BIOL199 BDB
Fall 2022

Lab 3: NGS2 Lab Part 2 (NGS2), Quantitatively assessing differences in data

(Learning) Objectives:

- Practice some more with using Linux commands to retrieve data
- Gain familiarity with using Microsoft Excel
- Understand and apply a Student's t-test

Pre-lab Readings/Assignments:

1. Make sure you have Microsoft Excel installed, see instructions at this link (http://tiny.cc/install_mic_office)
2. Read through the protocol below
3. If you have not used Microsoft Excel before, I encourage reading http://tiny.cc/excel_intro (needed after step A13)
4. Review your notes from Week 2 lab

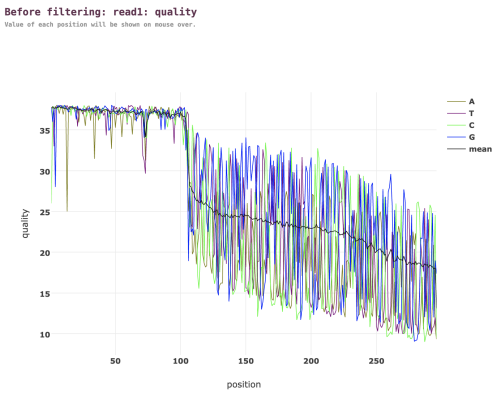
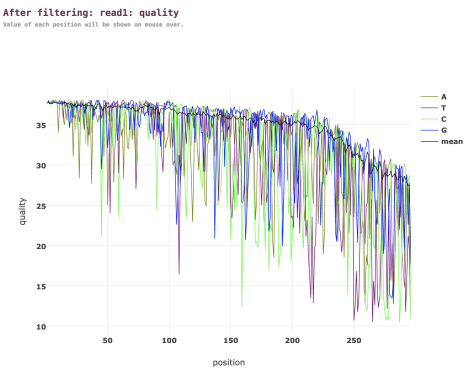
Post-lab Assignment:
Due by the end of the lab: Complete NGS2A and NGS2B worksheets - I will randomly choose worksheets in your group to check completion and give feedback.

A. Use quantitative means to compare (read) data

Last week, we examined FASTQ files containing read data from a NGS experiment on microbial communities of the oral and musk glands of a female garter snake. We learned how to interpret read quality scores, use useful Linux commands to navigate a remote server, and apply open-source software (*fastqe*, *fastp*) to assess read quality and filter FASTQ files to retain only high quality data. In this process, you examined four sets of figures - read quality before and after filtering for oral glands, and read quality before and after filtering for musk glands. Using these figures, you came to a conclusion about the change in read quality, but your answer was *qualitative* - you described a change, but it's unclear if that is a meaningful difference.

Statistical analyses are one way to provide a *quantitative* way of describing differences. That is, they are a way of assigning a number to your confidence that there is a meaningful difference between two populations under study. In this lab, you will treat the set of reads for each of your four conditions as a different population, and assess whether there is a meaningful change in the read qualities you observed. We will start with the same question, and after you're comfortable with the analysis, your group will determine another question to answer for the same dataset.

The question our class will start with is whether there is a difference between read quality scores for the beginning of a read before filtering versus after filtering. To answer this question, we will focus on position 50. I have put the figures related to this question below.

		**DISCUSS: What are your null and alternative hypotheses for the question "Is there a difference in read quality at position 50 before and after filtering for the oral gland data?"** **Answer question A1.**

As you work through the protocol, don't forget to add notes to your lab notebook for today (see 'Links' on Perusall). If you kept good records of the steps you used in Lab 2, that should help this week.

