QIAGEN

Tutorial

Working with large MLST schemes

June 30, 2020

Sample to Insight

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

Working with large MLST schemes

This tutorial is an introduction to multilocus sequence typing (MLST) with many loci, using the Large MLST Scheme tools of the CLC Microbial Genomics Module. These tools can be used for both core genome (cgMLST) and whole genome (wgMLST), as well as working with classic 7-gene schemes. Typing can be applied to NGS reads of an isolate, or to an assembly of an isolate.

In this tutorial, we cover:

- Creating a large MLST scheme from a set of references and adding sequence types
- Detecting resistance with a resistance database
- Typing reads using a large MLST scheme
- Adding a typing result to an existing scheme
- Downloading an existing large MLST scheme
- Exporting and importing large MLST schemes
- Adding annotations to references before using them to create a large MLST scheme

Please refer to the CLC Microbial Genomics Module manual for full information about the Large MLST tools.

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 by selecting the 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* 20.0 or higher and 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 *Klebsiella aerogenes*, a bacterium normally found in the gastrointestinal tract. It may cause opportunistic infections and has known antimicrobial resistance. NCBI lists hundreds of genomic references with varying degrees of similarity. For the sake of time and simplicity, we have selected nine assemblies to use. When building your own schemes, we advise using 30-50 high quality reference assemblies. The references should include as many strains as possible.

1. To get started, download the sample data from: http://resources.qiagenbioinformatics.com/testdata/Large_MLST_tutorial_data.zip and unzip it.
2. Open the *CLC Genomics Workbench*.
3. Create a new folder for the tutorial data, for example named "Large MLST tutorial".
4. Import the references using the standard importer:
   (a) Go to: **File** | **Import** | **Standard Import...**
   (b) Select the References folder from the tutorial data you downloaded and click on **Next**.
   (c) Select the folder you created earlier to save the imported data to and click on **Finish**.
5. Import the paired reads from the Reads folder of the tutorial data:
   (a) Go to:
   File | **Import** | **Illumina...**
   (b) Click on "Add folders" and select the Reads folder from the tutorial data you downloaded.
   (c) Enable the "Paired reads" option under General options. Leave the other options set to their default values and click on **Next**.
   (d) Choose to save the imported data and click on **Next**.
   (e) Create a new subfolder to save the imported data to, named for example, "Reads", and click on **Finish**.
6. (Optional) If you do not already have an antimicrobial resistance database, you can download the QMI-AR database. To do this, run the following tool from the Toolbox:
   Microbial Genomics Module ( ) | **Databases** ( ) | **Drug Resistance Analysis** ( ) | **Download Resistance Database** ( )
   As the use of QMI-AR database is not limited to this tutorial, you may wish to save it to your general database location.

You should now have two folders containing data you imported. The "References" folder should contain a sequence list of nine reference sequences downloaded from NCBI. The "Reads" folder should contain three sequence lists, one for each set of paired reads. These reads were originally downloaded from SRA. Two of the sequence lists contain reads from strains represented in the references you just imported. The third set of reads originates from a novel strain. For the sake of time, only 250,000 pairs are included per read set.

You can create your own set of references using the **Download Pathogen Reference Database** or **Download Microbial Reference Database** tool, or using assemblies created using other tools available in *CLC Genomics Workbench*. The references should be available as one assembly per sequence list.

Create and populate a large MLST scheme

In this section, we create a large MLST scheme and then add sequence types to it.

Creating the Large MLST scheme

1. From the Toolbox, choose:

   Microbial Genomics Module ( )| **Databases** ( ) | **Large MLST** ( ) | **Create Large MLST Scheme** ( ).

2. Select all the elements in the "References" folder and then click on **Next**.

3. Keep the Assembly Grouping option as the default: "Group sequences by annotation types", and keep the Assembly annotation type as the default, "Assembly ID". Click on **Next**.

4. Select the "Whole genome" option to include genes found in minimum 10% of input genomes and the "Search alleles before clustering" option to maximize gene detection, as shown in figure 1. Searching alleles before clustering is a thorough check but may be time consuming when creating a scheme with many genomes as input. Click on **Next**.

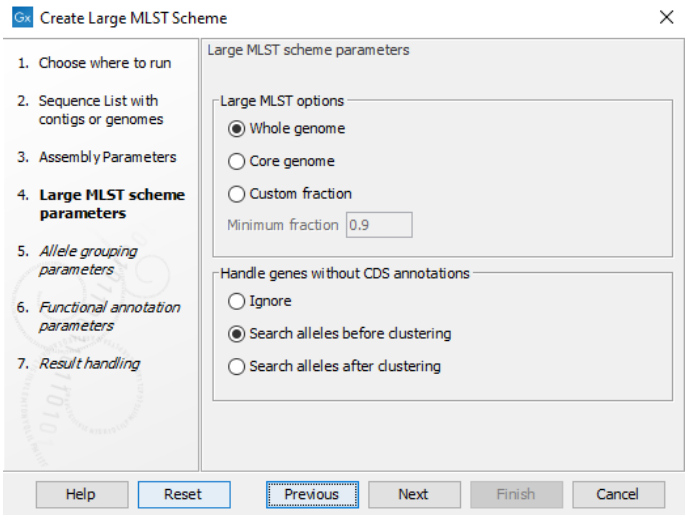

Figure 1: *Create Large MLST scheme options*

5. For the Translation table, select the genetic code "11 Bacterial, Archaeal and Plant Plastid", and enable the "Check codon positions" option.

