**生信分析报告**

**项目标题： 骨肉瘤 ;**

**单 号： BSZD231122 ;**

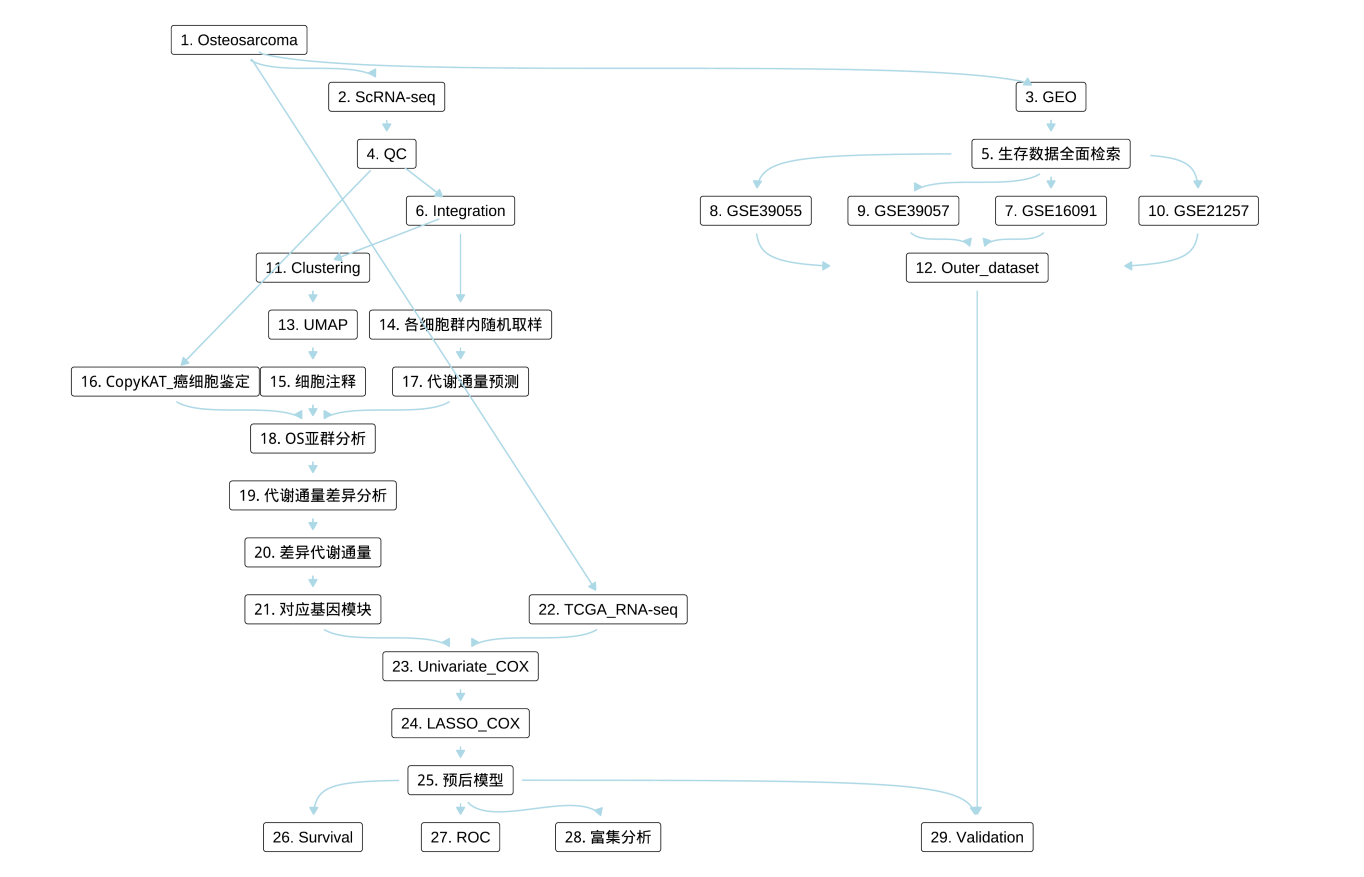
**分析人员： 黄礼闯 ;**

**分析类型： 分析优化 ;**

**委 托 人： 杨立宇 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程



**Fig.** **1** Route

**(File path: Figure+Table/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。 详细见各分析部分。

# 3 分析结果

## 3.1 GEO 数据获取 (OS)

以 R 包 GEOquery (2.74.0) 获取 GSE152048 数据集。

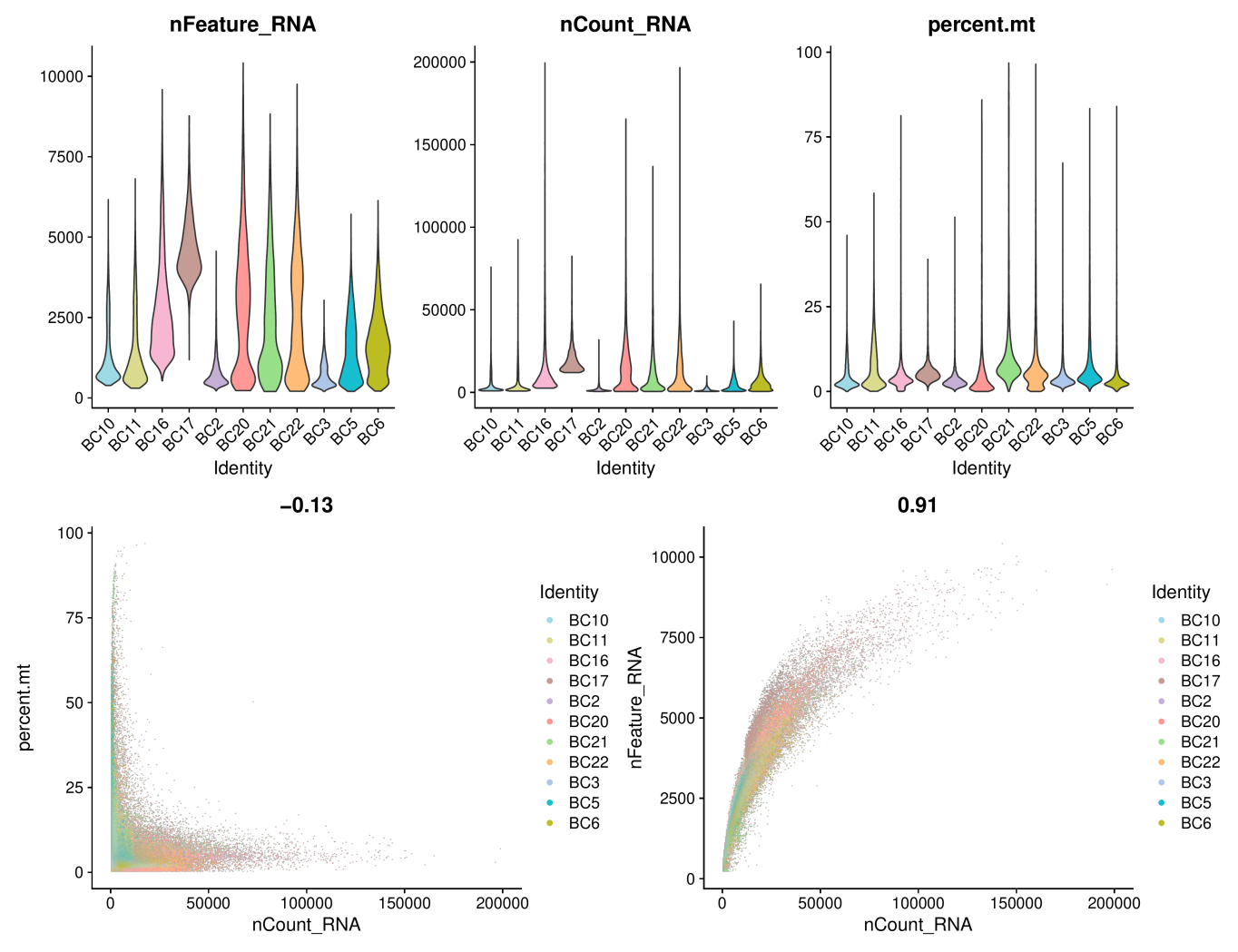
以 GEOquery 获取 GSE152048 的数据信息。

## 3.2 Seurat 集成单细胞数据分析 (OS)

使用 Seurat R 包 (5.1.0) 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 1000 个基因，并且基因数量小于 7000。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 1-10 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 0.01) 为所有细胞群寻找 Markers。以 Python 工具 SCSA ((2020, **IF:2.8**, Q2, Frontiers in genetics)1) (<https://github.com/bioinfo-ibms-pumc/SCSA>) 对细胞群注释。

读取 BC10, BC11, BC16, BC17, BC2, BC20, BC21, BC22, BC3, BC5, BC6 样本的数据集。前期质量控制，一个细胞至少应有 1000 个基因，并且基因数量小于 7000。线粒体基因的比例小于 10%。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 1-10 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。

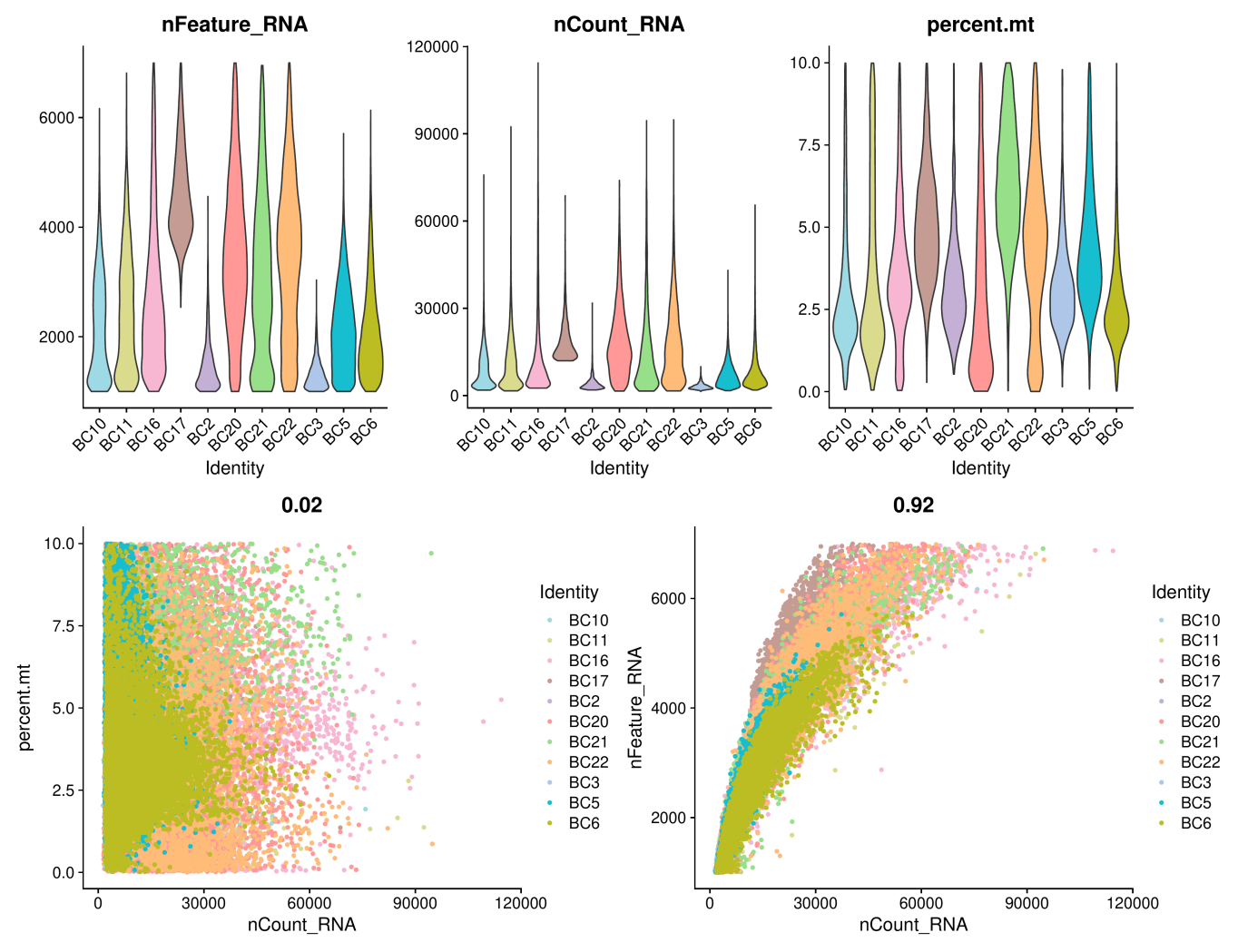
(鉴定所用 Markers 来源于原作文献 <PMID:33303760> (2020, **IF:14.7**, Q1, Nature communications)2)



**Fig.** **2** Pre Quality control

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/Pre-Quality-control.pdf)**

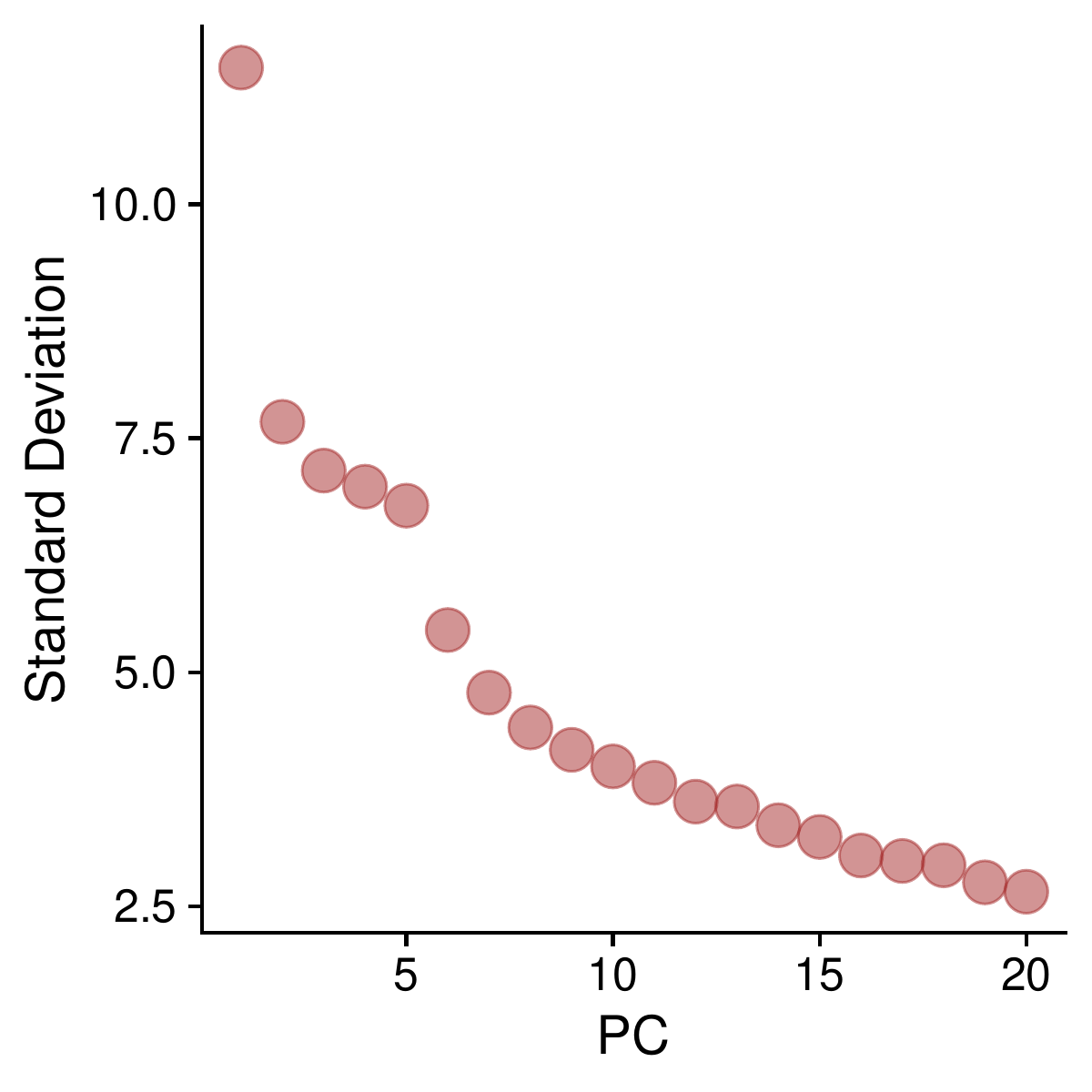
Fig. **[2](#Pre-Quality-control)** 为 QC (质量控制) 图 (数据过滤前) 。



**Fig.** **3** OS After Quality control

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-After-Quality-control.pdf)**

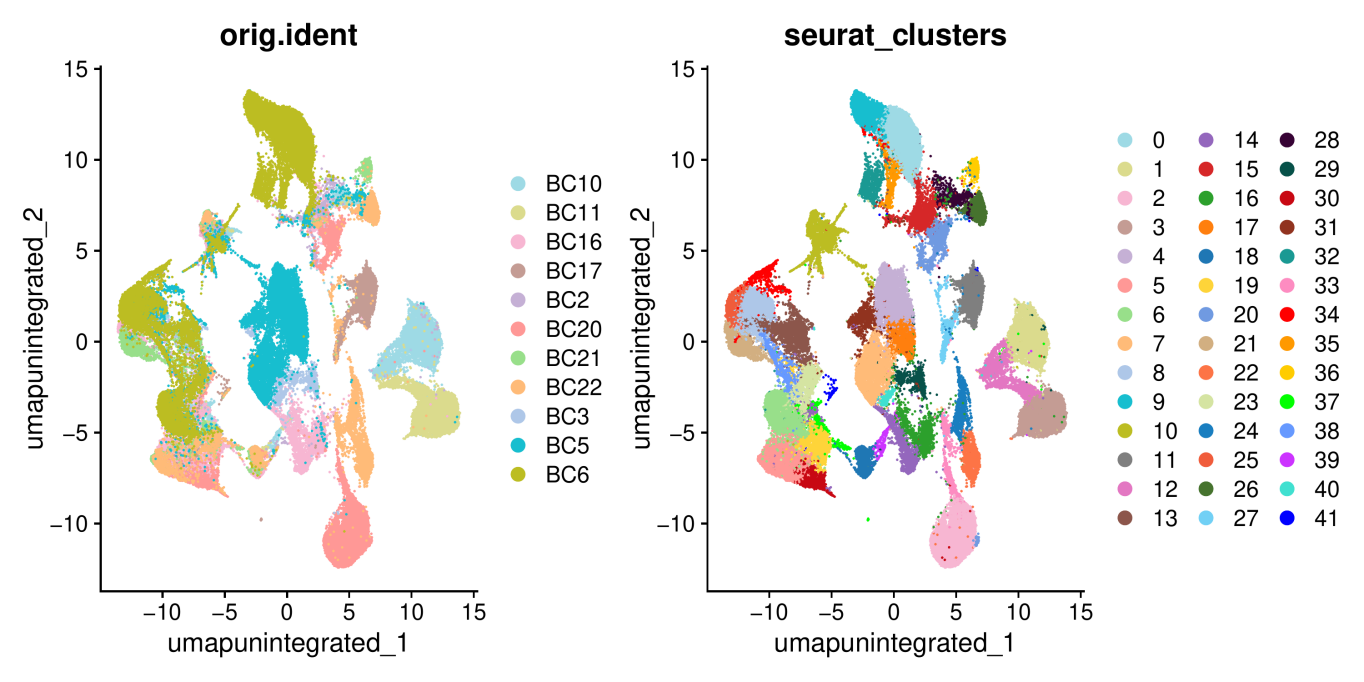
Fig. **[3](#OS-After-Quality-control)** 为数据过滤后的 QC 图。



