Tutorial

Microbiome Profiling using CLC Microbial Genomics Module workflows

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

Microbiome Profiling using CLC Microbial Genomics Module 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 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.

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).

All the files required throughout this tutorial have been packaged into a single zip-archive and made available via our website. Download the tutorial data from the following address: http://download.clcbio.com/testdata/MicrobialAnalysisData.zip. Once the download is completed, you can move the .zip file into an easily accessible location in your file-system (such as the Desktop for example) and unzip it. As a result, you should see a directory called "MicrobialAnalysisData" containing 27 files.

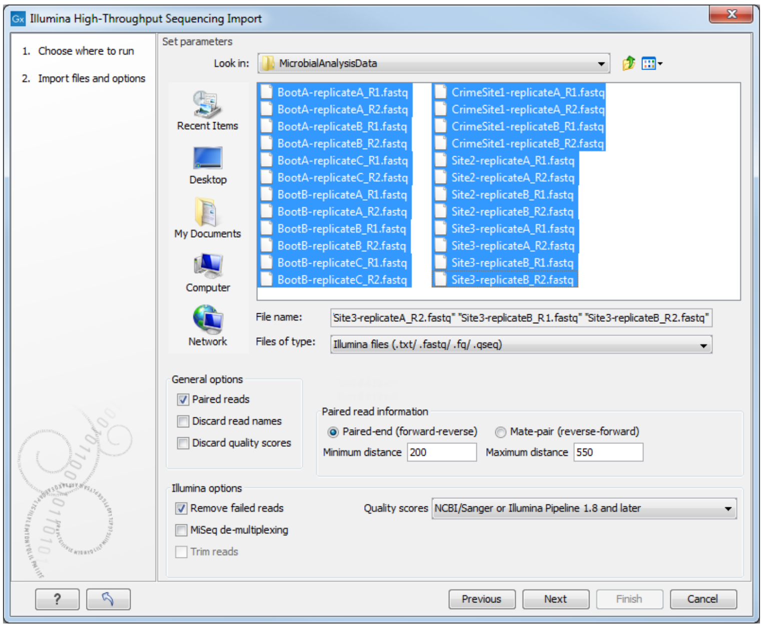
- **Sequence data**: 12 data sets (two each for soil from locations 1, 2 and 3, and three each for soil on the suspects boots A and B). The data was generated from the same MiSeq run and is composed of demultiplexed .fastq files. For the sake of speed, the original files have been down sampled to only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

Importing the example data

Now we have the data available locally, we can import them into the Workbench.

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).



Illumina High-Throughput Sequencing Import

1. Choose where to run
2. Import files and options

Set parameters

Look in: MicrobialAnalysisData

BootA-replicateA_R1.fastq
BootA-replicateA_R2.fastq
BootA-replicateB_R1.fastq
BootA-replicateB_R2.fastq
BootA-replicateC_R1.fastq
BootA-replicateC_R2.fastq
BootB-replicateA_R1.fastq
BootB-replicateA_R2.fastq
BootB-replicateB_R1.fastq
BootB-replicateB_R2.fastq
BootB-replicateC_R1.fastq
BootB-replicateC_R2.fastq
CrimeSite1-replicateA_R1.fastq
CrimeSite1-replicateA_R2.fastq
CrimeSite1-replicateB_R1.fastq
CrimeSite1-replicateB_R2.fastq
Site2-replicateA_R1.fastq
Site2-replicateA_R2.fastq
Site2-replicateB_R1.fastq
Site2-replicateB_R2.fastq
Site3-replicateA_R1.fastq
Site3-replicateA_R2.fastq
Site3-replicateB_R1.fastq
Site3-replicateB_R2.fastq

File name: Site3-replicateA_R2.fastq" "Site3-replicateB_R1.fastq" "Site3-replicateB_R2.fastq"
Files of type: Illumina files (.txt/ .fastq/ .fq/ .qseq)

General options
☑ Paired reads
□ Discard read names
□ Discard quality scores

Paired read information
● Paired-end (forward-reverse) ○ Mate-pair (reverse-forward)
Minimum distance 200 Maximum distance 550

Illumina options
☑ Remove failed reads
□ MiSeq de-multiplexing
□ Trim reads
Quality scores NCBI/Sanger or Illumina Pipeline 1.8 and later

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 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.

