QIAGEN

Tutorial

OTU Clustering and Analysis of Microbial Communities

September 15, 2016

Sample to Insight

CLC bio, a QIAGEN Company · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.clcbio.com · support-clcbio@qiagen.com

OTU Clustering and Analysis of Microbial Communities

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 "Microbial profiling using the CLC Microbial Genomics Module 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 Taxonomical 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 either CLC Genomics Workbench (Version 7.5 or higher), or Biomedical Genomics Workbench (Version 2.1 or higher), 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://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Installing_plugins.html, as well as in the module manual.

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 **Optional Merge Paired Reads**, Adapter Trimming, **Fixed Length Trimming** and the tool **Filter Samples Based on the Number of Reads** to obtain sequences that are comparable in length and 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 Phylogenic 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, Create Heat Map for Abundance Table .

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.csv 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. Note that several primer-databases are available for download in the Microbial Genomics Module.
- **Database** 16S_97_otus_GG.clc contains a database Operational Taxonomic Units (OTUs) to be used in the analysis.

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

1. Download the sample data from our website: http://download.clcbio.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 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. Click on the button labeled Next and select the location where you want to store the imported sequences. We recommend that you create a new folder called **Illumina reads** for example. You can check that you have now 12 files labeled as "paired".
3. Import the database sequence data by drag-and-drop the 16S_97_otus_GG.clc database and the 16s_primers_round_robin.clc primer sequences into your destination folder in your CLC Workbench, or by using the Standard Import button on top of the Navigation Area.

All of the data needed to get started is now imported; you can begin the steps leading to OTUs clustering.

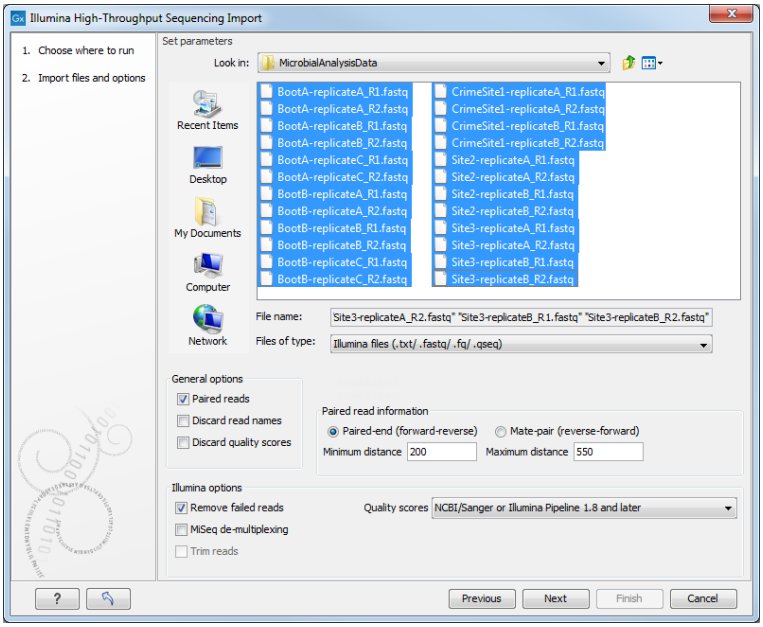

Figure 1: *Import the data from the samples collected on the sites and on the boots of the suspect.*

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 above 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, merged, and the reads with low coverage will be removed from the analysis.

The first step is to trim off the primer sequences.

1. Double click on the 16s_primers_round_robin file in the Navigation Area to open the trim adapter list.
2. Select the forward primer and click on the button labeled Edit Row to ensure that:
    - the parameter Strand is set to **Plus**
    - Action is again set to **Remove Adapter**.
    - Alignments scores costs are set to 2 for **Mismatch cost**.
    - Alignments scores costs are set to 3 for **Gap cost**.
    - **Match thresholds** are both checked and set to 8.
3. Similarly, select the reverse primer and click on the button labeled Edit Row to ensure that:
    - the parameter Strand is set to **Plus**
    - Action is again set to **Remove Adapter**.
    - Alignments scores costs are set to 2 for **Mismatch cost**.
    - Alignments scores costs are set to 3 for **Gap cost**.
    - **Match thresholds** are both checked and set to 8.

