This project is funded by the European Union

SEVENTH FRAMEWORK PROGRAMME

MiSAFE

The Development and Validation of Microbial Soil Community Analyses for Forensics Purposes

FP7 Theme 10: Security Call: FP7-SEC-2012-1

Work Programme: Topic SEC-2012.7.2-1. Open topic for Small and Medium Enterprises: "Advancing contemporary forensic methods and equipment"

Report D5.10
Documentation and instructional materials

Authors: Leif Schauser, CLC bio

Approved by Work Package Manager of WP5:

Date: 29th May 2015

Approved by Project Coordinator: **Edouard Jurkevitch, HUJI**

Date: 29th May 2015

This document was produced under the terms and conditions of Grant Agreement No. 313149 for the European Commission. It does not necessary reflect the view of the European Union and in no way anticipates the Commission's future policy in this area.

This page is left intentionally blank

Table of Contents

1. Objectives 4

2. Significant results: 4

Involved partners 01 HUJI, 03 CLCB, 04 ECL, 05 JHI

1. Objectives

The aim of this deliverable is to provide instructional material and thorough documentation of the software, now called the CLC Microbial Genomics Module (month 24).

2. Significant results:

The development of the software is now finished for release, and the software has been presented at The SFAF meeting in Santa Fe, 25th May 2015.

In the appendix, the functionality of the release of the software is presented in the form of a Quick start and the full documentation.

Tutorial

Tutorial: Analysis of Microbial Communities - Quick Guide

May 28, 2015

CLC bio, a QIAGEN Company
Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
Telephone: +45 70 22 32 44 · Fax: +45 86 20 12 22
www.clcbio.com · support@clcbio.com

Contents

Using Microbial Genomics Module for forensic investigation 3

Prerequisites 3

Results 6

Additional statistical analyses 8

Using Microbial Genomics Module for forensic investigation

Mr. X is a suspect in a murder. The body was found on site 1, but Mr. X claims he was never there because he spent his entire WE on site 2 and 3. Investigators have found at Mr. X's house a pair of rubber boot and a pair of hiking shoes, both dirty with soil on the soles. They took 3 samples of soil from each pair of shoes, and 2 samples of soil from each site: the crime scene (site 1) but also the 2 sites Mr. X claimed he was at (site 2 and 3).

This tutorial will guide you through the different tools you can use to transform NGS data into forensic evidence.

Indeed, each soil sample is characterized by a specific microbial community. To identify species present in soil samples, DNA is extracted from its microbial community, a region of the 16S gene is PCR amplified, and the resulting amplicon is sequenced using an NGS machine.

The Microbial Genomics Module will then assign taxonomy to the reads by clustering them with representative sequences of pseudo-species called Operational Taxonomical Units (OTUs), and tally the occurrences of each OTU. Secondary analyses will further describe microbial communities by estimating alpha and beta diversities in the context of sample metadata.

Prerequisites

For this tutorial, you must be working with the CLC Genomics Workbench version 8.0 or higher and you must have installed the Microbial Genomics Module. You 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**: 12 data sets (two each for soil from locations 1, 2 and 3, and three each for soil on the suspects rubber boots GT-A and hiking shoes GT-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, here only information about the origin of the samples.
- **Sequencing primer sequences**: 16s_primers_round_robin.clc for the 16S primers. Note that several databases are available for download in the CLC Genomics Workbench.
- **Database**: 16S_97_otus_GG.clc.n

Now that the prerequisites have been described, its time to start with the analyses of your samples.

1. Download the sample data from our website: ftp://LeifSchauser:ningohshi@upload.clcbio.com.

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

Now that all our data has been imported, you can start the workflows from Microbial Genomics Module.

The Data QC and OTU clustering workflow consists of 5 tools being executed sequentially (see a display of the workflow in figure 1). The input 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.

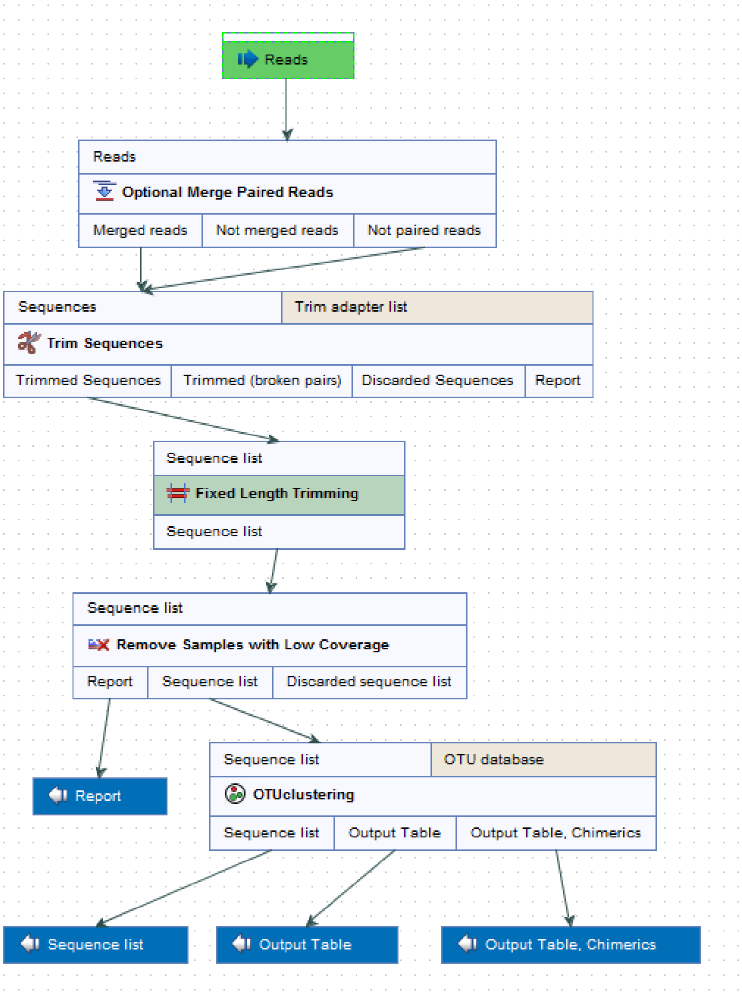


Figure 1: *Layout of the Data QC and OTU clustering workflow.*

1. Open the workflow **Toolbox** | **Workflows** | **Data QC and OTU clustering**.
2. Select the 12 sequence files form your folder called **Illumina reads**.
3. In the **Optional Merge Paired Reads** window, change the parameters to **Mismatch cost** 1, **Minimum score** 40, **Gap cost** 4 and **Maximum unaligned end mismatches** to 5 and click Next.
4. In the **Trim Sequences** window, select the list of primers sequences 16s_primers_round_robin.clc. Leave the other parameters as default and click on the button labeled Next.
5. In the **Fixed Length Trimming** window, leave the option as default as we want the length of trimming to be automatically detected by the software.
6. In the **OTUclustering** 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. Save your workflow outputs in a new folder called **Data QC and OTU clustering**. Click on the button labeled Finish.

You can follow the progress of the workflow in the Processes bar below the toolbox. When the workflow is done, you will have 4 output files in your folder (figure 2).

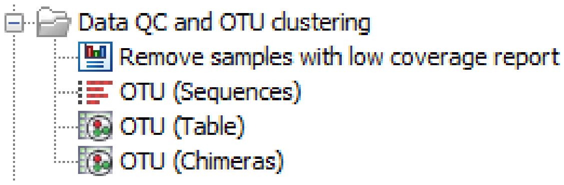

Data QC and OTU clustering
Remove samples with low coverage report
OTU (Sequences)
OTU (Table)
OTU (Chimeras)

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

The file OTU(Table) is the one you will use as input for the Estimate Alpha and Beta Diversities workflow, but as these secondary analyses require metadata you will first use the **Add Metadata to OTU Abundance Table** tool. You can then run the Estimate Alpha and Beta Diversities workflow which consists of 5 tools and your newly generated OTU(Table) as input file (figure 3).

1. Select **Toolbox** | **OTUclustering** ( ) | **Add Metadata to OTU 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 in the **Data QC and OTU clustering** folder. It is a table that will overwrite the previous OTU(Table) file. The tables are similar to each other, but you have now the option to **Aggregate samples** based on the headers of the columns of your metadata file. Note: if you had previously open the OTU(Table), close it and reopen it to be able to have the aggregate option on the right side panel of the workbench.
4. Open **Toolbox** | **Workflows** | **Estimate Alpha and Beta Diversities** and select the OTU (Table). Click on the button labeled Next.
5. In the **Alpha analysis** window, deselect Chaos 1 bias-corrected and Phylogenetic diversity, but keep **Number of OTUs** checked.

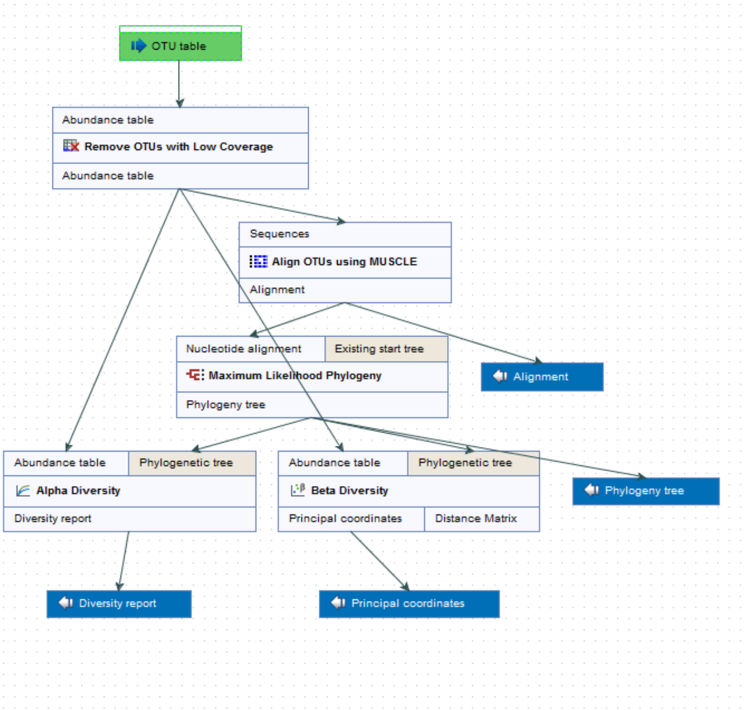
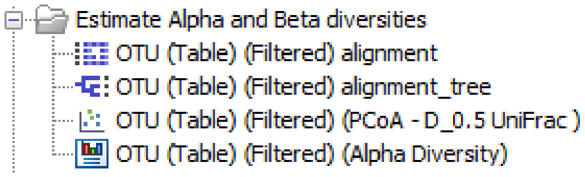


Figure 3: *Layout of the Alpha and Beta Diversities workflow.*

6. In the **Beta analysis** window select **D_0.5 UniFrac** but deselect Bray-Curtis and Jaccard measures, unweighted and weighted UniFrac.

7. Save the data in a new folder called **Estimate Alpha and Beta Diversities**.

Running this workflow will give at least 4 outputs (figure 4): an alignment of the OTUs, a phylogenetic tree of the OTUs, a diversity report for the alpha diversity and a PCoA for the beta diversity.



