



Tutorial

OTU Clustering Step by Step

June 29, 2023

— Sample to Insight —

OTU Clustering Step by Step

This tutorial will take you through the different tools available in CLC Microbial Genomics Module and CLC Genomics Workbench to perform OTU clustering and to estimate alpha and beta diversities in microbial samples. Note that CLC Microbial Genomics Module also contains two workflows that recapitulate the different steps of this tutorial. If you wish to only run the workflows, please refer to the tutorial called "OTU Clustering Using Workflows".

Introduction To identify species present in microbial samples, DNA is extracted from the sample(s) of interest, a region of the 16S gene is PCR amplified, and the resulting amplicon is sequenced using an NGS machine. The bioinformatics task is then to assign taxonomy to the reads and tally the occurrences of species. Due to the incomplete nature of bacterial taxonomy and presence of sequencing errors in the NGS reads, a common approach is to cluster reads at some level of similarity into representative sequences of pseudo-species called Operational Taxonomic Units (OTUs), where all reads within e.g. 97% similarity are clustered together and represented by a single OTU sequence. The primary output of the clustering and tallying process is an OTU table, listing the abundances of OTUs in the samples under investigation. Secondary analyses include estimations of alpha and beta diversities in the context of sample metadata, in addition to statistical tests for differential abundance.

Prerequisites For this tutorial, you will need CLC Genomics Workbench with CLC Microbial Genomics Module installed. Note that results may differ slightly depending on the workbench and module versions being used. How to install modules and plugins is described here: <http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html>.

Overview This tutorial will guide you through the different tools for analyzing NGS data from microbial communities.

- First you will learn how to import your data into your workbench and how to prepare your samples for analysis with **Adapter Trimming** and the tool **Filter Samples Based on Number of Reads** to obtain sequences that are comparable in coverage for clustering.
- You will then perform **OTU clustering**, **Add Metadata to Abundance Table** and **Remove OTUs with Low Abundance** to help visualize data in an intuitive way (Stack bars, Area charts and Zoomable Sunbursts).
- You will **Align OTUs using MUSCLE** and construct a Maximum Likelihood Phylogenetic tree before measuring biodiversity with **Alpha Diversity** and **Beta Diversity** statistical tests.
- Finally, you will use specific tools to perform statistical tests such as PERMANOVA Analysis, Differential Abundance Analysis, and Create Heat Map for Abundance Table.

General tips

- Tools can be launched from the Workbench Toolbox, as described in this tutorial, or alternatively, click on the Launch button () in the toolbar and use the Quick Launch tool to find and launch tools.

- Within wizard windows you can use the **Reset** button to change settings to their default values.
- You can access the in-built manual by clicking on **Help** buttons or going to the "Help" menu and choosing "Plugin Help" | "CLC Microbial Genomics Module Help".

Downloading and importing the data For this tutorial we will be using a data set containing the sequences and metadata from a round robin trial of several soil types generated in a mock crime investigation as part of the MiSAFE project (<http://forensicmisafe.wix.com/misafe>). DNA was extracted, and a region of the 16S gene was PCR amplified using standard primers. The resulting amplicon were sequenced on an Illumina MiSeq machine (300 cycles, forward and reverse).

The data set includes the following files:

- **Sequence data** As we are dealing with paired-end data, a pair of 2 files differing only in "_R1_" and "_R2_" in their names represent the forward and reverse reads of the 12 data sets, respectively. There are two replicates (A and B) for each of the 3 samples taken at locations 1, 2 and 3, Site 1 being the crime scene, Site 2 and Site 3 being the places where Mr. X claims to have spent the weekend. There are 3 replicates (A, B and C) for each of the two different pairs of boots found in Mr. X's home: boot A and boot B. The data was generated from the same MiSeq run and is composed of demultiplexed .fastq files. The original files have been down-sampled to only contain 1/10th of the original reads. This was done in the interest of reducing the time needed for the different subsequent analysis steps in the context of this tutorial.
- **Metadata** The spreadsheet MetadataRoundRobin.xls contains metadata information - in essence a tabular description of the datasets. In this case it contains information about the origin of the samples by associating each data-set with its source (i.e., site or boot).
- **Sequencing primer sequences** The 16S primer sequences are provided in .clc format in 16s_primers_round_robin.clc.
- **Database** 16S_97_otus_GG.clc contains a database Operational Taxonomic Units (OTUs) to be used in the analysis. Note that several OTU databases are available for download in the Microbial Genomics Module.

Now that the prerequisites have been described, it is time to start with the tutorial.

1. Download the sample data from our website: http://resources.qiagenbioinformatics.com/testdata/otuclustering_tutorial/otuclustering_tutorial.zip and unzip it.
2. Start your CLC Workbench and go to **File** | **Import**  | **Illumina**  to import the 24 sequence files (ending with "fastq") (figure 1).
 - Ensure that the import type under General options is set to **Paired reads** and that the radio button for **Paired-end** is selected.
 - Minimum distance must be set to **200** and Maximum distance to **550**.

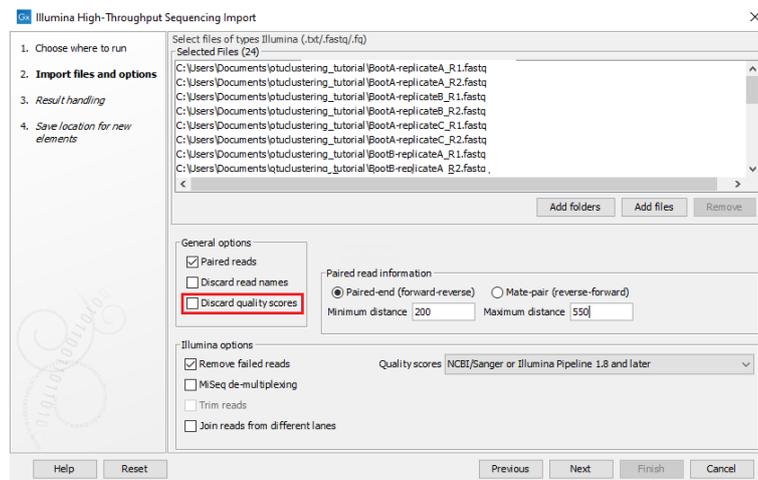


Figure 1: Import the data from the samples collected on the sites and on the boots of the suspect.

