BioNumerics Tutorial:

MLST analysis starting from whole genome sequences

1 Introduction

With the functionality present in the *Sequence extraction plugin* subsequences can be extracted from (whole genome) sequences and stored in BioNumerics. Any subsequence can be searched for (resistance gene sequences, virulence gene sequences, etc) and used for more in-depth study.

In this tutorial we will search for the sequences corresponding to the seven housekeeping genes used in the online MLST scheme of *Listeria monocytogenes* hosted at `http://bigsdb.web.pasteur.fr/`. The different steps are illustrated using the whole genome demonstration database of *Listeria monocytogenes*. This database is available for download on our website (see 2) and contains 51 publicly available sequence read sets of *Listeria monocytogenes* with already calculated de novo assemblies.

2 Preparing the database

2.1 Introduction to the demonstration database

The whole genome demonstration database of *Listeria monocytogenes* can be downloaded directly from the *BioNumerics Startup* window (see 2.2), or restored from the back-up file available on our website (see 2.3).

2.2 Option 1: Download demo database from the Startup Screen

1. Click the ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window.

This calls the *Tutorial databases* window (see Figure 1).

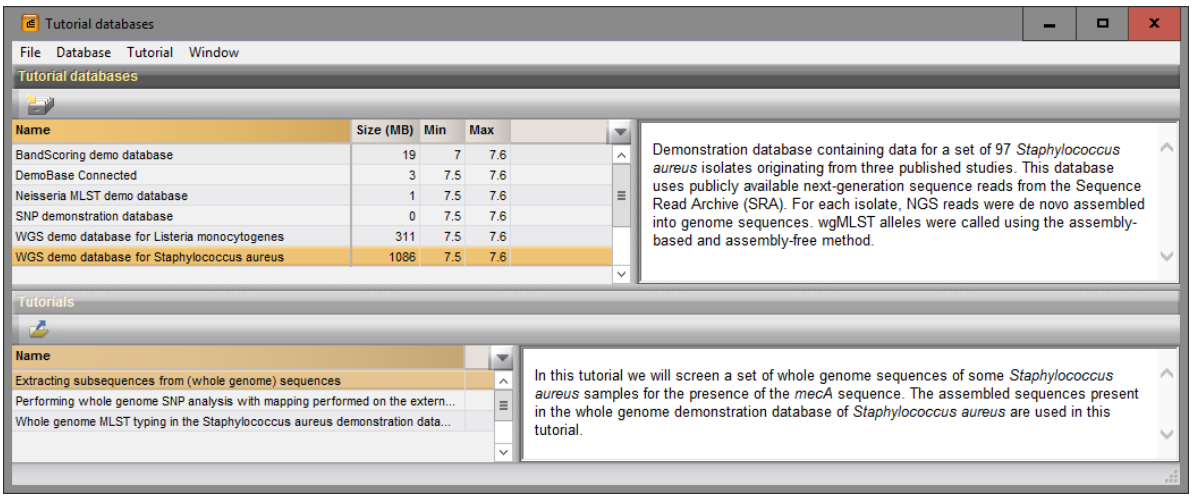

Figure 1: The *Tutorial databases* window, used to download the demonstration database.

2. Select the **WGS demo database for Listeria monocytogenes** from the list and select ***Database*** > ***Download*** ().

3. Confirm the installation of the database and press <***OK***> after successful installation of the database.

4. Close the *Tutorial databases* window with ***File*** > ***Exit***.

The **WGS demo database for Listeria monocytogenes** appears in the *BioNumerics Startup* window.

5. Double-click the **WGS demo database for Listeria monocytogenes** in the *BioNumerics Startup* window to open the database.

2.3 Option 2: Restore demo database from back-up file

A BioNumerics back-up file of the WGS demo database for Listeria monocytogenes is also available on our website. This backup can be restored to a functional database in BioNumerics.

6. Download the file `wgMLST_LMO.bnbk` file from `http://www.applied-maths.com/download/sample-data`, under 'WGS demo database for Listeria monocytogenes'.

In contrast to other browsers, some versions of Internet Explorer rename the `wgMLST_LMO.bnbk` database backup file into `wgMLST_LMO.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

7. In the *BioNumerics Startup* window, press the  button. From the menu that appears, select ***Restore database...***.

8. Browse for the downloaded file and select ***Create copy***. Note that, if ***Overwrite*** remains selected, an existing database will be overwritten.

9. Specify a new name for this demonstration database, e.g. "WGS Listeria demobase".

10. Click <***OK***> to start restoring the database from the backup file (see Figure 2).

11. Once the process is complete, click <***Yes***> to open the database.

The *Main* window is displayed (see Figure 3).

3 About the demonstration database

The whole genome demonstration database of *Listeria monocytogenes* contains links to sequence read set data on NCBI's sequence read archive (SRA) for 51 publicly available sequencing runs. The sequence read set experiment type **wgs** contains the link with some raw data statistics.

1. Click on the green colored dot for one of the entries in the first column in the *Experiment presence* panel. Column 1 corresponds to the first experiment type listed in the *Experiment types* panel, which is **wgs** in the default configuration.

In the *Sequence read set experiment* window, the link to the sequence read set data on NCBI (SRA) with a summary of the characteristics of the sequence read set is displayed: *Read set size*, *Sequence length statistics*, *Quality statistics*, *Base statistics* (see Figure 4).

2. Close the *Sequence read set experiment* window.

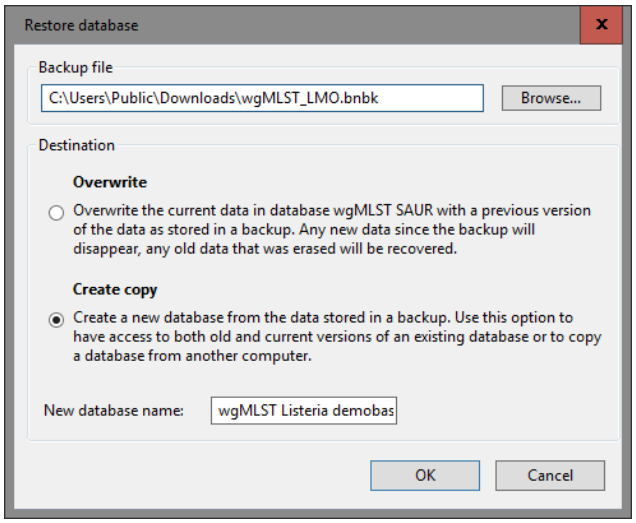

Figure 2: Restoring the WGS demonstration database from the BN backup file `wgMLST_LMO.bnbk`.

