Lab Book

Postdoc in metagenomics

Maria Beatriz Walter Costa

Federal University of Bahia

2018 - 2019

Sumário

Capítulo 1

May 2018

The information of this lab book refer to my analysis on the project *Microbiomes of Aquifers*. I started working on the project previously, reading bibliography on metagenomics, and getting acquainted with the project, which is not mentioned here (written notebook). Informations especially relevant for analysis is described here, along with locations (folders) of working folders, libraries, etc.

1.1 18 - Aquifer microbiomes project - short description

As a postdoc of Professor Pedro Meirelles at the Federal University of Bahia, I will lead the project "Aquifer Microbiomes", which is part of the Serrapilheira Initiative. The main objective of the project is to predict microbiome systems in aquifer water, given a model to be built using a large collection of different metagenomes.

The first step of the project is to develop ami 1.1 of the Serrapilheira project, constituting of data analysis. The list of metagenomes to be analysed is ready, and some initial steps have already been done by some students and collaborators of the project.

This is the link to videochat meetings with Pedro:
`https://appear.in/bia_pedro`

1.2 9 - Skype meeting with Leticia

Today I talked to Leticia, a student that will use the same tools as we are for the project. Her project is not directly related to aquifers, but rather to marine corals microbiomes. Leticia is working with FOCUS and with Kraken to predict microbiome phyla from a collection of metagenomics transcriptomes. She can run FOCUS without difficulties, but Kraken is jammed in one of the running steps, likely because of lack of disk space, which was explicitly written in the output error. I gave directions to her on how to solve the problem, in a file from OSF.

1.3 11 - Model for microbiome prediction - notes on the common meeting

This section is based on the meeting of May 11 with many collaborators of the project. The prediction model is described in aim 2.2 of the Serrapilheira project, and involve genetic programming and metabolic networks, but the details are still to be discussed and established. Based on today's discussion, I have a suggestion to develop the model using Bayesian statistics, similar to the MEBS tool developed by Valerie de Anda, or to my undergraduation project of protein folding.

Given a conditional probability of an environment containing some microbiome $P(B|A)$, one could predict the probability of a certain microbiome $P(A|B)$ of occuring in an environment. This idea is very initial, and needs to be further developed to see its viability. It is based on my project of protein folding prediction, in which we predicted the atomic distance of an atom given the conditional probability of their neighbours.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1.1}$$

In Val's work, they collected several microbiomes know to be involved in Sulfur cycle, and given their characteristics, they could predict if an unknown environment would be rich in Sulfur cycle as well. They also used Bayesian statistics with a simple information measure.

1.4 22 - Bas talk in Jena

I attended Bas' talk in Jena, about metagenomics analysis.

1.5 Plans for the two week internship in Utrecht - meeting with Pedro

In this two weeks I will learn about metagenomics techniques. For that I will process a small sample of the metagenomes that we will process at the Lab in UFBA. The goal here is to follow through the pipeline described in aim 1.1 of the Serrapilheira proposal, in a way that it is automated or semi-automated.

The samples and steps for this work will be discussed in the Netherlands, with the expert team of Bas' to optimize time and get input from their experience working with metagenomics.

Capítulo 2

June 2018

2.1 7 - Access to the Santos Dumont Cluster

I have now an open account to the Santos Dumont Cluster (Petropolis - RJ), opened to me by the Helpdesk of Santos Dumont. I should set up the VPN first, access by ssh and change my password.

Login: maria.costa Password: EBIoDIV

This is the url with the user manual: `sdumont.lncc.br`

2.2 08 - Meeting with Pedro about visit at Utrecht

For my stay at Utrecht, Pedro told me to focus on the assembly part of our pipeline. So I should use as input the reference DB (we have 700 genomes plus the database of Kraken, for instance) and the query metagenomics samples to be probed (we have 135 metagenome files of aquifers to be probed). As an output, we want the annotation of the organisms of the query files. The tools for this should be: Kraken, FOCUS and others.

2.3 12 - First day at Utrecht - installing profiling tools

For the next two weeks, I will be staying as a guest at the University of Utrecht, with Dr Bas Dutilh's group. I will learn metagenomics techniques and start my project for the postdoc.

Today I met with Bastiaan, a PhD student of Bas'. He helped me to create a small subset of 5Gb of the **nr** database from NCBI, that I will work on during my stay. I will work with maya here, so when I go to Salvador, I can easily transfer the files and start running the real data.

I installed the following software, for the binning course tomorrow and for my own project:

- BWA
- samtools

- metabat2
- spades
- checkM (not installed, since it requires 40 Gb memory)
- prokka
- kaiju (tool for taxonomic classification for metagenomics)
- Kraken (tool for taxonomic classification for metagenomics) `http://ccb.jhu.edu/software/kraken/MANUAL.html#installation`
  - To run kraken:

    `~bia/bin/kraken/kraken`
- MMseqs2 (tool for taxonomic classification for metagenomics)
  - To run mmseqs2:

    `~bia/bin/mmseqs2/bin/mmseqs`

Importantly, to make the programs visible in every folder:

```
export PATH="/path/to/dir:$PATH"
source ~/.bashrc
```

2.4 13 - second day at Utrecht - workshop

2.4.1 Workshop on metagenomics binning

Today I participated on a workshop to do the binning of metagenomics samples. This step is composed on assembly, and classification of the assembled contigs into different bins of organisms. After that is done, the function of the different bins can be assessed by retrieving the IDs of the species/contigs and visualization in KEGG for instance.

2.5 14 - third day at Utrecht - working at Bas' group

2.5.1 Journal club

Today we discussed two different articles that talks about profiling tools: FOCUS2 [Silva et al., 2018], a paper in preparation of Bas' group and the CAMI initiative that compares metagenomics tools [Sczyrba et al., 2017].

We discussed several aspects of profiling tools, but we ended up with a plan for me to do this and next week. I will run a couple of tools and compare their results, to run on the data we have. If we decide to benchmark a tool using the CAMI golden dataset, it is available at: `https://data.cami-challenge.org/participate`. These are pretty big (more than 100Gb, so not feasable for me to run in my laptop).

2.5.2 TODO's for next steps

- Analysis folder (maya): *Documents/posDoc/Profiling*

After discussing with Bas' group in the journal club and talking to them about using the tools, I came up with a pipeline of study to apply to a sample metagenomics file and to a reduced reference DB, since I will be running the tools in my laptop, for practical reasons, which has obvious limitations in memory (8Gb) and space (around 40Gb).

FOCUS2 is already available in GITHub: `https://github.com/metageni/FOCUS2`

kaiju: `https://github.com/bioinformatics-centre/kaiju/blob/master/README.md`

I will run the tools and compare the results so we can choose the best to use in our data (Fig. 2.1). If I can reach the end of the analysis, I can bring tables of abundances to run correlations with Julia Engelmann next week (from the Sea Research).

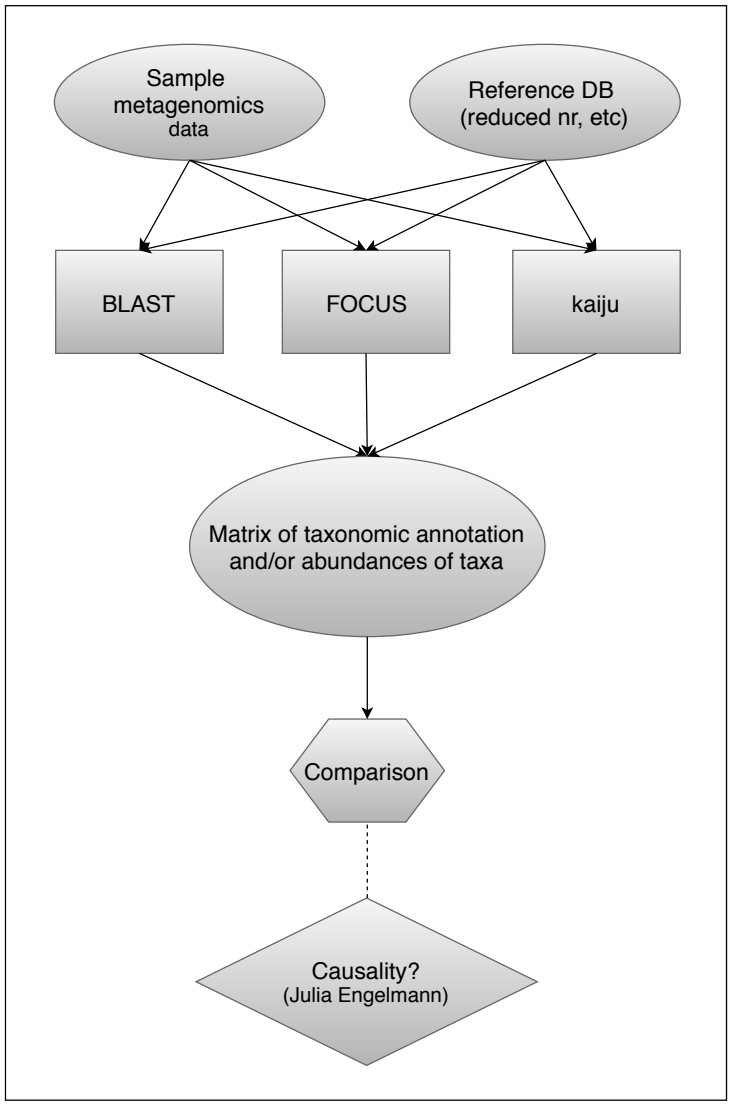

Figura 2.1: Pipeline of profiling tools for comparison and choice of best tool.

2.5.3 Preparation of next steps

- Analysis folder (maya): *Documents/posDoc/Profiling*

Sample metagenomics data: from the CAMI initiative `https://data.cami-challenge.org/participate`. I downloaded only the *sample_1* of the metagenomes of Gastrointestinal tract, for space reasons.

Sample_1 has the following folders (with sizes:)

Tabela 2.1

Size	Folder
259M	hybrid
11G	pacbio
16G	short_read

The idea for meeting Julia is to bring abundance of microbiomes to build networks or bring the networks and come back with causality models. The link to go to the Sea Research Institute in Texel (Netherlands):

`https://9292.nl/reisadvies/station-utrecht-centraal/den-hoorn-texel_t-horntje/aankomst/2018-06-18T1155`

2.5.4 Preparation for running FOCUS2

For running FOCUS, I went to the FOCUS2 website: `https://github.com/metageni/FOCUS2`, and followed the directions. Apart from Python, that I already have, I installed the dependencies:

- Jellyfish: `http://www.cbcb.umd.edu/software/jellyfish` (installed in *bia/bin*)
- Numpy: `http://sourceforge.net/projects/numpy/files/NumPy` (from `http://scipy.org/install.html`)
- SciPy: `http://sourceforge.net/projects/scipy` (from `http://scipy.org/install.html`)

Then I downloaded the GIT repository in *bia/bia/FOCUS2-master* and used the following script to download the FOCUS2 database.

Inputs and directories for FOCUS2:

- Input metagenome data: *Documents/posDoc/Profiling/sample_data_CAMI/short_read/2017.12.04_18.45.54_sample_1/reads/anonymous_reads.fq*
- Output: *Documents/posDoc/Profiling/FOCUS2_run*
- FOCUS tool and usage:

```
bin/FOCUS2-master$ python focus2.py
```

The file *anonymous_reads.fq* has 10 Gb size and 133,329,312 lines. I created a subset of that with a tenth of that (First 13,332,928 lines - a multiple of four lines for the fastq format to be correct at the end). The file is: **anonymous_reads_subset.fq**, which was moved to *Documents/posDoc/Profiling/FOCUS2_run* folder.

I tried running FOCUS2, but I had problems installing HS_BLASTN. `https://github.com/chenying2016/queries`

2.6 15 - working at Bas' group

2.6.1 Getting FOCUS2 to work

- Analysis folder (maya): *Documents/posDoc/Profiling* - Running folder (maya): *bin/FOCUS2-master*

I tried getting FOCUS2 to work, the SciPy package was not correctly installed, so I just installed it again using 'sudo apt get install scipy', which worked fine. After that jellyfish was also incorrectly installed, which I installed using the same approach. After that there was an error on the code. Line 242 has a function called 'loadDB', it opens a "pickle" file, which is only available if the whole 50Gb database of FOCUS2 was downloaded fully. Since I don't have space, I downloaded only a part, which did not contained these files. Bastiaan has helped me debug this. These files are not normal txt files, so it is not possible for me to use FOCUS2 now.

As an alternative, I will try to get FOCUS to run. First I will git clone this repository, and follow the instructions. Then I will try to get it to run on my sample data.

2.6.2 Installing FOCUS instead

- Repository folder (maya): *bin/FOCUS-master*

Since it was not possible to get FOCUS2 to work on maya, I am trying FOCUS. First, I git clone it to the repository folder from: `https://github.com/metageni/FOCUS`. I got the FOCUS repository, and tried to run:

```
python3 focus.py -q input -k -o output
```

It complained about a package, that I installed successfully using the following:

```
sudo pip install pathlib
```

Now it complained about numpy and scipy, so I installed it with:

```
sudo apt install python3-pip
pip3 install numpy
ModuleNotFoundError: No module named 'scipy'
pip3 install scipy
```

I had an error now:

```
[2018-06-15 12:13:21,369 - INFO]    Counting k-mers
count: unrecognized option '--disk'
Use --usage or --help for some help
terminate called after throwing an instance of
       'jellyfish::compacted_hash::ErrorReading'
what():  'kmer_counting_0.62600186751217': File truncated
Aborted (core dumped)
```

I went to the sript an saw that the error came from the line calling jellyfish. I saw the version was wrong, so I downloaded and installed the right one required by FOCUS (2.2.6) from: `https://github.com/gmarcais/Jellyfish`.

2.6.3 Running FOCUS successfully for the probe data

- Repository folder (maya): *bin/FOCUS-master*
- Analysis folder (maya): *Documents/posDoc/Profiling/FOCUS_run/output*

Now FOCUS ran without any problems with my sample input file. I had to run the tool on the same Repository folder from FOCUS, but I moved the input and output back to the working folder for the analysis. So now I have the abundances and the phyla, up to species level.

To visit Julia, I should have a matrix of abundances, so I can run co-relations between different samples. Right now I have only the profiles of this sample I got from CAMI (10% of file 'anonymous_reads_subset.fq', that I transformed to FASTA for FOCUS). I will run it again for more 11 samples, so that I can get a matrix of $12 \times n$, with n profiling groups.

$python3 focus.py -q input/ -o output/

2.6.4 Running kaiju for the probe data - unsuccessful

- kaiju running folder (maya): *Documents/posDoc/Profiling/kaiju_run*

Apparently, I installed kaiju correctly. Even if I am running it with the small input databases that Bastiaan gave me, I still could not successfully run it in my machine for

the sample data. I think the problem is low power and/or memory. I will drop this for now and follow the analysis with the FOCUS output. I won't use BLAST nor now, since it is not directly comparable with FOCUS, without kaiju as well.

2.6.5 Planning next steps with Bas and Bastiaan

- Repository folder (maya): *bin/FOCUS-master* - FOCUS matrices folder (maya): *Documents/posDoc/Profiling/FOCUS_output_bigData_Bastiaan*

The plan now is to build the matrix to bring to Julia. My matrix is quite small, so Bastiaan is kindly providing me with a 'real' bigger matrix, to show to Julia (FOCUS matrices folder). The second thing is that I give them the 700 genomes list that the students of Pedro built, to check their integrity, I asked Leticia about it.

2.6.6 Login in the cluster of Santos Dumont - unsuccessful

- Login tutorial (maya): *Documents/posDoc/SantosDumont_Cluster*

I am trying to get a login on the cluster now, to try to speed up the runs after we decide on a software. I followed their directions on first to access the cluster by VPN, but they required two passwords, and I only have one. I sent an e-mail to the helpdesk, so I will wait for their answer.

2.6.7 Preparing for meeting Dr Julia Engelmann

- Abundance data for metagenomes folder (maya): *Documents/posDoc/Profiling/FOCUS_output_bigData_Bastiaan*

On monday, I will visit Dr Julia Engelmann, a researcher working on marine ecosystems who is specialized in causality analysis. I will bring to her matrices with microbiomes abundances, so that I can discuss with her methods for running corelations and infer causality.

I have two different types of tables. Kaiju's report, that Laura Dijkhuizen kindly provided me. These refer to only one metagenomics sample, from a plant, and do not qualify to run such analysis, since we want to know what differs between two different metagenomics states.

I think the best choice here is to use the data Bastiaan gave me from FOCUS (abundance data for metagenomes folder). These comprehend 8,091 metagenomes. The namings are arbitrary, since we have little knowledge about where the genomes are coming from. There is some metadata in case we need it. I have 999 of the outputs, which should correspond to 999 input metagenomes. Each one has a FOCUS output folder, with files inside that correspond to:

- GROUP-LEVEL__out__STAMP_tabular.tsv

- out__STAMP_tabular.spf

The GROUP-LEVEL can be Kingdom, Order, Family, Phylum, Genus, Species. The second type of file correspond to a big summary, its header is:

```
Kingdom Phylum Class Order Family Genus Species Strain /
MGXDB000003_filtered.fastq
```

And it seems like the description and abundance of each strain, with all its higher phylogenetic classification details and abundances.

I talked to Bas about the input matrices for the correlation tools. The matrices whould contain in each column one of the metagenomes' abundances and in each line one of the phylogenetics groups. Some of the groups will have zero abundance, in case they were not observed, but I can discuss with Julia on how to deal with such cases.

I will make 7 matrices, each for a phylogenetic "GROUP-LEVEL", and I will ignore for now the "out" files. For that I used the following command in a loop.

2.7 16 - working at Bas' group

2.7.1 Finishing the script for creating matrices of abundances

- Lib (maya): *Documents/posDoc/Lib-Profiling*
- Script (maya): *buildMatrices.pl*
- Working folder (maya): *Documents/posDoc/Profiling/FOCUS_output_bigData_Bastiaan*

I finished the script that creates matrices of abundances, taking into consideration the IDs of the metagenomic sample (files), and the IDs of the taxonomic groups (lines). There were some tricky details, since the abundance matrices do not always contain only TWO columns, sometimes it contains more, but I re-named the groups to not contain any spaces, and these are on the first column. The rest correspond only to *digits*, or abundances. When the taxonomic group is not present in the file sample, a zero is printed.

To get the usage of the script:
perl ~bia/Documents/posDoc/Lib-Profiling-Processing/buildMatrices.pl --help

I ran it using the following parameters:
perl ~bia/Documents/posDoc/Lib-Profiling-Processing/buildMatrices.pl --input input-Folder/ --level Class --reg_exp _filtered.fastq > output.matrix

And for the following taxonomic groups (argument --level):

- Class
- Family

- Genus
- Kingdom
- Order
- Phylum
- Species

The seven matrices I produced correspond to files in folder:
FOCUS_output_bigData_Bastiaan_abundance_matrices:

- Class.matrix
- Family.matrix
- Genus.matrix
- Kingdom.matrix
- Order.matrix
- Phylum.matrix
- Species.matrix

2.8 18 - working at Bas' group

2.8.1 Cancel of my meeting Dr Julia Engelmann

Today morning I spontenously decided not to go to Texel to visit Dr Julia. I worried that I did not have enought time to do all the tasks I am supposed to do here before my visit is over. Julia works with networks and causality, which is a second step in the project. So I decided to focus on the first part (profiling metagenomes).

2.8.2 List of 700 genomes

Bas will take a look at the 700 genomes that the students of Pedro have compiled, to see if there is any addition he would suggest. Leticia emailed me and said those 700 genomes came from the **Bioproject PRJNA273161** of the article [Brown et al., 2015]. These came from microbial communities from an aquifer adjacent to the Colorado River in the US. The objective of the research was to better understand about the radiation of phyla that may comprise $> 15\%$ of the bacterial domain. The genomes come from part of > 35 of these phyla.

Looking at the bioproject description, they could reconstruct only 8 complete genomes and 793 draft genomes from a Candidate Phyla Radiation (CPR). Leticia has downloaded the genomes from the following link:

```
https://www.ncbi.nlm.nih.gov/assembly/?term=PRJNA273161
```

2.8.3 Try the login in the Server made available by Pablo Ivan - unsuccessfully

Pablo has kindly made access to us temporarily in a server that I can use to test the profiler tools, at his system.

The access should be done by ssh with the following:

```
ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br
l: aquifermicrobiomes
p: microbiome
```

Which unfortunetaly did not work out. I wrote to Pedro and he will check it with Pablo.

