**生信分析报告**

**项目标题： 黑色素瘤 ;**

**单 号： BSHQ240813 ;**

**分析人员： 黄礼闯 ;**

**分析类型： 分析优化 ;**

**委 托 人： 王凯 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程

1. 通过生物信息分析分析 CAP2 在黑色素瘤异常表达且与预后相关。HPA数据库分析。
2. 为了研究CAP2影响黑色素瘤进展的机制，进行GSEA等生物信息分析，寻找与CAP2R显著相关的下游信号通路

# 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: SKCM)

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: XENA\_SKCM)

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：SKCM 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 TCGA 数据获取 (Dataset: TCGA\_SKCM)

以 R 包 TCGAbiolinks (2.35.1) (2015, **IF:16.6**, Q1, Nucleic Acids Research)4 获取 TCGA-SKCM 数据集。

## 2.5 Survival 生存分析 (Dataset: TCGA\_SKCM)

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。

## 2.6 Limma 差异分析 (Dataset: TCGA\_SKCM)

使用 Survival 分析 (TCGA\_SKCM) 时定义的分组数据 (根据 CAP2 的表达，分为 High (n=236) , Low (n=237) )。 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：High vs Low。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。

## 2.7 ClusterProfiler GSEA 富集分析 (Dataset: TCGA\_SKCM)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)5 按 GSVA 算法 (clusterProfiler::gseGO, ClusterProfiler::gseKEGG)，进行 KEGG 和 GO 富集分析 (P-value Cutoff = 0.05) 。 以 R 包 msigdbr (10.0.1) 获取 MSigDB 数据库基因集，用于 clusterProfiler GSEA 富集分析。

# 3 分析结果

## 3.1 UCSCXenaTools 癌症相关数据获取 (SKCM)

获取 UCSC Xena 的 TcgaTargetGtex 数据。共 1023 个数据。样本组织为：Skin (n=1023) 。样本类型为：Metastatic (n=366) , Normal Tissue (n=555) , Primary Tumor (n=102) 。数据来源为：GTEX (n=555) , TCGA (n=468) 。

## 3.2 Limma 差异分析 (XENA\_SKCM)

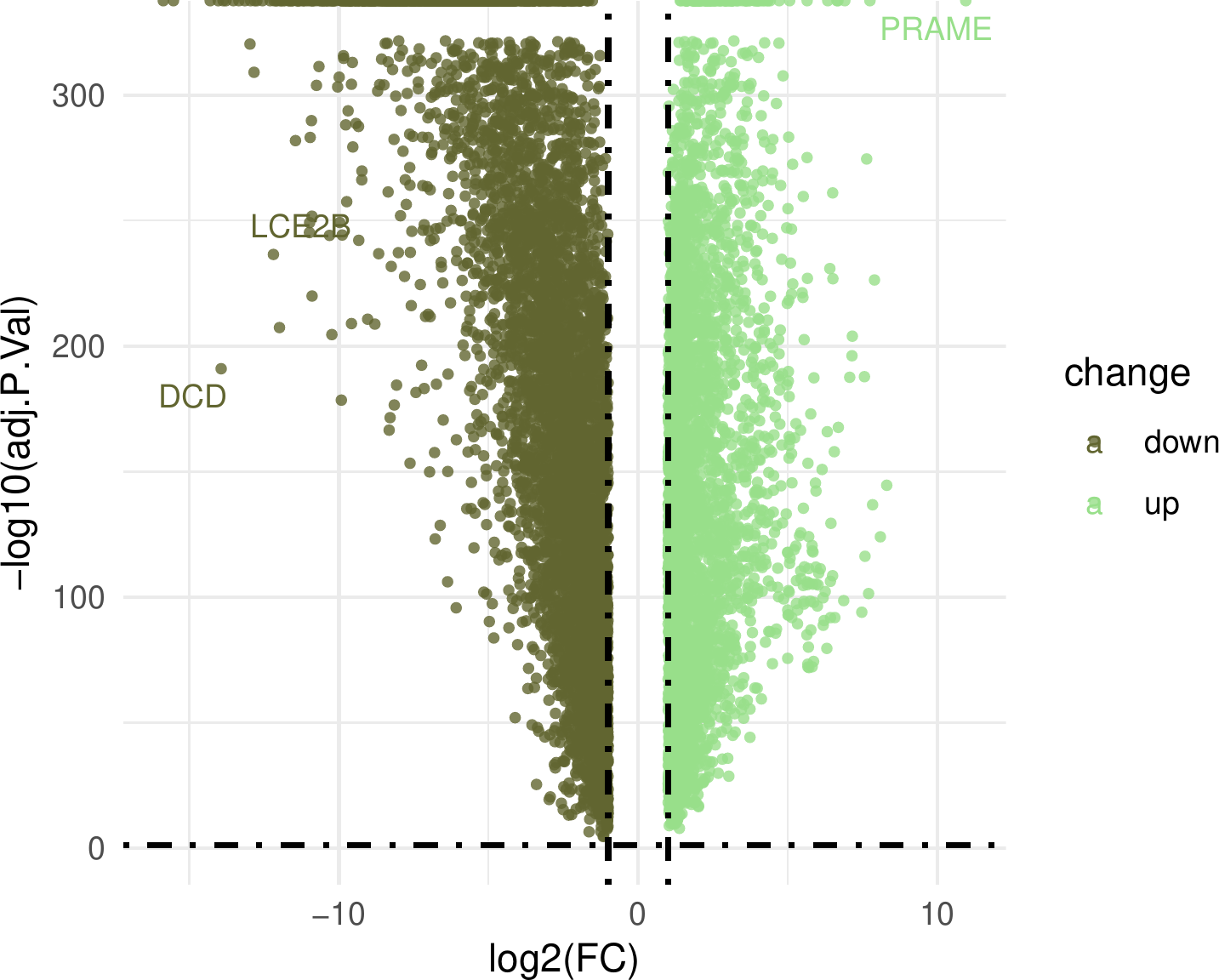
将 TcgaTargetGtex-SKCM 数据转化为 counts 型数据 (原数据为 log2(counts + 1)) 。样本分组：Normal (n=555) , SKCM (n=468) 。以 公式 ~ 0 + group + batch 创建设计矩阵 (design matrix) (Batch: GTEX (n=555) , TCGA (n=468) )。差异分析：SKCM vs Normal。(若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B)。上调或下调 DEGs 统计：up (n=3819) , down (n=5963)

