**生信分析报告**

**项目标题： CAP2通过靶向EMT信号通路促进黑色素瘤进展 ;**

**单 号： BSHQ240813 ;**

**分析人员： 黄礼闯 ;**

**分析类型： 生信分析 ;**

**委 托 人： 王凯 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

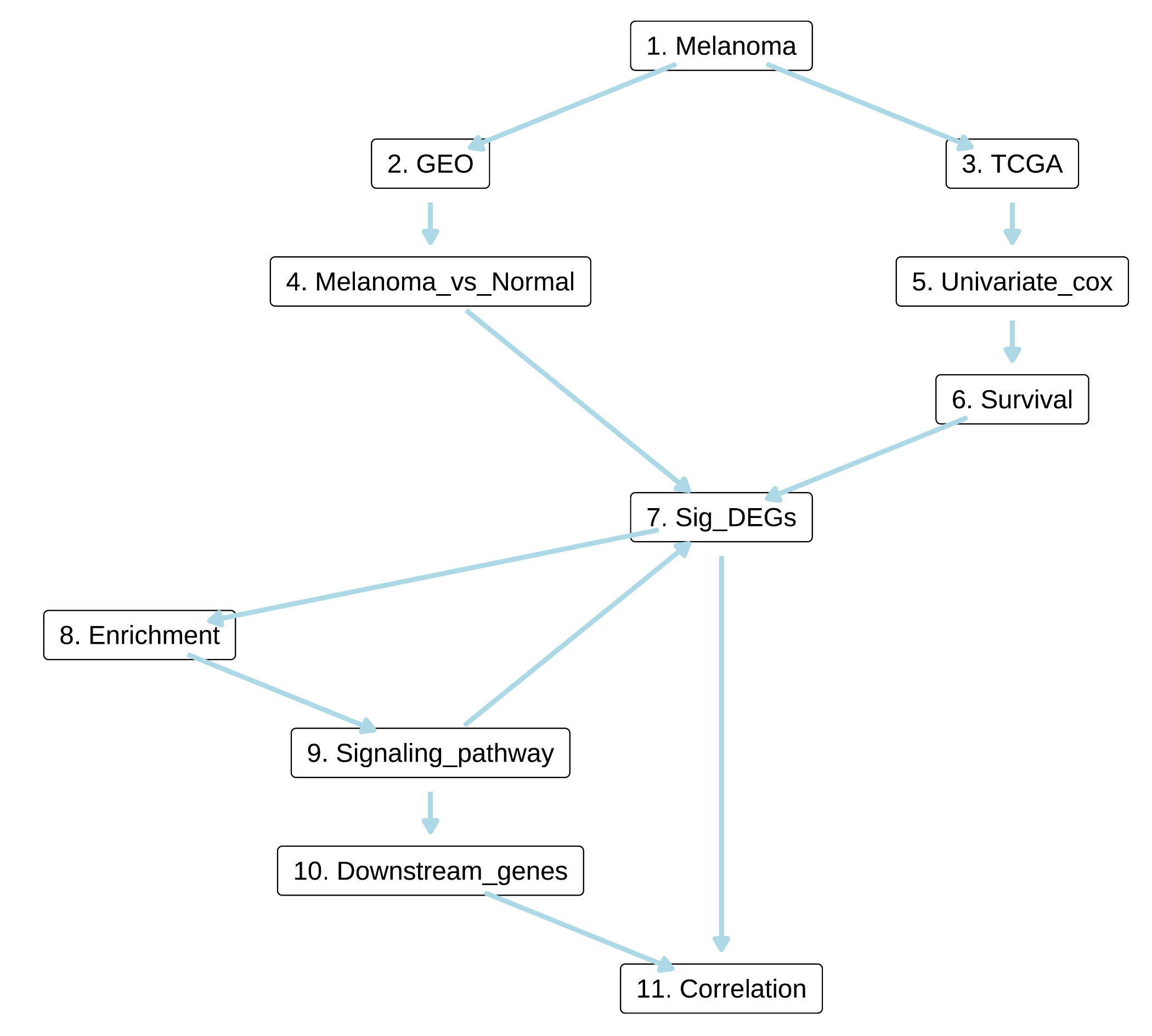
# 1 分析流程

## 1.1 需求

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

## 1.2 实际分析

因为 CAP2 未找到符合要求的数据集，改换寻找其他满足条件的基因，并寻找下游通路。



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

**(File path: Figure+Table/1.2\_实际分析/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 GEO 数据获取 (Dataset: GSE11907)

以 R 包 GEOquery (2.74.0) 获取 GSE11907 数据集。

## 2.3 Limma 差异分析 (Dataset: GSE11907)

使用 log2 和 limma::normalizeBetweenArrays 对数据标准化。 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：Melanoma vs Healthy。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 (2.34.0) (2015, **IF:16.6**, Q1, Nucleic Acids Research)2 获取 TCGA-SKCM 数据集。

