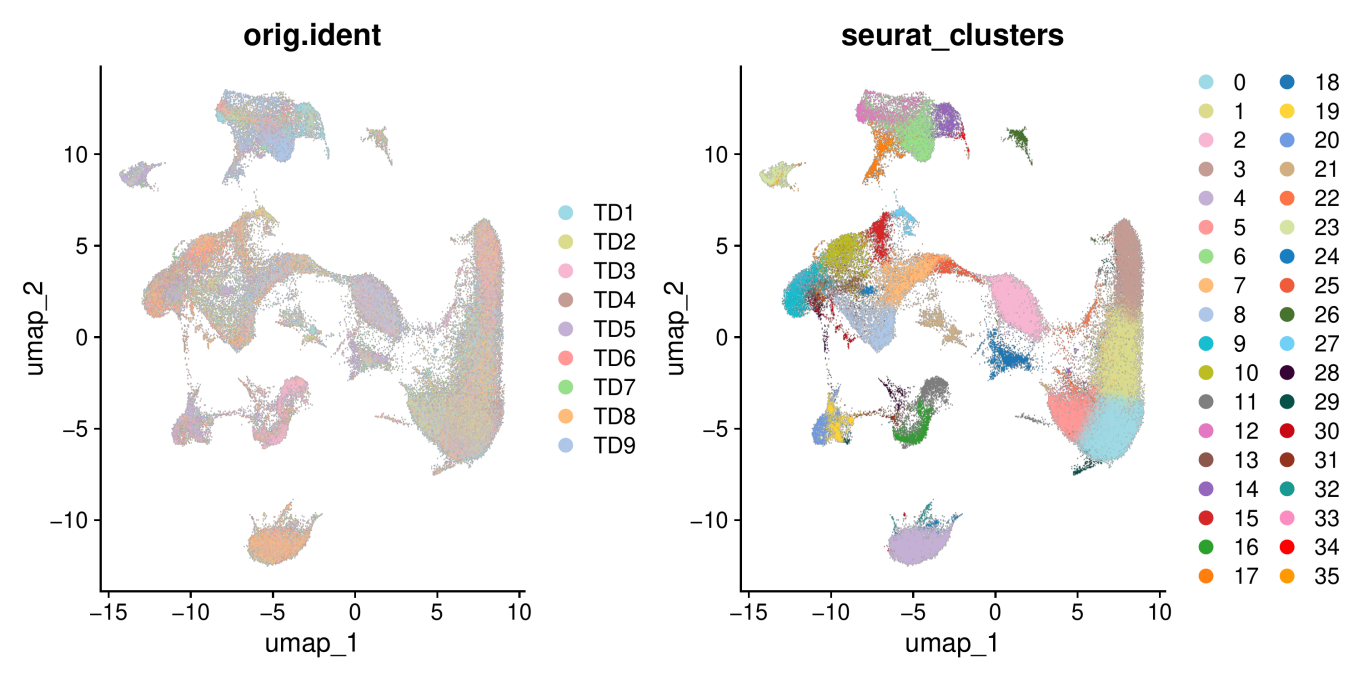
Fig. **[21](#LUAD-Standard-deviations-of-PCs)** 为主成分 (PC) 的 Standard deviations。