Running the workflows

The Data QC and OTU Clustering workflow consists of 5 steps that are executed sequentially (see a display of the workflow in figure 2). 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.

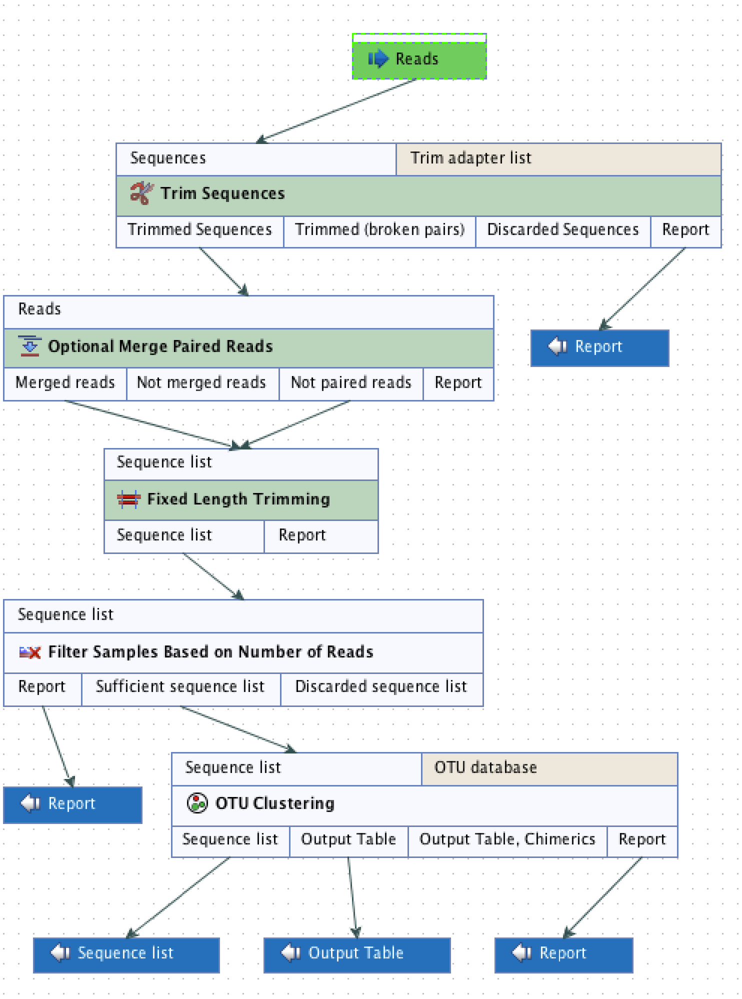
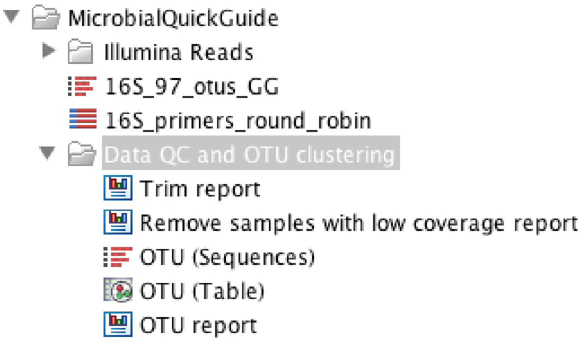


Figure 2: *Layout of the Data QC and OTU clustering workflow.*

1. Launch the workflow **Toolbox** | **OTU clustering** (🔘) | **Workflows** | **Data QC and OTU clustering**.
2. Select the 12 sequence files from your folder called **Illumina reads** and click Next.
3. In the **Trim Sequences** window, select the list of primer sequences 16s_primers_round_robin.clc. Leave the remaining parameters as default and click on the button labeled Next.
4. In the **Optional Merge Paired Reads** window, make sure that the parameters are set to **Mismatch cost** 1, **Minimum score** 40, **Gap cost** 4 and **Maximum unaligned end mismatches** to 5 and click Next.