**Tab.** **1** XENA SKCM data SKCM vs Normal

| Hgnc symbol | LogFC | Adj.P.Val | Rownames | Id |
| --- | --- | --- | --- | --- |
| AC009065.4 | 6.644 | 0 | ENSG00000279473.1 | ENSG00000279473.1 |
| RP11-40C6.2 | 7.747 | 0 | ENSG00000219928.2 | ENSG00000219928.2 |
| VIPR1-AS1 | -7.064 | 0 | ENSG00000232354.7 | ENSG00000232354.7 |
| HNRNPCP2 | 3.7 | 0 | ENSG00000204253.4 | ENSG00000204253.4 |
| IL20RB-AS1 | -6.995 | 0 | ENSG00000249407.1 | ENSG00000249407.1 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.2\_Limma\_差异分析\_(XENA\_SKCM)/XENA-SKCM-data-SKCM-vs-Normal.csv)**

Tab. **[1](#XENA-SKCM-data-SKCM-vs-Normal)** 为 SKCM - Normal 差异分析统计表格。



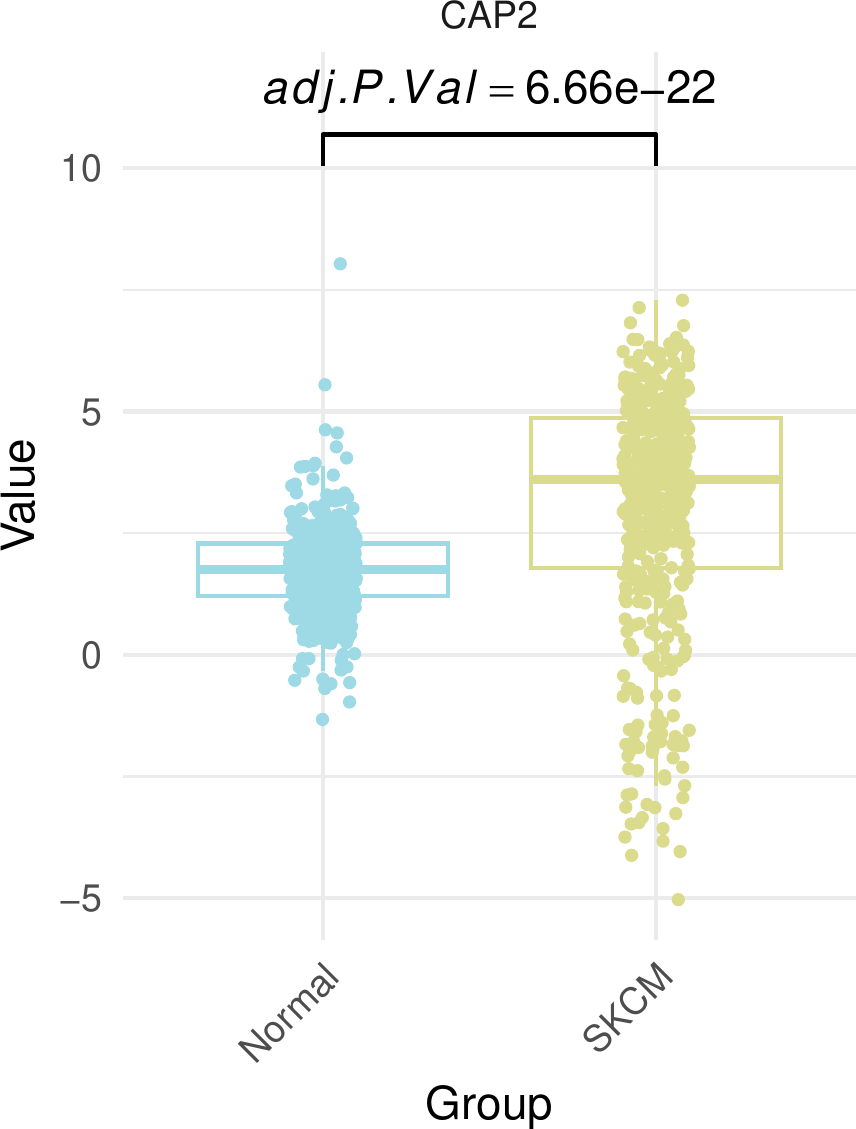
**Fig.** **1** XENA SKCM SKCM vs Normal

**(File path: Figure+Table/3.2\_Limma\_差异分析\_(XENA\_SKCM)/XENA-SKCM-SKCM-vs-Normal.pdf)**

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

**(See: Figure+Table/3.2\_Limma\_差异分析\_(XENA\_SKCM)/XENA-SKCM-SKCM-vs-Normal-content)**

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



**Fig.** **2** XENA SKCM Box Plot Of DEGs

**(File path: Figure+Table/3.2\_Limma\_差异分析\_(XENA\_SKCM)/XENA-SKCM-Box-Plot-Of-DEGs.pdf)**

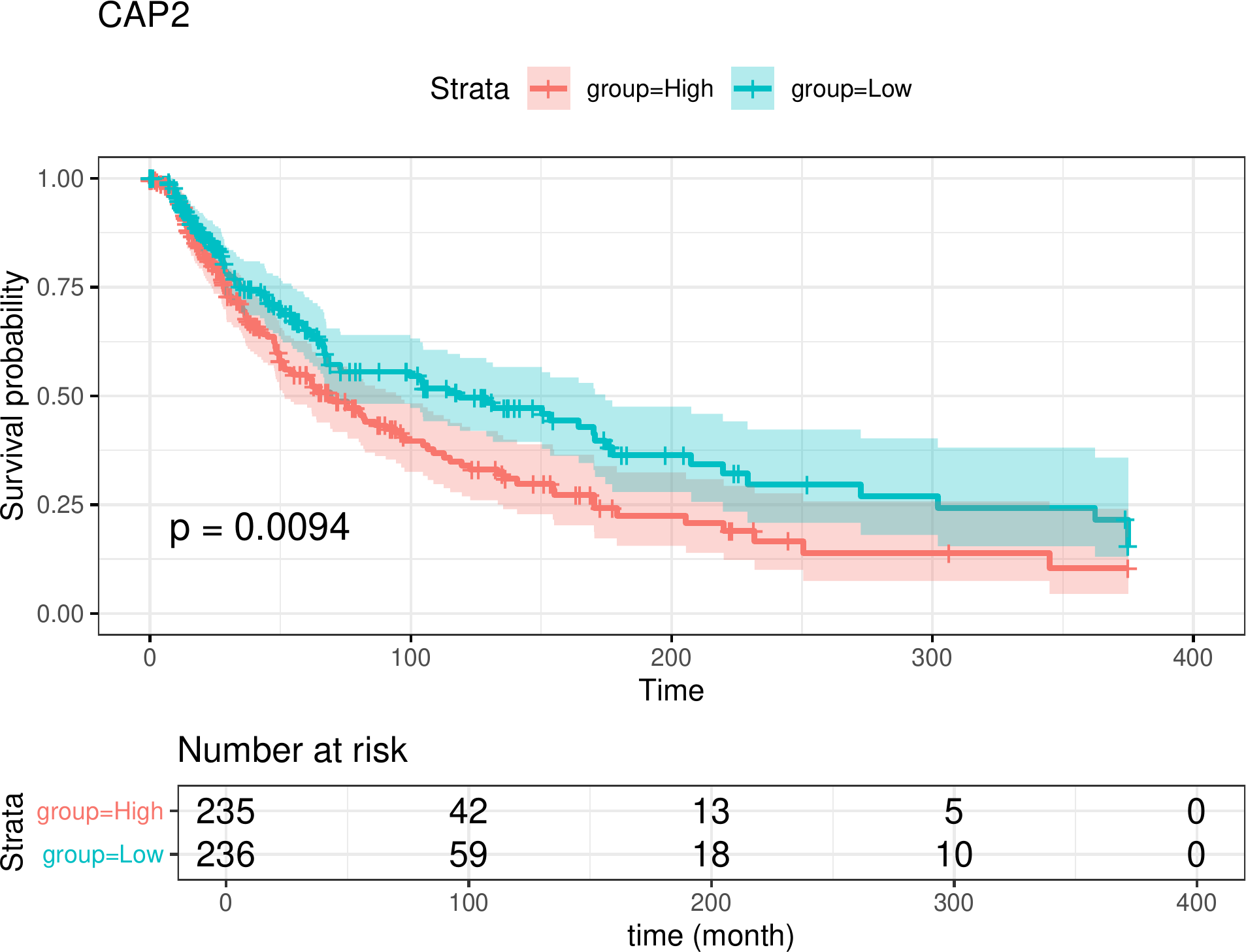
Fig. **[2](#XENA-SKCM-Box-Plot-Of-DEGs)** 基因 CAP2 表达水平，以及对应的 limma 差异分析显著水平。

