Tutorial

Updating and using attributed sequences lists as microbial reference data

February 7, 2023

Sample to Insight

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

Updating and using attributed sequences lists as microbial reference data

In this tutorial we introduce the tool **Update Sequence Attributes in List**, delivered by the *CLC Genomics Workbench* 22.0, which can be used to set up and annotate your own databases for various downstream analyses.

In this tutorial, we cover how to create the following custom databases:

- Creating a Microbial Reference Database.
- Creating a Gene Database for antimicrobial resistance analysis.
- Creating a Protein Database for functional annotations.
- Creating an Amplicon Database for OTU clustering.

We also include optional examples showing how to use these databases in downstream analyses and, in an optional advanced section, we cover creating a Gene Database for virulence resistance analysis.

Please refer to the `CLC Microbial Genomics Module manual` for detailed descriptions of the tools mentioned in this tutorial.

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".

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

Please refer to the `CLC Microbial Genomics Module manual` for information about module installation and licensing.

Download and import the tutorial data

The data used in this tutorial is from a selection of microbes, covering genes and full assemblies from organisms commonly studied in the literature.

1. Download the sample data from: `http://resources.qiagenbioinformatics.com/testdata/Annotated_sequence_list_example_data.zip` and unzip it.

Import the sequences list to be updated

2. Open the *CLC Genomics Workbench*.
3. Create a new folder for the tutorial data, for example named "Attributed sequence list tutorial".
4. Import the sequence lists to be updated using the standard importer:
   (a) Go to: **File** | **Import** | **Standard Import...**
   (b) Select the five files with names ending in ".fa" from the folder you downloaded and click on **Next**.
   (c) Save the imported data in the folder you created earlier and click on **Finish**.

You should now see the following elements in the tutorial folder:

- **Microbial genomes**, containing 500 microbial reference genomes.
- **Resistance genes**, containing the NCBI subset of resistance genes from QMI-AR Nucleotide Database.
- **16S amplicons**, containing the SILVA 16s RNA amplicon database downsampled to contain 50% of the original sequences.
- **Protein sequences**, containing 10.000 protein sequences from SwissProt.
- **Virulence genes**, containing a subset of the Virulence Factor Database.

Optional: Import the example reads

We have included a data set of simulated paired-end Illumina reads from reference genomes of bacteria commonly found in wastewater. Follow the import steps below if you wish to complete the optional sections on using the updated sequence attribute lists as databases for sample analysis. If not, this section can be skipped.

5. Import the example paired-end reads by going to: **File** | **Import** ( ) | **Illumina** ( )
6. Select the files "Simulated_wastewater_reads_R1.fastq" and "Simulated_wastewater_reads_R2.fastq" and leave the settings as the defaults. Click **Next**.
7. Specify where to save the reads and click on **Finish**.

You should now see a data element called **Simulated_wastewater_reads (paired)** in the Navigation Area.

General information about using the tool Update Sequence Attributes in Lists

- A heading column used to match the attributions of the sequence is required in each input file. Attributions are added to the sequence with the corresponding name or annotation. For example, a column with "Name", containing the sequence names, can be used to add attributions based on sequence names.

  This is covered further in this tutorial, and full details can be found `in the manual`.

When creating custom databases, there are additional requirements for particular database types. These are described in the examples in this tutorial.

Creating a microbial reference database

Updated sequence attribute lists intended for use as databases for taxonomic profiling must contain taxonomy information. This can be supplied in two ways. Here, we will download the taxonomy from the NCBI using the TaxID information from the attributed sequence list to update the Taxonomy field of the sequence list we are creating.

Optionally, when one or more of the reference assemblies consist of several contigs, an Assembly ID annotation should also be provided. Assembly IDs are used in Taxonomic Profiling to calculate the abundance of each assembly by summing up the read counts for a given Assembly ID. This is described further in the `Using the Assembly ID Annotation` section of the manual.

Creating a custom microbial reference database

1. To create an updated sequence attribute list to use as a microbial reference database, choose the following from the Toolbox:

   Utility Tools ( )| **Sequence Lists** ( ) | **Update Sequence Attributes in Lists** ( ).

2. Select "Microbial genomes" from the tutorial folder and then click on **Next**.

3. Select the "Microbial_genomes_annotations.xlsx" table from your local folder. "Column to match on" should be set to "Name" and select all 9 column to be included. Check the "Download Taxonomy" option and uncheck other options as shown on figure 1. Click on **Next**.

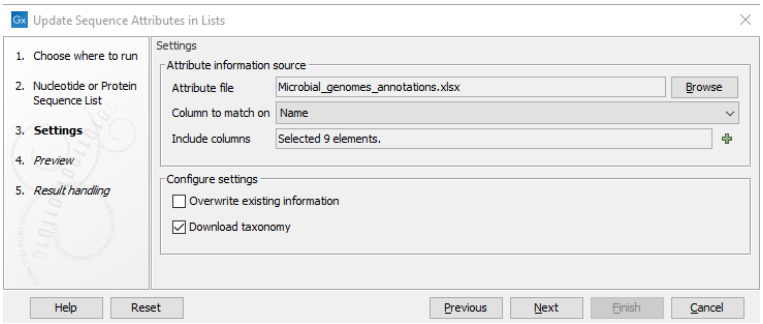

Figure 1: *Check Download Taxonomy in the attribute settings.*

4. In the "Preview" section, we get a view of the incoming data as seen in figure 2. The "Name", "Size", "Accession", "Start of sequence", "Linear", "Assembly ID", "FTP Path" and "Source" columns are column names that are known to the software and contain information consistent with the expected for this column type. When Download taxonomy checkbox is enabled and valid taxonomic identifiers are found in the "TaxID" and "Taxonomy" columns, then a 7-step taxonomy is then downloaded from the NCBI.

