RNAseq数据，下载GEO中的FPKM文件后该怎么下游分析

原创 泥人吴 生信技能树 2019-12-13

收录于话题
#RNA 36 #GEO 27

我们有很多学徒数据挖掘任务，已经完成的目录见：学徒数据挖掘专题半年目录汇总(生信菜鸟团周一见) 欢迎大家加入我们的学习团队，下面看FPKM文件后该怎么下游分析

- 文献标题是:Oncogenic lncRNA downregulates cancer cell antigen presentation and intrinsic tumor suppression不过不需要看文章，大家只需要做差异分析即可，这个时候需要注意的是，作者提供的是RPKM值表达矩阵！
- 6个样本，分成2组，是RPKM值表达矩阵，做差异分析，看GO通路，跟文章比较
- 作业:(f) Enrichment of GO biological process (BP) terms for up-regulated genes (red) and down-regulated genes in tumor versus normal samples (n = 3, 3 animals). (g-i) Log2 of fold changes of indicated metabolites in MMTV-Tg(LINK-A) breast tumor compared to that of Tg(LINK-A) mammary gland (n = 3 animals respectively).
- 首先需要去GEO数据库下载文件GSE113143_Normal_Tumor_Expression.tab.gz

1.下载数据GSE113143并加载数据

```
a=read.table('GSE113143_Normal_Tumor_Expression.tab.gz',sep='\t',quote = "",fill = T,
             comment.char = "!",header = T) # 提取表达矩阵
rownames(a)=a[,1]
a <- a[,-1]
```

- TPM值就是RPKM的百分比:关于TPM的解释可以看看这个
- What the FPKM? A review of RNA-Seq expression units
- Question: Differential expression analysis starting from TPM data

---

2.将FPKM转换为TPM

```
expMatrix <- a
fpkmToTpm <- function(fpkm)
{
  exp(log(fpkm) - log(sum(fpkm)) + log(1e6))
}
tpms <- apply(expMatrix,2,fpkmToTpm)
tpms[1:3,]
colSums(tpms)
#输出结果:
> tpms[1:3,]
                  N1      N2    N3    T1    T2    T3
0610005C13Rik  0.232  0.1715  0.00  0.00  0.00  0.00
0610007P14Rik 48.391 39.2632 46.04 50.04 59.05 67.29
0610009B22Rik 47.491 58.5954 54.27 49.79 53.13 58.00
> colSums(tpms)
   N1    N2    N3    T1    T2    T3
1e+06 1e+06 1e+06 1e+06 1e+06 1e+06
```

3.差异分析

```
group_list=c(rep('Normal',3),rep('Tumor',3))
## 强制限定顺序
group_list <- factor(group_list,levels = c("Normal","Tumor"),ordered = F)
#表达矩阵数据校正
exprSet <- tpms
boxplot(exprSet,outline=FALSE, notch=T,col=group_list, las=2)
library(limma)
exprSet=normalizeBetweenArrays(exprSet)
boxplot(exprSet,outline=FALSE, notch=T,col=group_list, las=2)
#判断数据是否需要转换
exprSet <- log2(exprSet+1)
#差异分析:
dat <- exprSet
design=model.matrix(~factor( group_list ))
fit=lmFit(dat,design)
fit=eBayes(fit)
options(digits = 4)
topTable(fit,coef=2,adjust='BH')
bp=function(g){
  library(ggpubr)
  df=data.frame(gene=g,stage=group_list)
  p <- ggboxplot(df, x = "stage", y = "gene",
                 color = "stage", palette = "jco",
                 add = "jitter")
  #  Add p-value
  p + stat_compare_means()
}
deg=topTable(fit,coef=2,adjust='BH',number = Inf)
head(deg)
#save(deg,file = 'deg.Rdata')
```

划重点：以下代码、方法全来自生信技能树的最新推文：为R包写一本书（向Y叔致敬）

这里面重点就是：RPKM矩阵可以转为TPM后,再使用limma进行差异分析哦!

4.做完差异分析

- GEO数据挖掘代码，很容易得到上下调基因，而且转为ENTREZID，后续分析都以这个为主线。
- 根据原文文献中： Differential gene expression was defined if the fold change >1.5 and P < 0.05 between tumor and normal samples 找差异基因

