Comparative Microbiome Analysis – version 3.0

Index

i. What is CoMA?

Nowadays, modern biological or medical research is scarcely conceivable without the deep insights that are gained from next-generation sequencing (NGS). These technologies have been constantly improved and yield up to 20 billion reads per run today. This allows for studying highly complex environments such as gut, soil or biogas reactor sludge at a high resolution, but also requires an adequate analysis of the huge amount of generated data. A plethora of software packages for NGS data analysis are available today; however, most of them demand extensive knowledge in computer sciences or even need a bioinformatic specialist to be executed.

CoMA (Comparative Microbiome Analysis) is a free pipeline for intuitive and user-friendly analysis of amplicon-sequencing data, available for all common computer platforms. The software package uses various open-source third party tools and combines them with own scripts into a linear analysis workflow in the form of a Bash script, starting with the raw input files (in FASTQ format) and resulting in aesthetically pleasing and publication-ready graphics. In addition, output files in standardized formats, such as a tab-delimited abundance table, an abundance table in BIOM format and a tree file in NEWICK format, are provided. These allow for subsequent secondary analysis using R for example.

The operation of this tool is remarkably intuitive and makes it accessible even for entry-level users. A graphical user interface facilitates the handling, representing a major advantage compared with command-line based applications. Nevertheless, multiple adjustment parameters and the high degree of automation make CoMA also suitable for advanced users who are looking for an efficient and streamlined workflow. The tool is capable of handling data from today's most important NGS platforms, including Illumina MiSeq, Illumina HiSeq, Illumina NextSeq, and Illumina NovaSeq, but also from the former 454 pyrosequencing technology, which was, in fact, terminated in 2016 but data are still around and analysis tools are still needed.

CoMA can be run on every common computer operating system: Linux, Windows, and macOS. With version 3.0, four different options for installation are available: a virtual appliance (which can be imported with tools like VMware Workstation, Oracle VM Virtualbox or Parallels Desktop for macOS), a Singularity image, a Docker image, and a direct Linux installer (Bash script). Each option provides specific advantages and the user can select the most suitable one in order to meet his needs. CoMA can also be used on high performance computer systems (HPC cluster). In this case, we suggest using the Singularity option. The following graphical summary shall help finding the ideal usage strategy for every user.

	Linux - Debian	*Linux - Other*	*macOS*	*Windows*
VirtualBox	✓	✓	✓	✓
Singularity	✓	✓	✓	✓
Docker	✓	✓	✓	✓
Bash script	✓			

✓ Possible
✓ Recommended

CoMA will be updated regularly to ensure an excellent performance also in the future, but also to implement additional features (e.g. a web-based platform for online data analysis). These updates will always include the newest versions of the taxonomic databases at the release date to guarantee the best possible and most current results.

ii. 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 operating system. It demands relatively large system requirements but the operation is intuitive and easy to learn for anyone with basic computer knowledge.

ii.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. We highly recommend this!

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. Download "**CoMA_3.0.ova.gz**" from the homepage (https://www.uibk.ac.at/microbiology/services/coma) and unpack the zipped file (e.g. using *7-zip*).

4. Run the unzipped "**CoMA_3.0.ova**" file 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 up to 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_3.0**"). 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 unsure.

OPTIONAL: The system memory for the installed virtual environment ("CoMA_3.0") is initially set to 2048 MB. This is sufficient for most analysis within the CoMA analysis pipeline. Nevertheless, some procedures (e.g. clustering huge datasets or taxonomic assignment) need more memory and you need to increase it manually. This is also recommended if you want to increase generally the performance of CoMA. **ATTENTION**: 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 be too little memory left for the host operating system. You can also change the number of CPU cores (default setting: 2 CPU 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_3.0**") and hit the "**Settings**" button (orange cogwheel)

   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.

5. Start now the virtual environment by hitting the "**Start**" button.

   If you receive an error message similar to "*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.

6. Login with the password: **Passw0rd!**

7. OPTIONAL: All CoMA tools are pre-installed except for "USEARCH". You can use CoMA also without "USEARCH"; however, we strongly recommend installing it since various steps of the pipeline are making use of it and you would end up with limited options for sequence clustering (not available: "UPARSE", "UNOISE"), chimera removal (not available: "USEARCH") and taxonomic assignment (not available: "USEARCH")! Due to license restrictions, each user must install "USEARCH" individually. Open the Linux "**Terminal**" and type in the code written in bold and italics (AFTER the $ sign!):

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

      $ cd /usr/local/Pipeline/lotus2/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 *

ii.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 are at the bottom of your screen, depending whether you are using the window or full screen mode) 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 may freeze your mouse or your keyboard, or make your mouse cursor invisible. However, this problem shall be fixed after restarting the virtual system.

ii.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).

iii. 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 bash script for a native installation (Chapter v) 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!

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). Here, we focus on the installation and usage on a Linux system:

iii.i. Installation

1. First, install Singularity using your distribution's package manager (e.g. **apt-get**, **yum**) or from source. For more information, consult the products web page: https://sylabs.io/guides/3.0/user-guide/installation.html.

2. Download the CoMA Singularity image (**CoMA_3.0.tar.gz**) from the CoMA webpage (https://www.uibk.ac.at/microbiology/services/coma). This can be done either manually or using a command line downloader like **wget**.