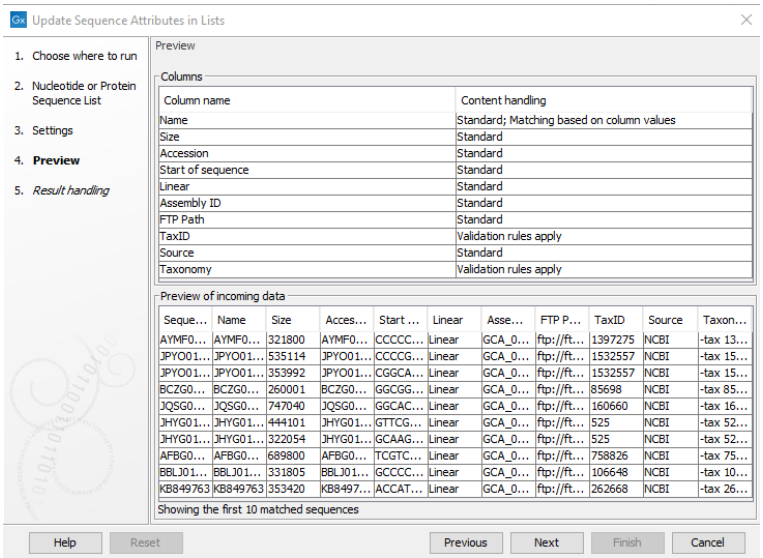

Figure 2: *Preview of the incoming attribute data.*

5. Click on **Next**. Check "Create log" and choose to save the output to a new subfolder, for example named "Attributed Microbial Reference DB".

Depending on your hardware and internet connection, the tool may take several minutes to run.

Reviewing the outputs

8. Open the log.

   In the log you can see how many sequences the tool traversed. We see that this is the number of sequences in the sequences list. This means the operation was successful.

9. Close the log when you are done.

10. Open the output sequence list from the "Attributed Microbial Reference DB" folder.

11. Switch to the Table view by clicking on () in the bottom left corner, as seen in figure 3 to see a table of the attributions present on each sequence.

12. Inspect the taxonomy column. The taxonomy matching the TaxID for each sequence was downloaded from the NCBI and then added as a Taxonomy attributions to the sequence.

You now have an updated sequence attribute list which can be used as a microbial reference database. In the following optional section, we will try using it to analyze the simulated wastewater reads.

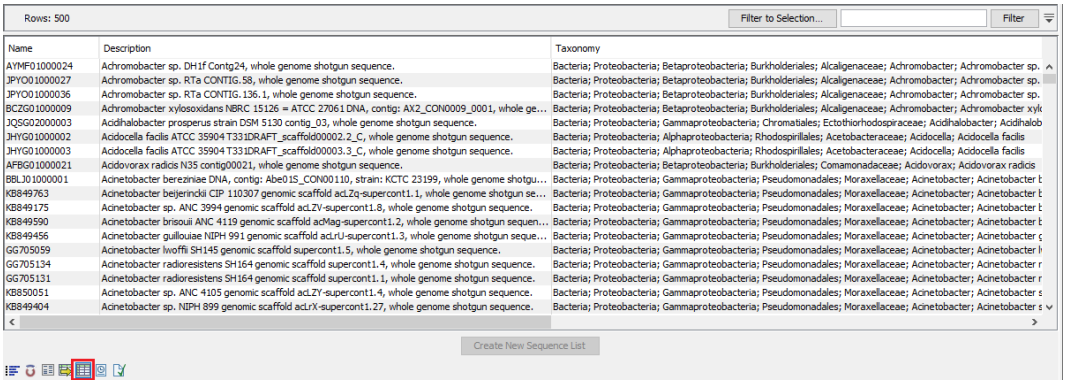

Figure 3: *Click on the Table view icon, highlighted by a red box here, to see a table of the attributions on each sequence*

Optional: Using the updated sequence attribute list as taxonomic profiling database for taxonomic profiling

You can run taxonomic profiling on the simulated wastewater reads by using the the sequence list you just updated to create a taxonomic profiling index. To do so, follow the steps below:

1. From the Toolbox, choose:
   Databases ( ) | **Taxonomic analysis** ( ) | **Create Taxonomic Profiling Index** ( )
2. Select the "Microbial genomes (Updated Attributes)" from the "Attributed Microbial Reference DB" as input.
3. Choose to **Save** the index in the "Attributed Microbial Reference DB" folder and click **Finish**. The tool will take several minutes to run. When it is done, you now have an index for taxonomic profiling.
4. Next, we will use this index to analyse the taxonomies of the simulated wastewater sample. From the Toolbox, choose:
   Metagenomics ( ) | **Taxonomic analysis** ( ) | **Taxonomic Profiling** ( )
5. As input, select the "Simulated_wastewater_reads (paired)" and click on **Next**.
6. Select the index created in the previous step by clicking on ( ). Leave the other settings on default (figure 4). Click on **Next** and save the output to a new subfolder, for example named "Taxonomic profile". The tool will now run and may take several minutes to complete.
7. Inspect the taxonomic profile ( ) in the output folder. In the Stacked visualisation, aggregate and color features by Species ( ) (figure 5). You will see there are 8 different species represented.

For more information on taxonomic profiling, we recommend you complete the **Taxonomic Profiling of Whole Shotgun Metagenomic Data tutorial** which can be found here: https://resources.qiagenbioinformatics.com/tutorials/Taxonomic_Profiling.pdf.

