网络药理学分析 + 蛋白对接模拟

2024-01-10

LiChuang Huang

@ 立效研究院

Contents

List of Figures

List of Tables

1 摘要

脓毒症肺损伤 + 血管重塑 + 基因 + 糖酵解

- 糖酵解与肺血管病理性重塑（如果比较少，放宽到血管重塑 remodeling）相关的基因集
  - 血管重塑基因 (6.1.1) 和脓毒症肺损伤基因 (6.1.2) 以及糖酵解相关基因 (6.2) 取全交集 (Fig. 3)
- 对基因集做功能通路富集分析
  - 分别对败血性肺损伤 (septic lung injury, SLI) 数据集 (Fig. 4) 和上述交集后的基因集 (Fig. 5 和 Fig. 6) 做了富集分析
- 在这些基因集中找到 kif2c（如果包含可能名次比较靠后了），及 kif2c 相关的基因，做 PPI 网络图
  - 找不到 KIF2F 基因。交集基因 PPI 图见 Fig. 7
- 目标靶基因是 MYC，用分子对接模拟 KIF2C 与 MYC 蛋白互作
  - KIF2C 与 MYC 蛋白互作模拟结果见 6.6
- 其他，看有能满足思路的花里胡哨的图都可以放上来

2 前言

3 材料和方法

3.1 材料

All used GEO expression data and their design:

- **GSE165226**: we divided 6-8 weeks mice into two groups - control and septic model and 6 mice per group.

3.2 方法

Mainly used method:

- R package `biomaRt` used for gene annotation[1].
- The `biomart` was used for mapping genes between organism (e.g., mgi_symbol to hgnc_symbol)[1].
- R package `ClusterProfiler` used for gene enrichment analysis[2].
- GEO https://www.ncbi.nlm.nih.gov/geo/ used for expression dataset aquisition.
- Databses of `DisGeNet`, `GeneCards`, `PharmGKB` used for collating disease related targets[3–5].
- R package `ClusterProfiler` used for GSEA enrichment[2].
- R package `Limma` and `edgeR` used for differential expression analysis[6,7].
- `LZerD` and `HawkDock` webservers used for protein–protein docking[8,9].
- R package `STEINGdb` used for PPI network construction[10,11].
- Other R packages (eg., `dplyr` and `ggplot2`) used for statistic analysis or data visualization.

4 分析结果

5 结论

6 附：分析流程

6.1 Disease (database: PharmGKB, DisGeNet, GeneCards)

6.1.1 血管重塑 (Vascular Remodeling, VR)

从三个数据库获取相关基因：

Figure 1 (下方图) 为图 VR Overall targets number of datasets 概览。

(对应文件为 `Figure+Table/VR-Overall-targets-number-of-datasets.pdf`)

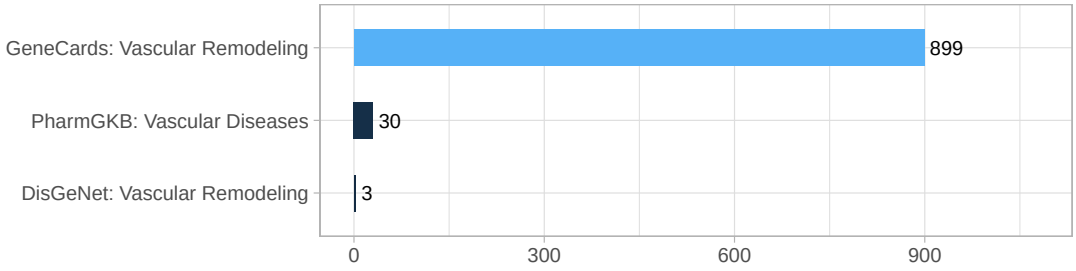

Figure 1: VR Overall targets number of datasets

'VR targets of datasets' 数据已全部提供。

(对应文件为 `Figure+Table/VR-targets-of-datasets`)

> 注：文件夹 Figure+Table/VR-targets-of-datasets 共包含 3 个文件。
>
> 1. 1_t.pharm.csv
> 2. 2_t.dis.csv
> 3. 3_t.genecard.csv

6.1.2 败血性肺损伤 (septic lung injury, SLI) GEO

6.1.2.1 DEGs-mice

Figure 2 (下方图) 为图 SLI Model vs Control DEGs 概览。

(对应文件为 `Figure+Table/SLI-Model-vs-Control-DEGs.pdf`)