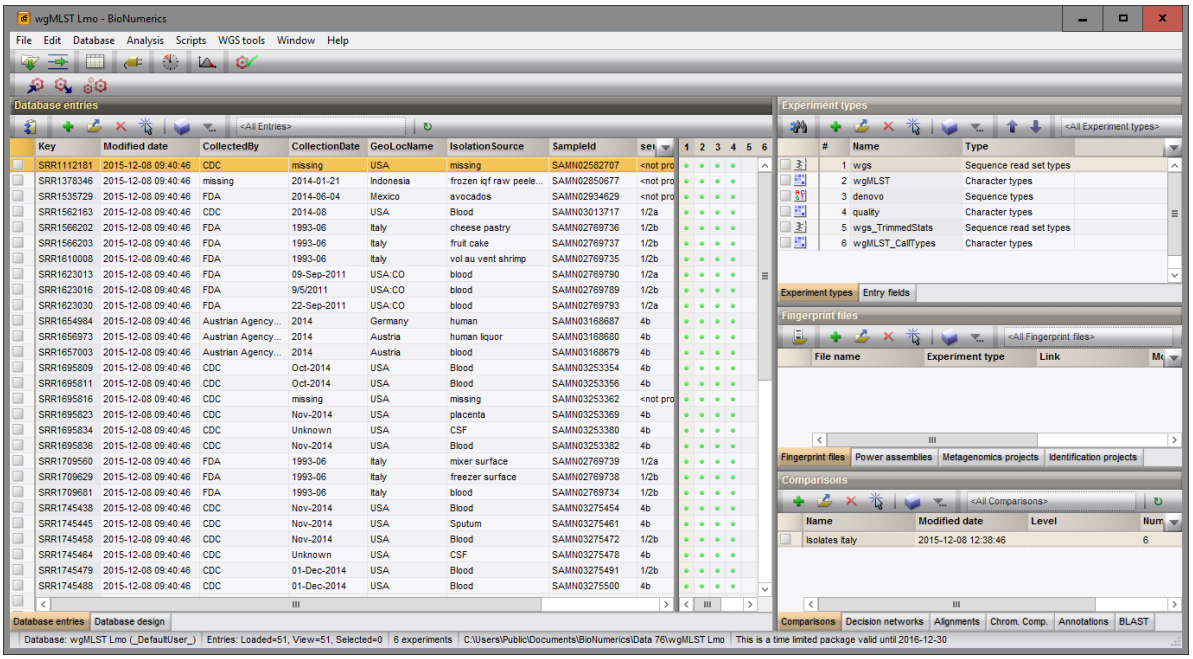

Figure 3: The *Listeria monocytogenes* demonstration database: the *Main* window.

The sequence experiment type **denovo** contains the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences.

3. Click on the green colored dot for one of the entries in the third column in the *Experiment presence* panel. Column 3 corresponds to the third experiment type listed in the *Experiment types* panel, which is **denovo** in the default configuration.

The *Sequence editor* window opens, containing the results from the de novo assembly algorithm, i.e. concatenated de novo contig sequences (see Figure 5).

4. If not all panels are in place select ***Window*** > ***Restore default configuration*** to restore the default configuration of the *Sequence editor* window.

In this tutorial we will extract subsequences in batch from the sequences stored in the **denovo** sequence

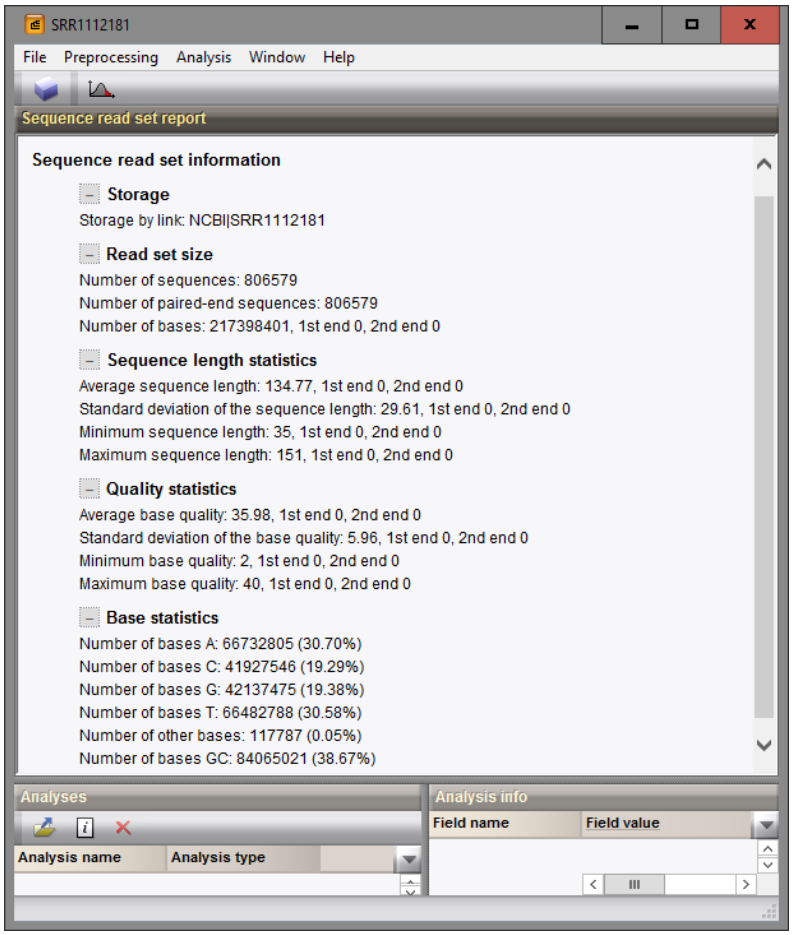

Figure 4: The sequence read set experiment card for an entry.

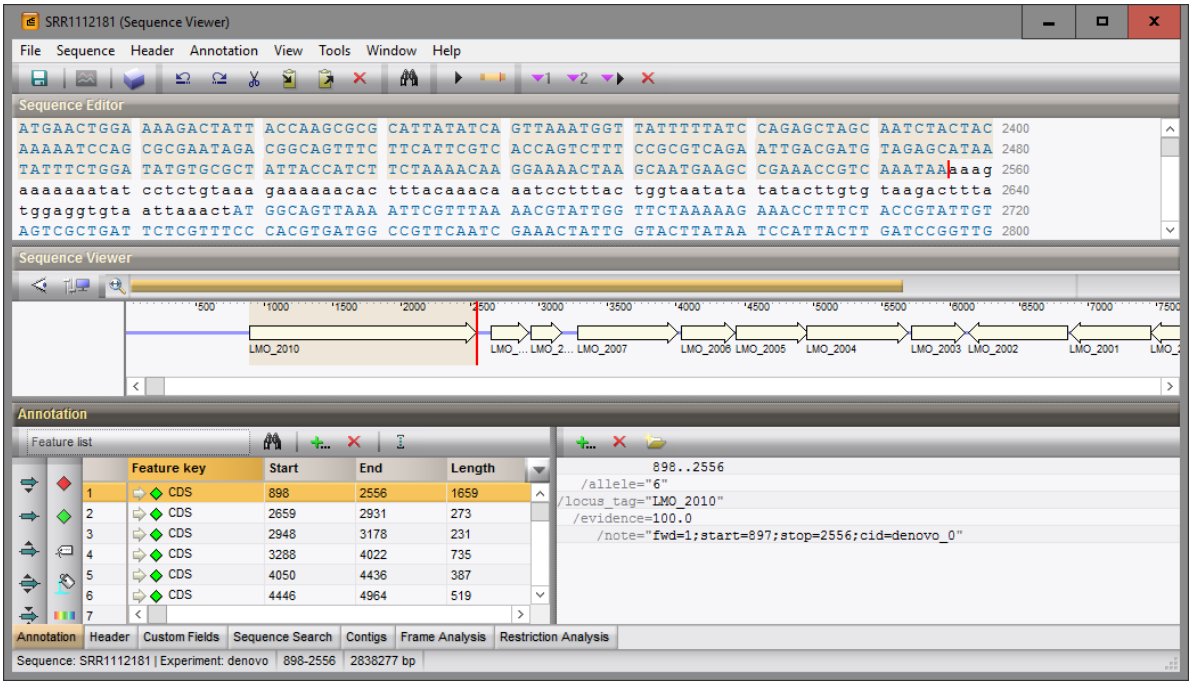

Figure 5: The *Sequence editor* window.

experiment. We will search for the sequences corresponding to the seven housekeeping genes used in the online MLST scheme of *Listeria monocytogenes* hosted at http://bigsdb.web.pasteur.fr/.

