```
### Script 1: Pre-preprocessing pipeline of whole metagenome sequences
(WMS)

## Author: Shalvi Chirmade
## Date created: June 27, 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_script1_cc.txt)
        # Change permissions of the file by running this command:
                # chmod a=rwx edit_script1_cc.txt

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

# This script is for the pre-processing steps to be run in Compute
Canada using the software:
        # 1. FastQC - Checking the quality of the sequences received
from MiGS
        # 2. Trimmomatic - Trimming the sequences based on the quality
of the reads analyzed
        # 3. fastq-join - Combining the forward and reverse reads into
a single file for use by 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.

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

# How to enter the Compute Canada Graham Cluster:
# Replace "user" with your username.

ssh user@graham.computecanada.ca


# Enter into your desired directory.

cd scratch/

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

```
# Create a folder for each of these steps, so the data can be
organized and easily found.

mkdir fastqc trimmomatic fastqjoin


# To look at the contents of your file, run this command:

ls -l

# Add the folder containing the WMS files into this directory. So
currently, your working directory should contain the subsequent
directories: WMS, fastqc, trimmomatic, fastqjoin

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

# Step 1: checking the quality of the WMS file using the software
FastQC.
# Make sure to use the newest version of fastqc, this can be done
using the first command. Replace the version number with the latest
version for you in the second command. Spider allows you to read more
about the version stated.

module keyword fastqc
module spider fastqc/0.11.9

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

# You will have to enter into an interactive node before running each
step. The interactive node allows you to carry out simple commands
that do not require high computational power.

salloc --time=1:0:0 --ntasks=2 --account=def-tvanraay

# --time= is given by hours:minutes:seconds
# --ntasks= is the number of MPI processes
# --account=def- is the user your account is under: tvanraay is Terry
Van Raay's username
# Additional arguments can be added if necessary, see https://
docs.alliancecan.ca/wiki/Running_jobs

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

# If you have more than ten samples, the time can be adjusted for
1:30:0 or another appropriate time.
# Now begin the first pre-processing step. Start by loading in the
```

```
latest version of the module.

module load fastqc/0.11.9

# To confirm that the module has loaded.

module list

# Make sure you are in the folder that contains the raw WMS files. If
you are not, manoeuvre your way back by using:
cd

# Run the command. Replace "sample" with your appropriate file names.
Make sure to include both the forward (R1) and reverse (R2) files for
each sample being analyzed.

fastqc sample_R1_001.fastq.gz sample_R2_001.fastq.gz -o ../fastqc

# -o argument is stating where the output file is being saved: here is
it is in the fastqc directory
# ../ is telling bash to go back a directory where it will find
fastqc. It can be changed accordingly.

# Repeat this last command for all the samples. It will rename the
output file based on the input filenames. Once all the files are
completed, make sure to exit the interactive node by saying:

exit

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

# The .html files created by fastqc now need to copied onto your
computer, so that they can opened and analyzed in an internet browser.
# Open a new terminal window and make your way using "cd" into the
folder where you want to copy the .html files. Once you are in the
appropriate folder, run this command.

scp user@graham.computecanada.ca:~/scratch/filename/fastqc/*.html .