**Fig.** **4** OS Standard deviations of PCs

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-Standard-deviations-of-PCs.pdf)**

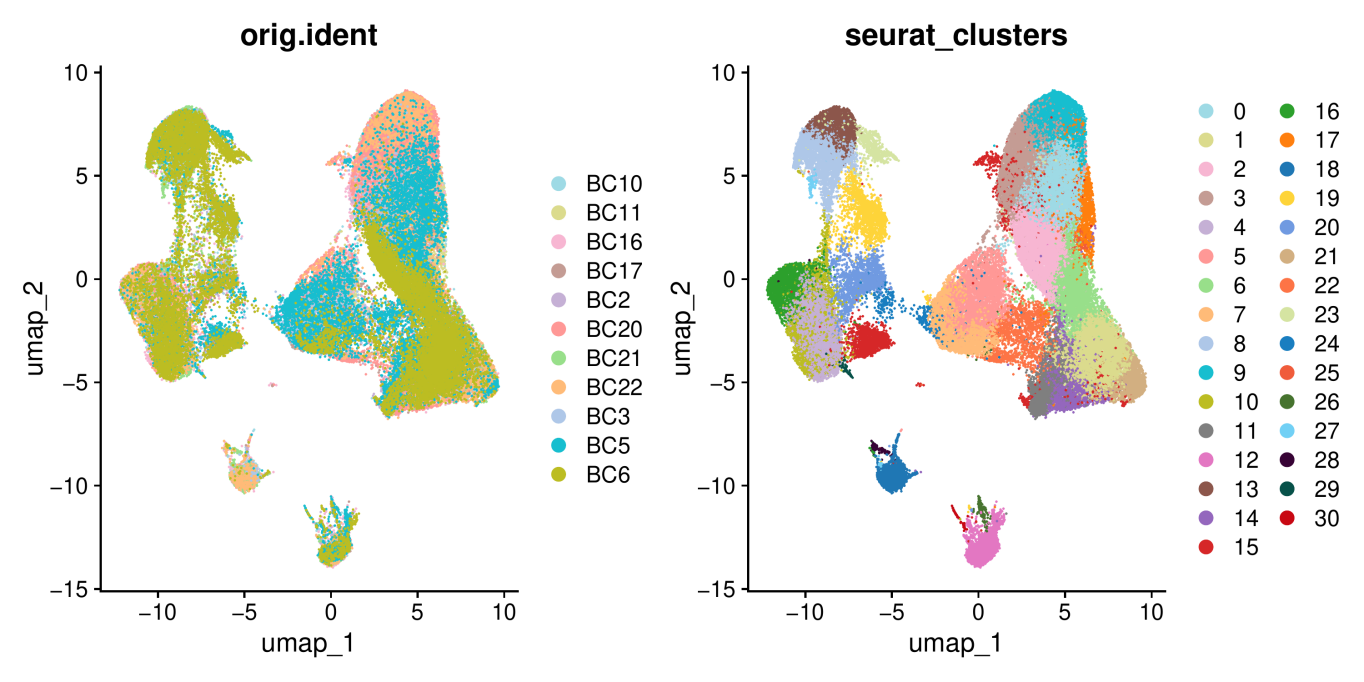
Fig. **[4](#OS-Standard-deviations-of-PCs)** 为主成分 (PC) 的 Standard deviations。



**Fig.** **5** OS UMAP Unintegrated

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-UMAP-Unintegrated.pdf)**

Fig. **[5](#OS-UMAP-Unintegrated)** 为去除批次效应之前的 UMAP 聚类图。



**Fig.** **6** OS UMAP Integrated

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-UMAP-Integrated.pdf)**

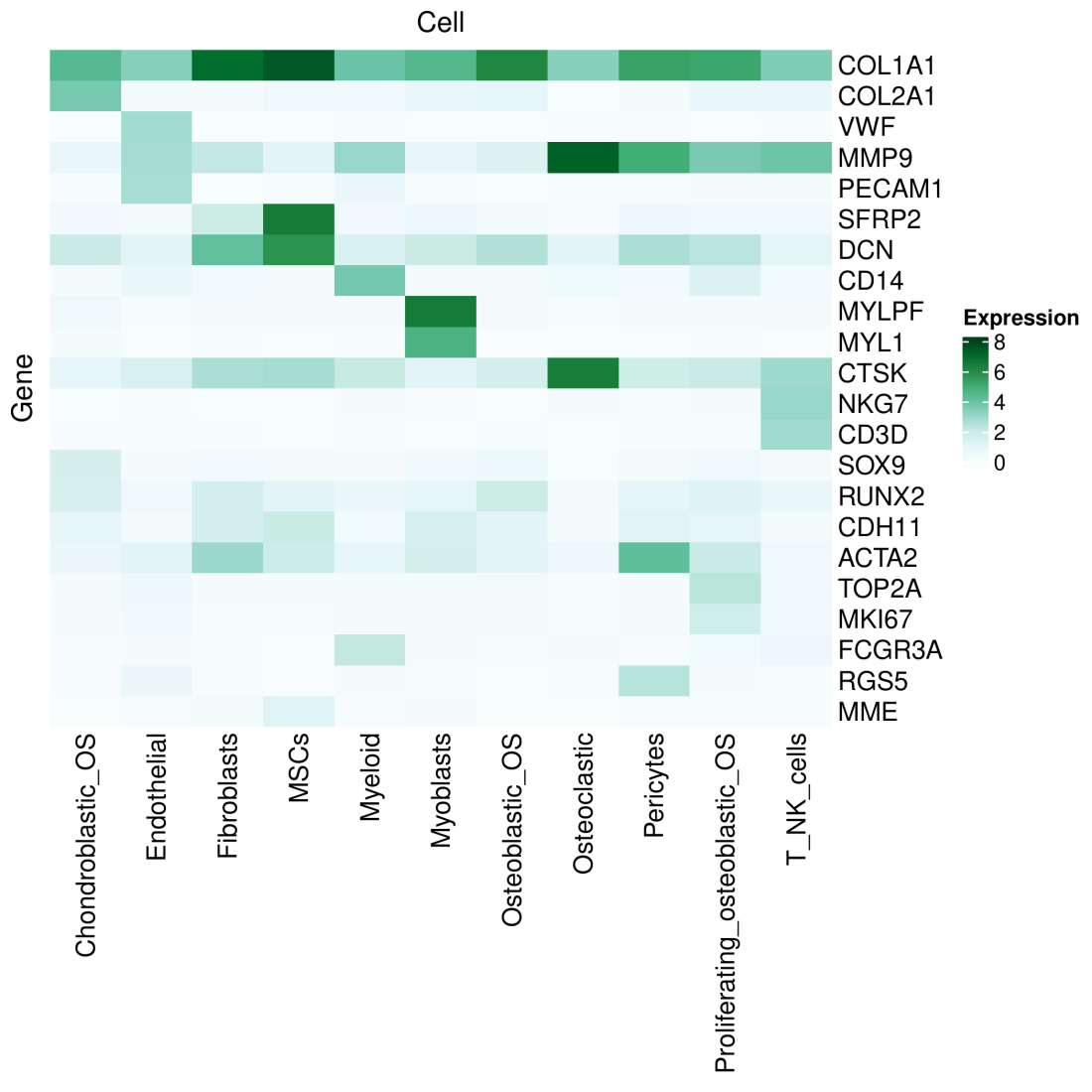
Fig. **[6](#OS-UMAP-Integrated)** 为 去除批次效应之后的 UMAP 聚类图。

**Tab.** **1** OS significant markers of cell clusters

| Rownames | P val | Avg log2FC | Pct.1 | Pct.2 |
| --- | --- | --- | --- | --- |
| ALPL | 0 | 1.494 | 0.915 | 0.412 |
| PANX3 | 0 | 2.181 | 0.645 | 0.156 |
| IFITM5 | 0 | 2.458 | 0.802 | 0.323 |
| LY6K | 0 | 1.191 | 0.77 | 0.314 |
| RHBDL2 | 0 | 1.733 | 0.644 | 0.21 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-significant-markers-of-cell-clusters.csv)**

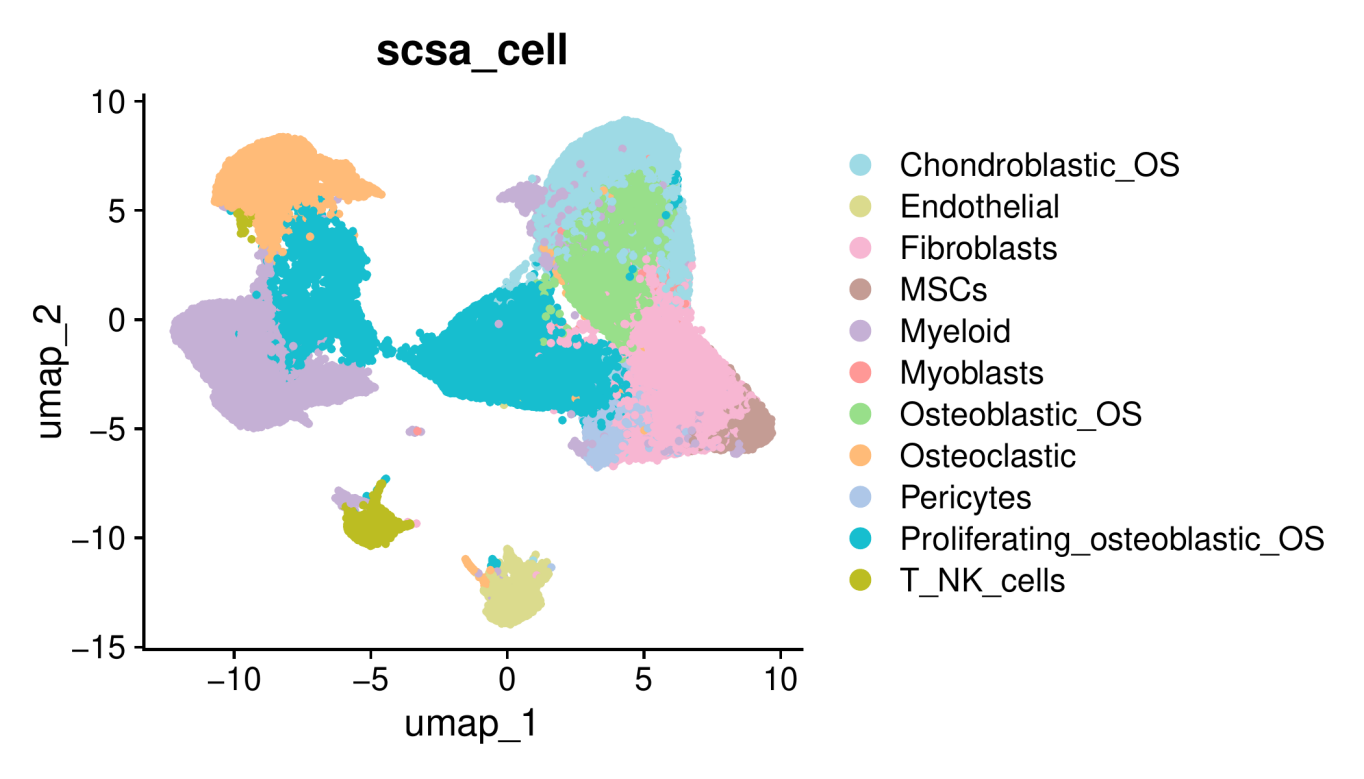
Tab. **[1](#OS-significant-markers-of-cell-clusters)** 为所有细胞群的 Marker (LogFC 阈值 0.25; 最小检出率 0.01; 矫正 P 值阈值 0.05)



**Fig.** **7** OS Marker Validation

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-Marker-Validation.pdf)**

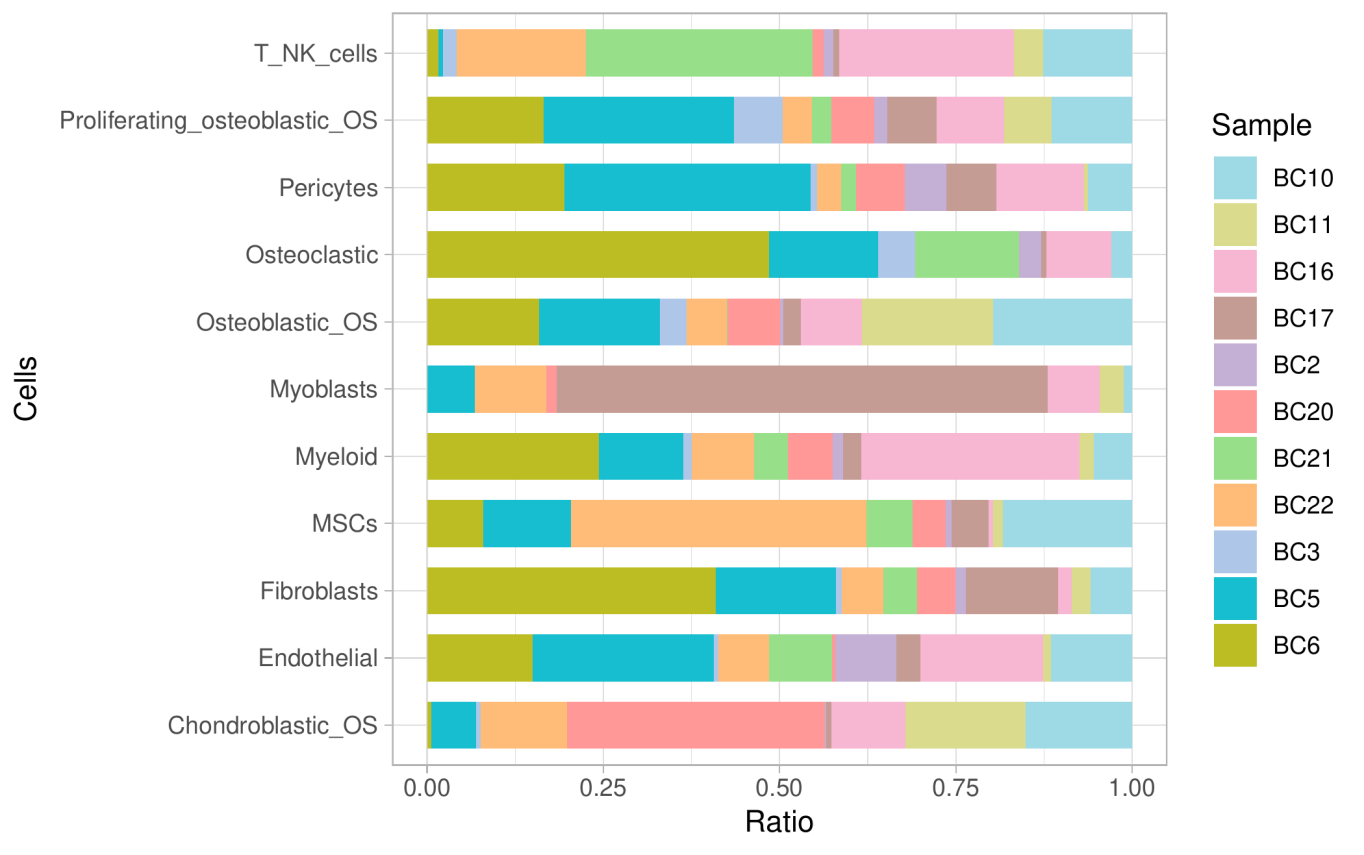
Fig. **[7](#OS-Marker-Validation)** 使用特异性 Marker 对细胞注释结果的验证热图。



**Fig.** **8** OS SCSA Cell type annotation

**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-SCSA-Cell-type-annotation.pdf)**

Fig. **[8](#OS-SCSA-Cell-type-annotation)** 为 SCSA 细胞注释结果的 UMAP 图。



**Fig.** **9** OS SCSA Cell Proportions in each sample

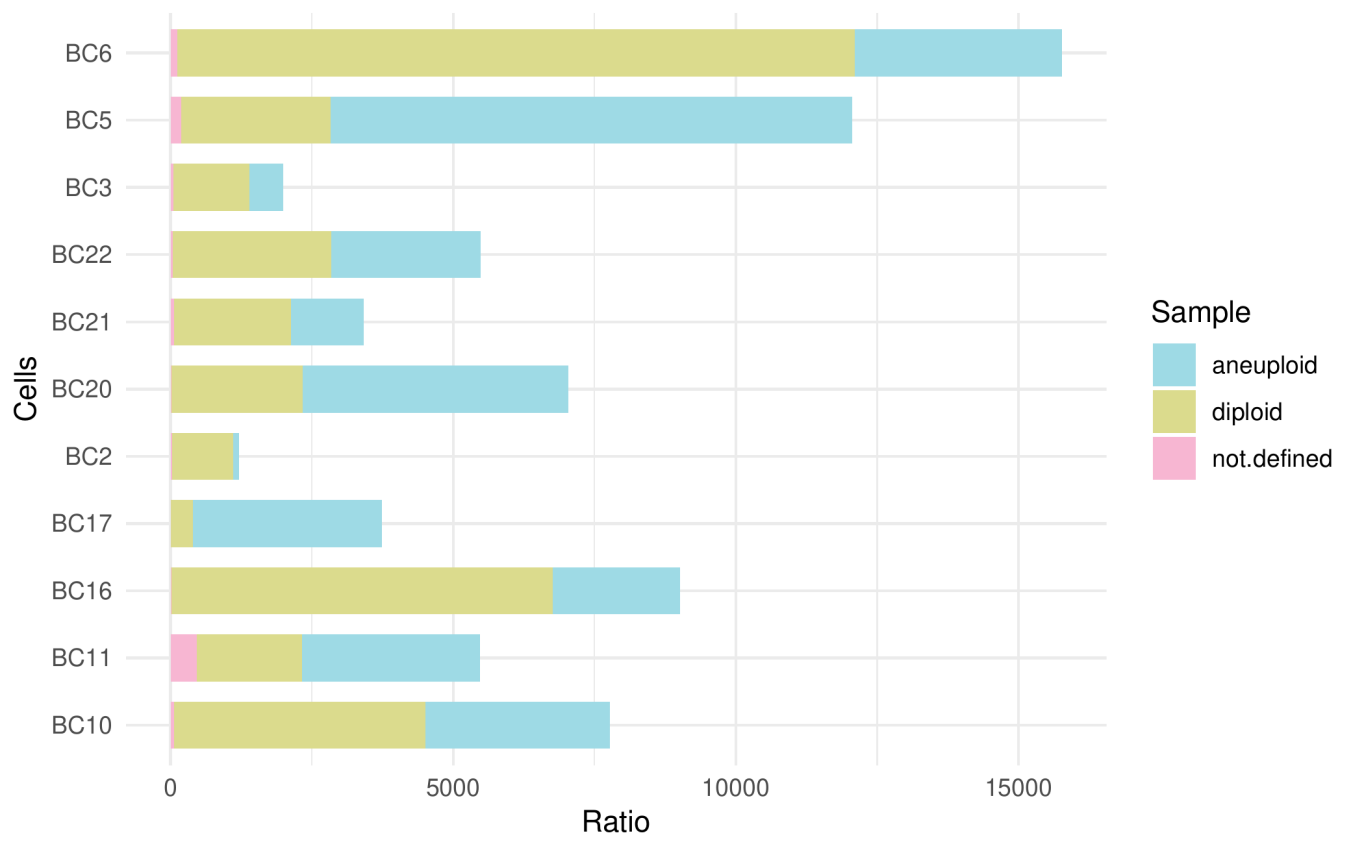
**(File path: Figure+Table/3.2.0\_Seurat\_集成单细胞数据分析\_(OS)/OS-SCSA-Cell-Proportions-in-each-sample.pdf)**

