QIAGEN

Tutorial

Taxonomic Profiling of Whole Shotgun Metagenomic Data

September 2, 2019

Sample to Insight

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

Taxonomic Profiling of Whole Shotgun Metagenomic Data

This tutorial will give an introduction to the use of the CLC Microbial Genomics Module Taxonomic Profiling tool on whole shotgun metagenomic data.

To demonstrate how to use the tools we will analyze a subset of the data from the publication by Willmann et al., 2015. This paper describes two healthy male subjects (S1 and S2) who were treated with the antibiotic ciprofloxacin (Cp) for 6 days. Stool samples were taken on day 0 (before treatment), the first, third and sixth (last) day of treatment, and two and 28 days after the treatment. Metagenomic shotgun sequencing was performed on all samples on an Illumina HiSeq 2000 platform using a paired-end sequencing approach with a targeted read length of 100 bp and an insert size of 180 bp. In this tutorial, we will analyze the samples that were collected before treatment (day 0), the last day of treatment (day 6) as well as 28 days after treatment (day 34). We will investigate how the composition of the gut microbiota develops over time in response to the treatment and in how far it recovers after the treatment.

Prerequisites For this tutorial, you will need CLC Genomics Workbench with CLC Microbial Genomics Module 3.5 or higher installed. How to install modules and plugins is described here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Install.html.

Overview In this tutorial we will go through the use of several different tools in order to monitor the evolution of the gut microbiota of the two subjects before, during and after a ciprofloxacin treatment.

- First, we will **import the NGS reads** from the 6 samples to the workbench and prepare them for analysis. We will also **import and create reference databases indexes** of microbial genomes and **metadata**.
- Then we will run the **Data QC and Taxonomic Profiling** workflow to build a profile of the bacteria and their abundances in each sample.
- We will then run the **Merge and Estimate Alpha and Beta Diversities** workflow in order to merge abundance profiles from each sample into a single table and measure diversity both within and between samples.
- We then look at the tables, visualizations and plots that we have created and make some interesting observations on the data.
- And finally, we create a heat map that shows how the samples cluster and how the different organism abundances correlate across the samples.

Downloading and importing the data The data for this tutorial consist of NGS data files from the Willmann et al., 2015 publication. The original files, the publication abstract and the full metadata are available directly from the workbench using the Search for Reads in SRA tool and looking for the study accession number ERP011645.

However, to ensure a reasonable analysis time for this tutorial, we provide a dataset where each sample has been reduced to a million paired-end reads, a metadata spreadsheet, and a customized reference database for identifying the composition of the gut microbiome.

1. Click on the following link or paste it into your web browser to download the tutorial data: `http://resources.qiagenbioinformatics.com/testdata/taxpro.zip`. The zip-file being downloaded is ~900MB, so depending on your internet connection, this may take a while to download.
2. Start your workbench and create a folder for storing input data and results, named for example **Profiling tutorial**.
3. Go to **Import** | **Illumina** to import the 12 sequence files (ending with "fastq") (figure 1). Make sure that:
   - The import type under General Options is set to **Paired reads**.
   - Check **Discard read names** and **Discard quality scores**.
   - **Paired-end (forward-reverse)** is selected.
   - Set the minimum distance to 50 and the Maximum distance to 500. Note that you could also use default distances at this step, because the workflow we will use later on recalculates all distances for input reads anyway.

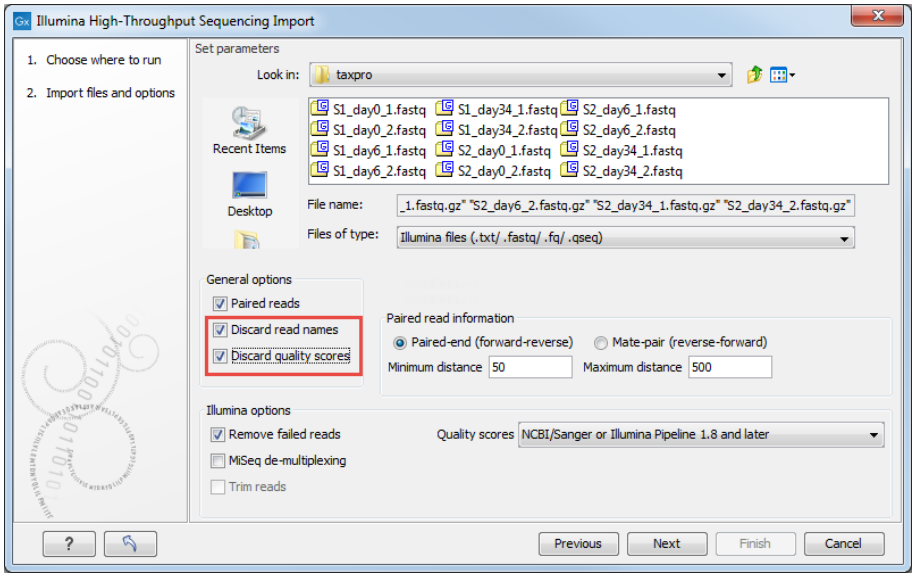

Figure 1: *Import paired-end reads for the six samples.*

4. Click **Next** and select the location where you want to store the imported sequences. You can check that you have now 6 files labeled as "paired".
5. Import the metadata by clicking **Import | Import Metadata** on top of the Navigation Area.