## 2.5 COX 回归 (Dataset: TCGA\_SKCM)

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。

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

以 R 包 edgeR (4.4.0) ()3 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 (3.7.0) 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。

## 2.7 富集分析 (Dataset: SIG)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)4进行 KEGG 和 GO 富集分析。以 p.adjust 表示显著水平。

## 2.8 Pathview 通路可视化 (Dataset: SIG)

以 R 包 pathview (2013, Bioinformatics (Oxford, England))5 将 clusterProfiler 富集的 KEGG 通路可视化。

# 3 分析结果

## 3.1 GEO 数据获取 (GSE11907)

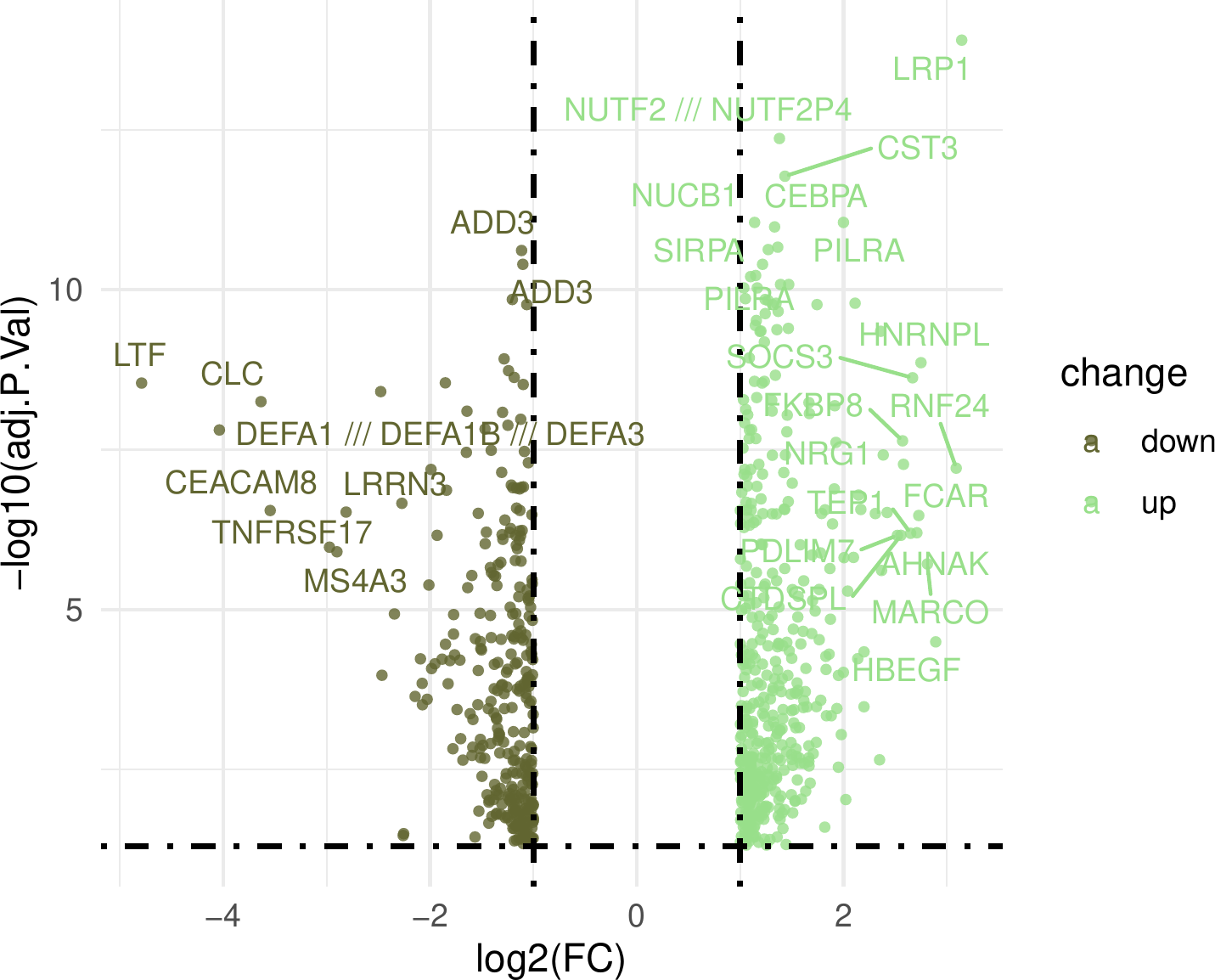
以 GEOquery 获取 GSE11907 的数据信息。

* Data Source ID: GSE11907
* data\_processing: The data were analyzed with Microarray Suite version 5.0 (MAS 5.0) using Affymetrix default analysis settings and global scaling as normalization method. The trimmed mean target intensity of each array was arbitrarily set to 500.

**(见Figure+Table/3.1\_GEO\_数据获取\_(GSE11907)/GSE11907-GSE11907-content)**

## 3.2 Limma 差异分析 (GSE11907)

匹配 group 中包含 “Healthy|Melanoma” 字符的数据，最终得到 51 例数据。匹配 group 中包含 “Healthy|Melanoma” 字符的数据，未过滤掉任何数据 (n = 51)。样本分组：Healthy (n=12) , Melanoma (n=39) , Others (n=0) 。以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 。差异分析：Melanoma vs Healthy。(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=392) , down (n=245)



