Tutorial

Metatranscriptome Analysis

March 1, 2022

Sample to Insight

QIAGEN Aarhus · Silkeborgvej 2 · Prismet · 8000 Aarhus C · Denmark
digitalinsights.qiagen.com · ts-bioinformatics@qiagen.com

Tutorial

Metatranscriptome Analysis

This tutorial will take you through the tools available in CLC Microbial Genomics Module to perform a metatranscriptome functional analysis.

Introduction The main goal of the tutorial is to profile metatranscriptomes derived from two different groups of samples and the subsequent investigation of functional differences between the groups. It serves as a template for performing a comparative investigation into the functional composition and diversity of microbial communities. The tools provide a way of looking at different samples in aggregate views and to drill down into differentiating functional categories that result from the comparative analysis.

This tutorial is designed to work with metatranscriptomic data (RNA-Seq) while the "Whole metagenomic functional analysis" tutorial works with metagenomic data (DNA-Seq).

Although these tutorials conceptually have a lot in common, there are differences in some key aspects of the analysis, e.g. gene finding is not very accurate for metatranscriptomic assemblies, instead DIAMOND is used to search in a reference database. Sometimes the concepts may also be complementary and easily carried over, e.g. for the metatranscriptomic analysis we suggest a hybrid approach of mapping and assembly similar to HumanN (Abubucker S, 2012, Franzosa EA, 2018, Beghini et al., 2021), a concept that could also be used for functional metagenomic work.

Prerequisites For this tutorial, you must be working with CLC Genomics Workbench 22.0 or higher and you must have installed CLC Microbial Genomics Module 22.0 or higher.

Overview In this tutorial we will go through a suite of useful components in pipelines for analyzing metatranscriptome NGS data from microbial communities.

- First, we will import "raw" NGS sequencing data into the workbench and prepare the samples for analysis.
- We will then use the tool **Taxonomic Profiling** to find the relevant annotated genomes with curated QMI-PTDB as the reference database while removing ribosomal RNA with the SILVA database (database of rRNAs).
- We will then assemble the unclassified reads using **De Novo Assemble Metagenome** into contigs to also find transcripts of taxa not present in the reference database.
- We will map the reads to the assembled contigs and the QMI database using **Map Reads to Reference**.
- We will then build a functional profile of the samples to get abundances of GO-terms using **Build Functional Profile**.
- We will do a statistical analysis of the groups of samples using **Differential Abundance Analysis**.

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 and importing the data

For this tutorial we will make use of a dataset generated by Peng et al., 2018 that consists of samples taken from flooded rice field soil and exposed to two different temperatures (30°C and 45°C) for 30 days. The original dataset contains 30 samples, divided in two groups corresponding to the two temperatures and sequences at day 5, 11, 16, 23 and 30. The goal of the analysis in this tutorial is to find those functional differences from metatranscriptome sequencing data. I.e. to find the active microorganisms and which expressed transcript are present. For sake of speed, the dataset used for this tutorial is a small subset of reads from the original dataset only for the temperature at 45°C and day 5 and 30. In Peng et al., 2018 they show at temperatures of 45 °C there is significant increase in the functional profile for methanogenesis between day 5 and day 30.

The data for this tutorial includes the following files:

- **samplenr_subset_RX.fastq**: four pairs of fastq files. The files contain paired-end sequencing reads.
- **MetadataTable_subset**: Metadata of the samples.
- **SILVA Tutorial Subset (taxpro index).clc**: index of a subset of SILVA database version 138 SSURef NR99.

Please note that this tutorial contains only a small subset of the SILVA database. The full database can be downloaded with the tool Download Amplicon-Based Reference Database. For applying the functional analysis pipeline to real datasets, please use the complete databases. Now that the prerequisites have been described, it is time to start importing the input data.

1. **Download** the data from our website: https://resources.qiagenbioinformatics.com/testdata/metatranscriptome_analysis_tutorial_data.zip. Unzip and save the files locally.
2. Start your CLC Workbench and create a folder for storing input data and results, named for example "Metatranscriptome analysis".
3. Go to **Import** | **Illumina** to import the 2 · 4 sequence files (ending with "fastq") (figure 1)
    - The import type under Options is set to Paired reads.
    - Paired-end (forward-reverse) is selected.
    - Discard read names and Discard quality scores are not checked.