6. A wizard opens (figure 2). Select the spreadsheet Metadata_Willman.xlsx in the first field. The content of the Excel spreadsheet populates the table situated at the bottom of the dialog. Click **Next**.

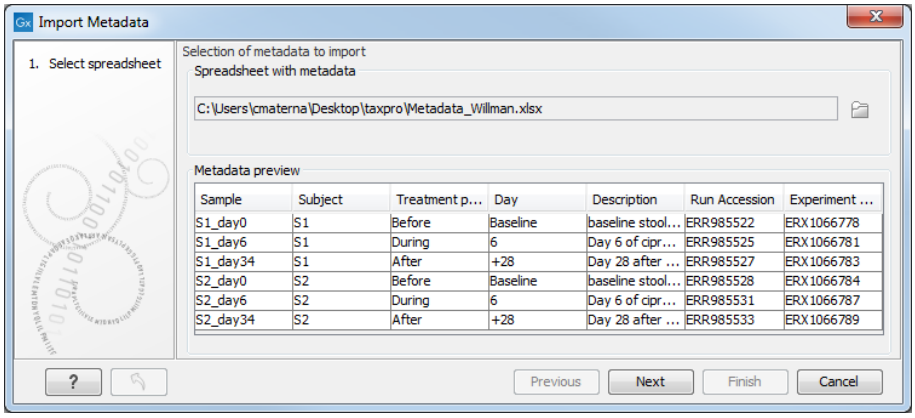

Figure 2: *Import the metadata.*

7. Click on the Navigation button next to **Location of data** (see figure 3), and select the reads you imported earlier. Click **OK**. The successful association between the data and the reads is not complete yet. Check the option **Partial** and the Data association preview will fill up, thereby confirming that association is now successful. Click **Next**.

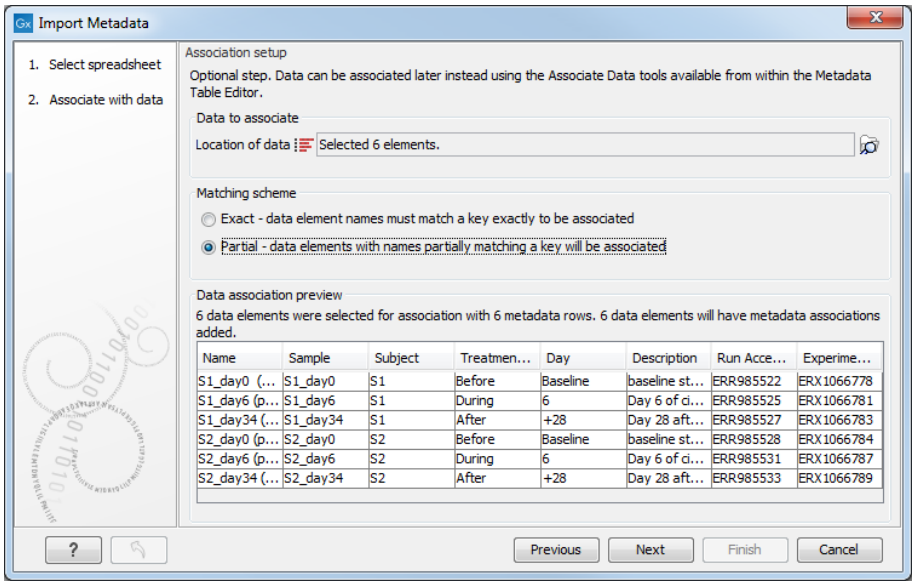

Figure 3: *The metadata and the reads are now associated.*

8. Save the metadata table in the folder created earlier (its default name is "Samples").
9. Go to **Import** | **Standard import** to import the reference data Reference database.clc. When working with your own data later on, you can select and download references from NCBI using the Download Microbial Reference Database tool.
10. Choose to **Save** it in the tutorial folder and click **Finish**.
11. Go to **Databases | Taxonomic Analysis | Create Taxonomic Profiling Index** to build an index file from the reference database.

12. Choose to **Save** the index in the tutorial folder and click **Finish**.

Taxonomic profiling

The Taxonomic Profiler tool maps each read against a reference database of complete genomes and assigns the read to a taxon in the database if a match is found. It produces an abundance table with a list of the taxons and their estimated abundances. It is associated with quality control tools in the Data QC and Taxonomic Profiling workflow, to produce a number of quality reports in addition to the abundance table.

We will demonstrate in this tutorial how to use the metadata table to select the relevant files before starting a workflow.

