**生信分析报告**

**项目标题： 膀胱癌 ;**

**单 号： BSHQ230805 ;**

**分析人员： 黄礼闯 ;**

**分析类型： 生信分析 ;**

**委 托 人： 刘炀 ;**

**受 托 人： 杭州铂赛生物科技有限公司 .**

# 1 分析流程

## 1.1 需求

### 1.1.1 第一部分

利用生物信息学筛选出与CLSPN联系最紧密的下游信号通路或机制。（参考PMID: 35149175，2022，IF9.1，Cancer Letters；PMID: 36627634，2023，11.4，Journal of Experimental & Clinical Cancer Research）

1. 根据CLSPN表达将TCGA-BLCA患者分为CLSPN高表达组和CLSPN低表达组
2. 利用GSVA富集分析CLSPN高表达组表达差异显著的通路（选择差异最显著的前3个通路）
3. 筛选出的前三条差异基因通路中有显著差异的基因的折叠变化进行了计算和排序，找出与CLSPN低表达组相比差异基因倍数变化最大的通路。
4. 对 3 中筛出的差异表达变化最显著的通路中显著上调的基因进行分类富集，选择较多基因富集的通路后续研究通路

### 1.1.2 第二部分

为了探究CLSPN在膀胱癌中的相互作用蛋白，我们利用不同在线软件对CLSPN可能的互作蛋白进行了探究。 分析

1. 通过PPI，STRING，Genemania等网站预测差异基因中具有与CLSPN互作性的蛋白，明确直接与CLSPN相互作用蛋白P
2. 利用TCGA数据库分析蛋白P在OS中的表达、相关性及与预后的关系

# 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 TCGA 数据获取 (Dataset: BLCA)

以 R 包 TCGAbiolinks (2.34.0) (2015, **IF:16.6**, Q1, Nucleic Acids Research)1 获取 TCGA-BLCA 数据集。

## 2.3 Limma 差异分析 (Dataset: BLCA)

以 R 包 limma (3.62.1) (2005, **IF:**, , )2 edgeR (4.4.0) (, **IF:**, , )3 进行差异分析。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)。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。随后，以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 用于线性分析。 使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 差异分析对比组：High vs Low。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。

## 2.4 ClusterProfiler GSEA 富集分析 (Dataset: BLCA)

以 ClusterProfiler R 包 (4.15.0.2) (2021, **IF:33.2**, Q1, The Innovation)4 按 GSVA 算法 (clusterProfiler::gseGO, ClusterProfiler::gseKEGG)，进行 KEGG 和 GO 富集分析。

## 2.5 富集分析 (Dataset: CELL)

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 pathview R 包 (1.46.0) 对选择的 KEGG 通路可视化。

## 2.6 Limma 差异分析 (Dataset: BLCA\_EX)

以 R 包 limma (3.62.1) (2005, **IF:**, , )2 edgeR (4.4.0) (, **IF:**, , )3 进行差异分析。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)。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。随后，以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 用于线性分析。 使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 差异分析对比组：Dead vs Alive。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 Limma 差异分析 (Dataset: BLCA\_TUMOR)

以 R 包 limma (3.62.1) (2005, **IF:**, , )2 edgeR (4.4.0) (, **IF:**, , )3 进行差异分析。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)。分析方法参考 <https://bioconductor.org/packages/release/workflows/vignettes/RNAseq123/inst/doc/limmaWorkflow.html>。随后，以 公式 ~ 0 + group 创建设计矩阵 (design matrix) 用于线性分析。 使用 limma::lmFit, limma::contrasts.fit, limma::eBayes 差异分析对比组：tumor vs normal。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。

## 2.8 STRINGdb PPI 分析 (Dataset: CLSPN)

以 R 包 STEINGdb (2.18.0) (2021, **IF:16.6**, Q1, Nucleic Acids Research)5 构建 PPI 网络。数据版本为 12.0，互作类型为 full。以 Cytohubba (2014, **IF:NA**, NA, BMC Systems Biology)6 的算法计算 MCC score (在 R 中计算) 。随后，以 ggraph 可视化网络 (2.2.1)。

## 2.9 Survival 生存分析 (Dataset: COL)

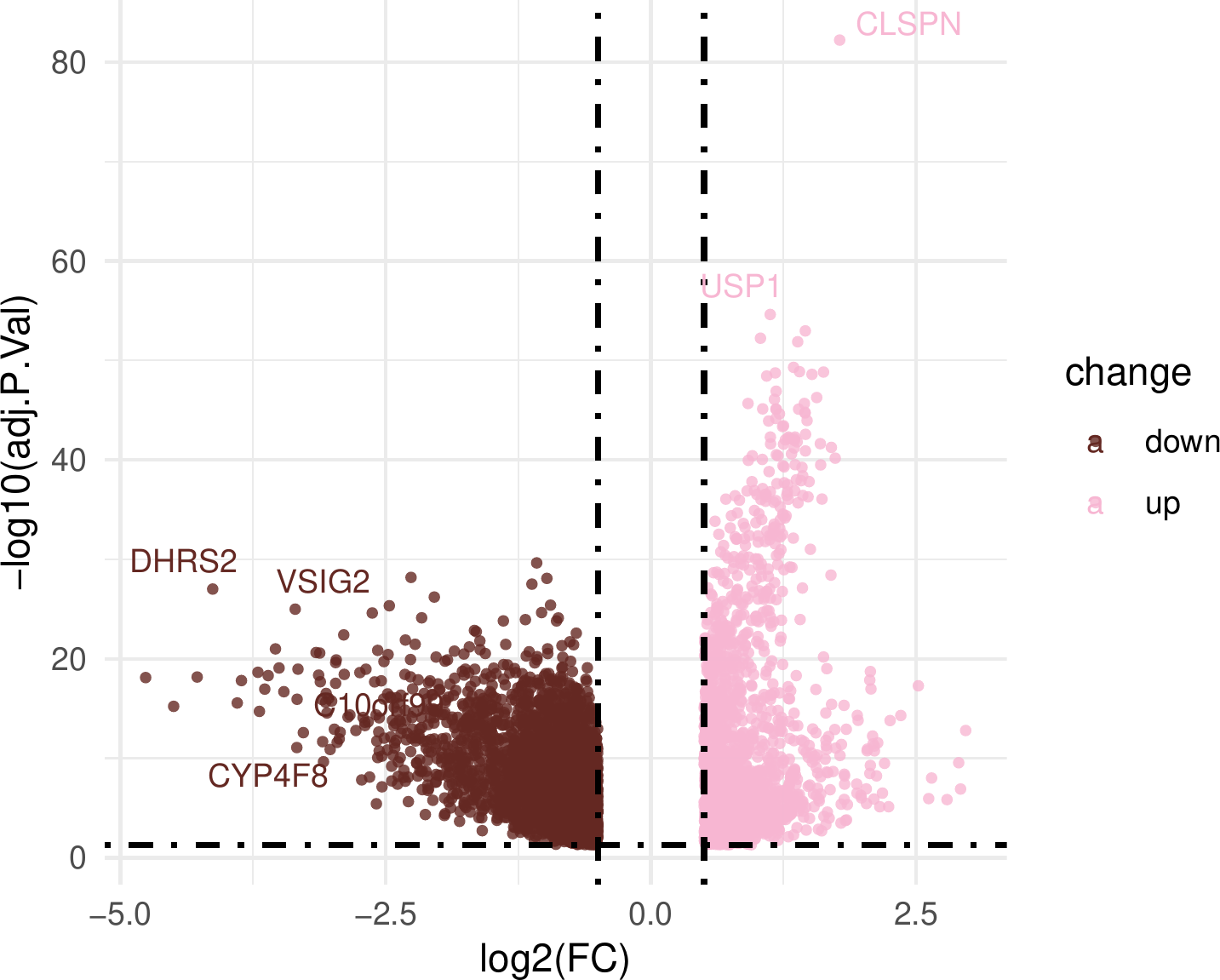
去除了生存状态未知的数据。 以 R 包 survival (3.7.0) 生存分析，以 R 包 survminer (0.5.0) 绘制生存曲线。以 R 包 timeROC (0.4) 绘制 1, 3, 5 年生存曲线。