   $ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_3.0.tar.gz

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_3.0.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_3.0.ovl CoMA_3.0.sif

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

5. To quit the Singularity session:

   $ exit

6. OPTIONAL: All CoMA tools are pre-installed except for "USEARCH". You can use CoMA indeed without "USEARCH"; however, we strongly recommend installing it since various steps of the pipeline are making use of it and you would end up with limited options for sequence clustering ("UPARSE" and "UNOISE" are not available in this case), chimera removal (no "USEARCH") and taxonomic assignment (no "USEARCH")! Due to license restrictions, each user must install "USEARCH" individually:

   a. Start a new Singularity session:

      $ singularity shell -B ~/.Xauthority --overlay CoMA_3.0.ovl CoMA_3.0.sif

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

      $ cd /usr/local/Pipeline/lotus2/usearch

   c. Download USEARCH from the tool's webpage:

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

d. Unzip the USEARCH file:

$ gunzip *

e. Rename the file:

$ mv * usearch

f. Give execution rights to the file:

$ chmod 111 *

7. OPTIONAL: The data files for the CoMA tutorial need to be downloaded from the CoMA server. This step is optional and can be skipped if you do not want to go through the CoMA tutorial.

a. Start a CoMA Singularity session:

$ singularity shell -B ~/.Xauthority --overlay CoMA_3.0.ovl CoMA_3.0.sif

b. Download the "CoMA_Tutorial" folder containing the example dataset for the tutorial of the pipeline (section vi):

$ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_Tutorial.tar.gz

c. Unpack the zipped archive to the current working directory:

$ tar -xzf CoMA_Tutorial.tar.gz

iii.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 results 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 (***$ coma***). 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 able to see the graphical user interface of CoMA, resulting in a termination of the pipeline.

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

iv. Using CoMA with Docker

Docker is recommended for all users irrespective of the operating system. It supports all common operating systems including Windows, macOS and Linux. Even though the usage of Docker is not complicated, we recommend this option for users who are already familiar with command line-based applications. Moreover, we recommend using Docker for all macOS users since it is the most convenient option for their system.

ATTENTION: You need to have administration rights on your system! Otherwise, it is not possible to either install Docker or start a Docker image!

The installation is exemplary described for macOS and Linux. For detailed information on the installation on Windows, please refer to the respective documentations on the Docker webpage (https://docs.docker.com/get-docker/).

iv.i. Installation on macOS

1. If you are using a Mac with Apple chip, you firstly need to install Rosetta 2 using the macOS terminal:

   $ softwareupdate --install-rosetta

   REMARK: Mac users with an Intel chip may skip this step.

2. Download and install the newest version of Docker Desktop, choosing the correct file for your system (either Apple or Intel chipset) from: https://docs.docker.com/desktop/mac/install/

3. Start Docker Desktop.

   ATTENTION: The Docker Desktop application needs to run permanently when working in a Docker container!

4. Pull the newest CoMA3 Docker image (**sebh87/coma3**) from the Docker Hub (https://hub.docker.com/). This can be done with the following command:

   $ docker pull sebh87/coma3:latest

5. Download and install XQuartz from: https://www.xquartz.org/. Reboot the system after the installation.

   REMARK: XQuartz is an X11 display server for macOS, allowing displaying GUI (X11) applications for example from a Linux Docker container.

6. Start XQuartz and navigate to **Preferences**, sub section **Security**. Check *"Allow connections from network client"* and restart the application.

   ATTENTION: The XQuartz application needs to run permanently when working with CoMA in a Docker container!

7. Store the systems IP address in a variable called **IP**:

   $ export IP=$(ifconfig en0 | grep inet | awk '$1=="inet" {print $2}')

8. Allow X11 connections from the Docker container to the host system:

   $ xhost +$IP

ATTENTION: Steps 7 and 8 need to be repeated each time the terminal (or system) was restarted! Otherwise, the X11 connection between the Docker container and the host will fail.

9. To start the CoMA3 Docker container, execute the following command:

   $ docker run -it -e DISPLAY=$IP:0 -e XAUTHORITY=/.Xauthority --net=host --ipc=host -v /tmp/.X11-unix:/tmp/.X11-unix -v ~/.Xauthority:/.Xauthority -v $(pwd):/home sebh87/coma3 bash

10. To quit the Docker container, use the following command:

    $ exit

11. OPTIONAL: All CoMA tools are pre-installed except for "USEARCH". You can use CoMA indeed without "USEARCH"; however, we strongly recommend installing it since various steps of the pipeline are making use of it and you would end up with limited options for sequence clustering ("UPARSE" and "UNOISE" are not available in this case), chimera removal (no "USEARCH") and taxonomic assignment (no "USEARCH")! Due to license restrictions, each user must install "USEARCH" individually:

    a. Start a new CoMA3 Docker container:

       $ docker run -it --name c3 sebh87/coma3 bash

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

       $ cd /usr/local/Pipeline/lotus2/usearch

    c. Download USEARCH v.11 from the tool's webpage:

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

    d. Unzip the USEARCH file:

       $ gunzip *

    e. Rename the file:

       $ mv * usearch

    f. Give execution rights to the file:

       $ chmod 111 *

    g. Quit the Docker container:

       $ exit

    h. Save the changes to the CoMA3 Docker image:

       $ docker commit c3 sebh87/coma3

12. OPTIONAL: The data files for the CoMA tutorial need to be downloaded from the CoMA server. This step is optional and can be skipped if you do not want to go through the CoMA tutorial.

    a. Start a new CoMA3 Docker container:

       $ docker run -it -e DISPLAY=$IP:0 -e XAUTHORITY=/.Xauthority --net=host --ipc=host -v /tmp/.X11-unix:/tmp/.X11-unix -v ~/.Xauthority:/.Xauthority -v $(pwd):/home sebh87/coma3 bash

b. Download the "CoMA_Tutorial" folder containing the example dataset for the tutorial of the pipeline (section vi):

$ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_Tutorial.tar.gz

c. Unpack the zipped archive to the current working directory:

$ tar -xzf CoMA_Tutorial.tar.gz

iv.ii. Installation on Linux

1. First, install Docker using your distribution's package manager (e.g. apt-get, yum) or from source. For Debian/Ubuntu for example run:

   $ sudo apt-get install -y docker.io

2. Pull the newest CoMA3 Docker image (**sebh87/coma3**) from the Docker Hub (https://hub.docker.com/). This can be done with the following command:

   $ sudo docker pull sebh87/coma3:latest

3. Allow X11 connections from the Docker container to the host system:

   $ xhost local:root

   ATTENTION: This step needs to be repeated each time the terminal (or system) was restarted! Otherwise, the X11 connection between the Docker container and the host system may fail.

4. To start the CoMA3 Docker container, execute the following command:

   $ sudo docker run -it -e DISPLAY=$DISPLAY -e XAUTHORITY=/.Xauthority -v /temp/.X11-unix:/temp/.X11-unix -v ~/.Xauthority:/.Xauthority -v $(pwd):/home --net=host --ipc=host sebh87/coma3 bash

5. To quit the Docker container, use the following command:

   $ exit

6. OPTIONAL: All CoMA tools are pre-installed except for "USEARCH". You can use CoMA indeed without "USEARCH"; however, we strongly recommend installing it since various steps of the pipeline are making use of it and you would end up with limited options for sequence clustering ("UPARSE" and "UNOISE" are not available in this case), chimera removal (no "USEARCH") and taxonomic assignment (no "USEARCH")! Due to license restrictions, each user must install "USEARCH" individually:

   a. Start a new CoMA3 Docker container:

      $ sudo docker run -it --name c3 sebh87/coma3 bash

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

      $ cd /usr/local/Pipeline/lotus2/usearch

   c. Download USEARCH v.11 from the tool's webpage:

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

   d. Unzip the USEARCH file:

      $ gunzip *

e. Rename the file:

$ mv * usearch

f. Give execution rights to the file:

$ chmod 111 *

g. Quit the Docker container:

$ exit

h. Save the changes to the CoMA3 Docker image:

$ sudo docker commit c3 sebh87/coma3

7. OPTIONAL: The data files for the CoMA tutorial need to be downloaded from the CoMA server. This step is optional and can be skipped if you do not want to go through the CoMA tutorial.

   a. Start a new CoMA3 Docker container:

   $ sudo docker run -it -e DISPLAY=$DISPLAY -e XAUTHORITY=/.Xauthority -v /temp/.X11-unix:/temp/.X11-unix -v ~/.Xauthority:/.Xauthority -v $(pwd):/home --net=host --ipc=host sebh87/coma3 bash

   b. Download the "CoMA_Tutorial" folder containing the example dataset for the tutorial of the pipeline (section vi):

   $ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_Tutorial.tar.gz

   c. Unpack the zipped archive to the current working directory:

   $ tar -xzf CoMA_Tutorial.tar.gz

iv.iii. Notes

The CoMA Docker 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.

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 able to see the graphical user interface of CoMA, resulting in a termination of the pipeline.

There might be a problem with the size of the Docker image when trying to pull it from the Docker Hub using a Linux system. In this case, try the following steps and re-run the Docker pull command.

1. ***$ echo "{"storage-driver": "overlay"}" | sudo tee /etc/docker/daemon.json***
2. ***$ sudo service docker stop***
3. ***$ sudo service docker start***

Instead of pulling the Docker image from the Docker Hub (e.g. in case of a server problem), it can also be created out of a Docker file when working on Linux. For that, download "**Dockerfile**" from the CoMA web page (https://www.uibk.ac.at/microbiology/services/coma) and build a new CoMA3 Docker image:

1. ***$ cd***
2. ***$ mkdir coma3***
3. ***$ cd coma3***
4. ***$ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_3.0.docker***
5. ***$ sudo docker build -t sebh87/coma3 .***

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

v. Install CoMA natively with Bash

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

ATTENTION: You need to have administration rights on your system! Otherwise, it is not possible to install the tools and packages needed for CoMA!

v.i. Installation

1. Download the Bash script for installation (**CoMA_3.0_installer.sh**) from the CoMA webpage (https://www.uibk.ac.at/microbiology/services/coma). This can be done either manually or using a download manager like **wget**.

   $ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_3.0_installer.sh

2. Execute the Bash script. You need administration rights to continue!

   $ sudo bash CoMA_3.0_installer.sh

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

3. OPTIONAL: All CoMA tools are pre-installed except for "USEARCH". You can use CoMA indeed without "USEARCH"; however, we strongly recommend installing it since various steps of the pipeline are making use of it and you would end up with limited options for sequence clustering ("UPARSE" and "UNOISE" are not available in this case), chimera removal (no "USEARCH" option) and taxonomic assignment (no "USEARCH" option)! Due to license restrictions, each user must install "USEARCH" individually. You need administration rights for the upcoming steps!

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

      $ cd /usr/local/Pipeline/lotus2/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 *

4. OPTIONAL: The data files for the CoMA tutorial need to be downloaded from the CoMA server. This step is optional and can be skipped if you do not want to go through the CoMA tutorial.

a. Download the “CoMA_Tutorial” folder containing the example dataset for the tutorial of the pipeline (section vi):

   $ wget https://www2.uibk.ac.at/downloads/CoMA/CoMA_Tutorial.tar.gz

b. Unpack the archive to the current working directory:

   $ tar -xzf CoMA_Tutorial.tar.gz

v.ii. Notes

The CoMA installer for Linux is intended to be used on Debian-based systems. We thoroughly tested it on Ubuntu 18.04 and 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 or Docker option.

