Tutorial

OTU Clustering Using Workflows

November 21, 2017

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · www.qiagenbioinformatics.com ·
AdvancedGenomicsSupport@qiagen.com

OTU Clustering Using Workflows

This tutorial provides a quick-guide through the different workflows and tools available in CLC Microbial Genomics Module.

CLC Microbial Genomics Module assigns taxonomy to the reads from different samples by clustering them with representative sequences of pseudo-species called Operational Taxonomical Units (OTUs), and compute the abundance of each OTU. Secondary analyses will further describe microbial communities by estimating alpha and beta diversities in the context of sample metadata.

Introduction As an example for the data analysis, we will assume here that Mr. X is a suspect in a robbery at site 1. He claims his innocence by saying he has never been at site 1 but that he spent the entire weekend at sites 2 and 3. Investigators found two pairs of boots in Mr. X's house. Both were dirty with soil on the soles. The investigators obtained 3 samples of soil from each pair of boots, and 2 samples of soil from each of the 3 sites: the crime scene (site 1) and the 2 sites Mr. X claimed he was at (sites 2 and 3).

Each soil sample is characterized by a specific microbial community. In order to identify species present in the samples, DNA is extracted from its microbial community. Subsequently a region of the 16S gene is PCR amplified, and the resulting amplicon is sequenced using an NGS machine. The question we are going to adress here is how likely the samples from Mr X's boots did originate from the crime scene versus the 2 sites Mr. X claims to have been at.

Prerequisites This tutorial was done using CLC Genomics Workbench (Version 11 or higher), or Biomedical Genomics Workbench (Version 5 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://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Installing_plugins.html, as well as in the module manual.

Downloading the dataset In this analysis, we will be using a data-set containing sequences and metadata from a round-robin trial of several soil types generated in a mock crime-investigation as part of the EU FP7 MiSAFE project (see the project webpage under http://forensicmisafe.wix.com/misafe for further details). DNA was extracted, and a region of the 16S gene was PCR amplified using standard primers. The resulting amplicon was sequenced on an Illumina MiSeq machine (300 cycles, forward and reverse).

Importing the example data

1. Download the sample data from our website: http://resources.qiagenbioinformatics.com/testdata/otuclustering_tutorial/otuclustering_tutorial.zip and unzip it. As a result, you should see a directory called "MicrobialAnalysisData" containing the following:

    - **Sequence data**: 12 data sets (two each for soil from locations 1, 2 and 3, and three each for soil on the suspects boots A and B). The data was generated from the same MiSeq run and is composed of demultiplexed .fastq files. For the sake of speed, the original files have been down sampled to only contain 1/10th of the reads.

- **Metadata**: the spreadsheet MetadataRoundRobin.csv contains metadata information.
- **Primer sequences**: 16s_primers_round_robin.clc for the 16S primers.
- **Database**: 16S_97_otus_GG.clc contains a database Operational Taxonomic Units (OTUs) to be used in the analysis.

2. Start your CLC Workbench and go to **File** | **Import** ( ) | **Illumina** ( ) to import the 24 sequence files (ending with "fastq") (figure 1).

