



**Comparative Microbiome
Analysis – version 2.0**

Index

| | | |
|---------|--|----|
| i. | Using CoMA with VirtualBox | 3 |
| i.i. | Installation | 3 |
| i.ii. | Using an USB device | 4 |
| i.iii. | Notes | 4 |
| ii. | Using CoMA with Singularity | 5 |
| ii.i. | Installation | 5 |
| ii.ii. | Notes | 6 |
| iii. | CoMA installer for Linux | 7 |
| iii.i. | Installation | 7 |
| iii.ii. | Notes | 8 |
| iv. | Tutorial | 9 |
| v. | Citation | 21 |
| vi. | The CoMA output | 21 |
| vii. | What is new in CoMA 2.0? | 23 |
| viii. | Software list | 24 |

i. Using CoMA with VirtualBox

The VirtualBox version of CoMA is recommended for all entry-level users as well as for users, who are using a Windows or Mac operating system. It demands relatively large system requirements but the operation is intuitive and easy to learn for anyone with basic computer knowledge.

i.i. Installation

1. First, you need to download the free “[VirtualBox](#)” software. Choose the correct platform package for your operating system.

OPTIONAL: Download the *VirtualBox* Extension Pack if you want to have additional features, including support for USB 2.0/3.0 devices.

2. Install *VirtualBox* on your PC using the downloaded installer file.

OPTIONAL: Install the *VirtualBox* Extension Pack by running the downloaded file. Make sure to install the extension pack with the same version as your installed version of *VirtualBox*.

3. Run the downloaded file “**CoMA_2.0.ova**” and import it to *VirtualBox* without changing any settings. Make sure that there is at least 50 GB free disk space available. This process may take a few minutes.

OPTIONAL: To transfer files from the virtual environment to your hosting system (e.g. Windows), you need to activate the USB ports. Start “**VirtualBox**” and select the newly installed virtual environment (“**CoMA_2.0**”). Hit now the “**Settings**” button and go to the “**USB**” section. Check the box “**Enable USB Controller**” and select either USB 2.0 or USB 3.0, depending on your computer hardware. Chose USB 2.0 if you are not sure.

OPTIONAL: The system memory for the installed virtual environment (“**CoMA_2.0**”) is initially set to 2048 MB. This is sufficient for most analysis within the CoMA analysis pipeline. Nevertheless, some procedures (e.g. blasting huge databases) need more memory and you need to increase it manually. This is also recommended if you want to increase generally the performance of CoMA. Please be aware that your system must be capable of providing the additional memory capacity! It is not recommended to assign more than 50% of your system’s memory to the virtual machine! Otherwise, there might not be enough memory left for the host operating system. You can also change the number of CPU cores (default setting: 2 cores). Again, assigning too many resources to the virtual environment may severely weaken the performance of your host system.

- a. Select the CoMA virtual environment (“**CoMA_2.0**”) and hit the “**Settings**” button
 - b. Go to “**System**”
 - c. Under the subsection “**Motherboard**”, you can adjust the base memory according to your requirements. Be aware that the machine might not start if you assign too much memory! Furthermore, assigning a large proportion of your system memory to the virtual machine may considerably slow down your computer.
 - d. If you want to adjust the number of CPU cores assigned to the virtual machine, go to the subsection “**Processor**” and apply your changes.
 - e. After all changes are done, hit the **OK** button.
4. Start now the virtual environment by hitting the “**Start**” button.

If you receive the error message “*VT-x/AMD-V hardware acceleration is not available on your system*”, you need to enable virtualization in the BIOS of your system:

- a. Reboot your computer and enter the BIOS (Consult the manual of your computer or of your mainboard if you are not able to enter the BIOS)
 - b. Find the configuration items related to the CPU (maybe also under the headings “Processor”, “Chipset” or “Northbridge”)
 - c. Enable virtualization for your system. The setting may be called “**VT-x**” or “**AMD-V**”.
 - d. Enable “**Intel VT-d**” or “**AMD IOMMU**” if it is available.
 - e. Save your changes and reboot the system.
5. Login with the password: **Passw0rd!**
6. All programs needed are pre-installed except for “USEARCH”. Each user must install it individually due to license restrictions:
 - a. Move to the **usearch** directory:


```
$ cd /usr/local/Pipeline/lotus_pipeline/usearch
```
 - b. Download USEARCH from the tool’s webpage:


```
$ wget https://drive5.com/downloads/usearch11.0.667_i86linux32.gz
```
 - c. Unzip the USEARCH file:


```
$ gunzip *
```
 - d. Rename the file:


```
$ mv * usearch
```
 - e. Give execution rights to the file:


```
$ chmod 111 *
```

i.ii. Using an USB device

For file import and export, we suggest using an USB drive. Prior the first use, you need to activate the device manually. Thereafter, the USB device will be recognized automatically after plugging it in and you can use it immediately. To activate a new device, go to “**Devices**” (either in the task bar or at the bottom of your screen) and select “**USB**”. Here, all connected USB devices are listed and you can tick the entry for your USB drive. **ATTENTION:** in some cases, this step may freeze your mouse or your keyboard. However, this problem shall be fixed after restarting the virtual system.

i.iii. Notes

The installed virtual environment represents a complete Ubuntu 20.04 LTS Linux installation with unrestricted functionality. The system is appropriate for further usage; nevertheless, no support can be offered on topics unrelated to CoMA.

Some users reported problems when using VirtualBox on MacOS. This is most likely related to conflicting audio settings in VirtualBox. Disabling specific audio devices (e.g. the microphone) may resolve the issue.

If there are any problems with the installation or the use of CoMA, please contact the CoMA support (coma-mikrobiologie@uibk.ac.at).

ii. Using CoMA with Singularity

Singularity is recommended for all Linux users but particularly for those who are using a non-Debian-based system since the Linux installer (Chapter iii) may not work properly in this instance. We strongly encourage the use of Singularity when running CoMA on a HPC (High Performance Computing) cluster system.

ii.i. Installation

1. First, install Singularity using your distribution's package manager (e.g. **apt-get**, **yum**) or from source.
2. Download the CoMA Singularity image (**CoMA_2.0.tar.gz**) from the CoMA webpage (<https://www.uibk.ac.at/microbiology/services/coma.html>). This can be done either manually or using a command line downloader like **wget**.
3. Unpack the zipped archive including the CoMA image as well as an overlay image, which is needed for storing user-specific data:

```
$ tar -xzf CoMA_2.0.tar.gz
```

4. For continuation of the installation (and to use CoMA later on), you need to start a virtual Singularity session:

```
$ singularity shell -B ~/.Xauthority --overlay CoMA_2.0.ovl CoMA_2.0.sif
```

ATTENTION: Both, the CoMA image as well as the CoMA overlay image must be in the current working directory!

5. Due to license restrictions, USEARCH must be downloaded and installed individually by the user:

- f. Move to the **usearch** directory inside the virtual CoMA environment:

```
$ cd /usr/local/Pipeline/lotus_pipeline/usearch
```

- g. Download USEARCH from the tool's webpage:

```
$ wget https://drive5.com/downloads/usearch11.0.667_i86linux32.gz
```

- h. Unzip the USEARCH file:

```
$ gunzip *
```

- i. Rename the file:

```
$ mv * usearch
```

- j. Give execution rights to the file:

```
$ chmod 111 *
```

6. OPTIONAL: Finally, the data files for the CoMA tutorial need to be copied to the user's Home directory. This step is optional and can be skipped if you do not want to go through the CoMA tutorial.