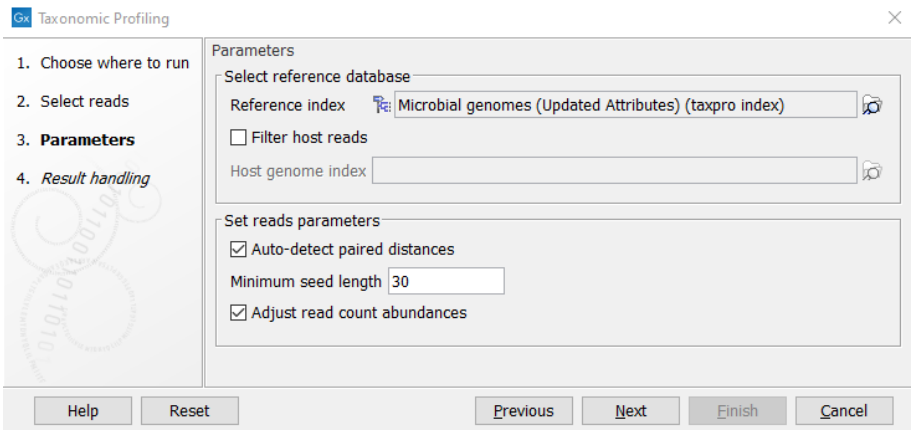

Figure 4: *Select the taxonomic profiling index created*

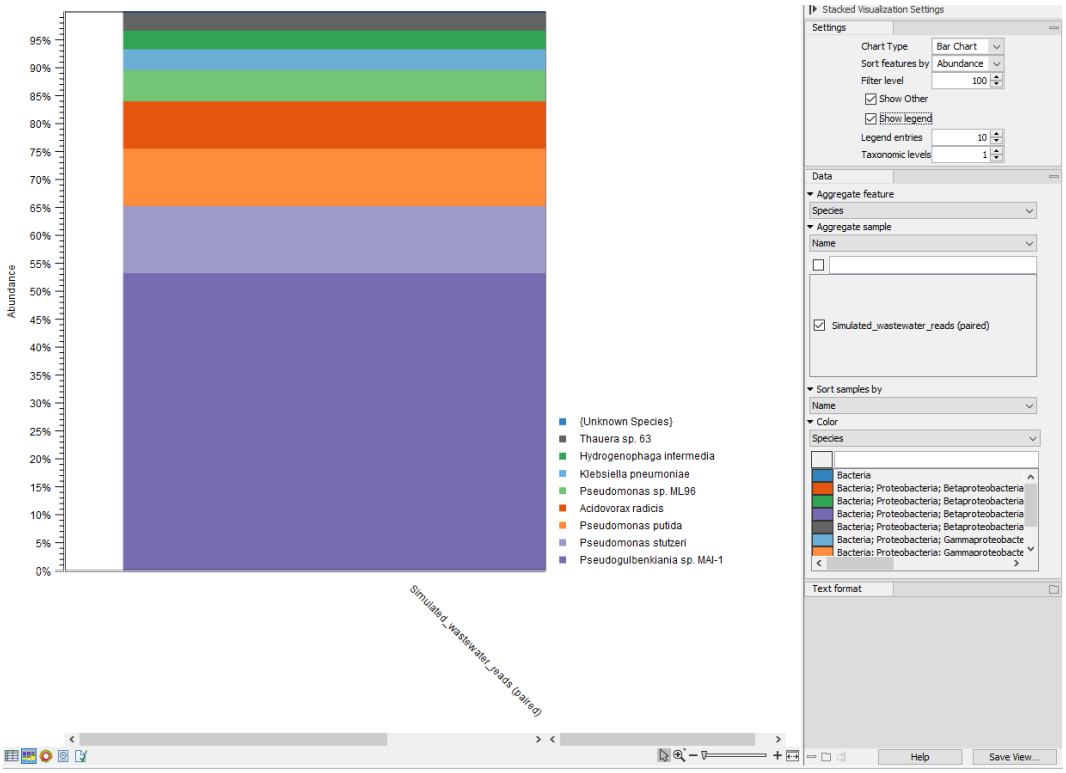

Figure 5: *Stacked visualization of the taxonomic profile*

Creating a custom gene database for antimicrobial resistance analysis

Sequence with attributed lists intended for use as resistance databases with the **Find Resistance with Nucleotide DB** tool must contain a Phenotype field which provides resistance information for a given gene.

Creating a gene database

1. To create an updated sequence attribute list to use as a nucleotide resistance database, choose the following from the Toolbox:

Utility Tools ( )| **Sequence Lists** ( ) | **Update Sequence Attributes in Lists** ( ).

2. Select "Resistance genes" from the tutorial folder location and then click on **Next**.

3. Select the "Resistance_genes_annotations.xlsx" table in your local folder. For the "Column to match on" select "Name" and include all columns in the "Include columns". Uncheck all other options as shown in figure 6. Click on **Next**

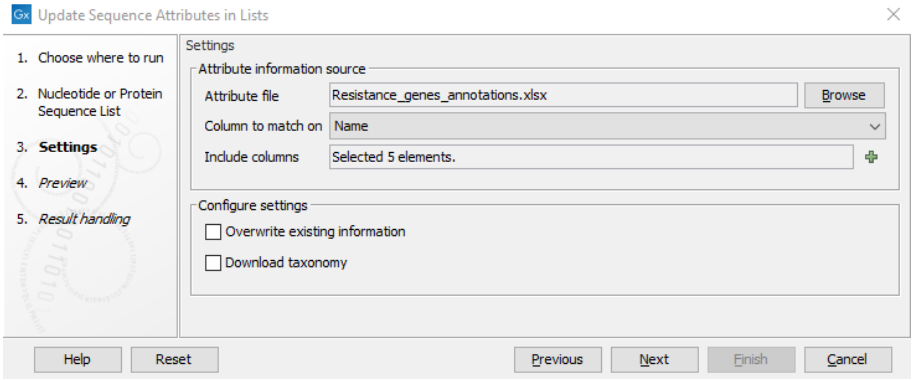

Figure 6: *Select the file containing the annotation table.*

4. In the Preview area, inspect the columns of the table as seen in figure 7. The headings are checked by the software and handled accordingly.

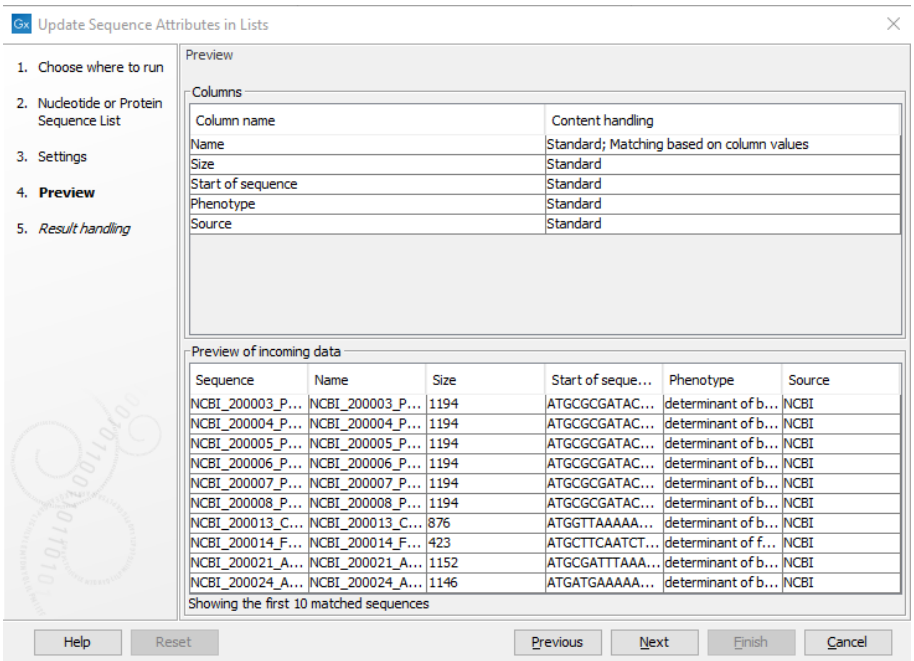

Figure 7: *Preview of the incoming attribute data.*

5. After confirming that the preview looks as expected with a Name and Phenotype field click on **Next**.

6. Keep the "Create log" checked, and choose to save the output to a new subfolder, for example titled "Attributed resistance genes".