**Fig.** **2** GSE11907 Melanoma vs Healthy

**(File path: Figure+Table/3.2\_Limma\_差异分析\_(GSE11907)/GSE11907-Melanoma-vs-Healthy.pdf)**

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

**(See: Figure+Table/3.2\_Limma\_差异分析\_(GSE11907)/GSE11907-Melanoma-vs-Healthy-content)**

## 3.3 TCGA 数据获取 (TCGA\_SKCM)

获取 TCGA-SKCM 数据。

## 3.4 COX 回归 (TCGA\_SKCM)

将基因集 (Melanoma - Healthy, 来自于Limma 差异分析(Section: GSE11907)) 用于 COX 回归分析。共 576 个基因在数据集 TCGA-SKCM 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=244) , Dead (n=218) )。执行单因素 COX 回归，筛选 P 值 < 0.05，共筛选到 21 个基因。 Tab. **[1](#TCGA-SKCM-sig-Univariate-Cox-Coefficients)** 。

**Tab.** **1** TCGA SKCM sig Univariate Cox Coefficients

| Feature | Coef | Exp(coef) | Se(coef) | Z | Pvalue | P.adjust |
| --- | --- | --- | --- | --- | --- | --- |
| ACSM3 | -0.1052 | 0.9002 | 0.04506 | -2.334 | 0.0196 | 0.936 |
| ADRB1 | -0.05621 | 0.9453 | 0.027 | -2.082 | 0.03734 | 0.936 |
| SYT1 | -0.06554 | 0.9366 | 0.02583 | -2.537 | 0.01117 | 0.936 |
| NME8 | -0.06992 | 0.9325 | 0.03384 | -2.066 | 0.03879 | 0.936 |
| CLC | -0.122 | 0.8851 | 0.05348 | -2.281 | 0.02253 | 0.936 |
| ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/3.4\_COX\_回归\_(TCGA\_SKCM)/TCGA-SKCM-sig-Univariate-Cox-Coefficients.csv)**

