GEO 数据快速查询与获取

2025-01-02

LiChuang Huang

1 安装依赖

对于该文档所述的功能，只需要两个工具，EDirect，以及我的 R 包 utils.tool (从 github 获取)。utils.tool 需要很多 R 包依赖，例如这里主要的 GEOquery R 包。还可能存在一些你没有安装过的 R 包。

对于该功能，我已经在服务器 (账号:

```
HostName ssh.cn-zhongwei-1.paracloud.com
User t0s000324@BSCC-T
```

) 中部署完毕 (conda: r4-base)，可直接使用，无需再安装了。

R input

```r
## 这是一个 Linux 命令行工具
system('sh -c "$(wget -q https://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/install-edirect.sh -O -)"')
## 获取 github 中的 R 包。
system("git clone --depth 1 https://github.com/shaman-yellow/utils.tool ~/utils.tool")
## 由于该包很多方法没有导出，所以无法通过 install 来使用。请使用以下方式加载:
devtools::load_all("~/utils.tool")
```

2 基本方法

2.1 加载包

R input

```r
## devtools::load_all, 相当于 `library` 这个命令
devtools::load_all("~/utils.tool")
## 设置缓存路径
set_prefix("~/cache")
```

2.2 主要方法

以下所有，除了 job_gds，都是 S4 泛型方法。

R input

```
## 以 EDirect 查询 GEO 信息，整理成数据框导入
job_gds
## 预设的一些过滤条件
step1
## 预设的一些过滤方法
step2
## 一种交互式操作，快速格式化元数据的列，整理出 sample, group 列
expect
## 针对 group 列，形成 markdown 格式文本
anno
```

2.3 查看方法本体

S4 在查看函数本体上不如普通的 function 方便，但你可以用以下方法查看：

R input

```
selectMethod(step1, "job_gds")
selectMethod(step2, "job_gds")
selectMethod(expect, "job_gds")
selectMethod(anno, "job_gds")
```

或者，也可以直接查看 R 包中的源代码。

R input

```
readLines("~/utils.tool/R/workflow_88_gds.R")
```

3 使用示例

3.1 最初

先加载这个包。

R input

```
devtools::load_all("~/utils.tool")
set_prefix("~/cache")
```

3.2 查询睡眠呼吸暂停症

以“睡眠呼吸暂停症”为例子，查询 GEO 可用数据。

R input

```r
## org 参数可以指定物种，例如 Humo Sapiens，这里不指定
## 这里还有一些默认参数，例如 n, 指定样本数量，默认是 6:1000
## 第一个参数，即 c("Sleep apnea")，可以指定多个关键词，例如 c("Sleep apnea", "Healthy")
gds.sa <- job_gds(c("Sleep apnea"), org = NULL)
```

注意，job_gds 会形成本地缓存，下次免于联网搜索，如果你一定要重新搜索，请指定参数：force = TRUE。

3.3 得到的结果（已经得到了数据表）

可以从 gds.sa@object 中查看运行结果。

R input

```r
gds.sa@object
```

```
## # A tibble: 12 x 7
##    GSE       title                    summary taxon gdsType n_samples PubMedIds
##    <chr>     <chr>                    <chr>   <chr> <chr>       <int> <chr>
##  1 GSE242668 Long-term intermittent h~ Obstru~ Mus ~ Expres~        10 PRJNA101~
##  2 GSE235867 Low testosterone and int~ Interm~ Mus ~ Expres~        20 PRJNA987~
##  3 GSE215935 mRNA, lncRNA, and circRN~ Backgr~ Mus ~ Expres~         6 PRJNA891~
##  4 GSE189958 Combined intermittent an~ Study ~ Mus ~ Expres~        16 PRJNA785~
##  5 GSE145435 Short-term exposure to i~ Obstru~ Mus ~ Expres~         6 PRJNA607~
##  6 GSE145434 Short-term exposure to i~ Obstru~ Mus ~ Expres~        12 PRJNA607~
##  7 GSE145221 Expression data from mou~ Athero~ Mus ~ Expres~        23 PRJNA606~
##  8 GSE21409  Chronic Intermittent Hyp~ Obstru~ Mus ~ Expres~        10 PRJNA126~
##  9 GSE14981  Distinct Mechanisms Unde~ Backgr~ Dros~ Expres~         9 PRJNA111~
## 10 GSE2271   Gene expression and phen~ Chroni~ Mus ~ Expres~        18 PRJNA913~
## 11 GSE1873   The effect of intermitte~ All an~ Mus ~ Expres~        10 PRJNA905~
## 12 GSE480    Sleep apnea and glucose ~ This s~ Mus ~ Expres~        20 PRJNA851~
```