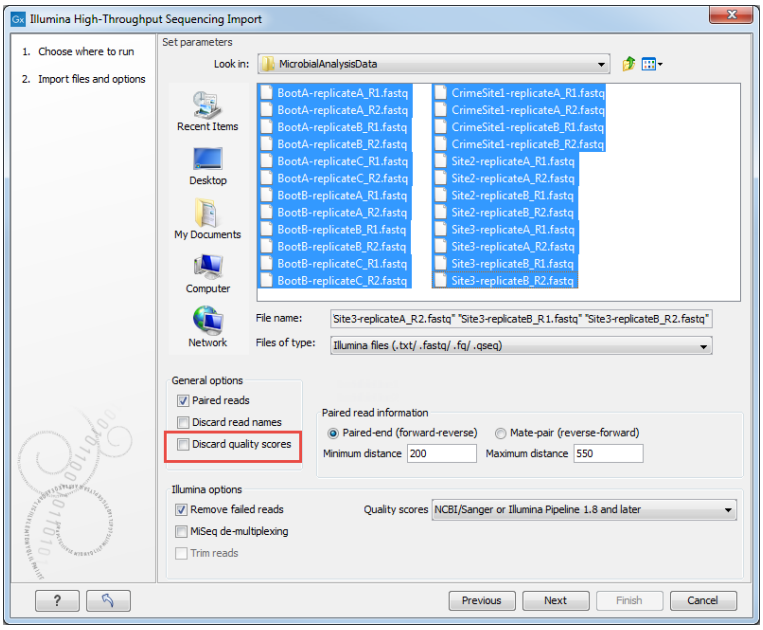

Figure 1: *Import the data from the samples collected on the sites and on the boots of the suspect.*

- 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.
- 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 using the **Import** | **Standard Import** button on top of the Navigation Area.

4. Import the metadata using the **Import** | **Import Metadata...** button.

- First select the MetadataRoundRobin.xls file. The contents of the Excel spreadsheet populates the table situated at the bottom of the dialog (figure 2). Click **Next**.
- Then select the 12 samples the metadata should be associated with: Click on the Navigation button next to "Location of data", and select the imported reads in your Navigation Area. Click OK.

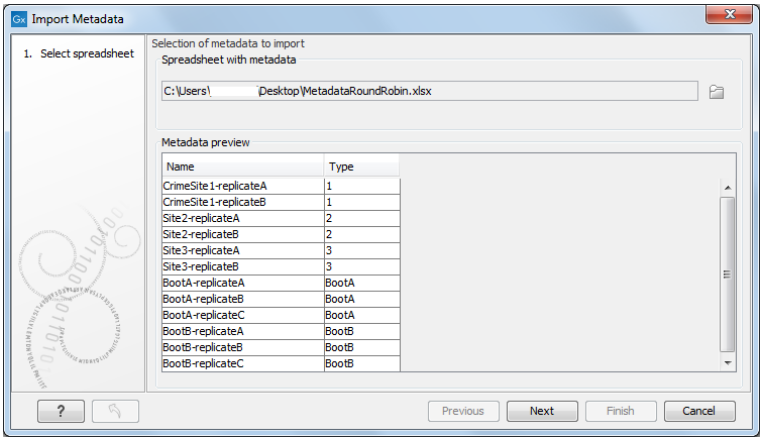

Figure 2: *Import the metadata spreadsheet.*

- In the "Data association preview", now you can see that the data association is not successful. It is because the option "Exact" matching of the names is checked by default. Choose instead "**Partial**" name matching. This results in successful associations for all data items (figure 3). Click **Next**.

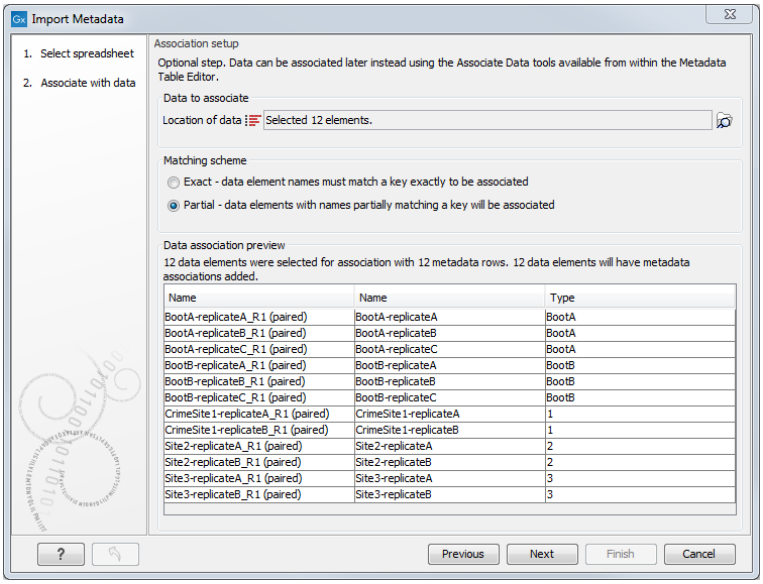

Figure 3: *Associate the metadata to the paired reads.*

- Select the folder you created for this tutorial to save the resulting metadata table called "Samples".