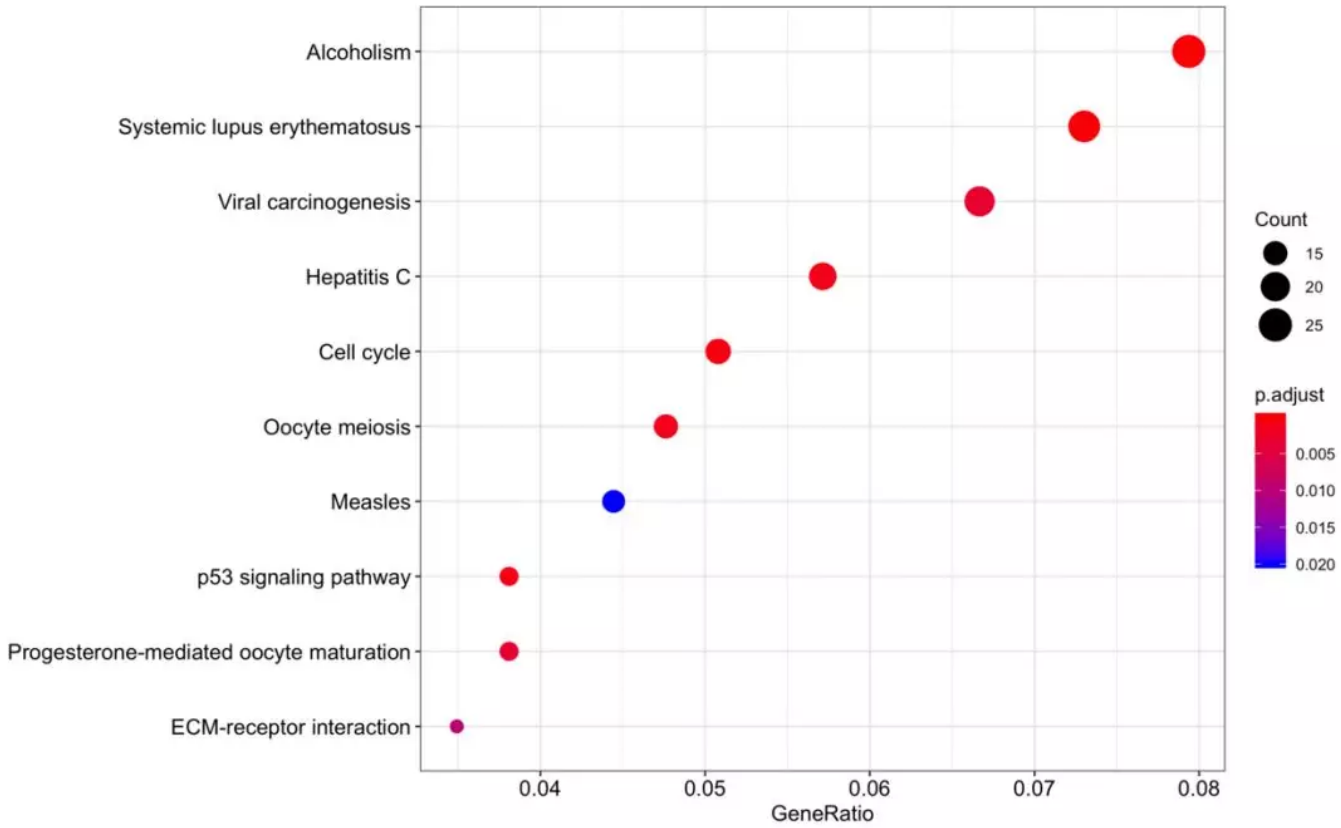
```
## 不同的阈值，筛选到的差异基因数量就不一样，后面的超几何分布检验结果就大相径庭。
if(T){
  logFC_t=1.5
  deg$g=ifelse(deg$P.Value>0.05,'stable',
               ifelse( deg$logFC > logFC_t,'UP',
                       ifelse( deg$logFC < -logFC_t,'DOWN','stable') )
  )
  table(deg$g)
  head(deg)
  deg$symbol=rownames(deg)
  library(ggplot2)
  library(clusterProfiler)
  library(org.Mm.eg.db)
  df <- bitr(unique(deg$symbol), fromType = "SYMBOL",
             toType = c( "ENTREZID"),
             OrgDb = org.Mm.eg.db)
  head(df)
  DEG=deg
  head(DEG)

  DEG=merge(DEG,df,by.y='SYMBOL',by.x='symbol')
  head(DEG)
```

```
  save(DEG,file = 'anno_DEG.Rdata')
  gene_up= DEG[DEG$g == 'UP','ENTREZID']
  gene_down=DEG[DEG$g == 'DOWN','ENTREZID']
}
```

5.最简单的超几何分布检验

```
# 最简单的超几何分布检验
###这里就拿KEGG数据库举例吧，拿自己判定好的上调基因集进行超几何分布检验，如下
if(T){
  gene_down
  gene_up
  enrichKK <- enrichKEGG(gene         =  gene_up,
                         organism     = 'mmu',
                         #universe     = gene_all,
                         pvalueCutoff = 0.05,
                         qvalueCutoff =0.05)
  head(enrichKK)[,1:6]
  browseKEGG(enrichKK, 'hsa04512')
  dotplot(enrichKK)
  ggsave("enrichKK.png")
  enrichKK=DOSE::setReadable(enrichKK, OrgDb='org.Mm.eg.db',keyType='ENTREZID')
  enrichKK
}
##最基础的条形图和点图
#条带图
barplot(enrichKK,showCategory=20)
#气泡图
dotplot(enrichKK)
```



