

Updating and using attributed sequences lists as microbial reference data

February 7, 2023

Sample to Insight -

QIAGEN Aarhus \cdot Silkeborgvej 2 \cdot Prismet \cdot 8000 Aarhus C \cdot Denmark digitalinsights.qiagen.com \cdot 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.giagenbioinformatics.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 ())
- 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.
- 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.

Gx (Jpdate Sequence Att Choose where to run	ributes in Lists Settings Attribute information	source			×
2. N S	Nucleotide or Protein Sequence List	Attribute file	Microbial_genomes_annota	tions.xlsx		Browse
3. 9	Settings	Include columns	Selected 9 elements.			4
5. /	Result handling	Configure settings	g information			
Concerning of the local data	2 Manantus				 	

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.

1 Choose where to run	Preview											
1. Choose where to run	- Columns -											
2. Nucleotide or Protein	Columns						_					
Sequence List	Column na	ame					Cont	tent handli	ng			
2 Cattings	Name						Stand	lard; Matc	hing bas	ed on colum	n values	
5. Setungs	Size						Stand	lard				
4. Preview	Accession						Stand	lard				
	Start of sequence					Stand	lard					
5. Result handling	Linear				Stand	lard						
	Assembly I	D					Stand	lard				
	FTP Path						Stand	lard				
	TaxID						Valida	ation rules	apply			
	Source						Stand	lard				
	Taxonomy						Valida	ation rules	apply			
	Preview of	f incoming o	data									
	Seque	Name	Size	Acces	Start	Linea	ar	Asse	FTP P.	TaxID	Source	Taxon
	AYMF0	AYMF0	321800	AYMF0	ccccc	Linea	r (GCA_0	ftp://ft	1397275	NCBI	-tax 13
	JPYO01	JPYO01	535114	JPYO01	CCCCG	Linea	r (GCA_0	ftp://ft	1532557	NCBI	-tax 15
	JPYO01	JPYO01	353992	JPYO01	CGGCA	Linea	r (GCA_0	ftp://ft	1532557	NCBI	-tax 15
	BCZG0	BCZG0	260001	BCZG0	GGCGG	Linea	r (GCA_0	ftp://ft	85698	NCBI	-tax 85
	JQSG0	JQSG0	747040	JQSG0	GGCAC	Linea	r (GCA_0	ftp://ft	160660	NCBI	-tax 16
	JHYG01	JHYG01	444101	JHYG01	GTTCG	Linea	r (GCA_0	ftp://ft	525	NCBI	-tax 52
	JHYG01	JHYG01	322054	JHYG01	GCAAG	Linea	r (GCA_0	ftp://ft	525	NCBI	-tax 52
	AFBG0	AFBG0	689800	AFBG0	TCGTC	Linea	r (GCA_0	ftp://ft	758826	NCBI	-tax 75
	BBLJ01	BBLJ01	331805	BBLJ01	GCCCC	Linea	r (GCA_0	ftp://ft	106648	NCBI	-tax 10
	KB849763	KB849763	353420	KB8497	ACCAT	Linea	r (GCA_0	ftp://ft	262668	NCBI	-tax 26

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 (E) 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.



