How to analyze SRA data – an introduction

PREFACE

Many research groups have over the years uploaded experimental data to the Gene Expression Omnibus repository (GEO). In many cases data have been preprocessed and normalized allowing others to easily download and start working with the data.

With the increased number of experiments using RNA-seq data, the results are often saved as raw data files in the Sequence Read Archive (SRA) hosted by NCBI. Data is then also available as copies on multiple servers in the cloud. SRA stores raw sequencing data and sometimes also alignment information.

To start working with data from SRA in Qlucore Omics Explorer more work is required. The steps include download, pre-processing and converting from SRA format to aligned BAM files.

This document shall be viewed as a template and a framework on how the data preparation can be done. It is not a complete guide and the workflows include open source tools which are outside the control of Qlucore and all use of this document as well as the mentioned tools are is up to the user. Qlucore takes no responsibility for results or outcomes.

Contents

1. Requirements 2
2. The steps that are required 2
3. Understanding the SRA HIERARCHY 2
4. Download files from the SRA repository 2
5. The NCBI SRA Toolkit 10
5.1. A General note on MacOS 10
5.2. Set the path in macOS 10.15 12
5.3. The VDB-CONFIG command 13
5.4. The PREFETCH command in the SRA Toolkit 15
5.5. The Fastq-dump command in the SRA Toolkit 16
6. Use an Aligner to convert FASTQ files to aligned BAM files 18
6.1. The STAR aligner 18
6.2. Download and prepare a fasta file 18
6.3. Run STAR on a single sample and on a folder with samples 19
7. Importing aligned BAM files into Omics Explorer 20
8. Change sample identifier and add annotations 22
9. Usage, Acknowledgements etc 24

10. Disclaimer 24

1. REQUIREMENTS

This document is written for users that would like to download and analyze SRA stored data on their own computer. The following tools are described:

1. NCBI SRA Toolkit (free tool). See section 4 for installation
2. STAR alignment tool (open source software). See section 5 for installation
3. Qlucore Omics Explorer (3.6 or later)

Later in the document you will get more info on where you can find these tools. References files (fasta and gtf) are also required. These should be of the same version as was used when the data was originally processed. Reference can be downloaded for instance from www.gencodegenes.org .

Further, this document requires that you are familiar with Mac OS X and optionally Microsoft Windows.

2. THE STEPS THAT ARE REQUIRED

This list shows the steps that are required and which we will go through below.

1. Identify the experiment and download the "Accession List" - a file called SraAccList.txt
2. Download all fastq files
3. Align the fastq files and store in BAM files
4. Load the BAM files into Qlucore Omics Explorer

3. UNDERSTANDING THE SRA HIERARCHY

In SRA there is a hierarchy defined like this:

- All data is organized into studies
- Studies contain samples
- Samples have experiments performed on them
- Experiments have results recorded by runs
- Runs describe results, how, and what made them
- Altogether, there is a hierarchy of provenance

There are a number of identifiers in SRA:

- Individual data files are called runs
- Runs have run IDs beginning with an "SRR" prefix
- Runs are described by other IDs such as Biosample
- Runs are contained in an SRA study
- An SRA study simply describes the raw data
- A similar entity might exist in GEO
- The GEO study would typically describe analyses
- Both datasets would be linked by a BioProject
- A BioProject ID encapsulates an entire study

4. DOWNLOAD FILES FROM THE SRA REPOSITORY

In GEO a series record links together a group of related Samples and provides a focal point and description of a whole study. Series records may also contain tables describing extracted

data, summary conclusions, or analyses. Each series record is assigned a unique and stable GEO accession number (GSExxxxx). The GSE number can be linked to a GEO dataset (GDS), a GEO platform (GPL) and GEO samples (GSM).

Let´s take a complex example, the GSE147507. In this example we have samples from two species (Homo sapiens and Mustela putorius furo), adding to the complexity.

More information is available at NCBI, you can use the link below:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147507

Information available on this web page is the following (note, the information may change for a project, this is a snapshot):

Series GSE147507 Query DataSets for GSE147507

Status	Public on Mar 25, 2020
Title	Transcriptional response to SARS-CoV-2 infection
Organisms	Homo sapiens; Mustela putorius furo
Experiment type	Expression profiling by high throughput sequencing
Summary	Viral pandemics pose an imminent threat to humanity. The ongoing COVID-19 pandemic, caused by the SARS-CoV-2 virus, requires the urgent development of anti-viral therapies. Because of its recent emergence, there is a paucity of information regarding viral behavior and host response following SARS-CoV-2 infection. Here, we offer an in-depth analysis of the host response to SARS-CoV-2 as it compares to other respiratory infections. Cell and animal models of SARS-CoV-2 infections, in addition to transcriptional profiling of a COVID-19 lung biopsy consistently revealed a unique and inappropriate inflammatory response defined by elevated chemokine expression in the absence of Type I and III interferons. Our identification of a muted transcriptional response to SARS-CoV-2 supports a model in which initial failure to rapidly respond to infection results in prolonged viral replication and an influx of proinflammatory cells that induce alveolar damage and manifest in COVID-19 lung pathology.
Overall design	Cell lines: Independent biological triplicates of primary human lung epithelium (NHBE) were mock treated or infected with SARS-CoV-2 (USA-WA1/2020), IAV (A/Puerto Rico/8/1934 (H1N1)), a IAV that lacks the NS1 protein (IAVdNS1) and treated with human interferon-beta. Independent biological triplicates of transformed lung alveolar (A549) cells were mock treated or infected with SARS-CoV-2 (USA-WA1/2020), RSV (A2 strain) or IAV (A/Puerto Rico/8/1934 (H1N1)). Additionally, Independent biological triplicates of transformed lung alveolar (A549) transduced with a vector expressing human ACE2, were also mock treated or infected with SARS-CoV-2 (USA-WA1/2020) with or without Ruxolitinib pre-treatment (500 nM). Finally transformed lung-derived Calu-3 cells were mock treated or infected with SARS-CoV-2 (USA-WA1/2020). Ferrets: 4 month old ferrets were infected intranasally with 105 PFU of influenza A/California/04/2009 (pH1N1) virus and nasal washes were collected from anesthetized ferrets on day 7 post infection. Additionally, another group of 4 month old ferrets were infected intranasally with 5 × 104 PFU of SARS-CoV-2 isolate USA-WA1/2020 and nasal washes were collected from anesthetized ferrets on days -1, 1, 3 and 7 post-infection. Finally, a separate group of 4 month old ferrets were mock treated (intranasally) with PBS. COVID19 patient samples: Uninfected human lung biopsies were derived from one male (age 72) and one female (age 60) and used as biological replicates. Additionally, lung samples derived from a single male COVID19 deceased patient (age 74) were processed in technical replicates. Experiments using samples from human subjects were conducted in accordance with local regulations and with the approval of the institutional review board at the Icahn School of Medicine at Mount Sinai under protocol HS#12-00145.
Contributor(s)	tenOever BR, Blanco-Melo D
Citation(s)	Blanco-Melo D, Nilsson-Payant BE, Liu WC, Uhl S et al. Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* 2020 May 28;181(5):1036-1045.e9. PMID: 32416070 https://doi.org/10.1101/2020.03.24.004655
Submission date	Mar 24, 2020
Last update date	May 26, 2020
Contact name	Daniel Blanco Melo
Organization name	Icahn School of Medicine at Mount Sina
Department	Microbiology
Lab	tenOever Lab
Street address	One Gustave L. Levy Place, Box 1124
City	New York
State/province	NY
ZIP/Postal code	10029
Country	USA
Platforms (2)	GPL18573 Illumina NextSeq 500 (Homo sapiens) GPL28369 Illumina NextSeq 500 (Mustela putorius furo)
Samples (110)	GSM4432378 Series1_NHBE_Mock_1 GSM4432379 Series1_NHBE_Mock_2 GSM4432380 Series1_NHBE_Mock_3
⊞ More...	

Relations

BioProject	PRJNA615032
SRA	SRP253951

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE147507_RawReadCounts_Ferret.tsv.gz	857.4 Kb	(ftp)(http)	TSV
GSE147507_RawReadCounts_Human.tsv.gz	1.8 Mb	(ftp)(http)	TSV

SRA Run Selector

Raw data are available in SRA

Processed data are available on Series record

Figure: Screenshot of Series GSE147507 at NCBI

There is a general introduction and then information about related information, like:

GEO platforms (GPL):
- GPL18573 - Illumina NextSeq 500 (Homo sapiens)
- GPL28369 - Illumina NextSeq 500 (Mustela putorius furo)

GEO samples (GSM):
There are in total 110 samples, of which some comes from Homo sapiens and some from Mustela putorius furo.