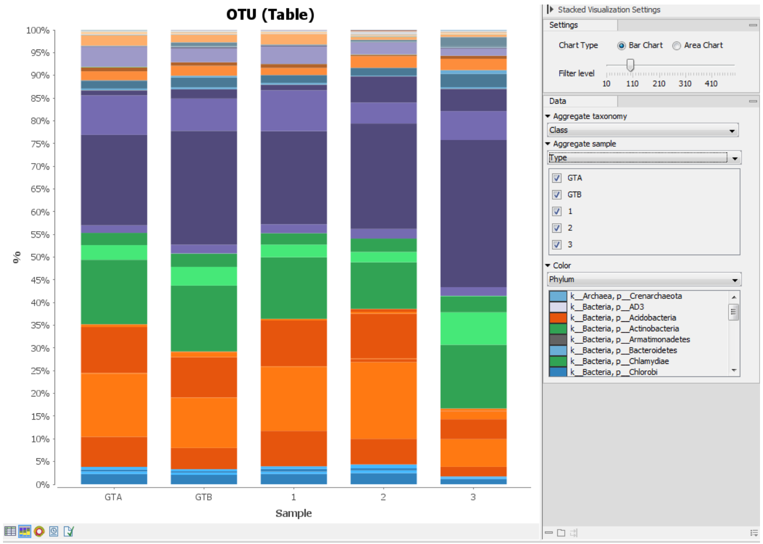
Estimate Alpha and Beta diversities
- OTU (Table) (Filtered) alignment
- OTU (Table) (Filtered) alignment_tree
- OTU (Table) (Filtered) (PCoA - D_0.5 UniFrac)
- OTU (Table) (Filtered) (Alpha Diversity)

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

Results

The primary output of your analysis is an OTU abundance table annotated with the 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 (◎).

Open the OTU(Table)file from your **Data QC and OTU clustering** folder and click on the Stacked Chart icon in the lower part of the workbench. In the right side panel, choose to aggregate samples by Type (figure 5). We observe a striking similarity between the GT-A profile found on the suspects rubber boots and the profile of the soil from Site 1, indicating that Mr. X was most likely lying when he pretended he had never been on Site 1.



Stacked Visualization Settings

Settings

Chart Type ● Bar Chart ○ Area Chart

Filter level 10 110 210 310 410

Data

Aggregate taxonomy: Class

Aggregate sample: Type

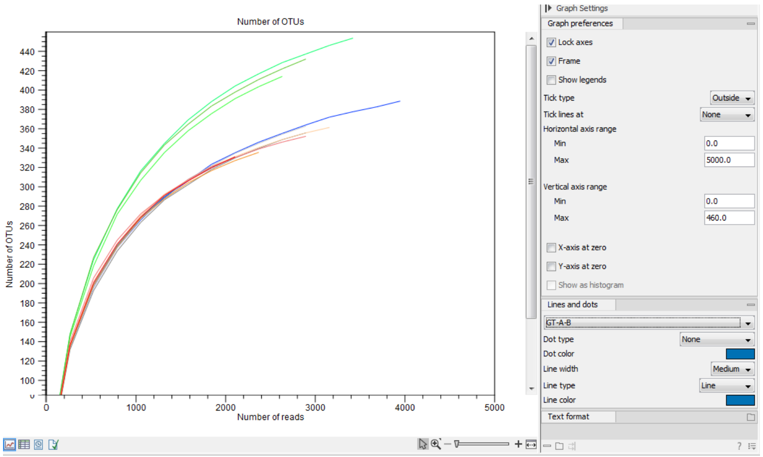
☑ GTA
☑ GTB
☑ 1
☑ 2
☑ 3

Color: Phylum

- k__Archaea, p__Crenarchaeota
- k__Bacteria, p__AD3
- k__Bacteria, p__Acidobacteria
- k__Bacteria, p__Actinobacteria
- k__Bacteria, p__Armatimonadetes
- k__Bacteria, p__Bacteroidetes
- k__Bacteria, p__Chlamydiae
- k__Bacteria, p__Chlorobi

Figure 5: *Aggregate samples based on metadata information.*

Open now the 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 one sample. The coloring scheme can be set by double clicking on a graph and changing the color for lines one sample at a time. Note: you need to set different shades of a particular color to the different samples as the samples cannot have the exact same color. In the following graph (figure 6) we have chosen shades of blue for GT-A and shades of green for GT-B. GT-A samples seem to have similar measures of alpha diversity as the sites 1, 2, and 3 while GT-B samples are clearly apart. We can conclude that Mr.X was not wearing his hiking boots on any of the sites sampled.



Graph Settings

Graph preferences

☑ Lock axes
☑ Frame
☐ Show legends

Tick type: Outside
Tick lines at: None
Horizontal axis range
Min: 0.0
Max: 5000.0
Vertical axis range
Min: 0.0
Max: 460.0

☐ X-axis at zero
☐ Y-axis at zero
☐ Show as histogram

Lines and dots

GT-A-B

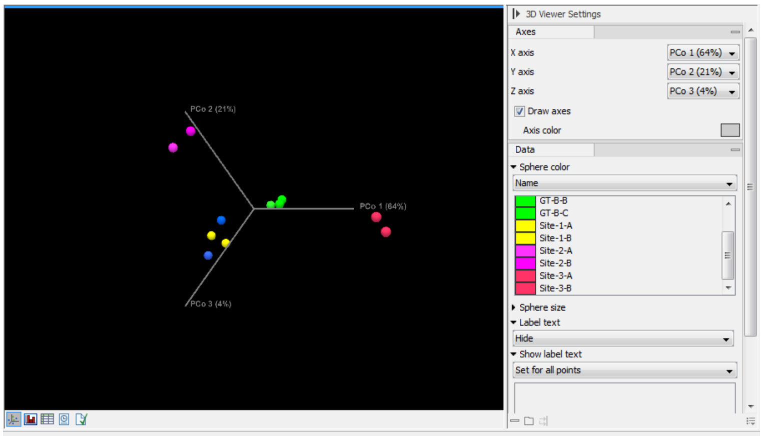
Dot type: None
Dot color: 🟦
Line width: Medium
Line type: Line
Line color: 🟦

Text format

Figure 6: *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 Component Analysis PCoA ( ) on the UniFrac distances (figure 7).

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



3D Viewer Settings

Axes

X axis: PCo 1 (64%)
Y axis: PCo 2 (21%)
Z axis: PCo 3 (4%)
☑ Draw axes
Axis color

Data

Sphere color: Name

🟩 GT-B-B
🟩 GT-B-C
🟨 Site-1-A
🟨 Site-1-B
🟪 Site-2-A
🟪 Site-2-B
🟥 Site-3-A
🟥 Site-3-B

Sphere size
Label text: Hide
Show label text: Set for all points

Figure 7: *Result of the beta diversity analysis.*

selecting PCoA as X axis and PCoA1 as Y axis allows for a better visualization of Site 1 and GT-A being clustered together, confirming a similarity between the 2 soils and thus confirming our suspicion that Mr. X was on site 1 with his rubber boots.

Additional statistical analyses

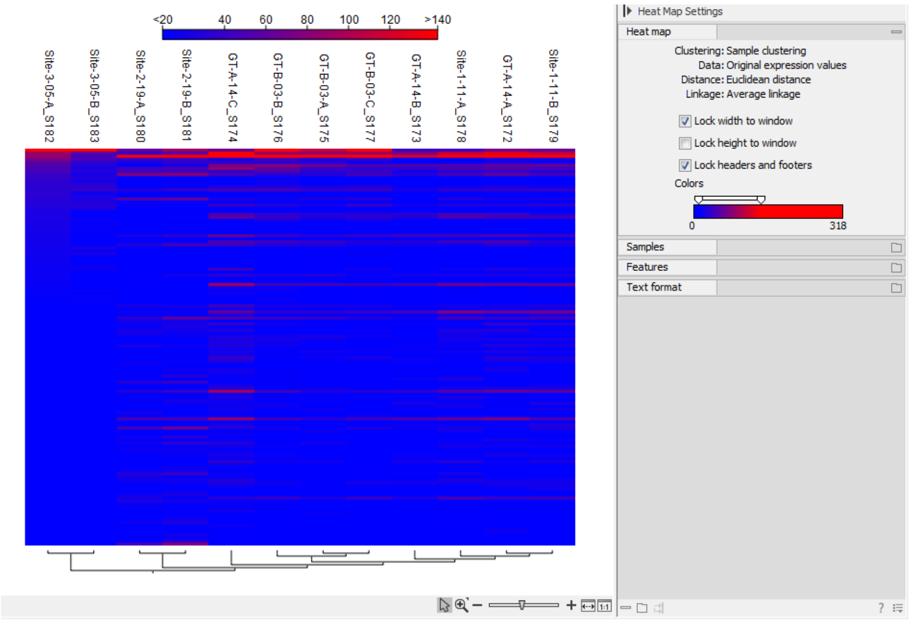
As a tool for assessing similarity between samples, a heat map and dendrogram can be helpful. In order to perform additional statistical tests, you can use the **Convert to Experiment** tool to assign categories to the samples from your OTU abundance table.

1. Select **Toolbox** | **OTUclustering** ( ) | **Convert to Experiment** ( )
2. Choose OTU(Table) from the **Data QC and OTU clustering** folder as input, and which Metadata group to consider as factors (here **Type**) before saving the resulting experiment in a new folder you can call **Statistics**. The file is called OTU(Table)(experiment).
3. Select **Toolbox** | **Transcriptomics Analysis** ( ) | **Quality control** ( ) | **Hierarchical Clustering of Samples** ( )
4. Choose the OTU(Table)(experiment) as input. Click on the button labeled Next.
5. Select **Euclidian distance** and **Average linkage** and click on the button labeled Next.
6. Save your result in the **Statistics** folder. It will overwrite the OTU(Table)(experiment) file.

If the OTU(Table)(experiment) table was previously opened in your workbench, close it and open it again. The Experiment table now displays a button at the bottom with a heat map icon. Selecting this view will display the heat map (figure 8). We can see that TG-A is again nested together with Site 1, confirming once more that the soil found on the rubber boots 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 Analysis can be used to measure effect size and significance of beta diversity.

1. Select **Toolbox** | **OTUclustering** ( ) | **PERMANOVA Analysis** ( ).



Heat Map Settings

Heat map

Clustering: Sample clustering
Data: Original expression values
Distance: Euclidean distance
Linkage: Average linkage

☑ Lock width to window

□ Lock height to window

☑ Lock headers and footers

Colors

Samples

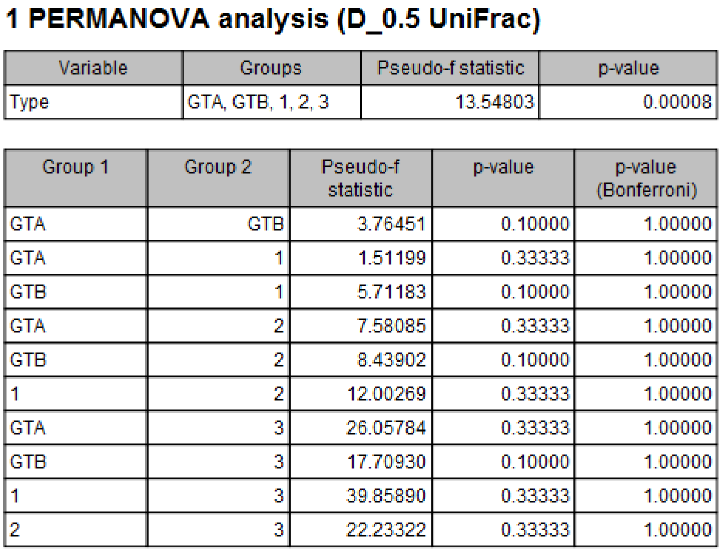
Features

Text format

Figure 8: *Heat map of the Hierarchical Clustering of Samples analysis.*

2. Choose OTU(Table) from the **Data QC and OTU clustering** folder as input and select Type as Metadata group.
3. Deselect Bray-Curtis and Jaccard measures and specify the phylogenetic tree (OTU(Table)(Filtered)aligr from the **Estimate Alpha and Beta Diversities** folder. Select D_0.5 UniFrac and 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 9).