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 able to see the graphical user interface of CoMA, resulting in a termination of the pipeline.

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

vi. 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 (and often to a termination of the workflow)! This manual demonstrates the usage of CoMA with a detailed systematic tutorial using a small example dataset.

1. Start CoMA from the 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. **ATTENTION**: Keep in mind that you need writing permissions in the selected working directory! It is not recommended to run CoMA with **sudo**!

   $ 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 77.

3. Enter the title of your project and confirm with the **OK** button. For the tutorial, we use "**demo**" and confirm with **OK**.

   ATTENTION: Keep in mind that sample names must not include any special characters other than underscore ("_")! Moreover, it is not possible to have a number on the first position!

   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_Tutorial/* 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. Where possible, all upcoming steps will use the given number of CPU cores!

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 the quality of your forward or reverse sequences was 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 (together with a log 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:

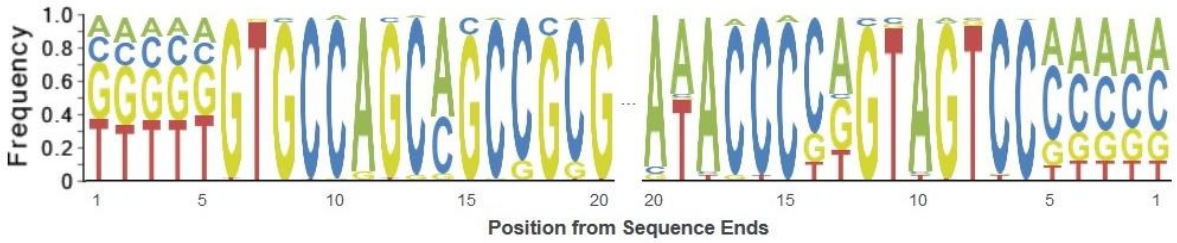
    .../Data/quality_reports/quality_before_filtering.

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).

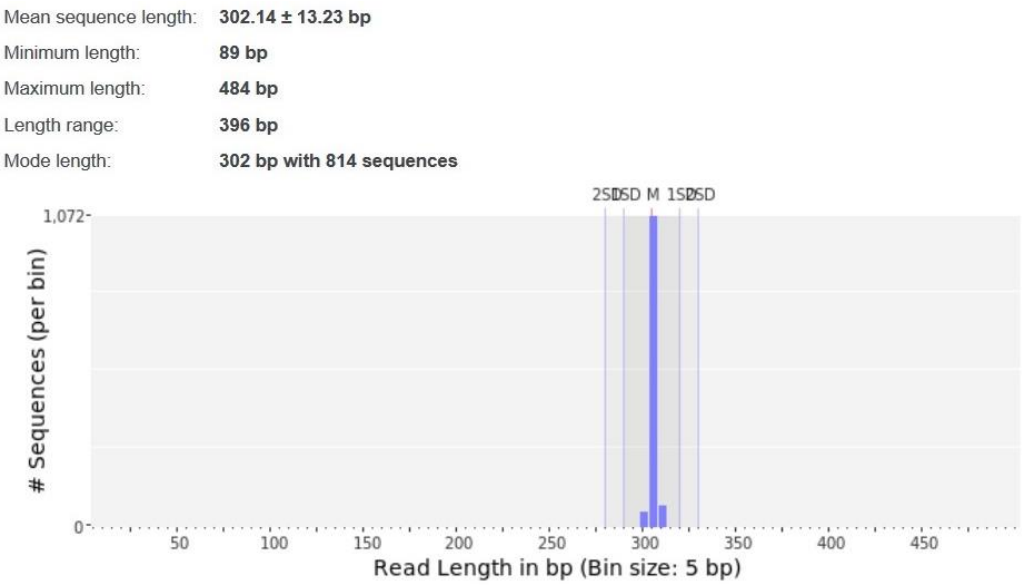
    Quality check of sample "a":



    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

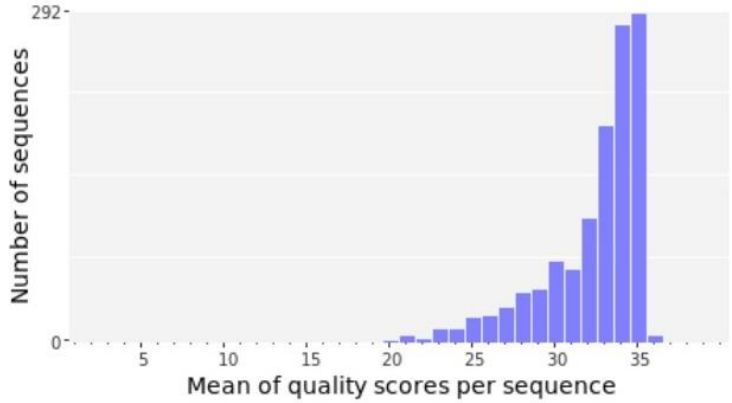
sequences, which are common for all reads (and thus show a high frequency of one single nucleotide at each position). You need to add the length of the primer to the length of the barcode to get the correct length for the trimming process.

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).



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 +/- 5%.

The mode length of all reads in sample "a" is 302 bp. After subtracting both trimmed sequences (302 - 24 – 25 = 253) you can set the limits to **266** 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 DNA 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. Choose an algorithm for the sequence clustering. CoMA support algorithms creating either OTUs (UPARSE, Swarm, CD-Hit), ASVs (DADA2) or ZOTUs (UNOISE3). All tools have their pros and cons, for detailed information consult the respective documentations. For the tutorial, we select "**cdhit**" and click **OK** (or double-click on **"cdhit"**).