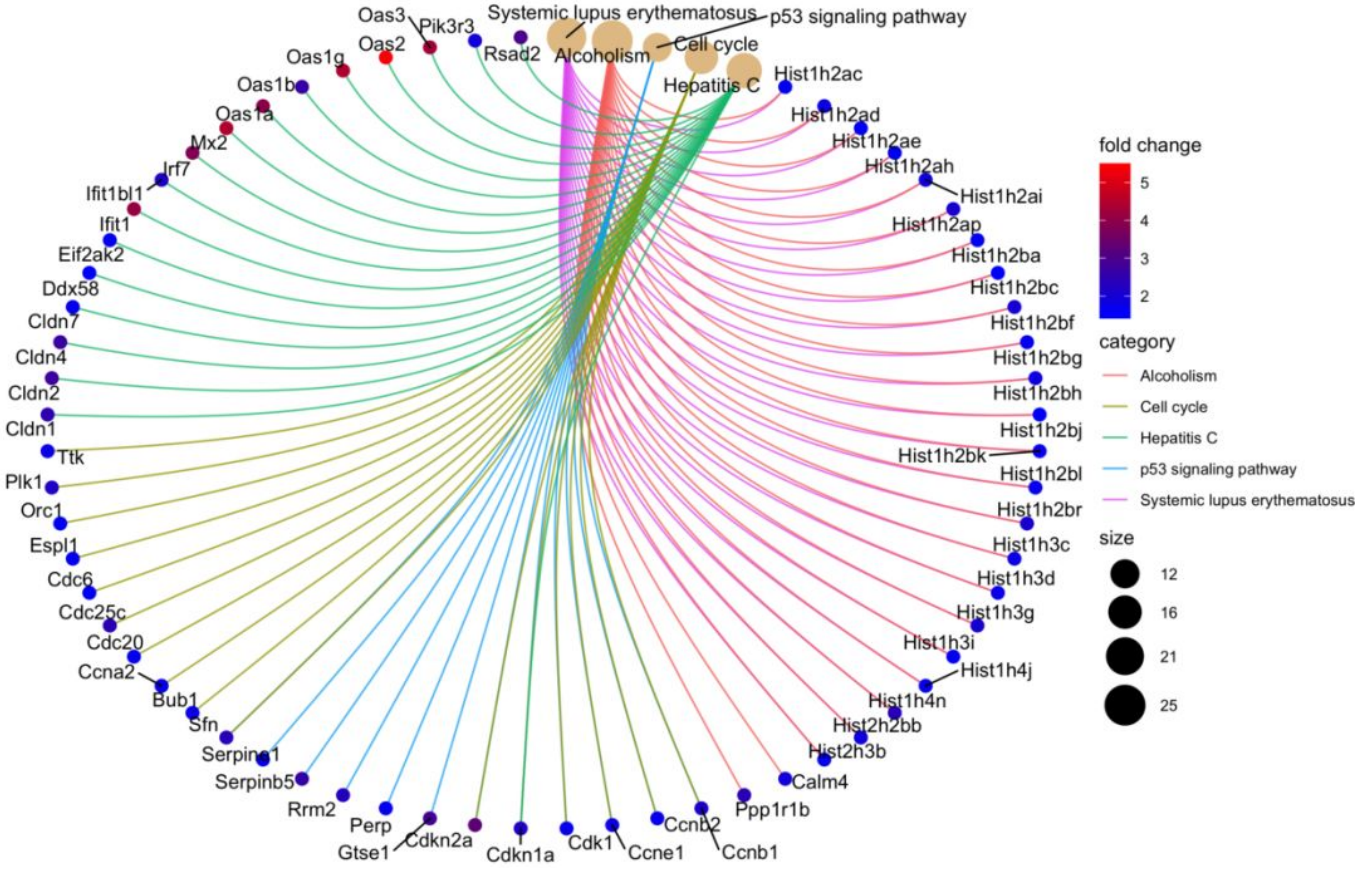
enrichKK.png

- 通路与基因之间的关系可视化

```
#通路与上调基因之间的关系可视化
###制作genlist三部曲：
## 1.获取基因logFC
DEG_up <- DEG[DEG$g == 'UP',]
geneList <- DEG_up$logFC
## 2.命名
names(geneList) = DEG_up$ENTREZID
## 3.排序很重要
geneList = sort(geneList, decreasing = TRUE)
head(geneList)

cnetplot(enrichKK, categorySize="pvalue", foldChange=geneList,colorEdge = TRUE)
```