Project name:
- PRJNA615032

SRA identifier:
- SRP253951

There is some supplementary information, the project data is available in two raw read count matrixes (in compressed format):
- GSE147507_RawReadCounts_Ferret.tsv.gz
- GSE147507_RawReadCounts_Human.tsv.gz

If you now would like to further investigate the content, you can go directly to the SRA using the SRA identifier:

https://www.ncbi.nlm.nih.gov/sra?term=SRP253951

Here you get an overview of all sample files in the SRA, in this case it looks like this (note, the information may change for a project, this is a snapshot):

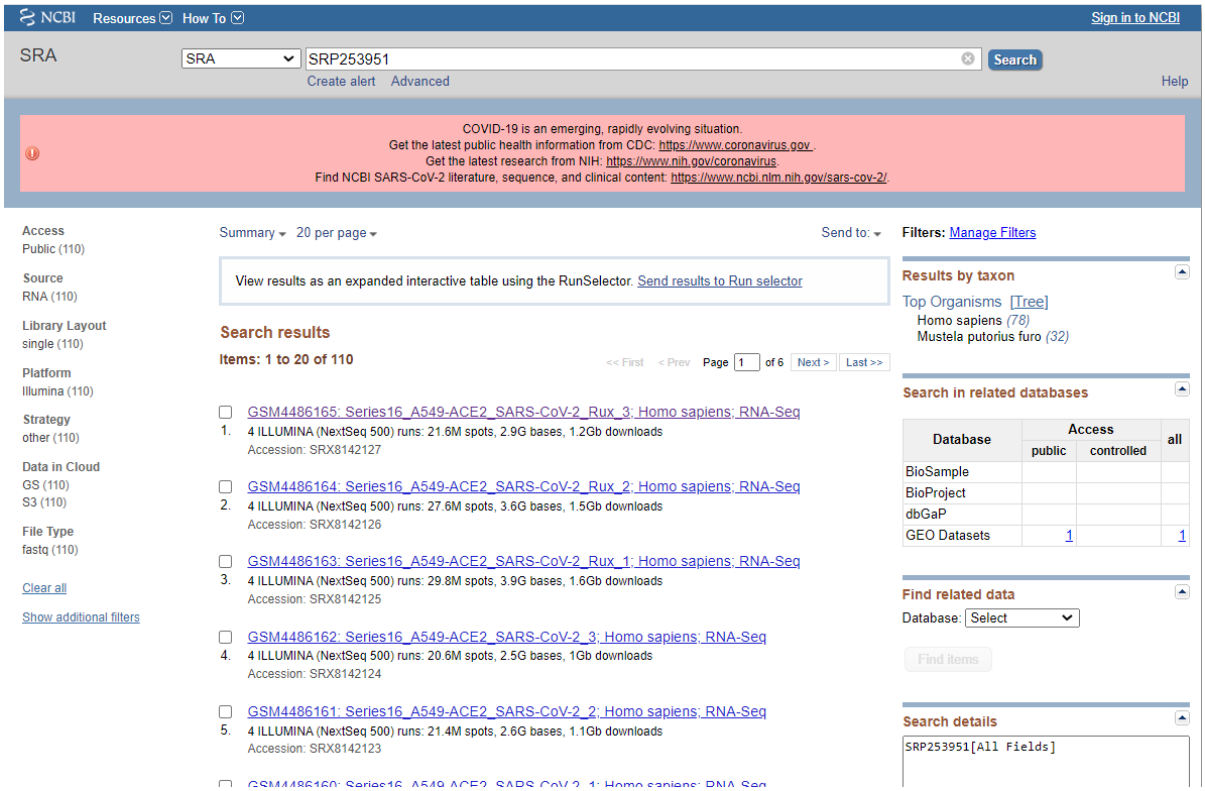


Figure: Screenshot of Samples in GSE147507 at NCBI

On this page you can also get a summary downloaded, that then can be used to download the individual files. You can also select it you would like to look only at the 78 Homo sapiens files, the 32 Mustela putorius furo files or all samples.

There is a hyperlink called “Send to” where you can select how you would like to download the summary.

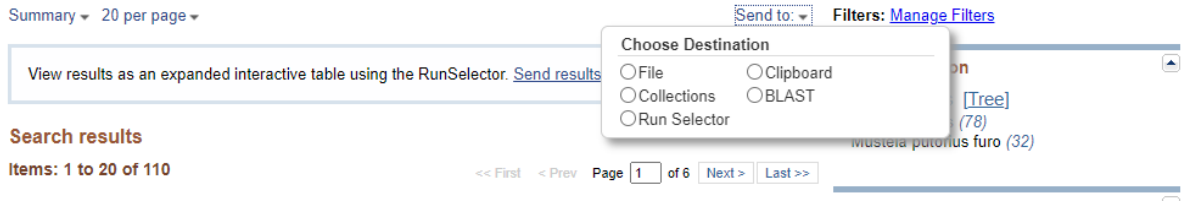


Figure: Screenshot of information download for GSE147507 at NCBI

Note that there also nowadays also alternative ways to connect, you can use the online SRA selector, found here:

https://www.ncbi.nlm.nih.gov/Traces/study/

You can then enter the GEO accession number, here GSE147507. You can then also select filters, like species, see below:

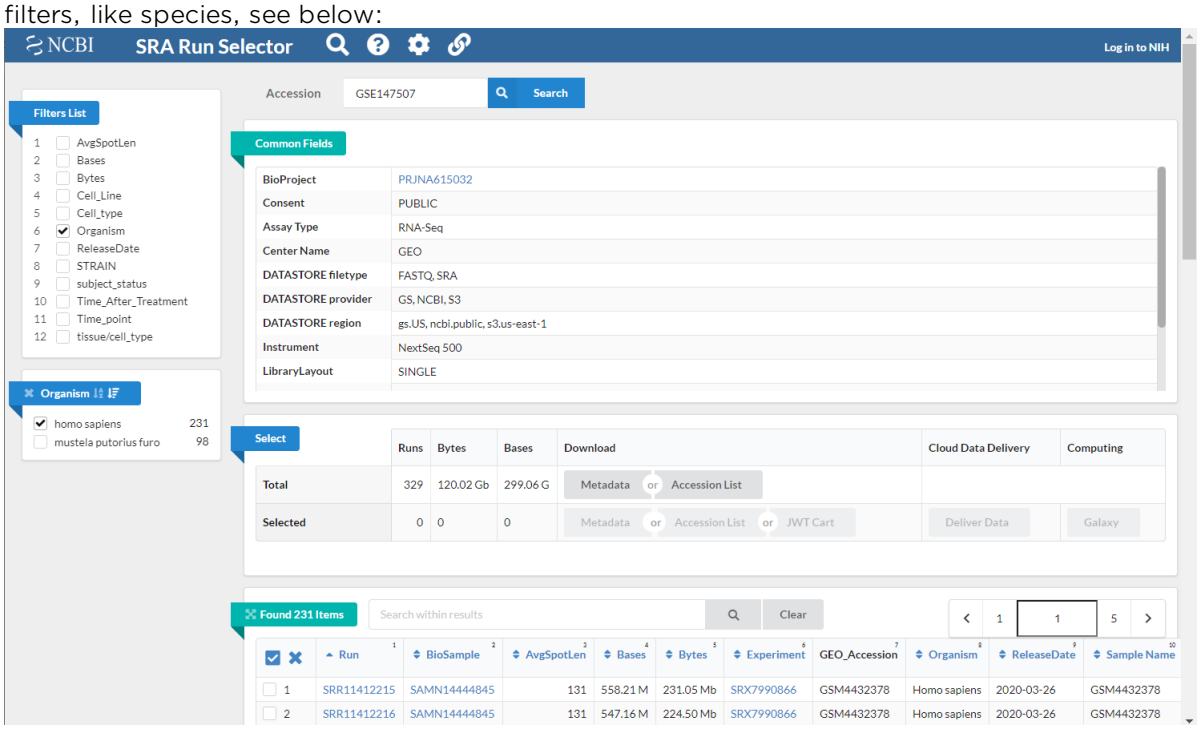


Figure: Screenshot of SRA Run Selector for GSE147507 at NCBI

You can select different filters, like species:

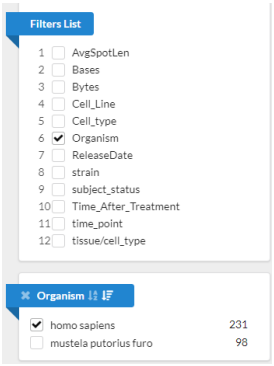
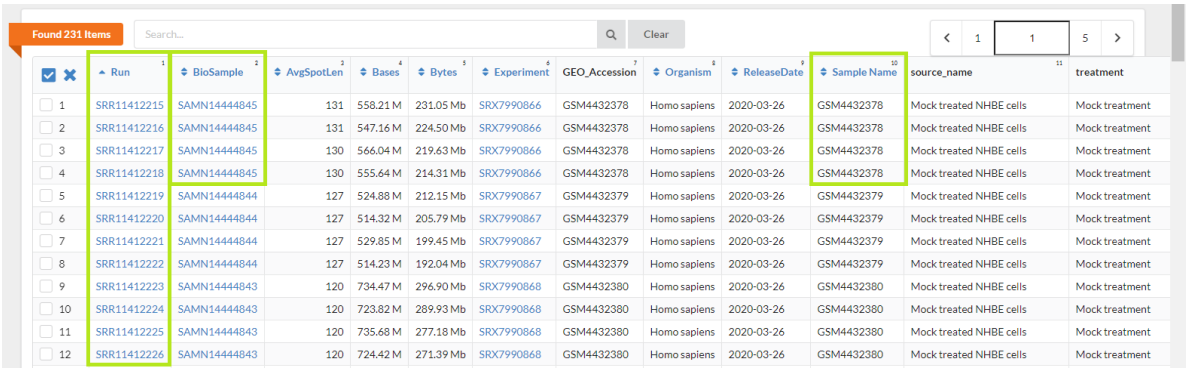