- It is crucial that quality scores are also imported, so make sure the option "Discard quality scores" is not checked. Click **Next**.
- Select the location where you want to store the imported sequences. We recommend that you create a new folder called **OTU clustering tutorial** for example, and a subfolder called **Illumina reads**. You can check that you have now 12 files labeled as "paired".

3. Import the database sequence data 16S_97_otus_GG.clc and the 16s_primers_round_robin.clc primer sequences by using the **Import | Standard Import** functionality.

It would be possible to import the metadata at this point, as described in the "OTU Clustering Using Workflows" tutorial, but we choose to import them later using the Add Metadata to Abundance Table tool.

OTU clustering

Remember that all the steps of the section below are included in the "Data QC and OTU Clustering" workflow for a convenient and automated way to perform your analyses. The detailed section below allows you to understand what the workflow does step by step.

Before you can start to cluster your reads they will need to be trimmed, and the reads with low coverage will be removed from the analysis.

The first step is to trim off the primer sequences.

1. Use the Launch button to start the Trim Reads () tool. Select the 12 sequences from your **Illumina reads** folder and click **Next**.
2. Leave parameters as default, i.e., **Trim using quality scores** with a limit set to 0.05 and **Trim ambiguous nucleotides** with a Maximum number of ambiguities set to 2. Click **Next**.
3. Leave the "Automatic read-through adapter trimming" option enabled, and use in addition the "16S_primers_round_robin" Trim Adapter List () (see figure 2) before clicking **Next**.

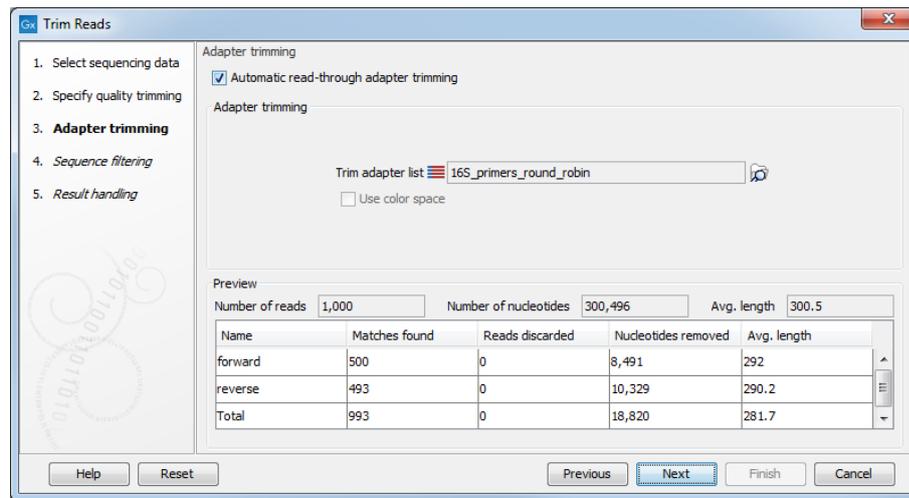


Figure 2: Setting up parameters for the adapter trimming step.

- In the wizard step called "Sequence Filtering", check **Discard short reads** and set it to 5 before clicking **Next**.
- Save the results using the default set up of the "Result handling" dialog. Create a new folder (called **trimmed** for example) before clicking **Finish**.

Only samples with similar coverage should be used for OTU clustering. However DNA extraction, PCR amplification, library construction, and sequencing may have introduced biases in the sequence data, with some samples represented by only few reads. Samples with many fewer reads than the others may be a biased representation of what was in the original sample and should be removed.

- Launch the tool **Metagenomics**  | **Amplicon-Based Analysis**  | **Filter Samples Based on Number of Reads** 
- Select the sequences from your **trimmed** folder. Click **Next**.
- Leave the parameters as default with 100 as the minimum number of reads and 50 the minimum percent from the median. The tool will discard samples that do not fulfill both requirements. Click **Next**.
- We would like a list of samples with sufficient coverage. To do this, check the option **Copy samples with sufficient coverage** but not **Copy discarded samples**. Select the option to save the results and provide a new folder to save them to (here **high coverage**) before clicking **Finish**.

The output of this tool is a list of the samples that passed the coverage cut-off, as well as a table. You can review these by opening them from the Navigation Area or by using the little arrow to the right of the analysis name in the Processes tab and by choosing the option "Show results" (see figure 3).

They will open in the View Area. The table (see figure 4) reports the samples name, how many reads they each contain, and which of these samples passed the **Filter Samples Based on Number of Reads** tool cut-off. The ones that did not meet the criteria are noted in the table

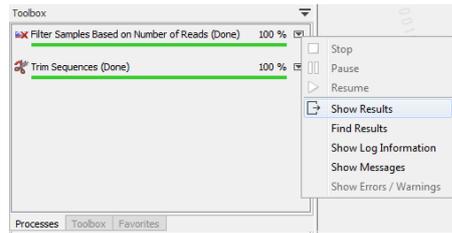


Figure 3: Find your results using the little arrow to the right of the analysis name in the Processes tab.

with the text "Number of reads too low". These samples will not be included in the subsequent analysis.

1 Number of reads

| Sample | Number of reads | Notes |
|--|-----------------|-------------------------|
| BootA-replicateA_R1 (paired) trimmed (paired) | 2260 | Number of reads too low |
| BootA-replicateB_R1 (paired) trimmed (paired) | 15732 | Passed |
| BootA-replicateC_R1 (paired) trimmed (paired) | 26500 | Passed |
| BootB-replicateA_R1 (paired) trimmed (paired) | 18040 | Passed |
| BootB-replicateB_R1 (paired) trimmed (paired) | 22572 | Passed |
| BootB-replicateC_R1 (paired) trimmed (paired) | 25042 | Passed |
| CrimeSite1-replicateA_R1 (paired) trimmed (paired) | 18716 | Passed |
| CrimeSite1-replicateB_R1 (paired) trimmed (paired) | 20370 | Passed |
| Site2-replicateA_R1 (paired) trimmed (paired) | 19034 | Passed |
| Site2-replicateB_R1 (paired) trimmed (paired) | 24096 | Passed |
| Site3-replicateA_R1 (paired) trimmed (paired) | 27074 | Passed |
| Site3-replicateB_R1 (paired) trimmed (paired) | 17960 | Passed |

Figure 4: Output table from the Filter Samples Based on Number of Reads tool.

