

# 生信考核答卷

黄礼闯

## Contents

<b>1 生信背景介绍</b>	<b>1</b>
1.1 主攻方向	1
1.2 生信知识和相关软件	1
1.3 编程语言	1
<b>2 分析的案例</b>	<b>2</b>
2.1 转录组学案例	2
2.2 代谢组学案例	3
<b>3 GEO 数据分析</b>	<b>5</b>
3.1 数据集和相关背景	5
3.2 软件、工具、语言包	6
3.3 分析流程	7
3.4 数据分析	9
3.4.1 Set-up	9
3.4.2 获取数据	9
3.4.3 差异性分析	10
3.4.4 功能注释	13
3.4.5 通路分析	13
3.4.6 结论与临床转化	17
3.5 附：操作截图	17
<b>Reference</b>	<b>20</b>

## 1 生信背景介绍

### 1.1 主攻方向

- 代谢组学: 在校主攻非靶向 LC-MS/MS 分析方法开发, 该方法 (R 包) 正投稿 *Analytical Chemistry* (Q1, top, IF=8.008) (详情见1.3)。
- 转录组学: 熟悉 Bioconductor, 用 ‘limma’ 分析过 GEO 转录组数据集。

## 1.2 生信知识和相关软件

- 对代谢组学分析最为熟悉，能独立完成整个流程的分析：
  1. 熟悉质谱数据格式和转化。相关软件：Proteowizard
  2. 数据预处理：峰检测、对齐等。相关软件：MZmine, XCMS 等。
  3. 统计分析：T-test, PCA, PLS-DA, OPLS-DA 等。相关软件：R。
  4. 数据可视化：常规做图（条形图、点图等），热图，网络图等。相关软件：R。
  5. 结构鉴定：机器预测，SIRIUS；可视化分析和光谱匹配，GNPS, CompMass, MCnebula2等。
  6. 通路分析富集分析。相关软件：R, MetaboAnalyst等。
  7. 数据库：原始数据库MASSIVE；参考光谱库：GNPS, HMDB, MassBank等；分子结构库：PubChem。
- 转录组分析：
  1. 平台：R 的Bioconductor。
  2. R 包：能熟练使用limma。
  3. 数据库：GEO。

## 1.3 编程语言

- 语言
  - R: 精通 R 语言，从函数式编程到面向对象编程，从 R 包的开发、测试到编写说明文档。独立开发 R 包 MCnebula2 (<https://github.com/Cao-lab-zcmu/MCnebula2>，近期独立完成了为其宣传的静态网站 (<https://mcnebula.netlify.app/>，由于文章还在投稿，请勿宣传)。
  - 其他: 使用 Bash 操作 Linux 系统，擅长 VIM 编辑，熟悉 Python 的使用，涉猎过 Java。
- 系统
  - Linux: 学习、工作于 Ubuntu 发行版（使用 Bash 语言）。
- 科研绘图、办公
  - ggplot2, grid: 擅长 ggplot2 (R 包) 结合 grid 进行简单或复杂的科研绘图，编写新的可视化工具。
  - Rmarkdown / Markdown / Latex: 替代 Microsoft 系列高效编写 word、ppt、pdf。

## 2 分析的案例

### 2.1 转录组学案例

曾以基因数据库 (<https://tfbsdb.systemsbio.net/>) 结合 GEO 数据库，再结合用于基因的 Java 自然语言处理工具 (<https://julielab.de/Resources/JCoRe.html>) 处理文献，用以筛选泛组织芳烃受体 (AHR) Signature (分析思路参考文献<sup>1</sup>)。

以下说明流程 (Fig. 1):

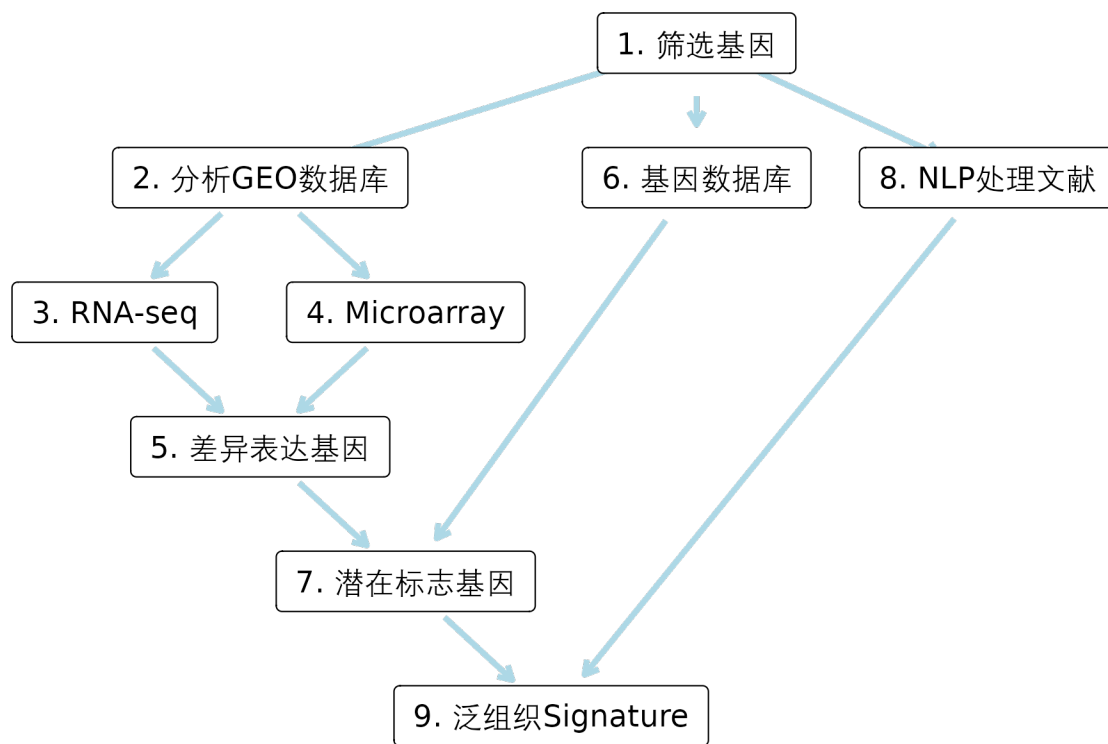


Figure 1: 转录组分析案例流程示例

1. 筛选基因结合：分析 GEO 数据库、筛选基因数据库、NLP 处理文献。
2. 分析 GEO 数据库，分析与 AHR 研究相关的疾病模型的数据集，根据差异性分析筛选基因。
3. RNA-seq，分析的 GEO 数据集的类型。
4. Microarray，分析的 GEO 数据集的类型。
5. 差异表达基因，使用 'limma' 包分析得出，或者根据原研究者的研究数据筛选，根据 'Q-value'、 $\log_2(FC)$  筛选。
6. 检索转录因子靶向基因数据库 (<https://tfbsdb.systemsbiology.net/>)
7. 将第 5 步筛选的差异表达基因与第 6 步检索到的基因取合集。
8. 使用 Java 包 (<https://julielab.de/Resources/JCoRe.html>) 处理报道有关 AHR 的文献。
9. 将第 7 步和第 8 步取得的基因集取交集，获得泛组织 AHR Signature 基因

## 2.2 代谢组学案例

为了示例编写的 R 包 MCnebula2 和其工作流的应用，在研究中曾重新分析 MASSIVE 中的代谢组数据集 (MSV000083593)<sup>2</sup>。流程如下所述 (Fig. 2，代码和详细说明可见于 [https://mcnebula.netlify.app/docs/workflow/serum\\_report\\_biocstyle](https://mcnebula.netlify.app/docs/workflow/serum_report_biocstyle))：