## 3.5 Survival 生存分析 (TCGA\_SKCM)

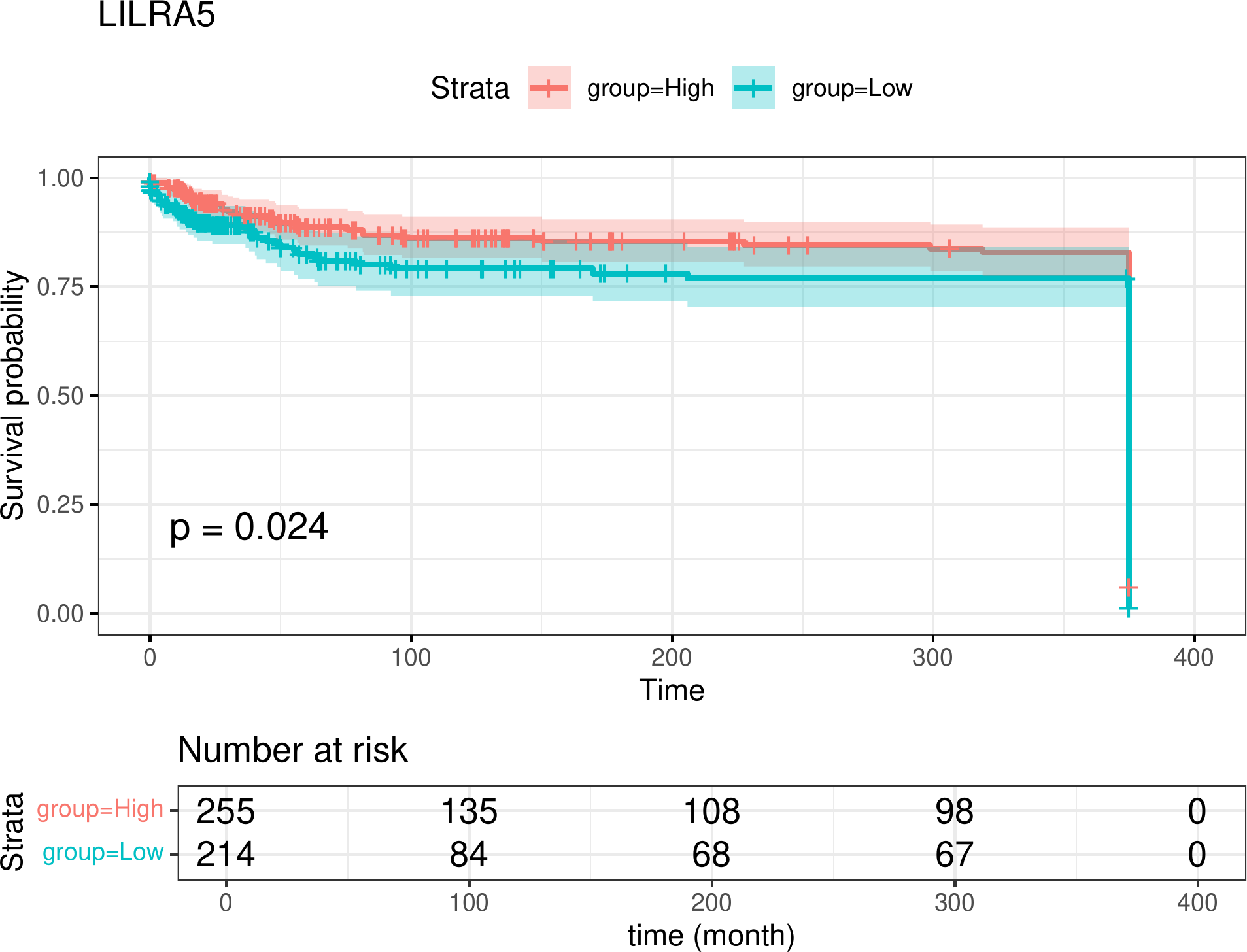
对基因集 (ACSM3, ADRB1, etc., 来自于COX 回归(Section: TCGA\_SKCM)) 进行Survival 生存分析。按 survminer::surv\_cutpoint 计算的 cutoff，将样本分为 Low 和 High 风险组。生存数据为TCGA-SKCM，使用标准化过的基因表达数据。根据元数据信息 (即临床数据) ，去除了生存状态未知的样例。根据 P value < 0.05, 共筛到 13 个特征。 分别为 ACSM3, SYT1, CLC, MS4A4A, ECE1, HIVEP3, MYBPH, FOLH1B, DSC1, CLIC2, DEFA4, DLK2, LILRA5。

Tab. **[2](#TCGA-SKCM-Significant-Survival-PValue)** 。 Fig. **[3](#TCGA-SKCM-survival-curve-of-LILRA5)** 为 LILRA5 生存曲线。

**Tab.** **2** TCGA SKCM Significant Survival PValue

| Name | Pvalue |
| --- | --- |
| ACSM3 | 1.735e-06 |
| SYT1 | 0.02008 |
| CLC | 0.01507 |
| MS4A4A | 0.0001311 |
| ECE1 | 0.02166 |
| ... | ... |

