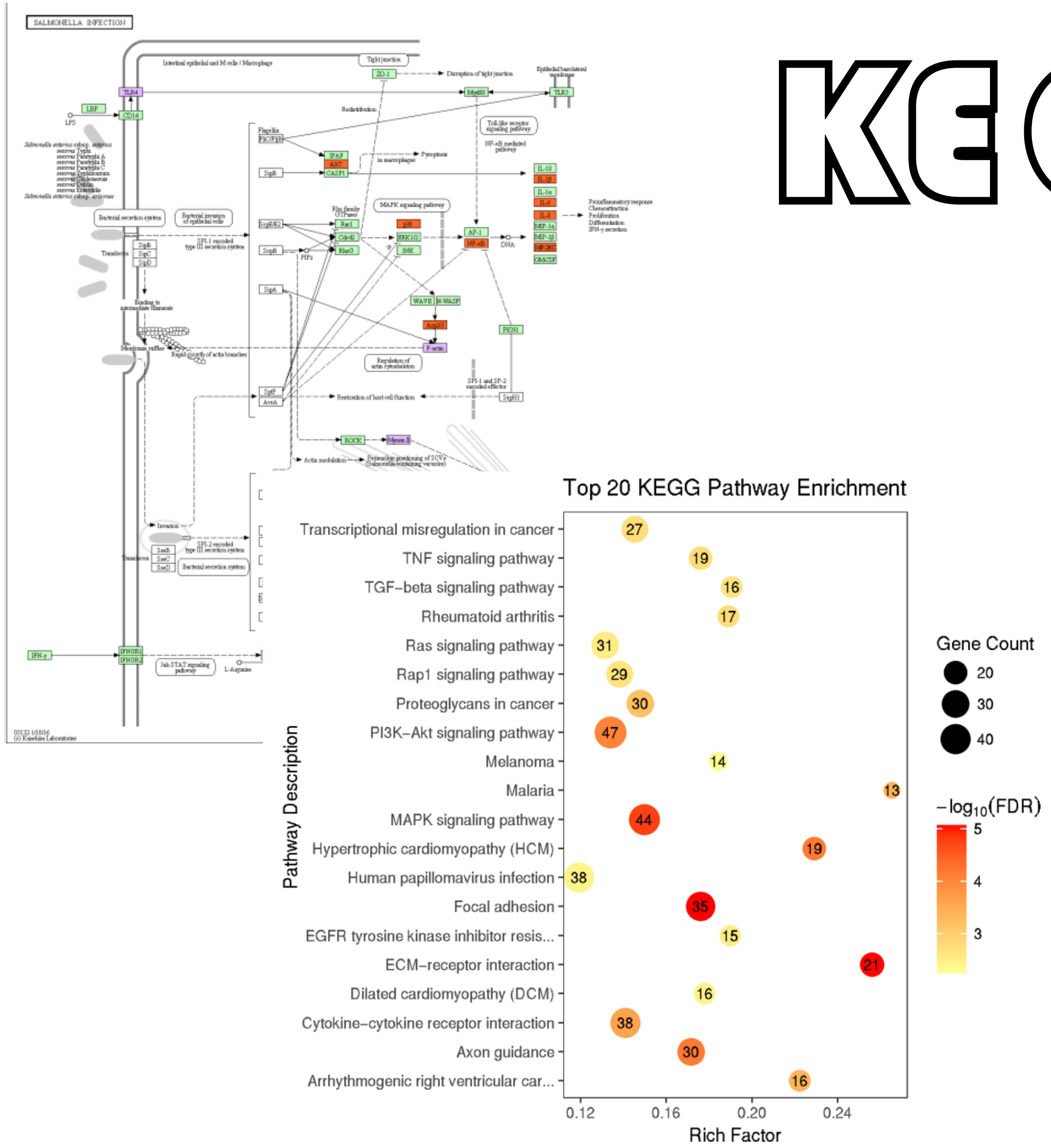
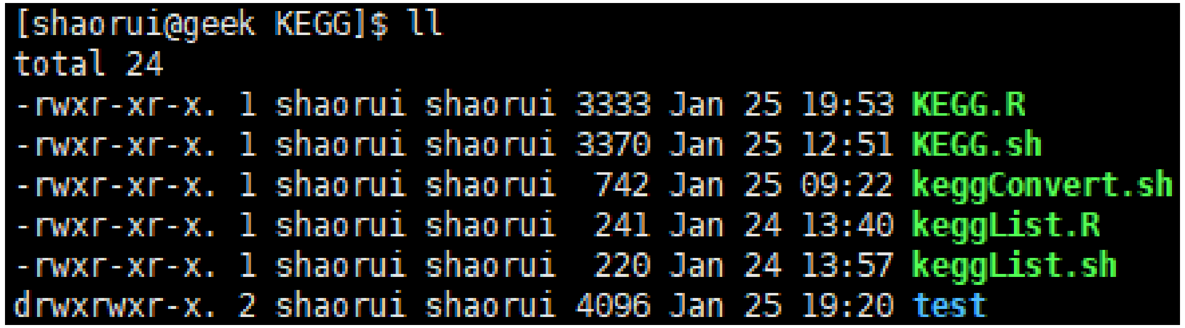
KEGG 结题报告

lsr —— 2018.1.26



经过八周的练手，趟过无数的坑，终于做出一点东西来了。实战是最好的训练，虽然还是很菜，但如今遇到问题不会像最开始那样无从着手了。

后面是相关脚本的总结说明。



```
[shaorui@geek KEGG]$ ll
total 24
-rwxr-xr-x. 1 shaorui shaorui 3333 Jan 25 19:53 KEGG.R
-rwxr-xr-x. 1 shaorui shaorui 3370 Jan 25 12:51 KEGG.sh
-rwxr-xr-x. 1 shaorui shaorui  742 Jan 25 09:22 keggConvert.sh
-rwxr-xr-x. 1 shaorui shaorui  241 Jan 24 13:40 keggList.R
-rwxr-xr-x. 1 shaorui shaorui  220 Jan 24 13:57 keggList.sh
drwxrwxr-x. 2 shaorui shaorui 4096 Jan 25 19:20 test
```