Instructions	**Comments**
1. Log into your account on the Spydur cluster, create an interactive session, and enter our shared class directory where you made a folder last week. The quick version is here, but if you're not sure what these commands are and want more details, you can return to last week's lab for more information. • Access Terminal and type `ssh <username>@spydur` • Enter your password (you won't see it) • Type `srun --nodes=1 --ntasks=1 --time=04:00:00 --partition yang1 --pty bash -i` • Type `cd /scratch/myang_shared/classes/BDB_F22/`	If you successfully complete this, you should see yourself in spdr59.
2. Type `ls -lrth` or `pwd` to make sure you are where you want to be	(i.e. you see a bunch of folders belonging to you and your classmates)
3. Use `cd <yourfirstname>` to change into your folder (and double check with *pwd*).	
4. We will now grab the script you need to pull out read quality scores, *readquality_bypos.py*, as our previous software *fastp* does not easily generate a text file of this type of data. Let's copy the script into your directory. Make sure you are in your directory, and if so, type `cp /home/myang/shared/readquality_bypos.py .` → don't forget the period - it tells Spydur to copy the file to your current directory, which should be your personal folder.	Remember to use `ls -lrth` to check that the file is in your folder → this is always a good thing to do to make sure you're in the right folder and have every file you need.
5. Now type `less readquality_bypos.py`	

6. Remember the question the class is asking is whether there's a difference between read quality at position 50 before and after filtering for the oral gland data. We want to sample a reasonable number of reads from each dataset and assess whether there is a difference using a statistical comparison. For our analysis, we will use a sample size of 100 random reads for each dataset. Examine the header of the PY file (readquality_bypos.py), that is, the description at the top in lines starting with "##". When done, you can use 'q' to exit the file.

DISCUSS & CHECK-IN: Read the description after the '##' at the top of the PY file. You can ignore everything that does not start with a '##'. What does this file do? Show your TA or instructor an example of the full command you need to type to run this script and retrieve 100 random read quality scores at position 50 of each read.

Command: _____

You don't need to understand everything below the '##', but I wanted you to see an example of Python code.

Mine's a pretty messy one, but I included comments (in the '##') so you can see what each section of the code does.

If you're interested in Python, CS150 teaches the basics.

7. To generate a file of 100 random read quality scores at position 50, type the command that was approved in the previous step's check-in → you should get four lines of output, summarizing your analysis.

```
(base) [myang@spdr59]: python readquality_bypos.py 200 female_oral2.fastq 10
   Sampling 200 read qualities for pos10 in
   female_oral2.fastq
   provided in
   female_oral2.10.txt
(base) [myang@spdr59]:
```

Note that the image above does NOT show the command you need to run. It shows the command in the example from the PY file.

Note this won't work if your fastq file is zipped (has .gz at the end). Check your file, and if it's zipped, you can use `gunzip female_oral2.fastq.gz` to unzip the file

Remember after running the PY file to check and see if your new datafile is generated (using `ls -lrth`).

8. Use `less female_oral2.50.txt` to look inside and see what you have.

DISCUSS: How many lines do you think this file should have? How can you check this and see how many lines you have? (throwback to last week's lab)

Remember that this file has printed out the quality score for each randomly sampled read at the position you designated.

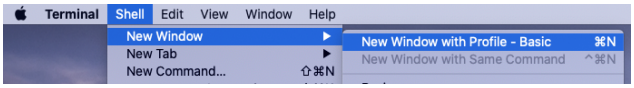
9. Repeat step A7, but now for your filtered data from last week. The filename should look something like 'female_oral2.q15.fastq' or 'female_oral2.Q15.fastq', but check your folder using `ls -lrth` if you aren't sure.

The new file you generate should have the same filename, with '.fastq' replaced by '.txt'

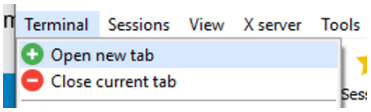
10. Our goal is to compare the data in these two files. For that purpose, we will use Microsoft Excel on our computers. Thus, we need to download these two data files onto our computer. Download both files, following the steps shown for the

'female_oral2.50.txt' file:
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



- 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` or `pwd`

iv. Type `scp -r <username>@spydur:/scratch/myang_shared/classes/BDB_F22/<your first name>/female_oral2.50.txt .` → this should copy your file from the Spydur cluster into the folder you're in on Terminal, hopefully your Downloads/ folder.