6. For the DIAMOND options, keep the default "Minimum identity" value of 0.2, and select the "More sensitive search" option, as shown in figure 2.

   Note: when creating larger schemes, you may wish to use the "Sensitive search" option in DIAMOND settings in the interest of time.

7. Click on **Next**.

8. (Optional) Enter the QMI-AR Nucleotide Database for the "Antimicrobial resistance database" option, as shown in figure 3.

   Doing this will lead to the annotation of loci with known resistance information.

9. Click on **Next**.

10. Choose to save the scheme to a new subfolder, for example named "wgMLST schemes" and click on **Finish**.

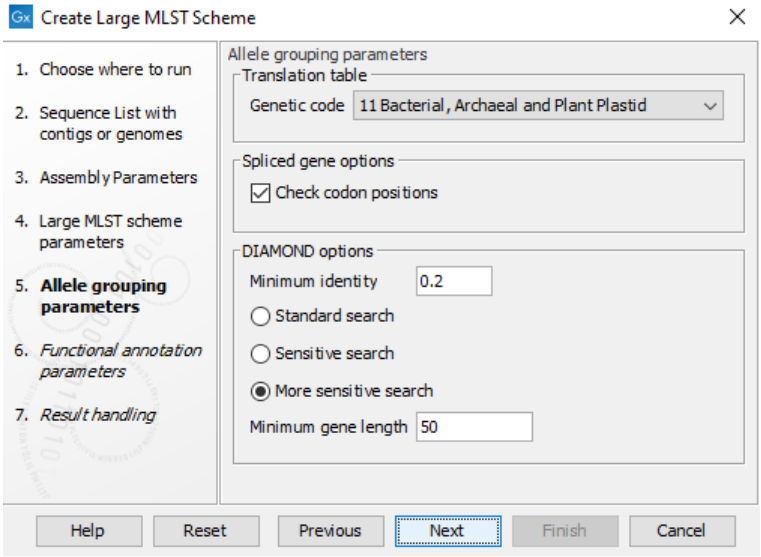

Figure 2: *Create Large MLST scheme allele grouping options*

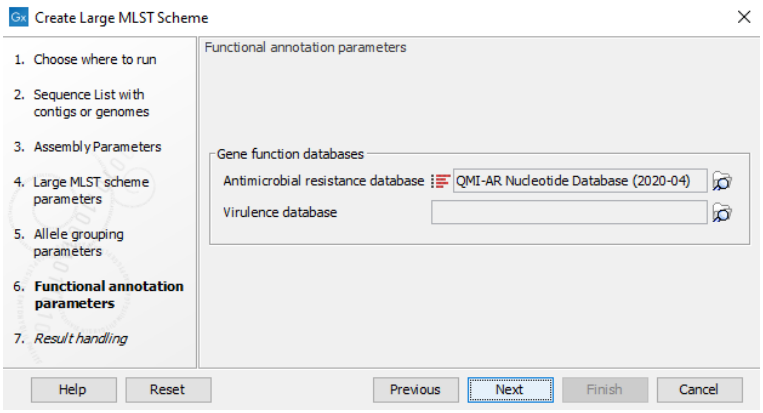

Figure 3: *Select the QMI-AR resistance database*

Create Large MLST Scheme will now run. It can take some time. You can monitor its progress in the Processes tab, located next to the Toolbox in the bottom left side of the Workbench. The **Create Large MLST Scheme** tool uses DIAMOND to establish the loci and create an initial set of alleles. It creates two outputs: a report and a scheme. We will look at those quickly before proceeding further.

The outputs of Create Large MLST Scheme

11. Open the report.
12. Under "Sequence and assembly information", you can check the number of input sequence and input assemblies. If the number of input assemblies do not match that expected, check that your sequences are correctly annotated.
13. Under "Allele filtering information", you get an overview of the alleles excluded from the scheme and the reason why.

14. Under "Allele grouping information", you get an overview of the loci and the results after loci filtering. This information can be used to assess your input assemblies. In wgMLST scheme creation for example, if a large number of loci are not present in at least 10% of the assemblies, you might need higher quality assemblies as basis for the scheme.

15. Open the MLST scheme ( ). As expected, it does not yet contain any sequence types and the only available view is the allele table. We will walk through the different views in the scheme after adding sequence types.