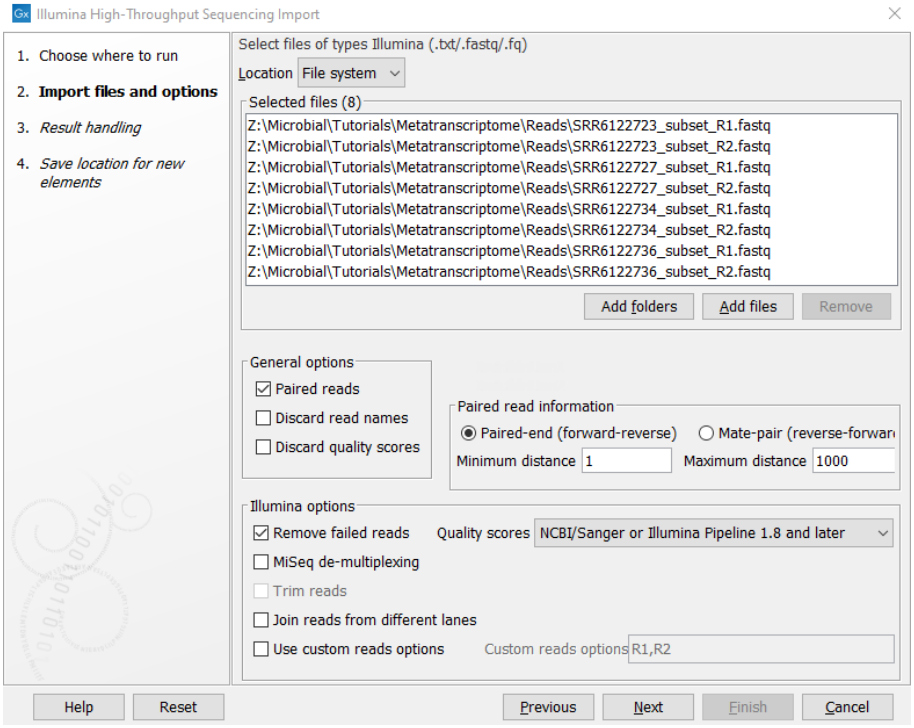

Figure 1: *Select the 8 files ending in ".fastq".*

- Leave the Paired read information as default.

4. Click Next and select the location where you want to store the imported sequences. You can check that you have four files labeled as "paired".
5. Import the metadata by clicking **Import** | **Import Metadata** on top of the Navigation Area.
6. Associate the read files with the metadata and select Prefix as seen in figure 2.
7. Go to **File** | **Import** | **Standard Import**. In the wizard, leave the import option to **Automatic Import**. Choose the file ending in ".clc" to import the SILVA database as seen in figure 3.

Your "Metatranstriptome tutorial" folder should look like figure 4.

For this tutorial we will work with QMI-PTDB and GO annotations. We therefore need to download those databases.

1. Go to **Microbial Genomic Module** ( ) | **Databases** ( ) | **Taxonomic Analysis** ( ) | **Download Curated Microbial Reference Database** ( ) to download the QMI-PTDB database.
2. Choose to download the 16GB QMI-PTDB database.
3. Check both Download Database as Sequence List and Download Database as Taxonomic Profiling Index, see figure 5.
4. Choose to save the data in the Database folder and click finish.

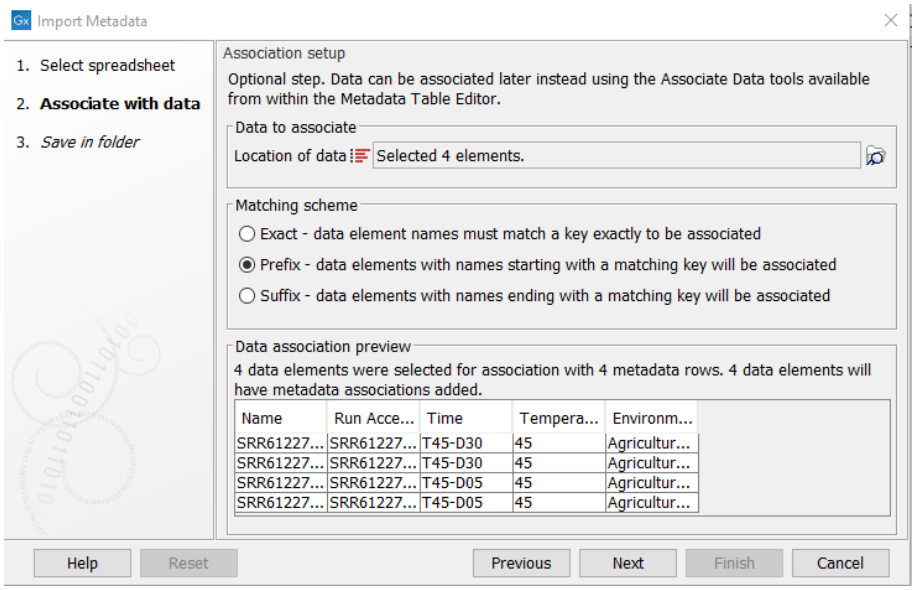

Figure 2: *Import the metadata.*

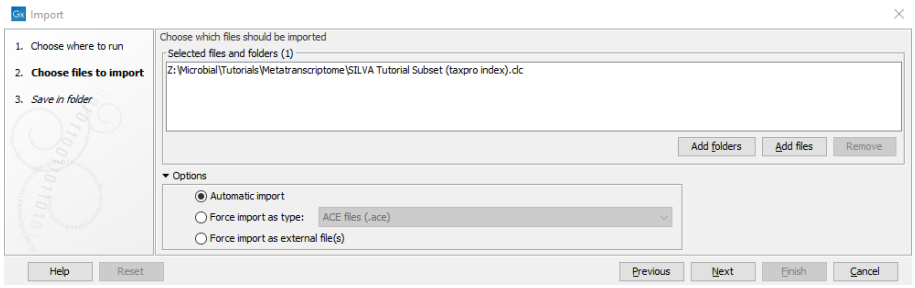

Figure 3: *Import the SILVA database.*

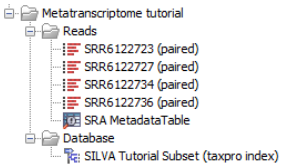

Figure 4: *Metatranscriptome analysis folder after import of all files.*

5. To download the two databases used for annotating GO-terms go to: **Microbial Genomic Module** ( ) | **Databases** ( ) | **Functional Analysis** ( ) | **Download Ontology Database** ( ).
6. Choose to download Gene Ontology with Pfam2GO mapping (GO) see figure 6.
7. Save the database in the Database folder.
8. And for the other one go to **Microbial Genomic Module** ( ) | **Databases** ( ) | **Functional Analysis** ( ) | **Download Protein Database** ( )
9. Select UniProt90 with GO annotation, see figure 7.
10. Save the database in the Database folder.