# 3 分析结果

## 3.1 TCGA 数据获取 (BLCA)

## 3.2 Limma 差异分析 (BLCA)

按 CLSPN 表达量，共 409 个样本，分 2 组，分别为 High (204) , Low (205) 。。 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)， 所有上调 DEGs 有 2131 个，下调共 3353；一共 5484 个 (非重复)。



**Fig.** **1** BLCA High vs Low

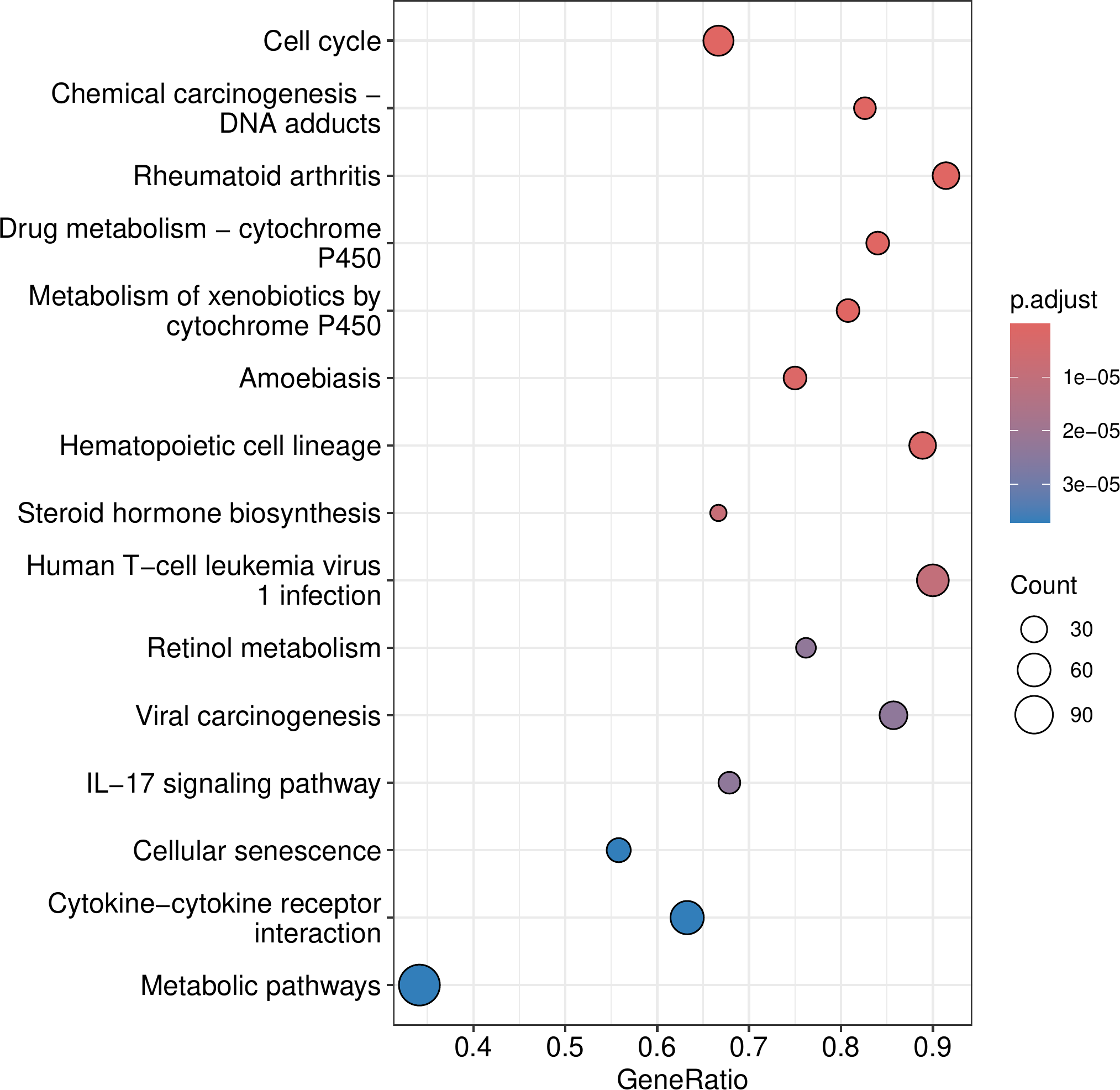
**(File path: Figure+Table/BLCA-High-vs-Low.pdf)**

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

**(See: Figure+Table/BLCA-High-vs-Low-content)**

## 3.3 ClusterProfiler GSEA 富集分析 (BLCA)

以 GSEA 算法，对Fig. **[1](#BLCA-High-vs-Low)** 的基因富集分析。 选择三条最显著的通路，见Tab. **[1](#Used-pathways)**



