IFN646 - Portfolio item 1

Key points

- Individual assignment
- 6 marks
- Release date: 3 August 2023
- Due date: 13 August 2023

Overview

For this portfolio item, we will ask you to align sequencing reads to a reference genome and analyse the results using a genome annotation.

You **do not** need to implement your own algorithms to align and count these reads (we will do that in the week 3 lab). Instead, you will use Bowtie2, and the *featureCounts* function from the subread package. You will only write the code to analyse the results. We recommend that you do this in Python, but you are free to use another language if you prefer.

We strongly recommend that you install both tools and start reading their manuals as soon as possible. You do not need to wait until the week 3 lecture for these initial steps.

Installation steps are given as an appendix to this file, for the virtual machine only. You are free to use a different environment, but you will have to find the instructions for your operating systems. Both tools are available for Linux, Mac OS and Windows.

1. Data for the portfolio item

- The complete genome for *E. coli* strain K-12, substrain MG1655
- A complete annotation (in the GTF format) for this substrain
- A pre-computed Bowtie2 index for this genome
- A set of reads (artificially generated to simulate a sequencing run; you each get a different set of reads)

All the data can be downloaded from Canvas. Gzip files need to be extracted to your working directory.

Remember that the genome and annotation are shared, but the set of reads is different. Only download the set of reads that matches the last digit of your student number. For instance, if your student number is n12345678, you need to work with file `reads_008_R1.fastq.tgz`. If you do not work with the assigned set of reads, your results will be incorrect and you may lose marks.

2. Tasks

Overall goal

Your objective is to map the reads to the reference genome, use the annotation to count the number of reads that are mapped to each gene, and finally report the three genes with the highest number of mapped reads.

Preliminary task

1. Familiarise yourself with:
   - The file formats used for the different steps:
     - FASTA
     - FASTQ
     - SAM
     - GTF
   - How to install and use `bowtie2`
   - How to install and use `featureCounts`
2. Make sure you have paid attention to the tasks details, including how to submit.
3. Download the genome, the annotation, the Bowtie2 index and your read set. Make sure that you are using the correct set of reads.

Task 1 [1 mark]

The reads are provided in the FASTQ format. Select one read from your set, and explain **in your own words** what the information given for that read means (including in terms of the quality of that read).

Task 2 [1 mark]

The mapping results you obtain from Bowtie2 are in the SAM format. Select one read from your results, and explain **in your own words** what the information given for that read means. Make sure not to forget any important fields.

Task 3 [1 mark]

Not all your reads are counted by featureCounts. Can you explain what happened to the others, and why?

Task 4 [3 marks]

Using `bowtie2`, `featureCounts` and your own processing of the results, identify the three genes with the highest number of mapped reads. Report both the genes and their read counts.

Apart from the input and output files, for this task you should use the default parameters of both tools.

3. Submission

You will submit your answers through the quiz environment on Canvas, in the "Portfolio Item 1" folder.

Remember that you can save and edit your answers as many times as you want, but **you can only submit once**.

Submission will close at 11.59pm on the due date.

4. Academic honesty

This is an individual assessment, and you need to submit your own work. We reserve the right to select some submissions and ask students to explain the reasoning behind their answers.

Appendix - Installation guides

Install Bowtie2 (v2.4.1) in Ubuntu (VM)

1. We are going to install Bowtie2 in the home directory.

   Open a Terminal shell (`ctrl + alt + t`) and navigate to your home directory using `cd ~` (tilde refers to home).

2. Download Bowtie2 (v2.4.1) from SourceForge using `wget` (docs).

   This will download a zip-archive of precompiled binaries.

   We will name the download (using flag `-O`) as `bowtie2-2.4.1.zip` :

   ```
   wget https://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.4.1/bowtie2-2.4.1-linux-x86_64.zip/download -O bowtie2-2.4.1.zip
   ```

3. Extract the contents of the zip-archive using `unzip` (docs).

   ```
   unzip bowtie2-2.4.1.zip
   ```

4. Add the directory containing the Bowtie2 executables to the `PATH` environment variable so that Ubuntu knows where to find them.

   Open your bash profile using gedit, a graphical text editor (docs):

   ```
   gedit ~/.profile
   ```

   When you invoke a terminal shell, the `~/.profile` file is read as a part of configuring the shell.

   Add the following code to the end of the file:

   ```
   export PATH=$HOME/bowtie2-2.4.1-linux-x86_64:$PATH
   ```

   Save and close the file.

5. Close and reopen the terminal window, or restart the VirtualMachine, for the change to take affect.

6. Check that you can execute Bowtie2 by typing `bowtie2 --version` into Terminal.

```
ifn646@ifn646-VirtualBox:~$ bowtie2 --version
/home/ifn646/bowtie2-2.4.1-linux-x86_64/bowtie2-align-s version 2.4.1
64-bit
Built on
Fri Feb 28 22:21:25 UTC 2020
Compiler: gcc version 7.3.1 20180303 (Red Hat 7.3.1-5) (GCC)
Options: -O3 -msse2 -funroll-loops -g3 -g -O2 -fvisibility=hidden -
I/hbb_exe_gc_hardened/include -ffunction-sections -fdata-sections -fstack-
protector -D_FORTIFY_SOURCE=2 -fPIE -DPOPCNT_CAPABILITY -DWITH_TBB -
std=c++11 -DNO_SPINLOCK -DWITH_QUEUELOCK=1
Sizeof {int, long, long long, void*, size_t, off_t}: {4, 8, 8, 8, 8, 8}
```

Install Subread (v2.0.1) in Ubuntu (VM)

1. We are going to install Subread in the home directory.

   Open a Terminal shell (`ctrl + alt + t`) and navigate to your home directory using `cd ~` (tilde refers to home).

2. Download Subread (v2.0.1) from SourceForge using `wget`

   This will download a zip-archive of precompiled binaries.

   We will name the download (using flag `-O`) as `subread-2.0.1.tar.gz` :

```
wget https://sourceforge.net/projects/subread/files/subread-2.0.1/subread-
2.0.1-Linux-x86_64.tar.gz/download -O subread-2.0.1.tar.gz
```

3. Extract the contents of the zip-archive using `tar` (docs).

```
tar -xf subread-2.0.1.tar.gz
```

4. Add the directory containing the Subread executables to the `PATH` environment variable so that Ubuntu knows where to find them.

   Open your bash profile using gedit, a graphical text editor (docs):

```
gedit ~/.profile
```

When you invoke a terminal shell, the `~/.profile` file is read as a part of configuring the shell.

Add the following code to the end of the file:

```
export PATH=$HOME/subread-2.0.1-Linux-x86_64/bin/:$PATH
```

Save and close the file.

5. Close and reopen the terminal window, or restart the VirtualMachine, for the change to take affect.

6. Check that you can execute featureCounts by typing `featureCounts -v` into Terminal.

```
ifn646@ifn646-VirtualBox:~$ featureCounts -v

featureCounts v2.0.1
```