**Fig.** **22** LUAD UMAP Unintegrated

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-UMAP-Unintegrated.pdf)**

Fig. **[22](#LUAD-UMAP-Unintegrated)** 为去除批次效应之前的 UMAP 聚类图。



**Fig.** **23** LUAD UMAP Integrated

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-UMAP-Integrated.pdf)**

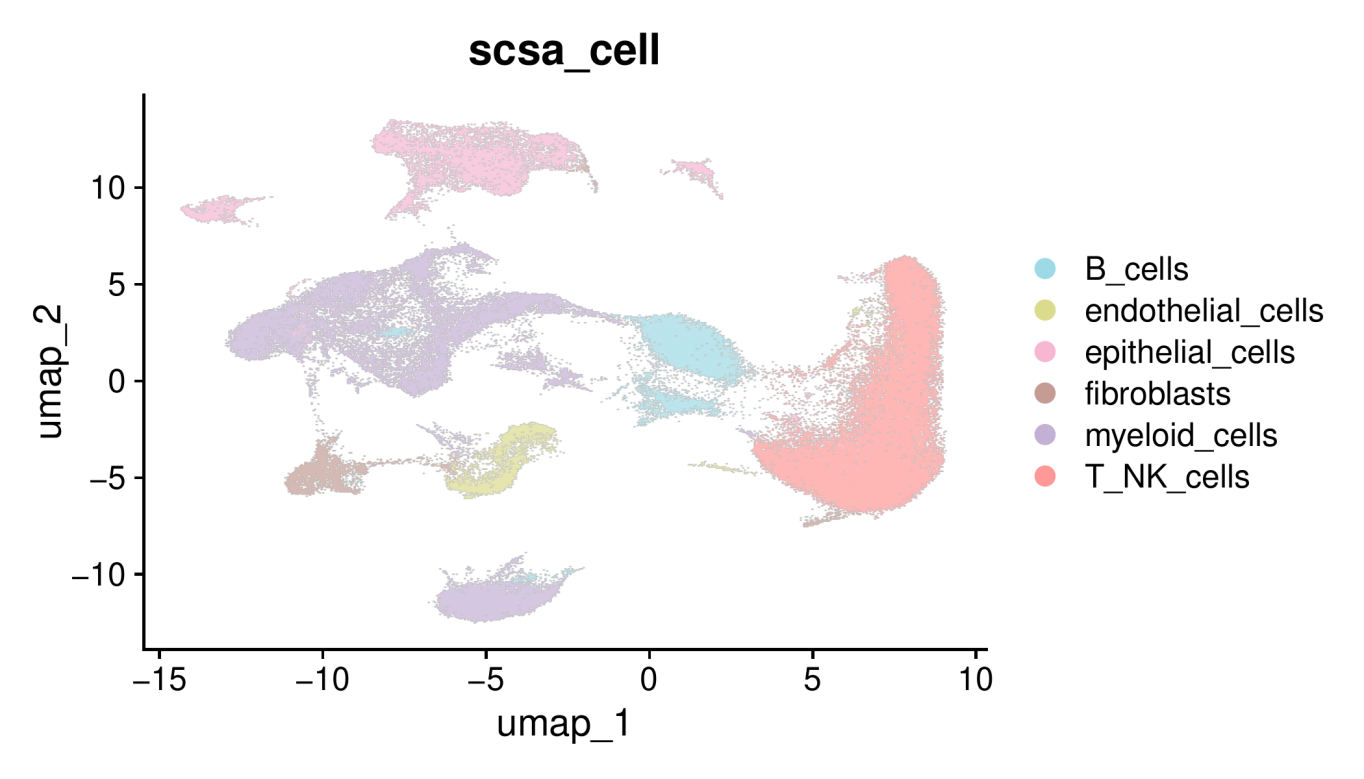
Fig. **[23](#LUAD-UMAP-Integrated)** 为 去除批次效应之后的 UMAP 聚类图。

**Tab.** **8** LUAD significant markers of cell clusters

| Rownames | P val | Avg log2FC | Pct.1 | Pct.2 |
| --- | --- | --- | --- | --- |
| IL7R | 0 | 2.179 | 0.884 | 0.342 |
| CD3D | 0 | 1.365 | 0.736 | 0.313 |
| CD2 | 0 | 1.513 | 0.639 | 0.282 |
| CD3E | 0 | 1.379 | 0.632 | 0.28 |
| LTB | 0 | 1.672 | 0.617 | 0.286 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-significant-markers-of-cell-clusters.csv)**