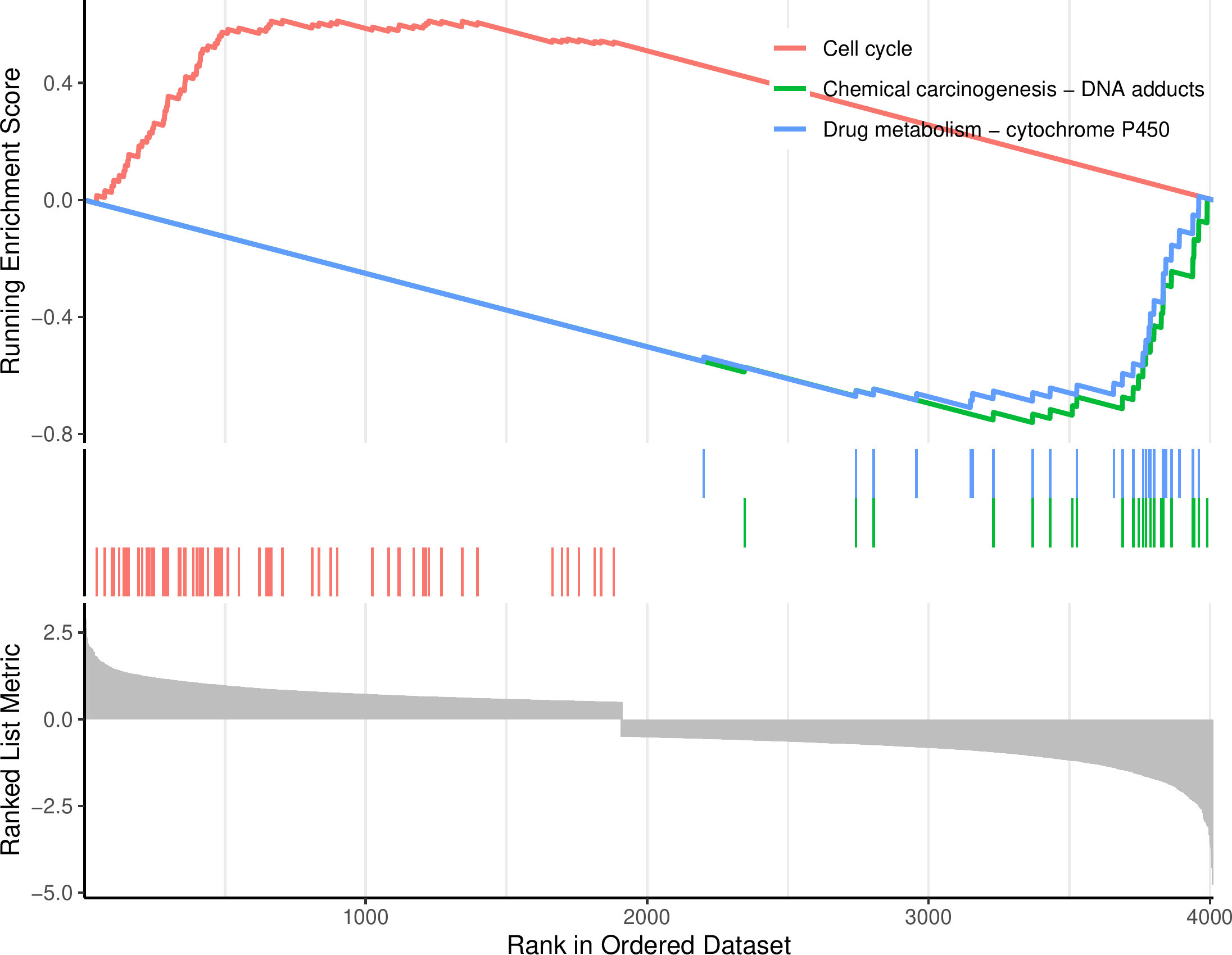
**Fig.** **2** BLCA KEGG enrichment

**(File path: Figure+Table/BLCA-KEGG-enrichment.pdf)**

**Tab.** **1** Used pathways

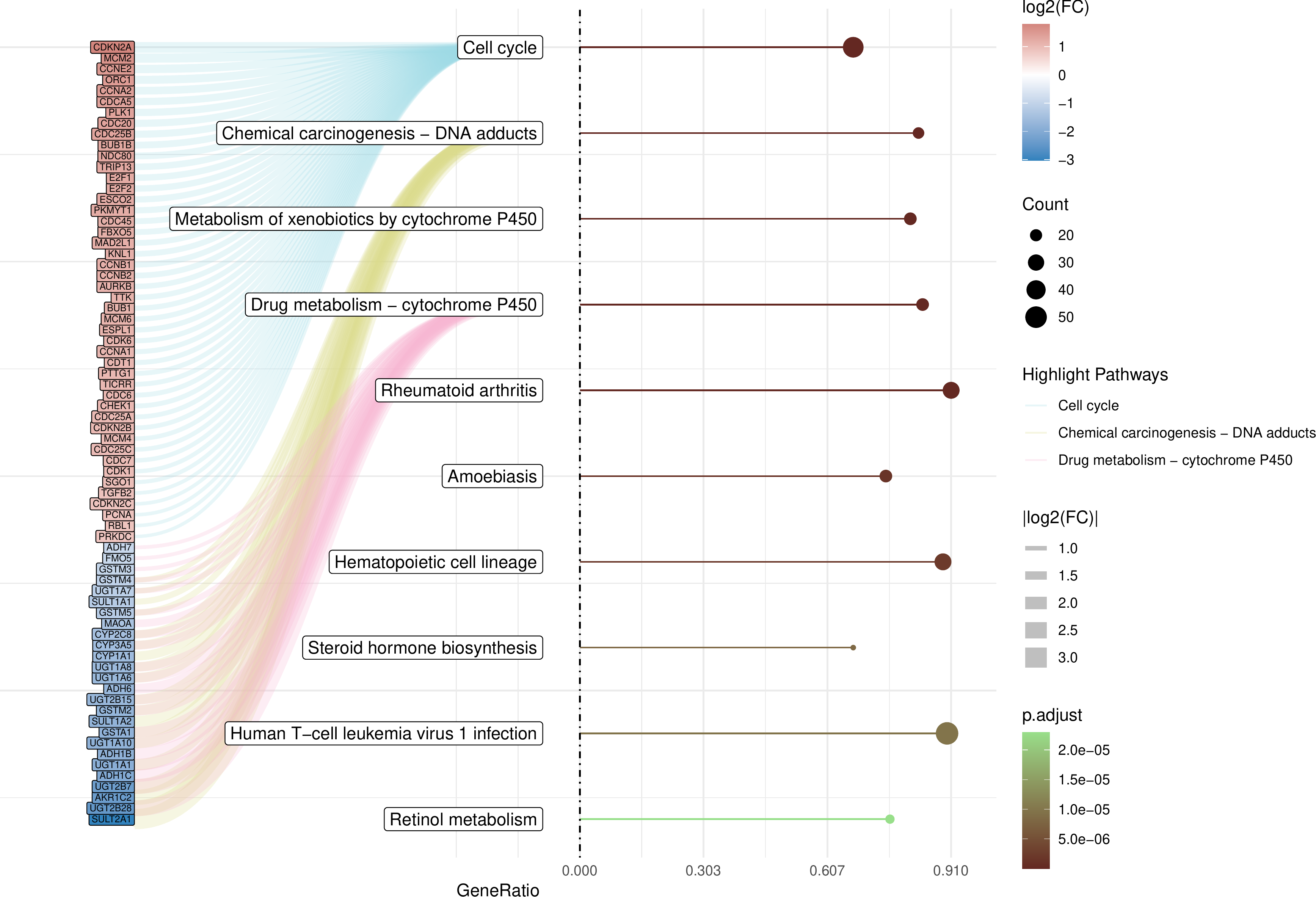
| ID | Descri... | SetSize | Enrich... | NES | Pvalue | P.adjust | Qvalue | Rank | Leadin... |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hsa04110 | Cell c... | 69 | 0.6142... | 4.0675... | 1e-10 | 1.4907... | 9.4759... | 704 | Tags=6... |
| Hsa05204 | Chemic... | 23 | -0.760... | -3.040... | 1.1467... | 1.4907... | 9.4759... | 643 | Tags=8... |
| Hsa00982 | Drug m... | 25 | -0.709... | -2.891... | 2.3980... | 1.3724... | 8.7239... | 864 | Tags=8... |