5. In the **Fixed Length Trimming** window, leave the option "Automatic read length" checked as we want the length of trimming to be automatically detected by the software.
6. In the **OTU clustering** window, choose from the drop-down menu **Reference based OTU clustering** and select the file called 16S_97_otus_GG. Click on the button labeled Next.
7. 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 3.



- MicrobialQuickGuide
  - Illumina Reads
  - 16S_97_otus_GG
  - 16S_primers_round_robin
  - Data QC and OTU clustering
    - Trim report
    - Remove samples with low coverage report
    - OTU (Sequences)
    - OTU (Table)
    - OTU report

Figure 3: *Outputs of the Data QC and OTU clustering workflow.*

The file OTU (Table) is the result you will use as input for the Estimate Alpha and Beta Diversities workflow. But as these secondary statistical analyses require metadata, you need first to use the **Add Metadata to Abundance Table** tool. Only after the OTU (Table) is decorated with the metadata can you run the Estimate Alpha and Beta Diversities workflow, which consists of 5 tools as seen on figure 4.

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 and click on the button labeled Next.
3. Save your result. It is a table that will overwrite the previous OTU (Table) file. The tables are similar to each other, but you now have the option to **Aggregate samples** based on the headers of the columns of your metadata file. Note: if you had previously opened the OTU (Table), close it and reopen it to be able to see the aggregate option on the right side panel of the workbench.
4. Launch **Toolbox** | **OTU clustering (** **)** | **Workflows** | **Estimate Alpha and Beta Diversities** and select the OTU (Table). Click on the button labeled Next.
5. In the **Alpha analysis** window, deselect everything except **Number of OTUs**.
6. In the **Beta analysis** window, deselect everything except **D_0.5 UniFrac**.
7. 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 on the button labeled Finish.

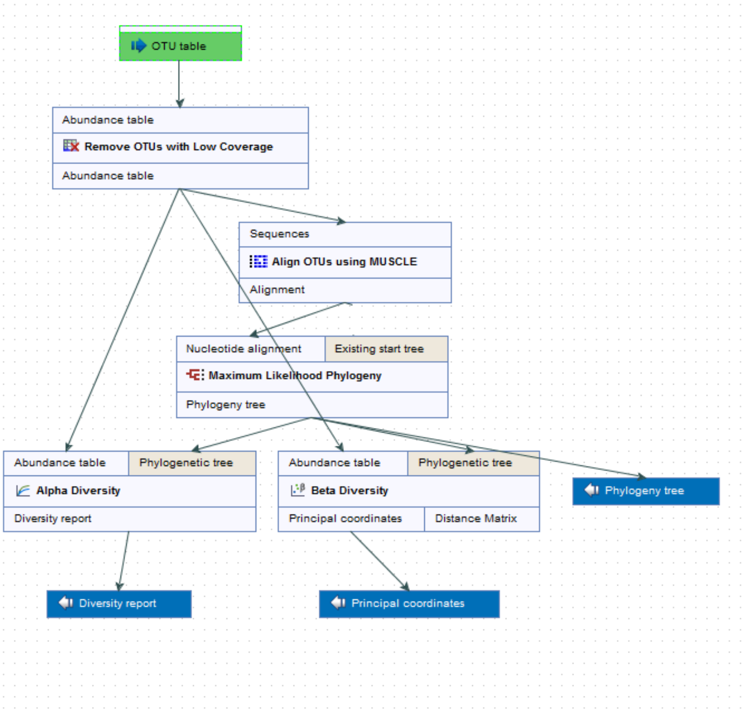
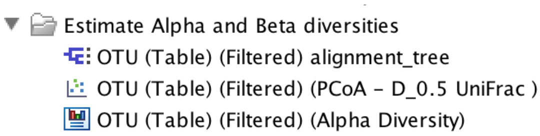


Figure 4: *Layout of the Alpha and Beta Diversities workflow.*

Running this workflow will give at least 3 outputs (figure 5): a phylogenetic tree of the OTUs, a diversity report for the alpha diversity and a Principal Coordinate Analysis (PCoA) chart for the beta diversity.



