**Assignment One: Exploration of Species Composition of Gut Microbiomes in Healthy Patients**

**/21**

**Data and Project Description:**

*The data presented here was generated by amplifying and sequencing DNA from the V4 16S ribosomal subunit of various samples which were collected from the colon microbiomes of healthy patients. This region is commonly sequenced because it is well-conserved within bacterial species while also containing enough nucleotide diversity to distinguish different species. The files LB.fasta and RB.fasta contain the DNA sequences sourced from the distal and proximal mucosa respectively. LS.fasta and RS.fasta contain the sequences from the distal and proximal lumen respectively. The sequences in reference.fasta are annotated to taxonomic order. For further reading on this method, please see the following article:* [*https://www.nature.com/articles/ismej2017119*](https://www.nature.com/articles/ismej2017119)*. Note, this project will mostly use commands you have learned in class. There will be commands for which you will need to look online. Also, there is an expectation that the entire project is completed using the bash shell. No other programming languages can be used. This data was obtained from the following study:* [*https://www.ncbi.nlm.nih.gov/pubmed/29636352*](https://www.ncbi.nlm.nih.gov/pubmed/29636352)

**Part One: Installing NCBI Blast**

Install Blast by following the instructions at the following site:

<https://github.com/enormandeau/ncbi_blast_tutorial/blob/master/README.md>

**Part Two: File Statistics**

1. Complete the chart below. Provide the command you used to complete the chart (2 marks):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LB.fasta | RB.fasta | RS.fasta | LS.fasta |
| Number of Unique Headers |  |  |  |  |

1. Copy and paste the command to output a sorted list of the orders found in “reference.fasta” to a file “output.txt”. Display the first 20 lines of the file “output.txt” (include the command) (3 marks).
2. Create a Blast database using the file reference.fasta. Copy and paste the commands used. How many unique headers are found in the reference file? How many sequences were added to the database? Provide an explanation for this difference (2 mark).
3. Use Blastn to search each of the test files (LB.fasta, RB.fasta, SS.fasta, RS.fasta, LS.fasta) for matches to the database that you created in question 3. The results of this search should be in tabular output with comments and the output should be redirected to a “Results” directory. Provide the commands used and the output of the first 10 lines of each file. On line 9, which fasta sequence was aligned to a reference sequence? How well? Does the same sequence align to other genera? (4 marks).
4. Sort each blast results file and select the top scoring hit for each sequence. Output a list of these hits to a text file. Provide the commands used (1 mark).
5. Create a file which contains the number of query sequences found for each reference genera. Copy the output from this file as an answer to this question. Include the commands you used and, if necessary, the source for where you found them (4 marks).
6. Are there any obvious differences in the genera compositions? Given your knowledge of where these files came from, speculate on what could be driving these differences (if any) (2 marks).
7. What is the purpose of the reference database? Describe some problems and advantages that could be encountered by using a reference database for elucidating community composition (3 marks).