**生信分析报告**

**项目标题： 肺癌 ;**

**单 号： BSXG240327 ;**

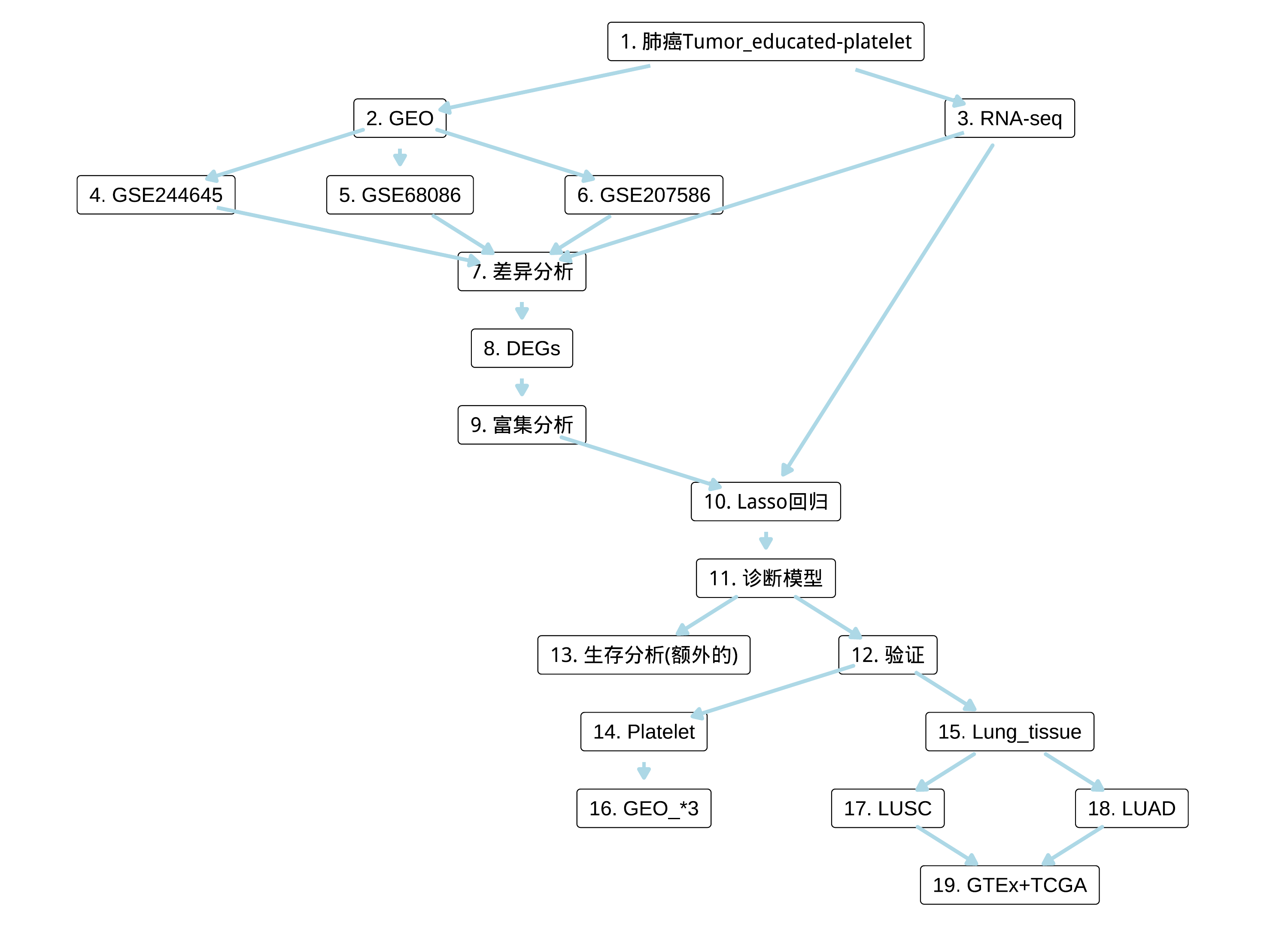
**分析人员： 黄礼闯 ;**

**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



**Fig.** Route

**(File path: Figure+Table\_scRNA/1.0.0\_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 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 (platelet[Description]) AND ((6:1000[Number of Samples]) AND (GSE[Entry Type]) AND (Homo Sapiens[Organism]))。 以正则匹配，滤除包含 ‘single cell’ 或 ‘scRNA’ 的数据例。仅获取类型包含 ‘Expression profiling by high throughput sequencing’ 或 ‘Expression profiling by array’ 的数据例。

## 2.3 GEO 数据获取 (Dataset: LUNG\_GSE244645)

以 R 包 GEOquery (2.74.0) 获取 GSE244645 数据集。

## 2.4 Limma 差异分析 (Dataset: LUNG\_GSE244645)

以 limma (3.62.1) (2005)1 差异分析。创建设计矩阵，对比矩阵，差异分析：tumour\_presence vs tumour\_free。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。

## 2.5 GEO 数据获取 (Dataset: LUNG\_GSE68086)

以 R 包 GEOquery (2.74.0) 获取 GSE68086 数据集。 以 GEOquery:::getRNASeqQuantResults 获取 RNA count 数据 (NCBI-generated data, 参考 <https://www.ncbi.nlm.nih.gov/geo/info/rnaseqcounts.html>) 缺失样本: GSM1662721 (‘NCBI-generated data’ 缺失样本计数数据的原因包括运行未通过 50% 的对齐率或由于技术原因处理失败) 以及基因注释。

## 2.6 Limma 差异分析 (Dataset: LUNG\_GSE68086)

以 R 包 edgeR (4.4.0) ()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 (3.62.1) (2005)1 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：Blood\_Platelets\_Lung vs Blood\_Platelets\_HC。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。

## 2.7 GEO 数据获取 (Dataset: LUNG\_GSE207586)

以 R 包 GEOquery (2.74.0) 获取 GSE207586 数据集。 以 GEOquery:::getRNASeqQuantResults 获取 RNA count 数据 (NCBI-generated data, 参考 <https://www.ncbi.nlm.nih.gov/geo/info/rnaseqcounts.html>) 缺失样本: GSM6300398, GSM6300420, GSM6300459, GSM6300460, GSM6300462 (‘NCBI-generated data’ 缺失样本计数数据的原因包括运行未通过 50% 的对齐率或由于技术原因处理失败) 以及基因注释。

## 2.8 Limma 差异分析 (Dataset: LUNG\_GSE207586)

以 R 包 edgeR (4.4.0) ()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 (3.62.1) (2005)1 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：NSCLC vs Control。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。

## 2.9 Limma 差异分析 (Dataset: LUNG\_SEQ)

以 R 包 edgeR (4.4.0) ()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 (3.62.1) (2005)1 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：Lung\_cancer vs Control。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。

## 2.10 ClusterProfiler 富集分析 (Dataset: VENN)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)3进行 KEGG 和 GO 富集分析。以 p.adjust 表示显著水平。

## 2.11 Lasso 诊断模型建立 (Dataset: LUNG\_SEQ)

以 R 包 glmnet (4.1.8) 作 lasso 处罚的 binomial 回归，以 cv.glmnet 函数作 10 交叉验证获得模型。

## 2.12 UCSCXenaTools 癌症相关数据获取 (Dataset: LUNG\_LUSC)

以 R 包 UCSCXenaTools (1.6.0) (2019, Journal of Open Source Software)4 获取 TcgaTargetGtex 类型数据。

## 2.13 UCSCXenaTools 癌症相关数据获取 (Dataset: LUNG\_LUAD)

以 R 包 UCSCXenaTools (1.6.0) (2019, Journal of Open Source Software)4 获取 TcgaTargetGtex 类型数据。

## 2.14 Survival 生存分析 (Dataset: LUSC)

以 R 包 survival (3.7.0) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

## 2.15 Survival 生存分析 (Dataset: LUAD)

以 R 包 survival (3.7.0) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

# 3 分析结果

## 3.1 Lung cancer (GEO Platelet)

## 3.2 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) 。

使用的 GEO 数据集见 Tab.

**Tab.** Lung cancer EDirect query

| GSE | Title | Summary | Taxon | GdsType |
| --- | --- | --- | --- | --- |
| GSE244645 | Expression data f... | Gene expression p... | Homo sapiens | Expression profil... |
| GSE207586 | Tumor-educated pl... | We report RNA-seq... | Homo sapiens | Expression profil... |
| GSE68086 | RNA-seq of tumor-... | We report RNA-seq... | Homo sapiens | Expression profil... |

**(File path: Figure+Table\_scRNA/3.2.0\_GSE\_数据搜索\_(LUNG)/Lung-cancer-EDirect-query.xlsx)**

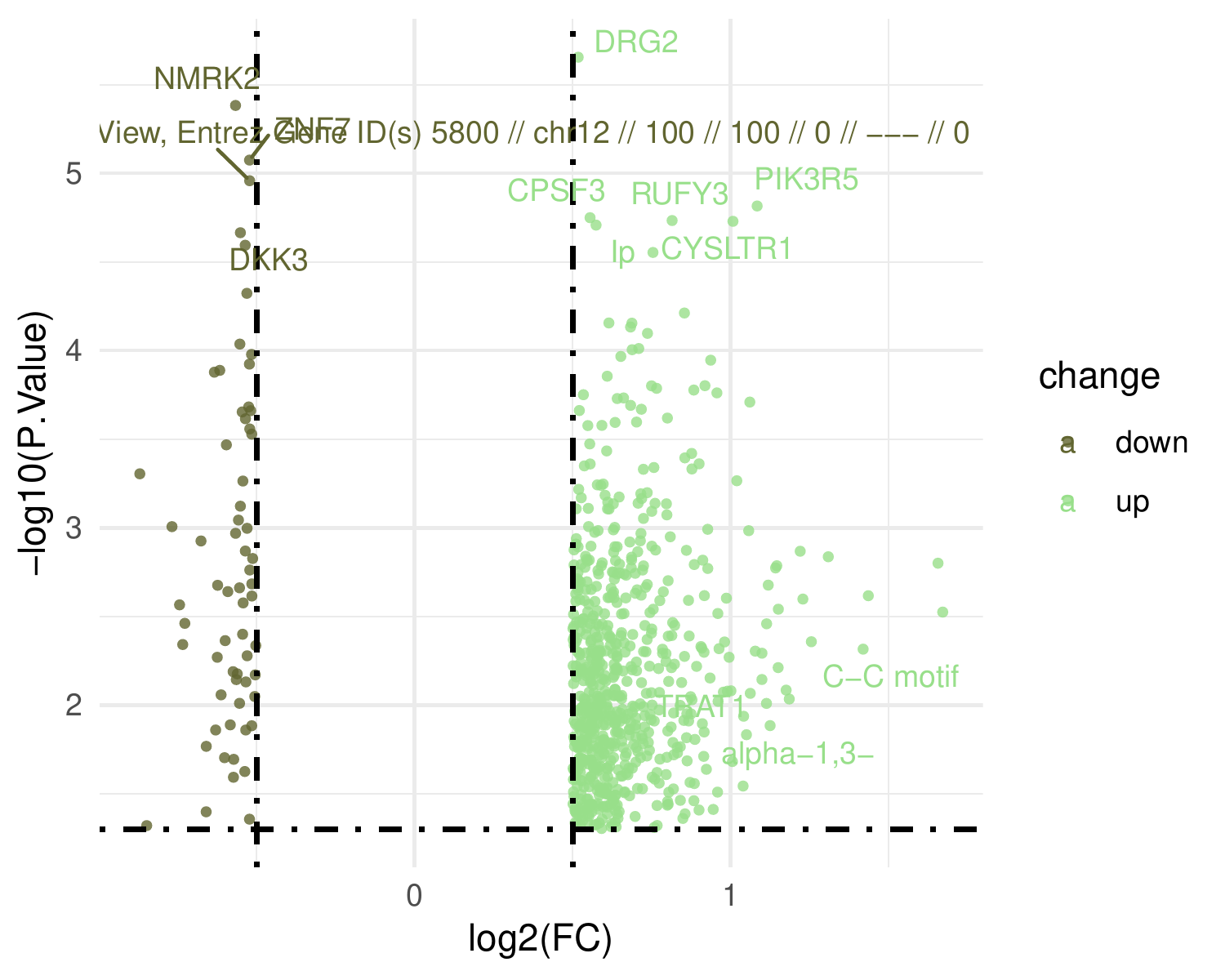
### 3.2.1 GEO 数据获取 (LUNG\_GSE244645)