4. Launch the Trim sequences tool **Toolbox** | **NGS Core Tools ( )** in CLC Genomics Workbench or **Preparing Raw Data ( )** in Biomedical Genomics Workbench | **Trim Sequences ( )** and select only the all sequences from your **Illumina reads** folder. Click on the button labeled Next.

5. 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 on the button labeled Next.

6. Select the "16S_primers_round_robin" Trim Adapter List ( ) (see figure 2), check the option **Search on both strands**, and click on the button labeled Next.

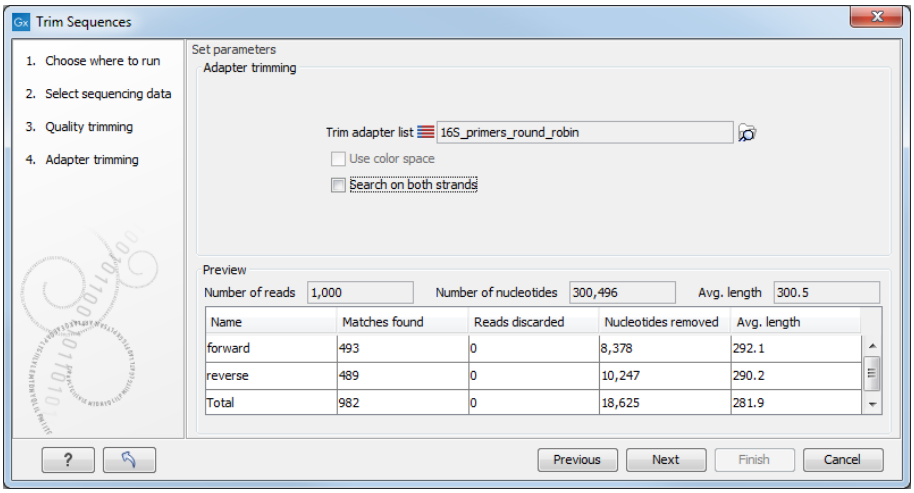

Figure 2: *Setting up parameters for the adapter trimming step.*

7. Check **Discard reads below length** and set it to 5 in the wizard step called "Sequence Filtering" and click on the button labeled Next.

8. Select the option to save the results and provide a new folder to save them to a new folder (here **trimmed**) before clicking on the button labeled Finish.

Second, in the case of paired reads, one can merge the forward and reverse reads to yield one high quality representative.

1. Launch the OTU clustering tool **Toolbox** | **Microbial Genomics Module ( )** | **OTU clustering ( )** | **Optional Merge Paired Reads ( )**, select the 12 paired-end sequences lists in the **trimmed** folder and click on the button labeled Next.

2. Relax alignments parameters to 1 for **Mismatch cost**, 40 for **Minimum score**, 4 for **Gap cost** and 5 as the **Maximum unaligned end mismatches** (see figure 3). Note that these values are based on previous analyses of the data and adjusted to provide the most relevant results based on the data set at hand. When working with your own data set, we recommend using the default values when analyzing the data for the first time, and to potentially re-adjust parameters in subsequent analysis runs. Click on the button labeled Next.

3. Select the option to save the results and provide a new folder to save them to (here **merged pairs**) before clicking on Finish.

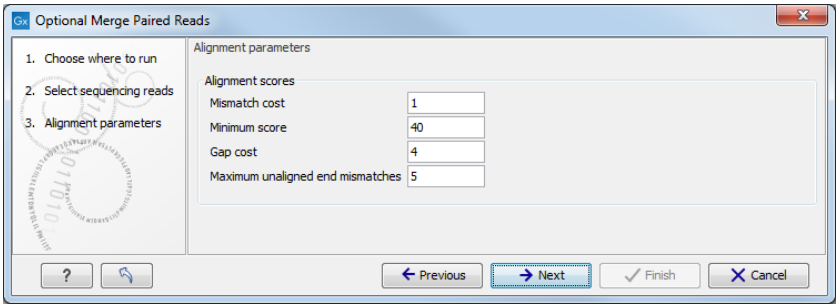

Figure 3: *Setting up the merge parameters.*

The merged read pairs are saved into sequence lists under their original name with the text "merged" appended. The sequences that were not merged will also appear in this folder with "not merged" appended to the end of the original sample names. The "not merged" reads are those that do not satisfy the requirements specified when we set up the merging analysis. These unmerged sequences will not be used further in this tutorial.

In order to compare sequences and cluster them, they all need to be trimmed to the exact same length.

