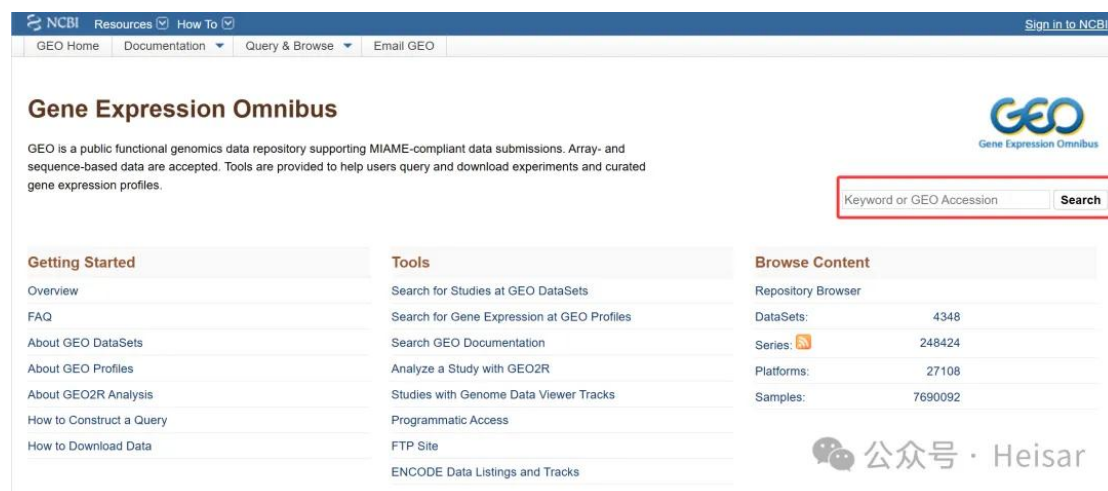


# 【保姆级教程】bulk-RNAseq 的一般技术流程（1）——原始数据（.fastq 文件）的获取

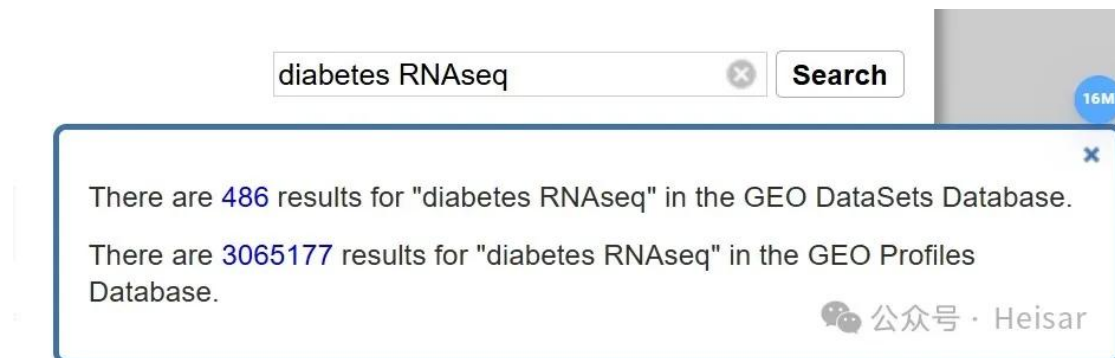
bulk-RNAseq 是最普通，最廉价的 RNA 测序技术（相比 scRNAseq），同时也是分析起来最简单的转录组数据类型（大概）。