1. Open the metadata table called **Samples**. Select the six rows and click on **Find Associated Data**.
2. A new table called Metadata Elements opens in split view below the metadata table (figure 4). Select the six rows whose role is set to "Sample".

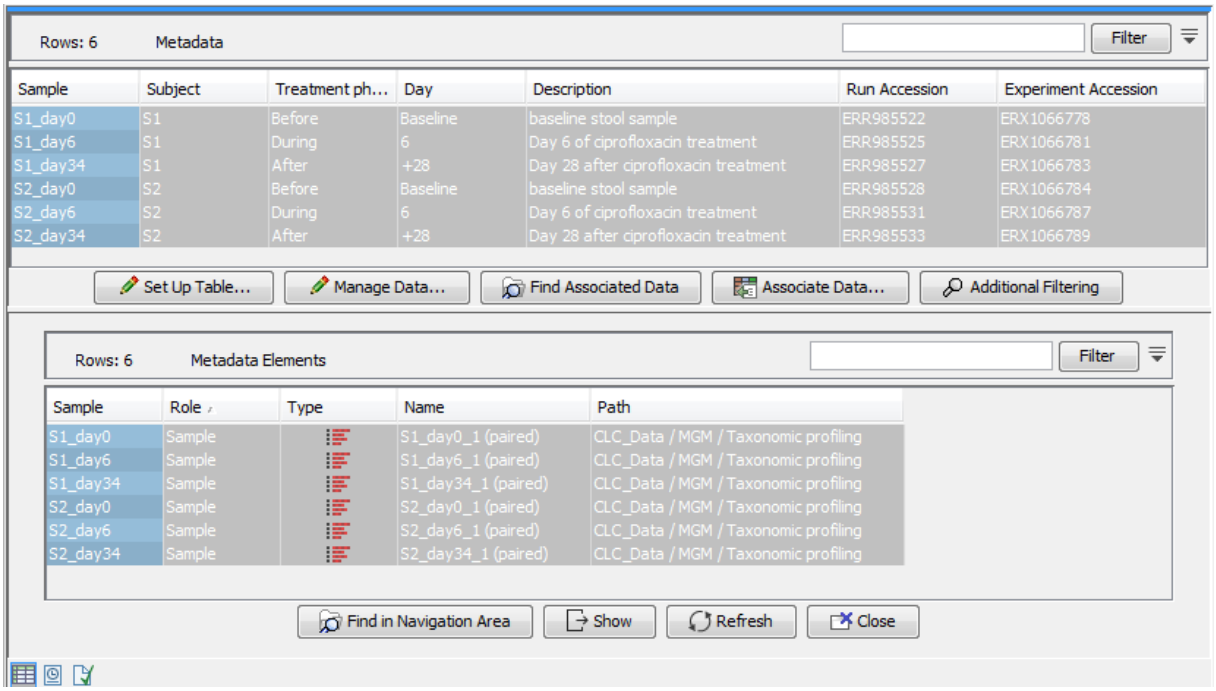

Figure 4: *Use the metadata table to find the relevant files for the workflow.*

3. Then go to the Toolbox and double-click on the workflow here:

   Metagenomics | Taxonomic Analysis | Workflows | Data QC and Taxonomic Profiling.
4. The six paired reads files to be analyzed are already selected. Check the **Batch** option (figure 5), and click **Next**.
5. The "Batch overview" dialog indicates the various batch units selected, and allow you to potentially choose which to include in the analysis. For this tutorial we have already selected all individual paired reads files needed, so you can just click **Next**.

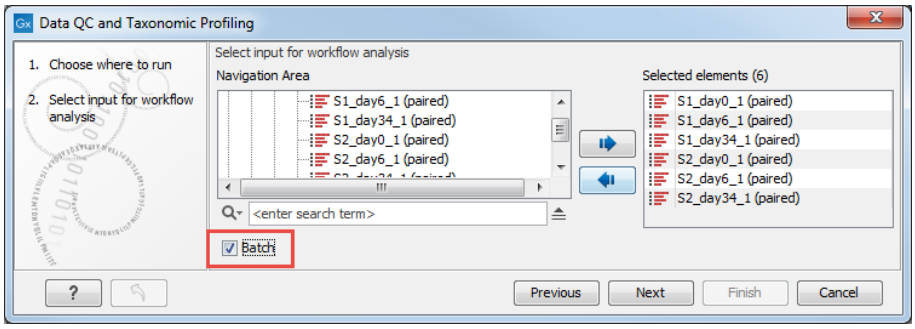

Figure 5: *Batch reads.*

6. The reads were already trimmed before they were published, so you can click **Next** to skip the "Trim Sequences" step.

7. In the next dialog, choose the list of references that you wish to map the reads against (in this case use the Reference database index file). You could also remove host DNA by specifying a reference genome index for the host. Leave the option unchecked for this tutorial (figure 6) and click **Next**.