1. Launch **Toolbox** | **Microbial Genomics Module (**) |**OTU clustering (**) | **Fixed Length Trimming (**)
2. Select the lists of trimmed merged sequences from your **merged pairs** folder (leave the sequences that are "not merged"). Note that a warning message informs you that "Paired reads are not allowed" but it does not concerns the reads from the merged pairs folder since they have been merged. Click on the button labeled Next.
3. Leave the length trimming set to **Automatic read length**. We already trimmed the primer sequences from the reads, so leave the "Read offset primer" setting blank.
4. Select the option to save the resulting fixed length sequences and provide a new folder to save them to (here **trimmed fixedLength**) before clicking on 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 then the others may be a biased representation of what was in the original sample and should be removed.

1. Launch the tool **Toolbox** | **Microbial Genomics Module (**) | **OTU clustering (**) | **Filter Samples Based on the Number of Reads (**)
2. Select the sequences from your **trimmed fixedLength** folder. Click on the button labeled Next.
3. 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 on the button labeled Next.

4. 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 on the button labeled 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 4).

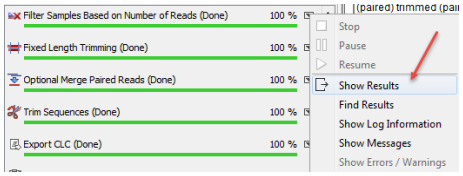

Figure 4: *Find your results using the little arrow to the right of the analysis name in the Processes tab.*

They will open in the View Area. The table (see figure 5) reports the samples name, how many reads they each contain, and which of these samples passed the **Filter Samples Based on the Number of Reads** tool cut-off. The ones that did not meet the criteria are noted in the table with the text "Number of reads too low". These samples will not be included in the subsequent analysis.

Sample	Number of reads	Notes
BootA-replicateA_R1 (paired) trimmed (paired) merged fixedLength	760	Number of reads too low
BootA-replicateB_R1 (paired) trimmed (paired) merged fixedLength	5692	Passed
BootA-replicateC_R1 (paired) trimmed (paired) merged fixedLength	9567	Passed
BootB-replicateA_R1 (paired) trimmed (paired) merged fixedLength	6635	Passed
BootB-replicateB_R1 (paired) trimmed (paired) merged fixedLength	8338	Passed
BootB-replicateC_R1 (paired) trimmed (paired) merged fixedLength	8318	Passed
CrimeSite1-replicateA_R1 (paired) trimmed (paired) merged fixedLength	6856	Passed
CrimeSite1-replicateB_R1 (paired) trimmed (paired) merged fixedLength	7353	Passed
Site2-replicateA_R1 (paired) trimmed (paired) merged fixedLength	6704	Passed
Site2-replicateB_R1 (paired) trimmed (paired) merged fixedLength	8569	Passed
Site3-replicateA_R1 (paired) trimmed (paired) merged fixedLength	9595	Passed
Site3-replicateB_R1 (paired) trimmed (paired) merged fixedLength	6305	Passed

Figure 5: *Output table from the Filter Samples Based on the 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 **Toolbox** | **Microbial Genomics Module ()** | **OTU clustering ()** | **OTU clustering ()**.
2. Select as input the sequence lists from the **high coverage** folder, which contains the sample(s) that passed the **Filter Samples Based on the Number of Reads** tool.
3. Set the parameters as follows (figure 6):

- 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 **Use the similarity percent specified by the reference database** as we wish to perform the OTU clustering using the same similarity threshold used to build the reference database (i.e., 97%)
- Due to time-limit for this tutorial, ensure that **Allow creation of new OTUs** is NOT selected.
- All other parameters are left as default.

