Tutorial

Typing and Epidemiological Clustering of Common Pathogens

March 1, 2022

Sample to Insight

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

Tutorial

Typing and Epidemiological Clustering of Common Pathogens

This tutorial will take you through the tools available in CLC Microbial Genomics Module 22.0 (or higher) to perform typing and epidemiological studies of cultured bacteria.

Introduction Typing bacteria like *Salmonella enterica*, *Listeria monocytogenes*, *Vibrio parahaemolyticus*, *Escherichia coli*, *Shigella*, *Campylobacter* and *Cronobacter* allows surveillance in food safety and public health. Molecular methods using Next Generation Sequencing (NGS) data from whole pathogen genomes are increasingly used for outbreak detection of common pathogens. This tutorial will guide you through the different workflows and tools included in CLC Microbial Genomics Module to analyze NGS data from isolated and cultivated bacterial samples.

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.

Overview Using NGS data of cultured *Salmonella enterica*, this tutorial will guide you through the following:

- Creating metadata and analysis results tables.
- Customizing provided template workflows in order to:
  - Identify the best matching reference and its taxonomy.
  - Perform NGS-based Multilocus Sequence Typing (MLST).
  - Find antimicrobial resistance genes.
  - Identify potential contaminants in a sample.
- Performing outbreak analysis based on phylogenetic trees.
- Visualizing associated metadata in the context of the phylogenetic tree.

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 use a *Salmonella enterica* data set originally described by Leekitcharoenphon et al., 2014. To ensure a reasonable analysis time for the tutorial, only 5 of 47 samples

are included in this tutorial, and each read file has been reduced to include only 20% of the original reads.

The data for this tutorial includes the following files:

- **"MGM_metadata.xlsx"**: The metadata spreadsheet includes the sample metadata as stated in the original reference by Leekitcharoenphon and co-workers. Note that a metadata spreadsheet is different from a metadata table, and you will learn in this tutorial how to convert the former into the latter.
- **Raw reads**: 5 sequence data files in CLC format containing each 20% of the original *Salmonella enterica* reads.
- A **reference genome** NZ_CP014971 used for re-mapping
- **Databases**: All databases needed for typing are easily downloadable through specific tools of CLC Microbial Genomics Module. The data included in this folder is provided for users who wish to bypass the different download steps of this tutorial.

We can now get started.

1. Download the data from our website: `http://resources.qiagenbioinformatics.com/testdata/typing_tutorial/typing_tutorial_5.zip`. Unzip and save the files locally.
2. Start your CLC Workbench and go to **File** | **Import** | **Standard Import**. In the wizard, leave the import option to **Automatic Import**. Choose the folders called **Raw reads**, **Databases** and the **reference NZ_CP014971** (see figure 1) and save the imported files into a new folder you can call for example "Typing tutorial".

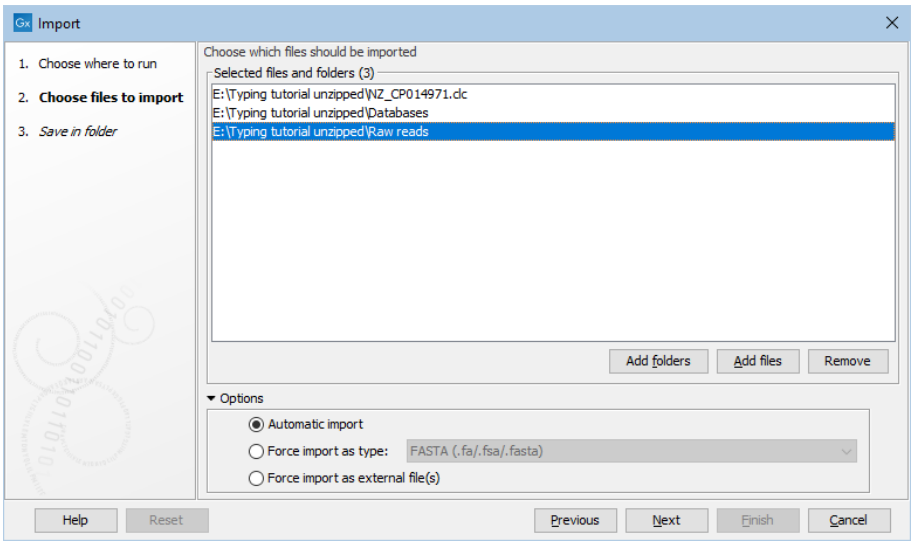

Figure 1: *Importing the reads and the reference*

3. Now import the metadata table via the toolbar: **File** | **Import** | **Import Metadata**.

- A wizard opens. In the first field (figure 2), select the spreadsheet saved on your local computer that contains the sample information "MGM_Metadata.xls". The contents of the Excel spreadsheet populates the table situated at the bottom of the dialog. Click **Next**.

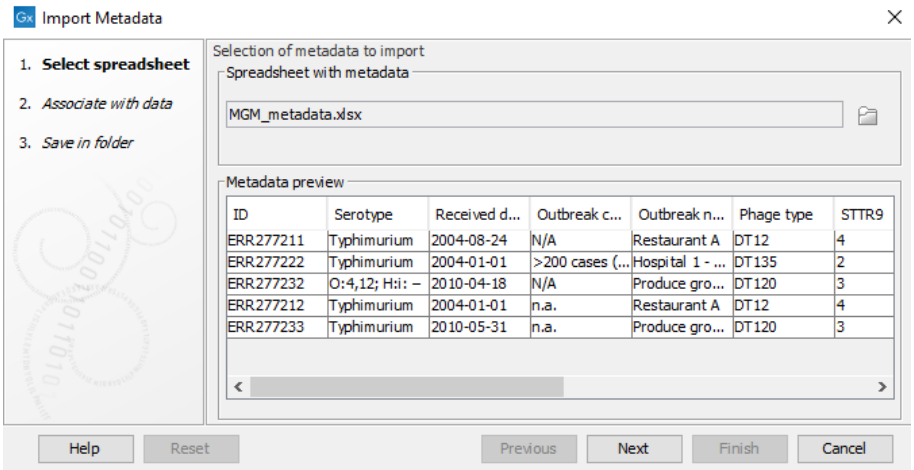

Figure 2: *Import the metadata spreadsheet.*

- Click on the Navigation button next to "Location of data", and select the imported reads in your Navigation Area. Click **OK**.
- The "Data association preview" table at the bottom of the dialog shows that the association between reads and metadata is successful (figure 3). Click **Next**.