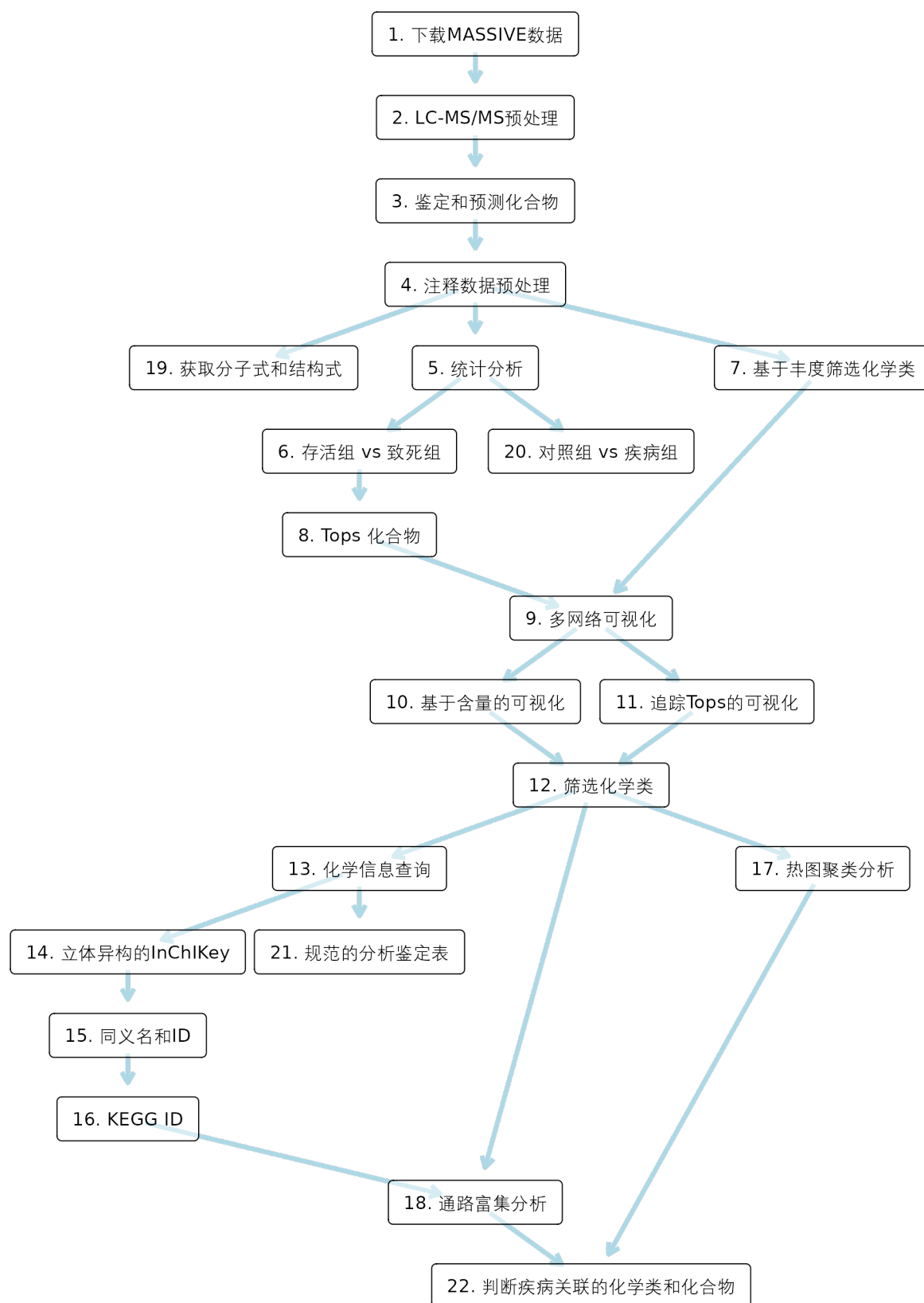


Figure 2: 代谢组分析案例流程示意

1. 下载 MASSIVE 数据, 即MSV000083593
2. LC-MS/MS 预处理, 使用 MZmine2 进行峰检测和对齐等处理, 并导出 MS/MS 信息的.mgf, 定量信息的.csv。
3. 鉴定和预测化合物, 使用 SIRIUS 软件预测化合物的分子式、结构式、化学类等。
4. 注释数据预处理, 使用 MCnebula2 (以下分析均采用) R 包整合并处理 SIRIUS 的注释数据。
5. 统计分析, 包括步骤 6 和步骤 20, 进行差异性分析。
6. 存活组 vs 致死组, 筛选两组间的差异代谢物。
7. 基于丰度筛选化学类, MCnebula2 提供的算法, 根据化合物的化学类在整体数据集中的丰富程度筛选。
8. Tops 化合物, 根据差异性分析的得分排名 (例如 Q-value), 得到高排名的化合物。
9. 多网络可视化, 基于化学类将化合物聚类可视化为多个网络。
10. 基于含量的可视化, 将代谢物的含量变化水平呈现在网络图中。
11. 追踪 Tops 的可视化, 将 Tops 化合物标记在多网络图中, 辅助筛选化学类。
12. 筛选化学类, 结合上述可视化筛选化学类。
13. 化学信息查询, 检索 PubChem 数据库, 获取各类名称、ID 和同义名信息 (使用 MCnebula2 提供的函数)。
14. 立体异构的 InChIKey, 质谱鉴定的程度达到分子骨架水平, 可以以 InChIKey2D 表示, 这一步在 PubChem 搜索所有 InChIKey2D 涵盖下的 InChIKey (使用 MCnebula2 提供的函数)
15. 同义名和 ID, 检索得到的化合物信息。
16. KEGG ID, 代谢物的 ID, 从步骤 15 得到的 ID 信息的 CID (PubChem CID) 通过'MetaboAnalystR'包转化为 KEGG ID。
17. 热图聚类分析, 根据步骤 12 筛选的化学类绘制聚类热图, 判断该化学类与疾病的关联性。
18. 通路富集分析, 通过步骤 16 得到的 KEGG ID 进行富集分析, 可以使用 R 的'FELLA'包, 也能使用 MetaboAnalyst 网站提供的服务。
19. 获取分子式和结构式, 得到化合物的鉴定数据, 用于注释网络图。
20. 对照组 vs 疾病组, 筛选两组间的差异代谢物。
21. 规范的分析鉴定表, 将鉴定数据和统计分析数据整合, 得到可以用于论文发表的规范表格。
22. 判断疾病关联的化学类和化合物, 结合上述得出分析的结论, 以备进一步验证。

### 3 GEO 数据分析

#### 3.1 数据集和相关背景

数据编号为: GSE223325 (Fig. 3)。本数据集为较新的数据集, 未见研究报道的记录。研究类风湿性关节炎单核细胞衍生巨噬细胞的促炎症和代谢功能。对极化的健康和类风湿关节炎单核细胞衍生的巨噬细胞进行了 RNA 测序。组别:

- 对照组, 24 hrs IL-4 (20ng/ml)。注: IL-4 极化的巨噬细胞脂多糖 (LPS) 暴露后可建立一个高炎症基因表达程序<sup>3</sup>。
- 模型组, 24 hrs LPS (100ng/ml) and IFN $\gamma$  (20ng/ml)。注: 由 IFN- 激活并由 STAT1 介导的前馈环路会放大细胞因子的信号, 还会增强巨噬细胞对微生物诱导剂如 Toll 样受体 (TLR) 配体 (例如 LPS) 的反应<sup>4,5</sup>。

Scope:  Format:  Amount:  GEO accession:

Series GSE223325

Query DataSets for GSE223325

Status

Public on Jan 20, 2023

Title

Rheumatoid Arthritis macrophages are primed for inflammation and display bioenergetic and functional alterations

Organism

[Homo sapiens](#)

Experiment type

Expression profiling by high throughput sequencing

Summary

To investigate the pro-inflammatory and metabolic function of Rheumatoid Arthritis monocyte derived macrophages  
RNA sequencing was performed on polarised healthy and Rheumatoid Arthritis monocyte-derived macrophages

Overall design

Comparative gene expression profiling analysis of RNA-seq data for polarised healthy and Rheumatoid Arthritis monocyte-derived macrophages

Contributor(s)

[Hanlon M](#), [Fearon U](#)

Citation missing

*Has this study been published? Please [login](#) to update or [notify GEO](#).*

Submission date

Jan 20, 2023

Last update date

Jan 24, 2023

Contact name

Megan Mary Hanlon

E-mail(s)

[hanlonme@tcd.ie](mailto:hanlonme@tcd.ie)

Phone

0873186702

Organization name

Molecular Rheumatology research group

Figure 3: GSE223325 数据集概览

## 3.2 软件、工具、语言包

来源于 CRAN 的 R 包。

```
pkgs <- c("dplyr", "data.table", "R.utils",
  "ggplot2", "BiocManager", "ggrepel")
lapply(pkgs,
  function(pkg) {
    if (!requireNamespace(pkg, quietly = T))
      install.package(pkg)
  })
```

来源于 Bioconductor 的包。

```
pkgs.bio <- c("GEOquery", "edgeR", "limma", "pathview",
  "clusterProfiler", "biomaRt")
lapply(pkgs.bio,
  function(pkg) {
    if (!requireNamespace(pkg, quietly = T))
      BiocManager::install(pkg)
  })
```

### 3.3 分析流程

本次分析的流程图见 Fig. 4。详情见3.4。

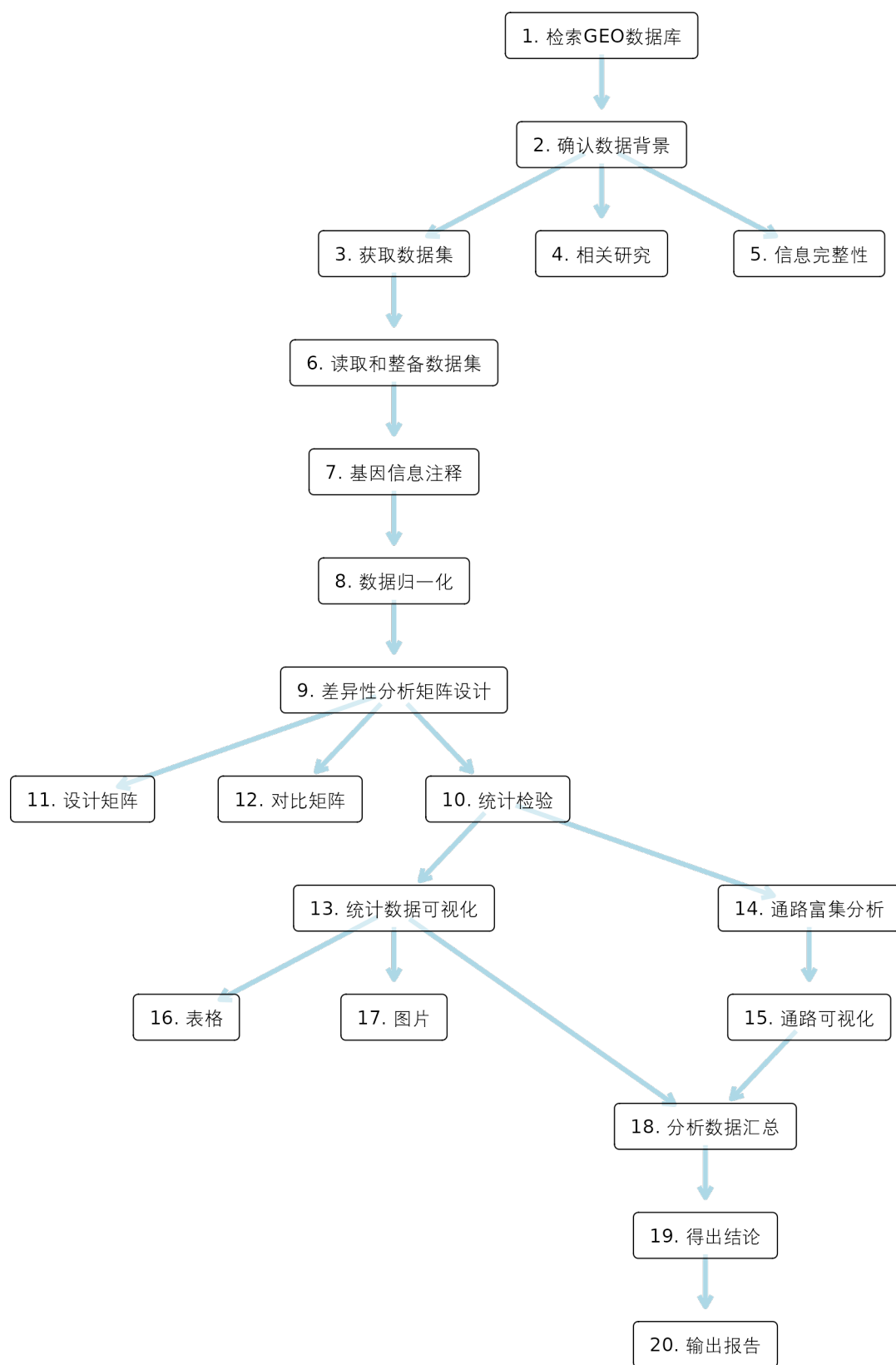


Figure 4: 分析流程示意图



## 3.4 数据分析

### 3.4.1 Set-up

R 的设定。只加载了 `ggplot2` 包，其他的包在使用时通过`::`调用。另外，个别函数用了自定义的函数，以免代码过于琐碎（对于关键性的步骤，没有使用自定义的函数）

```
library(ggplot2)
## The custom functions
bin <- RCurl::getURLContent(
  paste0("https://raw.githubusercontent.com/Cao-lab-zcmu/utils_tool/",
    "master/R/tmp.ahr.R")
)
writeBin(bin, tmp <- tempfile(fileext = ".R"))
source(tmp)
```

### 3.4.2 获取数据

使用 R 包 `GEOquery` 获取 GEO 数据。

```
gse <- "GSE223325"
about <- GEOquery::getGEO(gse)
```

查看数据已进行过的处理。

```
about[[1]]$data_processing.3[1]
```

```
## [1] "Supplementary files format and content: tab-delimited text files of TPM for each Sample"
```

获取‘GPL’的注释信息。

```
gpl <- about[[1]]@annotation
anno <- GEOquery::getGEO(gpl)
org <- anno@header$organism
```

将补充材料信息下载到本地（原作者处理过的包含 TPM 的矩阵数据）。

```
GEOquery::getGEOSuppFiles(gse)
utils::untar(list.files(gse, full.names = T), exdir = gse)
lapply(list.files(gse, "\\..gz$", full.names = T), R.utils::gunzip)
```

为样品信息创建元数据表格。

```
metadata <- data.frame(files = list.files(gse, "\\..txt$", full.names = T)) %>%
  dplyr::mutate(group = ifelse(grepl("M1\\.txt", files), "M1", "M2"),
    group.anno = ifelse(group == "M1", "LPC + IFN", "IL-4"),
    sample = gsub("^.*|\\.txt$", "", files)
  )
```