- a. Go to the Home directory:

```
$ cd
```

- b. Unpack the zipped archive to the current working directory:

```
$ tar -xzf /usr/local/Pipeline/CoMA.tar.gz
```

ii.ii. Notes

The CoMA Singularity image includes a minimalized version of Ubuntu 20.04, which is embedded in the host system. Inside the virtual system, you have access to the Home directory of the host system. This can be used for providing input data files as well as for storing the result files of the CoMA analysis runs.

ATTENTION: Do not start the virtual system with administration rights (i.e. using *sudo*)! Otherwise, the host's Home directory will not be accessible from inside the virtual system.

There are known issues with the overlay filesystem of Singularity v3.x in combination with AppArmor in the host system (AppArmor is used in Debian-based distributions). You can check for these issues by running

```
$ libreoffice --calc
```

in the container before starting the CoMA pipeline. If no window is opening and you receive an error message (similar to "Fatal exception: Signal 11") instead, quit the Singularity session and run the following command on your host system:

```
$ sudo apparmor_parser -R \  
    /etc/apparmor.d/usr.lib.libreoffice.program.oosplash \  
    /etc/apparmor.d/usr.lib.libreoffice.program soffice.bin
```

Now you can start a new Singularity session and run CoMA.

ATTENTION: This step needs to be repeated after a reboot of the system (or after restarting or reloading the AppArmor service on the host)!

If you are using a SSH connection (e.g. to work on an HPC cluster), keep in mind that you need to activate X11 forwarding in your SSH client! Otherwise, you might not be possible to see the graphical user interface of CoMA, resulting in a termination of the pipeline.

Singularity can also be used on Windows or MacOS, for detailed information please refer to the respective documentations on the Singularity webpage (<https://singularity.lbl.gov/install-windows>, <https://singularity.lbl.gov/install-mac>).

If there are any problems with the installation or the use of CoMA, please contact the CoMA support (coma-mikrobiologie@uibk.ac.at).

iii. CoMA installer for Linux

This option is recommended for all users with a Debian-based Linux system (e.g. Ubuntu, Linux Mint, etc.). Moreover, the direct installation of CoMA on a local Linux system is the only way of using CoMA without any kind of virtualization.

iii.i. Installation

1. Download the CoMA installer (**CoMA_2.0_installer.sh**) from the CoMA webpage (<https://www.uibk.ac.at/microbiology/services/coma.html>). This can be done either manually or using a download manager like **wget**.

2. Execute the installer. You need administration rights to continue!

\$ sudo bash CoMA_2.0_installer.sh

ATTENTION: The CoMA installer file must be in the current working directory!

3. During the installation, the user is asked several questions. Give your answer and press **ENTER** to continue. We recommend the following answers in order to guarantee the most flexibility of CoMA later on:

- a. “Continue (y/n)?”. Enter “**y**” and press **ENTER** to continue.

- b. “For similarity based taxonomic assignments LotuS can either use”

\$ 3

- c. “Do you want to install a reference database 16S database for similarity based 16S annotations?”

\$ 8

- d. You need to read and except the licence restrictions of the SILVA database now. Enter “**y**” and press **ENTER** to continue.

- e. “Do you want to”

\$ 1

- f. “Do you want to”

\$ 1

- g. Finally, we need to confirm that we want to install USEARCH later:

\$ 0

4. Due to license restrictions, USEARCH must be downloaded and installed individually by the user. You need administration rights for the upcoming steps!

- a. Move to the **usearch** directory:

\$ cd /usr/local/Pipeline/lotus_pipeline/usearch

- b. Download USEARCH from the tool’s webpage:

\$ sudo wget https://drive5.com/downloads/usearch11.0.667_i86linux32.gz

- c. Unzip the USEARCH file:
*\$ sudo gunzip **
 - d. Rename the file:
*\$ sudo mv * usearch*
 - e. Give execution rights to the file:
*\$ sudo chmod 111 **
5. OPTIONAL: Finally, the data files for the CoMA tutorial need to be copied to the user's Home directory. This step is optional and can be skipped if you do not want to go through the CoMA tutorial.
- a. Go to the Home directory:
\$ cd
 - b. Unpack the archive to the current working directory:
\$ cp /usr/local/Pipeline/CoMA.tar.gz .

iii.ii. Notes

The CoMA installer for Linux is intended to be used on Debian-based systems. We thoroughly tested it on Ubuntu 18.04/Ubuntu 20.04 and are confident that it also works properly on other Debian derivatives. However, we cannot guarantee its functionality on other Linux distributions! If you are encountering problems with the CoMA installer, we recommend considering the Singularity option.

If there are any problems with the installation or the use of CoMA, please contact the CoMA support (coma-mikrobiologie@uibk.ac.at).

iv. Tutorial

CoMA is a pipeline for amplicon sequencing data analysis with input files in FASTQ format (either uncompressed or zipped with **gzip**). The pipeline represents a series of steps from the processing of the supplied data to the computation of results. Each step can be skipped if you are continuing a former project or if you want to leave out specific steps. However, keep in mind that some steps are mandatory when running the dataset for the first time and leaving them out leads to severe problems due to missing dependencies! This manual demonstrates the usage of CoMA with a detailed systematic tutorial using a preinstalled (reduced) example dataset.

1. Start CoMA from Linux terminal. **REMARK:** You can start CoMA from any directory; however, if you start a new project, your project folder will be created in the current working directory. You may want to go to another directory (using **\$ cd**) first and start CoMA afterwards.

\$ coma

2. “Do you want to start a new project?” → **Yes**

Click **No** if you want to continue or recalculate an already existing project and select afterwards the folder of the project. **ATTENTION:** To do so, go to the directory of your project, click on it and click the **OK** button. Do not enter the project directory! This is also demonstrated in step 74.

3. Enter the title of your project and confirm with the **OK** button. Keep in mind that the title must not include any space characters (you maybe want to replace them with underscores: “_”). **ATTENTION:** An old project with the same title as a new project will be overwritten and all your data will be irrevocably lost! However, CoMA will show a warning message in this case.
4. Select the forward and reverse sequence files for your project. For paired-end sequences, be aware that forward and reverse files must be located in the same directory to import them! The files for the tutorial can be found in the home directory (e.g. */home/coma/CoMA/Example/* when using VirtualBox). We select all of them (six files in total) and hit the **OK** button.

5. “Do you want to assign your files and choose the number of CPUs?” → **Yes**

ATTENTION: This step is mandatory when analysing a dataset for the first time! Your inputs are then saved for all upcoming runs and you do not need to repeat this step unless you want to change any settings.