1 PERMANOVA analysis (D_0.5 UniFrac)

Variable	Groups	Pseudo-f statistic	p-value
Type	GTA, GTB, 1, 2, 3	13.54803	0.00008

Group 1	Group 2	Pseudo-f statistic	p-value	p-value (Bonferroni)
GTA	GTB	3.76451	0.10000	1.00000
GTA	1	1.51199	0.33333	1.00000
GTB	1	5.71183	0.10000	1.00000
GTA	2	7.58085	0.33333	1.00000
GTB	2	8.43902	0.10000	1.00000
1	2	12.00269	0.33333	1.00000
GTA	3	26.05784	0.33333	1.00000
GTB	3	17.70930	0.10000	1.00000
1	3	39.85890	0.33333	1.00000
2	3	22.23322	0.33333	1.00000

Figure 9: *Result of the PERMANOVA analysis.*

The PERMANOVA confirms that the clusters are significant (p=0,00008), 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 - and in particular from the rubber boots solesand site 1 - to transform this analysis into actual evidence!

MGM

CLC **Microbial Genomics** Module

USER MANUAL

User manual for

CLC Microbial Genomics Module 1.5

Windows, Mac OS X and Linux

May 29, 2015

This software is for research purposes only.

CLC bio, a QIAGEN Company
Silkeborgvej 2
Prismet
DK-8000 Aarhus C
Denmark

Contents

1 Introduction to the Microbial Genomics Module **5**

1.1 The concept of the Microbial Genomics Module 5

1.2 Contact information 6

2 System requirements and installation **7**

2.1 System Requirements 7

2.2 Installation of plugin 8

2.3 Uninstall of plugin 8

3 OTU clustering **10**

3.1 Download OTU Reference Database 10

3.2 Format Reference Database 10

3.3 Optional Merge Paired Reads 11

3.4 Fixed Length Trimming 12

3.5 Filter Samples Based on Number of Reads 13

3.6 OTUclustering 13

3.6.1 OTUclustering parameters 13

3.6.2 Add Metadata to OTU Abundance Table 16

3.6.3 Remove OTUs with Low Abundance 16

3.6.4 Visualization of the OTU abundance table 17

4 Estimation of Alpha and Beta diversity **20**

4.1 Align OTUs with MUSCLE 20

4.2 Alpha Diversity 21

4.2.1 Alpha diversity measures 21

4.3 Beta Diversity 22
4.3.1 Beta diversity measures 23
4.4 PERMANOVA Analysis 24
4.5 Convert to Experiment 25
4.6 Importing OTU abundance tables 25

5 Workflows **27**
5.1 Data QC and OTU clustering workflow 27
5.2 Estimate Alpha and Beta Diversities workflow 27

Bibliography **30**

Index **30**

Chapter 1

Introduction to the Microbial Genomics Module

1.1 The concept of the Microbial Genomics Module

The majority of microbial species present in the human body or indeed anywhere in the environment have never been isolated, cultured or sequenced, due to our inability to reproduce necessary growth conditions in the lab. Therefore, there are huge amounts of organismal and functional novelty still to be discovered. Two central questions in microbial community analysis ask: Which microbial species are present in a sample from a given habitat, and at what frequencies? Microbiome analysis takes advantage of DNA molecular techniques and sequencing technology in order to comprehensively retrieve specific regions of microbial genomic DNA useful for taxonomical identification. For bacteria, the most widely used regions are parts of the 16S rRNA gene. In a microbiome analysis workflow, total genomic 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 their occurrences. 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 pseudospecies called Operational Taxonomical Units (OTUs), where all reads within e.g. 97% similarity are clustered together and represented by a single sequence. PCR amplification can introduce artefacts in the form of chimeric sequences, where template swapping results in a sequence having two or more parental templates. These can be identified during the clustering step.

The features of the Microbial Genomics Module include:

- Trimming and merging of reads.
- Clustering of reads into OTUs.
- Generation of OTU tables.
- Working with metadata.
- Visualization options (stacked barplots, zoomable sunbursts).
- Multiple sequence alignment using MUSCLE.

- Phylogenetic tree construction using maximum likelihood.
- Estimation of alpha diversity.
- Rarefaction analysis.
- Estimation of beta diversity.
- Principal coordinates analysis
- PERMANOVA test
- Statistical tests for differential abundance

The primary output of the clustering and tallying process is an OTU table, listing the abundances of OTUs in the samples under investigation, as well as new features allowing clear visualization of the results. Secondary analyses include estimations of alpha and beta diversities, in addition to statistical tests for differential abundance.

1.2 Contact information

The CLC Workbench is developed by:

CLC bio, a QIAGEN Company
Silkeborgvej 2
Prismet
8000 Aarhus C
Denmark

http://www.clcbio.com

VAT no.: DK 28 30 50 87

Telephone: 45 70 22 32 44
Fax: +45 86 20 12 22

E-mail: info@clcbio.com

If you have questions or comments regarding the program, you can contact us through the support team as described here: http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Getting_help.html.

Chapter 2

System requirements and installation

2.1 System Requirements

With the exception of the two editors below, the system requirements of Microbial Genomics Module are the same as the ones required for the CLC Genomics Workbench:

- Windows Vista, Windows 7, Windows 8 or Windows Server 2008.
- Mac OS X 10.7 or later.
- Linux: RHEL 5.0 or later. SUSE 10.2 or later. Fedora 6 or later.
- 2 GB RAM required.
- 4 GB RAM recommended.
- 1024 x 768 display required.
- 1600 x 1200 display recommended.
- Intel or AMD CPU required.
- CLC Genomics Workbench version 8.0 or higher.

The PCoA 3D viewer requirements are the same as the 3D Molecule Viewer:

- **System requirements**
  - A graphics card capable of supporting OpenGL 2.0.
  - Updated graphics drivers. Please make sure the latest driver for the graphics card is installed.
- **System Recommendations**
  - A discrete graphics card from either Nvidia or AMD/ATI. Modern integrated graphics cards (such as the Intel HD Graphics series) may also be used, but these are usually slower than the discrete cards.
  - A 64-bit workbench version is recommended for working with large complexes.

The Sunburst viewer makes use of JavaFX and may not work on older Linux kernels. Un updated list of requirements for JavaFX can be found at http://www.oracle.com/technetwork/java/javafx/downloads/supportedconfigurations-1506746.html.

2.2 Installation of plugin

Plugins are installed using the Plugins and Resources Manager[1], which can be accessed via the menu in the Workbench

Help | Plugins and Resources ( )

or via the **Plugins ( )** button on the Toolbar.

From within the Plugins and Resources Manager, choose the Download Plugins tab and click on the CLC Workbench Client Plugin. Then click on the button labeled **Download and Install**.

If you are working on a system not connected to the internet, then you can also install the plugin by downloading the cpa file from the plugins page of our website

http://www.clcbio.com/clc-plugin/

Then start up the Plugin manager within the Workbench, and click on the button at the bottom of the Plugin manager labeled **Install from File**.

You need to restart the Workbench before the plugin is ready for use.

2.3 Uninstall of plugin

Plugins are uninstalled using the plugin manager:

Help in the Menu Bar | Plugins and Resources... ( )

or **Plugins ( ) in the Toolbar**

This will open the dialog shown in figure 2.1.

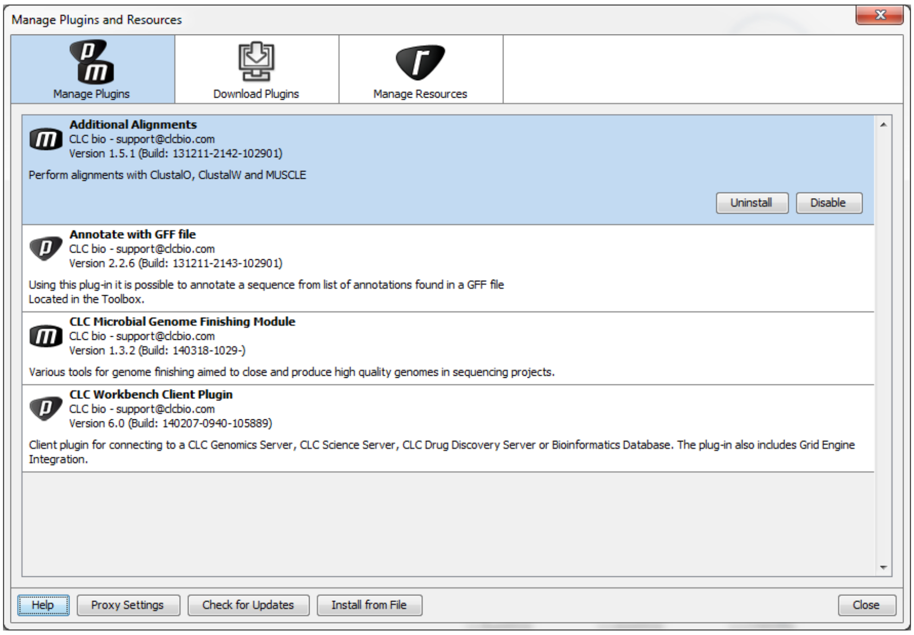
The installed plugins are shown in this dialog. To uninstall:

Click the Microbial Genomics Module | Uninstall

If you do not wish to completely uninstall the plugin but you don't want it to be used next time you start the Workbench, click the **Disable** button.

When you close the dialog, you will be asked whether you wish to restart the workbench. The plugin will not be uninstalled until the workbench is restarted.

[1] In order to install plugins on many systems, the Workbench must be run in administrator mode. On Windows Vista and Windows 7, you can do this by right-clicking the program shortcut and choosing "Run as Administrator".



Additional Alignments
CLC bio - support@clcbio.com
Version 1.5.1 (Build: 131211-2142-102901)

Perform alignments with ClustalO, ClustalW and MUSCLE

Annotate with GFF file
CLC bio - support@clcbio.com
Version 2.2.6 (Build: 131211-2143-102901)

Using this plug-in it is possible to annotate a sequence from list of annotations found in a GFF file
Located in the Toolbox.

CLC Microbial Genome Finishing Module
CLC bio - support@clcbio.com
Version 1.3.2 (Build: 140318-1029-)

Various tools for genome finishing aimed to close and produce high quality genomes in sequencing projects.

CLC Workbench Client Plugin
CLC bio - support@clcbio.com
Version 6.0 (Build: 140207-0940-105889)

Client plugin for connecting to a CLC Genomics Server, CLC Science Server, CLC Drug Discovery Server or Bioinformatics Database. The plug-in also includes Grid Engine Integration.

Figure 2.1: *The plugin manager with plugins installed.*

Chapter 3

OTU clustering

In this chapter we will describe the tools included in Microbial Genomics Moduleand that you can use to prepare your data for OTU clustering. First we will explain you how to download preexisting database, or how to format your own. We will then talk about tools that trim and filter your data so that you can cluster only the reads that are of best quality. We will finally describe the OTUclustering tool itself, as well the different ways you can visualize your results.

3.1 Download OTU Reference Database

OTU reference databases contain representative OTU sequences and their taxonomy. They are needed to perform reference-based OTU clustering. Three popular reference OTU databases, clustered at various similarity percentages, can be downloaded using the Download OTU Reference Database tool:

- Greengenes: 16S rRNA gene from ribosomal Small Subunit for Prokaryotic taxonomic assignment clustered OTUs at different percentages. http://greengenes.secondgenome.com/downloads
- Silva/Arb SSU: 16S/18S rRNA from ribosomal Small Subunit for Prokaryotic and Eukaryotic taxonomic assignment clustered OTUs at different percentages. http://www.arb-silva.de/download/archive/qiime/
- UNITE: ITS spacer clustered OTUs at different percentages for fungal taxonomic assignment. https://unite.ut.ee/repository.php

To run the tool, go to **Toolbox** | **OTUclustering ()** | **Download OTU Reference Database ()**, then select the database you want to download and where to store the result.

3.2 Format Reference Database

In addition to the download of reference databases using the **Download OTU Reference Database** tool, you can format other databases by running the Format Reference Database tool

Toolbox | **OTUclustering ()** | **Format Reference Database ()**.