16. Close the report and MLST scheme.

We will now call the remaining alleles and add sequence types to the MLST scheme. We do this by typing all the references against the empty scheme and then adding those results to the scheme.

Calling remaining alleles

17. From the Toolbox, choose:

    Microbial Genomics Module ( )| **Typing and Epidemiology** ( ) | **Large MLST Typing** ( ) | **Type with Large MLST Scheme** ( )

18. Select the nine references from the "References folder", check the **Batch** box in the bottom, left side of the wizard and click on **Next**.

    When you choose to run a tool in batch mode, it will run once for each "batch unit". At this step of the wizard, you are presented with the list of batch units, so you can check that the tool will run as expected. Here, you should see a list of the 9 references you selected. This means the tool will run nine times, each time using one of the references as input.

19. Locate and select the large MLST scheme created earlier by clicking on the ( ) icon.

20. Leave the other settings as default, as shown in figure 4 and click on **Next**.

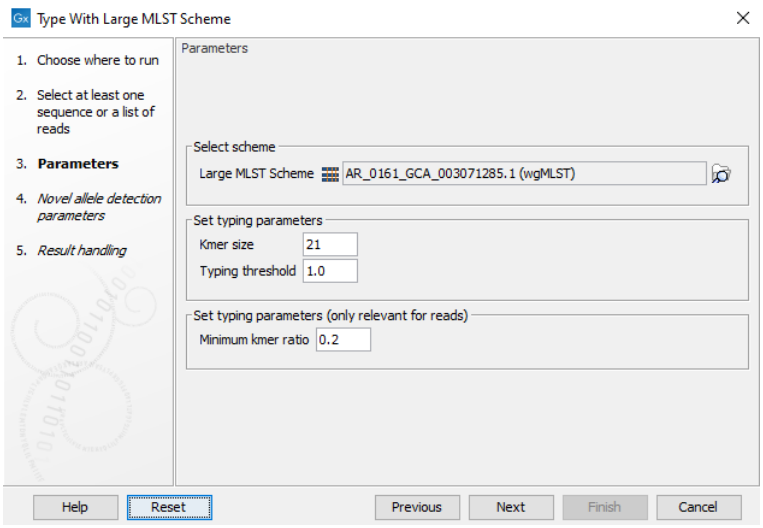

Figure 4: *Select the scheme created in previous steps*

21. Enable the "Search for novel alleles" option and lower the value for "Minimum required fraction of kmers" to 0.2, as shown in figure 5.

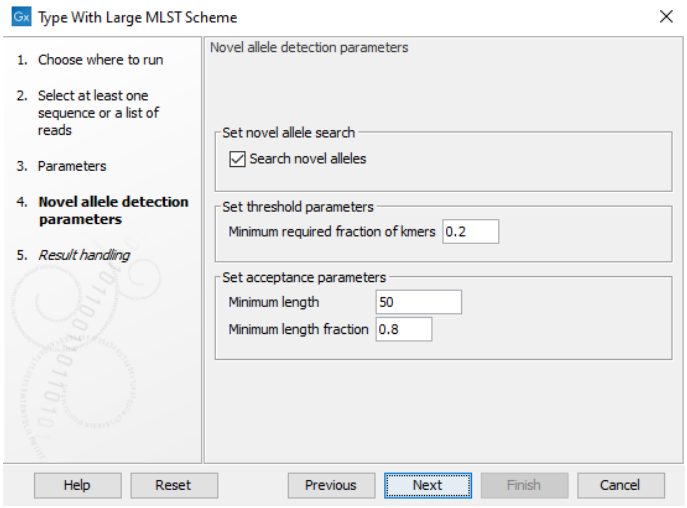

Figure 5: *Search for novel alleles in the created scheme*

22. Click on **Next**

23. Choose to save the results to a location that you will specify and click on **Next**.

24. Specify where results should be saved to, for example a new folder called "Typing for wgMLST creation", and click on **Finish**.

The tool will now run 9 times, once for each of the references input. You can monitor its progress in the Processes tab, located next to the Toolbox in the bottom left side of the Workbench.

Each run of the tool will generate a typing report and a typing result. Feel free to review the reports if you wish. We will use the typing results in the next section.

Adding sequence types to the scheme

We will now add sequence types to the scheme. We do this by using the 9 typing results as input to a single run of the **Add Typing Results to Large MLST scheme** to create a single updated scheme.

25. From the Toolbox, choose:

    Microbial Genomics Module ( )| **Typing and Epidemiology** ( ) | **Large MLST Typing** ( ) | **Add Typing Results to Large MLST scheme** ( )

26. Select as input the nine typing results you just created from the "Typing for wgMLST creation" folder and click on **Next**.

    (Do not check the Batch box. We are using all the inputs in a single run of this tool.)