5. Close the *Sequence editor* window.

Additional information, stored in entry info fields (CollectionDate, CollectedBy, serovar, etc.) was collected from the corresponding publications and added to the demonstration database.

The wgMLST analysis settings and results (assembly-based calls and assembly-free calls) performed on the Applied Maths Calculation Engine are in depth discussed in the tutorial "Whole genome MLST typing in the Listeria monocytogenes demonstration database" available on our website.

4 Installing the Sequence extraction plugin

In this section we will install the *Sequence extraction plugin* in our demonstration database.

1. The *Plugins* dialog box is called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ).

2. Select the *Sequence extraction plugin* from the list in the *Utilities tab* and press the <***Activate***> button.

The program will ask to confirm the installation of the plugin.

3. Press <***OK***> to continue with the installation of the plugin.

4. When the installation is complete, press <***Exit***> to close the *Plugins* dialog box.

The plugin provides three additional menu items in the *Main* window (see Figure 6).

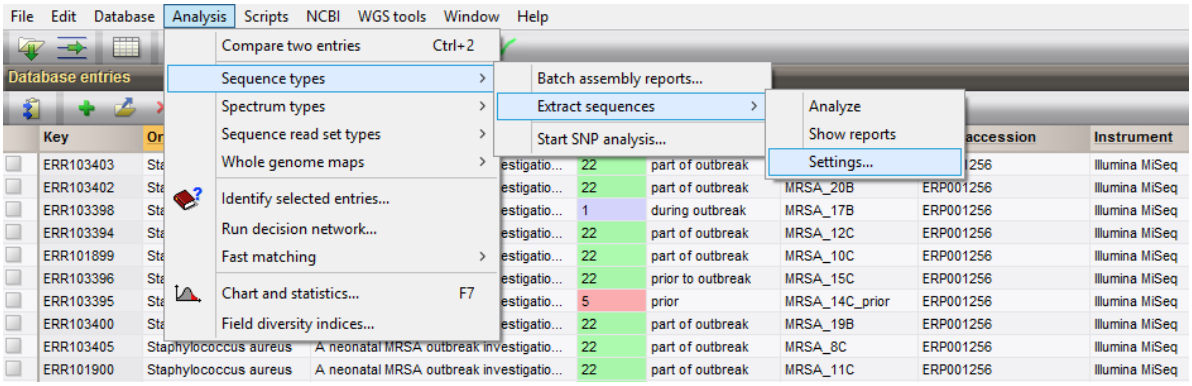

Figure 6: Additional menu items installed by the *Sequence extraction plugin*.

The commands ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Analyze*** and ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Settings...*** can also be executed from the *Process data* dialog box (see Figure 7). This dialog is called via ***File*** > ***Process...*** ( ).

5 Installing the MLST online plugin

In this section we will install the *MLST online plugin* and set it up to use the publicly available *Listeria monocytogenes* MLST schema hosted at http://bigsdb.web.pasteur.fr/. For each housekeeping

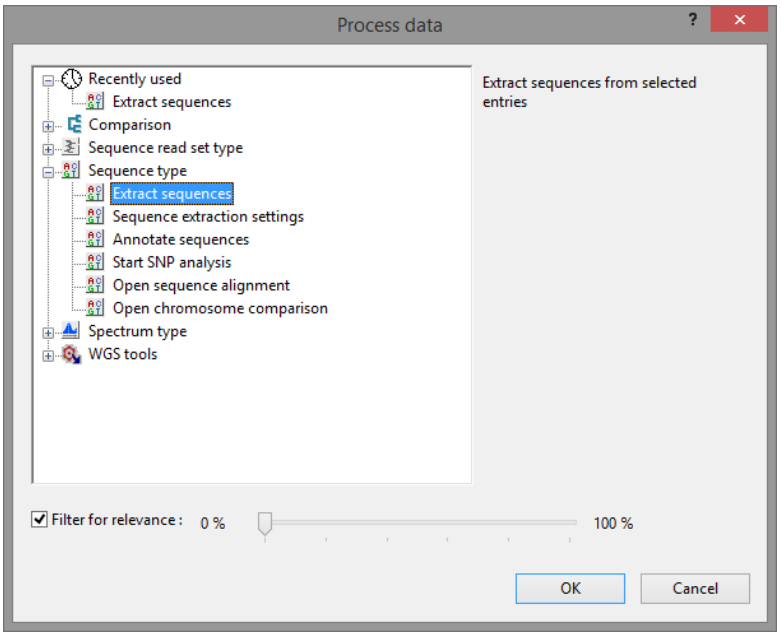

Figure 7: The *Process data* dialog box, displaying the two items (***Extract sequences*** and ***Sequence extraction settings***) that are injected by the *Sequence extraction plugin*.

gene, a sequence type will be created by the plugin that will hold the subsequences, extracted from the whole genome sequences (see 6).

1. The *Plugins* dialog box is called from the *Main* window by selecting ***File*** > ***Install / remove plugins...*** ( ).

2. Select the *MLST online plugin* from the list in the *Applications tab* and press the <***Activate***> button.

The next dialog asks to confirm the installation of the *MLST online plugin*. Installation of the plugin requires administrator privileges on the relational database.

3. Press <***Yes***> to confirm the installation of the *MLST online plugin*.

Since *Listeria monocytogenes* has an MLST repository online we will link our BioNumerics database to this online database.

4. Choose the option ***Select organism from on-line list*** and press <***Next***>.

Any organism for which an MLST repository is available online, is listed in the next step.

5. Select *Listeria monocytogenes* from the list (see Figure 8) and press <***Next***>.

The option ***Update profiles and alleles at database startup*** can be checked, to avoid having to do a manual update.

6. Press <***Next***> to continue.

7. In the next step, leave ***Calculate trimming patterns automatically*** checked and press <***Next***>.

In the final step, the program prompts for database information fields to store the ***Sequence types*** and ***Clonal complexes*** information.

