University of Guelph

Bioinformatics

Assignment 1:

**Exploration of Species Composition of Gut Microbiomes in Healthy Patients**

BINF 6410: Bioinformatics Programming

[Code Document]

Dilma Karunathilake (0969209)

Wednesday, October 9, 2024

#I attempted to use the command **module load blast+/2.14.0**, however, this led to an error, and it was suggested by the terminal that I use **module spider blast+/2.14.0** instead which ended up working.

A screenshot of a computer

Description automatically generated

#I downloaded the reference.ft file (renamed A1\_reference.ft) from CourseLink into my Ubuntu file system and then used the **scp** command to secure copy the file to my compute Canada account on PuTTY. A screenshot of a computer program

Description automatically generated

#I downloaded the two FASTQ file for the distal lumen and mucos from the NCBI database. I then used **rsync -e ssh** to secure (provided by the -e ssh option) copy these large files from my local file system (Ubuntu) to my compute Canada account on PuTTY.

A screenshot of a computer

Description automatically generated A screenshot of a computer

Description automatically generated





#All three required files are now available on PuTTY. After creating a directory called **Assignment1** with **mkdir**, I moved (with **mv**) the three files into the directory. Using **ls -l**, I listed the contents of the directory to make sure that they are there.

A screenshot of a computer

Description automatically generated

#Using **gunzip**, I decompressed the two **.gz** (compressed) FASTA files and then used **ls -l** to list them to make sure they had been successfully decompressed.

A screen shot of a computer

Description automatically generated

#Using the **makeblastdb** command (), the **reference.ft** subject file is converted into a set of files that are indexed (allowing for quick searches) that contain the same information (Oregon State University, n.d.; National Center for Biotechnology Information [NCBI], 2008).

**A close-up of a text

Description automatically generated**

A screenshot of a computer program

Description automatically generated

#Using **ls -l** to view the files (that make up the database) created. These files contain nucleotide sequences (due to the .n\_\_ file extensions) (Oregon State University, n.d.).

A screenshot of a computer

Description automatically generated

#Using **blastn**, compared the sequences from **DL\_sequences.txt** (for the distal lumen) and **PM\_sequences.txt** (for the proximal mucosa) against the BLAST database created from the **reference.fasta** file (National Center for Biotechnology Information [NCBI], 2008).

A close-up of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

#Used **ls -l** to see the new files created. **DL\_match.txt** and **PM\_match.txt** hold the matches between the **reference.ft** BLAST Database and each of the SRR files. For each query sequence, the best match (top hit based on the bit score) is saved in the output files **DL\_match.txt** and **PM\_match.txt**.

A screen shot of a computer

Description automatically generated

#A string of commands is used to create the files **DL\_genera\_list** and **PM\_genera\_list** (which will contain unique genera lists) from the second columns of the **DL\_match.txt** and **PM\_match.txt** files (which contains matched sequence IDs).

A close-up of a diagram

Description automatically generated

A screen shot of a computer program

Description automatically generated

#A string of commands is used to create the file **DL\_PM\_common.txt** (which will contain the common genera between **DL\_genera\_list** and **PM\_genera\_list**) from those list files (which contains the unique genera list from the DL and PM samples).

A close-up of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

# A string of commands is used to create the two files: **DL\_unique.txt** and **PM\_unique.txt** in the pwd.

The final result of the strings of commands used to create both files:

* **DL\_unique.txt**: Identifies the genera that are unique to the DL sample but not in the common genera file (i.e., not shared with PM).
* **PM\_unique.txt**: Identifies the genera that are unique to the PM sample but not in the common genera file (i.e., not shared with DL).

A close-up of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

A screen shot of a computer

Description automatically generated

[Transferring the Five Files to Submit]

#Created a directory called **FiveFiles** and copied (using the command **cp**) the necessary five files into this directory from the **Assignment1** directory.

#Using **rsync** and the options -**a** (archive option to allow for recursive copying – allowing entire directories to be copied, -**v** (verbose option to showyou the file(s) being copied and how the new copy will look, and -**e ssh** (allows for secure copying), the entire FiveFiles directory was copied to my Ubuntu file system from computecanada.



A screenshot of a computer

Description automatically generated

**References**:

National Center for Biotechnology Information. (n.d.). *Trace NCBI: Run browser for*

*SRR6288933*. <https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR6288933&display=download>

National Center for Biotechnology Information. (n.d.). *Trace NCBI: Run browser for*

*SRR6288926*. <https://trace.ncbi.nlm.nih.gov/Traces/?view=run_browser&acc=SRR6288926&display=download>

Oregon State University. (n.d.). *Command-line BLAST*. In *Computational biology*. Open Oregon

State. <https://open.oregonstate.education/computationalbiology/chapter/command-line-blast/>

National Center for Biotechnology Information (NCBI). (2008). *BLAST® Command Line*

*Applications User Manual* [Internet]. Bethesda, MD: National Center for Biotechnology Information (US). Available from <https://www.ncbi.nlm.nih.gov/books/NBK279690/>