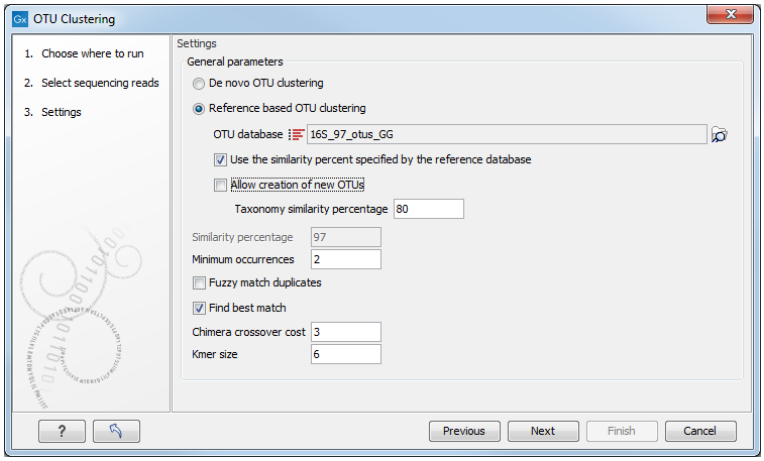

Figure 6: *Settings of the OTU clustering tools.*

4. Select the option to save the results and provide a new folder to save them to (here **OTU and statistics**) before clicking on 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 **Toolbox** | **Microbial Genomics Module ( )** | **General Tools ( )** | **Add Metadata to Abundance Table ( )** and choose the OTU (Table) as input.

2. Select the file describing the metadata on your local computer (MetatdataRoundRobin.csv) as shown in figure 7 and click Next.

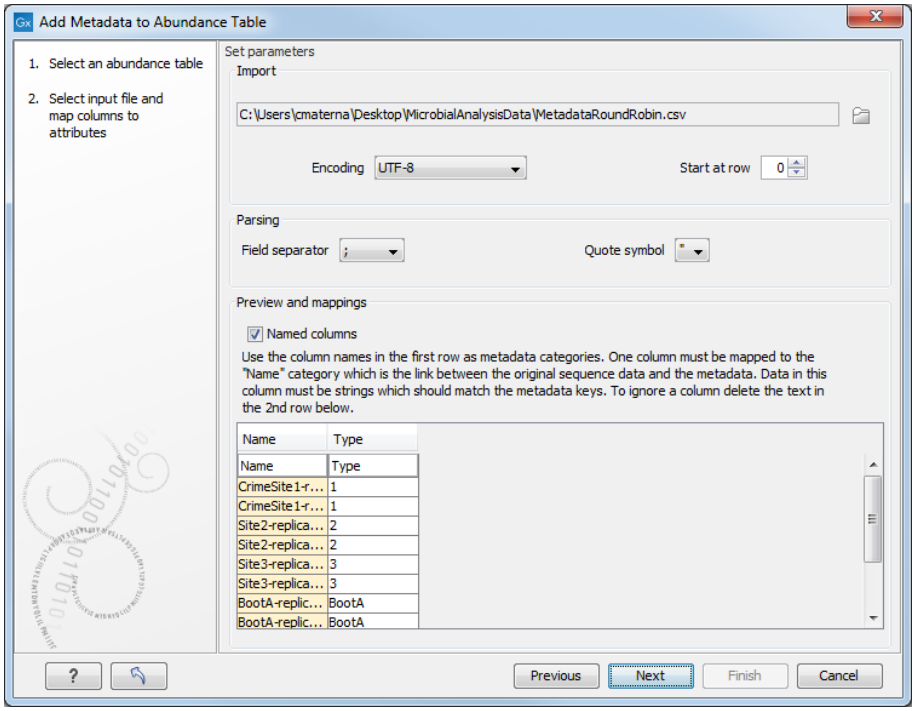

Figure 7: *Setting up metadata parameters.*

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**.

3. Save the table. This will overwrite the previous OTU (Table) such that the tables and graphs remain the same as before, but in the right hand settings area, the option in the Data tab allowing you to **Aggregate samples** will now include the metadata attributes. In this case, we can aggregate the data based on Type (seen as a Stacked Area Charts ( ) in figure 8). 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.

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 **Toolbox** | **Microbial Genomics Module ( )** | **OTU clustering ( )** | **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).

