BIOL199 BDB
Fall 2022

Lab 2: NGS Lab Part 1 (NGS1), Analyzing Read Data using Linux and Open-Source Software

(Learning) Objectives:

- Understand how terminals/command lines work on computers
- Use Linux commands to examine a FASTQ file and pull out information on the file
- Use *fastqe* and *fastp* to learn about read quality
- Interpret data and communicate findings

Pre-lab Readings:

1. "FASTQ Files" (https://compgenomr.github.io/book/fasta-and-fastq-formats.html)

Post-lab Assignment:

Due by the end of lab: Complete NGS1 worksheet - I will randomly choose one person's worksheet in your group to check completion and give feedback.

As you work through the protocol, don't forget to add notes to your lab notebook for today (link saved on Perusall). You may find it useful to copy commands you used into the lab notebook and add some notes, to help you quickly use these commands again (you will use some variant of these commands again in Lab 3).

A. Understanding the biology and practicing forming a hypothesis

Let's consider the biological question related to the two data files we are looking at today. These two data files were generated from sampling microbial DNA from the tissue of a female garter snake - specifically the tissue of the oral and musk glands. They derive from an experiment where researchers were interested in the differences in microbial communities in the mouth (oral glands) and the cloaca (musk glands). The glands contain pheromones generated by the garter snake to attract a mate.

For this lab, we are examining the data quality of their sequencing data. That is, the quality of the base pair calls for each base pair sequence. The DNA was cut into fragments, each fragment was sequenced (and each base pair in the fragment was assigned a quality score, and the resulting sequence fragments we will examine are referred to as **sequence reads**. Thus, we are interested in assessing read quality.

While we won't be fully analyzing the dataset, we can practice making a (biological) hypothesis. Note that a **hypothesis** is a testable explanation for a narrow set of phenomena. For "Are microbial communities found on an individual organism consistent across body parts?", form a hypothesis and put it in your worksheet **(Answer question A1)**.

DISCUSS: Based on your hypothesis, what do you think you would predict from an experiment comparing the sequence data? Share ideas with your group, and come to a consensus. After everyone in the class has completed Part A, we'll briefly discuss our responses. If everyone is satisfied with their thoughts here, jot some notes down in your lab notebook and continue on to Part B.

B. Accessing our Data Files

Instructions	**Comments**
1. Use instructions from Lab 1 or your lab notebook to sign back into the Spydur cluster. Remember you can check to see if you've signed in by looking for <username>@spydur in the bottom of your Terminal.	Remember that <username> should be completely replaced, including arrows. For me, I would use myang@spydur
2. (ONE-TIME STEP). For today's lab, we need to install a piece of software. Type `pip install fastqe` → this should cause a bunch of words to pop up on your screen, ending with successful installation. **Once installed successfully, you will not need to do this step again.**	The cluster has installed a tool (pip) that allows fast installation of programs that follow a particular format, set up by the writer of the code. These types of shared structures allow easy sharing of software between those using the software.
3. Type `srun --nodes=1 --ntasks=1 --time=04:00:00 --partition yang1 --pty bash -i` → this allows you to jump into an interactive Terminal session on one of my nodes for 4 hrs, where each of you will have the ability to do one job at a time. **DISCUSS & CHECK-IN: Look at the <username>@____ on the left side before your cursor. Scroll up and look at where you input the srun command. What is different after the @ sign? What do you think this reflects?**	We share the cluster with many other members of the UR community. Thus, instead of defaulting to using the first Terminal screen, we are using an interactive session that allows the cluster to officially provide time and computing power to us.
4. Type `cd /scratch/myang_shared/classes/BDB_F22/`	Remember that 'cd' means change directory.
5. Type `ls -lrth` **DISCUSS: What do the options -lrth mean? You can try dropping options or only typing `ls`, and use `man ls` to see description of options.**	Remember that 'ls' means list files. 'man' means to look at the manual of the given command. **Answer question B1.**
6. Use `cd <your first name>` to enter the directory you made last week.	e.g. I used `cd Mel` to enter my directory
7. Now, let's retrieve a data file filled with raw sequencing reads to examine, i.e. a FASTQ file. I have placed two files into /scratch/myang_shared/classes/BDB_F22/data/. Copy them into your folder with a command similar to the following (make sure you are doing this from inside your folder): ● `cp /scratch/myang_shared/classes/BDB_F22/data/female_oral2.fastq.gz .`	Note the period needs to be included here → the command is saying 'copy the specified file into my current working directory', where the period indicates the current working directory. These files have GZ at the end. This