2.8.4 Try login again to the Santos Dumont - unsuccessfully

I couldn't access the server last week, and the helpdesk told me any extra space or wrong char at the file */etc/vpnc/sdumont.conf* would cause problems. So I tried again, carefully re-writing the file, but the same error I had before occured. I wrote them again, sending a copy of the file and my command lines.

2.8.5 Prepare approaches for testing profiler tools

These are the important things that Pedro recommended me to have attention to when using the profile tools (email from July 15):

We should be able to (in order of importance):

- customize the reference Database (to add genomes/taxa)
- retrieve back the ID of the sequences in the output and the sequence as well to re-annotate interesting taxa
- Be fast, so it can be reasonable to run with our volume of data

So far, it looks like Kraken and kaiju attend these criteria. I will check this this week.

Another thing I should check with the research team of Bas is about quality control tools. Pedro used **PRINSEQ**. Are there any other tools?

Importantly, **Bacteria** and **Archea** are the organisms of interest to us.

2.8.6 Trying to run Kraken - successfully

- Location of tool (maya): *bia/kraken/*
- Location of Reference Database 4Gb (maya): *bin/kraken/minikraken_20171013_4GB*
- Working folder (maya): *Documents/posDoc/Profiling/*
- Input at WF (maya): *input_subsample_data_CAMI/file.fasta*
- Output at WF (maya): *Kraken_run*

Pedro said Kraken has a nice advantage over the other tools, since it outputs the ID/sequence of the microbiome. Which is super useful for us for later taxonomic analysis. I will thentry to run it. It is installed, but I still have to figure out how to use it, since it requires a DB of reference. The Kraken tool requires a server, at least 500 GB of disk space. But I am running miniKraken in my laptop, which require 4Gb space, 4Gb of size of the reference DB and 8 Gb of RAM.

According to their manual `http://ccb.jhu.edu/software/kraken/MANUAL.html#installation`, a Kraken database is a directory containing at least 4 files:

- database.kdb: Contains the k-mer to taxon mappings
- database.idx: Contains minimizer offset locations in database.kdb
- taxonomy/nodes.dmp: Taxonomy tree structure + ranks
- taxonomy/names.dmp: Taxonomy names

The small reference DB for the miniKraken is available at their website: `http://ccb.jhu.edu/software/kraken/`. I unzipped it and put at the same folder as kraken. To use it I followed their recommendation at the manual site:

$kraken --db $DBNAME seqs.fa, I ran the following very successfully in my laptop :)
$ bia/bin/kraken/kraken --db bia/bin/kraken/minikraken_20171013_4GB/ input_subsample_data_CAMI/file.fasta

```
3333232 sequences (499.98 Mbp) processed in 166.069s (1204.3 Kseq/m, 180.64 Mbp/m).
2166605 sequences classified (65.00%)
1166627 sequences unclassified (35.00%)
```

2.8.7 Try the login in the Server made available by Pablo Ivan - successful

Pablo has passed me the login again, like below:

```
ssh -X aquiferspablo@bioinfo03.bahia.fiocruz.br
l: aquiferspablo
p: microbiome
```

now I can run the tests to choose the best tools. I asked Pablo where can I run the experiments and how I can get the tools, and as soon as he gets back to me, I will start running the tests. While they run, I can start preparing to compare the results. I suggested trying the following tools:

- Kraken
- kaiju
- MMseqs2 (in a second moment)

2.8.8 Prepare data for comparison

- Location of control data (maya): *Documents/posDoc/Profiling/sample_ data_ CAMI*

After I get the results, I will compare two things between the result and the control to choose the best tool: (i) the classification (if the tools could identify correctly the microbiome groups) and (ii) the distributions of abundances. For this the VEGAN package of R should be useful.

For the first comparison, I basically have to compare the name strings I got for each ID with the real one and see if they match. I will count the matches and get the percentage of matches. Afterwards I will compare two vectors corresponding to the distributions of abundances (one of the results and the other of the real data - using R).

For now I will check the files and see how the tools deliver the results (format, if I have to transform the data, etc).

CAMI control - files and formats

I took data from the CAMI initiative of: *2nd CAMI Challenge Human Microbiome Project Toy Dataset*, Gastrointestinal tract, sample_1. I used the Illumina set, found at subfolder *short_ reads/*. Inside these folders, there are three subfolders, *bam/*, *contigs/* and *reads/*. The *bam/* contains mapping, which we did not do. *contigs/* contains assembly, which we also did not do.

reads/ contains Illumina data, with one file being the FASTQ reads and the other the mapping of every single read to the genome it originated from along with the original read IDs (pre anonymisation): **reads_mapping.tsv.gz**.

Every dataset of the CAMI initiative should contain one abundance file per sample mapping OTUs to genomes: **abundance#.tsv**. It is missing from my data, so I downloaded only it using:

```
wget https://openstack.cebitec.uni-bielefeld.de:8080/swift/v1/CAMI_Gastro \
intestinal_tract/short_read/abundance1.tsv
```

In the same way, I downloaded the taxonomic profile with file: **taxonomic_profile_1.txt**. If I understand the format correctly, the first column (named **TAXID**) corresponds to

the taxonomic ID, which is (hopefully) teh same one used by Kraken. It is a number that refer to the organisms.

The input multi-FASTA has the following headers:

```
>S1R0/1
>S1R0/2
>S1R1/1
>S1R1/2
```

Header of **taxonomic_profile_1.txt**:

```
@SampleID: 1
@Version: 0.9.1
@Ranks:superkingdom|phylum|class|order|family|genus|species|strain
@@TAXID RANK TAXPATH TAXPATHSN PERCENTAGE _CAMI_GENOMEID

2 superkingdom 2 Bacteria 100.0
544448 phylum 2|544448 Bacteria|Tenericutes 0.0
32066 phylum 2|32066 Bacteria|Fusobacteria 0.0
```

Kraken - files and formats

The headers of the output of Kraken, with the third column being the taxonomic ID of the read:

```
C S1R0/1 517 150 0:86 517:1 0:33
U S1R0/2 0 150 0:120
C S1R1/1 94624 150 0:37 94624:1 0:13 94624:1 0:26 94624:1 0:41
C S1R1/2 94624 150 0:105 94624:1 0:14
```

In this output, the first column refer to either "Classified" or "Unclassified". The second to the ID of the FASTA header. The third to the **labelled taxonomy ID** Kraken used to label the sequence (with 0 if the sequence is unclassified. The fourth to indicate how the k-mers mapped). I can take only the 1st column and the third, to get the classification.

If necessary, I can also translate the label to taxonomic names using a Kraken argument.

kaiju - files and formats

2.8.9 File transformation for tool comparison

- Working folder (maya): *Documents/posDoc/Profiling/tools_comparison*

After I studied the types of files and formats that the CAMI, the Kraken and the kaiju use, I can already start preparing the data for comparison. First I will do the easiest

which is to compare the taxonomic IDs between files and see what percentage did the Kraken/kaiju correctly and incorrectly profiled.

I put in the working folder the files in reference to the tool they came from. Then I can compare them in RStudio. I will start with Kraken, since I already have its output:

cut -f1,5 ../sample_data_CAMI/short_read/taxonomic_profile_1.txt > CAMI_gastro_sample1_taxonomic_profile.txt

Very importantly! I checked with Bastiaan about the taxonomic IDs. So, they indeed mean the same thing in Kraken and in CAMI. So in principle, I can directly compare them to evaluate the best tool. **However,** the *dates* of *when* the databases were generated could change the taxonomic IDs. For them to be completely compatible, I have to first check if the taxonomic IDs of the "sientific_name" match to the names of the microbiomes.

In practical terms, that means that I have to compare the *names.dmp* files between both CAMI and Kraken tools.

These files look like this:

```
bia@maya:~/bin/kraken/minikraken_20171013_4GB/taxonomy$ head names.dmp
1 | all | | synonym |
1 | root | | scientific name |
2 | Bacteria | Bacteria <prokaryotes>| scientific name |
2 | Monera | Monera <Bacteria>| in-part |
2 | Procaryotae | Procaryotae <Bacteria>| in-part |
2 | Prokaryota | Prokaryota <Bacteria>| in-part |
2 | Prokaryotae | Prokaryotae <Bacteria>| in-part |
2 | bacteria | bacteria <blast2>| blast name |
2 | eubacteria | | genbank common name |
2 | not Bacteria Haeckel 1894 | | synonym |
```

The steps for this comparison according to Bastiaan are:

- Grep only 'scientific name' tags in the last column (4th element)
- Get the taxonomic ID from the first column (e. g. '2')
- See if the regular expression in the second column is the same between both files (e. g. 'Bacteria' in the example above)

If so, they are compatible, if not, I have to transform the formats. If this happens, I will ask Bastiaan again on how to proceed.

2.9 19 - working at Bas' group

2.9.1 File transformation for tool comparison

- Lib (maya): *Documents/posDoc/Lib-Profiling*
- Script (at Lib): *compareTaxID-versions.pl*
- Working folder (maya): *Documents/posDoc/Profiling/tools_ comparison*

I wrote a script to compare versions of taxonomic IDs. For now I added options so that the script can parse CAMI and 'names.dmp' types of files. If there are any other formats in the future, it will be easy to add them as subroutines to my script.

For usage:

perl bia/Documents/posDoc/Lib-Profiling/compareTaxID-versions.pl --help

The command line I used to compare the formats of the CAMI file and the 'names.dmp' of the Kraken DB reference:

perl ~bia/Documents/posDoc/Lib-Profiling/compareTaxID-versions.pl --fileQuery ../sample_data_CAMI/short_read/taxonomic_profile_1.txt taxonomic_ profile_1.txt --formatQuery CAMI --fileDB ~bia/bin/kraken/minikraken_20171013_4GB/taxonomy/names.dmp --formatDB namesDMP

With this I got the following output:

```
228400 Query: Histophilus somni DB: Haemophilus somnus 2336
205914 Query: Histophilus somni DB: Haemophilus somnus 129PT
642492 Query: Cellulosilyticum lentocellum DB: Clostridium lentocellum DSM 5427
1852377 Query: Actinomyces pacaensis DB: Actinomyces sp. Marseille-P2985
1306519 Query: Flavobacterium commune DB: Flavobacterium communis

Number of matches: 1135
Number of unmatches: 5
```

Considering these results, I would say that the databases seem at first glance compatible, with very few small changes in the names.

2.9.2 Customize Databases to add more genomes to the reference DB of Kraken

One of the crucial steps of the project now is to customize DB using Kraken build. In this way, we add more gemomes to the Kraken DB. The DB of Kraken uses the complete RefSeq database of all genomes contained in RefSeq at the time of: bacteria, archea and virus.

Of the list of 700 genomes that Leticia passed to me, some are already annotated in RefSeq, so I suppose these would also be in the Kraken DB. Some of these genomes that are only at contig level are not in RefSeq yet, so these need to be included once the server is available.

The RefSeq is a very well curated database, while nr for instance is not, since it also includes sequences that were submitted by individual laboratories without further curation.

2.9.3 Talk at the University of Utrecht

Today I gave a talk at the Institute about my PhD work, entitled "Adaptive Evolution of Long Non-Coding RNAs".

2.10 20 - working at Bas' group

2.10.1 Transfering the list of 700 metagenomes of aquifers to Pablo at the Fiocruz server - successful

Pablo from the FioCruz Institute is helping us to create a customized DataBase of reference for Kraken. He is installing the tool at the fiocruz server and he asked me to transfer the metagenomes there. These metagenomes are the ones from the list that Leticia has given me. This list contains genomes comming from metagenomic samples of aquifers. 8 are complete and a bit more than 700 are incomplete. The BioProject refering to this dataset is: **PRJNA273161**.

I transfered the files using scp from maya to the FioCruz server to a folder specified by Pablo:

scp -r genomes_BioProject_NCBI_PRJNA273161/ aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/microbiomes/tools/krakenfiles/dbases/stdScript/.

The website I downloaded the fasta sequences first (to maya) is:
`https://www.ncbi.nlm.nih.gov/assembly/?term=PRJNA273161`

2.10.2 File transformation for tool comparison - assessment of results

- Lib (maya): *Documents/posDoc/Lib-Profiling*
- Script (at Lib): *get-lineage.py*

I checked the results of the comparison between taxID vesrions of teh CAMI query against the Kraken DB, and saw 5 mismatches (see June 19). I checked them with Bastiaan, and these are actually "synonyms" or "include" that are part of the organism as well.

I restricted my search to "scientific names". But it could be that even if these differ, the nomenclatures in reference to "synonym" or "include" match. In the case of the query and DB I probed before, they match, so I can proceed with the comparison of results.

Importantly, if the versions *differ*, then I have to make the conversion with a script. In this case, I would then pattern match using the name, than after I match I can get the "correct" taxID from the DB for the given name. Then, both versions would be compatible.

In addition, I got a nice script from Bastiaan that can retrieve the full phylogenetic tree of a given taxID. This script is at the Lib and can be used as the example for taxID 155.

$python3 ~bia/Documents/posDoc/Lib-Profiling/get_lineage.py 2 ~bia/bin/kraken/minikraken_20171013_4GB/taxonomy/nodes.dmp

2.10.3 File transformation for tool comparison

- Working folder (maya): *Documents/posDoc/Profiling/*
- Golden standard CAMI, taxIDs and abundances (at *WF*): *tools_comparison/CAMI_gastro_sample1_taxonomic_profile.txt*
- Golden standard CAMI, readIDs and taxIDs (at *WF*): *sample_data_CAMI/short_read/2017.12.04_18.45.54_sample_1/reads*
- Results of Kraken run (at *WF*): *Kraken_run/kraken_run_CAMI_2018-1806.txt*

Now that I checked that the versions of the taxIDs are compatible between the CAMI dataset and the Kraken dataset, I can already compare the results of the Kraken run on annotating the CAMI input with its golden standard.

So basically, I want to compare two inputs (tables of Read IDs and correspondent taxIDs) and report the rights and wrongs of Kraken. First of all, I want to extract these columns from the data. From the original file of *Golden standard CAMI, readIDs and taxIDs* to a two-column-table easy input at the *WF*:

$cut -f1,3 ../sample_data_CAMI/short_read/2017.12.04_18.45.54_sample_1/reads/reads_mapping.tsv > CAMI_readID2taxID.txt

Now from the original Kraken result file to a two-column-table easy input at the *WF*:

$grep '^C' ../Kraken_run/kraken_run_CAMI_2018-1806.txt | cut -f2,3 > Kraken_readID2 taxID.txt

Obs: the grep for the 'C' characters was done to pre-filter the results of Kraken for only the reads that have been assigned a TaxID.

Now the comparisons will be between these two files.

- Lib (maya): *Documents/posDoc/Lib-Profiling*
- Script (at Lib): *compare_ ReadID2TaxID.pl*
- Golden standard CAMI, readIDs and taxIDs (at *WF*): *tools_ comparison/CAMI_ readID2taxID.txt*
- Results of Kraken run (at *WF*): *tools_ comparison/Kraken_ readID2 taxID.txt*

perl ~bia/Documents/posDoc/Lib-Profiling/compare_ReadID2TaxID.pl --fileQuery Kraken_readID2taxID.txt --fileDB CAMI_readID2taxID.txt --fileNodes bia/bin/kraken/minikraken_20171013_4GB/taxonomy/nodes.dmp

2.11 21 - working at Bas' group

2.11.1 VPN access to Santos Dumont - requested new password

I requested a new password, according to the Helpdesk's recommendation.

2.11.2 Asked Bas about TrimSeq

I asked Bas about the quality filtering tools. He uses Trim-galore instead of TrimSeq, but said essentially they do very similar things (filter reads by quality q, say $q < 20$). So it shouldn't matter much which tool we use.

2.11.3 General tips

I had a meeting with Bas and got a lot of useful tips for our future analysis. He recommended taking a closer look at these profiling tools:

- Kraken
- kaiju
- MMSeqs2.0
- BLAST (for analysis of small subsets due to run time)

If we use BLAST at some point, the **CAT** tool could be very useful. It takes around 10% of the top hits, and gets the last common ancestor of all of them to assign the taxID.

Importantly, MMSeqs2.0 do not output the taxID or abundance directly, but we can get these back using scripts. First, the taxID can be recuperated from the protaccessioID using the file with the names conversion **protaccession2taxID.dmp** that can be downloaded from the NCBI. The abundance can be inferred from the output file.

2.11.4 Working on *compare_ ReadID2TaxID.pl* script

Check yesterday for more info.

2.12 22 - Prepare for moving from Germany to Salvador

Today I am going back to Leipzig and during next week I will prepare for moving from Germany to Salvador. I will arrive there at the 4th of July and start working at Pedro's Lab.

Capítulo 3

July 2018

3.1 9 - Setting up the computer

- Workstation (Meirelles Lab): **user-B150M-Gaming-3**

- Setting up the system's configurations
- Installing programs (TexWorks)
- Call Pablo about customizing the Databases
- Call LNCC for Santos Dumont access

The priority this week is to customize the DB for Kraken and to prepare the documentation for the supercomputer. First, we will test the program at Fiocruz using minikraken, then make sure the results are fine, and finally send the job to the supercomputer.

I talked to Pablo from Fiocruz, and we agreed on running minikraken. He told me to transfer the database to the same folder I used before and run the job at the same folder. His commands are two folders up, and I can use the same ones. He said that we need to format the FASTA files of the 700 genomes so that Kraken can read it. He will custom the script in python so that it can use only the required function to transform the format. Then I will transform the formats and run the kraken customize DB with the smaller DB from RefSeq plus the 700 genomes.

3.2 Add smaller database at Fiocruz server for minikraken - Kraken2

- Fiocruz server access: **ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br**
- File location: */media/bioinfoserver3/Bkp1/microbiomes/tools/krakenfiles/dbases/stdScript/*

I was going to download the smaller database of Kraken, but apparently they released Kraken2 in the past few weeks, which they clame is more efficient. I will check with Pablo if it is indeed more feasible.

Pablo's command:

/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/install-dir/kraken2-build --standard --threads 24 --db /media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/std-database

3.2.1 VPN access to Santos Dumont - successful

- Fiocruz server access: **ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br**
- Login tutorial for vpnc connection (user-B150M-Gaming-3): *Documents/SantosDumont/How To VPN SDUMONT LINUX.pdf*
- User guide: `http://sdumont.lncc.br/support_manual.php`

Today I could access the Santos Dumont cluster successfully with my new passwors, with their help by phone-line. My new password and login follow below:

```
login: maria.costa
password: look up at notebook/notes at mobile
```

To access the supercomputer, I have to first establish a vpnc connection in the background, and then do the ssh to the cluster, like below:

sudo vpnc /etc/vpnc/sdumont.conf
enter passwrd <local root>
enter passwrd for maria.costa <santos dumont>

After that, the vpn connection should be established, and then I can ssh to the cluster, as in:
ssh maria.costa@login.sdumont.lncc.br
enter passwrd for maria.costa <santos dumont>

After that I have an ssh connection. When I exit the ssh connection, then I have to log out of the vpn connection with:

sudo vpnc-disconnect

3.2.2 Report - request for premium account Santos Dumont

- Review with comments from Pedro: */home/biawalter/Documents/SantosDumont/Reviews_SDUMONT2018-CHAMADA1 paper 182361.docx*

Our review for requesting a premium account at the supercomputer was denied. The review with comments is at my workstation, with comments from Pedro to improve the proposal and re-submit it. As a TODO for me, there are specific points to be covered. Overall, I should focus on specifying the computational needs for each program and how to justify our need of the STDU.

3.3 10

3.3.1 Working with Kraken2

- Fiocruz server access: **ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br**
- Working folder: */media/bioinfoserver3/Bkp1/microbiomes/tools/krakenfiles/dbases/stdScript/*
- DB folder: */media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/std-database*
- Names.dmp file from NCBI - Kraken2 standard DB (at DB folder): *taxonomy/names.dmp*

Today I am working at customizing the standard DB of Kraken. Pablo yesterday installed Kraken2 and we agreed on trying to run it before making the test with a smaller subset of the standard DB. He tried to download the standard DB, but could not do it for the full set. I'll try again for the whole database using the same command from yesterday, but re-directing it to a different folder, like below:

/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/install-dir/kraken2-build --standard --threads 24 --db /media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/std-database2

This script first downloads all the data from RefSeq and afterwards builds the database. I'll see if I can run the process without the error Pablo had yesterday. Unfortunately, I had the same problem, which appears to be a problem of connectivity with the internet, in the step of downloading the sequences, before building the databases.

Pablo will try tommorrow to dowload the sequences in another computer and then transfer the sequences to the fiocruz server, to build the standard database there. In case this doesn't work, we can think of using the Santos Dumont or using kraken1.