Your data is now ready for OTU clustering. The OTU clustering tool clusters the reads and reduces the read collection in each sample to representative sequences (cluster centroids) that are 97% similar to any member of the cluster they represent. Database sequences can be included in this clustering and can also be specified as centroids, with their annotations inherited by the cluster they represent. In addition, chimeric sequences, which are frequent artifacts of PCR reactions, are detected. This is done by assessing whether a sequence is likely to be made up of segments of two different sequences, where those sequences appear more frequently in the current OTU collection than the suspected chimeric sequence.

1. Launch **Metagenomics** (🌿) | **Amplicon-Based Analysis** (🔬) | **OTU clustering** (🌐).
2. Select as input the sequence lists from the **high coverage** folder, which contains the 11 samples that passed the **Filter Samples Based on Number of Reads** tool.
3. Set the parameters as follows (figure 5):
 - Check the radio button called **Reference based OTU clustering** and select the file 16S_97_otus_GG as your annotated reference database.
 - Check the checkbox **Similarity percent specified by the OTU database** as we wish to perform the OTU clustering using the same similarity threshold used to build the reference database (i.e., 97%)
 - To make the calculations faster for this tutorial, ensure that **Allow creation of new OTUs** is **NOT** selected.
 - All other parameters are left as default.

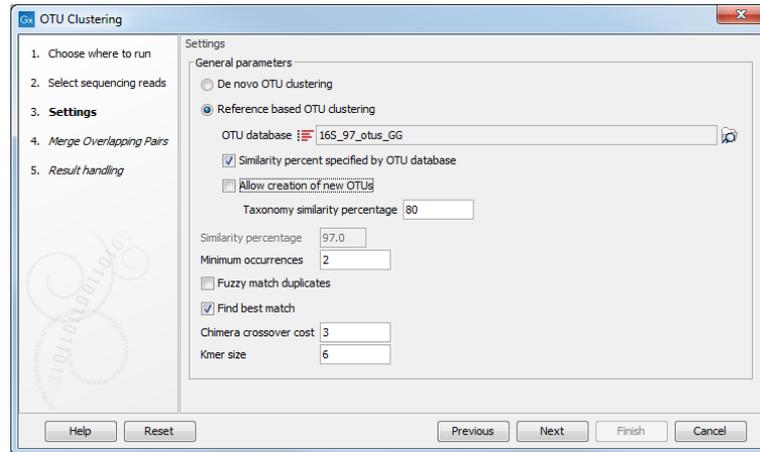


Figure 5: Settings of the OTU clustering tools.

- In the "Merge Overlapping Pairs" step, forward and reverse reads of paired reads are merged to yield one high quality representative. Use the default parameters values as seen in figure 6. Note that when working with your own data set, we recommend using the default values when analyzing the data for the first time, and potentially re-adjust parameters in subsequent analysis runs. Click **Next**.

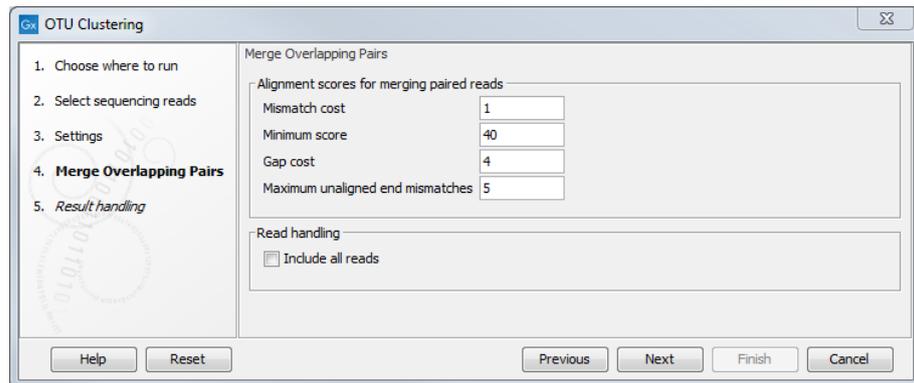


Figure 6: Setting up the merge parameters.

- Select the option to save the results and provide a new folder to save them to (here **OTU and statistics**) before clicking **Finish**.

If you open the Processes tab, you will see that the first task of OTU clustering tool is to map the reads before clustering them. Three output files are produced, named using the input filename plus an extension indicating what type of output it contains:

- **(Sequence)** output contains a list of the sequences that were clustered.
- **(Table)** output contains the abundance table for the sequences that clustered with OTUs from the annotated reference database. The table has a column defining the taxonomic name of the OTU, sample-specific columns for abundance counts, and a column for total abundance.
- **(Chimeras)** output contains a table listing the sequences found to be chimeras and their abundance in the different samples.

There are several ways to look at your newly generated OTU clusters in addition to the table view: Stacked Bar Charts, Stacked Area Charts () and Zoomable Sunbursts ().

Visualization of the OTU abundance table and metadata

To enhance the visualization of the OTU abundance table, it is useful to decorate it with metadata. This allows multiple samples to be aggregated based on particular attributes, for example, location.

1. Select **Metagenomics** () | **Abundance Analysis** () | **Add Metadata to Abundance Table** ()
2. Specify the OTU (Table) as input.
3. Select the file describing the metadata on your local computer (MetatdataRoundRobin.xls) as shown in figure 7 and click **Next**.

