

## **OTU Clustering Using Workflows**

March 1, 2022

Sample to Insight –

-

QIAGEN Aarhus  $\cdot$  Silkeborgvej 2  $\cdot$  Prismet  $\cdot$  8000 Aarhus C  $\cdot$  Denmark digitalinsights.qiagen.com  $\cdot$  ts-bioinformatics@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 Taxonomic 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 20.0, 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: <a href="http://resources.giagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html">http://resources.giagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html</a>.

#### **General tips**

- Tools can be launched from the Workbench Toolbox, as described in this tutorial, or alternatively, click on the Launch button (*(*) in the toolbar and use the Quick Launch tool to find and launch tools.
- Within wizard windows you can use the **Reset** button to change settings to their default values.
- You can access the in-built manual by clicking on **Help** buttons or going to the "Help" menu and choosing "Plugin Help" | "CLC Microbial Genomics Module Help".

**Downloading 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

- Download the sample data from our website: http://resources.giagenbioinformatics. 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 suspect's 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).

🐼 Illumina High-Throughput	Sequencing Import			
1. Choose where to run	Select files of types Illumina (.txt/.fastq/.fq) Selected Files (24)			
2. Import files and options	C:\Users\Documents\ptudustering_tutorial\BootA-replicateA_R1.fastq C:\Users\Documents\ptudustering_tutorial\BootA-replicateA_R2.fastq			
3. Result handling	C: Users \Documents \otudustering_tutorial \BootA-replicateB_R1.fastq			
<ol> <li>Save location for new elements</li> </ol>	C: Users Documents butuatiseting, tutorial BootA-replicate: Q. A.Lastq C: Users Documents butuatiseting, tutorial BootA-replicate: Q.R.Lastq C: Users Documents butuatiseting, tutorial BootA-replicate: Q.R.Lastq C: Users Documents butuatiseting, tutorial BootA-replicate: Q.R.Lastq			
	<			
	Add folders Add files Remove			
	General options Paired reads Discard read names Otiscard quality scores Minimum distance 200 Maximum distance 550			
	Illumina options Remove failed reads Quality scores NCBI/Sanger or Illumina Pipeline 1.8 and later ↓ 			
Help Reset	Previous Next Finish Cancel			

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**.

1. Select spreadsheet	Selection of metadata to impo Spreadsheet with metadata	rt		
	C:\Users\ Desktop\MetadataRoundRobin.xlsx			
	Metadata preview			
	Name	Type		
	CrimeSite 1-replicateA	1		
	CrimeSite 1-replicateB	1	_	
	Site2-replicateA	2		
NO NO	Site2-replicateB	2		
, NED	Site3-replicateA	3		
61	Site3-replicateB	3		-
mana	BootA-replicateA	BootA		
All and and a state of the stat	BootA-replicateB	BootA		
11	BootA-replicateC	BootA	_	
100 J	BootB-replicateA	BootB		
T C	BootB-replicateB	BootB		
All PALENTO .	BootB-replicateC	BootB		

Figure 2: Import the metadata spreadsheet.

- 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.
- 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 "**Prefix**" name matching. This results in successful associations for all data items (figure 3). Click **Next**.

Gx Import Metadata			×		
1. Select spreadsheet	Association setup Optional step. Data can be associated later instead using the Associate Data tools available from within the Metadata Table Editor.				
2. Associate with data	Data to associate				
3. Save in folder	Save in folder     Location of data ﷺ Selected 12 elements.     Matching scheme     Exact - data element names must match a key exactly to be associated     Prefix - data elements with names starting with a matching key will be associated     Suffix - data elements with names ending with a matching key will be associated				
Data association preview 12 data elements were selected for association with 12 metadata rows. 12 data elements will have metada associations added.					
	Restà replicateà D1 (paired)	PeetA replicateA	Pasta		
	BootA-replicateA_R1 (paired)	BootA-replicateR	BootA		
	BootA-replicateS_R1 (paired)	(paired) BootA-replicateD BootA			
	PootR replicateC_R1 (pared)	BootR replicated	Reat R		
	BootB-replicateB_P1 (paired)	BootB-replicateB	BootB		
	BootB-replicateC_P1 (paired)	BootB-replicateC	BootB		
	CrimeSite1-replicateA_R1 (paired)	CrimeSite1-replicateA	1		
	CrimeSite1-replicateB_B1 (paired)	CrimeSite 1-replicateB	1		
	Site2-replicateA_R1 (paired)	Site2-replicateA	2		
	Site2-replicateB_B1 (paired)	Site2-replicateB	2		
	Site3-replicateA_R1 (paired)	Site3-replicateA 3			
Site3-replicateB R1 (paired) Site3-replicateB 3					
Help Rese	t	Previous	Next Finish Cancel		

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 Template Workflows | Microbial Workflows (a) | Metagenomics (a) | Amplicon-Based Analysis (a) | 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**. 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 database 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.





Figure 5: 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, which consists of 5 tools as seen on figure 6. These tools make use of the metadata imported earlier.