6. You can select the number of CPU cores assigned to the analysis. For the tutorial, we select **2** (with either using the mouse or the arrow keys on your keyboard) and hit the **OK** button.
7. “Are you using paired-end reads?” → **Yes**

Click **No** if you are analysing single-end reads or paired-end reads that were already merged. You may also consider doing a single-end run for an unmerged paired-end dataset if your results were not satisfying. In this case, you can select either your forward or your reverse sequence files. Keep in mind that reverse sequences are often worse in terms of sequence quality.

8. Enter the (longest) common part of all your forward sequence files as well as of all your reverse sequence files and hit the **OK** button. For the tutorial we enter the following lines:

_R1.fastq.gz

_R2.fastq.gz

ATTENTION: If you are analysing single-end reads (answer **No** in the previous step), this step will ask you for the common sequence of all your files, since you do not have forward and reverse files.

You can directly verify your input by checking the output in the Linux terminal. There should be one sample name for each forward/reverse sequence pair or for each single-end sequence file, respectively.

9. “Do you want to merge the files?” → **Yes**

Forward and reverse reads of all files detected in step 8 are now merged. This leads to a single FASTQ file for each sample, which you can find in the following directory:

.../Data/processed_reads

ATTENTION: This step is mandatory when analysing a dataset for the first time!

ATTENTION: This step is entirely skipped in case of a single-end run (please see step 7)!

10. “Do you want to check the quality of the input files?” → **Yes**

If you want to see the results of the quality check, open the .html files in this folder:

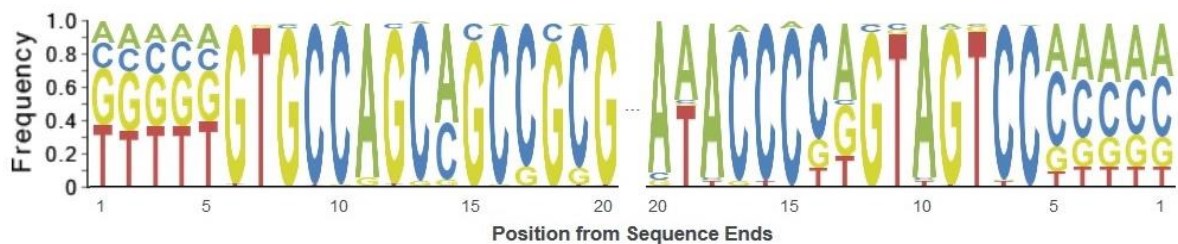
.../Data/quality_reports/quality_before_filtering.

11. “Do you want to trim your files and/or apply quality filtering?” → **Yes**

ATTENTION: This step is mandatory when analysing a dataset for the first time! If you do not want to do any kind of trimming or quality filtering, type in **0**, **0**, **999999**, **0**, **0** and **999999**, and proceed (with **OK**). This will leave your sequences untouched.

12. Enter all the information for the trimming/quality filtering process. You can get the information you need from the quality check in step 10. Keep in mind that it is important to know the length of your forward and reverse primers, as well as of potential barcode sequences. Moreover, the reports may help defining suitable parameters for quality filtering based on fragment length and PHRED quality score (PQS).

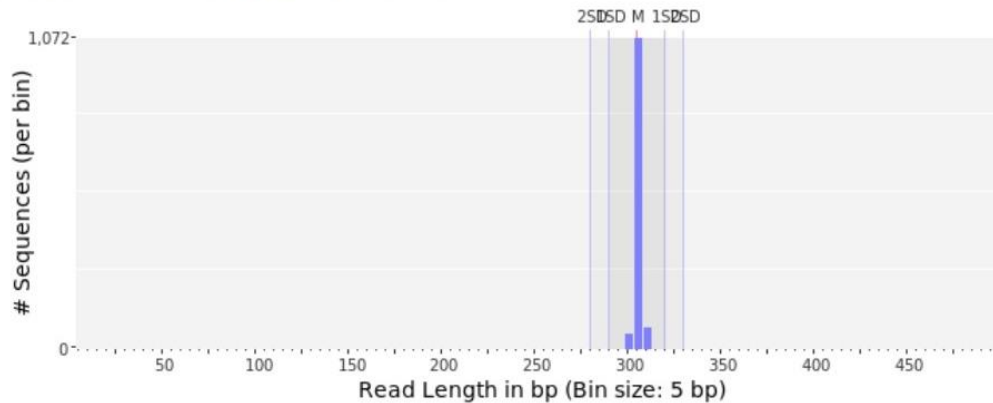
Quality check of sample 1:



This plot shows the first/last 20 base pairs (bp) of your merged reads. It reveals that there are barcode sequences with a length of 5 bp each extended to both sides (to be recognized by an even distribution of ACGT between all positions). Subsequently, you can find the primer sequences, which are common for all reads (and thus show a high frequency of one single nucleotide at each position). You need to add the length of the primer to the length of the barcode to get the correct length for the trimming process.

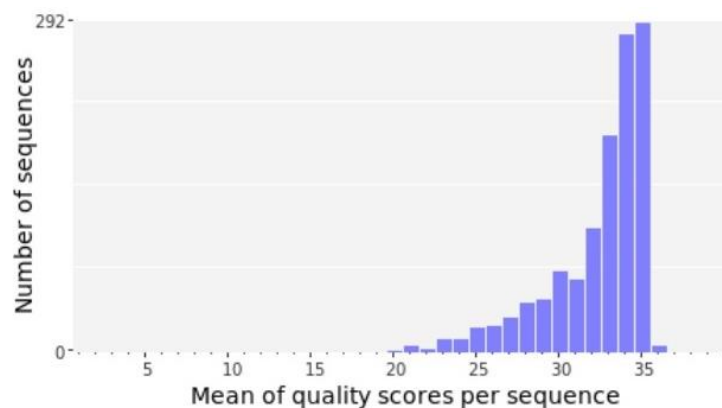
In our example, the length for trimming is **24** (19 + 5) on the forward side (5' end) and **25** (20 + 5) on the reverse side (3' end).

Mean sequence length: **302.14 ± 13.23 bp**
 Minimum length: **89 bp**
 Maximum length: **484 bp**
 Length range: **396 bp**
 Mode length: **302 bp with 814 sequences**



With the information from the length distribution analysis, we can set the limits for the fragment length filtering. You need to subtract the length of both sequences trimmed away from the mode length of all reads. Afterwards you can define a minimum and maximum fragment length for the filtering based on the histogram shown above. A convenient value would be $\pm 5\%$.

The mode length of all reads in sample 1 is 302 bp. After subtracting both trimmed sequences ($302 - 24 - 25 = 253$) you can set the limits to **270** and **240** for example.



You can define a minimum mean PQS for each read based on the PQS histogram. All reads with mean PQS below the threshold will be discarded. A high mean PQS leads to confident results but keep in mind that potentially many sequences are excluded from analysis and therefore information might get lost. We type in **30**, a typical value for minimum average PQS, representing a base call accuracy of 99.9%. For more information on PQS, visit the [Illumina homepage](#) or the [information file](#) given by the company.

Finally, you can define a maximum number of allowed ambiguous bases (Ns). Ambiguous bases are introduced to sequences as soon as specific positions could not be identified as particular base (A, C, G, T). Sequences with a high proportion of ambiguous bases are generally considered as of poor quality and the frequent occurrence of Ns may lead to erroneous results. We type in **5**, which should be a suitable value for most datasets, and click the **OK** button. **REMARK:** You do get information on ambiguous bases also in the quality reports (section: *Occurrence of N*). This may help you setting a suitable cut-off level for your dataset.