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

Running the workflows

The Data QC and OTU Clustering workflow consists of three steps that are executed sequentially (see a display of the workflow in figure 4). The inputs necessary to run the workflow are the reads

you want to cluster. You can also specify a list of the primers that were used to sequence these reads.

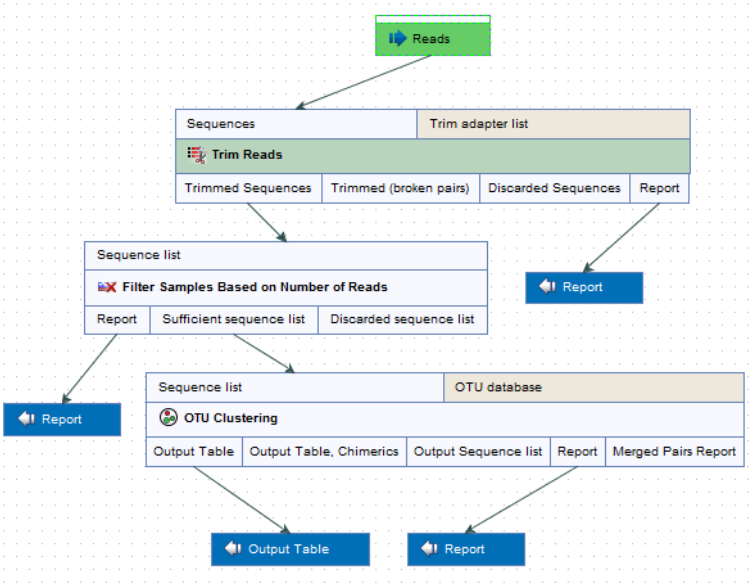

Figure 4: *Layout of the Data QC and OTU clustering workflow.*

1. Launch the workflow **Microbial Genomics Module** ( ) | **Metagenomics** ( ) | **Amplicon-based OTU clustering** ( ) | **Workflows** | **Data QC and OTU clustering**.
2. Select the 12 sequence files from your folder called "Illumina reads" and click **Next**. Make sure the "Batch" function is not checked.
3. In the **Trim Reads** window, select the list of primer sequences **16s_primers_round_robin.clc**. Leave the remaining parameters as default and click **Next**.
4. In the **OTU clustering** window, choose from the drop-down menu **Reference based OTU clustering** and select the file called **16S_97_otus_GG**. Uncheck the option **Allow creation of new OTUs** and click **Next**.
5. Choose to save your workflow outputs and click on the button labeled **Finish**. You can create a new folder in which you can save your results (here called "Data QC and OTU clustering").

You can follow the progress of the workflow in the Processes tab below the toolbox. When the workflow is done, you will see the output files as shown in figure 5.

The file OTU (Table) is the result you will use as input for the Estimate Alpha and Beta Diversities workflow, which consists of 5 tools as seen on figure 6. These tools make use of the metadata imported earlier.

Note: In cases where the metadata was not imported and associated with the reads before running the Data QC and OTU Clustering workflow, it is still possible to do it now using the Add Metadata to Abundance Table tool as described in the tutorial "OTU Clustering Step by Step".