Fig. **[9](#OS-SCSA-Cell-Proportions-in-each-sample)** 为 SCSA 注释的细胞群在各个样本中的占比。

## 3.3 CopyKAT 癌细胞鉴定 (OS)

R 包 CopyKAT 用于鉴定恶性细胞 (2021, **IF:33.1**, Q1, Nature Biotechnology)3。CopyKAT 可以区分整倍体与非整倍体，其中非整倍体被认为是肿瘤细胞，而整倍体是正常细胞 (2012, **IF:39.1**, Q1, Nature Reviews Genetics)4。由于 CopyKAT 不适用于多样本数据 (批次效应的存在) ，因此，对各个样本独立鉴定。

以 CopyKAT 鉴定恶质细胞。



**Fig.** **10** OS proportions of aneuploid and diploid

**(File path: Figure+Table/3.3.0\_CopyKAT\_癌细胞鉴定\_(OS)/OS-proportions-of-aneuploid-and-diploid.pdf)**

Fig. **[10](#OS-proportions-of-aneuploid-and-diploid)** 为 copyKAT 注释的所有样本中 aneuploid (Malignant cell) 与 diploid (Benign cell) 的细胞比例。

Note: The directory 'Figure+Table/OS-all-malignant-cells-heatmap' contains 11 files.  
  
1 BC10.png  
2 BC11.png  
3 BC16.png  
4 BC17.png  
5 BC2.png  
6 ...

**(File path: Figure+Table/3.3.0\_CopyKAT\_癌细胞鉴定\_(OS)/OS-all-malignant-cells-heatmap)**

**Tab.** **2** OS all copyKAT prediction data

| Orig.ident | Cell.names | Copykat.pred | Copykat cell |
| --- | --- | --- | --- |
| BC10 | AAACCTGAGACTCGGA-1 1 | Aneuploid | Cancer cell |
| BC10 | AAACCTGAGGAACTGC-1 1 | Diploid | Normal cell |
| BC10 | AAACCTGAGGATGGAA-1 1 | Diploid | Normal cell |
| BC10 | AAACCTGAGGTGCTTT-1 1 | Diploid | Normal cell |
| BC10 | AAACCTGAGTAGCGGT-1 1 | Diploid | Normal cell |
| ... | ... | ... | ... |

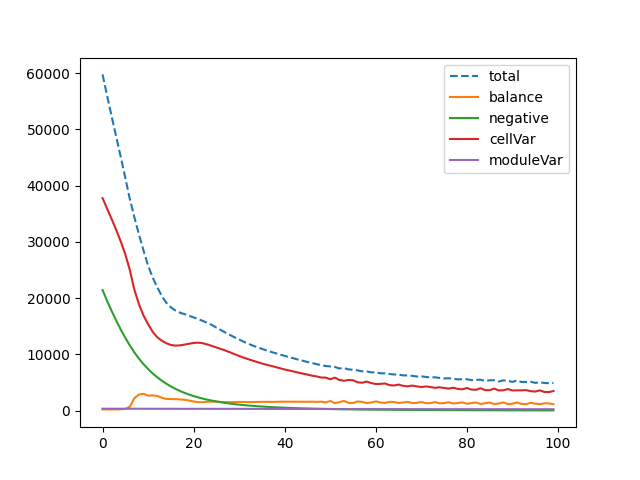
**(File path: Figure+Table/3.3.0\_CopyKAT\_癌细胞鉴定\_(OS)/OS-all-copyKAT-prediction-data.csv)**

Tab. **[2](#OS-all-copyKAT-prediction-data)** 为 copyKAT 所有样本注释结果附表。

## 3.4 scFEA 单细胞数据的代谢通量预测 (OS\_SAMPLE)

根据样本和细胞类型分组，将细胞随机抽样 (各组比例为：0.5) (细胞数量较多，通过随机抽样的方式减少计算负担) (随机种子：987456)。 将 Seurat 的 RNA Assay (‘counts’) 作为输入数据，以 scFEA 预测细胞的代谢通量 (2021, **IF:6.2**, Q1, Genome research)5。参考 <https://github.com/changwn/scFEA/blob/master/scFEA_tutorial1.ipynb> 和 <https://github.com/changwn/scFEA/blob/master/scFEA_tutorial2.ipynb>。

将 Seurat (所有细胞) 以 scFEA 预测代谢通量。



**Fig.** **11** OS SAMPLE Convergency of the loss terms during training

**(File path: Figure+Table/3.4.0\_scFEA\_单细胞数据的代谢通量预测\_(OS\_SAMPLE)/OS-SAMPLE-Convergency-of-the-loss-terms-during-training.png)**

Fig. **[11](#OS-SAMPLE-Convergency-of-the-loss-terms-during-training)** 为 scFEA 训练过程的收敛曲线。

**Tab.** **3** OS SAMPLE annotation of metabolic flux

| V1 | Module id | Compound IN name | Compound IN ID | Compound OUT name |
| --- | --- | --- | --- | --- |
| M 1 | 1 | Glucose | C00267 | G6P |
| M 2 | 2 | G6P | C00668 | G3P |
| M 3 | 3 | G3P | C00118 | 3PD |
| M 4 | 4 | 3PD | C00197 | Pyruvate |
| M 5 | 5 | Pyruvate | C00022 | Acetyl-Coa |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.4.0\_scFEA\_单细胞数据的代谢通量预测\_(OS\_SAMPLE)/OS-SAMPLE-annotation-of-metabolic-flux.xlsx)**

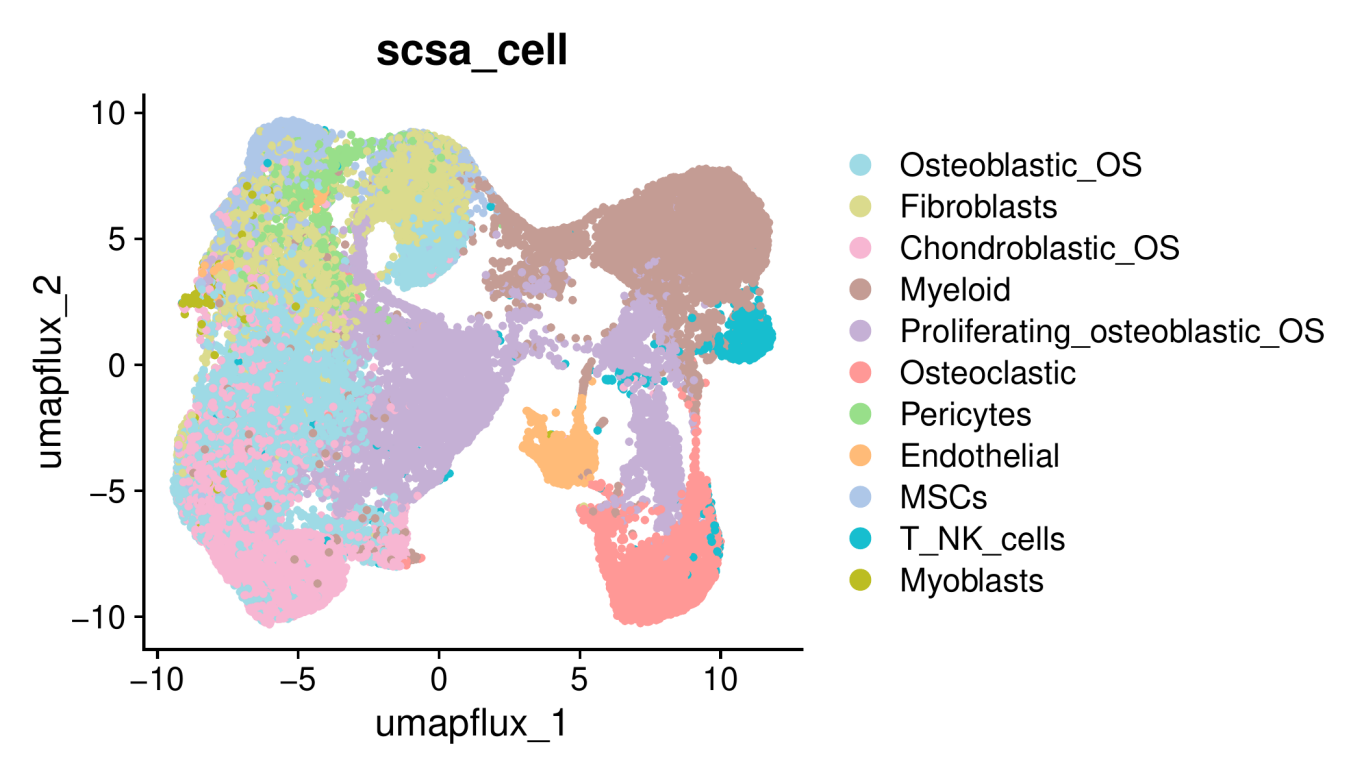
Tab. **[3](#OS-SAMPLE-annotation-of-metabolic-flux)** 各代谢模块的注释。

**Tab.** **4** OS SAMPLE metabolic flux matrix

| V1 | M 1 | M 2 | M 3 | M 4 |
| --- | --- | --- | --- | --- |
| AAACCTGAGGATGGAA-1 1 | 0.01224 | 0.01727 | 0.0649 | 0.1126 |
| AAACCTGCACAACGCC-1 1 | 0.01224 | 0.02175 | 0.06084 | 0.07629 |
| AAACCTGTCATCATTC-1 1 | 0.01618 | 0.02072 | 0.02948 | 0.03468 |
| AAACCTGTCGTCCGTT-1 1 | 0.01832 | 0.07996 | 0.1203 | 0.2113 |
| AAACCTGTCTTGCATT-1 1 | 0.01683 | 0.06143 | 0.09314 | 0.1331 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.4.0\_scFEA\_单细胞数据的代谢通量预测\_(OS\_SAMPLE)/OS-SAMPLE-metabolic-flux-matrix.csv)**

Tab. **[4](#OS-SAMPLE-metabolic-flux-matrix)** 为细胞代谢通量矩阵 (各 M\_ 为代谢模块)。



**Fig.** **12** OS SAMPLE cells metabolic flux

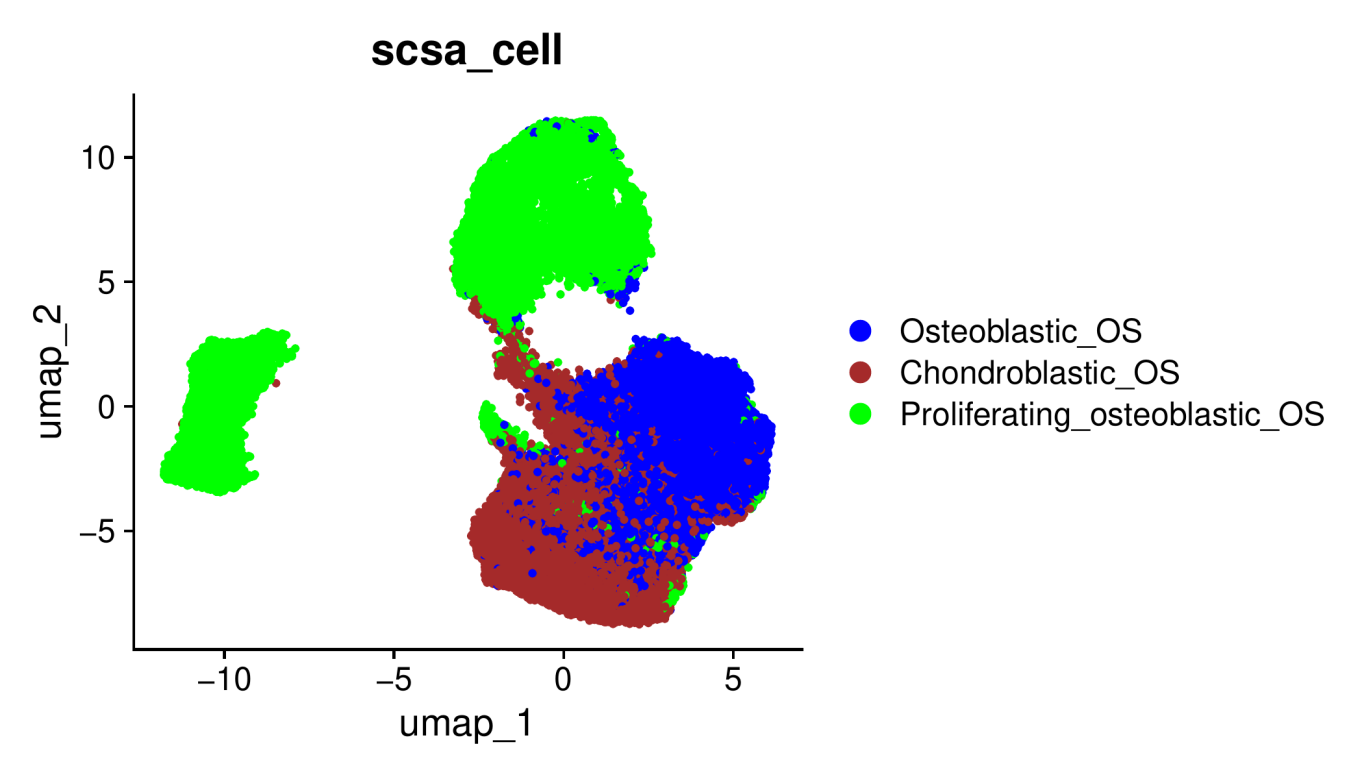
**(File path: Figure+Table/3.4.0\_scFEA\_单细胞数据的代谢通量预测\_(OS\_SAMPLE)/OS-SAMPLE-cells-metabolic-flux.pdf)**

Fig. **[12](#OS-SAMPLE-cells-metabolic-flux)** 为细胞代谢通量 (scFEA 预测，输入 Seurat) 的 UMAP 聚类。

## 3.5 Seurat 细胞亚群分析 (OS\_CANCER)

成骨细胞和软骨细胞骨肉瘤是临床上常见的两种主要骨肉瘤类型(2020, **IF:14.7**, Q1, Nature communications)2。 在这里，聚焦于注释结果中的 Proliferating\_osteoblastic\_OS, Chondroblastic\_OS, Osteoblastic\_OS 细胞，重新聚类分析。 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。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。

匹配 scsa\_cell 中包含"\_OS$"的描述，最终得到 34230 例数据。分析其亚群。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。在 1-15 PC 维度，1.2 分辨率下，对细胞群 UMAP 聚类。