Reviewing the outputs

After the tool has finished running, we will briefly inspect the output.

7. Open the log.

   In the log you can see how many sequences the tool traversed. We see that this is the number of sequences in the sequences list. This means the operation was successful.

8. Close the log when you are done.

9. Open the output sequence list from the "Attributed resistance genes" folder.

10. Switch to the Table view by clicking on ( ) in the bottom left corner to see a table of attributions present on each sequence.

11. Inspect the Phenotype column.

    The phenotype attributions have been transferred to the sequences by matching the contents of the "Name" column from the attribution table with the sequence names.

Optional: Using the annotated sequence list as a gene database for finding resistance

You can find resistance genes in the simulated wastewater reads using the updated sequence attribute list you just created. First, the metagenome reads must be assembled. To do so, follow the steps below:

1. From the Toolbox, choose: **Metagenomics** ( ) | **De Novo Assemble Metagenome** ( )

2. As input select the "Simulated_wastewater_reads (paired)" and click on **Next**.

3. Set execution mode to Longer contigs and leave the other settings on default as seen in (figure 8). Click on **Next**

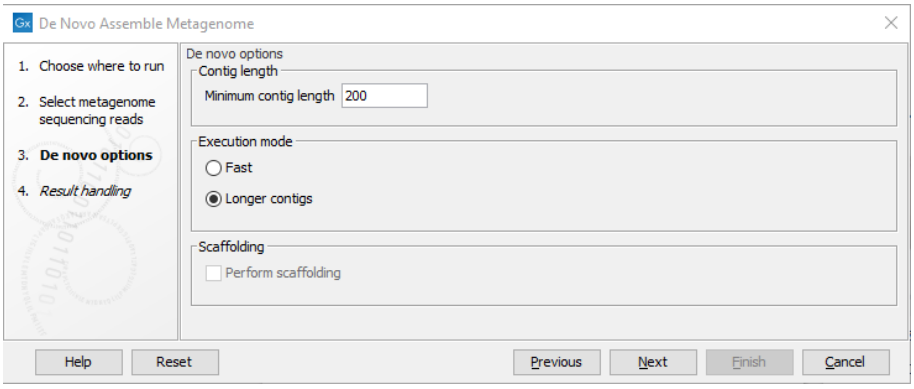

Figure 8: *De novo assemble metagenome settings*

4. Save the assembled metagenomes in a new subfolder, for example named "Assembled metagenome".

The tool will run and output a contig list. We will use the attributed sequence list we created as the nucleotide resistance database to search for resistance genes in the metagenome assembly.

5. From the Toolbox, choose:
   Drug Resistance Analysis ( ) | **Find Resistance with Nucleotide DB** ( )
6. As input select "Simulated_wastewater_reads (paired) contig list" from the "Assembled metagenome" folder. Click on **Next**.
7. Select the "Resistance genes" from the "Attributed resistance genes" folder as seen in (figure 9). Leave the other settings on default. Click on **Next**

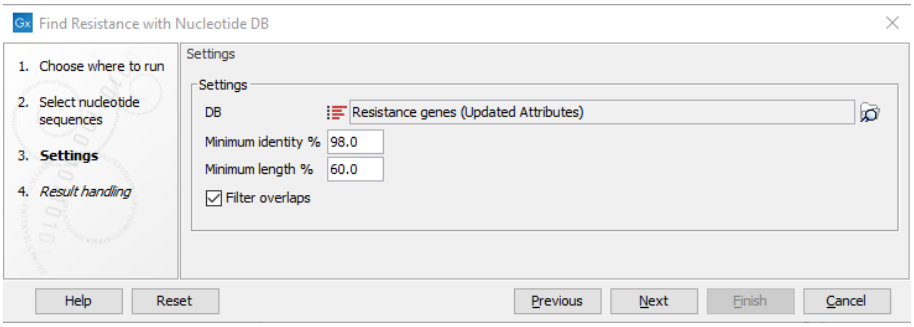

Figure 9: *Select the attributed resistance genes to search for resistance genes*

8. Save the output in the "Assembled metagenome" folder.

The tool outputs a resistance table. Open and inspect the table. You will observe that a number of resistance genes were found.

For more information on the tools for detecting antibiotic resistance, we recommend you complete the **Antibiotic Resistance Analysis tutorial** which can be found here: `https://resources.qiagenbioinformatics.com/tutorials/Antimicrobial_Resistance.pdf`.

Creating a protein database for functional annotations

There are several ways to create attributed protein sequence lists for use as protein databases. Here, we will go through how to attribute a protein sequence list with GO terms. GO term attributions are required in order to create a functional profile for GO terms.

Creating a custom protein database

1. To create an updated sequence attribute list to use as a protein database, choose the following from the Toolbox:
   Utility Tools ( )| **Sequence Lists** ( ) | **Update Sequence Attributes in Lists** ( ).
2. Select "Protein sequences" from the tutorial folder location and then click on **Next**.
3. Select the "Protein_sequences_annotations.xlsx" table in your local folder. For the "Column to match on" select "Name" and include all columns in the "Include columns". Uncheck all other options as shown in figure 10. Click on **Next**
4. In Preview, inspect the columns of the table. The headings are checked by the software and handled accordingly as seen in figure 11. In order for GO-terms to be recognized the input file must contain a column named "GO-terms".