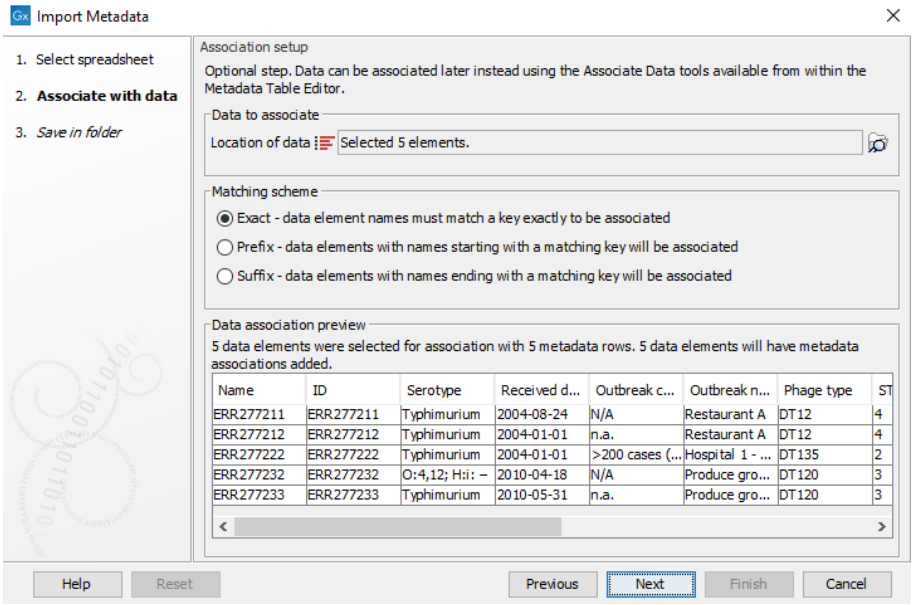

Figure 3: *The Import Metadata wizard showing a successful association between the reads and the metadata.*

- Select the "Typing tutorial" folder to save the "MGM_metadata" table.

4. The typing workflows in the Microbial Genomics Module require the use of a genome reference list, a resistance database and/or MLST schemes. The databases and schemes

needed to complete this tutorial are included in the "Databases" folder.

Remember that when you will work with your own data, you will download databases and schemes using the following tools found in the **Microbial Genomics Module | Databases ( )** folders of the Toolbox:

- **Drug Resistance Analysis ( ) | Download Resistance Database ( )**
- **Databases ( ) | MLST Typing ( ) | Download MLST Schemes ( )**
- **Taxonomic Analyses ( ) | Download Pathogen Reference Database ( )**

Creating the analysis Result Metadata Table

To proceed with the analyses, we need to generate a Result Metadata Table from the Metadata Table "MGM_metadata" imported earlier.

1. Go to:

   Typing and Epidemiology ( ) | Result Metadata ( ) | Create Result Metadata Table ( )

2. A dialog as shown in figure 4 is then displayed. Select the tutorial metadata table imported earlier, click **Next**, select **Save**, specify the location (the folder "Typing tutorial") and click **Finish**.

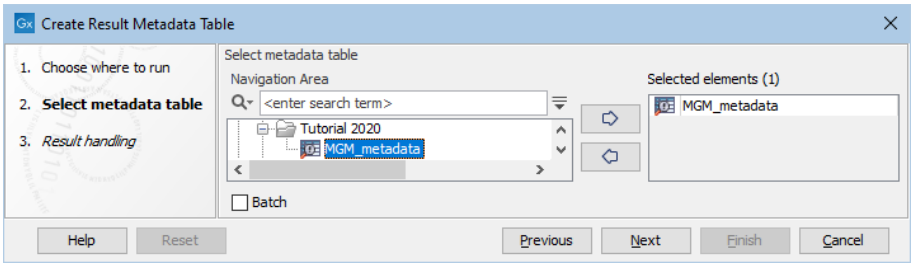

Figure 4: *Creation of a Result Metadata Table from a Metadata Table.*

   A new file called "MGM_metadata results" has now been created.

3. Open "MGM_metadata results". It is empty, as no analysis results have yet been generated.

4. Click the **Add Novel Samples** ( ) button to add your novel (i.e., not yet analyzed) samples to the Result Metadata Table. The following message will appear: *Any available novel samples have now been added* and all available novel samples are now shown with the available Metadata Table information in yellow (see figure 5).

5. Save the updated Result Metadata Table.

Now that all the reference data, databases and sample data have been downloaded/imported, and the Result Metadata Table created and organized as in figure 6, it is time to configure the provided template workflows so you can perform batch analysis of the example sequence data.

Rows: 5 Result Metadata Filter

ID	Serotype	Received date	Outbreak cases	Outbreak no. (Demo)	Phage type	STTR9	MLVA pattern	Accession
ERR277211	Typhimurium	2004-08-24	N/A	Restaurant A	DT12	4	JPX.0056.DK	ERR277211
ERR277222	Typhimurium	2004-01-01	>200 cases (Outbreak)	Hospital 1 - Ward 6	DT135	2	JPX.0855.DK	ERR277222
ERR277232	O:4,12; H:i: –	2010-04-18	N/A	Produce grower	DT120	3	JPX.0005.DK	ERR277232
ERR277212	Typhimurium	2004-01-01	n.a.	Restaurant A	DT12	4	JPX.0056.DK	ERR277212
ERR277233	Typhimurium	2010-05-31	n.a.	Produce grower	DT120	3	JPX.0005.DK	ERR277233

Figure 5: *Clicking on the Add Novel Samples button and all the novel Salmonella samples are added to the previously empty Result Metadata Table.*

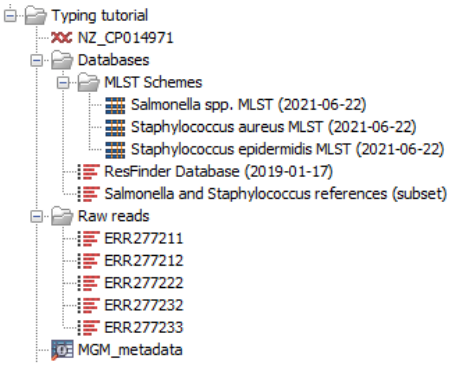

Figure 6: *Example data has been organized within the Navigation Area.*

How to run the Type Among Multiple Species workflow on a batch of samples:

The **Type Among Multiple Species** workflow is designed for typing a sample among multiple predefined species. The workflow allows identification of the closest matching reference species among the user-specified reference list(s), and of potential contaminants. The workflow also identifies the associated MLST scheme and type, determines variants found when mapping the sample data against the identified best matching reference, and finds occurring antibiotic resistance genes if they match genes within the user-specified resistance database.

To ensure the same workflow parameters are used each time it is beneficial to make a copy of the template workflow and save the copy in the Navigation Area before running it. This is described in detail below. Note that while having a saved copy of the workflow in the Navigation Area, and an open view of the layout open in the View Area are necessary to run the workflows, the configuration is optional and can be done later in the successive dialogs of the workflow wizard.

1. Select the workflow **Type Among Multiple Species ( )** found in the **Template Workflows** | **Microbial Workflows** ( ) | **Typing and Epidemiology** ( ) folder with one click (do not open the wizard yet with a double click). Right-click on the name of the workflow and choose the option **Open Copy of Workflow**.
2. This opens a copy of the workflow in the view area of your workbench.
3. Configure and possibly lock any parameters and inputs that remain the same for each use of the workflow. In this tutorial nothing has to be configured.
4. Make sure not to configure the green tile representing the Result Metadata Table input file (see figure 7).