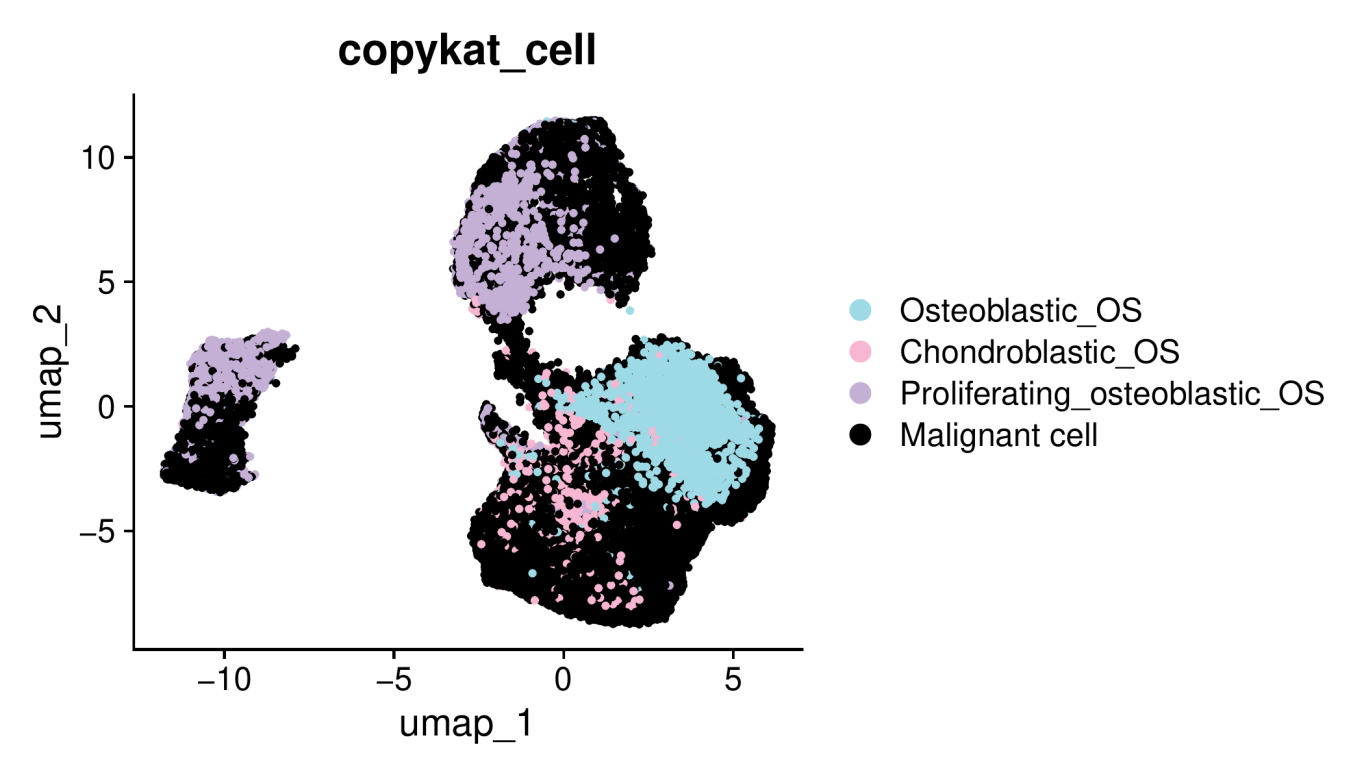
**Fig.** **13** OS CANCER The scsa cell

**(File path: Figure+Table/3.5.0\_Seurat\_细胞亚群分析\_(OS\_CANCER)/OS-CANCER-The-scsa-cell.pdf)**

Fig. **[13](#OS-CANCER-The-scsa-cell)** 为 scsa\_cell 的 umap 聚类图。

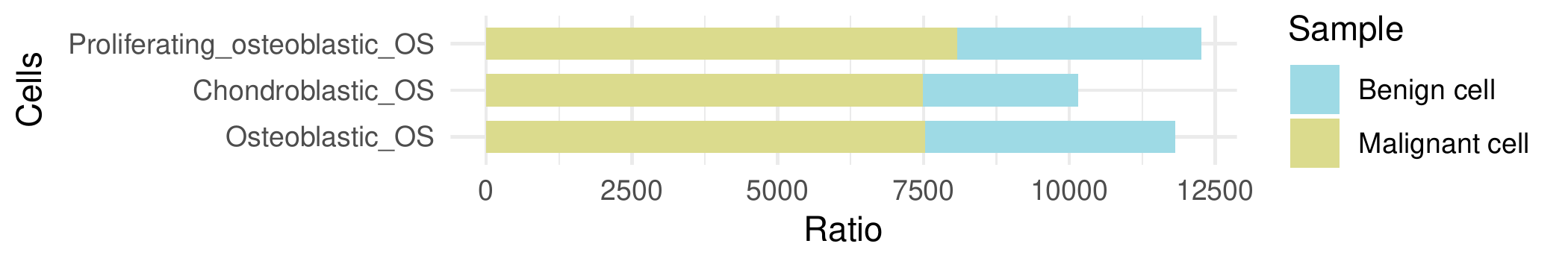
### 3.5.1 Seurat-copyKAT 癌细胞注释 (OS\_CANCER)

将 CopyKAT 的预测结果映射细胞注释中。



**Fig.** **14** OS CANCER Cancer Cell type annotation

**(File path: Figure+Table/3.5.1\_Seurat-copyKAT\_癌细胞注释\_(OS\_CANCER)/OS-CANCER-Cancer-Cell-type-annotation.pdf)**



**Fig.** **15** OS CANCER cancer cell proportions

**(File path: Figure+Table/3.5.1\_Seurat-copyKAT\_癌细胞注释\_(OS\_CANCER)/OS-CANCER-cancer-cell-proportions.pdf)**

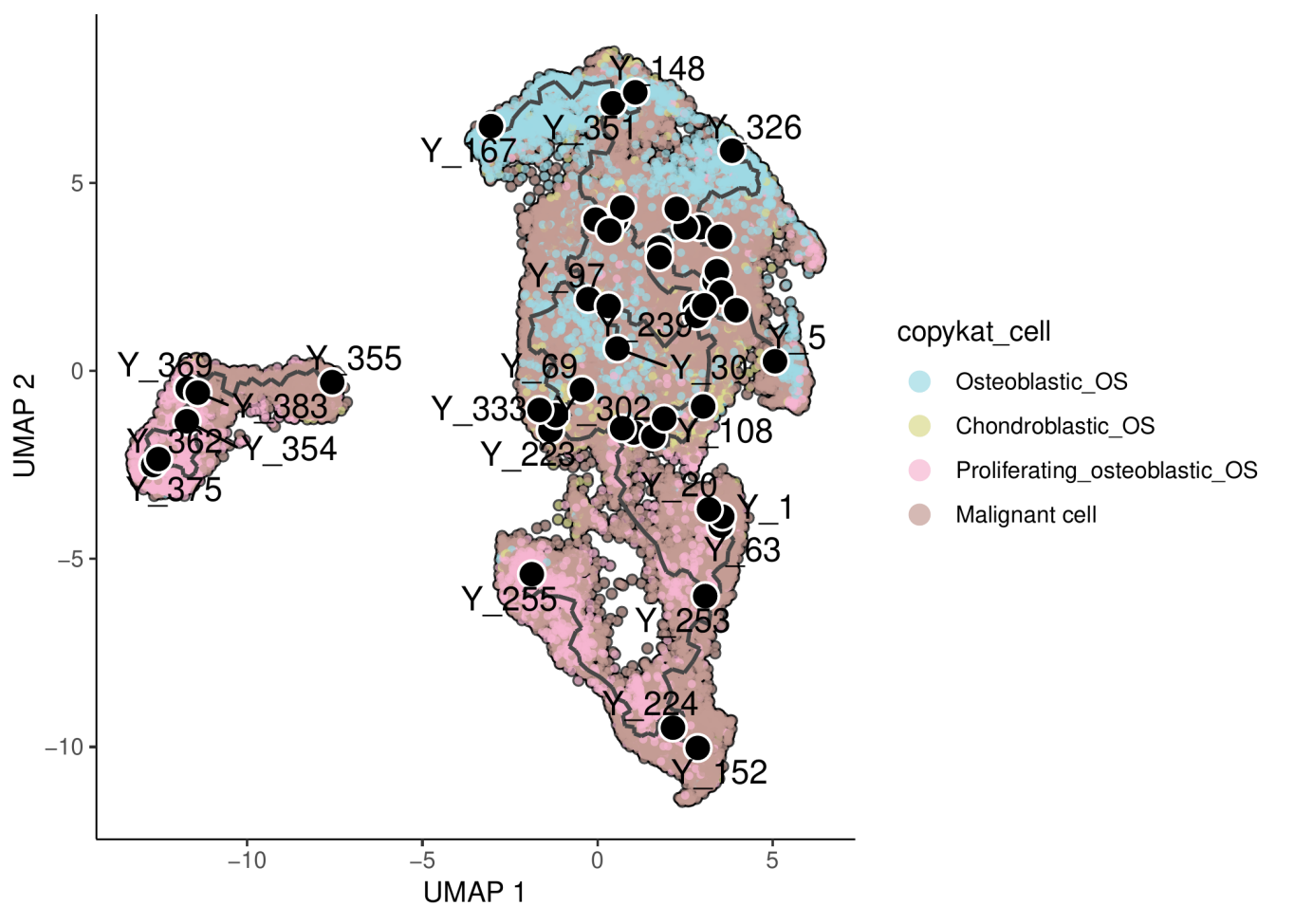
Fig. **[15](#OS-CANCER-cancer-cell-proportions)** 为 copyKAT 注释的恶质细胞在各个细胞类型中的占比。

### 3.5.2 Monocle3 拟时分析 (OS\_CANCER)

从 Seurat 数据对象 scsa\_cell 中提取 OS 类型的细胞，对其重新聚类分析。使用 Seurat::SCTransform 对数据集归一化。随后 PCA 聚类 (RunPCA)。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。使用 SeuratWrappers (SeuratWrappers::as.cell\_data\_set, 参考 <http://htmlpreview.github.io/?https://github.com/satijalab/seurat-wrappers/blob/master/docs/monocle3.html>) 将 Seurat 转化为 Monocle3 的 cell\_data\_set 数据。该转化将继承 Seurat 前期分析的 PCA、UMAP 等聚类结果，用于 Monocle3 的拟时分析，使前后分析一致。以 monocle3::cluster\_cells 计算细胞群的 ‘clusters’ 和 ‘partitions’。以 monocle3::learn\_graph 从高维空间 (high-dimensional space) 中构建 ‘trajectory’。选择 Y\_375, Y\_255, Y\_167 (principle points) 为拟时起点，以 monocle3::order\_cells 将细胞排序，随后构建细胞拟时变化图。以 monocle3::graph\_test 寻找单细胞拟时轨迹中差异表达的基因。

将 Seurat 数据对象转化为 Monocle3 数据对象 (详见方法章节)。构建轨迹图 (Trajectory)。选择 Y\_375, Y\_255, Y\_167 (principle points) 为拟时起点。构建细胞的拟时变化图 (Pseudotime)。寻找单细胞拟时轨迹 (monocle3::graph\_test) 中差异表达的基因。探究 **基因集** (TOP2A, VAMP8, CENPF, …[n = 20], 来自于Monocle3 拟时分析 Graph Test Top Significant genes[Section: OS\_CANCER]) 在拟时轨迹中的表达变化。这些基因中，共 20 属于 Moran’s I test (Graph Test) 差异表达基因。仅分析这些基因。

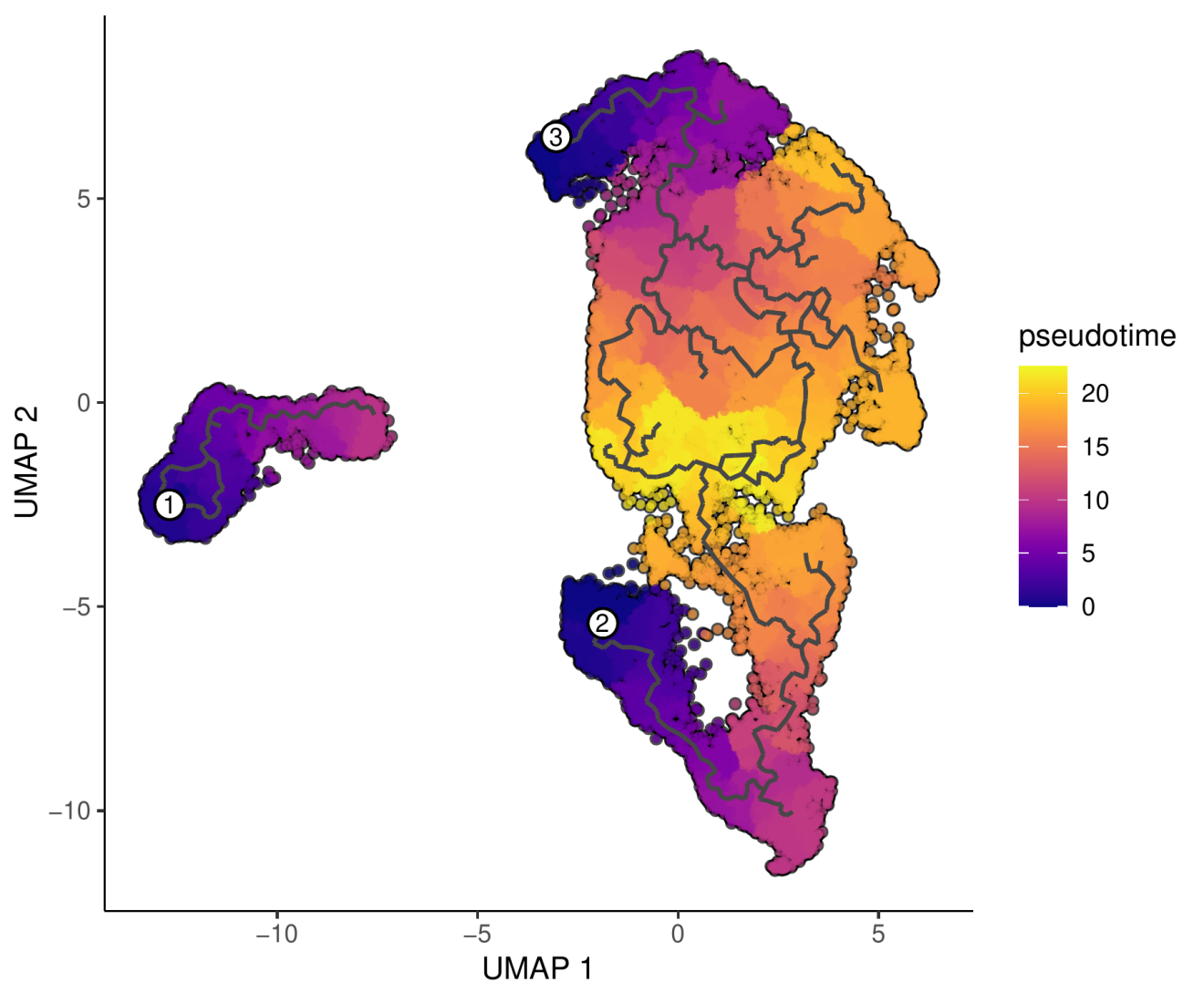
(依据正常细胞向癌细胞转变选择拟时起点)



**Fig.** **16** OS CANCER principal points

**(File path: Figure+Table/3.5.2\_Monocle3\_拟时分析\_(OS\_CANCER)/OS-CANCER-principal-points.pdf)**

Fig. **[16](#OS-CANCER-principal-points)** 为拟时轨迹与 principal point 示意。



**Fig.** **17** OS CANCER pseudotime

**(File path: Figure+Table/3.5.2\_Monocle3\_拟时分析\_(OS\_CANCER)/OS-CANCER-pseudotime.pdf)**

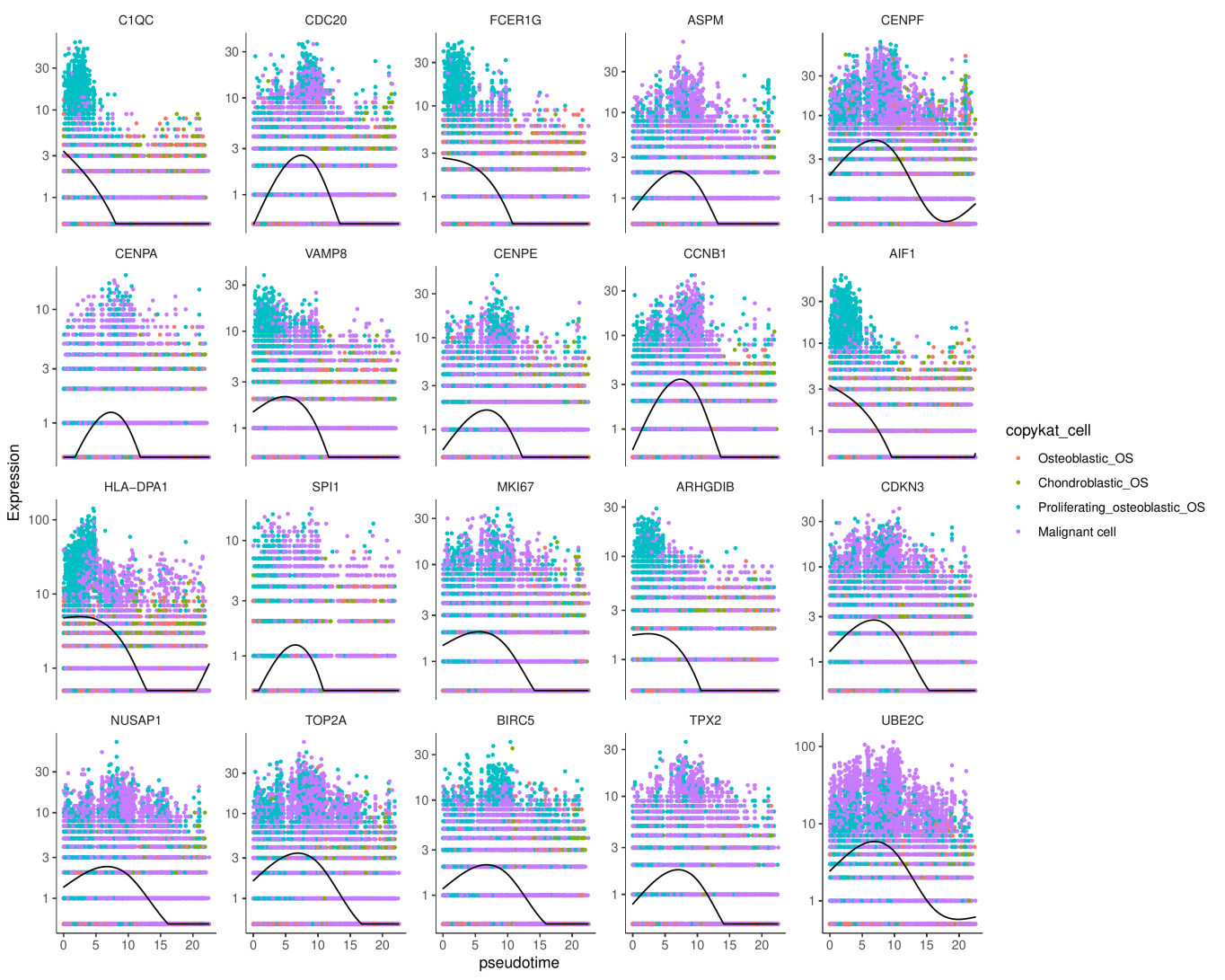
Fig. **[17](#OS-CANCER-pseudotime)** 为细胞拟时图。

**Tab.** **5** OS CANCER Graph Test Significant genes

| Gene id | Status | P value | Morans test stati... | Morans I |
| --- | --- | --- | --- | --- |
| TOP2A | OK | 0 | 409.3 | 0.6054 |
| VAMP8 | OK | 0 | 394.1 | 0.5829 |
| CENPF | OK | 0 | 392.9 | 0.5811 |
| NUSAP1 | OK | 0 | 384.1 | 0.568 |
| CDC20 | OK | 0 | 379.1 | 0.5607 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.5.2\_Monocle3\_拟时分析\_(OS\_CANCER)/OS-CANCER-Graph-Test-Significant-genes.csv)**

Tab. **[5](#OS-CANCER-Graph-Test-Significant-genes)** 为 Moran’s I test (Graph Test) 筛选的差异表达基因 (Q-value cutoff: 0.05)。