**(File path: Figure+Table/3.5\_Survival\_生存分析\_(TCGA\_SKCM)/TCGA-SKCM-Significant-Survival-PValue.csv)**



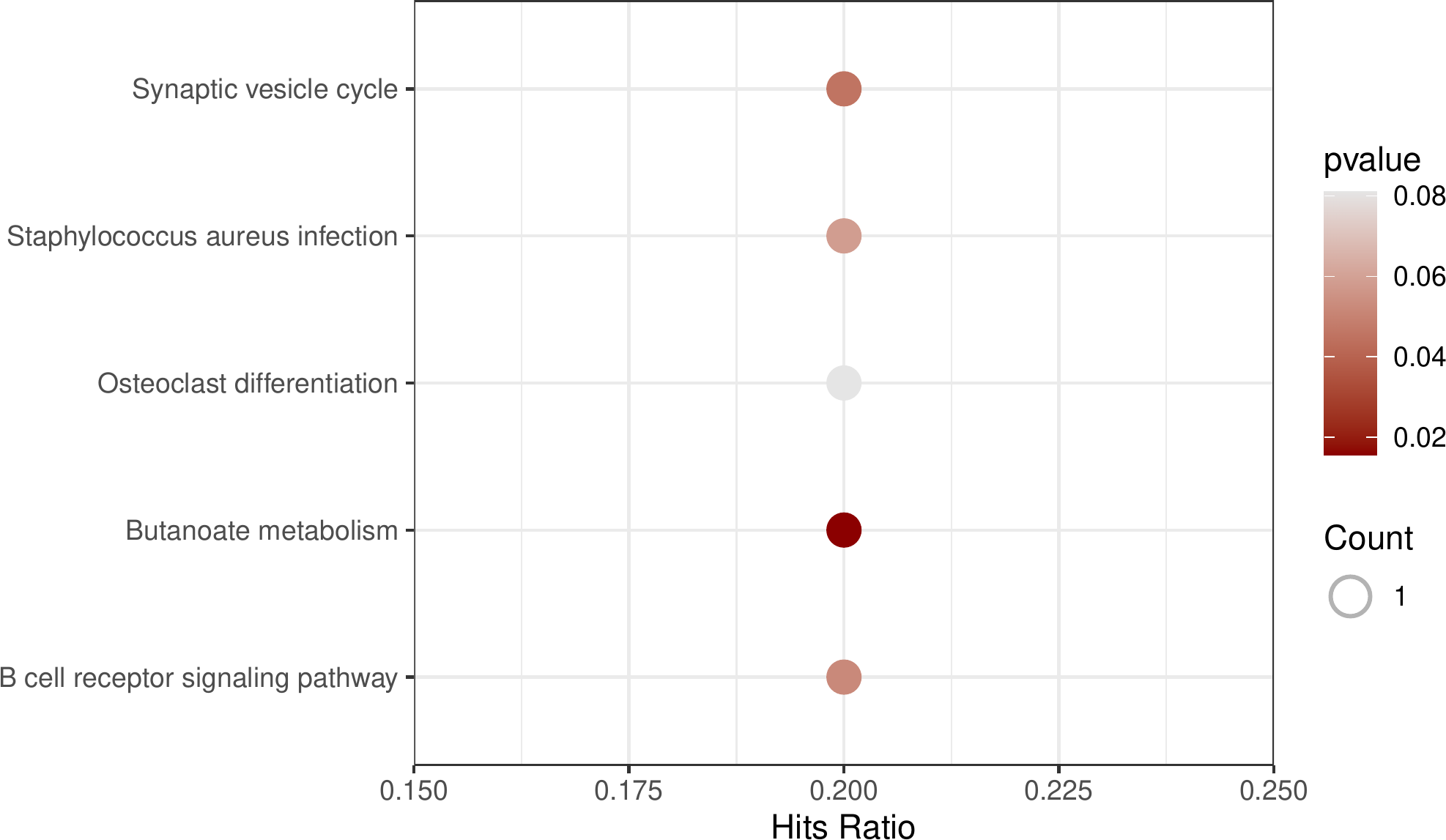
**Fig.** **3** TCGA SKCM survival curve of LILRA5