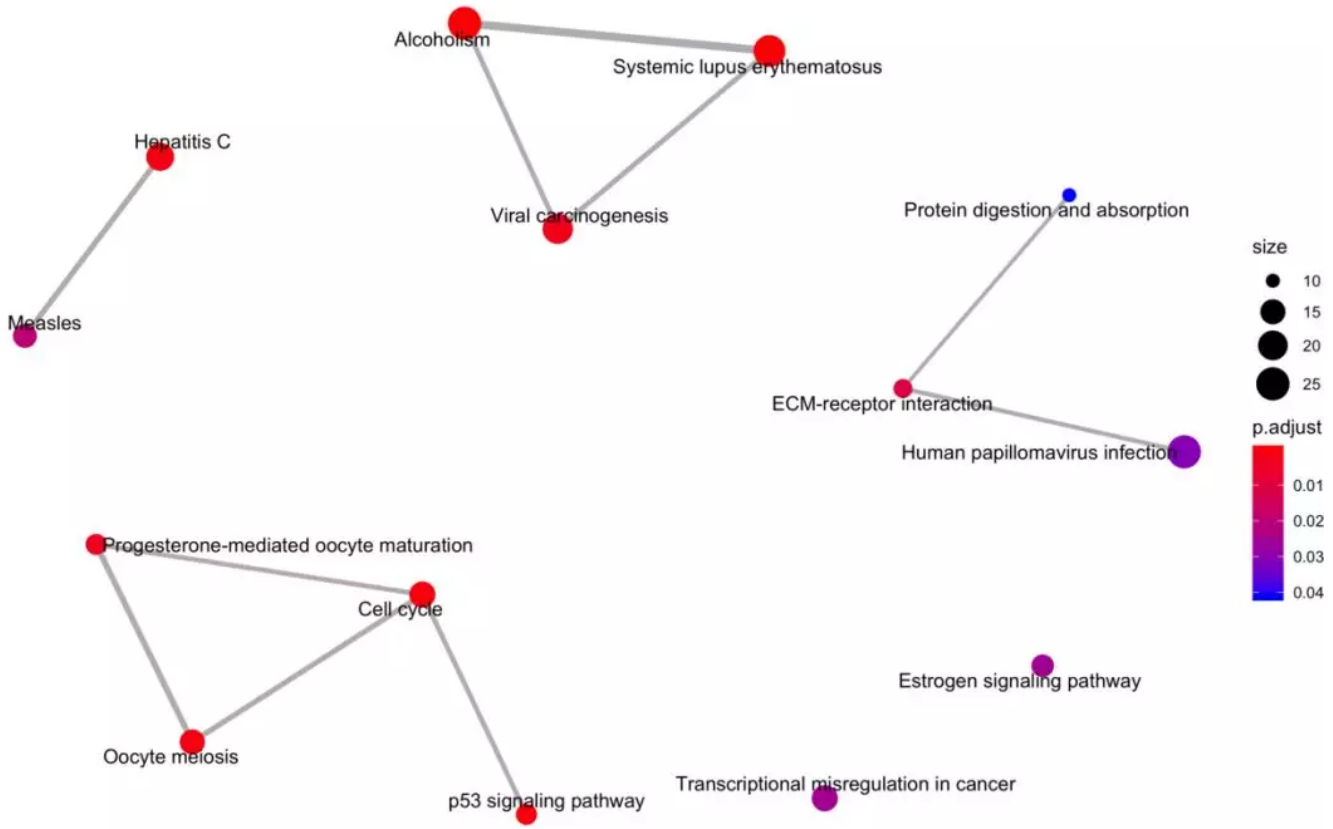
```
cnetplot(enrichKK, foldChange=geneList, circular = TRUE, colorEdge = TRUE)
ggsave("enrichKK_cnetplot.png")
```



