**MIP 280A4: Microbial sequence analysis**

**Snake exercise: downloading and cleaning data, In-class exercise questions**

*All questions 1 point unless noted*

**Before conda**

1. Is fastqc in your PATH? How do you know?
2. What happens if you try to run the command fastqc? (What does the shell output)?

**With bio\_tools conda environment activated:**

1. Now, what is the output of the command which fastqc?
2. Now, what is the output of the command echo $PATH?
3. Based on the previous answers, how do you think conda environments relate to your PATH?
4. Another tool in this environment that we will use today is named cutadapt. What version of cutadapt is in your bio\_tools conda environment? How did you figure that out?
5. What are the first 6 bases of the first read in SRR1984309\_1.fastq?
6. What are the quality scores *in encoded single characters* of the first 6 bases?
7. What quality scores do these characters represent?
8. What are the probabilities of incorrect basecall associated with these particular Q scores?
9. How does the first read in SRR1984309\_1.fastq relate to the first read in SRR1984309\_2.fastq?
10. How many reads are in each file? (Hint: the wc -l name\_of\_file command will tell you the number of *lines* in the file)?

**After trimming**

1. What does | tee cutadapt.log do in the cutadapt command we ran?
2. What percentage of all bases were quality-trimmed (hint: see cutadapt.log)?
3. What percent of read1 reads contained adapter sequence?
4. What percent of read pairs made it through the filtering?
5. What command did you use to run fastqc on the trimmed fastq?
6. How many read pairs remain after trimming?
7. Did the trimming remove residual adapter sequences from reads?