You can find the files containing all accepted (“good”) and all discarded reads in the “[good_sequences](#)” or “[discarded_sequences](#)” folders in this directory:

[.../Data/filtered_reads](#)

13. “Do you want to check the quality of your trimmed/quality filtered files?” → **Yes**

If you want to see the results of the quality check for the accepted (“good”) and the discarded reads, open the .html files in the “[good_sequences](#)” or “[discarded_sequences](#)” folders in this directory:

[.../Data/quality_reports](#)

14. “Do you want to align your sequences and make a taxonomic assignment?” → **Yes**

ATTENTION: This step is mandatory when analysing a dataset for the first time!

15. Select an aligner tool for the process. You can choose between three different options: [RDP](#), [blast](#) and [lambda](#). Be aware that the RDP aligner tool can only search against the RDP database. For both of the other tools, you can choose between different databases, including a custom database, in the upcoming step.

For the tutorial, we choose “**blast**” and hit the **OK** button (or double-click on “**blast**”).

ATTENTION: There might be a problem when analysing sequencing files with very long description lines (> 1000 characters) with blast. Truncating them might solve the problem, otherwise, please contact the CoMA support (coma-mikrobiologie@uibk.ac.at)!

16. Choose between various reference databases, based on your preferences and the type of your samples (bacteria, fungi, protists): **SLV**: [Silva](#) LSU (23/28S) or SSU (16/18S), **GG**: [greengenes](#) (only 16S available), [UNITE](#) (ITS focused on fungi), [PR2](#) (SSU focused on protists) or **HITdb** ([HITdb](#) is a database specialized only on human gut microbiota). Databases can be combined with “,”, with the first having the highest priority (e.g. “*PR2,SLV*” would first use PR2 to assign OTUs and all unassigned OTUs would then be searched for with SILVA). If you want to use a custom database, enter **CD** and select the database file and the taxonomy file in the following steps. For more information on custom databases and how the files need to be formatted, please refer to the online documentation:

http://psbweb05.psb.ugent.be/lotus/images/CustomDB_LotuS.pdf

ATTENTION: you cannot combine a custom database with any other database!

ATTENTION: you need to have write permissions in the directory of your custom database and taxonomy files!

For the tutorial, we enter “**SLV**” in the dialog and hit the **OK** button.

17. “Which Silva database do you want to use?” → **SSU**

Keep in mind that *Silva* is the only database within CoMA, which allows you to choose between SSU (small ribosomal subunit) and LSU (large ribosomal subunit) options. If multiple databases are selected in step 16, only *Silva* is affected. This step is skipped if *Silva* is not among the selected databases.

REMARK: This step is computationally intensive and may take up to several hours, depending on your hardware and the dataset you are analysing. The small example dataset of this tutorial, however, should be finished in a few minutes.

18. “Do you want to remove rare OTUs from your dataset?” → **Yes**

This step can be used to discard very rare OTUs from analysis. OTUs can be omitted due to a low number of total reads on the one hand or due to rare occurrence within the samples on the other hand.

19. Firstly, choose a minimum number of reads for an OTU to be retained. OTUs with fewer observations are excluded from the OTU-table (and therefore from further analysis). Be aware that the minimum number of reads is defined as sum of reads within all samples and not as reads per individual sample! For the tutorial, we choose **2** (which is a convenient value, often called *doubletons*) and hit the **OK** button. You can change the number of OTUs with either moving the button of the scale dialog with your mouse or your left/right arrow keys.

20. Secondly, choose a minimum number of samples in which an OTU must be present. OTUs which were found in fewer samples are excluded from the OTU-table (and therefore from further analysis). For the tutorial, we choose **0** (meaning that no OTUs will be removed based on this criterion) and hit the **OK** button.

21. “Do you want to generate rarefaction curves?” → **Yes**

Rarefaction curves are important to determine, how many reads are required to cover all or at least most of the information in a sample. The computation is based on a randomization procedure. You can access the plot of the rarefaction curves in the following directory:

.../Results/rarefaction_curves

In the same folder, you can also access the underlying data (including the higher (*hci*) and lower (*lci*) confidence interval) for secondary analysis.

22. You can select between different calculators based on which the rarefaction curves will be constructed: *otu* (OTU count; often also referred to as “Sobs”: species observed), *chao* (Chao1 richness), *shannon* (Shannon-Wiener index), *simpson* (Simpson index), and *coverage* (Good’s coverage of counts). For the tutorial, we chose “**otu**”, which is nowadays the most common calculator for rarefaction curves. Click the **OK** button.

23. You can now choose a file format for your plots. Available file formats include raster graphics as well as vector graphics. We select “**TIFF**” and click the **OK** button (or double-click on “**TIFF**”).

24. The next dialog asks for the pixel density of your graphics. We type in “**300**” (dpi) as pixel density for high quality figures and click **OK**.

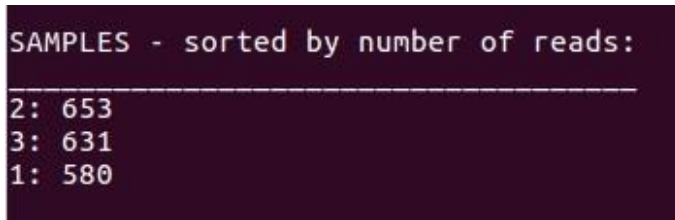
REMARK: This option is not available if a vector graphic format (e.g. EPS, PDF, SVG) was selected in the previous step!

25. “Do you want to make a subsampling?” → **Yes**

Within this step, the total number of reads of each sample is reduced to a desired count (= subsampling depth). This is necessary for most statistical approaches in order to compare the samples directly. Nevertheless, keep in mind that OTUs might be lost during this process, potentially affecting your results. Therefore, please check the rarefaction curves created in the previous steps and apply subsampling only in case of flatten curves! Samples that have fewer sequences than the requested subsampling depth are entirely excluded from further analysis. The removal of sequences is completely randomized and the outcome may differ from run to run. Randomization is done with a pseudo-random number generator, which itself is an implementation of the Mersenne twister PRNG.

26. “Please enter the number of reads for the subsampling:”

Look at the output in the Linux terminal to get a list of all samples together with the number of reads within each sample (sorted in decreasing order).



```
SAMPLES - sorted by number of reads:
-----
2: 653
3: 631
1: 580
```

We choose the smallest number of reads for the subsampling of our test results (always assuming, that no OTUs are getting lost; this assumption may be wrong in this reduced dataset for the tutorial!): **580** (you may get slightly different read counts; in this case, use your lowest value) and hit the **OK** button.

27. “Do you want to rename your samples?” → **Yes**

You can enter a new name for each of your samples in the Linux Terminal and accept with the enter key. Be aware that all upcoming analyses will be done using the new sample names! This is often useful since sample names provided by the sequencing company are often long and confusing. For demonstration purposes, we are giving new names to our samples even though they are already labelled in a simple way. This is done by typing in the new name of each sample printed in the Terminal window and accepting it with the **ENTER** key. We select **S1**, **S2** and **S3** as new sample names.