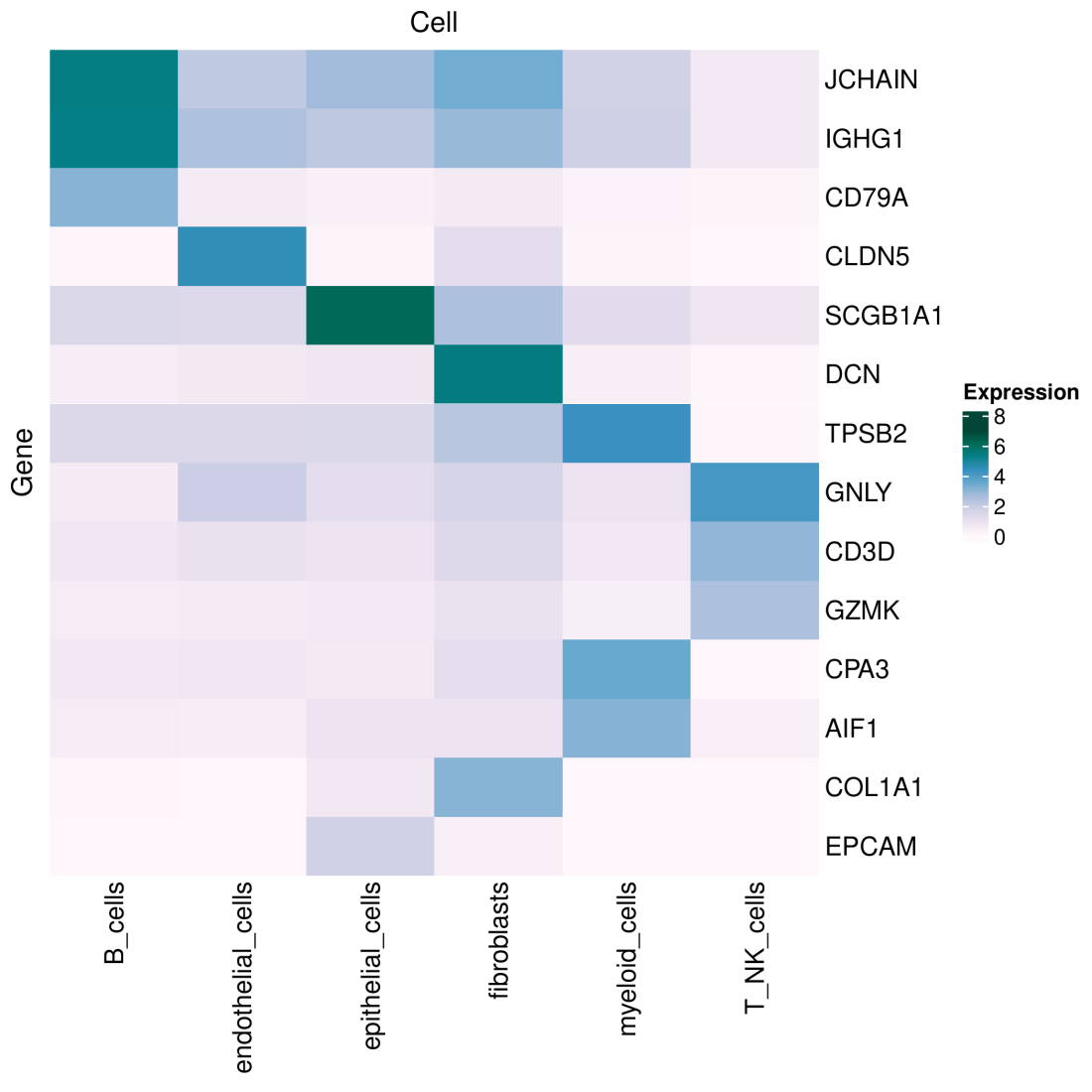
Tab. **[8](#LUAD-significant-markers-of-cell-clusters)** 为所有细胞群的 Marker (LogFC 阈值 0.25; 最小检出率 0.25; 矫正 P 值阈值 0.05)



**Fig.** **24** LUAD SCSA Cell type annotation

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-SCSA-Cell-type-annotation.pdf)**