以 GEOquery 获取 GSE244645 的数据信息。

### 3.2.2 Limma 差异分析 (LUNG\_GSE244645)

匹配 title 中包含“NSCLC”的描述，最终得到 26 例数据。样本分组：tumour\_free (n=11) , tumour\_presence (n=15) 。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：tumour\_presence vs tumour\_free。(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=658) , down (n=63)

Fig. 为 tumour\_presence - tumour\_free 差异分析火山图。 Tab. 为 tumour\_presence - tumour\_free 差异分析统计表格。



**Fig.** LUNG GSE244645 tumour presence vs tumour free

**(File path: Figure+Table\_scRNA/3.2.2\_Limma\_差异分析\_(LUNG\_GSE244645)/LUNG-GSE244645-tumour-presence-vs-tumour-free.pdf)**

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

**(See: Figure+Table\_scRNA/3.2.2\_Limma\_差异分析\_(LUNG\_GSE244645)/LUNG-GSE244645-tumour-presence-vs-tumour-free-content)**

**Tab.** LUNG GSE244645 data tumour presence vs tumour free

| GENE SYMBOL | LogFC | P.Value | Rownames | ID |
| --- | --- | --- | --- | --- |
| DRG2 | 0.5176 | 2.216e-06 | TC1700007099.hg.1 | TC1700007099.hg.1 |
| NMRK2 | -0.5662 | 4.145e-06 | TC1900006662.hg.1 | TC1900006662.hg.1 |
| ZNF7 | -0.5221 | 8.429e-06 | TC0800009291.hg.1 | TC0800009291.hg.1 |
| PTPRO.mAug10-unsp... | -0.5216 | 1.102e-05 | TC1200006959.hg.1 | TC1200006959.hg.1 |
| PIK3R5 | 1.084 | 1.529e-05 | TC1700009708.hg.1 | TC1700009708.hg.1 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.2.2\_Limma\_差异分析\_(LUNG\_GSE244645)/LUNG-GSE244645-data-tumour-presence-vs-tumour-free.tsv)**

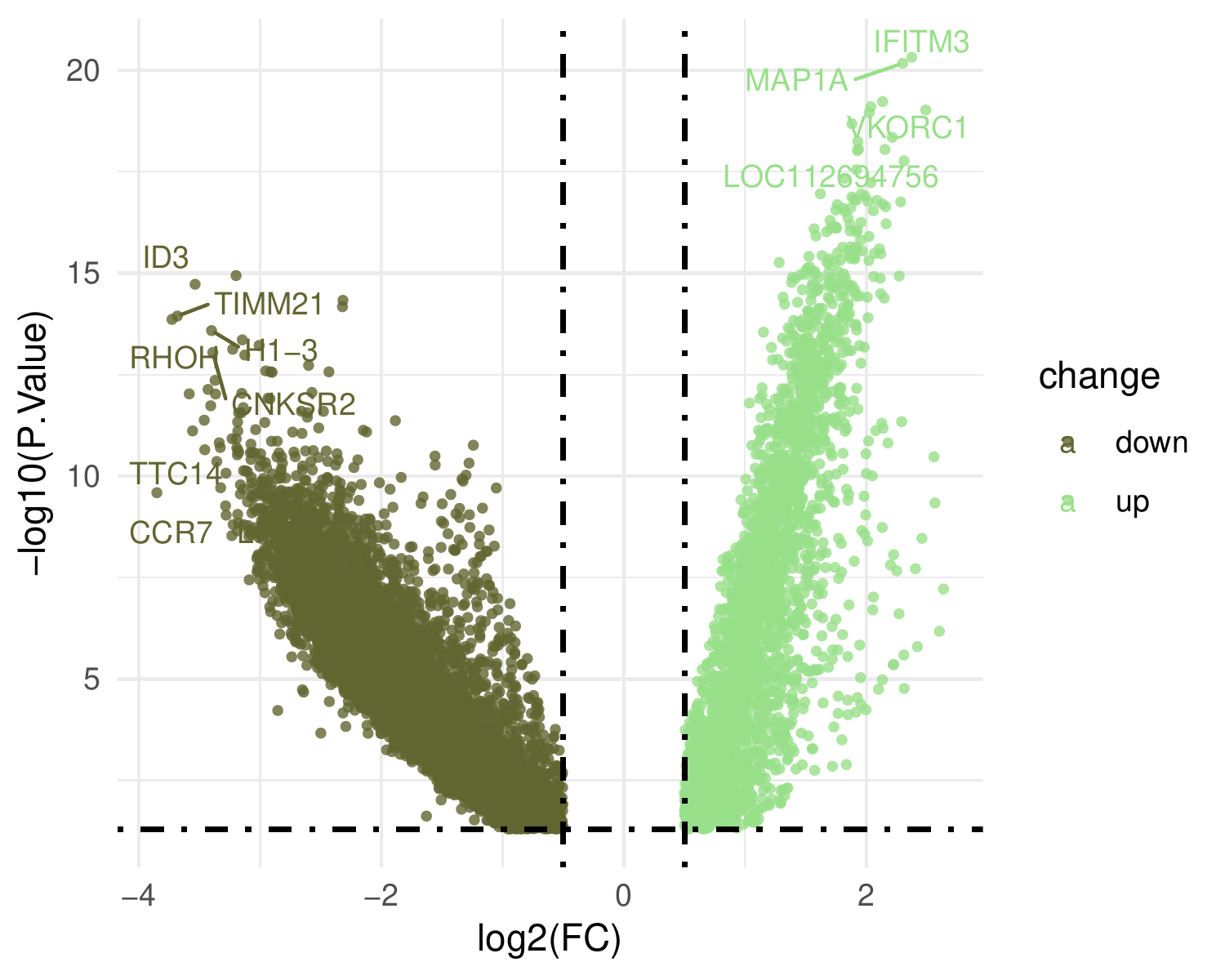
### 3.2.3 GEO 数据获取 (LUNG\_GSE68086)

以 GEOquery 获取 GSE68086 的数据信息。

### 3.2.4 Limma 差异分析 (LUNG\_GSE68086)

匹配 group 中包含“Blood\_Platelets\_Lung|Blood\_Platelets\_HC”的描述，最终得到 114 例数据。以 edgeR 将GSE68086 RNA-seq 数据标准化 (详见方法章节)。以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) (Batch: Batch02 (n=34) , Batch03 (n=22) , Batch04 (n=38) , Batch05 (n=3) , Batch06 (n=17) )。差异分析：Blood\_Platelets\_Lung vs Blood\_Platelets\_HC。(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=2636) , down (n=10184)

Fig. 为 Blood\_Platelets\_Lung - Blood\_Platelets\_HC 差异分析火山图。 Tab. 为 Blood\_Platelets\_Lung - Blood\_Platelets\_HC 差异分析统计表格。



**Fig.** LUNG GSE68086 Blood Platelets Lung vs Blood Platelets HC

**(File path: Figure+Table\_scRNA/3.2.4\_Limma\_差异分析\_(LUNG\_GSE68086)/LUNG-GSE68086-Blood-Platelets-Lung-vs-Blood-Platelets-HC.pdf)**

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

**(See: Figure+Table\_scRNA/3.2.4\_Limma\_差异分析\_(LUNG\_GSE68086)/LUNG-GSE68086-Blood-Platelets-Lung-vs-Blood-Platelets-HC-content)**

**Tab.** LUNG GSE68086 data Blood Platelets Lung vs Blood Platelets HC

| Hgnc symbol | LogFC | P.Value | Rownames | GeneID |
| --- | --- | --- | --- | --- |
| IFITM3 | 2.374 | 4.777e-21 | 10410 | 10410 |
| MAP1A | 2.3 | 6.724e-21 | 4130 | 4130 |
| HSPB1 | 2.134 | 5.915e-20 | 3315 | 3315 |
| TIMP1 | 2.036 | 7.933e-20 | 7076 | 7076 |
| ITGA2B | 2.49 | 9.523e-20 | 3674 | 3674 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.2.4\_Limma\_差异分析\_(LUNG\_GSE68086)/LUNG-GSE68086-data-Blood-Platelets-Lung-vs-Blood-Platelets-HC.tsv)**

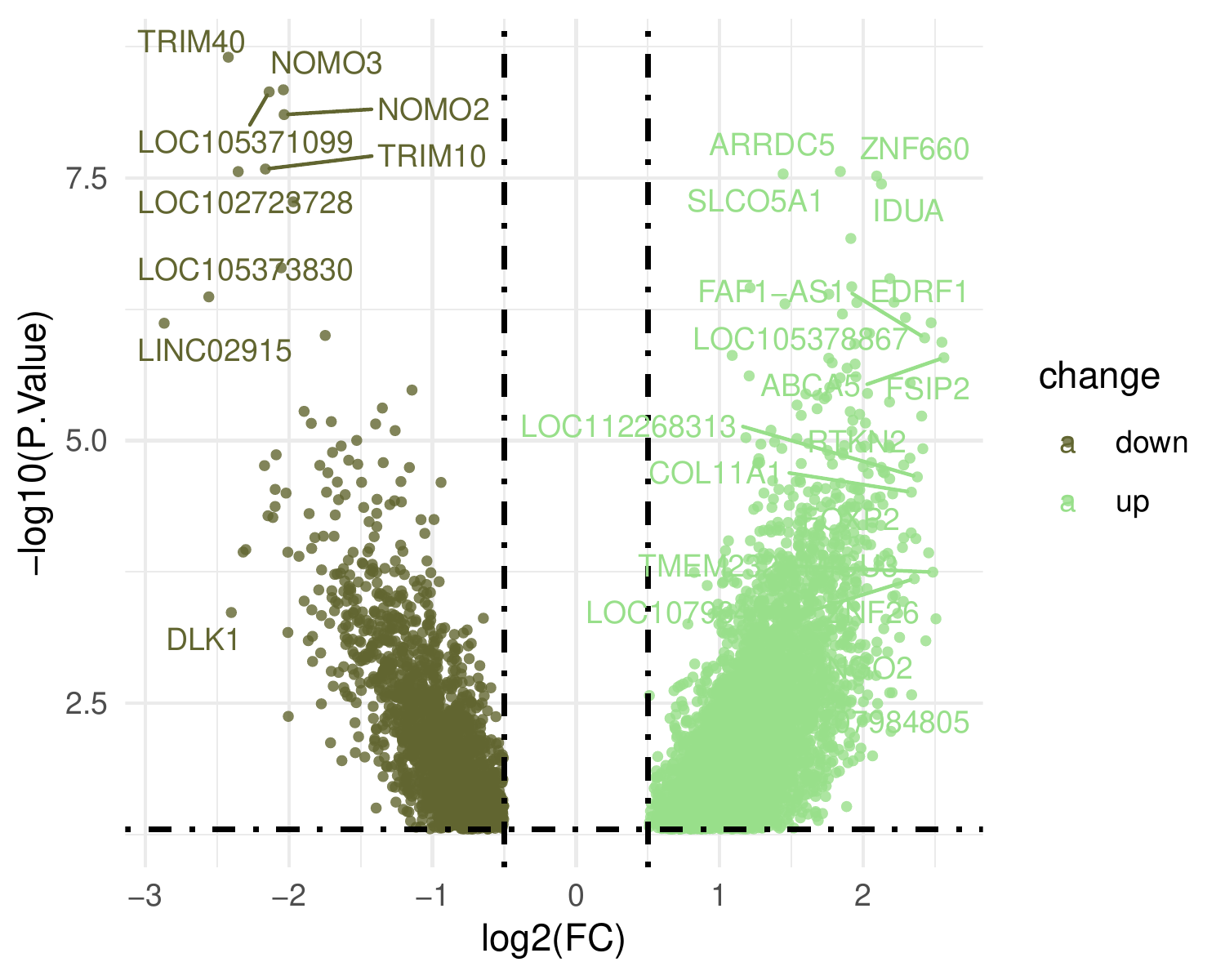
### 3.2.5 GEO 数据获取 (LUNG\_GSE207586)

以 GEOquery 获取 GSE207586 的数据信息。

### 3.2.6 Limma 差异分析 (LUNG\_GSE207586)

样本分组：Control (n=18) , NSCLC (n=153) 。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：NSCLC vs Control。(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=5331) , down (n=2026)