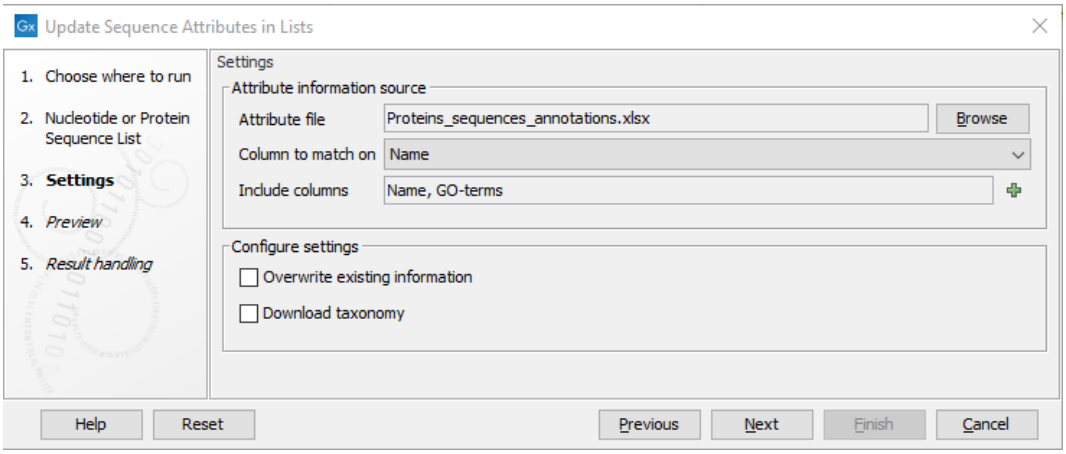

Figure 10: *Select the file containing the annotation table.*

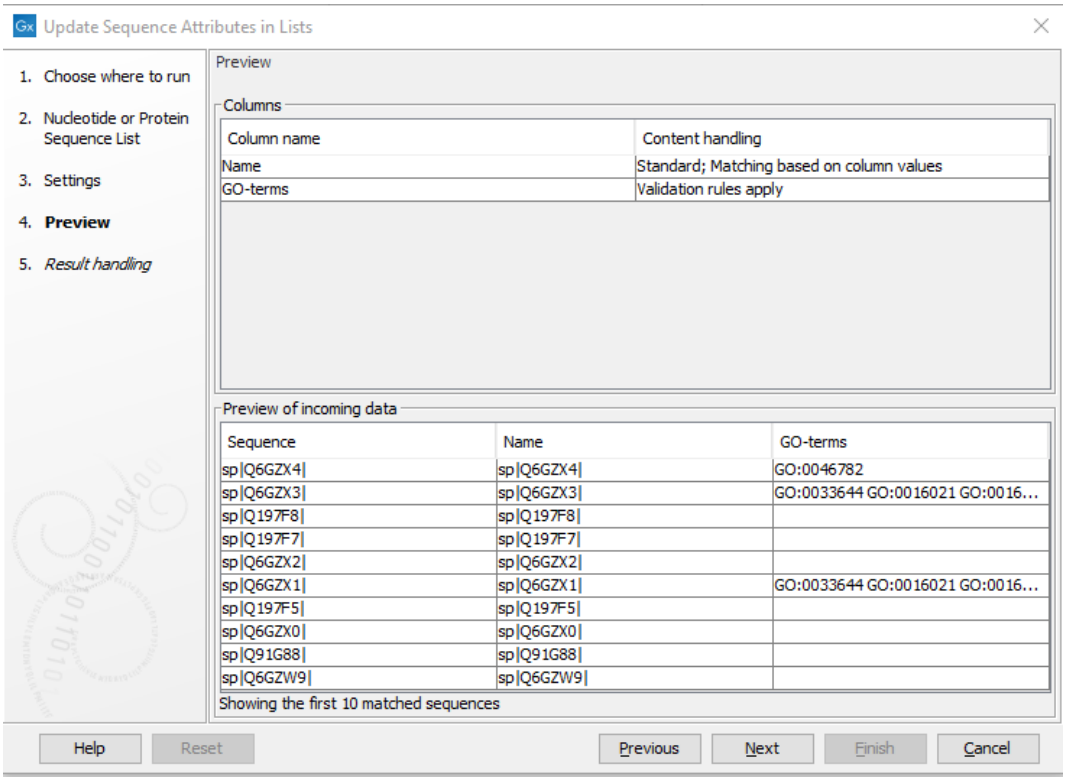

Figure 11: *Preview of the incoming data.*

5. After confirming that the preview looks as expected click on **Next**.
6. Keep the "Create log" checked, and choose to save the output to a new subfolder, for example titled "Attributed Protein DB".

Reviewing the outputs

7. Open the log. In the log you can see how many sequences the tool traversed. We see that this is the number of sequences in the sequences list. This means the operation was

successful.

8. Close the log when you are done.
9. Open the output sequence list from the "Attributed Proteins" folder.
10. Switch to the Table view by clicking on ( ) in the bottom left corner.
11. Inspect the GO-terms column. The sequences have been attributed with GO-terms. The GO-terms annotation has special meaning which can be seen by clicking on a row in the "GO-terms" column. This will take you to the GO description of this gene.

Optional: Using the attributed protein sequence list to build a functional profile

We will use the metagenome assembly of the wastewater sample we built previously with the updated protein sequence attribute list to build a GO functional profile.

In order to do so, the assembly must first be annotated with cds regions containing GO annotations. We will use the Annotate with DIAMOND tool for this.

1. From the Toolbox, choose: **Functional Analysis** ( ) | **Annotate with DIAMOND** ( )
2. As input select the "Simulated_wastewater_reads (paired) contig list" and click on **Next**.
3. Select "Protein sequence list" as the reference sequence then click ( ) to locate the "Protein sequences (Updated Attributes)" from the "Attributed Protein DB" folder. Leave the other options as default. The wizard parameters should appear as on figure 12. Click on **Next**.