Figure: Screenshot of filters in SRA Run Selector for GSE147507 at NCBI



	Run	BioSample	AvgSpotLen	Bases	Bytes	Experiment	GEO_Accession	Organism	ReleaseDate	Sample Name	source_name	treatment
1	SRR11412215	SAMN14444845	131	558.21 M	231.05 Mb	SRX7990866	GSM4432378	Homo sapiens	2020-03-26	GSM4432378	Mock treated NHBE cells	Mock treatment
2	SRR11412216	SAMN14444845	131	547.16 M	224.50 Mb	SRX7990866	GSM4432378	Homo sapiens	2020-03-26	GSM4432378	Mock treated NHBE cells	Mock treatment
3	SRR11412217	SAMN14444845	130	566.04 M	219.63 Mb	SRX7990866	GSM4432378	Homo sapiens	2020-03-26	GSM4432378	Mock treated NHBE cells	Mock treatment
4	SRR11412218	SAMN14444845	130	555.64 M	214.31 Mb	SRX7990866	GSM4432378	Homo sapiens	2020-03-26	GSM4432378	Mock treated NHBE cells	Mock treatment
5	SRR11412219	SAMN14444844	127	524.88 M	212.15 Mb	SRX7990867	GSM4432379	Homo sapiens	2020-03-26	GSM4432379	Mock treated NHBE cells	Mock treatment
6	SRR11412220	SAMN14444844	127	514.32 M	205.79 Mb	SRX7990867	GSM4432379	Homo sapiens	2020-03-26	GSM4432379	Mock treated NHBE cells	Mock treatment
7	SRR11412221	SAMN14444844	127	529.85 M	199.45 Mb	SRX7990867	GSM4432379	Homo sapiens	2020-03-26	GSM4432379	Mock treated NHBE cells	Mock treatment
8	SRR11412222	SAMN14444844	127	514.23 M	192.04 Mb	SRX7990867	GSM4432379	Homo sapiens	2020-03-26	GSM4432379	Mock treated NHBE cells	Mock treatment
9	SRR11412223	SAMN14444843	120	734.47 M	296.90 Mb	SRX7990868	GSM4432380	Homo sapiens	2020-03-26	GSM4432380	Mock treated NHBE cells	Mock treatment
10	SRR11412224	SAMN14444843	120	723.82 M	289.93 Mb	SRX7990868	GSM4432380	Homo sapiens	2020-03-26	GSM4432380	Mock treated NHBE cells	Mock treatment
11	SRR11412225	SAMN14444843	120	735.68 M	277.18 Mb	SRX7990868	GSM4432380	Homo sapiens	2020-03-26	GSM4432380	Mock treated NHBE cells	Mock treatment
12	SRR11412226	SAMN14444843	120	724.42 M	271.39 Mb	SRX7990868	GSM4432380	Homo sapiens	2020-03-26	GSM4432380	Mock treated NHBE cells	Mock treatment

Figure: Screenshot of Samples/Runs overview in SRA Run Selector for GSE147507 at NCBI

Each sample may have several Runs. As an example, in this project, the sample GSM4432378 has 4 Runs, SRR11412215, SRR11412216, SRR11412217 and SRR11412218

Click on the first run there: SRR11412215 . You get this info:

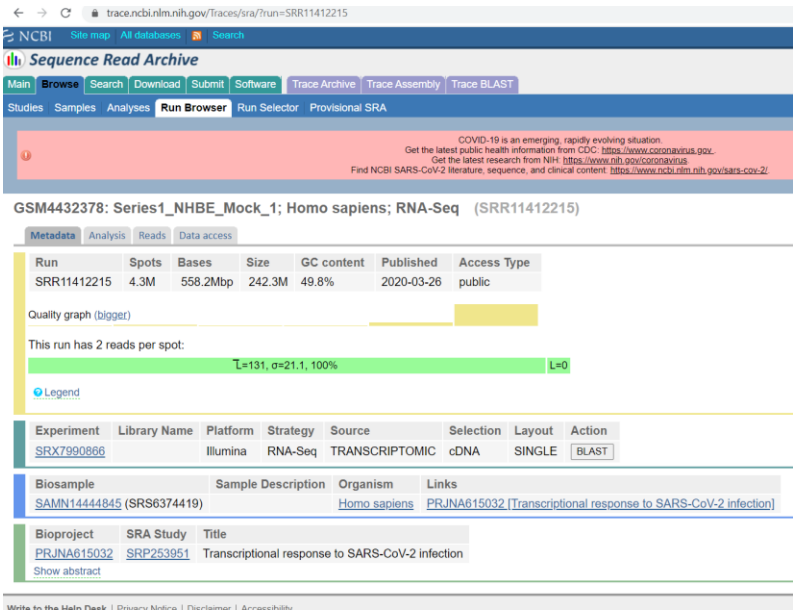


Figure: Screenshot of Samples/Runs overview in SRA Run Selector for GSE147507 at NCBI

Click on the first BioSample SAMN14444845 you get this info:

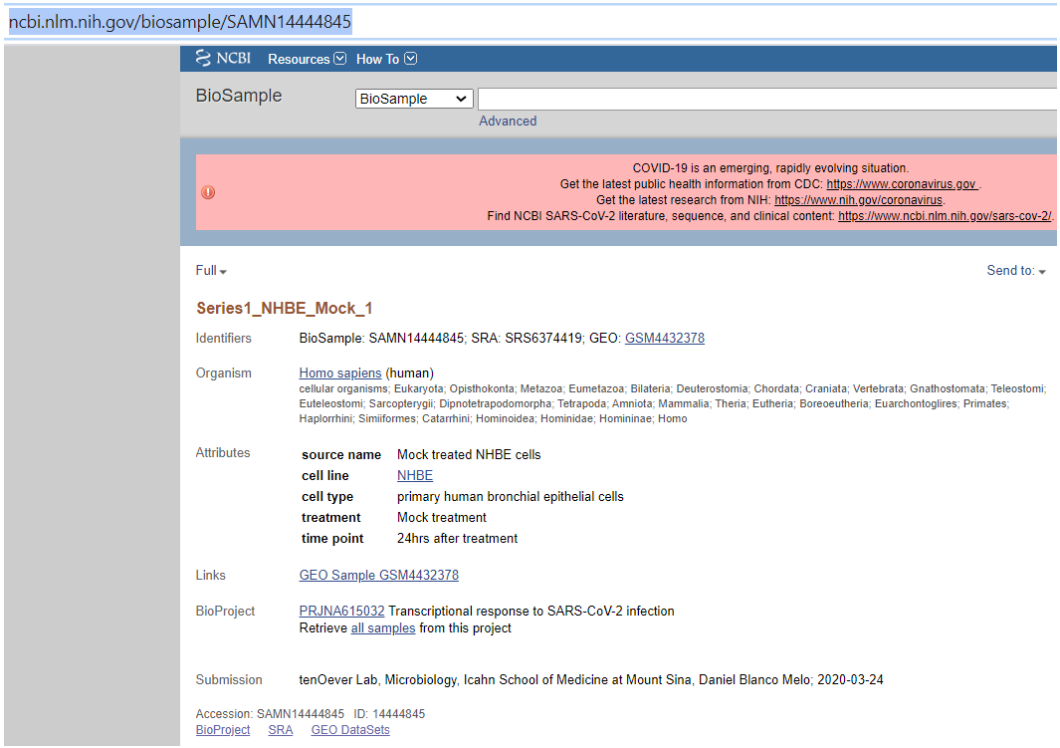


Figure: Screenshot of a BioSample

If you then click on the first Experiment: SRX7990866, you will get info on the Runs:

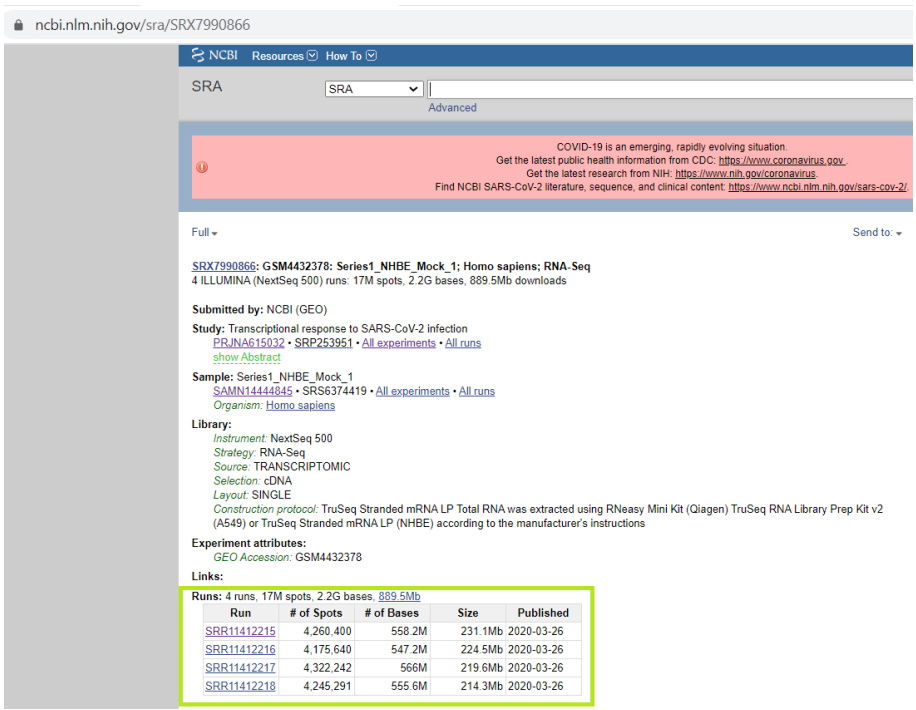


Figure: Screenshot of Runs for one Sample

As you can see, in this experiment, the runs for the samples in GSM4432378 all have about 200Mbyte of data each. Some samples have much larger data files, the SRR11517679 run belonging to the sample GSM4462341 has a size of 2,49 GByte indicating a very large file. This means that it is likely that it will take quite some time both to download and process a dataset like this.

Now, if we go back to https://www.ncbi.nlm.nih.gov/sra?term=SRP253951 we can filter by the organism too, like here "Homo sapiens".

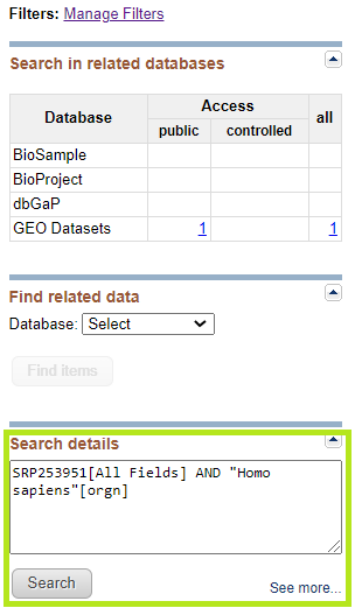


Figure: Screenshot of Search String

You can now select to get the info not in the Run Selector, but rather in a file. I can select which fomat to get ("Summary", "RunInfo", "Accession List" and "Full XML").

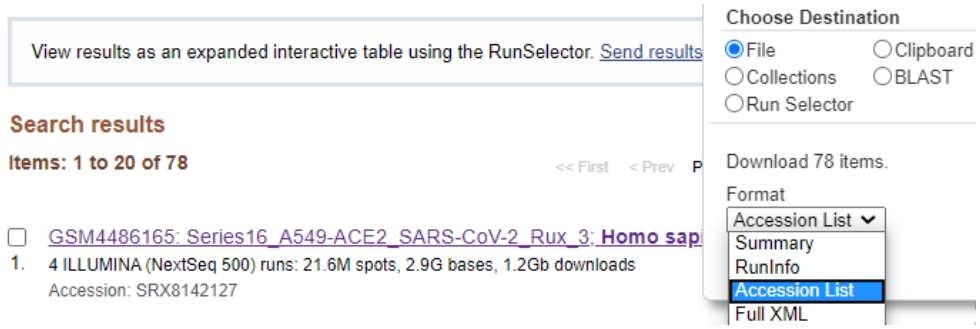


Figure: Screenshot of destination selection

If I select the format "Accession List" a file called SraAccList.txt is downloaded, in this case with one per line, for all the 231 SRRs available.

- SRR11412215
- SRR11412216
- SRR11412217

- SRR11412218
- SRR11412219
- SRR11412220
- Etc ..

You can also select to download RunInfo and get a much more detailed list.

The file SraAccList.txt can be used as an input to download of the individual files using the SRA toolkit software (available for PC, Mac etc).

5. THE NCBI SRA TOOLKIT

The SRA toolkit is available both as source files and pre-compiled versions at GitHub. The precompiled versions you find here.

The SRA Toolkit allows you to access data from the SRA and convert it from the SRA format to ta number of formats, like fasta, fastq and sam (human-readable bam, aligned or unaligned).

https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit

The SRA toolkit is available on macOS 64 bit architecture, MS Windows 64 bit architecture and several Linux distributions (CentOS Linux, Ubuntu Linux etc)

In this document the version used is sratoolkit.2.10.9. You will probably have a later version, which also means that the path names etc will be different compared to the ones in this document.

Also note that the paths will of course be different since you are likely to install the software in a different folder and also to keep your sra, fastq and bam files in other paths than the ones in the examples below.

Note that the files are compressed with tar, so you may need to decompress them using tar on macOS or other software on a PC (like 7-Zip).

macOS:

```
tar -xvf 2.10.9sratoolkit.2.10.9-mac64.tar
```

Once you have installed the toolkit you have several different commands available.

5.1. A GENERAL NOTE ON MACOS

It can be useful to add the directories where you have installed the sratoolkit and the STAR aligner to the $PATH, so that the operating system can find the executables directly.

To add a a directory to path on macOS 10.15 you do like this:
In the Finder, choose Go > Go to Folder. Type the name of the folder you would like to add to the path, like ~/STAR/, then click Go.

When you run commands on macOS (in this document macOS 10.15 (Catalina), both the sratools and later the STAR aligner you may get a security warning preventing you to execute the commands, like this:

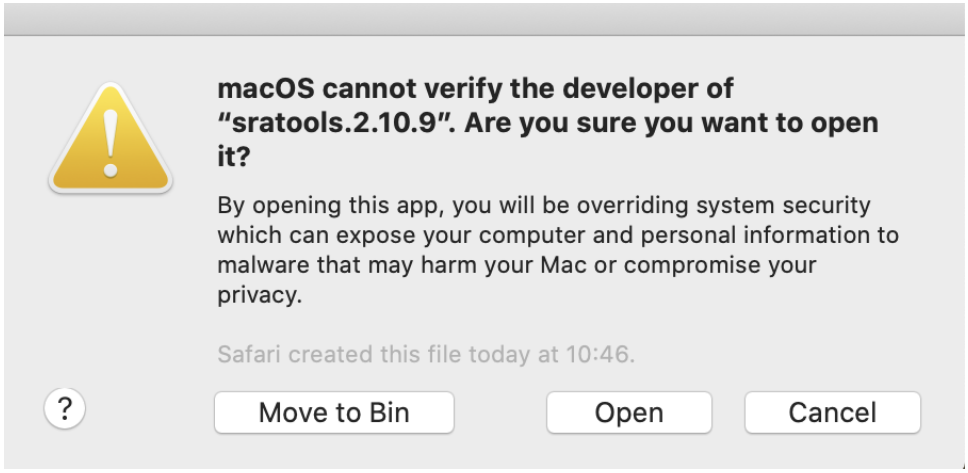


Figure: *Screenshot of security pop-up override*

Click the "Open" button to proceed.

Go to System Preferences -> Security & Privacy ->General

Figure: *Screenshot of Security & Privacy*

where you click the "Allow Anyway" button right to the message, something like this: "XXX2.10.9" was blocked from use because it is not from an identified developer" or something similar.

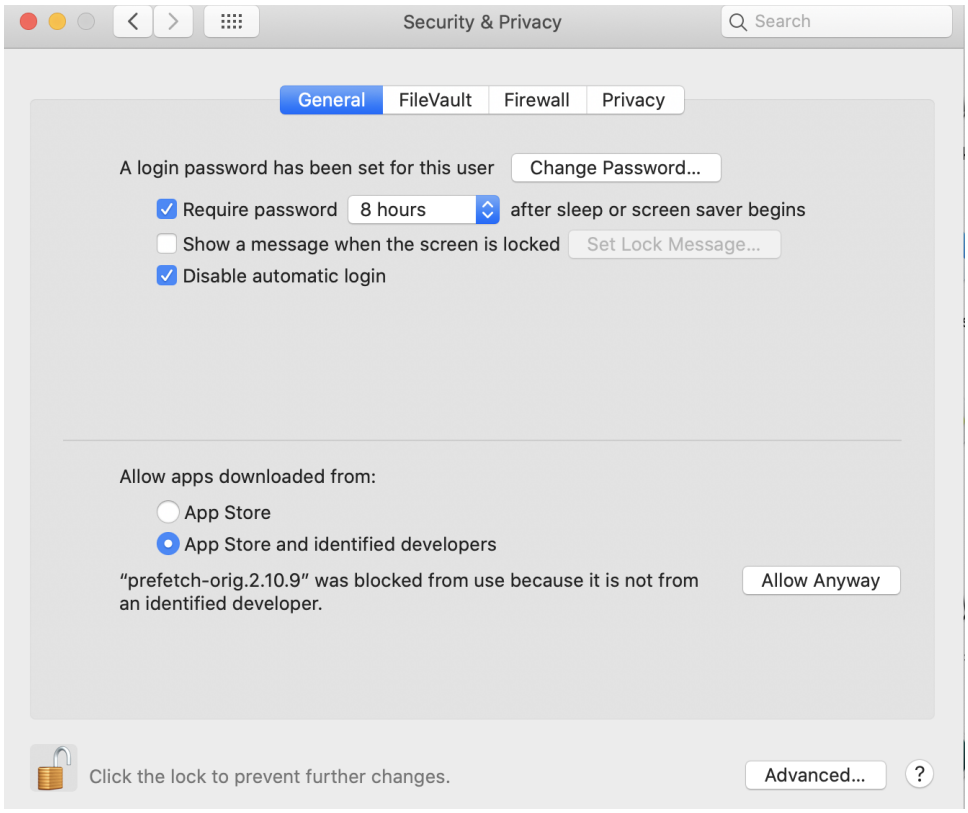


Figure: Screenshot of Security & Privacy ->General

Since some of the commands use underlying tools, this may have to be repeated, you probably need to do this repeatedly, also for for vdb-config, prefetch and for fastq-dump.