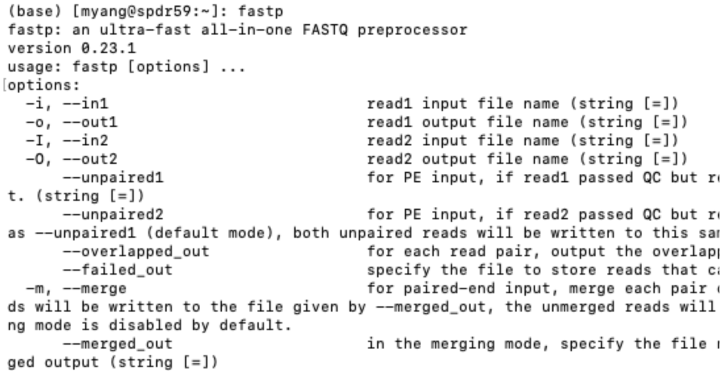
● cp /scratch/myang_shared/classes/BDB_F22/data/female_musk2.fastq.gz . ● When typing the above commands, there is a 'space' between cp and the file path.	means the files are zipped, or compressed to save storage space on the computer. This makes them not human readable, but still readable by a computer.
8. Check to make sure you now have the files you copied in your directory and note the file sizes on your worksheet. **DISCUSS: What is the command you use to look at file sizes? Check the command with each other, and make sure to check you all found the same file sizes before you move to step 9.**	**Answer question B2a for the oral2 column.**
9. Let's unzip these files so we can look at the data inside of them easily. For each file, type gunzip <filename> → after doing so, check how the file changed and the new file size.	'gunzip' unzips a compressed file→a file that has been made human unreadable to reduce its size. **Answer question B2b for the oral2 column.**
10. We will focus first on the female_oral2 file. Type less female_oral2.fastq → this allows you to look inside of the file. ● Try typing -S, followed by the Return button - this switches from wrapping text to not wrapping text (use right/left arrows to see end of line) ● Use your spacebar to jump large pieces of the file ● Use up and down arrows to move up and down ● Type 'q' to exit.	'less' allows you to open a file for read-access only. Thus, you cannot edit the file contents.
11. Now type head -n4 female_oral2.fastq → this determines the first four lines of the file and prints them to your Terminal screen. Review the pre-lab reading on FASTA/FASTQ files and use it to examine the format of female_oral2.fastq. **DISCUSS & CHECK-IN: For the THIRD read in the oral2 file, the last three bases are ACG -- what are the corresponding quality scores listed in the file? (HINT: you may want to change the number after the -n in the head command to help you answer this question).**	Note that you can also use tail -n4 female_oral2.fastq to look at the LAST four lines of the file. The number added after -n indicates the number of lines to show. **Answer question B3.**
12. Use wc -l female_oral2.fastq to determine the number of lines in your FASTQ file. Use this to determine the total number of reads in the file (what do you have to divide by?).	'wc' means word count, and the option '-l' means lines. Thus, 'wc -l' means count the number of lines in a given file. **Answer questions B2c-B2d for the oral2 column.**
13. Now repeat the above step for female_musk2.fastq.	**Answer questions B2a-B2d for the musk2 column**

14. If you haven't been taking notes on the commands above in your lab notebook, take the time to do so now. Jot down the commands you learned, when you had to put a filename, an example of how to use the commands, and perhaps an explanation of what the command does. Feel free to paste any text or figures from this document or the Internet as is helpful.	Adding a link to the original source of anything you reference or paste is helpful as well, in case you need the original documentation. Remember in next week's lab, you will have to run through a very similar protocol.

C. Analyzing read quality using information from the FASTQ file

With the sequencing data, the very first question to examine is what is the quality of the reads. We will use software that helps to analyze the read quality in the FASTQ file. In Part B, after step 2, you should have already downloaded the software we will use first to analyze read quality: **fastqe**. We will use this software to examine the quality of reads in two FASTQ files - one derived from sequence data for microbial communities in mouth glands (female_oral2.fastq) and one derived from sequence data for microbial communities in musk glands (female_musk2.fastq).

Instructions	**Comments**
1. Let's determine the average read quality at each position in the FASTQ file. In your folder, type `fastqe   female_oral2.fastq`.	
2. Now, run 'fastqe' for the female_musk2.fastq file as well. Describe what you see - what do each of the three columns refer to? Just based on your general understanding of emoticons, how would you interpret average read quality for each of your datasets (i.e., do you think the quality is good or bad for that base)? **DISCUSS & CHECK-IN: Share your answers to the above questions with your TA or instructor.**	For understanding the output, consider the following: Each read is the same length. An emoticon in position 1 indicates the average quality at position 1 across ALL reads in the file.
3. Now type `fastqe   --scale   female_oral2.fastq` → this adds an option allowing you to see what each emoticon stands for.	Look through the following link (http://tiny.cc/qualscores) - can you describe what a quality score means? **Answer question C1.**
4. Use `fastqe   --help` to explore different options available. Take some notes on different options available.	**Answer question C2.**
5. Researchers typically don't use 'fastqe'. One commonly used software that performs a similar function as 'fastqe' is 'fastp'. fastp is already installed, and you can test this by typing `fastp` and running this command → you should see a very large list of potential options to use.	



```
(base) [myang@spdr59:~]: fastp
fastp: an ultra-fast all-in-one FASTQ preprocessor
version 0.23.1
usage: fastp [options] ...
options:
  -i, --in1                             read1 input file name (string [=])
  -o, --out1                            read1 output file name (string [=])
  -I, --in2                             read2 input file name (string [=])
  -O, --out2                            read2 output file name (string [=])
      --unpaired1                       for PE input, if read1 passed QC but r
t. (string [=])
      --unpaired2                       for PE input, if read2 passed QC but r
as --unpaired1 (default mode), both unpaired reads will be written to this sa
      --overlapped_out                  for each read pair, output the overlap
      --failed_out                      specify the file to store reads that c
  -m, --merge                           for paired-end input, merge each pair
ds will be written to the file given by --merged_out, the unmerged reads will
ng mode is disabled by default.
      --merged_out                      in the merging mode, specify the file
ged output (string [=])
```