After clicking next, specify a taxonomy file and the similarity percentage threshold used to cluster sequences into OTUs. Each line of the taxonomy file should contain an OTU name and its taxonomy, where taxonomy levels are semicolon-separated. For example, the following line

```
o123  k__Bacteria;p__Bacteroidetes;c__Sphingobacteria;o__;f__;g__;s__.
```

indicates that the OTU o123 belongs to the class Sphingobacteria and that its taxonomy is specified only up to the class level.

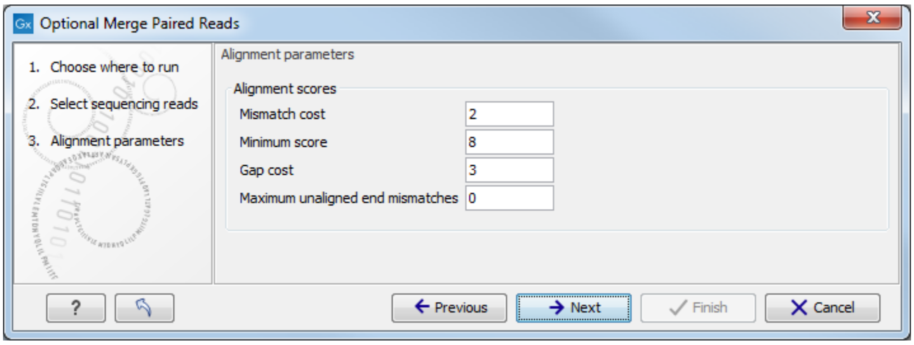
3.3 Optional Merge Paired Reads

In order to use the highest quality sequences for clustering, it is recommended to merge paired data. Indeed, if the read length is smaller than the amplicon size, forward and reverse reads are expected to overlap in most of their 3' regions. Therefore, one can merge the forward and reverse reads to yield one high quality representative using the Optional Merge Paired Reads tool. The Optional Merge Paired Reads tool will merge paired-end reads according to some pre-selected merge parameters: the overlap region and the quality of the sequences. For example, for a designed 150 bp overlap, a maximum score of 150 is achievable, but as the real length of the overlap is unknown, a lower minimum score should be chosen. Also, some mismatches and InDels should be allowed, especially if the sequence quality is not perfect. You can also decide penalties for mismatch, gap and unaligned ends.

To run the Optional Merge Paired Reads tool, go to **Toolbox** | **OTUclustering ()** | **Optional Merge Paired Reads ()**.

Select any number of sequences as input. The tool accepts both paired and unpaired reads but will only merge the paired reads while returning the unpaired ones as "not merged" reads in the output. Note that paired reads have to be in forward-reverse orientation. After merging, the merged reads will always be in the forward orientation.

Click Next to open the dialog shown in figure 3.1.



Alignment parameters

Alignment scores

Mismatch cost 2

Minimum score 8

Gap cost 3

Maximum unaligned end mismatches 0

Figure 3.1: *Alignment parameters.*

In order to understand how these parameters should be set, an explanation of the merging algorithm is needed: Because the fragment size is not an exact number of base pairs and is different from fragment to fragment, an alignment of the two reads has to be performed. If the alignment is *good and long enough*, the reads will be merged. *Good enough* in this context means that the alignment has to satisfy some user-specified score criteria (details below). Because of sequencing errors that typically are more abundant towards the end of the read, the alignment is not expected always to be perfect, and the user can decide how many errors are acceptable. *Long enough* in this context means that the overlap between the reads has to be non-coincidental.

Merging two reads that do not really overlap, leads to errors in the downstream analysis, thus it is very important to make sure that the overlap is big enough. If only a few bases overlap was required, some read pairs will match by chance, so this has to be avoided.

The following parameters are used to define what is *good enough* and *long enough*.

- **Mismatch cost**: The alignment awards one point for a match, and the mismatch cost is set by this parameter. The default value is 2.
- **Minimum score**: This is the minimum score of an alignment to be accepted for merging. The default value is 8. As an example: with default settings, this means that an overlap of 11 bases with one mismatch will be accepted (10 matches minus 2 for a mismatch).
- **Gap cost**: This is the cost for introducing an insertion or deletion in the alignment. The default value is 3.
- **Maximum unaligned end mismatches**: The alignment is local, which means that a number of bases can be left unaligned. If the quality of the reads is dropping to be very poor towards the end of the read, and the expected overlap is long enough, it makes sense to allow some unaligned bases at the end. However, this should be used with great care which is why the default value is 0. As explained above, a wrong decision to merge the reads leads to errors in the downstream analysis, so it is better to be conservative and accept fewer merged reads in the result. Please note that even with the alignment scores above the minimum score specified in the tool setup, the paired reads also need to have the number of end mismatches below the "Maximum unaligned end mismatches" value specified in the tool setup to be qualified for merging.

The main result will be two sequence lists for each sample selected as input to the tool: one containing the merged reads (labeled as "merged"), and one containing the reads that could not be merged (labeled as "not merged"). Note that low quality can be one of the reasons why a pair cannot be merged. Hence, the list of reads that could not be paired is more likely to contain more reads with errors than the one with the merged reads.

3.4 Fixed Length Trimming

In order to compare sequences and cluster them, they all need to be of exact same length. All reads which are shorter than the cut-off are discarded, and reads longer than that are trimmed back to the chosen length. Note: we recommend to perform a step in which adapter sequences are removed from the reads. For more information, read the adaptor trimming step in the Microbial Genomics Module tutorial, or the Adapter Trimming section of the CLC Genomics Workbench manual (http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Adapter_trimming.html).

To run the tool, go to **Toolbox** | **OTUclustering ( )** | **Fixed Length Trimming ( )** and select the sequences you would like to trim.

In the next wizard window you can enter manually the desired length for the trimmed reads. Alternatively, the Fixed length Trimming algorithm can calculate the trimming cut-off value as the mean length of the merged reads minus one standard deviation. If this option is chosen, it is important that all samples are trimmed at the same time as the mean and standard deviation for the combined reads in all samples needs to be estimated at once.

3.5 Filter Samples Based on Number of Reads

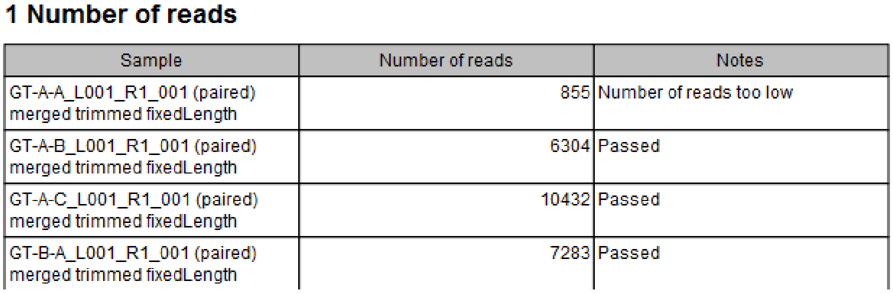
In order to cluster accurately samples, they should have comparable coverage. Sometimes, however, DNA extraction, PCR amplification, library construction or sequencing has not been entirely successful, and a fraction of the resulting sequencing data will be represented by too few reads. These samples should be exclude from further analysis using the Filter Samples Based on Number of Reads tool.

To run the tool, go to **Toolbox** | **OTUclustering ()** | **Filter Samples Based on Number of Reads ()**.

The tool requires that the input reads from each sample must be either all paired or all single. This check ensures that the samples are comparable, as the number of reads before merging paired reads is twice as great as the number of merged reads. The preferred way to run this tool with OTU sequencing data is to use single reads obtained from the Optional Merge Paired Reads tool.

The threshold for determining whether a sample has sufficient coverage is specified by the parameters **minimum number of reads** and **minimum percent from the median**. The algorithm filters out all samples whose number of reads is less than the **minimum number of reads** or less than the **minimum percent from the median** times the median number of reads across all samples.

The primary output is a table describing how many reads are in a particular sample and if they passed or failed the quality control (see figure 3.2).



1 Number of reads

Sample	Number of reads	Notes
GT-A-A_L001_R1_001 (paired) merged trimmed fixedLength	855	Number of reads too low
GT-A-B_L001_R1_001 (paired) merged trimmed fixedLength	6304	Passed
GT-A-C_L001_R1_001 (paired) merged trimmed fixedLength	10432	Passed
GT-B-A_L001_R1_001 (paired) merged trimmed fixedLength	7283	Passed

Figure 3.2: *Output table from the Filter Samples Based on Number of Reads tool.*

In the next wizard window you can decide to **Copy samples with sufficient coverage** as well as to **Copy the discarded samples**. Copying the samples with sufficient coverage will give you a new list of sequences that you can use in your following analyses because it does not contain the reads of poor quality that failed the Remove the samples with Low Coverage analysis.

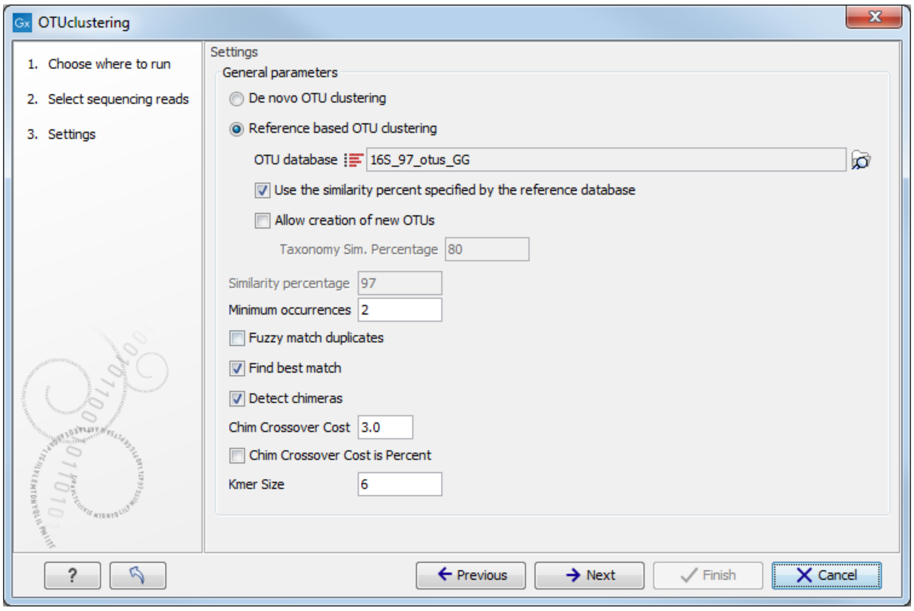
3.6 OTUclustering

The OTUclustering tool clusters a collection of fixed length trimmed reads in operational taxonomy units.

To run the tool, go to **Toolbox** | **OTUclustering ()** | **OTUclustering ()**.

3.6.1 OTUclustering parameters

After having selected the sequences you would like to cluster, the wizard offers to set some general parameters (see figure 3.3).



OTUclustering

1. Choose where to run
2. Select sequencing reads
3. Settings

Settings

General parameters

○ De novo OTU clustering

● Reference based OTU clustering

OTU database 16S_97_otus_GG

☑ Use the similarity percent specified by the reference database

□ Allow creation of new OTUs

Taxonomy Sim. Percentage 80

Similarity percentage 97

Minimum occurrences 2

□ Fuzzy match duplicates

☑ Find best match

☑ Detect chimeras

Chim Crossover Cost 3.0

□ Chim Crossover Cost is Percent

Kmer Size 6

Figure 3.3: *Outputtable from the Remove Low Coverage Samples tool.*

You can choose to perform a **De novo OTU clustering**, or you can perform a **Reference based OTU clustering**. The following parameters can then be set:

- **OTU database** The reference database to be used for Reference based OTU clustering. Reference databases can be created by the Download OTU Reference Database or the Format Reference Database tools. A database must be specified to perform a Reference based OTU clustering.

