AnnotationHub 获取kegg org数据库：除了公开的19个之外，其他的也都可以获取、下载

笔记本：	R		
创建时间：	2020/7/14 16:46	**更新时间：**	2020/7/15 11:09
作者：	干冰		
URL：	https://www.bioinfo-scrounger.com/archives/512/		

(1) 加载R包、创建链接

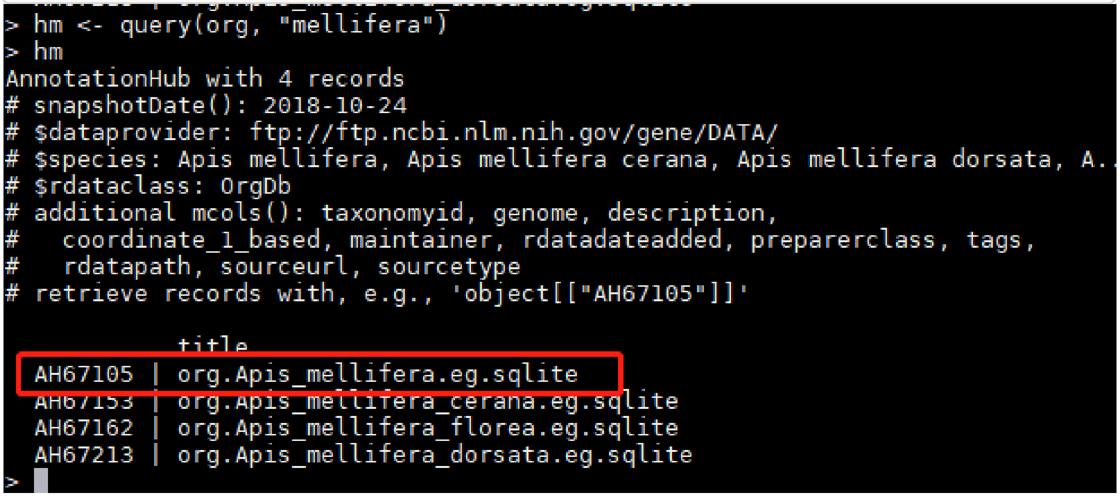
```
library(AnnotationHub)
library(AnnotationDbi)
ah <- AnnotationHub()
```

(2) 搜索org数据库

```
# 获取所有orgdb
org <- ah[ah$rdataclass == "OrgDb",]

# 搜索 物种
hm <- query(org, "Homo sapiens") # 人
hm <- query(org, "mellifera")  # 蜜蜂
hm  # 查看搜索结果
# 结果见下图，得到了很多个，第一个就是目标

# query的完整写法
#  query(x, pattern , ignore.case=TRUE)  pattern 是正则匹配
# 这里我找一个特殊的物种蜜蜂做示例 https://www.ncbi.nlm.nih.gov/genome/?
term=txid7460[orgn] Apis mellifera (honey bee)
```

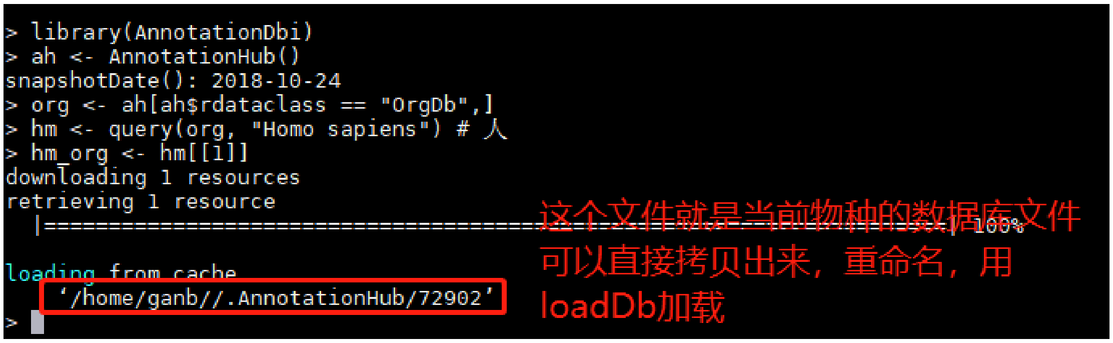


```
> hm <- query(org, "mellifera")
> hm
AnnotationHub with 4 records
# snapshotDate(): 2018-10-24
# $dataprovider: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
# $species: Apis mellifera, Apis mellifera cerana, Apis mellifera dorsata, A..
# $rdataclass: OrgDb
# additional mcols(): taxonomyid, genome, description,
#   coordinate_1_based, maintainer, rdatadateadded, preparerclass, tags,
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH67105"]]'

            title
  AH67105 | org.Apis_mellifera.eg.sqlite
  AH67153 | org.Apis_mellifera_cerana.eg.sqlite
  AH67162 | org.Apis_mellifera_florea.eg.sqlite
  AH67213 | org.Apis_mellifera_dorsata.eg.sqlite
>
```