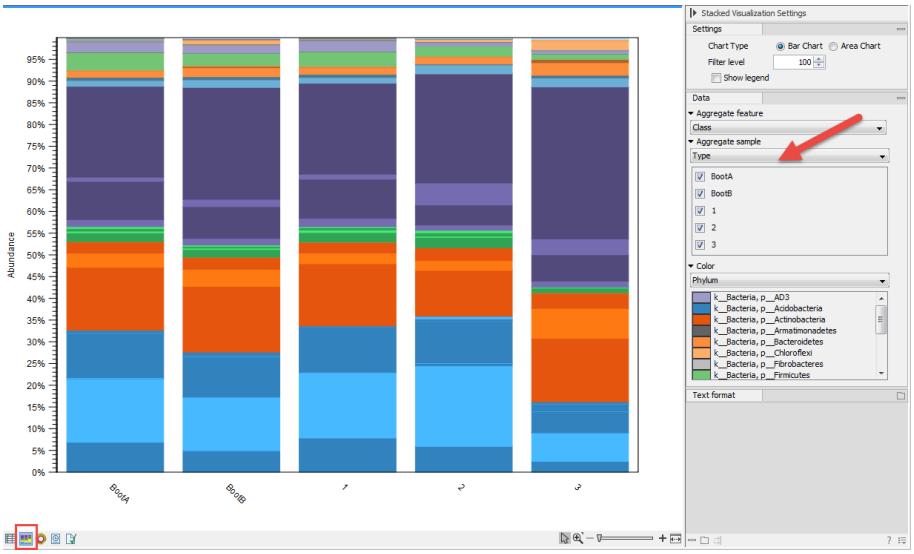

Figure 8: *Aggregate samples based on metadata information.*

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 BootA profile found on the suspects 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 above 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 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 **Toolbox** | **Microbial Genomics Module ( )** | **OTU clustering ( )** | **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 **Toolbox** | **Microbial Genomics Module ( )** | **OTU clustering ( )** | **Alpha Diversity ( )** and choose the abundance table **OTU (Table)** as input.
11. Click Next and select **Number of OTUs** and **Chao 1 bias-corrected** measures. In addition, specify the phylogenetic tree reconstructed in the previous step **OTU (Table) alignment_tree** and select **Phylogenetic diversity**. In the parameters box, change the **Maximum depth to sample** to 5,000 (figure 9).

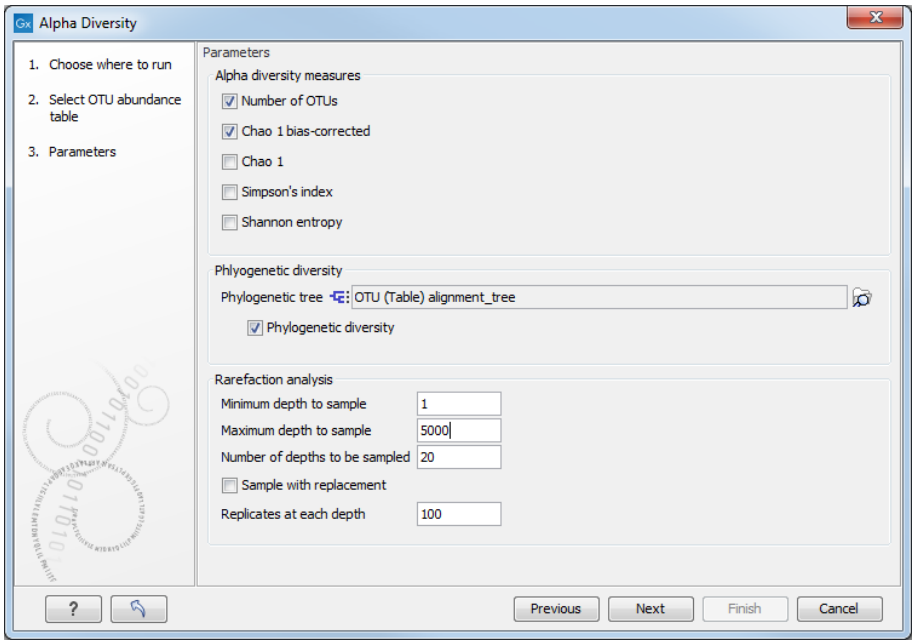

Figure 9: *Parameters for alpha diversity analysis.*

12. Click Next and 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 three).

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 of a metadata layer, or of all samples at once. In the following graph (figure 10) we have chosen blue lines for BootA and green lines for BootB. The lines do not plateau, indicating that we would need more samples to reach a definite conclusion, but BootA samples seem to have similar measures of alpha diversity as the sites 1, 2, and 3 while BootB samples look closer to site 3 when it comes to alpha diversity measures.