**(File path: Figure+Table/3.5\_Survival\_生存分析\_(TCGA\_SKCM)/TCGA-SKCM-survival-curve-of-LILRA5.pdf)**

## 3.6 ClusterProfiler 富集分析 (SIG)

对基因集 (ACSM3, SYT1, etc., 来自于Survival 生存分析(Section: TCGA\_SKCM)) 进行富集分析。

Fig. **[4](#SIG-KEGG-enrichment)** , 观察到， B cell receptor signaling pathway 富集到一个基因，为信号通路 (P-value: 0.0522，其实这里 P 值不重要，主要是找到基因所在的通路)。



**Fig.** **4** SIG KEGG enrichment

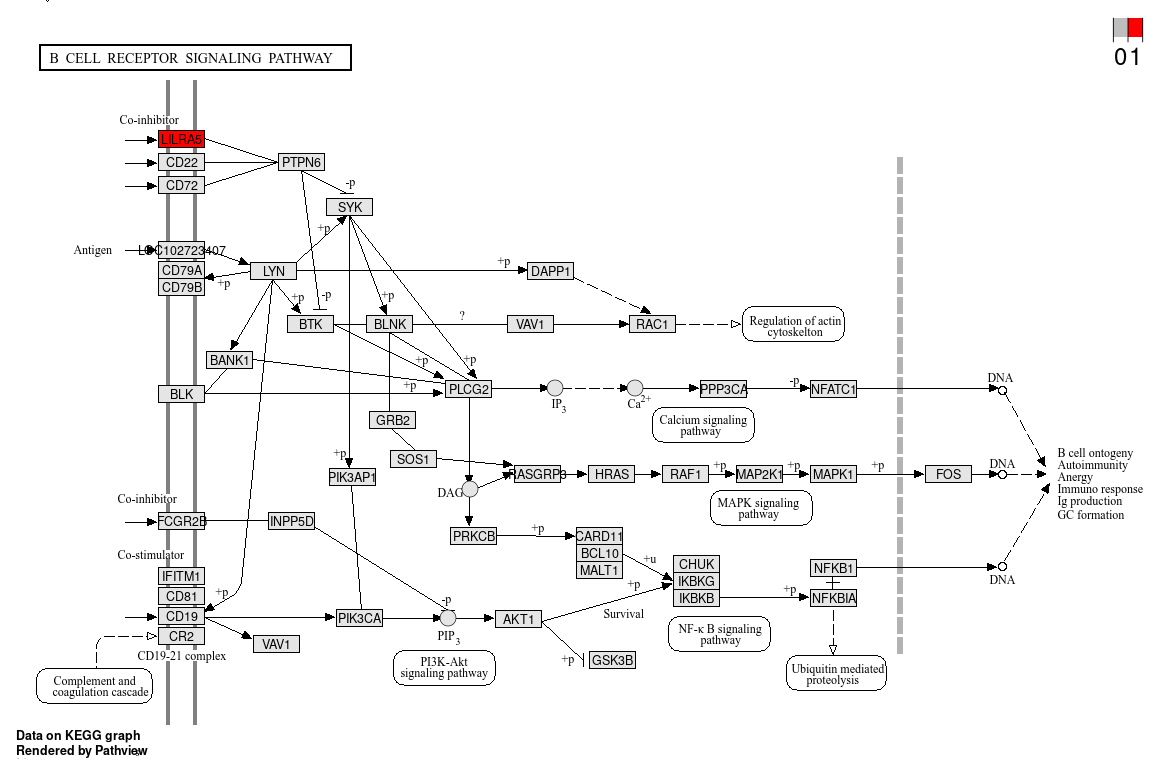
**(File path: Figure+Table/3.6\_ClusterProfiler\_富集分析\_(SIG)/SIG-KEGG-enrichment.pdf)**

## 3.7 Pathview 通路可视化 (SIG)

以 pathview 探究基因集在通路 hsa04662 中的上下游关系。

Fig. **[5](#SIG-pathviews)** KEGG 通路可视化 (hsa04662) 展示了富集基因在该通路的上下游关系。通路图中的基因的映射颜色表示是否显著富集。 富集的基因为 LILRA5。

