**Breast Cancer RNAseq data extraction using SRA toolkit**

**RNA seq data download**

NCBI GEO website - <https://www.ncbi.nlm.nih.gov/geo/>

Accession number – GSE183947

Click SRA Run Selector > click Accession list under Download tab > downloaded a txt file with SRR ids of 60 samples

These SR ids can be used to fetch the RNAseq data using the SRA toolkit

**SRA toolkit Download**

SRA toolkit download link (Github)- <https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit>

Downloaded for Arm64 architecture Mac

**SRA toolkit installation –**

1. Open mac terminal by entering the folder of saved file (right click on pathway toolbar > open Terminal)
2. Make a temporary directory

jagjotarora@Jagjots-MacBook-Air Breast cancer RNAseq data extraction (SRA) % mkdir tmp

1. Check the contents of folder

jagjotarora@Jagjots-MacBook-Air Breast cancer RNAseq data extraction (SRA) % ls

SRR\_Acc\_List.txt tmp

sratoolkit.3.1.0-mac-arm64.tar

so we have 3 files here 1)downloaded SRR file from NCBI geo

2) our newly made tmp directory

3) downloaded SRA github toolkit file (decompressed)

1. extract the material from github file.

jagjotarora@Jagjots-MacBook-Air Breast cancer RNAseq data extraction (SRA) % tar -xvzf sratoolkit.3.1.0-mac-arm64.tar

extraction complete, list is visible (to understand this code, view the additional info at end of this word file)

1. checking folder contents again

jagjotarora@Jagjots-MacBook-Air Breast cancer RNAseq data extraction (SRA) % ls

Breast cancer RNAseq data extraction.docx sratoolkit.3.1.0-mac-arm64.tar

SRR\_Acc\_List.txt tmp

sratoolkit.3.1.0-mac-arm64 ~$adder cancer project 2.docx

two new files are here 1)the docx file is this word file 2) our extracted sratoolkit file

1. Now we enter/change directory to this sratoolkit file

jagjotarora@Jagjots-MacBook-Air Breast cancer RNAseq data extraction (SRA) % cd sratoolkit.3.1.0-mac-arm64

1. Lets view its contents

jagjotarora@Jagjots-MacBook-Air sratoolkit.3.1.0-mac-arm64 % ls

CHANGES README.md schema

README-blastn bin

README-vdb-config example

1. Next we need to configure this toolkit. The executables are usually present in the “bin” folder which can help in configuring this toolkit

We do this using this command :- ./vdb-config -i

However macOS usually doesn’t let us open this interactive interface due to security reasons. Therefore we need to bypass the Gatekeeper security function in mac (temporarily disable it)

jagjotarora@Jagjots-MacBook-Air bin % sudo spctl --master-disable

It asks for password sometimes. Entered the computer password. Now we run our code.

jagjotarora@Jagjots-MacBook-Air bin % ./vdb-config -i

(For more info on this code, view additional info)

1. An blue interactive display is opened

A computer screen shot of a blue screen

Description automatically generated

We need to change some things here before we can start the SRA toolkit

Use the Tab button on the keyboard to move the red block and access the various panels and tabs. Press Enter to selct tab

1. Go to Main tab

Make sure the “Enable remote Access” is selected (X)+

1. Go to Cache tab

Make sure “Enable local file catching“ is selected (X)

But we need to change the location of the “User repository”

Go to Choose button and press enter

Ad location - /Users/jagjotarora/Desktop/data science /Breast cancer RNAseq data extraction (SRA)/tmp

We can copy the pathname from Finder and add here.

1. Save changes by pressing Tab key and going to the end of display > save configuration
2. After saved changes press the Exit button and end the display.
3. Now that we have extracted and configured our SRA toolkit, we can finally use it.

Dowload the sequencing files using a function/executable called as fasterq dump (available in the bin folder)

Going back one level

jagjotarora@Jagjots-MacBook-Air bin % cd –

Downloading 1 file (refer tot he txt file with SRR ids)

jagjotarora@Jagjots-MacBook-Air sratoolkit.3.1.0-mac-arm64 % bin/fasterq-dump --split-files SRR15852393

it usually takes time to download these files. There will be no apparent activity in terminal. To check if it is functioning, go to spotlight seach in mac and type “Activity monitor” to view if fasterq is using up the cpu ram actively or not.

**Additional Info**

1. **Tar Command**: The **tar -xvzf sratoolkit.2.11.3-mac64.tar.gz** portion is a command for extracting a tarball (a compressed archive file) in a Unix-based system like macOS or Linux. Here's a breakdown of the command:
   * **tar**: The program used to manipulate tarball files.
   * **-x**: Extract files from the archive.
   * **-v**: Verbosely list files processed (optional; it shows the progress).
   * **-z**: Filter the archive through gzip, decompressing it.
   * **-f**: Use the archive file **sratoolkit.2.11.3-mac64.tar.gz**.

This command extracts the contents of the **sratoolkit.2.11.3-mac64.tar.gz** file, which is a compressed archive file for the SRA Toolkit version 2.11.3 designed for macOS (64-bit).

1. The command **./vdb-config -I** is used in the context of configuring the SRA Toolkit. Here’s a detailed explanation:

* **./vdb-config**: This runs the **vdb-config** executable located in the current directory (**./**). The **vdb-config** tool is part of the SRA Toolkit and is used for configuring the local settings for the toolkit.
* **-I Option**: The **-I** (or **--interactive**) option opens the interactive configuration mode. This mode allows users to interactively set up and modify the configuration settings for the SRA Toolkit.

In summary, the command **./vdb-config -I** initiates the interactive configuration interface for the SRA Toolkit. This allows you to customize settings such as the location of the cache directory, proxy settings, and other parameters necessary for accessing and processing data from the Sequence Read Archive.