11. Now open the tool **Microbial Genomic Module** ( ) | **Databases** ( ) | **Functional Analysis** ( ) | **Create DIAMOND Index** ( )
12. Select the just downloaded UniProt database.
13. Save the DIAMOND index in the database folder.

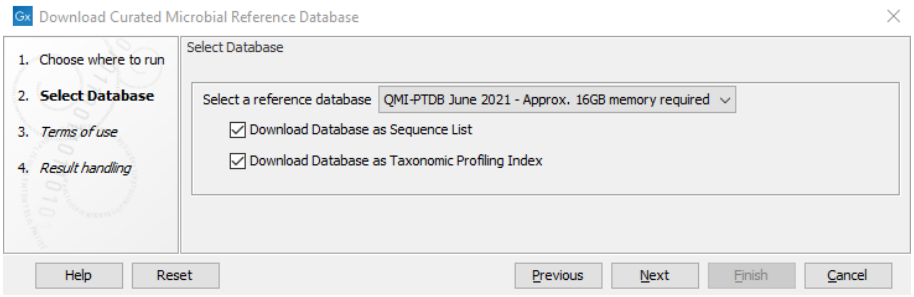

Figure 5: *Download the QMI-PTDB.*

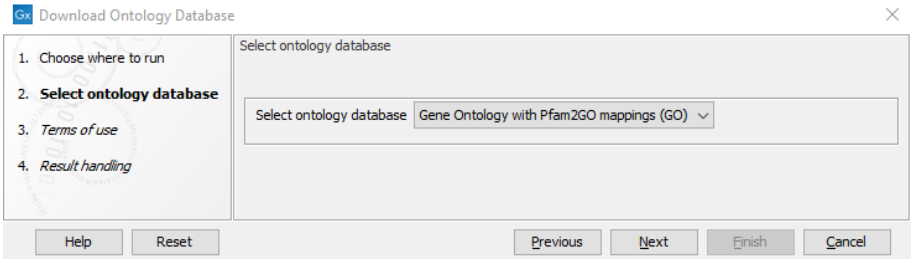

Figure 6: *Download the GO database.*

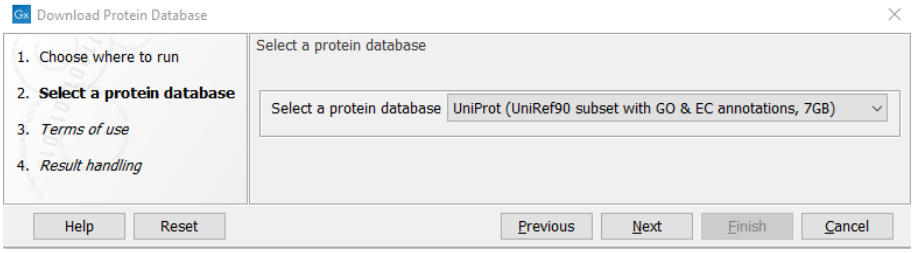

Figure 7: *Download the Protein database.*

Your folder should now look as in figure 8.

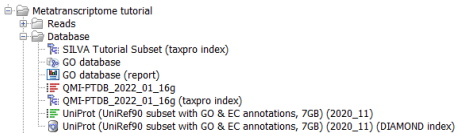

Figure 8: *Metatranscriptome analysis folder after import and download of all files.*

Remember that when working with your own data, you can create, customize or download databases using the following tools found in the **Microbial Genomic Module** ( ) | **Databases** ( ) | **Taxonomic Analysis** ( ) and the **Tools** ( ) folders of the **Microbial Genomics Module** Toolbox:

- **Create Sequence List**
- **Download Custom Microbial Reference Database** ()
- **Download Curated Microbial Reference Database** ()

Taxonomic profiling indexes can be created with **Create Taxonomic Profiling Index** tool found in **Databases** () | **Taxonomic Analysis** ().

Now that all the reference data, databases and sample data have been imported, we are ready to start the analysis.

Trim, Taxonomic Profiling, assemble and annotating

In this section, we will start by trimming the reads, then make taxonomic profiling of the samples and filter out reads that map to an rRNA database to see which microorganisms that are present in our samples to later subset the QMI-PTDB for mapping. We will assemble the reads that cannot be assigned to any of the reference genomes in QMI-PTDB into contigs and annotate them with functional information.

Trimming In this section, we will trim the reads.

1. Create a new folder, for example called "Trim" to store the results.
2. Go to the tool **Prepare Sequencing Data** () | **Trim Reads** ()
3. Select all four samples and check batch as seen in figure 9.
4. Click Next and leave the rest as default.
5. Save in specified folder.

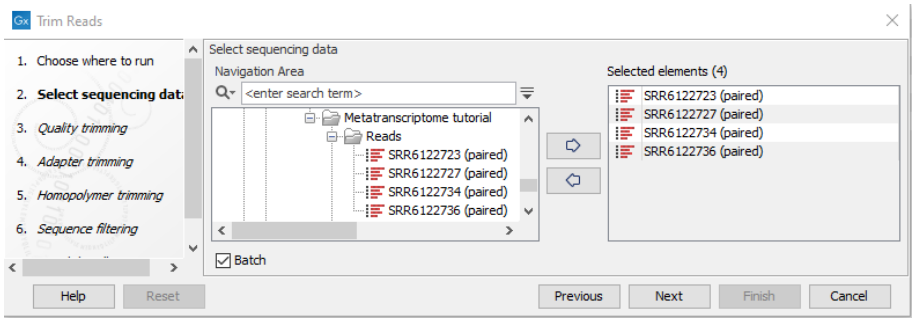

Figure 9: *Input files for trimming.*

Taxonomic Profiling In this section, we will create a taxonomic profile of the samples and filter out reads that map to a rRNA database to see which microorganisms that are present in our samples.