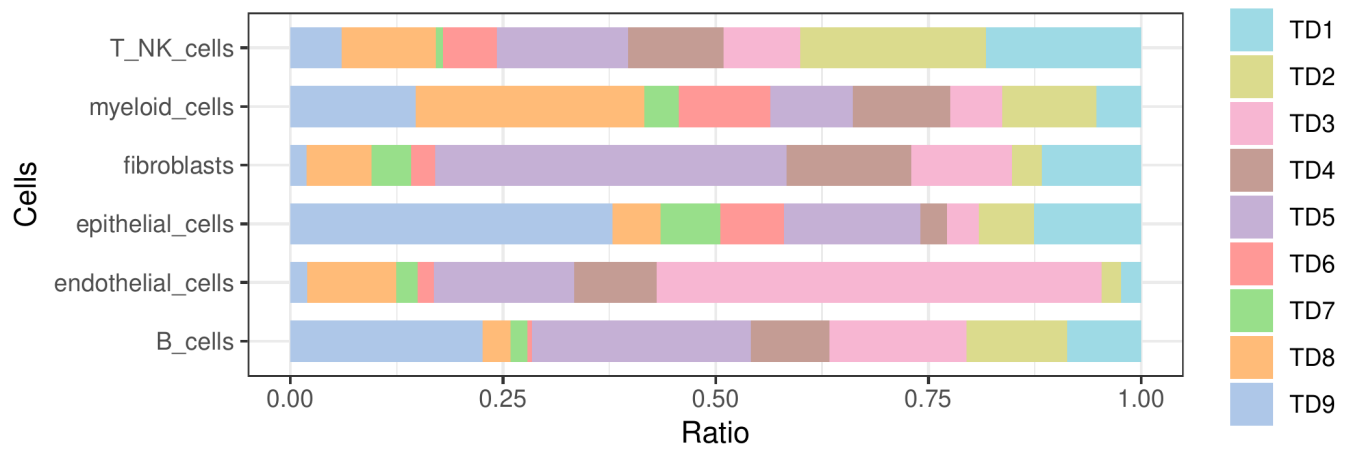
Fig. **[24](#LUAD-SCSA-Cell-type-annotation)** 为 SCSA 细胞注释结果的 UMAP 图。