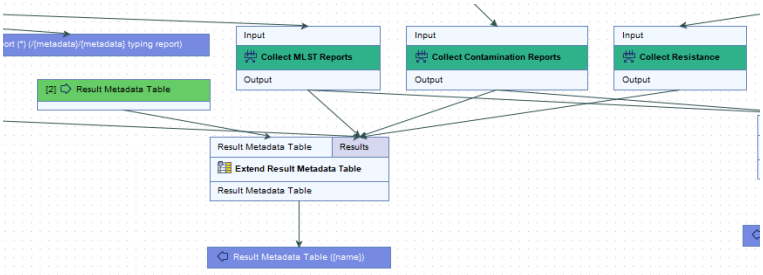

Figure 7: *Do not preconfigure the Result Metadata Table green input file tile as an updated version has to be used each time the workflow is run.*

5. Save your workflow in your Navigation Area by dragging the workflow tab to the relevant location in your Navigation Area (here in the folder called "Typing tutorial"). You can also rightclick on the workflow copy tab and select "Save as...".

6. Switch back to your Result metadata table and select the 5 samples to be typed (figure 8).

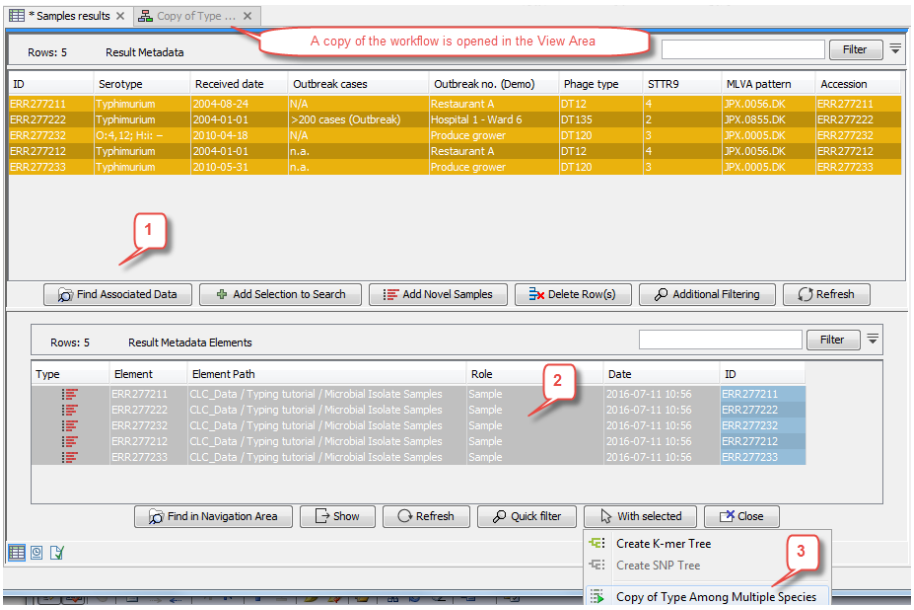

Figure 8: *In just a few clicks, select your samples, find the associated data and start the workflow with the relevant input files directly from the Result Metadata Table.*

7. Click on **Find Associated Data** ( ) button. It opens a Result Metadata Elements table.

8. Select the five reads files that have a role defined as "Sample" in the Result Metadata Elements table. Click on the **With selected** ( ) button and select the **Copy of Type Among Multiple Species** workflow. Note that a copy of this workflow needs to be opened in the View Area for this option to be available.

9. It will open a wizard where the 5 samples are pre-selected. As the workflow performs internal batching make sure not to tick the option **Batch** (highlighted in figure 9) before clicking on the button labeled **Next**.

10. The next wizard window gives you an overview of the samples selected. We want to analyze all 5 samples so we just click **Next**.

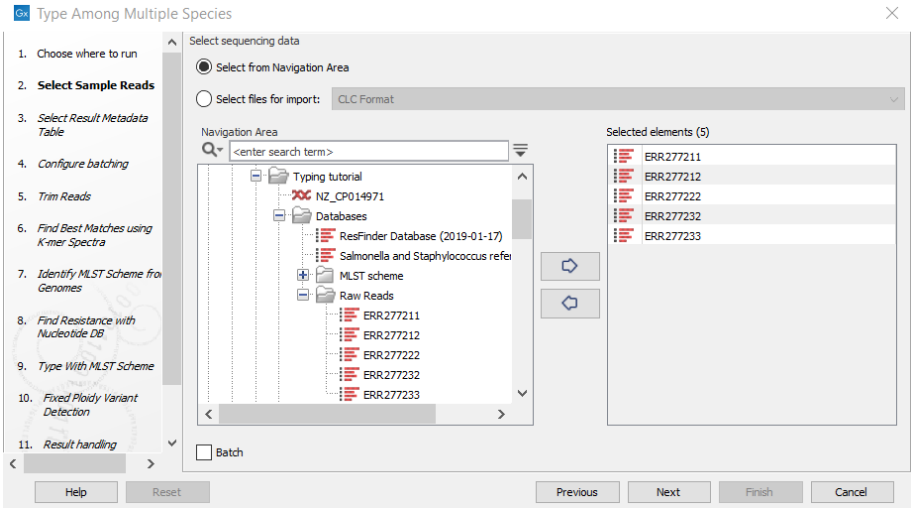

Figure 9: *Remember not to check the button labeled Batch highlighted at the bottom of the wizard window when selecting multiple samples as the workflow performs internal batching.*

11. In the next dialog you must select "MGM_metadata results" created earlier (see figure 10). The workflow will create a copy of "MGM_metadata results". This newer version must be used in the next analysis steps. Click **Next**.

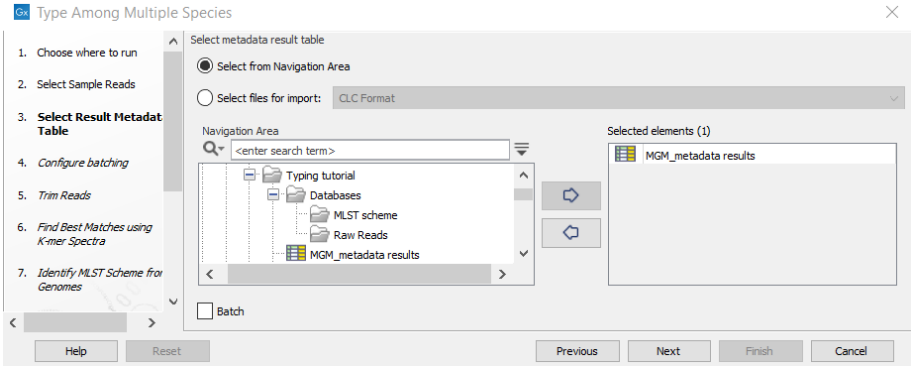

Figure 10: *Select the Result Metadata Table you created for running this workflow (here called "MGM_metadata results")*

12. In the batching dialog, select "Use organization of input data" as we have an input file for each sample (see figure 11). If multiple files existed per sample, "Use metadata" could be used and a column uniquely identifying each sample should be selected.