KEGG

脚本目录暂时是在： /home/shaorui/Install/KEGG/

富集分析的主体是： KEGG.sh KEGG.R

简易使用 ：

① 环境变量设置（能正常使用R/Rscript）

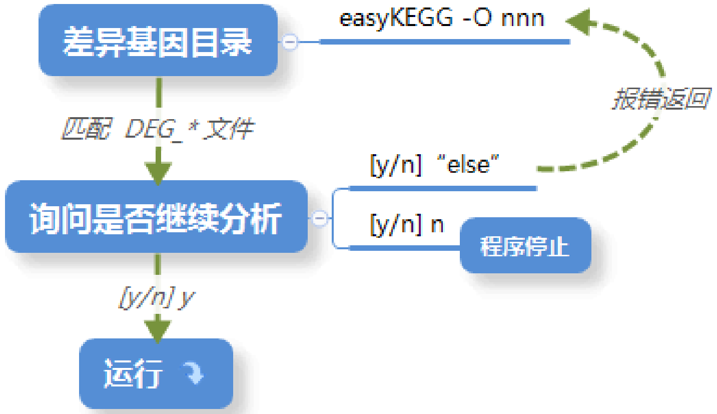
添加如下信息到~/.bash_profile并source

```
# KEGG
export PATH=$PATH:/home/shaorui/Install/KEGG
alias easyKEGG=KEGG.sh
```

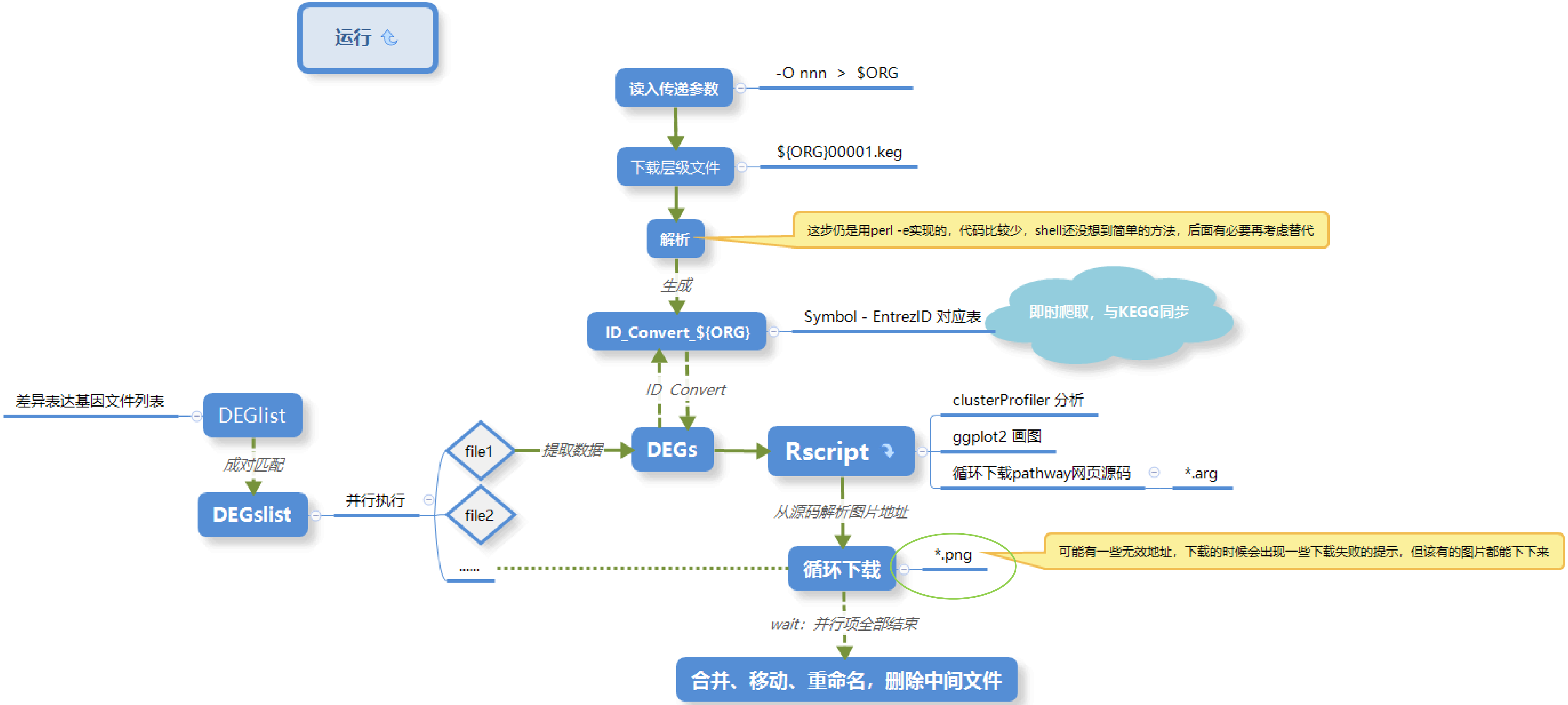
② 安装R依赖包 “KEGGREST” “clusterProfiler” “ggplot2”

③ 运行

$cd 20170712C_Report   #到差异基因文件目录
$easyKEGG –O nnn   # nnn: 3-4个字符，KEGG的org code
“匹配当前目录的DEG_*文件，显示差异基因数，询问是否继续分析”
[y/n]   # 选y继续；选n程序停止；其余输入报错重选
（y：运行时间取决于网速，大约一个小时）



差异基因目录
easyKEGG -O nnn
匹配 DEG_*文件
询问是否继续分析
[y/n] “else”
报错返回
[y/n] n
程序停止
[y/n] y
运行



运行

读入传递参数

-O nnn > $ORG

下载层级文件

${ORG}00001.keg

解析

这步仍是用perl -e实现的，代码比较少，shell还没想到简单的方法，后面有必要再考虑替代

生成

ID_Convert_${ORG}

Symbol - EntrezID 对应表

即时爬取，与KEGG同步

ID Convert

差异表达基因文件列表

DEGlist

成对匹配

DEGslist

并行执行

file1

file2

......

