Maria Beatriz Walter Costa: bia.walter@gmail.com
GITHub Repository: `https://github.com/waltercostamb/Profiling_metagenomes`

Tutorial for profiling metagenomes using the Kraken2 tool

With the present tutorial, you will be able to profile metagenomes using the Kraken2 tool (Manual: `http://ccb.jhu.edu/software/kraken/MANUAL.html`) [Wood and Salzberg, 2014]. Figure 1 shows the graphical representation of the pipeline. Note that the pre-processing of the files will change according to the input file you have.

The general steps of the pipeline follow in the itemize environment below, with the details and command lines in the following sections:

1. Check the **available space** at the machine you are running your experiment
2. Check the size of your files, and make sure the files fit in your machine. If they do not fit, you need to run the pipeline for fewer files at a time. Take into account the size of **input** as well as **output** files
3. What **type of data** do you have? Is it pre-processed or raw NGS data? In the example pipeline from figure 1, the NCBI/SRA are raw NGS and the MG-RAST are pre-processed. Depending on the type of data you have, you will have to adapt the pipeline to best suit your needs
4. Now that you made sure you have enough space for running your data, start with the **download**
5. Now that you have adapted the pipeline and know which steps you need to follow, **go to the appropriate step** for you (pre-processing or uniformity filter step?)
6. In this step **you should have pre-processed data**. In figure 1, this corresponds to step "Uniformity filter"
7. Perform the **Uniformity filter** step
8. Proceed to the **profiling with Kraken2** with the Reference Database that is appropriate for you (Kraken2 DB or a customized DB?)
9. Proceed to the filtering of the Kraken2 output. Choose appropriate Domains of life to analyse, and run the script **selectGroups.pl**
10. After you selected the groups you are interested in studying, proceed to the step of **creating a BIOM table** of abundances
11. Afterwards, add taxonomic ranks to your BIOM table **using kraken-biom**
12. **Create abundance matrices** for the taxonomic rank you want (e. g. "p" for Phylum, "c" for class, etc)
13. Transpose the matrix if needed and you are done!