检查元数据表格，并查看矩阵数据状态。

```
metadata
```

```
##               files group group.anno      sample
## 1 GSE223325/GSM6945621_RA1M1.txt    M1 LPC + IFN GSM6945621_RA1M1
## 2 GSE223325/GSM6945622_RA1M2.txt    M2      IL-4 GSM6945622_RA1M2
## 3 GSE223325/GSM6945623_RA2M1.txt    M1 LPC + IFN GSM6945623_RA2M1
## 4 GSE223325/GSM6945624_RA2M2.txt    M2      IL-4 GSM6945624_RA2M2
## 5 GSE223325/GSM6945625_RA3M1.txt    M1 LPC + IFN GSM6945625_RA3M1
## 6 GSE223325/GSM6945626_RA3M2.txt    M2      IL-4 GSM6945626_RA3M2
## 7 GSE223325/GSM6945627_RA4M1.txt    M1 LPC + IFN GSM6945627_RA4M1
## 8 GSE223325/GSM6945628_RA4M2.txt    M2      IL-4 GSM6945628_RA4M2
## 9 GSE223325/GSM6945629_RA5M1.txt    M1 LPC + IFN GSM6945629_RA5M1
## 10 GSE223325/GSM6945630_RA5M2.txt   M2      IL-4 GSM6945630_RA5M2
## 11 GSE223325/GSM6945631_RA6M1.txt   M1 LPC + IFN GSM6945631_RA6M1
## 12 GSE223325/GSM6945632_RA6M2.txt   M2      IL-4 GSM6945632_RA6M2
## 13 GSE223325/GSM6945633_RA7M1.txt   M1 LPC + IFN GSM6945633_RA7M1
## 14 GSE223325/GSM6945634_RA7M2.txt   M2      IL-4 GSM6945634_RA7M2
## 15 GSE223325/GSM6945635_RA8M1.txt   M1 LPC + IFN GSM6945635_RA8M1
## 16 GSE223325/GSM6945636_RA8M2.txt   M2      IL-4 GSM6945636_RA8M2
## 17 GSE223325/GSM6945637_RA9M1.txt   M1 LPC + IFN GSM6945637_RA9M1
## 18 GSE223325/GSM6945638_RA9M2.txt   M2      IL-4 GSM6945638_RA9M2
```

```
tibble::as_tibble(data.table::fread("./GSE223325/GSM6945621_RA1M1.txt"))
```

```
## # A tibble: 60,504 x 4
##   V1          V2    V3    V4
##   <chr>      <int> <int> <int>
## 1 ENSG00000223972      0      0      0
## 2 ENSG00000227232      0      0      0
## 3 ENSG00000278267      1      1      0
## 4 ENSG00000243485      0      0      0
## 5 ENSG00000274890      0      0      0
## 6 ENSG00000237613      0      0      0
## 7 ENSG00000268020      0      0      0
## 8 ENSG00000240361      0      0      0
## 9 ENSG00000186092      0      0      0
## 10 ENSG00000238009      3      2      1
## # ... with 60,494 more rows
```

### 3.4.3 差异性分析

读取表达数据集。

```
dge.list <- edgeR::readDGE(metadata$files, columns = c(1, 2))
dge.list <- re.sample.group(dge.list, metadata)
```

使用 R 包 'biomaRt' 对基因信息进行注释。

```
ensembl <- biomaRt::useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
attr <- c("ensembl_gene_id", "hgnc_symbol", "chromosome_name",
         "start_position", "end_position", "description")
gene.anno <- biomaRt::getBM(attr, mart = ensembl)
gene.anno <- tibble::as_tibble(gene.anno)
dge.list <- anno.into.list(dge.list, gene.anno, "ensembl_gene_id")
```

创建设计矩阵和对比矩阵<sup>6</sup>。

```
group. <- dge.list$samples$group
design <- model.matrix(~ 0 + group.)
contr.matrix <- limma::makeContrasts(
  M1_vs_M2 = group.M1 - group.M2,
  levels = design
)
```

滤除低表达的基因信息。

```
keep.exprs <- edgeR::filterByExpr(dge.list, group = group., min.count = 10)
dge.list <- edgeR::`[.DGEList` (dge.list, keep.exprs, , keep.lib.sizes = F)
```

数据归一化。

```
dge.list <- edgeR::calcNormFactors(dge.list, method = "TMM")
dge.list <- limma::voom(dge.list, design)
```

统计检验。

```
fit <- limma::lmFit(dge.list, design)
fit.cont <- limma::contrasts.fit(fit, contrasts = contr.matrix)
ebayes <- limma::eBayes(fit.cont)
```

根据 Q-value (FDR 校正的 P-value) 和  $\log_2(\text{FC})$  获取高排名的基因。

```
res <- limma::topTable(ebayes, coef = 1, number = Inf)
res <- tibble::as_tibble(res)
res.tops <- dplyr::filter(res, adj.P.Val < .05, abs(logFC) > 1)
res.top30 <- head(res.tops, n = 30)
```

检查结果。

```
res.top30
```

```
## # A tibble: 30 x 12
```

##	ensembl_gene_id	hgnc_symbol	chromosome_name	start_position	end_position	description	logFC
##	<chr>	<chr>	<chr>	<int>	<int>	<chr>	<dbl>
##	1	ENSG00000207508	"RNU6-1237P"	1	40177843	40177949 RNA, U6 small nu~	3.02
##	2	ENSG00000154162	"CDH12"	5	21750673	22853344 cadherin 12 [Sou~	4.45
##	3	ENSG00000235711	"ANKRD34C"	15	79282722	79298239 ankyrin repeat d~	6.80
##	4	ENSG00000150045	"KLRF1"	12	9827481	9845007 killer cell lect~	6.31
##	5	ENSG00000233996	"KDM3AP1"	2	189486480	189487992 KDM3A pseudogene~	8.99
##	6	ENSG00000280344	"	16	51176066	51178898 TEC	-4.10
##	7	ENSG00000189367	"KIAA0408"	6	127438406	127459389 KIAA0408 [Source~	3.23
##	8	ENSG00000120162	"MOB3B"	9	27325209	27529814 MOB kinase activ~	11.8
##	9	ENSG00000234855	"SLIT1-AS1"	10	97102756	97104355 SLIT1 antisense ~	2.42
##	10	ENSG00000137970	"RPL7P9"	1	96678874	96679620 ribosomal protei~	7.83

## # ... with 20 more rows, and 5 more variables: AveExpr <dbl>, t <dbl>, P.Value <dbl>,  
## # adj.P.Val <dbl>, B <dbl>

将结果可视化火山图 (Fig. 5)。

```
data <- dplyr::mutate(
  res.tops, change = ifelse(logFC < -1, "down",
    ifelse(logFC > 1, "up", "stable"))
)
p <- ggplot(data, aes(x = logFC, y = -log10(adj.P.Val), color = change)) +
  geom_point(alpha = 0.8, stroke = 0, size = 3) +
  scale_color_manual(values = c("down" = "#4DBBD5FF",
    "stable" = "#8491B4FF",
    "up" = "#DC0000FF")) +
  ylim(1, max(-log10(data$adj.P.Val))) +
  geom_hline(yintercept = -log10(0.05), linetype = 4, size = 0.8) +
  geom_vline(xintercept = c(-1,1), linetype = 4, size = 0.8) +
  labs(x = "log2(FC)", y = "-log10(Q-value)") +
  ggrepel::geom_text_repel(
    data = data[-log10(data$adj.P.Val) > 7.5 & abs(data$logFC) >= 7.5,],
    aes(label = hgnc_symbol,
      size = 3,family="Times") +
    theme(text = element_text(family = "Times"))
)
ggsave(p, file = paste0("volcano.png"), height = 5.5)
```