Rows: 500		Filter to Selection Filter
Name	Description	Taxonomy
AYMF01000024	Achromobacter sp. DH1f Contg24, whole genome shotgun sequence.	Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae; Achromobacter; Achromobacter sp.
JPYO01000027	Achromobacter sp. RTa CONTIG.58, whole genome shotgun sequence.	Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Alcalgenaceae; Achromobacter; Achromobacter sp.
JPYO01000036	Achromobacter sp. RTa CONTIG. 136.1, whole genome shotgun sequence.	Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae; Achromobacter; Achromobacter sp.
BCZG01000009	Achromobacter xylosoxidans NBRC 15126 = ATCC 27061 DNA, contig: AX2_CON0009_0001, whole ge	Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Alcaligenaceae; Achromobacter; Achromobacter xyk
JQSG02000003	Acidihalobacter prosperus strain DSM 5130 contig_03, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Chromatiales; Ectothiorhodospiraceae; Acidihalobacter; Acidihalob
JHYG01000002	Acidocella facilis ATCC 35904 T331DRAFT_scaffold00002.2_C, whole genome shotgun sequence.	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Acetobacteraceae; Acidocella; Acidocella facilis
JHYG01000003	Acidocella facilis ATCC 35904 T331DRAFT_scaffold00003.3_C, whole genome shotgun sequence.	Bacteria; Proteobacteria; Alphaproteobacteria; Rhodospirillales; Acetobacteraceae; Acidocella; Acidocella facilis
AFBG01000021	Acidovorax radicis N35 contig00021, whole genome shotgun sequence.	Bacteria; Proteobacteria; Betaproteobacteria; Burkholderiales; Comamonadaceae; Acidovorax; Acidovorax radicis
BBL J0 100000 1	Acinetobacter bereziniae DNA, contig: Abe015_CON00110, strain: KCTC 23199, whole genome shotgu	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter t
KB849763	Acinetobacter beijerinckii CIP 110307 genomic scaffold acl.Zq-supercont1.1, whole genome shotgun se	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter t
KB849175	Acinetobacter sp. ANC 3994 genomic scaffold acLZV-supercont1.8, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter t
KB849590	Acinetobacter brisouii ANC 4119 genomic scaffold acMag-supercont1.2, whole genome shotgun sequen	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter t
KB849456	Acinetobacter guillouiae NIPH 991 genomic scaffold acLrU-supercont1.3, whole genome shotgun seque	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter ç
GG705059	Acinetobacter lwoffii SH145 genomic scaffold supercont1.5, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter I
GG705134	Acinetobacter radioresistens SH164 genomic scaffold supercont1.4, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter r
GG705131	Acinetobacter radioresistens SH164 genomic scaffold supercont1.1, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter r
KB850051	Acinetobacter sp. ANC 4105 genomic scaffold acLZY-supercont1.4, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter s
KB849404	Acinetobacter sp. NIPH 899 genomic scaffold acLrX-supercont1.27, whole genome shotgun sequence.	Bacteria; Proteobacteria; Gammaproteobacteria; Pseudomonadales; Moraxellaceae; Acinetobacter; Acinetobacter s
<		>
	Create New Serus	ance list

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 (IIII) in the output folder. In the Stacked visualisation, aggregate and color features by Species (IIIII) (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.



. Choose where to run	Parameters Select reference database	
. Select reads	Reference index 🛛 👫 Microbial genomes (Updated Attributes) (taxpro index)	ø
. Parameters	Filter host reads	
. Result handling	Host genome index	Q
	Set reads parameters Auto-detect paired distances Minimum seed length 30	
	Adjust read count abundances	

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**

Gx	Update Sequence Att	ributes in Lists		\times
1.	 Choose where to run Nucleotide or Protein Sequence List 	Settings Attribute information	source	
2.		Attribute file Column to match on	Resistance_genes_annotations.xlsx Brow	se V
3.	Settings	Include columns	Selected 5 elements.	•
4. 5.	Preview Result handling	Configure settings	g information	
		Download taxond	my	
	Help Res	et	Previous Next Einish Can	cel

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.