1. Create a new folder, for example called "Taxonomic Profile" to store the results.

2. Go to the tool **Microbial Genomic Module** ( ) | **Metagenomics** ( ) | **Taxonomic analysis** ( ) | **Taxonomic Profiling** ( )
3. Select all four trimmed reads and check batch.
4. Click Next.
5. Now select the QMI-PTDB (taxpro index) as reference index and check filter for host and select SILVA tutorial (taxpro index) as host genome index see in figure 10. Click Next.

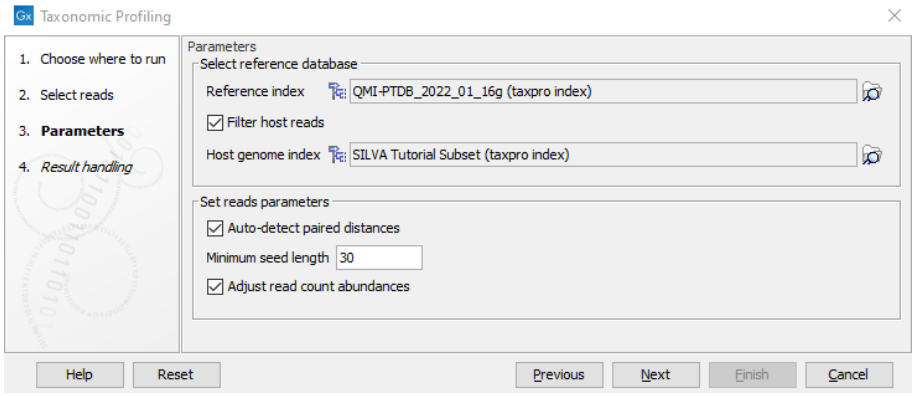

Figure 10: *Input files for taxonomic profiling.*

6. For output options check Abundance table, Unclassified reads and Report and check "Create subfolder" see in figure 11. The taxonomic profiling will take some minutes to run.

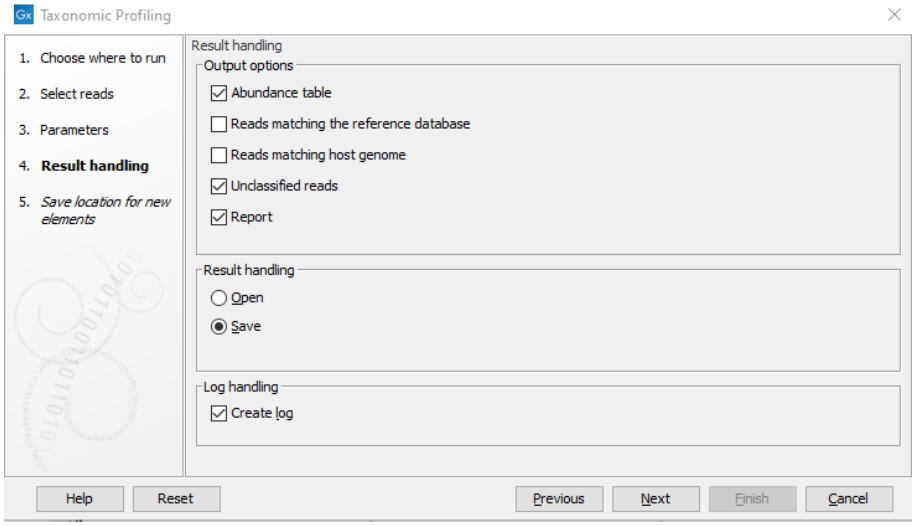

Figure 11: *Output settings for taxonomic profiling.*

Merge Abundance tables

1. Create a new folder, for example "Merged Taxa Profiling", to store the merge table. Open the tool **Microbial Genomic Module** ( ) | **Metagenomic** ( ) |**Abundance Analysis** ( ) | **Merge abundance Tables** ( ).

2. Select the four abundance tables made from the Taxonomic Profile and do not check Batch.
3. Click Next and leave the rest as default and save in specified folder, we will come make to the use of this profile later.

Assembling and annotating In this section, we will pool and assemble the unclassified reads into contigs and annotate them with functional GO-terms.

De novo assembles metagenomes

1. Create a new folder, for example "Assembly", to store the results. We are now ready to assemble the unclassified reads into contigs using the De Novo Assemble Metagenome tool: **Microbial Genomic Module** ( ) | **Metagenomics** ( ) | **De Novo Assemble Metagenome** ( )
2. Select the unclassified reads as input see figure 12. Do not select Batch.