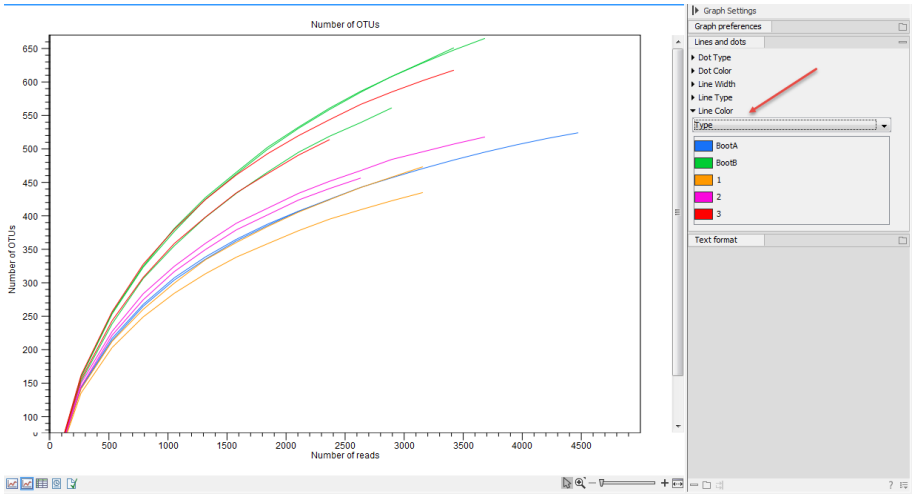

Figure 10: *Results of the alpha diversity analysis measured using Number of OTUs as parameters.*

Estimation of Beta Diversity

Beta diversity estimates differences in species diversity between samples.

1. Select **Toolbox** | **Microbial Genomics Module ( )** | **OTU clustering ( )** | **Beta Diversity ( )**
2. Choose the OTU (Table) as input.
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 11) and deselect all other distance measures.

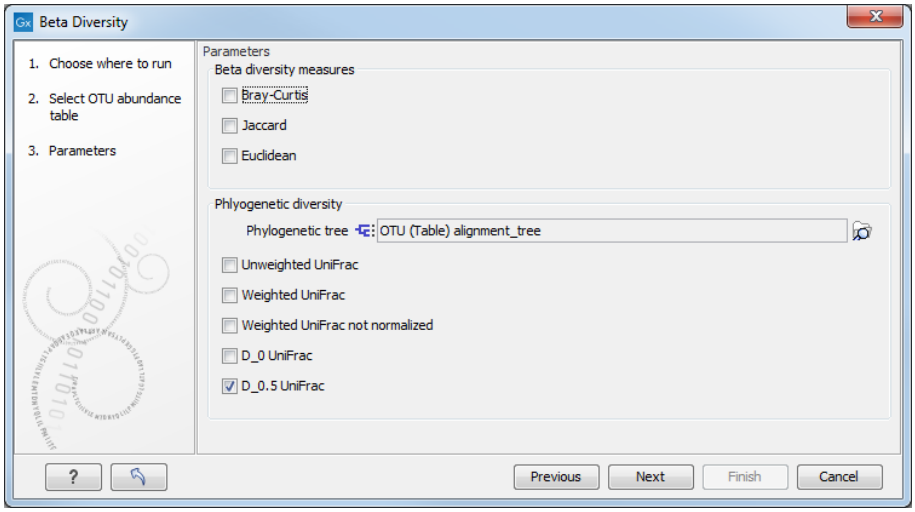

Figure 11: *Setting up parameters for the beta diversity analysis.*

4. Click Next and 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 12).

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 BootA are clearly clustered together, supporting our suspicion that the soil on the Boots A is the same as the one found on Site 1.