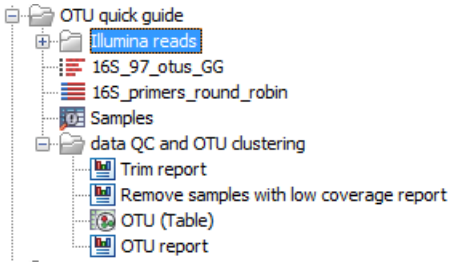

Figure 5: *Outputs of the Data QC and OTU clustering workflow.*

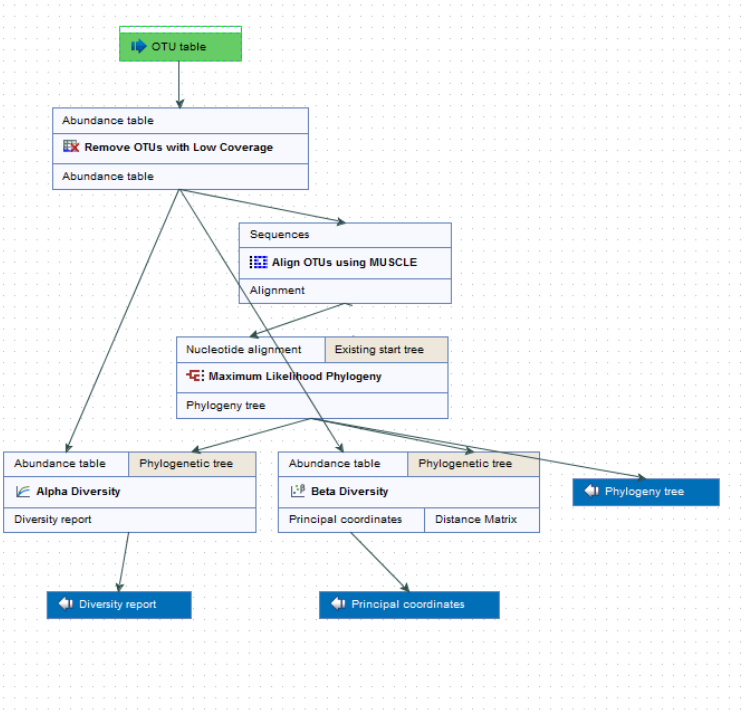

Figure 6: *Layout of the Alpha and Beta Diversities workflow.*

1. Launch the workflow **Microbial Genomics Module ( )** | **Metagenomics ( )** | **Amplicon-based OTU clustering ( )** | **Workflows** | | **Estimate Alpha and Beta Diversities**
2. In the first dialog, select the OTU (Table) and click **Next**.
3. In the **Alpha analysis** window, deselect everything except **Total number**.
4. In the **Beta analysis** window, deselect everything except **D_0.5 UniFrac**.
5. Choose to save your workflow outputs. You can create a new folder in which you can save your results (here called "Estimate Alpha and Beta Diversities"). Click **Finish**.

Running this workflow will give at least 3 outputs (figure 7): a phylogenetic tree of the OTUs, a diversity report for the alpha diversity and a Principal Coordinate Analysis (PCoA) chart for the beta diversity.

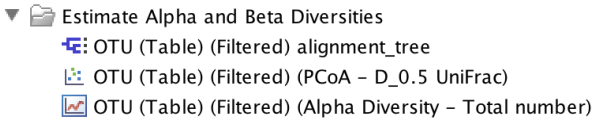

Figure 7: *Outputs of the Alpha and Beta Diversities workflow.*

Results

The primary output of your analysis is the OTU abundance table annotated with metadata. In this investigation, the metadata defines the origin of the different soil samples and allows the aggregation of the results to improve visualization of the results. In addition, the module offers several ways to look at your newly generated OTU clusters: the table itself, but also Stacked Bar Charts and Stacked Area Charts ( ) as well as Zoomable Sunbursts ( ).

To simplify the visualization of the OTU clustering results, you can filter out low abundance OTUs from the OTU table.

1. Select **Microbial Genomics Module ( )** | **Metagenomics** ( ) | **Amplicon-based OTU clustering ( )** | **Remove OTUs with Low Abundance ( )**
2. Choose the OTU(Table) as input.
3. Leave the parameters as default, i.e., the "Minimum combined abundance threshold for removal of OTUs is set to 10.
4. Save your result in the "Data QC and OTU clustering" folder and click **Finish**.

The new table will be labeled as (Filtered). Open it and click on the Stacked Bar Chart icon ( ) in the lower part of the workbench. In the right side panel, choose to **aggregate samples by Type** (figure 8). We observe a striking similarity between the Boot A profile found on the suspect's boots and the profile of the soil from Site 1, indicating that Mr. X was most likely lying when he said he had never been at Site 1.