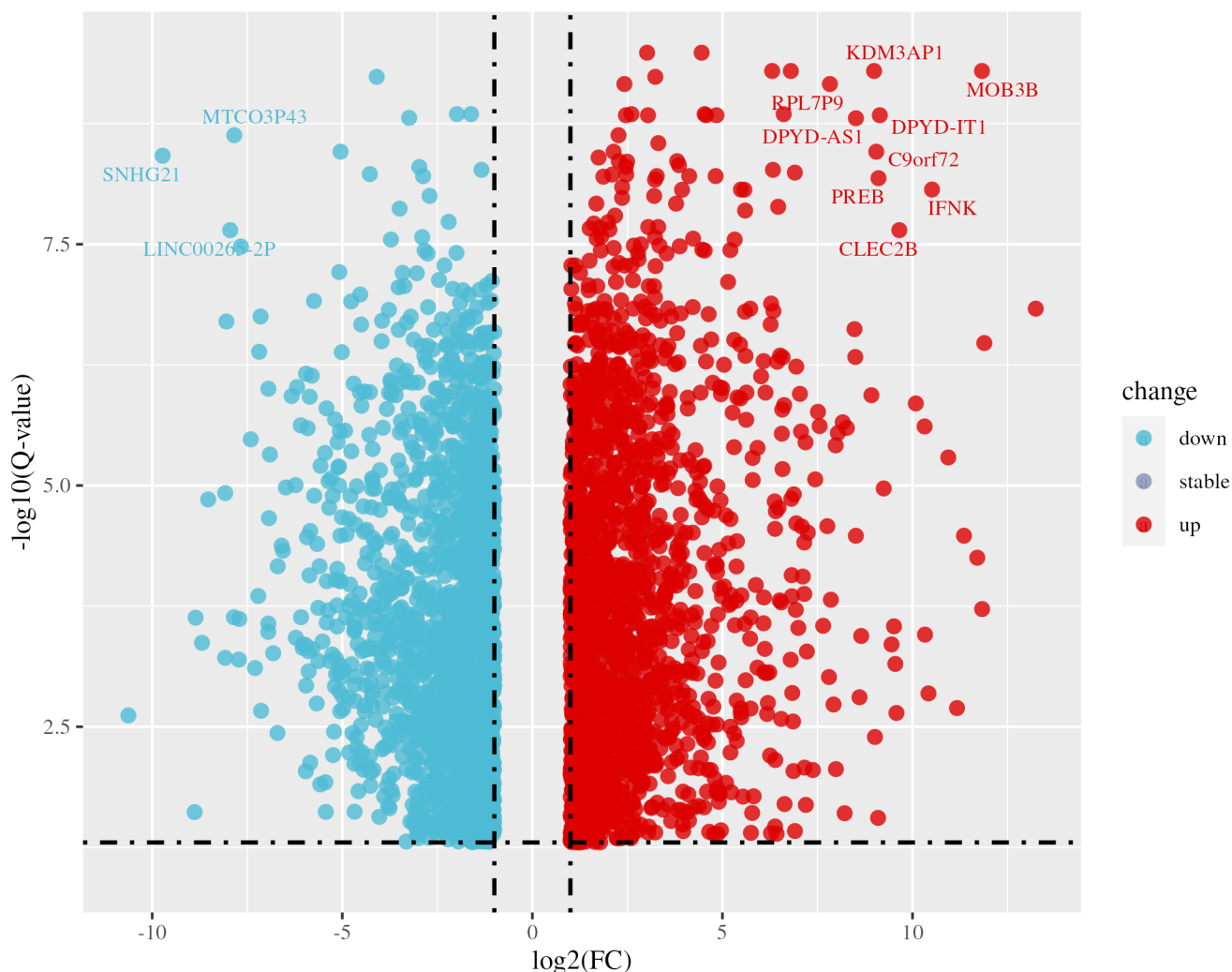


Figure 5: 差异性分析结果的火山图可视化

### 3.4.4 功能注释

在上一节3.4.3中，已经使用了 biomaRt 对基因进行了功能注释。Tab. 1显示 Top 30 的注释结果。

### 3.4.5 通路分析

**3.4.5.1 使用 clusterProfiler 通路富集分析** 使用 R 包 clusterProfiler 进行通路富集分析，在此之前，需要先将'ensembl ID' 转化为'entrez ID'（一并展示在了 Tab. 1中）。

```
ids <- clusterProfiler::bitr(
  res.top30$ensembl_gene_id, "ENSEMBL",
  "ENTREZID", "org.Hs.eg.db", F
)
ids <- dplyr::distinct(ids, ENSEMBL, .keep_all = T)
res.top30.ex <- dplyr::mutate(res.top30, entrezid = ids[[2]]) %>%
  dplyr::filter(!is.na(entrezid))
```

Table 1: Top 30 的基因功能注释 (此处仅显示 3 列)

Ensembl gene id	Hgnc symbol	Entrezid	Description
ENSG00000207508	RNU6-1237P	106480651	RNA, U6 small nuclear 1237, pseudogene [Source:HGNC Symbol;Acc:HGNC:48200]
ENSG00000154162	CDH12	1010	cadherin 12 [Source:HGNC Symbol;Acc:HGNC:1751]
ENSG00000235711	ANKRD34C	390616	ankyrin repeat domain 34C [Source:HGNC Symbol;Acc:HGNC:33888]
ENSG00000150045	KLRF1	51348	killer cell lectin like receptor F1 [Source:HGNC Symbol;Acc:HGNC:13342]
ENSG00000189367	KIAA0408	9729	KIAA0408 [Source:HGNC Symbol;Acc:HGNC:21636]
ENSG00000120162	MOB3B	79817	MOB kinase activator 3B [Source:HGNC Symbol;Acc:HGNC:23825]
ENSG00000234855	SLIT1-AS1	100505540	SLIT1 antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:51198]
ENSG00000137970	RPL7P9	653702	ribosomal protein L7 pseudogene 9 [Source:HGNC Symbol;Acc:HGNC:37028]
ENSG00000251223	MTCO1P9	107075139	MT-CO1 pseudogene 9 [Source:HGNC Symbol;Acc:HGNC:52011]
ENSG00000242423	RPL12P36	100271408	ribosomal protein L12 pseudogene 36 [Source:HGNC Symbol;Acc:HGNC:36598]
ENSG00000265420	MIR4779	100616159	microRNA 4779 [Source:HGNC Symbol;Acc:HGNC:41747]
ENSG00000117569	PTBP2	58155	polypyrimidine tract binding protein 2 [Source:HGNC Symbol;Acc:HGNC:17662]
ENSG00000176410	DNAJC30	84277	DnaJ heat shock protein family (Hsp40) member C30 [Source:HGNC Symbol;Acc:HGNC:16410]

Table 1: Top 30 的基因功能注释 (此处仅显示 3 列)

Ensembl gene id	Hgnc symbol	Entrezid	Description
ENSG00000270631		100287840	tetraspanin 12 (TSPAN12) pseudogene
ENSG00000232878	DPYD-AS1	100873932	DPYD antisense RNA 1 [Source:HGNC Symbol;Acc:HGNC:40195]
ENSG00000232400	RAD17P1	9207	RAD17 pseudogene 1 [Source:HGNC Symbol;Acc:HGNC:9808]
ENSG00000117000	RLF	6018	RLF zinc finger [Source:HGNC Symbol;Acc:HGNC:10025]
ENSG00000138111	MFSD13A	79847	major facilitator superfamily domain containing 13A [Source:HGNC Symbol;Acc:HGNC:26196]
ENSG00000092010	PSME1	5720	proteasome activator subunit 1 [Source:HGNC Symbol;Acc:HGNC:9568]