- **Use the similarity percent specified by the reference database** Allows to use the same similarity percent value (see below) that was used when creating the reference database. This parameter is available only when performing a Reference based OTU clustering. Selecting this parameter will disable the similarity percent parameter.

- **Allow creation of new OTUs** Allows sequences which are not already represented at the given similarity distance in the database to form a new cluster, and a new centroid is chosen. This parameter can be set only when performing a Reference based OTU clustering, disallowing the creation of new OTUs is also known as closed reference OTU picking.

- **Taxonomy similarity percentage** Specifies the similarity percentage to be used when creating new OTUs. This parameter is available only when **Allow creation of new OTUs** is selected.

- **Similarity percentage**: The required percentage of the identity for the alignment between a read and the centroid of an OTU for the read to join the OTU. An initial screening of the potential OTU's to align is done by excluding OTU's based on the number of kmer hits.

- **Minimum occurrences**: The minimum number of duplicates for specific read-data before the data will be included in further study. For instance, if set to 2, at least two reads with the same exact nucleotides needs to exist in the input for the data to propagate to further analysis. Other data will be thrown away. This can for instance be used to filter out singletons. Note that matches does not need to be exact when the **Fuzzy match duplicates** option is used.

- **Fuzzy match duplicates**: Specifies how duplicates are defined. If the option is not selected two reads are only duplicates if they are exactly equal. If the option is selected, two reads are duplicates if they are almos? equal, i.e. all differences are SNVs and there are not too many of them ($\leq 2\%$). This pseudo-merging is done by lexicographically sorting the input and looking in the neighborhood of the read being processed. The reads are processed from most abundant (in a completely equivalent sense) to the least. In this way two singletons can for instance be pseudo-merged together and be included for further study despite the **Minimum occurrences** option having specified 2. Upon further analysis a group can be split into several OTUs if not all members are within the specified threshold from the ''OTU-leader''.

- **Find best match**: If the option is not selected, the read becomes a member of the first OTU-database entry found within the specified threshold. If the option is selected all database entries are tested and the read becomes a member of the best matching result. Note that ''First'' and ''all'' are relative terms in this case as kmer-searches are used to speed up the process. ''All'' only includes the database entries that the kmer search deems close enough (i.e. database entries that cannot be within the specified threshold will be filtered out at this step). ''First'' is the first matching entry as returned by the kmer-search which will sort by the number of kmer-matches (most first).

- **Detect chimeras**: Chimeric sequences are frequent artifacts of PCR reactions. They are detected by assessing whether a sequence is likely to have two different and more frequent "parent" sequences in the current collection of OTUs, meaning that two fragments of the sequence map to different sequences.

- **Chimera crossover cost**: The cost of doing a chimeric crossover, i.e. the higher the cost the less likely it is that a read is marked as chimeric.

- **Chimera crossover cost is percent**: Whether the above cost is an absolute value or a percentage value (i.e. an absolute value will be automatically calculated based on the read length).

- **Kmer size**: The size of the kmer to use in regards to the kmer usage in finding the best match.

Chimera detection is performed by kmer searches as follows:

- All database entries and the read being processed are split into 4 equally sized portions with an additional 3 half-way-shifted to cover the merge-points. This results in 7 different kmer-search-options. Each of these is queried for matches within some threshold, and only the results are processed further. A database entry has to fit well in at least one of these 7 portions for the database-entry to be relevant.

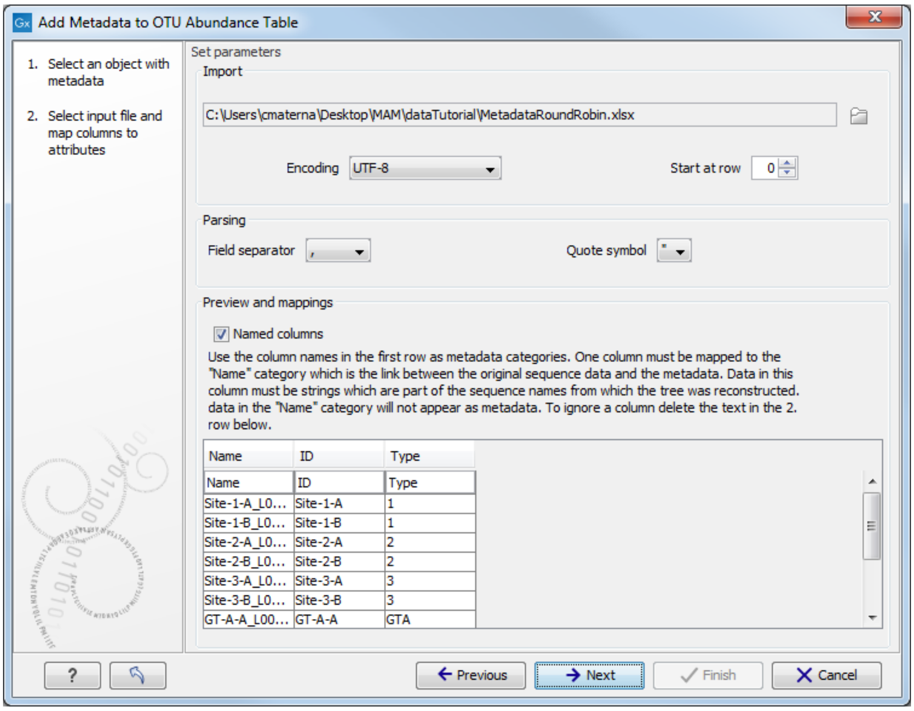
- Given the 7 sets of database entries some entries may just be present in one of the sets. For these entries, some may be duplicates in the region they represent. The duplicates are filtered out and only an arbitrary representative of the duplicates is kept.

The OTU clustering tool produces several outputs: a sequence list of the OTU centroids or of the Chimeras, and abundance tables with the newly created OTUs or the chimeras. Each table give abundance of the OTU or chimeras at each site as well as the total abundance for all samples.

3.6.2 Add Metadata to OTU Abundance Table

In order to enhance the visualization functionality of the OTU abundance table, it is useful to decorate it with metadata on the samples. To run the tool go to: **Toolbox** | **OTUclustering** ( ) | **Add Metadata to OTU Abundance Table** ( )

Choose a OTU table as input. In the next wizard window you can select a file describing the metadata on your local computer (figure 3.4).



Add Metadata to OTU Abundance Table

1. Select an object with metadata
2. Select input file and map columns to attributes

Set parameters

Import

C:\Users\cmaterna\Desktop\MAM\dataTutorial\MetadataRoundRobin.xlsx

Encoding UTF-8

Start at row 0

Parsing

Field separator ,

Quote symbol "

Preview and mappings

☑ Named columns

Use the column names in the first row as metadata categories. One column must be mapped to the "Name" category which is the link between the original sequence data and the metadata. Data in this column must be strings which are part of the sequence names from which the tree was reconstructed. data in the "Name" category will not appear as metadata. To ignore a column delete the text in the 2. row below.

Name	ID	Type
Name	ID	Type
Site-1-A_L0...	Site-1-A	1
Site-1-B_L0...	Site-1-B	1
Site-2-A_L0...	Site-2-A	2
Site-2-B_L0...	Site-2-B	2
Site-3-A_L0...	Site-3-A	3
Site-3-B_L0...	Site-3-B	3
GT-A-A_L00...	GT-A-A	GTA

Figure 3.4: *Setting up metadata parameters.*

The metadata should be formatted in a tabular file format (i.e. .xls, .xlsx, .csv). The first row of the table should contain column headers. There should be one column called ''Name'' and the entries in this column should match the names of the inputs to OTUclustering. This column is used to match row in the table with samples present in the OTU table, so if the names do not match you will not be able to aggregate your data at all. You can have as many other columns as you would like, and these information can be used as grouping variables to improve visualization or to perform analyses. If you wish to ignore a column without deleting it from your file, simply delete the text in the header row.

3.6.3 Remove OTUs with Low Abundance

Low abundance OTUs can eliminated from the OTU table if they have fewer than a given count across all the samples in the experiment.

To run the tool, go to **Toolbox** | **OTUclustering** ( ) | **Remove OTUs with Low Abundance** ( )

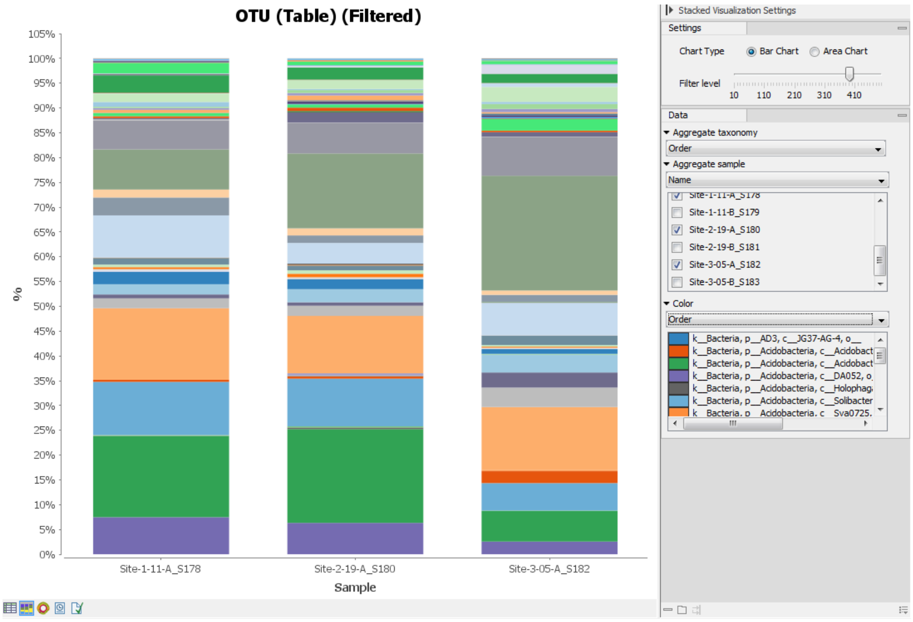
Choose one OTU table as input, select the filtering parameters and save the table. The threshold for determining whether an OTU has sufficient abundance is specified by the parameters **minimum combined abundance** and **minimum combined abundance (% of all the reads)**. The algorithm filters out all OTUs whose combined abundance across all samples is less than the **minimum combined abundance** or whose combined abundance is less than the **minimum combined**

abundance (% of all the reads) across all samples. The default value for the Minimum combined abundance is set at 10.

3.6.4 Visualization of the OTU abundance table

The OTU clustering tool produces several outputs: a sequence list of the OTU centroids or of the Chimeras, and abundance tables with the newly created OTUs or the chimeras. Each table give abundance of the OTU or chimeras at each site as well as the total abundance for all samples. There are a number of ways of visualizing the contents of an OTU abundance table:

- **Table view** (▦) The table display the following columns:
  - **Name** The name of the OTU, specified by either the reference database or by the OTU representative.
  - **Taxonomy** The taxonomy of the OTU, as specified by the reference database when a database entry was used as Reference.
  - **Combined Abundance** The total number of reads belonging to the OTU across all samples.
  - **Abundance for each sample** The number of reads belonging to the OTU in a specific sample.
  - **Sequence** The sequence of the centroid of the OTU.
- **Stacked Bar Chart and Stacked Area Chart** (▦) In the Stacked Bar (figure 3.5) and **Stacked Area Charts** (figure 3.6), the metadata can be used to aggregate groups of columns (samples) by selecting the relevant metadata category in the right hand side panel. Also, the data can be aggregated at any taxonomy level selected. The relevant data points will automatically be summed accordingly.