Now open the results of the alpha diversity analysis, called OTU (Table) (Filtered) (Alpha diversity - Total number): the plot contains the rarefaction results of the specified alpha diversity measure while each line corresponds to a 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 9) we have chosen red lines for BootA and pink 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 and 2 while BootB samples appear to be more distinct from the other sites when it comes to alpha diversity measures. We suspect that Mr. X was not wearing his boot B at any of the sites sampled.

Finally, beta diversity estimates differences in species diversity between samples. The beta diversity analysis tool performs a Principal Coordinate Analysis (PCoA) using the UniFrac distances (figure 10).

In the PCoA of the beta diversities, the soil samples cluster according to their origin. In this case

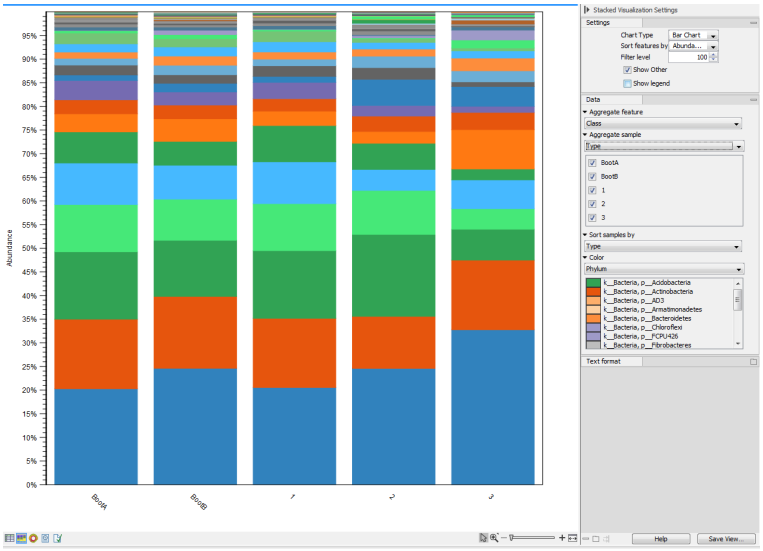

Figure 8: *Aggregate samples based on metadata information.*

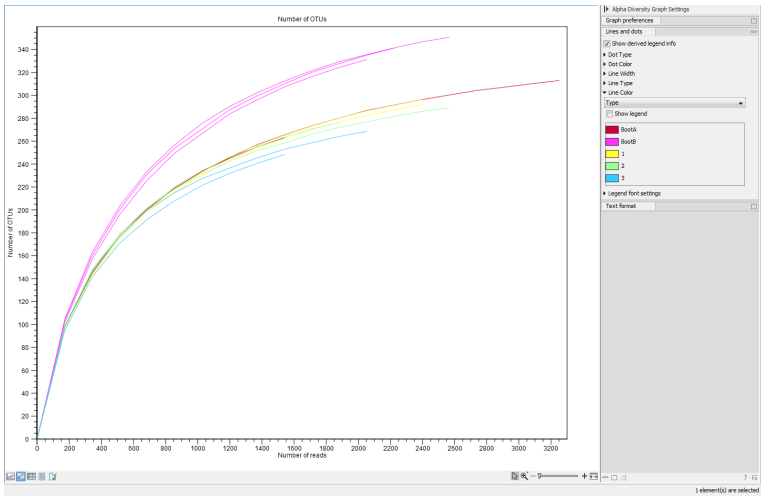

Figure 9: *Results of the alpha diversity analysis measured using Total number as parameter.*

all samples from Site 1 and Boot A cluster together, confirming a similarity between the 2 soils and thus confirming our suspicion that Mr. X was on site 1 with his boot A.

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 **Microbial Genomics Module | Metagenomics | Abundance Analysis | Differential Abundance Analysis**.
2. Choose OTU (Table) as input.
3. Choose **Type** as Metadata factor and **Across all samples (ANOVA-like)** as Comparisons (see figure 11). Click **Next**.