使用 KEGG 的数据库富集分析 (结果 `res.kegg@result` 见 Tab. 2)。

```
res.kegg <- clusterProfiler::enrichKEGG(res.top30.ex$entrezid)
```

Table 2: KEGG 通路富集分析结果

ID	p.adjust	geneID	Description
hsa03050	0.00931009787538772	5720	Proteasome
hsa04612	0.00931009787538772	5720	Antigen processing and presentation

注: 'geneID' 为 'entrezid'

**3.4.5.2 使用 pathview 将富集分析结果绘制成通路。** 这里选择将  $\log_2(\text{FC})$  数据可视化在通路图中, 因为 pathview 仅支持范围  $[-1, 1]$ , 所以先将  $\log_2(\text{FC})$  归一化到  $[-1, 1]$ 。

```
gene.data <- res.top30.ex$logFC / max(abs(res.top30.ex$logFC))
names(gene.data) <- res.top30.ex$entrezid
pathways <- gsub("[a-z]*", "", res.kegg@result$ID)
```

绘制显著的两条通路图。

```
data(bods, package = "pathview")
res.pathv <- sapply(pathways, simplify = F,
  function(id) {
    pathview::pathview(gene.data, pathway.id = id, species = "hsa")
  })
```

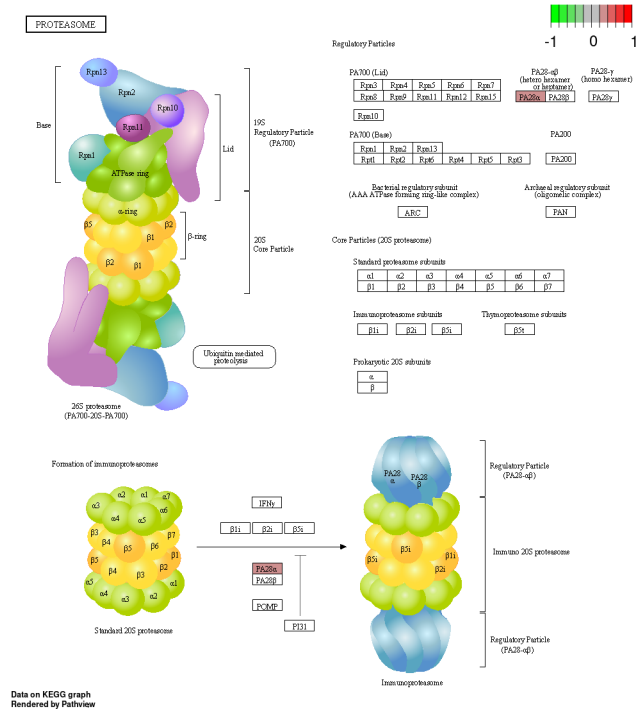


Figure 6: 通路'Proteasome' (ID:hsa03050)

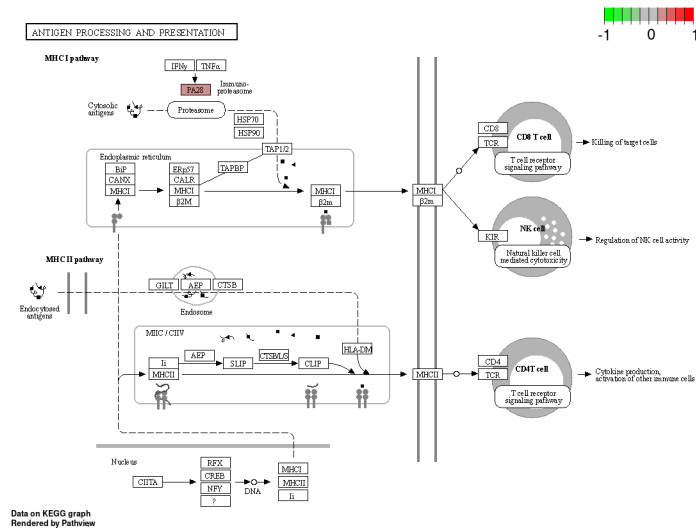


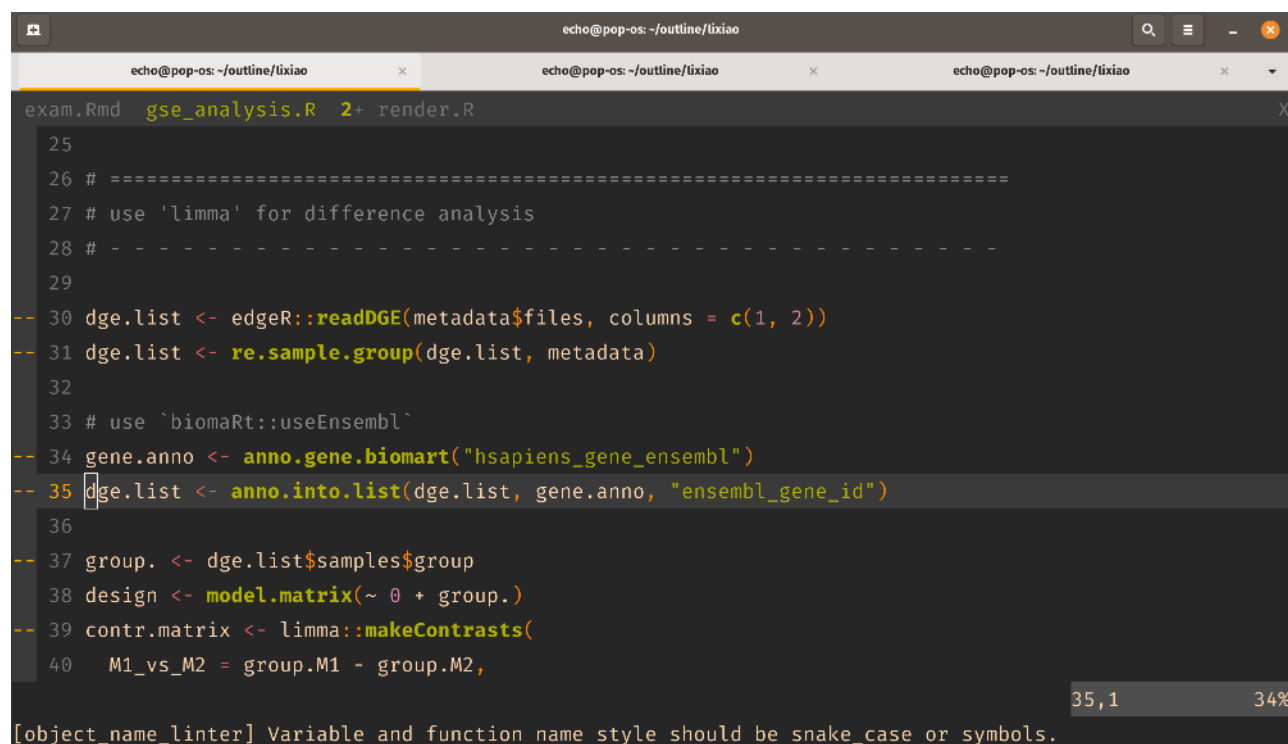
Figure 7: 通路'Antigen processing and presentation' (ID:hsa04612)