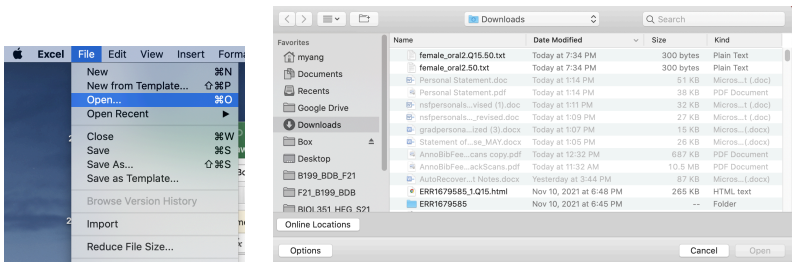
v. Then repeat the above for your second file containing your post-filtering reads (the Q15 file).

11. On your computer, navigate to your Downloads/ folder and make sure the two files are present.

12. Open Microsoft Excel and click File → Open…. In the new pop-up, navigate to Downloads/ and click on female_oral2.50.txt. In another pop-up that mentions the word 'Delimited', just click 'Finish'. This will open the TXT file in Excel. Repeat for the post-filtering file as well.

If you cannot access the file, adjust so that you look at 'All Files', not just 'Excel Files'.

In Windows, the option to do this is usually to change 'Excel Files' to 'All Files', on the bottom right.

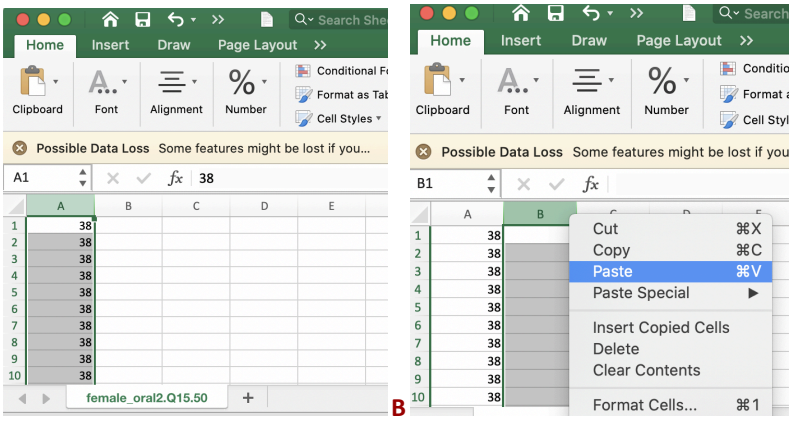


In Macs, the option to do this is usually with the 'Options' button on the bottom left.

13. Microsoft Excel is a useful tool for analyzing data. Looking at either file, if you scroll down, you should see that the data goes to row 100, with each number in one cell in the first 'A' column. Let's copy the data from the filtered file into the unfiltered file, pasting them in the 'B' column. The fastest way to do this is as follows:

- Go to your filtered (q15) file in excel.
- Hover over the column name ('A') until you see a black arrow pointing down.
- Click, which should highlight the entire column
- Copy the highlighted data
- Switch to the unfiltered file, and highlight the 'B' column
- Paste your data
- Make sure that both 'A' and 'B' columns have data from rows 1 to 100
- Exit out of the filtered data file, as you have copied it over

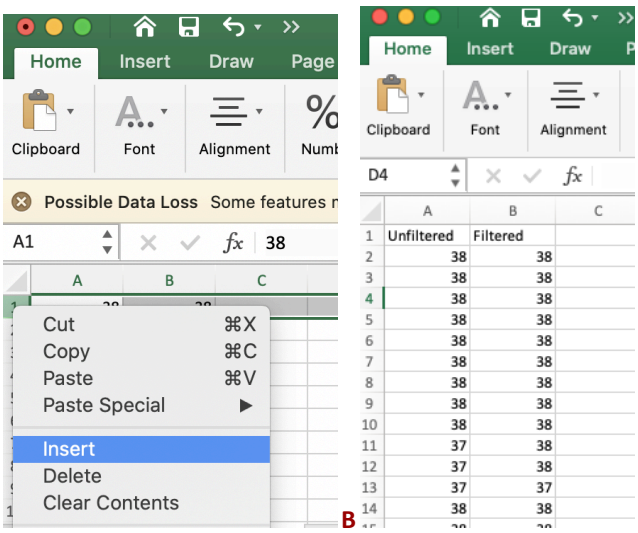
If you aren't familiar with Microsoft Excel, this is a good link to start from (http://tiny.cc/excel_intro).