8. For this exercise, use the default ”MLST ST” and ”MLST CC” fields and press <***Next***>.

9. Pressing <***Finish***> starts with the installation of the *MLST online plugin*.

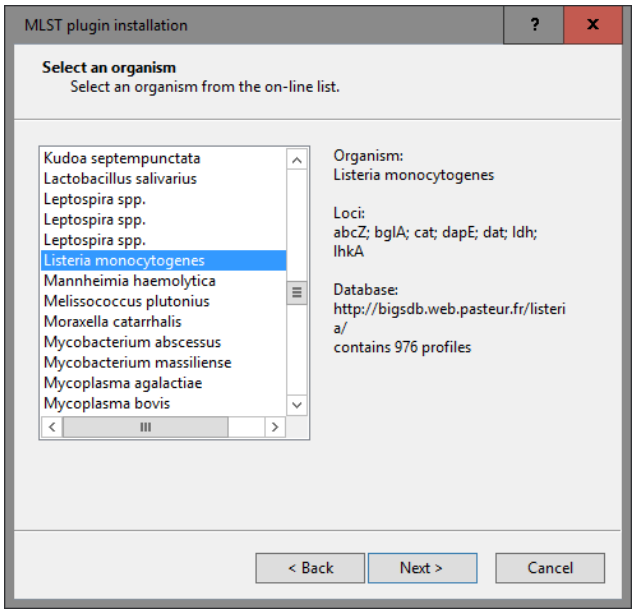

Figure 8: Online MLST repository for Listeria monocytogenes.

The online profile page of *Listeria monocytogenes* contains the sequence type numbers (and corresponding allele numbers), the clonal complexes, and the lineage information (see Figure 9). An extra dialog will pop up asking to link this information to the corresponding BioNumerics MLST ST and MLST CC information fields.

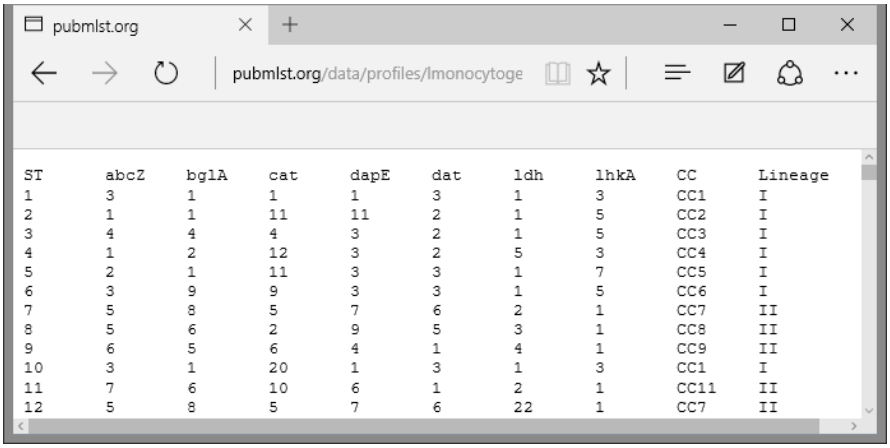

Figure 9: The online profile information for *Listeria monocytogenes*.

10. Select the ***ST*** and ***CC*** fields (see Figure 10) and press <***OK***>.

All remotely stored MLST information (allele numbering, sequence types and clonal complexes) is downloaded in the BioNumerics database during installation of the plugin. This might take several minutes. When querying for allele numbers, sequence types and clonal complexes in BioNumerics, this locally stored information will be used.

When the *MLST online plugin* is successfully installed, a confirmation message is displayed.

11. Press <***OK***> to close this message and press <***Exit***> to close the *Plugins* dialog box.

12. Close and reopen the database to activate the features of the *MLST online plugin*.

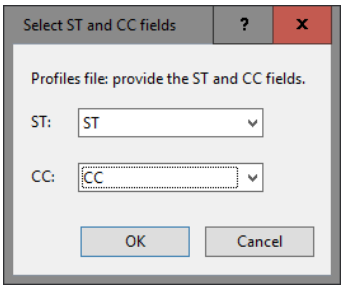

Figure 10: Select ST and CC fields.

The *MLST online plugin* installs menu items in the main menu of the software under ***MLST*** (see Figure 11). In the *Main* window, the *MLST online plugin* has installed following items:

- Extra information fields in the *Database entries* panel (default names: **MLST ST**; **MLST CC**).
- One character type called **MLST**, one composite dataset called **MLST_CMP**, and seven sequence types, each named after a housekeeping gene.

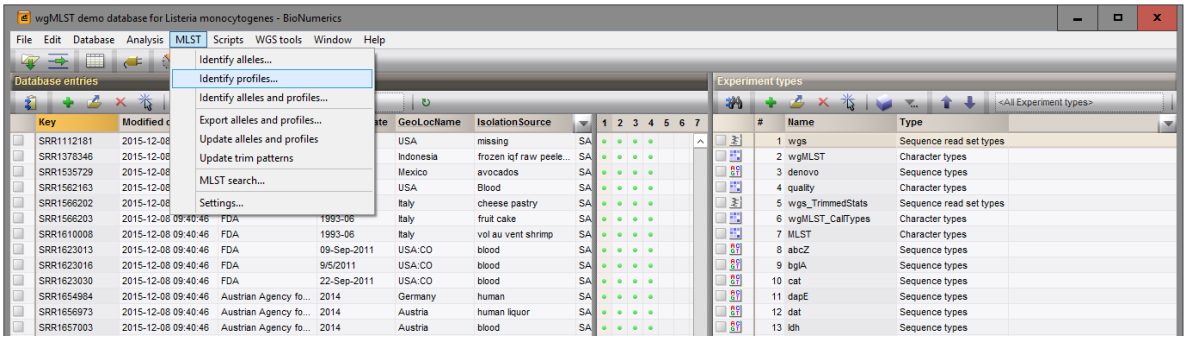

Figure 11: The *Main* window after installation of the plugin.

6 Extracting subsequences

6.1 Principles

The *Sequence extraction plugin* uses a BLAST approach to extract subsequences in batch from sequences stored in an ***Origin*** experiment type and saves the retrieved subsequences in a ***Destination*** experiment type. The BLAST search is based on a single *query sequence* per destination experiment type.

Before we can extract subsequences from the sequences stored in the **denovo** experiment, we first need to specify a query sequence in our demonstration database for each housekeeping gene experiment (see 6.2), and specify the sequence extraction settings (see 6.3).

6.2 Provide query sequences

A FASTA formatted text file can be found on our website, containing a sequence for each housekeeping gene used in the publicly available MLST scheme of *Listeria monocytogenes*. Each sequence corresponds to the allelic sequence with allele number 1 as defined in the online repository (see Figure 12). The example

Figure 12: MLST query sequences.

file can be found on the download page on our website (http://www.applied-maths.com/download/sample-data, "MLST query sequences").

In order to use these sequences as query sequences we first need to import these sequences in our demonstration database.

1. Select ***File*** > ***Import...*** (, **Ctrl+I**) to open the *Import* dialog box.

2. Choose the option ***Import FASTA sequences from text files*** under the ***Sequence type data*** item in the tree and click <***Import***>.

3. Press <***Browse***>, navigate to the downloaded file, select the `Allele1.txt` file and press <***Open***>.

4. With the option ***Preview sequences*** checked, press <***Next***>.

The import wizard now displays a preview of the sequence data in the FASTA file. From this preview, it is clear that the first (and only) FASTA field contains the name of the housekeeping gene experiments (see Figure 13).

5. Press <***Next***>.

The next step of the import wizard lists the templates that are present to import sequence information in the database. As this is the first time we import FASTA formatted sequences in the database, we need to create a new import template by specifying ***Import rules***.

6. Click <***Create new***> to create a new import template.

7. Select ***Field 1*** in the list and click <***Edit destination***> or simply double-click on "Field 1". Select ***Sequence type*** from the list and press <***OK***>.

8. Scroll down the list in the grid using the scroll bar on the right and select the last row in the grid, ***File Name***, and press <***Edit destination***>. Choose ***Key*** and press <***OK***>.

