```
### Script 2: Identification of taxonomy pipeline

## Author: Shalvi Chirmade
## Date created: July 4, 2022

# If the user would like to edit anything in this file, the user will
need to copy this file and change the permissions:
        # Copy file with a different name (for example:
edit_script2_mac.txt)
        # Change permissions of the file by running this command:
                # chmod a=rwx edit_script2_mac.txt

#
---------------------------------------------------------------------

# This script is for identification of taxonomy from whole metagenome
sequencing using the software:
        # 1. MetaPhlAn

# This script should be utilized after running the necessary steps
from script1_cc.txt. The last file created from fastq-join labeled
sample_merged.fastq.gz for each sample is to be used as input for
MetaPhlAn.


# All lines beginning with hashtags are comments, they are
explanations of when and how to use each command including what each
argument in the command corresponds to. This way, the arguments can be
manipulated by the user if required.

# All lines that do no begin with hashtags are lines of code that need
to be run. Once copied into the terminal, they can be edited as
required.

#
---------------------------------------------------------------------

# This script will consist of Part A and Part B:
        # Part A: How to install the necessary software required to
run this analysis. This part can be skipped if the software has
previously been loaded on your computer.
                # Python
                # Miniconda
                # Biopython
                # Apple M1 chip alteration
                # MetaPhlAn
        # Part B: How to run MetaPhlAn on sample_merged.fastq.gz

#
---------------------------------------------------------------------
```

```
## PART A: Installing required software.

# Please ensure you are using a Mac or Linux computer otherwise have Ubuntu installed on your Windows along with the latest version of Python and alter the download steps accordingly.


## PYTHON

# Use this link to download the latest version of Python https://www.python.org/downloads/
# Make sure you have all the requirements before opening the installer.
# Once downloaded, check if it has been installed properly by running the command:

python3 --version

# The output should be the software name and version number, for me it is:
        Python 3.9.12


## MINICONDA

# Use this link to download the latest version of Miniconda https://docs.conda.io/en/latest/miniconda.html
# Make sure you have all the requirements before opening the installer. You can choose the installer based on the version of Python installed, if not the newest.
# Use the "pkg" installer for macOS depending on whether your computer has the Intel/Apple M1 chip.
# Once downloaded, check if it has been installed properly by running the command:

which conda

# The output should be the file path for the location of the software, for me it is:
        /Users/shalvichirmade/miniconda3/bin/conda


## BIOPYTHON

pip install biopython


##  Apple M1 chip computers
```

```
# If your computer uses the Apple M1 chip, you will have to execute
these lines of code before moving forward.
# https://github.com/Haydnspass/miniforge#rosetta-on-mac-with-apple-
silicon-hardware

CONDA_SUBDIR=osx-64 conda create -n mpa
conda activate mpa
conda env config vars set CONDA_SUBDIR=osx-64
conda deactivate

# We will be using the mpa conda environment for our MetaPhlAn
analysis.
# If you run into this error for other software installations, you
will have to repeat these steps in the environment you have created
for running those particular analyses.


## METAPHLAN

# More information about this software can be found at:
        # https://github.com/biobakery/MetaPhlAn/wiki/MetaPhlAn-3.0
        # https://github.com/biobakery/biobakery/wiki/metaphlan3

# Check which version of Python you have in your conda environment:

conda list python -f

# Download MetaPhlAn and all its dependencies in a conda environment
(we have named it "mpa"):

conda activate mpa
conda install -c bioconda metaphlan

# A lot of packages will be downloaded, the terminal will be updated
based on the recurring downloads
# When prompted to proceed, press "y" and then "Enter"
# The last output to the screen should say: "Executing transaction:
done"

# To make sure MetaPhlAn has successfully downloaded in your conda
environment, run the command.

which metaphlan

# The output to the terminal for the location of MetaPhlAn should be
displayed, mine looks like this:
        /Users/shalvichirmade/miniconda3/envs/mpa/bin/metaphlan


# Next we need to install the Bowtie databases required by MetaPhlAn.
```

```
Please allocate a specific directory for these files; it takes about 3
GB of space. The path to my directory looks like this:
        /Users/shalvichirmade/Documents/MBinf/BINF 6999/Weston
Project/MetaPhlan 3.0

# Here is the command to download the databases, please edit
accordingly. It took about 15 minutes to be completed. The terminal
will output the progression of the download.

metaphlan --install --bowtie2db .

# . - is the current directory, please edit based on the path where
you have created the specific directory to store these files or you
can migrate to this directory using "cd" and keep "." in your argument
# The last output to the terminal will be:
        Download complete
        The database is installed

# Exit out of the conda environment after you are done.

conda deactivate


#
---------------------------------------------------------------------

## PART B: Running MetaPhlAn.

# First, enter into the conda environment we created:

conda activate mpa

# Make sure the command exists in the environment:

which metaphlan

# My output:
        /Users/shalvichirmade/miniconda3/envs/mpa/bin/metaphlan

# Navigate, using "cd" into the directory containing the Bowtie
database files. Run the command:

metaphlan sample_merged.fastq.gz --input_type fastq --bowtie2db . --
nproc 2 > sample_merged_profile.txt

# sample - replace with your sample/file name
# sample_merged.fastq.gz - if your input files are not in the same
directory, enter the file path here
# . - as we are in the directory containing the database files. If you
are not, then add the file path here
```

```
# --nproc - is telling the command to use two cores of my computer for
processing. Before running this command, find the number of cores your
computer has. Mine has four, so using two is appropriate. Do not use
the full capacity of your computer; if you only have a two cores
(DualCore) then change 2 to 1
# > - this is telling the command to save the output file in the same
directory. If you want the file saved in another folder, add the file
path before the output file name. For example: ../filename/
sample_merged_profile.txt
# This command took about 8 minutes to run without any output messages
to the terminal.
# At the end, the output to the terminal for me said this:
        WARNING: The metagenome profile contains clades that represent
multiple species merged into a single representant.
        An additional column listing the merged species is added to
the MetaPhlAn output.

# If you received this same message, this is the explanation:
# The meaning of this warning was found on https://
forum.biobakery.org/t/unexpected-output-format/658
# "These are from MetaPhlAn, they just inform you that some species
found can have “alternative” taxonomies (the list of species in the
additional_species column). All the species listed under
additional_species are not represented by any markers but they were
found to be <5% ANI distant from the “reference” species
(clade_name)."


# We can take a look at the output files created to make sure they are
not blank and we have what was expected. The "less" command allows you
to look at the files: press "Enter" to see more lines, press "q" to
quit the file. The "wc -l" command tells you the number of lines in
the file.

less -S sample_merged_profile.txt
wc -l sample_merged_profile.txt
less -S sample_merged.fastq.gz.bowtie2out.txt
wc -l sample_merged.fastq.gz.bowtie2out.txt

# Bowtie output contains the intermediate mapping results to unique
sequence markers.


# If you have more than one sample and would like to analyze them
together, their output files can be merged together with this command:

merge_metaphlan_tables.py *_profile.txt > merged_abundance_table.txt

# * - uses all the files that end with _profile.txt in the directory.
Make sure the sample files you want to merge are the only ones in this
```

```
directory otherwise specify the files you want to combine and their
file path
# > - is the name of the output file, this can be edited and the file
path changed based on your needs
# Can edit the output file name accordingly


# Deactivate conda environment after you are done running the
analysis.

conda deactivate


#
---------------------------------------------------------------------

# End of Script 2
```