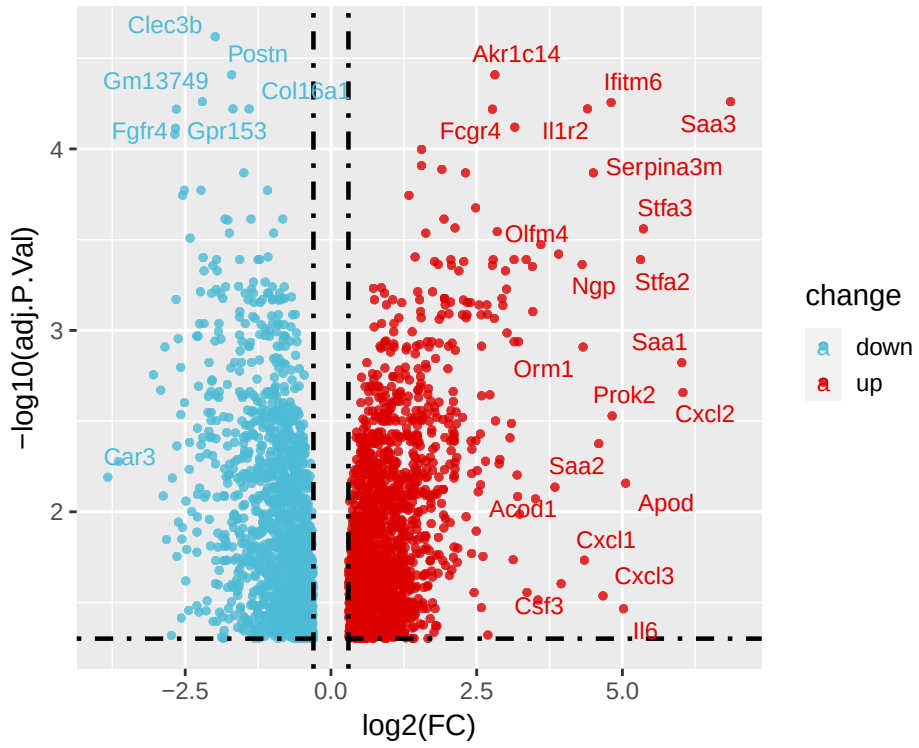

Figure 2: SLI Model vs Control DEGs

Table 1 (下方表格) 为表格 SLI data Model vs Control DEGs 概览。

(对应文件为 `Figure+Table/SLI-data-Model-vs-Control-DEGs.csv`)

注：表格共有 3242 行 7 列，以下预览的表格可能省略部分数据；表格含有 3242 个唯一 'rownames'。

1. logFC: estimate of the log2-fold-change corresponding to the effect or contrast (for 'topTableF' there may be several columns of log-fold-changes)
2. AveExpr: average log2-expression for the probe over all arrays and channels, same as 'Amean' in the 'MarrayLM' object
3. t: moderated t-statistic (omitted for 'topTableF')
4. P.Value: raw p-value
5. B: log-odds that the gene is differentially expressed (omitted for 'topTreat')

Table 1: SLI data Model vs Control DEGs

rownames	logFC	AveExpr	t	P.Value	adj.P.Val	B
Clec3b	-1.9848536...	7.33363889...	-17.498042...	7.86788380...	2.40662829...	12.2053789...
Akr1c14	2.81261598...	8.11883664...	15.3675634...	3.45102470...	3.90133653...	11.0637187...
Postn	-1.7034833...	9.91717469...	-15.228085...	3.82634026...	3.90133653...	10.9810734...
Gm13749	-2.2052006...	2.83326873...	-14.209839...	8.35339935...	5.48966386...	10.3444807...

rownames	logFC	AveExpr	t	P.Value	adj.P.Val	B
Saa3	6.85486850…	10.8792202…	14.1195879…	8.97355804…	5.48966386…	10.2850970…
Ifitm6	4.80701587…	9.89669026…	13.8796505…	1.08773578…	5.54527703…	10.1247543…
Il1r2	4.40115929…	7.69398107…	13.5044213…	1.47848787…	6.01193210…	9.86660629…
Gpr153	-1.6801679…	8.62129465…	-13.430255…	1.57236356…	6.01193210…	9.81448518…
Col16a1	-1.4026272…	7.91707219…	-13.287008…	1.77240443…	6.02381186…	9.71276692…
Fgfr4	-2.6504553…	6.15492443…	-13.121581…	2.03819307…	6.03742187…	9.59355318…
Fcgr4	2.77105360…	8.56146562…	13.0473547…	2.17116649…	6.03742187…	9.53944390…
Itgam	3.15398369…	10.0245737…	12.6817126…	2.97807344…	7.59110922…	9.26717907…
Ces2e	-2.6636707…	2.86681445…	-12.575583…	3.26902738…	7.69176996…	9.18633199…
Scube2	-2.6771258…	6.83822578…	-12.405540…	3.80107102…	8.30479717…	9.05504426…
Tgfbi	1.55482395…	10.2352866…	12.1160976…	4.93388623…	0.00010061…	8.82648927…
…	…	…	…	…	…	…