**(File path: Figure+Table/Used-pathways.xlsx)**



**Fig.** **3** BLCA GSEA plot of the pathways

**(File path: Figure+Table/BLCA-GSEA-plot-of-the-pathways.pdf)**



**Fig.** **4** BLCA KEGG enrichment with enriched genes

**(File path: Figure+Table/BLCA-KEGG-enrichment-with-enriched-genes.pdf)**



**Fig.** **5** BLCA hsa04110 visualization

**(File path: Figure+Table/BLCA-hsa04110-visualization.png)**

* Interactive figure:
* Enriched genes: CDKN2A, MCM2, ORC1, CDK1, BUB1B, PCNA, PLK1, MAD2L1, BUB1, CHEK1, PRKDC, PKMYT1, PTTG1, ESPL1, CDKN2B, TGFB2, CDC7, CDC20, CDC25B, CDC45, CDC6, CDC25A, CCNB1, CCNA2, CDK6, CCNE2, TTK, CDKN2C, RBL1, E2F1, NDC80, AURKB, KNL1, TRIP13, CDT1, TICRR, ESCO2, CDCA5, SGO1, FBXO5

**Tab.** **2** All pathways

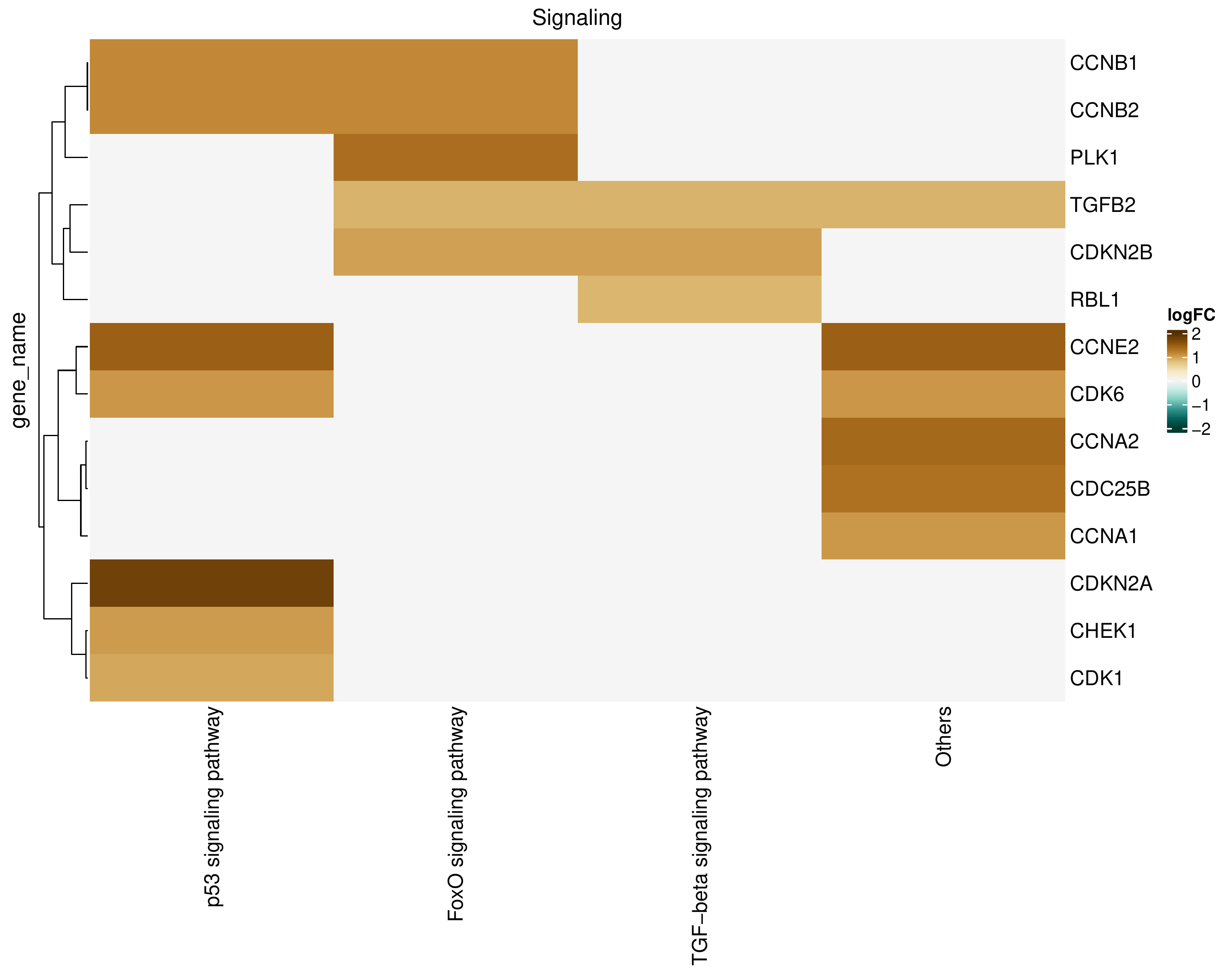
| ID | Descri... | SetSize | Enrich... | NES | Pvalue | P.adjust | Qvalue | Rank | Leadin... |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Hsa04110 | Cell c... | 69 | 0.6142... | 4.0675... | 1e-10 | 1.4907... | 9.4759... | 704 | Tags=6... |
| Hsa05204 | Chemic... | 23 | -0.760... | -3.040... | 1.1467... | 1.4907... | 9.4759... | 643 | Tags=8... |
| Hsa05323 | Rheuma... | 35 | 0.5859... | 3.2617... | 1.6229... | 1.3724... | 8.7239... | 1415 | Tags=9... |
| Hsa00982 | Drug m... | 25 | -0.709... | -2.891... | 2.3980... | 1.3724... | 8.7239... | 864 | Tags=8... |
| Hsa00980 | Metabo... | 26 | -0.698... | -2.873... | 2.6394... | 1.3724... | 8.7239... | 864 | Tags=8... |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**(File path: Figure+Table/All-pathways.xlsx)**