ATTENTION: If you are changing sample names, a previously created mapping file is no longer working! You need to either create a new mapping file (Step 28) or update the sample names in the existing one.

28. “Do you want to add metadata to your samples?” → **Yes**

Metadata are data describing your samples such as environmental conditions (e.g. season, temperature, pH, altitude) or process/sampling characteristics (e.g. stirring intensity, feeding rate, body site, drug load). These data can be used in the upcoming steps in order to group the samples based on a chosen mapping variable. We will do this step for demonstration purposes, however, keep in mind that these metadata are just created randomly and do not have any scientific meaning!

ATTENTION: All mapping variables must be categorical! Metric mapping variables are currently not supported. If you do have metric variables, make sure to classify them to suitable categories and provide them this way.

29. You can now enter the names of different mapping variables (separated with “,”), which you would like to include in order to characterize your sequencing data. We type in “**season,year**” and hit the **OK** button.

30. A new dialog window informs us that we now need to fill in the mapping file. After clicking **OK**, *LibreOffice Calc* (an open-source alternative to *Microsoft Excel*) will open and ask you for some settings for the import. Make sure that “**Separated by**” is selected as separator option and “**Tab**” as the separator token. Click **OK**.

ATTENTION: We recommend closing all other *LibreOffice Calc* windows before starting the input of your metadata, otherwise the checking step for proofing your input will fail and you

will not get a verification even if your data were entered correctly. However, this does not affect your further analyses at all and you do not have to repeat this step!

31. This opens the mapping file, with samples in rows and mapping variables in columns. For the tutorial, we enter “**spring**”, “**summer**” and “**spring**” in the column “season”, and “**y2019**”, “**y2019**” and “**y2020**” in the column “year”. Save the changes with “**File → Save As ..**” and tick “**Edit filter settings**” (lower left corner of the window). Leave all other settings (including the file name) as they are and click the **Save** button. Click now “**Use Text CSV Format**” and select in the next window “**{Tab}**” as *Field delimiter*.

ATTENTION: All cells of the mapping data matrix need to be filled in! Leaving cells without information may cause problems during the upcoming steps.

ATTENTION: It is not recommended to enter mapping variables starting with a number! Ignoring this may lead to problems in the upcoming steps. If you do have variables represented as numbers, simply put a character in front (e.g. “y2020” instead of “2020” or “pH_7” instead of “7”).

32. Close now the program with the **X button** in the upper right corner of the window. A short checking step will follow, proving if your inputs were given appropriately. In our case, all variables were entered correctly and we can proceed by clicking the **OK** button.
33. “Do you want to generate a summary report?” → **Yes**

This step creates text files containing the most abundant taxa from different taxonomic level (kingdom, phylum, class, order, family, genus, species) for each sample. The reports show absolute read counts as well as relative abundancies of the investigated taxa. Unassigned taxa are symbolized with a question mark (“?”). You can find the summary report files (starting with “summary_report_”) in the following directory:

.../Results

ATTENTION: No file will be created if the summary report would be empty anyway (because there are no taxa left fitting the selection criteria provided in the upcoming steps)!

34. “Do you want to use the information in the mapping file to group your samples?” → **Yes**

This allows you to get summary reports for groups based on a selected mapping variable rather than for individual samples. For the tutorial, we want to get summary reports for the years 2019 and 2020, and thus type in “**year**” as mapping variable. Click the **OK** button.

35. “Do you want a general summary?”

CoMA offers general summaries as well as summaries for a specific taxon (e.g. *Firmicutes* or *Enterobacterales*). For a specific summary report, you need to click “**No**” and enter the name of the taxon you want to focus on in an upcoming step. For the tutorial, however, we will compute a general summary with all detected taxa, and hence click “**Yes**”.

36. Choose now which summary reports do you want to create: *Archaea*, *Bacteria*, *Fungi*, *Eukaryota* or *Total*. Since we are analysing 16S samples, we are selecting “**Archaea**”, “**Bacteria**” and “**Total**”. Hit the **OK** button.
37. You can now decide how many taxa will be shown for each taxonomic level (as long as there are enough entries available). We type in “**999999**” because we do not want to limit the output and click the **OK** button.

38. “Do you want to create plots of the most important taxa?” → **Yes**

This step creates taxonomic plots for each taxonomic levels. CoMA offers all plots as bar charts as well as heatmaps. You can find the plots in the following directory:

.../Results/taxa_plots/

39. “Do you want to use the information in the mapping file to group your samples?” → **Yes**

This groups your samples in the graphics based on a selected mapping variable and no individual samples are shown anymore. We want to group our plots by season and enter “**season**” as mapping variable. Click the **OK** button. The plots will now depict data for spring as well as for summer.

40. “Do you want general plots?”

As previously described in step 35 for the summary reports, CoMA offers plots either generally or based on a specific taxon. For demonstration purposes, we want to compute now specific plots and click “**No**”.

41. Provide now the taxon based on which you want to create your plots. For the tutorial, we type in “**Firmicutes**” since it is the most diverse phylum in our example dataset. However, in other instances you may also provide a class name or a taxon corresponding to any other taxonomic level. Hit the **OK** button.

42. “Do you want to include unassigned taxa in the plots?” → **Yes**

REMARK: Including unassigned taxa is often reducing the clarity of taxonomic plots. However, excluding them always means losing information, possibly resulting in misleading results. This is particularly problematic in cases you are analysing habitats, which are not well investigated so far.

43. You can now choose a file format for your plots. Available file formats include raster graphics as well as vector graphics. We select “**TIFF**” and click the **OK** button (or double-click on “**TIFF**”).

44. The next dialog asks for a threshold as well as for the pixel density of your graphics (the second option is not available if a vector graphic format was selected in the previous step!). Taxa with a relative abundance below the threshold are excluded from depiction. We type in “**1**” (%) as threshold (which is a common value for such plots) and “**300**” (dpi) as pixel density for high-quality figures; click **OK**.

45. “Do you want to create Venn plots?” → **Yes**

Venn plots are used in CoMA to compare the taxonomy of different groups with each other. You can compare either two or three groups, and these groups are defined either by a mapping variable or by a specific grouping step. Venn plots are shown for each taxonomic level. You can find the plots in the following directory:

.../Results/Venn_plots

46. “Do you want to use the information in the mapping file to group your samples?” → **No**

This would group your samples based on a selected mapping variable. Grouping the samples manually, however, provides more flexibility. You can for instance leave specific samples out of the depiction, or group samples that are not related to each other, at least based on the provided mapping data.

ATTENTION: Only mapping variables leading to 2 or 3 different groups can be used for Venn plots since CoMA currently does not support comparison of more than 3 groups!

47. “How many groups do you want to compare?”

Type in “2” and click the **OK** button. This step is skipped if you are grouping your samples based on mapping data! CoMA automatically recognizes the number of different groups in this case.

48. Choose which file format do you want to use. Enter “tiff” and click **OK**.

REMARK: Choosing EPS as file format currently leads to an insufficient image quality since semi-transparent patches are not depicted properly.

