

Steps for downloading Sequence Data from NCBI

This example goes over downloading sequences used in this study:

<https://www.frontiersin.org/articles/10.3389/fcimb.2020.00063/full#h6>

The authors specifically state:

“Sequence data are available from SRA BioProject PRJNA607849

(<http://www.ncbi.nlm.nih.gov/bioproject/607849>).”

1. Getting Specific Accession Codes

- Click on the ncbi link above
- Click on the number **291** under the **Number of Links** for **SRA Experiments**

The screenshot displays the NCBI BioProject page for PRJNA607849. The page title is "Profiling the gut microbiome of healthy and type 2 diabetes individuals in urban Nigeria." The project description states: "The composition of gut microbiome has been shown to be associated with metabolic disorders, including type 2 diabetes (T2D). More..."

Project Details:

- Accession: PRJNA607849
- Data Type: Raw sequence reads
- Scope: Multispecies
- Grants:
 - "The Africa America Diabetes Mellitus Study" (Grant ID 3T37TW00041-0352, National Institute on Minority Health and Health Disparities, NIH)
 - "MINORITY INTERNATIONAL RESEARCH TRAINING GRANT" (Grant ID T37 TW000041, Fogarty International Center)
- Submission: Registration date: 20-Feb-2020, NIH
- Relevance: Medical

Project Data:

Resource Name	Number of Links
Sequence data	
SRA Experiments	291
Other datasets	
BioSample	291

A red arrow points to the number 291 under the SRA Experiments row.

SRA Data Details:

Parameter	Value
Data volume, Gbases	20
Data volume, Mbytes	13057

Related information: BioSample, SRA

Recent activity:

- SRP250213[All Fields] (291) SRA
- SRP250213[All Fields] AND diabetic (98) SRA
- SRA Links for BioProject (Select 607849) (291) SRA
- SRP250213 (291) SRA

See more...

c. This will take you here:

SRA

SRP250213[All Fields]

Search

Create alert Advanced

Summary 20 per page

Send to: Filters: Manage Filters

Find related data

Database: Select

Find items

Search details

SRP250213[All Fields]

Search

See more...

Recent activity

Turn Off Clear

Q SRP250213[All Fields] (291) SRA

Q SRP250213[All Fields] AND diabetic (98) SRA

Q SRA Links for BioProject (Select 607849) (291) SRA

Q SRP250213 (291) SRA

See more...

Access Public (291)

Source DNA (291)

Library Layout paired (291)

Platform Illumina (291)

Strategy other (291)

Data in Cloud GS (291) S3 (291)

File Type fastq (291)

Clear all

Show additional filters

Search results

Items: 1 to 20 of 291

<< First < Prev Page 1 of 15 Next > Last >>

1. ☐ Fecal microbiome (16S rRNA gene amplicon) of a healthy adult
1 ILLUMINA (Illumina MiSeq) run: 148,809 spots, 74.7M bases, 46.4Mb downloads
Accession: SRX7764438

2. ☐ Fecal microbiome (16S rRNA gene amplicon) of a healthy adult
1 ILLUMINA (Illumina MiSeq) run: 131,835 spots, 66.2M bases, 40.7Mb downloads
Accession: SRX7764437

3. ☐ Fecal microbiome (16S rRNA gene amplicon) of a healthy adult
1 ILLUMINA (Illumina MiSeq) run: 170,693 spots, 85.7M bases, 53.7Mb downloads
Accession: SRX7764436

4. ☐ Fecal microbiome (16S rRNA gene amplicon) of a diabetic adult
1 ILLUMINA (Illumina MiSeq) run: 177,181 spots, 88.9M bases, 55.4Mb downloads
Accession: SRX7764435

5. ☐ Fecal microbiome (16S rRNA gene amplicon) of a healthy adult
1 ILLUMINA (Illumina MiSeq) run: 129,679 spots, 65.1M bases, 41.5Mb downloads
Accession: SRX7764434

6. ☐ Fecal microbiome (16S rRNA gene amplicon) of a healthy adult
1 ILLUMINA (Illumina MiSeq) run: 97,797 spots, 49.1M bases, 25.4Mb downloads
Accession: SRX7764433

d. We are going to take only a subset of the data available to us here. Specifically, we only want sequence data taken from adults who are diabetic. We can subset the search results by typing “diabetic” into the **Search details** box and hit **enter**.

Search details

SRP250213[All Fields] diabetic

Search

See more...

e. Now on our current **Search results** page we are only seeing sequence data taken from diabetic adults. Click on the **Send to:** (in the top right corner) and

Send to: Filters: Manage F

Choose Destination

☒ File ☐ Clipboard

☐ Collections ☐ BLAST

☐ Run Selector

select **File**.

- f. In the dropdown window, change the **Format** to **Accession List** and click **Create File**. This will download all accession codes for diabetic adults in the current study. Save this file for later. It will be named something similar to `SraAccList.txt`.
2. **Download and install the NCBI SRA Toolkit**
(Note: If working on the CGRB, the `sratoolkit` is pre-installed)
 - a. Got to this link: <https://github.com/ncbi/sra-tools/wiki>
 - b. Follow steps 1, 2, and possibly 3

MAKE SURE THE TOOLKIT IS FUNCTIONAL BEFORE MOVING FORWARD

3. Download all sequence data

(Reference: https://www.reneshbedre.com/blog/ncbi_sra_toolkit.html#customized-download-of-sra-datasets)

- a. Write a bash script that downloads all sequence data for the diabetic adults in the study. Here, the list of accession codes is `SraAccList.txt` and it is stored in the folder `~/Documents/retrieve_ncbi_files`.

```
# program goes through NCBI Accession list and downloads all data pertaining to
the Accession code

cd ~/Documents/retrieve_ncbi_files

echo "Start!"
for line in $(cat SraAccList.txt)
do
    echo "Current Accession Code: " $line
    fastq-dump $line
done
~
~
~
```

Note:

1. If the FASTQ files are paired-end, use `--split-files`.
EX: `fastq-dump --split-files $line`
If you don't use `--split-files` for paired-ends, the reads will be merged from both ends.
2. You can also convert the FASTQ files to FASTA while downloading.
EX: `fastq-dump --fasta $line`

4. Upload the data to the CGRB

- a. Useful documentation for easy upload/download to CGRB:
https://shell.cqls.oregonstate.edu/files/cgrb_files_access.pdf

Note: The `sratoolkit` is already installed on the CGRB in `/local/cluster`

No need to download it directly to your own computer if you don't need to.

Note: If downloading directly to the CGRB, you will need to run the command: `vdb-config --interactive`

<https://github.com/ncbi/sra-tools/wiki/03.-Quick-Toolkit-Configuration>

You will see a screen where you operate the buttons by pressing the letter highlighted in red, or by pressing the tab-key until the wanted button is reached and then pressing the space- or the enter-key.

1. You want to enable the "Remote Access" option on the Main screen.
2. Proceed to the "Cache" tab where you will want to enable "local file-caching" and you want to set the "Location of user-repository".

a) The repository directory needs to be set to an empty folder. This is the folder where prefetch will deposit the files.

3. Go to your cloud provider tab and accept to "report cloud instance identity".

NOTE: You do not need to have an AWS or GCP account to download the data. Just choose a random one and you should be fine. The cloud instance identity only reports back in what cloud (AWS v GCP) you are working so you can access data for free.