(3) 下载数据库

```
org_db <- ah[["AH67105"]]  # 这一步会使用网络下载数据，缓存文件通常保存在个人
~/.AnnotationHub/id 文件中。见下方截图示例
```



```
> library(AnnotationDbi)
> ah <- AnnotationHub()
snapshotDate(): 2018-10-24
> org <- ah[ah$rdataclass == "OrgDb",]
> hm <- query(org, "Homo sapiens") # 人
> hm_org <- hm[[1]]
downloading 1 resources
retrieving 1 resource
  |==================================================================| 100%

loading from cache
    '/home/ganb//.AnnotationHub/72902'
> 
```

这个文件就是当前物种的数据库文件
可以直接拷贝出来，重命名，用
loadDb加载

(4) 数据库保存、加载（saveDb不能用了，报错，原因未知，此时请采用4.1备选方案）

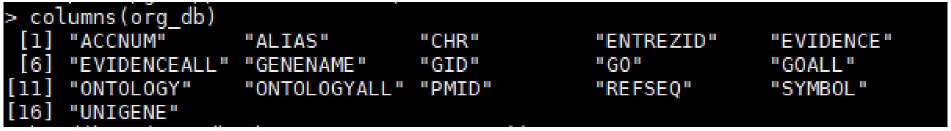
```
# 保存到文件，下次直接加载即可
saveDb(org_db, file = "mellifera.orgdb")

# 加载
org_db = loadDb(file = "mellifera.orgdb")
```

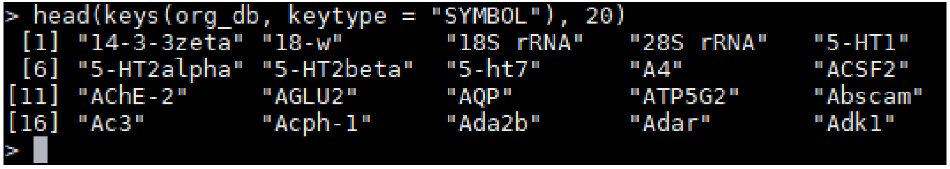
(4.1) 备选数据保存方案

```
cp ~/.AnnotationHub/id abc.orgdb  # 直接拷贝数据库缓存文件
org_db = loadDb(file = "abc.orgdb")
```

数据库相关操作

- columns(org_db)
  - 查看数据库包含哪些信息
  - 

```
> columns(org_db)
 [1] "ACCNUM"       "ALIAS"        "CHR"          "ENTREZID"     "EVIDENCE"
 [6] "EVIDENCEALL"  "GENENAME"     "GID"          "GO"           "GOALL"
[11] "ONTOLOGY"     "ONTOLOGYALL"  "PMID"         "REFSEQ"       "SYMBOL"
[16] "UNIGENE"
```

- head(keys(org_db, keytype = "SYMBOL"))
  - 获取所有SYMBOL信息
  - 

```
> head(keys(org_db, keytype = "SYMBOL"), 20)
 [1] "14-3-3zeta" "18-w"       "18S rRNA"   "28S rRNA"   "5-HT1"
 [6] "5-HT2alpha" "5-HT2beta"  "5-ht7"      "A4"         "ACSF2"
[11] "AChE-2"     "AGLU2"      "AQP"        "ATP5G2"     "Abscam"
[16] "Ac3"        "Acph-1"     "Ada2b"      "Adar"       "Adk1"
> 
```

注意：要完成GO/KEGG 分析，org数据库要包含 SYMBOL/ENTREZID/GO 这三个信息
SYMBOL/ENTREZID

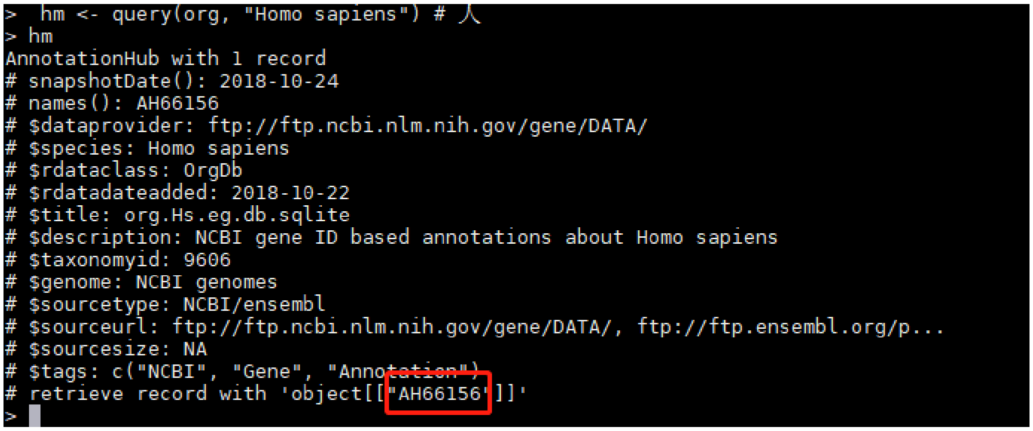
- **这两个信息主要用来把输入基因SYMBOL转化为ENTREZID(ENTREZID编号唯一， 且GO/KEGG富集分析用的都是用这个编号，而不是SYMBOL)**
- **SYMBOL严格区分大小写，一定要保证与NCBI一致**