Stacked Visualization Settings

Settings

Chart Type ● Bar Chart ○ Area Chart

Filter level 10 110 210 310 410

Data

Aggregate taxonomy: Order

Aggregate sample: Name

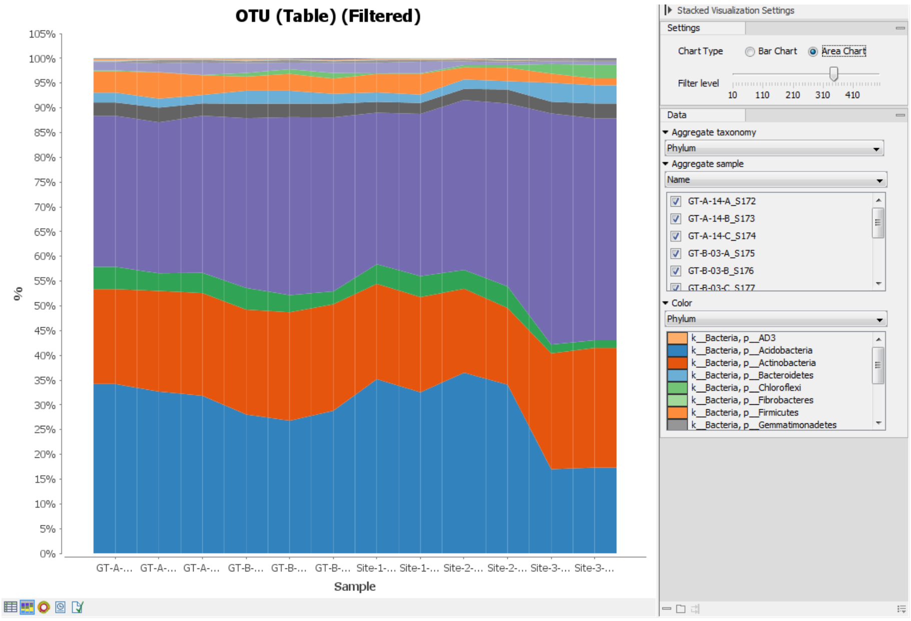
☑ Site-1-11-A_S178
☐ Site-1-11-B_S179
☑ Site-2-19-A_S180
☐ Site-2-19-B_S181
☑ Site-3-05-A_S182
☐ Site-3-05-B_S183

Color: Order

🟦 k__Bacteria, p__AD3, c__JG37-AG-4, o__
🟧 k__Bacteria, p__Acidobacteria, c__Acidobact
🟩 k__Bacteria, p__Acidobacteria, c__Acidobact
🟪 k__Bacteria, p__Acidobacteria, c__DA052, o
⬛ k__Bacteria, p__Acidobacteria, c__Holophag
🟦 k__Bacteria, p__Acidobacteria, c__Solibacter
🟧 k__Bacteria, p__Acidobacteria, c__Sva0725.

Figure 3.5: *Stacked bar of the microbial community at the order level for 3 different sites.*

Holding the pointer over a colored area in any of the plots will result in the display of the corresponding taxonomy label and counts. The slider **Filter level** allows the setting of a



Stacked Visualization Settings

Settings

Chart Type ○ Bar Chart ● Area Chart

Filter level 10 110 210 310 410

Data

Aggregate taxonomy: Phylum

Aggregate sample: Name

☑ GT-A-14-A_S172
☑ GT-A-14-B_S173
☑ GT-A-14-C_S174
☑ GT-B-03-A_S175
☑ GT-B-03-B_S176
☑ GT-B-03-C_S177

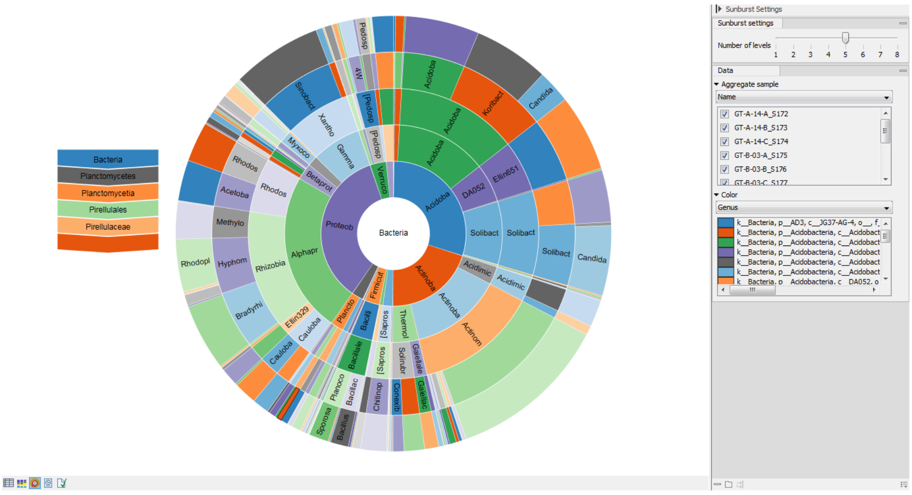
Color: Phylum

🟧 k__Bacteria, p__AD3
🟦 k__Bacteria, p__Acidobacteria
🟥 k__Bacteria, p__Actinobacteria
🟦 k__Bacteria, p__Bacteroidetes
🟩 k__Bacteria, p__Chloroflexi
🟩 k__Bacteria, p__Fibrobacteres
🟧 k__Bacteria, p__Firmicutes
⬛ k__Bacteria, p__Gemmatimonadetes

Figure 3.6: *Stacked area of the microbial community at the phylum level for 12 different sites.*

viewing filter on the minimum counts for individual viewing instead of being aggregated in the "Other" category displayed in the top field of the chart. One can select which taxonomy level to color, and change the default colors manually. Colors can be be specified at the same taxonomy level as the one use to aggregate the data or at a lower level. When lower taxonomy levels are chosen in the data aggregation field, the color will be inherited in alternating shadings. Using the bottom right-most button (**Save/restore settings** (☰)), the settings can be saved and applied in other plots, allowing visual comparisons across analyses.

- **Zoomable Sunbursts** (◎) The Zoomable Sunburst viewer lets the user select how many taxonomy level counts to display, and which level to color. Again, lower levels will inherit the color in alternating shadings. The metadata can be used to select which sample or group of samples to show in the sunburst (figure 3.7).



Sunburst Settings

Sunburst settings

Number of levels 1 2 3 4 5 6 7 8

Data

Aggregate sample: Name

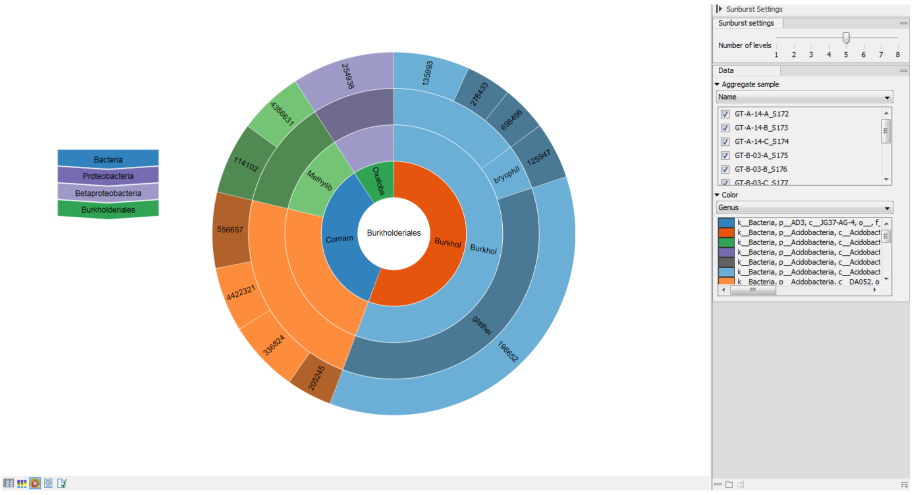
☑ GT-A-14-A_S172
☑ GT-A-14-B_S173
☑ GT-A-14-C_S174
☑ GT-B-03-A_S175
☑ GT-B-03-B_S176
☑ GT-B-03-C_S177

Color: Genus

🟦 k__Bacteria, p__AD3, c__JG37-AG-4, o__, f_
🟥 k__Bacteria, p__Acidobacteria, c__Acidobact
🟩 k__Bacteria, p__Acidobacteria, c__Acidobact
🟪 k__Bacteria, p__Acidobacteria, c__Acidobact
⬛ k__Bacteria, p__Acidobacteria, c__Acidobact
🟦 k__Bacteria, p__Acidobacteria, c__Acidobact
🟧 k__Bacteria, p__Acidobacteria, c__DA052, o

Figure 3.7: *Sunburst view of the microbial community showing all taxa belonging to the kingdom bacteria.*

Clicking on a lower level field will render that field the center of the plot and display lower level counts in a radial view. Clicking on the center field will render the level above the current view the center of the view (figure 3.8).



Sunburst Settings

Sunburst settings

Number of levels: 5 (1 2 3 4 5 6 7 8)

Data

Aggregate sample: Name

☑ GT-A-14-A_S172
☑ GT-A-14-B_S173
☑ GT-A-14-C_S174
☑ GT-B-03-A_S175
☑ GT-B-03-B_S176
☑ GT-B-03-C_S177

Color: Genus

🟦 k__Bacteria, p__AD3, c__JG37-AG-4, o__, f_
🟧 k__Bacteria, p__Acidobacteria, c__Acidobact
🟩 k__Bacteria, p__Acidobacteria, c__Acidobact
🟪 k__Bacteria, p__Acidobacteria, c__Acidobact
⬛ k__Bacteria, p__Acidobacteria, c__Acidobact
🟦 k__Bacteria, p__Acidobacteria, c__Acidobact
🟧 k__Bacteria, p__Acidobacteria, c__DA052, o

Figure 3.8: *Sunburst view of the microbial community zoomed to show all taxa belonging to the phylum bacteroidetes.*

Chapter 4

Estimation of Alpha and Beta diversity

Two levels of diversity are typically considered in microbial ecology: alpha- and beta-diversity. Alpha-diversity estimates describe the number of species (or similar metrics) in a single sample, whereas beta-diversity compares the number of species (or similar metrics) across samples [Whittaker, 1972].

Some measures of estimate alpha and beta diversity require a phylogenetic tree of all OTUs (Phylogenetic diversity and UniFrac distances). The phylogenetic tree is reconstructed based on a multiple sequence alignment (MSA) of the OTU sequences. Therefore, as a pre-requisite, a MSA needs to be created and a phylogeny reconstructed. Note that by default only the top 100 most abundant OTUs are aligned using MUSCLE and used to reconstruct the phylogeny tree in the next step. This phylogenetic tree is used for calculating the phylogenetic diversity and the UniFrac distances, so these measures disregard 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.

4.1 Align OTUs with MUSCLE

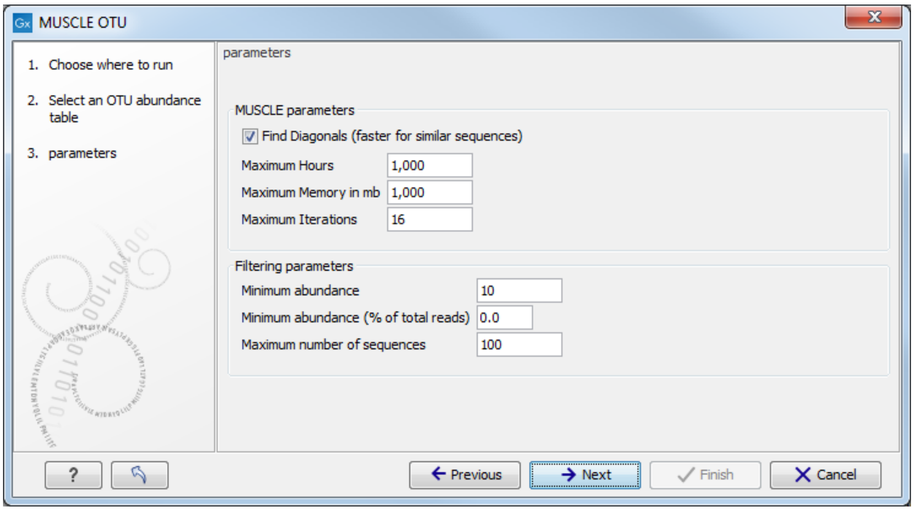
In order to estimate Alpha and Beta diversity, you must use first use the Align OTUs with MUSCLE tool of the Microbial Genomics Module:

Toolbox | OTUclustering ( ) | Align OTUs using MUSCLE ( ).