A. Highlighting data from filtered file to copy. B. In the unfiltered file, highlighting column B to paste the copied data.

14. We want to make sure we don't mix up the data. Right click on top of the '1' for Row 1, and click 'Insert' to make a new row above Row 1. In the empty cells, give your two columns of data headers.

Perhaps 'Unfiltered' and 'Filtered'? Make sure you don't accidentally mix up your columns.



A. Insert row above row 1. B. Added column headers to identify data

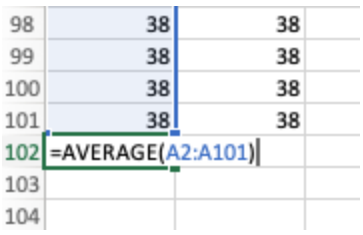
15. We will use the Student's t-test to examine whether there is a **significant** difference between read quality at position 50 before and after filtering. The equation to calculate our observed t-value is as follows, which we will calculate using Excel formulas.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

The x1 and x2 in the image are the average for each dataset. s1 and s2 are the standard deviations for each dataset, and N1 and N2 are the sample size for each dataset.

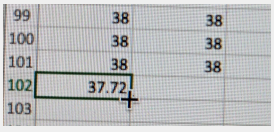
A Student's t-test calculates a t-value that helps us assess the probability that the two datasets are the same with respect to the tested variable (i.e. quality score). This link (http://tiny.cc/ttest) provides additional information on a Student's t-test.

16. To use formulas in Excel, double click a cell and start by entering an equal sign, followed by the formula. For instance, =AVERAGE(A2:A101) would give me the average of the first 2nd-101st rows of data in column A (or the average of our unfiltered data at position 50). You can either type the cell range in, or you can click the first cell and drag down until all cells are highlighted so Excel can autopopulate the range). For A102 and B102, determine the averages.



Note that the = sign is necessary for the formula. The highlighted range has a color that matches the range written in the parentheses. Press enter to complete the formula calculation.

Note that a fast way of applying the same formula to multiple columns or rows is to hover over the bottom right of the cell until you get a black plus sign. Then, click and drag to the empty cell. The formula will adjust to grab the corresponding row or column (so you should check to make sure it grabbed the data you want).



Note the black plus sign, which is my cursor when it hovers over the bottom right of the cell.

17. We also need the spread of the data, as denoted by standard deviation. The formula in this case is =STDEV(cells you want to include). In row 103, show the standard deviations for each dataset.

This link (http://tiny.cc/stdev) has a short explanation of standard deviation.

Answer question A2. Note that your answers here and after may look different, because each of you have a different random sample.

18. In row 104, include the number of samples, using =COUNT(cells you want to include).

This presumably should be 100 for each column.

19. Now, in C102, C103, and C104, add a label to indicate the three things you just calculated.

This is to make sure we don't forget what our numbers mean.

20. Now, examine the t-value equation from step A15. In several cells, use formulas to calculate your observed t-value. Note that =SQRT(value) determines the square root and =SUM(cells you want to include) determines the sum of all included cells.

A. Difference of two means	
B. s1^2	**C. s2^2**
D. B/N1	**E. C/N2**
F. D+E	
G. sqrt(F)	
H. A/G	

Schematic of how to break up the equation. The formulas after A-H indicate the operations used. Note that H needs all the previous cells.

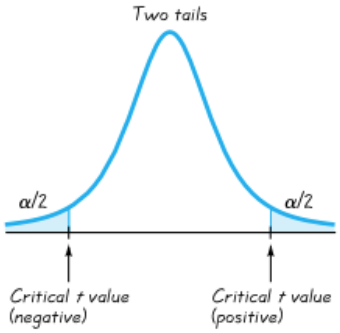
Note 1: I encourage taking each part of the equation and calculating them in separate cells, and then performing the final calculations, perhaps following the schematic in figure on left. That way, it's easier to correct typos along the way.

Note 2: Your t-value should not match that of your labmate's, as you each started with a different random sample.

DISCUSS: Double-check with each other that everyone understood these steps and have a similar Excel format before moving forward.