When you run the second time you may get the message "macOS cannot verify the developer of "sratoolkit.2.10.9"". Are you sure you want to open it? Then just click the "Open" button.

Also note that the default download directory is Users/[user name]/ncbi/public/sra in macOS unless you change it with the vdb-config command, see below.

5.2. SET THE PATH IN MACOS 10.15

macOS will find executable files that are available in folders defined by the PATH environment variable.

The sratool binaries and aligners you download will be stored in other folders. In order to avoid specifying the full path for a command you can update the $PATH variable.

You can start a program with the full path to the command, like below prefetch:

~/sratoolkit.2.10.9-mac64/bin/prefetch

If the folder where the executable resides is a part of $PATH you instead just type

prefetch

You can see what you PATH variable contains by starting a terminal window and write:

echo $PATH

It may look like this if not updated before:

/usr/local/bin:/usr/bin:/bin:/usr/sbin:/sbin

Now we want to update PATH.

As of macOS 10.15 (Catalina) the default shell is zsh (previously the bash shell).

In order to permanently add your own folder with executable programs or scripts to $PATH you need to create a .zsh file in your home directory and set the path there.

You can use the nano editor to edit PATH.

Move into your home directory and start nano like this:

nano .zsh

export PATH=/usr/local/sratoolkit.2.10.9-mac64/bin/:$PATH
Add in the above line which declares the new location /usr/local/sratoolkit.2.10.9-mac64/bin/ as well as the original path declared as $PATH.

Save the file in nano by clicking "control" + "o" and when the file name is shown .zsh the press return. Exit the nano editor.

To check that it works you can start a new terminal window and write:

echo $PATH

5.3. THE VDB-CONFIG COMMAND

When you use the SRA toolkit to collect the files, they will be dumped into your home folder. To change this behavior, run the following command:

Windows:
vdb-config.exe -interactive

"C:\Users\qlujani\Documents\SRA Windows toolkit\sratoolkit.2.10.9-win64\sratoolkit.2.10.9-win64\bin\vdb-config.exe" --interactive

macOS:
~/sratoolkit.2.10.9-mac64/bin/vdb-config -interactive

If vdb-config does not execute, you will need to make the file executable, like this:

% chmod a+x vdb-config
% ./vdb-config -interactive

You move between commands with the “Tab” key, and select the commands with a highlighted letter or “Enter”.

First the Cache needs to be set up, otherwise the results will be saved in a default folder .ncbi folder.

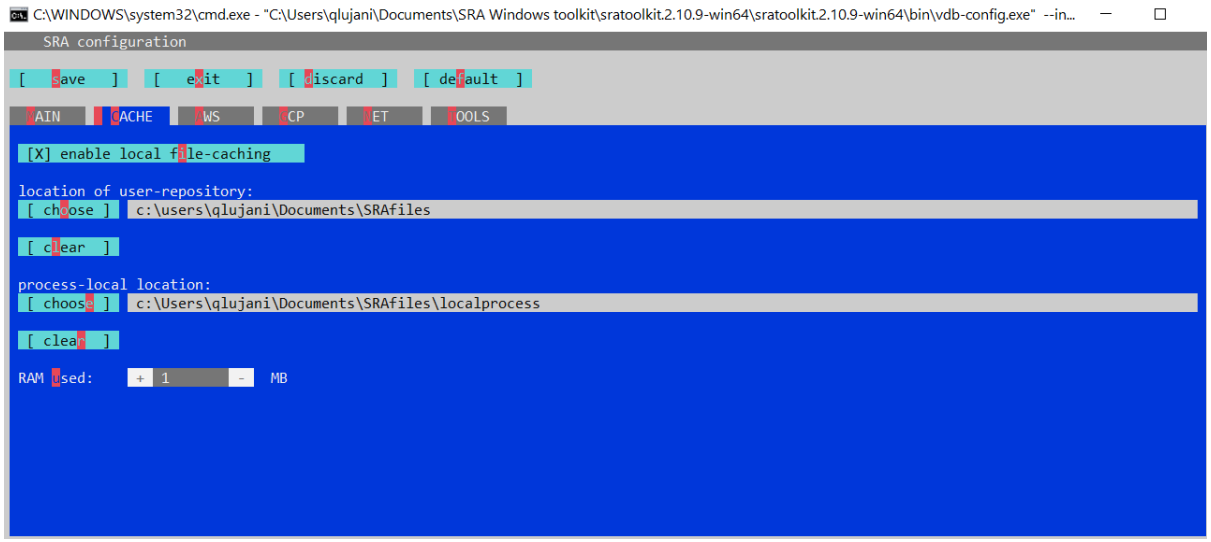


Figure: Screenshot of vdb-config

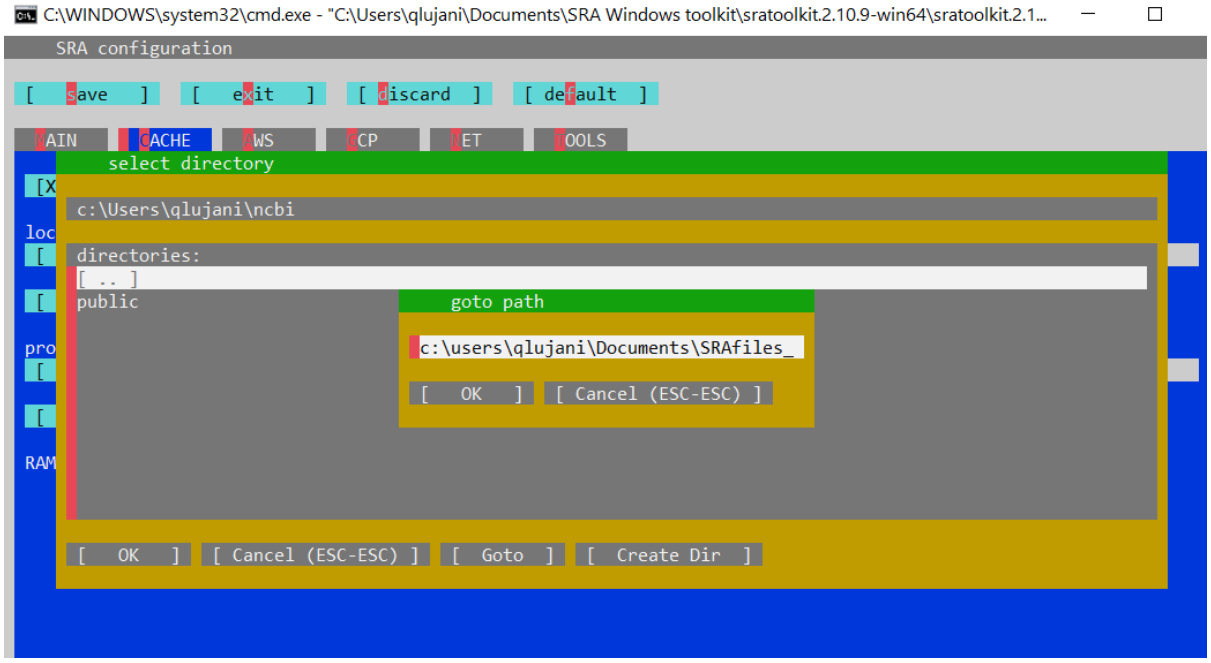


Figure: Screenshot of vdb-config

You can then use the graphical interface to alter the path under which the SRA files are stored by default (option number 5).