    ATTENTION: If you want to use either the UPARSE or the UNOISE3 algorithm, USEARCH needs to be installed!

16. Now, you can enter the sequence identity used for OTU clustering between 80-100%. In most studies, OTUs are clustered at a 97% level; hence, we keep the default value of **0.97** and hit the **OK** button.

    ATTENTION: You can only define a sequence identity in CoMA when clustering with CD-Hit! UPARSE and Swarm are currently limited to 97%, and ASVs/ZOTUs can be interpreted as OTUs at a sequence identity of 100%.

17. Choose an algorithm for removing chimeric sequences. CoMA supports either USEARCH or VSEARCH; we select "**VSEARCH"** and click **OK** (or double-click on **"VSEARCH"**).

    REMARK: USEARCH is automatically selected if UPARSE was used as clustering algorithm!

    ATTENTION: If you want to use the USEARCH algorithm for chimera removal, the USEARCH tool package needs to be installed!

18. Select an aligner tool for taxonomic assignment. You can choose between five different options: RDP, blast, lambda, USEARCH and VSEARCH. Be aware that the RDP aligner tool can only search against the RDP database. For all 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)!

19. Choose between various reference databases, based on your preferences and the type of your samples (bacteria, fungi, protists, etc.): **SLV**: Silva LSU (23/28S) or SSU (16/18S), **GG**: greengenes (only 16S available), UNITE (ITS focused on fungi), PR2 (SSU focused on protists), HITdb (specialized on human gut microbiota), or beetax (specialized on bee 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/ASVs/ZOTUs and all unassigned OTUs/ASVs/ZOTUs 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.

20. "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 19, only *Silva* is affected. This step is skipped if *Silva* is not among the selected databases. At the end of this step, multiple files are created in a new directory:

    .../Results

    The most important ones are "abundance.biom" and "abundance.txt" since they represent the feature table including the taxonomic assignment. However, there are various other useful files; for a detailed description, please consult the respective section (chapter viii) in this manual.

    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.

    REMARK: During this step, "rem_seq.txt" is created. 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).

21. "Do you want to remove rare OTUs/ASVs/ZOTUs from your dataset?" → **Yes**

    This step can be used to discard very rare OTUs/ASVs/ZOTUs from analysis. OTUs/ASVs/ZOTUs 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.

22. Firstly, choose a minimum number of reads for an OTU/ASV/ZOTU to be retained. OTUs/ASVs/ZOTUs with fewer observations are excluded from the abundance file (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 **3** (which is a convenient value, often referred to as *doubletons*) and hit the **OK** button. You can change the number of OTUs/ASVs/ZOTUs with either moving the button of the scale dialog with your mouse or your left/right arrow keys (if the dialog window is active).

    REMARK: If you enter either 0 or 1 (or click the Cancel button), your data remains untouched and no OTUs/ASVs/ZOTUs are getting removed from the abundance table!

23. Secondly, choose a minimum number of samples in which an OTU/ASV/ZOTU must be present. OTUs/ASVs/ZOTUs which were found in fewer samples are excluded from the abundance file (and therefore from further analysis). For the tutorial, we choose **0** (meaning that no OTUs will be removed based on this criterion; this is equal to choosing "1") and hit the **OK** button.

24. "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: e.g. "abundance.groups.rarefaction", "abundance.groups.r_shannon" or "abundance.groups.r_coverage".

25. You can select between different calculators based on which the rarefaction curves will be constructed: *otu* (OTU/ASV/ZOTU 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.

26. 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"**).

27. 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**.

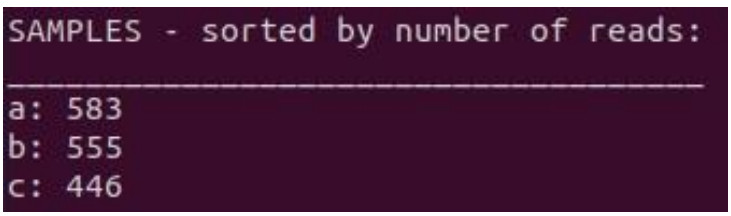
    REMARK: This option is not available if a vector graphic format (EPS, PDF, PS, SVG, SVGZ) was selected in the previous step!

28. "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/ASVs/ZOTUs 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.

29. "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).



We choose the smallest number of reads for the subsampling of our test results (always assuming, that no/only few OTUs are getting lost; this assumption may be wrong in this reduced dataset for the tutorial!): **446** (you may get slightly different read counts; in this case, use your lowest value) and hit the **OK** button.

30. "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 ("a", "b", "c"). 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 choose **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 31) or update the sample names in the existing one!

    ATTENTION: Keep in mind that sample names must not include any special characters other than underscore ("_")! Moreover, it is not possible to have a number on the first position!

31. "Do you want to add metadata to your samples?" → **Yes**

    Metadata are data describing your samples such as environmental conditions (e.g. temperature, pH, O_2 content), process parameters (e.g. stirring intensity, feeding rate, flow rate) or experimental characteristics (e.g. body site, drug load, season, altitude). 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 (e.g. pH_6_7 and pH_7_8).

32. 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.

    ATTENTION: Names for metadata variables must not include any special characters other than underscore ("_")! Moreover, it is not possible to have a number on the first position! Ignoring this may lead to severe problems during the upcoming steps.