enrichKK_cnetplot.png

- 通路与通路之间的连接展示

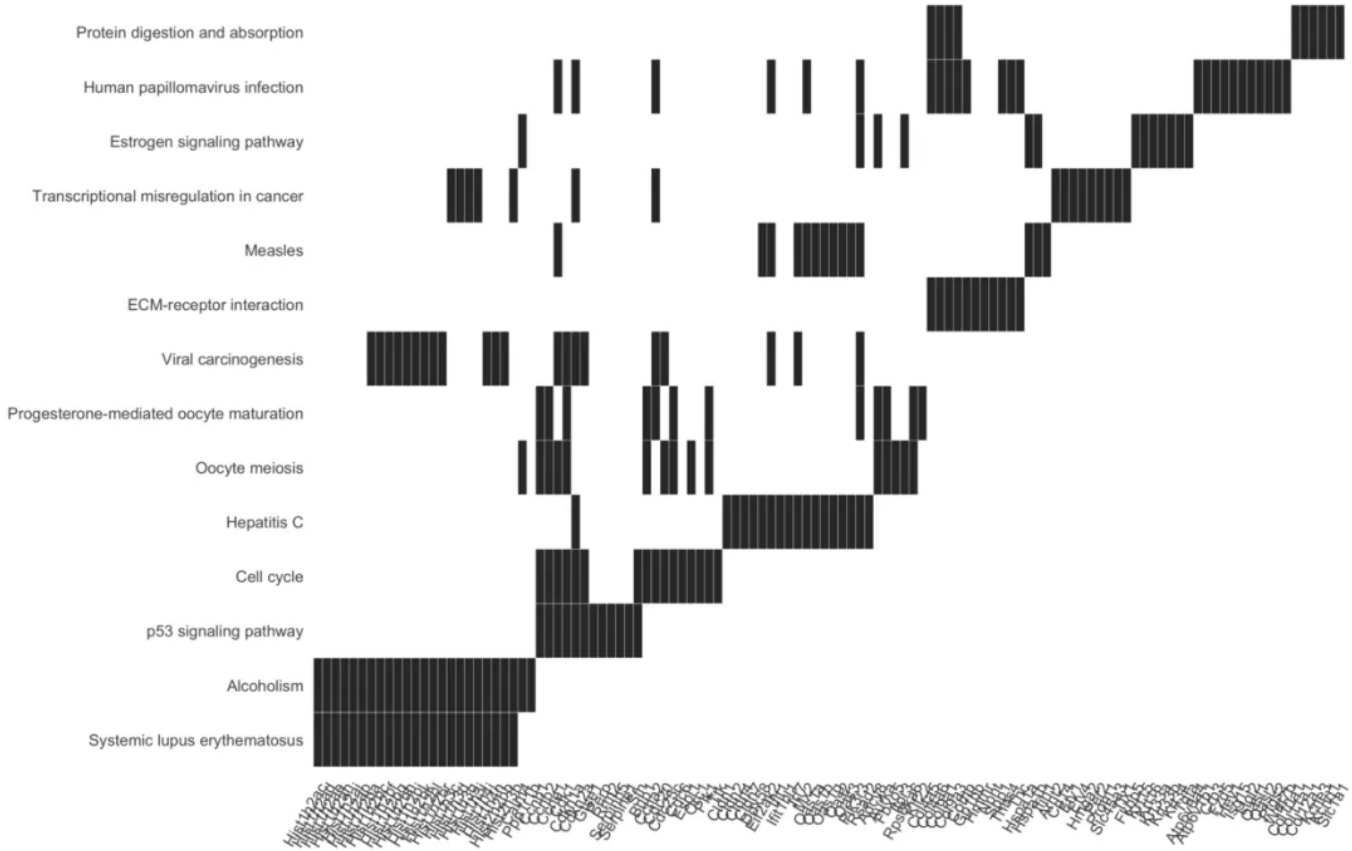
#通路与通路之间的连接展示
```
emapplot(enrichKK)
ggsave("enrichKK_emapplot.png")
```



enrichKK_emapplot.png

- 热图展现通路与基因之间的关系

```
#热图展现通路与基因之间的关系
heatplot(enrichKK)
ggsave("enrichKK_heatplot.png")
```

