# Bat gut microbio metagenomic analysis

TJ Rogers

import pandas as pd

## 8-Oct-2024

* To do:
  + Find good host sequences so that I can remove host reads from the metagenomes.
  + Run snakemake workflow to trim reads and retrieve host reference sequences.
  + Create script to remove host reads from the fastq files.
  + Update the main README.md file
  + Check the reference file to see if they all end in "\_mPhyDis1.pri.v3\_genomic.fna". If not, need to figure out how to reference them in the params section of the rule remove\_host\_reads in readQC.smk

### Find good host sequences so that I can remove host reads from the metagenomes.

Below, I have listed the sample name (host) along with the species name and the accession number for the reference seuquence:

metadata={'sample': ['NBS1051F-30-542494038','NBS1051F','NBS1079E-30-542494038','NBS1079E-30-550131958','080613-8-pd','080813-19-pd','080813-7-mc','080913-2-mc','080913-7-ph'],  
 'species': ['Pteronotus parnellii','Pteronotus parnellii','Myotis elegans', 'Myotis elegans','Phyllostomus discolor','Phyllostomus discolor','Mimon crenulatum','Mimon crenulatum','Phyllostomus hastatus'],  
 'accession': ['GCA\_036768555.1','GCA\_036768555.1','NaN','NaN','GCA\_004126475.3','GCA\_004126475.3','NaN','NaN','GCA\_019186645.2']}  
  
# Create DataFrame  
metadata = pd.DataFrame(metadata)  
print(metadata)

sample species accession  
0 NBS1051F-30-542494038 Pteronotus parnellii GCA\_036768555.1  
1 NBS1051F Pteronotus parnellii GCA\_036768555.1  
2 NBS1079E-30-542494038 Myotis elegans NaN  
3 NBS1079E-30-550131958 Myotis elegans NaN  
4 080613-8-pd Phyllostomus discolor GCA\_004126475.3  
5 080813-19-pd Phyllostomus discolor GCA\_004126475.3  
6 080813-7-mc Mimon crenulatum NaN  
7 080913-2-mc Mimon crenulatum NaN  
8 080913-7-ph Phyllostomus hastatus GCA\_019186645.2

* I already have the script written to download sequences NCBI, but the genome of some host have not been sequenced and are not found on any data base. For instance, neither Myotis elegans nor Mimon crenulatum have had their genomes sequences.
  + However, according to \*\*\*\*\*\* Mimon crenulatum is not an actual species. It is actually a clade that is composed of multiple species. One such species is M. crenulatum keenani. I found that the database 'bat1k' has the genome of M. crenulatum keenani sequenced. So I will retrieve the sequence from there.
  + I am currently waiting on bat1k to approve my membership so I can download the sequence data I need.

Below, I have updated the table to reflect this addition as well as added a column to identify where each genome will be taken from:

metadata\_update={'sample': ['NBS1051F-30-542494038','NBS1051F','NBS1079E-30-542494038','NBS1079E-30-550131958','080613-8-pd','080813-19-pd','080813-7-mc','080913-2-mc','080913-7-ph'],  
 'species': ['Pteronotus parnellii','Pteronotus parnellii','Myotis elegans', 'Myotis elegans','Phyllostomus discolor','Phyllostomus discolor','Mimon crenulatum','Mimon crenulatum','Phyllostomus hastatus'],  
 'accession': ['GCA\_036768555.1','GCA\_036768555.1','NaN','NaN','GCA\_004126475.3','GCA\_004126475.3','NaN','NaN','GCA\_019186645.2']}  
  
# Create DataFrame  
metadata\_update = pd.DataFrame(metadata\_update)  
print(metadata\_update)

sample species accession  
0 NBS1051F-30-542494038 Pteronotus parnellii GCA\_036768555.1  
1 NBS1051F Pteronotus parnellii GCA\_036768555.1  
2 NBS1079E-30-542494038 Myotis elegans NaN  
3 NBS1079E-30-550131958 Myotis elegans NaN  
4 080613-8-pd Phyllostomus discolor GCA\_004126475.3  
5 080813-19-pd Phyllostomus discolor GCA\_004126475.3  
6 080813-7-mc Mimon crenulatum NaN  
7 080913-2-mc Mimon crenulatum NaN  
8 080913-7-ph Phyllostomus hastatus GCA\_019186645.2

### Run snakemake workflow to trim reads and retrieve host reference sequences.

* JobID: 11161798:
  + Apparently this version of snakemake can not take in bash script in the arguement "script:". I got the following error:
    - "Unsupported script: Expecting either Python (.py), R (.R), RMarkdown (.Rmd) or Julia (.jl) script."
  + I added the script path to "params" and changed the argument "script" to "shell". Going to try again.
* JobID: 11161965
  + I made a mistake by giving the same output read file for the fr1 and fr2 slot in the in the shell section of the clean\_raw\_reads rule in the file readQC.smk.
  + Made the required change and running again.
* JobID: 11162659
  + For most of the samples, the adapter and phix removal worked. However, for some reason the temp fq file for sample 080913-2-mc isnt being created correctly or some other error is happening with it as snakemake is removing it and faulting out.
  + I am troubleshooting now:
    - Going to set up the rule to where it will create a log so I can use that to hopefully troubleshoot this sample.
* JobID: 11163418
  + I noticed that I was having the script place the trimmed reads in the 'tmp' directory. I have changed that in the script.
  + Once this job has stopped I need to:
    - mv the trimmed fq files to the clean\_reads directory.
  + This job failed as well for the same reason as the previous job. It said that line 23 in trimming.sh was not a command. Problem is I had already modified that script and was unable to figure out what line it was talking about. I am trying to run again to see what will happen.
* JobID: 11163492

### Create script to remove host reads from the fastq files.

* While I am currently creating this script, it will not be able to be tested until bat1k has given me access to their data base.
  + Script name: host\_read\_removal.sh
  + path: workflow/scripts/