Fig. 为 NSCLC - Control 差异分析火山图。 Tab. 为 NSCLC - Control 差异分析统计表格。



**Fig.** LUNG GSE207586 NSCLC vs Control

**(File path: Figure+Table\_scRNA/3.2.6\_Limma\_差异分析\_(LUNG\_GSE207586)/LUNG-GSE207586-NSCLC-vs-Control.pdf)**

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

**(See: Figure+Table\_scRNA/3.2.6\_Limma\_差异分析\_(LUNG\_GSE207586)/LUNG-GSE207586-NSCLC-vs-Control-content)**

**Tab.** LUNG GSE207586 data NSCLC vs Control

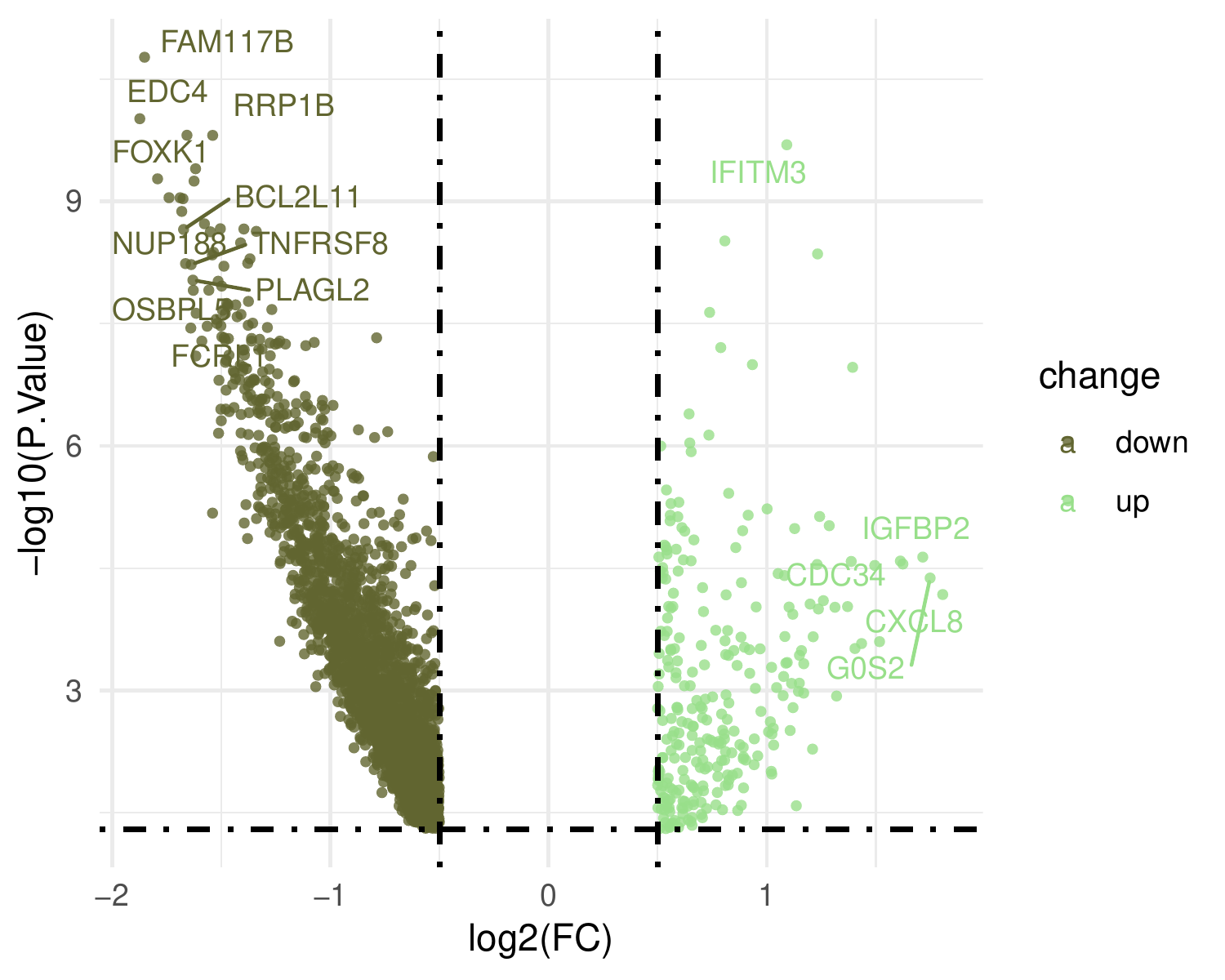
| Hgnc symbol | LogFC | P.Value | Rownames | GeneID |
| --- | --- | --- | --- | --- |
| TRIM40 | -2.423 | 2.242e-09 | 135644 | 135600 |
| NOMO3 | -2.04 | 4.578e-09 | 408050 | 408000 |
| LOC105371099 | -2.139 | 4.791e-09 | 105371099 | 105400000 |
| NOMO2 | -2.034 | 7.884e-09 | 283820 | 283800 |
| TRIM10 | -2.165 | 2.601e-08 | 10107 | 10110 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.2.6\_Limma\_差异分析\_(LUNG\_GSE207586)/LUNG-GSE207586-data-NSCLC-vs-Control.tsv)**

### 3.2.7 Limma 差异分析 (LUNG\_SEQ)

测序数据肺癌 RNA-seq, 以 mRNA (protein\_coding) 差异分析。 样本分组：Control (n=81) , Lung\_cancer (n=166) 。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Lung\_cancer vs Control。(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=273) , down (n=2538)

Fig. 为 Lung\_cancer - Control 差异分析火山图。 Tab. 为 Lung\_cancer - Control 差异分析统计表格。



**Fig.** LUNG SEQ Lung cancer vs Control

**(File path: Figure+Table\_scRNA/3.2.7\_Limma\_差异分析\_(LUNG\_SEQ)/LUNG-SEQ-Lung-cancer-vs-Control.pdf)**

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

**(See: Figure+Table\_scRNA/3.2.7\_Limma\_差异分析\_(LUNG\_SEQ)/LUNG-SEQ-Lung-cancer-vs-Control-content)**

**Tab.** LUNG SEQ data Lung cancer vs Control

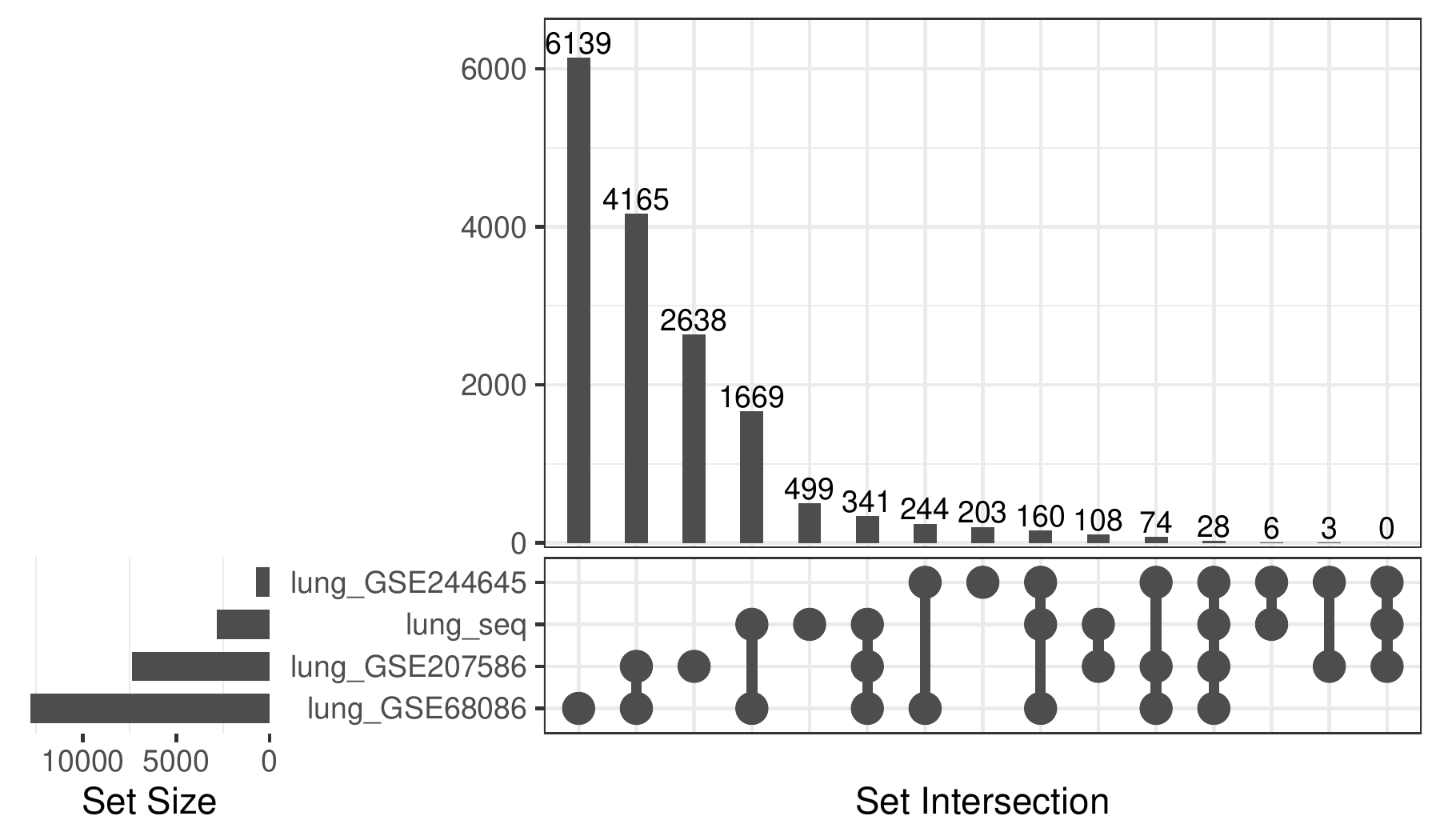
| Hgnc symbol | LogFC | P.Value | Rownames | Ensembl gene id |
| --- | --- | --- | --- | --- |
| FAM117B | -1.851 | 1.712e-11 | ENSG00000138439 | ENSG00000138439 |
| EDC4 | -1.873 | 9.72e-11 | ENSG00000038358 | ENSG00000038358 |
| RRP1B | -1.539 | 1.555e-10 | ENSG00000160208 | ENSG00000160208 |
| FOXK1 | -1.657 | 1.548e-10 | ENSG00000164916 | ENSG00000164916 |
| IFITM3 | 1.092 | 2.028e-10 | ENSG00000142089 | ENSG00000142089 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.2.7\_Limma\_差异分析\_(LUNG\_SEQ)/LUNG-SEQ-data-Lung-cancer-vs-Control.tsv)**

## 3.3 交集: lung\_GSE207586 + lung\_GSE244645 + lung\_GSE68086 + lung\_seq (LUNG)

以下取交集： - 基因集 (NSCLC - Control, 来自于Limma 差异分析[Section: LUNG\_GSE207586]) - 基因集 (tumour\_presence - tumour\_free, 来自于Limma 差异分析[Section: LUNG\_GSE244645]) - 基因集 (Blood\_Platelets\_Lung - Blood\_Platelets\_HC, 来自于Limma 差异分析[Section: LUNG\_GSE68086]) - 基因集 (Lung\_cancer - Control, 来自于Limma 差异分析[Section: LUNG\_SEQ])

Fig. 将lung\_GSE207586, lung\_GSE244645, lung\_GSE68086, lung\_seq 取交集。



**Fig.** Intersection of lung GSE207586 with lung GSE244645 with lung GSE68086 with lung seq

**(File path: Figure+Table\_scRNA/3.3.0\_交集:\_lung\_GSE207586\_+\_lung\_GSE244645\_+\_lung\_GSE68086\_+\_lung\_seq\_(LUNG)/Intersection-of-lung-GSE207586-with-lung-GSE244645-with-lung-GSE68086-with-lung-seq.pdf)**