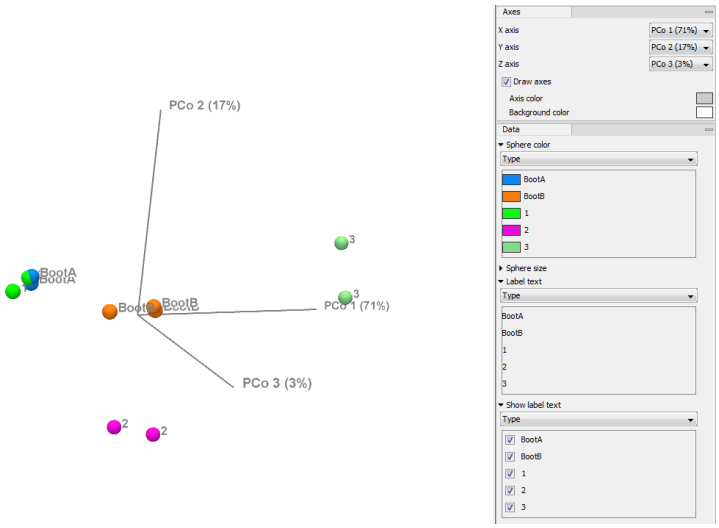

Figure 10: *Result of the beta diversity analysis.*

4. Save the result in the tutorial folder.

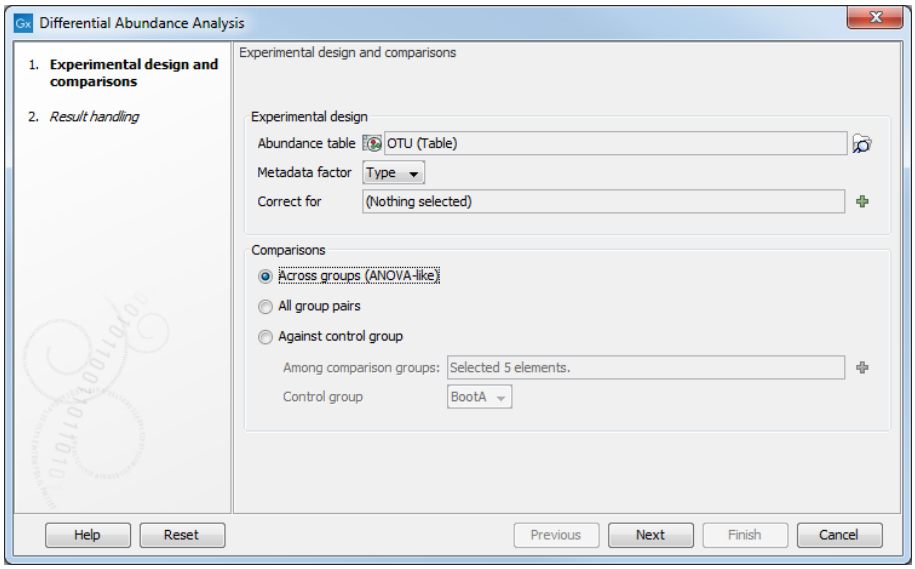

Figure 11: *Set up a differential abundance analysis.*

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 and press 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 **Microbial Genomics Module ( )** | **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 12). Click **Next**.