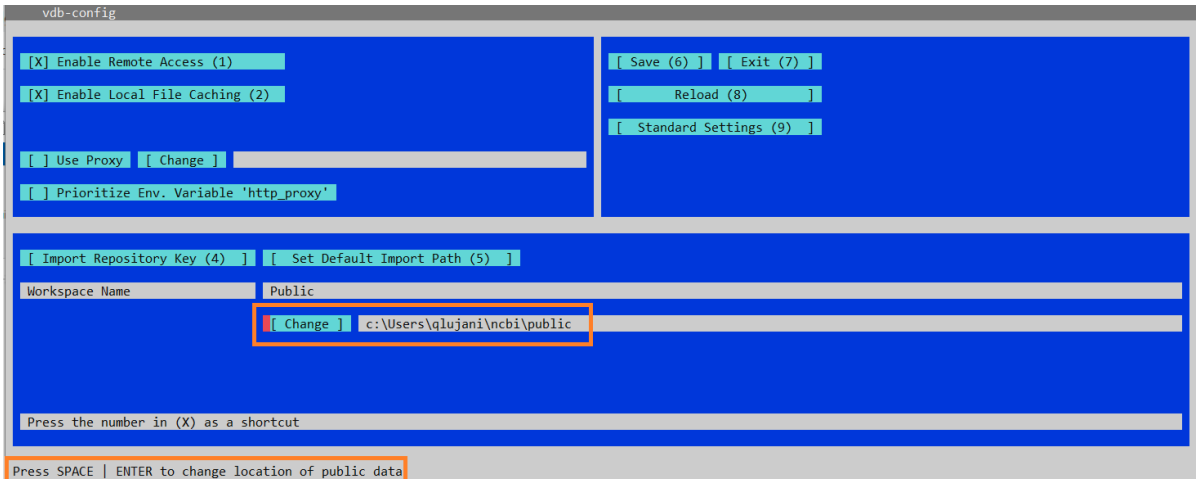


Figure: Screenshot of vdb-config

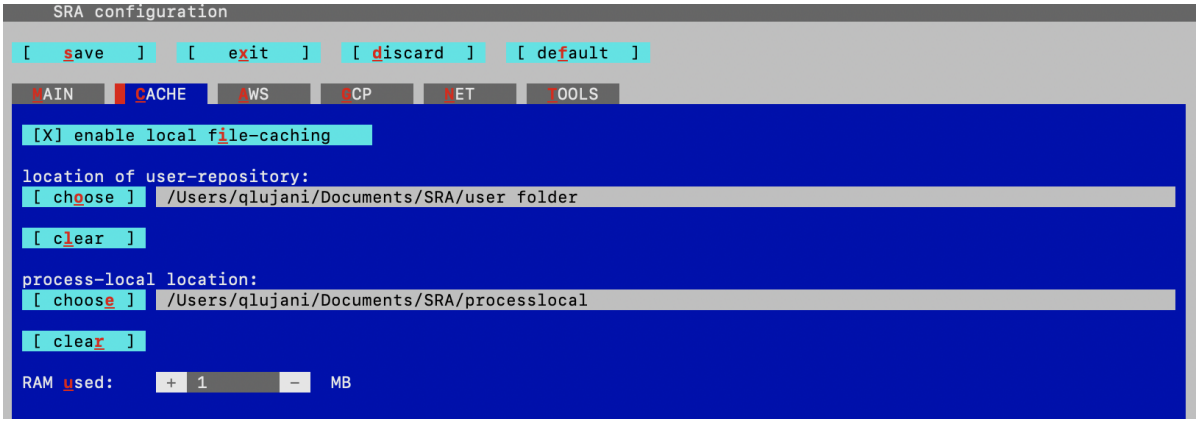


Figure: Screenshot of vdb-config

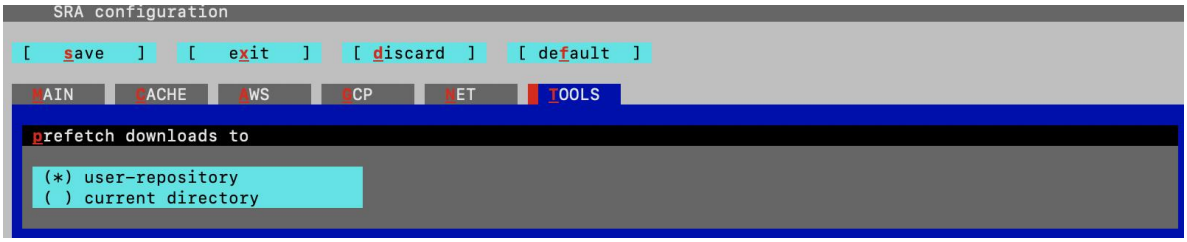


Figure: Screenshot of vdb-config

5.4. THE PREFETCH COMMAND IN THE SRA TOOLKIT

The SRA toolkit is available both as source files and pre-compiled versions at GitHub.

The **prefetch** command can be used to download data from the SRA. It can take the SraAccList.txt as input, and you can also specify the type of file that the SRA data shall be converted to, like a fastq file.

A command can look like this:
"prefetch –option-file SraAccList.txt –type fastq"
The command assumes that the file txt is in the folder from which you run the command, otherwise you need to specify the complete path to the file, like below:

Windows:
"C:\Users\qlujani\Documents\SRA Windows toolkit\sratoolkit.2.10.9-win64\sratoolkit.2.10.9-win64\bin\prefetch.exe" --option-file SraAccList.txt --type fastq

NOTE: With many large SRA files this may take considerable time. It is therefore can be a good idea to initiate the process and then let the computer work, perhaps overnight if you have a 100 files or so.

NOTE: that due communication errors etc, you may encounter files that are not successfully downloaded, so that the command must be run again, for individual files.

MacOS:

~/sratoolkit.2.10.9-mac64/bin/prefetch --option-file SraAccList.txt --type fastq

Figure: Screenshot of prefetch (macOS)

Corresponding on Windows:

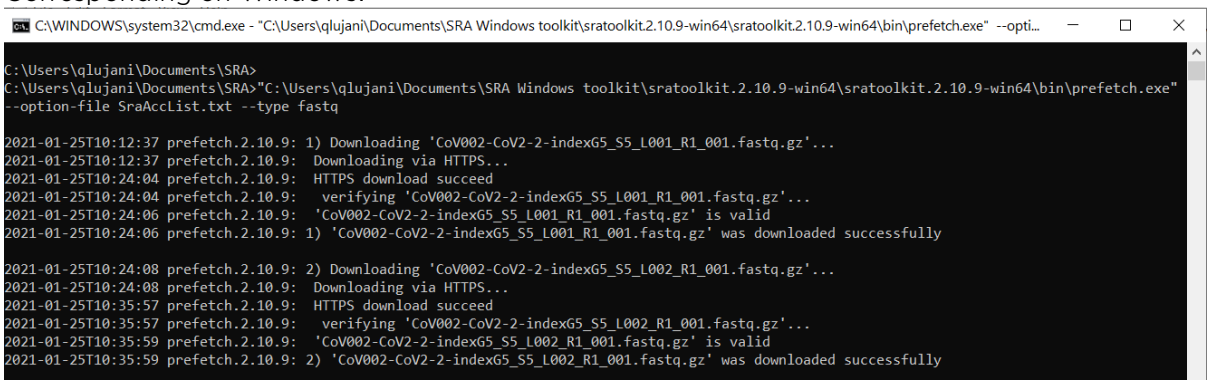


Figure: Screenshot of prefetch (Windows)

The fast-q files downloaded will be saved as gz (compressed files), one file per folder, in the folder you have specified in the vdb-config tool.

If you do not use the –fast-q option you will get the files as ".sra" files, whereafter you can use the commands fastq-dump and fasterq-dump, see in the next chapter.

5.5. THE FASTQ-DUMP COMMAND IN THE SRA TOOLKIT

If you have selected to download the files as sra files, you will get a folder with sra files, all ending with the suffix ".sra".

These files shall now be converted into fastq files. For this the command **fastq-dump.exe**, also from the SRA toolkit, can be used. NOTE: there is also a newer command, fasterq-dump, which will replace fastq-dump. The fasterq-dump is at this date not yet released on Windows.

The fastq-dump command takes several options, one is related to splits:

--split-files
Dump each read into separate file. Files will receive suffix corresponding to read number

--split-3
Legacy 3-file splitting for mate-pairs: First biological reads satisfying dumping conditions are placed in files *_1.fastq and *_2.fastq If only one biological read is present it is placed in *.fastq. Biological reads and above are ignored.

An example when you just want to use one SRA file can look like this:

Windows:
"C:\Users\qlujani\Documents\SRA Windows toolkit\sratoolkit.2.10.9-win64\sratoolkit.2.10.9-win64\bin\fastq-dump.exe" --split-files SRR11412215.sra

Then you may need to create a script to process several files in one go

Windows:
A script on Windows, can for an example be called myscript.bat, can look like this:

```
@echo off
for %%A in (*.sra) do (
echo Processing %%A....
"C:\Users\qlujani\Documents\SRA Windows toolkit\sratoolkit.2.10.9-win64\sratoolkit.2.10.9-win64\bin\fastq-dump.exe" --split-files %%A
)
@echo on
```

macOS:
A script on macOS, can for an example be save in a file called myscript, can look like this. If you save it in a file, remember to make it executable, with chmod a+x myscript

for FILE in *.sra; do ~/2.10.9sratoolkit.2.10.9-mac64/bin/fastq-dump --split-files ./$FILE ; done;

for FILE in *.gz ; do ~/sratoolkit.2.10.9-mac64/bin/fastq-dump --split-files ./$FILE ; done;

for DIR in SRR* ; do ~/Documents/STAR/sratoolkit.2.10.9-mac64/bin/fasterq-dump $DIR --split-files ; done ;

The you can run the file with ./myscript

An alternative, if you are used to run “R” is to run a script like this in “R”:

```
# Loop through all files and run fastq-dump
stopifnot( all(file.exists(files)) )
for(f in files) {
  cmd = paste("C:/Users/qlujani/Documents/SRA Windows toolkit/sratoolkit.2.10.9-win64/sratoolkit.2.10.9-win64/bin/fastq-dump.exe" --split-files ", f)
```

```
  cat(cmd,"\n") #print the current command
  system(cmd) # invoke command
}
```

NOTE: With many large SRA files the fastq-dump command will take considerable time.