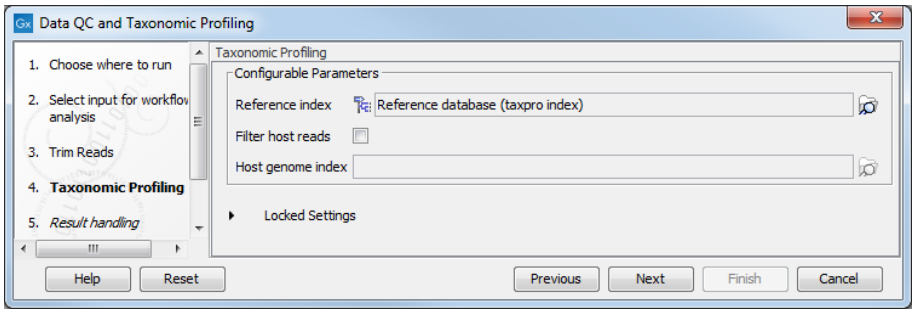

Figure 6: *Specify the reference database index. You can also check the option "Filter host reads" and specify the host genome index (in the case of human microbiota, the Homo sapiens hg38 for example). However to keep analysis time low for this tutorial, we choose not to filter.*

8. In the Result handling dialog, choose **Save in a specified location** and **Create subfolders per batch unit**.

9. Click **Next**, choose where to **Save** the results and click **Finish**.

The Data QC and Taxonomic Profiling workflow starts analyzing the specified data file, and you can follow the analysis progress in the Processes tab of the Toolbox. The first sample takes longer to analyze than the remaining samples because the reference database is being indexed and cached. The cached index is reused for the remaining samples. A completed batch unit offers the following results (figure 7):

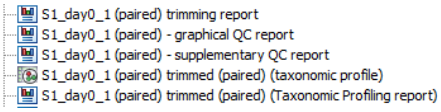

Figure 7: *Results from the Data QC and Taxonomic Profiling workflow.*

If you fail to see the 5 outputs listed in this folder, you probably forgot to check the "Discard quality scores" during the reads import step.

The results are also available from the metadata table: simply click **Refresh** under the Metadata Elements table to see them.

Merge results and statistical analyses

While it is possible to review each set of batch results one after the other, it makes more sense to merge the abundance tables into one, and to perform statistical tests on the merged table, later using metadata layers for improved visualization. You can do that with the **Merge and Estimate Alpha and Beta Diversities** workflow.

1. Once you have refreshed the Metadata Elements table, click on the header "Role" to sort elements. Select all six elements whose role is set to "Abundance table" (figure 8).