## 3.3 TCGA 数据获取 (TCGA\_SKCM)

获取 TCGA-SKCM 数据。

## 3.4 Survival 生存分析 (TCGA\_SKCM)

按中位风险评分，将样本分为 Low 和 High 风险组。生存数据为TCGA-SKCM，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。



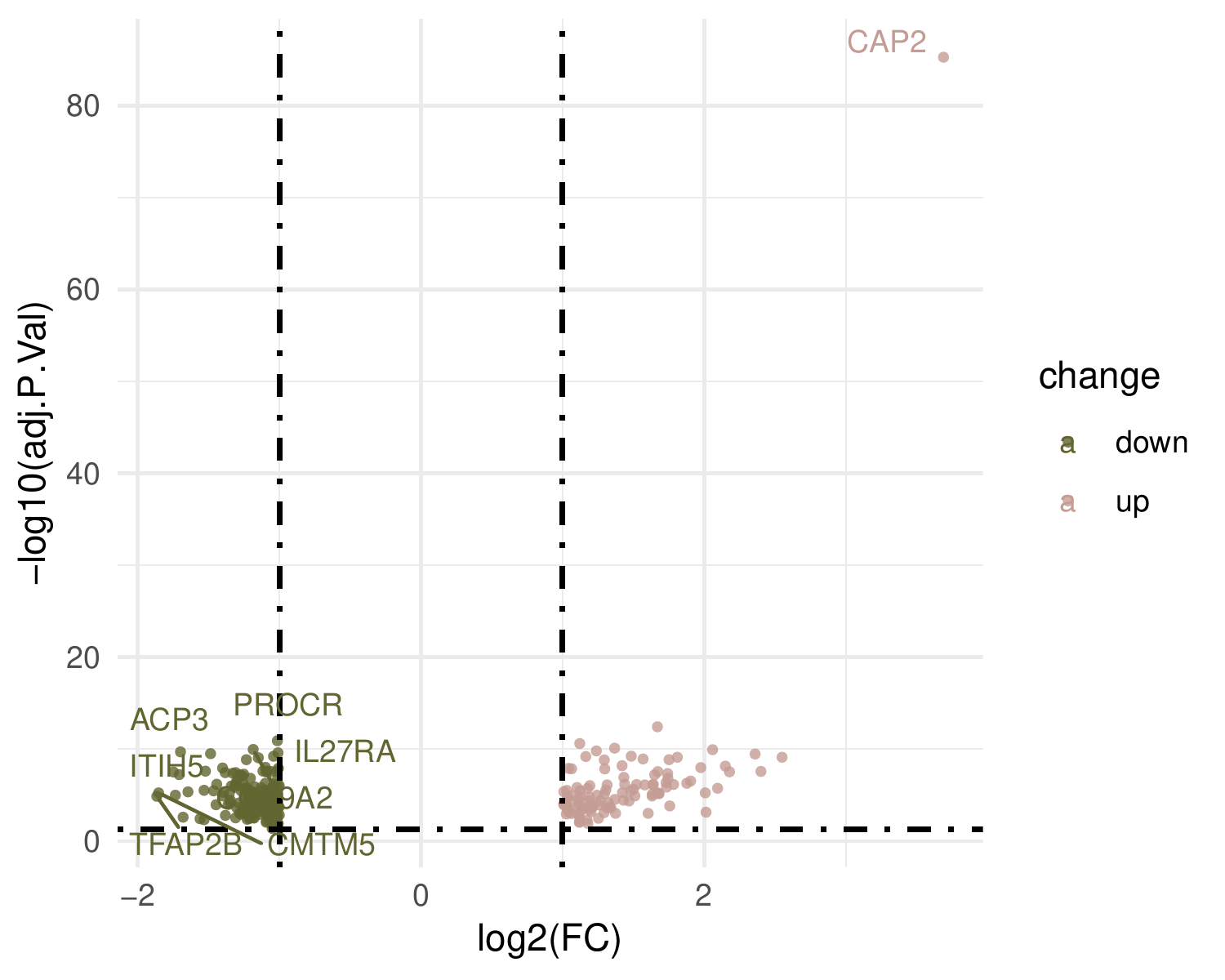
**Fig.** **3** TCGA SKCM survival curve of CAP2

**(File path: Figure+Table/3.4\_Survival\_生存分析\_(TCGA\_SKCM)/TCGA-SKCM-survival-curve-of-CAP2.pdf)**

Fig. **[3](#TCGA-SKCM-survival-curve-of-CAP2)** 为 CAP2 生存曲线。

## 3.5 Limma 差异分析 (TCGA\_SKCM)

使用 Survival 分析 (TCGA\_SKCM) 时定义的分组数据 (根据 CAP2 的表达，分为 High (n=236) , Low (n=237) )。差异分析：High vs Low。(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=105) , down (n=162)