In parallel, I will already send an email for the Santos Dumont helpdesk, so that they install the Kraken2. I read the manual for usage of the cluster, found at:

`http://sdumont.lncc.br/support_manual.php`

3.3.2 Documentary filming & form

Victor and Yuri are making a documentary about the aquifer's project. They made some filming in the lab today and are preparing some animations. I filled the form they sent

asking some questions about how can we contribute to transfer our knowledge to them. The form was at google docs at link:

`https://goo.gl/forms/a2DwBal4uJ9hFD552`

3.4 11 to 15

3.4.1 Working on the review of the proposal for a premium account at the STU

- Review: *Documents/Reports/Reviews_proposal for premium account at STU.docx*

Today I am working on the review of the STU premium account. I am attending the points requested by Pedro, adding my comments as track changes with LibreOffice.

3.4.2 700 genomes to customize with Kraken2 - formatting

- Fiocruz server access: **ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br**
- Working folder: */media/bioinfoserver3/Bkp1/microbiomes/tools/krakenfiles/dbases/stdScript/*
- FASTA files of the 700 genomes (at WF): *genomes_BioProject_NCBI_PRJNA273161/ncbi-genomes-2018-06-20/GCA*fna*
- DB folder: */media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/std-database*
- *accession2taxid files from NCBI - Kraken2 standard DB (at DB folder): *taxonomy/*accession2taxid*
- Lib: *Documents/Lib-Profiling*
- Script to format FASTA files to Kraken2 format (Lib): perl fromNCBI_ID2taxID.pl --help
- Bash loop script to run the perl script for multiple files (Lib): loop.sh

Pablo was successful to download the database of Kraken2 in a different computer from the Fiocruz cluster (the one with problems of connectivity). An now he transfered the files to the Fiocruz server and is building the database.

In parallel, I am formating the FASTA files of our 700 genomes. The FASTA headers of the genomes must have a specific format, according to the Kraken2 manual `http://ccb.jhu.edu/software/kraken/MANUAL.html#custom-databases`. I will transform their formats, including the taxID in the header of the FASTA file (instead of the NCBI ID that is right now). First I need to go from the NCBI ID to the taxID, using a big hash table and then re-write the FASTA files with the correct format. The perl script that does that follow below:

fromNCBI_ID2taxID.pl --fileDB FILE_DB --regDB REG_EXP_DB --fileQuery FILE_QUERY --regQuery REG_QUERY --outputFolder --jobID JOB_ID --number_sequence NB_SEQ

The --number_sequence argument is an argument to number the sequences in increasing order, according to the Kraken2 format, as shown below (in the example this

argument is 16):

```
 >sequence16|kraken:taxid|32630  Adapter sequence
CAAGCAGAAGACGGCATACGAGATCTTCGAGTGACTGGAGTTCCTTGGCACCCGAGAATTCCA
```

The script is now running for one file, and if the NCBI ID is not found in the DB file, it writes it in a report file (identified by the ID, which can be used multiple times). The bash script for the loop runs the script for multiple sequences:

```
#Usage: arguments are
#$1: folder in which the query files are (only folder, no name)
#$2: Database (full path, example: teste_DB)
#$3: jobID (simple string)
#$4: regDB for perl script DB
#$5: regQuery for perl script Query
#$6: output Folder
```

The exact loop I used in the Fiocruz server at the WF:

./loop.sh genomes_BioProject_NCBI_PRJNA273161/ncbi-genomes-2018-06-20 /media/ bioinfoserver3/Bkp1/microbiomes/tools/kraken2/std-database/taxonomy 13-07-18-16h accession2taxid fna genomes_BioProject_NCBI_PRJNA273161/ncbi-genomes-2018-06-20-formatted

And the perl line inside this script:

perl fromNCBI_ID2taxID.pl --fileDB $db --regDB $regDB --fileQuery $query --outputFolder $outputFolder --jobID $jobID --number_sequence $n --regQuery $regQuery

I tested the script and de-bugged it with less files and a sample of the accession2taxID file in the workstation and is seems to be working fine. I ran it in the server with the accessionb2taxid files that have in total 27Gb (arguments of the script: --fileDB FILE_DB --regDB REG_EXP_DB), that were stored in a hash in the perl script.

Today (15/07/18), the formatting finished (all formatted genomes are in one unique file of 500 Mb) and I moved the formatted genomes to the folder Pablo asked me to:

/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/std-database/library/metagenomes/ncbi-genomes-2018-06-20.formatted

Now he will add the genomes to the kraken original DB and try formatting them.

3.5 16

3.5.1 Pipeline of Amanda & Leticia to homogenize metagenomics files

Amanda and Leticia are having problems with filtering some files. They have a pipeline that first downloads metagenomes from MG-RAST and NCBI and then homogenizes the files, so they all can be used in the same pipelines afterwards.

The NCBI files are raw FASTQ files, unfiltered and some are paired-end while some others are single end. The MG-RAST files are already downloaded from a post-processed step (step 299.1), which means the files were already processed and filtered. In this case, we only need to homogenize the files.

Steps:

- NCBI: filter for quality AND homogenize (exclude N's from the end of the files and exclude sequences $>$ 100 bp)
- MG-RAST: homogenize ONLY (exclude N's from the end of the files and exclude sequences $>$ 100 bp)

3.5.2 Next steps - customize DBs of kaiju and Kraken2 and run metagenomes in Santos Dumont

These files came from a table made by Suzana. Leticia will be the main responsible for running the commands (at the STU), supervised by me.

3.5.3 Customizing DBs - kaiju

- STU access info (UFBA workstation): */Documents/SantosDumont*
- Working folder (STU): */scratch/maria.costa/*
- Lib (STU /home/): *DB_ custom/*
- Script for kaiju DB maker (at Lib STU): kaijuDB_nr.sh

In parallel to the Kraken2, we will also use the kaiju tool (`http://kaiju.binf.ku.dk/`) for profiling metagenomes. I will check the manual for how to customize DBs and if our 700 genomes are already contained in the DB (if so, we don't need to cusomize their DB). Kaiju will assign reads (Illumina or 454) directly to taxa using the NCBI taxonomy and a reference database of protein sequences from microbial and viral genomes (RefSeq or nr).

Before classification of reads, Kaiju's database index needs to be built from the reference protein database.

This can be done with the command makeDB.sh, which will download a specific DB (RefSeq, nr, etc), extract the proteins and index it. There are five available "base" DBs,

(i) RefSeq, (ii) proGenomes, (iii) virus genomes, (iv) nr, (v) MarRef and MarDB databases. The nr is more complete, so we will use this one.

To custom databases, we need a FASTA file of proteins in which the headers are the numeric NCBI taxon identifiers of the protein sequences.

The taxon identifiers must be contained in the NCBI taxonomy files nodes.dmp and names.dmp. Then, Kaiju's index is created using the programs mkbwt and mkfmi. For example, if the database FASTA file is called proteins.faa, then run:

mkbwt -n 5 -a ACDEFGHIKLMNPQRSTVWY -o proteins proteins.faa mkfmi proteins

which creates the file proteins.fmi that is used by Kaiju. Note that the protein sequences may only contain the uppercase characters of the standard 20 amino acids, all other characters need to be removed.

From what I understood of the manual, I could create two different DBs, one of nr, which they recommend as being the most complete one, and another of the 700 genomes, in case they are not included in nr. Then, I can run kaiju to profile the matagenomes for both DBs and afterwards merge the output, which is possible with kaiju. To create a customized DB, I need to get the proteins from my 700 genomes first. After that I have to create a FASTA file, each with a protein.

I connected to the STU via vpnc/ssh, and read the manual of slurm queueing systems to submit jobs. I can already start kaiju, since the lates version is already installed (kaiju 1.6.2). Gabriel Bertolino, a student from Pedro, taught me how to submit jubs using the slurm queue system `https://slurm.schedmd.com/srun.html`. The STU has a manual of usage, but when the tool supports MPI `http://sdumont.lncc.br/support_manual.php?pg=support#6`. For instance, when submitting the same command line to several files, it is good to use the sbatch and the script that Leticia passed to me. But for unique commands, I can submit it to slurm directly in the command line via srun. The command I used was:

nohup srun -N 1 -c 24 -p cpu_long /scratch/app/kaiju/1.6.2/bin/makeDB.sh -n -t 24 > out.txt &

To keep an eye on the job: `squeue-umaria.costa`
scontrol show jobid 188502
squeue #to keep an eye on other users' job submissions.
tail -f out.txt #To keep an eye on the job status

3.6 17 - 18

3.6.1 700 genomes to customize with Kraken2 DB - successful

- Fiocruz server access: **ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br**
- Customized DB: */media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB*

Pablo ran the commands for DB customization of the Kraken2 DB (from RefSeq) with the 700 genomes we wanted of aquifers (add genomes and customize the DB). The next step is to check if indeed the 700 genomes are not duplicated (were originally in RefSeq). After making that sure, I can test the tool in STU with the metagenomes Pedro asked (one small, one big and one huge sized), so we can calculate run time to add in the request for a premium account at the STU.

The kaiju tool did not work at the STU because of problems with wget. I wrote to them for a solution.

```
[maria.costa@sdumont13 kaijudb]$ cat out.txt
srun: job 188502 queued and waiting for resources
srun: job 188502 has been allocated resources
Error: wget not found
srun: error: sdumont1363: task 0: Exited with exit code 1
```

3.7 Seminar prep

Tomorrow I will present at the seminars of the Meirelles Lab. Summary of my presentation below:

"Vou contar um pouco da minha trajetÃşria, incluindo trabalhos com BiofÃŋsica teÃş-rica, Imunologia e EvoluÃğÃčo, com um resumo de 10' sobre o meu projeto de Doutorado: Adaptive evolution of long non-coding RNAs. Contarei sobre as minhas especialidades dentro da BioinformÃątica: design de algoritmos, construÃğÃčo de pipelines e anÃąlise de dados, a minha experiÃłncia de estudante em Leipzig, como Ãľ a Universidade, como funciona o LaboratÃşrio de BioinformÃątica, interaÃğÃčo entre os pesquisadores.

E depois vou detalhar como eu administro as informaÃğÃţes do meu trabalho. Incluindo como eu estruturo e implemento meu Lab Book usando a linguagem LaTex - criei um tutorial para isso na minha pÃągina do GITHub `https://github.com/waltercostamb/Lab-Book`. Contarei como eu estruturo a minha $home, guardando as cÃşpias principais de arquivos na workstation, mas com cÃşpias em dois outros lugares usando a ferramenta GIT, que uso para controle de versÃţes e tambÃľm como backup. Na $home eu estruturo os arquivos separando diretÃşrios de scripts (Libs/), papers, apresentaÃğÃţes, etc. Eu uso a workstation para escrever scripts e testar dados pequenos e rodo as ferramentas com os dados de interesse e tambÃľm guardo big data nas servidoras.

No final da apresentaÃğÃčo vou dar um resumo das minhas responsabilidades novas como posdoc do Meirelles Lab. Espero que com isso vocÃłs possam me conhecer um pouco mais, e saibam que podem contar comigo para qualquer problema que tiverem."

3.8 19

3.8.1 Presentation - "My Scientific history"

Today I presented at the seminars of the Meirelles Lab about my scientific history, for the group to get to know me better and also about my organization system.

3.8.2 Kaiju base DB customizing STU - successfull

- Job submission: **STU**
- Job location: *&home*

The helpdesk of the STU wrote me back about the error from kaiju. The /scratch/ nodes do not have internet access. So they told me to run this command at the login node and then process the DB at /scratch/. It is the same problem Pablo had when customizing the DB at the server from Fiocruz. I loged in at STU and ran the following command directly at my $home, to later on transfer to my /scratch folder. If necessary, I will only download the files and afterwards move the processing to the queueing system at the /scratch/ folder.

nohup /scratch/app/kaiju/1.6.2/bin/makeDB.sh -n -t 10 > outKaiju.txt 2>&1 &
#with the 2>&1 part to redirect the error messaged

Now it seems to be working. After it finishes, I will transfer it to /scratch and start making tests. **Remembering that files at /scratch/ are deleted after 60 days!**

3.8.3 Transfer of customized DB of Kraken2 from Fiocruz to STU - unsuccessful

- Customized Kraken2 DB: *Fiocruz server*
- New location: *STU*

For the profiling step of aim 1.1 of the aquifer project, we need to profile metagenomes. For that we needed first to customize the DB of Kraken2. In parallel also work with kaiju. We did that and I want to now move the customized DB from Fiocruz, where we ran the task, to the STU, where we will run most heavy tasks.

I wanted to do that by ssh transfer. After we can download the metagenomics data at STU (me, Leticia and Amanda), I can test the profiling command of Kraken2, and kaiju (as soon as we have its DB). The command line to transfer the files from the server at Fiocruz to the server of STU:

pending

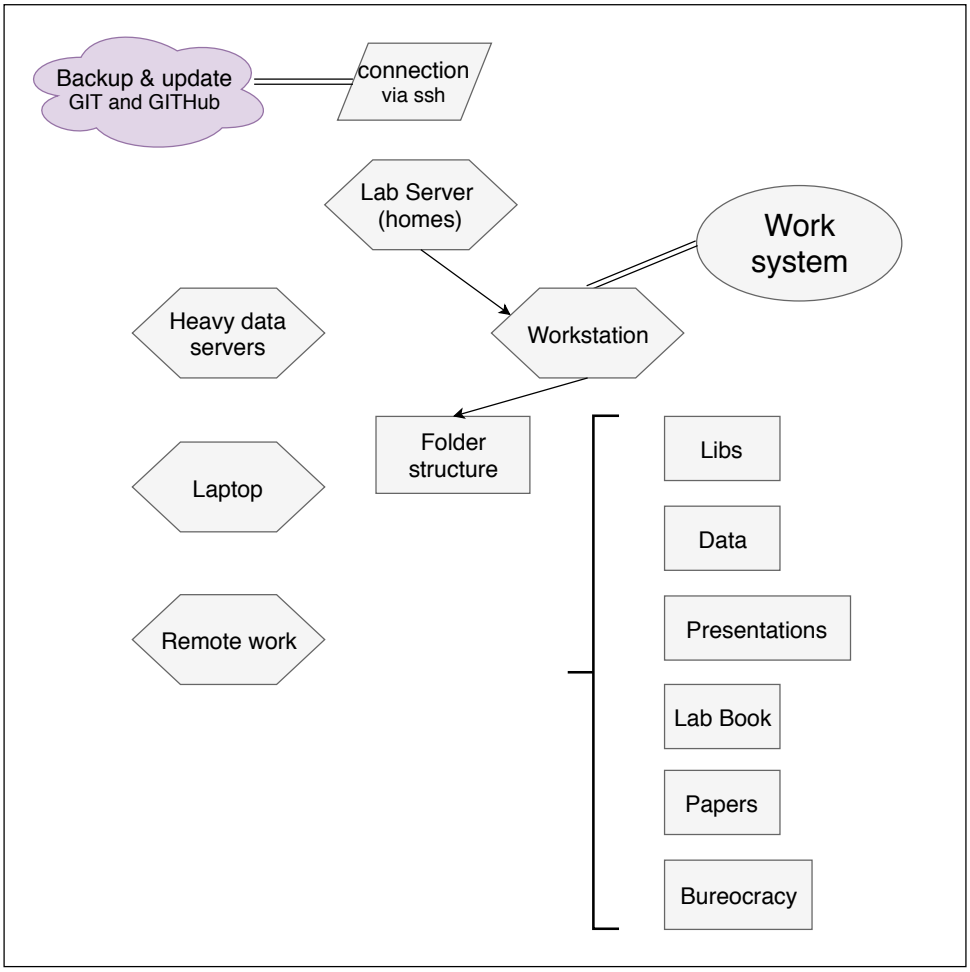

Figura 3.1: My organization scheme of a Bioinformatics work, based on GIT repositories that are updated in any machine I use for work and the folders in my home divided between Libs and Data for the scripts and tests and other folders.

I tried to ssh to Fiocruz, but the connection was timed out. I asked Pablo about it, maybe the server is down for some reason.

3.9 23

3.9.1 Installing TexLive at skywalker

I received a new workstation (skywalker) with Ubuntu 17.04 that I will use for my work during the postdoc. I configured it, by installing some software and updating the Ubuntu version from 17.04 to 17.10 and then to 18.04 LTS, and I also transferred my files from the old workstation by cloning my repositories from my GITHub (Lab Book, Libs, Presentations, Reports, etc), according to my organization scheme (Fig 3.1 and 3.2).

Today I installed texworks and TexLive in skywalker, my workstation of the Meirelles Lab so that I can update my Lab Book.

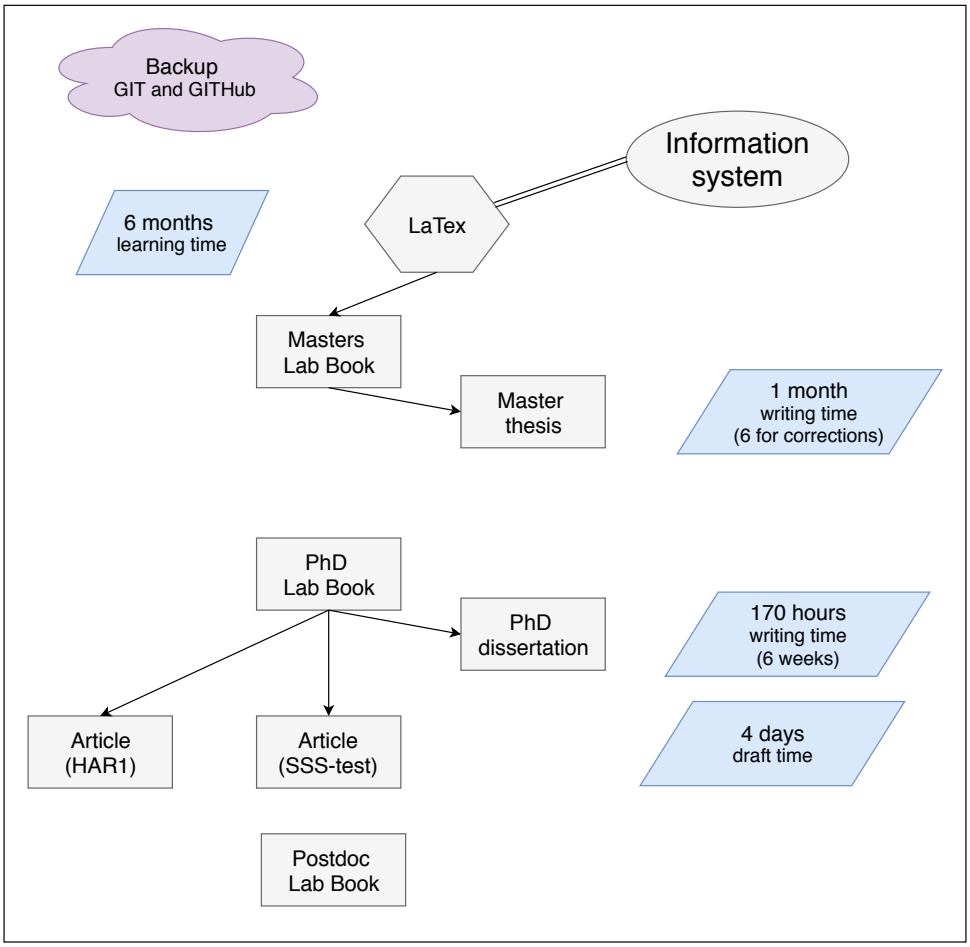

Figura 3.2: My organization scheme of the Lab Book using the Latex language, the exchange of information to other text sources and reference time of learning and writing of documents.

3.9.2 New account request for the cluster at UFBA

Pedro requested a new account for me at the cluster of UFBA to Prof Clemente. It has only 1 Tb of space for us, but can serve purposes of testing and running of data divided into parts. It doesn't have much support since it is being used mostly for a project of Prof Clemente, so we will have to install the software ourselves without root permissions.

3.9.3 Discussing organization system with Amanda & organizing download of 214 metagenomes (Suzana)

- Suzana's table of 214 metagenomes (skywalker): /Documents/Metagenomes_Suzana/Spreadsheet of Aquifers - MG-RAST and SRA (PPM).xlsx
- STU folder of Data: *Fiocruz server: /scratch/amanda/aquifer_suzana*
- status: MG-RAST files downloaded, paired from NCBI had problems with the sra-toolkit
- GITHub Lib (owned by Amanda and with me as a collaborator): `https://github.com/camposamanda/lib_aquifers`
- Script used (successfully) for download of the MG-RAST data (GITHub Rep of Amanda): download_mgm_23jul18.sh

I passed on my organization system to Amanda, a Master student of Pedro's, about GIT and the Lab Book using Latex (Fig 3.1 and 3.2). We created a GITHub page together for Amanda, and the Lib for these experiments of filtering of the 214 metagenomes will be owned by her, with me and Leticia (the undergrad student also working in this experiment) as a contributor. The files from MG-RAST were downloaded successfully, with her script proofread by me, but the ones from NCBI did not work out, so I will check this out by myself tomorrow and pass on to the students after I fix the download.

