Bioinformatics, BSC 4930/6932, Midterm 1 Name: Kyle Cameron
Take-home practical

Write your name at the top right of this page.

This midterm is out of 100 points. 65 of those points were from the in-class exam sheet that you already completed. The remaining 35 are from this take-home, open-book practical exam distributed on Canvas that is due on **Tuesday, 20 February, at 5 pm**.

You are welcome to work together on this practical. However, **the work that you submit must be your own**. Submitting identical work will be considered plagiarism and you will receive zero points for this part of the exam.

This practical is largely based on coding in R, so will be similar to the homeworks that you have completed. As Bioinformaticians, we often re-use code, and you are welcome to do so here. I recognize that much of the code will be identical to both that of your classmates and to the code you've already written. That is fine. For every line (or couple of lines) of code, you must write a comment line describing what your code is doing. This is what must be your own work. See the answer key to homework 5 as an example. The "#" at the start of a line of code makes it a comment line.

For any questions in this practical, put your answers in a comment in your R script following the relevant code (so, write your answers in an R script, not here). Save this R script to your Bioinformatics GitHub page. Submit a link to your GitHub page in the Midterm 1 assignment on Canvas. Make sure that the GitHub repository is public (otherwise I can't access it).

To make a GitHub repository public, go to your repository on the website, click 'settings', scroll to the bottom under 'Danger Zone' and click 'Change visibility', then 'change to public'. There will be a few more steps to verify that you want to make this change.

Push all your scripts, data files, and results plots to GitHub.

The questions in this practical are on the following page.

Take Home practical questions:

Your research lab is conducting a project that involves sequencing a gene in a population of people. You've sequenced this gene in 20 people so far, and you want to find out if there is any variation in this gene in your population. There is a file on Canvas containing all the DNA data from your sequencing efforts. Go ahead and download it to your Bioinformatics folder.

1. Import and align your DNA sequences

2. Check to see how different your samples are from one another. Are any of them different from the rest? If so, what kinds of mutations do you observe in this individual (or individuals)?

Homo 6 is the most different from the rest starting with an obvious deletion and some missmatches

3. You suspect that an individual (or individuals) in this population might have some mutations in this gene, but you don't know what this gene might be. Compare your sequences to a database to figure out what the gene is. Export your data, paste it into the relevant database search engine, and add your results to a comment line in R. What is the gene? What is the accession number of the best match to your search?

I added the file to blast and found Homo sapiens HBB gene for beta globin to be the closest match. Accession number GenBank LC121775

4. Find the individual that is the most different from the rest of the individuals in your dataset. Translate that sequence to protein. Write it to a fasta file.

homo sapiens 6

5. Use a database to figure out what your protein matches to. Click on the record for the best match. What is the accession number of this entry?

Using blast my protein matches Hemoglobin subunit beta, HBB, Homo sapiens (A0A0J9YWK4)

6. Either using R or by searching in the database, what disease(s) is this gene associated with? Does this person have the disease?

HBB is associated with some form of Beta thalassemia, I think they do have it!

7. What is the 3-dimensional structure of this protein? You can include a screenshot or download of a photo of this structure in your GitHub repository.