13. The next wizard window gives you an overview of the samples selected. We want to analyze all 5 samples so we just click **Next**.

14. Leave the parameters as default in the "Trim Reads" window and click **Next**.

15. In the next wizard window, select from the "Databases" folder the list called "**Salmonella and Staphylococcus reference (subset)**" (figure 12) to be used by the Find Best Matches using K-mer Spectra tool.

16. In the "Identify MLST Scheme from Genomes" dialog, you should select the downloaded

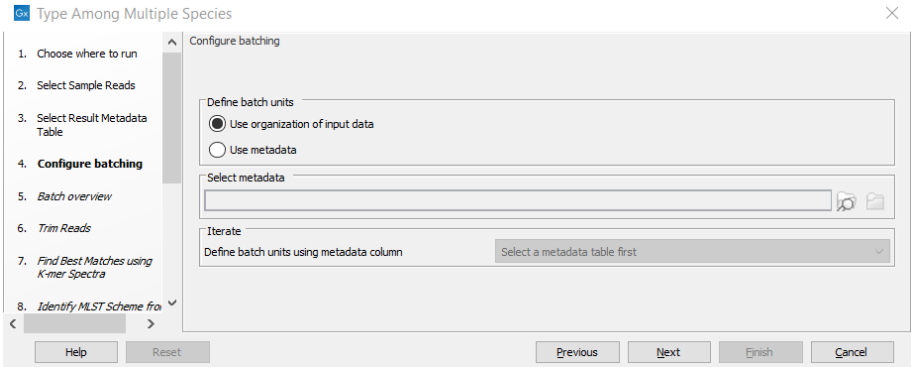

Figure 11: *As each sample only has one input file the option "Use organization of input data" can be used to define each batch.*

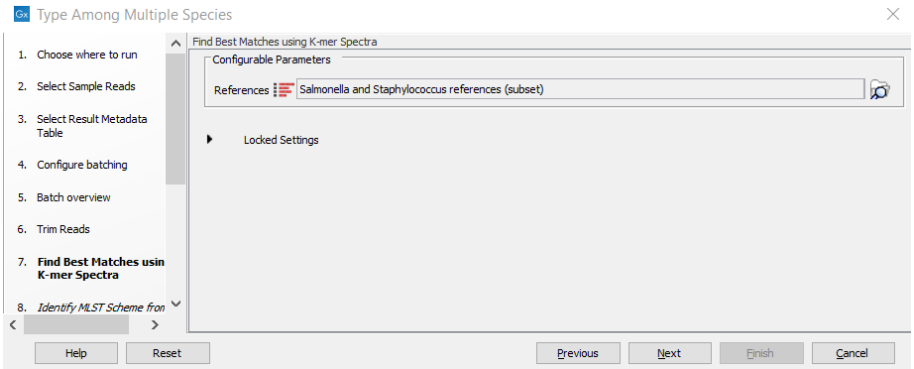

Figure 12: *Choose the genome references called "Salmonella and Staphylococcus reference (subset)".*

MLST schemes for *Staphylococcus aureus*, *Staphylococcus epidermis* and *Salmonella enterica* (figure 13).

17. In the "Find Resistance with Nucleotide DB" window you should specify the provided resistance database called "ResFinder Database (2019-01-17)" (figure 14).

18. Leave parameters as they are set by default in the "Type With MLST Scheme" and "Fixed Ploidy Variant Detection" windows, and click **Next**.

19. In the last wizard window "Result handling" simply click on **Finish** (figure 15). Choose to save the workflow output to a new folder, for example titled "Analysis results". A folder with results for each input sample will automatically be created by the workflow.

The output of the workflow consists of:

- A folder with results for each input sample.
- A combined report with overview tables from all of the samples.
- A copy of the Result Metadata Table (originally named "MGM_metadata results") with all new results produced by the workflow and an updated timestamp in the name.

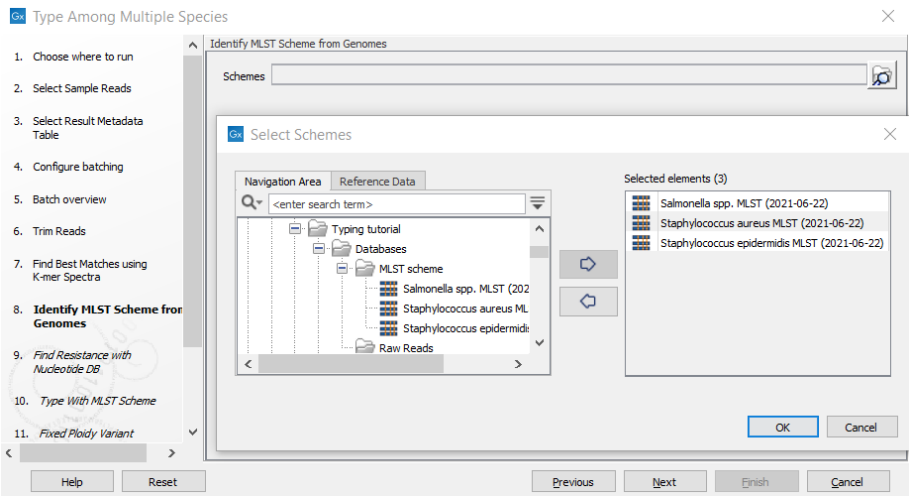

Figure 13: *Select the downloaded MLST schemes for Salmonella and Staphylococcus.*

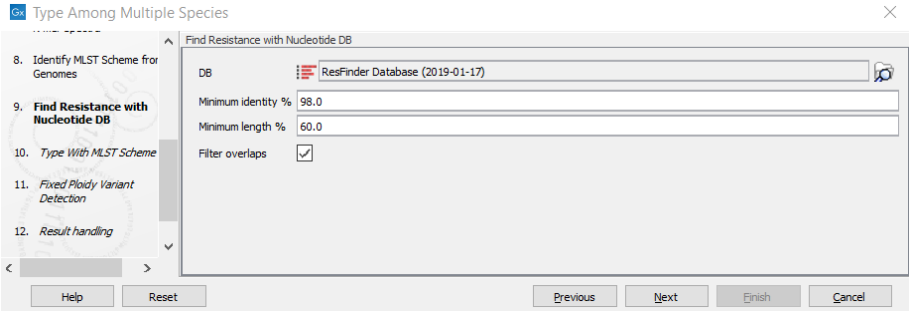

Figure 14: *Select the resistance database called "ResFinder Database (2019-01-17)".*