Choose the OTU abundance table as input. the next wizard window allows you to set up the alignment parameters with MUSCLE (figure 4.1).

- **Find Diagonals**: you can decide on some restrictive parameters for your analysis: the **Maximum Hours** the analysis should last, the **Maximum Memory in mb** that should be used for the analysis, or the **Maximum Iterations** the analysis should make. The latter is set to 16 by default.

- **Filtering Parameters**: The algorithm filters out all OTUs whose combined abundance across all samples is less than the **minimum combined abundance** or whose combined abundance is less than the **minimum combined abundance (% of all the reads)** across all samples. The default value for the Minimum combined abundance is set at 10. Moreover, you can specify the **Maximum number of sequences to be aligned**, so that only the sequences with the highest combined abundances will be used. Note that reducing the number of



MUSCLE OTU

1. Choose where to run
2. Select an OTU abundance table
3. parameters

parameters

MUSCLE parameters

☑ Find Diagonals (faster for similar sequences)

Maximum Hours: 1,000

Maximum Memory in mb: 1,000

Maximum Iterations: 16

Filtering parameters

Minimum abundance: 10

Minimum abundance (% of total reads): 0.0

Maximum number of sequences: 100

Figure 4.1: *Set up parameters for aligning sequences with MUSCLE.*

sequences will speed up the alignment and the construction of phylogeny trees.

For further analysis with the Alpha and Beta diversity tools, save the alignment and construct a phylogeny using the Maximum Likelihood Phylogeny tool from CLC Workbench core tools in **Toolbox** | **Classical Sequence Analysis ()** | **Alignments and Trees ()** | **Maximum Likelihood Phylogeny ()**. For more information, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Maximum_Likelihood_Phylogeny.html.

4.2 Alpha Diversity

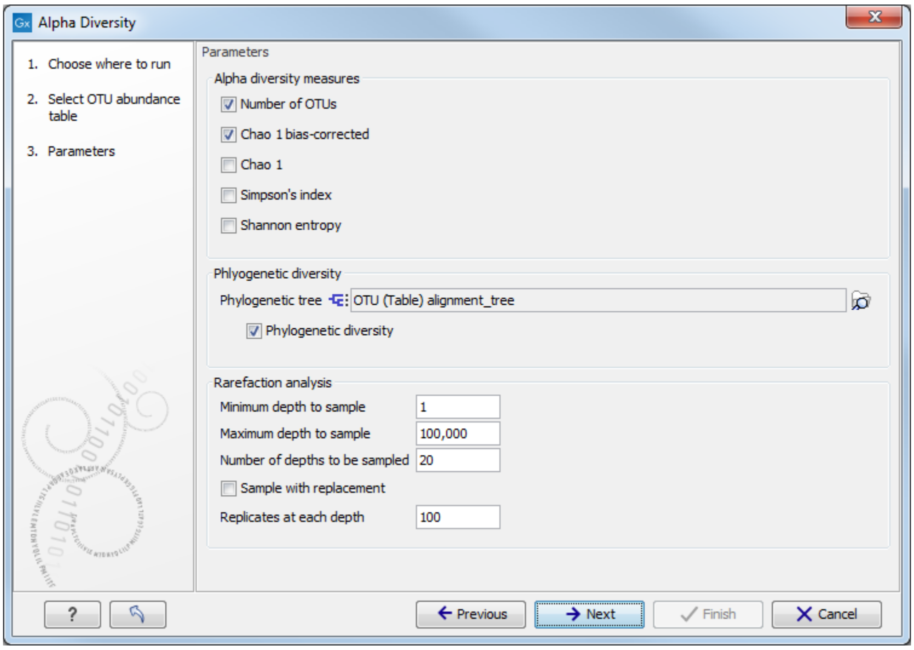
Alpha diversity is the diversity within a particular area or ecosystem; usually expressed by the number of species (i.e., species richness) in that ecosystem. Alpha diversity estimates are dependent on sampling depth, and hence rarefaction analysis is integral to this analysis step.

To run the tool go to **Toolbox** | **OTUClustering ()** | **Alpha Diversity ()**. Choose an OTU table to use as input. The next wizard window offers you to set up different analysis parameters (figure 4.2). For example, you can select which diversity measures to calculate (see section 4.2.1), specify the phylogenetic tree for computing phylogenetic diversity (you can use the tree reconstructed in the previous step), and parameterize the rarefaction analysis.

The rarefaction analysis is done by sub-sampling the OTU abundances in the different samples at different depths. The range of depths to be sampled is defined by the parameters **Minimum depth to sample** and **Maximum depth to sample**. If the maximum depth is set to 0, the number of reads of the most abundant sample is used. The number of different depths to be sampled is specified by the **Numbers of depths to be sampled parameters**. For example, if you choose to sample 5 depths between 1000 and 5000, the algorithm will sub-sample each sample at 1000, 2000, 3000, 4000, and 5000 reads. At each depth, the algorithm subsamples the data several times, according to the **Replicates at each depth**. You can choose whether the sampling should be performed with or without replacement by setting the **Sample with replacement** parameter.

4.2.1 Alpha diversity measures

The available diversity measures are:



Parameters

Alpha diversity measures

☑ Number of OTUs

☑ Chao 1 bias-corrected

□ Chao 1

□ Simpson's index

□ Shannon entropy

Phlyogenetic diversity

Phylogenetic tree: OTU (Table) alignment_tree

☑ Phylogenetic diversity

Rarefaction analysis

Minimum depth to sample	1
Maximum depth to sample	100,000
Number of depths to be sampled	20
□ Sample with replacement	
Replicates at each depth	100

Figure 4.2: *Set up parameters for the Alpha Diversity tool.*

- Number of OTUs: The number of OTUs observed in the sample.
- Chao 1 bias-corrected: Chao1-bc $= D + \frac{f_1(f_1-1)}{2(f_2+1)}$.
- Chao 1: Chao1 $= D + \frac{f_1^2}{2f_2}$.
- Simpson's index: SI $= 1 - \sum_{i=1}^{n} p_i^2$.
- Shannon entropy: H $= \sum_{i=1}^{n} p_i \log_2 p_i$.

where n is the number of OTUs; D is the number of distinct OTUs observed in the sample; f_1 is the number of OTUs for which only one read has been found in the sample; f_2 is the number of OTUs for which two reads have been found in the sample; and p_i is the fraction of reads that belong to OTU i.

If a phylogenetic tree is provided as input, the following distance is also available:

- Phylogenetic diversity: $PD = \sum_{i=1}^{n} b_i I(p_i > 0)$

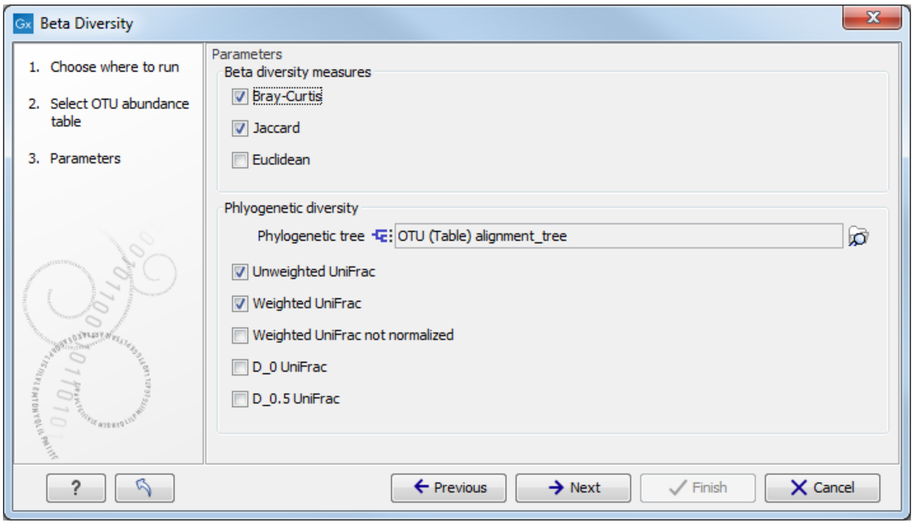
where n is the number of branches in the phylogenetic tree, b_i is the length of branch i; p_i is the proportion of taxa descending from branch i; and the indicator function $I(p_i > 0)$ and $I(p_i^B > 0)$ assumes the value of 1 if any taxa descending from branch i is present in the sample or 0 otherwise.

4.3 Beta Diversity

Beta diversity examines the change in species diversity between ecosystems. The analysis is done in two steps. First, the tool estimates a distance between each pair of samples (see the list of available distances below). Once the distance matrix is calculated, the beta diversity analysis tool performs Principal Coordinate Analysis (PCoA) on the distance matrices. These can be visualized by selecting the PCoA icon () in the bottom of the Beta Diversity results ().

To run the tool, open **Toolbox** | **OTUclustering ()** | **Beta Diversity ()** and select an OTU abundance table before clicking on the button labeled Next.

The following wizard window is shown in figure 4.3.



Beta Diversity

1. Choose where to run
2. Select OTU abundance table
3. Parameters

Parameters

Beta diversity measures

☑ Bray-Curtis

☑ Jaccard

□ Euclidean

Phylogenetic diversity

Phylogenetic tree: OTU (Table) alignment_tree

☑ Unweighted UniFrac

☑ Weighted UniFrac

□ Weighted UniFrac not normalized

□ D_0 UniFrac

□ D_0.5 UniFrac

Figure 4.3: *Set up parameters for the Alpha diversity tool.*

As in section 4.2, you must have previously aligned the OTUs and constructed a phylogeny that can be used as input in in the Beta Diversity tool.

4.3.1 Beta diversity measures

The following beta diversity measures are available:

- Bray-Curtis: $B = \frac{\sum_{i=1}^{n} |x_i^A - x_i^B|}{\sum_{i=1}^{n} (x_i^A - x_i^B)}$

- Jaccard: $J = 1 - \frac{\sum_{i=1}^{n} \min(x_i^A, x_i^B)}{\sum_{i=1}^{n} \max(x_i^A, x_i^B)}$

- Euclidean: $E = \sum_{i=1}^{n} \sqrt{(x_i^A - x_i^B)^2}$

where n is the number of OTUs and x_i^A and x_i^B are the abundances of OTU i in samples A and B, respectively.

If a phylogenetic tree is provided as input, the following distances are also available:

- Unweighted UniFrac: $d^{(U)} = \frac{\sum_{i=1}^{n} b_i |I(p_i^A > 0) - I(p_i^B > 0)|}{\sum_{i=1}^{n} b_i}$

- Weighted UniFrac: $d^{(W)} = \frac{\sum\limits_{i=1}^{n} b_i \left|p_i^A - p_i^B\right|}{\sum\limits_{i=1}^{n} b_i (p_i^A + p_i^B)}$
- Weighted UniFrac not normalized: $d^{(w)} = \sum\limits_{i=1}^{n} b_i \left|p_i^A - p_i^B\right|$
- D_0 UniFrac: The generalized UniFrac distance $d^{(0)} = \frac{\sum\limits_{i=1}^{n} b_i \left|\frac{p_i^A - p_i^B}{p_i^A + p_i^B}\right|}{\sum\limits_{i=1}^{n} b_i}$
- D_0.5 UniFrac: The generalized UniFrac distance $d^{(0.5)} = \frac{\sum\limits_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B} \left|\frac{p_i^A - p_i^B}{p_i^A + p_i^B}\right|}{\sum\limits_{i=1}^{n} b_i \sqrt{p_i^A + p_i^B}}$

where n is the number of branches in the phylogenetic tree, b_i is the length of branch i; p_i^A and p_i^B are the proportion of taxa descending from branch i for samples A and B, respectively; and the indicator functions $I(p_i^A > 0)$ and $I(p_i^B > 0)$ assume the value of 1 if any taxa descending from branch i is present is samples A and B, respectively, or 0 otherwise. The unweighted UniFrac distance gives comparatively more importance to rare lineages, while the weighted UniFrac distance gives more important to abundant lineages. The generalized UniFrac distance $d^{(0.5)}$ offers a robust tradeoff [Chen et al., 2012].