The size of the input is 1,7Tb for all of the 214 files and the expected since of output is something around 0,7 Tb (700 Gb), that if compressed should occupy something like 500 Gb of space (Fig 3.3). After we clean this data, I can proceed with the other experiments (first profiling and assembly, Fig 3.4).

3.10 24

3.10.1 Transfer of customized DB of Kraken2 from Fiocruz to STU - ongoing

- Fiocruz server access: **ssh -X aquifermicrobiomes@bioinfo03.bahia.fiocruz.br**
- Customized DB (Fiocruz): */media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB*
- New location: *STU*

Like I mentioned in the previous sections, I had a problem transfering the customized DB of Kraken2. Indeed the Fiocruz server was down for the week. Pablo told me the server is back and I will transfer the files to STU now.

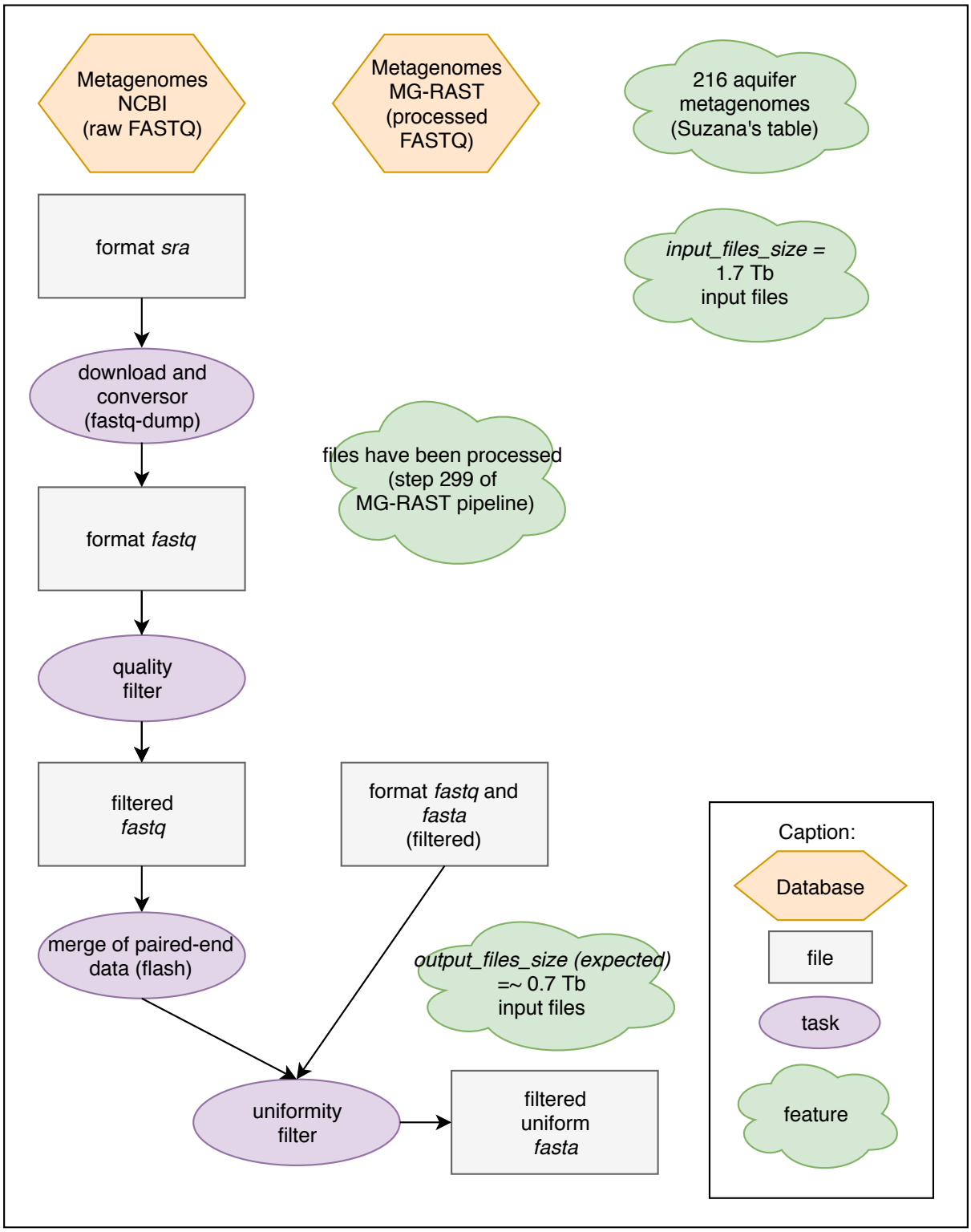

Figura 3.3: This is the workflow to be followed to filter and uniform of the metagenomes prepared by Suzana at her table.

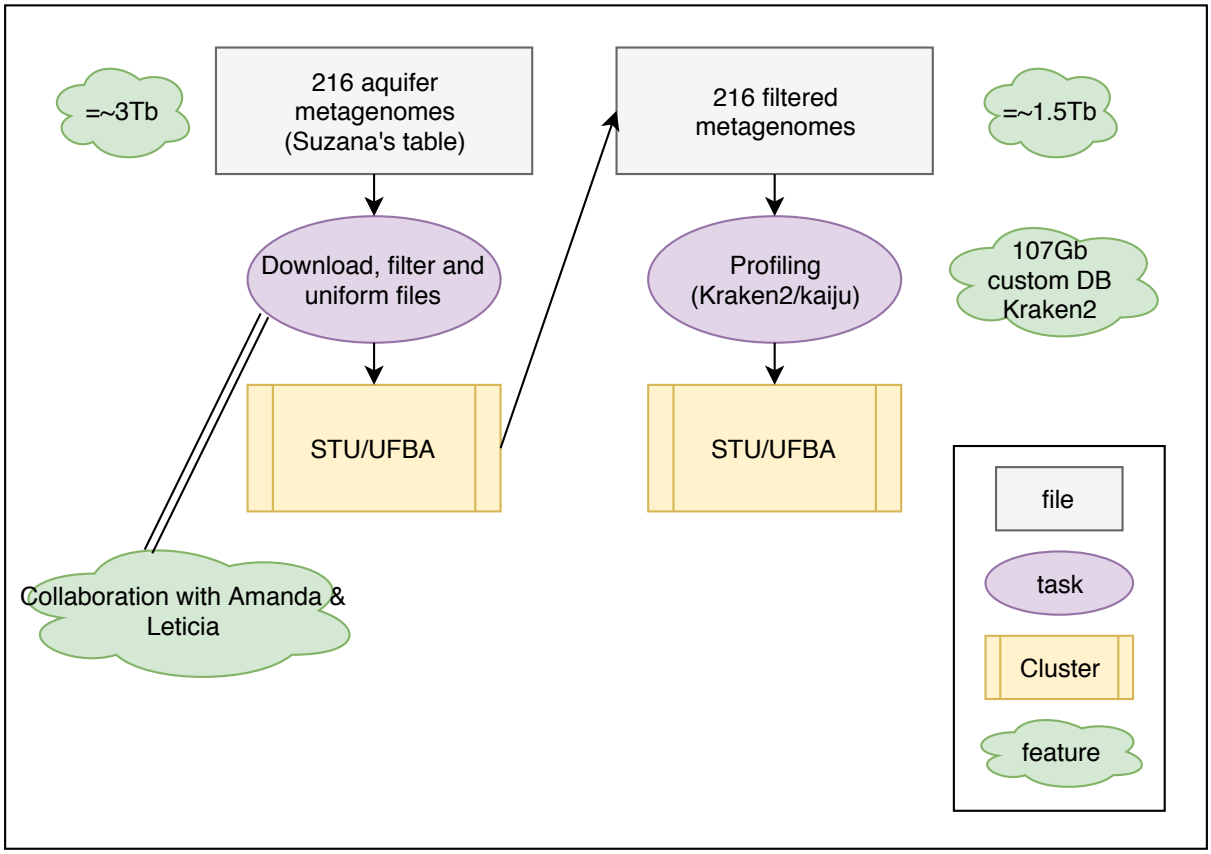

Figura 3.4: This is a global view of the experiments and the cluster in which I will run each analysis. Details of each experiment are not depicted in this scheme.

The files are very large, so I will compress the entire folder with the command Pablo suggested me, transfer it by ssh and then decompress it at the SDU, like below:

tar cf DB.tar.bz2 --use-compress-prog=pbzip2 DB/#to compress the folder, from originally 107 Gb to 40 Gb - *successfull!*

Then I logged in to SDU and moved the file to my home using the indications from this site `http://charmyin.github.io/scp/2014/10/07/run-scp-in-background/`:

nohup scp aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/ microbiomes/tools/kraken2/DB.tar.bz2 . > scp.out 2>&1

pbzip2 -dc -p24 DB.tar.bz2 | tar x #which will use 24 threads to decompress

To use the pbzip2:
module load pbzip2/1.1

3.10.2 Organizing download of 214 metagenomes (Suzana) - NCBI files

- Suzana's table of 214 metagenomes (skywalker): /Documents/Metagenomes_Suzana/Spreadsheet of Aquifers - MG-RAST and SRA (PPM).xlsx
- STU: *Fiocruz server: /scratch/amanda/aquifer_suzana*
- status: MG-RAST files downloaded, paired from NCBI had problems with the sra-toolkit
- GITHub Lib (owned by Amanda and with me as a collaborator): `https://github.com/camposamanda/lib_aquifers`, it contains the scripts we used for download

Yesterday, Amanda and I downloaded the metagenomes from MG-RAST. Today I will proceed with the NCBI files and check why sra-tools did not work yesterday. I tested again at the SDU, both in my home and scratch and I got the same error, which could be a problem of the sratoolkit installation or program. I will test this experiment in skywalker to see if this is indeed the problem and not our command line.

For this I tried installing sratoolkit in skywalker following the developer's directions: `https://ncbi.github.io/sra-tools/install_config.html`, which did not work, so I used instead: sudo apt install sra-toolkit, which worked fine with the test they suggested. While reading the REAMDE.md, I saw that **fastq-dump** was replaced by **fasterq-dump**quoting from the README:

"With release 2.9.1 of 'sra-tools' we have finally made available the tool 'fasterq-dump', a replacement for the much older 'fastq-dump' tool"

I wanted to try the example they gave ($fasterq-dump SRR000001) at the SDU before trying the test skywalker, since it might work better, but they don't have this new tool installed. This command line is supposed to also work with paired-end data.

The tests I did in skywalker (successfull):

fastq-dump --split-files ERR2136697#a small paired end file from the 214 set, already separates the strands into two files

wget ftp://ftp-trace.ncbi.nih.gov/sra/sra-instant/reads/ByRun/sra/SRRSRR390/SRR390728/SRR390728.sra#Downloads the sra, conversion to fastq is required later

The version I have of sratoolkit is 2.9.1-1 and is very recent (from last month), and already has this fasterq-dump. The version of the SDU is the 2.8.2 and doesn't have this new tool. I asked the helpdesk of SDU to install the newest version, since I don't have enough space to do this in my workstation.

They answered me very rapidly and reported we don't have enough space in our group (**ebiodiv** at the SDU), and that this caused the error and not the sratoolkit version:

"O erro "No space left on device"abaixo indica que o seu projeto utilizou todo o espaÃğo de armazenamento.

Verificando a quota de armazenamento do projeto ebiodiv:

SCRATCH - Utilizando 6.63206 TB de um limite total de 10 TB HOMEDIR - Utilizando 500G de um limite total de 500G"

We have a total of 10Tb of space in /scratch/ and 500Gb in our /home/ folders for the whole group. 70% is being used in /scratch/ and 100% is being used in our homes/. I wrote to Pedro to discuss how to proceed regarding removing files.

For our experiment, we need to calculate 3Tb of input data (that can be removed after processing) and around 1.5Tb of output data.

Importantly, to use the tool in the command line (even when not submitting the jobs to the queue), it is necessary to load the module previosuly!

```
module load sratoolkit/2.8.2
$ which fastq-dump
/scratch/app/sratoolkit/2.8.2/bin/fastq-dump

$ fastq-dump --split-files ERR2136697
Read 373378 spots for ERR2136697
Written 373378 spots for ERR2136697
```

3.10.3 Size problems of the 214 set

I calculated the sizes of Suzana's table with the raw values given by her. But for the NCBI files, they are in sra format, which means we are dealing with a substantial larger set. We will need to calculate the space, because the space we currently have might not be enough for our purposes. For an idea of sizes I took the last file of Suzana's table:

Tabela 3.1: Sizes of sra and an example NCBI/SRA paired end file.

Size	File
43Mb	ERR2136697.sra
6,7G	ERR2136697_1.fastq
6,7G	ERR2136697_2.fastq

This means that instead of dealing with a set of 1,7Tb of data, we will have to deal with a lot more, once they are in fastq format (also accounting for paired end files).

In table 3.2 I explicited the sizes we are dealing with at this step of the metagenomics analysis. The helpdesk of SDU also said we reached the limit of used space, so the group will have to remove files, so we can move on with the analysis.

Tabela 3.2: Sizes of files of the first step of the pipeline (Fig. 3.4). Consider that the sizes indicated by a $\sim$ symbol are predicted sizes.

File	Sizes
214 metagenomics set (Suzana) raw fastq	$\sim$ 3Tb
214 metagenomics set (Suzana) processed FASTA	1.5Tb
Kraken2 customized DB	107Gb
kaiju custom DB	80 Gb
profiling matrix Kraken2	?
profiling matrix kaiju	?

3.10.4 Uniformity filter of MG-RAST files of the 214 set

- GITHub Lib (owned by Amanda and with me as a collaborator): `https://github.com/camposamanda/lib_aquifers`
- script to submit prinseq-lite to the queue (Lib):

While I wait for the helpdesk of the SDU to update the sratoolkit, I can filter the files from MG-RAST to uniform them (uniformity filter **prinseq-lite.pl** from Fig. 3.3), and start the next step of the pipeline "Profiling" (Fig. 3.4).

The script and command line to do the uniformity filtering are described by Amanda and Leticia at the file *tutorial_ bioinfo_ pipeline.md* at the folder *Ferramentas_ Metagenomicas* that is at the App Box of the Lab: `https://app.box.com`:

perl prinseq-lite.pl -verbose -fastq file_1.fastq -fastq2 file_2.fastq -min_len 100 -ns_max_p 1 -out_format 3 -out_good good_file -seq_id file_id

Amanda will put the script to submit the jobs of the prinseq-lite at the GITHub repository.

3.11 25

3.11.1 Configuring skywalker

Today I added page up and page down to autocomplete my commands in the shell using the following tutorial:

`https://askubuntu.com/questions/308603/auto-complete-for-often-used-command-line-c`

3.11.2 Upgrading workstation - Ubuntu 17.04 - 17.10 - 18.04 LTS

Today I upgraded the workstation that Suzana wants to use. From 17.04 I had to upgrade to 17.10 using some workarounds since zsync was not working in 17.04. We decided to name the workstation **guarani**, a brazilian aquifer. I changed the workstation name and password and put both in a post-it on it, and wrote it down as well on my notebook and in a paper I attached in the internal door of the wooden bookcase of the lab.

3.11.3 Formatting scolymia

Bertolino, a student of Pedro's, formatted **scolymia**, our first server, to CentOS server without graphic interface. It will be later on transferred to the cluster of UFBA by the STI team, so we can access it by ssh.

3.11.4 Organizing space at the SDU

I sent an email yesterday to the group about space administration at the SDU. There is almost 80% of space occupied and I need half of it free (both at the scratch/ and homes/) for the FASTQ filtering of the 214 metagenomics set in the next couple weeks (Table 3.3). I will organize this with the students and other users of the cluster to free up the required space I will need.

After organizing a big cleanup with the users that needed to remove files from their accounts, we could free a lot of space at the SDU. The new values are in table 3.4.

3.11.5 Checking MG-RAST files of the 214 set

- GITHub Lib (owned by Amanda and with me as a collaborator): `https://github.com/camposamanda/lib_aquifers`
- Old location of MG-RAST files of the 214 set (SDU): */scratch/ebiodiv/amanda.campos/aquifer_db_suzana_table_23jul*
- New location of MG-RAST files of the 214 set (SDU): */scratch/ebiodiv/maria.costa/aquifer_db_suzana_table_23jul*
- ending of FASTQ files: ?file=299.1

I checked the data from MG-RAST [Meyer et al., 2008] me and Amanda downloaded yesterday. They are all single end, and some in FASTA and some in FASTQ format. There are 30 IDs from Suzana's table that are supposed to be files, but we could download only 28 files. I checked the output of the nohup command and apparently, the three last files

Tabela 3.3: Administration of **$home**/ and /**scratch**/ files at the SDU for our group: **ebiodiv**. My space is the requested one for the next couple of weeks for the filtering of FASTQ data and the other users spaces are the ones currently being occupied. * Size requested by Rilquer. In bold are users that need to remove files to free up space. Status: pend: pending, rm: file removal needed. I do not have permissions to use *du* for the individual users of our group, so I asked for the helpdesk of the SDU to check this for me.

User	Name	Used space /scratch/	Used space $home/	status
ebiodiv - group	ebiodiv	10 Tb	500Gb	-
occupied		70%	100%	-
maria.costa	Bia Walter	5 Tb	250 Gb	pend
(requested)	Bia Walter	(50%)	(50%)	pend
amanda.campos	Amanda	1008G	34G	**rm**
bastiaan.dutilh	Bas & Bastiaan	40K	40K	ok
diogo.rocha	Diogo	133G	128K	ok
guilherme.gall	Guilherme	5.9 G	1018M	ok
leticia.cavalcante	LetÃŋcia	2,3T	291G	**rm**
luiz.gadelha2	Luiz	22G	120M	ok
maria.costa	Bia	59 G	40 Gb	pend
marinez.siqueira	Marinez	946 G	920K	ok
matheus.souza	Matheus	4 K	128M	ok
mercia.santos	MÃľrcia	522G	32K	**rm**
pedro.meirelles	Pedro	595G	117G	**rm**
raquel.costa	Raquel	1.6 G	96K	ok
rilquer.silva	Rilquer	2Tb (20%)*/1.3T	25 Gb(10%)*/412M	ok
softwares	-	14G	-	-
Total	13 users	6.7T	481 Gb	-

Tabela 3.4: Administration of **$home**/ and /**scratch**/ files at the SDU for our group: **ebiodiv**. Freed space after re-organization of our users.

User	Name	Used space /scratch/	Used space $home/	status
ebiodiv - group	ebiodiv	10 Tb	500Gb	-
occupied		70%	100%	-
maria.costa	Bia Walter	5 Tb	250 Gb	pend
(requested)	Bia Walter	(50%)	(50%)	pend
amanda.campos	Amanda	714G - **294G** freed	34G - **0G** freed	**rm**
leticia.cavalcante	LetÃŋcia	2,3T - **1438G** freed	291G - **238G** freed	**rm**
mercia.santos	MÃľrcia	522G	32K - ok	**rm**
pedro.meirelles	Pedro	595G - **?**	13G - **104** freed	**rm**
Total	13 users	5.0T	139 Gb	-

had problems. I think that there was a problem in the connection and I will submit these three file fetching again separately with nohup.

Another issue is that there are more files than expected. When we download one ID, it may be we get more than one file per ID from the MG-RAST server. Most have the ending: **?file=299.1**, that is expected because the students chose this step of the MG-

RAST pipeline (check manual for version 4 of 2013). But some files end in **?file=299.1.1** or **?file=299.1.2**, which I think may be due to further steps after the 299. These numbers are not specified in the manual of 2017 for MG-RAST pipeline version 4. I think I might maintain only the original files **?file=299.1** and remove the others, so we have only one representative per ID. But just in case, I sent an email to MG-RAST.

3.12 26

3.12.1 Transfer of customized DB of Kraken2 from Fiocruz to SDU - ongoing

- Check July 24th 2018 for details

On the 24th of July I tried transfering the compressed folder from Fiocruz to SDU. I used the pbzip2 tool like Pablo suggested, but when I decompressed the DB I received an error:

```
cat: cat: No such file or directory
tar: Unexpected EOF in archive
tar: Unexpected EOF in archive
tar: Error is not recoverable: exiting now
```

I decided to try the nohup scp again but with the complete 107 Gb folder. If that doesn't work, I'll try the pbzip2 decompress command again. The full scp didn't work and I realized the number of used cores were too much (-p24), so I changed it to -p10 (command from the 24th).