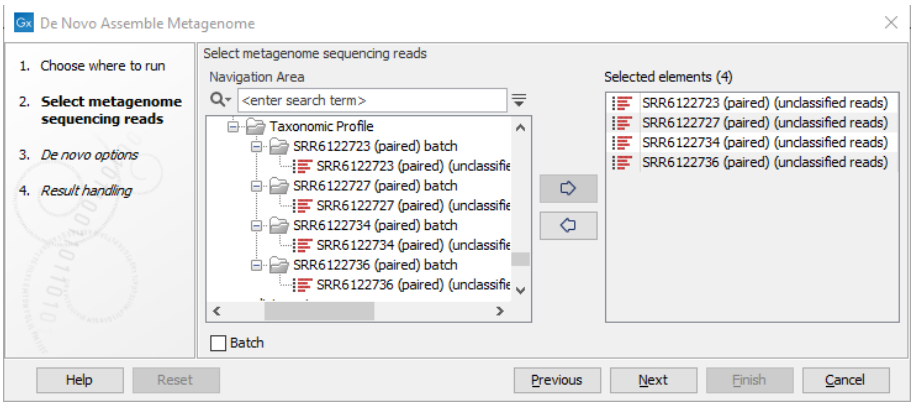

Figure 12: *Input files for the de novo assembly.*

3. Click Next.
4. In the "De novo options" dialog, make sure Minimum contig length is set to 200, choose the Longer contigs execution mode and make sure the Perform scaffolding checkbox is not checked. Click Next.
5. Choose to save the results in the "Assembly" folder. The tool will output one assembly and a report if this was enabled.

Annotate with DIAMOND

1. Create a new folder, for example "Annotated", to store the results. We are now ready to annotate the assembled contigs with the tool **Annotate with DIAMOND** ( ).
2. Open the **Microbial Genomic Module** ( ) | **Functional Analysis** ( ) | **Annotate with DIAMOND** ( ).
3. Select the just created file with the contigs assembled from metatranscriptomic reads.

4. Check DIAMOND index and select the UniProt Diamond index file see figure 13. Choose the sensitivity to be "Standard search" for this tutorial. When running Annotate with DIAMOND choosing "More sensitive" or higher can be necessary to achieve the desired sensitivity when running your own data, while other matching options may have to be adjusted.

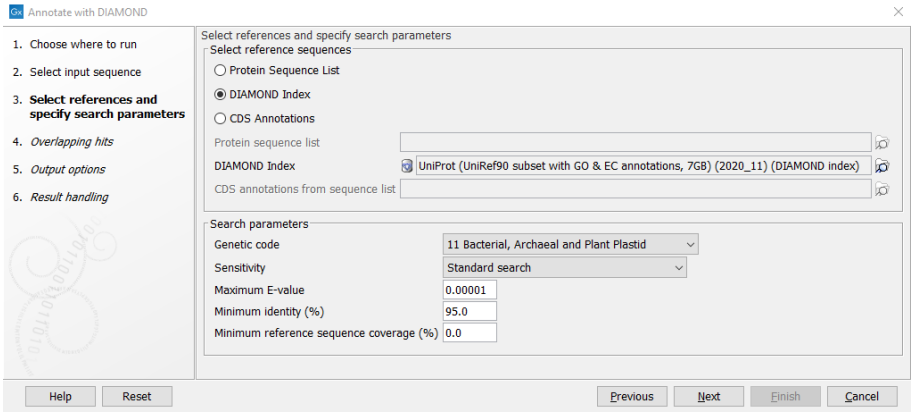

Figure 13: *Database input for the annotation of contigs.*

5. Click Next and save in the specified folder. The tool will output a sequence list of contigs annotated with CDS matching the UniProt database. Since the UniProt database was enriched with GO terms, these terms are now available for GO functional profiling using this annotated sequence list of contigs. Note that the QMI-PTDB is already annotated with such terms with UniRef90 which have GO and EC annotations.

Create a Reference List for functional profiling In this section we will create a new sequence list of the annotated Contig List and a sequence list from the QMI-PTDB.

1. Open the merged Taxonomic profile. This is a collection of the organisms found in the samples, which can be used to create a subset of the QMI-PTDB for the fast functional profiling.
2. On the right hand side we have a menu with Table Settings. Deselect all tacks in Show column except for the Assembly ID.
3. Mark the Assembly ID's and copy them as seen in figure 14.
4. Open the tool **Utilities** ( ) | **Sequence List** ( ) | **Split Sequence List** ( ) and select the downloaded QMI-PTDB. Click Next.
5. In Settings select the Define Splitting to Split based on attribute values and choose Assembly ID, and paste in the Assembly ID's from the merged Taxonomic Profile as seen in figure 15.
6. Click Next and save in a folder fx called Split Sequence List. We are now getting a file for each assembly ID.
7. Now open the tool **Utilities** ( ) | **Sequence Lists** ( ) | **Create Sequence List**.
8. Select the files in the folder Split Sequence List and the annotated Contig List as seen in figure 16.