39e			
age			
290			
rage			
Convenience			
aequences			
Align Olus using MUSC	LE		
Alignment			
· · · · · · · · · · · · · · · · · · ·			
· · · · · · · · · · · · · · · · · · ·			
Existing start	tree		
3			
m Likelihood Phylogeny			
Abundanas tabla	Phylogenetic tree		
Abundance table	Phylogenetic tree		
118 Pate Diversity			Division
Beta Diversity			Phylogeny tree
Dringing Langeding to a	Distance Matrix		
Principal coordinates	Distance Matrix		
	A		
🖓 Principal	coordinates		
	Alignment Existing start  Ilignment Existing start  Likelihood Phylogeny  ee Abundance table  Deta Diversity Principal coordinates  I Principal	Alignment Lignment Existing start tree m Likelitood Phylogeny ee Abundance table Phylogenetic tree Life Beta Diversity Principal coordinates Distance Matrix	Alignment  Existing start tree  Lignment  Existing start tree  LikeNtood Phylogeny  e  Abundance table Phylogenetic tree  Principal coordinates Distance Matrix  Principal coordinates

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

**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".

- 1. Launch the workflow Template Workflows | Microbial Workflows (a) | Metagenomics (a) | Amplicon-Based Analysis (a) | 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.

Estimate Alpha and Beta Diversities
 CTU (Table) (Filtered) alignment\_tree
 OTU (Table) (Filtered) (PCoA - D\_0.5 UniFrac)
 OTU (Table) (Filtered) (Alpha Diversity - Total number)

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 (**11**) as well as Zoomable Sunbursts (**0**).

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

- 1. Select Metagenomics (a) | Amplicon-Based Analysis (a) | Remove OTUs with Low Abundance (a)
- 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 Boot A and pink lines for BootB.

The lines do not plateau, indicating that we would need more samples to reach a definite conclusion, but Boot A samples seem to have similar measures of alpha diversity as the sites 1 and 2 while BootB samples appear to be more distinct from the other sites when it comes to





Figure 8: Aggregate samples based on metadata information.

alpha diversity measures. We suspect that Mr. X was not wearing his boot B at any of the sites sampled.



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

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 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 Metagenomics | Abundance Analysis (
  ) | Differential Abundance Analysis.
- 2. Choose OTU (Table) as input.





Figure 10: Result of the beta diversity analysis.

- Choose Type as Metadata factor and Across groups (ANOVA-like) as Comparisons (see figure 11). Click Next.
- 4. Save the result in the tutorial folder.

Differential Abundance Analy     Experimental design and     comparisons	sis Experimental design and comparisons	
2. Result handling	Experimental design Abundance table  CTU (Table) Metadata factor Type Correct for (Nothing selected)	ରି କ
	Comparisons  Across groups (ANOVA-like)  All group pairs  Against control group  Among comparison groups: Selected 5 elements.  Control group  BootA	] +
Help Reset	Previous Next Finish Car	cel

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 **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**.

Gx Create Heat Map for Abund	Jance Table
1. Choose where to run	Set filtering
2. Select an abundance table	Filter settings Specify features
3. Set clustering	Keep fixed number of features
4. Set filtering	Fixed number of features 25
5. Result handling	Minimum counts in at least one sample 10
6. Save location for new elements	Specify abundance table Abundance table for filtering
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	Specify features Features Features f_Sinobacteraceae, 150982 f_Rhodospirillaceae, 114037 c_Ellin6529, 112867 o_Ellin6513, 112828 g_Phenylobacterium, 108986 f_Koribacteraceae, 1135026 g_Candidatus Koribacter, 812560 f_Koribacteraceae, 3163059 f_Gaiellaceae, 1094433 g_Mycobacterium, 815498 o_Ellin6513, 4166998 o_Solibacterales, 4374729 f_Sinobacteraceae, 4468101 *
Help Reset	Previous Next Finish Cancel

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 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.
- 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).

Variable	Groups	Groups		Pseudo-f statistic		p-value	
Туре	BootA, BootB, 1, 2, 3			14.84554		0.00068	
Group 1	Group 2	Pseudo	-f statistic	p-value		p-value (Bonferroni)	
BootA	BootE	1	4.21139	0	20000	1.00000	
BootA	1		1.09631	0	.66667	1.00000	
BootB	1		5.49145	0	20000	1.00000	
BootA	2	:	16.60434	0	.33333	1.00000	
BootB	2	2	7.10021	0	.10000	1.00000	
1	2	1	19.97387	0	.33333	1.00000	
BootA	3		31.60816	0	.33333	1.00000	
BootB	3		16.18129	0	.10000	1.00000	
1	3		36.99953	0	.33333	1.00000	
2	3		18.18110	0	.33333	1.00000	

1 PERMANOVA analysis (D\_0.5 UniFrac)

Figure 14: Result of the PERMANOVA analysis. Values can differ slightly depending on the plugin version used.

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