9. Press <***Preview***> to obtain a preview of the data you are about to import (see Figure 14).

10. Close the preview and click <***Next***> and <***Finish***>.

11. Specify a template name, e.g. "FASTA", and optionally enter a description. Press <***OK***>.

12. Highlight the newly created template (see Figure 15) and press <***Next***>.

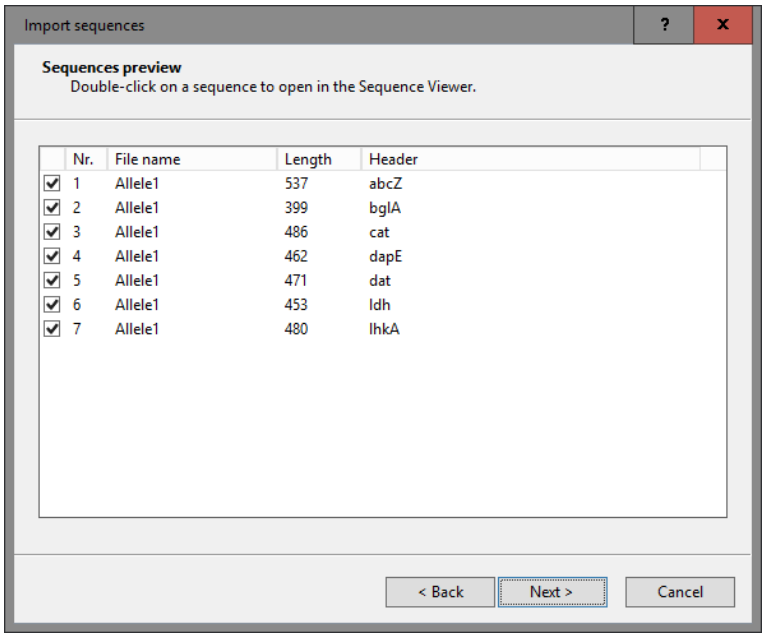

Figure 13: Preview.

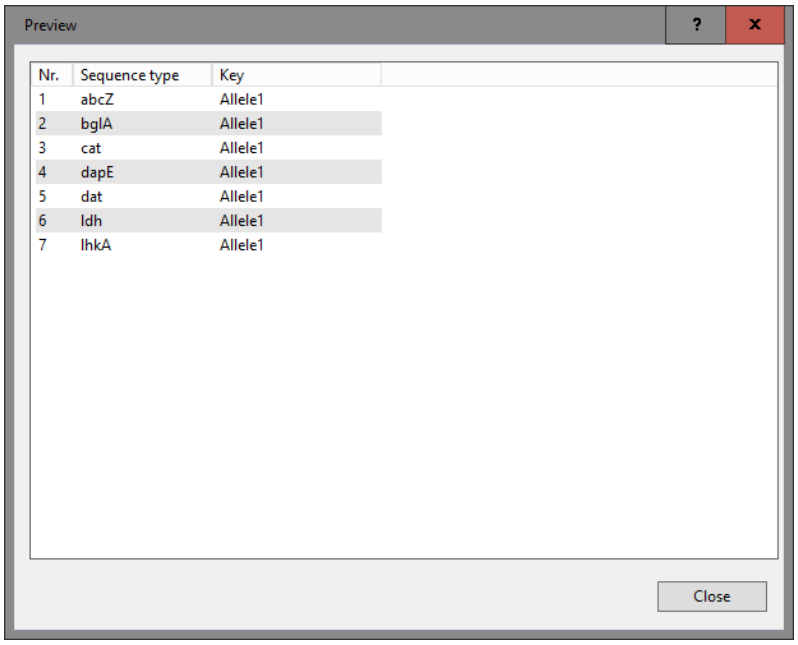

Figure 14: Preview.

The *Database links* wizard page will indicate that 1 new entry will be created during import.

13. Press <***Finish***> to start the import into the database.

An entry with key ***Allele1*** is created in the database and the seven sequences from the text file are linked to the corresponding housekeeping gene experiments of this entry (see Figure 16).

6.3 Specify sequence extraction settings

14. Select ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Settings...*** in the *Main* window to call the *Sequence extraction settings* dialog box.

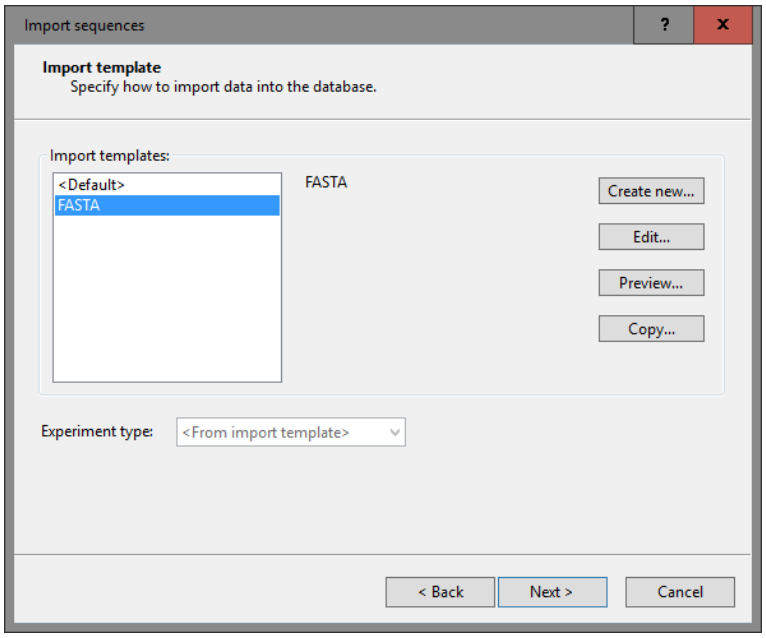

Figure 15: Import template.

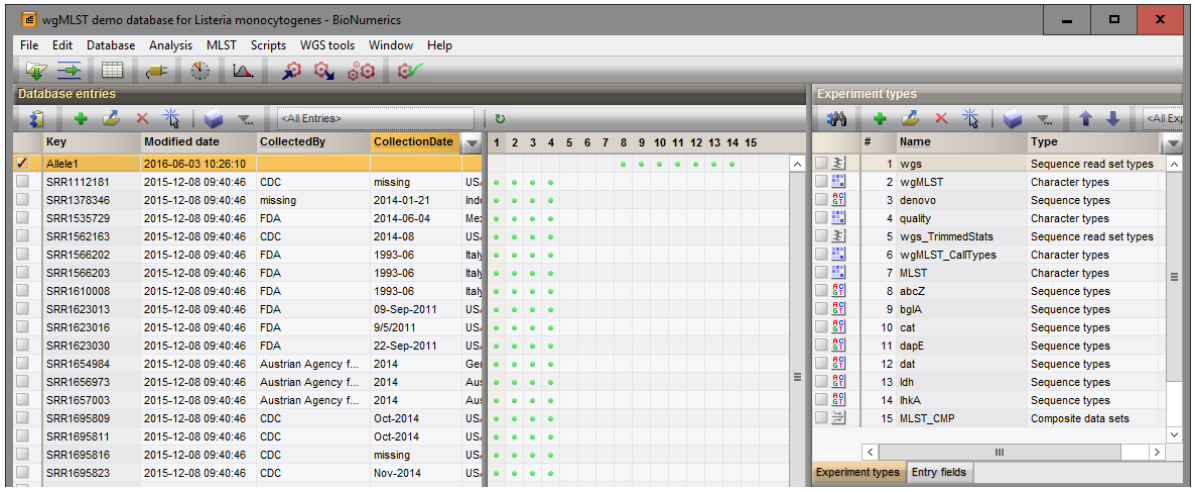

Figure 16: The *Main* window after import of the query sequences.

