Workbook

Basic Bioinformatics for Biologist

1. Linux basic commands

Command	Function
`pwd`	Print the current working directory (where you are)
`ls`	List files and directories in the current directory
`cd`	Change directory (move to another folder)
`mkdir`	Create a new directory
`rmdir`	Remove an empty directory
`cp`	Copy files or directory
`mv`	Move or rename files or directories
`rm`	Remove (delete) files
`cat`	Display the entire content of a life
`less`	View file content page by page (scrollable)
`head`	Show the first lines of a file
`tail`	Show the last lines of a file
`grep`	Search for patterns/keyword inside files
`wc -l`	Count the number of lines in a file
`cut`	Extract specific columns or fields from text
`sort`	Sort lines of a file
`uniq`	Remove duplicate lines (work after sort)
`echo`	Print a massage or variable to the terminal
`man`	Show the manual/help page of command

*) put `-h` or `--help` option to show the short instruction for each command

2. Obtain training data

2.1 Update package list

```
$ sudo apt-get update
```

2.2 Upgrade general packages

```
$ sudo apt-get upgrade
```

2.3 Install `git` using `apt` package manager

```
$ sudo apt-get install git
```

2.4 Clone training data from github

```
$ git clone https://github.com/reditama/bioinformatics-workshop.git
```

2.5 Check if the cloning is complete

```
$ ls -la
```

3. Basic Linux operation

3.1 Change directory to bioinformatics-workshop

```
$ cd bioinformatics-workshop
```

3.2 List files and folder in the directory

```
$ ls
```

3.3 Show the file size of files

```
$ ls -la
```

3.4 View the content of ecoli_500kb.fasta

```
$ less ecoli_500kb.fasta
```

3.5 View the content of ecoli_500kb.fastq.gz

```
$ less ecoli_500kb_hifi.fastq.gz
```

3.6 Use the `-S` option to unwrap the content

```
$ less -S ecoli_500kb.fastq.gz
```

3.7 View the content of gff file using head

```
$ head ecoli_K12_MG1655_example.gff3
```

3.8 Find gene annotation using grep

```
$ grep gene ecoli_K12_MG1655_example.gff3
```

3.8 Count the number of gene (Use pipe to perform commands simultaneously)

```
$ grep gene ecoli_K12_MG1655_example.gff3 | wc -l
```

4. Software installation

4.1 Seqkit installation

Hint: ask chatgpt

4.2 fastp installation

Hint: ask chatgpt

4.3 flye installation

```
$ cd
$ git clone https://github.com/fenderglass/Flye
$ cd Flye
$ make
```

5. Assembly preparation

5.1 Change directory to home (cd default argument is home/username)

```
$ cd
```

5.2 Create new folder called assembly and move inside (use && to perform a series of commands)

```
$ mkdir assembly && cd assembly
```

5.3 Copy ecoli_500kb.fasta to assembly folder

```
$ cp ../bioinformatics-workshop/ecoli_500kb_hifi.fastq.gz .
```

5.4 Inspect the statistics of fastq

```
$ seqkit stats ecoli_500kb_hifi.fastq.gz
```

5.5 Put the statistics in a file

```
$ seqkit stats ecoli_500kb_hifi.fastq.gz > ecoli_500kb_hifi.stats.txt
```

5.6 Perform qc using fastp, put your cleaned fastq in a new file (`ecoli_500kb_hifi_clean.fastq.gz`)

Hint: use fastp `--help` to learn how to.

5.7 Inspect the statistics of cleaned fastq

```
$ seqkit stats ecoli_500kb_hifi_clean.fastq.gz >
ecoli_500kb_hifi_clean.stats.txt
```

5.8 Assembly the clean reads using flye

```
$ /home/reditama/Flye/bin/flye --pacbio-raw
ecoli_500kb_reads_hifi_clean.fastq.gz --genome-size 5m --out-dir
flye_output
```

5.9 Inspect the output directory

```
$ cd flye_output
$ ls -la
```

5.10 Inspect the statistics of assembled contigs

```
$ seqkit stats assembly.fasta > assembly.stats.txt
```

***