49. Provide the pixel density for your figures. Be aware, this step is skipped if a vector graphic format was selected in the previous step! We enter “300” (dpi) for high-resolution and publication-ready graphics and hit the **OK** button.

50. Now, you can define the groups you want to compare. First, you need to assign samples to a group (**IMPORTANT:** always use the sample number, not the sample name!) and thereafter, you need to provide a name for the newly formed group.

51. Now, a series of dialogs ask you to enter the sample numbers (again, not the sample names!) for each group, separated with “,”. For demonstration, we are combining samples 1 and 3 to one group and take sample 2 as a second group (consisting of only one sample). Enter “1,3” and click the **OK** button.

52. After assigning samples to a group, a second dialog for providing the group name appears. We type in “A” and click the **OK** button.

53. These steps are repeated two or three times, depending on your decision in step 47. Samples that are already assigned to a group are not shown in the dialog window anymore. Samples that were not assigned to any group are excluded from the depiction. For the definition of our second (and last) group, we type in “2” (which is the only remaining sample anyway) and hit the **OK** button. Then we enter “B” and click again the **OK** button.

54. “Do you want to label all areas with the exact number of taxa?” → **Yes**

REMARK: Usually, we recommend labelling of the Venn plots. However, in some situations, labels will overlap, particularly when the areas are overlapping to a large degree. In this case, the user may want to exclude them and label the plot manually with a graphic tool of choice.

55. “Do you want to calculate/plot the alpha diversity of your samples?” → **Yes**

This step calculates alpha diversity (or within-sample diversity) using an OTU table created during the previous steps. The plot as well as a supporting text file containing the measured diversity can be found in the following directory:

.../Results/alpha_diversity

ATTENTION: Always keep in mind that the removal of rare OTUs and/or subsampling may significantly affect the results of your alpha diversity analysis!

ATTENTION: In case of an inhomogeneous distribution of reads among your samples, alpha diversity analysis may lead to erroneous results and thus wrong conclusions! You can check the rarefaction plot in order to see the distribution of reads in your dataset.

56. Different metrics are available for the calculation: *observed OTUs*, *Shannon-Wiener index*, *Simpson’s index*, *Pielou’s evenness index*, *Good’s coverage of counts*, *Chao1 richness estimator*

and *Faith's phylogenetic diversity*. In contrast to the others, the latter is a phylogeny-based method, getting the required phylogenetic information from CoMA's *Tree.tre* file. Every metric has different strengths and limitations - technical discussion of each metric is readily available [online](#) and in ecology textbooks, but it is beyond the scope of this manual. For the tutorial, we select "**Shannon**" and click **OK** (or double-click on "**Shannon**").

57. "Do you want to use the information in the mapping file to group your samples?" → **Yes**

This groups your samples in the graphic based on a selected mapping variable and no individual samples are shown anymore. If grouping is applied, the (arithmetic) mean diversity (\pm standard deviation) is shown for each of the groups. We want to group our plots by season and enter "**season**" as mapping variable. Click the **OK** button. Using these settings, the plots will include the Shannon-Wiener index for the groups: *spring* and *summer*.

58. You can now choose a file format for your plot. Available file formats include raster graphics as well as vector graphics. We type in "**tiff**" and click the **OK** button.

59. The next dialog asks for the pixel density of your graphic (this option is not available if a vector graphic format was selected in the previous step!). We type in "**300**" (dpi) for high-quality figures and click **OK**.

60. Now you can select the colour of the bars in the plot. You can provide the colour either as hex code (e.g. #000000 for black), by adjusting RGB values, or by adjusting *hue*, *saturation* and *value*. Alternatively, you can also use the pipette tool, which allows you to select any colour appearing on your screen. We type in "**#3A86D1**" for a nice blue colour and click the **Select** button. You can find the alpha diversity plot in the following directory:

[.../Results/alpha_diversity](#)

In the same directory, you also find a file containing the raw diversity data as well as a file containing statistics (Kruskal-Wallis H-test, Conover post-hoc test with Benjamini-Hochberg correction).

ATTENTION: Statistics can only be calculated if metadata were used for the grouping of the samples! If no metadata were used, the statistics file is missing.

61. "Do you want to calculate/plot the beta diversity of your samples" → **Yes**

62. To analyse beta diversity, you can choose between two different approaches: ordination using Principal Coordinates Analysis (PCoA) and Hierarchical Cluster Analysis (HCA). For now, we select "**PCoA**" and click **OK** (or double-click on "**PCoA**").

63. Now you can select a metric for the calculation of the distance between your samples and hence the creation of a distance matrix. CoMA offers non-phylogenetic (*Minkowski*, *Euclidean*, *Manhattan*, *Cosine*, *Jaccard*, *Dice*, *Canberra*, *Chebyshev*, *Braycurtis*) as well as phylogenetic (*Weighted UniFrac*, *Unweighted UniFrac*) metrics. For the phylogeny-based methods, CoMA's *Tree.tre* file provides the required phylogenetic information. Every metric has different strengths and limitations - technical discussion of each metric is readily available [online/online](#) and in ecology textbooks. We select "**Minkowski**" and click **OK** (or double-click on "**Minkowski**").

64. For *Minkowski* distance, you need to provide also a p-norm (be aware that this dialog does not show up if another metric was selected!). The p-norm must be a natural number (= positive integer), we type in "**3**" and hit the **OK** button.

REMARK: *Minkowski* analysis with $p = 1$ corresponds to the *Manhattan* distance and with $p = 2$ to the *Euclidean* distance. You can choose either *Minkowski* distance with $p = 1 / 2$ or directly *Manhattan* / *Euclidean* distance; both settings are leading to the identical results!

65. “Do you want to use metadata from the mapping file in order to colour your samples?” → **Yes**

This colours your samples in the PCoA plot based on a selected mapping variable. If no metadata are used, all samples are depicted in the same colour, which can be selected as previously described in step 60 for the alpha diversity plot. We want to group our plots by year and enter “**year**” as mapping variable. Click the **OK** button. Using these settings, samples from 2019 and 2020 will be shown in two different colours.

66. “Do you want to see a 3D illustration of your ordination?” → **Yes**

REMARK: Three-dimensional graphics are often helpful for discovering structures in your dataset. For plots, however, they are unsuitable and not recommendable. Therefore, CoMA offers live graphics in 3D and plots for publication in 2D.

67. “Do you want to create a two-dimensional plot of your ordination?” → **Yes**

68. You can now choose a file format for your plot. Available file formats include raster graphics as well as vector graphics. We type in “**tiff**” and click the **OK** button.

69. The next dialog asks for the pixel density of your graphic (this option is not available if a vector graphic format was selected in the previous step!). We type in “**300**” (dpi) for high-quality figures and click **OK**.

70. You can now have a look at the three-dimensional PCoA plot. You can rotate the graphic by holding the left mouse button and move the mouse in any direction. By holding the right mouse button and moving the mouse forward/backward, you can zoom in/out. If you want to save the graphic, click on the floppy disk symbol. If you want to continue, close the window by clicking on the **X** button.

