1. **Bioinformatics analysis of metagenomics (*adapted from Oksana Lukjancenko and Pimlapas Leekitcharoenphon, Technical University of Denmark; pile@food.dtu.dk*)**

*The objectives of this section are to run MGmapper and understand its output.*

Follow the instructions below to run MGmapper command line. Now you will run MGmapper with the sample that you trimmed in section II (Quality control) - Sample.R1.trim.fastq.gz and Sample.R1.trim.fastq.gz.

1. **Brief MGmapper description**

The MGmapper software performs three main steps: pre-processing of raw reads, mapping of reads to the reference sequence databases, and post-processing of mapping results.

The pre-processing step is done using cutadapt. Cutadapt performs common adapter removal and trims the low-quality ends from reads before adapter removal (default cutoff value of 30).

**NB: Since you have already trimmed the samples in section II using bbduk, you can skip this pre-processing step in MGmapper!**

In the second step, already trimmed reads are aligned to a pre-defined set of reference sequence databases using bwa mem. Samtools are used to remove singletons and filter reads where neither a read nor its mate is mapped. Reads can be mapped in either fullmode or bestmode. In bestmode, mapping is performed against all databases, simultaneously, and for each read pair the best hit among all alignments is chosen. A pair of reads is considered as a best hit only if the sum of the alignment scores (SAS) is higher than any SAS values from other database hits (*for more details on best mode and full mode, read the complentary document in module 2*).

In the last, post-processing step, alignments are filtered based on a matches/mis-matches threshold, which is specified as a fraction or absolute number (default matches/mis-matches threshold is 80%). Additionally, a combination of four criteria (I-IV) is used to identify a positive taxonomy annotation.

1. Minimum ReadCount of 10
2. Mismatch ratio < 0.01, defined as MismatchesNucloetides
3. S\_Abundance, the size normalized abundance > 0.01
4. Unique read count fraction > 0.5%, defined as ReadCount uniq/ReadCount
5. **Running MGmapper**

During this exercise you will run MGmapper on a small sample, which you have trimmed in section 2 of the exercise. The sample files are Sample.R1.trim.fastq.gz and Sample.R1.trim.fastq.gz and they are located in the shared folder /home/user/Vboxshare/

Commands shown in blue are the ones you are going to use during the exercise.

Set-up the MGmapper enviroment by running the following command in the (bash) shell:

export MGmap\_HOME=/home/user/tools/MGmapper

In the folder /home/user/Vboxshare/ create a MGmapper folder, where MGmapper results will be kept.

cd /home/user/Vboxshare (make sure you are in the right folder)

mkdir MGmapper

MGmapper can map reads against multiple databases. In this example you are going to use 2 databases: *ResFinder* to be mapped in fullmode and *Bacteria* to be mapped in bestmode. A database list can be found in /home/user/tools/MGmapper/databases.txt. You can note that ResFinder database is number 1 in the list and Bacteria database is number 2. Those numbers are going to be used in the MGmapper running command line.

MGmapper can be run on both pair-end reads (MGmapper\_PE.pl) and single-end reads (MGmapper\_SE.pl). MGmapper\_PE.pl and MGmapper\_SE.pl options can be seen using help (-h) option:

perl /home/user/tools/MGmapper/MGmapper\_PE.pl -h

perl /home/user/tools/MGmapper/MGmapper\_SE.pl -h

Here you will run MGmapper for pair-end reads **(remember to paste the command in a single line):**

perl /home/user/tools/MGmapper/MGmapper\_PE.pl -i /home/user/Vboxshare/Sample.R1.trim.fastq.gz -j /home/user/Vboxshare/Sample.R2.trim.fastq.gz -d /home/user/Vboxshare/MGmapper/Sample -F 1 -C 2

In this command, the following options are used:

-i*= forward reads*

-j *= reverse reads*

-d *= output folder*

-F *= full mode*

-C *= best mode*

1 *= Resfinder database*

2 *= Bacteria database*

MGmapper run might take some time. Do not write anything in the command line until the run is finished.

**NB: You do NOT need to run this command!**

Alternatively, a MGmapper run for single-end reads can be done with the command (in a single line):

perl /home/user/tools/MGmapper/MGmapper\_SE.pl -I /home/user/Vboxshare/Sample.R1.trim.fastq.gz

-d /home/user/Vboxshare/MGmapper/Sample\_se -F 1 -C 2

1. **Understanding the MGmapper output**

After MGmapper run is done, it is important to understand its output.

Enter the folder where the MGmapper run output is stored. It was previously defined by the -d option:

cd /home/user/Vboxshare/MGmapper/Sample

List the content of the MGmapper run output:

ll

A successful run will contain the following folders and files (figure a). Files and folders in the Sample folder can also be seen using Files from the Launcher menu (figure b).

