DE NOVO GENOME ASSEMBLY

We will analyze Illumina reads deriving from this paper:

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3960531/

Step that we will do:

1. Download reads

2. Quality control (explain which command you used and describe different plots you see):

3. Trimming (explain the command that you used with the parameters):

4. Assembly with SPADES

```
#Install

conda/mamba/micromamba install -c bioconda spades=3.15.5

##command


# Question: What's a kmer? In which analysis could be used?
How many contigs have you obtained?
```

5. Evaluation of the assembly with QUAST (find a solution for the hypothetical command):

```
#QUAST
conda create -n quast
conda install -c bioconda quast

conda activate quast
```

Question: What is N50? And L50? Describe these parameters.

6. Evaluation of the assembly with BUSCO:

#BUSCO
https://academic.oup.com/bioinformatics/article/31/19/3210/211866

##Try to see if the following command works.
conda install -c bioconda busco

Otherwise, we could install it through following command
conda env create -f busco.yaml

conda activate busco

find the right dataset to use
busco --list-datasets

##BUSCO command
busco -m genome -l lineage -c 4 -f -o busco_output -i scaffolds.fasta

CheckM

It is a software package designed to evaluate the quality and completeness of microbial genome assemblies based on the presence of single-copy marker genes.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4484387/

#INSTALL

conda install -c bioconda checkm-genome

#Create a directory and copy the fasta file inside that directory.
##Note: you have to rename the contigs.fasta in order that it finishes with .fna

checkm lineage_wf -t 2 folder_with_assembly CheckM_out > CheckM.txt

Inspect the CheckM.txt

RUN BUSCO again with a more appropriate (specific) lineage
Explain the differences between the BUSCO analysis using bacteria_odb10 dataset and a more specific dataset.

Only if BUSCO analysis doesn't work on your computer check the following

Filter fasta

mamba/micromamba/conda install -c bioconda seqkit
seqkit seq -m 1000 --remove-gaps contigs.fasta > filtered_contigs.fasta

Go on https://usegalaxy.org/ and do BUSCO analysis with filtered_contigs.fasta
But before check the taxonomy on NCBI, writing the name of the species.

Create an account on the Galaxy Server

1) **Upload filtered_contigs.fasta**
2) **BUSCO on Galaxy Server**

If you encounter any issues or if the BUSCO tool does not work on Galaxy, follow the instructions to rename and send your contigs file to me.

- **Rename your file to** `contigs_Name.fasta`**.**
- **Send the renamed file to my mail for further analysis.**
- **I will run the BUSCO tool on the server.**

If you need specific help with any of these steps or encounter issues, let me know!