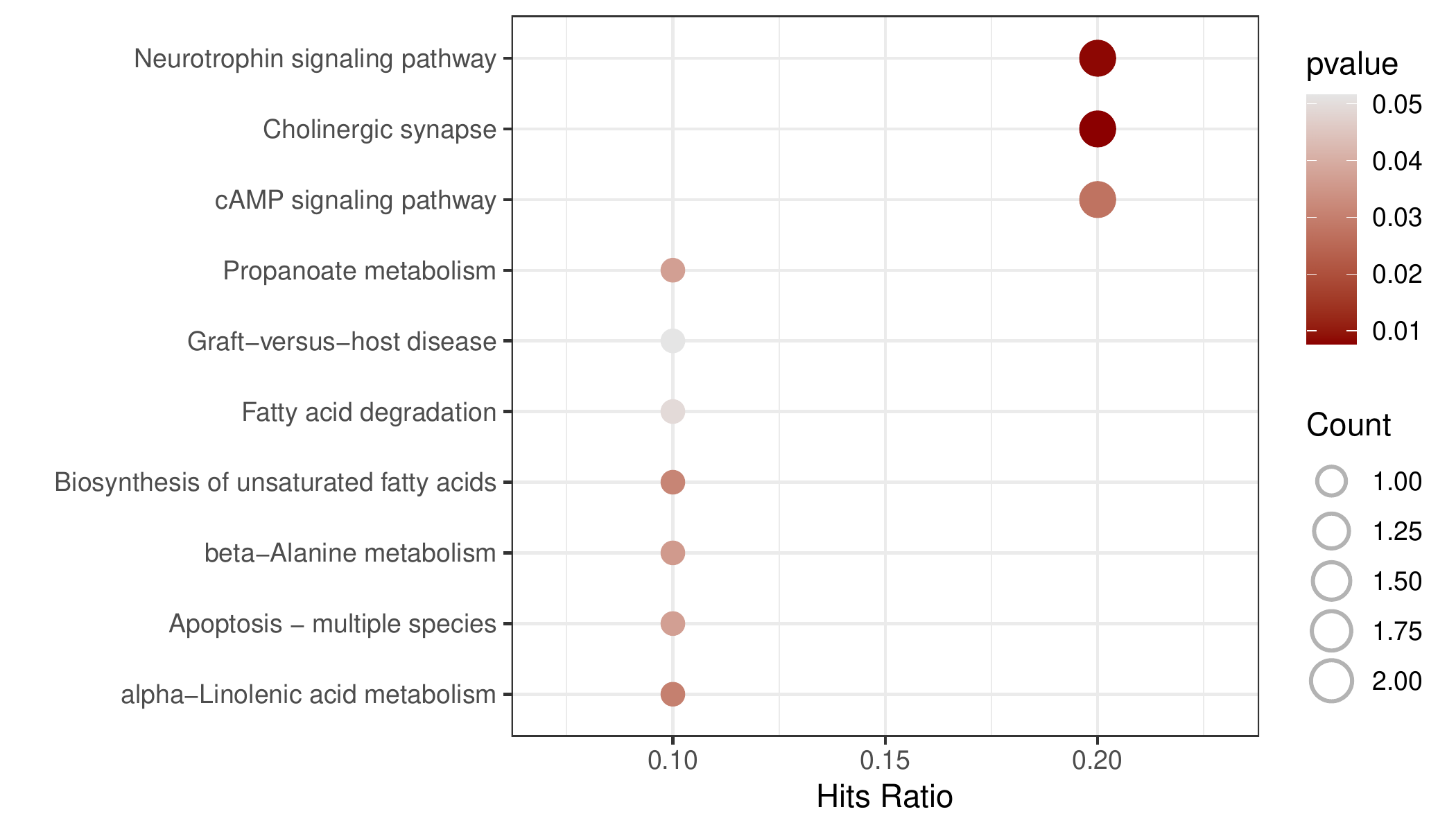
* All\_intersection: HEATR6, SLFN13, LY9, ZKSCAN8, MPZL1, CCNT1, KLRC1, KLHL20, MAML2, WDR55, TRAT1, PRDM1, CAMK4, ZNF398, LINS1, CAMKMT, ACOX3, BCL2, MCM9, SIPA1L3, ZNF830, ZBTB21, CRTC3, N4BP2, TLR10, FAM111A, GAPT, IRF8

**(See: Figure+Table\_scRNA/3.3.0\_交集:\_lung\_GSE207586\_+\_lung\_GSE244645\_+\_lung\_GSE68086\_+\_lung\_seq\_(LUNG)/Intersection-of-lung-GSE207586-with-lung-GSE244645-with-lung-GSE68086-with-lung-seq-content)**

## 3.4 ClusterProfiler 富集分析 (VENN)

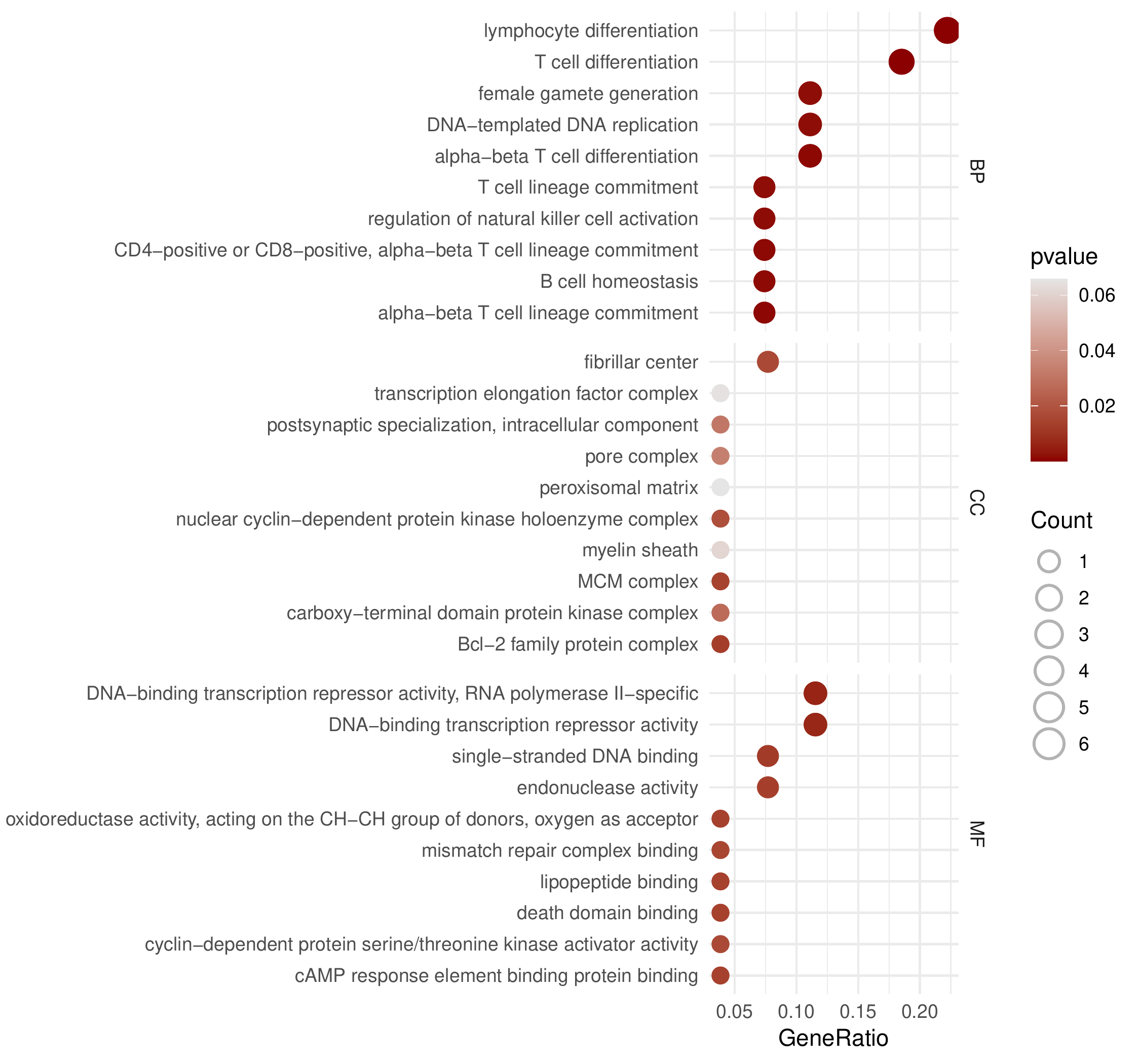
对基因集 (HEATR6, SLFN13, LY9, …[n = 28], 来自于Venn 交集[Section: LUNG]) 进行ClusterProfiler 富集分析。

Fig. 为 GO 富集分析气泡图。 Fig. 为 GO 富集分析气泡图。 Tab. 为 KEGG 富集分析统计表。 Tab. 为 GO 富集分析统计表。



**Fig.** VENN KEGG enrichment

**(File path: Figure+Table\_scRNA/3.4.0\_ClusterProfiler\_富集分析\_(VENN)/VENN-KEGG-enrichment.pdf)**



**Fig.** VENN GO enrichment

**(File path: Figure+Table\_scRNA/3.4.0\_ClusterProfiler\_富集分析\_(VENN)/VENN-GO-enrichment.pdf)**

**Tab.** VENN KEGG enrichment data

| Category | Subcategory | ID | Description | GeneRatio |
| --- | --- | --- | --- | --- |
| Organismal Systems | Nervous system | Hsa04725 | Cholinergic synapse | 2/10 |
| Organismal Systems | Nervous system | Hsa04722 | Neurotrophin sign... | 2/10 |
| Environmental Inf... | Signal transduction | Hsa04024 | CAMP signaling pa... | 2/10 |
| Metabolism | Lipid metabolism | Hsa00592 | Alpha-Linolenic a... | 1/10 |
| Metabolism | Lipid metabolism | Hsa01040 | Biosynthesis of u... | 1/10 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.4.0\_ClusterProfiler\_富集分析\_(VENN)/VENN-KEGG-enrichment-data.xlsx)**

**Tab.** VENN GO enrichment data

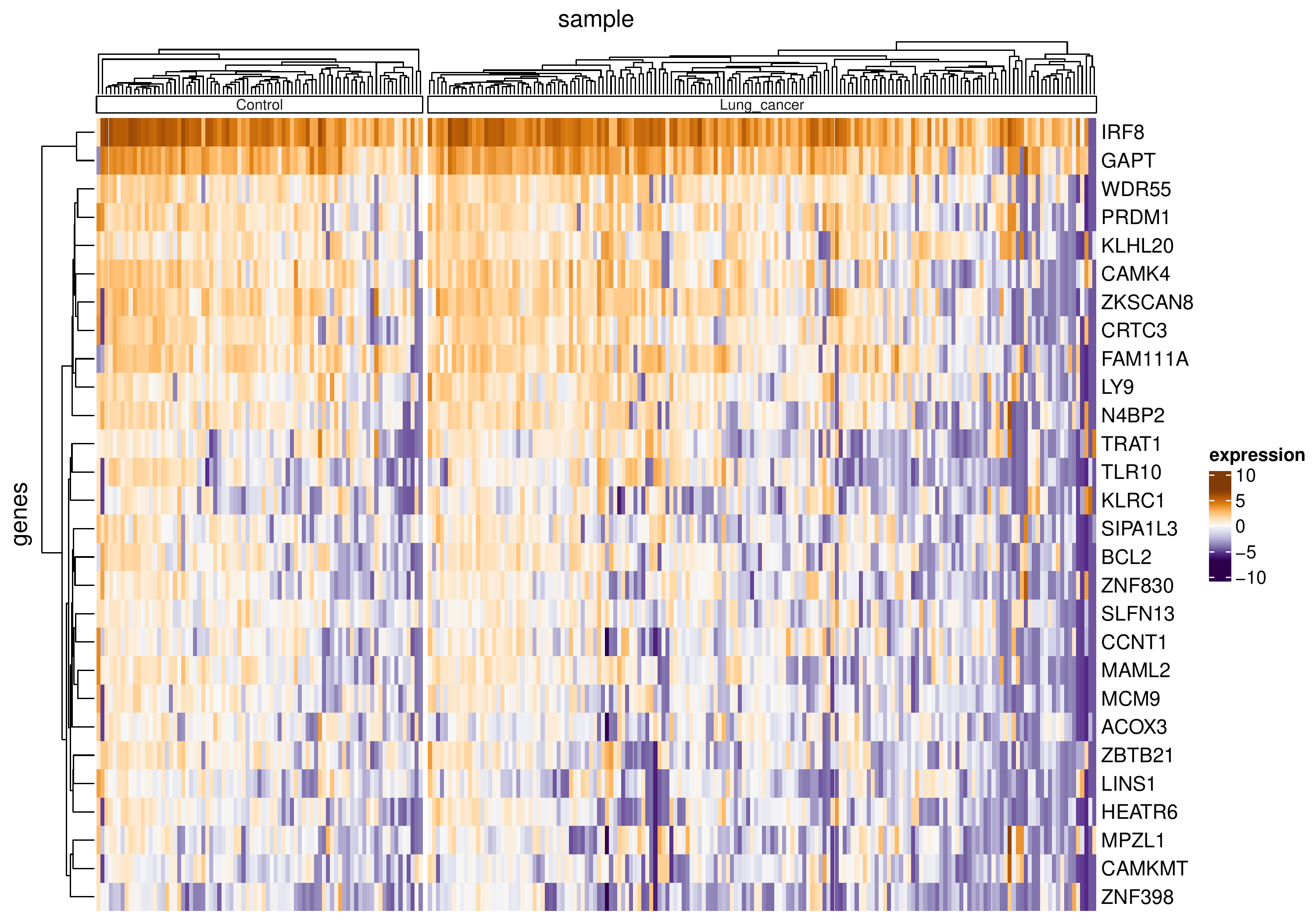
| Ont | ID | Description | GeneRatio | BgRatio |
| --- | --- | --- | --- | --- |
| BP | GO:0030098 | Lymphocyte differ... | 6/27 | 426/18986 |
| BP | GO:0030217 | T cell differenti... | 5/27 | 320/18986 |
| BP | GO:0046632 | Alpha-beta T cell... | 3/27 | 133/18986 |
| BP | GO:0002363 | Alpha-beta T cell... | 2/27 | 33/18986 |
| BP | GO:0043369 | CD4-positive or C... | 2/27 | 35/18986 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.4.0\_ClusterProfiler\_富集分析\_(VENN)/VENN-GO-enrichment-data.xlsx)**