## 3.4 富集分析 (CELL)

对Tab. **[1](#Used-pathways)** 所示的三条通路，归类富集 (重新富集分析一边，提取信号通路) ， 见 Fig. **[6](#Re-classification-pathways)**。

其中，p53 通路含最多的基因。



**Fig.** **6** Re classification pathways

**(File path: Figure+Table/Re-classification-pathways.pdf)**

## 3.5 Limma 差异分析 (BLCA\_EX: Dead\_vs\_alive)

对 Dead vs alive 做了差异分析。

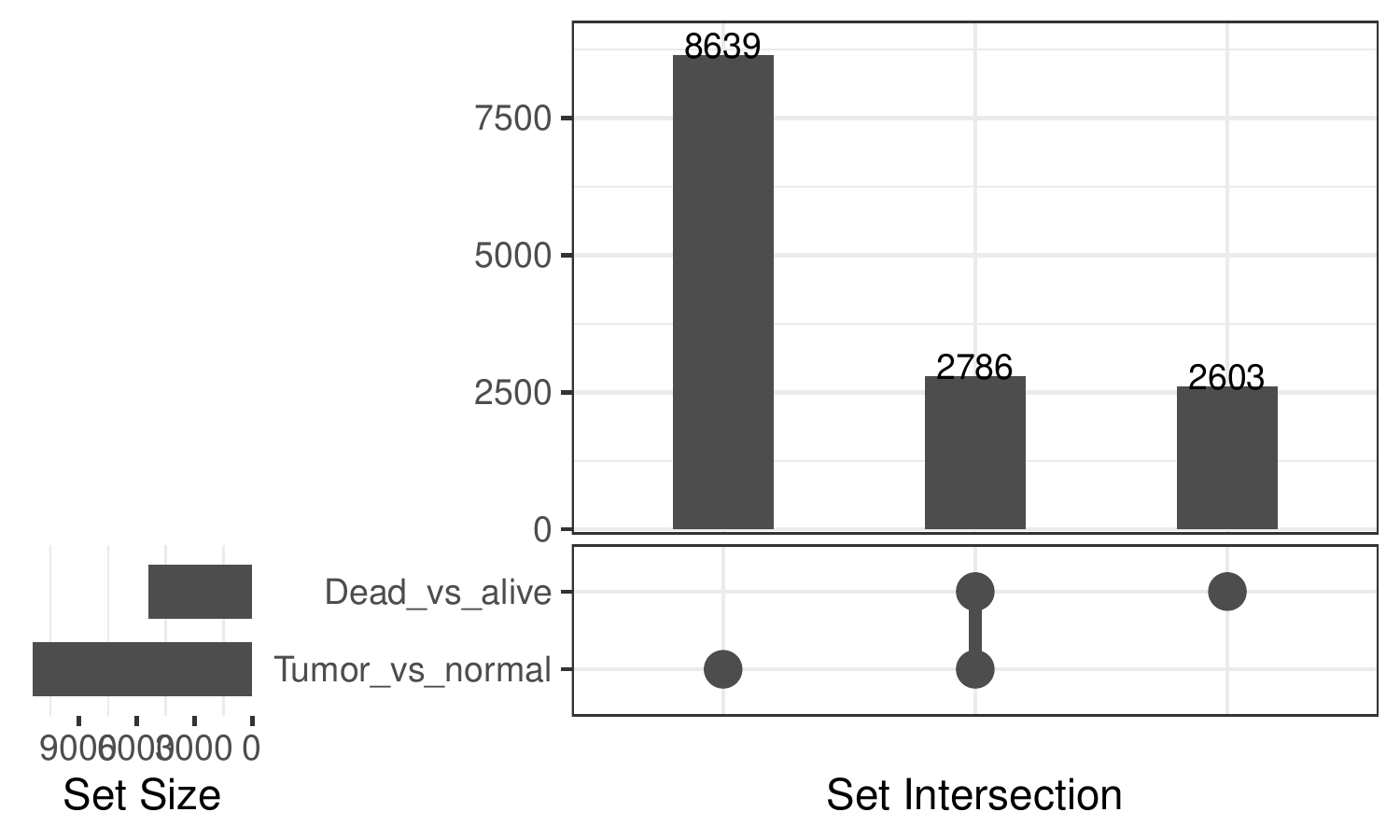
共 428 个样本，分 3 组，分别为 Alive (235) , Dead (192) , Not.Reported (1) 。 差异分析 Dead vs Alive (若 A vs B，则为前者比后者，LogFC 大于 0 时，A 表达量高于 B) 所有上调 DEGs 有 1501 个，下调共 979；一共 2480 个 (非重复)。

## 3.6 Limma 差异分析 (BLCA\_TUMOR: Tumor\_vs\_normal)

对 Tumor\_vs\_normal 做了差异分析。

共 428 个样本，分 2 组，分别为 normal (19) , tumor (409) 。 差异分析 tumor 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) 所有上调 DEGs 有 5899 个，下调共 5941；一共 11840 个 (非重复)。

