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 甘斌

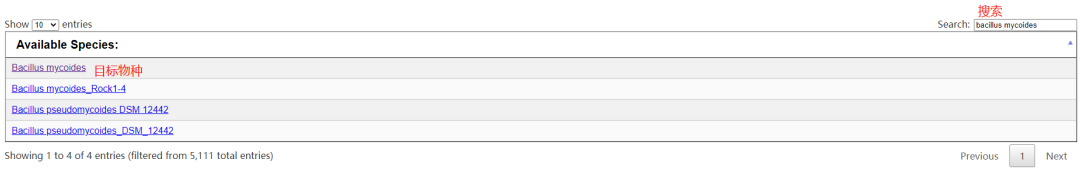
笔记本：	R		
创建时间：	2020/7/14 16:46	**更新时间：**	2020/12/29 13:51
作者：	干冰		
URL：	https://www.bioinfo-scrounger.com/archives/512/		

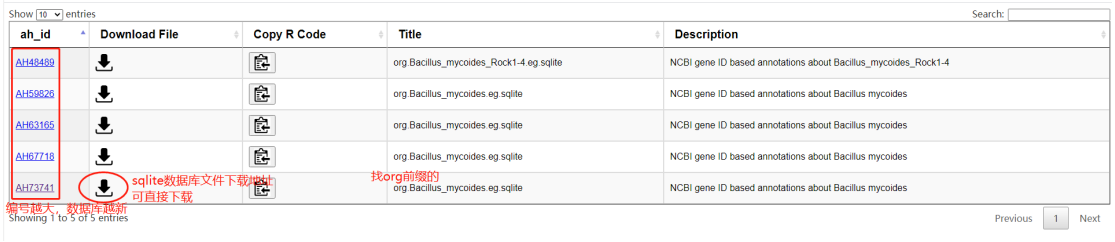
GO KEGG 富集分析数据库准备

注意：

- 不同的R版本下载到的数据库文件是不同的（创建日期不同）
  - 查看历史版本记录 possibleDates(ah)
- 最新版的库不一定齐全，如果新版找不到，可以用旧版找找，但是终究是**不方便**
  - 例如：bacillus mycoides 以及candida tropicalis， 这两个菌在R4.0.3里找不到，但是3.5.1中有

推荐搜索、下载方式：
https://annotationhub.bioconductor.org/species

- 

- 

```
{
  "id": 73741,
  "ah_id": "AH73741",
  "title": "org.Bacillus_mycoides.eg.sqlite",
  "dataprovider": "ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/",
  "species": "Bacillus mycoides",
  "taxonomyid": 1405,
  "genome": "NCBI genomes",
  "description": "NCBI gene ID based annotations about Bacillus mycoides",
  "coordinate_1_based": true,
  "maintainer": "Bioconductor Package Maintainer <maintainer@bioconductor.org>",
  "rdatadateadded": "2019-05-02",
  "rdatadateremoved": null,
  "preparerclass": "NCBIImportPreparer",
  "location_prefix": "http://s3.amazonaws.com/annotationhub/",
  "recipe": "AnnotationHubData::NCBIToOrgDbs",
  "recipe_package": "AnnotationHubData",
  "rdatapaths": [
    {
      "rdatapath": "ncbi/uniprot/3.9/org.Bacillus_mycoides.eg.sqlite",
      "rdataclass": "OrgDb",
      "dispatchclass": "SQLiteFile"
    }
  ],
  "input_sources": [
    {
      "sourcesize": null,
      "sourceurl": "ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/",
      "sourceversion": "NCBI gene annotations as of Mon May 20 12:42:58 2019",
      "sourcemd5": null,
      "sourcelastmodifieddate": null,
      "sourcetype": "NCBI/UniProt"
    },
    {
      "sourcesize": null,
      "sourceurl": "ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz",
      "sourceversion": "NCBI gene annotations as of Mon May 20 12:42:58 2019",
      "sourcemd5": null,
      "sourcelastmodifieddate": null,
      "sourcetype": "NCBI/UniProt"
    }
  ],
  "tags": [
    "NCBI",
    "Gene",
    "Annotation"
  ],
  "biocversions": [
    "3.9"
  ]
}
```