27. Locate and select the large MLST scheme created earlier by clicking on the ( ) icon.

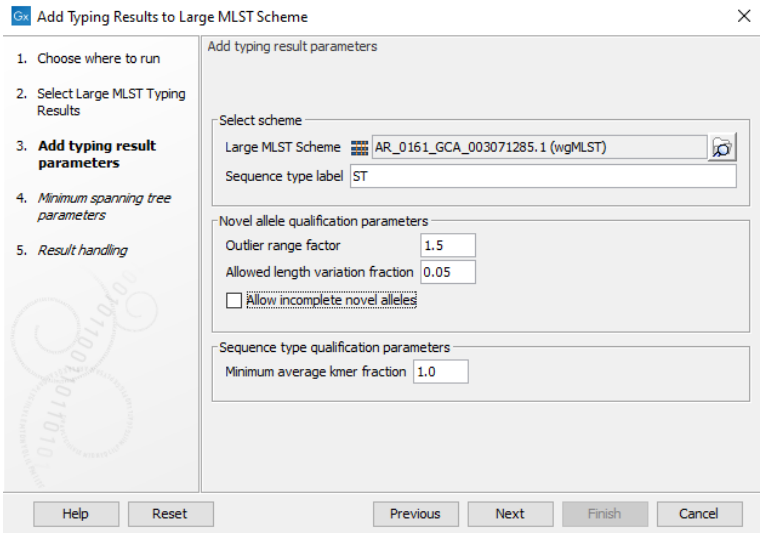

Figure 6: *Select the scheme used for typing*

28. Leave the "Sequence type label" as ST and uncheck the "Allow incomplete novel alleles" option, as shown in figure 6. Click on **Next**.
29. Click on **Next**.
30. Leave the Minimum spanning tree settings as the defaults and click on **Next**.
31. Choose to save the results and click on **Next**.
32. Specify where results should be saved to, for example, in the "wgMLST schemes" folder you made earlier, and click on **Finish**.

An updated large MLST scheme and a report is generated. We will now take a look at the updated MLST scheme.

Inspecting the updated large MLST scheme

A large MLST scheme contains several types of information, as described in the CLC Microbial Genomics Module manual. There are several views available, which we explore here.

The different views are opened by clicking on the small icons in the bottom left corner of the viewing area.

33. Click on the leftmost icon (), at the bottom of the viewing area, to open the **Heat Map** view.
34. Hover the mouse cursor over a particular location in the heatmap to reveal the locus name and sequence type in a tooltip.
35. Click on the second icon from the left (), at the bottom of the viewing area, to open the **Allele Table** view.

    This view shows an overview of the loci. If you used an Antimicrobial Resistance Database, you can see resistance annotations in the locus category column, if any were found.

36. Click on a locus row to select it.

    An overview of the alleles and associated sequence types in that locus are then shown in the lower section of the view.

37. Click on the third icon from the left ( ) to open the **Sequence Type Table** view.

    This provides an overview of the sequence types. You can use this view to help detect outliers, such as sequence types with a low number of loci associated with them.

    There is also a number of additional metadata columns such as "Assembly ID" and "Latin Name". The metadata is from the initial references and was added in **Add Typing Result to Large MLST Scheme**.

    To create a new large MLST scheme containing a subset of the sequence types, just highlight those rows and click on the "Create Large MLST Sub Scheme". For this tutorial, we will proceed with the full scheme.

38. Click on the fourth icon from the left ( ) to open the **Minimum Spanning Tree** view.

    This view is useful for visualizing relationships between strains or isolates.

There are 2 further views available, the **History** and **Element Info** views. These provide general information about the data element. These are described in the ***CLC Genomics Workbench*** manual.

Automating scheme creation using a workflow

In the above section, we stepped through each of the tools needed to create and populate a large MLST scheme. However, the set of steps can be run more efficiently by using a workflow. Such a workflow can be found at:

Microbial Genomics Module ( ) | **Typing and Epidemiology** ( ) | **Workflows** ( ) | **Create Large MLST Scheme With Sequence Types** ( ).

To inspect the workflow, right-click on the workflow in the Toolbox and select "Open Copy of Workflow". The workflow is shown in figure 7. By double-clicking on the tools, it is possible to view and change the settings of the workflow. When you are done, press Ctrl+S (⌘ + S on a Mac) or right-click on the workflow tab and select "Save" to save a copy of the workflow including any modifications made.

The workflow can be run by clicking "Run" in the bottom right corner of the workflow view. You can also run the workflow directly from the toolbox. This is described in the CLC Microbial Genomics Module manual

Feel free to run the workflow if you wish. The workflow outputs 2 schemes; an initial scheme corresponding to the output from "Create Large MLST Scheme" and a final scheme corresponding to the output from "Add Typing Results to Large MLST Scheme".

