GEO-TCGA-数据下载

阎俊安 2022-01-08

目录

本文档主要来介绍如何使用 **R** 代码与 **Python** 脚本轻松下载数据, 下面来看具体代码

1 安装并加载 R 包

```
package.list=c("tidyverse","GEOquery")

for (package in package.list) {
  if (!require(package,character.only=T, quietly=T)) {
    install.packages(package)
    library(package, character.only=T)
  }
}
```

2 GEO 数据下载

2.1 下载 GSE33126 数据

通过 GEOquery 包可以轻松下载 GEO 数据

```
gset <- getGEO("GSE33126",getGPL = FALSE)
```

通过终端命令行下载可以看到输出以下报错，如下继续操作即可

```
错误: The size of the connection buffer (131072) was not large enough
to fit a complete line:
  * Increase it by setting `Sys.setenv("VROOM_CONNECTION_SIZE")`
```

```
Sys.setenv("VROOM_CONNECTION_SIZE" = 131072 * 5)
```

2.2 保存数据

```
gset <- getGEO("GSE33126",getGPL = FALSE)
save(gset,file="GSE33126.rdata")
```

2.3 加载 GEO 数据

```
load(file ="GSE33126.rdata")
```

2.4 查看数据

```
gset <- gset[[1]]
class(gset)

# 可以看到数据格式为 ExpressionSet，此数据格式同时包含表达矩阵与样本信息表
```

2.5 清洗样本信息表

```
sampleinfo <- pData(gset) %>%# 提取样本信息表
  select(source_name_ch1,characteristics_ch1.1) %>%
  rename(group = source_name_ch1,patient=characteristics_ch1.1) %>%
  mutate_at(vars(patient),~str_split(.," ",simplify = T)[,2])
```

2.6 查看表达矩阵

```
gene_exp <- exprs(gset)
```

通过以上代码我们轻松下载到了 GEO 数据, **最关键的点当然是网速了**

下面来介绍通过 Python 下载 TCGA 数据

- (https://github.com/vappiah/DataMiner) ,
- TCGA 官网下载样本清单, 运行下列几行代码即可轻松下载 TCGA 数据
- **最关键的点当然还是网速了**

3 TCGA 数据下载

作者有配套的视频教程，后续我会上传到 B 站供大家观看

```
import os
os.chdir("~/Desktop/TCGA/DataMiner-main")
from tcga_downloader import *
ids=get_ids('gdc_manifest.txt')
payload=prepare_payload(ids,data_type='Gene Expression Quantification')
metadata=get_metadata(payload)
download_data(metadata,sep="\t",outdir="BRCA")
```