6.1.2.2 DEGs-human

使用 Biomart 将 mice 基因 (mgi symbol) 映射为 human 基因名 (hgnc symbol)

Table 2 (下方表格) 为表格 SLI Genes mapping 概览。

(对应文件为 `Figure+Table/SLI-Genes-mapping.csv`)

> 注：表格共有 2471 行 8 列，以下预览的表格可能省略部分数据；表格含有 2362 个唯一 'mgi_symbol'。

1. hgnc_symbol: 基因名 (Human)
2. mgi_symbol: 基因名 (Mice)
3. logFC: estimate of the log2-fold-change corresponding to the effect or contrast (for 'topTableF' there may be several columns of log-fold-changes)
4. AveExpr: average log2-expression for the probe over all arrays and channels, same as 'Amean' in the 'MarrayLM' object
5. t: moderated t-statistic (omitted for 'topTableF')
6. P.Value: raw p-value
7. B: log-odds that the gene is differentially expressed (omitted for 'topTreat')

Table 2: SLI Genes mapping

mgi_sy…	logFC	AveExpr	t	P.Value	adj.P.Val	B	hgnc_s…
Clec3b	-1.984…	7.3336…	-17.49…	7.8678…	2.4066…	12.205…	CLEC3B
Postn	-1.703…	9.9171…	-15.22…	3.8263…	3.9013…	10.981…	POSTN
Ifitm6	4.8070…	9.8966…	13.879…	1.0877…	5.5452…	10.124…	IFITM1
Ifitm6	4.8070…	9.8966…	13.879…	1.0877…	5.5452…	10.124…	IFITM3

mgi_sy…	logFC	AveExpr	t	P.Value	adj.P.Val	B	hgnc_s…
Ifitm6	4.8070…	9.8966…	13.879…	1.0877…	5.5452…	10.124…	IFITM2
Gpr153	-1.680…	8.6212…	-13.43…	1.5723…	6.0119…	9.8144…	GPR153
Il1r2	4.4011…	7.6939…	13.504…	1.4784…	6.0119…	9.8666…	IL1R2
Col16a1	-1.402…	7.9170…	-13.28…	1.7724…	6.0238…	9.7127…	COL16A1
Fcgr4	2.7710…	8.5614…	13.047…	2.1711…	6.0374…	9.5394…	FCGR3B
Fcgr4	2.7710…	8.5614…	13.047…	2.1711…	6.0374…	9.5394…	FCGR3A
Fgfr4	-2.650…	6.1549…	-13.12…	2.0381…	6.0374…	9.5935…	FGFR4
Itgam	3.1539…	10.024…	12.681…	2.9780…	7.5911…	9.2671…	ITGAM
Ces2e	-2.663…	2.8668…	-12.57…	3.2690…	7.6917…	9.1863…	CES2
Scube2	-2.677…	6.8382…	-12.40…	3.8010…	8.3047…	9.0550…	SCUBE2
Tgfbi	1.5548…	10.235…	12.116…	4.9338…	0.0001…	8.8264…	TGFBI
…	…	…	…	…	…	…	…

6.2 糖酵解 (Glycolysis, G)

Table 3 (下方表格) 为表格 Glycolysis related genes from genecards 概览。

(对应文件为 `Figure+Table/Glycolysis-related-genes-from-genecards.xlsx`)

> 注：表格共有 1362 行 7 列，以下预览的表格可能省略部分数据；表格含有 1362 个唯一 'Symbol'。

Table 3: Glycolysis related genes from genecards

Symbol	Description	Category	UniProt_ID	GIFtS	GC_id	Score
TIGAR	TP53 Induc…	Protein Co…	Q9NQ88	42	GC12P033681	22.41
PKM	Pyruvate K…	Protein Co…	P14618	53	GC15M072199	19.66
HK2	Hexokinase 2	Protein Co…	P52789	50	GC02P074833	19.44
GAPDH	Glyceralde…	Protein Co…	P04406	54	GC12P033726	17.19
LDHA	Lactate De…	Protein Co…	P00338	54	GC11P018394	15.83
RRAD	RRAD, Ras …	Protein Co…	P55042	42	GC16M067144	15.10
HIF1A	Hypoxia In…	Protein Co…	Q16665	52	GC14P061695	14.96
HK1	Hexokinase 1	Protein Co…	P19367	53	GC10P069269	14.28
ENO3	Enolase 3	Protein Co…	P13929	51	GC17P004948	13.56
TPI1	Triosephos…	Protein Co…	P60174	51	GC12P006867	13.21
ENO1	Enolase 1	Protein Co…	P06733	51	GC01M008861	13.07
GLTC1	Glycolysis…	RNA Gene		2	GC11U909607	12.97
PFKP	Phosphofru…	Protein Co…	Q01813	49	GC10P003066	12.85
PGK1	Phosphogly…	Protein Co…	P00558	53	GC0XP078104	12.76
GCK	Glucokinase	Protein Co…	P35557	53	GC07M044978	12.38

Symbol	Description	Category	UniProt_ID	GIFtS	GC_id	Score
...	...	...	...	...	...	...