## 3.7 共同显著的基因



**Fig.** **7** Intersection of Dead vs alive with Tumor vs normal

**(File path: Figure+Table/Intersection-of-Dead-vs-alive-with-Tumor-vs-normal.pdf)**

* All\_intersection: ORC1, DNMT1, WDR76, DTL, CCNA2, MELK, WDHD1, DLGAP5, MCM6, SMC2, FOXM1, FBXO5, NCAPG, DCLRE1B, NCAPG2, KIF4A, RACGAP1, KIF18A, HASPIN, STIL, PRIM1, TPX2, SGO2, CDCA8, UHRF1, RRM2, LMNB1, TYMS, CENPI, KIFC1, RAD54L, MKI67, KIF23, PLK1, ZNF367, MCM2, ECT2, E2F7, EXO1, DEPDC1, RAD51AP1, KIF11, CIP2A, BRCA1, NEIL3, NUP155, CDCA5, FAM111B, KIF2C, SPC25, MCM10, ATAD2, CENPA, CKAP2L, ASPM, ANLN, NCAPD2, NCAPH, AURKA, BUB1B, INCENP, CDKN3, CENPL, CHAF1B, NDC80, FEN1, ARHGAP11A, MCM4, KIF18B, DIAPH3, …

**(See: Figure+Table/Intersection-of-Dead-vs-alive-with-Tumor-vs-normal-content)**

## 3.8 STRINGdb PPI 分析 (CLSPN)

对上述 Fig. **[7](#Intersection-of-Dead-vs-alive-with-Tumor-vs-normal)** 交集基因，与 CLSPN 以 PPI 网络分析。

**Tab.** **3** CLSPN to others full link

| From | To |
| --- | --- |
| CLSPN | INMT |
| CLSPN | RFC2 |
| CLSPN | TRIP13 |
| CLSPN | CDC6 |
| CLSPN | CENPM |
| ... | ... |

**(File path: Figure+Table/CLSPN-to-others-full-link.csv)**

## 3.9 Survival 生存分析 (COL)

对Tab. **[3](#CLSPN-to-others-full-link)** 中的基因做了生存分析。

**Tab.** **4** COL Survival PValue

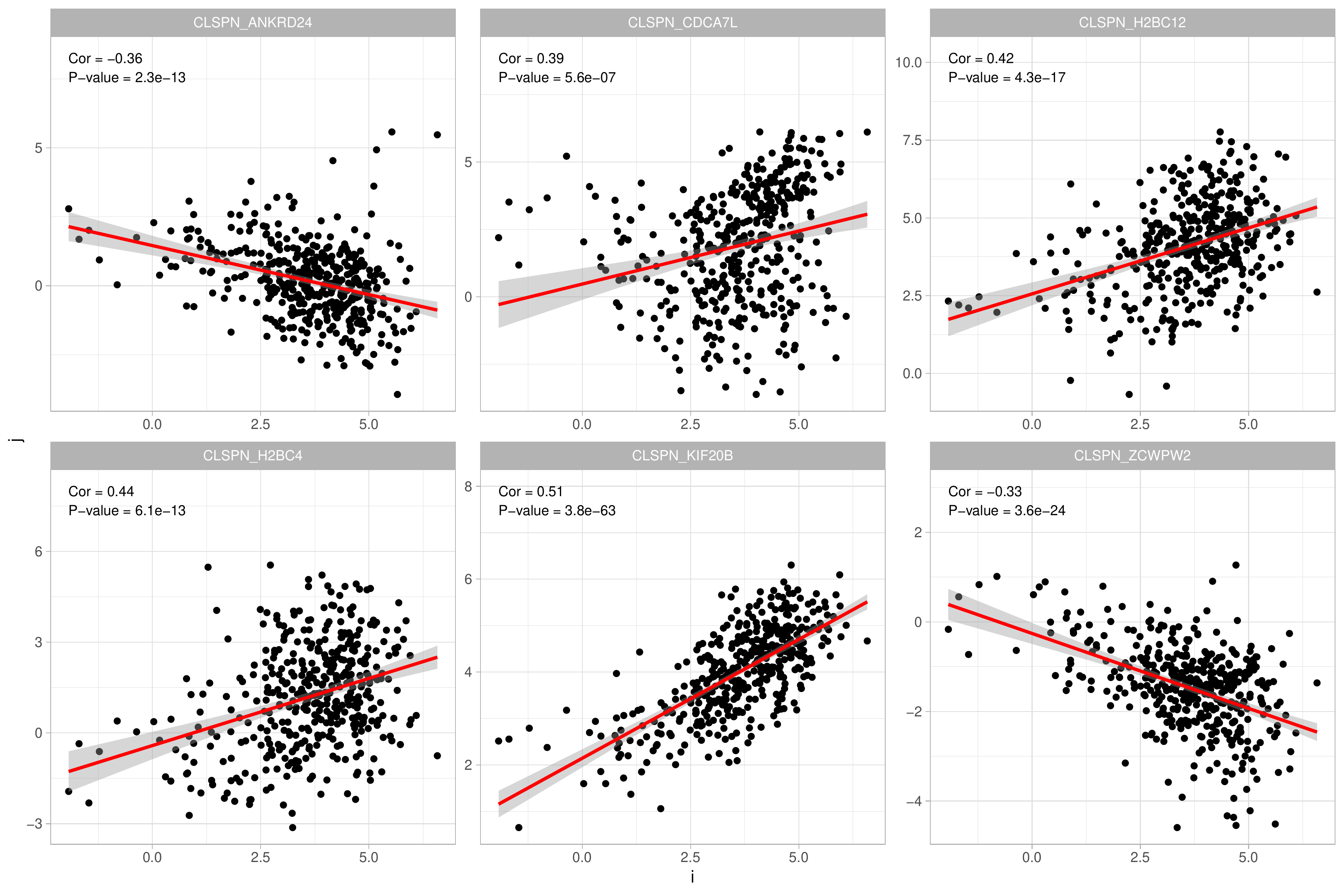
| Name | Pvalue |
| --- | --- |
| H2BC4 | 0.00398214930334519 |
| H2BC12 | 0.0130119858910616 |
| KIF20B | 0.0453907412167212 |
| ZCWPW2 | 0.0324567913269916 |
| CDCA7L | 0.0200568412210497 |
| ... | ... |

**(File path: Figure+Table/COL-Survival-PValue.csv)**

Note: The directory 'Figure+Table/Significant-Survival' contains 7 files.  
  
1 1\_H2BC4.pdf  
2 2\_H2BC12.pdf  
3 3\_KIF20B.pdf  
4 4\_ZCWPW2.pdf  
5 5\_CDCA7L.pdf  
6 ...