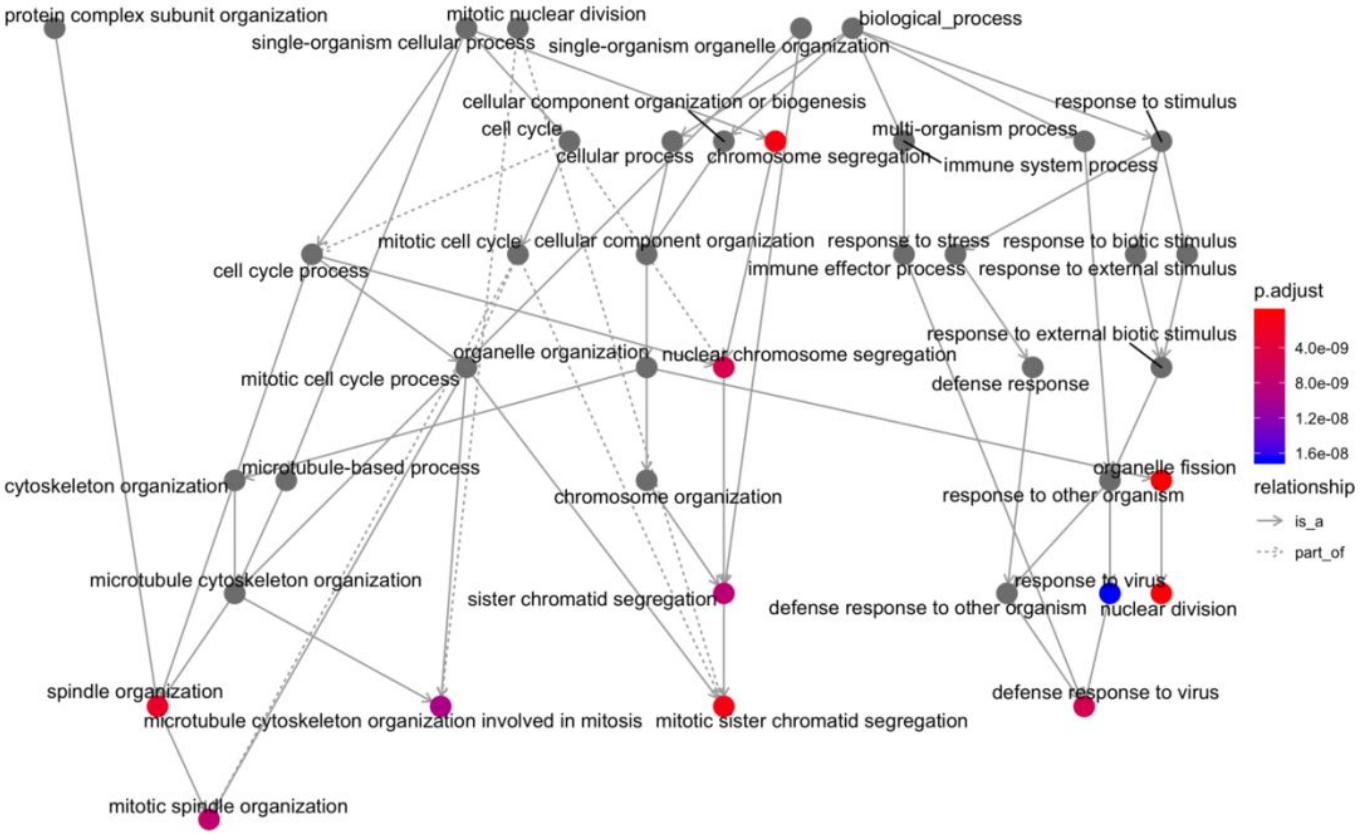


enrichKK_heatplot.png

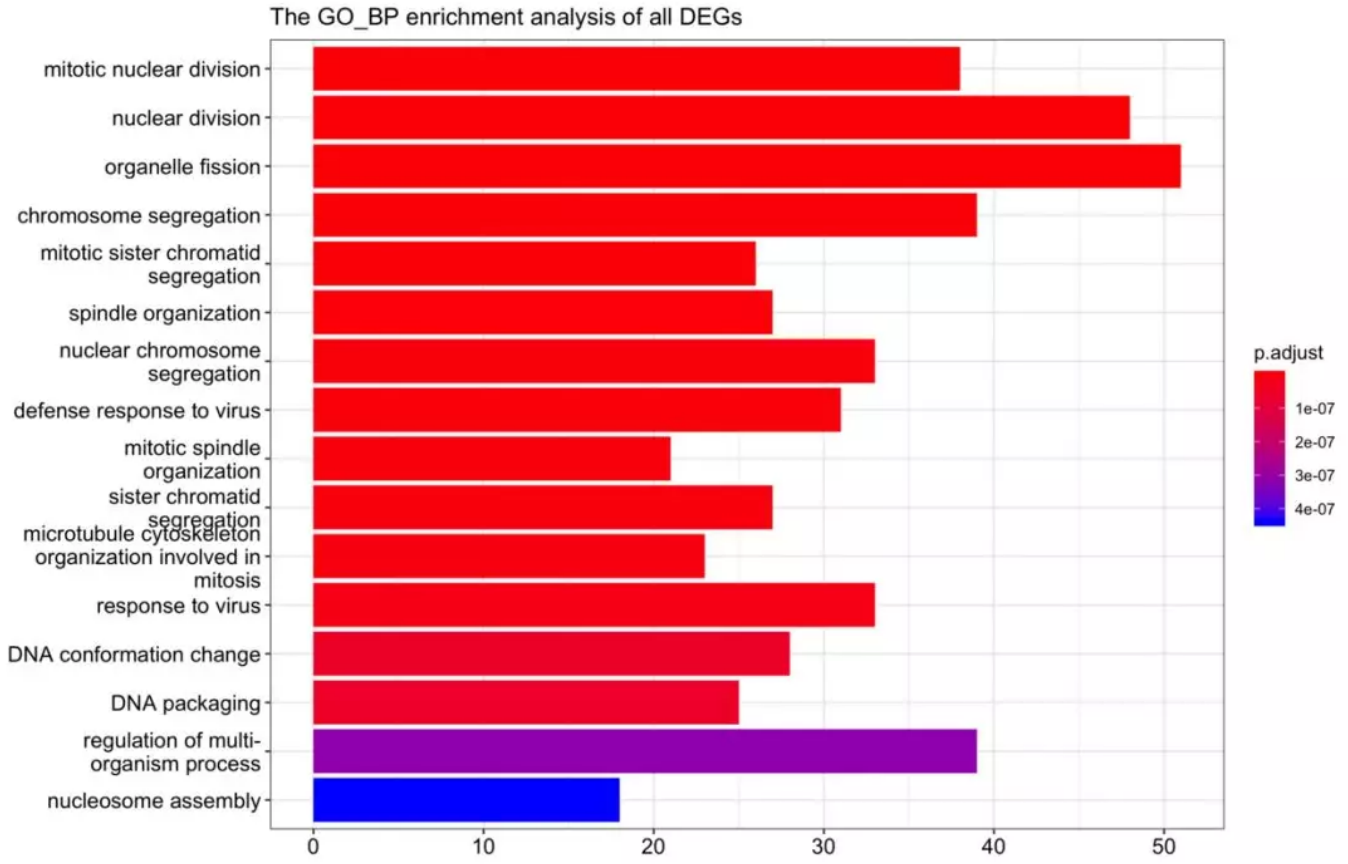
- 如果你是做GO数据库呢，其实还有一个goplot可以试试看，当然是以Y叔的书为主啦。

```
#如果你是做GO数据库呢，其实还有一个goplot可以试试看
ego_bp_up<-enrichGO(gene         = DEG_up$ENTREZID,
                OrgDb      = org.Mm.eg.db,
                keyType    = 'ENTREZID',
                ont        = "BP",
                pAdjustMethod = "BH",
                pvalueCutoff = 0.01,#0.01
                qvalueCutoff = 0.05)
goplot(ego_up)
ggsave("ego_bp_up_goplot.png")
head(ego)
library(stringr)
```

```
barplot(ego_bp_up,showCategory = 16,title="The GO_BP enrichment analysis of all DEGs ")+
  scale_size(range=c(2, 12))+
  scale_x_discrete(labels=function(ego_bp) str_wrap(ego_bp,width = 25))
ggsave("ego_bp_up_barplot.png")
```