6.3 基因集 (Filtered-DEGs)

Figure 3 (下方图) 为图 Filtered DEGs intersection 概览。

(对应文件为 `Figure+Table/Filtered-DEGs-intersection.pdf`)

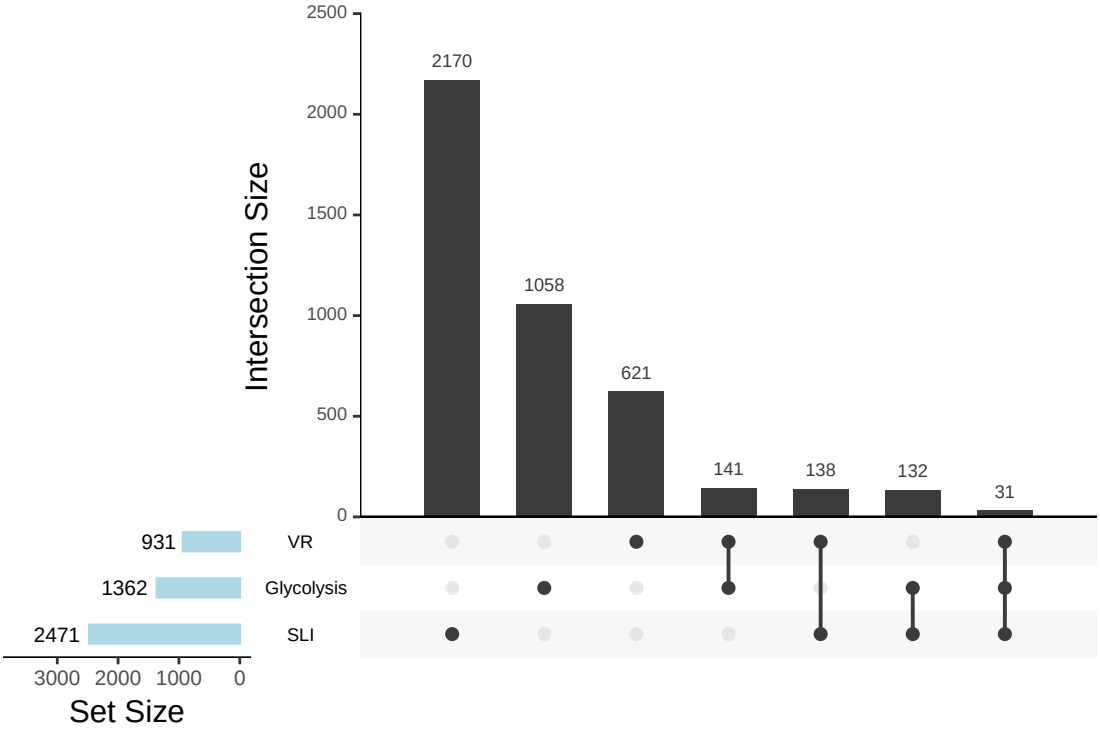

Figure 3: Filtered DEGs intersection

All_intersection :

SPHK1, COL4A5, HSPA4, PARP1, RUVBL2, TNFAIP3, PPBP, MYH10, RAB4B-EGLN2, ANGPTL4, INSR, HMGA2, KDR, TNFSF10, WWTR1, PARK7, HIF1AN, TKT, CA9, TRAF6, CASP1, AHR, NR4A1, CHUK, MUC1, ROCK1, CDK2, TSC1, IRS1, PRKACA, MDM2

(上述信息框内容已保存至 `Figure+Table/Filtered-DEGs-intersection-content`)

6.4 富集分析

6.4.1 SLI-DEGs

Figure 4 (下方图) 为图 KEGG enrichment with enriched genes 概览。

(对应文件为 `Figure+Table/KEGG-enrichment-with-enriched-genes.pdf`)