提取数据

DEGs

Rscript

clusterProfiler 分析

ggplot2 画图

循环下载pathway网页源码

*.arg

从源码解析图片地址

循环下载

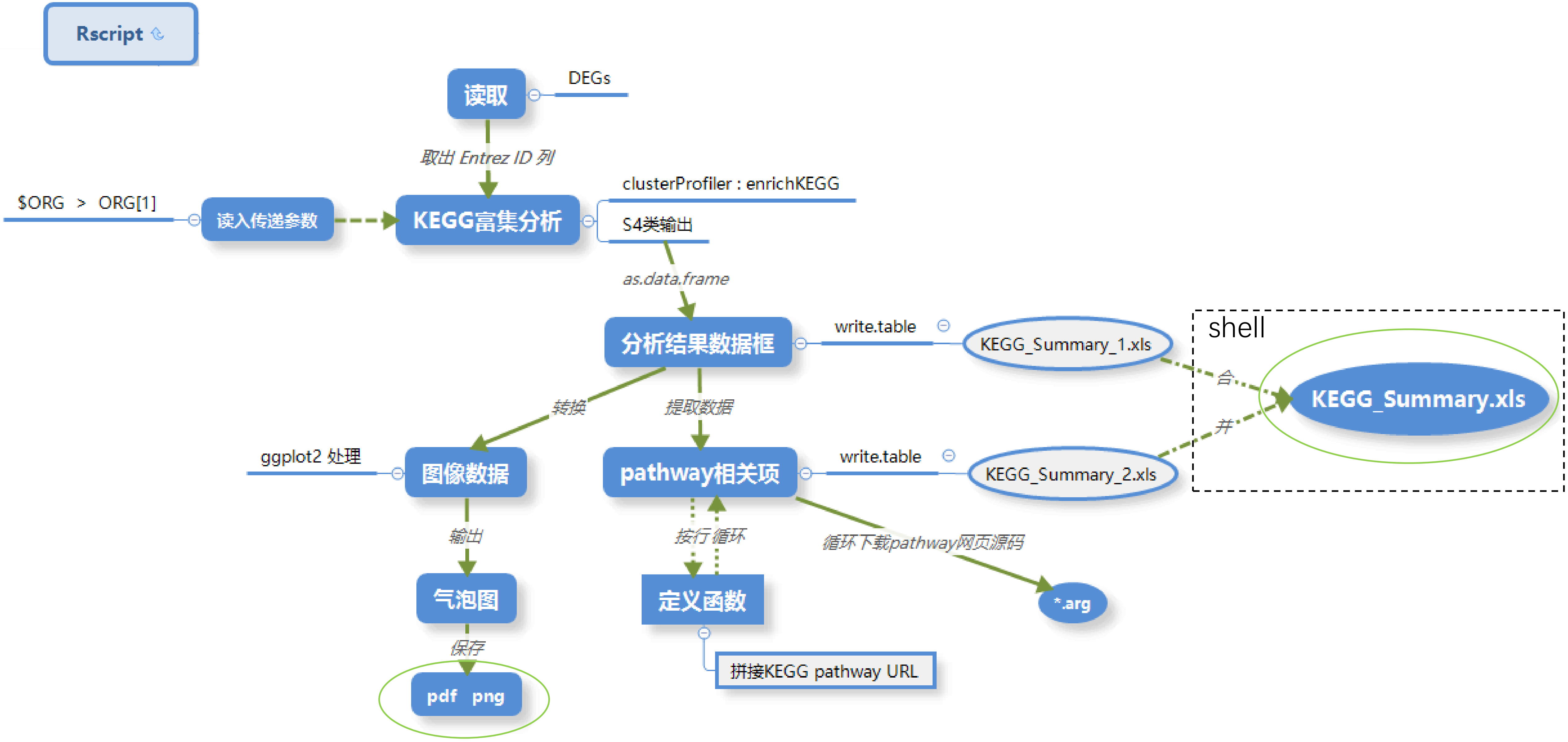
*.png

可能有一些无效地址，下载的时候会出现一些下载失败的提示，但该有的图片都能下下来

wait：并行项全部结束

合并、移动、重命名，删除中间文件

代码详见脚本



Rscript

读取

DEGs

取出 Entrez ID 列

$ORG > ORG[1]