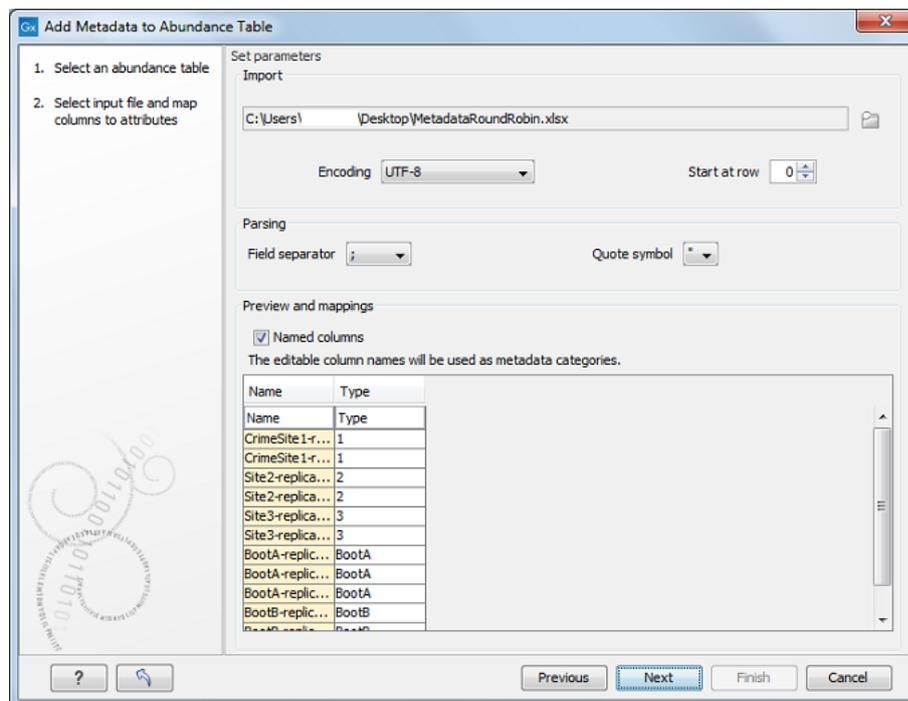


Figure 7: Setting up metadata parameters.

Note that in this table, the names in the Name column must be the same as the ones used to describe each sample initially. Each following column can be a different metadata category. Here we have only one metadata category, which is called **Type**.

4. Save the table. This will overwrite the previous OTU (Table) such that the tables and graphs remain the same as before, but in side panel on the right, the option in the Data tab allowing you to **Aggregate samples** will now include the metadata attributes. If the (Table) data element was open for viewing when you ran the above tool, then, you need to close it and then open it again to see the changes.
5. Using the icons from the left hand side lower corner of the View, display the results as a Stacked Area Charts (). Then choose to aggregate features as Phylum, and aggregate samples on Type (as in figure 8).

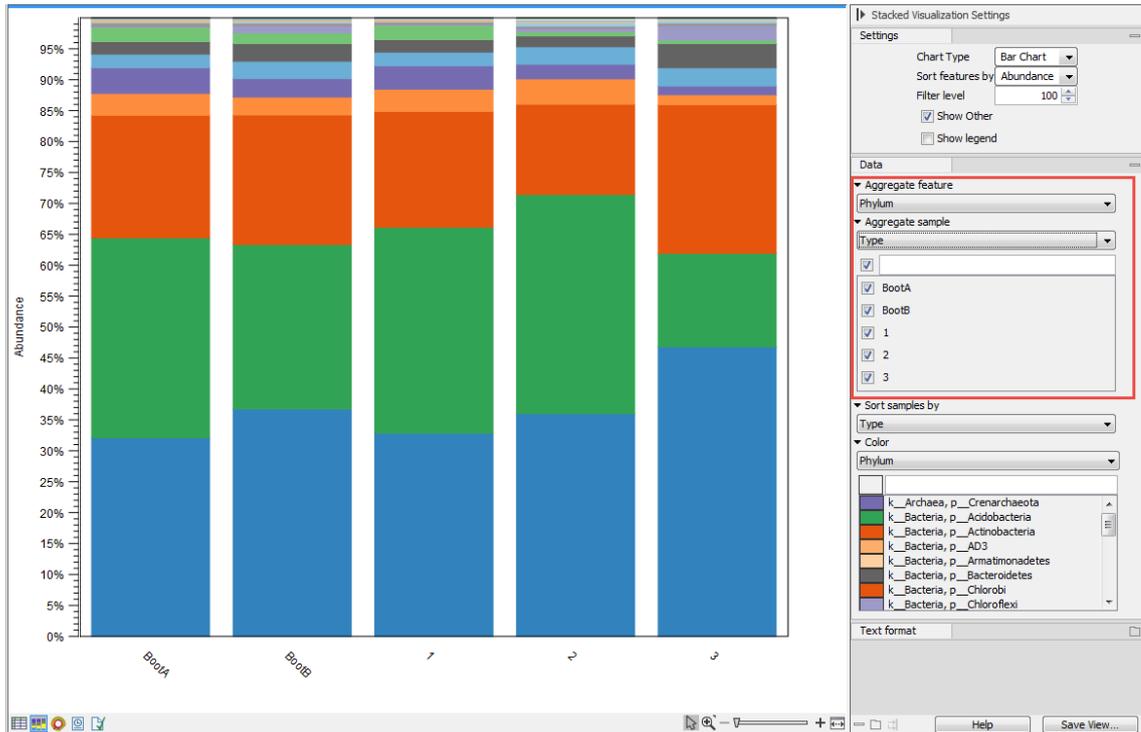


Figure 8: Aggregate samples based on metadata information.

To simplify the visualization of the OTU clustering results even more, you can filter out low abundance OTUs from the OTU table. You define the count level that must be exceeded across all the samples in the experiment for an OTU to be included in the results.

1. Select **Metagenomics** (🌿) | **Amplicon-Based Analysis** (🔬) | **Remove OTUs with Low Abundance** (🗑️)
2. Choose the OTU (Table) as input.
3. Select a threshold Minimum combined abundance for removal of OTU (here 10) and save your result. The new table will be labeled as (Filtered).

Open OTU (Table) (Filtered) and click on the Stacked Bar Chart icon to visualize your results. In the right side panel, choose to aggregate samples by Type as you did previously. You can see a striking similarity between the Boot A profile found on the suspect's boots and the profile of the soil from Site 1. We wish to evaluate this similarity further by performing statistical analyses on the different samples.

Estimation of Alpha Diversity

Remember that all the steps of the two sections below are included in the "Estimate Alpha and Beta Diversities" workflow for a convenient and automated way to perform your analyses. The detailed section below allows you to understand what the workflow does step by step.