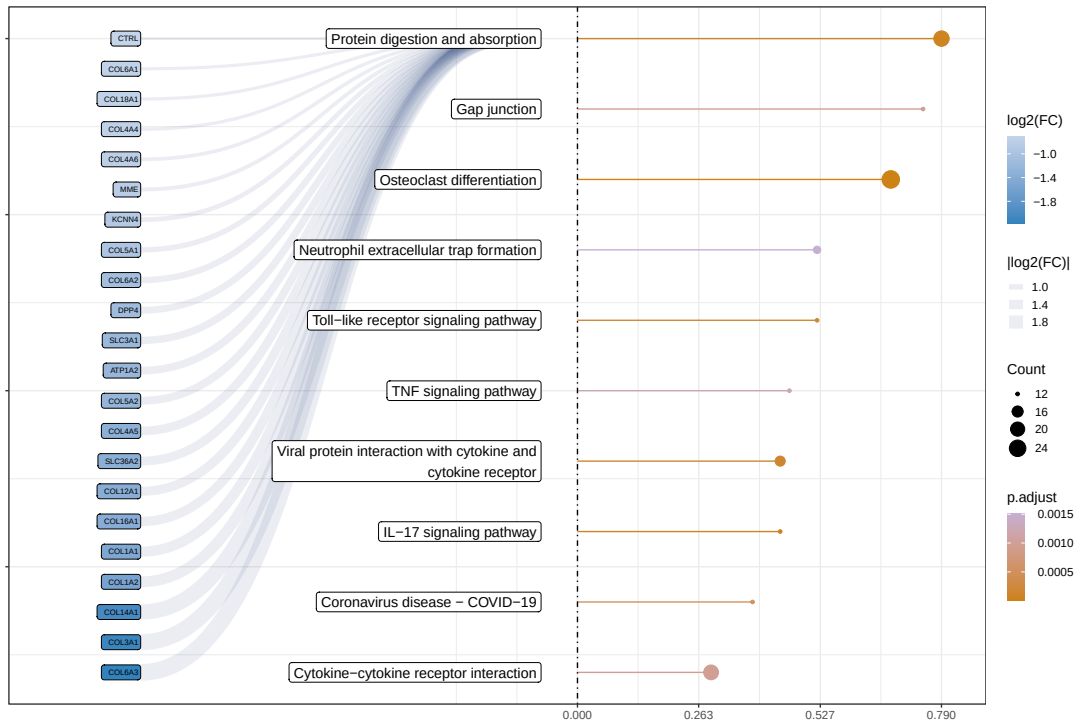

Figure 4: KEGG enrichment with enriched genes

6.4.2 Filtered-DEGs (FDEGs)

Figure 5 (下方图) 为图 FDEGS ids KEGG enrichment 概览。

(对应文件为 `Figure+Table/FDEGS-ids-KEGG-enrichment.pdf`)

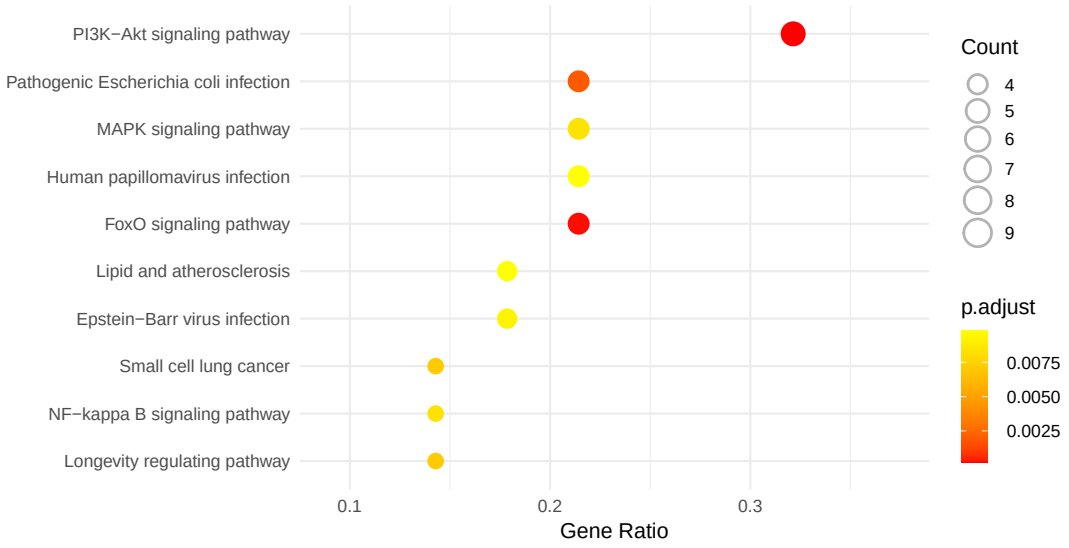

Figure 5: FDEGS ids KEGG enrichment

Figure 6 (下方图) 为图 FDEGS ids GO enrichment 概览。

(对应文件为 `Figure+Table/FDEGS-ids-GO-enrichment.pdf`)