9. Save it in new folder fx called Sequence List.

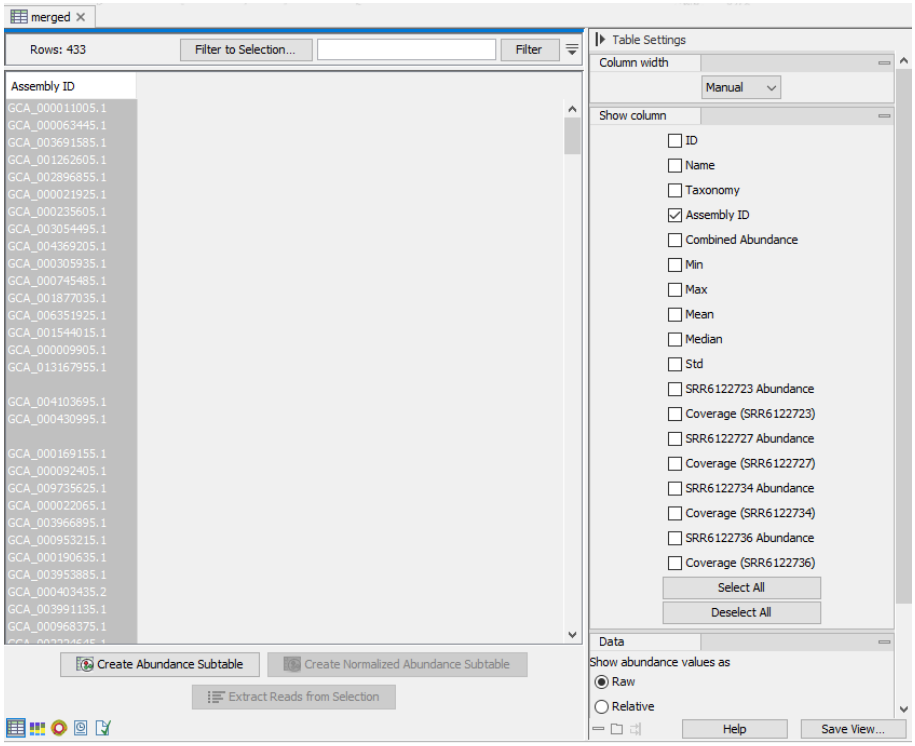

Figure 14: *Selection of Assembly ID's.*

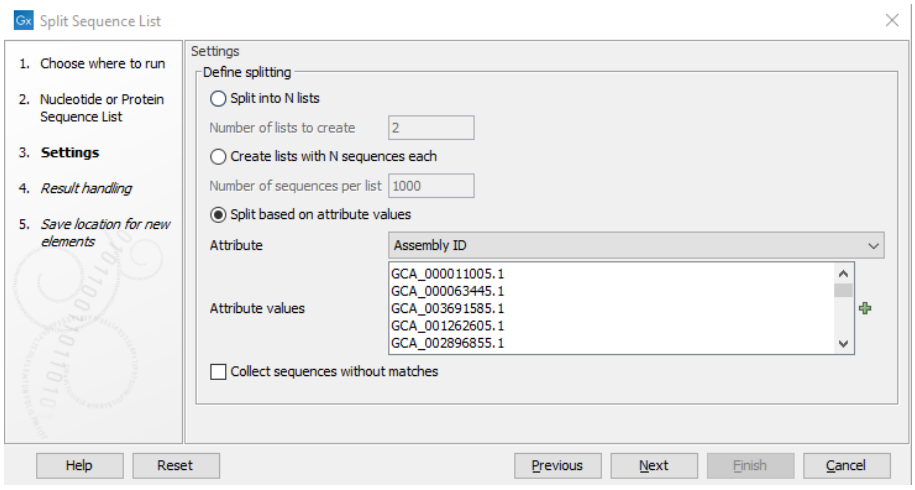

Figure 15: *Define Splitting settings to Assembly ID.*

Map and build functional report

In this section, we will map reads back to the just made Sequence List and build a functional report with the GO database.

Map Reads to Reference

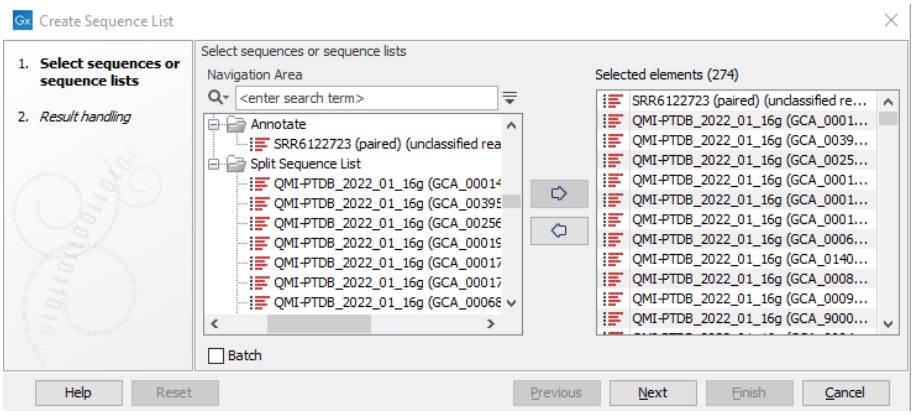

Figure 16: *Input files for creating a new reference list.*

1. Create a new folder, for example "Mapped reads", to store the results. We are now ready to map the reads back to the Sequence List using the **Map Reads to Reference** tool.
2. From the Toolbox, choose: **Resequencing Analysis** ( ) | **Map Reads to Reference** ( )
3. Select the trimmed reads as input and select the Batch option. Click Next.
4. Check that everything looks as expected in "Batch overview". If you do not see this, you likely forgot to select Batch.
5. Click ( ) and locate the new sequence list from the "Sequence List" folder see figure 17.

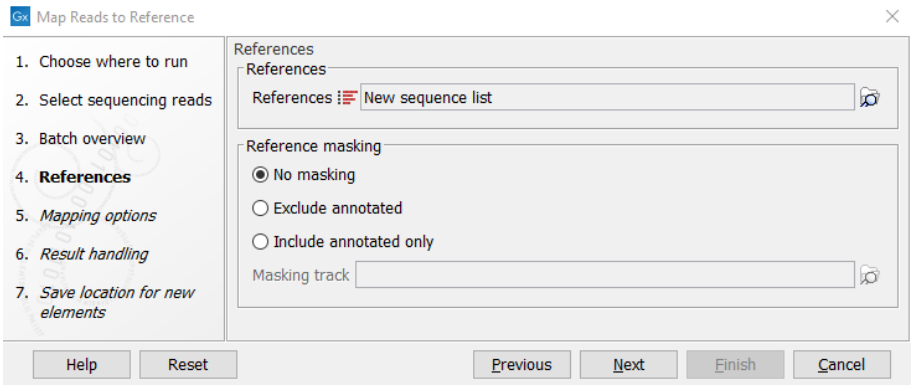

Figure 17: *New sequence list for input to the mapping to reference tool.*

6. Leave the mapping output options as default and in result handling select Save in specified location and check the Create subfolders per batch unit, see figure 18 and click Next.
7. Choose to save the result in the "Mapped reads" folder. This will take several minutes to run.