Note that the Create Large MLST Scheme tool requires that at least one of the input genomes has CDS annotations to serve as the basis for the loci. The data for this tutorial is already annotated. If you wish to use unannotated data as the basis of a large MLST scheme, such as a de novo assembly, you should first annotate it. We describe two ways of doing so in the section (Optional) Annotating genomes for use in creating large MLST schemes

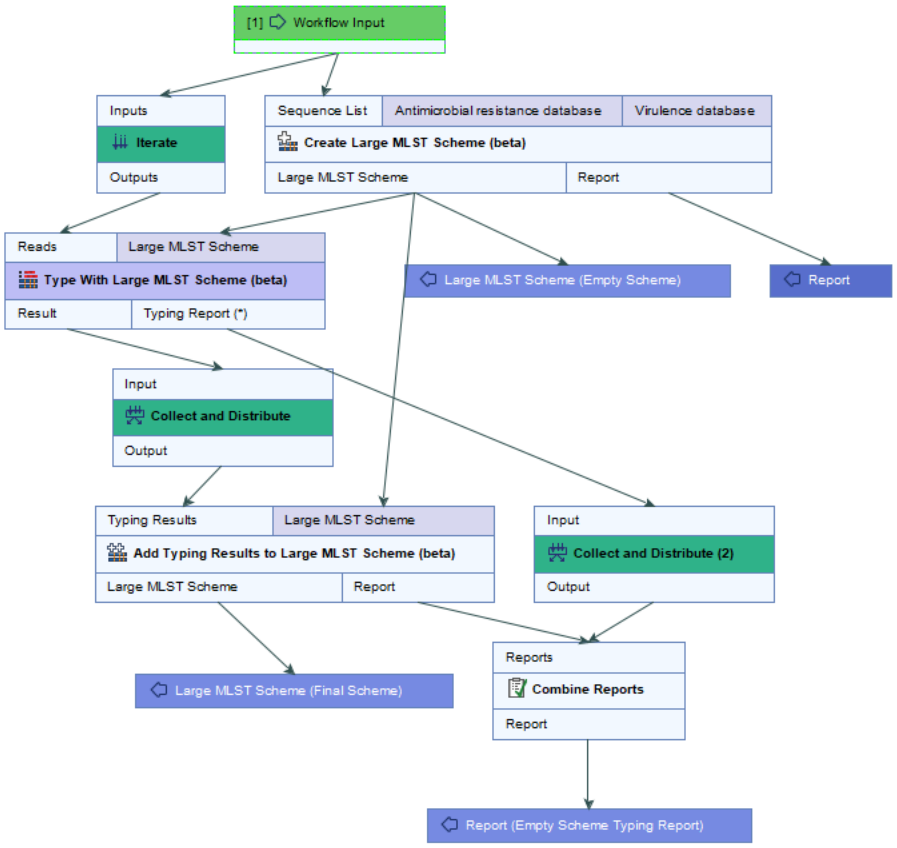

Figure 7: *Select the scheme created in previous steps*

Generating and interpreting typing results

In the following section, we type a set of reads using the updated large MLST scheme just created. We will type reads from a sequence type that is present in that scheme and reads from a sequence type that is not. We will then add the typing results to the large MLST scheme.

Typing using the large MLST scheme

1. From the Toolbox, choose:

   Microbial Genomics Module ( )| **Typing and Epidemiology** ( ) | **Large MLST Typing** ( ) | **Type with Large MLST Scheme** ( )

2. Select the three sequence lists in the "Reads" folder, check the **Batch** box in the bottom, left side of the wizard and click on **Next**.

3. Check the batch units are as you expect.

   Here, you should see the three sequence lists you selected, which means the tool will run three times, once for each of these inputs.

4. Select the updated scheme as the Large MLST Scheme, as shown in figure 8.

5. Set the "Typing threshold" to 0.99.

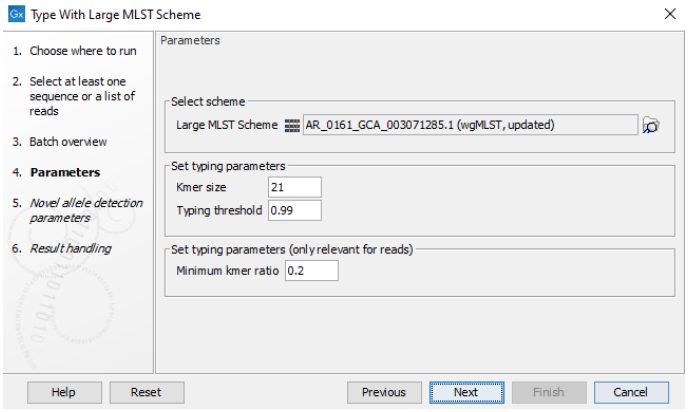

Figure 8: *Select the updated scheme created in previous steps*

As we are working with reads, we have lowered this threshold to accept very closely related hits as conclusive. When typing is not possible, that is, a sample does not match any sequence type in the scheme, this is noted in the report.

