Downloading SRA data using the SRA toolkit

James Thornton

Abstract

The Sequence Read Archive (SRA) is a database for biological sequence data and is maintained by the National Center for Biotechnology Information (NCBI). Sequence files can be obtained by using the SRA Toolkit. This protocol provides the steps necessary to use the SRA toolkit to get sequence data in fastq format.

Citation: James Thornton Downloading SRA data using the SRA toolkit. **protocols.io**
dx.doi.org/10.17504/protocols.io.frsbm6e
Published: 12 Sep 2016

Guidelines

SRA Toolkit Documentation

Before start

Login to the UA hpc. This protocol will begin in your home directory.

Protocol

Step 1.

Make sure you have /rsgrps/bh_class/bin in your path:

If you don't, you will get an error message like this:

2016-09-12T17:15:17 prefetch.2.4.4 int: path not found while resolving tree - cannot get cache location for SRR1647046

cmd COMMAND

```
$ cd
$ nano .bashrc

export PATH=/rsgrps/bh_class/bin:$PATH
```

```
$ source .bashrc
```
export PATH=/rsgrps/bh_class/bin:$PATH is copied into .bashrc. Then save and quit nano to source it.

NOTES
James Thornton Jr 07 Sep 2016

This step allows you to execute the executable files found in /rsgrps/bh_class/bin. Executable files appear green on the HPC.

Step 2.

Utilize the "prefetch" command from the SRA toolkit to get your SRA file.

cmd COMMAND
```
$ prefetch SRR1647145
```
NOTE: make sure to use your SRR number

NOTES
James Thornton Jr 08 Sep 2016

Your SRR numbers are found in the google drive sheet shared with the class under column 'M' . There should be a total of 8 files you need to download. See next step on how to download multiple files at once.

Step 3.

You can pass 'prefetch' multiple arguments to download all data files at once:

cmd COMMAND
```
$ prefetch SRR1647238 SRR1647240 SRR1647144 SRR1647260 SRR1647239 SRR1647236 SRR1647237
```
NOTE: make sure you use your SRR numbers .

ANNOTATIONS
Bonnie Hurwitz 12 Sep 2016

Rather than copying and pasting each file name, you can use Unix to help you!

make a file with the list of SRR files copied from the excel spread sheet

% nano list

use the translate command to convert new lines to spaces. Note the space in the second set of quotes.

% tr '\n' ' ' < list

then copy the line with the file names separated by space into the prefetch command, as below.

Step 4.

The .sra files will be stored in /ncbi/public/sra

Move into that directory, then make sure all 8 files are present:

cmd COMMAND

```
$ cd ~/ncbi/public/sra
$ ls
```

EXPECTED RESULTS

SRR390728.sra SRR1647238.sra SRR1647240.sra SRR1647144.sra SRR1647260.sra
SRR1647239.sra SRR1647236.sra SRR1647237.sra

Step 5.

Convert the .sra file into fastq format using the fastq-dump command from the SRA toolkit. All files can be converted in one command by passing fastq-dump all files with the .sra extension.

cmd COMMAND

```
$ fastq-dump *.sra
```

*.sra defines all files with a .sra extension NOTE: make sure you are in ~/ncbi/public/sra when you execute this command.

EXPECTED RESULTS

Read 2533849 spots for SRR1647144.sra
Written 2533849 spots for SRR1647144.sra
Read 3649566 spots for SRR1647145.sra
Written 3649566 spots for SRR1647145.sra
Read 3051288 spots for SRR1647236.sra
Written 3051288 spots for SRR1647236.sra
Read 1856522 spots for SRR1647237.sra
Written 1856522 spots for SRR1647237.sra
Read 492203 spots for SRR1647238.sra
Written 492203 spots for SRR1647238.sra
Read 1191553 spots for SRR1647239.sra
Written 1191553 spots for SRR1647239.sra
Read 1527542 spots for SRR1647240.sra
Written 1527542 spots for SRR1647240.sra
Read 39872 spots for SRR1647260.sra
Written 39872 spots for SRR1647260.sra
Read 14342395 spots total
Written 14342395 spots total

Step 6.

Now check to see you have 8 .fastq files, 1 for each .sra file. Make a /rsgrps/bh_class/<user>/fastq directory. Where you will replace <user> with your github id. Then move all of the all fastq files there for later use.

cmd COMMAND

```
ls
mkdir -p /rsgrps/bh_class/bhurwitz/fastq
mv *fastq !$
cd !$
ls
```

use mkdir -p to create all directories listed. In this case, we are creating bhurwitz (my user id) and the fastq directories. Note that you should use your github id here, so we can track your user id easily, and so it is consistent with your homework. Note that I am using !$ to use the argument from the last command line.

EXPECTED RESULTS

SRR390728.fastq SRR1647238.fastq SRR1647240.fastq SRR1647144.fastq SRR1647260.fastq SRR1647239.fastq SRR1647236.fastq SRR1647237.fastq