### 3.4.6 结论与临床转化

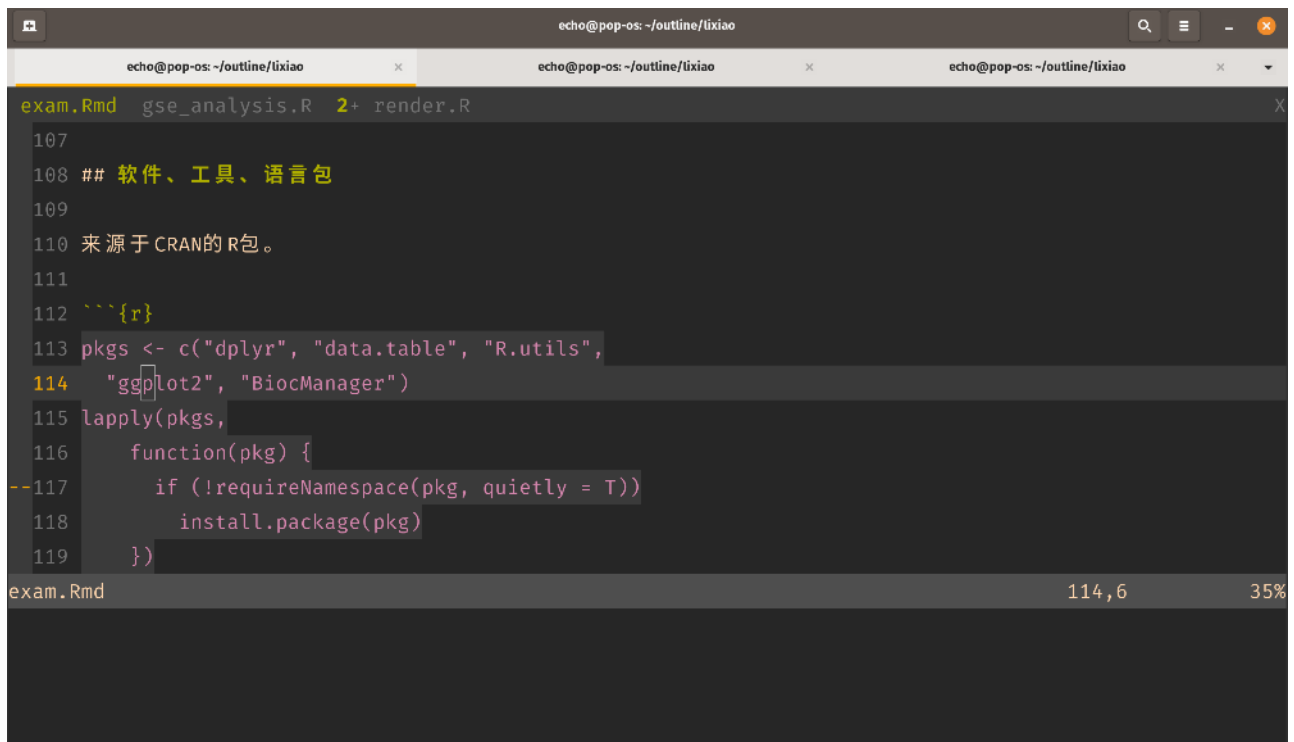
本次分析目的为探究类风湿性关节炎单核细胞衍生巨噬细胞的促炎症和代谢功能。巨噬细胞的激活和极化在炎症发展的前馈和反馈机制中起到了重要作用<sup>4</sup>，了解其极化后功能的改变对治疗类风湿性关节炎具有启发意义。本次分析结果表明，相比于极化健康巨噬细胞（M2, IL-4 处理组），类风湿性关节炎单核细胞衍生的巨噬细胞（M1, LPC + IFN 处理组）具有显著高表达的‘PA28’基因（entrezid: 5720，见 Tab. 1），富集于‘Antigen processing and presentation’和‘Proteasome’通路（Tab. 2）。‘PA28’基因可能成为治疗类风湿性关节炎的临床用药靶点。

## 3.5 附：操作截图



```
exam.Rmd  gse_analysis.R  2+  render.R
25
26 # =====
27 # use 'limma' for difference analysis
28 # - - - - -
29
-- 30 dge.list <- edgeR::readDGE(metadata$files, columns = c(1, 2))
-- 31 dge.list <- re.sample.group(dge.list, metadata)
32
33 # use `biomaRt::useEnsembl`
-- 34 gene.anno <- anno.gene.biomart("hsapiens_gene_ensembl")
-- 35 dge.list <- anno.into.list(dge.list, gene.anno, "ensembl_gene_id")
36
-- 37 group. <- dge.list$samples$group
38 design <- model.matrix(~ 0 + group.)
-- 39 contr.matrix <- limma::makeContrasts(
40   M1_vs_M2 = group.M1 - group.M2,
                                     35,1  34%
[object_name_linter] Variable and function name style should be snake_case or symbols.
```

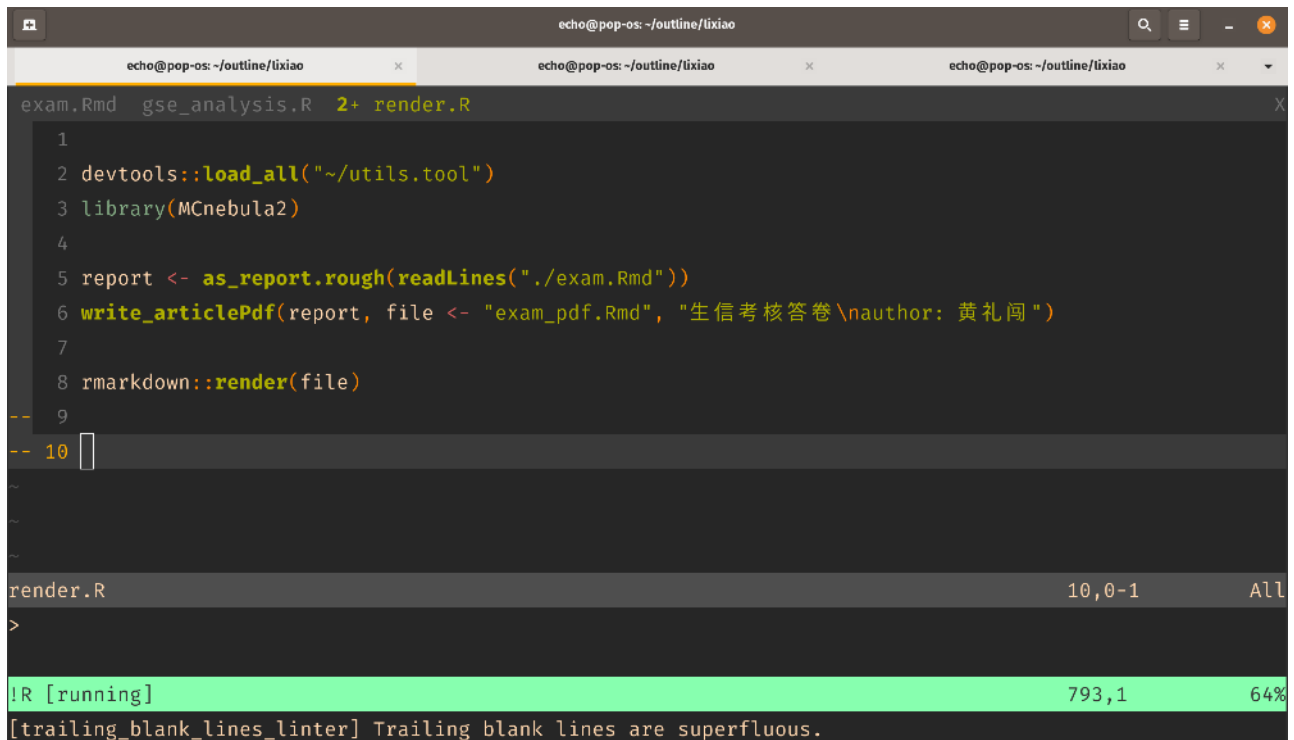
Figure 8: 编写分析脚本（VIM 界面）



```
107
108 ## 软件、工具、语言包
109
110 来源于CRAN的R包。
111
112 ```{r}
113 pkgs <- c("dplyr", "data.table", "R.utils",
114 "ggplot2", "BiocManager")
115 lapply(pkgs,
116   function(pkg) {
117     if (!requireNamespace(pkg, quietly = T))
118       install.package(pkg)
119   })
```

exam.Rmd 114,6 35%

Figure 9: 编写分析报告



```
1
2 devtools::load_all("~/utils.tool")
3 library(MCnebula2)
4
5 report <- as_report.rough(readLines("./exam.Rmd"))
6 write_articlePdf(report, file <- "exam_pdf.Rmd", "生信考核答卷\nauthor: 黄礼闯")
7
8 rmarkdown::render(file)
9
10
```

render.R 10,0-1 All

!R [running] 793,1 64%

[trailing\_blank\_lines\_linter] Trailing blank lines are superfluous.