Estimate Alpha and Beta diversities
- OTU (Table) (Filtered) alignment_tree
- OTU (Table) (Filtered) (PCoA – D_0.5 UniFrac)
- OTU (Table) (Filtered) (Alpha Diversity)

Figure 5: *Outputs of the Alpha and Beta Diversities workflow.*

Results

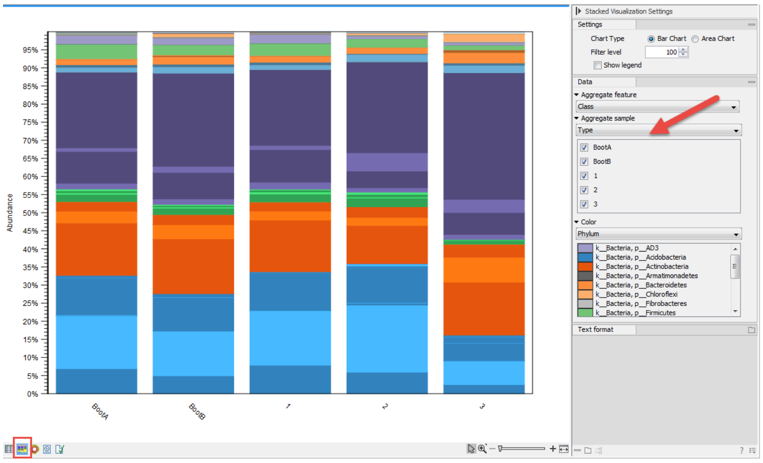
The primary output of your analysis is the OTU abundance table annotated with metadata. In this investigation, the metadata defines the origin of the different soil samples and allows the aggregation of the results to improve visualization of the results. In addition, the module offers several ways to 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 **Toolbox** | **Microbial Genomics Module (** **)** | **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 6). 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.



Stacked Visualization Settings

Settings
Chart Type ● Bar Chart ○ Area Chart
Filter level 100
□ Show legend

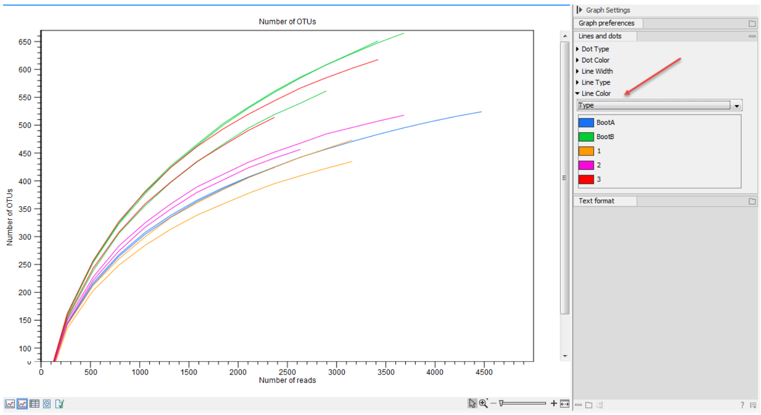
Data
Aggregate feature
Class
Aggregate sample
Type
☑ BootA
☑ BootB
☑ 1
☑ 2
☑ 3
Color
Phylum
k__Bacteria, p__AD3
k__Bacteria, p__Acidobacteria
k__Bacteria, p__Actinobacteria
k__Bacteria, p__Armatimonadetes
k__Bacteria, p__Bacteroidetes
k__Bacteria, p__Chloroflexi
k__Bacteria, p__Fibrobacteres
k__Bacteria, p__Firmicutes

Text format

Figure 6: *Aggregate samples based on metadata information.*

Now open the results of the alpha diversity analysis, called OTU (Table) (Filtered) (Alpha diversity): 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 7) 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 the Boot A samples seem to have similar measures of alpha diversity as the sites 1, 2, and 3 while the Boot B samples are clearly apart. We suspect that Mr. X was not wearing his boot B at any of the sites sampled.