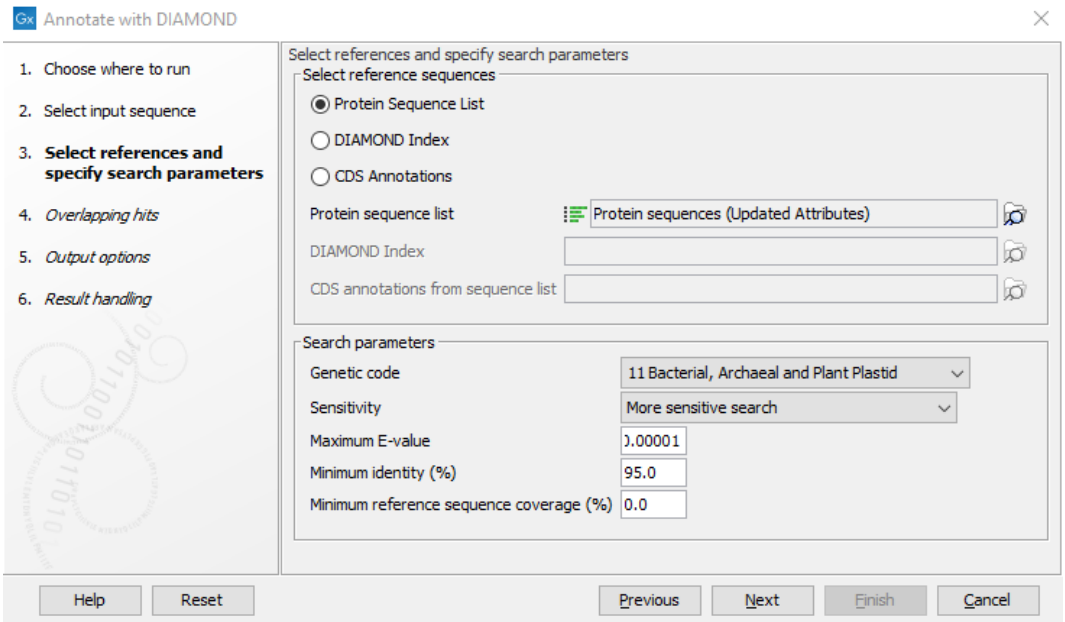

Figure 12: *Annotate the metagenome assembly with DIAMOND*

4. Leave the next two wizard steps on default by clicking **Next** twice.

5. Choose to save the annotated metagenome assembly in the "Assembled metagenome" folder.

The tool will run and output a contig list with "(DIAMOND annotations)".

Open the output list and switch to Annotation Table view by clicking on ( ) to see a number of cds annotations. We will use these annotations to build a functional profile.

The first step in building a functional profile is mapping the reads to the annotated contigs.

6. From the Toolbox, choose:
   Resequencing Analysis ( ) | **Map Reads to Reference** ( )
7. As input select the raw "Simulated_wastewater_reads (paired)" reads and click on **Next**.
8. As reference, select the "Simulated_wastewater_reads (paired) contig list (DIAMOND annotations)" from the "Assembled metagenome" folder. Click **Next**.
9. Leave the mapping options as default and click on **Next**.
10. Save the read mapping in the "Assembled metagenome" folder.

We now have a read mapping and are ready to build the GO functional profile. If you have do not already have a GO database downloaded, you should do so now using
Databases ( ) | **Functional Analysis** ( ) | **Download GO Database** ( )

This database is not limited to this tutorial so save it in your general database location.

11. From the Toolbox, choose:
    Functional Analysis ( ) | **Build Functional Profile** ( )
12. As input select the read mapping created in the previous step and click on **Next**.
13. As Reference, select the "Simulated_wastewater_reads (paired) contig list (DIAMOND annotations)" from the "Assembled metagenome" folder. In GO database, locate your GO database. Leave the other settings as default (see figure 13). Click **Next**.
14. Uncheck all output options except Create GO functional profile.
15. Choose to save the output in a new location for example named "Wastewater functional profile".

Inspect the output profile to see that a number of different GO terms are represented.

For more information on functional analysis including how to compare different samples, we recommend you complete the **Whole Metagenome Functional Analysis tutorial** which can be found here: https://resources.qiagenbioinformatics.com/tutorials/Microbial_Analysis_Functional.pdf.

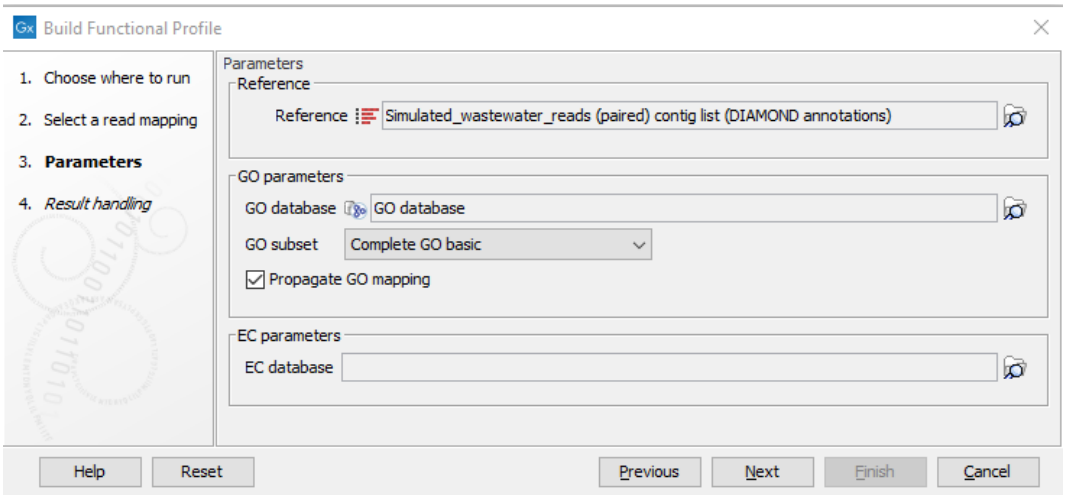

Figure 13: *Use the annotated contig list and downloaded GO database to build the functional profile*

Creating a 16S database for OTU clustering

In this section, you will create an attributed sequence list that can be used as a database for OTU clustering. Attributed sequence lists intended for use as databases for OTU clustering must contain taxonomy information. Here we provide the taxonomies directly in the Taxonomy field.

Create a custom 16S database

1. To create an updated sequence attribute list to use as a database for OTU clustering, choose the following from the Toolbox::

   Utility Tools ( )| **Sequence Lists** ( ) | **Update Sequence Attributes in Lists** ( ).