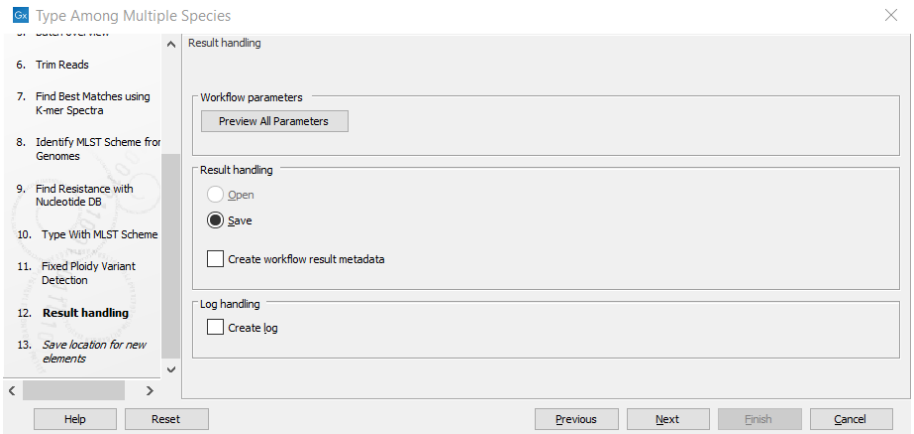

Figure 15: *A folder with results for each input sample will automatically be created by the workflow*

Checking results in the Result Metadata Table - optional

Once the typing analyses are done, each typing analysis results in a series of output files (elements), accessible both from the Navigation Area, and from the lower part of the Result Metadata Table (click on the **Refresh** button to see them). In addition, the main findings are also summarized in the Result Metadata Table as new columns are added next to the yellow

metadata columns.

Open the Result Metadata Table (for instance the latest copy could be named "MGM_metadata results 2021-06-22 10:57") and look at the results columns (in white). Adjust the visualization of the table according to your own preference using the panel to the right. For example, you can change column width and columns displayed. Use "Show column" (for individual columns) or "Show column groups" (for selecting entire groups of columns). See the following examples:

- "Resistances found" to highlight selected resistances.
- "Best matches" to see the best matching reference and its taxonomy in the "Best match, Description" column, as well as to identify potential contamination in the samples.
- "MLST Scheme" to see which schemes were associated with the data, and whether this association was conclusive or not.
- "Metadata" for all metadata columns that were included in the original spreadsheet.

Finding contamination using the functionalities of the Result Metadata Table

For a better overview, first click on "Deselect All", then select the columns "Best match", "Best match, Species", "Best match, % mapped" and "Contaminating species, % mapped" (figure 16). In the **Show column groups** section of the Side Panel, check the group "Metadata".

Best match	Best match, Species	Best match,...	Contaminating species, % ma...	MLST Scheme	ID	Serotype
Salmonella enterica subsp. enterica serovar Typhimurium str. USDA-ARS-USMARC-1898	Salmonella enterica	94		Salmonella spp. MLST (2021-06-22)	ERR277211	Typhimurium
Salmonella enterica subsp. enterica serovar Typhimurium str. UK-1	Salmonella enterica	98		Salmonella spp. MLST (2021-06-22)	ERR277222	Typhimurium
Salmonella enterica subsp. enterica serovar Typhimurium str. USDA-ARS-USMARC-1898	Salmonella enterica	49	41 (Staphylococcus aureus)	Salmonella spp. MLST (2021-06-22)	ERR277232	O:4,12; H:i: -
Salmonella enterica subsp. enterica serovar Typhimurium str. USDA-ARS-USMARC-1898	Salmonella enterica	93		Salmonella spp. MLST (2021-06-22)	ERR277212	Typhimurium
Salmonella enterica subsp. enterica serovar Typhimurium	Salmonella enterica	96		Salmonella spp. MLST (2021-06-22)	ERR277233	Typhimurium

Figure 16: *Finding contamination in the Result Metadata Table.*

You can see that the sample ERR277232 has only half of its reads mapped to *Salmonella enterica* (figure 16), while others mapped to *Staphylococcus aureus*. Select the ERR277232 row In the Elements table, click on "Find associated Data" and in the bottom table, find the element whose role is "Contamination report". Double-click on it to learn more about this particular sample, and decide whether you want to exclude it from subsequent analyses.

Note that the above conclusion could also be made from looking at the Contamination section of the combined report which provides an overview of the samples analyzed as part of the same workflow run.

Exploring the obtained Best match results and identifying a common reference

Go back to the updated Result Metadata Table and clear all filters before looking at the "Best match" column.

If all entries in the "Best match" column are the same: This indicates that the read files represent a single clade, and it is possible to create a SNP tree directly from the typing analysis data. Note that this is not the case for this tutorial.

If the "Best match" column includes a different strain: This indicates that the *Salmonella* read files in this tutorial do not represent a single clade. However, creating a SNP tree including multiple clusters requires that all read files were mapped and variants called using the same reference.

If you are under time constraints, you can go directly to the section on "Compare Variants Across Samples" and use the NZ_CP014971 genome sequence provided in the folder "Databases" for re-mapping and variant calling of the read files. Otherwise, follow the steps below to learn how you can identify a common reference by creating a k-mer tree.

First, you need to create a *Salmonella*-specific reference subset to be used for k-mer tree generation. In this tutorial, you will use the "Salmonella and Staphylococcus reference (subset)", but otherwise you would can create your own reference database using the **Download Custom Microbial Reference Database** tool.

1. Open the "Salmonella and Staphylococcus reference (subset)". This file opens as a Sequence list.
2. Click on the **Show table** icon at the bottom of the View Area (highlighted in red in figure 17).

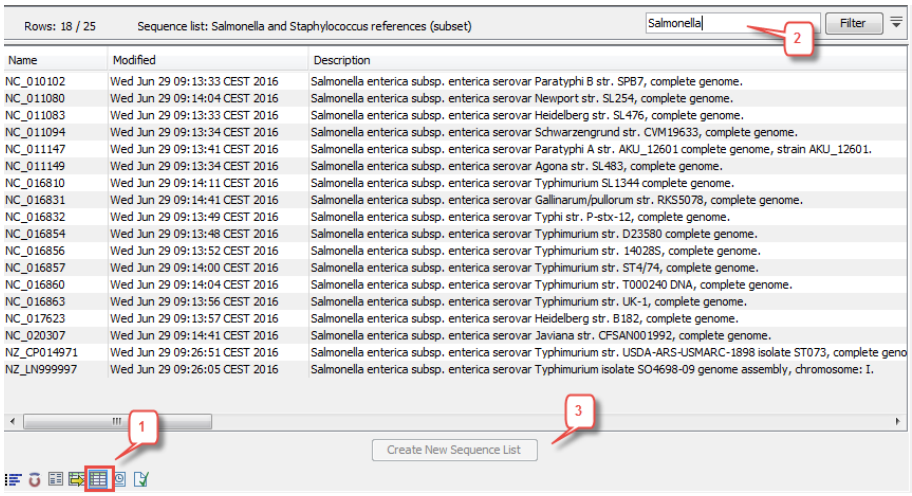

Figure 17: *Selection of Salmonella specific genomes for subset reference list to be used for k-mer creation.*

3. Filter on the term "*Salmonella*".
4. Select the remaining sequences, click on the **Create New Sequence List** button. It opens a new tab called "Salmonella and Staphylococcus reference list (subset) subset". **Save** the list by dragging it to the Navigation Area and rename it to "Salmonella references subset" or save by rightclicking the tab and selecting "Save as...".