Graph Settings

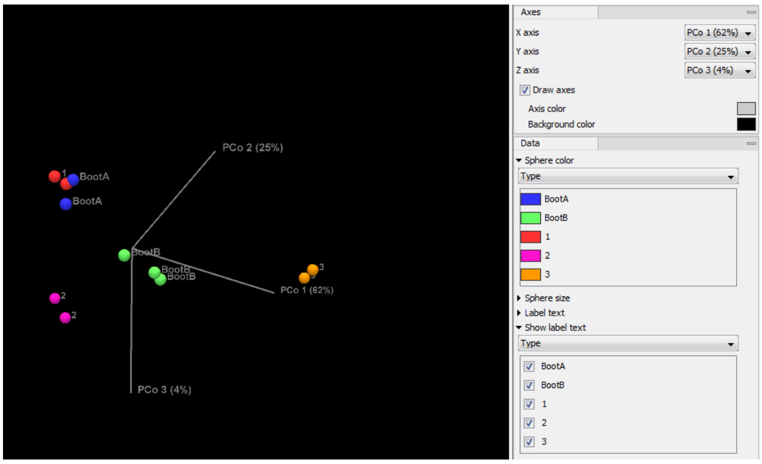
Graph preferences

Lines and dots
Dot Type
Dot Color
Line Width
Line Type
Line Color
Type
🟦 BootA
🟩 BootB
🟧 1
🟪 2
🟥 3

Text format

Figure 7: *Results of the alpha diversity analysis measured using Number of OTUs as parameters.*

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 8).



X axis PCo 1 (62%)
Y axis PCo 2 (25%)
Z axis PCo 3 (4%)
☑ Draw axes
Axis color
Background color
☑ BootA
☑ BootB
☑ 1
☑ 2
☑ 3

Figure 8: *Result of the beta diversity analysis.*

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

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 9).

Set the visualisation parameters like in the side panel in (figure 9). 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 ( ).**

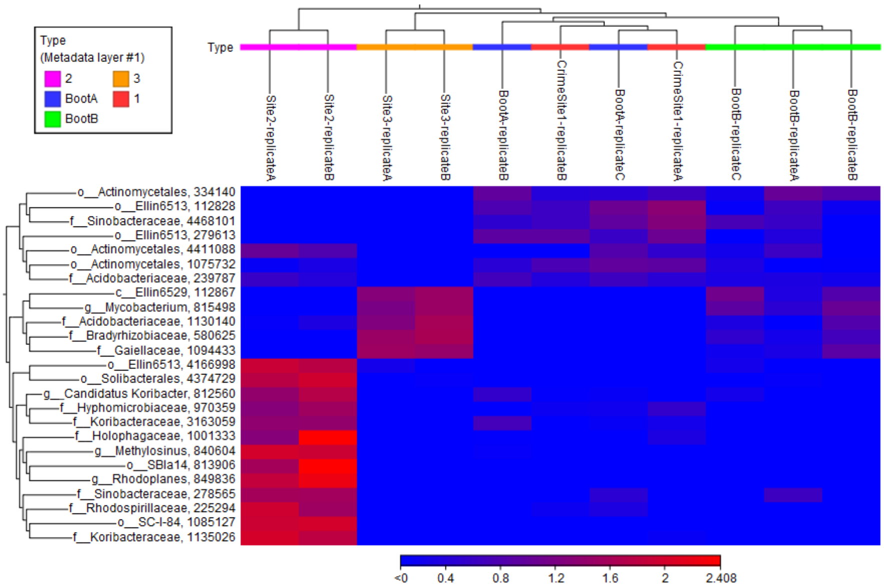


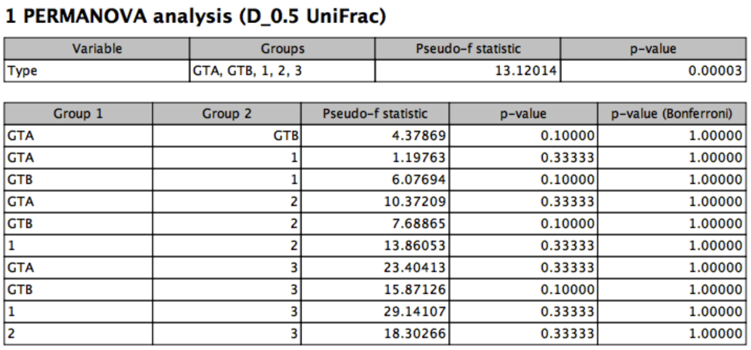
Figure 9: *Heat map from the abundance table.*

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.

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

The result of the PERMANOVA analysis is a table (figure 10).



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 10: *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 will need more samples - in particular from the soles of the Boots A and from site 1 - to transform this analysis into actual evidence!