## 查找所需的数据集

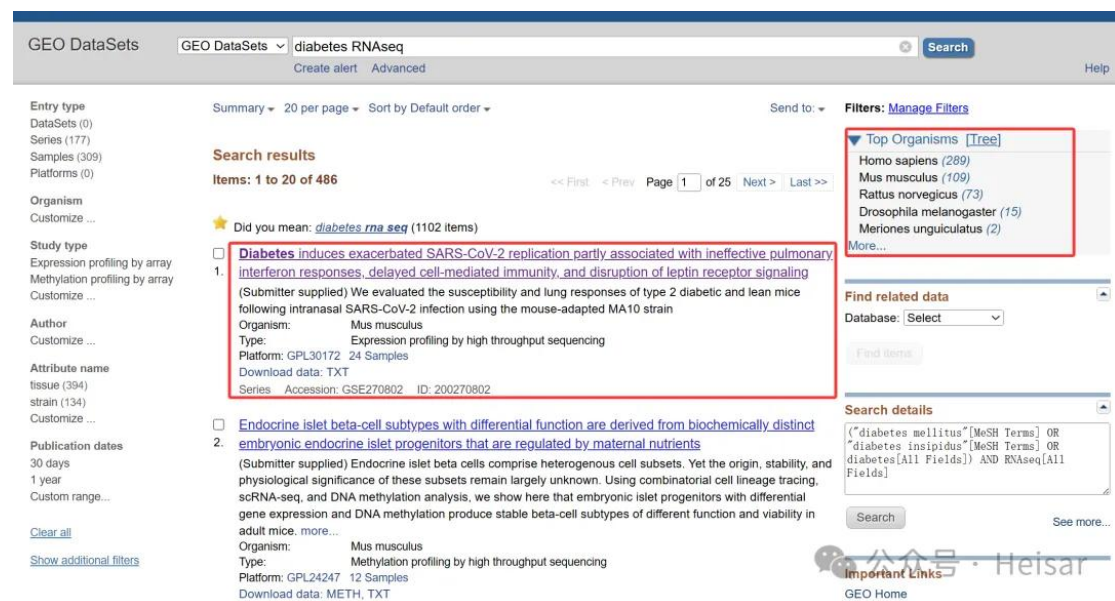
通常来说，一个 bulk-RNAseq 样品的送样分析会需要 500 元 RMB 左右，如果您有 3 个分组，每组 4 个样本，则需要 6000 元 RMB 左右。当然，在您正式送样之前，您也可以先在公开的数据库 (GEO) 中淘一淘金子，这有可能为您减少一部分不必要的开支。以 GEO 数据库为例，首先您需要打开 GEO 的网站 (Home - GEO - NCBI)，然后在它的搜索框里输入关键词进行搜索（如下图）。关键词记得带上 RNAseq 这样更容易搜到 bulk-RNAseq 的结果。





例如，我在这里键入的关键词是 diabetes 和 RNAseq，点击 Search 后会有下方两个选项，这里我们一般选上面那个来搜索数据集（486 个结果），下面那个搜索的则是具体的样本（3065177 个结果）。



你可以在搜索结果界面右上方筛选你需要的物种（选对物种非常重要）。结果条目中也会显示数据集的物种来源（Organism），以及很重要的数据集类型（Type）。一般包括 bulk-RNAseq 在内的转录组数据都会被描述为 Expression profiling by high throughput sequencing (高通量测序得到的转录本)，而具体是 bulk 还是 single cell，你需要查看数据集的详情才能确定。



点击上图中第一个条目进入到 GSE270802 的详情页，在 Overall design 一栏中，上传者描述道 Total RNA was ... and subjected to RNAseq，此时基本可以认为这是一个 bulk-RNAseq 的数据集（因为 scRNAseq 一般还会叙述分离单细胞的过程）。



  
Gene Expression Omnibus

[HOME](#) | [SEARCH](#) | [SITE MAP](#)
[GEO Publications](#) | [FAQ](#) | [MIAME](#) | [Email GEO](#)

NCBI > GEO > [Accession Display](#)
Not logged in | [Login](#)

Scope:  Format:  Amount:  GEO accession:

Series **GSE270802**
[Query DataSets for GSE270802](#)

Status	Public on Feb 19, 2025
Title	Diabetes induces exacerbated SARS-CoV-2 replication partly associated with ineffective pulmonary interferon responses, delayed cell-mediated immunity, and disruption of leptin receptor signaling
Organism	<a href="#">Mus musculus</a>
Experiment type	Expression profiling by high throughput sequencing
Summary	We evaluated the susceptibility and lung responses of type 2 diabetic and lean mice following intranasal SARS-CoV-2 infection using the mouse-adapted MA10 strain
Overall design	Diabetic Leprdb mice and lean heterozygote mice were intranasally inoculated with 20,000 PFU of mouse-adapted SARS-CoV-2 MA10 strain and lungs were collected and processed for analysis at 2 and 4 days post-infection. <u>Total RNA was isolated from homogenized lung and subjected to RNAseq</u>
Contributor(s)	<a href="#">Thieulent CJ</a> , <a href="#">Balasuriya UB</a> , <a href="#">Carossino M</a>
Citation missing	Has this study been published? Please <a href="#">login</a> to update or <a href="#">notify GEO</a> .