4.4 PERMANOVA Analysis

PERMANOVA Analysis (PERmutational Multivariate ANalysis Of VAriance, also known as non-parametric MANOVA [Anderson, 2001]), can be used to measure effect size and significance on beta diversity for a grouping variable. For example, it can be used to show whether OTU abundance profiles of replicate samples taken from different locations vary significantly according to the location or not. The significance is obtained by a permutation test.

To perform a PERMANOVA analysis, go to:

Toolbox | OTUclustering ( ) | PERMANOVA Analysis ( ).

Choose an OTU abundance table as input. In the next wizard window you can specify the phylogenetic tree reconstructed from the alignment of the most abundant OTUs in the previous step and select the beta-diversity and the phylogenetic diversity measures you wish to use for this analysis (see section 4.3.1 for definitions).

The output of the analysis is a report which contains 2 tables for each beta diversity measure used:

- A table showing the metadata variable used, its groups and the results of the test (pseudo-f-statistic and p-value)
- A PERMANOVA analysis for each pair of groups and the results of the test (pseudo-f-statistic and p-value). Bonferroni-corrected p-values (which correct for multiple testing) are also shown.

4.5 Convert to Experiment

This tool takes an OTU abundance table as input and assigns categories to the samples, allowing users to use the thus-generated table to perform statistical tests. To use the tool, go to:

Toolbox | **OTUclustering ()** | **Convert to Experiment ()**

Choose an OTU table as input, and define which metadata group is to be consider as factors.

The tool output is a table labeled (experiment). The first column is the name of the group used as factor in the analysis. For each OTUs there will be the following data

- Range
- IQR
- Difference
- Fold change
- Taxonomy
- Expression values and Means for each site

You can create subexperiment by selecting only a some of the OTUs from your experiment table (the most abundant ones for example).

Once you have created your experiment table, you can perform several statistical analyses with the following tools of the CLC Genomics Workbench:

Toolbox | **Transcriptomics Analysis ()** | **Statistical Analysis ()** | **Empirical Analysis of DGE ()**

For more information about Empirical analysis of DGE, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Empirical_analysis_DGE.html

Toolbox | **Transcriptomics Analysis ()** | **Quality control ()** | **Hierarchical Clustering of Samples ()**

For more information about Hierarchical Clustering of Samples, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Result_hierarchical_clustering samples.html

Toolbox | **Transcriptomics Analysis ()** | **Feature Clustering ()** | **Hierarchical Clustering of Features ()**

For more information about Hierarchical Clustering of Features, see http://www.clcsupport.com/clcgenomicsworkbench/current/index.php?manual=Hierarchical_clustering_features.html

4.6 Importing OTU abundance tables

It is possible to import a csv or excel file as an OTU abundance table, by going to **File** | **Import ()** | **Standard Import... ()** and force the input as type "OTU abundance table(.xls, .xlsx, .csv)".

This importer allow to perform statistical analyses on abundance tables that were not generated by OTUclustering tool.

For example, Terminal Restriction Fragment Length Polymorphism (TRFLP) data can be imported and treated in the same way as OTU abundance tables. However, all sequence-based actions cannot obviously be applied to this data (i.e. multiple sequence alignment, tree reconstruction and Phylogenetic tree measure estimation).

Chapter 5

Workflows

In the CLC Genomics Workbench, you can link tools to one another to be processed in sequential order enabling repeated execution of a workflow. Working with workflows is described in detail in http://www.clcbio.com/files/tutorials/Workflow-intro.pdf.

The Microbial Genomics Module contains two workflows that you can start here:

Toolbox | **Workflows** | **Data QC and OTU clustering** or **Estimate Alpha and Beta Diversities**.

They require that you provide the necessary input files and edit the parameter settings, and will output all relevant analysis results. As many secondary analyses require metadata, this assignment of OTU table and metadata has to be done between execution of the two workflows as described in 3.6.2.

5.1 Data QC and OTU clustering workflow

The **Data QC and OTU clustering** workflow consists of 5 tools being executed sequentially (figure 5.1). The only necessary input to run the workflow are the reads you want to cluster. You also have the option to provide a list of the primers that were used to sequence these reads if you wish to perform the adapters trimming step.

The first tool is the **Optional Merge Paired Reads** that will output 2 sets of sequences, the merged reads and the not paired reads. Both will be used as input in the **Trim Sequences** tool together with the primer list. This tool provides a list of trimmed sequences that will be the input of the **Fixed Length Trimming** tool. Again, the output is a list of trimmed sequences, used as input file for the **Filter Samples Based on the Number of Reads** tool. The results of the filter are complied in a report and the tool generates a sequence list that does not contain the reads of poor quality. This filtered list will be used for the final tool of the workflow, the **OTUclustering** tool. This tool will give 2 outputs: a sequence list of the OTU centroids and an abundance table with the newly created OTUs, their abundance at each site as well as the total abundance for all samples.

5.2 Estimate Alpha and Beta Diversities workflow

The **Estimate Alpha and Beta Diversities** workflow consists of 5 tools and requires only the OTU table as input file (figure 5.2). Remember to add metadata to your output table before starting

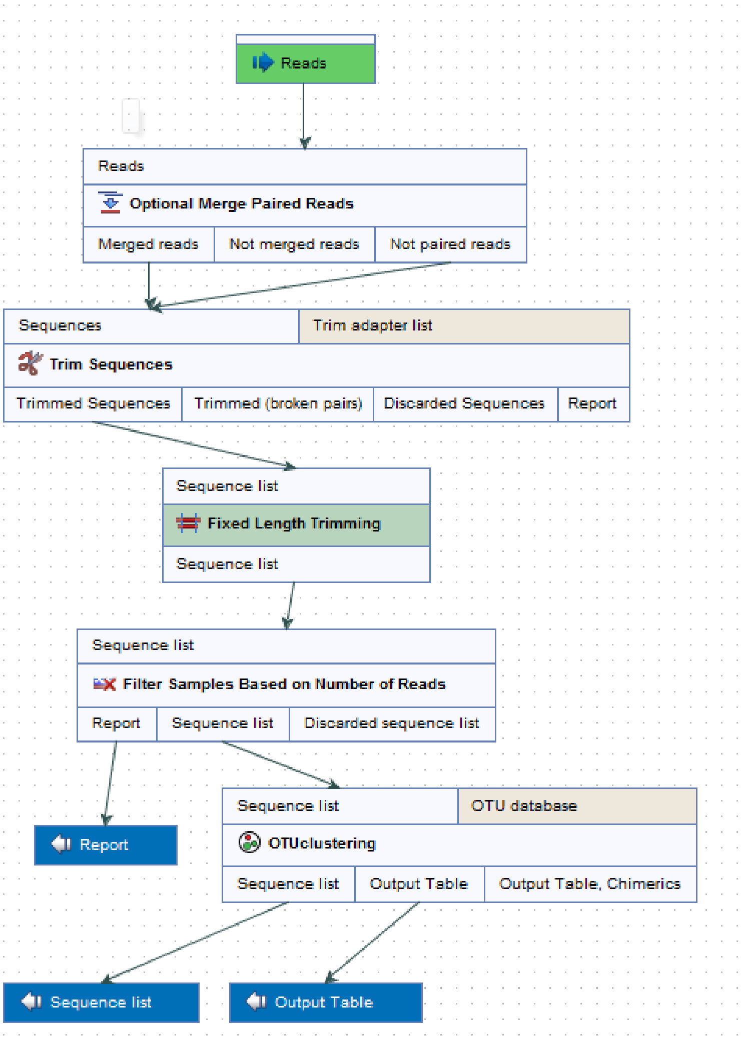


Figure 5.1: *Layout of the Data QC and OTU clustering workflow.*

the workflow.

The first tool of the workflow is the Remove OTUS with Low Abundance. The output is a reduced abundance table that will be used as input for 3 other tools:

- **Align OTUs with MUSCLE**, a tool that will produce an alignment used to reconstruct a Maximum Likelihood Phylogeny, which will in turn output a phylogenetic tree also used as input in the following 2 tools.
- **Alpha diversity** tool
- **Beta diversity** tool

Running this workflow will therefore give 4 outputs: an alignment of the OTUs, a phylogenetic tree of the OTUs, a diversity report for the alpha diversity and a PCoA for the beta diversity.

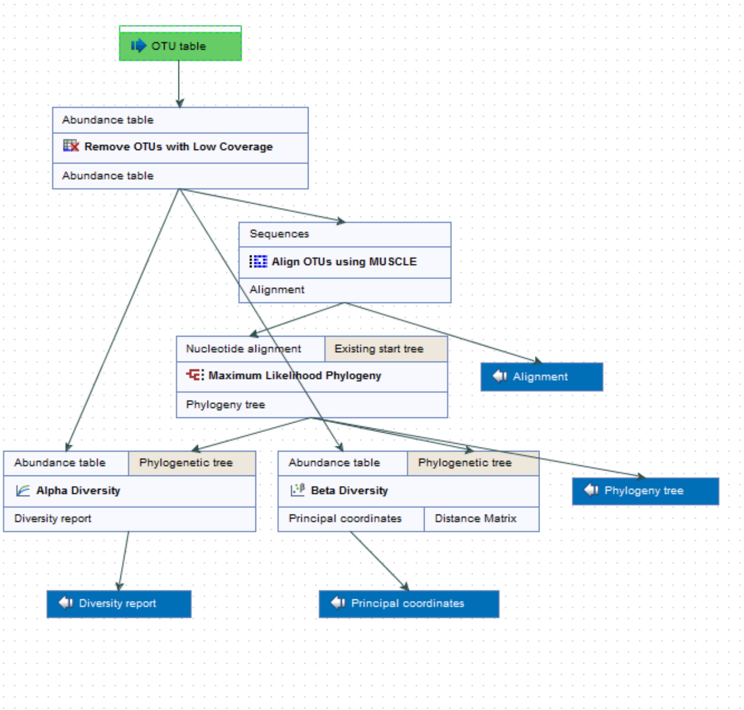


Figure 5.2: *Layout of the Alpha and Beta Diversities workflow.*

Bibliography

[Anderson, 2001] Anderson, M. (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology*, 26(1):32–46.

[Chen et al., 2012] Chen, J., Bittinger, K., Charlson, E. S., Hoffmann, C., Lewis, J., Wu, G. D., Collman, R. G., Bushman, F. D., and Li, H. (2012). Associating microbiome composition with environmental covariates using generalized unifrac distances. *Bioinformatics*, 28(16):2106–13.

[Whittaker, 1972] Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon*, pages 213–251.

Index

Alpha diversity, 21
Alpha diversity measures, 21

Beta diversity, 22
Beta diversity measures, 23
Bibliography, 30

Contact information, 6

Download OTU Reference Database, 10

Filtering, remove OTUs with low abundance, 16
Filtering, remove samples with low coverage, 12
Fixed length trimming, 12
Format reference database, 10

Importing OTU abundance tables, 25
Installation, 8

Metadata, import, 15

OTU clustering, 13
Overview, 5

Permanova, 24

rarefaction analysis, 21
References, 30

Uninstallation, 8

Visualization, abundance table, 17
Visualization, hierarchical clustering of samples, 24