6. Click on **Next**.
7. Click on the "Reset" button to set all the novel allele detection parameters back to their defaults, as shown in figure 9, and click on **Next**.

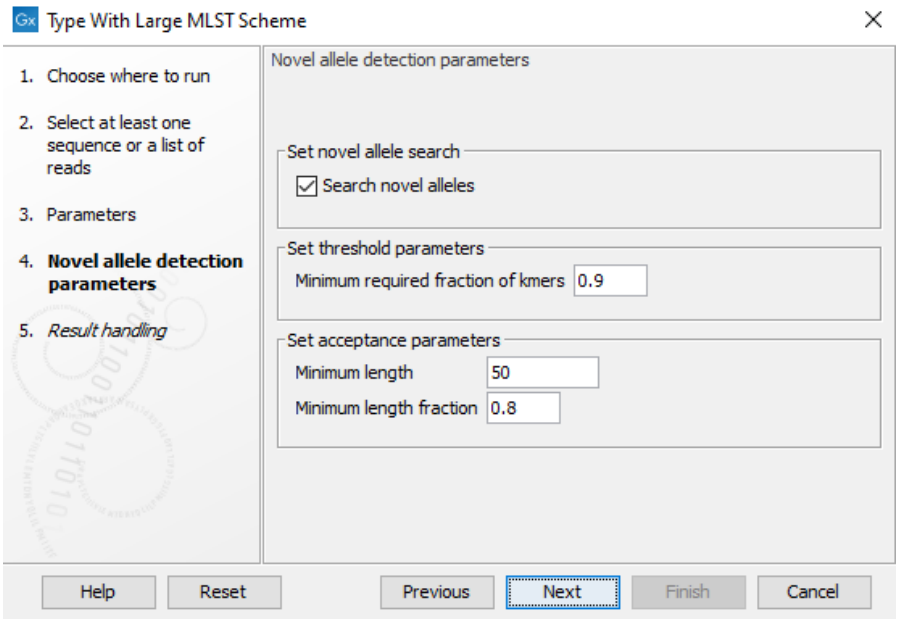

Figure 9: *Reset the novel allele detection settings*

8. Choose to save the results to a location you will specify and click on **Next**.
9. Specify where results should be saved to, for example, to a new subfolder called "Typing Results", and click on **Finish**.

The tool outputs typing results and a report containing information to help you interpret the results. We explore these outputs below, first for strains present in the scheme, and then for a

strain not found in the scheme.

Inspecting the typing results for strains found in the scheme

10. Open the typing reports for the samples SRR2960071 and SRR8268828. These will both show a conclusive typing result.

11. Open the typing result to see the detailed results.

    By default, this opens to the Sequence Type Table view (), where details for each sequence type in the scheme are listed, including average kmer fraction, hit counts, allele count, and alleles identified.

    The most likely sequence type for the sample is based on the average kmer fraction. This number is the average fraction of the number of kmers detected in all alleles for the listed sequence type.

    (a) Open the large MLST scheme.

    (b) Right-click on the tab of the large MLST scheme and select "View" | "Split Horizontally".

12. Switch to the Sequence Type Table () view for the large MLST scheme.

    The open typing result and the large MLST scheme are linked, so selecting information in one of these views will highlight associated information in the other.

13. Select a row in the typing results and click on "Select Sequence Types in Other Views".

    You should see the sequence type(s) selected in both the typing results and the large MLST scheme.

    The different views for each data element are also linked. For example, closely related sequence types can be easily identified using the Heat Map and Minimum Spanning Tree views, as shown in figure 10.

14. Switch to the Show Allele Table () view of the typing result and the large MLST scheme.

15. Look up alleles in the typing result by selecting a locus and then pressing "Select Loci in Other Views".

    You can also use the Heat Map view to look up alleles.

16. Switch to the Typing Result Novel Allele Table () view in the typing result. In this view, novel alleles found are listed. Even though these samples had a conclusive result, there may still be several novel alleles. Listed here are things to consider when evaluating the novel alleles:

    (a) The quality of the assemblies: if the corresponding assembly in the scheme is not a perfect assembly, all alleles may not have been called in a locus.

    (b) The settings used in **Add Typing Result to Large MLST scheme** for "Incomplete novel alleles": This filters out incomplete novel alleles when adding results to a scheme.

    (c) The settings used in **Add Typing Result to Large MLST scheme** for "Outlier range factor" and "Allowed length variation fraction": These affect which alleles are considered outliers and are thus filtered out of the results when adding to a scheme.