2. Select "16S amplicons" from the tutorial folder location and then click on **Next**.

3. In the import area click **Browse** and select the "16S_amplicons_annotations.xlsx" table, as shown in figure 14.

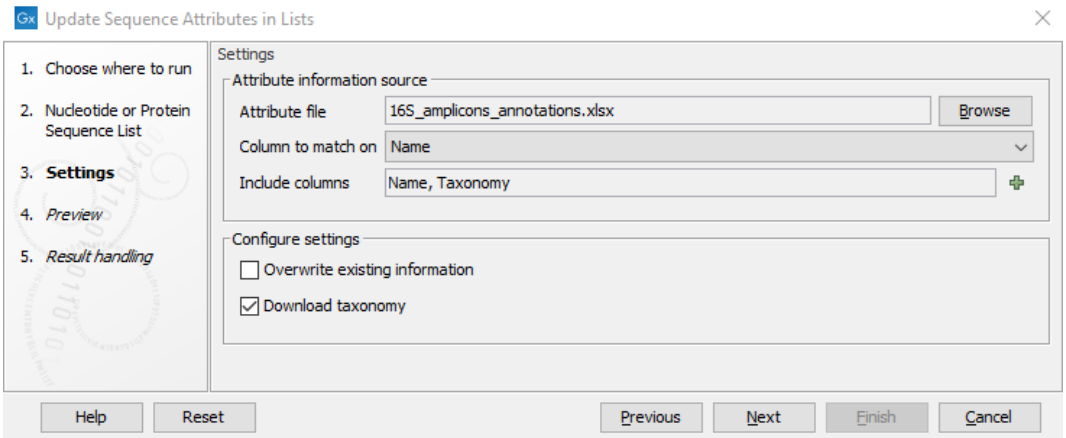

Figure 14: *Select the file containing the annotation table.*

4. In Preview, inspect the columns of the table. The headings are checked by the software and handled accordingly as seen in figure 15. Then click on **Next**.

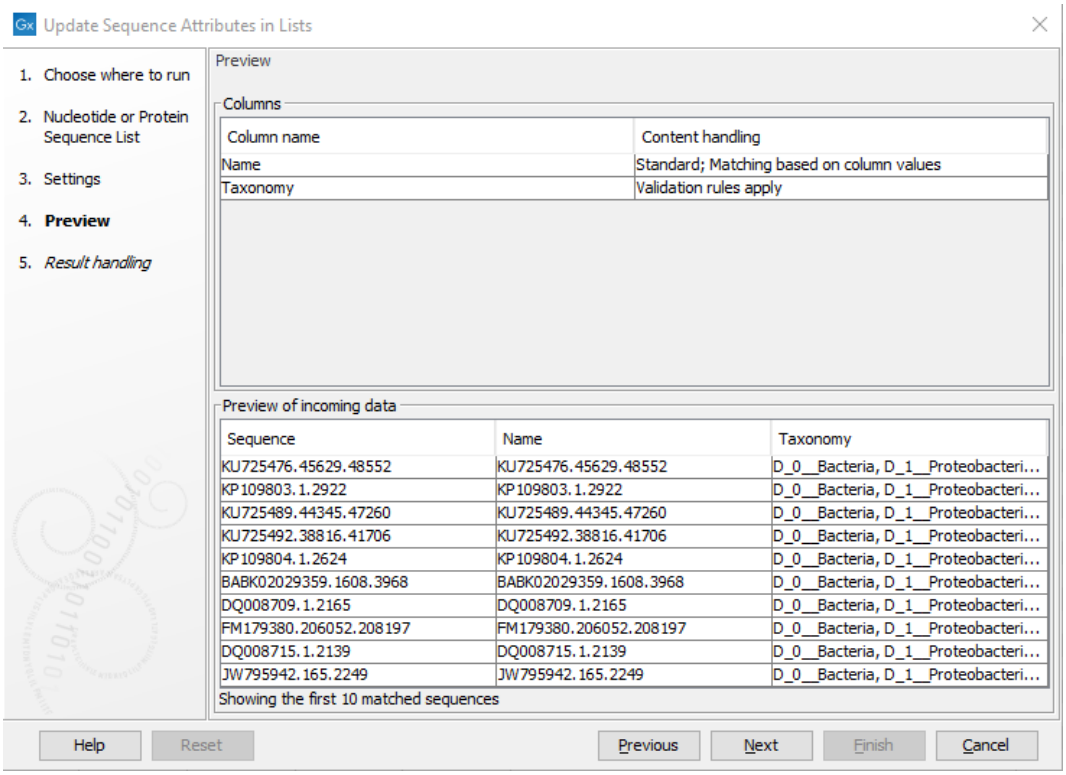

Figure 15: *Preview of the incoming data.*

5. Keep the "Create log" checked, and choose to save the output to a new subfolder, for example titled "Attributed 16S DB".

Reviewing the outputs

6. Open the log. In the log you can see how many sequences the tool traversed. We see that this is the number of sequences in the sequences list. This means the operation was successful.

7. Close the log when you are done.

8. Open the output sequence list from the "Attributed 16S DB" folder.

9. Switch to the Table view by clicking on ( ) in the bottom left corner to see a table of attributes present on each sequence.

10. Inspect the taxonomy column. The taxonomies were automatically detected as being QIIME formatted and converted to 7-step taxonomy.

    This conversion allows the taxonomies to be used as database input for both OTU clustering and to create taxonomic profiling indexes. Taxonomies can be specified in QIIME format (starting with "k__ " and comma or semi-colon separated) as seen here or as a semi-colon separated strings.

Optional: Using the attributed sequence list as reference database for OTU clustering

If you wish to try using the created database for OTU clustering, we recommend using the data from the **OTU clustering step by step tutorial** which can be found here: `https://resources.qiagenbioinformatics.com/tutorials/OTU_Clustering_Steps.pdf`. Simply replace the 16S_97_otus_GG database with the one you just created.

Optional: Creating a virulence database for cds annotation

In this optional section we will go through how to create an attributed sequence list that can be used as a virulence database. Attributed sequence lists intended for use as virulence databases with the **Find Resistance with Nucleotide DB** tool must contain the following four columns in addition to the Name column: Virulence factor, Virulence factor ID, Virulence gene and Gene ID.