21. We now have an observed t-value. Our question is if this t-value is large enough that it goes beyond a critical t-value we specify.

To determine our critical t-value, we need to determine the degrees of freedom of our comparison, or (N1-1)+(N2-1).



Going beyond the critical t-value (entering the shaded blue area in the figure) means our two datasets have a difference that has a very low probability of occurring if they represent the same original population (i.e., the null hypothesis).

Staying lower than the critical t-value (entering the non-shaded area) means our two datasets have a difference that has a high probability of occurring if they represent the same original population (i.e. the null hypothesis).

Answer question A3.

22. We can use the t-test table (http://tiny.cc/t-table) to find our critical t-value. Here, the numbers under 'Area in Two Tails' are the probabilities of interest (that is, potential p-values). **For instance, p=0.05 means a 5% chance that the data would show the observed differences if the null hypothesis were true.** For many people, a 5%, or 1 in 20, chance of observing this difference in means assuming the null hypothesis is true seems very unlikely, so they might be inclined to argue that the observed differences can't be due to the null hypothesis. Then, they would conclude that the null hypothesis should be rejected. We will use this cut-off of p=0.05, but note that sometimes, more or less stringent p-value cut-offs are preferred. To use the table, determine the row corresponding to your degree of freedom (d.f.), or the row just below your d.f.. Then, locate between which two columns your t-value falls. The column headers under 'Area under Two Tails' are the corresponding p-value ranges. For instance, a t-value of 2.2 with a d.f. of 3 would correspond to a p-value between 0.1 and 0.2 (i.e. $0.1<p<0.2$).

DISCUSS & CHECK-IN: For a degree of freedom of 92 and a t-value of 2.0, what is the corresponding p-value range in which your t-value falls? What is the meaning of this p-value? Discuss with each other, and double-check with your instructor or TA.

'Area in Two Tails' refers to both of the shaded blue regions. Our alternative hypothesis is that the filtered and unfiltered datasets are different, but we don't a priori know if one might be higher than the other. If we are confident about one dataset always being greater than the other dataset, then we might use an 'Area in One Tail' analysis.

Others might prefer a more strict boundary, such as a 1% cut-off, or a 1 in 100 chance of the observed difference if the null hypothesis were true (p=0.01). These levels of confidence we build around a particular p-value are dependent on norms within the field of study and personal preferences. Remember, it's all about how you communicate your data, so others can decide whether to go with your confidence or not.

Answer question A4.

23. On Microsoft Excel, in a new cell, type =TTEST(cells with dataset1, cells with dataset2, 2, 3) → this will calculate your p-value directly! Check that the exact p-value here matches the p-value range you found in the step above.

First and second entries are your two datasets, third entry (2) indicates you want a two-tailed t-test, and fourth entry (3) indicates you have two independent datasets of different variances.

24. Based on what is written in step A22, write a sentence describing the technical meaning of your calculated p-value. Then, write one sentence communicating your research finding (i.e. answer your question of whether there is a difference between your read quality before and after filtering at position 50). Usually you put your p-value in parentheses (p=##) immediately after this sentence.

Answer questions A5-A6.

25. Save your Excel workbook by clicking File → Save As. Give a name to your file (e.g. B199_Lab3_PartA.xlsx) and make sure you save the file format as an Excel Workbook (XLSX or XLS).

Take a break! Talk with your group about a 3-5 min break to rest your eyes. Then come back and start on Part B.

B. Try a different comparison with your group

Now that you know the above steps, you and your group can pick another question to examine regarding our read quality data for the oral and musk glands. Look at the figures from last week for your oral and musk glands and develop a question comparing two positions and whether there is a difference in quality scores. **I encourage looking for an area where you would hypothesize that there is a difference in quality scores.** Discuss with each other what question you want to examine.

DISCUSS & CHECK-IN: Run your question by your instructor to make sure it is appropriate before doing the statistical analysis. Upon completion of the check-in, you will get a hard copy of the Part B worksheet.

Pose your question and follow the steps of the above protocol to determine whether there is a difference in quality scores related to your question. **Fill out questions B1-B5 on the worksheet.**

Your first paper will be to communicate your findings from Part B of this worksheet.