Figure 6: FDEGS ids GO enrichment

6.5 PPI

Figure 7 (下方图) 为图 PPI of Filtered DEGs 概览。

(对应文件为 `Figure+Table/PPI-of-Filtered-DEGs.pdf`)

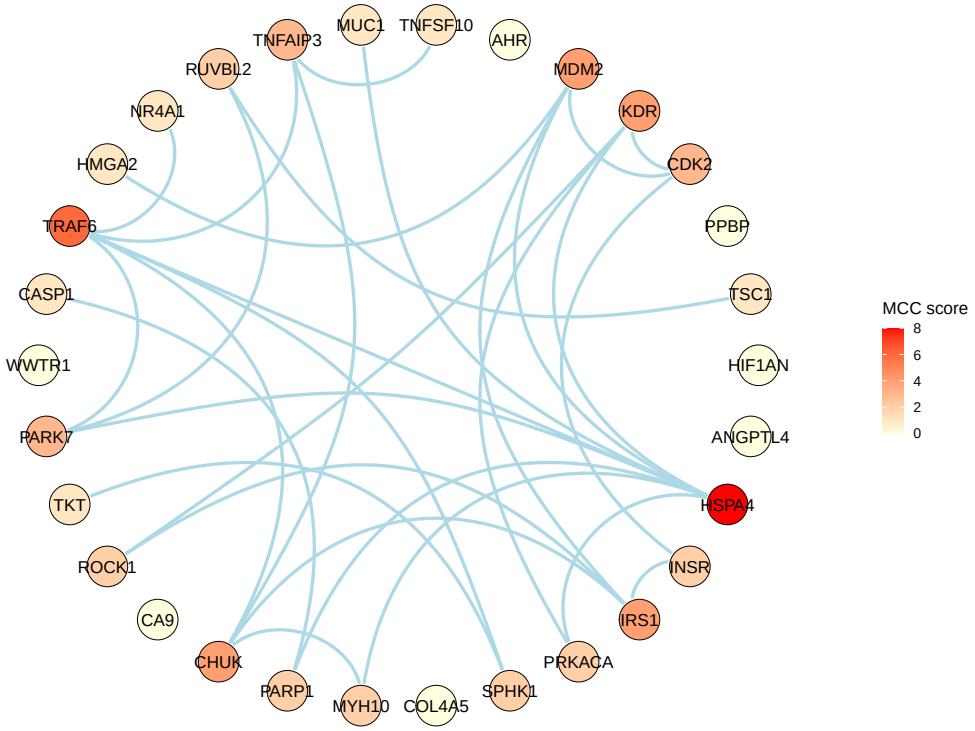

Figure 7: PPI of Filtered DEGs

6.6 蛋白互作模拟

使用了两种方法模拟对接 (LZerD 的服务器目前还没有出结果 (运行太久了)；HawkDock 的结果已出，已整理)

- Results (可以到如下网址查看结果):
  - LZerD: https://lzerd.kiharalab.org/view/b6748c34192e445686eec93fd455ce7a
  - HawkDock: http://cadd.zju.edu.cn/hawkdock/result/liwenhua-1704765524163

6.6.1 HawkDock results

Figure 8 (下方图) 为图 HawkDock ranking of all top 10 docking 概览。

(对应文件为 `Figure+Table/HawkDock-ranking-of-all-top-10-docking.pdf`)