读入传递参数

KEGG富集分析

clusterProfiler : enrichKEGG

S4类输出

as.data.frame

分析结果数据框

write.table

KEGG_Summary_1.xls

shell

合

并

KEGG_Summary.xls

转换

提取数据

ggplot2 处理

图像数据

pathway相关项

write.table

KEGG_Summary_2.xls

输出

按行 循环

循环下载pathway网页源码

气泡图

定义函数

*.arg

保存

拼接KEGG pathway URL

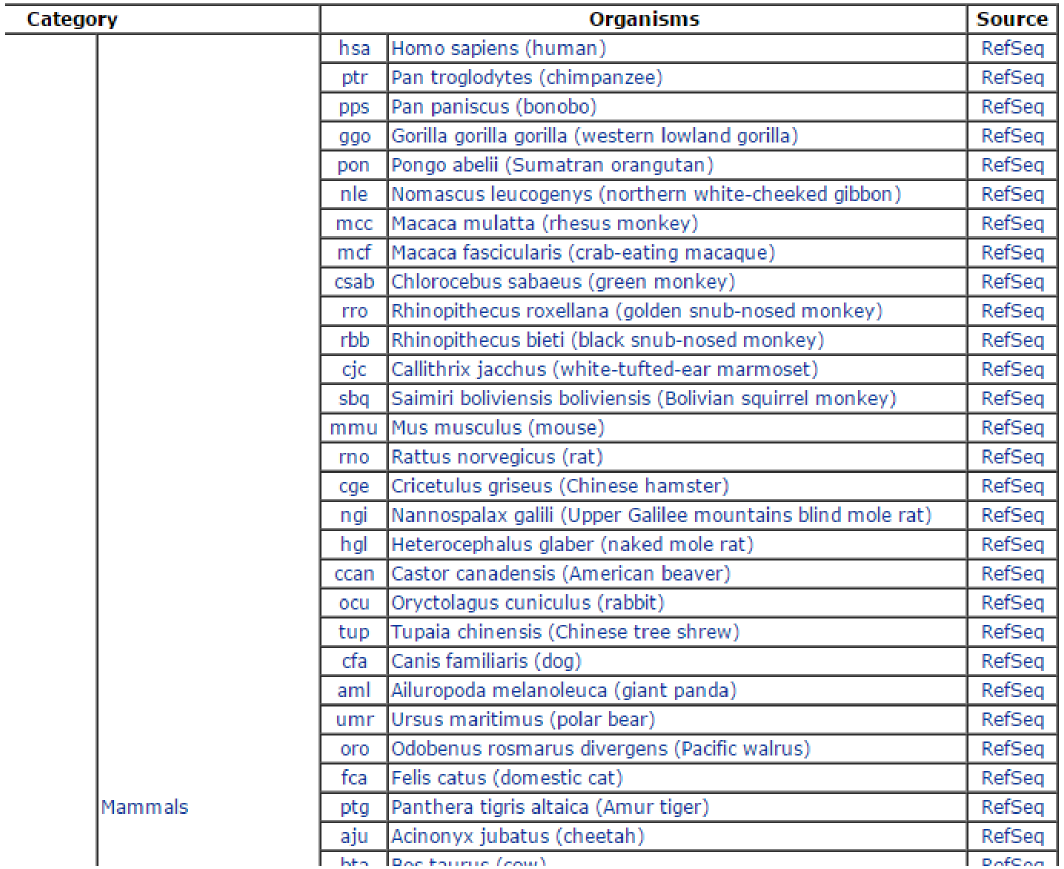
pdf png

代码详见脚本

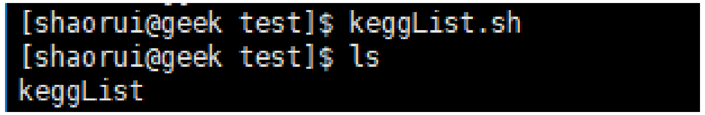
KEGG的org code

网页版见 http://www.kegg.jp/kegg/catalog/org_list.html

或者任意目录直接在命令行敲 keggList.sh ， 会生成一个对应code和organism的keggList文档 (脚本keggList.sh keggList.R在/home/shaorui/Install/KEGG，①已添加目录到环境变量)

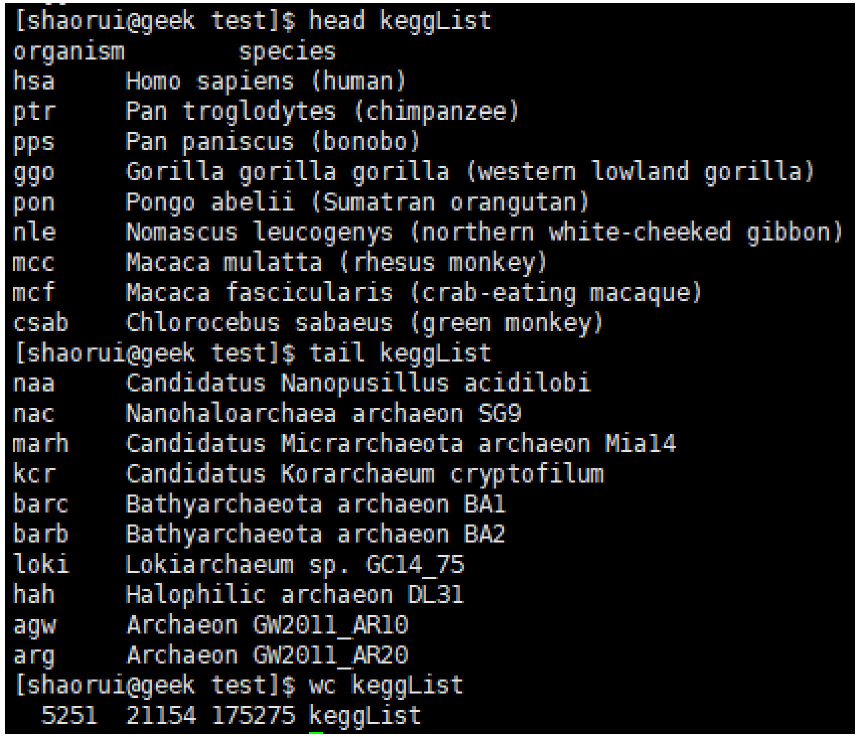


Category	Organisms		Source
Mammals	hsa	Homo sapiens (human)	RefSeq
	ptr	Pan troglodytes (chimpanzee)	RefSeq
	pps	Pan paniscus (bonobo)	RefSeq
	ggo	Gorilla gorilla gorilla (western lowland gorilla)	RefSeq
	pon	Pongo abelii (Sumatran orangutan)	RefSeq
	nle	Nomascus leucogenys (northern white-cheeked gibbon)	RefSeq
	mcc	Macaca mulatta (rhesus monkey)	RefSeq
	mcf	Macaca fascicularis (crab-eating macaque)	RefSeq
	csab	Chlorocebus sabaeus (green monkey)	RefSeq
	rro	Rhinopithecus roxellana (golden snub-nosed monkey)	RefSeq
	rbb	Rhinopithecus bieti (black snub-nosed monkey)	RefSeq
	cjc	Callithrix jacchus (white-tufted-ear marmoset)	RefSeq
	sbq	Saimiri boliviensis boliviensis (Bolivian squirrel monkey)	RefSeq
	mmu	Mus musculus (mouse)	RefSeq
	rno	Rattus norvegicus (rat)	RefSeq
	cge	Cricetulus griseus (Chinese hamster)	RefSeq
	ngi	Nannospalax galili (Upper Galilee mountains blind mole rat)	RefSeq
	hgl	Heterocephalus glaber (naked mole rat)	RefSeq
	ccan	Castor canadensis (American beaver)	RefSeq
	ocu	Oryctolagus cuniculus (rabbit)	RefSeq
	tup	Tupaia chinensis (Chinese tree shrew)	RefSeq
	cfa	Canis familiaris (dog)	RefSeq
	aml	Ailuropoda melanoleuca (giant panda)	RefSeq
	umr	Ursus maritimus (polar bear)	RefSeq
	oro	Odobenus rosmarus divergens (Pacific walrus)	RefSeq
	fca	Felis catus (domestic cat)	RefSeq
	ptg	Panthera tigris altaica (Amur tiger)	RefSeq
	aju	Acinonyx jubatus (cheetah)	RefSeq
	bta	Bos taurus (cow)	RefSeq



```
[shaorui@geek test]$ keggList.sh
[shaorui@geek test]$ ls
keggList
```