This dialog box gives access to ***Sequence extraction*** settings per sequence experiment type and the general ***Reports*** settings. Initially, the tree control on the left will be empty.

15. Press the <***Add***> button to call the *Extract gene from genome* dialog box.

16. Select ***denovo*** as ***Origin experiment type*** (see Figure 17). This is the sequence experiment, containing the whole genome sequences, that will be screened and from which a subsequence will be copied from.

17. Check all seven housekeeping gene experiments as ***Destination experiment types*** (see Figure 17).

18. Press <***Next***> to call the second step of the wizard.

The ***Search sequence*** is what the BLAST algorithm will use to screen the origin experiment type (here ***denovo***) for. In our demonstration database, entry with key ***Allele 1*** contains the query sequences, stored in the ***Destination experiment types***.

19. Press <***Pick***> to open the *Select entry* dialog box.

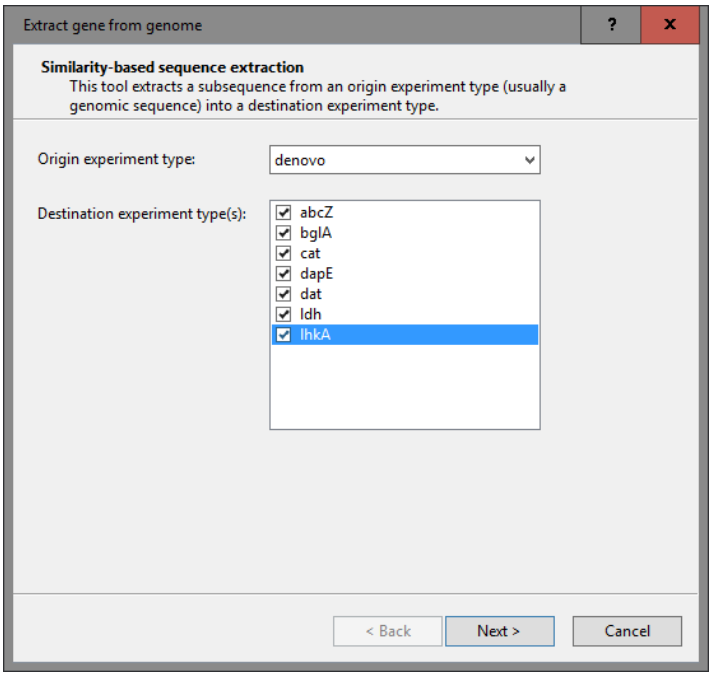

Figure 17: Specify the origin and destination sequence types.

20. Scroll down the list, highlight ***Allele1*** and press <***OK***>. Alternatively, start typing ***Allele1*** in the ***Search for*** text box, highlight ***Allele1*** and press <***OK***>.

The ***BLAST settings*** include two thresholds that a BLAST hit should fulfill in order to be considered:

- A ***Minimum sequence identity (%)*** between the search sequence and the matched subsequence in the origin sequence experiment, expressed as a percentage.
- A ***Minimum length for coverage (%)***, i.e. a minimum overlap between the search sequence and the matched subsequence.

In case more than one BLAST result is found that fulfills both criteria, the best match will be copied to the destination experiment.

Optionally, the length of the extracted sequence can be corrected (see ***Extracted sequence correction*** options).

21. For this exercise, make sure the ***Allele1*** entry is specify as query entry, leave the other settings at their defaults and press <***Next***>.

The tree in the *Extract gene from genome* dialog box is updated (see Figure 18).

Default report settings will be applied when running a report (see 6.4), but can be modified by highlighting ***Reports*** in the tree and pressing <***Edit***>.

22. Press <***OK***> to close the *Extract gene from genome* dialog box.

6.4 Sequence extraction analysis

Now that we have specified the sequence extraction settings (see 6.3), we can now start the actual sequence extraction process.

23. Select all entries in the *Main* window with ***Edit*** > ***Select all*** (**Ctrl+A**) and unselect entry with key ***Allele1***.

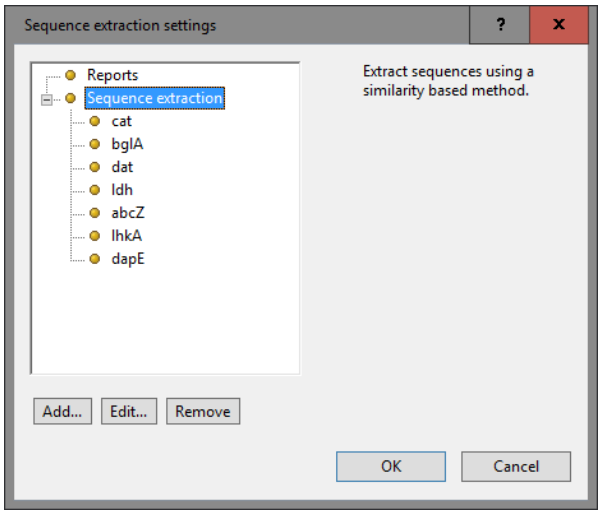

Figure 18: Sequence extraction settings.

The status bar, displayed at the bottom of the *Main* window will indicate that **51** entries are selected.

24. Select ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Analyze*** or use the *Process data* dialog box: select ***File*** > ***Process...*** (), highlight ***Extract sequences*** under ***Sequence type*** and press <***OK***>.

A progress bar appears. The complete analysis may take up to several minutes. When the analysis is finished, the question "Do you want to open the reports?" pops up.

25. Press <***Yes***> to open the *Report* window. Alternatively, a sequence extraction report can be opened for the selected entries with ***Analysis*** > ***Sequence types*** > ***Extract sequences*** > ***Show reports***.

The *Report* window displays a summary of the extraction results (see Figure 19).