**Fig.** **4** TCGA SKCM High vs Low

**(File path: Figure+Table/3.5\_Limma\_差异分析\_(TCGA\_SKCM)/TCGA-SKCM-High-vs-Low.pdf)**

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

**(See: Figure+Table/3.5\_Limma\_差异分析\_(TCGA\_SKCM)/TCGA-SKCM-High-vs-Low-content)**

Fig. **[4](#TCGA-SKCM-High-vs-Low)** 为 High - Low 差异分析火山图。

**Tab.** **2** TCGA SKCM data High vs Low

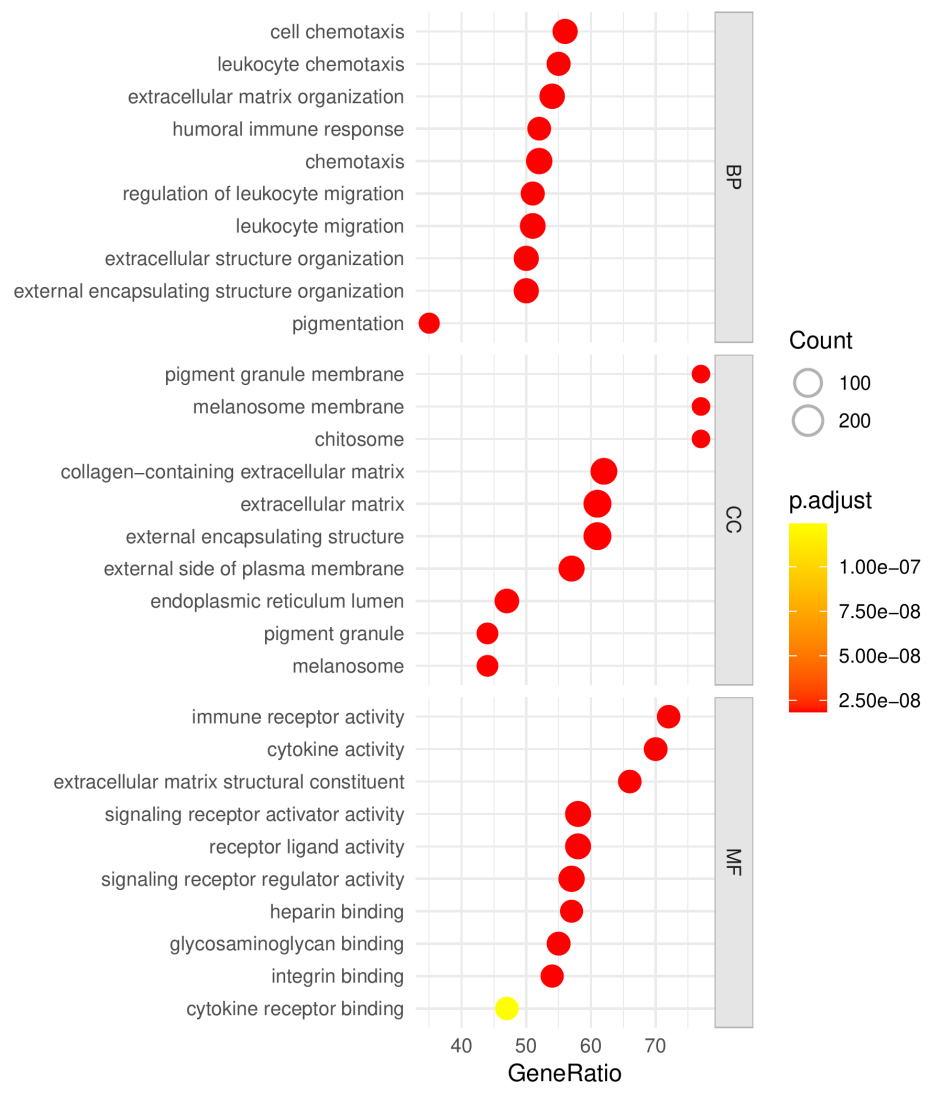
| Gene name | LogFC | Adj.P.Val | Rownames | Gene id |
| --- | --- | --- | --- | --- |
| CAP2 | 3.69 | 5.477e-86 | ENSG00000112186.12 | ENSG00000112186.12 |
| SOX6 | 1.669 | 3.727e-13 | ENSG00000110693.18 | ENSG00000110693.18 |
| PROCR | -1.014 | 1.238e-11 | ENSG00000101000.6 | ENSG00000101000.6 |
| RAB6B | 1.121 | 2.503e-11 | ENSG00000154917.11 | ENSG00000154917.11 |
| AL138885.3 | 1.367 | 7.912e-11 | ENSG00000285759.1 | ENSG00000285759.1 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.5\_Limma\_差异分析\_(TCGA\_SKCM)/TCGA-SKCM-data-High-vs-Low.csv)**