5250个物种可用



```
[shaorui@geek test]$ head keggList
organism        species
hsa     Homo sapiens (human)
ptr     Pan troglodytes (chimpanzee)
pps     Pan paniscus (bonobo)
ggo     Gorilla gorilla gorilla (western lowland gorilla)
pon     Pongo abelii (Sumatran orangutan)
nle     Nomascus leucogenys (northern white-cheeked gibbon)
mcc     Macaca mulatta (rhesus monkey)
mcf     Macaca fascicularis (crab-eating macaque)
csab    Chlorocebus sabaeus (green monkey)
[shaorui@geek test]$ tail keggList
naa     Candidatus Nanopusillus acidilobi
nac     Nanohaloarchaea archaeon SG9
marh    Candidatus Micrarchaeota archaeon Mia14
kcr     Candidatus Korarchaeum cryptofilum
barc    Bathyarchaeota archaeon BA1
barb    Bathyarchaeota archaeon BA2
loki    Lokiarchaeum sp. GC14_75
hah     Halophilic archaeon DL31
agw     Archaeon GW2011_AR10
arg     Archaeon GW2011_AR20
[shaorui@geek test]$ wc keggList
  5251  21154 175275 keggList
```

ID Convertion

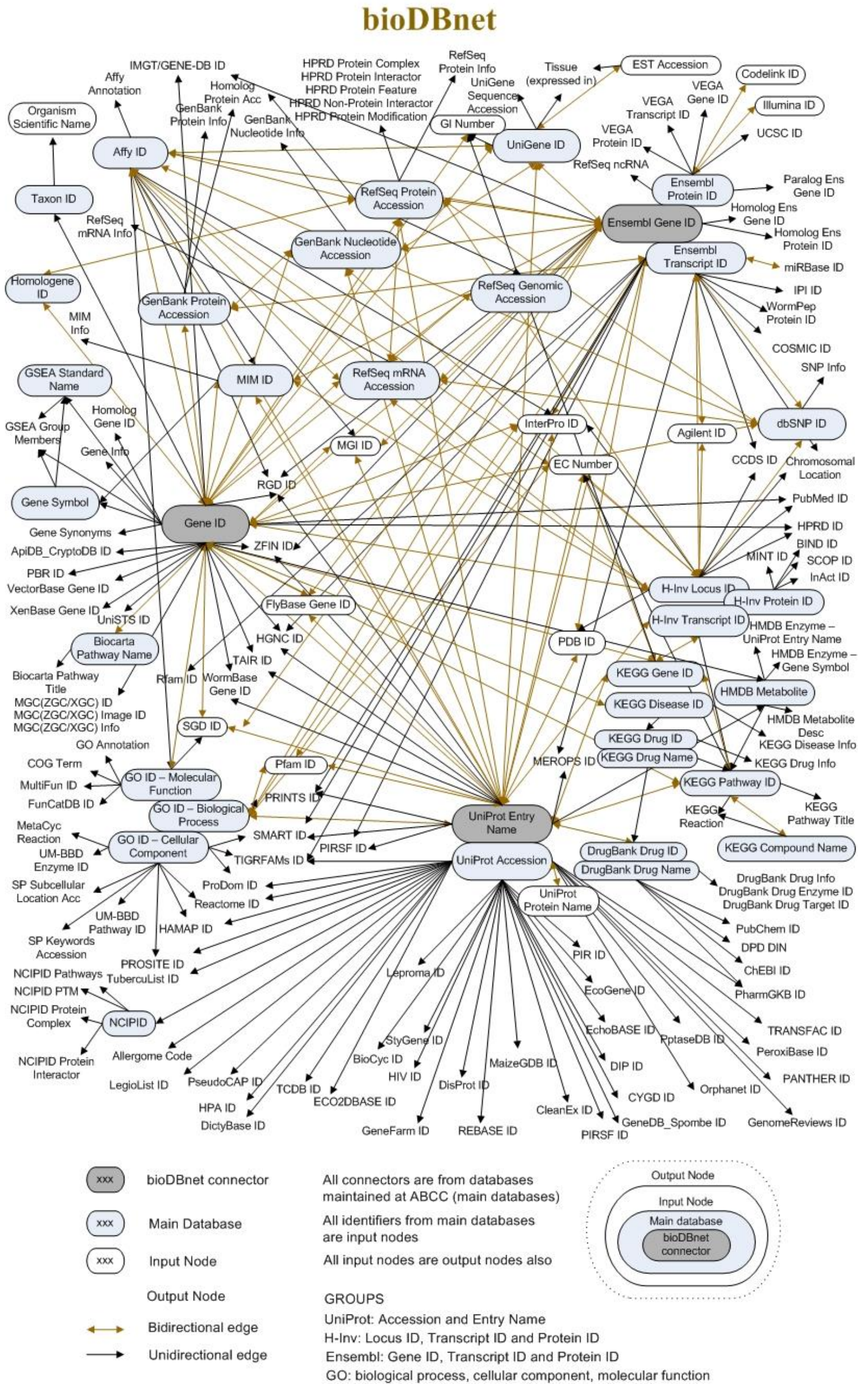
生物数据库的ID非常繁杂
各类型的ID转换据说是生物信息入门级的编程题目

KEGG数据库只支持三种ID进行pathway检索
一般使用 Entrez ID （NCBI Gene ID）
人和小鼠有各自的基因命名委员会 HGNC和MGI，命名比较规范
利用HGNC和MGI的Gene信息可以很方便地进行Symbol-ID互换