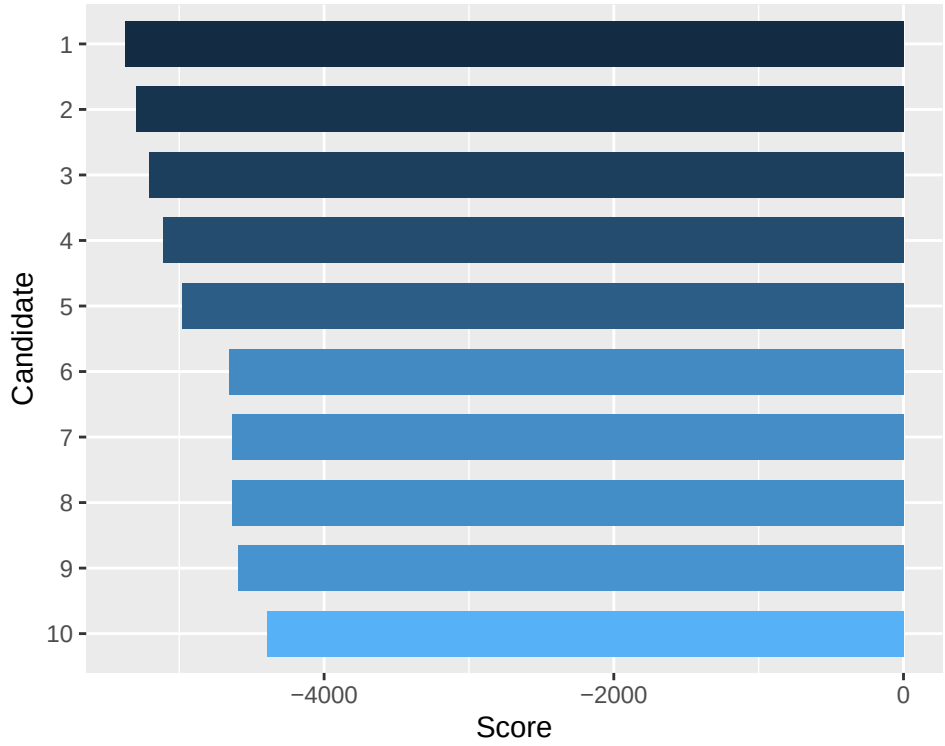

Figure 8: HawkDock ranking of all top 10 docking

Figure 9 (下方图) 为图 HawkDock docking top 1 概览。

(对应文件为 `Figure+Table/MYC..7T1Y_with_KIF2C..2HEH_top1.png`)

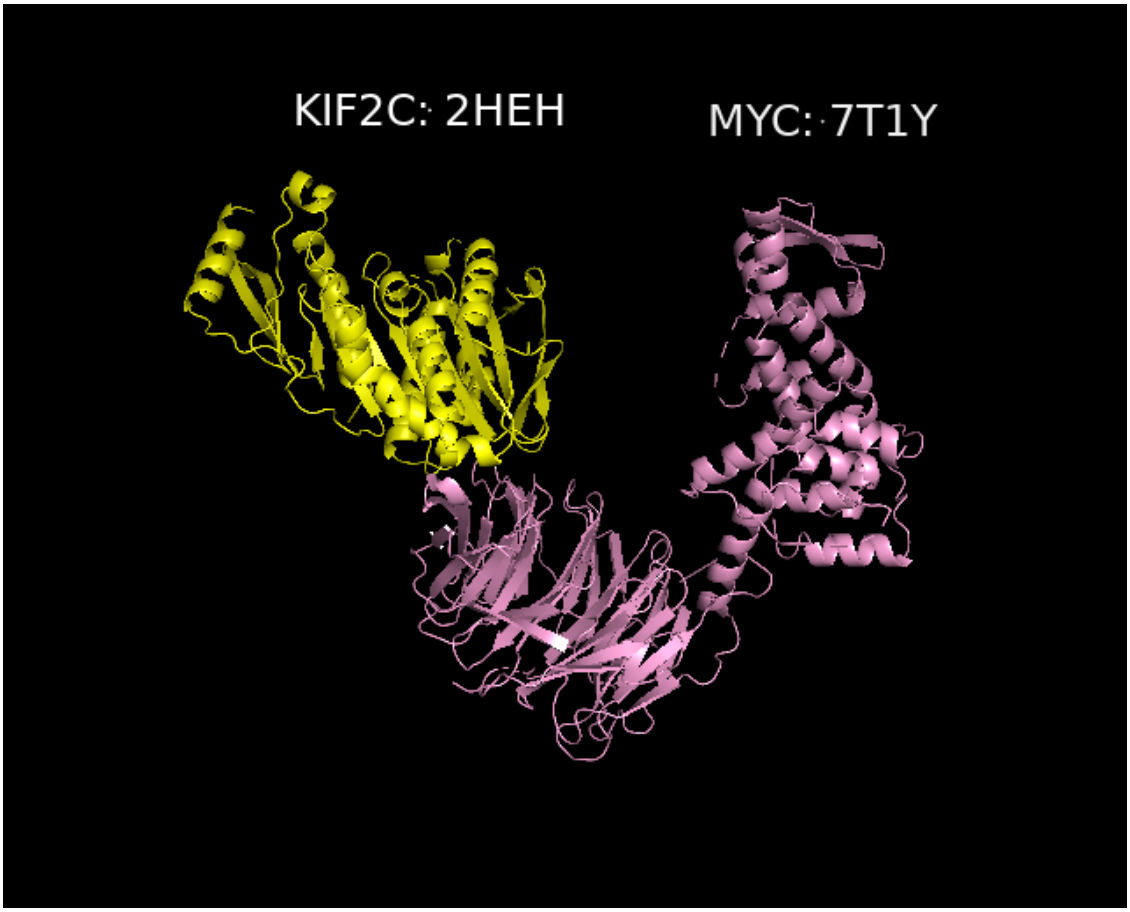

Figure 9: HawkDock docking top 1

Figure 10 (下方图) 为图 HawkDock docking top 4 概览。

(对应文件为 `Figure+Table/MYC..7T1Y_with_KIF2C..2HEH_top4.png`)