71. The PCoA plot can be found in the following directory:

.../Results/beta_diversity/ordination

In the same directory, there are also files containing the distance matrix as well as the eigenvalues, which determine how much of the variation is explained by each of the computed PC axis. CoMA also provides statistics quantifying the strength of the grouping/clustering seen in the PCoA. This is done on the one hand with ANOSIM and on the other hand with PERMANOVA. Both approaches are using multiple permutations in order to calculate their R/F statistics. Irrespective of the method, 999 permutations are pre-set in CoMA; however, experienced users may change the number of permutations by manipulating the underlying code.

ATTENTION: Statistics can only be calculated if metadata were used for the grouping of the samples! If no metadata were used, the statistics file is missing.

72. Click **OK**. In fact, this was the last step of the CoMA pipeline. However, we also want to do a cluster analysis. To do so, we need to start a new CoMA run. Let’s go!

73. Start CoMA from Linux terminal:

\$ coma

74. “Do you want to start a new project?” → **No**

75. Search and select the project folder that you created in course of this tutorial (do **not** double-click on it!). Click **OK**.
76. Skip now each step until “Do you want to calculate/plot the beta diversity of your samples” by clicking **No** (= 15 times). Now, click **Yes**.
77. Select “**HCA**” and click **OK** (or double click on “**HCA**”).
78. Choose between the following [linkage methods](#) for the cluster analysis: *single*, *complete*, *average*, *weighted*, *centroid*, *median* and *ward*. These methods are used to compute the distance $d(s,t)$ between two clusters s and t . The algorithm begins with a forest of clusters that have yet to be used in the hierarchy being formed. When two clusters s and t from this forest are combined into a single cluster u , s and t are removed from the forest, and u is added to the forest (bottom-up approach). When only one cluster remains in the forest, the algorithm stops, and this cluster becomes the root. Be aware that some methods (*centroid*, *median*, *ward*) are limited to Euclidean distance metric (in step 79). We choose “**average**” as linkage method for our tutorial and hit the **OK** button.
79. Choose between the following [metrics](#) for the cluster analysis: *Euclidean*, *cosine*, *cityblock*, *correlation*, *jaccard*, *braycurtis* and *dice*. The metric determines how to measure the distance between two points. For detailed information on the different metrics, follow the link above or other relevant literature. We choose the “**braycurtis**” metric for the tutorial and hit the **OK** button. **ATTENTION:** This dialog will not show up if either *centroid*, *median* or *ward* was selected as linkage method since they all require the Euclidean distance metric!
80. “Do you want to plot the distance of each node in the dendrogram?” → **Yes**
Keep in mind that displaying the distances of each note can lead to overlapping issues in datasets including many samples!
81. You can now choose a file format for your plot. Available file formats include raster graphics as well as vector graphics. We select “**TIFF**” and click the **OK** button (or double-click on “**TIFF**”).
82. The next dialog asks for the pixel density of your graphic (this option is not available if a vector graphic format was selected in the previous step!). We type in “**300**” (dpi) for a high-quality figure and click the **OK** button. The dendrogram can be found in the following directory:

.../Results/beta_diversity/cluster_analysis

Congratulations! You finished the tutorial successfully. Thank you for using CoMA, the pipeline for intuitive analysis of your NGS data! You can access a log file of your run (a copy of the Linux command line dialog that was shown during the analysis) to recap your analysis and check for all the important settings given by the user. Moreover, there is a detailed log file containing the complete standard output of this run. This log file also includes warnings and error messages that were not shown during the CoMA run, and may therefore be helpful for advanced users or the CoMA support for solving problems. You can find both log files in the main directory of your project.

Thank you again for using the CoMA pipeline!

v. Citation

If you use CoMA for any published research, please include the following citation:

Sebastian Hupfauf, Mohammad Etemadi, Marina Fernández-Delgado Juárez, María Gómez-Brandón, Heribert Insam, Sabine Marie Podmirseg. 2020. CoMA – an Intuitive and User-friendly Pipeline for Amplicon-Sequencing Data Analysis. PLOS One. *submitted*.

vi. The CoMA output

This chapter describes the output of the CoMA pipeline in detail and shall help the user to navigate through his/her results. It also treats files, which were computed during the CoMA run but which are not used by the pipeline directly. However, these files are in standardized formats and may be used by advanced users for secondary analysis (e.g. using R). All results can be found in the following directory:

.../Results

Directories:

alpha_diversity: Alpha diversity plot(s) and text file(s) containing the underlying data of the diversity analysis. Moreover, you can find the results of the statistical tests (Kruskal-Wallis H-test, Conover post-hoc test with Benjamini-Hochberg correction) if metadata were used for sample grouping.

beta_diversity: Here you can find two sub-directories including all files from the ordination analysis as well as from the cluster analysis. In the “*ordination*” directory, you find the PCoA plot, the distance matrix and a file containing the eigenvalues for all PC axis. Moreover, you can find the results of the statistical tests using ANOSIM and PERMANOVA if metadata were used for sample grouping. In the “*cluster_analysis*” directory, you find the dendrogram.

ExtraFiles: Here you can find files summarizing the chimera removal step.

higherLvl: This directory includes Species, Genus, Family, Class, Order and Phylum abundance matrices.

LotuSLogS: Several log files are stored here. These files are usually not needed; however, they may be helpful in case of unexpected results or other problems. For detailed explanation, please refer to the online documentation of *LotuS* (the software package that is involved in OTU clustering and taxonomic assignment): <http://psbweb05.psb.ugent.be/lotus/documentation.html>.

primary: Here you can find an option file for the *sdm* tool, which is responsible for the demultiplexing and quality filtering of sequences. In addition, you can find a copy of the map file, which was constructed in course of the analysis.

rarefaction_curves: This directory includes the rarefaction plot(s), a *Mothur* log file of the rarefaction analysis, and the underlying data files. The number of files as well as their labelling depends on the chosen calculator. “*otu.groups.rarefaction*” summarizes for instance all data when *otu* was the calculator and “*otu.groups.r_shannon*” would include the results of a rarefaction analysis based on Shannon-Wiener diversity.

taxa_plots: Here you can find all your taxonomic plots (bar charts, heatmaps).

Venn_plots: Here you can find the Venn plots.

Files:

cnadjusted_hierachy_cnt.tax: This file includes taxonomic information that is needed for the RDP aligner tool in order to assign the reads. **IMPORTANT:** This file is missing when RDP was not the selected aligner tool (e.g. when doing the tutorial)!

hierachy_cnt.tax: This file includes taxonomic information that is needed for the RDP aligner tool in order to assign the reads. **IMPORTANT:** This file is missing when RDP was not the selected aligner tool (e.g. when doing the tutorial)!

hiera_BLAST.txt: This OTU file shows the taxonomic assignments based on *BLAST*. **IMPORTANT:** Be aware that this file is missing (and replaced by another, e.g. *hiera_RDP.txt*) when *BLAST* was not the selected aligner tool!

mapping.txt: This file includes all the metadata that were provided by the user. Keep in mind that this file is missing if no metadata were entered.