Listing of some of the options shown when I typed in fastp.

EDIT: Prior to this step, install fastp by typing cp ~/shared_myang/fastp . (don't forget the period!)

6. Type ./fastp -i female_oral2.fastq -o female_oral2.q15.fastq ⟶ this uses your FASTQ file as input and outputs a trimmed FASTQ file where low quality bases or reads are removed.

Reads are only kept is: (1) the average quality score is ⩾15 across all bases, and (2) if no more than 40% of the bases have a quality score <15. It also filters reads with too many bases not sequenced (N) and reads that are too short (<50 bp long).

7. Type ls -lrth ⟶ you should see that three new files were made (JSON, HTML, and the new FASTQ file you specified with -o).

JSON files are written for easy parsing in several computing languages, esp. Javascript ⟶ we will not be using this. HTML files are designed to be visualized on a web browser.

8. Examine the number of lines from your original FASTQ file and compare it to the number of lines in your new Q15 FASTQ file. How many reads were filtered out of your original dataset? **DISCUSS: What is the command you use to look at the number of lines? You used this command towards the end of Part B. Check the command with each other, and make sure to check you all found the same number of lines for your original FASTQ file and new Q15 FASTQ file before you move to step 9.**

Answer question C3a-C3c for the oral2.q15 file.

9. We want to save the HTML file, but it was given a generic name of fastp.html.
To change its name, type mv fastp.html female_oral2.q15.html ⟶ this renames your file to one matching your new output file.

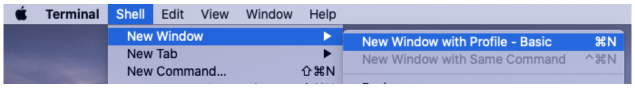
10. To open the HTML file, we need to download the file onto our computer and open it in a web browser. To do this:

i. Open a new Terminal window where you did NOT sign into the Spydur cluster.

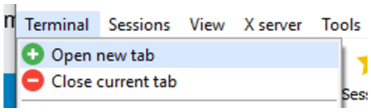
- Macs: Make sure you're in Terminal. Click Shell ⟶New Window ⟶ New Window with Profile-Basic

Definitely ask if something goes wrong here, as not all computers might be set up the same.

For Windows Users, the <username> should be the username for your personal laptop. Ask your instructor if



- Windows: Make sure you're in MobaXterm. Click Terminal → Open new tab



ii. Change into a folder that you can easily access.

- ○ Macs: Type cd   ~/Downloads/ to change into your Downloads directory.
- ○ Windows: Type cd /cygdrive/c/Users/<username>/Downloads/ to change into your Downloads directory.

iii. Check and make sure you're in your Downloads/ directory using ls   -lrth

iv. Type scp   -r <username>@spydur:/scratch/myang_shared/classes/BDB_F22/<your first name>/female_oral2.q15.html   .   → this should copy your file from the Spydur cluster into the folder you're in on Terminal, hopefully your Downloads/ folder

v. Click the link from your Downloads/ folder OR type open female_oral2.q15.html to tell your computer to open the file → your computer should know if it sees an HTML file to open it in your default web browser.

vi. A lot of metrics are shown → our focus is on the tables in the 'Summary' section (ignoring GC content row), the first figure under 'Before filtering', and the first figure under 'After filtering'. Take screenshots of each of these graphs to store in your lab notebook.

DISCUSS & CHECK-IN: In the 'Summary' section, discuss what you think Q20 and Q30 might mean. Then, check with your instructor or TA to see if you are correct.

you're not sure what this is.

Answer question C3d for the oral2 file.

11. Now repeat the process starting from step C6 for the female_musk2.fastq file.

Answer question C3a-C3d for the musk2 file.

12. Answer the remaining worksheet questions for Part C.

Answer questions C4-C6.

If you finished early, make sure you've completed all parts of the post-lab. Then, check in with your instructor, to have one worksheet chosen at random for submission in your group.