Choose where to ru	Preview					
	Columns					
Nucleotide or Protei	n la			C I It.		
Sequence List	Column name			Content handling		
Settings	Name			Standard; Matchir	ig based on column v	alues
	Size			Standard		
Preview	Start of sequence			Standard		
	Phenotype			Standard		
Result handling	Source			standard		
	Preview of incomin	iy uata				
	Sequence	Name	Size	Start of seque	Phenotype	Source
	Sequence NCBI 200003 P.,	Name	Size	Start of seque	Phenotype determinant of b	Source NCBI
	Sequence NCBI_200003_P NCBI_200004_P	Name NCBI_200003_P NCBI_200004_P	Size 1194 1194	Start of seque ATGCGCGATAC ATGCGCGATAC	Phenotype determinant of b determinant of b	Source NCBI NCBI
	Sequence NCBI_200003_P NCBI_200004_P NCBI_200005_P	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P	Size 1194 1194 1194	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC	Phenotype determinant of b determinant of b determinant of b	Source NCBI NCBI NCBI
	NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200006_P	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200006_P	Size 1194 1194 1194 1194	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC	Phenotype determinant of b determinant of b determinant of b determinant of b	Source NCBI NCBI NCBI NCBI
	VERIE Sequence NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200005_P NCBI_200006_P NCBI_200007_P	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200006_P NCBI_200007_P	Size 1194 1194 1194 1194 1194 1194	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC	Phenotype determinant of b determinant of b determinant of b determinant of b determinant of b	Source NCBI NCBI NCBI NCBI NCBI
	NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200005_P NCBI_200007_P NCBI_200007_P NCBI_200008_P	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200006_P NCBI_200007_P NCBI_200008_P	Size 1194 1194 1194 1194 1194 1194 1194	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC	Phenotype determinant of b determinant of b determinant of b determinant of b determinant of b	Source NCBI NCBI NCBI NCBI NCBI NCBI
	Preview of incomin Sequence NCBI_200003_P NCBI_200006_P NCBI_200006_P NCBI_200007_P NCBI_200007_P NCBI_200008_P NCBI_200008_P	Name NCBI_200003_P NCBI_200003_P NCBI_200005_P NCBI_200006_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_P	Size 1194 1194 1194 1194 1194 1194 1194 876	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC	Phenotype determinant of b determinant of b determinant of b determinant of b determinant of b determinant of b	Source NCBI NCBI NCBI NCBI NCBI NCBI NCBI
	Preview of incomin Sequence NCBI_200003_P NCBI_200005_P NCBI_200006_P NCBI_200006_P NCBI_200008_P NCBI_200008_P NCBI_2000014_F	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200006_P NCBI_200007_P NCBI_200008_P NCBI_200003_P NCBI_200013_C NCBI_200014_F	Size 1194 1194 1194 1194 1194 1194 1194 876 423	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCTTAAAA ATGCTTCAATCT	Phenotype determinant of b determinant of b determinant of b determinant of b determinant of b determinant of b determinant of f	Source NCBI NCBI NCBI NCBI NCBI NCBI NCBI
	Preview of incomin Sequence NCBI_200003_P NCBI_200005_P NCBI_200006_P NCBI_200008_P NCBI_200008_P NCBI_2000014_F NCBI_200014_F NCBI_200014_F	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200006_P NCBI_200006_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200017_P NCBI_200017_P NCBI_200017_P NCBI_200014_F NCBI_200014_F NCBI_200012_A	Size 1194 1194 1194 1194 1194 1194 876 423 1152	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGTCAATCT ATGCTTCAATCT ATGCTTCAATCT	Phenotype determinant of b determinant of b determinant of b determinant of b determinant of b determinant of f determinant of f	Source NCBI NCBI NCBI NCBI NCBI NCBI NCBI NCBI
	Preview of incomi Sequence NCBI_200003_P NCBI_200006_P NCBI_200006_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_P NCBI_200007_N NCBI_200013_C NCBI_200014_F NCBI_200014_F NCBI_200014_A	Name NCBI_200003_P NCBI_200004_P NCBI_200005_P NCBI_200005_P NCBI_200007_P NCBI_200007_P	Size 1194 1194 1194 1194 1194 1194 1194 876 423 1152 1146	Start of seque ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCGCGATAC ATGCTTAAAA ATGCTTCAATCT. ATGCTAAAAA ATGCTAAAAA ATGCTAAAAAA	Phenotype determinant of b determinant of b determinant of b determinant of b determinant of b determinant of b determinant of f determinant of f determinant of b determinant of b	Source NCBI NCBI NCBI NCBI NCBI NCBI NCBI NCBI

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

Gx De Novo Assemble N	letagenome	\times
 Choose where to run Select metagenome sequencing reads De novo options <i>Result handling</i> 	De novo options Contig length Minimum contig length 200 Execution mode Fast O Fast O Longer contigs	
	Perform scaffolding	
Help Res	et <u>Previous Next</u> <u>Finish</u> <u>Cancel</u>	

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**

Find Resistance with Choose where to run Select nudeotide sequences Settings	Nucleotide DB Settings DB Fesistance genes (Updated Attributes) Minimum identity % [98.0]	×
4. <i>Result handling</i> Help Re	Minimum length % 60.0 Filter overlaps set Previous Next Enish C	ancel

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".