Alpha diversity estimates describe the number of species (or similar metrics) in a single sample, while beta diversity estimates differences in species diversity between samples. In CLC Microbial Genomics Module, measures of alpha and beta diversity require a phylogenetic tree of all OTUs

(Phylogenetic diversity and UniFrac distances). The phylogenetic tree is reconstructed using a Maximum Likelihood approach based on a Multiple Sequence Alignment (MSA) of the OTU sequences generated by MUSCLE in the workbench.

Note: The default behavior of the MSA generation from OTU tables is to only include the 100 most abundant OTUs. Hence, the phylogenetic tree used for calculating the phylogenetic diversity and the UniFrac distances disregards the low abundance OTUs by default. If more OTUs are to be included, the default settings for the MUSCLE alignment need to be changed accordingly.

1. Select **Metagenomics**  | **Amplicon-Based Analysis**  | **Align OTUs using MUSCLE** .
2. Choose an OTU abundance table as input. Note that by default only the top 100 most abundant OTUs are aligned using MUSCLE and then used to reconstruct the phylogeny tree in the next step. Thus, you can choose either the OTU (Table) or the OTU (Table) (Filtered) as input.
3. Leave the parameters set to the defaults.
4. Choose to save the MSA in the **OTU and statistics** folder. The output will be given the name of the input data with the word alignment appended (the names in the section below reflect the use of the OTU table as input).
5. Use the Launch button  to look for the **Maximum Likelihood Phylogeny**  tool. Double click on the name in the list to get it started.
6. Choose the alignment **OTU (Table) alignment** as input.
7. Construct a **Neighbor Joining** tree with a **Jukes Cantor** nucleotide substitution model. You can uncheck the other parameters.
8. Leave the option to run a bootstrap analysis in the next wizard window unchecked.
9. Save your tree in the **OTU and statistics folder**. It will be given the name OTU (Table) alignment_tree.
10. Select **Metagenomics**  | **Abundance Analysis**  | **Alpha Diversity**  and choose the abundance table **OTU (Table)** as input. Click **Next**.
11. Select **Total number** as Alpha diversity measure (figure 9). In addition, specify the phylogenetic tree reconstructed in the previous step **OTU (Table) alignment_tree** and select **Phylogenetic diversity**.
In the Rarefaction analysis parameters box, check the **Set maximum depth** option and set it to 5,000 (figure 10). Click **Next**.
12. Save the resulting file. It is called OTU (Table)(Alpha Diversity) and contains as many graphs as there were parameters selected in the previous step (in this case one for Total number, and one for Phylogenetic diversity).

Each plot contains the rarefaction results of the specified alpha diversity measure. Each line corresponds to one sample. The coloring scheme can be set by using the Lines and dots settings in the right hand side panel. It is possible to change the line color of each sample one by one, or

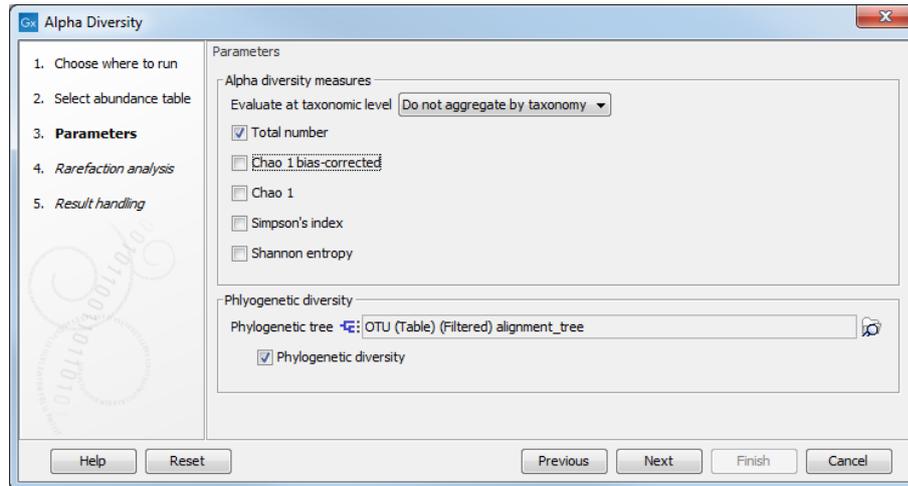


Figure 9: Parameters for alpha diversity analysis.

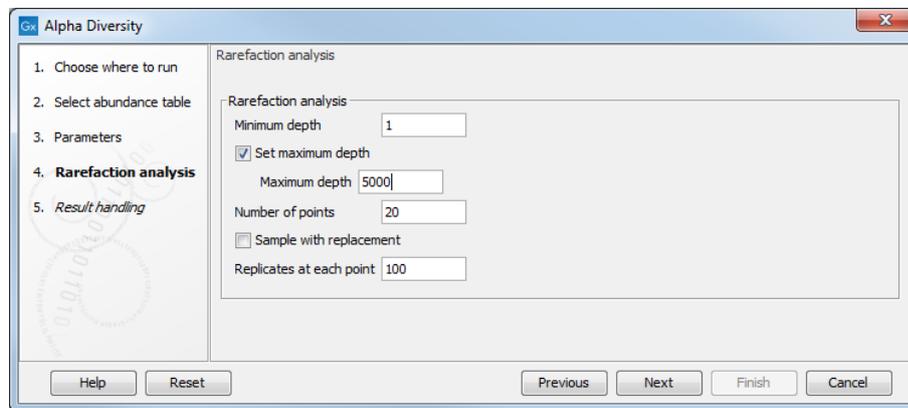


Figure 10: Parameters for rarefaction analysis.

of a metadata layer, or of all samples at once. In the following graph (figure 11) the lines do not plateau, indicating that we would need more samples to reach a definite conclusion, but Boot A samples seem to have similar measures of alpha diversity as the sites 1 and 2, while BootB samples look closer to site 3 when it comes to alpha diversity measures.

Switch the view to box plot () . Here, you can display additional statistics.