 公众号 · Heisar

当然，如果您很急，加上数据集的 Overall design 写得又臭又长，让人不忍卒读，那么您也可以直接下滑到最后，检查它的 Samples 和 Supplementary file 栏目。一般来说，bulk-RNAseq 的 Samples 数量会相当多，如 GSE270802 就有 24 个 Samples（而 scRNAseq 一般只会有个位数，毕竟一个样就要上万元，这么多钱都能全款拿下小米跑车了）。其次，bulk-RNAseq 的数据集如果在 Supplementary file 栏目中提供 count 文件的话（即图中的 GSE270802\_raw\_counts.txt.gz），文件大小一般在几百 Kb 到几 Mb（而 scRNAseq 一般在几十 Mb 到几 Gb，并且还会细分出 matrix，features 和 barcodes 文件，以后讲到 scRNAseq 的话会细说）。

Platforms (1) [GPL30172](#) NextSeq 2000 (Mus musculus)

Samples (24) [GSM8352308](#) HET\_M\_01\_OD2  
[GSM8352309](#) HET\_M\_02\_OD2  
[GSM8352310](#) HET\_M\_03\_OD2

#### Relations

BioProject [PRJNA1128525](#)

#### Download family

[SOFT formatted family file\(s\)](#)  
[MINiML formatted family file\(s\)](#)  
[Series Matrix File\(s\)](#)

#### Format

[SOFT](#) [?](#)  
[MINiML](#) [?](#)  
[TXT](#) [?](#)

Supplementary file	Size	Download	File type/resource
GSE270802_raw_counts.txt.gz	869.0 Kb	<a href="#">(ftp)</a> <a href="#">(http)</a>	TXT

[SRA Run Selector](#) [?](#)

Raw data are available in SRA

公众号 · Heisar

## 下载所需的数据集

此时，如果您对 bulk-RNAseq 的上游分析不感兴趣，您可以直接下载上传者提供的 count 文件（点击上图的 ftp 或 http 就可以下载了），随后就能开展差异分析（以后会细说差异分析如何开展）。当然，因为本文旨在教学如何获取.fastq 文件，所以继续点击上图的 SRA RUN selector，进入到数据集的 SRA RUN selector 页面，然后下滑到 Select 栏目，点击下图中 Accession List，下载名为 SRR\_Acc\_List.txt 的文本文件。

SRA Run Selector

Common Fields

BioProject	PRJNA1128525
Consent	PUBLIC
AGE	10-week old
Assay Type	RNA-Seq
AvgSpotLen	74
Center Name	GENOMICS CORE FACILITY, PENNINGTON BIOMEDICAL RESEARCH CENTER
Collection Date	missing
DATASTORE filetype	FASTQ, RUN, ZQ, SRA
DATASTORE provider	GS, NCBI, S3

Select

	Runs	Bytes	Bases	Download	Cloud Data Delivery	Computing
Total	24	15.80 Gb	37.83 G	Metadata or <b>Accession List</b>		
Selected	0	0	0	Metadata or Accession List or JWT Cart	Deliver Data	Galaxy

Found 24 Items

Search within results

Q Clear

公众号 · Heisar

SRR\_Acc\_List.txt 的内容如下图所示，每个 SRR 条目都代表一个数据集的 Sample，具体哪个 SRR 对应哪个 Sample (GSM) 可以在上图的 SRA RUN selector 页面的最后

查看（识别哪个 SRR 是哪个 GSM 非常重要，因为有些愚蠢的上传者会打乱 GSM 和 SRR 的顺序）。



```
SRR29553223
SRR29553224
SRR29553225
SRR29553226
SRR29553227
SRR29553228
SRR29553229
SRR29553230
SRR29553231
SRR29553232
SRR29553233
SRR29553234
SRR29553235
SRR29553236
SRR29553237
SRR29553238
SRR29553239
SRR29553240
SRR29553241
SRR29553242
SRR29553243
SRR29553244
SRR29553245
SRR29553246
```
























接下来的操作需要在 Linux 系统下进行（当然最佳选择是在服务器上操作，如果您没有一个有容乃大的硬盘的话）。

首先您需要安装一个叫做 `sra-tools` 的软件（GitHub - ncbi/sra-tools: SRA Tools），然后在命令行中用以下代码下载您需要的 SRR（在运行代码之前最好检查一下 SRR 的文件大小，如果有任何一个 SRR 超过了 25Gb，则必须调整代码中 XXGB 到您要下载的最大 SRR 文件的文件大小以上，比如您要下载的 SRR 中最大的一个有 75Gb 大，则设置 XX 为 75 以上）。因为该代码引入了无挂断+后台运行的功能，所以跑上之后可以不管，您只需要每隔一段时间检查一下下载进度就可以了。

```
1 #SRA RUN selector 上获取 SRR 列表 SRR_Acc_List.txt
2 nohup prefetch -O . $(<SRR_Acc_List.txt) --max-size XXGB &
```