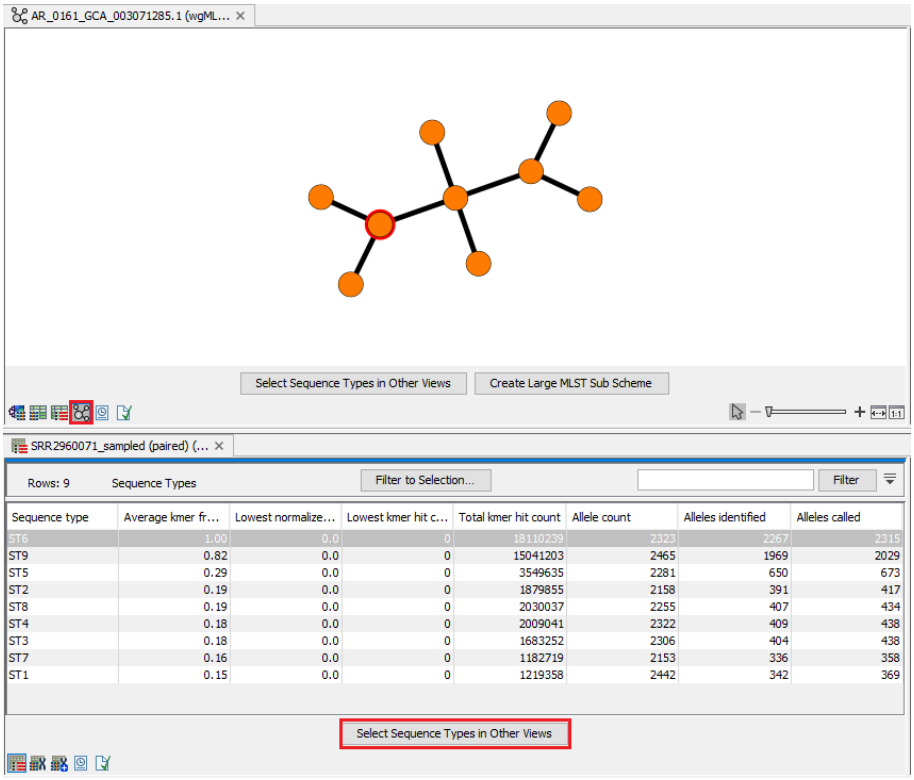

Figure 10: *View the most likely sequence types in the minimum spanning tree*

17. From the Toolbox, choose:

    Microbial Genomics Module ( )| **Typing and Epidemiology** ( ) | **Large MLST Typing** ( ) | **Add Typing Results to Large MLST scheme** ( )

18. Select the typing results for "SRR10905814" as input and click on **Next**.

19. Select the updated large MLST scheme made earlier.

20. Leave the "Sequence type label" as ST and uncheck the "Allow incomplete novel alleles" option.

    You can adjust the "Outlier range factor" and "Allowed length variation fraction" to be more or less strict depending on the organism you are working with. Here, we will leave them as defaults.

21. Click on **Next**.

22. Leave the Minimum spanning tree settings as the defaults and click on **Next**.

23. Choose to save the results and click on **Next**.

24. Specify where results should be saved to and click on **Finish**.

Typing the reads once more using this newly updated scheme would result in the newly added sequence type being identified.

(Optional) Download Large MLST Scheme

You can find and download multiple Large MLST Schemes using the **Download Large MLST Scheme** tool. A number of classic 7-gene schemes are also available.

Classic MLST schemes are a powerful tool for typing but cannot always capture the nuances in closely related strains such as those originating from the same outbreak. In these cases, cgMLST or wgMLST schemes are useful. In the section below, you will try downloading a 7-gene scheme.

1. From the Toolbox, choose:

   Microbial Genomics Module ( )| **Databases** ( ) | **Large MLST** ( ) | **Download Large MLST Scheme** ( )

2. Click on ( ) to select schemes to download.
3. In the search box enter "Klebsiella aerogenes" to search for relevant MLST schemes.
4. Select the Klebsiella aerogenes MLST scheme.

   This is a 7-gene scheme you can use to compare with your wgMLST scheme.
5. Select schemes to download by double-clicking or using the arrow.
6. Click on **Done** and then click on **Next**.
7. Leave the clustering settings with the default values and click on **Next**.
8. Choose whether you would like to create minimum spanning tree and click on **Next**.
9. Choose to save the results and click on **Next**.
10. Choose a location to save the results to and click on **Finish**..

You can run "Type with large MLST" with SRR2960071 and SRR8268828 reads as input to see that these samples cannot be separated with the 7-gene scheme.

(Optional) Export and import Large MLST schemes

Here we step through the export and import of MLST schemes, using the scheme you created as an example.

Exporting large MLST schemes

1. Go to:

   File | **Export**
2. Type "mlst" into the search box at the top.
3. Click on "Large MLST Scheme" in the list and then click on **Select**.
4. Select the large MLST scheme you wish to export and click on **Next**.

5. Specify the filename to export to and click on **Next**.
6. Specify the location to save the file to and click on **Finish**.

The output can then be shared and uploaded to other MLST resources. Note that you can also export schemes as a .clc object for easy sharing with other Workbench users.