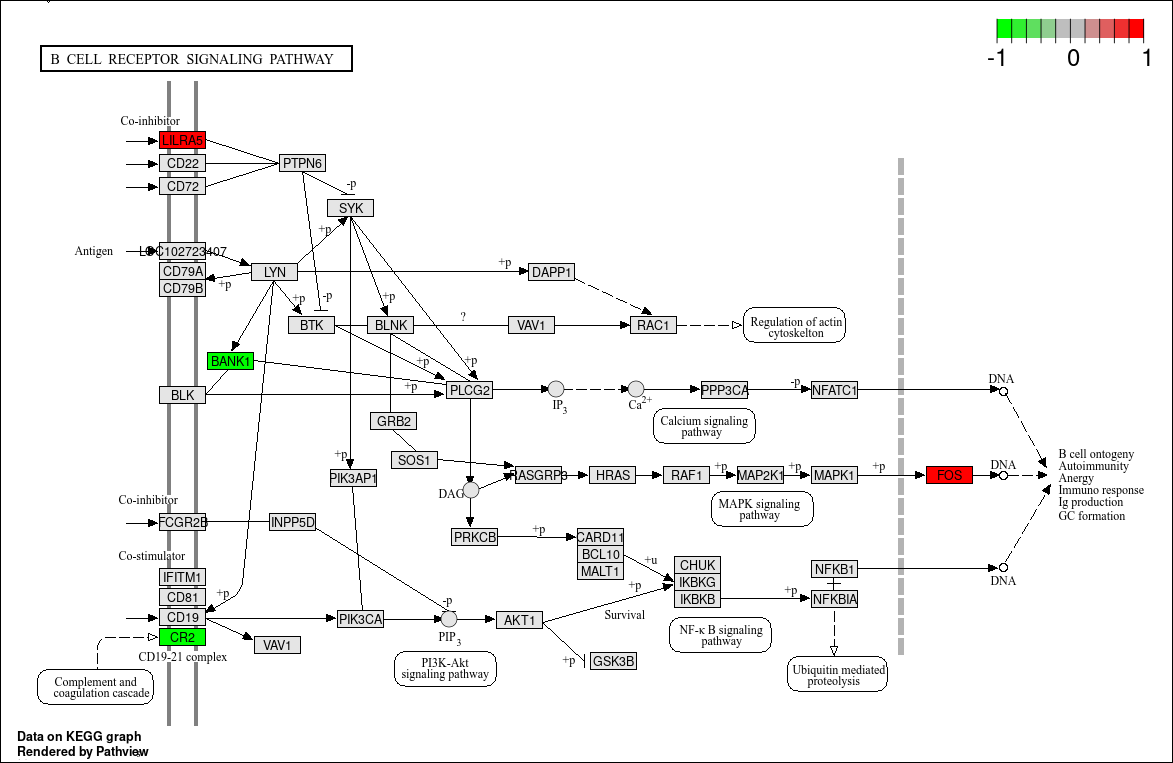
Fig. **[6](#SIG2-pathviews)** , 发现下游基因 FOS。



**Fig.** **5** SIG pathviews

**(File path: Figure+Table/3.7\_Pathview\_通路可视化\_(SIG)/SIG-pathviews.png)**

* Interactive figure: <https://www.genome.jp/pathway/hsa04662>
* Enriched genes: LILRA5