下载完成后，您会得到一系列 SRR 开头的文件夹，每个文件夹下是一个.SRA 文件。这些.SRA 文件本质是压缩的.FASTQ 文件，后者正是您后续所需要的文件格式，所以您要做的就是将.SRA 重新转换成.FASTQ。这里需要用到另一个叫做 `fasterq-dump` 的软件。



名字	大小	已改变
 SRR24709225		16/8/2024 13:05:08
 SRR24709224		16/8/2024 13:01:36
 SRR24709223		16/8/2024 12:57:35
 SRR24709222		16/8/2024 12:48:03
 SRR24709221		16/8/2024 12:46:18
 SRR24709220		16/8/2024 12:42:47
 SRR24709219		16/8/2024 12:38:27
 SRR24709218		16/8/2024 12:34:20
 SRR24709217		16/8/2024 12:30:20
 SRR24709216		16/8/2024 12:28:40
 SRR24709215		16/8/2024 12:26:50
 SRR24709214		16/8/2024 12:23:13
 SRR24709213		16/8/2024 12:21:53
 SRR24709212		16/8/2024 12:18:44
 SRR24709211		16/8/2024 12:14:31
 SRR24709210		16/8/2024 12:13:25
 SRR24709209		16/8/2024 12:12:04
 SRR24709208		16/8/2024 12:11:07
 SRR24709207		16/8/2024 12:07:27
 SRR24709206		16/8/2024 12:02:50
 SRR24709205		16/8/2024 11:59:02
 SRR24709204		16/8/2024 11:57:10
 SRR24709203		16/8/2024 11:55:06
























安装完 `fasterq-dump` 后运行以下代码即可转换所有.SRA 文件。

```

1 # 设置输出路径（即您想放置转换好的.fastq 文件的路径）
2 OUTPUT_DIR="./GSEXXXXXXXX"
3 # 运行 fasterq-dump 命令
4 fasterq-dump ./${<SRR_Acc_List.txt} -O $OUTPUT_DIR

```

转换后的 fastq 如图所示（如果样本是采取双端测序法测序的则一个 SRR 会被转换成尾缀为\_1 和\_2 两个.fastq 文件，单端测序法测序的则只产生一个）。

名字	大小	已改变
 SRR24709226_1.fastq	10,705,4...	16/8/2024 17:42:23
 SRR24709226_2.fastq	10,707,9...	16/8/2024 17:42:22
 SRR24709225_1.fastq	11,894,3...	16/8/2024 17:38:34
 SRR24709225_2.fastq	11,899,7...	16/8/2024 17:38:22
 SRR24709224_1.fastq	15,604,3...	16/8/2024 17:33:50
 SRR24709224_2.fastq	15,604,5...	16/8/2024 17:33:39
 SRR24709223_1.fastq	27,457,8...	16/8/2024 17:27:09
 SRR24709223_2.fastq	27,473,8...	16/8/2024 17:27:07
 SRR24709222_1.fastq	5,316,07...	16/8/2024 17:15:17
 SRR24709222_2.fastq	5,318,77...	16/8/2024 17:15:16
 SRR24709221_1.fastq	14,796,5...	16/8/2024 17:13:17
 SRR24709221_2.fastq	14,803,2...	16/8/2024 17:13:11
 SRR24709220_1.fastq	12,954,7...	16/8/2024 17:07:28
 SRR24709220_2.fastq	12,958,7...	16/8/2024 17:07:18
 SRR24709219_1.fastq	14,107,9...	16/8/2024 17:03:56
 SRR24709219_2.fastq	14,109,6...	16/8/2024 17:03:55
 SRR24709218_2.fastq	14,154,7...	16/8/2024 17:01:51
 SRR24709218_1.fastq	14,146,5...	16/8/2024 17:01:50
 SRR24709217_2.fastq	5,199,97...	16/8/2024 16:57:45
 SRR24709217_1.fastq	5,197,82...	16/8/2024 16:57:42
 SRR24709216_1.fastq	4,140,06...	16/8/2024 16:55:38
 SRR24709216_2.fastq	4,140,63...	16/8/2024 16:55:33
 SRR24709215_2.fastq	12,001,8...	16/8/2024 16:53:55

通过以上步骤，想必您一定能获取到自己需要的.fastq 文件了，这是一切 RNA 或 DNA 测序结果的原始形态，您可以在命令行中运行以下代码查看其中的内容。不过您也不必细究.fastq 里到底有什么，毕竟光凭人类的肉眼和直觉，就算.fastq 里的序列有问题，您也是看不出来的。您如果怀疑测序的质量问题，需要更加专业的 QC 软件，例如 fastqc。

```
1 head -n 10 SRR24709226_1.fastq
```