Importing large MLST schemes

1. Unzip the zip file you just exported.

   We will use this data to illustrate the import of large MLST schemes.
2. From the Toolbox, choose:

   Microbial Genomics Module ( )| **Databases** ( ) | **Large MLST** ( ) | **Import Large MLST Scheme** ( )
3. For the "Allele folder" option, specify the unzipped scheme folder.
4. For the "Sequence types" option, select the "(schemename)_sequencetypes.txt" file.
5. Optionally, you can add the locus file, which will be called "(schemename)_loci.txt" in your scheme files.

   This file contains metadata for the loci.
6. Click on **Next**.
7. Leave the settings for the clustering parameters set to their default values and click on **Next**.
8. Choose to create a minimum spanning tree and click on **Next**.
9. Choose to save the scheme and click on **Next**.
10. Choose a location to save the results to and click on **Finish**.
11. Open the scheme you just imported.

    It should be identical to the one you exported earlier.

(Optional) Annotating genomes for use in creating large MLST schemes

Two options for adding annotations prior to creating a large MLST Scheme are described below, using **Find Prokaryotic Genes** and using **Annotate with DIAMOND**.

Adding annotations using Find Prokaryotic Genes

We recommend using the *CLC Genomics Server* for this activity, if you are able to.

1. From the Toolbox, choose:

   Microbial Genomics Module ( )| **Functional Analysis** ( ) | **Find Prokaryotic Genes** ( )

2. Select the references you wish to annotate and click on **Next**.
3. Select "Learn one gene model for each assembly" as the "Model training" option.

   If you are using data from the same organism, an alternative is to learn and save one gene model. This can be reused for additional assemblies.
4. Change the assembly grouping to best match your data and leave the other settings with the default values.
5. Click on **Next**.
6. Choose to save the results and click on **Next**.
7. Specify where results should be saved to and click on **Finish**.

Adding annotations using Annotate with DIAMOND

To add annotations using DIAMOND, you need the relevant protein data to be available to refer to, for example, the SwissProt database. Protein databases can be downloaded using the tool:

Microbial Genomics Module ( )| **Databases** ( ) | **Functional Analysis** ( ) | **Download Protein Database** ( ).

Once you have the relevant protein databases available:

1. From the Toolbox, choose:

   Microbial Genomics Module ( )| **Functional Analysis** ( ) | **Annotate with DIAMOND** ( ).
2. Select as input the sequences you wish to have annotations added to and click on **Next**.
3. Select "Protein Sequence List" as the reference sequence type, and select the protein database you wish to use, as shown in figure 11.

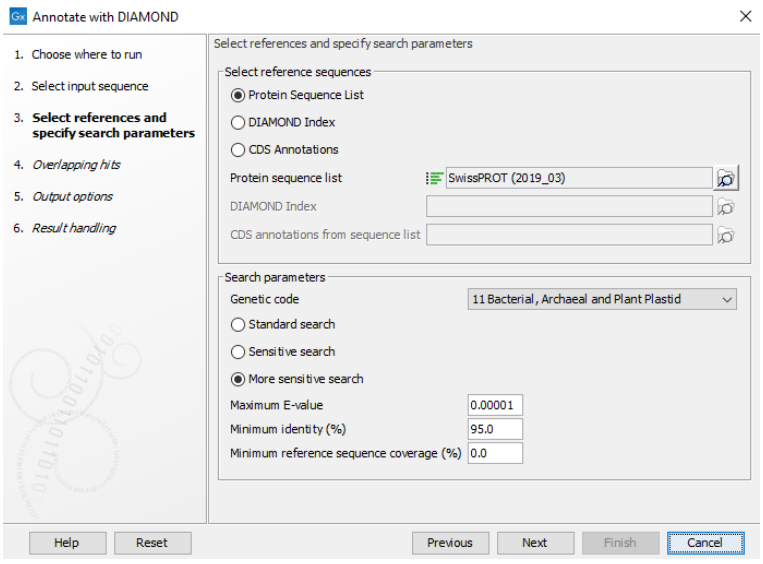

Figure 11: *Annotate with DIAMOND using a protein sequence list*

4. Adjust the "Minimum Identity (%)" to match the input database. Since we are looking for genes without needing to be specific, you should lower the Minimum Identity to match the input database.

   For example, when using SwissProt, the "Minimum Identity (%)" can be lowered to 80%.

   When using clustered databases (e.g. UniRef50 is clustered at the 50% sequence identity level), the "Minimum identity (%)" option for the Annotate with DIAMOND tool should be adjusted to match that clustering level.

5. Leave other settings as the default values and click on **Next**.
6. Leave hit-related settings as the default values and click on **Next**.
7. Leave the output settings as the default values and click on **Next**.
8. Choose to save the results and click on **Next**.
9. Specify where results should be saved to and click on **Finish**.