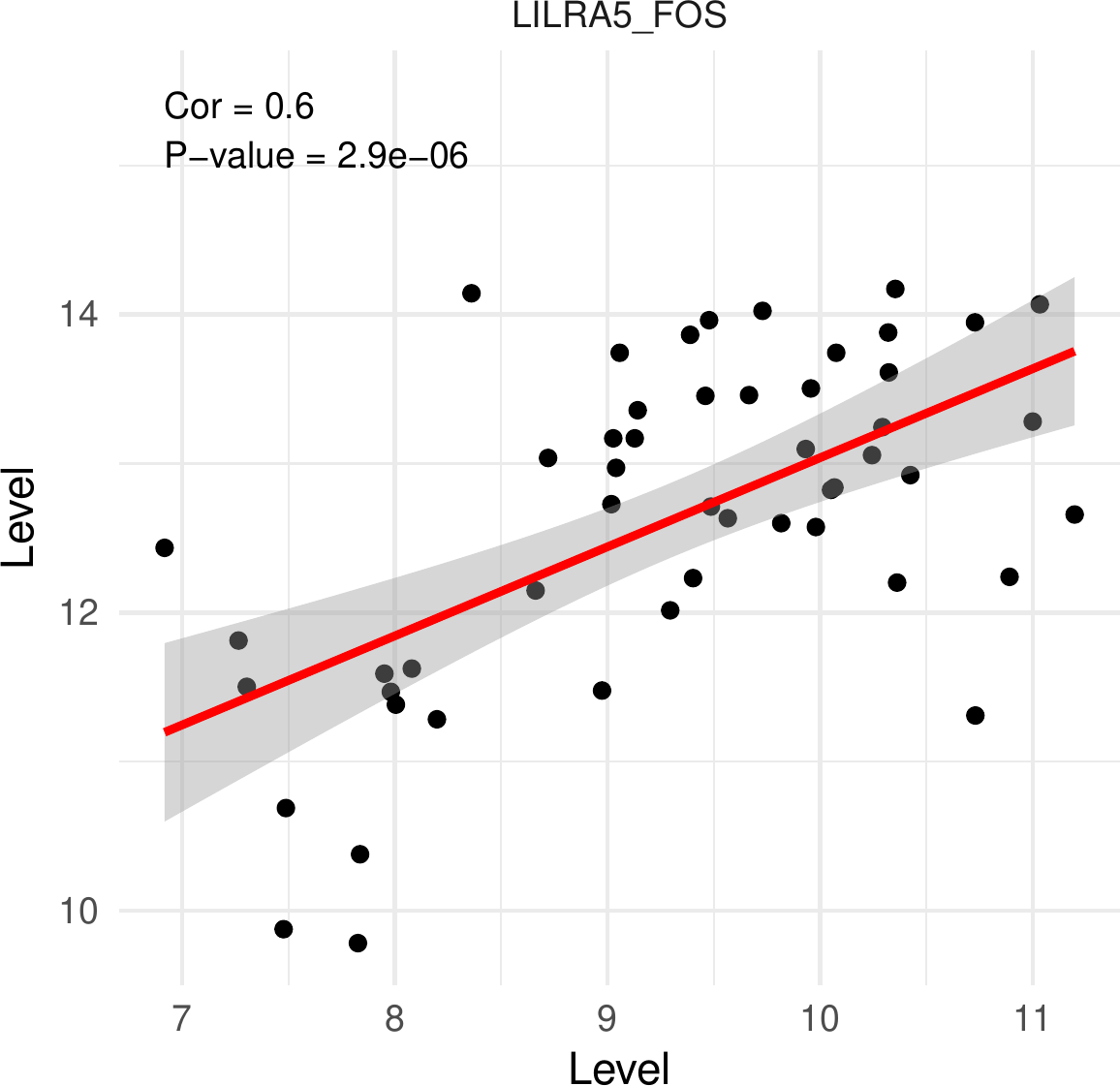
**Fig.** **6** SIG2 pathviews

**(File path: Figure+Table/3.7\_Pathview\_通路可视化\_(SIG)/SIG2-pathviews.png)**

* Interactive figure: <https://www.genome.jp/pathway/hsa04662>
* Enriched genes: CR2, FOS, LILRA5, BANK1

## 3.8 关联分析 (GSE11907)

将基因 (LILRA5 -> FOS) 关联分析， Fig. **[7](#Linear-regression)** ，两者显著关联。



**Fig.** **7** Linear regression

**(File path: Figure+Table/3.8\_关联分析\_(GSE11907)/Linear-regression.pdf)**

# 4 总结

筛选的基因为 LILRA5, 相关通路为 B cell receptor signaling pathway， 下游通路为Fig. **[6](#SIG2-pathviews)** KEGG 通路可视化 (hsa04662) 展示了富集基因在该通路的上下游关系。通路图中的基因的映射颜色表示基因显著富集，并在数据集中有上调或下调变化趋势。 中的 FOS，两者显著关联， 见Fig. **[7](#Linear-regression)** 。

FOS 在多数 Melanoma 的文献中被报导，例如 (2021, **IF:3.9**, Q1, Journal of photochemistry and photobiology. B, Biology)6。 而 LILRA5 还未被报道。

HPA 数据库未包含 LILRA5 的 Melanoma 组织染色等数据，仅 TCGA 表达量数据。未进一步分析。

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

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

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

5. Luo, W. & Brouwer, C. Pathview: An r/bioconductor package for pathway-based data integration and visualization. *Bioinformatics (Oxford, England)* **29**, 1830–1831 (2013).

6. Shao, J. *et al.* The role of fos-mediated autophagy activation in the indocyanine green-based photodynamic therapy for treating melanoma. *Journal of photochemistry and photobiology. B, Biology* **214**, (2021).