Figure 10: 输出分析报告为 PDF 格式

```

echo@pop-os: ~/outline/lixiao
$ cp ~/Pictures/Screenshots/Screenshot\ from\ 2023-04-02\ 21-08-19.png writeScript_capture.png
21:09 [pop-os] echo@172.31.130.40 ~/outline/lixiao
$ cp ~/Pictures/Screenshots/Screenshot\ from\ 2023-04-02\ 21-07-54.png writeReport_capture.png
21:10 [pop-os] echo@172.31.130.40 ~/outline/lixiao
$ cp ~/Pictures/Screenshots/Screenshot\ from\ 2023-04-02\ 21-12-43.png outputReport_capture.png
21:13 [pop-os] echo@172.31.130.40 ~/outline/lixiao
$
21:14 [pop-os] echo@172.31.130.40 ~/outline/lixiao
$ ls
analysis_case2.png  exam_pdf.tex          hsa03050.pathview.png  hsa04612.xml
analysis_case.png   exam.Rmd              hsa03050.png           outputReport_capture.png
exam_pdf_files      flowChart.R          hsa03050.xml           render.R
exam_pdf.log        GSE223325            hsa04612.pathview.pdf  thesis_fig
exam_pdf.pdf        gse223325_capture.png hsa04612.pathview.png  writeReport_capture.png
exam_pdf.Rmd        gse_analysis.R        hsa04612.png           writeScript_capture.png
21:14 [pop-os] echo@172.31.130.40 ~/outline/lixiao
$ o exam_pdf.pdf
21:14 [pop-os] echo@172.31.130.40 ~/outline/lixiao
$

```

Figure 11: 工作目录

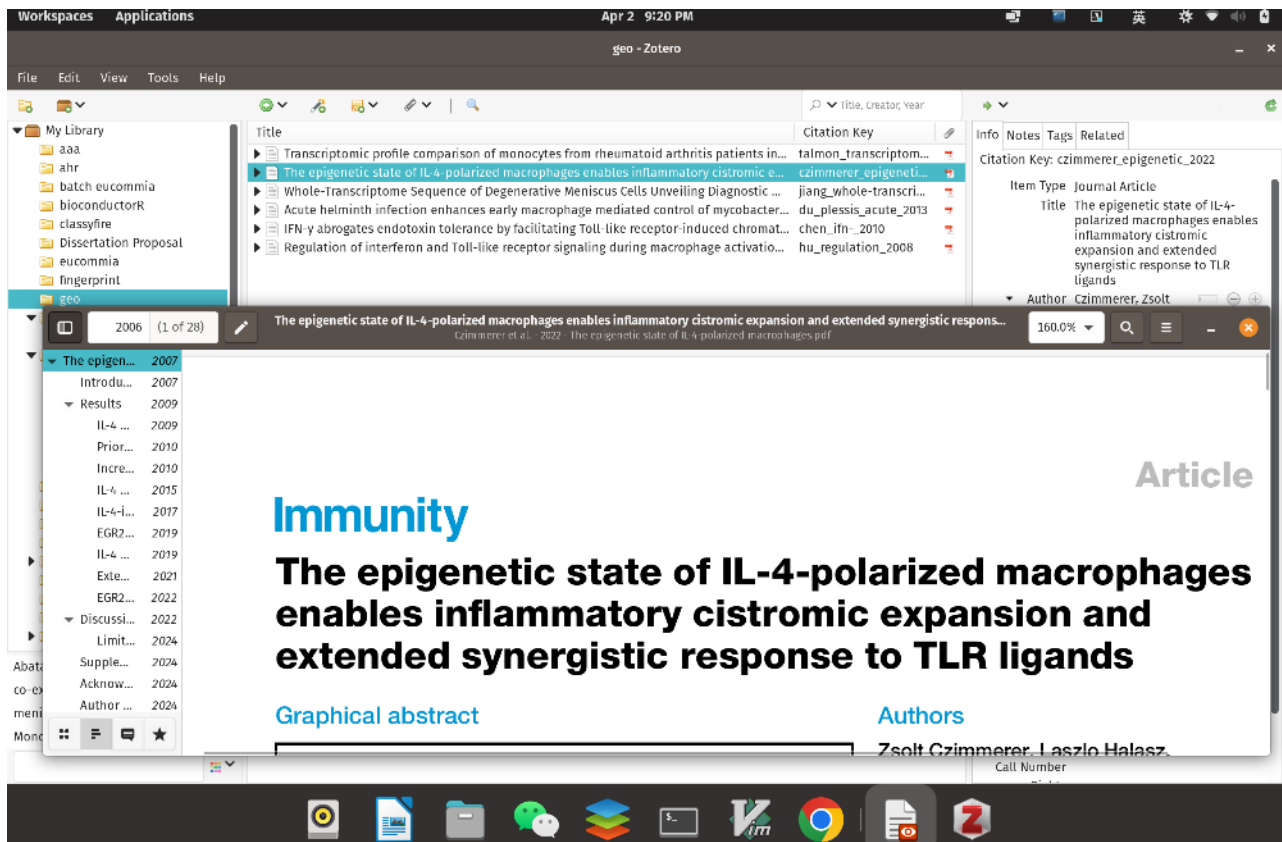


Figure 12: 查阅和管理文献

## Reference

1. Sadik, A. *et al.* IL4I1 Is a Metabolic Immune Checkpoint that Activates the AHR and Promotes Tumor Progression. *Cell* **182**, 1252–1270.e34 (2020).
2. Wozniak, J. M. *et al.* Mortality Risk Profiling of Staphylococcus aureus Bacteremia by Multi-omic Serum Analysis Reveals Early Predictive and Pathogenic Signatures. *Cell* **182**, 1311–1327.e14 (2020).
3. Czimmerer, Z. *et al.* The epigenetic state of IL-4-polarized macrophages enables inflammatory cistromic expansion and extended synergistic response to TLR ligands. *Immunity* **55**, 2006–2026.e6 (2022).
4. Hu, X., Chakravarty, S. D. & Ivashkiv, L. B. Regulation of interferon and Toll-like receptor signaling during macrophage activation by opposing feedforward and feedback inhibition mechanisms. *Immunological Reviews* **226**, 41–56 (2008).
5. Chen, J. & Ivashkiv, L. B. IFN- $\gamma$  abrogates endotoxin tolerance by facilitating Toll-like receptor-induced chromatin remodeling. *Proceedings of the National Academy of Sciences* **107**, 19438–19443 (2010).
6. Law, C. W. *et al.* A guide to creating design matrices for gene expression experiments. *F1000Research* **9**, 1444 (2020).