Midterm Exam

# Bio200: Introduction to Bioinformatics

This exam is given as a “take-home” exam with no sitting time limit imposed. You have until 11:59PM on October 22, 2018 to complete the questions below to the best of your abilities. No exam will be accepted after this time. You may consult the course materials, textbook, and online resources but the answers you give must be in your own words. You may not discuss this exam or any of its questions with any other person until after October 23, 2018. If you reference information that is not given in any of the course materials (textbook or posted lecture notes/slides) then you must cite the source of your information using in text and full APA bibliographic citation.

To acknowledge that you have read and understand the above please sign or type your name here:

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\* Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\*Typed name constitutes your electronic signature.

Instructions: Complete the following questions as best you can. When you are finished please submit the exam and any associated files. Email the files to [rharbert@stonehill.edu](mailto:rharbert@stonehill.edu) by 11:59PM on October 22, 2018.

Points: Each question is assigned a point value. Use this to guide your effort. A 10 point question should require about twice as much response effort as a 5 point question. You may be able to answer a 5 point question in a few sentences but a 10 point question may take a couple paragraphs.

Questions: (Unless instructed otherwise type your answers under each question)

1. Define the term “big-data” as it relates to biology. Provide at least two examples of what can be considered “big-data” in biology and summarize the challenges associated with each. (10pts)
2. What is meant by “next generation sequencing” (NGS)? Which technology or technologies for DNA sequencing are most often referred to by this term? (10pts)
3. Pick one DNA sequencing technology to describe in detail. How does it work? What are the primary data being recorded by the instrument? What is the scale of data being generated by this technology? What are the costs associated with it? What is it really good at? Are there drawbacks? (You may want to look at the collaborative document you all contributed to for some starting points for this question). (20pts)
4. The mitochondrial sequences that we have been working with for our select mammals are all about 16,000 base pairs long. In a sequence that length how often would you expect to find any arbitrary sequence of length 6 bases? (SHOW YOUR THOUGHT PROCESS) Discuss why this is important when considering analyses like BLAST and multiple sequence alignment (HINT: think about the many parameters we can set for those programs). (5pts)
5. How much RAM does your laptop have? If each cell in an R table uses 16 bytes per decimal number, how many rows can a 10 column table have in R before you run out of space in memory on your computer? Explain why this is useful information. (10pts)
6. What does the Unix command ‘cut’ do and how is it used? (HINT: try ‘man cut’ or ‘cut --help’ on the command line). Download the file at: <https://raw.githubusercontent.com/rsh249/bioinformatics/master/data/ebd_trim3.csv> and use ‘cut’ to print columns 3, 5, and 6 in a new file. Include that file in your exam submission named “cutout.csv”. Record the command that worked below. (5 pts)
7. Perform a BLAST search using the file <https://raw.githubusercontent.com/rsh249/bioinformatics/master/data/sequence.fasta> as your query against the human genome database we created in class. Record the series of commands that worked for you here. Use –outfmt 6 and save the output table to a file to be included with your exam named “blasthits.tab”. Where in the human genome did you find a match for that sequence? (15pts)
8. Download some data with:

surveys <- read.csv("https://ndownloader.figshare.com/files/2292169”)

These data are from a survey of animal trappings. Pick one categorical column (plot\_id, plot\_type, sex, genus, taxa) and one measurement column (hindfoot\_length, weight) to plot in ggplot. It is up to you to decide on a suitable plot method for these data. Paste your plot into the exam document. (5pts)

1. Download the fasta file of HoxC8 sequences: <https://github.com/rsh249/bioinformatics/blob/master/data/hoxC8.txt>
   1. Align this file using Muscle using whatever settings you see fit. These are DNA sequences of a highly conserved Hox gene for a few canonical animal genomes. Include the aligned fasta file in your submission. (5pts)
   2. Copy the exact command you ran here. (5pts)
   3. Explain the command and why you made any choices you made. (10pts)
2. Extra Credit: With sources cited – What is the default kmer length used in the kmer clustering step(s) of muscle and how is this kmer length used? For full points provide details about how you could change the kmer length and what effect you expect that to have on the algorithm. (5pts)