- Set the Boxplot color and Point color to "Type".
- Set the **Rarefaction level** to 1210. The values of "Rarefaction level" are the same as the horizontal axis in the Alpha Diversity Graph view.
- Check "Show p-values" and set the Mann-Whitney U test reference to BootA (figure 12) to perform a pair-wise **Mann-Whitney U test** of BootA to the order types. While BootA and BootB appear different, we would need more samples in order to reach statistical significance.

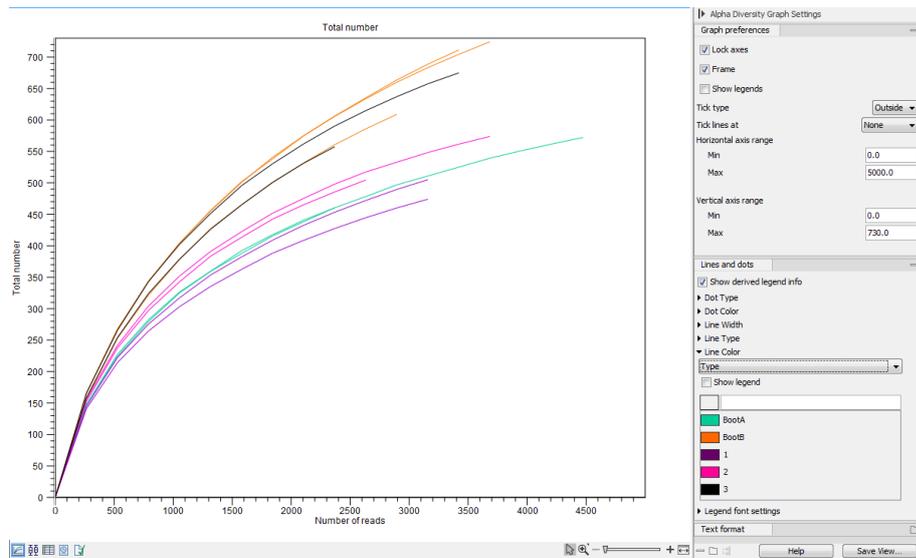


Figure 11: Results of the alpha diversity analysis measured using Total number as parameter.

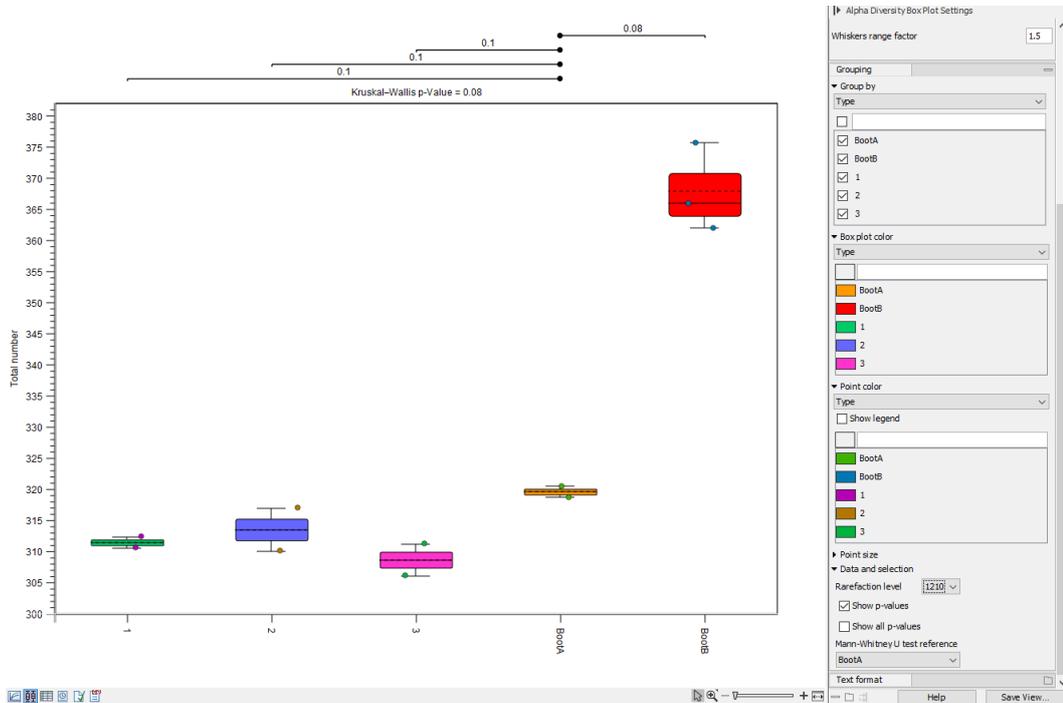


Figure 12: Alpha diversities shown in a box plot..

Estimation of Beta Diversity

Beta diversity estimates differences in species diversity between samples.

1. Select **Metagenomics** | **Abundance Analysis** | **Beta Diversity**
2. Choose the OTU (Table) as input and click **Next**.
3. Specify the phylogenetic tree reconstructed from the alignment of the most abundant OTUs in the previous step, OTU (Table) (Filtered) alignment_tree. Select D_0.5 UniFrac (figure 13) and deselect all other distance measures. Click **Next**.

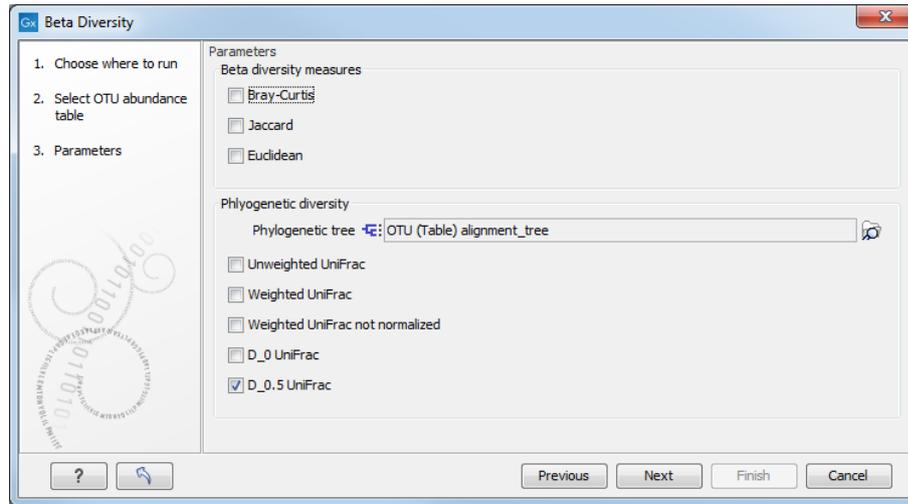


Figure 13: Setting up parameters for the beta diversity analysis.