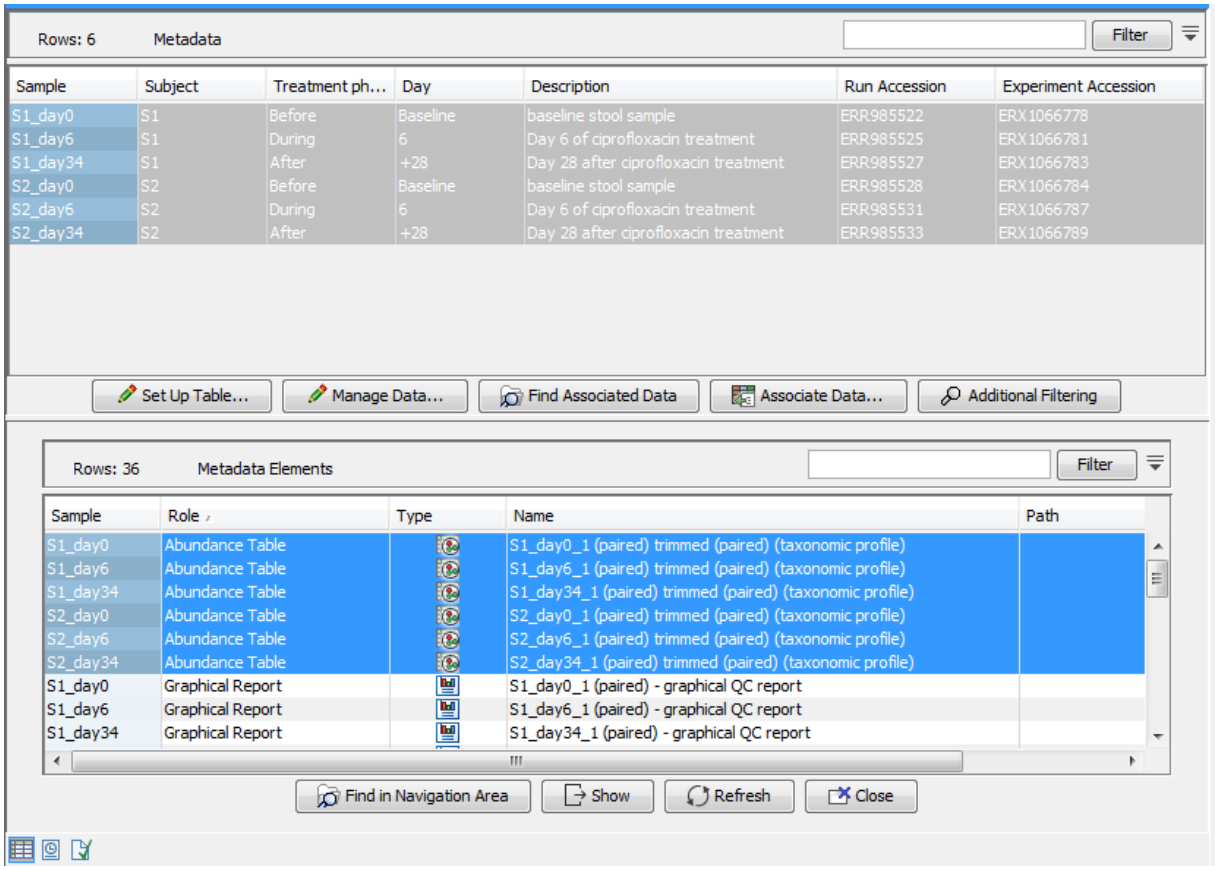

Figure 8: *Use the metadata table to find the relevant files for the workflow.*

2. Then go to the toolbox and double-click on the workflow

   Metagenomics | Taxonomic Analysis | Workflows | Merge and Estimate Alpha and Beta Diversities.

3. The six abundance tables produced by the previous workflow are already selected (figure 9) so you can just click **Next**.

4. Choose **Total number** as the parameter for the Alpha Diversity analysis and click **Next**.

5. Choose **Bray-Curtis** as the parameter for the Beta Diversity analysis and click **Next**.

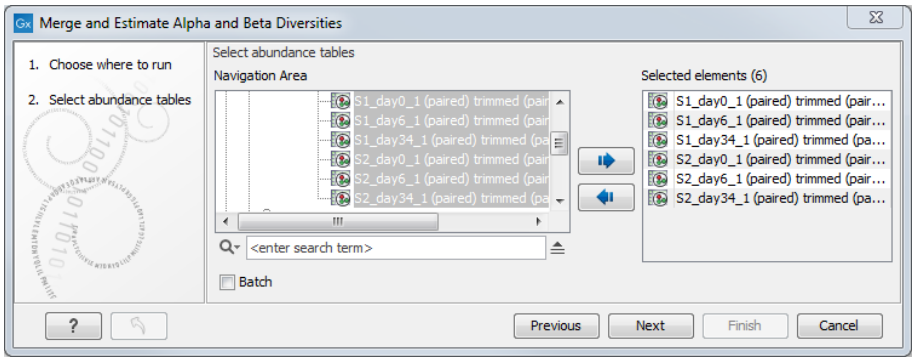

Figure 9: *Select abundance tables.*

6. Choose where to Save the workflow outputs and click **Finish**.

The Merge and Estimate Alpha and Beta Diversities workflow generates the results seen in figure 10. We will check each one of them in the next part of this tutorial.

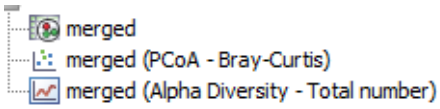

Figure 10: *Results from the Merge and Estimate Alpha and Beta Diversities workflow.*

Alpha- and beta-diversity analysis

Open the workflow output called merged (Alpha Diversity - Total number) (figure 11). Rarefaction curves plot the number of species as a function of the number of reads in a sample. The goal is to compare the diversities of different samples in relation to each other.

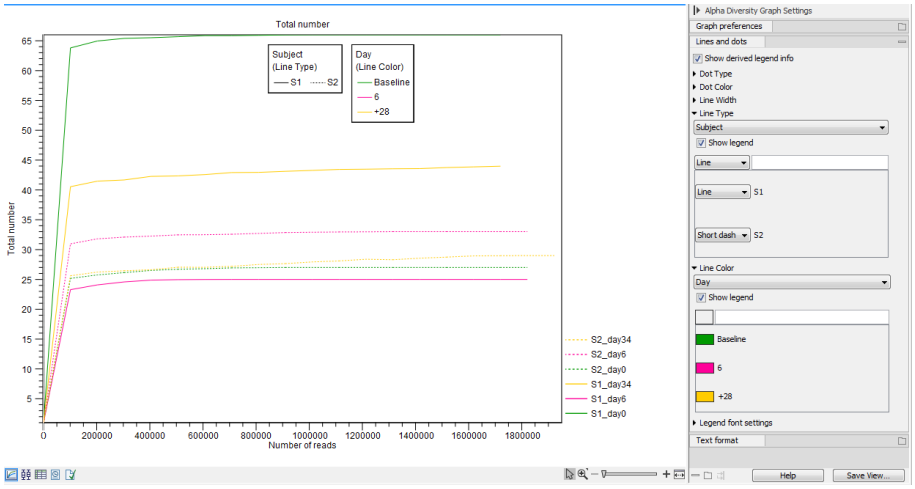

Figure 11: *Alpha diversity plot.*

You can augment the plot with metadata by setting the **Line type** and **Line color** in the "Lines and dots" palette of the side panel as shown in figure 11. From this plot, you can make the following observations:

- S1 has much higher baseline diversity than S2.