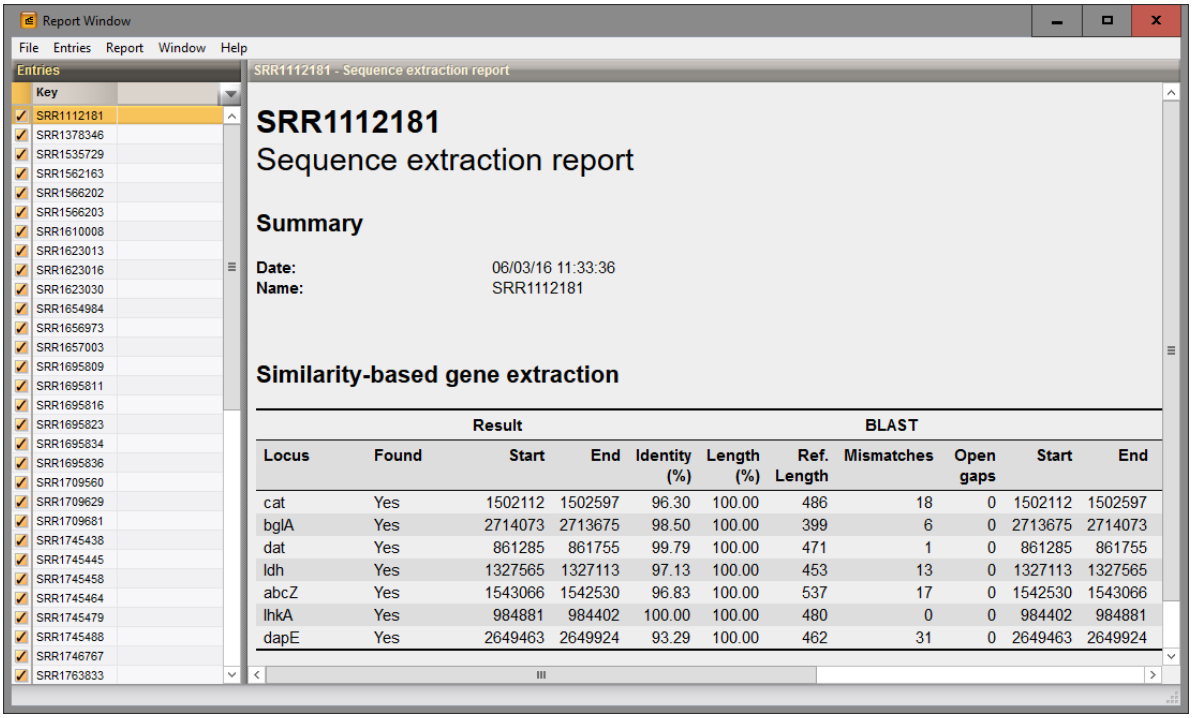

Figure 19: Summary of the sequence extraction results.

The *Report* window contains a gene extraction report for each of the selected entries. For each destination

experiment type ('Locus') that has sequence extraction settings, it is indicated whether or not a BLAST hit was found, its position on the origin sequence ('Start' and 'Stop'), sequence identity ('Identity (%)') and sequence overlap ('Length (%)'). Furthermore, the length of the retrieved subsequence is reported ('Ref length'), the number of mismatches with the query sequence ('Mismatches'), number of gaps ('Open gaps') and length correction applied.

26. Close the *Report* window.

The extracted sequences are stored in the seven corresponding ***destination*** sequence experiment types for all 51 selected entries in our database.

27. Clicking on a green colored dot in the *Experiment presence* panel for one of the housekeeping gene experiments will open the *Sequence editor* window, containing the extracted sequence (see Figure 20).

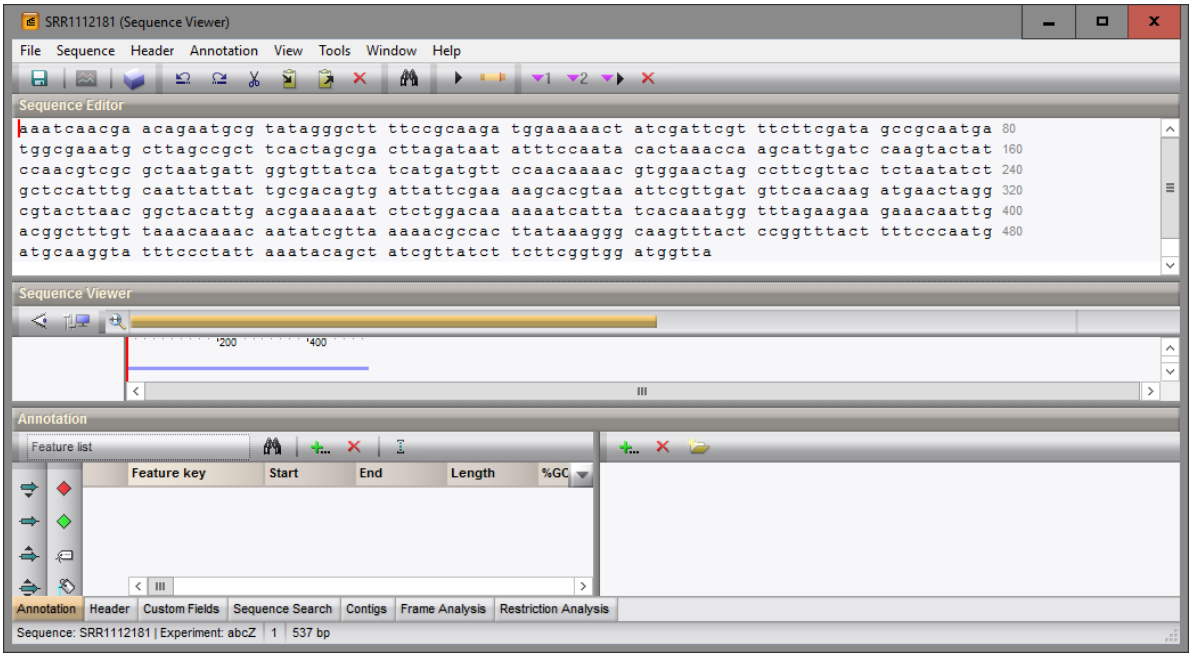

Figure 20: The *Sequence editor* window.

28. Close the *Sequence editor* window.

7 Follow-up analysis

1. Make sure the 51 entries are still selected in the *Main* window and select ***MLST*** > ***Identify alleles and profiles***.

In a first step, the sequences stored in the seven housekeeping gene sequence experiments are screened against the allele information that was downloaded from the online MLST repository of *Listeria monocytogenes* (see 5). The matched allele IDs are stored in the corresponding character fields of the **MLST** character type.

In a second step, the allelic profiles are screened against the downloaded profile information of *Listeria monocytogenes* (see 5). The matched sequences types and clonal complexes are displayed in the MLST information fields (default "MLST ST" and "MLST CC" respectively).

2. Click on the colored dot in the ***MLST*** column of the *Experiment presence* panel to open the character *Experiment card* window for an entry (see Figure 21).

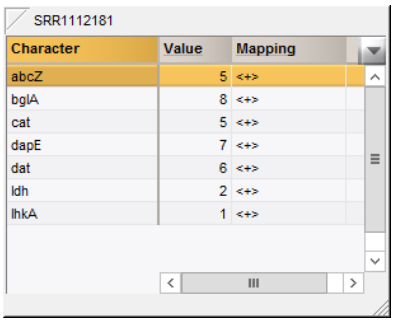

Figure 21: The MLST character card.

3. Close the *Experiment card* window by clicking in the small triangle-shaped button in the left upper corner.

4. Click the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** ( ) to create a new comparison with the 51 selected entries.

A *Comparison* window opens.

5. In the *Comparison* window, right-click in the header of the "MLST ST" field and select ***Create groups from database field*** from the floating menu. Alternatively select ***Groups*** > ***Create groups from database field***.

6. In the *Group creation preferences* dialog box, leave the default settings unaltered and press <***OK***>.

Every ST is now assigned to a unique group. The groups appear in the *Groups* panel along with their color, size and name (see Figure 22).