33. A new dialog window informs us that we now need to fill in the information 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 only 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 may fail and you may 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!

34. 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"** (or Ctrl + S) and click **"Use Text CSV Format"**.

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 values 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"). Moreover, we recommend avoiding special characters other than "_" and "%".

35. Close now the program with the **X button** in the upper right corner of the window (or Ctrl + Q). A short checking step will follow, proving if your inputs were given appropriately. In our case, all variables should have been entered correctly and we can proceed by clicking the **OK** button. You can find (and edit) your metadata file in the following directory:

    .../Results

    REMARK: You can edit your metadata at any time later on; however, keep in mind that the structure of the file (sample names in the first column, variable names in the first row, tab-delimited, etc.) needs to be maintained! If you change sample names (step 30) after creating the metadata file, do not forget to update the "mapping.txt" file with the new sample names (either manually or by re-executing this step)!

36. "Do you want to generate a summary report?" → **Yes**

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

    .../Results/summary_reports

    ATTENTION: No file will be created if the summary report would be empty anyways (because there are no taxa left fitting the selection criteria provided in the upcoming steps)! In this case, you will get a warning message in the terminal window.

37. "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 metadata variable (arithmetic mean of the individual samples) 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.

    ATTENTION: If the entered variable name cannot be found in the mapping file, an error message appears in the terminal and the CoMA workflow is terminated!

    REMARK: If only a single metadata variable is included in the mapping file, the variable is selected automatically! In this case, a message is prompted in the terminal.

    REMARK: If the user chooses summary reports for individual samples (rather than for groups defined by metadata variables), the key word "individual" is used for labelling the files!

38. "Do you want a general summary?"

    CoMA offers general summaries as well as summaries for a specific taxon. 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 (e.g. *Firmicutes*, *Bacilli*, *Enterobacterales*, *Methanosarcinaceae*). For the tutorial, however, we will compute a general summary with all detected taxa, and hence click **"Yes"**.

39. Choose now which summary reports do you want to create (*Archaea*, *Bacteria*, *Fungi*, *Eukaryota*, and/or *Total)*. Since we are analysing 16S rRNA samples, we are selecting **"Archaea"**, **"Bacteria"** and **"Total"**. Hit the **OK** button.

40. "How many taxa do you wish to show for each taxonomic level?"

    You can now decide how many taxa will be shown for each taxonomic level (as long as there are enough entries available). If you are interested only in key organisms, you can focus on the 5 most abundant taxa for instance. For the tutorial, however, we want to see the entirety of all taxa, and hence type in **"999999"** and click the **OK** button.

41. "Do you want to create plots of the most important taxa?" → **Yes**

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

    .../Results/taxa_plots

    ATTENTION: No plots will be created if there are no taxa left fitting the selection criteria provided in the upcoming steps! In this case, you will get a warning message in the terminal.

    REMARK: Both, bar charts and heatmaps have pros and cons, and which option is preferable in your case depends on the research question and the investigated dataset. Generally, we suggest using heatmaps if there are many samples and/or taxa displayed in the graphic.

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

    This will group your samples in the graphics (arithmetic mean of the individual samples) 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.

    ATTENTION: If the entered variable name cannot be found in the mapping file, an error message appears in the terminal and the CoMA workflow is terminated!

    REMARK: If only a single metadata variable is included in the mapping file, the variable is selected automatically! In this case, a message is prompted in the terminal.

    REMARK: If the user chooses plots for individual samples (rather than for groups defined by metadata variables), the key word "individual" is used for labelling the taxa plot directory!

43. "Do you want general plots?"

    As previously described in step 38 for the summary reports, CoMA offers either general plots or plots based on a specific taxon. For demonstration purposes, we want to compute now specific plots and click **"No"**.

44. 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.

45. "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 if you are analysing habitats, which are not well investigated so far.

46. 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"**).

47. The next dialog asks for a threshold as well as for the pixel density of your graphics (**ATTENTION**: the second option is not available if a vector graphic format (EPS, PDF, PS, SVG, SVGZ) was selected in the previous step!). Taxa with a relative abundance below the threshold are excluded from the depiction. We type in **"1"** (%) as threshold (which is a common value for taxa plots) and **"300"** (dpi) as pixel density for high-quality and publication-ready graphics; click **OK**.

48. "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

49. "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 with two or three different categories can be used for Venn plots since CoMA currently does not support comparison of more than three groups.

    ATTENTION: If the entered variable name cannot be found in the mapping file, an error message appears and the workflow is terminated (no plots are created)!

    REMARK: If only a single metadata variable is included in the mapping file, the variable is selected automatically! In this case, a message is prompted in the terminal. Keep in mind that this may lead to an error if this variable has other than two or three categories!

50. "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.

    ATTENTION: Remember that only the comparison of two or three groups is currently possible with CoMA; entering another number leads to an error!

51. Choose which file format do you want to use. Enter **"tiff"** and click **OK**.

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

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

53. 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.

54. 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.

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

56. These steps are repeated two or three times, depending on your decision in step 50. 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 anyways) and hit the **OK** button. Then we enter **"B"** and click again the **OK** button.

57. "Do you want to label all areas with the exact number of taxa?" → **Yes**

    REMARK: Usually, we recommend labelling 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 an image-editing tool of choice (e.g. Adobe Photoshop).

58. "Do you want to calculate/plot the alpha diversity of your samples?" → **Yes**

    This step calculates alpha diversity (= within-sample diversity) using the abundance file created during the previous steps. The plot as well as supporting text files containing the calculated diversity for each sample/group on the one hand and statistics on the other hand can be found in the following directory:

    .../Results/alpha_diversity

    ATTENTION: Always keep in mind that the removal of rare OTUs/ASVs/ZOTUs (step 21) and/or subsampling (step 28) 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.