## 3.5 Lasso 诊断模型建立 (LUNG\_SEQ)

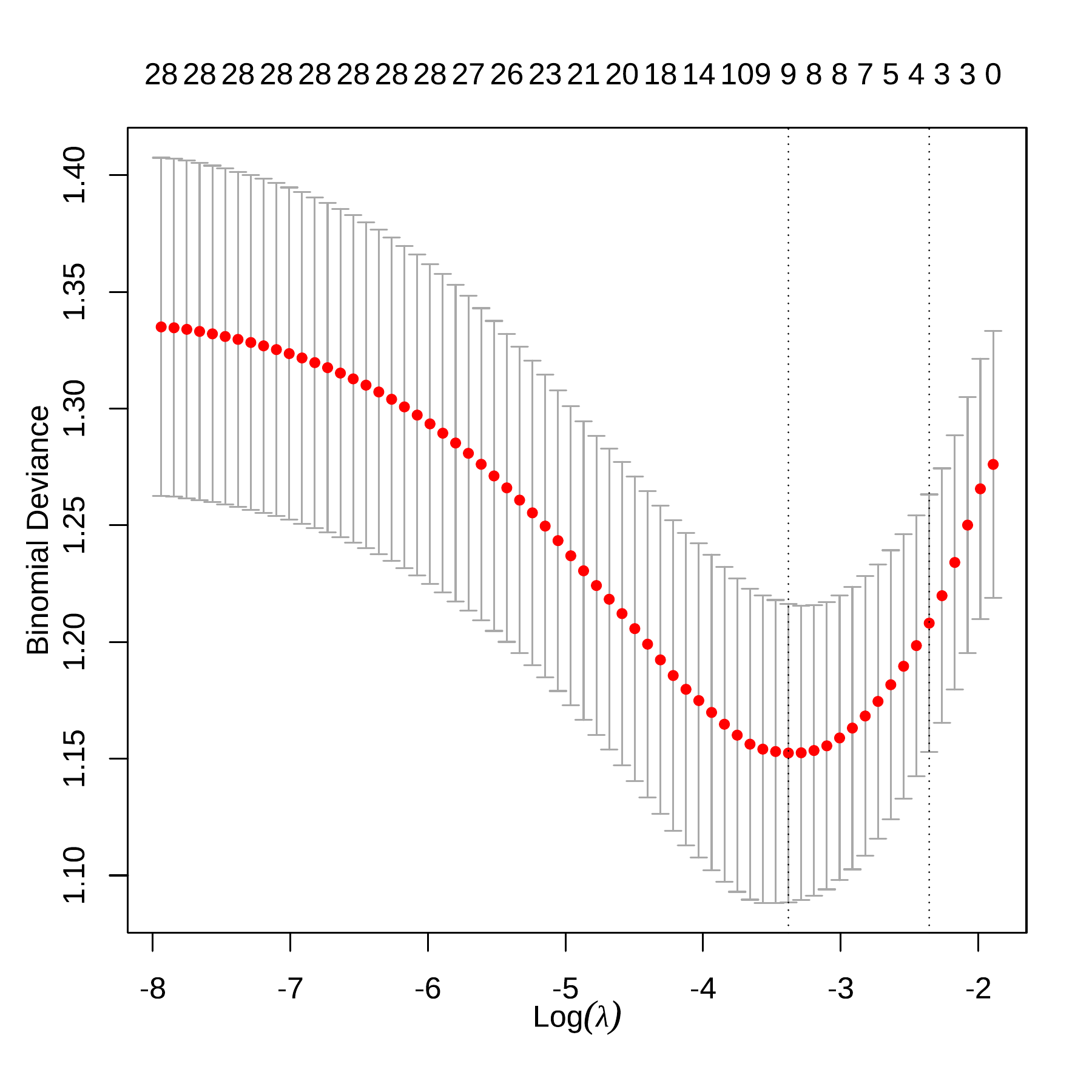
对基因集 (HEATR6, SLFN13, LY9, …[n = 28], 来自于Venn 交集[Section: LUNG]) 进行Lasso 诊断模型建立。使用 glmnet::cv.glmnet 作 10 倍交叉验证 (评估方式为 Binomial Deviance)，筛选 lambda 值。lambda.min, lambda.1se 值分别为 0.034, 0.095 (R 随机种子为 987456)。对应的特征数 (基因数) 分别为 9, 3。

Fig. 为输入基因的表达热图。 Fig. 为模型的交叉验证曲线 (cross-validation curve)。 Fig. 为 lambda.min 下各变量的系数。 Fig. 为 lambda.1se 下各变量的系数。 Tab. 为模型的系数附表。 Fig. 为 lambda.1se 下，模型预测 (以内部数据) 的 ROC 曲线。 Fig. 为 lambda.min 下，模型预测 (以内部数据) 的 ROC 曲线。



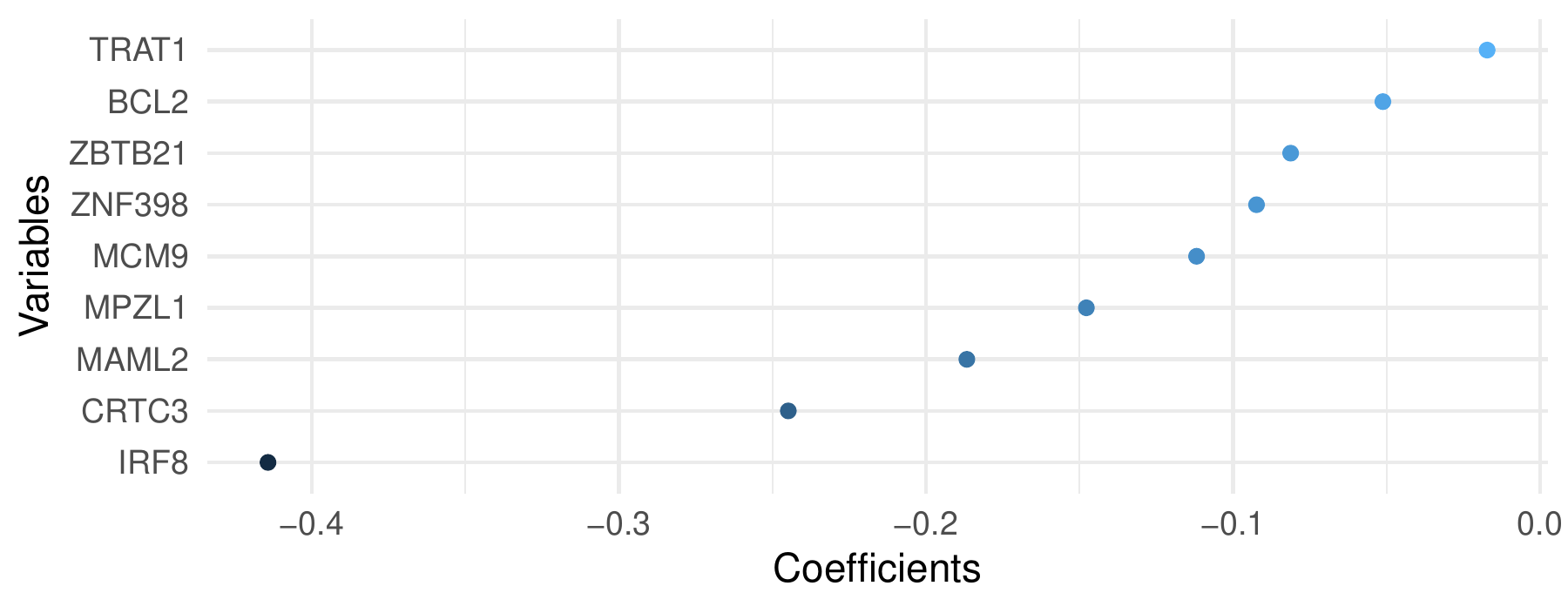
**Fig.** Heatmap of input genes

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/Heatmap-of-input-genes.pdf)**



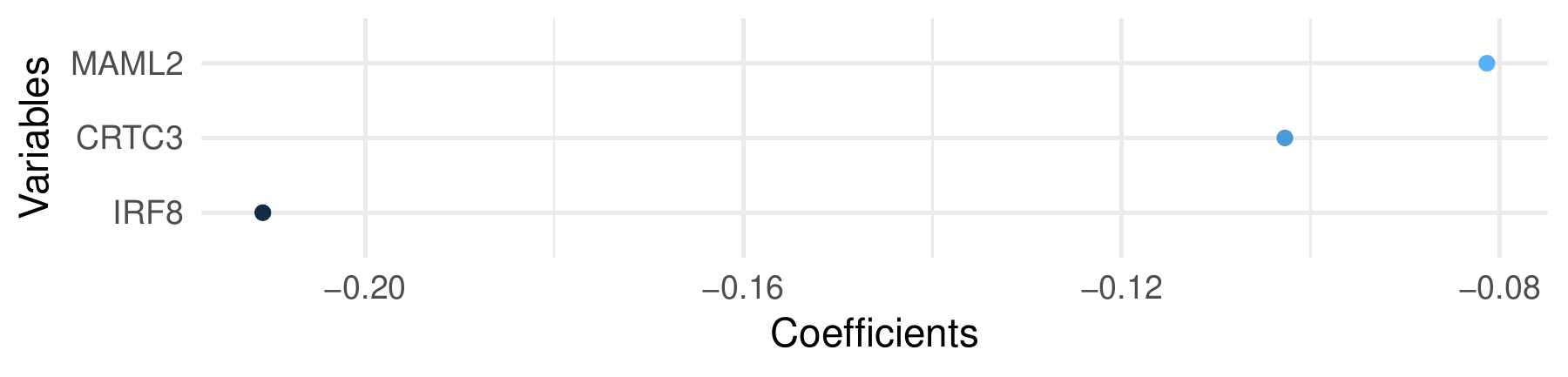
**Fig.** LUNG SEQ lasso model

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/LUNG-SEQ-lasso-model.pdf)**



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

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/LUNG-SEQ-lasso-coeffients-lambda-min.pdf)**