点击AH****编号，查看该数据库的详情

(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)
```

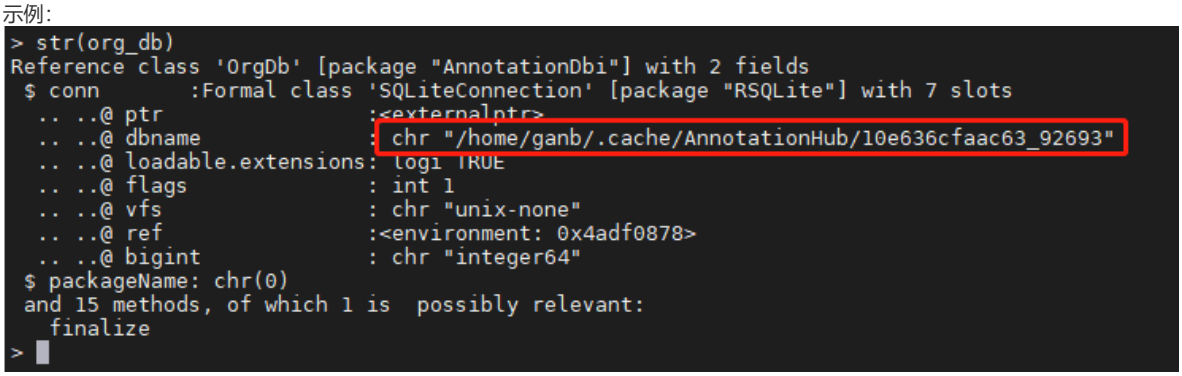
```
> hm <- query(org, "mellifera")
> hm
AnnotationHub with 4 records
# snapshotDate(): 2018-10-24
# $dataprovider: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/
# $species: Apis mellifera, Apis mellifera cerana, Apis mellifera dorsata, A..
# $rdataclass: OrgDb
# additional mcols(): taxonomyid, genome, description,
#   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
```

(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）

查看缓存目录：getAnnotationHubOption("CACHE")
通常在这里：个人家目录 ~/.cache/AnnotationHub
查看当前数据库缓存文件地址： str(org_db)
示例：




```
# 保存到文件，下次直接加载即可

saveDb(org_db, file = "mellifera.orgdb")
# 有时候会报错，分时间，多试几次

# 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")

文件名建议统一命名为： 物种名称.kegg简称.orgdb
escherichia_coli.eco.orgdb 大肠杆菌
honey_bee_apis_mellifera.ame.orgdb 蜜蜂
human.hsa.orgdb 人
mouse.mmu.orgdb 小鼠
rabbit_oryctolagus_cuniculus.ocu.orgdb 兔子
rat.rno.orgdb 大鼠
zea_mays.zma.orgdb 玉米
zebrafish_danio_rerio.dre.orgdb 斑马鱼
```

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)**
  - **注意：有的物种不支持ENTREZID转换，下面做详细说明**
- **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**
- **查看kegg对应物种的ncbi-geneid转换api能否打开，例如：http://rest.kegg.jp/conv/hsa/ncbi-geneid**
  - 能打开，一切正常
  - 不能打开，则需要换一种映射方式了，请仔细阅读后面的资料

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

```
# 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
```

大鼠

```
> 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
```

关于KEGG富集分析支持的输入ID说明

kegg支持的物种列表： https://www.genome.jp/kegg/catalog/org_list.html
kegg物种基因id使用： https://www.genome.jp/kegg/genes.html

Data Source of KEGG GENES

The following table shows the data source of the KEGG GENES database.

Category	Original DB[1]	Content[2]	Genome identifier	Gene identifier
Eukaryotes	RefSeq	RefSeq release (complete)	T0 numbers (three or four letter organism codes)	GeneID
Prokaryotes	RefSeq	NCBI reference genomes		Locus_tag
	GenBank	Other complete genomes		Locus_tag
Viruses	RefSeq	Refseq release (viral)	T40000 (vg) T4 numbers	GeneID
Addendum	KEGG	Functionally characterized proteins	T10000 (ag)	ProteinID

[1] Original DB name is shown in the definition field of each GENES entry.

[2] RefSeq bimonthly releases are used to update eukaryotes and viruses.
Prokaryoteic genomes are selected from ftp://ftp.ncbi.nlm.nih.gov/genomes/ASSEMBLY_REPORTS/.

（1）通常情况下，富集分析时，我们输入的都是基因symbol
（2）GO/KEGG 基本不支持symbol的识别，因为symbol存在别名问题

- 那么就得换一个id作为唯一识别码

（3）我们需要org_db数据库，把symbol映射为GO/KEGG可识别的标签，**通常情况下**，就是**ENTREZID**

- 查看物种在kegg中使用的基因名称（可以看到，都是一个数字编号，但是也有特殊情况，例如下面的bmyo）

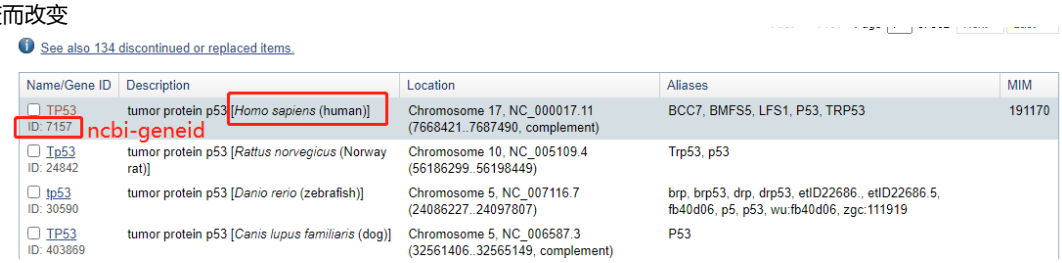
- 人 http://rest.kegg.jp/list/hsa