It still didn't work, so I went back to the Fiocruz server and tried compressing files and folders by parts. Even so, the files get corrupted, so I tried transferring the complete files one by one (the large ones). It still did not work. I am checking with Leticia if the problem is indeed in the connection stability of SDU/LNCC. Tomorrow we will know and I ask for an alternative to the helpdesk.

3.12.2 MG-RAST files of the 214 set - problem detected

- Location of test-download - MG-RAST 214 set (SDU): */scratch/ebiodiv/maria.costa/aquifer_db_suzana_table_23jul_scripts*
- ending of FASTQ files: ?file=299.1
- curl used instead of wget for download

I noticed that the sizes of our downloaded files did not match Suzana'a table! I checked the output reports of nohup and saw that for the bigger files, they were incomplete. **wget**

gives a report when it downloads files with time left and percent of file retrieved. In one of the files, it stopped at 86%. It might be connection problems at the SDU. In any case, I tried **curl** instead to see if it works better. Tomorrow I will be able to see if it worked.

3.13 27

3.13.1 Transfer of customized DB of Kraken2 from Fiocruz to SDU with rsync - ongoing

- Check July 24th 2018 for details

I will try using rsync instead of scp. From discussion on online forums, it seems a better option for unstable connections, and it also seems to be faster:

`https://superuser.com/questions/482134/how-to-send-huge-files-from-one-server-to-a`
`https://www.tecmint.com/rsync-local-remote-file-synchronization-commands/`

nohup rsync --partial aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB/hash.k2d . > rsync_hash.out 2>&1
<input server password>#Control+Z
bg#To initiate the nohup command

The file arrived apparently intact. It has the same size as the original (32 Gb). So I followed with the remaining files of the DB. All small ones have been transfered and the two big remaining ones are folders *library/* and *taxonomy/*, which I ran one after the other using 'bg' with:

nohup rsync -r --partial aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB/library/ . > rsync_library.out 2>&1

3.13.2 MG-RAST files of the 214 set - solved!

- Location of data - 30 files of MG-RAST 214 set (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Lib for download of data (SDU): */scratch/ebiodiv/maria.costa/lib_aquifer_db_suzana_table_scripts*
- Script for single data (at Lib) - 30 IDs: *download_mgm_26jul18_curl.sh*
- ending of FASTQ files: 299.1.fastq
- Size of dataset (30 files): 94 Gb - uncompressed FASTQ files
- curl used instead of wget for download (script at the lib)

The **curl** strategy from yesterday worked very nicely! I downloaded only 30 files, as it was supposed to be, the sizes are all exactly as specified by Suzana's table and the

output of the nohup command accused that 100% of all files were correctly read by curl. All files end with **.299.1.fastq**. Now the files are ready to go to the uniformity filter.

3.13.3 NCBI/SRA files of the 214 set - ongoing

- Location of data - 184 files of NCBI/SRA 214 set (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_ncbi_single*
- Lib for download of data (SDU): */scratch/ebiodiv/maria.costa/lib_aquifer_db_suzana_table_scripts*
- Script for single data (at Lib) - 14 IDs: *download_sra_single_27jul18_dump.sh*
- Script for paired data (at Lib) - 169 IDs: *download_sra_paired_27jul18_dump.sh*
- For unknow data (at Lib) - 1 ID: *download_sra_paired_27jul18_dump.sh*
- ending of FASTQ files: .fastq
- Size of dataset (30 files): XXX Gb - uncompressed FASTQ files
- for downloading the files (original sra) and converting them to fastq: *fastq-dump*

Now that the space is freed in our group's **scratch**/ and **home**/ folders, I can go back to downloading the rest of the data (remaining 184 files from the 214 set). The MG-RAST files have been obtained with *curl*, and for the NCBI files I will use *fastq-dump*. To test the behaviour of the tool I tested separate downloads for a single and a paired end file. Both tests worked out fine. Since there is an extra argument for paired end data, I will run them separately.

module load sratoolkit/2.8.2
fastq-dump SRR4343431
fastq-dump --split-files ERR1527244

The command for obtaining NCBI single-end data using nohup:

nohup ./download_sra_single_27jul18_dump.sh > download_sra_single_dump.nohupout 2>&1 &

The command for obtaining NCBI paired-end data using nohup:

nohup ./download_sra_paired_27jul18_dump.sh > download_sra_paired_dump.nohupout 2>&1 &

The command for obtaining NCBI unkown data (1 file):

./download_sra_unknown_27jul18_dump.sh

This last ID: **SAMN02954299** was not possible to be downloaded. I received the following error:

```
[maria.costa@sdumont13 lib_aquifer_db_suzana_table]$ ./download_sra_unknown \
_27jul18_dump.sh
2018-07-27T18:08:15 fastq-dump.2.8.2 err: item not found while constructing \
within virtual database module - the path 'SAMN02954299' cannot be opened as \
```

`database or table`

I went to the NCBI website to try to retrieve it by hand, this ID is from a BioSample, and the link to "retrieve all samples" is broken. I wrote an email to Suzana to correct the table, but I think we should exclude this ID from the analysis, since the author's data submission is not correct. The only data I could find related to this project is from an assembly: GCA_000205945.1, which we do not want for the present purposes.

3.13.4 Filtering MG-RAST files of the 214 set (uniformity filter)

- GITHub Lib (owned by Amanda and with me as a collaborator): `https://github.com/camposamanda/lib_aquifers`
- New location of MG-RAST files of the 214 set (SDU): */scratch/ebiodiv/maria.costa/aquifer_db_suzana_table_23jul*
- ending of FASTQ files: ?file=299.1

I will use the PRINSEQ tool `http://prinseq.sourceforge.net/manual.html`, as planned by Amanda and Leticia, but only as a simple uniformizing filter. This tool basically does a simple data processing. I will not use the quality in any way for the MG-RAST data, since I want to apply all files to the same process, and will output in FASTA format.

According to their website, "For metagenomic datasets, the exact and 5' duplicates should be removed"

I revised the command line the students passed on to me yesterday to:

Original line:
perl prinseq-lite.pl -verbose -fastq file_1.fastq -fastq2 file_2.fastq -min_len 100 -ns_max_p 1 -out_format 3 -out_good good_file -seq_id file_id

Changed line:
perl prinseq-lite.pl -verbose -fastq file.fastq -min_len XXX -ns_max_p 1 -out_format 1/3 -seq_id file_id

Change reasons:

- Since all MG-RAST files are single ended, there is no need for the -fastq2 argument (which is for paired end data)
- 100 was way to restrictive, but to set a good min_len I need to first examine the distribution of read sizes, since most data come from Illumina
- -out_format I changed to 1 to output in FASTA for decreasing the size by 50%, now that we do not need the quality anymore. It has to be FASTQ for FLASH, since it merges paired end data using the quality
- I have to check an output to understand what exactly is this unique -seq_id

The distribution of the lenghts of the reads of all 30 files of MG-RAST was calculated with:

```
nohup awk 'NR%4 == 2 {lengths[length($0)]++} END {for (l in lengths) {print l, \
 lengths[l]}}' mgm*.3.299.1.fastq > size_distribution_mgm_files.txt | sort -nr > \
 awk_read_length_distribution_nohup.out 2>&1 &
```

3.14 30

3.14.1 TCC defense - Erick Pinheiro

Today one of the students of Pedro defended his undergraduation final project about the evaluation of different configurations of Data Bases into multi-dimensional metrics.

3.14.2 Size problems SDU - $homes solved/ & /scratch/ ongoing

- Experiment 1: download NCBI/SRA data 169 IDs (paired and single)
- Experiment 2: transfer of customized DB Fiocruz - SDU

I am having problems downloading data from NCBI (**fastq-dump** - experiment 1) and transfering data from the Fiocruz server (**rsync** - experiment 2). From the output of both nohups, I see that I have problems of space at /scratch/ and at the $homes. Our $home was full again with 500G, because of my tests with the Fiocruz transfer. So, I excluded old data, and now I have 380G free space. The DB is 109 Gb, so it should be enough space now.

I already downloaded some of the DB, including 32 Gb file **hash.k2d**. I checked with command **md5sum** and saw that the files are indeed the same, which means that rsync was transfered successfully!

1ec6830b56f8684d85ce5ba779578b12 hash.k2d

I further checked online for examples of rsync commands for entire folders and I modified the nohup script a bit for one of the two missing folders to:

nohup rsync --progress --recursive --partial aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB/library/ . > rsync_library.out 2>&1#with the passw and bg

After transfering the library folder, I will do the same with the taxonomy folder (last missing one, if the present command works out). This means that experiment 2 should be easier to solve.

About experiment 1, the problem is a lot more seriuous. The sizes of the files are way larger than I calculated before. I downloaded 52 files, that are likely to be corrupted, and they amount to 1T of data. If we extrapolate, we need at least 7T for these 169 files. Maybe run the files in parts of XT? I will check with Pedro the best strategy to use now.

3.14.3 NCBI/SRA files of the 214 set - size problem detected

- Location of data - files of NCBI/SRA 214 set (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_ncbi_single and _paired*

I detected problems when downloading the data. On the output of nohup, I see the error: *storage exhausted*, which likely means lack of space. I checked online for the errors and I got to the same conclusion. On the online forums, I also saw that when I get statistics, it is likely that the file was correctly downloaded. But for two examples, I get different sizes for paired-end data, which raises the concern wether the data has indeed been correctly downloaded.

For 52 files that were downloaded, for 26 IDs, I have 1T of data. Since I have a list with 169 IDs, the size is going to be a lot larger than I anticipated.

3.14.4 Transfer of customized DB of Kraken2 from Fiocruz to SDU with rsync - ongoing

- Check July 24th 2018 for details

After I checked with **md5sum**, described two sections ago, that rsync works on transfering the custom DB, now I am transfering the two remaining folders. If everything goes well, I can use this DB to profile the 30 files of MG-RAST tomorrow, after filtering them.

3.14.5 Meeting with Pedro: *strategy definition* - MG-RAST file profiling

I had a meeting with Pedro and we defined the next steps, considering the space constraints we have for the NCBI data in /scratch/. For now, since I am having problems downloading the NCBI data, we decided to leave that aside and go on with the MG-RAST 30 files and go on with profiling them.

3.14.6 Meeting with Pedro - size problems at SDU to process 214 files

I told Pedro about the storage problems we are having with the NCBI files at SDU. For now, since we need results, I will focus only on the 30 files of MG-RAST, filter and

profile them with Kraken2 and kaiju. And afterwards, I will come back to the NCBI files, download, filter and profile them in the same way. Since we have storage issues, the strategy is to divide the files and process them in parts.

As a concrete example, I have three of the IDs of Suzana's list (of the 214 files): **SRR944699** in the table below in addition to one I had probed before.

Tabela 3.5: Example of size transformation of one of the files from the NCBI group of the 214 metagenomics file set: file **SRR944699**and file **ERR2136697** that I tested at skywalker (already cited in table 3.1). fastq-dump line: fastq-dump --split-files SRR500735, fasterq-dump equivalent line. fasterq-dump is faster than fastq-dump by at least a factor of 4-5. All these experiments were tests to try calculating file sizes, and the files have been erased, due to storage issues.

Size	File	Stage	Machine	run time
43Mb (311x)	ERR2136697.sra	Suzana's table (**sra** file)	-	-
6,7G	ERR2136697_1.fastq	fastq-dump	skywalker	6h
6,7G	ERR2136697_2.fastq	fastq-dump	skywalker	6h
4.14 Gb (7,7x)	SRR500735.sra	Suzana's table (**sra** file)	-	-
16 Gb	SRR500735_1.fastq	fastq-dump	skywalker	10h
16 Gb	SRR500735_2.fastq	fastq-dump	skywalker	10h
55.07 Gb (4,3x)	SRR944699.sra	Suzana's table (**sra** file)	-	-
118 Gb	SRR944699_1.fastq	fasterq-dump*	guldendraak	12h* ($\geq$2-3 days)
118 Gb	SRR944699_2.fastq	fasterq-dump*	guldendraak	12h* ($\geq$2-3 days)

Considering Suzana's table, the size of all 214 samples given by the table is: **1.7Tb**. That is only considering the sra sizes of the NCBI files. Excluding the MG-RAST 30 files, we have: **169** paired-end files from NCBI and **14** single-end files from NCBI and **1** unknown that I could not download or find any correspondent SRA file, which amount to around 800 Gb of sra files.

If we extrapolate from my table 3.5, we see that the size grows from sra to fastq files by a factor dependent on the sra size. From what we can see, it goes from a factor of 300x for small files to 4x for larger files. If we take a rough mean of the sizes of the 184 files of the NCBI group (including paired and single end), we have a mean size of **8.6Gb**. If we take the closest factor of the table, we have a factor of **8x**.

The total of giga bytes from this set is: **1589Gg**, or **1.6Tb**, and if we multiply this by the factor, we have:

$$1.6 \times 8 = 12.8\ Tb$$

3.14.7 Meeting with Pedro - discussing issue on downloading NCBI group from the 214 set - unsuccessfull

- Folder (skywalker): *Documents/lib_aquifers*
- Script: *download_sra_paired_27jul18_dump.sh*

On our meeting, I also discussed with Pedro about the **fastq-dump** issue.
Issues are:

- Download errors (lack of space in SDU)
- Files arrive corrupted (different sizes between forward and reverse fastq files)

He sent me a script *download_sra_paired_27jul18_dump.sh* that details a pipeline similar to the one I am running for our 214 set. It does not help directly my issues, but have nice tips for future work.

Pedro also sent me e-mail discussing this pipeline, and a very important remark has been made that I shall include in my command line:

```
The fastq-dump command should be run with the split-3 option. If not done \
the reads are not paired properly leading to assembly problems.
```

The email detailing this process is named "Pipeline viromes 2" sent to my by Pedro on July 30, and stored in my gmail folder (posDoc-Salvador).

12.8 Tb of input data from the NCBI files. Importantly, this is a rough estimation, based on the mean size of the files. I have only **5Tb available** at the SDU for processing of these files.

Two links that discuss the fastq-dump problem that may have some useful tips are:

`https://edwards.sdsu.edu/research/fastq-dump/`
`https://edwards.sdsu.edu/research/getting-data-from-the-sra/`

From the orientations I had with Pedro, I will process the files by parts after I process the MG-RAST 30 file group.

3.14.8 Filtering MG-RAST data

- Folder (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Script for slurm queueing: *job_prinseq_submit_paired_metagenomes.bash*

I ran the uniform filtering script for cleaning up the data of the MG-RAST files.

3.15 31 - sick leave

Today I was on sick leave.

Capítulo 4

August 2018

4.1 1

4.1.1 Next TODO plans - ongoing

For the next TODO plans, I based on the last meeting I had with Pedro. I will leave aside for now the problem with fastq-dump and focus on the profiling of the 30 files of MG-RAST.

Tasks:

- Check transfer of Kraken2 custom DB - Fiocruz -> SDU (ongoing)
  - rsync transfer of **library**/ folder: (done)
  - md5sum check of largest file of library: (done)
  - md5sum report for the largest file of library/ folder: 1463a8f58b95789dcddf37922ba322cd Wr6iLltz64.fna (both match - done)
  - rsync is working with nohup to transfer the taxonomy/ folder (done)
  - md5sum report for the largest file of taxonomy/ folder: 9dfb216afc9af45764f38aa5b462a568 nucl_wgs.accession2taxid (both match - done)
  - after **taxonomy**/ folder is transfered, check with rsync *hash.k2d* (done)
  - md5sum report for file **hash.k2d**: 1ec6830b56f8684d85ce5ba779578b12 (both match - done)
  - transfer the rest of the small files of the custom DB (ongoing)
- Check filter of MG-RAST 30 file data
- Prepare profiling with Kraken2 of the filtered data
- Prepare kaiju DB making (part1): RefSeq DB
- Prepare kaiju DB making (part2): custom DB - 700 genomes from aquifer metagenomic data

Scripts for resuming incomplete transfers with rsync using **--append-verify**:

nohup rsync --progress --append-verify aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB/hash.k2d . > rsync_hash_part2.out 2>&1

nohup rsync --progress --recursive --append-verify aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB/taxonomy . > rsync_taxonomy_part2.out 2>&1

4.1.2 Filtering MG-RAST data - ongoing

- Folder (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Script for slurm queueing: *job_prinseq_submit_paired_metagenomes.bash*

The filtering produced an error. I corrected the script and ran again to the slurm queue, with status of job checking: **sacct 197282**.

4.1.3 Meeting about the modelling section of the aquifer's project

Today we had a very productive meeting about the modelling part of the project, and I presented to the group my results and struggles with size and file integrity. We discussed the different methods that the different groups have developed and results with other datasets.

4.2 2

4.2.1 Meeting with Pedro - looking for solutions to problems of file transfer (Kraken DB) and fastq-dump

I had a very productive meeting with Pedro about the problems of file transfer (likely connectivity) of the custom DB of Kraken to SDU and about the fastq-dump download of the NCBI files. We came up with a good strategy for moving forward taking into consideration priorities.

The focus until the next deadline in August is to process the 30 files of MG-RAST. After that we will deal with the NCBI files.

For now, I tested kraken2 at the Fiocruz server, to check if the DB is working. The next step is to complete the filtering of the 30 files and profiling with Kraken2. After that I will check how to build the kaiju ref DB and custom.

4.2.2 Test one small metagenomics file if Kraken2 custom DB is working fine at Fiocruz - successfull

- Server: Fiocruz
- Working folder: */media/bioinfoserver3/Bkp1/microbiomes/tools/kraken2/DB*

According to directions Pedro and I arranged in our meeting, I tested a single file at the Fiocruz server to check if the custom DB was correctly built.

I ran the following:

```
aquiferspablo@bioinfoserver3{DB} /media/bioinfoserver3/Bkp1/microbiomes/tools/ /
kraken2/install-dir/kraken2 --db /media/bioinfoserver3/Bkp1/microbiomes/tools/ /
kraken2/DB mgm4536100.3\?file=299.1 > mgm4536100.3.profiled
```

And got the following output:

```
Loading database information... done.
390745 sequences (42.94 Mbp) processed in 4.472s (5242.4 Kseq/m, 576.11 Mbp/m).
  118122 sequences classified (30.23%)
  272623 sequences unclassified (69.77%)
```

```
aquiferspablo@bioinfoserver3{DB} head mgm4536100.3.profiled
C       HISEQ2_0992:4:1101:9118:2227#TCACCA/1   1392877 157     1392877:2 0:66 A:36 0:19
U       HISEQ2_0992:4:1101:4958:2492#TGACTA/1   0       181     0:147
C       HISEQ2_0992:4:1101:17919:2317#TGACAA/1  80864   146     0:51 80864:6 1224:16 0:10 1224:1 28216:1 0:27
U       HISEQ2_0992:4:1101:20342:2496#TGATCA/1  0       104     0:70
C       HISEQ2_0992:4:1101:2713:2598#TGACAA/1   356     167     0:121 356:5 0:7
U       HISEQ2_0992:4:1101:3504:2664#TGATCA/1   0       141     0:107
U       HISEQ2_0992:4:1101:12527:2722#TAACCA/1  0       132     0:98
U       HISEQ2_0992:4:1101:16920:2721#TGTCCA/1  0       149     0:115
C       HISEQ2_0992:4:1101:2695:2856#TGTCCA/1   33057   173     0:30 33057:17 0:92
C       HISEQ2_0992:4:1101:3286:2778#TGAACA/1   572477  135     0:63 572477:10 0:28
```

The next step regarding this output is to write a script to transform the output into a table with the organism's name, the raw count, the taxID and the reads. Including the unclassified organisms in the table.

Then to check if our 700 genomes are there.

4.2.3 Reading articles for Lab Meeting

Today Mercia will present a paper at our Lab Meeting [Konstantinidis et al., 2017], which we will discuss in the group. In this paper, the authors propose a standardized system to classify and name uncultivated bacteria and archea, mainly based on sequence manipulation using Bioinformatics techniques. Pedro sent around two follow-up short papers to the article, one with a critique to the proposed system by another research group [Oren and Garrity, 2017] and the other a reply to the critique by the authors of the original paper [Konstantinidis et al., 2018].

4.2.4 Lab Meeting

Today MÃľrcia presented the article mentioned in the previous section. We also discussed some issues about the lab.

4.2.5 Feedback from MG-RAST team

I sent an e-mail to MG-RAST about the file downloads (some arrived as $i.1, $i.1.1, $i.1.2), and I was confused about these different files (see July 25th for more details). The team has written me back and said that there was only one file, and that these other endings mean that the download was incomplete and the file arrived truncated. This is a further indication that the internet connection at the SDU is unstable.

