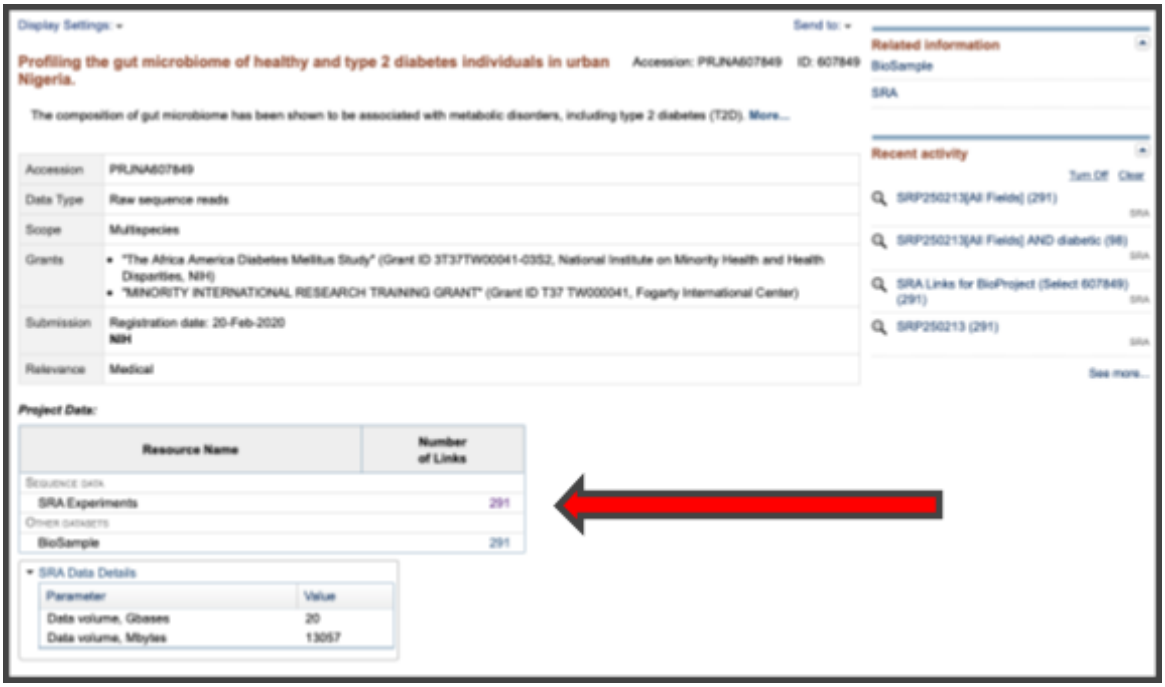
Steps for downloading Sequence Data from NCBI

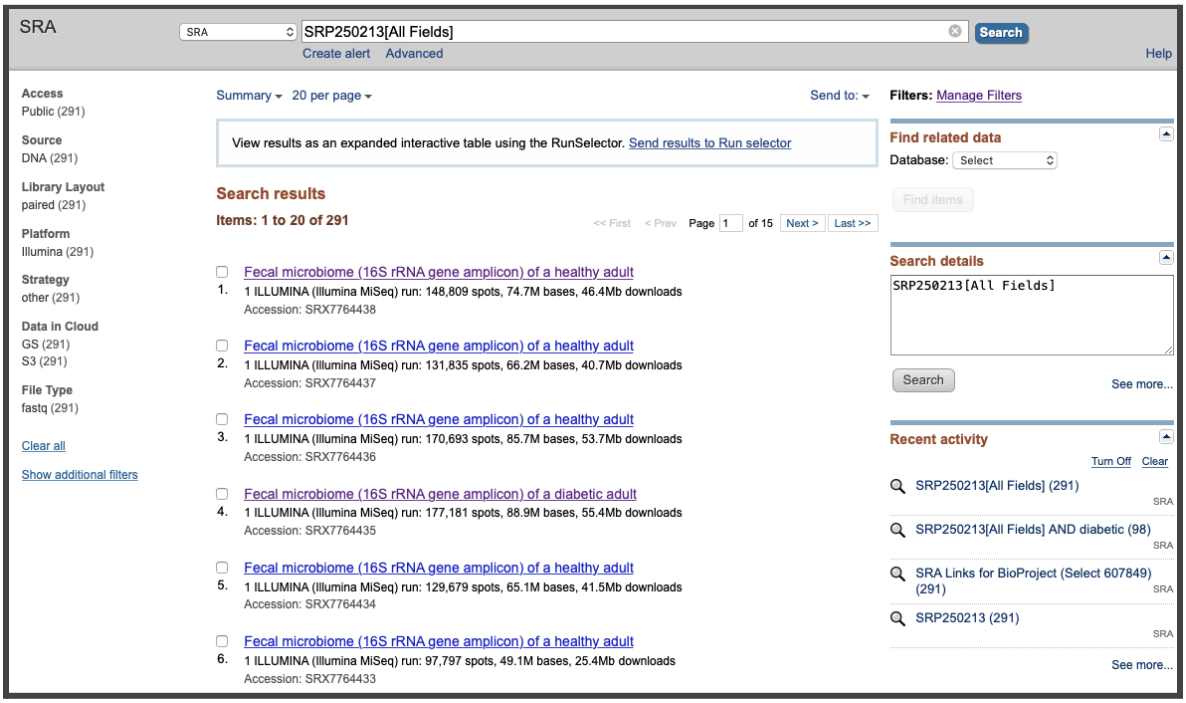
This example is goes over downloading sequences used in this study:
https://www.frontiersin.org/articles/10.3389/fcimb.2020.00063/full#h6

The authors specifically state:
“Sequence data are available from SRA BioProject PRJNA607849 (http://www.ncbi.nlm.nih.gov/bioproject/607849).”

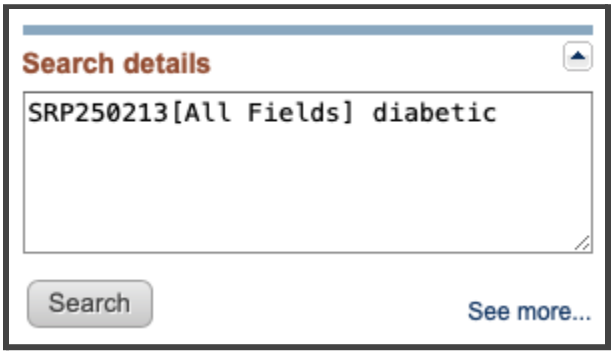
1. **Getting Specific Accession Codes**
    a. Click on the ncbi link above
    b. Click on the number **291** under the **Number of Links** for **SRA Experiments**



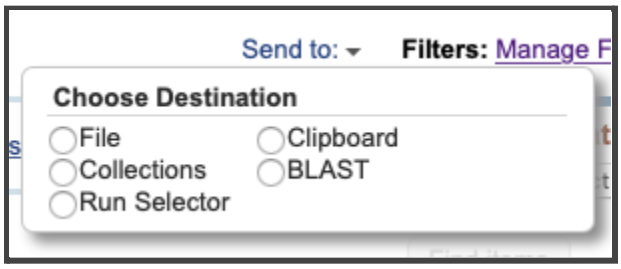
c. This will take you here:



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**.



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 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.