You can now create a K-mer tree through the Result Metadata Table:

1. Open the latest Result Metadata Table and select the five samples to which a common best matching references should be identified. Note that we decided to leave the contaminated sample in the analysis, but when working with your own data, you could sort the table based on the "Contaminating species, % mapped" column and select only the samples that are below a certain threshold of contamination for example.
2. Click on the **Find Associated Data** ( ) button to find the 65 associated Metadata Elements.
3. Click on the **Advanced filter drop-down button** ( ) button and filter for Role contains "sequence list" and Element contains "paired".

4. Select all remaining Metadata Element files.

5. Click on the **With selected** ( ) button and select the **Create K-mer Tree** action to open the tool's wizard (figure 18.

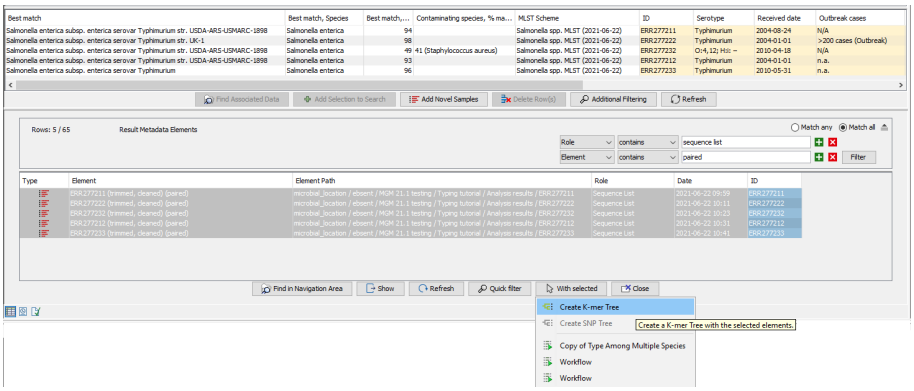

Figure 18: *Select sample reads as well as the Salmonella specific reference list.*

6. Leave the parameters as set in the second wizard window (figure 20).

7. In the first wizard window, the samples are pre-selected. Do not forget to add the *Salmonella* specific reference list (figure 19).

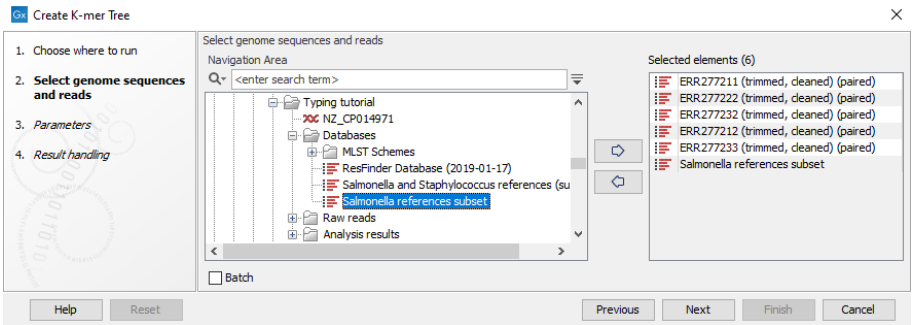

Figure 19: *Select sample reads as well as the Salmonella specific reference list.*

8. Leave the parameters as set in the second wizard window (figure 20).

9. **Save** your results in the folder "Typing tutorial".

Open the K-mer tree. Choose a reference genome that shares the closest common ancestor with the clade of isolates under study. For this tutorial we choose to use reference NZ_CP014971 (highlighted in figure 21) as the common reference in the following sections.

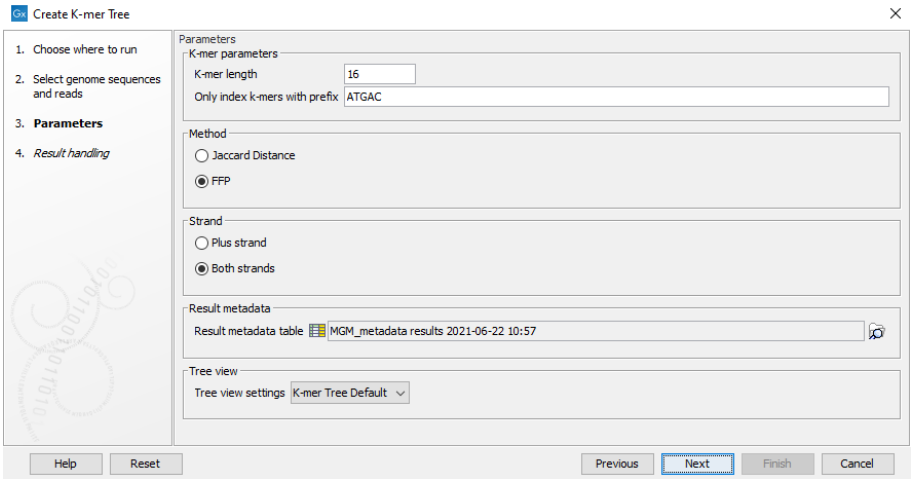

Figure 20: *Default parameters for the "Create K-mer Tree" tool, including the view setting set to K-mer Tree Default.*

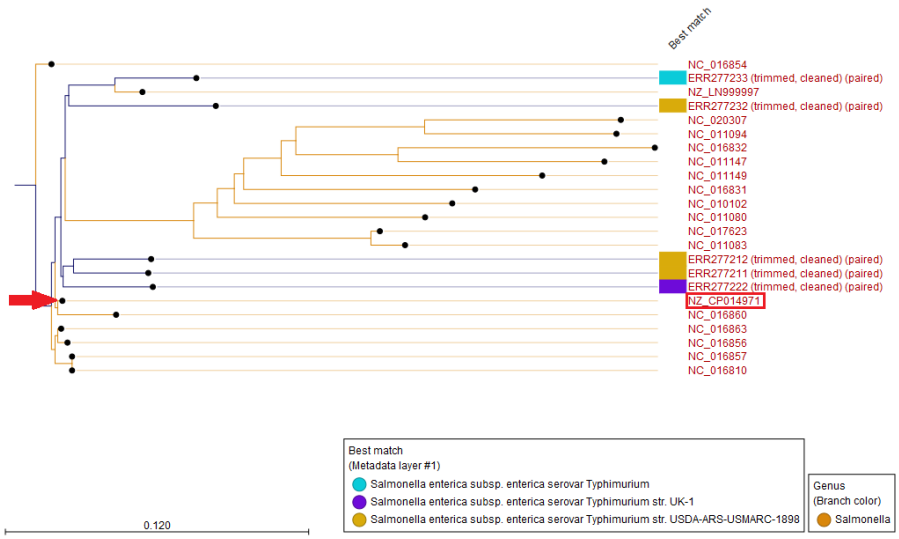

Figure 21: *K-mer Tree with NZ_CP014971 highlighted in red.*

Compare Variants Across Samples

As we saw with previous workflows, if certain parameters of the **Compare Variants Across Samples** workflow are customized it can be beneficial to make a copy and save it in the navigation area. However remember not to specify your Result Metadata Table.