**Fig.** **18** OS CANCER Set1 genes in pseudotime

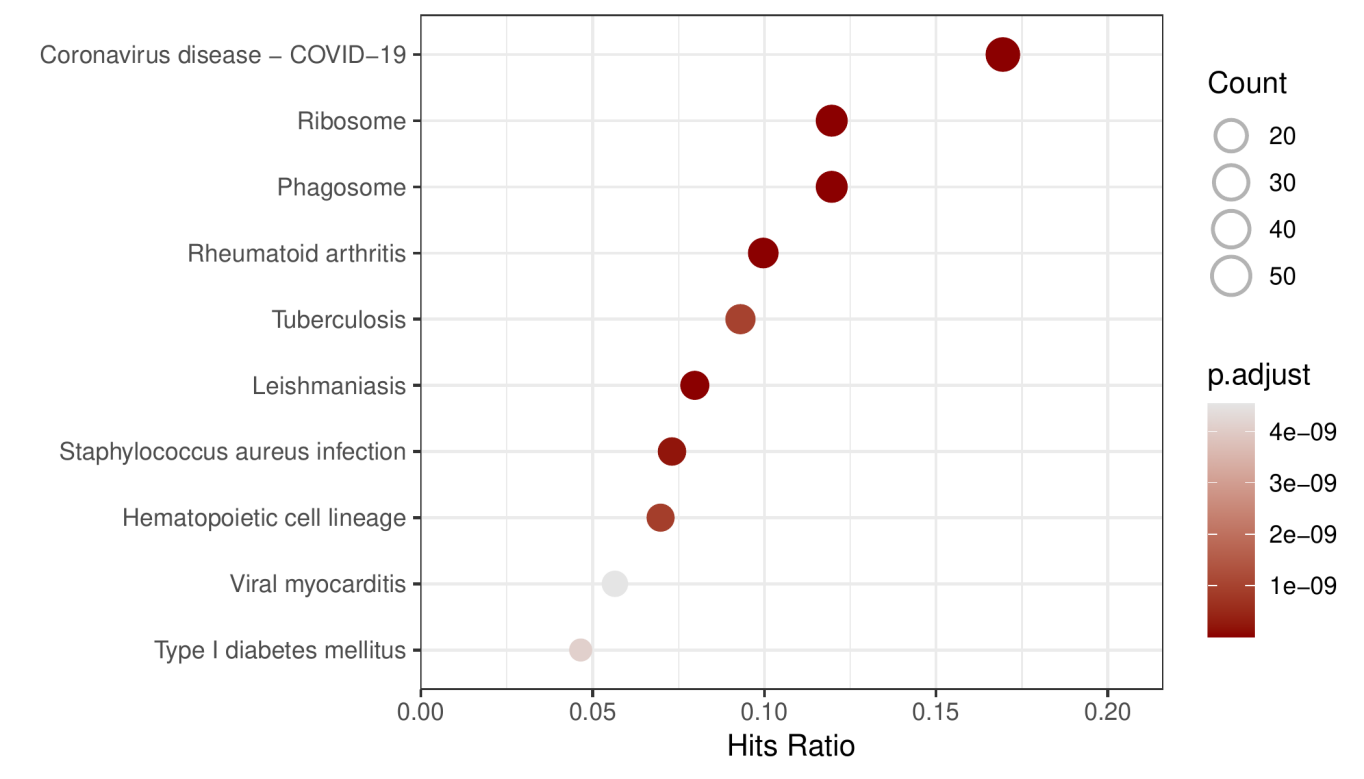
**(File path: Figure+Table/3.5.2\_Monocle3\_拟时分析\_(OS\_CANCER)/OS-CANCER-Set1-genes-in-pseudotime.pdf)**

Fig. **[18](#OS-CANCER-Set1-genes-in-pseudotime)** 为 TOP2A, VAMP8, CENPF, …(n = 20) 在拟时 (Pseudotime) 中的表达变化。

### 3.5.3 ClusterProfiler 富集分析 (OS\_CANCER)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)6进行 KEGG 和 GO 富集分析。以 p.adjust 表示显著水平。

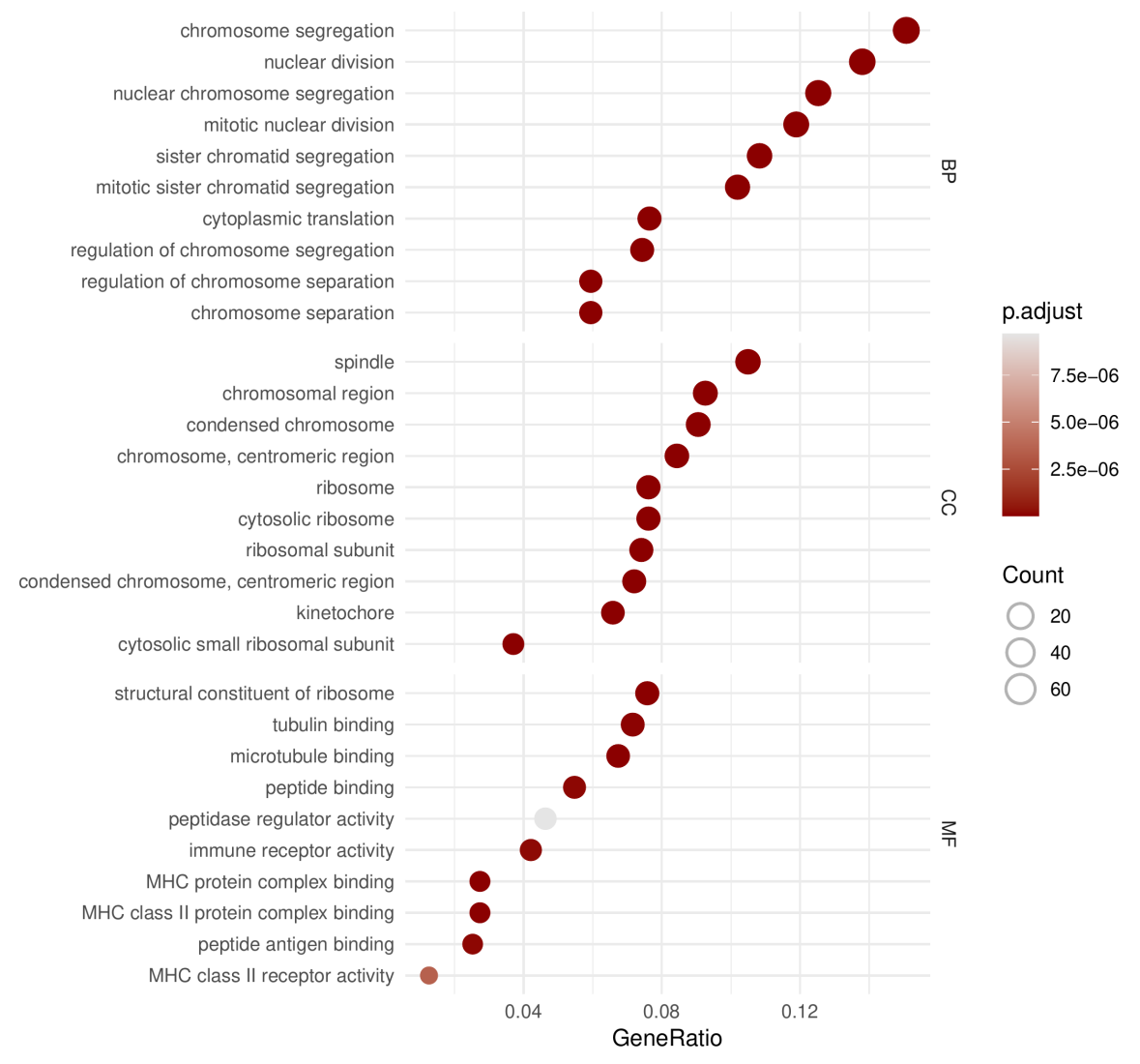
对**基因集** (TOP2A, VAMP8, CENPF, …[n = 500], 来自于Monocle3 拟时分析 Graph Test Top Significant genes[Section: OS\_CANCER]) 进行ClusterProfiler 富集分析。



**Fig.** **19** OS CANCER KEGG enrichment

**(File path: Figure+Table/3.5.3\_ClusterProfiler\_富集分析\_(OS\_CANCER)/OS-CANCER-KEGG-enrichment.pdf)**

Fig. **[19](#OS-CANCER-KEGG-enrichment)** 为 GO 富集分析气泡图。



**Fig.** **20** OS CANCER GO enrichment

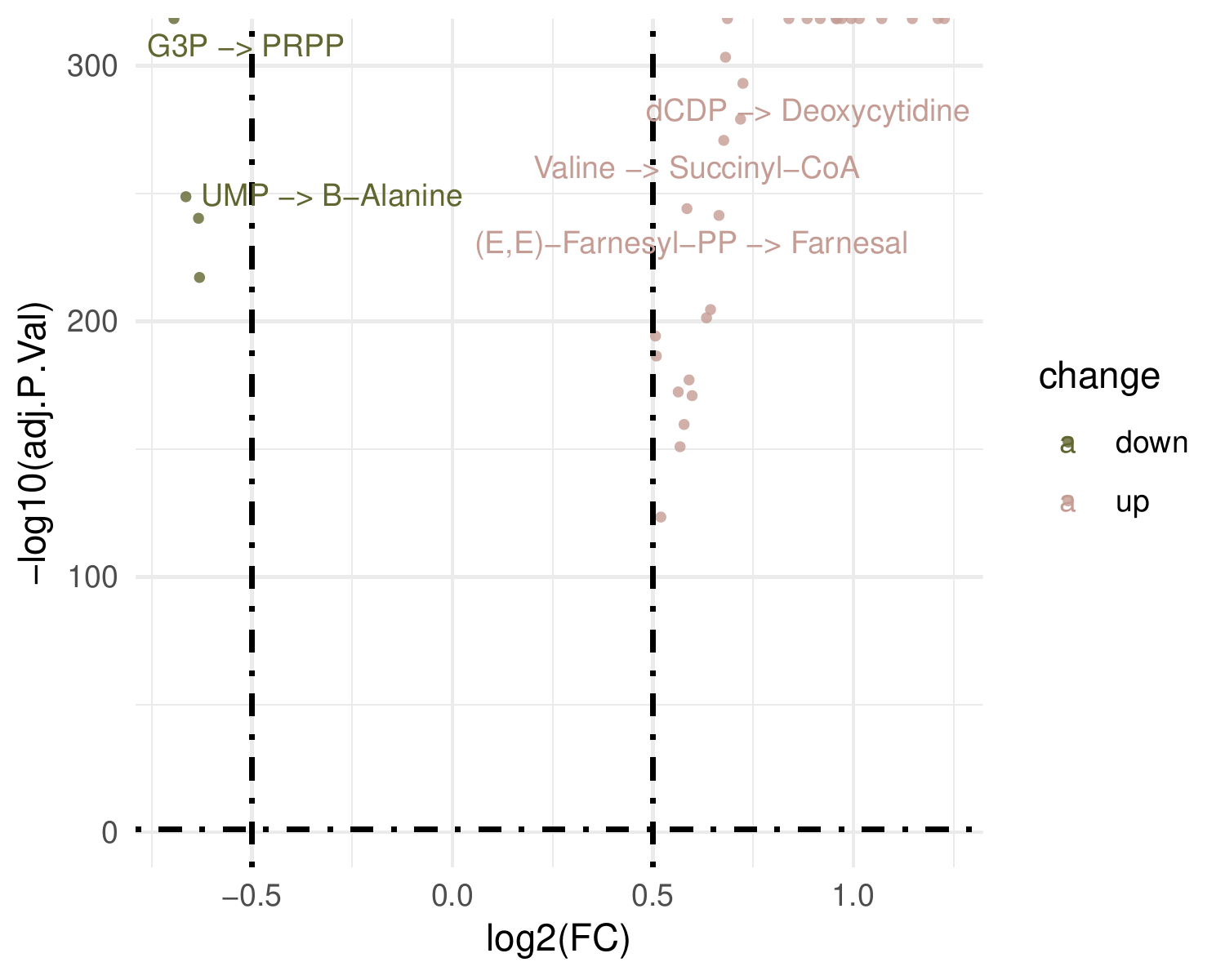
**(File path: Figure+Table/3.5.3\_ClusterProfiler\_富集分析\_(OS\_CANCER)/OS-CANCER-GO-enrichment.pdf)**

Fig. **[20](#OS-CANCER-GO-enrichment)** 为 GO 富集分析气泡图。

### 3.5.4 Limma 代谢通量差异分析 (OS\_CANCER\_FLUX)

以 limma (3.62.2) (2005)7 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：Malignant\_cell\_BC vs Benign\_cell\_BC。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 0.5 的统计结果。

匹配 scsa\_cell 中包含"\_OS$"的描述，最终得到 17122 例数据。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Malignant\_cell\_BC vs Benign\_cell\_BC。(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 DMFs 统计：up (n=29) , down (n=4)



**Fig.** **21** OS CANCER FLUX Malignant cell BC vs Benign cell BC

**(File path: Figure+Table/3.5.4\_Limma\_代谢通量差异分析\_(OS\_CANCER\_FLUX)/OS-CANCER-FLUX-Malignant-cell-BC-vs-Benign-cell-BC.pdf)**

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

**(See: Figure+Table/3.5.4\_Limma\_代谢通量差异分析\_(OS\_CANCER\_FLUX)/OS-CANCER-FLUX-Malignant-cell-BC-vs-Benign-cell-BC-content)**

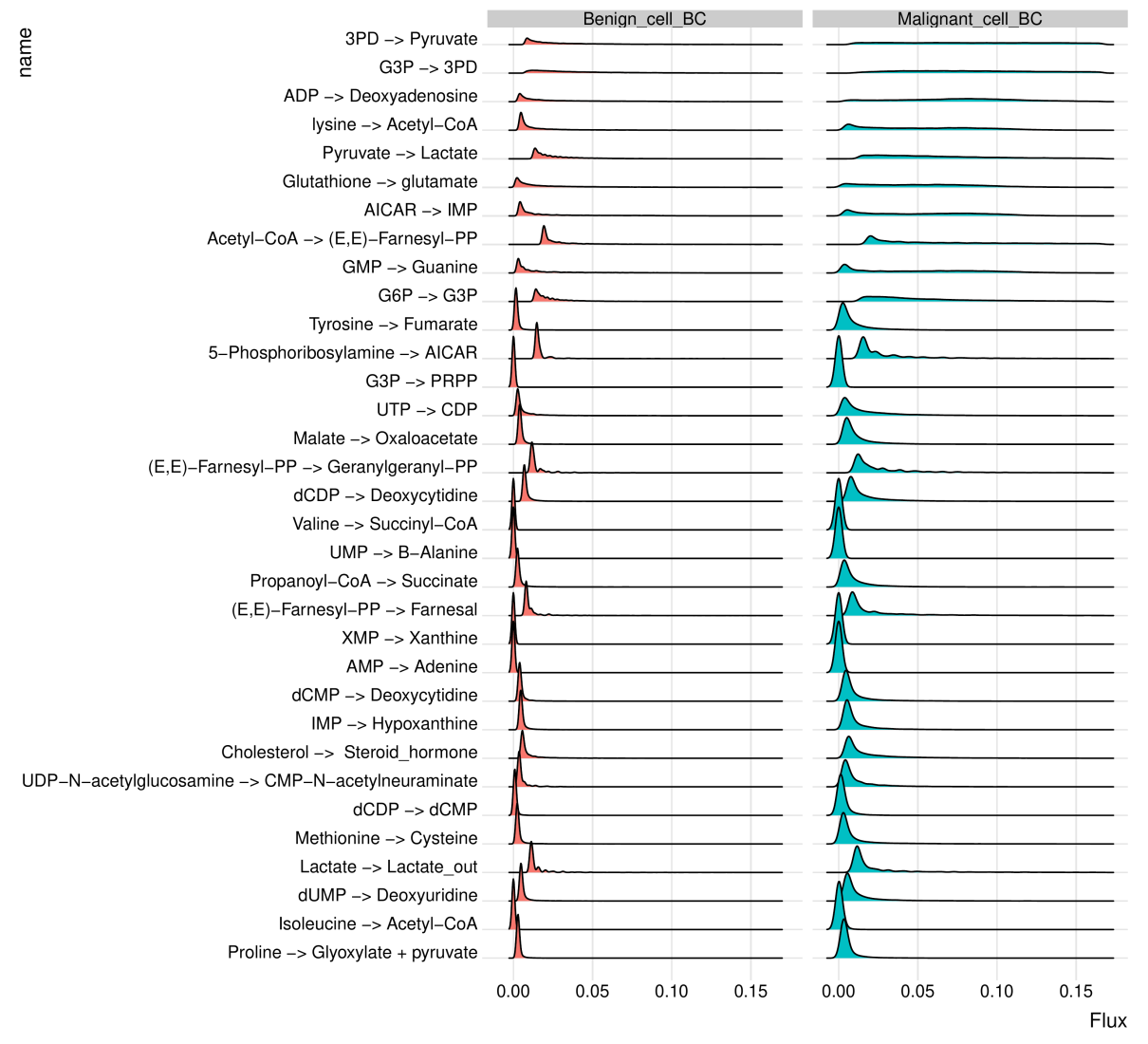
Fig. **[21](#OS-CANCER-FLUX-Malignant-cell-BC-vs-Benign-cell-BC)** 为 Malignant\_cell\_BC - Benign\_cell\_BC 差异分析火山图。

**Tab.** **6** OS CANCER FLUX data Malignant cell BC vs Benign cell BC

| Name | LogFC | Adj.P.Val | Rownames | Module id |
| --- | --- | --- | --- | --- |
| 3PD -> Pyruvate | 1.227 | 0 | M 4 | 4 |
| G3P -> 3PD | 1.211 | 0 | M 3 | 3 |
| ADP -> Deoxyadeno... | 1.015 | 0 | M 140 | 140 |
| Lysine -> Acetyl-CoA | 1.07 | 0 | M 60 | 60 |
| Pyruvate -> Lactate | 1.147 | 0 | M 6 | 6 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.5.4\_Limma\_代谢通量差异分析\_(OS\_CANCER\_FLUX)/OS-CANCER-FLUX-data-Malignant-cell-BC-vs-Benign-cell-BC.xlsx)**

Tab. **[6](#OS-CANCER-FLUX-data-Malignant-cell-BC-vs-Benign-cell-BC)** 为 Malignant\_cell\_BC - Benign\_cell\_BC 差异分析统计表格。