4.2.6 Test one small metagenomics file if Kraken2 custom DB is working fine at the SDU - failed

- Server: SDU
- Working folder: */scratch/ebiodiv/maria.costa/test_mgm_file*
- Script (at wk): *job_kraken2_submit.bash*
- Probe file, 1st ID of Suzana's table (at wk): *mgm4536100.3?file=299.1*

Now that I tested Kraken2 at the Fiocruz server and it worked out fine, I am testing the same file at the SDU. I submited the job to the test queue: **cpu_dev**, for the unique file.

4.2.7 Filtering MG-RAST data - failed

- Folder (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Script for slurm queueing: *job_prinseq_submit_paired_metagenomes.bash*

The filtering still produced a similar error as before, that the file was not found.

4.2.8 Transfer of 700 genomes data Fiocruz - SDU - successfull

- Working Folder (SDU): */prj/ebiodiv/maria.costa/Kraken2_DB_700_genomes_custom_data*

I transfered the scripts and the genomes (raw and formatted by me) from the Fiocruz to the SDU.

```
scp aquiferspablo@bioinfo03.bahia.fiocruz.br:/media/bioinfoserver3/Bkp1/ \
microbiomes/tools/kraken2/std-database/library/metagenomes/ncbi-genomes-2018 \
```

```
-06-20.formatted .
```

4.2.9 Check if the 700 genomes have been correctly added to Kraken's default DB - successfull

Check if our 700 genomes are indeed being profiled by Kraken. They are, as can be seen below, by greping the ID '1619077' of our 700 genomes into the profiled test file.

```
aquiferspablo@bioinfoserver3[DB] grep 1619077 mgm4536100.3.profiled
C HISEQ2_0992:4:1101:2712:67351#TGACCA/1.1 2 100 0:9 317025:5 0:30 580331:5 0:1 /
 491077:1 2:1 0:1 2:5 1619077:1 0:7
C HISEQ2_0992:4:1101:6444:94611#TGACCA/2.2 345632 100 0:13 345632:5 1224:2 0:31  /
1619077:2 0:1 1263:5 2:7
```

4.2.10 Help student with MG-RAST email

I helped Suzana today to write an e-mail to the MG-RAST team about retrieving metadata from NCBI samples.

4.3 3

4.3.1 Test one small metagenomics file if Kraken2 custom DB is working fine at the SDU - solved

- Server: SDU
- Working folder: */scratch/ebiodiv/maria.costa/test_mgm_file*
- Script (at wk): *job_kraken2_submit.bash*
- Probe file, 1st ID of Suzana's table (at wk): *mgm4536100.3?file=299.1*
- My test perl script based on a Kraken2 module (at wf): *test_kraken2.pl*

I am trying to test Kraken2 in the SDU using the same file as I did yesterday, when I performed the test at the Fiocruz. I had an error from the slurm output, that Kraken2 has not correctly recognized the path and files of the Database:

```
[maria.costa@sdumont13 test_mgm_file]$ cat slurm-197866.out
sdumont[1024,1190-1193,1271-1272,1436-1437,1472-1473,3137-3146]
sdumont1024 sdumont1190 sdumont1191 sdumont1192 sdumont1193 sdumont1271 \
 sdumont1272 sdumont1436 sdumont1437 sdumont1472 sdumont1473 sdumont3137 \
sdumont3138 sdumont3139 sdumont3140 sdumont3141 sdumont3142 sdumont3143 \
sdumont3144 sdumont3145 sdumont3146
srun: Warning: can't run 1 processes on 4 nodes, setting nnodes to 1
kraken2: database ("/prj/ebiodiv/maria.costa/Kraken2_DB") does not contain \
 necessary file taxo.k2d
```

```
srun: error: sdumont1024: task 0: Exited with exit code 2
```

I went back to the original main Kraken2 perl script, and saw that this error is produced in a module (a separate script called *kraken2lib.pm*). Kraken2 GITHub Lib: `https://github.com/DerrickWood/kraken2/blob/master/scripts/kraken2`

I retrieved the exact *for* block in which the Kraken script tests if all files are contained in the DB (exactly at the *kraken2lib.pm* module), and made a separate perl script test, to see if this small perl script recognizes the files. I ran it off the queueing, to check if (i) my DB is somehow incorrect or if (ii) the queuing system recognized the path wrong.

Running it like this, the script recognizes all files, so I think it is a problem of being able to run the script correctly on the queue.

```
[maria.costa@sdumont13 test_mgm_file]$ perl test_kraken2.pl --db /prj/ebiodiv/ \
maria.costa/Kraken2_DB
Check!
Check /prj/ebiodiv/maria.costa/Kraken2_DB
database ("/prj/ebiodiv/maria.costa/Kraken2_DB") contains necessary file taxo.k2d
database ("/prj/ebiodiv/maria.costa/Kraken2_DB") contains necessary file hash.k2d
database ("/prj/ebiodiv/maria.costa/Kraken2_DB") contains necessary file opts.k2d
```

I wrote to the helpdesk of the SDU about the error, and according to the manual: `http://sdumont.lncc.br/support_manual.php?pg=support#6`, all files required for running jobs need to be in the /scratch/ folder. So I am copying the Kraken DB to the working folder from my home and I am submitting the job again to the development queue.

[maria.costa@sdumont12]$ cp -r /prj/ebiodiv/maria.costa/Kraken2_DB /scratch/ebiodiv/maria.costa/.

The perl script now works, it can find the Database files. I am now running the command of Kraken with a sample file and it worked!

```
[maria.costa@sdumont13 test_mgm_file]$ srun -N 1 -n 1 -c 1 -p cpu_dev kraken2 \
 --db /prj/ebiodiv/maria.costa/Kraken2_DB mgm4536100.3\?file\=299.1 > \
 mgm4536100.profiled
390745 sequences (42.94 Mbp) processed in 11.234s (2087.0 Kseq/m, 229.35 Mbp/m).
  118122 sequences classified (30.23%)
  272623 sequences unclassified (69.77%)
```

I checked the integrity of the output file above with the one I ran at the Fiocruz, and they both match the md5sum profile, so it seems all is working out fine.

4.3.2 Filtering MG-RAST data - ongoing

- Working Folder (SDU): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Script for slurm queueing: *job_prinseq_submit_paired_metagenomes.bash*

I tested one file only at one node at the development queue and it worked out:

```
[maria.costa@sdumont13 data_aquifer_db_suzana_table_mgm]$ srun -N 1 -n 1 -c 1 \
-p cpu_dev /scratch/app/prinseq/0.20.4/bin/prinseq-lite.pl -verbose -fasta \
 mgm4451759.3.299.1.fastq -min_len 80 -ns_max_p 2 -out_format 1 > teste.stat
Estimate size of input data for status report (this might take a while for large \
 files)
done
Parse and process input data
done
Clean up empty files
done
Input and filter stats:
Input sequences: 355,371
Input bases: 144,002,614
Input mean length: 405.22
Good sequences: 344,512 (96.94%)
Good bases: 143,364,844
Good mean length: 416.14
Bad sequences: 10,859 (3.06%)
Bad bases: 637,770
Bad mean length: 58.73
Sequences filtered by specified parameters:
min_len: 10730
ns_max_p: 129
```

I submitted the same script from yesterday, just modifying it to adding a line to cd into the working directory, and submitted it to the queue with sbatch script.sh.

4.3.3 Login to the Buriti server at Rio de Janeiro

The admin of the Buriti server created an account for me there, which I accessed today.

login: mbcosta
senha: bae9Vi2u

I changed my password and could access it by simple ssh from inside the SDU supercomputer.

ssh mbcosta@buriti.lncc.br

To access it from the skywalker workstation, I have to configure the vpnc.

4.3.4 Organizing the files at the SDU

- Server: SDU
- Working folder: */scratch/ebiodiv/maria.costa/test_ mgm_file*
- Script (at wk): *job_ kraken2_ submit.bash*
- Probe file, 1st ID of Suzana's table (at wk): *mgm4536100.3?file=299.1*
- My test perl script based on a Kraken2 module (at wf): *test_ kraken2.pl*

According to the manual file of the SDU: `http://sdumont.lncc.br/support_manual.php?pg=support#6`, the files should be used in a specific manner. Files at the **$homes** are protected and files at /**scratch**/ are to be removed after 60 days after the last modification of it. That means that all files that are to be kept have to remain at the **$homes** folder. Including Libs, results and special DBs. I will create Libs, Data and other file structures at my home and move the scripts to /scratch/ only if I need to run them.

First I asked the SDU if I can clone GITHub repositories at my $home.

4.4 6

4.4.1 Important issues regarding my salary

Today Pedro and I discussed some important issues about my postdoc salary (see Appendix for details).

4.4.2 Organization - reproductible science

A very important concern of Pedro's is to make reproducible science in the aquifer's project. Thinking carefully about implementing this right away, I decided to use private repositories from my GITHub account and to update them both at my workstation and at the SantosDumont server. In this way, I will always keep my scripts updated and create backups. To clone my private repository, I had to (i) create a public key at my user account at the SDU, (ii) add this key to my GITHub and then I could (iii) clone the repository. Guidelines:

`https://help.github.com/articles/adding-a-new-ssh-key-to-your-github-account/`
`https://github.com/settings/keys`

With this I could clone my Lib with:

git clone git@github.com:waltercostamb/Lib-Profiling

This Lib will contain all scripts refering to the pipeline I am currently doing (Fig. 4.1).

4.4.3 Filtering MG-RAST data - 50% done and going

- Working Folder (SDU /scratch/): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Script for slurm queueing for prinseq on FASTA (Lib, temporarily copied for working folder for running purposes only): *job_prinseq_submit_single_metagenomes_MGM_FASTA_03-08-18.bash* - it ws re-named, but is the exact same one I ran on 03/08/18
- Script for slurm queueing for prinseq on FASTQ (Lib, temporarily copied for working folder for running purposes only) *job_prinseq_submit_single_metagenomes_MGM_FASTQ_06-08-18.bash*
- Lib (GITHub rep ($home of SDU)): */prj/ebiodiv/maria.costa/Lib-Profiling*
- Data Folder (SDU $home): */prj/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Report from slurm queue process and stats of prinseq filtering (data folder): *prinseq-stats-FASTA-files-slurm-198350-03-08-18.out*

I checked the data I submitted to the queue last week, and I solved the problem of the system finding my files. The previous errors were due to trying to run the process in the incorrect folder (a cd command was missing). I added a cd working-directory-path to the script for slurm queueing, and the perl filtering tool ran without problems.

It filtered the FASTA files over the weekend (Table 4.1), but there were some FASTQ files within the 30 file set. So I will separate and correctly name the groups (FASTA and FASTQ) and run the tool for the missing files, transforming them to FASTA in the output, since we won't need these qualities for the analysis. We will though for the NCBI/SRA data.

In regard to the FASTQ files that were already filtered, I have the statistics to summarize percentages of filtered (bad) and kept (good) sequences. I cannot know from which file, since the tool did not report input file (Table 4.4).

I modified the script for running the filtering on the FASTQ files to try retrieving the statistics refering to which file. I submitted the job to the queueing system with sbatch. After this step is done, I can follow up with the nest one that is to submit the kept sequences to the profiling step with the customized DB of Kraken2.

4.4.4 Preparing for profiling filtered sequences with Kraken2

- Working Folder (SDU /scratch/): */scratch/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm/filtered_prinseq_good*
- Script for slurm queueing for Kraken2 (Lib, temporarily copied for working folder for running purposes only): *job_kraken2_submit_metagenomes_MGM_06-08-18.bash*
- Lib (GITHub rep ($home of SDU)): */prj/ebiodiv/maria.costa/Lib-Profiling*

After the filtering is finished for all of the 30 files of MG-RAST from Suzana's table, I can profile them with Kraken2 using our customized DB. I am preparing the script today, to run tomorrow right after the filtering is done.

Tabela 4.1: Set of the MG-RAST, 30 files, of Suzana's table. 11 files are in FASTQ format and 19 files are in FASTA format. The biggest files are highlighted in bold.

File	Size	Status
mgm4529964.3.299.1.fastq	9,4M	not filtered
mgm4529965.3.299.1.fastq	18M	not filtered
mgm4536074.3.299.1.fastq	178M	not filtered
mgm4536100.3.299.1.fastq	102M	not filtered
mgm4536472.3.299.1.fastq	15M	not filtered
mgm4536473.3.299.1.fastq	279M	not filtered
mgm4536476.3.299.1.fastq	6,0M	not filtered
mgm4569549.3.299.1.fastq	368M	not filtered
mgm4569550.3.299.1.fastq	61M	not filtered
mgm4569551.3.299.1.fastq	87M	not filtered
mgm4569552.3.299.1.fastq	41M	not filtered
mgm4451759.3.299.1.fasta	144M	filtered
mgm4451761.3.299.1.fasta	130M	filtered
mgm4453297.3.299.1.fasta	91M	filtered
mgm4739174.3.299.1.fasta	213M	filtered
mgm4739175.3.299.1.fasta	826M	filtered
mgm4739176.3.299.1.fasta	52M	filtered
mgm4739177.3.299.1.fasta	2,0G	filtered
mgm4739178.3.299.1.fasta	122M	filtered
mgm4739179.3.299.1.fasta	459M	filtered
mgm4739180.3.299.1.fasta	**28G**	filtered
mgm4739181.3.299.1.fasta	771M	filtered
mgm4739182.3.299.1.fasta	1000M	filtered
mgm4739183.3.299.1.fasta	**29G**	filtered
mgm4739184.3.299.1.fasta	247M	filtered
mgm4739185.3.299.1.fasta	865M	filtered
mgm4739186.3.299.1.fasta	95M	filtered
mgm4739187.3.299.1.fasta	**29G**	filtered
mgm4739188.3.299.1.fasta	371M	filtered
mgm4739189.3.299.1.fasta	898M	filtered

Tabela 4.2: Set of the MG-RAST, 30 files, of Suzana's table. Statistics regarding the 19 files in FASTA format that were filtered.

Good sequences	Bad sequences
96.94%	3.06%
97.01%	2.99%
99.99%	0.01%
95.93%	4.07%
97.47%	2.53%
91.75%	8.25%
97.34%	2.66%
93.93%	6.07%
97.14%	2.86%
92.45%	7.55%
97.48%	2.52%
97.43%	2.57%
92.68%	7.32%
94.53%	5.47%
97.04%	2.96%
87.08%	12.92%
90.72%	9.28%
95.74%	4.26%
97.13%	2.87%

4.4.5 Organizing next steps - Producing matrices of abundances for the MG-RAST 30 file set

Now that I have worked with the data and got a bit more acquainted with the SDU slurm system, I have a better grasp on the pipeline. Right now I am filtering the MG-RAST data, and the next step is to profile the microorganisms with Kraken2, write the script for producing the matrices and running the script to produce the final matrices. After that, I will hand them over to the modelling team that will create the interaction networks. If everything goes according to plan, I can have the matrices by the end of this week.

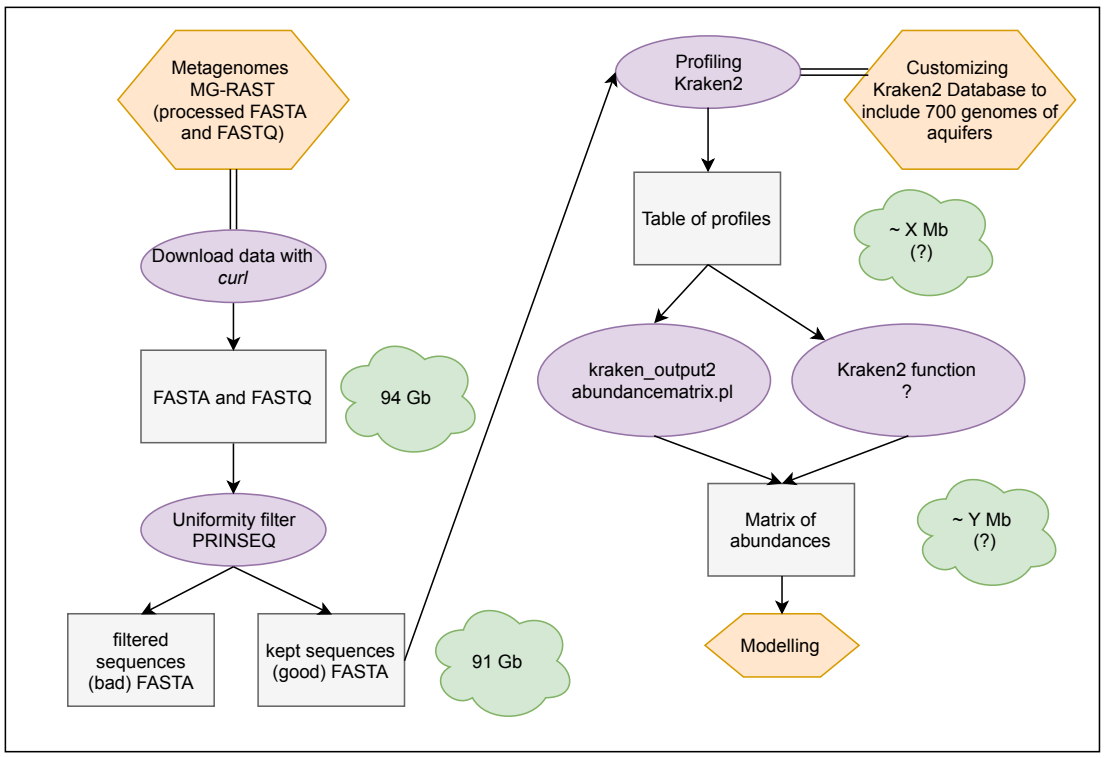

Figura 4.1: Pipeline to produce abundance matrices for the 30 file set of the MG-RAST files of Suzana's table. The input are the downloaded metagenomics data (NGS) of MG-RAST and the output are matrices of abundances of microorganisms. These matrices will be handed over to the modelling team of the Serrapilheira aquifer's project. This pipeline was done at the SDU supercomputer, with the exception of the Kraken2 DB customization, that was done in collaboration with Pablo at the Fiocruz server.

4.5 7

4.5.1 Control over experiments

The job for filtering I submitted yesterday's afternoon is still on queue awaiting for resources. Meanwhile I will test the script of Kraken2. I had a meeting with Pedro and we went together through all steps of figure 4.1 and agreed on the pipeline. I also went through the listings of 06/08/2018 (Appendix) and we agreed on all steps. It might be that Kraken2 has a function to transform the output table into the abundance matrix we want for the modelling. I will check into that.

4.5.2 Filtering MG-RAST data - solved

- Working Folder (SDU /scratch/): */scratch/ebiodiv/maria.costa/data_ aquifer_ db_ suzana_ table_ mgm*
- Script for slurm queueing for prinseq on FASTQ (Lib, temporarily copied for working folder for running purposes only): *job_ prinseq_ submit_ single_ metagenomes_ MGM_ FASTQ_ 06-08-18.bash*
- Lib (GITHub rep ($home of SDU)): */prj/ebiodiv/maria.costa/Lib-Profiling*
- Data Folder (SDU $home): */prj/ebiodiv/maria.costa/data_ aquifer_ db_ suzana_ table_ mgm*
- Report from slurm queue process and stats of prinseq filtering (data folder): *prinseq-stats-FASTQ-files-slurm-cpu_ dev-07-08-18.out*

I waited until 11h and the job was still waiting, so I ran each of the 11 jobs by hand in the queue of development **cpu_dev**. I got all of the files filtered, exactly on the same fashion as the FASTA files, and I kept them in the same folder as yesterday (at the wf):

- filtered_prinseq_good/
- filtered_prinseq_bad/

With these new results, I got all of the statistics, and updated table 4.1. The complete table is table 4.3.

4.5.3 Profiling the filtered sequences with Kraken2 - solved

- Working Folder (SDU /scratch/): */scratch/ebiodiv/maria.costa/data_ aquifer_ db_ suzana_ table_ mgm/filtered_ prinseq_ good*
- Script for slurm queueing for Kraken2 (Lib, temporarily copied for working folder for running purposes only): *job_ kraken2_ submit_ metagenomes_ MGM_ 06-08-18.bash*
- Lib (GITHub rep ($home of SDU)): */prj/ebiodiv/maria.costa/Lib-Profiling*

I tested the script line with the development tool with success:

```
[maria.costa@sdumont13 filtered_prinseq_good]$ srun -N 1 -n 1 -c 4 -p cpu_dev /
 kraken2 --db /scratch/ebiodiv/maria.costa/Kraken2_DB --threads 4 mgm4739176_ /
prinseq_good_vYWn.fasta --output test2.profiled
232330 sequences (40.54 Mbp) processed in 1.912s (7289.2 Kseq/m, 1271.90 Mbp/m).
  70113 sequences classified (30.18%)
  162217 sequences unclassified (69.82%)
```

After that I ran the pipeline with the script for slurm queueing in the development queue, and after I made sure it worked fine, I submitted it to the cpu queue. All "good"files (filtered ones with PRINSEQ) will be profiled with Kraken2. After that I will produce the abundance matrix (Fig. 4.2).