# user - needs to be replaced with your username
# filename - needs to replaced with where the fastqc directory is
located
# *.html - this will copy every .html file in the fastqc directory
onto your computer. If a specific file is required, replace * with the
appropriate file
# . - the dot at the end is stating you want the files to be copied in
the directory you are currently in. If another directory is required,
put the path here
```

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

# After analyzing each of the .html files for your sequences, take note on the number of bases you want trimmed for each end of each sequence file.

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

# Step 2: trimming each sequence based on the quality of reads using the software Trimmomatic.
# MiGS, where the sequencing is carried out, has already trimmed the adapters before sending the sequence files. Please check the documentation to make sure this has been done, if not, you will have to trim the adapters as well. This can be seen in your fastqc output as well.

# Make sure you are in the directory containing the raw WMS files. If not, edit the command to add the path.

# Check the latest version of the software:

module spider trimmomatic

# Enter into the interactive node. Edit time and software version if necessary.

salloc --time=1:0:0 --ntasks=2 --account=def-tvanraay
module load trimmomatic/0.39
module list

java -jar $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar PE -phred33 sample_R1_001.fastq.gz sample_R2_001.fastq.gz ../trimmomatic/sample_R1_001-pe.fastq.gz ../trimmomatic/sample_R1_001-se.fastq.gz ../trimmomatic/sample_R2_001-pe.fastq.gz ../trimmomatic/sample_R2_001-se.fastq.gz LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50

# $EBROOTTRIMMOMATIC/trimmomatic-0.39.jar - is the path to trimmomatic in Compute Canada
# sample - replace with appropriate filename
# -phred33 - using phred +33 scores for base encoding (use -phred64 if using +64 scores)
# The first files in the command are the R1 and R2 files of the sample
# The last four files in the command are the paired and unpaired reads for both the forward and reverse sequence files
# SLIDINGWINDOW - range where sequence is cut based on average read quality --> windowsize:requiredquality
# MINLEN - will remove reads that are lower than this basepair length
# LEADING - low quality bases that will be removed from the beginning;
```

```
value is the quality score
# TRAILING - low quality bases that will be removed from the end;
value is the quality score
# CROP - removes stated number of bases from the end regardless of
quality
# HEADCROP - removes stated number of bases from the beginning
regardless of quality
# ILLUMINACLIP - this would be added if adapters needed to be trimmed.
Read the manual to see how to identify which argument to add

# If you would like to edit any of the arguments, here is the manual:
https://thesequencingcenter.com/wp-content/uploads/2019/04/
Trimmomatic_Manual_v.0.32.pdf

# The output will state how many reads were paired, unpaired and
dropped.

# Remember to "exit" out of the interactive node once completed. If
your time in the node has run out, re-enter the "salloc" command, re-
load the required modules, and continue on.

# As the sequence quality from MiGS is usually very high, the amount
of paired reads was over 97% in my experience. Hence, only the paired
forward and reverse reads will be used for each sample moving forward.
As a side note, only paired reads can be "merged" together as it looks
for overlapping bases. If you want to include the unpaired reads in
the analysis as well, you can concatenate the files together
afterwards. Look for (****) in the later steps for an example.

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

# Step 3: combining the paired forward and reverse reads into a single
file. This file will be used by MetaPhlAn.

# Make sure to be in the directory where the "pe" files are kept (this
should be in the trimmomatic directory). If not, edit the path to the
file accordingly.

# Check for the latest version for the software.

module spider fastq-join

# Enter into the interactive node. Edit time and software version if
necessary.

salloc --time=1:0:0 --ntasks=2 --account=def-tvanraay
module load fastq-join/1.3.1
module list
```

```
fastq-join sample_R1_001-pe.fastq.gz sample_R2_001-pe.fastq.gz -o ../
fastqjoin/sample_%.fastq.gz

# FYI, this step takes a few minutes and does not print anything to
the terminal for confirmation of success. It may seem like it's not
working, but the output statements will show up within five minutes or
so. The output statements should state:
        # Total reads
        # Total joined
        # Average join len
        # Stdev join len
        # Version
# sample - replace with appropriate name
# The first two input files are the paired-end "pe" for both R1 and R2
of the sample
# -o is telling the command the output file name
# ../ - is telling the command to go up a directory where the
fastqjoin folder will be found. If this is different for you, please
edit accordingly
# % is a placeholder where the three files created will be renamed
accordingly inplace of this symbol
# The three files created: joined, forward unjoined and reverse
unjoined

# Remember to close the interactive node by typing "exit" once all
samples are completed.

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

# A merged file containing all three of the output files will be
created using a simple bash command. Make sure you are in the
directory containing the fastq-join output files ONLY (this should be
called fastqjoin if the steps in the beginning were followed).

cat sample*.gz > sample_merged.fastq.gz

# sample - replace with appropriate name
# * - means "all". Every file with the name starting with sample will
be included
# > - telling the command to output the file with this name

# Moving forward, the merged.fastq.gz files will be used by MetaPhlan
for each sample.

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

# Copy the zipped merged files for each sample onto your personal
computer.
```

```
# Open a new terminal window and make your way using "cd" into the
folder where you want to copy the merged files. Once you are in the
appropriate folder, run this command.

scp user@graham.computecanada.ca:~/scratch/filename/fastqjoin/
*merged.fastq.gz .

# user - needs to be replaced with your username
# filename - needs to replaced with where the fastqjoin directory is
located
# *merged.fastq.gz - this will copy every merged.fastq.gz file in the
fastqjoin directory onto your computer. If a specific file is
required, replace * with the appropriate file
# . - the dot at the end is stating you want the files to be copied in
the directory you are currently in. If another directory is required,
put the path here

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

# End of Script 1
```