59. Different metrics are available for the calculation: *observed OTUs/ASVs/ZOTUs*, *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 file ("OTUphylo.nwk"). 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**").

60. "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 (± 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*.

    ATTENTION: If the entered variable name cannot be found in the mapping file, an error message appears and the workflow is terminated (no plot is created)!

REMARK: If only a single metadata variable is included in the mapping file, the variable is selected automatically! In this case, a message is prompted in the terminal.

61. 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.

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

63. 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 or #FF0000 for red), 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 (either individual results or arithmetic means ± standard deviation in case of groups) 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.

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

65. To analyse beta diversity, you can choose between two different options: ordination using Principal Coordinates Analysis (PCoA) and Hierarchical Cluster Analysis (HCA). For now, we select "**PCoA**" and click **OK** (or double-click on "**PCoA**"). A demonstration of the HCA follows later on (step 79).

66. Now you can select a metric for the calculation of the distance between your samples and hence the creation of a distance/similarity matrix. CoMA offers non-phylogenetic (*Minkowski*, *Euclidean*, *Manhattan*, *Cosine*, *Jaccard*, *Dice*, *Canberra*, *Chebyshev*, *Bray-Curtis*) as well as phylogenetic (*Weighted UniFrac*, *Unweighted UniFrac*) metrics. For the phylogeny-based methods, CoMA's tree file ("OTUphylo.nwk") 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**").

    REMARK: Please keep in mind that the *Minkowski* distance may not be the optimum choice for your data and we do not suggest it for analysing sequencing data! We are using it here for demonstration purposes since it is the only metric that allows for different P-norms (next step). For sequencing data, we generally suggest *Bray-Curtis* distance, *Jaccard* distance or weighted/unweighted *UniFrac* distance.

67. 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! Furthermore, the *Chebyshev* distance can be interpreted as *Minkowski* distance with an infitely high P-norm.

68. "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 63 for the alpha diversity plot. We want to group/colour 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.

    ATTENTION: If the entered variable name cannot be found in the mapping file, an error message appears and the workflow is terminated (no plots are created)!

    REMARK: If only a single metadata variable is included in the mapping file, the variable is selected automatically! In this case, a message is prompted in the terminal.

69. "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.

70. "Do you want to create a two-dimensional plot of your ordination?" → **Yes**

71. 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.

    REMARK: This dialog only shows up if two-dimensional plots were selected in the previous step!

72. 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**.

    REMARK: This dialog only shows up if two-dimensional plots were selected in the previous step!

73. 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.

74. 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 (*ordination.py*, *ordination_mapped.py*).

    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.

    REMARK: The additional files (distance matrix, eigenvalues, statistics) are always created, even if no plots were selected!

75. Click **OK**. In fact, this was the last step of the CoMA pipeline. However, we also want to do a hierarchical cluster analysis. To do so, we need to start a new CoMA run. Let's go, it's going to be fast!

76. Start CoMA from Linux terminal:

    $ coma

77. "Do you want to start a new project?" → **No**

78. Search and select the project folder that you created in course of this tutorial (do **not** double-click on it!). Click **OK**.

79. 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**.

80. Select "**HCA**" and click **OK** (or double click on "**HCA**").

81. 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. We choose "**average**" (better known as UPGMA method) as linkage method for our tutorial and hit the **OK** button (or double click on "**average**").

    REMARK: Be aware that some methods (*centroid*, *median*, *ward*) are limited to Euclidean distance metric (in step 82).

    REMARK: Every method has different strengths and limitations - technical discussion of each metric is available online and in ecology textbooks, but it is beyond the scope of this manual.

82. 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 and hit the **OK** button (or double click on "**braycurtis**").

    REMARK: 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!

    REMARK: 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.

83. "Do you want to plot the distance of each node in the dendrogram?" → **Yes**

    REMARK: Keep in mind that displaying the distances of each note can lead to overlapping issues in huge datasets with many samples!

84. 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"**).

85. The next dialog asks for the pixel density of your graphic (this option is not available if a vector graphic format ("EPS", "PDF", "PS", "SVG", "SVGZ") was selected in the previous step!). We type in **"300"** (dpi) for a high-quality figure and click the **OK** button.

86. "Do you want to use metadata to colour your sample names?" → **Yes**

    This shall help identifying structures in the dataset. If no metadata are used, all sample names are written in black. We want to colour our names 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.

    ATTENTION: If the entered variable name cannot be found in the mapping file, an error message appears and the workflow is terminated (no plots are created)!

    REMARK: If only a single metadata variable is included in the mapping file, the variable is selected automatically! In this case, a message is prompted in the terminal.

    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 here:

.../ (= main directory of your project)

Moreover, there is a detailed log file containing the complete standard output of this run stored at the same place. 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 in order to solve problems.

Thank you again for using the CoMA pipeline!

vii. 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, 15(12): e0243241. https://doi.org/10.1371/journal.pone.0243241

For a complete summary of all used third party software including their references, please read the **"citations.txt"** file, which can be found in the following directory:

.../Results

Keep in mind that a proper citation strategy helps the development teams to maintain and improve their products!

viii. 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 tackles files, which were computed during the CoMA run but which are not used by the pipeline directly. However, all files are stored in standardized formats and may be used by advanced users for secondary and more sophisticated analysis (e.g. using R). All result files 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 hierarchical 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(s).