Tabela 4.3: Set of the MG-RAST, 30 files, of Suzana's table. 11 files are in FASTQ format and 19 files are in FASTA format. The biggest files are highlighted in bold.

File	Size	Status
mgm4529964.3.299.1.fastq	9,4M	filtered
mgm4529965.3.299.1.fastq	18M	filtered
mgm4536074.3.299.1.fastq	178M	filtered
mgm4536100.3.299.1.fastq	102M	filtered
mgm4536472.3.299.1.fastq	15M	filtered
mgm4536473.3.299.1.fastq	279M	filtered
mgm4536476.3.299.1.fastq	6,0M	filtered
mgm4569549.3.299.1.fastq	368M	filtered
mgm4569550.3.299.1.fastq	61M	filtered
mgm4569551.3.299.1.fastq	87M	filtered
mgm4569552.3.299.1.fastq	41M	filtered
mgm4451759.3.299.1.fasta	144M	filtered
mgm4451761.3.299.1.fasta	130M	filtered
mgm4453297.3.299.1.fasta	91M	filtered
mgm4739174.3.299.1.fasta	213M	filtered
mgm4739175.3.299.1.fasta	826M	filtered
mgm4739176.3.299.1.fasta	52M	filtered
mgm4739177.3.299.1.fasta	2,0G	filtered
mgm4739178.3.299.1.fasta	122M	filtered
mgm4739179.3.299.1.fasta	459M	filtered
mgm4739180.3.299.1.fasta	**28G**	filtered
mgm4739181.3.299.1.fasta	771M	filtered
mgm4739182.3.299.1.fasta	1000M	filtered
mgm4739183.3.299.1.fasta	**29G**	filtered
mgm4739184.3.299.1.fasta	247M	filtered
mgm4739185.3.299.1.fasta	865M	filtered
mgm4739186.3.299.1.fasta	95M	filtered
mgm4739187.3.299.1.fasta	**29G**	filtered
mgm4739188.3.299.1.fasta	371M	filtered
mgm4739189.3.299.1.fasta	898M	filtered

Tabela 4.4: Set of the MG-RAST, 30 files, of Suzana's table. Statistics regarding the 19 files in FASTA format that were filtered.

File	Good sequences	Bad sequences
unknown FASTA	96.94%	3.06%
unknown FASTA	97.01%	2.99%
unknown FASTA	99.99%	0.01%
unknown FASTA	95.93%	4.07%
unknown FASTA	97.47%	2.53%
unknown FASTA	91.75%	8.25%
unknown FASTA	97.34%	2.66%
unknown FASTA	93.93%	6.07%
unknown FASTA	97.14%	2.86%
unknown FASTA	92.45%	7.55%
unknown FASTA	97.48%	2.52%
unknown FASTA	97.43%	2.57%
unknown FASTA	92.68%	7.32%
unknown FASTA	94.53%	5.47%
unknown FASTA	97.04%	2.96%
unknown FASTA	87.08%	12.92%
unknown FASTA	90.72%	9.28%
unknown FASTA	95.74%	4.26%
unknown FASTA	97.13%	2.87%
mgm4529964 FASTQ	100.00%	0.00%
mgm4529965 FASTQ	99.99%	0.01%
mgm4536074 FASTQ	99.73%	0.27%
mgm4536100 FASTQ	99.72%	0.28%
mgm4536472 FASTQ	99.35%	0.65%
mgm4536473 FASTQ	99.71%	0.29%
mgm4536476 FASTQ	99.02%	0.98%
mgm4569549 FASTQ	100.00%	0.00%
mgm4569550 FASTQ	100.00%	0.00%
mgm4569551 FASTQ	100.00%	0.00%
mgm4569552 FASTQ	100.00%	0.00%

4.5.4 Discussion with Suzana about the pipeline of analysis of the MG-RAST set

I discussed with Suzana in detail about the pipeline of analysis and made a new more general figure of the pipeline (old fig 4.1 and new fig 4.2). She will put that into her scientific report.

The current step is the one of microorganisms profiling with Kraken2. Next I will check if I have to run a script or if Kraken2 has an option I can use to produce this matrix.

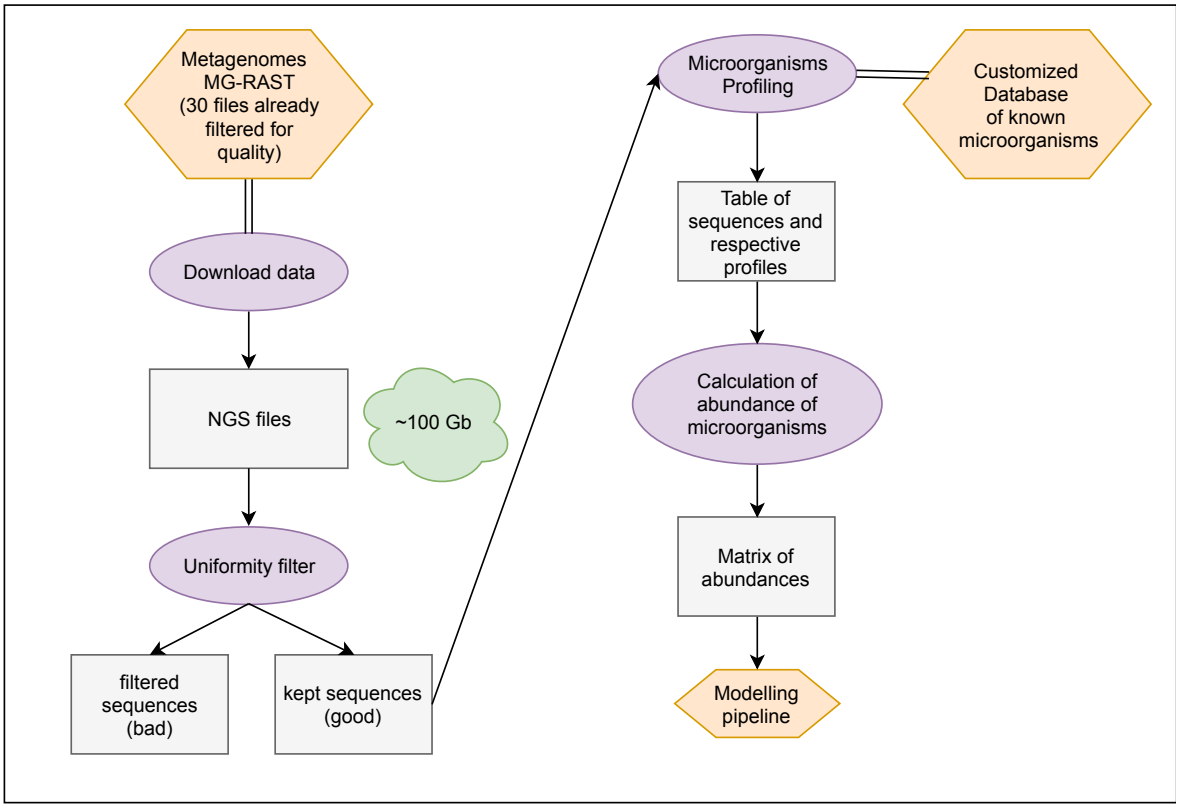

Figura 4.2: Pipeline to produce abundance matrices for the 30 file set of the MG-RAST files of Suzana's table. The input are the downloaded metagenomics data (NGS) of MG-RAST and the output are matrices of abundances of microorganisms. These matrices will be handed over to the modelling team of the Serrapilheira aquifer's project. This pipeline was done at the SDU supercomputer, with the exception of the Kraken2 DB customization, that was done in collaboration with Pablo at the Fiocruz server.

4.6 8

4.6.1 Investigating Kraken2 for matrix function - solved

- Kraken2 manual: `http://ccb.jhu.edu/software/kraken/MANUAL.html`

Before I develop a script to produce the abundance matrix for the modelling project using the output of Kraken2 as input, I checked the manual for extra functions (Fig. 4.2). The argument –**report file.report** does exactly that. I tested with the following command line, and it produces a second output with the abundances.

```
[maria.costa@sdumont13 filtered_prinseq_good]$ srun -N 1 -n 1 -c 4 -p cpu_dev /
 kraken2 --db /scratch/ebiodiv/maria.costa/Kraken2_DB --threads 4 mgm4739176_ /
prinseq_good_vYWn.fasta --output test4.profiled --use-names --report test4.report
232330 sequences (40.54 Mbp) processed in 2.165s (6439.4 Kseq/m, 1123.61 Mbp/m).
  70113 sequences classified (30.18%)
  162217 sequences unclassified (69.82%)

[maria.costa@sdumont13 filtered_prinseq_good]$ head test4.report
 69.82% 162217 162217 U 0 unclassified
 30.18% 70113 1091 R 1 root
 29.67% 68938 64 R1 131567   cellular organisms
 27.59% 64097 989 D 2     Bacteria
 23.52% 54643 1286 P 1224       Proteobacteria
 16.54% 38427 136 C 28216         Betaproteobacteria
 16.29% 37844 307 O 80840           Burkholderiales
 15.09% 35051 301 F 119060             Burkholderiaceae
  8.40% 19508 684 G 106589               Cupriavidus
  7.98% 18534 13876 S 119219                 Cupriavidus metallidurans
```

I asked the modelling group if this output is what they need. If so, the matrices will be ready after the job reaches the queue, and I will pass that on to them (awaiting resources since yesterday).

4.6.2 Developing script for making abundance matrix

- Lib (at GITHub): *Lib-Profiling*
- Script (at Lib): *abundanceMatrix.pl*

I got inputs from the modelling team and Pedro, and they oriented me with the appropriate format of the abundance matrix. I should make two matrices, one with percentages and the other with the raw counts at the given group. The format is exemplified in tables 4.5 and 4.6.

Tabela 4.5: Matrix of abundances (i) with percentages. This matrix is a subset of the output of Kraken2, after it was filtered by Phyla, Family, Order, etc. I named "Group" as a general term, which will be substituted by the specific "Phyla", "Order", etc.

Metagenome ID	Group1	Group2	GroupN
mgmXXXX1	x1%	y1%	z1%
mgmXXXX2	x2%	y2%	z2%

Tabela 4.6: Matrix of abundances (ii) with raw counts. This matrix is a subset of the output of Kraken2, after it was filtered by Phyla, Family, Order, etc. I named "Group" as a general term, which will be substituted by the specific "Phyla", "Order", etc.

Metagenome ID	Group1	Group2	GroupN
mgmXXXX1	a1	b1	c1
mgmXXXX2	a2	b2	c2

4.6.3 Initial results about the 700 genomes influence on the aquifer profiles

- Lib (at GITHub): *Lib-Profiling*
- Script (at Lib): *which_700_genomes_are_present_in_the_profiles.sh*
- Argument needed for script (file of taxIDs): *unique_taxIDs_from_the_700_genomes.txt*

I made an experiment today that tells us about the influence of the 700 genomes in the metagenomes profiles that I am making with Kraken2.

I added these 700 genomes into the DB of Kraken previously, because they came from aquifer metagenomes and we want to improve the accuracy of our profiling experiment. I wrote a bash script that outputs exactly the taxIDs that are present in the aquifers. As an example, one of our metagenomes (from the 30 file set of the MG-RAST of Suzana's table), got the output below, showing that our 700 genomes are indeed present, but at a very small percentage. After I get all the 30 metagenomes profiled, I will make a report for each one.

```
0.00% 2 2 S 1619077               candidate division TM6 bacterium GW2011_GWF2_28_16
0.01% 12 12 S 1618707                 Candidatus Moranbacteria bacterium GW2011_GWE1_35_17
0.00% 3 3 S 1618333             Berkelbacteria bacterium GW2011_GWA2_35_9
0.00% 6 6 S 1618895                Parcubacteria group bacterium GW2011_GWC1_36_9
0.00% 6 6 S 1618811                Parcubacteria group bacterium GW2011_GWA2_38_13
0.00% 6 6 S 1618945                Parcubacteria group bacterium GW2011_GWD2_43_10
0.00% 3 3 S 1618926                Parcubacteria group bacterium GW2011_GWC2_42_12
0.00% 1 1 S 1618820                Parcubacteria group bacterium GW2011_GWA2_42_14
0.00% 3 3 S 1618341             candidate division CPR1 bacterium GW2011_GWA2_42_17
0.00% 1 1 S 1618651                Candidatus Giovannonibacteria bacterium GW2011_GWB1_43_13
0.01% 12 12 S 1618443                 Candidatus Gottesmanbacteria bacterium GW2011_GWA2_43_14
0.00% 8 8 S 1618846                Parcubacteria group bacterium GW2011_GWA2_47_7
0.00% 9 9 S 1618830                Parcubacteria group bacterium GW2011_GWA2_44_13
0.00% 6 6 S 1618502                Microgenomates group bacterium GW2011_GWA2_46_7
0.00% 1 1 S 1618847                Parcubacteria group bacterium GW2011_GWA2_47_8
0.00% 4 4 S 1619044                Candidatus Magasanikbacteria bacterium GW2011_GWA2_56_11
0.00% 2 2 S 1619077               candidate division TM6 bacterium GW2011_GWF2_28_16
0.00% 1 1 S 1618337             Berkelbacteria bacterium GW2011_GWE1_39_12
```

4.6.4 Talk of Prof Flora Bacelar

Today I attended a talk of Prof Flora Bacelar about her work. She is a physicist with a collaboration with Pedro about complex systems. Her student Rafael is building a network to calculate interaction between microorganisms of the aquifer's project.

4.7 9

4.7.1 Developing script for making abundance matrix

- Lib (at GITHub): *Lib-Profiling*
- Script (at Lib): *abundanceMatrix.pl*

Pseudocode 1 contains the code for making the script of the abundance matrices.

Algorithm 1 AbundanceMatrix Algorithm

```
1:  procedure PRODUCE ABUNDANCE MATRICES FROM THE OUTPUT OF KRAKEN
2:      list of all Kraken reports ← list of files
3:      for files do
4:          open file
5:          if line matches rank then
6:              get metagenome ID
7:              get specific group
8:              get abundance
9:              add metagenome ID, group and abundance to hash tree
10:     Get list of groups from hash tree
11:     Use hash tree and list of groups to print matrix of abundances
```

4.8 10

4.8.1 Preparing for the Lab Meeting of the 15th of August

On the 15th of August we will have a general Lab Meeting, in which we will discuss organization things of the Lab. I have some suggestions in regard to the organization of our workstations and our groceries. I am preparing two spreadsheets that will help us organize these issues. I am also making a plan of our system architecture.

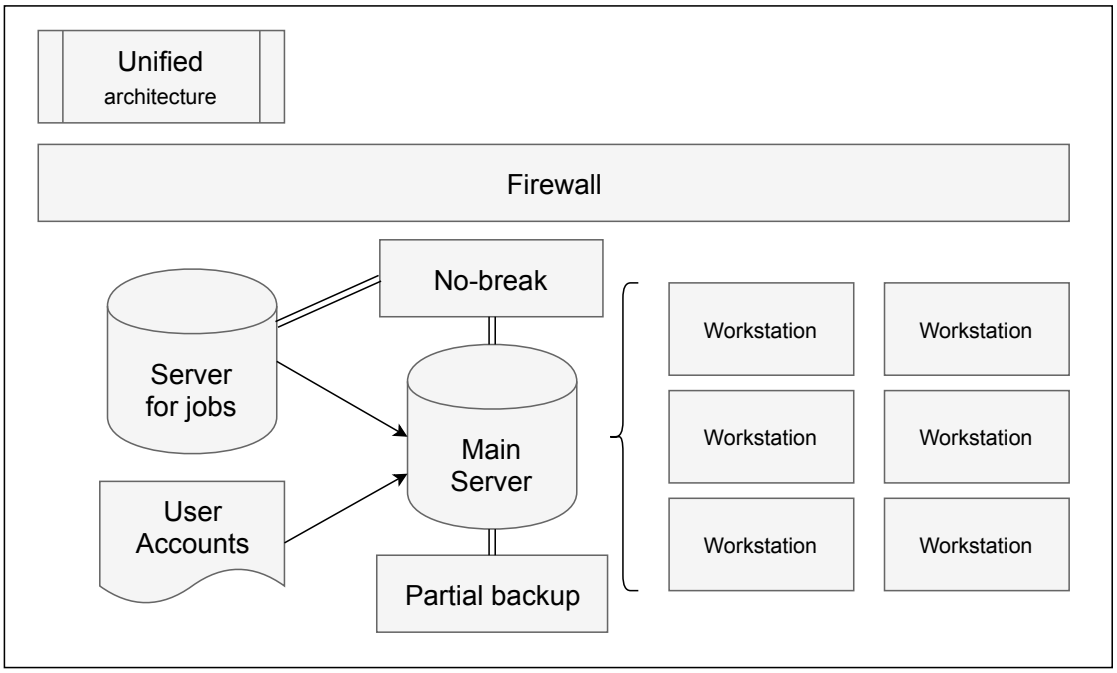

Figura 4.3: Architecture of a unified system.

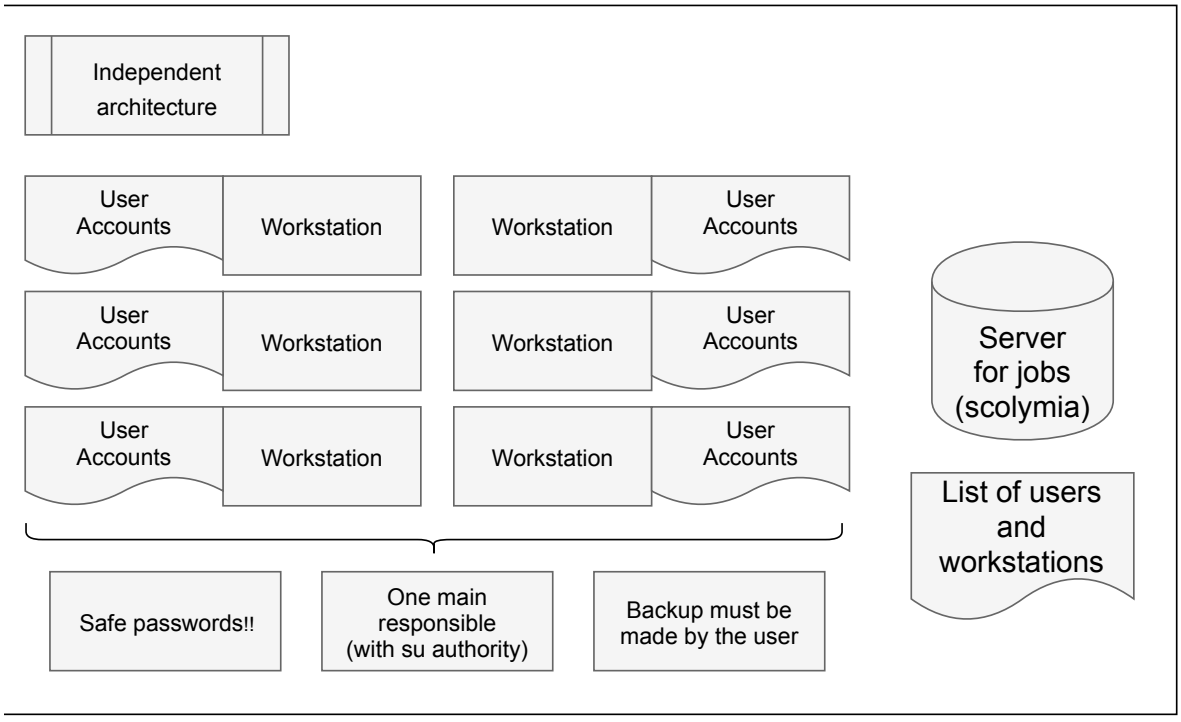

Figura 4.4: Architecture of an independent system.