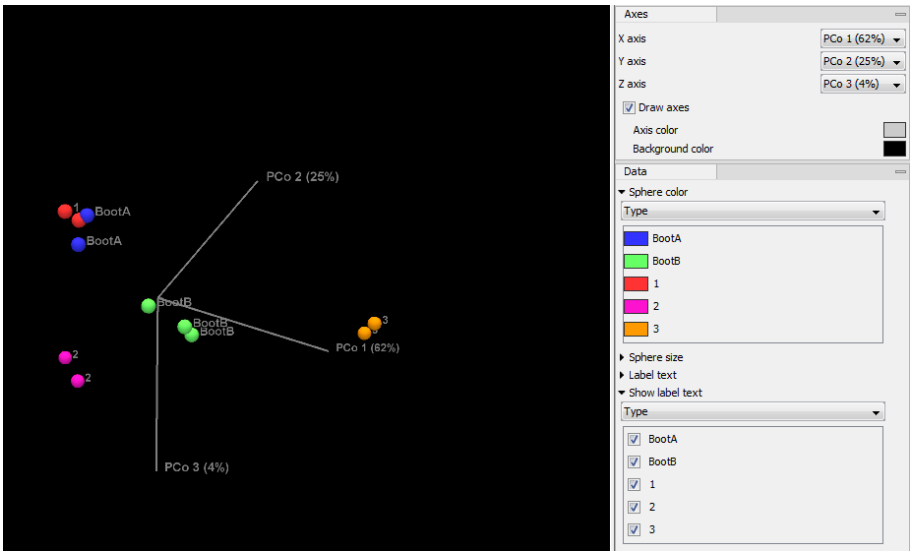

Figure 12: *Result of the beta diversity analysis.*

Additional statistical analyses

As a tool for assessing similarity between samples, a heat map and dendrogram can be helpful.

1. Open the **Toolbox** | **Microbial Genomics Module ( )** | **General Tools ( )** | **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, do not set any particular filter and 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 13).

Set the visualisation parameters like in the side panel in (figure 13). 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 **Toolbox** | **Microbial Genomics Module ( )** | **OTU clustering ( )** | **PERMANOVA ( )**.

2. Choose OTU (Table) from the **Data QC and OTU clustering** folder as input and select Type as Metadata group.

3. Specify the phylogenetic tree (OTU (Table) alignment_tree) from the **Estimate Alpha and Beta Diversities** folder. Select D_0.5 UniFrac and deselect all other distance measures. Leave the number of permutations to 99,999. Click on the button labeled Next.

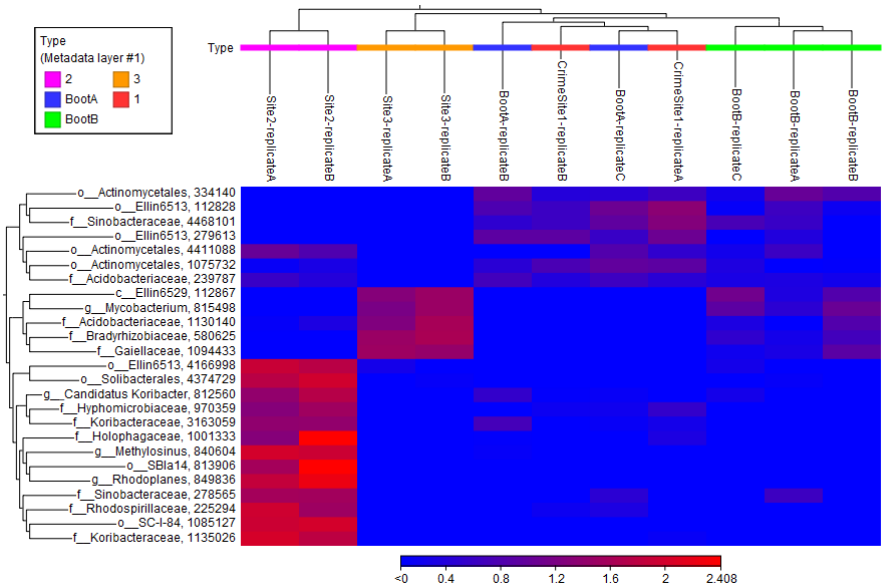

Figure 13: *Heat map from the abundance table.*

4. Save the report in the **Statistics** folder.

The result of the PERMANOVA analysis is a table (figure 14).

1 PERMANOVA analysis (D_0.5 UniFrac)

Variable	Groups	Pseudo-f statistic	p-value
Type	GTA, GTB, 1, 2, 3	13.12014	0.00003

Group 1	Group 2	Pseudo-f statistic	p-value	p-value (Bonferroni)
GTA	GTB	4.37869	0.10000	1.00000
GTA	1	1.19763	0.33333	1.00000
GTB	1	6.07694	0.10000	1.00000
GTA	2	10.37209	0.33333	1.00000
GTB	2	7.68865	0.10000	1.00000
1	2	13.86053	0.33333	1.00000
GTA	3	23.40413	0.33333	1.00000
GTB	3	15.87126	0.10000	1.00000
1	3	29.14107	0.33333	1.00000
2	3	18.30266	0.33333	1.00000

Figure 14: *Result of the PERMANOVA analysis.*

The PERMANOVA confirms that the clusters are significant (p=0,00003), 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 Boots A and from site 1 - to transform this analysis into actual forensic evidence!