NGS3_DNA-SEQ: Preparation before class

Author: Duy Dao

Date: May 2nd, 2023.

~

Hi guys, welcome to the new chapter of our class - Module II: DNAseq. In this Module, we will work with real-world NGS data. But before that, please prepare some data and tools needed for the lecture DNAseq: Upstream Analysis (2023, May 7th) by following the below guidelines.

And make sure that you have at least 50 GB of free disk space for this.

Have fun with the installation!

Setup working directory

Create directories and organize our workplace. We will do all the work here.

```
mkdir dnaseq_work/

cd dnaseq_work/
mkdir tools/ work/

cd work/
mkdir 1_raw/ 2_trim/ 3_align/
mkdir ref_genome

mkdir -p 1_raw/sample1/ 1_raw/sample2
mkdir -p 2_trim/sample1/ 2_trim/sample2/
mkdir -p 3_align/sample1/ 3_align/sample2/
```

Install "tree" tools to view better:

```
sudo apt install tree
```

And here is what our working directory looks like:

```
tree dnaseq_work/
#
dnaseq_work/
├── tools
└── work
    ├── 1_raw
    │   ├── sample1
    │   └── sample2
    ├── 2_trim
    │   ├── sample1
    │   └── sample2
    ├── 3_align
    │   ├── sample1
    │   └── sample2
    └── ref_genome
```

INSTALL TOOLS

Note: All tools will be installed at the tools/ dir

List of tools

Tools	Description	file format	Preferences
FastQC	Sequencing quality control	fastq	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Trimmomatic	Useful trimming tasks for illumina paired-end and single ended data.	fastq	https://github.com/usadellab/Trimmomatic
BWA mem	Mapped reads to reference genome	sam/bam	https://bio-bwa.sourceforge.net/bwa.shtml
GATK	A genomic analysis toolkit focused on variant discovery.		https://gatk.broadinstitute.org/hc/en-us
samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format	sam/bam/cram	http://www.htslib.org/

FastQC

This tool use to check your reads quality.

Download: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

```
#Install & unzip
wget https://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.12.1.zip

unzip fastqc_v0.12.1.zip

# Write at the end of the .bashrc file. This command will let you export the path to FastQC, and
execute it everywhere
nano ~/.bashrc

export PATH='path/to/FastQC/':$PATH
# For example:
export PATH='/home/duydao/dnaseq_work/tools/FastQC':$PATH

source ~/.bashrc

#Try it by running
fastqc
```

Your .bashrc will look like this.

And when you type this command. The tools become executable.

```
source ~/.bashrc
```

Do this for the rest.

Trimmomatic

Useful trimming tasks for illumina paired-end and single ended data.

Download: https://anaconda.org/bioconda/trimmomatic/0.39/download/noarch/trimmomatic-0.39-hdfd78af_2.tar.bz2

```
# Move the file to tools/ and then unzip
mkdir trimmomatic
tar -xvf trimmomatic-0.39-hdfd78af_2.tar.bz2 -C trimmomatic/
#
nano ~/.bashrc
#
export PATH='/home/duydao/dnaseq_work/tools/trimmomatic/bin/':$PATH
#
source ~/.bashrc
```

BWA

Used to map your reads to the reference genome.

If you don't have "git", please install it by:

```
sudo apt install git-all
```

```
git clone https://github.com/lh3/bwa.git

cd bwa/

make

# export to PATH
nano ~/.bashrc

export PATH='/home/duydao/dnaseq_work/tools/bwa/':$PATH #Write at the end of the file.

source ~/.bashrc

# done.
```