3.4 正则匹配过滤（可选的）

R input

```r
## 函数 `grpl` 是 `grepl` 的封装，只是改变了参数顺序。
## clinical 会按照预设的一些条件，过滤掉一些数据，请查看 `selectMethod(step1, "job_gds")`
## 这里筛选了包含 `Intermittent hypoxia` 的数据。
gds.sa <- step1(
  gds.sa, clinical = FALSE, !grpl(taxon, "Homo Sapiens", TRUE),
  grpl(summary, "Intermittent hypoxia", TRUE)
)
```

可以以下方式，跳过这一步。

R input

```
gds.sa@step <- 1L
```

3.5 获取元数据

R input

```
## 会下载数据集，请注意，尽量避免一次性下载过多，所以过滤 `ges.sa@object` 中的数据是必要的
## 需要等待一会儿
ges.sa <- step2(gds.sa)
```

3.6 查看元数据

基本上，该数据集能否用于你的分析，看一下这个结果就能知道了。例如生存分析，你至少要在结果中找到 survival 对应的数据记录。

R input

```
head(gds.sa@params$res$metas, n = 1)
```

```
## $GSE242668
## # A tibble: 10 x 11
##    sample     group  rownames title age.at_the_treatment~1 diet.ch1 genotype.ch1
##    <chr>      <chr>  <chr>    <chr> <chr>                  <chr>    <chr>
##  1 GSM7766596 inter~ GSM7766~ ih36~ 16 weeks               regular~ C57BL/6JRj ~
##  2 GSM7766597 inter~ GSM7766~ ih37~ 16 weeks               regular~ C57BL/6JRj ~
##  3 GSM7766598 inter~ GSM7766~ ih38~ 16 weeks               regular~ C57BL/6JRj ~
##  4 GSM7766599 inter~ GSM7766~ ih39~ 16 weeks               regular~ C57BL/6JRj ~
##  5 GSM7766600 inter~ GSM7766~ ih40~ 16 weeks               regular~ C57BL/6JRj ~
##  6 GSM7766601 normo~ GSM7766~ no31~ 16 weeks               regular~ C57BL/6JRj ~
##  7 GSM7766602 normo~ GSM7766~ no32~ 16 weeks               regular~ C57BL/6JRj ~
##  8 GSM7766603 normo~ GSM7766~ no33~ 16 weeks               regular~ C57BL/6JRj ~
##  9 GSM7766604 normo~ GSM7766~ no34~ 16 weeks               regular~ C57BL/6JRj ~
## 10 GSM7766605 normo~ GSM7766~ no35~ 16 weeks               regular~ C57BL/6JRj ~
## # i abbreviated name: 1: age.at_the_treatment_onset.ch1
## # i 4 more variables: Sex.ch1 <chr>, tissue.ch1 <chr>,
## #   treatment.duration.ch1 <chr>, treatment.ch1 <chr>
```

3.7 快速格式化分组信息（可选）

这是一个极其方便的工具，查找可能存在的 “group” 列，交互式 (并生成本地记录) 提示，可能让你手动指定。请自行探索。

R input

```r
gds.sa <- expect(gds.sa, geo_cols())
## 结果请查看
gds.sa@params$res$metas
```

3.8 对 Group 列生成总结 (可选)

R input

```r
gds.sa <- anno(gds.sa)
```

可以通过以下方式查看结果

R input

```r
gds.sa@snap
```

或者:

R input

```r
writeLines(snap(gds.sa, "a"))
```

3.9 最终效果展现

上述步骤都运行后，可得到:

R input

```r
writeLines(snap(gds.sa, "a"))
```

- **GSE242668**, **Type**: RNA-seq
  - intermittent_hypoxia (n = 5)
  - normoxic_control (n = 5)
- **GSE235867**, **Type**: RNA-seq
  - ORX-IH (n = 5)
  - ORX-Nx (n = 5)
  - Sham-IH (n = 5)
  - Sham-Nx (n = 5)
- **GSE215935**, **Type**: Microarray; Non-coding RNA-seq
  - chronic intermittent hypoxia (CIH) system combined with Ang II (n = 3)
  - normal saline (n = 3)
- **GSE189958**, **Type**: RNA-seq
  - Intermittent hypoxia (n = 4)
  - Overlap hypoxia (n = 4)

  - Room air (n = 4)
  - Sustained hypoxia (n = 4)
- **GSE145435**, **Type**: (scRNA-seq) RNA-seq
  - Ctrl (n = 3)
  - Hypo (n = 3)
- **GSE145434**, **Type**: RNA-seq
  - CTRL (n = 6)
  - HYPO (n = 6)
- **GSE145221**, **Type**: Microarray
  - CIH for 12 (n = 4)
  - CIH for 8 (n = 5)
  - CIH for 8 weeks followed by normoxia for 4 (n = 4)
  - normoxia for 12 (n = 5)
  - normoxia for 8 (n = 5)
- **GSE21409**, **Type**: Microarray
  - Interm Hypoxia (n = 5)
  - Normoxia (n = 5)
- **GSE14981**, **Type**: Microarray
  - CH (n = 3)
  - IH (n = 3)
  - NC (n = 3)
- **GSE2271**, **Type**: Microarray
  - mouse subjected (n = 2)
  - mouse subjected 1 week to chronic constant hypoxia (n = 1)
  - mouse subjected 1 week to chronic intermittent hypoxia (n = 2)
  - mouse subjected 2 week to chronic constant hypoxia (n = 1)
  - mouse subjected 2 week to chronic intermittent hypoxia (n = 2)
  - mouse subjected 4 week to chronic constant hypoxia (n = 2)
  - mouse subjected 4 week to chronic intermittent hypoxia (n = 2)
  - mouse, 1 week control (n = 2)
  - mouse, 2 week control (n = 2)
  - mouse, 4 week control (n = 2)
- **GSE1873**, **Type**: Microarray
  - Liver, Intermittent Hypoxia (n = 5)
  - Liver, Normoxia (n = 5)
- **GSE480**, **Type**: Microarray
  - PGA-MGM-ConBrain (n = 1)
  - PGA-MGM-ConHeart (n = 1)
  - PGA-MGM-ConHyp (n = 2)
  - PGA-MGM-ConLung (n = 1)
  - PGA-MGM-ConMuscle (n = 1)
  - PGA-MGM-ConNonHyp (n = 2)

- PGA-MGM-FragBrain (n = 1)
- PGA-MGM-FragHeart (n = 1)
- PGA-MGM-FragLung (n = 1)
- PGA-MGM-FragMuscle (n = 1)
- PGA-MGM-GlucoseHyp (n = 2)
- PGA-MGM-GlucoseNonHyp (n = 2)
- PGA-MGM-HypoxiaBrain (n = 1)
- PGA-MGM-HypoxiaHeart (n = 1)
- PGA-MGM-HypoxiaLung (n = 1)
- PGA-MGM-HypoxiaMuscle (n = 1)

4 补充说明

4.1 关于 job_gds

`gds.sa <- job_gds(c("Sleep apnea"), org = NULL)` 的运行效果，相当于以下：

R input

```
system("esearch -db gds -query '(Sleep apnea[Description]) AND ((6:1000[Number of Samples]) AND (GSE
```

请参考 https://www.ncbi.nlm.nih.gov/books/NBK3837/ 官方文档说明。

4.2 关于 step2 获取数据

是以下函数的封装：

R input

```
GEOquery::getGEO
```