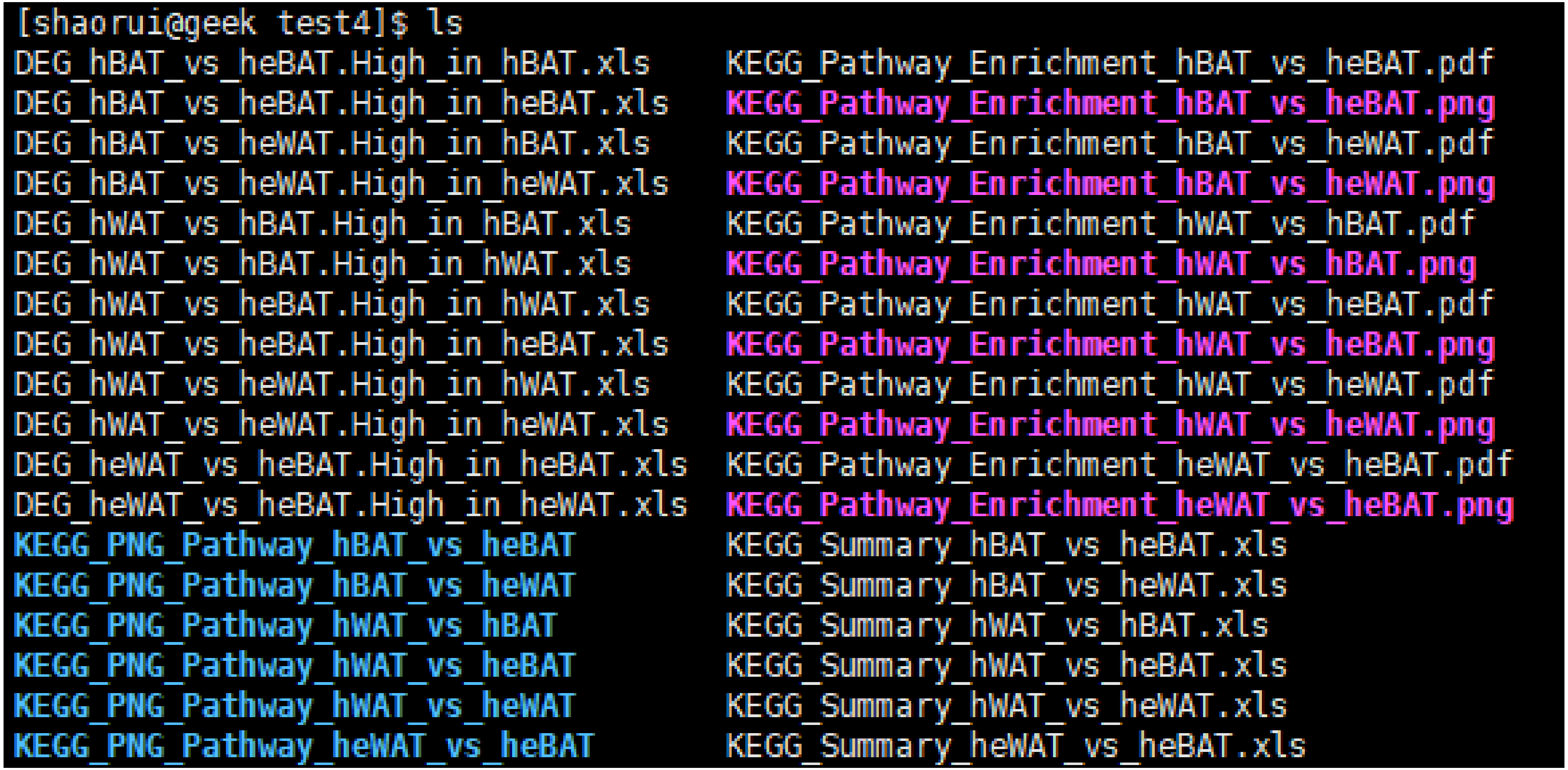
其他没有类似组织的物种，ID转换没有那么容易
如果不想ID转换时出现1对多，多对1的情况，需要作数据清洗
每个物种的情况可能都不一样，相当麻烦，短期内我无法解决

利用KEGG的层级文件，可以很快解析出各个物种的Symbol-ID对应表
这个用来做KEGG分析是完全够用的，但不能用作ID Convertion
（某个物种有KEGG注释的基因可能不到所有基因的四分之一）

$KEGG_Convert.sh -O nnn #可以生成转换对应表 ID_Convert_nnn
(脚本keggConvert.sh 在/home/shaorui/Install/KEGG)



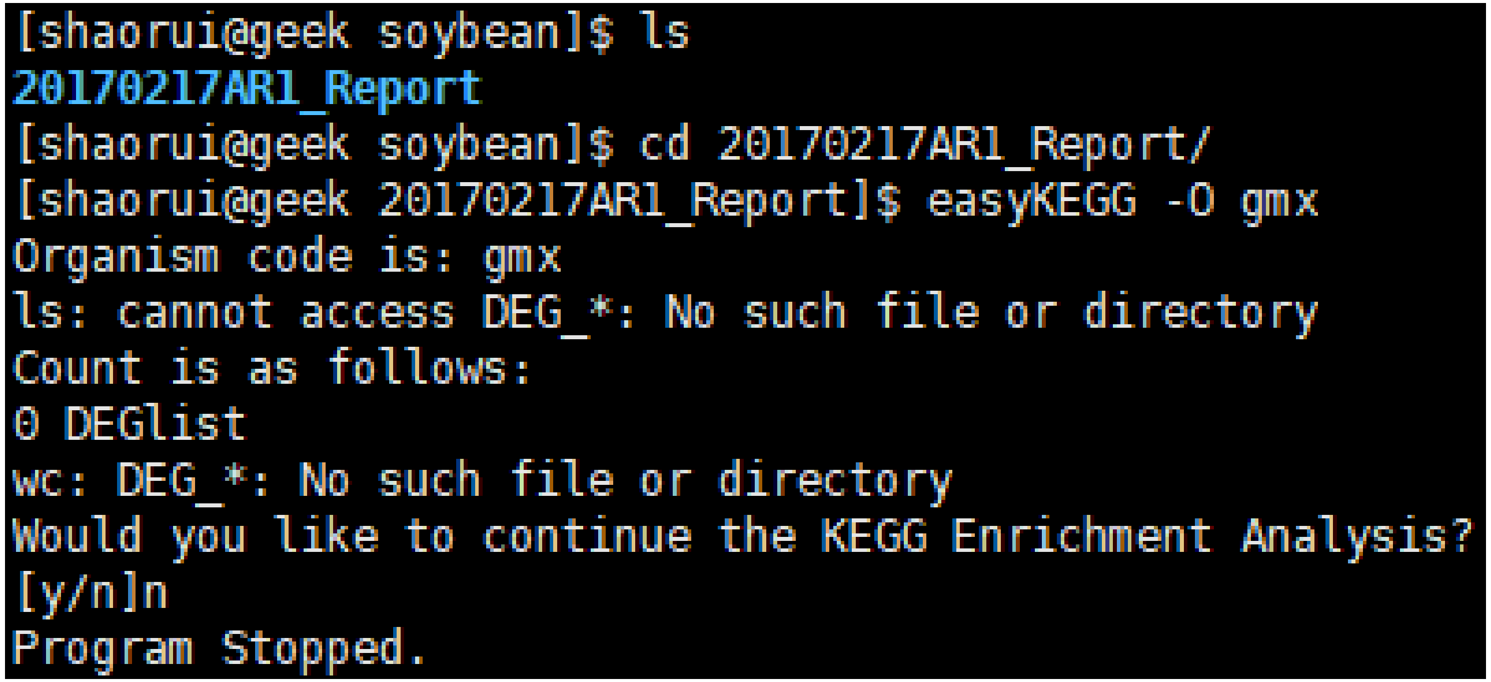
Output



```
[shaorui@geek test4]$ ls
DEG_hBAT_vs_heBAT.High_in_hBAT.xls    KEGG_Pathway_Enrichment_hBAT_vs_heBAT.pdf
DEG_hBAT_vs_heBAT.High_in_heBAT.xls   KEGG_Pathway_Enrichment_hBAT_vs_heBAT.png
DEG_hBAT_vs_heWAT.High_in_hBAT.xls    KEGG_Pathway_Enrichment_hBAT_vs_heWAT.pdf
DEG_hBAT_vs_heWAT.High_in_heWAT.xls   KEGG_Pathway_Enrichment_hBAT_vs_heWAT.png
DEG_hWAT_vs_hBAT.High_in_hBAT.xls     KEGG_Pathway_Enrichment_hWAT_vs_hBAT.pdf
DEG_hWAT_vs_hBAT.High_in_hWAT.xls     KEGG_Pathway_Enrichment_hWAT_vs_hBAT.png
DEG_hWAT_vs_heBAT.High_in_hWAT.xls    KEGG_Pathway_Enrichment_hWAT_vs_heBAT.pdf
DEG_hWAT_vs_heBAT.High_in_heBAT.xls   KEGG_Pathway_Enrichment_hWAT_vs_heBAT.png
DEG_hWAT_vs_heWAT.High_in_hWAT.xls    KEGG_Pathway_Enrichment_hWAT_vs_heWAT.pdf
DEG_hWAT_vs_heWAT.High_in_heWAT.xls   KEGG_Pathway_Enrichment_hWAT_vs_heWAT.png
DEG_heWAT_vs_heBAT.High_in_heBAT.xls  KEGG_Pathway_Enrichment_heWAT_vs_heBAT.pdf
DEG_heWAT_vs_heBAT.High_in_heWAT.xls  KEGG_Pathway_Enrichment_heWAT_vs_heBAT.png
KEGG_PNG_Pathway_hBAT_vs_heBAT        KEGG_Summary_hBAT_vs_heBAT.xls
KEGG_PNG_Pathway_hBAT_vs_heWAT        KEGG_Summary_hBAT_vs_heWAT.xls
KEGG_PNG_Pathway_hWAT_vs_hBAT         KEGG_Summary_hWAT_vs_hBAT.xls
KEGG_PNG_Pathway_hWAT_vs_heBAT        KEGG_Summary_hWAT_vs_heBAT.xls
KEGG_PNG_Pathway_hWAT_vs_heWAT        KEGG_Summary_hWAT_vs_heWAT.xls
KEGG_PNG_Pathway_heWAT_vs_heBAT       KEGG_Summary_heWAT_vs_heBAT.xls
```