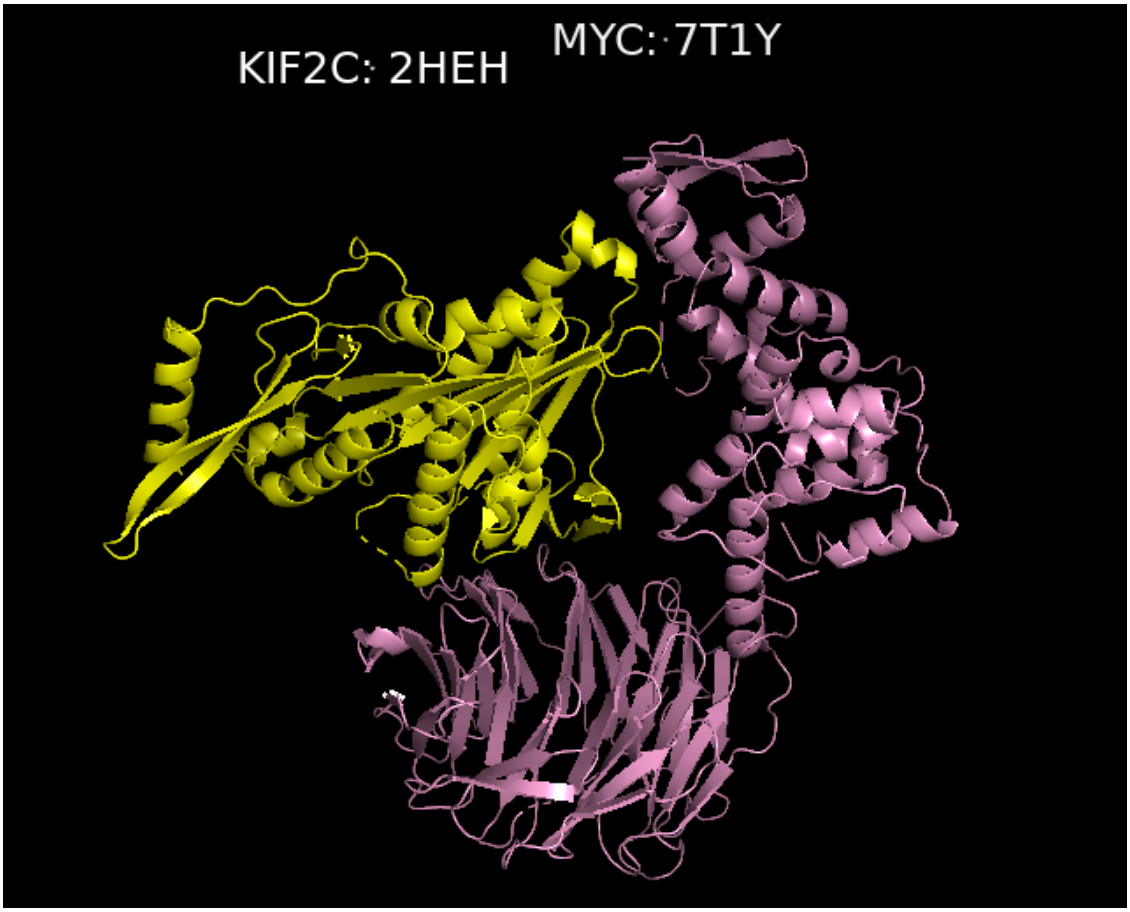

Figure 10: HawkDock docking top 4

Reference

1. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. *Nature protocols* **4**, 1184–1191 (2009).

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

3. Piñero, J. *et al.* The disgenet knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2019) doi:10.1093/nar/gkz1021.

4. Stelzer, G. *et al.* The genecards suite: From gene data mining to disease genome sequence analyses. *Current protocols in bioinformatics* **54**, 1.30.1–1.30.33 (2016).

5. Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley interdisciplinary reviews. Systems biology and medicine* **10**, (2018).

6. Ritchie, M. E. *et al.* Limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research* **43**, e47 (2015).

7. Chen, Y., McCarthy, D., Ritchie, M., Robinson, M. & Smyth, G. EdgeR: Differential analysis of sequence

read count data user's guide. 119.

8. Christoffer, C. *et al.* LZerD webserver for pairwise and multiple protein-protein docking. *Nucleic acids research* **49**, W359–W365 (2021).

9. Weng, G. *et al.* HawkDock: A web server to predict and analyze the protein-protein complex based on computational docking and mm/gbsa. *Nucleic acids research* **47**, W322–W330 (2019).

10. 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).

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