Using SRA toolkit to download sequencing data for CellRanger

This tutorial introduces how to download sequencing data from NCBI using SRA Toolkit and obtain the UMI count matrix.

Install SRA Toolkit

- Method 1. Download the binary files of SRA Toolkit from Github https://github.com/ncbi/sra-tools/wiki/01.-Downloading-SRA-Toolkit
- Method 2. If you use SeaDragon, load SRA Toolkit using `module load sratoolkit`

Download sequencing data using SRA Toolkit

For example, we want to download the sequencing data from NCBI project `PRJNA725335` https://www.ncbi.nlm.nih.gov/Traces/study/?acc=PRJNA725335&o=acc_s%3Aa.

The accession numbers are:

SRR14710616
SRR14710618
SRR14710619
SRR14710620
SRR14710621
SRR14710622
SRR14710623
SRR14710624
SRR14710625
SRR14710626
SRR14710627
SRR14710628
SRR14710629
SRR14710617

For each accession number, use the `fastq-dump` command from SRA Toolkit with the `–split-files`, `–origfmt`, `–gzip` arguments to retrieve the FASTQ files:

```
fastq-dump --split-files --origfmt --gzip SRR14710616
```

The output would be two FASTQ files:

```
SRR14710616_1.fastq.gz

SRR14710616_2.fastq.gz
```

Sometimes, the authors may upload 3 FASTQ files for a sample by including index reads. For example, if we try to retrieve FASTQ files from `SRR9291388`: The output would be three FASTQ files:

```
SRR9291388_1.fastq.gz

SRR9291388_2.fastq.gz

SRR9291388_3.fastq.gz
```

If there are three FASTQ files generated, we need to determine which one is the index file. The size of the index file is generally much smaller than the reads FASTQ files. For the above one, we can have

```
SRR9291388_1.fastq is Read 1

SRR9291388_2.fastq is Read 2

SRR9291388_3.fastq is Index 1
```

Cell Ranger requires FASTQ file names to follow the `bcl2fastq` file naming convention.

```
[Sample Name]_S1_L00[Lane Number]_[Read Type]_001.fastq.gz
```

Where Read Type is one of:

```
I1: Sample index read (optional)

I2: Sample index read (optional)

R1: Read 1 (required)

R2: Read 2 (required)
```

Therefore, we need to change the FASTQ file names for Cell Ranger:

```
SRR14710616_1.fastq.gz to SRR14710616_S1_L001_R1_001.fastq.gz

SRR14710616_2.fastq.gz to SRR14710616_S1_L001_R2_001.fastq.gz
```

Lastly, run CellRanger by

```
cellranger count --id=SRR14710616 --fastqs=PATH_TO_SRR14710616 --sample=SRR1471061
6 --transcriptome=CellRanger_Reference_Genome(e.g., refdata-gex-GRCh38-2020-A) --c
hemistry=threeprime
```

Note: –chemistry argument, threeprime or fiveprime can be found in the related paper or the project page in NCBI website.

Please visit the CellRanger website for detailed instructions about how to run CellRanger.