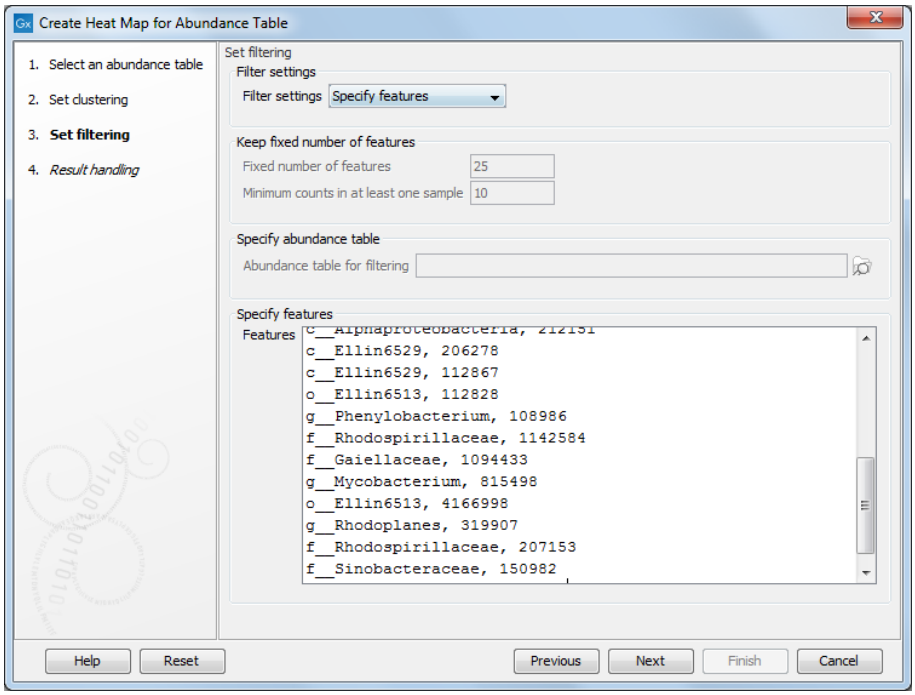

Figure 12: *Filter based on the 25 most different OTU.*

4. **Save** the result in the Navigation Area.

Display the heat map by double-clicking on it in the Navigation Area (figure 13).

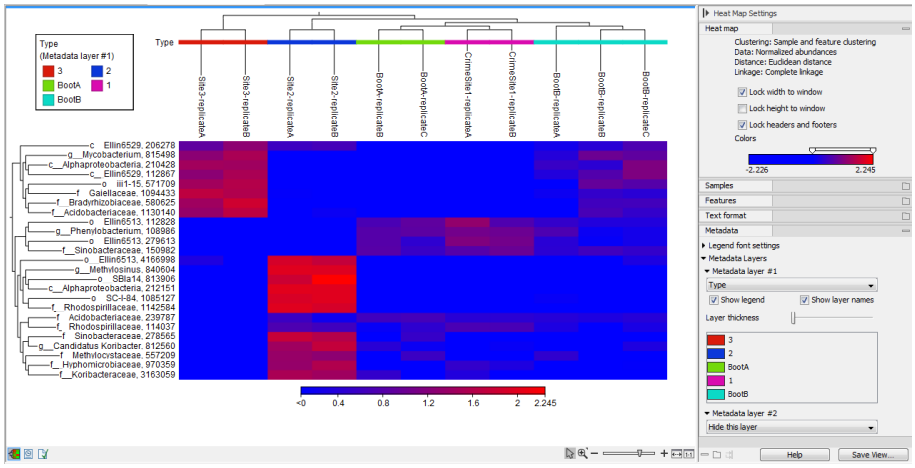

Figure 13: *Heat map from the abundance table.*

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 **Microbial Genomics Module ( )** | **Metagenomics ( )** | **Abundance Analysis ( )** | **PERMANOVA Analysis ( )**.

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 **Next**.
4. Choose to **Open** the report.

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	BootA, BootB, 1, 2, 3	15.74532	0.00006

Group 1	Group 2	Pseudo-f statistic	p-value	p-value (Bonferroni)
BootA	BootB	4.80698	0.10000	1.00000
BootA	1	1.12252	0.33333	1.00000
BootB	1	6.58892	0.10000	1.00000
BootA	2	13.54072	0.33333	1.00000
BootB	2	7.71485	0.10000	1.00000
1	2	16.86622	0.33333	1.00000
BootA	3	29.39652	0.33333	1.00000
BootB	3	18.90927	0.10000	1.00000
1	3	35.30066	0.33333	1.00000
2	3	18.73599	0.33333	1.00000

Figure 14: *Result of the PERMANOVA analysis.*

The PERMANOVA confirms that the clusters are significant (p=0,00006), 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 will need more samples - in particular from the soles of the Boots A and from site 1 - to transform this analysis into actual evidence!