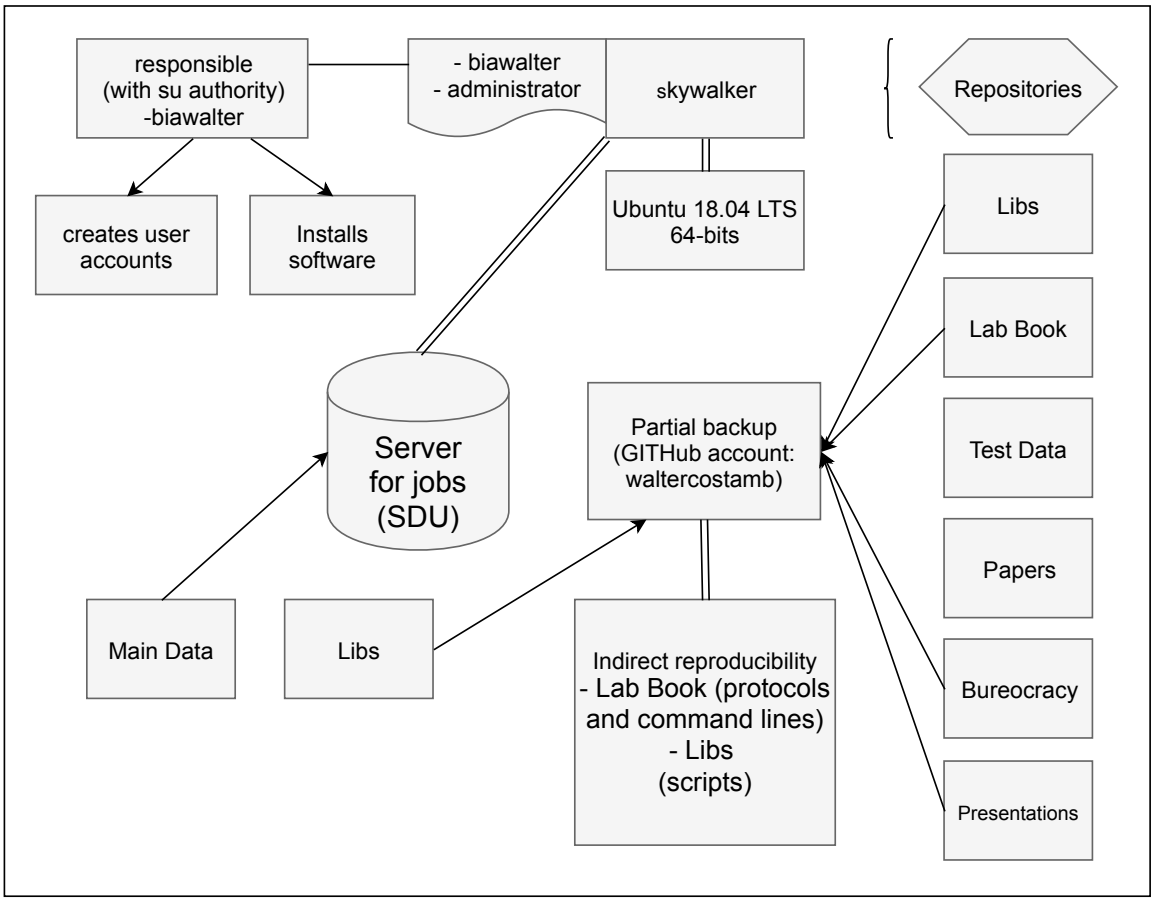

Figura 4.5: Architecture of my workstation - skywalker.

4.9 13 - 14

4.9.1 Debugging *abundanceMatrix.pl* and producing abundance matrices - solved

- Lib (at GITHub): *Lib-Profiling*
- Script (at Lib): *abundanceMatrix.pl*
- Matrices and README (skywalker): */home/biawalter/Documents/Metagenomes_Suzana/abundance_matrices*
- Matrices and README (SDU): */prj/ebiodiv/maria.costa/data_aquifer_db_suzana_table_mgm*
- Matrices and README (OSF): *OSF Storage/Bioinformatics_Results/Aquifers/matrices_MG_RAST.zip*

I was generating the matrices and found an error in the script. It was printing the same specific group more than once (multiple occurances of groups of header of tables 4.5 and 4.6). This occured because I was using normal arrays to store these groups. I changed the internal data structures from normal arrays to hashes (associative arrays), and avoided the repetition of the groups.

To see the usage of the script, one should run:
perl abundanceMatrix.pl --help

I ran the script 20 times for all of the 10 groups specified by the Kraken2 manual, with both options of abundances: (i) raw percentaged (field zero of Kraken2 output) and (ii) raw count (field one of Kraken2 output). I only changed the appropriate field, but using the general command below:

```
perl /prj/ebiodiv/maria.costa/Lib-Profiling/abundanceMatrix.pl --reg_exp_input /
mgm --reg_exp_ext matrix --field FIELD --group GROUP > metagenomes_abundances_ /
MG_RAST_Unclassified_field0_raw_percentages.tsv
```

With, **FIELD** being: (i) 0 or (ii) 1, and **GROUP** being: (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies.

4.10 15 - 2018.02 organizing Meeting for the Meirelles-lab

Today all of us from the Meirelles Lab met to discuss common organization issues. We set up protocols and TODO's. I was responsible for organizing the computer system. I defined our architecture and set up one administrator per workstation, and users for them. The administrators will be responsible for maintaining and administrating the machines, adding users, among other administrative tasks.

Importantly, I brainstormed about our system architecture and organization, taking into account the use of the machines. We have an independent architecture of the lab currently (Fig. 4.6), so I suggested one administrator per workstation and that everyone update the systems to the latest stable Ubuntu version: Ubuntu 18.04 LTS 64-bits. I

will be on this organization task for the next days, until we have all workstations named, taged, updated and with an administrator.

4.11 16 - 17

4.11.1 TODOs from yesterday - all soved

- Report from 2018.02 Meeting (skywalker): *Documents/Meirelle_ Lab/Ata-Reuniao-2018_ 2_ Matheus.docx*

I did all TODO's related to yesterday's organizing meeting. Organized my slack, OSF and Trello accounts, to communicate and store files in the right channels. All details related to the organization of our Lab is in the report.

4.11.2 Graphical representation of the pipeline

- Workflow (draw.io): *postdoc/*
- Figures (pdf, jpg and png) - GIT: *Lab_ Book_ aquifers*
- OSF: Storage -> Bioinformatics Results -> Aquifers

I prepared a figure for Pedro's talk at Campinas, a workflow of the aquifer metagenomes profiles. For the next ones, Pedro asked me to build them in a way that they fit better in a power point slide, and also that the letter sizes are bigger.

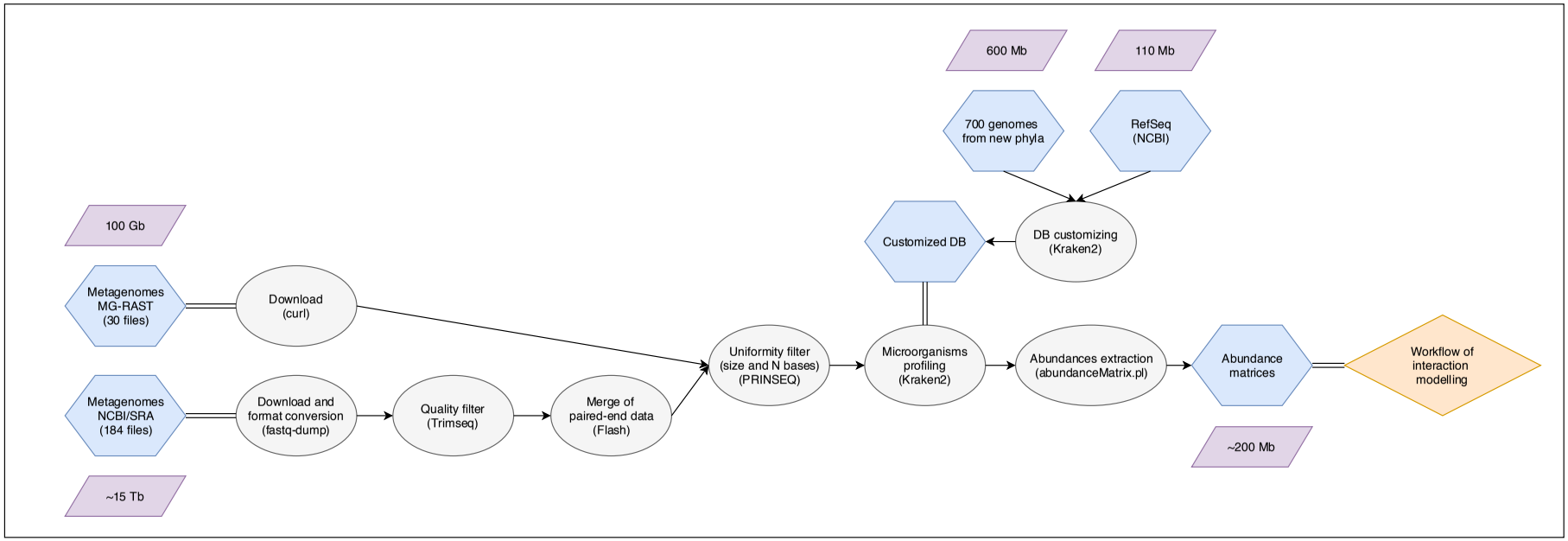

Figura 4.6: Pipeline of microorganisms profiling of aquifer metagenomes. Slide for Pedro's presentation on Sunday at Campinas. Pipelines (workflows) are marked in yellow, Databases in blue, sizes as purple and steps of the pipeline in grey.

4.11.3 Formatting of treponema - solved

Today I formatted skywalker according to directions we agreed upon on the meeting of August 15th 2018.2. The new name of the machine is **treponema**, it has Ubuntu 18.04 LTS 64-bits installed now. I backed-up all data into an external HD, and after it was finished, I copied files again to the machine. I was automatically assigned to superuser.

4.11.4 Spontanous meeting - extra step of filtering required into the pipeline

We had an extraordinary meeting with Pedro and the group, because Pedro noticed something important on the networks of Gabriel Bertolino. Gabriel used my matrices to produce his networks, and Pedro noticed eukaryotes were also contained there. They should be removed, so I will include a further step into the pipeline to exclude all eukaryotes from the matrices.

I will re-do the pipeline (Fig. 4.6) to include the filtering step.

4.12 20

4.12.1 Configuring new workstation treponema - solved

Today I installed Texlive, Texworks, R, Rstudio, GIT and vpnc on treponema. Data seems to be fine.

4.12.2 Update graphical representation of the pipeline - include eukarya filter step

- Workflow (draw.io): *postdoc/*
- Figures (pdf) - GIT: *Lab_Book_aquifers*
- OSF: Storage -> Bioinformatics Results -> Aquifers

I updated my graphical representation of the pipeline to include the filtering step we discussed last friday (Fig. 4.7). Also increasing the letter sizes and formatting the visualization to better fit into power point slides.

4.12.3 How to format a workstation to install an Ubuntu system

I will need to format another workstation, that Suzana uses. For that I learned how to format a machine with Ubuntu 18.04 LTS 64-bits and erase the older partitions. One needs to first download the the correct Ubuntu ISO version, store it in a flashdrive, restart the computer, enter the ISO system in the flashdrive and then follow up with the installation. I learned this following the Ubuntu tutorials: `https://tutorials.ubuntu.com/` and also with the help of Gabriel Bertolino.

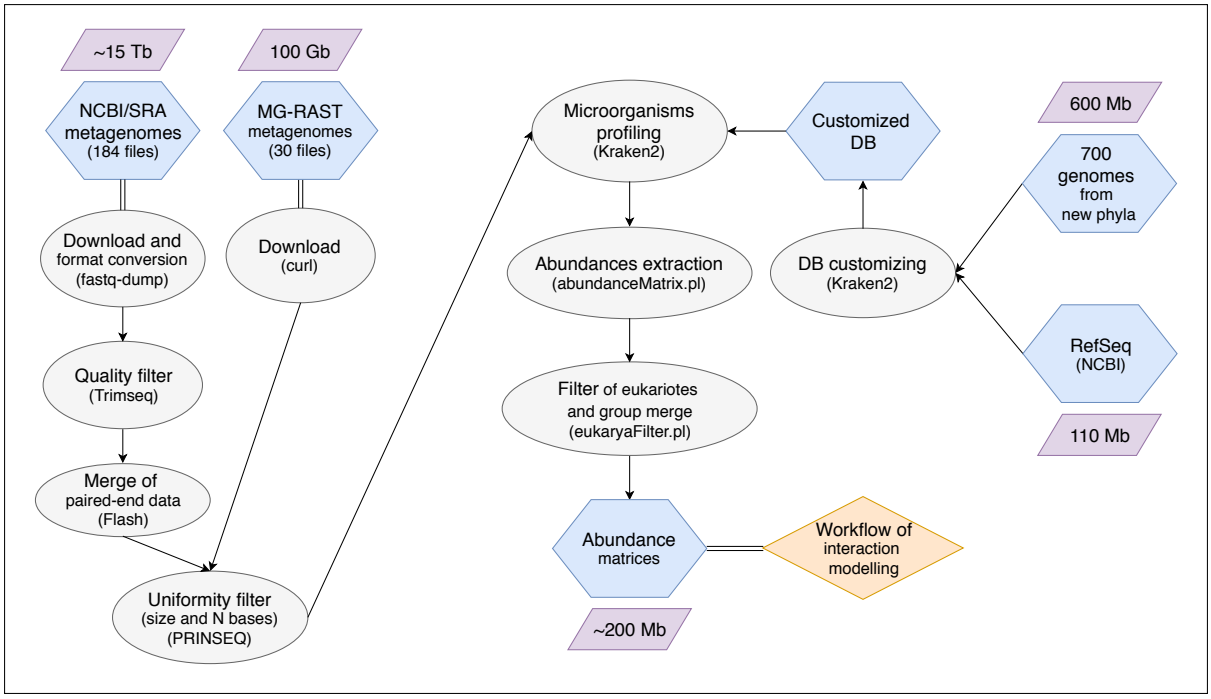

Figura 4.7: Pipeline of microorganisms profiling of aquifer metagenomes. Slide for Pedro's presentation on Sunday at Campinas. Pipelines (workflows) are marked in yellow, Databases in blue, sizes as purple and steps of the pipeline in grey.

After downloading the ISO version of Ubuntu, I transfered it to my flashdrive with the following commands (passed on to me by Gabriel):

```
lsblk
umount /dev/sdx
dd if=FILE of=/dev/sdx bs=2MiB
```

4.12.4 Meeting with Leticia to understand Kraken2 files part1 - solved

4.12.5 Meeting with Rafael about the workstations of the Lab

4.13 21

4.13.1 Organizing the workstations of the Lab - solved

After all users put their scheduling times into the form sent out by Matheus: `https://docs.google.com/spreadsheets/d/1n59ntGgx8StEYy0vc1mwXYejpwJa97m82V93S11PSME/edit?ts=59caa8bb#gid=0`.

`https://docs.google.com/spreadsheets/d/135cc6uT2u5jQX3DBJJSZ3KpL6RHJJZ0c5vZuFu8tVX/edit#gid=0`

Tags into Trello

```
0.00% 4       0       P       74152         Elusimicrobia
0.00% 3       0       C       447830          Endomicrobia
0.00% 3       0       O       1783344           Endomicrobiales
0.00% 3       0       F       1783343             Endomicrobiaceae
0.00% 3       0       G       1408194               Endomicrobium
0.00% 2       2       S       1408281                 Endomicrobium proavitum
0.00% 1       1       S       1408204                 Candidatus Endomicrobium trichonymphae
0.00% 1       0       P1      99260           environmental samples
0.00% 1       0       S       167965            uncultured Termite group 1 bacterium
0.00% 1       1       S1      471821              uncultured Termite group 1 bacterium phylotype Rs-D17
0.26% 891     0       D       2759        Eukaryota
0.26% 891     0       D1      33154         Opisthokonta
0.26% 891     0       K       33208           Metazoa
0.26% 891     0       K1      6072              Eumetazoa
0.26% 891     0       K2      33213               Bilateria
0.26% 891     0       K3      33511                 Deuterostomia
0.26% 891     0       P       7711                    Chordata
0.26% 891     0       P1      89593                     Craniata
0.26% 891     0       P2      7742                        Vertebrata
0.26% 891     0       P3      7776                          Gnathostomata
0.26% 891     0       P4      117570                          Teleostomi
0.26% 891     0       P5      117571                            Euteleostomi
0.26% 891     0       P6      8287                                Sarcopterygii
0.26% 891     0       P7      1338369                               Dipnotetrapodomorpha
0.26% 891     0       P8      32523                                   Tetrapoda
0.26% 891     0       P9      32524                                     Amniota
0.26% 891     0       C       40674                                       Mammalia
0.26% 891     0       C1      32525                                         Theria
0.26% 891     0       C2      9347                                            Eutheria
0.26% 891     0       C3      1437010                                           Boreoeutheria
0.26% 891     0       C4      314146                                              Euarchontoglires
0.26% 891     0       O       9443                                                Primates
0.26% 891     0       O1      376913                                                Haplorrhini
0.26% 891     0       O2      314293                                                  Simiiformes
0.26% 891     0       O3      9526                                                    Catarrhini
0.26% 891     0       O4      314295                                                    Hominoidea
0.26% 891     0       F       9604                                                      Hominidae
0.26% 891     0       F1      207598                                                      Homininae
0.26% 891     0       G       9605                                                        Homo
0.26% 891     891     S       9606                                                          Homo sapiens
0.06% 210     1       D       2157        Archaea
0.05% 188     0       P       28890         Euryarchaeota
```

Figura 4.8: Part of the output file of Kraken2 that contains the data we want to filter (all eucariotes), to be implemented on the new filtering step of the pipeline (Fig 4.7), specifically on script *eukaryaFilter.pl*.

4.13.2 Meeting with Leticia 1 - Organizing

4.13.3 Meeting with Leticia 2 - understand Kraken2 files part2-solved

Check with groups to gather into one in the matrix
filtrar com eucariotos: environmental samples

4.13.4 Think about script to filter matrices

4.13.5 Seminar presentation

https://docs.google.com/document/d/1v4NN4vD7B9b6rQqtwWeUlrsXc3k0569WCQWXMkmHZbY/edit?ts=5b7d6f99

https://docs.google.com/document/d/1libTn2zYp0s-GwgRGWWIoK_7Kl6I4vYiKq3FsQblQqw/edit?ts=5b7d70cf

4.14 TODO's - from July and August

- Check data for download of metagenomes with Leticia, Amanda & Suzana, check steps of filtering - done
- Download metagenomics data with Leticia and Amanha, and check if it fits in the small server of the Lab - done
- Solve problem of data download at STU (Leticia, Amanda) - done
- Filter metagenomics files - MG-RAST STU (Leticia, Amanda) - done
- Test Kraken2 with three files (small, large, huge sizes) - done
- Customize DB of kaiju STU - done
- Run Kraken with the metagenomes that we filtered (Leticia, Amanda) - done
- Review proposal, send it to Pedro (deadline: 20/07/18) - done (sent by deadline)
- erase Data from maya: *Documents/posDoc/Profiling/tools_comparison*
- Check implementation of pipeline that Pedro sent (`https://www.nature.com/articles/s41564-018-0171-1`) - on hold (priority is producing the matrices for the modelling project of the 30 files)
- Check if Kraken2 has a function to transform its output into an abundance matrix - done
- Make report on my difficulties throughout this first part of the project: mainly **storage space** - to add into the proposal for premium account at the SDU - ongoing
- Check label maker price in Germany to bring to the lab (130 R$ in Salvador) - done
- Lab Meeting on the 15th: bring organization suggestions: (i) workstations, su, names, users and administrators, (ii) lab organization (water, tea, etc), responsibles and lists with people names - done
- Lab Meeting on the 15th: meeting of 30' with Pedro once/twice a week with me to discuss updates on the project - ongoing

Referências Bibliográficas

[Brown et al., 2015] Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H., and Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain bacteria. *Nature*, 523(7559):208.

[Konstantinidis et al., 2017] Konstantinidis, K. T., Rosselló-Móra, R., and Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *The ISME Journal*, 11(11):2399.

[Konstantinidis et al., 2018] Konstantinidis, K. T., Rosselló-Móra, R., and Amann, R. (2018). Reply to the commentary âĂIJuncultivated microbesâĂŤin need of their own nomenclature?âĂİ. *The ISME journal*, page 1.

[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.

[Oren and Garrity, 2017] Oren, A. and Garrity, G. M. (2017). Uncultivated microbesâĂŤin need of their own nomenclature? *The ISME journal*, 12(2):309.

[Sczyrba et al., 2017] Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., et al. (2017). Critical assessment of metagenome interpretationâĂŤa benchmark of metagenomics software. *Nature methods*, 14(11):1063.

[Silva et al., 2018] Silva, G. G., Haggerty, J. M., Cuevas, D. A., Doane, M., Dinsdale, E. A., Dutilh, B. E., and Edward, R. A. (2018). Ecological implications of metagenomics data analysis. *in preparation*.