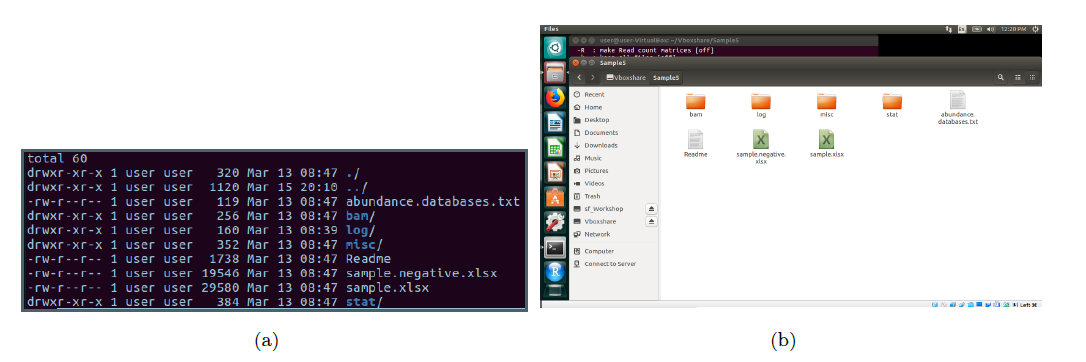
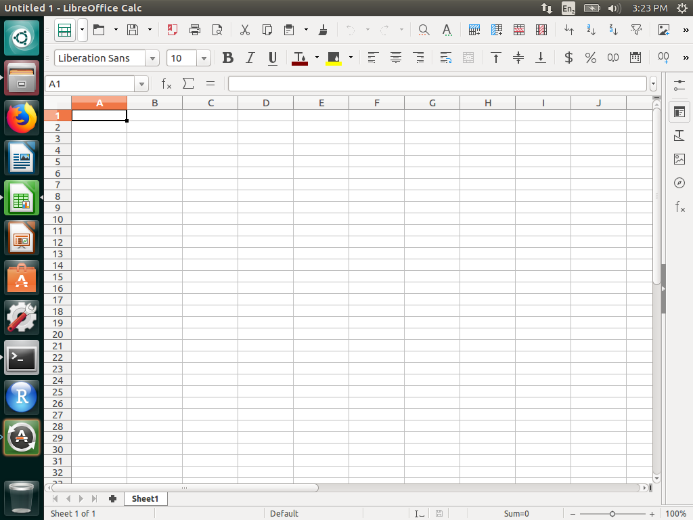
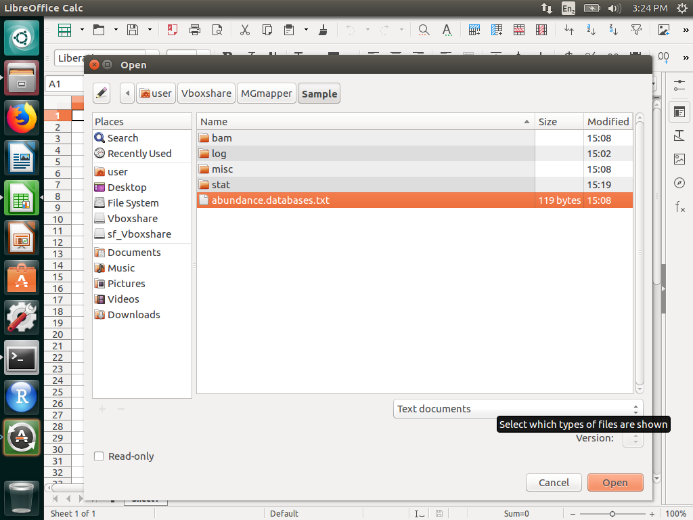
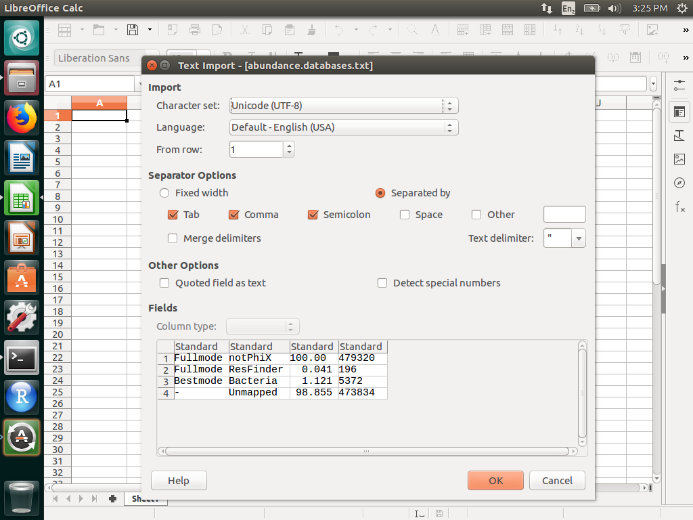


Figure: Contents of the MGmapper output folder.

Look into the output files *(remember you can open .txt files with the command* less *or using LibreOffice Calc (Ubuntu equivalent to Microsoft Excel) – see figure below):*





Output files:

* **abundance.databases.txt** file contains overall mapping information
* misc among others, includes **stat.ResFinder.annot** and **stat.Bacteria.annot** files, which contain read counts for each reference organism/gene in the database, after post-processing step
* stat includes **positive.strain.\*.txt** and **negative.strain.\*.txt files**. Positive include the reference organisms/genes that satisfied the cut-offs of four criteria, described above. Negative include the reference organisms/genes that did not satisfy the cut-offs of those four criteria.

Additionaly to those files, you also find among the MGmapper output the following:

* bam contains files in bam format, the result of reads mapping against reference databases using bwa mem
* log contains log files, where each step of MGmapper run can be seen

QUESTIONS:

1. overall mapping information: How many reads were classified as notPhix?
2. overall mapping information: How many reads were mapped to ResFinder?
3. overall mapping information: What is the percentage of unmapped reads?
4. *misc* output: against how many macrolide resistance genes did the reads map?
5. *stat* output: what is the relative abundance of the bacteria strain *Treponema succinifaciens* DSM 2489 in the sample?

In this exercise, you obtained the output of MGmapper for a single sample. In the next exercise section IV, you will use the output from multiple MGmapper runs, for different samples, combined into one table.