- The gut microbiome of S1 was profoundly disturbed over the course of Ciprofloxacin exposure with the lowest diversity occurring at the last day of treatment (day 6).
- S2 generally was much less affected, and diversity remained almost the same in all samples at values comparable to those for subject 1 under treatment.

Open the workflow output called merged (PCoA - Bray-Curtis) (figure 12). Change the **Sphere color** to be based on the Subject metadata column, and choose the Day metadata column as Label text for the spheres.

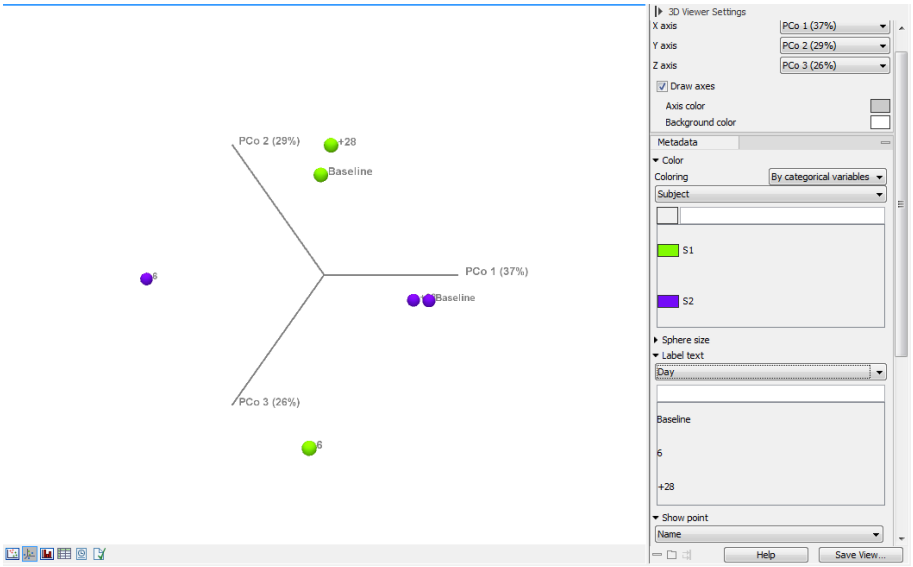

Figure 12: *Beta diversity 3D plot.*

Now the following observations can be made:

- Subject 1 and 2 are clearly separated.
- For both subjects, the baseline and the last sample time are located closely together, while the sample from the 6th day of treatment is clearly separated from these. This illustrates that treatment has affected the microbial flora, but also that recovery after treatment is observed.

Additional visualization options

Open the workflow output called "merged", which is the merged abundance table. Explore the visualization options by clicking on the **Stacked Visualization** button in the bottom left corner of the view (figure 13). Choose "Bar chart" and to "Sort samples by subject" (highlighted in red in the figure).

From the figure, we can see that the microbial diversity was higher in S1 at the beginning of the treatment than it was in S2. For both subjects, the diversity is reduced to almost exclusively Firmicutes and Bacteroidetes during treatment. After recovery, we note that the abundance of Verrucomicrobiae has increased (in green).

Click the **Show Sunburst** button in the bottom left corner of the view to investigate further the taxonomic diversity of each subject. Choose to "Aggregate samples by Subject" and play with

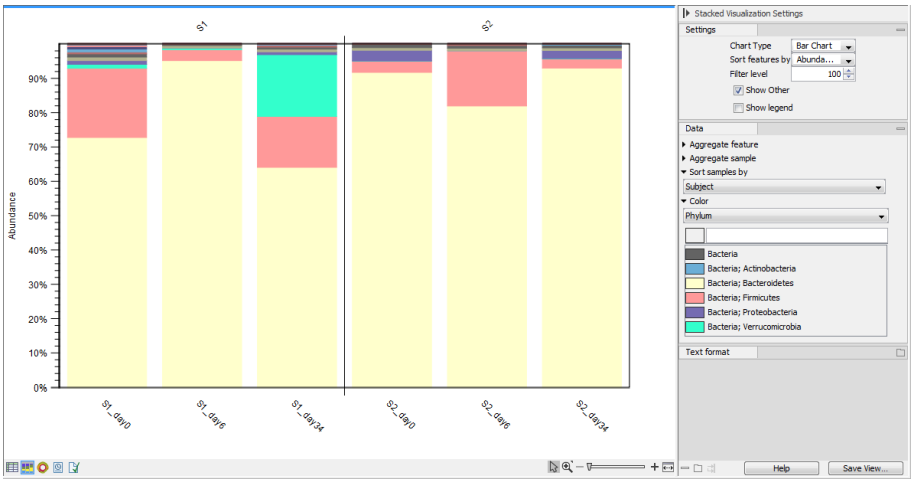

Figure 13: *Relative abundances of microbial diversity over the course of treatment.*

the number of levels to explore the taxonomy of each subject's microbiome.