ExtraFiles: Here you can find files summarizing the chimera removal step. Moreover, a file including the multiple sequence alignment of OTUs/ASVs/ZOTUs ("*OTU.MSA.fna"*) is stored here in FASTA format. Divergent positions are indicated with hyphens ("-").

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/ASV/ZOTU clustering and taxonomic assignment): http://lotus2.earlham.ac.uk/.

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. "*abundance.groups.rarefaction*" summarizes for instance all data when *otu* was the calculator and "*abundance.groups. r_shannon*" would include the results of a rarefaction analysis based on the Shannon-Wiener diversity.

summary_reports: Here you can find all your summary reports including either all detected taxa (key word *"Total"*) or entries associated only with a specific taxon. Files including the key word *"individual"* show results for each individual sample whereas files without show results for groups defined by metadata variables.

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

Venn_plots: Here you can find the Venn plots.

Files:

abundance.biom: This is the current abundance 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 abundance 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 "*abundance.txt"*.

abundance.txt: This is the most current abundance file of your analysis (including the identical information as the "*abundance.biom"* file but stored in a different format). 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!

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

abundance_original.txt: This is the original abundance 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 (equal information to "*abundance_original.biom"*).

abundance_without_subsampling.biom: This is the BIOM-converted abundance table after the removal of rare OTUs/ASVs/ZOTUs, but without subsampling or sample renaming.

abundance_without_subsampling.txt: This is the abundance table after the removal of rare OTUs/ASVs/ZOTUs, but without subsampling or sample renaming (equal information to "*abundance_without_subsampling.biom"*).

citations.txt: This file contains a summary of all third party software that was used for the current CoMA run. Please keep in mind that a proper citation strategy helps the development teams to maintain and improve their products!

hiera_BLAST.txt: This abundance 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 for taxonomic assignment!

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 (step 31).

OTU.fna: This file shows the extended OTU/ASV/ZOTU seed sequences in FASTA format. The sequences given here are identical to those in "*ExtraFiles/OTU.MSA.fna"*, but without the hyphens.

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

OTUphylo.nwk: This file represents a taxonomic tree in NEWICK format. It is used for phylogeny-based alpha/beta diversity analyses (*Faith's PD*, *weighted/unweighted UniFrac* distance). Moreover, advanced users may use this for even more-sophisticated

secondary analyses based on taxonomic information and the tree structure (e.g. with *FastTree*, *FigTree*, *GraPhlAn*).

phyloseq.Rdata: This is a *Phyloseq* object, which is ready to be loaded directly in *R* for secondary data analysis.

rem_seq.txt: This file summarized the amount of sequences that were dropped/lost 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).

ix. Update notes

ix.i. 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/ASV/ZOTU count, Chao1 diversity, Shannon-Wiener index, Simpson index, and Good's coverage for OTU/ASV/ZOTU (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.

ix.ii. CoMA 3.0

- A fourth installation option for CoMA was added, allowing the usage of the tool from within a Docker container. This option is available for users working on a Linux, macOS, or Windows system.
- CoMA can now also construct ASVs and ZOTUs (which can be interpreted as OTUs with a sequence similarity of 100%) in addition to OTUs.
- CoMA now offers different algorithms for sequence clustering in addition to UPARSE (constructing OTUs): Swarm (OTUs), CD-Hit (OTUs), UNOISE3 (ZOTUs), and DADA2 (ASVs).
- The sequence identity for clustering OTUs can now be adjusted between 80-100% (previously set to 97%). REMARK: This is currently only possible when using CD-Hit!
- The user can now also select VSEARCH for the removal of chimeric sequences (previously only USEARCH). REMARK: This is only possible for clustering algorithms other than UPARSE!
- Two additional aligning tools are now available for taxonomic assignment: USEARCH and VSEARCH (previously only Blast, Lambda and RDP).

- CoMA can now be used without installing USEARCH! **ATTENTION**: Keep in mind that doing so limits the options for sequence clustering (UPARSE and UNOISE3 are not available in this case), chimera removal (no USEARCH) and taxonomic assignment (no USEARCH)!
- A metadata variable is automatically selected if "*mapping.txt*" includes only a single one. REMARK: This is only relevant for steps in which the user decides to use metadata for structuring the data (summary reports, taxa plots, Venn plots, alpha diversity plots, ordination plots, and dendrograms).
- CoMA now shows a warning message if summary reports or taxa plots cannot be created because no data were found for the selected taxon.
- Summary reports are now stored in a dedicated directory (*/summary_reports*).
- Summary reports are now labelled based on the used metadata variable. This allows creating multiple summary reports without overwriting the previous ones.
- Taxa plots are now stored in different directories for each combination of settings (metadata variable, abundance threshold, including/excluding unassigned sequences). This allows creating multiple sets of plots without overwriting the previous ones.
- Running the step for removing rare OTUs/ASVs/ZOTUs with zero no longer leads to an error and the termination of the workflow.
- When doing alpha diversity analysis with metadata, underlying data are now stored as means ± standard deviation (previously, results were showed only for individual samples).
- Orientation lines at x=0 and y=0 were added to the PCoA plots.
- Sample names in dendrograms can now be labelled based on metadata. This shall help identifying structures within the dataset.
- A file including all used third party software and their references is now provided in order to guarantee a proper citation.
- All CoMA scripts were generally revised and updated. The files can now be downloaded from a new GitHub repository: https://github.com/SebH87/CoMA3.
- The CoMA manual was improved and updated.

x. 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.

- Qiime
- Mothur
- Pandaseq
- Prinseq-lite
- LotuS2
- USEARCH (has to be personally installed by the user)
- BiOM-format-tools