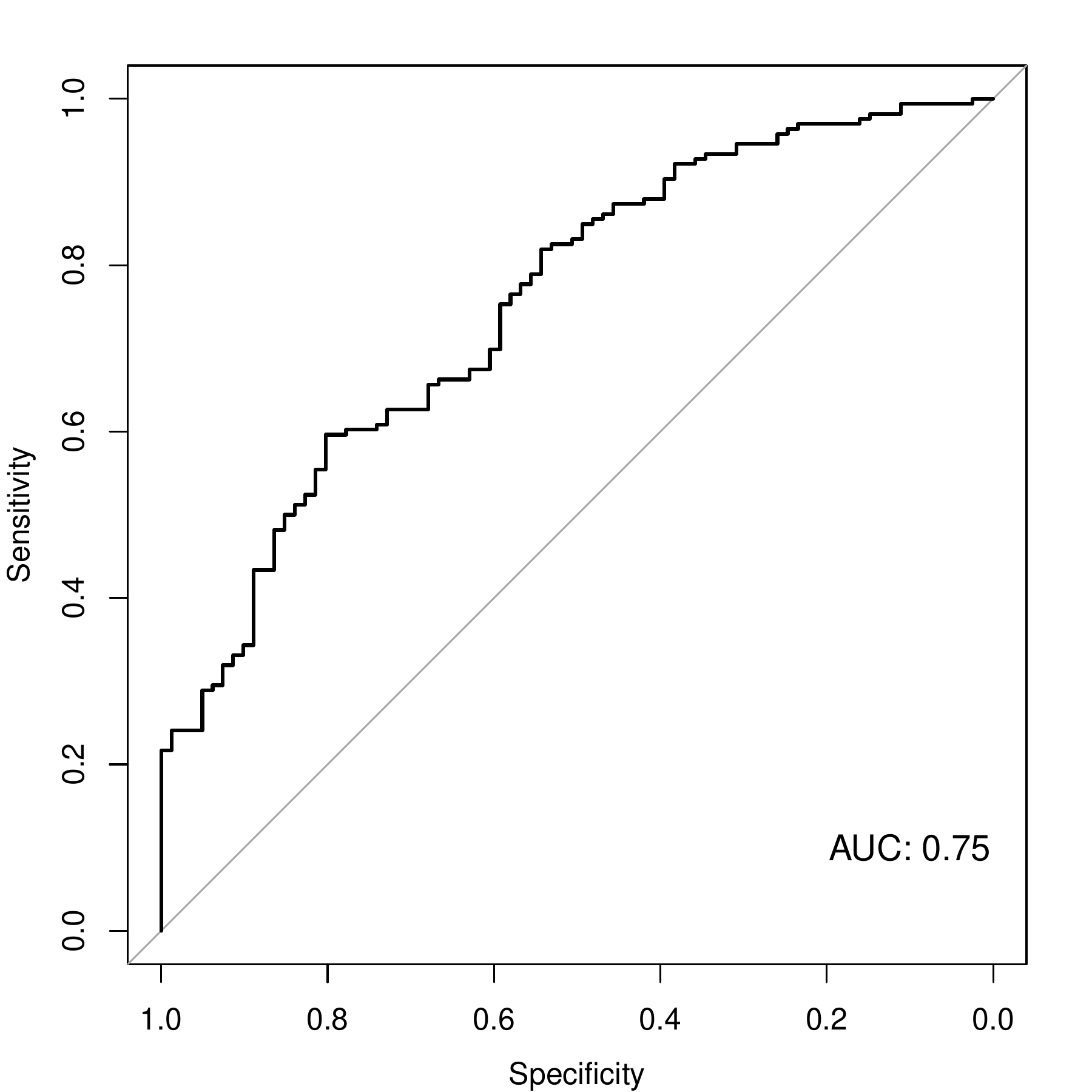
**Fig.** LUNG SEQ lasso coeffients lambda 1se

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/LUNG-SEQ-lasso-coeffients-lambda-1se.pdf)**

**Tab.** LUNG SEQ sig Coefficients

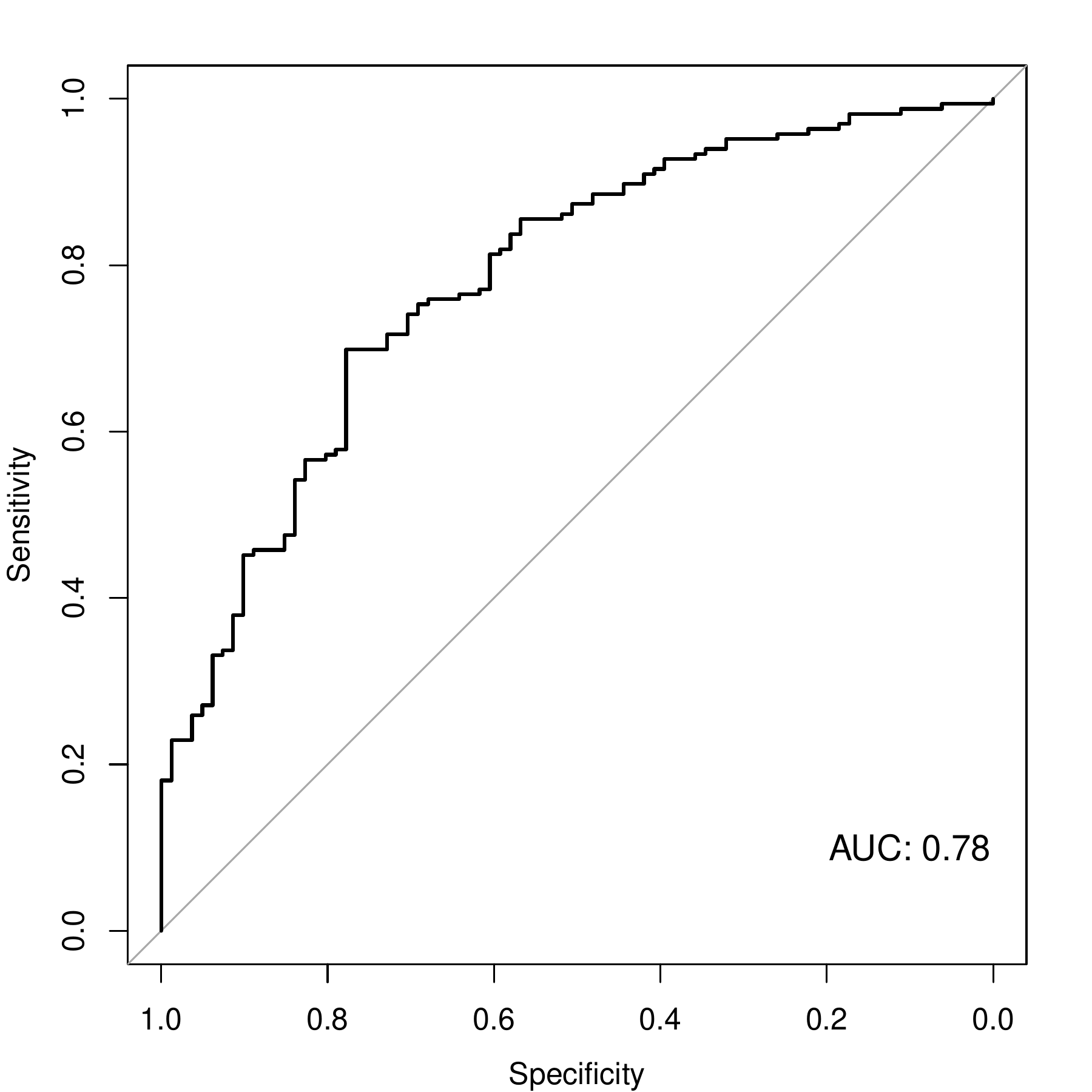
| Feature | Coef.lambda.min | Coef.lambda.1se |
| --- | --- | --- |
| ZNF830 | 0 | 0 |
| ZKSCAN8 | 0 | 0 |
| MPZL1 | -0.1478 | 0 |
| ZNF398 | -0.09235 | 0 |
| MAML2 | -0.1867 | -0.08134 |
| ... | ... | ... |

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/LUNG-SEQ-sig-Coefficients.csv)**



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

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/LUNG-SEQ-lasso-ROC-lambda-1se.pdf)**



