How to use NCBI SRA Toolkit effectively?

Renesh Bedre 5 minute read

What is NCBI Sequence Read Archive (SRA) Toolkit?

- NCBI SRA toolkit is a set of utilities to download, view and search large volume of high-throughput sequencing data from NCBI SRA database at faster speed
- SRA database has several accessions including, SRR (run accession for actual sequencing data for the particular experiment), SRX (experiment accession representing the metadata for study, sample, library, and runs), SRP (study accession representing the metadata for sequencing study and project abstract), SAMN/SRS (BioSample/SRA accession representing the metadata for biological sample).

Applications

- Effectively download the large volume of high-throughput sequencing data (eg. FASTQ, SAM)
- Convert SRA file into other biological file format (eg. FASTA, ABI, SAM, QSEQ, SFF)
- Retrieve a small subset of large files (e.g. sequences, alignment)
- Search within SRA files and fetch specific sequences

Download and install NCBI SRA toolkit

```
# I am using Ubuntu Linux 20.04.1 LTS
# download latest version of compiled binaries of NCBI SRA toolkit
# (June 29, 2020, version 2.10.8) for Ubuntu Linux
# Compiled binaries for other OS visit: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software
$ wget https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.10.8/sratoolkit.2.10.8-ubuntu64.tar.gz

# extract tar.gz file
$ tar -zxvf sratoolkit.2.10.8-ubuntu64.tar.gz

# add binaries to path using export path or editing ~/.bashrc file
$ export PATH=$PATH:/home/ren/software/sratoolkit.2.10.8-ubuntu64/bin
# Now SRA binaries added to path and ready to use
```

Download SRA datasets using NCBI SRA toolkit

> **Note:** Current SRA toolkit version 2.10.8 does not support Aspera client (ascp). Even though </i>ascp</i> can run with older versions, it will download the data by *https* mode and not by *FASP* mode.

```
# download file: prefetch will download and save SRA file related to SRR accession in
# the current directory under newly created SRA accession directory
$ prefetch  SRR5790106  # for a single file
$ prefetch  SRR5790106 SRR5790104  # multiple files

# convert to FASTQ: fastq-dump will convert SRR5790106.sra to SRR5790106.fastq
$ fastq-dump  SRR5790106  # single file
$ fastq-dump  SRR5790106  SRR5790104 # multiple files

# now you can also replace fastq-dump with fasterq-dump which is much faster
# and efficient for large datasets
# by default it will use 6 threads (-e option)
$ fasterq-dump  SRR5790106  # single file
$ fasterq-dump  SRR5790106  SRR5790104 # multiple files

# for paired-end data use --split-files (fastq-dump) and -S or --split-files (fasterq-dump) option
$ fastq-dump --split-files SRR8296149
$ fasterq-dump -S SRR8296149

# download alignment files (SAM)
# make sure corresponding accession has alignment file at SRA database
$ sam-dump --output-file SRR1236468.sam SRR1236468
```

Note: With *fastq-dump* and *fasterq-dump*, *prefetch* step is unncessary and you can directly download sequence data in FASTQ format

Batch download SRA datasets

- Sometimes, we need to download hundreds or thousands of FASTQ files from the SRA database and it would be inconvenient to directly use the SRA toolkit for batch download
- I have added a wrapper script for `fasterq-dump` in `bioinfokit` (v0.9.7 or later) for easy download of a large number of FASTQ files from the SRA database
- Check bioinfokit documentation (https://github.com/reneshbedre/bioinfokit) for installation and documentation
- Download test SRA accession file containing accessions for both single and paired-end FASTQ datasets

```
# tested on Linux and Mac. It may not work on Windows
>>> from bioinfokit.analys import fastq

# batch download fastq files
# make sure you have installed the latest version of NCBI SRA toolkit (version 2.10.8) and added
binaries in the
# system path
>>> fastq.sra_bd(file='sra_accessions.txt')

# increase number of threads
>>> fastq.sra_bd(file='sra_accessions.txt', t=16)

# use fasterq-dump customized options, you can see more options for fas terq-dump as
# fasterq-dump -help
fastq.sra_bd(file='sra_accessions.txt', t=16, other_opts='--outdir temp --skip-technical')

# multiple FASTQ (technical and biological)  files from from
# 10x chromium single cell 3' RNA-seq data
# if you provide file containing SRA accessions for 10x chromium
# single cell 3' RNA-seq data, it will give multiple FASTQ files
# for example, SRA accession SRR12564282 will give  three FASTQ files
# (sample barcode,  cell barcode, and biological read FASTQ files)
fastq.sra_bd(file='path_to_sra_file', t=16, other_opts='--include-technical --split-files')
```