Tab. **[2](#TCGA-SKCM-data-High-vs-Low)** 为 High - Low 差异分析统计表格。

## 3.6 ClusterProfiler GSEA 富集分析 (TCGA\_SKCM)

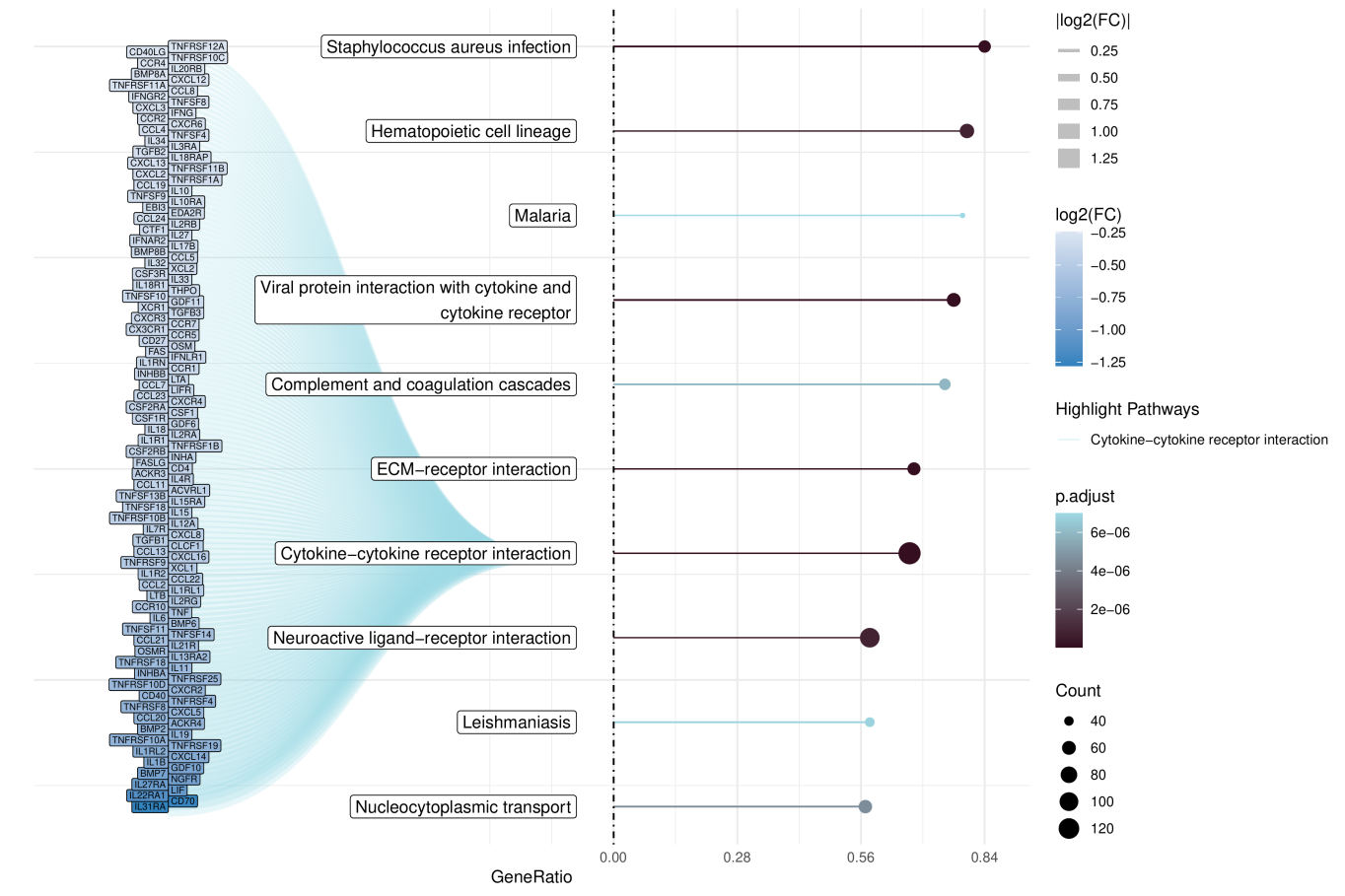
对**基因集** (CAP2, E2F3, SOX6, …[n = 20257], 来自于Limma 差异分析[Section: TCGA\_SKCM]) 进行ClusterProfiler GSEA 富集分析。以 KEGG、GO 数据集，对基因集富集分析。以 msigdbr 获取 H (hallmark gene sets) 基因集。使用 H 数据集, 以 clusterProfiler::GSEA 对 CAP2, ABCB5, TRPM1, …(n = 17374) 富集分析。