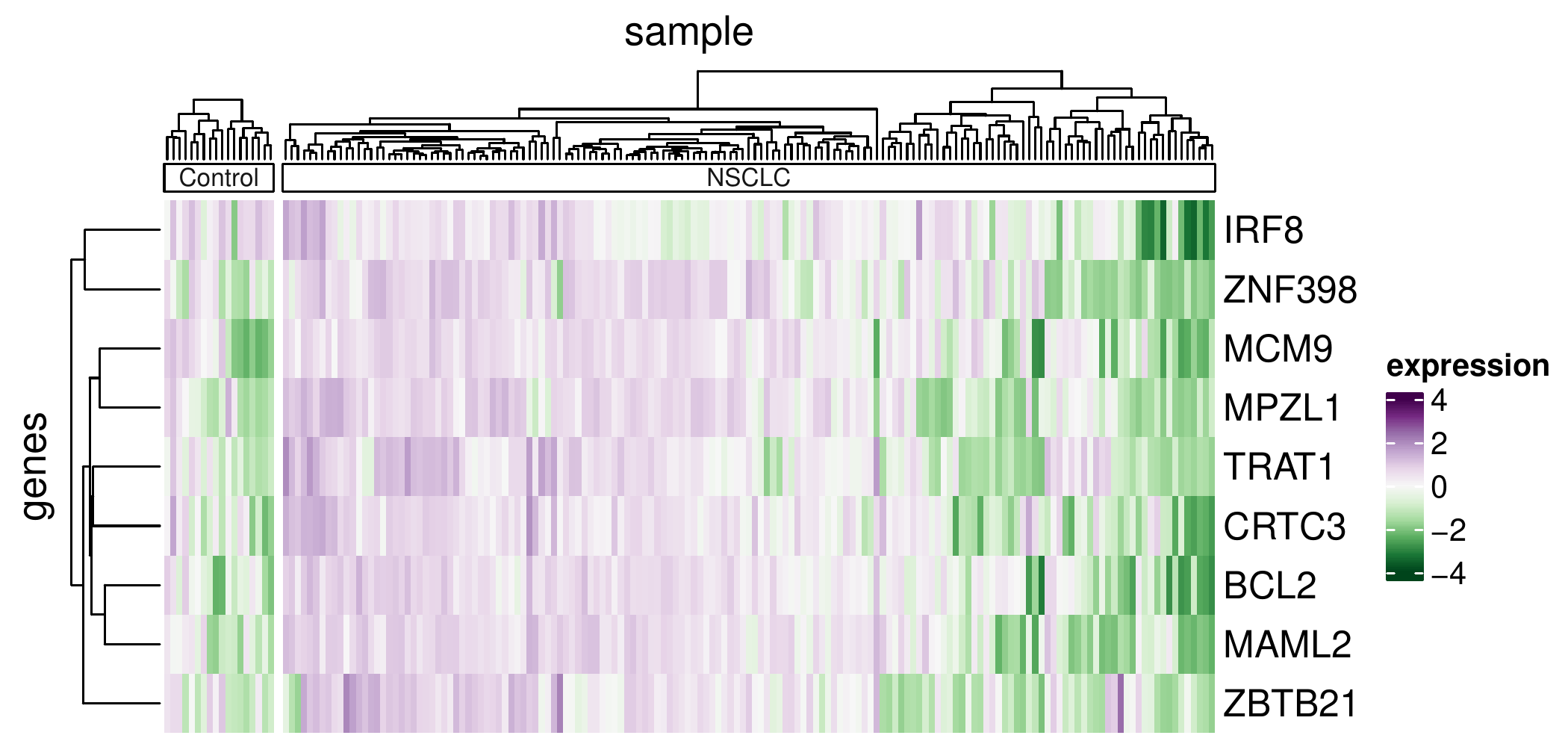
**Fig.** LUNG SEQ lasso ROC lambda min

**(File path: Figure+Table\_scRNA/3.5.0\_Lasso\_诊断模型建立\_(LUNG\_SEQ)/LUNG-SEQ-lasso-ROC-lambda-min.pdf)**

## 3.6 RNA 数据集 (Platelet) 验证 (LUNG\_VALID)

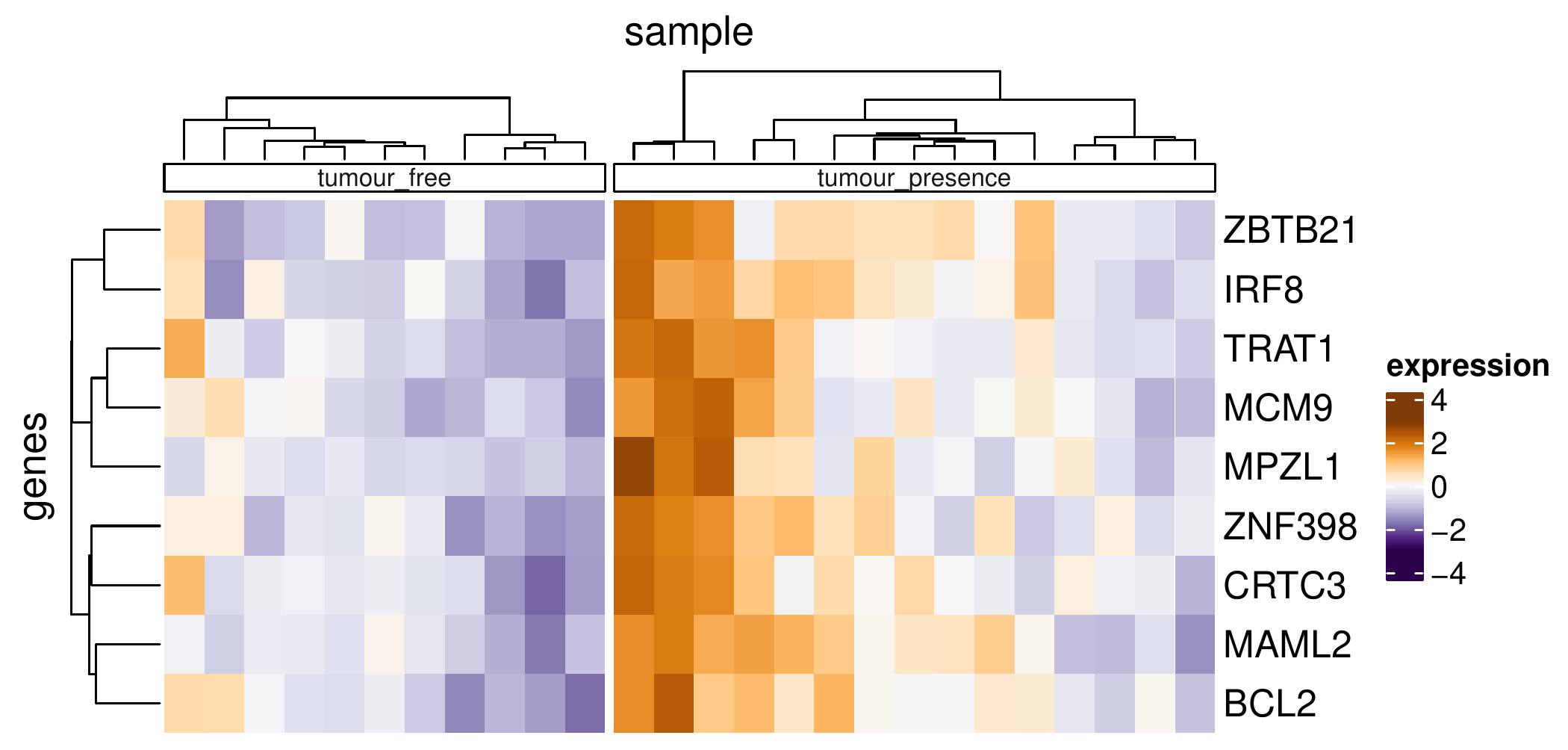
选择 lambda.min 时的诊断模型，使用数据集 GSE207586, GSE244645, GSE68086 验证。

Fig. 为特征基因在验证数据集 (GSE207586) 的表达热图。 Fig. 为特征基因在验证数据集 (GSE244645) 的表达热图。 Fig. 为特征基因在验证数据集 (GSE68086) 的表达热图。 Fig. 为 lambda.min ROC 曲线。



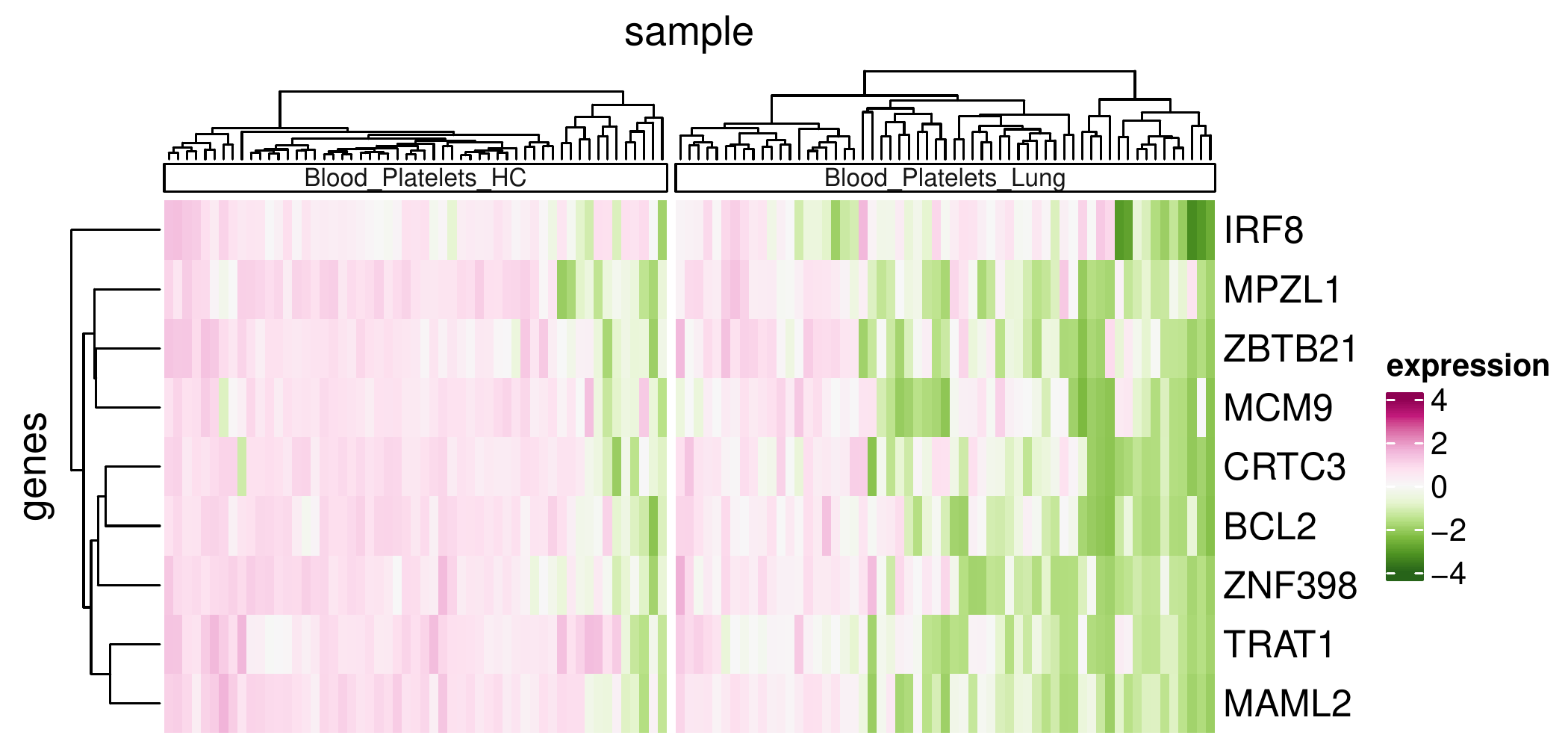
**Fig.** LUNG VALID GSE207586 feature heatmap in validation dataset

**(File path: Figure+Table\_scRNA/3.6.0\_RNA\_数据集\_(Platelet)\_验证\_(LUNG\_VALID)/LUNG-VALID-GSE207586-feature-heatmap-in-validation-dataset.pdf)**



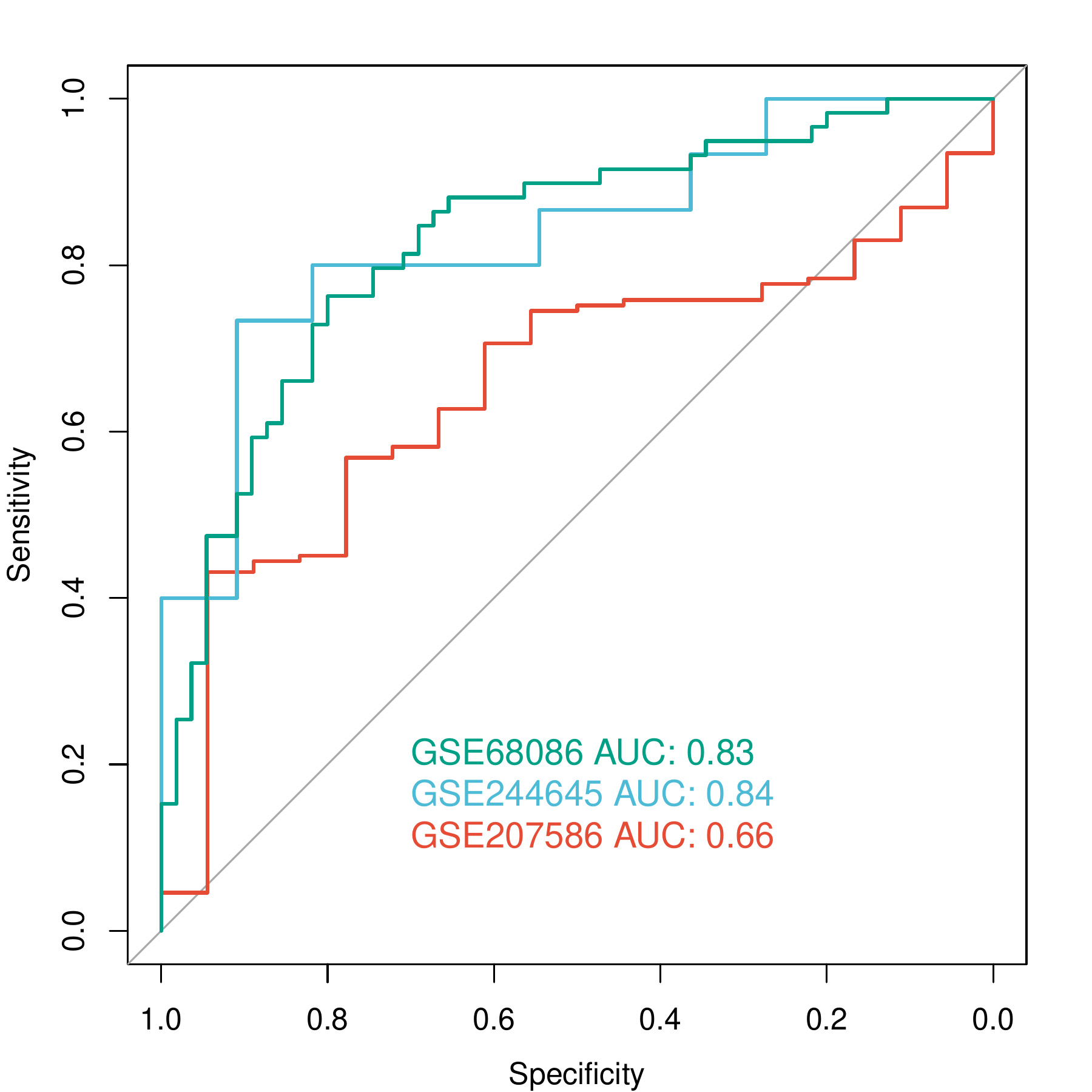
**Fig.** LUNG VALID GSE244645 feature heatmap in validation dataset

**(File path: Figure+Table\_scRNA/3.6.0\_RNA\_数据集\_(Platelet)\_验证\_(LUNG\_VALID)/LUNG-VALID-GSE244645-feature-heatmap-in-validation-dataset.pdf)**



**Fig.** LUNG VALID GSE68086 feature heatmap in validation dataset

**(File path: Figure+Table\_scRNA/3.6.0\_RNA\_数据集\_(Platelet)\_验证\_(LUNG\_VALID)/LUNG-VALID-GSE68086-feature-heatmap-in-validation-dataset.pdf)**



**Fig.** LUNG VALID lambda min ROC

**(File path: Figure+Table\_scRNA/3.6.0\_RNA\_数据集\_(Platelet)\_验证\_(LUNG\_VALID)/LUNG-VALID-lambda-min-ROC.pdf)**

## 3.7 肺组织 RNA-seq 数据集验证

以上模型的构建和验证均选取肺癌 Tumour-educated Platelet 基因表达数据。 以下，尝试将模型推广于肺组织样本基因表达数据 (来自于 GTEx 的正常组织样本和 TCGA 的癌组织样本) 。

### 3.7.1 UCSCXenaTools 癌症相关数据获取 (LUNG\_LUSC)

获取 UCSC Xena 的 TcgaTargetGtex 数据。共 786 个数据。样本组织为：Lung (n=786) 。样本类型为：Normal Tissue (n=288) , Primary Tumor (n=498) 。数据来源为：GTEX (n=288) , TCGA (n=498) 。