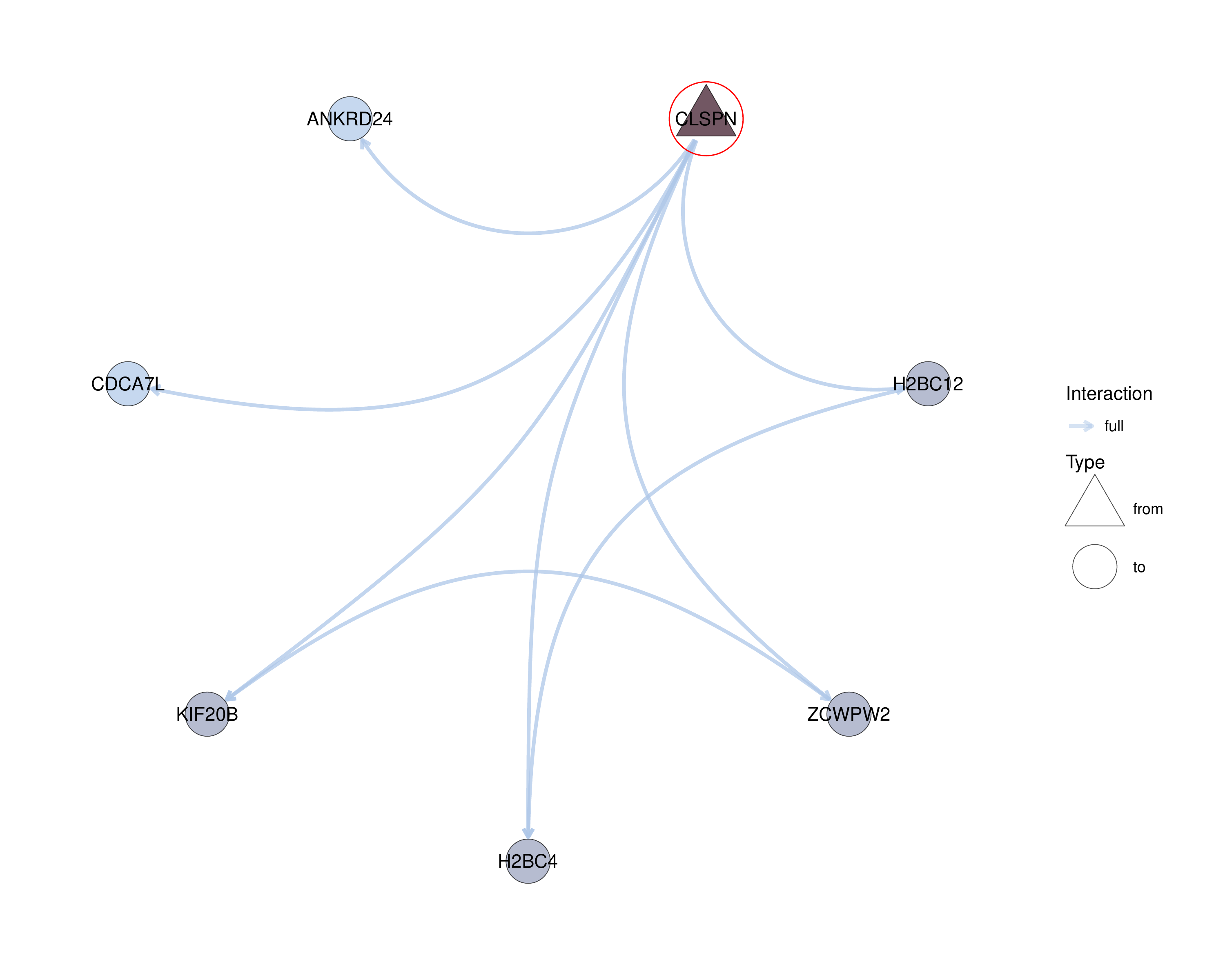
**(File path: Figure+Table/Significant-Survival)**