问题节选

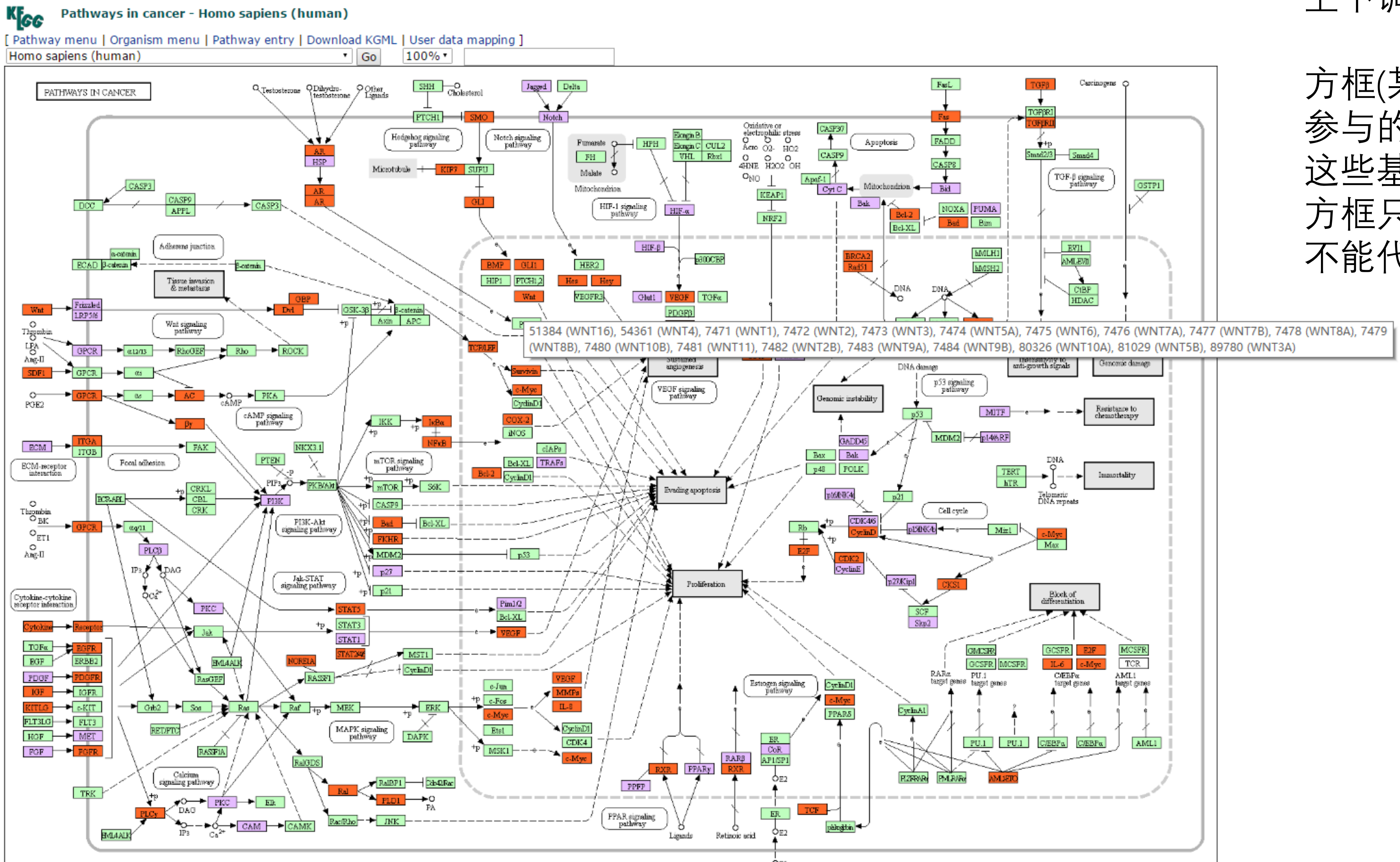
本来想试一个植物的，找了一个去年2月大豆的报告



```
[shaorui@geek soybean]$ ls
20170217AR1_Report
[shaorui@geek soybean]$ cd 20170217AR1_Report/
[shaorui@geek 20170217AR1_Report]$ easyKEGG -O gmx
Organism code is: gmx
ls: cannot access DEG_*: No such file or directory
Count is as follows:
0 DEGlist
wc: DEG_*: No such file or directory
Would you like to continue the KEGG Enrichment Analysis?
[y/n]n
Program Stopped.
```