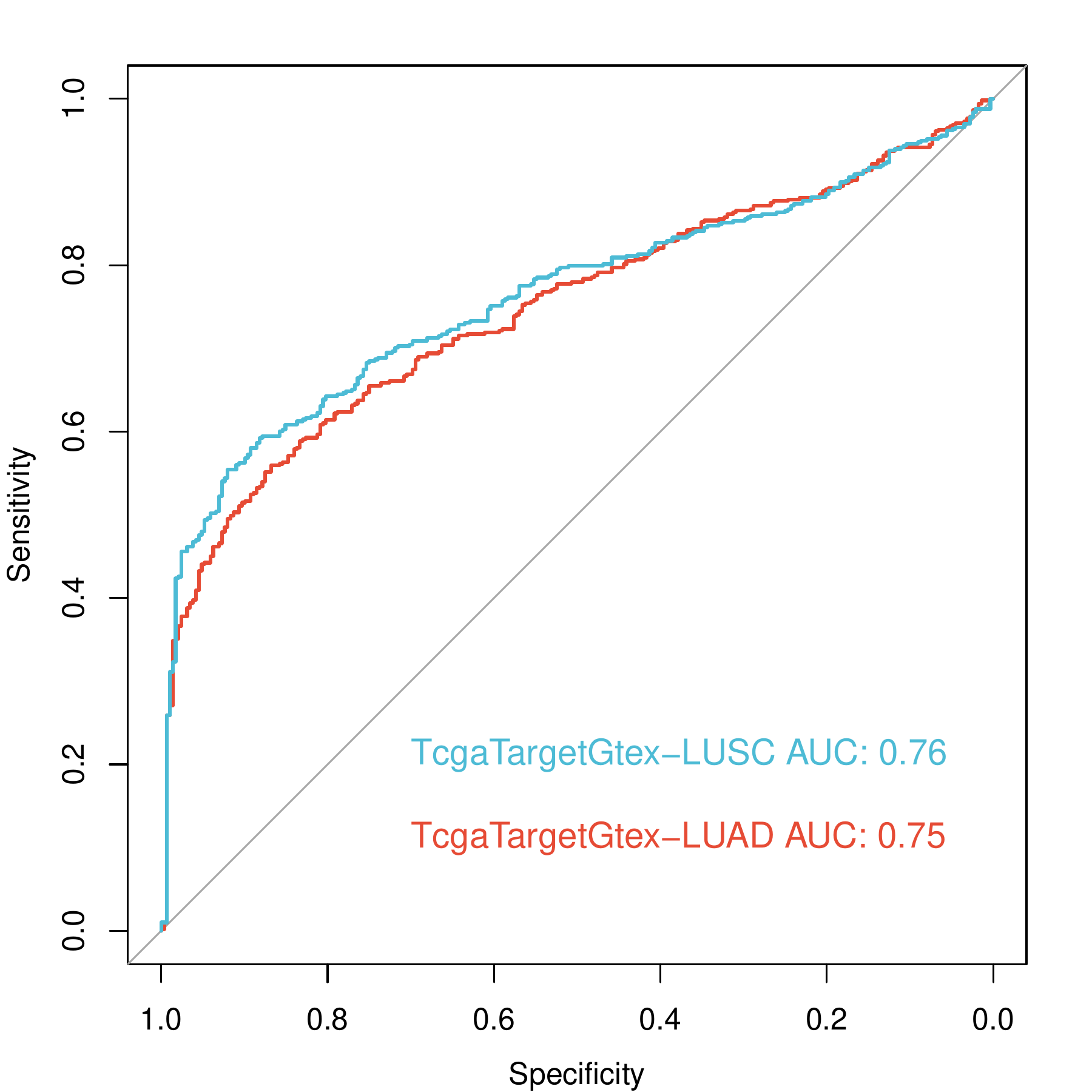
### 3.7.2 UCSCXenaTools 癌症相关数据获取 (LUNG\_LUAD)

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) 。

### 3.7.3 RNA 数据集验证 (TCGA\_VALID)

使用数据集 TcgaTargetGtex-LUAD, TcgaTargetGtex-LUSC 验证。

Fig. 为 lambda.min ROC 曲线。



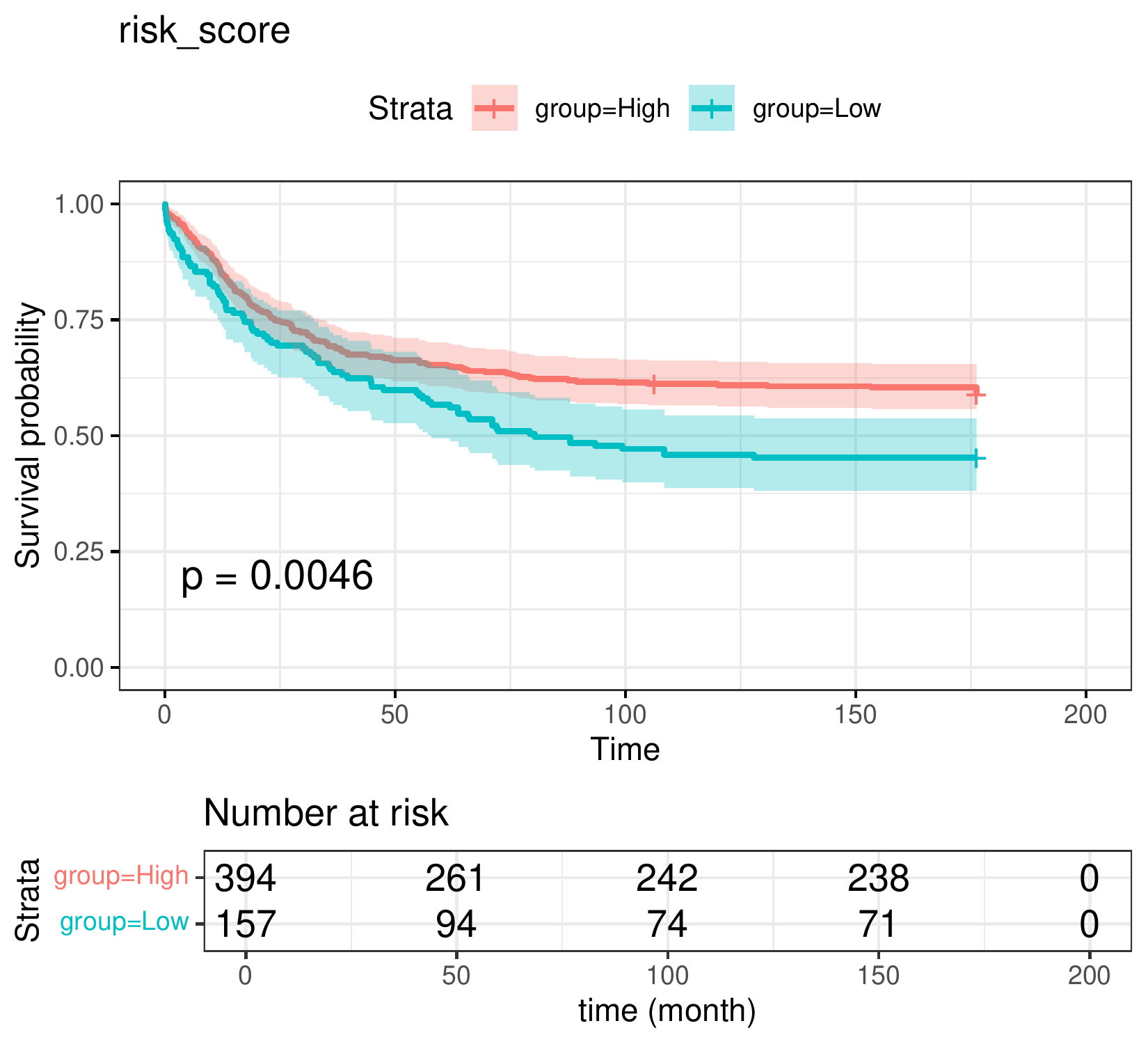
**Fig.** TCGA VALID lambda min ROC

**(File path: Figure+Table\_scRNA/3.7.3\_RNA\_数据集验证\_(TCGA\_VALID)/TCGA-VALID-lambda-min-ROC.pdf)**

## 3.8 Survival 生存分析 (LUSC)

将风险评分模型运用于数据集 (dataset: LUSC)。按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: -0.408749592774634) (High (n=394) , Low (n=157) )， 随后进行生存分析。

Fig. 为 risk\_score 生存曲线。



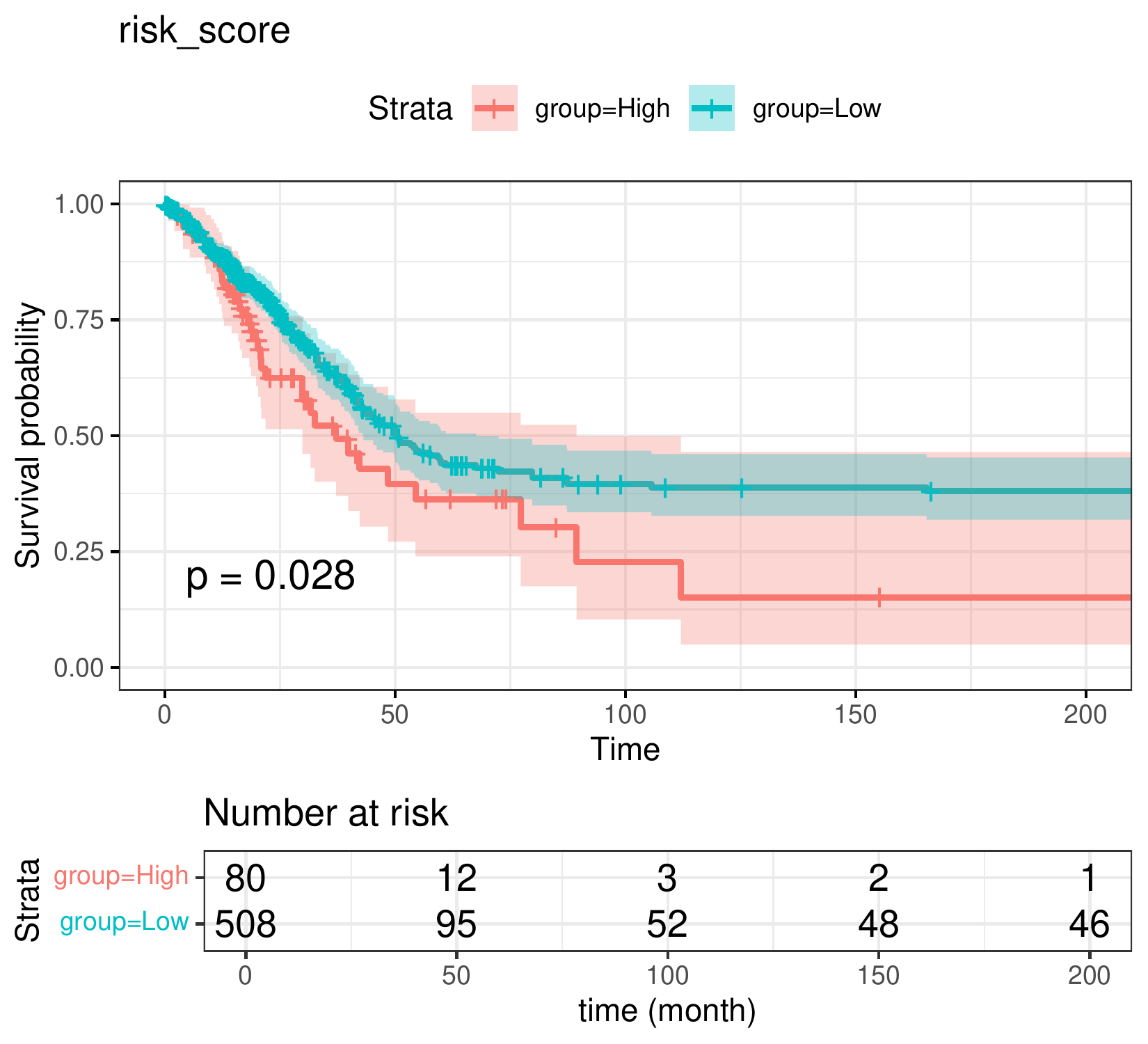
**Fig.** LUSC survival curve of risk score

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

## 3.9 Survival 生存分析 (LUAD)

将风险评分模型运用于数据集 (dataset: LUAD)。按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 0.768015003838708) (High (n=80) , Low (n=508) )， 随后进行生存分析。

Fig. 为 risk\_score 生存曲线。



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

**(File path: Figure+Table\_scRNA/3.9.0\_Survival\_生存分析\_(LUAD)/LUAD-survival-curve-of-risk-score.pdf)**

# 4 总结

本研究从多组 Tumor-educated platelet 数据集 (3 组GEO 和 1 组临床样本RNA-seq) 出发，差异分析， 发现共同 DEGs，随后以机器学习 (Lasso) 建立诊断模型，用以肺癌 (包括 LUAD 和 LUSC) 诊断。 随后又从多个数据集上外部验证了模型的稳定性。本研究还将 Tumor-educated platelet 测序数据，推广 到肺组织样本的基因表达数据的诊断，LUSC 和 LUAD 的结果均表明，该诊断模型具有优异的诊断性能。

# Reference

1. 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).

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

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

4. 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).