**Fig.** **5** TCGA SKCM GSEA GO enrichment

**(File path: Figure+Table/3.6\_ClusterProfiler\_GSEA\_富集分析\_(TCGA\_SKCM)/TCGA-SKCM-GSEA-GO-enrichment.pdf)**

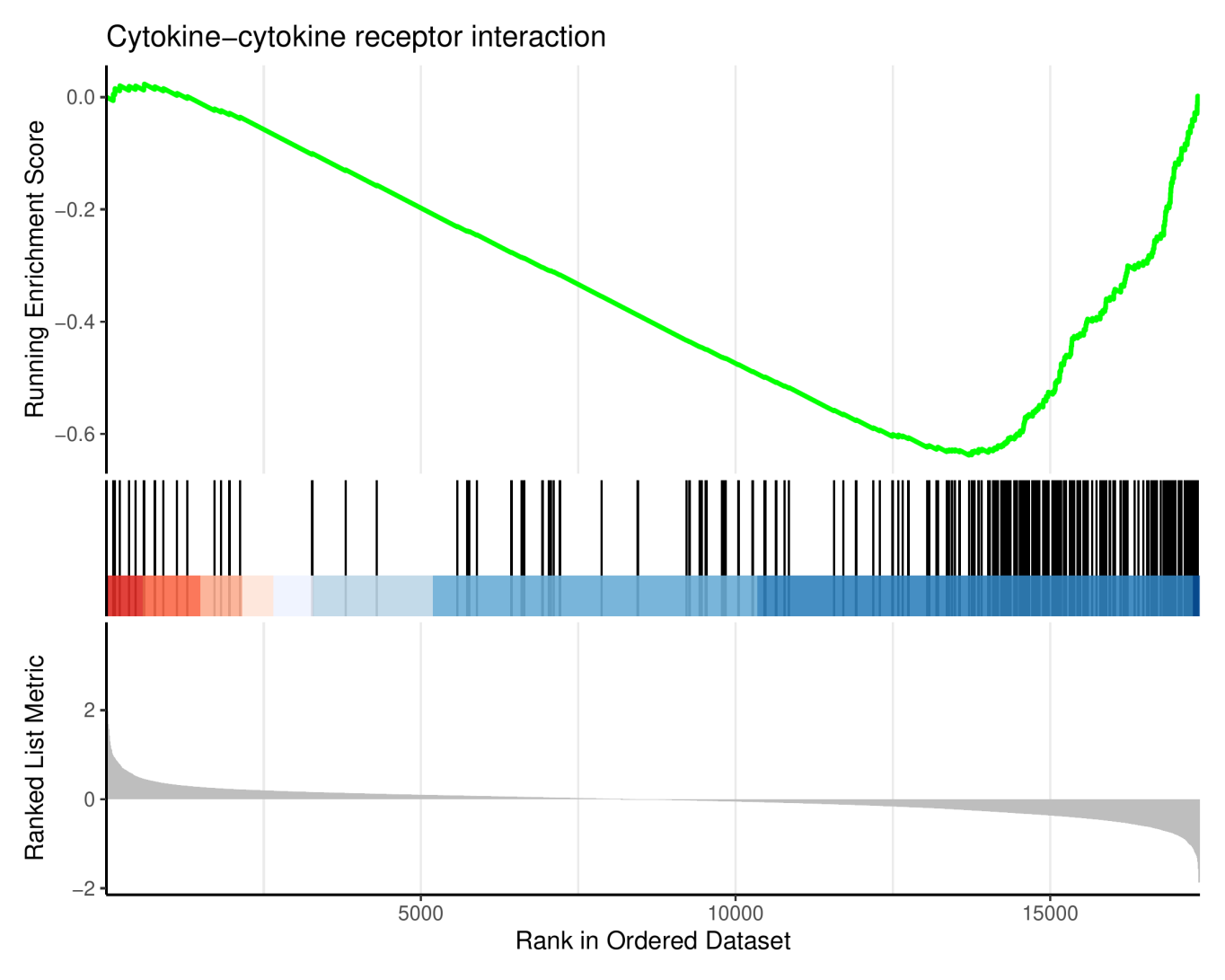
Fig. **[5](#TCGA-SKCM-GSEA-GO-enrichment)** GSEA GO 富集分析气泡图。



**Fig.** **6** TCGA SKCM KEGG enrichment with enriched genes

**(File path: Figure+Table/3.6\_ClusterProfiler\_GSEA\_富集分析\_(TCGA\_SKCM)/TCGA-SKCM-KEGG-enrichment-with-enriched-genes.pdf)**