. Choose where to run	Settings Attribute information s	iource	
Nucleotide or Protein	Attribute file	Proteins_sequences_annotations.xlsx	Browse
Sequence List	Column to match on	Name	~
Settings	Include columns	Name, GO-terms	4
. Preview			
. Result handling	Configure settings	g information	
	Download taxono	my	



Gx Update Sequence At	tributes in Lists					\times
1. Choose where to run	Preview					
	Columns					
2. Nucleotide or Protein Sequence List	Column name		Content ha	ndling		
	Name		Standard; M	latching base	ed on column value	s
3. Settings	GO-terms		Validation ru	iles apply		
4. Preview						
5. Result handling						
	Preview of incoming data					
	Sequence	Name		GO-	-terms	
	sp Q6GZX4	sp Q6GZX4		GO:0	0046782	
	sp Q6GZX3	sp Q6GZX3		GO:0	0033644 GO:0016	021 GO:0016
	sp Q197F8	sp Q197F8				
	sp Q197F7	sp Q197F7				
	sp Q6GZX2	sp Q6GZX2				
	sp Q6GZX1	sp Q6GZX1		GO:0	0033644 GO:0016	021 GO:0016
	sp Q197F5	sp Q197F5				
	sp Q6GZX0	sp Q6GZX0				
	sp Q91G88	sp Q91G88				
	sp Q6GZW9	sp Q6GZW9				
	Showing the first 10 match	ned sequences				
Help Re	eset		Previous	<u>N</u> ext	<u>F</u> inish	<u>C</u> ancel

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 (IIII) 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 (a) 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 ((a) 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**.

Gx Annotate with DIAMOND			\times
1. Choose where to run	Select references and specify search parameter Select reference sequences	'S	
2. Select input sequence	Protein Sequence List		
3. Select references and specify search parameters	DIAMOND Index OCDS Annotations		
4. Overlapping hits	Protein sequence list	tein sequences (Updated Attributes)	R
5. Output options	DIAMOND Index		Ø
6. Result handling	CDS annotations from sequence list		ର
1001101 1001100 1000000	Search parameters Genetic code Sensitivity Maximum E-value Minimum identity (%) Minimum reference sequence coverage (%)	11 Bacterial, Archaeal and Plant Plastid More sensitive search 0.00001 95.0 0.0	
Help Reset		Previous Next Einish Ca	incel

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 (E) 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** (Solar) | **Functional Analysis** (Solar) | **Download GO Database** (Solar)

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.



Choose where to run	ParametersReference	
Select a read mapping	Reference :: Simulated_wastewater_reads (paired) contig list (DIAMOND annotations)	ø
Parameters	⊂GO parameters	
Result handling	GO database 🕼 GO database	Ø
	GO subset Complete GO basic 🗸	
	Propagate GO mapping	
	_ EC parameters	
	EC database	Q

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.

Gx	Gx Update Sequence Attributes in Lists X						
1.	Choose where to run	Settings Attribute information s	ource				
2.	Nucleotide or Protein Sequence List	Attribute file 16S_amplicons_annotations.xlsx Column to match on Name				Browse	
3.	3. Settings	Include columns	Name, Taxonomy				4
 Preview Result handling 	Preview Result handling	Configure settings	y information my				
	Help Reset Previous Next Finish Cancel						<u>C</u> ancel

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**.