OTU.biom: This is the current OTU-table in BIOM format. BIOM (Biological Observation Matrix) was designed in order to represent a widely accepted and supported file format for contingency tables of biological samples. BIOM files can be easily converted to tab-delimited OTU-tables and vice versa using the *biom-convert* tool. For more information, please refer to the BIOM webpage: <http://biom-format.org/>. Please see also the information provided for *otu_table.txt*.

OTU.txt: This file represents an OTU abundance matrix, showing which OTU appears in which sample/group and at which number.

otuMultiAlign.fna: This file shows the results of the multiple sequence alignment of OTUs in FASTA format. Divergent positions are indicated with hyphens (“-”).

OTU_original.biom: This is the original OTU-table in BIOM format for your analysis, which was constructed immediately after OTU clustering and taxonomic assignment. It does not include any post-processing steps such as subsampling.

otus.fa: This file shows the extended OTU seed sequences in FASTA format. The sequences given here are identical to those in *otuMultiAlign.fna*, but without the hyphens.

otu_table.txt: This is the most current OTU-table of your analysis. Depending on your settings, this table may include rarefied, subsampled, renamed or grouped samples/replicates/groups. **IMPORTANT:** it is indispensable to know which steps have been performed during the CoMA analysis if you want to do further analysis based on this file!

otu_table_original.txt: This is the original OTU-table for your analysis, which was constructed immediately after OTU clustering and taxonomic assignment. It does not include any post-processing steps such as subsampling.

otu_table_without_subsampling.txt: This is the OTU-table after the removal of rare OTUs, but without any other post-processing steps such as subsampling.

OTU_without_subsampling.biom: This is the BIOM-converted OTU-table after the removal of rare OTUs, but without any other post-processing steps such as subsampling.

rem_seq.txt: This file summarized the amount of sequences that were dropped in course of the analysis. With this information, the user can estimate the overall loss of reads as well as determine which step removed the most sequences (e.g. quality filtering, chimera removal, phiX contaminants).

summary_report_*TAXON*.txt: These files are summary reports of your samples, which only include entries associated with a specific taxon (e.g. *summary_report_Archaea.txt* will only include archaeal entries, whereas *summary_report_Total.txt* will include all of them).

Tree.tre: This file represents a taxonomic tree in NEWICK format. It is used for phylogeny-based alpha/beta diversity analyses (Faith's PD, weighted or unweighted Unifrac distance). Moreover, advanced users may use this for even more-sophisticated analyses based on taxonomic information and the tree structure (e.g. with [FastTree](#), [FigTree](#), [GraPhlAn](#)).

vii. What is new in CoMA 2.0?

- Two additional options for installation are provided now: a Singularity image and a direct Linux installer.
- The Ubuntu operating system was updated to version 20.04 LTS (in CoMA 1.0: Ubuntu 16.04 LTS).
- All CoMA source files are now available at GitHub: <https://github.com/SebH87/coma>.
- Support of multithreading. You can assign now multiple CPU cores to CoMA leading to a considerably better performance resulting in shorter computation times.
- Support of USEARCH v11. For more information, please visit the online documentation: <https://drive5.com/usearch/manual/whatsnewv11.html>.
- All taxonomic databases were updated. This includes also the newest release of the SILVA database (version 138).
- CoMA now offers two log files for each run: a compact log file including the user settings and inputs, and a detailed log file with all the information including potential warnings and error messages.
- Sequences can now be filtered based on the number of ambiguous bases (Ns) in course of the trimming/quality filtering step.
- CoMA now supports the analysis of single-end data or of sequence files that were already merged.
- Sample registration is now a separate step und no longer connected to sequence merging. With that, the merging step can be repeated multiple times without the necessarily to provide the common part of the sample names each time.
- CoMA now checks if input files are missing when analysing paired-end data. Moreover, CoMA now detects wrongly assigned pattern strings in course of the sample registration step.
- Several steps checking for missing dependencies have been implemented in order to help the user locating a problem.
- Rarefaction curves can now be computed based on five different calculators: OTU count, Chao1 diversity, Shannon-Wiener index, Simpson index, and Good's coverage for OTU (compared to only OTU count in CoMA 1.0).
- Samples can now be renamed in order to avoid confusing names, which are often assigned by NGS machines or sequencing companies.
- CoMA now supports metadata, which can be used for grouping samples based on specific parameters.
- CoMA now supports general summary reports (for Archaea, Bacteria, Fungi, Eukaryota, and all data) as well as specific summaries based on a given taxon. The summary reports are provided as tab-delimited text file, where the information can be easily extracted and used for further analysis. In addition, the user can now select the numbers of entries for each taxonomic level (in CoMA 1.0, it was pre-set to 5).

- CoMA now supports general (for Archaea, Bacteria, Fungi, Eukaryota, and all data) and specific taxonomic plots (bar charts, heatmaps) based on a given taxon. In addition, unassigned taxa can now be included or excluded from the depiction.
- CoMA now supports Venn plots for the taxonomic comparison of two or three groups.
- The calculation step for alpha diversity is now improved and the user can choose between various metrics. CoMA supports now also phylogeny-based alpha diversity calculation (Faith's PD).
- CoMA now offers statistical tests for alpha diversity (Kruskal-Wallis H-test, Conover post-hoc test with Benjamini-Hochberg correction), as long as samples were grouped based on metadata.
- Aside Hierarchical Cluster Analysis (HCA), CoMA now supports ordination (Principal Coordinates Analysis, PCoA) as second option for beta diversity analysis. The user can choose between various metrics, including phylogeny-based weighted/unweighted UniFrac distance.
- CoMA now offers also statistical tests for beta diversity (ANOSIM, PERMANOVA) in order to quantify the strength of clustering determined with PCoA.
- You can now select a file format and, in case of raster graphic formats, the pixel density (DPI) for all CoMA graphics (rarefaction curves, bar charts, heatmaps, Venn plots, alpha diversity plots, PCoA plots, dendrograms). CoMA supports 10 different file formats, including the most popular raster- (e.g. "JPEG", "PNG", "TIFF") and vector- (e.g. "EPS", "PDF", "SVG") graphic formats.
- We established an overall CoMA colour code. All CoMA graphics are now using the same colour palette (ranging from blue to green to yellow to beige).
- All CoMA 1.0 scripts were revised and optimised where needed.
- CoMA now uses Python 3 (in comparison to Python 2 in CoMA 1.0).
- Any directory can now be used for CoMA projects, allowing much more flexibility and facilitates the usage of CoMA on HPC systems.
- CoMA now shows a warning message if a new project is started using an already existing project name. The user can decide if he wants to keep the old data or overwrite it with the new project.
- The CoMA manual was completely overworked and improved. Moreover, a chapter describing the CoMA output in detail was included.

viii. Software list

CoMA is a pipeline for NGS data analysis, using various different applications, tools and scripts originating from internal and external sources (open-source third party software). Within this section, all external tools are listed. Please follow the web links if you want to get more information on a specific tool. Keep in mind that some of these tools may also be using third party software, for more information consult the prevailing manuals.

- [Python](#)
- [Qiime](#)
- [Mothur](#)
- [Pandaseq](#)
- [Prinseq-lite](#)
- [LotuS](#)

- [USEARCH](#) (has to be installed by the user: page 4, step 6)
- [BiOM-format-tools](#)