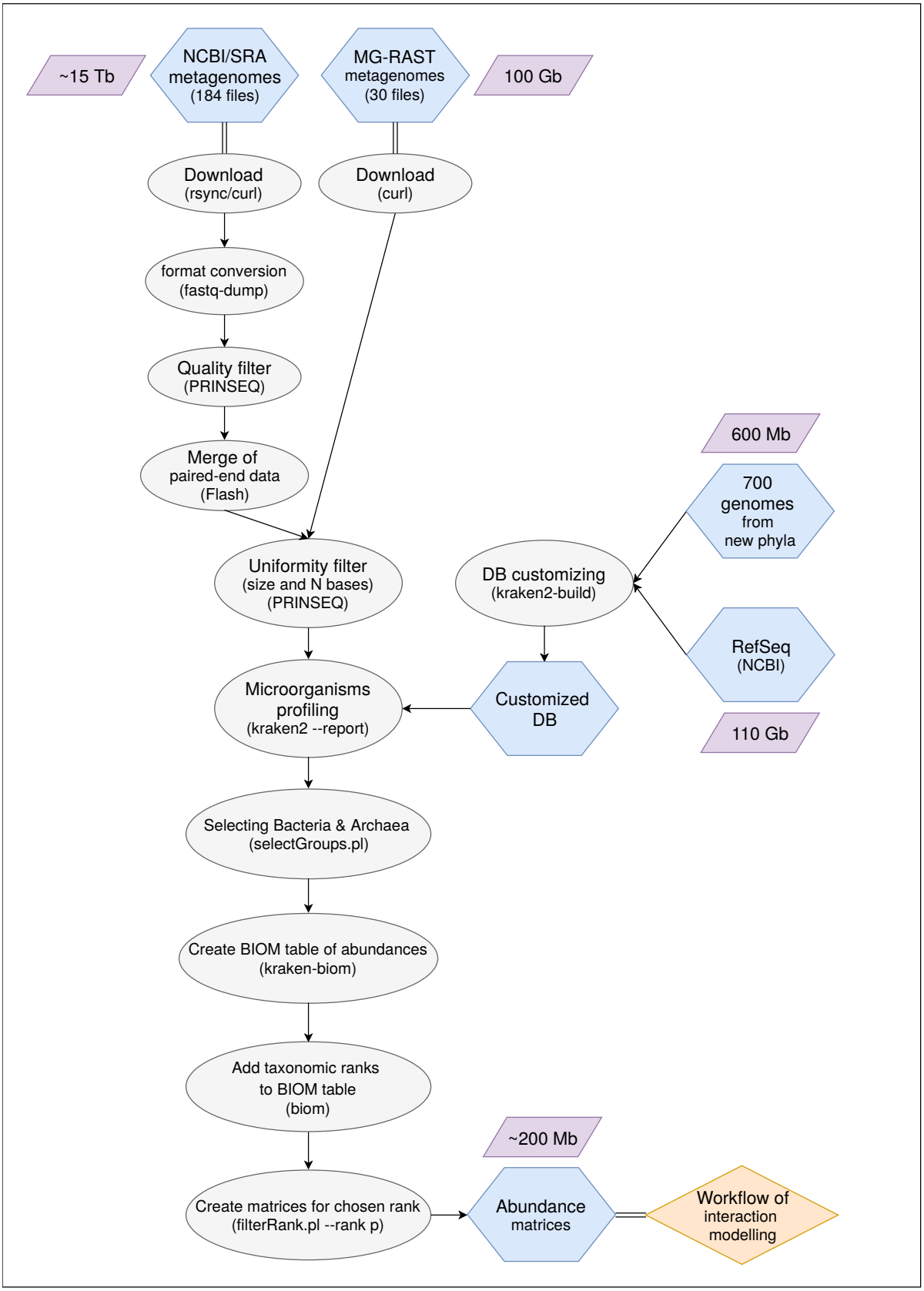

Figure 1: Pipeline of microorganisms profiling of aquifer metagenomes. Note that NCBI/SRA files require pre-processing of quality filters, while MG-RAST files might not require this step. If one retrieves MG-RAST samples from step 299 (as done in the present pipeline), the files have already been filtered at the MG-RAST server.

Cloning the Repository

To clone the repository in a server, try:

```
git clone git@github.com/waltercostamb/Profiling_metagenomes
```

Testing the Pipeline

You can test the pipeline with a small dataset, to better understand the whole process. When you cloned the repository, you should have gotten a folder: **sequencing_examples/**. Inside this folder, you have two metagenomes as examples (downloaded from the MG-RAST server):

1. mgm4529964.3.299.1.fastq
2. mgm4529965.3.299.1.fastq

You can use these two metagenomes to manually do all the following steps (without submitting the scripts and programs to the slurm queue) and obtain a small abundance matrix. After that, the process for multiple files will become a lot easier. In addition, you will not have to wait for queueing time, so the running time of the pipeline will be very small.

To test the pipeline with these example files, you should adapt the general command lines mentioned in each of the steps and run the steps manually.

Download of data from already pre-processed files

The 30 files from MG-RAST (`https://www.mg-rast.org/`) of the pipeline of figure 1 are already pre-processed by the MG-RAST server, so we can follow with the download and uniformity filter. In the example of the figure, I retrieved the files already from step 299 of the MG-RAST pipeline [Meyer et al., 2008], which means they were already filtered by quality and submitted to initial filtering steps at the MG-RAST server.

To download this type of data individually, use the *curl* command and the following general command line:

```
$curl http://api.metagenomics.anl.gov/download/"${i}"?file=299.1 > ${i}.299.1.gz
```

To use this command line for multiple files, adapt script download_mgm_curl.sh to contain your IDs in the array of the script and run with the *nohup*, as below:

```
$nohup bash download_mgm_curl.sh > download_mgm.nohupout &
```

Uniformity filter

Given that the data has been pre-processed for quality filter, you can proceed to the uniformity filter step. You can use the PRINSEQ tool [Schmieder and Edwards, 2011] with the parameters below (general command line for fastq files):

```
$prinseq-lite.pl -verbose -fastq $file -min_len 80 -ns_max_p 2 \
-out_format 1
```

Note that you must specify if your input is FASTQ or FASTA. In the general command line above and in the example below, you have a FASTQ input:

```
$prinseq-lite.pl -verbose -fastq ${i}.3.299.1.fastq -min_len 80 -ns_max_p 2 \
-out_format 1
```

Note that in the example command line above, the variable ${i} is an ID for the file.

To run PRINSEQ for multiple files, you must adapt script *slurm_job_prinseq_single_FASTA.sh* (if you have FASTA files) or *slurm_job_prinseq_single_FASTQ.sh* (if you have FASTQ files). These two scripts submit jobs to the slurm queue. You can run them with the following commands:

For FASTA files:

```
$sbatch slurm_job_prinseq_single_FASTA.bash
```

or for FASTQ files:

```
$sbatch slurm_job_prinseq_single_FASTQ.bash
```

Profiling of microorganisms

After the uniformity filter, we could profile the filtered files using the Kraken2 tool. The filtered files can now be submitted to profiling with the Kraken2 tool [Wood and Salzberg, 2014] using our customized DB, or the appropriate DB. The general command line is:

```
$kraken2 --db customized_Kraken2_DB $filtered_file --output $profiled_file \
-report $report_file
```

The output file $profiled_file is the one that contains reads and microorganisms, while the report file $report_file is the one that contains the abundances of microorganisms. The following steps ("Selection of Groups of Interest", "Creation of BIOM Table of abundances" and others) have to be done with the $report_files as input.