1. Select the workflow **Compare Variants Across Samples** in the Toolbox, right-click on the name and choose the option **Open Copy of Workflow** (figure 22).
2. This opens a copy of the workflow in the view area of your workbench. Since we are working with downsampled data, we will lower the coverage required to construct the SNP tree. To do so, double-click to open the "Create SNP Tree" element (figure 23).
3. Set the Minimum coverage required in each sample to 10 (figure 24). Optionally, you can also click (🔒) to unlock the Result metadata table.

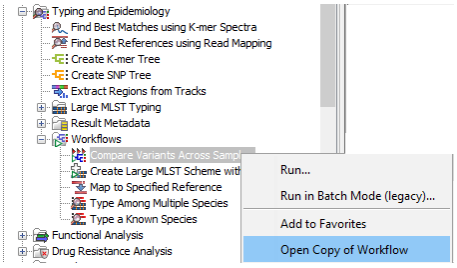

Figure 22: *Open a copy of the Compare Variants Across Samples workflow.*

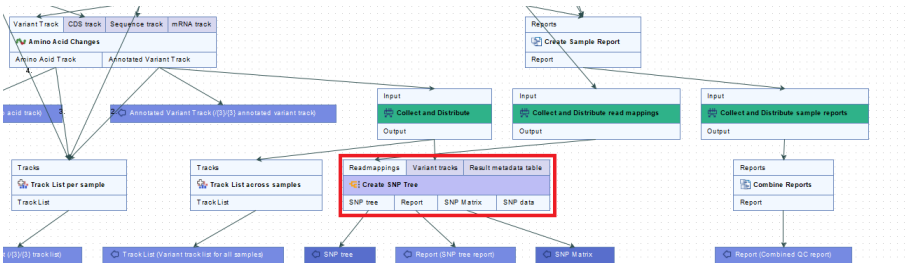

Figure 23: *Double-click to configure the "Create SNP Tree" element*

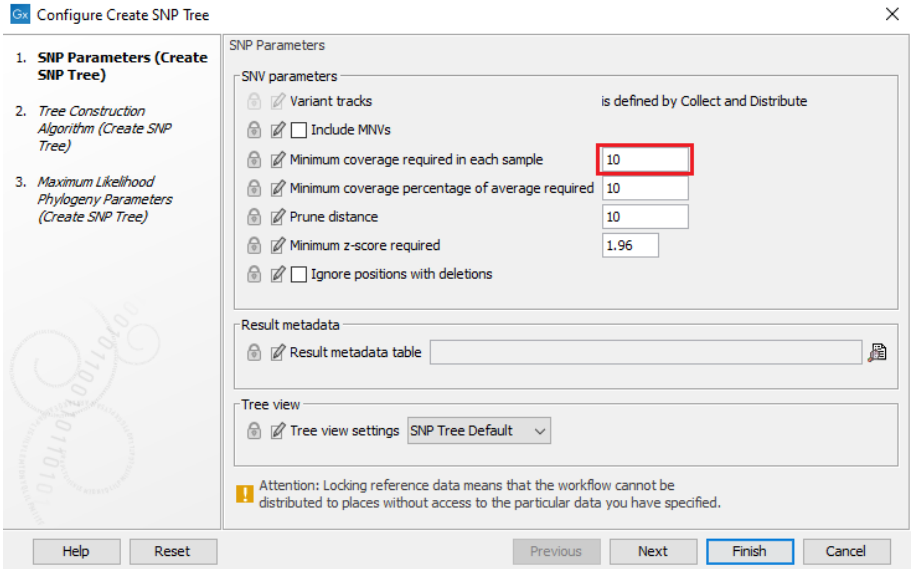

Figure 24: *Set "Minimum coverage required in each sample" to 10.*

4. **Save** the workflow by simply dragging the tab to the relevant location in your Navigation Area.

5. The workflow requires a reference genome track and a CDS. This can easily be created from the NZ_CP014971 sequence list. To do so, run **Track Tools ( ) | Track Conversion ( ) | Convert to Tracks ( )**

6. Select the "NZ_CP014971" sequence list as input and click **Next**.

7. Check "Create sequence track" and "Create annotation tracks". In "Annotation types", select CDS (figure 25). Click **Next** and save the tracks in the "Typing tutorial" folder.

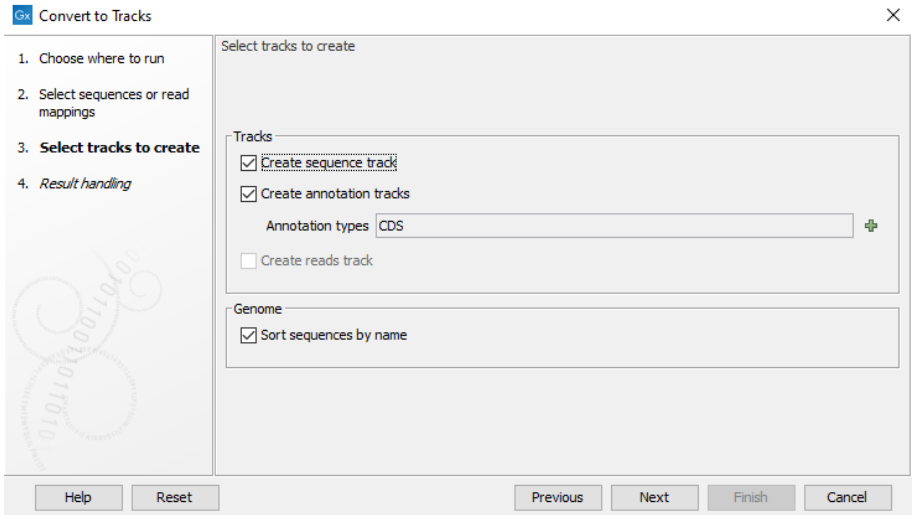

Figure 25: *Create sequence and CDS tracks from a sequence list.*

8. Switch back to the result metadata table.
9. Click on the **Filter** button and filter for Role contains "sequence list" and Element contains "paired".
10. Click on the **With selected** ( ) button and select the **Copy of Compare Variants Across Samples** workflow (see figure 26).