6. USE AN ALIGNER TO CONVERT FASTQ FILES TO ALIGNED BAM FILES

Now that you have fastq files the next step is to process them to get aligned BAM files. To do so, you will need an aligner program.

There are many free aligners available.

Some examples are STAR, HISAT2, Bowtie2 and BWA (Burrows-Wheeler Aligner). The aligners are normally available as source code, some have pre-compiled libraries for some operating systems, normally Linux and in some cases macOS. Microsoft Windows support is rare, although some may have support for building the aligner on Windows.

Many aligners are open source software, see respective software for license terms.

In this document only one example of an aligner will be introduced, the STAR aligner (see chapter Usage, Acknowledgements for more info).

6.1. THE STAR ALIGNER

The STAR aligner is available at github.com, where you will find both source code that you can compile yourself, and for some platforms also binaries (i.e., executables that you can download and run directly). You will also find documentation etc. Note that the information is a snapshot, it may change.

STAR has only pre-compiled executables for (currently) Linux and macOS. This introduction will focus on the macOS platform. There are currently no pre-compiled executables available for Windows.

First visit https://github.com/alexdobin/STAR

Here you proceed to the folder bin/MacOSX_x86_64 and download the file that is listed as an "Executable file". Note that when you download on macOS using Safari, the file will get the extension STAR.dms . macOS does not know what a dms file is, it is simply an extension that is given at download.

Please remove the extension by renaming the file to STAR only. Then change the permission to an executable file.

- Start a terminal window
- Go to the folder where you have placed the file STAR
- Write the following command in the terminal window: chmod a+x STAR

Now STAR should be executable. You should be able to start it by writing (in a macOS terminal window):

./STAR

In the folder where STAR is located. You can also add the STAR folder to the PATH variable, see previous chapter. Now we also need to prepare a fasta file (reference genome file) that is needed when aligning the fastq files.

6.2. DOWNLOAD AND PREPARE A FASTA FILE

In this example we will work with fastq files from humas, and will need a fasta file for Homo

sapiens

Start by go to https://www.gencodegenes.org/human/

Here we will need to download the following

- Comprehensive gene annotation CHR - GTF
- Genome sequence, primary assembly (GRCh38) PRI – Fasta

Then run the following command (after you have changed the paths to the correct ones for you for the .gtf and the .fa-files for the reference genome:

./STAR --runThreadN 4 --runMode genomeGenerate --genomeDir GRCh38_index --genomeFastaFiles /path/to/GRCh38.primary_assembly.genome.fa --sjdbGTFfile /path/to/gencode.v34.annotation.gtf

This will create the folder GRCH38_index that we will need when we run STAR to align our fastq files later on.

6.3. RUN STAR ON A SINGLE SAMPLE AND ON A FOLDER WITH SAMPLES

Now that STAR has been downloaded and we have prepared the folder GRCH38_index we are ready to run a test. Note that the paths to the files may be different, here we just name them /path/to/ . Also note that you may have your fastq files in separate directories and may need to re-arrange them. You may also need to de-compress the gz files either before the command or as a part of the command line.

Run star on a single samle called SRR11517755_1.fastq:
./STAR --runThreadN 4 --genomeDir /path/to/GRCh38_index --readFilesIn /path/to/ SRR11517755_1.fastq

In some cases, you need to add the following two options to, may be needed on a computer with limited internal memory:
--genomeSAsparseD 3 --genomeSAindexNbases 12

If you now would like to have BAM files sorted by as output, you will add these two options:
--outSAMtype BAM SortedByCoordinate

If you would like to add a prefix o the output file (BAM file) you can use this option:
--outFileNamePrefix yourprefix

Now the command will then be:
./STAR --runThreadN 4 --genomeSAsparseD 3 --genomeSAindexNbases 12 --genomeDir ./GRCh38_index --readFilesIn ./SRR11517755_1.fastq --outSAMtype BAM SortedByCoordinate --outFileNamePrefix SRR11517755_1_

Often you have many files and then you can write a script which you can past in the terminal window to run, like this one on macOS in a terminal window:

for FILE in *.fastq; do ./STAR --runThreadN 4 --genomeSAsparseD 3 --genomeSAindexNbases 12 --genomeDir ./GRCh38_index --readFilesIn ./$FILE --outSAMtype BAM SortedByCoordinate --outFileNamePrefix $FILE ; done;

The script will for all files ending with the suffix .fastq in the current folder execute STAR and produce a BAM file. Note that you may need to adjust the folder to the one you have. For each fastq file you will get these files as output (here the file SRR11517755_1 is the input):

SRR11517755_1.fastqSJ.out.tab
SRR11517755_1.fastqLog.progress.out

SRR11517755_1.fastqLog.out
SRR11517755_1.fastqLog.final.out
SRR11517755_1.fastqAligned.sortedByCoord.out.bam

The file SRR11517755_1.fastqAligned.sortedByCoord.out.bam is the aligned bam file sorted by coordinate.

```
qlujani@Q00029 SRA % for FILE in *.fastq; do ./STAR --runThreadN 4 --genomeSAsparseD 3 --genomeSAindexNbases
 12 --genomeDir ./GRCh38_index --readFilesIn ./$FILE --outSAMtype BAM SortedByCoordinate --outFileNamePrefix
 $FILE ; done;
Jan 15 12:34:10 ..... started STAR run
Jan 15 12:34:10 ..... loading genome
Jan 15 12:37:41 ..... started mapping
Jan 15 12:44:43 ..... finished mapping
Jan 15 12:44:45 ..... started sorting BAM
Jan 15 12:44:50 ..... finished successfully
Jan 15 12:44:51 ..... started STAR run
Jan 15 12:44:51 ..... loading genome
Jan 15 12:45:40 ..... started mapping
Jan 15 12:48:48 ..... finished mapping
Jan 15 12:48:49 ..... started sorting BAM
Jan 15 12:49:01 ..... finished successfully
Jan 15 12:49:01 ..... started STAR run
Jan 15 12:49:01 ..... loading genome
Jan 15 12:49:30 ..... started mapping
Jan 15 12:51:26 ..... finished mapping
Jan 15 12:51:27 ..... started sorting BAM
Jan 15 12:51:35 ..... finished successfully
Jan 15 12:51:36 ..... started STAR run
Jan 15 12:51:36 ..... loading genome
Jan 15 12:52:05 ..... started mapping
Jan 15 12:55:09 ..... finished mapping
Jan 15 12:55:10 ..... started sorting BAM
Jan 15 12:55:21 ..... finished successfully
```

Figure: Screenshot of output from the STAR aligner

Note: When you have processed all your files this way (this will take time) using the script that takes all fastq files in the specified folder, then you will have a your bam files.

```
qlujani@Q00029 SRA % ls *.bam
CoV002-CoV2-1-indexG4_S4_L001_R1_001.fastqAligned.sortedByCoord.out.bam
CoV002-CoV2-1-indexG4_S4_L002_R1_001.fastqAligned.sortedByCoord.out.bam
CoV002-CoV2-1-indexG4_S4_L003_R1_001.fastqAligned.sortedByCoord.out.bam
CoV002-CoV2-2-indexG5_S5_L001_R1_001.fastqAligned.sortedByCoord.out.bam
```

Figure: Screenshot of aligned and sorted BAM files

7. IMPORTING ALIGNED BAM FILES INTO OMICS EXPLORER

Now it´s time to import the BAM files into Omics Explorer. To do so, you will need the corresponding gtf file. One site where you can find the information is here:

https://www.gencodegenes.org/pages/data_access.html .

One way to download the files is to use ftp, then select "Download from the FTP site" (ftp://ftp.ensembl.org/pub) where you then can navigate to the correct gtf file to use. In this case the latest gtf for humans, at ftp://ftp.ensembl.org/pub/current_gtf/homo_sapiens/ , where Homo_sapiens.GRCh38.102.gtf.gz is available.

Now you can start Qlucore Omics Explorer. Select "File" and the "Open BAM Files…":

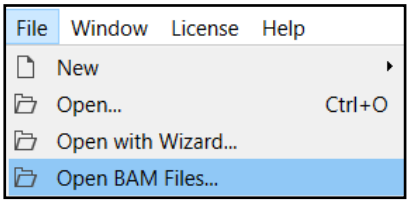


Figure: Screenshot of File->Open BAM Files in Qlucore Omics Explorer

You can now specify the gtf file to be used and the BAM files (or a folder, where all BAM files will then be selected.