GO:

- **GO富集分析要使用**

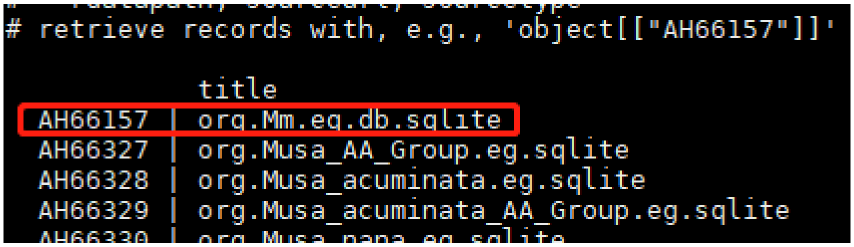
常见的几个数据库

人

- 

```
>  hm <- query(org, "Homo sapiens") # 人
> hm
AnnotationHub with 1 record
# snapshotDate(): 2018-10-24
# names(): AH66156
# $dataprovider: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
# $species: Homo sapiens
# $rdataclass: OrgDb
# $rdatadateadded: 2018-10-22
# $title: org.Hs.eg.db.sqlite
# $description: NCBI gene ID based annotations about Homo sapiens
# $taxonomyid: 9606
# $genome: NCBI genomes
# $sourcetype: NCBI/ensembl
# $sourceurl: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, ftp://ftp.ensembl.org/p...
# $sourcesize: NA
# $tags: c("NCBI", "Gene", "Annotation")
# retrieve record with 'object[["AH66156"]]'
> 
```

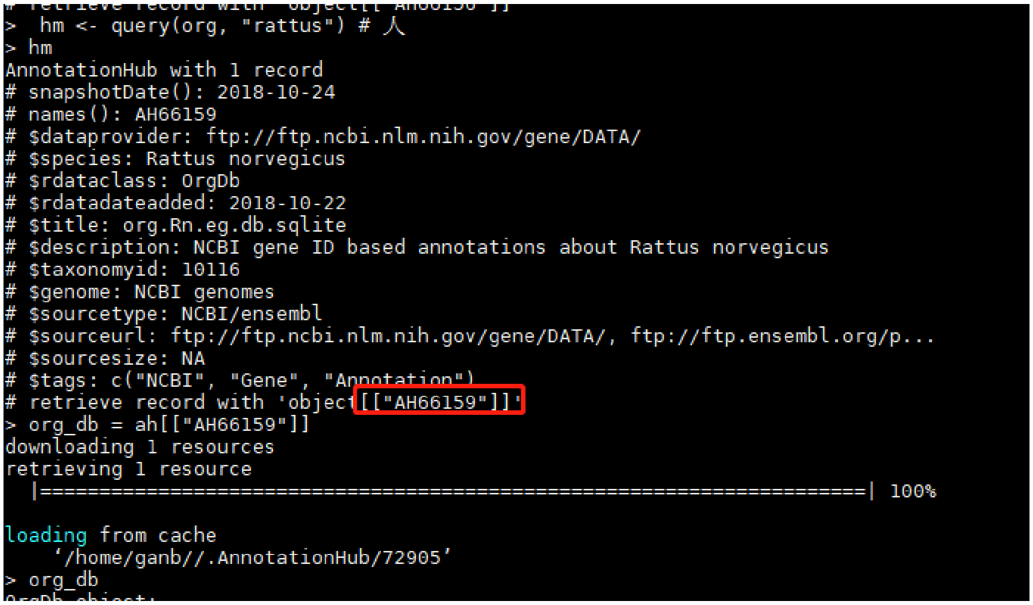
小鼠

- 

```
#   rdatapath, sourceurl, sourcetype
# retrieve records with, e.g., 'object[["AH66157"]]'

             title
  AH66157 | org.Mm.eg.db.sqlite
  AH66327 | org.Musa_AA_Group.eg.sqlite
  AH66328 | org.Musa_acuminata.eg.sqlite
  AH66329 | org.Musa_acuminata_AA_Group.eg.sqlite
  AH66330 | org.Musa_nana.eg.sqlite
```

大鼠

- 

```
# retrieve record with 'object[["AH66156"]]'
>  hm <- query(org, "rattus") # 人
> hm
AnnotationHub with 1 record
# snapshotDate(): 2018-10-24
# names(): AH66159
# $dataprovider: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
# $species: Rattus norvegicus
# $rdataclass: OrgDb
# $rdatadateadded: 2018-10-22
# $title: org.Rn.eg.db.sqlite
# $description: NCBI gene ID based annotations about Rattus norvegicus
# $taxonomyid: 10116
# $genome: NCBI genomes
# $sourcetype: NCBI/ensembl
# $sourceurl: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/, ftp://ftp.ensembl.org/p...
# $sourcesize: NA
# $tags: c("NCBI", "Gene", "Annotation")
# retrieve record with 'object[["AH66159"]]'
> org_db = ah[["AH66159"]]
downloading 1 resources
retrieving 1 resource
  |======================================================================| 100%

loading from cache
    '/home/ganb//.AnnotationHub/72905'
> org_db
OrgDb object:
```

end