**Fig.** **22** OS SAMPLE Malignant cell Benign cell Cell flux ridge plot

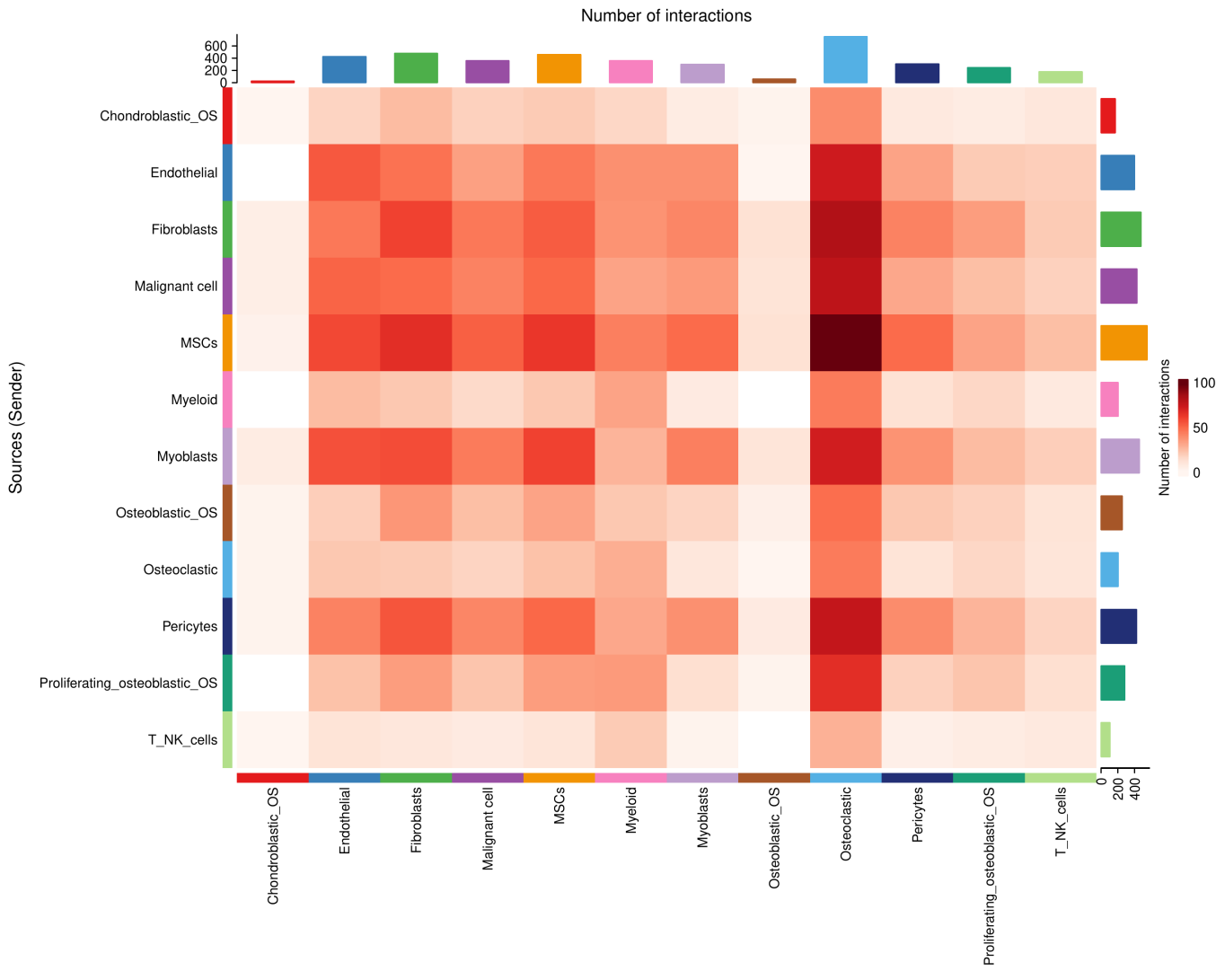
**(File path: Figure+Table/3.5.4\_Limma\_代谢通量差异分析\_(OS\_CANCER\_FLUX)/OS-SAMPLE-Malignant-cell-Benign-cell-Cell-flux-ridge-plot.pdf)**

Fig. **[22](#OS-SAMPLE-Malignant-cell-Benign-cell-Cell-flux-ridge-plot)** 为 Malignant\_cell\_Benign\_cell 的代谢通量山脊图。

### 3.5.5 CellChat 细胞通讯分析 (OS)

以 CellChat R 包 (1.6.1) (2021, **IF:14.7**, Q1, Nature Communications)8 对单细胞数据进行细胞通讯分析。以 CellChat::createCellChat 将 Seurat 对象的 RNA Assay 转化为 CellChat 对象。参照 <https://htmlpreview.github.io/?https://github.com/sqjin/CellChat/blob/master/tutorial/CellChat-vignette.html> 分析 scRNA-seq 数据。

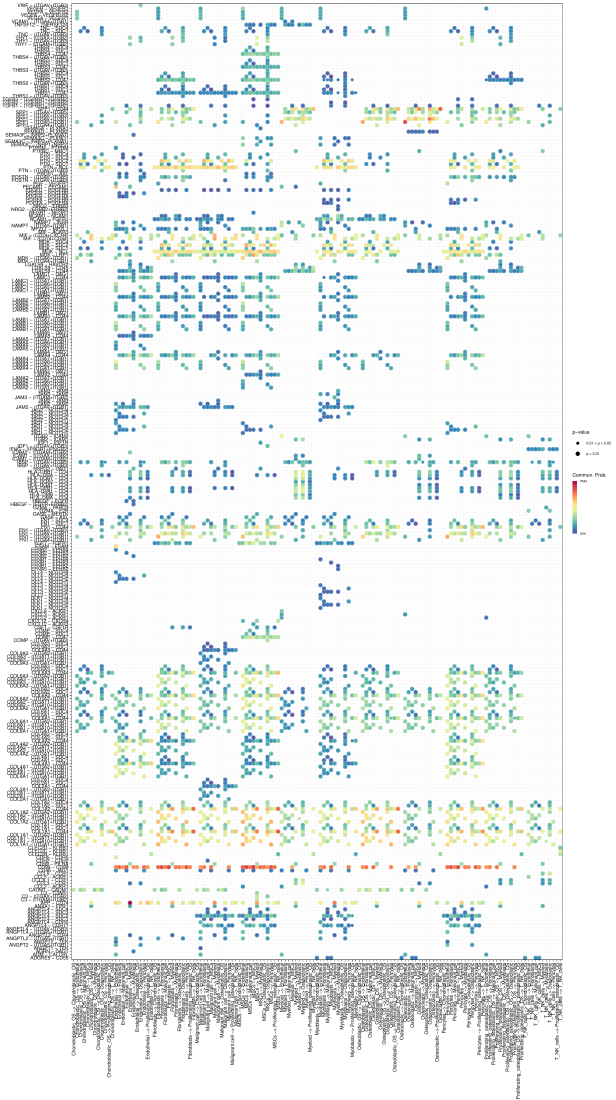
根据样本和细胞类型分组，将细胞随机抽样 (各组比例为：0.5) (细胞数量较多，通过随机抽样的方式减少计算负担) (随机种子：987456)。对 ‘Chondroblastic\_OS (n=1343)’, ‘Endothelial (n=1204)’, ‘Fibroblasts (n=6111)’, ‘Malignant cell (n=11523)’, ‘MSCs (n=744)’, ‘Myeloid (n=6075)’, ‘Myoblasts (n=165)’, ‘Osteoblastic\_OS (n=2127)’, ‘Osteoclastic (n=3001)’, ‘Pericytes (n=1089)’, ‘Proliferating\_osteoblastic\_OS (n=2110)’, ‘T\_NK\_cells (n=992)’ 细胞通讯分析。



**Fig.** **23** OS All Cell communication heatmap

**(File path: Figure+Table/3.5.5\_CellChat\_细胞通讯分析\_(OS)/OS-All-Cell-communication-heatmap.pdf)**

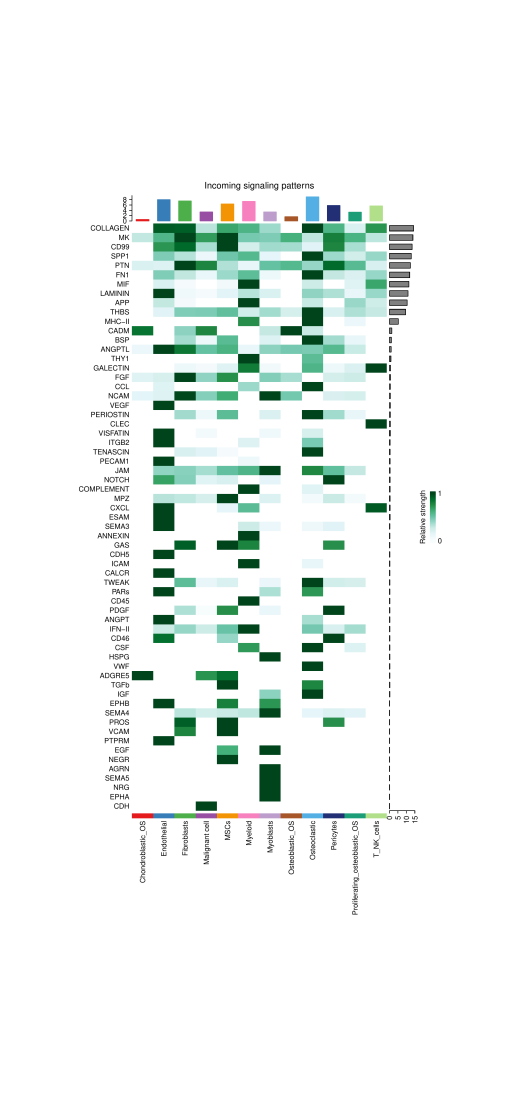
Fig. **[23](#OS-All-Cell-communication-heatmap)** 为所有细胞的通讯热图。



**Fig.** **24** OS communication probability and significant

**(File path: Figure+Table/3.5.5\_CellChat\_细胞通讯分析\_(OS)/OS-communication-probability-and-significant.pdf)**

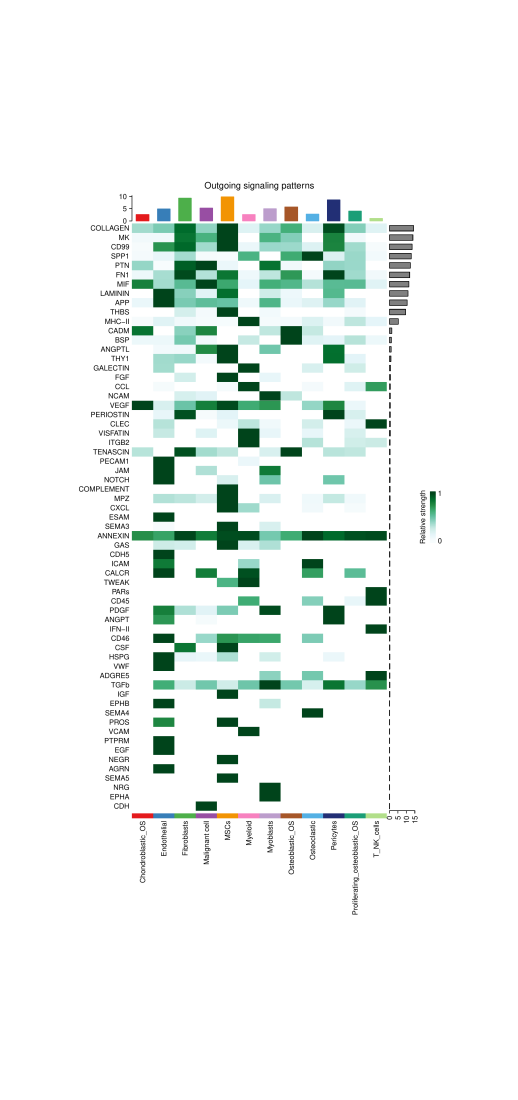
Fig. **[24](#OS-communication-probability-and-significant)** 为细胞通讯概率以及显著性。



**Fig.** **25** OS incoming ligand receptor role

**(File path: Figure+Table/3.5.5\_CellChat\_细胞通讯分析\_(OS)/OS-incoming-ligand-receptor-role.pdf)**

Fig. **[25](#OS-incoming-ligand-receptor-role)** 为细胞间的 incoming 类型通路信号强度



**Fig.** **26** OS outgoing ligand receptor role

**(File path: Figure+Table/3.5.5\_CellChat\_细胞通讯分析\_(OS)/OS-outgoing-ligand-receptor-role.pdf)**

Fig. **[26](#OS-outgoing-ligand-receptor-role)** 为细胞间的 outgoing 类型通路信号强度

## 3.6 TCGA 数据获取 (OS)

以 R 包 TCGAbiolinks (2.35.1) (2015, **IF:16.6**, Q1, Nucleic Acids Research)9 获取 TARGET-OS 数据集。

获取 TARGET-OS 数据。

## 3.7 COX 回归 (TCGA\_OS)

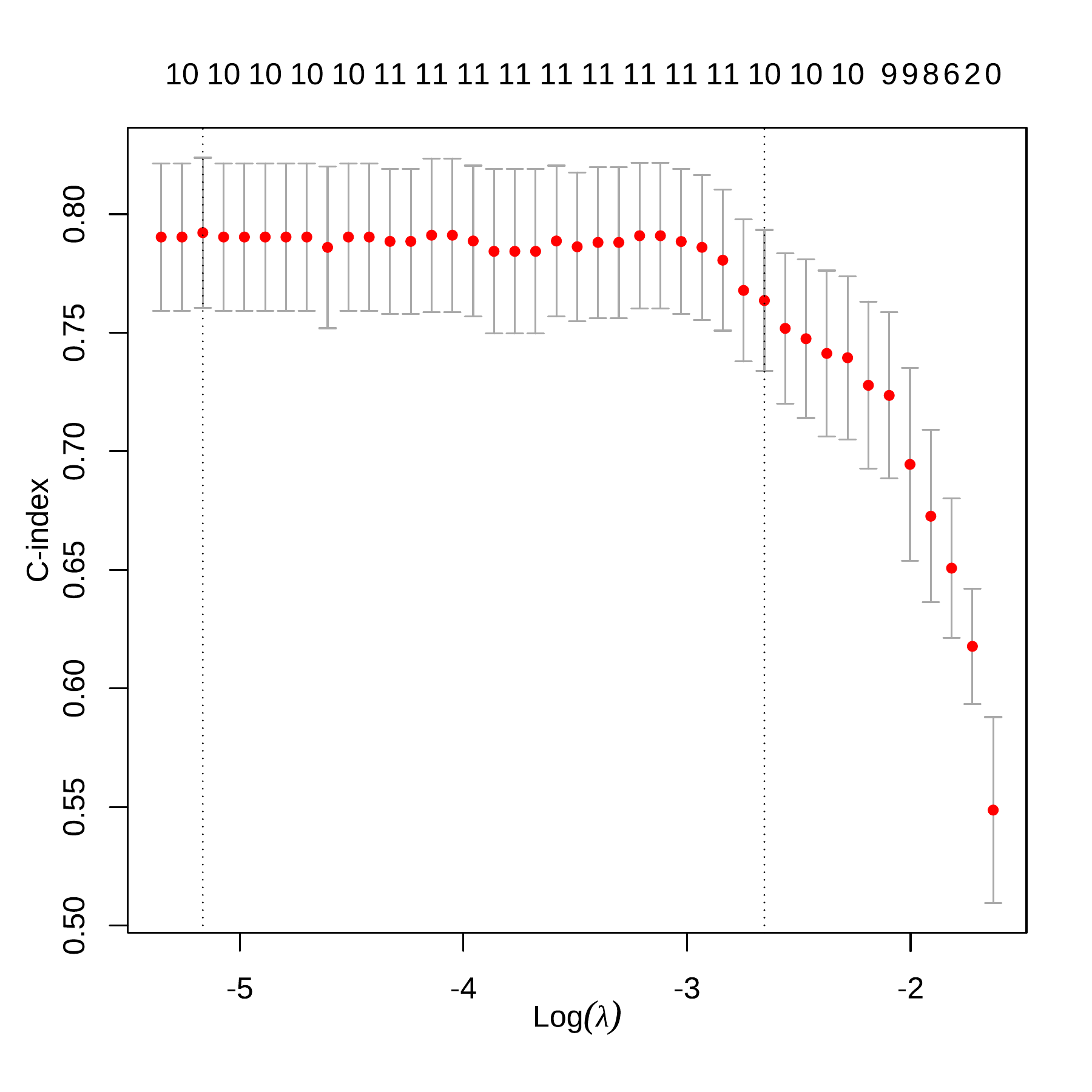
以 R 包 survival (3.8.3) 进行单因素 COX 回归 (survival::coxph)。筛选 Pr(>|z|) < .05的基因。以 R 包glmnet(4.1.8) 作 lasso 处罚的 cox 回归，以cv.glmnet函数作 5 交叉验证获得模型。以 R 包glmnet(4.1.8) 作 lasso 处罚的 cox 回归，以cv.glmnet函数作 5 交叉验证获得模型。以 R 包glmnet(4.1.8) 作 lasso 处罚的 cox 回归，以cv.glmnet` 函数作 5 交叉验证获得模型。

将**基因集** (Malignant\_cell\_Benign\_cell, 来自于scFEA 单细胞数据的代谢通量预测[Section: OS\_SAMPLE]) 用于模型建立。共 298 个基因在数据集 TARGET-OS 中找到 (根据基因名匹配)。所有数据生存状态 (去除生存状态未知的数据)，(Alive (n=57) , Dead (n=29) )。执行单因素 COX 回归，筛选 P 值 < 0.05，共筛选到 25 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 5 倍交叉验证 (评估方式为 C-index)，筛选 lambda 值。lambda.min, lambda.1se 值分别为 0.006, 0.07 (R 随机种子为 987456)。对应的特征数 (基因数) 分别为 10, 10。