Finally, we are going to visualize our results in a heat map.

1. Go back to the **table** view of the "merged" abundance table and choose to "Aggregate features" by **Class** using the drop down menu in the right hand side panel (in the Data section of the settings).
2. The resulting abundance table has 12 rows. Select them all and click **Create Abundance Subtable** below the table.
3. A new table called Merged (Filtered) opens in split view (figure 14). Save it in the Navigation Area by dragging the tab to the location you want to save it.
4. In the Toolbox, go to

   Metagenomics | Abundance Analysis () | Create Heat Map for Abundance Table
5. Select the merged (Filtered) table you just created (figure 15).
6. Set the heat map parameters as in figure 16, i.e., choosing **1-Pearson correlation** as a distance and clustering based on **Average linkage**.
7. We choose to use "No filtering" from the Filter settings drop-down menu. Click **Next**.
8. Choose to **Open** the heat map and click **Finish**.
9. On the heat map that opens in the View Area, use metadata layers as seen in figure 17.

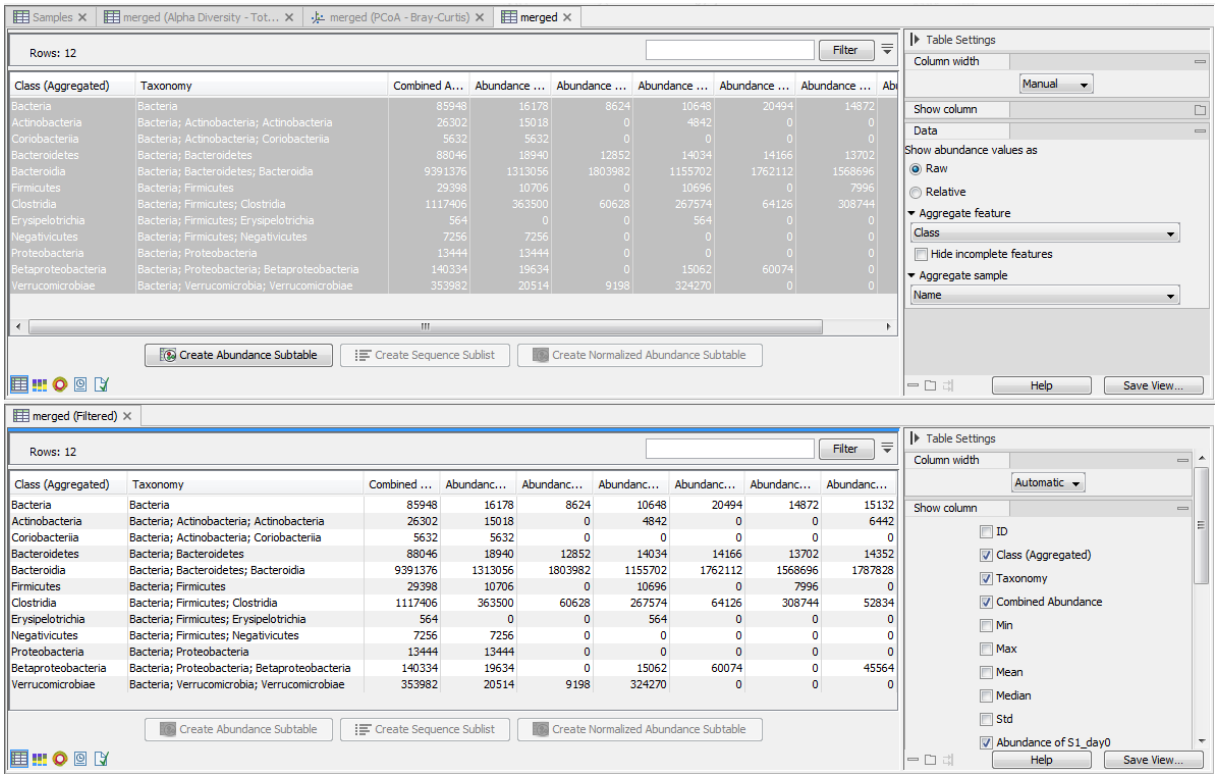

Figure 14: *Create an abundance table where features are aggregated by class.*

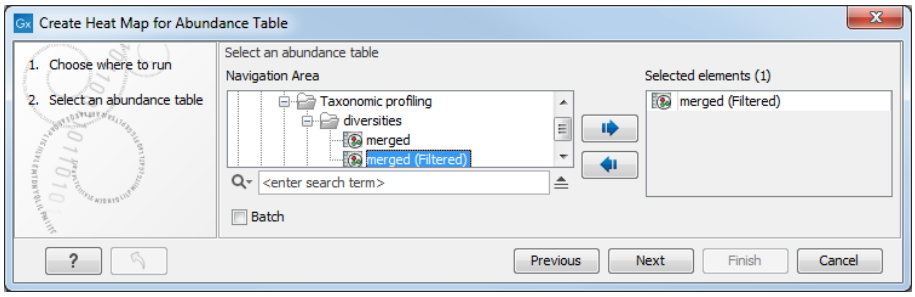

Figure 15: *Using an aggregated abundance table helps define how many features are included in the heat map.*