Build functional Profile

1. Create a new folder, for example "Build function", to store the results. We are now ready to build functional profiles for the samples with the tool **Build Functional Profile**.

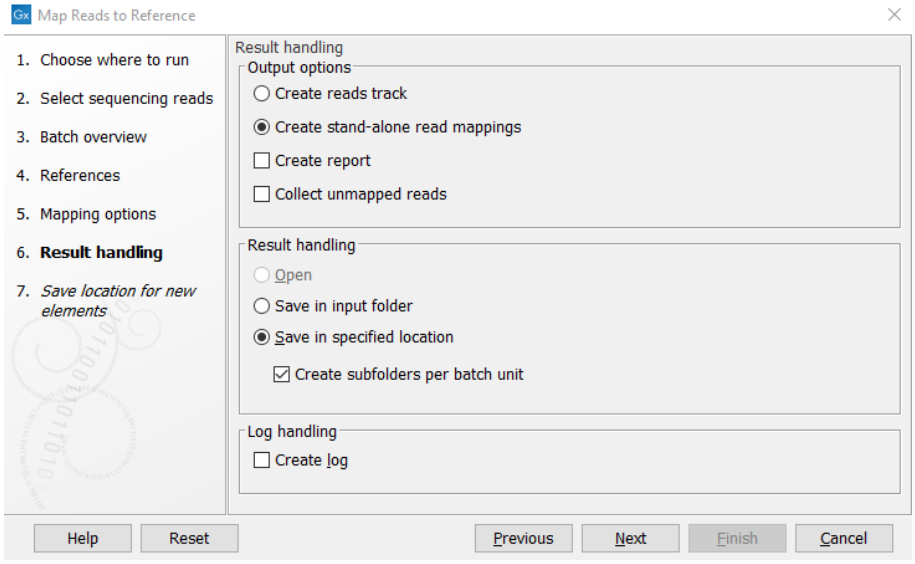

Figure 18: *Settings for output of the mapping to reference tool.*

2. From the Toolbox, choose: **Microbial Genomic Module** ( ) | **Functional Analysis** ( ) | **Build Functional Profile** ( )
3. Select the Read Mapping from the "Mapped reads" folder as input and select the Batch option. Click Next.
4. Check that everything looks as expected in "Batch overview". If you do not see this, you likely forgot to select Batch.
5. Click ( ) and locate the GO database see figure 19.

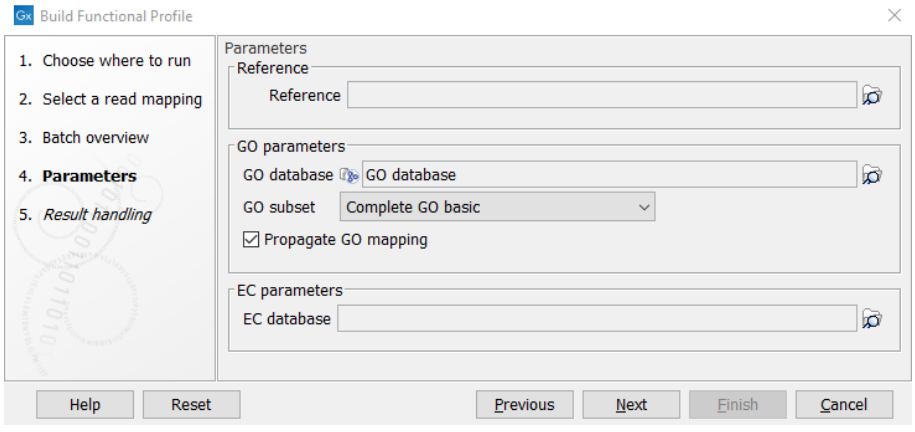

Figure 19: *Database input for the Build Functional Profile tool.*

6. Select the output to be a GO functional profile, see figure 20.
7. Choose to save the result in the "Build function" folder.