**Fig.** **25** LUAD Marker Validation

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-Marker-Validation.pdf)**

Fig. **[25](#LUAD-Marker-Validation)** 使用特异性 Marker 对细胞注释结果的验证热图。

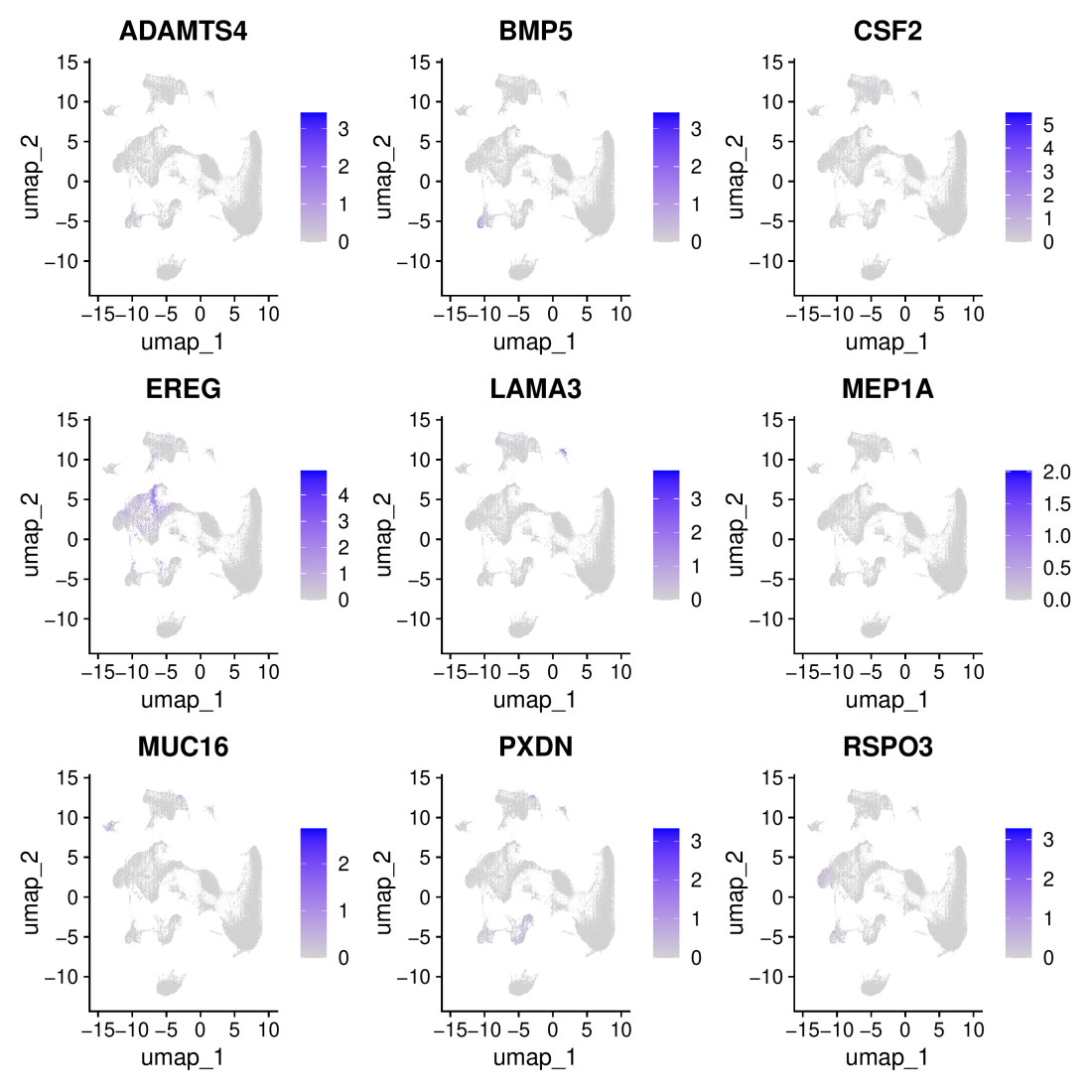


**Fig.** **26** LUAD SCSA Cell Proportions in each sample

**(File path: Figure+Table/3.13.0\_Seurat\_集成单细胞数据分析\_(LUAD)/LUAD-SCSA-Cell-Proportions-in-each-sample.pdf)**

Fig. **[26](#LUAD-SCSA-Cell-Proportions-in-each-sample)** 为 SCSA 注释的细胞群在各个样本中的占比。

## 3.14 (预后模型基因) 首要表达的细胞



**Fig.** **27** LUAD dimension plot of expression level genes

**(File path: Figure+Table/3.14.0\_(预后模型基因)\_首要表达的细胞/LUAD-dimension-plot-of-expression-level-genes.pdf)**

Fig. **[27](#LUAD-dimension-plot-of-expression-level-genes)** 基因 **基因集** (ADAMTS4, BMP5, CSF2, …[n = 9], 来自于Survival 生存分析[Section: TCGA\_LUAD]) 表达水平的 Dimension reduction plot.

# 4 总结

本分析聚焦于基质体，分析肺癌 (LUAD)，通过差异分析、WGCNA 等方式，筛选出关键基质体基因，随后以COX、LASSO 等 算法，构建预后模型，再以外部数据集验证。随后，我们在单细胞数据集 (LUAD) 中，确认了这些基因的表达。 其中，EREG 集中表达于 myeloid\_cells，该细胞类型可能是基质体在 LUAD 发挥作用的关键细胞。

# Reference

1. Wang, S. & Liu, X. The ucscxenatools r package: A toolkit for accessing genomics data from ucsc xena platform, from cancer multi-omics to single-cell rna-seq. *Journal of Open Source Software* **4**, 1627 (2019).

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

3. 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).

4. Langfelder, P. & Horvath, S. WGCNA: An r package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).

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

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

7. Cao, Y., Wang, X. & Peng, G. SCSA: A cell type annotation tool for single-cell rna-seq data. *Frontiers in genetics* **11**, (2020).

8. Naba, A. *et al.* The matrisome: In silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Molecular & cellular proteomics : MCP* **11**, (2012).