```
hsa:65249       ZSWIM4; zinc finger SWIM-type containing 4
hsa:100462981   MTRNR2L2, HN2; MT-RNR2 like 2
hsa:100526771   SMIM35, TMPRSS4-AS1; small integral membrane protein 35
hsa:3441        IFNA4, IFN-alpha4a, INFA4; interferon alpha 4
hsa:317754      POTED, A26B3, ANKRD21, CT104.1, POTE, POTE-21, POTE21; POTE ankyrin domain family member D
hsa:149685      ADIG, SMAF1; adipogenin
hsa:134864      TAAR1, TA1, TAR1, TRAR1; trace amine associated receptor 1
hsa:9720        CCDC144A; coiled-coil domain containing 144A
hsa:3442        IFNA5, IFN-alpha-5, IFN-alphaG, INA5, INFA5, leIF_G; interferon alpha 5
hsa:84210       ANKRD20A1, ANKRD20A; ankyrin repeat domain 20 family member A1
hsa:101928917   HSFX3; heat shock transcription factor family, X-linked member 3
hsa:139741      ACTRT1, AIP1, ARIP1, ARPT1, HSD27; actin related protein T1
hsa:100302174   MIR1307, MIRN1307, hsa-mir-1307, mir-1307; microRNA 1307
```

- 小鼠 http://rest.kegg.jp/list/mmu
- 大鼠 http://rest.kegg.jp/list/rno
- 蕈状芽孢杆菌 http://rest.kegg.jp/list/bmyo

```
bmyo:BG05_1     glutamine amidotransferases class-II family protein
bmyo:BG05_2     aluminum activated malate transporter family protein
bmyo:BG05_3     ABC transporter family protein
bmyo:BG05_4     hypothetical protein
bmyo:BG05_5     ygaB-like family protein
bmyo:BG05_6     small, acid-soluble spore, gamma-type family protein
bmyo:BG05_7     putative yfhS
bmyo:BG05_8     mutY; A/G-specific adenine glycosylase
bmyo:BG05_9     hypothetical protein
bmyo:BG05_10    WVELL family protein
bmyo:BG05_11    sspK; small, acid-soluble spore protein K
bmyo:BG05_12    ypzG-like family protein
bmyo:BG05_13    hypothetical protein
bmyo:BG05_14    recX family protein
bmyo:BG05_15    NAD dependent epimerase/dehydratase family protein
bmyo:BG05_16    amidohydrolase family protein
bmyo:BG05_17    acetyltransferase domain protein
bmyo:BG05_18    yfhE-like family protein
bmyo:BG05_19    ugtP; processive diacylglycerol glucosyltransferase
bmyo:BG05_20    pflA; pyruvate formate-lyase 1-activating enzyme
bmyo:BG05_21    pflB; formate acetyltransferase
bmyo:BG05_22    hypothetical protein
bmyo:BG05_23    glycosyltransferase like 2 family protein
bmyo:BG05_24    3-beta hydroxysteroid dehydrogenase/isomerase family protein
bmyo:BG05_25    nucleotide sugar dehydrogenase family protein
bmyo:BG05_26    glycosyltransferase like 2 family protein
```

- 通常情况下，KEGG使用的唯一基因标签是ENTREZID，这个ID就是ncbi-geneid，这个ID是唯一的，不会因为别名改变而改变



- 使用 http://rest.kegg.jp/conv/hsa/ncbi-geneid 可以查询物种在kegg中，keggid与ncbi-geneid的映射关系，可以发现，数字编号id是完全一样的，那么，我们就可以直接使用ENTREZID做富集分了