**Tab.** **7** TCGA OS sig Univariate Cox Coefficients

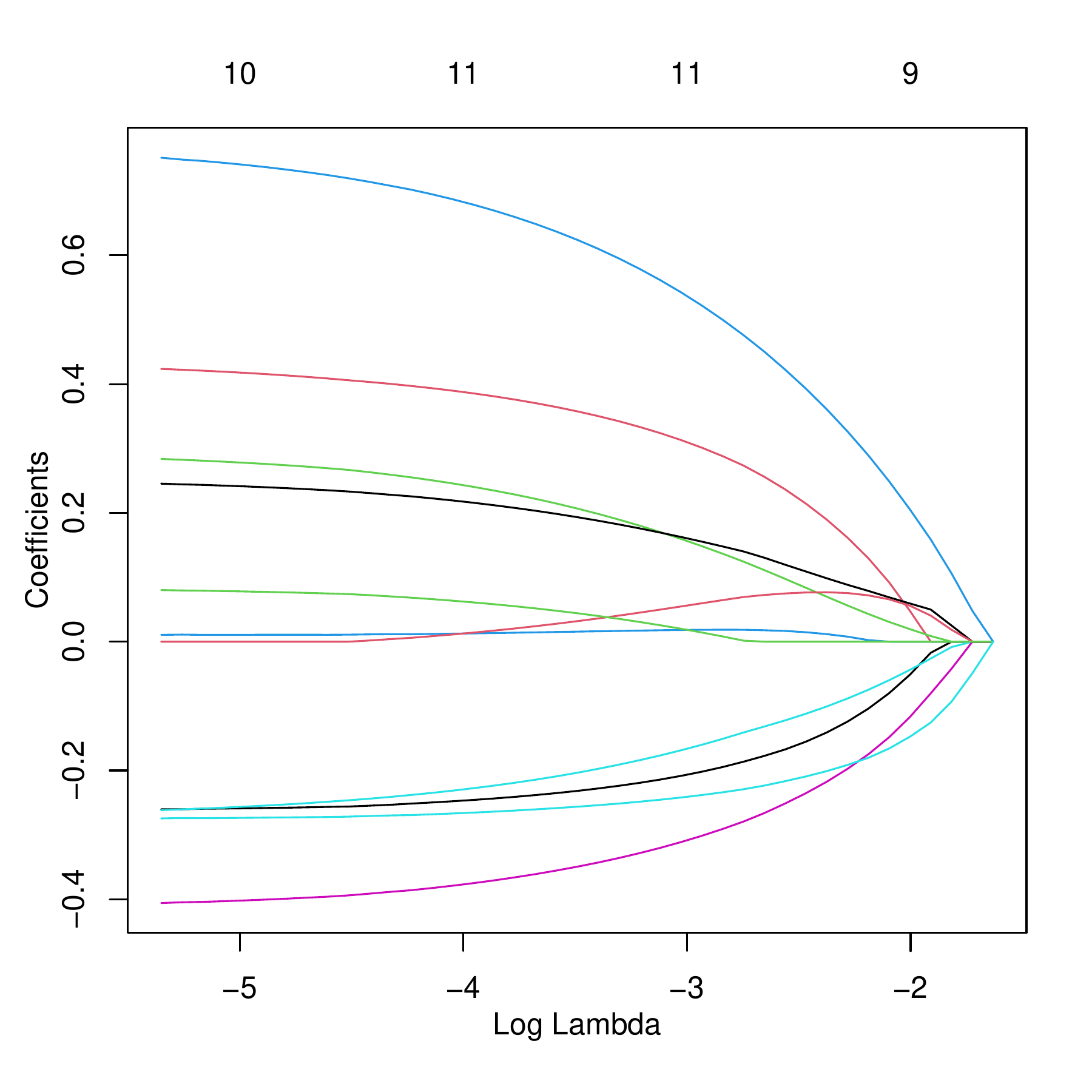
| Feature | Coef | Exp(coef) | Se(coef) | Z |
| --- | --- | --- | --- | --- |
| ACAT1 | 0.4221 | 1.525 | 0.2025 | 2.084 |
| UPRT | -0.6027 | 0.5473 | 0.208 | -2.897 |
| UGT2B10 | 0.3296 | 1.39 | 0.1409 | 2.339 |
| PCCB | 0.4687 | 1.598 | 0.175 | 2.679 |
| PGLS | -0.3914 | 0.6761 | 0.1773 | -2.207 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_OS)/TCGA-OS-sig-Univariate-Cox-Coefficients.csv)**



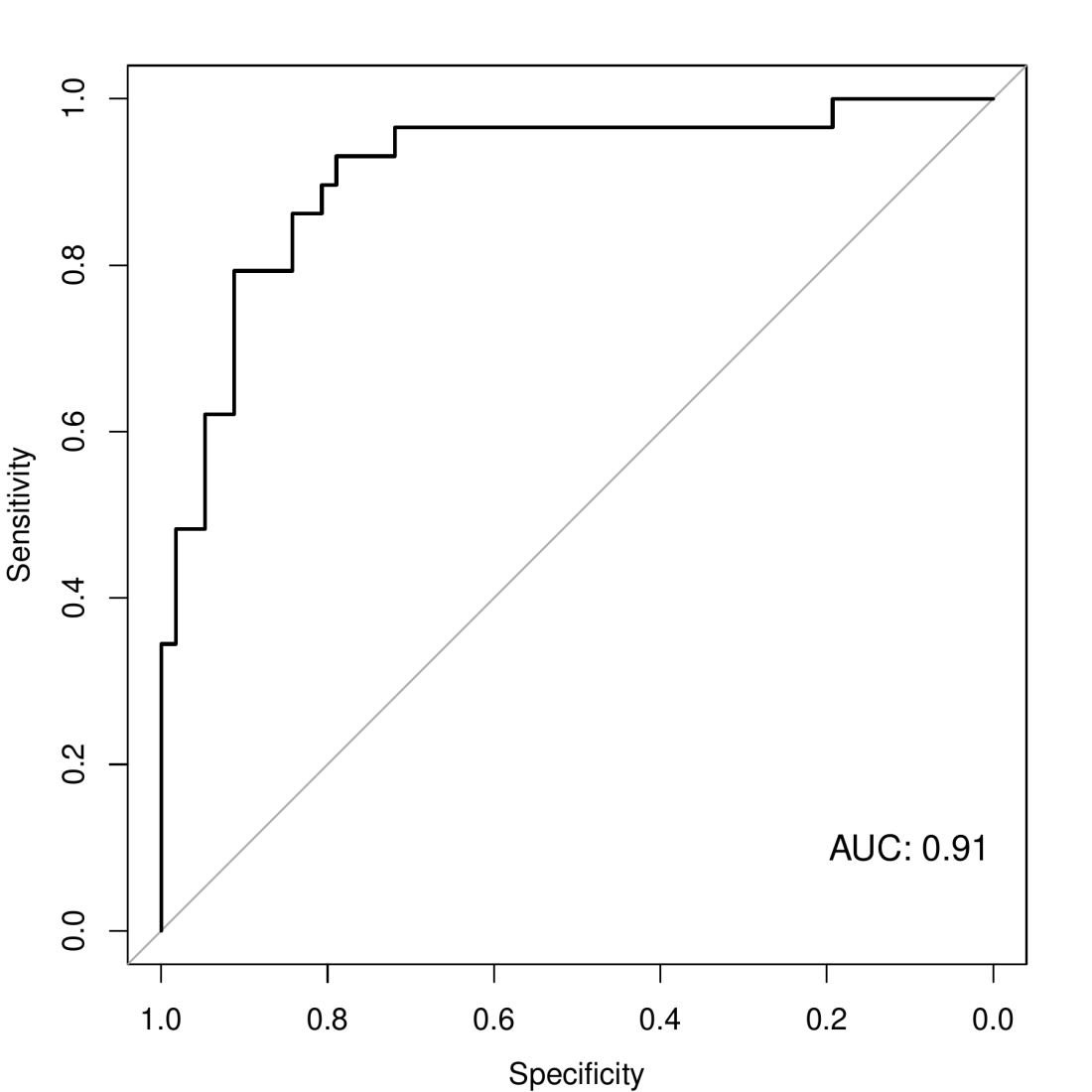
**Fig.** **27** TCGA OS lasso COX model

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



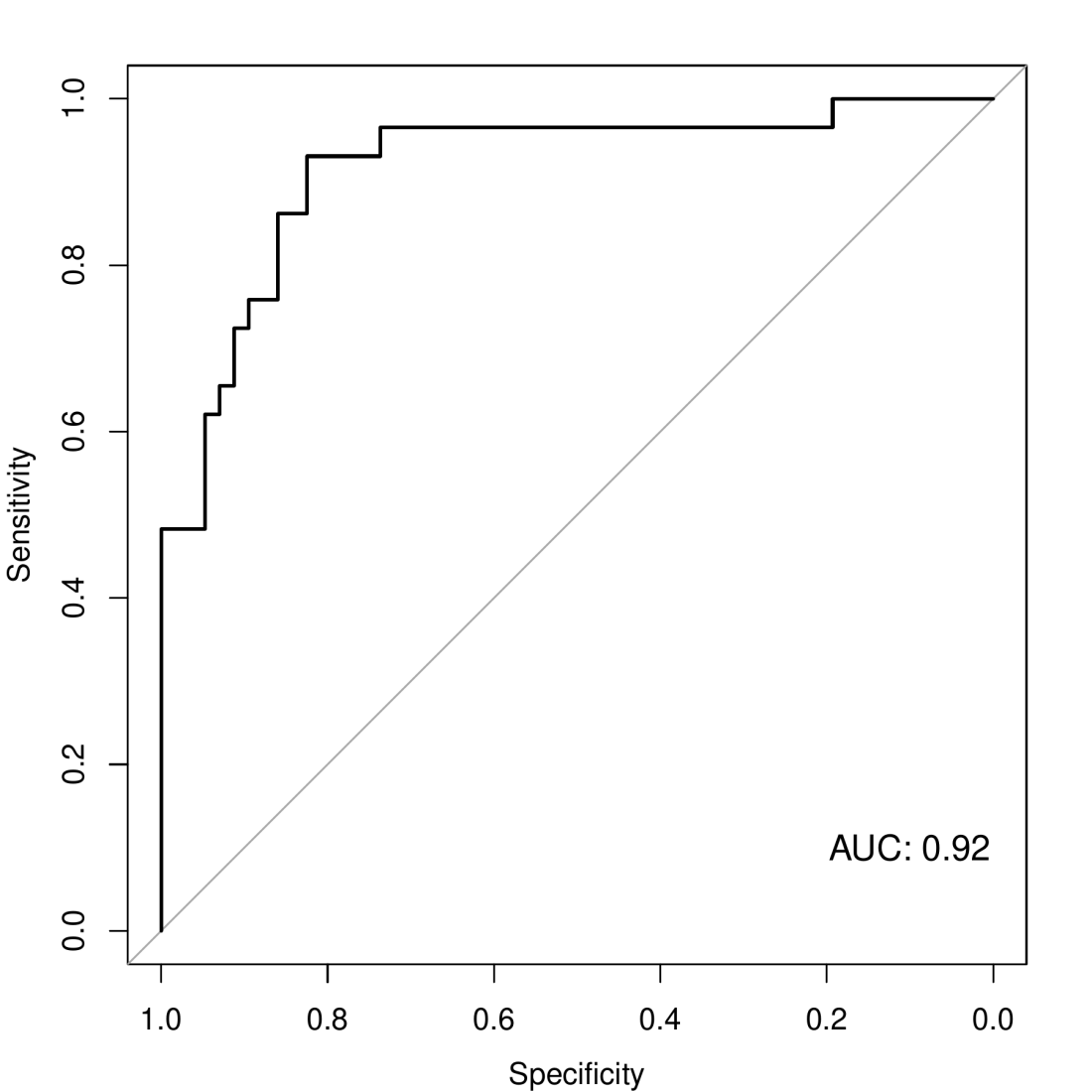
**Fig.** **28** TCGA OS lasso Cox coefficient

**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_OS)/TCGA-OS-lasso-Cox-coefficient.pdf)**



**Fig.** **29** TCGA OS lasso COX ROC lambda min

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



**Fig.** **30** TCGA OS lasso COX ROC lambda 1se

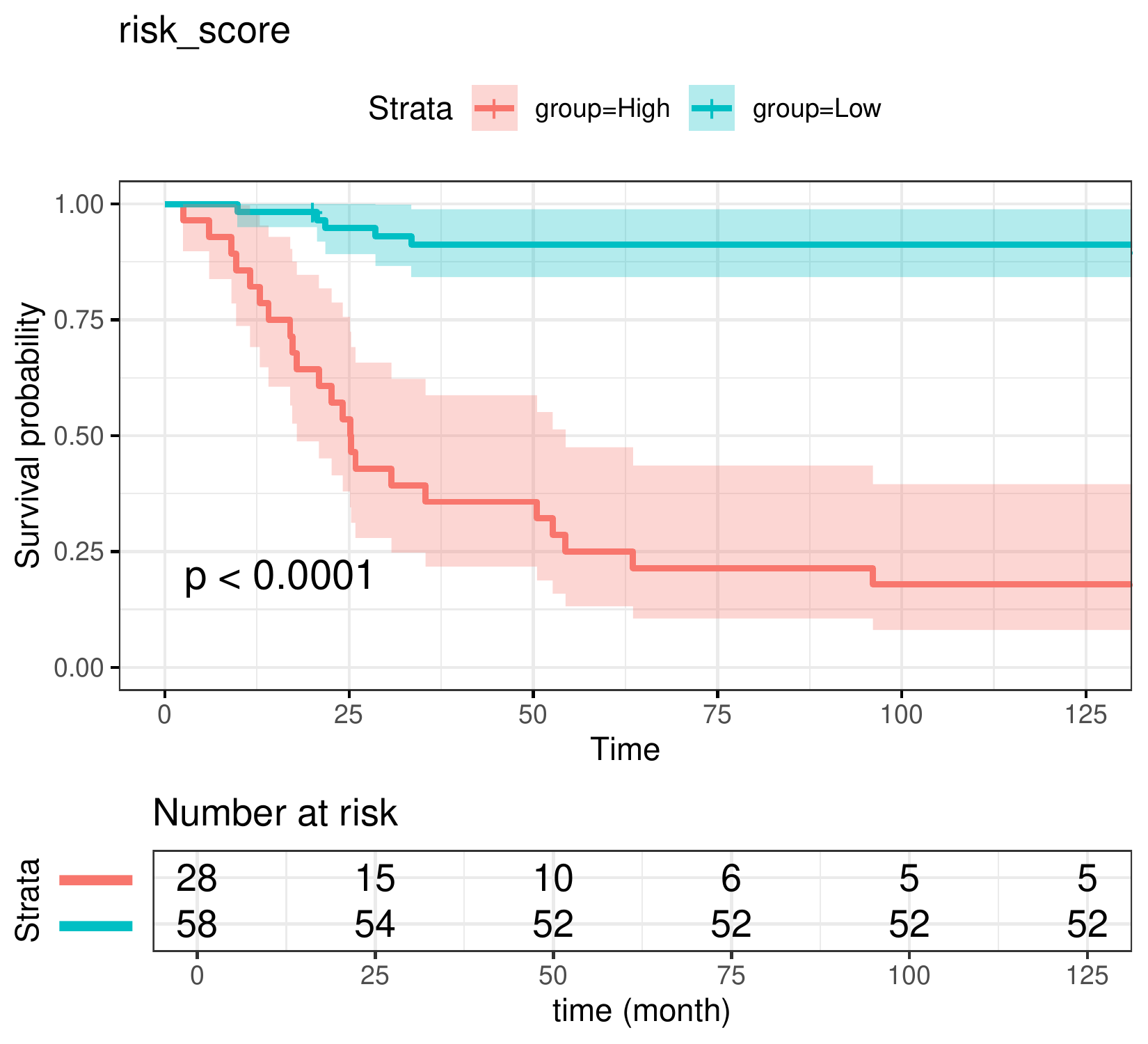
**(File path: Figure+Table/3.7.0\_COX\_回归\_(TCGA\_OS)/TCGA-OS-lasso-COX-ROC-lambda-1se.pdf)**

## 3.8 Survival 生存分析 (TCGA\_OS)

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

选择 lambda.min 时得到的特征集，包含 10 个基因， 分别为: UPRT, PCCB, CYP2C8, UCK2, ALDH4A1, G6PD, FDPS, CTPS1, HSD11B2, ABAT。以回归系数构建风险评分模型。

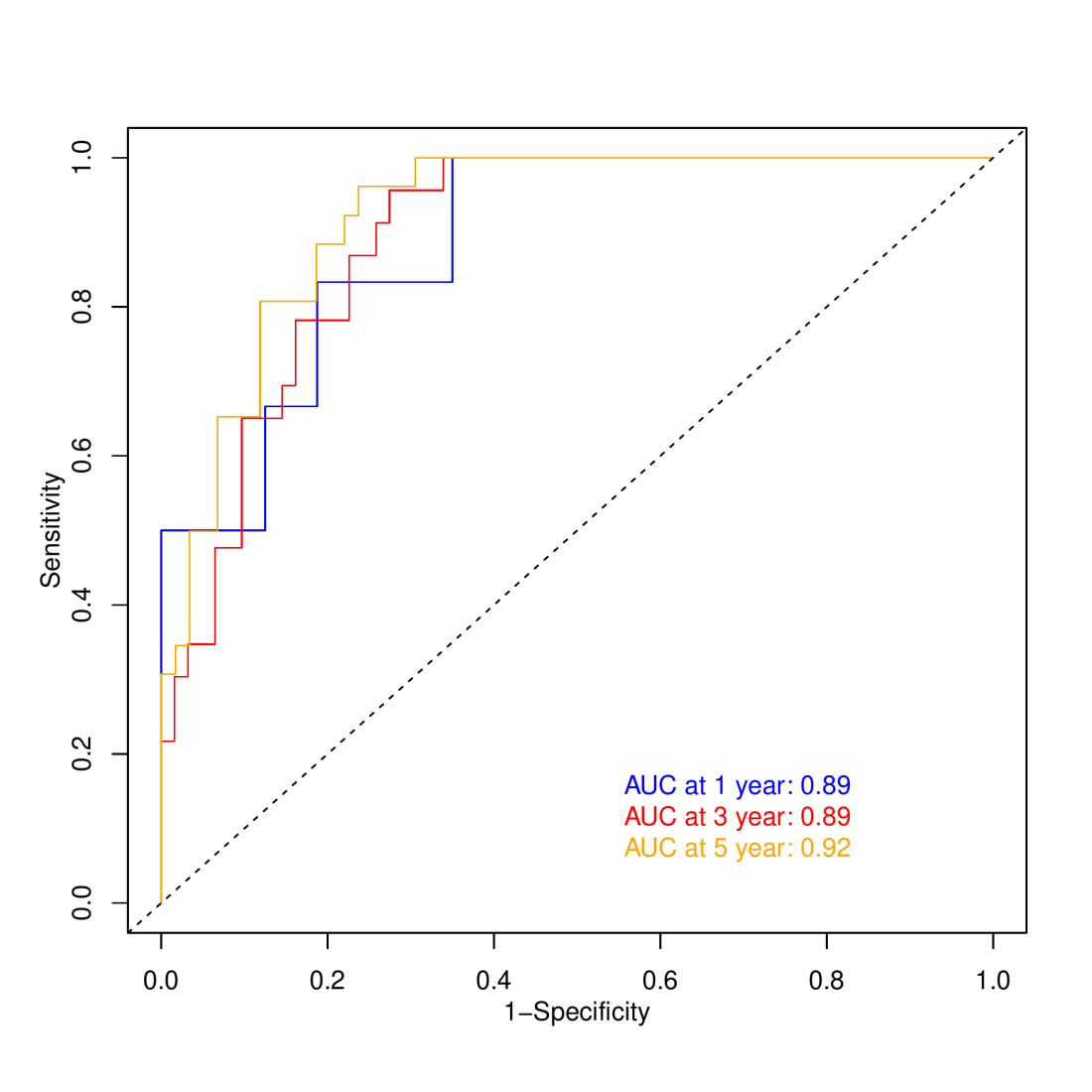
按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 0.55168695257864) (High (n=28) , Low (n=58) )， 随后进行生存分析。