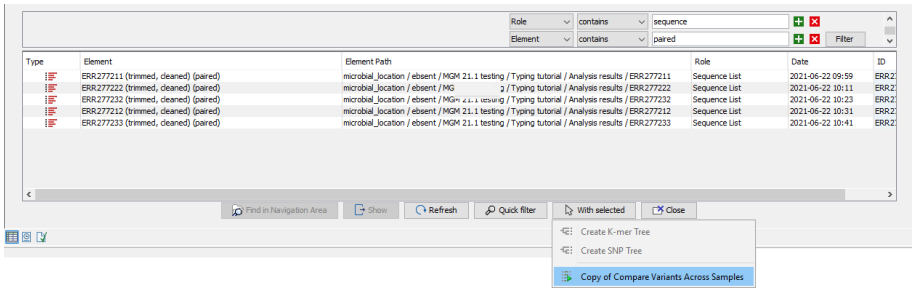

Figure 26: *The reads you want to map to the reference will be pre-selected in launched workflow wizard.*

11. A wizard will appear with the samples pre-selected because you started the workflow directly from the Metadata Elements table. Remember not to check the option **Batch**.
12. In the next step specify the reference sequence to "NZ_CP014971 (Genome)" (figure 27). Click on **Finish**.
13. In the next step specify the "NZ_CP014971 (CDS)" track (figure 27).
14. Next batch units must be defined. Leave it set as "Use organization of input data". Click **Next**.
15. The next wizard window gives an overview of the selected samples. Click **Next**.
16. If you unlocked the Result metadata table, you see the "Create SNP tree". Here, you will be able to select the metadata table from the Analysis folder.

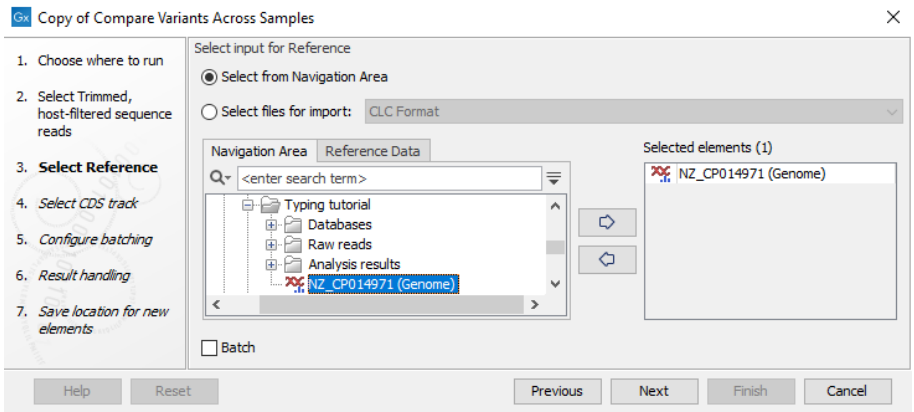

Figure 27: *Specify the reference genome track of "NZ_CP014971".*

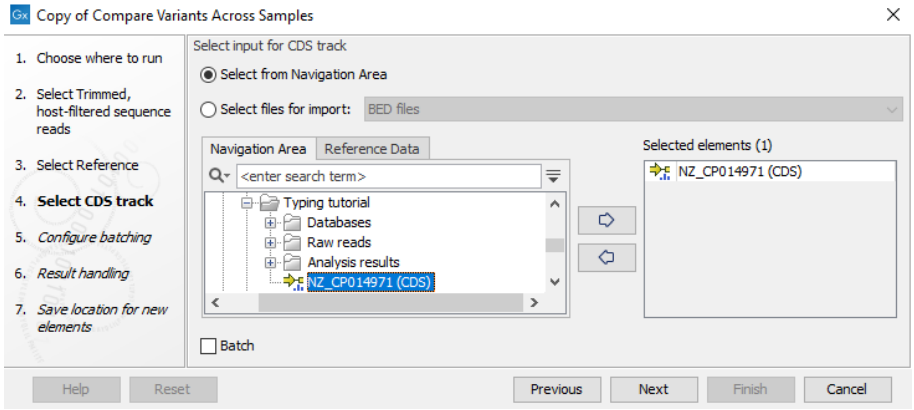

Figure 28: *Specify the CDS track of "NZ_CP014971".*

17. In the last wizard window "Result handling" simply click on **Finish** and choose a save location, for example you can create a "SNP tree" folder.

The workflow creates a subfolder for each sample with variant calls and read mappings. In the top folder, you will find a SNP tree and SNP matrix as well as a combined report. Note that the tool will output, among other files, variant tracks. It is possible to export multiple variant track files from monoploid data into a single VCF file with the Multi-VCF exporter. This exporter is installed as part of the Microbial Genomics Module. All variant track files must have the same reference genome for the Multi-VCF export to work.

Tree visualization

Once the Compare Variants Across Samples has finished you will have a SNP tree for the 5 samples.

Open the generated SNP tree file and explore the settings in the right hand side panel to fit the visualization to your needs. An example is shown in figure 29. The following settings where chosen:

- Tree layout - Ordering as Increasing

- Metadata - Label text as Name
- Metadata layer #1 Sequence type (available if the SNP tree was created using generated Result metadata table)
- Metadata layer #2 Serotype
- Metadata layer #3 Outbreak no. (Demo)

Color coding can be modified by clicking on the associated color marks. Read more on Tree Settings in general here: http://resources.qiagenbioinformatics.com/manuals/clcgenomicsworkbench/current/index.php?manual=Tree_Settings.html.

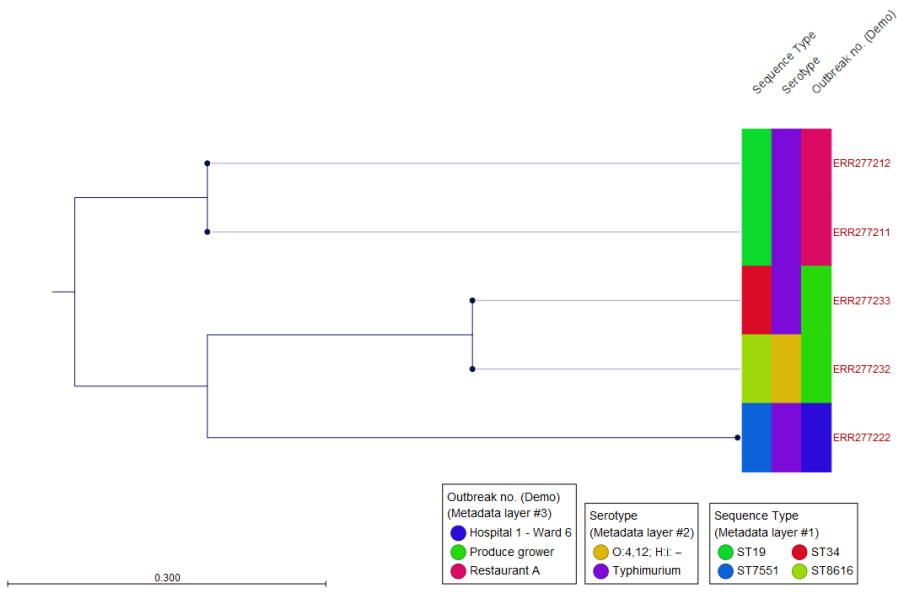

Figure 29: *Visualization of metadata as well as analysis results on the SNP tree.*

Results and metadata available during tree generation can also be used to explore and decorate this epidemiologically relevant information on the phylogenetic tree.

References

[Leekitcharoenphon et al., 2014] Leekitcharoenphon, P., Nielsen, E. M., Kaas, R. S., Lund, O., and Aarestrup, F. M. (2014). Evaluation of whole genome sequencing for outbreak detection of salmonella enterica. *PLoS One*, 9(2):e87991.