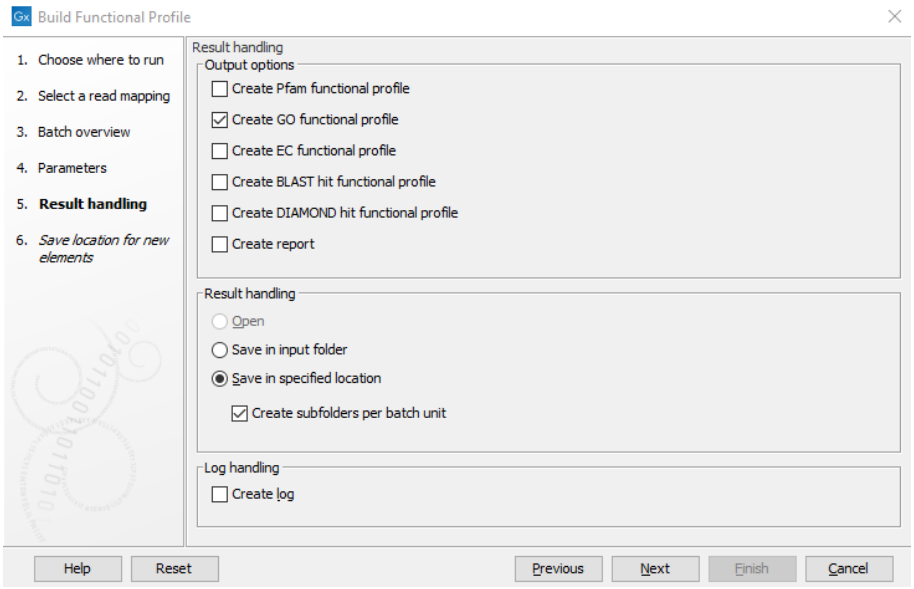

Figure 20: *Settings for output of the Build Functional Profile tool.*

Merge Abundance tables

1. Create a new folder, for example "Merged build function", to store the merge table. Open the tool **Microbial Genomic Module** ( ) | **Metagenomic** ( ) | **Abundance Analysis** ( ) | **Merge abundance Tables** ( ).
2. Select the four GO abundance tables made from the Build functional Profile and do not check batch.
3. Click Next and save in specified folder.

Differential abundance analysis and Heat map

1. Create a new folder, for example "Differential abundance analysis GO-terms", to store the differential analysis.
2. Open the tool **Microbial Genomic Module** ( ) | **Metagenomic** ( ) | **Abundance Analysis** ( ) | **Differential Abundance Analysis** ( ).
3. Select the just created merged abundance table and select Metadata factor to "Time", to find the GO terms that are differential expressed between the two time points, see figure 21.
4. Click Next and save in specified folder.
5. Open the file created. In the Filter to Selection in the top right search for "Methanogenesis" as in figure 22.

In figure 22 we see that both P-value, FDR P-vlaue and Bonferroni values are below 0.05 and there is a fold change of -177.25 meaning we have methanogenisis significantly more expressed at day 30 than at day 5 at a temperature of 45 °C.

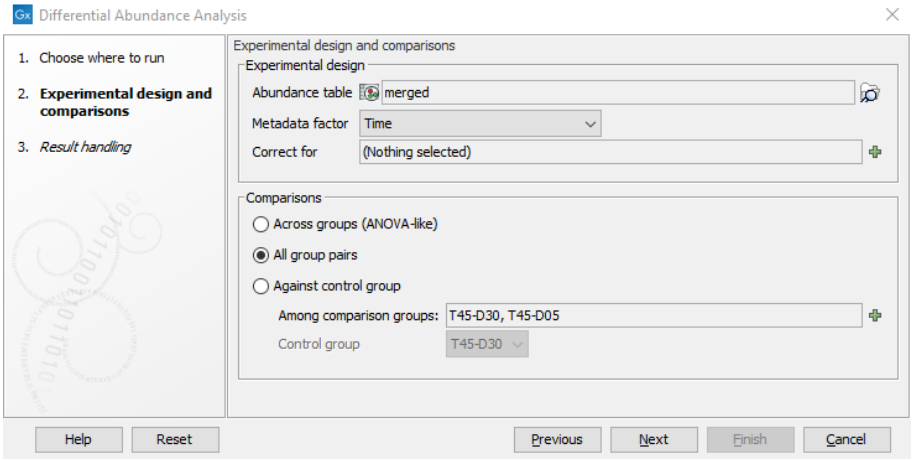

Figure 21: *Settings for the differential abundance tool.*

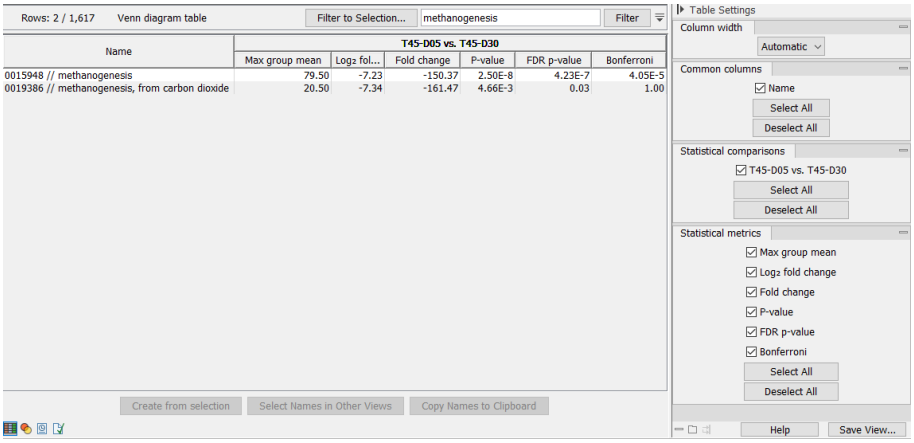

Figure 22: *Methanogenesis differential expression.*

Last comment Since we are using a subset of the full dataset both in regards to the number of samples but also the sizes of the samples, we will not get the true picture of the information the dataset holds.

References

[Abubucker S, 2012] Abubucker S, H. C. (2012). Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol. 2012 Jun;8(6):e1002358*.

[Beghini et al., 2021] Beghini, F., Huttenhower, C., Franzosa, E. A., and Segata, N. (2021). Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery. *eLife 2021;10:e65088*.

[Franzosa EA, 2018] Franzosa EA, H. C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nat Methods 15: 962-968 (2018)*.

[Peng et al., 2018] Peng, J., Wegner, C.-E., Bei, Q., Liu, P., and Liesack, W. (2018). Metatranscriptomics reveals a differential temperature effect on the structural and functional organization of the anaerobic food web in rice field soil. *Microbiome*.