如果没有差异表达基因文件DEG_*，目前的脚本无法进行分析

问题节选



Pathways in cancer - Homo sapiens (human)

[Pathway menu | Organism menu | Pathway entry | Download KGML | User data mapping]

Homo sapiens (human) Go 100%

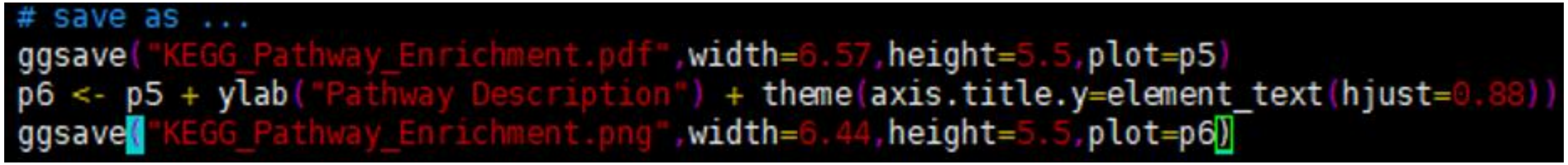
51384 (WNT16), 54361 (WNT4), 7471 (WNT1), 7472 (WNT2), 7473 (WNT3), 7474 (WNT5A), 7475 (WNT6), 7476 (WNT7A), 7477 (WNT7B), 7478 (WNT8A), 7479 (WNT8B), 7480 (WNT10B), 7481 (WNT11), 7482 (WNT2B), 7483 (WNT9A), 7484 (WNT9B), 80326 (WNT10A), 81029 (WNT5B), 89780 (WNT3A)

上下调显示

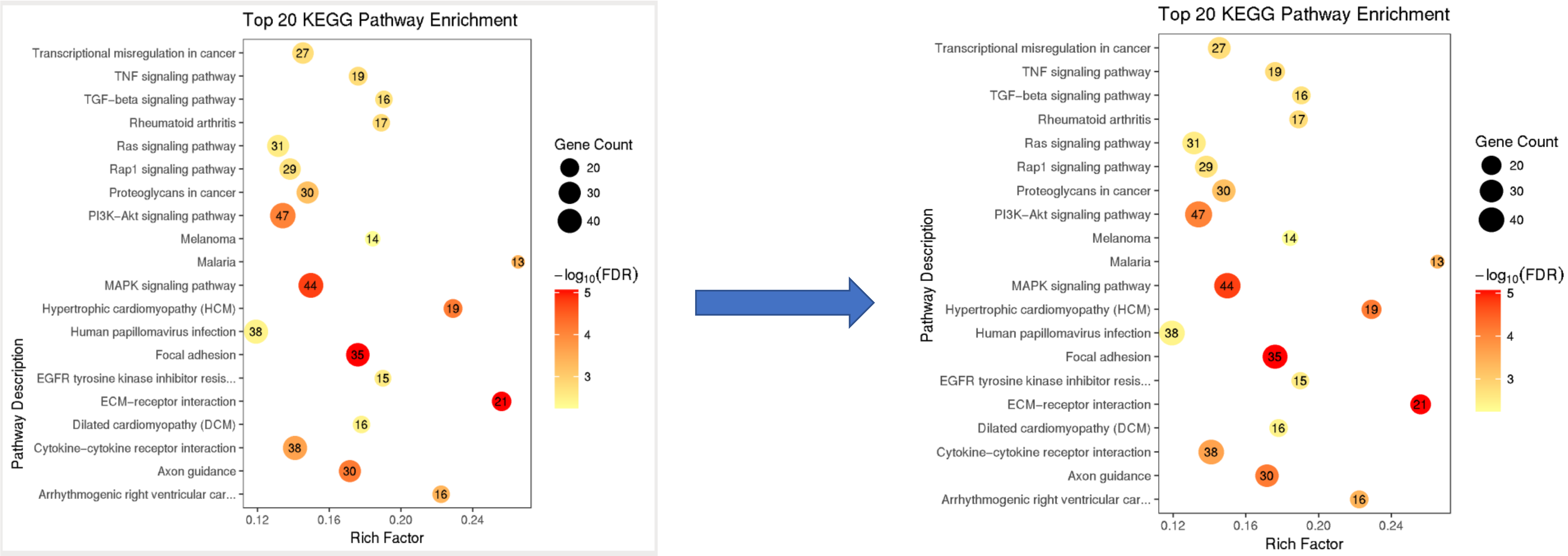
方框(某种蛋白)只有一种颜色
参与的基因可能有很多
这些基因如果既有上调又有下调
方框只显示第一个设置的颜色
不能代表整体的趋势

问题节选

ggplot2 画气泡图，使用pdf格式预览本来已经把图调好了
保存为png，出现明显位移，纵坐标标签位置也变尴尬，保存前重新做了调整



```
# save as ...
ggsave("KEGG_Pathway_Enrichment.pdf",width=6.57,height=5.5,plot=p5)
p6 <- p5 + ylab("Pathway Description") + theme(axis.title.y=element_text(hjust=0.88))
ggsave("KEGG_Pathway_Enrichment.png",width=6.44,height=5.5,plot=p6)
```



服务器环境设置

不要在系统上花费过多时间，专注 *重要问题！！！*

模块化

不要去深究上一个模块存在的bug，发现是好的，但不要去考虑怎么解决，这是上一个模块的事情
在假设它正确的前提下进行开发

先把一件事情做好，再进行下一个阶段，不然什么也做不好

一个好的程序，开发永远不会结束，更多的bug有待发现，更多的需求有待发掘

THANKS!