Validation of downloaded SRA data integrity

It is essential to check the integrity and checksum of SRA datasets to ensure successful download

```
# download FASTQ file
$ prefetch SRR5790104
# fastq-dump  SRR5790104

# check integrity of downloaded SRR5790106.fastq file
# output from vdb-validate should report 'ok' and 'consistent' for all parameters
# Note: make sure you have .sra (not .cache) file for corresponding accession in
# sra accession directory
$ vdb-validate SRR5790104
2020-08-31T22:46:27 vdb-validate.2.10.8 info: Table 'SRR5790104.sra' metadata: md5 ok
2020-08-31T22:46:27 vdb-validate.2.10.8 info: Column 'ALTREAD': checksums ok
2020-08-31T22:46:29 vdb-validate.2.10.8 info: Column 'QUALITY': checksums ok
2020-08-31T22:46:30 vdb-validate.2.10.8 info: Column 'READ': checksums ok
2020-08-31T22:46:30 vdb-validate.2.10.8 info: Column 'X': checksums ok
2020-08-31T22:46:30 vdb-validate.2.10.8 info: Column 'Y': checksums ok
2020-08-31T22:46:30 vdb-validate.2.10.8 info: Table 'SRR5790104.sra' is consistent
```

Customized download of SRA datasets

You can use SRA tools for customized output of large SRA datasets without downloading complete datasets (NOTE: some options are not available in `fasterq-dump`)

```
# print first 10 reads from single-end FASTQ file
# -Z option will print output on screen (STDOUT)
$ fastq-dump -X 10 -Z SRR5790106

# save FASTQ file to specified directory
$ fastq-dump -O temp SRR5790106
$ fasterq-dump -O temp SRR5790106

# compress FASTQ file gzip or bzip2
$ fastq-dump -O temp SRR5790106
$ fastq-dump --gzip SRR5790106
$ fastq-dump --bzip2 SRR5790106
# Note: --gzip or --bzip2 options are not available with fasterq-dump

# Multithreading
$ fasterq-dump -e 10 SRR5790106
```

Convert SRA data into other biological formats

SRA tools allow you to convert SRA files into FASTA, ABI, Illumina native (QSEQ), and SFF format

```
# convert to FASTA
# you need to first download the FASTQ file to convert to FASTA file
$ fastq-dump --fasta 60 SRR5790106
# if you have paired-end FASTQ, use --split-files -fasta 60
# if you don't use --split-files for paired-ends, the reads will be merged from both ends
# number 60 represents number of bases per line
# Note: --fasta options is not available with fasterq-dump

# convert to ABI (CSFASTA and QVAL)
$ abi-dump  SRR5790106

# convert to QSEQ
# SRA database should have alignment information submitted for corresponding accession
$ illumina-dump --qseq 2 SRR1236472 # 2 for paired-end and 1 for single-end

# convert to SFF
# SFF is a binary file format related to 454 high-throughput sequencing
$ sff-dump SRR996630
```

Search within SRA files

You can search specific sequences or subset of sequences in SRA files

```
# search within SRA files
# output will be sequence read IDs
$ sra-search  GATGCCGCGCC SRR5790104
```

NOTE: For every SRA tools, you can check all options by providing `-h` parameter (eg. `fasterq-dump -h`)

References

- Understanding SRA Search Results (https://www.ncbi.nlm.nih.gov/books/NBK56913/)

If you have any questions, comments or recommendations, please email me at **reneshbe@gmail.com**

Tags: Bioinformatics

Updated: May 30, 2021