4. Save the results in the **OTU and statistics** folder.

The beta diversity analysis tool performs a Principal Coordinate Analysis PCoA () on the UniFrac distances (figure 14).

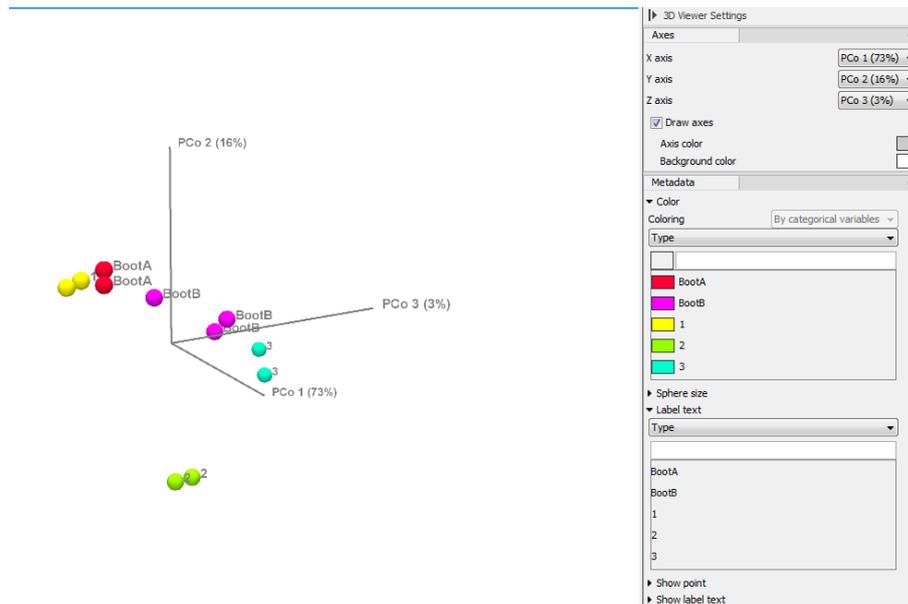


Figure 14: Result of the beta diversity analysis.

In a PCoA of the beta diversities, spheres in the plot can be colored according to metadata. Soil samples cluster according to their origin, revealing which of the samples from the Boot A are from which site. In this case, Site 1 and Boot A are clearly clustered together, supporting our suspicion that the soil on the Boot A is the same as the one found on Site 1.

Additional statistical analyses

To further assess the similarity between samples, run a differential abundance analysis to find the OTU's which have the most significantly different abundance across all samples.

1. Open the **Metagenomics | Abundance Analysis (📁) | Differential Abundance Analysis** tool.
2. Choose OTU (Table) as Abundance table.
3. Choose **Type** as Metadata factor and **Across groups (ANOVA-like)** as Comparisons (see figure 15). Click **Next**.
4. Save the result in the OTU and statistics folder.

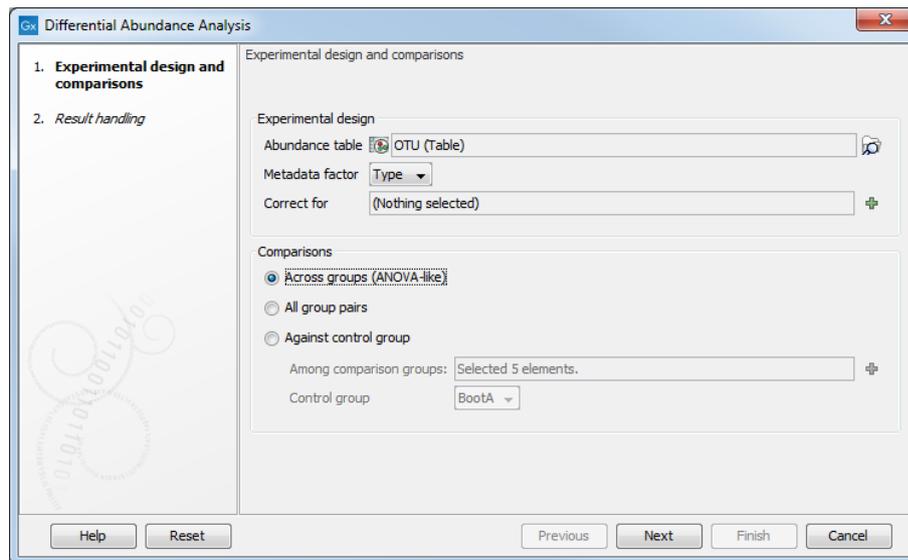


Figure 15: Set up a differential abundance analysis.

Open the OTU (Table) (differential abundance analysis) and sort the table in ascending order for the FDR p-value column. Highlight the 25 most different OTU's across all samples (FDR p-value=0). The number of rows selected is visible in the right hand side corner of the workbench. Click on **Copy Names to Clipboard**.

Constructing a heat map and dendrogram from these 25 OTU's will help in assessing similarity between samples.

1. Open the **Metagenomics (📁) | Abundance Analysis (📁) | Create Heat Map for Abundance Table (📁)** and choose OTU (Table) as input.
2. Leave the parameters as set by default, i.e., the distance to **Euclidean** and clusters to **Complete linkage**. Click **Next**.
3. In the next wizard window, select **Specify features** as Filter settings, and paste in the names of the 25 most different OTU's in the Specify features field (figure 16). Click **Next**.
4. **Save** the result in the "OTU and statistics" folder.

Display the heat map by double-clicking on it in the Navigation Area (figure 17).