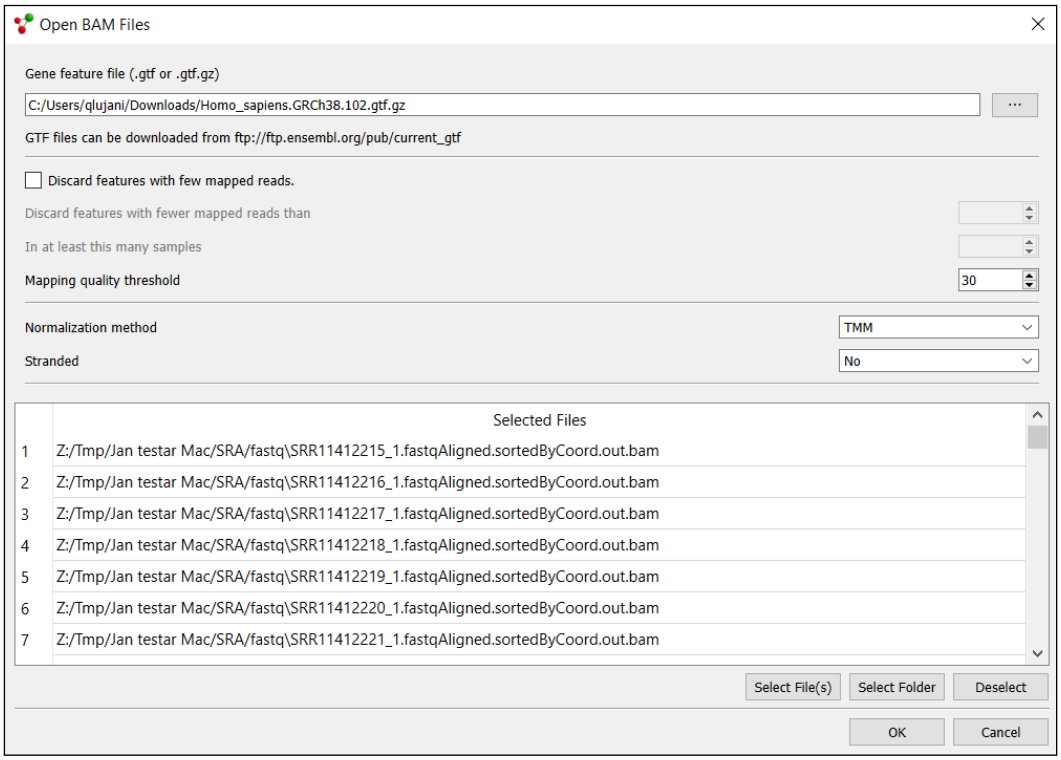


Figure: Screenshot of Open BAM Files in Qlucore Omics Explorer

You can also select to discard features and to set a Threshold value. The normalization method can be selected, and information about if a strand-specific protocol was used.

When you press OK the import process starts. NOTE: With many BAM large files this may take considerable time. It is therefore can be a good idea to initiate the process and then let the computer work, perhaps overnight if you have a 100 BAM files or so.

When all BAM files have been imported, you will get a quality window pop-up.

Remember to go to “File” and the “Save As…” to save the result in a gedata file.

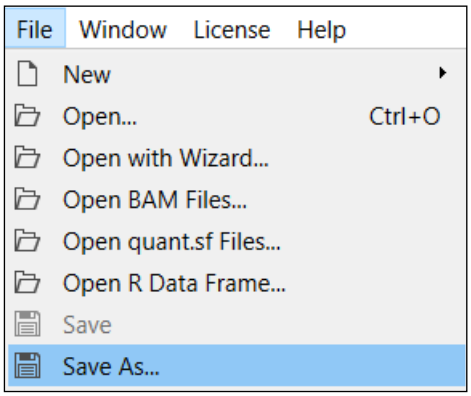


Figure: Screenshot of "Save As..." in Qlucore Omics Explorer

The next time you just select "File" and the "Open..." to open the gedata file, a process that just takes a few seconds.

8. CHANGE SAMPLE IDENTIFIER AND ADD ANNOTATIONS

After BAM file import the sample identifier is normally the full path BAM file path. This is probably inconvenient, and you would like to change this.

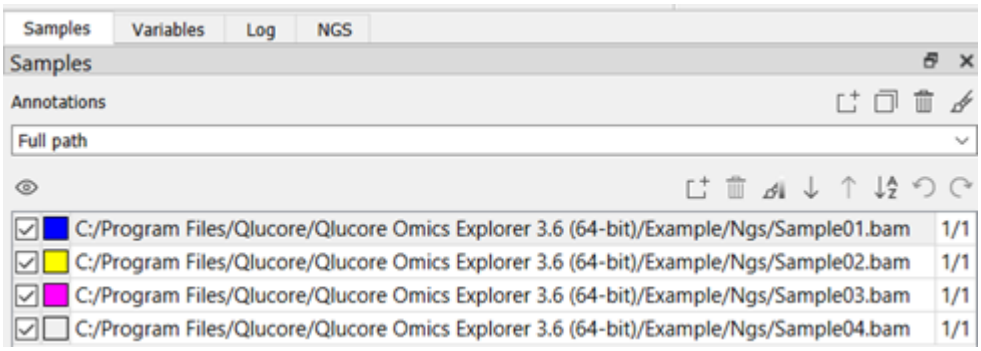


Figure: Screenshot of the unique identifier, here the Full path name

You can change to the shorter filename in the Data tab:

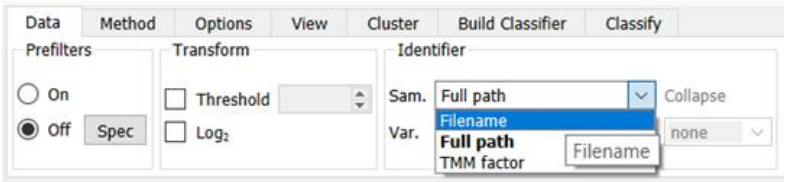


Figure: Screenshot of the Data tab, changing the Sample Identifier to Filename

After that, you save the dataset with the new identifier, with "File->Export->Data".

Figure: Screenshot of "File->Export->Data".

The BAM files do not have any annotations, but you can add that now to the dataset.

Create a file with the unique sample identifier and the annotations in Excel or another tool and

save as a tab separated txt file.

Filename	Age	Gender	Treatment	Rank	Censor
File1	20	Female	Drug 2	Very low	1
File2	26	Female	Placebo	Very low	1
File3	28	Male	Drug 2	Low	1
File4	30	Male	Drug 1	Low	1
File5	40	Male	Drug 1	Medium	1
File6	40	Male	Placebo	Medium	0
File7	43	Female	Drug 1	Medium	1
File8	48	Male	Placebo	High	1

Figure: Screenshot of a matrix with annotations

Now you can add the annotation file to the dataset with "Import->Sample Annotations...".

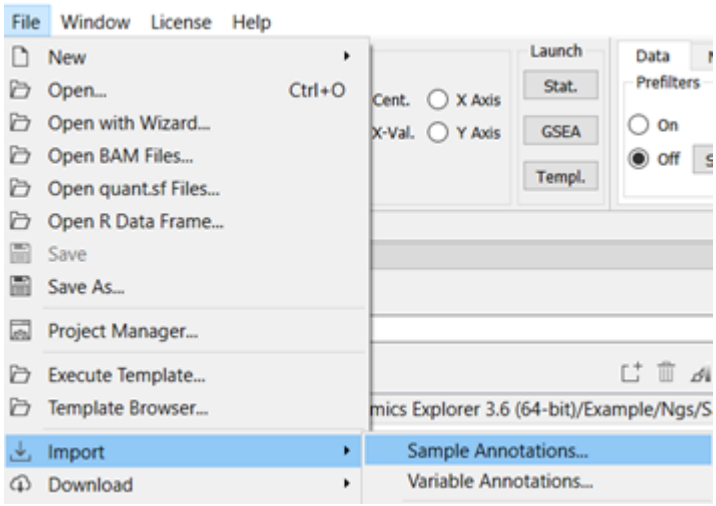


Figure: Screenshot of "Import->Sample Annotations..."

Then save the dataset again, with "File->Save As...".

9. USAGE, ACKNOWLEDGEMENTS ETC

Note that it is required that you accept the Licenses, Data Use Policy and Publication Guidelines at NCBI and other sources referenced herein in order to be able to use their publicly available software, information and data.

For information about the STAR aligner, please see: A. Dobin et al, Bioinformatics 2012; doi: 10.1093/bioinformatics/bts635

10. DISCLAIMER

The contents of this document are subject to revision without notice due to continuous progress in methodology, design, and manufacturing.
Qlucore shall have no liability for any error or damages of any kind resulting from the use of this document.
Qlucore Omics Explorer is only intended for research purposes.