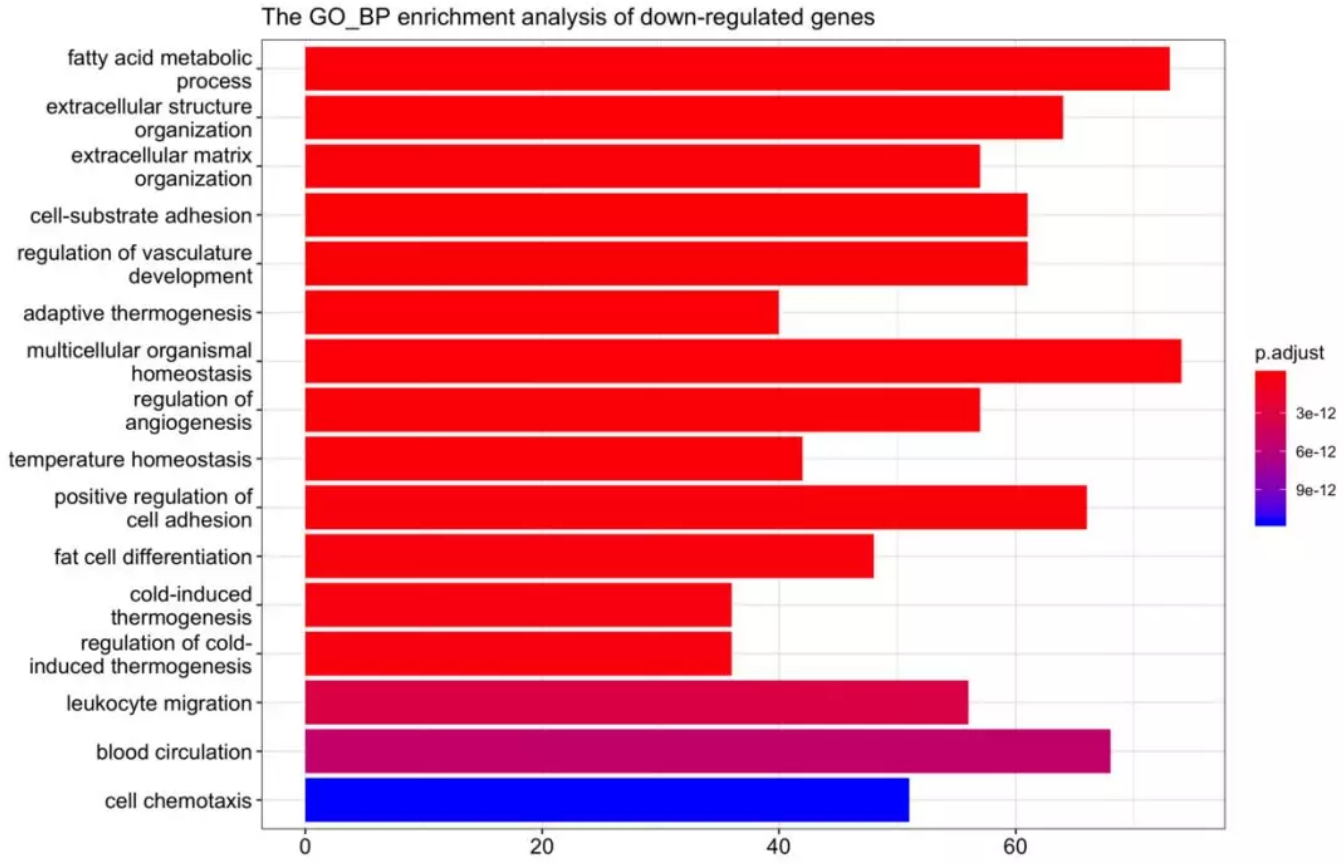


ego_up_goplot.png



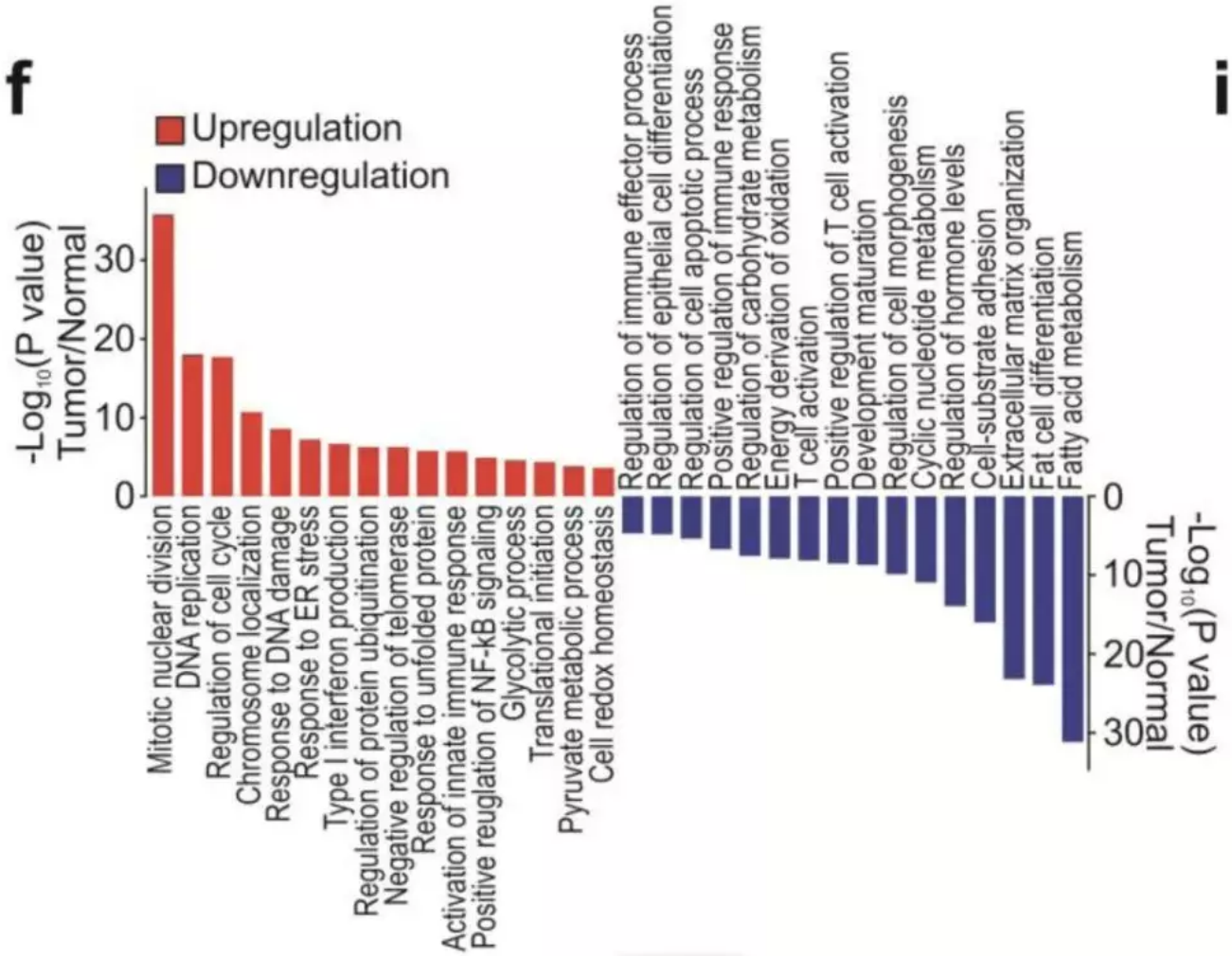
ego_up_barplot.png

- 同样的方式看看下调基因的GO_BP:



down_regulated_genes.png

- **和文献中的GO_BP比较一下**



友情宣传

- 生信入门课全国巡讲2019收官--长沙站

- 广州专场（全年无休）GEO数据挖掘课，带你起飞

收录于话题 #GEO·27个

上一篇
GEO数据库中国区镜像奔走相告啊

下一篇
GEO数据库中国区镜像横空出世