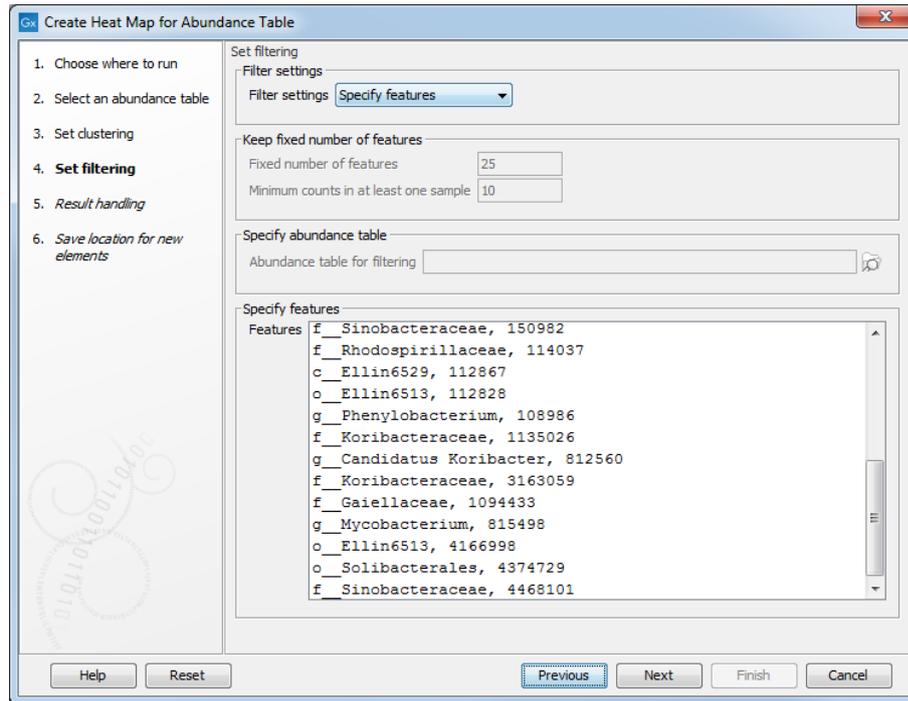


Figure 16: Filter based on the 25 most different OTU.

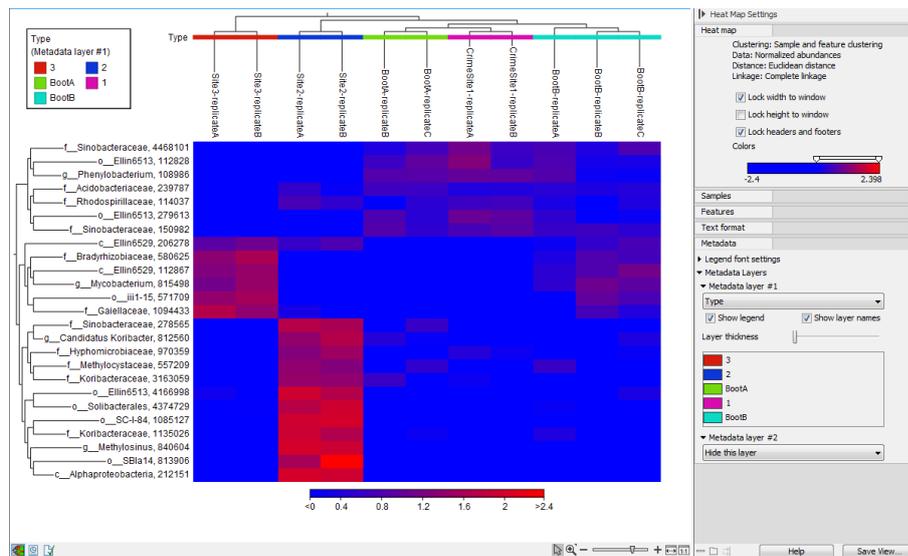


Figure 17: Heat map from the abundance table.

Set the visualisation parameters like in the side panel in figure 17. We can now see that Boot A is again nested together with Site 1, confirming once more that the soil found on Boot A is extremely similar to the one sampled from Site 1.

Finally, you can assess the robustness of your results by running a PERMANOVA analysis on your samples. PERMANOVA can be used to measure the effect size and significance of beta diversity.

1. Select **Metagenomics** (🗄️) | **Abundance Analysis** (🌐) | **PERMANOVA Analysis** (🎯).
2. Choose OTU (Table) from the **OTU and statistics** folder as input and click **Next**.

3. Select **Type** as Metadata group and click **Next**.
4. Specify the phylogenetic tree (OTU (Table) alignment_tree) from the **OTU and statistics** folder. Select D_0.5 UniFrac and deselect all other distance measures. Leave the number of permutations to 99,999. Click **Next**.
5. Save the report in the **OTU and Statistics** folder.

The result of the PERMANOVA analysis is a table (figure 18).

1 PERMANOVA analysis (D_0.5 UniFrac)

| Variable | Groups | Pseudo-f statistic | p-value |
|----------|-----------------------|--------------------|---------|
| Type | BootA, BootB, 1, 2, 3 | 14.84554 | 0.00068 |

| Group 1 | Group 2 | Pseudo-f statistic | p-value | p-value (Bonferroni) |
|---------|---------|--------------------|---------|----------------------|
| BootA | BootB | 4.21139 | 0.20000 | 1.00000 |
| BootA | 1 | 1.09631 | 0.66667 | 1.00000 |
| BootB | 1 | 5.49145 | 0.20000 | 1.00000 |
| BootA | 2 | 16.60434 | 0.33333 | 1.00000 |
| BootB | 2 | 7.10021 | 0.10000 | 1.00000 |
| 1 | 2 | 19.97387 | 0.33333 | 1.00000 |
| BootA | 3 | 31.60816 | 0.33333 | 1.00000 |
| BootB | 3 | 16.18129 | 0.10000 | 1.00000 |
| 1 | 3 | 36.99953 | 0.33333 | 1.00000 |
| 2 | 3 | 18.18110 | 0.33333 | 1.00000 |

Figure 18: Result of the PERMANOVA analysis. Values can differ slightly depending on the plugin version used.

The PERMANOVA confirms that the clusters are significant, but with only two to three replicates for each sample or group, the clustering is not significant on pair-wise comparisons of the Types. The investigators would need more samples - in particular from the soles of the Boot A and from site 1 - to transform this analysis into actual forensic evidence!