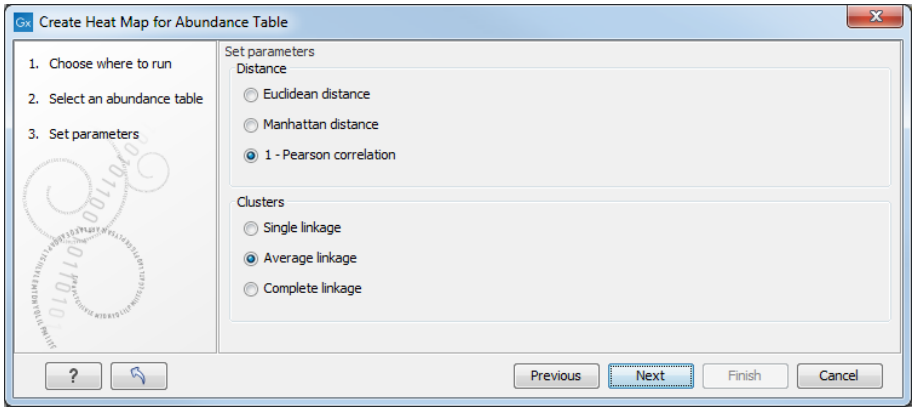

Figure 16: *Set the heat map parameters.*

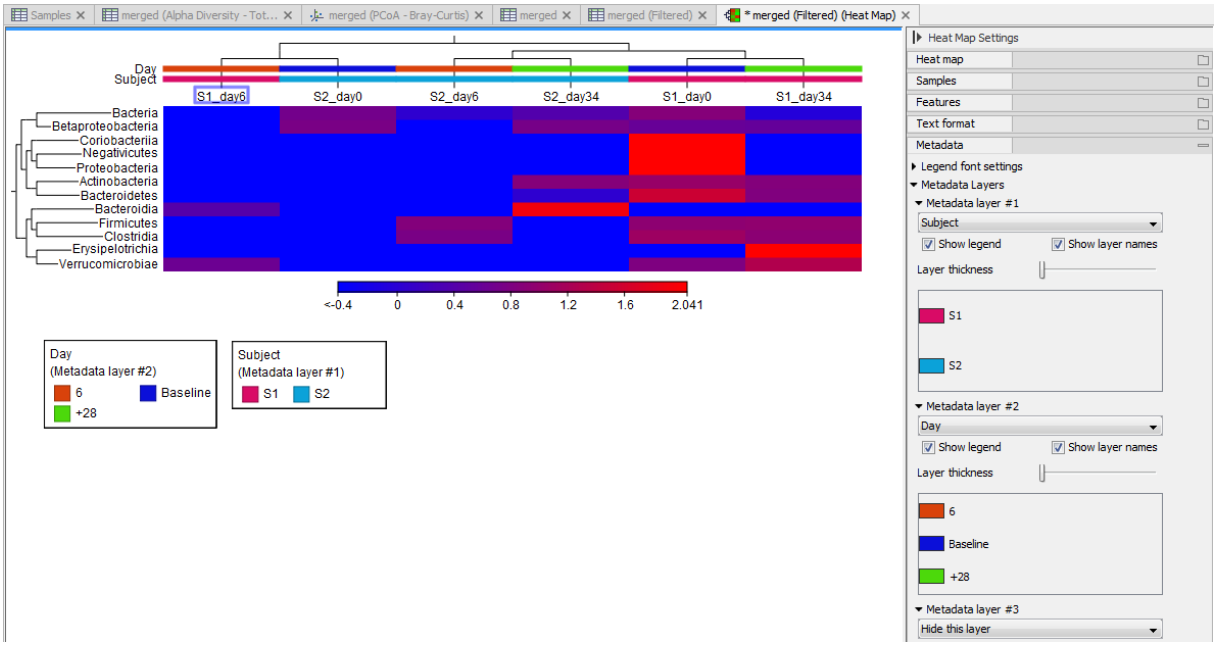

Figure 17: *Metadata layers facilitate the reading of the heat map.*

Bibliography

[Willmann et al., 2015] Willmann, M., El-Hadidi, M., Huson, D., Schuetz, M., Weidenmaier, C., Autenrieth, I., and Peter, S. (2015). Antibiotic selection pressure determination through sequence-based metagenomics. *Antimicrobial Agents and Chemotherapy*, 59(12):7335–7345. doi: 10.1128/AAC.01504-15.