## 3.10 关联性分析



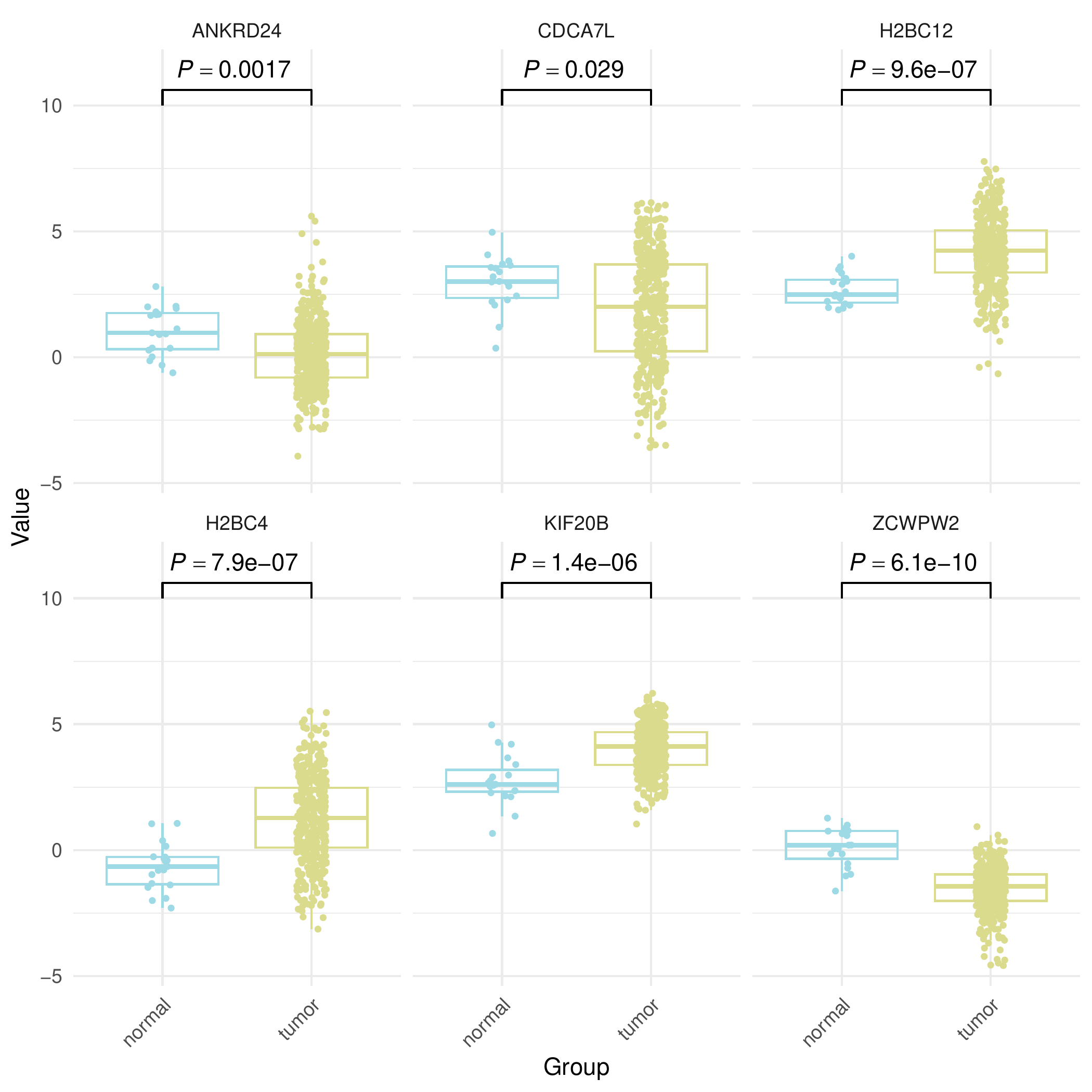
**Fig.** **8** Linear regression

**(File path: Figure+Table/Linear-regression.pdf)**



**Fig.** **9** CLSPN interaction to Significant genes

**(File path: Figure+Table/CLSPN-interaction-to-Significant-genes.pdf)**

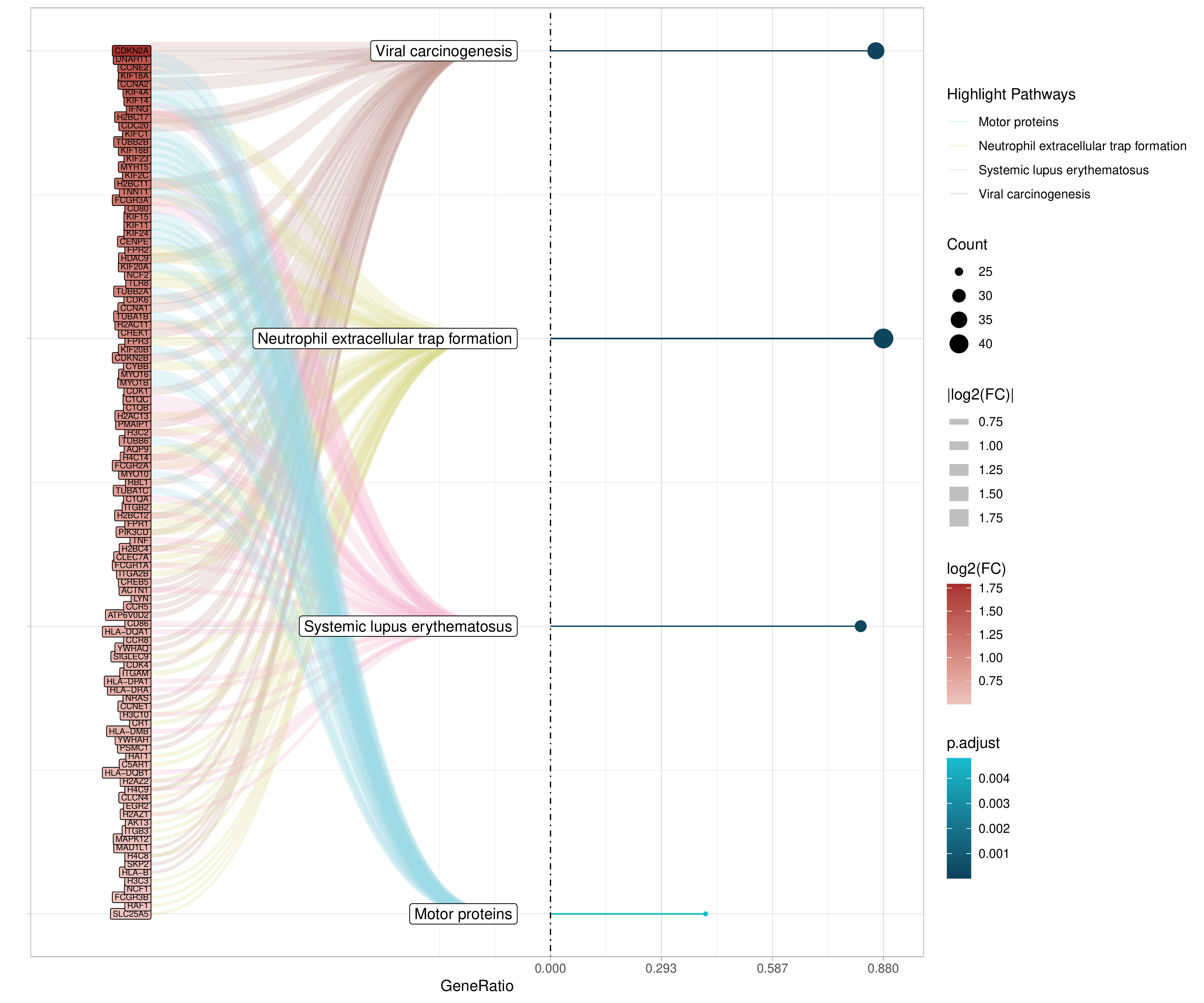


**Fig.** **10** Boxplot of Significant genes

**(File path: Figure+Table/Boxplot-of-Significant-genes.pdf)**

## 3.11 信号通路的选择

注：原思路无法在最显著的通路 (前三条) 与后续的 生存、差异分析、互作之间，找到共同的基因， 因此，以后续的分析，反向寻找通路可能更好。 以下是可行的通路。 (以 Fig. **[8](#Linear-regression)** 的基因，在 Tab. **[2](#All-pathways)** 中寻找通路) 见 Fig. **[11](#BLCA-Significant-genes-pathway)**



**Fig.** **11** BLCA Significant genes pathway

**(File path: Figure+Table/BLCA-Significant-genes-pathway.pdf)**

# 4 总结

原思路的第一部分与第二部分分别分析，第一部分主要结果见 Fig. **[6](#Re-classification-pathways)** 第二部分主要结果见： Fig. **[8](#Linear-regression)** Fig. **[9](#CLSPN-interaction-to-Significant-genes)** Fig. **[10](#Boxplot-of-Significant-genes)**

两部分的基因没有共同交集，因此建议，按第二部分的基因，在第一部分重新寻找其它通路，见 Fig. **[11](#BLCA-Significant-genes-pathway)**

# Reference

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

2. 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).

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. Szklarczyk, D. *et al.* The string database in 2021: Customizable proteinprotein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605–D612 (2021).

6. Chin, C.-H. *et al.* CytoHubba: Identifying hub objects and sub-networks from complex interactome. *BMC Systems Biology* **8**, S11 (2014).