```
ncbi-geneid:1    hsa:1
ncbi-geneid:10   hsa:10
ncbi-geneid:100  hsa:100
ncbi-geneid:1000         hsa:1000
ncbi-geneid:10000        hsa:10000
ncbi-geneid:100008586    hsa:100008586
ncbi-geneid:100008587    hsa:100008587
ncbi-geneid:100008588    hsa:100008588
ncbi-geneid:100008589    hsa:100008589
ncbi-geneid:10001        hsa:10001
ncbi-geneid:10002        hsa:10002
ncbi-geneid:10003        hsa:10003
ncbi-geneid:100033411    hsa:100033411
ncbi-geneid:100033413    hsa:100033413
ncbi-geneid:100033414    hsa:100033414
ncbi-geneid:100033415    hsa:100033415
ncbi-geneid:100033416    hsa:100033416
ncbi-geneid:100033417    hsa:100033417
```

- 但是，有的物种，在kegg中是不存在ncbi-geneid映射关系的，例如前面说的细菌 bmyo, http://rest.kegg.jp/conv/bmyo/ncbi-geneid 你会发现无法访问。

- 面对非ncbi-geneid的情况，你就不能用ENTREZID了，需要换一种ID。这就需要你自己找一下org_db，哪一个库中是keggid
  - 这里以bmyo为例
    1. 查看包含的数据库 columns(org_db)

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

    2. 查看数据库的中内容 head(keys(org_db, keytype = "ALIAS"))

```
> head(keys(org_db, keytype = "ALIAS"))
[1] "BG05_10"   "BG05_100"  "BG05_1000" "BG05_1001" "BG05_1002" "BG05_1003"
>
```

    3. 发现，上面的ALIAS就是KEGGID，那么，富集分析时，上面的id就是转换目标

KEGG目前的总结：

1. 查看物种在kegg里的kegg id http://rest.kegg.jp/list/hsa
2. 查看物种在kegg中是否存在ncbi-geneid 映射关系 http://rest.kegg.jp/conv/hsa/ncbi-geneid
   1. 如果有，那么基本上id映射方式就是ENTREZID
   2. 如果没有，那么，就需要打开org_db， 查一下哪个数据库能与 list/物种 列表里的keggid对应上
      1. 目前来看，通常是ALIAS数据库对应kegg编号
3. 注意：并不是所有kegg列出来的物种，都有org数据库，只是某一个菌群存在数据库，例如 Rahnella只有 raa有orgdb

Chania	sfo	Chania multitudinisentens	2014	GenBank
Rahnella	rah	Rahnella sp. Y9602	2011	GenBank
	raq	Rahnella aquatilis CIP 78.65 = ATCC 33071	2012	GenBank
	raa	Rahnella aquatilis HX2	2012	GenBank
	rox	Rahnella sp. ERMR1:05	2018	GenBank

   1. 所以，在创建物种数据库前，先看一下kegg支持哪一个菌群

工具：

kegg物种列表	https://www.genome.jp/kegg/catalog/org_list.html	
查看物种所有的通路	http://rest.kegg.jp/list/pathway/eco	
查看物种的所有kegg gene id-> symbol映射关系	http://rest.kegg.jp/list/eco	
查看物种的ncbi-geneid -> kegg gene id映射关系	http://rest.kegg.jp/conv/eco/ncbi-geneid	
下载通路图片	http://rest.kegg.jp/get/eco00010/image	
下载通路conf文件	http://rest.kegg.jp/get/eco00010/conf	
查看交互图	http://www.kegg.jp/kegg-bin/show_pathway?eco00010	

annotationhub 数据库下载地址	https://annotationhub.bioconductor.org/species	**library(AnnotationHub)** **library(AnnotationDbi)** **org_db = loadDb(file = "mellifera.orgdb") # 加载** columns(org_db) # 查看包含哪些列 head(keys(org_db, keytype = "SYMBOL")) # 查看某一列

代表物种	**kegg id**	**输入**	**orgdb转换**	**clusterprofile任务**	**富集分析返回geneid**	**http颜色标记**	注意
human	hsa	symbol	ncbi-geneid	转换为kegg id	**orgdb转换**	**ncbi-geneid/symbol都可以**	这里 ncbi-geneid 与 kegg id一样
Escherichia coli	eco	symbol	ncbi-geneid	转换为kegg id	**orgdb转换**	**kegg id/symbol 都可以**	
Bacillus mycoides ATCC 6462	bmyo	symbol	ALIAS/kegg id	无	**ALIAS/kegg id**	**kegg id/ 特殊的symbol，至少目前与ncbi下载的gtf无法对应**	

end