Creating a custom virulence database

1. To create an attributed sequence list to use as a virulence database, choose the following from the Toolbox:

   Utility Tools ( )| **Sequence Lists** ( ) | **Update Sequence Attributes in Lists** ( ).

2. Select "Virulence genes" from the tutorial folder location and then click on **Next**.

3. Uncheck all options. Click on **Next**.

4. Click **Reset** to clear the previous input.

5. In the import area **Browse** and select the "Virulence_genes_annotations.xlsx" table, as shown in figure 16.

6. In Preview, inspect the columns of the table. The headings are checked by the software and handled accordingly as seen in figure 17. Then click on **Next**

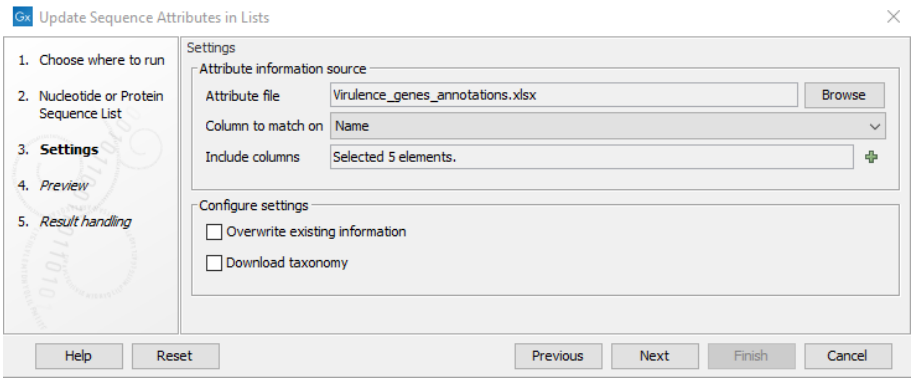

Figure 16: *Select the file containing the annotation table.*

7. Keep the "Create log" checked, and choose to save the output to a new subfolder, for example titled "Attributed virulence genes".

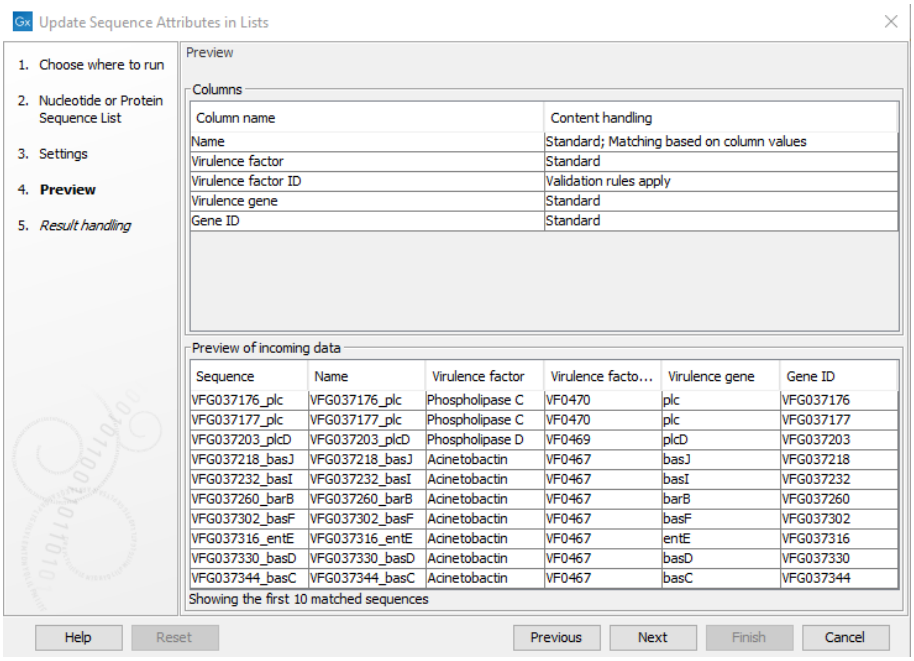

Figure 17: *Preview of the incoming metadata.*

Reviewing the outputs

8. Open the log. In the log you can see how many sequences the tool traversed. We see that this is the number of sequences in the sequences list. This means the operation was successful.
9. Close the log when you are done.
10. Open the output sequence list from the "Attributed virulence genes" folder.
11. Switch to the Table view by clicking on ( ) in the bottom left corner.
12. Inspect the Virulence factor and Gene ID columns. These field have special meaning. Clicking on a row in the " Virulence factor" or " Gene ID" columns will take you to a description of this virulence gene.

Optional: Using the updated sequence attribute list as a virulence database for finding virulence

We will use the Microbial genome database we imported and attributed previously and add virulence attributions.

1. From the Toolbox, choose:
   Drug Resistance Analysis ( ) | **Find Resistance with Nucleotide DB** ( )
2. As input select the "Microbial genomes" from the "Attributed Microbial Reference DB" folder and click on **Next**.
3. Select the "Virulence genes" nucleotide sequence list from the "Attributed virulence genes" folder as the DB by clicking ( ). Leave the other options as default. The wizard parameters should appear as on figure 18. Click on **Next**

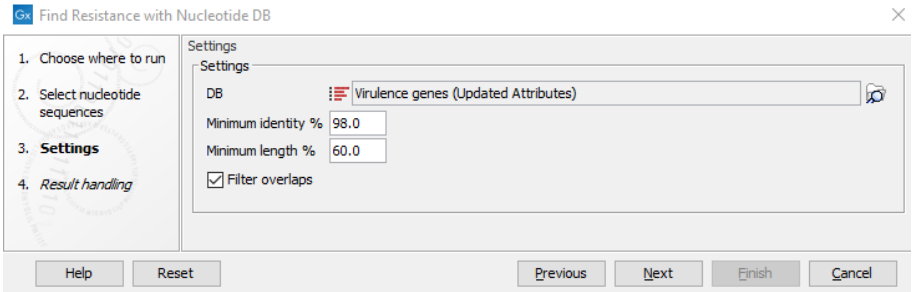

Figure 18: *Find virulence genes in the reference database*

4. In the last step, save the output table in the "Attributed Microbial Reference DB" folder.

The tools runs and may take several minutes to complete. Open and inspect the Find resistance table. In the contigs column, we can see that three of the references were found to have virulence genes. None of these were detected in taxonomic profiling and it is therefore unlikely that the sample contains any particularly virulent strain.