GATK

A genomic analysis toolkit focused on variant discovery, include lots of useful tools.

```
Note: This tool quite heavy and took a lot of diskspaces (15 G.B)
```

https://github.com/broadinstitute/gatk/releases

```
git clone https://github.com/broadinstitute/gatk.git

# Install git-lfs
wget https://github.com/git-lfs/git-lfs/releases/download/v3.3.0/git-lfs-linux-amd64-v3.3.0.tar.gz

tar -xvf git-lfs-linux-amd64-v3.3.0.tar.gz

# Set up java version 17
## Download java 17
sudo apt install openjdk-17-jdk

## Switch to java 17
sudo update-alternatives --config java
## Type selection number 1,2,3,... that correspond to Java 17 and press Enter.
  Selection    Path                                            Priority   Status
```

```
-----------------------------------------------------------
  0            /usr/lib/jvm/java-19-openjdk-amd64/bin/java   1911      auto mode
* 1            /usr/lib/jvm/java-17-openjdk-amd64/bin/java   1711      manual mode
  2            /usr/lib/jvm/java-18-openjdk-amd64/bin/java   1811      manual mode
  3            /usr/lib/jvm/java-19-openjdk-amd64/bin/java   1911      manual mode

--> 1


# Build gatk
cd gatk/
sudo ./gradlew bundle # This may take a while.

# export to PATH
nano ~/.bashrc
#
export PATH='/home/duydao/dnaseq_work/tools/gatk/':$PATH #Write at the end of the file.
#
source ~/.bashrc
# done.
```

Samtools

Use for Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format

```
wget https://github.com/samtools/samtools/releases/download/1.17/samtools-1.17.tar.bz2

tar -xvf samtools-1.17.tar.bz2

cd samtools-1.17/

./configure

make

make install

# export to PATH
nano ~/.bashrc
#
export PATH='/home/duydao/dnaseq_work/tools/samtools-1.17/':$PATH #Write at the end of the file.
#
source ~/.bashrc
# done.
```

Tabix

```
wget https://github.com/samtools/htslib/releases/download/1.17/htslib-1.17.tar.bz2

tar -xvf htslib-1.17.tar.bz2

cd htslib-1.17/
make

# export to PATH
nano ~/.bashrc
#
export PATH='/home/duydao/dnaseq_work/tools/htslib-1.17/':$PATH #Write at the end of the file.
#
source ~/.bashrc
# done.
```

When complete installing all the tools, our tools/ dir will look like this

```
tree -L 1 tools

#
tools
├── bwa
├── FastQC
├── gatk
├── git-lfs-3.3.0
├── htslib-1.17
├── samtools-1.17
└── trimmomatic
```

~~

DOWNLOAD DATA: Raw fastq & Reference genome data

Download raw fastq:

For the scope of this lecture, we will working on 2 samples

```
cd 1_raw
mkdir sample1/ sample2/
```

Sample1

Download: https://zenodo.org/record/3531578/files/LowQuality_Reads.fastq.gz

```
cd sample1/
wget https://zenodo.org/record/3531578/files/LowQuality_Reads.fastq.gz
LowQuality_Reads.fastq.gz
```

Sample2

Download link:
https://ftp.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R1_001.fastq.gz

https://ftp.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R2_001.fastq.gz

Or use command lines:

```
cd sample2/

sudo apt install lftp

lftp -e "pget -n 20 ftp://ftp-
trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R1_
001.fastq.gz; bye"

lftp -e "pget -n 20 ftp://ftp-
trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R2_
001.fastq.gz; bye"
```

Reference genome

Download link: https://hgdownload-test.gi.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/hg38.fullAnalysisSet.chroms.tar.gz

Or use command lines:

```
cd ref_genome/
```

```
wget https://hgdownload-
test.gi.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/hg38.fullAnalysisSet.chroms.tar.gz

tar xvzf hg38.fullAnalysisSet.chroms.tar.gz

cd hg38.fullAnalysisSet.chroms
cat *.fa > hs38DH.fa

cd ref_genome/

mkdir hg38/

mv hs38DH.fa ../hg38

# We just need hs38DH.fa. All the unnessesary files could be removed.
rm -rf  hg38.fullAnalysisSet.chroms
```

Index the reference genome (take a long time)

```
cd hg38/

bwa index -a bwtsw hs38DH.fa
```

COMPLETED

```
# This is our working dir looks like when completed.
tree -L 2 dnaseq_work/
#
dnaseq_work/
├── tools
│   ├── bwa
│   ├── FastQC
│   ├── gatk
│   ├── git-lfs-3.3.0
│   ├── htslib-1.17
│   ├── samtools-1.17
│   └── trimmomatic
└── work
    ├── 1_raw
    ├── 2_trim
    ├── 3_align
    └── ref_genome
```