. Choose where to ru	Preview					
. Nucleotide or Prote	in Column and	Contractions	dia a			
Sequence List	Column name	Content nan				
. Settings	Name	Standard; Ma	atching based on column values			
	Taxonomy	Validation rule	Validation rules apply			
Preview						
Result handling						
	Preview of incoming data					
	Preview of incoming data					
	Preview of incoming data	Name	Taxonomy			
	Preview of incoming data Sequence KU725476.45629.48552	Name KU725476.45629.48552	Taxonomy D 0 Bacteria, D 1 Proteobacteri			
	Preview of incoming data Sequence KU725476.45629.48552 KP109803.1.2922	Name KU725476.45629.48552 KP109803.1.2922	Taxonomy D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri			
	Preview of incoming data Sequence KU725476.45629.48552 KP109803.1.2922 KU725489.44345.47260	Name KU725476,45629.48552 KP109803.1.2922 KU725489,44345.47260	Taxonomy D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706	Name KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624	Name KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968	Name KU725476.45629.48552 KP109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP109804.1.2624 BABK02029359.1608.3968	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165	Name KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM179380.206052.208197	Name KU725476.45629.48552 KP109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM179380.206052.208197	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM179380.206052.208197 DQ008715.1.2139	Name KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM179380.206052.208197 DQ008715.1.2139	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725499.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM 179380.206052.208197 DQ008715.1.2139 JW795942.165.2249	Name KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38316.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM179380.206052.208197 DQ008715.1.2139 JW795942.165.2249	Taxonomy D_0_Bacteria, D_1_Proteobacter D_0_Bacteria, D_1_Proteobacter			
	Preview of incoming data Sequence KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725499.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM 179380.206052.208197 DQ008715.1.2139 JW 795942.165.2249 Showing the first 10 matched seq	Name KU725476.45629.48552 KP 109803.1.2922 KU725489.44345.47260 KU725492.38816.41706 KP 109804.1.2624 BABK02029359.1608.3968 DQ008709.1.2165 FM179380.206052.208197 DQ008715.1.2139 JW795942.165.2249	Taxonomy D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri D_0_Bacteria, D_1_Proteobacteri			