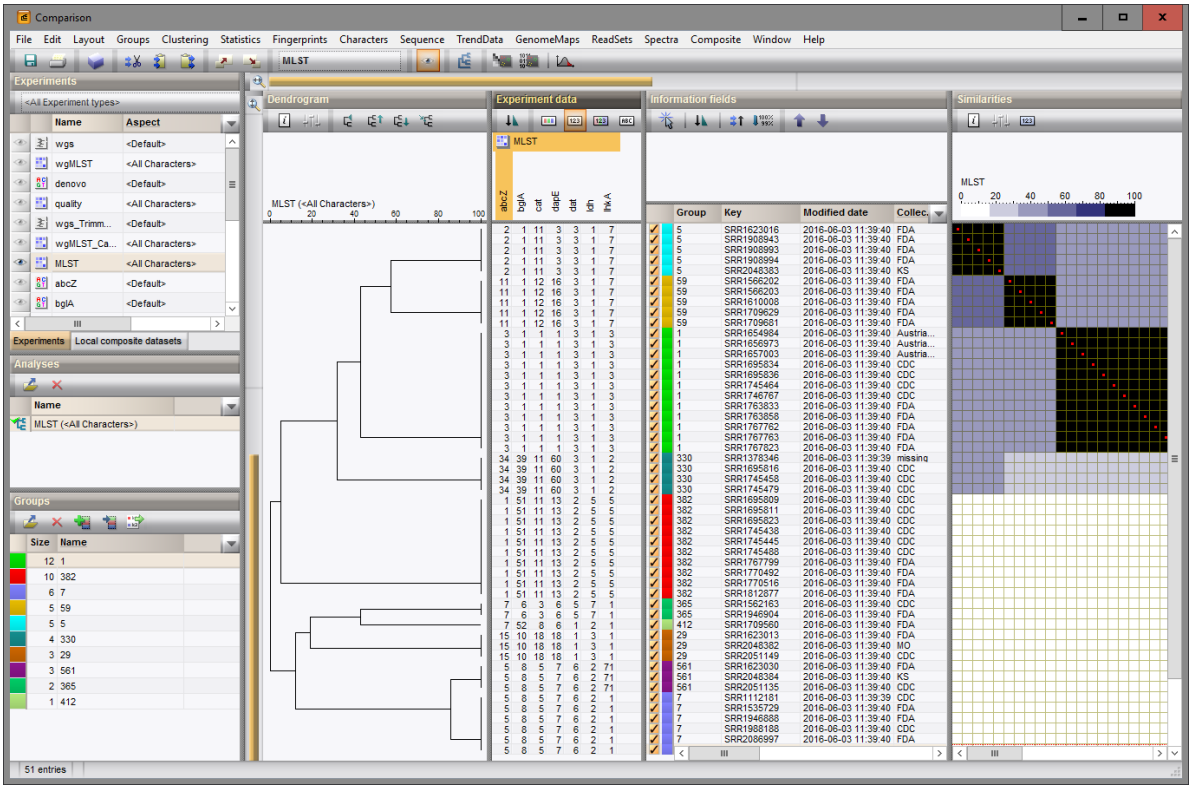

Figure 22: The *Comparison* window with groups defined.

7. Click on the eye icon next to ***MLST*** in the *Experiments* panel and display the values in the *Experiment data* panel with ***Characters*** > ***Show values*** ( ).

8. Make sure ***MLST*** is selected in the *Experiments* panel and select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***.

The first step deals with the similarity coefficient for the calculation of the similarity matrix. Due to the arbitrariness of the allele numbers, the similarity coefficient for clustering MLST data is the categorical coefficient. The categorical coefficient compares the allele numbers to see if they are the same or different but does not quantify the difference.

9. Select ***Categorical (values)*** from the list and press <***Next***>.

10. Select ***UPGMA***, change the ***Dendrogram name*** (e.g. **MLST extract**) and press <***Finish***> to start the cluster analysis.

When finished, the dendrogram and the similarity matrix are displayed in their corresponding panels (see Figure 22). The cluster analysis is listed in the *Analyses* panel of the *Comparison* window.

In our demonstration database a full wgMLST analysis (assembly-based calls and assembly-free calls) was performed on all 51 samples on the Applied Maths Calculation Engine. The character experiment type ***wgMLST*** contains the allele calls for detected loci in each sample, where the consensus from assembly-based and assembly-free calling resulted in a single allele ID.

11. Click on the drop-down list in the **Aspect** column of **wgMLST** in the *Experiments* panel.

Next to the view **All loci**, the **MLST** view has been created by the curator, only containing the seven housekeeping loci used in the online MLST scheme of *Listeria monocytogenes* hosted at `http://bigsdb.web.pasteur.fr/` (see Figure 23).

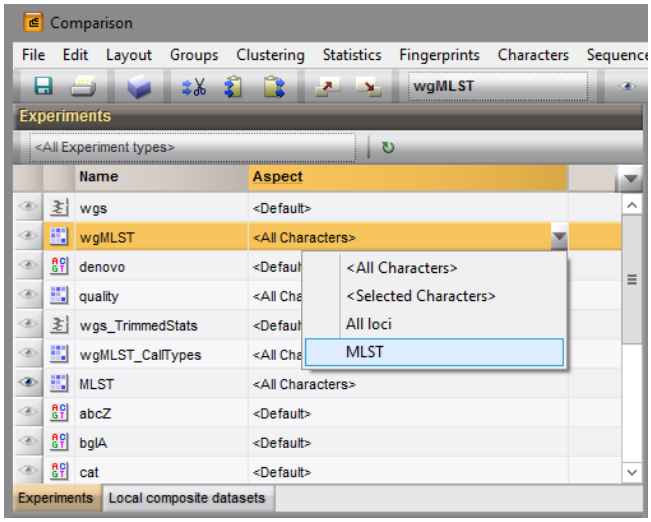

Figure 23: wgMLST character views.

12. Select the **MLST** view from the list.

We can now check if there are differences between the allele IDs present in the ***MLST*** character type (results obtained using the *Sequence extraction plugin* and *MLST online plugin*) and the ***wgMLST*** character type (results obtained using the *WGS tools plugin*).

13. Click on the eye icons next to ***wgMLST*** and ***MLST*** in the *Experiments* panel and display for both experiments the values in the *Experiment data* panel with ***Characters*** > ***Show values*** ( ).

The same values are present in both experiments. We can have an additional check by clustering the MLST data in the ***wgMLST*** experiment:

14. Make sure ***wgMLST*** is selected in the *Experiments* panel, select ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***, select ***Categorical (values)*** from the list and press <***Next***>.

15. Select ***UPGMA***, change the name of the analysis (e.g. **MLST wgMLST**) and <***Finish***> to start the cluster analysis.

The same tree as displayed in Figure 22 is displayed in the *Comparison* window. Switching between the different dendrograms can be done by clicking on the analyses in the *Analyses* panel.

16. The sequences of a housekeeping gene can be displayed in the *Experiment data* panel by clicking the corresponding eye icon ( ) in the *Experiments* panel.

17. Select ***Sequence*** > ***Multiple alignment...*** ( ) to calculate a multiple alignment based on the sequences present in the highlighted sequence experiment.

18. Clicking on the eye icon next to ***MLST_CMP*** in the *Experiments* panel displays the mutation list of the concatenated MLST sequences in the *Experiment data* panel.

More information about dendrograms, minimum spanning trees and other follow-up analyses on MLST data can be found in the tutorial ”Follow-up analysis of MLST data” which can be found on our website.