To run the script for multiple files and submit the jobs to the slurm queue, use the following script:

```
 $sbatch slurm_job_kraken2.bash
```

At the end of this step, you will have files *profiled and *report. Follow up with the *report files in the next step.

Selection of Groups of Interest

The profiled files contain all Domains of Life. Since we are generally interested in microorganisms, more specifically Bacteria and Archaea, we need to filter the raw output of Kraken2.

For this, you will use script **selectGroups.pl** with the following general command line:

```
$perl selectGroups.pl --input $report_file --file_groups groups.txt > \
$selected_file
```

File **groups.txt** is a file contaning the keywords for Bacteria and Archaea - the Domains of life we want to include in the further Bioinformatics steps. In case you have other groups, you can change this file.

To get its usage and manual page:

```
$perl selectGroups.pl --help
```

To run this script for multiple files with *nohup* (run it at your $home at the SDU server):

```
$nohup bash select_groups_perl.sh > select_groups_perl.nohupout &
```

Creation of BIOM table of abundances

After we filtered the Kraken2 outputs with the **selectGroups.pl** script, we can use the tools:

- BIOM `http://biom-format.org/`;
- **Kraken-biom** `https://github.com/smdabdoub/kraken-biom`

to manipulate the formats and collapse similar groups of microorganisms (e. g. "Candidatus Peregrinibacteri" and "unclassified Candidatus Peregrinibacteria" as a unique group).

First, you need to check if the tools are installed.

After loading the appropriate modules, you can run the tools with the following general command lines (you can run this at your $home folder):

```
$kraken-biom selected_file1 selected_file2 selected_file3 ... selected_fileN \
-o table.biom --max HIGHER_RANK --min DESIRED_RANK
```

With the arguments:

- HIGHER_RANK being the taxonomic rank that is immediatly superior to your desired rank (e. g. D for "Domain" if your desired rank is Phylum)
- DESIRED_RANK being the desired rank (e. g. P for "Phylum")

After running kraken-biom, you can convert your "kraken-biom" table with the following script:

```
$biom convert -i table.biom -o table.from_biom_with_taxonomy.txt --to-tsv \
--header-key taxonomy
```

After that, you will end up with the output file *table.from_biom_with_taxonomy.txt*, which you will use as input to create your matrix of abundance.

Creation of matrices of abundances

After we collapsed the similar groups and retrieved the taxonomy in the previous step, we obtained file *table.from_biom_with_taxonomy.txt*. We can now chose a specific rank, e. g. Phylum, to proceed with producing the final abundance matrix. For that, you will use the script **filterRank.pl**, that filters files from BIOM chosing only the rank that was specified by the user.

For that, you can use the following general command line:

```
$perl filterRank.pl --input table.from_biom_with_taxonomy.txt --rank p > \
abundance.matrix
```

To get its usage and manual page:

```
$perl filterRank.pl --help
```

With this pipeline, you now have obtained an abundance matrix for microorganisms with the desired rank. This matrix can now be further processed with modelling or other methods.

Database customization - Kraken2

To customize the DB, you should follow the manual from Kraken2:
Kraken manual: `http://ccb.jhu.edu/software/kraken/MANUAL.html`

From the indications of the manual, to customize a database you should: (i) download your genomes of interest (the ones you want to add to the Kraken2 Reference DB); (ii) format the genomes of interest to a format accepted by Kraken2; (iii) add the genomes to the Kraken2 Reference DB, following the manual.

To format the genomes of interest (step (ii)), you can use the script *fromNCBI_ID2taxID.pl*. This script formats genomes downloaded from the NCBI to the Kraken2 format.

If you want to test this script with a small dataset, you can use the example inputs you got with the GITHub Repository. You should have gotten folders: **test_genomes/** and **accessionDB_test/**, which contain respectively (i) the genomes you wish to include in the Reference DB and (ii) the accession files (for the Kraken2 formatting). To custom this small DB, you can run the command below:

```
perl fromNCBI_ID2taxID.pl --accessionFile accessionDB_test --regExpAccessionFile \
accession --fileQuery test_genomes --regQuery fna --outputFolder test_genomes_\
out --jobID test_run --number_sequence 0
```

To get the usage of the script:

```
perl fromNCBI_ID2taxID.pl --help
```

References

[Meyer et al., 2008] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., et al. (2008). The metagenomics rast server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, 9(1):386.

[Schmieder and Edwards, 2011] Schmieder, R. and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27(6):863–864.

[Wood and Salzberg, 2014] Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.