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 (\mathbf{F}) Sequence Lists (\mathbf{F}) | Update Sequence Attributes in Lists (\mathbf{F}).
```

- 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**

Gx	Update Sequence Att	ributes in Lists					\times
1.	Choose where to run	Settings Attribute information s	source				
2.	Nucleotide or Protein	Attribute file	Virulence_genes_annotations.xlsx				Browse
	Sequence List	Column to match on Name				~	
3.	Settings	Include columns	Selected 5 elements.				4
4.	Preview						
5.	Result handling	Configure settings	g information				
		Download taxono	omy				
	Help Res	set		Previous	Next	Finish	Cancel

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".



Choose where to run	Preview						
	Columns						
Nucleotide or Protein Sequence List	Column name			Content handling			
bequerice bot	Name			Standard: Matching	a based on column	values	
Settings	Virulence factor	Virulence factor			Standard, Hatching based on column values		
	Virulence factor ID	Virulence factor ID			Validation rules apply		
Preview	Virulence gene	Virulence gene					
Result handling	Gene ID	Gene ID					
	-Preview of incomin	g data					
	Preview of incomin	g data	Virulence factor	Virulence facto	Virulence gene	Gene ID	
	Preview of incomin Sequence VFG037176_plc	g data Name VFG037176_plc	Virulence factor Phospholipase C	Virulence facto VF0470	Virulence gene plc	Gene ID VFG037176	
	Preview of incomin Sequence VFG037176_plc VFG037177_plc	g data Name VFG037176_plc VFG037177_plc	Virulence factor Phospholipase C Phospholipase C	Virulence facto VF0470 VF0470	Virulence gene plc plc	Gene ID VFG037176 VFG037177	
	Preview of incomin Sequence VFG037176_plc VFG037177_plc VFG037203_plcD	g data Name VFG037176_plc VFG037177_plc VFG037203_plcD	Virulence factor Phospholipase C Phospholipase C Phospholipase D	Virulence facto VF0470 VF0470 VF0469	Virulence gene plc plc plcD	Gene ID VFG037176 VFG037177 VFG037203	
	Preview of incomin Sequence VF6037176_plc VF6037177_plc VF6037203_plcD VF6037218_basJ	g data Name VFG037176_plc VFG037177_plc VFG037203_plcD VFG037218_basJ	Virulence factor Phospholipase C Phospholipase C Phospholipase D Acinetobactin	Virulence facto VF0470 VF0470 VF0469 VF0469 VF0467	Virulence gene plc plc plcD basJ	Gene ID VFG037176 VFG037177 VFG037203 VFG037218	
	Preview of incomin Sequence VFG037176_plc VFG037203_plcD VFG0372218_basJ VFG0372232_basI	g data Name VFG037176_plc VFG037203_plcD VFG037218_basJ VFG037232_basI	Virulence factor Phospholipase C Phospholipase D Acinetobactin Acinetobactin	Virulence facto VF0470 VF0469 VF0467 VF0467 VF0467	Virulence gene plc plcD basJ basI	Gene ID VFG037176 VFG037177 VFG037203 VFG037218 VFG037232	
	Preview of incomin Sequence VFG037176_plc VFG037203_plcD VFG037218_bas3 VFG037222_bas1 VFG037220_bar8	g data Name VFG037176_plc VFG037177_plc VFG037203_plcD VFG037218_basJ VFG037232_basI VFG037260_barB	Virulence factor Phospholipase C Phospholipase D Acinetobactin Acinetobactin	Virulence facto VF0470 VF0470 VF0469 VF0467 VF0467 VF0467 VF0467	Virulence gene plc plcD basJ basI basF	Gene ID VFG037176 VFG037177 VFG037203 VFG037218 VFG037232 VFG037260	
	Preview of Incomin Sequence VFG037176_plc VFG037203_plcD VFG037203_plcD VFG037228_basI VFG037250_barB VFG037260_barB	g data Name VFG037176_plc VFG037177_plc VFG037203_plcD VFG037218_basI VFG037260_bar8 VFG037302_basF	Virulence factor Phospholipase C Phospholipase D Acinetobactin Acinetobactin Acinetobactin	Virulence facto VF0470 VF0470 VF0469 VF0467 VF0467 VF0467 VF0467 VF0467	Virulence gene plc plcD basJ basI barB barB	Gene ID VFG037176 VFG037177 VFG037203 VFG037218 VFG037232 VFG037260 VFG037302	
	Preview of incomin Sequence VFG037176_plc VFG037213_plc VFG03722_bas1 VFG03722_bas1 VFG03720_bar8 VFG03730_bas5 VFG03730_bas5	g data Name VFG037176_plc VFG037177_plc VFG037203_plcD VFG037226_basI VFG037206_barB VFG037302_basF VFG037316_entE	Virulence factor Phospholipase C Phospholipase D Acinetobactin Acinetobactin Acinetobactin Acinetobactin	Virulence facto VF0470 VF0470 VF0469 VF0467 VF0467 VF0467 VF0467 VF0467 VF0467	Virulence gene plc plc plcD basJ basI basF entE	Gene ID VFG037176 VFG037177 VFG037218 VFG037218 VFG037232 VFG037260 VFG037302 VFG037316	
	Preview of incomin Sequence VFG037176_plc VFG037203_plc VFG037223_basJ VFG037223_basJ VFG037232_basF VFG037302_basF VFG037330_basB VFG037330_basA	g data Name VFG037176_plc VFG037177_plc VFG037203_plcD VFG037232_basJ VFG0372302_basF VFG037302_basF VFG037316_entE VFG037330_basD	Virulence factor Phospholipase C Phospholipase D Acinetobactin Acinetobactin Acinetobactin Acinetobactin Acinetobactin	Virulence facto VF0470 VF0470 VF0467 VF0467 VF0467 VF0467 VF0467 VF0467 VF0467 VF0467	Virulence gene plc plcD basJ basE basF entE basD	Gene ID VFG037176 VFG037177 VFG037203 VFG037218 VFG03728 VFG037202 VFG037302 VFG037302 VFG037316 VFG037330	

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 (11) 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**



Gx Find Resistance with	Nucleotide DB	×		
1. Choose where to run	Settings			
2. Select nucleotide sequences	DB IF Virulence genes (Updated Attributes)	3		
3. Settings	Minimum length % 60.0			
4. Result handling	Filter overlaps			
Help Re	set Previous Next Einish Cancel			

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.