**Fig.** **31** TCGA OS Survival plots

**(File path: Figure+Table/3.8.0\_Survival\_生存分析\_(TCGA\_OS)/TCGA-OS-Survival-plots.pdf)**

Fig. **[31](#TCGA-OS-Survival-plots)** 为 risk\_score 生存曲线。



**Fig.** **32** TCGA OS time ROC

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

Fig. **[32](#TCGA-OS-time-ROC)** 为 risk\_score 1, 3, 5 年生存分析 ROC 曲线。

## 3.9 外部数据集验证

### 3.9.1 GSE 数据搜索 (OS)

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: ((Osteosarcoma[Description]) AND ((6:300[Number of Samples]) AND (GSE[Entry Type]) AND (Homo sapiens[Organism]))，转化为数据表格。以正则匹配，滤除 ‘summary’ 或 ‘title’ 中包含 ‘single cell’ 或 ‘scRNA’ 的数据例。仅查询临床数据，因此滤除匹配到关键词 in vitro, cell line, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。 (注：以上仅为查找合适的 GEO 数据所做的数据筛选，与实际分析无关) 。仅获取类型包含 ‘Expression profiling by high throughput sequencing’ 或 ‘Expression profiling by array’ 的数据例。此外，排除 summary 或 title 中匹配到字符集 EX (KO, WT, *WT*, *KO*, wildtype, mutant, knock, deficien, absen, SuperSeries, transgenic, CD[0-9]+) 的数据例。上述得到 147 个 GSE 数据集。以 R 抓取网页 (例如，RCurl::getURL 抓取 )，解析 ‘Overall design’ 和 ‘Samples’ 模块，匹配字符集 EX，排除匹配到的数据例。排除 ‘Overall design’ 中包含 in vitro, cell line, CD[0-9]+, vehicle, vector, DMSO, /ml, nm 的数据例。仅获取包含 ‘protein coding’ 测序的数据集，排除 ‘Samples’ 和 ‘Overall design’ 中包含 siRNA, miRNA, miR, lncRNA 字符的数据例。余下共 73 个。以 GEOquery 获取 GSE 数据集 (n=73)。从元数据中匹配包含关键词的数据：‘Survival|Event|Dead|Alive|Status|Day|Time’，共得到 17 个数据集。

以 Entrez Direct (EDirect) 搜索 GEO 数据库 (检索条件见方法章节) 。 在检索匹配后，经人工确认，全部带有生存数据的 Osteosarcoma 为：GSE16091, GSE39055, GSE39057, GSE21257

### 3.9.2 GEO 数据获取 (OS\_GSE39057)

以 R 包 GEOquery (2.74.0) 获取 GSE39057 数据集。

以 GEOquery 获取 GSE39057 的数据信息。

### 3.9.3 GEO 数据获取 (OS\_GSE39055)

以 R 包 GEOquery (2.74.0) 获取 GSE39055 数据集。

以 GEOquery 获取 GSE39055 的数据信息。

### 3.9.4 GEO 数据获取 (OS\_GSE16091)

以 R 包 GEOquery (2.74.0) 获取 GSE16091 数据集。

以 GEOquery 获取 GSE16091 的数据信息。

### 3.9.5 GEO 数据获取 (OS\_GSE21257)

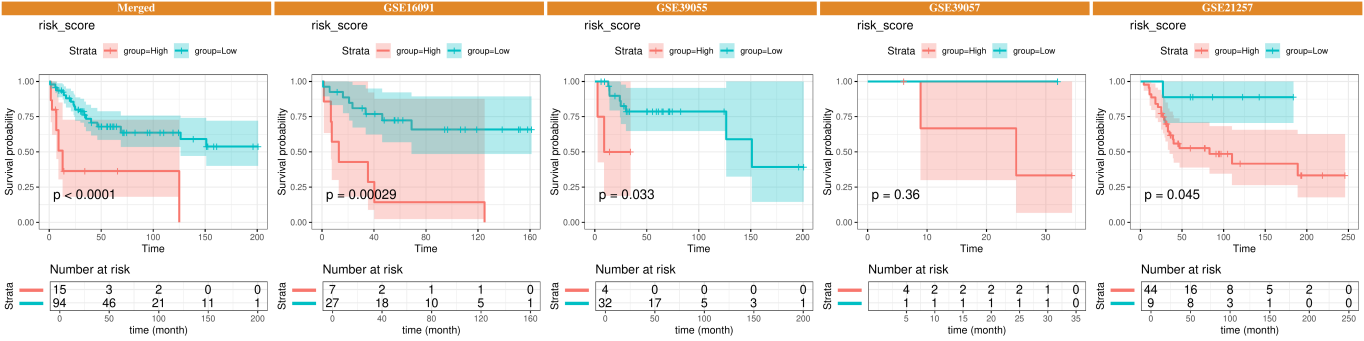
以 R 包 GEOquery (2.74.0) 获取 GSE21257 数据集。

以 GEOquery 获取 GSE21257 的数据信息。

### 3.9.6 Survival 生存分析 (OS\_OUTER)

对在 GEO 找到的所有具备生存信息的 Osteosarcoma 数据集做了外部验证。 以 R 包 survival (3.8.3) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。

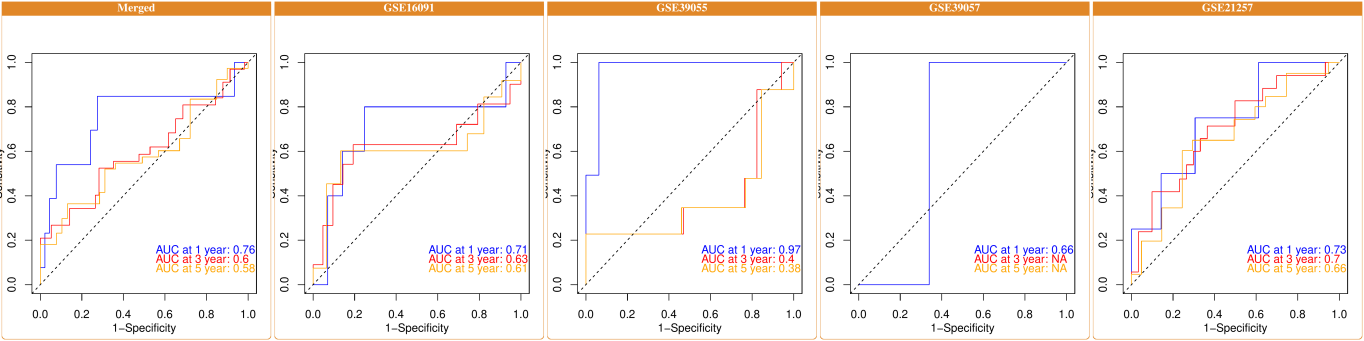
合并数据集 (GSE16091, GSE39055, GSE39057, GSE21257)。对于不同注释来源的基因名，以 org.Hs.eg.db::org.Hs.eg.db 获取基因的别名 (ALIAS) ，根据 (ALIAS) 的一致性合并。查找预后模型中基因的 ALIAS，在未找到对应基因的情况下，使用该基因的 ALIAS 查找。 (原模型基因：UPRT, PCCB, CYP2C8, UCK2, ALDH4A1, G6PD, FDPS, CTPS1, HSD11B2, ABAT；以 ALIAS 匹配后，基因为：UPRT, PCCB, CYP2C8, UCK2, ALDH4A1, G6PD, FDPS, CTPS1, HSD11B2, ABAT) 。随后，将基因表达数据归一化 (Z-score)。按 survminer::surv\_cutpoint 计算的 cutoff， 将样本分为 Low 和 High 风险组 (cutoff: 1.12270604442518) (High (n=15) , Low (n=94) )， 随后进行生存分析。此外，对于未合并前的各个数据集，以相同的方式生存分析。



**Fig.** **33** OS OUTER all datasets survival plot

**(File path: Figure+Table/3.9.6\_Survival\_生存分析\_(OS\_OUTER)/OS-OUTER-all-datasets-survival-plot.pdf)**

Fig. **[33](#OS-OUTER-all-datasets-survival-plot)** 为所有数据集的生存分析图。



**Fig.** **34** OS OUTER all datasets ROC validation

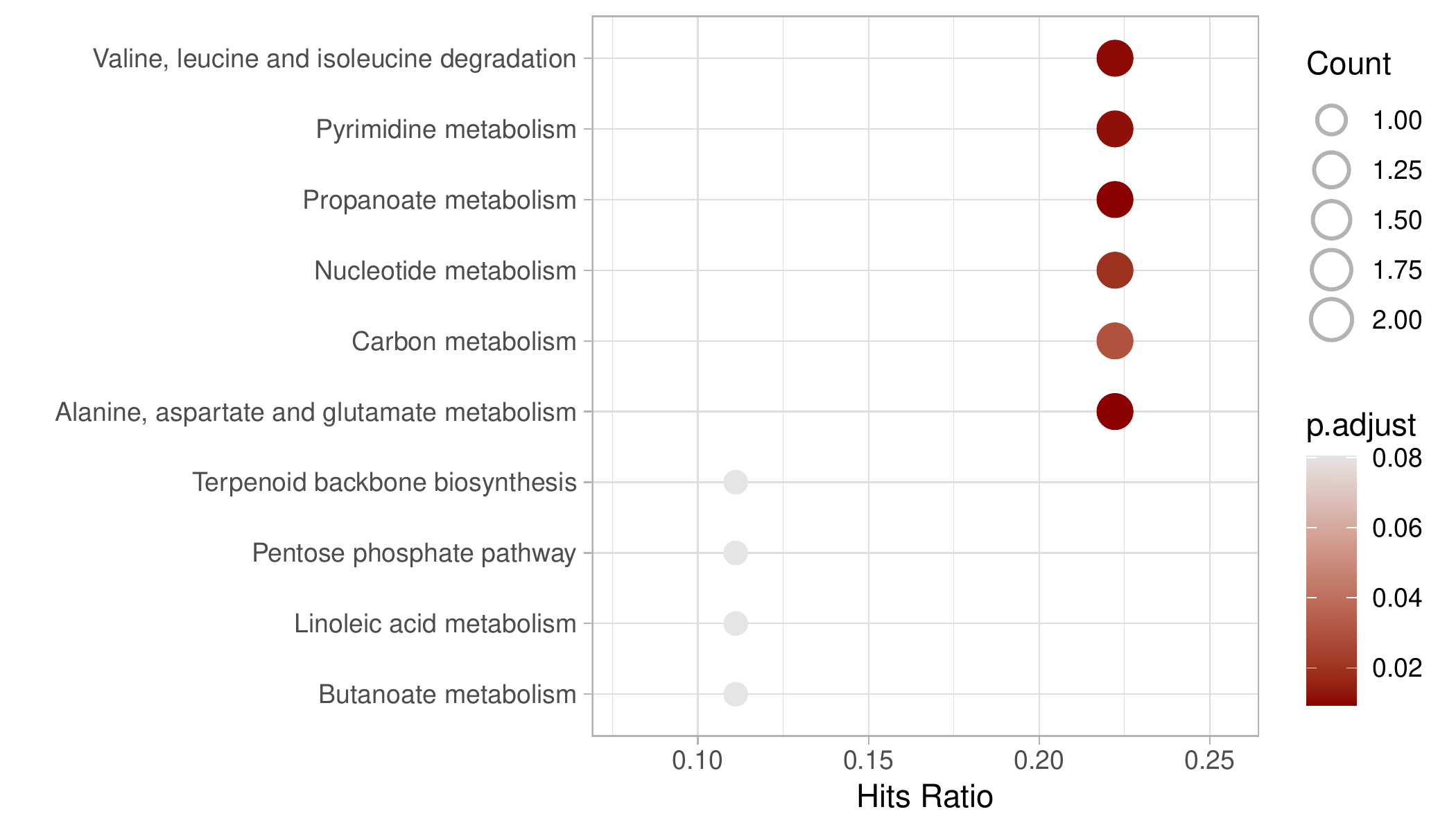
**(File path: Figure+Table/3.9.6\_Survival\_生存分析\_(OS\_OUTER)/OS-OUTER-all-datasets-ROC-validation.pdf)**

Fig. **[34](#OS-OUTER-all-datasets-ROC-validation)** 为所有数据集的 ROC 图。

## 3.10 ClusterProfiler 富集分析 (PROG)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)6进行 KEGG 和 GO 富集分析。以 p.adjust 表示显著水平。

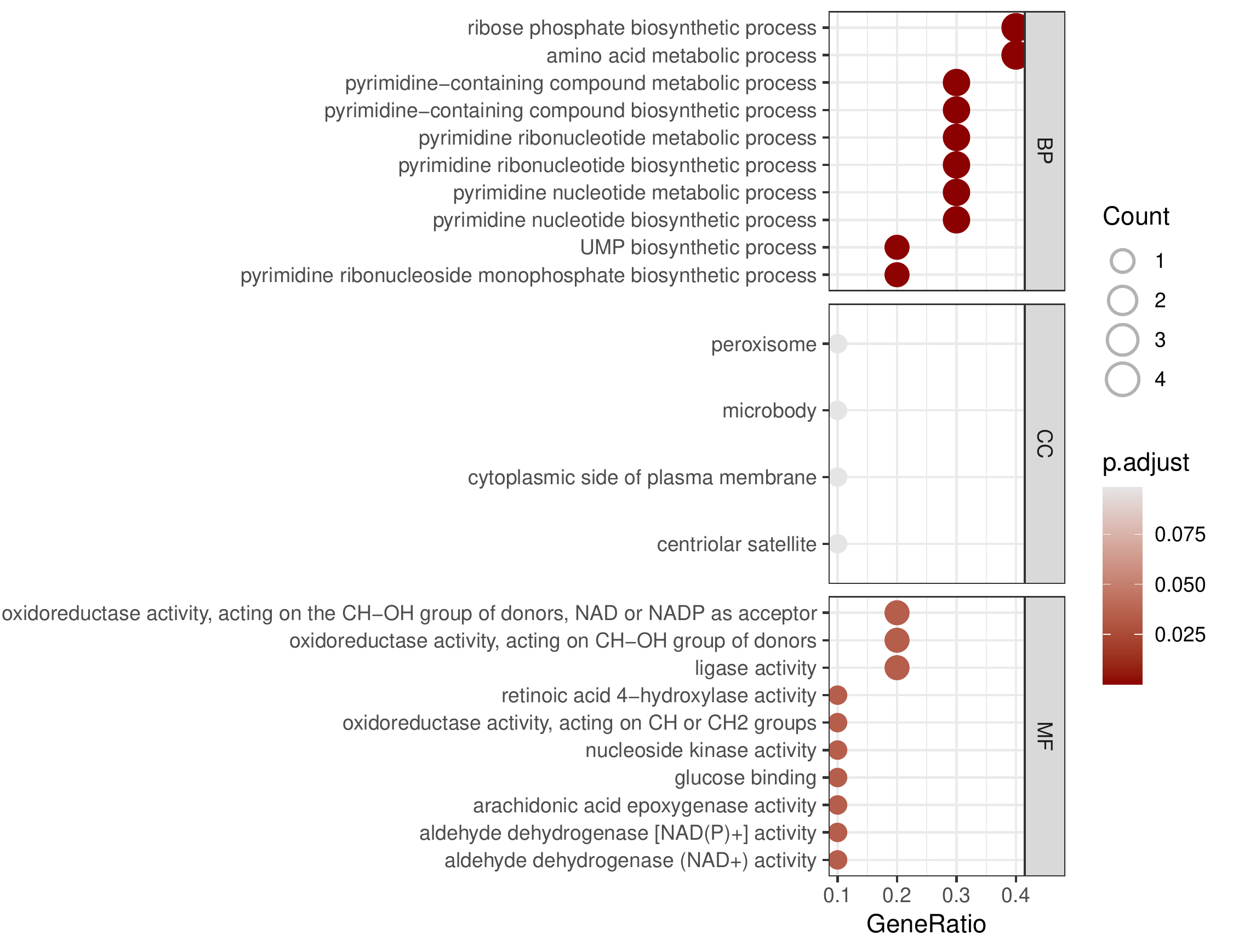
对**基因集** (UPRT, PCCB, CYP2C8, …[n = 10], 来自于Survival 生存分析[Section: TCGA\_OS]) 进行ClusterProfiler 富集分析。



**Fig.** **35** PROG KEGG enrichment

**(File path: Figure+Table/3.10.0\_ClusterProfiler\_富集分析\_(PROG)/PROG-KEGG-enrichment.pdf)**

Fig. **[35](#PROG-KEGG-enrichment)** 为 GO 富集分析气泡图。



**Fig.** **36** PROG GO enrichment

**(File path: Figure+Table/3.10.0\_ClusterProfiler\_富集分析\_(PROG)/PROG-GO-enrichment.pdf)**

Fig. **[36](#PROG-GO-enrichment)** 为 GO 富集分析气泡图。

# 4 总结

癌症的本质特征与癌细胞自身代谢的改变息息相关 (2008, **IF:48.8**, Q1, Cancer cell)10。 对于骨肉瘤，目前仍缺少研究从单细胞水平探究癌症的代谢变化。 本分析从单细胞水平鉴定的恶质细胞 (肿瘤细胞) 出发，分析正常细胞与癌症细胞之间的代谢通量差异， 进而获取对应代谢模块的基因，建立预后模型，以代谢改变的角度，预测疾病的进展。 模型以 TARGET-OS 数据集建立，进而在 GEO 数据库搜索了所有可用的带有生存信息的基因表达数据集， 用以验证预后模型的可靠性。

# Reference

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

2. Zhou, Y. *et al.* Single-cell rna landscape of intratumoral heterogeneity and immunosuppressive microenvironment in advanced osteosarcoma. *Nature communications* **11**, (2020).

3. Gao, R. *et al.* Delineating copy number and clonal substructure in human tumors from single-cell transcriptomes. *Nature Biotechnology* **39**, 599–608 (2021).

4. Gordon, D. J., Resio, B. & Pellman, D. Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics* **13**, 189–203 (2012).

5. Alghamdi, N. *et al.* A graph neural network model to estimate cell-wise metabolic flux using single-cell rna-seq data. *Genome research* **31**, 1867–1884 (2021).

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

7. 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).

8. Jin, S. *et al.* Inference and analysis of cell-cell communication using cellchat. *Nature Communications* **12**, (2021).

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

10. Kroemer, G. & Pouyssegur, J. Tumor cell metabolism: Cancers achillesheel. *Cancer cell* **13**, 472–482 (2008).