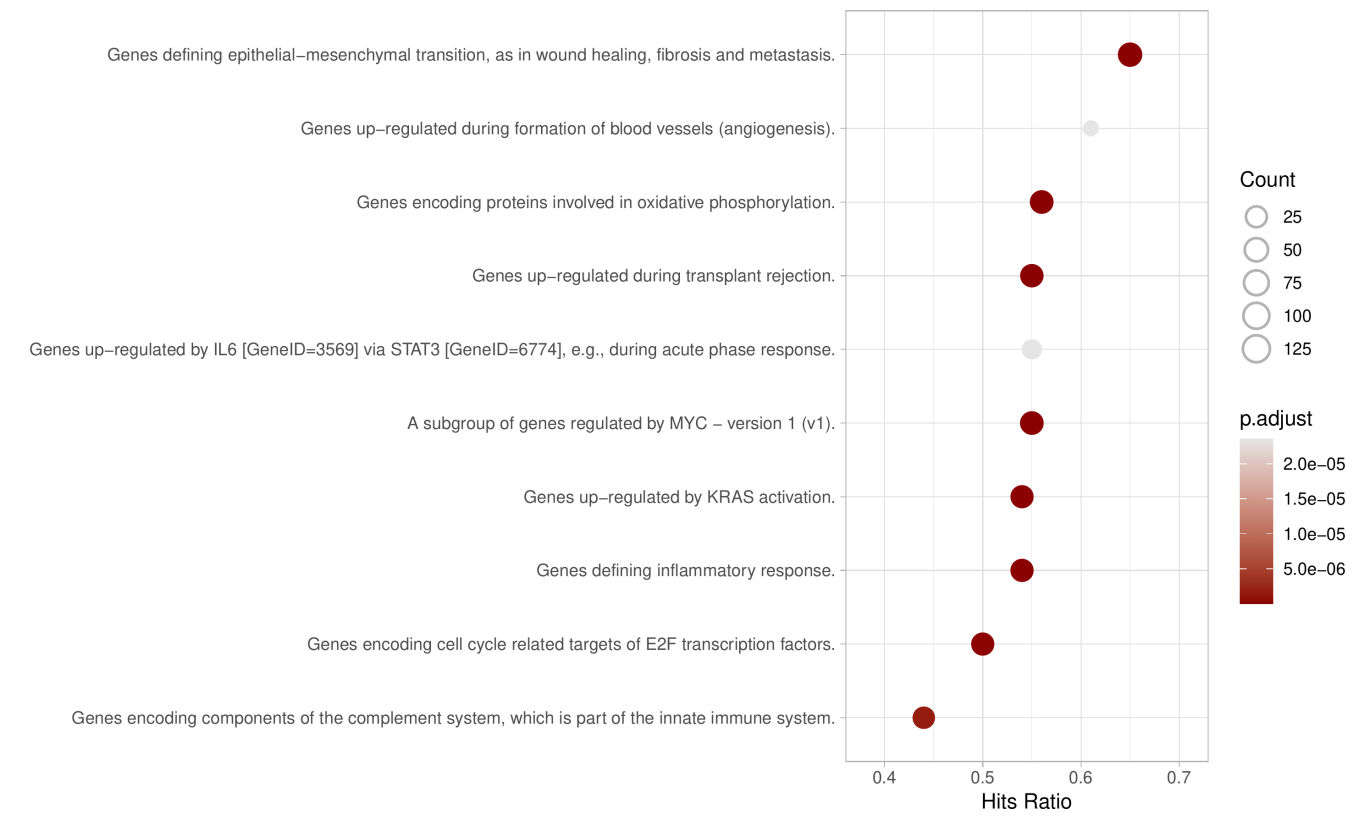
Fig. **[6](#TCGA-SKCM-KEGG-enrichment-with-enriched-genes)** 为通路富集图，兼部分通路的富集基因表达。



**Fig.** **7** TCGA SKCM GSEA plot of the pathways

**(File path: Figure+Table/3.6\_ClusterProfiler\_GSEA\_富集分析\_(TCGA\_SKCM)/TCGA-SKCM-GSEA-plot-of-the-pathways.pdf)**

Fig. **[7](#TCGA-SKCM-GSEA-plot-of-the-pathways)** 为 GSEA KEGG Cytokine-cytokine receptor interaction 富集条码图 (以 clusterProfiler 仿 GSEA 软件绘图)。



**Fig.** **8** TCGA SKCM GSEA pathway list of H

**(File path: Figure+Table/3.6\_ClusterProfiler\_GSEA\_富集分析\_(TCGA\_SKCM)/TCGA-SKCM-GSEA-pathway-list-of-H.pdf)**

Fig. **[8](#TCGA-SKCM-GSEA-pathway-list-of-H)** 为 GSEA 按 H (hallmark gene sets) 数据集富集图。

**Tab.** **3** GSEA pathway list of H data

| ID | Description | SetSize | EnrichmentScore | NES |
| --- | --- | --- | --- | --- |
| M5930 | Genes defining ep... | 197 | -0.6242 | -2.243 |
| M5950 | Genes up-regulate... | 189 | -0.602 | -2.156 |
| M5932 | Genes defining in... | 185 | -0.5849 | -2.09 |
| M5953 | Genes up-regulate... | 182 | -0.5597 | -1.996 |
| M5926 | A subgroup of gen... | 199 | 0.4595 | 2.103 |
| ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.6\_ClusterProfiler\_GSEA\_富集分析\_(TCGA\_SKCM)/GSEA-pathway-list-of-H-data.xlsx)**

Tab. **[3](#GSEA-pathway-list-of-H-data)** 为 GSEA 按 H (hallmark gene sets) 数据集富集附表。

# 4 总结

按 CAP2 表达分组，筛选与 CAP2 显著相关的通路，GSEA 的 KEGG、GO 富集见 Fig. **[6](#TCGA-SKCM-KEGG-enrichment-with-enriched-genes)** Fig. **[5](#TCGA-SKCM-GSEA-GO-enrichment)** 。 hallmark 基因集富集，见Fig. **[8](#TCGA-SKCM-GSEA-pathway-list-of-H)**

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

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